TESIS DE GRADO
EN INGENIERÍA INDUSTRIAL

APLICACIÓN DE MINERÍA DE DATOS PARA LA EXPLORACIÓN Y DETECCIÓN DE PATRONES DELICTIVOS EN ARGENTINA

Autor: Ignacio Perversi

Directores de Tesis:
M. Ing. Enrique Fernández
Dr. Ramón García Martínez

2007

RESÚMEN

A partir de la crisis de finales de 2001, Argentina se vio afectada por una creciente ola de inseguridad caracterizada por un aumento en los índices delictivos y los niveles de violencia. En el plano internacional, los ataques terroristas del 11 de septiembre han aumentado significativamente la preocupación por la seguridad interna en EEUU.

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. En Argentina este tipo de análisis se ha realizado históricamente mediante herramientas estadísticas descriptivas básicas, considerando fundamentalmente variables y relaciones primarias. Sin embargo, muchas veces la estadística descriptiva clásica no refleja la verdadera interrelación de las variables y por lo tanto, el problema real. Esto requiere un tratamiento más complejo.

En este contexto, el objetivo de este trabajo es realizar una implementación de minería de datos en el análisis de información criminal en Argentina y comprobar su efectividad y valor agregado.

Para ello se trabajará en la identificación y detección de patrones de homicidios dolosos cometidos en Argentina durante 2005 en base a información suministrada por la Dirección Nacional de Política Criminal del Ministerio de Justicia y Derechos Humanos de la Nación, organismo encargado de realizar las estadísticas oficiales de criminalidad en Argentina.

Palabras clave: *minería de datos, información criminal, patrones delictivos, clustering, clasificación, homicidios dolosos, inseguridad*.

ABSTRACT

After the economic crisis of 2001, Argentina was shocked by an increasing crime wave, which resulted on high criminal and violence levels. On a worldwide context, the concern about national security has increased significantly in USA since the 9/11 terrorist attacks.

Criminal report analysis is crucial to crime prevention. It enables polices´ design as well as the development of effective preventive measures. In Argentina, this kind of analysis has been made by means of basic descriptive statistic tools, focusing on primary variables and relationships**.** Nevertheless, classic descriptive statistics often fail to reflect the real interrelation between variables -and therefore the real problem-. A more complex treatment is required.

The goal of this paper is to analyze criminal databases in Argentina utilizing data mining in order to corroborate its effectiveness as well as its value added.

The analysis will consist in identifying and detecting murders patterns committed in Argentina during 2005 on the basis of the information provided by the National Direction of Criminal Policy of the National Ministry of Justice and Human Rights. This organism is in charge of the official crime statistics in Argentina.

Keywords: *data mining, criminal information, criminal data mining, criminal patterns, clustering, classification, intentional homicide, murder, security*.

AGRADECIMIENTOS

En el camino recorrido hasta aquí tuve la oportunidad de conocer a muchas personas que contribuyeron en mayor o menor medida a la realización de este trabajo. Quiero agradecerles a todas ellas por el tiempo y dedicación prestados. En orden cronológico:

- Prof. Dr. Ramón García Martínez y M. Ing. Enrique Fernández del Centro de Ingeniería de Software e Ingeniería del Conocimiento (CAPIS) del Instituto Tecnológico de Buenos Aires (ITBA).
- Ing. Juan Carlos Blumberg y Claudia Menin de Gallego Cima de la Fundación Axel Blumberg.
- Susana y Carlos Garnil y Sebastián Luchessa del Equipo de Prevención Ciudadana de la Pastoral de San Isidro.
- Arq. DPU María Adela Igarzábal de Nistal, Directora Ejecutiva del Centro de Información Metropolitana (CIM).
- Federico Valenga, alumno de grado de la Universidad de Morón.
- Dr. Alejandro Slokar, Secretario de Política Criminal y Asuntos Penitenciarios de la Nación.
- Hernán Olaeta y María del Pilar Gándaras Costa de la Secretaría Nacional de Política Criminal.

TABLA DE CONTENIDOS

1. INTRODUCCIÓN

1.1 PRESENTACIÓN DEL PROBLEMA

A partir de la crisis de finales de 2001, Argentina se vio afectada por una creciente ola de inseguridad caracterizada por un aumento en los índices delictivos y los niveles de violencia. Esta situación fue más profunda en los principales centros urbanos y llevó a tomar acciones coordinadas a nivel nacional tendientes a prevenir el delito. Una de estas medidas fue el impulso del Sistema de Alerta Temprana (SAT) por parte del Ministerio de Justicia y Derechos Humanos. En el plano internacional, los ataques terroristas del 11 de septiembre han aumentado significativamente la preocupación por la seguridad interna en EEUU. Las agencias de inteligencia como la CIA o el FBI procesan y analizan información activamente en busca de actividad terrorista [Chen *et al*, 2004].

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. En Argentina este tipo de análisis se ha realizado históricamente mediante herramientas estadísticas descriptivas básicas, considerando fundamentalmente variables y relaciones primarias. Sin embargo, muchas veces la estadística descriptiva clásica no refleja la verdadera interrelación de las variables y por lo tanto, el problema real. Este contexto requiere un tratamiento más complejo que obliga a evolucionar en el análisis de información criminal.

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 [Kantardzic, 2002]. Por esta razón, en muchos casos, los registros almacenados son demasiado grandes o complejos como para analizar [Kantardzic, 2002] y superan el alcance de la estadística [Hand, 1997]. La Minería de Datos (*Data Mining*) es un proceso iterativo de búsqueda de información no trivial en grandes volúmenes de datos [Kantardzic, 2002]. Busca generar información similar a la que podría generar un experto humano: patrones, asociaciones, cambios, anomalías y estructuras significativas [Ochoa, 2004].

En el caso de la inteligencia criminal, la gran cantidad de información y de variables intervienientes justifican el uso de herramientas más potentes que la estadística convencional que permitan determinar relaciones multivariantes subyacentes. La minería de datos aplicada a la inteligencia criminal es un campo bastante nuevo y ha tenido un gran impulso en los últimos años en EEUU [Chen *et al*, 2004].

En este contexto, el objetivo de este trabajo es evaluar una implementación de minería de datos en el análisis de información criminal en Argentina y comprobar su efectividad y valor agregado.

1.2 MOTIVACIONES Y OBJETIVOS SUPERIORES

El fin último de este trabajo es realizar una contribución a la comunidad mediante un aporte para la modernización de las prácticas de información criminal en Argentina. Para ello se procuró mantener un permanente contraste con la "realidad" mediante entrevistas con especialistas y referentes, y al mismo tiempo que la información fuera realista y tuviera el mayor alcance posible. Al respecto se trabajó en conjunto con la Dirección Nacional de Política Criminal del Ministerio de Justicia y Derechos Humanos de la Nación, organismo encargado de realizar las estadísticas oficiales de criminalidad en Argentina.

1.3 ESTRUCTURA DEL TRABAJO

El trabajo esta estructurado de la siguiente forma. En el capítulo 2 se presenta el estado de la cuestión. Comienza con una introducción a la minería de datos, luego se comenta el uso de la informática en el análisis de la información criminal y sus antecedentes, haciendo hincapié en la minería de datos, y finalmente se desarrolla el marco contextual de la información criminal en Argentina. En el capítulo 3 se presenta el problema. En el capítulo 4 se propone una solución al problema basada en minería de datos. En el capítulo 5 se expone la estructura de la información recibida y se desarrolla el pre-procesamiento de la misma hasta obtener el conjunto de datos definitivo para el tratamiento. En el capítulo 6 se presenta un ejemplo didáctico de minería de datos con el objetivo de que el lector se familiarice con algunas herramientas útiles para el análisis del siguiente capítulo. En el capítulo 7 se desarrolla el proceso de minería de datos llevado a cabo y se exponen los resultados experimentales. En el capítulo 8 se presentan las conclusiones. En el capítulo 9 se encuentran las referencias bibliográficas. En el capítulo 10 se presenta una lista de los acrónimos utilizados para una lectura ágil. Finalmente en el capítulo 11 están los anexos.

2. ESTADO DE LA CUESTIÓN

2.1 MINERÍA DE DATOS

Se denomina Minería de Datos [Servente & García-Martínez, 2002; Perichinsky & García-Martínez, 2000; Perichinsky *et al.,* 2000; Perichinsky *et al.,* 2001; Perichinsky *et al.,* 2003] 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 de forma automatizada tendencias y comportamientos; y describir de forma automatizada modelos previamente desconocidos [Piatetski-Shapiro *et al.,* 1991; Chen *et al.,* 1996; Mannila, 1997]. El término Minería de Datos Inteligente [Evangelos & Han, 1996; Michalskí *et al.,* 1998] refiere específicamente a la aplicación de métodos de aprendizaje automático [Michalski *et al.,* 1983; Holsheimer & Siebes, 1991], para descubrir y enumerar patrones presentes en los datos, para estos, se desarrollaron un gran número de métodos de análisis de datos basados en la estadística [Michalski *et al,* 1982]. 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. Una de las diferencias entre al análisis de datos tradicional 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 supone que los patrones e hipótesis son automáticamente extraídos de los datos [Hernández Orallo, 2000].

La Minería de Datos es un proceso completo de descubrimiento de conocimiento que involucra varios pasos [Morales, 2003]:

i Entendimiento del dominio de aplicación, el conocimiento relevante a utilizar y las metas del usuario.

ii Seleccionar un conjunto de datos en donde realizar el proceso de descubrimiento.

iii Limpieza y preprocesamiento de los datos, diseñando una estrategia adecuada para manejar ruido, valores incompletos, valores fuera de rango, valores inconsistentes, etc..

iv Selección de la tarea de descubrimiento a realizar, por ejemplo, clasificación, agrupamiento o clustering, reglas de asociación, etc..

v Selección de los algoritmos a utilizar.

vi Transformación de los datos al formato requerido por el algoritmo especifico de explotación de datos, hallando los atributos útiles, reduciendo las dimensiones de los datos, etc..

vii Llevar a cabo el proceso de minería de datos para encontrar patrones interesantes.

viii Evaluación de los patrones descubiertos y presentación de los mismos mediante técnicas de visualización. Quizás sea necesario eliminar patrones redundantes o no interesantes, o se necesite repetir algún paso anterior con otros datos, con otros algoritmos, con otras metas o con otras estrategias.

ix 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. Un patrón interesante representa conocimiento [Ale, 2005a].

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

- Agrupación o *clustering*.
- Clasificación.
- Asociación.

A continuación se realiza una breve descripción de cada una de estas técnicas y los algoritmos más utilizados.

2.1.1 Agrupación de datos

La agrupación o *clustering* 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 [Chen *et al*., 1996]. De esta manera se busca maximizar la similitud de las instancias en cada cluster y minimizar la similitud entre clusters [Han & Kamber, 2001].

La técnica de *clustering* ha sido estudiada en las áreas de la estadística [Cheeseman & Stutz, 1996; Jain & Dubes, 1988], *machine learning* [Fisher, 1996], base de datos espaciales y minería de datos [Cheeseman & Stutz, 1996; Ester *et al*., 1995; Ng & Han, 1994; Zhang *et al*., 1996].

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 [Kohonen, 1995]. 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 [Vesanto & Alhoniemi, 2000], y además tiene la capacidad de generalizar [Essenreiter *et al.*, 1999], 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 al cluster 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. *K-means* se explica en mayor detalle en el capítulo 4.

Otros algoritmos de *clustering* son *K-medoids* o *PAM* (*Partition around medoids*) y *CLARA* (*Clustering LARge Applications*) [Kaufman & Rousseeuw, 1990]. Este último permite manejar conjuntos de datos más grandes que el primero. *CLARANS* [Ng & Han, 1994] 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 [Chen *et al*., 1996].

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, el presente trabajo hará hincapié en aquellos 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* [Quinlan, 1993a], 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 [Han & Kamber, 2001].

Los algoritmos *ID3* y *C4.5* se detallan en el capítulo 4.

Existen muchos otros algoritmos de clasificación de datos, incluyendo métodos estadísticos [Cheeseman & Stutz, 1996], como el análisis de regresión lineal [Elder IV & Pregibon, 1996]; algoritmos de *machine learning* [Cheeseman & Stutz, 1996]; redes neuronales [Lu *et al*., 1995], algoritmos genéticos y lógica difusa.

2.1.3 Reglas de asociación

La minería de reglas de asociación consiste en encontrar reglas de la forma (A1yA2y...yAm) => (B1yB2y...yBn), donde Ai y Bj son valores de atributos del conjunto de datos [Chen *et al*., 1996]. 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 [Hand *et al*., 2001].

Existen varios algoritmos que realizan el descubrimiento de reglas de asociación, uno de los más utilizados es *Apriori*.

2.2 APLICACIONES INFORMÁTICAS EN EL ANÁLISIS DE INFORMACIÓN CRIMINAL

En los últimos tiempos 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 [Chen *et al.*, 2004]. 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

2.2.1.1 Técnicas geográfico-visuales: el Mapa del Delito

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

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* [Mapinfo, 2007] y *Arcview* [ESRI, 2007].

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 podemos mencionar [Eck, 2005]:

- elipses de desvío estándar (s*tandard deviation ellipses)*: utilizados para delimitar agrupaciones de hechos identificadas mediante técnicas de *clustering* [Figura 2.1 [a]];

- mapas coloreados (*chloropeth maps*): en donde la escala de color representa la cantidad de hechos registrados en una determinada jurisdicción geográfica [Figura 2.1 [b]];

- 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 [Figura 2.1 [c]];

- mapas de contorno suavizado (*kernel density estimation*): similar al anterior pero con un efecto continuo logrado mediante el uso de algoritmos [Figura 2.1 [d]]. Este último es el más apropiado de todos los métodos para la visualización del delito [Eck, 2005].

Existen varios paquetes de análisis estadístico espacial para información criminal que trabajan sobre GISs. Uno de los primeros fue *STAC (Spatial and Temporal Análisis of Crime*), desarrollado por la Autoridad para la Información de Justicia Criminal de Illinois [ICJIA, 2007]. Luego le siguieron *CompStat* y *CrimeStat* [CrimeStat, 2007]. Este último contiene un amplio set de algoritmos y si bien sus resultados pueden ser visualizados en GIS, esta orientado hacia un usuario con determinados conocimientos técnicos. Por esta razón su uso queda limitado al analista criminal más que al policía tradicional [Oatley *et al.*, 2004].

En Reino Unido la herramienta más utilizada como complemento de GISs es el *i2 Analyst´s Workstation* [i2, 2007]. Si bien su capacidad es limitada, posee un módulo muy útil (llamado *PatternTracer)* que permite detectar patrones en los registros de llamadas telefónicas [Oatley *et al.*, 2004].

Es importante aclarar que si bien todo 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 [Chen *et al.*, 2004]. Básicamente todas las técnicas de minería de datos descriptas 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.

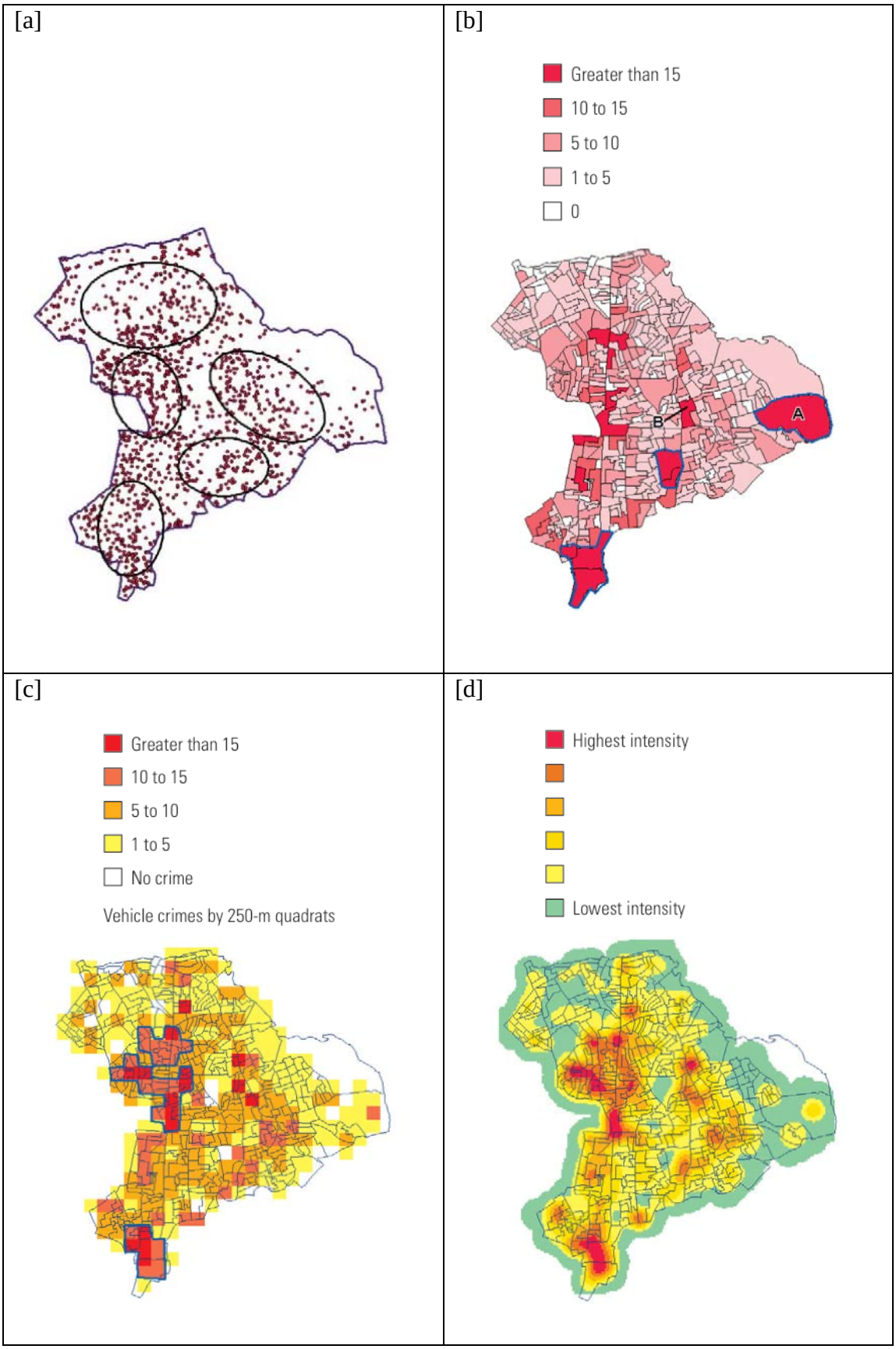


Figura 2.1: Diferentes tipos de mapas del delito delimitando *hotspots* para el robo de vehículos [Eck, 2005]

Existen ciertos desarrollos que integran tanto información *hard* (ADN, huellas digitales, huellas de calzado) como *soft* (relativa al *modus operandi*), como por ejemplo el software FLINTS (Forensic Led Intelligence System). El mismo en su primera versión permitía construir asociaciones gráficas entre crímenes y criminales, muchas de las cuales eran desconocidas hasta el momento. Luego en una segunda versión, se automatizó esta función, trabajando *online* con distintas bases de datos y en tiempo real. La ultima versión incorpora la capacidad de detectar redes criminales y una visualización geográfica, en forma de mapas y animaciones, que permite identificar *hotspots* y realizar comparaciones temporales [Oatley *et al.*, 2004].

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 ellos incorporan a su vez herramientas de visualización geográfica.

2.2.2.1 Proyecto COPLINK

El Proyecto COPLINK fue creado en el año 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, 2007].

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 interfase sencilla, integrando distintas fuentes de información. El segundo esta diseñado para detectar de forma automática distintos tipos de asociaciones entre las bases de datos mediante técnicas de minería de datos. Ambos sistemas presentan una interfase visual amigable [Coplink, 2004].

Algunas de las aplicaciones de minería de datos desarrolladas por Coplink son las siguientes:

Análisis de Redes Criminales [Chen *et al.*, 2004]: 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 en base a 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 baricentro de cada subgrupo para determinar su miembro clave o líder.

El resultado en base a registros del Departamento de Policía de Tucson sobre hechos cometidos entre 1985 y 2002 reveló 16 miembros clave [Figuras 2.2 y 2.3]. En la Figura 2.3 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.

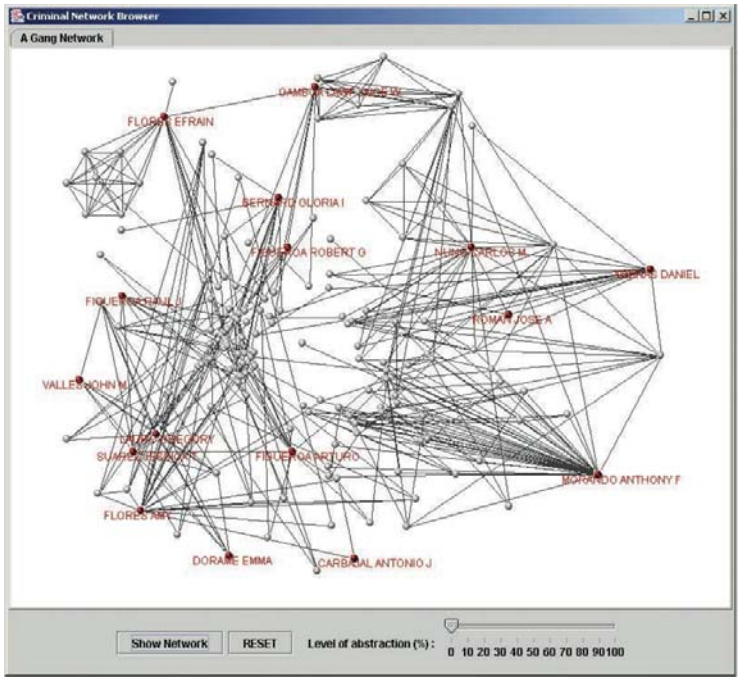


Figura 2.2: Análisis de Redes Criminales: vínculos entre sospechosos [Coplink, 2004]

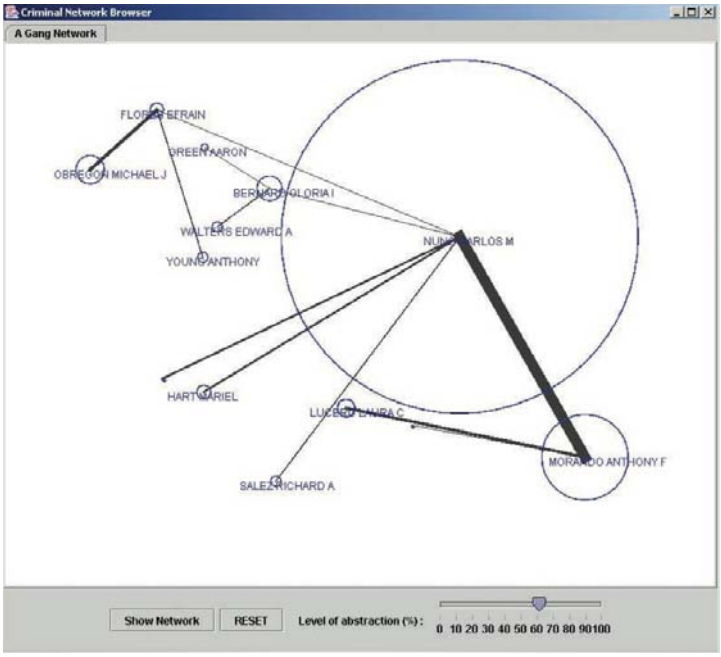


Figura 2.3: Análisis de Redes Criminales: subgrupos identificados por Coplink [Coplink, 2004]

Extracción automática de entidades [Chau *et al.*, 2002]: consiste en extraer automáticamente determinada información criminal de los reportes policiales (en este caso, nombres propios, direcciones, características personales, información de vehículos y nombres de drogas). Para esto se utiliza un algoritmo que funciona de la siguiente manera:

i. identifica las oraciones que poseen sustantivos de acuerdo a un análisis sintáctico basado en determinadas reglas lingüísticas;

ii. compara cada palabra de la oración con una base de entidades (por ejemplo apellidos, nombres de calles, etc.);

iii. calcula un puntaje para cada oración en función a la cantidad de coincidencias encontradas;

iv. 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 [Wang *et al.*, 2004]: 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 año 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 esta enfocado en

los casos de robo a domicilio particulares. Sus principales objetivos son [Zeleznikow, 2005]:

- identificar los recursos críticos para establecer estrategias de prevención y detección más eficientes;
- proveer de fundamentos empíricos para el desarrollo de planes inter-departamentales 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 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 como por ejemplo la visualización geo-referenciada 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* [Sentient, 2007] 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 casos 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 de la performance 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 iterfase desarrollada por *RTI Internacional* [RTI, 2007]. El principal objetivo es optimizar la alocación de recursos, en base a una modalidad preactiva y no reactiva. Por ejemplo durante año nuevo se identificaron las zonas que habían tenido un aumento en los casos de heridos de 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) [SPSS, 2007].

- La Policía Estatal de Illinois adquirió en 2005 un software de minería de datos del compañía RiverGlass Inc. [RiverGalss, 2007] 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* [NYC, 2007]. 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.

2.2.3 Antecedentes en Argentina

En Argentina no se conoce ninguna experiencia de aplicación de minería de datos a información criminal. Sin embargo hay dos proyectos relacionados que merecen ser mencionados.

2.2.3.1 El Proyecto SURC

A comienzos de 2004 el entonces Ministerio de Justicia, Seguridad y Derechos Humanos lanzó el proyecto del Sistema Unificado de Registros Criminales (SURC). El objetivo era interconectar y articular las instituciones del Sistema Policial y el Sistema Judicial mediante una red en la cual todos tuvieran acceso a un banco de datos común, alimentado en tiempo real y del cual se pudieran realizar consultas *online*.

Este banco de datos contemplaba información diversa [SSI-MI, 2004]:

- Registro de hechos: características generales del hecho denunciado (lugar, día, hora, delito denunciado y comisaría interviniente).
- Registro de denunciantes: identidad y características de la víctima o denunciante.
- Registro de autores identificados: identidad, características, historial criminal e imágenes de los autores.
- Registro de autores no identificados: descripción de los NN (contextura física, edad aproximada, estatura, color de pelo, señas particulares, frases frecuentes, etc.).
- Registro de elementos robados: información útil para la identificación de los objetos robados.
- Registro de autos robados: marca, modelo, color, número de patente, número de motor, características particulares, etc.
- Registro de armas secuestradas: características de las armas secuestradas, vinculando esta base con otros sistemas como el Ibis.
- Registro de evidencias: descripción de huellas y pistas relevadas en la escena del crimen.
- Mapa del delito: presentación de los hechos en forma gráfica y geo-referenciada mediante GIS.

Las principales mejoras que se conseguían eran:

- Sistematización del proceso iniciado con la denuncia policial: la fiscalía podía acceder mediante Internet a la denuncia registrada por la policía en tiempo real,

eliminando el soporte papel y la correspondencia como forma de intercambio de información.

- Seguimiento de los expedientes de las fiscalías y su instancia administrativa.
- Integración con el Sistema de Intercambio de Información de Seguridad del Mercosur (SISME).
- Estandarización de la información delictiva a nivel nacional y alimentación automática de los sistemas de estadísticas como el Sistema Nacional de Información Criminal (SNIC).
- Confección del Mapa del Delito a nivel nacional y análisis espacial de la información mediante GIS, optimizando la alocación de recursos y la eficiencia de los planes de prevención.
- Asociación de hechos con autor NN según las características del autor y el *modus operandi*, colaborando con las tareas de investigación criminal.

Este proyecto de gran alcance contemplaba una implementación progresiva, comenzando por la Ciudad Autónoma de Buenos Aires y avanzando hacia las provincias. Sin embargo tras la salida del entonces Ministro de Justicia, Seguridad y Derechos Humanos, Gustavo Béliz, y gran parte de su equipo de trabajo, en julio de 2004, el proyecto quedó congelado. Tiempo más tarde se trasladaron las funciones de seguridad de la esfera del Ministerio de Justicia al Ministerio del Interior, y consigo el proyecto SURC. Actualmente el proyecto permanece vigente, con radicación en la Secretaría de Inteligencia Criminal del Ministerio del Interior de la Nación, pero relegado y con un cambio de enfoque respecto al original.

2.2.3.2 El Mapa del Delito de la Ciudad Autónoma de Buenos Aires

El Ministerio Público Fiscal de la Nación (MPFN) es una de las pocas instituciones judiciales de Argentina que posee un sistema de información digitalizada. Cuenta con una base de datos de los hechos delictivos de autoría desconocida (NN) registrados en Capital Federal. Esta base contiene información referida al hecho, como ser: tipo de delito, fecha, lugar y cantidad de víctimas.

Asimismo el Centro de Información Metropolitana (CIM), radicado en la Facultad de Arquitectura, Diseño y Urbanismo (FADU) de la Universidad de Buenos Aires (UBA), posee el Sistema de Información Territorial del Área Metropolitana de Buenos Aires (SAT/AMBA). Este sistema consiste en la base cartográfica digital de todo el AMBA para ser utilizada bajo GISs (*Geographical Information Systems*). No sólo posee los elementos tradicionales (calles, avenidas, vías del ferrocarril, plazas, etc.) sino también la visualización de las demarcaciones zonales (barrios, centros de gestión y participación, comisarías, etc.) y gran parte del equipamiento urbano (escuelas, clubes, bancos, etc.).

En 2002 ambas instituciones firmaron un "Convenio de Asistencia, Complementación y Cooperación" con el objetivo de que el CIM elaborase el Mapa del Delito de la Ciudad Autónoma de Buenos Aires con la información suministrada por el MPFN.

Si bien la existencia de este mapa no es muy conocida y pese a que cuenta con ciertas limitaciones (únicamente hechos ocurridos en Capital Federal con autoría desconocida), su aporte en el análisis de la situación criminal es muy valioso. Por ejemplo en la Figura 2.4 se puede ver la variación en la distribución geográfica de los robos de automotor de 2002 a 2003, posiblemente atribuible a mayor control sobre los accesos a la ciudad desde la Av. General Paz [Behar & Lucilli, 2003].

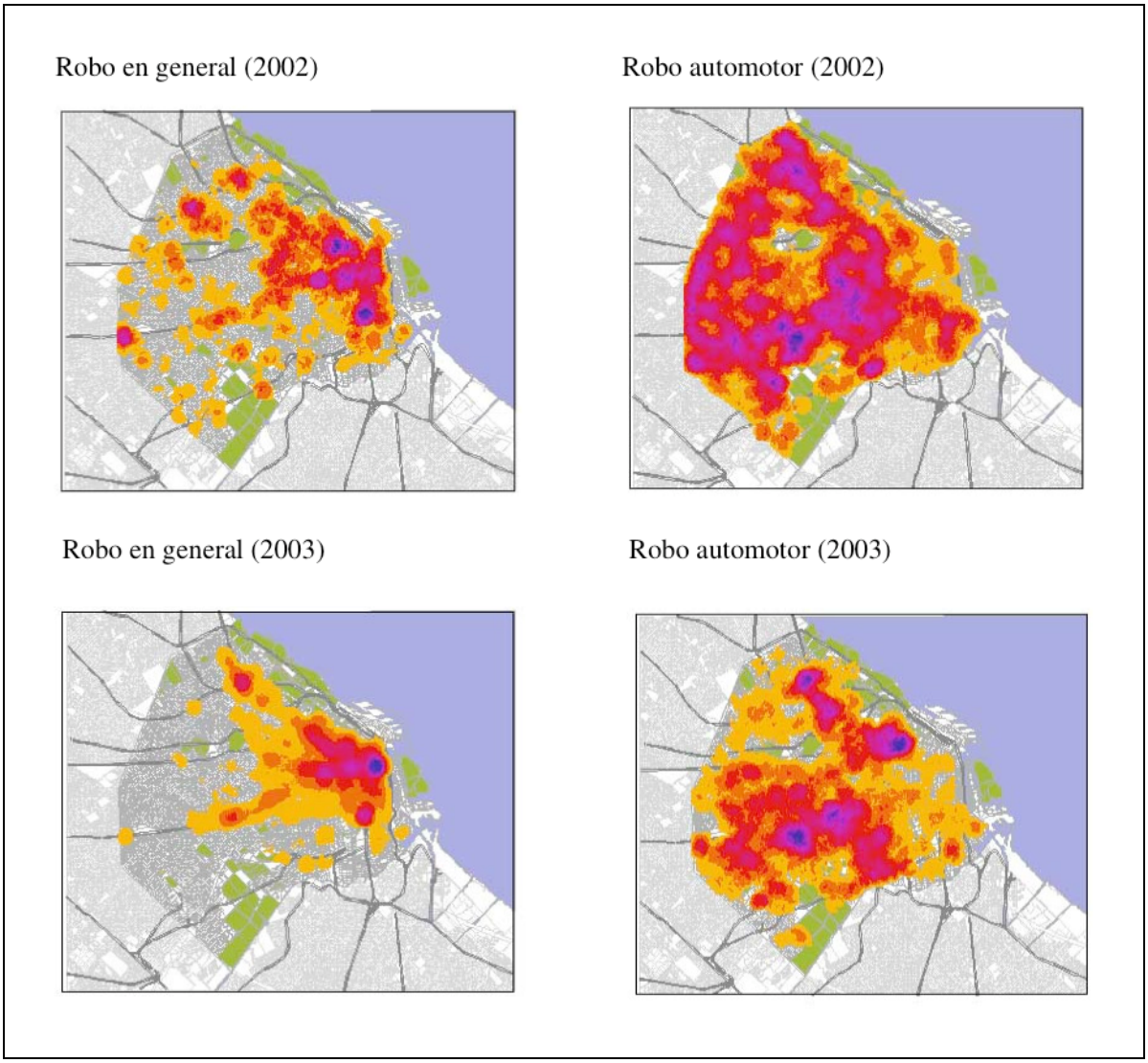


Figura 2.4: Ejemplo de visualización del Mapa del Delito de la Ciudad Autónoma de Buenos Aires [Behar & Lucilli, 2003]

2.3 LA INFORMACIÓN CRIMINAL EN ARGENTINA

Antes de presentar el problema e iniciar el análisis es necesario presentar, a modo de marco de referencia, una breve descripción de la información criminal en Argentina. La presente sección esta basada en las reuniones mantenidas con la gente de la Dirección Nacional de Política Criminal del Ministerio de Justicia y Derechos Humanos de la Nación.

2.3.1 Las fuentes de información criminal

Entendemos por información criminal a toda aquella información resultante a partir de un presunto delito o de sus componentes (victima, victimario, propiedades, vehículos, etc.) que sea relevante para la toma de decisiones a posteriori. Ya sea en la prevención, detección y esclarecimiento del delito como en la prosecución 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 delictuoso desde que es registrado por el Estado.

Podemos subdividir al Sistema Penal según las distintas instancias en Sistema Policial, Sistema Judicial y Sistema Penitenciario. Como muestra la Figura 2.5, una gran cantidad de hechos ingresa por el Sistema Policial, atraviesa el cuello de botella del Sistema Judicial y egresa a través del Sistema Penitenciario.

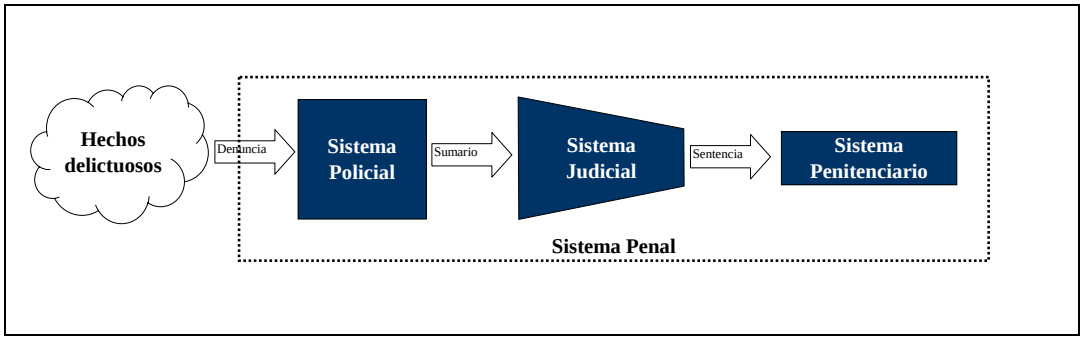


Figura 2.5: Esquema del Sistema Penal

En Argentina debido al sistema de gobierno federal adoptado, este esquema se replica en cada una de las provincias del territorio nacional.

Este sistema de gobierno tiene fuertes implicancias en la consolidación de la información a nivel nacional. No sólo por la falta de homogeneidad entre las distintas provincias, sino fundamentalmente porque cada provincia tiene autonomía sobre la información generada bajo su jurisdicción y el Estado Nacional no tiene injerencia sobre la misma.

Las funciones de consolidación de información criminal a nivel nacional y confección de la estadística general fueron delegadas, a través de la Ley Nacional 25.266, a la

Dirección Nacional de Política Criminal (DNPC) del Ministerio de Justicia y Derechos Humanos de la Nación (MJDHN) [Anexo 1].

Esta ley establece a la DNPC como **única fuente oficial de información criminal a nivel nacional**. Cabe mencionar que esta capacidad de la DNCP no va en desmedro de que las distintas instituciones del Sistema Penal posean su propia información, e incluso sectores de estadística, para analizar su propia gestión. De hecho, la gran mayoría de las policías nacionales poseen una división de estadística.

Si bien el Sistema Penal en su conjunto constituye la fuente natural de información criminal, existen otras fuentes diseminadas que aportan información parcial y que sirven para complementar o contrastar la información del Sistema Penal. Por ejemplo, las estadísticas del Ministerio de Salud reflejan la cantidad de muertes violentas que se registran según los certificados de defunción y que los médicos caratulan según su posible causa (accidente, suicidio, homicidio y no determinada) [DNPC, 2002].

A fines del presente trabajo nos concentraremos en general en la información criminal oficial relevada a nivel nacional por la DNPC, y particularmente en la que se encuentra a la entrada del Sistema Penal.

2.3.2 Información criminal del Sistema Policial

Surge de la registración de hechos delictuosos por parte de las instituciones policiales y las fuerzas de seguridad: Policía Federal Argentina, Prefectura Naval Argentina, Gendarmería Nacional y 24 policías provinciales.

2.3.2.1 Breve historia

Las primeras estadísticas policiales en Argentina comenzaron a ser elaboradas por la Policía de la Capital Federal en 1887 [Blackwelder & Jonson, 1984; Rubial, 1993; Sozzo, 2000]. A partir de 1971 toda la información proveniente del Sistema Policial pasó a ser consolidada a nivel nacional por el Registro Nacional de Reincidencia y Estadísticas Criminales (RNREC). Los siguientes 30 años se caracterizaron por estadísticas incompletas, de poca calidad y sin un análisis posterior.

A partir de 1999 se intentó revertir esta situación mediante la creación del Sistema Nacional de Información Criminal (SNIC) y el Sistema de Alerta Temprana (SAT), y la transferencia de las funciones de consolidación y análisis de información criminal a la Dirección Nacional de Política Criminal (DNPC). En julio de 2000 se formalizó esta transferencia mediante la Ley Nacional 25.266 anteriormente mencionada.

2.3.2.2 El Sistema Nacional de Información Criminal (SNIC) y el Sistema de Alerta Temprana (SAT)

El SNIC se nutre de una planilla heredada del RNREC que consiste en una sumarización de los presuntos delitos ocurridos en una determinada jurisdicción y en un

determinado mes, según la siguiente tipificación basada en el código penal [Tabla 2.1 y Anexo 2.1]:

Sistema Nacional de Información Criminal (SNIC)	
Sumarización general	**Sumarización de casos especiales**
Delitos contra las personas	Homicidios dolosos
	Homicidios dolosos en grado de tentativa
	Homicidios culposos en accidentes de tránsito
	Homicidios culposos por otros hechos
	Lesiones dolosas
	Lesiones culposas en accidentes de tránsito
	Lesiones culposas por otros hechos
Delitos contra el honor	
Delitos contra la integridad sexual	Violaciones
Delitos contra el estado civil	
Delitos contra la libertad	Amenazas
Delitos contra la propiedad	Robos
	Tentativas de robo
	Hurtos
	Robos agravados por lesiones y/o muertes
	Tentativas de robo agravadas por lesiones y/o muertes
	Hurtos
	Tentativas de hurto
Delitos contra la seguridad pública Delitos contra el orden público Delitos contra la seguridad de la nación Delitos contra los poderes públicos y el orden constitucional Delitos contra la administración pública Delitos contra la fe pública	
Delitos previstos en leyes especiales	Delitos de estupefacientes
Figuras contravencionales	
Suicidios	

Tabla 2.1: Hechos delictuosos sumarizados a través de las planillas del SNIC

Desde que el sistema entró en funcionamiento, en enero de 1999, se ha hecho hincapié en obtener una mayor desagregación, tanto geográfica como temporal.

El Sistema de Alerta Temprana (SAT) se nutre de cuatro planillas complementarias a la del SNIC que relevan información detallada sobre cuatro aspectos en particular:

- homicidios dolosos [Anexo 2.2];
- homicidios culposos en accidentes de tránsito [Anexo 2.3];

- suicidios [Anexo 2.4];
- delitos contra la propiedad [Anexo 2.5].

En los primeros tres casos se releva información puntual de cada hecho, mientras que el último consta de sumarizaciones parciales según distintas variables. La información relevada en cada caso es la siguiente [Tabla 2.2]:

Sistema de Alerta Temprana (SAT)	
Planilla	**Información Relevada**
Homicidios dolosos	Fecha y hora
	Calle, altura y localidad
	Tipo de lugar
	Clase de arma
	Otro delito
	Datos de la víctima
	Datos del imputado
Homicidios culposos en accidentes de tránsito	Fecha y hora
	Calle, altura y localidad
	Intersección de calles
	Semáforo
	Modo de colisión
	Condiciones climáticas
	Datos de la víctima
	Datos del imputado
Suicidios	Fecha y hora
	Localidad
	Tipo de lugar
	Modalidad
	Datos del suicida
Delitos contra la propiedad	Cantidad de hechos según Tipo de lugar
	Cantidad de hechos según Horario
	Cantidad de hechos según Arma
	Cantidad de inculpados según Sexo
	Cantidad de inculpados según Edad

Tabla 2.2: Información relevada por SAT

Actualmente estas cinco planillas (1 de SNIC y 4 de SAT) son completadas mensualmente por cada dependencia del Sistema Policial (en la mayoría de los casos comisarías).

2.3.2.3 Ciclo de registración

El proceso de registración comienza con una denuncia. Esta denuncia puede ser hecha por un particular o por actuación de oficio de la fuerza de seguridad (este caso se da casi exclusivamente en delitos de estupefacientes, robos y homicidios). La denuncia es

enviada por la fuerza de seguridad a la fiscalía que corresponda, ingresando en el sistema judicial en forma de sumario.

Una vez por mes cada comisaría completa la planilla de SNIC realizando un recuento de la cantidad de denuncias que tuvieron lugar durante ese mes según cada tipo de delito, junto con las planillas complementarias de SAT. Estas planillas son remitidas al Ministerio de Justicia y Derechos Humanos de la Nación (MJDHN) ya sea en formato papel, en cuyo caso son ingresadas al sistema en la DNPC por 2 *data entries* de dedicación exclusiva, o en formato digital (planilla de Excel), en cuyo caso son incorporadas al sistema automáticamente. Este proceso se repite mensualmente. A fin de cada año calendario la DNPC comienza a elaborar el informe anual que es finalizado y elevado para su publicación aproximadamente a mediados de abril.

2.3.2.4 Algunas observaciones

Es importante destacar ciertos aspectos de este sistema de registración que derivan básicamente de que el mismo se encuentra en la entrada al Sistema Penal.

En primer lugar en esta instancia hablaremos de hecho delictuoso o presunto delito ya que la definición de si realmente fue o no un delito corresponde a instancias superiores del proceso penal.

Con este mismo criterio debemos ser cuidadosos con las caratulaciones que se realizan en el Sistema Policial, ya que son de carácter provisorio y pueden ser modificadas posteriormente. Esta modificación se puede deber a la evolución temporal de los hechos (por ejemplo casos caratulados por el Sistema Policial como “Lesiones graves” que, tras la muerte de la víctima, devienen en “Homicidio”) o como resultado de la investigación judicial (por ejemplo casos caratulados por el Sistema Policial como “Muerte dudosa” que devienen en “Homicidio”).

Otro punto importante es que se debe diferenciar el hecho o denuncia del presunto delito, ya que la relación no es biunívoca, y una misma denuncia podría contener más de un delito.

2.3.2.5 Auditoria y validación de la información

Esporádicamente la DNPC realiza viajes de capacitación y auditoria a las respectivas dependencias provinciales. El objetivo de estos viajes es capacitar al personal policial en el llenado de las planillas y corroborar que la información histórica interna de cada policía provincial coincida con la que fue enviada oportunamente a la DNPC. No existe ningún tipo de validación oficial externa de los datos suministrados por cada dependencia.

Este control debería ser realizado con la información existente en el Sistema Judicial, teniendo en cuenta las salvedades mencionadas en la sección 2.3.2.4. Sin embargo la ausencia de un sistema nacional de información judicial, consecuencia en parte del

sistema federal de gobierno y la falta de digitalización de los expedientes, hacen que esta tarea sea imposible en la práctica.

En 2002 la DNPC hizo un estudio sobre homicidios dolosos ocurridos en Capital Federal sobre la base de unos 200 expedientes del Sistema Judicial y comparó tanto la información como la tendencia respecto a la misma información provista por el Sistema Policial.

En algunas regiones del país existe cierto control informal entre las policías de provincias lindantes y los ciudadanos que garantiza que al menos los delitos más graves no pasen desapercibidos por el SNIC. Por ejemplo en 2002 hubo un sólo homicidio en Tierra del Fuego que fue omitido en la planilla SNIC por la policía y por lo tanto inadvertido en el informe anual elaborado por la DNPC. Sin embargo al ver este informe los allegados a la víctima detectaron esta omisión y reclamaron la rectificación.

2.3.2.6 Limitación de la información: la "cifra negra del delito"

Todas las fuentes de información oficial tienen la limitación de que consideran únicamente los hechos delictuosos que ingresaron efectivamente al Sistema Penal y no la totalidad de los hechos. Esto significa que son una visión sesgada de la realidad. La fracción de hechos que no ingresa al Sistema Penal es lo que se denomina comúnmente la "cifra negra del delito". El origen de esta situación suele estar en la omisión de la denuncia policial debido a diversas razones [Sozzo, 2000]:

- Creencia en que determinados delitos menores no justifican el trámite administrativo;
- Creencia en que la policía o la justicia son ineficientes y no van a solucionar el problema;
- Desaliento por parte de la policía a realizar la denuncia de determinados delitos para alivianar sus propias estadísticas;
- Creencia en que las fuerzas de seguridad locales pueden estar involucradas en el hecho;
- Cierto grado de involucramiento de la víctima en el hecho;
- Temor por parte de la víctima a eventuales represalias o a atravesar situaciones de humillación o dolor.

El nivel de la "cifra negra" depende del tipo del delito. Por ejemplo en el caso de homicidios, robos a propiedades privadas (comercios, entidades bancarias, domicilios particulares) y robos de automotores, la "cifra negra" es muy baja. En cambio en los casos de robos o hurtos a personas en la vía pública y delitos sexuales la cifra es muy alta [Sozzo, 2000].

Para estimar la "cifra negra del delito" surge la encuesta de victimización. Esta metodología comenzó a ser utilizada por el Ministerio de Justicia de Estados Unidos en la década del 70`. En Argentina las primeras encuestas de este tipo las realizó la DNPC en 1996 sobre los delitos ocurridos en Capital Federal en 1995.

Consiste en una encuesta realizada en domicilios particulares sobre una muestra de la población en donde se pregunta al entrevistado si él o alguno de los miembros de su familia conviviente fue víctima de algún delito en el último año y, de resultar afirmativo, se indaga sobre las características del mismo. También se hace hincapié en si se hizo la denuncia correspondiente y en caso de omisión, las razones de la misma. La encuesta concluye con preguntas acerca de la confianza en las distintas fuerzas de seguridad, la sensación de inseguridad y las medidas de autoprotección adoptadas.

Es importante destacar que las encuestas de victimización no cubren todo el espectro de delitos, sino que ponen foco sobre los delitos comunes en los que existe una víctima, como ser robos, hurtos o violaciones sexuales (y no sobre otro tipo de delitos como estafas o posesión de drogas).

2.3.3 Información criminal del Sistema Judicial y Penitenciario

La Información del Sistema Judicial surge de la registración de los presuntos delitos por parte de las instituciones judiciales penales en forma de sumarios. En la Argentina existen una administración federal con jurisdicción sobre todo el territorio nacional con competencia sobre determinados tipos de delitos, una administración nacional con jurisdicción en Capital Federal y 24 administraciones provinciales [Sozzo, 2000].

A diferencia del Sistema Policial, aquí todavía no se cuenta con un sistema centralizado de información a nivel nacional. Hay un proyecto para crear un Sistema Nacional de Estadísticas Judiciales (SNEJ) de similares características al SNIC. Por el momento existen diferentes fuentes de información desarticuladas; por un lado el poder judicial federal y por el otro cada uno de los poderes judiciales provinciales. Esto genera falta de criterios comunes y por lo tanto de homogeneidad a la hora de hacer comparaciones o consolidar la información.

En cuanto a la Información del Sistema Penitenciario, existe el Sistema Nacional de Estadísticas sobre Ejecución de la Pena (SNEEP). El mismo tiene por objetivo recolectar periódicamente información estadística sobre la población privada de la libertad en todo el país [DNPC, 2007]. Esta recolección se hace anualmente mediante cuestionarios que constan de dos partes. En la primera se recopila información acerca de lo acontecido durante el año, tomando como unidad de análisis el establecimiento. La segunda, consiste en un censo de la población detenida al 31 de diciembre de cada año [Anexo 2.6].

De esta forma la DNPC pretende relevar información criminal en todas las instancias del Sistema Penal [Figura 2.6].

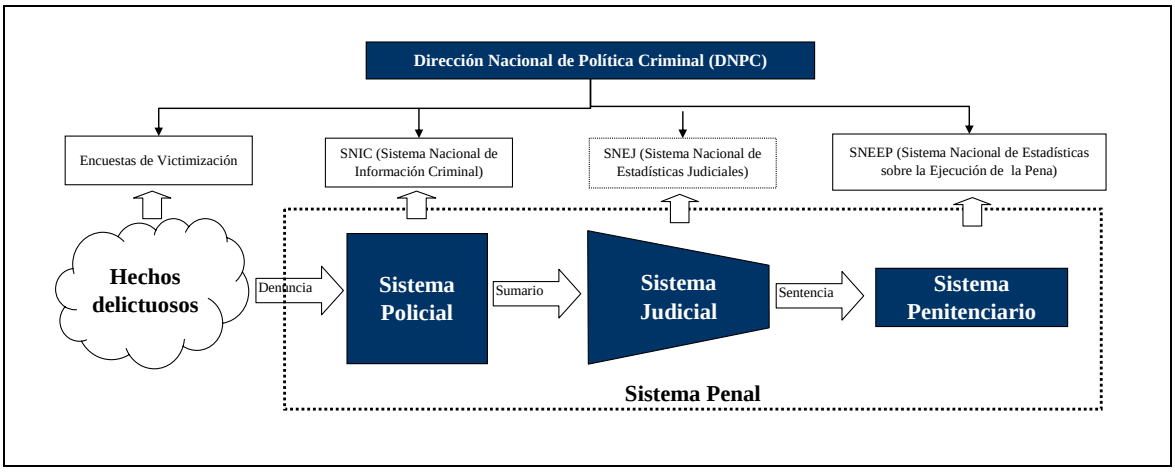


Figura 2.6: Esquema de relevamiento de la información criminal por parte de la Dirección Nacional de Política Criminal

3. DEFINICIÓN DEL PROBLEMA

3.1. EL PROBLEMA DEL TRATAMIENTO DE LA INFORMACIÓN

Actualmente la Dirección Nacional de Política Criminal (DNPC) del Ministerio de Justicia y Derechos Humanos de la Nación (MJDHN) analiza la información proveniente de sus sistemas (SNIC, SAT y SNEEP) mediante un análisis estadístico básico, sin hacer un aprovechamiento exhaustivo de la información. Los resultados de este análisis son utilizados con distintos objetivos:

- emitir un informe anual a nivel nacional que refleje la situación de los hechos delictuosos, en cumplimiento con la ley nacional 25.266;
- servir de fundamento para la creación de planes de prevención y el diseño de políticas criminales;
- nutrir de información a otros actores públicos y privados, como ser a legisladores para el impulso de nuevas leyes.

El gran alcance de estos objetivos, sumado a la gran cantidad de información y de variables intervinientes, justifican el uso de herramientas más potentes que la estadística descriptiva convencional.

En resumen podemos decir que:

"la DNPC no cuenta en la actualidad con un sistema que le permita realizar un análisis exhaustivo de la información recolectada, determinando relaciones multivariantes subyacentes y extrayendo conclusiones de mayor valor agregado para la toma de decisiones".

3.2. EL PROBLEMA ESPECÍFICO

A partir de las reuniones mantenidas, la gente de la DNPC nos ha señalado su especial interés en el análisis de determinados tipos de delitos que revisten prioridad en función de su gravedad y frecuencia. Estos son los "homicidios dolosos" y los "homicidios culposos en accidentes de tránsito". Ambos son relevados por SAT, por lo que se cuenta con cierta información puntual sobre cada hecho. Por recomendación de la DNPC, el presente trabajo aborda el primer tipo de delito.

Los "homicidios dolosos", es decir intencionales, son el resultado de una de las mayores problemáticas de América Latina: la violencia. A fines del siglo XX esta era la primera causa de muerte en América Latina de las personas de 15 a 44 años [Spinelli *et al.*, 2006; OPS/OMS, 2003].

Dentro de este tipo de homicidios presentan principal relevancia aquellos cometidos con armas de fuego. La OMS señala que, a fines del siglo XX, el 63% de los homicidios

mundiales eran ocasionados por armas de fuego. En América Latina esta cifra sobrepasaba el 80% [Spinelli *et al.*, 2006; Briceño-León, 2001]. 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 [Spinelli *et al.*, 2006; EAP, 2004].

El problema de las armas de fuego radica principalmente en su alto grado de efectividad y letalidad. En algunos trabajos 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 autodefensa aumentan las probabilidades de ser una víctima de homicidio [Spinelli *et al.*, 2006; Szwarewald & Castillo, 1998; Briceño-León, 2001; Cardona *et al*., 2005; Peres & Santos, 2005].

La Argentina no es la excepción en cuanto a la problemática de la violencia y su relación con las armas de fuego [Spinelli *et al.*, 2006]. Para combatir este problema, en agosto de 2006 se lanzó un Plan Nacional de Desarme Civil [La Nación, 2006].

Desde este punto de vista el problema particular es:

“la DNPC esta interesada en encontrar patrones de homicidios dolosos, vinculados con el tipo de arma empleada, que permitan generar nuevo conocimiento sobre la problemática y/o validar los conocimientos adquiridos hasta el momento”.

4. SOLUCIÓN PROPUESTA

4.1. SOLUCIÓN PROPUESTA AL PROBLEMA DEL TRATAMIENTO DE LA INFORMACIÓN

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

El software propuesto es *Weka 3.5.5* [Weka, 2007] por contar con las siguientes ventajas:

- es de distribución libre y gratuita;
- cuenta con una interfase gráfica amigable y es fácil de usar;
- tiene incorporado un amplio set de algoritmos de minería de datos;
- esta programado en código abierto, permitiendo al usuario programador agregar nuevas funciones según su necesidad.

4.2. SOLUCIÓN PROPUESTA AL PROBLEMA ESPECÍFICO

Se propone llevar adelante un proceso de minería de datos sobre la base de datos de homicidios dolosos de SAT 2005 mediante *Weka 3.5.5* para identificar patrones de homicidios dolosos. El proceso propuesto es el siguiente:

i. confeccionar un *data set* a partir de la base de datos de homicidios dolosos SAT 2005;

ii. aplicar el algoritmo *K-means* para agrupar los hechos según su similitud en grupos o *clusters* distintos;

iii. interpretar y convalidar los resultados obtenidos con los usuarios, haciendo uso de los informes emitidos por *Weka*;

iv. utilizar el algoritmo de inducción C4.5 para identificar reglas de pertenencia a cada uno de los grupos o clusters;

v. brindar una interpretación definitiva y extraer conclusiones.

4.3. ALGORITMOS A UTILIZAR

4.3.1 Algoritmo *K-means*

K-means en 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 [Figura 4.1]:

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

Figura 4.1: Error cuadrático

Donde existen k grupos S_i, i=1,2,...,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 se debe establecer 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, se suelen realizar 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 [Ale, 2005b]:

i. Particionar los objetos en k subconjuntos no vacíos.
ii. Computar los centroides de los clusters de la partición corriente. El centroide es el centro (punto medio) del cluster.
iii. Asignar cada objeto al cluster cuyo centroide sea más cercano.
iv. 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.

4.3.2 Algoritmos de inducción

4.3.2.1 Algoritmo *ID3*

4.3.2.1.1 Introducción

El algoritmo *ID3*, diseñado en 1993 por J. Ross Quinlan [Quinlan, 1993a], 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 [Quinlan, 1993b]. 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 masas 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 [Quinlan, 1993b].

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 esta forma [Quinlan, 1993b]:

- si C está vacío, entonces se lo 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,....., 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.

El resultado es un árbol en el cual cada hoja contiene un nombre de clase y cada nodo interior especifica una variable para ser testeada con una rama correspondiente al valor de la variable.

4.3.2.1.2 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.

Dados:

- 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* [B1urock, 1996], 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.

4.3.2.1.3 Pseudo-código del algoritmo *ID3*

A continuación, en la Figura 4.2, se presenta el algoritmo del método *ID3* para la construcción de árboles de decisión en función de un conjunto de datos previamente clasificados.

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 $\{d_j \mid j=1,2,...., m\}$ los valores del atributo D;
Sean $\{d_j \mid j=1,2,...., m\}$ los subconjuntos de S correspondientes a los valores de d_j respectivamente;
Devolver un árbol con la raíz nombrada como D y con los arcos nombrados d_l, d_2,...., d_m que van respectivamente a los árboles
ID3(R-{D}, C, Sl), ID3(R-{D}, C, S2), .., ID3(R-{D}, C, Sm);
Fin

Figura 4.2: Pseudocódigo del Algoritmo de *ID3*

4.3.2.1.4 Limitaciones de *ID3*

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 en [Gallion *et al.*, 1993] o la discretización de los mismos explicada en [B1urock, 1996; Quinlan, 1993c]. El *C4.5* resolvió el problema de los atributos continuos mediante la discretización.

4.3.2.2 Algoritmo *C4.5*

4.3.2.2.1 Introducción

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 explicarán las características particulares de este método que lo diferencian de su antecesor.

4.3.2.2.2 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 [Figura 4.3].

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 $\{d_j \mid j=1,2,...., m\}$ los valores del atributo D;
Sean $\{d_j \mid j=1,2,...., m\}$ los subconjuntos de S correspondientes a los valores de d_j respectivamente;
Devolver un árbol con la raíz nombrada como D y con los arcos nombrados d_1, d_2,....,d_m, que van respectivamente a los árboles
C4.5(R-{D}, C, Sl), C4.5(R-{D}, C, S2), C4.5(R-{D}, C, Sm);
Fin

Figura 4.3: Pseudocódigo del Algoritmo de *C4.5*

4.3.2.2.3 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 [Quinlan, 1993c]:

i. la prueba "estándar" para las variables discretas, con un resultado y una rama para cada valor posible de la variable;

ii. 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;

iii. si una variable A tiene valores numéricos continuos, se realiza una prueba binaria con resultados A <= 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 tenida en cuenta solamente cuando el conjunto T es pequeño.

5. FUENTES DE INFORMACIÓN PARA EL ANÁLISIS

5.1. INTRODUCCIÓN

En el presente capítulo se describe la estructura de la información recibida y la conformación del *data set*. Se denomina *data set* al conjunto de datos a analizar con el *software* de minería de datos. El proceso de conformación del *data set* a partir de una base de datos involucra diversas etapas:

i. consolidación de la información de interés en una única tabla;

ii. selección de los campos de interés;

iii. depuración de registros en busca de completitud y consistencia;

iv. 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 en base a la información suministrada por la DNPC.

5.2. CONSOLIDACIÓN DE LA INFORMACIÓN EN UNA ÚNICA TABLA

La base de datos recibida contiene los **1810 hechos de homicidios dolosos ocurridos en 2005 registrados en el Sistema de Alerta Temprana (SAT)**. Está compuesta por una tabla principal, dos tablas secundarias y tres tablas de referencia.

La tabla principal es:

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

Las tablas secundarias son:

- *Imputados*: contiene información referida a los imputados por un determinado hecho (sexo, edad y una referencia a si se trata de un civil o de una agente de alguna fuerza de seguridad).

- *Victimas*: contiene la misma información que la tabla *Imputados*, pero referida a las víctimas de un determinado hecho.

Las tablas de referencia describen la codificación de determinados campos de la tabla *Hechos*. Estas son:

- *Provincias*: contiene la descripción de las provincias que se encuentran codificadas en la tabla *Hechos*.

- *Departamentos*: contiene la descripción de los departamentos regionales que se encuentran codificados en la tabla *Hechos*.
- *Seccionales*: contiene la descripción de las seccionales policiales codificadas en la tabla *Hechos* y su agrupamiento por departamento y por provincia.

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

Según la opinión de los especialistas de la DNPC las tablas secundarias son de baja calidad (hay muchos registros incompletos) y aportan poca información sobre la víctima y el imputado. Por esta razón, sumado a la dificultad para consolidarlas junto a la tabla principal de forma que cada registro represente un hecho (un único hecho puede contener más de un imputado y/o víctima), estas tablas se excluyeron del análisis.

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

5.3. SELECCIÓN DE LOS CAMPOS DE INTERÉS

La tabla *Hechos* presenta una gran cantidad de campos, muchos de los cuales aportan información redundante sobre un determinado aspecto del hecho. Por esta razón se intentó preservar para el análisis los campos más representativos de cada aspecto.

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

- estados posibles: valores que puede tomar cada campo;
- 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.

5.3.1. Campos seleccionados

Provincia: contiene la provincia en la cual se cometió el hecho. Su codificación (contenida en la tabla de referencia *Provincias*) y la distribución de los hechos se muestran respectivamente en la Tabla 5.1 y la Figura 5.1.

Se puede observar que la mayor cantidad de casos ocurre en pocas provincias (Buenos Aires, Santa Fé, Ciudad de Buenos Aires y Córdoba), mientras que el resto se encuentra muy atomizado.

Estado	Descripción	Frecuencia
1	Buenos Aires	794
2	Catamarca	7
3	Córdoba	117
4	Corrientes	58
5	Chaco	61
6	Chubut	36
7	Entre Rios	59
8	Formosa	42
9	Jujuy	20
10	La Pampa	5
11	La Rioja	8
12	Mendoza	0
13	Misiones	75
14	Neuquén	30
15	Rio Negro	31
16	Salta	0
17	San Juan	10
18	San Luis	10
19	Santa Cruz	9
20	Santa Fe	234
21	Santiago del Estero	34
22	Tierra del Fuego	1
23	Tucumán	44
26	Ciudad de Buenos Aires	125
	Total	1810

Tabla 5.1: Codificación y frecuencias del campo *Provincia*

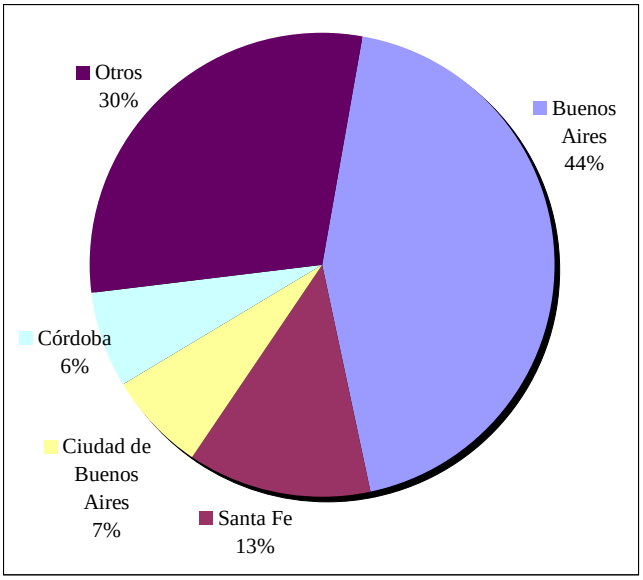


Figura 5.1: Distribución del campo *Provincia*

Hay dos provincias que no presentan casos de homicidios dolosos en 2005, Mendoza y Salta. Según lo expuesto en la sección 2.3, esto hace suponer que, al menos los homicidios dolosos de estas provincias, no fueron incorporados al SAT en ese año.

F_Mes: contiene el mes en que se cometió el hecho. La distribución se es la siguiente [Tabla 5.2 y Figura 5.2]:

Estado	Descripción	Frecuencia
1	Enero	201
2	Febrero	191
3	Marzo	183
4	Abril	153
5	Mayo	143
6	Junio	118
7	Julio	142
8	Agosto	106
9	Septiembre	120
10	Octubre	136
11	Noviembre	153
12	Diciembre	164
	Total	1810

Tabla 5.2: Codificación y frecuencias del campo *F_Mes*

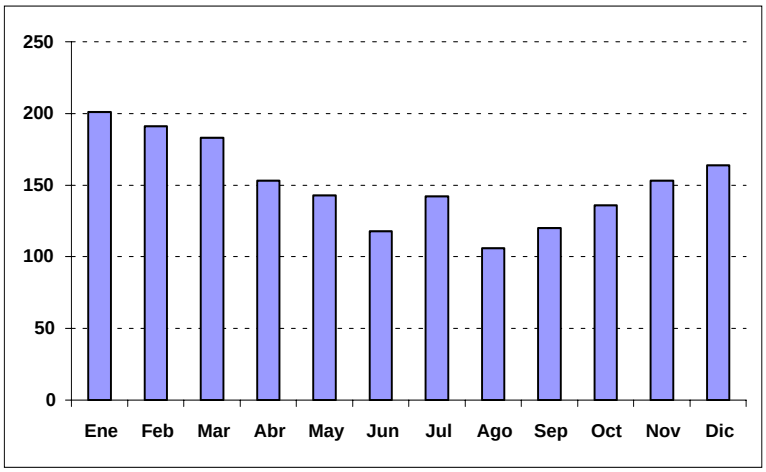


Figura 5.2: Distribución del campo *F_Mes*

Existe cierto comportamiento estacional de los hechos, registrándose una mayor cantidad de casos en los primeros meses del año. La relación entre el mes de mayor cantidad de casos (Enero, 201 casos) y el de menor (Agosto 106) es prácticamente el doble.

F_Día: contiene el día del mes en el cual se cometió el hecho. La distribución es la siguiente [Tabla 5.3 y Figura 5.3]:

Estado	Frecuencia	Estado	Frecuencia
1	95	17	60
2	50	18	55
3	62	19	75
4	69	20	60
5	60	21	61
6	64	22	61
7	70	23	51
8	51	24	61
9	54	25	67
10	75	26	55
11	63	27	56
12	47	28	56
13	62	29	44
14	45	30	40
15	41	31	33
16	67		
		Total	1810

Tabla 5.3: Codificación y frecuencias del campo *F_Día*

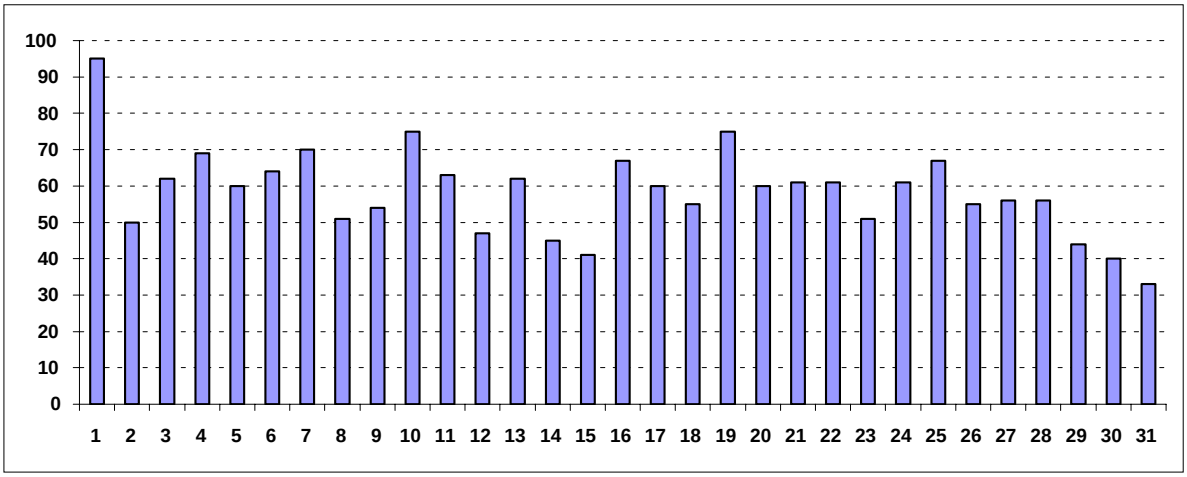


Figura 5.3: Distribución del campo *F_Día*

En este caso la distribución de los hechos es más aleatoria (no se observa un comportamiento estacional).

Día: contiene el día de la semana en que se cometió el hecho. La clasificación y su distribución son las siguientes [Tabla 5.4 y Figura 5.4]:

Estado	Descripción	Frecuencia
1	Domingo	432
2	Lunes	242
3	Martes	187
4	Miércoles	180
5	Jueves	196
6	Viernes	242
7	Sábado	331
	Total	1810

Tabla 5.4: Codificación y frecuencias del campo *Día*

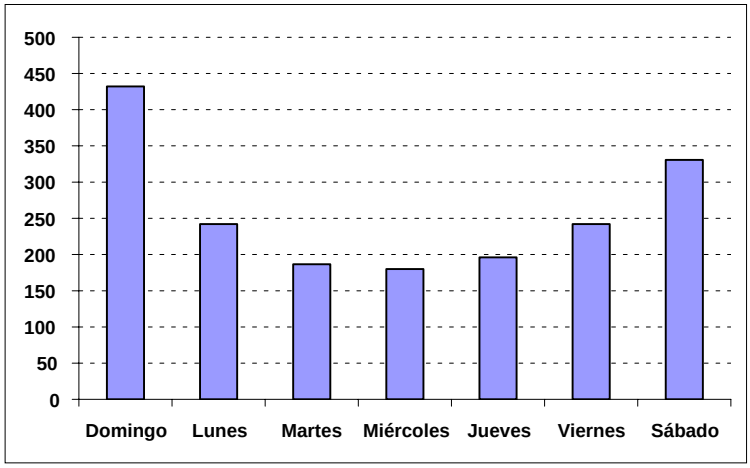


Figura 5.4: Distribución del campo *Día*

Nuevamente se observa un comportamiento estacional en la distribución temporal de los hechos, con una alta concentración durante el fin de semana. El día de la semana con mayor cantidad de casos (Domingo, 432 casos) presenta un 140% más que el de menor (Miércoles, 180 casos). La gran cantidad de casos el día domingo esta afectada por los casos ocurridos en la noche del sábado.

HoraRNue: contiene el rango horario en la cual se cometió el hecho. Los rangos y su distribución son los siguientes [Tabla 5.5 y Figura 5.5]:

Estado	Descripción	Frecuencia
0	NA	90
1	0-4 hs	301
2	4-8 hs	313
3	8-12 hs	222
4	12-16 hs	210
5	16-20 hs	264
6	20-24 hs	410
	Total	1810

Tabla 5.5: Codificación y frecuencias del campo *HoraRNue*

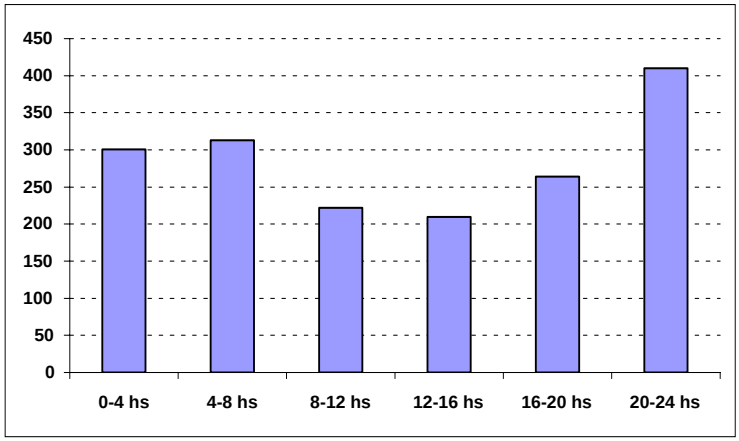


Figura 5.5: Distribución del campo *HoraRNue*

Se puede observar que existe una tendencia a una mayor cantidad de casos en horas de la noche. Hay 90 registros (5% del total) en los que no se informa la hora (identificados como NA).

TipoLugar: contiene una clasificación del tipo de lugar donde se cometió el hecho. La clasificación y su distribución se muestran en la Tabla 5.6 y la Figura 5.6, respectivamente.

Se puede observar que la mayor cantidad de casos fueron en la vía pública, seguido de domicilio particular. Hay 31 registros (2% del total) en los que no se informa el tipo de lugar.

Estado	Descripción	Frecuencia
0	NA	31
1	Vía Pública	893
2	Domicilio Particular	607
3	Comercio	88
4	Interior de Rodados	21
5	Cárcel o comisaría	38
6	Otro Lugar	132
	Total	1810

Tabla 5.6: Codificación y frecuencias del campo *TipoLugar*

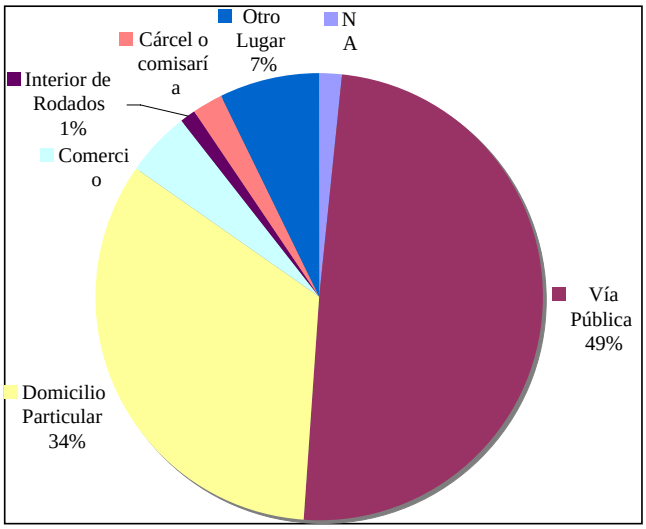


Figura 5.6: Distribución del campo *TipoLugar*

EnOcasion: da información sobre si el homicidio se cometió en ocasión de algún otro delito. La clasificación y su distribución son las siguientes [Tabla 5.7 y Figura 5.7]:

Estado	Descripción	Frecuencia
0	NA	241
1	Robo	384
2	Violación	7
3	Otro delito	189
4	No hubo otro delito	989
	Total	1810

Tabla 5.7: Codificación y frecuencias del campo *EnOcasion*

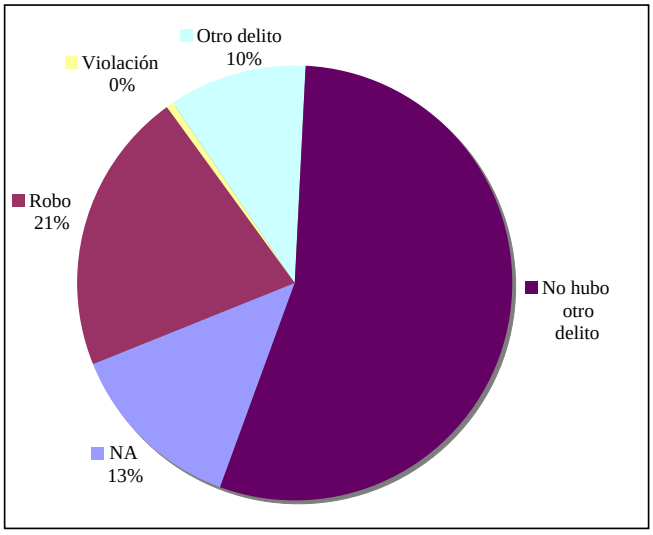


Figura 5.7: Distribución del campo *EnOcasion*

La mayor cantidad de casos no ocurrieron en ocasión de otro delito. A diferencia de los otros campos, en este caso la cantidad de registros no informados es significativa (13%). Esto se debe a la forma en que esta diseñada la planilla de SAT, por lo que en la gran mayoría de los casos la omisión de completar este campo corresponde a la omisión del hecho en sí, es decir, *no hubo otro delito*.

ClaseArma: contiene el tipo de arma con la que se cometió el homicidio. La clasificación y su distribución se muestran en la Tabla 5.8 y la Figura 5.8, respectivamente.

Se puede observar que la mayor cantidad de casos con arma de fuego. Hay 54 registros en los que no se informa el tipo de arma.

Estado	Descripción	Frecuencia
0	NA	54
1	de Fuego	919
2	Blanca	492
3	Otra	118
4	Ninguna	227
	Total	1810

Tabla 5.8: Codificación y frecuencias del campo *Arma*

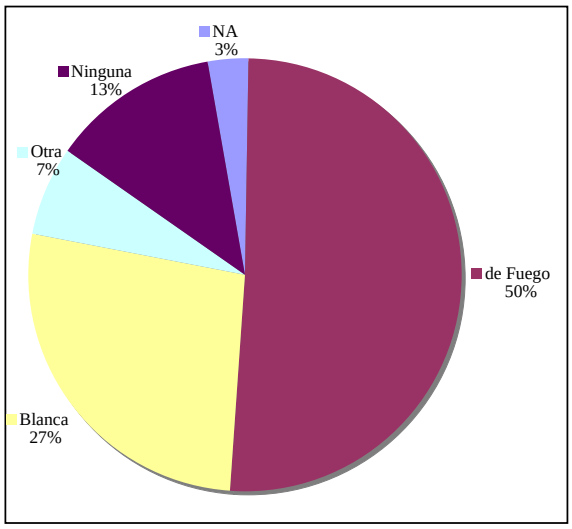


Figura 5.8: Codificación y frecuencias del campo *Arma*

5.3.2. Campos omitidos

Id: contiene el número de registro de cada hecho. Se omitió porque no aporta información referida al hecho.

Anio: contiene el año en que se cometió el hecho. Se omitió debido a que todos los hechos a analizar pertenecen al mismo año.

Sumario: contiene un número de uso interno para el Sistema Judicial Se omitió porque no aporta información referida al hecho.

CodSeccional: contiene el código de la Seccional en la cual se cometió el hecho. Se omitió porque presenta una gran cantidad de registros no informados (27%).

Calle: contiene el nombre de la calle en la cual se cometió el hecho. Se omitió porque no se cuenta con los sistemas para geo-referenciar la información.

Altura: contiene la numeración de la calle en la cual se cometió el hecho Se omitió porque no se cuenta con los sistemas para geo-referenciar la información.

Intersección: contiene el nombre de la calle de intersección en la cual se cometió el hecho. Se omitió porque no se cuenta con los sistemas para geo-referenciar la información.

CodDepa: contiene el código del departamento regional en el cual ocurrió el hecho. Se omitió porque presenta demasiados estados diferentes y aporta información similar a la del campo *Provincia*.

Localidad: contiene el nombre de la localidad en la cual se cometió el hecho. Se omitió porque no pertenece a una clasificación formal (no hay ninguna tabla que indique la pertenencia de las distintas localidades a los departamentos), contiene demasiados

estados diferentes (610) y estados ambiguos (por ejemplo “Lomas de Zamora”, “L. Zamora” y “L. de Zamora”).

FechaHecho: 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 (*F_Día*) y al mes (*F_Mes*) por separado.

Mes: contiene el mes en el cual ocurrió el hecho. Se omitió porque esta información ya esta considerada en el campo *F_Mes*.

F_Anio: contiene el año en que se cometió el hecho extraído del campo FechaHecho. Se omitió debido a que, como se mencionó anteriormente, todos los hechos a analizar pertenecen al mismo año.

HoraHecho: contiene la hora en la cual se cometió el hecho en formato hh:mm a.m./p.m. Se omitió porque esta información ya se encuentra representada por el campo *HoraRNue*.

HoraR: contiene una categorización de la hora según 4 intervalos de 6 horas (0-6; 6-12; 12-18;18-24). Se omitió porque se prefirió al campo *HoraRNue* para representar esta información, ya que es similar pero con mayor nivel de detalle.

Otro TipoLugar: contiene una descripción del campo *TipoLugar* en los casos en que se seleccionó la opción otro lugar. Se omitió porque la poca cantidad de registros en relación a la gran cantidad de estados diferentes no justifican una re-categorización del estado otro tipo lugar.

Otra ClaseArma; contiene una descripción del campo *Arma* en los casos en que se seleccionó la opción otro tipo de arma. Se omitió porque la poca cantidad de registros en relación a la gran cantidad de estados diferentes no justifican una re-categorización del estado otra clase arma.

OtroDelito: contiene una descripción del campo *EnOcasion* en los casos en los que se seleccionó la opción otro delito Se omitió porque la poca cantidad de registros en relación a la gran cantidad de estados diferentes no justifican una re-categorización del estado otro delito.

5.4. DEPURACIÓN DE REGISTROS

Como se pudo ver algunos 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 algunos especialistas de minería de datos [Ochoa, 2004] de 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:

- *HoraRNue:* se completaron los 90 registros que no tenían asignado un intervalo horario con la media total los intervalos (3,61);
- *TipoLugar*: se completaron los 31 registros que no tenían especificado el tipo de lugar con la moda del campo: *vía pública*;
- *EnOcasion*: se completaron los 241 registros que no tenían especificado si hubo otro delito con la moda del campo: *no hubo otro delito*. Como se mencionó anteriormente, por el tipo de planilla utilizada, este reemplazo tiene sentido más allá del criterio adoptado desde el punto de vista de la minería de datos;
- *ClaseArma*: se completaron los 54 registros que no tenían especificado el tipo de arma con la moda del campo: *arma de fuego*.

5.5. MODIFICACIÓN DE LOS ESTADOS ORIGINALES DE CADA CAMPO

A partir de este momento, para ser consistentes con la terminología de *Weka*, hablaremos de atributo para referirnos a la información aportada por un determinado campo en el *data set*. Dicho de otra forma, lo que en el contexto de una tabla llamamos campo, en el contexto del *data set* denominaremos 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 (*F_Mes*, *F_Día*, *Día* y *HoraRNue*) es de naturaleza cíclica (el último estado antecede al primero). Esto no puede ser representado correctamente en *Weka*, sino que debemos conformarnos con una asignación lineal. Esto presenta un problema a la hora de agrupar los registros ya que, 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 si se puede morigerar, 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 interfase 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 Provincia (Prov)

Corresponde al campo *Provincia*. Debido a que se trata de un atributo categórico, se han reemplazado los estados originales por los siguientes [Tabla 5.9]:

Estado Original	Nuevo Estado	Descripción
1	BsAs	Buenos Aires
2	Cata	Catamarca
3	Cord	Córdoba
4	Corr	Corrientes
5	Chac	Chaco
6	Chub	Chubut
7	E.Rios	Entre Rios
8	Form	Formosa
9	Juj	Jujuy
10	L.Pam	La Pampa
11	L.Rio	La Rioja
12	Mend	Mendoza
13	Misio	Misiones
14	Neuq	Neuquén
15	R.Neg	Rio Negro
16	Salt	Salta
17	S.Jua	San Juan
18	S.Lui	San Luis
19	S.Cru	Santa Cruz
20	S.Fe	Santa Fe
21	S.Est	Santiago del Estero
22	T.Fue	Tierra del Fuego
23	Tucu	Tucumán
26	CBA	Ciudad de Buenos Aires

Tabla 5.9: Nuevos estados del atributo *Provincia*

Atributo Mes (Mes)

Corresponde al campo *F_Mes*. Debido a que se trata de un atributo continuo con comportamiento estacionario, para minimizar el efecto antes comentado, se ha decidido comenzar la escala en el mes que presenta la menor cantidad de casos (Agosto). De esta forma, el reemplazado de los estados originales se muestra en la Tabla 5.10.

Atributo Día del Mes (DMes)

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

Estado Original	Nuevo Estado	Descripción
1	6	Enero
2	7	Febrero
3	8	Marzo
4	9	Abril
5	10	Mayo
6	11	Junio
7	12	Julio
8	1	Agosto
9	2	Septiembre
10	3	Octubre
11	4	Noviembre
12	5	Diciembre

Tabla 5.10: Nuevos estados del atributo *Mes*

Atributo Día de la Semana (DSem)

Corresponde al campo *Día*. 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 (Miércoles). De esta forma se han reemplazado los estados originales por los siguientes [Tabla 5.11]:

Estado Original	Nuevo Estado	Descripción
1	5	Domingo
2	6	Lunes
3	7	Martes
4	1	Miércoles
5	2	Jueves
6	3	Viernes
7	4	Sábado

Tabla 5.11: Nuevos estados del atributo *Día de la Semana*

Atributo Hora (Hora)

Corresponde al campo *HoraRNue*. En este caso puntual se ha decidido comenzar la escala en el intervalo 8-12hs ya que existe una barrera natural entre este intervalo y el anterior (originada en el hecho de que alrededor de las 8hs comienza la actividad laboral). Por esta razón se han reemplazado los estados originales por los siguientes [Tabla 5.12]:

Estado Original	Nuevo Estado	Descripción
1	5	0-4 hs
2	6	4-8 hs
3	1	8-12 hs
4	2	12-16 hs
5	3	16-20 hs
6	4	20-24 hs

Tabla 5.12: Nuevos estados del atributo *Hora*

Atributo Lugar (Lugar)

Corresponde al campo *TipoLugar*. Debido a que se trata de un atributo categórico, se han renombrado los estados originales por los siguientes [Tabla 5.13]:

Estado Original	Nuevo Estado	Descripción
1	V.Púb	Vía Pública
2	D.Par	Domicilio Particular
3	Com	Comercio
4	Rod	Interior de Rodados
5	C-C	Cárcel o comisaría
6	Otro	Otro Lugar

Tabla 5.13: Nuevos estados del atributo *Lugar*

Atributo Otro Delito (Ot.Del)

Corresponde al campo *EnOcasion*. Debido a que se trata de un atributo categórico, se han renombrado los estados originales por los siguientes [Tabla 5.14]:

Estado Original	Nuevo Estado	Descripción
1	Robo	Robo
2	Viol	Violación
3	Otro	Otro delito
4	NoHub	No hubo otro delito

Tabla 5.14: Nuevos estados del atributo *Otro Delito*

Atributo Arma (Arma)

Corresponde al campo *TipoArma*. Debido a que se trata de un atributo categórico, se han renombrado los estados originales por los siguientes [Tabla 5.15]:

Estado Original	Nuevo Estado	Descripción
1	Fuego	de Fuego
2	Blanca	Blanca
3	Otra	Otra
4	Ning	Ninguna

Tabla 5.15: Nuevos estados del atributo *Arma*

5.6. *DATA SET* DEFINITIVO

Finalmente el *data set* queda compuesto de los 1810 registros originales y 8 atributos que aportan información de distintas dimensiones de los hechos:

- Espacial: información sobre la distribución geográfica de los hechos representada por el atributo *Provincia*.
- 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 *Arma, Lugar* y *Otro Delito*.

6. HERRAMIENTAS PARA EL ANÁLISIS

6.1. INTRODUCCIÓN

En este capítulo se explican 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. A modo didáctico se presenta un caso de aplicación.

6.2. PRESENTACIÓN DEL CASO

Se extrajo una muestra de la base de datos, correspondiente a la totalidad de homicidios dolosos ocurridos en la comisaría 36 de la Ciudad Autónoma de Buenos Aires durante 2005. Se eligió la Ciudad de Buenos Aires por ser un lugar de conocimiento general y la comisaría 36 por ser la que más casos presenta en dicho año (15 casos).

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 consideradas.

Los atributos seleccionados para el ejemplo fueron: *Lugar*, *Arma*, *Día de la semana* (*DSem*) y *Hora* (*Hora*) [Tabla 6.1].

Atributo	Tipo	Variables	Rango	Descripción
Lugar	Categórico	D.Part.		Domicilio Particular
		V.Publ.		Vía Pública
		Otro		Otro lugar (por ej. Hospital Municipal)
Arma	Categórico	Fuego		Arma de fuego
		Blanca		Arma blanca
		Ninguna		Ningun arma
DSem	Continuo		1 al 7	1 corresponde al miércoles
Hora	Continuo		1 al 6	6 intervalos de 4 horas comenzando a las 0 hs

Tabla 6.1: Descripción de atributos de la muestra

En la Tabla 6.2 se detalla el *data set* con los 15 registros y sus respectivos atributos.

Tras aplicar el algoritmo *K-means* se obtuvieron 2 clusters. En la Tabla 6.2 se muestra la pertenencia de cada registro a cada cluster (denominados 0 y 1).

6.3. DESCRIPCIÓN DE LAS HERRAMIENTAS

6.3.1. Tabla de centroides

La tabla de centroides permite conocer cuál es el centroide de cada cluster. En un sentido geométrico, el centroide es el lugar del hiper-espacio de posibles estados que

equidista de todos los casos que corresponden a un determinado cluster. En un sentido práctico no es mas que la media o la moda de cada atributo para cada cluster.

Caso	Lugar	Arma	DSem	Hora	Cluster
1	D.Part.	Fuego	6	3	1
2	Otro	Fuego	6	2	1
3	D.Part.	Ninguna	2	2	0
4	D.Part.	Blanca	5	2	0
5	Otro	Fuego	4	1	1
6	D.Part.	Fuego	2	3	1
7	V.Publ.	Blanca	1	2	0
8	D.Part.	Fuego	5	5	1
9	D.Part.	Fuego	7	6	1
10	D.Part.	Fuego	4	5	1
11	D.Part.	Fuego	5	6	1
12	D.Part.	Fuego	4	5	1
13	D.Part.	Fuego	5	4	1
14	V.Publ.	Fuego	1	6	1
15	D.Part.	Ninguna	3	3	0

Tabla 6.2: Asignación de clusters al Data Set

En el caso en cuestión la tabla de centroides es la siguiente [Tabla 6.3]:

		Atributos categóricos (modas)		Atributos continuos (medias)	
	Cant. (%)	Lugar	Arma	Hora	Día Semana
Cluster 0	27%	Domicilio Particular	Ninguna	2,25	2,75
Cluster 1	73%	Domicilio Particular	de Fuego	4,18	4,45
General	100%	Domicilio Particular	de Fuego	3,66	4,00

Tabla 6.3: Tabla de centroides

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 que por ejemplo para el cluster 1 la lectura debe ser “la mayoría de los casos fue en *domicilio particular* e, 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 *domicilio particular* y con *arma de fuego*” (cosa que para este caso puntual también se cumple pero no necesariamente es así).

Otro problema que presenta esta visión es que por definición la moda indica una mayoría relativa pero no la cuantifica, por lo que a priori no conocemos 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, sin embargo *ninguna arma* no es tan representativo del cluster 0 (corresponde sólo al 50% de los casos).

6.3.2. Diagramas de Venn

Como se presentó en el punto anterior, un problema que presenta la tabla de centroides para extraer conclusiones es la representatividad de los atributos categóricos para cada cluster y su nivel de solapamiento. Los Diagramas de Venn nos 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 [Figuras 6.1 y 6.2].

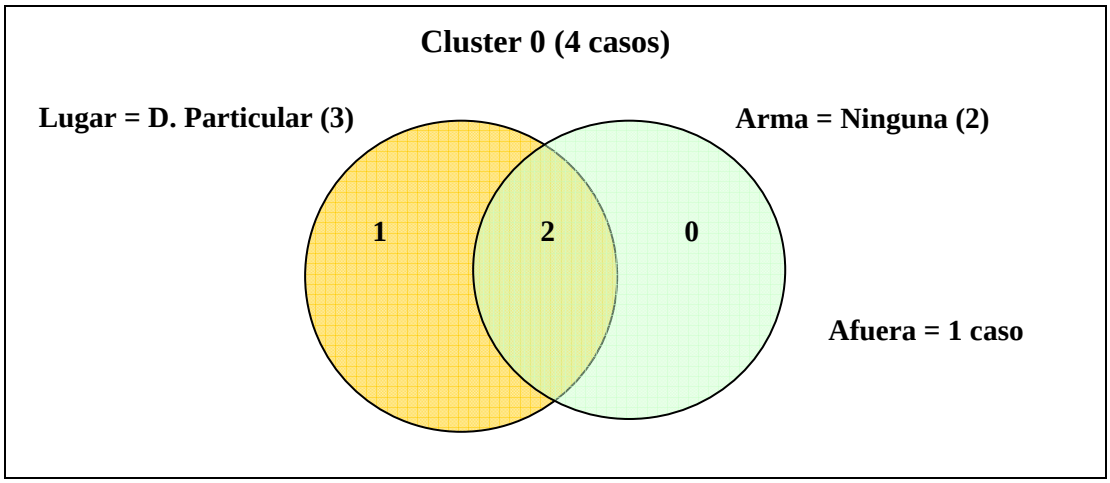


Figura 6.1: Diagrama de Venn para el cluster 0

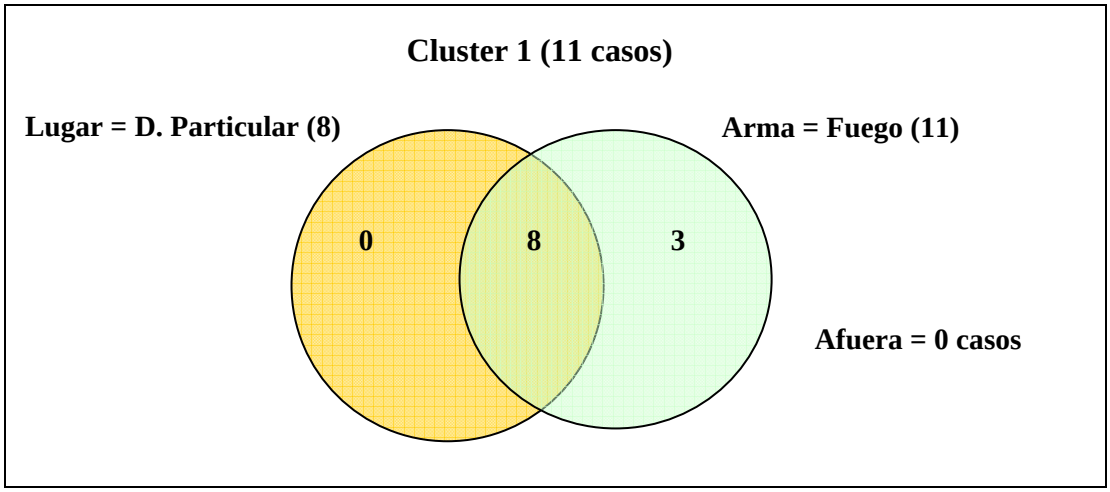


Figura 6.2: Diagrama de Venn para el cluster 1

6.3.3. Gráficos de barras

Los gráficos de barras permiten visualizar cual es la distribución de un determinado atributo según las variables de los demás atributos. En este caso utilizaremos el atributo cluster [Figura 6.3].

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 cluster (en este caso 27% para el cluster 0 y 73% para el cluster 1), se mantenga aproximadamente igual independientemente de a través de que

variable lo segmentemos. 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.

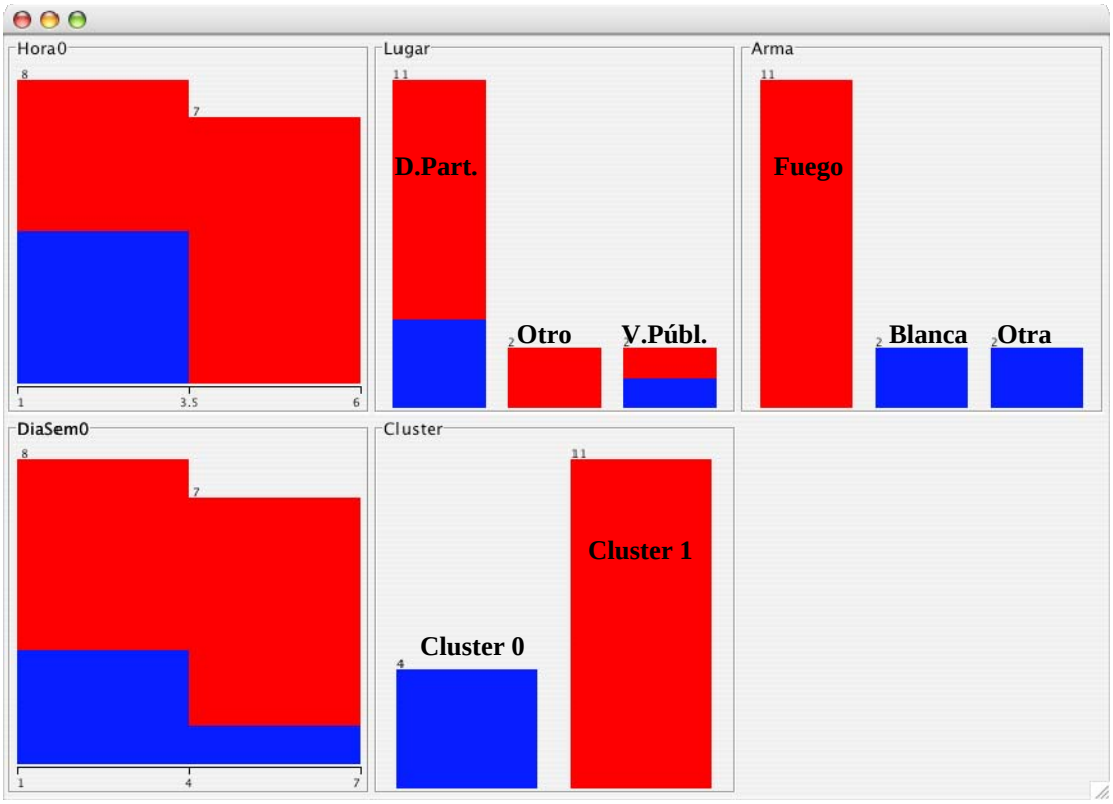


Figura 6.3: Gráficos de barras: distribución de los clusters según el resto de los atributos

Por esta razón toda distribución dentro de una determinada variable que se aparte de la distribución global (27-73 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 biunívoca entre *arma de fuego* y el cluster 1) pero no así tanto en el lugar (por ejemplo en *domicilio particular* se mantiene exacta la proporción 27-73). También hay cierto patrón de comportamiento en los atributos continuos, fundamentalmente en la hora (una alta incidencia del cluster 0 antes de las 12hs y una alta incidencia del cluster 1 después de las 12hs).

Este gráfico nos 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 nos permite visualizar dos atributos por gráfico (el cluster y otro atributo).

6.3.4. Gráficos de dispersión

Los gráficos de dispersión se representan en ejes cartesianos: cada eje representa un atributo y cada punto un hecho. Estos gráficos 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.

6.3.4.1. Gráficos de distribución

Son un caso especial de los gráficos de dispersión, en 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 6.4]:

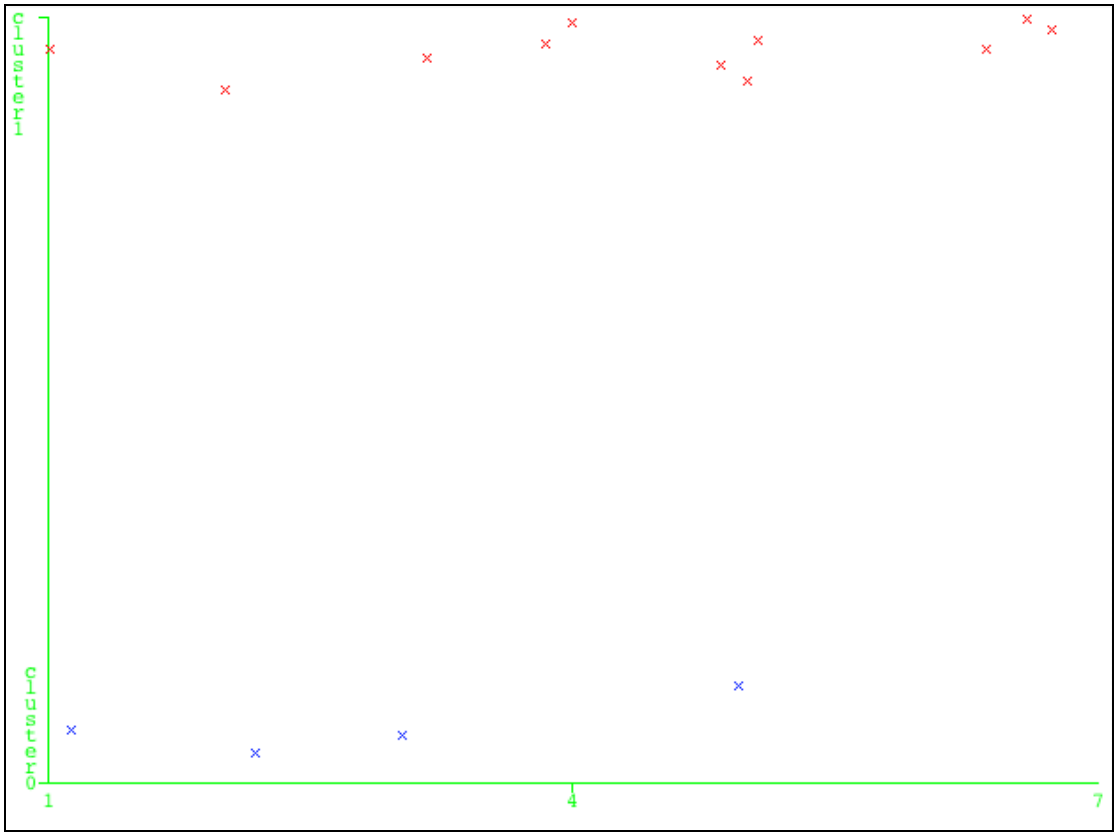


Figura 6.4: Gráfico de distribución de los clusters según *día de la semana*

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 miércoles, podemos ver que el cluster 1 presenta una alta proporción de casos durante el fin de semana (entre el sábado y el lunes).

Es importante aclarar que las coordenadas de cada punto no son precisas. Por ejemplo se ve que para cada cluster los puntos no están alineados. Este es un efecto hecho ad hoc ya que al tratarse de variables discretas, coordenadas precisas llevarían a la superposición de puntos, perdiendo noción de la cantidad. De no ser por este efecto, por

ejemplo, los tres casos que corresponden al cluster 1 y al día 5 estarían superpuestos y se verían como un único punto.

6.3.4.2. Gráficos de interrelaciones

Estos gráficos de dispersión permiten visualizar 3 atributos al mismo tiempo e identificar cual 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 6.5]:

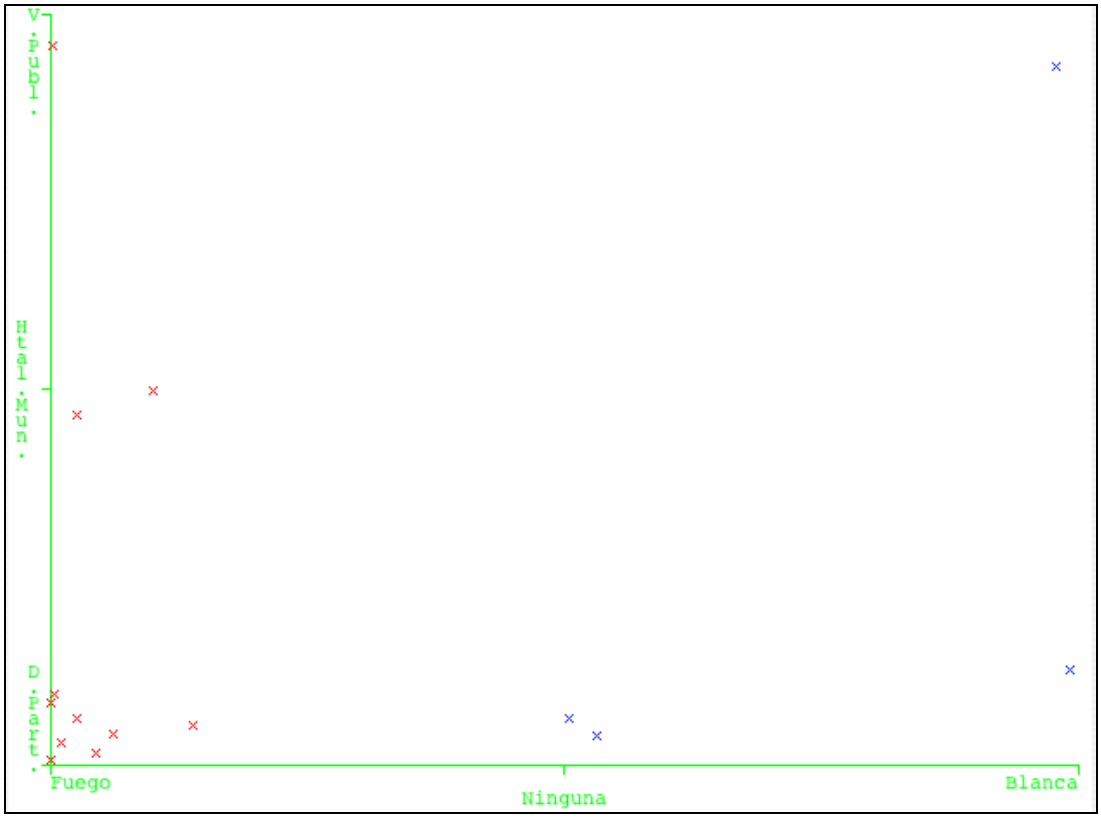


Figura 6.5: Gráfico de interrelación *arma-lugar*

Cuando para un determinado cruce de variables observamos una acumulación de puntos de un determinado tipo, entonces diremos 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 diremos que hay una fuerte interacción entre *arma de fuego* y *domicilio particular* para el cluster 1.

6.3.5. Árbol de clasificación

Luego del análisis anterior utilizaremos 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 6.6].

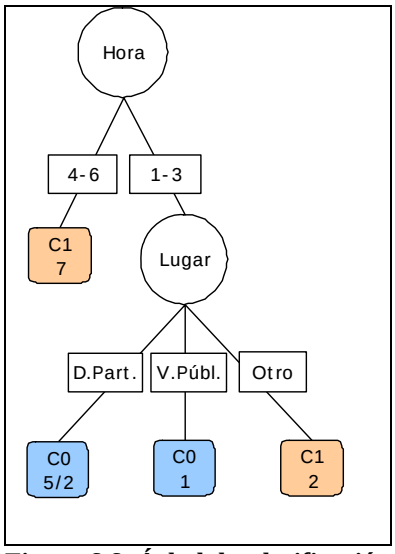


Figura 6.6: Árbol de clasificación

Los árboles presentan nodos, en donde se evalúa un determinado atributo (en este caso *hora* y *lugar*); 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 clusters 0 y 1). En las hojas se muestra la cantidad total de registros clasificados y, separado 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. En este caso las reglas son:

Regla 1

SI Hora = 4-6
ENTONCES Cluster 1 (7)

Regla 2

SI Hora = 1-3
Y Lugar = D.Part.
ENTONCES Cluster 0 (5/2)

Regla 3

SI Hora = 1-3
Y Lugar = V.Públ.
ENTONCES Cluster 0 (1)

Regla 4

SI Hora = 1-3
Y Lugar = Otro
ENTONCES Cluster 1 (2)

6.3.6. Matrices de confusión

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

Para este caso la matriz de confusión es [Tabla 6.4]:

		Clasificado como		TP Rate
		0	**1**	
Real	**0**	4	0	**100%**
	1	2	9	**82%**

Tabla 6.4: Matriz de confusión

Si la clasificación hubiera sido perfecta esperaríamos encontrar únicamente elementos en la diagonal. Como se puede ver en este caso hay 2 hechos que corresponden al cluster 1 que fueron mal clasificados por el árbol (registros 6 y 14). El TP Rate (*Total Predictive Rate*) indica cual es la probabilidad de clasificar correctamente un registro para cada cluster. Por ejemplo para el cluster 1 esta probabilidad es 9/11 = 82%.

7. RESULTADOS EXPERIMENTALES

7.1. INTRODUCCIÓN

Se el analizó el *data set* obtenido en el capitulo 5 con el *software Weka 3.5.5*. En primer lugar se aplicó el algoritmo *K-means* para agrupar los 1810 registros en 3 clusters. Luego, con las herramientas descriptas 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. En el Anexo 3 se muestran todos los pasos de este proceso desde la interfase de *Weka*.

7.2. CLUSTERING

Se procedió a agrupar el *data set* en 3 grupos 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 sucesivas iteraciones. Para la selección de este número se realizaron 20 corridas consecutivas 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 [Tabla 7.1] se presentan los resultados obtenidos para las 20 corridas:

Semilla	Suma error cuadrático	Número de iteraciones	Semilla (cont)	Suma error cuadrático	Número de iteraciones
1	3055	14	11	3203	12
2	3281	22	12	3105	10
3	3088	17	13	3122	17
4	3133	13	14	3471	14
5	3147	19	15	3212	11
6	3285	22	16	3105	19
7	3285	23	17	3277	14
8	3278	50	18	3242	13
9	3593	22	19	3467	16
10	3171	11	20	3112	17

Tabla 7.1: Resultado de *K-Means* para 3 clusters con varias semillas

Como se puede ver la menor suma de error cuadrático se obtuvo con una semilla de 1.

7.2.1. Tabla de centroides

El resultado obtenido con *Weka* tras la ejecución de *simple K-means* con 3 clusters y una semilla de 1 se sintetiza en la tabla de centroides [Tabla 7.2].

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*, *arma* y *otro delito* que parecerían estar identificando a los clusters.

		Atributos categóricos (modas)				Atributos continuos (medias)			
	Cant. (%)	Provincia	Lugar	Arma	Otro Delito	Hora	Día Semana	Día Mes	Mes
Cluster 0	22%	BsAs	Vía Pública	de Fuego	Robo	19	Sábado	16	7
Cluster 1	43%	BsAs	Vía Pública	de Fuego	No Hubo	17	Sábado	15	7
Cluster 2	35%	BsAs	Domicilio Particular	Blanca	No Hubo	21	Sábado	15	7
General	100%	BsAs	Vía Pública	de Fuego	No Hubo	19	Sábado	15	7

Tabla 7.2: Tabla de centroides

7.2.2. Diagramas de Venn

Como se explicó en la sección 6.3.1, los centroides no necesariamente representan la combinación de atributos más frecuente, y por lo tanto es necesario un análisis más detallado para caracterizar a los clusters. A continuación se presentan diagramas de Venn indicando la composición real de cada cluster según los 3 atributos de interés mencionados [Figura 7.1, Figura 7.2 y Figura 7.3].

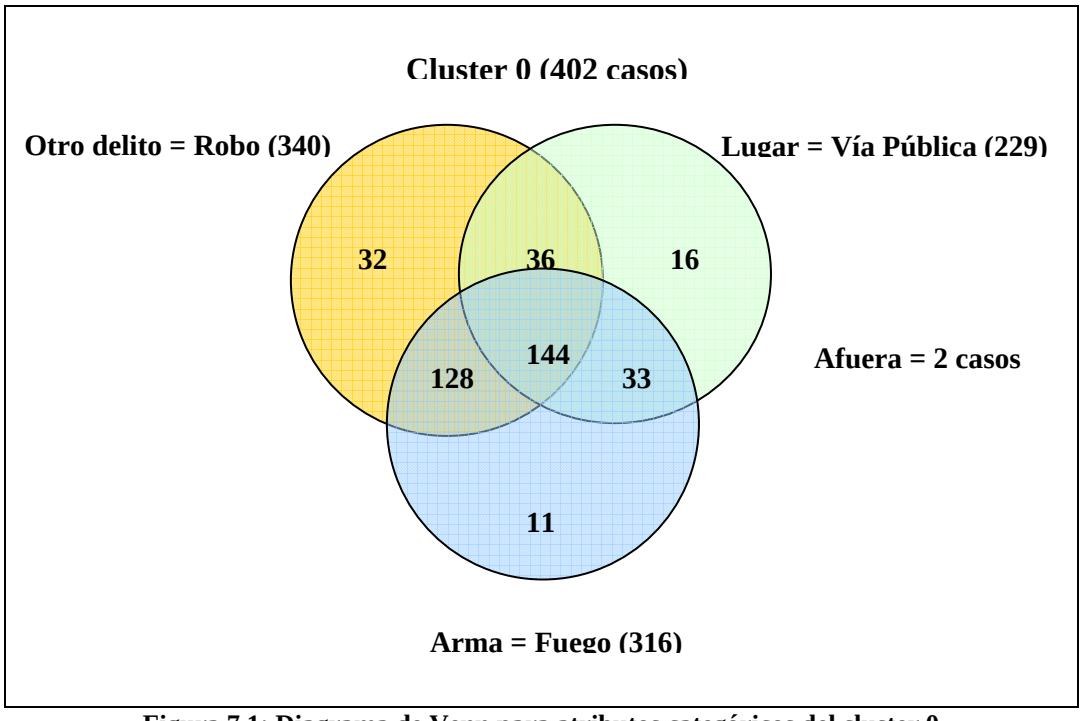


Figura 7.1: Diagrama de Venn para atributos categóricos del cluster 0

El 85% de los registros agrupados en este cluster cumple con al menos 2 de los atributos. Este cluster parecería estar caracterizado por **homicidios en ocasión de robo con arma de fuego (68% de los casos)**. La vía pública no parecería ser característica de este cluster (el 43% de los casos no fueron la vía pública).

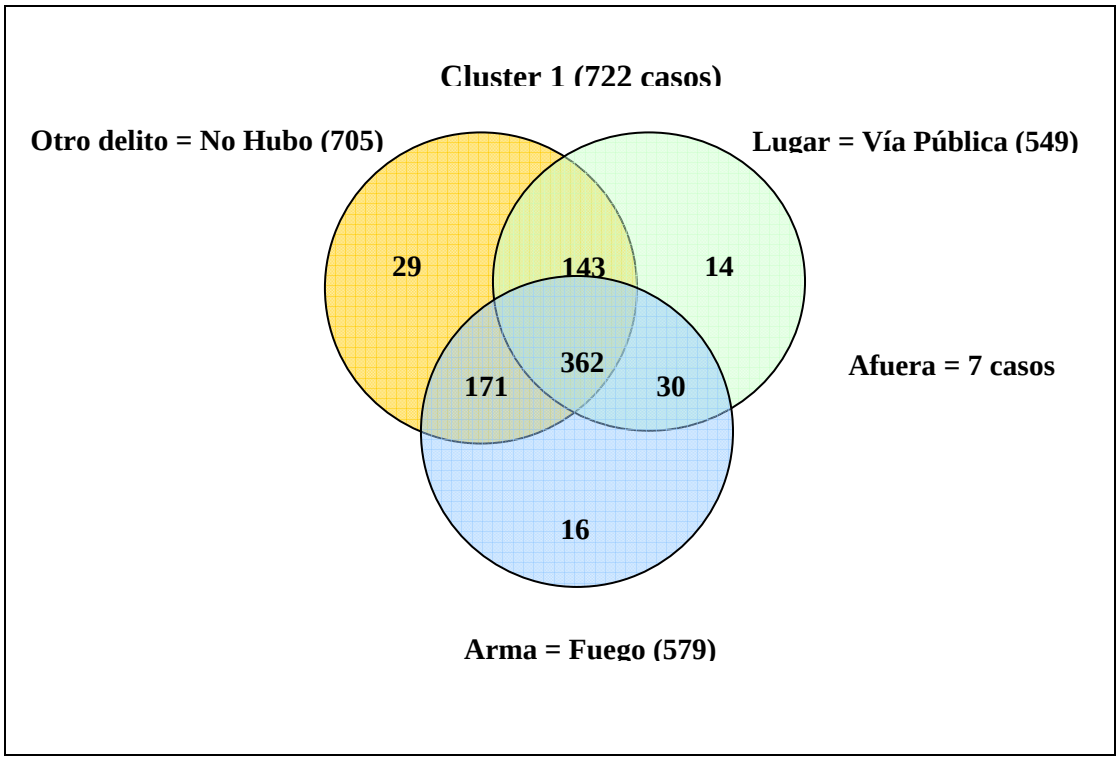


Figura 7.2: Diagrama de Venn para atributos categóricos del cluster 1

El 98% de los registros agrupados en este cluster cumple con al menos 2 de los atributos. Este cluster parecería estar caracterizado por **homicidios con arma de fuego sin ocasión de otro delito (74% de los casos)**.

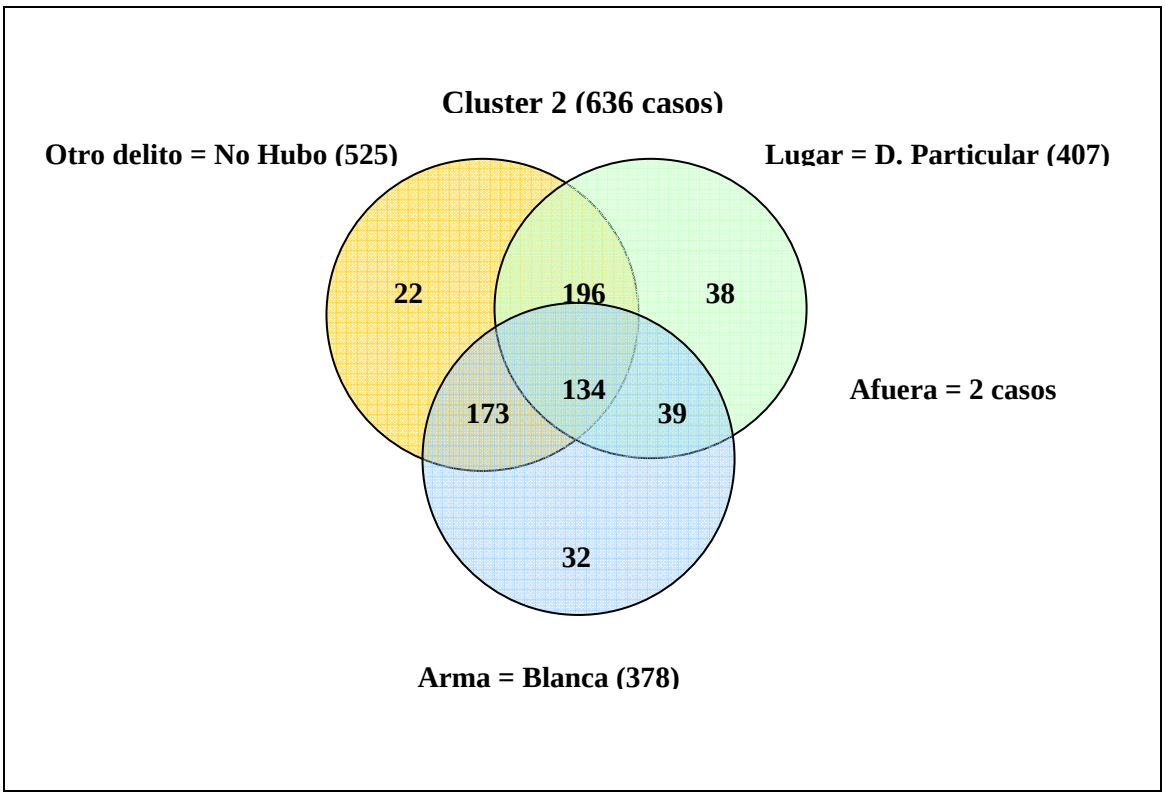


Figura 7.3: Diagrama de Venn para atributos categóricos del cluster 2

El 85% de los registros agrupados en este cluster cumple con al menos 2 de los atributos. Este cluster parecería estar caracterizado, de forma más difusa que los clusters anteriores, por **homicidios ocurridos en domicilio particular sin ocasión de otro delito (52% de los casos**).

7.2.3. Gráficos de barras

Como hemos dicho 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, esperaríamos encontrar una proporción aproximada de 43% rojo (cluster 1); 22% azul (cluster 0) y 35% turquesa (cluster 2) en cada variable de cada atributo. Si bien en algunos atributos esta proporción se cumple (*día mes* y *provincia*) en otros existen interacciones significativas (por ejemplo cluster 2 con *arma blanca* y *domicilio particular*). [Figura 7.4]

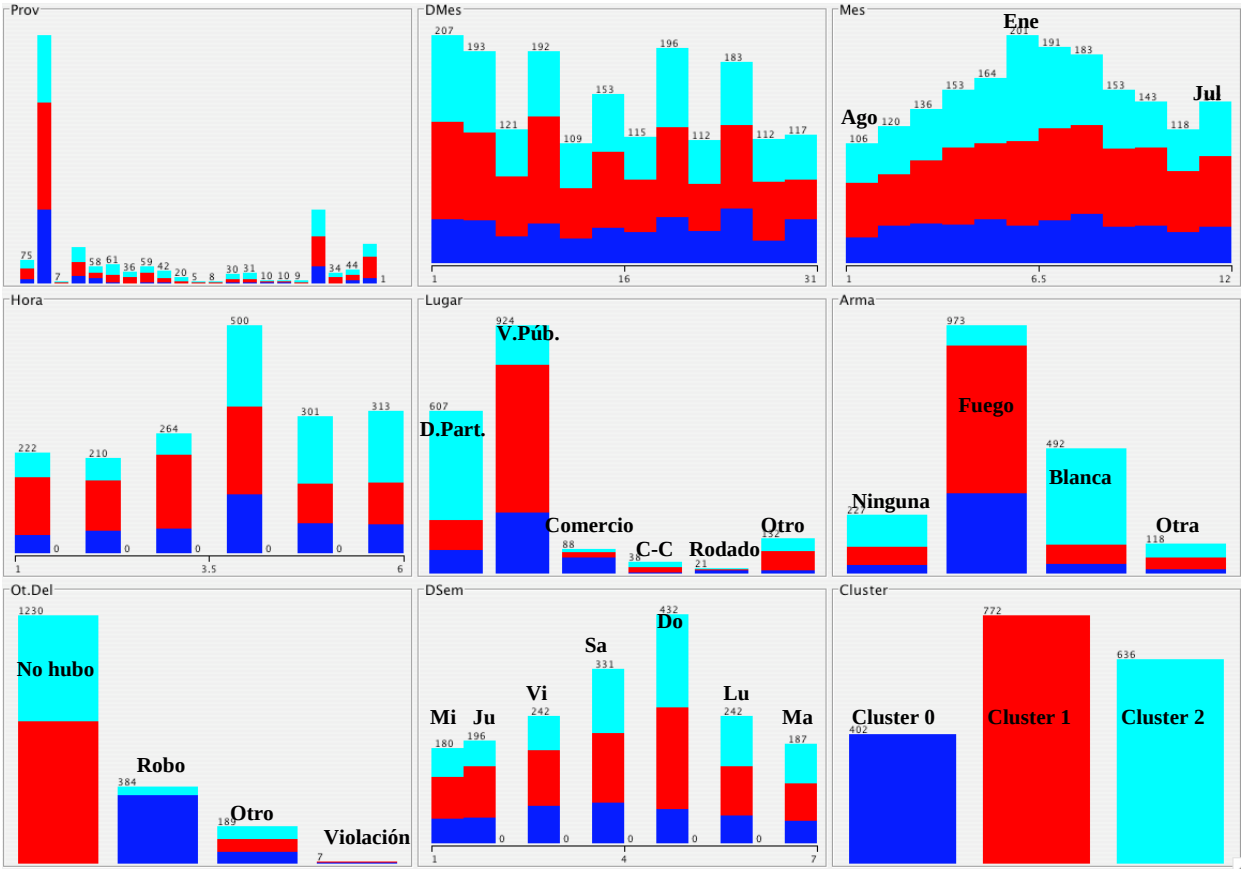


Figura 7.4: Distribución de los clusters según atributos

Los atributos donde se observan más interacción entre las variables y los clusters son: *lugar, arma, otro delito* y *día de la semana*. Examinaremos detenidamente estas interacciones.

7.2.4. Gráficos de dispersión

7.2.4.1. Distribución de los clusters según el atributo lugar

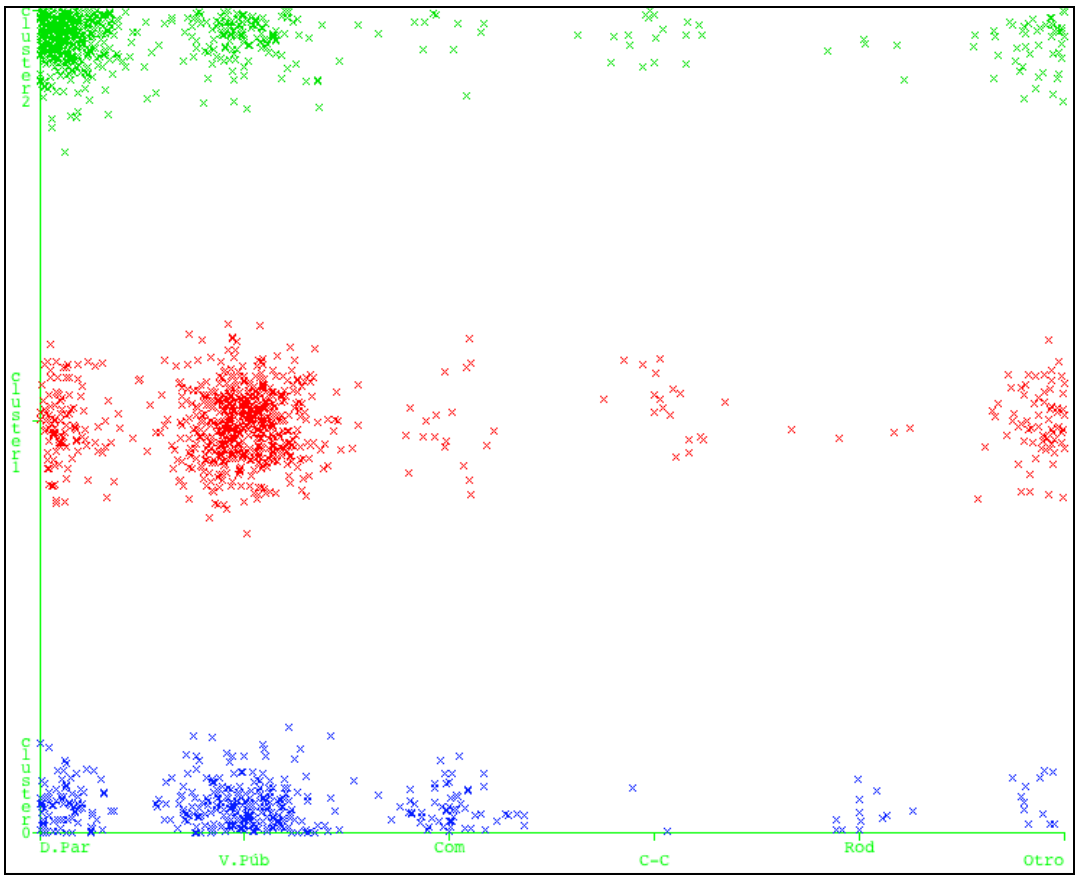


Figura 7.5: Distribución de clusters según atributo *lugar*

Mientras el cluster 2 esta muy concentrado en *domicilio particular* y el cluster 1 en *vía pública*, el cluster 0 se encuentra distribuido más homogéneamente [Figura 7.5]. Si bien este último presenta la mayoría de registros en *vía pública*, tiene una alta proporción de homicidios en comercios respecto a los otros clusters.

7.2.4.2. Distribución de los clusters según el atributo arma

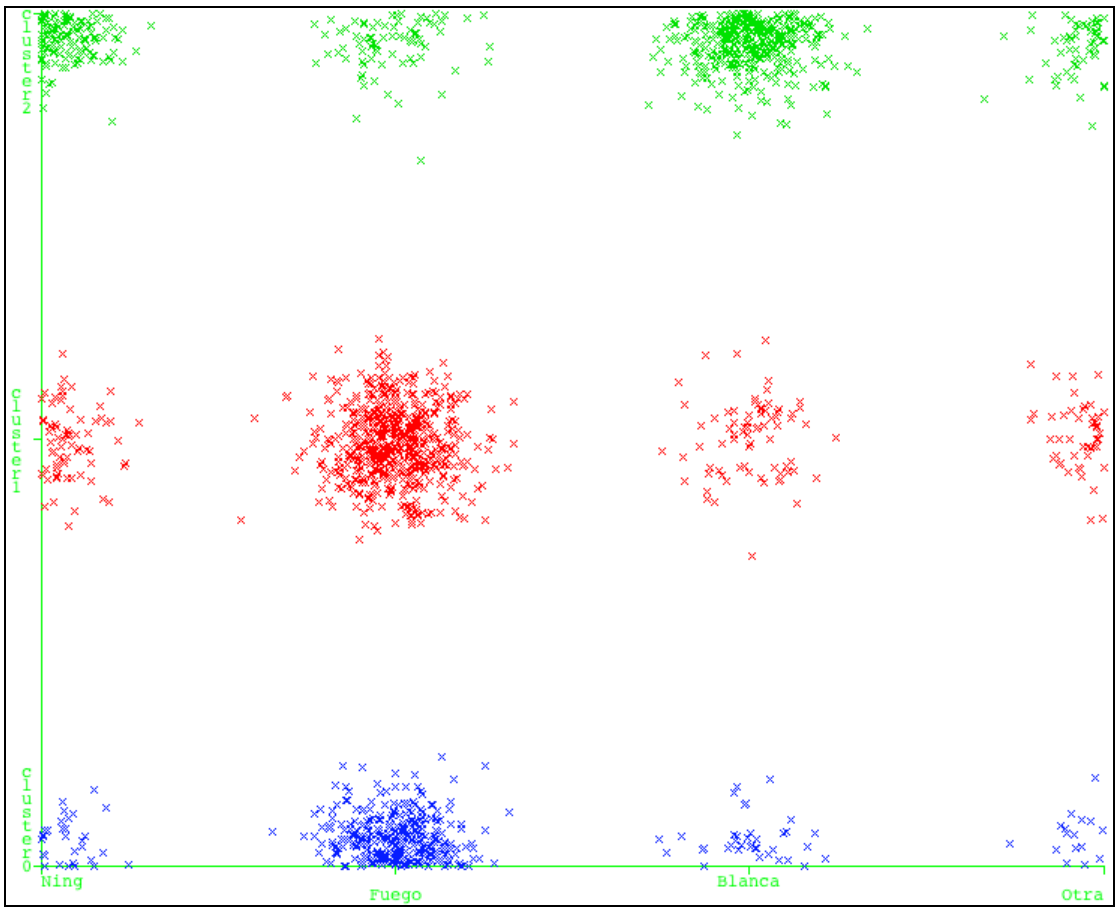


Figura 7.6: Distribución de clusters según atributo *arma*

El cluster 1 y el cluster 0 presentan una distribución similar, con una alta concentración en *arma de fuego*, seguida por *arma blanca* y prácticamente muy pocos casos *sin arma* [Figura 7.6]. En contraposición el cluster 2 presenta muy pocos casos con *arma de fuego* (una proporción muy baja respecto a la proporción global) y muchos casos *con arma blanca* y *sin arma* (una proporción muy alta respecto a la proporción global).

7.2.4.3. Distribución de los clusters según el atributo otro delito

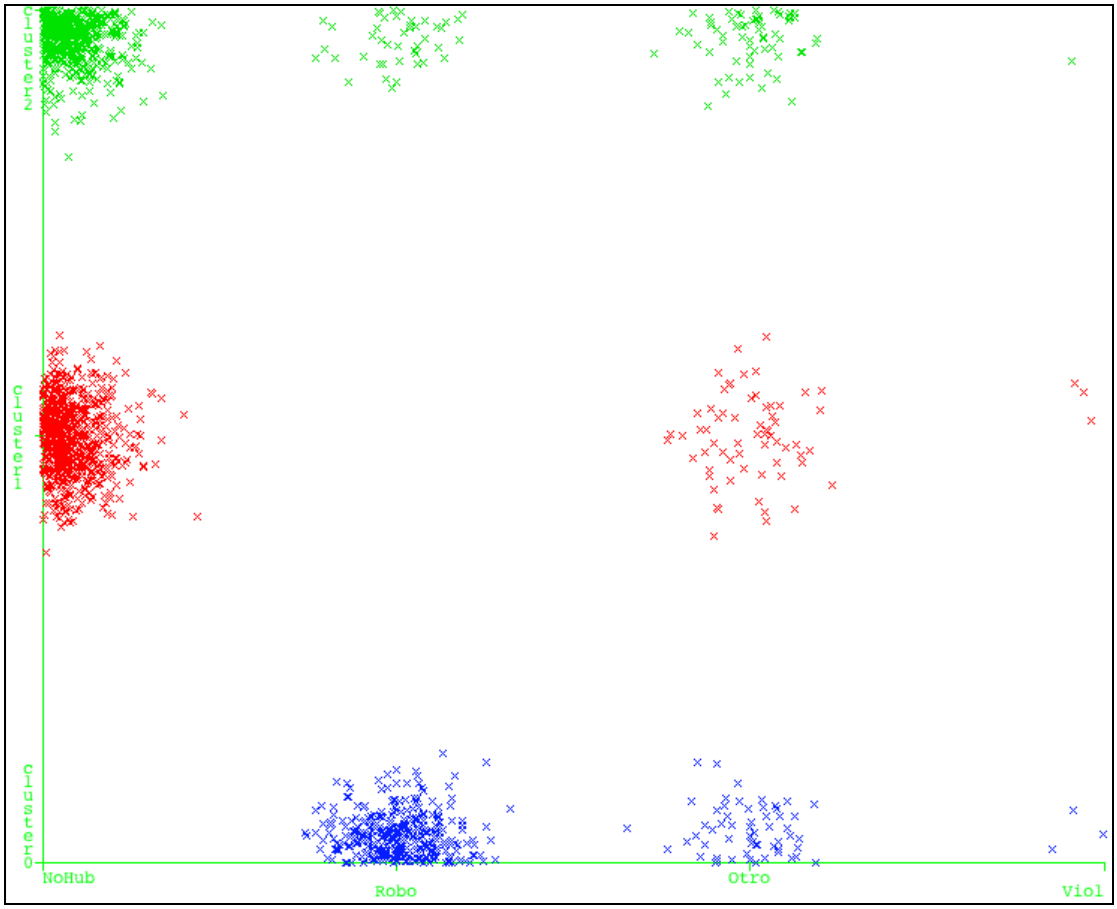


Figura 7.7: Distribución de clusters según atributo *otro delito*

El cluster 1 presenta una muy alta proporción de hechos donde *no hubo* otro delito y ningún caso de *robo*. Si bien el cluster 2 sigue esta tendencia, está menos polarizado, presentando algunos casos de *robo* [Figura 7.7]. El cluster 0 presenta una proporción muy alta de homicidios en ocasión de *robo* y ningún caso donde *no hubo* otro delito.

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

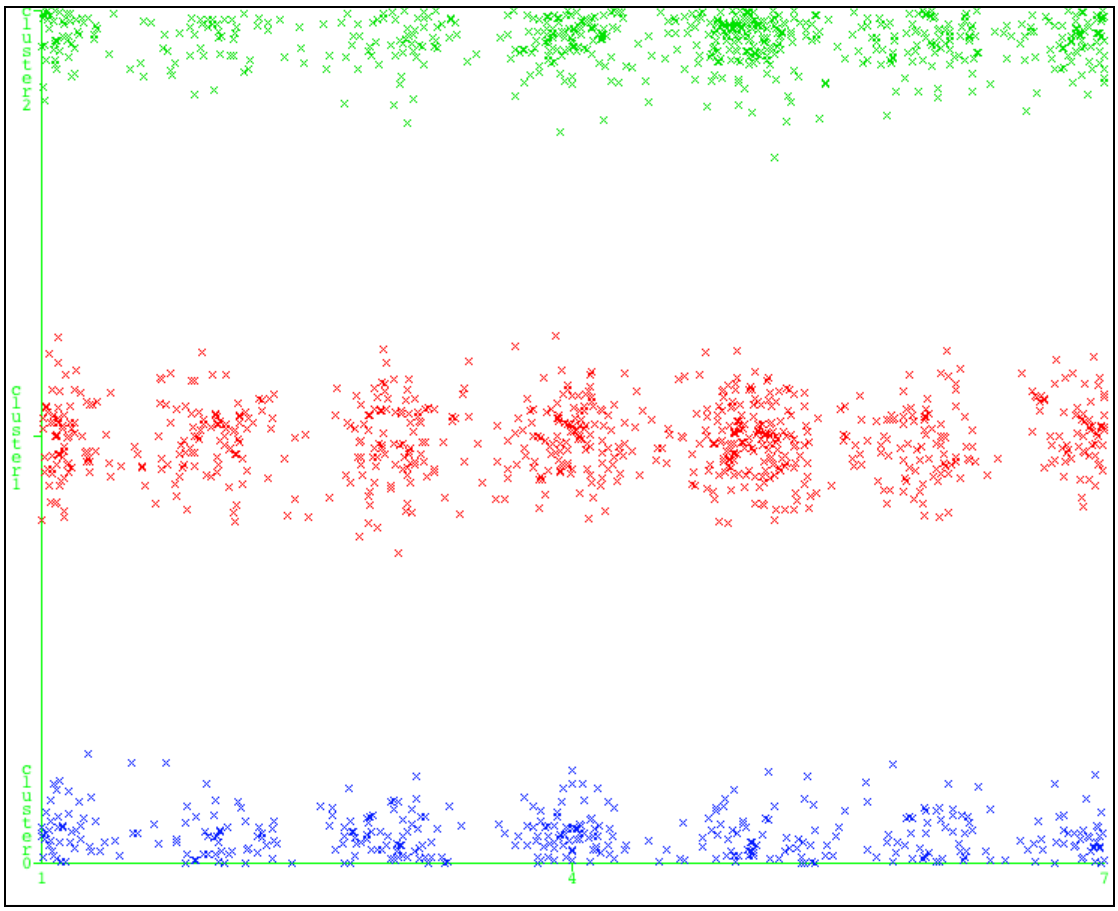


Figura 7.8: Distribución de clusters según atributo *día de la semana*

El cluster 0 presenta una distribución relativamente homogénea [Figura 7.8].

El cluster 2, y en menor medida el cluster 1, presentan una alta concentración de casos los sábados y domingos.

7.2.4.5. Interrelación lugar-arma

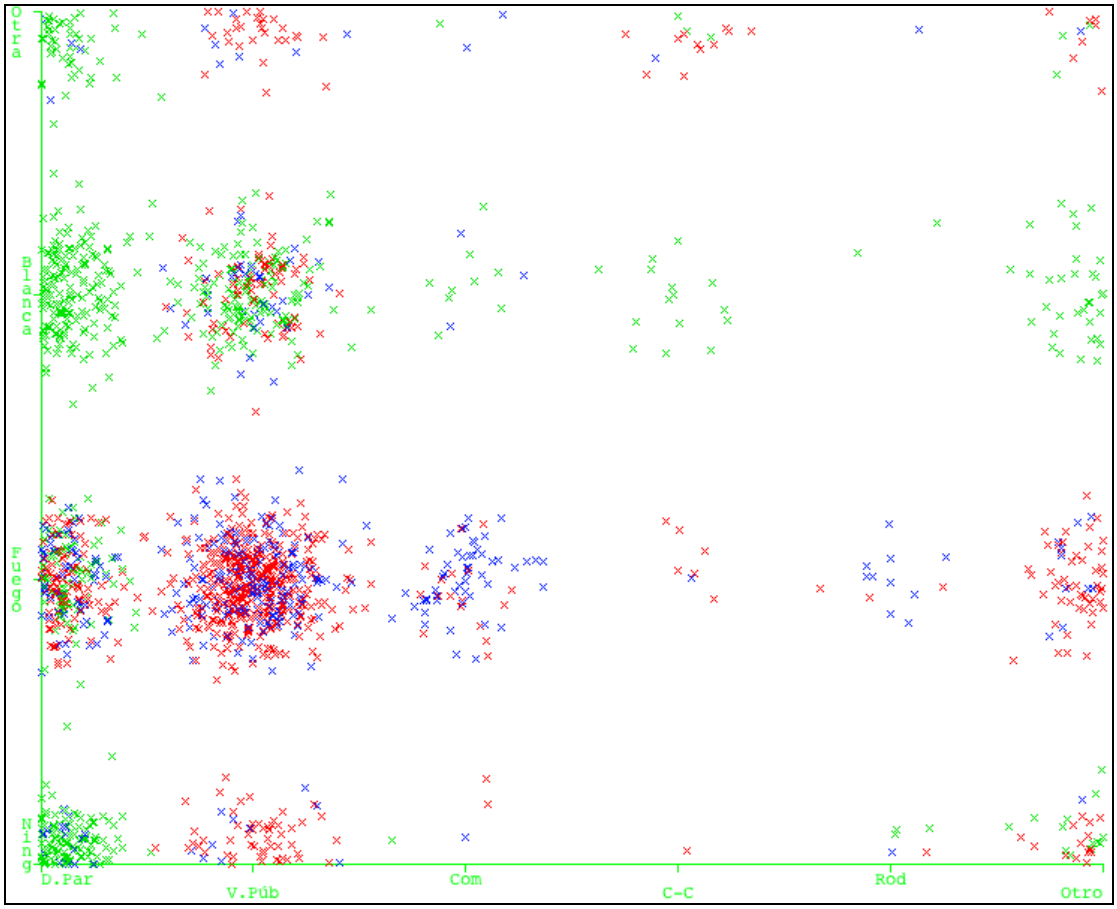


Figura 7.9: Interrelación *lugar-arma*

Existe una fuerte interacción entre *domicilio particular*, *arma blanca* y cluster 2 [Figura 7.9]. En un nivel más general podríamos interpretar al cluster 2 como homicidios en *domicilio particular* donde el arma *no es arma de fuego*.

También se observa interacción entre *vía pública*, *arma de fuego* y cluster 1, aunque con cierto solapamiento con el cluster 0.

7.2.4.6. Interrelación lugar-otro delito

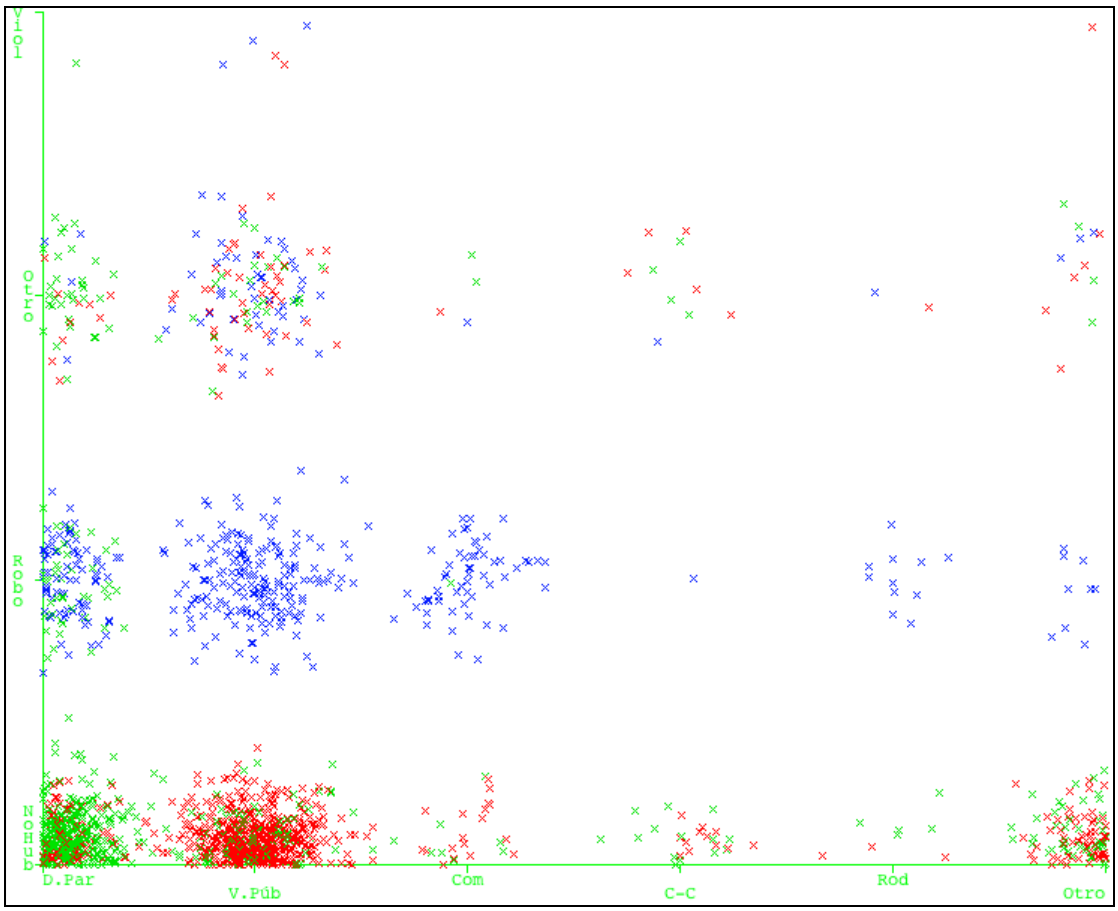


Figura 7.10: Interrelación *lugar-otro delito*

Se observa interacción entre *domicilio particular*, *no hubo otro delito* y cluster 2 [Figura 7.10].

Asimismo existe otra fuerte interacción entre *vía pública*, *no hubo otro delito* y cluster 1.

Finalmente existe cierta asociación entre *robo* y el cluster 0, con un leve ruido por parte del cluster 2 en los casos ocurridos en *domicilio particular*.

7.2.4.7. Interrelación arma-otro delito

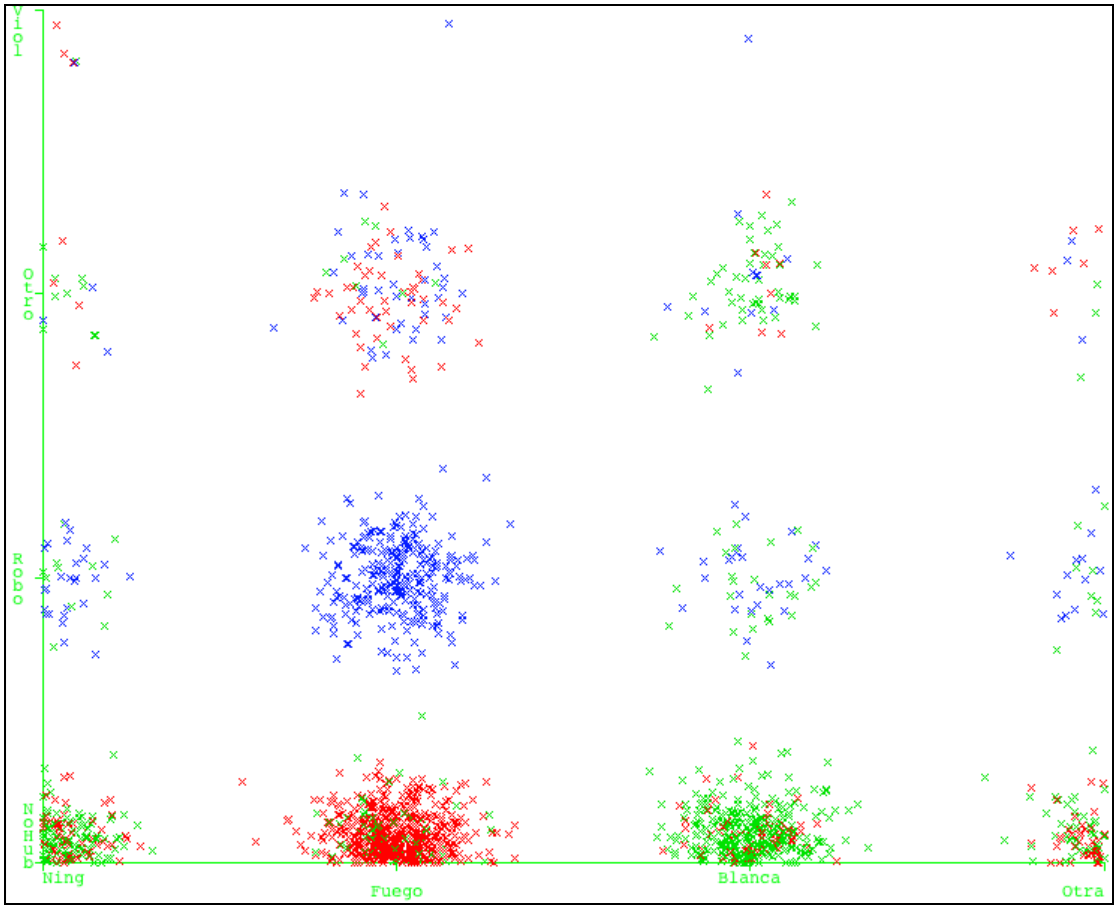


Figura 7.11: Interrelación *arma-otro delito*

Existe interacción entre *arma blanca*, *no hubo otro delito* y cluster 2 [Figura 7.11]. En un nivel general en este cluster se observa prácticamente la ausencia de *arma de fuego* comentada en el punto 6.2.4.2, sumada a una ausencia de *otro delito*.

Otra fuerte interacción se da entre *arma de fuego*, *no hubo otro delito* y cluster 1.

Finalmente hay otra fuerte interacción entre *arma de fuego*, *en ocasión de robo* y cluster 0. Esto confirma la tendencia observada en los puntos 7.2.4.2 y 7.2.4.3, aunque también se destacan algunos casos de *otro delito* con *arma de fuego* y de *robo* con *arma blanca*.

7.2.5. Primera interpretación

En base a la información que surge de este análisis podemos dar una primera interpretación a los clusters:

- **Cluster 0 (22%):** esta caracterizado por homicidios mayoritariamente en ocasión de robo y con arma de fuego. En principio diremos que se trata de *"homicidios en ocasión de robo"*.

- **Cluster 1 (43%):** es el que más registros agrupa y el más parecido a la media global. Está caracterizado por homicidios mayoritariamente en la vía pública con arma de fuego y sin la existencia de otro delito. Se podrían interpretar como *"homicidios en ocasión de riña o ajuste de cuentas"*.

- **Cluster 2 (35%):** es el más particular de los clusters, ya que la mayoría de sus registros presentan casos de homicidios sin arma de fuego y en domicilio particular. Los denominaremos *"homicidios en ocasión de emoción violenta"*.

7.3. APLICACIÓN DE C4.5 PARA LA CLASIFICACIÓN DE LOS CLUSTERS

7.3.1. Selección de atributos

Weka cuenta con un set de métodos para preseleccionar los atributos que serán utilizados posteriormente en algoritmos TDIDT. Algunos actúan a priori 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 7.3 se muestran los resultados obtenidos (en los casos de ranking se muestra entre paréntesis el puntaje obtenido para los principales atributos).

		Atributos							
Evaluator	**Search Method**	**Prov.**	**Lugar**	**Arma**	**Otro Delito**	**Hora**	**Día Semana**	**Día Mes**	**Mes**
Wrapper	Exahustive		X	X	X	X	X	X	
Classifier	Exahustive		X	X	X	X	X	X	X
Consistency	Exahustive	X	X	X	X	X	X		
Cfs	Exahustive		X	X	X	X	X		
ChiSquare	Ranker	5 (137)	3 (572)	2 (737)	1 (1353)	4 (140)			
Gain Ratio	Ranker	5 (0,019)	3 (0,131)	2 (0,195)	1 (0,479)	4 (0,035)			

Tabla 7.3: Resultado de selección de atributos

Como se puede observar, todos los métodos coinciden en tomar al menos los atributos *lugar, arma, otro delito* y *hora*.

7.3.2. Comparación de resultados de C4.5 con distintos atributos

Se corrió el algoritmo C4.5 (J48 en la terminología de *Weka*) con distintos atributos preseleccionados. En la Tabla 7.4 se observan el porcentaje de registros correctamente clasificados con C4.5 en base a las distintas combinaciones de atributos.

Si bien la combinación óptima es la B, logrando aún un mejor resultado que con la combinación de todos los atributos (combinación A), no se pierde mucho poder de clasificación al descartar el atributo *día mes* (combinación C) pero si bastante más al descartar el atributo *día de la semana* (combinación D). Pese a esto sigue manteniendo un buen poder clasificatorio con pocos atributos. Por último, si dejamos los 3 atributos

principales según los métodos de ranking (combinación E) el poder clasificatorio decae abruptamente.

	Atributos								
Combinación	**Prov.**	**Lugar**	**Arma**	**Otro Delito**	**Hora**	**Día Semana**	**Día Mes**	**Mes**	**Resultado**
A	X	X	X	X	X	X	X	X	98,7%
B		X	X	X	X	X	X		98,8%
C		X	X	X	X	X			98,1%
D		X	X	X	X				95,5%
E		X	X	X					86,2%

Tabla 7.4: Combinaciones de atributos

Las matrices de confusión para estas combinaciones [Tabla 7.5] muestran que, a excepción de la última combinación, en todos los casos se conserva un muy buen poder clasificatorio de todos los clusters en forma homogénea.

A

		Clasificado como 0	1	2	TP Rate
Real	**0**	396	5	1	**99%**
	1	2	769	1	**100%**
	2	4	10	622	**98%**

D

		Clasificado como 0	1	2	TP Rate
Real	**0**	384	5	13	**96%**
	1	12	743	17	**96%**
	2	11	23	602	**95%**

B

		Clasificado como 0	1	2	TP Rate
Real	**0**	399	2	1	**99%**
	1	2	768	2	**99%**
	2	4	11	621	**98%**

E

		Clasificado como 0	1	2	TP Rate
Real	**0**	342	46	14	**85%**
	1	4	692	76	**90%**
	2	13	97	526	**83%**

C

		Clasificado como 0	1	2	TP Rate
Real	**0**	395	5	2	**98%**
	1	13	758	1	**98%**
	2	4	9	623	**98%**

Tabla 7.5: Matrices de confusión para las principales combinaciones

7.3.3. Árbol definitivo

Como se demostró en la sección anterior, la combinación óptima es la B, que involucra los atributos: *lugar, arma, otro delito, hora, día de la semana* y *día del mes*. El excelente poder clasificatorio de esta combinación (98,8% de instancias bien clasificadas) confirma que la los *clusters* determinados por *K-means* responden a un criterio determinado subyacente a los datos.

Por cuestiones dimensionales, el árbol generado con el algoritmo *C4.5* no puede ser presentado en una única hoja. La estructura global del mismo es la siguiente [Figura 7.12]:

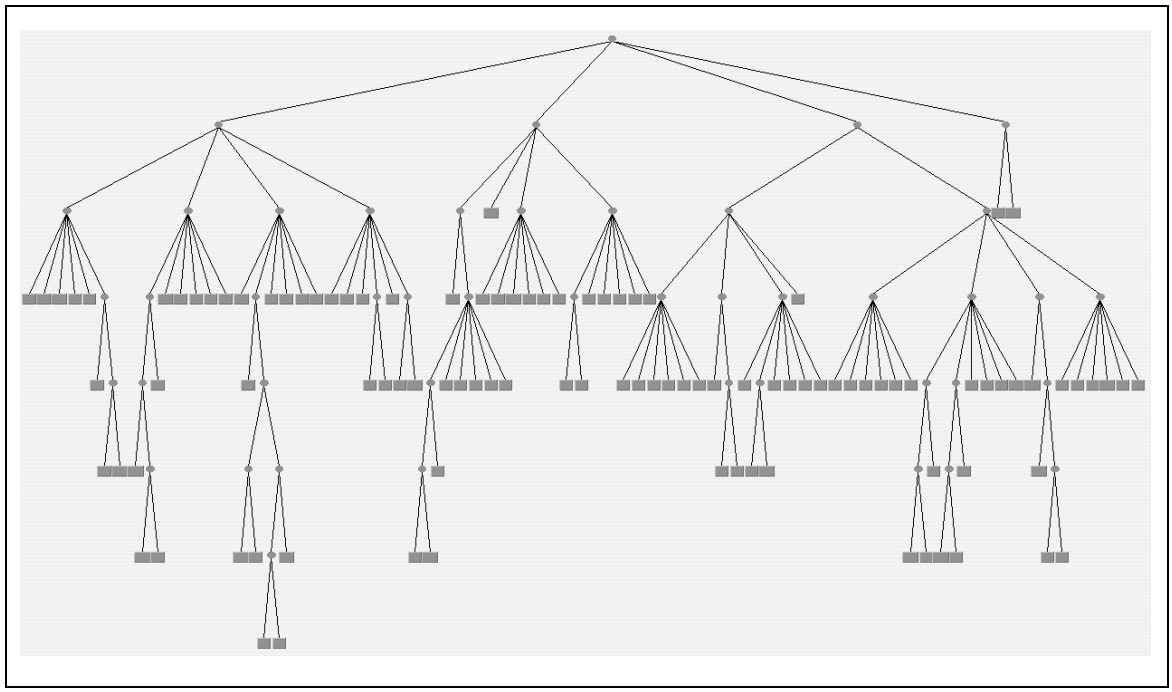


Figura 7.12: Estructura del árbol generado con C4.5

El árbol presenta 103 hojas. El nodo principal es el atributo *otro delito*. Las ramas correspondientes a *no hubo otro delito* y *robo* se dirigen a nodos referidos al atributo *arma* [Figura 7.13]. Por otro lado, las ramas correspondientes a *otro delito* y *violación* se dirigen a nodos referidos al atributo *hora*. En el caso de *violación*, del nodo *hora* surge la rama *entre las 8 y las 20 hs* en cuyo caso corresponde al cluster 1 (3 instancias clasificadas correctamente), y la rama complementaria, *entre las 20 y las 8hs*, en cuyo caso corresponde al cluster 0 (4 instancias clasificadas, pero una de ellas incorrectamente). El resto del árbol se puede ver en el Anexo 4.

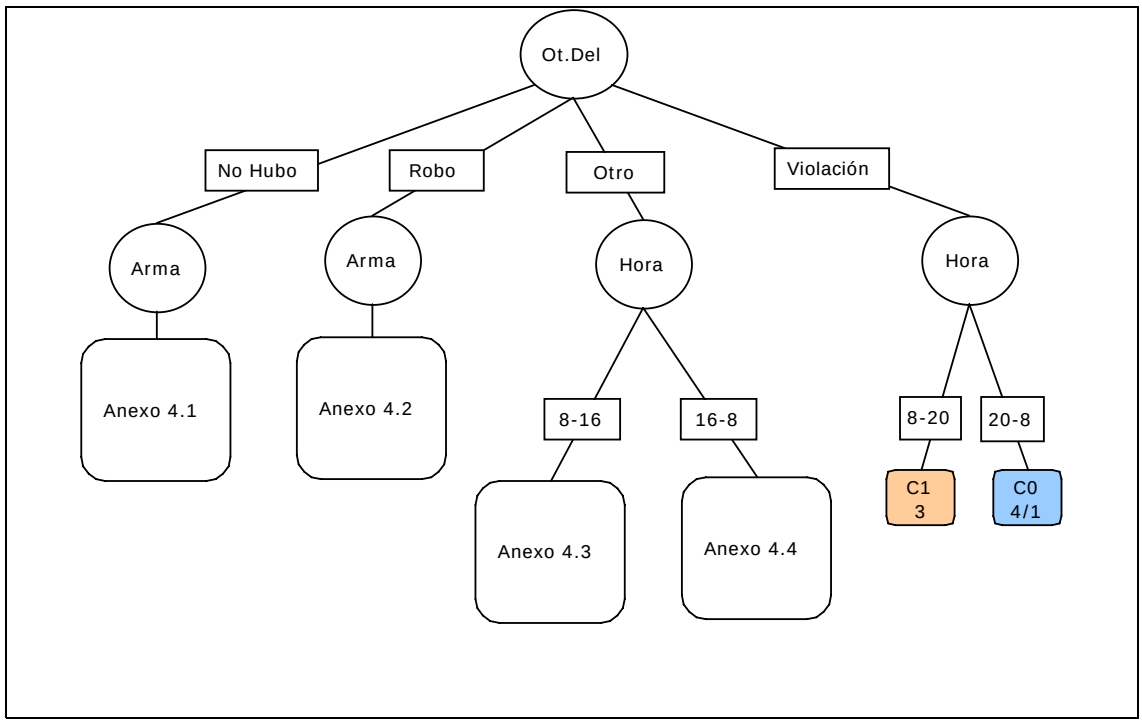


Figura 7.13: Referencias del árbol generado con C4.5

El 9% de las reglas de clasificación extraídas del árbol clasifican el 66% de las instancias (1200). Estas reglas son las siguientes:

Regla 1

SI otro delito = no hubo
Y arma = fuego
Y lugar = V.Púb.
ENTONCES Cluster 1 (362)

Regla 2

SI otro delito = robo
Y arma = fuego
ENTONCES Cluster 0 (272)

Regla 3

SI otro delito = no hubo
Y arma = blanca
Y lugar = D.Part.
ENTONCES Cluster 2 (134)

Regla 4

SI otro delito = no hubo
Y arma = blanca
Y lugar = V.Pub.
Y DSem = Sa-Ma
Y hora = 19-8
ENTONCES Cluster 2 (109)

Regla 5

SI otro delito = no hubo
Y arma = ninguna
Y lugar = D.Part.
ENTONCES Cluster 2 (87)

Regla 6

SI otro delito = no hubo
Y arma = fuego
Y lugar = D.Part.
Y hora = 8-19
ENTONCES Cluster 1 (85/3)

Regla 7

SI otro delito = no hubo
Y arma = blanca
Y lugar = V.Pub.
Y hora = 8-16
ENTONCES Cluster 1 (55)

Regla 8

SI otro delito = no hubo
Y arma = ninguna
Y lugar = V.Pub.
ENTONCES Cluster 1 (48)

Regla 9

SI otro delito = no hubo
Y arma = fuego
Y lugar D.Part.
Y hora = 20-8
ENTONCES Cluster 1 (48)

Estas reglas fueron consultadas con los especialistas y permitieron confirmar la interpretación hecha anteriormente. Al respecto, los especialistas comentaron que hasta el momento ellos solían clasificar a los homicidios 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 esta representado por el cluster 0, mientras que el segundo por el cluster 1 y 2.

Sin embargo, las diferencias entre el cluster 1 y 2 indican cierta relación entre el lugar y el arma de los *"homicidios en conflictos interpersonales"* (1230 casos en total). Si el arma es de fuego (560 casos), entonces es más probable que sea en la vía pública (330 casos vs 157 en domicilio particular). Sin embargo, si no hay arma (167 casos), entonces es más probable que sea en domicilio particular (87 casos vs 45 en la vía pública). Si el arma es blanca (375 casos) entonces ya no existe gran diferencia (191 casos en la vía pública vs. 134 en domicilio particular).

La conclusión arribada junto con los especialistas es que se trata de **dos tipos de conflictos interpersonales distintos**, uno más bien **familiar** (dentro del domicilio particular) e **impulsivo** (sin arma) y otro más bien **vecinal** o de **ajuste de cuentas** (vía pública) y con cierto nivel de **premeditación o pre-intencionalidad** (arma de fuego). En el medio de estos dos grupos extremos están los casos de armas blancas, difíciles de asignar a priori a una u otra modalidad.

8. CONCLUSIONES

8.1 CONCLUSIONES GENERALES

El presente trabajo ha demostrado no sólo que es factible aplicar minería de datos a la información criminal en Argentina, sino también su alto valor agregado para el análisis y la generación de nuevo conocimiento.

La experiencia realizada en conjunto con la Dirección Nacional de Política Criminal (DNPC) del Ministerio de Justicia y Derechos Humanos de la Nación (MJDHN), basa la factibilidad en los siguientes puntos:

- existe gran cantidad de información que actualmente no esta siendo aprovechada en toda su dimensión;
- existe un software de minería de datos de distribución libre y gratuita, fácil de usar y que contiene las herramientas necesarias para el análisis;
- este software de minería de datos puede ser utilizado por una persona ajena al ámbito informático con una capacitación básica;
- los recursos humanos y tecnológicos necesarios son mínimos y están a disposición de la DNPC.

Los resultados experimentales obtenidos han sido validados por los especialistas de la DNPC. Estos resultados han permitido tanto confirmar conceptos preexistentes (pero con una justificación sustentada en los datos), como generar nuevas piezas de conocimiento. Al respecto se han identificado tres patrones distintos de homicidios dolosos en base a los hechos ocurridos en Argentina durante 2005.

A su vez el presente trabajo ha contribuido a la comunidad científica, participando en el *IX Workshop de Investigadores en Ciencias de la Computación (WICC)* realizado en la ciudad de Trelew, Chubut, Argentina, el 3 y 4 de mayo de 2007.

8.2 FUTURAS LÍNEAS DE INVESTIGACIÓN

En primer lugar se propone aumentar el alcance de la información de la DNPC a ser analizada con este tipo de técnicas. Esto implica tanto una expansión transversal, haciendo uso de otras bases de datos como la de “homicidios culposos en accidentes de tránsito”; como longitudinal, analizando la información histórica existente para detectar patrones de evolución temporal en cuanto a las modalidades delictivas.

En segundo lugar se sugiere el diseño de procedimientos estándar de minería de datos con *Weka* para ser implementados en la DNPC. Esta batería de procedimientos les permitiría a los analistas de la DNPC extraer e identificar patrones y asociaciones en forma automatizada y estandarizada.

En tercer lugar se propone proceder al análisis de la información geográfica relevada por la DNPC (que hoy no es aprovechada) mediante GISs (*Geographical Information Systems*). Este tipo de análisis permitiría detectar, por ejemplo, zonas de alta densidad de homicidios en accidentes de tránsito.

Finalmente se propone expandir el uso de estas técnicas a las fuerzas de seguridad, en donde estas aplicaciones han encontrado su mayor aplicación a nivel mundial.

8.2 REFLEXIONES

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 la alocación de los recursos policiales y detectar grupos delictivos organizados. Si bien esta disciplina ha tenido mucho auge en los últimos tiempos, en especial a partir de los atentados del 11 de septiembre, aún se encuentra en desarrollo.

En este sentido la Argentina presenta la oportunidad única, no sólo de formar parte de este cambio, sino también de participar activamente en el proceso de desarrollo de nuevas tecnologías asociadas. Para ello es necesario superar ciertas barreras.

En primer lugar es necesario sistematizar los procesos de información que surgen a partir de una denuncia.

En la actualidad las denuncias recibidas por las fuerzas de seguridad no se registran en ningún sistema informático. Las estadísticas nacionales de criminalidad se basan en planillas elaboradas por cada dependencia policial en donde se contabiliza la cantidad de denuncias realizadas según cada tipo de delito. La información contenida en estas planillas, al tratarse de sumarizaciones y no de hechos puntuales, es prácticamente incontrastable con cualquier otra fuente de información. Por lo tanto la validez y veracidad de estas planillas queda a merced de las autoridades policiales de cada dependencia. A su vez la recolección y consolidación de esta información por parte de la DNPC es bastante engorrosa.

Frente a esta situación resulta indispensable la creación de un sistema interconectado nacional, que vincule a las fuerzas de seguridad y a las autoridades judiciales. En donde cada denuncia sea ingresada *online* a partir de un formulario digital estándar con un código de identificación único que se mantenga durante el resto del proceso penal. Esto permitiría conformar un único *data warehouse* a nivel nacional en base al cual todas las instituciones policiales y judiciales del país puedan realizar consultas según un sistema de perfiles de usuario. El Proyecto SURC demostró que existen los recursos para establecer un sistema de estas características.

En segundo lugar, es necesario que la dirigencia política modifique su postura frente a la información criminal.

A partir de la crisis de 2001, la “inseguridad” pasó a ser uno de los principales temas que preocupan a la gente según las encuestas de opinión popular. En virtud de esto, la publicación de las estadísticas criminales tomó carácter político, al mismo tiempo que fueron relegadas. Es importante que la postura oficial sea concebir a la información criminal no solamente como “termómetro” de la inseguridad, sino también como herramienta fundamental para la toma de decisiones. Cuanto mejor sea la calidad de la información, más acertadas serán las decisiones y más efectivamente se podrá reducir el nivel delictivo (“*no se puede mejorar lo que no se mide: basura entra, basura sale*”). Este cambio de postura implica la voluntad política de desarrollar los sistemas de información criminal a nivel nacional, en principio no para publicar mejores estadísticas, sino para obtener mejores resultados.

Finalmente es importante destacar que la utilización de minería de datos para reducir el delito no es la panacea, sino que debe ser considerada como una herramienta poderosa en la medida que sea acompañada de políticas de estado de largo alcance y efectivas que actúen en consecuencia.

9. REFERENCIAS

Ale, J., 2005a. *Análisis de Clusters*.

Ale, J., 2005b. *Introducción a Data Mining*.

Behar, A. M., P. Lucilli, 2003. *Mapa del delito de la Ciudad Autónoma de Buenos Aires*. Terceras Jornadas de Jóvenes Investigadores, Instituto Gino Germani.

Blackwelder, J.K., L.L. Jonson, 1984. *Estadística Criminal y Acción Policial en Buenos Aires, 1887-1914*. Desarrollo Económico, 93, Vol. 24, 1984, pp. 109-122.

Blurock, E., 1996. *The ID3 Algorith*. Research Institute for Symbolic Computation.

Briceño-León R., 2001. *Violencia, sociedad y justicia en América Latina*.

Cardona M. *et al*., 2005. *Homicidios en Medellín, Colombia, entre 1990 y 2002: actores móviles y circunstancias.* Cadernos de Saúde Pública 21(3). Páginas: 840-851.

Coplink, 2004. *Crime Data Mining and Visualization for Intelligence and Security Informatics: The COPLINK Research*. University of Arizona Artificial Intelligence Lab.
URL: http://ai.bpa.arizona.edu/research/coplink/index.htm. Acceso mayo 2007.

Coplink, 2007. COPLINK Solution Suite.
URL: http://www.coplink.com. Acceso mayo 2007.

CrimeStat, 2007. CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incident Locations.
URL: http://www.icpsr.umich.edu/NACJD/crimestat.html. Acceso mayo 2007.

Chau, M., J. J. Xu, H. Chen, 2002. *Extracting Meaningful Entities from Police Narrative Reports*.
URL: http://ai.elller.arizona.edu/go/intranet/Publication/COPLINKEE.pdf.Acceso mayo 2007.

Cheeseman, P., J. Stutz, 1996. *Bayesian classification (AutoClass): Theory and results*. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, Advances in Knowledge Discovery and Data Mining. Páginas 153-180. AAAI/MIT Press, California.

Chen, H., W. Chung, J. Xu, G. Wang, Y. Qin, M. Chau, 2004. *Crime Data Mining: A General Framework and Some Examples*. IEEE Computer Society, vol. 37, no. 4. Páginas 50-56.

Chen, M., J. Han, 1996. *Data mining: An overview from database perspective*. IEEE Transactions on Knowledge and Data Eng.

Desarme.org, 2002. Qué son las armas pequeñas?
URL: http://www.desarme.org. Acceso mayo 2007.

DNPC, 2002. *Investigación sobre homicidios dolosos en la Ciudad de Buenos Aires, año 2002*. Dirección Nacional de Política Criminal.
URL: http://www.policrim.jus.gov.ar. Acceso mayo 2007.

DNPC, 2007. *Sistema Nacional de Estadísticas sobre Ejecución de la Pena (SNEEP)*. Dirección Nacional de Política Criminal.
URL: http://www.policrim.jus.gov.ar. Acceso mayo 2007.

EAP, 2004. Encuesta de armas pequeñas: derechos en riesgo.
URL: http://samllarmssurvey.org. Acceso mayo 2007.

Eck, J. E., S. Chainey, J. G. Cameron, M. Leitner, R. E. Wilson, 2005. *Mapping Crime: Understanding Hot Spots*. Crime Mapping Research Center U.S. Department of Justice. Office of Justice Programs.

Elder IV, J., D. Pregibon, 1996. *A statistical perspective on knowledge discovery in databases*. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, Advances in Knowledge Discovery and Data Mining. Páginas 83-115. AAAI/MIT Press, California, EEUU.

ESRI, 2007. Enviromental Systems Research Institute, Inc.
URL: http://www.esri.com/software/arcview. Acceso mayo 2007.

Essenreiter R., M. Karrenbach, S. Treitel, 1999. *Identification and classification of multiple reflections with selforganizing maps*. Geophysical Institute, University of Karlsruhe, Germany.

Ester, M., H. P. Kriegel, X. Xu, 1995. *Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification*. In Proc. 4th Int. Symp. on Large Spatial Databases (SSD'95), Portland, Maine, EEUU. Páginas 67-82.

Evangelos, S., J. Han, 1996. *Proceeding of the Second International Conference on Knowledge Discovery and Data Mining*. Portland, EEUU.

Fisher, D., 1996. *Iterative optimization and simplification of hierarchical clusterings*. Departament of Computer Science. Vanderbilt University, Nashville, EEUU.

Gallion, R., D. St Clair, C. Sabharwal, W. Bond, 1993. *Dynamic ID3: A Symbolic Learning Algorithm for Many-Valued Attribute Domains*. Engineering Education Center, University of Missouri-Rolla, St. Luis, EEUU.

Han, J., M. Kamber, 2001. *Data mining: Concepts and techniques*. Morgan Kauffmann Publishers.

Hand, D. J., 1997. *Data Mining: Statistics and More?*. The American Statistician.

Hand, D., H. Mannila, P. Smyth, 2001. *Principles of data mining*. The MIT Press, California, EEUU.

Hernández Orallo, J., 2000. *Extracción Automática de Conocimiento de base de datos e ingeniería del software.* Programación declarativa e ingeniería de la programación.

i2, 2007. i2 Analyst´s Workstation.
URL:http://www.i2.co.uk/Products/Analysts_Workstation/default.asp. Acceso mayo 2007.

IB, 2007. Information Builders.
URL: http://www.informationbuilders.com. Acceso mayo 2007.

IBM, 2007. Internacional Business Machines.
URL:http://www03.ibm.com/industries/goverment/doc/content/news/pressrelease/1019264109.html. Acceso mayo 2007.

ICJIA, 2007. Illinois Criminal Justice Information Authority.
URL:http://www.icjia.state.il.us/public/index.cfm?metasection=Data&metapage=StacFacts. Acceso mayo 2007.

Jain, A. K., R. C. Dubes, 1988. *Algorithms for Clustering Data*. Prentice Hall.

Kantardzic, M. 2002. *Data Mining: Concepts, models, methods and algorithms.* Wiley-IEEE Press. ISBN 0-471-22852-4.

Kaufman, L., P. J. Rousseeuw, 1990. *Finding Groups in Data: an Introduction to Cluster Analysis*. Wiley-Interscience.

Kohonen, T., 1995. *Self-Organizing Maps*. Springer-Verlag.

La Nación, 2006. *El Gobierno canjeará armas por dinero*. Diario La Nación, 9 agosto de 2006.
URL:http://www.lanacion.com.ar/Archivo/nota.asp?nota_id=830078. Acceso mayo 2007.

Lu H., R. Setiono, H. Liu, 1995. *NeuroRule: A Connectionist Approach to Data Mining*. Proceedings of the 21th International Conference on Very Large Data Bases. Páginas 478/489.

Mannila, H., 1997. *Methods and problems in data mining.* In Proc. of International Conference on Database Theory, Delphi, Greece.

Mapinfo, 2007. MapInfo Corporation.
URL:http://www.mapinfo.com/location/integration. Acceso mayo 2007.

Michalski R., 1983. *A Theory and Methodology of Inductive Learning*. Morgan-Kauffman, EEUU.

Michalski R., A. Baskin, K. Spackman, 1982. *A Logic-Based Approach to Conceptual Database Analysis*. Sixth Annual Symposium on Computer Applications on Medical Care, George Washington University, Medical Center, Washington, DC, EE.UU.

Michalski R., I. Bratko, M. Kubat, 1998. *Machine Learning and data mining: Methods and Applications*. Wiley & Sons Ltd., EE.UU.

Morales, E., 2003. *Descubrimiento de Conocimiento en Bases de Datos*.

Ng, R., J. Han, 1994. *Efficient and effective clustering method for spatial data mining*. In Proc. 1994 Int. Conf. Very Large Data Bases, (Páginas 144/155), Santiago de Chile, Chile.

NYC, 2007. New York Police Department Real Time Crime Center. URL:http://www.nyc.gov/html/nypd/html/dcpi/RTCCRevisedFINALWEB_files/frame.htm. Acceso mayo 2007.

Oatley, G.C., B.W. Ewart, J. Zeleznikow, 2004. *Decisión Support Systems for Police: Lessons from the application of Data Mining Techniques to "Soft" forensic Evidence*. URL:http://www.aic.gov.au/conferences/occasional/2005-04.zeleznikow.html. Acceso mayo 2007.

Ochoa, M. A. 2004. *Herramientas Inteligentes para la Explotación de Información*. Trabajo Final: Especialidad en Ingeniería en Sistemas Expertos, Instituto Tecnológico de Buenos Aires (ITBA).

OPS/OMS, 2003. *Informe Mundial sobre la violencia y la salud*. Organización Panamericana de la Salud. Publicación científica y técnica Nro. 588. Washington DC, EEUU.

Peres, M.F.T., P.C. Santos, 2005. *Mortalidade por homidios no Brasil na década de 90: o papel das armas de fogo*. Revista de Saúde Pública 39(1). Páginas 58-66.

Perichinsky, G. y R. Garcia Martinez, 2000. *A Data Mining Approach to Computational Taxonomy*. Proceedings del Workshop de Investigadores en Ciencias de la Computación. Páginas 107-110. Editado por Departamento de Publicaciones de la Facultad de Informática. Universidad Nacional de La Plata, Buenos Aires, Argentina.

Perichinsky,G., R. García-Martínez, A. Proto, 2000. *Knowledge Discovery Based on Computational Taxonomy And Intelligent Data Mining*. CD del VI Congreso Argentino de Ciencias de la Computación. Ushuaia, Argentina.

Perichinsky, G., R. Garcia-Martínez, A. Proto, A. Sevetto, D. Grossi, 2001. *Integrated Environment of Systems Automated Engineering*. Proceedings del II Workshop de

Investigadores en Ciencias de la Computación. Editado por Universidad Nacional de San Luis en el CD Wicc 2001.

Perichinsky, G., M. Servente, A. Servetto, R. García-Martínez, R. Orellana, A. Plastino, 2003. *Taxonomic Evidence and Robustness of the Classification Appying Intelligent Data Mining*. Proceedings del VIII Congreso Argentino de Ciencias de la Computación. Pág. 1797-1808.

Piatetski-Shapiro, G., U. Fayyad, P. Smith, 1996. *From data mining to Knowledge discovery*. AAAI Press/MIT Press, California, EEUU.

Piatetski-Shapiro, G.; W. Frawley, C. Matheus, 1991. *Knowledge discovery in databases: an overview*. AAAI-MIP Press, Menlo Park, California, EEUU.

Quinlan J., 1993a. *Learning Efficient Clasifications Procedures and Their Application to Chess Games*. En R. Michalski, J. Carbonell & T. Mitchells (Eds.) Machine Learning, The Artificial Intelligence Approach. Morgan Kaufmann, Vol. II, Capítulo 15, páginas 463 a 482. EE.UU.

Quinlan J., 1993b. *The effect of Noise on Concept Learning*. En R. Michalski, J. Carbonell & T. Mitchells (Eds.) Machine Learning, The Artificial Intelligence Approach. Morgan Kaufmann, Vol. I, Capítulo 6, páginas 149 a 167. San Mateo, California, EE.UU.

Quinlan, J., 1993c. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

RiverGlass, 2007. RiverGlass Inc.
URL:http://www.riverglassinc.com/solutions/intelligence.php. Acceso mayo 2007.

RTI, 2007. Research Triangle Institute.
URL:http://www.rti.org. Acceso mayo 2007.

Rubial B.C., 1993. *Ideología del Control Social, 1880-1920*. Centro Editor de América Latina, Buenos Aires, Argentina.

Sain M, 2003. *Recuperación y destrucción de armas en circulación ilícita*. Proyecto Construyendo Seguridad Ciudadana del Programa de Apoyo a la Reforma Estatal y Fortalecimiento Fiscal. Informe de Avance 1 y 2.

Sentient, 2007. Sentient Information Systems.
URL:http://www.sentient.nl. Acceso mayo 2007.

Servente, M., R. García-Martínez, 2002. *Tesis Doctoral Algoritmos TDIDT aplicados a la minería de datos inteligente*. Universidad de Buenos Aires, Argentina.

Sozzo, M., 2000. *Pintando a Través de Números: Fuentes Estadísticas de Conocimiento y Gobierno Democrático de la Cuestión Criminal en Argentina*.

URL:http://www.ilsed.org/index.php?option=com_docman&task=doc_view&gid=159&itemid=44. Acceso mayo 2007.

Spinelli H., M. Alazraqui, M.G. Zunino, H. Olaeta, H. Poggiese, C. Concaro, S. Porterie, 2006. *Muertes y delitos por armas de fuego en la Ciudad Autónoma de Buenos Aires, año 2002*. Dirección Nacional de Política Criminal. URL:http://wwwpolcrim.jus.gov.ar/Otras_Invest/Armas/Muertes%20y%20Armas%20de%20Fuego_WEB.pdf. Acceso Mayo 2007.

SPSS, 2007. SPSS Inc.. URL:http://www.spss.com/success/pdf/CS%20%20Richmond%20PD%20LR.pdf. Acceso mayo 2007.

SSI-MI, 2004. *Presentación Institucional Proyecto SURC*. Secretaría de Seguridad Interior, Ministerio del Interior de la República Argentina.

Szwarewald, C.L., E.A. Castillo, 1998. *Mortalidade por armas de fogo o estado do Rio de Janeiro, Brasil: uma análise espacial*. Rev Panam Salud/ Pan Am J Public Health 4(3).

Vesanto J., E. Alhoniemi, 2000. *Clustering of the Self-Organizing Map*. IEEE transactions on neural networks, Vol 11, No. 3.

Wang, G.A., H. Atabakhsh, T.Petersen, H.Chen, 2004. *Discovering Identity Problems: a Case Study*. URL:http://ai.elller.arizona.edu/COPLINK/publications/Identity_ISI05_final.pdf. Acceso mayo 2007.

Weka, 2007. University of Waikato. URL:http://www.cs.waikato.ac.nz/~ml/weka/index.html. Acceso mayo 2007.

Zeleznikow, J., 2005. *Using Data Mining to Detect Criminal Networks*. URL:http://www.aic.gov.au/conferences/occasional/2005-04.zeleznikow.html. Acceso mayo 2007.

Zhang, T., R. Ramakrishnan, M. Livny, 1996. *BIRCH: an efficient data clustering method for very large databases*. In Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data, Montreal, Canadá.

10. TABLA DE ACRÓNIMOS

AMBA: Área Metropolitana de Buenos Aires

CABA: Ciudad Autónoma de Buenos Aires

CAPIS: Centro de Ingeniería de Software e Ingeniería del Conocimiento

CIA: Central Intelligence Agency

CIM: Centro de Información Metropolitana

DNPC: Dirección Nacional de Política Criminal

FADU: Facultad de Arquitectura, Diseño y Urbanismo

FBI: Federal Bureau of Investigation

GIS: Geographical Information Systems

ITBA: Instituto Tecnológico de Buenos Aires

MJDHN: Ministerio de Justicia y Derechos Humanos de la Nación

MPFN: Ministerio Público Fiscal de la Nación

OMS: Organización Mundial de la Salud

RNREC: Registro Nacional de Reincidencia y Estadísticas Criminales

SAT: Sistema de Alerta Temprana

SNEEP: Sistema Nacional de Estadísticas sobre Ejecución de la Pena

SNEJ: Sistema Nacional de Estadísticas Judiciales

SNIC: Sistema Nacional de Información Criminal

SOM: Self Organized Maps

SURC: Sistema Unificado de Registros Criminales

UBA: Universidad de Buenos Aires

11. ANEXOS

ANEXO 1 - LEY NACIONAL 25.266

ARTICULO 1º — Sustitúyese la denominación "Registro Nacional de Reincidencia y Estadística Criminal", Ley Nº 22.117, por la de "Registro Nacional de Reincidencia".

ARTICULO 2º — Sustitúyese el artículo 13 de la Ley Nº 22.117 por el siguiente:
Artículo 13. — Todos los tribunales del país con competencia en materia penal, así como los representantes del Ministerio Público ante los tribunales con competencia en materia penal de todo el país, la Policía Federal Argentina, las policías provinciales, las demás fuerzas de seguridad y los servicios penitenciarios y, en su caso, las Fuerzas Armadas de la Nación, remitirán a la Dirección Nacional de Política Criminal del Ministerio de Justicia y Derechos Humanos de la Nación los datos que esta dependencia les requiera con el fin de confeccionar anualmente la estadística general sobre la criminalidad en el país y el funcionamiento de la justicia.
El requerimiento de datos será realizado trimestralmente por resolución fundada del director del organismo. Los datos requeridos, que no serán personales en caso alguno, sólo podrán ser utilizados con fines estadísticos - criminales.
El requerimiento deberá ser preciso procurando que no obstaculice la tarea cotidiana del personal de los organismos requeridos. A tal efecto, el requerimiento podrá estar acompañado de planillas de recolección de datos con una indicación precisa del mecanismo a utilizar para ser completadas.
Quienes por esta ley resulten obligados a suministrar información estadística deberán disponer lo necesario para que, eventualmente y con el único fin de verificar la exactitud de los datos brindados, la Dirección Nacional de Política Criminal pueda acceder a los registros pertinentes.
Sobre esta base, y la información que le suministre el Registro Nacional de Reincidencia, la Dirección Nacional de Política Criminal del Ministerio de Justicia y Derechos Humanos de la Nación, confeccionará anualmente la estadística general sobre la criminalidad en el país y el funcionamiento de la justicia, única que será considerada estadística criminal oficial de la Nación.

ARTICULO 3º — Incorpórase como artículo 13 bis de la Ley Nº 22.117 el siguiente:
Artículo 13 bis. — Será reprimido con multa de uno a tres de sus sueldos el funcionario público que, en violación al deber de informar establecido en el artículo precedente, no proporcione la información estadística requerida o lo haga de modo inexacto, incorrecto o tardío, siempre que no cumpla correctamente con el deber de informar dentro de los cinco días de haber sido interpelado de la falta por la Dirección Nacional de Política Criminal a través de cualquier forma documentada de comunicación.

ARTICULO 4º — Comuníquese al Poder Ejecutivo.
DADA EN LA SALA DE SESIONES DEL CONGRESO ARGENTINO, EN BUENOS AIRES, A LOS VEINTIDOS DIAS DEL MES DE JUNIO DEL AÑO DOS MIL.

REGISTRADA BAJO EL Nº 25.266
RAFAEL PASCUAL. — JOSE GENOUD. — Guillermo Aramburu. — Mario L. Pontaquarto.

ANEXO 2 - PLANILLAS DE RECOLECCIÓN DE LA DNPC

Anexo 2.1 - Planilla del Sistema Nacional de Información Criminal (SNIC)

Ministerio de Justicia, Seguridad y Derechos Humanos — S N I C — Dirección Nacional de Política Criminal

PLANILLA DE HECHOS DELICTUOSOS

Provincia:	**Mes:**
Departamento:	**Año:**
Municipalidad:	
Seccional:	

Tipo de Delito	Cantidad de hechos			Cantidad de Víctimas			
	Por denuncia particular	Por interv. policial	Total de hechos	Masculino	Femenino	No consta	Total de víctimas
1. Homicidios dolosos							
2. Homicidios dolosos en grado de tentativa							
3. Homicidios culposos en accidentes de tránsito							
4. Homicidios culposos por otros hechos							
5. Lesiones dolosas							
6. Lesiones culposas en accidentes de tránsito							
7. Lesiones culposas por otros hechos							
8. Otros delitos contra las personas							
9. Delitos contra el honor							
10. Violaciones							
11. Otros delitos contra la integridad sexual							
12. Delitos contra el estado civil							
13. Amenazas							
14. Otros delitos contra la libertad							
15. Robos (**excluye** los agravados por el resultado de lesiones y/o muertes)							
16. Tentativas de robo (**excluye** las agravadas por el res.de lesiones y/o muertes)							
17. Robos agravados por el resultado de lesiones y/o muertes							
18. Tentativas de robo agravado por el resultado de lesiones y/o muertes							
19. Hurtos							
20. Tentativas de hurto							
21. Otros delitos contra la propiedad							
22. Delitos contra la seguridad pública							
23. Delitos contra el orden público							
24. Delitos contra la seguridad de la nación							
25. Delitos contra los poderes públicos y el orden constitucional							
26. Delitos contra la administración pública							
27. Delitos contra la fe pública							
28. Ley 23.737 (estupefacientes)							
29. Otros delitos previstos en leyes especiales							
30. Figuras contravencionales							
31. Suicidios (consumados)							

Lugar y fecha de remisión

Firma y sello del responsable

IMPORTANTE: ESTE FORMULARIO DEBE SER DEVUELTO DENTRO DE LOS DIEZ PRIMEROS DIAS POSTERIORES AL MES A QUE SE REFIERE LA INFORMACIÓN. ORIGINAL PARA LA DIRECCION NACIONAL DE POLITICA CRIMINAL, SARMIENTO 329 6° PISO (1041) CIUDAD DE BUENOS AIRES.

POR CONSULTAS COMUNICARSE AL 11 4328 4674 (FAX) O AL 11 4328 3015 INT. 2659

Anexo 2.2 - Planilla del Sistema de Alerta Temprana (SAT) para homicidios dolosos

S A T

HOMICIDIOS DOLOSOS

Ministerio de Justicia de la Nación. Dirección Nacional de Política Criminal

Provincia: **Departamento:** **Año:** **Mes:**

DATOS DEL HECHO											DATOS DE LA VÍCTIMA			DATOS DEL IMPUTADO		
Nº de Sumario	Fecha del hecho	Hora del hecho	Nombre de la calle o ruta	Altura / km / entre calles	Localidad	Tipo de lugar (A)	Tipo de lugar Especificar sólo código 6	Clase de Arma (B)	Clase de Arma Especificar sólo códigos 3 ó 4	En ocasión de otro delito (C)	En ocasión de otro delito Especificar sólo código 3	Sexo (D)	Edad	Clase (E)	Sexo (D)	Edad	Clase (E)
	/ /	:															
	/ /	:															
	/ /	:															
	/ /	:															
	/ /	:															
	/ /	:															
	/ /	:															
	/ /	:															
	/ /	:															
	/ /	:															
	/ /	:															
	/ /	:															
	/ /	:															
	/ /	:															
	/ /	:															

TABLAS DE CODIFICACIÓN DE DATOS

A	B	C	D	E
Tipo de Lugar	**Clase de Arma**	**En ocasión de otro delito?**	**Sexo**	**Clase de Vict./ Imp.**
1 - Vía Pública	**1** - Arma de Fuego	**1** - Sí, ROBO	**1** - Masculino	**1** - Civil
2 - Domicilio Particular	**2** - Arma Blanca	**2** - Sí, VIOLACIÓN	**2** - Femenino	**2** - Policía en Servicio
3 - Comercio	**3** - Otra arma (Especificar CUÁL?)	**3** - Sí, Otro delito (Especificar)		**3** - Policía fuera de Servicio
4 - Interior de Rodados	**4** - Sin armas (Especificar CÓMO?)	**4** - NO fue en ocasión de otro delito		**4** - Seguridad Privada
5 - Cárcel o Comisaría				**5** - Otra fuerza de seguridad
6 - Otro Lugar (Especif.)				

Lugar y Fecha de Emisión

Firma y Sello del Responsable

NOTA DE IMPORTANCIA: Este formulario deberá ser devuelto dentro de los 10 primeros días posteriores al mes a que se refiere la información. ORIGINAL para de Dirección Nacional de Política Criminal.

Anexo 2.3 - Planilla del Sistema de Alerta Temprana (SAT) para homicidios culposos en accidentes de tránsito

S A T

HOMICIDIOS CULPOSOS EN ACCIDENTES DE TRÁNSITO

Ministerio de Justicia de la Nación. Dirección Nacional de Política Criminal

Provincia: **Departamento:** **Año:** **Mes:**

DATOS DEL HECHO										DATOS DE LA VÍCTIMA				DATOS DEL IMPUTADO			
Nº de Sumario	Fecha del hecho	Hora del hecho	Nombre de la calle o ruta	Altura / km / entre calles	Localidad	Intersección (A)	Semáforo (B)	Modo de Producción del Hecho (C)	Modo de Producción del Hecho Especificar sólo código 5	Condiciones Climátic. (D)	Sexo (E)	Edad	Clase de Víctima (F)	Clase de Víctima Especificar sólo código 5	Vehículo Víctima (excluir peatón) (G)	Vehículo Víctima (excluir peatón) Especificar sólo código 10	Sexo (E)	Edad	Vehículo del Imputado (G)	Vehículo del Imputado Especificar sólo código 10
	/ /	:																		
	/ /	:																		
	/ /	:																		
	/ /	:																		
	/ /	:																		
	/ /	:																		
	/ /	:																		
	/ /	:																		
	/ /	:																		
	/ /	:																		
	/ /	:																		
	/ /	:																		
	/ /	:																		
	/ /	:																		
	/ /	:																		

TABLAS DE CODIFICACIÓN DE DATOS

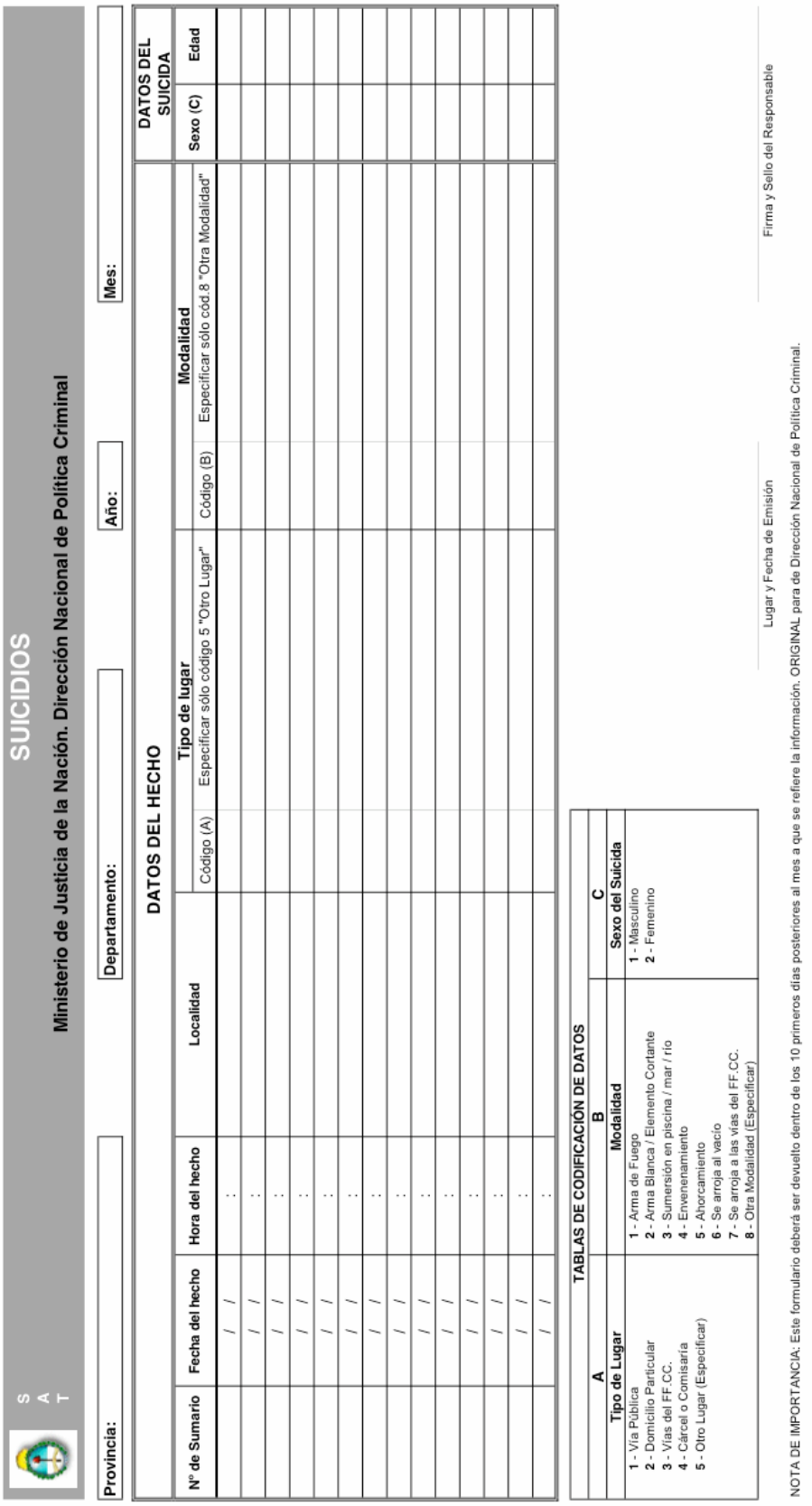
A	B	C	D	E	F	G
Intersec.	**Semáforo**	**Modo de Producción del Hecho** (variables no excluyentes)	**Condiciones Climáticas**	**Sexo de la Víctima o del Imputado**	**Clase de Víctima**	**Vehículo de la Víctima o del Imputado**
1 - SI **2** - NO	**1** - Funciona bien **2** - No funciona **3** - No hay	**1** - Colisión Vehículo / persona **2** - Colisión Vehículo / vehículo **3** - Colisión Vehículo / objeto **4** - Vuelco / despistes **5** - Otro modo (Especificar)	**1** - Normal **2** - Niebla **3** - Lluvia/llovizna **4** - Nieve/granizo **5** - Otra condición	**1** - Masculino **2** - Femenino	**1** - Conductor **2** - Acompañante **3** - Pasajero **4** - Peatón **5** - Otra (Especif.)	**1** - Micro larga dist **2** - Colectivo **3** - Camión **4** - Camioneta **5** - Automóvil **6** - Motocicleta **7** - Ciclomotor **8** - Bicicleta **9** - Tren **10** - Otro (Especif.)

NOTA DE IMPORTANCIA: Este formulario deberá ser devuelto dentro de los 10 primeros días posteriores al mes a que se refiere la información. ORIGINAL para Dirección Nacional de Política Criminal.

Lugar y Fecha de Emisión

Firma y Sello del Responsable

Anexo 2.4 - Planilla del Sistema de Alerta Temprana (SAT) para suicidios



SUICIDIOS

S A T

Ministerio de Justicia de la Nación. Dirección Nacional de Política Criminal

Provincia: | Departamento: | Año: | Mes:

DATOS DEL HECHO

Nº de Sumario	Fecha del hecho	Hora del hecho	Localidad	Tipo de lugar		Modalidad		DATOS DEL SUICIDA	
				Código (A)	Especificar sólo código 5 "Otro Lugar"	Código (B)	Especificar sólo cód.8 "Otra Modalidad"	Sexo (C)	Edad
	/ /	:							
	/ /	:							
	/ /	:							
	/ /	:							
	/ /	:							
	/ /	:							
	/ /	:							
	/ /	:							
	/ /	:							
	/ /	:							
	/ /	:							
	/ /	:							
	/ /	:							
	/ /	:							

TABLAS DE CODIFICACIÓN DE DATOS

A

Tipo de Lugar
1 - Vía Pública
2 - Domicilio Particular
3 - Vías del FF.CC.
4 - Cárcel o Comisaría
5 - Otro Lugar (Especificar)

B

Modalidad
1 - Arma de Fuego
2 - Arma Blanca / Elemento Cortante
3 - Sumersión en piscina / mar / río
4 - Envenenamiento
5 - Ahorcamiento
6 - Se arroja al vacío
7 - Se arroja a las vías del FF.CC.
8 - Otra Modalidad (Especificar)

C

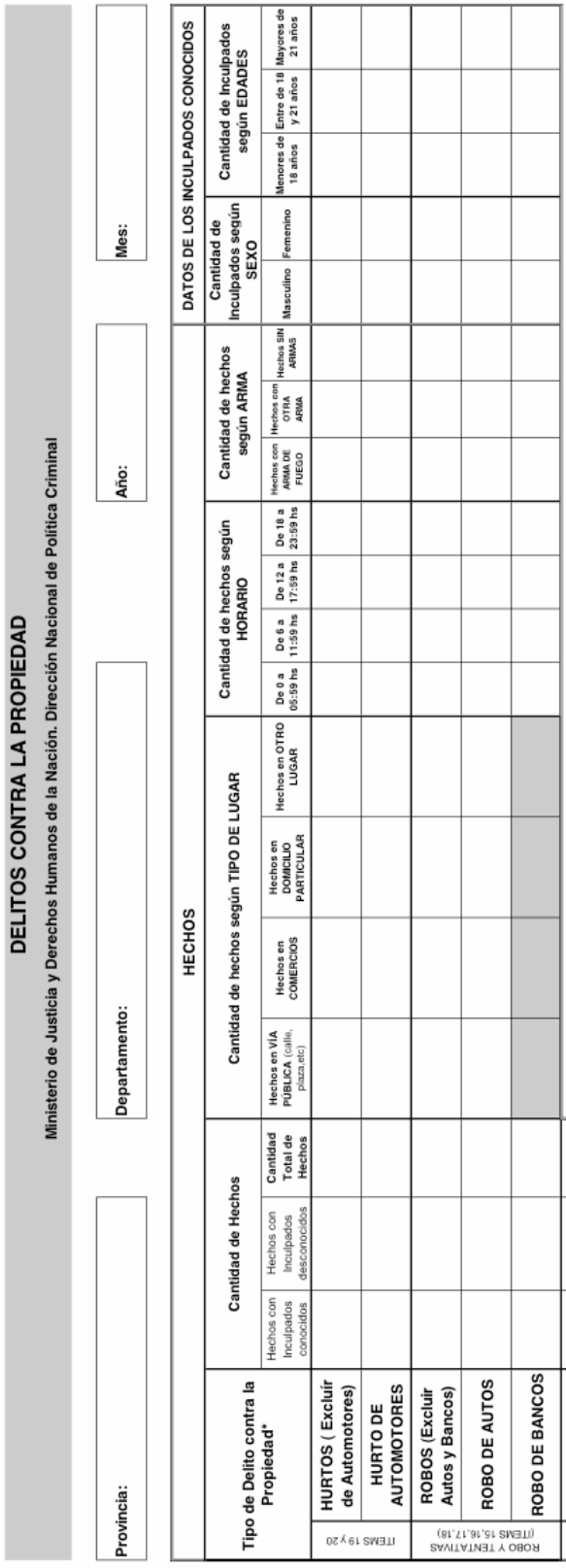
Sexo del Suicida
1 - Masculino
2 - Femenino

Lugar y Fecha de Emisión

Firma y Sello del Responsable

NOTA DE IMPORTANCIA: Este formulario deberá ser devuelto dentro de los 10 primeros días posteriores al mes a que se refiere la información. ORIGINAL para de Dirección Nacional de Política Criminal.

Anexo 2.5 - Planilla del Sistema de Alerta Temprana (SAT) para delitos contra la propiedad



DELITOS CONTRA LA PROPIEDAD

Ministerio de Justicia y Derechos Humanos de la Nación. Dirección Nacional de Política Criminal

Provincia: **Departamento:** **Año:** **Mes:**

	Tipo de Delito contra la Propiedad*	HECHOS														DATOS DE LOS INCULPADOS CONOCIDOS					
		Cantidad de Hechos			Cantidad de hechos según TIPO DE LUGAR				Cantidad de hechos según HORARIO				Cantidad de hechos según ARMA			Cantidad de Inculpados según SEXO		Cantidad de Inculpados según EDADES		
		Hechos con Inculpados conocidos	Hechos con Inculpados desconocidos	Cantidad Total de Hechos	Hechos en VÍA PÚBLICA (calle, plaza,etc)	Hechos en COMERCIOS	Hechos en DOMICILIO PARTICULAR	Hechos en OTRO LUGAR	De 0 a 05:59 hs	De 6 a 11:59 hs	De 12 a 17:59 hs	De 18 a 23:59 hs	Hechos con ARMA DE FUEGO	Hechos con OTRA ARMA	Hechos SIN ARMAS	Masculino	Femenino	Menores de 18 años	Entre de 18 y 21 años	Mayores de 21 años
ITEMS 19 y 20	HURTOS (Excluir de Automotores)																				
	HURTO DE AUTOMOTORES																				
ROBO Y TENTATIVAS (ITEMS 15,16,17,18)	ROBOS (Excluir Autos y Bancos)																				
	ROBO DE AUTOS																				
	ROBO DE BANCOS																				
OTROS DELITOS CONTRA LA PROPIEDAD (ITEM 21)	EXTORSIONES																				
	SECUESTROS																				
	ESTAFAS																				
	USURA																				
	QUIEBRAS																				
	USURPACIÓN																				
	DAÑOS																				

Nota: LOS DATOS A TRANSCRIBIR A LA PLANILLA DE DELITOS CONTRA LA PROPIEDAD SE OBTIENEN DE LA SIGUIENTE MANERA:

* HURTOS: ES LA SUMA DE LOS ITEMS 19 Y 20 (HURTOS Y TENTATIVAS DE HURTO) DE LA PLANILLA DE SNIC. EL HURTO DE AUTOMOTORES SE DETALLA EN RENGLON APARTE.

* ROBOS: ES LA SUMA DE LOS ITEMS 15, 16, 17, 18 (ROBOS, TENTATIVAS DE ROBO, ROBO AGRAVADO, TENTATIVA DE ROBO AGRAVADO) DE LA PLANILLA DE SNIC. LOS ROBOS DE AUTO Y ROBOS DE BANCO SE COLOCAN APARTE.

* LA SUMA DE EXTORSIONES, SECUESTROS, ESTAFAS, USURA, QUIEBRAS, USURPACIÓN Y DAÑOS DEBE SER IGUAL AL ITEM 21 (OTROS DELITOS CONTRA LA PROPIEDAD) DE LA PLANILLA DE SNIC

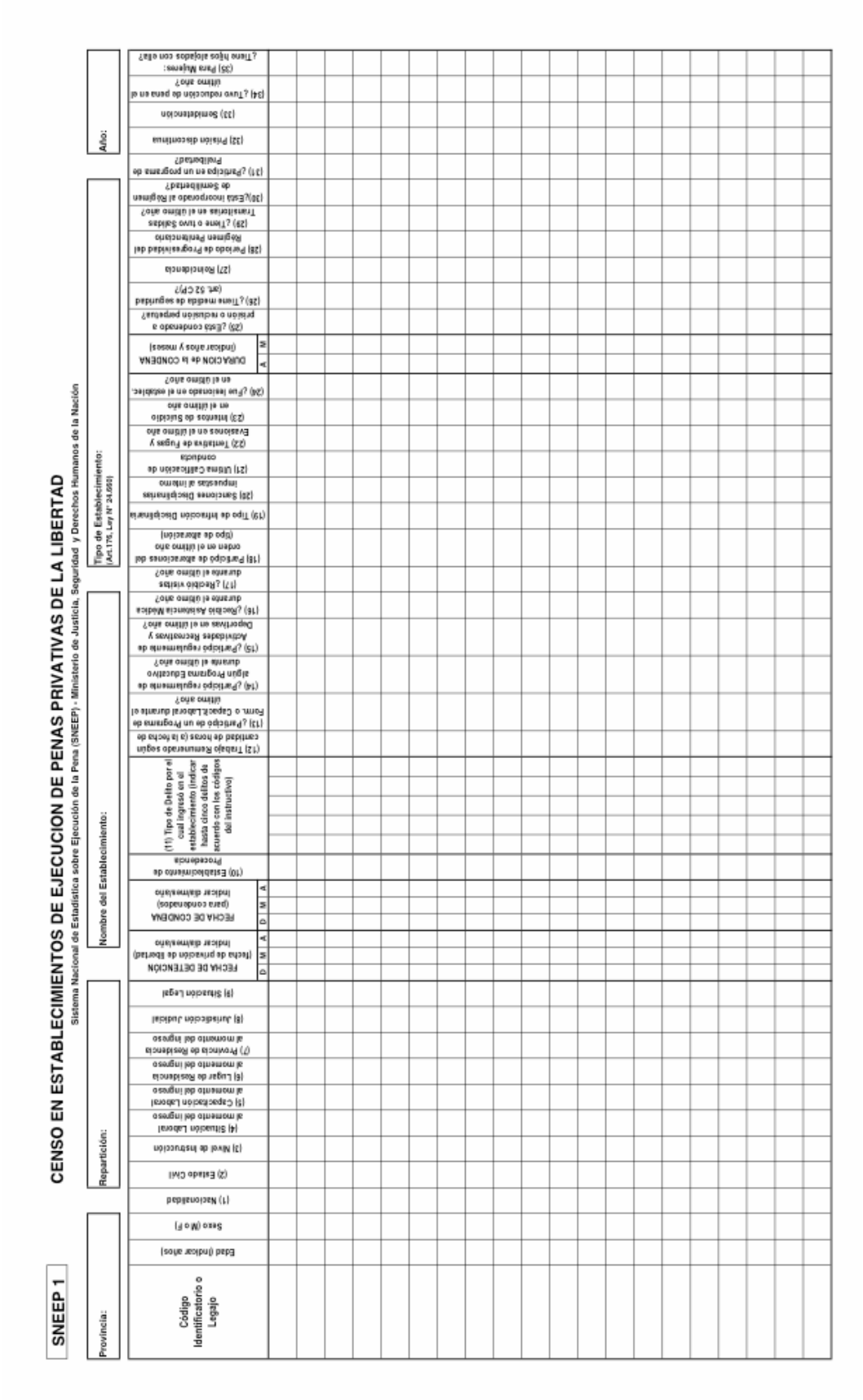
*Todos los tipos delictivos incluyen sus tentativas

NOTA DE IMPORTANCIA: Este formulario deberá ser devuelto dentro de los 10 primeros días posteriores al mes a que se refiere la información. ORIGINAL para de Dirección Nacional de Política Criminal.

Lugar y Fecha de Emisión

Firma y Sello del Responsable

Anexo 2.6 - Planilla del Sistema Nacional de Estadísticas sobre Ejecución de la Pena (SNEEP)



ANEXO 3 – VENTANAS DE *WEKA*

En el siguiente anexo se detalla el proceso realizado en *Weka 3.5.5* que permitió llegar a los resultados experimentales expuestos en el trabajo.

En primer lugar se debe mencionar que la confección del *data set*, desarrollado en el capítulo 5, se realizó en *MS Excel* y se guardó bajo el formato CSV (*comma separated values*).

La ventana de inicio de *Weka* es la siguiente [Figura 10.1]:

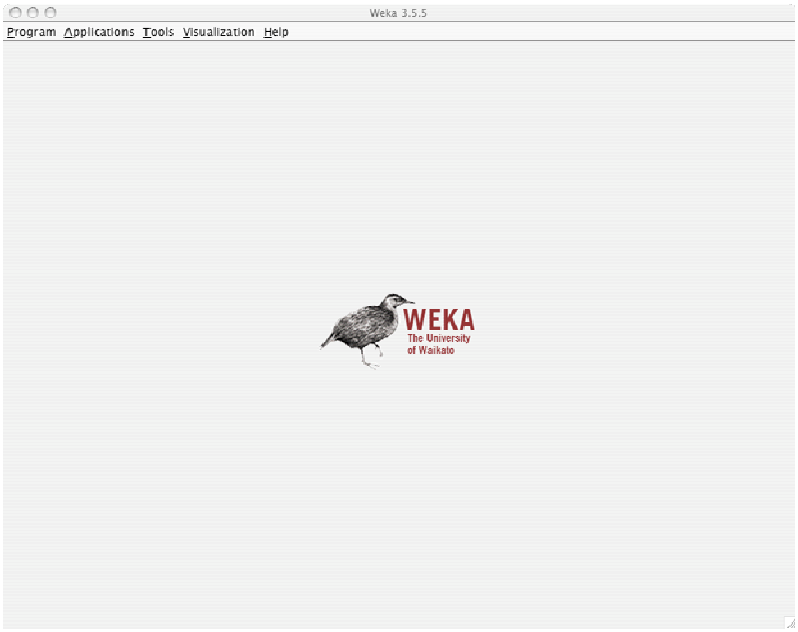


Figura 10.1: Ventana de inicio de *Weka*

Se seleccionó la opción *Explorer* del menú *Applications* [Figura 10.2]:

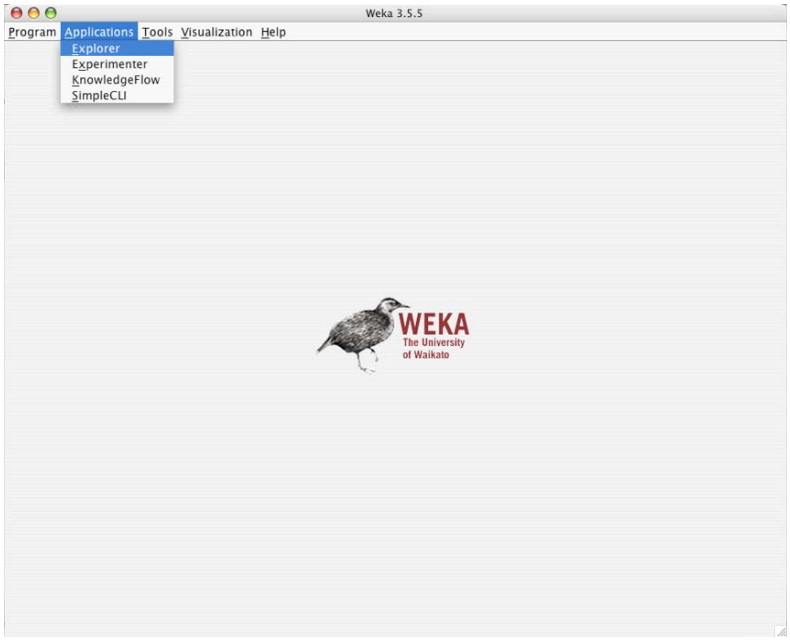


Figura 10.2: Selección del *Explorer*

Dentro de la ventana del *Explorer* de *Weka* se visualizan distintas solapas para realizar distintos procesos de minería de datos (*Preporcess*, *Classify*, *Cluster*, *Associate*, *Select attributes* y *Visualize*). El *Explorer* se inicia en la solapa *Preprocess* [Figura 10.3].

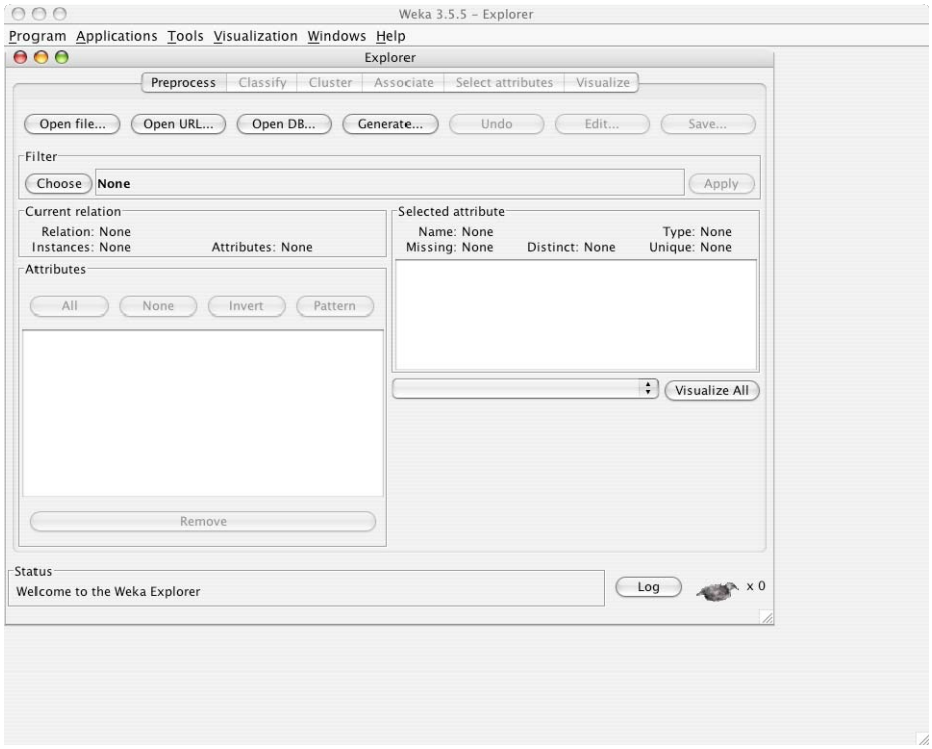


Figura 10.3: Visualización inicial del *Explorer*

Dentro de esta solapa se seleccionó el botón *Open file…* y luego se buscó el archivo del *data set* en formato CSV generado previamente con *MS Excel* [Figura 10.4]

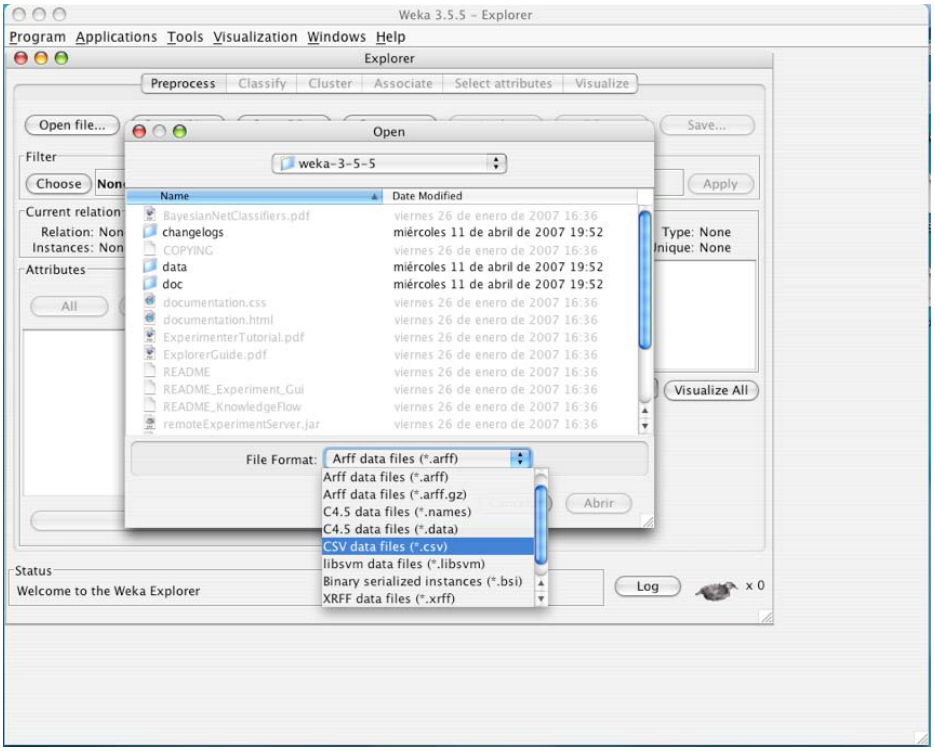


Figura 10.4: Ventana de inicio de *Weka*

Una vez abierto el archivo, en la solapa *Preprocess* se visualizan los atributos y sus variables o estados [Figura 10.5]:

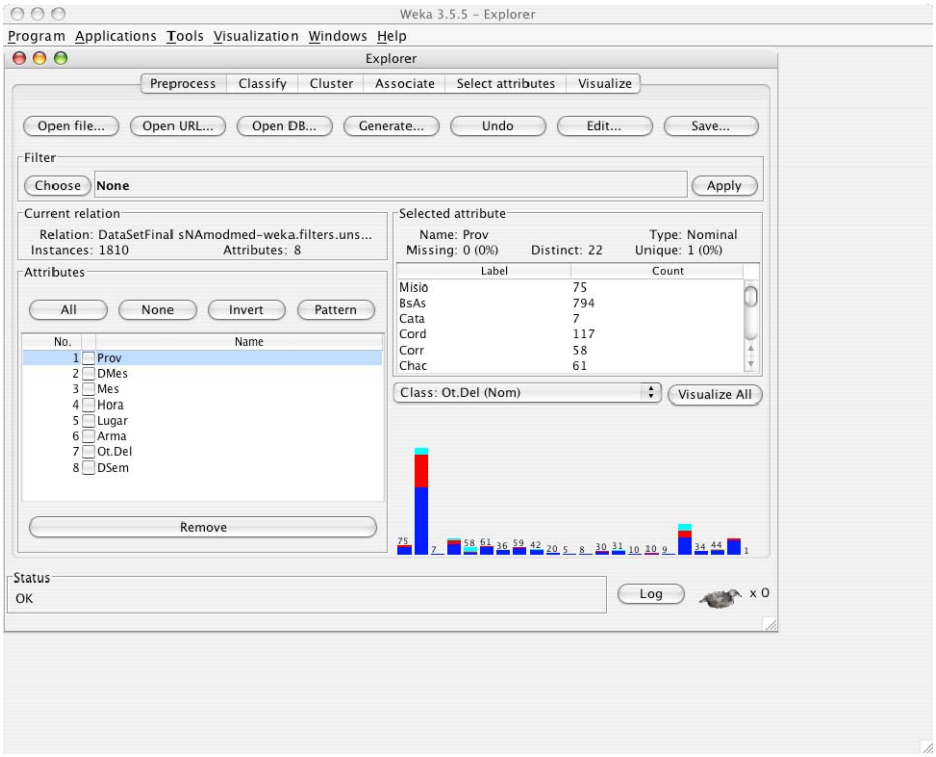


Figura 10.5: Visualización de la solapa *Preprocess* una vez abierto el archivo

Luego se seleccionó la solapa *Cluster* para iniciar el proceso de *clustering*. La visualización inicial de esta solapa es la siguiente [Figura 10.6]:

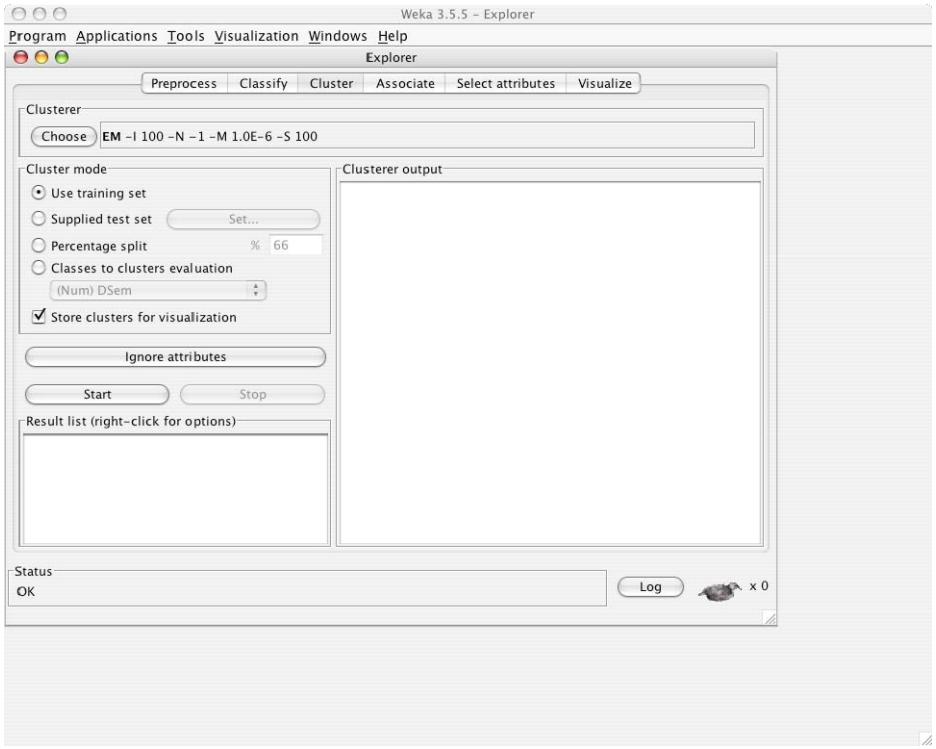


Figura 10.6: Visualización inicial de la solapa *Cluster*

Por *default* se encuentra activado el algoritmo *EM*. Para activar *K-means* se seleccionó la opción *SimpleKMeans* a través del botón *Choose* [Figura 10.7] y luego se configuraron los parámetros del algoritmo: cantidad de clusters (3) y semilla (1) [Figura 10.8].

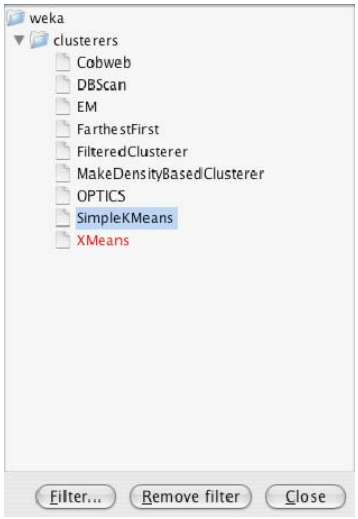


Figura 10.7: Selección de *K-means*

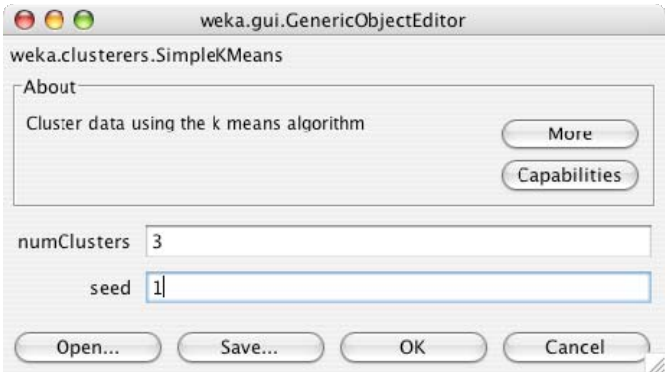


Figura 10.8: Configuración de los parámetros de *K-means*

Los parámetros referidos al modo de entrenamiento (*Cluster mode*) no se modificaron [Figura 10.9]. Luego se inició la corrida presionando el botón *Start* y se mostró el resultado [Figura 10.9].

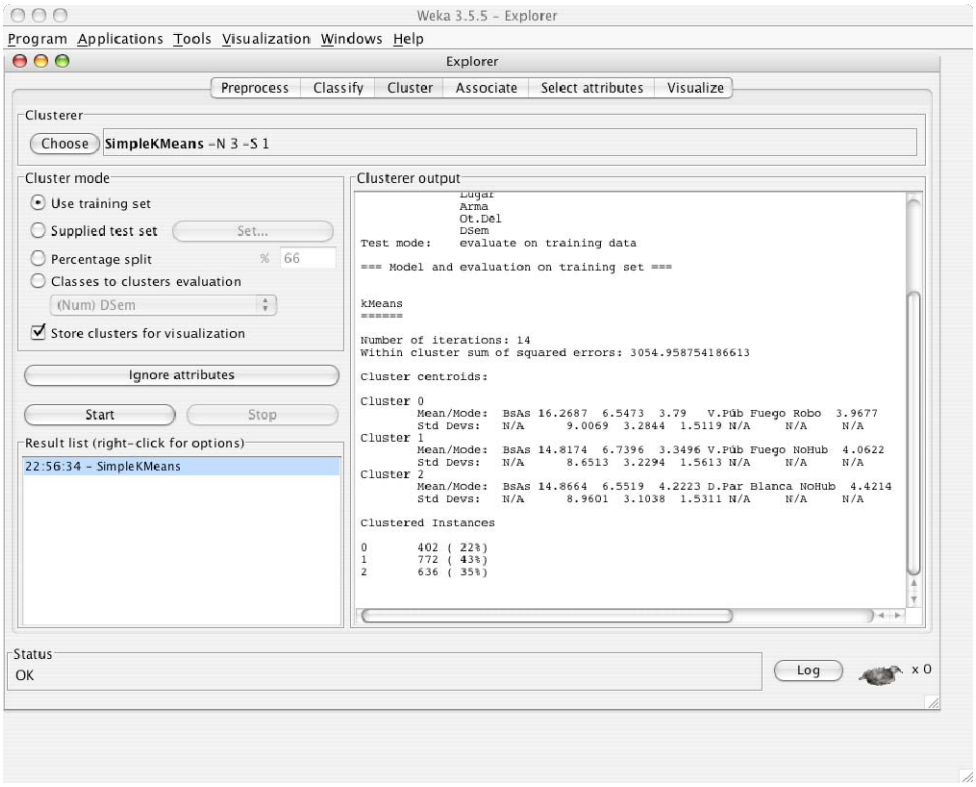


Figura 10.9: Resultado de la corrida de *K-means*

Para visualizar la asignación a los clusters en los gráficos de dispersión se hizo clic derecho sobre la lista de resultados (*Result list)*, que aparece en la zona inferior izquierda, y se seleccionó la opción *Visualize cluster assignments* [Figura 10.10].

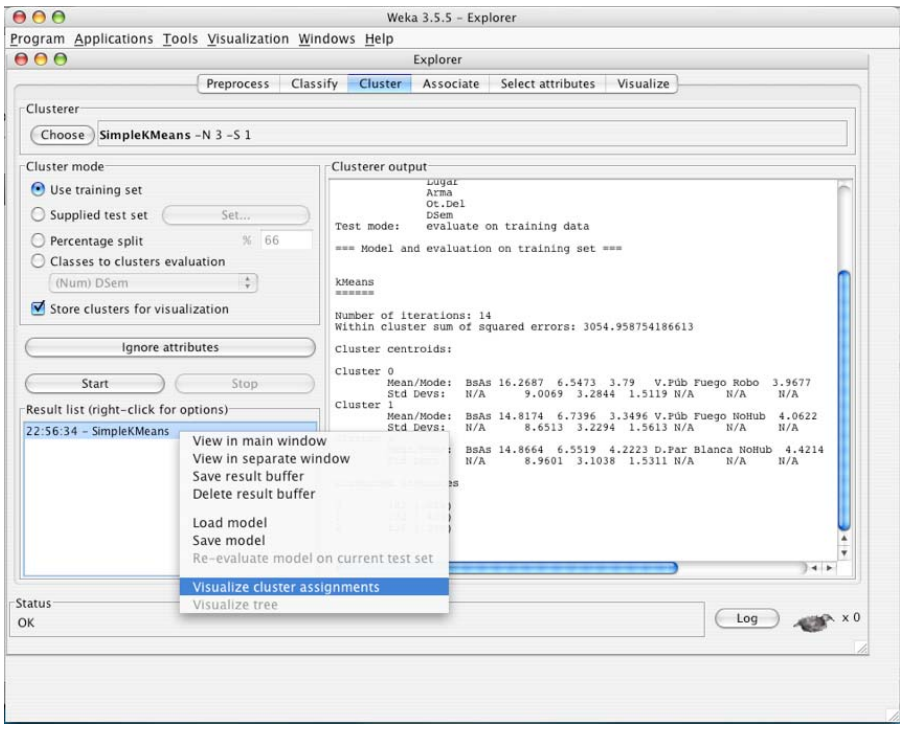


Figura 10.10: Menú desplegable para la visualización de los clusters

Dentro de la ventana de visualización se fueron seleccionando las variables de interés analizadas en el capítulo 7. Para una mejor visualización se estableció el nivel de *Jitter* (opción que permite una mejor separación de puntos) en su punto máximo [Figura 10.11].

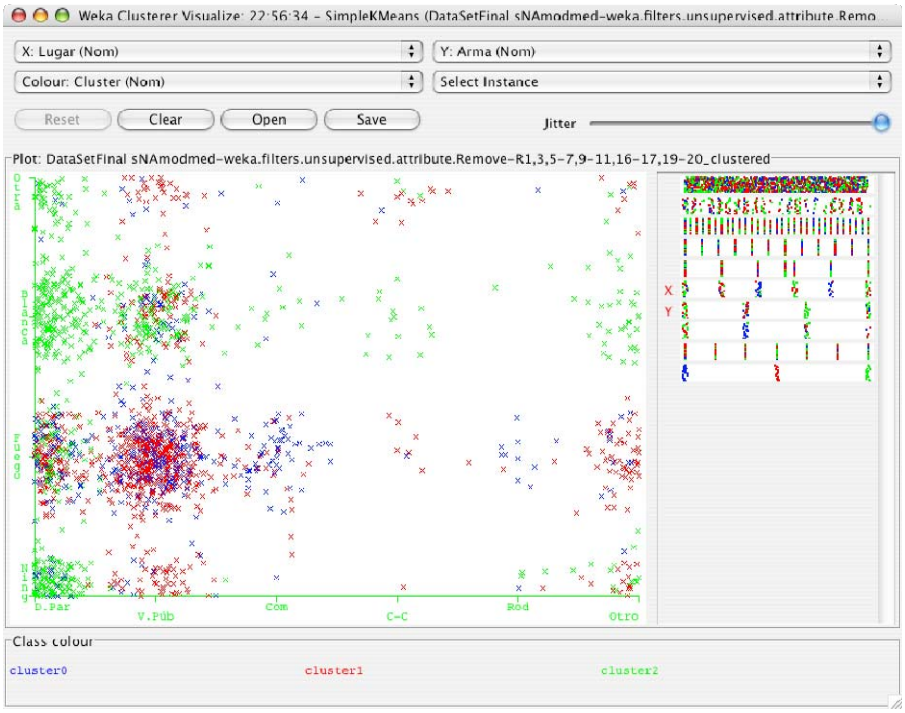


Figura 10.11: Visualización de los clusters

Para poder guardar el archivo con las nuevas asignaciones de *clusters* como nuevos atributos se presionó el botón *Save* presente en la ventana de visualización [Figura 10.11] y luego se guardó con un nuevo nombre bajo el formato Arff [Figura 10.12].

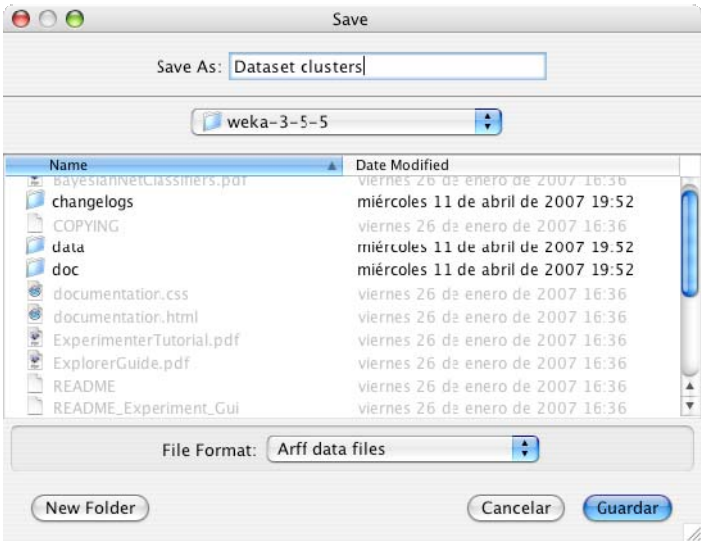


Figura 10.12: Ventana para guardar la asignación de clusters

Se abrió este nuevo archivo desde la solapa *Preprocess* (ahora con el atributo cluster) y se presionó el botón *Visualize All* situado en el extremo superior derecho del gráfico [Figura 10.13].

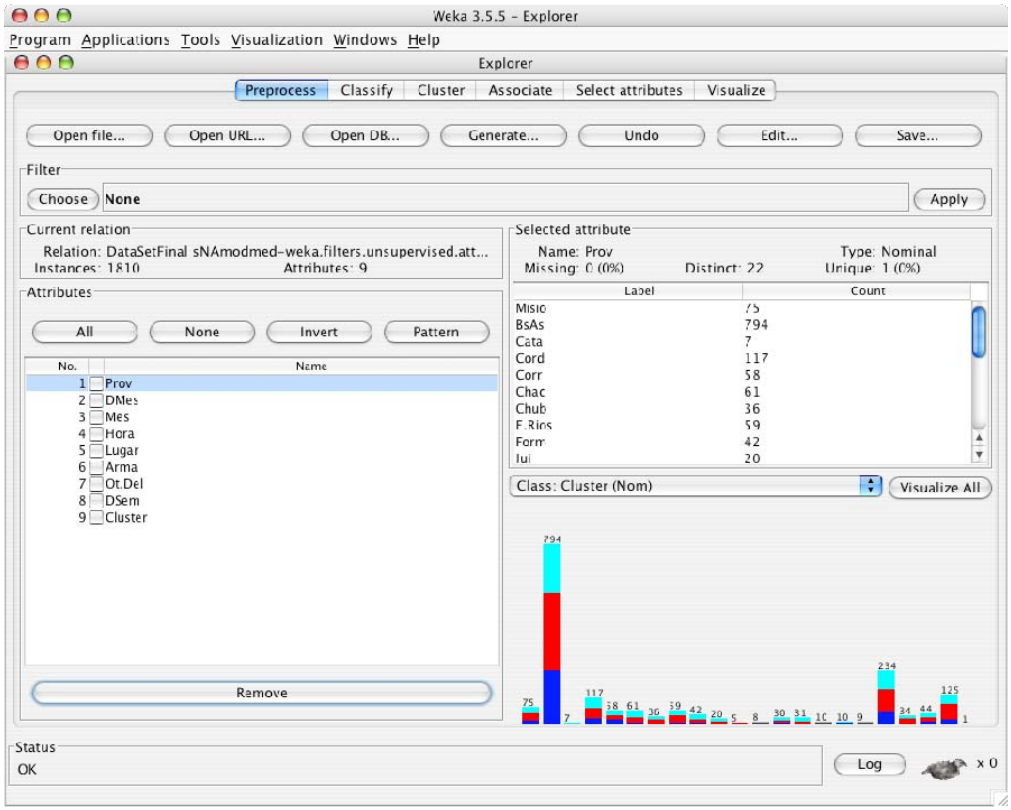


Figura 10.13: Visualización de la solapa *Preprocess* con el cluster como atributo

Esta opción permitió acceder a la visualización de la distribución de los clusters en todas las variables de los atributos en forma de gráficos de barras [Figura 10.14].

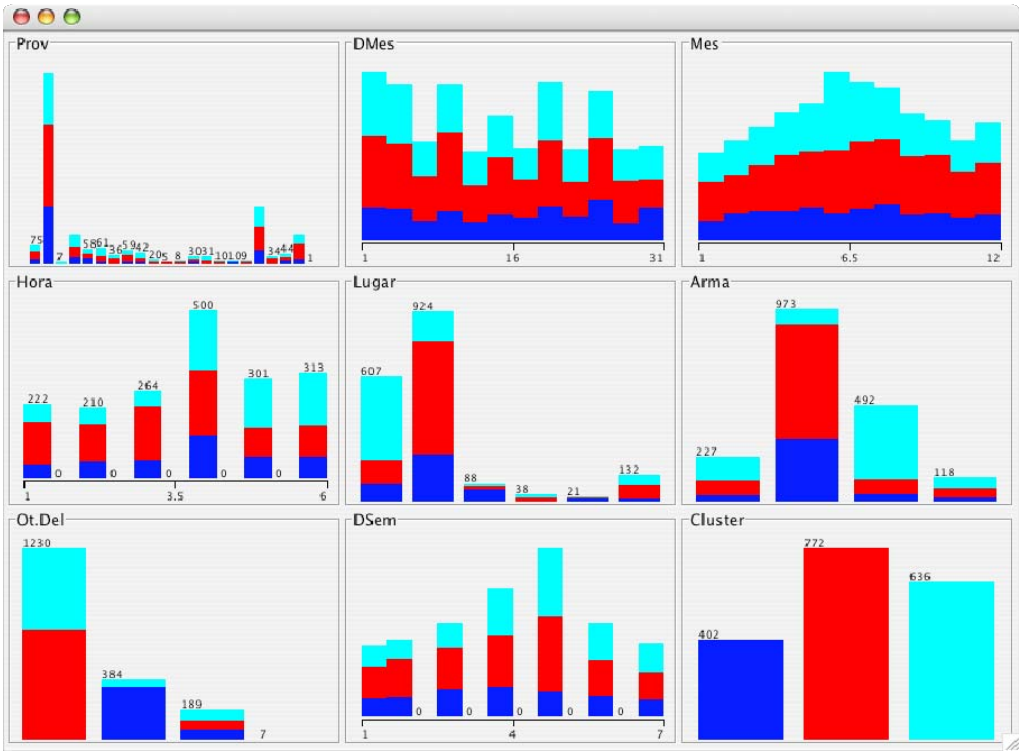


Figura 10.14: Visualización de la distribución de los clusters

Con el objetivo de seleccionar los atributos a ser utilizados para el proceso de clasificación, desde la ventana del *Explorer*, se seleccionó la solapa *Select attributes*. Su visualización inicial es la siguiente [Figura 10.15]:

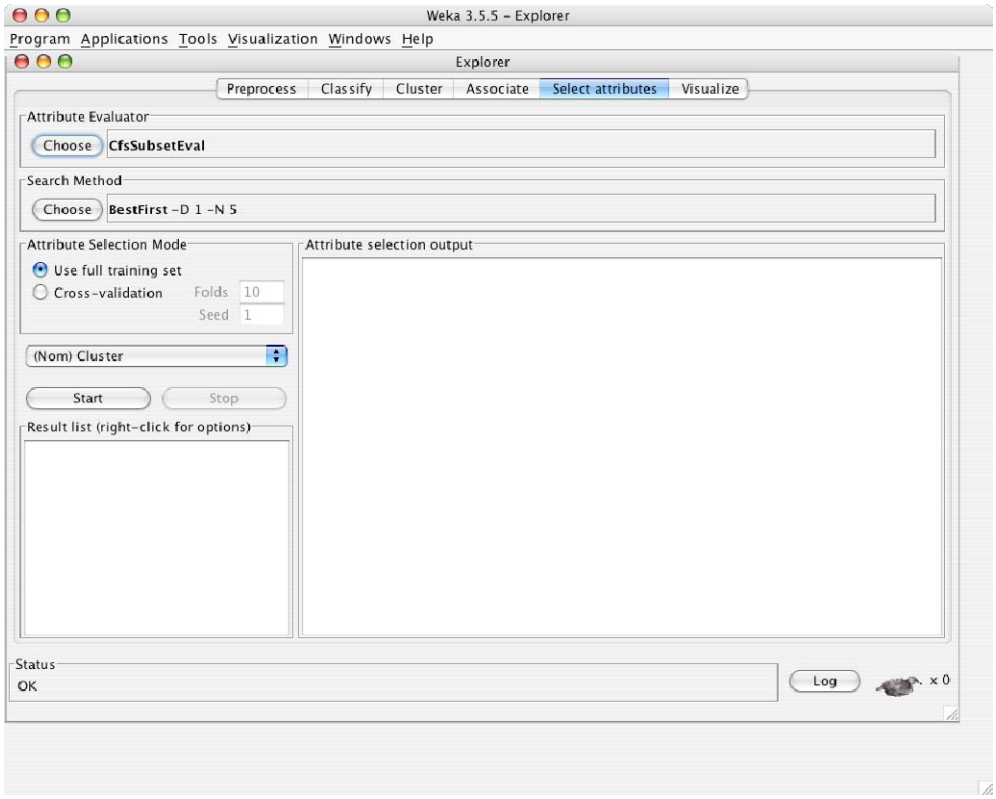


Figura 10.15: Visualización inicial de la solapa *Select attributes*

Desde esta ventana, presionando los respectivos botones *Choose*, se fueron seleccionando los distintos algoritmos evaluadores [Figura 10.16] y métodos de búsqueda [Figura 10.17] presentados en el capítulo 7.

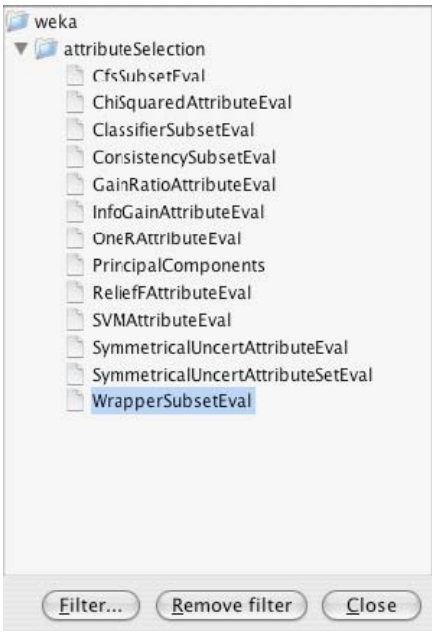


Figura 10.16: Selección de algoritmos evaluadores de atributos

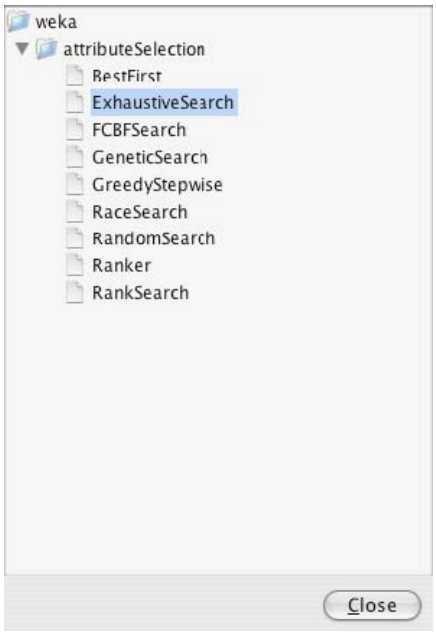


Figura 10.17: Selección de los métodos de búsqueda

Para el caso especial de los algoritmos evaluadores empleados que funcionan a priori del algoritmo de clasificación a utilizar (como *WrapperSubsetEval*), fue necesario especificar este algoritmo (en nuestro caso J48) [Figura 10.18] y determinar sus parámetros a ser utilizados durante la clasificación (se dejaron los determinados por *default*) [Figura 10.19].

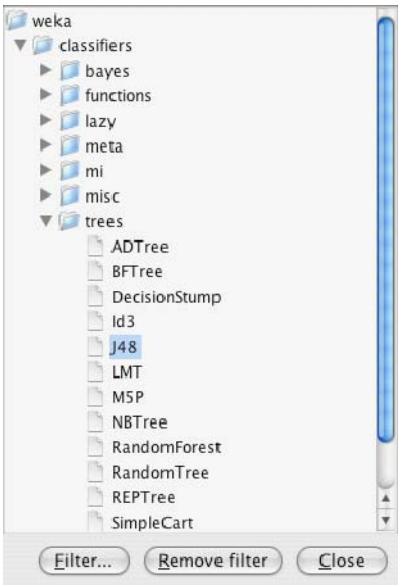


Figura 10.18: Selección del algoritmo de clasificación a utilizar

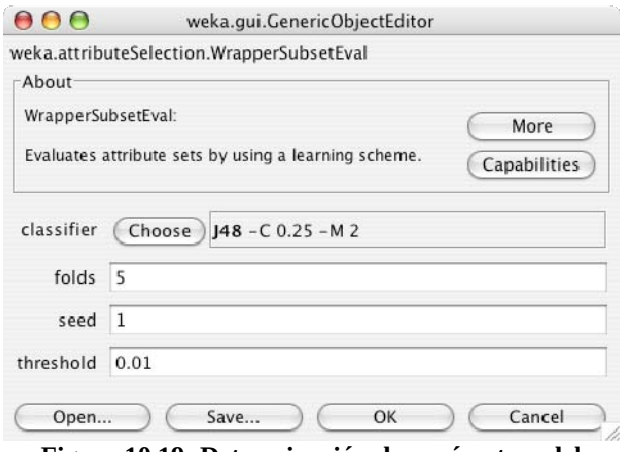


Figura 10.19: Determinación de parámetros del algoritmo de clasificación a utilizar

El resultado de la corrida para el caso puntual del algoritmo *WrapperSubsetEval* se visualiza de la siguiente forma [Figura 10.20]:

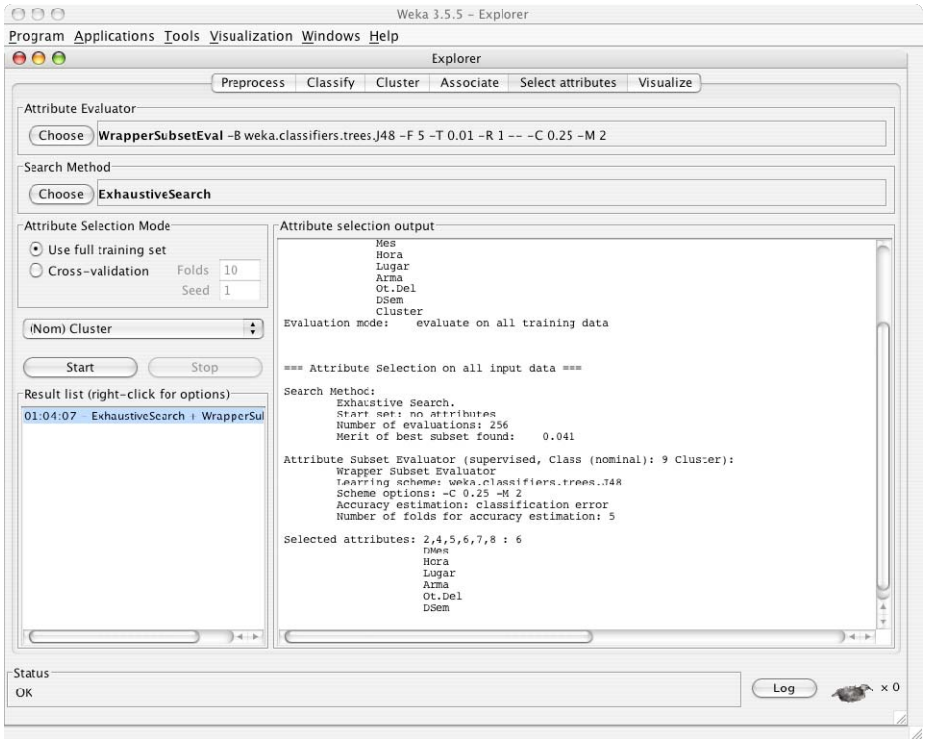


Figura 10.20: Visualización de los resultados de la selección de atributos óptimos según el algoritmo *WrapperSubsetEval*

Finalmente se inició el proceso de clasificación mediante la selección de la solapa *Classify* desde la ventana del *Explorer*. Su visualización inicial es la siguiente [Figura 10.21]:

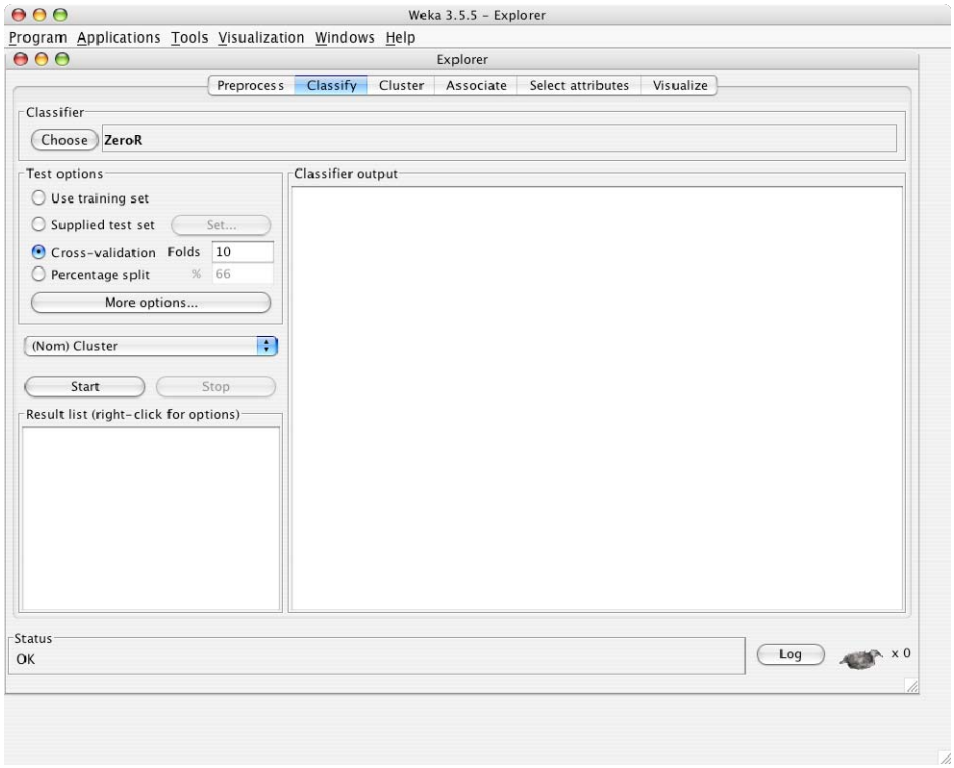


Figura 10.21: Visualización inicial de la solapa *Classify*

Por *default* se encuentra seleccionado el algoritmo *ZeroR*, por lo que se presionó el botón *Choose* para seleccionar J48 (equivalente a C4.5 en *Weka*), ubicado en el conjunto de clasificadores *trees* [Figura 10.22].

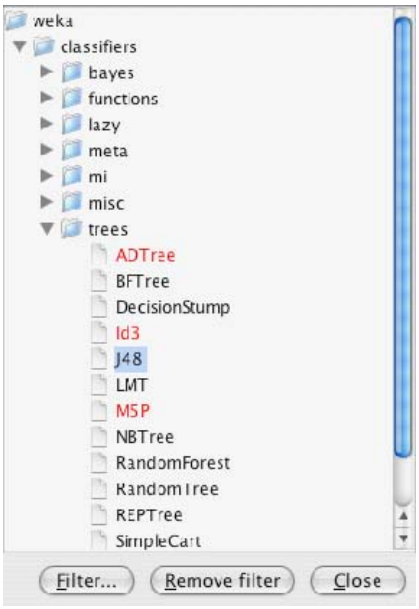


Figura 10.22: Selección del algoritmo de clasificación J48

Los parámetros de J48 no fueron modificados (se dejaron los presentes por *default*). Dentro de las opciones de entrenamiento (*Test options*) se seleccionó: *Use training set* [Figura 10.23].

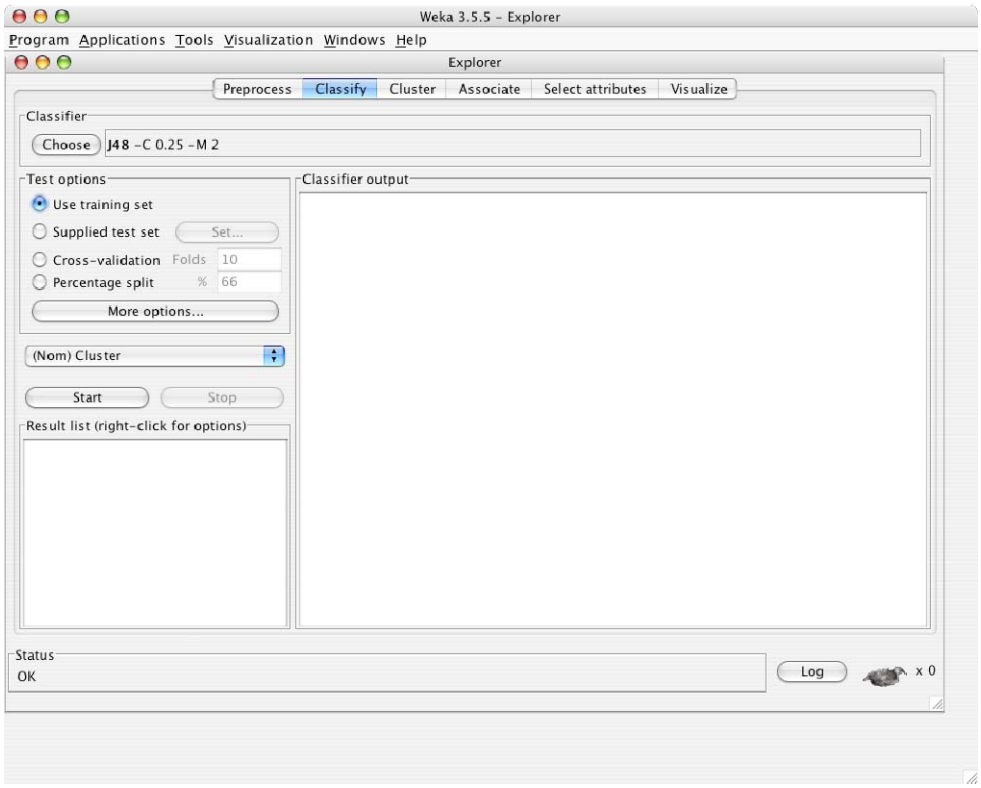


Figura 10.23: Visualización la solapa *Classify* antes de la corrida con J48

Tras inicializar la corrida mediante el botón *Start* se visualizaron los resultados de la clasificación [Figura 10.24].

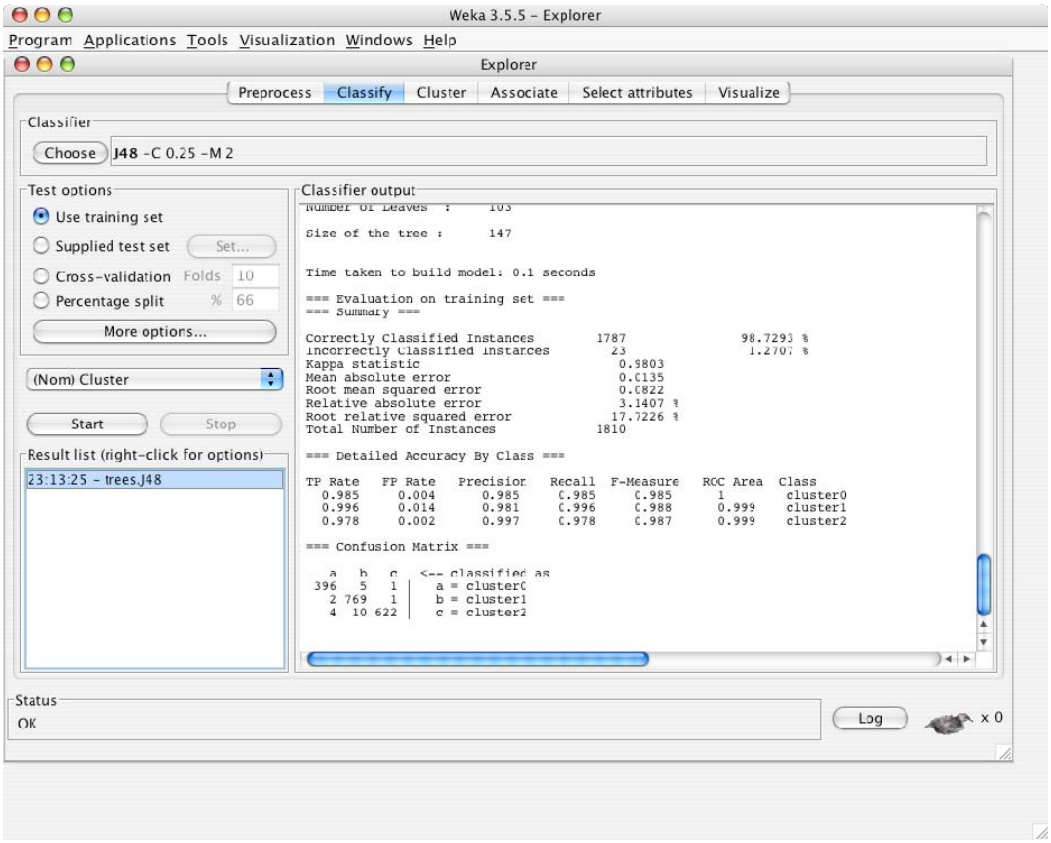


Figura 10.24: Visualización de los resultados de la clasificación con el algoritmo J48

Para visualizar el árbol se hizo clic derecho sobre la lista de resultados (*Result list)*, que aparece en la zona inferior izquierda [Figura 10.24], y de la lista desplegada se seleccionó la opción *Visualize tree* [Figura 10.25]. La visualización definitiva del árbol se ve en la Figura 10.26.

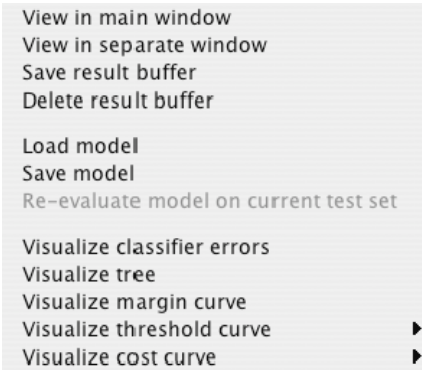


Figura 10.25: Menú desplegable para la visualización del árbol

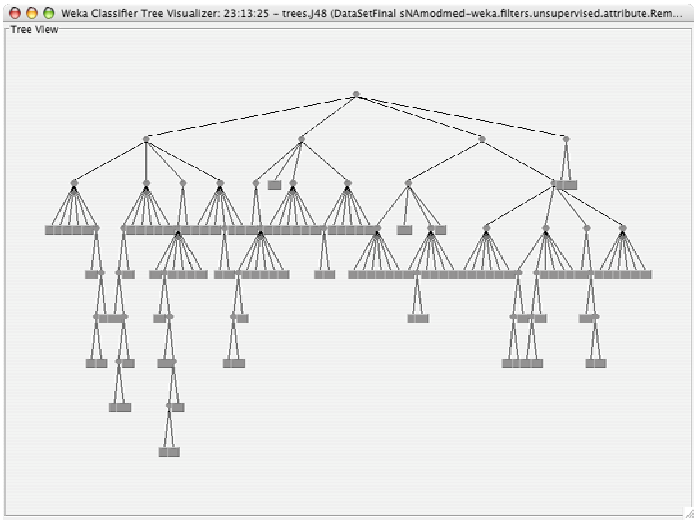
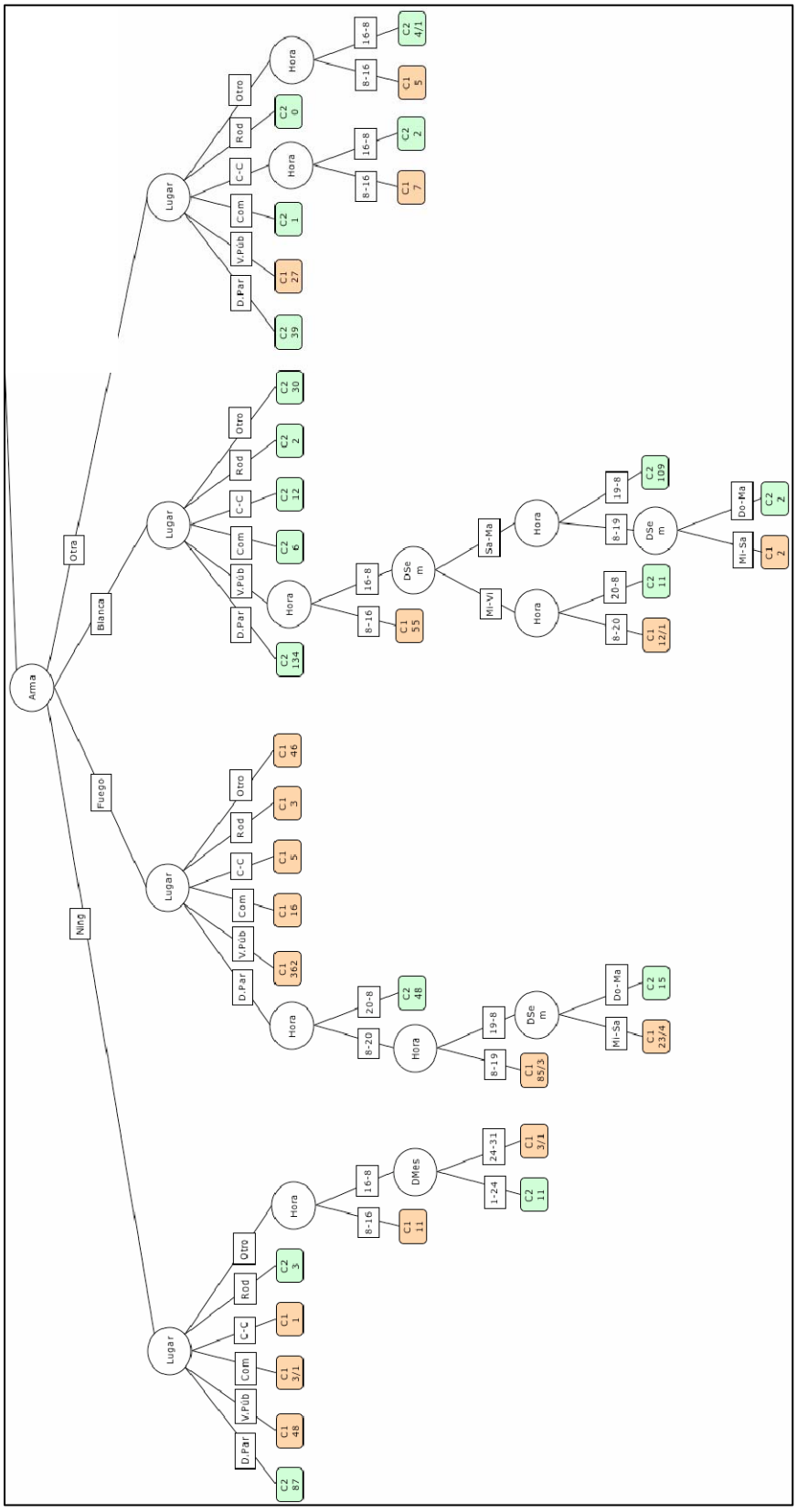


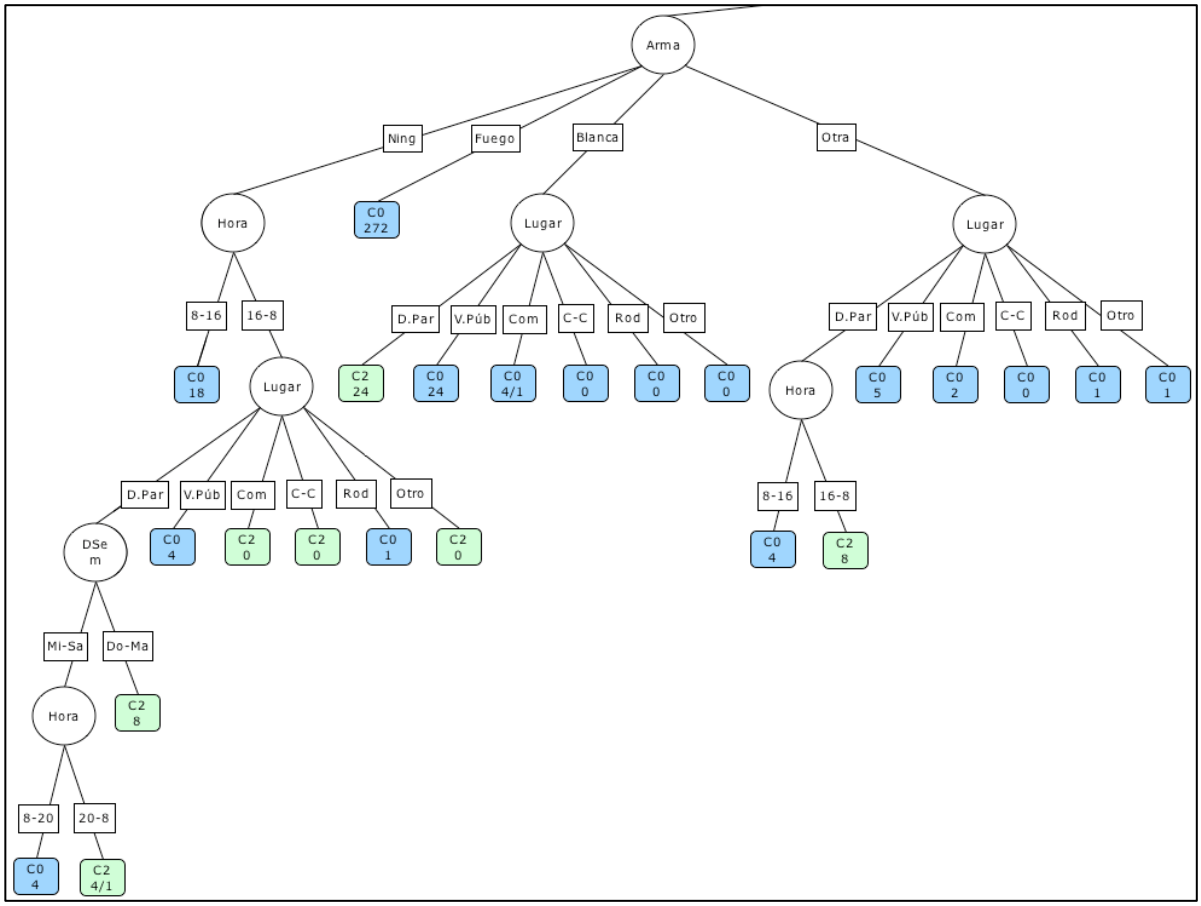
Figura 10.26: Visualización del árbol obtenido

ANEXO 4 – ÁRBOL GENERADO CON C4.5

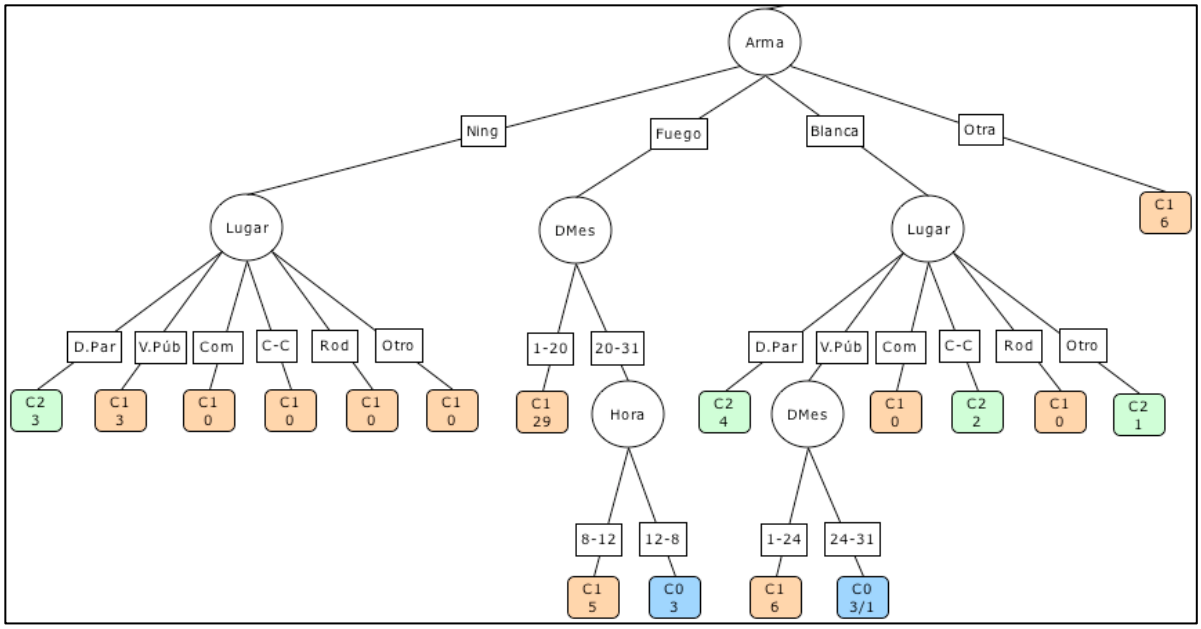
Anexo 4.1 – Rama *no hubo otro delito*



Anexo 4.2 – Rama *robo*



Anexo 4.3 – Rama *otro delito / hora 8-20*



Anexo 4.4 – Rama *otro delito* / *hora 20-8*