

Descripción del Dataset y Origen

Data set que se usará en el proyecto:

El data set está disponible en la carpeta “data” del git compartido al final de este documento. En dicha carpeta se encontrarán tanto los documentos originales en las carpetas “external” y “raw”, así como el material con el que se realizará la modelización, el cual se encuentra disponible en la carpeta “interim” con el nombre de *df_completo.csv*.

Descripción completa del dataset, incluyendo la cantidad de instancias, características (columnas), tipos de datos e información relevante:

La siguiente descripción se encuentra disponible en la carpeta “references” del git compartido:

Descripción de la base de datos

El data set está compuesto por 89541 observaciones o instancias y un total de 19 variables o características. El mismo se ha construido a partir de la concatenación de cuatro data frames base (data_2020, data_2021, data_2022, data_2023).

La siguiente tabla describe las variables presentes en el dataframe, la cantidad de datos (no nulos) que se encuentran en la misma

n° de variable	variable	Conteo no nulos	Tipo de dato
0	fecha	89541 non-null	object
1	prov_residencia_persona_en_situacion_violencia	88377 non-null	object
2	genero_persona_en_situacion_de_violencia	87637 non-null	object
3	edad_persona_en_situacion_de_violencia	72263 non-null	float64
4	pais_nacimiento_persona_en_situacion_de_violencia	59260 non-null	object
5	tipo_de_violencia_fisica	89541 non-null	object
6	tipo_de_violencia_psicologica	89541 non-null	object
7	tipo_de_violencia_sexual	89541 non-null	object
8	tipo_de_violencia_economica_y_patrimonial	89541 non-null	object
9	tipo_de_violencia_simbolica	89541 non-null	object

10	tipo_de_violencia_domestica	89541 non-null	object
11	modalidad_de_violencia_institucional	89541 non-null	object
12	modalidad_de_violencia_laboral	89541 non-null	object
13	modalidad_violencia_contra_libertad_reproductiva	89541 non-null	object
14	modalidad_de_violencia_obstetrica	89541 non-null	object
15	modalidad_de_violencia_mediatica	89541 non-null	object
16	modalidad_de_violencia_otras	89541 non-null	object
17	vinculo_con_la_persona_agresora	86146 non-null	object
18	genero_de_la_persona_agresora	80328 non-null	object

dtypes: float64(1), object(18)

Diccionario de datos


Este diccionario de datos se encuentra también disponible en la página web desde donde se extrajeron los datos. En ella, cada uno de los dataframes usados, comparten las mismas variables que se detallan a continuación:

Nombre de la columna	Tipo de dato	Descripción
fecha	Fecha ISO-8601 (date)	Fecha del ingreso de la consulta
prov_residencia_persona_en_situacion_de_violencia	Texto (string)	Provincia en donde la persona en situación de violencia declara estar residiendo al momento de la consulta
genero_persona_en_situacion_de_violencia	Texto (string)	Identidad de género autopercebida de la persona en situación de violencia
edad_persona_en_situacion_de_violencia	Texto (string)	Edad declarada de la persona en situación de violencia al momento en que se ingresó la consulta
pais_nacimiento_persona_en_situacion_de_violencia	Texto (string)	País de nacimiento de la persona en situación de violencia
tipo_de_violencia_fisica	Texto (string)	Tipo de violencia, contemplado por la ley 26.485, que motiva la consulta
tipo_de_violencia_psicologica	Texto (string)	Tipo de violencia, contemplado por la ley 26.485, que motiva la consulta
tipo_de_violencia_sexual	Texto (string)	Tipo de violencia, contemplado por la ley 26.485, que motiva la consulta

tipo_de_violencia_economica_y_patri monial	Texto (string)	Tipo de violencia, contemplado por la ley 26.485, que motiva la consulta
tipo_de_violencia_simbolica	Texto (string)	Tipo de violencia, contemplado por la ley 26.485, que motiva la consulta
tipo_de_violencia_domestica	Texto (string)	Modalidad de violencia, contempladas por la ley 26.485, que motivan la consulta bajo la cual se desarrollaron
modalidad_de_violencia_institucional	Texto (string)	Modalidad de violencia, contempladas por la ley 26.485, que motivan la consulta bajo la cual se desarrollaron
modalidad_de_violencia_laboral	Texto (string)	Modalidad de violencia, contempladas por la ley 26.485, que motivan la consulta bajo la cual se desarrollaron
modalidad_violencia_contra_libertad_ reproductiva	Texto (string)	Modalidad de violencia, contempladas por la ley 26.485, que motivan la consulta bajo la cual se desarrollaron
modalidad_de_violencia_obstetrica	Texto (string)	Modalidad de violencia, contempladas por la ley 26.485, que motivan la consulta bajo la cual se desarrollaron
modalidad_de_violencia_mediatica	Texto (string)	Modalidad de violencia, contempladas por la ley 26.485, que motivan la consulta bajo la cual se desarrollaron
modalidad_de_violencia_otras	Texto (string)	Categoría "Otras modalidades" para dejar registro de la presencia de otras modalidades distintas a las mencionadas (por ejemplo, violencia política, acoso callejero)
vinculo_con_la_persona_agresora	Texto (string)	Vínculo que tiene o tenía la persona en situación de violencia con quien ejerce la agresión en el momento en que sucedieron las situaciones de violencia declaradas
genero_de_la_persona_agresora	Texto (string)	Identidad de género de la persona agresora

Informe sobre el origen del dataset (fuente, la fecha de adquisición y recopilación o preprocesamiento realizado:

El origen del data set es la página de datos públicos de Argentina <https://datos.gob.ar/dataset/generos-base-datos-linea-144> en donde podemos observar la siguiente información respecto de los datos.

Información adicional		Mantenedor	Ministerio de las Mujeres, Géneros y Diversidad de la Nación - Dirección Técnica de Registros y Bases de Datos. ✉
Temas			
Etiquetas	144 Género género línea 144 violencia	Fecha de publicación	1 de septiembre de 2022
		Fecha de actualización	28 de julio de 2023
Licencia	Creative Commons Attribution 4.0	Página de referencia	https://www.argentina.gob.ar/linea144
Frecuencia de actualización	Trimestralmente	Fuente primaria	Buenos Aires, CABA, y sede de gestión nacional

De dicha web se extrajeron las cuatro bases de datos disponibles, a saber:

- ☐ Línea 144 -2020
- ☐ Línea 144 -2021
- ☐ Línea 144 - enero-diciembre - 2022
- ☐ Línea 144 - enero-junio - 2023

Los mismos fueron extraídos en el mes de mayo del año 2024, e incorporados al repositorio de github el día 15 de junio del mismo año.

A partir del set de datos, se construyó un data frame final con una concatenación donde se incorporaron los cuatro años disponibles respecto al tema. El data frame se encuentra disponible en la carpeta *interim* como se mencionó anteriormente.

En las carpeta *notebook* de github se hallan los documentos de jupyter con los que se ha realizado el preprocesamiento. Se decidió realizar la limpieza en dos etapas para aportar claridad al trabajo, por lo que se realizaron dos documentos *Preprocesamiento.ipynb* y *Descripción base_limpia.ipynb*. En la primera se encontrará la unión de los data sets originales y la descripción general de la “base_completa.csv”. En el segundo, se realizó la limpieza de la base anteriormente nombrada y se realizó la descripción de cada una de las variables que compone la “base_limpia.csv”

Enlace a github: <https://github.com/Nadpess/linea144.git>