Tecnológico de Monterrey

Predictor sobre la Dinámica de las Relaciones en los Hogares

Johanna C. Willis-Ruiz[1], Victoria González-González[2], Sara Rivera-Méndez[3], Valeria Aguilar-Meza[4] and Karla R. Munguía-Romero[4]

[1] *Tecnologico de Monterrey, Escuela de Ingeniería y Ciencias, Campus Guadalajara*

Fecha: 14/03/2024

Abstract—En el presente reporte se analizan los resultados de la Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares realizada por el INEGI para construir un modelo predictivo que determine si una mujer está siendo víctima de violencia en base a diversos factores que describen la dinámica en su hogar. La metodología utilizada se divide en tres secciones: *data engineering*, *feature selection* y modelado. En la primera sección se realizó la exploración y limpieza de los datos, el s*ampling distribution*, se observó el sesgo de los datos, se creó la variable objetivo. En la segunda sección se utilizaron tres métodos de filtrado y uno de envoltura. Por último, en la tercera sección se aplicaron dos modelos, de los cuales *Random Forest* demostró tener un mejor desempeño ya que con este se obtuvo una mayor exactitud. Este modelo tuvo una sensibilidad del 20%, una especificidad del 89.67%, un F1-score del 21.57%, una precisión del 38%, y una exactitud del 72.92%. Por otro lado, se detectó que las variables más significativas en la presencia de esta problemática son: número de focos en la vivienda, número de personas en vivienda, la edad a la que tuvieron su primera relación sexual, la edad de cuando se juntaron con su pareja actual, y estado civil. En base a este análisis se pudo concluir que los resultados obtenidos muestran una clave directriz acerca de los enfoques que las reformes públicas y organizaciones gubernamentales deben tomar para mejorar la situación de violencia de género del país.

Keywords—violencia, mujer, pareja, predicción, vivienda.

I. Introducción

En el presente reporte se analizan los resultados de la Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares (ENDIREH) realizada por el INEGI a mujeres mexicanas de 15 años en adelante en 2021. El objetivo es construir un modelo predictivo que sea capaz de determinar si una mujer está siendo víctima de violencia en base a diversos factores que describen la dinámica en su hogar. La base de datos utilizada contempla experiencias de violencia física, económica, sexual, psicológica y patrimonial en diversos contextos. Sin embargo, el enfoque de este entregable es la violencia en pareja. Por consiguiente, la delimitación de la información relevante para este ámbito es esencial para realizar el modelo adecuadamente.

II. PROBLEMÁTICA

La violencia contra la mujer ejercida por su pareja infringe los derechos humanos de las mujeres y se considera un grave problema social y de salud pública. En México, el 43.9% de mujeres de 15 años en adelante han sufrido agresiones por parte de su pareja más reciente [1]. Además, se ha demostrado que esta problemática afecta significativamente la salud física y mental de las víctimas, a corto y largo plazo, ya que afecta su funcionamiento social y suele impactar de manera negativa su situación financiera porque, dependiendo del nivel de violencia sufrida, las víctimas pueden perder su capacidad para trabajar y generar ingresos activos [2]. Por otra parte, la depresión, la ansiedad, el estrés postraumático y la sintomatología de tipo somático son los problemas de salud mental de mayor presencia en mujeres violentadas por sus parejas son: depresión, ansiedad, estrés postraumático y sintomatología de tipo somático [2]. En el caso de violencia psicológica, las secuelas pueden incluir pérdida del apetito y trastornos del sueño.

Los factores más influyentes en la presencia de este problema son el estado residencial, sector socioeconómico, consumo frecuente de alcohol por parte de la pareja, si la mujer es migrante o no, y el apoyo social que tiene [3].

III. METODOLOGÍA

a. Data Engineering

1. Exploración de los datos

En priemra instancia, se intentó juntar en una sola base de datos, los 28 archivos proporcionados por el INEGI, sin embargo, no contamos con la capacidad computacional necesaria para hacerlo, ya que, la plataforma "Colab" en la que trabajamos a lo largo de todo este reto, no contaba con la suficiente memoria RAM para realizar el combinado de las bases de datos, por lo mismo, se decidió elegir las seis bases de datos con base en la temática que deseamos específicamente estudiar para este trabajo: la violencia en pareja sufrida por la mujer.

Con esto en mente, se eligieron los siguientes archivos para conformar nuestra nueva base de datos:

1. TVIV: tabla de información acerca de la vivienda de la

entrevistada y de las personas que viven con ella.

2. TSDem: contiene las características sociodemográficas, económicas y culturales de cada integrante de la vivienda.
3. TB_SEC_III: describe la situación conyugal de las mujeres de 15 años o más de la vivienda.
4. TB_SEC_IV: contiene información básica de la pareja, esposo, novio o ex-pareja que no reside en la vivienda de la mujer elegida.
5. TB_SEC_XIII: contiene información sobre las características de las relaciones de pareja de las mujeres de 15 años o más.
6. TB_SEC_XIV: contiene información sobre la relación de pareja actual o última de las mujeres de 15 años y más, y de las situaciones de violencia derivadas de las posibles experiencias de agresiones, emocionales, económicas, físicas o sexuales.

Una vez unidas todas las bases de datos pasamos a trabajar con la misma.

2. Limpieza de los datos

Para poder tener una mejor interpretación de las variables, es necesario iniciar por limpiar nuestras columnas y sus registros, por lo que para no sacrificar tantas filas, se sacó un porcentaje individual de cada columna de sus respectivos resgistros con valores nulos, y aquelllas variables con un porcetaje mayor al 10% de cada archivo antes mencionado, fueron eliminadas de la base de datos.

Una vez haciendo esto, nos quedamos con únicamete con 172 columnas de haber tenido en un principio más de 300 variables.

Después, se eliminaron por completo aquellas columnas con un porcentaje menor al 10%, donde solo se perdieron pocos registos y nuestra base de datos ya no tenía ninguna fila con datos vacíos.

Como parte final de la limpieza, se eliminaron tanto los registros duplicados así como las columnas duplicadas de la base de datos, es decir, aquellas variables que tenían en común los distintos archivos; quedándonos de esta manera con únicamente 82 variables a estudiar.

3. Sampling Distribution

El *sampling distribution* es la distribución de frecuencias de un estadístico muestral sobre muchas muestras extraídas del conjunto de datos. Los pasos para hacer esto fueron: obtener una muestra, computar una métrica estadística en esa muestra (como la media) y registrarla, repetir estos pasos muchas veces, y realizar un histograma que muestre la distribución de la métrica. Según el teorema de límite central, el tamaño de muestra es directamente proporcional a la tendencia de los datos a adoptar una distribución normal.

Como se puede ver en la mayoría de los histogramas anteriores, ocurre un claro sesgo positivo o hacia la derecha en las gráficas. Muchos factores contribuyen a este sesgo,

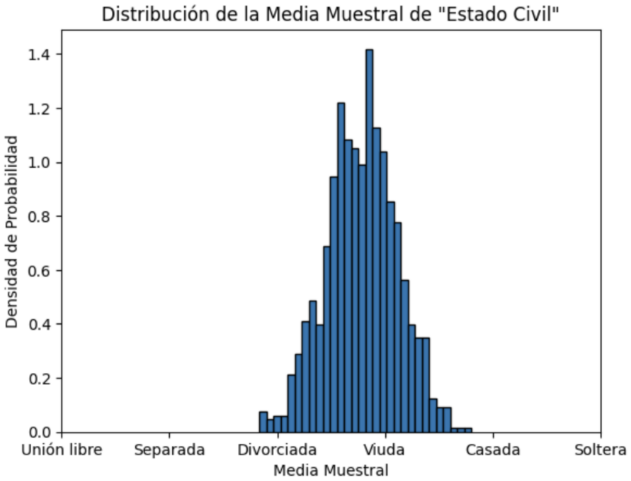

Fig. 1: Sample distribution de "Estado Civil"

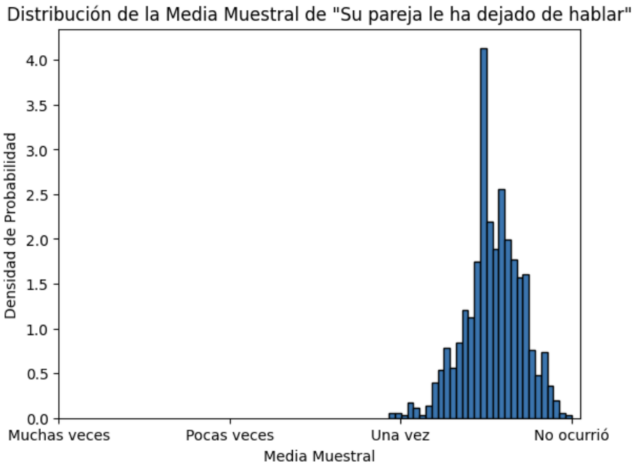

Fig. 2: Sample distribution de "Su pareja le ha dejado de hablar"

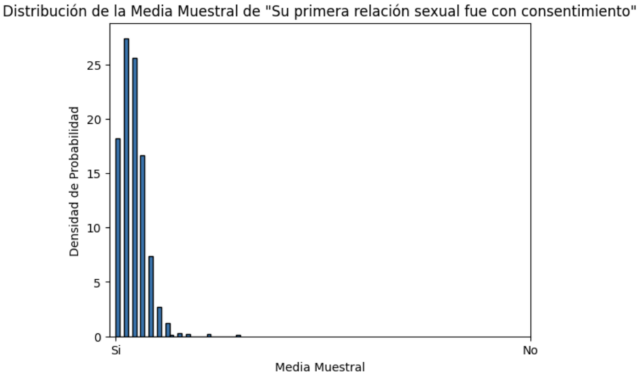

Fig. 3: Sample distribution de "Su primera relación sexual fue con su consentimiento"

como la distribución de la población mexicana, que hace que los valores se inclinen más a los jóvenes, la aplicación de la encuesta en mayormente ciudades dejando a un lado una muy grande parte de la población rural, así como los datos extremos que afectan severamente la distribución de los histogramas. En general, nuestros datos tienen un sesgo postivo por las características generales de la mujer mexicana, así como el tipo de persona que tendría la disponibilidad para poder participar en la encuesta.

4. Sesgo

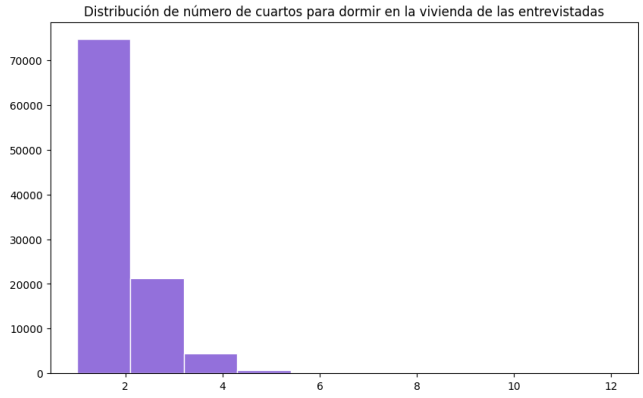

Fig. 4: Cuartos para dormir en la vivienda de las entrevistadas

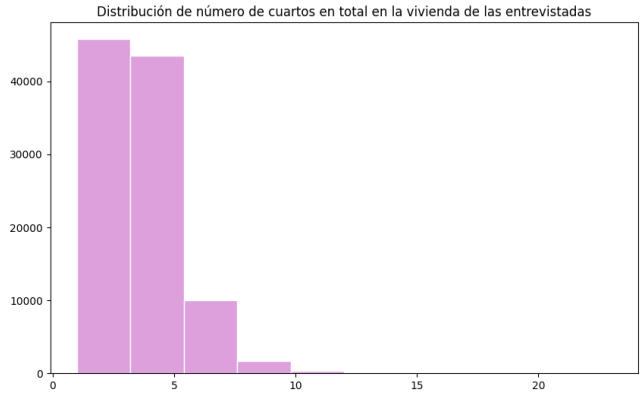

Fig. 5: Cuartos en total en la vivienda de las entrevistadas

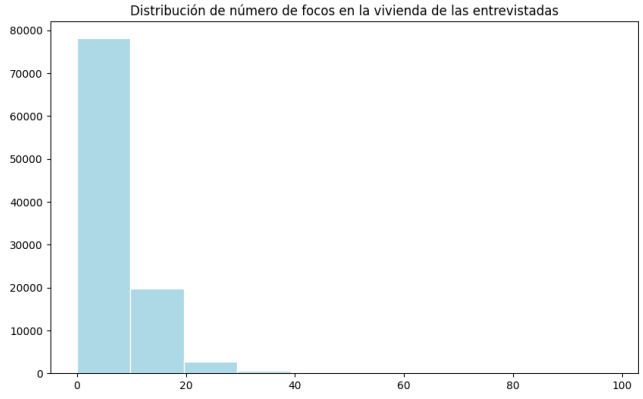

Fig. 6: Distribución de número de focos en la vivienda de las entrevistadas

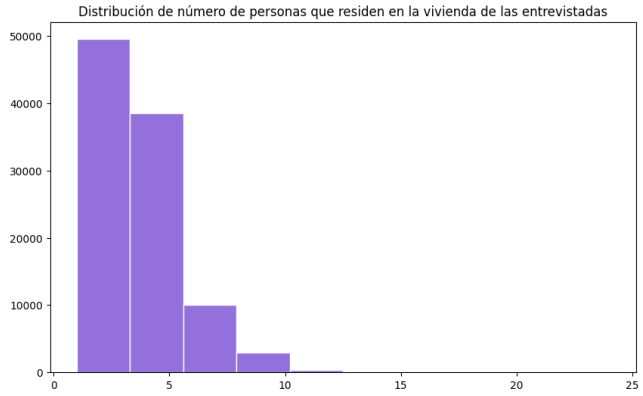

Fig. 7: Distribución de número de personas que residen en la vivienda de las entrevistadas

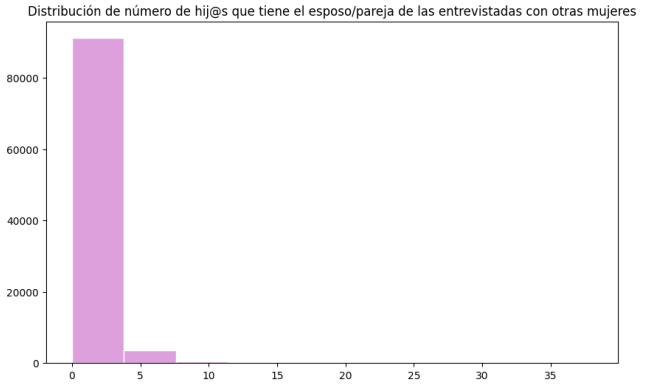

Fig. 8: Distribución de número de hij@s que tiene el esposo/pareja de las entrevistadas con otras mujeres

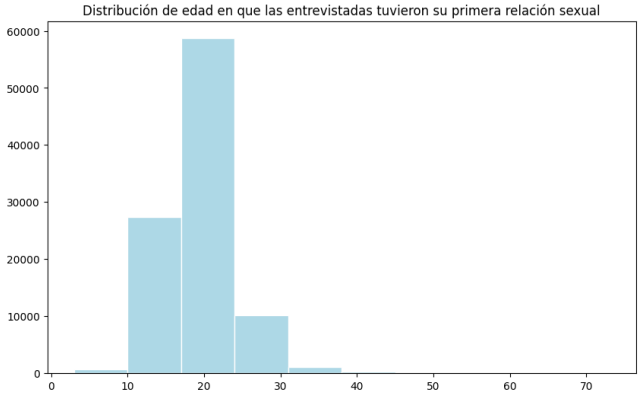

Fig. 9: Distribución de edades de las entrevistadas

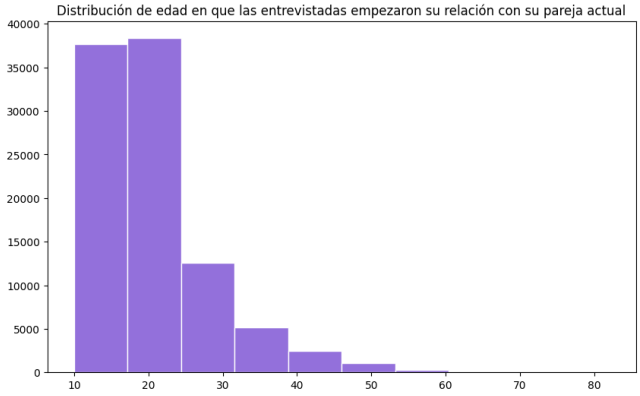

Fig. 10: Distribución de edad en que las entrevistadas tuvieron su primera relación sexual

5. Variable a predecir

Para diseñar nuestro modelo fue necesario crear una nueva variable que especifique si una persona entrevistada fue víctima de violencia o no. Para crear esta variable se utilizó la tabla TB_SEC_XIV, en ella se hacen preguntas explícitas sobre haber experimentado violencia de pareja. Se decició considerar la ocurrencia de cualquiera de estos incidentes es suficiente para catalogar a la persona como víctima de violencia.

6. Preguntas

1. ¿Cuáles son las dimensiones del dataset?
   110,126 x 172

2. ¿Cuántas variables categóricas existen?

26 variables categóricas.

3. ¿Cuántas variables numéricas hay?
   146 variables numéricas.

4. ¿Cuántas posibles variables objeto hay?
   Antes de realizar las modificaciones adecuadas había 26 variables de tipo objeto.

5. ¿Cuánto ruido existe en el dataset?
   Muchos, ya que el dataset tienen demasiados valores atípicos. Tanto así que algunos histogramas son afectados por ello.

6. ¿Cuántos valores nulos existen en la base de datos?
   Hay 8,645 de un total de 110127 valores.

7. ¿Qué porcentajes por cada variable?
   Estas varían dependiendo de la variable, teniendo variables para la mayoría de datos que llegan a más de 97 mientras otras de entrada no tenían ningún valor nulo. Al final nuestras variables quedaron sin datos nulos.

8. ¿A partir de que valor de "n" la distribución tiende a verse "normal"?
   En el Teorema del límite central, nos indica que la distribución de la media tiende a verse normal cuándo la muestra es mayor que 50 dónde perfectamente tenemos más de 50 registros.

b. Feature Selection

1. Método de Filtrado: Varianza

Los métodos de filtado se centran en características esenciales de las variables, es decir, en lugar de utilizar un modelo, utilizando medidas como la correlación o la varianza. En este caso se seleccionaron las variables númericas del training set para poder aplicar el método de *variance threshold*, el cual asume que las variables con mayor varianza contienen información más útil. No obstante, todas tenían una varianza muy elevada, por lo que la dimensión de la base de datos permaneció igual.

2. Método de Filtrado: Linear Discriminant Analysis

Se utilizó el método de análisis de discriminante lineal para considerar la precisión que tendría nuestro modelo en caso de filtrar nuestras variables numéricas haciendo un cambio de dimensionalidad. Sin embargo, los resultados obtenidos no demostraron generar un impacto positivo relevante en la precisión final del modelo, en comparación con los métodos descritos a continuación.

3. Método de Filtrado: Chi-squared

La mayor parte de las variables finales de nuestra base de datos son variables categóricas. Por lo tanto, consideramos de mayor importancia utilizar un método de filtrado exclusiva para ellas. La métrica de chi-cuadrada nos indica la relación entre las variables categóricas predictoras y la variable categórica a predecir. Con este método, obtuvimos una lista de las variables categóricas más significativas.

4. Método de envoltura

Un método de envoltura, es aquel que con un modelo de "machine learning", utiliza de este mismo su rendimiento como criterio de evaluación. Este método busca una característica más adecuada para el algoritmo, ya que tiene como objetivo, mejorar su rendimiento.

Para nuestra base de datos, elegimos la técnica relacionada a este método denominada como: "backforward selection". Esta técnica consiste en eliminar la característica menos significativa, comenzando en primera instancia, con todas las características. Lo que mejora el rendimiento del modelo. Se repite esto mismo hasta que no se observe ninguna mejora en la eliminación de características. En nuestro caso, con esta técnica logramos extraer las 5 mejores variables para predecir si la mujer sufre violencia de pareja.

Las 5 mejores variables fueron:

- P3_1: estado civil de la entrevistada
- P1_3: cantidad de focos en la vivienda
- P1_7: número de personas viviendo actualmente con ella
- P13_6: años que tenía cuando tuvo su primera relación sexual
- P13_8: edad que tenía la entrevistada cuando empezó su noviazgo con su pareja actual

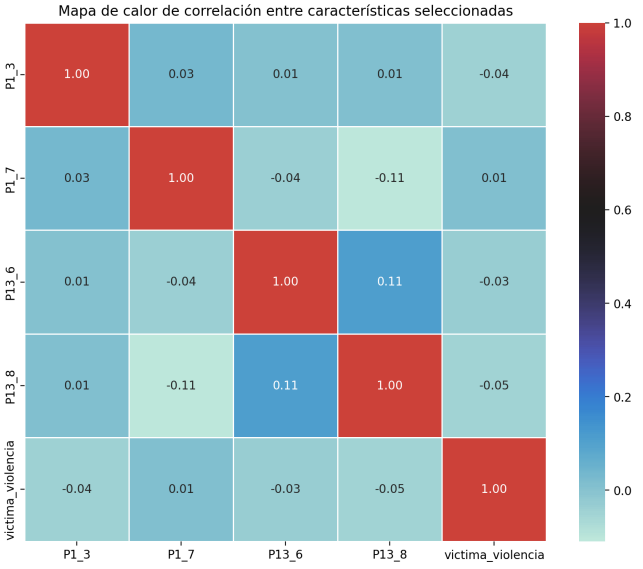

Fig. 11: Matriz de correlación para las últimas cinco características

Como podemos observar en la matriz de correlación, ninguna de las 5 variables están relacionadas entre sí, lo cual las hace válidas para ser utilizadas en el modelado de las mismas. Según esta técnica, tenemos una precisión del 73.21% al usar estas variables en un futuro modelo que prediga el número de mujeres entrevistadas que sufren violencia de pareja.

c. Modelado

Después de extraer las mejores variables encontradas para construir el modelo, se realizaron dos modelos distintos ex-

plorando diferentes algoritmos de clasificación. Para entrenar el modelos, se separaron los registros en conjuntos de prueba y entrenamiento con 20% y 80% de los datos, respectivamente.
El primer modelo utiliza un *Bosque Aleatorio* como algoritmo de clasificación con 100 árboles. La precisión obtenida para este modelo fue de 72.92% y su matriz de confusión se presenta a continuación:

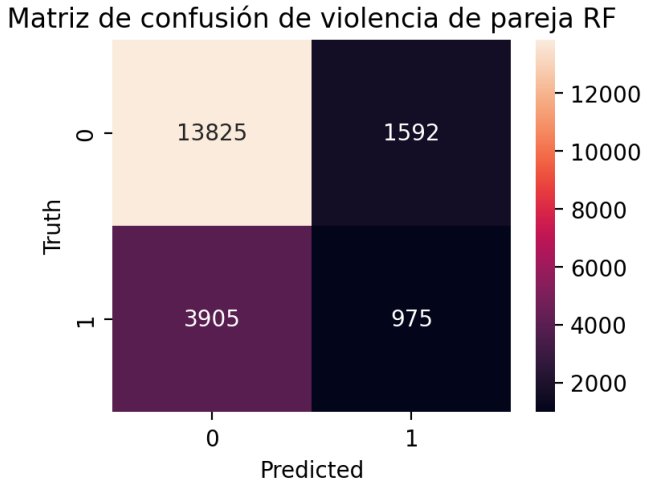

Fig. 12: Matriz de confusión del modelo de *bosque aleatorio*

El segundo modelo utiliza *Adaptative Boosting* como algoritmo de clasificación, con 100 árboles de decisión. La precisión obtenida para este modelo fue de 76.59% y su matriz de confusión se presenta a continueción:

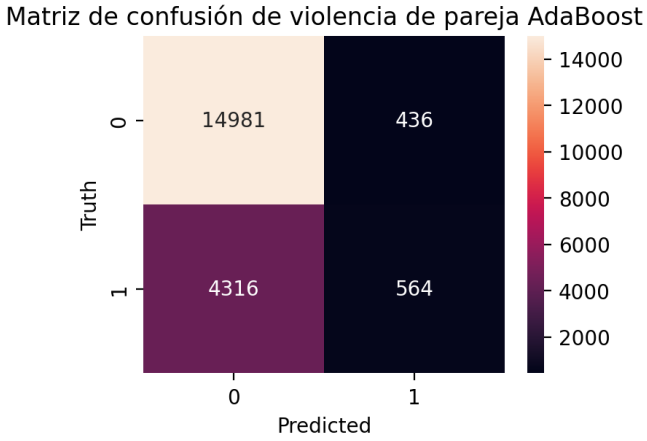

Fig. 13: Matriz de confusión del modelo de *Adaptative Boosting*

IV. ANÁLISIS DE RESULTADOS

Las bases de datos utilizadas permitieron construir modelos con exactitudes buenas, para esto se tuvo que seleccionar las características con mayor importancia para la predicción, para esto se usaron 4 métodos para tener una comparación acerca de las carcaterísticas más importantes, dónde los métodos que tuvieron más impacto fueron el método de *Filtrado de Chi-squared* y por metodo de envoltura usamos la tecnica *backforward*, dónde nos quedamos con las 5 variables todas categoricas más importantes, las cuales son: estado civil de la entrevistada, cantidad de focos de la vivienda, número de personas viviendo actualmente con ella, años que tenía cuando tuvo su primera relación sexual, edad de la entrevistada cuando empezó su noviazgo con su pareja actual.
Teniendo en cuenta las 5 principales variables se utilizarón modelos con diferentes algoritmos de clasificación donde se uso como primer modelo *Bosque aleatorio* y el segundo *Adaptative Boosting*, se puede determinar que es mejor modelo es *Bosque aleatorio*, ya que este predica de mejor manera los verdaderos positivos, este modelo tiene una la habilidad de detectar correctamente a las observaciones en la categoría de éxito de 20%, mientras que la habilidad de detectar correctamente las observaciones de categoría fracaso es de 89.67%, el F1-score de dicho modelo es de 21.57%, la precisión es de 38%,mientras la exactitud es de 72.92%. Con el segundo modelo de *Adaptative Boosting* este modelo tiene una habilidad de detectar correctamente las observaciones en la categoría de éxito de 11.56%, mientras que la habilidad de detectar correctamente las observaciones de categoría fracaso es de 97.17%, el F1-score de dicho modelo es de 20.65%, la precisión es de 56.4%, mientras la exactitud es de 76.59%.

V. CONCLUSIONES

El análisis realizado en esta modelación, así como las variables más significativas que dieron como resultado, muestran un patrón claro de grupos vulnerables con respecto a esta problemática. El número de focos en la vivienda, una de las preguntas claves para determinar el nivel socioeconómico de una persona, presenta una clara correlación negativa. El número de personas en vivienda, una estadística relacionada a la pobreza presenta una correlación positiva. Estos datos confirman las preocupaciones que se han escrito en múltiples textos sobre la violencia hacia las mujeres en situación vulnerable económicamente, al experimentar retos únicos de su situación como el encontrar mucho más difícil poder salir de una situación de violencia o presentar una mayor inseguridad en su día a día. Por otro lado, la correlación negativa de la edad de cuando tuvieron su primera relación sexual y la edad de cuando se juntaron con su pareja actual dan pruebas concretas del gran problema con el "grooming" que existe en México, con 80 mil casos reportados en el país en solo los últimos 2 años, y una gran normalización tanto en algunos sectores de la población como el gobierno. Finalmente, el estado civil de la persona puede estar aludiendo a una evidencia clara de cómo los valores familiares en México normalmente relacionados con la subordinación de la mujer terminan en casos reales y graves de violencia en las mujeres casadas, además de mostrar la falta de protección a la violencia en el entorno de los hogares y familia que se tiene en el país, donde los datos muestran ocurre la mayor violencia.

Este análisis muestra una clave directriz en los enfoques a los que las reformes públicas y organizaciones gubernamentales se les debería dar enfoque si se quieren hacer cambios en la situación de violencia de género del país. Así como la modelación general de los datos ha sido utilizada para diseñar y dar seguimiento a políticas públicas, se espera que

investigaciones como esta puedan ser utilizadas como alertas para que estas entidades den la continuación que es debida y ayudar tanto a cada una de estas mujeres que presentaron su caso en esta estadística como a aquellas cuya voz todavía no ha podido ser escuchada.

REFERENCIAS

[1] I. Medina Núñez and A. Medina Villegas, "Violencias contra las mujeres en las relaciones de pareja en méxico," *Intersticios sociales*, no. 18, pp. 269–302, 2019.

[2] M. J. García Oramas and M. P. Matud Aznar, "Salud mental en mujeres maltratadas por su pareja. un estudio con muestras de méxico y españa," *Salud mental*, vol. 38, no. 5, pp. 321–327, 2015.

[3] J. C. Ramírez-Rodríguez, "La violencia de varones contra sus parejas heterosexuales: realidades y desafíos. un recuento de la producción mexicana," *Salud pública de México*, vol. 48, pp. s315–s327, 2006.