Detección de riesgo de violencia física contra niños

Jeremias Morlandi

March 13, 2025

1 Introducción

La violencia física contra niños es un problema persistente a nivel global, y su detección temprana es crucial para la protección de los más vulnerables. En Argentina, la Encuesta de Indicadores Múltiples por Conglomerados (MICS) realizada por UNICEF entre 2019 y 2020 reveló que un 59.4% de los niños menores de 15 años experimentaron algún tipo de violencia disciplinaria en sus hogares, y un 35.4% fueron sometidos a castigos corporales. A pesar de la gravedad de esta problemática, los niveles de denuncia siguen siendo alarmantemente bajos.

Dado que las intervenciones en muchos casos dependen de la denuncia por parte de terceros, se plantea la necesidad de utilizar herramientas predictivas para identificar a los niños en riesgo de violencia física antes de que ocurran situaciones graves. El objetivo de este trabajo es aplicar técnicas de aprendizaje automático (machine learning) para desarrollar un modelo que pueda predecir el riesgo de violencia física contra niños a nivel de hogar. Al hacerlo, se busca contribuir al diseño de políticas públicas que permitan una intervención temprana, aumentando la cobertura de protección infantil y mitigando los problemas de subregistro de casos.

Este trabajo utiliza un dataset proveniente de la encuesta MICS para entrenar y evaluar diferentes modelos de clasificación, centrándose en la capacidad predictiva de estos para detectar hogares con alto riesgo de violencia. Los resultados obtenidos ayudarán a mejorar la asignación de recursos y la eficiencia en la detección temprana de posibles víctimas, proporcionando una herramienta de apoyo para los trabajadores sociales en sus intervenciones.

2 Descripción del Dataset

El dataset analizado contiene 5150 filas y 644 columnas. Definimos la columna binaria "violence" como la variable target para este estudio. A su vez, encontramos que el conjunto de datos no posee valores faltes.

Selección de variables:

El proceso realizado con el algoritmo Random Forest tuvo como objetivo identificar las variables más importantes entre un conjunto inicial de 643 variables, utilizando la importancia de las características basada en el índice Gini.

1. Entrenamiento del modelo: Comenzamos ajustando un modelo de Random Forest a los datos, utilizando Grid Search con validación cruzada de K-Folds para optimizar los hiperparámetros más relevantes. Estos hiperparámetros incluyen el número de árboles en el bosque, la profundidad máxima de los árboles, y el número mínimo de muestras requeridas para dividir un nodo. Este proceso asegura que el modelo se entrena con los mejores parámetros para maximizar la precisión.

2. Importancia de características: Una vez entrenado el modelo con los mejores hiperparámetros, se calculó la importancia de las características. El índice Gini se utiliza como criterio para evaluar el poder predictivo de cada variable: mide cuánto cada característica contribuye a mejorar la pureza de los nodos en los árboles del bosque. Las variables que causan la mayor disminución del índice Gini se consideran las más importantes.

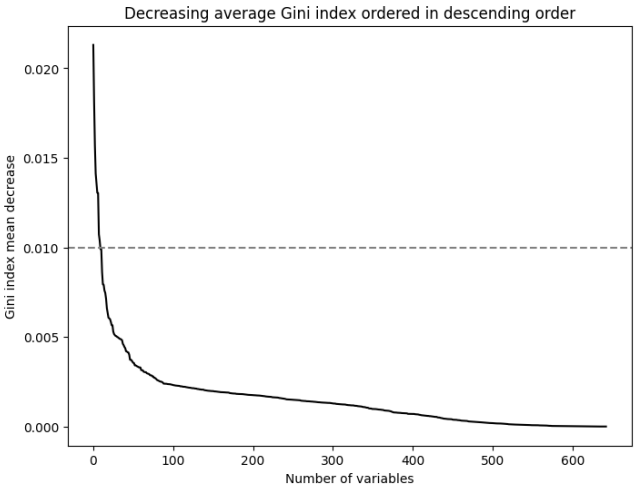

Figure 1: Disminución del índice Gini promedio en función de la cantidad de variables

3. Selección de características: Se identificaron las 50 variables más importantes (de un total de 643) ordenadas en función de su importancia Gini. Estas variables fueron seleccionadas porque aportaron la mayor cantidad de información para predecir correctamente los resultados en el modelo de Random Forest.

4. Visualización: Se generó un gráfico de la importancia acumulada de las variables, lo que permite visualizar cómo el índice Gini disminuye a medida que se consideran más variables. Este gráfico muestra claramente que, más allá de cierto número de variables, la ganancia adicional en la precisión del modelo es marginal, justificando la selección de un subconjunto de 50 variables.

Este conjunto de 50 variables seleccionadas fue luego utilizado para ajustar y evaluar otros modelos de clasificación: KNN, SVM, y Regresión Logística.

Variables de interés:

El dataset analizado está compuesto diversas variables que describen características sociodemográficas, educativas, de infraestructura y salud en los hogares. Estas variables han sido categorizadas en varias áreas clave:

1. Demografía del Hogar: Incluye variables como la cantidad de miembros del hogar, la edad promedio de los integrantes, así como el porcentaje de varones y la distribución por edades. Estas variables nos permiten entender la composición de los hogares, tanto en términos de género como de estructura etaria.

2. Educación: Enfocada en el nivel educativo de los adultos y los padres/madres del hogar. Además, se evalúa el nivel máximo de educación alcanzado por los miembros del hogar, lo cual es fundamental para medir el acceso y calidad educativa.

3. Infraestructura y Recursos del Hogar: Estas variables reflejan las condiciones de vida en los hogares, incluyendo el número de habitaciones, el acceso a tecnología (como computadoras o tablets) y la cantidad de libros disponibles para los niños.

4. Salud: Las variables de esta categoría están centradas en el estado de salud de los niños del hogar, el acceso a información sobre métodos anticonceptivos y la prevalencia de enfermedades, como la tos o diarrea. Estas métricas nos ayudan a evaluar el acceso a servicios de salud y la prevalencia de condiciones básicas de salud en los hogares.

5. Ingreso y Asistencia Social: Esta sección aborda variables relacionadas con el índice de riqueza de los hogares y la dependencia de transferencias sociales, brindando una visión de la estabilidad financiera y el acceso a ayudas gubernamentales o sociales.

6. Violencia y Percepciones: En esta categoría se incluyen variables que miden las percepciones y creencias sobre la violencia en la crianza, así como el acceso de los niños a programas educativos y de socialización, lo que refleja las actitudes y el entorno de desarrollo infantil en los hogares.

7. Interacciones Familiares: Estas variables miden la presencia de los padres en el hogar y el estatus de vida de los padres de los niños, proporcionando información sobre la estructura familiar y el apoyo parental.

3 Clasificadores Usados

Para este tipo de problema de clasificación, enfocado en la predicción del riesgo de violencia física contra niños, se utilizaron estas familias de clasificadores:

Random Forest

El algoritmo **Random Forest** es un modelo de ensamblaje que construye múltiples árboles de decisión de forma independiente y luego combina sus predicciones para obtener una clasificación más precisa. Además, permite manejar datasets con muchas características y puede capturar interacciones complejas entre variables. Es robusto a datos faltantes y menos propenso a overfitting.

En este caso, utilizamos **GridSearchCV** para ajustar los hiperparámetros, como el número de árboles en el bosque (`n_estimators`), la profundidad máxima de los árboles (`max_depth`), y el número mínimo de muestras para dividir un nodo (`min_samples_split`). El mejor conjunto de hiperparámetros encontrados fue:

- `max_depth`: None (sin límite)
- `min_samples_split`: 10
- `n_estimators`: 200

Support Vector Machine (SVM)

El algoritmo **SVM** busca encontrar un hiperplano que maximice el margen entre diferentes clases. Utiliza transformaciones mediante kernels para clasificar datos no lineales. En este caso, se utilizó el kernel radial (RBF) y el ajuste de hiperparámetros como `C` (que controla el margen del hiperplano) y `gamma` (que controla la influencia de cada punto de datos). Se utilizó **RandomizedSearchCV** con validación cruzada K-Fold, y los mejores hiperparámetros encontrados fueron:

- `kernel`: 'rbf' (radial basis function)
- `gamma`: 0.01
- `C`: 0.01

K-Nearest Neighbors (KNN)

El algoritmo **KNN** clasifica un punto basándose en los puntos más cercanos a él. Se utilizan las distancias entre puntos para determinar la clase más común entre los vecinos cercanos. Se probaron distintos números de vecinos (`n_neighbors`) y pesos para cada vecino (`weights`), manteniendo la distancia euclidiana. Los mejores hiperparámetros encontrados mediante K-Fold cross-validation fueron:

- `n_neighbors`: 11
- `weights`: 'uniform' (todos los vecinos tienen el mismo peso)

Regresión Logística

El modelo de **Regresión Logística** es un modelo lineal que estima la probabilidad de que un punto de datos pertenezca a una clase. Utiliza una función sigmoidal para transformar las predicciones a probabilidades. Se ajustó el hiperparámetro `C`, que controla la regularización (una forma de evitar el sobreajuste), a través de **K-Fold cross-validation**. Los valores probados de `C` determinaron que el mejor hiperparámetro fue:

- `C`: 0.01 (mayor regularización, evitando sobreajuste)

4 Resultados

Métrica	RF	SVM	KNN	Logística
Umbral (Índice de Youden)	0.4212	0.4095	0.4545	0.3645
Accuracy	0.6149	0.5786	0.5464	0.6031
Precision	0.5461	0.5088	0.4768	0.5291
Recall	0.6232	0.6053	0.5505	0.7084
F1 Score	0.5821	0.5529	0.5110	0.6058
ROC AUC	0.6543	0.6007	0.5739	0.6639

Table 1: Métricas por Clasificador con Umbral Óptimo

	No Castigo	Castigo
No Castigo	658	599
Castigo	277	673

Table 2: Matriz de Confusión Regresion Logistica con Umbral Ajustado (0.3645)

A partir de las métricas de rendimiento obtenidas, podemos observar que la Regresión Logística ha mostrado el mejor desempeño global entre los clasificadores evaluados, considerando que todos utilizaron las mismas 50 variables. Este modelo logró el mayor valor de Sensibilidad (Recall), con un 0.7084, lo que indica que es el más efectivo detectando correctamente los casos de "Castigo". En la matriz de confusión, se observa que logró el mejor balance en las predicciones correctas de "Castigo" y "No Castigo". Además, la Regresión Logística obtuvo un F1 Score de 0.6058 y un ROC AUC de 0.6639, ambos superiores a los demás modelos, lo que sugiere un buen balance entre precisión y sensibilidad.

Por otro lado, el Random Forest también tuvo un buen rendimiento en términos de precisión (Precision) con un valor de 0.5461 y un F1 Score de 0.5821, lo que lo convierte en un modelo competitivo.

Sin embargo, el SVM y el KNN se quedaron rezagados en la mayoría de las métricas, siendo notable que ambos tienen menores valores en ROC AUC y F1 Score en comparación con la Regresión Logística y el Random Forest.

4.1 Predictores

Variables Socioeconómicas

El índice de riqueza (`wscore`) tiene un impacto negativo en la probabilidad de violencia, lo cual es consistente con lo que se podría esperar: los hogares más pobres son más propensos a la violencia física contra los niños. Este hallazgo refleja la vulnerabilidad de los sectores más desfavorecidos, donde las limitaciones económicas y las condiciones de vida precarias podrían generar entornos asociados con conductas violentas. La pobreza no solo afecta el acceso a recursos materiales, sino también la capacidad del hogar para acceder a apoyo psicosocial y servicios que podrían mitigar estas situaciones.

Aprobación de la violencia

La variable `endorse_violence_any`, que mide si en el hogar se aprueba el uso de la violencia como método de crianza, resulta ser uno de los factores más relevantes en el modelo, con un coeficiente muy significativo. Esto refleja que las creencias y actitudes hacia la violencia juegan un papel crucial en la perpetuación de la misma. En hogares donde los cuidadores justifican o aprueban el uso de castigos físicos, la probabilidad de que los niños sean víctimas de violencia física es mucho mayor.

Dinámica del hogar

La composición y dinámica del hogar también influyen significativamente en la probabilidad de violencia. Variables como el número de niños en el hogar, la proporción de varones, y el número de miembros mayores de 14 años afectan de manera diversa la probabilidad de violencia física. Por ejemplo, se observa que un mayor número de niños (particularmente aquellos de entre 1 y 14 años) está asociado con un mayor riesgo de violencia.

La proporción de varones en el hogar también resulta significativa, especialmente entre los niños de 5 a 17 años, lo que sugiere que los hogares con una mayor proporción de niños varones son más propensos a la violencia física. En contraste, la presencia de miembros mayores de 14 años en el hogar parece reducir la probabilidad de violencia.

Salud y educación

La variable relacionada con la cobertura de salud en el hogar (`cobertura_salud_pct`) también muestra una relación significativa, lo que sugiere que los hogares con mayor acceso a servicios de salud presentan una mayor probabilidad de reportar violencia.

Además, la falta de conocimiento sobre métodos anticonceptivos (`contracep_unkown_any`) y la discriminación relacionada con el VIH también aparecen como factores significativos.

Al comparar los resultados obtenidos en nuestro trabajo con los presentados en el estudio *"Machine learning and public policy: Early detection of physical violence against children" de María Edo, Victoria Oubiña y Marcela Svarc*, surgen las siguientes observaciones clave para los modelos de Random Forest y Regresión Logística:

Random Forest: En el paper, el umbral óptimo seleccionado fue del 48%, mientras que en nuestro trabajo fue del 42.12%, lo que sugiere una estrategia de optimización similar. La exactitud obtenida en ambos estudios es comparable, con un 63% en el paper y un 61.49% en nuestro análisis. Sin embargo, nuestro modelo mostró una sensibilidad significativamente mayor (62.32% frente a 49%), lo que sugiere una mejor detección de casos positivos de violencia. La precisión fue similar, con un 59% en el paper y 54.61% en nuestro trabajo. Además, el ROC AUC de nuestro modelo fue de 0.6543, lo que indica una capacidad razonable para discriminar entre las clases, aunque esta métrica no fue reportada en el paper.

Regresión Logística (Lasso-logit en el paper): En el paper, el umbral óptimo fue del 40.1%, mientras que en nuestro trabajo fue del 36.45%, lo que indica estrategias de optimización alineadas. La exactitud fue del 64% en el paper, frente al 60.31% en nuestro análisis, mostrando que ambos modelos son competitivos. En términos de sensibilidad, nuestro trabajo alcanzó un 70.84%, superior al 63% reportado en el paper, lo que sugiere una mejor capacidad de identificación de hogares en riesgo. La precisión fue ligeramente menor en nuestro análisis (52.91% frente al 57%), pero sigue siendo coherente con los hallazgos del paper. Además, nuestro modelo de Regresión Logística mostró un ROC AUC de 0.6639, lo que refleja una buena capacidad discriminatoria, aunque esta métrica no fue reportada en el estudio original.

Ambos modelos, Random Forest y Regresión Logística, en nuestro trabajo mostraron desempeños que son comparables a los presentados en el paper. Sin embargo, en nuestro análisis, la Regresión Logística se destacó por su mayor sensibilidad, lo que sugiere una mayor capacidad para identificar correctamente los casos de violencia física en los hogares. Este resultado es particularmente importante dado que la identificación temprana de estos casos es el objetivo principal del estudio.

El modelo de Random Forest también mostró un buen rendimiento, especialmente en términos de precisión, pero la Regresión Logística demostró ser la opción preferida en este contexto, al equilibrar adecuadamente precisión y sensibilidad.

5 Conclusiones

El presente trabajo intentó demostrar la eficacia de los modelos de aprendizaje automático para la predicción del riesgo de violencia física contra niños en hogares de Argentina, utilizando datos de la Encuesta MICS. La aplicación de técnicas de clasificación como Regresión Logística y Random Forest permitió identificar importantes predictores de violencia, como las creencias en torno a la violencia, el nivel socioeconómico y la estructura del hogar.

Entre los modelos evaluados, la Regresión Logística ha demostrado el mejor desempeño general, con la mayor sensibilidad (0.7084), lo que la convierte en la opción más eficaz para detectar correctamente los hogares en los que es más probable que ocurra violencia física contra niños. Este modelo, además, mostró un buen equilibrio entre precisión y sensibilidad, reflejado en su F1 Score y ROC AUC superiores a los obtenidos por otros clasificadores. El Random Forest, aunque ligeramente inferior en algunos aspectos, también mostró un rendimiento competitivo, con una precisión (0.5461) que lo convierte en una opción válida para este tipo de problemas.

El análisis de los coeficientes de la Regresión Logística ha permitido identificar que variables como el nivel socioeconómico, la aprobación de la violencia como método de crianza y la estructura demográfica del hogar juegan un papel crucial en la predicción de la violencia. Los hogares más pobres y aquellos donde se justifica el uso de la violencia presentan un mayor riesgo, mientras que la presencia de adultos mayores en el hogar parece tener un efecto protector.

Estos resultados sugieren que las intervenciones más efectivas deben enfocarse en varios frentes desde una perspectiva de políticas públicas. En primer lugar, es crucial abordar las creencias culturales en torno a la violencia mediante campañas de concientización que promuevan métodos de disciplina no violentos. En segundo lugar, es necesario mejorar las condiciones económicas de los hogares más vulnerables mediante políticas de inclusión social y económica. Finalmente, el fortalecimiento de los servicios de salud y educación, junto con el acceso a apoyo psicosocial, es fundamental para reducir los factores de riesgo asociados con la violencia infantil.

En conclusión, este estudio demostró el potencial de los modelos de aprendizaje automático para identificar hogares en riesgo de violencia física contra niños. El uso de estas herramientas predictivas puede ser un gran apoyo para los trabajadores sociales y policy-makers, mejorando la detección temprana de casos de violencia y facilitando la asignación más efectiva de recursos para la protección infantil.