Universidad Católica Argentina
Facultad de Ingeniería y Ciencias Agrarias

Homicidios dolosos en Argentina: Predicción de vínculo víctima-victimario

Trabajo Final de Ingeniería Industrial

Francisco Gallegos Luque - Ignacio Félix Massun

Tutor: Emilio Osvaldo Picasso
Cotutor: Emilio Pugnaloni
Buenos Aires, Junio de 2025

Índice

1. Resumen

El presente trabajo tiene como objetivos principales analizar y desarrollar un modelo de predicción para determinar el vínculo entre la víctima y el victimario en homicidios dolosos ocurridos en Argentina durante los años 2017 a 2023, a partir de una base de datos del Ministerio de Seguridad. Se busca identificar patrones en las variables disponibles que permitan inferir si el vínculo es familiar, no familiar, o si no hay vínculo, empleando modelos estadísticos de clasificación, con énfasis en técnicas accesibles para su implementación en contextos judiciales o criminológicos. Se desarrollan dos modelos: uno explicativo (regresión logística multinomial) y otro predictivo (gradient boosting) para identificar y anticipar relaciones entre los protagonistas del crimen (relación familiar, no familiar, o sin relación)

El trabajo se estructura en varias etapas. En primer lugar, se realiza una exploración y limpieza de los datos, dado que la base original presenta valores faltantes, inconsistencias y una estructura que requiere preprocesamiento para su análisis. Posteriormente, se seleccionan y transforman variables clave que pudieran tener relación con el tipo de vínculo entre víctima y victimario. Entre estas se encuentran el sexo y la edad de la víctima, el día del hecho, el uso de armas, entre otras.

A continuación, se procede a una etapa de análisis descriptivo, donde se identifican diferencias entre los tres tipo de vínculos. Se observa que en los homicidios con vínculo familiar hay mayor presencia de víctimas mujeres, suelen ocurrir en el hogar y no se asocian con un delito. También se evidencia que el uso de armas de fuego es más frecuente en homicidios a manos de desconocidos, por ejemplo.

En cuanto a la modelización, se analizaron distintos algoritmos de clasificación, evaluando su desempeño mediante técnicas de validación cruzada y métricas como el acierto general, el F_1-score y el índice Kappa de Cohen. Uno de los principales hallazgos es que ciertas variables tienen un peso considerable en la predicción del vínculo. Por ejemplo, el sexo de la víctima, el lugar del hecho y la presencia o no de arma de fuego resultan ser factores determinantes. A partir de esto, se construye un modelo final que logra una precisión aceptable (acierto general: 65,8 %, F_1: 65 %) y que podría ser implementado como herramienta de apoyo para fiscales, jueces o analistas en el abordaje inicial de un caso de homicidio, facilitando una mejor comprensión del entorno del hecho.

El trabajo concluye destacando la importancia de aplicar herramientas estadísticas en el análisis de fenómenos criminales, no como sustituto del juicio humano, sino como complemento que permita guiar líneas de investigación con mayor eficiencia. Asimismo, se reconocen las limitaciones del estudio, tales como la calidad de los datos, la ausencia de información sobre contexto socioeconómico o histórico de violencia previa, y la naturaleza inherentemente compleja de los vínculos humanos.

Finalmente, se sugiere continuar esta línea de investigación incorporando técnicas más avanzadas, además de promover la mejora en la recolección y sistematización de datos por parte de los organismos estatales. Esto permitiría no solo mejorar la predicción, sino también comprender mejor los factores estructurales detrás de la violencia en Argentina.

2. Introducción

La Real Academia Española define homicidio doloso como la "causación de la muerte de otra persona de forma consciente y voluntaria" (Diccionario panhispánico del español jurídico, 2025). Entre los crímenes que Argentina y América Latina en general sufren, los homicidios representan uno de los peores (sino el peor) y más perturbadores para la sociedad.

Argentina presenta indicadores tanto alarmantes como esperanzadores en cuanto a homicidios respecta cuando se los compara con el resto de los países latinoamericanos. Las estadísticas de homicidios en Argentina y en el mundo se presentan en la Tabla 1.

Zonas Geográficas	**Región**	**Homicidios totales**		**Homicidios cada 100 mil personas**		**Variación (%)**
		2000	2018	2000	2018	
África		108100	158200	13.3	12.2	-8.3
América	EEUU - Canadá	16100	17000	5.1	4.6	-9.8
	Latinoam. y Caribe	114300	151900	21.0	23.9	13.7
Asia		125000	98200	3.4	2.2	-35.3
Australia y NZ		400	300	1.8	1.0	-43.7
Europa		56500	20500	7.7	2.7	-64.9
Argentina		3129	2387	8.4	5.4	**-35.6**

Tabla 1: Fuente: Adaptado de Bergman et al. (2023) con base a UNODC (2022).

América Latina tiene la tasa de homicidios más alta del mundo. En el año 2018, hubo 23,9 homicidios cada 100 mil personas: casi el doble que en África y más de 8 veces superior a la de Europa. Además, entre los años 2000 y 2018, Latinoamérica y el Caribe fueron las únicas regiones que aumentaron esta tasa, alcanzando un aumento del 13,7 %. Con solo el 8 % de la población mundial, el 33 % de los homicidios globales ocurren en la región: una tasa de homicidios cada 100.000 habitantes 20 veces superior a la de Reino Unido (Jaitman & Ajzenman, 2016).

En cuanto a la Argentina en 2018, la cantidad de homicidios cada 100 mil personas fue de 5,4, más de 4 veces menor a la media de la región (23,9 homicidios cada 100 mil personas). Además, la variación de esta tasa entre 2000 y 2018 se vio disminuida en un 35,6 %, una evolución opuesta a la de Latinoamérica y el Caribe. Siguiendo esta tendencia favorable, Argentina presentó en 2024 una tasa de 3,8 homicidios por cada 100.00 habitantes, muy inferior a la de la región que alcanza los 20,2 homicidios por cada 100.000 habitantes (Manjarrés et al., 2025). Para el país, según el Sistema Nacional de Información Criminal (2025), esta tasa es la más baja en los últimos 25 años. A pesar de los datos esperanzadores, Argentina presenta una tasa de homicidios considerablemente superior a la de países como España (0,69 homicidios cada 100.000 habitantes en 2023) o Australia (0,8 homicidios cada 100.000 habitantes en 2023)(UNODC, 2025).

Un homicidio doloso es procesado por el sistema penal, que se desarrolla en distintas etapas. Al momento de un homicidio, primero acuden generalmente las fuerzas policiales, quienes informan a la fiscalía sobre el hecho. En último lugar, se incorpora la instancia judicial (juez/jueza) quien pasa a ser responsable de la causa (Bergman et al., 2021). Como cabe esperar, una de las herramientas más efectivas para disminuir la cantidad de homicidios dolosos es un efectivo uso de la justicia penal. Por lo tanto, es importante la relación entre el número de homicidios y la cantidad de condenas firmes. Esta relación conlleva una dinámica dual en donde una medida afecta directamente a la otra. Por un lado, debido a un efecto de disuasión, según Patternoster (2010) y Nagin (1998), el número de homicidios se verá disminuido si la tasa de condenas firmes aumenta, ya que una mayor probabilidad de ser detenido desalienta a las personas a cometer este delito. Por otro lado, el aumento en la cantidad de homicidios dolosos en un país reduce el número de

condenas firmes, puesto que los limitados recursos del sistema penal imposibilitan el estudio en profundidad para cada caso. En el período 2002-2021, por cada 100 víctimas de homicidio hubo 65 condenas firmes (Bergman et al., 2021). En 2022, según datos del Registro Nacional de Reincidencia, un homicidio demoró en promedio cuatro años y cinco meses en proferir una sentencia.

En este contexto, surge la necesidad de aumentar el número de herramientas para detectar y condenar efectivamente a los responsables de homicidios dolosos. En el último tiempo, se ha incorporado el uso de Inteligencia Artificial y modelos descriptivos y predictivos a partir de grandes volúmenes de datos para pronosticar y formular potenciales resultados (Rigano, 2019). En particular, en la justicia penal, este trabajo depende principalmente de las fuerzas policiales, peritos y fiscales que deben obtener experiencia durante muchos años, lo cual conlleva mucho tiempo y está sujeto a objetividad y error (Rigano, 2019). En consecuencia, la implementación de estas nuevas herramientas podría potencialmente ayudar a obtener un mejor entendimiento de la realidad delictiva para así lograr disminuirla.

Particularmente, en materia de análisis de datos en el proceso penal, se han realizado múltiples estudios que abordan la relación que existe entre las características de un hecho delictivo, como un homicidio doloso, y su presunto autor. En estos estudios se sugiere que distintas variables sociodemográficas y variables situacionales del homicidio juegan distintos roles en la probabilidad de que la relación entre la víctima y el victimario sea de algún tipo en específico (Cao et al., 2007). Esto quiere decir que el vínculo entre los protagonistas de un homicidio doloso puede determinar las características del hecho, ya sea desde el lugar y la hora del asesinato hasta el método utilizado para matar, entre otros. Este tipo de análisis permite orientar la investigación en situaciones donde no se cuenta con la información necesaria para realizar una sentencia definitiva. Como plantea Heymann (1985), identificar posibles indicadores de la relación víctima-victimario puede ser útil para facilitar las investigaciones criminales, al reducir el número de posibles sospechosos. No solo se reduciría el tiempo del proceso judicial, sino que también se evitan subjetividades y falsas interpretaciones. Aunque cada caso es particular y conlleva un análisis en profundidad, este enfoque permite obtener ciertos indicios y herramientas para una mejor resolución del caso.

El siguiente estudio busca aplicar lo investigado sobre vínculos víctima-victimario en el análisis de homicidios dolosos al caso particular de la Argentina, mediante el desarrollo de un modelo con carácter explicativo que permita identificar relaciones significativas entre variables, y un modelo predictivo orientado a clasificar dicho vínculo en función de las características del homicidio.

El presente trabajo se organiza en varias secciones. A continuación, se presenta un resumen de antecedentes de estudio similares en Argentina y el resto del mundo. En la sección 2 se formalizan los objetivos. En la sección 3 se describe la metodología respecto a la base de datos empleada, el análisis exploratorio de la misma, el modelado explicativo y el modelado predictivo. En la sección 4, se presentan los resultados obtenidos y sus conclusiones se desarrollan en la sección 5. Por último, se detalla la bibliografía utilizada en la sección 6 y se anexan los archivos de código en lenguaje R del tratamiento de los datos y los modelos empleados en la sección 7.

Antecedentes en Argentina y el resto del mundo

El vínculo víctima-victimario está sujeto a distintas interpretaciones en cuanto a su clasificación. Por un lado, basado en estudios previos, Decker (1993) propuso una clasificación de acuerdo con el grado de cercanía entre víctima y victimario, discriminando en cinco grupos: desconocidos, conocidos, amigos, vínculos románticos y parientes no románticos. No obstante, el mismo Decker cuestionó su clasificación alegando que resulta muy difícil separar entre conocidos y amigos, y que la clasificación puede resultar muy arbitraria. Riedel (1987), en cambio, propone la clasifi-

cación que adoptará nuestro estudio: desconocidos, vínculos no familiares (amigos, conocidos, empleadores, etc.) y vínculos familiares/románticos.

Existen trabajos similares realizados en otras partes del mundo. Destaca el estudio de Cao et al. (2008), que utiliza una clasificación de los homicidios basada en la relación víctima–victimario mediante una regresión logística multinomial, para identificar de manera más precisa los factores asociados a tres tipos diferentes de relación en homicidios. Los tres tipos de relación son los propuestos por Riedel previamente mencionados. El estudio se limita a los homicidios registrados en Taiwán entre 1994 y 1998. Presenta un análisis descriptivo de los factores en homicidios, pero no detalla un modelo predictivo.

En líneas similares, el trabajo de Kim, Chopin y Beauregard (2024) se centra en el uso de características victimológicas y de la escena del crimen para predecir la relación víctima-victimario en homicidios sexuales. Se acota a homicidios ocurridos en Francia y Canadá entre 1948 y 2018. A partir de los datos obtenidos, el estudio utiliza análisis bivariado, regresión logística binaria secuencial y un modelo de red neuronal artificial para examinar las correlaciones y los valores predictivos de diversos factores para determinar la naturaleza de la relación víctima-victimario. Utiliza un total de 23 variables dicotómicas vinculadas con la escena y el contexto del crimen clasificadas según: actividad y lugar en el que estaba involucrada la víctima, comportamientos sexuales y comportamiento post-mortem del victimario. A pesar de estar acotado al ámbito de los homicidios sexuales, propone una metodología concreta para entender mejores patrones de comportamiento y relación en este tipo de crímenes.

En un tenor más descriptivo, Drawdy y Myers (2004) examinan las correlaciones entre la relación víctima/victimario y tres variables: motivo, elección de arma y número de lesiones infligidas. La muestra consistió en 57 casos de homicidios intencionales en Florida entre los años 1992 y 1996. Al igual que Wolfgang (1958), las relaciones se dividieron en categorías primarias (familiares y amigos) y secundarias (conocidos y desconocidos).

En cuanto a estudios basados en la situación criminalística en Argentina, se logra ahondar en el análisis de la geolocalización de los homicidios en distintas partes del país. Por un lado, Ramos et al. (2022) presentan una investigación cualitativa en la ciudad de Córdoba en donde refuerzan la idea de que la desorganización social en una zona urbana por problemáticas socioeconómicas incrementa la violencia interpersonal, la cual puede desencadenar, en situaciones extremas, en homicidio. Sin embargo, con respecto a los homicidios realizados dentro del hogar, la predicción de los mismos resulta más compleja ya que su motivación podría tener un entramado de características contextuales, sociales, psicológicas y de género. Por otro lado, se han realizado estudios descriptivos de la situación criminalística en la Ciudad Autónoma de Buenos Aires, identificando las áreas con mayor probabilidad de ocurrencia de los delitos en cuestión. No obstante, todavía no se ha estudiado la posible relación entre las características del homicidio y el vínculo entre la víctima y el victimario en base a los datos estadísticos de Argentina.

América Latina y Argentina en particular no cuentan con un destacado historial en materia de ciencia de datos y estadística para la resolución de problemas corrientes, y sí cuentan con un triste historial de tasas de homicidios extremadamente altas. Los estudios descriptivos y predictivos en materia de homicidios, si bien son cada vez más frecuentes, no parecen estar a la altura de la situación. A nivel mundial, a pesar de haber muchos antecedentes, priman los estudios de carácter descriptivo por sobre los predictivos. En particular, ningún estudio previo ha intentado hacer un análisis descriptivo y un modelo predictivo del vínculo víctima-victimario con la información disponible de homicidios dolosos en Argentina recientemente.

3. Objetivos

El objetivo general de este trabajo es profundizar en los análisis de los homicidios dolosos en Argentina para aportar herramientas a los agentes responsables del proceso penal. Específicamente, definiremos el objetivo del siguiente estudio en dos grandes categorías:

1. Describir y demostrar la existencia de patrones de correlación entre factores del homicidio y el vínculo víctima-victimario en homicidios dolosos en Argentina
2. Desarrollar un modelo capaz de predecir con cierto grado de significancia el vínculo víctima-victimario en homicidios dolosos en Argentina a partir de factores propios del homicidio.

4. Metodología

Con el fin de cumplir con los objetivos propuestos, se partió de una base de datos para realizar un análisis exploratorio de los homicidios del país, un modelo explicativo y otro predictivo. Debido a la complejidad de la recolección de información en el proceso penal, la fuente de datos criminalísticos en Argentina es incompleta y desorganizada. En consecuencia, previo al análisis de datos, fue conveniente realizar una propia preparación de los mismos.

Luego de adaptar y hacer un análisis exploratorio de la base de datos, se desarrolló un modelo a partir de una regresión logística multinomial para realizar un análisis explicativo de los homicidios en el país. También, se realizó un modelo predictivo en base a algoritmos de *gradient boosting* y se lo evaluó para corroborar su nivel de concordancia con la realidad.

A continuación se desarrollan en profundidad cada uno de los pasos realizados para el cumplimiento de los objetivos propuestos.

4.1. Base de datos

4.1.1. Origen

Según los datos más recientes provistos por el Ministerio de Seguridad de la Nación, Dirección Nacional de Estadística Criminal (2024), los homicidios dolosos cometidos a una persona en Argentina mostraron una leve disminución a los largo de los años. En 2023 se registraron 1 699 casos, casi un 18 % menos a los 2 071 casos ocurridos en 2017. Entre estos años los homicidios dolosos con una sola víctima y victimario alcanzaron 2 122 casos en 2018, 1 909 en 2019, 1 946 en 2020, 1 685 en 2021 y 1 579 en 2022.

Durante estos 7 años de registro, fueron documentadas 14 999 observaciones con 46 variables distintas. Entre estas, se incluyen homicidios dolosos de una o múltiples víctimas y de uno o múltiples victimarios. A continuación se detallan las variables de la base de datos:

1. id_hecho
2. tipo_hecho_segun_victima
3. tipo_persona_id
4. tipo_persona
5. cant_inc
6. cant_vic
7. federal
8. provincia_id
9. provincia_nombre
10. departamento_id
11. departamento_nombre
12. localidad_id
13. localidad_nombre
14. radio_censal
15. latitud_radio
16. longitud_radio
17. anio
18. mes
19. fecha_hecho
20. hora_hecho
21. tipo_lugar
22. tipo_lugar_otro
23. tipo_lugar_ampliado
24. clase_arma
25. clase_arma_otro
26. en_ocasion_otro_delito
27. en_ocasion_otro_delito_otro
28. motivo_origen_registro
29. motivo_origen_registro_otro
30. victima_sexo
31. victima_identidad_genero
32. victima_identidad_genero_otro
33. victima_tr_edad
34. victima_18_años_o_mas
35. victima_clase
36. victima_clase_otro
37. victima_situacion_ocupacional
38. victima_situacion_ocupacional_otro
39. **victima_relacion_inculpado**
40. inculpado_sexo
41. inculpado_identidad_genero
42. inculpado_identidad_genero_otro
43. inculpado_tr_edad
44. inculpado_clase
45. inculpado_otro_clase
46. inculpado_relacion_victima

4.1.2. Procesamiento

Sin embargo, no todos estos datos fueron utilizados para el desarrollo del modelo. En primer lugar, se consideraron exclusivamente los homicidios dolosos en los cuales hubo una sola víctima y un solo inculpado, es decir, se descartaron aquellos casos en los cuales hubo más de una víctima y/o inculpado.

En segundo lugar, se seleccionaron aquellas variables con mayor relevancia para los objetivos planteados y se descartaron las variables con más de un 40 % de observaciones faltantes.
Las variables remanentes resultaron:

1. localidad_nombre
2. mes
3. fecha_hecho
4. tipo_lugar
5. clase_arma
6. en_ocasion_otro_delito
7. motivo_origen_registro
8. victima_identidad_genero
9. victima_tr_edad
10. victima_clase
11. **victima_relacion_inculpado**

Las 11 variables resultantes fueron adaptadas y renombradas de acuerdo a los requerimientos del modelo en 11 variables explicativas y una variable respuesta:

- Primera variable explicativa

Nombre de la variable	poblacion
Descripción	Cantidad de habitantes en la localidad del hecho
Tipo de variable	Numérica
Comentarios	A partir de la variable explicativa **localidad_nombre** y los resultados del Censo 2022 del Instituto Nacional de Estadística y Censos (INDEC), se reemplaza la variable existente por la cantidad de habitantes de la localidad donde ocurre el homicidio doloso.

- Segunda variable explicativa

Nombre de la variable	mes
Descripción	Mes de ocurrencia del hecho
Tipo de variable	Numérica
Comentarios	Valores con los que aparece: 1 a 12.

- Tercera variable explicativa

Nombre de la variable	dia
Descripción	Día del mes de ocurrencia del hecho
Tipo de variable	Numérica
Comentarios	Valores con los que aparece: 1 a 31.

- Cuarta variable explicativa

Nombre de la variable	dia_mes
Descripción	Día de la semana de ocurrencia del hecho
Tipo de variable	Numérica
Comentarios	Valores con los que aparece: 1 (lunes) a 7 (domingo).

- Quinta variable explicativa

Nombre de la variable	lugar
Descripción	Lugar de ocurrencia del hecho
Tipo de variable	Texto
Valores con los que aparece	hogar - otro - via_publica

- Sexta variable explicativa

Nombre de la variable	arma
Descripción	Tipo de arma o mecanismo que se utilizó para provocar la muerte de la víctima
Tipo de variable	Texto
Valores con los que aparece	ahorcamiento - arrollamiento - blanca - envenenamiento - fuego - golpes - objeto_contundente - otro - quemadura - salto_al_vacio

- Séptima variable explicativa

Nombre de la variable	delito
Descripción	Este campo busca identificar los hechos de homicidios dolosos que se producen en el contexto de otro tipo de delitos
Tipo de variable	Texto
Valores con los que aparece	no - otro - robo - violacion

- Octava variable explicativa

Nombre de la variable	registro
Descripción	Motivo que origina el registro del hecho
Tipo de variable	Texto
Valores con los que aparece	orden_judicial - otro - denuncia - policial

- Novena variable explicativa

Nombre de la variable	genero
Descripción	Identidad de género de la víctima
Tipo de variable	Texto
Valores con los que aparece	mujer - trans - varon

- Décima variable explicativa

Nombre de la variable	edad
Descripción	Edad de la víctima
Tipo de variable	Numérica
Comentarios	La variable explicativa original **victima_edad_tr** se encontraba en tramos de 5 años (15 - 19). En la nueva variable **edad**, se tomó como edad el límite inferior de cada tramo (15).

- Onceava variable explicativa

Nombre de la variable	clase
Descripción	Esta variable busca identificar algunos tipos de víctimas en función de la relevancia de sus categorías para el análisis de los hechos
Tipo de variable	Texto
Valores con los que aparece	civil - detenido - otra_fuerza_seguridad - policia_en_servicio - policia_no_servicio - policia_no_servicio - seguridad_privada

- Variable dependiente

Nombre de la variable	victima_relacion_inculpado
Descripción	Esta variable releva el tipo de vínculo de la víctima con el inculpado, y se asigna a cada víctima
Tipo de variable	Texto
Valores con los que aparece	relacion_familiar - relacion_no_familiar - sin_relacion

Se debe considerar que, con el fin de obtener una mejor predicción por parte del modelo y tal como propuso Riedel (1987), se agruparon los factores de esta última variable de la siguiente manera:

relacion_familiar	Conyuge/ pareja - Hermano/a - Hijo/a - Otros vínculos familiares - Padre/Madre
relacion_no_familiar	Cliente / proveedor - Empleado - Empleador - Ex conyuge / ex pareja - Otras relaciones no familiares - Socio (Relación comercial)
sin_relacion	Sin relacion

Considerando todos los cambios descriptos anteriormente, se muestra a continuación una porción de la base de datos a modo ilustrativo:

Índice	Población	Mes	Día	Día Sem.	Lugar	Arma	Delito	Registro	Género	Edad	Clase	Relación Inculpado
1	288.896	1	7	6	Vía Pública	Fuego	No	Policial	Varón	25	Civil	Sin Rel.
2	1.438	1	2	1	Hogar	Ahorcamiento	No	Policial	Varón	70	Civil	Rel. Familiar
3	288.896	4	3	1	Vía Pública	Fuego	No	Policial	Varón	15	Civil	Sin Rel.
4	40.472	3	17	5	Vía Pública	Blanca	No	Policial	Varón	40	Detenido	Sin Rel.
5	895	3	11	6	Hogar	Fuego	No	Policial	Mujer	40	Civil	Rel. No Familiar
6	36.913	3	12	7	Campo	Blanca	Otro	Denuncia	Varón	20	Civil	Sin Rel.
7	3.921	6	8	4	Hogar	Fuego	No	Policial	Mujer	5	Civil	Rel. Familiar
8	5.951	4	22	6	Hogar	Ahorcamiento	No	Policial	Mujer	0	Civil	Rel. Familiar
9	591.163	1	1	7	Hogar	Blanca	No	Denuncia	Varón	45	Civil	Rel. Familiar
10	591.163	1	5	4	Comercio	Objeto Contundente	No	Denuncia	Varón	35	Civil	Rel. No Familiar
...	...	...	...	...	...	...	...	...	...	...	...	...
16	106.214	2	24	5	Hogar	Objeto Contundente	No	Denuncia	Varón	*NA*	*NA*	Sin Rel.
...	...	...	...	...	...	...	...	...	...	...	...	...
7128	598.835	8	13	7	Hogar	Quemadura	No	Policial	Mujer	55	Civil	Rel. Familiar

Tabla 2: Base de datos final con mayor tamaño y ancho completo.

La base de datos, luego de su debida preparación, cuenta con 7 128 observaciones y 12 variables explicativas. Sin embargo, la misma posee datos faltantes que son indeseados para el correcto desarrollo de los modelos. En la siguiente tabla se vuelcan la cantidad de datos faltantes en cada variable explicativa:

lugar	770
arma	281
delito	195
genero	2
edad	169
clase	131
Total	**1 548**

Tabla 3: Cantidad de datos faltantes por variable.

Esta cantidad de datos faltantes se encuentran distribuidos en 1 392 observaciones, lo cual indica que hay 5 736 observaciones completas. Los datos faltantes no pueden permanecer en la base para el correcto desarrollo de los modelos explicativos y predictivos, por ende, se puede optar por dos alternativas: eliminar todas aquellas observaciones con datos faltantes y obtener una base de datos más reducida, o completar los datos faltantes con algún método que lo permita. A priori, pareció conveniente utilizar la base de datos más extensa posible, por ende, se decidieron evaluar los siguientes métodos para tratar los datos faltantes:

1. *Imputación de categoría "faltante" para variables categóricas y valor ilógico para variables numéricas*: En este intento, para aquellas variables categóricas (*lugar*, *arma*, *delito*, *genero* y *clase*), se imputó una nueva categoría denominada "faltante". A su vez, para la única variable numérica incompleta (*edad*), se imputó un valor ilógico, en este caso se utilizó -1 (menos uno).

2. *Imputación de categoría "faltante" para variables categóricas y mediana para variables numéricas*: Nuevamente se imputó una categoría "faltante" para las variables categóricas, pero para la variable *edad*, se utilizó la mediana de los valores disponibles. Además, se añadió una variable indicadora que adoptaba un valor *TRUE*, cuando la observación había recibido un valor imputado, y un valor *FALSE* cuando la observación estaba completa y no fue imputada.

3. *Imputación con el algoritmo MICE:* En último lugar, se realizaron imputación a los datos faltantes a partir del método MICE (multivariate imputation by chained equations). El mismo ofrece una serie de herramientas para la imputación de datos faltantes dependiendo de la naturaleza de las observaciones y las relaciones entre las variables a imputar. MICE crea una serie de m bases de datos imputadas, que difieren únicamente entre sí en los valores faltantes. En cada base de datos imputada, se reemplazan los valores faltantes por valores predichos a partir del resto de las variables completas, usando distintos modelos predictivos según el tipo de variable. A partir de cada base de datos imputada, se realiza una predicción usando el modelo predictivo propuesto y se calcula un promedio de dichas predicciones. (Buuren et al., 2011)

Entre todas las alternativas del tratamiento de datos faltantes, se optó por eliminarlos en los modelos explicativos y predictivos, y reemplazarlos por una categoría denominada *desconocido* en el análisis exploratorio. Así, la base de datos con los registros incompletos eliminados resultó de 5 736 observaciones, con 11 variables explicativas y 1 variable respuesta.

4.1.3. Descripción General

Los datos usados cuentan con un total de 7 128 observaciones, cada una con 11 variables explicativas y 1 variable a analizar (*victima_relacion_inculpado*).

Del total de las victimas identificadas, un $21,8\,\%$ fueron mujeres, un $77,9\,\%$ fueron varones y $0,98\,\%$ fueron transexuales. El $21,1\,\%$ de las víctimas tenían entre 0-20 años, el $37,1\,\%$ entre 20-40 años, el $13\,\%$ entre 40-60 años y el $28,8\,\%$ más de 60 años. El $78,1\,\%$ de los homicidios no fueron en ocasión de otro delito, mientras que un $0,36\,\%$ fueron en ocasión de violación, un $18,7\,\%$ de robo y un $2,84\,\%$ en otro tipo de delito. En cuanto a los vínculos victima-inculpado, el $26,5\,\%$ de los inculpados mantenían una relación familiar, el $33,3\,\%$ una relación no familiar y el restante $40,2\,\%$ era sin relación.

4.2. Estrategia de Análisis Exploratorio

A partir de los datos recogidos, se realizaron una serie de comparaciones entre las distintas variables y la variable respuesta (*victima_relacion_inculpado*). Para eso, se tomó un gráfico de la frecuencia de cada tipo de relación de acuerdo a los distintos factores posibles de las variables independientes más relevantes (*poblacion*, *dia_sem*, *lugar*, *arma*, *delito*, *genero* y *edad*).
En cada gráfico, la amplitud de clase está definida por los factores de la variable analizada. Es decir, si tomamos la variable *dia_sem* como ejemplo, vemos que el gráfico tiene 7 barras verticales: una por cada factor (cada día de la semana). En el caso de las variable numéricas, la amplitud de clase de *poblacion* es de 100 000 habitantes, y la de *edad* es de 5 años.

Con el objetivo de tener un panorama más completo de la información, se optó por no eliminar las observaciones incompletas en la confección de los gráficos. Para eso, se tuvieron en cuenta los datos faltantes y se agregaron (cuando fuera necesario) como una categoría nueva catalogada "desconocida".

Estas representaciones visuales de la base de datos permiten observar la distribución de los factores en las distintas categorías de relación, lo cual muestra las tendencias en las características de los homicidios en Argentina. A su vez, este análisis posibilita la detección de desequilibrios entre las categorías de la variable respuesta, es decir, que alguna sea mucho más "pequeña"que las otras. Estas características deterioran la capacidad predictiva del modelo, por lo cual es importante utilizar una base de datos bien estructurada y con categorías equilibradas. Una vez que se realiza este análisis descriptivo y se obtienen un mayor conocimiento acerca de los datos, se puede avanzar hacia el desarrollo de los modelos explicativos y predictivos.

4.3. Modelado Explicativo

El primer objetivo de este estudio es describir y demostrar la existencia de patrones de correlación entre factores del homicidio y el vínculo víctima-victimario en homicidios dolosos en Argentina. Para eso, es conveniente desarrollar algún modelo explicativo que permita analizar cada variable independiente y su influencia en los distintos posibles resultados de la variable respuesta. La ventaja de este enfoque consiste en poder analizar el impacto individual de cada variable de manera aislada, manteniendo todas las otras variables explicativas fijas

En líneas generales, los modelos que sueles ser mejores para realizar predicciones (random forest, redes neurales), no evidencian de forma clara las relaciones entre las variables y los resultados. Como plantean Haddouchi y Berrado (2024), los modelos resultantes de random forest se consideran una "caja negra"debido a sus numerosos árboles de decisión profundos, obtener una visión clara de todo el proceso que conduce a las decisiones finales explorando cada árbol de decisión es complicado, si no imposible. De esta manera, no son los modelos óptimos para realizar un análisis explicativo del fenómeno que se quiere predecir y entender. Por otro lado, algunos modelos que no destacan por su capacidad predictiva son mejores para entender la naturaleza e influencia de las variables, como las regresiones lineales o logísticas. En nuestro caso, resulta más útil usar un modelo con mejores características explicativas que predictivas para realizar una interpretación de las variables y la manera en que afectan el vínculo víctima-victimario en homicidios. Por este motivo, se utilizó un modelo logístico multinomial.

4.3.1. Modelo Logístico Multinomial

El modelo logístico plantea una función lineal para la probabilidad de cada clase y luego, según un corte de probabilidad aplicado, clasifica la variable respuesta a predecir. La estimación se hace por máxima verosimilitud aplicando el algoritmo de Newton & Raphson, el cual asegura la consistencia y normalidad asintótica del estimador.

Este modelo posee la desventaja de que no se puede utilizar para variables de respuesta discretas múltiples y solo permite que sean binarias. Por ende, debido a que en el caso analizado la variable respuesta tiene tres factores posibles, se debe utilizar un modelo alternativo conocido como la **regresión logística multinomial**, que generaliza el método de regresión logística para problemas con más de dos posibles resultados discretos.

Este modelo define a J como el número de categorías de la variable Y y a $\{\pi_1, \cdots, \pi_J\}$ como las probabilidades de respuesta, cumpliendo que $\sum_j \pi_j = 1$. Con n observaciones independientes, la distribución de probabilidad para el número de resultados en las J categorías es la multinomial. Esta especifica la probabilidad de cada posible manera en que las n observaciones pueden distribuirse entre las J categorías.

Los modelos logit para variables de respuesta nominal emparejan cada categoría con una *categoría de referencia*. Dado que la respuesta cae en la categoría j o en la categoría J, esta es la *razón logarítmica de probabilidades* (log-odds) de que la respuesta sea j. Por ejemplo, en nuestro caso con $J = 3$, el modelo utiliza $\log(\pi_1/\pi_3)$ y $\log(\pi_2/\pi_3)$. El modelo logit con una variable predictora x se expresa como:

$$\log\left(\frac{\pi_j}{\pi_J}\right) = \alpha_j + \beta_j x, \quad j = 1, \ldots, J-1 \tag{1}$$

El modelo logístico multinomial también puede expresarse en términos de las *probabilidades de respuesta* de la siguiente manera:

$$\pi_j = \frac{e^{\alpha_j + \beta_j x}}{\sum_{h=1}^{J} e^{\alpha_h + \beta_h x}}, \quad j = 1, \ldots, J \tag{2}$$

El denominador es el mismo para cada probabilidad y asegura que la suma de las probabilidades para todas las categorías sea 1. En esta formulación, los *parámetros* α_j y β_j representan los efectos específicos para cada categoría. Cuando todos los parámetros se igualan a cero, la expresión resulta en una distribución uniforme de probabilidades, es decir, cada categoría tiene la misma probabilidad (Agresti, 2007).

Los elementos extraídos del modelo explicativo que permitirán analizar los homicidios dolosos son dos. Por un lado, se considerarán los coeficientes de regresión o estimaciones de cada variable, β_j. Por otro lado, se utilizarán los efectos marginales promedio sobre la probabilidad de que la variable respuesta sea cada una de las tres relaciones posibles. Este valor indica la variación en la probabilidad de cada relación cuando varía cada variable, manteniendo a las otras constantes.

Dado que en los modelos logísticos multinomiales los efectos marginales no son iguales entre individuos (ya que dependen de las características particulares de cada observación i), se calculan de forma individual y luego se promedian para obtener un valor global. La fórmula para el efecto marginal de la variable x_k sobre la probabilidad de que la respuesta sea la categoría j para el individuo i es: :

$$\frac{\partial \pi_{ij}}{\partial x_{ik}} = \pi_{ij} \left(\beta_{jk} - \sum_r \pi_{ir} \beta_{rk} \right) \tag{3}$$

Luego, el efecto marginal promedio se calcula como el promedio de estos valores individuales para todos los individuos i. Los efectos marginales fueron calculados con la ayuda de Excel.

En nuestro caso, para realizar un análisis explicativo más claro y en función de los datos obtenidos en el análisis exploratorio, se realizaron algunas variaciones a la base de datos a utilizar. En primer lugar, en base a los gráficos de frecuencia realizados de cada variable, se mantuvieron las variables independientes que mostraban alguna tendencia entre sus factores: *población*, *dia_sem*, *lugar*, *arma*, *delito*, *genero* y *edad*. Las variables numéricas (*población* y *edad*) se adaptaron tal que *población* esta expresada en 100 000 habitantes, y *edad* está expresada cada 10 años. La variable *arma* fue reducida en sus factores a los siguientes 3: arma_fuego, arma_blanca y sin_arma. Por último, las variables restantes fueron modificadas para convertirlas en binarias, lo que permite realizar un análisis más claro del impacto de cada variable independiente en la variable respuesta. Sus cambios fueron:

- dia_sem: se la renombró como "fin_de_semana", donde los días del lunes al jueves adoptaron el valor 0 y los días del viernes al domingo adoptaron el valor 1.
- lugar: se la renombró como "hogar", donde los homicidios ocurridos fuera del hogar adoptaron el valor 0 y los que fueron dentro del hogar adoptaron el valor 1.
- delito: los homicidios ocurridos en ocasión de robo u otro delito adoptaron el valor 1 y los casos en los que no hubo otro delito adoptaron el valor 0.
- genero: se la renombró "mujer", donde los homicidios a mujeres adoptaron el valor 1 y los homicidios a hombres adoptaron el valor 0. Los casos que involucran transexuales fueron eliminados para este análisis debido a que representan una porción muy pequeña de los datos.

Luego, a partir del paquete **mlogit** se definió el modelo logístico multinomial, donde se establece cuál es la variable respuesta y cuáles las variables explicativas. Para cada variable independiente se determinó un factor de referencia, siendo sin_arma para *arma* y el valor 0 para las restantes. Además, se determinó "sin_relacion" como factor de referencia del modelo. Esto implica que los coeficientes de regresión de cada factor de cada variable indican qué tanto más o menos probable es que el vínculo sea de "relacion_familiar" o "relacion_no_familiar" comparado a "sin_relacion".

Cada coeficiente es testeado con la *prueba de Wald*, utilizando un valor de significancia de 0,05. Aquellos significativos validan la existencia de una relación entre estos y la variables respuesta, que es luego interpretada.

4.4. Modelado Predictivo

El segundo objetivo de este estudio es desarrollar un modelo capaz de predecir con cierto grado de significancia el vínculo víctima-victimario en homicidios dolosos en Argentina a partir de factores propios del homicidio. Con este fin, se busca realizar un modelado predictivo con los datos disponibles. El objetivo de la modelización es encontrar un modelo parsimonioso que ajuste bien a los datos. Por "parsimonioso" se entiende, según la *Lex Parsimoniæ*, a aquel modelo que, frente a otro que represente de igual manera la realidad, lo haga de la forma más simple. En este caso, debido a la naturaleza de los datos, se desarrolla una clasificación supervisada a través de métodos de predicción discreta. Esto quiere decir que frente a nuevas observaciones sin variable respuesta se las debe clasificar entre varios grupos de unidades. Para esto se desean

hallar las variables que permiten discriminar o discernir de la mejor manera. Algunos ejemplos de modelos predictivos discretos son el reconocimiento de caracteres de escritura manual, la discriminación de mails en deseados, spam o notificación, la identificación de fraude en tarjetas de crédito a partir de las características de la operación y la historia del cliente, entre otros.

En este caso de estudio, se desea hallar un modelo que frente a las características de un homicidio doloso pueda identificar la relación entre la víctima y el inculpado. Debido a que la variable respuesta puede adoptar tres categorías ("relacion_familiar", "relacion_no_familiar" y "sin_relacion"), el modelo debe ser discreto múltiple y no binario o continuo. Algunos modelos que se pueden utilizar son K-Nearest Neighbours, Naïve Bayes, Random Forest, Neural Network, Logístico Multinomial, entre otros. Contrario a lo mencionado en el apartado anterior (subsección 4.3), en este segundo objetivo es conveniente utilizar algoritmos con mejores capacidades predictivas. En general, estos algoritmos no evidencian de forma tan clara las relaciones entre las variables y los resultados. Considerando esto y en base a resultados obtenidos en estudios propios anteriores, se optó por utilizar modelos de *Gradient Boosting*, los cuales emplean una serie de arboles de clasificación en sus implementaciones.

4.4.1. Gradient Boosting

La técnica de *Gradient Boosting* en *machine learning* se basa en combinar las predicciones de varios modelos débiles para obtener un solo modelo más preciso. Se considera modelo débil a aquel que predice ligeramente mejor que adivinar al azar y en general, para estos modelos se suelen utilizar árboles de decisión pequeños. Luego, se combinan decenas o cientos de estos modelos que se ajustan de forma adaptativa para reducir el sesgo. Esto quiere decir que cada árbol intenta corregir los errores del árbol anterior para así poder alcanzar predicciones con mayor capacidad predictiva (Hastie, Tibshirani, & Friedman, 2009).

Los modelos de *Gradient Boosting* poseen ciertos hiperparámetros que se deben ajustar para obtener los mejores resultados posibles. Entre estos se encuentran:

1. *Máxima profundidad (max_prof)*: La cantidad de niveles de cada modelo débil estará limitada por este hiperparámetro. Si el mismo, por ejemplo, toma un valor de 3, todos los árboles de decisión podrán adoptar hasta 3 niveles, incluyendo los nodos terminales. Con una mayor profundidad de los árboles, aumenta el costo computacional y la complejidad del modelo. Se debe tener en cuenta que una profundidad máxima cercana a 3 ayudará a prevenir el *overfitting* y que no es recomendable utilizar más de 10 niveles.

2. *Cantidad de árboles (n_arb)*: El tercer hiperparámetro indica la cantidad de modelos débiles, en este caso árboles de decisión, que tendrá el modelo. Con una mayor cantidad de árboles, se obtiene mayor robustez y rendimiento pero se aumenta la complejidad y la probabilidad de *overfitting*. Para evitar esto se suelen utilizar bajas tasas de aprendizaje. Se suelen usar cantidades de árboles cercanas a los 20, 50, 100, 500, 1000, dependiendo de la tasa de aprendizaje a utilizar.

3. *Tasa de aprendizaje (eta)*: Este hiperparámetro controla la contribución de cada modelo débil. Los valores más pequeños disminuyen el peso que tiene cada modelo en el ensamble final. En consecuencia, si se utiliza un valor pequeño, se precisarán más modelos débiles y más entrenamiento, pero se obtendrá un resultado final más potente y menos propenso a *overfitting*. Los valores que se suelen utilizar son 0,1 , 0,05 , 0,01 y 0,001.

El algoritmo *Gradient Boosting* se ha implementado en varias librerías de *Python* entre las cuales se destacan *eXtreme Gradient Boosting (XGBoost)*, *Light Gradient Boosting Machine (LGBM)* y *Categorical Boosting (CatBoost)*. En este trabajo se evaluarán los primeros dos mencionados

para abordar el segundo objetivo propuesto: *XGBoost* y *LGBM*. Una de las mayores diferencias entre estas librerías es que *XGBoost* realiza el crecimiento de los árboles nivel a nivel, mientras que *LGBM* lo hace hoja por hoja. Esto conlleva a que en *LGBM* los árboles tengan estructuras más irregulares, con gran profundidad pero baja cantidad de hojas, por ejemplo. A su vez, se obtienen menores tiempos de entrenamiento lo que lo hace mejor para trabajar con grandes bases de datos que el *XGBoost*. Por otro lado, el algoritmo *LGBM* es más propenso a tener *overfitting* que el *XGBoost* (Florek et al., 2023). Para el caso específico a tratar en este trabajo, predicción del vínculo víctima-victimario en los homicidios dolosos, se desarrollaran ambos algoritmos y se los evaluará para elegir el que mejor se ajuste a la realidad.

4.4.2. Criterio de Elección del Modelo

Con el fin de poder medir el nivel predictivo del modelo y evaluar su fiabilidad, se dividió aleatoriamente la base de datos en 3 grupos: *Muestra de aprendizaje*, *Muestra de prueba Muestra de validación*. Los tamaños de estas muestras fueron del 70 %, 15 % y 15 % de las observaciones, respectivamente. Dado que la base de datos cuenta con 5 736 observaciones, la *Muestra de aprendizaje* resultó con un tamaño de 4 016 observaciones, la *Muestra de prueba* de 859 observaciones y la *Muestra de validación* de 861 observaciones.

Para ajustar el modelo correctamente, se debe elegir los hiperparámetros del mismo siguiendo una serie de pasos. En primer lugar, se definen 4 valores posibles para la *máxima profundidad* (2, 3, 5 y 10), 3 valores posibles para la *cantidad de árboles* (20, 50 y 100) y 4 valores posibles para la *tasa de aprendizaje* (0,1 , 0,05 , 0,01 y 0,001). A partir de estos posibles hiperparámetros se toman las $4 \cdot 3 \cdot 4 = 48$ combinaciones y se define un modelo para cada combinación de parámetros. Con cada uno de ellos, se lo entrena a partir de la *muestra de aprendizaje* y se evalúa su performance utilizando la *muestra de validación*. De entre los 48 modelos evaluados se selecciona aquel con mejores indicadores. A este modelo definitivo, se lo entrena con la *muestra de aprendizaje* y *muestra de validación* unificadas, y se evalúa su capacidad predictiva con la *muestra de prueba*, cotejando las predicciones obtenidas con las observaciones reales. Esta manera de seleccionar los hiperparámetros con la *muestra de validación* y *muestra de prueba* busca reportar la capacidad predictiva en una base de datos aún no vista.

#	max_prof	n_arb	eta
1	2	20	0,100
2	3	20	0,100
3	5	20	0,100
4	10	20	0,100
5	2	50	0,100
6	3	50	0,100
7	5	50	0,100
8	10	50	0,100
9	2	100	0,100
10	3	100	0,100
11	5	100	0,100
12	10	100	0,100
13	2	20	0,050
14	3	20	0,050
15	5	20	0,050
16	10	20	0,050
17	2	50	0,050
18	3	50	0,050
19	5	50	0,050
20	10	50	0,050
21	2	100	0,050
22	3	100	0,050
23	5	100	0,050
24	10	100	0,050

#	max_prof	n_arb	eta
25	2	20	0,010
26	3	20	0,010
27	5	20	0,010
28	10	20	0,010
29	2	50	0,010
30	3	50	0,010
31	5	50	0,010
32	10	50	0,010
33	2	100	0,010
34	3	100	0,010
35	5	100	0,010
36	10	100	0,010
37	2	20	0,001
38	3	20	0,001
39	5	20	0,001
40	10	20	0,001
41	2	50	0,001
42	3	50	0,001
43	5	50	0,001
44	10	50	0,001
45	2	100	0,001
46	3	100	0,001
47	5	100	0,001
48	10	100	0,001

Tabla 4: Combinaciones de los posibles hiperparámetros

La evaluación del modelo se debe realizar con parámetros que permitan analizar el ajuste de las predicciones con la realidad. En este sentido, dos aspectos distintos entran a formar parte típicamente del estudio de fiabilidad de un modelo. Por un lado, el sesgo entre la variable respuesta predicha y la variable respuesta observada, es decir, la tendencia del modelo a dar consistentemente el mismo error. Por el otro, la concordancia entre las variables mencionadas, que refleja hasta qué punto las predicciones coinciden con la realidad (López de Ullibarri Galparsoro & Pita Fernández, 2001). La manera de poder abordar estos parámetros depende de la naturaleza de los datos a estudiar. En este trabajo se utilizarán tres indicadores para evaluar los modelos propuestos: acierto general, F_1-score y kappa de Cohen. Cada uno de ellos se calcula a partir de la matriz de confusión del modelo, un método de visualización para los resultados del algoritmo clasificador, reflejado en la Tabla 5.

		Clase real			
		relacion_familiar	**relacion_no_familiar**	**sin_relacion**	Total
Clase predicha	**relacion_familiar**	x_{11}	x_{12}	x_{13}	X_1
	relacion_no_familiar	x_{21}	x_{22}	x_{23}	X_2
	sin_relacion	x_{31}	x_{32}	x_{33}	X_3
	Total	$X_{.1}$	$X_{.2}$	$X_{.3}$	N

Tabla 5: Matriz de confusión para la variable respuesta predicha.

1. Acierto general (AG): mide la proporción de predicciones correctas sobre el total de predicciones realizadas. Es un indicador simple e intuitivo, pero puede ser engañosos cuando una clase (factor) es mucho más frecuente que otra. En nuestro caso, las clases están relativamente balanceadas, con una distribución de 2 426, 1 916 y 1 394 observaciones entre los tres tipos de vínculos. Se calcula a partir de:

$$AG = \frac{x_{11} + x_{22} + x_{33}}{N} \tag{4}$$

2. F_1-score: mide la media armónica entre la precisión y la sensibilidad. Es un indicador que toma en consideración la sensibilidad y la precisión y es particularmente relevante en casos donde los falsos negativos y falsos positivos tienen consecuencias a tomar en cuenta. En este caso, donde la variable respuesta tiene 3 factores, es necesario calcular la media armónica de cada clase y realizar luego un promedio entre estos 3 valores (F_1 macro). Se a calcula a partir de:

$$F_{1.rel.fam} = 2 \cdot \frac{\frac{x_{11}}{X_1} \cdot \frac{x_{11}}{X_{.1}}}{\frac{x_{11}}{X_1} + \frac{x_{11}}{X_{.1}}} \tag{5}$$

$$F_{1.rel.no.fam} = 2 \cdot \frac{\frac{x_{22}}{X_2} \cdot \frac{x_{22}}{X_{.2}}}{\frac{x_{22}}{X_2} + \frac{x_{22}}{X_{.2}}} \tag{6}$$

$$F_{1.sin.rel} = 2 \cdot \frac{\frac{x_{33}}{X_3} \cdot \frac{x_{33}}{X_{.3}}}{\frac{x_{33}}{X_3} + \frac{x_{33}}{X_{.3}}} \tag{7}$$

$$F_{1.macro} = \frac{F_{1.rel.fam} + F_{1.rel.no.fam} + F_{1.sin.rel}}{3} \tag{8}$$

3. Kappa de Cohen (κ): este coeficiente es más robusto que el acierto general debido a que también ajusta el efecto del azar en la proporción del acierto observado. Según Cantor (1995), el cual se basa en lo definido por Cohen en 1960, el índice *kappa* se toma de la siguiente forma:

$$\kappa = \frac{\pi_o - \pi_e}{1 - \pi_e} \tag{9}$$

Donde π_0 es la proporción de coincidencias entre la clase real y la predicha, y π_e es la proporción esperada de coincidencias debido al azar. Estas probabilidades se calculan de la siguiente manera:

$$\pi_o = \frac{x_{11} + x_{22} + x_{33}}{N} \tag{10}$$

$$\pi_e = \frac{X_1}{N} \cdot \frac{X_{.1}}{N} + \frac{X_2}{N} \cdot \frac{X_{.2}}{N} + \frac{X_3}{N} \cdot \frac{X_{.3}}{N} \tag{11}$$

La ventaja de este índice por sobre la simple concordancia (porcentaje de observaciones bien clasificadas) es que considera la concordancia por azar y *"penaliza"* en consecuencia. A modo ilustrativo, se tomaran dos ejemplos de posibles matrices de confusión de dos factores:

		Clase real		
		1	**0**	Total
Clase predicha	**1**	10	80	90
	0	0	10	10
	Total	10	90	100

Tabla 6: Matriz de confusión caso A.

		Clase real		
		1	**0**	Total
Clase predicha	**1**	80	10	90
	0	10	0	10
	Total	90	10	100

Tabla 7: Matriz de confusión caso B.

En el *Caso A* la concordancia es de 20 % mientras que en el *Caso B* es de 80 %, con lo cual se podría concluir que el segundo modelo se ajusta mejor a la realidad. Sin embargo al observar las probabilidades esperadas por efecto del azar, el resultado cambia. A continuación se calcula el índice *kappa* para ambas situaciones:

Caso A $\quad \pi_o = \frac{10+10}{100} = 0,2 \quad \pi_e = \frac{90}{100} \cdot \frac{10}{100} + \frac{10}{100} \cdot \frac{90}{100} = 0,18 \quad \kappa = \frac{0,2-0,18}{1-0,18} = 0,024$

Caso B $\quad \pi_o = \frac{80+0}{100} = 0,8 \quad \pi_e = \frac{90}{100} \cdot \frac{90}{100} + \frac{10}{100} \cdot \frac{10}{100} = 0,82 \quad \kappa = \frac{0,8-0,82}{1-0,82} = -0,111$

En los ejemplos provistos, se puede observar que el modelo con menor concordancia obtiene un índice *kappa* mayor. Incluso el segundo modelo, cuya concordancia es significativamente elevada, obtuvo un valor negativo de *kappa*. En conclusión se puede considerar al índice *kappa* un coeficiente robusto y de mayor utilidad para evaluar los resultados del modelo.

A partir de estos 3 indicadores se logra obtener una evaluación integral de los modelos analizados. Se selecciona el modelo que obtenga los valores más altos para los indicadores descriptos, suponiendo que será el que mejor se ajuste a la realidad.

5. Resultados

5.1. Análisis Exploratorio

Los resultados de los gráficos de frecuencia de cada variable explicativa según el tipo de vínculo víctima-victimario se presentan a continuación. De los mismos es posible esbozar conclusiones generales acerca de la relación entre la variable explicativa y el vínculo.

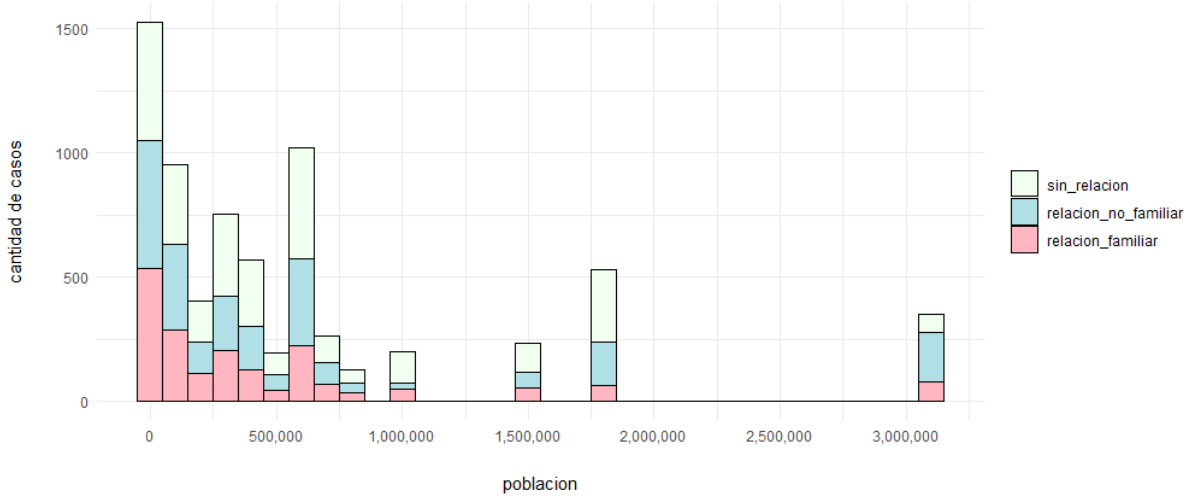

Figura 1: Histograma de la variable *poblacion* con respecto al vinculo víctima-inculpado.

En lo que respecta a la población del lugar de homicidio, en la Figura 1 se puede notar que en localidades con menor población, el porcentaje de homicidios causados por un familiar es mayor que en localidades con mayor población. Además, el grueso de los homicidios está concentrado en las localidades de menor población. Se reconoce, de derecha a izquierda, las barras correspondientes a los homicidios ocurridos en la Ciudad de Buenos Aires, La Matanza, Córdoba y Rosario, respectivamente. Cabe destacar que en las últimas 3 ciudades, aproximadamente la mitad de los homicidios tienen vínculo **sin relación**.

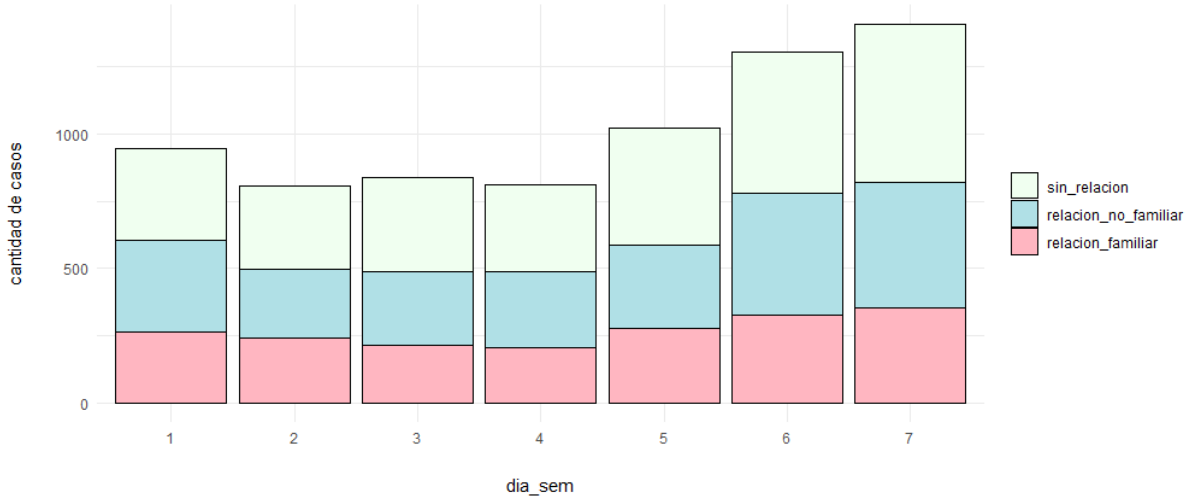

Figura 2: Gráfico de barras de la variable *dia_sem* con respecto al vinculo víctima-inculpado.

En la Figura 2, 1 representa al lunes y 7 a domingo. Se observa que los fines de semana ocurre una mayor cantidad de homicidios, siendo el domingo el día con más casos registrados, seguido por el sábado y luego por el viernes. De los días de semana, el lunes es el que más casos tiene

y el martes el que menos tiene. No se identifican relaciones muy claras respecto a los vínculos víctima-victimario en cada día.

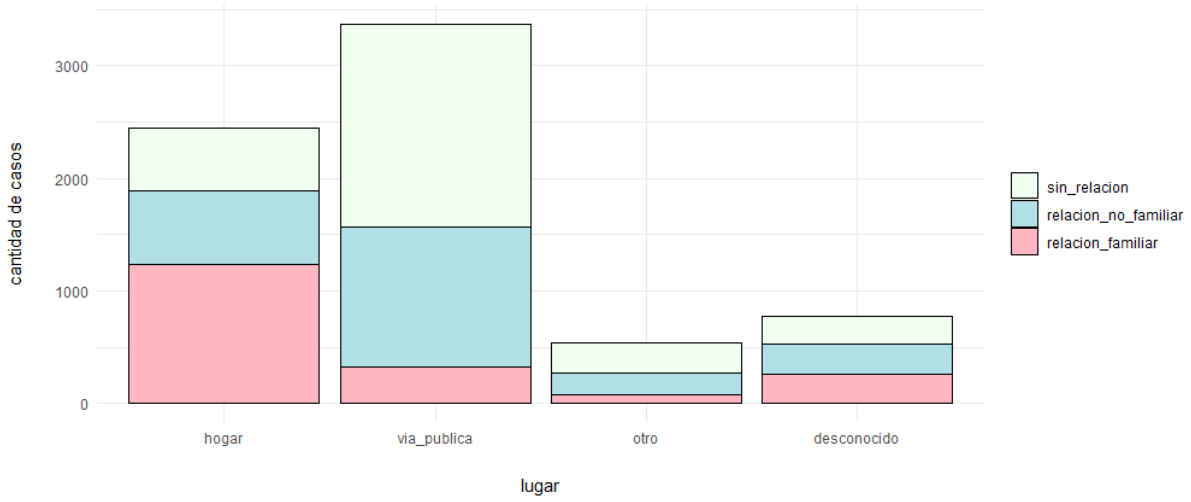

Figura 3: Gráfico de barras de la variable *lugar* con respecto al vinculo víctima-inculpado.

Como indica la Figura 3, la mayor parte de los homicidios registrados fueron en la vía pública y, en segundo lugar, en domicilios particulares. Es evidente la tendencia al homicidio de desconocidos en la vía pública, mientras que el grueso de los homicidios en hogares fueron causados por familiares. Cabe destacar el bajo porcentaje de homicidios familiares en la vía pública y el alto porcentaje relativo de homicidios causados por conocidos no familiares.

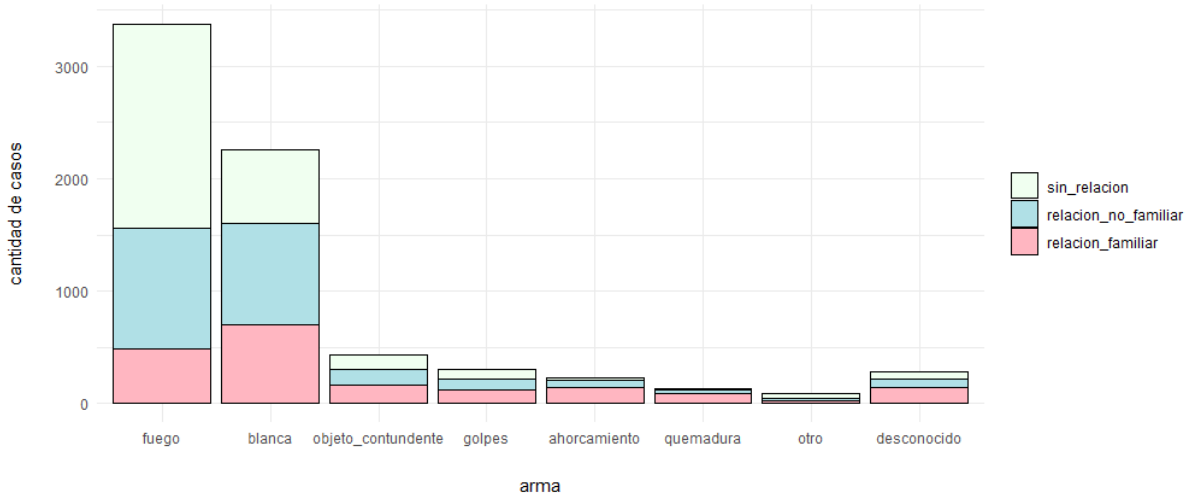

Figura 4: Gráfico de barras de la variable *arma* con respecto al vinculo víctima-inculpado.

En cuanto a las armas utilizadas, se observa en la Figura 4 que la mayor cantidad de homicidios fueron cometidos con armas de fuego y en segundo lugar con armas blancas. Aquellos homicidios realizados con un arma de fuego corresponden mayoritariamente al vínculo **sin_relacion**, con más de la mitad de los casos. Por otro lado, los homicidios cometidos por un familiar son con mayor frecuencia con un arma blanca que con cualquier otra arma. Los homicidios ocasionados con un objeto contundente o a golpes presentan una distribución relativamente equitativa entre las tres categorías de vínculo posibles. Sin embargo, en el caso de ahorcamiento y quemadura, el vínculo **relacion_familiar** parecería ser más frecuente que los otros, superando la mitad de los casos.

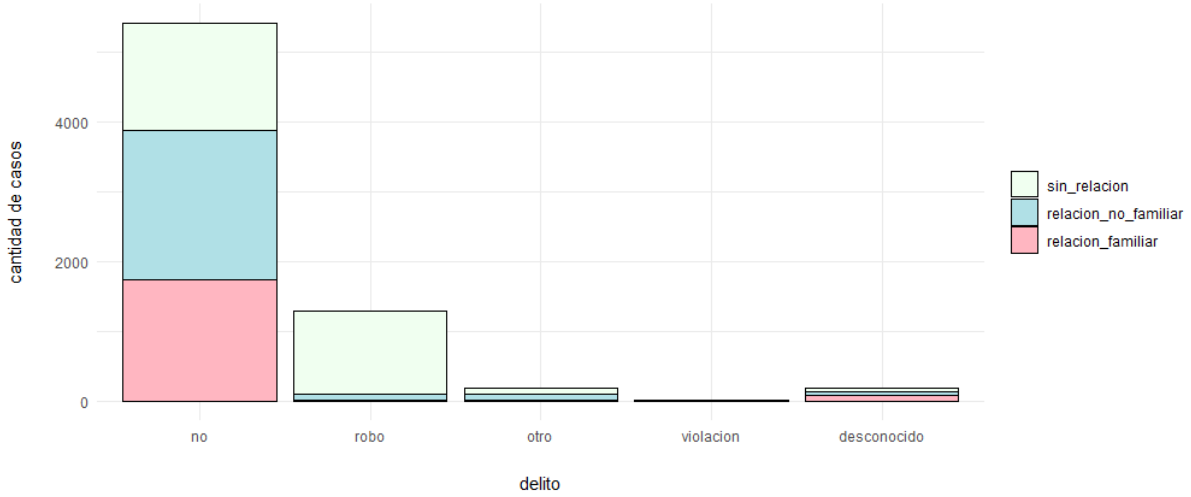

Figura 5: Gráfico de barras de la variable *delito* con respecto al vinculo víctima-inculpado.

Los homicidios que no fueron realizados en el contexto de otro delito presentan aproximadamente frecuencias similares para las tres categorías de vínculo. No obstante, este no es el caso para cuando el homicidio se comete junto a un robo. En estas situaciones el vínculo que prima por sobre los otros es **sin_relacion**, tomando casi la totalidad de los casos. En los casos en los que el homicidio fue cometido por un familiar, este se presenta casi totalmente en la ausencia de cualquier otro delito (Figura 5).

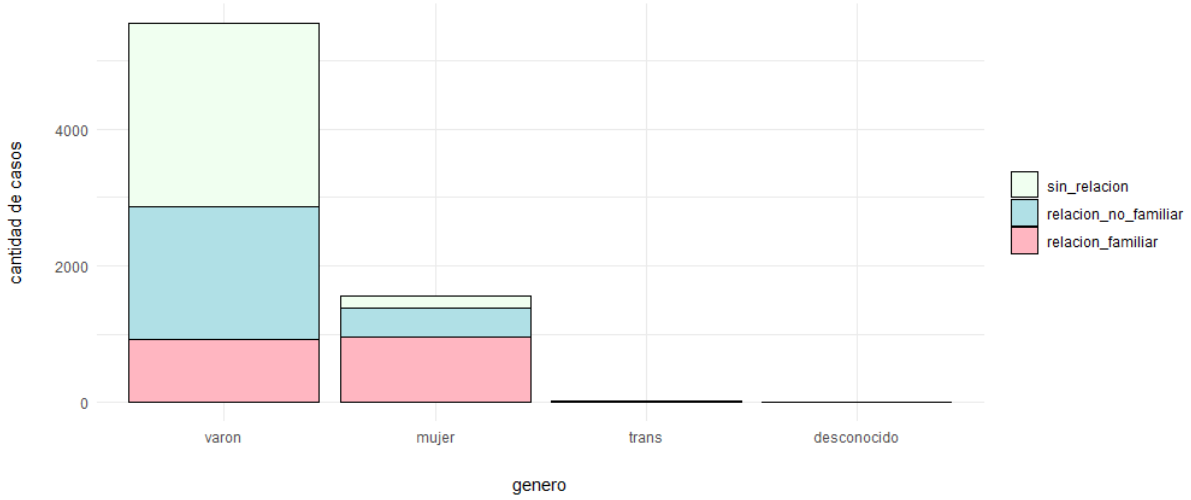

Figura 6: Gráfico de barras de la variable *genero* con respecto al vinculo víctima-inculpado.

En cuanto al genero de las víctimas, se observa en la Figura 6 que en la mayoría de los casos es varón, en cuya categoría predominan los homicidios de vínculo **sin_relacion**. Por otro lado, más de la mitad de las mujeres asesinadas son víctimas de algún familiar. A su vez, la cantidad total de homicidios con vínculo **relacion_familiar** son aproximadamente iguales en varones que en mujeres. En relación a la distribución de casos por genero de los otros vínculos, ambos tienen ampliamente mayor cantidad en la categoría **varón** por sobre **mujer**. Por último, los homicidios hacia las personas trans representan una cantidad casi insignificante de casos.

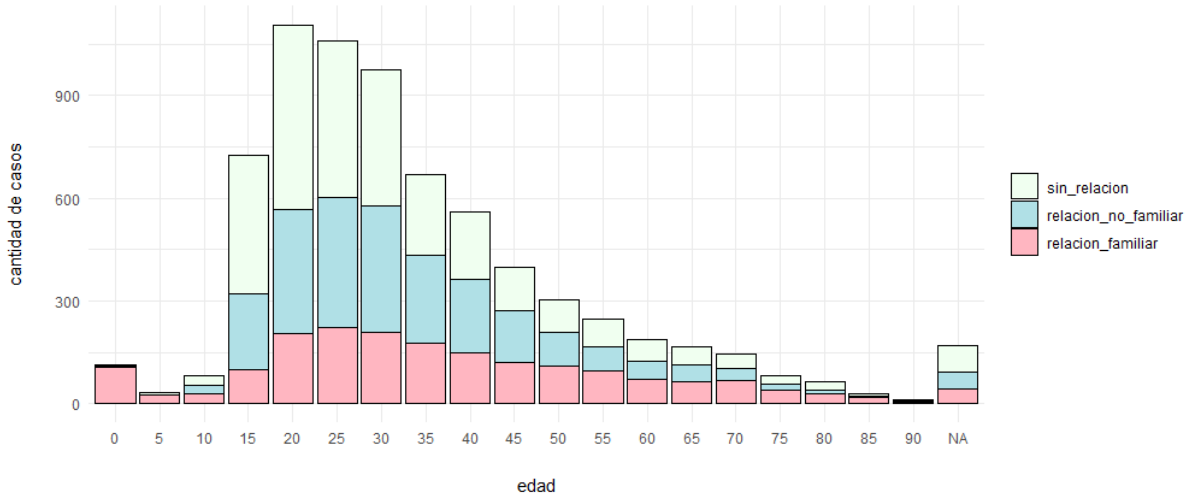

Figura 7: Gráfico de barras de la variable *edad* con respecto al vinculo víctima-inculpado.

La edad de las víctimas se encuentra concentrada principalmente entre los 15 y los 35 años, con el rango etario de los 20 años siendo la edad con más víctimas (Figura 7). A partir de los 20 años, siempre disminuye la cantidad de homicidios registrados a medida que la edad es mayor. Cabe destacar que en homicidios de criaturas menores a los 10 años, prácticamente la totalidad de los vínculos son de clase **relación_familiar**. A partir de los 65 años, el porcentaje de homicidios por familiares es cada vez mayor sobre el total de los casos, mientras que entre los 15 y los 30 años, la mayor parte de los vínculos son **sin_relacion**.

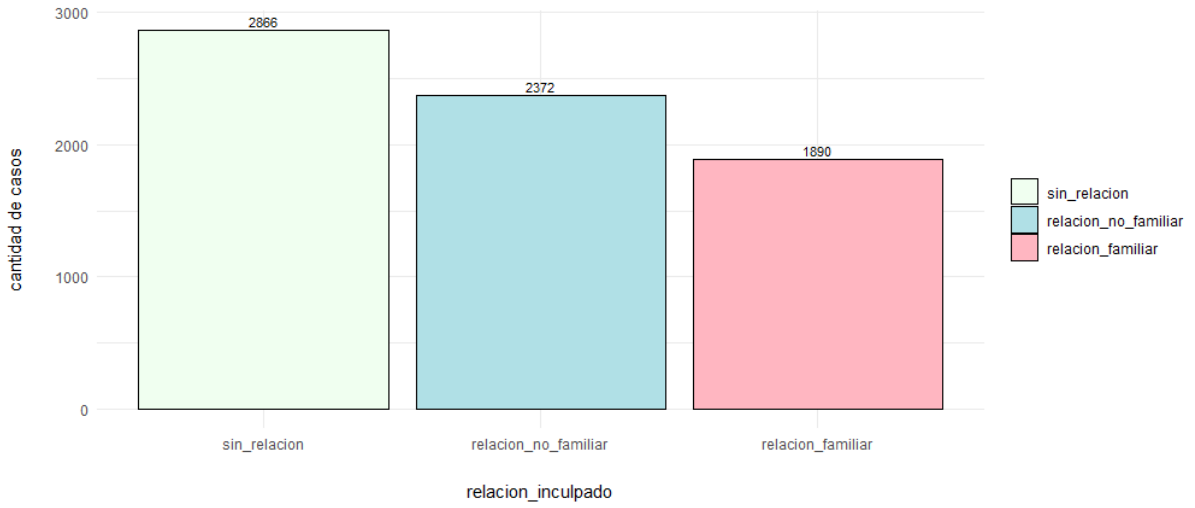

Figura 8: Gráfico de barras de la variable *relacion_inculpado* con respecto al vinculo víctima-inculpado.

Por último, vemos en la Figura 8 que la mayor cantidad de casos son de vínculo **sin_relacion**, seguido por **relacion_no_familiar** y en menor medida, por **relacion_familiar**. Sin embargo, las tres categorías presentan una cantidad de casos considerable.

5.2. Modelo explicativo

Los resultados se reflejan en la Tabla 8. La primera columna indica la variable que se está analizando y el intercepto, que refleja el modelo con todas las categorías de referencia mencionada en la sección 4.3.1. La segunda y la tercera columna representan los coeficientes β_k obtenidos a partir del modelo logístico multinomial para cada variable y en cada categoría de vínculo. A la derecha del valor obtenido para cada coeficiente se aclara la significancia del mismo, de acuerdo a los códigos explicados al pie de la tabla. Las últimas tres columnas indican el promedio de los efectos marginales de la probabilidad en cada variable de cada una de las categorías.

	Estimación		**Efectos marginales en la probabilidad**		
Variable	**Relación Familiar**	**Relación No Familiar**	p_1	p_2	p_3
(Intercept)	-1.0189 ***	-0.1020			
poblacion	0.0138 *	0.0296 ***	-0.0000	0.0062	-0.0062
fin_de_semana	-0.1656 ·	-0.1451 *	-0.0130	-0.0242	0.0372
hogar	1.9727 ***	0.3297 ***	0.2434	-0.0464	-0.1970
arma_blanca	-0.1167	0.2847 *	-0.0337	0.0757	-0.0419
arma_fuego	-1.0433 ***	-0.2570 *	-0.1235	0.0047	0.1187
delito	-2.8440 ***	-2.1780 ***	-0.2426	-0.3413	0.5840
mujer	2.0934 ***	1.0461 ***	0.2140	0.1177	-0.3317
edad	0.0273	0.0399 ·	0.0011	0.0078	-0.0089

Nota: Categoría de referencia: sin_relacion. Significancia: $^{***}p < 0.001$, $^{**}\ p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$
p_1: prob. de relación familiar, p_2: prob. de relación no familiar, p_3: prob. de sin relación.

Tabla 8: Modelo multinomial logístico.

En cuanto a la población, ambos coeficientes son significativos. De aquí en adelante, cada vez que se habla de los efectos marginales de la probabilidad en cada variable, se habla del promedio entre todas las observaciones para cierta variable y cierta categoría, por lo explicado en la sección 4.3.1. Por cada 100 000 habitantes, hay 0,06 % más de probabilidades de que sea causado por un conocido no familiar, y un 0,06 % menos de probabilidades de que sea causado por un desconocido. Por lo tanto, a mayor cantidad de habitantes en un centro urbano, menor probabilidad hay de que ocurran homicidios por parte de desconocidos. Los resultados son algo sorprendentes, porque sería lógico creer que en grandes ciudades es más frecuente el homicidio por parte de desconocidos. Sin embargo, analizando el gráfico de frecuencias de la variable *poblacion* con respecto a las categorías de vínculo (figura 1), se observa que la Ciudad de Buenos Aires (la de mayor población por diferencia) tiene un porcentaje menor de homicidios del tipo *sin_relacion* que de los otros dos tipos de vínculo. Esto no es del todo congruente con la tendencia general que se observa en la figura, donde el porcentaje de homicidios *sin_relacion* parece ir aumentando (o a lo sumo, mantenerse constante) a medida que aumenta la población. Sería prudente hacer un análisis similar, tratando a la Ciudad de Buenos Aires como un caso aislado. De todos modos el efecto es prácticamente nulo, por lo que no parece ser crítico su impacto en el modelo.

La variable *fin_de_semana* tiene coeficientes menos significativos que *poblacion*, con el coeficiente relacionado a *relacion_familiar* con un p-value entre 0,05 y 0,1. Si los consideramos, el modelo indica que los homicidios ocurridos durante el fin de semana tienen un 3,7 % más de probabilidades de ser causados por un desconocido. Las probabilidades de que la relación sea familiar o no familiar disminuyen en un 1,3 % y un 2,4 % respectivamente. Al ser variaciones poco considerables y los coeficientes poco significativos, no tiene demasiado sentido su interpretación.

Respecto al lugar del homicidio, los resultados son considerablemente más valiosos. Los coeficientes del modelo para la variable *hogar* son ambos muy significativos, con un p-value menor a 0,001. Los homicidios que ocurren en el hogar tienen un 24,3 % más de probabilidades de presentar un vínculo familiar. Esto es concordante con lo esperado: los homicidios familiares suelen ocurrir en el hogar. A su vez, disminuye mucho (19,7 %) las probabilidades de que el homicida sea desconocido, indicando que este tipo de homicidio suele ocurrir ya sea en la vía pública o en lugares como descampados, baldíos o similar. Los resultados además concuerdan con lo analizando en la figura 3.

El uso de arma blanca, por su parte, implica un aumento del 7,57 % de probabilidades de relación no familiar. El arma de fuego, en cambio, es un indicador bastante claro de homicidio por parte de un desconocido, con un 11,9 % de aumento en la probabilidad. Además el uso de arma de fuego se desasocia con un asesino familiar, con un 12,3 % menos de probabilidades.

Los resultados arrojados por la variable *delito* también son muy significativos y claros. El hecho de que el homicidio ocurra a la vez que un delito (ya sea robo, violación u otro), aumenta en un 58,4 % las probabilidades de que se trate de un desconocido. Es, por diferencia, la variable con mayor efecto marginal. Entre los efectos causados a las relaciones familiares y las no familiares, resulta menos probable que el homicida sea no familiar (34,1 %) que familiar (24,3 %) en el caso de delito. En general, los homicidios familiares y por conocidos nos familiares son por motivos diferentes que el robo. Es muy poco probable que un alguien asesine a un familiar para luego cometer un robo o violación. Sin embargo, como indica el gráfico de la variable *delito* (figura 5), existe una alta proporción de los homicidios sin delito que corresponden a desconocidos. Esto implica que no todos los desconocidos asesinan meramente para causar un robo o violación. Un delito en el homicidio implica casi siempre un desconocido, pero no viceversa.

En cuanto al género de la víctima, el hecho de que sea una mujer implica un significativo aumento del 21,4 % de probabilidades de que el homicida sea un familiar y un 11,8 % de que sea un conocido. Se demuestra que los homicidios a manos de desconocidos son un indicador de que la víctima probablemente fuera hombre y no mujer, con una baja del 33,2 % de las probabilidades en la categoría *sin_relacion*. En cambio, cuando existe algún vínculo entre víctima y victimaria, es más probable que se trate de una mujer. Los resultados también se condicen con lo comentado en el análisis exploratorio, donde se veía que cuando la víctima era mujer, era muchísimo más alto el porcentaje de homicidios a manos de familiares que de desconocidos.

Por último, la edad no presenta resultados demasiado valiosos para nuestro análisis. Los coeficientes no son significativos, e incluso si lo fueran, los efectos marginales en la probabilidad son menores al 1 % cada 10 años en los 3 casos.

5.3. Modelo predictivo

Los resultados de los modelos predictivos se basan en tres elementos. En primer lugar se detallan los indicadores obtenidos para cada combinación de hiperparámetros posibles. En segundo lugar se presenta la matriz de confusión del modelo final elegido. En tercer y último lugar se muestran los indicadores resultantes del modelo final, calculados a partir de lo explicado en la sección 4.4.2.

Light Gradient Boosting Machine

Los indicadores de cada combinación de hiperparámetros fueron ordenados de mayor a menor. Se mantuvo el índice utilizado para identificar cada combinación de hiperparámetros posible, el cual fue definido en la sección 4.4.2. Se presentan los resultados de los 30 mejores modelos en la Tabla 9.

#	**max_prof**	**n_arb**	**eta**	AG	F_1	κ
20	10	50	0.050	0.6643438	0.6566075	0.48330810
22	3	100	0.050	0.6562137	0.6535630	0.47775838
9	2	100	0.100	0.6573751	0.6535508	0.47597591
24	10	100	0.050	0.6620209	0.6530104	0.47872247
23	5	100	0.050	0.6608595	0.6526513	0.48045282
7	5	50	0.100	0.6596980	0.6519602	0.47785781
5	2	50	0.100	0.6527294	0.6516895	0.47263073
19	5	50	0.050	0.6550523	0.6502547	0.47445296
21	2	100	0.050	0.6504065	0.6490728	0.46884857
4	10	20	0.100	0.6550523	0.6471530	0.46900191
6	3	50	0.100	0.6492451	0.6467605	0.46697907
36	10	100	0.010	0.6550523	0.6416853	0.46697975
10	3	100	0.100	0.6457607	0.6416345	0.45756442
18	3	50	0.050	0.6399535	0.6408899	0.45599403
11	5	100	0.100	0.6492451	0.6406156	0.46136446
3	5	20	0.100	0.6434379	0.6398251	0.45688502
8	10	50	0.100	0.6492451	0.6392834	0.45901107
16	10	20	0.050	0.6480836	0.6332759	0.45568408
12	10	100	0.100	0.6434379	0.6328578	0.45138634
2	3	20	0.100	0.6318235	0.6320059	0.44503004
17	2	50	0.050	0.6248548	0.6287662	0.43151937
15	5	20	0.050	0.6387921	0.6226663	0.44539512
35	5	100	0.010	0.6376307	0.6213193	0.44347696
1	2	20	0.100	0.6167247	0.6195609	0.41989306
32	10	50	0.010	0.6236934	0.5977160	0.40658205
34	3	100	0.010	0.5981417	0.5976000	0.38790882
14	3	20	0.050	0.5981417	0.5974554	0.38761440
31	5	50	0.010	0.5934959	0.5499817	0.34626273
13	2	20	0.050	0.5412311	0.5259635	0.27666756
33	2	100	0.010	0.5412311	0.5259635	0.27666756

Tabla 9: Indicadores para cada combinación posible de hiperparámetros.

La combinación de hiperparámetros con mejores resultados fue la número 20, que incluye una *máxima profundidad* de 10, una *cantidad de árboles* de 50 y una *tasa de aprendizaje* de 0,050. Los modelos siguientes obtienen indicadores relativamente similares a los del primero pero luego van empeorando significativamente. Los hiperparámetros de la combinación 20 fueron utilizados para realizar el modelo final de *LGBM* y sus predicciones fueron:

		Clase real		
		relacion_familiar	**relacion_no_familiar**	**sin_relacion**
Clase predicha	**relacion_familiar**	151	58	25
	relacion_no_familiar	51	149	73
	sin_relacion	16	71	265

Tabla 10: Matriz de confusión para el modelo *LGBM*.

A partir de los resultados expuestos en la Tabla 10 se calcularon los indicadores de ajuste correspondientes.

$$AG = 0,6577416$$
$$F_1 = 0,6500784$$
$$\kappa = 0,4773458$$

eXteme Gradient Boosting

Los resultados de este algoritmo fueron tratados de igual manera que los anteriores.

#	**max_prof**	**n_arb**	**eta**	AG	F_1	κ
7	5	50	0.100	0.6666667	0.6614719	0.4919038
9	2	100	0.100	0.6620209	0.6598266	0.4854698
23	5	100	0.050	0.6573751	0.6526782	0.4778053
10	3	100	0.100	0.6550523	0.6524804	0.4734335
6	3	50	0.100	0.6515679	0.6506859	0.4707954
22	3	100	0.050	0.6480836	0.6479029	0.4653160
11	5	100	0.100	0.6492451	0.6445230	0.4638700
3	5	20	0.100	0.6411150	0.6404805	0.4546209
19	5	50	0.050	0.6399535	0.6384654	0.4525428
8	10	50	0.100	0.6434379	0.6325228	0.4504625
24	10	100	0.050	0.6399535	0.6300650	0.4450336
15	5	20	0.050	0.6271777	0.6292263	0.4354133
21	2	100	0.050	0.6248548	0.6273817	0.4319548
5	2	50	0.100	0.6236934	0.6263940	0.4308030
35	5	100	0.010	0.6236934	0.6257630	0.4298048
18	3	50	0.050	0.6236934	0.6255709	0.4334284
12	10	100	0.100	0.6364692	0.6254570	0.4410714
39	5	20	0.001	0.6248548	0.6250896	0.4305695
4	10	20	0.100	0.6329849	0.6236720	0.4358547
43	5	50	0.001	0.6236934	0.6235047	0.4287399
47	5	100	0.001	0.6236934	0.6235047	0.4287399
16	10	20	0.050	0.6306620	0.6222204	0.4324748
20	10	50	0.050	0.6318235	0.6205802	0.4320982
2	3	20	0.100	0.6190476	0.6205589	0.4280541
31	5	50	0.010	0.6190476	0.6202220	0.4221701
36	10	100	0.010	0.6271777	0.6175524	0.4257460
48	10	100	0.001	0.6248548	0.6162717	0.4215335
17	2	50	0.050	0.6132404	0.6145068	0.4213710
32	10	50	0.010	0.6213705	0.6126553	0.4169974
28	10	20	0.010	0.6213705	0.6116078	0.4163027

Tabla 11: Indicadores para cada combinación posible de hiperparámetros.

En esta oportunidad, la combinación de hiperparámetros con mejores resultados fue la número 7, que incluye una *máxima profundidad* de 5, una *cantidad de árboles* de 50 y una *tasa de aprendizaje* de 0,100. Los modelos siguientes obtienen indicadores relativamente similares a los del primero y al contrario que el algoritmo anterior, los resultados no empeoran drásticamente.

Los hiperparámetros de la combinación 7 fueron utilizados para realizar el modelo final de *XGBoost* y sus predicciones fueron:

		Clase real		
		relacion_familiar	**relacion_no_familiar**	**sin_relacion**
Clase predicha	**relacion_familiar**	162	59	29
	relacion_no_familiar	42	151	89
	sin_relacion	14	68	245

Tabla 12: Matriz de confusión para el modelo XGBoost.

A partir de los resultados expuestos en la Tabla 12 se calcularon los indicadores de ajuste correspondientes.

$$AG = 0,6495925$$
$$F_1 = 0,6472461$$
$$\kappa = 0,4682963$$

Elección del modelo predictivo final

Habiendo elegido los dos mejores modelos predicitvos, uno para cada algoritmo de *Gradient Boosting*, se puede avanzar en la elección del modelo predictivo que mejor se ajuste a la realidad de los homicidios dolosos en Argentina. Aunque ambos modelos obtengan prácticamente los mismos indicadores de ajuste, los del modelo *LGBM* con la combinación #20 de hiperparámetros son levemente mayores a los del modelo *XGBoost* con la combinación #7 de hiperparámetros. Se podrían utilizar cualquiera de los dos algoritmos propuestos si se seleccionan correctamente los hiperparámetros, ya que ambos demostraron que pueden obtener resultados relativamente semejantes. Sin embargo, en esta oportunidad, se selecciona como modelo predictivo final al algoritmo de ***Light Gradient Boosting Machine* con una *máxima profundidad* de 10, una *cantidad de árboles* de 50 y una *tasa de aprendizaje* de 0,050.** Estos hiperparámetros denotan que el modelo posee una cantidad moderada de árboles (el intermedio entre los tres posibles: 20, 50 y 100) con una profundidad elevada (la mayor entre las 4 posibles: 2, 3, 5 y 10). A su vez, la tasa de aprendizaje seleccionada es de un rango medio. Se puede deducir de esta combinación de hiperparámetros, que el modelo aprende con fuerza desde cada árbol.

Los indicadores de ajuste del modelo predictivo final obtienen resultados aceptables pero con espacio a mejora. En primer lugar, el acierto general alcanzó un valor de $65,8\,\%$, lo que indica que el modelo puede predecir correctamente el $65,8\,\%$ de los casos. Aunque idealmente se esperaría obtener un acierto general del $100\,\%$, para un primer estudio y con los datos disponibles al momento, acertar en dos tercios de los casos parecería representar una gran ayuda como primer indicio en un proceso de investigación penal.

Por otro lado, el F_1-score obtuvo un valor de $65\,\%$, la cual coincide con el acierto general. Esto sugiere que la predicciones correctas representadas por el acierto general no están dominadas por una sola clase sino que se reparten equilibradamente entre los tres factores de la variable respuesta. A su vez, se puede interpretar que la precisión y la sensibilidad de cada categoría no tienen grandes discrepancias, lo que implica que los aciertos en las predicciones de una clase son similares a los aciertos en los casos reales de la misma clase. Estas interpretaciones se logran visualizar en la matriz de confusión del modelo, estructurada en la Tabla 10. Los casos en los

que el modelo predijo erróneamente que el vínculo era **relacion_familiar**, cuando en realidad el vínculo era **relacion_no_familiar**, fueron 58; un valor muy cercano a los 51 casos en los que el modelo predijo erróneamente que el vínculo era **relacion_no_familiar**, cuando en realidad el vínculo era **relacion_familiar**. Lo mismo sucede para **sin_relacion** y **relacion_no_familiar** donde 25 y 16 casos fueron clasificados de manera equívoca una clase con la otra. En el tercer cruce, **relacion_no_familiar** con **sin_relacion**, la cantidad de casos son aún más semejantes, alcanzando valores de 71 y 73 predicciones incorrectas. En consecuencia, se puede suponer que el rendimiento del modelo no está sesgado por ninguna de las tres clases de la variable respuesta, lo que representa equilibrio y consistencia en el modelo planteado.

En último lugar, el índice Kappa de Cohen obtuvo un $47,7\,\%$ para el modelo predictivo final. La fuerza de concordancia según este indicador se puede clasificar como lo hicieron Landis y Koch (1977):

Estadístico *kappa*	**Concordancia**
<0.00	Pobre
0.00 - 0.20	Ligera
0.21 - 0.40	Aceptable
0.41 - 0.60	Moderada
0.61 - 0.80	Substancial
0.81 - 1.00	Casi Perfecta

Tabla 13: Fuerza de concordancia según *kappa*

El valor obtenido entraría en el rango entre 0.41 y 0.60, lo cual denota una fuerza de concordancia moderada. El índice elimina el efecto del azar, por lo cual es más "exigente" obtener valor elevados. Teniendo esto en cuenta, y que el resultado es más que aceptable, se puede considerar que el modelo ha aprendido patrones relevantes a partir de la base de datos provista y que es capaz de predecir significativamente a partir de estos patrones y no aleatoriamente.

6. Conclusiones

Los objetivos de este trabajo eran describir y demostrar las tendencias de los homicidios dolosos en Argentina y desarrollar un modelo capaz de predecir el vínculo entre la víctima y el homicida (familiar, conocido o desconocido).

En relación al primer objetivo, los homicidios en Argentina presentan ciertos patrones fácilmente reconocibles. La mayor cantidad de los casos son ocasionados por personas sin relación con la víctima (40 %). Estos incidentes se caracterizan por ocurrir en conjunto con otro delito (robo, violación u otro) y suelen ser causados por el uso de un arma de fuego. A su vez, las víctimas de estos homicidios son en su mayoría de genero masculino, probablemente por estar más expuestos a situaciones de riesgo fuera del hogar. Por el contrario, los homicidios entre familiares (26, 5 %) representan la categoría con menor cantidad de casos. Los mismos suelen ocurrir, lógicamente, en los hogares y en ausencia de cualquier otro delito. En estos casos, no se suelen utilizar armas de fuego para acabar con la vida de las víctimas, las cuales suelen ser mujeres. En último lugar y con un 33, 5 % de los casos, se encuentran los homicidios ocasionados por un conocido fuera de la familia. Este tipo de asesinato es similar al ocurrido entre familiares, pero en estos casos en particular, no hay una relación tan estrecha con que el hecho ocurra dentro del hogar. Además, es la categoría con menor presencia de otro delito al momento del homicidio. De manera similar que los homicidios familiares, las mujeres son más propensas que los hombres a ser asesinadas por una persona conocida. Cabe destacar que los análisis realizados en el modelo exploratorio son validados por los resultados del modelo explicativo, lo que hace más robusto al estudio y su capacidad de explicar la realidad.

Por otro lado, con respecto al segundo objetivo, el modelo predictivo seleccionado obtuvo un acierto general de 65, 8 %, un F_1-score de 65 % y un índice Kappa de 47, 7 %. Los resultados alcanzados se pueden considerar moderadamente buenos, lo que representa un ajuste aceptable del modelo con la realidad. Se pudo comprobar que el mismo no se comporta aleatoriamente y está balanceado respecto a las tres categorías de la variable respuesta. En consecuencia, se interpreta que el modelo aprende correctamente los patrones aportados por la base de datos provista y que es capaz de predecir en función a estos.

En conclusión, los objetivos planteados fueron cumplidos satisfactoriamente. El análisis exploratorio y explicativo es capaz de detectar y describir tendencias que permiten entender mejor la situación homicida en el país. Los resultados del modelo explicativo son significativos y congruentes a la realidad. El desarrollo el modelo predictivo demostró obtener resultados positivos. Los indicadores del mismo revelan que sería capaz de dar una buena primera aproximación o primeros indicios en una investigación penal. Sin embargo, con el fin de mejorar los índices de este modelo se podría considerar: ampliar la base de datos incluyendo más años, procurar no tener datos faltantes y/o incluir más variables explicativas como la calidad de vida, entre otros. A su vez, sería interesante replicar este estudio para otros países y comparar los resultados obtenidos. Más allá de los objetivo cumplidos, este trabajo demuestra que a partir del uso de la tecnología se puede ir en búsqueda de un sistema penal y judicial más trasparente y eficiente.

7. Bibliografía

Agresti, A. (2007). An introduction to categorical data analysis (2nd ed.). Wiley-Interscience.

Ajzenman, N., y Jaitman, L. (2016). Crime Concentration and Hot Spot Dynamics in Latin America.

Alan B. Cantor (1995). Sample-Size Calculations for Cohen's Kappa. American Psychological Association.

Anny Kim, Julien Chopin & Eric Beauregard (2024). Tell Me What You Do, I'll Tell You Who You Are: Predicting Offender-Victim Relationships in Sexual Homicide. Victims & Offenders.

Bergman, M., Ambrogi, J., Bruno, M., Croci, G., Prueger, E. A. (2023). Estudio sobre homicidios en Argentina: un análisis del periodo 2001–2021. Universidad de Tres de Febrero.

Buuren, S., Groothuis-Oudshoorn, C. (2011). MICE: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software.

Cao, L., Hou, C., Huang, B. (2008). Correlates of the victim-offender relationship in homicide. International Journal of Offender Therapy and Comparative Criminology, 52(6), 658–672.

Decker, S. H. (1993). Exploring victim-offender relationships in homicide: The role of individual and event characteristics. Justice Quarterly, 10(4), 585–612.

Drawdy, S. M., Myers, W. C. (2004). Homicide victim/offender relationship in Florida Medical Examiner District 8. Journal of Forensic Sciences, 49(1), 150–154.

Florek, P., & Zagdański, A. (2023). Benchmarking state-of-the-art gradient boosting algorithms for classification.

Haddouchi, M., & Berrado, A. (2024). A survey and taxonomy of methods interpreting random forest models.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

Heymann, P. B. (1985). Understanding Criminal Investigations. Harvard Journal on Legislation, 22, 315.

INDEC. (2023). Total del país. Viviendas particulares y población en viviendas particulares, según área de gobierno local. Buenos Aires, Argentina.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. Biometrics, 33(1), 159.

López de Ullibarri Galparsoro, I., Pita Fernández, S. (2001). Medidas de concordancia: el índice de Kappa. Unidad de Epidemiología Clínica y Bioestadística, Complexo Hospitalario-Universitario Juan Canalejo, A Coruña (España).

Manjarrés, J., Newton, C., & Cavalari, M. (2025, May 16). Balance de InSight Crime de los homicidios en 2024. InSight Crime. https://insightcrime.org/es/noticias/balance-insight-crime-homicidios-2024/

Ministerio de Seguridad de la Nación. Dirección Nacional de Estadística Criminal. (2024, 19 de diciembre). Homicidios dolosos. Sistema de Alerta Temprana. Estadísticas criminales en la República Argentina. [Conjunto de datos]. Recuperado el 28 de mayo de 2025, de https://www.argentina.gob.ar/seguridad/estadisticascriminales

Nagin, D. (1998). Criminal Deterrence Research at the Outset of the Twenty-First Century. In M. Tonry (Ed.), Crime and Justice: A Review of Research (Vol. 23). University of Chicago Press.

Oficina de las Naciones Unidas contra la Droga y el Delito. (2023). Estudio Mundial sobre el Homicidio 2023. https://dataunodc.un.org/content/country-list

Paternoster, R. (2010). How Much Do We Really Know about Criminal Deterrence? Journal of Criminal Law and Criminology, 100(3), 765–823.

Ramos, L. A., Olmedo, M. C., & Ditz, Y. A. (2022). Modelos criminológicos aplicados a la criminalidad de robos, estupefacientes y homicidios ocurridos en la Ciudad de Córdoba. In III Congreso Internacional de Victimología (Ensenada, October 28–30, 2021).

Riedel, M. (1988). Stranger Violence: Perspectives, Issues, and Problems. Journal of Criminal Law & Criminology, 78, 223.

Rigano, C. (2019). "Using Artificial Intelligence to Address Criminal Justice Needs."NIJ Journal, 280, January.

Van Buuren, S. (2018). Flexible Imputation of Missing Data (2nd ed.). Chapman and Hall/CRC.

Wolfgang, M. (1958). Patterns in Criminal Homicide. University of Pennsylvania Press.

Zhang, R. (2021, November 23). Guide to the gradient boosting algorithm. DataCamp. https://www.datacamp.com/tutorial/guide-to-the-gradient-boosting-algorithm

8. Apéndice

Código de R utilizado para el análisis exploratorio

```
install.packages("ggplot2")
install.packages("readr")
install.packages("dplyr")
library(ggplot2)
library(readr)
library(dplyr)

#POBLACION
ggplot(data = datos, aes(x = poblacion, fill = relacion_inculpado)) +
  geom_histogram(binwidth = 100000, color = "black") +
  scale_fill_manual(values = c(
    "relacion_familiar" = "lightpink",
    "relacion_no_familiar" = "powderblue",
    "sin_relacion" = "honeydew"
  )) +
  theme_minimal() +
  theme(legend.position = "right",
        axis.title.x = element_text(margin = margin(t = 20)),
        axis.title.y = element_text(margin = margin(r = 20))
  ) +
  labs(x = "poblacion", y = "cantidad de casos", fill = NULL) +
  scale_x_continuous(
    breaks = seq(0, 10000000, by = 500000),
    labels = scales::comma
  )


#DIA_SEM
datos$dia_sem <- factor(datos$dia_sem, levels = as.character(1:7))
ggplot(data = datos, aes(x = dia_sem, fill = relacion_inculpado)) +
  geom_bar(color = "black") +
  scale_fill_manual(values = c(
    "relacion_familiar" = "lightpink",
    "relacion_no_familiar" = "powderblue",
    "sin_relacion" = "honeydew"
  )) +
  theme_minimal() +
  theme(legend.position = "right",
        axis.title.x = element_text(margin = margin(t = 20)),
        axis.title.y = element_text(margin = margin(r = 20))
  ) +
  labs(x = "dia_sem", y = "cantidad de casos", fill = NULL)


#LUGAR
datos <- datos %>%
  mutate(lugar = if_else(is.na(lugar), "desconocido", lugar)) %>%
  mutate(lugar_filtrado = if_else(
    lugar %in% c("hogar", "via_publica"),
    lugar,
    if_else(lugar == "desconocido", "desconocido", "otro")
  ))
datos$lugar_filtrado <- factor(datos$lugar_filtrado,
levels = c("hogar", "via_publica", "otro","desconocido"))
ggplot(datos, aes(x = lugar_filtrado, fill = relacion_inculpado)) +
  geom_bar(color = "black") +
  scale_fill_manual(values = c(
    "relacion_familiar" = "lightpink",
    "relacion_no_familiar" = "powderblue",
```

```
    "sin_relacion" = "honeydew"
  )) +
  theme_minimal() +
  theme(legend.position = "right",
        axis.title.x = element_text(margin = margin(t = 20)),
        axis.title.y = element_text(margin = margin(r = 20))
  ) +
  labs(x = "lugar", y = "cantidad de casos", fill = NULL)



#ARMA
datos <- datos %>%
  mutate(arma = if_else(is.na(arma), "desconocido", arma)) %>%
  mutate(arma_filtrado = if_else(
    arma %in% c("fuego", "blanca", "objeto_contundente",
    "golpes", "ahorcamiento", "quemadura"),
    arma,
    if_else(arma == "desconocido", "desconocido", "otro")
  )) %>%
  mutate(arma_filtrado = factor(
    arma_filtrado,
    levels = c("fuego", "blanca", "objeto_contundente",
    "golpes", "ahorcamiento", "quemadura", "otro", "desconocido")
  ))

ggplot(datos, aes(x = arma_filtrado, fill = relacion_inculpado)) +
  geom_bar(color = "black") +
  scale_fill_manual(values = c(
    "relacion_familiar" = "lightpink",
    "relacion_no_familiar" = "powderblue",
    "sin_relacion" = "honeydew"
  )) +
  theme_minimal() +
  theme(legend.position = "right",
        axis.title.x = element_text(margin = margin(t = 20)),
        axis.title.y = element_text(margin = margin(r = 20))
  ) +
  labs(x = "arma", y = "cantidad de casos", fill = NULL)


#DELITO
datos <- datos %>%
  mutate(delito = if_else(is.na(delito), "desconocido", delito)) %>%
  mutate(delito = factor(delito, levels = c("no", "robo",
  "otro", "violacion", "desconocido")))
ggplot(data = datos, aes(x = delito, fill = relacion_inculpado)) +
  geom_bar(color = "black") +
  scale_fill_manual(values = c(
    "relacion_familiar" = "lightpink",
    "relacion_no_familiar" = "powderblue",
    "sin_relacion" = "honeydew"
  )) +
  theme_minimal() +
  theme(legend.position = "right",
        axis.title.x = element_text(margin = margin(t = 20)),
        axis.title.y = element_text(margin = margin(r = 20))
  ) +
  labs(x = "delito", y = "cantidad de casos", fill = NULL)


#GENERO
datos <- datos %>%
```

```
  mutate(genero = if_else(is.na(genero), "desconocido", genero)) %>%
  mutate(genero = factor(genero,
  levels = c("varon", "mujer", "trans", "desconocido")))
ggplot(data = datos, aes(x = genero, fill = relacion_inculpado)) +
  geom_bar(color = "black") +
  scale_fill_manual(values = c(
    "relacion_familiar" = "lightpink",
    "relacion_no_familiar" = "powderblue",
    "sin_relacion" = "honeydew"
  )) +
  theme_minimal() +
  theme(legend.position = "right",
        axis.title.x = element_text(margin = margin(t = 20)),
        axis.title.y = element_text(margin = margin(r = 20))
  ) +
  labs(x = "genero", y = "cantidad de casos", fill = NULL)


#EDAD
ggplot(data = datos, aes(x = factor(edad), fill = relacion_inculpado)) +
  geom_bar(color = "black") +
  scale_fill_manual(values = c(
    "relacion_familiar" = "lightpink",
    "relacion_no_familiar" = "powderblue",
    "sin_relacion" = "honeydew"
  )) +
  theme_minimal() +
  theme(legend.position = "right",
        axis.title.x = element_text(margin = margin(t = 20)),
        axis.title.y = element_text(margin = margin(r = 20))
  ) +
  labs(x = "edad", y = "cantidad de casos", fill = NULL)


#RELACION_INCULPADO
datos$relacion_inculpado <- factor(datos$relacion_inculpado,
levels = c("sin_relacion", "relacion_no_familiar", "relacion_familiar"))
ggplot(data = datos, aes(x = relacion_inculpado, fill = relacion_inculpado)) +
  geom_bar(color = "black") +
  scale_fill_manual(values = c(
    "relacion_familiar" = "lightpink",
    "relacion_no_familiar" = "powderblue",
    "sin_relacion" = "honeydew"
  )) +
  geom_text(stat = "count", aes(label = ..count..), vjust = -0.3, size = 3) +
  theme_minimal() +
  theme(legend.position = "right",
        axis.title.x = element_text(margin = margin(t = 20)),
        axis.title.y = element_text(margin = margin(r = 20))
  ) +
  labs(x = "relacion_inculpado", y = "cantidad de casos", fill = NULL)
```

Código de R utilizado para el modelado explicativo

```
library(mlogit)
library(caret)
library(foreign)

set.seed(0)

# Keep selected columns
cols <- c("poblacion", "fin_de_semana", "hogar", "arma", "delito","mujer", "edad",
  "relacion_inculpado")
data <- data[, cols]
data$fin\_de\_semana <- as.factor(data$fin_de_semana)
data$hogar <- as.factor(data$hogar)
data$arma <- as.factor(data$arma)
data$delito <- as.factor(data$delito)
data$mujer <- as.factor(data$mujer)
data$relacion_inculpado <- as.factor(data$relacion_inculpado)


# Relevel categorical variables
data$fin_de_semana <- relevel(data$fin_de_semana, ref = "0")
data$hogar   <- relevel(data$hogar, ref = "0")
data$arma    <- relevel(data$arma, ref = "sin_arma")
data$delito  <- relevel(data$delito, ref = "0")
data$mujer  <- relevel(data$mujer, ref = "0")
data$relacion_inculpado <- relevel(data$relacion_inculpado, ref = "sin_relacion")

# === Mlogit ===
data = mlogit.data(data,shape="wide",choice="relacion_inculpado")
model <- mlogit(relacion_inculpado ~ 0 | poblacion + fin\_de\_semana + hogar + arma
  + delito + mujer + edad , data = data,  reflevel = "sin_relacion")
summary(model)
```

Código de R utilizado para el modelado predictivo

Light Gradient Boosting Machine

```
# Install required packages if not already installed
if (!requireNamespace("lightgbm", quietly = TRUE)) install.packages("lightgbm",
repos = "https://cran.r-project.org")
if (!requireNamespace("foreign", quietly = TRUE)) install.packages("foreign")
if (!requireNamespace("caret", quietly = TRUE)) install.packages("caret")
if (!requireNamespace("data.table", quietly = TRUE)) install.packages("data.table")
if (!requireNamespace("stringi", quietly = TRUE)) install.packages("stringi")
if (!requireNamespace("dplyr", quietly = TRUE)) install.packages("dplyr")
if (!requireNamespace("MLmetrics", quietly = TRUE)) install.packages("MLmetrics")

library(MLmetrics)
library(dplyr)
library(lightgbm)
library(foreign)
library(caret)
library(data.table)
library(psych)

# Keep selected columns
cols <- c("poblacion", "mes", "dia", "dia_sem", "lugar", "arma", "delito",
"registro", "genero", "edad", "clase", "relacion_inculpado")
data <- data[, cols]
data$lugar <- as.factor(data$lugar)
data$arma <- as.factor(data$arma)
data$delito <- as.factor(data$delito)
data$registro <- as.factor(data$registro)
data$genero <- as.factor(data$genero)
data$clase <- as.factor(data$clase)
data$relacion_inculpado <- as.factor(data$relacion_inculpado)

# Inspect target
table(data$relacion_inculpado)

# Encode categorical variables with dummy variables
dummies <- model.matrix(~ lugar + arma + delito + registro + genero + clase,
data = data)[, -1]
features <- cbind(dummies, data[, c("poblacion", "mes", "dia", "dia_sem", "edad")])
target <- as.numeric(data$relacion_inculpado) - 1

# Train/Val/test split: 70%, 15%, 15%
set.seed(0)
train_index <- createDataPartition(target, p = 0.7, list = FALSE)
val_index <- createDataPartition(target[-train_index], p = 0.5, list = FALSE)
X_train <- features[train_index, ]
X_test <- features[-train_index, ][-val_index, ]
X_val <- features[-train_index, ][val_index, ]
y_train <- target[train_index]
y_test <- target[-train_index][-val_index]
y_val <- target[-train_index][val_index]

# Create lgb.Dataset
lgb_train <- lgb.Dataset(data = as.matrix(X_train), label = y_train)

# === LGBM Grid-Search ==== #

# Define grid search parameters
grid <- expand.grid(
  max_depth = c(2,3, 5, 10),
  num_iterations = c(20,50, 100),
```

```
  learning_rate = c(0.1, 0.05, 0.01, 0.001)
)


# Run Grid-Search
results <- data.frame()

for (i in 1:nrow(grid)) {
  params <- list(
    objective = "multiclass",
    num_class = length(unique(target)),
    metric = "multi_logloss",
    max_depth = grid$max_depth[i],
    num_iterations = grid$num_iterations[i],
    learning_rate = grid$learning_rate[i],
    verbosity = -1,
    seed=0,
    is_unbalance = TRUE #si el dataset esta desbalanceado
  )

  model <- lgb.train(params = params, data = lgb_train, verbose = -1)

  # Predict
  preds <- predict(model, as.matrix(X_val))
  pred_labels <- apply(matrix(preds, ncol = length(unique(target))),
  1, which.max) - 1

  # Indicators
  acc <- mean(pred_labels == y_val)
  f1a <- F1_Score(y_pred = pred_labels, y_true = y_val, positive = 0)
  f1b <- F1_Score(y_pred = pred_labels, y_true = y_val, positive = 1)
  f1c <- F1_Score(y_pred = pred_labels, y_true = y_val, positive = 2)
  f1 <- (f1a + f1b + f1c)/3
  kappa <- cohen.kappa(x = cbind(pred_labels, y_val))$kappa

  results <- rbind(results, data.frame(
    max_depth = params$max_depth,
    num_iterations = params$num_iterations,
    learning_rate = params$learning_rate,
    accuracy = acc,
    f1 = f1,
    kappa = kappa
  ))
}

# Print results
print(results[order(-results$f1), ])

# === LGBM Final Model ==== #

# Get final train dataset: train + val
X_train_val <- bind_rows(X_train, X_val)
y_train_val <- c(y_train, y_val)
lgb_train_val <- lgb.Dataset(data = as.matrix(X_train_val), label = y_train_val)

# Set params (best)
best_params <- list(
  objective = "multiclass",
  num_class = length(unique(target)),
  metric = "multi_logloss",
  max_depth = 10,
  num_iterations = 50,
  learning_rate = 0.05,
```

```
  verbosity = -1,
  seed=0,
  is_unbalance = TRUE

)

# Train LGBM
model <- lgb.train(params = best_params, data = lgb_train_val, verbose = -1)

# Predict
preds <- predict(model, as.matrix(X_test))
pred_labels <- apply(matrix(preds, ncol = length(unique(target))), 1, which.max) - 1

# Indicators
acc <- mean(pred_labels == y_test)
f1a <- F1_Score(y_pred = pred_labels, y_true = y_test, positive = 0)
f1b <- F1_Score(y_pred = pred_labels, y_true = y_test, positive = 1)
f1c <- F1_Score(y_pred = pred_labels, y_true = y_test, positive = 2)
f1 <- (f1a + f1b + f1c)/3
kappa <- cohen.kappa(x = cbind(pred_labels, y_test))$kappa
print(acc)
print(f1)
print(kappa)

# Confussion Matrix
table(Predicted = pred_labels, Actual = y_test)
```

eXtreme Gradient Boosting

```
# Install required packages if not already installed
if (!requireNamespace("xgboost", quietly = TRUE)) install.packages("lightgbm",
repos = "https://cran.r-project.org")
if (!requireNamespace("foreign", quietly = TRUE)) install.packages("foreign")
if (!requireNamespace("caret", quietly = TRUE)) install.packages("caret")
if (!requireNamespace("data.table", quietly = TRUE)) install.packages("data.table")
if (!requireNamespace("stringi", quietly = TRUE)) install.packages("stringi")
if (!requireNamespace("dplyr", quietly = TRUE)) install.packages("dplyr")
if (!requireNamespace("MLmetrics", quietly = TRUE)) install.packages("MLmetrics")

library(MLmetrics)
library(dplyr)
library(xgboost)
library(foreign)
library(caret)
library(data.table)
library(psych)

set.seed(0)

# Keep selected columns
cols <- c("poblacion", "mes", "dia", "dia_sem", "lugar", "arma",
"delito", "registro", "genero", "edad", "clase", "relacion_inculpado")
data <- data[, cols]
data$lugar <- as.factor(data$lugar)
data$arma <- as.factor(data$arma)
data$delito <- as.factor(data$delito)
data$registro <- as.factor(data$registro)
data$genero <- as.factor(data$genero)
data$clase <- as.factor(data$clase)
data$relacion_inculpado <- as.factor(data$relacion_inculpado)

# Inspect target
table(data$relacion_inculpado)

# Encode categorical variables with dummy variables
dummies <- model.matrix(~ lugar + arma + delito + registro + genero + clase,
data = data)[, -1]
features <- cbind(dummies, data[, c("poblacion", "mes", "dia", "dia_sem", "edad")])
target <- as.numeric(data$relacion_inculpado) - 1

# Train/Val/test split: 70%, 15%, 15%
train_index <- createDataPartition(target, p = 0.7, list = FALSE)
val_index <- createDataPartition(target[-train_index], p = 0.5, list = FALSE)
X_train <- features[train_index, ]
X_test <- features[-train_index, ][-val_index, ]
X_val <- features[-train_index, ][val_index, ]
y_train <- target[train_index]
y_test <- target[-train_index][-val_index]
y_val <- target[-train_index][val_index]

# Create XGBoost matrix
dtrain <- xgb.DMatrix(data = as.matrix(X_train), label = y_train)


# === XGBoost Grid-Search ==== #

# Define grid search parameters
grid <- expand.grid(
  max_depth = c(2,3, 5, 10),
  nrounds = c(20,50, 100),
```

```
  eta = c(0.1, 0.05, 0.01, 0.001)
)

# Run Grid-Search
results <- data.frame()

for (i in 1:nrow(grid)) {
  params <- list(
    objective = "multi:softmax",  # Multi-class classification
    num_class = length(unique(target)),
    max_depth = grid$max_depth[i],
    eta = grid$eta[i],
    eval_metric = "mlogloss",
    seed = 0
  )

  model <- xgb.train(params = params, data = dtrain, verbose = 0,
  nrounds = grid$nrounds[i])

  # Predict
  preds <- predict(model, as.matrix(X_val))

  # Indicators
  pred_labels <- preds
  acc <- mean(pred_labels == y_val)
  f1a <- F1_Score(y_pred = pred_labels, y_true = y_val, positive = 0)
  f1b <- F1_Score(y_pred = pred_labels, y_true = y_val, positive = 1)
  f1c <- F1_Score(y_pred = pred_labels, y_true = y_val, positive = 2)
  f1 <- (f1a + f1b + f1c)/3
  kappa <- cohen.kappa(x = cbind(pred_labels, y_val))$kappa

  results <- rbind(results, data.frame(
    max_depth = params$max_depth,
    nrounds = grid$nrounds[i],
    eta = params$eta,
    accuracy = acc,
    f1 = f1,
    kappa = kappa
  ))
}

# Print results
print(results[order(-results$f1), ])

# === XGBoost Final Model ==== #

# Get final train dataset: train + val
X_train_val <- bind_rows(X_train, X_val)
y_train_val <- c(y_train, y_val)

dtrain_val <- xgb.DMatrix(data = as.matrix(X_train_val), label = y_train_val)

# Set params (best from grid search)
best_params <- list(
  objective = "multi:softmax",  # Multi-class classification
  num_class = length(unique(target)),
  max_depth = 5,
  eta = 0.1,
  eval_metric = "mlogloss",
  seed = 0
)
```

```r
# Train XGBoost
model <- xgb.train(params = best_params, data = dtrain_val, verbose = 0,
nrounds = 50)

# Predict
preds <- predict(model, as.matrix(X_test))
pred_labels <- preds

# Indicators
acc <- mean(pred_labels == y_test)
f1a <- F1_Score(y_pred = pred_labels, y_true = y_test, positive = 0)
f1b <- F1_Score(y_pred = pred_labels, y_true = y_test, positive = 1)
f1c <- F1_Score(y_pred = pred_labels, y_true = y_test, positive = 2)
f1 <- (f1a + f1b + f1c)/3
kappa <- cohen.kappa(x = cbind(pred_labels, y_test))$kappa
print(acc)
print(f1)
print(kappa)

# Confusion Matrix
table(Predicted = pred_labels, Actual = y_test)
```