Facultad de Ciencias

Grado de Estadística Aplicada

Homicidios y drogas: un estudio basado en redes bayesianas

Blanca Domènech Pellejà

Tutora: Rosario Delgado de la Torre

Barcelona, 29 de enero de 2021

Agradecimientos

A Rosario Delgado, por su magnífica dedicación y asesoramiento, sin sus correcciones y consejos no hubiera sido posible realizar el proyecto.

Al Ministerio del Interior, por dejarme usar su base de datos para aprender más sobre el uso de la estadística.

A todas las compañeras del doble grado, que sin su compañía el camino hubiera sido más difícil.

A toda mi familia, abuelos, padre, madre, hermano y hermana, por confiar en mí desde el principio de la etapa universitaria y estar a mi lado en todas las decisiones que tomo.

Por último, a Maria Queralt, mi amiga de toda la vida y para toda la vida, por las horas de escucha, sus opiniones y su apoyo en los momentos de mayor dificultad y debilidad.

Gracias a todas las personas aquí mencionadas por hacer este camino más agradable.

Resumen

El estudio del fenómeno del homicidio principalmente se ha desarrollado desde la psicología y la criminología, aunque con la mayor disponibilidad de herramientas tecnológicas e informáticas se han podido recopilar bases de datos y acercase al fenómeno desde la estadística. El objetivo de este trabajo es utilizar las redes bayesianas para realizar el perfil criminológico, teniendo en cuenta la relación con las drogas, y predecir las características del autor desconocido dado un nuevo homicidio.

Se han construido cuatro modelos diferentes teniendo en cuenta las situaciones en que algunas variables del hecho se usan como evidencias y en otras como variables a predecir. Además, como se quiere predecir más de una característica del autor desconocido, se han comparado los modelos obtenidos con procedimientos diferentes, el *Binary Relevance* y el *Chain Classifier*, obteniendo mejores resultados con el segundo.

Los resultados indican es más probable que el autor beba alcohol con habitual frecuencia cuando se tiene evidencia de que la víctima bebe alcohol y lo hace habitualmente.

Hay que tener en cuenta que los resultados no representan la realidad al 100 %, sino que se trata de un modelo probabilístico y su uso es orientativo para la resolución de un nuevo homicidio y para que las personas investigadoras puedan combinar con otras técnicas policiales.

Resum

L'estudi del fenomen de l'homicidi principalment s'ha desenvolupat des de la psicologia i la criminologia, encara que amb la disponibilitat major d'eines tecnològiques i informàtiques s'han pogut recopilar bases de dades i aproximar-se al fenomen des de l'estadística. L'objectiu d'aquest treball és usar les xarxes bayesianes per realitzar el perfil criminològic, tenint en compte la seva relació amb les drogues, i predir les característiques de l'autor desconegut donat un nou homicidi.

S'han construït quatre models diferents tenint en compte les situacions que algunes variables del fet s'usen com a evidències i en d'altres com a variables a predir. A més, com que es vol predir més d'una característica de l'autor desconegut, s'han comparat els models obtinguts amb procediments diferents, el Binary Relevance i el Chain Classifier, obtenint millors resultats amb el segon.

Els resultats indiquen que és més probable que l'autor begui alcohol amb una freqüència habitual quan es té l'evidència que la víctima beu alcohol i ho fa habitualment.

Cal tenir present que els resultats no representen la realitat al 100 %, sinó que es tracta d'un model probabilístic i el seu ús és orientatiu per a la resolució d'un nou homicidi i que les persones investigadores poden combinar amb altres tècniques policials.

Abstract

The study of the phenomenon of homicide has mainly been developed from psychology and criminology, although with the greater availability of technological and computer tools it has been possible to compile databases and approach the phenomenon from statistics. The aim of this work is to use Bayesian networks to perform criminal profiling, taking into account the relationship with drugs, and to predict the characteristics of the unknown perpetrator of a new homicide.

Four different models have been built taking into account situations where some variables of the event are used as evidence and others as variables to be predicted. Moreover, as more than one characteristic of the unknown perpetrator is to be predicted, the models obtained with different procedures, the Binary Relevance and the Chain Classifier, have been compared, obtaining better results with the latter.

The results indicate that perpetrator is more likely to drink alcohol regularly when there is evidence that the victim drinks alcohol and does so regularly.

It should be borne in mind that the results do not represent 100 % reality, but they are a probabilistic model and their use is for guidance for the resolution of a new homicide and for investigators to combine with other police techniques.

Índice

Índice de cuadros

Índice de figuras

Lista de símbolos y abreviaturas

Variable	**Definición**
Θ	Dominio de todas las CPTs de un DAG específico, donde $\Theta = (\theta_1, ..., \theta_n)$
Θ^{Γ}	Vector de parámetros asociado al DAG
Θ^{Γ}_{MLE}	Vector de parámetros estimado a partir del método MLE
θ_i	$\theta_i \in \Theta$ es una CPT para cada variable X_i
Γ	Directed Acyclic Graph, formado por ($\mathbf{V}$, $\mathbf{E}$)
Γ^n	Conjunto de DAGs, cada uno con un total de n nodos
α	Nivel de significación de la prueba
A	Matriz de confusión, donde $A = (a_{ij})_{i,j=1,...,r}$
AIC	Akaike Information Criterion
b^t	$b^t = \{x^t_1, x^t_2, ..., x^t_{m^t}\}$. Bolsa o *bag* t, que es un conjunto de casos de tamaño m^t
BIC	Bayesian Information Criterion
BR	Binary Relevance
c_i	Caso i-éssimo del conjunto de datos D
CC	Chain Classifier
CL	Confidence Level
CPT	Conditional Probability Table
d	Dimensión de Γ, es decir, los parámetros no redundantes en Θ^{Γ}
D	Conjunto de datos. $D = \{c_1, ..., c_M\}$
DAG	Directed Acyclic Graph
$\mathbf{E}$	Conjunto de pares de nodos ordenados de distintos elementos de $\mathbf{V}$. Son las aristas (o arcos) de un DAG
i.i.d.	Independientes e identicamente distribuidas
I_P	Independencia condicional
IPA	Individual Predictive Accuracy
k	Número de particiones que se hacen en la muestra original para realizar la validación cruzada (*cross validation*)
$L^D(\Theta^{\Gamma}_{MLE})$	Valor de la función de verosimilitud evaluada por Θ^{Γ}_{MLE} a partir del conjunto de datos D
M	Número total de casos en el conjunto D
m^t	Número total de casos en el *bag* t
MI	Multi-Instance
ML	Multi-Label
MLE	Maximum Likelihood Estimation

Variable	**Definición**
n	Número total de variables
N	Suma de los elementos de la matriz de confusión, es decir, el número total de predicciones. $N = \sum_{i=1}^{r} \sum_{j=1}^{r} a_{ij}$
ND_{X_i}	Conjunto de no descendientes del nodo (variable) X_i
OPA	Overall Predictive Accuracy
P	Distribución de probabilidad conjunta para un conjunto de variables aleatorias $X_i \in \mathbf{V}$ para $i = (1, ..., n)$
PA_{X_i}	Conjunto de padres del nodo (variable) X_i
r	Número de categorías de la variable clase
r^t	Clase del *bag* b^t
RB	Red Bayesiana
T	Conjunto de datos de entrenamiento (*training data set*)
T_h	Conjunto de datos de entrenamiento para k = h cuando se usa la técnica (*k-fold cross validation*)
$\mathbf{V}$	Conjunto finito de variables aleatorias discretas
V	Conjunto de datos de validación (*validation data set*)
V_h	Conjunto de datos de validación para k = h cuando se usa la técnica (*k-fold cross validation*)
X_i	Variable aleatoria discreta i que pertenece a $\mathbf{V}$, con $i = (1, ..., n)$
$\{X\}$	Conjunto de X

1. Introducción

El homicidio es el hecho delictivo que consiste en terminar con la vida de otra persona de forma dolosa o culposa (González et al., 2018). En los últimos años el interés sobre el fenómeno ha aumentado porque en las sociedades modernas se considera la manifestación más violenta del comportamiento criminal (Liem, 2013) y, además, impacta en los aspectos psicosociales, políticos y socioeconómicos de un país (González et al., 2018). Asimismo, la Oficina de las Naciones Unidas contra la Droga y el Delito (ONUDD[1]) afirma que el homicidio constituye uno de los indicadores más potentes para medir el nivel de violencia de un país porque su impacto puede generar un entorno de miedo, incertidumbre e inseguridad (UNODC, 2013).

Aunque en España la tasa de homicidios sea de las más bajas no significa que su estudio sea menos importante porque, como se ha dicho anteriormente, se vulnera el derecho fundamental de la vida, que se pone de manifiesto en el artículo 15 de la Constitución Española[2] (González et al., 2018).

En el momento en que se comete un homicidio es crucial que el sistema de justicia penal sea efectivo y dicte sentencia justa para las personas presuntas homicidas porque la impunidad de estas puede propiciar que se cometan más delitos (González et al., 2018). Por esa razón es de interés recopilar datos sobre las características del autor, hecho y víctima con el objetivo de estudiar casos esclarecidos en el pasado para construir modelos con fines identificativos y predictivos. Así, un investigador criminal podría usar dicho modelo para hipotetizar el perfil de los autores desconocidos de nuevos homicidios. Es decir, cuál es el tipo general de persona que suele cometer este tipo de hecho, orientando la búsqueda y priorizando sospechosos.

Desde hace unas décadas, las personas investigadoras han contemplado la posibilidad de hacer conjeturas acerca del tipo de persona que ha podido cometer un homicidio o si existe la posibilidad de identificar los elementos que diferencian a los autores (Pecino, 2019). En este contexto se ha desarrollado la técnica de la perfilación criminal, que principalmente se basa en la experiencia personal de los investigadores criminales y los psicólogos forenses, más que en métodos científicos empíricos (Palermo and Kocsis, 2004). Diferentes autores han defendido la aproximación estadística para la elaboración de instrumentos de ayuda en la decisión para las investigaciones policiales, para analizar el homicidio desde otro enfoque y reducir los errores causados por prejuicios culturales y malas interpretaciones de los investigadores (Baumgartner et al., 2008).

[1] Las siglas en inglés UNODC: *United Nations Office on Drugs and Crime*

[2] Todos tienen derecho a la vida y a la integridad física y moral, sin que, en ningún caso, puedan ser sometidos a tortura ni a penas o tratos inhumanos o degradantes. Queda abolida la pena de muerte, salvo lo que puedan disponer las leyes penales militares para tiempos de guerra.

Gracias a la mayor disponibilidad de tecnologías informáticas y de la información, los organismos encargados de hacer cumplir la ley han podido recopilar bases de datos con información detallada del delincuente y del lugar del delito. Por consiguiente, importantes autores han defendido que las técnicas de aprendizaje automático desempeñan un papel importante en el desarrollo de herramientas de ayuda a la decisión para las investigaciones policiales (Baumgartner et al., 2008).

Uno de los enfoques consiste en aprender un modelo de red bayesiana (RB) de comportamiento del delincuente a partir de los datos y, posteriormente, implementar el modelo de elaboración de perfiles mediante un motor de inferencia. Además, las características del delincuente inferidas incluyen niveles de confianza que representan su exactitud esperada. Así pues, la persona investigadora puede establecer cuáles son las predicciones fiables (Baumgartner et al., 2008).

En el presente trabajo se hace una aproximación estadística al fenómeno del homicidio mediante un análisis descriptivo y la elaboración de un clasificador bayesiano de tipo RB que modela el comportamiento del homicida de España, con mayor atención a su relación con las drogas, con el fin de predecir el perfil del delincuente en casos no resueltos. Esta técnica se conoce como perfilación criminal. Para hacerlo se usa una base de datos de homicidios esclarecidos de la Secretaría de Estado de Seguridad del Ministerio del Interior.

Dado que se pretende predecir más de una característica del homicida, se han usado dos procedimientos: el *Binary Relevance* y el *Chain Classifier*. La diferencia es que en el primero se asume independencia entre las variables del autor y se construye una RB para cada variable independientemente de la otra. En el segundo se construye una RB para cada variable en cadena, es decir, las variables predichas anteriormente se tienen en cuenta para predecir las que faltan.

Se ha visto que algunas características del hecho a veces son conocidas y a veces no, lo que hace que se necesiten diferentes modelos que tengan en cuenta en qué situación se encuentra la persona investigadora. Para ello se han construido 4 modelos diferentes que tienen en cuenta si el número de autores y/o la tipología del hecho son desconocidos.

Una vez ajustadas las RBs mediante *Binary Relevance* y *Classifier Chain* para las 4 situaciones diferentes, se han evaluado y validado los resultados para elegir el procedimiento que tenga mayor rendimiento predictivo.

Para la realización del trabajo se usa el software libre  , que tiene múltiples paquetes para aprender y ajustar redes bayesianas. En este estudio se ha trabajado con el paquete `bnlearn` por ser uno de los más versátiles (Scutari, 2010). También se han usado paquetes como `gRain` y `gRbase` para la predicción de las redes y `Rgraphviz` para la creación de gráficos. Finalmente, para la realización de la aplicación web Shiny se han usado los paquetes `shiny` y `shinydashboard`.

2. El homicidio

El anhelo internacional por el estudio del homicidio se manifiesta, por una parte, por la creación de diferentes instituciones como la ONUDD y la Organización Mundial de la Salud (OMS), y por otra, por la creación de múltiples organizaciones, formadas por investigadores y profesionales especializados en homicidios, dedicadas al estudio empírico del homicidio, como el Homicide Working Group, el European Homicide Research Group o el Murder Accountability Project (Pecino, 2019).

La repercusión de los homicidios va más allá de las personas fallecidas, extendiéndose la victimización a los familiares y personas cercanas a la víctima mortal. Además, los homicidios generan alarma social porque se mantienen a largo plazo en la memoria colectiva, y tienen consecuencias directas en las valoraciones que hace la ciudadanía sobre la eficiencia de los cuerpos policiales. Todo esto tiene un gran efecto en la percepción de inseguridad que tiene la sociedad (González et al., 2018).

Es por esto que un estudio pormenorizado de todos los aspectos de la conducta homicida es una herramienta fundamental para monitorizar la seguridad, la justicia y el desarrollo de los países (UNODC, 2015).

Las disciplinas científicas dedicadas al estudio de los homicidios son heterogéneas, pero la criminología y la psicología son las que han tenido un papel más relevante y las que han generado múltiples teorías criminológicas. Igualmente, se han desarrollado distintas tendencias en el estudio del homicidio: estudio focalizado en los factores de riesgo asociados a la etiología de los homicidios, estudio de las modalidades y clasificaciones de los homicidios y de los homicidas, estudio de la víctima y, finalmente, el estudio de la predicción, donde se ha hecho uso de la técnica de la perfilación criminal (Ioannou and Hammond, 2015).

También han sido ampliamente estudiados los factores de riesgo de los homicidios procedentes del ámbito familiar que ejercen una influencia negativa en el desarrollo y maduración de los infantes. La crianza en familias donde es habitual el consumo de alcohol y drogas, la presencia de enfermedades mentales y la experiencia criminal de los progenitores, están significativamente asociados con los homicidios en edades adultas (Pecino, 2019). Además, hay estudios que señalan que los patrones de adicción y consumo son más comunes entre las personas condenadas por homicidio que entre la población general, pudiendo ser un factor de riesgo de victimización (González et al., 2018).

2.1. La perfilación criminal

Las técnicas policiales usadas en las investigaciones criminales contribuyen de manera significativa al esclarecimiento de delitos perseguidos por las Fuerzas y Cuerpos de Seguridad, ya que las evidencias

permiten realizar una adecuada reconstrucción de los hechos e informar sobre la identidad de los posibles autores (Jiménez, 2012).

La perfilación criminal se conoce por múltiples términos anglosajones, como por ejemplo: *offender profiling, criminal profiling, psychological profiling, crime scene profiling* y se puede definir como una "técnica psicológica dirigida a servir de ayuda a la identificación y detención de delincuentes dentro del proceso de una investigación criminal" (Farrington and Lambert, 2017, p. 135).

Entre los principales objetivos de la perfilación criminal destaca el de predecir las características de un delincuente desconocido en un caso particular, con el fin de reducir el número de posibles sospechosos, a partir de determinados aspectos del crimen cometido (Garrido and Sobral, 2008). De este modo, no se encamina a señalar a una persona en concreto, sino a sugerir qué tipo de persona es la que más probablemente sea la autora de un delito, lo que permite orientar la investigación. Por lo tanto, esta técnica es útil para incrementar la eficacia de los cuerpos policiales y el sistema judicial, ya que los investigadores priorizan sus actuaciones y recursos, reduciendo los tiempos de esclarecimiento y la tasa de hechos sin esclarecer (González et al., 2018).

Por esas razones es de interés recopilar datos sobre las características del autor, hecho y víctima con el objetivo de estudiar casos esclarecidos en el pasado para construir modelos con fines identificativos y predictivos, hipotetizando el perfil de los autores de casos nuevos (González et al., 2018).

En la práctica actual, la elaboración de perfiles delictivos se basa principalmente en la experiencia personal de los investigadores criminales y de los psicólogos forenses, más que en métodos científicos empíricos (Palermo and Kocsis, 2004), aunque diferentes autores han defendido la aproximación matemática. Es por ello que en este trabajo se utiliza el enfoque de la red bayesiana, un modelo estadístico con el fin de realizar la perfilación criminal del homicida en España.

2.2. El homicidio en España

En España el Código Penal recoge las definiciones de homicidio doloso, homicidio imprudente y asesinato. Siguiendo la tradición internacional, para el estudio del homicidio se tiene en cuenta los homicidios dolosos y los asesinatos, dejando de lado los homicidios imprudentes porque no existe intención o dolo en su comisión. Por lo tanto, de aquí en adelante el término homicidio hará referencia a estos dos tipos de hechos delictivos (puede consultar las definiciones en el Anexo A).

Para que un homicidio se considere intencional o doloso son necesarias las tres condiciones siguientes: que haya un sujeto que lleve a cabo la acción; un sujeto que resulte muerto; y que haya voluntad de matar, que es lo que se conoce como "dolo", con independencia del grado de premeditación (González et al., 2018).

En España la tasa de homicidio es de 0.66 por cada 100,000 habitantes (INE, 2017), cifra que se sitúa por debajo de la media europea y mundial, donde la tasa de homicidio en el 2017 era de 3 y 6.1 por cada 100,000 habitantes, respectivamente (UNODC, 2019). A nivel europeo la comparativa sitúa a España como uno de los países con la tasa más baja.

3. Las Redes Bayesianas

Las Redes Bayesianas (RB) son modelos gráficos de relaciones probabilísticas entre un conjunto de variables finito. Estas estructuras gráficas se utilizan para representar el conocimiento sobre situaciones de incertidumbre y comprender las relaciones entre las variables que afectan a un determinado fenómeno, el homicidio en este caso.

Esta metodología permite hacer inferencia bayesiana, es decir, estimar la probabilidad posterior de las variables desconocidas basándose en las evidencias sobre las variables conocidas. A medida que se conocen nuevas evidencias, las RBs actualizan sus probabilidades (a esto se le conoce como propagación bayesiana). Además, pueden dar información sobre cómo se relacionan las variables del dominio, las cuales pueden ser interpretadas en ocasiones como relaciones de causa-efecto.

En este documento se utilizan la RBs para elaborar el perfil criminológico a partir de los datos y, posteriormente, implementar el modelo para identificar el tipo de persona que es más probable que sea la autora.

3.1. Conceptos técnicos de las Redes Byesianas

Una RB es una pareja (Γ, P), donde Γ es un gráfico acíclico dirigido (DAG[3]), que representa las relaciones de dependencia entre las variables, y P la distribución de probabilidad conjunta de las n variables aleatorias $\mathbf{V} = \{X_1, ..., X_n\}$. Las variables pueden ser discretas, binarias u ordinales, pero no continuas.

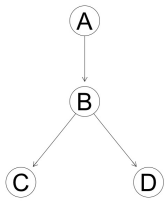
Formalmente, un gráfico dirigido es un par ($\mathbf{V}$, $\mathbf{E}$), donde $\mathbf{V}$ es un cojunto finito de variables aleatorias discretas, cuyos elementos son llamados nodos (o vértices), y $\mathbf{E}$ es un conjunto de pares de nodos ordenados de distintos elementos de $\mathbf{V}$, llamados arcos (o aristas), y representan las relaciones condicionales. Al conjunto de arcos que conectan dos nodos se les llama camino, y en los gráficos acíclicos no existen caminos que empiecen en un nodo y terminen en este mismo nodo.

En la Figura 1 se ejemplifica un DAG $\Gamma = (\mathbf{V}, \mathbf{E})$, donde $\mathbf{V} = \{A, B, C, D\}$ son las variables o nodos, y $\mathbf{E} = \{(A, B), (B, C), (B, D)\}$ los arcos dirigidos. En esta ocasión, el nodo A es padre del nodo B

[3]Por sus siglas en inglés *Directed Acyclic Graph*.

porque hay un arco dirigido de A a B. El conjunto de padres de B se denota como PA_B. El nodo A es un antepasado de D porque existen un conjunto de arcos dirigidos que conectan los nodos y, por consiguiente, D un descendiente de A. El nodo A es un nodo raíz, ya que no tiene padres, y los nodos C y D son nodos hojas, porque no tienen hijos. Por último, B es un nodo intermedio porque no es ni un nodo raíz ni un nodo hoja.

Figura 1: Ejemplo de un gráfico acíclico dirigido (DAG)



Fuente: Elaboración propia.

A los nodos que les llega algun arco están asociados a una tabla de probabilidad condicionada (CPT[4]), que muestra la probabilidad de cada valor que toma el nodo, dados los posibles valores de los padres. A los nodos que no les llega ninguna flecha, que no tienen padres, tienen asociada una tabla de probabilidad (PT[5]) que muestra la probabilidad de cada valor del nodo.

3.1.1. Independencia condicional

La estructura gráfica de las RBs, el DAG, permite ver las dependencias e independencias entre las variables aleatorias del dominio **V**. Si dos variables son independientes dado el estado de una tercera variable, entonces se dice que son condicionalmente independientes.

Específicamente, sean X, Y y Z variables aleatorias, X e Y son condicionalmente independientes dado Z si y solo si, dado cualquier valor de Z, digamos z, la distribución de probabilidad de X e Y es la misma independientemente del valor de Z, siempre y cuando la probabilidad de Z sea diferente de zero. Es decir:

$$P(\, X = x, Y = y \,|\, Z = z) = P(X = x \,|\, Z = z)\; P(Y = y \,|\, Z = z)$$

La notación para expresar que "X e Y son condicionalmente independientes dado Z" es:

$$I_P(X, Y \,|\, Z)$$

[4]Por sus siglas en inlgés significa "Conditional Probability Table".
[5]Por sus siglas en inglés significa "Probability Table".

3.1.2. Condición de Markov

La condición de Markov caracteriza las RBs y se define de la siguiente forma:

Definición 1 (Def. 1.9 Neapolitan, 2004) *Se supone que se tiene la distribución de probabilidad conjunta P de las variables aleatorias en algún cojunto de $\mathbf{V}$ y un DAG $\Gamma = (\mathbf{V}, \mathbf{E})$. Se dice que (Γ, P) satisface la condición de Markov si para cada variable $X \in \mathbf{V}$, $\{X\}$ es condicionalmente independiente del conjunto de todos sus no descendientes dado el conjunto de todos sus progenitores. Esto significa que si se denota el conjunto de padres de X y los no descendientes de X como PA_X y ND_X, respectivamente, entonces:*

$$I_P(\{X\}, ND_X \,|\, PA_X)$$

Si X es un nodo raíz, entonces el conjunto de padres de X, PA_X, está vacio. Entonces para esta situación la condición de Markov significa que el conjunto de X, $\{X\}$, y ND_X son independientes.

Con la Definición 1 se puede conocer si se satisface la *condición de Markov* a partir de un DAG y la distribución de probabilidad conjunta P, es decir, dada la pareja (Γ, P).

Teorema 1 (Th. 1.4 Neapolitan (2004)) *Si (Γ, P) satisface la condición de Markov, entonces P es igual al producto de las distribuciones condicionales de todos los nodos dados los valores de sus padres, siempre que estas distribuciones condicionales existan. Es decir, si $\boldsymbol{V} = \{X_1, ..., X_n\}$ para todos los valores posibles x_i de X_i, se tiene que:*

$$P(X_1 = x_1, ..., X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i | PA_{X_i})$$

La igualdad anterior se conoce como **regla de la cadena**. Cabe mencionar que si una distribución de probabilidad conjunta P satisface la condición de Markov en un DAG, entonces también se cumple la **regla de la cadena**.

El Teorema 1 dice que si se empieza con una distribución conjunta que satisface la condición de Markov con algún DAG, los valores en esa distribución conjunta estarán dados por el producto de las distribuciones condicionales. Sin embargo, se debe trabajar a la inversa. Se debe empezar con un DAG, Γ, que represente las relaciones de dependencia entre las variables del problema $\mathbf{V}$, después encontrar o estimar las distribuciones de probabilidad condicionales y, finalmente, poder concluir que el producto de estas distribuciones es una distribución conjunta P que satisface la condición de Markov con algún DAG. El teorema siguiente permite hacer justamente esto:

Teorema 2 (Th. 1.5 Neapolitan, 2004) *Se da un DAG Γ donde cada nodo es una variable aleatoria discreta y se especifica una distribución de probabilidad condicional de cada nodo dados los valores de sus padres en Γ. Entonces el producto de estas distribuciones condicionales dan una*

distribución de probabilidad conjunta P de las variables (según la regla de la cadena), y la pareja (Γ, P) satisface la condición de Markov.

El Teorema 2 indica que si se especifica la distribución de probabilidad condicional de cada nodo de un DAG, que ha sido dado, respecto sus padres y, luego se construye una distribución de probabilidad conjunta P sobre las variables del DAG mediante la **regla de la cadena**, entonces el par (Γ, P) es una RB. Esta es la forma en que las RBs se construyen en la práctica.

3.2. Inferencia bayesiana

El uso de una RB para calcular la probabilidad a posteriori de las variables no conocidas basándose en las variables conocidas se denomina inferencia bayesiana, propagación o actualización de creencias. Las inferencias involucran variables de consulta[6] (X) y variables de evidencia (E), y corresponde a calcular las probabilidades $P(X|E)$. Dicho de otra manera, a partir de una evidencia de la forma $E = \{X_{i1} = x_{i1}, ..., X_{it} = x_{it}\}$, donde $\{X_{i1}, ..., X_{it}\} \subset \mathbf{V}$ son las variables de *evidencia*, una inferencia consiste en calcular las probabilidades de la siguiente forma:

$$P(\ X_{j1} = x_{j1}, ..., X_{js} = x_{js} \mid E)$$

con $\{X_{j1}, ..., X_{js}\} \subset \mathbf{V} \setminus \{X_{i1}, ..., X_{it}\}$ las variables de consulta (*queries*). Las variables de la RB que no aparecen como variables de consulta ni evidencia, son tratadas como no observadas.

Dado que se conocen los valores de las variables de evidencia E, se quiere encontrar la predicción para la variable de consulta X, que es el valor de la variable X con mayor probabilidad a posteriori (estimación MAP[7]). Es decir, si $x_1, ..., x_r$ son los diferentes valores que puede tener X, entonces la predicción para X es:

$$x^* = arg\, max_{k=1,...,r}\, P(X = x_k|E)$$

Y se dice que $P(X = x^*|E)$ es el nivel de confianza (CL[8]) de la predicción.

3.3. Aprendizaje de los parámetros

Los parámetros de una RB, que son las PT y CPTs de cada nodo, pueden obtenerse a partir de los conocimientos y opinión de los expertos, a partir de los datos o una combinación de las dos opciones. Frecuentemente es difícil definir los parámetros cuando hay muchos nodos (variables) y conexiones.

[6] De la traducción del inglés *query*.
[7] Por sus siglas en inglés significa "Maximum a Posteriori".
[8] Por sus siglas en inglés significa "Confidence Level".

A partir de una estructura fija de una RB (DAG) es necesario estimar los parámetros para obtener el modelo completo. El objetivo es representar la distribución de probabilidad conjunta P, pero primero hay que especificar para cada nodo X_i la distribución de probabilidad condicional dado sus padres.

Para ello se usa el método de la estimación por máxima verosimilitud (MLE[9]) y así obtener los valores de los parámetros más verosímiles que hacen que los datos observados tengan mayor probabilidad de ocurrencia.

Una RB, como se ha comentado anteriormente, está formada por el par (Γ, P) donde $\Gamma = (\mathbf{V},\mathbf{E})$ es un DAG y $\mathbf{V} = \{X_1, ..., X_n\}$ es un conjunto de variables aleatorias discretas. El conjunto de datos $D = \{c_1, ..., c_M\}$ está formado por un total de M casos, es decir, cada caso (c_i) corresponde a una realización del vector aleatorio $(X_1, ..., X_n)$ en el caso de que los datos sean completos y una realización parcial del vector en el caso que los datos sean incompletos. La MLE define un vector de parámetros $\Theta = (\theta_1, ..., \theta_n)$ de la RB cuyos valores hacen que el conjunto de datos de D tenga mayor probabilidad de ocurrencia.

Los parámetros que se quieren estimar en la RB son definidos como:

$$\Theta_{x|u} = P(X = x \,|\, PA_X = u)$$

donde X es cualquier nodo de la RB ($X \in \mathbf{V}$) y $\mathrm{PA_X}$ es el conjunto de sus padres ($\mathrm{PA_X} \in \mathbf{V}$). El vector de todos los parámetros de la RB se define como Θ y la función de verosimilitud asociada al conjunto de datos D como $L(\Theta|D)$. Entonces, el MLE que maximiza la función de verosimilitud se defie como:

$$\widehat{\Theta} = \underset{\Theta}{arg\,max}\, L(\Theta|D) = \underset{\Theta}{arg\,max}\, \log(\, L(\Theta|D)\,)$$

Teorema 3 (Th. 17.1 Darwiche (2009)) *Si el conjunto de datos de entrenamiento D del modelo es completo (sin datos faltantes), entonces el MLE de $\widehat{\Theta}$ verifica que sus componentes se obtienen de la función de distribución empírica, es decir:*

$$\widehat{\Theta}_{x|u} = \#D(x,u) \,|\, \#D(u)$$

donde $\#D(x,u)$ es el número de casos de c_i en el conjunto D por los cuales $X = x$ y $PA_X = u$, y $\#D(u)$ el número de casos por los cuales $PA_X = u$.

A partir del Teorema 3 se sabe que $\widehat{\Theta}$ (que es el MLE de Θ) existe y es único cuando el conjunto de datos usado para ajustar el modelo es completo.

[9]Por sus siglas en inglés significa "Maximum Likelihood Estimation".

3.4. Aprendizaje de la estructura

La estructura de una RB, el DAG, puede obtenerse a partir de los conocimientos y opinión de los expertos, a partir de los datos o una combinación de ambas opciones, como pasa con los parámetros. Frecuentemente es difícil definir la estructura, por eso a menudo se opta por una opción híbrida. Entonces, se construye la estructura de la RB a partir de los datos, pero los expertos pueden aplicar restricciones.

En la Sección 3.3 se ha detallado cómo se aprenden los parámetros a partir de un DAG conocido, pero muchas veces es desconocido. Entonces, primero hay que aprender la estructura de la RB (DAG) y luego estimar los parámetros mediante el método de MLE, y así hacer inferencias a través de la RB completa.

El problema del aprendizaje de la estructura de una RB puede definirse de la siguiente manera:

Definición 2 (Eq. 2 Gómez et al. 2011) *Dado un conjunto de datos de entrenamiento* $D = \{c_1, ..., c_M\}$*, formado por M casos, donde cada caso* (c_i) *corresponde a una realización del vector aleatorio* $\boldsymbol{V} = (X_1, ..., X_n)$*, es decir* $c = (x_1, ..., x_n)$*, se quiere encontrar el DAG tal que:*

$$\Gamma^* = \underset{\Gamma \in \Gamma^n}{arg\,max}\ f(\Gamma : D)$$

donde $f(\Gamma : D)$ *es una función de puntuación o score que evalúa la calidad de cualquier DAG* Γ *candidato que hace referencia al conjunto de datos D, y* Γ^n *es el conjunto que contiene todos los DAGs con un total de n nodos.*

Hay que tener en cuenta que el aprendizaje de la estructura de una RB es un problema computacional complicado, y por ello resulta inviable encontrar una solución óptima. Para encontrarla sería necesario calcular todas las estructuras posibles que se pueden obtener de los datos, y escoger la que maximiza la función de puntuación o *score*. Esta función se basa en calcular cómo la estructura de la RB se adapta a los datos. Para resolver este problema se usan modelos heurísticos que maximizan la función *score* localmente, encontrando una solución local ótpima y útil para los propósitos de la investigación.

El algoritmo *Hill-Climbing*

El algoritmo *Hill-Climbing* es un método *scored-based* o *search-and-score* que asigna una función de puntuación o *score* a cada DAG con el objetivo de maximizarla para encontrar una solución local óptima.

El algoritmo realiza un bucle que se mueve continuamente en la dirección del aumento del valor, es decir, cuesta arriba. El algoritmo parte de una solución inicial (un DAG) y realiza el bucle un

número finito de pasos. En cada paso el alogritmo considera los DAGs vecinos, y elige el que tiene mayor valor en la función de puntuación o *score* $f(\Gamma : D)$. El algoritmo se detiene cuando no hay ningún cambio local del que resulta una mejora de f.

En el aprendizaje de la estructura de la RB se suele considerar como solución de partida la red vacía (sin arcos), aunque también se utilizan DAGs elegidos al azar. Sea cual sea el DAG inicial, tiene que cumplir la restricción de no tener ciclos dirigidos. En cuanto a las opciones habituales para los cambios locales en el espacio de los DAGs son: la adición de un arco, la eliminación de un arco o el cambio de sentido de un arco. Por supuesto, hay que tener cuidado de no introducir ciclos dirigidos en el gráfico durante el cambio (Gómez et al., 2011).

3.5. Función de puntuación o *score*

Antes de definir las funciones de puntuación más relevantes se especifica como se escriben distintos aspectos de las RBs.

Sea $D = \{c_1, ..., c_M\}$ un cojunto de datos de entrenamiento, formado por M casos, y $\mathbf{V} = (X_1, ..., X_n)$ un conjunto de variables aleatorias discretas, para cada DAG Γ sobre $\mathbf{V}$, se denomina:

- Θ^{Γ} el vector de parámetros asociado al DAG.
- Θ^{Γ}_{MLE} el vector de parámeros estimado por máxima verosimilitud (MLE). Por lo tanto, es el valor que más probablemente haya generado el conjunto de datos D.
- $L^D(\Theta^{\Gamma}_{MLE})$ el valor de la función de verosimilitud evaluada por Θ^{Γ}_{MLE} a partir del conjunto de datos D.
- d la dimensión de Γ, es decir, los parámetros no redundantes en Θ^{Γ}. Por lo tanto, es una medida de la complejidad de la RB con DAG Γ.

3.5.1. Score BIC

El criterio de información Bayesiano (BIC[10]) se usa para la selección de un modelo entre un conjunto finito de modelos y se prefiere el que tiene el BIC más alto. Cuando se ajusta la estructura de la RB es posible aumentar la verosimilitud agregando parámetros, pero hacerlo puede resultar un ajuste excesivo (*overfitting*) y predecir mal para los nuevos casos, aquellos que no se encuentran en el conjunto de entrenamiento.

[10]Por sus siglas en inglés significa "Bayesian Information Criterion"

El *score* BIC se define de la siguiente forma:

$$BIC\left(\Gamma, D\right) = \ln\left(L^D\left(\Theta^{\Gamma}_{MLE}\right)\right) - \frac{d}{2}\ln(N)$$

La primera parte de la fórmula se observa el logaritmo de función de verosimilitud. Esta medida muestra como de bien se ajusta el modelo (la RB) a los datos. La segunda parte de la fórmula, la resta, es una penalización por la complejidad del modelo.

3.5.2. Score AIC

El criterio de información Akaike (AIC[11]) se usa para la selección de un modelo entre un conjunto finito de modelos y se prefiere el que tiene el AIC más alto. El AIC maneja una compensación entre la bondad de ajuste y la complejidad del modelo. Es un *score* similar al BIC, pero penaliza menos por la complejidad del modelo y, por lo tanto, el resultado de la estructura de la RB (el DAG) suele ser más conectada y más compleja que la que se obtiene con el *score* BIC.

El *score* AIC se define de la siguiente forma:

$$AIC\left(\Gamma, D\right) = \ln\left(L^D\left(\Theta^{\Gamma}_{MLE}\right)\right) - d$$

3.6. Clasificadores bayesianos

La construcción de una RB a partir de los datos es útil para usarla como clasificador, un modelo que a partir de los valores de un conjunto de variables *input* es capaz de asignar una clase a la variable predicha u *output*.

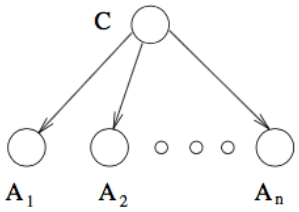
El clasificador bayesiano puede variar según la estructura del clasificador, del algoritmo de aprendizaje, de la función *score*, etc. A continuación se detallan dos clasificadores bayesianos: el *Naive Bayes* y el *Augmented Naive Bayes*.

3.6.1. Clasificador Naive Bayes

El clasificador Naive Bayes es uno de los más eficaces, en el sentido de que tiene buen rendimiento predictivo comparado con otros clasificadores más avanzados. Este clasificador aprende del conjunto de datos de entrenamiento T la probabilidad condicional a cada atributo X_i dada la clase C_k. La estructura del DAG de la RB está fijada anteriormente a la búsqueda de la distribución de probabilidad condicional, donde se obligan las flechas desde la clase C_k a todos los atributos y no existen relaciones entre los atributos, tal como se observa a continuación:

[11] Por sus siglas en inglés significa "Akaike Information Criterion"

Figura 2: Estructura de un clasificador Naive Bayes



Fuente: Friedman et al. (1997) The structure of the naive Bayes network. [Figura]. Recuperado de: `https://www.researchgate.net/publication/220343395_Bayesian_Network_Classifiers`.

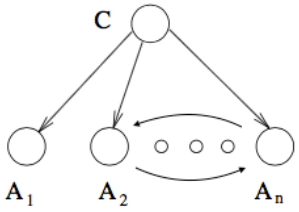
El clasificador Naive Bayes usa la Fórmula de Bayes para calcular la probabilidad de la variable predicha C a partir de las probabilidades $P(X_i|C)$ i $P(C)$ y eligiendo como clasificación la clase con mayor probabilidad (a posteriori).

La condición de Markov (ver Sección 3.1.2) implica asumir que todos los atributos son condicionalmente independientes entre ellos dados el valor de la variable clase C. Aunque parezca una suposición fuerte y poco realista, a la práctica este clasificador resulta útil y tiene buenos resultados.

3.6.2. Clasificador Augmented Naive Bayes

El clasificador Augmented Naive Bayes, como dice su nombre, se basa en la estructura del clasificador Naive Bayes. En este caso la variable predicha C también tiene que ser padre de todos los atributos, pero se permiten arcos (flechas) entre los atributos (aunque no son obligatorios). Es decir, a partir de los datos se modela la influencia entre los atributos. Por lo tanto, se fija la estructura del DAG anteriormente con la obligación de flechas desde la variable predicha C hacia los atributos y posibles flechas entre los atributos, elegidas por el algoritmo de aprendizaje. Con todo esto, se obtiene un DAG con una estructura como la siguiente:

Figura 3: Estructura de un clasificador Augmented Naive Bayes



Fuente: Elaboración propia a partir de Friedman et al. (1997) The structure of the naive Bayes network. [Figura]. Recuperado de: `https://www.researchgate.net/publication/220343395_Bayesian_Network_Classifiers`.

Cabe mencionar que la obligación de las relaciones de la variable clase C a los atributos se realiza mediante la opción **whitelist** de la función del paquete `bnlearn` de R que se usa para el aprendizaje de la estructura, y la prohibición de flechas de los atributos a la variable clase C con la opción **blacklist**. Visto que la estructura final tiene que ser un grafo dirigido acíclico (DAG), al definir la **whitelist**, indirectamente se prohíben las flechas desde los atributos hacia la variable clase C porque el resultado sería un grafo cíclico.

El aprendizaje de este clasificador tiene costos computacionales adicionales al clasificador Naive Bayes, ya que la estructura sí se ha de aprender.

3.7. Evaluación y validación de los clasificadores

3.7.1. *K-fold cross validation*

El *k-fold cross validation* es una técnica para evaluar la capacidad predictiva de un modelo, en este caso un clasificador de tipo red bayesiana. Esta técnica consiste en dividir aleatoriamente la muestra original D en k submuestras[12], con tamaño similares. Después el conjunto de datos formado por $k-1$ submuestras T, de tamaño $\frac{k-1}{k}$ aproximadamente, se usa para el entrenamiento del modelo. De aquí que T se llame conjunto de entrenamiento (en inglés *training data set*). El conjunto restante V, formado por una submuestra de tamaño $\frac{1}{k}$ aproximadamente, se usa para la validación del modelo y se le llama conjunto de validación (en inglés *validation data set*).

Como hay un total de k submuestras, siendo cada vez una submuestra diferente la que forme el conjunto de validación V, existen los conjuntos T_h y V_h con $h = 1, ..., k$. Para cada h se ajusta un modelo de clasificación a partir de los datos de entrenamiento T_h. Luego se lleva a cabo el proceso de validación, en el cual se usa el modelo ajustado para realizar las predicciones para nuevos casos (no usados en el entrenamiento y vistos por primera vez para el modelo), que son los que pertenecen al conjunto de validación V_h. Realizadas las predicciones para cada caso, dado que se conoce el valor observado (el real), se puede construir una matriz de confusión y calcular las métricas de rendimiento del modelo para evaluar su capacidad predictiva.

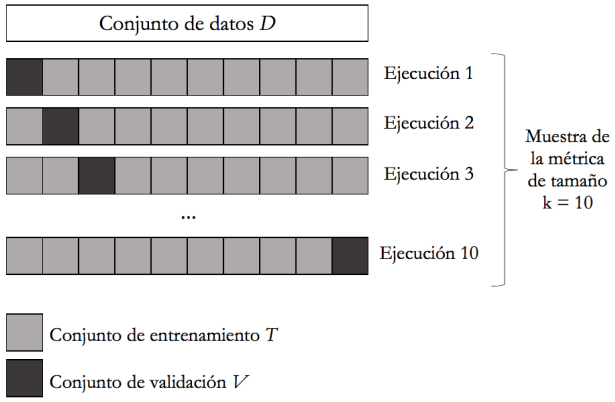
Dado que se ajusta un modelo para cada h, por cada métrica de rendimiento se obtiene una muestra de tamaño k. Este hecho es útil para posteriormente calcular la media, mediana, desviación estándar y/o realizar pruebas de hipótesis para cada métrica.

En el estudio presente se define $k = 10$, por lo que cada conjunto T_h representa el 90 % de la muestra original y cada conjunto V_h un 10 %, aproximadamente. Una vez ajustados todos los modelos posibles,

[12] k es una constante que normalmente toma valor 3, 5 o 10.

un total de 10, se hace una media de los resultados para encontrar una única estimación de la métrica de interés.

Figura 4: Ejemplificación del método *k-fold cros validation* con $k = 10$



Fuente: Elaboración propia.

3.7.2. Métricas de comportamiento

Existen múltiples métricas de rendimiento para evaluar la capacidad predictiva de un modelo. Aquí se explican las más usuales y las que se emplearán más adelante.

Para los problemas de clasificación se usa la matriz de confusión, una tabla donde las clasificaciones observadas se representan en las columnas y las predichas en las filas. La matriz de confusión se denomina $A = (a_{ij})_{i,j=1,\dots,r}$, donde r es el número de categorías de la variable clase y $\sum_{i=1}^{r} \sum_{j=1}^{r} a_{ij} = N$.

Cuadro 1: Matriz de confusión general

$A =$

Predicción \ Observado	C_1	C_2	$\cdots$	C_r
C_1	a_{11}	a_{12}	$\cdots$	a_{1r}
C_2	a_{21}	a_{21}	$\cdots$	a_{2r}
$\dots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
C_r	a_{r1}	a_{r2}	$\cdots$	a_{rr}

Fuente: Elaboración propia.

A partir de la matriz de confusión se pueden obtener las siguentes métricas:

i) *Accuracy*: es la proporción de predicciones correctas que ha hecho el modelo de todas las posibles. Esta métrica se puede usar para variables binarias, categóricas y ordinales.

$$\text{Accuracy} = \frac{\sum_{i=1} a_{ii}}{N}$$

ii) *Matthews Correlation Coefficient (MCC)*: es una extensión para el caso de múltiples categorías de lo que se conoce como el coeficiente ϕ para el caso binario, es decir, la raíz cuadrada de la media del estadístico χ^2 sobre el número de casos observados para una matriz de confusión 2x2 de la clasificación binaria. La fórmula MCC es la siguiente:

$$MCC^{13} = \frac{\sum_{k,l,m=1}^{r} \left(a_{kk}\, a_{lm} - a_{mk}\, a_{kl}\right)}{\sqrt{\sum_{k=1}^{r} \left(\left(\sum_{l=1}^{r} a_{kl}\right) \left(\sum_{u,v=1,u\neq k}^{r} a_{uv}\right)\right)} \sqrt{\sum_{k=1}^{r} \left(\left(\sum_{l=1}^{r} a_{lk}\right) \left(\sum_{u,v=1,u\neq k}^{r} a_{vu}\right)\right)}}$$

El valor de MCC se encuentra entre el rango $[-1, 1]$, donde +1 indica una clasificación perfecta. Cuando mayor sea el valor de MCC, mejor será el rendimiento del clasificador.

iii) *Mean Absolut Error (MAE)*: es la media de la diferencia absoluta entre la observación (y_j) y la predicción ($\hat{y}_j$), y penaliza más para las preidcciones que se equivocan más.

$$MAE = \frac{\sum_{j=1}^{n} |y_j - \hat{y}_j|}{n}$$

El MAE solo tiene sentido para variables ordinales. Por ejemplo, si la variable tiene las clases baja, media y alta, el error de predecir baja cuando es media se contabiliza como "1" y el error de predecir baja cuando es alta se contabiliza como "2".

Para las variables binarias se obtiene una matriz de confusión 2x2, y se pueden calcular las métricas habituales de rendimiento para un problema de clasificación. Para las variables categóricas también se puede construir una matriz de confusión 2x2 para cada clase mediante la unión de las otras (matriz de confusión "colapsada"), y obtener las mismas métricas que en una variable binaria, mediante la media que se ha obtenido con cada clase. A continuación se muestra un ejemplo de matriz de confusión 2x2, con las categorías más (+) y menos (-):

Cuadro 2: Matriz de confusión 2x2

Predicción \ Observado	+	-
+	TP	FP
-	FN	TN

Fuente: Elaboración propia.

donde:

- TP = True Positive. Número de casos donde la predicción y la clase real coinciden y son (+).
- FP = False Positive. Número de casos donde la predicción es (+) y el valor real observado es (-).

[13] Fórmula del artículo Delgado and Tibau (2019).

- FP = False Negative. Número de casos donde la predicción es (-) y el valor real observado es (+).
- TN = True Negative. Número de casos donde la predicción y la clase real coinciden y son (-).

A partir de la matriz de confusión 2x2 se pueden obtener las siguentes métricas:

iv) *Precision*: mide la proporción de predicciones correctas de la clase + de entre todos los casos predichos con dicha clase.

$$\text{Precision} = \frac{TP}{TP + FP}$$

iv) *Recall*: mide la proporción de predicciones correctas de la clase + de entre todos los casos reales con dicha clase.

$$\text{Recall} = \frac{TP}{TP + FN}$$

iv) *F-score*: es una medida que combina la *precision* y el *recall* en un número (se usa la media armónica, como es habitual, porque tanto las métricas *precision* como *recall* son proporciones entre 0 y 1, y luego se pueden interpretar como tasas). Su definición es:

$$\text{F} - \text{score} = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2\,TP}{(2\,TP) + FP + FN}$$

4. Base de datos

Para construir el modelo probabilístico se emplea una base de datos de la Secretaría de Estado de Seguridad del Ministerio del Interior del Gobierno de España, que contiene información de homicidios esclarecidos cometidos en España entre los años 2010 y 2012, ambos inclusive.

Entre los años 2010 y 2012 en España se registraron 1150 homicidios, y se pudieron reunir los atestados de 682 casos (59.3 %), ocurridos en las demarcaciones de la Guardia Civil y la Policía Nacional. De estos, únicamente se seleccionaron los casos esclarecidos policialmente, entendiendo que son aquellos para los que se llegó a conocer la identidad del autor o autores, incluyéndose los casos en que siendo varios autores solamente se conoció la identidad de alguno de ellos, dejando un total de 632 casos.

Hay que tener en cuenta que la base de datos no contiene homicidios de las siguientes comunidades autónomas: Cataluña, País Vasco y Extremadura. Por un lado, la policía autonómica de Cataluña y País Vasco consideraron que su participación no afectaba a la representatividad de la muestra. Por otro lado, en los atestados de Extremadura hubo problemas técnicos (González et al., 2018).

En resumen, el siguiente estudio se realiza sobre un total de 632 homicidios, cometidos por 871 autores que dejaron un total de 662 víctimas desde el 2010 hasta el 2012. Es decir, en algunos casos

hay más de un autor y/o más de una víctima (para más información consulte la Sección 6.1). En el Cuadro 3 se resume esta información por años.

Cuadro 3: Muestra del estudio por años

Año	Homicidios en España	Atestados recopilados	Hechos esclarecidos	Autores	Víctimas
2010	401	233	216	272	224
2011	485	235	220	334	233
2012	364	214	196	265	205
Total	1150	682	632	871	662

Fuente: Tabla 1.1. de González et al., 2018.

Una vez obtenidos los atestados policiales, un equipo universitario se encargó de leer y vaciar la información que contenían, creando una base de datos con 255 variables. Estas contienen información sobre las características del hecho (año, día de la semana, número de autores, número de víctimas, lugar, etc.), características de la víctima (edad, sexo, situación laboral, etc.) y características del autor (edad, sexo, situación laboral, etc.).

A partir de todas las variables se ha hecho una selección de las más útiles y relevantes para el estudio del homicidio con mayor atención a la relación con las drogas. Estas variables corresponden a características del hecho ($H_1, ..., H_{14}$), de la víctima ($V_1, ..., V_9$), del autor ($A_1, ..., A_{10}$) y la relación víctima - autor (R). Destacar que todas ellas son categóricas, particularidad necesaria para realizar la red bayesiana. En el Cuadro 4 se resumen las variables utilizadas y en el Anexo B se especifican las definiciones de las variables y sus categorías.

Cuadro 4: Resumen de las variables del modelo y sus categorías

Variable	Categorías
H_1 = año del hecho	2010/2011/2012
H_2 = estación del año	primavera/verano/otoño/invierno
H_3 = día de la semana	lunes/martes/miércoles/jueves/viernes/sábado/domingo
H_4 = franja horaria	madrugada/mañana/tarde/noche
H_5 = núm. víctimas	una/varias
H_6 = núm. autores	uno/varios
H_7 = testigos	sí/no
H_8 = tipología del hecho	discusión/violencia género/violencia familiar/ otras interpersonales/robo/actividades criminales
H_9 = arma	objeto contundente/blanca/fuego/fuerza agresor/asfixia/otros

continuación Cuadro 4: Resumen de las variables del modelo y sus categorías

Variable	Categorías
H_{10} = desplazamiento del arma	sí/no/cuerpo del agresor
H_{11} = tipo de lugar	interior/exterior/vehículo/agua
H_{12} = escenas	única/multiescena
H_{13} = movimiento del cuerpo	desplazado y no oculto/oculto y no desplazado/ desplazado y oculto/ni desplazado ni oculto
H_{14} = método de huida	a pie/escena/vehículo/suicidio
V_1 = edad víctima	0/menores/18-20/21-30/31-40/41-50/51-64/65 o más
V_2 = sexo víctima	hombre/mujer
V_3 = origen víctima	España/resto de Europa/resto del mundo
V_4 = situación laboral víctima	ocupado/parado/estudiante/jubilado/otra
V_5 = antecedentes víctima	homicidio/contra las personas/sí pero desconocidos/no
V_6 = adicciones víctima	alcohol/drogas/alcohol y drogas/no
V_7 = frecuencia de consumo víctima	ocasional/habitual/no/frecuencia consumo desconocida
V_8 = sustancia consumida por la víctima durante el hecho	alcohol/drogas/alcohol y drogas/sí pero desconocida/no
V_9 = actividades ilegales víctima	sí/no
A_1 = edad autor	menores/18-20/21-30/31-40/41-50/51 o más
A_2 = sexo autor	hombre/mujer
A_3 = origen autor	España/resto de Europa/resto del mundo
A_4 = situación laboral autor	ocupado/parado/estudiante/jubilado/otra
A_5 = antecedentes autor	homicidio/contra las personas/sí pero desconocidos/no
A_6 = adicciones autor	alcohol/drogas/alcohol y drogas/no
A_7 = frecuencia de consumo autor	ocasional/habitual/no/frecuencia consumo desconocida
A_8 = sustancia consumida por el autor durante el hecho	alcohol/drogas/alcohol y drogas/sí pero desconocida/no
A_9 = trastorno mental autor	sí/no consta
A_{10} = actividades ilegales autor	sí/no

continuación Cuadro 4: Resumen de las variables del modelo y sus categorías

Variable	Categorías
R = relación víctima - autor	pareja o expareja/conocido o vecino/amistad o familia/otra/no

Fuente: Elaboración propia.

4.1. Pre-processing

Para obtener las variables y categorías presentadas en el Cuadro 4, se ha explorado profundamente la base de datos y elegido las variables más relevantes.

4.1.1. Agrupación y discretización de las categorías de las variables

Inicialmente se ha visto una incongruencia entre la clasificación de un homicidio como doble (más de una víctima y más de un autor) cuando en realidad solo había una víctima.

La variable tipología del hecho (H_8) se ha categorizado a partir de una combinación de la clasificación de la ONUDD y del Ministerio del Interior de España. La ONUDD clasifica el homicidio de tres maneras: interpersonal, actividades criminales y sociopolítico. El Ministerio del Interior cataloga el homicidio como: discusión/reyerta, violencia de género, violencia familiar/doméstica, otros motivos interpersonales, robo, otras actividades criminales, criminalidad organizada, prostitución y bandas. A partir de aquí, se ha establecido las siguientes categorías: discusión/reyerta, violencia de género, violencia doméstica/familiar, otras interpersonales, robo y actividades criminales (que agrupa los hechos de grupos criminales, prostitución, bandas y otras actividades criminales).

La variable desplazamiento del arma (H_{10}) inicialmente tenía dos categorías: sí y no. Dado que existe una variable que informa que a veces el autor usa medios asfixiantes para la perpetración del homicidio, se ha añadido la categoría "cuerpo del agresor" para indicar que el arma no puede ser desplazada porque el autor ha usado su propio cuerpo.

Para la variable movimiento del cuerpo (H_{13}) se han usado dos variables dicotómicas que indicaban si el cuerpo había sido ocultado y si había sido desplazado, dando como resultado cuatro categorías.

Las franjas de edad son diferentes para la víctima y el autor. En la edad de la víctima se ha conservado la edad de 0 años, ya que indica que la víctima era un recién nacido, y luego existe la categoría "menores" que tiene en cuenta la franja de 1 año hasta los 17 años. En cambio, para el autor no hay una categoría explícita para los de 0 años, ya que no hay ningún caso con dicha característica.

La variable origen víctima (V_3) y origen autor (A_3) se ha creado a partir de la información del

país de origen y el continente de ellos. Con el objetivo de obtener categorías más equitativas se han creado tres categorías: España, resto de Europa y resto del mundo.

La variable referente a los antecedentes (V_5 y A_5) se ha categorizado a partir de las tres variables dicotómicas (sí/no) sobre los antecedentes por homicidio, antecedentes contra las personas y antecedentes. Como resultado se han creado cuatro categorías: antecedentes por homicidio, antecedentes contra las personas, antecedentes pero se desconoce de qué tipo y que no tiene antecedentes.

Respecto a la variable de la frecuencia de consumo (V_7 y A_7), se ha creado la categoría "consume pero se desconoce", que tiene en cuenta las víctimas y autores que se sabe que consumen alcohol y/o drogas pero no su frecuencia de consumo.

Referente al consumo de sustancias momentos antes del homicidio (V_8 y A_8), se ha tenido en cuenta dos variables. Una informaba si habían consumido o no, y la otra qué habían consumido. Con esto se han creado cinco categorías finales: consumió alcohol, consumió drogas, consumió ambas, consumió pero no se sabe qué y no consumió.

Finalmente, para la variable sobre la relación entre la víctima y el autor (R) se han definido como "pareja o ex" las relaciones entre víctima y autor de tipo pareja, cónyuge, expareja y separado/divorciado. La relación de amistad y familiar se han unido para crar una única categoría: "amistad/familiar". Y, para las categorías iniciales de relación laboral/comercial, escolar, conocido (pero sin especificar de qué), se ha creado la categoría "otra".

4.1.2. Tratamiento de los datos faltantes (NA)

El paquete `bnlearn` del software `R` que se usa en este trabajo para ajustar las redes bayesianas no permite trabajar con valores faltantes (NA) en la base de datos y, por tanto, se ha hecho un tratamiento de ellos.

Antes al tratamiento, se ha realizado el análisis descriptivo y el análisis de valores faltantes (NA). Después, se ha generado una categoría artificial en todas las variables con NA, sustituyendo NA por la categoría "Desc", indicando que se desconoce la categoría real. A partir de esta base de datos se han entrenado las RBs.

Posteriormente al entrenamiento de los modelos se realizan las predicciones. Para aquellas variables con la categoría artificial "Desc" se ha hecho un cambio de escala en la distribución de probabilidad asociada a la variable que se quiere predecir condicionada a la evidencia (variables *input*). Por ejemplo, si hay una variable con las categorías "A", "B", "C" y la categoría artificial "Desc" y a partir de la evidencia se encuentra la siguiente distribución de probabilidad: "A" (0.3), "B" (0.2), "C" (0.4)

y “Desc” (0.1), entonces se divide cada probabilidad obtenida entre la suma de las probabilidades de las categorías no artificiales (que sean diferente a “Desc”). En este ejemplo se obtiene la siguiente distribución de probabilidad: “A” (3/9), “B” (2/9), “C” (4/9). Finalmente, se elige como predicción la categoría con probabilidad superior, que en este caso sería la categoría “C”.

Una vez hechas las predicciones hay que calcular las métricas de comportamiento a partir de la matriz de confusión. Hay que tener en cuenta que la categoría “Desc” significa que no se conoce el valor real, y nunca se sabrá si la predicción es correcta o no. Es por ese motivo que no se usa para la validación del modelo. Por lo tanto, si la columna de la matriz de confusión representa las clasificaciones observadas, se quita dicha columna y se realiza el cálculo de la métrica pertinente.

5. Análisis de los valores faltantes

Se ha comentado que la base de datos se ha creado a partir de los atestados policiales, documentos escritos por la policía que lleva el caso del homicidio, y vaciados por un grupo de universitarios. Esta particularidad, junto con otros inconvenientes, afecta en la información recogida, habiendo variables de algunos hechos donde no se conoce su valor. A continuación se realiza un análisis más exhaustivo para ver qué variables tienen un porcentaje mayor de valores faltantes para tenerlo en cuenta en la realización del perfil predictivo.

5.1. Valores faltantes en los hechos

En general, las variables referentes a los hechos no destacan por tener un porcentaje elevado de valores faltantes, ya que todos los casos son esclarecidos. En el Cuadro 5 se puede ver que el método de huida del autor de la escena es la variable con mayor desconocimiento, en un 35.8 % de los hechos, seguido por el desplazamiento del arma, desconocido en un 22.8 % de los casos. También se observa que en un 12.8 % se desconoce la franja horaria del hecho porque a veces no se puede dar una hora exacta del homicidio, pero el día se conoce en todos los casos. Finalmente, decir que en el 3.6 % de los casos recogidos no se especifica la tipología del hecho.

Cuadro 5: Valores faltantes en las variables de los hechos

Variable	n (N = 632)	%
H_{14} = método de huida	226	35.8
H_{10} = desplazamiento del arma	144	22.8
H_4 = franja horaria	81	12.8
H_9 = arma	29	4.6

continuación Cuadro 5: valores faltantes en las variables de los hechos

Variable	n (N = 632)	%
H_8 = tipología del hecho	23	3.6
H_{13} = movimiento del cuerpo	18	2.8
H_{11} = tipo de lugar	2	0.3
H_{12} = escenas	2	0.3
H_1 = año	0	0.0
H_2 = estación del año	0	0.0
H_3 = día de la semana	0	0.0
H_5 = número de vícitimas mortales	0	0.0
H_6 = número de autores	0	0.0
H_7 = testigos	0	0.0

Fuente: Elaboración propia.

5.2. Valores faltantes en las víctimas

En el Cuadro 6 se resumen los valores faltantes de las variables de las víctimas. Las características que menos se conocen son aquellas relacionadas con las drogas. Aproximadamente en el 80 % de las víctimas no se conoce si eran adictos a alguna sustancia, ni el patrón de su consumo ni si había consumido momentos previos a su homicidio.

Después del consumo de sustancias, en la mayoría de víctimas no se sabe si tenía antecedentes penales, exactamente en 531 (80.2 %) de ellas. La situación laboral no se conoce en un 61.6 % de las víctimas, esto se explica porque es una variable que no se suele contemplar en los atestados policiales (González et al., 2018).

En el caso del sexo de la víctima, únicamente se desconoce en un caso, ya que era un bebé y no se detalló el sexo en el atestado policial.

Cuadro 6: Valores faltantes en las variables de la víctima

Variable	n (N = 662)	%
V_9 = sustancia en el hecho	537	81.1
V_6 = adicciones	535	80.8
V_7 = frecuencia de consumo	535	80.8
V_5 = antecedentes	531	80.2
V_4 = situación laboral	408	61.6

continuación Cuadro 6: valores faltantes en las variables de la víctima

Variable	n (N = 662)	%
V_1 = franjas de edad	31	4.7
V_3 = origen	17	2.6
V_2 = sexo	1	0.2
V_9 = actividades ilegales o no registradas	0	0.0

Fuente: Elaboración propia.

5.3. Valores faltantes en los autores

En el Cuadro 7 se resume los valores faltantes de las variables de los autores. Respecto al consumo de sustancias, de nuevo, hay un número elevado de valores faltantes, exactamente en 648 (74.4 %) autores se desconoce su adicción y su frecuencia de consumo. En el momento de perpetrar el homicidio se desconoce en 670 (76.9 %) autores si estaban bajo los efectos de las drogas.

Destacar que en el 62.3 % de los autores no se conoce su situación laboral y en un 22 % no se sabe si tenía antecedentes.

Cuadro 7: Valores faltantes en las variables del autor

Variable	n (N = 871)	%
A_8 = sustancia en el hecho	670	76.9
A_6 = adicciones	648	74.4
A_7 = frecuencia de consumo	648	74.4
A_4 = situación laboral	543	62.3
A_5 = antecedentes	192	22.0
A_3 = origen	16	1.8
A_1 = franjas edad	12	1.4
A_2 = sexo	0	0.0
A_9 = trastorno mental	0	0.0
A_{10} = actividades ilegales o no registradas	0	0.0

Fuente: Elaboración propia.

6. Análisis descriptivo

Anteriormente se ha hecho un análisis detallado de los porcentajes de los valores faltantes, es por eso que para la descriptiva se han calculado los porcentajes a partir de los valores conocidos (válidos) de las variables. En Anexo C se muestran las tablas de frecuencia para todas las variables.

6.1. El hecho

Los homicidios analizados (N = 632) se distribuyeron anualmente de la siguiente forma: 216 (34.2 %) tuvieron lugar en el año 2010, 220 (34.8 %) en el 2011 y 196 (31 %) en 2012. Agrupándolos por estaciones del año, en verano se produjeron 180 casos (28.5 %), 164 (25.9 %) en primavera, 150 (23.7 %) en invierno y 138 (21.9 %) en otoño. El día de la semana con un porcentaje mayor de homicidios es el domingo, con un 16.3 %, y el menor lunes, con un 11.6 %. Respecto a la franja horaria de comisión delictiva la noche (32.8 %) y la madrugada (27 %) tienen porcentajes mayores.

En el 95.6 % de los homicidios hay una víctima, siendo poco frecuentes aquellos hechos donde hay más de una víctima, el 4.4 % restante. En el 79.1 % de los hechos únicamente hay un autor. Cuando se analiza el número de víctimas y autores que han intervenido en un hecho, se observa que en 477 (75.5 %) casos hay una víctima y un autor. Los casos donde hay más de una víctima y más de un autor son poco frecuentes, sumando un total de 5 casos en la base de datos.

Cuadro 8: Número de casos (y %) según el nº de autores y víctimas

	Un autor	Más de un autor
Una víctima	477 (75.5)	127 (20.1)
Más de una víctima	23 (3.6)	5 (0.8)

Fuente: Elaboración propia.

Respecto a la tipología de homicidios, la mayoría de los casos estudiados son en contextos de discusión/reyerta (22.7 %), seguidos por los homicidios de violencia de género y violencia doméstica o familiar, con un 21.3 % ambos. Los homicidios relacionados con la comisión de la actividad criminal o la comisión de un robo representan porcentajes menores, un 10.8 % y un 8 % respectivamente.

Las armas blancas son las mayormente empleadas en los homicidios, en un 47.4 %, seguidas por las de fuego con un 15.6 %. Los autores hacen menor uso de medios asfixiantes para acabar con la vida de la víctima, un 5.3 % de los hechos. Sobre el desplazamiento del arma, en un 48 % de los homicidios el agresor no desplaza el arma de la escena del crimen respecto de un 36.7 % de los homicidios en los que sí que se desplaza.

En relación con el tipo de lugar donde se cometen los homicidios, la mayoría de los casos se producen en escenas interiores (62.9 %), seguidas de escenas exteriores (33.2 %), siendo el 3.3 % del total de los casos los que se producen en el interior de vehículos y el 0.6 % en localizaciones acuáticas. Además, el hecho se realiza en una única escena el 89 % y el resto en más de una escena.

En la mayoría de los hechos (59 %) existen testigos principales que puedan acreditar la comisión del hecho delictivo y que puede aportar datos sobre el/los autor/es.

En lo que se refiere al método de huida, el 40.9 % de los autores lo hizo a pie, el 32 % fue detenido en el lugar de los hechos y el 3.3 % utilizó un vehículo.

Atendiendo a algunos comportamientos más concretos del agresor posterior el homicidio, se ha encontrado que en un 88.9 % el agresor ni desplazó ni oculto el cadáver, en un 4.7 % solo lo desplazó, en un 3.9 % lo desplazó y ocultó y, finalmente, en un 2.3 % únicamente lo ocultó.

6.2. La víctima

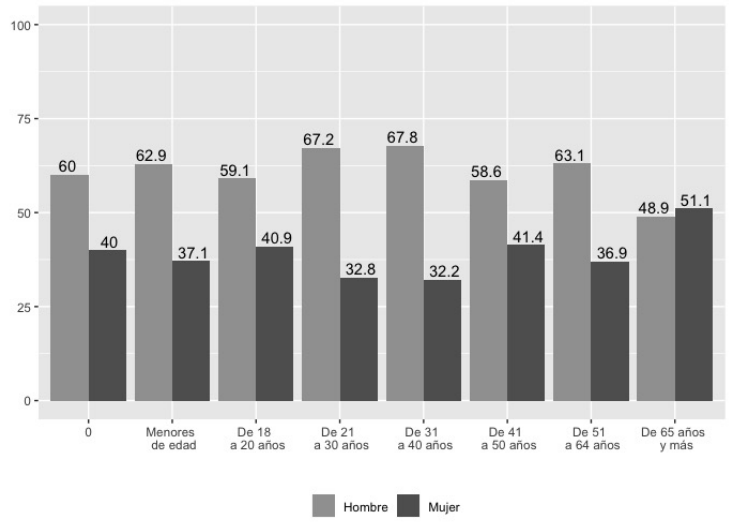
El número de víctimas totales es de 662, de las cuales 406 son hombres (61.4 %) y 255 mujeres (38.6 %). La edad media de las víctimas es de 42 años, pero si se distingue por sexos los hombres tienen una media de 40 años y las mujeres de 44. En general, el grupo de edad más frecuente de las víctimas es el de 21-30 años con 122 víctimas (19.3 %), seguido de los grupos de 31-40 años con 118 (18.7 %) y el de 41-50 años con 116 (18.4 %). En la Figura 5 se muestra la distribución de las víctimas por franjas de edad y sexo, y se observa que en todas las franjas hay un porcentaje mayor de hombres víctimas, excepto en el grupo de 65 años y más, donde las mujeres tienen 2 puntos porcentuales más que los hombres.

En la Figura 6 se segrega el sexo de las víctimas según la tipología del hecho, y se observa que para casi todas las categorías la víctima es mayormente un hombre. De las víctimas que murieron por un contexto de discusión/reyerta, el 92.3 % eran hombres. Destacar que las víctimas de los homicidios de violencia de género son todas mujeres, ya que se define como tal los hechos donde el autor (hombre) mata a la víctima (mujer) con la que mantienen o mantenía una relación de afectividad o análoga (ver definición en el Anexo B).

En cuanto a la nacionalidad, las víctimas son principalmente españolas, exactamente 464 (71.9 %), 66 víctimas (10.2 %) son de otro país de Europa y 115 (10.2 %) de un país no europeo. Destacar que existe un porcentaje mayor de víctimas extranjeras en los hechos relacionados con actividades criminales comparado con las otras tipologías (Fig. 7).

Figura 5: Edad de las víctimas por franjas según su sexo

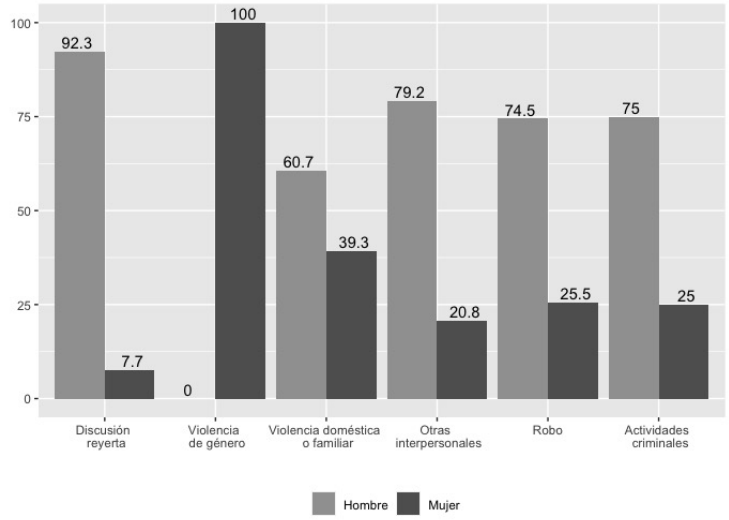
Unidades: porcentaje



Fuente: Elaboración propia.

Figura 6: Tipología del hecho según sexo de la víctima

Unidades: porcentaje



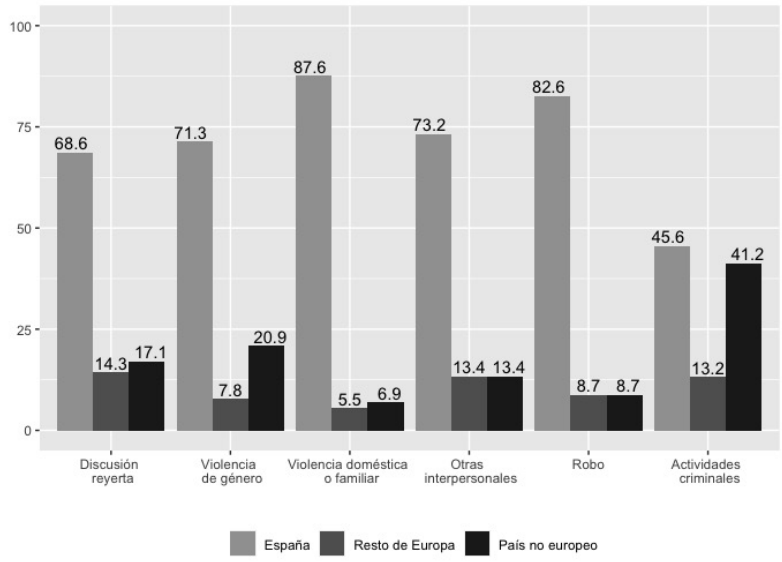
Fuente: Elaboración propia.

En relación con la situación laboral, casi la mitad de las víctimas (47.6 %) tienen trabajo en el momento de su homicidio y un 25.6 % de las víctimas tienen otra situación laboral, es decir, realizaban alguna actividad no registrada o ilegal que les servía de sustento.

Respecto a las adicciones, 48 (37.8 %) víctimas son adictas al alcohol, 39 (30.7 %) a las drogas y 25 (19.7 %) de ellas a ambas cosas. De las víctimas que se conoce su adicción, 78 (61.4 %) consumen habitualmente. En el momento del homicidio 80 (64 %) víctimas habían consumido alcohol, y 17 (13.6 %) drogas. Hay que tener en cuenta que en la mayoría de las víctimas (80.8 %) no se conoce la adicción a las drogas ni su frecuencia de consumo.

Figura 7: Tipología del hecho según la nacionalidad de la víctima

Unidades: porcentaje



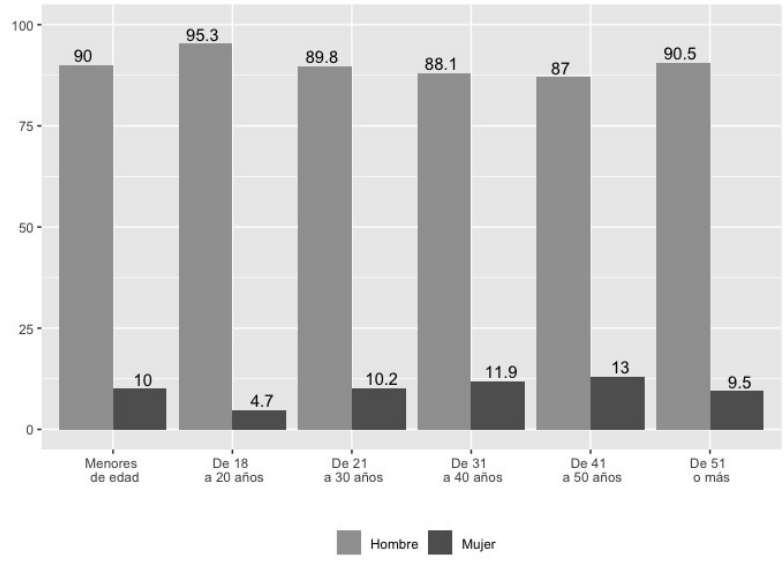
Fuente: Elaboración propia.

6.3. El autor

El número de autores totales es de 871, de los cuales 778 son hombres (89.3 %) y 93 mujeres (10.7 %). La media de edad de los autores es de 36.4 años, pero si se distingue por sexos los hombres tienen una media de 36 años y las mujeres de 38. En general, el grupo de edad del autor más frecuente es el de 31-40 años con un 27 %, seguido por el de 21-30 años con un 26.2 %. Hay 40 (4.7 %) autores que son menores y 126 (14.7 %) que tienen 51 años o más. En la Figura 8 se segrega el sexo de los autores según la franja de edad y se observa que en todas las franjas el sexo que predomina son los hombres. En el caso de las mujeres, se observa mayor porcentaje de mujeres en la franja de edad de los 41 a 50 años, con un 13 %.

Figura 8: Edad de los autores por franjas según su sexo

Unidades: porcentaje



Fuente: Elaboración propia.

En la Figura 9 se segrega el sexo de los autores según la tipología del hecho, y se observa que para todas las categorías el autor es mayormente hombre. Destacar que el porcentaje de mujeres autoras es mayor para los hechos clasificados como violencia doméstica o familiar, ya que recoge los casos en los que la mujer mata a la pareja. Cuando el hombre es el que mata a la pareja (mujer) se clasifica como violencia de género, y es por eso que en esta tipología el 100 % de los autores son hombres.

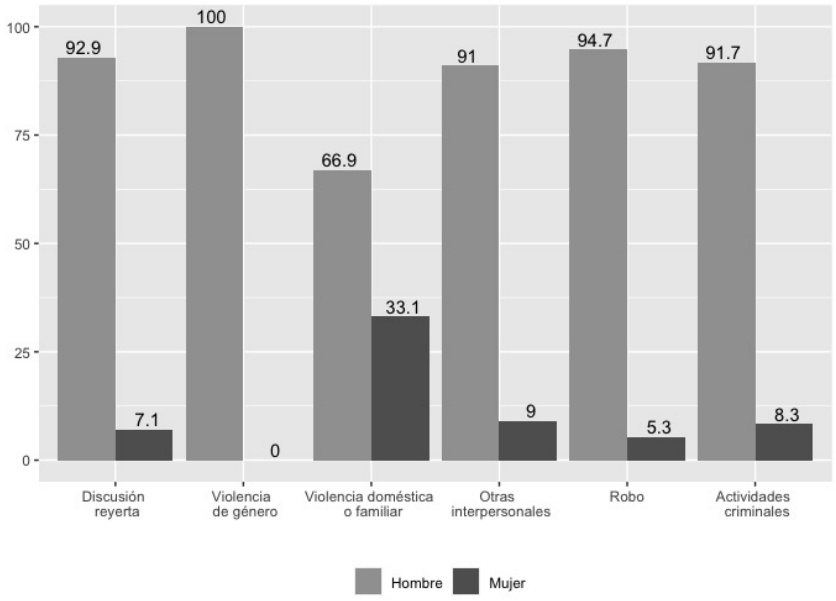
En cuanto a la nacionalidad, los autores son principalmente españoles, exactamente 553 (64.7 %), 131 (15.3 %) autores son de otro país de Europa y 171 (20 %) de un país no europeo. Destacar que existe una presencia mayor de autores de fuera de España (51.8 %) en los hechos clasificados como actividades criminales (ver Fig. 10).

En relación con la situación laboral, la mayoría de autores tienen trabajo en el momento de cometer el homicidio y un 33.3 % de los autores tienen otra situación laboral, es decir, realizaban alguna actividad no registrada o ilegal que les servía de sustento. Casi una cuarta parte de los autores (23.2 %) carecía de empleo remunerado en el momento de los hechos.

Por lo que hace a la existencia de antecedentes penales y policiales del autor, la mayoría de ellos (37.6 %) tiene antecedentes contra las personas y un 24 % tiene antecedentes, pero se desconoce de qué tipo. Aun así, un 24.4 % de los autores no tiene antecedentes.

Figura 9: Tipología del hecho según el sexo del autor

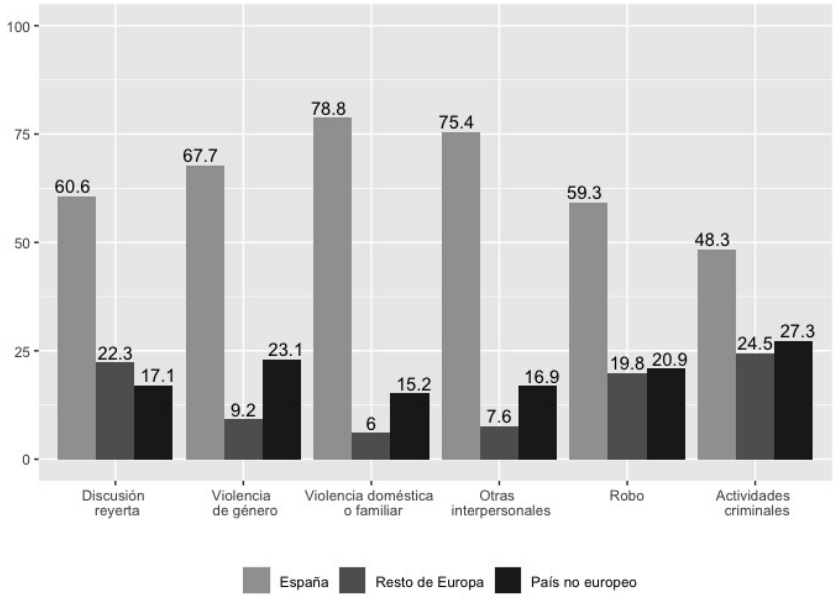
Unidades: porcentaje



Fuente: Elaboración propia.

Figura 10: Tipología del hecho según la nacionalidad de la víctima

Unidades: porcentaje



Fuente: Elaboración propia.

Respecto a las adicciones, 90 (40.4 %) autores son adictos al alcohol, 55 (24.7 %) a las drogas y 52 (23.3 %) a ambas cosas. De los que se conoce su adicción, 130 (58.3 %) consumen habitualmente. En el momento del homicidio 111 (55.2 %) autores habían consumido alcohol, 32 (16 %) alcohol y drogas y 24 (11.9 %) solo droga. Hay que tener en cuenta que en la mayoría de autores (74.4 %) no se conoce su adicción ni su frecuencia de consumo. Por último, mencionar que en un 9.8 % de los autores se conoce si tienen trastorno mental y en el 90.2 % restante no consta tal información.

6.4. Relación víctima y autor

Anteriormente se ha visto que hay un total de 871 autores y 662 víctimas (Cuadro 3), esto genera una suma de 909 relaciones diferentes entre víctima y autor, ya que hay casos donde hay más de un autor y/o más de una víctima. A continuación se resume el tipo de relación, pero esta no se conoce en 61 casos, es decir, un 6.7 % del total. Por lo tanto, solamente se conoce la relación entre víctima y autor en 848 relaciones de las 909 totales.

En el Cuadro 9 vemos que la mayoría de autores y víctimas eran conocidos y/o vecinos, en un 26.4 %. Seguidamente, en un 21.3 %, el autor y la víctima tenían una relación de amistad y/o familiar. En un 16.9 % la víctima y el autor eran pareja o expareja y en un 19.3 % la víctima y el autor no se conocían. Finalmente, en un 9.4 % de las relaciones son de otro tipo de las mencionadas anteriormente.

Cuadro 9: Análisis de la relación entre víctima y autor

Relación	n	%
Conocidos y/o vecinos	240	26.4
Amistad y/o familiar	194	21.3
Pareja o expareja	154	16.9
Otra	85	9.4
No se conocen	175	19.3
Se desconoce la relación	61	6.7
Total	909	100

Fuente: Elaboración propia.

7. Clasificación *Multi-Instance* en redes bayesianas

Tal como se ha explicado, hay homicidios perpetrados por un único autor o por varios autores, con una víctima o más de una (ver Cuadro 8). Entonces, habrá un conjunto de víctimas que tendrán el mismo autor y, por tanto, los valores de las variables del autor coincidirán, siendo a la vez las que se quieren predecir.

Primero hay que identificar el conjunto de víctimas que tienen el mismo autor y tener en cuenta que todo el conjunto de víctimas solo puede predecir un valor para cada variable del autor. Este tipo de clasificación se conoce como clasificación de **casos múltiples** (MI[14]).

En la clasificación MI cada objeto de entrada o *input* está representado por un conjunto de casos b^t, denominados bolsas o *bags*, y es la bolsa la que predice un valor para la variable clase. Además, existen diferentes bolsas y pueden contener un número diferente de casos: $b^t = \{x_1^t, x_2^t, ..., x_{m^t}^t\}$, donde m^t es el número de casos en el *bag* t. En esta clasificación, el conjunto de datos de entrenamiento se denota como $T^t = (b^t, r^t)_{t=1}^N$, donde r^t es la clase del *bag* b^t y el clasificador genera una única decisión para el *bag* (Alpaydin et al., 2015).

En este trabajo los *bags* b^t son todos los conjuntos de víctimas que tienen el mismo autor. Para identificar los *bags* se ha creado una variable identificadora en la base de datos, llamada bag.

Por un lado, hay un total de 23 homicidios con múltiples víctimas y cometidos por un autor, que generan un total de 23 *bags* diferentes. La mayoría de homicidios cometidos por un autor y con más de una víctima el número total de víctimas son dos, siendo poco frecuentes las situaciones donde un autor acaba con la vida de más de dos personas (en dos homicidios hay un total de tres víctimas).

Por otro lado, hay un total de 5 homicidios con múltiples víctimas y múltiples autores. Para esta situación, para un hecho concreto se crearán tantos *bags* como número de autores haya. Así pues, el número total de *bags* obtenidos será la suma del número de autores que haya en cada homicidio con múltiples víctimas y múltiples autores.

En definitiva, el número total de *bags* en el presente estudio es de 36, obtenidos a partir de la siguiente fórmula:

$$\text{bags total} = \sum_{i=1}^{28} \text{autores}_i$$

donde *autores* es el número de autores que hay en los homicidios donde hay más de una víctima y definidos como hecho de caso primario, es decir, que es la primera vez que se registra en la base de datos aunque luego haya más casos que hacen referencia al mismo homicidio.

[14]Por sus siglas en inglés significa "Multi-Instance".

Una vez identificados los *bags* hay que realizar el entrenamiento y validación del clasificador bayesiano con la técnica *k-fold corss validation* (ver Sección 3.7.1). Seguidamente se detalla el proceso de entrenamiento y validación para una clasificación MI.

7.1. Entrenamiento para la clasificación MI

En el entrenamiento del clasificador bayesiano los elementos de una misma *bag* b^t no pueden separarse. Como en este trabajo se ha utilizado la técnica de *k-fold cross validation* para la evaluación de los modelos, con $k = 10$, se han repartido equitativamente los 36 *bags* en los 10 grupos, obteniendo seis *folds* con 4 *bags* y cuatro *folds* con 3 *bags*. Después de realizar la partición de la base de datos, teniendo en cuenta lo que se acaba de explicar, se hace el aprendizaje del clasificador bayesiano. Para el aprendizaje cada caso del *bag* se trata como un caso más del conjunto de entrenamiento T.

7.2. Validación para la clasificación MI

Una vez efectuado el entrenamiento del clasificador bayesiano se realiza la predicción para el conjunto de datos de validación V. Para los casos que no forman parte de un *bag* se predice la clase de la variable a partir de las evidencias entradas en el modelo, y la clase con mayor probabilidad es la predicción final. Para los casos que forman parte de un *bag* se complica un poco más. En esta ocasión también se hace una predicción para cada caso del *bag*, un total de m^t, pero como todos estos casos están intentando predecir los mismos valores de las variables del autor, se escoge la clase que tenga una mayoría de votos[15] y dicha predicción se establece para todos los elementos del *bag*. Para las situaciones donde haya un empate entre dos o más clases en la mayoría de votos, se escoge al azar una de ellas.

8. Proceso de construcción del clasificador bayesiano

El objetivo principal de este trabajo es poner múltiples etiquetas a un nuevo homicidio, donde las diferentes etiquetas son todas las variables del autor que se quieren predecir (sexo, franja de edad, situación laboral, etc.). Dado que no existe una única variable, el problema que se plantea es la clasificación de **múltiples etiquetas** (ML[16]).

[15] En inglés "Majority Vote"
[16] Por sus siglas en inglés significa "Multi-Label".

Además, en la sección anterior se ha visto que se trata de una clasificación Multi-Instance (hay conjuntos de casos que quieren predecir un único valor para la variable del autor). Por lo tanto, el modelo que se implementa es **Multi-Instance Multi-Label** (MI-ML).

Para la resolución de problemas ML existen varios enfoques, entre ellos el *Binary Relevance* y el *Chain Classifier*, que se explican a continuación.

8.1. Binary Relevance

El enfoque *Binary Relevance* consiste en construir un clasificador, en este caso será bayesiano y de tipo *Augmented Naive Bayes*, para cada etiqueta (variable clase u *output*) independientemente de las otras, asumiendo independencia entre ellas aunque en la realidad no sea así. Entonces para cada variable que se quiere predecir se aprende, a partir de los datos, una RB, sin tener en cuenta las otras variables que también se quieren predecir. En este trabajo se ha fijado, para cada RB, el algoritmo de aprendizaje *Hill-Climbing* con función *score* BIC (ver Sección 3.5.1) y el ajuste de los parámetros se realiza mediante el método de EMV (ver Sección 3.3).

Una vez realizada la predicción para todas las variables se pueden comparar los resultados que ha dado el modelo con las clases reales y calcular métricas de comportamiento interesantes.

8.2. Chain Classifier

El enfoque *Chain Classifier* consiste en construir clasificadores de las etiquetas (variables clase u *output*) en cadena, una detrás de otra. Primero hay que saber con qué orden se construirán los clasificadores de las variables, es decir, las RBs.

8.2.1. El orden ancestral

Para este estudio se ha usado el **orden ancestral** de las variables *output* como orden del clasificador en cadena. Para determinarlo se ajusta una RB "libre" para el conjunto de entrenamiento T_h (con h = 1, ..., 10), indicando la prohibición de flechas en el DAG desde las variables *input* (que son las variables del hecho y de la víctima que se conocen) hacia las *output*, es decir, creando una **blacklist**. Finalmente se encuentra el orden ancestral de las variables que se quieren predecir en la RB "libre" ajustada.

Para el ajuste de la RB "libre" se ha usado el algoritmo de aprendizaje *Hill-Climbing* con función *score* AIC, para obtener DAGs más conectados que empleando el *score* BIC. En el Anexo D se muestran las estructuras de las RBs "libres" para T_1 (cuando $k = 1$ en la técnica *k-fold cross*

validation) y para los cuatro modelos diferentes (descritos en la Sección 9). Se observa que para cada modelo el orden ancestral varia.

8.2.2. Construcción de las redes bayesianas en cadena

Después de obtener el orden ancestral se realiza la *Chain Classifier*. Primero se selecciona la primera variable *output* del orden y se ajusta una RB de tipo *Augmented Naive Bayes* a partir del conjunto de entrenamiento T. En este caso se utiliza el algoritmo de aprendizaje *Hill-Climbing* con función *score* BIC, y el ajuste de parámetros se realiza con el método de EMV. Segundo, se usa el conjunto de validación V para predecir la variable específica y se guardan sus resultados para cada caso del conjunto V. En la Figura 20 del Anexo E se puede ver cómo es la estructura de la RB para la primera variable *output*.

Para las variables siguientes se realizan los mismos pasos pero con variantes, ya que se trata de una clasificación en cadena.

Primero, para el ajuste de la estructura de la RB, de tipo *Augmented Naive Bayes*, las variables *output* previas en el orden ancestral de la variable que se quiere predecir se tratan como una variable *input* más. Es por eso que la estructura de la RB varía, obligando la flecha desde la variable *output* actual hacia las variables *output* previas al orden ancestral. En la Figura 21 del Anexo E se muestra la estructura del DAG para la segunda variable *output* del orden ancestral, donde se ve que la primera variable del orden se trata como una variable *input*.

Segundo, para la predicción también se usa el conjunto de validación V, pero en este caso las predicciones que se han obtenido anteriormente de las variables previas en el orden ancestral, se usan como si cada una de ellas fuera una evidencia más del homicidio.

Una vez realizada la predicción para todas las variables se pueden comparar las predicciones obtenidas con los valores observados y calcular diferentes métricas.

9. Modelos y resultados

Tal como se ha detallado anteriormente, el objetivo del trabajo es construir un modelo capaz de predecir las variables del autor dada la comisión de un nuevo homicidio. Así, la persona investigadora del caso puede orientar la investigación criminal y saber qué tipo de persona es la que más probablemente sea la autora del delito (González et al., 2018).

En el Cuadro 4 de la Sección 4 se sintetizan las variables elegidas para el desarrollo del modelo de clasificación bayesiano. Como se ha visto, hay variables referentes a las características del hecho ($H_1, ..., H_{14}$), de la víctima ($V_1, ..., V_9$), del autor ($A_1, ..., A_{10}$) y la relación víctima - autor (R). De todas ellas, unas variables se usan como evidencias (variables clase o *input*) y las otras son las que se quieren predecir (variables clase u *output*).

Inicialmente, las variables que se quieren predecir son las características del autor ($A_1, ..., A_{10}$) y la relación entre la víctima y el autor (R), pero existen algunas situaciones en las que se desconocen algunas características del hecho. Estas variables son: H_6 = número de autores y H_8 = tipología del hecho. Entonces, se construyen cuatro modelos diferentes según si las variables H_6 y H_8 son desconocidas. En el caso de que se conozcan, estas se tratan como variables *input* y, en caso contrario, como *output*, siendo una característica más del autor que se quiere predecir. En el Cuadro 10 se resume dicha información.

Cuadro 10: Variables *input* y *output* de los modelos

Modelo	Variables *input* o predictoras	Total v. *input*	Variables *output* o de clase	Total v. *output*
M1	H_2, H_3, ..., H_{14} V_1, V_2, ..., V_9	23	A_1, A_2, ..., A_{10}, R	11
M2	H_2, H_3, H_4, H_5, H_7, ..., H_{14} V_1, V_2, ..., V_9	22	A_1, A_2, ..., A_{10}, R, H_6	12
M3	H_2, ..., H_7, H_9, ..., H_{14} V_1, V_2, ..., V_9	22	A_1, A_2, ..., A_{10}, R, H_8	12
M4	H_2, ..., H_5, H_7, H_9, ..., H_{14} V_1, V_2, ..., V_9	21	A_1, A_2, ..., A_{10}, R, H_6, H_8	13

Fuente: Elaboración propia.

Los cuatro modelos resultantes se basan en problemas **Multi-Instance Multi-Label**. Para abordar el problema *Multi-Label* se utiliza el procedimiento *Binary Relevance* (detallado en 8.1) y el procedimiento *Chain Classifier* (detallado en 8.2) para realizar el clasificador bayesiano. Esto quiere decir que para cada modelo, de los cuatro totales, se obtienen dos clasificadores bayesianos diferentes creados con procedimientos diferentes. Después se comparan y discuten los resultados obtenidos para elegir el procedimiento que tenga mejores resultados en las métricas de comportamiento.

9.1. Modelo 1

El modelo 1 (M1) pretende predecir las características del autor ($A_1, ..., A_{10}$) y la relación entre víctima y autor (R), un total de 11 variables. Para hacerlo se usan las características del hecho ($H_2, ..., H_{14}$) y de la víctima ($V_1, ..., V_9$) como evidencias.

Visto que es un problema ML se construyen dos clasificadores bayesianos, uno mediante el procedimiento *Binary Relevance* (BR) y el otro con *Chain Classifier* (CC). En los dos procedimientos se usa la técnica *k-fold cross-validation* (ver Sección 3.7.1).

Para cada una de las divisiones (un total de $k = 10$) se obtiene la frecuencia con la que las variables *output* se predicen correctamente en los casos del conjunto de validación V. Una variable *output* se predice correctamente cuando la clase observada para un caso específico de V coincide con la clase predicha. Se denomina IPA[17] la proporción de predicciones correctas de una variable concreta sobre el número total de predicciones para esta variable, que corresponde al número de casos del conjunto de validación V, es decir, es su *accuracy*. Se denomina OPA[18] la proporción de predicciones correctas sobre el número total de predicciones totales, que para una ejecución de *k-fold cross validation* corresponde al número de variables a predecir multiplicado por el número de casos del conjunto de validación V.

Cuadro 11: Media del IPA y OPA usando la *k-fold cross validation*. Resultados referentes a los enfoques *Binary Relevance* (BR) y *Chain Classifier* (CC) del modelo 1

Variables *output* Modelo 1	*k-fold* IPA(%) media	
	BR	CC
A_1 = edad	52	59.8
A_2 = sexo	91.5	93.3
A_3 = origen	81.4	82.8
A_4 = situación laboral	66.6	80.8
A_5 = antecedentes	57.9	64.7
A_6 = adicciones	60.8	77.3
A_7 = frecuencia de consumo	77.8	83.9
A_8 = sustancia consumida	71.7	87.5
A_9 = trastorno mental	91.3	93.6
A_{10} = actividades ilegales	94.6	95.7
R = relación víctima - autor	75.1	77.8
Total OPA (%) media	54.7	58.5

Fuente: Elaboración propia.

A partir de los k valores obtenidos, se calcula la media de la IPA y del OPA. En el Cuadro 11 se muestran los resultados y se observa que la media del IPA para cada variable predictora es mayor en el enfoque CC. También se observa que para todas las predicciones posibles, el OPA, el enfoque CC supera de media cerca de 4 puntos porcentuales el BR.

[17]Por sus siglas en inglés significa "Individual Predictive Accuracy".
[18]Pos sus siglas en inglés significa "Overall Predictive Accuracy".

Después de ver los resultados generales, se realizan las pruebas pertinentes con la finalidad de constatar, para cada variable *output*, si la diferencia de los valores de la *accuracy* de cada variable (IPA) obtenidos con BR y CC es significativa.

En primer lugar, recordar que para cada enfoque (BR o CC) se han obtenido 10 valores de la *accuracy* para cada variable *output*, fruto de la *k-fold cross validation*. Por tanto, existen dos muestras de tamaño 10 de datos apareados, ya que cada pareja corresponde a un mismo *fold*. A partir de estas dos muestras se hace una resta en orden y se obtiene una nueva muestra de tamaño 10 con la que se realiza la prueba de Shapiro-Wilk, que plantea si dicha muestra proviene de una distribución normal.

En segundo lugar, se realiza una prueba para ver si la media o mediana de la muestra de la diferencia es cero, es decir, si existen diferencias significativas entre la media o mediana de la *accuracy* obtenida con BR y CC. La prueba que se usa depende de si anteriormente se ha aceptado o no normalidad en la muestra. Para las situaciones donde se acepta normalidad, se usa la prueba paramétrica t-test para hacer una comparación de las medias. Para las situaciones donde no se acepta normalidad, se usa la prueba no paramétrica de Wilcoxon, donde se comparan las medianas.

En el Cuadro 12 se muestran los resultados de la prueba unilateral de comparación de medias o medianas de la *accuracy* para cada variable *output*. En la mayoría de variables la media o mediana de la *accuracy* es mayor significativamente en el modelo obtenido con el enfoque *Chain Classifier*. Para las variables A_7 y A_{10} no se acepta una mediana y una media de la *accuracy* mayor significativamente con una confianza del 95 %, respectivamente, aunque sí con una del 90 %.

En resumen, con una confianza del 95 % los resultados de la CC son significativamente superior para 8 variables de las 11 totales, y con una confianza del 90 %, son significativamente superiores para 10 variables de las 11. Para A_3 no lo es en ningún caso.

Seguidamente se analizan otras métricas de comportamiento, realizando las pruebas de comparación de media o mediana pertinentes. Para las variables *output* ordinales se analiza el MAE, para las binarias el F-score y para las categóricas (con más de dos categorías) el MCC. Los resultados se detallan en el Cuadro 13 y hacen referencia a la media, calculada a partir de los $k = 10$ valores obtenidos con la técnica *k-fold cross validation*.

A partir del Cuadro 13 se observa que para la variable A_1 la media del MAE (que interesa que tenga valores cercanos a cero) es más baja significativamente en el modelo ajustado a partir del procedimiento CC.

También se ve que la media del MCC obtenida con CC es mayor significativamente para las variables A_4, A_5, A_6, A_8 y R. Además, para la variable A_9 la media del F-score es significativamente superior en CC con una confianza del 95 %, y para A_{10} con una confianza del 90 %.

Cuadro 12: Resultados de la prueba unilateral de comparación de medias o medianas de la *accuracy* (IPA) de datos apareados para el modelo 1

Variables *output*	Media o mediana	Media o mediana en % con BR	Media o mediana en % con CC	P-valor de la prueba unilateral de comparación de media o mediana
A_1 = edad	media	52	59.8	<.001 ***
A_2 = sexo	mediana	93.9	94.5	.046 *
A_3 = origen	media	81.4	82.8	.142
A_4 = situación laboral	media	66.6	80.8	<.001 ***
A_5 = antecedentes	media	57.9	64.7	.002 **
A_6 = adicciones	media	60.8	77.3	<.001 ***
A_7 = frecuencia de consumo	mediana	81.3	92	.051 ·
A_8 = sustancia consumida	media	71.7	87.5	<.001 ***
A_9 = trastorno mental	media	91.3	93.6	.012 *
A_{10} = actividades ilegales	mediana	95.6	95.7	.068 ·
R = relación víctima - autor	media	75.2	77.8	.005 **

Fuente: Elaboración propia.

Cuadro 13: Resultados de la prueba unilateral de comparación de medias o medianas de la métrica pertinente (MAE, F-score o MCC) de datos apareados para el modelo 1

Variables *output*	Métrica	Media o mediana	Media o mediana con BR	Media o mediana con CC	P-valor de la prueba unilateral de comparación de media o mediana
A_1 = edad	MAE	media	0.87	0.71	< .001 ***
A_2 = sexo	F-score	mediana	0.97	0.98	.284
A_3 = origen	MCC	media	0.62	0.66	.103
A_4 = situación laboral	MCC	media	0.50	0.71	.003 **
A_5 = antecedentes	MCC	media	0.38	0.48	.002 **
A_6 = adicciones	MCC	media	0.46	0.69	< .001 ***
A_7 = frecuencia de consumo	MCC	mediana	0.64	0.87	.253
A_8 = sustancia consumida	MCC	media	0.48	0.77	.001 **
A_9 = trastorno mental	F-score	media	0.58	0.71	.011 *
A_{10} = actividades ilegales	F-score	media	0.97	0.98	.074 ·
R = relación víctima - autor	MCC	media	0.68	0.71	.006 **

Fuente: Elaboración propia.

9.2. Modelo 2

El modelo 2 (M2) pretende predecir las características del autor $(A_1, ..., A_{10})$, la relación entre víctima y autor (R) y el número de autores (H_6), un total de 12 variables. Para hacerlo se usan las características del hecho $(H_2, ..., H_5, H_7, ..., H_{14})$ y de la víctima $(V_1, ..., V_9)$ como evidencias.

En este modelo se observa que para todas las variables la media o mediana de la *accuracy* es mayor significativamente en el clasificador bayesiano *Chain Classifier* con una confianza del 95 % (ver Cuadro 15).

Cuadro 14: Media del IPA y OPA usando la k-fold cross validation. Resultados referentes a los enfoques *Binary Relevance* (BR) y *Chain Classifier* (CC) del modelo 2

Variables *output* Modelo 2	*k-fold* IPA(%) media	
	BR	CC
A_1 = edad	51.7	58.3
A_2 = sexo	91	92.2
A_3 = origen	81.1	83.7
A_4 = situación laboral	68.6	80.8
A_5 = antecedentes	56.8	62.5
A_6 = adicciones	60.1	82.3
A_7 = frecuencia de consumo	77.1	84.0
A_8 = sustancia consumida	73.6	86.4
A_9 = trastorno mental	92.1	93.8
A_{10} = actividades ilegales	94.5	95.4
R = relación víctima - autor	74.1	78.1
H_6 = número de autores	86.2	88.3
Total OPA (%) media	57.1	60.8

Fuente: Elaboración propia.

Cuadro 15: Resultados de la prueba bilateral de comparación de medias o medianas de la *accuracy* (IPA) de datos apareados para el modelo 2

Variables *output*	Media o mediana en %	Media o mediana en % con BR	Media o mediana en % con CC	P-valor de la prueba unilateral de comparación de media o mediana
A_1 = edad	media	51.7	58.3	.003 **
A_2 = sexo	mediana	91.7	94	.018 **
A_3 = origen	mediana	80.6	83.8	.004 **
A_4 = situación laboral	media	68.6	80.8	.002 **
A_5 = antecedentes	media	56.8	62.5	.003 **
A_6 = adicciones	media	60.1	82.3	< .001 **
A_7 = frecuencia de consumo	media	77.1	84	.009 **
A_8 = sustancia consumida	media	73.6	86.4	.002 **
A_9 = trastorno mental	mediana	92.4	94	.018 *
A_{10} = actividades ilegales	media	94.5	95.4	.049 *
R = relación víctima - autor	media	74.1	78.1	.003 **
H_6 = número de autores	media	86.2	88.3	.017 *

Fuente: Elaboración propia.

A partir del Cuadro 16 se observa que la media del MCC es superior significativamente en CC para las variables *output* A_4, A_5, A_6, A_7, A_8, R y H_6. Los resultados de las variables *output* binarias, donde se muestra el F-score, se ve que los resultados para el modelo con el procedimiento CC son mejores, pero no significativamente. Aunque para A_{10} sí sería significativo con una confianza del 90 %.

Para la variable A_1 se ve que tiene un resultado mejor con el modelo ajustado con el procedimiento CC y, además, es significativo.

Cuadro 16: Resultados de la prueba unilateral de comparación de medias o medianas de la métrica pertinente (MAE, F-score o MCC) de datos apareados para el modelo 2

Variables *output*	Métrica	Media o mediana	Media o mediana con BR	Media o mediana con CC	P-valor de la prueba unilateral de comparación de media o mediana
A_1 = edad	MAE	media	0.89	0.75	.007 **
A_2 = sexo	F-score	mediana	0.95	0.97	.324
A_3 = origen	MCC	mediana	0.61	0.69	.128
A_4 = situación laboral	MCC	media	0.53	0.72	.002 **
A_5 = antecedentes	MCC	media	0.37	0.45	.003 **
A_6 = adicciones	MCC	media	0.45	0.75	< .001 ***
A_7 = frecuencia de consumo	MCC	mediana	0.55	0.70	.020 *
A_8 = sustancia consumida	MCC	media	0.50	0.75	.003 **
A_9 = trastorno mental	F-score	mediana	0.66	0.71	.200
A_{10} = actividades ilegales	F-score	media	0.96	0.97	.053 ·
R = relación víctima - autor	MCC	media	0.67	0.72	.002 **
H_6 = número de autores	MCC	media	0.88	0.90	.016 *

Fuente: Elaboración propia.

9.3. Modelo 3

El modelo 3 (M3) pretende predecir las características del autor $(A_1, ..., A_{10})$, la relación entre víctima y autor (R) y la tipología del hecho (H_8), un total de 12 variables. Para hacerlo se usan las características del hecho $(H_2, ..., H_7, H_9, ..., H_{14})$ y de la víctima $(V_1, ..., V_9)$ como evidencias.

En el M3 la *accuracy* obtenida con el procedimiento CC es superior significativamente casi para todas las variables, excepto para A_7 y A_9. A pesar de eso, las dos variables tienen un valor de la mediana y media de la *accuracy* mayor, respectivamente.

En el Cuadro 19 se observa que A_7 no tiene una mediana del MCC superior significativamente y A_9 no tiene una media del F-score superior significativamente. Aun así, el resultado de A_9 es significativo con una confianza del 90 %.

Para las otras variables todos los resultados son significativamente mejores para el enfoque *Chain Classifier*.

Cuadro 17: Media del IPA y OPA usando la k-fold cross validation. Resultados referentes a los enfoques *Binary Relevance* (BR) y *Chain Classifier* (CC) del modelo 3

Variables *output* Modelo 3	*k-fold* IPA(%) media	
	BR	CC
A_1 = edad	48.2	55.5
A_2 = sexo	92.3	94.8
A_3 = origen	80.4	82.9
A_4 = situación laboral	65.6	81.7
A_5 = antecedentes	56.6	64.8
A_6 = adicciones	56.4	74.3
A_7 = frecuencia de consumo	75.9	80.9
A_8 = sustancia consumida	74.6	87.7
A_9 = trastorno mental	92	92.6
A_{10} = actividades ilegales	93.8	95.6
R = relación víctima - autor	65.3	68.9
H_8 = tipología del hecho	66.6	70.8
Total OPA (%) media	57.1	58.3

Fuente: Elaboración propia.

Cuadro 18: Resultados de la prueba unilateral de comparación de medias o medianas de la *accuracy* (IPA) de datos apareados para el modelo 3

Variables *output*	Media o mediana en %	Media o mediana en % con BR	Media o mediana en % con CC	P-valor de la prueba unilateral de comparación de media o mediana
A_1 = edad	media	48.2	55.5	< .001 ***
A_2 = sexo	media	92.3	94.8	.007 **
A_3 = origen	media	80.4	82.9	.013 *
A_4 = situación laboral	media	65.6	81.7	.001 **
A_5 = antecedentes	media	56.6	64.8	< .001 ***
A_6 = adicciones	media	56.4	74.3	< .001 ***
A_7 = frecuencia de consumo	mediana	82.1	88	.101
A_8 = sustancia consumida	media	74.6	87.7	.001 **
A_9 = trastorno mental	media	92	92.6	.199
A_{10} = actividades ilegales	media	93.8	95.6	.004 **
R = relación víctima - autor	media	65.3	68.9	.003 **
H_8 = tipología del hecho	media	66.6	70.8	.004 **

Fuente: Elaboración propia.

Cuadro 19: Resultados de la prueba unilateral de comparación de medias o medianas de la métrica pertinente (MAE, F-score o MCC) de datos apareados para el modelo 3

Variables *output*	Métrica	Media o mediana	Media o mediana con BR	Media o mediana con CC	P-valor de la prueba unilateral de comparación de media o mediana
A_1 = edad	MAE	media	0.92	0.76	$< .001$ ***
A_2 = sexo	F-score	media	0.96	0.97	.008 **
A_3 = origen	MCC	media	0.60	0.65	.021 *
A_4 = situación laboral	MCC	media	0.49	0.73	.001 **
A_5 = antecedentes	MCC	media	0.36	0.48	$< .001$ ***
A_6 = adicciones	MCC	media	0.40	0.64	$< .001$ ***
A_7 = frecuencia de consumo	MCC	mediana	0.69	0.81	.283
A_8 = sustancia consumida	MCC	media	0.53	0.79	.002 **
A_9 = trastorno mental	F-score	media	0.60	0.65	.069 ·
A_{10} = actividades ilegales	F-score	media	0.97	0.98	.004 **
R = relación víctima - autor	MCC	media	0.55	0.60	.009 **
H_8 = tipología del hecho	MCC	media	0.59	0.64	.005 **

Fuente: Elaboración propia.

9.4. Modelo 4

El modelo 4 (M4) pretende predecir las características del autor ($A_1, ..., A_{10}$), la relación entre víctima y autor (R), el número de autores (H_6) y la tipología del hecho (H_8), un total de 13 variables. Para hacerlo se usan las características del hecho ($H_2, ..., H_5, H_7, H_9, ..., H_{14}$) y de la víctima ($V_1, ..., V_9$) como evidencias.

En el M4 la media o mediana de la *accuracy* del procedimiento CC es significativamente superior casi para todas las variables, menos para A_9, A_{10} y H_6. Aunque con una confianza del 90 %, A_9 y H_6 sí que tienen una media de la *accuracy* superior significativamente con CC.

A partir del Cuadro 22 se observa que para todas las variables con más de dos categorías la media del MCC es superior significativamente en el modelo realizado con el procedimiento CC, estas variables son: A_3, A_4, ..., A_8, R y H_8. Para las variables binarias, con una confianza del 95 %, la media del F-score del CC es mayor significativamente para A_2 y A_9, y con una confianza del 90 % también para A_{10} y H_6.

Para la variable A_1 se ve que tiene un resultado mejor con el modelo ajustado con el procedimiento CC, ya que se aproxima más a cero que con BR, y además es significativo.

Cuadro 20: Media del IPA y OPA usando la k-fold cross validation. Resultados referentes a los enfoques *Binary Relevance* (BR) y *Chain Classifier* (CC) del modelo 4

Variables *output* Modelo 4	*k-fold* IPA(%) media	
	BR	CC
A_1 = edad	49.6	55.9
A_2 = sexo	91.6	94.3
A_3 = origen	80.2	82.4
A_4 = situación laboral	66.9	81.2
A_5 = antecedentes	55.6	62.6
A_6 = adicciones	59	76.8
A_7 = frecuencia de consumo	75.3	83.1
A_8 = sustancia consumida	74.4	88.2
A_9 = trastorno mental	92	93.1
A_{10} = actividades ilegales	94.2	94.8
R = relación víctima - autor	62.7	65.8
H_6 = número de autores	83.8	85.8
H_8 = tipología del hecho	63.5	69.3
Total OPA (%) media	56.1	59.9

Fuente: Elaboración propia.

Cuadro 21: Resultados de la prueba unilateral de comparación de medias o medianas de la *accuracy* (IPA) de datos apareados para el modelo 4

Variables *output*	Media o mediana en %	Media o mediana en % con BR	Media o mediana en % con CC	P-valor de la prueba unilateral de comparación de media o mediana
A_1 = edad	media	49.6	55.9	$< .001$ ***
A_2 = sexo	mediana	91.6	94.3	$< .001$ ***
A_3 = origen	media	80.2	82.4	.008 **
A_4 = situación laboral	media	68.6	80.8	$< .001$ **
A_5 = antecedentes	media	55.6	62.6	$< .001$ ***
A_6 = adicciones	media	59	76.8	.001 **
A_7 = frecuencia de consumo	mediana	76.4	92	.007 **
A_8 = sustancia consumida	media	74.4	88.2	.001 **
A_9 = trastorno mental	media	92	93.1	.063 ·
A_{10} = actividades ilegales	mediana	96.1	96.1	.984
R = relación víctima - autor	mediana	63.8	67.9	.026 *
H_6 = número de autores	media	83.8	85.8	.065 ·
H_8 = tipología del hecho	mediana	63.5	69.3	.004 **

Fuente: Elaboración propia.

Cuadro 22: Resultados de la prueba unilateral de comparación de medias o medianas de la métrica pertinente (MAE, F-score o MCC) de datos apareados para el modelo 4

Variables *output*	Métrica	Media o mediana	Media o mediana con BR	Media o mediana con CC	P-valor de la prueba unilateral de comparación de media o mediana
A_1 = edad	MAE	media	0.90	0.80	.005 **
A_2 = sexo	F-score	mediana	0.95	0.97	< .001 **
A_3 = origen	MCC	media	0.59	0.64	.011 *
A_4 = situación laboral	MCC	media	0.52	0.73	< .001 ***
A_5 = antecedentes	MCC	media	0.35	0.45	< .001 ***
A_6 = adicciones	MCC	media	0.43	0.67	.001 **
A_7 = frecuencia de consumo	MCC	media	0.52	0.69	.009 **
A_8 = sustancia consumida	MCC	media	0.53	0.79	.001 **
A_9 = trastorno mental	F-score	media	0.59	0.66	.022 *
A_{10} = actividades ilegales	F-score	media	0.98	0.98	.650 ·
R = relación víctima - autor	MCC	media	0.52	0.56	.016 *
H_6 = número autores	F-score	media	0.86	0.88	.080 ·
H_8 = tipología del hecho	MCC	media	0.55	0.004	.002 **

Fuente: Elaboración propia.

10. Shiny

Shiny es un paquete libre de  que proporciona un entorno web elegante para implementar aplicaciones web usando el lenguaje de programación `R`. Shiny ayuda a convertir los análisis en aplicaciones interactivas sin ningún tipo de conocimiento de HTLM, CSS o JavaScript.

En este trabajo, la aplicación de Shiny se realiza a partir de `shinydashboard`, una extensión de Shiny que permite hacer tableros interactivos con capacidad de modificaciones, para así obtener diferentes resultados y personalizarlos. El código `R` para crear el Shiny se diferencia en dos funciones, la **ui** y la **server**.

- **ui**: se encarga de crear la interfaz visual de la aplicación. La función se compone de tres partes, la *dashboardHeader*, que corresponde a la cabecera del entorno web, la *dashboardSidebar*, que corresponde a la pestaña lateral, y la *dashboardBody*, que corresponde al contenido de la aplicación.
- **server**: en esta función se programa la lógica del servidor y se genera el contenido dinámico que depende de las interacciones con la pantalla.

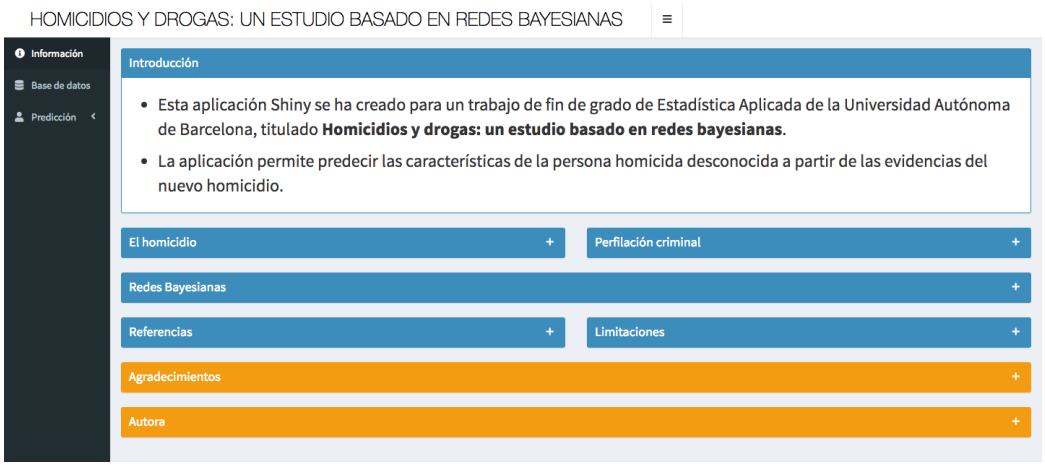
El objetivo principal de la creación de la aplicación de Shiny es poner a disposición a cualquier persona el modelo entrenado para que pueda ver las predicciones y probabilidades del perfil del autor desconocido dado un nuevo crimen. Cabe mencionar que a partir de los resultados obtenidos, el clasificador bayesiano final se entrena a partir de todos los casos de la base de datos y usando

el procedimiento *Chain Classifier*, porque los resultados son mejores si se comparan con *Binary Relevance*.

Para empezar, al abrir la aplicación se puede ver arriba del todo el título del trabajo y un icono con tres rayas paralelas que permite esconder la pestaña lateral. El usuario puede acceder rápidamente a cualquier opción de la pestaña, formada por **información**, que se muestra por defecto al abrir la aplicación, **base de datos** y **predicción**.

En el apartado de **información** (ver Figura 11) se muestra una breve introducción acerca del objetivo de la aplicación. Además, hay diferentes desplegables con información sobre los aspectos que se han estudiado en el presente trabajo, como son: el homicidio, la perfilación criminal y las redes bayesianas. También se puede consultar información de las principales referencias y limitaciones del trabajo. Finalmente, hay un desplegable con los agradecimientos y otro sobre la autoría del trabajo.

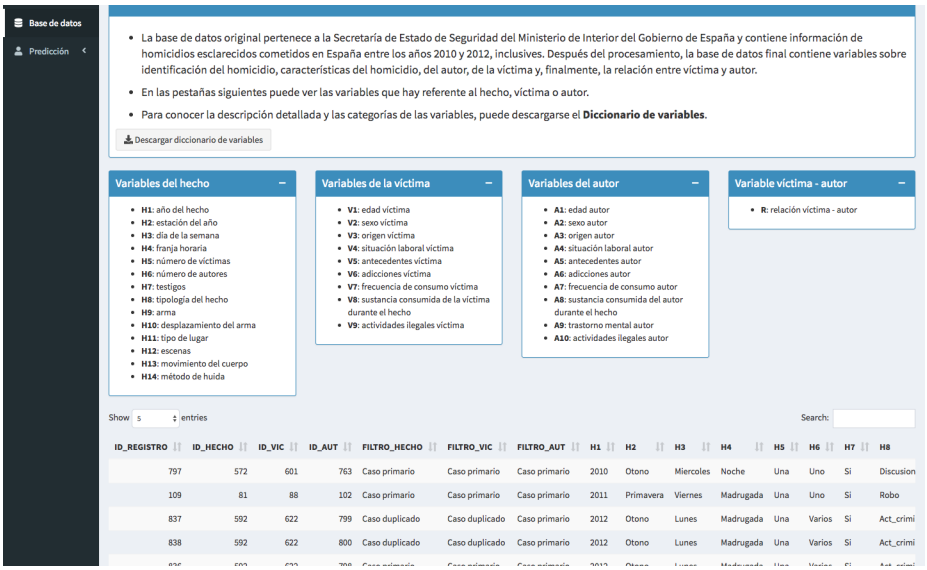
Figura 11: Pestaña de inicio por defecto de la aplicación Shiny



Fuente: Captura de pantalla de la aplicación Shiny creada por la autora.

En el apartado de **base de datos** (ver Figura 12) hay una breve explicación de dónde pertenece la base de datos, qué se puede encontrar y la posibilidad de descargarse un fichero en formato PDF con la explicación detallada de todas las variables y sus categorías (como el del Anexo B). Además, hay cuatro desplegables que muestran las variables según si hacen referencia a las características del hecho, del autor, de la víctima o de la relación víctima - autor. También se puede ver un total de 10 casos para hacerse una idea de cómo es la base de datos que se emplea para la realización del modelo probabilístico.

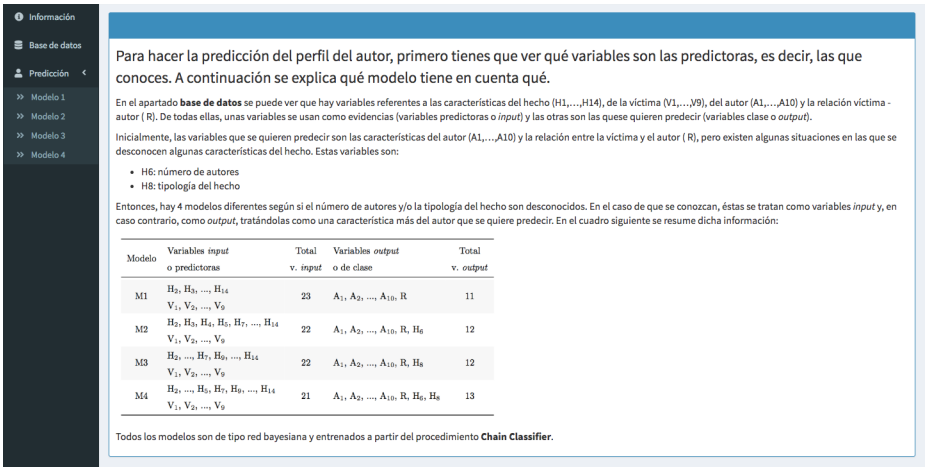
Figura 12: Pestaña “base de datos” de la aplicación Shiny



Fuente: Captura de pantalla de la aplicación Shiny creada por la autora.

En el tercer y último apartado, llamado **predicción** (ver Figura 13), se explica cuáles son las variables predictoras y las que se quieren predecir, de las presentadas en el apartado **base de datos**. Además, se expone la necesidad de construir 4 modelos diferentes según cuáles son las variables *input*. Para realizar la predicción de cada situación, cuando se pulsa el apartado **predicción** aparece un desplegable con subapartados, donde la persona puede elegir el modelo que necesita para hacer la predicción del perfil del autor dado un nuevo homicidio.

Figura 13: Pestaña “predicción” de la aplicación Shiny

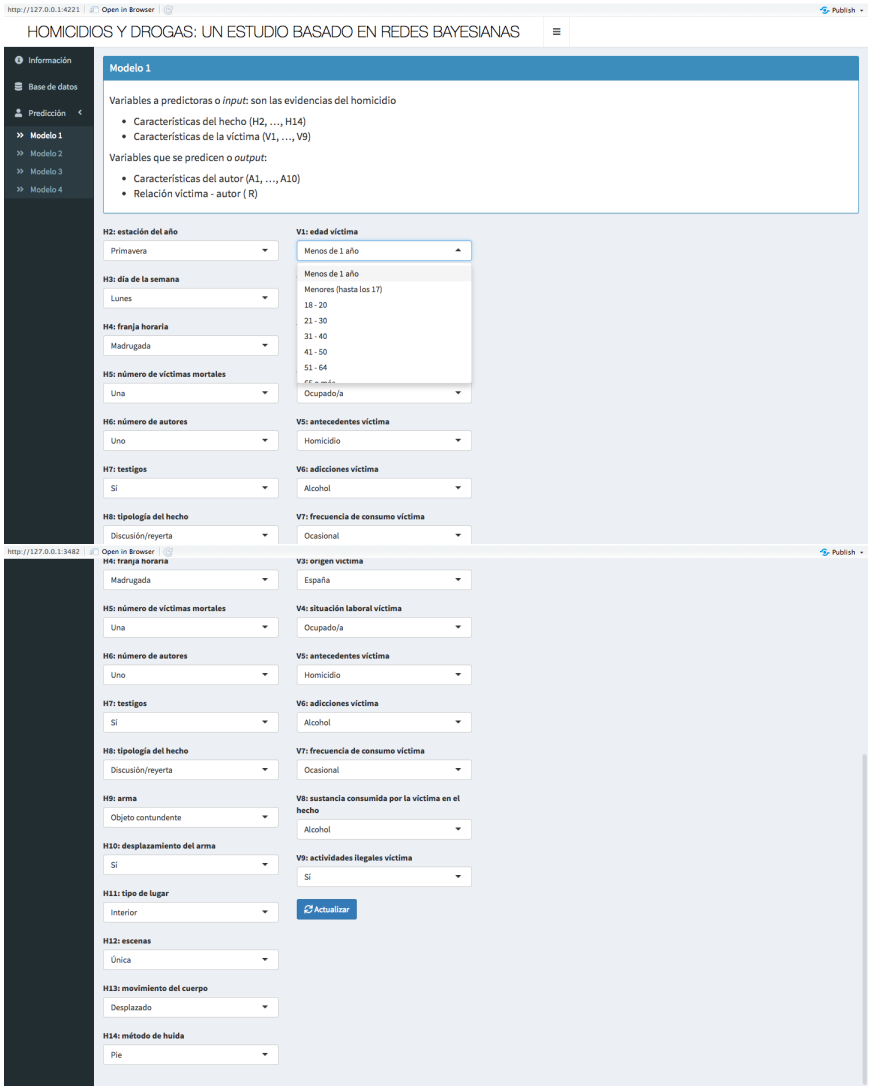


Fuente: Captura de pantalla de la aplicación Shiny creada por la autora.

Para los cuatro modelos se usa un clasificador de tipo red bayesiana y realizado con *Chain Classifier*, lo que significa que cada modelo tiene un orden ancestral propio (aprendido a partir de los datos) y cada variable *output* del modelo tiene su DAG específico. En el Anexo F se muestran los DAGs que se emplean para encontrar dicho orden (ver las Figuras 24, 25, 26 y 27).

En el subapartado **modelo 1** hay una caja de información sobre qué variables son las evidencias del homicidio y las que se quieren predecir. Luego hay un conjunto de desplegables debajo, que permiten elegir los valores correspondientes a cada variable de evidencia que se conoce (ver Figura 14).

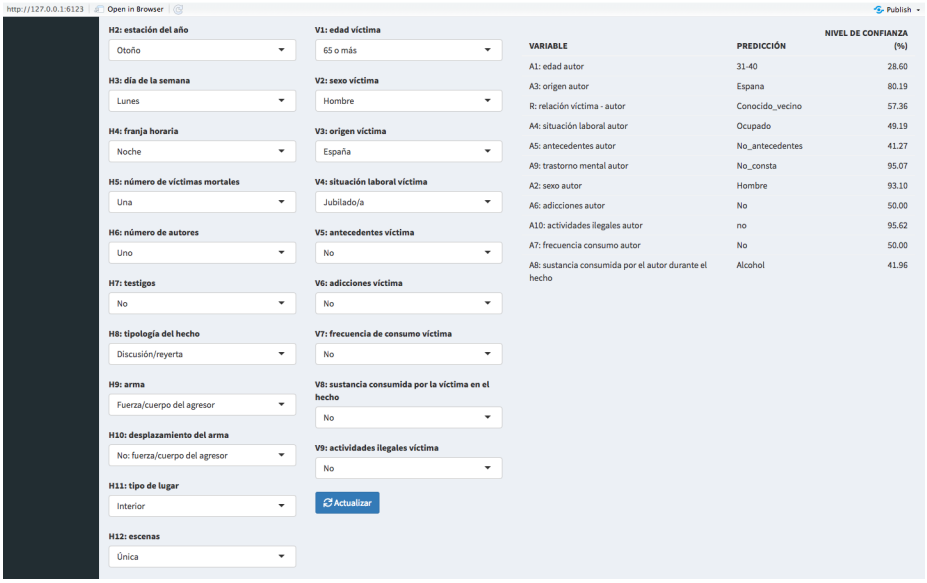
Figura 14: Pestaña "Modelo 1" del subapartado "predicción" de la aplicación Shiny



Fuente: Captura de pantalla de la aplicación Shiny creada por la autora.

Finalmente, dentro del subapartado hay un botón de actualizar para que el modelo calcule las predicciones de las variables del autor y su probabilidad. En la Figura 15 se muestra la predicción del autor del nuevo homicidio a partir de las evidencias entradas. Además, la columna nivel de confianza (en porcentaje) indica la probabilidad de que sea dicha categoría la predicción de la variable.

Figura 15: Ejemplo de predicción y nivel de confianza del perfil del autor a partir de unas evidencias



Fuente: Captura de pantalla de la aplicación Shiny creada por la autora.

En el Anexo G se puede ver más detalle sobre el diseño de la aplicación web Shiny, como la información que hay en los desplegables de la pestaña **introducción** y la cabecera de los demás subapartados de **predicción**.

11. Conclusiones

El estudio del fenómeno del homicidio principalmente se ha desarrollado desde la psicología y la criminología, aunque con la mayor disponibilidad de herramientas tecnológicas e informáticas se han podido recopilar bases de datos y acercase al fenómeno desde la estadística. El objetivo de este trabajo es utilizar las redes bayesianas para realizar el perfil criminológico, teniendo en cuenta la relación con las drogas, y predecir las características del delincuente dado un nuevo homicidio.

El problema a resolver es del tipo **Multi-Instance Multi-Label**. *Multi-Instance* porque hay conjuntos de casos que quieren predecir un único valor para las variables *output*, y *Multi-Label* porque hay que poner más de una etiqueta al fenómeno del homicidio para hacer el perfil del autor. Para abordar la problemática, se construyen dos clasificadores, formados por redes bayesianas, realizados con procedimientos diferentes: *Binary Relevance* y *Chain Classifier*.

Una vez construidas las RBs con procedimientos diferentes se ha visto que los resultados con el enfoque *Chain Classifier* son mejores. Por esta razón se utiliza este enfoque para la construcción del clasificador bayesiano definitivo, un total de 4 según el conocimiento o no de las dos variables del hecho: H_6 = número de autores y H_8 = tipología del hecho.

Los modelos finales se han entrenado a partir de todos los casos de la base de datos. Después se ha creado el entorno web con Shiny para que cualquier persona pueda entrar las variables del nuevo homicidio y observar las predicciones del perfil del autor y sus probabilidades. Esto permite extraer resultados y conclusiones más rápidamente y de forma sencilla.

De los modelos ajustados, en general, el que tiene mayor capacidad predictiva es el modelo 2 (M2) porque tiene un OPA mayor al resto de modelos, exactamente 60.8 %. Es decir, cuando la variable H_6 (número de autores) es desconocida y se quiere predecir. En cambio, para la variable H_8 (tipología del hecho) se obtienen mejores resultados cuando es conocida, que corresponde al M1, que cuando se quiere predecir, que corresponde al M3. En la situación donde se desconocen H_6 y H_8, el M4, los resultados generales son mejores que cuando se conocen las dos variables, el M1. Este acontecimiento se puede explicar porqué las dos variables tienen una IPA elevada y están por encima de la media, pero H_6 tiene una IPA más elevada que la de H_8.

Para el caso donde se conoce el número de autores y tipología del homicidio (corresponde al modelo 1), dadas las características del hecho y de la víctima se puede ver que para el caso donde la víctima no presenta adicción al alcohol ni a las drogas, no presenta un patrón de consumo y tampoco había consumido momentos antes de su muerte, el modelo indica que el autor tampoco tiene adicciones (con una probabilidad del 50 %) y no presenta un patrón de consumo (con un 50 % de probabilidad). Aun así, con un 42 % el modelo predice que el autor había consumido alcohol momentos antes de

perpetrar el homicidio. En cambio, cuando la víctima presenta evidencias de ser adicta al alcohol, con un consumo habitual y que momentos antes de su muerte había consumido alcohol, el modelo predice que el autor es adicto al alcohol (con una probabilidad del 53 %), su frecuencia de consumo es habitual (63 %) y momentos antes de cometer el homicidio había consumido alcohol (60 %). Por tanto, se puede ver que hay una asociación entre el consumo de alcohol de la víctima y el autor[19]

A partir de aquí los modelos pueden mejorar si se recogen más datos y más actuales, ya que son de hace una década. Además de ajustarse más al perfil del homicida, se podría estudiar si ha habido un cambio en la forma de perpetrar los homicidios y/o en el perfil del homicida.

Otro aspecto que puede mejorarse es la escritura de los atestados policiales y su vaciado para crear la base de datos, ya que en muchas ocasiones no se detalla específicamente información sobre las drogas ni trastorno mental, generando un gran porcentaje de valores desconocidos en las variables y, por tanto, una menor precisión en la perfilación criminal.

En este trabajo se ha centrado el estudio del homicidio en relación en las drogas, pero podrían predecirse otras variables y/o menos, según el tipo de homicidio que se ha cometido. Aunque se han creado cuatro modelos para tener en cuenta las diferentes evidencias que puede tener la persona investigadora, podría ser que algunas evidencias no fueran conocidas, por tanto se podrían construir otros modelos.

Hay que tener en cuenta que los resultados no representan la realidad al 100 %, sino que se trata de un modelo probabilístico y su uso es orientativo para la resolución de un nuevo caso y para que las personas investigadoras puedan combinar con otras técnicas policiales.

[19]Para ver las otras evidencias introducidas del hecho y de la víctima, consulte la Figura 35 del Anexo H.

Referencias

Alpaydin, E., Cheplygina, V., Loog, M., and Tax, D. D. J. (2015). Single- vs. multiple-instance classification. *Pattern Recognition*, (48):2831–2838.

Baumgartner, K., Serrari, S., and Palermo, G. (2008). Constructing bayesian networks for criminal profiling from limited data. *Knowledge - Based Systems*.

Darwiche, A. (2009). *Modeling and reasoning with Bayesian Networks*. Cambridge University Press.

Delgado, R. and Tibau, X.-A. (2019). Why cohen's kappa should be avoided as performance measure in classification. *PLOS ONE*, 14(9):1–26.

Farrington, D. P. and Lambert, S. (2017). Predicting offender profiles from offense and victims characteristics. *Criminal Profiling: International Theory Research and Practice*, pages 135–167.

Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, 29:131–163.

Garrido, V. and Sobral, J. (2008). *La investigación criminal. La psicología aplicada al descubrimiento, captura y condena de los criminales*. Barcelona, España: Nabla Ediciones.

González, J., Sánchez, F., López-Ossorio, J., Santos, J., and Cereceda, J. (2018). Informe sobre el homicidio. España 2010-2012. *Ministerio del Interior. Madrid, España*.

Gómez, J. A., Mateo, J. L., and Puerta, J. M. (2011). Learning bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood. *Data Min Knowl Disc*, (22):106–148.

INE (2017). Tasa de homicidio. *Instituto Nacional de Estadística. Madrid, España*.

Ioannou, M. and Hammond, L. (2015). The changing face of homicide research: The shift in empirical focus and emerging research trends. *Journal of Criminal Psychology*, 5(3):157–162.

Jiménez, J. (2012). *Manual Práctico del perfil criminológico. Criminal profiling*. Lex Nova.

Ley Orgánica (1/2004). , de 28 de diciembre, de medidas de protección integral contra la violencia de género, modificada el 4 de agosto de 2018.

Liem, M. (2013). Homicide offender recidivism: A review of literature. *Agression and Violent Behavior. Madrid, España*, 18:19–25.

Neapolitan, R. E. (2004). *Learning Bayesian Networks*. Prentice Hall Series in Artificial Intelligence.

Palermo, G. and Kocsis, R. N. (2004). *Offender profiling: An introduction to the sociopsychological analysis of violent crime*. Garland publishing.

Pecino, M. M. (2019). *La perfilación criminal en homicidios: implicaciones prácticas en el ámbito policial y educativo*. (Tesis doctoral). Universidad de Almería.

Scutari, M. (2010). Learning bayesian networks with the bnlearn r package. *Journal of Statistical Software*, VV.

UNODC (2013). Global Study on Homicide. Executive summary. *United Nations Office on Drugs and Crime*.

UNODC (2015). International Classification of Crimen for Statistical Proposes, Version 1.0. Retrieved September 4, 2019. *United Nations Office on Drugs and Crime*.

UNODC (2019). Global Study on Homicide 2019. *United Nations Office on Drugs and Crime. Viena*.

Anexos

A. Definición del homicidio en el Código Penal

El Código Penal Español (Ley Orgánica 10/1995), en su Libro II (Delitos y sus penas), Título I (Del homicidio y sus formas), se define el término homicidio dolodos y el de su versión agravada, el asesinato.

Artículo 138.

1. El que matare a otro será castigado, como reo de homicidio, con la pena de prisión de diez a quince años.

2. Los hechos serán castigados con la pena superior en grado en los siguientes casos:

a) cuando concurra en su comisión alguna de las circunstancias del apartado 1 del artículo 140, o

b) cuando los hechos sean además constitutivos de un delito de atentado del artículo 550.

Artículo 139.

1. Será castigado con la pena de prisión de quince a veinticinco años, como reo de asesinato, el que matare a otro concurriendo alguna de las circunstancias siguientes:

1.ª Con alevosía.

2.ª Por precio, recompensa o promesa.

3.ª Con ensañamiento, aumentando deliberada e inhumanamente el dolor del ofendido.

4.ª Para facilitar la comisión de otro delito o para evitar que se descubra.

2. Cuando en un asesinato concurran más de una de las circunstancias previstas en el apartado anterior, se impondrá la pena en su mitad superior.

Artículo 140.

1. El asesinato será castigado con pena de prisión permanente revisable cuando concurra alguna de las siguientes circunstancias:

1.ª Que la víctima sea menor de dieciséis años de edad, o se trate de una persona especialmente vulnerable por razón de su edad, enfermedad o discapacidad.

2.ª Que el hecho fuera subsiguiente a un delito contra la libertad sexual que el autor hubiera cometido sobre la víctima.

3.ª Que el delito se hubiera cometido por quien perteneciere a un grupo u organización criminal.

2. Al reo de asesinato que hubiera sido condenado por la muerte de más de dos personas se le impondrá una pena de prisión permanente revisable. En este caso, será de aplicación lo dispuesto en la letra b) del apartado 1 del artículo 78 bis y en la letra b) del apartado 2 del mismo artículo.

B. Diccionario de variables

En este anexo se detalla el nombre de las variables, la abreviatura utilizada en la base de datos, la definición y las categorías.

1. Identificación del hecho (ID_HECHO): número de registro asignado a cada homicidio.
2. Filtro del hecho (FILTRO_HECHO): identifica los hechos duplicados.
    - Caso primario.
    - Caso duplicado.
3. Identificación de la víctima (ID_VIC): número de registro asignado a cada víctima.
4. Filtro de la víctima (FILTRO_VIC): identifica las víctimas duplicadas.
    - Caso primario.
    - Caso duplicado.
5. Identificación del autor (ID_AUT): número de registro asignado a cada autor.
6. Filtro del autor (FILTRO_AUT): identifica los autores duplicados.
    - Caso primario.
    - Caso duplicado.
7. Año del hecho (H1): año en que se produjeron los hechos. La muestra está integrada por homicidios cometidos durante tres años.
    - 2010.
    - 2011.
    - 2012.
8. Estación del año (H2): estación del año en que se produjeron los hechos.
    - Primavera: del 20 de marzo hasta el 20 de junio.
    - Verano: del 21 de junio hasta el 21 de septiembre.
    - Otoño: del 22 de septiembre hasta el 20 de diciembre.

- Invierno: del 21 de diciembre hasta el 19 de marzo.

9. Día de la semana (H3): día de la semana en que se produjeron los hechos.
    - Lunes.
    - Martes.
    - Miércoles.
    - Jueves.
    - Viernes.
    - Sábado.
    - Domingo.
10. Hora (H4): franja horaria en la que se produjeron los hechos.
    - Madrugada: 00:00h - 05:59h.
    - Mañana: 06:00h - 11:59h.
    - Tarde: 12:00h - 17:59h.
    - Noche: 18:00h - 23:59h.
11. Número de víctimas mortales (H5): indica si hubo una o más de una víctima mortal en el hecho.
    - Una víctima.
    - Varias víctimas.
12. Número de autores (H6): indica si hubo uno o más de un autor. No se incluyen los cómplices.
    - Un autor.
    - Varios autores.
13. Testigos (H7): indica la existencia de testigos principales. Se entiende por testigo principal aquel que puede acreditar la comisión del hecho delictivo (homicidio) y que puede aportar datos sobre el autor o autores, pudiendo llegar a identificarlo.
    - Sí.
    - No.
14. Tipología del hecho (H8): indica el contexto en que se desarrolló el homicidio.

- Discusión/reyerta: derivados de la intención de resolver un conflicto o de castigar a la víctima (conocida o desconocida) mediante la violencia cuando las relaciones se tensan por diferentes causas y en los que se recurre a la violencia como estrategia de afrontamiento.
- Violencia de género: derivados de la reacción o estrategia violenta frente a un conflicto, en los que el autor (hombre) mata a la víctima mujer con la que mantiene o mantenía una relación de afectividad o análoga. Quedan excluidos los casos en los que la autora mata a un hombre y los casos de parejas homosexuales (Ley Orgánica , 2004).
- Violencia doméstica/familiar: se incluyen los casos en los que los implicados tienen relación familiar, o conviven en el mismo domicilio. También se incluye los casos en los que la mujer mata a la pareja (hombre) y los casos de parejas homosexuales masculinas y femeninas.
- Otras interpersonales: derivan de la intención de resolver un conflicto o de castigar a la víctima (conocida o desconocida) mediante la violencia cuando las relaciones se tensan por diferentes causas. Estos no se pueden clasificar en las categorías: discusión/reyerta, violencia de género ni resto de violencia doméstica/familiar.
- Robo: el homicidio está relacionado con la comisión de un robo. Con base en lo expresado en el artículo 268 del Código Penal, se considerará como actividad criminal la comisión de robo entre los cónyuges que no estuvieran separados legalmente o de hecho o en proceso judicial de separación, divorcio o nulidad de su matrimonio y los ascendientes, descendientes y hermanos por naturaleza o por adopción, así como los afines en primer grado si viviesen juntos, por hacer uso de violencia.
- Actividades criminales: el homicidio está relacionado con la comisión de la actividad criminal, bien cuando dicha actividad criminal sucede con anterioridad al homicidio o cuando el homicidio facilita la comisión de otro hecho delictivo. Incluye:
  - Organizaciones/Grupos criminales: el homicidio se da en el marco de las actividades criminales de grupos u organizaciones criminales. Entendiendo por organización criminal la agrupación formada por más de dos personas con carácter estable o por tiempo indefinido, que de manera concertada y coordinada se repartan diversas tareas o funciones con el fin de cometer delitos (artículo 570 bis del Código Penal). Y entendiendo por grupo criminal la unión de más de dos personas que, sin reunir alguna o algunas de las características de la organización criminal definida en el artículo 570 bis, tenga por finalidad o por objeto la perpetración concertada de delitos (artículo 570 ter del Código Penal).
  - Prostitución: víctima o autor ejercen la prostitución, y el homicidio se produce durante el ejercicio de esa actividad.

- Bandas: homicidios cometidos entre bandas, sobre todo de carácter juvenil.
- Otras actividades criminales: otro tipo de contexto criminal que no se puede clasificar en tipologías anteriores.

15. Arma (H9): arma o mecanismo utilizado por el autor para acabar con la vida de la víctima.
    - Objeto contundente: objeto que carece de punta y/o filo y puede presentar aristas romas que puede ser utilizado para golpear y producir lesiones contusas.
    - Arma blanca: arma constituida por una hoja metálica u otro material de características físicas semejantes, cortante o punzante.
    - Arma de fuego: toda arma portátil que tenga cañón y que lance, esté concebida para lanzar o pueda transformarse fácilmente para lanzar un perdigón, bala o proyectil por la acción de un combustible propulsor. De este modo, se considerará que un objeto es susceptible de transformarse para lanzar un perdigón, bala o proyectil por la acción de un combustible propulsor cuando tenga la apariencia de una arma de fuego y debido a su construcción o al material con el que está fabricada.
    - Fuerza/cuerpo del agresor: el autor emplea su cuerpo para agredir mortalmente a la víctima, asimismo, se han tenido en cuenta los casos en los que el agresor agarra a la víctima precipitándola al vacío.
    - Medios asfixiantes: el autor hace uso de alguna herramienta/instrumento para asfixiar a la víctima.
    - Otros: cualquier otro medio utilizado para la comisión del homicidio que no se encuentre descrito en las categorías anteriores.
16. Desplazamiento del arma (H10): indica si el arma fue desplazada por el agresor del lugar de los hechos.
    - Sí
    - No
    - Fuerza/cuerpo del agresor: el autor no hace uso de una arma, sino que emplea su cuerpo para agredir mortalmente a la víctima, asimismo, se han tenido en cuenta los casos en los que el agresor agarra a la víctima precipitándola al vacío.
17. Tipo de lugar (H11): tipo de lugar donde se comete el homicidio.
    - Escena del crimen interior: homicidios cometidos en escenas protegidas, como el domicilio.
    - Escena del crimen exterior: homicidios cometidos en lugares expuestos a la naturaleza.

- Escena del crimen en vehículo: homicidios cometidos en el interior de vehículos móviles.
- Escena del crimen en el agua: homicidios cometidos cerca de una zona acuática o debajo del agua.

18. Escenas (H12): indica si el hecho es de escena única o multiescena.
    - Escena única
    - Multiescena
19. Movimiento del cuerpo (H13): indica si el autor desplazó y/o ocultó el cuerpo de la víctima de la escena del crimen.
    - Desplazado, pero no oculto: el autor desplazó pero no ocultó el cuerpo de la víctima de la escena del crimen.
    - Oculto, pero no desplazado: el autor ocultó pero no desplazó el cuerpo de la víctima de la escena del crimen.
    - Desplazado y oculto: el autor desplazó y ocultó el cuerpo de la escena del crimen.
    - Ni desplazado ni oculto: el autor no desplazó ni ocultó el cuerpo de la víctima de la escena del crimen.
20. Método de huida (H14): método de huida empleado por el autor del homicidio.
    - A pie: el autor se marcha del lugar de los hechos andando o corriendo, sin usar ningún tipo de vehículo.
    - Se queda/es detenido en la escena del crimen.
    - Vehículo: el autor del crimen usa cualquier tipo de vehículo para huir de la escena del crimen, sea motorizado o no.
    - Suicidio/tentativa de suicidio: siempre y cuando el suicidio o la tentativa se produzcan en la escena. Esta categoría difiere de la expresada en consecuencias del autor, pues aquí solo se clasifica de esta manera si lo ejecuta espacial y temporalmente próximo al hecho. En caso de que el autor se marchara de la escena e intetara suicidarse en otro lugar, se recogerá únicamente la manera en la que se marcha.
21. Franjas de edad de la víctima (V1): edad de la víctima agrupada en franjas.
    - 0 años: primer año de vida.
    - Menores: desde 1 año hasta los 17 años.
    - 18 a 20 años.

- 21 a 30 años.
- 31 a 40 años.
- 41 a 50 años.
- 41 a 64 años.
- Más de 64 años.

22. Sexo víctima (V2): variable dicotómica que indica el sexo de la víctima.
    - Hombre.
    - Mujer.
23. País origen de la víctima (V3): nacionalidad de origen de la víctima.
    - España.
    - Resto de Europa.
    - Resto del mundo: África, América, Asia y resto de países.
24. Situación laboral de la víctima (V4): situación laboral de la víctima en el momento de los hechos.
    - Estudiante: persona escolarizada y que no realiza otra actividad o trabajo, y que se encuentra cursando algún estudio desde el ciclo de infantil al universitario. En caso de compatibilizar esta situación con la de ocupado, deberá darse prioridad a la que abarque mayor duración en la jornada.
    - Ocupado: persona asalariada ya sea por cuenta propia o cuenta ajena, tanto en el sector público como en el privado. Excepcionalmente se incluye a las personas que trabajan en una empresa familiar aunque sin remuneración.
    - Parado: persona mayor de 16 años que forma parte de la población activa, pero carece de empleo remunerado.
    - Jubilado: persona que percibe la pensión de jubilado, independientemente de la edad o motivación.
    - Otra situación laboral: persona que su situación laboral no se puede categorizar en las clasificaciones definidas anteriormente. La persona tiene un beneficio económico, pero no implica una situación laboral regulada (ejemplo: prostitución, tráfico de drogas, etc.).
25. Antecedentes de la víctima (V5): existencia de antecedentes penales y policiales de la víctima, exluyendo las sanciones administrativas y las denuncias que no acabaran en condena o detención.

- Homicidio o tentativa: la víctima tiene antecedentes por homicidios, asesinatos o algunos de éstos.
- Contra las personas: la víctima tiene antecedentes contra la vida, integridad y libertad de las personas. Se han incluido aquellas tipologías criminales que implican comportamiento violento o intimidatorio, de cualquier tipo, contra una persona física.
- Sí, pero desconocidos: la víctima tiene antecedentes pero no son contra las personas ni por homicidio, sino que son de otro tipo.
- No: la víctima no tiene antecedentes.

26. Adicciones de la víctima (V6): indica la adicción de la víctima.
    - Solo alcohol.
    - Solo drogas.
    - Ambos.
    - No.
27. Frecuencia de consumo de la víctima (V7): indica el patrón de consumo de sustancias de la víctima.
    - Ocasional: se puede discernir que la víctima consume de forma esporádica, sin que sea recurrente este consumo.
    - Habitual: se puede discernir que la víctima tiene un patrón de consumo más o menos constante.
    - No.
    - Frecuencia desconocida: la víctima consume sustancias pero se desconoce el patrón de consumo.
28. Sustancia en el hecho de la víctima (V8): indica el tipo de consumo de la víctima momentos antes de los hechos o durante los mismos.
    - Solo alcohol.
    - Solo drogas.
    - Ambos.
    - Sí, pero desconocidos: la víctima había consumido momentos antes de los hechos o durante los mismos, pero se desconoce el tipo de sustancia consumida.
    - No.

29. Actividades ilegales o registradas de la víctima (V9): indica si la víctima realizaba alguna actividad no registrada o ilegal que le sirviera de sustento (ejemplo: prostitución, tráfico de drogas, etc.).
    - Sí.
    - No.
30. Franjas de edad del autor (A1): edad del autor agrupada en franjas.
    - Menores: hasta los 17 años.
    - 18 a 20 años.
    - 21 a 30 años.
    - 31 a 40 años.
    - 41 a 50 años.
    - Más de 50 años.
31. Sexo autor (A2): variable dicotómica que indica el sexo del autor.
    - Hombre.
    - Mujer.
32. País origen del autor (A3): nacionalidad de origen del autor.
    - España.
    - Resto de Europa.
    - Resto del mundo: África, América, Asia y resto de países.
33. Situación laboral del autor (A4): situación laboral del autor en el momento de los hechos.
    - Estudiante: persona escolarizada y que no realiza otra actividad o trabajo, y que se encuentra cursando algún estudio desde el ciclo de infantil al universitario. En caso de compatibilizar esta situación con la de ocupado, deberá darse prioridad a la que abarque mayor duración en la jornada.
    - Ocupado: persona asalariada ya sea por cuenta propia o cuenta ajena, tanto en el sector público como en el privado. Excepcionalmente se incluye a las personas que trabajan en una empresa familiar aunque sin remuneración.
    - Parado: persona mayor de 16 años que forma parte de la población activa, pero carece de empleo remunerado.

- Jubilado: persona que percibe la pensión de jubilado, independientemente de la edad o motivación.
- Otra situación laboral: persona que su situación laboral no se puede categorizar en las clasificaciones definidas anteriormente. La persona tiene un beneficio económico, pero no implica una situación laboral regulada (ejemplo: prostitución, tráfico de drogas, etc.).

34. Antecedentes del autor (A5): existencia de antecedentes penales y policiales del autor, excluyendo las sanciones administrativas y las denuncias que no acabaran en condena o detención.
    - Homicidio o tentativa: el autor tiene antecedentes por homicidios, asesinatos o algunos de estos.
    - Contra las personas: el autor tiene antecedentes contra la vida, integridad y libertad de las personas. Se han incluido aquellas tipologías criminales que implican comportamiento violento o intimidatorio, de cualquier tipo, contra una persona física.
    - Sí, pero desconocidos: el autor tiene antecedentes pero no son contra las personas ni por homicidio, sino que son de otro tipo.
    - No: el autor no tiene antecedentes.
35. Adicciones del autor (A6): indica la adicción del autor.
    - Solo alcohol.
    - Solo drogas.
    - Ambos.
    - No.
36. Frecuencia de consumo del autor (A7): indica el patrón de consumo de sustancias del autor.
    - Ocasional: se puede discernir que el autor consume de forma esporádica, sin que sea recurrente este consumo.
    - Habitual: se puede discernir que el autor tiene un patrón de consumo más o menos constante.
    - No.
    - Frecuencia desconocida: el autor consume sustancias pero se desconoce el patrón de consumo.
37. Sustancia en el hecho del autor (A8): indica el tipo de consumo del autor momentos antes de los hechos o durante los mismos.
    - Solo alcohol.
    - Solo drogas.

- Ambos.
- Sí, pero desconocidos: el autor había consumido momentos antes de los hechos o durante los mismos, pero se desconoce el tipo de sustancia consumida.
- No.

38. Trastorno mental del autor (A9): indica si el autor tiene algún tipo de trastorno mental.
    - Sí: se conoce que el autor tiene algún trastorno mental.
    - No consta: no consta que el autor tenga un trastorno mental.
39. Actividades ilegales o registradas de la víctima (A10): indica si la víctima realizaba alguna actividad no registrada o ilegal que le sirviera de sustento (ejemplo: prostitución, tráfico de drogas, etc.).
    - Sí
    - No
40. Relación víctima - autor (R): indica la relación que mantenían víctima y autor en el momento de los hechos. La relación se indica des del punto de vista de la víctima.
    - Pareja/expareja: relación afectiva o no y matrimonial o no entre víctima y autor. Incluye:
        - Pareja: persona que mantiene relaciones afectivas con otra, con cierta estabilidad temporal, y mantiene expectativas de futuro. Se excluyen las relaciones puramente esporádicas.
        - Cónyuge: cada una de las personas que integran el matrimonio celebrado y reconocido oficialmente en España.
        - Expareja: la persona que ha cesado su relación afectiva mantenida con estabilidad, donde no existían vinculos regulados por la ley.
        - Separado/divorciado: el divorcio es la extinción del vínculo matrimonial y la separación es la suspensión de la convivencia conyugal declarados judicialmente en sentencia.
    - Conocido/vecino: víctima y autor se conocen, o viven cerca, en el mismo barrio o localidad.
    - Amistad/familia: la relación que mantiene la víctima con el autor es de carácter afectivo o están unidos por algún lazo familiar (se han tenido en cuenta como relaciones familiares las de compañeros de piso).
    - Otra: víctima y autor mantienen otro tipo de relación que no se encuentra especificada anteriormente. Incluyen los casos en los que la víctima mantenga una relación de trabajo

respecto el autor o que mantenga una relación en el ámbito educativo (de infantil a universitario).

- No: víctima y autor no mantienen ninguna relación.

C. Análisis de las frecuencias de las variables

Cuadro 23: Variables del hecho y frecuencias

Variable	Categorías	Frecuencia (N = 632)	Porcentaje	Porcentaje válido (sin NA)
H_1 = año del hecho	2010	216	34.2	34.2
	2011	220	34.8	34.8
	2012	196	31.0	31.0
H_2 = estación del año	Primavera	164	25.9	25.9
	Verano	180	28.5	28.5
	Otoño	138	21.9	21.9
	Invierno	150	23.7	23.7
H_3 = día de la semana	Lunes	73	11.6	11.6
	Martes	81	12.8	12.8
	Miércoles	97	15.3	15.3
	Jueves	88	13.9	13.9
	Viernes	98	15.5	15.5
	Sábado	92	14.6	14.6
	Domingo	103	16.3	16.3
H_4 = franja horaria	Madrugada	149	23.6	27.0
	Mañana	101	16	18.3
	Tarde	120	19	21.9
	Noche	181	28.6	32.8
	NA	81	12.8	
H_5 = núm. víctimas	Una	604	95.6	95.6
	Varias	28	4.4	4.4
H_6 = núm. autores	Uno	500	79.1	79.1
	Varios	132	20.9	20.9
H_7 = testigos	Sí	373	59	59
	No	259	41	41

continuación Cuadro 23: variables del hecho y frecuencias

Variable	Categorías	Frecuencia (N = 632)	Porcentaje	Porcentaje válido (sin NA)
H_8 = tipología del hecho	Discusión	138	21.8	22.7
	Violencia de género	130	20.6	21.3
	Violencia familiar	129	20.4	21.3
	Otras interpersonales	97	15.3	15.9
	Robo	49	7.8	8
	Actividades criminales	66	10.4	10.8
	NA	23	10.8	
H_9 = arma	Objeto contundente	62	9.8	10.3
	Blanca	286	45.3	47.4
	Fuego	94	14.9	15.6
	Fuerza agresor	75	11.9	12.4
	Asfixia	32	5.1	5.3
	Otros	54	8.5	8.9
	NA	29	4.6	
H_{10} = desplazamiento del arma	Sí	179	28.3	36.7
	No	234	37	48
	Cuerpo del agresor	75	11.9	15.4
	NA	144	22.8	
H_{11} = tipo de lugar	Interior	396	62.7	62.9
	Exterior	209	33.1	33.2
	Vehículo	21	3.3	3.3.
	Agua	4	0.6	0.6
	NA	2	0.3	
H_{12} = escenas	Única	561	88.8	89
	Multiescena	69	10.9	11
	NA	2	0.3	
H_{13} = movimiento del cuerpo	Desplazado y no oculto	30	4.9	4.7
	Oculto y no desplazado	14	2.2	2.3
	Desplazado y oculto	24	3.8	3.9
	Ni desplazado ni oculto	546	86.4	88.9
	NA	18	2.8	
H_{14} = método de huida	Pie	166	26.3	40.9
	Escena	130	20.6	32

continuación Cuadro 23: variables del hecho y frecuencias

Variable	Categorías	Frecuencia (N = 632)	Porcentaje	Porcentaje válido (sin NA)
	Vehículo	103	16.3	25.4
	Suicidio	7	1.1	1.7
	NA	226	35.8	

Fuente: Elaboración propia.

Cuadro 24: Variables de la víctima y frecuencias

Variable	Categorías	Frecuencia (N = 662)	Porcentaje	Porcentaje válido
V_1 = edad víctima	0	21	3.2	3.3
	Menores	35	5.3	5.5
	18-20	22	3.3	3.6
	21-30	122	18.4	19.3
	31-40	118	17.8	18.7
	41-50	116	17.5	18.4
	51-64	103	15.6	16.3
	65 o más	94	14.2	14. 9
	NA	31	4.7	
V_2 = sexo víctima	Hombre	406	61.3	61.4
	Mujer	255	38.5	38.6
	NA	1	0.2	
V_3 = origen víctima	España	464	70.1	71.9
	Resto de Europa	66	10	10.2
	Resto del mundo	115	17.4	17.8
	NA	17	2.6	
V_4 = situación laboral	Ocupado	121	18.3	47.6
víctima	Parado	27	4.1	10.6
	Estudiante	16	2.4	6.3
	Jubilado	25	3.8	9.8
	Otra	65	9.8	25.6
	NA	408	61.6	
V_5 = antecedentes	Homicidio	7	1.1	5.3
víctima	Personas	39	5.9	29.8

continuación Cuadro 24: variables de la víctima y frecuencias

Variable	Categorías	Frecuencia (N = 662)	Porcentaje	Porcentaje válido
	Sí pero desconocidos	30	4.5	22.9
	No	55	8.3	42
	NA	531	80.2	
V_6 = adicciones víctima	Alcohol	48	7.3	37.8
	Drogas	39	5.9	30.7
	Alcohol y drogas	25	3.8	19.7
	No	15	2.3	11.8
	NA	535	80.8	
V_7 = frecuencia de	Ocasional	15	2.3	11.8
consumo víctima	Habitual	78	11.8	61.4
	No	15	2.3	11.8
	Sí pero desconocida	19	2.9	15
	NA	535	80.8	
V_8 = sustancia consumida	Alcohol	80	12.1	64
de la víctima durante el	Drogas	17	2.6	13.6
hecho	Alcohol y drogas	10	1.5	8
	Sí pero desconocida	4	0.6	3.2
	No	14	2.1	3.2
	NA	537	11.2	
V_9 = actividades ilegales	Sí	68	10.3	10.3
víctima	No	594	89.7	89.7

Fuente: Elaboración propia.

Cuadro 25: Variables del autor y frecuencias

Variable	Categorías	Frecuencia (N = 871)	Porcentaje	Porcentaje válido
A_1 = edad autor	Menores	40	4.6	4.7
	18-20	64	7.3	7.5
	21-30	225	25.8	26.2
	31-40	235	27	27.4
	41-50	169	19.4	19.7
	51 o más	126	14.5	14.7

continuación Cuadro 25: variables del autor y frecuencias

Variable	Categorías	Frecuencia (N = 871)	Porcentaje	Porcentaje válido
	NA	12	1.4	
A_2 = sexo autor	Hombre	778	89.3	89.3
	Mujer	93	10.7	10.7
A_3 = origen autor	España	553	63.5	64.7
	Resto de Europa	131	15	15.3
	Resto del mundo	171	19.6	20
	NA	16	1.8	
A_4 = situación laboral	Ocupado	125	14.4	38.1
autor	Parado	76	8.7	23.2
	Estudiante	7	0.8	2.1
	Jubilado	11	1.3	3.4
	Otra	109	12.5	33.2
	NA	543	62.3	
A_5 = antecedentes autor	Homicidio	26	3	3.8
	Personas	255	29.3	37.6
	Sí pero desconocidos	166	19.1	24.4
	No	232	26.6	34.2
	NA	192	22	
A_6 = adicciones autor	Alcohol	55	6.3	24.7
	Drogas	90	10.3	40.4
	Alcohol y drogas	52	6	23.3
	No	26	3	11.7
	NA	648	74.4	
A_7 = frecuencia de	Ocasional	18	2.1	8.1
consumo autor	Habitual	130	14.9	58.3
	No	26	3	11.6
	Sí pero desconocida	49	5.6	22
	NA	648	74.4	
A_8 = sustancia consumida	Alcohol	111	12.7	55.2
del autor durante el hecho	Drogas	24	2.8	11.9
	Alcohol y drogas	32	3.7	16

continuación Cuadro 25: variables del autor y frecuencias

Variable	Categorías	Frecuencia (N = 871)	Porcentaje	Porcentaje válido
	Sí pero desconocida	9	1	4.5
	No	25	2.9	12.4
	NA	670	76.9	
A_9 = trastorno mental autor	Sí	85	9.8	9.8
	No consta	786	90.2	90.2
A_{10} = actividades ilegales autor	Sí	110	12.6	12.6
	No	761	87.4	87.4

Fuente: Elaboración propia.

Cuadro 26: Variable relación víctima - autor y frecuencias

Variable	Categorías	Frecuencia (N = 909)	Porcentaje	Porcentaje válido
R = relación víctima - autor	Pareja o expareja	154	16.9	18.2
	Conocido o vecino	240	26.4	28.3
	Amistad o familia	194	21.3	22.9
	Otra	85	9.4	10
	No	175	19.3	20.6
	NA	61	6.7	

Fuente: Elaboración propia.

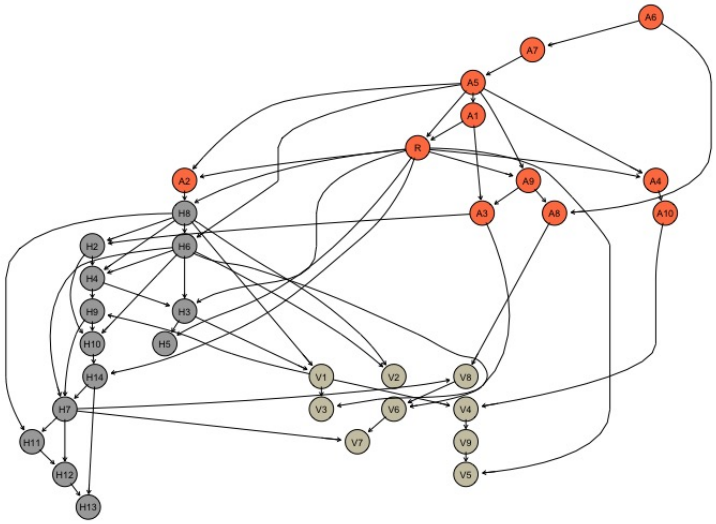
D. Estructura del DAG para el orden ancestral del procedimiento *Chain Classifier* usando la técnica *k-fold cross validation*

A continuación se muestran las estructuras de la RB cuando $k = 1$ en la técnica *k-fold cross validation* para hallar el orden ancestral, que será el orden del clasificador en cadena. Hay cuatro gráficos, ya que hace referencia a los cuatro modelos diferentes que existen (explicados en la Sección 9). Se distingue en diferentes colores según si las variables son del hecho (en gris), del autor (en rojo) o de la víctima (gris más claro).

El orden ancestral varía en los cuatro modelos (cuando $k = 1$). Recordar que el orden ancestral se obtiene a partir de la RB libre ajustada, la cual produce un DAG. A continuación se detalla el orden ancestral de cada modelo y se indica la figura donde se observa la estructura de la RB (el DAG):

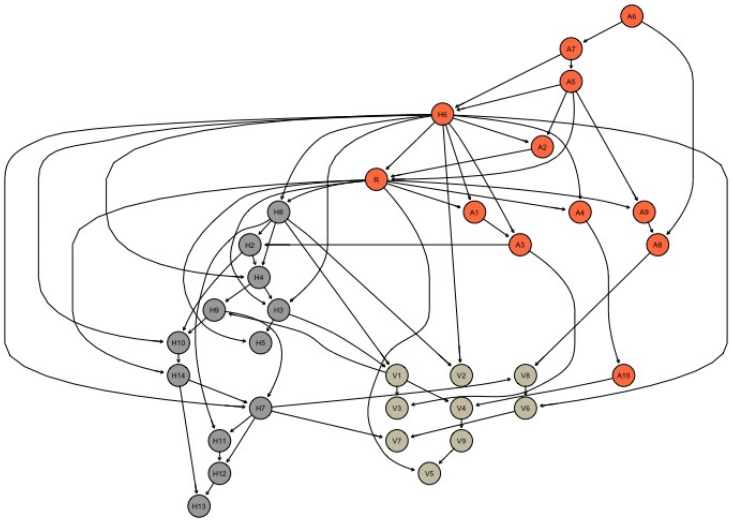
- Modelo 1 (Fig. 16): A_6, A_7, A_5, A_1, R, A_2, A_4, A_9, A_3, A_8 y A_{10}.
- Modelo 2 (Fig. 17): A_6, A_7, A_5, A_6, A_2, R A_1, A_4, A_9, A_3, A_8 y A_{10}.
- Modelo 3 (Fig. 18): H_8, A_1, A_4, A_9, R, A_3, A_{10}, A_5, A_2, A_6, A_7 y A_8.
- Modelo 4 (Fig. 19): A_6, A_7, A_5, H_6, H_8, A_1, A_2, A_4, A_9, R, A_3, A_8 y A_{10}.

Figura 16: DAG del orden ancestral para el modelo 1 cuando $k = 1$



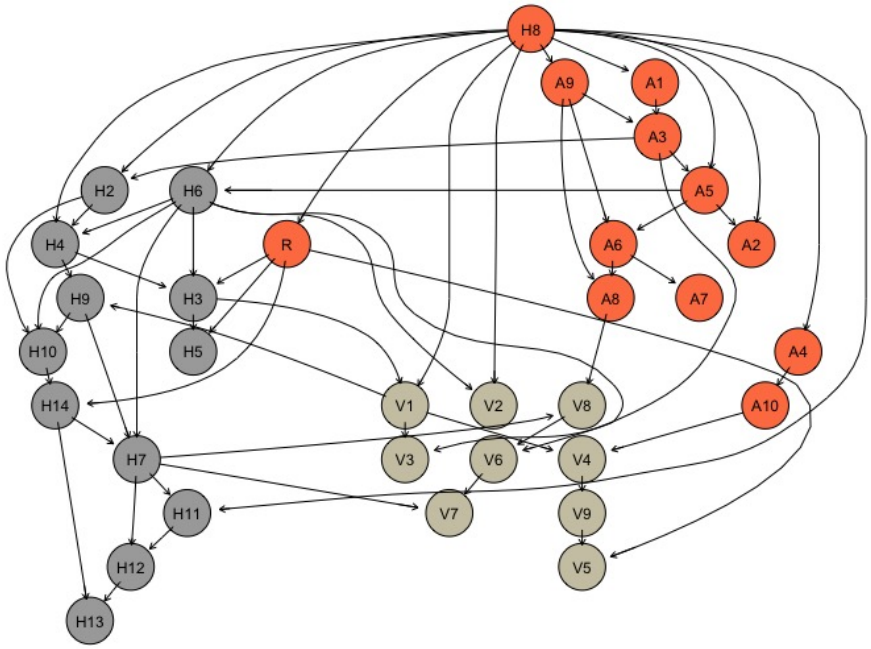
Fuente: Elaboración propia.

Figura 17: DAG del orden ancestral para el modelo 2 cuando $k = 1$



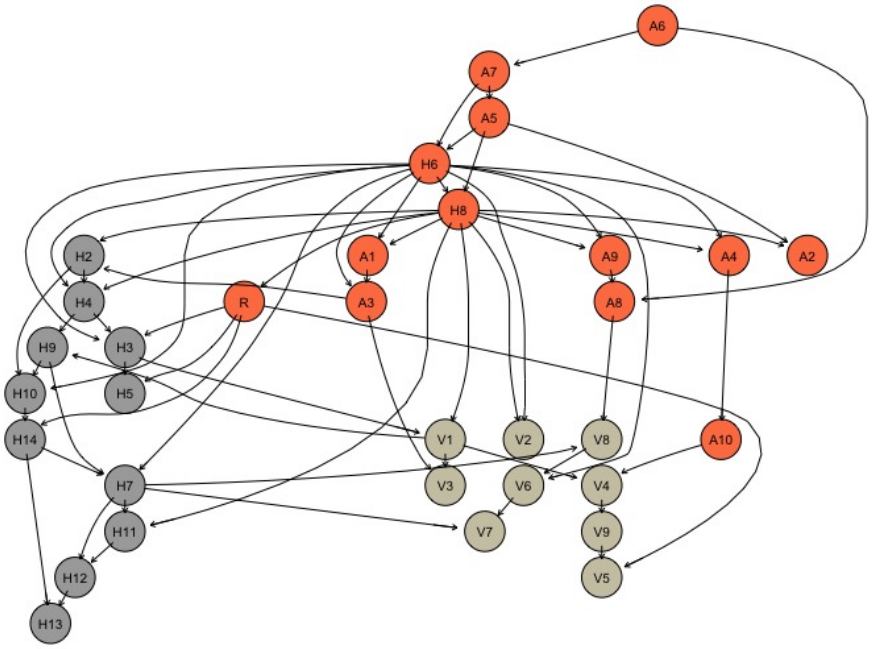
Fuente: Elaboración propia.

Figura 18: DAG del orden ancestral para el modelo 3 cuando $k = 1$



Fuente: Elaboración propia.

Figura 19: DAG del orden ancestral para el modelo 4 cuando $k = 1$



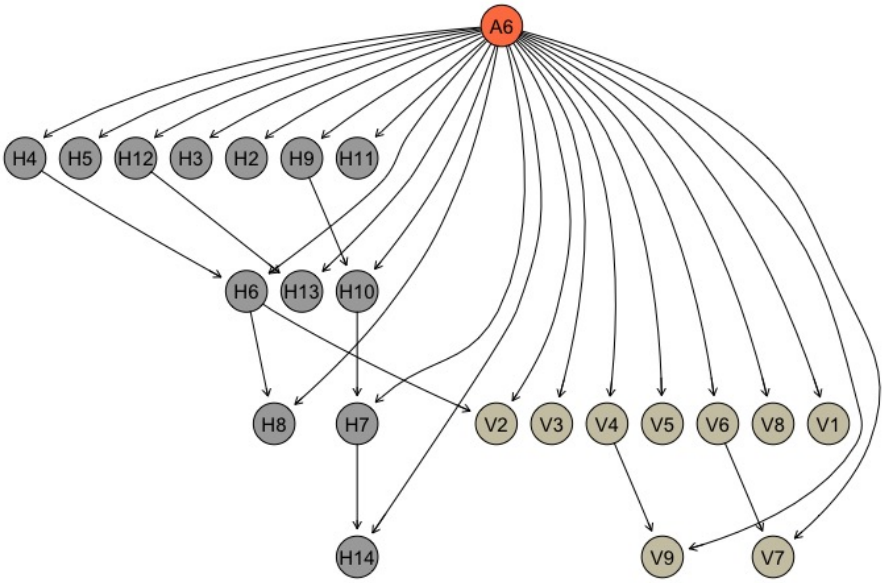
Fuente: Elaboración propia.

E. Estructura del DAG para las variables *output*

A continuación se muestran las estructuras del DAG de las dos primeras variables del orden ancestral obtenidas con el procedimiento *Chain Classifier* para el modelo 1 y cuando $k = 1$.

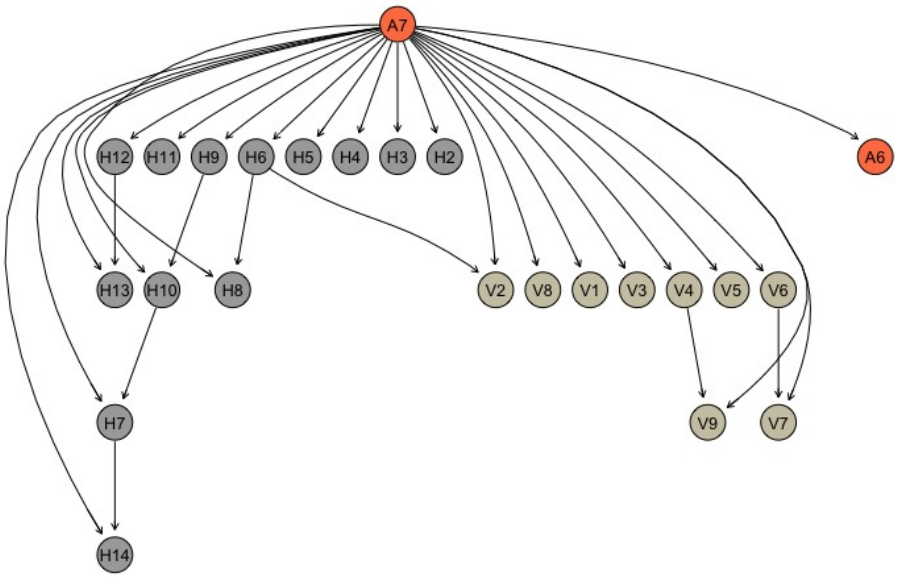
La Figura 20 representa la estructura *Augmented Naive Bayes* para la variable A_6, que corresponde a las adicciones del autor, y es la primera variable del orden ancestral en el modelo 1. La Figura 21 representa la estructura, también de tipo *Augmented Naive Bayes*, para la segunda variable del orden ancestral en el modelo 1, A_7, que corresponde a la frecuencia de consumo del autor. En este caso se observa como la variable anterior en el orden (la A_6) se trata como una variable *input* más, tal como se ha explicado en la Sección 8.2.

Figura 20: DAG de la varable A_6 = adicciones autor en el modelo 1 y con el procedimiento *Chain Classifier*



Fuente: Elaboración propia.

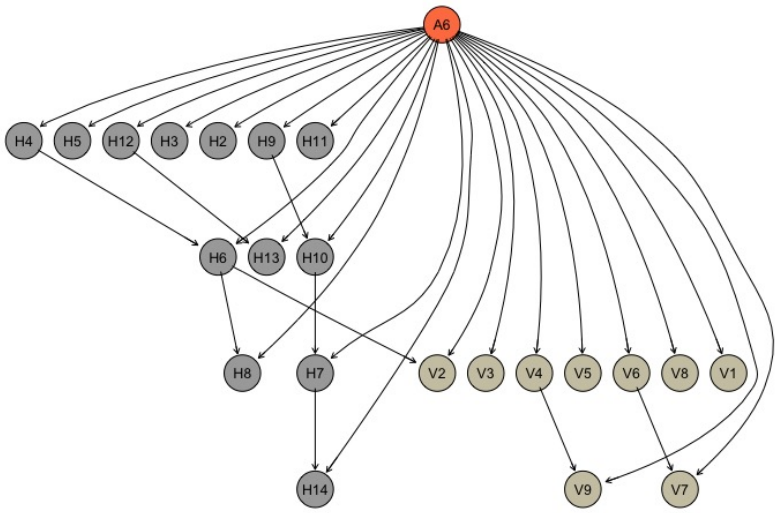
Figura 21: DAG de la varable A_7 = frecuencia de consumo del autor en el modelo 1 y con el procedimiento *Chain Classifier* y $k = 1$



Fuente: Elaboración propia.

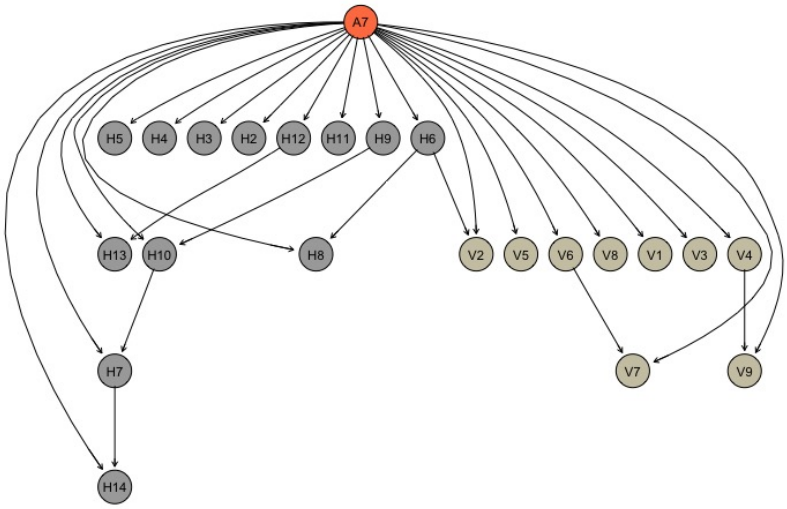
A continuación se muestran las estructuras del DAG de las variables A_6 y A_7 obtenidas a partir del procedimiento *Binary Relevance*. En este caso se quiere destacar el DAG de la variable A_7 (ver Figura 23), porque en este caso se construye sin tener en cuenta las otras variables que se quieren predecir y, es por eso, que en ningún caso las otras variables *output* forman parte del DAG.

Figura 22: DAG de la varable A_6 = adicciones autor en el modelo 1 y con el procedimiento *Binary Relevance*



Fuente: Elaboración propia.

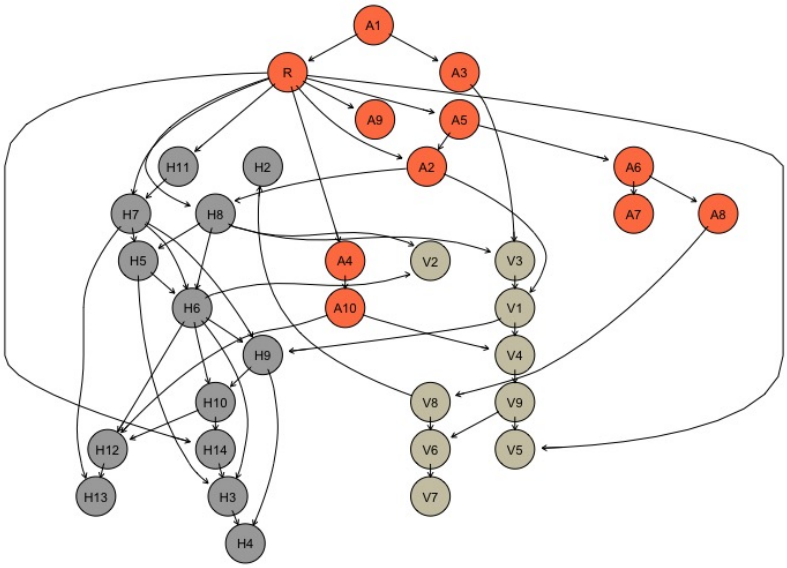
Figura 23: DAG de la varable A_7 = frecuencia de consumo del autor en el modelo 1 y con el procedimiento *Binary Relevance Classifier* y $k = 1$



Fuente: Elaboración propia.

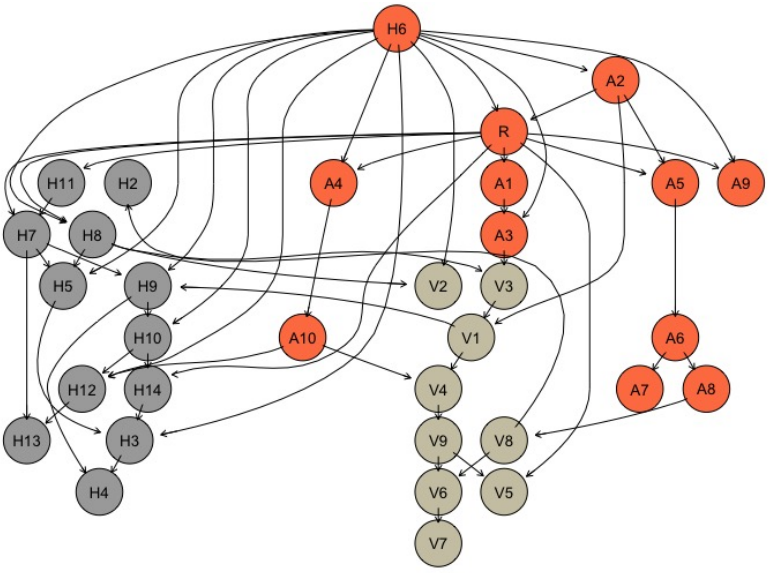
F. Estructura del DAG para el orden ancestral del Shiny

Figura 24: DAG del orden ancestral del modelo 1 usado en el Shiny



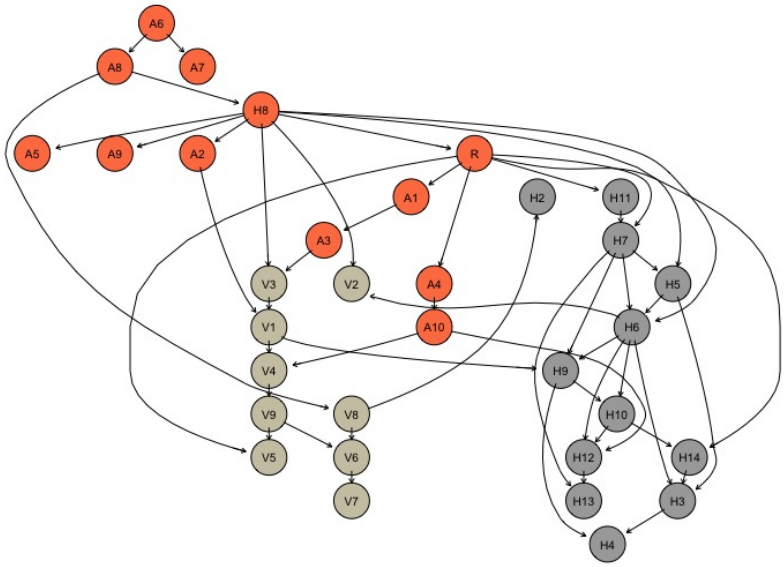
Fuente: Elaboración propia.

Figura 25: DAG del orden ancestral del modelo 2 usado en el Shiny



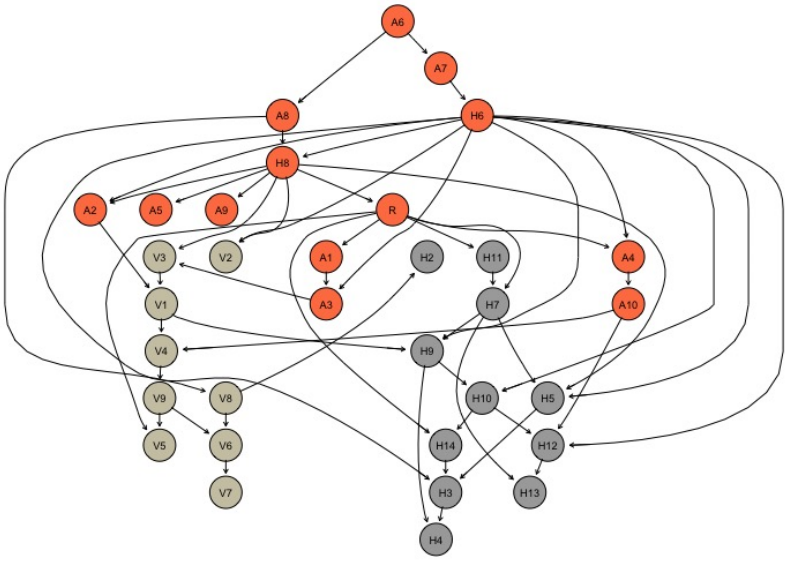
Fuente: Elaboración propia.

Figura 26: DAG del orden ancestral del modelo 3 usado en el Shiny



Fuente: Elaboración propia.

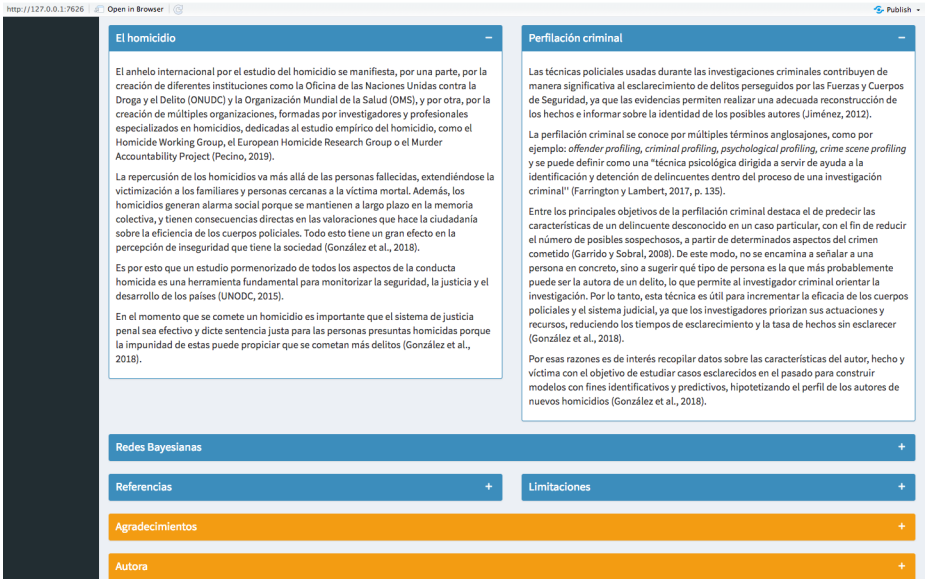
Figura 27: DAG del orden ancestral del modelo 4 usado en el Shiny



Fuente: Elaboración propia.

G. Aplicación Shiny

Figura 28: Introducción: el homicidio y la perfilación criminal

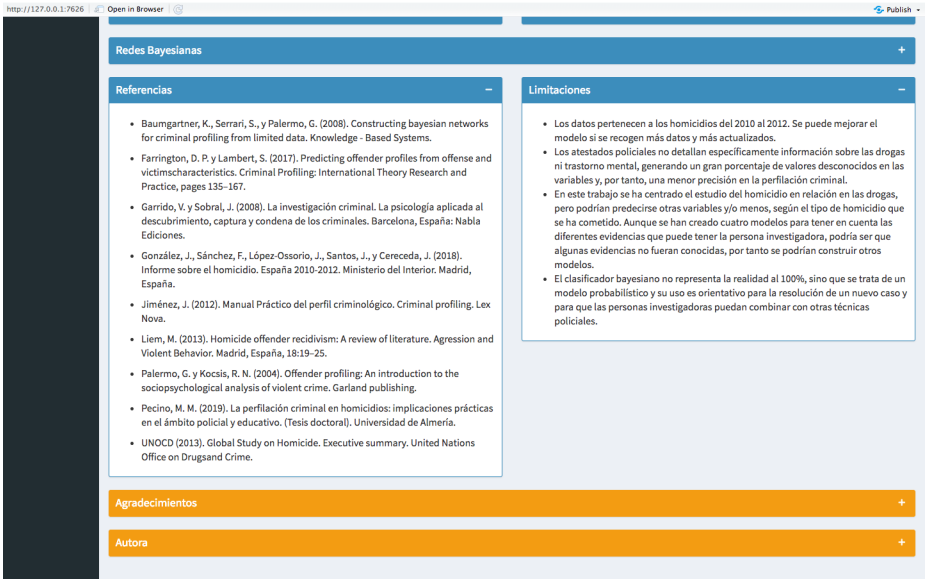


Fuente: Captura de pantalla de la aplicación Shiny creada por la autora.

Figura 29: Introducción: redes bayesianas

Fuente: Captura de pantalla de la aplicación Shiny creada por la autora.

Figura 30: Introducción: referencias y limitaciones

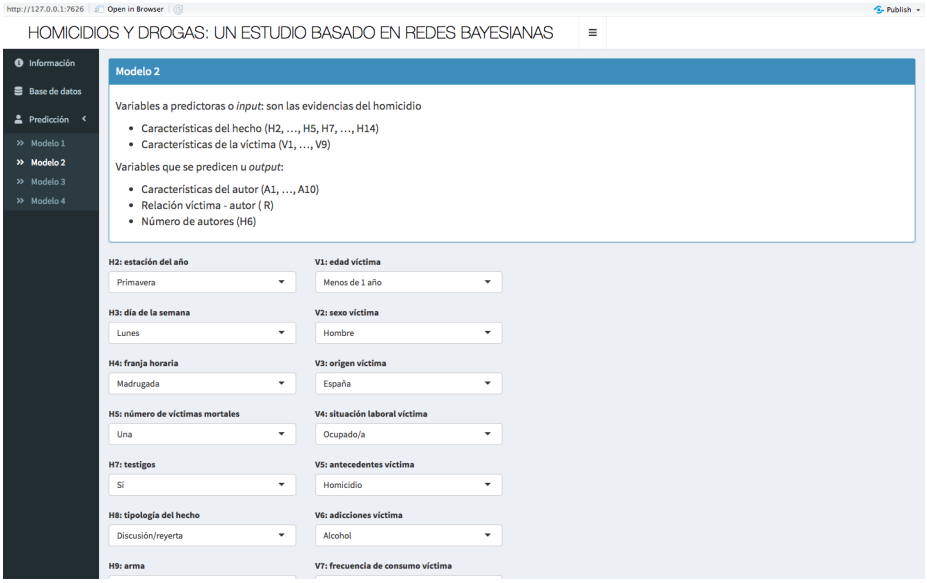


Fuente: Captura de pantalla de la aplicación Shiny creada por la autora.

Figura 31: Introducción: agradecimientos y autora

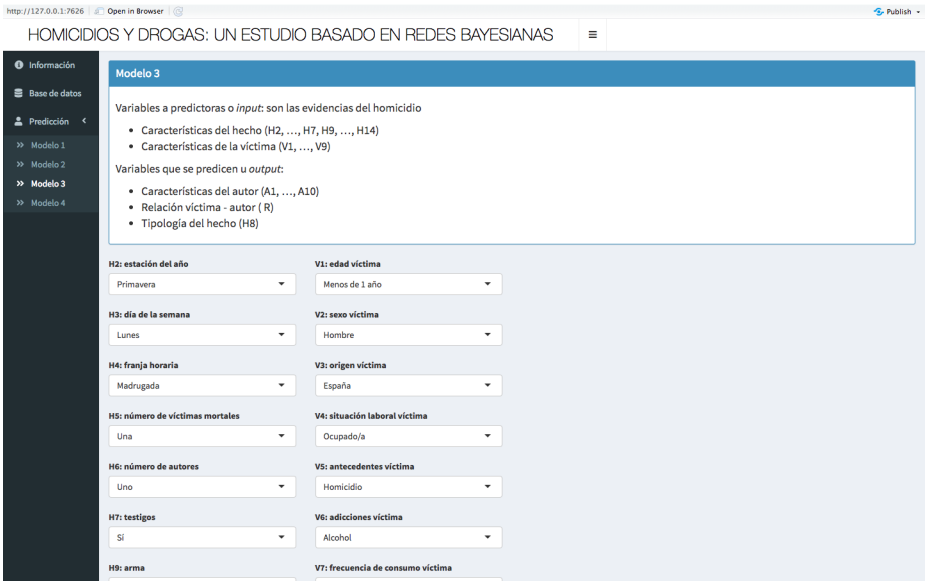
Fuente: Captura de pantalla de la aplicación Shiny creada por la autora.

Figura 32: Subapartado Modelo 2



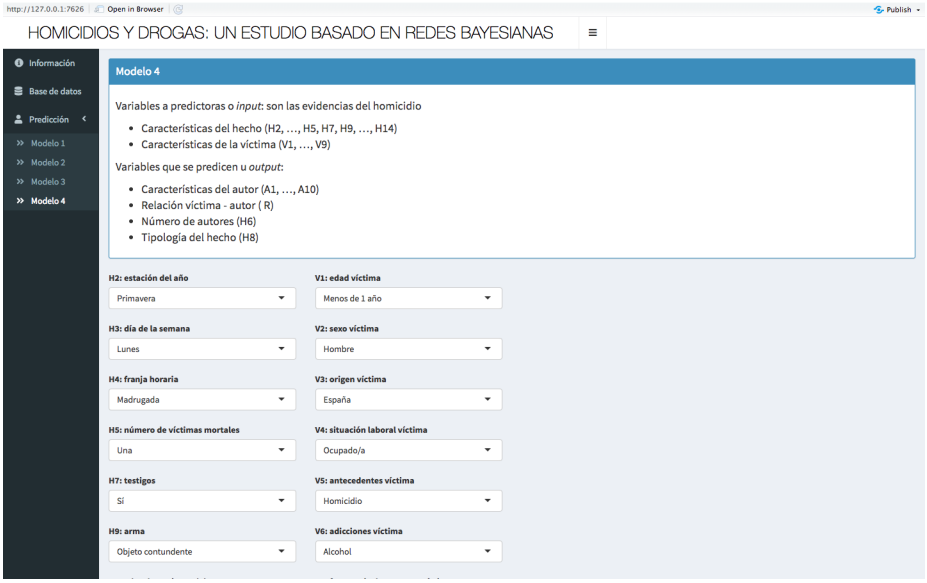
Fuente: Captura de pantalla de la aplicación Shiny creada por la autora.

Figura 33: Subapartado Modelo 3



Fuente: Captura de pantalla de la aplicación Shiny creada por la autora.

Figura 34: Subapartado Modelo 4

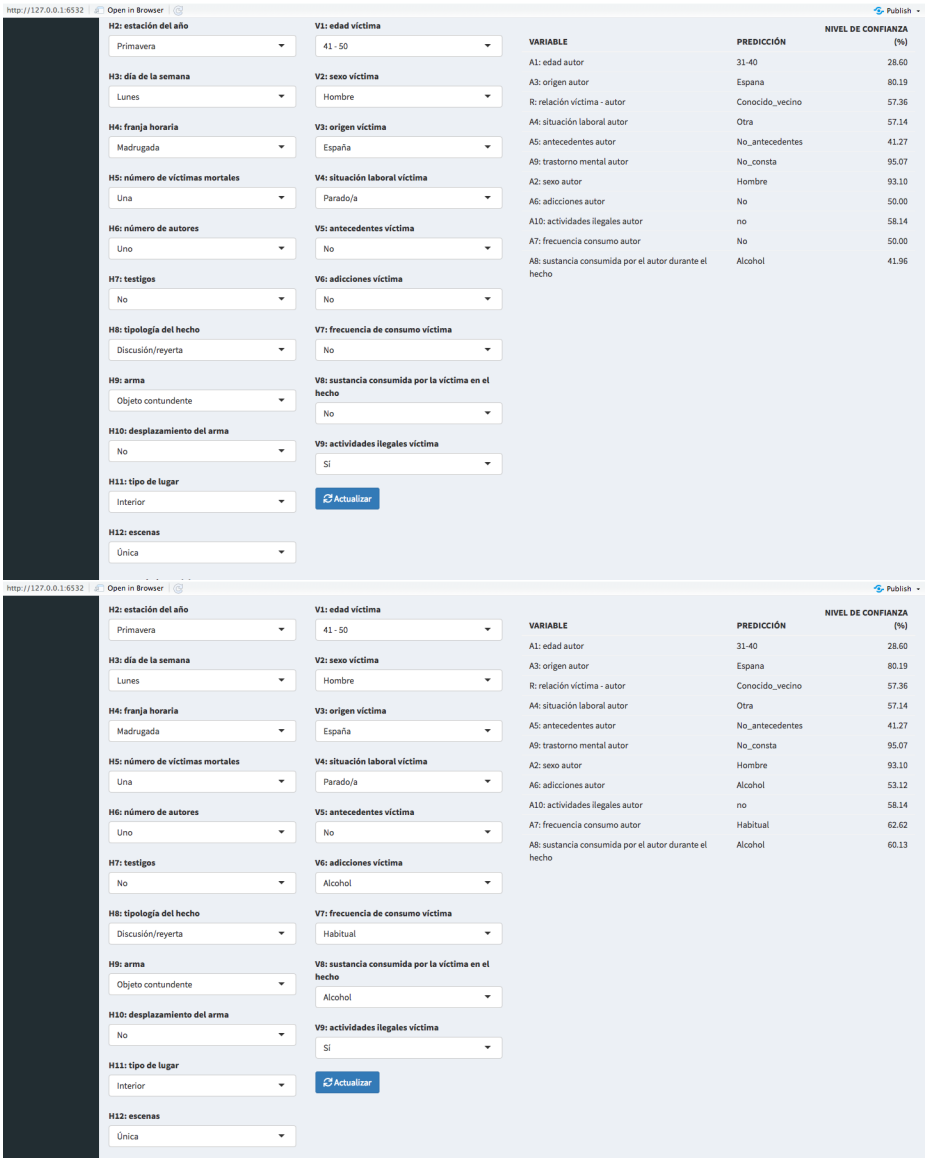


Fuente: Captura de pantalla de la aplicación Shiny creada por la autora.

H. Ejemplo de la predicción del perfil del homicidia

En este anexo se muestran dos ejemplos de perfilación criminal a partir del modelo 1.

Figura 35: Ejemplo de perfilación criminal a partir del modelo 1. En la imagen de arriba la víctima no presenta relación con las drogas, en la de abajo sí.



Fuente: Captura de pantalla de la aplicación Shiny creada por la autora.