Todo list

Maestría en Explotación de Datos y Descubrimiento del Conocimiento

Universidad de Buenos Aires

Trabajo integrador

Victoria Colombo

fecha

Resumen

Índice

Introducción

La violencia sexual comprende una multiplicidad de conductas o intentos de conductas, que van desde actos hasta comentarios sexuales, dirigidos contra la sexualidad de otra persona de manera coercitiva. El trabajo con datos sobre violencia sexual presenta complicaciones porque los datos suelen ser escasos o presentar gran cantidad de faltantes (Ferris, 2002, p. 150). Uno de los motivos es que las víctimas o su entorno a menudo se rehúsan a denunciar o participar en encuestas sobre este tipo de agresiones, o proveen información incompleta. Esto puede deberse a la vergüenza y el estigma social frecuentemente asociado no solo con la violencia sexual sino con la sexualidad en general, pero también a la falta de acceso a la justicia, al temor a las represalias por parte de los agresores, o el temor a que la denuncia no sea creída (Murphy-Oikonen, McQueen, Miller, Chambers, and Hiebert, 2022). Otros posibles motivos para la escasez y/o mala calidad de los datos pueden ser la falta de vías adecuadas para recabar esta información, o la negligencia o desconocimiento de procedimientos adecuados por parte de oficiales de policía encargados de recibir denuncias. A pesar de las dificultades en la recolección de datos, diversos estudios a nivel mundial logran identificar patrones frecuentes en la violencia sexual. Para este trabajo, resultan relevantes dos de ellos: la mayoría de las víctimas son mujeres, mientras que los perpetradores suelen ser hombres (Ferris, 2002, p. 149; Contreras et al., 2016, p. 15); y en la mayoría de los casos, los agresores son personas conocidas por las víctimas, como parejas, exparejas u otros conocidos (García-Moreno et al., 2005, p. 9,Unicef et al., 2018, p. 22, Ferris, 2002, p. 151).

La clasificación de las identidades de género de víctimas y perpetradores es compleja. Por un lado, muchos estudios clasifican a las personas únicamente como hombres o mujeres, omitiendo las identidades de género disidentes.[1]. Por otro lado, aunque se reportan pocos casos de violencia sexual contra hombres cisgénero, es probable que estén subrepresentados debido a los prejuicios y estigmas sociales sobre la masculinidad que dificultan las denuncias y el acceso a la justicia para estas víctimas (Ferris, 2002, p. 149). Analizar esas complejidades excede a este trabajo de especialización. En mi análisis las categorías de género de víctimas, victimarios y llamantes se limitan a las registradas en el *dataset*: hombre, mujer, y transgénero, sin especificar si es un hombre o una mujer transgénero. Reconozco esto como una limitación no solo de mi trabajo sino también de los datos disponibles.

La recopilación, sistematización, y análisis de datos sobre violencia sexual por parte de los Estados es crucial para planificar y llevar adelante políticas efectivas de prevención, asistencia, y erradicación de la violencia sexual. En Argentina, si bien no hay un sistema estatal único y centralizado de este tipo de información, existen entidades judiciales y programas estatales que, además de ofrecer auxilio, asistencia y/o acceso a la justicia, recaban datos sobre violencia sexual, y mantienen un registro público de ellos. Unos de esos programas es Las Víctimas contra las Violencias.

Desde el año 2016, en el marco del programa Las Víctimas contra las Violencias, dependiente del Ministerio de Justicia de la Nación, la línea 137 funciona las 24 horas del día para solicitar asistencia en casos de violencia sexual o familiar[2]. El programa cuenta con equipos de intervención de abogadas, psicólogas, y trabajadoras sociales. Al

[1] Entre los estudios e informes consultados para este trabajo, solamente el *Relevamiento de fuentes secundarias de datos sobre violencia sexual* de la Unidad Fiscal Especializada en Violencia contra las Mujeres (UFEM) (2023) menciona identidades de género cuando especifica que la violencia sexual "afecta particularmente a las mujeres cis y personas LGBTI+" (p.7).

[2] Además, desde 2020 cuenta también con el canal de *Whatsapp* (54911) 3133-1000.

recibir una llamada solicitando asistencia se coordina el envío de equipos móviles para proveer a la víctima, en base a las necesidades del caso, de contención emocional, acompañamiento a un hospital y/o a radicar una denuncia, y/o a un lugar seguro donde pueda alojarse (Ministerio de Justicia de la República Argentina, 2022).
rever cómo sigue el programa ahora

Los registros de las llamadas a la línea 137 se encuentran digitalizados desde 2017 y están disponibles en el Portal de Datos Abiertos de la Justicia Argentina. Allí se encuentran publicados cuatro tipos de *datasets* por año: llamados e intervenciones domiciliarias por situaciones de violencia familiar, y llamados e intervenciones domiciliarias por situaciones de violencia sexual. Los registros no están exentos de los problemas frecuentes antes mencionados en los datos sobre violencia sexual, presentan información faltante de dos maneras: celdas vacías en el caso de las variables numéricas de edad, y respuestas NS/NC (no sabe-no contesta) en lugar de SI o NO en el resto de las variables categóricas. Teniendo en cuenta la clasificación de datos faltantes que se origina en Rubin (1976), los datos faltantes en el *dataset* de llamados son posiblemente del tipo *missing at random* (MAR) y *missing not at random/ non-ignorable missing data* (MNAR). Es decir, o bien los datos faltan por motivos que tienen que ver con otras variables (MAR), o bien el valor de los datos que faltan está relacionado con el motivo mismo por el que faltan (MNAR). En este trabajo analizo llamados para reportar violencia sexual a la línea 137 entre 2017 y 2021, e intento predecir valores faltantes de la variable "víctima convive con el agresor".
esto así declarado es una reverenda poronga

Datos

Obtención y limpieza

Para este trabajo descargué del Portal de Datos Abiertos mencionado arriba 5 *datasets* en formato *csv* de llamados a la línea 137 para solicitar asistencia por violencia sexual. Los archivos pertenecen, a razón de uno de por año, al período entre enero de 2017 y julio de 2021.

Una vez descargados, unifiqué los 5 archivos en un solo *dataset*. Para eso fue necesario realizar una primera limpieza destinada a dejar consistentes los distintos archivos en términos de cantidad y nombre de columnas[3]:

- Eliminé la variable *caso_id*, que solo existe a partir de 2020.
- Cambié el nombre de la variable *llamado_provincia_indec_id* en los *datasets* de 2017, 2018, y 2019 a su equivalente en 2020 y 2021: *llamado_provincia_id*.

El siguiente paso fue limpiar el *dataset* unificado de inconsistencias y errores de carga varios:

- Unifiqué para todas las variables pertinentes los valores *SI*, *NO*, y *NS/NC* dejándolos en mayúscula, ya que aparecían en distintos formatos: minúscula, mayúscula inicial, etc.

[3]La limpieza, normalización, y preprocesamiento del *dataset* y la aplicación de los métodos exploratorios y predictivos fueron realizados en Python

- Unifiqué en la variable *victima_vinculo_agresor* el valor *Ex pareja de la víctima* que aparecía también como *Ex pareja*, *Ex-pareja de la víctima* y *Expareja de la víctima*, otro tanto hice con *Pareja de la víctima* que presentaba variaciones similares.
- Unifiqué en *hecho_lugar* dos variaciones de una misma categoría: *Otra institución*, y *Otra Institución*, optando por la primera forma.
- Sustituí todos los valores *Sin datos* por *NS/NC* por considerarlos equivalentes.
- Quité espacios de comienzo y final de *strings* para solucionar problemas del tipo *Madre* $=/=$ *Madre*
- Convertí en la variable *llamante_vinculo* el valor *Vecino* a *Vecina/o*, ya que no necesariamente se refiere unívocamente a personas de género masculino.
- Unifiqué en *llamado_provincia* "Ciudad Autónoma de Buenos Aires" y "CABA" optando por "CABA".
- Corregí en *llamado_provincia* las instancias de "Santa Fé" a "Santa Fe".

Exploración

El *dataset* final unificado consta de 19143 observaciones y 54 variables, en su mayoría categóricas, que aportan información sobre la víctima, el agresor, la persona que llama para reportar el hecho, el contexto del hecho y el tipo de violencia sufrida. En el cuadro 1 se puede ver un detalle de las variables y su tipo.

Cuadro 1: Resumen de las variables.

Descriptor	**Tipo variable**	**Variable(s)**
Víctima	Cuantitativa	victima_edad
	Cualitativa	victima_genero, victima_nacionalidad, victima_discapacidad, victima_vinculo_agresor, victima_convive_agresor, victima_a_resguardo
Llamante	Cuantitativa	llamante_edad
	Cualitativa	llamante_genero, llamante_vinculo
Llamado	Ordinal	llamado_fecha_hora
	Cualitativa	caso_id, llamado_provincia llamado_provincia_id, caso_judicializado, hecho_lugar
Violencia sexual	Cualitativa	vs_violacion_via_vaginal, vs_violacion_via_anal, vs_violacion_via_oral, vs_tentativa_violacion, vs_tocamiento_sexual, vs_intento_tocamiento, vs_intento_violacion_tercera_persona, vs_grooming, vs_exhibicionismo, vs_amenazas_verbales_contenido_sexual, vs_explotacion_sexual, vs_explotacion_sexual_comercial, vs_explotacion_sexual_viajes_turismo, vs_sospecha_trata_personas_fines_sexuales, vs_existencia_facilitador_corrupcion_nnya, vs_obligacion_sacarse_fotos_pornograficas, vs_eyaculacion_partes_cuerpo, vs_acoso_sexual, vs_iniciacion_sexual_forzada_inducida, vs_otra_forma_violencia_sexual, vs_no_sabe_no_contesta
Otras violencias	Cualitativa	ofv_sentimiento_amenaza, ofv_amenazas_explicitas, ofv_violencia_fisica, ofv_intento_ahorcar, ofv_intento_quemar, ofv_intento_ahogar, ofv_amenaza_muerte, ofv_uso_sustancias_psicoactivas, ofv_intento_privacion_libertad, ofv_privacion_libertad, ofv_uso_arma_blanca, ofv_uso_arma_fuego, ofv_enganio_seduccion, ofv_intento_matar, ofv_uso_animal_victimizar, ofv_grooming, ofv_otra_forma_violencia, ofv_no_sabe_no_contesta

Las variables que describen la violencia sexual sufrida y otras formas de violencia reportadas pueden tomar los valores *SI* o *NO*, siendo este último el valor más común, como se aprecia más abajo en las figuras 1 y 2 que muestran la distribución de respuestas para violencia sexual y otras formas de violencia respectivamente. Es interesante notar el volumen de respuestas positivas de las categorías *vs_no_sabe_no_contesta* y *ofv_no_sabe_no_contesta*. Es decir, en gran cantidad de llamados se reporta una forma de violencia (sexual o no) sufrida, pero no se puede reportar qué forma. A lo largo de esta sección se observa esta prevalencia de respuestas de tipo *NS/NC* en casi todas las variables.

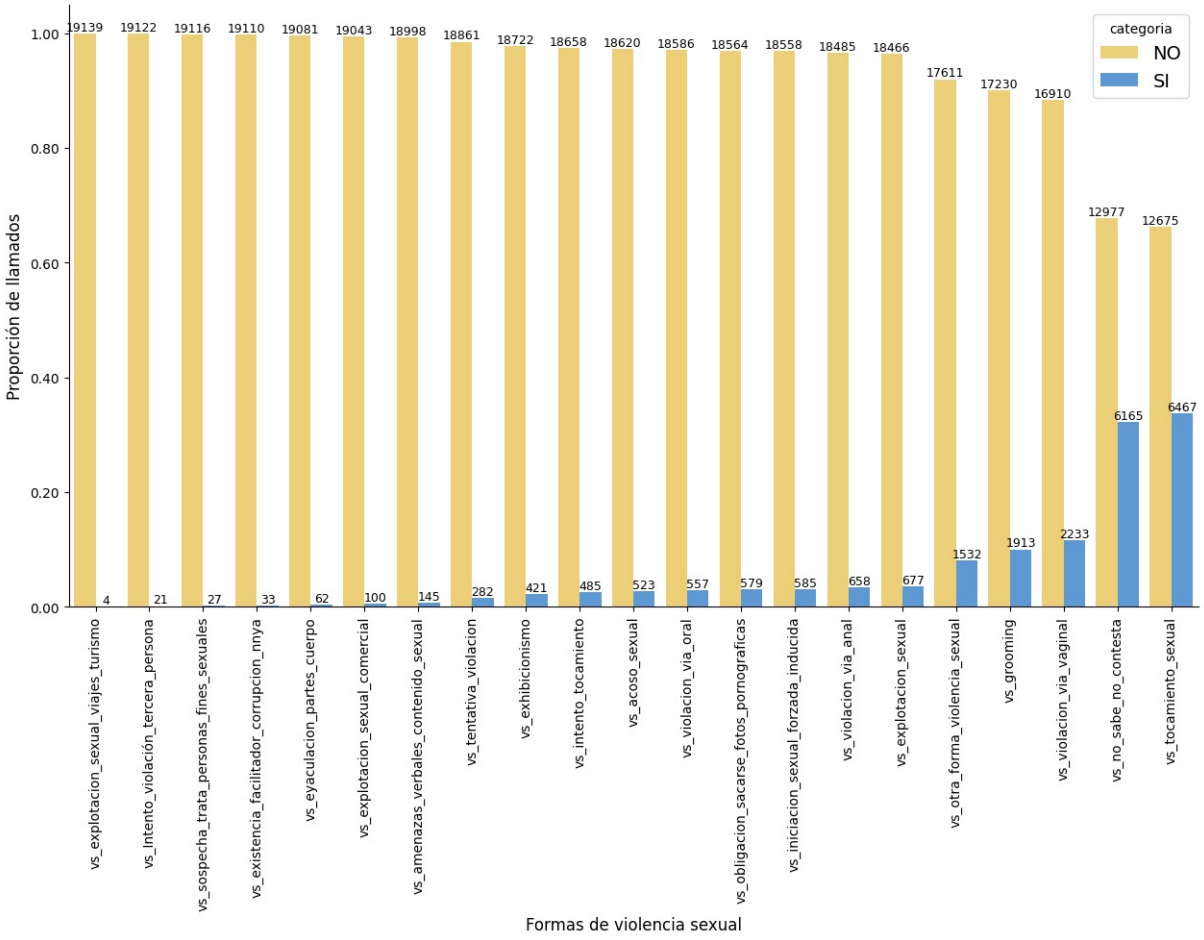

Figura 1: Tipos de violencia sexual reportada en los llamados.

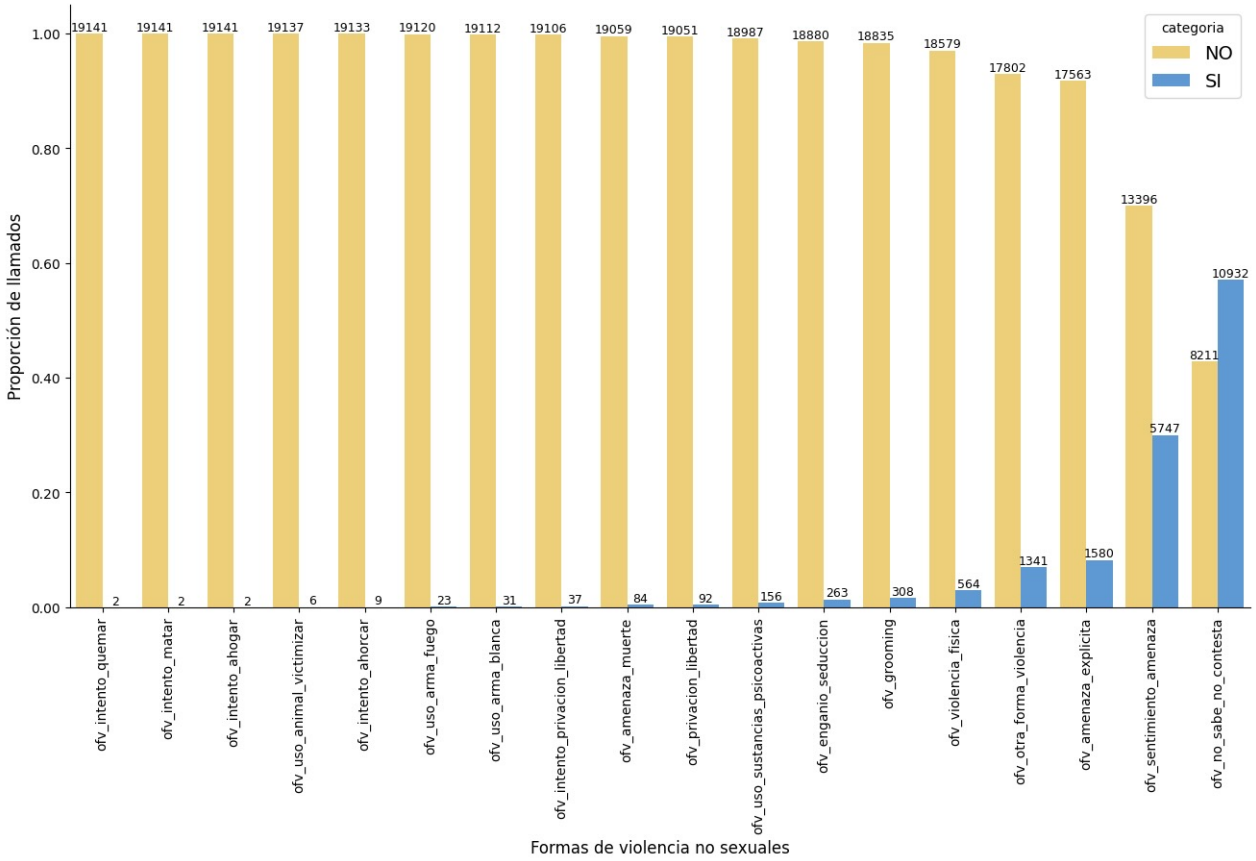

Figura 2: Tipos de violencia no sexual reportada en los llamados.

Las variables *victima_edad* y *llamante_edad* presentaban valores atípicos positivos, no solo identificables por superar la barrera de $3 * IQR$, sino también y principalmente por ser valores incongruentes con la edad de una persona. Por lo tanto, removí todos los valores por encima de 110 para ambas variables, y todos los valores por debajo de 1 para *llamante_edad* (si existen víctimas que presentan edad 0, considero que se trata de menores que aún no alcanzan el año). Los valores removidos y su cantidad para cada variable pueden verse en el cuadro 2, Outliers en variables de edad. Se puede comprobar allí que la mayoría eran 999 en ambas variables, muy probablemente un valor por defecto ingresado para no dejar el campo vacío. En total, removí 195 valores en *llamante_edad*, y 101 valores en *victima_edad*.

Cuadro 2: Outliers en variables de edad.

Variable	**Outlier**	**Cantidad de filas**
llamante_edad	999	192
	0	3
victima_edad	999	98
	224	1
	125	1
	111	1

Una vez removidos estos valores, tomé las medidas descriptivas de las variables de edad que se observan en el cuadro 3. Se puede ver que la mayoría de las víctimas no supera los 21 años, con una media de 17 y una moda de 14. Las personas que llaman para reportar los casos, en cambio, son en su mayoría adultos, con una media de 36 años, y una moda de 40. Esto refuerza lo mencionado en la introducción de hallazgos de otros estudios de que las personas más jóvenes y sobre todo los adolescentes e infantes son los grupos más en riesgo de ser víctimas de violencia sexual.

Cuadro 3: Medidas descriptivas de las variables de edad.

Descriptor	**Edad de quien llama**	**Edad de la víctima**
Media	36.25	17.17
Moda	40	14
Desvío Est.	11.41	11.91
Min.	3	0
25 %	29	10
50 %	35	14
75 %	42	21
Max	99	99

Para explorar patrones en la distribución temporal de los llamados realicé el gráfico de tendencia de la figura 3 con los datos agregados mensualmente y una media móvil de 4 meses. Se puede ver claramente en este gráfico una tendencia creciente en la cantidad de llamados desde mediados de 2017, que podría estar asociada a campañas de concientización sobre el programa y la línea y también sobre la violencia doméstica en general. Hay picos de llamados que se repiten alrededor de finales de cada año, entre los meses de octubre y enero en 2016, 2017, 2018 y 2020 aunque no parecen ser consistentes en tamaño como para considerarlos una tendencia clara. Por otro lado, hay una gran suba entre finales de 2018 y comienzos de 2019 que puede estar asociada a factores externos como los que mencioné antes. Se observa, luego de una baja y período de estabilización en 2019, una suba marcada en 2020. Un factor externo que podría estar relacionado con este patrón, es la implementación de políticas de ASPO (Aislamiento Social Preventivo y Obligatorio) durante la epidemia de COVID-19 de 2020 que obligó a la población a permanecer en sus hogares y entornos más cercanos. Si tenemos en cuenta la mayor prevalencia de la violencia sexual en ámbitos cercanos y por parte de agresores conocidos a la víctima, podría explicarse la suba de cantidad de llamados durante esta época. Sin embargo, cabe aclarar, que todas las posibles asociaciones que planteo como interpretación de esta figura deben ser contrastadas con un análisis en profundidad de las series temporales del *dataset*, que excede los objetivos de este trabajo.

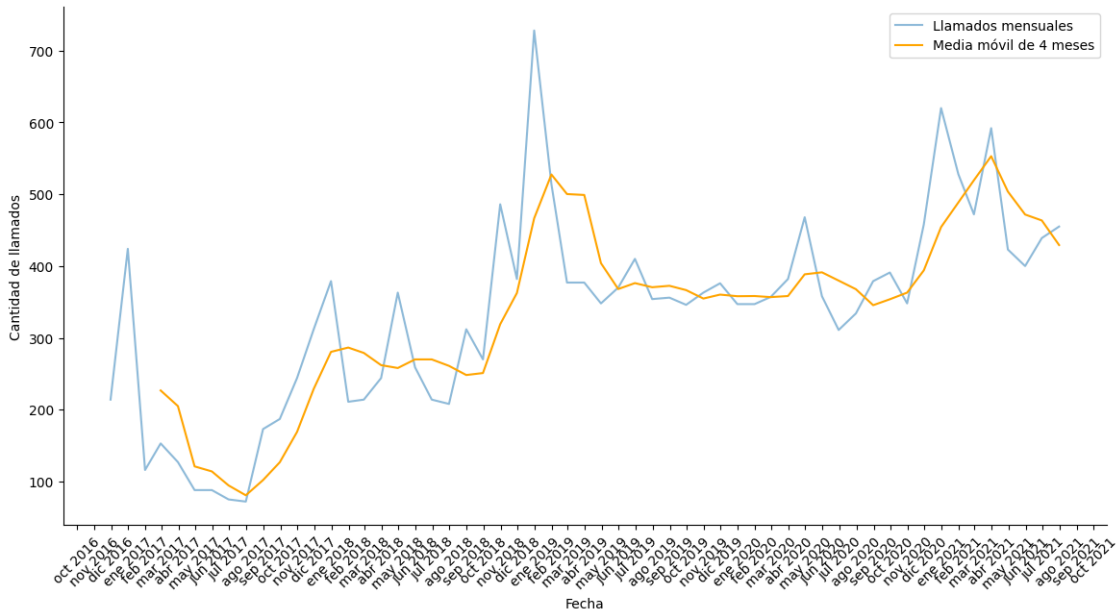

Figura 3: Cantidad de llamados en el tiempo con media móvil de 4 meses.

Además, construí las variables *estación del año*, *fin de semana*, y *momento del día* para explorar la posibilidad de otros patrones en los llamados. Observé que una mayor proporción de llamados ocurren durante la semana (80 %) y por la tarde (38 %). No observé disparidad significativa en la distribución de llamados de acuerdo a las estaciones del año.

Según la distribución de la variable *llamado_provincia*, la mayoría de los llamados provienen de la Ciudad Autónoma (37 %) y la Provincia de Buenos Aires (36 %). Del 9 % no se cuentan con datos (respuestas *NS/NC*); y el restante 18 % se reparte entre las restantes provincias del país, siendo de ese grupo Córdoba y Santa Fé las que más llamados tienen, con un 3 % cada una.[4]

Según la distribución de la variable *caso_judicializado*, el 46.7 % de los llamados no está asociado a un caso ya judicializado, el 39.7 % sí, y en el restante 13.4 % no se cuenta con datos de este tipo.[5]

En cuanto a la variable *hecho_lugar*, como ilustra el gráfico de barras de la figura 4, para aproximadamente el 30 % de los llamados no se cuenta con datos (respuestas *NS/NC*); luego, el 25 % los hechos suceden en la vivienda de la víctima y el 13 % en la vivienda del agresor. La cuarta categoría más reportada, con el 12 %, es *redes sociales*. El restante 20 % se divide entre categorías de espacios públicos (plazas, descampados, etc.), transporte, y ámbito educativo, entre otros sitios. La elevada proporción de casos que suceden en la vivienda de la víctima, es un dato que acompaña lo ya dicho en la Introducción sobre la mayoría de los hechos de violencia sexual ocurriendo más bien en el entorno de la víctima antes que involucrar personas y lugares desconocidos.

[4]Ver figura 15 en el Anexo.
[5]Ver figura 16 en el Anexo.

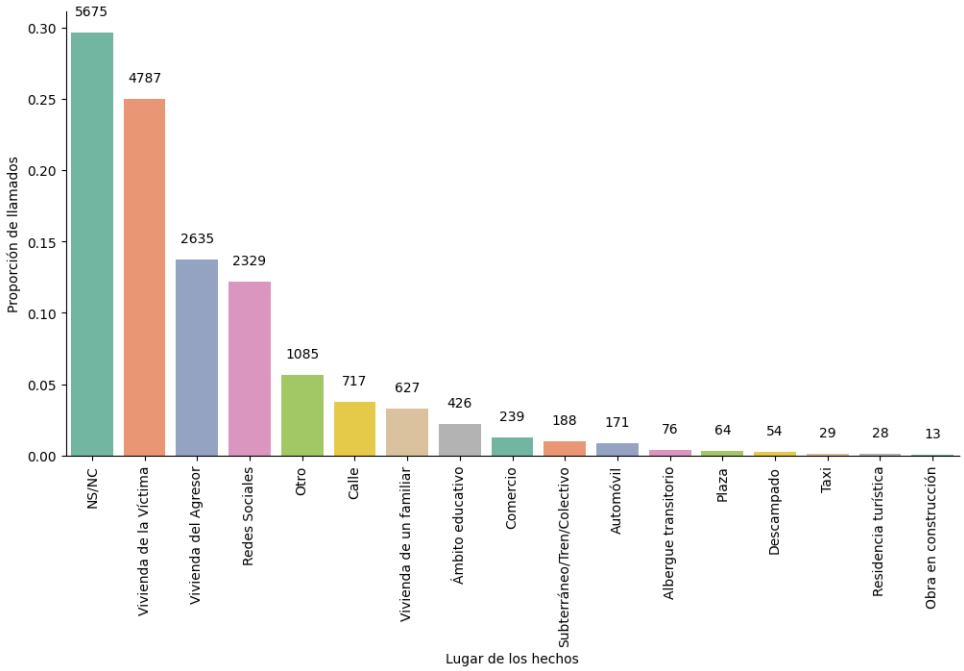

Figura 4: Lugar de los hechos.

Llama la atención la cuarta categoría más presente en *hecho_lugar*, los casos sucedidos en redes sociales. Dada la media de edad de las víctimas que reporté más arriba, se podría hipotetizar sobre la relación entre estas variables: la población joven pasa más tiempo en redes sociales y entonces es más propensa a sufrir violencia sexual en ese lugar; o las redes sociales son lugares donde proliferan más los actos de violencia sexual por alguna(s) característica(s) intrínseca(s) de las redes mismas.

La nacionalidad de las víctimas, informada por *victima_nacionalidad*, se distribuye de la siguiente manera: el 80 % de las víctimas son argentinas; del 15 % no se cuenta con datos; y el restante 5 % se divide entre las nacionalidades boliviana, paraguaya, peruana, brasileña, uruguaya, chilena, y la categoría "otra".[6]

reduje víctima nacionalidad para SVM?

Según la distribución de *victima_discapidad*, para el 53.7 % de las víctimas no se cuenta con datos, el 43.2 % no posee discapacidad, y el 2.9 % sí[7].

En cuanto al género de las víctimas, en el gráfico de barras de la figura 5, se ve reforzado lo establecido en la Introducción sobre la distribución de género en las víctimas: el 77.6 % de las víctimas son mujeres, el 18.4 % hombres, del 3.7 % no se tienen datos, y el 0.14 % son personas transgénero.

[6] Ver figura 17 en el Anexo
[7] Ver figura 18 en el Anexo

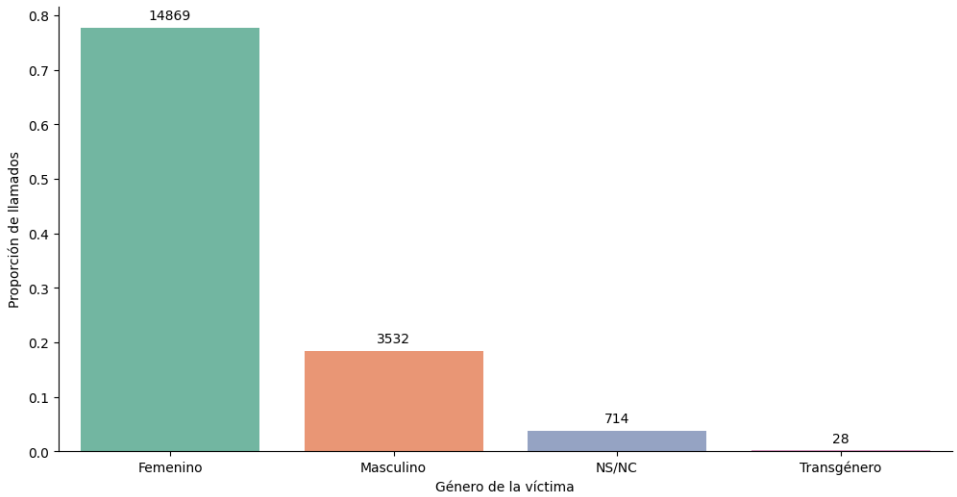

Figura 5: Género de las víctimas.

Los vínculos entre víctimas y agresores nuevamente reflejan la persistencia de los hechos de violencia sexual perpetuados por personas del entorno de las víctimas. En el gráfico de barras de la figura 6 para la variable *victima_vinculo_agresor* se observa la distribución en las diferentes categorías vinculares. Pero además la tendencia se evidencia aún más al reagrupar las categorías de la variable en *Conocido familiar*, *Conocido no familiar* (categoría ya presente en la variable original) *Desconocido*, y *NS/NC*. Mientras que 15.4 % de los agresores son declarados como desconocidos; entre familiares (47.4 %) y no familiares (19.7 %), los agresores conocidos por la víctima suman un 67.1 %. El número podría incluso ser más elevado si consideramos que podría haber agresores conocidos también como parte del 17.2 % de los *NS/NC*.

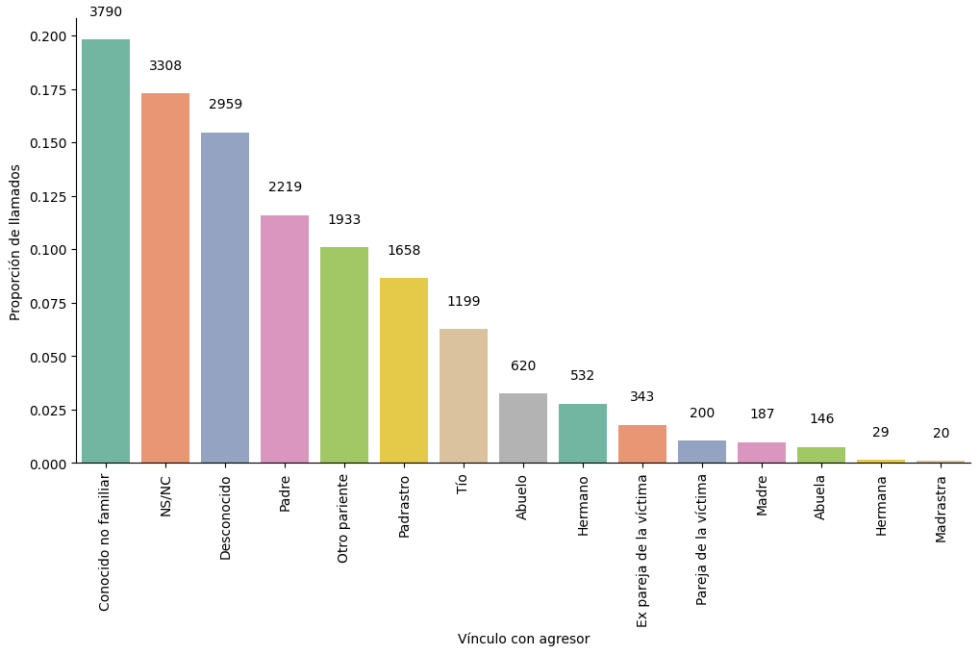

Figura 6: Vínculos víctima-agresor.

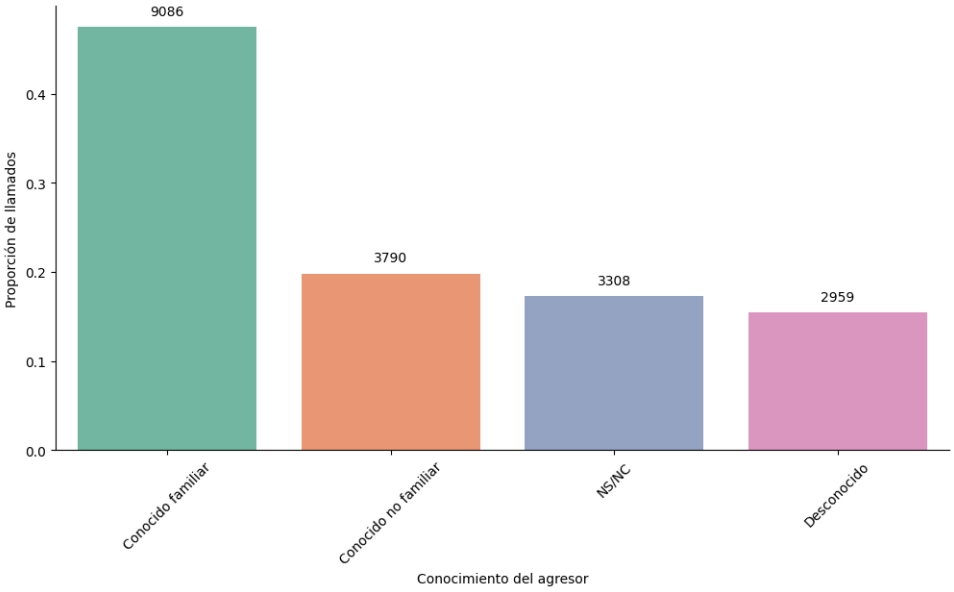

Figura 7: Agresor conocido o no por la víctima.

En la variable *vinculo_llamante_victima*, el 24.9 % de los llamados provienen de comisarías, el 17.2 % de un familiar de la víctima (otro familiar que no pertenezca a las categorías: *Madre*, *Padre*, *Abuela/o*, o *Hermana/o*), el 16 % de los llamantes son madres de las víctimas, y el 14.2 % lo constituyen las propias víctimas. El resto de las categorías son otros conocidos de las víctimas, padres, vecinos, abuelos, hermanos, otras instituciones, o *NS/NC* todas con menos del 10 %. Por último, los llamados provenientes de escuelas, defensorías y los mismos agresores suman menos del 1 %.[8]

En cuanto a la variable de interés *victima_convive_agresor*, encontré en el análisis univariado que puede verse en el gráfico de barras de la figura 8, que quiénes sí conviven con su agresor son minoría, 14.3 %; mientras que el 64.4 % no convive con su agresor. El restante 21.19 % las respuestas son *NS/NC*, la categoría que más adelante intento predecir como *SI* o *NO*.

[8] Ver figura 19 en el Anexo

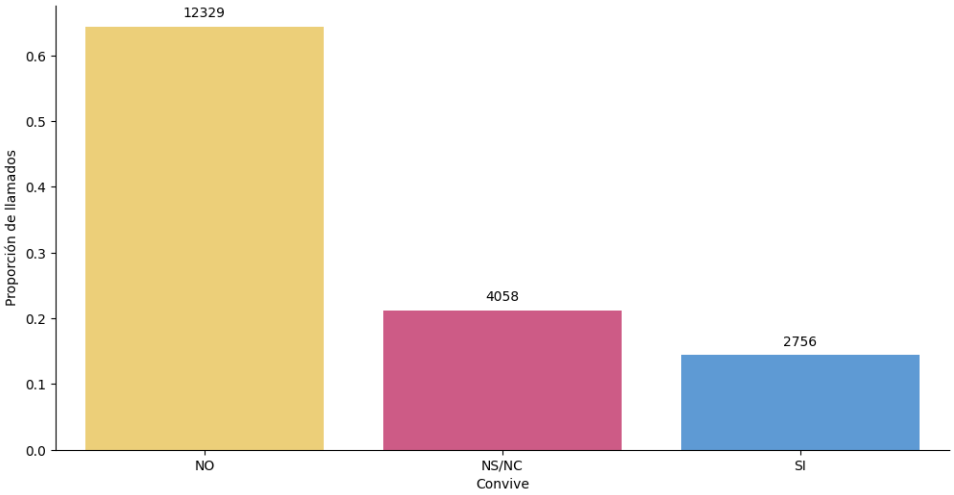

Figura 8: Convivencia víctima-agresor.

Análisis multivariado de *victima_convive_agresor*

Para un análisis multivariado seleccioné algunas variables que podían estar más relacionadas con *victima_convive_agresor*: *hecho_lugar*, *momento_dia*, *victima_vinculo_agresor*, *llamante_vinculo*, y *victima_edad*.

Primero, realicé gráficos de barras para explorar la relación entre *victima_convive_agresor* y las variables categóricas. Observé que la distribución original de *victima_convive_agresor* (mayoría de respuestas *NO* y minoría de *SI*, con *NS/NC* posicionado ordinalmente en el medio), se mantiene para casi todas las categorías de estas variables con las siguientes excepciones. En primer lugar, como se ve en la figura 9, cuando los hechos suceden en la vivienda de la víctima, hay más casos en los que la víctima sí convive con el agresor y la distribución pasa a ser *NO*, *SI*, *NS/NC*. Lo mismo sucede cuando los hechos ocurren en la vivienda del agresor, aunque en ese caso la proporción de *SI* supera por muy poco la proporción de *NS/NC*.

En segundo lugar, en la figura 10 se observa que para la mayor parte de los casos en los que el agresor es parte de la familia de la víctima (*Abuelo*, *Hermana*, *Hermano*, *Madrastra*, *Madre*, *Padrastro*, *Pareja de la víctima*), los casos en que la víctima convive son más que los casos en los que no hay respuesta sobre la situación convivencial. Sin embargo, solo en las categorías *Madre*, *Padrastro*, y *Pareja de la víctima* los casos en que las víctimas conviven con sus agresores superan a los casos en los que no lo hacen.

Por último, en la figura 11, se puede ver que para la categoría *vecina/o* de *llamante_vinculo*, la tendencia de respuestas positivas y negativas también se invierte. Sobre esta variable se observa además que los valores de *NS/NC* para *victima_convive_agresor* son notablemente más altos cuando el llamado proviene de *Otra institución* que no sea una escuela, comisaría, u hospital.

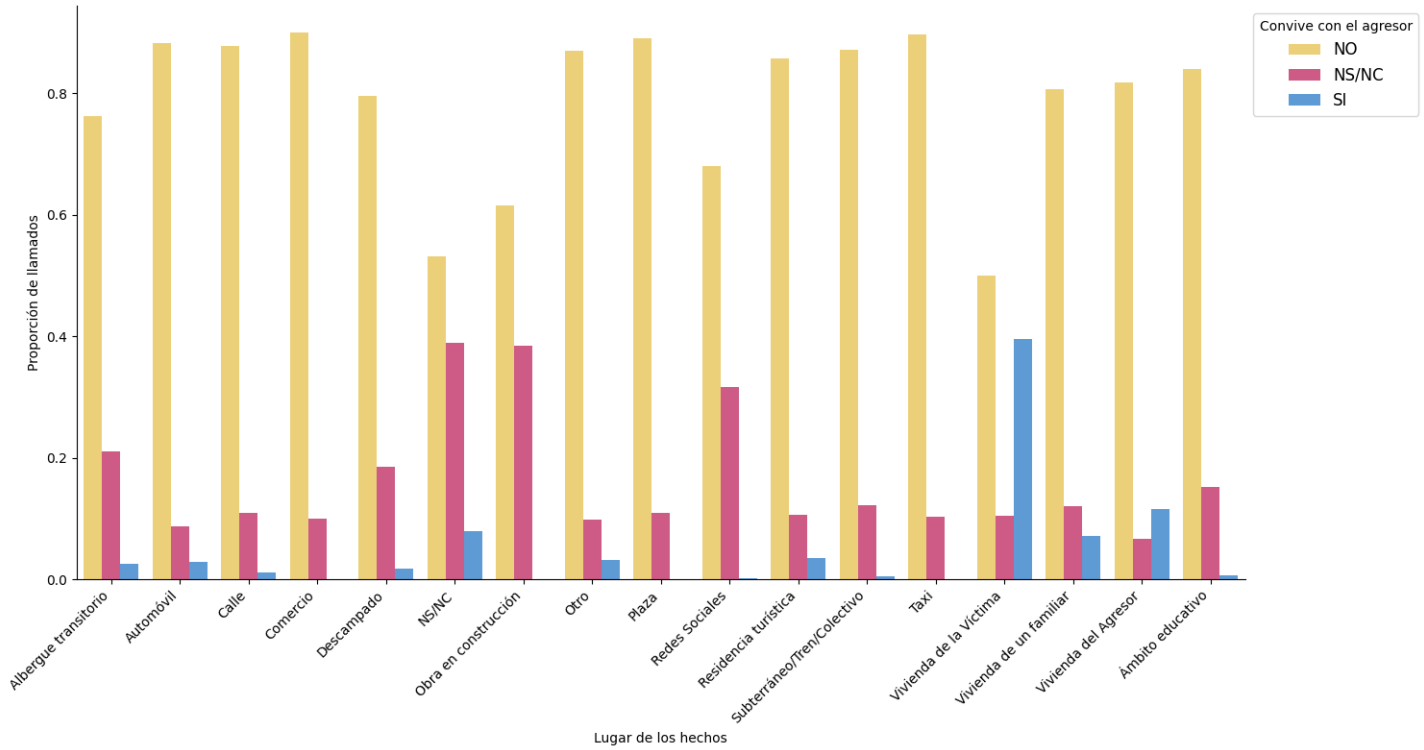

Figura 9: Convivencia con el agresor según lugar de los hechos.

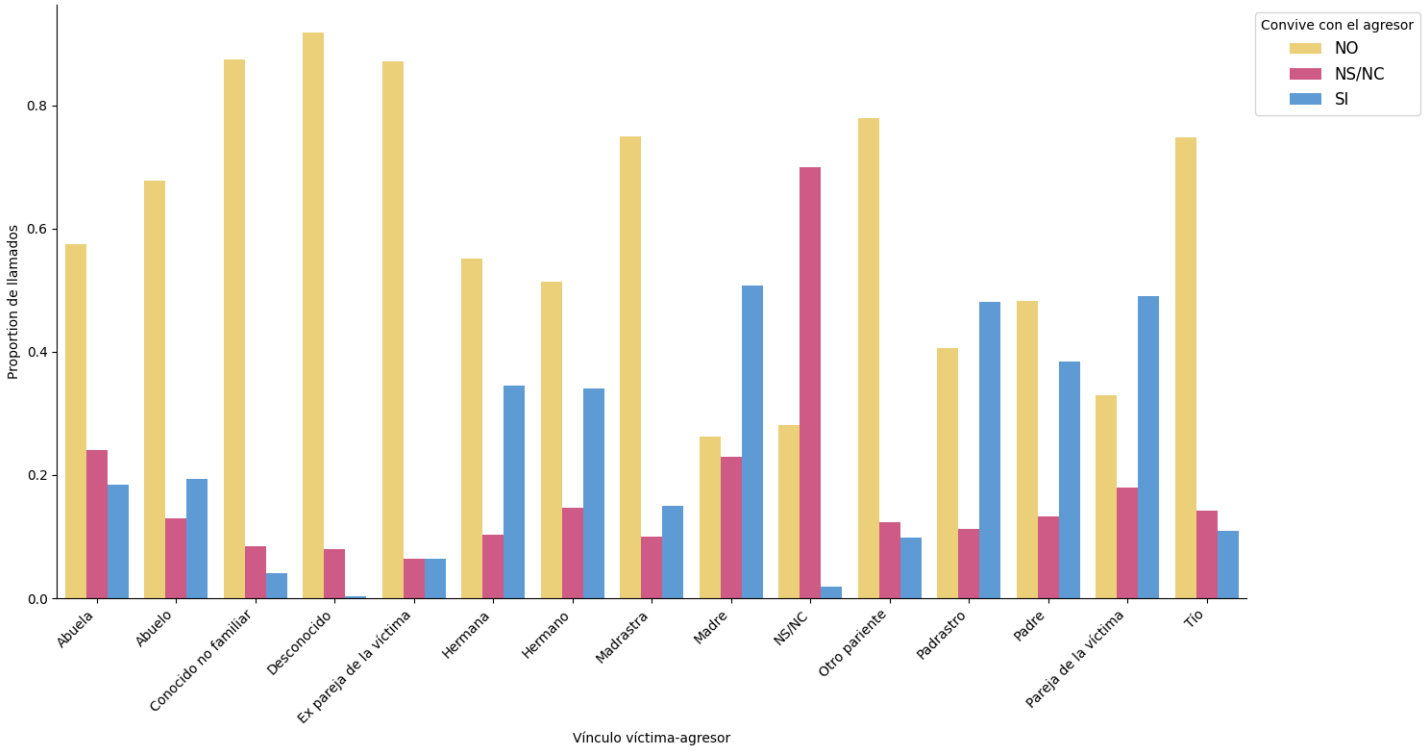

Figura 10: Convivencia con el agresor según vínculos víctima-agresor.

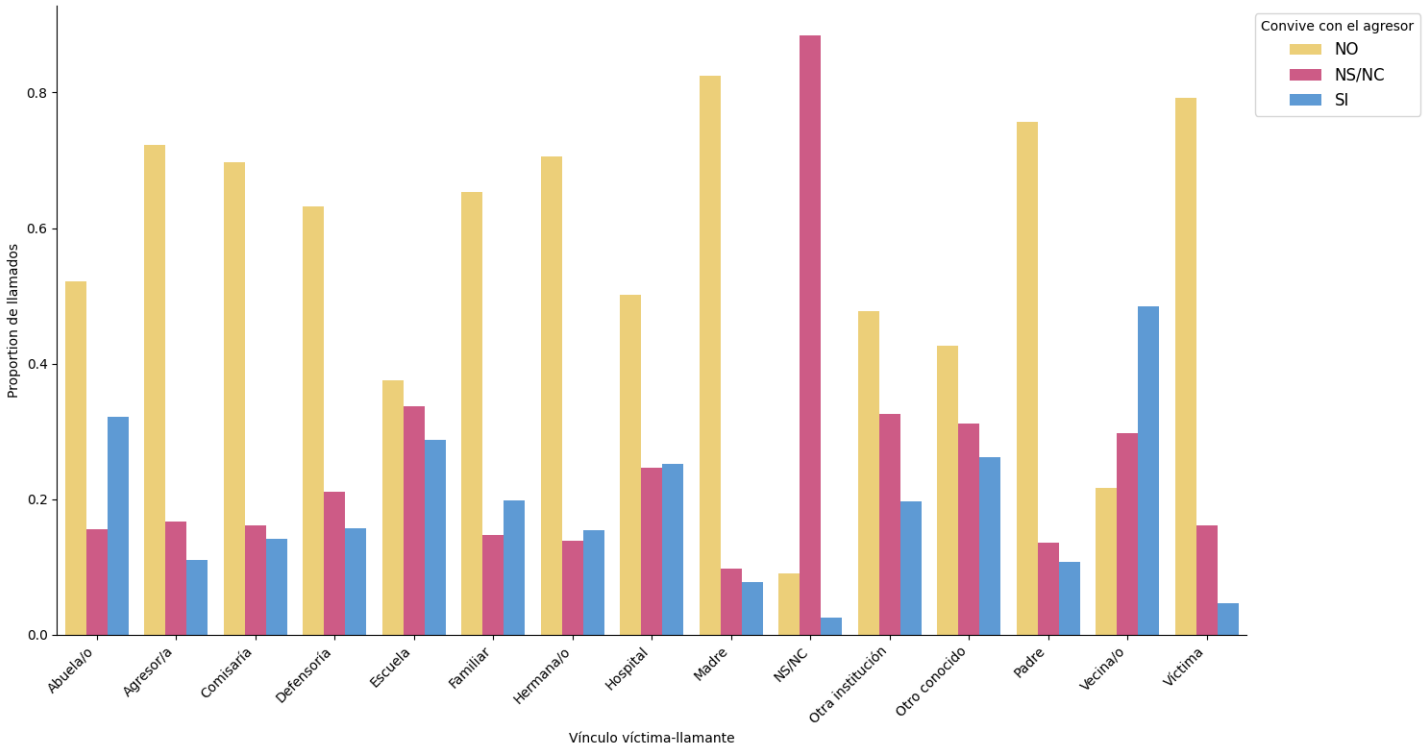

Figura 11: Convivencia con el agresor según vínculos víctima-llamante.

Luego realicé *boxplots* comparativos y un detalle del análisis de cuartiles de *victima_edad* según cada categoría de *victima_convive_agresor*. Como se ve en la figura 12 y el cuadro 4, las víctimas que conviven con el agresor son ligeramente más jóvenes que las que no lo hacen; y las víctimas de las que no se cuenta con datos sobre la convivencia parecen estar más cerca en edad de las que conviven. Sin embargo, estas diferencias en edad no parecen significativas[9].

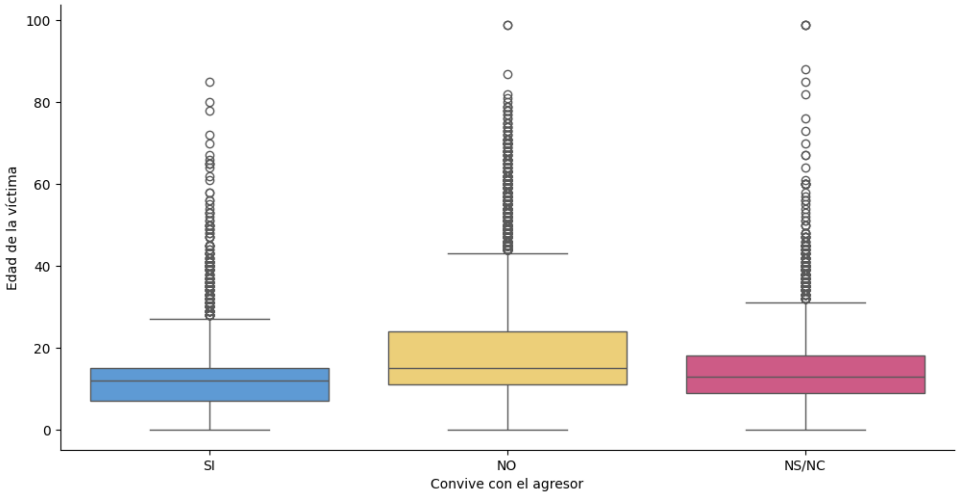

Figura 12: Distribución de la edad de la víctima según su convivencia o no con el agresor.

[9]Realizo en la sección Metodología un análsis más detallado con respecto a la correlación estadística entre la edad de la víctima y la variable de convivencia, entre otras.

Cuadro 4: Cuartiles de edad según categoría de *victima_convive_agresor*.

	Convive	**No Convive**	**NS/NC**
Q1	7	11	9
Media	12	15	13
Q3	15	24	18
IQR	8	13	9

Evalué también una posible relación entre la edad de la víctima, el vínculo con el agresor, y la situación de convivencia o no con este. En la figura 13 se observa la tendencia que ya se presentó en los *boxplots* de 12: las medias de edades de las víctimas según sus situaciones convivenciales con el agresor son similares entre sí. Destaco, sin embargo, que para las categorías *Pareja* y *Ex-pareja de la víctima* la media de edad de las víctimas es ligeramente más alta en comparación a las otras categorías de vínculos; y que específicamente la media de edad de las víctimas que sí conviven con sus agresores es más alta que la de las que no conviven o aquellas para las que no se cuenta con datos sobre la convivencia. La media de edad también se dispara para la categoría vincular *Madrastra* en los casos en que no se tienen datos sobre la situación convivencial. Por último, quiero señalar que las medias de edad más bajas ocurren con agresores *Abuelo*, *Abuela*, *Madrastra*, y *Padre*, donde ninguna media de edad supera los 10 años en las víctimas que sí conviven con sus agresores.

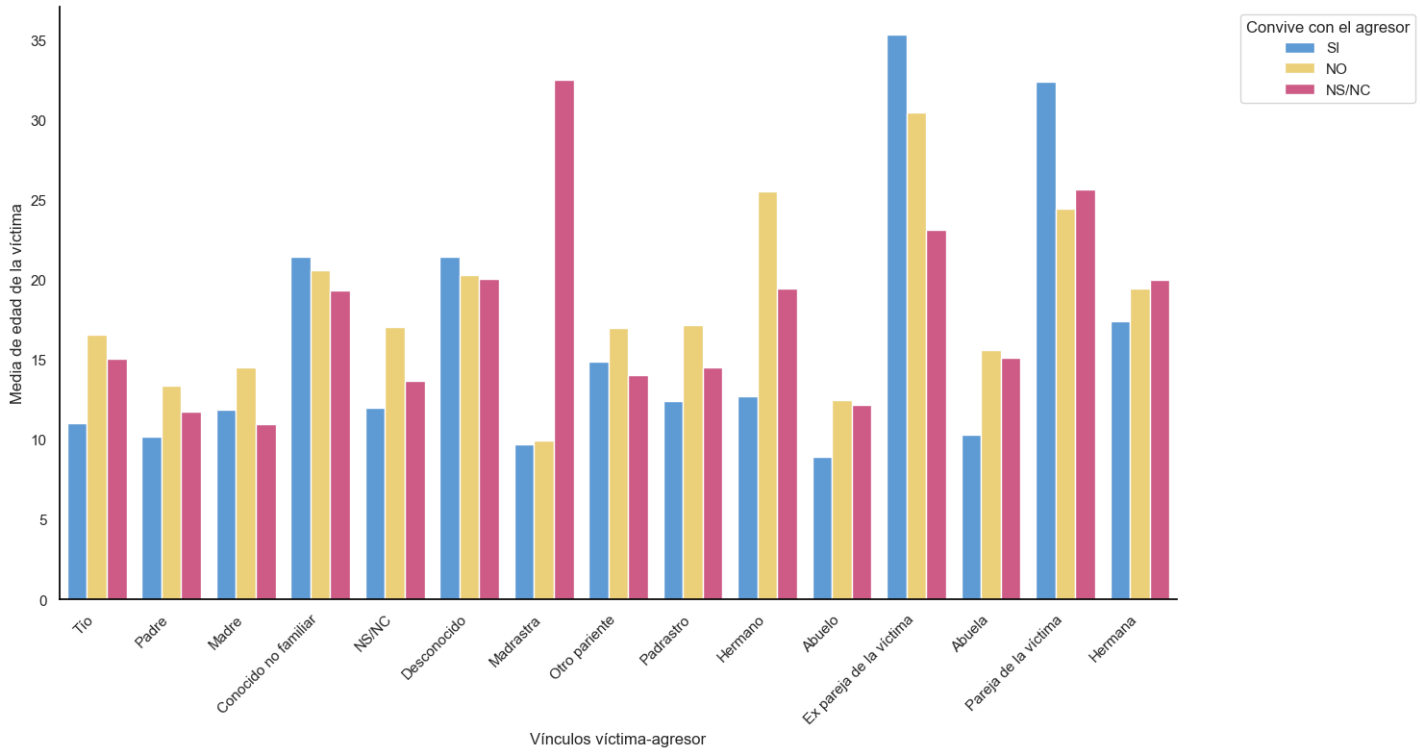

Figura 13: Edad de la víctima según su vínculo y convivencia o no con el agresor.

Por último, puse en relación *victima_convive_agresor* con los valores faltantes en *victima_edad*, que representan

el 9.82 %. Apliqué al conjunto de datos un filtro para incluir solamente las filas con casos vacíos de *victima_edad*, y generé el mismo gráfico de barras de la figura 8 con esos datos. El resultado, que puede verse abajo en la figura 14, muestra un aumento de los casos de respuesta *NS/NC* para *victima_convive_agresor*. En *dataset* completo, *NS/NC* representaba el 21.19 % en *victima_convive_agresor*, cuando solo se observan los casos con datos faltantes de edad de la víctima, ese porcentaje sube a 57.49 %. Es decir, cuando no se tienen datos sobre la edad de la víctima, tampoco se los tiene en mayor medida sobre la situación convivencial entre la víctima y el agresor.

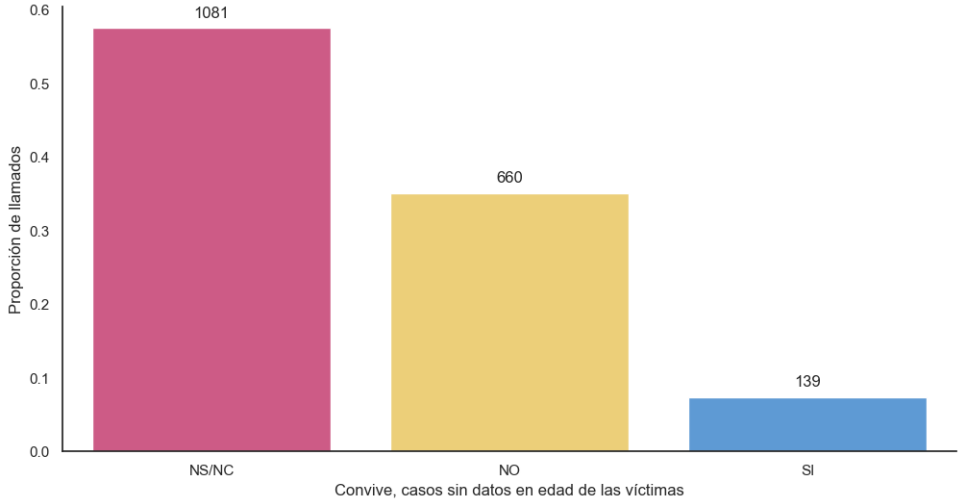

Figura 14: Convivencia víctima-agresor en las filas de datos faltantes para *victima_edad*.

Metodología

El conjunto de datos tiene gran cantidad de variables, no todas informativas, y además algunas de esas variables tienen una cardinalidad alta (por ejemplo la variable *victima_vinculo_agresor* tiene 15 niveles, la variable *llamado_provincia* tiene 25, etc.). Esto podría acarrear algunos problemas de interpretabilidad o incluso de performance del modelo predictivo. Por lo tanto, pre procesé los datos con el objetivo de reducir, por un lado, las dimensiones en general de los datos, es decir, obetener menos variables; y por otro, la cardinalidad de algunas variables. Como se verá a lo largo de esta sección, los preprocesamientos que apliqué resultan en dos tipos de *datasets* distintos que luego utilizo para entrenar modelos de SVM y evaluar las diferencias entre los tipos de preprocesamiento.

0.1. Reducción de los datos con NMDS

Acá pongo todo lo de NMDS pero nada de la reducción manual de eliminar poco informativas y agrupar por dominio. Hago nmds solo así

El siguiente paso para reducir la dimensionalidad de los datos fue aplicar un algoritmo de escalamiento multidimensional no métrico (en adelante "NMDS" por sus siglas en inglés: *non-metric multidimensioal scaling*). Este método se utiliza a menudo como método de ordenamiento para mostrar similitudes y diferencias entre los datos. Lo elegí porque, al ser un método flexible con respecto al cálculo de la matriz de distancias, me permitió calcular entre

mis datos la distancia de Gower y así trabajar con variables de distinto tipo. La desventaja de esta metodología es que la matriz de Gower no puede calcularse con datos vacíos, por lo tanto las variables de edad, las únicas con celdas vacías, resultaron un problema. La solución fue crear dos tipos de *datasets*: uno en que la edad está categorizada y los datos faltantes clasificados como *NS/NC*; y uno en que descarto la variable *llamante_edad* y trabajo únicamente con los casos completos de la variable *victima_edad*. Esta decisión que general el segundo conjunto de datos está guiada por el supuesto de que la edad de quien llama es menos relevante para la variable *target* que la edad de la víctima; y por otro lado, por el hecho de que la edad de la víctima falta en el 9.82 % de los casos, mientras que la edad de quien llama falta en el 44.82 % y dejar afuera estos casos reduciría demasiado el conjunto de datos.

Además de utilizar la técnica de NMDS para reducir la dimensionalidad del dataset y entrenar luego un modelo de SVM con el dataset reducido, aproveché las bondades de la técnica para graficar el ordenamiento del dataset reducido distinguiendo por color las respuestas SI, NO, y NS/NC de víctima_convive_agresor, con el objetivo de observar agrupamiento en los datos en torno a estas tres categorías.

Tanto el entrenamiento de modelos como los gráficos de ordenamiento de los datos utilizaron las distintas versiones del dataset reducido que mecioné: datos completos con la variable edad cagegorizada, y una versión con la variable llamante edad eliminada y solo los casos completos de edad de la víctima.

0.2. Reducción manual de los datos

Como se puede apreciar en los gráficos de las figuras 1 y 2 de la sección Datos, muchas de las variables tienen en su mayoría respuestas *NO*, por lo que resultan poco informativas. Además, muchas comparten dominio semántico y jurídico, como por ejemplo *ofv_uso_arma_blanca*, y *ofv_uso_arma_fuego*.

Reduje entonces la cantidad de variables que describen los tipos de violencias sufridas, en primer lugar, agrupando algunas de ellas en cuatro nuevas variables por dominio. En el cuadro 5 a continuación resumo cada grupo de variables y la nueva variable que las reemplaza.

Cuadro 5: Agrupación de variables de violencia por dominio.

Nueva variable agrupadora	**Variables agrupadas**
vs_explotación_sexual	vs_explotación_sexual
	vs_explotación_sexual_comercial
	vs_explotación_sexual_viajes_turismo
	vs_sospecha_trata_personas_fines_sexuales
vs_violacion	vs_violancion_via_vaginal
	vs_violancion_via_anal
	vs_violancion_via_oral
ofv_intento_violencia_fatal	ofv_intento_ahogar
	ofv_intento_quemar
	ofv_intento_matar
	ofv_intento_ahorcar
ofv_uso_arma	ofv_uso_arma_blanca
	ofv_uso_arma_fuego

En segundo lugar, eliminé las variables cuya tasa de respuesta *SI* representaba menos del 1 % de los casos. Las variables eliminadas fueron:

- *vs_amenazas_verbales_contenido_sexual*
- *vs_existencia_facilitador_corrupcion_nnya*
- *vs_eyaculacion_partes_cuerpo*
- *ofv_amenaza_muerte*
- *ofv_uso_sustancias_psicoactivas*
- *ofv_intento_privacion_libertad*
- *ofv_privacion_libertad*
- *ofv_uso_animal_victimizar*

rEDUCCI´ON DE LA CARDINALIDAD DE LAS VARIABLES

0.3. Modelos SVM

Vamos a reducirlo antes de aplicar un método predictivo. Dos maneras de reducirlo, NMDS y manual A través del análisis que llevé a cabo durante la epxloración de los datos Con el objetivo de visualizar el dataset en dimensiones reducidas me da una idea de agrupamientos con repsecto a las tres categorías de convive. Elijo NMDS porque me

permite trabajar con variables de distinto tipo sin transformaciones. Luego, para intentar predecir los NSNC como si o no usé SVM.

A modo de un segundo preprocesamiento para poder llevar a cabo estas tareas, decidí reducir las dimensiones del dataset a mano primero agrupando variables y reduciendo la cantidad de categorías en algunas otras variables. lA REDUCCI´OND E CARDINALIDAD DE algunas variables es especialemnte útil para la aplicación de SVM ya que esta conlleva encodear las features y para uno de los encoders elegidos, one-hot, la alta cardinalidad de features puede resultar problemática.

Habiendp hecho los gráficos de más arriba para explorar posibles interacciones entre dos o tres variables, me pareció valioso explorar más dimensiones y plasmarlo en dos dimensiones. Para eso apliqué NMDS

Reducción manual de dimensiones

Despues medí correlación para ver si podía sacar más variables pero al final no saqué ninguna. Explicar cómo queda el dataset final

Después apliqué encoders para hacer SVM e hice SVM con la librería tal y con una búsqueda de hiperparámetros. Además experimenté con diferentes versiones del dataset cambiando la variable edad numérica por su contraparte categórica. Esto me peritió medir el posible impacto de los datos faltantes de edad. Cuando la edad era numérica, debía dejar afuera los datos faltantes ya que SVM no puede utilizarlos. Para poder utilizar todos los datos completos de edad, pasé la edad a categorica utilizando las categorías tal tal y tal y dejado como NSNC los datos faltantes. Teniendo en cuenta estas variaciones en el tratamiento de la variable edad, los experimentos que realicé con SVM fueron: tal tal tal

cada uno probando los siguientes hiperparámentros

VER DE ARMAR UNA TABLA QUE RESUMA ESTAS VARIANTES

Resultados

NMDS vemos que no hay en ninguna de las veriones del dtaset que usé una separación clara entre las categorías de interés.

VER DE RE ARMAR NMDS y que separe solo SI de NO y luego un tercero que haga SI NO NSNC

Todos dieron bien y luego el mejor modelo lo apliqué a

Discusión y conclusiones

cruzamiento de datos ovd líneas de asistencia, observatorio de género. acceso y análisis de datos extensivo a provincias, no solo benos aitres

VER DE EN SVM RESULTANTE FINAL A LOS QUE LES PUSO S´I CU´AL ES LA EDAD DE V´ICTIMA Y A LOS QUE LES PUSO NO, CU´AL ES

en *La guerra contra las mujeres*, (2016), Rita Segato habla de la violencia sexual como algo siempre dirigido hacia cuerpos femeninos y *feminizados* (resaltado propio). Con esto último quiere decir cuerpos percibidos o construidos por los abusadores como femeninos con respecto a posiciones de poder: menores, débiles, racializados, pertenecientes

a disidencias sexuales. Esto se condice con datos sobre la mayor incidencia de la violencia sexual contra identidades masculinas durante la niñez y la adolescencia, es decir, en períodos en que los cuerpos y los sujetos son más vulnerables, y por lo tanto, también percibidos como feminizados (Contreras, Both, Guedes, and Dartnall, 2016; Unidad Fiscal Especializada en Violencia contra las Mujeres (UFEM), 2023; Ferris, 2002).

Si las denuncias de violencia sexual contra disidencias de género representan una minoría en los datos, ¿quiere decir esto que esas personas sufren menos violencia sexual?, ¿O quiere decir que, como minoría social, están subrepresentados en general y que tienen menos acceso a la justicia?

Análisis de series temporales, porsibilidad de hacer forcasting

Las víctimas que sí conviven suelen ser más jóvenes pero por muy pocos, no a significativa la correlación, salvo en el caso de agresores parejas o ex parejas de las víctimas. Pero además hay vínculos con agresores más comuens que otros para esas víctimas que sí conviven y son jóvenes.

Referencias

Juan Manuel Contreras, S Both, A Guedes, and E Dartnall. Violencia sexual en latinoamérica y el caribe: análisis de datos secundarios. iniciativa de investigación sobre la violencia sexual., 2016.

Lorraine E Ferris. World report on violence and health: Edited by etienne g. krug, linda l. dahlberg, james a. mercy, anthony zwi and rafael lozano. geneva: World health organization, 2002. *Canadian Journal of Public Health= Revue Canadienne de Santé Publique*, 93(6):451, 2002.

Claudia García-Moreno, Henrica AFM Jansen, Mary Ellsberg, Lori Heise, Charlotte Watts, et al. *WHO multi-country study on women's health and domestic violence against women.* World Health Organization, 2005.

Ministerio de Justicia de la República Argentina. Nueva Línea 137: ampliación de servicios de atención contra las violencias y para el acceso a derechos. `https://www.argentina.gob.ar/noticias/nueva-linea-137-ampliacion-de-servicios-de-atencion-contra-las-violencias-y-para-el-acceso`, 2022.

Jodie Murphy-Oikonen, Karen McQueen, Ainsley Miller, Lori Chambers, and Alexa Hiebert. Unfounded sexual assault: Women's experiences of not being believed by the police. *Journal of interpersonal violence*, 37(11-12): NP8916–NP8940, 2022.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Rita Laura Segato. *La guerra contra las mujeres.* Traficantes de sueños, 2016.

Unicef et al. Un análisis de los datos del programa «las víctimas contra las violencias", 2018.

Unidad Fiscal Especializada en Violencia contra las Mujeres (UFEM). Relevamiento de fuentes secundarias de datos sobre violencia sexual Información a nivel nacional y de la Ciudad Autónoma de Buenos Aires. `https://www.mpf.gob.ar/ufem/violencia-sexual/`, 2023.

Anexo

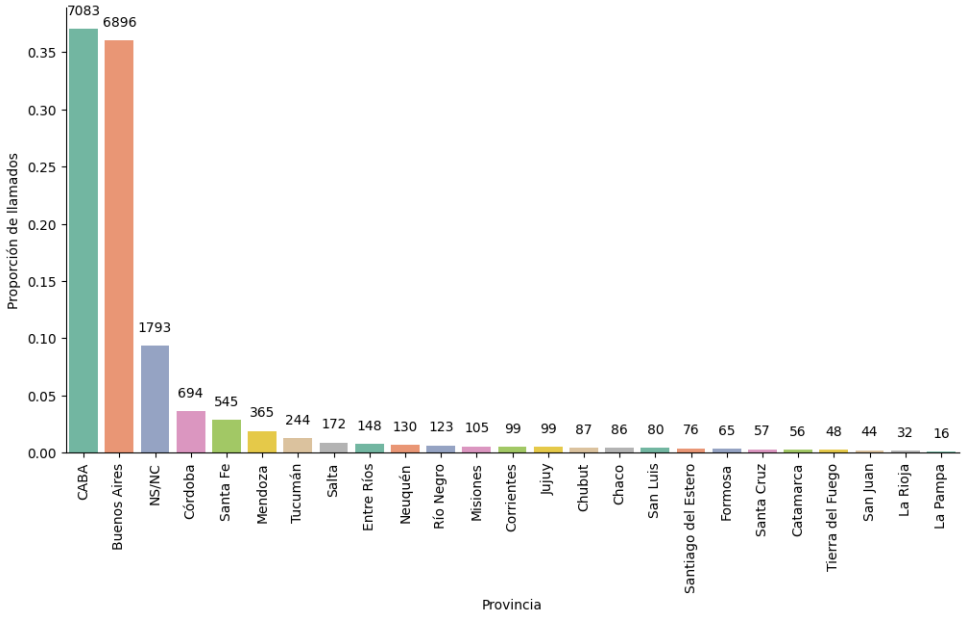

Figura 15: Llamados por provincia.

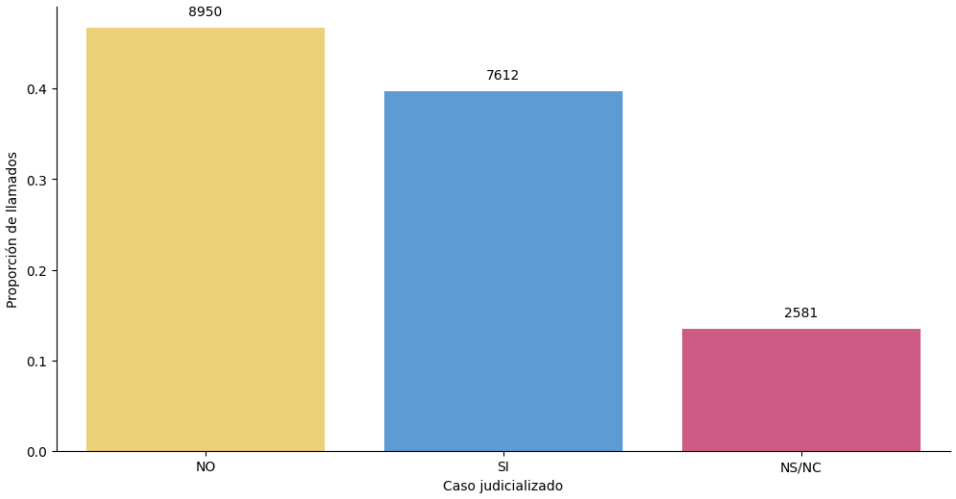

Figura 16: Caso judicializado.

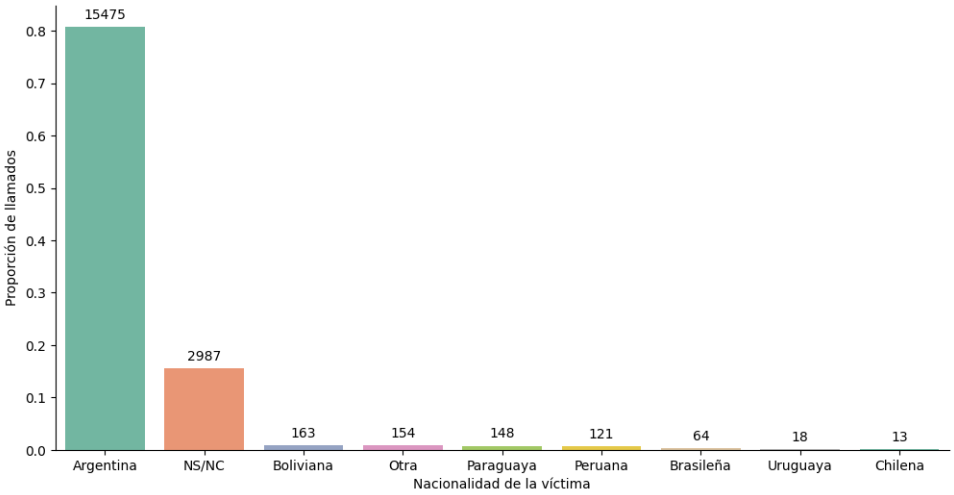

Figura 17: Nacionalidad de las víctimas.

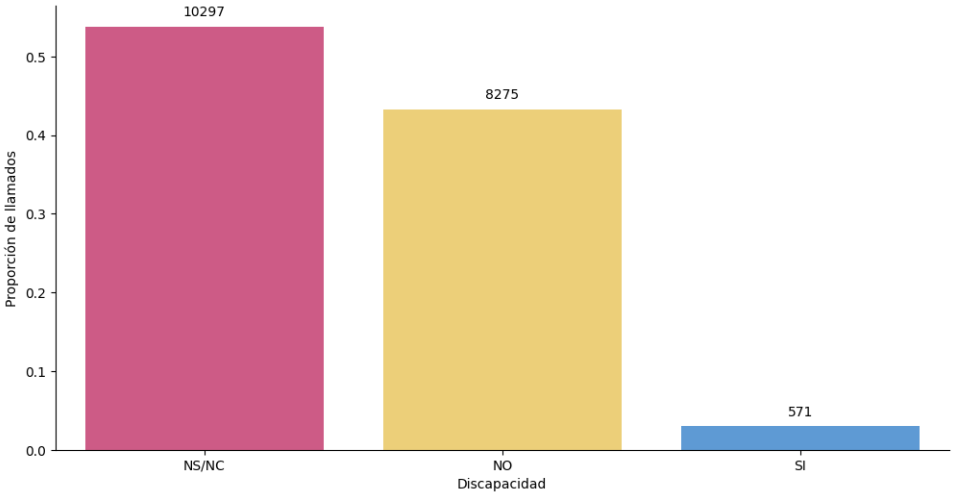

Figura 18: Presencia de discapacidad en las víctimas.

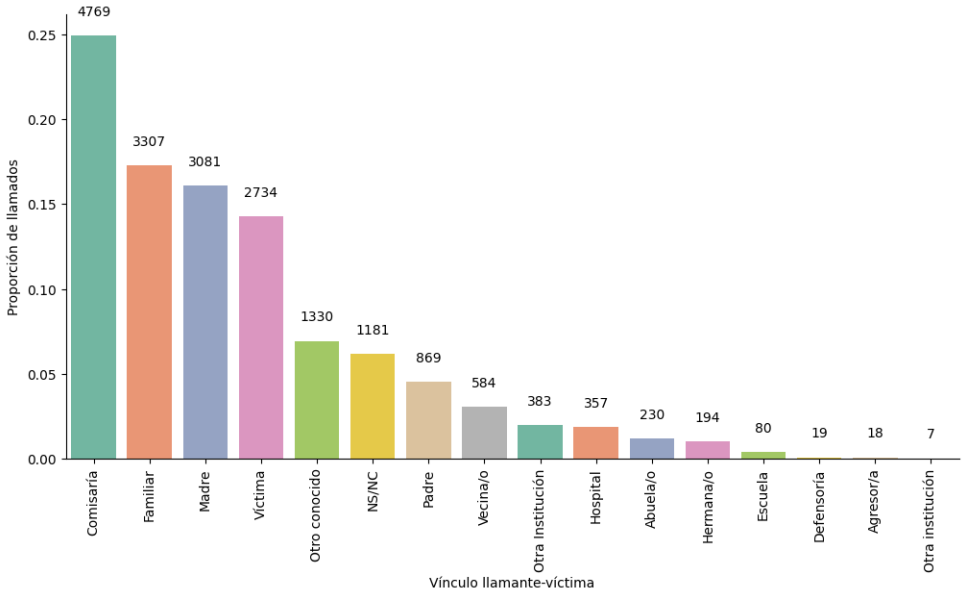

Figura 19: Vínculos víctima-llamante.