Instituto tecnológico y de Estudios Superiores de Monterrey

Análisis de ciencia de datos

Marzo 2024

Adamaris Leticia De Dios Ramos A01643931
Santiago Mora Cruz A01369517
Adara Luisa Pulido Sanchez A01642450
Gabriel Reynoso Escamilla A01643561

1 Introducción

En México, donde una cultura machista predomina desde tiempos de la conquista, la violencia de género prevalence como un grave problema contemporáneo. Apesar del incremento de la preocupación y el interés de combatir esta problemática, cada año un gran número de mujeres son victimas de violencia.

La violencia de género se define como la realización de cualquier acto dañino dirigido hacia una persona o un grupo de personas en razón de su género esto es bastante consternante pues la violencia de género es una grave violación a los derechos humanos que puede llegar a poner la vida de la victima en riesgo. En el caso de las mujeres es cualuiqer acto de violencia que tenga o pueda tener como resultando daño o sufrimiento físico, mental o sexual hacia la misma, llegando a abarcar la violencia física, sexual, psicológica, entre otras [3].

La violencia psicológica consiste en dañar la estabilidad mental mediante insultos, abandono, entre otros, llevando a depresión, aislamiento y deterioro de la autoestima en la víctima, e incluso al suicidio. Por otro lado la violencia física implica cualquier acción intencional que cause daño, pudiendo ocasionar lesiones internas, externas o ambas. Así mismo la violencia economica toda acción u omisión que afecta la supervivencia económica de la víctima. Continuando la violencia sexual es cualquier acto que degrada o daña el cuerpo y/o la sexualidad de la Víctima y que por tanto atenta contra su libertad, dignidad e integridad física. La violencia familiar abuso intencional de poder para controlar o agredir a mujeres, ya sea física, verbal, psicológica, económica o sexualmente, dentro o fuera del ámbito familiar, por parte de personas con vínculos familiares o sentimentales [2].

2 Problemática

Una de las formas más comunes de violencia contra la mujer es aquella infligida por la pareja, es por esto que en el presente trabajo se busca, debido al gran impacto que tiene, indagar con profundidad las características que promueven esta forma de violencia. El propósito es construir un modelo predictivo que permita anticipar si las dinámicas y condiciones presentes en los hogares pueden determinar si una mujer está siendo víctima de violencia o no. Este trabajo busca ofrecer una herramienta efectiva para identificar y abordar de manera temprana y precisa los casos de violencia de pareja, contribuyendo así a la prevención y protección de las mujeres en situaciones de riesgo.

Para lograr este objetivo, el primer paso consistirá en identificar y recopilar información relevante, centrando el análisis en el contexto de la violencia de pareja y examinando los factores que la perpetúan. Además, se buscará desarrollar modelos predictivos que ayuden a determinar si una mujer está siendo

víctima de violencia, lo que permitirá abordar este problema de manera más efectiva.

3 Metodología

Para la exploración de este problema se usó como fuente de datos los resultados de la Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares (ENDIREH) de 2021 proporcionada por el INEGI. Esta consiste de 28 bases de datos que cubren desde identificadores personales, del hogar, hasta las condiciones en las que se desarrolló la entrevista además del descriptor de los archivos [1].

El manejo de una base de datos de este tamaño resultó ser difícil por lo que fue necesario una primera selección de las bases de datos que serían útiles de los 28 archivos en total solo 14 eran concernientes a la problematica a tratar, estas se unieron usando métodos de *pandas* y tomando como clave la variable ID_PER, una variable numérica que identifica a cada entrevistada, esta será el *Data Frame* maestro y como variable respuesta se eligió VPAR_12M que identifica "Condición de violencia total en el ámbito de pareja a lo largo de su relación actual o última en los últimos 12 meses".

Por el diseño de la entrevista y sus preguntas se permitía dejar un gran número de observaciones vacias en algunas preguntas [Fig. 1], para manejar esto primero se eliminaron las observaciones con datos faltantes en la variable respuesta, y los predictores si tenian datos faltantes los eliminamos del *Data Frame*, con excepciones, esto para no tener una imputación mal manejada y afectar el rendimiento del modelo.

Utilizamos Cross-validation para ver cual valor de C daría mejores resultados en el modelo, esto se realizó con un GridSearch, al realizar esto se encontró que no había mucha mejora en el modelo al cambiar este parámetro.

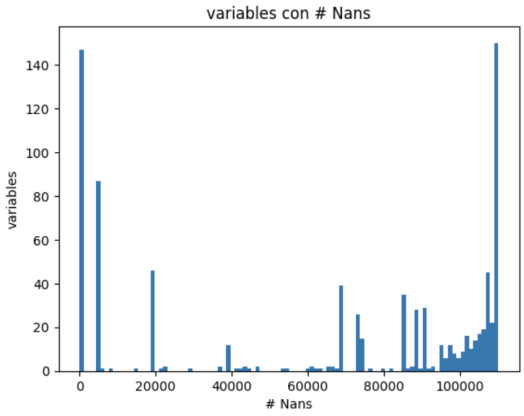

Figure 1: Cantidad de variables con cierto número de datos faltantes

Después fue necesario una selección de características con el método de envoltura, la cantidad de características que mejor resultados tenía era cinco pero tambien se realizaron modelaciones con veinte caracterñísticas.

Para los fines del proyecto se decidió hacer un modelo de regresión logística, este modelo además de realizar predicciones sobre la situación de violencia nos permite observar los coeficientes del modelo lo que permite un análisis de los predictores de esta situación.

4 Feature Engineering

La base de datos contenía alrededor de 849 columnas, ya que eran demasiadas columnas se realizó una selección de variables, esto fue de acuerdo a nuestra variable respuesta VPAR 12M "Condición de violencia total en el ámbito de pareja a lo largo de su relación actual o última en los últimos 12 meses". Se seleccionaron las variables que estaban relacionadas con la problemática a tratar, esto redució la base de datos a 190 columnas.

Se decidió eliminar las filas de la base de datos en la que la columna contenga como valor "b", ya que se encontró que se utilizaba para representar información en particular por parte de la entrevistada.La base de datos seguía teniendo muchas columnas por lo que se decidió disminuir aún más la base de datos, esta selección se hizo aún más específica a la variable de respuesta, lo que redujo la base de datos a 50 columnas, esto también lo hace más fácil de manipular. Se

utilizó el método de envoltura para poder hacer un feature selection. El método de envoltura se probó con 20 y 5 features, de aquí se realizó el feature selection con este método, para aplicar el modelo con las 20 y 5 features.

5 Modelado

Se aplicó el modelo de Logistic Regression, primero para la base de datos que se obtuvo después de reducirla a 50 variables que tenían relación con la variable respuesta, toma variables que deberían ser categóricas como númericas, entonces estas fueron separadas y con esto se codificaron las variables. Se realizó el entrenamiento de la base de datos y con esto un pipeline, el score que resultó de este modelo fue alrededor de 0.85. Después del feature selection se obtuvo del

método de envoltura 5 y 20 features y se aplicó el modelo de Logistic Regression a los dos resultados, se volvieron a codificar las varibles, entrenar el modelo y a crear un pipeline para los dos casos. Para 20 features se obtuvo un score alrededor de 0.8485 y para 5 features se obtuvo alrededor de 0.8454.

6 Resultados

El modelado usando 5 variables resultó en una regresión como la siguiente:

$$\begin{aligned} \text{logit}(p) = -0.64 + 6.58\text{V\_TOT1} + 2.96\text{V\_TOT9}... \\ ... - 0.39\text{P14\_1\_192} - 0.58\text{P14\_1\_191} \end{aligned} \tag{1}$$

Donde V_TOT1 y V_TOT9 es la respuesta afirmativa (1) o no especificada (9) a cualquier condición de violencia durante la vida de la entrevistada, P14_1_192 y P14_1_191 son amenazas de muerte a la entrevistada, la pareja o los hijos pocas (2) o muchas (1) veces. Este modelo obtuvo una precisión de 0.8454.

El modelado usando 20 variables resulto en una regresión como la siguiente:

$$\begin{aligned} \text{logit}(p) = -1.43.0.23\text{POB\_E\_A} + 0.13\text{POB\_L\_A}... \\ ... + 0.13\text{PE5\_12} + 0.03\text{PE5\_14} \end{aligned} \tag{2}$$

Donde POB_E_A es la variable que indica si la entrevista ha estudiado o no durante su vida, POB_L_A indica si la entrevista ha trabajado durante su vida, PE5_12 indica que la entrevistada no tuvo buena disposición para responder y PE5_14 indica intranquilidad durante la entrevista. Este model tuvo una presición de 0.8485 un poco más alto que aquel con 5 variables. Por su parte el modelo con las variables completas obtuvo un puntaje de 0.8536 el menor de los tres. Aunque la finalidad del modelo es predecir situaciones de violencia de pareja, debemos tomar en cuenta que los datos tratan situaciones complejas como lo son las emociones y relaciones humanas, pero podemos usar las funciones de probabilidad para rescatar cuales podrían ser indicadores de una situación de violencia. Por una parte la prescencia de cualquier tipo de violencia en la vida de las mujeres también la respuesta positiva a preguntas como "Cuando tienen relaciones sexuales la ha obligado hacer cosas que no le gustan" o "Llama o manda mensajes por teléfono todo el tiempo, para saber dónde y con quién está y qué está haciendo" y el nivel de educación y la situación laboral de la mujer .

7 Conclusión

Al realizar el feature selection con el método de envoltura y entrenar el modelo con los features seleccionados, se pudo observar que no hubo un gran cambio en el score. Al no obtener un gran cambio entre 20 y 5 features, se podría utilizar el modelo con menos características puede ofrecer ventajas en eficiencia computacional, pero se podrían realizar métodos diferentes que den un mejor score y obtener una mejor predicción. La implementación de un chatbot o línea telefónica para la prevención de la violencia de pareja con acceso directo a elementos de autoridad y seguridad es una medida crucial en la lucha contra este problema. Esta herramienta ofrece una vía de acceso rápida y confidencial para aquellas personas que enfrentan situaciones de violencia en sus relaciones.

8 Referencias

1. De Estadística Y, I. N. (s. f.). Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares (ENDIREH) 2021. https://www.inegi.org.mx/programas/endireh/2021/

2. Ley general de acceso de las mujeres a una vida libre de violencia, Reformada, Diario Oficial de la Federación [D.O.F.], 14 de junio de 2012, (México).

3. Preguntas frecuentes: Tipos de violencia contra las mujeres y las niñas. (s. f.). ONU Mujeres. https://www.unwomen.org/es/what-we-do/ending-violence-against-women/faqs/types-of-violence