Violencia intrafamiliar y comunitaria en Nuevo León

Gabriela Sánchez Yepez

Posgrado en Ingeniería de Sistemas
Facultad de Ingeniería Mecánica y Eléctrica
Universidad Autónoma de Nuevo León

Abstract

Se presenta un análisis de incidentes de violencia intrafamiliar y comunitaria en el estado de Nuevo León reportados durante el año 2018. Es importante identificar los grupos vulnerables y posibles causas al problema que permitan planear estrategias de prevención. Para realizar dicho análisis, se utilizan distintas herramientas que proporciona el lenguaje de programación Python.

Keywords:

1. Introducción

Este trabajo busca aplicar herramientas de ciencia de datos que permitan analizar el efecto de distintos factores presentes en incidentes de violencia intrafamiliar y comunitaria en el estado de Nuevo León, con el objetivo de identificar vulnerables y posibles causas al problema que permitan planear estrategias de prevención.

La estructura del artículo es la siguiente: en la sección 2 se describen las características de los datos utilizados en el estudio así como las herramientas que se usan en los distintos tipos de análisis, los resultados de dichos análisis se presentan y discuten en la sección 3. Finalmente, en la sección 4 se presentan las conclusiones.

2. Metodología

En esta sección se especifican las carácteristicas de los datos con los que se trabaja así como las herramientas utilizadas para el análisis de dichos datos.

2.1. Datos

Los datos fueron proporcionados por la doctora Patricia L. Cerda Pérez, en formato xlsx.

Los datos se presentan en dos archivos, el primero contiene los incidentes de violencia intrafamiliar y comunitaria presentados durante los meses de enero a noviembre en el año 2018 y el segundo, los incidentes del resto del año, teniendo un total de 16 410 reportes.

Los registros contienen 47 columnas de las cuales únicamente 17 proporcionan información relevante, sin datos personales, que serán usados en el análisis. Estos registros proveen datos sobre lugar, fecha y hora de los incidentes, así como información sobre la(s) víctima(s) y agresor(es).

2.2. Herramientas

En esta subsección se explica la selección de herramientas para los distintos análisis que fueron realizados.

Preprocesamiento de los datos

Para realizar un preprocesamiento de los datos primero se utilizó la herramienta BASH que permitió convertir el archivo de los datos a un formato manejable, esto es, csv, utilizando la instrucción

```
sscovert datos.xlsx vf.csv
```

La siguiente fase de preprocesamiento se realizó con la librería PANDAS. El primer paso es unir los datos en un solo archivo, eliminar la información que no se utiliza y cambiar los tipos de los datos para no tener problemas una vez que se empiece a procesar la información.

Estadística descriptiva y visualización de la información

El análisis estadístico básico así como una visualización de la información se realizó con distintas herramientas proporcionadas por la librería PANDAS.

Para poder trabajar con esta librería es necesario cargar los datos para que se lean en el formato que se requiere, esto se logra con

```
import pandas as pd
vf = pd.read_csv("vf.csv")
```

En el análisis descriptivo se revisan los datos que pueden estudiarse cuantitativamente tales como la cantidad de incidentes reportados por mes, hora y fecha, entre otros. Esto se puede realizar utilizando la siguiente instrucción

```
vf.mes.value_counts(sort=True, normalize=False,
    dropna=True)
```

los resultados nos permiten observar el impacto de dichos factores en el total de los incidentes.

Estos y otros resultados, tales como las edades de las víctimas y agresores, pueden visualizarse gráficamente utilizando la librería PLOTLY.

Para poder graficar estos datos, la información correspondiente debe ser procesada ya que hay reportes para los cuales no se especifica la edad del agresor o la víctima; en algunos casos el reporte fue llenado con un *NE* y en otros casos se reportó un rango por ejemplo *40 a 45 años*. La forma en que se procede a limpiar estos datos es la siguiente

```
edadv = vf.edad_v # edades de las victimas
edad_v = []
for dato in edadv:
    s = str(dato).replace(",", " ")
    pedazos = s.split()
    while "a" in pedazos:
        pos = pedazos.index("a")
        desde = int(pedazos[pos - 1])
        hasta = int(pedazos[pos + 1])
        prom = (desde + hasta) // 2
        edad_v.append(prom)
        pedazos = pedazos[:(pos - 1)] + pedazos[(pos
            + 2):]
    edad_v += pedazos
edad_v = list(filter(lambda dato: dato != "NE",
    edad_v))
edad_v = list(filter(lambda dato: dato != "nan",
    edad_v))
```

Una vez que los datos estén en el formato adecuado se procede a graficar. Por ejemplo, teniendo la información de las edades de las víctimas y agresores en los *diccionarios di_edadA* y *di_edadV*, se puede graficar un diagrama de barras:

```
import plotly
import plotly.plotly as py
import plotly.graph_objs as go

trace1 = go.Bar(
    x = list(di_edadA.keys()),
    y = list(di_edadA.values()),
    name = 'Edad agresores'
)
trace2 = go.Bar(
    x = list(di_edadV.keys()),
    y = list(di_edadV.values()),
    name = 'Edad victimas'
)

data = [trace1, trace2]
layout = go.Layout(
    barmode = 'group'
)
```

Pruebas estadísticas

Para realizar pruebas estadísticas como determinar la normalidad de los datos, así como probar modelos lineales se utiliza la librería *scipy.stats*.

Los modelos lineales sirven para representar una variable de interés, por ejemplo la cantidad de incidentes reportados, como una funcin de uno o más factores conocidos. Algunos de los modelos lineales que se analizan son: la relación entre la cantidad de incidentes y la hora, mes y municipio.

También se analizan modelos de regresión múltiple, los cuales tienen más de un factor modelando conjuntamente la variable de interés. Esto se obtiene utilizando la librería *statsmodels*, que también se utiliza para realizar análisis de varianza que ayudan a cuantificar sí o no una variable o factor tiene un efecto estadísticamente significativo en la variable de interés.

Uno de los modelos que se analiza el efecto de los factores *edad de la víctima, hora y mes* en la cantidad de incidentes. Para este y la siguiente experimentación solo se utilizan los datos correspondientes con una víctima y un agresor teniendo en total 1479 reportes.

```
import statsmodels.api as sm
from statsmodels.formula.api import ols

datos = pd.read_csv("vf_1-1.csv")
datos['cuantos'] = 1
c = datos.groupby(['edad_v', 'mes', 'hora'],
    as_index=False).agg({"cuantos": "sum"})
m = ols('cuantos ~ edad_v + hora + mes', data =
    c).fit()
a = sm.stats.anova_lm(m, typ = 2)

print(a)
n = len(a)
alpha = 0.05
for i in range(n):
    print("{:s} {:s}es
        significativo".format(a.index[i], "" if
        a['PR(>F)'][i] < alpha else "NO "))
```

Pronósticos

Estudiar pronósticos de los datos es importante ya que permite predecir un valor que tomará una variable a partir de valores que haya tomado en la anterioridad, si se analiza la cantidad de incidentes de acuerdo a distintos factores podría ayudar a planear estrategias de solución. Los pronósticos también se realizan con la librería *statsmodels*.

3. Resultados

En esta sección se presentan los resultados del análisis y experimentos que se realizan.

Primero se muestran gráficos que permiten una visualización de la información con la que se trabaja.

Las Figuras 1 y 2 muestran un histograma de la cantidad de incidentes reportados de acuerdo a las horas y de acuerdo a los meses respectivamente. De ellas podemos observar que el rango de horarios de las 20:00 - 01:00 horas es el que reporta una mayor cantidad de incidentes mientras que en el caso de los meses, julio fue el mes más violento de ese año.

Además, la figura 2 parece indicar que la cantidad de incidentes de acuerdo a los meses parece seguir una distribución normal. Al realizar las pruebas estadísticas de normalidad se obtiene que el *p-valor* es menor que el estadístico de prueba

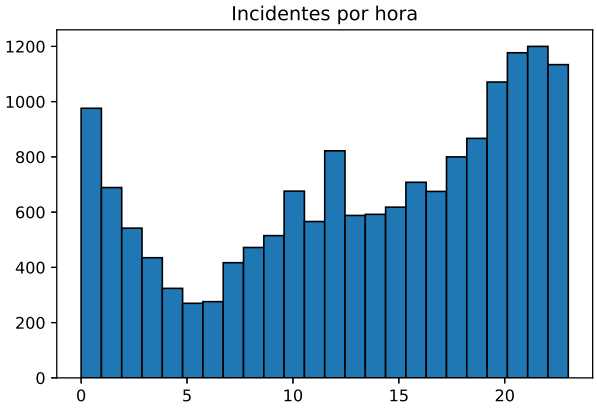

Figura 1: Cantidad de incidentes reportados por hora durante el año 2018.

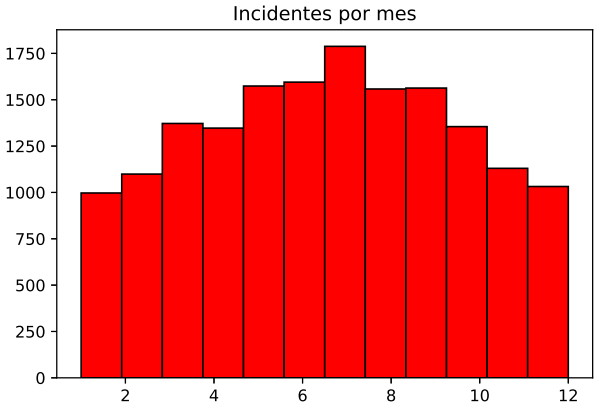

Figura 2: Cantidad de incidentes reportados por mes durante el año 2018.

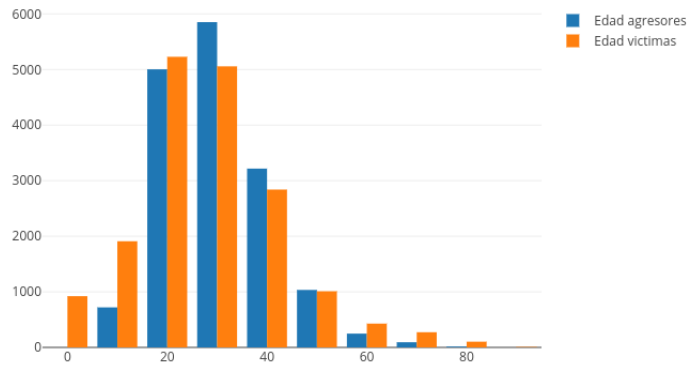

Figura 3: Rango de edades de víctimas y agresores.

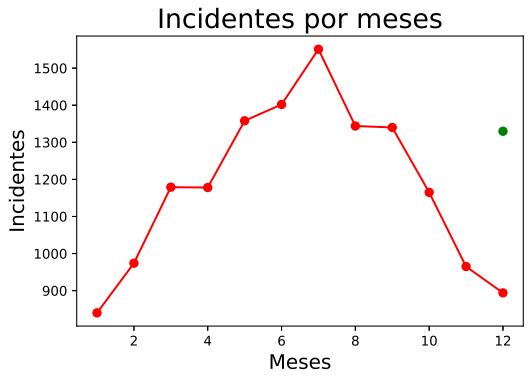

Figura 4: Prediccin del número de indicentes en el mes de diciembre usando el método de Holt.

por lo que *se rechaza la hipótesis de que los datos sigan una distribución normal*.

Para visualizar el rango de edades de víctimas y agresores, primero se agrupan en décadas para evitar obtener *peines* por la falta de información reportada. La Figura 3 muestra la gráfica que corresponde a esta información. De esta figura se resalta la cantidad de incidentes con vícimas entre 0-10 años pues se reportan 922 casos, además se reportan 111 casos de víctimas mayores de 80 años.

También es notorio que la mayoría de los incidentes ocurre con agresores y víctimas de edades de 30-50 años.

Una de las predicciones que se realiza con el método de Holt, utiliza los reportes de los meses de enero a noviembre y predice los reportes que se tendrn en el mes de diciembre, los resultados obtenidos se muestran en la Figura 4, donde los datos reales se grafican en rojo y la predicción en verde. No se logró identificar porqué la predicción no se ajusta muy bien a lo real, podría tratarse de un error al momento de procesar la información, sin embargo no se determinó el error.

Analizando únicamente los datos correspondientes a las víctimas con los factores edad, estado civil, escolaridad, hora y mes,

se realiza un análisis de varianza, aunque los resultados no parecen ayudar a concluir muchas cosas ya que los residuales son muy altos.

```
m = ols('cuantos ~ edad * edo_civil + edad * hora +
    edad * escolaridad', data = c).fit()
a = sm.stats.anova_lm(m, typ = 2)
print(a)
n = len(a)
alpha = 0.05
for i in range(n):
    print("{:s} {:s}es
        significativo".format(a.index[i], "" if
        a['PR(>F)'][i] < alpha else "NO "))
```

```
                      sum_sq     df          F
                        PR(>F)
edad              28047.938691    1.0  218.876307
    2.726351e-47
edo_civil         30247.078317    1.0  236.037623
    1.103524e-50
edad:edo_civil     6916.724660    1.0   53.975701
    2.843324e-13
hora               3020.502432    1.0   23.570945
    1.288347e-06
edad:hora           645.207120    1.0    5.034971
    2.493937e-02
escolaridad        6346.876698    1.0   49.528807
    2.599533e-12
edad:escolaridad    500.301989    1.0    3.904182
    4.829056e-02
Residual         282560.071260 2205.0        NaN
             NaN
edad es significativo
edo_civil es significativo
edad:edo_civil es significativo
hora es significativo
edad:hora es significativo
escolaridad es significativo
edad:escolaridad es significativo
Residual NO es significativo
```

Con esta misma información se realiza un análisis de componentes principales intentanto separar los datos de acuerdo a la escolaridad de la víctima. Para el caso del estado civil, se consideran solo *casado o soltero*. El resultado del análisis se observa en la Figura 5.

Un procesamiento que debe realizarse es analizar aquellos puntos que visualmente están separados del resto y estudiar cuál es la diferencia entre ellos, ya que este resultado no logra aportar mucho.

4. Conclusiones

Se presenta un análisis de incidentes de violencia intrafamiliar y comunitaria reportados en municipios del estado de Nuevo León.

Los resultados que se pudieron obtener de dicho análisis indican que hay un grupo específico de edades en víctimas y agresores que parecen ser el foco de las agresiones. Como trabajo a

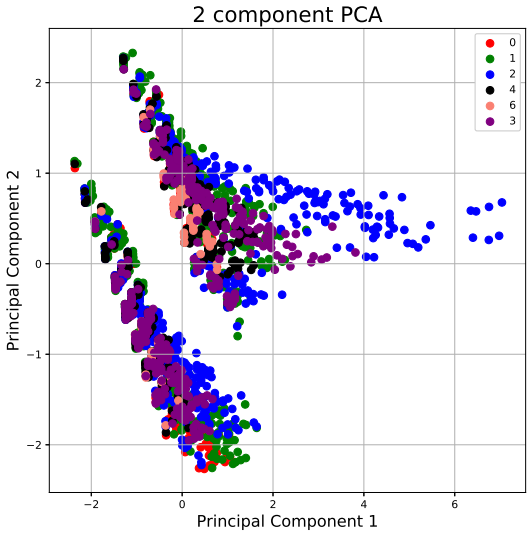

Figura 5: Análisis de componentes principales.

futuro se podría realizar un análisis específico para este grupo casos, así como también se pueden analizar aquellos casos en los que las víctimas son menores de edad ya que es inaceptable que existan casos de violencia con víctimas menores de 5 años al igual que los casos con víctimas mayores de 90.

El manejo de la información no resultó ser el adecuado ya que los resultados no permiten concluir de manera eficaz los factores que más influencia tienen en el número de incidentes, por lo que no se determinan posibles causas del problema.

Agradecimientos

Agradezco al Consejo Nacional de Ciencia y Tecnología (CONACyT) por la beca otorgada. A la doctora Patricia L. Cerda Pérez por proveer los datos utilizados en el estudio y a la doctora Elisa Schaeffer por la guía proporcionada para la realización de este trabajo.