This specialization thesis was originally written in Spanish. It was translated to English using an LLM and posterior human supervision.

Specialization in Data Mining and Knowledge Discovery

Facultad de Ciencias Exactas y Naturales

Universidad de Buenos Aires

Specialization Thesis

Exploratory Analysis and Missing Data Imputation Using Support Vector Machines in Calls Reporting Sexual Violence to Helpline 137

Victoria Colombo

17.01.2025

Resumen

Statistical reporting on sexual violence is often hindered by data collection challenges, which frequently result in missing information. This study examines calls to Argentina's national helpline 137 concerning incidents of sexual violence. Focusing on the cohabitation status between victim and perpetrator, I employ NMDS to investigate potential clustering patterns in the cases through visualizations. Then, I implement SVM to impute missing values for the cohabitation variable. In doing so, I compare two preprocessing strategies for training the classifier: one using NMDS and another developed manually. The NMDS-based visualizations offer little interpretive value, while the manual preprocessing approach achieves superior classification performance. Nonetheless, the pronounced class imbalance in the target variable substantially undermines the accuracy of the classifier.

1. Introduction

1.1. Background

Sexual violence encompasses a wide range of behaviors—or attempted behaviors—coercively directed at another person's sexuality. Data on sexual violence are often scarce or contain substantial missingness (Ferris, 2002, p. 150). One reason is that victims or their social circles often refuse to report these situations or to provide complete information. Motives range from the social stigma frequently associated not only with sexual violence but with sexuality in general, to barriers to justice, fear of retaliation by the aggressors, or fear that the report will not be believed (Murphy-Oikonen et al., 2022). On the other hand, additional reasons for scarce and/or low-quality data include the lack of appropriate channels for collecting this information, or negligence or unfamiliarity with proper procedures among officials responsible for receiving reports.

Despite these hurdles, studies around the world have identified patterns in sexual violence. Two of the most common are: most victims are women, while perpetrators are usually men (Ferris, 2002, p. 149; Contreras et al., 2016, p. 15); and in most cases the aggressors are people known to the victims, such as partners, ex-partners, or other acquaintances (García-Moreno et al., 2005, p. 9,Unicef et al., 2018, p. 22, Ferris, 2002, p. 151).

With regard to the first pattern, classifying the gender identities of victims and perpetrators is more complex than it may seem. Many studies classify subjects in a binary way, omitting dissident gender identities.[1] In addition, cases with cisgender male victims are likely underrepresented due to social prejudices related to masculinity that discourage reporting (Ferris, 2002, p. 149).

In my analysis, the gender categories of the subjects are limited to those recorded in the dataset I work with: male, female, and transgender, without specifying whether transgender refers to a trans man or a trans woman. I recognize this as a limitation of my work and of the available data.

The collection, systematization, and analysis of data on sexual violence by the State are crucial for planning and implementing effective policies for prevention, assistance, and eradication of sexual violence. In Argentina, although there is no single centralized state system for this type of information, there are judicial entities and public programs that, in addition to providing assistance and/or access to justice, collect data on sexual violence and maintain public records. One of these programs is Las Víctimas Contra las Violencias.

Since 2016, within the framework of the Las Víctimas Contra las Violencias program under the Ministry of Justice, helpline 137 has operated 24 hours a day to provide assistance in cases of sexual or domestic violence.[2] The program has intervention teams composed of attorneys, psychologists, and social workers. Upon receiving a call requesting assistance, mobile teams are dispatched to provide the victim—depending on the needs of the case—with emotional support; accompaniment to a hospital and/or to file a report; and/or accompaniment to a safe place for shelter (Ministerio de Justicia de la República Argentina, 2022).

Records of calls to helpline 137 have been digitized since 2017 and are available on the Open Data Portal of the Argentine Justice Ministry. Four types of datasets are published there:

[1] Among the studies and reports consulted for this work, only the *Relevamiento de fuentes secundarias de datos sobre violencia sexual* by Unidad Fiscal Especializada en Violencia contra las Mujeres (UFEM) (2023) mentions gender identities when it specifies that sexual violence "particularly affects cis women and LGBTI+ people" (p. 7).

[2] In addition, since 2020 it has also offered a WhatsApp channel: (54911) 3133-1000.

- Calls concerning domestic violence
- Calls concerning sexual violence
- In-home interventions related to domestic violence
- In-home interventions related to sexual violence

In this work, I analyse the call concerning sexual violence. These calls present missing data in two ways: empty cells in some numerical variables, and "doesn't know/no answer" responses (hereafter `N/A`) in categorical variables. I am particularly interested in the variable `victima_convive_agresor`, which encodes the cohabitation status between the victim and the aggressor and takes the values `YES`, `NO`, and `N/A`. The information this variable provides can be crucial for determining the type of assistance and the steps to follow to ensure the victim's well-being. Given the challenges noted at the beginning of the introduction regarding the collection of data on sexual violence, it is likely that the missingness in `victima_convive_agresor` is related to the very reasons for the missing data. However, identifying why values are missing is a challenge that falls outside the scope of this work.

After conducting an exploratory analysis of the data, I chose—due to its versatility in handling variables of different types—the technique of Non-Metric Multidimensional Scaling (hereafter *NMDS*) to generate two-dimensional visualizations of the dataset and to explore clustering among the classes of the target variable `victima_convive_agresor`. I then trained Support Vector Machines (hereafter *SVM*), classifiers with proven strong performance for both linear and non-linear data, to classify the missing values (`N/A` responses) in the target variable as `YES` or `NO`. I experimented with different preprocessing strategies on the data to train the classifiers.

1.2. Objectives

My general objectives for this work are:

- To broaden and contribute to the field of research on gender-based and sexual violence in Argentina.
- To explore the data for ordering patterns that distinguish the group of victims who cohabit with their aggressor from those who do not.
- To train a predictive classification model to impute missing data in the variable `victima_convive_agresor`.
- To contrast and compare different types of data preprocessing for training the predictive model according to their impact on the final model.

My specific objectives are:

- To visualize the `YES`, `NO`, and `N/A` classes of the variable `victima_convive_agresor` using the Non-Metric Multidimensional Scaling technique.
- To train a Support Vector Machine model to complete the missing data for the variable `victima_convive_agresor`.
- To contrast types of data preprocessing for model training:

    - Reduction via reordering with the NMDS technique.
    - Manual reduction of the data by deleting, grouping, and transforming variables in response to domain knowledge and the results of data exploration.
    - Deletion of missing data in numerical variables.
    - Categorical imputation of missing data in numerical variables.
- To evaluate and compare the performance of the different models according to data preprocessing using different metrics such as precision, recall, F1, and *accuracy*.

All the code generated for this specialization project was written in *python* and can be found in the public repository: `https://github.com/VicColombo/linea_137_llamados_vs`

1.3. Structure of the thesis

In section 2 I explain the process of obtaining, cleaning, and normalizing the data, and I provide a description and exploratory analysis. In section 3, I describe and explain the methods and implementations for NMDS, for manual data reduction, for handling missing values in specific variables, and for training and testing SVM models. In section 4 I present and briefly discuss the results of the 2D NMDS visualizations and the different experiments for training SVMs. Finally, in section 5 I reflect on the results presented and propose ideas to enrich future work on similar topics.

2. Data

2.1. Acquisition and cleaning

I downloaded from the Open Data Portal mentioned above five *csv* datasets of calls to helpline 137 to report and request assistance in situations of sexual violence. The calls span the period from January 2017 to July 2021.

Once downloaded, I unified the five files into a single dataset. This required an initial cleaning to make the files consistent with one another in terms of the number and names of columns:

- I removed the variable `caso_id`, which exists only from 2020 onward.
- I changed the name of the variable `llamado_provincia_indec_id` in the 2017, 2018, and 2019 data to match its 2020–2021 equivalent: `llamado_provincia_id`.

The next step was to clean the unified dataset of inconsistencies and various data-entry errors:

- I standardized, for all relevant variables, the values `YES`, `NO`, and `N/A` in uppercase, since they appeared in different formats: lowercase, title case, etc.
- In the variable `victima_vinculo_agresor`, I unified the value `Ex pareja de la víctima`, which also appeared as `Ex pareja`, `Ex-pareja de la víctima`, and `Expareja de la víctima`. I did the same with `Pareja de la víctima`, which presented similar variants.

- In `hecho_lugar`, I unified two variants of the same category: `Otra institución` and `Otra Institución`, opting for the first form.
- I replaced all `Sin datos` values with `N/A`, considering them equivalent.
- I trimmed leading and trailing spaces from strings to resolve issues such as `Madre` $=/=$ `Madre`.
- In the variable `llamante_vinculo`, I converted the value `Vecino` to `Vecina/o`, since it does not necessarily refer exclusively to people of the male gender.
- In `llamado_provincia`, I unified "Ciudad Autónoma de Buenos Aires" and "CABA", opting for "CABA".
- In `llamado_provincia`, I corrected instances of "Santa Fé" to "Santa Fe".

2.2. Exploratory analysis

The unified dataset consists of 19,143 observations and 54 variables, most of them categorical, providing information about the victim, the aggressor, the caller, the context of the incident, and the type of violence experienced. Table 1 details the variables and their types.

Table 1: Summary of variables.

Descriptor	Variable type	Variable(s)
Victim	Quantitative	victima_edad
	Qualitative	victima_genero, victima_nacionalidad, victima_discapacidad, victima_vinculo_agresor, victima_convive_agresor, victima_a_resguardo
Caller	Quantitative	llamante_edad
	Qualitative	llamante_genero, llamante_vinculo
Call	Ordinal	llamado_fecha_hora
	Qualitative	caso_id, llamado_provincia llamado_provincia_id, caso_judicializado, hecho_lugar
Sexual violence	Qualitative	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
Other forms of violence	Qualitative	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

The variables describing sexual violence experienced and other forms of reported violence can take the values `YES` or `NO`, with the latter being the most common, as shown in Figures 1 and 2, which display the distribution of responses for sexual and non-sexual violence respectively. It is noteworthy to observe the volume of positive responses for the categories `vs_no_sabe_no_contesta` and `ofv_no_sabe_no_contesta`. In other words, in a large number of calls, a form of violence (sexual or otherwise) is reported, but the specific type cannot be identified. Throughout this section, this prevalence of `N/A` responses can be observed in nearly all variables.

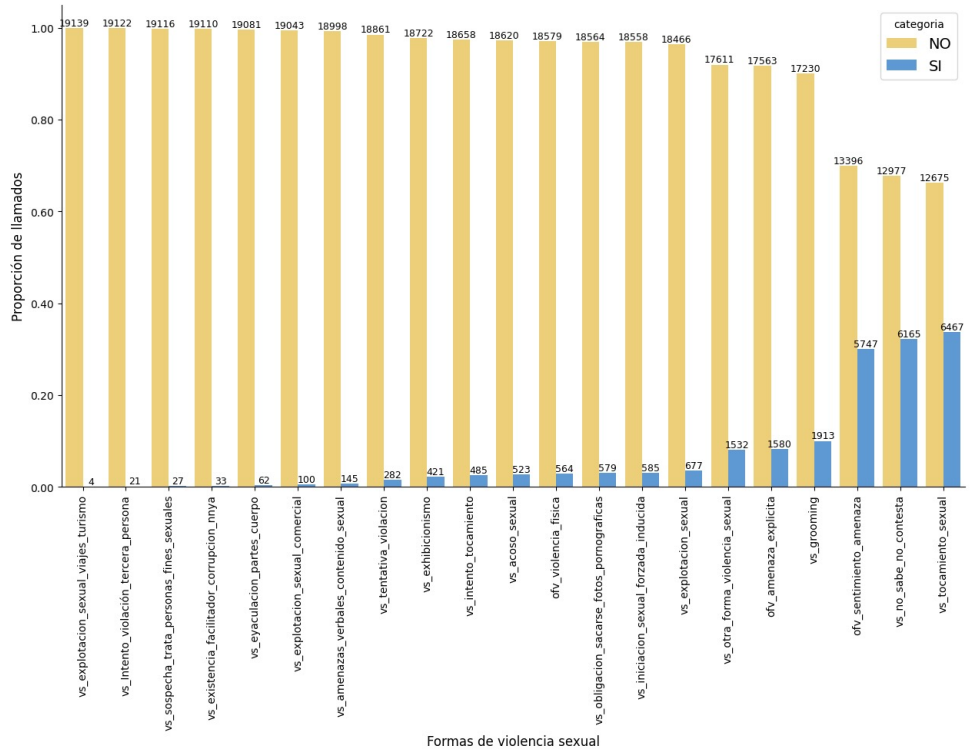

Figure 1: Types of sexual violence reported.

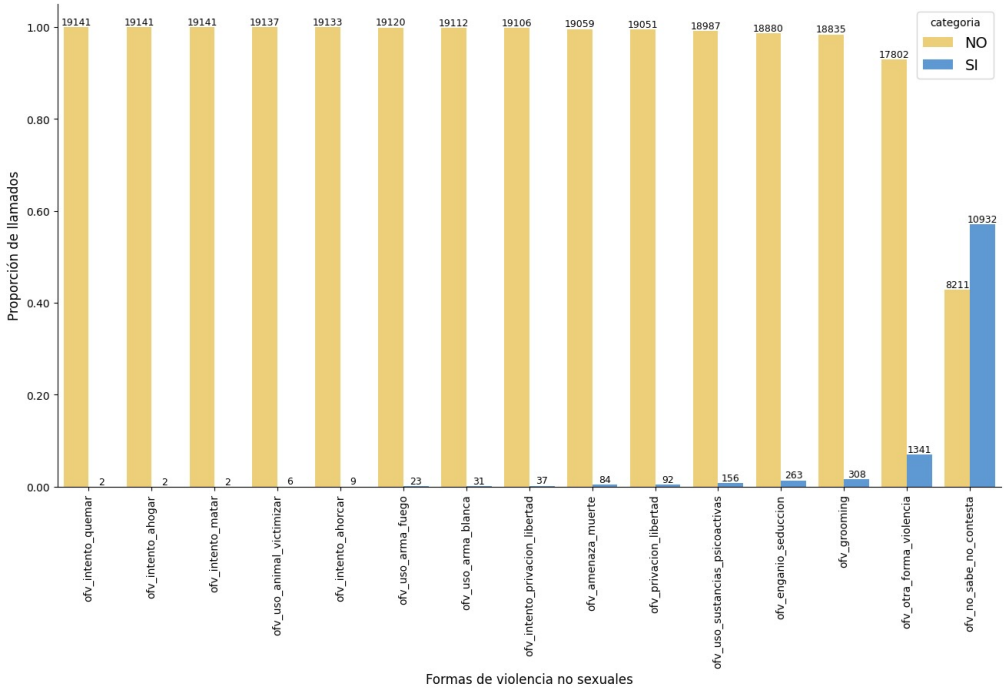

Figure 2: Types of non-sexual violence reported.

The variables `victima_edad` and `llamante_edad` contained extreme positive outliers, not only identifiable for

exceeding the $3 * IQR$ threshold but also—more importantly—because they were inconsistent with plausible human ages. I therefore removed all values above 110 for both variables, and all values below 1 for `llamante_edad` (if victims appear with age 0, I interpret them as infants not yet one year old). The removed values and their frequencies per variable can be seen in Table 2. As shown there, most were 999 in both variables, very likely a default entry used to avoid leaving the field empty. In total, I removed 195 values from `llamante_edad`, and 101 values from `victima_edad`.

Table 2: Outliers in age variables.

Variable	**Outlier**	**Row count**
llamante_edad	999	192
	0	3
victima_edad	999	98
	224	1
	125	1
	111	1

After removing these values, I calculated descriptive measures for the age variables, shown in Table 3. Most victims are under 21 years old, with a mean of 17 and a mode of 14. In contrast, the callers are mostly adults, with a mean age of 36 and a mode of 40. This reinforces what was mentioned in the Introduction and in other studies: young people, especially adolescents and children, are among the groups most at risk of sexual violence.

Table 3: Descriptive measures of age variables.

Descriptor	**Caller's age**	**Victim's age**
Mean	36.25	17.17
Mode	40	14
Std. Dev.	11.41	11.91
Min.	3	0
25 %	29	10
50 %	35	14
75 %	42	21
Max	99	99

To explore temporal patterns in the distribution of calls, I plotted the trend in Figure 3 using monthly aggregates and a four-month moving average. The chart clearly shows a rising trend in call volume since mid-2017, which could be linked to awareness campaigns about the program, the helpline, and domestic violence in general. There are spikes in calls that recur around the end of each year—between October and January in 2016, 2017, 2018, and 2020—although they do not appear consistent enough in size to be considered a clear trend. Conversely, there is a notable rise between late 2018 and early 2019 that could be associated with external factors such as those mentioned earlier. After a decline and stabilization period in 2019, there is another sharp increase in 2020. One external factor that might explain this pattern is the implementation of ASPO (which stands for Preventive and Mandatory Social Isolation in Spanish) policies during the 2020 COVID-19 pandemic, which forced the population to remain in their

homes and immediate surroundings. Given the greater prevalence of sexual violence in domestic settings and by aggressors known to victims, the rise in calls during this period could be explained. However, it should be clarified that all possible associations I suggest as interpretation of this figure would need to be validated with an in-depth time-series analysis, which goes beyond the scope of this work.

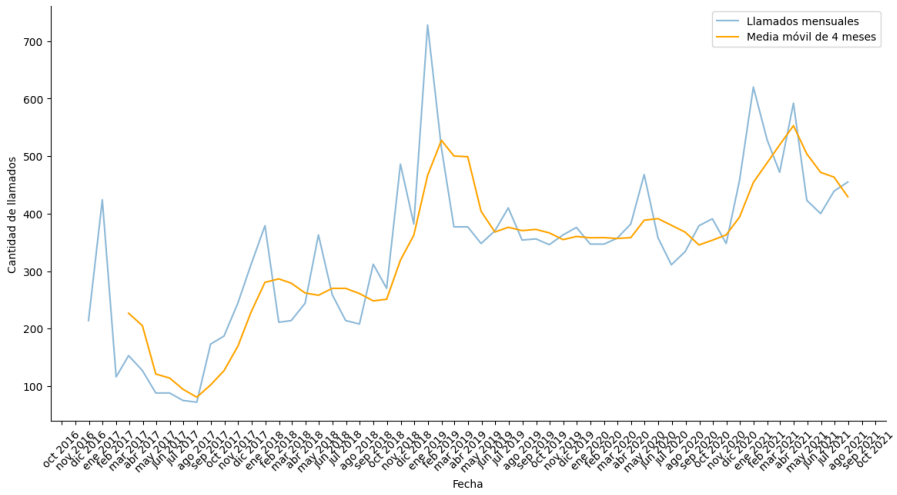

Figure 3: Number of calls over time with a four-month moving average.

I also constructed the variables `season`, `weekend`, and `time of day` to explore the possibility of additional patterns in the calls. I found that a higher proportion of calls occur during weekdays (80 %) and in the afternoon (38 %). No significant disparities were observed in the distribution of calls across seasons.

According to the distribution of the variable `llamado_provincia`, most calls originate from CABA (37 %) and Buenos Aires Province (36 %). For 9 % of cases, there are no data (`N/A`), and the remaining 18 % are distributed across the other provinces, with Córdoba and Santa Fe contributing the most calls at 3 % each.[3]

The variable `caso_judicializado` shows that 46.7 % of calls are not associated with an ongoing judicial case, 39.7 % are, and for the remaining 13.4 % there are no data.[4]

The variable `hecho_lugar` informs where the aggression took place. As illustrated in the bar chart in Figure 4, approximately 30 % of calls have no data (`N/A`); 25 % of incidents that occur at the victim's residence; and 13 % at the residence of the aggressor. The fourth most reported category, at 12 %, is *social media*. The remaining 20 % is distributed among categories such as public spaces (parks, open fields, etc.), transportation, and educational settings, among others. The high proportion of cases occurring at the victim's residence supports what was stated in the Introduction, namely that most cases of sexual violence occur in familiar environments rather than involving strangers or unknown places.

[3]See Figure 17 in the Appendix.
[4]See Figure 18 in the Appendix.

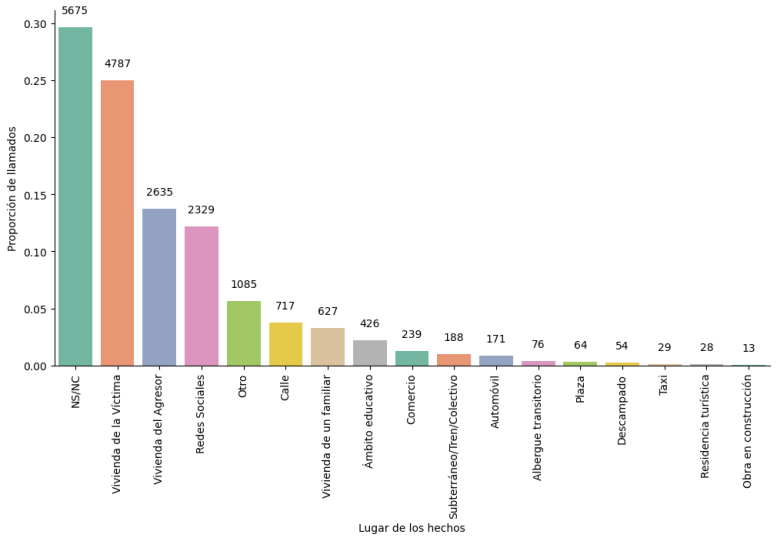

Figure 4: Place of occurrence.

Victims' nationality, reported in `victima_nacionalidad`, is distributed as follows: 80 % Argentine, 15 % no data, and the remaining 5 % divided among Bolivian, Paraguayan, Peruvian, Brazilian, Uruguayan, and Chilean, as well as the category "other."[5]

For `victima_discapidad`, 53.7 % of victims have no data, 43.2 % have no disability, and 2.9 % do.[6]

Regarding the gender of the victims, the bar chart in Figure 5 reinforces what was established in the Introduction regarding gender distribution among victims: 77.6 % are women, 18.4 % men, 3.7 % no data, and 0.14 % transgender.

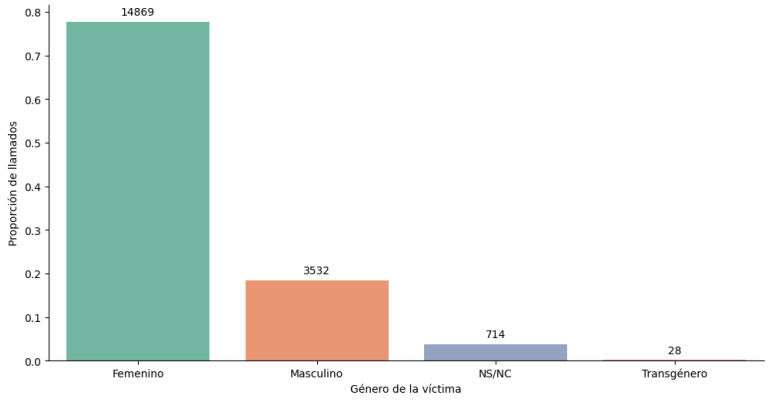

Figure 5: Gender of victims.

The links between victims and aggressors again reflect the persistence of sexual violence perpetrated by people within victims' social circles. In the bar chart of Figure 6 for the variable `victima_vinculo_agresor`, the distribution across different relational categories is shown. The trend becomes even clearer when regrouping the categories into `Known family member`, `Known non-family` (a category already present in the original variable), `Stranger`, and

[5] See Figure 19 in the Appendix.
[6] See Figure 20 in the Appendix.

N/A. While 15. 4 % of the aggressors are reported as strangers, 47.4 % are family members and 19. 7 % nonfamily acquaintances, meaning 67. 1 % of the aggressors are known to the victim. This number could be even higher if we consider that some aggressors could also be hidden within the 17.2 % of N/A.

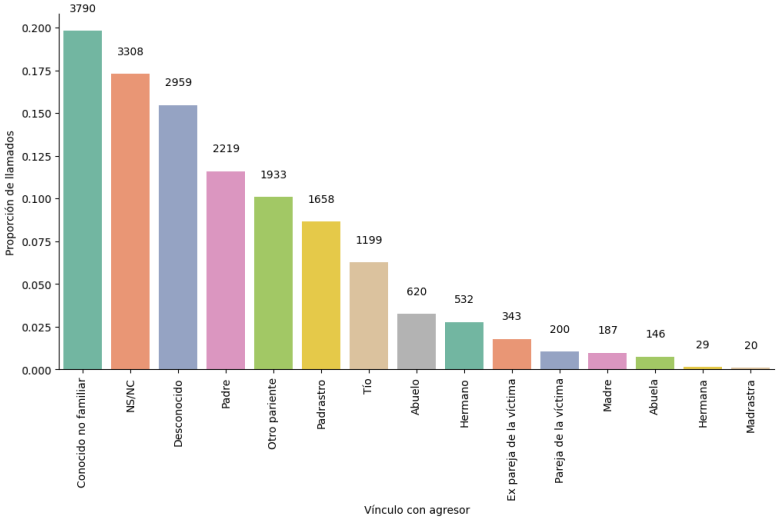

Figure 6: Victim-aggressor relationship.

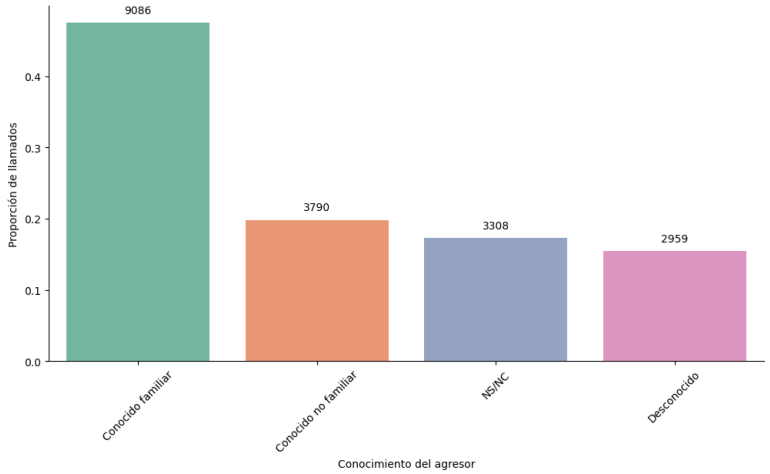

Figure 7: Aggressor known or unknown to the victim.

For the variable vinculo_llamante_victima which encodes the relationship between the caller and the victim, 24.9 % of calls come from police stations, 17.2 % from a family member of the victim (excluding the categories Mother, Father, Grandparent, or Sibling), 16 % from the victim's mother, and 14.2 % from the victims themselves. The rest of the categories include other acquaintances, fathers, neighbours, grandparents, siblings, other institutions, or N/A, all with less than 10 %. Finally, calls from schools, public defenders' offices, and the aggressors themselves total less than 1 %.[7]

[7]See Figure 21 in the Appendix.

Regarding the variable of interest `victima_convive_agresor`, the univariate analysis shown in the bar chart in Figure 8 reveals that those who cohabit with their aggressor are a minority at 14.3 %, while 64.4 % do not. The remaining 21.19 % are `N/A`, this the category I later attempt to predict as either `YES` or `NO`.

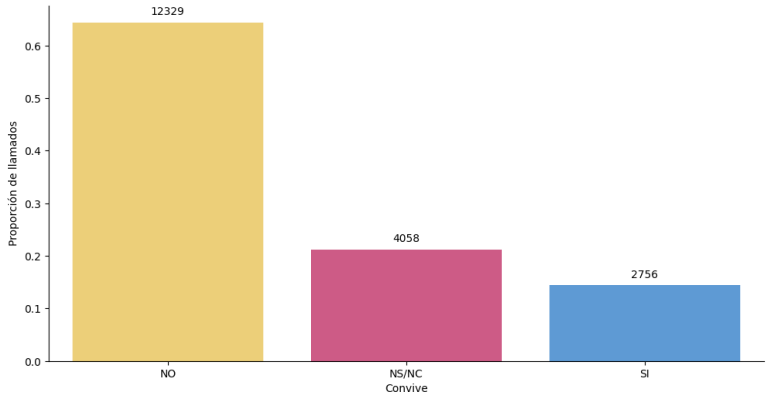

Figure 8: Victim-aggressor cohabitation.

2.2.1. Multivariate analysis of `victima_convive_agresor`

For the multivariate analysis, I selected some variables likely to be more related to `victima_convive_agresor`: `hecho_lugar`, `momento_dia`, `victima_vinculo_agresor`, `llamante_vinculo`, and `victima_edad`.

I created bar charts to explore the relationship between `victima_convive_agresor` and the categorical variables. I observed that the original distribution of `victima_convive_agresor`-a majority of `NO` responses, a minority of `YES`, and `N/A` positioned in between-holds for almost all categories of these variables, with the following exceptions.

First, as shown in Figure 9, when incidents occur at the victim's residence, there are more cases in which the victim cohabits with the aggressor and the distribution shifts, in decreasing order, to `NO`, `YES`, `N/A`. The same happens when incidents occur at the residence of the aggressor, although in this case the proportion of `YES` only slightly exceeds the proportion of `N/A`.

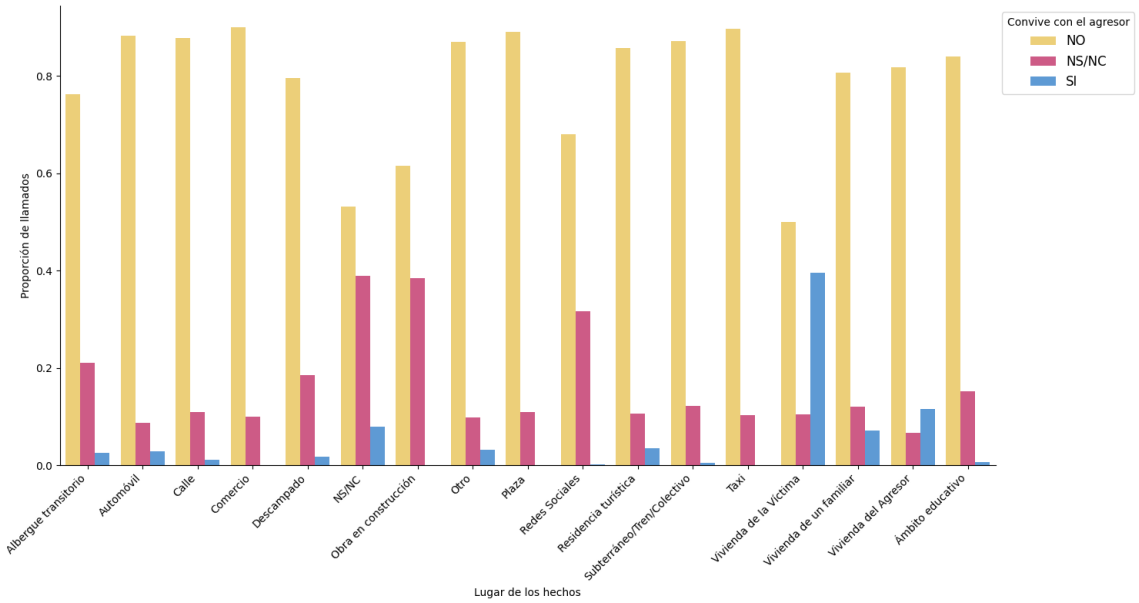

Figure 9: Cohabitation with the aggressor by place of occurrence.

Second, in Figure 10, we can see that for most cases in which the aggressor is a family member of the victim (`Grandfather`, `Sister`, `Brother`, `Stepmother`, `Mother`, `Stepfather`, `Victim's partner`), the cases where the victim does cohabit outnumber those in which there is no response about cohabitation. However, only in the categories `Mother`, `Stepfather`, and `Victim's partner` do co-habiting cases actually outnumber non-co-habiting ones.

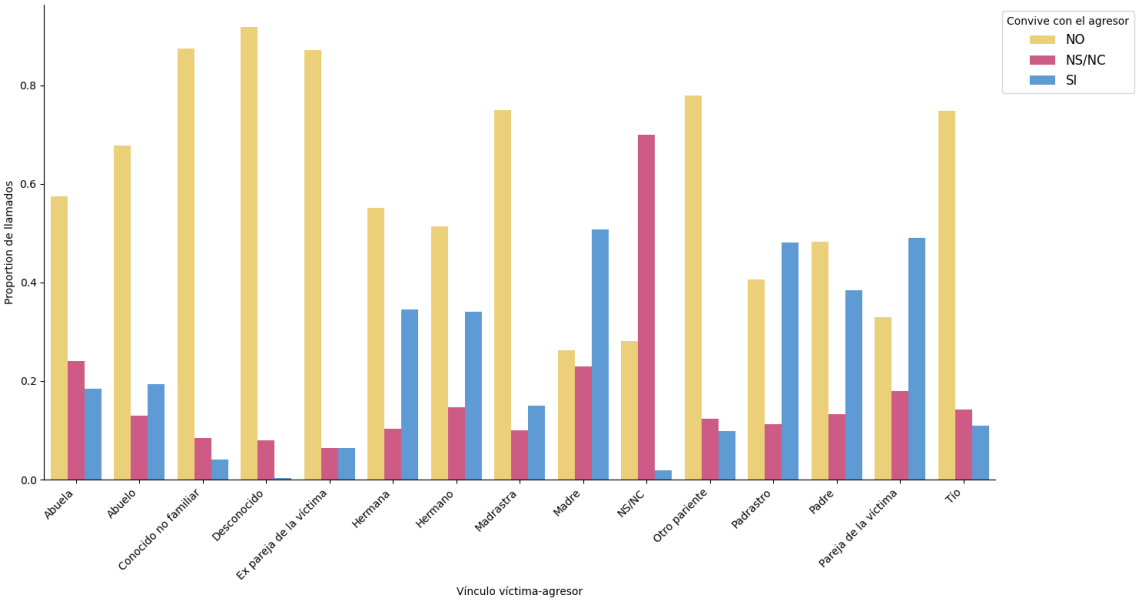

Figure 10: Cohabitation with the aggressor by victim–aggressor relationship.

Finally, in Figure 11, it can be seen that for the category `Neighbour` of `llamante_vinculo`, the trend in positive and negative responses is also reversed. It is also noteworthy that `N/A` values for `victima_convive_agresor` are

particularly high when the call originates from `Other institution` not including schools, police stations, or hospitals.

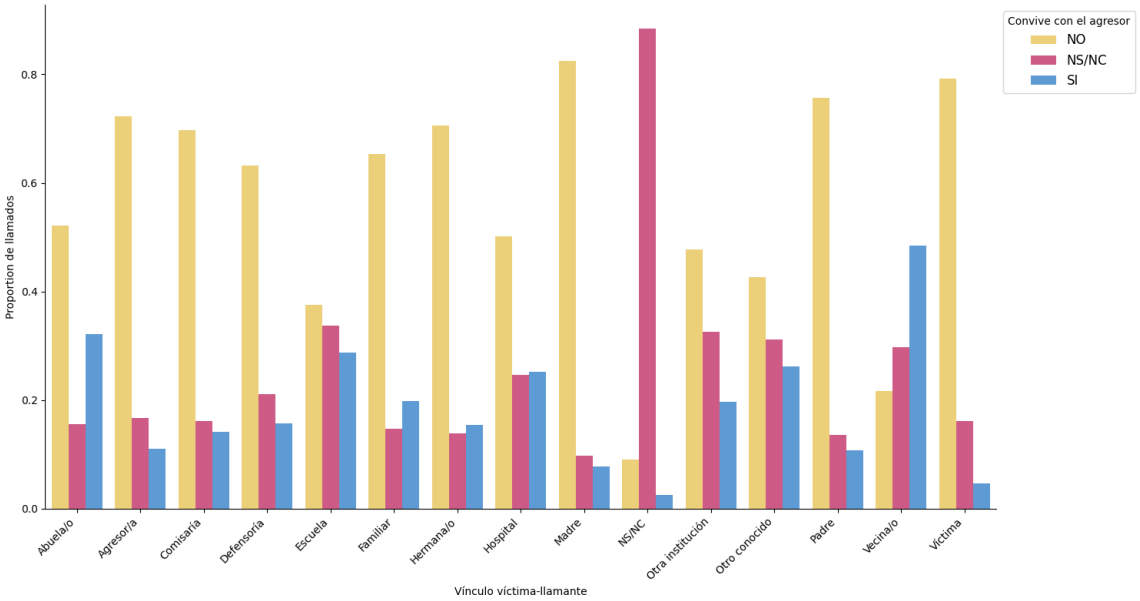

Figure 11: Cohabitation with the aggressor by caller–victim relationship.

I then produced comparative boxplots and quartile analyses of `victima_edad` for each category of `victima_convive_agresor`. As shown in Figure 12 and Table 4, victims who cohabit with their aggressor are slightly younger than those who do not; victims with `N/A` in the cohabitation variable appear closer in age to those who do cohabit. However, these differences in age do not appear significant.

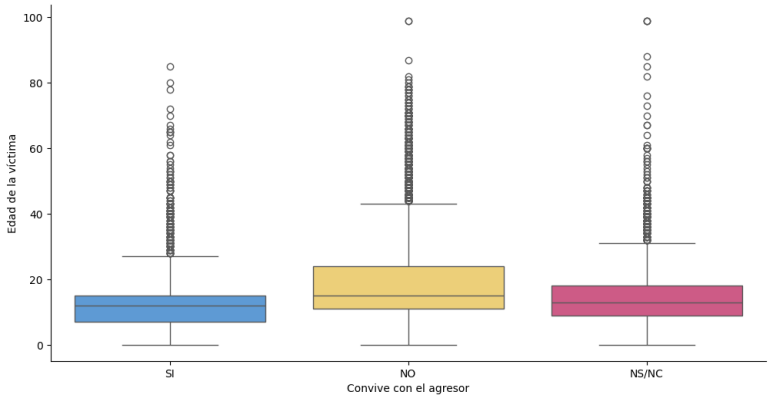

Figure 12: Distribution of victim's age by cohabitation status with the aggressor.

Table 4: Victim's age quartiles by category of `victima_convive_agresor`.

	Cohabits	**Does not cohabit**	**N/A**
Q1	7	11	9
Median	12	15	13
Q3	15	24	18
IQR	8	13	9

I also evaluated a possible relationship between victim's age, the aggressor's relationship to the victim, and whether or not they cohabit. In Figure 13, the same trend already shown in Figure 12 appears: average ages of victims across cohabitation categories are fairly similar. However, I highlight that for the categories `Partner` and `Ex-partner`, the average age of victims is slightly higher compared to other relationship categories; and specifically, the mean age of victims who do cohabit with their aggressors is higher than that of those who do not cohabit or whose cohabitation status is `N/A`. Mean age also spikes for the category `Stepmother` when cohabitation status is missing. Finally, I note that the lowest mean ages occur when aggressors are `Grandfather`, `Grandmother`, `Stepmother`, or `Father`, where no mean age surpasses 10 years for victims who cohabit with their aggressors.

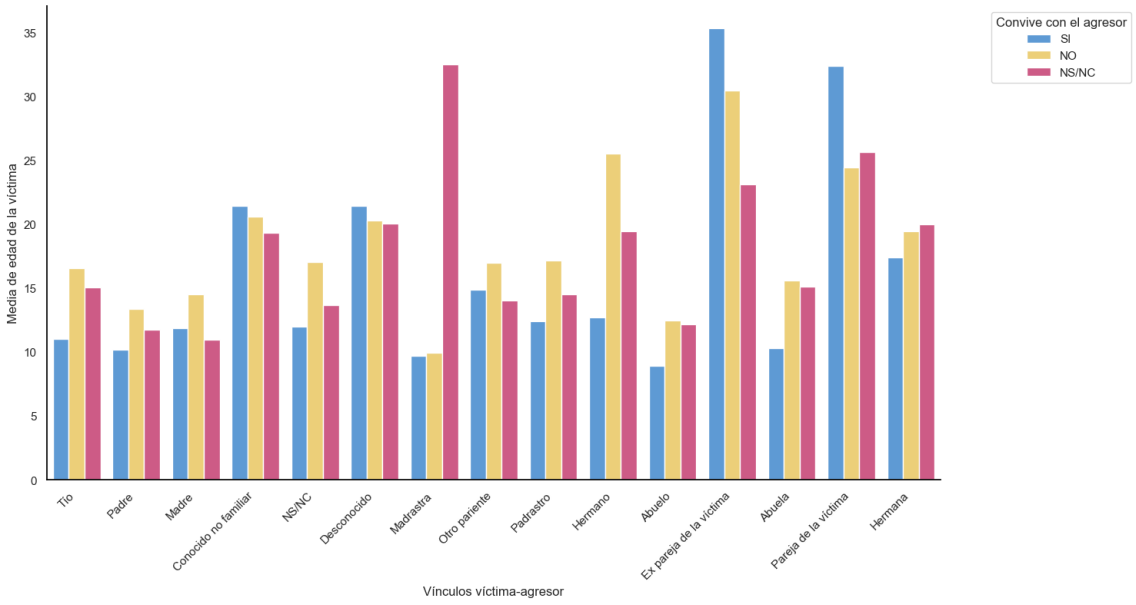

Figure 13: Victim's age by relationship and cohabitation with the aggressor.

Finally, I related `victima_convive_agresor` to missing values in `victima_edad`, which account for 9.82 %. I filtered the dataset to include only rows with missing `victima_edad`, and generated the same bar chart as in Figure 8 with those cases. The result, shown in Figure 14, indicates an increase in `N/A` cases for `victima_convive_agresor`. In the full dataset, `N/A` represents 21.19 % of `victima_convive_agresor`, but when looking only at cases with missing victim age, that proportion rises to 57.49 %. In other words, when victim's age is missing, cohabitation status with the aggressor is also more likely to be missing.

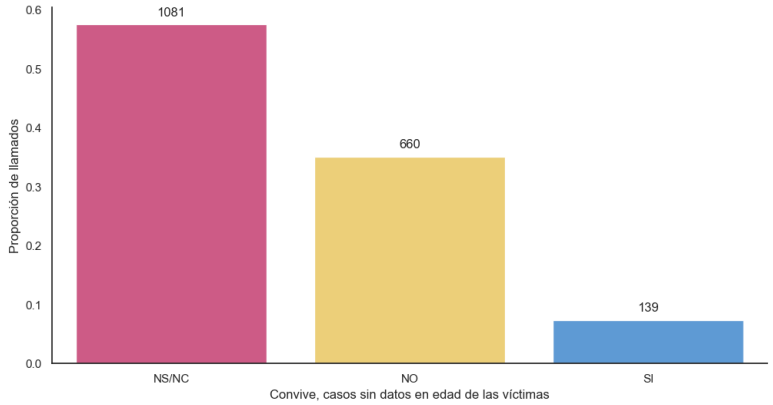

Figure 14: Victim–aggressor cohabitation in rows with missing `victima_edad`.

3. Methodology

The high dimensionality of the dataset and the multiplicity of levels in many variables present challenges for analyzing multivariate relationships with the target variable; in addition, they may cause processing issues when training the predictive model with SVM. I therefore applied two methods to reduce the dimensionality of the data:

1. **Non-Metric Multidimensional Scaling (NMDS)**: This method allows me to generate 2D visualizations to look for clustering of the classes to be predicted; it also functions as a preprocessing step for training SVMs.
2. **Manual data reduction**: I deleted, grouped, and transformed variables based on domain knowledge and the Exploratory analysis.

I complemented these reduction strategies with two different preprocessing approaches to handle missing data in the age variables, since the implementations I use for Gower distance and SVM do not accept empty cells. I generated two alternative datasets:

- Dataset A: categorized the age variables[8], and coded missing data as `N/A`.
- Dataset B: discarded the variable `llamante_edad`, and also dropped rows with missing data in the variable `victima_edad`.

The choice of Dataset B is based, on the one hand, on the assumption that the victim's age is more relevant for the target variable; and on the other hand, on the fact that caller's age is missing in 44.82 % of cases, so filtering would excessively shrink the dataset.

Finally, I evaluated the effects of these reduction methods and preprocessing strategies on the performance of the SVM model.

[8] 1–11 years = childhood; 12–18 years = adolescence; 19–30 years = youth; 31–65 years = adulthood; 66+ years = old age.

3.1. Reordering and reduction with NMDS

NMDS, a particular case of Multidimensional Scaling (MDS), is an ordination method commonly used to display similarities and differences between data by rearranging them in lower-dimensional spaces while preserving the relative order of the original distances (Chan et al., 2019, p. 218). As its name indicates, it accepts non-metric distance matrices as input, which is an advantage for the multivariate dataset I am working with.

The use of this method as preprocessing for SVM training was motivated by the experiments of Cai et al.(2019) in the field of microbiology, who obtained good *accuracy* results applying this methodology (p.69).

I implemented the `mds` method from the *scikit-learn* library (Pedregosa et al., 2011), adjusting the following parameters:

- `n_dimensions`: tested values between 2 and 7.
- `metric`: set to `False` to apply non-metric scaling.
- `dissimilarity`: set to `precomputed` to use the Gower distance matrix.
- `normalized_stress`: set to `True` to compute stress for NMDS.

Given the mixed nature of the dataset's variables, I chose Gower distance as the input for NMDS. Similarity is calculated pairwise for numerical and ordinal variables[9] with:

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

Where:

- s_{ijk} is the similarity (or distance) between two individuals or rows i, j in variable k.
- R_k is the range of k.

For categorical variables, similarity between two points i and j is computed binarily as 0 when values are identical (minimum distance), or 1 when they're not (maximum distance).

The final similarity matrix is then calculated as:

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

Where δ_{ijk} is the total number of variables or the number of variables for which comparison is possible (Gower, 1971, p. 859-860).

From the original distance matrix S, NMDS computes coordinates in a reduced n-dimensional space (typically $n = 2$ or $n = 3$ for easier visualization), and transforms them into a matrix of disparities $\hat{d}_{ij}$. In $\hat{d}_{ij}$, the magnitudes of the original distances are not preserved, but the rank ordering of those magnitudes is. Stress is then calculated as:

[9] Although Podani (1999) proposes a special treatment for ordinal variables that is widely used today, the method I use here treats ordinal variables as numeric, as in Gower's original work (see `https://sourceforge.net/projects/gower-distance-4python/files/`).

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

The algorithm iterates, recalculating d_{ij} and $\hat{d}_{ij}$, to minimize stress—that is, to minimize the difference between distances and disparities (Kruskal, 1964, p. 117-123).

I applied NMDS to datasets A and B mentioned earlier, varying the values of `n_components`. I generated visualizations from the resulting `n_components` = 2 projections to evaluate clustering based on `victima_convive_agresor`; and I trained SVM models for each dimensionality reduction.

3.2. Manual data reduction

I eliminated 12 of the variables describing violence experienced, considering them uninformative since they had positive response rates lower than 1 % (see again Figures Types of sexual violence reported and Types of non-sexual violence reported in section Data).

I also grouped some variables into three broader categories that share semantic and legal domains, as shown in Table 5:

Table 5: Grouping of violence variables by domain.

New grouped variable	**Grouped variables**
`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`

When choosing an encoding method to transform the data for SVM, I based my decision on the findings of Udilă (2023) in *Encoding methods for categorical data*. Comparing the performance of SVM models trained with different encoding methods, the author finds that one-hot encoding consistently yields models with higher *accuracy* (p.7).

However, in the same paper, the author warns about the potentially costly processing (in time and memory) that this method incurs for variables with high cardinality (p.7). Therefore, I reduced the cardinality of variables with more than five levels as follows:

- `llamado_provincia`: reduced from 25 levels to 6: `Buenos Aires`, `C.A.B.A.`, `Región Norte`, `Región Central`, `Región Patagónica`, and `N/A`.
- `victima_nacionalidad`: reduced from 9 levels to 3: `Argentina`, `N/A`, and `Other`.

- `hecho_lugar`: reduced from 17 levels to 6: `N/A`, `Victim's residence`, `Aggressor's residence`, `Social media`, `Public space/transport`, and `Other`.
- `llamante_vinculo`: reduced from 16 levels to 5: `Institution`, `Acquaintance of victim`, `Victim`, `Aggressor`, and `N/A`.
- `agresor_vinculo`: reduced from 16 levels to 4, using the grouping shown in Figure 7, Aggressor known or unknown to the victim: `Known family member`, `Known non-family`, `N/A`, and `Stranger`.

The dataset resulting from manual reduction contains 36 variables (compared to 54 originally). Table 10 in the Appendix shows a summary of deleted, grouped, and transformed variables.

I also split this dataset into types A and B, following the approach of the previous section, Reordering and reduction with NMDS.

3.3. SVM models

Support Vector Machines are supervised learning models widely used for classification tasks. They aim to find the hyperplane that best separates the classes in the original space if the data are linearly separable, or in a higher-dimensional transformed space when they are not, maximizing the margin between the closest data points of each class.

The ability of SVMs to handle data that are not linearly separable—through different kernel tricks—is the reason I chose them, given the complex relationships between the target variable and the rest of the variables (Chan et al., 2019, p. 323-331).

In implementing and training the SVM models, I used the `svc` method from *scikit-learn* (Pedregosa et al., 2011), optimizing the hyperparameters:

- `kernel`: defines the type of function to transform the input space.
- `C`: regularizes the balance between maximizing the margin between classes and minimizing classification errors.
- `gamma`: a weight controlling the influence of individual datapoints on the decision boundary of the model.

Table 6 summarizes the different datasets for the SVM training experiments.

Table 6: Preprocessing for training SVM models.

Preprocessing	Dataset	Specifications
1. NMDS Reduction	A	Age variables transformed to categories. Missing data coded as `N/A`
	B	Only complete data on victim age data (numeric). Caller's age removed.
2. Manual reduction using mixed methods	A	Age variables transformed to categories. Missing data coded as `N/A`. One-hot encoder applied to categorical variables.
	B	Only complete data on victim age data (numeric). Caller's age removed.

3.3.1. SVM training with NMDS reduction

I prepared the data by first replacing the literal textual values `N/A` in `victima_convive_agresor` with the `NA` code for missing data. Then, I separated the target variable `victima_convive_agresor` (y) from the rest of the data (X); and removed missing rows from y, saving their indices. I computed the Gower distance matrix from X for use in NMDS.

I generated different NMDS reductions by varying the `n_components` parameter. For each reduction, I created the final blind test set of unseen cases from the transformed rows of X corresponding to missing values in `victima_convive_agresor`. I stratified the main dataset (X and y) into training (80 %) and testing (20 %) using `StratifiedShuffleSplit` from *scikit-learn* (Pedregosa et al., 2011) to mitigate the imbalance between `YES` and `NO` classes.

For training, I performed a hyperparameter search (`kernel`, `C`, `gamma`) with 5-fold cross-validation, evaluating performance with the `f1_weighted` metric. I chose this metric both because of class imbalance (*accuracy* or *precision* can be skewed toward the dominant class) and to ensure a model balancing precision and recall, since I am interested not only in correctly predicting true positives but also in not excluding potential positive cases. I also calculated the stress associated with each NMDS reduction as an additional evaluation metric.

Finally, I tested the trained model with the best hyperparameters on the blind dataset. Although it is not possible to perform a traditional evaluation on this final set since these are truly missing values, I compared the proportions of `YES` and `NO` between the original dataset and the predictions to obtain an estimate of model performance.

3.3.2. SVM training with manual data reduction

For this experiment, I began by encoding the variables. For `llamado_fecha_hora`, `momento_dia`, `estacion_del_año`, and the age variables (in Dataset A) I used an ordinal encoder (`OrdinalEncoder` from *scikit-learn* (Pedregosa et al., 2011)) to preserve the ordinal nature of the data. In Dataset B, I scaled the victim's age.

I transformed binary variables (`YES`, `NO`), including the target variable, assigning `YES` = 1 and `NO` = 0. As in the previous experiment, I replaced `N/A` values in the target variable with missing data. For the other categorical variables, I applied `OneHotEncoder`.

I created the blind test set in the same way as for NMDS, using rows corresponding to `N/A` in the target variable. The model was then trained using the same methodology of cross-validation and hyperparameter search (`kernel`, `C`, `gamma`), optimizing for `f1_weighted`.

4. Results

4.1. Visualization of the ordering with NMDS

Figures 15 and 16 show that in the two-dimensional rearrangement of the data, the categories `YES`, `NO`, and `N/A` do not exhibit clear separation. `YES` and `NO` appear intermingled, and consequently, `N/A` does not cluster distinctly near either of the two classes, nor in isolation. In addition, the stress values obtained in both NMDS configurations are considerably high: 0.3, whereas an optimal value is usually around 0.05.

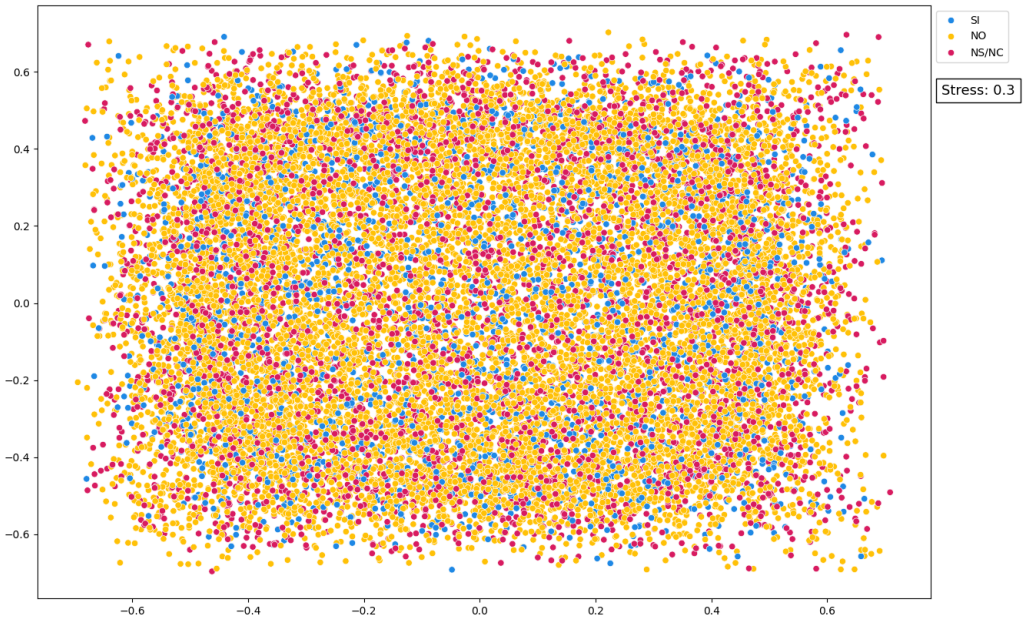

Figure 15: Visualization of data reduced with NMDS - dataset A

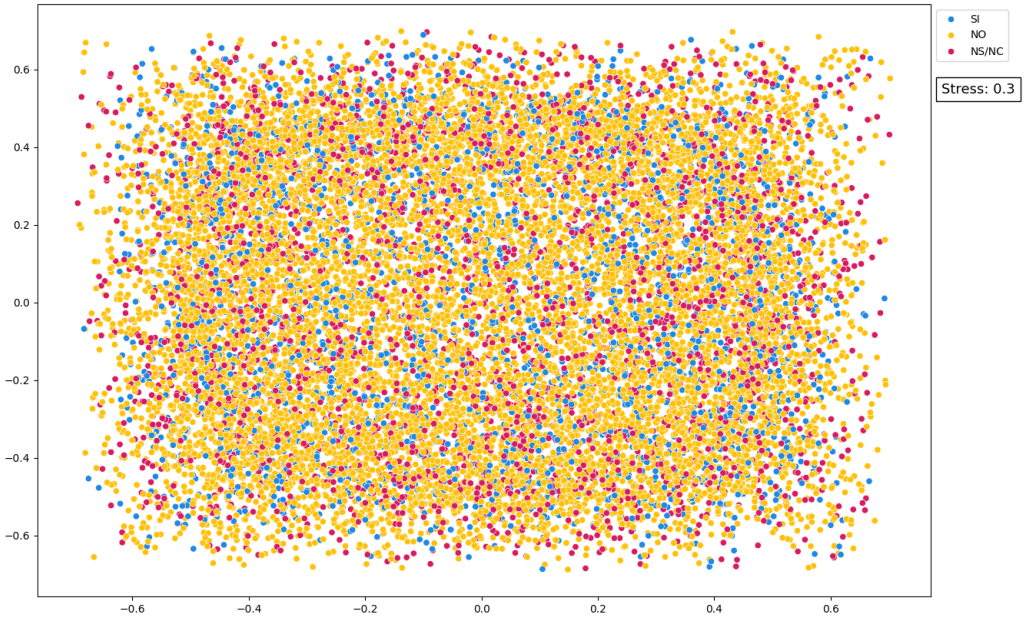

Figure 16: Visualization of data reduced with NMDS - dataset B

4.2. SVM predictive models

4.2.1. Preprocessing with NMDS

The stress values for the orderings, even in dimensions > 2, remained high, with the lowest being 0.16 for 7 dimensions.

For datasets A and B, and across all variations of `n_components`, hyperparameter optimization consistently found the best results with:

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

The selection of a polynomial `kernel` reflects the non-linearity of the data in the input space. The low value of the `C` penalty parameter indicates a preference for generalization over minimizing training errors. Configuring `gamma` as *auto* results in a value of $\frac{1}{\texttt{n\_components}}$: $0{,}5 \leq \gamma \leq \pi$, since $2 \leq$ `n_components` ≤ 7.

Model performance was also similar for both datasets and across all `n_components`, summarized in Table 7 below. The models are heavily affected by class imbalance and do not generalize well for the minority class. The F1, recall, and precision metrics are good for the majority class `NO` (0.90, 1.00, and 0.82 respectively), indicating that the models correctly identify most points in this class. For the minority class `YES`, while the models do not produce false positives

in classification (hence the precision of 1.0), they fail to correctly identify any cases belonging to this class, making recall and thus F1 equal to 0.

The macro F1 (0.45) and macro recall (0.50), which average performance across both classes, highlight the imbalance problem. The overall *accuracy*, by contrast, is misleading on its own: its high value (0.82) reflects overfitting to the majority class. Similarly, macro precision averages class precision, and as noted earlier, the minority class's high precision is only because the model produced no false positives.

Cuadro 7: Performance of SVM models using NMDS as preprocessing.

Class	F1	Recall	Precision	Macro F1	Macro Recall	Macro Precision	Accuracy
0 (NO)	0.90	1.00	0.82	0.45	0.50	0.91	0.82
1 (YES)	0.00	0.00	1.00				

When applying the best models to the final blind test sets for A and B, I found that they classified all cases as belonging to the majority class `NO`.

4.2.2. Preprocessing with manual reduction

Hyperparameter optimization yielded the best results with:

Dataset A

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

Dataset B

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

Both kernels are appropriate for non-linear data. The best regularization parameter `C` is significantly higher than in the previous experiment, especially for dataset B.

As in the previous experiment, model performance was similar for datasets A and B. Results are shown in Table 8. In this second experiment, the model classified the minority class better than in the previous experiment, with F1, recall, and precision for `YES` at 0.62, 0.54, and 0.72 respectively, while maintaining strong performance for the majority class (0.93, 0.95, and 0.90 respectively). While the improvement is notable, the achieved figures still do not indicate a reliable model for classifying `YES` cases; in fact, recall appears close to random classification. Precision of 0.72 for `YES` is the highest of the three metrics for this class, reflecting the same phenomenon as in the previous experiment.

Macro F1 (0.77) and macro recall (0.75) also improved compared to the previous experiment. Macro precision (0.81) and *accuracy* (0.88) reflect the same issues as before, although the inter-class performance gap is less extreme this time.

Cuadro 8: Performance of SVM models with manual reduction preprocessing.

Class	F1	Recall	Precision	Macro F1	Macro Recall	Macro Precision	Accuracy
0 (NO)	0.93	0.95	0.90	0.77	0.75	0.81	0.88
1 (YES)	0.62	0.54	0.72				

Comparing the proportions of `YES` and `NO` predicted by the best models in the blind test sets with the proportions in the original dataset[10], I found the results summarized in the table below:

Cuadro 9: Predicted vs. original proportions for each class.

Dataset	`YES` original	`YES` predicted	`NO` original	`NO` predicted
A	0.18	0.09	0.82	0.91
B	0.18	0.10	0.82	0.90

5. Conclusions

Given the shortcomings of the NMDS representations (visualizations that provided no useful insights, the seemingly random distribution of the data, and poor stress values), future work could explore alternative dimensionality reduction and visualization techniques, such as PCA, while ensuring appropriate data transformations. It is also possible that in this dataset, the patterns related to the target variable are too complex to be represented in two or three dimensions.

Although manual preprocessing of the dataset produced markedly superior results in training the classifiers/imputers for missing data, both experiments were hindered by class imbalance, and I cannot conclude that any of the models provide a reliable basis for imputing missing values.

NMDS preprocessing yielded the weakest models. Different values of `n_components` and the two strategies for handling missing age values did not affect model performance. The low regularization parameter `C` (0.1) may have contributed to overfitting the majority class in the first NMDS experiment. In the second experiment, where `C` values were higher (1 for dataset A and 100 for dataset B), the model performed better. However, this does not mean that `C` alone explains the performance difference, since higher `C` values were also tested during hyperparameter optimization in the first experiment without improvement.

The parameter `gamma`, set to `scale` in the second experiment and `auto` in the first, does not appear to have had a significant impact on its own.

During the exploratory analysis, I observed that the mean age of victims who cohabit with their aggressors was higher than that of those who do not cohabit **and** of those whose cohabitation status is N/A. This does make me wonder whether the missing data could belong mostly to victim's who do cohabit with their aggressors. Given the weakness of the resulting classifier models, this intuition can only remain as such, but it does pose an interesting path forward for the analysis of these data.

[10] I used the data after manual preprocessing, excluding the `N/A` rows in each case A and B.

For future work in that direction, I consider it important to implement strategies to mitigate class imbalance, such as artificially generating minority-class cases, or applying over- or under-sampling techniques.

More broadly, I also find it relevant to consider integrating data from different sources on gender and sexual violence available in Argentina, such as the 144 hotline for reporting gender-based violence, records from the OVD (Domestic Violence Office), and data from the Gender Data Observatory.

Finally, regarding the gender of victims and perpetrators—and returning to a point raised in the introduction—it is essential that data collection be carried out with gender awareness and perspective. As Rita Segato points out in La guerra contra las mujeres (2016), sexual violence is directed toward female or feminized bodies. This includes bodies perceived as young, weak, racialized, or belonging to sexual minorities, which aligns with data showing higher incidence of sexual violence against boys during childhood and adolescence, when they are more vulnerable and therefore feminized (Contreras et al., 2016; Unidad Fiscal Especializada en Violencia contra las Mujeres (UFEM), 2023; Ferris, 2002).

Based on these reflections, I propose some guiding questions for future studies:

How can the existing knowledge about the statistics of victim's living with their aggressors or being familiar with their aggressors inform the making of more accurate algorithms that help officials assess the risk of a situation when receiving a phone call?

If reports of sexual violence against gender minorities are scarce in the data, does this mean those groups experience less sexual violence? Or does it reflect a general underrepresentation due to their position as social minorities and their limited access to justice?

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. Appendix

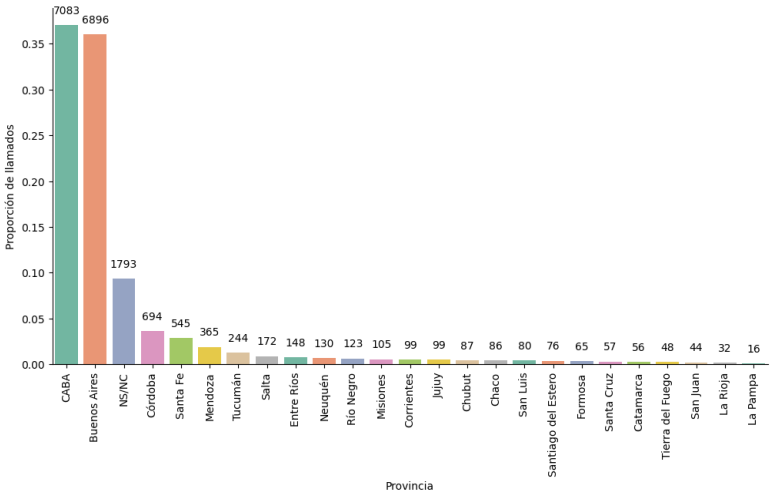

Figure 17: Calls by province.

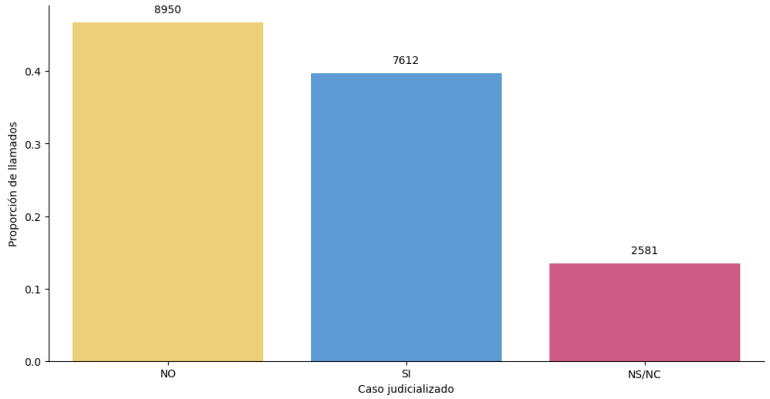

Figure 18: Cases that reached the judicial system.

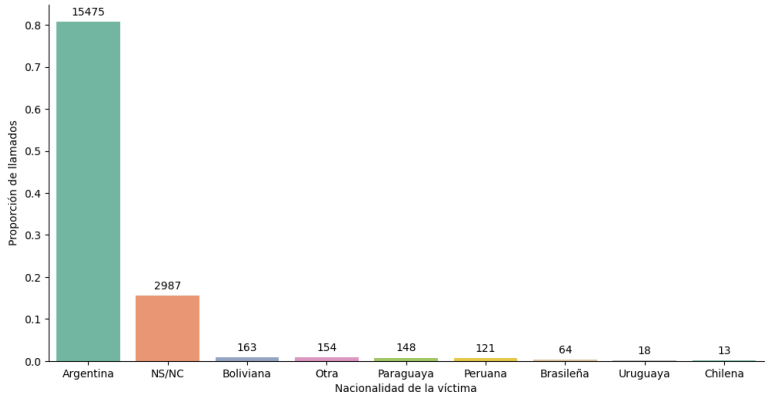

Figure 19: Victims' nationality.

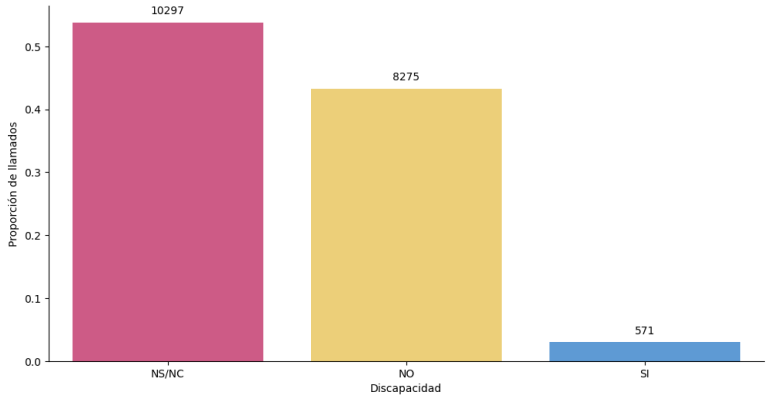

Figure 20: Presence of disability among victims.

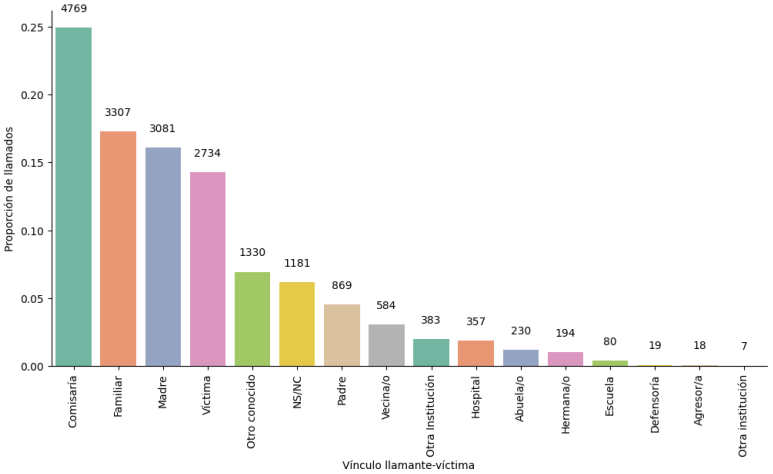

Figure 21: Victim–caller relationships.

Cuadro 10: Summary of variable transformations.

Original dataset	Reduced dataset	Transformation
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	Eliminated	Non informative (<1 %)
vs_explotación_sexual, vs_explotación_sexual_comercial, vs_explotación_sexual_viajes_turismo, vs_sospecha_trata_personas_fines_sexuales	vs_explotación_sexual	Grouped by domain
vs_violacion_via_vaginal, vs_violacion_via_anal, vs_violacion_via_oral	vs_violacion	Grouped by domain
vs_tentativa_violacion, vs_intento_violacion_tercera_persona	vs_tentativa_violacion	Grouped by domain
llamado_provincia	Buenos Aires, C.A.B.A., Región Norte, Región Central, Región Patagónica, N/A	Reduced levels
victima_nacionalidad	Argentina, Otra, N/A	Reduced levels
hecho_lugar	N/A, Victim's residence, Agressor's residence, Social Media, Public space/transport, Other	Reduced levels
llamante_vinculo	Institution, victim's acquaintance, Victim, Agressor, N/A	Reduced levels
agresor_vinculo	Known family member, Known non-family, Stranger, N/A	Reduced levels