CAPÍTULO II

Big data para la identificación de comportamiento criminal

ELWIN VAN 'T WOUT
Instituto de Ingeniería Matemática y Computacional UC

EDUARDO VALENZUELA
Instituto de Sociología UC

KENZO ASAHI
Escuela de Gobierno UC

CHRISTIAN PIERINGER
Instituto de Ingeniería Matemática y Computacional UC y
Universidad Tecnológica de Chile, INACAP

DAVID TORRES
Escuela de Psicología UC

PILAR LARROULET
Instituto de Sociología UC

CAPÍTULO II

Big data para la identificación de comportamiento criminal

INVESTIGADORES[1]

ELWIN VAN 'T WOUT
Instituto de Ingeniería Matemática y Computacional UC

EDUARDO VALENZUELA
Instituto de Sociología UC

KENZO ASAHI
Escuela de Gobierno UC

CHRISTIAN PIERINGER
Instituto de Ingeniería Matemática y Computacional UC y
Universidad Tecnológica de Chile, INACAP

DAVID TORRES
Escuela de Psicología UC

PILAR LARROULET
Instituto de Sociología UC

Resumen[2]

El *big data* ha revolucionado las ciencias y la industria con su impresionante poder de analizar grandes conjuntos de datos y algoritmos que pueden predecir eventos futuros con una alta precisión. Este capítulo estudiará el uso de herramientas automatizadas para identificar el comportamiento delictual, con el objetivo final de ayudar a la policía en su mejora continua de la gestión eficiente de los recursos gubernamentales para reducir la delincuencia. Se han desarrollado modelos matemáticos que permiten extraer información relevante de una base de datos, que contiene registros de detenciones por las policías e información sociodemográfica sobre los victimarios. El análisis descriptivo presenta los indicadores clave del estado actual de delincuencia en Chile y una tipología del comportamiento criminal. La predicción de detenciones futuras es desafiante: la precisión debe aumentarse, antes de usar

1 El equipo de investigadores extiende su profundo agradecimiento al Centro Nacional del Análisis Criminal de la Policía de Investigaciones por su fuerte compromiso al proyecto; particularmente, a los aportes del subprefecto Pedro Muñoz, de los subcomisarios Angelo Dini y Harmin Cottenie, y de Roberto Zulantay, María José Rojas y Carlos Sáez.

2 Esta propuesta fue presentada en un seminario organizado por el Centro de Políticas Públicas UC, realizado el 19 de noviembre de 2018, en el que participaron como panelistas Arturo Reyes, jefe del departamento de Reinserción Social de la Subsecretaría Prevención del Delito, Ministerio del Interior; Ana María Morales, directora del área de Justicia y Reinserción de Paz Ciudadana, y Cristóbal Weinborn, asesor del BID en materia de criminalidad y victimización.

los algoritmos en la práctica, especialmente porque una predicción errónea puede resultar en costos sociales demasiados altos. Para completar el estudio, se presenta una discusión sobre los alcances y posibles fuentes de sesgo en la predicción. Por último, se elaboran propuestas sobre cómo mejorar las políticas públicas y la organización policial basada en herramientas de *big data*.

Introducción

La delincuencia es un problema mundial que afecta transversalmente a toda la sociedad. En Chile, el crimen es el problema que más preocupa a los ciudadanos. En el Estudio Nacional de Opinión Pública, "la delincuencia" concentra un 27% de las respuestas, situándose como primera prioridad a nivel país, más alta incluso que "el desarrollo económico" (16%) y "las pensiones" (13%), entre otras respuestas (Centro de Estudios Públicos, 2017). El mismo estudio indica que para un 47% de los respondientes, la delincuencia es uno de los tres problemas a los que el Gobierno debería dedicar el mayor esfuerzo en solucionar. Estos indicadores muestran una fuerte necesidad de elaborar e implementar políticas públicas efectivas, orientadas a la disminución de actos delictivos. Tales políticas requieren que el conocimiento de la criminalidad en el país mejore continuamente y sea actual.

Las políticas –cuyo objetivo es disminuir los índices de victimización– ocupan principalmente dos fuentes de datos: encuestas y datos administrativos. En las últimas décadas, la elaboración de encuestas de victimización ha permitido mejorar la comprensión de qué tipos de delitos son más recurrentes y tienen una connotación social más alta. Los resultados permiten un mejor análisis de las principales políticas dirigidas a enfrentar la victimización y mejorar la seguridad percibida por los ciudadanos. No obstante, uno de los problemas de estas mediciones es su especial foco en los delitos y en las víctimas, pasando por alto los datos sobre los victimarios.

Otra fuente de información importante para entender la delincuencia proviene de registros administrativos. Estos datos son generados mediante el esfuerzo que realizan los distintos actores gubernamentales encargados de la gestión, control y prevención de delitos, como son la Subsecretaría de Prevención del Delito, Gendarmería, la Fiscalía, Carabineros de Chile y la Policía de Investigaciones. Su ventaja principal radica en que incluyen información clave de los victimarios, lo que permite estudiar temas relacionados con la reincidencia y los patrones de comportamiento criminal.

De las instituciones públicas que estudian la criminología actual en el país, la Policía de Investigaciones (PDI) ha conseguido posicionarse como referente nacional. Con el fin de modernizar su gestión, la PDI creó el Centro Nacional de Análisis Criminal (Cenacrim) en 2015, cuyo objetivo principal es la recolección, evaluación y análisis de información que sea de utilidad

en el trabajo policial y las estrategias que de él se desprendan. Esto otorga visiones globales y sistemáticas sobre los fenómenos delictuales. De esta forma, el Cenacrim apoya con políticas y conocimiento –a través del análisis de datos administrativos– a la PDI y a los organismos estatales encargados de velar por la seguridad interna del país como, por ejemplo, la Subsecretaría de Prevención del Delito. Por una parte, la creación del Cenacrim ha permitido que la PDI mejore el registro de actividades delictuales, aumentando considerablemente el volumen y variedad de los datos, haciendo impracticable su análisis manual. Por otra, la riqueza de la información consolidada –desde distintas bases de datos administrativas– hace posible la aplicación de nuevos métodos para la búsqueda de patrones, que permitirían al Estado realizar una labor más efectiva en la disminución del delito. El Cenacrim es actualmente una pieza clave para la generación de conocimiento en el área del análisis criminal nacional, aportando con la experiencia de campo y los datos recolectados.

La presente investigación está alineada con la misión del Cenacrim, que busca generar conocimiento criminológico a través del estudio de bases de datos administrativos, en particular, los registros de detenciones que puedan dar nuevas luces sobre la delincuencia en el país. El objetivo general es describir y predecir el comportamiento criminal de individuos que ya poseen registros de detención, con el fin de apoyar la elaboración de políticas públicas que apunten a una disminución en la delincuencia. Basada en la identificación de comportamiento criminal, se contempla fortalecer la capacidad de la PDI para elaborar estrategias de mediano y largo plazo en la mejora de su gestión interna. Ello contempla, entre otros, los mecanismos de priorización de las investigaciones y el apoyo a otros actores gubernamentales, a través de la entrega de herramientas y conocimiento en el área de la investigación policial. Metodológicamente, este proyecto estudia los alcances del uso de herramientas del *big data* en la identificación de comportamiento criminal.

Marco conceptual

1. El *big data* en la criminología

En los últimos años, ha existido un tremendo incremento de las capacidades de almacenamiento y generación de bases de datos en todas las áreas de conocimiento, incluida la criminología. El Estado recolecta información sobre todos los ciudadanos, manejada por los departamentos de policía, sistemas judiciales y organizaciones de desarrollo social. La agregación de esta información administrativa permite tener antecedentes en distintas dimensiones para todas las interacciones entre un individuo y el aparato estatal. Esto abre posibilidades insospechadas en cuanto al análisis de datos, que no existían hasta hace algunos años (Ridgeway, 2018).

Los gobiernos en distintos países del mundo están abordando el problema de la reincidencia delictual mediante el análisis de datos administrativos y el uso de modelos matemáticos, para comprender los patrones que pueden explicarla. Ozkan (2017) realizó un estudio sobre la aplicación de técnicas de aprendizaje automatizado para la predicción de reincidencia. Los resultados revelan que en Estados Unidos la tasa de reincidencia en promedio es de un 70% de los individuos que han sido liberados dentro de una ventana de tres años. Esta tasa varía según el tipo de delitos, siendo más alta para aquellos que afectan a la propiedad. Otros ejemplos del uso de aprendizaje automatizado han sido presentados por Bartolucci et al. (2007) y Zeng et al. (2017), quienes desarrollaron métodos matemáticos para la predicción de trayectorias y reincidencia criminal. A pesar del beneficio que este tipo de metodologías pueda traer en este contexto, también existen críticas sobre las limitaciones de los modelos predictivos y la extensión de las conclusiones que se puedan sacar a partir de estas herramientas de *big data* (Dressel y Farid, 2018).

Actualmente, en Chile, faltan herramientas robustas y sistemáticas de análisis que permitan al Estado priorizar el uso de recursos en determinados grupos de personas que tienen un riesgo alto de reincidencia. Sin embargo, esta brecha genera el desafío de desarrollar técnicas de *big data* para la creación de nuevo conocimiento y elaboración de políticas de seguridad. Los modelos matemáticos pueden ayudar a identificar asociaciones entre variables relevantes para el análisis criminal, como son las detenciones, inscripción de armas, grupo socioeconómico, viajes, bienes, red familiar, entre otras.

La existencia de una tipología de comportamiento criminal o trayectorias delictuales podría sugerir la priorización de recursos policiales y el desarrollo de programas gubernamentales de prevención de delitos. Además, el proyecto podría sugerir la existencia y aplicación de programas de reinserción social personalizados a las necesidades de cada perfil criminal que emerja del análisis. De esta forma, los algoritmos y modelos generados servirían como justificación para la creación de programas pilotos de reinserción social, elaborados por el Ministerio de Justicia o por la Subsecretaría de Prevención del Delito para los diferentes grupos de infractores.

2. Hechos empíricos en la criminología

La literatura criminológica apunta a diversos caminos que llevan al involucramiento en el delito y en actos violentos (Farrington, 2003); en general, dichos caminos no están asociados a causas únicas, sino a un conjunto de factores de riesgo. Entre los destacados en las investigaciones, se encuentran factores individuales (bajo autocontrol, impulsividad), familiares (herramientas parentales débiles o inconsistentes, exposición a violencia y maltrato), escolares (bajo involucramiento y rendimiento escolar) y de pares (vínculos

con pares involucrados en delito). Además, factores de contexto, como que el vivir en comunidades con una alta concentración de pobreza e inestabilidad residencial aumenta las probabilidades de involucrarse en delitos (Laub, 2015; Akers, 1999).

Adicionalmente, hay una serie de hechos empíricos reconocidos en la literatura que apuntan a los patrones de involucramiento delictual. Sabemos que existe una relación entre la edad y el delito; la comisión de delito se concentra mayoritariamente en la adolescencia tardía y la adultez temprana: el *peak* ocurre alrededor de los 20 años y la curva presenta una caída rápida a partir de los 25 años (Gottfredson y Hirschi, 1990; Farrington, 2003). Las publicaciones también señalan que un inicio temprano en el delito predice un involucramiento más largo y serio en el mismo.

La evidencia empírica apunta también a la relación entre delito y género. Gran parte de los primeros, y principalmente los más serios y violentos, son cometidos por hombres (Schwartz et al., 2009). Los estudios muestran que, en comparación con las mujeres, los hombres tienen una mayor prevalencia de involucramiento en el delito, se inician de manera más temprana, cometen una mayor cantidad de ellos, más violentos y graves, y tienden a tener carreras delictuales más extensas (Moffitt et al., 2001; Block et al., 2010; Britton, 2011).

Los datos que se presentan son, sin embargo, promedios sobre un total de población que difiere en sus probabilidades de involucrarse en comportamiento delictual, sea como consecuencia de un diferencial en los factores de riesgo o en las oportunidades para cometer crímenes.

3. Perfiles y trayectorias delictuales

La literatura en criminología indica que, en general, no existe un camino único al involucramiento delictual, sino que una multiplicidad de vías y trayectorias que se distinguen tanto en sus factores de riesgo como en sus patrones en el tiempo (Piquero, 2008). De la mano con el desarrollo de nuevos modelos estadísticos que han permitido estudiar patrones de actividad criminal individuales y de grupo, el uso de datos longitudinales ha permitido el desarrollo del área de trayectorias delictuales (Nagin y Land, 1993). Como plantea Piquero (2008), los estudios realizados muestran que se pueden identificar entre tres y cinco grupos de individuos que difieren en su probabilidad de involucramiento delictual, inicio en el delito y persistencia en el tiempo. Estos grupos parecieran estar asociados a factores de riesgo diversos, que explicarían tanto su edad de inicio como la forma de su trayectoria en términos de incidencia y persistencia. Como lo han planteado diversos autores (ver, por ejemplo, Sampson y Laub, 2005), es clave tener en mente que los grupos o clases de comportamiento delictual son solo un reflejo de la heterogeneidad

de la población en términos de su comportamiento criminal y no grupos reales. Sin embargo, la creación de grupos de perfiles criminales son una buena manera de representar un comportamiento que no varía de manera regular en la población sino que tiende a presentar distintos niveles de intensidad entre subgrupos de la misma (Nagin, 2005).

Si bien la literatura y teoría del delito reconocen que el comportamiento pasado es el mejor predictor del comportamiento futuro (Nagin y Paternoster, 1991; 2000), también destacan la dificultad de predecir el comportamiento en períodos más largos de tiempo. Como se señaló anteriormente, uno de los hechos empíricos más consistentes en la criminología es el vínculo entre edad y delito. Si bien algunos individuos persisten en el comportamiento delictual a lo largo del curso de la vida, la enorme mayoría de ellos desiste durante la edad adulta. Como señalan Laub y Sampson en su trabajo siguiendo a individuos desde los siete a los 70 años de edad (Sampson y Laub, 1993; Laub y Sampson, 2003), existe una enorme heterogeneidad en el comportamiento delictual a lo largo del curso de la vida y, si bien estos patrones delictuales son posibles de identificar retrospectivamente, no se pueden predecir prospectivamente (Petersilia, 1980; Auerhahn, 1999).

Durante la última década, ha existido bastante discusión en el ámbito académico sobre el poder predictivo de los modelos de comportamiento criminal. Estos han estado en la base de políticas fundadas en la incapacitación selectiva de individuos involucrados en comportamiento criminal. Sin embargo, si bien sabemos que un porcentaje menor de personas es responsable de la gran mayoría de los delitos cometidos (Wolfgang et al., 1987), también la evidencia señala que entre un 20 y un 30% de quienes cometen delito desertan de la actividad delictual luego de cada evento, y que este proceso es estocástico. Es decir, si bien podemos predecir que una proporción menor de individuos persistirá en el delito, la identificación de quiénes son esos individuos es sumamente compleja (Auerhahn, 1999).

4. Ética y marco legal

Si bien hay quienes abogan por el uso de herramientas computacionales bajo la lógica de aprendizaje automatizado, argumentando que son herramientas predictivas y no explicativas (ver, por ejemplo, Berk y Bleich, 2013), hay problemas éticos asociados al uso de herramientas predictivas de baja precisión. Dada la excepcionalidad de los eventos delictuales, la precisión de las predicciones pareciera limitada en el caso de criminalidad. El uso de instrumentos de medición de riesgo, por ejemplo, ha sido altamente discutido en el contexto norteamericano, principalmente cuando la estimación estadística del riesgo futuro –es decir, por un delito aún no cometido– es utilizada en la determinación de sentencias y libertades condicionales (Kleiman et al., 2007; Hannah-Moffat, 2013).

El uso de modelos predictivos para las detenciones policiales se ha experimentado en varios lugares norteamericanos y ha generado una discusión ética sobre su impacto en minorías de la población. Los patrones detectados por los métodos de aprendizaje automatizado pueden generar predicciones que se inclinen hacia las personas más vulnerables o poblaciones que ya son sesgadas actualmente. Una fuente principal de sesgo en los modelos automatizados es la disponibilidad limitada de datos administrativos y la recopilación de registros delictuales en un sistema policial y penal que ya podría tener sesgos históricamente. Por este motivo, es clave entender los potenciales sesgos generados por modelos matemáticos en comparación con los de la sociedad y los actores policiales. Recientemente, Brantingham et al. (2018) han comparado el sesgo racial en las detenciones en una situación, con acciones policiales basadas en modelos predictivos y los procedimientos actuales. Estos autores concluyen que no hay evidencia que indique que el uso de herramientas predictivas aumente el sesgo en las detenciones. Sin embargo, el carácter preliminar del estudio sugiere la necesidad de ser cuidadoso en el uso a gran escala de las herramientas predictivas en temas relacionados con crimen.

Por último, se tiene que considerar que el uso de herramientas predictivas basadas en datos administrativos, sea a través de aprendizaje automatizado o de forma manual, con el objetivo de disminuir la reincidencia criminal, genera una discusión legal nueva y poco explorada. La pregunta clave es si el fin pretendido de mejorar la seguridad de toda la sociedad justifica medios en los cuales la dignidad de una persona individual se ve afectada, ya sea en la propiedad de los datos sensibles, el consentimiento informado o la posible estigmatización (ver Lara et al. (2013) sobre la privacidad en el sistema legal chileno). El estudio presente se ejecutó bajo las normativas vigentes de la PDI y con el fin de apoyar en su plan de desarrollo.

Metodología

El presente estudio se ha desarrollado en base a los datos administrativos respecto de las detenciones en Chile provista por la PDI. Las bases de datos incluyen información recolectada por la PDI, combinada con datos administrativos provenientes de otras fuentes como, por ejemplo, la Fiscalía. La base de datos principal para el estudio es la que contiene las detenciones registradas por la PDI a través de los datos proporcionados por Fiscalía. Dicha base contiene 777.724 registros de detenciones realizadas en la Región Metropolitana de Santiago desde el año 2009 hasta enero del año 2018, lo cual corresponde a 332.609 infractores. Adicionalmente, se dispone de información sobre las armas inscritas en Chile y datos biográficos manejados por la PDI, que incluyen edad, género y número de familiares directos con antecedentes.

Una limitación de los datos disponibles es que solo contemplan detenciones. Es importante destacar que las detenciones se diferencian de actos criminales en cuanto que no toda persona que comete un crimen es detenida y no todo detenido ha cometido un crimen. No se dispone de información sobre condenas o periodos de encarcelamiento en el estudio presente. Por ende, se analizará el historial de detenciones, manteniendo el foco en identificar patrones de habitualidad delictual. Por la misma razón, al hablar de reincidencia, para ser más exactos, se refiere a detenciones repetidas en esta investigación.

Como primer paso en proyectos de *big data*, los datos deben ser preparados mediante la limpieza y preprocesamiento, en línea con los objetivos del estudio. Un ejemplo del adecuamiento de los datos es la categorización de los tipos de delito, que es clave para la evaluación del análisis descriptivo y la toma de decisiones basada en el estudio. Es decir, por cada registro de detención, los datos crudos tienen un campo de texto libre que describe los delitos por los cuales la persona fue detenida. Por ejemplo, una detención puede registrar simultáneamente los delitos de hurto simple y conducción en estado de ebriedad. La gran gama de delitos registrados en la base de datos en texto libre hace necesario el organizar estos datos en categorías relevantes para su interpretación. Este formato de registro ofrece grados de flexibilidad que entorpecen su estandarización y el agrupamiento automático. Por ejemplo, la base de datos contiene el delito de hurto simple, hurto falta (494 bis Código Penal) y hurto falta (494 no. 19 Código Penal), el cual debemos interpretar como una sola categoría: hurto. Para los fines del presente estudio, se ha creado una categorización propia para la identificación de comportamiento delictual, con métodos de reconocimiento de patrones de texto y reglas predefinidas. Esta agrupación, sintetizada en la Tabla 1, está basada en la experiencia de los investigadores en el Cenacrim y estudios internacionales (Dominguez y Raphael, 2015).

TABLA 1. **Categorización de los tipos de delitos**

Tipo de delito	Descripción
Robos	Incluye todos los tipos de robo como, por ejemplo: robo con intimidación, robo con violencia, entre otros.
Hurtos	Todos los hurtos tipificados en la ley: hurto simple, hurto falta, entre otros.
Homicidios	Involucra, por ejemplo, homicidios, parricidios, femicidios, infanticidios y combinaciones, como violación con homicidio.
Tráfico de drogas	Incluye los delitos relacionados con el tráfico de narcóticos y microtráfico.
Consumo de drogas	Considera delitos como tenencia y consumo ilegal de drogas.
Delitos sexuales	Agrupa delitos sexuales como, por ejemplo: todos los tipos de violación y abuso sexual, distribución de pornografía infantil, entre otros.
Porte de arma blanca	Incluye delitos que consideren el porte ilegal de armas cortopunzantes.
Porte de arma de fuego	Agrupa delitos de tenencia ilegal de armas de fuego (pistolas/escopetas).
Violencia	Incluye todos los delitos que involucran lesiones desde leves a graves y violencia intrafamiliar.
Falsificación y estafa	Contiene delitos asociados a la falsificación, uso malicioso de documentación, lavado de dinero, fraude al fisco, entre otros.
Terrorismo y vandalismo	Agrupa delitos relacionados con daño a la propiedad pública y privada.
Tránsito	Involucra todos los delitos correspondientes a infracciones del tránsito como manejo en estado de ebriedad.
Otros	Incluye todos los delitos fuera de alguna de las categorías mencionadas antes descritas, como es el abigeato, denuncia por alimentos devengados, entre otros delitos.

Fuente: elaboración propia a partir de experiencia de Cenacrim y Domínguez, 2015.

Para el estudio del fenómeno delictual en Chile a través de los datos administrativos, se utilizó una metodología moderna de ciencias de datos (Figure Eight, 2017). Los principales componentes de este trabajo son: 1) entendimiento del contexto, 2) adecuación de los datos, 3) análisis descriptivo, 4) modelamiento predictivo, 5) evaluación de la calidad de los resultados, 6) entrega de conocimiento sobre el fenómeno estudiado. Este proceso es altamente iterativo, lo que permite, por ejemplo, agregar nuevas variables para mejorar la calidad de las predicciones o bien ajustando los modelos utilizados en base a una realimentación. En este estudio, el análisis descriptivo se basa en métodos de conglomeración y el análisis predictivo en métodos de clasificación.

Resultados

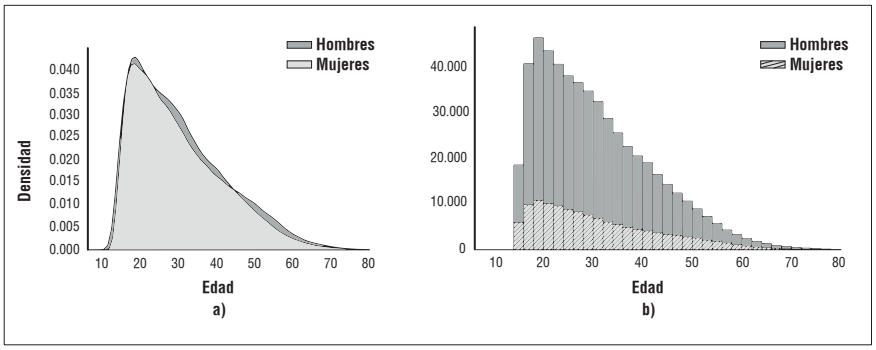
Los datos administrativos disponibles en este estudio permiten identificar patrones de detención en Chile. La investigación contrasta las estadísticas descriptivas obtenidas por medio del análisis de la base de datos, con el conocimiento general sobre delincuencia que está publicado en la literatura científica. Como se mencionó en el marco conceptual, la edad y el género son dos de los indicadores clave en estudios de comportamiento criminal. Ambos tienen correlaciones fuertes con la prevalencia e incidencia delictual, así como con el tipo de delito que se realiza. Mientras una carrera delictual evoluciona, algunos individuos pueden salir de la carrera criminal y otros pueden profundizar en dicha carrera, especializándose en delitos más elaborados y de mayor gravedad. Después de validar los hechos reconocidos en criminología, se estudian los perfiles y trayectorias de criminalidad en Chile, lo que incluye una caracterización de la habitualidad delictual y una tipología de comportamiento criminal. Para poder avanzar hacia una predicción de comportamiento criminal futuro, se analizan los alcances de métodos matemáticos que predicen la siguiente detención esperada, en base a la historia delictual y datos biográficos.

1. Análisis descriptivo

a. La influencia de la edad y género en la participación delictual

El género y la edad son dos de los indicadores más conocidos en el estudio del involucramiento delictual. Los datos administrativos muestran que, del total de individuos detenidos, un 86% corresponde a hombres y un 14% a mujeres. La Figura 1 muestra la frecuencia que tienen los delitos según la edad para cada género. La base de datos utilizada contiene 73.622 detenciones de menores entre los 14 y 18 años. El total de ellas corresponde a 35.849 detenidos en este rango de edad. De manera consistente con la literatura, se puede observar en la Figura 1 que la máxima actividad delictual se da en adolescentes y adultos jóvenes, marcando un máximo entre los 18 y 22 años. Luego de esa edad, la actividad decae rápidamente. Este descenso es relativamente más rápido en el caso de las mujeres, fenómeno que podría estar asociado con el inicio de la maternidad. En hombres se observa un patrón similar a partir de los cuarenta años. Temas como el comienzo de lazos familiares o la inserción en empleos formales han sido asociados en la literatura con este cambio (ver, por ejemplo, Laub y Sampson, 2003).

FIGURA 1. **Comparación de las distribuciones de la edad según género**



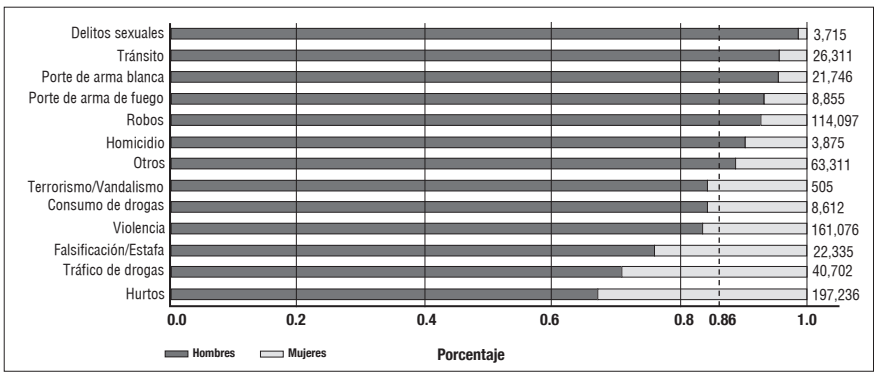
a) Distribución de delitos según edad y género. b) Histograma de frecuencia de delitos según edad y género.

Fuente: elaboración propia.

b. Distribución de tipos de delito según género

La variable género muestra también una clara relación con el tipo de actividad delictual. De manera acorde a lo esperado, los hombres tienden a presentar mayor prevalencia de delitos de más severidad, mientras que las mujeres se concentran en delitos menores contra la propiedad y el tráfico de drogas. La Figura 2 muestra la proporción de involucramiento en tipos de delitos según género. Se puede observar que, en general, los hombres participan en delitos más violentos que las mujeres, como son robos, homicidios o la posesión ilegal de armas de fuego. Al mismo tiempo, las mujeres están relativamente más involucradas en el tráfico de drogas.

FIGURA 2. **Resumen de los delitos agregados según categorías Resumen de los delitos agregados según categorías**

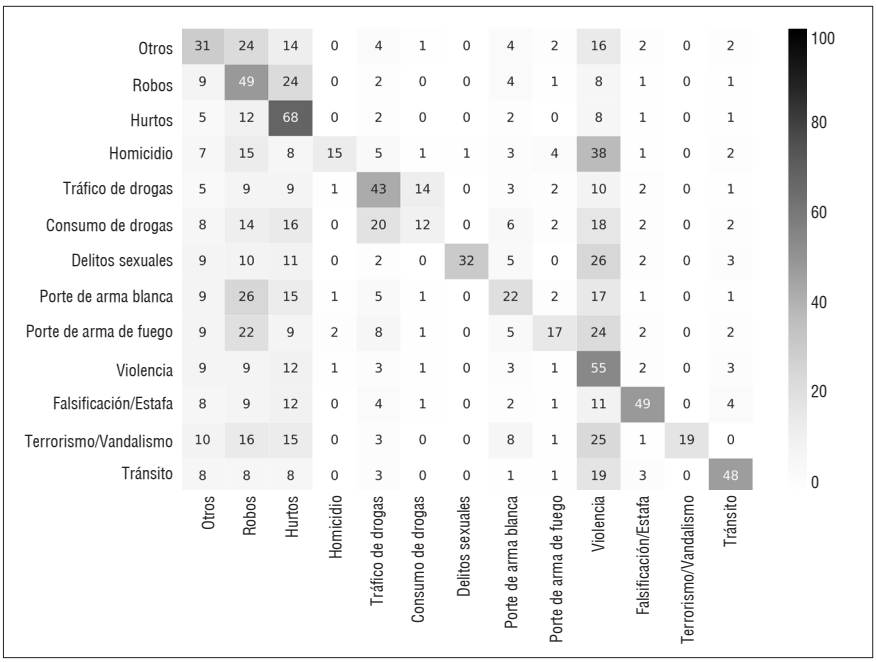


Fuente: elaboración propia.

c. Patrones de detención en el tiempo

Con el fin de comprender el fenómeno delictual, se estudian los patrones en las detenciones criminales sucesivas, para identificar una caracterización en comportamiento criminal. La Figura 3 muestra una matriz de transición entre las categorías de delitos, para todos los individuos que tienen al menos dos detenciones registradas en la base de datos. Se puede observar que la diagonal de la matriz concentra la mayor tasa de ocurrencia para cada tipo de delito, por lo tanto, que hay una tendencia a la especialización, lo que confirma evidencia previa para el caso chileno (Fábrega et al., 2014). Asimismo, hay ejemplos de transiciones habituales entre tipos de delitos, como la transición desde porte de armas a robos y hacia la violencia.

FIGURA 3. **Matriz de transición entre categorías de delitos**



Fuente: elaboración propia.

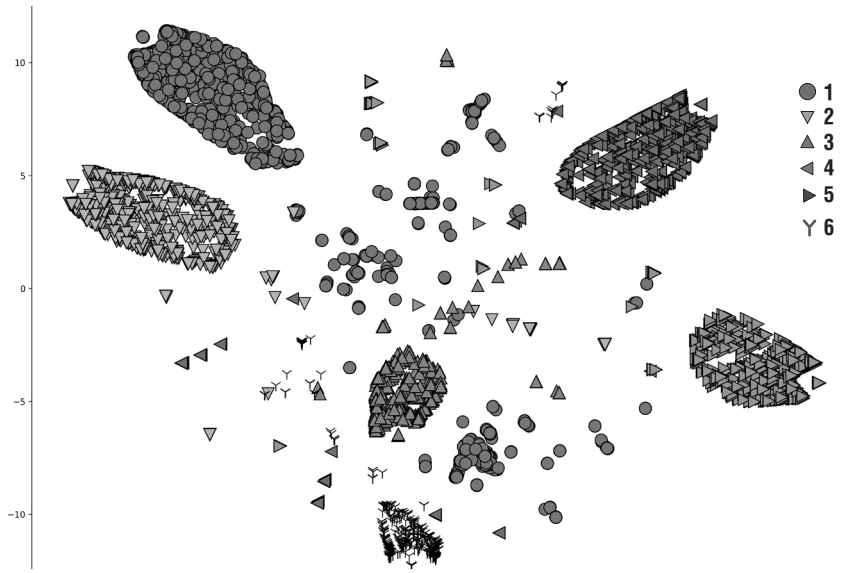
d. Tipología de incidencia delictual

Los resultados presentados indican que es posible identificar patrones empíricos significativos a partir de la base de datos de detenciones. Esto da soporte a la idea de que es posible generar tipologías de comportamiento criminal, en base a los patrones de detención, que permitan organizar la gran variedad de trayectorias delictuales en torno a un conjunto reducido de tipos o clases. La organización de patrones de detención –frente a un grupo reducido de clases

significativas– abre la posibilidad de optimizar el uso de recursos por parte de la PDI y otros actores gubernamentales involucrados en el combate a la delincuencia, al poder identificar y priorizar aquellos de mayor interés por sus altas consecuencias o connotación social.

Con el fin de buscar grupos de personas que tienen un comportamiento criminal parecido, se llevó a cabo un análisis descriptivo mediante técnicas de conglomeración (*cluster analysis*; Murphy, 2012). Para este estudio, se incluye la historia de detenciones de cada individuo, más las variables biográficas disponibles, tales como edad, género y número de armas inscritas. El algoritmo de conglomerados busca clasificar automáticamente los casos estudiados, en base a la similitud entre sus patrones de datos registrados. La Figura 4 muestra una visualización de cada individuo de acuerdo con el rótulo del conglomerado asignado por el algoritmo. Si bien el análisis de conglomerados se basa en el cálculo de distancias multidimensionales, este gráfico bidimensional se generó mediante el uso del método de t-SNE (Maaten y Hinton, 2008).

FIGURA 4. **Visualización de los conglomerados encontrados mediante reducción de dimensionalidad de los datos a través del método t-SNE**



Nota: las variables de los ejes corresponden a combinaciones de los indicadores utilizados en el método de conglomeración.

Fuente: elaboración propia.

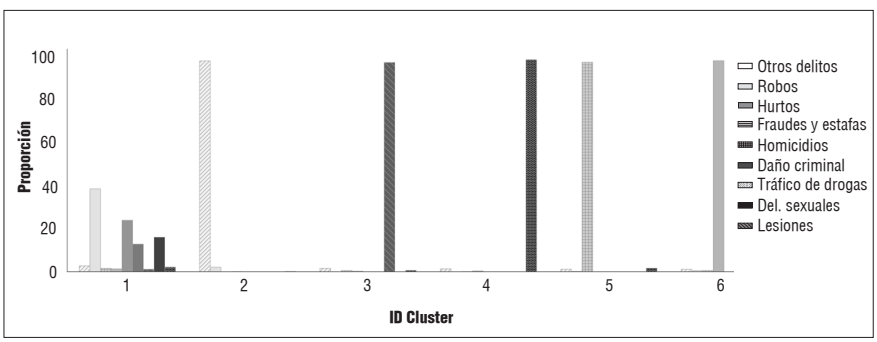
En base a los datos demográficos, podemos caracterizar cada conglomerado. La Tabla 2 resume el resultado de los determinados por el algoritmo. Se observa que los grupos tienen un tamaño similar salvo el primero, que concentra un 32% de las personas activas en la base de datos. Además, este es el que agrupa a todos los individuos que poseen armas.

TABLA 2. **Información demográfica de los conglomerados**

	Conglomerado					
	1	**2**	**3**	**4**	**5**	**6**
Porcentaje de individuos	32,14 %	14,81%	12,77%	14,71%	13,83%	11,75%
Porcentaje de armas inscritas	8,76%	0 %	0 %	0 %	0 %	0 %
Porcentaje de mujeres	8,92%	15,97%	29,47%	15,43%	30,55%	29,87%
Estado civil soltero	19,42%	25,27%	19,18%	23,43%	14,21%	28,44%

Fuente: elaboración propia.

FIGURA 5. **Diagrama con la conformación de los conglomerados según los tipos de delitos**



Fuente: elaboración propia.

Además de la información demográfica, otro aspecto central para la interpretación de cada conglomerado corresponde a los tipos de delitos más comunes asociados a las detenciones, los que son presentados en la Figura 5. Se observa que el primer conglomerado, en que se concentran las personas con armas, tiene una alta proporción de personas con antecedentes de detenciones por robos y homicidios y, en menor medida, por daño criminal y delitos

sexuales. Los otros conglomerados se caracterizan por una especialización en un tipo de delito, tales como el tráfico de drogas (tercer conglomerado), lesiones (cuarto), hurtos (quinto) y fraudes (sexto y último).

El análisis usando conglomerados ilustra el potencial detrás del objetivo de investigar tipologías de patrones de criminalidad, en base a los datos administrativos. Consolidar cientos de miles de patrones individuales en un número reducido de conglomerados que representan distintos tipos de patrones de conducta delictual, permite una reducción significativa en la complejidad asociada a la interpretación de los datos disponibles. Este número reducido de patrones podría ser utilizado, entonces, para clasificar de forma preliminar una persona de interés con un historial de detenciones y, de forma general, para focalizar y priorizar recursos tales como el tiempo y las experticias asociadas a la investigación de delitos específicos.

Los resultados presentados se basan en uno de los posibles algoritmos que pueden ser utilizados en un análisis de conglomerados, el algoritmo de "distancia del coseno". Es importante tener en cuenta que el análisis de conglomerados es una técnica en la cual el analista juega un rol central al definir el número de conglomerados a identificar. En este análisis, fue fijado en seis. Esta decisión es importante, ya que la cantidad de conglomerados definida tiene un impacto sobre las características de la tipología obtenida. Sin embargo, esta decisión se apoya en métodos auxiliares, que apuntan a soluciones de al menos seis conglomerados en la base de datos utilizada. Asimismo, este tipo de métodos son sensibles al tratamiento de los datos y al conjunto de hiperparámetros, los cuales pueden ser modificados para generar soluciones diferentes. Con el fin de identificar tipos o clases estables que permitan informar de forma consistente la toma de decisiones, es posible también complementar este tipo de técnicas con otras, tales como el análisis de clases latentes (LCA, por su nombre en inglés; Hagenaars et al., 2002) y su extensión a contextos longitudinales: el análisis de transición latente (LTA; Collins y Lanza, 2010) o análisis longitudinal de clases latentes (LLCA; Bartolucci et al., 2007).

2. Análisis predictivo

Análisis descriptivos, como los presentados hasta este punto, entregan información valiosa sobre los indicadores del comportamiento criminal y permiten el estudio de tipologías de incidencia delictual. Sin embargo, si se busca mejorar la priorización de recursos policiales y la elaboración de políticas públicas, sería de gran utilidad tener también a disposición herramientas automatizadas que puedan predecir comportamientos delictuales de personas que ya presentan antecedentes. Dado que las experiencias en criminología han planteado consistentemente la dificultad de predecir futuros delitos, se espera que la creación de modelos matemáticos de predicción sea desafiante.

Sin embargo, para estudiar si las herramientas computacionales pueden complementar los procesos manuales, es necesario cuantificar la calidad de los modelos predictivos. Por este motivo, se diseñó un experimento para contrastar las predicciones de diversos métodos con los datos administrativos disponibles, el que permitió cuantificar la precisión de los modelos predictivos en el ámbito de este estudio.

El objetivo del análisis predictivo es estimar el riesgo que tiene un individuo de volver a ser detenido por un nuevo delito (de cualquier tipo), en el futuro. Sin embargo, no hay conocimiento de lo que ocurrirá más allá del período de datos comprendido en la base actual. Para poder hacer la predicción y cuantificar su precisión, se reorganizó la base de datos de tal modo que, para cada individuo, la última detención registrada se separa de las detenciones previas. La historia de detenciones de cada individuo hasta su penúltima detención disponible en la base de datos fue utilizada como alimentación del modelo predictivo, con el objetivo de predecir si será detenido nuevamente y comparar esta predicción con la situación en la base de datos. La última detención en la base de datos no se utiliza para la predicción, sino para verificar la predicción del modelo.

En el experimento predictivo, cada individuo en la base de datos es etiquetado como "detenido" o "no-detenido". Un individuo tiene la etiqueta de "detenido" si la diferencia en días entre la penúltima y la última detención registrada es de entre 1 y 365 días. En caso contrario, cuando la diferencia es mayor a 365 días o tiene una sola detención registrada, el individuo es etiquetado como "no-detenido". La idea detrás de esta regla de etiquetar los individuos es que se puede predecir si un individuo será detenido dentro de una ventana de un año a partir de su última detención registrada en el modelo. Este periodo de un año se ha elegido como una ventana de tiempo en que ser detenido nuevamente podría ser interpretado como una indicación de habitualidad en el sistema y, por tanto, sugerir la priorización de recursos. Sin embargo, se puede usar los modelos predictivos con ventanas de tiempo distintas.

El modelo matemático predice si un individuo será detenido en el futuro a través de la información disponible de cada uno. Para poder validar la predicción, se excluye la última detención en la base de datos como atributo. Se utilizaron varios indicadores que se pueden extraer de los datos disponibles para predecir una redetención futura. Como caracterización de la historia delictual, se incluyó el número total de detenciones previas, la frecuencia –definida como el cociente entre el número de detenciones registradas y los años activos– y el número de detenciones por cada tipo de delito. Adicionalmente, se incluyeron atributos familiares y personales de cada individuo, tales como edad, género, registro de armas y número de detenciones de familiares.

La Tabla 3 muestra la matriz de confusión para el clasificador "árbol de decisión". Esta matriz permite visualizar el desempeño del algoritmo, en función de los rótulos reales y las predicciones hechas por él. Se observa que cuando el modelo predice que un sujeto no va a ser detenido, esta predicción es correcta en un 91% de los casos. Predecir las detenciones tuvo una precisión menor, con un 63% de éxito; es decir, cuando el modelo predice que un individuo será detenido en una ventana de un año posterior a la última detención, en un 63% de los casos esta predicción es correcta. Además de la precisión, es importante estudiar el *recall*, que mide cuántos casos reales el modelo es capaz de capturar en la predicción. Por ejemplo, de los individuos que tienen una nueva detención en la base de datos, un 57% fueron predichos como detenidos por el modelo. El modelo predictivo captura un 93% de los individuos no detenidos.

TABLA 3. **Matriz de confusión del modelo predictivo de árbol de decisión**

		clase predicha	
		no-detenido	detenido
clase real	no-detenido	76,6%	5,8%
	detenido	7,5%	10,1%

Fuente: elaboración propia.

Mientras la precisión y *recall* indican que el modelo automatizado tiene un valor predictivo significativo, hay limitaciones importantes en la predicción de ser detenido en el futuro. Por ejemplo, en un 5,8% de todos los casos, el modelo predice que el individuo será detenido, pero en realidad no lo fue (falsos positivos). En otras palabras, un 37% de las predicciones de ser detenido fueron erradas. En el caso de usar esta herramienta como apoyo en la toma de decisiones, los falsos positivos implican un alto costo social, ya que harían efectivo que policías investiguen a un individuo que, en realidad, no tenía una detención en la ventana de tiempo considerada. Esta proporción importante de falsos positivos presenta, en el contexto del fenómeno delictual, desafíos éticos mayores que en otros ámbitos del conocimiento. Asimismo, un 7,5% de los casos son falsos negativos, es decir, individuos que fueron detenidos, pero que el modelo predijo que no lo serían. Esto se relaciona con un 9% de las predicciones negativas, en que el modelo se confundió y en que el individuo no tiene la priorización de la policía que sería justificable, que es relevante especialmente en casos de delitos graves.

Hay una proporción significativa de predicciones falsas en el modelo que genera costos sociales altos, en caso de ser utilizado por el Estado. Esta limitación en el valor predictivo puede tener distintas causas como, por ejemplo, errores en los modelos, la falta de datos esenciales o tan solo que el crimen sea inherentemente difícil de predecir. Para estudiar si el modelo tiene un impacto grande en la precisión de la predicción, se comparó cinco distintos métodos de predicción: Naive Bayes, Regresión logística, Árbol de decisión, Random Forest y Multilayer perceptron (MLP). Mediante una validación cruzada, donde los datos son divididos en diez particiones estratificadas que preservan las particiones de las clases en la base de datos, cada método busca los parámetros que reproducen la realidad de la mejor forma. La Tabla 4 reporta las cifras de rendimiento para cada uno de los modelos, en términos de precisión y *recall*, y además el F-score y AUC, que son otras medidas de rendimiento de modelos predictivos más significativos. Se observa que los cinco tienen un desempeño muy similar.

TABLA 4. **Rendimiento alcanzado por modelos de predicción entrenados con una muestra de los datos, mediante validación cruzada estratificada con K=10 particiones**

Modelo	Precisión	*Recall*	F-Score	AUC
Naive Bayes	0,58	0,92	0,56	0,81
Regresión logística	0,69	0,41	0,51	0,81
Árbol de decisión	0,66	0,52	0,52	0,81
Random Forest	0,63	0,57	0,56	0,81
MLP	0,66	0,50	0,56	0,81

Fuente: elaboración propia.

La comparación del rendimiento de los modelos predictivos sugiere que, el solo considerar el modelo como tal, no puede explicar las limitaciones de las predicciones automatizadas. El rendimiento podría ser mejorado si se considerasen atributos adicionales como, por ejemplo, una historia de detenciones más larga o indicadores biográficos como educación o bienes. Sin embargo, un factor clave en la limitación es que el fenómeno de la delincuencia es simplemente difícil de predecir.

Por último, una de las ventajas en el modelo de Árbol de decisión es su capacidad para proveer un ordenamiento de los atributos relativo a la importancia de ellos en la clasificación. La Tabla 5 muestra algunos de los atributos más interesantes.

TABLA 5. **La importancia relativa de algunos atributos según el modelo Árbol de decisión**

Atributo	Importancia relativa
Período máximo entre detenciones	0,2089
Total de detenciones	0,1000
Frecuencia de detenciones	0,0530
Edad en el momento de la primera detención registrada	0,0030
Número de familiares con antecedentes	0,0010
Género	0,0001

Fuente: elaboración propia.

El atributo más relevante en la clasificación es el periodo máximo entre detenciones. Dado que el algoritmo puede generar reglas de decisión más elaboradas que una relación lineal, es muy probable que un periodo corto entre detenciones implique una probabilidad más alta de ser detenido en el futuro. En general, más que los datos biográficos, los indicadores relacionados con la historia de detenciones previas son los más significativos a la hora de predecir detenciones futuras. Mientras esta información es valiosa, la importancia relativa depende en gran parte del modelo utilizado, el experimento desarrollado, los datos disponibles y el tipo o tipos de delitos que se busca predecir.

Discusión

Los datos analizados confirman patrones ya conocidos en la literatura internacional: los hombres constituyen la mayor parte de quienes presentan alguna detención, la prevalencia de estas disminuye a mayor edad y hay diferencias según el tipo de delito. También se observa un alto nivel de especialización en ciertos tipos de delito, principalmente en aquellos contra la propiedad, de lesiones y violencia.

El estudio muestra que es posible caracterizar a la población en contacto con el sistema policial, a través del uso de datos administrativos tales como el historial de detenciones previas y variables sociodemográficas. Como se señaló, los resultados apuntan a seis conglomerados de individuos que se agrupan principalmente a partir del tipo de delito cometido.

En términos de predicción, los resultados del estudio indican una limitada capacidad predictiva a partir de los datos disponibles. Si bien, y de manera consistente con la literatura internacional (Laub y Sampson, 2003), es posible predecir quién no cometerá delitos, es difícil la generación de modelos predictivos precisos que permitan identificar a aquellos que tienen mayor

probabilidad de redetención. Esto conlleva a un análisis ético de las herramientas disponibles, dadas las consecuencias sociales asociadas a la identificación de falsos positivos.

1. Alcances del *big data* en criminología

Como se ha discutido en el marco conceptual, hay un debate relevante en la literatura internacional sobre la ética de usar el *big data* para predecir el crimen y su sesgo hacia grupos específicos de la población. Varios estudios indican que los modelos llevan problemas éticos (Kleiman et al., 2007; Hannah-Moffat, 2013) mientras otras sugieren que los sesgos no son más altos que los que ya presenta la policía (Brantingham et al., 2018). Dado que las herramientas automatizadas tienen la capacidad de procesar bases de datos enormes y encontrar patrones elaborados en un tiempo mucho más corto que una persona, las implicancias éticas deben ser incorporadas tanto en la decisión de usar estas técnicas como en la forma y propósito, en caso de hacerlo.

En general, se pueden distinguir tres etapas en el uso de big data en criminología: la recopilación de datos, el diseño del modelo matemático-estadístico y la incorporación de herramientas automatizadas en la toma de decisiones. Cada uno podría ser una fuente de sesgo y tener implicancias éticas.

a. Recopilación de datos

El rendimiento de las herramientas del *big data* está estrechamente vinculado con el volumen, variedad y veracidad de los datos disponibles. El volumen se refiere a la cantidad de datos disponibles. Mientras que hay una gran cantidad de registros en la base de datos, el estudio de trayectorias delictuales está limitado por la historia de detenciones desde 2009 y no más temprano, por lo cual no se puede investigar trayectorias de toda la vida de los individuos. Además, no se puede asegurar que todas las detenciones de una persona registrada están disponibles en la base de datos. Como ejemplo de variedad de los datos, la falta de información sobre las condenas y tiempo de incapacitación limita los alcances del modelo predictivo. En temas de veracidad, hay obstáculos porque la recopilación de los registros es semiautomatizada e involucra parcialmente un trabajo manual. Por ejemplo, la descripción del delito es declarada por el personal policial, que lleva a una ambigüedad entre delitos similares. Algunos registros tienen datos biográficos autodeclarados por el victimario como la educación. Dado que no se puede asegurar la veracidad de esta información, no se la incluyó en el análisis. Aún más sensible es información como la ubicación o comuna de la detención, que podrían inducir sesgos en desmedro de sectores vulnerables y ser un *proxy* para datos sociales o étnicos que no debieran ser parte del modelo predictivo por razones éticas. Por último, sirve señalar que cautelar la calidad de los datos es aún más desafiante si son proporcionados por actores externos.

Una limitación del estudio especialmente importante es que los resultados podrían tener sesgos generados por el tiempo de incapacitación por condenas. La literatura del involucramiento criminal y la reincidencia delictual señala la necesidad de considerar dicho factor en los modelos de análisis prospectivos (Nagin y Land, 1993). En el caso del análisis realizado, la información relativa a posibles encarcelamientos no estaba disponible. Esto implica que la predicción asume que los individuos, una vez detenidos, vuelven a quedar en libertad y, por tanto, vuelven a tener la oportunidad de delinquir. Como es de esperar, la probabilidad de encarcelamiento variará por el tipo de delito, por lo que conclusiones relativas a la falta de redetención de aquellos que, por ejemplo, han cometido un homicidio, puede ser consecuencia de la incapacitación de los mismos en el sistema penitenciario.

Como todos los estudios realizados con datos administrativos, el presente análisis incorpora dentro de su caracterización y predicción los sesgos propios imbuidos en el sistema de justicia. Por ejemplo, si hay diferencias en la probabilidad de ser detenido por un delito en función del lugar donde vives independiente del nivel de involucramiento delictual (algo que ha sido analizado en la literatura internacional, por ejemplo, por Sampson (1986)), este sesgo estaría siendo incorporado por la predicción automatizada, indicando más que una probabilidad de volver a cometer un delito, la probabilidad de ser detenido, independiente incluso de la comisión de un acto delictual. Esta observación, que también genera un sesgo hacia individuos ya sobrerrepresentados en las detenciones, hace imperativo el incluir datos de condenas en el modelo; ello permitirá distinguir entre las detenciones justificadas e injustificadas.

b. Diseño del modelo y algoritmo

Los modelos matemático-estadísticos que se usan normalmente en el *big data* tienen un fundamento científico sólido y son utilizados con gran éxito en distintas áreas de las ciencias y la industria. Si bien hay conocimiento científico del funcionamiento de la gran mayoría de los algoritmos, hay algunos métodos que no permiten la extracción de su funcionamiento en una forma que se puede interpretar como investigador como, por ejemplo, las versiones sofisticadas de redes neurales. Además, la extracción de información relevante de los modelos normalmente requiere una formación como experto en el tema del *big data*. Sin embargo, esta información es importante para comprender cómo los modelos generaron las predicciones, el grado de incertidumbre asociado a ellas, comprender sus limitaciones y estudiar si hay un posible sesgo.

c. El uso como herramienta automatizada

En términos de un potencial uso futuro de las herramientas automatizadas, en un escenario en que los actores policiales priorizan sus recursos en base

al modelo predictivo, hay un riesgo de crear una retroalimentación indeseable del sistema. Es decir, si hay un individuo marcado como un riesgo por el modelo, las policías aumentan los esfuerzos de investigación sobre dicho individuo y con ello aumentan su probabilidad de detención. En otras palabras, la policía podría ver que el modelo tiene una precisión alta en la predicción de detenciones, sin que esto signifique que se estén priorizando las investigaciones sobre los individuos con mayor potencial de daño social a través de actos delictuales.

Por último, hay una discusión ética sobre el propósito de usar las herramientas del *big data* en la toma de decisiones. Un ejemplo es la utilización de modelos estadísticos en el sistema penal como un apoyo en el juicio. El problema es que, si las herramientas automatizadas indican una probabilidad alta de ser el victimario, esto todavía no significa que se compruebe la culpabilidad de la persona. En el ambiente policial, la información entregada por el *big data* no debe ser utilizada como razón única de detener a una persona, sino como un apoyo adicional a los protocolos existentes para tomar decisiones.

2. Situación legal

La disponibilidad de datos y el poder descriptivo y predictivo de los modelos matemáticos han crecido exponencialmente y hoy se pueden usar herramientas computacionales en una variedad de entornos. En el ambiente de la criminología, la introducción de sistemas de información –que utilizan datos personales en gran escala y entrecruzan información de diversas fuentes administrativas– ha sido una demanda creciente de las policías y órganos de persecución penal, que observan en esta clase de sistemas una oportunidad para la prevención e investigación criminal. La utilización de estos sistemas ofrece, sin embargo, muchas interrogantes éticas y legales: ¿hasta qué punto el Estado puede utilizar información personal para propósitos de bien público, pero que no han sido informados ni consentidos por los ciudadanos? ¿Se puede utilizar el registro de detenciones anteriores de determinados individuos u otra información similar como base de acciones de prevención e investigación criminal?

Respecto de la recolección y uso de información personal, debe distinguirse entre datos personales de uso público y baja sensibilidad (la fecha de nacimiento, por ejemplo) y datos privados o de alta sensibilidad (como el diagnóstico de una determinada enfermedad o el historial penal de una persona). La primera clase de datos no ofrece inconveniente ético ni legal, pero los segundos deben obtenerse y usarse bajo condiciones reguladas. También los datos proporcionados por determinados organismos y con propósitos específicos (los puntajes obtenidos en pruebas estandarizadas de desempeño

académico) no podrían ser utilizados por órganos y propósitos diferentes (como oficinas que dirigen investigación criminal) sin los debidos resguardos y autorizaciones. El entrecruzamiento de información personal presenta diversas limitaciones éticas y legales (en un área, no obstante, poco tipificada legalmente), que resguardan los derechos de autonomía y privacidad que el estado constitucional moderno reconoce a sus ciudadanos.

El uso de información personal para propósitos de investigación criminal ofrece, asimismo, limitaciones más severas. En el derecho, se utiliza la expresión "derecho penal de autor" –que comprende los juicios que se realizan sobre un sujeto determinado no por los actos cometidos, sino por sus características personales– y el "derecho penal del enemigo" –que comprende la sanción que recae sobre un individuo en razón simplemente de su peligrosidad (por ejemplo, la internación en campos de detención de extranjeros en un estado de beligerancia). Tales formas del derecho son enteramente excepcionales y solo pueden aplicarse en condiciones extremas. En general, los estados constitucionales hacen prevalecer el principio de inocencia que resguarda a los individuos de cualquier pesquisa o sanción que no esté fundada en hechos debidamente acreditados ante un tribunal de justicia.

En el informe realizado para este estudio por Vergara, la autora concluye taxativamente que no se "podría iniciar una acción preventiva respecto de un individuo que aparezca dentro de una base de datos con un alto potencial de incidencia criminal, independiente de si nunca ha delinquido o si ya lo hizo y cumplió su condena"[3]. La criminología ha establecido claramente los sesgos de la investigación y persecución policial en contra de determinadas categorías de individuos estigmatizados socialmente (población que reside en barrios marginales, población inmigrante o de color, por ejemplo). La disposición de información personal podría sortear estos procesos de estigmatización (que siempre se basan en atributos sociales del individuo), pero se corre el riesgo de vulnerar derechos personales refrendados constitucionalmente.

Si bien uno de los mejores predictores de la conducta criminal es la historia delictiva previa de los individuos (y la reincidencia es un motivo de agravamiento de las sanciones judiciales), la ley concede a cualquier ciudadano el derecho de permanecer libre de toda pesquisa mientras no existan presunciones basadas en hechos y autorizadas legalmente. La colisión entre derechos personales y el resguardo de la seguridad de la sociedad constituye un problema ético y legal de gran consideración, que requiere un examen ético más detenido y definiciones legales más claras que las que existen actualmente.

3 Ignacia Vergara Caroca. Consideraciones ético-legales respecto del uso de datos personales para propósitos de prevención delictual. Concurso de Investigación de Pregrado en el Invierno, Vicerrectoría de Investigación, Pontificia Universidad Católica de Chile (2018). Para las consideraciones de este apartado se ha utilizado con mucho provecho el informe referido.

Propuestas

Los resultados del estudio en el uso del *big data* para entender de mejor forma la realidad de delincuencia en Chile sugieren las siguientes propuestas de políticas públicas.

1. Comenzar con un primer piloto de uso de herramientas automatizadas, como apoyo para que el personal de la PDI pueda mejorar la priorización de la asignación de recursos para sus labores de investigación. Condición necesaria para que un piloto sea éticamente viable es que los datos contengan información sobre condenas. De lo contrario, un algoritmo basado en detenciones solamente podría aumentar injustificadamente la probabilidad de ser detenido de individuos que no necesariamente representen una amenaza para su entorno.
2. Continuar con la investigación académica en la predicción de detenciones para evaluar si es factible utilizar el *big data* para predecir futuras condenas o, en su defecto, detenciones.
3. Agilizar el intercambio de la información administrativa que manejan actores gubernamentales, que permitan comprender el fenómeno criminal desde diferentes aristas.
4. Mejorar los procesos de recopilación de datos administrativos y estandarizar la clasificación de los tipos de delitos, para facilitar la integración y calidad de distintas bases de datos.
5. Abordar una discusión amplia e inclusiva sobre la ética del uso del *big data* para la identificación del comportamiento criminal y, de forma integrada, estudiar y modernizar el Código Penal para aclarar la usabilidad y limitaciones del aprendizaje automatizado.

1. Uso del *big data* en la organización de recursos policiales

El estudio permitió comparar el comportamiento criminal esperado de la literatura internacional con el análisis de los datos administrativos de detenciones en la Región Metropolitana de Santiago entre los años 2009 y 2018. Los resultados confirmaron la influencia de factores como edad y género en el tipo de delito que se comete, así como en la incidencia delictual. El análisis de las transiciones entre tipos de delitos de una misma persona indicó que la habitualidad criminal más común es la especialización en un tipo de delito. La generación automatizada de conglomerados de personas detenidas sugiere que la característica más diferenciadora es la historia delictual, especialmente el tipo de delito. Estos resultados confirman que se puede construir una tipología de comportamiento criminal para distintos grupos de detenidos, principalmente, pero no exclusivamente, en base a la historia de detenciones.

Las herramientas automatizadas pueden clasificar a cada persona que ingresa en la base de datos de detenciones como miembro de uno de los conglomerados. Esta tipología entrega información sobre el comportamiento criminal esperado del detenido como, por ejemplo, la habitualidad en términos del tipo de delito. Información de esta naturaleza es valiosa para la creación y funcionamiento de departamentos especializados en el análisis e investigación de delitos específicos. Al ser un proceso automatizado, permitiría una asignación más precisa de casos que la que hoy se realiza de manera manual, entregando mayor efectividad a los recursos policiales.

Mientras la asignación automatizada de recursos policiales puede implicar un funcionamiento más eficiente y adecuado, la experiencia del estudio también refuerza la necesidad de hacerlo de manera paulatina y de revalidar el análisis, dado que la calidad de la tipología depende de forma importante del volumen, variedad y veracidad de los datos disponibles. Asimismo, para evitar la toma de decisiones basadas en una mala interpretación de los resultados o en una excesiva confianza en sus resultados, se requiere personal con una formación en herramientas computacionales para el uso e interpretación de métodos matemático-estadísticos. Por este motivo, el piloto debería ser ejecutado por expertos de los actores policiales junto a otros que lo sean en herramientas de modelamiento del mundo académico, por ejemplo, en un centro especializado como el Cenacrimo. Además, las herramientas no deberían reemplazar los procedimientos actuales en esta fase inicial, sino incorporarlos como apoyo complementario a la toma de decisiones. Con estas precauciones, se puede desarrollar una herramienta que sea un apoyo real para mejorar los procedimientos de investigación policial.

Este piloto también es importante para una validación más profunda de los alcances del *big data* en criminología. Por ejemplo, en comparar la precisión de las herramientas automatizadas con la precisión de los protocolos actuales de priorización de recursos de investigación.

Para el piloto es necesario contar con datos de condenas. De lo contrario, trabajar con datos de solamente detenciones, que no cuentan con un juicio de por medio, puede aumentar la probabilidad de detención de un individuo inocente, pero con características que podrían aumentar su probabilidad de detención como, por ejemplo, vivir en un barrio vulnerable.

2. Estudios académicos en criminología predictiva

La experiencia internacional señala la dificultad de predecir trayectorias delictuales y reincidencia a nivel del individuo. Los resultados de los modelos predictivos realizados confirman esta complejidad. Es decir, la precisión de la

predicción de una detención futura, entre aquellas personas que ya han sido detenidas, es baja. Por un lado, la baja precisión de los modelos predictivos podría ser explicada por las limitaciones de los datos disponibles. Por otro, es posible que el fenómeno criminal tenga tal complejidad que sea difícil de predecir, en base a la información de la historia delictual y características personales disponibles. Este estudio no alcanza a responder el porqué detrás de la precisión de la predicción sobre futuras detenciones. Esto, en conjunto con los desafíos éticos que implica catalogar erradamente a un individuo como futuro reincidente (un falso positivo) con los costos sociales que esto implica, nos lleva a no recomendar aún el uso de herramientas de este tipo para la predicción del delito a nivel individual. Habiendo dicho esto, un estudio que queda pendiente y que sería de alta relevancia sería el poder comparar los alcances y el sesgo de la predicción automatizada con las predicciones que, actualmente, realizan las policías en terreno.

3. Intercambio de datos administrativos

Hay varios actores gubernamentales que administran datos de delito en Chile, como la PDI, Carabineros y la Fiscalía. Además, si se quieren cruzar estos datos con características personales tales como estado civil, propiedad de bienes raíces o automóviles, educación y situación financiera, la variedad de fuentes y la cantidad de organizaciones involucradas aumenta. El valor de los estudios de *big data* depende en gran medida de la disponibilidad de los datos que ayudan al analista a inferir probabilidades. Por este motivo, una centralización o agregación de la información administrativa tendría un impacto positivo en el alcance de modelos matemáticos para la descripción o, eventualmente, predicción del comportamiento delictual. Al mismo tiempo y por motivos de seguridad y privacidad, centralizar datos administrativos o compartir información sensible tiene que ser ejecutado con cuidado y solo si hay una necesidad. El entorno más adecuado para la centralización de datos sería una entidad especializada en el manejo de datos administrativos sobre delitos.

4. Estandarización de datos administrativos

Las distintas bases de datos administrativas son recopiladas por diferentes actores. Esta realidad resulta en problemas para la integración de las bases de datos y para el estudio integral del fenómeno criminal. Por ejemplo, el estudio de trayectorias criminales está principalmente limitado por la historia de eventos disponibles; en el caso del estudio presente, se remite a los últimos nueve años. Si bien hay datos históricos de detenciones anteriores, estos tienen una utilidad limitada, porque su formato es distinto y muchas veces han sido ingresados de forma manual. Los procesos manuales y poco estandarizados no solo generan más trabajo en su limpieza, sino también

introducen un sesgo en el proceso de generación de datos. Otro ejemplo para la necesidad de estandarización es que varios datos son elaborados manualmente: la información personal como educación y estado civil podrían ser autodeclarados por el victimario y la descripción del delito es normalmente registrada por el personal policial que hizo la detención. Una estandarización de los tipos de delitos y su categorización ayudaría a la agregación de distintas bases de datos y a la posibilidad de comparación entre distintos estudios. Hoy, la categorización está hecha de forma semiautomática, según distintas reglas formadas para el propósito de cada estudio.

5. Discusión ética y legal

Como se mencionó en la discusión sobre el uso del *big data* en criminología, los modelos matemáticos llevan alcances y limitaciones en ella. Dado que las implicancias de usar estas herramientas como apoyo en la toma de decisiones pueden tener costos sociales altos, se tiene que abordar una discusión amplia e inclusiva sobre la ética al utilizarlos y modernizar el sistema judicial a esta realidad. Este debate debe incluir a los distintos actores gubernamentales en el área, tales como las policías, la fiscalía y gendarmería, y los expertos académicos, entre otros. Dado que el poder computacional crece exponencialmente, hay una prioridad en abordar la discusión ética de manera amplia y temprana.

Conclusiones

El presente estudio tuvo como objetivo explorar los beneficios del uso de herramientas del *big data*, para entender el fenómeno de la delincuencia en Chile. En primer lugar, se analizaron los datos disponibles en cuanto a la consistencia con correlatos conocidos del delito, como son el género y la edad. Ambos factores se correlacionan tanto con la frecuencia como con el tipo de delito. Adicionalmente, se analizaron patrones de involucramiento delictual, a fin de detectar perfiles en torno al tipo de delito. Los datos muestran que existe una tendencia a la especialización, lo que se aprecia tanto en la matriz de transición como en los conglomerados estimados. Finalmente, el presente estudio muestra una predicción de detenciones futuras, con porcentajes significativos de falsos positivos y falsos negativos. Este último punto se conjuga con los desafíos éticos planteados sobre el uso del *big data* en predicción de la reincidencia delictual.

Si bien los datos administrativos utilizados permiten describir el fenómeno delictual y detectar patrones y perfiles entre quienes entran en contacto con la policía, en términos de predicción de detenciones futuras, la utilidad de dichos datos es restringida. Por lo mismo, el análisis del *big data* puede ser una herramienta útil para la priorización de los recursos de investigación de

las instituciones correspondientes. Sin embargo, si se quisiese utilizar dichos modelos para predecir detenciones futuras, debiésemos estar conscientes que su poder en ese sentido es limitado y, a la vez, los costos sociales asociados a un error en la predicción son extremadamente altos en el contexto de la justicia criminal, por lo que se no se aconseja hacerlo.

Para una mayor robustez en las conclusiones, un estudio futuro que podría ser de gran interés es comparar los alcances de los algoritmos del *big data*, tales como la precisión en la predicción de futuras detenciones y el sesgo hacia grupos de individuos específicos, con los protocolos que usan actualmente las policías al priorizar sus recursos. Asimismo, para disminuir posibles sesgos en los falsos positivos sobre poblaciones vulnerables e inocentes, poder distinguir entre detenciones justificadas e injustificadas sería esencial, por lo cual es necesario contar con datos de condenas.

Referencias

Akers, R.L., 1999. *Criminological Theories*. Fitzroy Dearborn Publishers, Chicago, IL.

Auerhahn, K., 1999. Selective incapacitation and the problem of prediction. *Criminology*, 37(4), pp. 703-734.

Bartolucci, F., Pennoni, F. y Francis, B., 2007. A latent Markov model for detecting patterns of criminal activity. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(1), pp. 115–132.

Berk, R.A. y Bleich, J., 2013. Statistical procedures for forecasting criminal behavior: A comparative assessment. *Criminology & Public Policy,* 12(3), 513-544.

Block, C., Arjan, R., Blockland, A.J., Van der Werff, C., Van Os, R. y Nieuwbeerta, P., 2010. Long-term patterns of offending in women. *Feminist Criminology* 5(1), pp. 73-107.

Brantingham, P.J., Valasik, M. y Mohler, G.O., 2018. Does Predictive Policing Lead to Biased Arrests? Results From a Randomized Controlled Trial. *Statistics and Public Policy*, 5(1), pp. 1-6.

Britton, D.M., 2011. *The Gender of Crime*. Lanham: Rowman & Littlefield.

Centro de Estudios Públicos, 2017. *Estudio Nacional de Opinión Pública*, Nº 81.

Collins, L.M. y Lanza, S.T., 2010. Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences, Vol. 718. John Wiley & Sons.

Domínguez, P. y Raphael, S., 2015. The role of the cost-of-crime literature in bridging the gap between social science research and policy making. *Criminology & Public Policy*, 14(4), pp. 589–632.

Dressel, J. y Farid, H., 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1), eaao5580.

Fábrega, J., Morales, A.M. y Muñoz, N., 2014. Delito y especialización en Chile. *Política criminal,* 9(18), pp. 521-542

Farrington, D.P., 2003. Key results from the first forty years of the Cambridge study in delinquent development. *In Taking stock of delinquency,* pp. 137-183. Springer, Boston, MA.

Figure Eight, 2017. *Data Scientist Report 2017*. URL: https://www.figure-eight.com/download-2017-data-scientist-report/.

Gottfredson, M.R. y Hirschi, T., 1990. *A General Theory of Crime.* Stanford, CA: Stanford University Press.

Hagenaars, J.A. y McCutcheon, A.L. (Eds.), 2002. *Applied latent class analysis.* Cambridge University Press.

Hannah-Moffat, K., 2013. Actuarial sentencing: An "unsettled" proposition. *Justice Quarterly*, 30(2), pp. 270-296.

Kleiman, M., Ostrom, B. J., y Cheesman, F. L., 2007. Using risk assessment to inform sentencing decisions for nonviolent offenders in Virginia. *Crime & Delinquency*, 53(1), pp. 106-132.

Lara, J.C., Pincheira, C. y Vera, F., 2013. La privacidad en el sistema legal chilena. Policy paper, ONG Derechos Digitales.

Laub, J., 2015. *Understanding inequality and the justice system response: Charting a new way forward.* William T. Grant Foundation.

Laub, J.H. y Sampson, R.J., 2003. *Shared beginnings, divergent lives: Delinquent boys to age 70.* Cambridge, MA: Harvard University Press.

Maaten, L.V.D. y Hinton, G., 2008. Visualizing data using t-SNE. J*ournal of machine learning research*, 9(Nov), pp. 2579-2605.

Moffitt, T.E., Caspi, A., Rutter, M. y Silva, P.A., 2001. *Sex Differences in Antisocial Behaviour: Conduct Disorder, Delinquency, and Violence in the Dunedin Longitudinal Study.* Cambridge: Cambridge University Press.

Murphy, K., 2012. *Machine Learning: A Probabilistic Perspective.* MIT Press.

Nagin, D.S., 2005. *Group-based modeling of development.* Harvard University Press.

Nagin, D.S. y Land, K.C., 1993. Age, criminal careers, and population heterogeneity: Specification and estimation of a nonparametric, mixed Poisson model, *Criminology*, 31(3), pp. 327-362.

Nagin, D.S. y Paternoster, R., 1991. On the relationship of past to future participation in delinquency. *Criminology*, 29(2), pp. 163-189.

Nagin, D. y Paternoster, R., 2000. Population heterogeneity and state dependence: State of the evidence and directions for future research. *Journal of Quantitative Criminology*, 16(2), pp. 117-144.

Ozkan, T., 2017. Predicting Recidivism *Through Machine Learning*. PhD thesis, The University of Texas at Dallas.

Petersilia, J., 1980. Criminal career research: A review of recent evidence. *Crime and justice*, 2, pp. 321-379.

Piquero, A.R., 2008. Taking stock of developmental trajectories of criminal activity over the life course'. *In The long view of crime: A synthesis of longitudinal research* (pp. 23-78). Springer, New York, NY.

Ridgeway, G., 2018. Policing in the Era of Big Data. *Annual Review of Criminology*, 1, pp. 401-419.

Sampson, R.J., 1986. Crime in cities: The effects of formal and informal social control. *Crime and justice*, 8, 271-311.

Sampson, R.J. y Laub, J.H., 1993. *Crime in the making: Pathways and turning points through life.* Cambridge, MA: Harvard University Press.

Sampson, R.J. y Laub, J.H., 2005. Seductions of Method: Rejoinder to Nagin and Tremblay's developmental Trajectory Groups: Fact or Fiction?. *Criminology*, 43, pp. 905-913.

Schwartz, J., Steffensmeier, D., Zhong, H., y Ackerman, J., 2009. Trends in the gender gap in violence: Reevaluating NCVS and other evidence. *Criminology*, 47(2), 401-425.

Shearer, C., 2000. The CRISP-DM model: the new blueprint for data mining. Journal of data warehousing, 5(4), pp. 13-22.

Wirth, R. y Hipp, J., 2000. CRISP-DM: Towards a standard process model for data mining. *In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* pp. 29-39.

Wolfgang, M.E., Figlio, R.M. & Sellin, T. 1987. *Delinquency in a birth cohort.* University of Chicago Press.

Zeng, J., Ustun, B. & Rudin, C., 2017, 'Interpretable classification models for recidivism prediction', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180(3), 689–722.

CÓMO CITAR ESTA PUBLICACIÓN:

Van ´t Wout, E., Valenzuela, E., Asahi, K., Pieringer, C., Torres, D. y Larroulet, P., 2019. Big data para la identificación de comportamiento criminal. En: Centro de Políticas Públicas UC (ed), Propuestas para Chile. *Concurso de Políticas Públicas 2018*. Santiago: Pontificia Universidad Católica de Chile, pp. 49-78.