EXPLOTACION DE INFORMACIÓN APLICADA A INTELIGENCIA CRIMINAL EN ARGENTINA

Britos, P., Fernández, E., Merlino, H., Pollo-Cataneo, F., Rodríguez, D., Procopio, C., Rancan, C., García-Martínez, R.

Centro de Ingeniería del Software e Ingeniería del Conocimiento. Escuela de Postgrado. ITBA.
Departamento de Ingeniería Industrial. ITBA.
Laboratorio de Sistemas Inteligentes. Facultad de Ingeniería. Universidad de Buenos Aires

rgm@itba.edu.ar

Resumen

El presente trabajo describe un Proyecto de Explotación de información en el ámbito de la información criminal, analizando homicidios dolosos, homicidios culposos en accidentes de tránsito y la población carcelaria Argentina. Los análisis se realizan a través de herramientas de explotación de información de distribución libre.

Palabras claves: Explotación de información, aplicación de explotación de información, criminología.

1. INTRODUCCION

Desde hace ya varios años la inseguridad pasó a ser un tema recurrente en la vida de los argentinos. Reducir el índice de delitos como el de muertes por accidentes de tránsito es un tema que tiene muy presente cada nuevo gobierno. La toma de decisiones eficientes y la implementación de medidas preventivas en lo que respecta a política criminal y social, es el primer paso que se debería analizar a fin de lograr reducir la cantidad de hechos delictivos [Valenga, F., et al; 2007a; Lázaro Castillo, J. 2007; Gutierrez-Rüegg, P. *et al*; 2008; Gutierrez-Rüegg, P.; 2008]. Se cree que analizar los registros criminales sobre hechos delictivos ocurridos en el pasado como también estudiar a los autores materiales de dichos hechos, es fundamental a la hora de intentar prevenir delitos. 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. Es necesario un tratamiento de la información estadística más complejo que obligue a evolucionar en el análisis de información criminal. Aún más cuando se trabajan con base de datos de más de 50,000 registros, como es, por ejemplo, el caso de la población carcelaria en Argentina. El objetivo primario de estas bases de datos es, como su nombre indica, almacenar grandes cantidades de datos organizados siguiendo un determinado esquema o modelo de datos que facilite su almacenamiento, recuperación y modificación, pero no así su posterior uso o análisis. En muchos casos los registros almacenados son demasiados grandes y complejos como para analizar [Kantardzic, 2003].

Una posible herramienta a utilizar para tratar los grandes volúmenes de información almacenados en dichas bases de datos es la técnica de Explotación de información. El termino *Explotación de información* (Data Mining) puede ser definido como la manera no trivial de extracción de información no implícita, previamente desconocida y potencialmente útil, de una base de datos [Frawley *et al*, 1992]. Explotación de información representa la posibilidad de buscar

exhaustivamente dentro de un gran volumen de datos, información y conocimiento que pueden resultar de mucho valor. Es considerada uno de los puntos más importantes de los sistemas expertos de base de datos, y uno de los desarrollos más prometedores en la industria del manejo de la información [Cartagenova, 2005]. En el marco internacional el uso de la explotación de información aplicado a la inteligencia criminal ha tenido un gran crecimiento en los últimos años en EEUU [Chen *et al*, 2004]. A modo de ejemplo, ha sido mencionado como el método mediante el cual supuestamente las fuerzas armadas estadounidenses habrían identificado al líder de los ataques del 11 de Septiembre de 2001, Mohamed Atta, junto a otros 3 terroristas como posibles miembros de una célula de *Al Qaeda* operando en ese país más de un año previo a los ataques.

En este contexto, se busca desarrollar modelos de explotación de información basados en sistemas inteligentes para el estudio de la información criminal relacionada con homicidios dolosos y en accidentes de tránsito, que permitan obtener patrones de comportamientos y faciliten la generación de estrategias de prevención. Al mismo tiempo, se busca estudiar la factibilidad y valor agregado de aplicar estas técnicas en la población carcelaria argentina a fines de poder caracterizarla y poder extraer conclusiones para la prevención de delitos. De esta forma se estaría abarcando un amplio abanico de información relacionada con los crímenes y delitos, comenzando por los hechos ocurridos hasta llegar a la persona que lo comete.

2. ESTADO DE LA CUESTIÓN

2.1. Introducción a la Explotación de información

El termino explotación de información es una etapa dentro de un proceso mayor llamado *Extracción de Conocimiento en Bases de Datos* (*Knowledge Discovery in Databases o KDD*). Lo que en verdad hace la explotación de información es reunir las ventajas de varias áreas como la estadística, la inteligencia artificial, la computación gráfica, las bases de datos y el procesamiento masivo, principalmente usando como materia prima las bases de datos. Una definición tradicional [Fayyad *et al*, 1996] podría ser la siguiente*: "Un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible, de patrones comprensibles que se encuentran ocultos en los datos"*. El creciente volumen de datos en todas las áreas de aplicación humana demanda nuevas y poderosas técnicas de transformación de los datos en conocimiento útil. Se puede decir que la explotación de información es un avance en dicha cuestión. Busca generar información similar a la que podría producir un experto humano, que además satisfaga el Principio de Comprensibilidad (utilizar lenguaje adecuado a la temática del trabajo) [Britos *et al*, 2005].

El objetivo de la explotación de información es descubrir comportamientos interesantes como lo son patrones, asociaciones, cambios, anomalías y estructuras significativas a partir de grandes cantidades de datos almacenados en bases de datos, *data warehouses* o cualquier otro medio de almacenamiento de información [Britos *et Al*, 2005].

2.2. Clustering o Agrupamiento de los Datos

El clustering consiste en agrupar un conjunto de datos sin tener clases predefinidas, basándose en la similitud de los valores de los atributos de los distintos datos. Este tipo de algoritmo se realiza en forma no supervisada ya que no se saben de antemano las clases del conjunto de datos de entrenamiento. El clustering identifica regiones densamente pobladas, de acuerdo a alguna medida de distancia, en un gran conjunto de datos multidimensional [Chen & Han, 1996]. El análisis de clusters se basa en maximizar la similitud de las instancias en cada cluster y minimizar la similitud

entre clusters [Han & Lamber, 2001]. Es utilizado en numerosas aplicaciones tales como reconocimiento de patrones, análisis de datos, procesamiento de imágenes e investigaciones de mercado. Como función de la *explotación de información*, el análisis de clusters puede ser utilizado como una herramienta independiente para obtener una visión de la distribución de los datos, para observar las características de cada cluster y enfocar un análisis más exhaustivo hacia un grupo o cluster determinado. Alternativamente, puede servir como un paso del preprocesamiento de los datos para otros algoritmos, como por ejemplo, el de clasificaciones en el cual se trabajaría luego sobre los clusters originados.

2.3. Clasificación de los Datos. Algoritmos de Inducción

Los algoritmos de clasificación se utilizan para clasificar un conjunto de datos basado en los valores de sus atributos [Servente & García Martínez, 2002]. 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 Los algoritmos más utilizados para la clasificación son los algoritmos de inducción. Aún cuando existen varios enfoques para los algoritmos de inducción, se trabajará con aquellos que generan árboles de decisión conocida como la familia TDIT (*Top Down Induction Trees*). En particular se utiliza el C4.5 [Quinlan, 1993]. El J48 es una implementación mejorada del algoritmo C4.5 funcionando bien tanto con atributos nominales como numéricos. Cabe destacar, que a modo de prueba para el estudio relacionado a la población carcelaria, se utilizó tanto el J48 como el CHAID (*Chi Squared Automatic Interaction Detection*) [Hartigan, 1975] obteniéndose resultados muy similares con ambos algoritmos.

2.4. Información Criminal en Argentina

Se define información criminal, en sentido general, a un conjunto organizado de datos de hechos delictivos que es alimentado por diversas autoridades, instituciones policiales y fuerzas de seguridad con el objetivo de llevar a cabo acciones concretas, entre las que se pueden mencionar [Valenga *et al*., 2007b; 2008]: generación de estadísticas, desarrollo de políticas de prevención y/o persecución penal, asistencia a las investigaciones criminales, generación lineamientos político-criminales, propuesta de leyes al Congreso.

A partir de la sanción de la Ley 25.266 se le otorgó a la Dirección Nacional de Política Criminal (DNPC), creada en el año 1991 y dependiente del Ministerio de Justicia y Derechos Humanos, la potestad no sólo de diseñar y producir las estadísticas de criminalidad, sino también las estadísticas sobre el sistema penal en la Argentina. El análisis de las estadísticas sobre criminalidad puede ser dividido en dos grandes áreas: las estadísticas oficiales, que toman como fuente el registro de agencias estatales (Policía, Poder Judicial, Penitenciarias) y los estudios de victimización, que se basan en encuestas poblacionales [Perversi *et al*, 2007].

Como puede observarse en la figura 1, la DNPC cuenta con cinco sistemas de información que se utilizan como fuente para la generación de estadísticas oficiales: Encuestas de Victimización, Sistema Nacional de Información Criminal (SNIC), Sistema de Alerta Temprano (SAT), Sistema Nacional de Estadísticas Judiciales (SNEJ), Sistema Nacional de Estadísticas sobre Ejecución de la Pena (SNEEP).

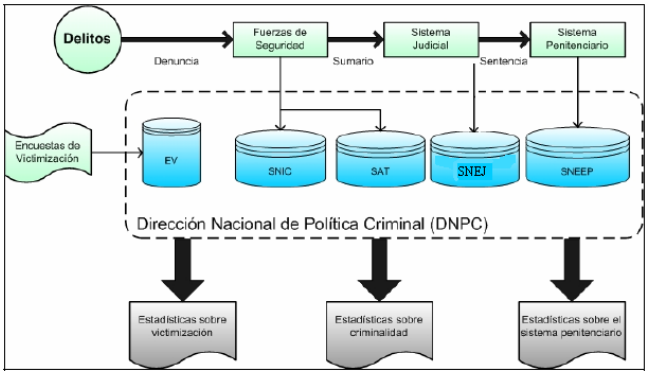

Figura 1. Flujo de información criminal. [DNPC, 2007]

3. DESCRIPCIÓN DEL PROBLEMA

3.1. Tratamiento de la Información

En tiempos en los que los volúmenes de información son muy grandes, se hace necesario un tratamiento de la información más complejo, esperando encontrar relaciones subyacentes y comportamientos en los datos que no pueden identificarse mediante un tratamiento estadístico clásico. Las técnicas estadísticas clásicas se centran generalmente en técnicas confirmatorias, mientras que las técnicas de explotación de información son generalmente exploratorias, pudiendo validar comportamientos ya conocidos como también plantear nuevas hipótesis. El creciente volumen de datos y la evolución de diversas técnicas para el descubrimiento del conocimiento hoy permiten realizar un análisis más detallado de la información disponible otorgando una mayor visión al área de "Homicidios Dolosos", "Homicidios Culposos en Accidentes de Tránsito" y "Población Carcelaria Argentina". 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 que le permita validar los conocimientos adquiridos y descubrir comportamientos desconocidos que se encuentran almacenados en sus sistemas.

3.2. Problemas Específicos

La DNPC 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". A su vez, se cree necesario llevar a cabo un estudio que involucre a la sociedad carcelaria como punto inevitable dentro de la política criminal.

Homicidios Dolosos [Perversi, 2007]: Este tipo de delito intencional 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]. Se busca poder 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.

Homicidios Culposos en Accidentes de Tránsito [Valenga, 2007]: El Plan Nacional de Seguridad Vial [ReNAT, 2007] sostiene que los accidentes de tránsito en nuestro país constituyen la cuarta causa de mortalidad luego de las enfermedades cardiovasculares, los tumores malignos y las

enfermedades cerebro vasculares. La cantidad de víctimas fatales es muy superior a las que se producen por los delitos dolosos tales como los realizados con armas de fuego (53% vs. 33%). Paradójicamente, este último problema ha generado mucha más preocupación y debate que el de los accidentes de tránsito, aunque los fallecimientos por esta causa sean muy superiores. Se busca desarrollar y proponer un modelo metodológico basado en explotación de información que permita asociar el comportamiento de diferentes variables relacionadas a los delitos provocados en accidentes de tránsito almacenados en el sistema de Alerta Temprana (SAT) que pudieran servir a los fines de la DNPC.

Población Carcelaria en Argentina [Lázaro Castillo, J. 2007; Gutierrez-Rüegg, P. *et al*; 2008; Gutierrez-Rüegg, P.; 2008]: Mientras que la cantidad de presos en establecimientos en la República Argentina continúa en ascenso continuo, lo mismo ocurre con las tasas de delitos registradas año tras año. Estamos frente a una situación en la cual los delincuentes entran a las penitenciarias para la readaptación y resocialización, y sin embargo cuando la mayoría de ellos cumplen sus penas y dejan las cárceles, vuelven a buscar al delito como medio de subsistencia. Existe dentro de estos establecimientos penitenciarios diversos factores como el ocio, la violencia, el hacinamiento, la promiscuidad, la soledad, que hacen que los detenidos sean despersonalizados moral, psíquica y físicamente. Se quiere Analizar la factibilidad de aplicar explotación de información en poblaciones carcelarias logrando entender mejor los motivos que pudieron haber motivado a esta fracción de la sociedad a delinquir y extraer conclusiones que sirvan de apoyo para la creación de políticas criminales reparadoras intramuros y preventivas fuera de las cárceles.

4. ABORDAJE DEL PROBLEMA

Se prosigue a encarar cada problemática planteada de la siguiente manera:

- Se utiliza como marco de trabajo la metodología **CRISP-DM** (*Cross Industry Standard Process for Data Mining*). La misma permite realizar el trabajo en forma ordenada y consistente pasando por cada una de sus fases: Compresión del Negocio, Comprensión de los Datos, Preparación de los Datos, Modelado, Evaluación e Implementación [Chapman, P, *et al;* 1999].
- Se busca proveer un software libre y gratuito de explotación de información que sirva de soporte a la explotación y análisis de información criminal. Se utilizan los siguientes: Weka 3.5.5 [Weka, 2007], RapidMiner [Rapid-i, 2008]
- Proceso de **Clustering** utilizando atributos significativos de cada *set* de datos. Se esperan encontrar patrones de comportamiento relevantes a través de la agrupación de casos.
- Análisis y validación de los clusters obtenidos con especialistas. Descubrimiento de comportamientos desconocidos.
- Aplicación de algoritmos de **Inducción** a cada cluster para la identificación de reglas de clasificación que ayuden a explicar la composición de cada grupo.

5. RESULTADOS EXPERIMENTALES

A continuación se procede a mostrar los resultados obtenidos aplicando explotación de información a cada una de las bases de “homicidios dolosos” [Perversi, I., et al, 2007], “homicidios culposos en accidentes de transito” [Valenga, F., 2007] y de “población carcelaria” [Lázaro Castillo, J. 2007; Gutierrez-Rüegg, P. *et al*; 2008; Gutierrez-Rüegg, P.; 2008].

5.1. Análisis de “Homicidios Dolosos”

El *data set* utilizado esta conformado por los siguientes atributos: *Provincia, Mes, Día del Mes, Día de la Semana, Hora, Arma, Lugar y Otro Delito.* En primer lugar se aplica el algoritmo *K-Means* para agrupar los 1810 registros originarios del SAT en 3 clusters (figura 2). Así se obtiene una primera caracterización de los clusters y finalmente se utiliza el algoritmo C4.5 para una interpretación formal y definitiva.

		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

Figura 2. Tabla de Centroides.

Se analizan los clusters formados mediante gráficos de barras como se muestra en la figura 3:

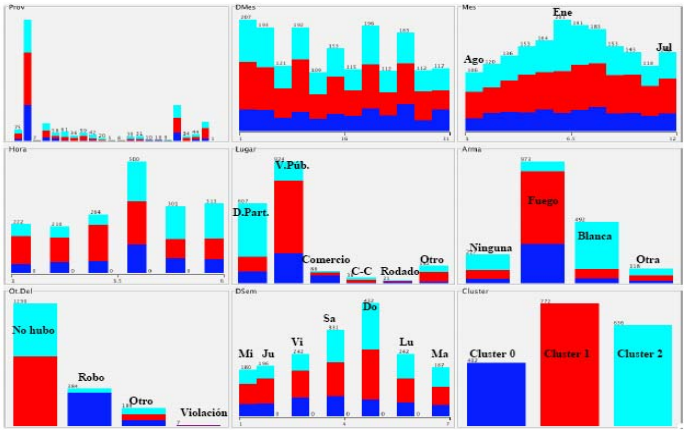

Figura 3. Distribución Clusters según Atributos

Si los clusters fueran irrelevantes, se esperaría 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*). Los atributos donde se observan más interacción entre las variables y los clusters son: *lugar, arma, otro delito* y *día de la semana*. Para estudiar estas interacciones se utilizan los gráficos de dispersión que se muestran en las figuras 4 y 5.

Existe una fuerte interacción entre *domicilio particular*, *arma blanca* y cluster 2 (verde) (Figura 4). En un nivel más general se podría 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 (rojo), aunque con cierto solapamiento con el cluster 0 (azul). Puede observarse interacción entre domicilio particular, no hubo otro delito y cluster 2 (Figura 5). 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.

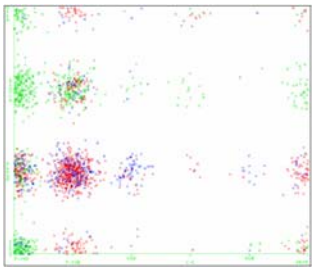

Figura 4. Interacción lugar-arma

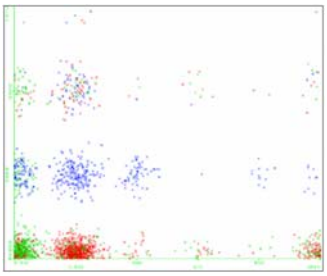

Figura 5. Interacción lugar-otro delitos

5.1.1 Interpretación de los clusters formados

En base a la información que surge del análisis se puede dar una primera interpretación a los clusters:

Cluster 0 (22%): caracterizado por homicidios mayoritariamente en ocasión de robo y con arma de fuego. En principio se tratarían 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".

5.2. Análisis de "Homicidios Culposos en Accidentes de Tránsito"

En este caso, el *data set* utilizado, proveniente del sistema SAT 2005 y ocurridos en la provincia de Buenos Aires, esta conformado por los siguientes atributos: *Mes, Día, Hora, Tipo de Arteria (calle o avenida, ruta o autopista, camino rural, otra), Existencia de intersección y finalmente, modo de colisión (Vehículo-Persona, Vehículo-Vehículo, Vehículo-Objeto, Vehículo-Despiste, Otra Colisión)*. Una vez preparado, seleccionados, limpiados y construidos los datos que fueran necesarios, se utiliza el algoritmo *K-Means* para agrupar los 881 valores resultantes del preproceso. Se obtienen 6 clusters como se muestra a continuación en la figura 6:

Cluster	Atrib. Continuos + Nominales	Distribución	Título Preeliminar
0	5.8284 4.2794 2.3382 CA/AV S VV	204 (23%)	*"Colisión de VV en CA/AV con Intersección"*
1	7.3467 3.7333 2.82 CA/AV S VP	150 (17%)	*"Colisión de VP en CA/AV con Intersección"*
2	7.123 5.1148 2.5328 RU/AU N VP	122 (14%)	*"Colisión de VP en RU/AU sin Intersección"*
3	4.79 3.96 2.39 RU/AU S VV	100 (11%)	*"Colisión de VV en RU/AU con Intersección"*
4	6.6618 3.4734 2.7053 RU/AU N VV	207 (23%)	*"Colisión de VV en RU/AU sin Intersección"*
5	6.7959 4.2041 2.1531 CA/AV N VP	98 (11%)	*"Colisión de VP en CA/AV sin Intersección"*

Figura 6. Modas de atributos nominales y promedios de atributos numéricos.

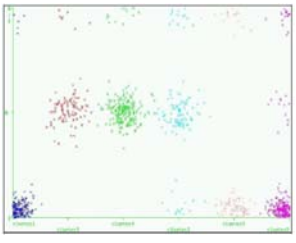

Figura 7. Relación Cluster-Intersección.

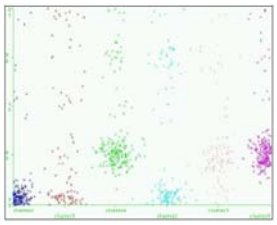

Figura 8. Relación Cluster-Modo Colisión.

Para las Intersecciones (Figura 7) se ven las siguientes distribuciones: **"Con Intersección"** (S) en los grupos *cluster0* (colisión Vehículo-Vehículo en Calle o Avenida), *cluster1* (colisión Vehículo-Persona en Calle o Avenida) y *cluster3* (colisión Vehículo-Vehículo en Ruta o Autopista). Los casos **"Sin Intersección"** (N) se agrupan en cluster2 (colisión Vehículo- Persona en Ruta o Autopista), *cluster4* (colisión Vehículo-Persona en Calle o Avenida) y *cluster5* (colisión Vehículo-Vehículo en Ruta o Autopista). Por último, en la (Figura 8), con los Modos de Colisión (no tan evidentes como los gráficos anteriores) existe una mayor concentración de las colisiones "Vehículo-Vehículo" (VV) en los grupos *cluster0* (colisión en Calle o Avenida con intersección), *cluster3* (colisión en Ruta o Autopista con intersección) y *cluster4* (colisión en Calle o Avenida sin intersección). Para las colisiones "Vehículo-Persona" (VP) los casos se concentran en *cluster1* (colisión en Calle o Avenida con intersección), *cluster2* (colisión en Ruta o Autopista sin intersección) y *cluster5* (colisión en Ruta o Autopista sin intersección). En cuanto a los tipos "Vehículo-Objeto" (VO), "Vehículo-Despiste" (VD) y "Otros" (OT) existe una distribución homogénea en los seis clusters. De esta forma, se puede asumir que los grupos reflejan comportamientos comunes dentro del total de las instancias analizadas o, dicho de otra manera, existen accidentes de tránsito con características comunes entre los casos registrados que pudieron ser eficazmente agrupados.
El paso siguiente es el de realizar la clasificación de los datos mediante métodos de inducción, con el objetivo de validar los grupos formados. Al mismo tiempo, se consulta con los especialistas en materia de accidentes de tránsito. A través de la clasificación, se obtienen las siguientes conclusiones para la Provincia de Buenos Aires en el año 2005:

- La mayor cantidad de accidentes ocurridos en Calles/Avenidas se dan en intersecciones.
- En la mayoría de los casos ocurridos en Rutas/Autopistas no hay intersecciones.
- De las colisiones vehículo-persona en Calles/Avenidas, la más frecuente se da cuando existe intersección
- En Rutas/Autopistas, es evidente el mayor grado de colisión entre vehículos (sin intersección) en comparación a la de vehículo-persona.
- Resulta llamativo, en Rutas/Autopistas, los más de 50 casos de colisión entre vehículos donde existe intersección y los 60 de vehículo-persona (en este caso sin intersección).
- Las reglas más específicas, por su escasa cantidad de instancias por caso, no prueban tener peso suficiente para considerarlas como hechos frecuentes. De todas formas pueden ser de gran utilidad en el estudio detallado de ciertas variables o en su aplicación a nivel nacional (obteniendo de esta forma, mayor cantidad de instancias en cada regla).

5.3. Análisis de la "Población Carcelaria en Argentina"

Se analizaron 50408 registros de presos masculinos pertenecientes a la base de datos "Censo Población Carcelaria" provenientes del SNEEP. Una vez que se realizada la fase de recolección, exploración y limpieza de los datos iniciales, se preparó el siguiente conjunto de datos (40928 registros) con sus respectivos atributos (tabla 1).

Edad	Estado Civil	Nivel de Instrucción	Situación Laboral
Lugar de Residencia	Capacitación Laboral	Delito Cometido	Reincidencia

Tabla 1. Atributos del Dataset

5.3.1 Resultados del Clustering de los Datos

En la tabla 2 se presentan los resultados obtenidos luego de aplicar la técnica de clusterización.

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Delito Cometido	Contra la Propiedad	Contra las Personas	Contra la Propiedad	Violación / Drogas
Nivel de Instrucción	Primario Completo	Primario Completo	Primario Completo	Primario Completo
Ultima Situacion Laboral	Tiempo Parcial	Desocupado	Desocupado	Tiempo Parcial
Capacitación Laboral	Oficio	Ni Oficio ni Profesión	Ni Oficio ni Profesión	Oficio / Profesión
Estado Civil	Soltero	Soltero	Soltero	Casado
Lugar de Residencia	Urbana	Urbana	Urbana	Urbana
Edad Promedio	31	34	27	43
Total	16849 (41%)	6513 (16 %)	14662 (36%)	2904 (7%)

Tabla 2. Centroides obtenidos mediante el clustering.

En primer lugar, y antes de analizar cada cluster por separado, se observa un *nivel de instrucción* muy pobre, en donde primario completo e incompleto agrupan al 75% de los casos. Los mismo ocurre con el atributo *último lugar de residencia*, donde aproximadamente el 90% de las instancias corresponden a Urbana.

5.3.2. Gráficos de Barras

Se puede observar (Figura 9) como se distribuyen significativamente las variables de los atributos en los distintos clusters.

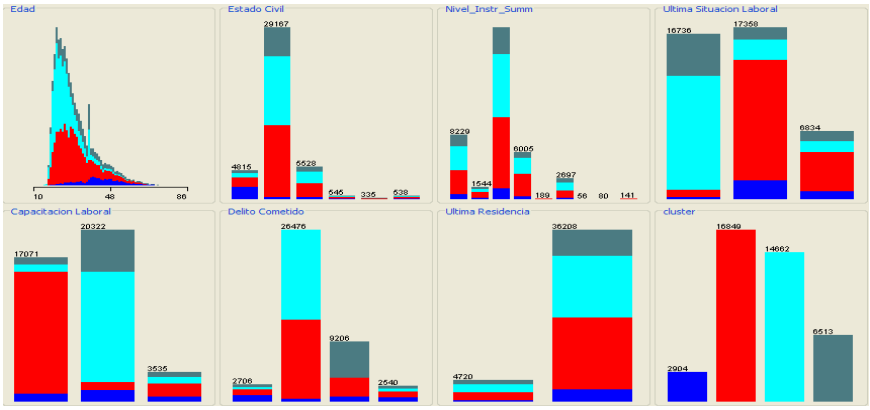

Figura 9. Distribución de los clusters entre las variables de los distintos atributos

Si bien en los atributos nivel de instrucción y ultima residencia la distribución en los clusters es irrelevante, ya que se cumple la proporción 41% rojo (cluster 0) 16% gris (cluster 1) 36% turquesa (cluster 2) 7% azul (cluster 3), en los otros atributos se pueden encontrar interacciones significativas.

5.3.3. Gráficos de dispersión

Tres de los atributos más significativos son el *delito cometido*, la *última situación laboral* y la *capacitación laboral***.** Tal como puede observarse en la figura 10, existe una interacción importante entre el cluster 2 (verde), no tiene oficio ni profesión y delito contra la propiedad. A su vez, en la figura 11 puede se aprecia que el cluster 2 esta caracterizado por personas desocupadas, mientras que el cluster 0 (rojo) concentra mayor cantidad de instancias en la situación de oficio de tiempo parcial/completo. El delito que caracteriza al cluster 0 es contra la propiedad (figura 10).

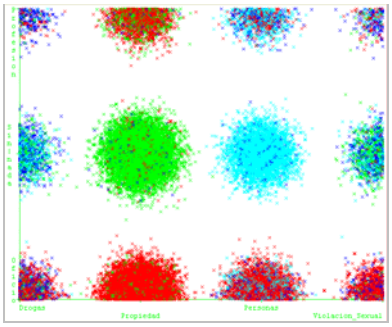

Figura 10. Distribución según Delito-Capacitación

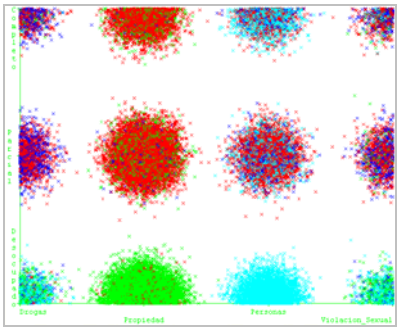

Figura 11. Distribución según Delito-Situación Laboral

En lo que respecta al cluster 1 (turquesa) se observa interacción con delito contra las personas y sin oficio ni profesión (figura 5.9), en su mayoría se trata de personas desocupadas (figura 5.10). A su vez, el cluster 3 (azul) se distribuye en delitos contra la integridad sexual, estupefacientes y en menor medida delitos contra las personas (figura 5.9). Generalmente se observan que son personas con trabajos de tiempo parcial o completo.

5.3.4. Primera Interpretación de los Clusters

Cluster 0 (41%): el que más registros agrupa**,** podría tratarse de personas, que aun cuando cometieron **delito contra la propiedad (mayormente robo)** no hay patrones que indiquen que lo hayan hecho por una necesidad marcada. Son personas que trabajan parcialmente y tenían algún oficio. Generalmente reincidentes, y que sus salarios les alcanza para lo básico, teniendo que salir a robar para complementar sus necesidades y la de sus hijos.
Cluster 1 (16%): cluster que agrupa a los presos que delinquieron **contra las personas**. Generalmente desocupados y sin oficio ni profesión, estaría caracterizado por un lado por las personas que cometieron homicidio en ocasión de robo. A su vez, podría decirse que al ser personas totalmente inactivas a una edad en donde conseguir trabajo se les hace casi imposible, pueden haber llegado a delinquir contra las personas en una reacción de emoción violenta.
Cluster 2 (36%): segundo grupo en importancia, agrupa a los jóvenes que no tienen estudios, trabajo, ni profesión alguna. En un principio, este estado de exclusión de un régimen laboral los llevaría a **robar y/o hurtar** como única salida para poder subsistir. Importante tener en cuenta que es uno de los clusters más preocupantes por tratarse de gente joven, en donde las drogas pesadas juegan un papel muy importante.
Cluster 3 (7%): cluster más difícil de interpretar, agrupa en mayor medida a las personas que cometieron delitos contra la **Integridad Sexual** y por quienes fueron procesados o condenados por **Estupefacientes**. Con un promedio de edad mayor a los 40 años y casados o en concubinato, no se observan patrones que los hayan llevado a delinquir por una necesidad especifica ya que se trata de personas con algún tipo de empleo de tiempo parcial o completo.

6. CONCLUSIONES

Queda demostrado que la explotación de información es una herramienta muy potente que permite explorar grandes bases de datos de manera rápida y eficiente, sin necesidad de ser un experto en el tema a investigar. En los 3 casos analizados, no se tenía entrenamiento ni experiencia alguna en las técnicas utilizadas como tampoco del tema encarado, y pudieron de todas formas obtenerse interacciones interesantes que no llegan a observarse a simple vista.

En particular, el uso combinado de técnicas de explotación de información ha demostrado ser útil en el descubrimiento de patrones de comportamiento en los registros de **"Homicidios Culposos por Accidentes de Tránsito"** [Valenga, F. 2007]. Con el agrupamiento o clustering se pudieron centralizar grupos con características comunes alcanzando una visión más clara y directa de los comportamientos almacenados en el SAT. A partir de estos grupos, se llevó a cabo una selección de atributos para descartar aquellos que puedan perjudicar la generación de reglas, tarea que se realiza con la técnica de clasificación. En este sentido, se obtuvo un porcentaje de instancias clasificadas correctamente que superó el 98% y se generaron 6 reglas generales con gran una cantidad de instancias en cada una, como así también 10 reglas específicas con escaso soporte.
En cuanto a los "**Homicidios Dolosos"** [Perversi, I. 2007], 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.
Finalmente, y haciendo referencia a la **"Población Carcelaria"** [Lázaro Castillo, J. 2007; Gutierrez-Rüegg, P. *et al*; 2008; Gutierrez-Rüegg, P.; 2008], se ha demostrado la factibilidad y valor agregado al aplicar sistemas inteligentes de procesamiento de información, como la explotación de información, a la información criminal en Argentina. En particular, se destacan los resultados obtenidos al utilizar métodos de explotación de información en poblaciones carcelarias con el fin de extraer conocimiento de los datos. Los resultados experimentales lograron obtener 4 grupos de presos con distintas características y conductas permitiendo validar conocimientos preexistentes.
Lógicamente, previo a la aplicación de la explotación de información, se tuvo que realizar una fuerte investigación sobre estadísticas vinculadas al crimen, al servicio penitenciario y a los procesos penales en Argentina. Esto fue el punto de partida del proyecto, el cual basa su factibilidad en los siguientes puntos:

- Existe muchísima información que no esta siendo aprovechada para la extracción de conocimiento útil (patrones, comportamientos, anomalías, etc.) en toda su dimensión.
- Existen programas de explotación de información de libre uso, relativamente fáciles de aprender y utilizar.
- No es necesario ser un experto ni en temas de sistemas inteligentes ni en el tema a explorar.
- Sin embargo, es condición necesaria contar con la experiencia de especialistas en la temática a analizar que validen las conclusiones extraídas de los modelos de explotación de información.

7. FINANCIAMIENTO Y AGRADECIMIENTOS

Este proyecto ha sido financiado parcialmente con subsidios UBACyT 2004-2007-I050, UBACyT 2008-2010-I012 y ANPC BID 1728/OC-AR PICT 02-13533. Los autores desean agradecer a la Secretaría de Política Criminal de la Nación por el apoyo proporcionado al facilitar las bases de datos con las cuales se desarrolla la línea de investigación en la que se inserta este proyecto.

8. REFERENCIAS

Britos, P., Hossian, A., García-Martínez, R. y Sierra, E., 2005. *Minería de Datos Basada en Sistemas Inteligentes*. Editorial Nueva Librería. Buenos Aires. ISBN 987-1104-30-8.

Cartagenova, S. G., 2005. *Detección Automática de Reglas de Asociación*. Trabajo Final de Especialidad en Ingeniería en sistemas Expertos, Instituto Tecnológico de Buenos Aires (ITBA).

Chapman, P., Clinton, J., Keber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. (1999). *CRISP-DM 1.0 Step by step BIguide*. Edited by SPSS. http://www.crisp-dm.org/ CRISPWP-0800.pdf. Ultimo acceso Junio 2008.

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

Chen, H., Chung, W., Xu, J., Wang, G., Qin, Y., Chau, M., 2004. *Crime Data Mining: A General Framework and Some Examples*. IEEE Computer Society, vol. 37, no.4, pp. 50-56.*data mining*. Cambridge (Massachussets): AAAI/MIT Press.

Dirección Nacional de Política Criminal. *http://wwwpolcrim.jus.gov.ar/.* Acceso Octubre 2007.

Fayyad, U.M., Piatetsky-Shapiro G., Smyth, P., Uthurusamy, R., 1996. *Advances in Knowledge and*

Frawley W., Piatetsky-Shapiro G., Matheus C. *Knowledge Discovery in Databases: An Overview.* AI Magazine: pp. 213-228. ISSN 0738-4602.

Gutiérrez Rüegg, P., Merlino, H., Rancan, C., Procopio, C., Rodríguez, D., Britos, P., García-Martínez, R. (2008). *Identificación de Patrones Característicos de la Población Carcelaria Mediante Minería de Datos*. Proceedings X Workshop de Investigadores en Ciencias de la Computación. Pág. 461-465. 978-950-863-101-5.

Gutierrez-Rüegg, P. 2008. Tesis de Grado en Ingeniería Industrial del Instituto Tecnológico de Buenos Aires en el área *Caracterización de la Población Carcelaria en Argentina Mediante la Aplicación de Minería de Datos para la Prevención de Hechos Delictivos*.

Han, J. y Lamber, M., 2001. *Data Mining: Concepts and techniques.* Morgan Kauffmann Publishers. Edición. 2001.

Hartigan, J.A., 1975. *Clustering algorithms*. John Wiley & Sons, New York.

Kantardzic, M., 2003. *Data Mining: Concepts, Models, Methods, and Algorithms*. John Wiley & Sons. ISBN 0471228524.

Lázaro Castillo, J. 2007. *Caracterización de la Delincuencia en Argentina a Través de Sistemas Inteligentes*. Tesis de Grado en Ingeniería Informática. Escuela Universitaria de Informática de la Universidad Politécnica de Madrid.

Perversi, I. 2007. *Aplicación de Minería de Datos para la Exploración y Detección de Patrones Delictivos en Argentina.* Tesis de Grado en Ingeniería Industrial. ITBA.

Perversi, I., Valenga, F., Fernández, E., Britos P., García-Martínez, R. 2007. *Identificación y detección de Patrones Delictivos basada en Minería de Datos*. Proceedings IX Workshop de Investigadores en Ciencias de la Computación. Pag. 385-389. ISBN 978-950-763-075-0.

Quinlan, J., 1993. *Programs for Machine Learning.* Morgan Kaufmann Publishers. Edición 1993.

Registro Nacional de Antecedentes de Transito. *Plan Nacional de Seguridad Vial*. http://www.renat .gov.ar/plan_nacional.pdf . Acceso Septiembre 2007.

Servente, M.; García-Martínez, R., 2002. *Algoritmos TDIDT Aplicados a la Minería Inteligente.* http://www.fiuba.ar/laboratorios/lsi/R-ITBA-26-datamining.pdf Acceso Enero 2008.

Valenga, F. 2007. *Minería de datos criminales. Aplicación a homicidios en accidentes de tránsito.* Tesis de Grado en Sistemas. Universidad de Morón.

Valenga, F., Fernández, E., Merlino, H., Rodríguez, D., Procopio, C., Britos, P., García-Martínez, R. 2008. *Minería de Datos Aplicada a la Detección de Patrones Delictivos en Argentina*. Proceedings VII Jornadas Iberoamericanas de Ingeniería de Software e Ingeniería del Conocimiento. Pág. 31-39. ISSN 1390-292X.

Valenga, F., Perversi, I., Fernández, E., Merlino, H., Rodríguez, D., Britos, P., García-Martínez, R. 2007a. *Estudio Preliminar: La Estadística Criminal y el Aporte de la Minería de Datos.* En Kaminsky, G., Kosovsky,D., Kessler, G. El Delito en la Argentina Post-crisis. pp 11-24. Editado por la Friedrich Ebert Stiftung.

Valenga, F., Perversi, I., Fernández, E., Merlino, H., Rodríguez, D., Britos, P., García-Martínez, R. 2007b. *Aplicación de Minería de Datos para la Exploración y Detección de Patrones Delictivos en Argentina*. Anales del XIII Congreso Argentino de Ciencias de la Computación. Pag. 258-270. ISBN 978-950-656-109-3.