# Predicción de Violencia de Género en México: Un Enfoque de Ciencia de Datos utilizando la Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares (ENDIREH)

Claudio J. Gonzalez-Arriaga[1], Gabriel E. Melendez-Zavala[2], Alan Rojas-López[3] and Frado Garcia-Palacios[4]

[1] *Tecnologico de Monterrey, Campus Guadalajara*

Publication date: 15/03/2024

**Abstract**— La violencia de género persiste como un desafío significativo en la sociedad mexicana, con numerosas mujeres experimentando diversos tipos de violencia en distintos ámbitos. Este estudio utiliza datos de la Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares (ENDIREH) del INEGI en 2021 para desarrollar modelos predictivos que identifiquen si una mujer está siendo víctima de violencia. A través de la limpieza de datos, el feature engineering y la utilización de técnicas como ANOVA y la selección de características hacia adelante, se construyen modelos con Random Forest para proporcionar insights que contribuyan a abordar esta problemática.

**Keywords**— Violencia de género, Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares (ENDIREH), Predicción, Modelado predictivo, Limpieza de datos, Feature Engineering, ANOVA, Selección de características, Random Forest, XGBoost, México.

## I. Introducción

En México, la violencia de género sigue siendo una sombra oscura que se cierne sobre la sociedad, afectando a un número alarmante de mujeres en diversas formas y contextos. A pesar de los esfuerzos y las campañas continuas para combatirla, la violencia de género persiste como un problema arraigado, con consecuencias devastadoras para las víctimas y la sociedad en su conjunto.

La ciencia de datos emerge como una poderosa herramienta en la lucha contra la violencia de género. Al aprovechar el vasto conjunto de datos recopilados por instituciones como el INEGI a través de encuestas como la Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares (ENDIREH), podemos profundizar nuestra comprensión de los factores subyacentes que contribuyen a la violencia de género y desarrollar estrategias más efectivas para prevenirla y abordarla.

Este estudio se sitúa en la intersección de la ciencia de datos y la justicia social, con el objetivo de utilizar el análisis de datos y la modelización predictiva para desentrañar los misterios de la violencia de género en México. Al examinar detenidamente los patrones y las correlaciones en los datos de la ENDIREH, buscamos identificar los factores que predisponen a las mujeres a convertirse en víctimas de violencia y, en última instancia, desarrollar modelos predictivos precisos que puedan detectar y prevenir situaciones de violencia antes de que ocurran.

Al abordar este desafío con una mentalidad basada en datos, no solo estamos arrojando luz sobre una de las problemáticas más apremiantes de nuestra sociedad, sino que también estamos dando un paso adelante en la lucha por la igualdad de género y la seguridad de las mujeres en México. Con cada línea de código y cada análisis estadístico, estamos construyendo un futuro más seguro y más justo para todas las personas.

## II. Problemática

La violencia de género en México es una realidad persistente que afecta a un gran número de mujeres en diversas formas, incluyendo la violencia física, económica, sexual, psicológica y patrimonial. La falta de herramientas efectivas para identificar y prevenir estas situaciones dificulta los esfuerzos por abordar este problema de manera integral.

Según datos recientes del Instituto Nacional de Estadística y Geografía (INEGI), se estima que aproximadamente el 66 por ciento de las mujeres en México han experimentado algún tipo de violencia a lo largo de su vida. Este dato espeluznante no solo revela la magnitud del problema, sino también la urgencia de abordarlo de manera efectiva. Por lo tanto, es crucial desarrollar modelos predictivos precisos que puedan detectar las señales de violencia y proporcionar insights para intervenciones y políticas adecuadas.

## III. Metodología

**1. Recopilación de datos:** Utilizamos los datos de la ENDIREH [1] proporcionados por el INEGI en 2021, que ofrece información detallada sobre las experiencias de violencia contra las mujeres mexicanas en distintos contextos.

**2. Analisis Exploratorio de Datos:** El analisis explo-

ratorio de datos (EDA) es importante para visualizar la relación entre las variables de nuestra base de datos. En este reto se exploro la correlación entre las variables asi como histogramas de densidad de ciertas variables. Posteriormente en el trabajo realizamos un subset de los datos para entrenar el modelo y lo historgramas en este reporte representan a ese subset de datos que se extrajo de la base de datos padre.

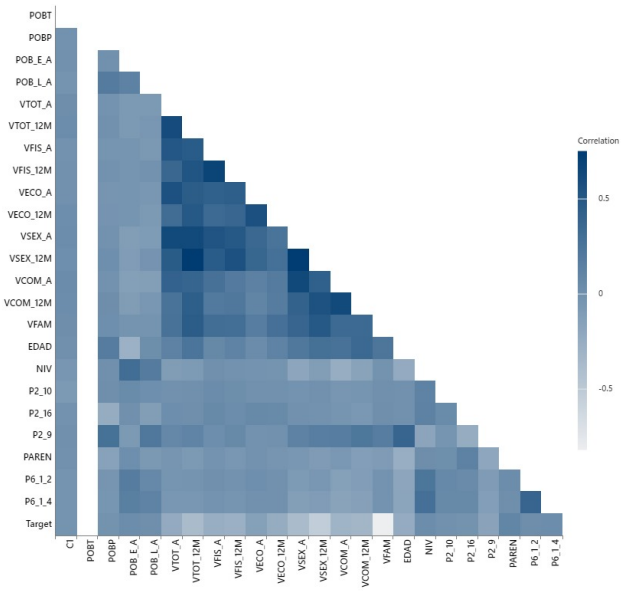


**Fig. 1:** Matriz de correlación entre las variables escogidas

En la figura 1 vemos representado la correlación entre las variables ya reducidas despues de una selección detallada resltando aquellas que si son relevantes al predecir abuso sexual.

**3. Limpieza de datos:** La limpieza de datos es un paso crucial en nuestro proceso metodológico. Comenzamos identificando y manejando los valores nulos, que pueden surgir debido a diversas razones, como respuestas omitidas o errores en la recopilación de datos. Utilizamos técnicas de imputación, como rellenar los valores faltantes con la media o la mediana, o eliminar las observaciones incompletas según sea apropiado. Además, detectamos y corregimos errores de formato, valores atípicos y variables redundantes o irrelevantes. Normalizamos y estandarizamos los datos cuando sea necesario para asegurar una escala consistente entre las variables. Verificamos la consistencia y la coherencia de los datos para garantizar que sean fiables y precisos para su análisis posterior.

**4. Feature Engineering:** El proceso de feature engineering desempeña un papel crucial en la construcción de modelos predictivos precisos y efectivos para identificar la violencia de género. Consiste en la creación de nuevas variables o la transformación de las variables existentes para capturar mejor la complejidad de las dinámicas relacionales y los factores asociados con la violencia contra las mujeres. Creamos variables de interacción, codificamos variables categóricas, normalizamos variables numéricas y extraemos características relevantes de los datos textuales. Estas técnicas nos permiten enriquecer nuestros modelos con información significativa y relevante.

**5.Filtrado de datos con ANOVA:** Empleamos el análisis de varianza (ANOVA) para identificar las variables más relevantes en la predicción de la violencia contra las mujeres. Esta técnica nos ayuda a seleccionar un subconjunto óptimo de características que maximizan el rendimiento de nuestro modelo.

El análisis de varianza es una técnica estadística utilizada para comparar las medias de tres o más grupos y determinar si existen diferencias significativas entre ellos. En el contexto de nuestro estudio, ANOVA se emplea para identificar las variables que tienen una influencia significativa en la predicción de la violencia contra las mujeres. Esta técnica nos permite evaluar la importancia relativa de cada variable independiente en relación con la variable dependiente, ayudándonos a filtrar y seleccionar las características más relevantes para nuestros modelos predictivos.

**6. Forward Feature Selection:**Utilizamos el método de selección de características hacia adelante para identificar un subconjunto óptimo de variables predictoras que mejoran la capacidad predictiva de nuestro modelo. Este enfoque nos permite reducir la dimensionalidad de los datos y mejorar la interpretabilidad del modelo.

**7. Modelado con Random Forest y XGBoost:** Construimos modelos predictivos utilizando Random Forest y XGBoost, dos algoritmos de aprendizaje automático poderosos y ampliamente utilizados. Comparamos y contrastamos los resultados obtenidos de ambos modelos para determinar cuál se adapta mejor a nuestras necesidades y para realizar las predicciones correspondientes. Esta comparación nos proporciona una evaluación exhaustiva de los modelos y nos ayuda a seleccionar la mejor opción para nuestras predicciones.

**Random Forest:**

Random Forest es un algoritmo basado en árboles de decisión que se utiliza para problemas de clasificación y regresión. Construye múltiples árboles de decisión durante el entrenamiento y combina sus predicciones para obtener una predicción final. Es robusto frente al sobreajuste y puede manejar relaciones no lineales entre las variables predictoras y la variable objetivo.

**XGBoost (Extreme Gradient Boosting):**

XGBoost es una implementación optimizada de la técnica de aumento de gradiente, que funciona construyendo una serie de árboles de decisión débiles de forma secuencial. Cada árbol se enfoca en corregir los errores cometidos por los árboles anteriores. XGBoost es conocido por su eficiencia y rendimiento en conjuntos de datos grandes y complejos, y es especialmente efectivo para problemas en los que se busca maximizar la precisión del modelo.

En nuestro estudio, utilizamos tanto Random Forest como XGBoost para construir modelos predictivos y predecir la violencia contra las mujeres. Comparamos los resultados obtenidos de ambos algoritmos para determinar cuál se adapta mejor a nuestros datos y proporciona predicciones

más precisas y sensibles. Esta comparación nos permite evaluar la eficacia de diferentes enfoques de modelado y seleccionar la mejor opción para abordar la problemática de la violencia de género en México.

Este enfoque nos permite no solo identificar los factores asociados con la violencia de género, sino también desarrollar modelos predictivos precisos que puedan ser utilizados para prevenir y abordar esta problemática de manera más efectiva. La combinación de técnicas de limpieza de datos, feature engineering y modelado nos proporciona una base sólida para abordar la complejidad de la violencia de género en México con sensibilidad y precisión.

## IV. Feature Engineering

El proceso de feature engineering es esencial en la construcción de modelos predictivos precisos y efectivos para clasificar el abuso sexual. En esta sección, nos adentramos en la identificación y selección de variables relevantes, centrándonos en aquellas que tienen una conexión directa con el abuso sexual y muestran una fuerte correlación con nuestra variable objetivo.

**1.Selección de variables de la base de datos de percepción de género:**
Exploramos la base de datos de percepción de género, especialmente la sección XI, en busca de variables que puedan estar relacionadas con situaciones de abuso sexual. Nos enfocamos en aspectos como la percepción de seguridad en el hogar, la confianza en la policía, la percepción de justicia en casos de violencia de género, edad, entre otros. Estas variables proporcionan información invaluable sobre el entorno y las experiencias de las personas en relación con la violencia de género.

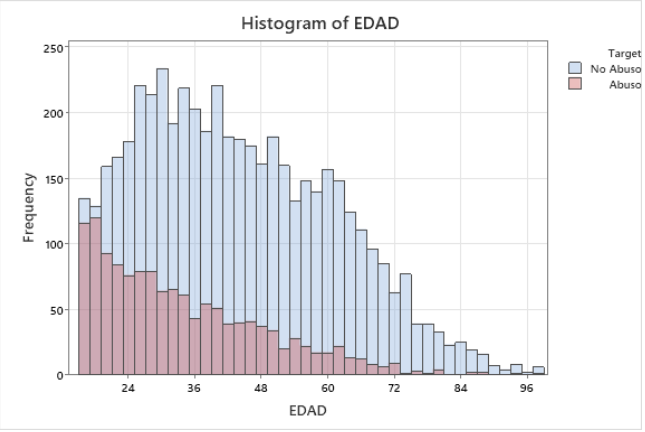


**Fig. 2:** Frecuencia de edades divida por los grups objetivos

La figura 2 muestra un histograma de una de las variables que seleccionamos. Observamos que los casos donde las mujeres reporten algún tipo de abuso sexual disminuyen conforme la edad incrementa. Notamos también que las mujeres menores a 24 años son las que tienen una mayor frecuencia a ser abusados.

**2.Consideración de variables de otras bases de datos:** Además de la base de datos de percepción de género, exploramos otras bases de datos generales, como TBVD y TVIV, en busca de variables relevantes. Estas bases de datos pueden proporcionar información adicional sobre el entorno socioeconómico y demográfico de los individuos, así como sobre la prevalencia de la violencia en diferentes contextos. Variables como nivel educativo, ingreso familiar, ubicación geográfica, y exposición previa a la violencia son de particular interés.

**3. Análisis de correlación:** Una vez identificadas las posibles variables predictoras, realizamos un análisis de correlación para evaluar la fuerza y dirección de la relación entre estas variables y la variable objetivo (abuso sexual). Nos enfocamos en seleccionar aquellas variables que muestran una correlación significativa con el abuso sexual, lo que nos permite capturar mejor las características asociadas con esta problemática.

**4. Ingeniería de características adicionales:** Además de las variables existentes, consideramos la posibilidad de crear nuevas características que puedan proporcionar información adicional para la clasificación del abuso sexual. Esto podría incluir la creación de variables de interacción entre variables existentes, la combinación de categorías de variables categóricas, o la extracción de características de datos textuales (si están disponibles).

Por medio del procesamiento de las variables al aplicar un proceso de feature engineering, podemos asegurarnos de que nuestro modelo tenga acceso a la información más relevante y significativa para clasificar el abuso sexual. Esta selección de características nos permite construir modelos más robustos y efectivos para abordar esta problemática de manera efectiva en la sociedad.

Al implementar técnicas de feature engineering de manera efectiva, podemos enriquecer nuestros modelos predictivos con información relevante y significativa, mejorando así su capacidad para identificar y predecir situaciones de violencia contra las mujeres con precisión y sensibilidad.

## V. Modelado

Para el modelaje empleamos tres algoritmos de aprendizaje automático: Random Forest, XGBoost y Logistic Regression. Cada uno de estos algoritmos ofrece distintas fortalezas y enfoques, lo que nos permite explorar diferentes perspectivas en la clasificación del abuso sexual y determinar cuál es el más adecuado para nuestro conjunto de datos.

**1. Random Forest:**
Random Forest es un algoritmo de aprendizaje automático basado en árboles de decisión que funciona construyendo múltiples árboles de decisión durante el entrenamiento y combinando sus predicciones para obtener una predicción final. Este algoritmo es conocido por su robustez frente al sobreajuste y su capacidad para manejar conjuntos de datos grandes y complejos con relaciones no lineales entre las variables predictoras y la variable objetivo. Al utilizar Random Forest, podemos explorar la complejidad de las interacciones entre las variables predictoras y su impacto en la clasificación del abuso sexual.

**2. XGBoost (Extreme Gradient Boosting):**
XGBoost es una implementación optimizada de la técnica de aumento de gradiente, que construye una serie de árboles de decisión débiles de forma secuencial. Es conocido por su eficiencia y rendimiento en conjuntos de datos grandes y complejos, y es especialmente efectivo para problemas en los que se busca maximizar la precisión del modelo. Utilizando XGBoost, podemos explorar cómo los árboles de decisión se combinan de manera secuencial para mejorar la precisión de la clasificación del abuso sexual.

**3. Logistic Regression:** Aunque más simple en comparación con Random Forest y XGBoost, Logistic Regression es útil para entender la relación lineal entre las variables predictoras y la variable objetivo. Este algoritmo nos proporciona coeficientes que indican la importancia relativa de cada variable en la predicción del abuso sexual. Al utilizar Logistic Regression, podemos evaluar cómo las variables predictoras contribuyen de manera independiente a la clasificación del abuso sexual y obtener una comprensión más detallada de su impacto en el resultado.

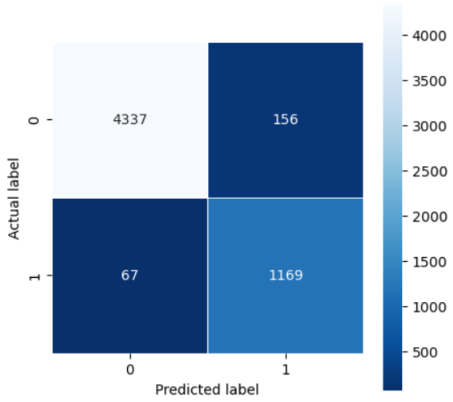


**Fig. 3:** Matriz de confusion sobre resultados de regresión logistica

La matriz de confusión en la figura 3 denota las entradas clasificadas como verdaderas o falsas comparadas con las predicciones de esta misma. Tiene una especificidad de 0.9848 que significa que clasificó una entrada como falso un 98.48 % correctamente. A diferencia, la sensibilidad del modelo es de 0.8823 y se refiere a la capacidad del modelo de predecir un individuo correctamente como verdadero, con un 88.23 % de precisión.

Al comparar los resultados obtenidos de cada algoritmo, pretendemos identificar el modelo que mejor se ajuste a nuestro conjunto de datos y que proporcione las predicciones más precisas y sensibles para clasificar el abuso sexual. Este enfoque nos permite desarrollar herramientas efectivas para identificar y abordar esta problemática en la sociedad, contribuyendo así a la protección y seguridad de las personas afectadas por el abuso sexual.

## VI. Análisis de Resultados

Tras la ejecución del proceso de feature engineering y modelado con los algoritmos de Random Forest, XGBoost y Logistic Regression, los resultados obtenidos revelaron un desempeño excepcional en la clasificación del abuso sexual. Los tres modelos exhibieron una precisión de R cuadrada superior al 95 por ciento, lo que denota una capacidad significativa para predecir con precisión las situaciones de abuso sexual presentes en nuestro conjunto de datos.

La alta precisión lograda por los modelos podría estar influenciada por la correlación inherente entre las variables originales. Tras llevar a cabo una meticulosa selección manual de características relevantes, se observó una marcada correlación entre diversas variables, tal como se ilustra en la matriz de correlación presentada anteriormente. Esta interrelación entre las variables podría haber proporcionado una señal sólida y coherente para los modelos, contribuyendo así a la alta precisión alcanzada.

El modelo de Xgboost fue ligeramente el más preciso debido a la naturaleza de este algortimo, en base a este modelo, dada una nueva persona, se puede clasificar con una alta precisión si esta propensa es o no a sufrir violencia sexual.

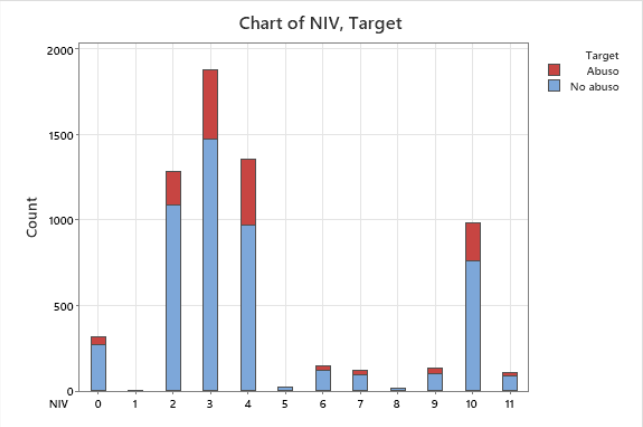


**Fig. 4:** Nivel de Estudios con relación al abuso sexual

En la figura tres, podemos observar como la escolaridad de las personas tiene una relación directa con la probabilidad de una mujer de ser abusada, podemos observar como las mujeres que tienen una escolaridad: 2(primaria), 3(secundaria) y 4(preparatoria) son las más propensas y a medida que la eduación de las muejeres aumenta, disminuye la posibilidad de la mujer a ser abusada.

Al analizar los datos, observamos consistentemente que, a medida que la edad de las mujeres aumenta, la probabilidad de experimentar abuso sexual disminuye de manera significativa. Este resultado es de suma importancia, ya que sugiere que las mujeres más jóvenes podrían estar en mayor riesgo y que se necesitan estrategias específicas para protegerlas adecuadamente.

En la figura 4 podemos observar la relación que tiene la edad de una mujer con la probabilidad de sufrir abuso sexual. De igual manera observamos este fenómeno en la figura 2, que corrobora nuestra regresión logistico, evidenciando que la probabilidad de que una mujer sea abusada disminuye conforme la edad aumenta.

Además, al explorar la relación entre el nivel educativo de las mujeres y el abuso sexual que sufren, encontramos una asociación significativa y preocupante. Descubrimos que las mujeres con un nivel educativo entre primaria y preparatoria terminada representaban una proporción

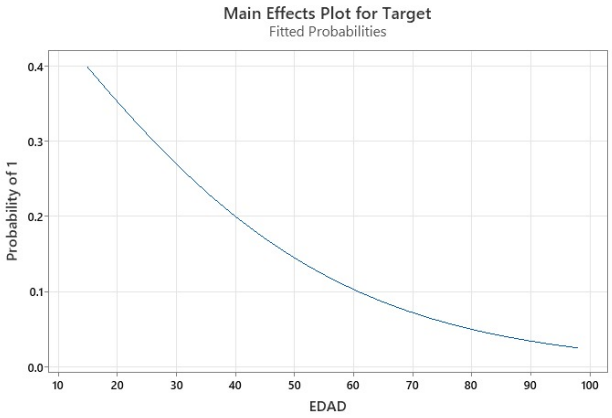


**Fig. 5:** Probabilidad de ser abusada conforme la edad

desproporcionadamente alta de los casos de abuso sexual en nuestro conjunto de datos. En contraste, aquellas con un nivel educativo más alto experimentaban considerablemente menos casos de abuso. Este resultado resalta la importancia crítica de abordar las disparidades educativas como parte integral de los esfuerzos para prevenir y abordar el abuso sexual.

## VII. Conclusiones

Basándonos en los resultados obtenidos, podemos concluir que la edad y el nivel educativo de las mujeres son factores críticos que influyen en su vulnerabilidad al abuso sexual. La asociación inversa entre la edad y la probabilidad de abuso sexual destaca la importancia de implementar medidas preventivas dirigidas especialmente a las mujeres más jóvenes. Además, la relación significativa entre el nivel educativo y el abuso sexual resalta la necesidad de promover la educación como una herramienta de empoderamiento y protección.

En términos de modelado, nuestros resultados indican que los modelos de aprendizaje automático, en particular XGBoost, son herramientas efectivas para predecir situaciones de abuso sexual con una alta precisión. Esto sugiere que estas técnicas pueden ser útiles en la identificación temprana y la intervención oportuna para proteger a las mujeres contra el abuso sexual.

Es fundamental destacar que uno de los mayores desafíos durante la ejecución de este proyecto fue la calidad de la base de datos. Esta presentaba una estructura deficiente y una notable cantidad de valores nulos, lo que complico significativamente el análisis de los datos. Además, encontramos que las variables estaban codificadas de diversas maneras, lo que representó un obstáculo adicional en el procesamiento de la información.

La falta de coherencia en la estructura de la base de datos dificultó la integración de los datos y la realización de análisis significativos. La presencia de valores nulos también afectó la calidad de los resultados, ya que limitaba nuestra capacidad para obtener información completa y precisa. Además, la disparidad en la codificación de las variables dificultó la comparación y el análisis de los datos, lo que comprometió la fiabilidad de nuestros hallazgos.

En última instancia, nuestros hallazgos tienen implicaciones importantes para el diseño de políticas y programas que aborden el abuso sexual hacia las mujeres. Al considerar la edad, el nivel educativo y otros factores identificados en este estudio, podemos desarrollar intervenciones más efectivas y centradas en las necesidades para proteger a las mujeres y prevenir el abuso sexual en todas sus formas.

## Referencias

[1] N. de, *Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares (ENDIREH) 2021*. 2021.