380.69788pt

Maestría en Explotación de Datos y Descubrimiento del Conocimiento

Universidad de Buenos Aires

TÍTULO DEL TRABAJO

Victoria Colombo

fecha

Lista de tareas pendientes

introducción . . . 3
no es una línea para denunciar . . . 3
ver y resumir acá toda la descripción de faltantes que hay en el documento notas . . . 4
análisis respecto a convive . . . 4
esto va en metodología . . . 4
por qué quiero reducir las dimensiones del dataset . . . 5
poner gráficos con respecto a estas nuevas variables . . . 6
esto va en datos o en metodología??? . . . 6
describir . . . 6
poner la cantidad de variables que quedan en este dataset (es el 5) y las que había en el original . . . 7

Estructura propuesta del trabajo

RESUMEN Introducción Datos Metodología Resultados Discusión y conclusiones Bibliografía

Resumen

1. Introducción

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

introducción

1. dificultad de recabar datos

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.

2. descripción del programa 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 una línea para denunciar

2. Datos

Para este trabajo he tomado en principio los llamados de denuncias por violencia sexual desde enero de 2017 hasta julio de 2021. El *dataset* se compone en total de 19143 observaciones y 54 variables, en su mayoría categóricas, que aportan información sobre la víctima, la persona denunciante, el contexto del hecho y el tipo de violencia sufrida. En la tabla **??** se detallan 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 1.1 nombres de columnas 1.1.1 caso$_i$*denlaunificacióndedatasetseldellamadosde2*

2017 y 2018 no tienen la columna caso$_i$*dperolosotrossí*

voy a droppear todos los caso id

1.1.2 llamante$_v$*inculo*

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

2017 2018: llamante$_{q}uien_{l}lama2019adelante$: $llamante_{v}inculo$

normalizo los nombres a llamante$_{v}inculo$

1.1.3 provincia$_{i}d$

llamado$_{p}rovincia_{i}d(string)$: $provinciadesdelaqueserealizaelllamadoalaLínea137/0800-222-1717,según$

2017 2018 2019 llamado$_{p}rovincia_{i}ndec_{i}d == 2021llamado_{p}rovincia_{i}d$

normalizo a llamado$_{p}rovincia_{i}d$

1.1.4 fecha$_{h}oracambiofechahorapordoscolumnasfechayhora$

1.2 Tipos de datos paso a integer: llamante$_{e}dad, victima_{e}dadparaesosolosfaltanteslospaséaNA, porqueestaban$

1.3 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$_{e}dad$: $carguécomoNAlosmayoresa100ymenoresa3porconsiderarloserrorSi$

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

ver y resumir acá toda la descripción de faltantes que hay en el documento notas

- Análisis con respecto a convive

análisis respecto a convive

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.

esto va en metodología

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

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

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.

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.

por qué quiero reducir las dimensiones del dataset

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_tercera_persona se agrupan en una sola variable de violación

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.

3. Metodología

1. Manipulación de variables y reducción de dimensiones - armado de variables para ver otros patrones: género agresor, agresor conocido/no conocido, agresor familiar no familiar, momento del día, estación del año OK

poner gráficos con respecto a estas nuevas variables

Manual: A. reducir la cardinalidad de hecho lugar, provincia, llamante vínculo y victima$_v inculo_a gresorvínculo$.

Si bien al principio la idea era que la primera prueba de svm fuera con el dataset completo, normalizado pero con poca o ninguna intervención en la construcción de variables; llegados a este punto, la cardinalidad de alta de estasvaraibles lleva a tomar la decisón de reducrilas sin antes correr el experimento con svm porque ya está probado en la literatura que alta cardinalidad con encoders tipo one hot es mala y el target o ordinal encoder que funcionan bien para alta cardinalidad no me convencve para etsas vaiables porque no hay ordinalidad que preserar y porque el target implica tener otros cuidados para no incurrir en data leackage

esto va en datos o en metodología???

- provincia: con porcentaje que aparece cada provincia o con agrupación por zona del país. OK Norte (NOA, NEA), Central (Cuyana, Pampeana), Patagónica, Bs as, CABA, NS/NC. Se podrían haber elegido otras formas de agrupar pero lo cierto es que CABA y Buenos Aires representan el 80 de los llamados, NS/NC el 9 por ciento, Córdoba, Santa Fe, Tucumán, y Mendoza el 7 por ciento, y el resto de la provincias representan individualmente menos del 1 pr ciento de los llamados recibidos.

- hecho lugar: ver porcentajes que representan y agrupar por dominio OK Otro: le sumé a la categoría Otro (5 pct), que representa el 5pct de los casos, Residencia turística (menos del 1 pct),Obra en construcción (menos del 1 pct), Taxi (menos del 1 pct),Albergue transitorio (menos del 1 pct),Automóvil (menos de 1 pct),Comercio (menos de 2pct),Ámbito educativo(menos de 3 pct),Vivienda de un familiar (3 pct) que están todas por debajo del 3 pct. Espacio público: Subterráneo/Tren/Colectivo menos del 1pct, Plaza y Descampado son menos del 1 pct, Calle 3pct Después quedaron las categoirías originales: vivienda de la víctima(25pct), vivienda del agresor(13pct), redes sociales(12), y NS/NC(29)

- llamante vínculo: agrupé por dominio y con vistas a porcentajes representados. OK Institición (Hospital 1.86pct pct, Comisaría 24.91 pct, Escuela 0.42 pct, Defensoría 0.10 pct, Otra Institución 2.04 pct), Conocido de la víctima (puede ser familiar o no familiar) (Madre 16.09 pct, Vecina/o 3.05, Padre 4.54 pct, Familiar 17.28 pct, Otro conocido 6.95 pct, Abuela/o 1.20 pct,Hermana/o 1.01 pct), Agresor 0.09 pct, víctima 14.28 pct, y NS/NC 6.17 pct.

- agresor vínculo: me quedé con la agrupación previa de conocido no conocido pero distinguiendo si el conocido es familiar o no, porque las categorías que ya existen en la variable vínculo cn el agresor me lo permiten y porque la cantidad de casos en que el agresor es conocido pero no familiar es mucho más alta que la cantidad de casos de cada fmailiar. OK Conocido: Conocido no familiar 19.79, Conocido familiar: Padre 11.59, Otro pariente 10.09, Padrastro 8.66, Tío 6.26, abuelo 3.23, Hermano 2.77, Ex pareja 1.79, Pareja 1.04, Madre 0.97, Abuela 0.76, Hermana 0.15, Madrastra 0.10. NS/NC 17.28. Desconocido 15.45

B. agrupar variables de violencia x dominio OK script pipeline

describir

violencia sexual:

se agrupan en una sola variable porque comparten dominio semántico (?) y jurídico:

vs explotación sexual vs explotación sexual comercial vs explotación sexual viajes turismo vs sospecha trata personas fines sexuales https://www.argentina.gob.ar/justicia/derechofacil/leysimple/trat de-personas https://www.argentina.gob.ar/trabajo/trata-de-personas

Nueva variable: explotación sexual

se agrupan en una nueva variable porque ... violación?:

vs violacion via vaginal vs violacion via anal vs violacion via oral

Nueva variable: violación

ofv:

se agrupan en una sola variable por dominio:

ofv intento ahogar ofv intento quemar ofv intento matar ofv intento ahorcar

Nueva variable: intento violencia potencialmente fatal

Se agrupan en una nueva variable por dominio:

ofv uso arma blanca ofv uso arma fuego

Nueva variable: uso de arma

C. quitar las variables poco informativas xq tienen baja tasa de rta.(menos de 191 SI, es decir 1 por ciento o menos de SI) Las borradas son: ['vs$_a menazas_v erbales_c ontenido_s exual'$,' $vs_e xistencia_f acilit$

Entonces el dataset reducido tiene: - reducción de llamante$_v inculo, victima_v inculo_a gresor, hecho_l ugar, yllama$ $nuevasvariables$: $violacion, usodearma, explotacionsexual, intentoviolenciapotencialmentefatalqueagrupanca$

Correlación: - correlacion entre numerica y binaria: point biserial -¿edad v, edad ll, fecha$_h oracontravictimaconvivesi/no, SINNS/NC$

Primero testeé la normalidad de las variables de edad con el test de shapiro wilkinson y con un histograma. Ninguna de las dos es normal, lo que hace perder fuerza a los resultados de la correlación point biserial. Sin embargo, la voy a hacer.

Los resultados de point biserial son: (tabla) Llamante edad - convive: correlation: 0.0377 p-value: 0.0004

Victima edad - convive: correlation: -0.2005 p-value: 0.0000

Para llamante edad el test indica una correlación positiva muy débil, es decir, a mayor edad del llamante, más probabilidad de que la víctima conviva con el agresor. El p-valor muy por debajo del estándar de 0.05 indicaría que la correlación observada, si bien débil, podría ser estadísticamente significativa.

Para victima edad el test indica una correlación negativa débil, es decir, a mayor edad de la víctima, menos probabilidad de que esta conviva con el agresor. El p-valor de 0.0000 indicaría también en este caso que la correlación observada podría ser estadísticamente significativa.

Nuevamente es necesario recordar que la confiabilidad en los resultados de estos tests se ve afectada por la falta de normalidad de las variables.

- correlacion entre binaria y binaria: phi coefficient -¿violencia (todas), fin de semana, resguardo

The phi coefficient is a measure of the degree of association between two binary variables. This measure is similar to the correlation coefficient in its interpretation.

Two binary variables are considered positively associated if most of the data falls along the diagonal cells (i.e., a and d are larger than b and c). In contrast, two binary variables are considered negatively associated if most of the data falls off the diagonal.

Se midió la correlación entre la variable victima$_c onvive_a gresorSINNS/NCylasvariablesbinarias$: $fin_d e_s emana, vs_e xplotacion_s exual_g roup, vs_v iolacion_g roup, vs_t entativa_g roup, vs_t ocamiento_s exual, vs_i ntento_t oc$

Los valores de phi sobre la correlación entre cada una de estas variables binarias y la variable target victima$_c onvive_a gresortodospositivosperotodosmuybajos, siendoelmásalto$0,41$paravictima_a resg$ $valoresde$0$.Puedeverselatablacompletaderesultadosenelanexo.$

- correlacion entre categorica y categorica: chi2 -¿provincia, ll genero, ll vinculo, caso judic, lugar, v genero, v nacionalidad, vic vinculo agr, vic discapacidad contra convive SIN NS/NC en convive

En la tabla resultante se ven los valores para X2 para cada variable en relación a la variable target çonvive", y el p-valor asociado. Todas las variables tienen una asociación estadísticamente significativa con la variable de convivencia.

Tiene en total [cantidad de variables], es decir [cantidad de variables] menos que el original —

poner la cantidad de variables que quedan en este dataset (es el 5) y las que había en el original

2. Visualización:

Intento ver si usando un método de ordenamiento para visualizar el dataset en dimensiones reducidas me da una idea de agrupamientos con repsecto a las tres categorías de convive. Elijo NMDS porque me permite trabajar con variables de distinto tipo sin transformaciones.

A. NMDS usando una matriz de distancias de gower: visualizar patrones en los casos.

versiones: - uso solo los datos completos de edad victima, ll2 ll5 - luego solo completos de edad llamante, ll2 ll5 - luego datos completos de ambos. ll2 ll5

3. Predictivo Encoding Methods for Categorical Data: A Comparative Analysis for Linear Models, Decision Trees, and Support Vector Machines

Finalmente aplico un SVM para predecir los NS/NC de convive como si fueran faltantes. Uso el dataset con todas las variables modificadas y etc. Preparación: A. reemplazo NS/NC por NA. B. Dados los faltantes en la variable edad, voy a usar mis dos versiones del dataset, una con edad pasada a categórica y luego a dummy, y otra con edad dejada en numérica con sus faltantes y todo C. aplico disitntos encoders a la variables, porque en algunas me interesa mantener la ordinalidad y en otras no

Ordinal encoder para timestampt y escalar. Lo de escalar lo saqué de Encoding Methods for Categorical Data: A Comparative Analysis for Linear Models, Decision Trees, and Support Vector Machines

Si bien al principio la idea era que la primera prueba de svm fuera con el dataset completo, normalizado pero con poca o ninguna intervención en la construcción de variables; llegados a este punto, la cardinalidad de alta de estasvaraibles lleva a tomar la decisón de reducrilas sin antes correr el experimento con svm porque ya está probado en la literatura que alta cardinalidad con encoders tipo one hot es mala y el target o ordinal encoder que funcionan bien para alta cardinalidad no me convencve para etsas vaiables porque no hay ordinalidad que preserar y porque el target implica tener otros cuidados para no incurrir en data leackage

D. Predictivo para llenar los NS/NC

4. Resultados

5. Discusión y conclusiones

Referencias

[Contreras et al.(2016)Contreras, Both, Guedes, and Dartnall] 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.

[Vassallo, L.(2020)] Vassallo, L. Documental línea 137, 2020.