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

## Estructura propuesta del trabajo

RESUMEN Introducción Datos Metodología Resultados Discusión y conclusiones Bibliografía

# Resumen

# 1. Introducción

INTRODUCCIÓN

# 2. Datos

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

Manual: reducir la cardinalidad de hecho lugar, provincia, llamante vínculo y victima_vinculo_agresor vínculo.

poner gráficos con respecto a estas nuevas variables

esto va en datos o en metodología???

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

- 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
Correlación:

A continuación, medí la correlación entre `victima_convive_agresor` (solo para los casos de `SI` y `NO`) y el resto de las variables con el objetivo de inspeccionar la fuerza y dirección de las posibles relaciones. Utilicé tres medidas de correlación distintas apropiadas para los distintos grupos de variables del *dataset*: 52

- Correlación puntual biserial para medir la correlación entre `victima_convive_agresor` (transformada a binaria) y las variables numéricas.
- Coeficiente fi para medir la correlación entre `victima_convive_agresor` (transformada a binaria) y las variables binarias.
- Coeficiente chi cuadrado para medir la correlación entre `victima_convive_agresor` y las variables categóricas.

En el cuadro se pueden apreciar los resultados de la correlación puntual biserial entre los valores `SI` y `NO` de `victima_convive_agresor` y las variables de edad. Para `llamante_edad` el test se encuentra una correlación positiva muy débil, es decir, a mayor edad quien llama, 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` se encuentra 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. Sin embargo, la confiabilidad de los resultados del test de correlación se ve afectada negativamente por la falta de normalidad de ambas variables de edad, comprobada con la aplicación del test Shapiro-Wilkinson, cuyos resultados fueron:

- Edad de la víctima: estadístico de S-W= 0.886, p= 0.000. Se rechaza H0.
- Edad del llamante: estadístico de S-W= 0.974, p= 0.000. Se rechaza H0.

Cuadro 1: Correlación puntual biserial `victima_convive_agresor` - variables numéricas.

| | `llamante_edad` | `victima_edad` |
|---|---|---|
| Correlación | 0.0377 | -0.2005 |
| P-valor | 0.0004 | 0.0000 |

El cálculo del coeficiente $\Phi$ de correlación entre la variable objetivo y las variables binarias (todas las indicadoras del tipo de violencia, `fin_de_semana`, `victima_a_resguardo`, y `caso_judicializado`) resultó en valores positivos para todas las asociaciones. Sin embargo, para algunas variables el p-valor resultante era muy elevado, por lo tanto decidí eliminarlas por considerar que la correlación insinuada por el estadítico no es estadísticamente significativa:

Cuadro 2: Cuartiles de edad según categoría de `victima_convive_agresor`.

| | Convive | No Convive | NS/NC |
|---|---|---|---|
| Q1 | 7 | 11 | 9 |
| Media | 12 | 15 | 13 |
| Q3 | 15 | 24 | 18 |
| IQR | 8 | 13 | 9 |

- `ofv_otra_forma_violencia`
- `ofv_amenaza_explicita`
- `vs_obligacion_sacarse_fotos_pornograficas`
- `vs_intento_tocamiento`
- `vs_violacion_group`
- `vs_tentativa_group`
- `fin_de_semana`

Para el resto de las varaibles los valores del estadístico son positivos aunque bajos y los p-valores aportan significacncia estadística.

lo que indicaría asociación positiva. Sin embargo, es necesario aclarar que todos los valores resultaron bajos, siendo el más alto 0.41 para victima_a_resguardo, 0.16 para vs_grooming, y 0.11 para vs_no_sa_no_contesta. En los tres casos el p-valor es de 0. Puede verse la tabla completa de resultados en el anexo. SACAR ofv_otra_forma_violencia 0.002097 0.796705 SACAR ofv_amenaza_explicita 0.014023 0.085017 SACAR vs_obligacion_sacarse_fotos_pornogra 0.015149 0.062793 SACAR vs_intento_tocamiento 0.003637 0.655076 SACAR vs_violacion_group 0.002538 0.755210 SACAR vs_tentativa_group 0.006390 0.432564 SACAR fin_de_semana 0.010664 0.190263

- 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