



Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ingeniería en Ciencias y Sistemas

MINERÍA DE DATOS PARA LA DETECCIÓN DE PATRONES CRIMINALÍSTICOS EN GUATEMALA

David Alberto García Santisteban

Asesorado por el Ing. César Rolando Batz Saquimux

Guatemala, abril de 2012

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA



FACULTAD DE INGENIERÍA

**MINERÍA DE DATOS PARA LA DETECCIÓN DE PATRONES
CRIMINALÍSTICOS EN GUATEMALA**

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERIA
POR

DAVID ALBERTO GARCÍA SANTISTEBAN

ASESORADO POR EL ING. CÉSAR ROLANDO BATZ SAQUIMUX

AL CONFERÍRSELE EL TÍTULO DE

INGENIERO EN CIENCIAS Y SISTEMAS

GUATEMALA, ABRIL DE 2012

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
FACULTAD DE INGENIERÍA



NÓMINA DE JUNTA DIRECTIVA

DECANO	Ing. Murphy Olympo Paiz Recinos
VOCAL I	Ing. Alfredo Enrique Beber Aceituno
VOCAL II	Ing. Pedro Antonio Aguilar Polanco
VOCAL III	Ing. Miguel Ángel Dávila Calderón
VOCAL IV	Br. Juan Carlos Molina Jiménez
VOCAL V	Br. Mario Maldonado Muralles
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

DECANO	Ing. Murphy Olympo Paiz Recinos
EXAMINADOR	Ing. Edgar Josué González Constanza
EXAMINADOR	Ing. Pedro Pablo Hernández Ramírez
EXAMINADOR	Ing. Oscar Alejandro Paz Campos
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:

MINERÍA DE DATOS PARA LA DETECCIÓN DE PATRONES CRIMINALÍSTICOS EN GUATEMALA

Tema que me fuera asignado por la Dirección de la Escuela de Ingeniería en Ciencias y Sistemas, con fecha marzo de 2011.

David Alberto García Santisteban

ACTO QUE DEDICO A:

- Dios** Padre celestial que me dio la vida, que me ha guiado durante todo este camino y me ha brindado la sabiduría, salud, fortaleza y sobre todo las lecciones necesarias para alcanzar esta meta.
- Mis padres** David Alberto García Ruiz y Kira Judith Santisteban de García, por ser ejemplo de lucha y sacrificio, por darme la herencia más valiosa que es la educación, su amor incondicional, los valores morales, y buenos principios. No tengo palabras para expresarles todo mi agradecimiento, gracias por todo, los amo.
- Mi hijo** David Sebastián y mi sobrina Shirley Elizabeth, por ser esos dos angelitos que han venido a iluminar mi vida.
- Mis abuelos** Que desde el cielo nos protegen y bendicen con su amor, y sé que están celebrando conmigo este éxito.
- Mis hermanos** Steven y Allan García Santisteban por todo el apoyo que me han brindado en los buenos y en los malos momentos, y porque sé que toman este éxito como suyo.

Mis tías

Mirna, María y Soraya por ser como mis madres, las amo.

Mi familia

Por todo el amor y el apoyo recibido. En especial a Paola, Ana Lucía, Héctor, Miguel, Enrique, Luis Orlando, mi cuñada Evelyn por siempre confiar en mí.

Mis amigos

Por la amistad sincera, el trabajo en equipo y todas las experiencias inolvidables que llevamos siempre en nuestra mente y corazones que nos permitirán siempre seguir unidos.

AGRADECIMIENTOS A:

**Universidad de San
Carlos de Guatemala**

Por haber sido la casa de estudios que me permitió formarme académicamente y permitir lograr mis sueños y aspiraciones profesionales.

Facultad de Ingeniería

Por haber sido el lugar que me permitió formarme como profesional, gracias por todas las lecciones y enseñanzas aprendidas.

Mi asesor

Por el apoyo brindado para la finalización exitosa del presente trabajo, y por aportar su conocimiento profesional.

Mis catedráticos

Por el conocimiento transmitido y por tener la virtud de poder educar.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES.....	V
GLOSARIO.....	IX
RESUMEN.....	XI
OBJETIVOS.....	XIII
INTRODUCCIÓN.....	XV
1. PLANTEAMIENTO DEL PROBLEMA	1
1.1 . Introducción	1
1.2. Motivaciones y objetivos superiores	3
1.3. Modelo de Aceptación de la Tecnología (TAM) aplicado a la investigación	3
1.4. La información criminal en Guatemala	6
1.4.1. Las fuentes de información criminal.....	9
1.4.2. Situación actual de la política criminal en Guatemala, en cuanto a la investigación criminal	12
1.4.2.1. La formulación político-criminal de la investigación criminal	13
1.4.2.2. Fines de la investigación criminal (ubicación dentro de la política criminal del Estado).....	20
1.4.3. Información criminal del sistema judicial y penitenciario ..	23
1.4.3.1. Roles de los órganos que intervienen en la investigación criminal	23
1.4.3.2. Funciones del Organismo Judicial	24
1.4.3.3. Funciones del Ministerio Público y de la Policía Nacional Civil	26

1.5.	El sistema informático de control de casos del Ministerio Público....	28
1.5.1.	El problema del tratamiento de la información.....	30
1.5.2.	El problema específico	31
2.	MINERÍA DE DATOS APLICADA A LA INVESTIGACIÓN CRIMINAL	33
2.1.	Minería de datos	33
2.1.1.	Agrupación de datos o <i>clustering</i>	39
2.1.2.	Clasificación de datos.....	40
2.1.3.	Reglas de asociación.....	42
2.2.	Aplicaciones informáticas en el análisis de información criminal....	43
2.2.1.	Técnicas informáticas utilizadas en el análisis de información criminal.....	44
2.2.1.1.	Técnicas geográfico-visuales: el Mapa del Delito.....	44
2.2.1.2.	Técnicas de minería de datos.....	47
2.2.2.	Principales experiencias a nivel mundial	48
2.2.2.1.	Proyecto <i>COPLINK</i>	48
2.2.2.2.	Proyecto <i>OVER</i>	53
2.2.2.3.	Otras experiencias.....	54
3.	SOLUCIÓN PROPUESTA.....	57
3.1.	Solución propuesta al problema del tratamiento de la información	57
3.2.	Herramienta a utilizar para el análisis de la información	57
3.3.	Solución propuesta al problema específico.....	59
3.4.	Algoritmos a utilizar	59
3.4.1.	Algoritmo <i>K-means</i>	60
3.4.2.	Algoritmos de inducción.....	62
3.4.2.1.	Algoritmo <i>ID3</i>	62

3.4.2.1.1.	Descripción de <i>ID3</i>	64
3.4.2.1.2.	Pseudo-código del algoritmo <i>ID3</i> ..	65
3.4.2.1.3.	Limitaciones de <i>ID3</i>	65
3.4.2.2.	Algoritmo <i>C4.5</i>	66
3.4.2.2.1.	Pseudo-código del algoritmo <i>C4.5</i>	67
3.4.2.2.2.	Características particulares de <i>C4.5</i>	68
4.	ANÁLISIS DE DATOS CON WEKA	69
4.1.	Conjunto de datos	69
4.2.	Consolidación de la información en una tabla única	70
4.3.	Selección de los campos de interés	71
4.3.1.	Campos seleccionados	72
4.3.2.	Campos omitidos	87
4.4.	Depuración de registros	88
4.5.	Modificación de los estados originales de cada campo	89
4.6.	<i>Data set</i> definitivo	97
5.	PRESENTACIÓN DEL CASO	99
5.1.	Introducción	99
5.2.	Descripción de las herramientas	101
5.2.1.	Tabla de centroides	101
5.2.2.	Diagramas de Venn	103
5.2.3.	Gráficos de barras	105
5.2.4.	Gráficos de dispersión	107
5.2.4.1.	Gráficos de distribución	107
5.2.4.2.	Gráficos de interrelaciones	109
5.2.5.	Árbol de clasificación	110

5.2.6.	Matrices de confusión	112
6.	RESULTADOS	113
6.1.	<i>Clustering</i>	113
6.1.1.	Tabla de centroides	115
6.1.2.	Diagramas de Venn	115
6.1.3.	Gráficos de barras	117
6.1.4.	Gráficos de dispersión	119
6.1.4.1.	Distribución de los <i>clusters</i> según el atributo lugar.....	119
6.1.4.2.	Distribución de los <i>clusters</i> según el atributo arma.....	120
6.1.4.3.	Distribución de los <i>clusters</i> según el atributo día de la semana	121
6.1.4.4.	Interrelación lugar-arma.....	122
6.1.4.5.	Interrelación hora-tipo-arma	123
6.1.4.6.	Interrelación hora-tipo-arma	124
6.1.5.	Primera interpretación	125
6.2.	Aplicación de c4.5 para la clasificación de los <i>clusters</i>	125
6.2.1.	Selección de atributos	125
6.2.2.	Resultados de C4.5 con todos los atributos	126
6.2.3.	Árbol definitivo	128
	CONCLUSIONES.....	135
	RECOMENDACIONES	137
	BIBLIOGRAFÍA	139

ÍNDICE DE ILUSTRACIONES

FIGURAS

1. Modelo de la investigación(basado en el modelo TAM).....	4
2. Diferentes tipos de mapas del delito delimitando hotspots para el robo de vehículos en Argentina	46
3. Interface <i>software Coplink</i> : formulario de búsqueda de personas.....	49
4. Análisis de Redes Criminales: vínculos entre sospechosos.....	50
5. Análisis de Redes Criminales: subgrupos identificados por <i>Coplink</i>	51
6. Error cuadrático	60
7. Pseudo-código del algoritmo de ID3.....	65
8. Pseudo-código del algoritmo de C4.5.....	67
9. Distribución del campo departamento	73
10. Distribución del campo mes	75
11. Distribución del campo día_mes.....	77
12. Distribución del campo del campo día_semana	78
13. Distribución del campo hora	80
14. Codificación y frecuencias del campo dirección_hecho	82
15. Codificación y distribución del campo arma	83
16. Codificación y distribución del campo sexo	85
17. Codificación y distribución del campo edad.....	86
18. Diagrama de Venn para el <i>cluster 0</i>	104
19. Diagrama de Venn para el <i>cluster 1</i>	104
20. Gráficos de barras: distribución de los <i>clusters</i> según el resto de los atributos.....	106
21. Gráfico de distribución de los <i>clusters</i> según día de la semana.....	108

22.	Gráfico de interrelación arma-lugar	109
23.	Árbol de clasificación	110
24.	Matriz de confusión	112
25.	Diagrama de Venn para atributos categóricos del <i>cluster</i> 1	116
26.	Diagrama de Venn para atributos categóricos del <i>cluster</i> 0	116
27.	Distribución de los clusters según atributos	118
28.	Distribución de clusters según atributo lugar	119
29.	Distribución de clusters según atributo arma	120
30.	Distribución de clusters según atributo día de la semana	121
31.	Interrelación lugar-arma	122
32.	Interrelación hora-tipo_arma	123
33.	Interrelación sexo-tipo_arma.....	124
34.	Matriz de confusión	127
35.	Estructura del árbol generado con C4.5.....	128
36.	Referencias del árbol generado con C4.5.....	129

TABLAS

I.	Cantidad de homicidios por año cometidos en el país	1
II.	Fuentes de información criminal	11
III.	Distribución del campo departamento	72
IV.	Codificación y frecuencias del campo mes	74
V.	Codificación y frecuencias del campo día_mes	76
VI.	Codificación y frecuencias del campo día_semana	78
VII.	Codificación y frecuencias del campo hora	79
VIII.	Codificación y frecuencias del campo dirección_hecho	81
IX.	Codificación y frecuencias del campo arma	83
X.	Codificación y frecuencias del campo sexo	84
XI.	Codificación y frecuencias del campo edad	86

XII.	Nuevos estados del atributo departamento91
XIII.	Nuevos estados del atributo mes	92
XIV.	Nuevos estados del atributo día de la semana.....	93
XV.	Nuevos estados del atributo hora	94
XVI.	Nuevos estados del atributo lugar	95
XVII.	Nuevos estados del atributo arma	96
XVIII.	Nuevos estados del atributo sexo.....	96
XIX.	Descripción de atributos de la muestra	100
XX.	Asignación del <i>cluster</i> al data set	101
XXI.	Tabla de centroides	102
XXII.	Resultado de <i>K-means</i> para 3 <i>clusters</i> con varias semillas	114
XXIII.	Tabla de centroides	115
XXIV.	Resultado de selección de atributos.....	126
XXV.	Combinación de atributos.....	127

GLOSARIO

<i>Algoritmo</i>	Conjunto de reglas o instrucciones, ordenadas y finitas que permiten realizar una actividad mediante la ejecución de pasos sucesivos.
<i>Clustering</i>	Procedimiento utilizado para clasificar un conjunto de elementos de muestra en un determinado número de grupos basándose en las semejanzas y diferencias existentes entre los componentes de la muestra.
<i>Criminalística</i>	Conjunto de técnicas y procedimientos de investigación cuyo objetivo es el descubrimiento, explicación y prueba de los delitos, así como la verificación de sus autores y víctimas.
<i>Java</i>	Lenguaje de programación orientado a objetos, proporciona a los programadores un entorno de desarrollo completo, así como una infraestructura.
<i>Minería de datos</i>	Conjunto de técnicas y herramientas utilizadas para la preparación, el sondeo y exploración de grandes volúmenes de datos y encontrar información oculta en ellos.

Patrón	Conjunto de sucesos u objetos recurrentes, a veces referidos a un conjunto de datos, los cuales se repiten de manera recurrente.
<i>PNUD</i>	Programa de las Naciones Unidas para el desarrollo en Guatemala.
<i>Software</i>	El sistema operativo que básicamente, permite al resto de los programas funcionar adecuadamente, facilitando también la interacción entre los componentes físicos y el resto de las aplicaciones, proporcionando una interfaz para el usuario.
<i>WEKA</i>	Herramienta utilizada para el análisis de datos, desarrollada por la Universidad de Waikato de Nueva Zelanda

RESUMEN

La detección de patrones criminalísticos puede ser de gran ayuda para disminuir los índices de criminalidad existentes en el país, por medio de la generación de nuevo conocimiento mediante herramientas más potentes que la estadística descriptiva, es importante y necesario encontrar información valiosa y oculta dentro de los grandes volúmenes de datos que almacenan actualmente los sistemas de información transaccional.

El enfoque del presente estudio consiste en una aplicación práctica de minería de datos para la detección de patrones criminalísticos por medio de herramientas informáticas disponibles, para ese objetivo como se presenta *Weka*, demostrando los resultados que puede proporcionar este tipo de técnicas al análisis de grandes cantidades de información.

OBJETIVOS

General

Realizar una contribución a la comunidad mediante un aporte práctico de la aplicación de minería de datos a una base de información, para la modernización de las prácticas de análisis de información criminal en Guatemala.

Específicos

1. Proporcionar información importante sobre la minería de datos como una técnica a considerar para la detección y análisis de información oculta en grandes volúmenes de datos.
2. Aplicar minería de datos a una base de datos de información sobre hechos delictivos cometidos durante el 2008 y demostrar los resultados que se pueden obtener con este tipo de análisis.

INTRODUCCIÓN

El análisis de los registros criminales es fundamental en la prevención del delito. Entre otras cosas, porque permite el diseño de políticas y planes de prevención efectivos. Es por esto que deben buscarse métodos y herramientas modernas que nos permitan obtener mejores resultados al momento de analizar la información. Este contexto requiere un tratamiento más complejo que obliga a evolucionar en el análisis de información criminal.

La minería de datos es una de las herramientas modernas que permiten el análisis de grandes volúmenes de información y permite el descubrimiento de patrones e información oculta que se encuentran en ellos; es por ello, que debe ser utilizada como herramienta para la detección y análisis de patrones criminalísticos.

1. PLANTEAMIENTO DEL PROBLEMA

1.1. Introducción

La situación de violencia e inseguridad que viven los habitantes de Guatemala, es un tema de análisis, el informe estadístico de homicidios registrado en el país entregado por el Programa de Seguridad Ciudadana y Prevención de la Violencia del PNUD de Guatemala, demuestra que el índice de la violencia homicida ha aumentado en promedio a un ritmo superior al 12% por año.

Los datos presentados por el informe se muestran en la tabla I:

Tabla I. **Cantidad de homicidios por año cometidos en Guatemala**

	AÑO 2000	AÑO 2001	AÑO 2002	AÑO 2003	AÑO 2004	AÑO 2005	AÑO 2006
HOMICIDIOS	2 904	3 230	3 631	4 327	4 507	5 338	5 885

Fuente: elaboración propia.

Existen diversas herramientas tecnológicas para el análisis de datos; tales como:

- La estadística descriptiva: técnica de análisis y representación de datos, pero lo hace de una manera básica ya que su estudio está basado en el cálculo de medidas de tendencia central, lo cual hace de sus resultados una aproximación.

- Las redes neuronales que son un paradigma de aprendizaje y procesamiento automático de datos inspirado en la forma del funcionamiento del sistema nervioso de los animales.
- El proceso completo de extracción y análisis de información KDD.
- La minería de datos, entre muchos otros que engloba resultados de investigación, técnicas y herramientas usadas para extraer información útil de grandes bases de datos.

En general, el tamaño de las bases de datos está basado en aspectos como la capacidad y eficiencia de almacenamiento y no en su posterior uso o análisis. Por esta razón, en muchos casos, los registros almacenados son demasiado grandes o complejos como para analizar y superan el alcance de la estadística y otras técnicas de análisis. La minería de datos (*Data Mining*) es un proceso iterativo de búsqueda de información no trivial en grandes volúmenes de datos. Busca generar información similar a la que podría generar un experto humano: patrones, asociaciones, cambios, anomalías y estructuras significativas.

En el caso de la inteligencia criminal, la gran cantidad de información y de variables intervinientes, justifican el uso de herramientas más potentes que la estadística convencional que permitan determinar relaciones multivariadas. La minería de datos aplicada a la inteligencia criminal es un campo bastante nuevo, ha tenido un gran impulso en los últimos años en EEUU y ha habido casos de éxito en países como Argentina.

El objetivo de este estudio es evaluar una implementación de minería de datos en el análisis de información criminal en Guatemala y comprobar, que esta herramienta tecnológica debe ser considerada por su alta efectividad y valor agregado, para la disminución significativa de los índices de criminalidad.

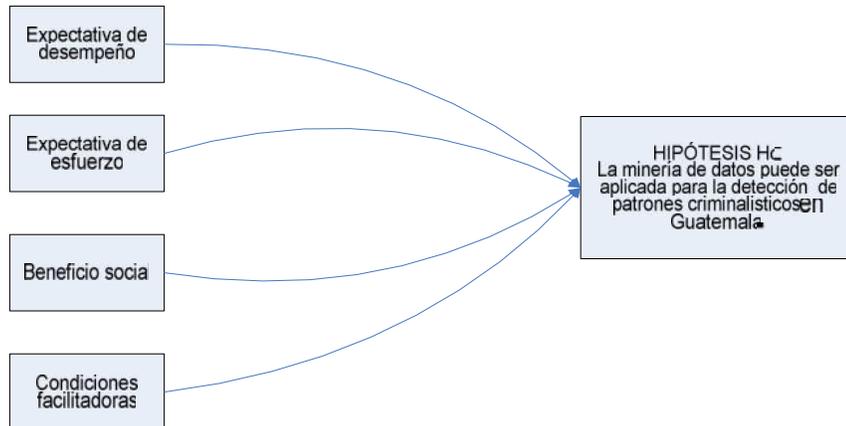
1.2. Motivaciones y objetivos superiores

Contribuir con la comunidad, mediante un aporte para la modernización de las prácticas de análisis de información criminal en Guatemala, demostrando cómo la tecnología puede ser aplicada en el procesamiento y análisis de información criminal para la exploración y detección de patrones criminalísticos, brindando apoyo importante a entidades vinculadas directamente con el sector justicia; tales como el Ministerio Público (MP) y la Policía Nacional Civil (PNC).

1.3. Modelo de aceptación de la tecnología (TAM), aplicado a la investigación

El modelo base utilizado para la investigación es el Modelo de Aceptación de la Tecnología (TAM), como se muestra en la figura 1.

Figura 1. **Modelo de la investigación basado en el TAM**



Fuente: elaboración propia.

- El modelo está compuesto por un factor dependiente y cuatro factores dependientes; la variable dependiente en este caso es:
 - Demostrar que la minería de datos es una herramienta de gran utilidad, para la detección de patrones criminales en Guatemala.
- Esta variable dependiente o hipótesis está basada en el objetivo principal del uso de la minería de datos, que consiste en la extracción no trivial de información que reside implícitamente en los datos, la cual es previamente desconocida, posterior a aplicarle la minería de datos puede presentar información de gran utilidad para cualquier tipo de proceso. Asimismo de la utilización que le dio la Agencia Central de Inteligencia (CIA) a la minería de datos para la detección de patrones relacionados con crímenes y terrorismos.

- Las variables o factores independientes a considerar para este caso son cuatro, las primeras dos corresponden a la percepción de la facilidad del uso; y las siguientes, corresponden a la percepción:
 - Expectativa de desempeño: qué aceptación le darán las entidades encargadas de registrar las estadísticas criminales en el país a la minería de datos como herramienta para la detección de patrones.
 - Expectativa de esfuerzo: el costo-beneficio que representa utilizar cierta herramienta de minería de datos. En este caso el mayor beneficio es el que se obtendría al usar herramientas de *software* libre por todas las ventajas que estas herramientas presentan.
 - Beneficio social: cuánto representa para la población este trabajo de investigación, en el área de criminalística económica y socialmente.
 - Condiciones facilitadoras: qué tan fácil sea aplicar la minería de datos para la detección de patrones criminalísticos en Guatemala. Además de la voluntad que tengan las personas de las entidades del país encargadas de llevar el registro de los crímenes de utilizar esta herramienta como apoyo en el procesamiento de información.

Las variables independientes y el modelo de investigación se basan en el modelo de la aceptación de la tecnología TAM, ya que esta investigación depende de la facilidad de uso y de la utilidad que el producto final pueda presentar.

Por lo tanto, la expectativa de desempeño y de esfuerzo están ligadas a la facilidad de uso que se tenga del proyecto; así como el beneficio social y las condiciones facilitadoras ligadas a la utilidad que puede presentar para la sociedad, en general la aplicación de minería de datos en la detección de patrones criminalísticos.

1.4. La información criminal en Guatemala

La investigación criminal es un conjunto de saberes interdisciplinarios y acciones sistemáticas, integrados para llegar al conocimiento de una verdad relacionada con el fenómeno delictivo.

La investigación criminal ha venido a constituirse como la columna vertebral del proceso penal moderno. Efectivamente, una vez tenido conocimiento de la probable comisión de un delito, se inicia un proceso penal con el correspondiente desarrollo de una investigación criminal, la que por su importancia llega a ocupar una fase completa dentro de ese proceso, independientemente que sea de carácter inquisitivo o acusatorio.

En tal sentido, la investigación criminal puede concebirse en dos sentidos:

- Restringido: la investigación criminal es la actividad técnica y científica que realizan los órganos del Estado delegados para ello, con el fin de recolectar los medios de prueba que permitan conocer y comprender un hecho delictivo.

- Amplio, es una fase del proceso penal en la que se desarrolla la actividad de investigación criminal y se liga a este proceso a una persona determinada con base en los hallazgos primarios que la investigación va aportando.

Una deficiente investigación criminal no logrará recolectar los suficientes medios de prueba que hagan razonable para el Estado invertir tiempo, recursos y esfuerzos en llevar a una persona a un juicio formal. Por ello se dice que en la fase intermedia se discute básicamente, si la investigación criminal ha sido capaz de recolectar los suficientes elementos de prueba que permitan proseguir con el proceso penal para una persona determinada.

Superada la fase intermedia, en el sentido que la valoración judicial es favorable a la existencia de medios de prueba racionalmente útiles para ser conocidos en juicio, se llega a la tercera fase del proceso penal. Es decir el juicio. En ella es donde, básicamente, se analizará y discutirá, si el contenido de los medios de prueba recolectados con la investigación criminal es suficiente, para demostrar la existencia del delito y la responsabilidad de la persona acusada.

Independientemente de la consideración que se adopte respecto a la sentencia, ya sea como parte de la fase de juicio o como una cuarta fase del proceso penal, el proceso de deliberación que los jueces realizan para llegar a ésta, no es más que un análisis técnico-jurídico que permite determinar si los elementos de prueba recolectados en la investigación criminal y puestos a su disposición en la etapa del juicio, pueden asegurar que el hecho juzgado es constitutivo de delito, y si la persona acusada es la responsable del mismo.

En la fase de impugnaciones, si bien existe una limitación técnica en cuanto a discutir nuevamente la prueba producida en el juicio, con base en la investigación criminal realizada, se permite discutir los errores en que el tribunal haya incurrido respecto a la valoración de dichos elementos de prueba durante el juicio.

Salvo en la fase de ejecución, donde únicamente se limita al cumplimiento de lo dictaminado durante el juicio y la fase de impugnaciones, todo el proceso penal se ve directamente influenciado por la investigación criminal, en tanto como el instrumento que permite reconstruir un pasado y, con base en ella, determinar la aplicación de la respuesta penal prevista por el Estado en la ley.

Por su importancia, en la investigación criminal interviene el accionar policial, fiscal y judicial, independientemente del modelo procesal penal vigente.

La investigación criminal es un proceso que puede ser caracterizado por:

- Es continuo, por ser un proceso concatenado de actividades que está en interrelación con los diversos aspectos que afectan al problema por investigar.
- Es un proceso especializado, porque requiere de un trabajo metodológico de rigor técnico y científico.
- Requiere de previsión, ya que cada acto o fase de este proceso requiere de un planeamiento específico.

- Es analítico y sintético, ya que necesita de un análisis permanente de los elementos de la realidad obtenida y la síntesis de la información que ella aporta.
- Es explicativo-causal, porque permite determinar a quién, dónde, cuándo, cómo, por qué y para qué se perpetró el delito y con qué medios.
- Es metódico, puesto que es un proceso que se plantea hipótesis y métodos para la comprobación de los hechos.
- Es legal, por regirse con base a los preceptos y límites establecidos en la ley y se sujeta al control de un órgano judicial.

1.4.1. Las fuentes de información criminal

Información criminal es la que se da a partir de un presunto delito o de sus componentes (víctima, victimario, propiedades, vehículos, etc.), que sea relevante para la toma de decisiones. Ya sea en la prevención, detección y esclarecimiento del delito como en la persecución de delincuentes, la mejora de procesos judiciales y la creación de nuevas leyes.

Según esta definición, la mayor fuente de información criminal es el Sistema Penal, entendido como el conjunto de instituciones y procedimientos presentes en el proceso que transita un hecho delictivo desde que es registrado por el Estado.

La investigación criminal requiere de la intervención de los órganos policial, fiscal y judicial. Cada uno de ellos tiene un rol dentro de este proceso, que sí varía, de acuerdo con el modelo procesal que se tenga.

Dentro del esquema lógico del proceso de investigación criminal, podemos encontrar las siguientes fases

- Conocimiento del hecho o comprobación: puede darse a través de una denuncia, referencia, flagrancia, o de oficio por los órganos de persecución penal.
- Diligencias preliminares: comprenden el aislamiento y protección del lugar de los hechos, verificación del acto delictivo, inspección, descripción del lugar, elaboración de croquis, toma de fotografías, recolección de evidencias, protección de huellas, inscripción técnico-criminalística, envíos a laboratorios.
- Planeamiento de la investigación: incluye la determinación del tipo de modalidad del crimen, formulación de hipótesis sobre la víctima, del autor, de las circunstancias y medios del crimen, establecimiento de estrategias de resolución, y la asignación de responsables para cada paso.
- Recolección de información: consta de la escucha de testimonios y versiones, desarrollo de entrevistas, el manejo de informantes, seguimientos, instalación de fachadas y utilización de archivos nacionales (propiedades, vehículos, bancos, impuestos, armas.).
- Sustentación de la prueba: comprende el estudio científico de la evidencia y el desarrollo de las pruebas periciales.

- Presentación de la prueba al Órgano Judicial: comprende la presentación de hipótesis del delito, presentación de medios de información recolectados, presentación de medios de prueba sustentados en laboratorios, y presentación de hipótesis de solución del caso.

Este proceso lógico es sustantivamente útil para poder visualizar, conforme al ordenamiento jurídico vigente, el rol de las distintas agencias del sistema en cada una de estas fases, para evitar de esta manera, la desviación de funciones con la correspondiente sobrecarga de trabajo que genera impunidad. Para el efecto, el orden lógico anterior se describe en la tabla II.

Tabla II. **Fuentes de información criminal**

Fase	Órganos	Roles
Conocimiento del hecho o comprobación	Policía Nacional Civil	Recibir denuncias y cursarlas al Ministerio Público. Conocimiento de oficio y cursar la información al Ministerio Público.
	Jueces	Recibir denuncias y querellas y cursarlas al Ministerio Público. Conocimiento de oficio y cursar la información al Ministerio Público.
	Ministerio Público	Recibir denuncias y ordenar la investigación.
Diligencias preliminares	PNC: patrulleros, Servicio de Investigación Criminal (SIC) y miembros del Gabinete de Identificación	Fijar la escena del crimen, protegerla y realizar el trabajo de la escena del crimen bajo las instrucciones del Fiscal. Dirigida a una Política de Seguridad Ciudadana
	Ministerio Público	Dirigir a la PNC, desde el inicio hasta el final, en su trabajo en la escena del crimen y la recolección de información en la misma.

Continuación tabla II.

	Jueces de paz	Sólo excepcionalmente, mientras no exista despliegue suficiente del Ministerio Público en el interior del país, comparecer en la escena del crimen para autorizar el levantamiento de cadáveres.
Planeamiento de la investigación	Ministerio Público	Éste es el protagonista del fiscal.
Recolección de información		Desarrollo de diligencias solicitadas por el Fiscal en el modo y tiempo planificados. Encargar, monitorear y supervisar periódicamente el desarrollo de las diligencias encargadas a la PNC.
	Ministerio Público	Excepcionalmente, recibir en su sede información que se brinde por las partes o por las personas localizadas por la PNC que aportarán información relevante del caso.
Sustentación de la prueba	Esta fase debería ser realizada específicamente, por un ente especializado y autónomo; es decir, El Instituto Nacional Autónomo de Ciencias Forenses.	Desarrollo de peritajes y elaboración de informes para las partes.
Presentación de la prueba al Órgano Judicial	Ministerio Público	Elaborar la solicitud de acto conclusivo de la fase de investigación, que podrá ser, entre otras, la acusación y petición de auto de apertura a juicio.

Fuente: elaboración propia.

1.4.2. Situación actual de la política criminal en Guatemala en cuanto a la investigación criminal

Para poder realizar un análisis efectivo de la investigación criminal dentro de la política criminal del Estado, es necesario reconocer, lo que los teóricos han denominado dos grandes momentos en que se define la política criminal.

1.4.2.1. La formulación político-criminal de la investigación criminal

El primer momento del proceso de definición de la política criminal es la formulación. Este inicio es el que comprende la adopción general de las grandes decisiones, que orientarán el uso del poder coercitivo del Estado para dar respuestas al fenómeno criminal.

En este plano hay que referirse concretamente, a las decisiones políticas que adopta el Estado con carácter general, y que por su naturaleza se recogen en instrumentos jurídicos como la Constitución Política, el Código Penal, el Código Procesal Penal, la Ley de Redención de Penas, y algunos otros instrumentos de menor importancia, pero que igualmente recogen decisiones generales para determinado sector del sistema penal. Tal es el caso de las instrucciones Generales del Fiscal General de la República, las circulares de la Corte Suprema de Justicia, o las instrucciones Generales del Ministerio de Gobernación, en tanto se refieren a la organización del aparato estatal policial, para responder al fenómeno criminal.

Para el caso de la investigación criminal, al plano de la formulación le corresponde la adopción de decisiones generales que orienten el uso y desarrollo de dicha actividad, como herramienta a disposición del Estado para responder al fenómeno criminal.

Corresponde, entre otras decisiones, definir los fines de la investigación criminal, la extensión de la misma, forma de desarrollarla, la definición de los órganos que deben intervenir, los roles de los distintos órganos en dicho proceso, los plazos, los medios para llevarla a cabo, las facultades de los órganos que intervienen, los límites y la forma de controlar su desarrollo.

Estas decisiones se recogen en cuerpos normativos de carácter legal Constitución Política, Código Procesal Penal, Ley Orgánica del Ministerio Público, Ley de la Policía Nacional Civil, y en instrumentos políticos de dirección institucional Manual del Fiscal, Instrucciones Generales del Fiscal General, Convenio de Cooperación Interinstitucional entre el Ministerio Público y el Ministerio de Gobernación, para la eficacia de la investigación criminal, entre otros.

En el marco jurídico constitucional se encuentra claramente una orientación político-criminal que determina al Ministerio Público como el órgano encargado del ejercicio de la acción penal pública. La acción penal, procesalmente hablando, puede concebirse desde un sentido estricto, o bien, en un sentido amplio.

- Sentido restringido: se comprende la acción penal únicamente, como la facultad de solicitar al órgano judicial correspondiente la determinación de una situación jurídico-penal de una persona y su consecuencia jurídica.

Bajo este enfoque, el Ministerio Público sería únicamente un órgano meramente jurídico, que se encargaría de impulsar el funcionamiento del Órgano Judicial para decidir la situación jurídica de una persona que fue sujeto de investigación criminal por un órgano distinto y fuera de su control y supervisión.

- Sentido amplio: se comprende no sólo la facultad de solicitar al Órgano Judicial la determinación de una situación jurídico-penal de una persona y su consecuencia jurídica, sino también la de procurar todas las diligencias necesarias para adoptar una determinada postura jurídica ante el órgano jurisdiccional.

Bajo este enfoque, el Ministerio Público sería un ente facultado para procurar la obtención de los medios de prueba necesarios que le permitan decidir sobre la posición que adoptará ante el Órgano Judicial. El marco constitucional del país, no refleja en qué sentido le otorga el ejercicio de la acción penal pública al Ministerio Público, es decir, no es taxativo en cuanto si éste tiene bajo su competencia funciones de investigación criminal y, en su caso, cuáles.

No obstante, el Código Procesal Penal sigue desarrollando la formulación político criminal de la investigación criminal, al indicar claramente que el Ministerio Público tendrá a su cargo el procedimiento preparatorio y la dirección de la Policía Nacional Civil en su función investigativa dentro del proceso penal (Código Procesal Penal, artículo 107, segundo párrafo. Decreto 51- 92, del Congreso de la República.)

Por su parte, la Corte de Constitucionalidad ha ratificado que dicha regulación no contradice, sino desarrolla, el marco constitucional que otorga al Ministerio Público el ejercicio de la persecución penal.

La Corte indicó que la persecución penal no es sino una manifestación de la acción penal, ya que el procedimiento preparatorio es el conjunto de actos, particularmente de investigación, que llevan a determinar si hay razones suficientes para someter a una persona al juicio penal; por lo mismo, es lógico que se atribuya al Ministerio Público esas funciones y la dirección de la policía en su aspecto de aparato investigador. (Ley de la Policía Nacional Civil, artículo 9, Decreto 11-97, Congreso de la República.).

En tal sentido, se puede asegurar que, efectivamente, el Ministerio Público ha recibido la delegación del Estado para procurar una efectiva investigación criminal, que le permita determinar objetivamente si es procedente solicitar una condena, absolución, o cualquier otra posibilidad de solución a un acto delictivo sometido a su competencia. Esta situación; sin embargo, aún no aclara otro aspecto. Es importante indicar la forma en la que se formuló cómo debería desarrollarse la investigación criminal y qué le compete al Ministerio Público realizar dentro de ese marco político formulado.

Por su parte, la Ley de la Policía Nacional Civil continúa elaborando el marco político formulado por el Estado en cuanto al desarrollo de la investigación criminal, al indicar que la Policía Nacional Civil es la institución encargada de proteger la vida, la integridad física, la seguridad de las personas y sus bienes, el libre ejercicio de los derechos y libertades, así como prevenir, investigar y combatir el delito preservando el orden y la seguridad pública. (Ley de la Policía Nacional Civil, artículo 9, Decreto 11-97, Congreso de la República.)

Este marco político-criminal también se encuentra en el Código Procesal Penal, el cual establece que la policía, por iniciativa propia, en virtud de una denuncia o por orden del Ministerio Público, deberá: 1) investigar los hechos punibles de oficio; 2)...; 3) individualizar a los sindicados; 4) reunir los elementos de investigación útiles para dar base a la acusación o determinar el sobreseimiento".(Código Procesal Penal, artículo 112; en igual sentido, los artículos 113, 114 y 115).

En cuanto a los instrumentos internos que recogen las decisiones políticas formuladas para el desarrollo de la investigación criminal, se puede observar un vacío fundamental que incide en gran medida en la situación actual del desarrollo de dicha actividad estatal.

Por parte de la PNC se observa un vacío en cuanto a la regulación oficial del nuevo Servicio de Investigación Criminal (sic), el que de conformidad con la Ley Orgánica de la Policía y el Acuerdo de Fortalecimiento del Poder Civil y función del Ejército en una Sociedad Democrática (AFPC), constituye una verdadera sección de trabajo policial especializada. Su estructura actual mantiene la misma lógica de las estructuras de los órganos de investigación criminal de la Policía Nacional en años anteriores.

Por su parte, en el Ministerio Público se observa igualmente, un vacío significativo de instrumentos que recojan las decisiones políticas internas, las que organicen el funcionamiento de dicho órgano en las tareas que se le han encomendado en el proceso de investigación criminal.

Las instrucciones generales del Jefe del Ministerio Público no son mecanismos hasta ahora utilizados para definir, desde el plano político, la actuación de los funcionarios fiscales en el desarrollo de la investigación criminal.

Durante los últimos años, sólo se puede observar la Instrucción General 13-2001 como instrumento que orienta la actuación fiscal en sus competencias de investigación. Dicha instrucción dicta las orientaciones necesarias a los fiscales para proceder a la persecución penal de delitos cometidos en las entidades del Estado, cuyas denuncias carecen de la documentación por parte de los denunciantes.

El instrumento más importante sobre la orientación política de los fiscales en su competencia del proceso de investigación, sigue siendo hasta hoy, el Manual del Fiscal; sin embargo, dicho instrumento se considera más un insumo académico que político, dado que no ha existido ningún acompañamiento político de la institución que garantice su aplicabilidad.

También debe señalarse el último instrumento denominado por el Fiscal General de la República como Plan de Política Criminal Democrático del Ministerio Público, en el que, según el propio documento, se plasman los objetivos, principios y cualidades que orientarán todas y cada una de las acciones del quehacer de de sus miembros, expresando la unidad de criterios frente al tema del delito, la persecución penal y el mantenimiento de la legalidad del país.

El segundo momento del proceso de definición de la política criminal es el denominado momento de la configuración, en el que se trasladan a la realidad los métodos y las decisiones políticas adoptadas de forma general por el Estado para dar respuesta al fenómeno criminal. Es decir, este segundo momento comprende el traslado de las decisiones formuladas con carácter general a la cotidianidad, en consecuencia, nos referimos a la aplicabilidad que hacen del poder coercitivo del Estado los funcionarios delegados por éste para responder al fenómeno criminal.

En el plano de la configuración se encuentra que la aplicabilidad del poder coercitivo del Estado para responder al fenómeno criminal haciendo uso de las herramientas con las que se ha provisto, entre ellas la investigación criminal, se hace efectiva a partir de ciertos elementos que deben considerarse, tales como, la organización de los órganos que intervienen en ella, el despliegue de dichos órganos, los recursos con que se desarrollan tales funciones, la coordinación entre ellos, la priorización racional de los casos a investigarse, el cumplimiento o incumplimiento de los procedimientos establecidos, el cumplimiento o incumplimiento de plazos establecidos, entre muchos otros.

La importancia de reconocer estos dos grandes momentos en que se define la política criminal de Estado es fundamental ya que, como señala Binder, las características distintivas de la decisión original se irán definiendo a través de este proceso y no sólo en aquella formulación.

Concretizando en el tema de la investigación criminal, quienes deseen incursionar en el análisis de la misma para su correspondiente evaluación y generación de propuestas, deberán considerar las decisiones adoptadas, tanto en el plano de la formulación como en el plano de la configuración.

Las decisiones adoptadas en el plano de la formulación no constituyen teoría, como suele confundirse repetidamente por los funcionarios del sistema de justicia. Tales decisiones están respaldadas, en mayor o menor medida, por fundamentos teóricos, pero en sí mismas las decisiones formuladas en la ley son situaciones prácticas y no teóricas. En tal sentido, la regulación legal de una actividad, como la investigación criminal, no se considera ni debe considerarse como teoría, sino como decisiones prácticas adoptadas por los funcionarios competentes.

A continuación se analizan algunas situaciones incoherentes que existen en el plano de la política criminal de Guatemala, con relación al caso concreto de la investigación criminal.

1.4.2.2. Fines de la investigación criminal (su ubicación dentro de la política criminal del Estado)

Una pregunta fundamental que en el análisis político-criminal debe responderse, en cuanto a la investigación criminal, es la relativa a la ubicación que ésta tiene dentro de la política criminal del Estado.

La política criminal es el conjunto de principios, métodos y respuestas que el Estado organiza para responder al fenómeno criminal. La pregunta, gira en torno a si la investigación criminal se ubica dentro de la política criminal del Estado como un método para llegar a una respuesta del Estado ante determinado fenómeno criminal, o más bien es una respuesta organizada por el Estado para responder a este fenómeno.

En síntesis, hay que responder, si de acuerdo con el marco político-criminal, ¿es la investigación criminal, como parte del proceso penal, un paso necesario para que el Estado pueda responder al fenómeno criminal, o es en sí misma una respuesta al fenómeno criminal?

Desde el plano normativo se observa que el Estado ha formulado el proceso penal con objetivos claramente definidos: "la averiguación de un hecho señalado como delito o falta y de las circunstancias en que pudo ser cometido, el establecimiento de la posible participación del sindicado, el pronunciamiento de la sentencia respectiva, y la ejecución de la misma.

En ese sentido, puede observarse que el proceso penal es el vehículo por medio del cual el Estado averigua, determina la situación jurídica, decide la probable imposición de una pena y, en su caso, la ejecuta. Es decir, que el proceso penal es, efectivamente, el vehículo que permite al Estado ejercitar su poder coercitivo contra las personas, cumpliéndose lo que alguna vez citara Von Litz. No obstante, el proceso penal así concebido, comprende diversas etapas, y cada una de ellas contemplará finalidades específicas.

Cabe entonces preguntarse sobre las finalidades de la investigación criminal. Al respecto, en el Código Procesal Penal está la formulación de los fines que el Estado le ha asignado a la etapa preparatoria del proceso penal, también denominada etapa de investigación. Los fines que el Estado le ha reconocido a esta etapa del proceso penal son:

- Determinar la existencia del hecho con todas las circunstancias de importancia para la ley penal.

- Establecer quiénes son los partícipes.
- Procurar su identificación y el conocimiento de las circunstancias personales que sirvan para valorar su responsabilidad o influyan en su punibilidad, la verificación del daño o daños producidos por el delito.

De esta forma se puede indicar que en el plano de la formulación de la política criminal del Estado, la investigación criminal, ya como primera fase del proceso penal o como actividad previa del Estado para decidir sobre una respuesta al fenómeno criminal, es concebida únicamente como un método o un instrumento por medio del cual, el mismo pretende garantizar que la respuesta penal prevista en el marco jurídico respectivo será aplicada de manera justa.

En consecuencia, no puede concebirse la investigación criminal ninguna otra finalidad que no sea la de procurar un diligenciamiento correcto de los medios de prueba, que permitan al Estado decidir sobre la correspondencia y la necesidad de responder según las formas previstas en la ley ante el quebrantamiento de valores fundamentales de convivencia social.

Por un lado, entonces, la investigación criminal es el instrumento que permitirá al Estado arribar a la aplicación de una respuesta formulada con anterioridad; pero por el otro, la investigación criminal debe ser vista como uno de los límites que el Estado debe superar para poder hacer uso de su poder coercitivo para enfrentar el fenómeno criminal.

1.4.3. Información criminal del sistema judicial y penitenciario

La investigación criminal, como ha venido considerándose, es más que una actividad destinada a la recolección de información que ha de ser ingresada mediante medios de prueba a un proceso penal. Es una fase fundamental del proceso penal, que contiene una serie de principios, medidas, finalidades, actividades de investigación, entre otros aspectos.

En dicho procedimiento intervienen distintos órganos del Estado que, para poder intervenir, deben tener una clara delegación de competencia en el marco legal. Esto corresponde, nuevamente, al plano de la formulación de la política criminal del Estado. Es decisión política del Estado definir quiénes han de tener competencia para desarrollar funciones en el proceso de la investigación criminal, la extensión de su competencia y la forma en que deben ejercerla.

1.4.3.1. Roles de los órganos que intervienen en la investigación criminal

En Guatemala se ha formulado la intervención del Estado mediante tres órganos específicos: el Organismo Judicial, quien actúa a través de los jueces correspondientes, el Ministerio Público y la Policía Nacional Civil.

1.4.3.2. Funciones del Organismo Judicial

La investigación criminal, *in strictu*, es una actividad ajena a la función judicial, al menos en el modelo de proceso penal acusatorio acorde a un Estado Democrático de Derecho. Sin embargo, dentro de este modelo de proceso penal, se considera la función de pesos y contrapesos como uno de los mecanismos más importantes para garantizar el apego irrestricto a la ley.

Dentro de un modelo inquisitivo, es el juez quien desarrolla la función de investigar, acusar y juzgar, situación que es actualmente incompatible con los postulados modernos de un sistema de justicia democrático.

En el modelo actual del proceso penal guatemalteco, no se ha encarado la actuación judicial en la etapa de investigación, sin embargo ésta se restringe a la custodia de las garantías procesales estipuladas para el desarrollo justo del proceso penal.

Las funciones de control o custodia, que han sido formuladas en el marco político-criminal del Estado, para ser desarrolladas por los jueces en la etapa de investigación se resumen en tres:

- Control de actos que implican una limitación o restricción de derechos fundamentales

Efectivamente, en el plano legal ha sido considerada la necesidad de contar con una autorización judicial para el desarrollo de actos necesarios para el correcto avance de la investigación criminal, cuando estos afecten derechos fundamentales de las personas.

Entre éstos se consideran, principalmente, las medidas de coerción que se pueden imponer al sujeto pasivo de la investigación criminal. Así, la detención de las personas, la imposición de prisión preventiva, la imposición de arraigos, y otras medidas de coerción, requieren de la autorización judicial.

También se consideran dentro de estos actos, aquéllos, que siendo de recopilación de información, limitan algún derecho fundamental, tal el caso de los allanamientos y de los secuestros de objetos.

- Control de la intervención de los sujetos procesales durante dicha etapa

Le concede al funcionario judicial la facultad de habilitar la intervención de distintas personas en el procedimiento de averiguación. Entre ellas, se mencionan al actor civil, al tercero civilmente demandado y al querellante adhesivo.

- Control y diligenciamiento de los actos definitivos irreproducibles

Función importante otorgada en el marco jurídico, al funcionario judicial. Consiste en la autorización y diligenciamiento de los denominados anticipos de prueba, cuya autorización se fundamenta en la naturaleza irreproducible del medio de prueba que puede aportar información relevante para el esclarecimiento del acto que se investiga.

1.4.3.3. Funciones del Ministerio Público y de la Policía Nacional Civil

La formulación política, respecto a los órganos que intervienen en la investigación criminal, inicia con la norma constitucional que otorga al Ministerio Público el ejercicio de la acción penal pública. Esto, complementado con lo interpretado por la Corte de Constitucionalidad, en el sentido de que la persecución penal (que incluye la investigación criminal), no es sino la manifestación de la acción penal. (Expediente 296-94, de fecha 26 de enero de 1995, en Gaceta Jurisprudencial 35, Corte de Constitucionalidad, Guatemala, 1995, pp. 14-15.)

Cuando el Estado decidió romper con el esquema del proceso penal inquisitivo, otorgó al Ministerio Público la función de impulsar la investigación criminal, para reunir los elementos de prueba que sirvan al órgano judicial correspondiente para decidir el fondo del asunto. De esta manera se abstrae del juez la función de impulsar y coordinar o dirigir la investigación criminal, la cual se traslada al Ministerio Público.

Desarrollando el marco anterior, la Ley Orgánica del Ministerio Público (LOMP), también estipula que compete a este órgano promover la persecución penal y dirigir la investigación de los delitos de acción pública. (Ley Orgánica del Ministerio Público (LOMP), artículo 1.) También contempla que son funciones del Ministerio Público:

- Investigar los delitos de acción pública y promover la persecución penal ante los tribunales, según las facultades que le confieren la Constitución y otros tratados.

- Dirigir a la policía y demás cuerpos de seguridad del Estado en la investigación de hechos delictivos.

El marco político-criminal se completa en su formulación cuando el Estado otorga a la PNC la función de desarrollar la investigación criminal, bajo la dirección y coordinación del Ministerio Público.

En el Código Procesal Penal, artículo 112, estipula que la policía, por iniciativa propia, en virtud de una denuncia o por orden del Ministerio Público, deberá:

- Investigar los hechos punibles de oficio.
- Individualizar a los sindicados.
- Reunir los elementos de investigación útiles para dar base a la acusación o determinar el sobreseimiento.

Debido a los inconvenientes que representó el nuevo paradigma formulado, hubo necesidad de introducir reformas en 1997, al Código Procesal Penal que entró en vigencia en 1994, con lo cual se ratificó el esquema lógico de separación de funciones entre el Ministerio Público y la Policía Nacional, pero garantizándose en todo momento, que el primero es el ente que debe ejercer la dirección e impulso de la investigación criminal.

1.5. El Sistema Informático de Control de Casos del Ministerio Público (SICOMP)

A partir del 2003 el Ministerio Público, por medio de la Instrucción General número 01-2003 instituyó el Sistema Informático de Control de Casos SICOMP, como sistema de información para el registro, investigación y seguimiento de todos los casos que se presentan en las diferentes fiscalías que operan en toda la República. Esta es una herramienta que ordena y registra la información correspondiente y brinda valiosa información en todos los niveles de esa entidad.

Este es un sistema transaccional y no un sistema de análisis de datos, el cual contempla información diversa:

- Registro de hechos: características generales del hecho denunciado (lugar, día, hora, delito denunciado, agencia de recepción, entre otros.
- Registro de persona: características detalladas e identidad, ya sea del agraviado o del sindicado.
- Registro de elementos robados: información útil para la identificación de los objetos robados.
- Registro de autos robados: marca, modelo, color, chasis, número de motor, características particulares, entre otros.
- Registro de armas robadas: sus características, vinculando esta base de datos con otros sistemas tales como el IBIS.

- Registro de evidencias: descripción de huellas y pistas relevadas en la escena del crimen.
- Otra información importante relevante del hecho.

Las ventajas y mejoras que se conseguían con este sistema eran:

- Atención de los casos que cada fiscal tiene a su cargo.
- La toma de decisiones cualitativas de orientación jurídica de la investigación que los agentes fiscales efectúan.
- La dirección y control que los Fiscales de Distrito y de Sección ejercen en sus respectivas unidades de trabajo.
- La toma de decisiones de desarrollo institucional y generación de políticas de persecución penal de aplicación general.
- Almacenamiento de información importante para la generación de reportes estadísticos acerca de la información criminal, a nivel república.

Este proyecto de gran alcance contemplaba una implementación progresiva, comenzando por todas las fiscalías que operan en la ciudad de Guatemala y avanzando hacia el interior. En la actualidad este sistema ya cuenta con una nueva versión llamada SICOMP 2, la cual fue desarrollada con nuevas tecnologías y mejoras.

1.5.1. El problema del tratamiento de la información

Actualmente, el Ministerio Público y el Organismo Judicial analizan la información proveniente de sus sistemas (SICOMP, SICOMP2 y el Informático del Organismo Judicial), mediante análisis estadístico, sin hacer un aprovechamiento exhaustivo de la información. Los resultados de estos análisis son utilizados con distintos objetivos:

- Emitir informes mensuales a nivel nacional que reflejen la situación de los hechos delictivos.
- Servir de fundamento para la creación de planes de prevención y diseño de políticas criminales.
- Nutrir de información a los distintos fiscales que operan en las distintas agencias para dar el debido seguimiento a cada uno de los hechos.

El gran alcance de estos objetivos, en conjunto con la gran cantidad de información y de variables intervinientes, justifica el uso de herramientas más potentes para el análisis de la información que la estadística descriptiva convencional.

Por lo que, es necesario el uso de herramientas más potentes para el análisis exhaustivo de la información criminal recolectada, determinando relaciones subyacentes multivariantes para extraer conclusiones de mayor valor agregado, para la toma de decisiones.

1.5.2. El problema específico

El objetivo de esta investigación es dar especial interés en el análisis de determinados tipos de delitos que tienen prioridad en función de su gravedad y frecuencia. Éstos son los homicidios dolosos. Delitos que se caracterizan porque el objetivo del criminal es buscar la muerte de su víctima, éste es relevado por el Ministerio Público, por lo que se cuenta con información puntual sobre cada hecho.

Los homicidios dolosos, es decir intencionales, son el resultado de una de los mayores problemas de América Latina: la violencia. A fines del siglo XX ésta era la primera causa de muerte en América Latina, en personas de 15 a 44 años. En la actualidad, en Guatemala constituye uno de los mayores problemas relacionados con la violencia.

Dentro de este tipo de homicidios presentan principal relevancia aquéllos cometidos con armas de fuego. La Organización Mundial de la Salud muestra que, a fines del siglo XX, el 63% de los homicidios mundiales eran cometidos por armas de fuego. En América Latina esta cifra sobrepasaba el 80%.

Se estima que América Latina tiene la tasa específica de homicidios por armas de fuego más alta del mundo; la misma es aproximadamente tres veces superior a la de África, cinco veces mayor a la de Norteamérica, Europa Central y Europa del Este, y cuarenta y ocho veces más alta que la de Europa del Oeste y se centra en gran número en América Central donde se dan las mayores tasas de estos delitos.

El problema de las armas de fuego radica principalmente en su alto grado de efectividad y letalidad. En algunos trabajos estadísticos en Guatemala se observó que las áreas con mayor número de armas presentan mayores tasa de homicidios por armas de fuego, así como la presencia doméstica de armas de fuego, para la defensa personal aumentan las probabilidades de ser una víctima de homicidio.

Guatemala no es la excepción, en cuanto a la problemática de la violencia y su relación con las armas de fuego. El Ministerio Público registró 255 208 hechos delictivos en Guatemala en el 2004, de ellos 4 507 fueron homicidios, y de esa cantidad el 80 por ciento de las muertes fueron ocasionadas con armas de fuego. Para combatir este problema, en el 2005 se lanzó una Comisión Nacional de Desarme.

Desde este punto de vista el problema particular es:

Utilizar herramientas tecnológicas para encontrar patrones de homicidios dolosos, vinculados con el tipo de arma empleada, con el objetivo de generar nuevo conocimiento sobre la problemática y/o validar los conocimientos adquiridos hasta el momento, y así; utilizarlos como una herramienta de apoyo para el combate al crimen.

2. MINERÍA DE DATOS APLICADA A LA INVESTIGACIÓN CRIMINAL

2.1. Minería de datos

Se denomina Minería de datos al conjunto de técnicas y herramientas aplicadas al proceso no trivial de extraer y presentar conocimiento implícito, previamente desconocido, potencialmente útil y humanamente comprensible, a partir de grandes conjuntos de datos, con objeto de predecir y describir de forma automatizada tendencias y comportamientos; como también modelos previamente desconocidos.

El término minería de datos inteligente se refiere, específicamente a la aplicación de métodos de aprendizaje automático, para descubrir y analizar patrones que pueden encontrarse en los datos, para éstos, se desarrollaron un gran número de métodos de análisis de datos basados en la estadística.

En la medida en que se incrementaba la cantidad de información almacenada en las bases de datos, estos métodos empezaron a enfrentar problemas de eficiencia y escalabilidad, y es aquí donde aparece el concepto de minería de datos.

Entre las diferencias del análisis de datos por estadística descriptiva y la minería de datos, es que el primero supone que las hipótesis ya están construidas y validadas contra los datos, mientras que el segundo, que los patrones son automáticamente extraídos de los datos analizados.

La minería de datos es un proceso completo de descubrimiento de conocimiento que involucra varios aspectos:

- Entendimiento del dominio de aplicación, el conocimiento relevante a utilizar y las metas del usuario.
- Seleccionar un conjunto de datos dónde realizar el proceso de descubrimiento.
- Limpieza y pre procesamiento de los datos, diseñando una estrategia adecuada para manejar ruido, valores incompletos, valores fuera de rango, valores inconsistentes, entre otros.
- Selección de la tarea de descubrimiento a realizar, por ejemplo, clasificación, agrupamiento o *clustering*, reglas de asociación, entre otros.
- Selección de los algoritmos a utilizar.
- Transformación de los datos al formato requerido por el algoritmo específico de explotación de datos, hallando los atributos útiles, reduciendo las dimensiones de los datos, entre otros.
- Llevar a cabo el proceso de minería de datos para encontrar patrones interesantes.

- Evaluación de los patrones descubiertos y presentación de los mismos mediante técnicas de visualización. Podría ser necesario eliminar patrones redundantes o no interesantes, o se necesite repetir algún paso anterior con otros datos, algoritmos, otras metas o con otras estrategias.
- Utilización del conocimiento descubierto, ya sea incorporándolo dentro de un sistema, o simplemente para almacenarlo y reportarlo a las personas interesadas.

Es muy importante la etapa del pre-procesamiento de los datos y su transformación al formato requerido por el algoritmo, ya que dependiendo de cómo se realicen estas tareas, va a depender la calidad final de los patrones descubiertos. Un patrón es interesante, si es fácilmente entendible por las personas, potencialmente útil, novedoso, o valida alguna hipótesis que el usuario busca confirmar.

El proceso de minería en sí involucra ajustar modelos o determinar patrones a partir de datos. Este ajuste normalmente es de tipo estadístico, en el sentido que se permite un cierto error dentro del modelo.

Los algoritmos de minería de datos realizan, en general, tareas de predicción (de datos desconocidos) y de descripción (de patrones).

El término patrón se refiere a cualquier relación entre los elementos de la base de datos. Pueden incluir medidas de incertidumbre. Aquí se aplican una gran cantidad de algoritmos de aprendizaje y estadísticos. Un patrón es interesante en la medida que sea confiable, novedoso y útil respecto al conocimiento y los objetivos del usuario.

La evaluación normalmente se le deja a los algoritmos de extracción de patrones que, generalmente, están basados en significación estadística. Suele decirse que un patrón representa conocimiento si su medida de interesante rebasa un cierto umbral, lo cual está estructurado únicamente en medidas definidas por el usuario.

Las tareas principales en la minería de datos son:

- Análisis de dependencias

El valor de un elemento puede usarse para predecir el valor de otro. La dependencia puede ser: probabilística, definir una red de dependencias o ser funcional. Se ha orientado mucho en los últimos años en el descubrimiento de redes bayesianas o causales en donde la dependencia se da a nivel estructural (dependencias e independencias entre variables), y cuantitativa (fuerza de las dependencias).

- Identificación de clases (agrupamiento de registros en clases)

Identifica un conjunto finito de categorías o *clusters* que describen los datos (pueden ser exhaustivas y mutuamente excluyentes o jerárquicas y con superposiciones). Las clases pueden ser relevantes en sí o servir como entradas a otros sistemas de aprendizaje. Se utilizan algoritmos de *clustering*. Normalmente el usuario tiene una buena capacidad de formar las clases y se han desarrollado herramientas visuales interactivas para ayudar al usuario.

- Descripción de conceptos

Se resume un cierto patrón. La descripción puede ser característica (qué registros son comunes entre clases), o discriminatoria (cómo difieren las clases). La mayoría de los sistemas de aprendizaje encuentran descripciones de conceptos y están enfocados a clasificación: aprender una función que mapea (clasificar) un dato dentro de un conjunto de posibles clases predefinidas. Otra técnica relacionada es regresión: aprender una función que mapea un dato a una variable real. A veces se trata de encontrar descripciones compactas de subconjuntos de datos (media y varianza, leyes físicas) que los resuman de alguna forma.

- Detección de desviaciones, casos extremos o anomalías

Detectar los cambios más significativos en los datos con respecto a valores pasados o normales. Sirve para filtrar grandes volúmenes de datos que son menos probables de ser interesantes. El problema está en determinar cuándo una desviación es significativa para ser de interés.

Los componentes básicos de los métodos de minería son:

- Lenguaje de representación del modelo

Es muy importante conocer las suposiciones y restricciones en la representación empleada, es decir; la selección del conjunto de datos a analizar, tanto las variables objetivo que se quieren calcular o predecir, como las variables independientes que sirven para hacer el cálculo o proceso.

- Evaluación del modelo

En cuanto a predictibilidad, se basa en técnicas de validación cruzada (*cross validation*), la cual es una práctica estadística que parte de una muestra de datos en subconjuntos, de tal modo que el análisis inicial es realizado en uno de ellos; en cuanto a calidad descriptiva del modelo se basan en principios como el de máxima verosimilitud (*maximum likelihood*), método habitual para ajustar un modelo y encontrar sus parámetros; o en el principio de descripción mínima o MDL (*minimum description length*), el cual consiste en elegir la explicación más corta a los datos observados.

- Método de búsqueda

Se puede dividir en búsqueda de parámetros y del modelo, y determina los criterios que se siguen para encontrar los modelos (hipótesis).

Las principales técnicas de minería de datos se suelen clasificar, según su tarea de descubrimiento en:

- Agrupación o *clustering*
- Clasificación
- Asociación

2.1.1. Agrupación de datos o *clustering*

Ésta consiste en agrupar un conjunto de datos basándose en la similitud de los valores de sus atributos. El *clustering* identifica regiones densamente pobladas, denominadas *clusters*, de acuerdo a alguna medida de distancia establecida. De esta manera se busca maximizar la similitud de las instancias en cada cluster y minimizar la similitud entre clusters.

Dos de los algoritmos de *clustering* más utilizados son *Self Organizing Maps* (SOM) y *K-means*. SOM, también denominado redes de Kohonen, fue creado por Teuvo Kohonen en 1982. Se trata de un modelo de red neuronal con capacidad para formar mapas de características de manera similar a como ocurre en el cerebro.

SOM está basado en el aprendizaje no supervisado y competitivo, lo cual quiere decir que no se necesita intervención humana durante el mismo y que se necesita saber muy poco sobre las características de la información de entrada. SOM provee un mapa topológico de datos, que se representan en varias dimensiones, utilizando unidades de mapa (las neuronas) para simplificar la representación.

Las neuronas usualmente forman un mapa bidimensional, por lo que el mapeo transforma un problema de muchas dimensiones en el espacio, a un plano. La propiedad de preservar la topología significa que el mapeo preserva las distancias relativas entre puntos.

Los puntos que están cerca unos de los otros en el espacio original de entrada son mapeados a neuronas cercanas en *SOM*. Por esta razón, *SOM* es muy útil como herramienta de análisis de clases de datos de muchas dimensiones, y además tiene la capacidad de generalizar, lo que implica que la red puede reconocer o caracterizar entradas que nunca antes ha encontrado.

K-means es un método iterativo que busca formar *k clusters*, con *k* predeterminado antes del inicio del proceso. *K-means* comienza particionando los datos en *k* subconjuntos no vacíos, calcula el centroide de cada partición como el punto medio del *cluster* y asigna cada dato a éste, cuyo centroide sea el más próximo. Luego vuelve a particionar los datos iterativamente, hasta que no haya más datos que cambien de *cluster* de una iteración a la otra.

Otros algoritmos de *clustering* son: *K-medoids* o *PAM (Partition around medoids)* y *CLARA (Clustering Large Applications)*. Este último permite manejar conjuntos de datos más grandes que el primero. *CLARANS* integra los algoritmos *PAM* y *CLARA* en uno.

2.1.2. Clasificación de datos

La clasificación se utiliza para clasificar un conjunto de datos basado en los valores de sus atributos. Por ejemplo, se podría clasificar a distintas personas para la otorgación de un préstamo en riesgo bajo, medio y alto, teniendo en cuenta información histórica de las mismas.

La clasificación encuentra las propiedades comunes entre un conjunto de objetos y los clasifica en diferentes clases, de acuerdo a un modelo de clasificación. Para construir este modelo, se utiliza un conjunto de entrenamiento, en el que cada instancia consiste en un conjunto de atributos y el valor de la clase a la cual pertenece. El objetivo de la clasificación es analizar los datos de entrenamiento y, mediante un método supervisado, desarrollar una descripción o un modelo para cada clase utilizando las características disponibles en los datos. Esta descripción o modelo permite clasificar otras instancias, cuya clase es desconocida.

El método se conoce como supervisado debido a que, para el conjunto de entrenamiento, se conoce la clase de pertenencia y se le indica al modelo si la clasificación que realiza es correcta o no. La construcción del modelo se realimenta de estas indicaciones del supervisor.

Los algoritmos mayormente utilizados para las tareas de clasificación son los algoritmos de inducción. En la actualidad existen numerosos enfoques de algoritmos de inducción y variedad en cada enfoque, en el presente estudio se hace hincapié en aquéllos orientados a generar árboles de decisión.

La clasificación basada en árboles de decisión es un método de aprendizaje supervisado que construye árboles de decisión a partir de un conjunto de entrenamiento.

Un sistema típico de construcción de árboles de decisión es *ID3*, que utiliza la teoría de la información para minimizar la cantidad de pruebas para clasificar un objeto. Al utilizar métodos heurísticos, *ID3* garantiza un árbol simple, pero no necesariamente el más simple. Una extensión de *ID3* es *C4.5*, que extiende el dominio de clasificación de atributos categóricos a numéricos.

Un paso importante en la construcción del árbol de decisión es la poda, la cual elimina las ramas no necesarias, resultando en una clasificación más rápida y una mejora en la precisión de la clasificación de datos.

Existen muchos otros algoritmos de clasificación de datos, incluyendo métodos estadísticos, como algoritmos de *machine learning*; análisis de regresión lineal; redes neuronales, algoritmos genéticos y lógica difusa.

2.1.3. Reglas de asociación

La minería de reglas de asociación consiste en encontrar reglas de la forma:

$$(A_1yA_2y...yA_m) \Rightarrow (B_1yB_2y...yB_n),$$

Donde:

A_i y B_j son valores de atributos del conjunto de datos

Por ejemplo, se podría encontrar en un gran repositorio de datos de compras en un supermercado, la regla de asociación correspondiente a que si un cliente compra leche, entonces compra pan. Una regla de asociación es una sentencia probabilística acerca de la co-ocurrencia de ciertos eventos en una base de datos, y es particularmente aplicable a grandes conjuntos de datos.

2.2. Aplicaciones informáticas en el análisis de información criminal

Actualmente, la cantidad de información criminal recogida ha experimentado un crecimiento exponencial. Por ejemplo, a partir de los ataques terroristas del 11 de septiembre las agencias de inteligencia de EEUU, como la CIA o el FBI, procesan y analizan información activamente en búsqueda de actividad terrorista. Este hecho provoca que los analistas encuentren cada vez más dificultad para reunir la información adecuada, en un momento determinado, para la toma de decisiones (la llamada paradoja de la información; hay más información pero menos conocimiento).

A su vez, las modalidades criminales son cada vez más complejas y dinámicas. Bajo esta perspectiva se hace indispensable la utilización de herramientas informáticas para el tratamiento y análisis de la información criminal.

En tal sentido, el aporte de la informática en este campo abarca un amplio espectro que va desde la simple visualización de los hechos en un mapa hasta el uso de técnicas complejas de minería de datos. Los países que más han contribuido al desarrollo de estas aplicaciones son: Estados Unidos y Reino Unido. A continuación se describen las principales técnicas y proyectos realizados a nivel mundial.

2.2.1. Técnicas informáticas utilizadas en el análisis de información criminal

Existen diversas técnicas tanto gráficas, como descriptivas para el análisis de la información criminal. Cada una proporciona distinta información valiosa para su análisis.

2.2.1.1. Técnicas geográfico-visuales: el mapa del delito

El mapa del delito consiste en referenciar geográficamente los hechos delictivos, obteniendo una visualización geográfica que contempla no sólo la distancia entre hechos, sino también el territorio urbano (bancos, comercios, plazas) y las marcaciones territoriales (comisarías, barrios, zonas marginales).

Para la realización del mapa del delito se utilizan sistemas informáticos que permiten integrar, almacenar, analizar y visualizar información geográfica. Estos sistemas se denominan *GISs (Geographic Information Systems)* y los más utilizados a nivel mundial son *MapInfo* y *Arcview*.

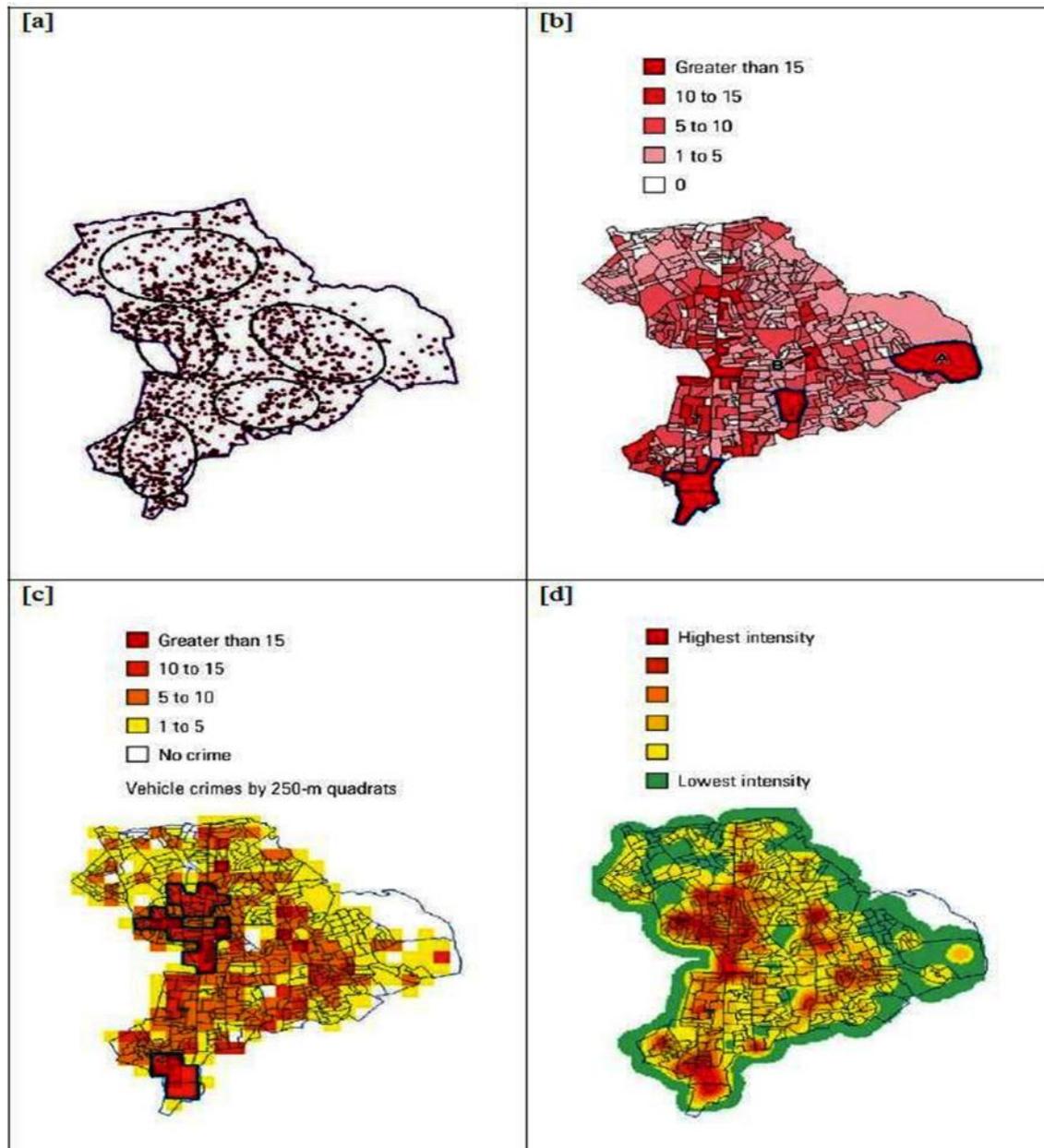
Existen muchas técnicas de análisis y visualización que trabajan sobre *GISs*. La mayoría de ellas buscan determinar y delimitar zonas de alta densidad delictiva, comúnmente llamadas zonas calientes o *hotspots*. Entre estas técnicas se pueden mencionar:

- Elipses de desvío estándar (*standard deviation ellipses*): utilizados para delimitar agrupaciones de hechos identificadas mediante técnicas de *clustering*.

- Mapas coloreados (*chloropeth maps*): en donde la escala de color representa la cantidad de hechos registrados en una determinada jurisdicción geográfica.
- Mapas de grillas (*quadrat maps*): similar al anterior, pero en este caso el mapa se fracciona según una grilla, y la escala de color representa la cantidad de hechos registrados en cada celda, asegurando igualdad de superficie.
- Mapas de contorno suavizado (*kernel density estimation*): similar al anterior pero con un efecto continuo logrado mediante el uso de algoritmos. Puede ser el más apropiado de todos los métodos para la visualización del delito.

Un caso de éxito de la utilización del mapa del delito, es el de Argentina donde esta técnica es muy utilizada para encontrar información importante sobre patrones de robos de vehículos:

Figura 2. **Diferentes tipos de mapas del delito delimitando hotspots para el robo de vehículos en Argentina**



Fuente: <http://laboratorios.fi.uba.ar/lsi/p-kogan-proyectodetesis.htm>. 08-09-11.

Existen varios paquetes de análisis estadístico espacial para información criminal que trabajan sobre *GISs*. Uno de los primeros fue *STAC (Spatial and Temporal Analysis of Crime)*, desarrollado por la Autoridad para la Información de Justicia Criminal de Illinois. Luego le siguieron *CompStat* y *CrimeStat*.

Este último contiene un amplio set de algoritmos, y si bien sus resultados pueden ser visualizados en *GIS*, está orientado hacia un usuario con determinados conocimientos técnicos. Por esta razón su uso queda limitado al analista criminal, más que al policía tradicional.

En Reino Unido la herramienta más utilizada como complemento de *GISs* es el *i2 Analyst's Workstation*. Si bien su capacidad es limitada, posee un módulo muy útil (llamado *PatternTracer*) que permite detectar patrones en los registros de llamadas telefónicas.

Es importante aclarar que, si este *software* encuentra su mayor utilización en el campo de la información criminal, la mayoría puede ser utilizado en cualquier otro campo en que se trate de información espacial, ya que en general, no incorpora ningún conocimiento específico del dominio criminal.

2.2.1.2. Técnicas de minería de datos

La minería de datos aplicada a la información criminal es un campo bastante nuevo y ha tenido un gran impulso en los últimos años en EEUU. Básicamente todas las técnicas de minería de datos descritas en la sección 2.1 pueden ser utilizadas en el análisis de información criminal.

Algunas de las aplicaciones más frecuentes son el uso de *clustering* particional para determinar *hotspots* y el uso de *SOM* para detectar grupos similares según el *modus operandi*. Esta última se basa en la idea de que cada grupo corresponda a una misma banda o delincuente.

2.2.2. Principales experiencias a nivel mundial

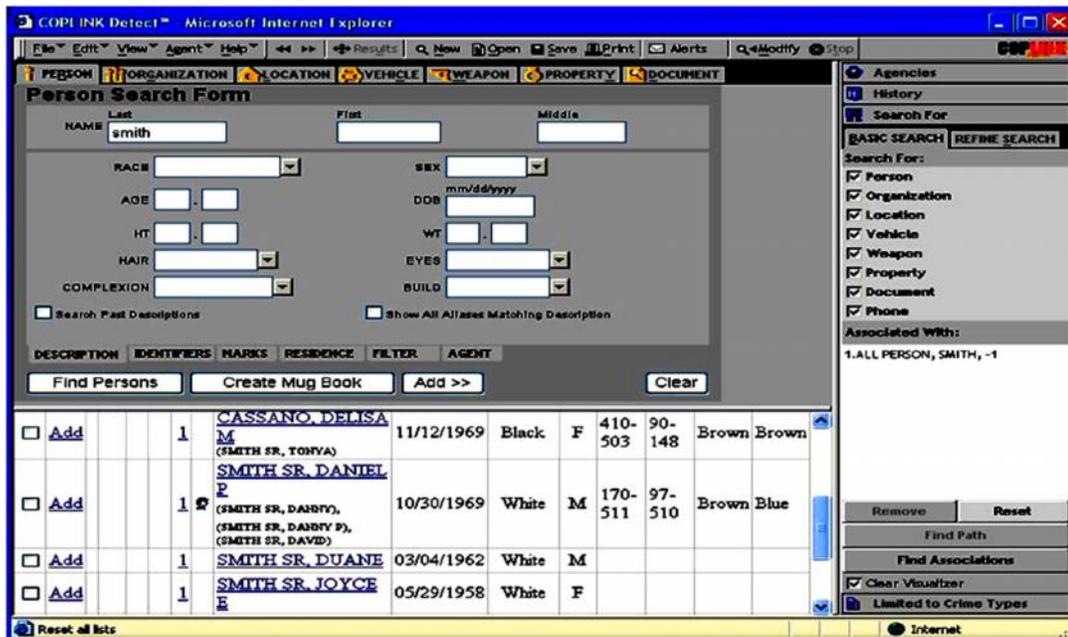
A continuación se describen las principales experiencias de aplicación de minería de datos en el análisis de información criminal. Es importante destacar que la mayoría de ellas incorporan a su vez, herramientas de visualización geográfica.

2.2.2.1. Proyecto *COPLINK*

El Proyecto *COPLINK* fue creado en 1997 en el Laboratorio de Inteligencia Artificial de la Universidad de Arizona, en Tucson, con el objetivo de servir de modelo para ser llevado a nivel nacional. Recientemente se ha desarrollado la versión comercial, denominada *COPLINK Solution Suite*.

Coplink está compuesto por dos sistemas integrados: *Coplink Connect* y *Coplink Detect*. El primero busca compartir información criminal entre distintos departamentos policiales, mediante un fácil acceso y una interface sencilla, integrando distintas fuentes de información. El segundo está diseñado para detectar de forma automática distintos tipos de asociaciones entre las bases de datos mediante técnicas de minería de datos.

Figura 3. Interface software Coplink: formulario de búsqueda de personas



Fuente: <http://bliss48.tripod.com/cpsc810/dg221.htm>. 08-09-2011.

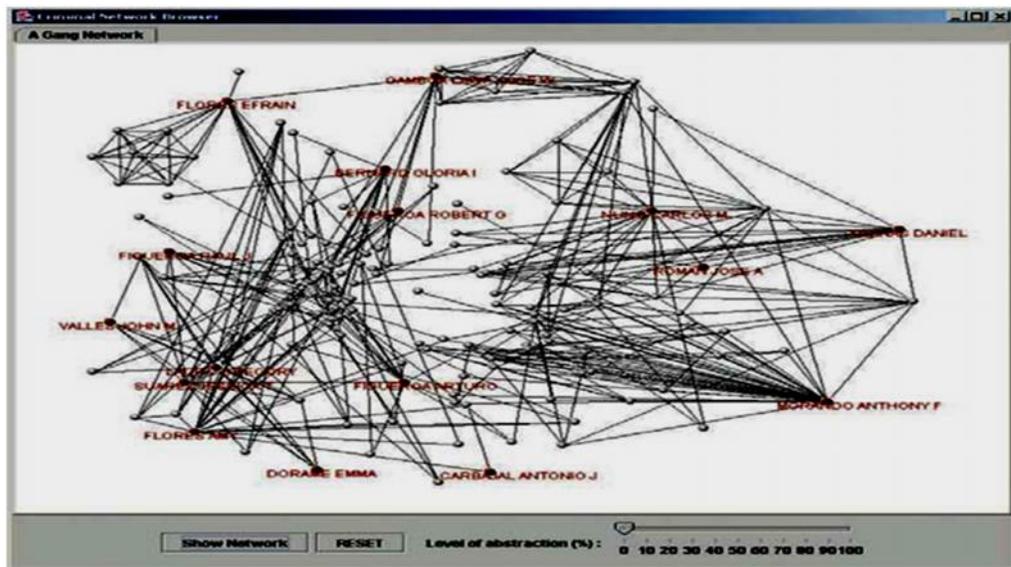
Ambos sistemas presentan una interface visual amigable y de fácil uso. Algunas de las aplicaciones de minería de datos desarrolladas por Coplink son las siguientes:

Análisis de redes criminales: consiste en identificar las redes o bandas criminales, sus líderes o integrantes clave y como se relacionan entre sí. En primer lugar se utiliza la técnica de *concept space* para extraer relaciones de los sumarios policiales y construir una posible red de sospechosos. La fuerza del vínculo entre dos sospechosos se mide con base en la frecuencia de hechos en los que participaron ambos. Luego se utiliza *clustering* jerárquico para partir la red en subgrupos y *block modeling*, para identificar patrones de interacción

entre los mismos. Finalmente se calcula el centro de cada subgrupo para determinar su miembro clave o líder.

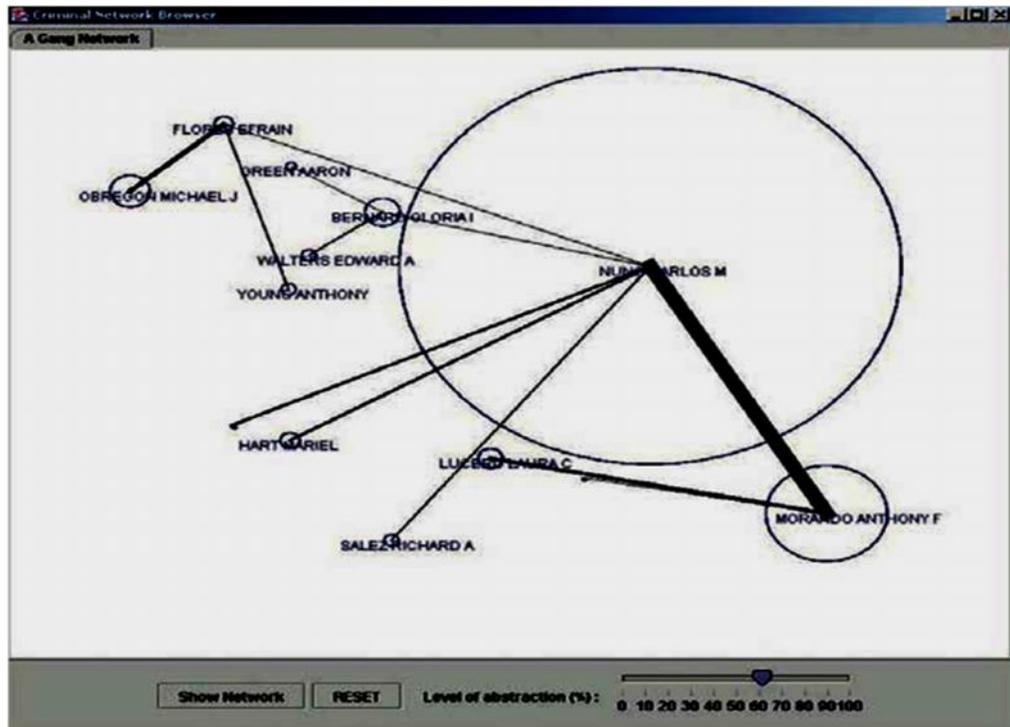
El resultado con base en registros del Departamento de Policía de Tucson sobre hechos cometidos entre 1985 y 2002 reveló 16 miembros clave (figuras 4. y 5). En la figura 5 se pueden ver representados los subgrupos encontrados y algunos de los miembros centrales o líderes. El tamaño del círculo es proporcional a la cantidad de miembros y el grosor de las líneas que vinculan los subgrupos representa la fuerza del vínculo. La validación con los expertos confirmó los resultados encontrados. Todos los expertos coincidieron en que esta herramienta aumentaría la productividad de los analistas criminales al mismo tiempo que favorecería a la prevención del crimen mediante una desarticulación efectiva de las bandas.

Figura 4. **Análisis de Redes Criminales: vínculos entre sospechosos**



Fuente: <http://ai.arizona.edu/research/coplink/crimenet.asp>. 08-09-2011

Figura 5. **Análisis de Redes Criminales: subgrupos identificados por Coplink**



Fuente: <http://ai.arizona.edu/research/coplink/crimenet.asp>. 08-09-2011

Extracción automática de entidades: consiste en extraer automáticamente determinada información criminal de los reportes policiales (nombres propios, direcciones, características personales, información de vehículos y nombres de drogas). Para esto se utiliza un algoritmo que funciona de la siguiente manera:

- Identifica las oraciones que poseen sustantivos, de acuerdo a un análisis sintáctico basado en determinadas reglas lingüísticas.

- Compara cada palabra de la oración con una base de entidades (por ejemplo apellidos, nombres de calles).
- Calcula un puntaje para cada oración en función a la cantidad de coincidencias encontradas.
- Utiliza una red neuronal para determinar el tipo de entidad.

El resultado sobre 36 sumarios de casos de narcóticos seleccionados al azar de la base de datos del Departamento de Policía de Phoenix demostró un buen poder para identificar nombres de personas (73,4% identificado) y nombres de drogas (77,9% identificado), aunque no tan bueno para direcciones (51,4% identificado) y características personales (47,8% identificado).

Detección automática de multiplicación de identidades: sobre una base de datos de sospechosos (nombre, sexo, documento, fecha de nacimiento, etc.) permite detectar casos en los que una misma persona se encuentra más de una vez con distinta identidad, ya sea intencionalmente (malversación de identidad) o por error en el ingreso de datos. Para este caso los especialistas seleccionaron únicamente 4 campos para determinar la identidad de una persona (por ser los menos ambiguos): nombre, fecha de nacimiento, dirección y número de seguridad social. El método consiste en tomar pares de registros, computar la similitud entre las cadenas de caracteres presentes en cada uno de los 4 campos mediante algoritmos especiales (*Phonetic Russell SoundEx Code* y *Agrep*) y luego calcular la distancia euclídea total entre registros.

El resultado sobre una muestra de 120 registros no únicos identificados manualmente sobre la base de datos del Departamento de Policía de Tucson, demostró un muy alto poder de precisión. Durante la fase de entrenamiento (con 80 registros) se logró un 97,4% de precisión, mientras que en la fase de prueba (con los 40 registros restantes) se obtuvo el 94,0% de precisión.

2.2.2.2. Proyecto OVER

El Proyecto *OVER* comenzó en el 2000 en Reino Unido como una iniciativa conjunta de la Policía de West Midlands y el Centro de Sistemas de Adaptación y División de Psicología de la Universidad de Sunderland. El proyecto está enfocado en los casos de robo a domicilio particulares. Sus principales objetivos son:

- Identificar los recursos críticos para establecer estrategias de prevención y detección más eficientes.
- Proveer de fundamentos empíricos para el desarrollo de planes interdepartamentales orientados a la reducción del delito.
- Identificar la información relevante a ser recolectada en el lugar del hecho, redundando en mejoras de eficiencia y reducción de tiempo del personal policial.
- Alimentar al sistema, tanto con información *hard* (información forense) como *soft* (información sobre la escena del delito).
- Analizar la distribución espacio-temporal de los hechos y confirmar las suposiciones sobre tendencias y patrones.

Las principales técnicas utilizadas para el análisis de la información son:

- Redes bayesianas.
- Redes neuronales de *Kohonen* (SOM), para la confección de perfiles de delincuentes, según el *modus operandi* y su asociación con delitos no resueltos.

Si bien el proyecto desarrolla principalmente capacidades predictivas, el *software* incorpora otras herramientas útiles, por ejemplo: la visualización referenciada geográficamente de los hechos.

2.2.2.3. Otras experiencias

A continuación se presentan otras experiencias menos difundidas de aplicaciones de este tipo:

- El Departamento de Policía de Ámsterdam utiliza el *software* de minería de datos *DataDetective* junto con *Mapinfo* para el análisis de registros criminales. Las principales técnicas empleadas son árboles de decisión y redes neuronales de *backpropagation*. Han unificado varias bases de datos policiales junto con información externa (clima, variables socioeconómicas y demográficas) en un único *data warehouse*. Los principales usos son:
 - Identificación de las causas del comportamiento criminal (por ejemplo casos de reincidencia).
 - Identificación de las causas del delito en un determinado barrio.

- Agrupamiento de delitos parecidos en *clusters* y su descripción, permitiendo un abordaje más efectivo.
- Identificación de delitos parecidos utilizando algoritmos *fuzzy search*, relacionando casos no resueltos con resueltos.
- Identificación de zonas de aumento del delito (por ejemplo se ha utilizado para la localización de equipos preventivos en operativos de búsqueda de armas).
- Evaluación del rendimiento policial.
- El Departamento de Policía de Richmond (Virginia) ha desarrollado una aplicación para el análisis de información criminal que combina minería de datos, mediante el *software Clementine* (SPSS, 2007), junto a un entorno visual aportado por *Information Builders* (IB, 2007) y una interface desarrollada por *RTI Internacional* (RTI, 2007). El principal objetivo es optimizar la ubicación de recursos, con base en una modalidad proactiva y no reactiva. Por ejemplo, durante 2007 se identificaron las zonas que habían tenido un aumento en los casos de heridos con arma de fuego el año anterior y para la noche se reforzaron exclusivamente esas zonas. El resultado obtenido fue una reducción del 49% en los casos de este tipo con un menor requerimiento de personal policial (aproximadamente 50 agentes menos).

- La Policía Estatal de Illinois adquirió en 2005 un *software* de minería de datos de la compañía RiverGlass Inc., con el objetivo de analizar la información criminal en tiempo real. El campo de aplicación es muy grande, y va desde la seguridad marítima en los puertos a la detección de casos de fraude financiero.
- El Departamento de Policía de San Francisco desarrolló junto a IBM la aplicación *CrimeMaps*, en base a la tecnología DB2 de IBM (IBM, 2007). Este *software* permite a los oficiales mediante un simple explorador web buscar un determinado tipo de crimen, realizar análisis de *clustering* y fijar niveles umbrales de alerta temprana, para un determinado delito en una determinada zona de acuerdo a una frecuencia histórica.
- El Departamento de Policía de Nueva York inició en julio de 2005 el *Real Time Crime Center*. Este ambicioso proyecto tiene como objetivo conformar un enorme *data warehouse* y cruzar información de todo tipo mediante herramientas de inteligencia de negocios (como *Repotnet 1.1* y *Accurint Pro*) de forma de detectar patrones de comportamiento y asociaciones antes desapercibidos.

3. SOLUCIÓN PROPUESTA

3.1. Solución propuesta al problema del tratamiento de la información

Se propone aplicar técnicas y herramientas de minería de datos sobre la información proporcionada por el MP mediante un *software* gratuito que permita a los analistas complementar el análisis actual con conclusiones de mayor valor agregado.

3.2. Herramienta a utilizar para el análisis de la información

La herramienta a utilizar para el análisis de datos será *Weka*, una herramienta perteneciente a la extensa colección de algoritmos de Máquinas de conocimiento desarrollados por la universidad de Waikato (Nueva Zelanda) implementados en Java; útiles para ser aplicados sobre datos mediante las interfaces que ofrece o para embeberlos dentro de cualquier aplicación.

Además, *Weka* contiene las herramientas necesarias para realizar transformaciones sobre los datos, tareas de clasificación, regresión, clustering, asociación y visualización. *Weka* está diseñado como una herramienta orientada a la extensibilidad por lo que añadir nuevas funcionalidades es una tarea sencilla.

Entre las ventajas y desventajas que esta herramienta posee, están:

Ventajas

- La licencia de *Weka* es GPL, significa que es *software* libre de distribución gratuita, por lo que el costo que representa el uso de esta herramienta es 0.
- *Weka* está programado en java, por lo que es independiente de la arquitectura, ya que funciona en cualquier plataforma sobre la que exista una máquina virtual de Java disponible.
- Cuenta con una interfaz gráfica amigable y fácil de usar.
- Tiene incorporado un amplio set de algoritmos de minería de datos.
- Está programado en código abierto, permitiendo al usuario programador agregar nuevas funciones según su necesidad.

Desventajas

- Es una herramienta que requiere conocimientos técnicos básicos, ya que presenta una interfaz bastante pobre, que la hace una herramienta difícil de comprender y manejar sin tener conocimientos previos de cómo se utiliza.
- Existe poca documentación orientada al usuario.

3.3. Solución propuesta al problema específico

Se propone llevar adelante un proceso de minería de datos sobre la base de datos de homicidios dolosos ocurridos en el país durante el 2008 mediante *Weka* 3.6.2 para identificar patrones de homicidios dolosos. El proceso propuesto es el siguiente:

- Construir un *data set* a partir de la base de datos de homicidios dolosos del 2008.
- Aplicar el algoritmo *K-means* para agrupar los hechos según su similitud en grupos o *clusters* distintos.
- Interpretar y convalidar los resultados obtenidos con los usuarios, haciendo uso de los informes emitidos por *Weka*.
- Utilizar el algoritmo de inducción C4.5 para identificar reglas de pertenencia a cada uno de los grupos o *clusters*.
- Proporcionar una interpretación definitiva y extraer conclusiones.

3.4. Algoritmos a utilizar

Se describen a continuación el conjunto de algoritmos a utilizar para realizar el análisis de datos.

3.4.1. Algoritmo *K-means*

K-means es un método particional de *clustering* donde se construye una partición de una base de datos D de n objetos en un conjunto de k grupos, buscando optimizar el criterio de particionamiento elegido. En *K-means* cada grupo está representado por su centro. El objetivo que se intenta alcanzar es minimizar la varianza total intra-grupo o la función de error cuadrático, ver figura 6.

Figura 6. **Error cuadrático**

$$V = \sum_{i=0}^k \sum_{j \in S_i} |x_j - \mu_i|^2$$

Fuente: elaboración propia.

Donde existen k grupos

S_i , $i=1,2,\dots, k$ y

μ_i es el punto medio o centroide de todos los puntos $X_j \in S_i$.

K-means comienza particionando los datos en k subconjuntos no vacíos, aleatoriamente o usando alguna heurística. Luego calcula el centroide de cada partición como el punto medio del cluster y asigna cada dato al *cluster* cuyo centroide sea el más próximo. Luego los centroides son recalculados para los grupos nuevos y el algoritmo se repite hasta la convergencia, la cual es obtenida cuando no haya más datos que cambien de grupo de una iteración a otra.

Para calcular el centroide más cercano a cada punto se debe utilizar una función de distancia. Para datos reales se suele utilizar la distancia euclídea. Para datos categóricos debe establecerse una función específica de distancia para ese conjunto de datos. Algunas de las opciones son utilizar una matriz de distancias predefinidas o una función heurística.

El algoritmo no garantiza que se obtenga un óptimo global. La calidad de la solución final depende principalmente del conjunto inicial de grupos. Debido a esto, suelen realizarse varias ejecuciones del algoritmo con distintos conjuntos iniciales, de modo de obtener una mejor solución.

Dado k , el algoritmo *K-means* se implementa en 4 pasos:

- Particionar los objetos en k subconjuntos no vacíos.
- Calcular los centroides de los *clusters* de la partición corriente. El centroide es el centro (punto medio) del *cluster*.
- Asignar cada objeto al cluster cuyo centroide sea más cercano.
- Volver al paso 2, parar cuando no haya más reasignaciones.

K-means es ampliamente utilizado en la explotación de datos, en la cuantificación de vectores, para cuantificar variable reales en k rangos no uniformes y para reducir el número de colores en una imagen.

3.4.2. Algoritmos de inducción

Son utilizados para inducir reglas a partir de datos históricos, clasificándolos en diferentes objetos, basándose en sus características y atributos.

3.4.2.1. Algoritmo *ID3*

El algoritmo *ID3*, diseñado en 1993 por J. Ross Quinlan, toma objetos de una clase conocida y los describe en términos de una colección fija de propiedades o de variables, produciendo un árbol de decisión sobre estas variables que clasifica correctamente todos los objetos. Hay ciertas cualidades que diferencian a este algoritmo de otros sistemas generales de inferencia.

La primera se basa en la forma en que el esfuerzo requerido para realizar una tarea de inducción crece con la dificultad de la tarea. El *ID3* fue diseñado específicamente para trabajar con cantidades grandes de objetos, y el tiempo requerido para procesar los datos crece sólo linealmente con la dificultad, como producto de:

- La cantidad de objetos presentados como ejemplos.
- La cantidad de variables dadas para describir estos objetos.
- La complejidad del concepto a ser desarrollado (medido por la cantidad de nodos en el árbol de decisión).

Esta linealidad se consigue a costa del poder descriptivo, ya que los conceptos desarrollados por el *ID3* sólo toman la forma de árboles de decisión basados en las variables dadas, y este lenguaje es mucho más restrictivo que la lógica de primer orden o la lógica multivaluada, en la cual otros sistemas expresan sus conceptos.

El *ID3* fue presentado como descendiente del *CLS* y, como contrapartida de su antecesor, es un mecanismo mucho más simple para el descubrimiento de una colección de objetos pertenecientes a dos o más clases. Cada objeto debe estar descrito en términos de un conjunto fijo de variables, cada una de las cuales cuenta con su conjunto de posibles valores. Por ejemplo, la variable humedad puede tener los valores: {alta, baja} y la variable clima: {soleado, nublado, lluvioso}.

Una regla de clasificación en la forma de un árbol de decisión puede construirse para cualquier conjunto C de variables de la siguiente forma:

- Si C está vacío, entonces se le asocia arbitrariamente a cualquiera de las clases.
- Si C contiene los representantes de varias clases, se selecciona una variable y se particiona C en conjuntos disjuntos C_1, C_2, \dots, C_n , donde C_1 contiene aquellos miembros de C_i que tienen el valor i para la variable seleccionada. Cada una de estos subconjuntos se maneja con la misma estrategia.

3.4.2.1.1. Descripción de *ID3*

El objetivo de *ID3* es crear una descripción eficiente de un conjunto de datos mediante la utilización de un árbol de decisión. Dados datos consistentes, es decir, sin contradicción entre ellos, el árbol resultante describirá el conjunto de entrada a la perfección. Además, el árbol puede ser utilizado para predecir los valores de nuevos datos, asumiendo siempre que el conjunto de datos sobre el cual se trabaja es representativo de la totalidad de los datos. Datos:

- Un conjunto de datos.
- Un conjunto de descriptores de cada dato.
- Un clasificador/conjunto de clasificadores para cada objeto.

Se desea obtener un árbol de decisión simple basándose en la entropía, donde los nodos pueden ser:

- Nodos intermedios: en donde se encuentran los descriptores escogidos según el criterio de entropía, que determina cuál rama es la que debe tomarse.
- Hojas: estos nodos determinan el valor del clasificador.

Este procedimiento de formación de reglas funcionará siempre, dado que no existen dos objetos pertenecientes a distintas clases, pero con idéntico valor para cada una de sus variables; si este caso llegara a presentarse, las variables son inadecuadas para el proceso de clasificación.

Hay dos conceptos importantes a tener en cuenta en el algoritmo *ID3*, la entropía y el árbol de decisión. La entropía se utiliza para encontrar el parámetro más significativo en la caracterización de un clasificador. El árbol de decisión es un medio eficiente e intuitivo para organizar los descriptores que pueden ser utilizados con funciones predictivas.

3.4.2.1.2. Pseudo-código del algoritmo *ID3*

A continuación, en la figura 7, se presenta el algoritmo del método *ID3* para la construcción de árboles de decisión en función de un conjunto de datos previamente clasificados.

Figura 7. Pseudo-código del Algoritmo de *ID3*

```
Función ID3
(R: conjunto de atributos no clasificadores,
C: atributo clasificador,
S: conjunto de entrenamiento) devuelve un árbol de decisión;
Comienzo
Si S está vacío,
Devolver un único nodo con Valor Falla;
Si todos los registros de S tienen el mismo valor para el atributo clasificador,
Devolver un único nodo con dicho valor;
Si R está vacío,
Devolver un único nodo con el valor más frecuente del atributo clasificador en los
registros de S [Nota: habrá errores, es decir, registros que no estarán bien
clasificados en este caso];
Si R no está vacío,
D ← atributo con mayor Ganancia (D,S) entre los atributos de R;
Sean {dj | j=1,2,..., m} los valores del atributo D;
Sean {Sj | j=1,2,..., m} los subconjuntos de S correspondientes a los valores de
dj respectivamente;
Devolver un árbol con la raíz nombrada como D y con los arcos nombrados d1,
d2,...,
dm que van respectivamente a los árboles
ID3(R-{D}, C, S1), ID3(R-{D}, C, S2), ..., ID3(R-{D}, C, Sm);
Fin
```

El *ID3* puede aplicarse a cualquier conjunto de datos, siempre y cuando las variables sean discretas. Este sistema no cuenta con la facilidad de trabajar con variables continuas, ya que analiza la entropía sobre cada uno de los valores de una variable, por lo tanto, tomaría cada valor de una variable continua individualmente en el cálculo de la entropía, lo cual no es útil en muchos de los dominios. Cuando se trabaja con variables continuas, generalmente se piensa en rangos de valores y no en valores particulares.

Existen varias maneras de solucionar este problema del *ID3*, como la agrupación de valores presentada, o la discretización de los mismos. El *C4.5* resolvió el problema de los atributos continuos mediante la discretización.

3.4.2.2. Algoritmo C4.5

El *C4.5* se basa en el *ID3*, por lo tanto, la estructura principal de ambos métodos es la misma. El *C4.5* construye un árbol de decisión y evalúa la información en cada caso utilizando los criterios de entropía y ganancia o proporción de ganancia, según sea el caso. A continuación, se describen las características particulares de este método que lo diferencian de su antecesor.

3.4.2.2.1. Pseudo-código del algoritmo C4.5

El algoritmo del método C4.5 para la construcción de árboles de decisión a grandes rasgos es muy similar al del ID3. Varía en la manera en que realiza las pruebas sobre las variables, ver figura 8.

Figura 8. Pseudo-código del Algoritmo de C4.5

```
Función C4.5
(R: conjunto de atributos no clasificadores,
C: atributo clasificador,
S: conjunto de entrenamiento) devuelve un árbol de decisión;
Comienzo
Si S está vacío,
Devolver un único nodo con Valor Falla;
Si todos los registros de S tienen el mismo valor para el atributo
clasificador,
Devolver un único nodo con dicho valor;
Si R está vacío,
Devolver un único nodo con el valor más frecuente del atributo
clasificador en los registros de S [Nota: habrá errores, es decir,
registros que no estarán bien clasificados en este caso];
Si R no está vacío,
D ← atributo con mayor Proporción de Ganancia(D,S) entre los
atributos de R;
Sean {dj | j=1,2,..., m} los valores del atributo D;
Sean {Sj | j=1,2,..., m} los subconjuntos de S correspondientes a los
valores de dj respectivamente;
Devolver un árbol con la raíz nombrada como D y con los arcos
nombrados d1,d2,...,dm, que van respectivamente a los árboles
C4.5(R-{D}, C, S1), C4.5(R-{D}, C, S2), C4.5(R-{D}, C, Sm);
Fin
```

Fuente: elaboración propia.

3.4.2.2.2. Características particulares de C4.5

En cada nodo, el sistema debe decidir cuál prueba escoge para dividir los datos. Los tres tipos de pruebas posibles propuestas por C4.5 son:

- La prueba estándar para las variables discretas, con un resultado y una rama para cada valor posible de la variable.
- Una prueba más compleja, basada en una variable discreta, en donde los valores posibles son asignados a un número variable de grupos con un resultado posible para cada grupo, en lugar de para cada valor.
- Si una variable A tiene valores numéricos continuos, se realiza una prueba binaria con resultados $A \leq Z$ y $A > Z$, para lo cual debe determinarse el valor límite Z .

Todas estas pruebas se evalúan de la misma manera, mirando el resultado de la proporción de ganancia, o alternativamente, el de la ganancia resultante de la división que producen. Ha sido útil agregar una restricción adicional: para cualquier división, al menos dos de los subconjuntos T_i deben contener un número razonable de casos. Esta restricción, que evita las subdivisiones casi triviales, es tomada en cuenta solamente cuando el conjunto T es pequeño.

4. ANÁLISIS DE DATOS CON WEKA

4.1. Conjunto de datos

Se denomina *data set* al conjunto de datos a analizar con el *software Weka* de minería de datos. El proceso de configuración del *data set* a partir de una base de datos involucra diversas etapas:

- Consolidación de la información de interés en una tabla única.
- Selección de los campos de interés.
- Depuración de registros en busca de integridad y consistencia.
- Modificación de las variables de los campos en función del *software* y los algoritmos a utilizar y/o de la visión del especialista.

En los siguientes puntos se desarrollan estos pasos con un conjunto de datos ficticios que simulan la información recabada por el Ministerio Público. La base de datos contiene 2855 hechos de homicidios dolosos ocurridos en el 2008.

4.2. Consolidación de la información en una tabla única

La base de datos está compuesta por una tabla principal, una tabla secundaria y tres tablas de referencia.

La tabla principal es:

- Incidente: contiene la información relevante del hecho en sí (fecha, lugar, hora, circunstancias).

La tabla secundaria es:

- Persona: contiene información referida a los involucrados por un determinado hecho (sexo, edad, nombre, características físicas, etc.), tanto del los agraviados como de los sindicados.

Las tablas de referencia describen la codificación de determinados campos de la tabla Incidente. Éstas son:

- Departamento: contiene la descripción de la división política de Guatemala.
- Municipio: contiene la descripción de los municipios que conforman cada uno de los departamentos contenidos en la tabla departamento.
- Tipo_delito: contiene la descripción de los delitos codificados en la tabla incidente.

Debido a que la tabla principal (incidente) contiene la mayor cantidad de información relevante, será la tabla base para conformar *data set*.

Las tablas de referencia son únicamente catálogos y serán consideradas en la medida que describan la codificación de algún campo de interés de la tabla principal.

4.3. Selección de los campos de interés

La tabla incidente presenta cierta cantidad de campos sobre los aspectos del hecho. Para objetivos del análisis se utilizarán los campos más representativos de cada aspecto y se utilizarán otros para dar mayor relevancia de la información.

A continuación se realiza una descripción de cada uno de estos campos. Para cada uno de los campos seleccionados se determinan:

- Delitos: valores que puede tomar cada delito
- Descripción de cada uno de los estados
- Frecuencia: cantidad de registros para cada estado
- Nivel de completitud: cantidad de registros vacíos o incompletos

Mientras que para los campos omitidos, se explica la razón de su omisión.

4.3.1. Campos seleccionados

Departamento: contiene el lugar en el cual se cometió el hecho. Su codificación (contenida en el catálogo departamento) y la distribución de los hechos se muestran respectivamente en la tabla III y la figura 9.

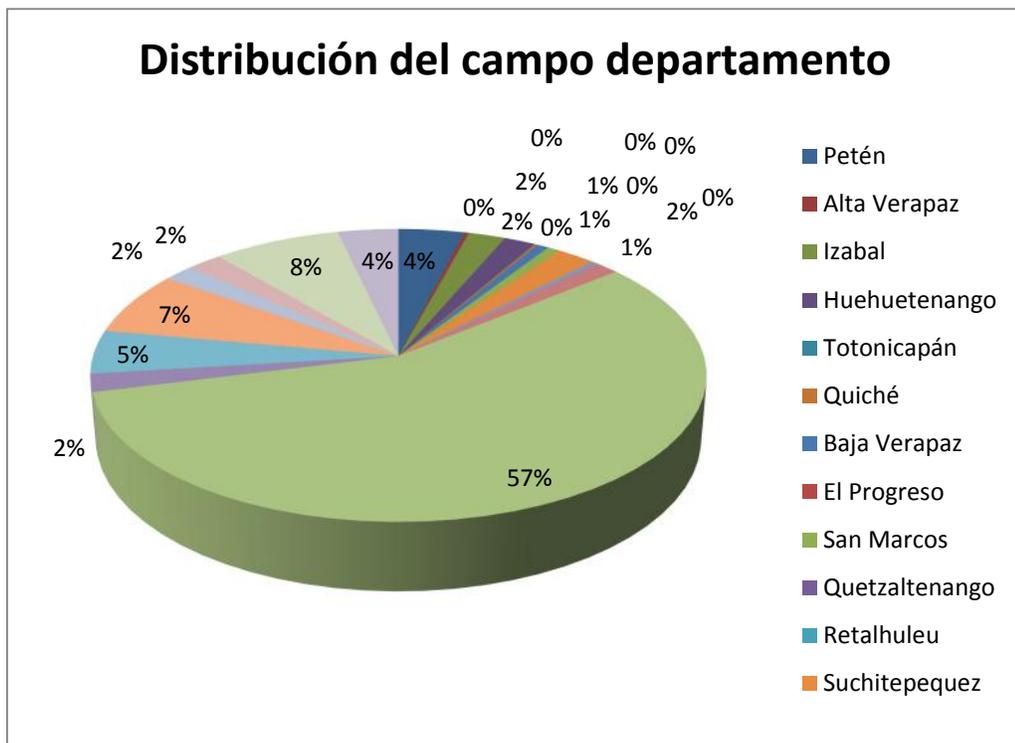
Tabla III. **Distribución del campo departamento**

Departamento	Cantidad
Petén	112
Alta Verapaz	8
Izabal	63
Huehuetenango	26
Totonicapán	0
Quiché	5
Baja Verapaz	20
El Progreso	28
San Marcos	22
Quetzaltenango	0
Retalhuleu	0
Suchitepéquez	66
Sololá	11
Chimaltenango	40
Guatemala	1 628
Sacatepéquez	57
Escuintla	137
Santa Rosa	200
Zacapa	52
Jalapa	59
Jutiapa	218
Chiquimula	103

Fuente: elaboración propia.

Se puede observar que la mayor cantidad de casos ocurre en pocos los departamentos de Guatemala, Santa Rosa, Escuintla, Jutiapa, Petén.

Figura 9. **Distribución del campo departamento**



Fuente: elaboración propia.

El campo fecha_hecho permite hacer una división importante para el análisis de datos correctos al aplicar minería de datos a los mismos, por lo tanto fue dividido en los siguientes campos:

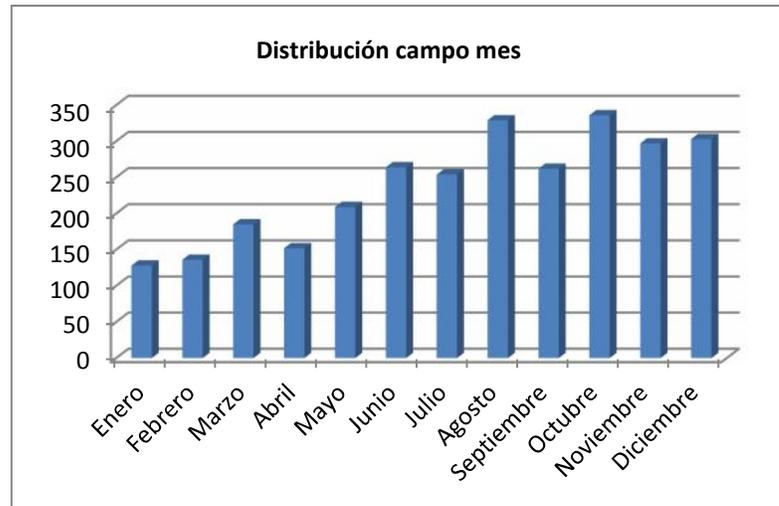
Mes: contiene el mes en que se cometió el hecho. La distribución es la siguiente tabla IV y figura 10:

Tabla IV. **Codificación y frecuencias del campo mes**

Estado	Mes	Cantidad
1	Enero	128
2	Febrero	136
3	Marzo	185
4	Abril	152
5	Mayo	209
6	Junio	264
7	Julio	254
8	Agosto	329
9	Septiembre	262
10	Octubre	336
11	Noviembre	297
12	Diciembre	303
	Total	2 855

Fuente: elaboración propia.

Figura 10. **Distribución del campo mes**



Fuente: elaboración propia.

Existe cierto comportamiento estacional de los hechos, registrándose una mayor cantidad de casos en los últimos meses del año. La relación entre el mes de mayor cantidad de casos (octubre, 336 casos) y el de menor (enero, 128 casos) es de más del doble.

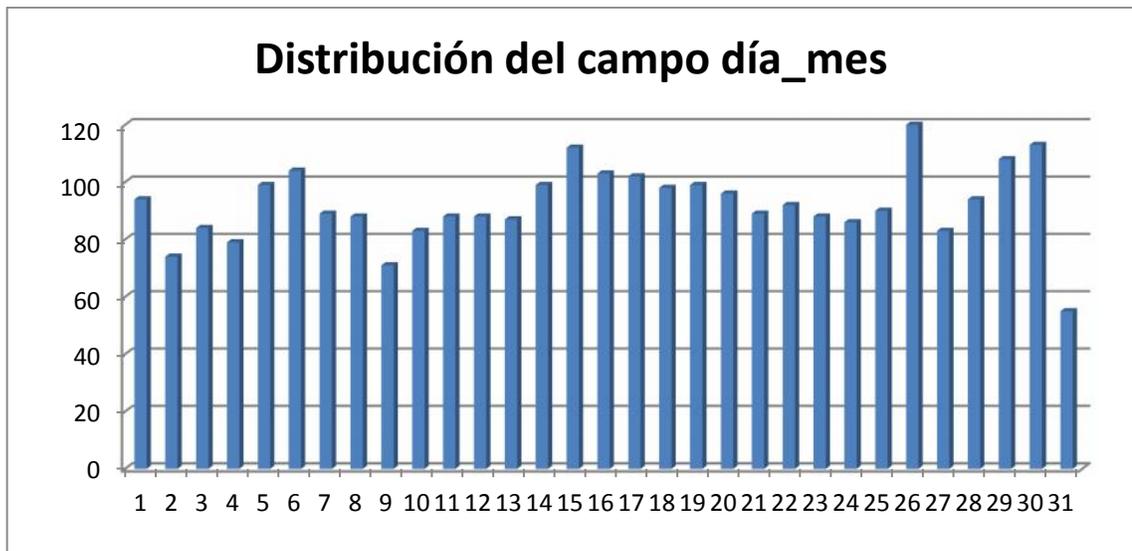
Día_mes: contiene el día del mes en el cual se cometió el hecho. La distribución es la siguiente, ver tabla V y figura 11.

Tabla V. **Codificación y frecuencias del campo día_mes**

Estado	Cantidad	Estado	Cantidad
1	94	17	102
2	74	18	98
3	84	19	99
4	79	20	96
5	99	21	89
6	104	22	92
7	89	23	88
8	88	24	86
9	71	25	90
10	83	26	120
11	88	27	83
12	88	28	94
13	87	29	108
14	99	30	113
15	112	31	55
16	103		
		Total	1 413

Fuente: elaboración propia.

Figura 11. Distribución del campo día_mes



Fuente: elaboración propia.

En este caso, la distribución de los hechos es más aleatoria (no se observa un comportamiento estacional) y el día 31 muestra la menor cantidad de casos registrada debido a que no todos los meses cuentan con 31 días. Se observan pequeños picos en los días quincenales del mes y los últimos días del mes, esto podría indicar que muchos de estos hechos podrían estar vinculados en ocasión de otros delitos, por ejemplo robo, ya que en estos días es cuando se dan los pagos de salarios.

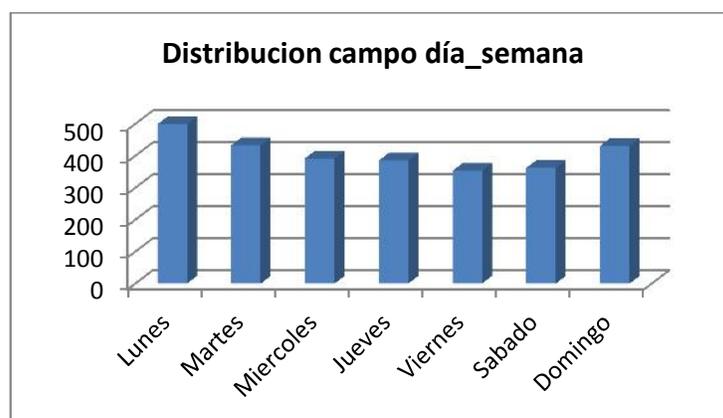
Día_semana: contiene el día de la semana en que se cometió el hecho.
La clasificación y su distribución son las siguientes, ver tabla VI y figura 12:

Tabla VI. **Codificación y frecuencias día_semana**

Estado	Día semana	Cantidad
1	Lunes	499
2	Martes	433
3	Miércoles	392
4	Jueves	385
5	Viernes	353
6	Sábado	362
7	Domingo	431
	Total	2 855

Fuente: elaboración propia.

Figura 12. **Distribución del campo del campo día_semana**



Fuente: elaboración propia.

En este comportamiento de los datos no se observa una variación muy definida entre los distintos días de la semana, se puede observar un leve aumento en la cantidad de casos entre los días domingo y martes y va disminuyendo la cantidad de casos entre los días miércoles y sábado. Por lo que, la mayor cantidad de delitos ocurren en los primeros días de la semana (lunes y martes).

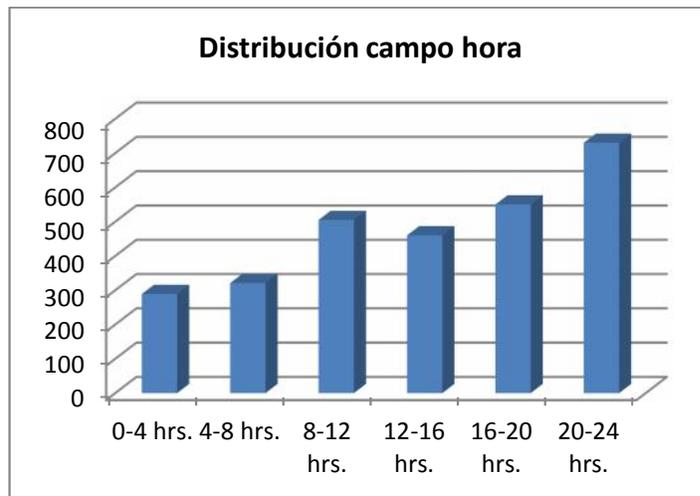
Hora: contiene el rango horario en la cual se cometió el delito, dividiendo el día en rangos de 4 horas lo cual servirá para el análisis posterior y brindará un mejor resultado. Los rangos y su distribución son los siguientes, ver tabla VII y figura 13:

Tabla VII. **Codificación y frecuencias del campo hora**

Estado	Hora	Cantidad
1	0-4 hrs.	289
2	4-8 hrs.	321
3	8-12 hrs.	505
4	12-16 hrs.	460
5	16-20 hrs.	550
6	20-24 hrs.	730
	Total	2 855

Fuente: elaboración propia.

Figura 13. **Distribución del campo hora**



Fuente: elaboración propia.

Se puede observar que existe una tendencia estacional a una mayor cantidad de casos en horas de la noche, entre las 16 y 24 horas es cuando más existen este tipo de hechos y es cuando más incidencia criminal existe en el país.

Este parámetro podría indicar que los delincuentes operan con mayor frecuencia en horas nocturnas aprovechando los lugares ocultos y solitarios, tomando ventaja del poco movimiento que existe en estas horas.

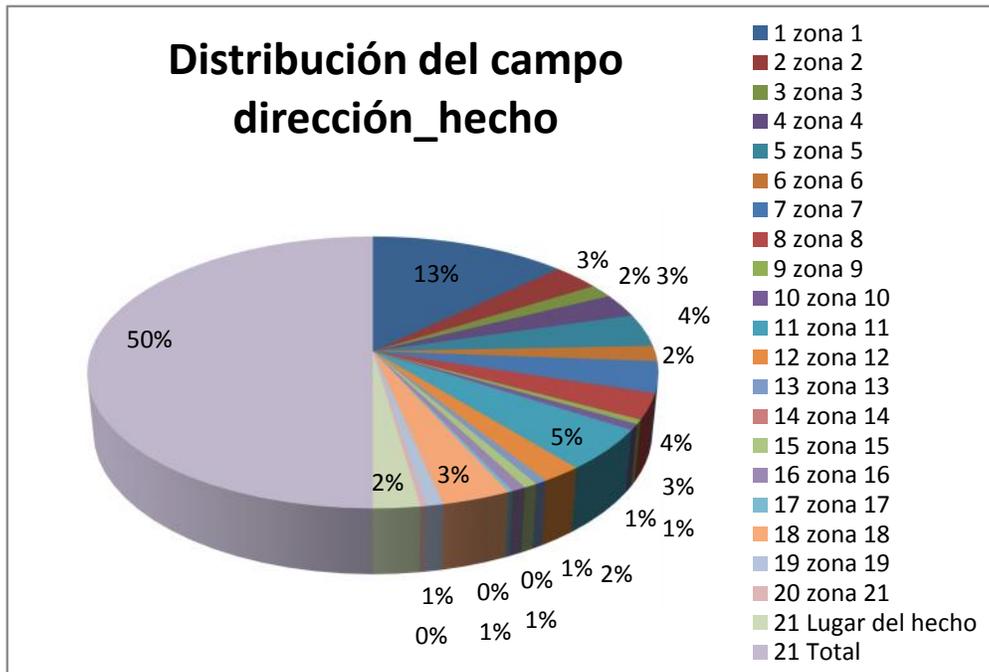
Dirección_hecho: contiene una clasificación del lugar donde se cometió el hecho. La clasificación y su distribución se muestran en la tabla VIII y la figura 14, respectivamente.

Tabla VIII. **Codificación y frecuencias del campo dirección_hecho**

Estado	Dirección_Hecho	Cantidad
1	zona 1	741
2	zona 2	178
3	zona 3	90
4	zona 4	159
5	zona 5	223
6	zona 6	104
7	zona 7	211
8	zona 8	175
9	zona 9	29
10	zona 10	41
11	zona 11	280
12	zona 12	114
13	zona 13	32
14	zona 14	1
15	zona 15	38
16	zona 16	40
17	zona 17	13
18	zona 18	192
19	zona 19	47
20	zona 21	17
21	Lugar del hecho	130
	Total	2 855

Fuente: elaboración propia.

Figura 14. Codificación y frecuencias del campo dirección_hecho



Fuente: elaboración propia.

Se muestran únicamente 20 zonas (de la zona 1 a la zona 19, y la zona 20), ya que se toma como referencia la ciudad de Guatemala que es donde ocurren más casos. La capital está compuesta por 25 zonas, pero la 20 sería el área de ciudad San Cristóbal, las zonas 22 y 23 se ubican en la salida al atlántico, y las 22 y 23 están ubicadas en la parte de San José Pínula.

Se registra también, el campo lugar del hecho, esto es porque existen lugares donde se cometió el delito registrados como cantones, asentamientos, barrancos, y carreteras en las cuales no está definida exactamente la zona del en que se cometió el hecho o en el momento no se registró exactamente.

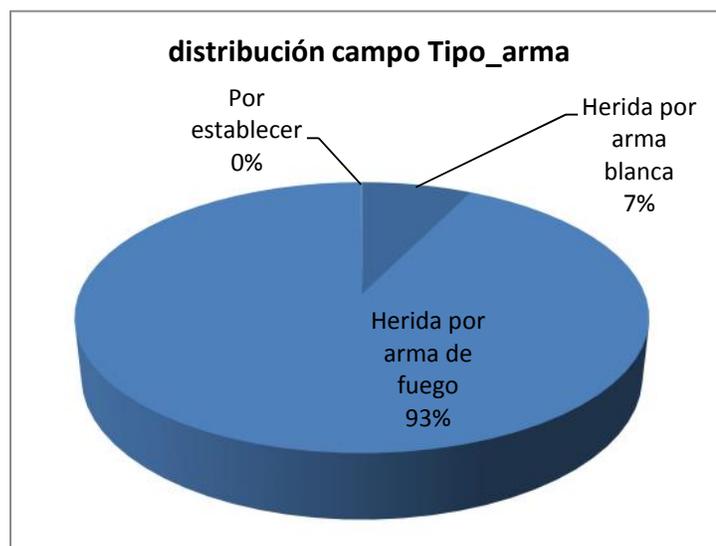
Tipo_arma: con la que se cometió el homicidio. La clasificación y su distribución se muestran en la tabla IX y la figura 15, respectivamente.

Tabla IX. **Codificación y frecuencias del campo Arma**

Estado	Descripción	Cantidad
1	Herida por arma blanca	209
2	Herida por arma de fuego	2645
3	Por establecer	1
	Total	2 855

Fuente: elaboración propia.

Figura 15. **Codificación y distribución del campo arma**



Fuente: elaboración propia.

La mayor cantidad de casos ocurre con arma de fuego (2645 registros), este es un indicador que en el país las armas de fuego tienen un papel protagónico en la violencia que afecta la sociedad.

Podrían ser varios factores los que expliquen este tipo de comportamiento: las políticas de permisividad en el uso y portación de armas, los mercados negros que ofrecen armas a la alta demanda que existe en el país por parte de la población, la inseguridad ciudadana; factor importante que existe actualmente y que ha promovido la cultura de violencia, obligando a la población a creer que mientras esté armada, estará más segura. Todos estos factores inciden en la elevada tasa de delitos cometidos con este tipo de armas (96%). Hay 1 registro en el que no se informa el tipo de arma, lo que indica que no se determinó o no se registró el tipo de arma con el cual se realizó el hecho.

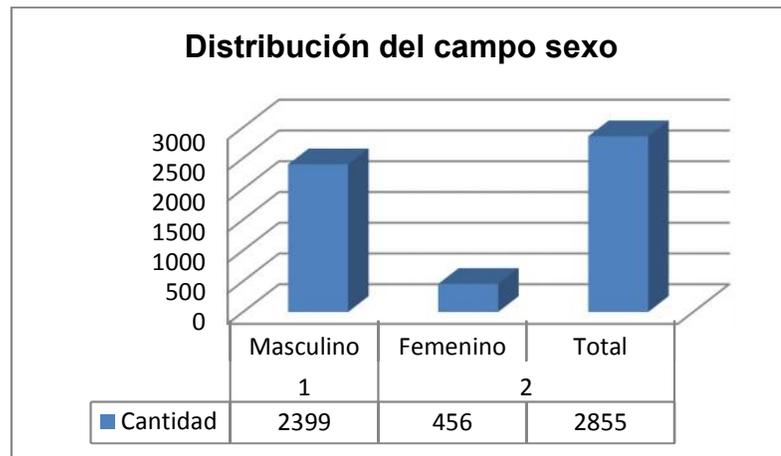
Sexo de la víctima. La clasificación y su distribución se muestran en la tabla X y la figura 16, respectivamente.

Tabla X. **Codificación y frecuencias del campo sexo**

Estado	Sexo	Cantidad
1	Masculino	2 399
2	Femenino	456
	Total	2 855

Fuente: elaboración propia.

Figura 16. **Codificación y distribución del campo sexo**



Fuente: elaboración propia.

El porcentaje de homicidios cometidos en contra de mujeres (15%) del total de los casos es bastante alto, ya que muestra que este tipo de delitos ha ido en aumento (se registró un porcentaje de 9.9% entre 2001 a 2006 y un alto repunte en el 2004 del 12.4%). Una apreciación real de las dimensiones reales de este fenómeno amerita una investigación independiente que contribuya a orientar a las autoridades públicas sobre este grave problema. Para efectos de este trabajo se analizarán los datos en toda su dimensión (masculinos y femeninos).

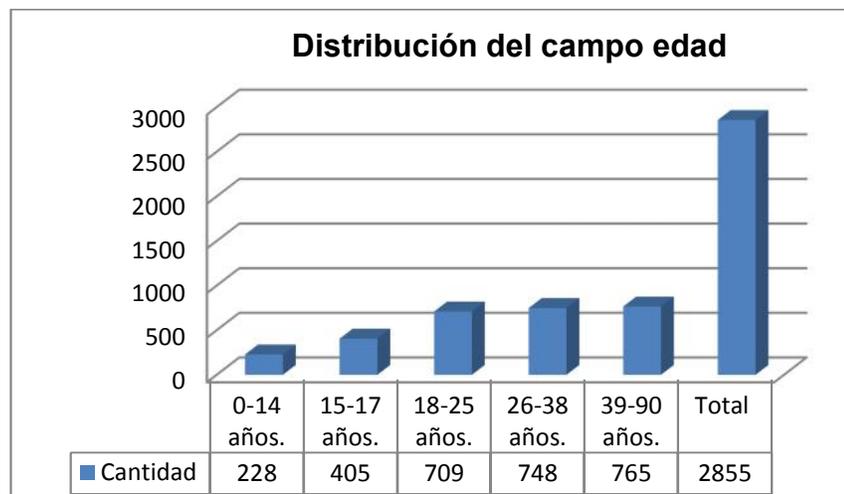
Edad: contiene el rango de edades de las víctimas, dividiéndolas en 5 rangos establecidos por las autoridades públicas para determinar por clasificación, la edad de las víctimas (niños, jóvenes, adultos, etc.). Los rangos y su distribución son los siguientes, ver tabla XI y figura 17.

Tabla XI. **Codificación y frecuencias del campo edad**

Estado	Edad	Cantidad
1	0-14 años.	228
2	15-17 años.	405
3	18-25 años.	709
4	26-38 años.	748
5	39-90 años.	765
	Total	2 855

Fuente: elaboración propia.

Figura 17. **Codificación y distribución del campo edad**



Fuente: elaboración propia.

4.3.2. Campos omitidos

- Id: contiene el número de registro de cada hecho. Se omitió porque no aporta información referida al hecho.
- Año: contiene el año en que se cometió el hecho. Se omitió debido a que todos los hechos a analizar pertenecen al mismo año.
- Folio: contiene un número de uso interno para el Sistema Judicial. Se omitió porque no aporta información referida al hecho.
- Cod_causa: contiene el código de la causa de muerte del agraviado. Se omitió porque únicamente representa un identificador asociado a una descripción de causa de muerte.
- Abreviatura: contiene la descripción abreviada de la causa de fallecimiento.
- Delito: contiene el código de delito establecido por el sistema judicial. Se omitió porque se utiliza para identificar los tipos de delitos.
- Grupo_delito: contiene la codificación en una escala superior del grupo de delitos a los cuales pertenecen los delitos individuales definidos por el Sistema Judicial.
- Tipo: contiene la codificación del tipo de delito.
- Estado: contiene el código del departamento regional en el cual ocurrió el hecho. Se omitió porque se utiliza únicamente para identificar a los departamentos de la República de Guatemala.

- Municipio: contiene el código del municipio en la cual se cometió el hecho. Se omitió porque se utiliza para identificar a cada municipio de la República de Guatemala.
- Fecha_Hecho: contiene la fecha en la cual se cometió el hecho en formato dd/mm/aaaa. Se omitió porque se analizan los campos que corresponden al Día (Día_Mes), Día semana (Dsem), Hora (Hora) y al mes (Mes) por separado.
- Delito: contiene la codificación de los demás delitos que tiene asociado un caso, ya que un sindicado puede ser acusado por más de un delito. Se omitieron los delitos que no pertenecían a los cometidos por arma blanca y arma de fuego.

4.4. Depuración de registros

Como puede verse muy pocos registros presentaban algún campo incompleto. Para solucionar este problema, sin perder la información aportada por el resto de los campos, se adoptó el criterio sugerido por especialistas de minería de datos, la cual consiste en reemplazar la información faltante por la media o la moda del campo en cuestión (según se trate de una variable continua o categórica, respectivamente).

De esta forma realizaron las siguientes modificaciones:

- Tipo_arma: se completó el único registro que no tenía asociado un tipo de arma por el tipo de arma moda del campo: arma de fuego(2645).
- Dirección_hecho: se completaron los 89 registros que no tenían especificado el tipo de lugar con la moda del campo: lugar del hecho.

4.5. Modificación de los estados originales de cada campo

A partir de este momento, para ser consistentes con la terminología de *Weka*, se tomará atributo, para referir a la información aportada por un determinado campo en el *data set*. Dicho de otra forma, lo que en el contexto de una tabla se llama campo, en el contexto del *data set* se denominará atributo.

En función del *software* y los algoritmos a utilizar se realizaron algunas modificaciones a los datos originales.

En primer lugar, *Weka* interpreta los campos que presentan estados numéricos como atributos continuos y aquellos que presentan estados alfa-numéricos, como atributos categóricos. En la base de datos original, todos los campos seleccionados presentan estados numéricos, por lo que aquellos atributos categóricos han de ser modificados.

Por otro lado, la información temporal aportada por determinados campos (*Mes, Día_mes, Día_semana y Hora*) es de naturaleza cíclica (el último estado antecede al primero). Esto no puede ser representado correctamente en *Weka*, sino que con una asignación lineal.

Esto presenta un problema a la hora de agrupar los registros, por ejemplo, la herramienta interpretará que el domingo (día 1), está muy cerca del lunes (día 2), pero muy lejos del sábado (día 7). Si bien este efecto no se puede anular por completo, sí se puede minimizar, como se verá más adelante, modificando la escala de asignación de los estados de forma de que el salto en la escala (del último al primero), tenga el menor impacto posible.

Por último, con el objetivo de hacer más fluida la lectura en la interface de *Weka*, tanto los nombres de los atributos como los de los estados de los atributos categóricos se renombraron de forma abreviada.

A continuación se explican las modificaciones realizadas a cada campo para obtener cada uno de los atributos del *data set* definitivo. Entre paréntesis se incluye el nombre de abreviado de cada atributo en el entorno de *Weka*.

- Atributo departamento (Depto)

Corresponde al campo departamento, debido a que se trata de un atributo categórico, se han reemplazado los estados originales por los siguientes, ver tabla XII.

Tabla XII. **Nuevos estados del atributo departamento**

Estado original	Cantidad	Cantidad
1	Pet	Petén
2	A.Ver	Alta Verapaz
3	Iza	Izabal
4	Huehue	Huehuetenango
5	Toto	Totonicapán
6	Qui	Quiché
7	B.Ver	Baja Verapaz
8	E.Pro	El Progreso
9	S.Mar	San Marcos
10	Quet	Quetzaltenango
11	Reu	Retalhuleu
12	Suchi	Suchitepéquez
13	Sol	Sololá
14	Chimal	Chimaltenango
15	Gua	Guatemala
16	Saca	Sacatepéquez
17	Esc	Escuintla
18	S.Rosa	Santa Rosa
19	Zac	Zacapa
20	Jal	Jalapa
21	Jut	Jutiapa
22	Chiqui	Chiquimula

Fuente: elaboración propia.

- Atributo mes

Corresponde al campo mes. Debido a que se trata de un atributo continuo con comportamiento estacionario, para minimizar el efecto antes comentado, se debe de comenzar la escala en el mes que presenta la menor cantidad de casos, en este hecho, concuerda con el mes que inicia el año (enero). De esta forma, el reemplazo de los estados originales se muestra en la tabla XIII.

Tabla XIII. **Nuevos estados del atributo mes**

Estado Original	Nuevo Estado	Mes	Cantidad
1	1	Enero	128
2	2	Febrero	136
3	3	Marzo	185
4	4	Abril	152
5	5	Mayo	209
6	6	Junio	264
7	7	Julio	254
8	8	Agosto	329
9	9	Septiembre	262
10	10	Octubre	336
11	11	Noviembre	297
12	12	Diciembre	303
		Total	2 855

Fuente: elaboración propia

- Atributo día del mes (DMes)

Corresponde al campo *Día_mes*. Si bien se trata de un atributo continuo presenta un comportamiento uniforme, por lo que no es necesaria una reclasificación.

- Atributo día de la semana (DSem)

Corresponde al campo *Día_semana*. Debido a que se trata de un atributo continuo con comportamiento estacionario, se ha decidido comenzar la escala en el día de la semana que presenta la menor cantidad de casos (viernes). De esta forma se han reemplazado los estados originales por los siguientes, ver tabla XIV.

Tabla XIV. **Nuevos estados del atributo día de la semana**

Estado Original	Nuevo Estado	Día semana	Cantidad
1	4	Lunes	499
2	5	Martes	433
3	6	Miércoles	392
4	7	Jueves	385
5	1	Viernes	353
6	2	Sábado	362
7	3	Domingo	431
		Total	2 855

Fuente: elaboración propia.

- Atributo hora

Corresponde al campo hora. En este caso puntual se decidió comenzar la escala en el intervalo 8-12 horas, ya que existe una barrera natural entre este intervalo y el anterior (originada por el hecho de que a las 8 horas comienza la actividad laboral en el Sistema Judicial del país). Por esta razón se han reemplazado los estados originales por los siguientes, ver tabla XV.

Tabla XV. **Nuevos estados del atributo hora**

Estado Original	Nuevo Estado	Hora	Cantidad
1	5	0-4 horas.	289
2	6	4-8 horas.	321
3	1	8-12 horas.	505
4	2	12-16 horas.	460
5	3	16-20 horas.	550
6	4	20-24 horas.	730
		Total	2 855

Fuente: elaboración propia.

- Atributo dirección

Corresponde al campo Direccion_hecho. Debido a que se trata de un atributo categórico, se han renombrado los estados originales por los siguientes, ver tabla XVI:

Tabla XVI **Nuevos estados del atributo lugar**

Estado	Nuevo Estado	Dirección_Hecho	Cantidad
1	z.1	zona 1	741
2	z.2	zona 2	178
3	z.3	zona 3	90
4	z.4	zona 4	159
5	z.5	zona 5	223
6	z.6	zona 6	104
7	z.7	zona 7	211
8	z.8	zona 8	175
9	z.9	zona 9	29
10	z.10	zona 10	41
11	z.11	zona 11	280
12	z.12	zona 12	114
13	z.13	zona 13	32
14	z.14	zona 14	1
15	z.15	zona 15	38
16	z.16	zona 16	40
17	z.17	zona 17	13
18	z.18	zona 18	192
19	z.19	zona 19	47
20	z.21	zona 21	17
21	L.Hech	Lugar del hecho	130
		Total	2 855

Fuente: elaboración propia.

- **Atributo arma**

Corresponde al campo Tipo_arma, debido a que se trata de un atributo categórico, se han renombrado los estados originales por los siguientes, ver tabla XVII:

Tabla XVII Nuevos estados del atributo arma

Estado Original	Nuevo Estado	Descripción	Cantidad
1	Blanca	Herida por arma blanca	209
2	Fuego	Herida por arma de fuego	2 645
3	Otra	Por establecer	1
		Total	2 855

Fuente: elaboración propia.

- **Atributo sexo**

Debido a que se trata de un atributo categórico, se han renombrado los estados originales por los siguientes, ver tabla XVIII.

Tabla XVIII Nuevos estados del atributo sexo

Estado Original	Nuevo Estado	Sexo	Cantidad
1	M	Masculino	2 399
2	F	Femenino	456
		Total	2 855

Fuente: elaboración propia

- Atributo edad

Al igual que el campo *Dia_mes* se trata de un atributo continuo que presenta un comportamiento uniforme, por lo que no es necesaria una reclasificación.

4.6. Data set definitivo

Finalmente el *data set* queda compuesto de los 2855 registros originales y 7 atributos que aportan información de distintas dimensiones de los hechos:

- Espacial: información sobre la distribución geográfica de los hechos representada por el atributo departamento.
- Temporal: información sobre la distribución temporal de los hechos representada por los atributos mes, día del mes, día de la semana y hora.
- Circunstancial: información sobre el modo en que ocurrieron los hechos representada por los atributos tipo arma y dirección hecho.

5. PRESENTACIÓN DEL CASO

5.1. Introducción

A continuación se interpretan un conjunto de tablas y gráficos que serán de utilidad para el análisis de los resultados en el próximo capítulo. El objetivo es que el lector se familiarice en la comprensión y lectura de los mismos, al mismo tiempo, que con la metodología de análisis. En forma didáctica se presenta un caso de aplicación.

Se extrajo una muestra de la base de datos, correspondiente a la totalidad de homicidios dolosos ocurridos en la fiscalía de Chimaltenango.

Cabe aclarar que, si bien se seleccionó un *data set* de pocos registros para que pueda ser visualizado e interpretado por el lector, no tiene sentido práctico aplicar técnicas de minería de datos sobre tan poca cantidad de información. Por lo tanto, el análisis y las conclusiones de este caso puramente didáctico no deben ser considerados.

Los atributos seleccionados para el ejemplo fueron: lugar, arma, día de la semana (DSem) y hora (Hora), ver tabla XIX.

Tabla XIX. Descripción de atributos de la muestra

Atributo	Tipo	Variables	Rango	Descripción
Lugar	Categórico	z.1		zona 1
		z.2		zona 2
		z.3		zona 3
		z.4		zona 4
		z.5		zona 5
		z.6		zona 6
		z.7		zona 7
		z.8		zona 8
		z.9		zona 9
		z.10		zona 10
		z.11		zona 11
		z.12		zona 12
		z.13		zona 13
		z.14		zona 14
		z.15		zona 15
		z.16		zona 16
		z.17		zona 17
		z.18		zona 18
		z.19		zona 19
				z.21
		L.Hech		Lugar del hecho
Arma	Categórico	Fuego		Arma de fuego
		Blanca		Arma blanca
		Otra		Ninguna arma
Dsem	Continuo		1 al 7	1 corresponde al viernes
Hora	Continuo		1 al 6	6 intervalos de 4 horas c/u empezando a las 0hs.

Fuente: elaboración propia.

Tras aplicar el algoritmo *K-means* se obtuvieron 2 *clusters*. En la tabla XX se muestra la pertenencia de cada registro a cada *cluster* (denominados 0 y 1).

5.2. Descripción de las herramientas

A continuación se describen las herramientas utilizadas para el análisis de los datos.

5.2.1. Tabla de centroides

La tabla de centroides permite conocer cuál es el centroide de cada *cluster*. En un sentido práctico no es más que la media o la moda de cada atributo para cada *cluster*, ver tabla XX.

Tabla XX. **Asignación de clusters al data set**

Caso	Lugar	Arma	Dsem	Hora	Cluster	Caso	Lugar	Arma	Dsem	Hora	Cluster
1	z.1	Fuego	1	0	1	21	z.5	fuego	4	20	1
2	z.1	Fuego	1	22	1	22	z.4	fuego	5	19	1
3	z.1	Fuego	1	19	1	23	z.4	fuego	5	20	1
4	z.1	Blanca	1	9	1	24	z.4	fuego	5	23	1
5	z.1	blanca	2	18	0	25	z.1	fuego	5	14	1
6	z.5	fuego	2	3	1	26	z.18	fuego	5	14	1
7	z.4	fuego	2	11	1	27	z.1	fuego	5	13	1
8	z.1	fuego	2	15	1	28	z.4	fuego	6	14	1
9	z.1	blanca	2	6	0	29	z.18	fuego	6	18	1
10	z.4	fuego	2	15	1	30	z.7	fuego	6	11	1
11	z.8	fuego	2	23	1	31	z.1	fuego	6	17	1
12	z.1	fuego	2	21	1	32	z.1	fuego	6	22	1

Continuación tabla XX.

13	z.5	fuego	2	11	1	33	z.7	fuego	7	16	1
14	z.7	fuego	3	12	1	34	z.7	fuego	7	19	1
15	z.2	fuego	3	20	1	35	z.5	fuego	7	21	1
16	z.2	blanca	3	15	0	36	z.8	fuego	7	23	1
17	z.4	fuego	3	8	1	37	z.12	fuego	7	9	1
18	z.5	fuego	3	20	1	38	z.8	fuego	7	11	1
19	z.18	fuego	4	16	1	39	z.8	fuego	7	18	1
20	z.4	fuego	4	20	1						

Fuente: elaboración propia

En el caso en cuestión la tabla de centroides es la siguiente.

Tabla XXI. **Tabla de centroides**

	Cantidad(%)	Atributos categóricos(modas)		Atributos Continuos (medias)	
		Lugar	Arma	Hora	Día de la Semana
Cluster 0	10%	z.1	Blanca	12	2
Cluster 1	90%	z.1	Fuego	16	3.22
General	100%	z.1	Fuego	14	3

Fuente: elaboración propia.

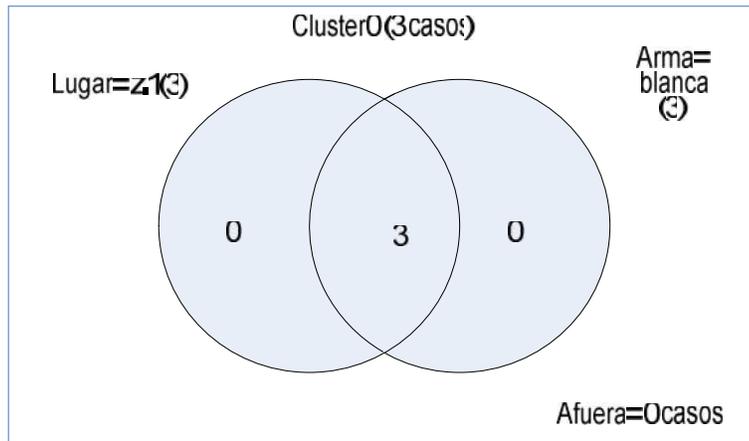
Es importante tener especial cuidado en la interpretación de las modas de los atributos categóricos. La correcta lectura debe hacerse en cada atributo por separado, independientemente del resto. Esto significa, por ejemplo, para el *cluster* 1 la lectura debe ser la mayoría de los casos fue en z.1, independientemente de esto, la mayoría de los casos fue con *arma de fuego*, lo que no es equivalente a decir que la mayoría de los casos fue *en z.1* y con *arma de fuego* (cosa que para este caso puntual también se cumple pero no necesariamente es así).

La moda indica una mayoría relativa pero no cuantifica, por lo que a priori no se conoce la representatividad de una determinada variable para identificar un determinado *cluster*. Por ejemplo, el arma de fuego para el *cluster* 1 es sumamente representativo, ya que corresponde al 100% de los casos, y blanca para el *cluster* 0 también representa el 100% de los casos, sin embargo debemos tener cuidado ya que si hubiera aparecido otra variable como la de por establecer hubiera modificado el estado de cualquiera de los *cluster* y el resultado pudiera no ser tan representativo.

5.2.2. Diagramas de Venn

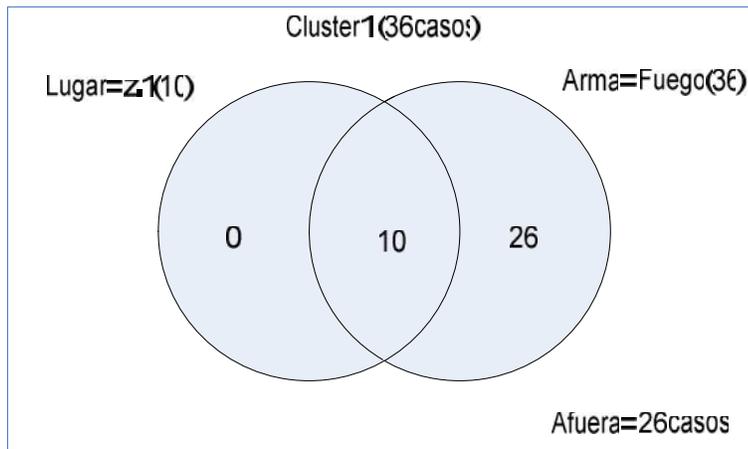
Los Diagramas de Venn ayudan a visualizar los niveles de representatividad y solapamiento. En este caso, al tratarse de pocos datos, quedan sub conjuntos vacíos que no son comunes cuando hay mayor cantidad de registros.

Figura 18. **Diagrama de Venn para el cluster 0**



Fuente: elaboración propia.

Figura 19. **Diagrama de Venn para el cluster 1**



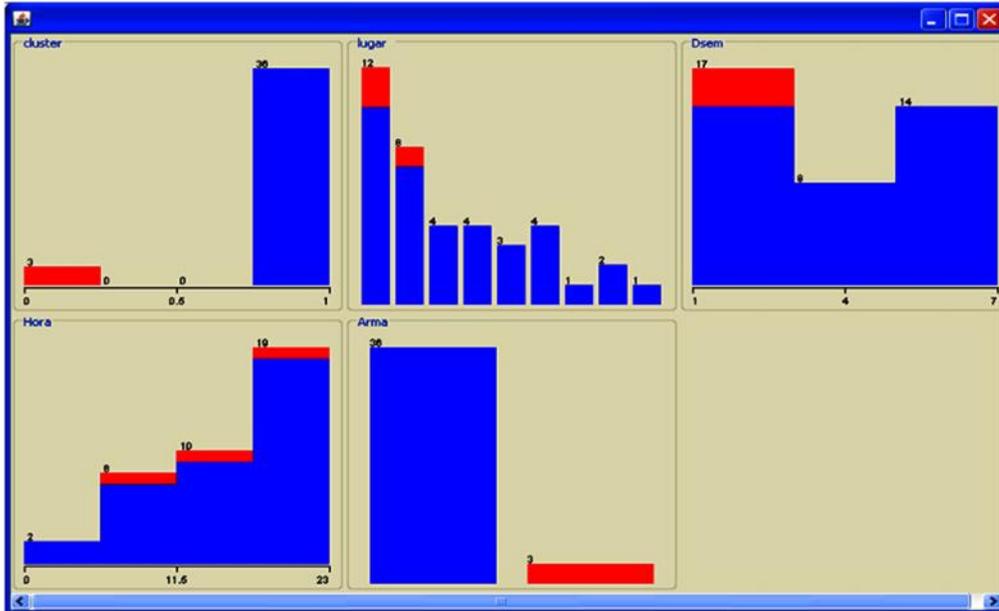
Fuente: elaboración propia.

5.2.3. Gráficos de barras

Los gráficos de barras permiten visualizar la distribución de un determinado atributo, según las variables de los demás. En este caso se utilizará el atributo *cluster*, ver figura 20.

Este tipo de gráfico es fundamental para comprender la relevancia de los *clusters*. Si la asignación a los *clusters* hubiera sido aleatoria, entonces sería lógico esperar que la proporción asignada a cada uno (en este caso 8% para el *cluster* 0 rojo y 82% para el *cluster* 1 azul), se mantenga aproximadamente igual independientemente de a través de qué variable se haga la segmentación. Dicho de otra forma, bajo la hipótesis de aleatoriedad no cabría esperar ningún patrón de distribución especial de los *clusters* en el resto de los atributos.

Figura 20. **Gráficos de barras: distribución de los *clusters* según el resto de los atributos**



Fuente: elaboración propia.

Por esta razón, toda distribución dentro de una determinada variable que se aparte de la distribución global (10-90 en este caso) será motivo de un análisis más detallado, ya que podría estar identificando la verdadera naturaleza del *cluster*.

En este caso, por ejemplo, se observa un comportamiento particular en el tipo de arma (hay una relación entre arma de fuego y el *cluster* 1 y arma blanca y *cluster* 1) pero no así, tanto en el lugar, ya que son varios donde ocurrieron los hechos. También hay cierto patrón de comportamiento en los atributos continuos, fundamentalmente en la hora (una alta incidencia del *cluster* 1 después de las 16 horas).

Este gráfico servirá para entender cuáles son los atributos que están caracterizando a los *clusters* y continuar el análisis poniendo énfasis en los mismos. Esta gráfico tiene la desventaja que sólo permite visualizar dos atributos por gráfico (el *cluster* y otro atributo).

5.2.4. Gráficos de dispersión

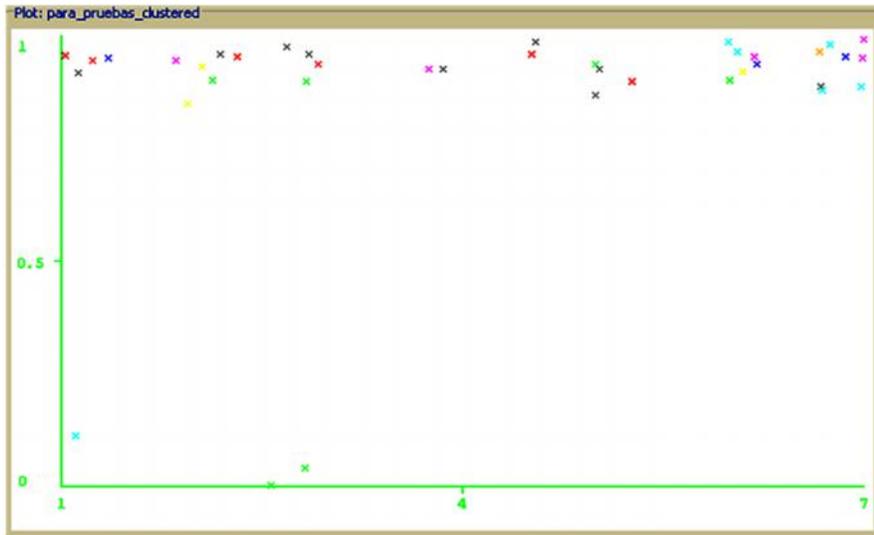
Los gráficos de dispersión se representan en ejes cartesianos: cada eje representa un atributo y cada punto un hecho. Éstos tienen la particularidad que permiten incorporar virtualmente una tercera dimensión mediante la asignación de distintos colores a los puntos. Existen dos tipos de gráficos que se describen a continuación.

5.2.4.1. Gráficos de distribución

Los gráficos de dispersión es donde la asignación del color coincide con el eje de ordenadas, permitiendo visualizar la distribución de un atributo en función de otro. Aportan información similar a los gráficos de barras, pero con otro enfoque.

Por ejemplo, el gráfico de distribución de los *clusters* a lo largo de la semana es el siguiente.

Figura 21. Distribución de los *clusters* según *día de la semana*



Fuente: elaboración propia.

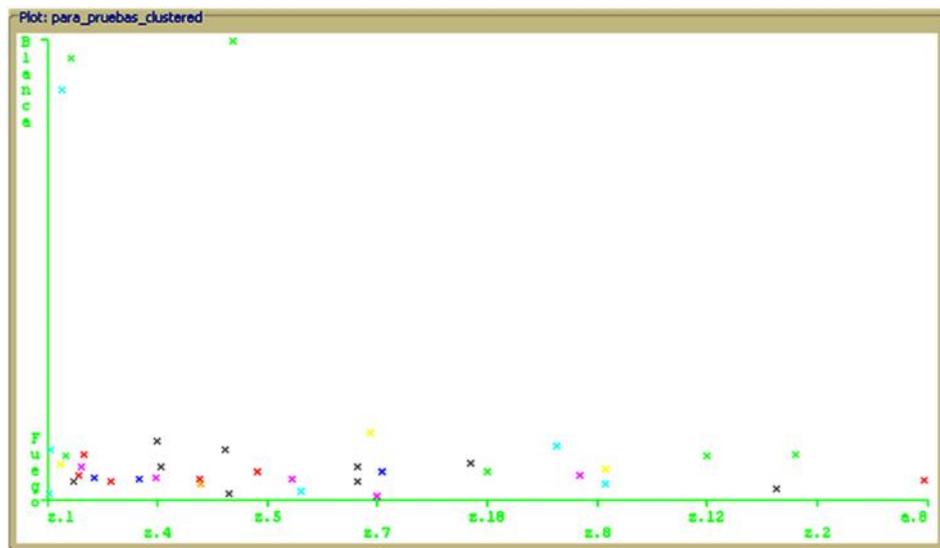
Como se puede ver en el eje de las abscisas se encuentran representados los días de la semana, mientras que en el de las ordenadas, ambos *clusters*. Dado que el día 1 representa al viernes, puede verse que el *cluster* 1 presenta una alta proporción de casos durante el intervalo de días domingo a martes.

5.2.4.2. Gráficos de interrelaciones

Estos gráficos de dispersión permiten visualizar 3 atributos al mismo tiempo e identificar cuál es la interrelación que subyace entre ellos. Por lo general el atributo que se encuentra en la dimensión de color es el *cluster* (variable a explicar).

Por ejemplo el gráfico de interrelación arma-lugar es el siguiente:

Figura 22. Interrelación arma-lugar



Fuente: elaboración propia.

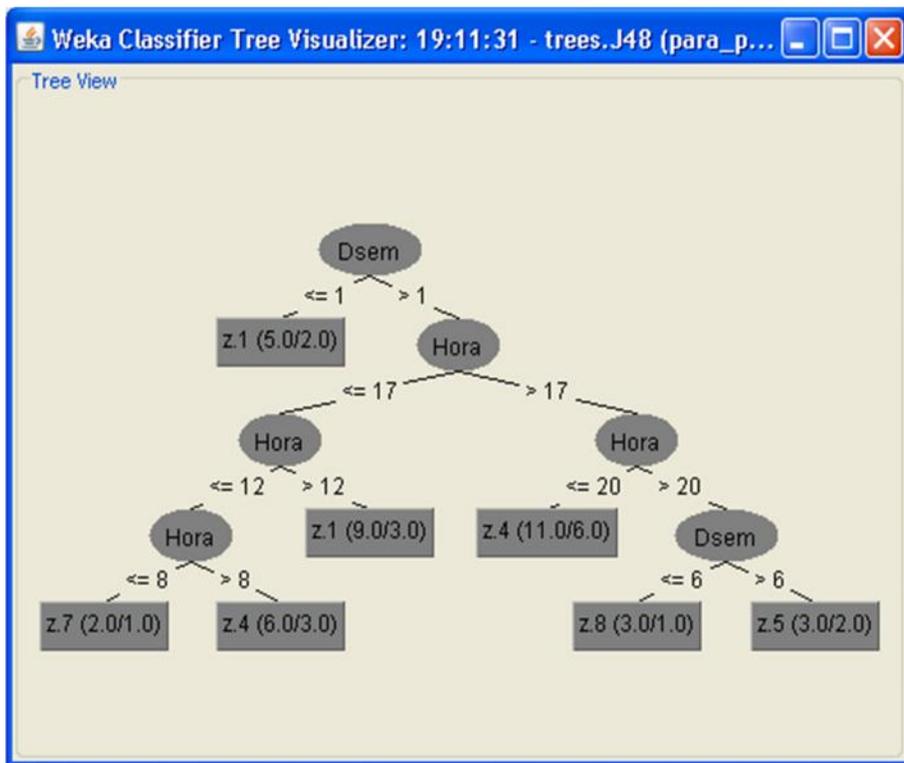
Para un determinado cruce de variables se observa una acumulación de puntos de un determinado tipo, entonces se dice que se trata de una interacción entre las variables. La misma podrá ser más fuerte o más débil, según la cantidad de puntos acumulados haya en relación a la cantidad total de puntos del mismo color.

En este caso hay una fuerte interacción entre arma de fuego y lugar del hecho, así como arma de fuego y lugar del hecho, ambas para el *cluster* 1.

5.2.5. Árbol de clasificación

Luego del análisis anterior se utiliza el algoritmo C4.5 para identificar las reglas de pertenencia a cada *cluster* de una manera formal. El resultado de este algoritmo se da en forma de árbol:

Figura 23. Árbol de clasificación



Fuente: elaboración propia.

Los árboles presentan nodos, en donde se evalúa un determinado atributo (en este caso Dsem y hora); las ramas que surgen de cada nodo, en donde se representan los estados posibles que puede tomar el atributo del nodo; y hojas, en donde se muestra la clasificación a cada clase (en este caso *dirección del hecho*). En las hojas se muestra la cantidad total de registros clasificados y, *separados* con una barra, la cantidad de registros mal clasificados (si los hubiera).

La lectura de un árbol se realiza en forma de reglas de clasificación. Existe una regla para cada hoja:

Regla 1

SI Dsem \leq 1

entonces z.1 (5.0/1.0)

Regla 2

SI Dsem $>$ 1

y Hora \leq 8

entonces z.7 (2.0/1.0)

Regla 3

SI Dsem $>$ 1

y Hora $>$ 12

entonces z.1 (9.0/3.0)

Regla 4

SI Dsem \leq 6

y Hora \leq 20

entonces z.8 (3.0/1.0)

Regla 5

SI DSem > 6

y Hora > 6

entonces z.5 (3.0/2.0)

5.2.6. Matrices de confusión

Las matrices de confusión permiten entender cuál es el error que comete un árbol de clasificación al intentar clasificar todos los registros.

Para este caso la matriz de confusión es:

Figura 24. **Matriz de confusión**

```
=== Confusion Matrix ===
  a b c d e f g h i  <-- classified as
5 5 1 1 0 0 0 0 0 | a = z.1
2 4 0 1 0 1 0 0 0 | b = z.4
3 1 0 0 0 0 0 0 0 | c = z.5
3 1 0 0 0 0 0 0 0 | d = z.7
3 0 0 0 0 0 0 0 0 | e = z.18
2 2 0 0 0 0 0 0 0 | f = z.8
0 1 0 0 0 0 0 0 0 | g = z.12
0 1 0 1 0 0 0 0 0 | h = z.2
0 0 0 0 0 1 0 0 0 | i = a.8
```

Fuente: elaboración propia.

Si la clasificación hubiera sido perfecta, se esperaría encontrar únicamente elementos en la diagonal.

6. RESULTADOS

Se analizó el *data set* obtenido en el capítulo 5 con el *software Weka*. En primer lugar se aplicó el algoritmo *K-means* para agrupar los 2855 registros en 2 *clusters*. Luego, con las herramientas descritas en el capítulo anterior, se obtuvo una primera caracterización de los *clusters* y finalmente se utilizó el algoritmo *C4.5* para una interpretación formal y definitiva.

6.1. *Clustering*

Procediéndose a agrupar el *data set* en 2 grupos (identificados por el tipo de arma) utilizando el algoritmo *K-means*.

Para la ejecución de este algoritmo es necesario seleccionar un número, denominado semilla, para realizar una distribución aleatoria inicial a partir de la cual el algoritmo comience las iteraciones sucesivas. En la selección de este número se realizaron 20 corridas consecutivas (después de esta cantidad se notó que la suma del error cuadrático tendía al aumento y no a la disminución) probando distintas semillas y se seleccionó aquella que minimizaba la suma del error cuadrático. Si bien este método heurístico no garantiza la semilla óptima, asegura una relativamente buena asignación. A continuación se presentan los resultados obtenidos para las 20 corridas:

Tabla XXII. **Resultado de *K-Means* para 3 clusters con varias semillas**

Semilla	Suma error Cuadrático	No. Iteraciones
1	4997.61	14
2	4993.24	8
3	4934.12	4
4	4976.56	10
5	4890.72	3
6	4993.24	20
7	4993.24	10
8	5064.58	9
9	4993.24	6
10	4993.24	6
11	4993.24	7
12	5060.45	8
13	4890.72	3
14	5068.86	10
15	4993.24	8
16	5068.86	10
17	4090.72	3
18	5006.04	7
19	4993.24	7
20	4993.24	7

Fuente: elaboración propia.

Como se puede ver la menor suma de error cuadrático se obtuvo con una semilla de 17 y 3 iteraciones.

6.1.1. Tabla de centroides

El resultado obtenido con *Weka* tras la ejecución de *simple K-means* con 2 *clusters*, una semilla de 17 y 3 iteraciones se resume en la tabla de centroides XXIII. Si bien las medias de los atributos continuos para cada *cluster* se encuentran muy cerca de la media global, no ocurre lo mismo con las modas de los atributos categóricos.

Existe cierta alternancia entre las modas de los atributos lugar y arma parecerían estar identificando a los *clusters*.

Tabla XXIII. Tabla de centroides

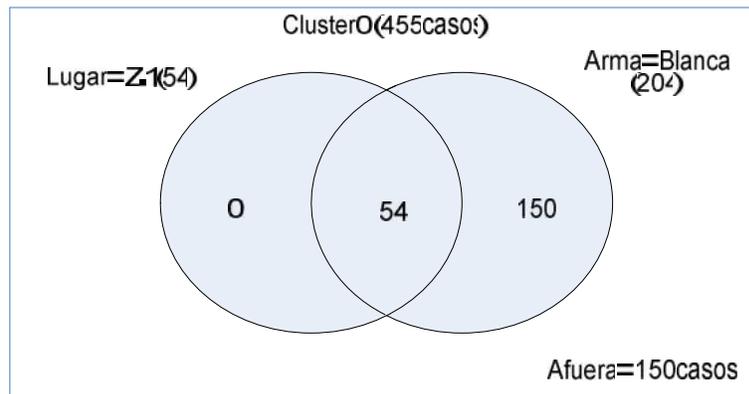
	Cantidad (%)	Atributos categóricos(modas)				Atributos Continuos (medias)				
		Depto	Lugar	Arma	Sexo	Hora	Día de la Semana	Día del mes	Mes	Edad
Cluster 0	84%	Gua	z.1	Fuego	M	16	Lunes	16	7	3
Cluster 1	16%	Gua	z.1	Fuego	F	16	Lunes	16	7	3

Fuente: elaboración propia.

6.1.2. Diagramas de Venn

Como se explicó en la sección 6.3.1, los centroides no necesariamente representan la combinación de atributos más frecuente, por lo tanto es necesario un análisis más detallado para caracterizar a los *clusters*. A continuación se muestran diagramas de Venn que indican la composición real de cada *cluster* según los atributos departamento, lugar y arma:

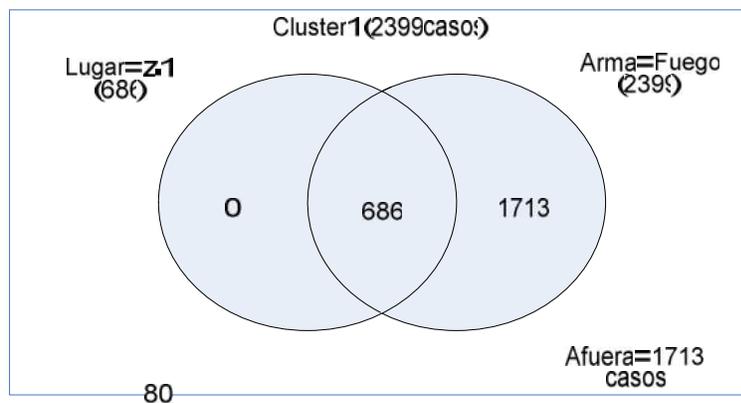
Figura 25. Diagrama de Venn para atributos categóricos del *cluster 1*



Fuente: elaboración propia.

El 26% de los registros agrupados en este *cluster*, cumple con el requisito de homicidios cometidos con arma blanca. Este *cluster* parecería estar caracterizado, de forma más difusa que el 0, ya que este tipo de crímenes pueden ser difícil de catalogarlos.

Figura 26. Diagrama de Venn para atributos categóricos del *cluster 0*



Fuente: elaboración propia.

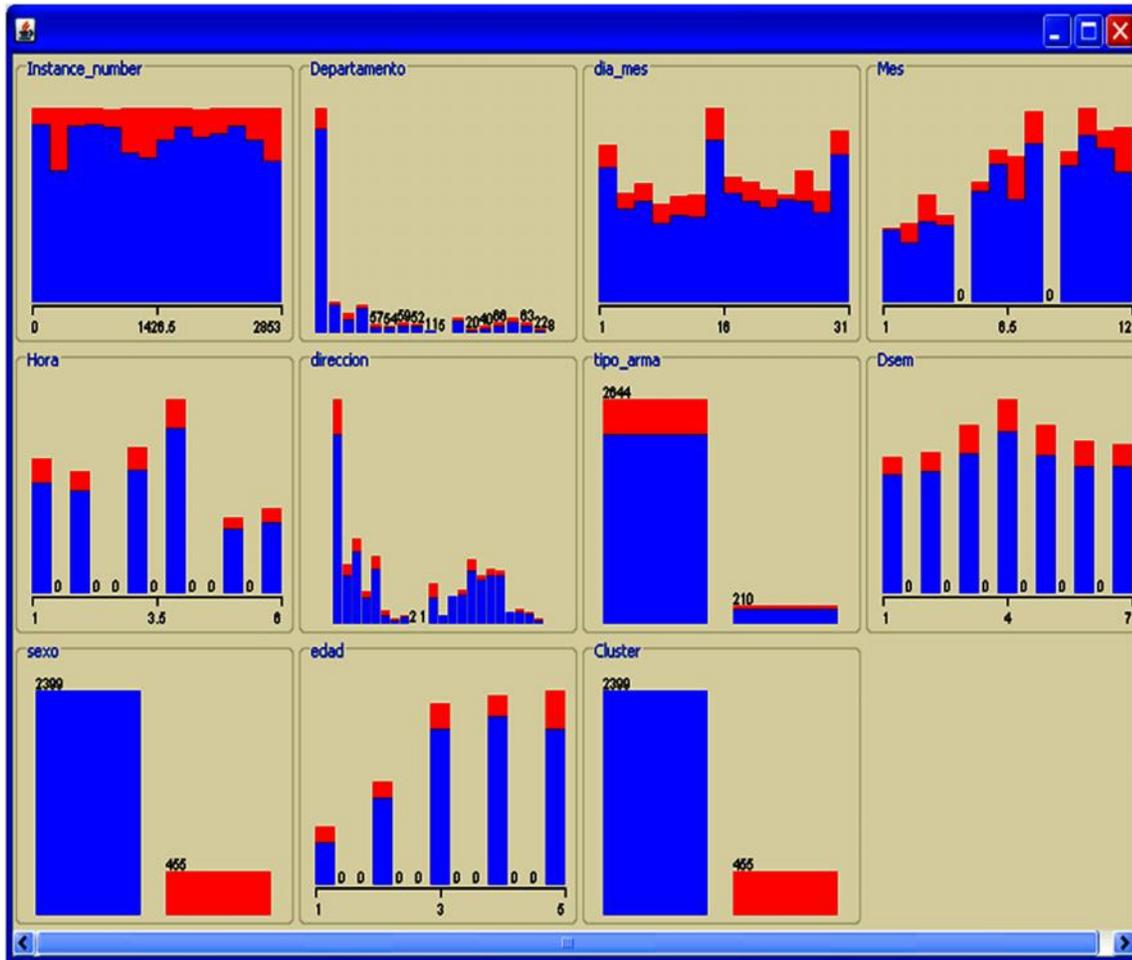
El 100% de los registros agrupados en este *cluster* cumple con el atributo de homicidios cometidos con arma de fuego. Este *cluster* parecería estar caracterizado por homicidios premeditados (75% de los casos). La z.1 es otra característica de este *cluster*, ya que el 28% de los casos fueron cometidos en este lugar.

6.1.3. Gráficos de barras

Como se mencionó en la sección 6.3.3, la distribución de los clusters entre las variables de los distintos atributos permite comprender el nivel de significancia de los mismos.

En este caso, si los *clusters* fueran irrelevantes, se esperaría encontrar una proporción aproximada de 7% rojo (*cluster* 1) y 93% azul (*cluster* 0) en cada variable de cada atributo. Si bien en algunos atributos esta proporción se cumple (día_mes y departamento), en otros existen interacciones significativas (por ejemplo cluster 2 con arma fuego y lugar del hecho).

Figura 27. Distribución de los *clusters* según atributos



Fuente: elaboración propia.

Los atributos donde se observan más interacción entre las variables y los *clusters* son: dirección_hecho, tipo_arma, departamento, hora, día_semana y edad.

6.1.4. Gráficos de dispersión

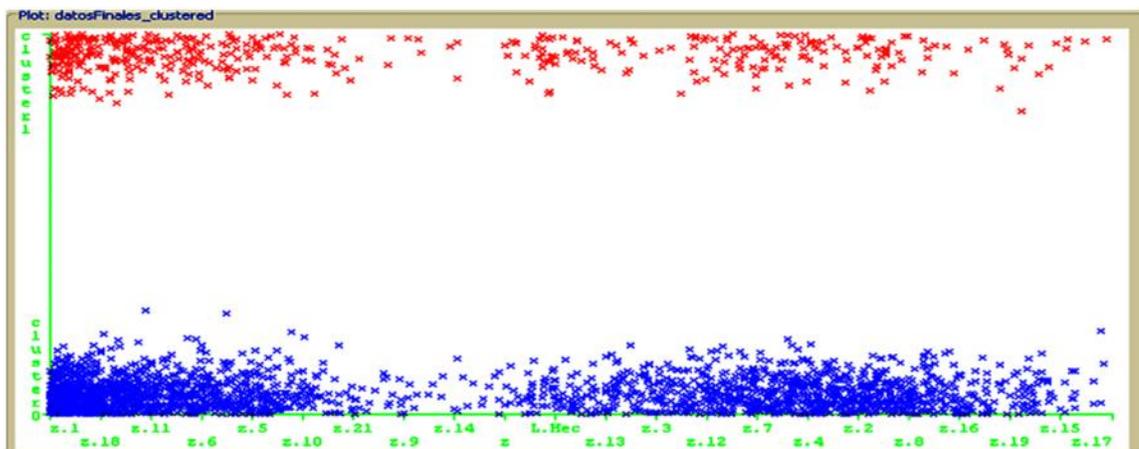
Son utilizados para representar las distribuciones de los *clusters*.

6.1.4.1. Distribución de los *clusters* según el atributo lugar

El *cluster* 0 y el *cluster* 1 están muy concentrados en el campo en ciertas áreas del campo lugar del hecho y su distribución en los demás lugares es un poco más homogénea, ambos se concentra en lugar del hecho {z.1, z.11, z.18}, y el resto de su distribución es un poco más homogéneo.

Esto indica que la mayor concentración de delitos se dan en las zonas 1, 11, 18 y en un poco menos de concentración en las zonas 7, 4, 2; como se muestra en la figura 28.

Figura 28. Distribución de *clusters* según atributo *lugar*

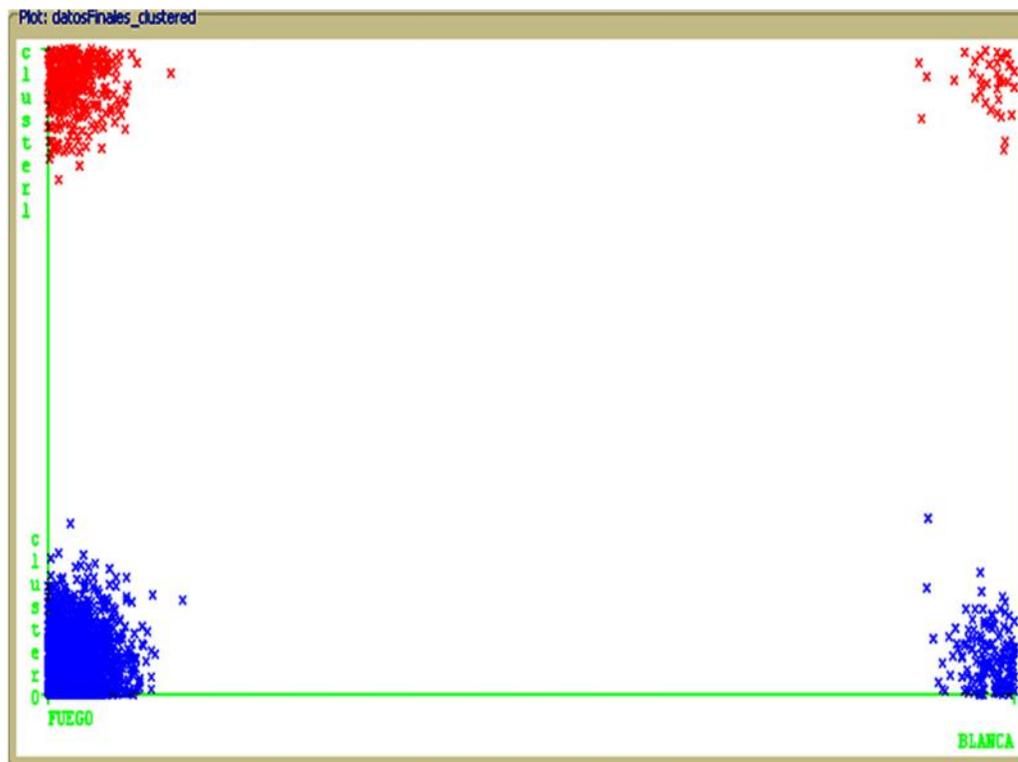


Fuente: elaboración propia.

6.1.4.2. Distribución de los *clusters* según el atributo arma

El *cluster* 1 y el *cluster* 0 presentan una distribución similar, con una alta concentración en arma de fuego, seguida por arma blanca. Podría haber otras distribuciones si se tuvieran registros de homicidios en los cuales no estuviera registrada el arma utilizada; esta sería otra variable a estudiar. La mayoría de homicidios se cometen con arma de fuego, que es donde se observa la mayor cantidad de concentración en el *cluster* 0.

Figura 29. Distribución de clusters según atributo *arma*

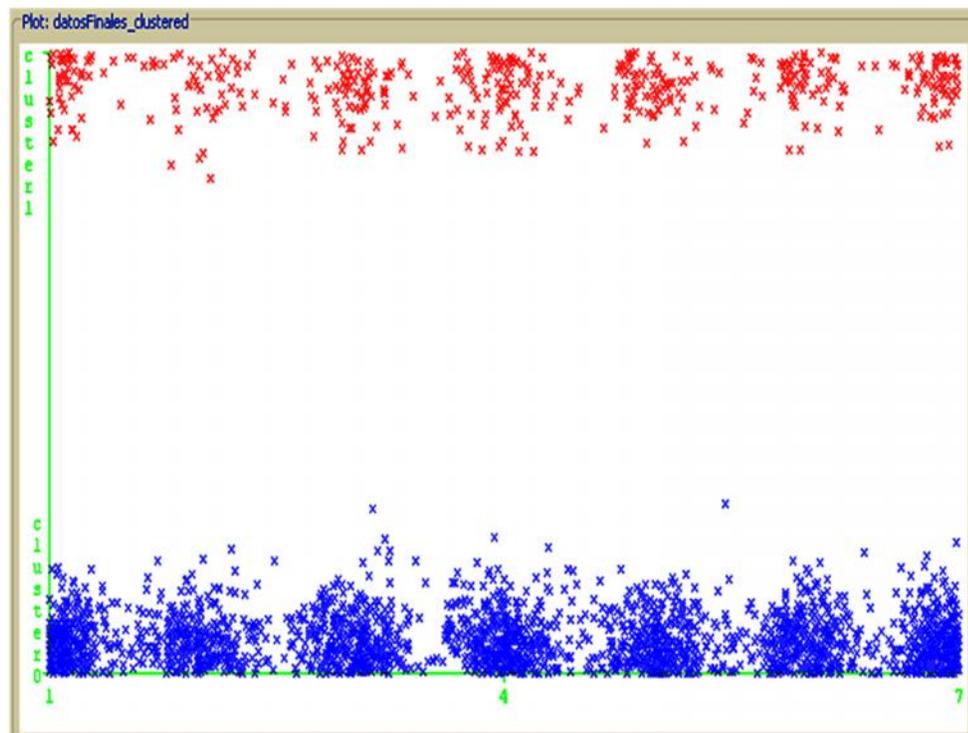


Fuente: elaboración propia.

6.1.4.3. Distribución de los clusters según el atributo día de la semana

El *cluster* 0 presenta una distribución poco homogénea, se centra más en los días del 1 al 4, lo que indica que en los últimos días de la semana es cuando más homicidios ocurren con este tipo de arma (recordando que para esta distribución el día 1 es viernes). Mientras el *cluster* 1, presenta una distribución más homogénea, con una mayor concentración de casos en los días 1 y 7; como se muestra a continuación:

Figura 30. Distribución de *clusters* según atributo *día de la semana*

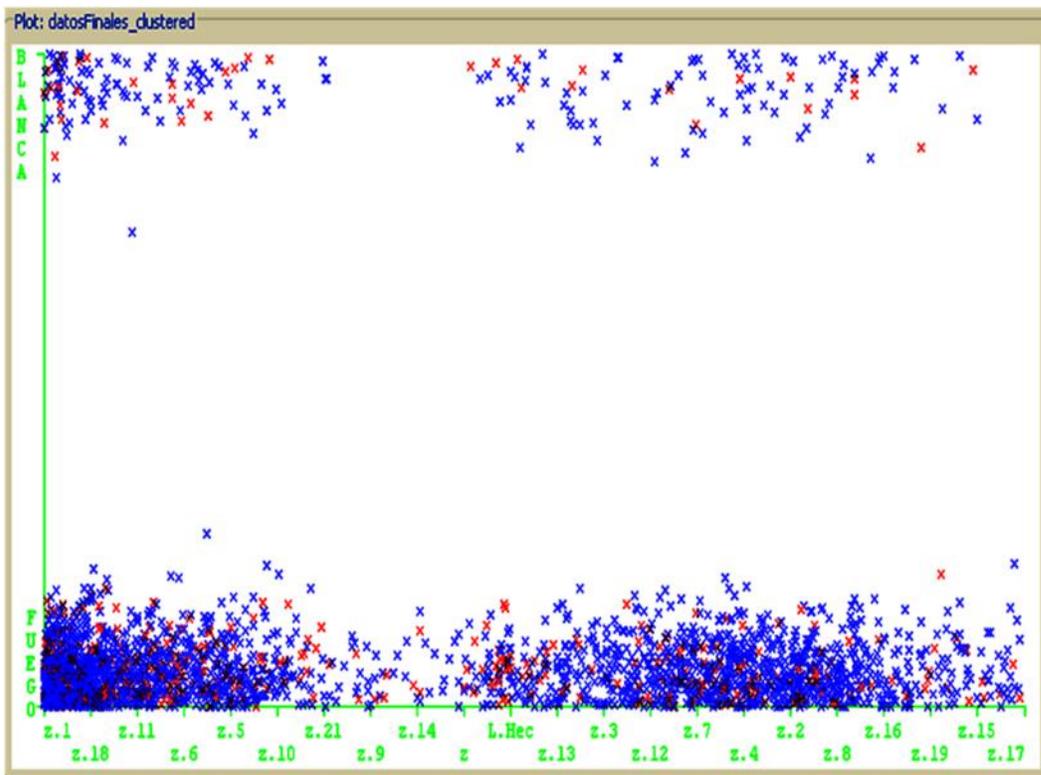


Fuente: elaboración propia.

6.1.4.4. Interrelación lugar-arma

Existe una fuerte interacción entre las zonas 1,11 y 18, arma de fuego y *cluster* 0. En un nivel más general se puede interpretar al *cluster* 1 como homicidios en zonas 1, 11 ,18 donde el arma no es arma de fuego. También se observa interacción muy grande entre zonas 1, 11 ,18 arma de fuego y *cluster* 0, así como una leve interacción entre zonas 7 ,4 ,8 y 2 y arma de fuego; como se muestra en la siguiente figura:

Figura 31. Interrelación *lugar-arma*



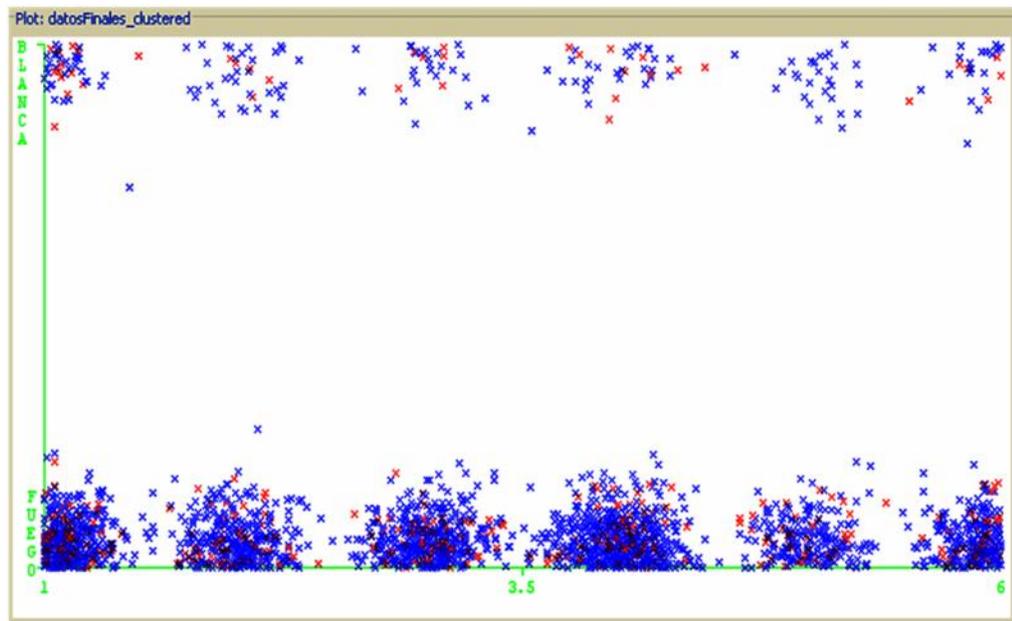
Fuente: elaboración propia.

Cabe resaltar que la incidencia entre arma de fuego y L.hech no tiene una interpretación real, debido a que el fallecimiento de la persona se registra en el lugar del hecho y no en la dirección real en la cual ocurrió.

6.1.4.5. Interrelación hora-tipo_arma

Se observa una distribución homogénea entre el tipo de arma y hora en la cual ocurre el hecho. Esta distribución muestra la alta incidencia de hechos que ocurren en las últimas horas del día (recordando que el valor 4 representa al intervalo de las 20-24 horas.) y el tipo de arma utilizada que es de fuego, podría interpretarse esta gráfica como la mayoría de homicidios a altas horas de la noche son cometidos con armas de fuego.

Figura 32. Interrelación *hora-tipo_arma*

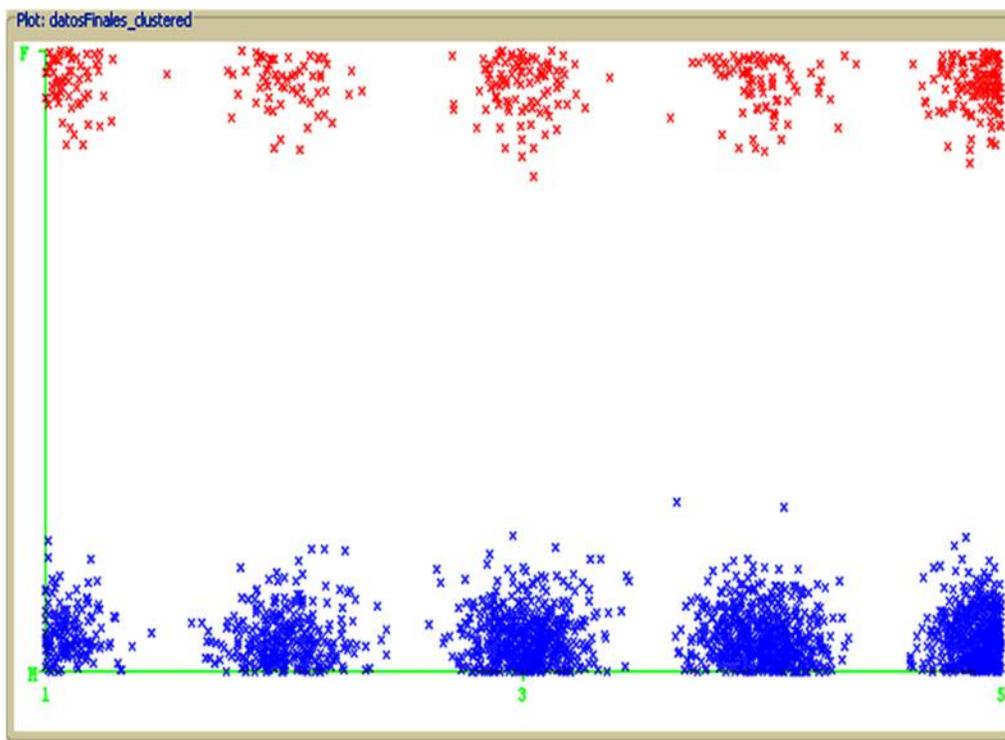


Fuente: elaboración propia.

6.1.4.6. Interrelación hora-tipo_arma

Para efecto del análisis de homicidios por sexo, se puede decir que la mayoría de mujeres asesinadas son de 29- 90 años, la mayoría de estos delitos son cometidos con armas de fuego y se muestran en la figura 33.

Figura 33. Interrelación sexo-tipo_arma



Fuente: elaboración propia.

6.1.5. Primera interpretación

Con base en la información que surge de este análisis puede darse una primera interpretación a los *clusters*:

- *Cluster 1* (16%): está caracterizado por homicidios mayoritariamente y con arma blanca. En principio se dice que se trata de homicidios en ocasión de robo.
- *Cluster 0* (84%): es el que más registros agrupa y el más parecido a la media global. Está caracterizado por homicidios mayoritariamente en la vía pública y lugar del hecho con arma de fuego. Estos registros se podrían interpretar como homicidios en ocasión de riña o ajuste de cuentas, ya que la mayoría de estos delitos ocurren en ataques a objetivos directos y a pilotos del transporte público.

6.2. Aplicación de c4.5 para la clasificación de los *clusters*

Para la aplicación del algoritmo C4.5 a los datos, se utiliza la selección de atributos descrita a continuación.

6.2.1. Selección de atributos

Weka cuenta con un set de métodos para preseleccionar los atributos que serán utilizados posteriormente en algoritmos TDIDT. Algunos actúan en función del algoritmo de inducción a utilizar, obteniendo como resultado los atributos óptimos, mientras que otros los hacen genéricamente, definiendo un ranking de atributos.

Se evaluó el set de los 8 atributos con los principales métodos. En la tabla XXIV se muestran los resultados obtenidos (en los casos de *ranking* se muestra entre paréntesis el puntaje obtenido para los principales atributos).

Tabla XXIV. **Resultado de selección de atributos**

Evaluador	Método de búsqueda	Atributos							
		depto	dirección hecho	arma	hora	dsem	mes	sexo	edad
Wrapper	Exhaustive	X		X		X	X		x
Clasifier	Exhaustive								
Consistency	Exhaustive	X	X		X	X	X	x	x
Cfs	Exhaustive	X				X			x
ChiSquare	Ranker	1519.466	220.401	19.237	15.308	600.784	68.774	2854	44.307
Gain Ratio	Ranker	0.18011	0.04484	0.01179	0.00451	0.08903	0.0134	1	0.0086

Fuente: elaboración propia.

Como se puede observar, todos los métodos coinciden en tomar al menos los atributos departamento, dirección_hecho, tipo_arma, dsem, mes, sexo y edad.

6.2.2. Resultados de C4.5 con todos los atributos

Se corrió el algoritmo C4.5 (J48 en la terminología de *Weka*) con todos los atributos analizados. En la tabla XXV se observan el porcentaje de registros correctamente clasificados con C4.5, con base en los atributos analizados.

Tabla XXV. **Combinaciones de atributos**

Comb.	depto	Atributos								Resultado
		dirección hecho	tipo arma	hora	dsem	dmes	mes	edad	sexo	
A	X	X	X	X	X	X	X	x	x	99%

Fuente: elaboración propia.

La selección de todos los atributos es la selección óptima, ya que con esta combinación se logra un resultado del 99% de clasificación de los mismos.

La matriz de confusión para esta combinación, muestran que, se conserva un muy buen poder clasificatorio de todos los *clusters* en forma homogénea.

Figura 34. **Matriz de confusión**

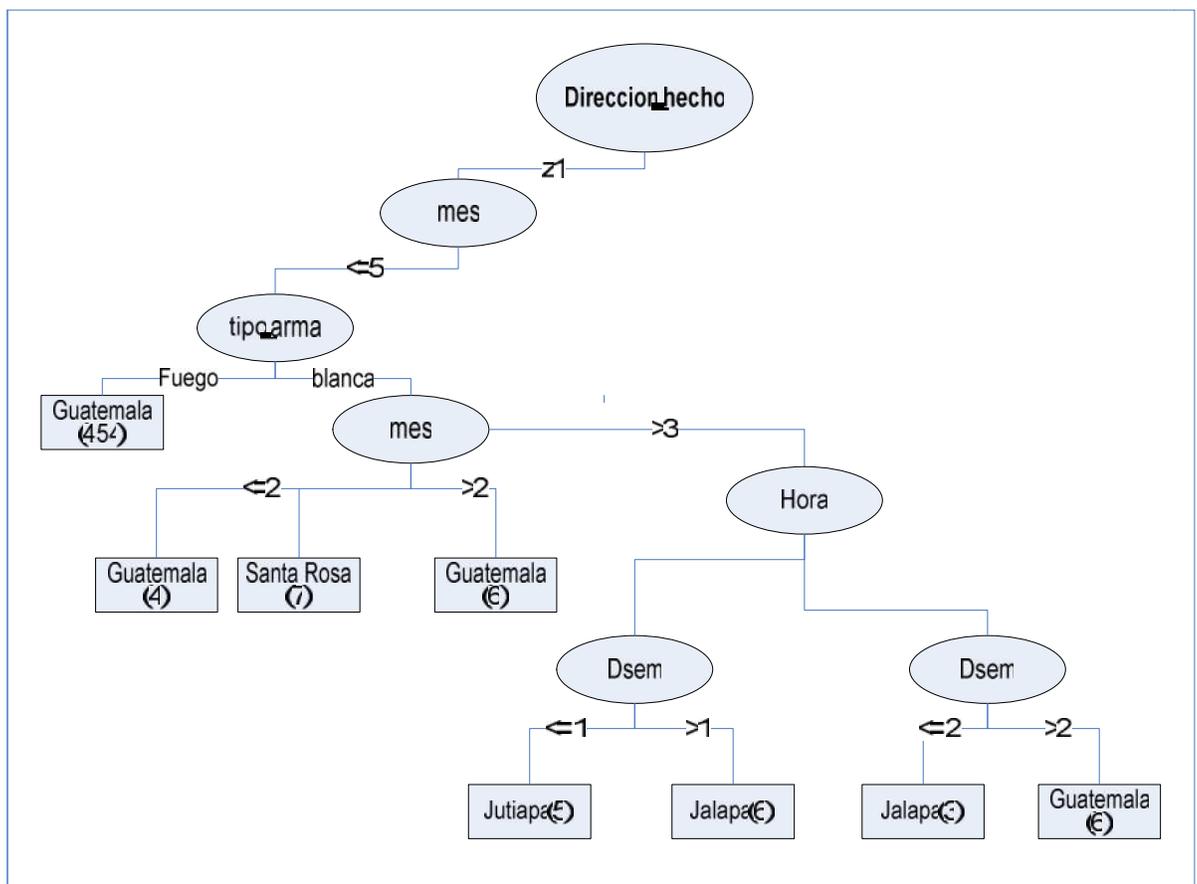
```

=== Confusion Matrix ===
      a  b  <-- classified as
2399  0  |  a = cluster0
  0 455 |  b = cluster1
    
```

Fuente: elaboración propia.

El árbol presenta 257 hojas. El nodo principal es el atributo dirección_hecho. Las ramas correspondientes a dirección_hecho se dirigen a nodos referidos al atributo mes. Por otro lado, las ramas correspondientes a tipo_arma se dirigen a nodos referidos al atributo mes, departamento, sexo y edad. En el caso de arma de fuego, del nodo mes y departamento Guatemala surge la rama entre ≤ 3 y > 3 en cuyo caso corresponde al *cluster* 1 (3 instancias clasificadas correctamente).

Figura 36. Referencias del árbol generado con C4.5



Fuente: elaboración propia.

Con el árbol generado podemos obtener distintas reglas para encontrar ciertos patrones en los datos, a continuación se muestran algunas de las reglas:

Regla 1

SI departamento = Guatemala

Y arma = fuego

Y lugar = z.1

entonces *Cluster 0* (411)

Regla 2

SI departamento = Guatemala

Y arma = blanca

Y lugar = z.1

entonces *Cluster 1* (17)

Regla 3

SI arma = Fuego

Y lugar in {z.1, z.18, z.11}

entonces *Cluster 0* (175)

Regla 4

SI departamento = Guatemala

Y sexo = F

Y hora = 4

entonces *Cluster 0* (22)

Regla 5

SI arma = fuego

Y sexo = F

Y hora = 4

entonces *Cluster 0* (177)

Regla 6

SI arma = Blanca

Y sexo = F

Y edad = 4

entonces *Cluster 1* (13)

Estas reglas permiten confirmar la interpretación hecha anteriormente. Al respecto, puede decirse que los homicidios pueden clasificarse en dos grupos, según el vínculo existente entre la víctima y el agresor:

- Los casos de robo, en los que víctima y agresor no se conocen.
- El resto de los casos, denominados homicidios en conflictos interpersonales.

El primer grupo está representado por el *cluster 0*, mientras que el segundo, por el *cluster 1*.

Los datos más importantes muestran una fuerte relación entre zonas 1, 11, 18, 7, 4 y 8 – arma (1757 casos en total). Si el arma es de fuego (1633 casos), entonces es más probable que sea en estas zonas que en el resto de toda la distribución geográfica (1098). Sin embargo, si el arma es blanca (124 casos) entonces la diferencia sigue predominando en este mismo sector (146) que los ocurridos en los demás (64 casos).

La conclusión a la que se llega es que se trata de dos tipos de conflictos interpersonales distintos, el primero y bien marcado es el de ajuste de cuentas y con cierto nivel de premeditación o pre-intencionalidad (arma de fuego) correspondientes a los casos cometidos en las zonas 1, y 18 mayoritariamente, este tipo de delitos puede marcarse por la amplia presencia de distintos grupos pandilleros que existen en el sector y podríamos decir que también están vinculados a la pelea de territorios.

El segundo que puede ser situacional, ya que los crímenes que ocurren en las distintas zonas y los registrados como lugar del hecho se derivan de asaltos cometidos por motoristas o asesinatos de choferes, ya sea por una extorsión o por un asalto.

El segundo caso es el de los crímenes cometidos con arma blanca, éstos son difíciles de asignar a *priori* a una u otra modalidad. De manera general podría decirse que los crímenes ocurridos en el lugar del hecho y con arma blanca tienen cierto sentido de premeditación, como es el caso de los crímenes ocurridos en las aldeas lejanas dentro de las fincas o cantones vecinales con armas blancas tales como machetes, cuchillos u cualquier otro objeto punzo cortante.

Los crímenes asociados a arma blanca, y a víctimas del sexo *femenino*, indican que la mayor incidencia se da entre las edades de 26-38 años, esta información podría indicar que se trata de crímenes pasionales, y mayormente se dan en el interior de la república en departamentos como Jutiapa y Chiquimula.

CONCLUSIONES

1. Existe gran cantidad de información a nivel nacional que actualmente no está siendo aprovechada y analizada en toda su dimensión, ya que los sistemas existentes son de índole transaccional y no para análisis de datos.
2. Al aprovechar la tecnología existente por medio de un *software* especializado para minería de datos, que contiene las herramientas necesarias para el análisis de datos, se obtiene el beneficio de reducción de costos y maximización de rendimiento en el análisis de los datos que se poseen actualmente.
3. El *software* de minería de datos puede ser utilizado por personas ajenas al ámbito de la informática con una capacitación básica y conocimientos relacionados a los temas de criminalística (personal de las distintas organizaciones del país relacionadas con el sector justicia).
4. Los recursos humanos y tecnológicos necesarios son mínimos y están a disposición de las organizaciones del país relacionadas al sector justicia.

RECOMENDACIONES

1. La utilización de minería de datos para el análisis de información criminal ha demostrado ser exitosa a nivel mundial. Sus distintas aplicaciones han permitido, por ejemplo, relacionar delitos de autoría desconocida según el *modus operandi*, optimizar los recursos policiales y detectar grupos delictivos organizados. Para cumplir con este objetivo, los sistemas transaccionales deben mejorarse para tener mejor calidad de información y así poder ser utilizada de mejor forma.
2. Para este tipo de estudio se podría agregar información muy importante para analizar y buscar patrones de homicidios dolosos asociados a otros delitos, es decir, muchos asesinatos ocurren en ocasión de otros: robos, violaciones, asaltos, extorsiones, etc. teniendo la información del listado de delitos asociado a cada caso, puede obtenerse este tipo de información y poder analizar esta data en un cluster más, aplicando minería de datos y obtener otra variable de estudio.
3. El uso de estas tecnologías también se puede aplicar a información del tránsito, es decir, utilizar la minería de datos para encontrar patrones del comportamiento del mismo, así como de patrones acerca de las zonas de reincidencia de accidentes.
4. La utilización de minería de datos debe ser utilizada como una herramienta poderosa de apoyo en temas de sector justicia.

BIBLIOGRAFÍA

1. CHAU, M.; J. J. Xu, H. Chen. *Extracting meaningful entities from police narrative reports*. [en línea].
URL:<http://ai.eeller.arizona.edu/go/intranet/Publication/COPLINKEE.pdf>. [Consulta: 6 de enero de 2011].
2. CrimeStat, 2007. *CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incident Locations*. [en línea].
URL: <http://www.icpsr.umich.edu/NACJD/crimestat.html>.
[Consulta: 12 de mayo de 2011].
3. Coplink, 2007. *COPLINK Solution Suite*. [en línea].
URL: <http://www.coplink.com>. [Consulta: 12 de mayo de 2011].
4. Expediente 296-94 de fecha 26 de enero de 1995, en Gaceta Jurisprudencial 35, Corte de Constitucionalidad, Guatemala, 1995, p. 14-15.
5. Guatemala. *Código Procesal Penal de la República de Guatemala*. Artículos 5, 112, 113, 114 y 115.
6. _____. Congreso de la República de Guatemala. *Decreto número 51-92. Código Procesal Penal de la República de Guatemala*. Artículo 107, segundo párrafo.

7. _____. Decreto número 11-97. *Ley de la Policía Nacional Civil*. Artículo 9.
8. _____. Informe estadístico de la violencia en Guatemala:
Programa de Seguridad Ciudadana y Prevención de la Violencia del PNUD. Guatemala: PNUD 2007. 80 p.
9. IBM, 2007. Internacional Business Machines. [en línea].
URL:<http://www03.ibm.com/industries/goverment/doc/content/news/pressrelease1019264109.html>. [Consulta: 6 enero 2011].
10. Laboratorio de sistemas inteligentes. Facultad de Ingeniería, Universidad de Buenos Aires. [en línea].
URL: <http://laboratorios.fi.uba.ar/lsi/p-kogan-proyectodetesis.htm>
[Consulta: 6 enero 2011].
11. Mapinfo, 2007. MapInfo Corporation. [en línea].
URL:<http://www.mapinfo.com/location/integration>.
[Consulta: 10 de julio de 2011].
12. PERVERSI, Ignacio. *Aplicación de minería de datos para la exploración y detección de patrones delictivos en Argentina*. Tesis de grado Industrial. Facultad de Ingeniería. Instituto Tecnológico de Buenos Aires, 2007. 105 p.
13. RTI, 2007. Research Triangle Institute. [en línea].
URL:<http://www.rti.org>.
[Consulta: 8 de agosto de 2011]

14. Sentient, 2007. Sentient Information Systems. [en línea].
URL:<http://www.sentient.nl>. [Consulta: 8 de agosto de 2011].

15. SPSS, 2007. SPSS Inc. [en línea].
URL:<http://www.spss.com/success/pdf/CS%20%20Richmond%20PD%20LR.pdf>. [Consulta: 8 de agosto de 2011].