DOI: http://dx.doi.org/10.22267/rus.171903.101

Universidad y Salud
ARTÍCULO ORIGINAL

Aplicación de los árboles de decisión en la identificación de patrones de lesiones fatales por causa externa en el municipio de Pasto, Colombia

Application of decision trees in the identification of patterns of fatal injuries by external cause in the municipality of Pasto, Colombia

Ricardo Timarán-Pereira[1*] orcid.org/0000-0002-0006-6654

Andrés Calderón-Romero[2] orcid.org/0000-0002-7396-413X

Arsenio Hidalgo-Troya[3] orcid.org/0000-0003-4080-118X

1 Grupo de Investigación GRIAS, Facultad de Ingeniería, Universidad de Nariño. San Juan de Pasto, Colombia
2 The University of Twente. Enschede, Países Bajos
3 Departamento de Matemáticas y Estadística, Facultad de Ciencias Naturales, Universidad de Nariño. San Juan de Pasto, Colombia

Fecha de recepción: Febrero 9 - 2017 Fecha de revisión: Agosto 9 - 2017 Fecha de aceptación: Diciembre 1- 2017

Timarán-Pereira R, Calderón-Romero A, Hidalgo-Troya A. Aplicación de los árboles de decisión en la identificación de patrones de lesiones fatales por causa externa en el municipio de Pasto, Colombia. Univ. Salud. 2017;19(3):388-399. DOI: http://dx.doi.org/10.22267/rus.171903.101

Resumen

Introducción: La Organización Panamericana de la Salud (OPS) desde el año 1993 y la Organización Mundial de la Salud (OMS) en 1996, aceptaron que la violencia es un problema de salud pública, situación que se corrobora en el Informe de Violencia y Salud, en el cual América Latina presentó una tasa de homicidios de 18 por cada 100.000 personas, y es considerada como una de las regiones más violentas del mundo. **Objetivo**: Detectar patrones delictivos con técnicas de minería de datos en el Observatorio del Delito del municipio de Pasto (Colombia). **Materiales y métodos**: Se aplicó Cross Industry Standard Process for Data Mining (CRISP-DM), una de las metodologías utilizadas en el desarrollo de proyectos de minería de datos en los ambientes académico e industrial. La fuente de información fue el Observatorio del Delito del municipio de Pasto, donde está almacenadas las cifras históricas, limpias y transformadas sobre las lesiones de causa externa (fatales y no fatales), registrados en 11 años. **Resultados:** Se construyó un modelo de clasificación basado en árboles de decisión que permitió descubrir patrones de muertes por causa externa. Para el caso de homicidios, estos sucedieron en su mayoría en la Comuna 5 de Pasto, los fines de semana, en la madrugada, en el segundo semestre del año, en la vía pública y las víctimas fueron hombres adultos, de oficios varios, la causa de los homicidios fueron riñas y se produjeron con arma de fuego. **Conclusión:** El conocimiento generado ayudará a los organismos gubernamentales y de seguridad a tomar decisiones eficaces en lo relacionado a la implementación de planes de prevención de delitos y seguridad ciudadana.

Palabras clave: Reconocimiento de normas patrones automatizadas; minería de datos; árboles de decisión; clasificación. (Fuente: DeCS, Bireme).

Abstract

Introduction: The Pan American Health Organization (PHO) and the World Health Organization (WHO) accepted, since the year 1993 and 1996 respectively, that violence is a public health problem, a situation that is corroborated in the report on violence and health, in which Latin America presented a homicide rate of 18 per 100,000 people, and it is considered one of the most violent regions in the world. **Objective:** To detect criminal patterns with data mining techniques in the Crime Observatory of the municipality of Pasto (Colombia). **Materials and methods:** Cross Industry Standard Process for Data Mining (CRISP-DM) was applied, which is one of the methodologies used in the development of data mining projects in academic and industrial environments. The source of information was the

***Autor de correspondencia**
Ricardo Timarán Pereira
e-mail: ritimar@udenar.edu.co

Crime Observatory of the municipality of Pasto, where the historical clean and transformed figures on the injuries of external cause (fatal and nonfatal) recorded in 11 years are stored. **Results:** A decision tree-based classification model was built that allowed the discovery of patterns of deaths from external causes. In the case of homicide, these happened mostly in the commune 5 in Pasto under the following circumstances: during the weekends, in the early morning, in the second semester of the year and in the public thoroughfare; besides, the victims were adult men of various professions; and the cause of the homicides were quarrels and they were produced with a fire gun. **Conclusion:** The generated knowledge will help government and security agencies make effective decisions regarding the implementation of crime prevention and citizen security plans.

Keywords: Pattern recognition, automated; data mining; decision trees; classification. (Source: DeCS, Bireme).

Introducción

Los ataques terroristas del 11 de septiembre de 2001 han aumentado significativamente la preocupación por la seguridad interna en todo el mundo. Organizaciones como la Agencia Central de Inteligencia (CIA, del inglés *Central Intelligence Agency*) o el Buró Federal de Investigaciones (FBI, del inglés *Federal Bureau of Investigation*) procesan y analizan información activamente en busca de actividad terrorista[1]. El análisis de los registros criminales es fundamental en la prevención del delito porque conlleva al diseño de políticas y planes de prevención efectivos[2].

Las lesiones de causa externa son definidas como el daño o lesión en una persona en forma intencional o no intencional, que puede originarse por un traumatismo, envenenamiento, agresión, accidentes, etc., ser mortal (lesión fatal) o no conducir a la muerte (lesión no fatal)[3]. Las lesiones por causa externa son consideradas desde hace dos décadas como un problema sanitario a nivel mundial. Según datos de la OMS, aproximadamente 5,8 millones de personas mueren por año por estas causas, cerca de 16000 personas al día, representando cerca del 10% del total de las muertes que se registran en el mundo, 32% más que el número de muertes que resultan por malaria, tuberculosis y el VIH/SIDA. Por cada persona que muere por esta causa, hay miles más lesionadas, muchas de ellas con secuelas permanentes[3].

En América Latina desde el año de 1993 se han establecido claros esfuerzos por anticipar acciones a la violencia marcada en todos los países, esto debido a que la OPS y la OMS presentaron estadísticas que determinaron que Latinoamérica es una de las regiones más violentas del mundo[4]. De acuerdo con estimativos de los años 2003 y 2005 algunos países como Colombia, Venezuela, El Salvador, Guatemala y Honduras reportan altos índices de homicidios con tasas iguales o superiores a 29 por 100.000 personas, mientras que Costa Rica reportó la menor tasa de Centroamérica con 9,2 homicidios por cada 100.000 personas[5].

En América Latina se reconoce que la ausencia de información confiable y oportuna es una limitante para avanzar en la identificación de la magnitud y características de las diferentes formas en que se expresa la violencia, así como el monitoreo y evaluación de los programas y proyectos para su prevención y control[4].

En Colombia una de las estrategias implementadas en vigilancia en salud pública, corresponde a los Observatorios de Muertes de Causa Externa[5], los cuales se han instaurado para el seguimiento y análisis en el nivel local (municipal) en casos de mortalidad por causa externa como: homicidios, suicidios, eventos de tránsito y muertes no intencionales. Gracias a ello, se cuenta con diversas experiencias en el ámbito municipal y departamental. Particularmente en el municipio de Pasto (Departamento de Nariño), el observatorio de muertes por causa externa, denominado: "Observatorio del Delito", nace en el segundo semestre del año 2002, como resultado de un proyecto conjunto con el Programa Colombia de la Universidad de Georgetown[6].

A pesar que se conoce cuál es el problema, las cifras disponibles no son de buena calidad. En la

mayoría de ocasiones es necesario desarrollar estudios específicos para conocer el problema o son datos oficiales que presentan subregistros o no coinciden al compararlos con otras fuentes de información. Por lo tanto, es necesario recopilar datos e información que permitan mejorar el conocimiento sobre la magnitud y las características de los hechos, orientar estudios para identificar los factores que inciden en la presencia o no del evento, evaluar políticas e intervenciones y hacer difusión de las mismas[4].

Mientras que la estadística plantea hipótesis que deben ser validadas a partir de las cifras disponibles, la minería de datos descubre patrones que mediante su interpretación permiten, en el caso del Observatorio del Delito, caracterizar los diferentes tipos de muertes por causa externa no previstos desde la estadística. En este contexto, la minería emerge como el siguiente paso evolutivo en el proceso de análisis de datos criminales[2,7]. Entre las experiencias de aplicación en la detección de patrones delictivos, está la del Departamento de policía de Ámsterdam, que utilizó el software *DataDetective*[8] junto con *Mapinfo* para el análisis de registros criminales. Las principales técnicas empleadas son árboles de decisión y redes neuronales de *backpropagation*. En una bodega de datos se ha unificado varias bases policiales junto con información externa (clima, variables socioeconómicas y demográficas). Esto permite la 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 (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)[9] ha desarrollado una aplicación para el análisis de información criminal que combina minería de datos, junto a un entorno visual y una interfaz. El principal objetivo fue optimizar la alocación de recursos con 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 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).

El proyecto COPLINK[10], creado en el año 1997, une la experiencia técnica del Laboratorio de Inteligencia Artificial de la Universidad de Arizona con el conocimiento de la aplicación de la ley del Departamento de Policía de Tucson (USA). COPLINK sirve a la comunidad al superar la brecha entre la realización de investigaciones en tecnologías de vanguardia y resolver problemas del mundo real, ayudando a los policías a combatir el crimen. Está compuesto por dos sistemas integrados: *Coplink Connect* y *Coplink Detect*, el primero busca compartir información criminal entre distintos departamentos policiales, mediante un fácil acceso y una interface sencilla, integrando distintas fuentes de información. El segundo está diseñado para detectar de forma automática distintos tipos de asociaciones entre las bases mediante técnicas de minería de datos[10,11].

En Argentina, Valenga *et al.*[12], aplicaron las técnicas de *clustering* utilizando el algoritmo *k-means* para la detección de patrones de homicidios dolosos. Se analizaron 1810 registros de la base de datos "homicidios dolosos" correspondientes a la totalidad de hechos registrados durante el año 2005. Se obtuvieron tres clusters: Cluster 0 (22%): caracterizado por homicidios mayoritariamente en ocasión de robo y con arma de fuego. Cluster 1 (43%): agrupa más registros y es el más parecido a la media global, caracterizado por homicidios principalmente en la vía pública con arma de fuego y sin la existencia de otro delito. 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.

En Chile, Montt *et al.*[13], reportan el análisis de accidentes de tránsito con inteligencia computacional cuyo objetivo fue estimar el número de personas lesionadas y fallecidas en accidentes de tránsito, relacionado con las causas que producen el accidente. Se utilizaron redes neuronales artificiales en combinación con algoritmos de técnicas de inteligencia artificial. Si bien en Latinoamérica se han adelantado estudios de minería de datos relacionados con actos criminales[12,13], en Colombia no se reportan estudios que apliquen esta técnica para detectar patrones de muertes por causa externa.

En cuanto a las metodologías para desarrollar análisis de minería de datos y en un intento de normalización del proceso, de forma similar a como se hace en ingeniería para normalizar el proceso de desarrollo software, surgieron a finales de los 90 dos metodologías principales: CRISP-DM (*Cross Industry Standard Process for Data Mining)* y SEMMA (*Sample, Explore, Modify, Model, and Assess*). Ambas especifican las tareas a realizar en cada fase descrita por el proceso, asignando tareas concretas y definiendo lo que es deseable obtener tras cada fase. Azevedo y Santos[14] comparan ambas implementaciones y llegan a la conclusión de que, aunque se puede establecer un paralelismo claro entre ellas, CRISP-DM es más completo porque tiene en cuenta la aplicación al entorno de negocio de los resultados, y por ello es la que se adoptó popularmente. En encuestas realizadas en KDNuggets en 2002, 2004, 2007 y 2014 se comprobó que CRISP-DM era la principal metodología utilizada, cuatro veces más que SEMMA[14]. La metodología CRISP-DM para proyectos de minería de datos no es la "más actual" o "la mejor", pero es muy útil para comprender esta tecnología o extraer ideas para diseñar o revisar métodos de trabajo para proyectos de similares características.

En este artículo se presentan los resultados del proyecto de investigación que tuvo como objetivo detectar patrones delictivos con técnicas de minería de datos en el Observatorio del Delito del municipio de Pasto. Como resultado de este, actualmente el Observatorio cuenta con un sistema de vigilancia de eventos violentos a partir de la implementación de un sistema de información georreferenciado denominado SIGEODEP[15], soportado en un mercado de datos (*datamart*)[16], donde se encuentra almacenada información histórica, limpia y transformada sobre lesiones de causa externa fatales y no fatales registradas en un periodo de 11 años, que permite disponer de información confiable, oportuna, de buena calidad y representativa de las lesiones de causa externa que ocurren en el municipio de Pasto.

Materiales y métodos

Se aplicó CRISP-DM que comprende seis fases: Análisis del problema, análisis de los datos, preparación de los datos, modelado, evaluación y explotación[17]. Se utilizó como fuente de información, el *datamart* construido en el proyecto[16], donde se almacena la información registrada en el Observatorio del Delito en un periodo de 11 años. Se construyó un modelo de clasificación basado en árboles de decisión, utilizando el algoritmo J48 de la herramienta Weka[18]. Se escogió este algoritmo por su simplicidad y facilidad para interpretar los patrones y por ser el más utilizado para este tipo de problemas[19,20].

En la fase de análisis del problema se identificó con exactitud la problemática que se solucionaría utilizando la minería de datos, esto permitió recolectar la información necesaria para interpretar con asertividad los resultados encontrados[18,21]. En la fase de análisis de los datos se realizó la recolección inicial de datos, para establecer un primer contacto con el problema, familiarizarse con ellos, identificando su calidad y establecer las relaciones más evidentes que permitieron definir las primeras hipótesis.

En la fase de preparación se seleccionó los datos a los cuales se les aplicaría una determinada técnica de modelado, limpieza, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato[18,21]. En la fase de modelado se seleccionaron las técnicas más apropiadas para el proyecto. En la fase de evaluación se verificó si el modelo se ajusta a las

necesidades establecidas en el proyecto. Se evaluaron los patrones encontrados con el fin de determinar su validez, remover los redundantes o irrelevantes y traducir los patrones útiles en términos que sean entendibles para el usuario. Finalmente, en la fase de implementación, el conocimiento obtenido se transformó en acciones dentro del proceso. Se trató de explotar la potencialidad de los modelos, integrarlos en los procesos de toma de decisión de la organización y difundir informes sobre el conocimiento extraído[18,21].

Resultados

Teniendo en cuenta las fases de la metodología CRISP-DM los siguientes son los resultados de cada una de las etapas:

Análisis del problema

Según el censo general realizado en Colombia en el año 2005, el municipio de Pasto contaba con 431.141 habitantes, de los cuales 382.618 pertenecían a la zona urbana y 48.500 a la parte rural[6]. En el periodo comprendido entre 2003 y 2013, las cifras registradas en el Observatorio del Delito del municipio de Pasto de muertes por causa externa como homicidios, suicidios, muertes en eventos de tránsito y muertes no intencionales muestran un comportamiento similar. El promedio de casos para homicidios es de 110, para accidentes de tránsito es de 50, para suicidios es de 40 y para muertes no intencionales principalmente por caídas o por asfixia (bronco aspiraciones) es de 55. La tasa promedio por cada 100.000 habitantes para homicidios es de 26, para accidentes de tránsito 12, para suicidios 9 y para muertes no intencionales 13.

Análisis de los datos

Del *datamart*[16] del Observatorio del Delito del municipio de Pasto, diseñado bajo el modelo multidimensional, se obtuvieron únicamente los datos de las muertes por causa externa desde el año 2003 hasta el año 2013. Después de consultar el número de caso de las tablas de hechos para los cuatro eventos en cuestión (Figura 1), se cuenta con el siguiente número de registros: homicidios: 1285; suicidios: 482; accidentes de tránsito: 287 y muertes no intencionales: 144. Sin embargo, no todos los eventos comparten el mismo número de atributos. Una primera selección de atributos se ejecutó para descartar aquellas variables utilizadas como llaves foráneas y otras que se usaban como indicadoras de orden dentro del *datamart*. Con el fin de obtener los atributos más representativos para el estudio, se realizó un ranking inicial de importancia de los atributos basado en la ganancia de información de cada uno con respecto al atributo clase, que en este caso es *Injury* (tipo de lesión). Como resultado se obtuvo un repositorio inicial con 45 atributos y 2198 registros, el cual se denominó T2198A45 y sirvió de base para las siguientes etapas de este proceso.

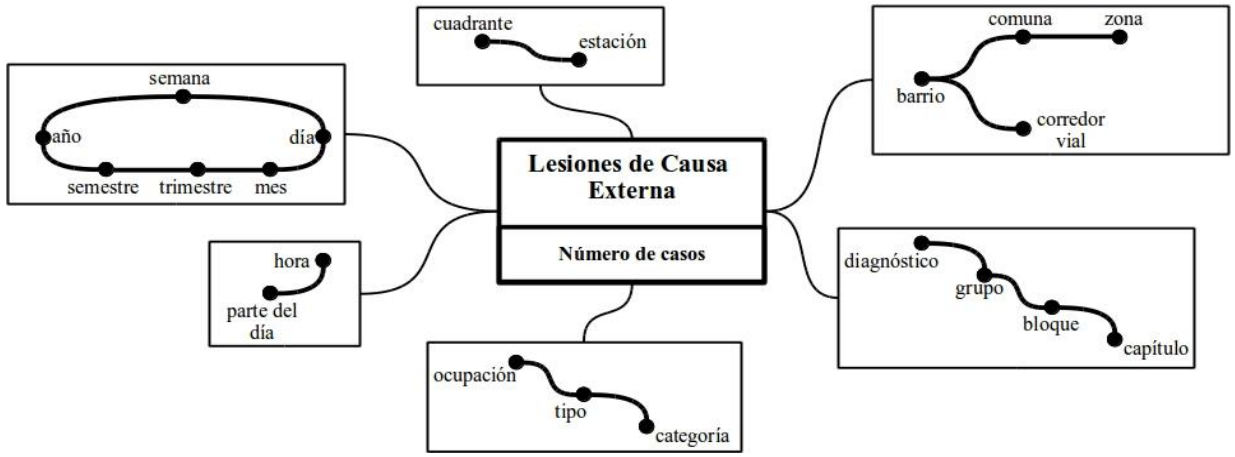

Figura 1. Topología en estrella del mercado de datos[12]

Preparación de los datos

Se realizó un análisis de la calidad de los datos del repositorio T2198A45, donde se identificó por cada atributo el número de valores distintos, el número de valores nulos, el valor máximo, valor mínimo, moda, media y un histograma para determinar cuáles técnicas de limpieza de datos se deben aplicar.

Teniendo en cuenta el análisis de datos y por la imposibilidad de encontrar sus valores con fuentes externas, se eliminaron los atributos con un alto porcentaje de nulos y los que presentaban un alto porcentaje con el valor "SIN DATO". Se descartaron aquellos atributos utilizados como llaves foráneas y otras que se usaban como indicadoras de orden dentro del *datamart*. Como resultado de este proceso, se seleccionaron los 12 más significativos, organizándolos en atributos de lugar, tiempo y víctima, se construyó con estos el conjunto de datos T2198A12 sobre el cual se aplicaron las técnicas de modelado. Todos estos atributos son de tipo categórico y se muestran junto con su descripción y su agrupación en la Tabla 1.

Tabla 1. Atributos conjunto de datos final T2198A12

	Atributos	Descripción	Grupo
1	Injury	Atributo clase	Clase
2	date_day_name	Día de la semana	Tiempo
3	date_month_name	Mes en que ocurrió la lesión	Tiempo
4	date_quarter_name	Trimestre en que sucedió el evento	Tiempo
5	date_semester_name	Semestre en que sucedió el evento	Tiempo
6	time_group	Parte del día en que ocurrió la lesión {madrugada, mañana, tarde o noche}	Tiempo
7	neighborhood_suburb_name	Comuna o Corregimiento donde ocurrió la lesión	Lugar
8	fatal_place_name	Lugar donde ocurrió la lesión {Casa, espacio o vía pública, lugar de trabajo, ...}	Lugar
9	victim_age_groups	Edad de la víctima {niño, joven, adulto, adulto mayor}	Víctima
10	victim_gender	Sexo de la víctima	Víctima
11	residence_neighborhood_zone_name	Zona de residencia de la víctima {rural o urbana}	Víctima
12	job_jobs_category_name	Ocupación de la víctima	Víctima

Modelado

Con el fin de descubrir patrones delictivos de muertes por causa externa, se seleccionó el modelo de clasificación basado en árboles de decisión. La clasificación con árboles de decisión considera clases disjuntas, de forma que el árbol conducirá a una y solo una hoja, asignando una única clase a la predicción[22]. Después de construido el modelo servirá para determinar en nuevos casos, el tipo de delito que causó la muerte[19,20]. Para esta tarea, se escogió como clase el atributo *Injury* que determina si la muerte fue por un homicidio, un accidente de tránsito, un suicidio o una muerte no intencional.

Para la poda del árbol se tuvo en cuenta el factor de confianza C (*confidence level*), que influye en el tamaño y capacidad de predicción del árbol construido. El valor por defecto de este factor es del 25% y conforme este va disminuyendo, se permiten más operaciones de poda y por lo tanto llegar a árboles cada vez más pequeños[23]. Otro factor que se utilizó para variar el tamaño del árbol fue a través del parámetro M que especifica el mínimo número de instancias o registros por nodo del árbol[24].

Para evaluar la calidad del modelo, se dividió el repositorio de datos en dos conjuntos: entrenamiento y prueba, se escogió el método validación cruzada con *n* pliegues (*n-fold cross validation*). Este método consiste en dividir el conjunto de entrenamiento en *n* subconjuntos disjuntos de similar tamaño llamados pliegues (*folds*) de forma aleatoria. Posteriormente se realizaron *n* iteraciones (igual al número de subconjuntos definido), donde en cada una se reservó un subconjunto diferente para el conjunto de prueba y los restantes *n-1* (uniendo todos los datos) para construir el modelo (entrenamiento). En cada iteración se calculó el error de muestra parcial del modelo. Por último se construyó el modelo con todos los datos y se obtuvo su error promediando los obtenidos anteriormente en cada una de las iteraciones[22].

Con el fin de detectar los patrones más confiables de muertes por causa externa, utilizando árboles de decisión con el repositorio T2198A12 y tomando como clase el atributo *Injury*, se realizaron diferentes pruebas para obtener los mejores parámetros de poda y por ende obtener el árbol con el mayor número y porcentaje de instancias correctamente clasificadas. Para evaluar la calidad del modelo y su validez se escogió el método de validación cruzada con 10 pliegues por presentar los mejores resultados obtenidos. Se utilizaron criterios de pospoda para dejar las mejores ramas del árbol y las reglas de clasificación más representativas, estos criterios fueron: el soporte, que es el porcentaje mínimo de registros del total de casos que deben estar en los nodos; y la confianza que es el porcentaje de casos por nodo correctamente clasificados con relación a los incorrectamente clasificados en el mismo nodo.

Se generaron 50 árboles variando el factor de confianza C de 0,1 hasta 0,5 con un incremento de 0,1 y el número de instancias por nodo M de 2 en 2 iniciando en 2 hasta 20. En la Figura 2 se muestra la precisión del árbol y su matriz de confusión. En la Figura 3 se puede visualizar el mejor árbol.

```
J48 pruned tree
------------------
Number of Leaves:     138

Size of the tree:         155

=== 10 Fold Cross Validation ===

=== Summary ===

Correctly Classified Instances        1810               82,3476 %
Incorrectly Classified Instances       388               17,6524 %
Kappa statistic                          0,6761
Mean absolute error                      0,1314
Root mean squared error                  0,2666
Relative absolute error                 44,6229 %
Root relative squared error             69,4944 %
Coverage of cases (0.95 level)          97,0428 %
Mean rel. region size (0.95 level)      46,4968 %
Total Number of Instances             2198

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0,368    0,016    0,624      0,368   0,463      0,452    0,854     0,395     fact_accidents
                 0,938    0,318    0,806      0,938   0,867      0,655    0,866     0,863     fact_murder
                 0,896    0,009    0,966      0,896   0,930      0,912    0,977     0,959     fact_suicides
                 0,418    0,027    0,702      0,418   0,524      0,492    0,868     0,488     fact_traffic
Weighted Avg.    0,823    0,192    0,816      0,823   0,810      0,677    0,890     0,804

=== Confusion Matrix ===

a   b   c   d   <-- classified as
 53  76   6   9 |   a = fact_accidents
 32 1205  9  39 |   b = fact_murder
  0  47  432  3 |   c = fact_suicides
  0 167   0 120 |   d = fact_traffic
```

Figura 2. Matriz de Confusión y precisión del mejor árbol

Evaluación

Analizando los resultados de las pruebas de clasificación con árboles de decisión realizadas con el conjunto de datos de lesiones fatales T2198A12, en el cual se almacenan los datos de 2198 casos de muertes fatales distribuidos en homicidios: 1285; suicidios: 482 casos; accidente de tránsito: 287 y muertes no intencionales: 144, se puede observar que el árbol de decisión construido con un factor de confianza C=0,5 y un número de casos por nodo M=8 es el mejor (Figura 2), con 1810 instancias correctamente clasificadas, que corresponden a un porcentaje de precisión del 82,35% y 388 instancias incorrectamente clasificadas, correspondiente a un porcentaje de error del 17,65%.

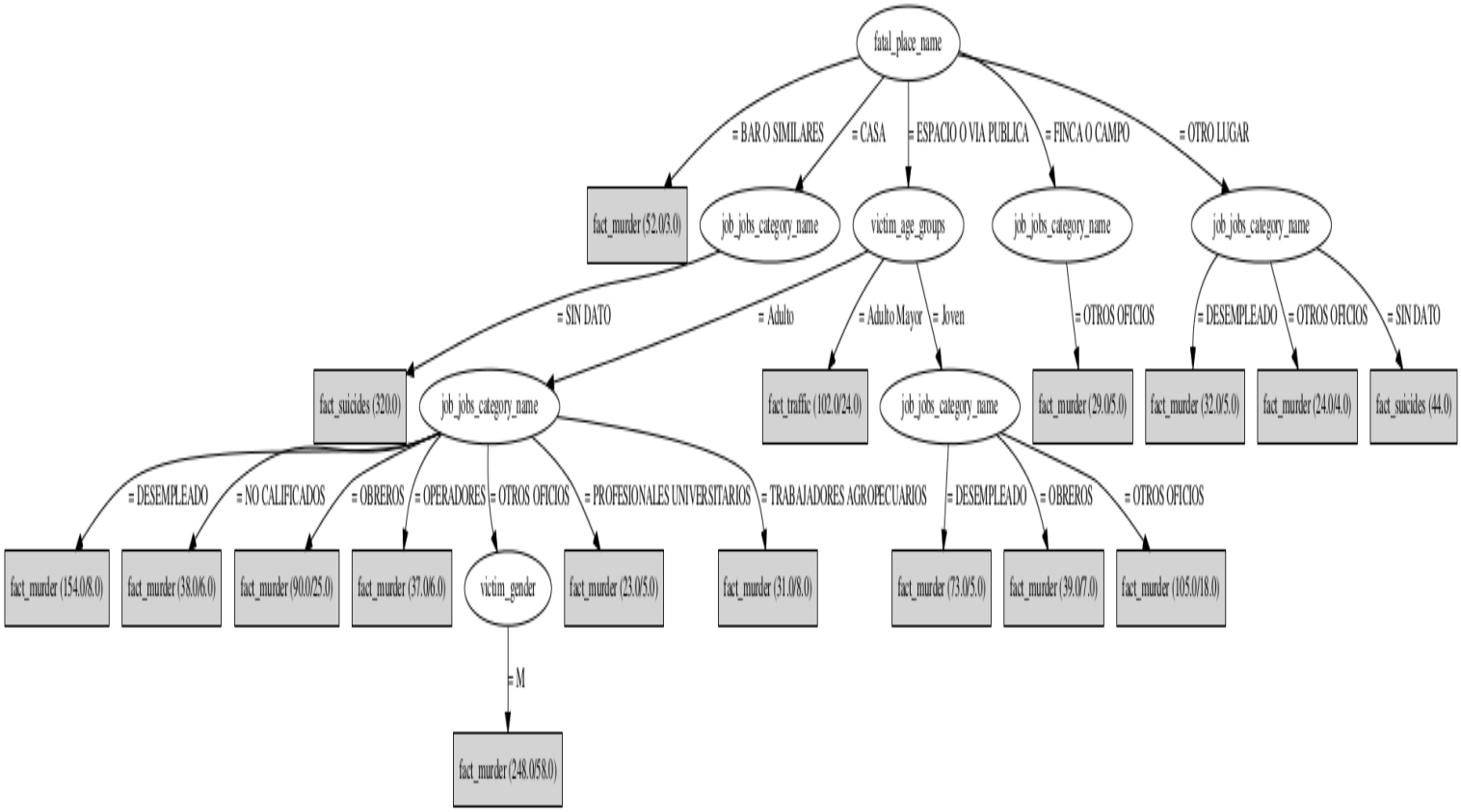

Figura 3. Mejor árbol de clasificación de muertes por causa externa

Implementación

Teniendo en cuenta los patrones descubiertos, es importante implementar planes de prevención que permitan disminuir los homicidios en las comunas 4 y 5 del municipio de Pasto. Implementar programas de sensibilización en colegios y universidades con estudiantes y padres de familia que permitan prevenir los suicidios entre los niños y jóvenes y finalmente realizar campañas de sensibilización en la comunidad con el fin de que los ancianos siempre salgan de sus casas acompañados de un familiar para evitar las muertes por accidentes de tránsito, a las cuales son propensas este tipo de población. De igual manera, se deben redoblar esfuerzos en educación vial y sanciones más enérgicas contra los infractores.

Este conocimiento descubierto se incorporará al existente y se integrará a los procesos de toma de decisiones de la Alcaldía de Pasto y de las instituciones gubernamentales que velan por la

seguridad ciudadana. Una vez los organismos gubernamentales del municipio de Pasto intervengan los factores que inciden en las muertes por causa externa, será posible analizar los resultados y determinar sus efectos.

Discusión

Según la matriz de confusión de la Figura 2, el modelo clasifica correctamente al 36,81% de casos de muertes no intencionales, al 93,77% de casos de homicidios, el 89,63% de casos de suicidios y al 41,81% de casos de muertes por accidentes de tránsito. Por otra parte, el estadístico Kappa, que mide la coincidencia de la predicción con la clase real de este modelo es de 0,67, que se considera aceptable (1,0 significa que ha habido coincidencia absoluta). Los porcentajes de instancias correctamente clasificados presentados en el árbol como en la matriz de confusión indican que el modelo tiene una precisión buena, por lo tanto es confiable y eficiente, para clasificar nuevos casos, especialmente homicidios y suicidios.

Como se puede observar en la Figura 3, algunos de los patrones más representativos son:

- Si la muerte sucede en un bar o similares entonces es un homicidio. El 2,37% de todas las muertes (2198) que sucedieron entre los años 2003 y 2013 se clasifican de esa manera y el 94,23% de los homicidios (1285) cumplen con este patrón.
- Si la muerte sucede en un espacio o vía pública y la víctima es adulto y está desempleado entonces es un homicidio. El 7,01% de todas las muertes que sucedieron entre los años 2003 y 2013 se clasifican de esa manera y el 94,81% de los homicidios cumplen con este patrón.
- Si la muerte sucede en casa, entonces es un suicidio. El 14,56% de las 2198 muertes que sucedieron entre los años 2003 y 2013 se clasifican de esa manera y el 100% de los 482 suicidios cumplen este patrón.
- Si la muerte sucede en la vía pública y la víctima es mayor que 65 años, entonces la causa es un accidente de tránsito. El 4,64% de las 2198 muertes que sucedieron entre los años 2003 y 2013 se clasifican de esa manera y el 76,47% de los 287 accidentes de tránsito cumplen este patrón.

En general, los homicidios suceden en lugares públicos como bares y vías públicas de las comunas 4 y 5 del municipio de Pasto. Las víctimas son jóvenes entre 20 y 25 años o adultos entre 30 y 50 años y desempleados. Los suicidios suceden en las casas y las víctimas son niños o jóvenes entre los 10 y 25 años. La población vulnerable a morir en un accidente de tránsito, en la vía pública son los adultos mayores. No se presentan patrones para muertes no intencionales, mediante la técnica de clasificación por árboles de decisión, por el menor número de casos que se han presentado entre los años 2003 al 2013 que es de 144.

Con respecto a homicidios, los patrones obtenidos en esta investigación se parecen a los obtenidos por Valenga *et al.*[7,12], aplicando *clustering* para la detección de patrones de homicidios dolosos en Argentina. Se obtuvieron tres clusters. El cluster 0 y cluster 1 que agrupan el 22% y el 43% respectivamente de los 1810 homicidios registrados en el año 2005, están caracterizados por homicidios mayoritariamente en la vía pública con arma de fuego.

En relación a muertes por accidentes de tránsito, la vía pública es el lugar del deceso para transeúntes con el que coincide el estudio realizado por Montt *et al.*[13], utilizando redes neuronales al concluir que para peatones las causas más frecuentes de las muertes por accidentes de tránsito son “no respetar derecho preferente de paso peatón” y “peatón, cruza calzada forma sorpresiva o descuidada”.

En cuanto a cifras, estos resultados se asemejan a los obtenidos en un estudio descriptivo longitudinal retrospectivo realizado por Cambell y Quintero[25] sobre lesiones fatales de causa externa en el municipio de Florencia (Caquetá), ocurridas durante los años 1991 a 1995, donde se clasificaron 810 casos, de acuerdo con edad, sexo, manera, causa de muerte y sitio donde ocurrió la lesión. Según la manera cómo ocurrieron las muertes, los homicidios ocuparon

el primer lugar con 530 eventos (65,4%), en segundo lugar los accidentes de tránsito 184 (22,7%), en tercer lugar los accidentes no intencionales 84(10,4%) y finalmente, en un porcentaje menor (1,4%) los suicidios con 11 casos. El grupo etario más comprometido por los homicidios fue el de 15 a 34 años con un total de 318 casos (60%). El medio principal por el cual se cometieron los homicidios fue el uso de armas de fuego. De 530 homicidios, 433 (81,7%) ocurrieron por la utilización de este tipo de armas, mientras que por asfixia mecánica (estrangulamiento) se encontraron ocho casos (1,5%). El sexo masculino fue el más afectado con 393 casos (91%) de 433. Los homicidios urbanos predominaron sobre los rurales con 465 casos contra 65, para un 88% y 12% respectivamente.

De manera similar a los resultados del estudio realizado en Florencia, en el municipio de Pasto, durante 2003 y 2013, los homicidios fueron la causa principal de las muertes por causa externa con un 58,46%, afectando principalmente la población joven y productiva, fenómeno que es generalizado en toda Colombia, como lo muestra un estudio realizado por Moya[26] sobre el comportamiento de lesiones de causa externa durante el año 2013. En este año se analizaron 26.623 necropsias medicolegales de muertes violentas ocurridas en el país. El homicidio se ubicó en la primera manera de muerte violenta con un total de 14.294 casos que representa un 53,7% de todas las muertes, con una tasa del 56,50 por 100 mil habitantes.

De acuerdo con informes oficiales recibidos por la OPS[27], entre los años 1992 y 2002, hubo en las Américas un promedio de casi 120.000 homicidios, 55.000 suicidios y unas 125.000 defunciones debidas a accidentes de tránsito. La tasa bruta de homicidios registrados es de 14 por 100.000 habitantes, una de las más altas notificadas en diferentes regiones del mundo. Los homicidios ocurren con mayor frecuencia en las zonas urbanas. Casi 81% del total de 120.000 homicidios en la región ocurre en Brasil (37.151), Colombia (23.466), Estados Unidos (20.984) y México (15.625). En estos informes se muestra que las muertes por accidentes de tránsito fueron levemente más altas que las debidas a homicidios. Lo mismo sucede en países como Argentina donde los accidentes de tránsito y no los homicidios son la principal causa de muerte. Según el informe de resultados de la Segunda Encuesta Nacional de Factores de Riesgo realizada en el año 2009 en Argentina[28], se registraron un total de 304.525 muertes de las cuales 18.860 fueron por lesiones de causa externa, equivalente al 6% del total, siendo estas la cuarta causa de muerte a nivel de la población general y la primera en personas de entre 1 y 44 años. De estas, el mayor porcentaje, 23%, corresponde a las lesiones por accidentes de tránsito, que son la principal causa de mortalidad por causas externas y representan aproximadamente el 2,2% de todas las muertes en el mundo. Según la OMS, en el mundo mueren más de 1,2 millones de personas al año por lesiones ocasionadas por accidentes de tránsito, cifra que equivale a aproximadamente 3000 muertes por día. Más del 90% de estas muertes ocurren en países de medianos y bajos recursos[29].

Finalmente, según el volumen de países el informe de Salud en las Américas de 2012 de la OPS[30] en Colombia se ha registrado una reducción en la proporción de muertes por homicidios, sobre todo en adultos varones jóvenes, pero junto con un aumento en la mortalidad y la discapacidad por accidentes de tránsito en adultos mayores, información que concuerda con el patrón general que se obtuvo en esta investigación para muertes por accidente de tránsito.

Conclusiones y trabajos futuros

Los resultados obtenidos con el modelo de clasificación por árboles de decisión, indican que este es capaz de generar modelos consistentes con la realidad observada y el respaldo teórico, basándose únicamente en los datos que se encuentran almacenados en el *datamart* del Observatorio del Delito del municipio de Pasto.

Los porcentajes de instancias correctamente clasificadas presentados en el árbol como en la matriz de confusión muestran que el modelo

tiene una precisión buena y por consiguiente es confiable para clasificar nuevos casos, especialmente las muertes por homicidios y suicidios. El proceso de pospoda del árbol y el factor de confianza y soporte establecidos no permitió generar reglas para las muertes no intencionales.

Como trabajos futuros se plantea utilizar otros clasificadores para comparar estos resultados y mejorar la precisión, especialmente en las muertes por accidentes de tránsito y no intencionales; aplicar tareas descriptivas de minería de datos como asociación y agrupación, con el fin de encontrar relaciones y similitudes, según el tipo de muerte por causa externa.

Financiación
Este proyecto de investigación se financió con recursos del Patrimonio Autónomo Fondo Nacional de Financiamiento para la Ciencia, la Tecnología y la Innovación Francisco José de Caldas -COLCIENCIAS y con recursos de contrapartida del Sistema de Investigaciones de la Universidad de Nariño y de la Alcaldía Municipal de Pasto.

Conflicto de intereses
Dentro de esta investigación no hubo conflicto de intereses.

Referencias

1. Chen H, Chung W, Qin Y, Chau M, Xu JJ, Wang G, et al. Crime Data Mining : An Overview and Case Studies. Commun ACM. 2002;2:165-276.
2. Perversi I, Valenga F, Fernández F, Britos P, Garcia-Martinez R. Identificación y Detección de Patrones Delictivos basada en Minería de Datos. En: IX Workshop de Investigadores en Ciencias de la Computación; Trelew, Argentina 2007. Trelew: Red de Universidades con Carreras en Informática (RedUNCI); 2007. p. 385-9.
3. Schotborgh M, Laverde N, Valbuena Y, Blandón A. Protocolo de Vigilancia en Salud Pública: Lesión de Causa Externa. Bogotá: INS; 2016.
4. Instituto CISALVA. Sistematización de Experiencias sobre Sistemas de Vigilancia, Observatorios o Sistemas de Información de Violencia en América Latina. Cali,: Centro Editorial CATORSE SCS; 2009. 62 p.
5. Instituto CISALVA. Guía Metodológica para la Replicación de Observatorios Municipales de Violencia. Cali: Centro Editorial CATORSE SCS; 2008. 45 p.
6. Betancourt C. Vigilancia de lesiones de causa externa para la toma de decisiones en el nivel local, experiencia de Pasto, Colombia, 1 diciembre de 2004 - enero 15 de 2005. Inf Quinc Epidemiol Nac. 2005;10(12):177-92.
7. Valenga F, Fernández E, Merlino H, Procopio C, Britos P, Garcia-Martinez R. Minería de Datos Aplicada a la Detección de Patrones Delictivos en Argentina. En: VI Jornadas Iberoamericanas de Ingeniería de Software e Ingeniería del Conocimiento; Guayaquil, 2008. Guayaquil: Escuela Superior Politécnica del Litoral Facultad de Ingeniería Eléctrica y Computación Área de Ingeniería en Software VLIR -ESPOL Componente 8; 2008. p. 427.
8. Sentient. DataDetective Sentient Information Systems [Internet]. Ámsterdam: Sentient; 2012. Disponible en: http://sentient.nl/docs/ReleaseNotes_DataDetective2012_NL.pdf
9. Chen H, Chung W, Xu J, Wan G, Qin Y, Chau M. Crime Data Mining: A General Framework and Some Examples. IEEE Comput Soc. 2004;37(4):50-6.
10. Hauck R, Atabakhsh H, Ongvasith P, Gupta H, Chen H. Using Coplink to Analyze CriminalJustice Data. IEEE Comput. 2002;35:30-7.
11. Reza K, Javideh M, Reza E. Detecting and investigating crime by means of data mining: a general crime matching framework. Procedia Computer Science. 2011;3:872-80.
12. Valenga F, Perversi I, Fernández E, Merlino H, Rodríguez D, Britos P. Aplicación de Minería de Datos para la Exploración y Detección de Patrones Delictivos en Argentina. En: XIII Congreso Argentino de Ciencias de la Computación; Argentina, 2007. Argentina: Universidad Nacional del Nordeste; 2008. p. 1868.
13. Montt C, Rubio J, Lanata S. Análisis de accidentes de tránsito con Inteligencia Computacional. XVI Congreso Chileno de Ingeniería de Transporte. 2013;(16):1-11.
14. Azevedo A, Santos M. KDD, SEMMA and CRISP-DM: a parallel overview. In: IADIS European Conference on Data Mining. Amsterdam, Netherlands; 2008. p. 182-5.
15. Timaran R, Baron A, Hernàndez G, Arsenio H, Betancourth C. SIGEODEP: Un primer paso para la Detección de Patrones Delictivos con Técnicas de Minería de Datos. In: Pow-Sang JA, Melgar A, editors. IX Jornadas Iberoamericanas de Ingeniería de Software e Ingeniería del Conocimiento. Lima, Perú: Pontificia Universidad Católica del Perú; 2012. p. 87-94.
16. Timaran R, Calderón A, Hidalgo A, Baron A, Hernández G. Construcción de un mercado de datos para el almacenamiento de lesiones de causa externa. Vent Inform. 2014;30:67-79.
17. Gallardo J. Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM. [Internet]. 2009. Disponible en: http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037. pdf
18. Waikato. Weka 3: Data Mining Software in Java [Internet]. Nueva Zelanda: Machine Learning Group at the University of Waikato. Disponible en: http://www.cs.waikato.ac.nz/ml/weka/

19. Han J, Kamber M. Data Mining: Concepts and Techniques. San Francisco, USA: Morgan Kaufmann Publishers; 2001. 550 p.
20. Sattler K, Dunemann O. SQL Database Primitives for Decision Tree Classifiers. In: Paques H, Liu L, Grossman D, editors. The 10th ACM International Conference on Information and Knowledge Management. Atlanta, USA: ACM New York; 2001. p. 379-86.
21. Villena J. CRISP-DM: La metodología para poner orden en los proyectos de Data Science. [Internet]. 2016. Disponible en: https://data.sngular.team/es/art/25/crisp-dm-la-metodologia-para-poner-orden-en-los-proyectos-de-data-science
22. Hernández J, Ramírez M, Ferri C. Introducción a la Minería de Datos. Fayerman D, editor. Madrid: Pearson Prentice Hall; 2004. 680 p.
23. García M, Álvarez A. Análisis de Datos en WEKA - Pruebas de Selectividad [Internet]. Disponible en: http://www.it.uc3m.es/jvillena/irc/practicas/06-07/28.pdf
24. Witten I, Frank E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. San Francisco: Morgan Kaufmann Publishers; 2000. 365 p.
25. Campbell S, Quintero M. Comportamiento de las lesiones fatales de causa externa en Florencia. Acta Médica Colombiana. 1997;22(4):161-166.
26. Moya D. Comportamiento de lesiones de causa externa, Colombia, 2013. Bogotá: Instituto Nacional de Medicina Legal y Ciencias Forenses.
27. Organización Panamericana de la Salud. La Salud en las Américas Edición 2002 Volumen I. Washington: OPS; 2002. 473 p.
28. Ministerio de Salud de la Nación. Lesiones por causa externa. En: Informe de resultados Segunda Encuesta Nacional de Factores de Riesgo. Buenos Aires: Ministerio de Salud; 2009. p. 182-225.
29. World Health Organization. The Global Burden of Disease 2004 update. Geneve: WHO; 2008. 160 p.
30. Organización Panamericana de la Salud. La Salud en las Américas Edición 2012 Volumen de Países: Colombia. Washington: OPS; 2012.