**Instituto Tecnológico**

**y de Estudios Superiores de Occidente**

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Matemáticas y Física

Maestría en Ciencia de Datos



**Aplicación de un modelo predictivo**

Presentan: Lilivette Cruz Aguirre

Ricardo de León Flores

Óscar Aguirre Zambrano

Tlaquepaque, Jalisco. 29 de noviembre de 2023.

TABLA DE CONTENIDO

[Maestría en Ciencia de Datos i](#_Toc152101122)

[1. INTRODUCCIÓN iii](#_Toc152101123)

[1.1. Contexto iii](#_Toc152101124)

[1.2. Justificación iv](#_Toc152101125)

[2. DEFINICIÓN DEL PROBLEMA v](#_Toc152101126)

[2.1. Objetivos v](#_Toc152101127)

[2.1.1. Objetivo General v](#_Toc152101128)

[2.1.2. Objetivos Específicos v](#_Toc152101129)

[3. PREPARACIÓN DE LOS DATOS vi](#_Toc152101130)

[3.1. Descripción ENDIREH vi](#_Toc152101131)

[3.1.1. Alcance vi](#_Toc152101132)

[3.1.2. Periodicidad vi](#_Toc152101133)

[3.1.3. Población y cobertura vi](#_Toc152101134)

[3.1.4. Cuestionarios vii](#_Toc152101135)

[3.2. Delimitación del conjunto de datos vii](#_Toc152101136)

[3.2.1. Condiciones sobre los datos ix](#_Toc152101137)

[3.3. Descripción del conjunto de datos de trabajo ix](#_Toc152101138)

[3.4. Tratamiento de datos x](#_Toc152101139)

[3.4.1. Datos perdidos x](#_Toc152101140)

[3.4.2. Sustitución de datos x](#_Toc152101141)

[3.4.3. Columnas eliminadas x](#_Toc152101142)

[3.4.4. Eliminación de registros x](#_Toc152101143)

[3.4.5. Tratamiento de variable categóricas ordinales x](#_Toc152101144)

[3.4.6. Creación de variables Dummy xi](#_Toc152101145)

[3.4.1. Análisis de variables numéricas xi](#_Toc152101146)

[3.4.2. Análisis de Distribuciones y Relaciones xiii](#_Toc152101147)

[3.4.3. Matriz de correlación xiii](#_Toc152101148)

[4. MODELOS PREDICTIVOS xv](#_Toc152101149)

[5. RESULTADOS xvi](#_Toc152101150)

[5.1. Primera evaluación xvi](#_Toc152101151)

[5.1.1. Matriz de confusión xvi](#_Toc152101152)

[5.1.2. Actual vs Predicted xvii](#_Toc152101153)

[5.1.3. Receiver Operating Characteristic xviii](#_Toc152101154)

[5.2. Segunda evaluación (Considerando ROC) xix](#_Toc152101155)

[5.2.1. Matriz de confusión xix](#_Toc152101156)

[5.2.1. Receiver Operating Characteristic xx](#_Toc152101157)

[6. CONCLUSIONES xxi](#_Toc152101158)

# INTRODUCCIÓN

## Contexto

La ONU (Organización de las Naciones Unidad) define la violencia de género como los actos dañinos dirigidos contra una persona o un grupo de personas en razón de su género. Tiene su origen en la desigualdad de género, el abuso de poder y la existencia de normas dañinas. Aunque en menor manera, los hombres y los niños también pueden ser blancos de este tipo de violencia por lo que en este trabajo se referirá de manera específica a la violencia contra mujeres y niñas.

La violencia contra las mujeres y las niñas se define como todo acto de violencia basado en el género que tenga o pueda tener como resultado un daño o sufrimiento físico, sexual o mental para la mujer, así como las amenazas de tales actos, la coacción o la privación arbitraria de la libertad, tanto si se producen en la vida pública como en la vida privada.

De estas amplias definiciones se ha derivado el reconocimiento de distintos tipos de violencia, según el ámbito en el que tiene lugar (doméstica o familiar, comunitaria, educativa, laboral o institucional) y la naturaleza de la agresión (psicológica o emocional, física, obstétrica, sexual, económica o patrimonial).

La violencia contra las mujeres y las niñas constituye un grave problema de salud pública y una violación de los derechos humanos. La violencia puede afectar negativamente la salud física, mental, sexual y reproductiva de las mujeres.

Estimaciones de la ONU indican que una de cada tres mujeres en las Américas ha sufrido violencia física o sexual por parte de sus parejas o de algún tercero.

En México, la generación de información y datos sobre la violencia contra las mujeres se inició en la década de los 90 como respuesta a numerosas demandas de diversos sectores sociales, en particular de académicas y activistas feministas que pugnaban por el reconocimiento y visibilización del problema tanto en espacios públicos como privados.

México fue adquiriendo la responsabilidad de atender este problema social —y de dar cuenta de ello— al suscribir diversos acuerdos internacionales, como la Primera Conferencia Internacional de la Mujer realizada en México en 1975, la Convención para la Eliminación de Todas las Formas de Violencia contra las Mujeres (CEDAW, por sus siglas en inglés) en 1979, la Cuarta Conferencia Internacional de la Mujer celebrada en Beijing en 1995 y la Convención Interamericana para Prevenir, Sancionar y Erradicar la Violencia Contra las Mujeres de Belem do Pará.

El 1ro de febrero de 2007, fue publicada en el Diario Oficial de la Federación, la Ley General de Acceso de las Mujeres a una Vida Libre de Violencia que tiene ene por objetivo prevenir, sancionar y erradicar la violencia contra las mujeres, mediante una coordinación entre el [gobierno federal](https://es.wikipedia.org/wiki/Pol%C3%ADtica_de_M%C3%A9xico), estados y municipios, y con ello garantizar para las mujeres el acceso a una vida libre de violencia

## Justificación

La violencia contra las mujeres y las niñas tiene un efecto multidimensional sobre la salud general de la economía, tanto a corto como a largo plazo.

La violencia contra las mujeres ocasiona pérdidas económicas a nivel individual, familiar e incluso comunitario. Son las propias mujeres violentadas quienes destinan parte de sus ingresos y mucho de su tiempo para acceder a tratamiento médico o psicológico, asesoría jurídica o cualquier otro tipo de apoyo.

Lesiones, problemas sexuales, enfermedades crónicas, desórdenes mentales como ansiedad y depresión, así como el abuso de sustancias, son solo algunas de las principales consecuencias del maltrato que reciben miles de mujeres en el mundo.

Estudios recientes de la Organización Internacional del Trabajo (OIT) añaden a la lista de efectos de esta violencia, la baja productividad, la pérdida de días laborales, la inasistencia, los retardos, entre otros. La misma OIT enfatiza que dichos efectos de la violencia llegan a costarle a las naciones alrededor de un 3.5% de su PIB; otras fuentes señalan que las pérdidas económicas rebasan el 5%.

Se estima que la violencia contra las mujeres tiene costos que aún no son medibles, derivados del daño emocional que afecta a las víctimas, a sus familiares y a la sociedad en su conjunto. De hecho, se afirma que por cada mujer que está inmersa en una situación de violencia, son varias generaciones posteriores que también padecerán las consecuencias.

Por lo anterior, erradicar la violencia contra las mujeres es un camino para garantizar una mejor calidad de vida para la sociedad en general, pero también una forma de impactar positivamente en la economía de las naciones.

# DEFINICIÓN DEL PROBLEMA

La Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares (ENDIREH) proporciona datos valiosos sobre la violencia de género en México. Sin embargo, estos datos son complejos y extensos. La ciencia de datos puede ayudar a analizarlos para comprender mejor la magnitud, las causas y las consecuencias de la violencia contra las mujeres en diferentes regiones y contextos socioeconómicos. Esto permitirá a los responsables de la toma de decisiones diseñar estrategias más efectivas para combatir este problema.

En el contexto de la materia de Modelado Predictivo, el problema a abordar es la generación de un modelo de regresión que nos permita anticipar o prever si una mujer en pareja es susceptible de sufrir violencia de género de acuerdo los resultados de ENDIREH.

## Objetivos

#### Objetivo General

Desarrollar un modelo predictivo que pueda predecir la incidencia de la violencia de género en diferentes áreas geográficas de México.

#### Objetivos Específicos

1. Realizar un análisis exploratorio de los datos de ENDIREH para identificar las variables clave relacionadas con la violencia de género.
2. Preparar los datos y hacer un análisis estadístico de los mismos.
3. Identificar el mejor modelo de regresión a implementar.

# PREPARACIÓN DE LOS DATOS

## Descripción ENDIREH

La Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares (ENDIREH) es una encuesta realizada en México para recopilar información relacionada con la violencia de género, la convivencia y las relaciones familiares en los hogares del país. La ENDIREH es una iniciativa del Instituto Nacional de Estadística y Geografía (INEGI) de México y tiene el propósito de proporcionar datos sobre la dinámica de las relaciones en los hogares mexicanos, incluyendo aspectos como la violencia de género, el control, la convivencia, y otros temas relacionados.

#### Alcance

La ENDIREH ofrece información sobre las experiencias de violencia física, económica, sexual, psicológica y patrimonial que han enfrentado las mujeres de 15 años y más en los ámbitos familiar, escolar, laboral, comunitario y de pareja. Asimismo, presenta datos sobre las personas agresoras y los lugares donde ocurrieron las agresiones, así como variables adicionales que permiten analizar la violencia contra las mujeres en México.

#### Periodicidad

La ENDIREH se realiza de manera periódica, con un aproximado de cinco años entre cada edición. Encuestas disponibles:

* 2021. Para este proyecto utilizaremos la última encuesta disponible.
* 2016
* 2011
* 2006
* 2003

#### Población y cobertura

Población objetivo.

* Mujeres de 15 años o más residentes habituales de las viviendas seleccionadas en la muestra.

Cobertura geográfica.

* Nacional
  + Urbano (107,680)
  + Rural (33,104)
* Entidad federativa

#### Cuestionarios

I Cuestionario general  
IV. Ingresos y recursos  
VI. Opinión sobre los roles masculinos y femeninos  
VII. Ámbito escolar  
VIII. Ámbito laboral  
IX. Ámbito comunitario  
X. Atención Obstétrica  
XI. Ámbito familiar  
XII. Familia de origen  
XIII. Vida en Pareja  
XIII.I. Tensiones y conflictos  
XIV. Relación actual o última  
XV. Decisiones y libertad personal  
XVI. Recursos sociales  
XVII. División del trabajo en el hogar  
XVIII. Mujeres de 60 años y más  
XIX. Discapacidad  
XX. Privacidad en la entrevista  
XXI. Prevalencia de la violencia

## Delimitación del conjunto de datos

La base de datos ENDIREH consta de 28 archivos CSV.

Tabla

Descripción generada automáticamente

Para este proyecto usaremos la información de los cuestionarios generales que nos brinden la información sobre la vivienda y características sociodemográficas. También utilizaremos los cuestionarios que describan la vida en pareja, los ingresos y recursos de las mujeres, la descripción de su relación actual y su condición de violencia. La información anterior se encuentra en los siguientes archivos:

|  |  |
| --- | --- |
| TVIV | Características de la Vivienda y Hogares en la Vivienda |
| TSDem | Características Sociodemográficas de Residentes de la Vivienda |
| TB\_SEC\_III | Elegibilidad y Verificación de Situación Conyugal de la Mujer Elegida de 15 años o más |
| TB\_SEC\_IV | Ingresos y recursos |
| TB\_SEC\_XIII | Vida en Pareja |
| TB\_SEC\_XIV | Relación actual o última |
| TB\_VD | Contiene información sobre la condición de violencia total |

#### Condiciones sobre los datos

En la ENDIREH las mujeres elegidas se clasifican con base en tres situa­ciones conyugales, que a su vez corresponden a cada tipo de cuestionario para la mujer elegida:

1. casadas o unidas,
2. separadas, divorciadas o viudas, y
3. solteras

Para este proyecto, seleccionaremos solo a las **mujeres casadas o unidas del estado de Jalisco.**

## Descripción del conjunto de datos de trabajo

Los archivos se cargaron completos en dataframes independientes y sobre estos se aplicaron los filtros de ESTADO, SITUACIÓN CONYUGAL, SEXO y COD\_RES (para determinar las encuestas completas).

Se analizaron todos los dataframes para identificar los IDs y las variables a utilizar (la descripción de las variables, su rango de valores y la descripción de estos se encuentra en el Colab). Dataframes resultantes:

|  |  |
| --- | --- |
| **Dataframe** | **Ids** |
| TVIV | ID\_VIV |
| TSDem | ID\_VIV, ID\_PER |
| TB\_SEC\_III | ID\_VIV, ID\_PER |
| TB\_SEC\_IV | ID\_VIV, ID\_PER |
| TB\_SEC\_XIII | ID\_VIV, ID\_PER |
| TB\_SEC\_XIV | ID\_VIV, ID\_PER |
| TB\_VD | ID\_VIV, ID\_PER |

Una vez que los dataframes fueron procesados se unieron en un solo dataframe con las siguientes características:

Nombre: VG

IDs: ID\_VIV, ID\_PER

Variables (X): 57 columnas (disponibles en el Colab)

Y: **VTOT\_A** (Variable que identifica si la informante ha sido víctima o no, de al menos una situación de violencia en alguno de los ámbitos escolar, laboral, comunitario, familiar o de pareja, a lo largo de su vida.)

Rango Y: [1, 2: Con incidencia de violencia, Sin incidencia de violencia]

Número de registros: 2013

## Tratamiento de datos

#### Datos perdidos

Se contabilizaron los valores perdidos y el porcentaje de estos por variable y se eliminaron los que tuvieron 40% o más de datos perdidos.

#### Sustitución de datos

Se sustituyeron los Nan de algunas variables que de acuerdo a la descripción de la base de datos deberían corresponder a algún valor en específico. Ejemplo:

Variable: P13\_2

Descripción: Edad al nacer el primer hijo

Rango: 10-99 ([98, 99]: No recuerda, No especificado)

Por lo tanto, los Nan se sustituyen por 99.

#### Columnas eliminadas

Las siguientes columnas se eliminaron:

IDs (ID\_VIV, ID\_PER). No aportan valor al modelo.

VTOT\_12M (Violencia en los últimos 12 meses). Nos interesa la violencia total a lo largo de toda la vida.

P4\_2\_1, P4\_7\_AB y P4\_5\_AB. No hay variedad relevante en los datos.

COD\_M15 y CODIGO. Variables usadas para filtrar los datos.

#### Eliminación de registros

Se eliminaron los registros para los cuales no hay valores en nuestra Y (**VTOT\_A**)

#### Tratamiento de variable categóricas ordinales

Las variables para las que existe un mapeo de los valores “No recuerda” o “No especificado” se hace la sustitución de estos por 0 ya que la escala si tiene importancia y con el 0 se reduce el impacto que pudieran tener en el modelo.

Ejemplo:

Variable: P4\_7\_AB

Descripción: Cantidad de la aportación económica

Rango: 0- 999997 ([999999]: No recuerda)

#### Creación de variables Dummy

Se seleccionaron las variables categóricas de interés, se generaron las correspondientes variables dummy y las columnas originales fueron eliminadas del dataset.

El resultado de la transformación fue un dataset con 77 columnas.

#### Análisis de variables numéricas

Se realizaron gráficas de cajas y bigotes para las variables númericas, en estás se pueden observar los valores atípicos.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

##### Valores atípicos

Total de outliers por variable:

NIV: 371 outliers

P13\_1: 303 outliers

P4\_9: 209 outliers

P13\_2: 194 outliers

P13\_8: 122 outliers

P13\_9: 88 outliers

P13\_6: 87 outliers

P4\_7\_AB: 81 outliers

P13\_10: 81 outliers

P1\_7: 34 outliers

EDAD: 4 outliers

ESTRATO: 0 outliers

GRA: 0 outliers

Los datos se estandarizaron utilizando StandardScaler().

#### Análisis de Distribuciones y Relaciones

Se realizó lo siguiente:

* Creación de la matriz de correlación
* Se hizo el aplanado de la matriz unstack() para examinar las relaciones entre pares de variables y detectar patrones de correlación en los datos.
* Se excluyeron las correlaciones perfectas (correlations = 1)

#### Matriz de correlación

Interfaz de usuario gráfica, Gráfico

Descripción generada automáticamente

**Correlaciones Fuertes**: Los cuadros que aparecen en rojo oscuro indican una correlación fuerte y positiva entre las variables.

Los cuadros rojos oscuros en la diagonal indican la correlación perfecta de una variable consigo misma.

**Correlaciones Débiles**: Los cuadros en tonos de azul representan una correlación débil o nula entre las variables. Esto sugiere que no hay una relación lineal fuerte entre estas variables.

**Correlaciones Negativas**: Los cuadros en tonos de azul oscuro fuera de la diagonal indican una correlación negativa fuerte, lo que significa que cuando una variable aumenta, la otra tiende a disminuir.

Patrones de Bloques: Los bloques de colores similares indican grupos de variables que se comportan de manera similar entre sí.

En este caso en las variables P13\_8, P13\_9 y P13\_10 se comportan similar ya que se refieren a la edad a la que se inició la última relación, la edad a la que inició la vida en pareja y la edad de la pareja al iniciar la relación, esto es entendible ya que solo se están analizando mujeres que actualmente están en una relación.

# MODELOS PREDICTIVOS

Realizamos un proceso de partición del conjunto de datos mediante la técnica de "split" con una proporción del 70% para el conjunto de entrenamiento y el 30% para el conjunto de pruebas.

Para nuestro proyecto utilizamos los siguientes modelos de clasificación para predecir si una mujer puede o no sufrir violencia con base en las respuestas a las preguntas que seleccionamos de la ENDIREH.

* LogisticRegression
* SVM
* MLPClassifier
* DecisionTreeClassifier
* BaggingClassifier
* RandomForestClassifier
* RandomForestClassifier

También utilizamos Grid Search para identificar el mejor modelo.

# RESULTADOS

## Primera evaluación

Se ejecutaron los modelos definidos anteriormente tomando como métrica para evaluarlos el Accuracy. Resultados:

Mejor Modelo: **SVM**

Mejor puntaje (**Accuracy**): **0.7182313267144234**

Mejor configuración: {'C': 10, 'kernel': 'linear'}

Accuracy en el conjunto de prueba: 0.7654584221748401

#### Matriz de confusión

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

**Nota**: En nuestra matriz de confusión los casos positivos se refieren a “Sin incidencia de violencia” y los negativos a “Con incidencia de violencia”.

**Verdaderos positivos** (TP): La celda inferior derecha (con el número 29) muestra el número de casos positivos (1) que fueron correctamente identificados por el modelo.

**Falsos positivos** (FP): La celda superior derecha (con el número 12) muestra el número de casos negativos (0) que fueron incorrectamente identificados como positivos (1) por el modelo.

**Verdaderos negativos** (TN): La celda superior izquierda (con el número 330) muestra el número de casos negativos (0) que fueron correctamente identificados por el modelo.

**Falsos negativos** (FN): La celda inferior izquierda (con el número 98) muestra el número de casos positivos (1) que fueron incorrectamente identificados como negativos (0) por el modelo.

En resumen, la matriz muestra que el modelo es bastante conservador al predecir la clase positiva (Sin incidencia de violencia), con muchos más falsos negativos que falsos positivos.

Esto podría ser indicativo de un modelo que tiene un umbral alto para predecir la clase positiva (Sin incidencia de violencia), o que simplemente es más preciso al predecir la clase negativa (Con incidencia de violencia) en comparación con la clase positiva.

#### Actual vs Predicted

En la siguiente gráfica podemos observar en azul los valores reales vs los valores que se predijeron.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Se observa que para la clase 1 (Con incidencia de violencia) se predicen más falsos positivos. En la clase 2 (Sin incidencia de violencia) se puede observar que se predicen menos falsos negativos.

#### Receiver Operating Characteristic

Eje X (Tasa de Falsos Positivos): Muestra la proporción de resultados negativos reales que se clasificaron incorrectamente como positivos. Va de 0 a 1, y un modelo ideal tendría una tasa de falsos positivos de 0.

Eje Y (Tasa de Verdaderos Positivos): Muestra la proporción de resultados positivos reales que se clasificaron correctamente. También va de 0 a 1, y un modelo ideal tendría una tasa de verdaderos positivos de 1.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

La curva ROC tiene un AUC de 0,60, lo que indica que el modelo tiene una capacidad de discriminación limitada, pero todavía es mejor que el azar. Un modelo ideal tendría la curva ROC más cerca de la esquina superior izquierda, maximizando la tasa de verdaderos positivos mientras minimiza la tasa de falsos positivos.

## Segunda evaluación (Considerando ROC)

Se volvieron a ejecutar los mismos modelos, pero está vez usando ROC como métrica para evaluarlos. Resultados:

Mejor Modelo: LogisticRegression

Mejor puntaje (AUC): 0.7540134937626967

Mejor configuración: {'C': 20}

AUC en el conjunto de prueba: 0.7463738085370908

#### Matriz de confusión

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

**Nota**: En nuestra matriz de confusión los casos positivos se refieren a “Sin incidencia de violencia” y los negativos a “Con incidencia de violencia”.

**Verdaderos positivos** (TP): La celda inferior derecha (con el número 36) muestra el número de casos positivos (1) que fueron correctamente identificados por el modelo.

**Falsos positivos** (FP): La celda superior derecha (con el número 31) muestra el número de casos negativos (0) que fueron incorrectamente identificados como positivos (1) por el modelo.

**Verdaderos negativos** (TN): La celda superior izquierda (con el número 311) muestra el número de casos negativos (0) que fueron correctamente identificados por el modelo.

**Falsos negativos** (FN): La celda inferior izquierda (con el número 91) muestra el número de casos positivos (1) que fueron incorrectamente identificados como negativos (0) por el modelo.

En resumen, en esta segunda ejecución, con LogisticRegression como el mejor modelo, se incrementó el número de casos “Con incidencia de violencia” que no se identificaron correctamente y se incremento el número de casos “Sin incidencia de violencia” incorrectamente identificados.

#### Receiver Operating Characteristic

Gráfico, Gráfico de líneas

Descripción generada automáticamente

La gráfica ROC de la primera evaluación con SVM tiene el mismo valor de ROC curve que con la segunda evaluación con LogisticRegression.

# CONCLUSIONES

En este trabajo evaluamos varios modelos utilizando dos métricas distintas, accuracy y AUC-ROC, para comprender mejor su rendimiento en un problema de clasificación binaria para determinar la incidencia de violencia basándonos en las respuestas obtenidas de ENDIREH 2021.

Los modelos evaluados incluyeron Logistic Regression, Support Vector Machine (SVM), MLPClassifier, Decision Tree Classifier, Bagging Classifier y Random Forest Classifier.

Resultados de la Primera Evaluación (**Accuracy**):

Mejor Modelo: SVM

Mejor Puntaje (Accuracy): 0.7182

Mejor Configuración: {'C': 10, 'kernel': 'linear'}

Accuracy en el Conjunto de Prueba: 0.7655

Resultados de la Segunda Evaluación (**AUC-ROC**):

Mejor Modelo: Logistic Regression

Mejor Puntaje (AUC): 0.7540

Mejor Configuración: {'C': 20}

AUC en el Conjunto de Prueba: 0.7464

Análisis de Matriz de Confusión (Primera Evaluación):

La matriz de confusión reveló que el modelo SVM fue conservador al predecir la clase positiva ("Sin incidencia de violencia"). Hubo más falsos negativos que falsos positivos, indicando que el modelo podría tener un umbral alto para predecir la clase positiva o ser más preciso al predecir la clase negativa ("Con incidencia de violencia").

Análisis de Matriz de Confusión (Segunda Evaluación):

Con Logistic Regression como el mejor modelo en la segunda evaluación, observamos un aumento en los falsos negativos ("Sin incidencia de violencia") y un aumento en los falsos positivos ("Con incidencia de violencia"). Esto podría indicar un desafío en encontrar un equilibrio óptimo entre las clases y señala la importancia de ajustar el umbral de decisión para adaptarse a los requisitos específicos del problema.

Después de comparar los modelos SVM y Logistic Regression concluimos que para nuestro caso el mejor modelo es SVM ya que es mejor prediciendo los casos "Con incidencia de violencia" y aunque haya más falsos negativos preferimos que nuestro modelo sea capaz de predecir correctamente cuando hay violencia, ya que dentro del contexto del problema que tratamos de resolver tiene más valor identificar los casos en los que una mujer puede ser víctima de violencia para poder tomar medidas preventivas las cuáles aún es los falsos negativos podrían tener un impacto positivo.