Modelo Predictivo de Violencia de Género en Mujeres de Jalisco, 15+ en Relaciones.

Cruz, Lilivette Zambrabno, Oscar De León, Ricardo

ITESO, Maestria en ciencia de datos

ITESO, Universidad Jesuita de Guadalajara

Introducción

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.

Objetivo

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

Análisis de Datos

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

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

Figure 1. Dataframe

Análisis de Distribuciones y Relaciones

- 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)

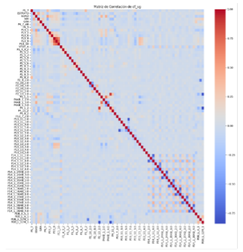

Figure 2. Matriz de correlación

En este caso en las variables **P138**, **P139** y **P1310** 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

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

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

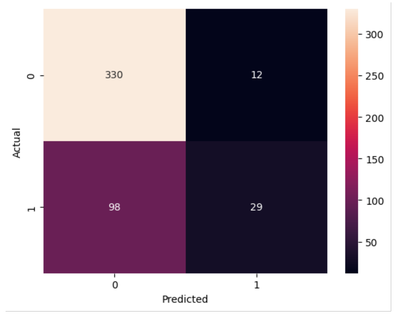

Figure 3. Matriz de confusión SVM

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

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.

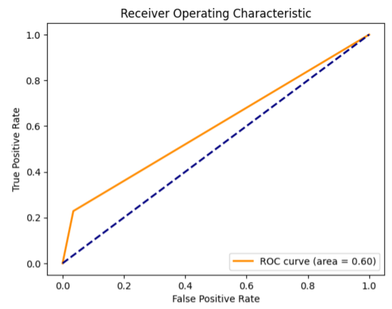

Figure 4. Curve rock svm

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.

Ejecución de un modelo diferente para hacer una comparación de modelos.

Mejor Modelo: **LogisticRegression**

Mejor puntaje (**AUC**): **0.7540134937626967**

Mejor configuración: 'C': 20

Accuracy en el conjunto de prueba: 0.7463738085370908

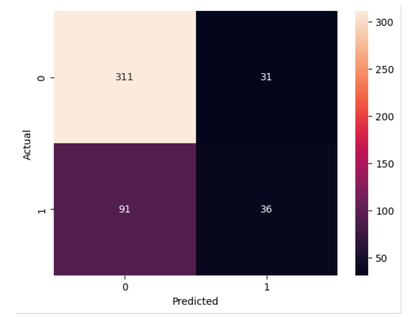

Figure 5. Matriz de confusión LogisticRegression

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.

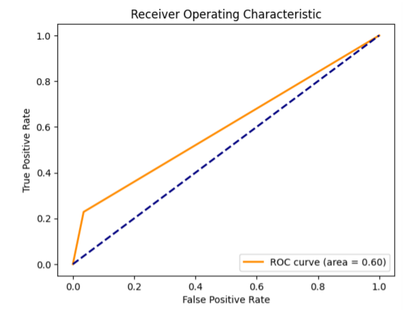

Figure 6. Curve rock LogisticRegression

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.

Evaluación (Accuracy):

Mejor Modelo: **SVM**

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

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

Accuracy en el conjunto de prueba: 0.7654584221748401

Evaluación (AUC-ROC):

Mejor Modelo: **LogisticRegression**

Mejor puntaje (**AUC**): **0.7540134937626967**

Mejor configuración: 'C': 20

Accuracy en el conjunto de prueba: 0.7463738085370908

En la evaluación de modelos, se encontró que el **SVM** fue conservador al predecir la clase positiva, mostrando más falsos negativos que falsos positivos. Aunque **Logistic Regression** fue el mejor modelo en una evaluación, presentó desafíos al equilibrar las clases en la segunda evaluación. Se concluyó que, a pesar de los falsos negativos, **SVM** es preferido porque destaca en predecir casos de violencia, lo cual es crucial para tomar medidas preventivas en el contexto del problema abordado.