

Metadatos

Nombre del dataset (si tiene nombre, si no asignarle uno):

Macroencuesta de Violencia contra la Mujer 2019 (MVM)

Fuente (incluir link a la fuente si está disponible y describirla):

https://violenciagenero.igualdad.gob.es/violenciaEnCifras/macroencuesta2015/pdf/Macroencuesta_2019_estudio_investigacion.pdf

Descripción general del dataset

La MVM es un documento disponible en formato PDF que contiene información realizada a una muestra de 9.568 mujeres, según nuestros análisis que pueden observarse en las notebooks y la documentación de la encuesta, la misma es representativa de la población femenina residente en España de personas mayores a 16 años.

Los tipos de violencia de género medidos en la encuesta son:

- Violencia de la pareja actual y violencia de parejas pasadas
 - Física
 - Sexual
 - Psicológica: comportamientos de control y violencia emocional
 - Económica
- Violencia fuera de la pareja
 - Física
 - Sexual
- Acoso sexual
- Acoso reiterado (stalking)

La MVM diferencia entre relaciones actuales, relaciones con ex parejas y violencias o acoso recibido por personas que no son pareja. A su vez, diferencia violencias actuales de violencias sucedidas durante el último año y de aquellas sucedidas durante los últimos 4 años. También hace hincapié en los y las menores de edad para comprender si también fueron víctimas o se vieron afectados por las situaciones de violencia.

Cuenta con 1121 preguntas, algunas de las cuales tienen saltos en función de la respuesta de la persona entrevistada. A su vez, cada pregunta tiene un código específico que luego será el nombre de las variables en el set de datos.

En este link están los códigos de las preguntas así como sus posibles respuestas:

https://drive.google.com/file/d/164Pvkly8ysFdwENgPmEURBxyXLbYuoN_/view?usp=sharing

Cabe aclarar que esta encuesta no tiene ningún tipo de pregunta abierta, son preguntas cerradas con respuestas limitadas y estandarizadas para que sea más fácil su análisis.

El set de datos cuenta con 9.568 filas, número que coincide con la cantidad de mujeres entrevistadas. A su vez, cuenta con 1.121 columnas que corresponden a las preguntas de la encuesta.

La gran mayoría de las variables son categóricas, es decir que toman valores limitados y fijos, ya que las respuestas están estandarizadas.

Hay pocas variables numéricas, la mayoría son strings, ya que para no dejar opción libre, se decidió que muchos valores numéricos se transformen en rangos. Por ejemplo: salario "de 0 a 500 euros"; "de 501 euros a 900"; etc.

Autoras del data statement: Susana Araujo, Ivana Feldfeber, Tamara Maggioni

Otros que contribuyeron a este documento: Gabriel Moyano, mentor

Versión (si está disponible) y fecha aproximada de creación: la MVM se realizó en el año 2019 y se publicó en el año 2020. Sólo hay una versión disponible.

Licencia bajo la que se distribuye el dataset: la MVM es de libre acceso y descarga, no contamos con datos sobre sus licencias.

Resumen del data statement (largo mínimo 50 palabras y máximo 70 palabras):

El conjunto de datos de Macroencuesta de Violencia contra la Mujer 2019 (MVM) es un conjunto de datos que recopila las respuestas de 9.568 mujeres a 1121 preguntas, estas abarcan desde información demográfica de la persona entrevistada a diversas situaciones donde puede haber existido violencia de género en sus distintas formas. El mismo recomendamos solamente emplear para obtener información estadística y descriptiva, pero no para realizar predicciones.

1. Motivación para la creación del dataset

Las preguntas de esta sección están destinadas principalmente a alentar a los creadores de datasets a articular claramente sus razones para crear el conjunto de datos y promover la transparencia sobre los intereses de financiación.

1.1 ¿Con qué propósito se creó el conjunto de datos? ¿Se tenía una tarea específica en mente? ¿Conoce otros datasets parecidos, y, en ese caso, cuáles son las diferencias? Proporcione una descripción.

De acuerdo al sitio web, la MVM se creó con el objetivo de “proporcionar una visión detallada y actual de la extensión y formas de violencia contra la mujer, así como de diversos aspectos relacionados con las mismas”. De forma más específica, los principales objetivos de la macroencuesta son:

1. Estimar la prevalencia de la violencia contra la mujer en España, tanto la que se produce en el ámbito de la pareja actual o de parejas pasadas como la sufrida por las mujeres fuera del ámbito de la pareja, distinguiendo entre violencia física, sexual, psicológica y económica en el caso de la violencia en la pareja, y entre violencia física y sexual en el caso de la violencia ejercida por terceras personas. También se investiga el acoso sexual y el stalking.
2. Conocer los tipos de agresores más frecuentes, obtener información sobre la denuncia, la búsqueda de ayuda formal e informal, la satisfacción con la ayuda recibida, recabar información relacionada con hijos e hijas de las víctimas,

especialmente los menores, en relación a su presencia durante los episodios de violencia, y conocer las consecuencias de la violencia sobre la salud física y mental de las mujeres afectadas.

En cuanto a otras encuestas similares, en Argentina, se realizó dos veces (2015 y 2018) la Encuesta Nacional de Violencia contra las mujeres (<https://www.argentina.gob.ar/justicia/politicacriminal/estadisticas/encuesta-violencia-mujeres>), la cual está basada en la Encuesta Internacional de Violencias contra las Mujeres impulsada por la ONU. La misma tiene 829 preguntas en vez de las 1.121 de la MVM, aunque incluye una cantidad mucho menor de mujeres encuestadas (aproximadamente una décima parte) e incluye el acoso cibernético además del callejero. También en esta encuesta sí se pregunta la nacionalidad y el rango de ingresos, en pesos, del hogar. Por otra parte, además de un instrumento descriptivo y estadístico, la implementación de esta encuesta en Argentina busca también “aportar información confiable para **propiciar medidas legislativas y recursos de intervención para las mujeres**” y “**promover el desarrollo y formulación de políticas y estrategias de prevención de la violencia**”, aunque no se explicita cómo.

1.2 ¿Quién creó el conjunto de datos (por ejemplo, qué equipo, grupo de investigación) y en nombre de qué entidad (por ejemplo, empresa, institución, organización)?

La MVM fue realizada por la Subdirección General de Sensibilización, Prevención y Estudios de la Violencia de Género de la Delegación del Gobierno contra la Violencia de Género del Ministerio de Igualdad del Gobierno de España.

1.3 ¿Quién financió la creación del conjunto de datos?

Si bien no hay información que corrobore el financiamiento de esta encuesta, suponemos que fue gestionado por el Estado español.

2. Composición del dataset

Estas preguntas están destinadas a proporcionar a los consumidores del conjunto de datos la información que necesitan para tomar decisiones informadas sobre el uso del conjunto de datos para tareas específicas.

2.1 ¿Qué representan las instancias que componen el conjunto de datos (por ejemplo, documentos, fotos, personas, países)? ¿Hay varios tipos de instancias (por ejemplo, películas, usuarios y calificaciones; personas e interacciones entre ellas; nodos y conexiones)? Proporcione una descripción.

Las instancias se corresponden con las personas entrevistadas.

2.2 ¿Cuántas instancias hay en total (y de cada tipo, si corresponde)?

Se encuestaron 9.568 personas.

2.3 ¿De qué datos consta cada instancia? ¿Datos "sin procesar" (por ejemplo, texto o imágenes sin procesar) o características? En cualquier caso, proporcione una descripción y algunos ejemplos.

Cada instancia consta de texto, siendo cada dato correspondiente la respuesta a una pregunta. Por Ejemplo: En la columna 'M0P1' de la primer instancia que mostramos a continuación, aparece que el dato es 'Varias veces a la semana'. Este ejemplo muestra con claridad cómo se constituye un dato. Para cada pregunta/columna, se le asignó un código que la representa, en este caso MP01, representa la pregunta: *Para empezar vamos a realizarle algunas preguntas sobre su vida cotidiana. ¿Con qué frecuencia se reúne con sus parientes o familiares o amistades con quienes no convive?*

MUN	DISTR	SECCION	CAPITAL	TAMUNI	COORDINADOR	ENTREV	M0P0	M0P1	M0P2	M0P3	M0P4	M0P4A	M0P4B	M0P4C	M0P5	M0P5A
Mun. =100.000 hab. no capitales de CA o provincia	Anonimizado	Anonimizado	Otros municipios	2.001 a 10.000 habitantes	C39	Anonimizado	Parada y ha trabajado antes	Varias veces a la semana	Sí	Sí	No	N.P.	N.P.	N.P.	No	N.P.
Girona / Madrid	Anonimizado	Anonimizado	Capital de CC.AA.	Más de 1.000.000 habitantes	C1707	Anonimizado	Estudiante	Varias veces al año	Sí	Sí	No	N.P.	N.P.	N.P.	No	N.P.
Mun. =100.000 hab. no capitales de CA o provincia	Anonimizado	Anonimizado	Otros municipios	50.001 a 100.000 habitantes	C405	Anonimizado	Estudiante	Varias veces a la semana	Sí	Sí	No	N.P.	N.P.	N.P.	No	N.P.

A la hora de trabajar con los datos, el nombre de las preguntas se reemplazó por un nombre más significativo, en este caso se cambió por 'info_personal_frec_reuniones_familiares'.

2.4 ¿Hay una etiqueta o anotación de supervisión asociada con cada instancia? Si es así, proporcione una descripción.

No se proporciona ninguna etiqueta o anotación para cada instancia.

2.5 ¿Cree que hay algún aspecto de los datos que no ha sido representado y que podría llegar a resultar relevante para alguna aplicación o por sus efectos en algún segmento de la población (p. ej., falta información de género)? De ser así, proporcione una descripción.

Tal vez se podrían haber incluido preguntas a la víctima sobre el agresor y sobre las preferencias en caso de requerir ayuda (tanto en casos de vivencia de violencia como en casos de terceros).

Sería bueno saber también si la encuestada se encontraba sola al momento de realizar la encuesta o si había más gente alrededor porque eso podría condicionar las respuestas.

2.6 ¿El conjunto de datos contiene datos que podrían considerarse confidenciales (por ejemplo, datos que están protegidos por privilegios legales o por la confidencialidad del médico-paciente, datos que incluyen el contenido de las comunicaciones no públicas de las personas)? Si es así, proporcione una descripción.

Sí, hay datos sensibles o de información personal sobre cada una de las encuestadas, algunos de los cuales han sido completados con “Anonimizado”. Por ejemplo, las categorías de distrito y sección, que detallan la ubicación geográfica de las personas encuestadas. También está anonimizado el encuestador.

3. Recopilación del dataset

Estas preguntas están destinadas a identificar características de la recopilación del dataset para advertir a consumidores del conjunto de datos sobre posibles sesgos emergentes debido a decisiones de recolección.

3.1 ¿Qué mecanismos se utilizaron para recopilar los datos, elegir la muestra y curar los datos (por ejemplo, aparatos o sensores de hardware, curación humana manual o automática, programa de software)? Esta pregunta es particularmente importante para datasets que son demasiado grandes para ser inspeccionados a mano. Explicar la racionalidad de la muestra puede ayudar a los usuarios de datasets a entender a qué tipos de usos el dataset puede generalizar bien o mal.

Los métodos utilizados no están explicitados. Sí para el análisis dice: Para realizar el análisis estadístico se ha empleado el programa SPSS v.21. Este programa redondea los resultados implicando que a veces los porcentajes finales de las tablas no van a sumar exactamente 100% sino que pueden sumar 99,9% o 100,1%. Lo mismo sucede con las frecuencias absolutas (hay que tener en cuenta que la muestra está ponderada para ajustarla a la población por lo que las frecuencias absolutas ponderadas en general son cifras con decimales). A veces la suma de frecuencias muestrales será una unidad más o una menos que el total que corresponda. En el estudio se muestran los resultados relativos a porcentajes con un decimal y el de las frecuencias muestrales sin decimales. Si se calculan los porcentajes a partir de las frecuencias muestrales que aparecen en el estudio, el resultado puede variar ligeramente del que aparece en las tablas debido al redondeo de cifras.

3.2 ¿Quiénes participaron en el proceso de recopilación de datos (p. Ej., Estudiantes, trabajadores digitales, contratistas) y cómo se les compensó (p. Ej., ¿Cuánto se les pagó a los trabajadores)?

La recopilación de datos, como así también el diseño de la muestra y las entrevistas, la codificación y la depuración, fueron realizadas por el Centro de Investigaciones Sociológicas de España. Se desconoce la formación previa de las personas participantes y el monto asignado a su pago.

3.3 ¿Durante qué período de tiempo se recopilaron los datos?

Los datos se recopilaron entre los meses de septiembre y diciembre de 2019.

3.4 ¿El conjunto de datos se relaciona con las personas? De lo contrario, puede omitir las preguntas de la sección 4.

Sí.

4. Datos de personas

4.1 ¿El conjunto de datos incluye datos demográficos (por ejemplo, por edad, sexo)? De ser así, enumere qué datos demográficos se incluyen.

Sí, se toman datos demográficos que incluyen: edad, nivel de estudios, situación laboral actual y pasada, tamaño familiar (cantidad de hijos con pareja actual o pasada, edad de los hijos, convivencia con pareja, hijos u otros parientes), fuentes de ingreso (incluyendo subsidios y becas), situación legal con la pareja. No se incluye orientación sexual, pero sí se pregunta sobre el sexo de parejas actuales o pasadas. Además se pregunta sobre si posee nacionalidad española y cuál es el país de nacimiento de cada persona entrevistada. No se pregunta sobre la raza, credo ni religión.

4.2 ¿Es posible identificar individuos (es decir, una o más personas físicas), ya sea directa o indirectamente (es decir, en combinación con otros datos) del conjunto de datos? Si es así, describa cómo.

No.

4.3 ¿El conjunto de datos contiene datos que podrían considerarse sensibles de alguna manera (por ejemplo, datos que revelan orígenes raciales o étnicos, orientaciones sexuales, creencias religiosas, etc.)?

Sí. Existen numerosos datos muy sensibles, aparte de la información personal, tales como información referente al estado de salud (incluyendo discapacidad, tendencias suicidas o problemas de salud mental) y a la vulnerabilidad (por ejemplo, si cuentan o no con allegados que puedan contenerlos emocionalmente o a quien puedan acudir en caso de algún tipo de necesidad, o las razones por las cuales no han denunciado casos de violencia -incluido el acoso sexual-).

4.4 ¿Recopiló los datos de las personas en cuestión directamente o los obtuvo a través de terceros u otras fuentes (por ejemplo, sitios web)?

Los datos están alojados en un sitio web.

4.5 ¿Se notificó a las personas en cuestión sobre la recopilación de datos? Si es así, describa cómo se envió la notificación.

Al tratarse de una macroencuesta, se asume que las personas entrevistadas eran conscientes de la recopilación de la información de sus respuestas al momento de efectuarse la misma. Desconocemos si se presentó algún tipo de *disclaimer* que asegurara a la entrevistada el anonimato de los datos proporcionados.

5. Usos

Estas preguntas están destinadas a alentar a los usuarios del dataset a reflexionar sobre las tareas para las que el dataset debe y no debe usarse. Al resaltar explícitamente estas tareas, los consumidores de datasets pueden tomar decisiones mejor informadas, previendo riesgos o daños potenciales.

5.1 ¿Para qué tareas creen que se puede usar este conjunto de datos? ¿Ya se ha utilizado el conjunto de datos para alguna tarea? Si es así, proporcione una descripción.

El conjunto de datos ya se ha empleado anteriormente, dado que se ha hecho un análisis de los datos por parte del gobierno español, dicho análisis es estadístico y descriptivo. Este tipo de macroencuestas se realizan cada 4 años con el fin de medir la prevalencia de la violencia contra la mujer. Su objetivo principal es conocer el porcentaje de mujeres mayores a 16 años que residen en España que sufren algún tipo de violencia por el hecho de ser mujeres. Es decir, violencia de género.

5.2 ¿Existe un repositorio que se vincule a alguno o todos los artículos o sistemas que utilizan el conjunto de datos? Si es así, proporcione un enlace u otro punto de acceso.

La página de la Delegación del Gobierno contra la Violencia de Género de España, tiene un apartado con las Macroencuestas que se han realizado, pero no parece vincular todos los artículos o sistemas que utilizan el conjunto de datos:

<https://violenciagenero.igualdad.gob.es/violenciaEnCifras/macroencuesta2015/home.htm>

5.3 ¿Para qué (otras) tareas se podría utilizar el conjunto de datos?

Creemos que este tipo de documentos, bajo este enfoque, sólo podría tener fines descriptivos o estadísticos, educativos o informativos.

5.4 ¿Hay algo sobre la composición del conjunto de datos o la forma en que se recopiló y preprocesó / limpió / etiquetó que pueda afectar los usos futuros? Por ejemplo, ¿hay algo que un futuro usuario pueda necesitar saber para evitar usos que podrían resultar en un trato injusto de individuos o grupos (por ejemplo, estereotipos, problemas de calidad del servicio) u otros daños indeseables (por ejemplo, daños financieros, riesgos legales)? Si es así, proporcione una descripción.

Existe un problema a la hora de trabajar con los datos y son algunas respuestas que se emplean. A la hora de armar el dataset, se introducen por valores nulos espacios (' '), además existen dos respuestas negativas a las preguntas, que son 'N.C.' y 'N.P.', la primera hace referencia a que *No corresponde*, y el segundo *No pertenece*. Ambas opciones refieren de que dicha pregunta no tiene relevancia o no corresponde hacerla por algo respondido anteriormente. Un problema entonces a la hora de emplear el dataset, es tener en cuenta todas las posibles respuestas negativas que pueden aparecer, dado que esto puede llevar a un análisis erróneo de los datos, y si se emplean los mismo para sacar conclusiones y querer establecer alguna política o acción, algo que nos parece que sería erróneo, se tiene que tener un cuidado mayor con respecto a las posibles respuestas que aparecen en el dataset.

5.5 ¿Hay tareas para las que no se debería utilizar el conjunto de datos? Si es así, proporcione una descripción.

No debería utilizarse con fines predictivos. Consideramos demasiado arriesgado, parcial e incluso inapropiado intentar buscar patrones o características de las mujeres violentadas, ya que los resultados pueden fomentar una mirada discriminatoria y sesgada sobre un fenómeno por demás complejo. Este tipo de interrogatorios nos proporciona un análisis descriptivo del fenómeno, en un determinado período de tiempo y área geográfica, pero no es posible hacer inferencias o extrapolaciones. Como mucho, sólo comparaciones.

5.6 ¿Algún otro comentario?