



Especialización en Explotación de Datos y Descubrimiento del Conocimiento

Facultad de Ciencias Exactas y Naturales

Universidad de Buenos Aires

Trabajo integrador

Exploración e imputación de datos faltantes utilizando máquinas de soporte vectorial en datos de  
llamados para reportar violencia sexual a la línea 137

---

María Victoria Colombo

17.01.2025

## Resumen

En las estadísticas sobre violencia sexual, las dificultades para recolectar datos a menudo resultan en información faltante. En este trabajo analizo llamados a la línea nacional 137 para reportar situaciones de violencia sexual. Con particular foco en la situación convivencial entre la víctima y su agresor, utilizo NMDS para buscar agrupamientos visuales en los casos según esta variable. Luego, implemento SVM con el objetivo de imputar los datos faltantes en la variable. Experimento con dos tipos de preprocesamientos para entrenar los clasificadores: uno con NMDS, y uno manual. Las visualizaciones con NMDS no resultan informativas. El preprocesamiento manual obtiene los mejores resultados en SVM. Sin embargo, el desbalance de las clases en la variable objetivo afecta largamente la performance del clasificador.

# Índice

<b>1. Introducción</b>	<b>3</b>
1.1. Antecedentes . . . . .	3
1.2. Objetivos . . . . .	4
1.3. Estructura del trabajo . . . . .	5
<b>2. Datos</b>	<b>5</b>
2.1. Obtención y limpieza . . . . .	5
2.2. Análisis exploratorio . . . . .	6
2.2.1. Análisis multivariado de <code>victima_convive_agresor</code> . . . . .	14
<b>3. Metodología</b>	<b>18</b>
3.1. Reordenamiento y reducción con NMDS . . . . .	18
3.2. Reducción manual de los datos . . . . .	19
3.3. Modelos SVM . . . . .	21
3.3.1. Entrenamiento de SVM con reducción NMDS . . . . .	21
3.3.2. Entrenamiento de SVM con reducción manual de los datos . . . . .	22
<b>4. Resultados</b>	<b>22</b>
4.1. Visualización del ordenamiento con NMDS . . . . .	22
4.2. Modelos predictivos de SVM . . . . .	24
4.2.1. Preprocesamiento con NMDS . . . . .	24
4.2.2. Preprocesamiento con reducción manual . . . . .	24
<b>5. Conclusiones</b>	<b>25</b>
<b>6. Anexo</b>	<b>28</b>

# 1. Introducción

## 1.1. Antecedentes

La violencia sexual comprende una multiplicidad de conductas o intentos de conductas dirigidos contra la sexualidad de otra persona de manera coercitiva. Los datos sobre violencia sexual 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 reportar estas situaciones, o proveer información completa. Los motivos pueden variar desde el estigma social frecuentemente asociado no solo con la violencia sexual sino con la sexualidad en general, hasta la falta de acceso a la justicia, el temor a las represalias por parte de los agresores, o al temor a que la denuncia no sea creída (Murphy-Oikonen et al., 2022). Por otro lado, otros 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 encargados de recibir denuncias.

A pesar de estas dificultades, diversos estudios a nivel mundial logran identificar patrones en la violencia sexual. Dos de los más frecuentes son: 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).

Con respecto al primer patrón, la clasificación de las identidades de género de víctimas y perpetradores es más compleja de lo que parece. Muchos estudios clasifican a los sujetos de manera binaria, omitiendo las identidades de género disidentes.<sup>1</sup> Por otro lado, los casos con víctimas hombres cis género probablemente estén subrepresentados debido a prejuicios sociales relacionados con la masculinidad que los desalientan a la hora de realizar denuncias (Ferris, 2002, p. 149).

En mi análisis las categorías de género de los sujetos se limitan a las registradas en el conjunto de datos con que trabajo: 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.<sup>2</sup> El programa cuenta con equipos de intervención de abogadas, psicólogas, y trabajadoras sociales. Al recibir una llamada solicitando asistencia se coordina el envío de equipos móviles para proveer a la víctima, en base

---

<sup>1</sup>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).

<sup>2</sup>Además, desde 2020 cuenta también con el canal de *Whatsapp* (54911) 3133-1000.

a las necesidades del caso, de contención emocional, acompañamiento a un hospital y/o a radicar una denuncia, y/o acompañamiento a un lugar seguro donde pueda alojarse (Ministerio de Justicia de la República Argentina, 2022).

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 los datos de llamados por situaciones de violencia sexual. Estos llamados presentan datos faltantes de dos maneras: celdas vacías en algunas variables numéricas, y respuestas no-sabe/no-contesta (en adelante NS/NC) en variables categóricas. Me resulta de particular interés 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. La información que provee esta variable puede ser crucial para determinar el tipo de asistencia y los pasos a seguir para asegurar el bienestar de la víctima. Teniendo en cuenta lo dicho al comienzo de la introducción sobre los desafíos para recopilar datos sobre violencia sexual, es probable que los datos faltantes en `victima_convive_agresor` estén relacionados con el motivo mismo por el que faltan. Sin embargo, determinar la razón por la que faltan es un desafío que no forma parte de este trabajo.

Luego de realizar un análisis exploratorio de los datos, elegí -por su versatilidad para operar con variables de distinto tipo- la técnica de Escalamiento Multidimensional No Métrico<sup>3</sup> para generar visualizaciones del conjunto de datos dos dimensiones y explorar agrupamientos entre las clases de la variable objetivo `victima_convive_agresor`. Luego, entrené Máquinas de Soporte Vectorial<sup>4</sup> por ser clasificadores con probado buen rendimiento tanto para datos lineales como no lineales, para clasificar los datos faltantes en la variable objetivo (respuestas NS/NC) como SI o NO, experimentando con distintas estrategias de preprocesamiento.

## 1.2. Objetivos

Mis objetivos generales para este trabajo son:

- Ampliar y aportar al campo de las investigaciones en materia de violencia de género y violencia sexual en Argentina.
- Explorar los datos en busca de patrones de ordenamiento que distingan a los grupos de víctimas que conviven con su agresor de los que no.
- Entrenar un modelo predictivo de clasificación para imputar los datos faltantes en la variable `victima_convive_agresor`.
- Contrastar y comparar distintos tipos de preprocesamientos de los datos para el entrenamiento del modelo predictivo según su impacto en el modelo final.

---

<sup>3</sup>En adelante *NMDS* por sus siglas en inglés *Non-Metric Multidimensional Scaling*

<sup>4</sup>en adelante *SVM* por sus siglas en inglés *Support Vector Machine*

Mis objetivos particulares:

- Visualizar las clases SI, NO, y NS/NC de la variable `victima_convive_agresor` utilizando la técnica de Escalamiento Multidimensional No Métrico.
- Entrenar un modelo de Máquinas de Soporte Vectorial para completar los datos faltantes de la variable `victima_convive_agresor`.
- Contrastar tipos de preprocesamientos de los datos para entrenar el modelo:
  - Reducción mediante un reordenamiento con la técnica de NMDS.
  - Reducción de los datos de manera manual, a través de la eliminación, agrupación y transformación de las variables respondiendo al conocimiento del dominio y los resultados de la exploración de los datos.
  - Eliminación de datos faltantes en variables numéricas.
  - Imputación categórica de datos faltantes en variables numéricas.
- Evaluar y comparar el rendimiento de los diferentes modelos según el preprocesamiento de los datos utilizando diferentes métricas como precisión, cobertura, F1, y *accuracy*.

Todo el código generado para este trabajo de especialización fue escrito en *python* y puede encontrarse en el repositorio público: [https://github.com/VicColombo/linea\\_137\\_llamados\\_vs](https://github.com/VicColombo/linea_137_llamados_vs)

### 1.3. Estructura del trabajo

En la sección 2 explico el proceso de obtención, limpieza, y normalización de los datos, y hago una descripción y análisis exploratorio de los datos. En la sección 3, describo y explico los métodos e implementaciones para NMDS, para la reducción manual de los datos, para el tratamiento de los valores faltantes en algunas variables específicas, y para el entrenamiento y testeo de modelos SVM. En la sección 4 expongo y comento brevemente los resultados de las visualizaciones con NMDS en 2D, y de los distintos experimentos para entrenar SVM. Finalmente, en la sección 5, reflexiono sobre los resultados expuestos y propongo ideas para enriquecer futuros trabajos de temática similar.

## 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 reportar y solicitar asistencia por situaciones de violencia sexual. Los llamados pertenecen al período entre enero de 2017 y julio de 2021.

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

- Eliminé la variable `caso_id`, que solo existe a partir de 2020.

- Cambié el nombre de la variable `llamado_provincia_indec_id` en los datos 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.
- Quitó 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. Análisis exploratorio

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.

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_eyacuacion_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_seducion, 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.



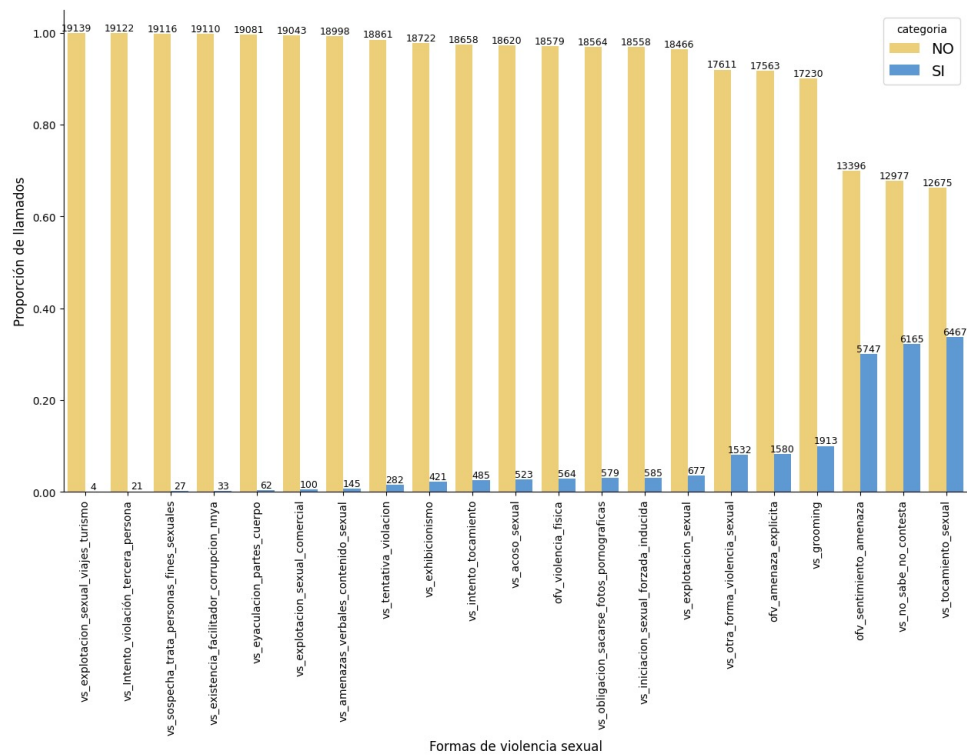


Figura 1: Tipos de violencia sexual reportada.

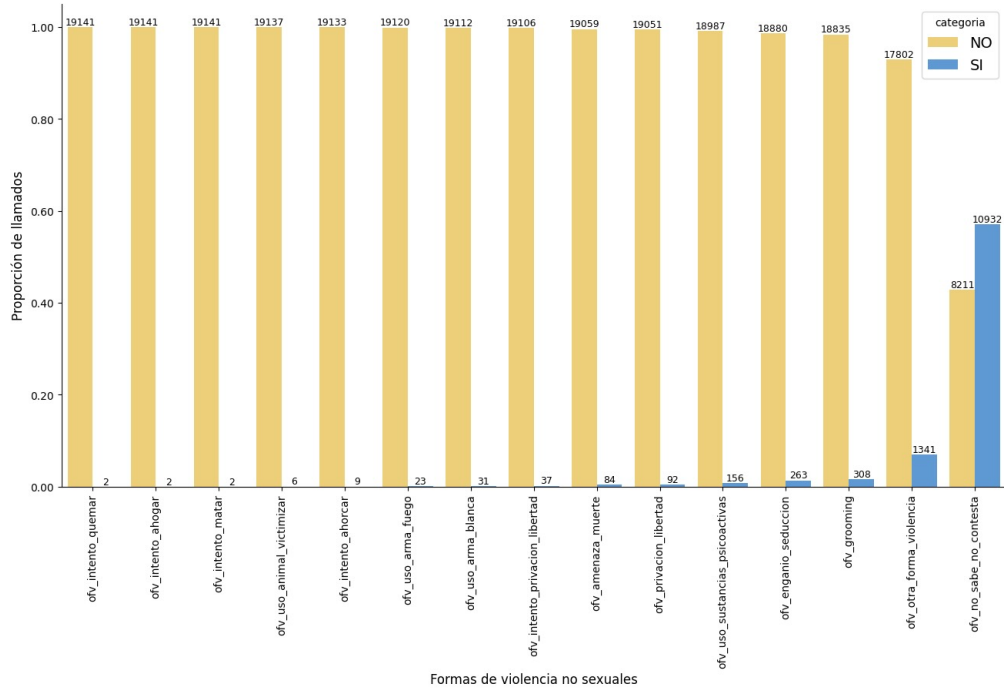


Figura 2: Tipos de violencia no sexual reportada.

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 de los datos, que excede los objetivos de este trabajo.

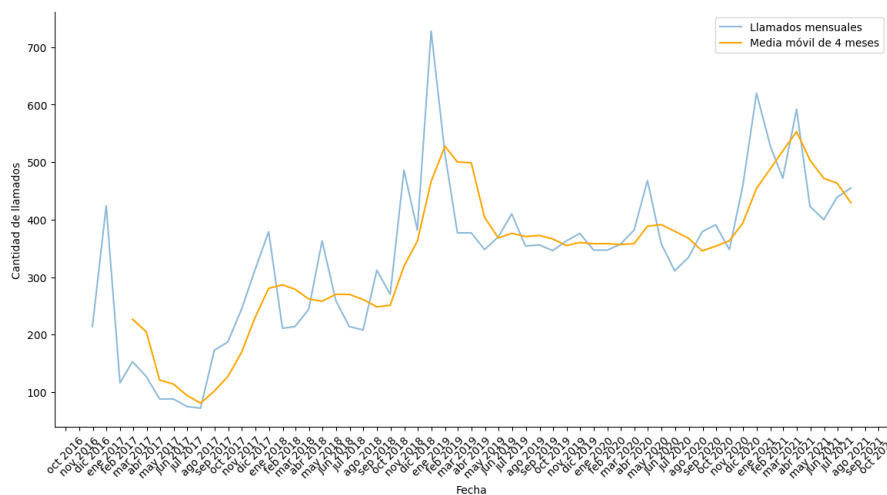


Figura 3: Cantidad de llamados en el tiempo con media móvil de 4 meses.

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.<sup>5</sup>

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.<sup>6</sup>

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

<sup>5</sup>Ver figura 17 en el Anexo.

<sup>6</sup>Ver figura 18 en el Anexo.

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.

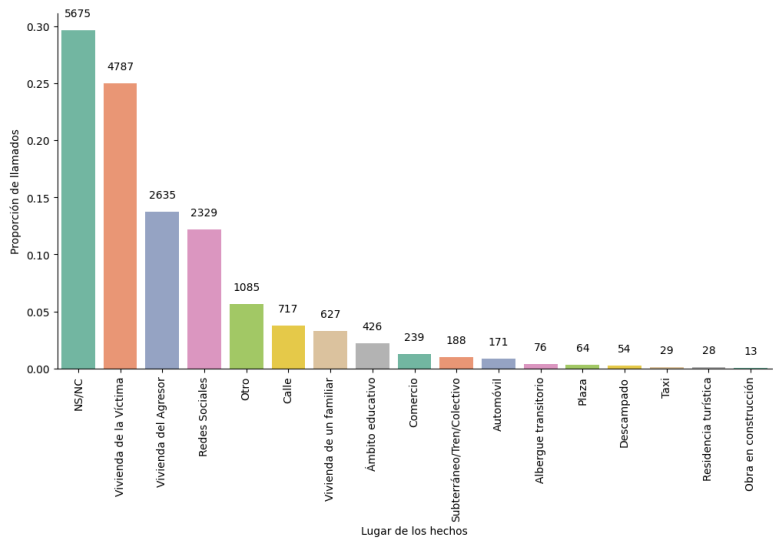


Figura 4: Lugar de los hechos.

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”.<sup>7</sup>

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<sup>8</sup>.

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.

<sup>7</sup>Ver figura 19 en el Anexo  
<sup>8</sup>Ver figura 20 en el Anexo

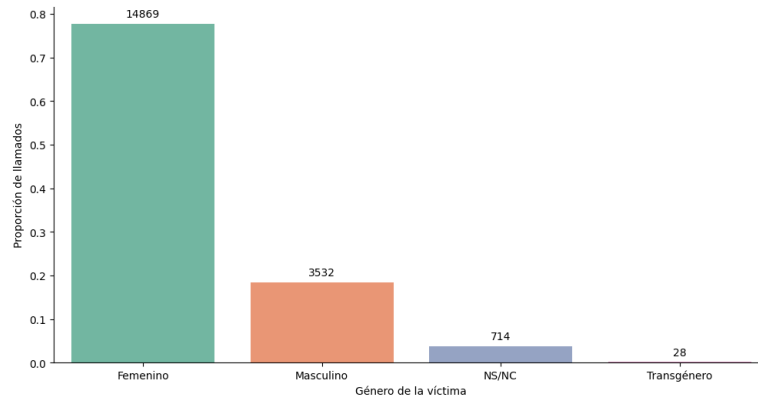


Figura 5: Género de las víctimas.

Los vínculos entre víctimas y agresores nuevamente reflejan la persistencia de los hechos de violencia sexual perpetrados 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.

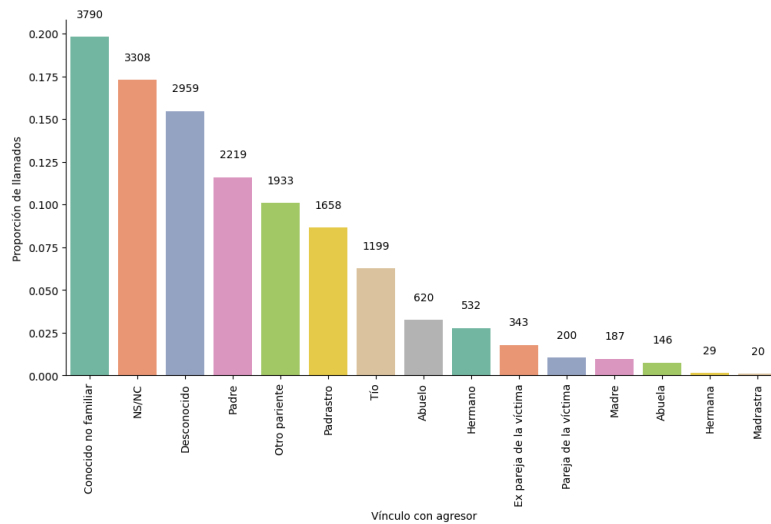


Figura 6: Vínculos víctima-agresor.

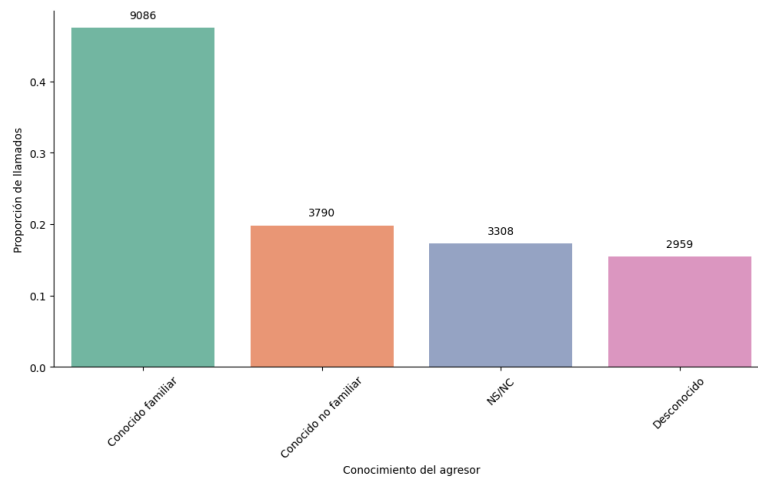


Figura 7: Agresor conocido o no por la víctima.

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 %.<sup>9</sup>

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**.

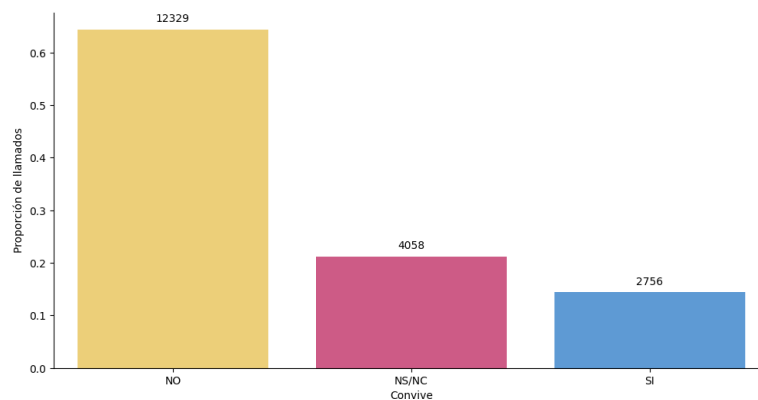


Figura 8: Convivencia víctima-agresor.

<sup>9</sup>Ver figura 21 en el Anexo

### 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.

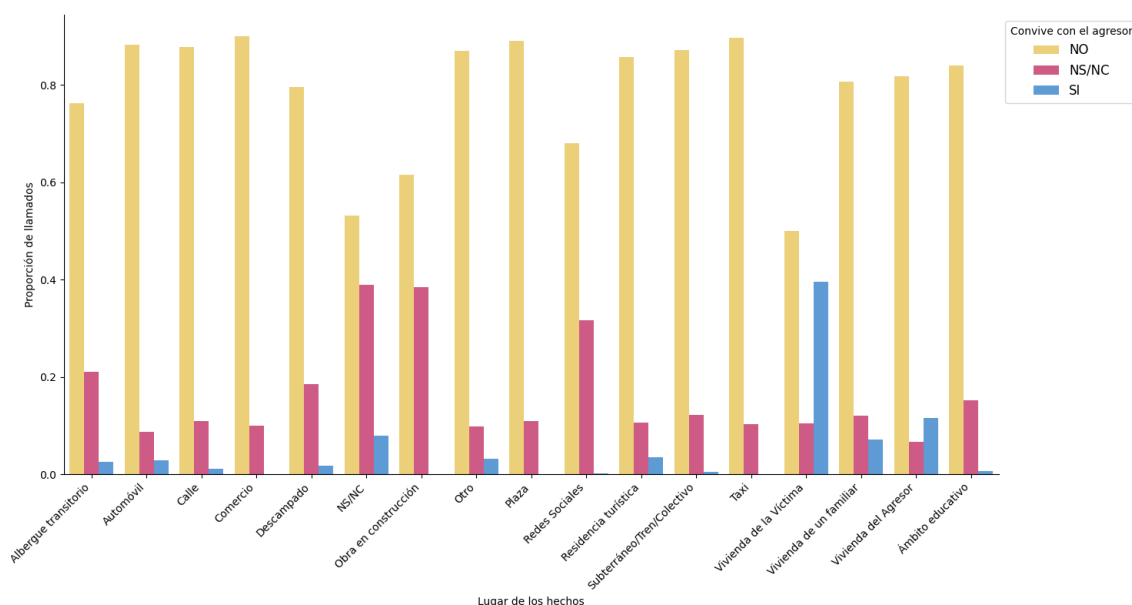


Figura 9: Convivencia con el agresor según lugar de los hechos.

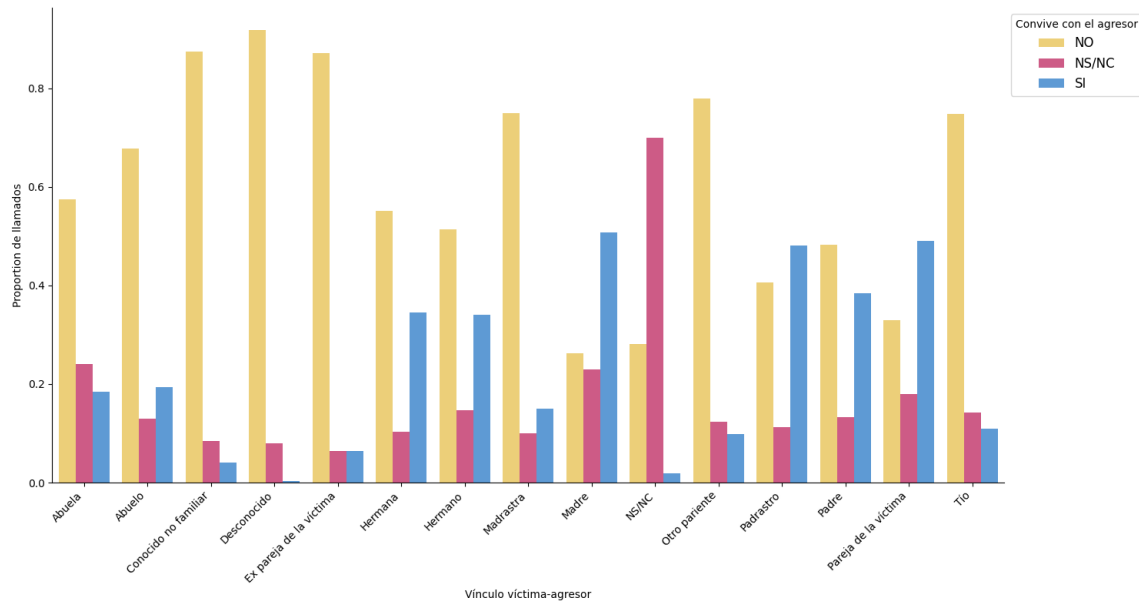


Figura 10: Convivencia con el agresor según vínculos víctima-agresor.

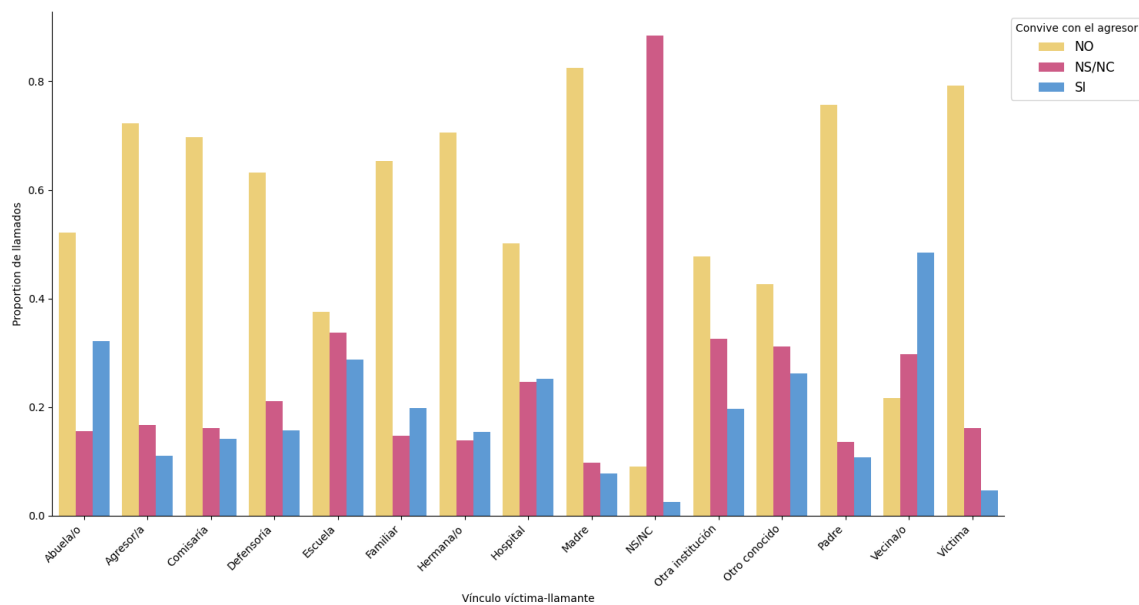


Figura 11: Convivencia con el agresor según vínculos víctima-llamante.

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 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.



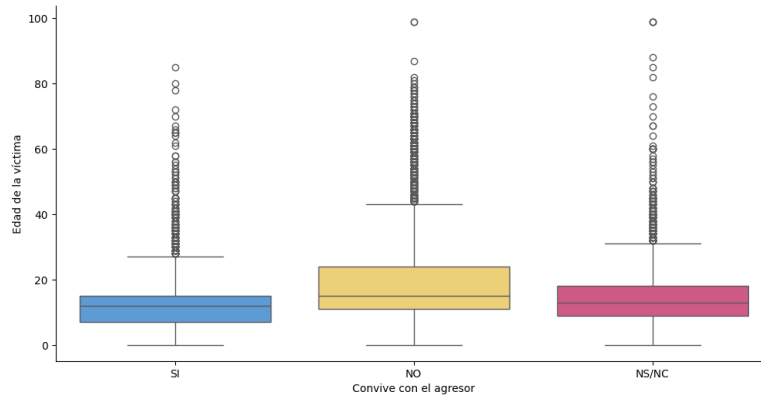


Figura 12: Distribución de la edad de la víctima según su convivencia o no con el agresor.

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.

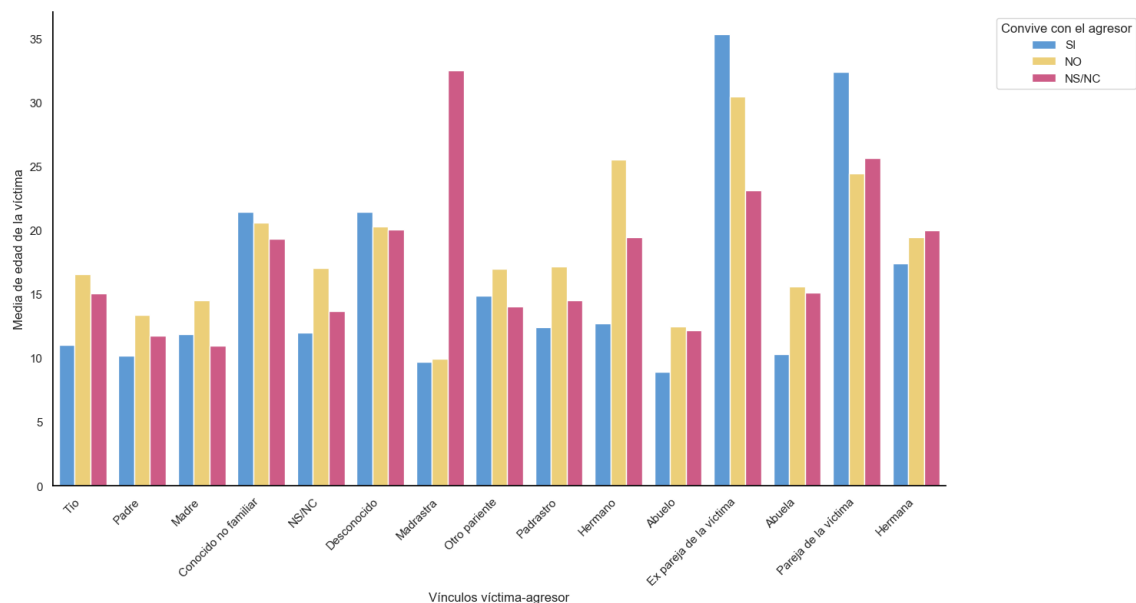


Figura 13: Edad de la víctima según su vínculo y convivencia o no con el agresor.

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 el *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.

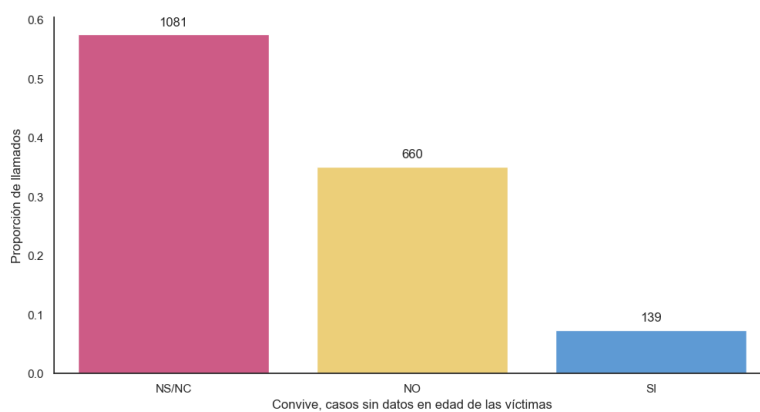


Figura 14: Convivencia víctima-agresor en las filas de datos faltantes para `victima_edad`.

### 3. Metodología

La alta dimensionalidad del conjunto de datos y la multiplicidad de niveles de muchas de las variables presenta desafíos para analizar relaciones multivariadas con la variable objetivo; además, puede causar problemas de procesamiento en el entrenamiento del modelo predictivo con SVM. Apliqué entonces dos métodos para reducir las dimensiones de los datos:

1. **Escalamiento Multidimensional No Métrico (NMDS)**: Este método me permite generar visualizaciones en 2 dimensiones para buscar agrupamientos de las clases a predecir; y también funciona como preprocesamiento para el entrenamiento de SVM.
2. **Reducción manual de los datos**: Eliminé, agrupé, y transformé las variables basándome en el conocimiento del dominio y el Análisis exploratorio.

Complementé estas estrategias de reducción con dos preprocesamientos distintos para tratar los datos faltantes en las variables de edad, ya que las implementaciones que utilizo para distancia de Gower y SVM no admiten datos vacíos. Generé dos conjuntos de datos alternativos:

- Conjunto A: categoricé las variables de edad<sup>10</sup>, y clasifiqué los datos faltantes como NS/NC.
- Conjunto B: descarté la variable `llamante_edad`, y también las filas con datos vacíos en la variable `victima_edad`.

La decisión sobre el conjunto B responde por un lado al supuesto de que la edad de la víctima es más relevante para la variable objetivo; y por otro lado, la edad de quien llama falta en el 44.82 %, y filtrar casos reduciría demasiado el conjunto de datos.

Finalmente evalué los efectos de estos métodos de reducción y los distintos preprocesamientos en el rendimiento del modelo SVM.

#### 3.1. Reordenamiento y reducción con NMDS

NMDS, un caso particular de Escalamiento Multidimensional (MDS, por sus siglas en inglés *Multidimensional Scaling*), es un método de ordenamiento que se suele utilizar para mostrar similitudes y diferencias entre los datos reorganizándolos en espacios de menor dimensión preservando el orden relativo de las distancias originales (Chan et al., 2019, p. 218). Como su nombre lo indica, acepta matrices de distancias no métricas como entrada, y eso resulta una ventaja para el conjunto de datos multivariado con el que trabajo.

La utilización de este método como preprocesamiento para entrenar SVM fue suscitada por los experimentos de Cai et al.(2019) en el campo de la microbiología, quienes encuentran resultados con buen *accuracy* aplicando esta metodología (p.69).

Implementé el método `mds` de la librería *scikit-learn*(Pedregosa et al., 2011) ajustando los siguientes parámetros:

- `n.dimensions`: probé los valores entre 2 y 7.
- `metric`: configurado `False` para aplicar escalamiento no-métrico.

---

<sup>10</sup>1 - 11 años = niñez; 12 - 18 años = adolescencia; 19 - 30 años = juventud; 31 - 65 años = adultez, más de 66 años = vejez.

- `dissimilarity`: configurado como `precomputed` para usar la matriz de distancias de Gower.
- `normalized_stress`: configurado en `True` calcula el estrés para NMDS.

Dada la naturaleza mixta de las variables del conjunto de datos, elegí la distancia de Gower para la matriz de entrada a NMDS. La similitud se calcula de a pares para variables numéricas y ordinales<sup>11</sup> con:

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

Donde:

- $s_{ijk}$  es la similitud (o distancia) entre dos individuos o filas  $i, j$  en la variable  $k$ .
- $R_k$  es el rango de  $k$ .

Para variables categóricas, la similitud entre dos puntos  $i$  y  $j$  se computa de manera binaria como 0 (los valores son idénticos, mínima distancia) o 1 (no hay similitud, máxima distancia).

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  $\delta_{ijk}$  es la cantidad total de variables o la cantidad total de variables en las que puede realizarse la comparación (Gower, 1971, p. 859-860).

A partir de la matriz de distancias originales  $S$ , NMDS calcula las coordenadas en un espacio reducido a  $n$ -dimensiones (en general  $n = 2$  o  $n = 3$  para facilitar la visualización), y a su vez las transforma 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. A continuación, se calcula el valor del estrés:

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

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

Apliqué NMDS a los conjuntos de datos A y B mencionados al comienzo de esta sección, variando los valores de `n_components`. Generé visualizaciones con las proyecciones resultantes de `n_components = 2` para evaluar la agrupación basada en `victima_convive_agresor`; y entrené modelos de SVM para cada reducción dimensional.

### 3.2. Reducción manual de los datos

Eliminé 12 de las variables que describen la violencia sufrida por considerarlas poco informativas, ya que tenían tasas de respuesta positivas menores a 1 % (ver nuevamente la figuras Tipos de violencia sexual reportada y Tipos de violencia no sexual reportada en la sección Datos).

<sup>11</sup>Si bien Podani (1999) propone un tratamiento especial para las variables ordinales largamente utilizado hoy en día, el método 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/>).

Además, agrupé en tres categorías más amplias algunas variables que comparten dominio semántico y jurídico, como se puede ver en el cuadro 5:

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

A la hora de elegir un método de encodeo para transformar los datos para SVM, basé mi elección de *one-hot encoder* en los hallazgos de Udilă (2023) en *Encoding methods for categorical data*. Comparando la performance de un modelo de SVM entrenado con distintos métodos de encodeo, el autor encuentra que *one-hot encoding* resulta en modelos con mejor *accuracy* de manera consistente (p.7).

En el mismo artículo, sin embargo, el autor advierte sobre el procesamiento potencialmente costoso (en tiempo y capacidad) en que este método incurre con variables de alta cardinalidad puede resultar demasiado alto (p.7). Por lo tanto, reduje la cardinalidad de las variables con más de 5 niveles de la siguiente manera:

- llamado\_provincia de 25 niveles a 6: Buenos Aires, C.A.B.A., Región Norte, Región Central, Región Patagónica, y NS/NC.
- victima\_nacionalidad de 9 niveles a 3: Argentina, NS/NC, y Otra.
- hecho\_lugar de 17 niveles a 6: NS/NC, Vivienda de la víctima, Vivienda del agresor, Redes sociales, Espacio/transporte público, y Otro.
- llamante\_vinculo de 16 niveles a 5: Institución, Conocido de la víctima, Víctima, Agresor, y NS/NC.
- agresor\_vinculo de 16 niveles a 4 tomando la agrupación mostrada en la figura 7, Agresor conocido o no por la víctima: Conocido familiar, Conocido no familiar, NS/NC, y Desconocido.

El conjunto de datos resultante de la reducción manual tiene 36 variables, (el original tenía 54). En el cuadro 10 del Anexo se puede ver un resumen de las variables eliminadas, agrupadas, y transformadas.

También dividí este conjunto en los tipos A y B, siguiendo el enfoque de la sección anterior, Reordenamiento y reducción con NMDS.

### 3.3. Modelos SVM

Las máquinas de soporte vectorial son modelos de aprendizaje supervisado ampliamente utilizados para tareas de clasificación que buscan encontrar el hiperplano que mejor separa las clases en el espacio original si los datos son linealmente separables, o en un espacio transformado de mayores dimensiones cuando no lo son, maximizando el margen entre los puntos de datos más cercanos de cada clase.

La capacidad de SVM para trabajar con datos que no son linealmente separables utilizando distintos *kernel tricks* es el motivo por el que las elegí, dadas las relaciones complejas entre la variable objetivo y el resto de las variables (Chan et al., 2019, p. 323-331).

En la implementación y entrenamiento de los modelos de SVM, utilicé el método `svc` de *scikit-learn* (Pedregosa et al., 2011) optimizando los hiperparámetros:

- **kernel**: define el tipo de función para transformar el espacio de entrada.
- **C**: regulariza el equilibrio entre maximizar el margen entre las clases y minimizar errores de clasificación.
- **gamma**: un peso que controla la influencia de *datapoints* individuales en la frontera de decisión del modelo.

El cuadro 6 ofrece un resumen de los distintos conjuntos de datos para los experimentos de entrenamiento de SVM.

Cuadro 6: Preprocesamientos para entrenar modelos de SVM.

Preprocesamiento	Dataset	Especificaciones
1. 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.
2. Reducción manual con métodos mixtos	A	Variables de edad transformadas a categóricas. Datos faltantes codificados como NS/NC. Aplicación de one hot encoder por categorización.
	B	Solo datos completos de edad de la víctima (numérica). Edad de quien llama eliminada.

#### 3.3.1. Entrenamiento de SVM con reducción NMDS

Preparé los datos comenzando por reemplazar los valores NS/NC en `victima_convive_agresor` por NA. Después, separé la variable objetivo `victima_convive_agresor` ( $y$ ) del resto de los datos  $X$ ; y quité a  $y$  las filas vacías, guardando sus índices. Calculé a partir de  $X$  label matriz de distancias de Gower para utilizar en NMDS.

Generé distintas reducciones NMDS variando el parámetro `n_components`. Para cada reducción, 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 vacía en `victima_convive_agresor`. Dividí conjunto principal  $X$  e  $y$  de manera estratificada en entrenamiento (80 %) y testeo (20 %) de manera estratificada utilizando `StratifiedShuffleSplit` de *scikit-learn* (Pedregosa et al., 2011) para mitigar el desbalance de las clases SI y NO.

Para el entrenamiento, implementé una búsqueda de hiperparámetros (`kernel`, `C`, `gamma`) mediante validación cruzada de 5 divisiones, evaluando el rendimiento con la métrica `f1_weighted`. Elegí esta métrica tanto por el desbalance entre las clases (*accuracy* o *precision* pueden arrojar valores para la clase predominante), como para intentar garantizar un modelo que balanceara precisión y cobertura, ya que me interesa no solo predecir de manera correcta los verdaderos positivos, sino también, no dejar afuera potenciales casos positivos. Además, calculé el estrés asociado a cada reducción NMDS como una métrica adicional de evaluación.

Finalmente, probé el modelo entrenado con los mejores hiperparámetros en el conjunto ciego. Aunque no me es posible realizar una evaluación tradicional del conjunto final porque son realmente datos faltantes, comparé las proporciones de clases SI y NO entre el conjunto original con las predicciones para obtener una estimación de la performance del modelo.

### 3.3.2. Entrenamiento de SVM con reducción manual de los datos

Para este experimento, comencé por codificar las variables. Para `llamado_fecha_hora`, `momento_dia`, `estacion_del_año`, y las variables de edad (en el caso del conjunto A) utilicé un encoder ordinal (`OrdinalEncoder` de *scikit-learn* (Pedregosa et al., 2011)) para preservar la naturaleza ordinal de los datos. En el caso del conjunto B, escalé la edad de la víctima.

Transformé las variables binarias (SI, NO), incluida la variable objetivo, asignando `SI = 1` y `NO = 0`. Al igual que en el experimento anterior, los valores NS/NC en la variable objetivo los reemplacé por datos vacíos. Para el resto de las variables categóricas, utilicé `OneHotEncoder`.

Generé el conjunto de prueba ciega de manera similar al enfoque NMDS, utilizando las filas correspondientes a casos con NS/NC en la variable objetivo. Entrené el modelo fue entrenado con la misma metodología de validación cruzada y búsqueda de hiperparámetros (`kernel`, `C`, `gamma`), optimizando `f1_weighted`.

## 4. Resultados

### 4.1. Visualización del ordenamiento con NMDS

Las figuras 15 y 16 muestran que en el reordenamiento en dos dimensiones de los datos las categorías SI, NO, y NS/NC no presentan una separación clara. SI y NO aparecen mezcladas entre sí, y, por consiguiente, NS/NC no se agrupa distintivamente cerca de ninguna de las dos clases, ni de forma aislada. Además, los valores de estrés obtenidos en ambas configuraciones de NMDS son considerablemente altos: 0.3, cuando que un valor óptimo suele encontrarse alrededor de 0.05.

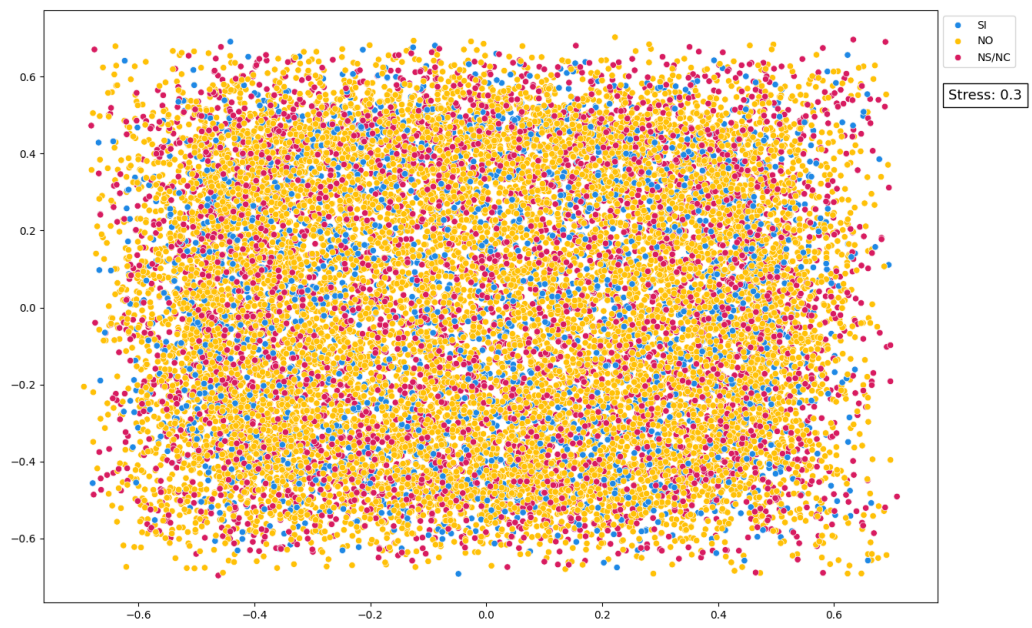


Figura 15: Visualización de los datos reducidos con NMDS - conjunto A



Figura 16: Visualización de los datos reducidos con NMDS - conjunto B



## 4.2. Modelos predictivos de SVM

### 4.2.1. Preprocesamiento con NMDS

Los valores de estrés para los ordenamientos aún en dimensiones  $> 2$  resultaron nuevamente altos, siendo el más bajo 0.16 para 7 dimensiones.

Para los conjuntos de datos A y B, y para todas las variaciones de `n_components`, la optimización de hiperparámetros encontró siempre los mejores resultados con:

- `kernel: polynomial`
- `C: 0.1`
- `gamma: auto`

La selección del `kernel` polinomial da cuenta de la no linealidad de los datos en el espacio de entrada. El valor bajo de la penalización `C` indica una preferencia por generalizar antes que por minimizar los errores de entrenamiento. La configuración de `gamma` en *auto* hace que el valor resultante para este parámetro resulte de  $\frac{1}{n\_components}$ :  $0,5 \leq \gamma \leq \pi$ , ya que  $2 \leq n\_components \leq 7$ .

La performance de los modelos también se igualó en ambos conjuntos de datos y para todos los `n_components`, puede verse resumida en el cuadro 7 a continuación. Los modelos resultan sumamente afectados por el desbalance de clases y no generalizan bien para la clase minoritaria. Las métricas F1, cobertura, y precisión son buenas para la clase mayoritaria NO (0.90, 1.00, y 0.82 respectivamente), indicando que son modelos capaces de identificar correctamente la mayoría de los puntos de esta clase. Para la clase minoritaria SI, si bien los modelos no incurren en falsos positivos en la clasificación (de ahí la precisión de 1.0), tampoco pueden identificar correctamente ningún caso perteneciente a la clase, haciendo que la cobertura, y por lo tanto F1, sean 0.

Las métricas F1 macro (0.45) y cobertura macro (0.50), que promedian el rendimiento en ambas clases, dan cuenta del problema del desbalance de los datos. La métrica general de *accuracy* por el contrario resulta engañosa por sí sola, pero su valor alto (0.82) responde al sobreajuste del modelo para la clase mayoritaria. Otro tanto sucede con la macro precisión que promedia la precisión por clase, y como comenté anteriormente, la precisión para la clase minoritaria solo es alta porque el modelo no incurre en falsos positivos.

Cuadro 7: Performance de modelos SVM utilizando NMDS como preprocesamiento.

Clase	F1	Cobertura	Precisión	Macro F1	Macro Cobertura	Macro Precisión	Accuracy
0 (NO)	0.90	1.00	0.82	0.45	0.50	0.91	0.82
1 (SI)	0.00	0.00	1.00				

Al aplicar los mejores modelos a los sets de testeo final ciegos para A y B, encontré que identificaban todos los casos como pertenecientes a la clase mayoritaria NO.

### 4.2.2. Preprocesamiento con reducción manual

La optimización de hiperparámetros encontró los mejores resultados con:

### Conjunto A

- `kernel: polynomial`
- `C: 1`
- `gamma: scale`

### Conjunto B

- `kernel: rbf`
- `C: 100`
- `gamma: auto`

Ambos *kernels* apropiados para datos no lineales. El mejor valor del parámetro de regularización `C` es significativamente más alto que para el experimento anterior, sobre todo en el caso del conjunto B.

Al igual que en el experimento anterior, las métricas de performance se igualaron para los modelos entrenados con cada conjunto A y B. Los resultados pueden verse en el cuadro 8. En este segundo experimento el modelo resulta mejor clasificador de la clase minoritaria que en el experimento anterior mirando F1, cobertura, y precisión para NO (0.62, 0.54, y 0.72 respectivamente), y mantiene la buena performance para la clase mayoritaria (0.93, 0.95, y 0.90 respectivamente). Si bien la mejora es notable, las cifras alcanzadas no hablan de un modelo confiable para clasificar casos de SI, de hecho la cobertura parece estar cerca de una clasificación por azar. La precisión de 0.72 en SI es la más alta de las tres métricas para esta clase porque responde al mismo fenómeno que en el experimento anterior.

Las métricas F1 macro (0.77), cobertura macro (0.75) también mejoraron con respecto al experimento anterior. La precisión macro (0.81) y la *accuracy* (0.88) responden a los mismos fenómenos que en el experimento anterior aunque esta vez la diferencia de performance por clase no es tan extrema.

Cuadro 8: Performance de modelos SVM reducción manual como preprocesamiento.

Clase	F1	Cobertura	Precisión	Macro F1	Macro Cobertura	Macro Precisión	Accuracy
0 (NO)	0.93	0.95	0.90	0.77	0.75	0.81	0.88
1 (SI)	0.62	0.54	0.72				

Al comparar las proporciones de SI y NO predichas por los mejores modelos en los sets de testeo ciegos con la proporción existente en el conjunto de datos original<sup>12</sup>, encontré los resultados resumidos en el cuadro a continuación:

Cuadro 9: Proporciones predichas vs. originales para cada clase.

Conjunto	SI original	SI predicho	NO original	NO predicho
A	0.18	0.09	0.82	0.91
B	0.18	0.10	0.82	0.90

## 5. Conclusiones

Dadas las falencias de las representaciones con NMDS (las visualizaciones que no aportaron información útil, la apariencia aleatoria de la distribución de los datos, y los valores de estrés deficientes), en el futuro podría explorar otras metodologías de reducción y visualización, como por ejemplo PCA, mediando siempre una transformación

<sup>12</sup>Tomo los datos post procesamiento manual, sin las filas NS/NC en cada caso A y B.

adecuada de los datos. Es posible también que en este conjunto de datos los patrones relacionados con la variable objetivo sean demasiado complejos para ser representados en dos o tres dimensiones.

Aunque el preprocesamiento manual del dataset produjo resultados notablemente superiores en el entrenamiento de los clasificadores/imputadores de datos faltantes, ambos experimentos sufrieron el desbalance de clases y no podría decir que ningún modelo constituye una fuente confiable para la imputación de los datos.

El preprocesamiento con NMDS fue el que produjo los peores modelos. Los distintos valores de `n_components` y las dos estrategias para manejar valores faltantes de edad no afectaron el desempeño de los modelos. El valor bajo de la regularización `C` (0.1) podría haber contribuido al sobreajuste a la clase mayoritaria en el primer experimento con NMDS. En el segundo experimento, con valores de `C` más altos (1 para el dataset A y 100 para el dataset B), el modelo funcionó mejor. Sin embargo, esto no implica que `C` sea el único factor que explica la diferencia de rendimiento, ya que durante la optimización de hiperparámetros en el primer experimento también probó valores más altos de `C` sin obtener mejores resultados.

El parámetro `gamma`, configurado como `scale` en el segundo experimento y como `auto` en el primero, parece no haber tenido un impacto significativo por sí solo.

Para trabajos futuros, considero importante implementar estrategias que mitiguen el desbalance de clases, como la generación artificial de casos de la clase minoritaria, o el sobremuestreo o el submuestreo de las clases.

A modo general, me resulta interesante también la posibilidad de cruzar datos de diferentes fuentes sobre violencia de género y violencia sexual disponibles en Argentina, como los datos de la línea 144 para denunciar violencia de género, los registros de la OVD (Oficina de Violencia Doméstica), y los datos que pudiera aportar el Observatorio de Datos de Género.

Finalmente, en lo referente al género de víctimas y victimarios, y volviendo a un comentario hecho en la introducción, es esencial que la recolección de datos se realice con conciencia y perspectiva de género. Como señala Rita Segato en *La guerra contra las mujeres* (2016), la violencia sexual está dirigida hacia cuerpos femeninos o *feminizados*. Esto incluye cuerpos percibidos como menores, débiles, racializados o pertenecientes a disidencias sexuales, lo cual coincide con datos que indican una mayor incidencia de violencia sexual contra hombres durante la niñez y adolescencia, cuando los sujetos son más vulnerables y, por lo tanto, feminizados (Contreras et al., 2016; Unidad Fiscal Especializada en Violencia contra las Mujeres (UFEM), 2023; Ferris, 2002).

A partir de estas reflexiones, propongo una pregunta para futuros estudios: si las denuncias de violencia sexual contra disidencias de género son una minoría en los datos, ¿significa esto que esas personas sufren menos violencia sexual? ¿O refleja una subrepresentación general debido a su posición como minorías sociales y su limitado acceso a la justicia?

## 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.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- János Podani. Extending gower’s general coefficient of similarity to ordinal characters. *Taxon*, 48(2):331–340, 1999.
- 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

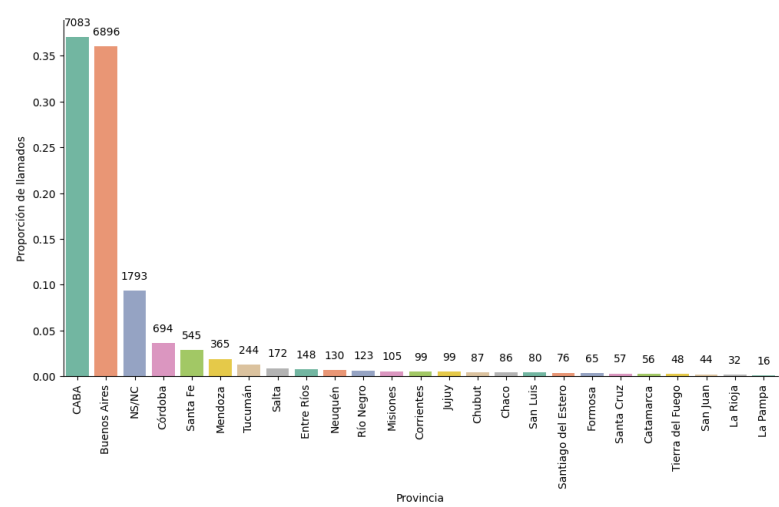


Figura 17: Llamados por provincia.

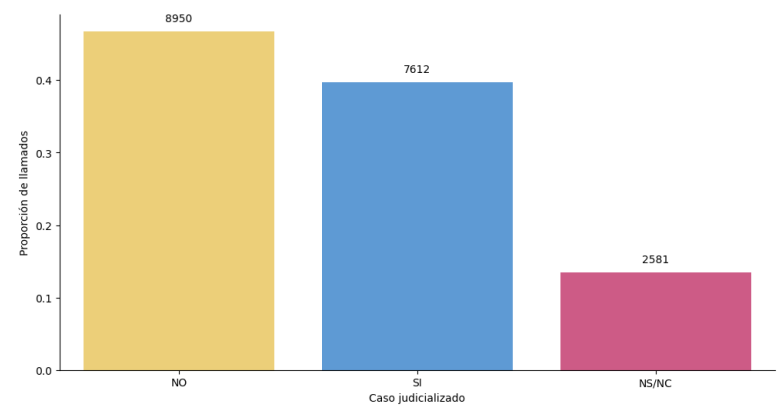


Figura 18: Caso judicializado.

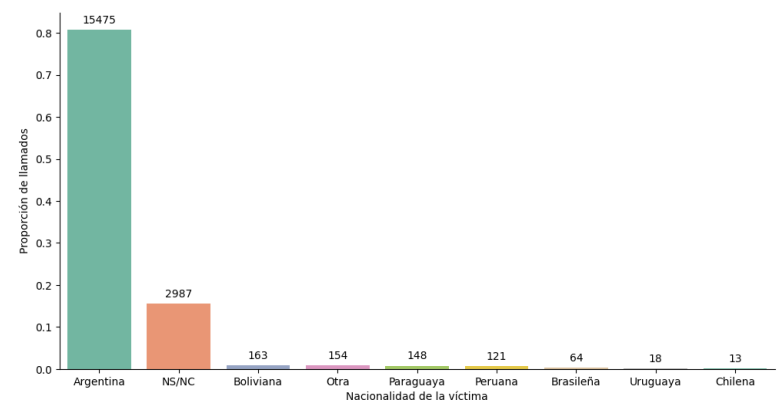


Figura 19: Nacionalidad de las víctimas.

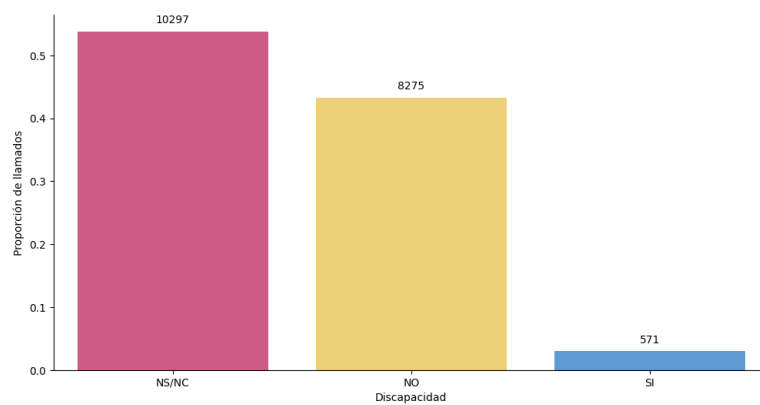


Figura 20: Presencia de discapacidad en las víctimas.

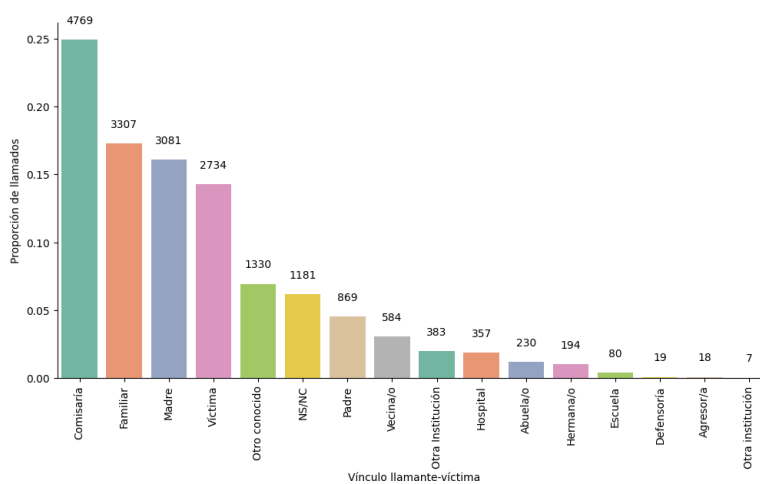


Figura 21: Vínculos víctima-llamante.

Cuadro 10: Resumen de transformaciones de variables.

<i>Dataset original</i>	<i>Dataset reducido</i>	<i>Transformación</i>
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