De la Educación a la Prevención: Una Mirada a la Correlación entre los Niveles de Educación y la Violencia de Género

Juan Manuel Hernández Solano

Jose Carlos Yamuni Contreras

Miguel Steven Nguyen

I. Contexto de la Problemática

La violencia de género es un problema grave en México y ha sido objeto de atención en los últimos años, tanto a nivel nacional como internacional. Según datos del Instituto Nacional de Estadística y Geografía (INEGI), en 2020 se registraron más de 130, 000 denuncias de violencia de género en el país, lo que representa un aumento del 3.7 % en comparación con el año anterior. Las mujeres son las principales víctimas de este tipo de violencia, y las estadísticas muestran que las mujeres indígenas, afrodescendientes y con discapacidades son particularmente vulnerables.

Además de la violencia en sí misma, también es un problema la falta de acceso a la justicia y la impunidad que prevalece en muchos casos de violencia de género. Muchas mujeres no denuncian la violencia por miedo a represalias o porque no creen que el sistema de justicia les brindará protección. La falta de atención a este problema también se ve reflejada en la falta de recursos destinados a la prevención y atención de la violencia de género en México.

Estos sucesos han sido objeto de protestas y movilizaciones sociales en los últimos años, lo que ha llevado a la implementación de algunas medidas de protección y prevención. Sin embargo, la violencia de género sigue siendo un problema grave que requiere una atención constante y un compromiso real por parte de las autoridades y la sociedad en su conjunto para abordar sus causas y encontrar soluciones efectivas.

2. Problemática a Analizar

Antes de empezar a trabajar, es necesario definir la problemática a mano. A través de un análisis de datos buscamos encontrar una correlación entre el tipo de violencia que pueden sufrir las mujeres con

la escolaridad de su pareja.

Esto con la finalidad de mostrar la extensión y gravedad de la violencia contra las mujeres además de proponer y exigir diseños y seguimientos de políticas públicas orientadas a atender y erradicar la violencia contra las mujeres por razones de género.

Nuestra pregunta es: *¿Existe correlación entre el nivel de escolaridad de las parejas de las mujeres con el tipo de violencia que estas les ejercen?*

La hipótesis con que empezamos nos dice que sí. El ambiente en donde crecen ciertos tipos de personas afecta la manera en que tratan a los demás; alguien con menor educación tiene comportamientos diferentes a alguien más educada, y esta diferencia se puede manifestar como formas de violencia diferentes. Sin embargo, este razonamiento no es suficiente, y buscamos responder esta pregunta mediante el análisis de la información que tenemos.

2.1. Procedimiento

Primero, buscaremos clasificar el nivel de educación de las parejas de las mujeres abusadas de acuerdo a sus respuestas en el cuestionario, ya que la información recaudada es incompleta. Con un modelo entrenado en la información que ya tenemos, podemos predecir con cierto nivel de confianza la escolaridad.

Ya que tenemos este modelo de clasificación, podemos empezar a responder la pregunta planteada. Podemos hacer un análisis numérico de las respuestas de la encuesta, y ver en qué tipos de violencia se puede dividir la violencia total que sufre. Finalmente, vemos si existen alguna relación entre la escolaridad y los tipos de violencia.

3. Extracción de Características

Con una base de datos obtenido del INEGI, se realizó un procesamiento de datos. Este paso es necesario para poder evaluar la información proporcionada. El procesamiento de los datos se logró en 4 pasos:

3.1. Agregación de las bases de datos

Originalmente teníamos 28 tablas con diferentes contenidos de información. Seleccionamos solo 18, ya que estas contenían información de violencia de género. Posteriormente los unimos en un solo DataFrame.

3.2. Filtrado de información por estado

Redujimos aún más nuestro DataFrame, filtrando los datos solamente con personas del estado de Jalisco.

3.3. Selección manual de variables

Para este punto, aún teníamos un número de variables enormes, así que nos dimos a la titánica tarea de seleccionar las variables que, según nuestro criterio, serán útiles para nuestro modelo.

3.4. Conversión de variables categóricas

Finalmente, fue necesario limpiar los datos, reduciéndolos a variables categóricas. Las respuestas los volvimos binarias: 1, en el caso de respuestas positivas y 2 para respuestas negativas. En el caso de encontrar valores vacíos, se reemplazaban con una

función de probabilidad, dependiendo de la distribución de respuestas de la pregunta. Para nuestra variable objetivo, hicimos el mismo procedimiento, terminando con respuestas categóricas. Usaremos este modelo estocástico para un primer llenado de nuestro DataFrame, en un proceso conocido como "Data Augmentation". Esta es la información utilizada para entrenar los primeros modelos de clasificación.

Selección de Datos		
P4BC_2	P7_6_I	P7_6_2
2	2	2
2	2	2
2	2	2
9	2	2
2	I	2

Tabla I: Algunos de las variables dentro del DataFrame. La columna P4BC_2 lo utilizaremos como nuestro objetivo e inicialmente fue llenada con una función estocástica.

4. Modelación

Para la solución del problema, tuvimos que escoger dos modelos que hemos visto en clase; decidimos escoger KNN y Decision Trees. Trabajaremos con las respuestas de las preguntas como nuestras características, con nuestra etiqueta siendo el nivel de escolaridad de la pareja, columna `P4BC_2`. No fue necesario estandarizar los datos iniciales, ya que son variables categóricas.

4.I. KNN

El algoritmo k-Nearest Neighbors es un algoritmo de aprendizaje supervisado que se utiliza para la clasificación y la regresión.

El funcionamiento básico del algoritmo KNN es encontrar las k instancias más cercanas (vecinos) a una instancia de prueba dada y utilizarlas para predecir la clase o el valor de la instancia de prueba. La elección del valor de k es un parámetro importante del algoritmo y puede afectar significativamente la precisión de las predicciones.

En el caso de la clasificación, el algoritmo KNN asigna a la instancia de prueba la clase que es más común entre sus k vecinos más cercanos. En el caso de la regresión,

el algoritmo KNN asigna a la instancia de prueba el valor medio de los valores de sus k vecinos más cercanos.

El algoritmo KNN se utiliza comúnmente en aplicaciones de reconocimiento de patrones, minería de datos, análisis de imágenes, recomendación de productos, entre otros.

Vimos que nos sirve para responder la pregunta de interés ya que es un algoritmo de clasificación, y nos puede ayudar a hacer un contraste de precisión y costo con otros modelos.

4.1.1. Parámetros de ajuste

El principal parámetro que ajustar en este modelo es el número de vecinos, k, que se necesitan para clasificar.

4.2. Random Forest Classification

El algoritmo Random Forest es un modelo de aprendizaje automático supervisado utilizado para tareas de clasificación, regresión y otros problemas de análisis predictivo.

Consiste en la creación de múltiples árboles de decisión (Decision Trees) y la combinación de sus predicciones para obtener una predicción final. Cada árbol en el bosque se construye utilizando una muestra aleatoria de datos y un subconjunto aleatorio de características. Esto permite que cada árbol en el bosque tenga una cierta cantidad de diversidad, lo que ayuda a prevenir el sobreajuste.

Durante la fase de entrenamiento, cada árbol en el bosque se entrena en una muestra aleatoria de datos y un subconjunto aleatorio de características. Luego, durante la fase de predicción, las predicciones de cada árbol se combinan para formar una predicción final. En la tarea de clasificación, la predicción final es la clase más común predicha por los árboles individuales, mientras que en la tarea de regresión, la predicción final es la media de las predicciones de los árboles individuales.

El algoritmo Random Forest es popular debido a su capacidad para manejar una gran cantidad de características y datos faltantes, así como su capacidad para detectar y manejar el sobreajuste. Además, el modelo es fácil de usar y puede ser implementado rápidamente en grandes conjuntos de datos.

Escogimos este algoritmo de classificación ya que da buenos resultados sin ser tan computacionalmente caro en comparación de algoritmos como XGBoost, y MLPs.

4.2.1. Parámetros de ajuste

`n_estimators`: El número de árboles totales en el modelo.

`max_depth`: Controla la profundidad máxima de los árboles. Un árbol más profundo puede capturar patrones más complejos en los datos de entrenamiento, pero también puede aumentar el riesgo de sobreajuste.

`min_samples_split`: Especifica el número mínimo de muestras necesarias para dividir un nodo interno del árbol. Un valor más alto reduce el riesgo de sobreajuste, pero puede hacer que el modelo sea menos flexible.

`min_samples_leaf`: Especifica el número mínimo de muestras necesarias en una hoja del árbol. Un valor más alto reduce el riesgo de sobreajuste, pero puede hacer que el modelo sea menos flexible.

`max_features`: Especifica el número máximo de características que se conside-

ran para dividir un nodo. Un valor más bajo reduce el riesgo de sobreajuste, pero también puede reducir el rendimiento del modelo.

4.3. Determinando parámetros de ajuste

Decidimos cambiar las parámetros de ajuste del algoritmo KNN porque este se presta para la experimentación. En cambio, no decidimos probar con diferentes parámetros en el algoritmo de Random Forest porque la función nos regresa los parámetros en función de cómo está distribuida nuestra información y no tiene mucha utilidad cambiar los parámetros con valores arbitrarios. Tomamos valores de $k = 2$, $k = 75$, $k = 100$.

5. Resultados

5.1. KNN

Valor de k	Score
$k = 2$	0.6677
$k = 51$	0.7166
$k = 75$	0.7166
$k = 100$	0.7166

Tabla 2: Valores de k con el puntaje de sus modelos

En la tabla 2, podemos ver cómo cambia el rendimiento del modelo con diferentes valores de k. Observamos que, al aumentar este valor, el rendimiento no mejora después de cierto punto.

Escogimos el valor default de k para crear y evaluar la matriz de confusión. De la matriz, vemos el sesgo del modelo para ciertos valores, en este caso el 2.

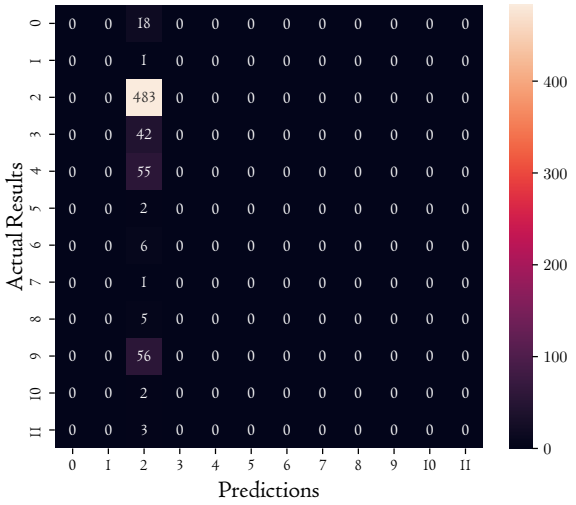

Figura 1: Matriz de confusión del modelo KNN

5.2. Random Forest

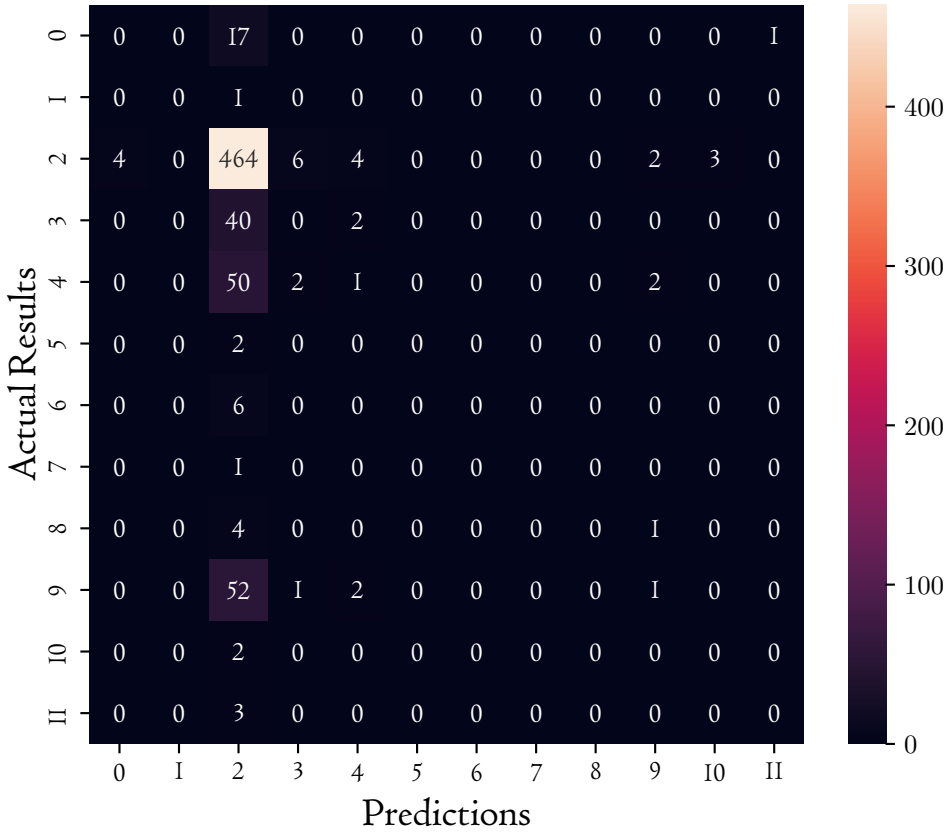

Figura 2: Matriz de confusión del modelo Random Forest

Con un puntaje de 0.6914, vemos que el modelo de Random Forest también se encuentra sesgada al mismo valor del modelo anterior, aunque podemos observar más variación en los errores.

5.3. Elección de Modelo

Decidimos quedarnos con Random Forest porque aunque KNN tiene un puntaje mayor, esto puede ser gracias al sobreajuste nos arroja una predicción muy sesgada en la que un solo valor predomina. Además, los modelos de Random Forest tienden a tener mejor rendimiento, con más hiperparámetros que ajustar.

6. Análisis Numérico

Ya teniendo una base de datos completa, gracias a nuestro modelo de clasificación, podemos empezar a buscar la correlación entre educación y violencia. Pero primero, necesitamos categorizar los tipos de violencia, y el impacto que tiene cada uno de ellos dentro de la encuesta.

6.1. Clasificación de Violencia

Identificamos cinco principales tipos de violencia: física, psicológica, sexual, económica y patrimonial. Cada una de estas violencias se encontraba representada dentro de la encuesta inicial, con cierto número de preguntas. A cada pregunta, le dimos un peso, de acuerdo a la cantidad de preguntas de su categoría; preguntas de categorías más grandes tienen menor peso, y vice versa.

Con estos pesos, pudimos dar un despliegue de los tipos y la cantidad de violencia que sufre cada mujer.

6.2. Agrupación por Escolaridad

Con estos porcentajes, podemos clasificar las violencias más comunes por nivel de educación. Estamos conscientes de que la violencia se puede manifestar de muchas formas diferentes, y que la generalización de este tema puede fomentar los prejuicios, sin embargo, creemos que este proyecto nos permite hacer un primer acercamiento a esta problemática.

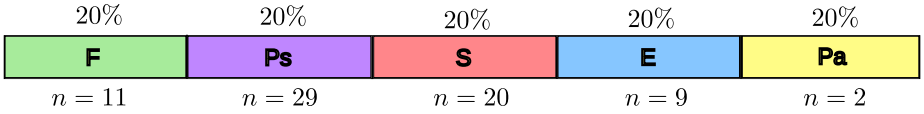

Figura 3: División de tipos de violencia con número de preguntas en la encuesta.

Violencia	Casos
Económica	1387
Psicológica	1268
No Violencia	379
Física	185
Sexual	148

Tabla 3: Tipos de violencia más comunes.

Escolaridad	Violencia
0: Ninguno	Psicológica
1: Preescolar	Económica
2: Primaria	Económica
3: Secundaria	Psicológica
4: Preparatoria	Psicológica
5: Primaria Técnica	Económica
6: Secundaria Técnica	Económica
7: Preparatoria Técnica	Psicológica
8: Primaria y Secundaria	Económica
9: Licenciatura	Psicológica
10: Carrera Professional	Psicológica
11: Posgrado	Psicológica
98: No Sabe	Psicológica

Tabla 4: Clasificación de violencia por educación.

7. Conclusiones

7.1. Interpretación de resultados

Los resultados que observamos en la tabla 4 nos mostraron que las mujeres que tienen parejas con un nivel de escolaridad bajo son más propensas a sufrir violencia económica proveniente de sus parejas, lo que nos indica que existe una correlación entre estas variables. A pesar que tipos de violencia específicos pueden variar según el grado de estudios, las mujeres siguen experimentando otras formas de violencia de género cotidianas. Esto nos da razón para creer que nuestra hipótesis era cierta: El ambiente en donde crecen ciertos tipos de personas afecta la manera en que tratan a los demás. Es importante recalcar que no estamos asociando estas conductas a la escolaridad per se, sino al contexto que una persona pudo haber tenido dependiendo su etapa de estudios. Además el hecho de que la violencia se siga presentando en todos los casos es preocupante. Los datos nos señalan que existe un problema estructural, donde en contextos diferentes de una misma cultura es un factor común la opresión hacia a las mujeres. Una opresión que les imposibilita tener una vida digna con igualdad de oportunidades a las de los hombres.

7.2. Alcance y Limitaciones

El principal alcance de nuestros datos es que pueden servir para mostrar la extensión y gravedad de la violencia contra las mujeres y esto a su vez puede resultar en el diseño y seguimiento de políticas públicas orientadas a atender y erradicar la violencia contra las mujeres por razones de género.

No obstante, nuestros resultados puede llegar a tener ciertas limitaciones:

Nos enfocamos en una parte de la encuesta muy específica y nuestros resultados pueden no ser representativos en situaciones diferentes. Así como también la confianza del modelo no es muy alta porque tuvimos muchos datos sesgados.

Además, la forma en la que seleccionamos las variables fue de forma manual y pudo haber sido con un método de selección de variables que quizá nos hubiera regresado variables que nos hubieran sido útiles.

Es importante tener en cuenta que la estadística no es un fin en sí mismo, sino una herramienta para ayudarnos a comprender el mundo y tomar decisiones informadas. Por lo tanto, es importante utilizar la estadística de manera rigurosa y ética, y reconocer sus limitaciones y sesgos potenciales.

7.3. Aplicaciones

Si bien una muestra de casi 4000 mujeres pareciera ser pequeña, en comparación con las 64 millones 540 mil 634 mujeres que reporta el Censo de Población y Vivienda de 2020 los resultados nos parecen estadísticamente significativos ya que los resultados se acercan. Si no nos ha bastado con las noticias, marchas y grupos activistas, debemos poner atención a los datos.

Algunas de las políticas públicas que se pueden implementar en México para crear espacios seguros para combatir la violencia de género pueden ser:

Campañas de concientización y educación: Se podrían implementar campañas de sensibilización en los medios de comunicación, redes sociales y en las escuelas, para prevenir la violencia de género y fomentar la igualdad de género. También se podrían promover espacios de diálogo y

debate sobre la violencia de género y su impacto en la sociedad.

Fortalecimiento de la participación política de las mujeres: Se podría implementar políticas para aumentar la representación de las mujeres en los cargos públicos y promover su participación activa en la toma de decisiones.

Fomento de la cultura de la denuncia: Es importante promover la denuncia de la violencia de género y garantizar que se investiguen y sancionen los casos de manera efectiva.

Creación de refugios y servicios de apoyo: Se podrían crear más refugios para mujeres que han sufrido violencia de género y servicios de atención médica y psicológica especializados para las víctimas.

Reforzar la legislación en materia de violencia de género: México ya cuenta con una Ley General de Acceso de las Mujeres a una Vida Libre de Violencia, pero es necesario fortalecer su implementación y garantizar que se cumpla. También se podrían implementar políticas para mejorar la atención a las víctimas y para sancionar con mayor severidad a los agresores.

Referencias

[1] *Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares*. URL: https://www.inegi.org.mx/programas/endireh/2021/#Documentacion.

[2] *Insumos Informativos*. URL: https://igualdad.jalisco.gob.mx/insumos-informativos/.