Maestría en Explotación de Datos y Descubrimiento del Conocimiento

Universidad de Buenos Aires

TÍTULO DEL TRABAJO

Victoria Colombo

fecha

Resumen

ACÁ VA EL RESUMEN

Introducción

- antedecentes: explicación del programa que origina los datos. Algunas estadísticas o patrones de datos comunes a esta problemática encontrados en otros trabajos a nivel mundial o a nivel local. Por aquí agregar algo sobre la frecuencia con que las víctimas conviven con sus agresores. Por ejemplo durante la pandemia los números de violencia de género específicamente contra las mujeres no bajaron subieron estaban en sus casas. Problemas con los tipos de datos falta de denuncias

- objetivos: imputar los NS/NC de convive no convive con probabilidad de SI/NO

En general los datos sobre delitos sexuales son difíciles de recabar, no necesariamente por su naturaleza sino por el contexto social que los rodea. A las víctimas a menudo no se les ofrece empatía, contención ni un lugar seguro para relatar y denunciar lo sucedido, muchas veces son re victimizadas por el sistema judicial y/o por la sociedad misma y algunas veces, como se muestra en el documental Línea 137 (**?**), no reconocen algunas situaciones de violencia sexual. Esto puede tener que ver en parte con el ocultamiento de ese tipo de violencia cuando ocurre, la naturalización si la violencia se da al interior de una pareja y la falta de educación sexual no solo o no necesariamente de la víctima sino de su entorno social en general (**?**). Al mismo tiempo, recopilar y analizar estos datos para tener información estadística confiable sobre la problemática es importante para poder pensar e implementar soluciones efectivas.

El programa Las Víctimas contra las Violencias depende del Ministerio de Justicia y Derechos Humanos de la Nación y fue creado en el año 2006 con el objetivo de brindar atención e intervención institucional a víctimas de abusos y violencia familiar o sexual[1]. Para denunciar y solicitar asistencia las víctimas cuentan, desde 2016, con la línea nacional de emergencia 137 que funciona las 24 horas del día, todo el año, y cuenta en al menos cinco ciudades del país con equipos especializados para llevar a cabo el acompañamiento y las intervenciones necesarias. Los registros de esas llamadas e intervenciones se encuentran digitalizados al menos desde 2017 y están disponibles en el Portal de Datos Abiertos de la Justicia Argentina. Allí se recopilan bajo la clasificación de llamados e intervenciones domiciliarias por situaciones de violencia familiar y llamados e intervenciones domiciliarias por situaciones de violencia sexual.

no es u
línea pa
denunc

1. Datos

Para este trabajo he tomado los llamados de denuncias por violencia sexual desde enero de 2017 hasta julio de 2021. Para cada año descargué El *dataset* se compone en total de 19143 observaciones y 54 variables, en su mayoría categóricas, que aportan información sobre la víctima, el victimario, la persona denunciante, el contexto del hecho y el tipo de violencia sufrida. En la tabla **??** detallo las variables y su tipo.

DATOS - por menor de la construcción del data set que es distinto para datos abiertos que es el original - preprocesamiento: prepro para juntarlos + limpieza –¿script pipeline

Descripción de la limpieza del dataset:

Notas sobre la normalización nombres de columnas

caso_id en la unificación de datasets el de llamados de 2019 tiene el encoding distinto/roto 2017 y 2018 no tienen la columna caso_id pero los otros sí voy a droppear todos los caso id

llamante_vinculo 2017 2018: llamante_quien_llama 2019 adelante: llamante_vinculo

normalizo los nombres a llamante_vinculo

provincia_id

llamado_provincia_id (string): provincia desde la que se realiza el llamado a la Línea 137/0800-222-1717, según la codificación de provincia implementada por INDEC (hasta 05/2019 nombre campo llamado_provincia_indec_id) 2017 2018 2019 llamado_provincia$_{i}ndec_{i}d2021llamado_provincia_idnormalizoallamado_provincia_id$

fecha_hora cambio fecha hora por dos columnas fecha y hora

Tipos de datos paso a integer: llamante_edad, victima_edad para eso los faltantes los pasé a NA, porque estaban como Sin datos, era una columna de mixed types. Al pasar a integer los valores string (sin dato, NS/NC etc pasaron solos a N/A con coerce errors) dejo todos los otros datos en categorical (cambio los que no estaban)

Valores erróneos

normalicé los valores que tomaban las columnas porque había errores de tipeo en la carga y valores no normalizados que eran sinónimos en al menos 9 columnas: normalizo variaciones de NS/NC: 'Ns/Nc' / 'NS/NS' / 'Sin datos' = 'NS/NC' normalizo No/no a NO normalizo si/Sí/Si a SI llamante$_{e}dad$: $carguécomoNAlosmayoresa100ymenoresa3porconsider$

[1]Dentro de la categoría de violencia familiar se incluyen varios tipos de violencia, entre ellos, la sexual

Cuadro 1: Resumen de las variables.

Descriptor	Tipo variable	Variables	Observaciones
Víctima	Cuantitativa	victima_edad	
	Cualitativa	victima_genero, victima_nacionalidad, victima_discapacidad, victima_vinculo_agresor, victima_convive_agresor, victima_a_resguardo	victima_genero toma los valores: masculino, femenino, trans, NS/NC.
Llamante	Cuantitativa	llamante_edad	
	Cualitativa	llamante_genero, llamante_vinculo	llamante_genero toma los valores: masculino, femenino, trans, NS/NC. llamante_vinculo_ refiere a vínculo con la víctima.
Llamado	Cuantitativa	llamado_fecha_hora	
	Cualitativa	caso_id, llamado_provincia, llamado_provincia_id, caso_judicializado, hecho_lugar	llamado_provincia_id refiere al id numérico para las provincias según codificación INDEC.
Violencia sexual	Cualitativa	vs_violacion_via_vaginal, vs_violacion_via_anal, vs_violacion_via_oral, vs_tentativa_violacion, vs_tocamiento_sexual, vs_intento_tocamiento, vs_intento_violacion_tercera_persona, vs_grooming, vs_exhibicionismo, vs_amenazas_verbales_contenido_sexual, vs_explotacion_sexual, vs_explotacion_sexual_comercial, vs_explotacion_sexual_viajes_turismo, vs_sospecha_trata_personas-_fines_sexuales, vs_existencia_facilitador-_corrupcion_nnya, vs_obligacion_sacarse_fotos_pornograficas, vs_eyaculacion_partes_cuerpo, vs_acoso_sexual, vs_iniciacion_sexual_forzada_inducida, vs_otra_forma_violencia_sexual, vs_no_sabe_no_contesta	vs_existencia_facilitador-_corrupcion_nnya refiere a la existencia de un facilitador de la corrupción de niños, niñas y adolescentes. vs_no_sabe_no_contesta refiere violencia sexual que se desconoce o que no hace referencia a los otros campos mencionados.
Otras violencias	Cualitativa	ofv_sentimiento_amenaza, ofv_amenazas_explicitas, ofv_violencia_fisica, ofv_intento_ahorcar, ofv_intento_quemar, ofv_intento_ahogar, ofv_amenaza_muerte, ofv_uso_sustancias_psicoactivas, ofv_intento_privacion_libertad, ofv_privacion_libertad, ofv_uso_arma_blanca, ofv_uso_arma_fuego, ofv_enganio_seduccion, ofv_intento_matar, ofv_uso_animal_victimizar, ofv_grooming, ofv_otra_forma_violencia, ofv_no_sabe_no_contesta	

victima edad: saqué los que parecían ser errores: 111 125 y 221

normalicé nombres de provincias uinifiqué caba con ciudad autonoma de buenos aires, santa fé y santa fe dejé solo santa fe

- Descripción: análisis medidas de centralidad, cantidad de faltantes y de NS/NC. –¿notebook epxloración
- Análisis con respecto a convive

2. Metodología

Los datos de los llamados desde 2017 hasta julio de 2021 fueron descargados del portal mencionado en la sección anterior en cinco archivos de formato csv separados, uno por año. La unificación de esos archivos en un solo *dataset* implicó realizar algunas modificaciones para sortear problemas de correspondencias entre años. La variable *caso_id* solo existe desde el primer trimestre de 2020, los casos anteriores a esa fecha no contaban con ella, por lo tanto tomé la decisión de eliminarla también para 2020 y 2021. La variable *llamado_provincia_id* llevaba otro nombre hasta el año 2019: *llamado_provincia_indec_id* y fue entonces modificada en 2017, 2018 y 2019 para llevar el nombre actual. Los *types* de las variables cualitativas fueron cambiados a *categorical*[2] . Además, los valores que tomaban al menos 9 de esas variables categóricas debieron ser normalizados por errores varios en la carga o valores cargados con sinónimos. Por ejemplo, muchos NS/NC fueron cargados en minúscula y en mayúscula en la misma columna y debieron ser normalizados a mayúscula; además, por ejemplo en la variable *victima_vinculo_agresor* se repetían algunos valores cargados con distinta ortografía como .Ex pareja de la víctimaz .Ex-pareja de la víctimaz .Expareja de la víctima"que debieron ser normalizados. Los *types* de las variables cuantitativas de edad fueron pasados a *integer*. Los valores numéricos de *victima_edad* y *llamante_edad* tenían errores de carga evidentes ya que aparecían valores numéricos demasiado altos para ser edades como: 125, 221, 324. Las filas con esos valores no fueron eliminadas por el momento porque considero que el resto de los datos de la fila no están errados y es posible que los necesite más adelante. En cambio, los datos fueron marcados para no ser utilizados en análisis que incluyan las variables de edad. A modo de análisis exploratorio, realicé histogramas univariados para ver la frecuencia de las categorías de las variables: *victima_genero*, *victima_discapacidad*, *victima_convive_agresor*, *victima_vinculo_agresor*, *llamante_edad*, *llamante_genero*, *llamante_vinculo* y *hecho_lugar*. Además, realicé un agrupamiento de las categorías de vínculos entre agresor y víctima para poder distinguir entre parejas, familiares y no familiares (conocidos). Algunos de estos histogramas se comentan en la sección siguiente.

Me propongo como continuación de este análisis explorar la fecha y hora de los llamados, las edades de las víctimas y llamantes, las formas de violencia más comunes, y construir una variable de género del agresor utilizando la variable que estipula el vínculo entre la víctima y el agresor, ya que en algunas de sus categorías el género se encuentra expresado inequívocamente (por ejemplo en las categorías padre, madre, hermano). Además, me interesa sumar análisis multivariados para ver la interacción entre algunas de las variables. Por último, tengo la intención de investigar asociaciones entre variables como edad de la víctima y vínculo con el agresor.

3. Reducción de dimensiones y análisis

Tratamient y reducción de variables que describen violencia sexual y otras formas de violencia con un criterio cualitativo

Las variables vs y ofv son muchas, hay muchas categorías dentro de cada una.

Las variables de vs y ofv toman los valores SI NO.

En general hay muchos más NO que sí.

Algunas resultan muy poco informativas como vs explotacion sexual viajes turismo (0,02) y ofv intento matar (0,01) resultan muy poco informativas (ocurrencia de 0,02 y 0,01).

Muchas pertenecen al mismo dominio de tipo de violencia, por ejemplo: vs violacion via vaginal vs violacion via anal vs violacion via oral.

Una forma de reducir las dimensaiones del dataset para ver tendencias (?) es agrupar variables similares entre sí.

También podría eliminar las poco o nada informativas pero por el momento voy a agruparlas nomás.

Los agrupamientos propuestos se basan en conocimiento de dominio: la pertenencia de las distintas variables dentro de un agrupamiento al mismo tipo de violencia ejercida sobre una víctima.

las variables vs_violacion_via_vaginal, vs_violacion_via_anal, vs_violacion_via_oral, vs_tentativa_violacion y vs_intento_violacion se agrupan en una sola variable de violación

[2]El análisis exploratorio y el resto del trabajo con datos fue y será realizado en Python

las variables vs_tocamiento_sexual y
vs_intento_tocamiento se agrupan en una sola variable de tocamiento sexual

las variables vs_explotacion_sexual, vs_explotacion_sexual_comercial y vs_explotacion_sexual_viajes_turismo se agrupan en una sola variable de explotación

Las variables ofv uso arma blanca ofv uso arma fuego se agrupan en una sola variable de uso de arma

Las varaibles ofv intento ahogar ofv intento quemar ofv intento matar ofv intento ahorcar se agrupan en una sola variable ofv intento violencia potencialmente fatal/ intento violencia extrema.

Candidatas a eliminarse si esa fuera la elección:

VS con un punto de corte de al menos 10 ocurrencias en todo el dataset: vs_explotacion_sexual_viajes_turismo

OFV con un punto de corte de al menos 10 ocurrencias de SI en todo el dataset: ofv uso animal victimizar ofv intento ahogar ofv intento quemar ofv intento matar

* del mail con soria:

3. Tengo un agrupamiento cualitativo pensado simplemente para achicar la dimensionalidad juntando variables entre sí. Las variables originales están en la imagen adjunta "variables_vs_ofv_original", y el agrupamiento propuesto está ejemplificado para las de violencia sexual aquí, para las de ofv es bastante similar. Lo que me gustaría es nuevamente algún material de apoyo bibliográfico para estas técnicas manuales de reducción de dimensionalidad. Quizás no haya o no sea necesario tener tanto basamento, si les parece que es así, también acepto esa respuesta.

Me parece bien el agrupamiento que proponés. Como te decía, acá es más importante poder justificar desde el dominio, y no tanto desde los datos en sí. No hay reglas escritas que te digan si una variable tiene una distribución, por ejemplo, 96 % SI y 4 % no, hay que descartarla. El hecho de que vos puedas justificar desde el dominio, después te facilita la interpretación. Por ejemplo, cuando juntás todos los tipos de explotación en una sola. Está bien, porque explotación es algo bien delimitado, y para un trabajo donde no hay tantos datos, no sería posible entrar a indagar mucho sobre la variante de explotación.

4. Construcción de un modelo predictivo

5. Resultados

6. Conclusiones

Referencias

Línea 137 (documental), 2020.

Juan Manuel Contreras, S Both, A Guedes, and E Dartnall. Violencia sexual en latinoamérica y el caribe: análisis de datos secundarios. iniciativa de investigación sobre la violencia sexual., 2016.