Fuente del dataset

* Dataset:<https://www.kaggle.com/datasets/aracelifernandez/base-de-datos-linea-144-argentina>
* Información de columnas:<https://www.datos.gob.ar/dataset/generos-base-datos-linea-144/archivo/generos_1.4>

Resumen del proyecto

El proyecto se centra en el análisis de la evolución de los casos de violencia de género a lo largo del tiempo, utilizando datos de la Línea 144 de Argentina. Esta línea proporciona asistencia telefónica especializada a mujeres víctimas de violencia de género, abarcando diversas formas de violencia como física, psicológica, sexual, económica, y simbólica.

El dataset, disponible en [Kaggle](https://www.kaggle.com/datasets/aracelifernandez/base-de-datos-linea-144-argentina) y consolidado desde 2020 hasta 2023, incluye información detallada sobre las denuncias recibidas. Se han aplicado técnicas de clasificación para desarrollar modelos predictivos que predicen si una persona ha recibido o no violencia. Además, se han utilizado métodos de clustering para identificar patrones y agrupar casos en función de variables como la edad y otros factores relevantes. Este enfoque proporciona una visión de cómo cambian y se distribuyen los tipos de violencia y permite prever sus posibles evoluciones en los próximos años.

Objetivo

El proyecto tiene como objetivo analizar cómo han variado los casos de violencia a lo largo del tiempo y determinar qué tipos de violencia (física, psicológica, sexual, económica, etc.) han mostrado mayores fluctuaciones.

Análisis exploratorio

Los datos revelan que el 97% de las personas que sufrieron violencia son mujeres, en su mayoría adultas de entre 30 y 60 años. Además, el 54% de las víctimas residen en la provincia de Buenos Aires.

Se identificó que cerca del 70% de los agresores fueron o son parejas de las víctimas y pertenecen mayoritariamente al género masculino. Las víctimas que denunciaron reportaron haber sufrido entre 3 y 5 tipos de violencia en la mayoría de los casos.

Las denuncias se concentran principalmente en casos de violencia psicológica y doméstica, seguidas por violencia física. Además, se registraron modalidades de violencia como la institucional, laboral, contra la libertad reproductiva, obstétrica y mediática.

Aprendizaje supervisado

Nuestro objetivo es predecir si una mujer ha sufrido o no violencia física, representado como 1 para "Sí" y 0 para "No". Para ello, utilizamos como variables predictoras todos los tipos de violencia existentes, la provincia, la edad de la mujer y su vínculo con el agresor.

Nuestros datos muestran que en el 70% de los casos la persona fue víctima de violencia física.

Para cumplir nuestro objetivo utilizamos el algoritmo conocido como KNN (Vecinos más cercanos). Este algoritmo clasifica un dato nuevo basándose en la "cercanía" de este a otros datos previamente clasificados. "K" se refiere al número de vecinos más cercanos que se consideran para tomar la decisión.

Para K=5:

Al probar con K=5, la matriz de confusión muestra que el modelo predijo correctamente 13.071 casos de violencia física y 5.464 casos de no violencia física, lo que da un total de 18.535 predicciones correctas de 18.944 en total. Esto resulta en una precisión (accuracy) del 97%.

Sin embargo, este alto rendimiento sugiere un posible sobreajuste (overfitting), donde el modelo podría estar aprendiendo demasiado bien las características específicas del conjunto de datos de entrenamiento, pero no generalizar bien a nuevos datos.

Para K=250:  
Por otro lado, al utilizar K=250 obtuvimos un rendimiento del 84%, con 13.046 casos de violencia física y 2.777 casos de no violencia física correctamente predichos. Aunque la precisión es menor, este resultado es más deseable porque indica un mejor equilibrio entre la capacidad del modelo para aprender y su capacidad para generalizar, reduciendo el riesgo de sobreajuste.

Aprendizaje no supervisado

El objetivo es encontrar patrones para agrupar a las mujeres y aportar información relevante para la formulación e implementación de políticas públicas que permitan mejorar la asistencia a quienes acuden a la Línea 144.

Para esto elegimos el algoritmo kmeans que se utiliza para agrupar datos en K grupos o clusters.

En este caso se utilizaron como variables los tipos de violencia, la edad y el vínculo con el agresor.

El primer paso fue discernir el mejor número de clusters K en los cuales podríamos dividir los datos. Para ello utilizamos los métodos denominados "Elbow" y "Silhouette".

Con el método de Elbow, obtuvimos que los números óptimos se encontraban entre 6 y 7 clusters.

En tanto con Silhouette se obtuvo una curva creciente, por tanto, se decidió avanzar con un análisis detallado de la composición de los clusters en cuanto a observaciones y score promedio para los siguientes valores de K= 3, 4, 5, 6, 7, 8, 9, 10, 14, 16, 18, 20. Los mejores agrupamientos se observaron para K=6 y K=7.

Para K=6 la distribución de los datos en los clusters no es muy uniforme. Hay una concentración significativa de datos en el Cluster 0 (22.194 datos) y una menor cantidad en el Clusters 3 con 4.264 datos.

Con K=7, aunque sigue habiendo variaciones, los tamaños de los clusters están más cerca unos de otros, con una menor disparidad entre el cluster más grande (11.630 datos en Cluster 5) y el más pequeño (4.249 datos en Cluster 2).

Por esta razón, mostramos el detalle de 7 clusters.

En cuanto a la edad, los clusters 0, 1, 5 y 6 tienen una media de edad cercana a los 33-34 años, lo que indica una población relativamente joven en estos grupos.

Los clusters 2, 3 y 4 presentan una media de edad mayor, con el cluster 3 teniendo la media más alta (39 años).

Con respecto a los tipos de violencia, se puede observar que en los dos primeros clusters tienen aproximadamente el mismo porcentaje las personas que sufrieron violencia psicológica, física y doméstica. En los clusters 2 y 3 predomina la violencia doméstica.

En el 4 y 5 tiene el mismo porcentaje psicológica, física, doméstica y simbólica, pero en el 5 cambia esta última por económica. Por último, en el 6 predomina la violencia sexual.

Líneas a seguir

Aprendizaje supervisado:

* Repasar el análisis exploratorio, enfocando en la interacción entre los tipos de violencia y las modalidades.
* Avanzar con modelos alternativos que consideren otras variables explicativas.
* Elección de diferentes variables predictivas.

Aprendizaje no supervisado:

* Los tipos de violencia se configuran como una variable relevante para el análisis de acciones a evaluar para cada grupo.
* La edad no resulta una variable significativa para caracterizar cada cluster.
* Se podrían incluir otras variables al análisis, como por ejemplo la distribución por provincias.