FACULTAD DE ESTUDIOS ESTADÍSTICOS

GRADO EN ESTADISTICA APLICADA

Curso 2024/2025

Trabajo de Fin de Grado

TÍTULO:
La generación expuesta: un análisis estadístico de la violencia sexual digital en jóvenes

***Alumna*: Olaya del Río García**

***Tutor*: Fernando Pérez Contreras**

Junio de 2025

UNIVERSIDAD COMPLUTENSE
MADRID

Contenido

Contenido tablas

Contenido figuras

1. Resumen

En la actualidad en la que vivimos, los jóvenes están cada vez más expuestos al mundo digital, y con ello a nuevas formas de violencia, entre ellas, la violencia sexual digital (VSD). Esta forma de violencia ha crecido de manera exponencial en los últimos años, abarcando desde el acoso en redes sociales, hasta la difusión no consentida de contenido íntimo. A pesar de su frecuencia y gravedad, la VSD no recibe aún la atención que merece y, por tanto, resulta muy complicado diseñar estrategias eficaces de prevención e intervención.

Este trabajo tiene como objetivo principal realizar un estudio exploratorio para identificar características asociadas a una mayor probabilidad de haber experimentado o ejercido VSD. Para ello se aplican técnicas de regresión logística binaria y métodos de selección automática de variables como Stepwise y LASSO, así como un árbol de clasificación como herramienta complementaria, con el fin de examinar posibles factores de riesgo. Además, se analiza el impacto en los jóvenes que la sufren a través de técnicas de análisis no supervisado.

2. Abstract

In today's world, young people are increasingly exposed to the digital world, and with it, to new forms of violence, including digital sexual violence (DSV). This form of violence has grown exponentially in recent years, ranging from harassment on social networks to the non-consensual dissemination of intimate content. Despite its frequency and seriousness, SGBV still does not receive the attention it deserves and, therefore, it is very complicated to design effective prevention and intervention strategies.

The main objective of this study is to conduct an exploratory study to identify characteristics associated with a higher probability of having experienced or exercised digital sexual violence. To this end, binary logistic regression techniques and automatic variable selection methods such as Stepwise and LASSO are applied, as well as a classification tree as a complementary tool, in order to examine possible risk factors. In addition, the impact on young people who suffer it is analyzed through unsupervised analysis techniques.

3. Introducción y objetivos

En los últimos años la forma en la que los jóvenes interactúan entre sí ha cambiado drásticamente, especialmente con el avance tecnológico, pasando de interacciones mayormente presenciales, a una diversidad de formas de comunicación, como las llamadas telefónicas, la mensajería instantánea o las redes sociales.

Si bien estos avances han facilitado el acceso a la información y ampliado las formas de comunicación, también han dado pie a la aparición de nuevas formas de violencia en el ámbito digital. Entre ellas se encuentra la violencia sexual digital (VSD), que afecta especialmente a los jóvenes.

Esta forma de violencia se manifiesta a través de interacciones y comportamientos en entornos digitales que reproducen dinámicas de agresión sexual o buscan dañar a una persona en este ámbito. Se presenta mediante actuaciones que van desde el uso de expresiones de abuso o acoso, hasta la difusión no consentida de contenido íntimo. Estas acciones vulneran derechos fundamentales de las víctimas, exponiéndolas a situaciones de humillación, invasión de la privacidad y riesgos para su seguridad. En el caso de niños, niñas y jóvenes, la VSD afecta también a su desarrollo, dejando consecuencias emocionales que, en algunos casos, resultan irreversibles.

A pesar de que cada vez es más frecuente, la VSD aún no recibe la atención que merece, lo que dificulta la implementación de estrategias eficaces para prevenirla.

Inicialmente, el objetivo del estudio era construir modelos capaces de predecir perfiles de riesgo de víctima o agresor de violencia sexual digital. Sin embargo, dada la complejidad del fenómeno y la baja capacidad predictiva que presentaban los modelos tras una primera evaluación, se modificó la perspectiva inicial con el fin de generar resultados más útiles en la lucha contra este problema, ajustados a la complejidad de los datos. Por tanto, se decidió adoptar un enfoque exploratorio, empleando modelos estadísticos habitualmente utilizados con fines predictivos, pero aquí orientados a la identificación e interpretación de factores asociados con la experiencia de haber sido víctima o agresor de VSD, con el objetivo de extraer información útil para la prevención e intervención.

Además, como objetivo secundario, se busca examinar el impacto que tiene este tipo de violencia en los jóvenes que la sufren. Para ello, se emplean métodos de análisis no supervisado con el fin de identificar posibles patrones de conducta y evaluar su efecto en distintos grupos, contribuyendo así al desarrollo de estrategias de apoyo eficaces que ayuden a mitigar sus consecuencias.

4. Metodología

4.1. Algoritmo MissForest

El algoritmo MissForest es una técnica de imputación no paramétrica basada en bosques aleatorios (random forest), diseñada para manejar conjuntos de datos con variables de tipo mixto (continuas y categóricas) y relaciones complejas entre variables.

El proceso de imputación con MissForest se realiza de manera iterativa y consta de los siguientes pasos

1. Se imputan los valores faltantes mediante métodos de imputación simples, como la mediana o la moda.
2. Las variables se ordenan en función de la cantidad de valores faltantes, comenzando por aquellas que tienen menos.
3. Para cada variable con datos faltantes, se entrena un modelo de random forest, utilizando sus observaciones completas como conjunto de entrenamiento.
4. Se obtiene una predicción de los valores faltantes de cada variable con los modelos RF y se actualizan en el conjunto de datos.
5. Los pasos anteriores se repiten hasta que la diferencia entre las imputaciones de dos iteraciones consecutivas sea insignificante o hasta alcanzar un número máximo predefinido de iteraciones.

La evaluación del rendimiento de la imputación se hace mediante dos métricas:

- Error cuadrático medio normalizado (NRMSE): Utilizado para variables continuas, se define como:

$$NRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i^{imp} - x_i^{true}\right)^2}}{\sigma}$$

  Donde x_i^{imp} es el valor imputado, x_i^{true} es el valor verdadero y σ es la desviación estándar de la variable.

- Proporción de clasificación errónea (PCF): Aplicable a variables categóricas, se calcula como:

$$PFC = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(x_i^{imp} \neq x_i^{true}\right)$$

  Donde $\mathbb{I}$ es la función indicadora que toma el valor 1 su la condición se cumple y 0 si no.

4.2. Regresión logística binaria

La regresión logística binaria es una técnica estadística empleada para modelar la relación entre una variable dependiente dicotómica (evento/no evento) y un conjunto de variables independientes, que pueden ser continuas o categóricas. Su objetivo es estimar la probabilidad de ocurrencia de un evento a partir de una combinación lineal de predictores, transformada mediante la función logística:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Donde p es la probabilidad de que ocurra el evento de interés, β_0 es el intercepto del modelo y $\beta_1 + \beta_2, \ldots, \beta_n$ son los coeficientes asociados a las variables independientes $x_1 + x_2, \ldots, x_n$.

Esta relación permite expresar directamente la probabilidad como:

$$p = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n}}$$

Interpretación de los coeficientes

La interpretación de los coeficientes suele hacerse a través del *odds ratio* (OR), calculado como $OR = e^{\beta_i}$. Un OR mayor que 1 indica un aumento en la probabilidad del evento asociado a un incremento en x_i; un OR menor que 1, lo contrario.

Evaluación del modelo

Para evaluar la calidad de los modelos, se ha dividido la muestra en un conjunto de entrenamiento (80%) y otro de prueba (20%). Las métricas empleadas han sido:

- Contraste de significación de Wald: evalúa si cada coeficiente es significativamente distinto de cero. Un valor p inferior a 0,05 indica que la variable tiene un efecto significativo sobre la variable dependiente.
- Análisis de tipo II: mide la pérdida de ajuste del modelo al eliminar cada variable independiente, generando una ordenación por importancia basada en la reducción de la verosimilitud. El contraste se basa en la distribución χ^2.
- Matriz de confusión: clasifica las observaciones como eventos (1) y no eventos (0) según un punto de corte y una vez obtenida esta clasificación se obtienen métricas como la tasa de acierto, la sensibilidad y la especificidad:

	Predicción = 0	Predicción = 1
Realidad = 0	VN verdadero negativo	FP falso positivo
Realidad = 1	FN falso negativo	VP verdadero positivo

Tasa de acierto: $\frac{VN+VP}{VN+FP+FN+VP}$ Sensibilidad: $\frac{VP}{FN+VP}$ Especificidad: $\frac{VN}{VN+FP}$

- Indice Kappa: ajusta la tasa de acierto eliminando el efecto del azar

$$k = \frac{acc - \sum_{i=1}^{k} p_{i\cdot}\, p_{\cdot i}}{1 - \sum_{i=1}^{k} p_{i\cdot}\, p_{\cdot i}}$$

  Donde acc es la tasa de acierto, $p_{i\cdot}$ la proporción real observada en la categoria i y $p_{\cdot i}$ la proporción predicha en la categoría i.

- Área bajo la curva ROC: la curva ROC representa la sensibilidad frente a 1 - especificidad para diferentes umbrales de clasificación. El AUC mide la capacidad global del modelo para discriminar entre eventos y no eventos. Un AUC cercano a 1 indica un modelo con excelente capacidad de clasificación, mientras que un valor próximo a 0'5 refleja un rendimiento similar al azar

4.2.1. Métodos automáticos de selección de variables

Se han empleado tres métodos automáticos para seleccionar el subconjunto óptimo de variables en los modelos:

- Método backward: parte de un modelo completo que incluye todos los predictores disponibles y elimina iterativamente las variables que aportan menor información al modelo. Este proceso continúa hasta que la eliminación de cualquier variable restante supone un empeoramiento significativo del ajuste. Una vez eliminada una variable, no se vuelve a considerar.
- Método forward: a diferencia del anterior, comienza con un modelo vacío y va incorporando, una a una, las variables que más mejoran el ajuste. Se detiene cuando ninguna de las variables restantes aporta una mejora significativa. Las variables que han sido incluidas en el modelo no se eliminan posteriormente.
- Método stepwise: combina las estrategias de los dos métodos anteriores. Parte de un modelo vacío y, en cada paso, evalúa tanto la inclusión como la eliminación de variables. Esto permite que, si la entrada de una nueva variable hace que otra deje de ser relevante, esta última pueda ser eliminada. El criterio de eliminación sigue el esquema del método backward, permitiendo una mayor flexibilidad en la construcción del modelo. En cada iteración, el algoritmo evalúa qué acción (añadir o eliminar una variable) produce la mayor mejora en el modelo, considerando tanto el ajuste como la complejidad.

Para evitar el sobreajuste, se utilizan criterios de penalización que equilibran el ajuste del modelo con su complejidad:

- AIC (Akaike information criterion): $-2\ \ln(L) + 2\tau$
- BIC (Bayesian information criterion): $-2\ \ln(L) + \tau\ \ln(n)$

Ambos criterios valoran la mejora en la verosimilitud (L), penalizando al mismo tiempo el número de parámetros incluidos en el modelo (τ). La principal diferencia entre ellos radica en el grado de penalización: el BIC penaliza más fuertemente la

complejidad (n), lo que tiende a producir modelos más simples con menos variables seleccionadas.

4.2.2. Método Lasso

LASSO (Least Absolute Shrinkage and Selection Operator) es una técnica de regularización que mejora la capacidad predictiva y evita el sobreajuste al introducir una penalización en la función de coste. Esta penalización reduce algunos coeficientes a cero, lo que equivale a eliminar las variables correspondientes:

$$\min_{\beta}\left\{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right\}$$

Donde y_i es el valor observado de la variable dependiente, $\hat{y}_i$ es el valor predicho por el modelo, β_j son los coeficientes de las variables predictoras y λ es el parámetro de regularización que controla la intensidad de la penalización.

4.2.3. Validación cruzada

La validación cruzada es una técnica utilizada para evaluar el rendimiento de los modelos y estimar su capacidad de generalización. Consiste en dividir el conjunto de datos en varias particiones (o folds), entrenar el modelo en todas menos una y evaluar en la restante. Este proceso se repite hasta que todas las particiones hayan sido utilizadas como conjunto de prueba.

En este trabajo, la validación cruzada se ha empleado tanto para la elección del valor óptimo de λ, como para comparar los modelos de regresión logística binaria (manual y automáticos). Las métricas utilizadas para evaluar los modelos han sido la tasa de acierto, el AUC y el índice Kappa.

4.3. Árboles de clasificación

Los árboles de clasificación son un método estadístico no paramétrico utilizado para predecir una variable dependiente categórica (como una variable binaria), a partir de un conjunto de variables explicativas. Su estructura consiste en una serie de divisiones binarias jerárquicas que segmentan los datos en grupos progresivamente más homogéneos respecto a la variable objetivo.

La construcción del árbol comienza con un nodo raíz que contiene todas las observaciones. En cada paso, el algoritmo selecciona la variable y el punto de corte que generan la mejor partición, según un criterio de impureza. En este trabajo se ha utilizado el índice de Gini, definido como:

$$IG(nodo_j)=1-\sum_{k=1}^{K}(p_{jk})^2$$

Donde p_{jk} es la proporción de observaciones de la categoría k en el nodo j.

Valores más bajos del índice de Gini indican nodos más homogéneos; un nodo "puro" tendrá un valor próximo a cero.

El árbol se expande recursivamente hasta cumplir un criterio de parada, como una profundidad máxima, un tamaño mínimo de nodo hoja o la ausencia de mejoras significativas. Cada observación se asigna a la categoría mayoritaria del nodo hoja al que pertenece.

Poda del árbol

Para evitar el sobreajuste, se aplica una técnica de poda, que consiste en eliminar ramas poco informativas del árbol completo (árbol maximal) con el objetivo de obtener un modelo más simple y generalizable. Esta operación se basa en la minimización de una función de coste que penaliza la complejidad del árbol:

$$R_\alpha(T) = error(T) + \alpha \cdot |T|$$

Donde α es el parámetro de penalización y $|T|$ es el número de hojas de un árbol.

En este trabajo, se ha generado una secuencia de los subárboles asociados a los valores críticos (α) y se ha empleado validación cruzada para seleccionar el árbol óptimo en función de su rendimiento.

Evaluación de árboles

Para la evaluación de los árboles se utiliza también la partición en datos de entrenamiento y prueba, y se obtiene el valor de la tasa de acierto, el índice Kappa, el AUC, la sensibilidad y la especificidad para evaluar la calidad del árbol.

4.4. Clustering no supervisado mediante PAM

El método PAM (Partitioning Around Medoids) es una técnica de agrupamiento no supervisado utilizada para clasificar observaciones en grupos homogéneos (clústeres) según su similitud. A diferencia del método k-means, que utiliza medias como centros de clúster, PAM emplea medoides, que son observaciones reales del conjunto de datos. Esto lo hace más robusto frente a valores atípicos y adecuado para trabajar con datos categóricos o mixtos.

El procedimiento busca identificar un conjunto de k medoides que minimicen la suma de las disimilitudes entre cada observación y el medoide más cercano. El algoritmo sigue los siguientes pasos:

1. Selección inicial de k observaciones como medoides.
2. Asignación de cada observación al clúster del medoide más cercano, en función de una medida de disimilitud.

3. Reasignación iterativa de medoides para minimizar la distancia total dentro de los clústeres.
4. Repetición hasta que las asignaciones no cambien o hasta alcanzar un número máximo de iteraciones.

En este trabajo, la elección del número óptimo de clústeres se ha realizado utilizando el coeficiente de silhouette, que evalúa la calidad de la agrupación considerando simultáneamente la cohesión interna de los clústeres y su separación respecto a otros.

4.5. Análisis de Correspondencias Múltiples (ACM)

El Análisis de Correspondencias Múltiples (ACM) es una técnica de reducción de dimensionalidad utilizada para representar gráficamente las relaciones entre variables categóricas. El objetivo del ACM es proyectar tanto las categorías de las variables como los individuos en un espacio de menor dimensión (generalmente dos dimensiones), de modo que se preserven las asociaciones más relevantes entre categorías. Cuanto más próximos aparecen dos elementos en el plano factorial, más similar es su comportamiento en los datos originales.

En este trabajo, el ACM se ha utilizado para visualizar los clústeres identificados mediante el método PAM, a partir de las variables relacionadas con los síntomas experimentados por las víctimas de violencia sexual digital y su disposición a compartir la experiencia. El porcentaje de variabilidad explicada por las dos primeras dimensiones se ha empleado como indicador de la calidad de la representación, permitiendo validar gráficamente la diferenciación entre los perfiles identificados.

5. Depuración

5.1. Lectura de datos

La base de datos utilizada en este trabajo proviene de la encuesta del Centro Reina Sofía sobre Adolescencia y Juventud, "Generación expuesta – Jóvenes frente a la violencia sexual digital". Este estudio fue publicado el 20 de noviembre de 2024 y tiene como objetivo identificar estrategias para hacer frente a la VSD en jóvenes, para definir agentes estratégicos en la lucha contra estas formas de violencia y para diseñar protocolos que ayuden a prevenir acompañar y reparar a las víctimas.

Esta investigación utilizó la técnica de encuesta online sobre una muestra de 1212 adolescentes y jóvenes de entre 16 y 29 años, residentes en España, seleccionados a partir de un panel online de participantes. El método de muestreo seguido ha sido un muestreo por cuotas, con afijación proporcional por género y edad, además de realizar una ponderación posterior por nivel de estudios. Se alcanzó un error muestral de ±2'81% bajo supuesto de muestreo aleatorio simple (MAS) y máxima heterogeneidad (p=q=0'5), con un nivel de confianza del 95%.

Al tratarse de un tema sensible, esta investigación se realizó siguiendo los principios éticos y legales, priorizando el bienestar de los participantes, respetando la Guía de ESOMAR sobre investigación con personas vulnerables. Además, la encuesta incluye un aviso previo sobre el contenido, para que los participantes decidan su participación de manera informada, en línea con criterios de transparencia y protección de datos personales.

5.1.1. Variables

Esta base de datos da lugar a un total de 332 variables, de las cuales se realizó una selección inicial de 97 de ellas.

Partiendo de esta selección inicial se crearon nuevas variables a partir del valor de otras, se modificaron los valores de algunas para facilitar su compresión y se agruparon aquellas con conceptos altamente relacionados, con el fin de optimizar el análisis y evitar redundancia en los datos.

Tras este proceso de depuración y transformación, la base de datos final utilizada en este trabajo cuenta con 55 variables, que han sido renombradas para facilitar su manejo durante el análisis.

Es importante destacar que varias de las variables incluidas en la base de datos derivan de una misma pregunta de respuesta múltiple, lo que ha dado lugar a la generación de múltiples variables asociadas a una única cuestión. Estas variables han sido renombradas de manera similar, permitiendo identificar fácilmente su relación dentro del estudio.

A continuación, se presentan las tablas de las variables empleadas en el estudio, organizadas por bloques temáticos.

Bloque I: Cuestiones sociodemográficas

Nombre	Descripción	Tipo
genero	Género del individuo	Dicotómica
edad	Edad del individuo	Numérica
nivel_estudios	Nivel de estudios más alto que ha finalizado el individuo	Factor ordinal
orientacion_sexual	Orientación sexual del individuo	Dicotómica
situacion_hogar	Indica con quien vive el individuo actualmente	Factor nominal
situacion_amorosa	Indica si el individuo tiene pareja o una relación estable en la actualidad	Factor nominal

Tabla 1: Variables sociodemográficas

Bloque II: Usos y prácticas tecnológicas

Nombre	Descripción	Tipo
uso_redes	Indica si el individuo utiliza habitualmente o no redes sociales	Dicotómica
uso_citas	Indica si el individuo utiliza habitualmente o no apps de citas	Dicotómica
veo_porno	Indica si el individuo ve porno habitualmente o no	Dicotómica

Tabla 2: Variables de usos y prácticas tecnológicas

Bloque III: Privacidad, intimidad y exposición online

Nombre	Descripción	Tipo
ciberacoso_sufrido	Indica si el incividuo ha sufrido ciberacoso (mensajes insistentes, insultos, stalkeo, que intenten localizarle...)	Dicotómica
cuenta_anonima	Indica la frecuencia con la que el individuo utiliza cuentas anónimas en redes que no le puedan identificar	Escala (1-4)
evitado_subir	Indica la frecuencia con la que el individuo evita subir contenido a Internet por miedo a que le insulten o acosen	Escala (1-4)
bloqueado_perfil	Indica la frecuencia con la que el individuo ha tenido que bloquear perfiles por que le insultaban o acosaban	Escala (1-4)
uso_responsable	Indica la frecuencia con la que el individuo utiliza internet de forma responsable y respetuosa	Escala (1-4)
creado_IA_realizado	Indica si el individuo ha creado contenido íntimo o sexual falso de personas públicas o famosas o de personas cercanas usando IA o similares alguna vez	Dicotómica
difundido_IA_realizado	Indica si el individuo ha difundido contenido íntimo o sexual falso de personas públicas o famosas usando IA o similares alguna vez	Dicotómica

Tabla 3: Variables de privacidad, intimidad y exposición online

Bloque IV: Situaciones de violencia sexual digital experimentadas

Nombre	Descripción	Tipo
insultado_entorno	Indica si alguien del entorno del individuo ha insultado a alguien ya sea por su apariencia física o por su vida sexual	Dicotómica
recibido_sexual_entorno	Indica si alguien del entorno del individuo ha recibido contenido de tipo sexual sin quererlo, sin consentimiento	Dicotómica
compartido_sexual_entorno	Indica si alguien del entorno del individuo ha mostrado fotos de alguien como un objeto sexual o ha difundido contenidos íntimos de alguien	Dicotómica
coaccionado_sexual_entorno	Indica si alguien del entorno del individuo ha ejercido extorsión sexual, amenazado online o presionado a alguien con un fin sexual	Dicotómica
acosado_menor_entorno	Indica si alguien del entorno del individuo ha acosado a un/a menor de edad	Dicotómica
entorno_algo	Indica si el individuo tiene a alguien de su entorno que haya vivido alguna de las prácticas y situaciones incómodas, desagradables o violentas relacionadas con la tecnología del listado	Dicotómica

entorno_reaccion_algo	Indica si el individuo reaccionó de alguna manera ante haber presenciado alguno de los comportamientos o situaciones previamente señalados	Dicotómica
insultado_sufrido	Indica si individuo ha sufrido insultos ya sea por su apariencia física o por su vida sexual	Dicotómica
recibido_sexual_sufrido	Indica si el individuo ha recibido contenido de tipo sexual sin quererlo, sin consentimiento	Dicotómica
compartido_sexual_sufrido	Indica si alguien ha mostrado fotos del individuo como un objeto sexual o ha difundido contenidos íntimos del individuo	Dicotómica
coaccionado_sexual_sufrido	Indica si alguien ha ejercido extorsión sexual, amenazado online o presionado al individuo con un fin sexual	Dicotómica
creado_IA_sufrido	Indica si alguien ha creado imágenes sexuales del individuo con inteligencia artificial (IA)	Dicotómica
acosado_menor_sufrido	Indica si al individuo le ha acosado una persona adulta siendo él menor de edad	Dicotómica
sufrido_algo	Indica si el individuo ha sufrido alguna las prácticas y situaciones incómodas, desagradables o violentas relacionadas con la tecnología del listado	Dicotómica
sufrido_contado	Indica si el individuo se lo contó a alguien tras sufrir alguno de los comportamientos o situaciones previamente señalados	Dicotómica
sufrido_contado_solucionado	Indica si tras contar que sufrió alguno de los comportamientos o situaciones previamente señalados el individuo recibió ayuda para solucionar la situación	Dicotómica
no_contar_no_grave	Indica si el individuo no contó haber sufrido alguno de los comportamientos o situaciones previamente señalados porque consideró que no era algo grave	Dicotómica
no_contar_verguenza	Indica si el individuo no contó haber sufrido alguno de los comportamientos o situaciones previamente señalados por vergüenza o incomodidad al contarlo	Dicotómica
no_contar_miedo	Indica si el individuo no contó haber sufrido alguno de los comportamientos o situaciones previamente señalados porque tenía miedo ya sea de que le hicieran sentir culpable, de que no le creyeran, de que le rechazasen o le marginasen o de sufrir más ataques o violencia	Dicotómica
no_contar_no_ayuda	Indica si el individuo no contó haber sufrido alguno de los comportamientos o situaciones previamente señalados porque no pensó que pudiesen ayudarle	Dicotómica
insultado_realizado	Indica si el individuo ha insultado a alguien ya sea por su apariencia física o por su vida sexual	Dicotómica
recibido_sexual_realizado	Indica si el individuo ha enviado contenido de tipo sexual sin quererlo, sin consentimiento	Dicotómica
compartido_sexual_realizado	Indica si el individuo ha mostrado fotos de alguien como un objeto sexual o ha difundido contenidos íntimos de alguien	Dicotómica
coaccionado_sexual_realizado	Indica si el individuo ha ejercido extorsión sexual, amenazado online o presionado a alguien con un fin sexual	Dicotómica
realizado_algo	Indica si el individuo ha realizado alguna de las prácticas y situaciones incómodas, desagradables o violentas relacionadas con la tecnología del listado	Dicotómica
agresor_pareja	Indica si la persona que provocó alguna de las situaciones desagradables anteriormente descritas al individuo o a alguien de su entorno era alguien con quien mantenía una relación sentimental o de pareja	Dicotómica
agresor_amigo	Indica si la persona que provocó alguna de las situaciones desagradables anteriormente descritas al individuo o a alguien de su entorno era un/a amigo/a	Dicotómica
agresor_familiar	Indica si la persona que provocó alguna de las situaciones desagradables anteriormente descritas al individuo o a alguien de su entorno era alguien del entorno familiar	Dicotómica

agresor_no_cercano	Indica si la persona que provocó alguna de las situaciones desagradables anteriormente descritas al individuo o a alguien de su entorno era conocido, pero no cercano	Dicotómica
agresor_conocido	Indica si la persona que provocó alguna de las situaciones desagradables anteriormente descritas al individuo o a alguien de su entorno era conocida	Dicotómica
canal_redes	Indica si el canal por el que ocurrió alguna de las situaciones desagradables anteriormente descritas al individuo o a alguien de su entorno eran redes sociales (Facebook, Instagram, Twitter-X, Snapchat, TikTok, BeReal...)	Dicotómica
canal_mensajeria	Indica si el canal por el que ocurrió alguna de las situaciones desagradables anteriormente descritas al individuo o a alguien de su entorno eran aplicaciones de mensajería instantánea (Whatsapp, Telegram, Messenger, Kik, Discord...)	Dicotómica
canal_otros	Indica si el canal por el que ocurrió alguna de las situaciones desagradables anteriormente descritas al individuo o a alguien de su entorno fue por otro canal distinto a los anteriores	Dicotómica

Tabla 4: Variables de situaciones de VSD experimentadas

Bloque V: Consecuencias de la violencia sexual digital

Nombre	Descripción	Tipo
sintomas_ansiedad	Indica si el individuo experimentó síntomas de ansiedad o estrés (ansiedad, anguestia, estres, miedo, confusión, problemas de concentración) tras sufrir alguna de las situaciones de vsd listadas.	Dicotómica
sintomas_depresivos	Indica si el individuo experimentó síntomas de depresión o desánimo (depresión, desmotivación, impotencia, soledad, deterioro de la autoestima) tras sufrir alguna de las situaciones de vsd listadas.	Dicotómica
sintomas_emocionales	Indica si el individuo experimentó síntomas de respuesta emocional intensa (enfado, ira, vergüenza, culpa) tras sufrir alguna de las situaciones de vsd listadas.	Dicotómica
sintomas_nada	Indica si el individuo no experimentó ningún síntoma tras sufrir alguna de las situaciones de vsd listadas.	Dicotómica

Tabla 5: Variables de consecuencias de la VSD

Bloque VI: Perspectivas de futuro

Nombre	Descripción	Tipo
evolucion_ultimos	Indica el grado de aumento o disminución que el individuo considera que ha habido en los últimos 10 años en la presencia de situaciones de violencia sexual digital	Escala (1-5)
evolucion_proximos	Indica el grado de aumento o disminución que el individuo considera que habrá en los próximos 10 años en la presencia de situaciones de violencia sexual digital	Escala (1-5)

Tabla 6: Variables de perspectivas de futuro

Para garantizar un tratamiento preciso y riguroso de los datos, las variables han sido clasificadas según su tipología, tal como se muestra en las tablas. Esta clasificación facilita la selección de los métodos estadísticos más adecuados y reduce posibles riesgos de interpretación. Para asegurar una codificación correcta, se ha realizado una recodificación manual basada en dicha clasificación.

Las variables de tipo escala que presentan una progresión lineal razonable entre niveles han sido tratadas como numéricas, con el fin de simplificar los modelos sin

comprometer su capacidad interpretativa. En particular, la variable *nivel_estudios* se ha considerado como un factor ordinal, ya que, aunque presenta un orden lógico, no puede asumirse que los saltos entre niveles tengan efectos equivalentes sobre el resultado.

5.2. Análisis de datos ausentes

5.2.1. Asignación y visualización de los datos ausentes

Teniendo en cuenta el diseño del cuestionario, en el que algunas preguntas solo deben ser respondidas por un subconjunto de la muestra según sus respuestas previas, se ha procedido a la recodificación de aquellas variables en las que algunas observaciones no son aplicables. En estos casos, se ha asignado el valor 99 a los individuos para los que la pregunta no corresponde, garantizando que solo los valores realmente ausentes sean codificados como *NA*. Además, se han considerado como datos faltantes las categorías NS/NC de las preguntas del cuestionario.

Para garantizar que los datos sean tratados de manera rigurosa, es fundamental evaluar la presencia y el patrón de valores ausentes en las variables seleccionadas. Para ello, se presenta el gráfico que visualiza el porcentaje de valores faltantes por variable.

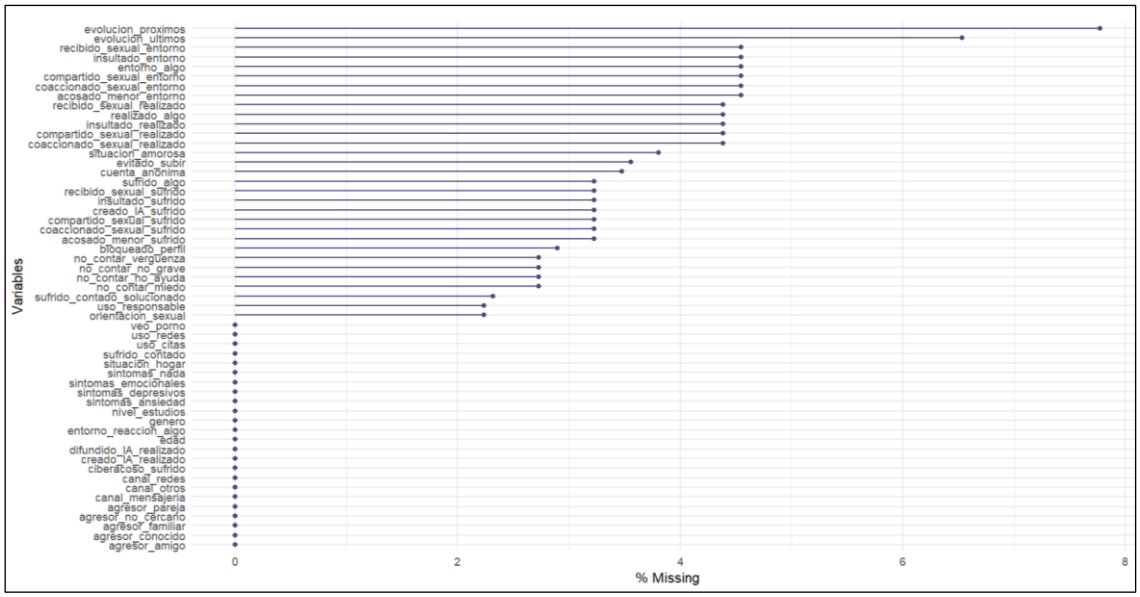


Figura 1: Porcentaje de datos faltantes por variable

Se observa que todas las variables tienen un porcentaje de valores ausentes inferior al 10%, lo que indica que ninguna variable tiene una cantidad de *missings* tan elevada como para eliminarla del análisis. Cabe destacar que las variables con mayor proporción de ausentes son las aquellas relacionadas con la percepción sobre la evolución de la VSD.

Se presenta a continuación una visualización que muestra cómo se distribuyen los valores faltantes a lo largo del conjunto de datos.

Figura 2: Distribución de datos faltantes antes de la imputación

Se observan varios aspectos relevantes en los datos. En primer lugar, ningún individuo cuenta con un número excesivo de variables con valor *NA*, lo que indica que todos los participantes son válidos para el análisis.

Por otro lado, el conjunto de datos presenta únicamente un 2’1% de valores ausentes, una proporción baja que será optimizada aún más mediante un tratamiento adecuado de los datos faltantes.

5.2.2. Tratamiento de los datos ausentes

Para abordar esta cuestión, se ha optado por aplicar técnicas de imputación que permitan preservar la estructura original de los datos sin comprometer la validez de las conclusiones. En particular, se utilizará el algoritmo MissForest, ya que es una técnica robusta que permite imputar datos mixtos (categóricos y numéricos) de forma no paramétrica. La imputación se realizará únicamente en variables de carácter sociodemográfico, uso tecnológico y percepción de evolución, mientras que aquellas relacionadas directamente con experiencias personales de violencia o consecuencias emocionales no serán imputadas, ya que no se quiere comprometer la integridad de la información sensible para el análisis.

Variables sometidas a la imputación	genero edad nivel_estudios orientacion_sexual situacion_hogar situacion_amorosa uso_redes uso_citas veo_porno cuenta_anonima evitado_subir bloqueado_perfil uso_responsable evolucion_ultimos evolucion_proximos

Tabla 7: Variables imputadas

Tras aplicar el algoritmo MissForest se han obtenido un error medio cuadrático normalizado (NRMSE) del 11’85% para las variables cuantitativas y un 9’29% de clasificación errónea para las variables categóricas. Ambos valores están dentro de un rango aceptable por lo que se da por buena la imputación realizada.

Se muestran los valores faltantes restantes, que corresponden a las variables que no han sido sometidas a imputación.

Figura 3: Distribución de datos faltantes después de la imputación

Después del proceso de imputación, el porcentaje de valores faltantes se ha reducido al 1’5%. Dado que las variables restantes no pueden ser imputadas por su sensibilidad en el análisis y las técnicas que se van a emplear no admiten valores *NA*, se ha optado por eliminar aquellas filas que contienen datos ausentes. Al tratarse de un número reducido de observaciones, eliminarlas mantiene la calidad del estudio sin comprometer significativamente la representatividad de los datos.

Tras todo el proceso de tratamiento de los datos ausentes la base de datos final queda conformada por 55 variables y 1063 observaciones.

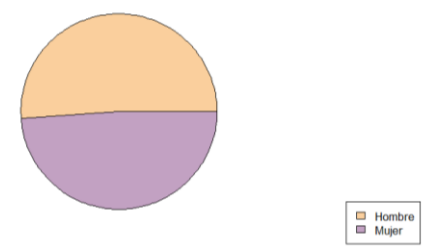
6. Análisis descriptivo

6.1. Comprensión de variables

Para garantizar una comprensión clara de las características de los individuos del conjunto de datos se lleva a cabo un análisis descriptivo de cada una de las variables.

6.1.1. Cuestiones sociodemográficas

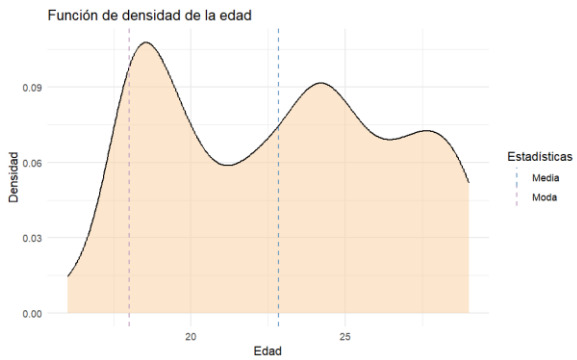
Género



Categoría	Porcentaje
Hombre	51,08%
Mujer	48,92%

En la muestra, hombres y mujeres están representados de forma prácticamente equitativa.

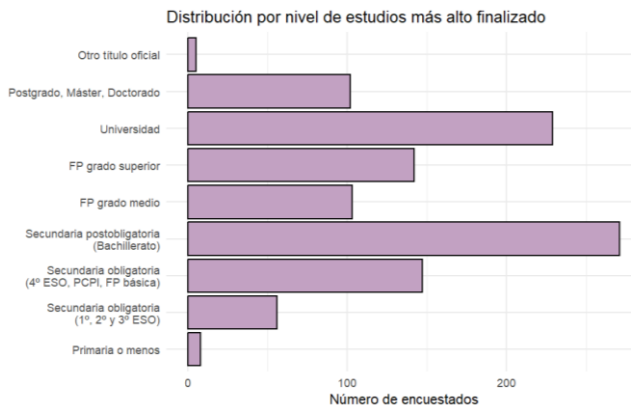
Edad



Estadístico	Valor
Mínimo	16
1º cuartil	19
Mediana	23
Media	22,84
3º cuartil	26
Máximo	29

La edad media de los individuos de la muestra es de 22’84 años, con una concentración mayoritaria en torno a los 18 años.

Nivel de estudios más alto finalizado



Categoría	Porcentaje
Primaria o menos	0,75%
Secundaria obligatoria (1º, 2º y 3º ESO)	5,27%
Secundaria obligatoria (4º ESO, PCPI, FP básica)	13,83%
Secundaria postobligatoria (Bachillerato)	25,49%
FP grado medio	9,69%
FP grado superior	13,36%
Universidad	21,54%
Postgrado, Máster, Doctorado	9,60%
Otro título oficial	0,47%

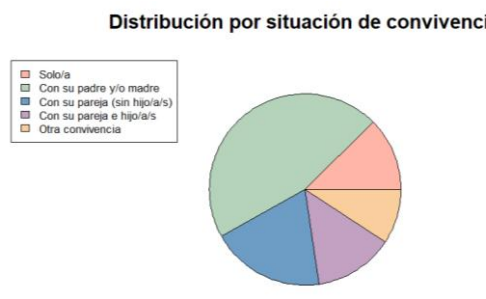
El 25’49% de los encuestados, han completado al menos hasta la educación secundaria postobligatoria.

Orientación sexual

Categoría	Porcentaje
Heterosexual	84,20%
Pertenece al colectivo LGTBIQ+	15,80%

La mayoría de los encuestados se identifica como heterosexual (84’20%). Además, se observa una representación minoritaria pero relevante de personas que pertenecen al colectivo LGTBIQ+.

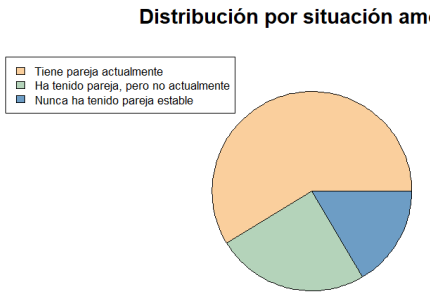
Situación de vivienda



Categoría	Porcentaje
Solo/a	12,42%
Con padre y/o madre	45,81%
Con pareja (sin hijo/a/s)	19,19%
Con pareja (con hijo/a/s)	13,36%
Otra convivencia	9,22%

El tipo de convivencia más frecuente es con padre y/o madre, con un 45’81% de la muestra.

Situación amorosa



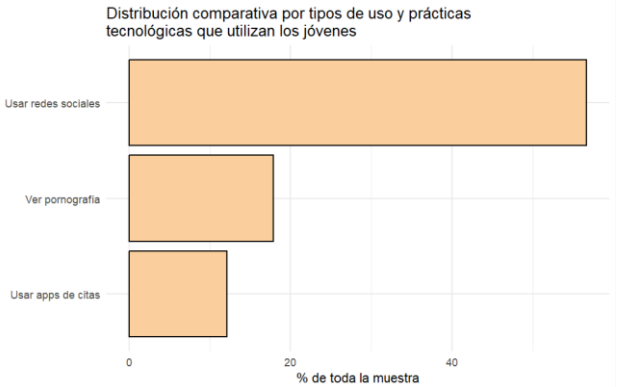
Categoría	Porcentaje
Tiene pareja actualmente	58,51%
Ha tenido pareja, pero no actualmente	24,74%
Nunca ha tenido pareja estable	16,75%

El 58’51% de los participantes indica tener pareja actualmente.

6.1.2. Uso y prácticas tecnológicas

Practicas tecnológicas utilizadas frecuentemente

Dado que las tres variables analizadas son dicotómicas y provienen de una misma pregunta de respuesta múltiple, se presentan de forma conjunta en un único gráfico.

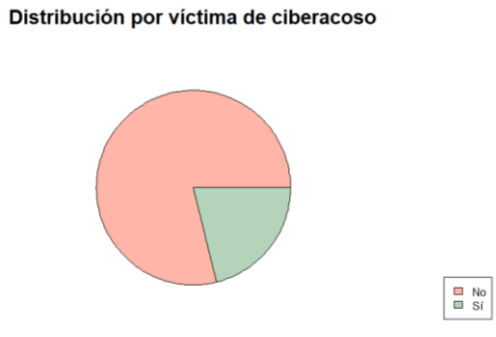


Variable	Porcentaje de sí
uso_redes	56,63%
uso_citas	12,04%
veo_porno	17,78%

El gráfico muestra que el 56’63% de los encuestados utiliza redes sociales de manera habitual, lo que sugiere que estas plataformas desempeñan un papel central en sus prácticas digitales.

6.1.3. Privacidad, intimidad y exposición online

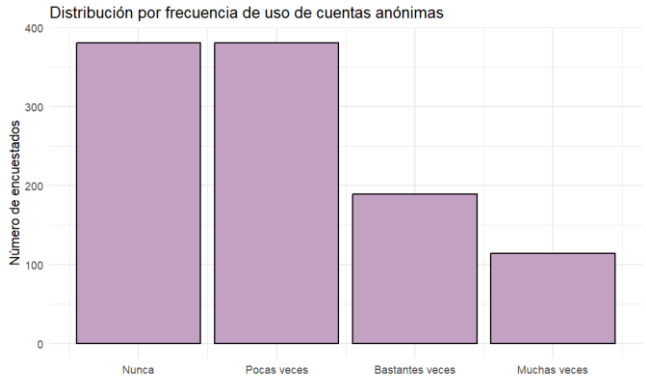
Víctimas de ciberacoso



Categoría	Porcentaje
No	78,74%
Sí	21,26%

En el conjunto de datos predominan los individuos que no han sufrido ciberacoso; solo el 21’22% declara haber sido víctima de este tipo de acoso.

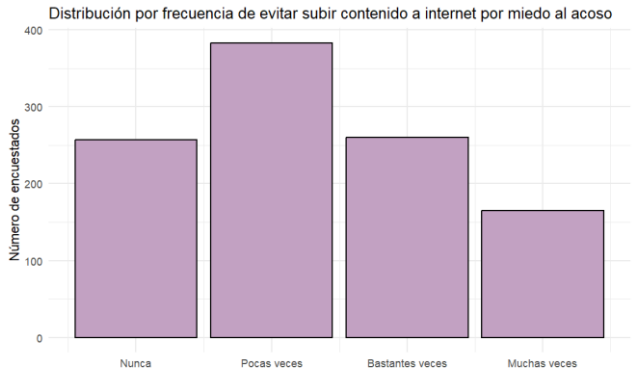
Frecuencia de uso de cuentas anónimas



Categoría	Porcentaje
Nunca	35,75%
Pocas veces	35,75%
Bastantes veces	17,78%
Muchas veces	10,72%

Aproximadamente el 70% de la muestra indica que no utiliza, o utiliza muy poco, cuentas anónimas.

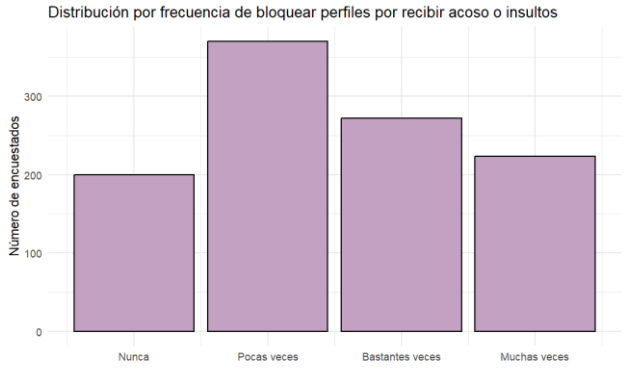
Frecuencia de evitar subir contenido a internet por miedo al acoso



Categoría	Porcentaje
Nunca	24,18%
Pocas veces	36,03%
Bastantes veces	24,27%
Muchas veces	15,52%

La mayoría de los participantes ha evitado subir contenido a internet por temor a ser acosados. Solo el 24’18% afirma no haber tenido que actuar de esta forma nunca.

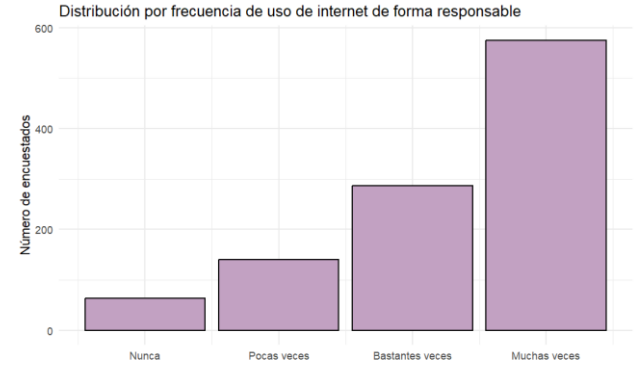
Frecuencia de bloquear perfiles por recibir acoso o insultos



Categoría	Porcentaje
Nunca	18,82%
Pocas veces	35,18%
Bastantes veces	25,02%
Muchas veces	20,98%

También es mayoritaria la proporción de personas que ha tenido que bloquear perfiles alguna vez por recibir acoso, siendo “pocas veces” la categoría más frecuente. Solo un 18’82% no ha tenido que hacerlo nunca.

Frecuencia de uso responsable y respetuoso de internet



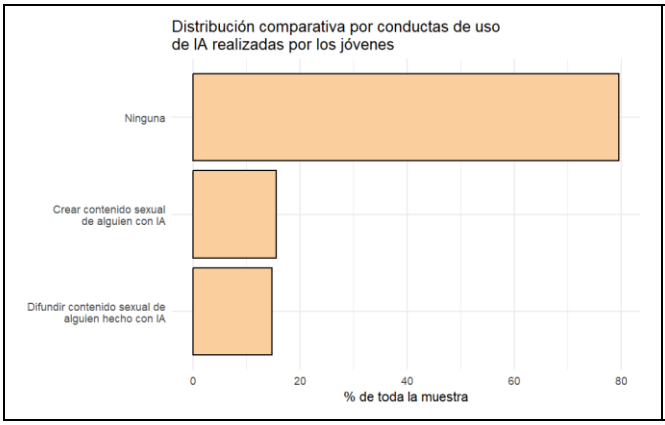
Categoría	Porcentaje
Nunca	5,93%
Pocas veces	13,17%
Bastantes veces	27,19%
Muchas veces	53,71%

El 53’71% de los encuestados declara utilizar internet de forma habitual de manera responsable y respetuosa.

Conductas de uso de IA

Dado que las dos variables analizadas son dicotómicas y provienen de una misma pregunta de respuesta múltiple, se presentan de forma conjunta en un único gráfico.

Para complementar este análisis, se ha generado la variable *nada_realizado*, derivada de las anteriores, con el objetivo de incluir en el gráfico también a aquellos individuos que no se han realizado ninguna de estas acciones.

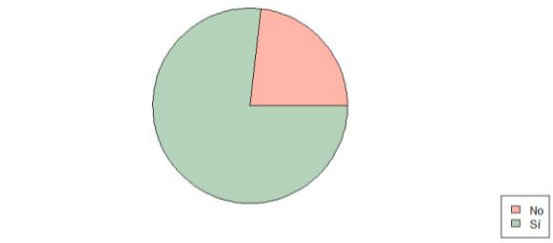


Variable	Porcentaje de sí
creado_IA_realizado	15,62%
difundido_IA_realizado	14,69%
nada_realizado	79,49%

El 79’49% de los participantes no ha llevado a cabo ninguna de las dos acciones. Entre quienes sí lo han hecho, se observa una proporción muy similar entre quienes han creado contenido sexual mediante IA y quienes lo han difundido.

6.1.4. Situaciones de violencia sexual digital experimentadas

Testigos de VSD

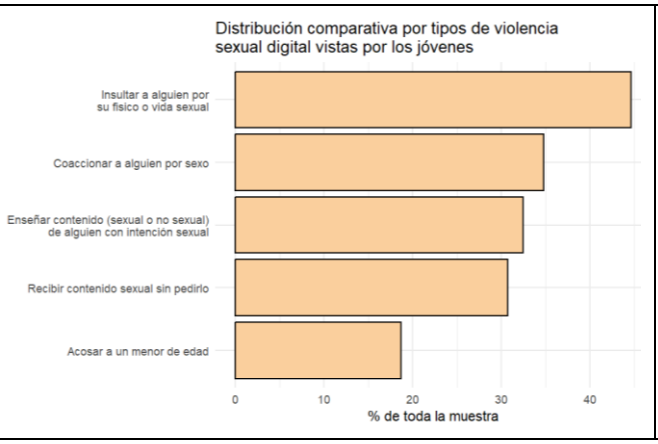


Categoría	Porcentaje
No	23,24%
Sí	76,76%

El 76’76% de los participantes afirma conocer algún caso de VSD en su entorno.

Tipos de VSD presenciada

Con el objetivo de visualizar la distribución de estos casos, todos los tipos de VSD presenciados se agrupan en un único gráfico, considerando que una misma persona puede haber estado expuesta a más de una situación.

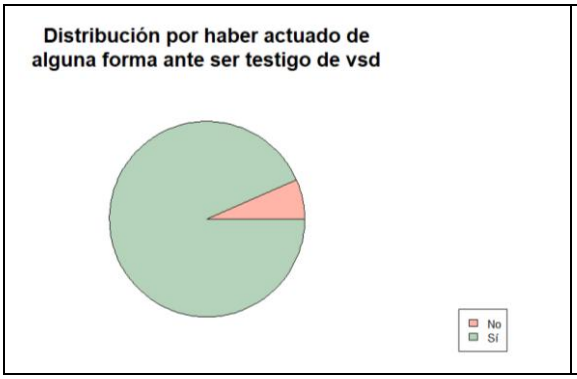


Variable	Porcentaje de sí
recibido_sexual_entorno	30,67%
insultado_entorno	44,59%
compartido_sexual_entorno	32,55%
coaccionado_sexual_entorno	34,81%
acosado_menor_entorno	18,63%

La forma más común de VSD presenciada por los encuestados ha sido insultar a alguien por su físico o vida sexual, una situación observada por el 44’59% de la muestra.

Reacción ante ser testigo de VSD

Para el análisis de esta variable, se ha seleccionado la submuestra de quienes han presenciado VSD en su entorno.



Categoría	Porcentaje
No	6,62%
Sí	93,38%

El 93’38% de los individuos que han presenciado VSD afirma haber reaccionado de alguna forma ante la situación.

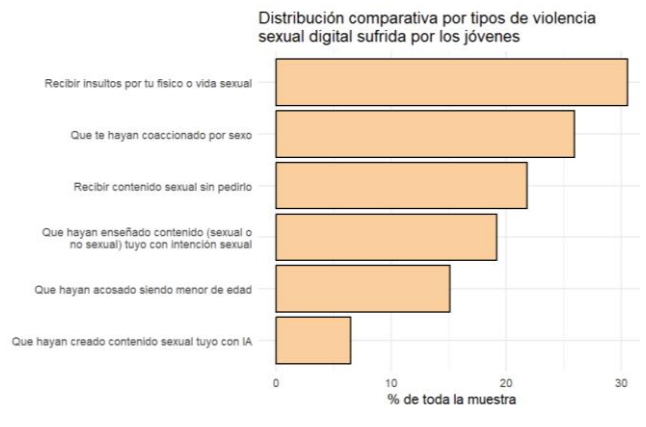
Víctimas de VSD

Categoría	Porcentaje
No	39,70%
Sí	60,30%

En cuanto a la experiencia directa, el 60’30% de los encuestados ha sufrido al menos una de las situaciones de VSD descritas en el cuestionario.

Tipos de VSD sufrida

Al igual que con las variables del entorno, todos los tipos de VSD sufridos se agrupan en un único gráfico, considerando que una misma persona puede haber sido víctima de más de una situación.

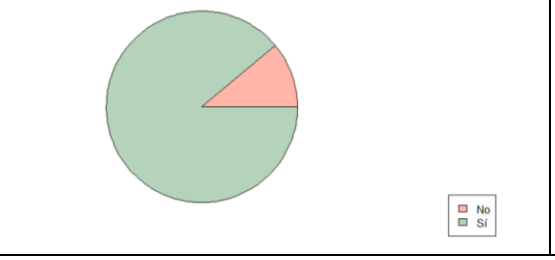


Variable	Porcentaje de sí
recibido_sexual_sufrido	21,82%
insultado_sufrido	30,48%
compartido_sexual_sufrido	19,19%
coaccionado_sexual_sufrido	25,96%
creado_IA_sufrido	6,49%
acosado_menor_sufrido	15,15%

La forma más común de VSD sufrida ha sido recibir insultos por la apariencia física o la vida sexual, con un 30’48% de representación en la muestra.

Reacción ante sufrir VSD (contar o no la experiencia)

Para el análisis de esta variable, se ha seleccionado la submuestra de quienes han sufrido VSD.

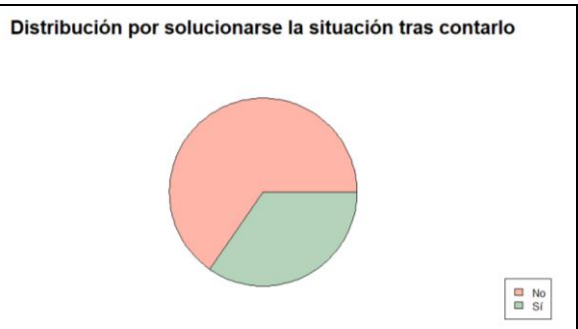


Categoría	Porcentaje
No	11,08%
Sí	88,92%

Entre quienes han sufrido VSD, el 88’92% lo ha contado a alguien.

Situaciones de VSD solucionadas tras contarlo

Para el análisis de esta variable, se ha seleccionado la submuestra de víctimas que han contado su experiencia.



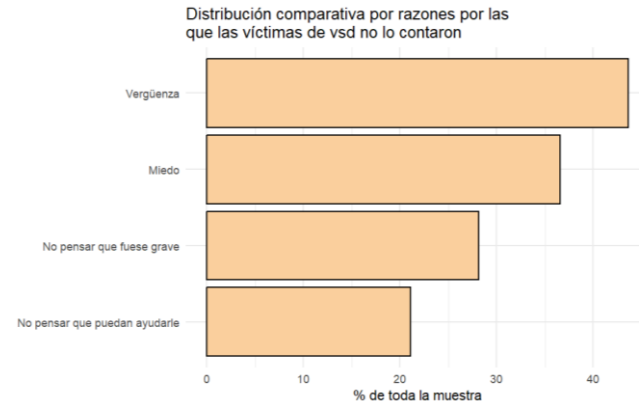
Categoría	Porcentaje
No	65,26%
Sí	34,74%

En el 34’74% de los casos en los que la víctima compartió su experiencia, la situación se solucionó.

Razones de no haber contado la experiencia sufrida

Para el análisis de las variables siguientes, se ha seleccionado la submuestra de personas que han sufrido VSD y no lo han contado a nadie.

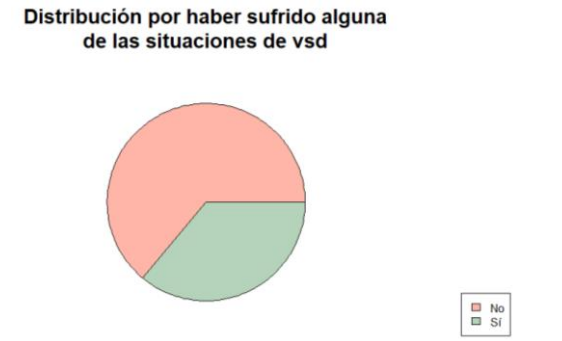
Además, al ser todas variables dicotómicas derivadas de una pregunta de respuesta múltiple, se presentan de forma conjunta en un único gráfico.



Variable	Porcentaje de sí
no_contar_no_grave	28,17%
no_contar_verguenza	43,66%
no_contar_no_ayuda	21,13%
no_contar_miedo	36,62%

La razón más frecuente para no compartir la experiencia fue sentir vergüenza (43’66%), seguida de sentir miedo (36’62%).

Agresores de VSD

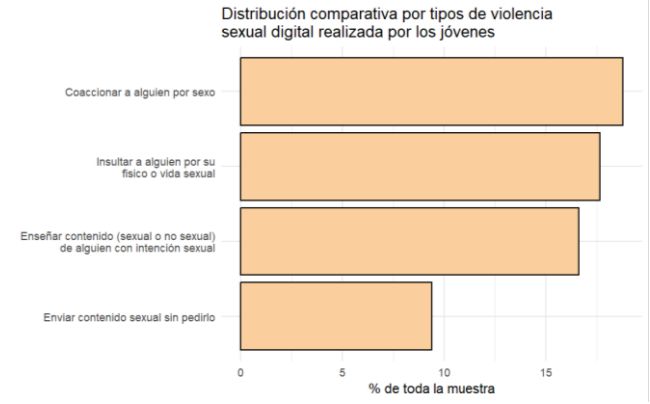


Categoría	Porcentaje
No	63,78%
Sí	36,22%

El 36’22% de los participantes declara haber ejercido algún acto de VSD sobre otra persona. Aunque la mayoría no lo ha hecho, se trata de una proporción considerable.

Tipos de VSD realizada

Todos los tipos de VSD ejercida se agrupan en un único gráfico, considerando que una misma persona puede haber realizado más de un tipo de acción.

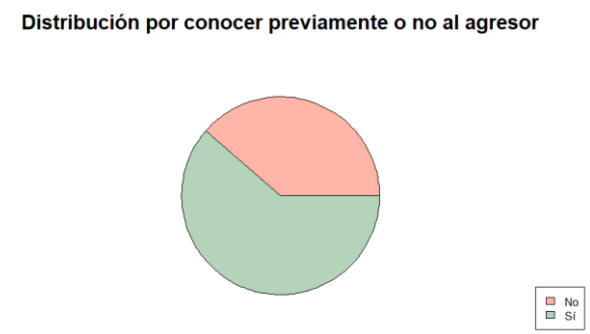


Variable	Porcentaje de sí
recibido_sexual_realizado	9,40%
insultado_realizado	17,69%
compartido_sexual_realizado	16,65%
coaccionado_sexual_realizado	18,81%

El acto más frecuente ha sido ejercer extorsión sexual, amenazar online o presionar con fines sexuales, con un 18’81%.

Victimas que guardaban una relación con el agresor

Para analizar las variables sobre el vínculo con el agresor, se ha seleccionado la submuestra de personas que han sufrido o presenciado VSD.



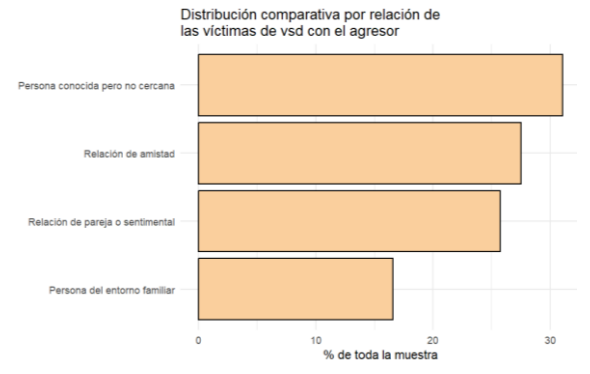
Categoría	Porcentaje
No	38,71%
Sí	61,29%

El 61’29% de los casos presenciados o sufridos de VSD fueron cometidos por alguien conocido.

Tipos de relación con el agresor

En este caso, también se analiza la submuestra de personas que han sufrido o presenciado VSD y conocían al agresor.

Al tratarse de variables dicotómicas derivadas de una misma pregunta multirrespuesta, se agrupan en un único gráfico.



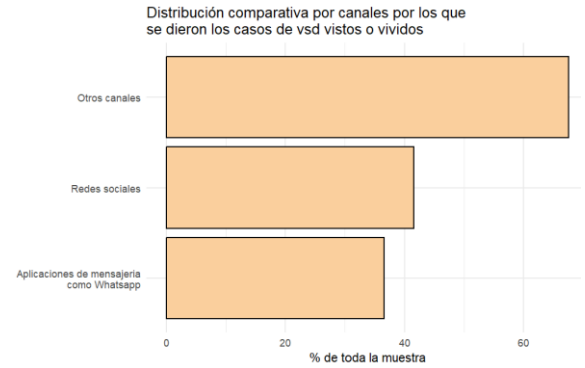
Variable	Porcentaje de sí
agresor_pareja	25,81%
agresor_amigo	27,58%
agresor_familiar	16,45%
agresor_no_cercano	30,97%

El 30’97% de las víctimas mantenía una relación no cercana con su agresor, seguido del 25’58% que indica haber tenido una relación de amistad.

Canales por los que ocurrió la experiencia de VSD

Por último, al analizar los canales a través de los cuales ocurrieron los casos de VSD, se selecciona nuevamente la submuestra de quienes han sufrido o presenciado este tipo de violencia.

Al igual que en los casos anteriores, las variables se agrupan en un único gráfico por su tipología dicotómica y su origen común en una pregunta multirrespuesta.



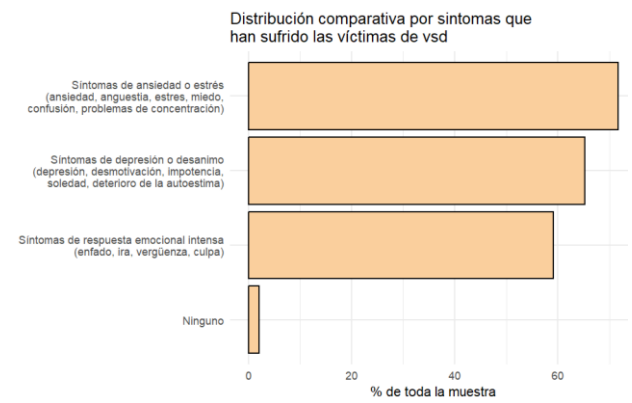
Variable	Porcentaje de sí
canal_mensajeria	36,61%
canal_redes	41,45%
canal_otros	67,74%

El 41’45% de los casos recogidos en la muestra ocurrieron a través de redes sociales.

6.1.5. Consecuencias de la violencia sexual digital

Síntomas experimentados tras sufrir VSD

Para el análisis de estas variables, se ha seleccionado la submuestra de personas que han sufrido violencia sexual digital.

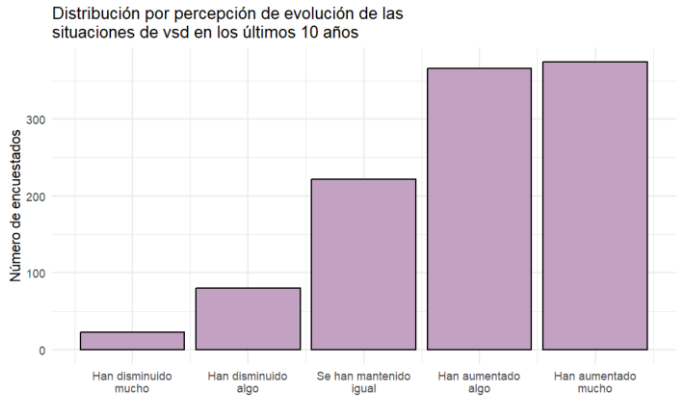


Variable	Porcentaje de sí
sintomas_nada	2,03%
sintomas_ emocionales	59,13%
sintomas_ depresivos	65,21%
sintomas_ ansiedad	71,61%

La mayoría de quienes han vivido este tipo de violencia ha experimentado algún síntoma asociado, siendo los síntomas de ansiedad los más frecuentes, con una prevalencia del 71’61%.

6.1.6. Perspectivas de futuro

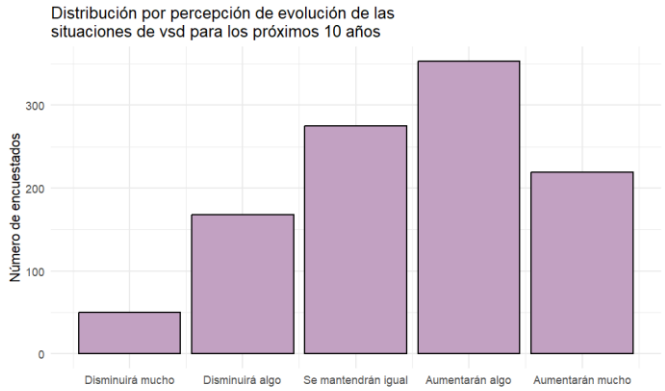
Percepción de evolución de la VSD en los últimos 10 años



Categoría	Porcentaje
Han disminuido mucho	2,16%
Han disminuido algo	7,53%
Se han mantenido igual	21,07%
Han aumentado algo	34,06%
Han aumentado mucho	35,18%

La mayoría de los participantes percibe que, en los últimos diez años, los casos de violencia sexual han aumentado significativamente.

Percepción de la evolución de la VSD en los próximos 10 años



Categoría	Porcentaje
Disminuirá mucho	4,70%
Disminuirá algo	15,90%
Se mantendrán igual	26,34%
Aumentarán algo	32,46%
Aumentarán mucho	20,60%

Asimismo, la mayoría opina que, en los próximos diez años, los casos de violencia sexual digital aumentarán moderadamente.

6.2. Exploración de relaciones entre variables

En esta sección, se presentan cruces simples pero visuales que permiten identificar patrones y posibles relaciones entre variables. Se incluyen los cruces que han presentado una relación de interés.

Prácticas y uso tecnológico según el género

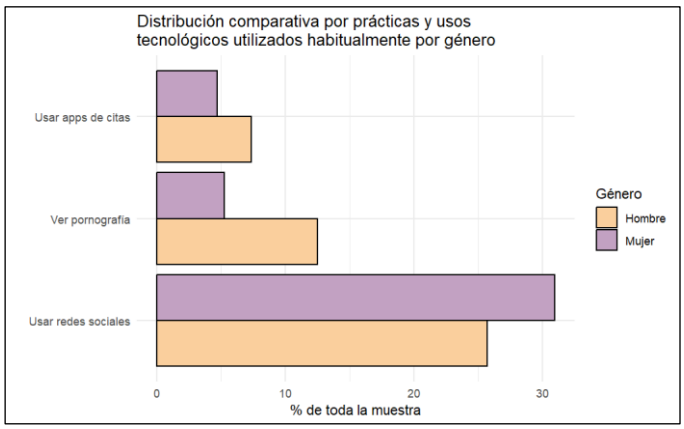


Figura 4: Distribución de prácticas y uso tecnológico según el género

Los hombres presentan una mayor frecuencia de uso de aplicaciones de citas en comparación con las mujeres. En cuanto al consumo de pornografía, la diferencia es especialmente notable, con un porcentaje significativamente más alto entre los hombres. Las mujeres, en cambio, destacan en el uso habitual de redes sociales.

Estos resultados sugieren que existen diferencias marcadas en las prácticas digitales en función del género.

Víctimas de ciberacoso según el género

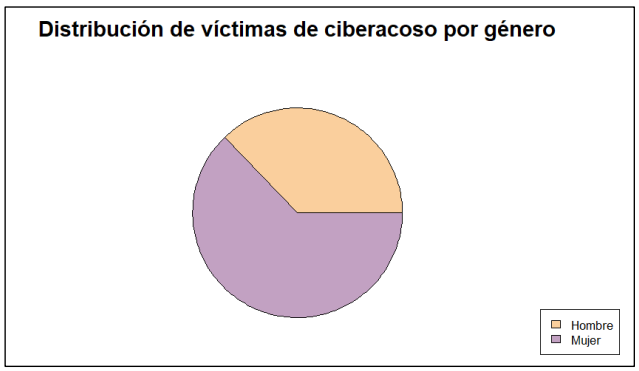


Figura 5: Distribución de víctimas de ciberacoso por género

En la muestra, se observa un porcentaje significativamente mayor de mujeres que han sido víctimas de ciberacoso.

Víctimas de violencia sexual digital según el género

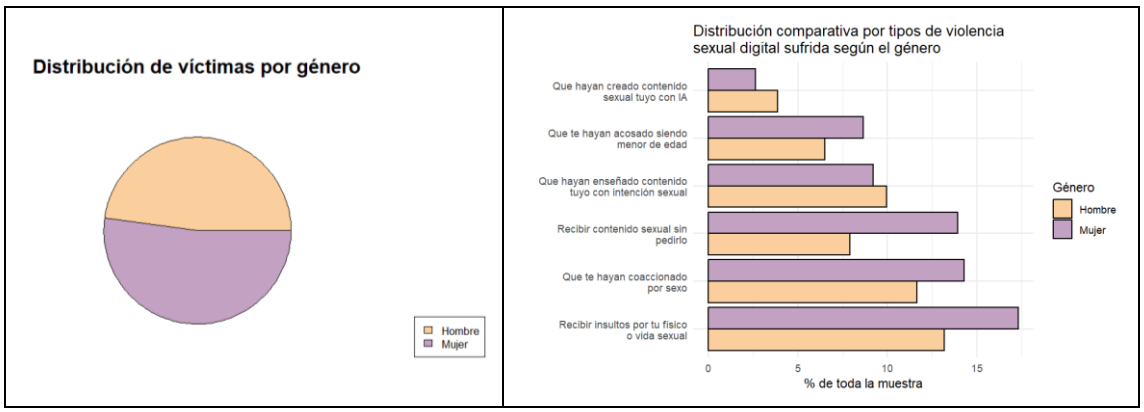


Figura 6: Distribución de víctimas de VSD por género

La mayoría de las personas que han sufrido violencia sexual digital son mujeres, mientras que el porcentaje de hombres afectados es ligeramente inferior.

Las mujeres son el grupo más afectado en casi todas las categorías, excepto en aquellas relacionadas con la creación o difusión de contenido sexual sobre la víctima, donde los hombres presentan una mayor proporción.

Actuación ante sufrir violencia sexual digital según el género

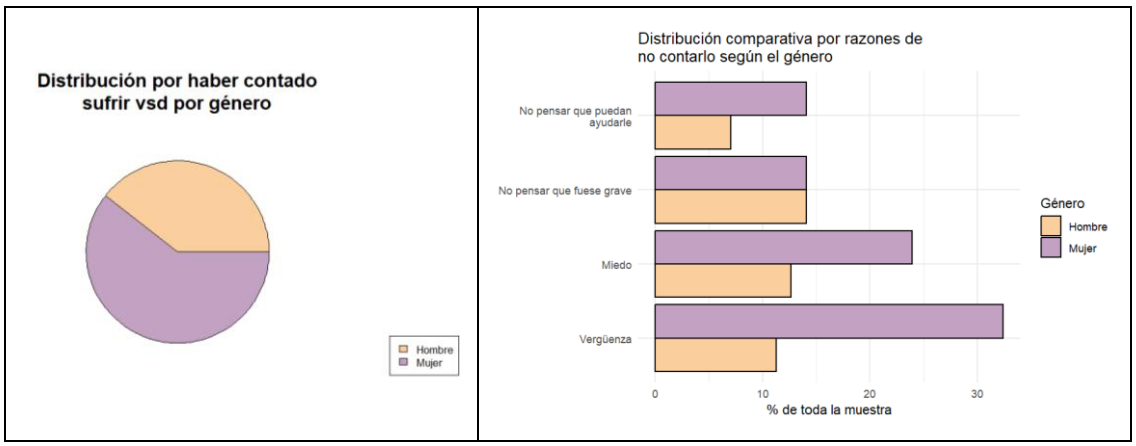


Figura 7: Distribución de compartir la experiencia sufrida por género

También se observa que la mayoría de las personas que no compartieron su experiencia de VSD son mujeres.

En cuanto a las razones, el motivo más común entre los hombres fue no considerar grave lo ocurrido, mientras que, entre las mujeres, predominó la vergüenza.

Síntomas experimentados tras sufrir violencia sexual digital según el género

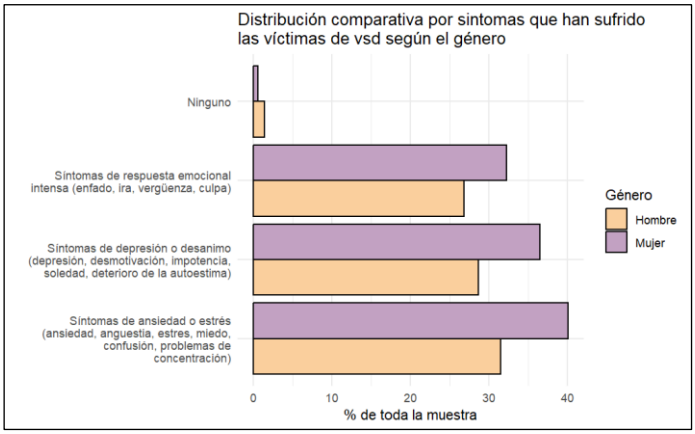


Figura 8: Distribución de síntomas experimentados tras sufrir VSD por género

Las mujeres reportan síntomas emocionales con mayor frecuencia en todas las categorías analizadas. Este patrón sugiere que el género influye en la forma en que se experimentan y procesan las consecuencias de la VSD.

Agresores de violencia sexual digital según el género

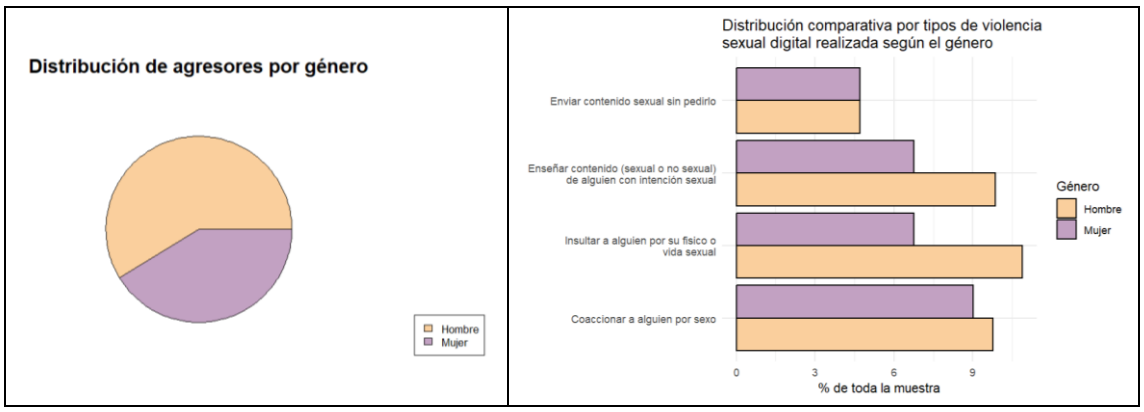


Figura 9: Distribución de agresores por género

Respecto al perfil de agresores, la mayoría son hombres. En todos los tipos de violencia sexual digital ejercida, los hombres presentan proporciones más altas.

7. Análisis estadístico

7.1. Modelización exploratoria de víctimas y agresores

En este apartado se analiza la relación entre distintas variables y la probabilidad de haber sido víctima o agresor de violencia sexual digital, mediante la aplicación de modelos de regresión logística binaria. El objetivo es explorar qué factores se asocian significativamente con cada uno de los perfiles. Para ello, se desarrollan dos modelos diferenciados: uno para el perfil de víctima, cuya variable dependiente es *sufrido_algo*, y otro para el perfil de agresor, cuya variable dependiente es *realizado_algo*.

Se ha seleccionado un subconjunto de variables explicativas candidatas para identificar posibles características asociadas a cada perfil.

Variables explicativas candidatas	genero edad nivel_estudios orientacion_sexual situacion_hogar situacion_amorosa uso_redes uso_citas veo_porno cuenta_anonima uso_responsable

Tabla 8: Variables explicativas candidatas para la regresión

Se han excluido del análisis aquellas variables que contienen información directamente relacionada con las variables dependientes, como los síntomas posteriores a la violencia o los detalles del acto cometido o sufrido, con el objetivo de garantizar que los modelos reflejen características previas al suceso, y no consecuencias del mismo.

Estrategia de modelización

Para cada uno de los perfiles (víctima y agresor) se construyen varios modelos iniciales de regresión logística binaria:

- Un modelo manual, que parte de un modelo saturado que incluye todas las variables candidatas, del cual se eliminan progresivamente las variables no significativas.
- Dos modelos backward, uno guiado por BIC y otro por AIC.
- Dos modelos forward, uno guiado por BIC y otro por AIC.
- Dos modelos stepwise, uno guiado por BIC y otro por AIC.

Una vez construidos estos modelos, se comparan en términos de calidad de ajuste y estabilidad de los coeficientes, seleccionándose el más adecuado desde un enfoque exploratorio que permita una interpretación detallada y fundamentada.

Posteriormente, se estima un modelo adicional mediante regresión logística con penalización LASSO. Esta técnica permite la selección automática de variables y reduce el riesgo de sobreajuste mediante la penalización de los coeficientes, lo que contribuye a identificar un conjunto de factores asociados más robusto. A continuación, se contrastan ambos enfoques (modelo clásico y modelo penalizado) para alcanzar una interpretación más completa.

Como técnica complementaria a los modelos logísticos, se construye un árbol de clasificación, con el objetivo de identificar perfiles de riesgo asociados a la victimización y la agresión en violencia sexual digital. Esta herramienta permite generar reglas de decisión jerárquicas e interpretar visualmente los factores más relevantes en la predicción de cada comportamiento. Además, los árboles facilitan la detección de interacciones no lineales entre variables, aportando una perspectiva adicional frente a los modelos lineales.

Validación de los modelos

Para evaluar la estabilidad y el rendimiento de los modelos, se divide previamente la muestra en dos subconjuntos: un conjunto de entrenamiento (80% de la muestra), utilizado para la construcción del modelo, y un conjunto de prueba (20%), reservado para su evaluación. Este procedimiento de validación cruzada simple permite controlar el riesgo de sobreajuste y proporciona una estimación más realista del comportamiento del modelo frente a nuevos datos.

La evaluación de los modelos se centra en tres métricas: el AUC (área bajo la curva ROC), la tasa de acierto y el índice Kappa. Estas métricas permiten valorar tanto la capacidad discriminativa del modelo como su concordancia en la clasificación.

Balance de clases

Dado que las variables dependientes son binarias, es importante tener en mente la proporción entre clases (eventos y no eventos) para cada una de ellas:

	sufrido_algo	realizado_algo
proporciones	0: 0,397 1: 0,603	0: 0,638 1: 0,362

Tabla 9: Balance de clases de las variables dependientes

Se observa que ambas variables están ligeramente desbalanceadas, lo que puede influir en las métricas de evaluación. Por este motivo, se ajustará el umbral de clasificación, por defecto 0’5, para obtener métricas más realistas y evitar que la clasificación esté sesgada hacia la clase mayoritaria.

7.1.1. Modelos de regresión logística binaria – víctimas

Construcción de los modelos

Tal como se ha mencionado, se han construido varios modelos iniciales para explorar los factores asociados al perfil de víctima de VSD.

Se observa tras la construcción que los modelos obtenidos mediante los métodos de stepwise, backward y forward incluyen las mismas variables, por lo que, para evitar redundancias, solo se presenta uno de ellos.

A continuación se muestran las variables que han sido incluidas en cada modelo.

Variable	manual (Mode1)	StepBIC (Model2)	StepAIC (Model3)
genero	✓	✓	✓
orientación_sexual	✓	×	✓
situacion_amorosa	✓	✓	✓
cuenta_anonima	✓	✓	✓
uso_responsable	✓	✓	✓
edad	×	×	✓
uso_redes	×	×	✓

Tabla 10: Variables incluidas en cada modelo - víctimas

Las variables que no aparecen en la tabla no han sido seleccionadas en ninguno de los modelos.

Comparación de los modelos

Se presenta a continuación una comparación visual de los tres modelos construidos. Cada uno ha sido evaluado mediante validación cruzada simple, utilizando como métricas la tasa de acierto, el AUC y el índice Kappa, con el objetivo de valorar su rendimiento.

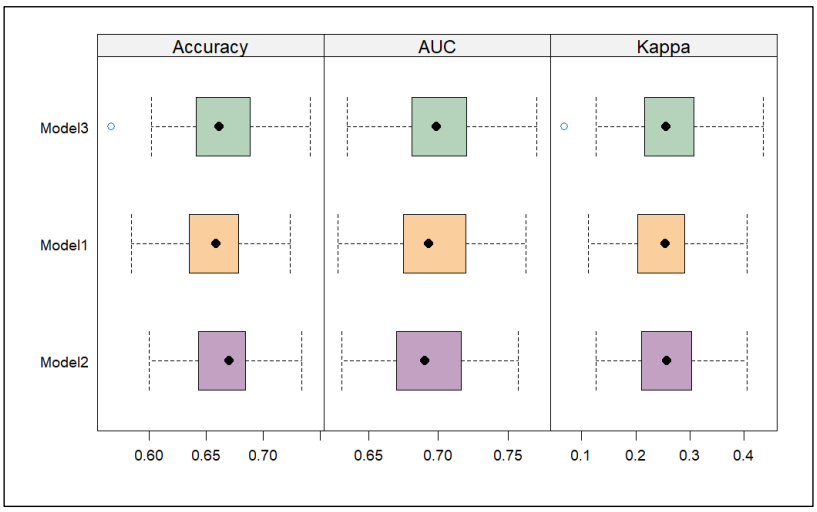


Figura 10: Comparación de modelos - víctimas

Los resultados muestran que no existen diferencias significativas entre los tres modelos en cuanto a las métricas obtenidas. En general, todos presentan un rendimiento predictivo limitado. No obstante, el tercer modelo (StepAIC) presenta

valores ligeramente superiores en las tres métricas evaluadas y, además, retiene el mayor número de variables. Esto lo convierte en la opción más adecuada para los fines del estudio, ya que permite una interpretación exploratoria más completa y detallada.

Por tanto, se selecciona el modelo StepAIC, compuesto por siete variables, como modelo de referencia para el análisis del perfil de víctima.

Evaluación del mejor modelo (modelo stepwise guiado por AIC)

El modelo ha sido evaluado sobre el conjunto de entrenamiento y el conjunto de prueba, obteniéndose los siguientes resultados:

	Datos de entrenamiento	Datos de prueba
Precisión (tasa de acierto)	0,67	0,59
AUC	0,71	0,64
Indice Kappa	0,33	0,18
Sensibilidad	0,65	0,58
Especificidad	0,68	0,61

Tabla 11: Métricas de evaluación del modelo stepwise - víctimas

Se observa una ligera diferencia entre los resultados obtenidos sobre ambos conjuntos, lo que indica que el modelo no es tan estable como se desearía.

Las métricas reflejan un rendimiento predictivo limitado. Una precisión del 59% sugiere que el modelo acierta solo ligeramente por encima del azar (50%), y un AUC de 0’64 indica una capacidad de discriminación baja entre quienes han sufrido violencia sexual digital y quienes no. El índice Kappa de 0’18 refuerza esta limitación, al evidenciar una escasa concordancia entre las predicciones y los valores reales. No obstante, las métricas de sensibilidad (0’58) y especificidad (0’61) indican que el modelo no está fuertemente sesgado hacia una de las clases, lo cual es positivo desde el punto de vista interpretativo.

A pesar de su baja capacidad predictiva, el modelo conserva un valor exploratorio importante, ya que permite identificar variables significativamente asociadas con la probabilidad de haber sido víctima de violencia sexual digital.

Conclusiones exploratorias sacadas del modelo

A continuación, se presenta la tabla con los odds ratios estimados, que permiten cuantificar la magnitud y dirección de la relación entre cada parámetro y la probabilidad de haber sido víctima de VSD.

Parámetro	odds ratio
cuenta_anonima	1,58
situacion_amorosa2	1,06
situacion_amorosa3	0,32
uso_responsable	0,67
genero2	1,71
orientación_sexual2	1,76
edad	0,97
uso_redes1	0,79

Tabla 12: Estimación de los odds ratio del modelo stepwise - víctimas

Los parámetros no significativos al 5%, marcados en rojo, no muestran una asociación estadísticamente relevante con la probabilidad de haber sufrido violencia sexual digital en este modelo.

Las conclusiones exploratorias obtenidas a partir del modelo son las siguientes:

- Las personas que utilizan cuentas anónimas con mayor frecuencia tienen un 58% más de probabilidad de haber sido víctimas de VSD respecto a quienes las utilizan con menor frecuencia.
- Las personas que no han tenido nunca una relación de pareja tienen un 68% menos de probabilidad de haber sido víctima de VSD que las que tienen pareja en la actualidad.
- Las personas que utilizan internet de forma responsable con mayor frecuencia tienen un 33% menos de probabilidad de haber sido sufrido VSD en comparación con quienes lo utilizan con menor frecuencia.
- Las mujeres tienen un 71% más de probabilidad de haber sido víctima de VSD en comparación con los hombres.
- Las personas que pertenecen al colectivo LGTBIQ+ presentan una probabilidad un 76% mayor de haber sufrido VSD respecto a las personas heterosexuales.

Estos resultados reflejan asociaciones estadísticas dentro del conjunto de datos analizado, pero no permiten establecer relaciones de causalidad. En particular, la asociación positiva entre el uso de cuentas anónimas y la victimización resulta contraintuitiva, ya que cabría esperar un efecto protector. Una posible explicación es que el uso de cuentas anónimas pueda ser, en algunos casos, una consecuencia posterior a haber sido víctima, más que una causa previa.

Por tanto, las interpretaciones deben entenderse como hipótesis exploratorias, útiles para identificar posibles factores de riesgo, pero no como evidencias causales.

7.1.2. Modelo LASSO – víctimas

Construcción del modelo

El modelo de regresión logística con penalización LASSO ha seleccionado un total de 8 variables.

Evaluación del modelo

Al igual que con el modelo StepAIC, se ha evaluado el modelo sobre el conjunto de datos de entrenamiento y de prueba y se han obtenido los siguientes resultados:

	Datos de entrenamiento	Datos de prueba
Precisión (tasa de acierto)	0,66	0,60
AUC	0,70	0,65
Indice Kappa	0,31	0,21
Sensibilidad	0,64	0,56
Especificidad	0,68	0,65

Tabla 13: Métricas de evaluación del modelo LASSO - víctimas

Las métricas muestran un rendimiento predictivo limitado, aunque ligeramente mejor que el anterior, especialmente en el conjunto de prueba.

A pesar de no ser adecuado como modelo predictivo, desde un enfoque exploratorio el modelo LASSO resulta muy útil, ya que permite identificar de forma automatizada variables potencialmente asociadas con la probabilidad de haber sido víctima de violencia sexual digital y puede aportar información no capturada en el modelo StepAIC.

Conclusiones exploratorias sacadas del modelo

A diferencia del modelo StepAIC, el modelo Lasso introduce una penalización que sesga los coeficientes hacia valores menores, por lo que la conversión directa a odds ratio puede generar interpretaciones distorsionadas. Por tanto, en este caso se analizan los coeficientes directamente, lo cual permite valorar la importancia relativa de las variables dentro del modelo sin introducir una transformación que podría generar distorsión.

Se muestra, por tanto, el valor de los coeficientes estimados del modelo.

Variable	Coeficientes
genero	0,39
edad	-0,02
orientacion_sexual	0,39
situacion_amorosa	-0,36
uso_redes	-0,16
uso_citas	0,14
cuenta_anonima	0,40
uso_responsable	-0,34

Tabla 14: Estimación de los coeficientes del modelo LASSO - víctimas

Este modelo ha permitido identificar y retener variables que no fueron seleccionadas en el modelo StepAIC, o que no pudieron ser interpretadas al no alcanzar la significación estadística. En concreto, las variables *edad*, *situacion_amorosa* y *uso_redes*, no pudieron ser interpretadas, pero en este nuevo modelo presentan coeficientes con magnitud suficiente como para ser consideradas relevantes en este modelo.

Las conclusiones adicionales que se pueden extraer:

- A mayor edad, la probabilidad de haber sido víctima de VSD se disminuye ligeramente.
- Las personas que no tienen pareja actualmente (ya sea porque nunca han tenido o porque han tenido, pero no en la actualidad) presentan una menor probabilidad de haber sufrido de VSD en comparación con quienes sí tienen pareja.
- Las personas que utilizan las redes sociales de forma frecuente presentan una menor probabilidad de haber sido víctimas de VSD que las que no.
- Las personas que utilizan aplicaciones de citas de forma frecuente presentan una mayor probabilidad de haber sido víctimas de VSD respecto a las que no.

En cuanto a las variables que también aparecían en el modelo StepAIC y resultaron significativas, los coeficientes estimados en el modelo LASSO presentan el mismo signo y sentido interpretativo. Esta coherencia refuerza la solidez de los resultados obtenidos, independientemente del enfoque de modelización utilizado.

7.1.3. Árbol de clasificación – víctimas

Construcción del árbol

Con el objetivo de complementar los resultados obtenidos a través de los modelos de regresión en la identificación de factores asociados a la probabilidad de haber sido víctima de violencia sexual digital, se ha construido un modelo de árbol de clasificación. Este enfoque permite representar de forma jerárquica y visual las variables más relevantes, facilitando la identificación de combinaciones de características asociadas a un mayor riesgo de victimización que podrían no haber sido detectadas en los modelos de regresión logística.

Inicialmente, se ha generado un árbol con un número elevado de hojas con el fin de maximizar su capacidad de ajuste y posteriormente, para evitar el sobreajuste y mejorar la interpretabilidad, se ha aplicado un proceso de poda que elimina ramas poco informativas manteniendo únicamente las divisiones más relevantes.

Elección del número óptimo de hojas

Para determinar el tamaño óptimo del árbol podado, se ha evaluado su rendimiento mediante validación cruzada, utilizando como métricas el área bajo la curva ROC (AUC) y el índice Kappa.

Se muestran los gráficos que ilustran cómo evolucionan ambas métricas en función del número de hojas:

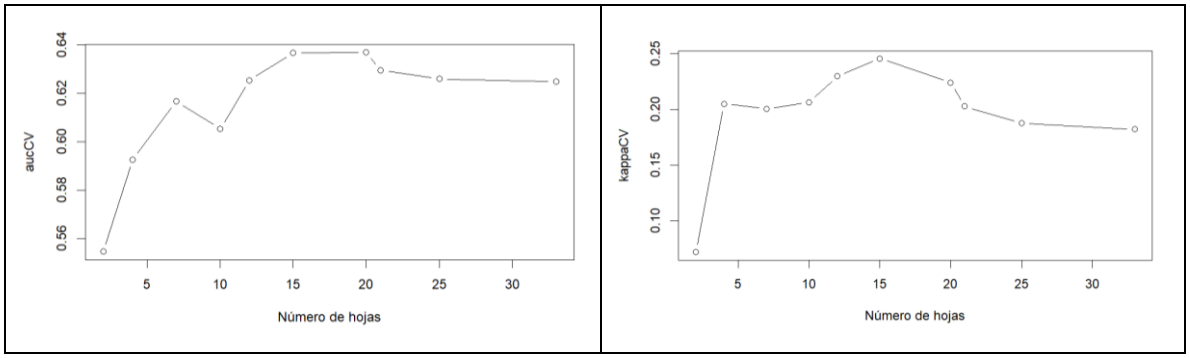


Figura 11: Evolución de las métricas en función del número de hojas - víctimas

Con el fin de obtener equilibrio entre simplicidad y calidad del modelo, se ha seleccionado la versión con 11 hojas, que ofrece un rendimiento adecuado sin introducir complejidad innecesaria, facilitando también la interpretación posterior.

Evaluación del árbol

Al igual que con los modelos de regresión logística, el árbol de clasificación ha sido avaluado sobre el conjunto de entrenamiento y prueba.

	Datos de entrenamiento	Datos de prueba
Precisión (tasa de acierto)	0,72	0,59
AUC	0,73	0,58
Indice Kappa	0,39	0,12
Sensibilidad	0,81	0,68
Especificidad	0,57	0,44

Tabla 15: Métricas de evaluación del árbol de clasificación - víctimas

Los resultados muestran que el modelo presenta un rendimiento considerablemente inferior sobre el conjunto de prueba, lo que sugiere la presencia de cierto sobreajuste. Una vez más, en términos predictivos este modelo no presenta un rendimiento suficiente, con un AUC demasiado próximo a 0’5 (0’59) y un índice Kappa muy reducido (0’12). No obstante, el árbol conserva cierta utilidad exploratoria.

Visualización del árbol

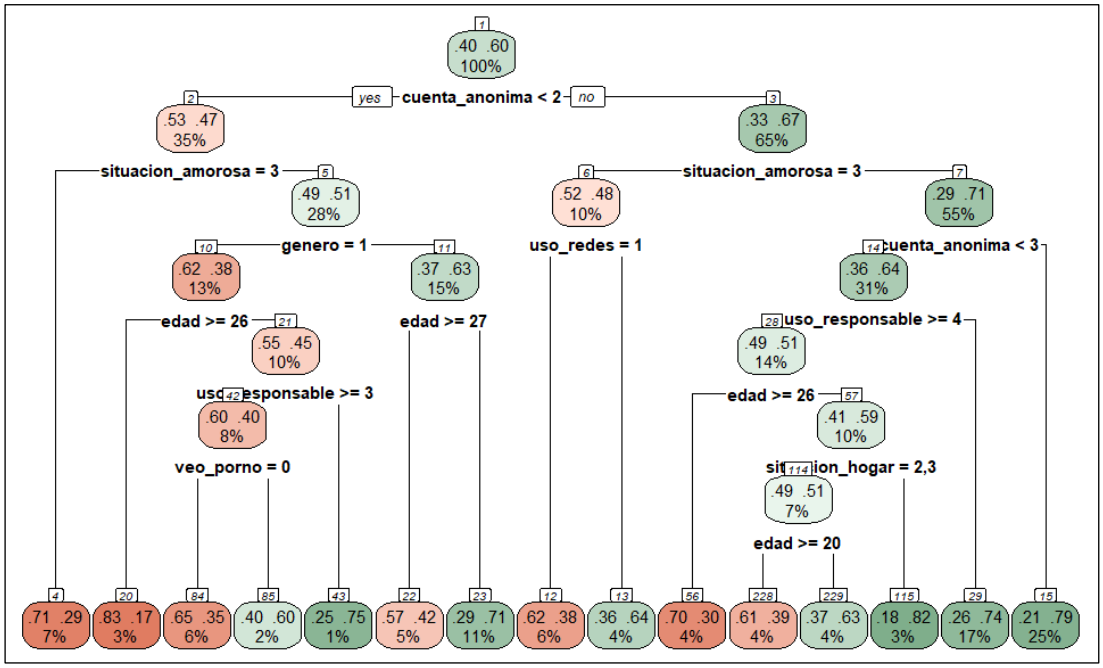


Figura 12: Árbol de clasificación - víctimas

El árbol resultante presenta un total de 11 hojas y permite identificar distintos perfiles asociados a una mayor o menos probabilidad de haber sido víctima de violencia sexual digital.

Aunque el análisis de la importancia relativa de las variables se abordará en el siguiente apartado, ya puede observarse que *cuenta_anonima* aparece como la variable de división principal, seguida de *situacion_amorosa_3* (no haber tenido nunca una pareja estable), lo que destaca su influencia en la construcción del árbol.

Entre los perfiles más relevantes se encuentran:

- El nodo 20 está compuesto por los individuos que presentan una menor probabilidad de haber sufrido violencia sexual digital (17%). Este grupo se caracterizan por estar compuesto por hombres de más de 26 años, que han tenido o tienen pareja estable, y que utilizan cuentas anónimas con poca frecuencia.
- El nodo 115 representa al grupo con mayor probabilidad de ser víctimas de VSD (82%). Incluye a individuos que viven solos, con su pareja e hijos o en otra forma de convivencia no recogida en el cuestionario, que tienen menos de 26 años, que hacen un uso responsable de internet con mucha frecuencia, que tienen o han tenido pareja estable y que utilizan cuentas anónimas pocas veces.
- Los nodos 15 y 29 son los que agrupan al mayor número de individuos del árbol. Ambos se asocian con una probabilidad alta de haber sufrido VSD y ambos incluyen individuos que tienen o han tenido pareja.

Este árbol complementa muy bien los resultados obtenidos en los modelos anteriores.

Importancia de las variables

La siguiente figura muestra la importancia relativa de cada variable en el árbol de clasificación.

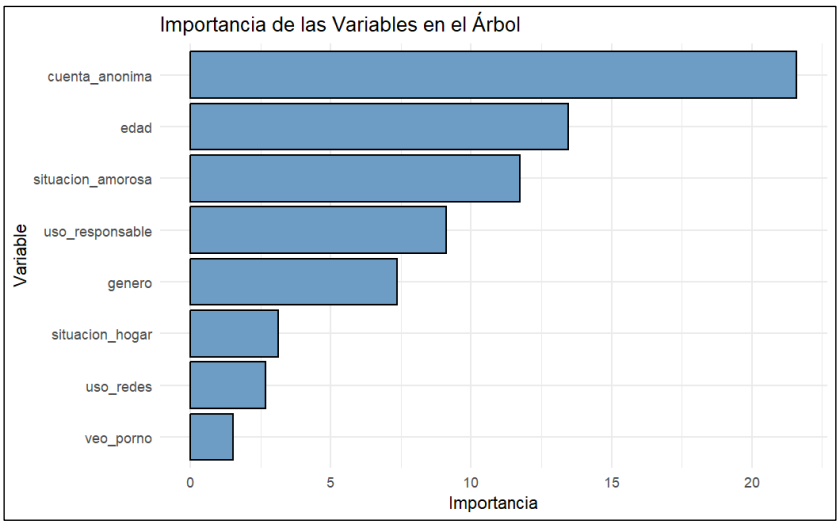


Figura 13: Importancia de las variables en el árbol - víctimas

Tal como se observa, la variable con mayor importancia es *cuenta_anonima*, lo que refuerza su papel central en la clasificación. Le siguen *edad* y *situacion_amorosa*, lo que indica que estas tres variables han sido determinantes en la segmentación jerárquica del conjunto de datos.

Es especialmente relevante destacar que la variable *edad*, aunque no alcanzó significación estadística en el modelo StepAIC, ni fue interpretable en términos de odds ratio, sí ha cobrado una importancia considerable en el árbol. Esto sugiere que el enfoque no lineal del árbol ha permitido detectar patrones complejos en la relación entre la edad y victimización que los modelos logísticos no lograron captar.

Por otro lado, el árbol no incluye algunas variables que habían resultado relevantes en los otros enfoques, como *orientacion_sexual*, que había sido significativa tanto en StepAIC como en el modelo LASSO. Esto puede deberse a que el árbol prioriza divisiones que generan la mayor ganancia informativa global en los primeros niveles, lo que no siempre coincide con la significación estadística individual de las variables.

En conjunto, este análisis complementa los resultados de los modelos anteriores al incorporar nuevas variables relevantes (como *edad*) y al ofrecer una representación jerárquica e interactiva de las combinaciones de factores asociados a la probabilidad de haber sufrido violencia sexual digital.

7.1.4. Modelos de regresión logística binaria – agresores

Construcción de los modelos

En este apartado se exploran los factores asociados al perfil de agresor de violencia sexual digital.

Al igual que en el análisis anterior, se han construido varios modelos de regresión logística binaria. Dado que los modelos obtenidos mediante los métodos stepwise, backward y forward incluyen las mismas variables, se presenta únicamente uno de ellos para evitar redundancias.

A continuación se presentan las variables seleccionadas por cada uno de los modelos:

Variable	manual (Mode1)	StepBIC (Model2)	StepAIC (Model3)
situacion_hogar	×	×	✓
cuenta_anonima	✓	✓	✓
uso_responsable	✓	✓	✓
edad	×	×	✓
uso_redes	×	✓	✓
uso_citas	×	×	✓

Tabla 16:Variables incluidas en cada modelo - agresores

Las variables que no aparecen en la tabla no han sido seleccionadas por ninguno de los métodos aplicados.

Comparación de los modelos

Siguiendo el mismo procedimiento que en el análisis del perfil de víctima, se ha realizado una comparación de los modelos mediante validación cruzada, utilizando como métricas la tasa de acierto, el AUC y el índice Kappa.

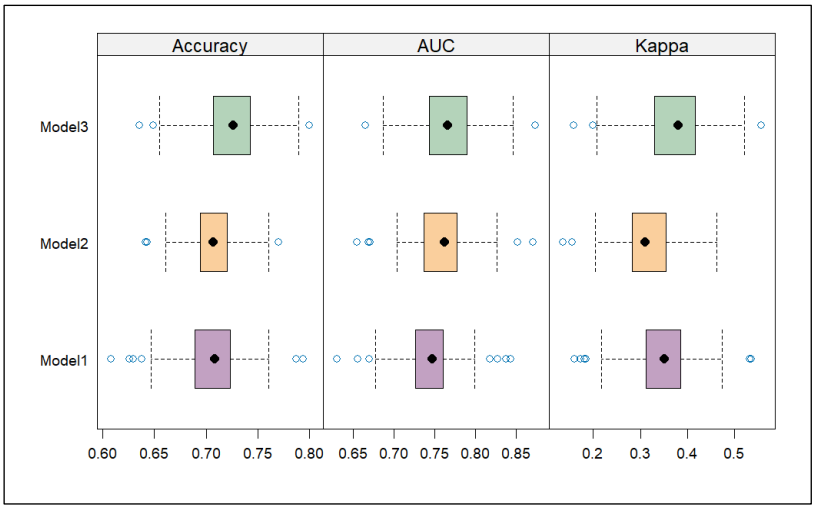


Figura 14: Comparación de modelos - agresores

En este caso, sí que se aprecian diferencias más marcadas entre los modelos, y en general, presentan una capacidad predictiva superior que la obtenida en los modelos para las víctimas.

El tercer modelo (StepAIC) es el que alcanza las métricas más altas y, además, es el que más variables ha retenido. Por tanto, se selecciona este modelo como el mejor de los tres para el análisis, ya que su valor exploratorio será el más completo.

Evaluación del mejor modelo (modelo stepwise guiado por AIC)

El modelo ha sido evaluado utilizando los datos de entrenamiento y prueba, de la misma manera que en el análisis del perfil de víctima, con el objetivo de comprobar la estabilidad de los resultados obtenidos:

	Datos de entrenamiento	Datos de prueba
Precisión (tasa de acierto)	0,72	0,67
AUC	0,78	0,76
Indice Kappa	0,42	0,35
Sensibilidad	0,72	0,75
Especificidad	0,72	0,63

Tabla 17: Métricas de evaluación del modelo stepwise - agresores

El modelo alcanza, en el conjunto de prueba, una precisión del 67%, un AUC de 0'76 y un índice Kappa de 0'35, valores que indican, en general, un rendimiento superior al observado en el modelo del perfil de víctima. Además, las métricas de sensibilidad (0'75) y especificidad (0'63) sugieren que el modelo mantiene un buen equilibrio en la clasificación de agresores y no agresores.

La similitud entre los resultados obtenidos en los conjuntos de entrenamiento y prueba sugiere que el modelo presenta una buena estabilidad.

Aunque el objetivo de este modelo no es predictivo, sus resultados muestran un rendimiento suficientemente robusto como para considerar su posible utilidad en futuros estudios orientados a la predicción. En cualquier caso, y en consonancia con el enfoque exploratorio de este trabajo, los resultados ofrecen una base sólida para identificar factores asociados a la probabilidad de haber ejercido violencia sexual digital.

Conclusiones exploratorias sacadas del modelo

A continuación, se presenta la tabla con los odds ratios estimados para cada variable incluida en el modelo, lo que permite cuantificar la magnitud y el sentido de la relación entre cada predictor y la probabilidad de haber ejercido VSD:

Parámetro	odds ratio
cuenta_anonima	1,83
uso_responsable	0,54
edad	0,96
uso_redes1	0,47
uso_citas1	1,60
situacion_hogar2	0,49
situacion_hogar3	0,58
situacion_hogar4	0,98
situacion_hogar5	0,63

Tabla 18: Estimación de los odds ratio del modelo stepwise - agresores

Los parámetros no significativos al 5%, marcados en rojo, no muestran una asociación estadísticamente relevante con la probabilidad de haber ejercido violencia sexual digital en este modelo.

Las conclusiones exploratorias obtenidas a partir del modelo son las siguientes:

- Las personas que utilizan cuentas anónimas con mayor frecuencia tienen un 83% más de probabilidad de haber ejercido VSD en comparación con las que las utilizan con menor frecuencia.
- Las personas que utilizan internet de forma responsable con mayor frecuencia tienen 46% menos probabilidad de haber ejercido VSD respecto a las que lo utilizan con menor frecuencia.
- Las personas que utilizan redes sociales frecuentemente tienen 53% menos probabilidad de haber ejercido VSD que las que no.
- Las personas que viven con su padre y/o madre tienen una probabilidad un 51% menor de haber ejercido VSD en comparación con las que viven solas.

Una vez más, es importante señalar que estas interpretaciones deben entenderse como hipótesis exploratorias dentro del conjunto de datos analizado, y no como evidencias de causalidad.

7.1.5. Modelo LASSO – agresores

Construcción del modelo

El modelo de regresión logística con penalización LASSO ha seleccionado un total de 8 variables.

Evaluación del modelo

El modelo ha sido evaluado sobre los conjuntos de datos de entrenamiento y de prueba, obteniéndose los siguientes resultados:

	Datos de entrenamiento	Datos de prueba
Precisión (tasa de acierto)	0,72	0,72
AUC	0,77	0,78
Indice Kappa	0,42	0,43
Sensibilidad	0,70	0,77
Especificidad	0,74	0,70

Tabla 19: Métricas de evaluación del modelo LASSO - agresores

Los resultados muestran una alta estabilidad del modelo, con valores muy similares entre el conjunto de entrenamiento y el de prueba.

En el conjunto de prueba, el modelo alcanza una precisión del 72%, un AUC de 0’78 y un índice Kappa de 0’43, lo que supone una mejora respecto al modelo StepAIC. Las métricas de sensibilidad (0’77) y especificidad (0’70) indican que el modelo mantiene un equilibrio adecuado entre la detección de agresores y no agresores.

Si bien el modelo anterior ya mostraba un rendimiento notable y podría considerarse útil para aplicaciones predictivas futuras, el modelo LASSO mejora ligeramente dichos resultados, por lo que ofrece una base aún más sólida para investigaciones orientadas a la predicción.

Conclusiones exploratorias sacadas del modelo

Siguiendo la misma lógica que en el modelo LASSO para el perfil de víctimas, en este caso también se analizan directamente los coeficientes sin transformarlos a odds ratios, debido a la penalización impuesta por el método LASSO.

Variable	Coeficientes
genero	-0,16
edad	-0,004
orientación_sexual	-0,008
situacion_amorosa	-0,07
uso_redes	-0,67
uso_citas	0,28
cuenta_anonima	0,57
uso_responsable	-0,57

Tabla 20: Estimación de los coeficientes del modelo LASSO - agresores

En comparación con el modelo StepAIC, el modelo LASSO ha permitido identificar algunas variables adicionales que no fueron seleccionadas anteriormente o que no resultaron significativas, lo cual aporta información complementaria sobre las características asociadas a la probabilidad de haber ejercido VSD.

No obstante, dado que los coeficientes asociados a las variables *edad* y *orientacion_sexual* son extremadamente bajos, no se van a considerar relevantes en el análisis interpretativo ya que su efecto es prácticamente nulo.

Las conclusiones adicionales que se recogen en este modelo son las siguientes:

- Las mujeres presentan una menor probabilidad de haber ejercido VSD en comparación con los hombres.
- Las personas que no tienen pareja actualmente (ya sea porque nunca han tenido o porque han tenido, pero no en la actualidad) presentan una probabilidad ligeramente menor de haber ejercido VSD en comparación con quienes sí tienen pareja.
- Las personas que utilizan aplicaciones de citas de forma frecuente tienen una mayor probabilidad de haber ejercido VSD que las que no.

Las variables que ya aparecían en el modelo StepAIC y fueron interpretables, y que además han sido seleccionadas en el modelo LASSO, presentan coeficientes con el mismo signo y sentido interpretativo, lo que refuerza la coherencia y solidez de los resultados obtenidos a través de ambos enfoques.

7.1.6. Árbol de clasificación – agresores

Construcción del árbol

Siguiendo el mismo enfoque utilizado en el análisis del perfil de víctima, se ha construido un modelo de árbol de clasificación para complementar los resultados obtenidos, mediante regresión logística.

Como en el análisis del perfil de víctima, se partió de un árbol inicial muy extenso, que posteriormente fue podado para reducir la complejidad y mejorar la interpretación, manteniendo únicamente las divisiones más relevantes.

Elección del número óptimo de hojas

Para determinar el tamaño óptimo del árbol podado, se ha evaluado su rendimiento mediante validación cruzada, utilizando como métricas el área bajo la curva ROC (AUC) y el índice Kappa.

En los siguientes gráficos se muestra cómo evolucionan el AUC y el índice Kappa en función del número de hojas:

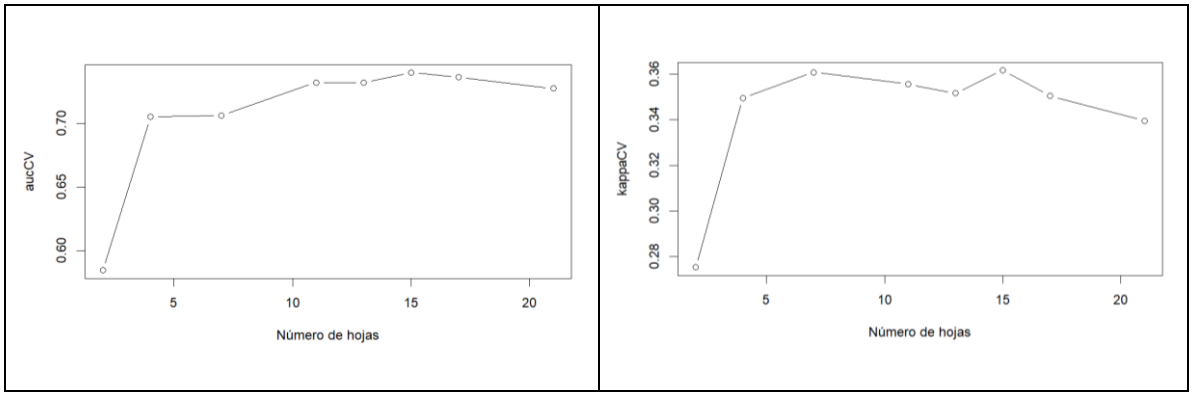


Figura 15: Evolución de las métricas en función del número de hojas - agresores

En este caso, se ha priorizado el equilibrio entre la simplicidad interpretativa y la calidad del modelo, por ello se han seleccionado 12 hojas.

Evaluación del árbol

Al igual que con los modelos de regresión logística, el árbol de clasificación ha sido avaluado sobre el conjunto de entrenamiento y prueba.

	Datos de entrenamiento	Datos de prueba
Precisión (tasa de acierto)	0,77	0,66
AUC	0,79	0,69
Indice Kappa	0,48	0,22
Sensibilidad	0,57	0,43
Especificidad	0,88	0,78

Tabla 21: Métricas de evaluación del árbol de clasificación - agresores

En el conjunto de prueba, el modelo alcanza una precisión del 66%, un AUC de 0’69 y un índice Kappa de 0’22. Aunque estas métricas reflejan un rendimiento inferior al observado en el modelo LASSO, el árbol conserva valor exploratorio.

Visualización del árbol

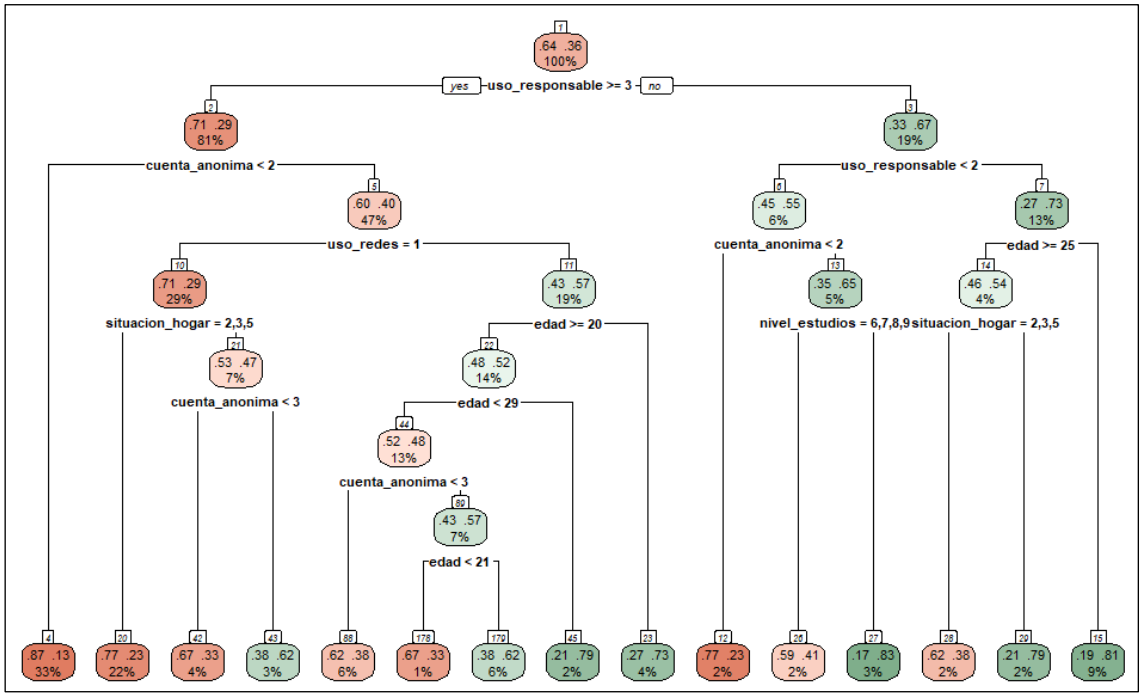


Figura 16: Árbol de clasificación - agresores

El árbol resultante presenta un total de 12 hojas y permite identificar distintos perfiles asociados a una mayor o menos probabilidad de haber ejercido violencia sexual digital.

En este caso, la variable de división principal es la de *uso_responsable*, lo que es la que más influye en la construcción del árbol.

Entre los perfiles más relevantes se encuentran:

- El nodo 4 está compuesto por los individuos que presentan una menor probabilidad de haber ejercido VSD (13%). Este grupo incluye a personas que no utilizan cuentas anónimas nunca y utilizan internet de forma responsable con frecuencia. Este nodo además es el que agrupa un número mayor de individuos dentro del árbol.
- El nodo 27 representa al grupo con mayor probabilidad de haber ejercido VSD (83%). Este grupo se caracteriza por estar compuesto por individuos que no han llegado a completar el Bachillerato, que utilizan, aunque sea ocasionalmente, cuentas anónimas y que nunca utilizan internet de forma responsable.
- El nodo 15 es el segundo grupo con mayor probabilidad de haber ejercido VSD y, además, está ligeramente mejor representado que el nodo 27. En este nodo se agrupan individuos que tienen menos de 25 años y utilizan internet de forma responsable solo de vez en cuando.

Importancia de las variables

La siguiente figura muestra la importancia relativa de cada variable en el árbol de clasificación.

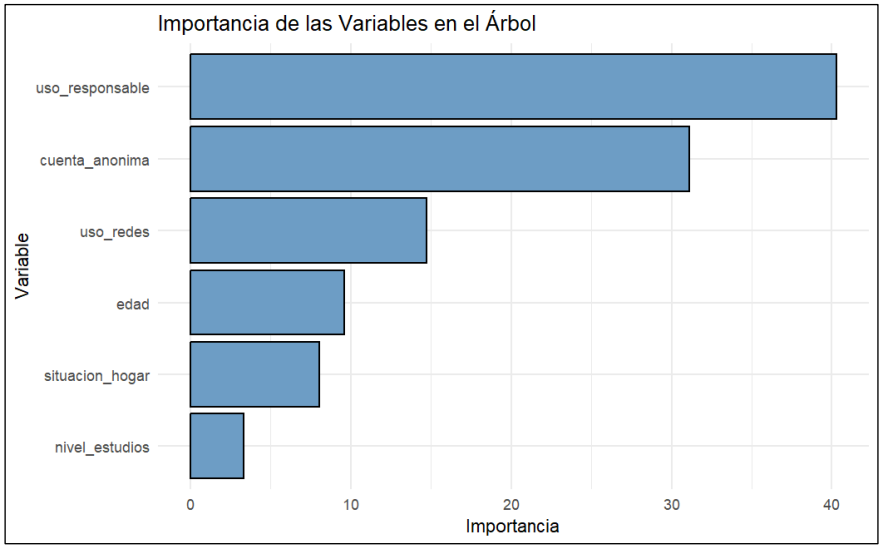


Figura 17: Importancia de las variables en el árbol - agresores

La figura muestra que, como ya se intuyó en la visualización del árbol, la variable más influyente es *uso_responsable*, seguida de *cuenta_anonima*.

El modelo ha incluido variables como *edad* y *nivel_estudios*, que, si bien no presentan la mayor importancia relativa, aportan información adicional que no pudo ser extraída en los modelos de regresión logística, donde no alcanzaron significación o fueron descartadas en el proceso de selección.

Por el contrario, variables como *genero* o *uso_citas*, que sí resultaron relevantes en el modelo StepAIC, no han sido seleccionadas en este modelo. Esto refleja una vez más, cómo distintos enfoques estadísticos pueden priorizar diferentes dimensiones del fenómeno.

7.2. Patrones de experiencia e impacto en víctimas de VSD

En este apartado se lleva a cabo un análisis descriptivo y multivariante, complementado con técnicas de clustering no supervisado, con el fin de identificar patrones entre las personas que han sido víctimas de VSD. El objetivo de este apartado es visualizar la relación entre sus experiencias y las consecuencias emocionales o psicológicas reportadas, para así obtener una visión más profunda del impacto de este tipo de violencia en la juventud.

7.2.1. Clustering con PAM

Se ha escogido las variables relacionadas con los síntomas experimentados tras haber sufrido VSD y la variable *sufrido_contado*, para hacer la agrupación de individuos en clústeres con características similares.

Variables utilizadas para hacer la agrupación	sintomas_ansiedad sintomas_depresivos sintomas_emocionales sintomas_nada sufrido_contado

Tabla 22: Variables utilizadas para el clustering

Elección del número óptimo de clústeres

Para determinar el número óptimo de clústeres antes de aplicar el método de clustering correspondiente, se ha utilizado el coeficiente de silhouette, una métrica que evalúa la calidad de la agrupación midiendo tanto la cohesión interna dentro de los clústeres como la separación entre ellos.

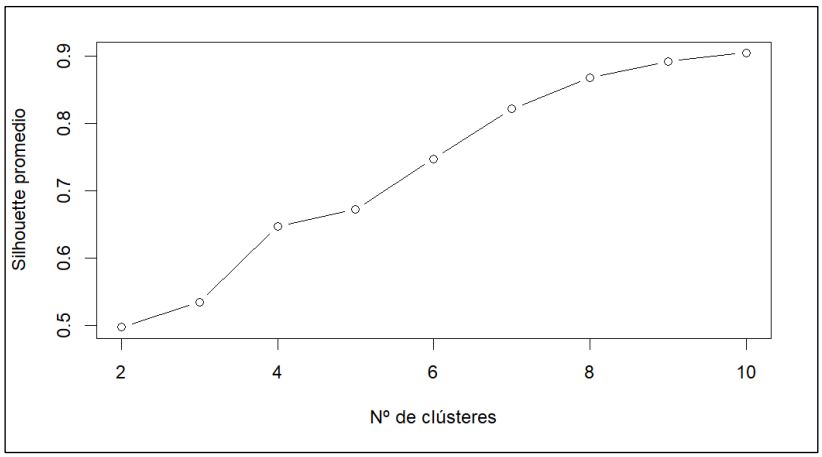


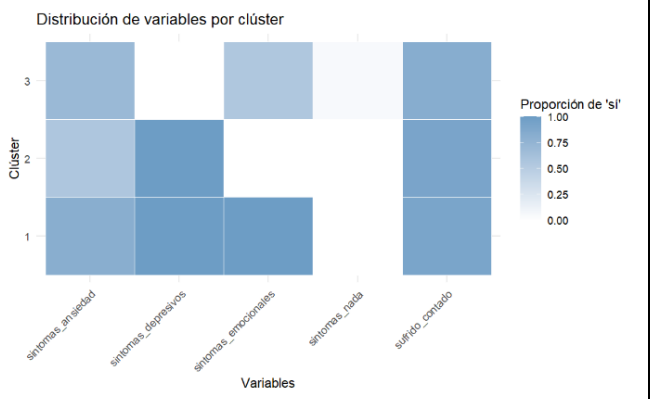
Figura 18: Evolución del coeficiente de silhouette en función del número de clústeres

Aunque el valor de silhouette promedio aumenta progresivamente con el número de clústeres, con el fin de facilitar una interpretación clara y manejable de los resultados, se ha optado por trabajar con tres clústeres, priorizando un equilibrio entre calidad de agrupación y simplicidad analítica.

Aplicación del método PAM para la identificación de los distintos perfiles

Para realizar la agrupación se ha empleado el método PAM (Partitioning Around Medoids), ya que, a diferencia de otros métodos como k-means, no depende de la media como centro del clúster, sino de observaciones reales (medoides). Esto lo hace menos sensible a valores atípicos y permite trabajar de forma adecuada con variables categóricas o mixtas, por lo que resulta más apropiado para este análisis.

A continuación, se muestra la distribución de variables por clúster:



	clúster 1	clúster 2	clúster 3
sintomas_ ansiedad	0,82	0,57	0,71
sintomas_ depresivos	1	1	0
sintomas_ emocionales	1	0	0,56
sintomas_ nada	0	0	0,058
sufrido_ contado	0,92	0,92	0,83

Figura 19: Distribución de variables por clúster

A partir de los patrones observados en el gráfico, podemos distinguir los clústeres, identificando tres perfiles diferenciados:

- El clúster 1 agrupa a personas que presentan síntomas de los tres tipos (ansiedad, depresión y emocionales) y que han contado su experiencia. Este grupo podría describirse como un perfil psicológica y emocionalmente afectado, con disposición a contarlo.
- El clúster 2 agrupa a personas que presentas síntomas de ansiedad y depresivos, pero no emocionales y que han contado su experiencia. Este segundo grupo podría describirse como un perfil psicológicamente afectado, pero emocionalmente contenido y con disposición a contarlo.
- El clúster 3 agrupa a personas que presentan menos síntomas en general, y con una menor disposición a contarlo. Este grupo podría describirse como un perfil parcialmente afectado y más reservado.

7.2.2. Análisis de Correspondencias Múltiples (ACM)

Tras haber identificado los perfiles, se ha llevado a cabo un análisis de correspondencias múltiples para explorar visualmente la relación entre las variables utilizadas para la agrupación.

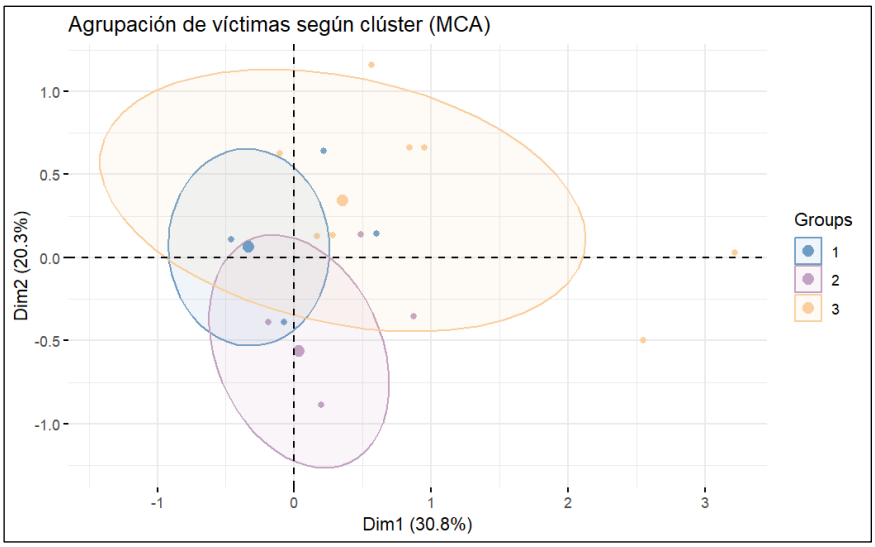


Figura 20: Agrupación de víctimas por clúster (ACM)

El gráfico muestra que las dos primeras dimensiones explican en conjunto un 51’1% de la variabilidad total, lo cual proporciona una representación suficientemente clara de las relaciones entre las categorías de las variables analizadas.

Además, se observa que los tres clústeres identificados tienden a agruparse de forma diferenciada. Los clústeres 1 y 2 aparecen relativamente próximos entre sí, aunque con centros bien diferenciados, lo cual resulta coherente, ya que ambos agrupan a individuos afectados psicológicamente y con disposición a contar su experiencia, pero se distinguen por la presencia o ausencia de síntomas emocionales. Por otro lado, el clúster 3 aparece claramente separado de los otros dos, lo que sugiere que corresponde a un perfil más distanciado en términos de sintomatología y de disposición a compartir la experiencia.

En conjunto, los resultados del clustering y del análisis de correspondencias múltiples permiten concluir que los perfiles de víctimas de violencia sexual digital presentan una consistencia interna y una diferenciación clara tanto en síntomas experimentados como en la manera de gestionarlo (contarlo o no).

7.2.3. Relación entre los perfiles identificados y variables de interés

Una vez diferenciados los perfiles de víctimas en función de las consecuencias experimentadas y evaluado su consistencia, se analiza cómo se relacionan con otras variables relevantes del estudio. Esto se hace con el objetivo de comprobar si existen diferencias significativas en el impacto y la forma de afrontar la VSD, en función de distintas características incluidas en el estudio.

A continuación, se presentan únicamente aquellos resultados que han mostrado una asociación estadísticamente significativa con los clústeres, según el contraste de chi - cuadrado, para poder extraer conclusiones reales acerca de factores que intervienen en la manera de afrontar la VSD.

Género por clúster

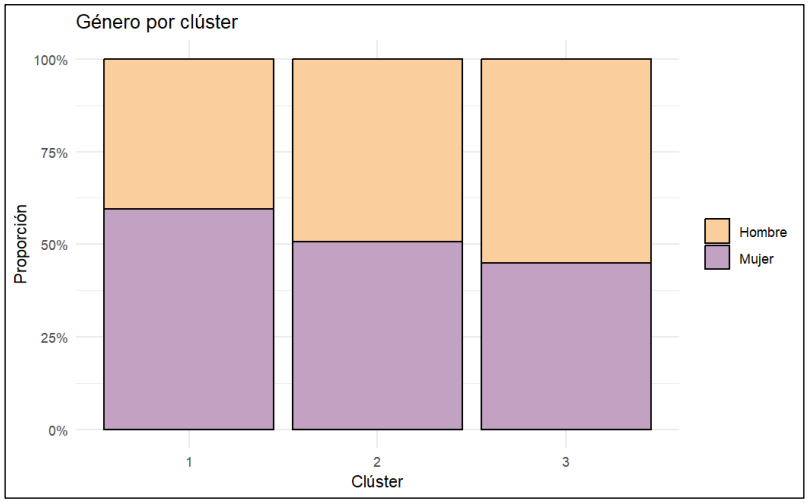


Figura 21: Distribución de género por clúster

Se ha detectado que la distribución de los hombres y mujeres no es homogénea entre los distintos perfiles.

- El clúster 1 presenta una mayor proporción de mujeres que de hombres.
- El clúster 2 presenta una distribución más homogénea.
- El clúster 3 presenta una mayor proporción de hombres que de mujeres.

Estos resultados sugieren que el género es un factor diferencial en la forma de afrontar la experiencia de violencia sexual digital sufrida.

Tipo de violencia sexual digital sufrida por clúster

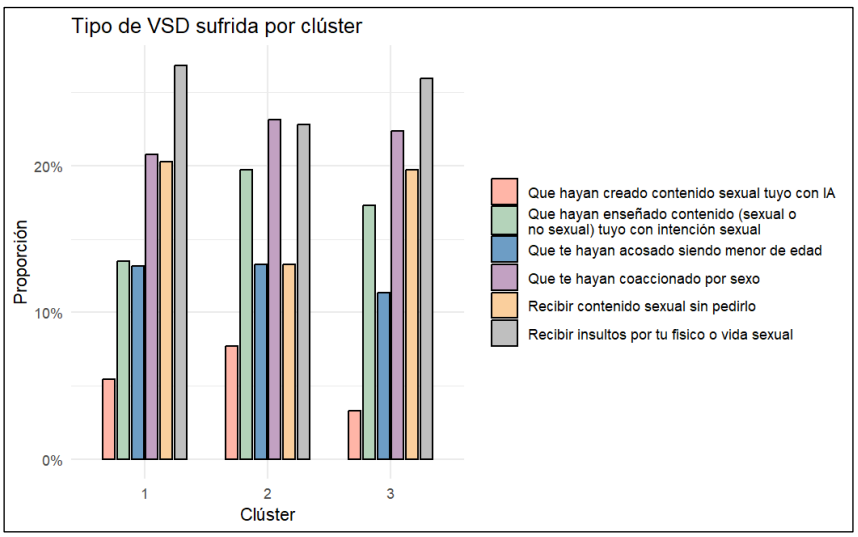


Figura 22: Distribución de tipo de VSD sufrida por clúster

Aunque las proporciones observadas entre clústeres son en general similares, se aprecian algunas diferencias relevantes en el impacto experimentado según el tipo de violencia sexual digital sufrida.

- El clúster 1 presenta una proporción más alta, en comparación con los otros clústeres, de personas que han recibido insultos por su físico o vida sexual, así como contenido sexual sin haberlo solicitado.
- El clúster 2 concentra las proporciones más elevadas de personas que han sido coaccionadas sexualmente, han sufrido la difusión no consentida de contenido con intención sexual, o han sufrido la creación de contenido sexual con IA.
- El clúster 3 no destaca en ninguna forma concreta de VSD, pero mantiene niveles intermedios en varios tipos, lo que sugiere una vivencia más dispersa o menos centrada en una agresión específica.

En conjunto, estos resultados indican que el tipo de violencia sufrida puede influir en el impacto y la forma de afrontar la violencia sufrida, no todos los tipos tienen el mismo impacto ni afectan del mismo modo a todas las víctimas.

Percepción de evolución de violencia sexual digital en los últimos 10 años por clúster

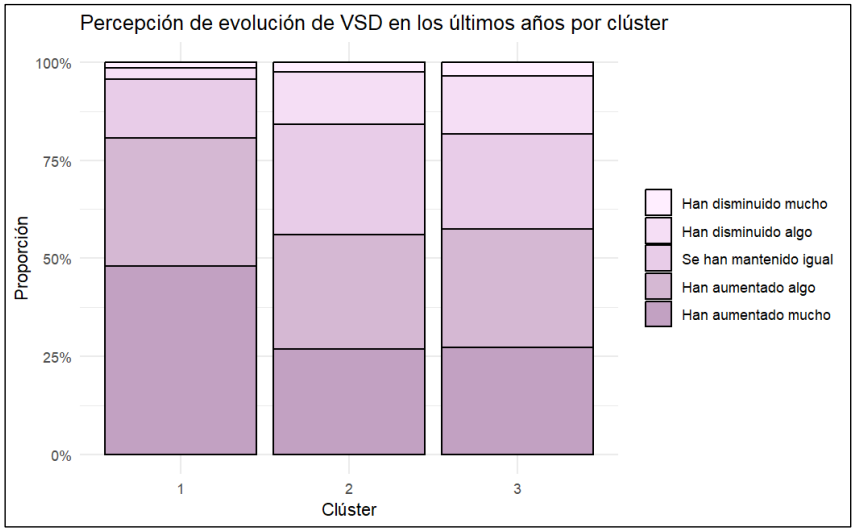


Figura 23: Distribución de percepción de evolución en los últimos años por clúster

El clúster 1 presenta la mayor proporción de personas que perciben que la VSD ha aumentado notablemente en los últimos años. El clúster 2 muestra una distribución más repartida, con cierta inclinación a pensar que la situación se ha mantenido igual. Por su parte, el clúster 3 contiene la proporción más alta de personas que creen que esta forma de violencia ha disminuido.

Estos resultados sugieren que la forma en que se vive y se afronta la VSD influye directamente en cómo se percibe su evolución, revelando un vínculo entre la experiencia personal y la percepción social del problema.

Percepción de evolución de violencia sexual digital en los próximos 10 años por clúster

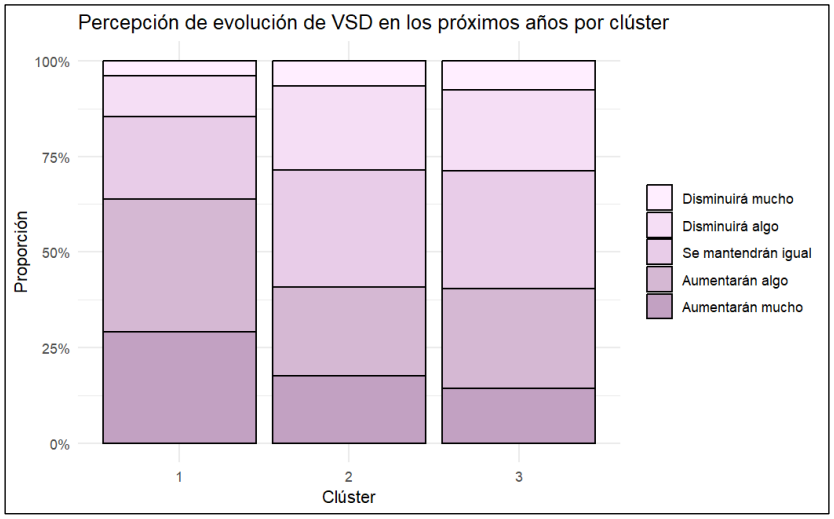


Figura 24: Distribución de percepción de evolución en los próximos años por clúster

En cuanto a la percepción de la evolución que experimentará la VSD en los próximos 10 años también se detecta una diferencia según el impacto y afrontamiento posterior a la experiencia sufrida. El clúster 1 es el grupo menos optimista, mostrando una mayor tendencia a creer que este tipo de violencia aumentará significativamente en los próximos años. En contraste, los clústeres 2 y 3, en especial el clúster 3, presentan una percepción algo más optimista, con mayor proporción de respuestas en las categorías de que la VSD se mantendrá igual o disminuirá.

Estos resultados refuerzan la idea de que la vivencia personal de la violencia y su impacto posterior influyen en la perspectiva futura que tienen las víctimas. En general se observa como la mayoría de las víctimas tienen una visión poco esperanzadora sobre la mejora de este problema, lo que reafirma la necesidad de diseñar intervenciones eficaces tanto para prevenir como para reparar.

8. Conclusiones

Este trabajo ha abordado la violencia sexual digital (VSD) en jóvenes desde una perspectiva estadística, combinando técnicas descriptivas, modelos de regresión, árboles de clasificación y análisis no supervisados, con el objetivo de identificar factores asociados a haber sido víctima o agresor, así como de explorar el impacto emocional de este fenómeno.

8.1. ¿Con qué se relaciona haber sufrido violencia sexual digital?

El análisis exploratorio realizado ha permitido identificar diversos factores asociados a una mayor probabilidad de haber sido víctima de violencia sexual digital. De manera destacada, el género y la orientación sexual aparecen como variables clave: ser mujer o formar parte del colectivo LGTBIQ+ incrementa significativamente la probabilidad de haber sufrido este tipo de violencia. Este hallazgo refuerza la idea de que las desigualdades estructurales presentes en otros ámbitos sociales también se manifiestan en los entornos digitales.

Asimismo, la situación amorosa muestra una relación relevante con la victimización. Las personas que nunca han tenido una pareja estable presentan una menor probabilidad de haber sido víctimas, lo que sugiere que ciertas formas de violencia digital se producen con mayor frecuencia en contextos íntimos o de confianza, como las relaciones de pareja.

Un resultado particularmente llamativo es la asociación entre el uso frecuente de cuentas anónimas y la probabilidad de haber sufrido VSD. Aunque podría interpretarse como un factor de riesgo, es posible que esta conducta refleje una estrategia de autoprotección adoptada tras haber vivido una situación de violencia, es decir, como una forma de resguardarse frente a nuevas exposiciones.

Por otra parte, el uso responsable de Internet se presenta como un factor protector claro, asociado con una menor probabilidad de victimización.

La edad también parece desempeñar un papel importante, ya que a medida que esta aumenta, la probabilidad de haber sido víctima disminuye ligeramente. Esto podría vincularse con una mayor madurez o experiencia en la gestión de riesgos digitales. Finalmente, ciertas prácticas tecnológicas, como el uso habitual de redes sociales o de aplicaciones de citas, también influyen, aunque su efecto es más moderado en comparación con otras variables.

Aunque no puede establecerse causalidad, los resultados ofrecen una base útil para orientar la prevención e investigar más a fondo este fenómeno. Comprender estos factores también ayuda a visibilizar desigualdades que siguen presentes, incluso en lo digital.

8.2. ¿Con qué se relaciona haber ejercido violencia sexual digital?

El perfil de quienes ejercen violencia sexual digital también presenta patrones reconocibles, aunque no se reduce a un único modelo. Uno de los factores más relevantes es el uso frecuente de cuentas anónimas, que, al desinhibir y reducir el temor a consecuencias, facilita comportamientos que difícilmente se darían de forma presencial. Esto no implica culpar al anonimato en sí, sino cuestionar el contexto de impunidad que a veces lo rodea.

Otro factor clave es la actitud hacia el uso de internet: quienes lo utilizan de forma responsable tienen una probabilidad considerablemente menor de ejercer este tipo de violencia. En cambio, el uso habitual de aplicaciones de citas se asocia con una mayor probabilidad de ser agresor, posiblemente porque estos espacios favorecen dinámicas rápidas y despersonalizadas, donde se diluyen los límites del consentimiento. No obstante, lo decisivo no es tanto su uso como la forma en que se usan.

De forma llamativa, el uso frecuente de redes sociales se relaciona con una menor probabilidad de ejercer violencia, un resultado inesperado dado su protagonismo en la vida digital.

También se observó que ser mujer se asocia con una menor probabilidad de ejercer VSD, lo que refuerza la consistencia del género como factor relevante a lo largo del análisis.

Por último, es interesante que vivir con los padres o madres actúe como un factor protector. Esta variable, que puede parecer menor, en realidad señala la importancia del entorno familiar y social en la construcción de referentes éticos, límites y cuidados.

En conjunto, estos hallazgos apuntan a la importancia de trabajar no solo sobre los comportamientos concretos, sino también sobre el contexto, las actitudes y los entornos que favorecen o frenan la aparición de la VSD. Solo a través de la comprensión de estos contextos podremos actuar sobre las raíces del problema y no solo sobre sus manifestaciones.

8.3. ¿Existen patrones en el impacto de la VSD en jóvenes?

Más allá de entender quiénes son las víctimas o los agresores, este trabajo ha permitido explorar cómo impacta la violencia sexual digital en quienes la sufren. Y, lo que es más importante aún, cómo se procesa y afronta esa experiencia.

Gracias al análisis no supervisado, se han identificado tres perfiles diferenciados entre las personas que han sufrido VSD, en función de los síntomas experimentados y la disposición a compartir la experiencia:

1. Perfil altamente afectado y comunicativo: con síntomas de ansiedad, depresión y emocionales, y con tendencia a contar lo vivido.
2. Perfil psicológicamente afectado, pero emocionalmente contenido: con síntomas de ansiedad y depresión, pero sin síntomas emocionales, también con disposición a hablar de su experiencia.
3. Perfil parcialmente afectado y reservado: con escasa sintomatología y menor disposición a contar lo sucedido.

En el primer grupo predominan las mujeres; en el segundo, hay una distribución equitativa entre hombres y mujeres; y en el tercero predominan claramente los hombres. Esto refuerza la idea de que el impacto emocional de la violencia sexual digital no es neutral al género. Las mujeres no solo tienden a manifestar un mayor grado de afectación, sino que también muestran una mayor disposición a compartir lo vivido. En cambio, los hombres aparecen como más reservados, lo cual no implica necesariamente que la violencia les haya afectado menos, sino que es posible que hayan reprimido, minimizado o naturalizado su experiencia, en parte por las normas sociales que desincentivan la expresión emocional en ellos.

También se han detectado diferencias según el tipo de violencia sufrida: por ejemplo, el perfil altamente afectado suele haber sido víctima de insultos por su físico o su vida sexual, así como del envío no consentido de contenido sexual. Por otro lado, el perfil emocionalmente contenido presenta una mayor proporción de víctimas de coacción sexual o difusión de contenido íntimo, mientras que el perfil más reservado muestra una distribución más dispersa y menos concentrada en un tipo específico de agresión. Esto sugiere que el tipo de violencia sufrida también influye en el nivel de impacto emocional y en la forma de afrontarlo.

Estos tres grupos difieren además en cómo perciben la evolución del problema en la sociedad. Cuanto mayor ha sido el impacto vivido, más pesimista es la visión sobre el presente y el futuro de la VSD. Esto tiene sentido ya que quienes han sufrido más tienden a reconocer con mayor claridad que este tipo de violencia no solo persiste, sino que adopta formas sutiles, complejas y muchas veces invisibles para quienes no las han experimentado.

En conclusión, no todas las víctimas se sienten igual, ni todas lo afrontan del mismo modo. Entender estas diferencias es clave si se quiere diseñar intervenciones eficaces, porque no basta con señalar que existe VSD sino que también hay que atender a cómo duele, cómo se vive, cómo se calla y cómo, a veces, se transforma en conciencia. Y es ahí donde la estadística deja de ser solo números, y se convierte en una herramienta para escuchar.

9. Bibliografía

1. Agencia de Gobierno Electrónico y Sociedad de la Información y el Conocimiento (AGESIC). (s.f.). *Red delante de las pantallas: acompañar y sostener a niñas, niños y adolescentes*. Gobierno del Uruguay. https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/comunicacion/publicaciones/red-delante-pantallas-acompanar-sostener-ninas-ninos-adolescentes-8
2. R-universe. (s.f.). *missForest manual*. https://cran.r-universe.dev/missForest/doc/manual.html
3. Daily Dose of Data Science. (2023, noviembre 17). *MissForest: A better alternative to mean/median imputation*. https://blog.dailydoseofds.com/p/missforest-a-better-alternative-to
4. Calviño, A. (2024). *Ejemplo regresión logística binaria* [Apuntes de clase, Curso Técnicas de Segmentación y Tratamiento de Encuestas]. Universidad Complutense de Madrid.
5. IBM. (s.f.). *Lasso regression*. https://www.ibm.com/mx-es/think/topics/lasso-regression
6. Calviño, A. (2024). *Tema 3: Árboles de clasificación y regresión y bosques aleatorios* [Apuntes de clase, Curso Técnicas de Segmentación y Tratamiento de Encuestas]. Universidad Complutense de Madrid.
7. Universidad Icesi. (s.f.). *Método de partición alrededor de los medoides (PAM)*. https://www.icesi.edu.co/editorial/intro-clustering-web/PAM.html
8. Ciencia de Datos. (s.f.). *Análisis de Componentes Principales (PCA)*. https://cienciadedatos.net/documentos/35_principal_component_analysis