Todo list

Maestría en Explotación de Datos y Descubrimiento del Conocimiento

Universidad de Buenos Aires

Trabajo integrador

Victoria Colombo

fecha

Resumen

Índice

1. Introducción

1.1. Antecedentes

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 et al., 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*:

- Llamados por situaciones de violencia familiar
- Llamados por situaciones de violencia sexual
- Intervenciones domiciliarias por situaciones de violencia familiar
- Intervenciones domiciliarias por situaciones de violencia sexual

En este trabajo analizo llamados a la línea 137 para reportar violencia sexual entre 2017 y 2021. Los registros presentan datos faltantes de dos maneras: celdas vacías en algunas variables numéricas, y respuestas no-sabe/no-contesta en lugar de sí o no en algunas variables categóricas. De particular interés para este trabajo es la variable `victima_convive_agresor`, que codifica la situación convivencial entre la víctima y el agresor, y toma los valores `SI`, `NO`, y `NS/NC`.

1.2. Objetivos

Mis objetivos para este trabajo son:

separar con subtítulos objetivos gnerales de particulares

- Analizar los datos para determinar si existen patrones y asociaciones que permitan caracterizar a la población de víctimas que convive con su agresor y la que no lo hace.
- Entrenar un modelo predictivo que trate como datos faltantes las respuestas `NS/NC` en la variable `victima_convive_agresor` y los clasifique como `SI` o `NO`.
- En particular, entrenar un modelo de Máquinas de Soporte Vectorial.[3]
- Contrastar dos tipos de preprocesamientos de los datos para entrenar el modelo: el primero utilizando el método de ordenamiento y reducción de dimensionalidad de Escalamiento Multidimensional No-Métrico.[4]

  donde menciono el paper que usa NMDS para SVM?

en algún lado listar código y librerías

[3] En adelante, *SVM* por sus siglas en inglés: *Support Vector Machine*.

[4] En adelante, *NMDS* por sus siglas en inglés *Non-Metric Multidimensional Scaling*; el segundo, reduciendo dimensiones de los datos de manera manual,a través de la eliminación, agrupación y transformación de las variables.

2. Datos

2.1. Obtención y limpieza

Descargué del Portal de Datos Abiertos mencionado arriba 5 conjuntos de datos 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[5]:

- 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 conjunto de datos 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.
- 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".

2.2. Exploración

El conjunto de datos 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.

[5]La limpieza, normalización, y preprocesamiento de los datos y la aplicación de los métodos exploratorios y predictivos fueron realizados en Python

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.

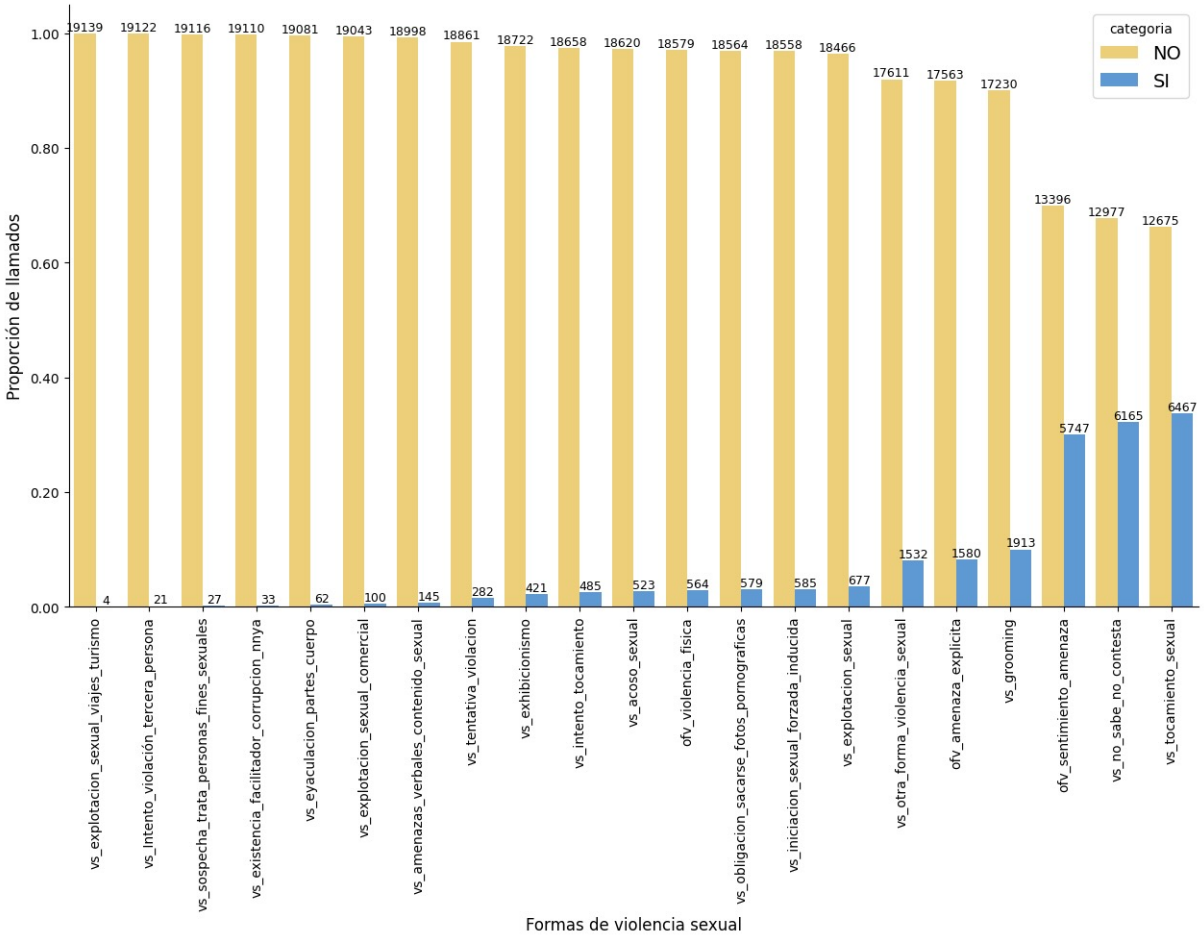

Figura 1: Tipos de violencia sexual reportada.

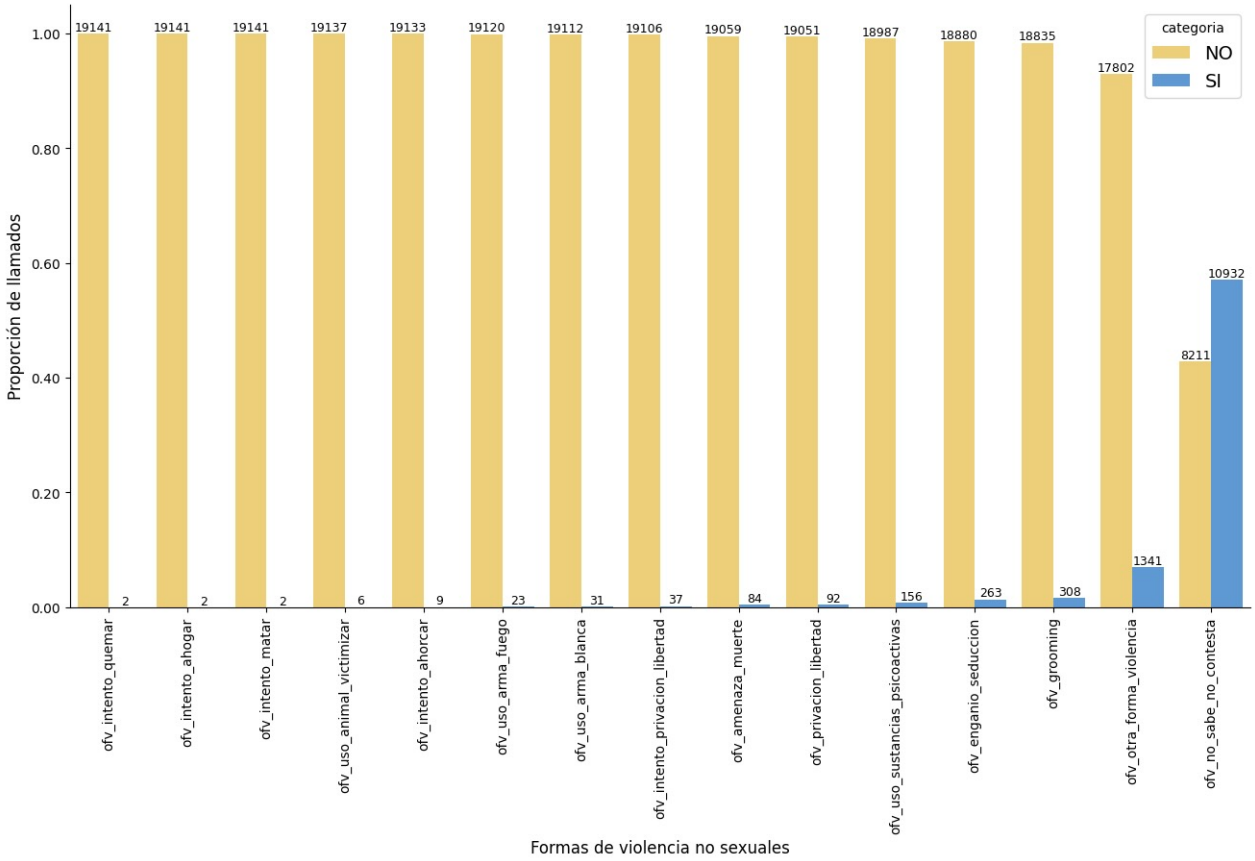

Figura 2: Tipos de violencia no sexual reportada.

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.

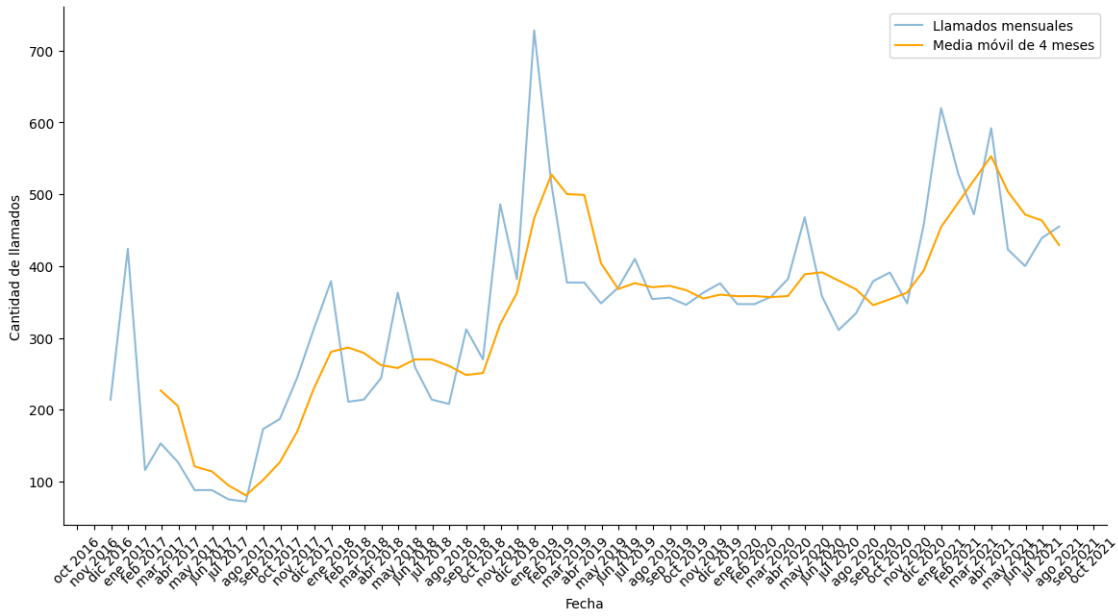

Figura 3: Cantidad de llamados en el tiempo con media móvil de 4 meses.

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.[6]

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.[7]

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.

[6] Ver figura 15 en el Anexo.
[7] Ver figura 16 en el Anexo.

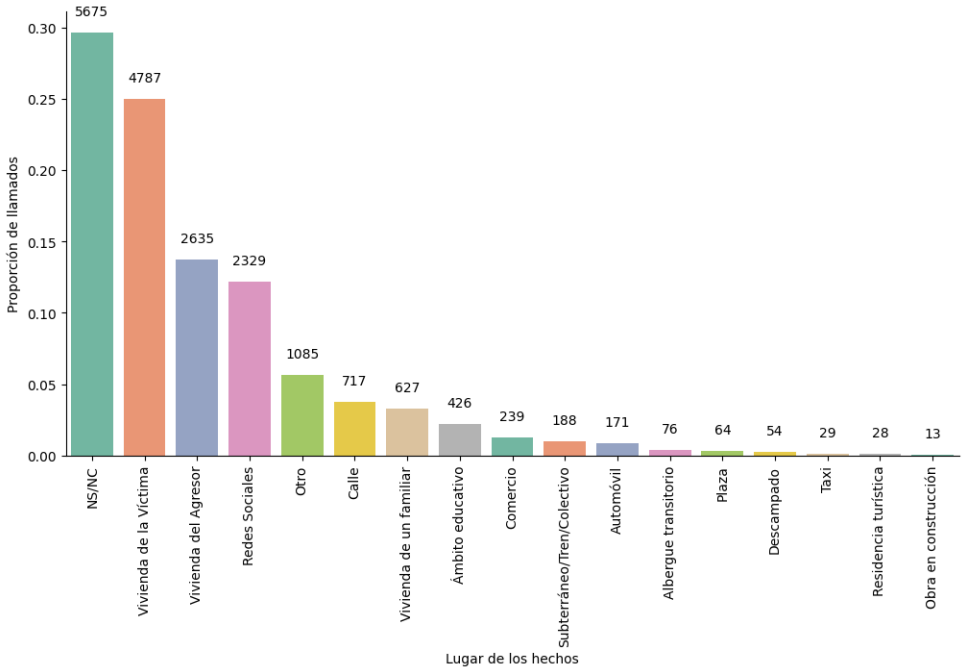

Figura 4: Lugar de los hechos.

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".[8]

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[9].

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.

[8] Ver figura 17 en el Anexo
[9] Ver figura 18 en el Anexo

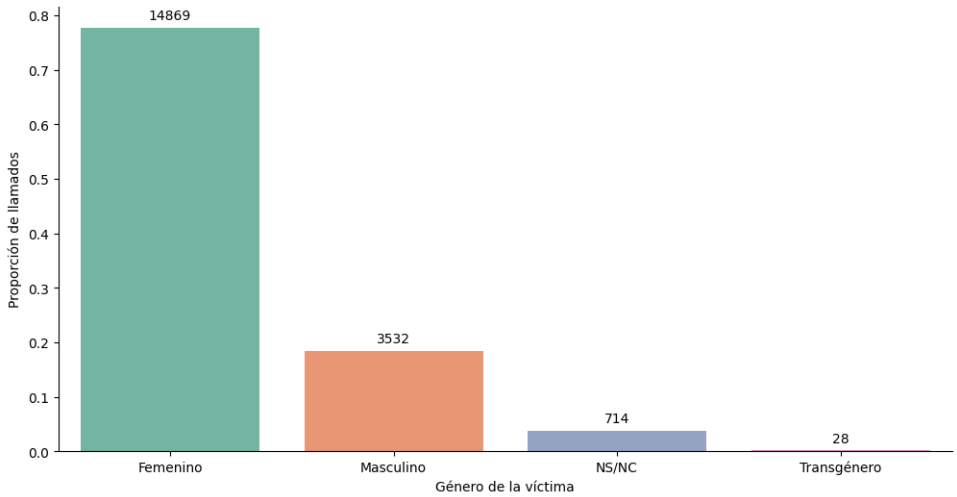

Figura 5: Género de las víctimas.

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`.

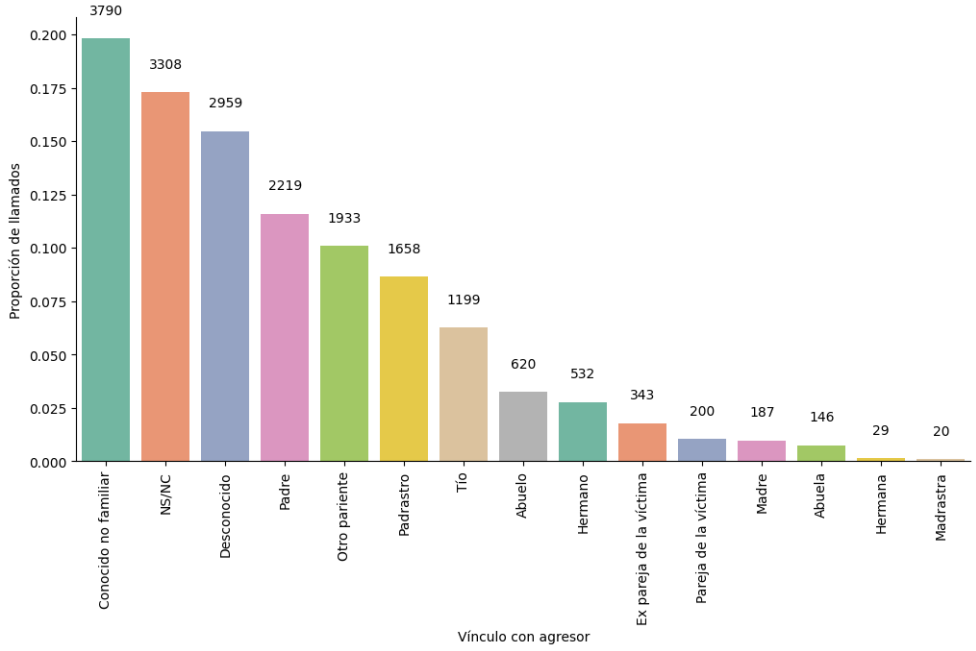

Figura 6: Vínculos víctima-agresor.

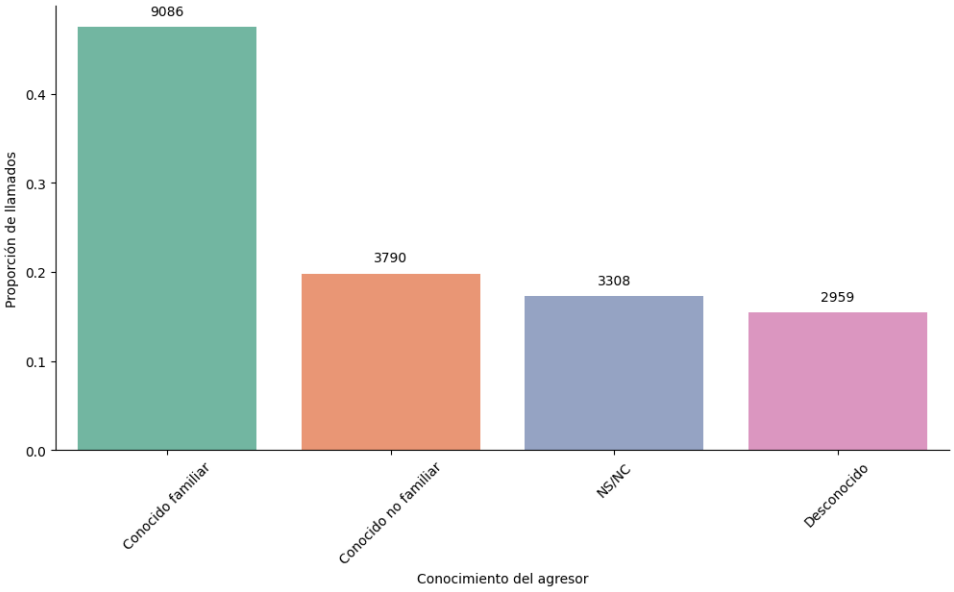

Figura 7: Agresor conocido o no por la víctima.

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 %.[10]

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`.

[10] Ver figura 19 en el Anexo

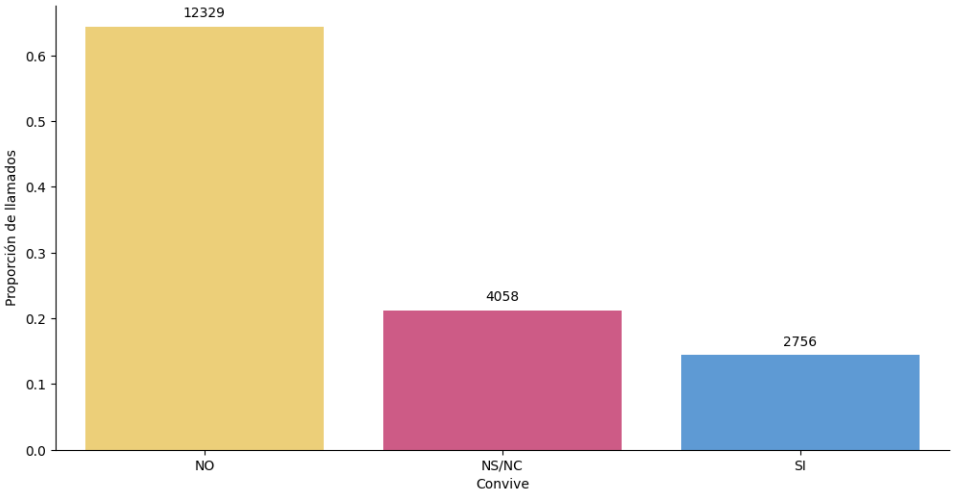

Figura 8: Convivencia víctima-agresor.

2.2.1. 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.

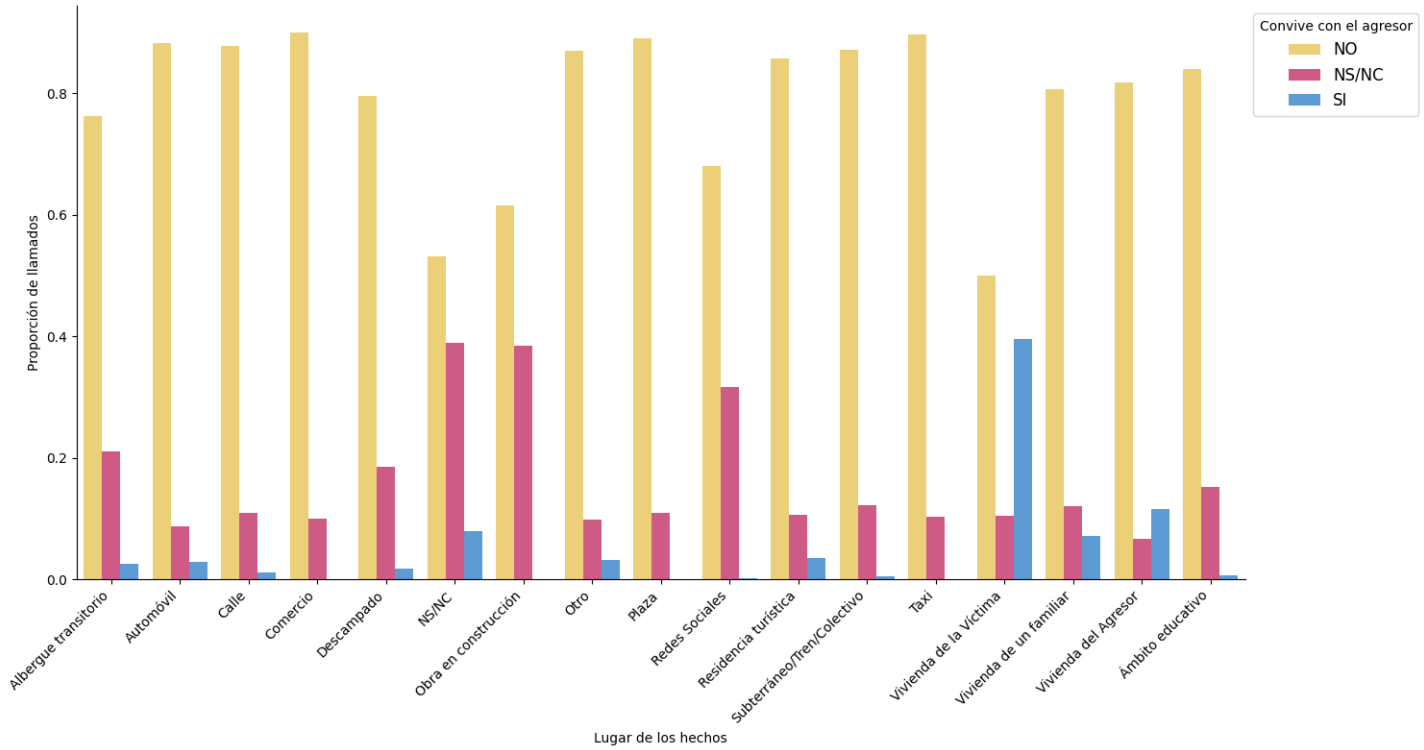

Figura 9: Convivencia con el agresor según lugar de los hechos.

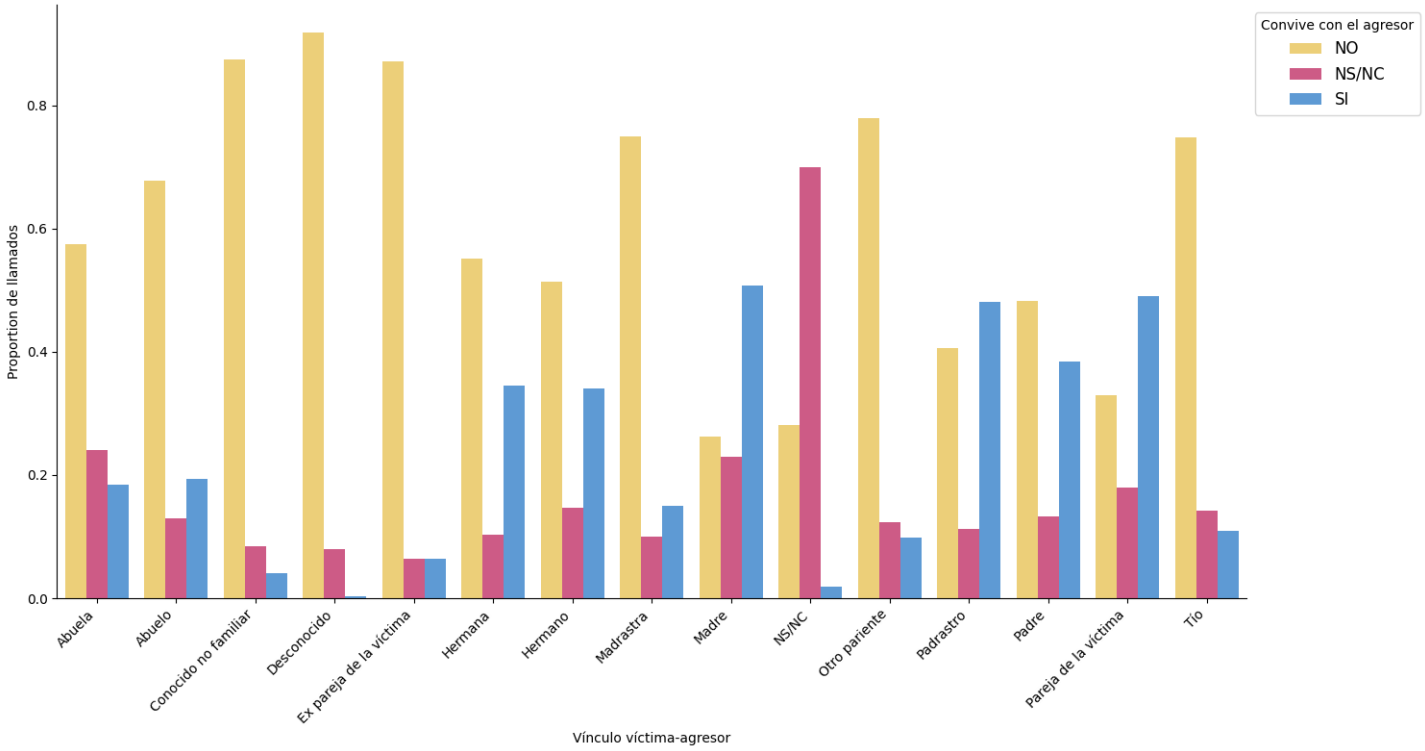

Figura 10: Convivencia con el agresor según vínculos víctima-agresor.

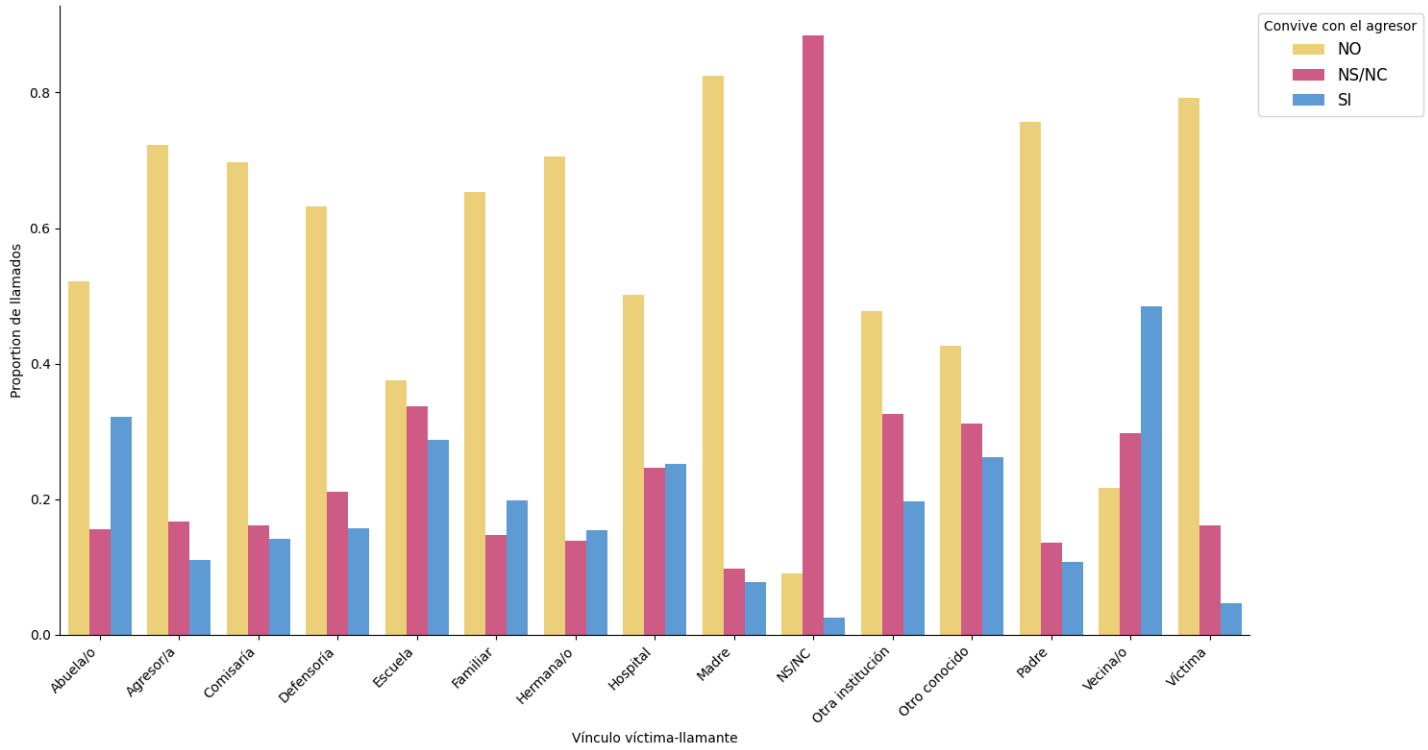

Figura 11: Convivencia con el agresor según vínculos víctima-llamante.

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.

agregar acá el análisis de correlaciones hecho para edad

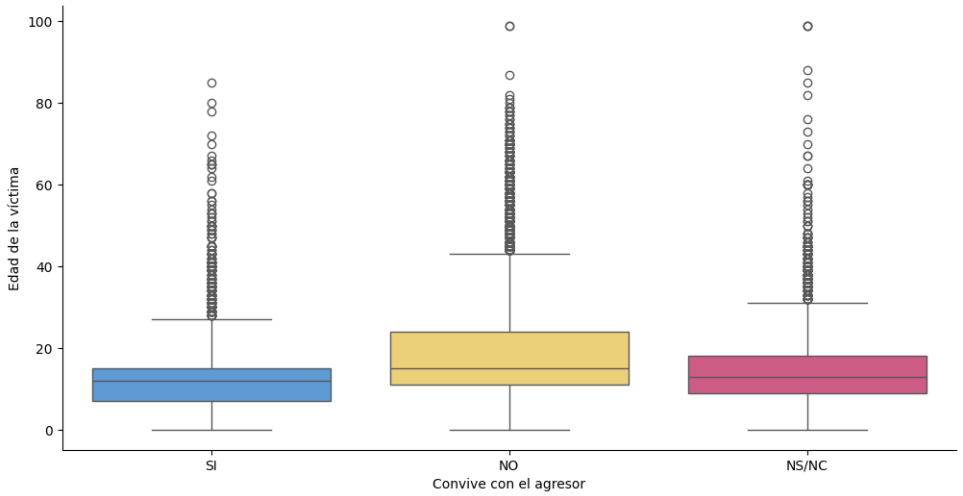

Figura 12: Distribución de la edad de la víctima según su convivencia o no con el agresor.

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.

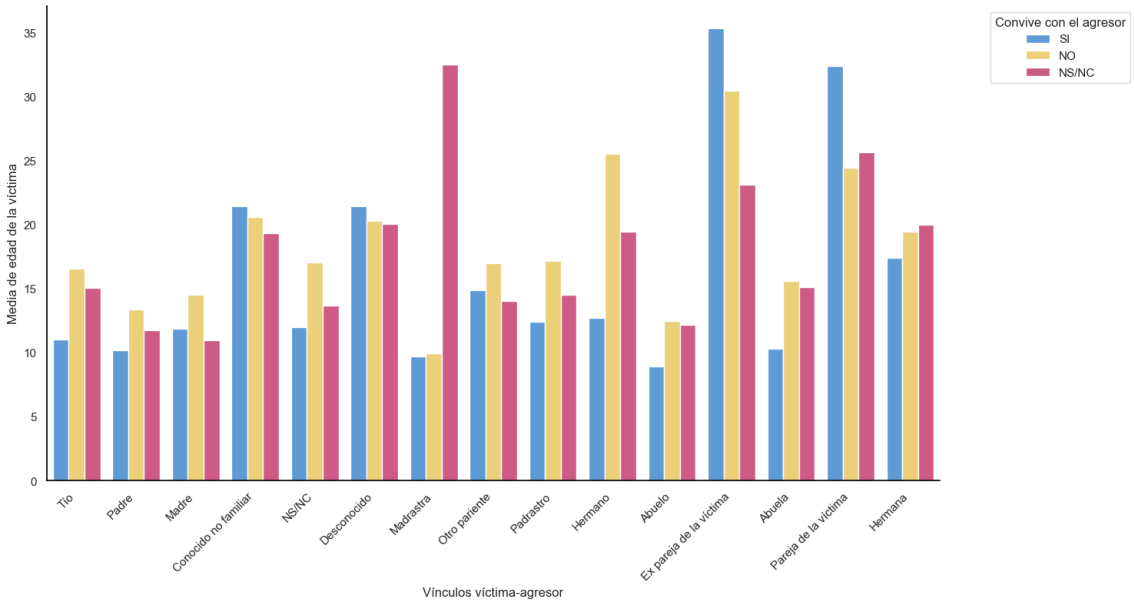

Figura 13: Edad de la víctima según su vínculo y convivencia o no con el agresor.

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.

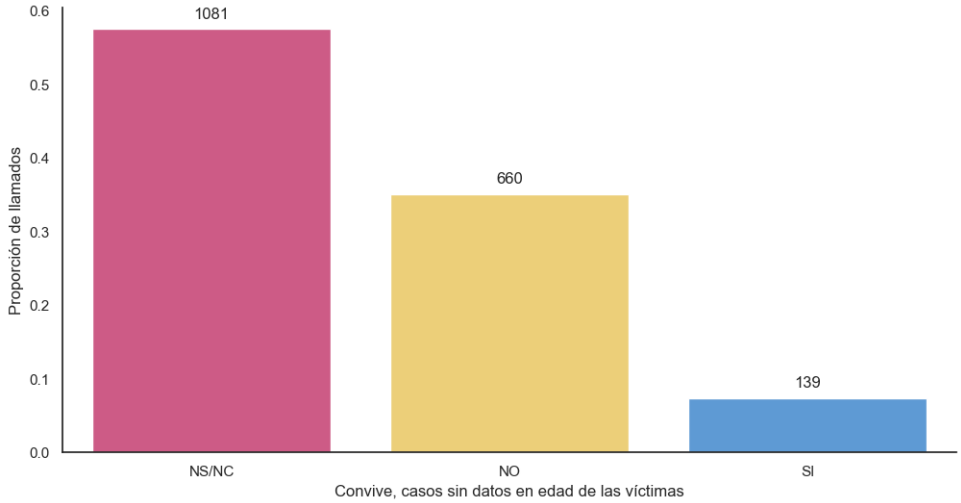

Figura 14: Convivencia víctima-agresor en las filas de datos faltantes para `victima_edad`.

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 faltante está relacionado con el motivo mismo por el que falta (MNAR).
tipo de datos faltantes que son podría ir en datos o en resultados también quizás

3. Metodología

La extensa cantidad de variables que presenta el conjunto de datos acarrea limitaciones a la hora de buscar relaciones de más de dos o tres variables con la variable objetivo, y puede causar problemas de procesamiento en el entrenamiento del modelo predictivo con SVM. Experimenté con dos metodologías para reducir las dimensiones de los datos. Primero apliqué el método de NMDS para obtener un nuevo ordenamiento de los datos en un espacio de menor dimensión. Este paso cumple la doble función de obtener un *dataset* reducido para entrenar un modelo de SVM[11], y, por otro lado, generar visualizaciones en 2 dimensiones del total del conjunto de datos para buscar agrupamientos entre las clases de `victima_conive_agresor`. En segundo lugar, realicé una reducción "manual" de los datos, eliminando, agrupando, y transformando variables de manera más controlada, respondiendo a una exploración a conciencia de los datos. Finalmente, para el entrenamiento de modelos de SVM que predigan los datos faltantes de la variable objetivo como `SI` o `NO`, utilicé distintos conjuntos de datos que varían según los dos tipos de reducciones aplicadas y distintos tipos de preprocesamientos para solventar los datos faltantes en las variables de edad.

[11] La idea de utilizar NMDS como preparación de los datos para entrenar un modelo de SVM fue suscitada por los experimentos de Cai et al.(2019) en el campo de la microbiología donde encuentran resultados moderadamente buenos en *accuracy* al entrenar un modelo de SVM con un *dataset* comprimido por NMDS (p.69).

3.1. Reducción de los datos con NMDS

NMDS

El algoritmo de NMDS es un caso particular del algoritmo MDS (por sus siglas en inglés *Multidimensional Scaling*), un método de ordenamiento que se suele utilizar para mostrar similitudes y diferencias entre los datos representándolos en espacios de menor dimensión. En comparación con otros métodos de ordenamiento y reducción, el Escalamiento Multidimensional prioriza preservar en el nuevo espacio de dimensiones reducidas el ordenamiento relativo de las distancias (o diferencias) que existen entre los puntos en el espacio original (Chan et al., 2019, p. 218). El algoritmo toma una matriz de distancias entre los puntos originales y calcula sus coordenadas en un espacio reducido a n-dimensiones (siendo en general $n = 2$ o $n = 3$ para facilitar la visualización), obteniendo la matriz de coordenadas X, que es a su vez transformada en una matriz con las disparidades $\hat{d}_{ij}$. En $\hat{d}_{ij}$ no se preservan las magnitudes de las distancias originales, pero sí se preserva el rango de ordenamiento de esas magnitudes(Chan et al., 2019, p. 219,220). A continuación, se calcula el valor del *stress*, cuya fórmula para el caso particular de NMDS[12] es:

$$Stress = \sqrt{\frac{\sum_{i<j} \left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_{i<j} d_{ij}^2}}$$

El algoritmo luego itera recalculando d_{ij} y $\hat{d}_{ij}$ para minimizar el *stress*, es decir, minimizar la diferencia entre las distancias y las disparidades (Kruskal, 1964, p. 117-123).

Para realizar el ordenamiento con NMDS utilicé el método `mds` de la librería Sklearn.[13] Los parámetros relevantes para este método fueron:

- `n_dimensions`: las dimensiones en las que al algoritmo proyecta el nuevo ordenamiento de los datos. Realicé pruebas con valores de 2 a 7, y 20 (para probar un valor extremo).
- `metric`: `False` para aplicar escalamiento no-métrico.
- `dissimilarity`: para especificar la medida de similitud o diferencia a utilizar con los datos. En este caso toma el valor `precomputed` que me permite pasarle directamente al método `fit_transform` la matriz de distancias de Gower.
- `normalized_stress`: `True` para calcular el *stress* normalizado para la versión no métrica, con la fórmula presentada más arriba.

Distancia de Gower

La elección del algoritmo de NMDS para este trabajo está motivada por la flexibilidad del método con respecto a la matriz de distancias inicial que admite. Dada la variabilidad en el conjunto de datos, elegí calcular entre ellos la distancia de Gower, una medida de similitud no métrica que puede tomarse entre variables de distinto tipo.

Para variables categóricas, la similitud entre dos puntos i y j se computa como 0 si los valores son idénticos (no hay distancia, máxima similitud) y 1 si no lo son (no hay similutud, máxima distancia). Para variables numéricas

[12] Ver Kruskal (1964), página 115.

[13] La función que se utiliza es para MDS (escalamiento métrico) pero se ajustan los parámetros necesarios para convertirla en NMDS.

se divide el valor absoluto de la diferencia entre dos puntos i y j por el rango de la variable, resultando en un valor entre 0 y 1:

$$s_{ijk} = 1 - \frac{|x_i - x_j|}{R_k}$$

Este mismo cálculo se aplica en variables ordinales.[14]

Luego, se calcula la matriz de similitud final con:

$$S(i,j) = \frac{\sum_{k=1}^{p} s_{ijk}}{\sum_{k=1}^{p} \delta_{ijk}}$$

Donde s_{ijk} es la similitud (o distancia) entre dos individuos o filas i, j en la variable k. Y δ_{ijk} es la cantidad total de variables o la cantidad total de variables en las que puede realizarse la comparación (Gower, 1971, p. 859-860). $S(i,j)$ es una matriz cuadrada con diagonal cero, que representa las distancias entre cada par de filas del conjunto de datos.

La librería que utilicé para el cálculo de la distancia de Gower

citar

no admite datos vacíos, lo que generaba un impedimento para incluir las variables de edad (las únicas con datos vacíos) en el análisis sin mediar algún tipo de transformación o filtrado. Una posible solución, que llamé solución `A`, es transformar las variables de edad de numéricas a categóricas y llenar los faltantes con `NS/NC`. Otra solución, que llamé solución `B`, es eliminar los casos vacíos en las variables de edad. Ambas tienen la desventaja de la pérdida de información, ya sea por la agrupación de rangos de edad bajo categorías unificadoras o por la literal eliminación de casos. Creé dos *datasets* siguiendo en cada uno una estrategia distinta:

- *dataset* A: las variables de edad están categorizadas, y los datos faltantes clasificados como `NS/NC`.
- *dataset* B: la variable `llamante_edad` es descartada, y también se descartan las filas con datos vacíos en la variable `victima_edad`.

La decisión que genera el *dataset* B responde por un lado al supuesto de que la edad de quien llama es menos relevante para la variable objetivo que la edad de la víctima; y por otro lado, a 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 últimos casos reduciría demasiado el conjunto de datos.

Una vez creados los conjuntos de datos A y B, calculé para cada uno la matriz de similitudes de Gower para luego aplicar NMDS y construir el modelo de SVM. En Resultados y discusión comparo y evalúo ambas soluciones según los valores de *stress* para NMDS y según la performance de los modelos de SVM.

Apliqué entonces NMDS sobre los *datasets* A y B variando los valores de `n_components`. Con el ordenamiento obtenido con `n_components` = 2 generé dos visualizaciones (una para cada conjunto de datos A y B) en las que mapeé los valores `SI`, `NO`, y `NS/NC` de `victima_convive_agresor` para buscar agrupamientos de los datos en torno

[14] Si bien Podani (1999) propone un tratamiento especial para las variables ordinales largamente utilizado hoy en día, la librería gower de sikit learn que utilizo en este trabajo trata a las variables ordinales como variables numéricas como en el trabajo original de Gower. (ver `https://sourceforge.net/projects/gower-distance-4python/files/`).

a estas categorías. Luego entrené, para cada *dataset* A y B, y para cada ordenamiento de datos según los distintos valores de `n_components` un modelo de SVM (ver más abajo la subsección Modelos SVM).

3.2. Reducción manual de los datos

Muchas de las variables que describen la violencia sufrida tienen en su mayoría respuestas `NO`, por lo que resultan poco informativas (ver nuevamente la figuras Tipos de violencia sexual reportada y Tipos de violencia no sexual reportada en la sección Datos). Además, muchas comparten dominio semántico y jurídico, como por ejemplo `vs_explotación_sexual`, `vs_explotación_sexual_comercial`, y `vs_explotación_sexual_viajes_turismo`.

Teniendo en cuenta esas características, en primer lugar eliminé 12 descriptores de violencia cuya tasa de respuesta `SI` representaba menos del 1 % de los casos.

En segundo lugar, agrupé algunos de los restantes en tres nuevas variables por dominio, como se puede ver en el cuadro 5 a continuación.

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_violacion_via_vaginal`
	`vs_violacion_via_anal`
	`vs_violacion_via_oral`
`vs_tentativa_violacion`	`vs_tentativa_violacion`
	`vs_intento_violacion_tercera_persona`

En *Encoding methods for categorical data*, (Udilă, 2023), el autor compara métodos de *encoding* según su impacto en la calidad de un modelo de SVM entrenado, y encuentra que *One-hot encoding* consistentemente resulta en modelos con mejor *accuracy* (p.7). Por otro lado, en el mismo artículo el autor expone una limitación conocida del método *One-hot*: con variables de alta cardinalidad el tiempo de procesamiento se vuelve potencialmente demasiado alto (p.7).

Por lo tanto, para aplicar *One-hot encoding* al conjunto de datos (paso necesario para entrenar un modelo de SVM que no puede aplicarse con variables categóricas) sin arriesgar problemas de procesamiento, reduje la cardinalidad de las variables con más de 5 niveles.[15]

- `llamado_provincia` tenía originalmente 25 niveles y quedó reducida a 6: `Buenos Aires`, `C.A.B.A.`, `Región Norte`, `Región Central`, `Región Patagónica`, y `NS/NC`.

[15] El desglose de cada una de las variables originales con la proporción de llamados por categoría se puede re visitar en la sección Exploración.

Cuadro 6: Experimentos para entrenar modelos de SVM.

Experimento	***Dataset***	**Especificaciones**
Reducción con NMDS	A	Variables de edad transformadas a categóricas. Datos faltantes codificados como `NS/NC`
	B	Solo datos completos de edad de la víctima (numérica). Edad de quien llama eliminada.
Reducción manual con métodos mixtos	A	Variables de edad transformadas a categóricas. Datos faltantes codificados como `NS/NC`. Aplicación de one hot encoder pot categorización.
	B	Solo datos completos de edad de la víctima (numérica). Edad de quien llama eliminada.

- `victima_nacionalidad` tenía originalmente 9 niveles y quedó reducida a 3: `Argentina`, `NS/NC`, y `Otra`.
- `hecho_lugar` tenía originalmente 17 niveles y quedó reducida a 6: `NS/NC`, `Vivienda de la víctima`, `Vivienda del agresor`, `Redes sociales`, `Espacio/transporte público`, y `Otro`.
- `llamante_vinculo` tenía originalmente 16 niveles y quedó reducida a 5: `Institución`, `Conocido de la víctima`, `Víctima`, `Agresor`, y `NS/NC`.
- Para `agresor_vinculo` tomé la agrupación mostrada en la figura 7, Agresor conocido o no por la víctima presentada en Exploración, que reduce la cantidad de niveles de 15 a 4: `Conocido familiar`, `Conocido no familiar`, `NS/NC`, y `Desconocido`.

El conjunto de datos resultante de la reducción manual tiene 36 variables, 18 menos que el original (54). En el cuadro 7 del Anexo se puede ver un resumen de las variables eliminadas, agrupadas, y transformadas.

Finalmente también dividí este conjunto de datos final en dos tipos, A y B, que responden a las transformaciones o filtrado de las variables de edad explicadas al final de la sección anterior, Reducción de los datos con NMDS.

3.3. Modelos SVM

importantísimo ver si accuracy tiene sentido calcular u otra métrica

Entrené modelos de SVM con dos las dos reducciones de datos explicadas, y las dos variaciones del conjunto de datos en las variables de edad A, y B. En el cuadro ref modelos SVM se detallan los experimentos. Los distintos modelos fueron entrenados con los datos "completos" de `victima_convive_agresor`, es decir, las filas con respuesta `SI` o `NO` en esa variable; y las filas con respuesta `NS/NC` fueron tratadas como los datos faltantes que el modelo debe aprender a completar como `SI` o `NO` y por lo tanto como un *test* final ciego.

3.3.1. Entrenamiento de SVM con reducción NMDS

Para el entrenamiento de los modelos de SVM con los *datasets* A y B:

1. Separé `victima_convive_agresor` (y) del resto de los datos X.

2. Quité a y las filas correspondientes a los casos con respuesta `NS/NC`.

3. Generé a partir de X una matriz de distancias de Gower que luego utilicé para la transformación con NMDS.

4. Generé distintas reducciones NMDS varaindo el parámetro `n_components`.

5. Creé el conjunto de prueba ciega final de casos no vistos a partir de las filas ya transformadas de X correspondientes a los casos con respuesta `NS/NC` en `victima_convive_agresor`.

6. Dividí X e y en entrenamiento (80 %) y testeo (%20) de manera estratificada[16] para mitigar el desbalance de las clases `SI NO`.[17]

7. Con cada variación de reducción por NMDS correspondiente a los distintos `n_components`, realicé una optimización de los hiperparámetros de SVM `kernel`, `C`, y `gamma` entrenando con validación cruzada de 5 divisiones. Medí también el `stress` para cada reducción.

8. Con los mejores hiperparámentros por f1_weighted entrené un modelo que apliqué al conjunto de *test* final no visto. Ante la imposibilidad de evaluar el resultado del mejor modelo sobre el test final con métodos tradicionales, medí la proporción de SI NO en el dataset original y en la predicción para compararlas y así estimar la performance.

Explicar el tipo de kernel, el vaor de C y gamma.

3.3.2. Entrenamiento de SVM con reducción manual de los datos

Para entrenar modelos de SVM con los datos reducidos a mano:

1. Encodeé con un método ordinal (ordinal encoder sklearn) las variables `llamado_fecha_hora`, `momento_dia`, `estacion_del_año` y ambas variables de edad previamente categorizadas (esto último solo en el caso del *dataset* A). La elección del tipo de *encoder* ordinal fue para preservar la naturaleza ordinal de las variables al ser transformadas.

2. Escalé las variables encodeadas con *ordinal encoder*, siguiendo las recomendaciones de Udilă 2023.
   agregar página

3. Encodeé las variables cuyas categorías eran solamente `SI`, `NO`, incluyendo la variable a predecir, mapeando `SI`= 1, y `NO`= 0. (En el caso de la variable objetivo, los `NS/NC` fueron previamente transformados a celdas vacías.)

4. Encodeé el resto de las variables con *one hot encoder*.

5. Creé el test final ciego ut

[16] Utilizando el método `StratifiedShuffleSplit` de sklearn.
[17] Ver nuevamente la figura 8, Convivencia víctima-agresor.

4. Resultados y discusión

4.0.1. Graficar ordenamiento con NMDS

NMDS graficar vemos que no hay en ninguna de las veriones del dtaset que usé una separación clara entre las categorías de interés. El valor del stress para dos dimensiones es malo.

4.0.2. Modelos predictivos de SVM

modelos entrenados con datos reducidos A y B. El gridsearch encontró los siguientes mejores parámetros, con los mismos valores para todas las versiones de n componenets. Es decir, la cantidad de dimensiones a la que se reducen los datos, y por lo tanto el valor stress de NMDS no afecta el entrenamiento del modelo positiva ni negativamente. Durante la etapa de entrenamiento y búsqueda de mejores parámentros, el modelo resultante consistentemente es únicamente bueno rediciendo respuestas NO, es decir,buen accuracy, del 80, pero el peor recall de la vida, peor precision de la vida, y peor F1 de la vida.

Al aplicar estos "mejores modelos" al set de testeo final ciego, es decir, al set de datos que no tienen etiqueta SI o NO en convive porque son los datos faltante, el modelo predice mayoría de NO.

Los modelos entrenados con el dataset reducido a mano dieron mejor. Igual no se puede saber porque es un test a ciegas al fin y al cabo, solo pude aplicarlos al test final y comparar la proporción de NO y SI que predijo con la proporción de NO y SI en el dataset original. Es parecida.

5. Conclusiones

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

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

Wenfang Cai, Keaton Larson Lesnik, Matthew J Wade, Elizabeth S Heidrich, Yunhai Wang, and Hong Liu. Incorporating microbial community data with machine learning techniques to predict feed substrates in microbial fuel cells. *Biosensors and Bioelectronics*, 133:64–71, 2019.

Débora Chan, Cristina Inés Badano, and Andrea Alejandra Rey. *Análisis inteligente de datos con R: con aplicaciones a imágenes*. edUTecNe, 2019.

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.

John C Gower. A general coefficient of similarity and some of its properties. *Biometrics*, pages 857–871, 1971.

Joseph B Kruskal. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2):115–129, 1964.

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.

János Podani. Extending gower's general coefficient of similarity to ordinal characters. *Taxon*, 48(2):331–340, 1999.

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.

Andrei Udilă. Encoding methods for categorical data: A comparative analysis for linear models, decision trees, and support vector machines. 2023.

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.

6. Anexo

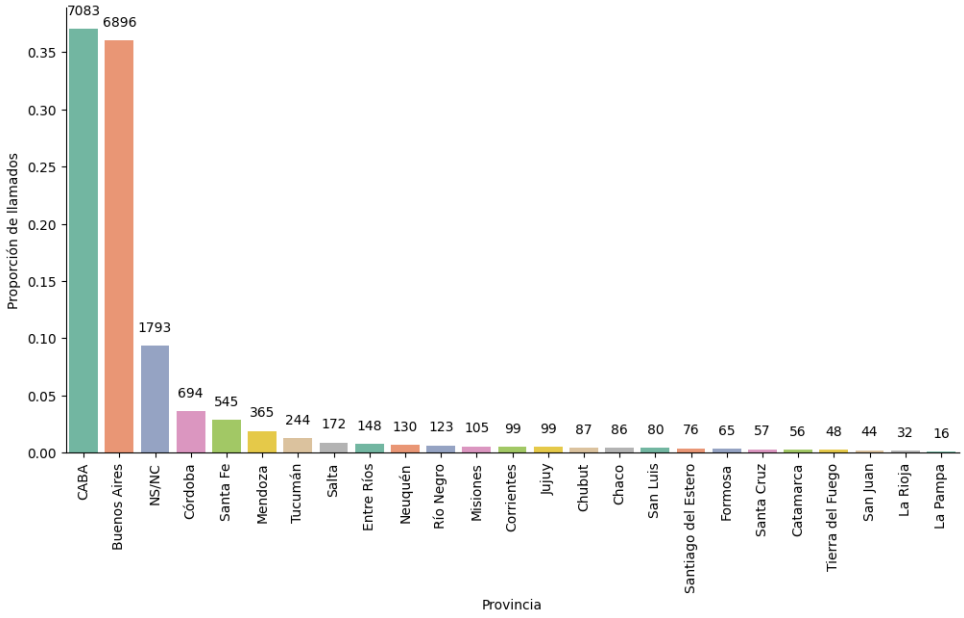

Figura 15: Llamados por provincia.

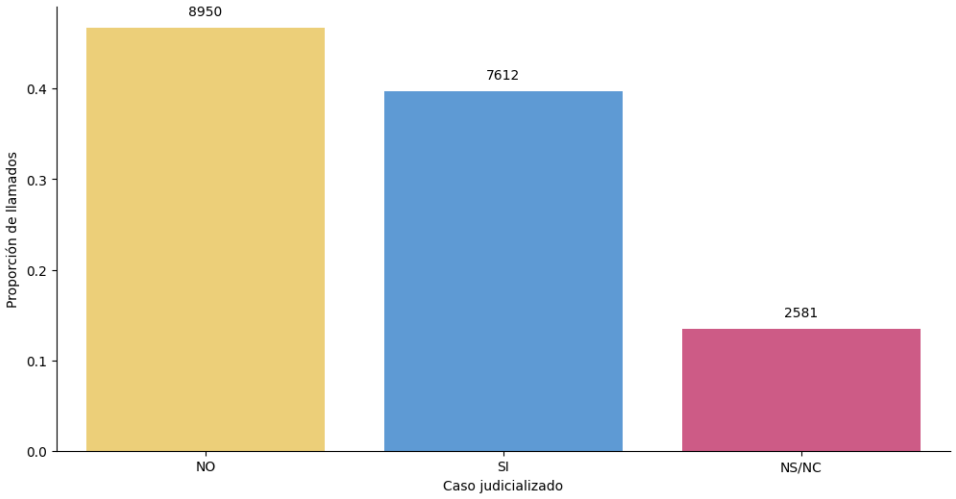

Figura 16: Caso judicializado.

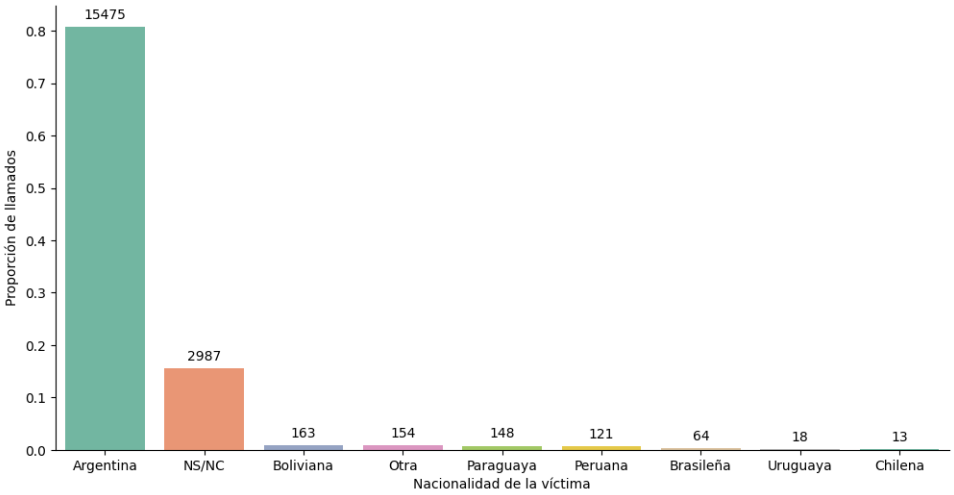

Figura 17: Nacionalidad de las víctimas.

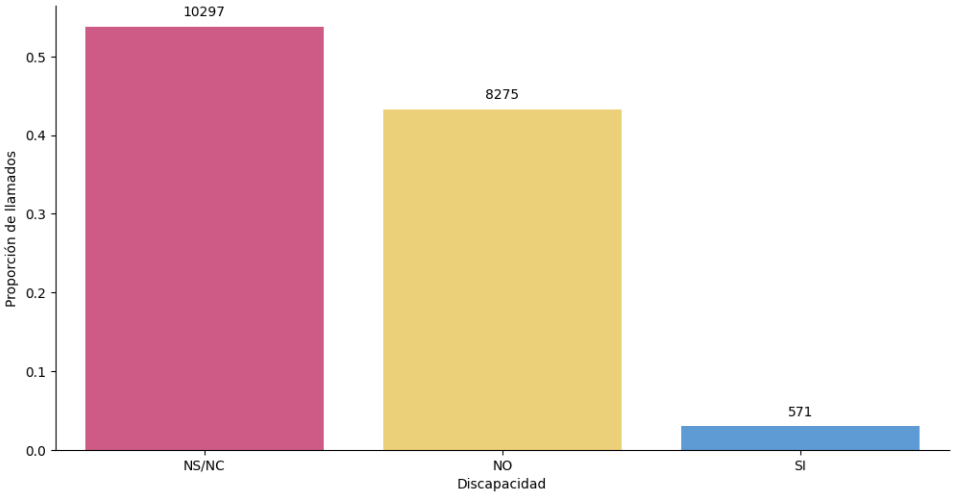

Figura 18: Presencia de discapacidad en las víctimas.

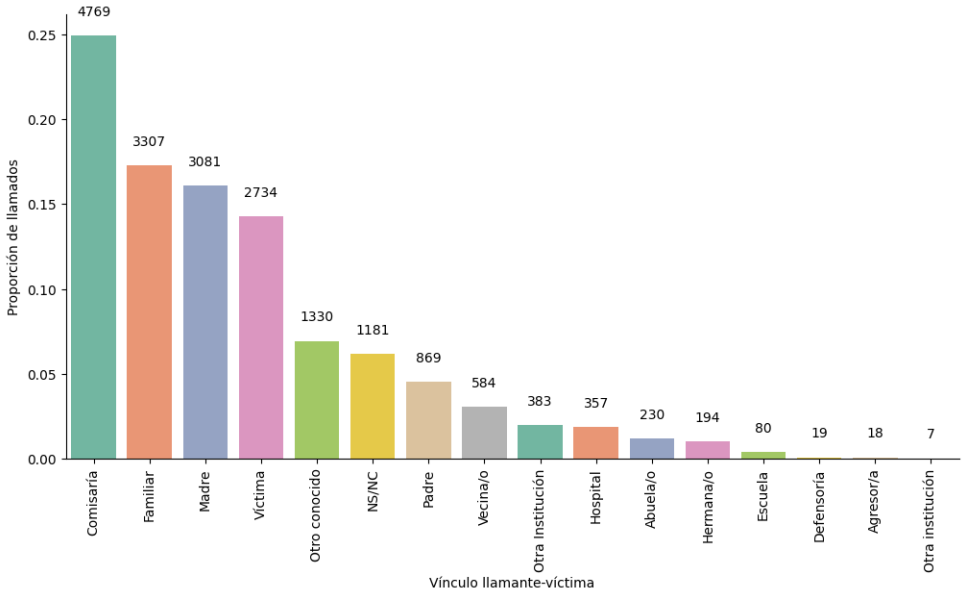

Figura 19: Vínculos víctima-llamante.

Cuadro 7: Resumen de transformaciones de variables.

***Dataset* original**	***Dataset* reducido**	**Transformación**
vs_amenazas_verbales_contenido_sexual, vs_existencia_facilitador_corrupcion_nnya, vs_eyaculacion_partes_cuerpo, 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_animal_victimizar, ofv_intento_matar	Eliminadas	No informativas (<1 %)
vs_explotación_sexual, vs_explotación_sexual_comercial, vs_explotación_sexual_viajes_turismo, vs_sospecha_trata_personas_fines_sexuales	vs_explotación_sexual	Agrupadas por dominio
vs_violacion_via_vaginal, vs_violacion_via_anal, vs_violacion_via_oral	vs_violacion	Agrupadas por dominio
vs_tentativa_violacion, vs_intento_violacion_tercera_persona	vs_tentativa_violacion	Agrupadas por dominio
llamado_provincia	Buenos Aires, C.A.B.A., Región Norte, Región Central, Región Patagónica, NS/NC	Reducción de niveles
victima_nacionalidad	Argentina, Otra, NS/NC	Reducción de niveles
hecho_lugar	NS/NC, Vivienda víctima, Vivienda agresor, Redes sociales, Espacio/transporte público, Otro	Reducción de niveles
llamante_vinculo	Institución, Conocido víctima, Víctima, Agresor, NS/NC	Reducción de niveles
agresor_vinculo	Conocido familiar, Conocido no familiar, Desconocido, NS/NC	Reducción de niveles