


## Determinantes y Modelos Predictivos de Violencia Basada en Género en Paraguay: Un Estudio sobre datos de la ENSIMUP 2021.

### *Determinants and Predictive Models of Gender-Based Violence in Paraguay: A Study on ENSIMUP 2021 Data.*

 Alfonso González, Agustina <sup>1</sup>

 Beck, Federico Javier <sup>2</sup>

1. Universidad Nacional de Concepción, Facultad de Ciencias Económicas y Administrativas, Concepción - Paraguay.
2. Universidad Nacional de Concepción, Facultad de Ciencias Económicas y Administrativas, Concepción - Paraguay.

### RESUMEN

Este estudio examina los determinantes de la violencia basada en género en Paraguay utilizando algoritmos de machine learning y datos de la Encuesta ENSIMUP 2021. Se analizaron esferas de violencia como la pública, privada y familiar, identificando factores clave como la edad, el nivel de ingresos y el número de hijos. A través del modelo predictivo Random Forest, la edad se destacó como el principal predictor de la violencia, seguida de variables socioeconómicas como el nivel educativo y las condiciones familiares. Los resultados del estudio sugieren que la implementación de modelos predictivos puede apoyar a las políticas públicas en la detección temprana de casos de violencia, optimizando la asignación de recursos para una actuación más eficaz en la prevención. Además, se sugiere la importancia de continuar investigando para abordar otros aspectos no considerados, como las repercusiones sociales y psicológicas de la violencia.

**Palabras clave:** VIOLENCIA BASADA EN GÉNERO, MACHINE LEARNING, ENSIMUP.

Cómo referenciar este artículo/  
How to reference this article

Alfonso González, A., Beck, F. Determinantes y Modelos Predictivos de Violencia Basada en Género en Paraguay: Un Estudio sobre datos de la ENSIMUP 2021. Rev. Cien. Humanidades 2024; 3(1): 48-63.


### ABSTRACT

This study examines the determinants of gender-based violence in Paraguay using machine learning algorithms and data from the ENSIMUP 2021 survey. Various spheres of violence, such as public, private, and family, were analyzed, identifying key factors like age, income level, and the number of children. The Random Forest predictive model highlighted age as the main predictor of violence, followed by socioeconomic variables such as education level and family conditions. The study suggests that predictive models can assist public policies in the early detection of violence cases, optimizing resource allocation for more effective action in prevention. Additionally, further research is recommended to explore social and psychological impacts not considered in this study.

**Key words:** GENDER-BASED VIOLENCE, MACHINE LEARNING, ENSIMUP

Fecha de recepción: 9 de octubre del 2024. Fecha de aceptación: 12 de noviembre del 2024

\*Autor de correspondencia: Alfonso González, Agustina. email: [agustinalf.aa@gmail.com](mailto:agustinalf.aa@gmail.com)

 Este es un artículo fue publicado en acceso abierto, bajo licencia de Creative Commons Attribution – Non Commercial (BY-NC) 4.0 Internacional.

## INTRODUCCIÓN

La Organización Mundial de la Salud (OMS) define la violencia como el uso intencional de la fuerza física o el poder, real o amenazado, contra otra persona, grupo o comunidad, lo que resulta en un alto riesgo de lesión, daño psicológico, muerte o privación. La violencia es prevenible y requiere un enfoque basado en datos para la formulación de políticas efectivas (World Health Organization, 2014). La percepción de lo que constituye un daño y los comportamientos aceptables dependen de contextos culturales, lo que complica la formulación de políticas universales. Para desarrollar políticas efectivas, es esencial reconocer las características específicas de las mujeres violentadas, incluyendo su contexto socioeconómico, nivel educativo y antecedentes culturales, estos influyen en cómo experimentan la violencia (Escobar Manero, et al.,2024).

La violencia basada en género (VBG) constituye un problema social crítico que afecta a mujeres en todo el mundo (World Health Organization, 2014). La violencia contra las mujeres sigue siendo uno de los problemas sociales más persistentes y devastadores en América Latina, afectando a millones de mujeres cada año. En Paraguay, la situación es particularmente alarmante. Según los datos de la Encuesta Nacional sobre Situación de las Mujeres en Paraguay (ENSIMUP) 2021, el 78.5% de las mujeres han experimentado algún tipo de violencia a lo largo de su vida, ya sea psicológica, física, sexual o económico-laboral. Estas cifras

indican que la prevalencia de la violencia en Paraguay es significativamente mayor que en muchos países vecinos. Estos resultados demuestran no solo la magnitud del problema, sino también la importancia de desglosar estos datos para comprender las características sociodemográficas de las mujeres afectadas. Identificar estos factores es crucial para desarrollar políticas públicas más efectivas y adaptadas a las necesidades locales.

Los datos proporcionados por la Encuesta Nacional sobre Situación de las Mujeres en Paraguay (ENSIMUP) 2021 exponen la magnitud de este problema: casi 8 de cada 10 mujeres han experimentado algún tipo de violencia a lo largo de su vida. Estos datos destacan que la violencia psicológica afecta al 57.8% de las mujeres, la violencia física al 25.7%, la violencia sexual al 60.9%, y la violencia económico-laboral al 25%. Estos hallazgos evidencian no solo la prevalencia de la violencia, sino también la necesidad de intervenciones específicas basadas en un análisis riguroso de las condiciones socioeconómicas y demográficas de las mujeres afectadas.

El presente estudio tiene como objetivo: identificar las características que inciden en la violencia de género en Paraguay mediante el uso de algoritmos de machine learning utilizando datos de la Encuesta Nacional sobre Situación de las Mujeres en Paraguay (ENSIMUP), para ello se emplea técnicas avanzadas de machine learning aplicadas a los datos de la Encuesta Nacional sobre

Situación de las Mujeres en Paraguay (ENSIMUP) para identificar las características específicas que inciden de manera significativa en la violencia basada en género. Al analizar estos datos de manera detallada, se busca identificar qué factores socioeconómicos, educativos, demográficos y de conocimiento de las leyes de protección están más estrechamente relacionados con la ocurrencia de violencia. Este enfoque no solo amplía nuestro conocimiento sobre los elementos que contribuyen a la violencia, sino que también proporciona una base sólida para el diseño de políticas públicas más efectivas y focalizadas.

Se plantea la siguiente hipótesis: las mujeres con un mayor nivel educativo y conocimiento de las leyes de protección contra la violencia presentan una menor prevalencia de violencia.

En este contexto, el estudio se enmarca dentro de los Objetivos de Desarrollo Sostenible (ODS), específicamente en el ODS 9: Industria, Innovación e Infraestructura, el ODS 5: Igualdad de Género y el ODS 16: Paz, Justicia e Instituciones Sólidas. La aplicación de técnicas avanzadas de machine learning y el uso de TICs no solo permiten un análisis más profundo de los datos relacionados con la violencia de género, sino que también facilitan la identificación de factores determinantes que influyen en su prevalencia. Estos hallazgos contribuirán directamente al diseño de políticas públicas más adaptadas, efectivas y focalizadas en la erradicación de la violencia de género. Además, la capacidad de las TICs para

procesar grandes volúmenes de información en tiempo real ofrece una base sólida para la creación de sistemas más eficientes que permitan mejorar la protección, la justicia y la seguridad de las mujeres. Al alinear este análisis con los objetivos globales de igualdad de género y paz social, el estudio no solo promueve el empoderamiento de las mujeres, sino también la construcción de sociedades más inclusivas, justas y seguras.

## METODOLOGIA

El uso de machine learning ha emergido como una herramienta crucial en la predicción y análisis de la violencia basada en género, proporcionando la capacidad de manejar grandes volúmenes de datos y descubrir patrones complejos que no son evidentes mediante métodos tradicionales. Los métodos híbridos de machine learning han demostrado ser particularmente efectivos en la evaluación del riesgo de reincidencia en crímenes de género, mostrando que la combinación de modelos predictivos, como Random Forest y Nearest Centroid, mejora significativamente la precisión y la eficacia de las predicciones de riesgo, al tiempo que optimiza los recursos de protección policial (González-Prieto et al., 2023).

En un contexto similar, se han utilizado técnicas de machine learning para clasificar y predecir la vulnerabilidad socioeconómica en áreas rurales de Indonesia. Aunque estos estudios no se centraron exclusivamente en la violencia de género, los métodos y enfoques aplicados pueden ser utilizados para entender cómo los factores

socioeconómicos y contextuales influyen en la vulnerabilidad a la violencia. La capacidad de estos modelos para manejar múltiples variables y su robustez en la clasificación los hace relevantes para la identificación de factores de riesgo en la violencia basada en género (Yuliawan et al., 2022).

Este estudio se basa en un análisis de datos secundarios provenientes de la Encuesta Nacional sobre Situación de las Mujeres en Paraguay (ENSIMUP) del año 2021. Se realiza mediante un enfoque descriptivo de los factores determinantes de la violencia basada en género.

Como parte de la metodología de trabajo se utiliza el Modelo CRISP-DM (Cross Industry Standard Process for Data Mining), el cual es un enfoque ampliamente reconocido en el campo del análisis de datos, que ofrece una estructura clara para la ejecución de proyectos de minería de datos. Este modelo organiza el proceso en seis fases fundamentales, lo que permite a los investigadores abordar sus estudios de forma estructurada y eficiente. Su flexibilidad facilita su aplicación en una variedad de proyectos, especialmente aquellos que requieren múltiples iteraciones para mejorar los resultados y estrategias (Marbán, Mariscal & Segovia, 2009).

El modelo CRISP-DM fue clave para organizar el análisis de los datos de la Encuesta Nacional sobre Situación de las Mujeres en Paraguay (ENSIMUP). El proceso comenzó con una comprensión detallada del problema de la violencia de género, definiendo objetivos concretos para las

predicciones y análisis posteriores. Seguidamente, se realizó una etapa exhaustiva de comprensión y preparación de los datos, que incluyó tareas como la limpieza, normalización y selección de las características más relevantes. Continuando con el proceso de análisis y construcción de los modelos predictivos, se procedió a aplicar una serie de técnicas avanzadas para optimizar el rendimiento de los algoritmos. Inicialmente, se realizó un proceso de selección de características, identificando las 15 más relevantes para la predicción, lo cual permitió enfocar el modelo en las variables que más influyen en la ocurrencia de violencia. Posteriormente, se aplicó un proceso de dummificación a las variables categóricas. Este paso fue fundamental para transformar las variables cualitativas en representaciones numéricas que los modelos pudieran interpretar eficazmente. La dummificación consiste en la creación de variables binarias (0 o 1) para representar cada una de las categorías, lo que asegura que los algoritmos no asignen relaciones arbitrarias entre categorías y, en cambio, traten cada categoría de manera independiente.

Tras este paso, se reentrenaron los modelos, ajustando los algoritmos a los datos dummificados, lo que permitió mejorar la precisión de las predicciones. Como resultado, se seleccionaron las 20 características más importantes identificadas por el modelo (Random Forest), las cuales se destacan por su influencia significativa en la predicción de la violencia. Estas características incluyen factores

demográficos y socioeconómicos como la edad, el número de hijos, el nivel educativo, y el ingreso total, entre otros. Estas variables representan patrones clave que reflejan una mayor o menor probabilidad de experimentar violencia en función de condiciones específicas, proporcionando un marco robusto para interpretar los resultados.

En la fase de modelado, se emplearon algoritmos avanzados de machine learning, como Random Forest, Support Vector Machine (SVM) y XGBoost, con el fin de desarrollar modelos predictivos. Estos modelos fueron evaluados minuciosamente a través de métricas de rendimiento tales como precisión, recall, F1-score y el área bajo la curva ROC (ROC-AUC). Aunque no se completó la fase de implementación, la evaluación realizada proporciona una base sólida para futuras aplicaciones prácticas y para el desarrollo de políticas públicas.

El dataset utilizado en este estudio proviene de la Encuesta Nacional sobre Situación de las Mujeres en Paraguay 2021. Esta encuesta recopila información detallada sobre las experiencias de las mujeres en relación con diferentes tipos de violencia, así como sus características sociodemográficas y económicas. El objetivo de la encuesta es proporcionar una visión comprensiva de la prevalencia y los factores asociados con la violencia basada en género en Paraguay. En este estudio, se trabajó con un conjunto de datos de 3,276 mujeres encuestadas. Originalmente, los datos incluían múltiples registros por cada participante, lo que requería una estructuración adecuada para

garantizar que cada fila del conjunto de datos representara los datos completos de una sola mujer encuestada. Este ajuste fue esencial para mantener la consistencia y precisión en el análisis. Las características utilizadas se centraron en aspectos demográficos, socioeconómicos y antecedentes de violencia previa (experiencia o no de violencia, sin profundizar en los tipos de violencia que pudiera experimentar) enfocándose en tres ámbitos definidos por el estudio: público familiar, privado familiar, y privado en la vida en pareja. El ámbito público familiar se refiere a la violencia sufrida en contextos fuera del hogar, como en lugares de trabajo o espacios públicos. El ámbito privado familiar abarca los eventos de violencia ocurridos dentro del hogar, pero en situaciones no directamente relacionadas con la pareja de la mujer, como conflictos con otros miembros de la familia. Finalmente, el ámbito privado en la vida en pareja comprende la violencia ejercida tanto por la pareja actual como por ex parejas, que para fines de este estudio fueron analizadas de manera conjunta. Este enfoque permitió un análisis específico de los tipos de violencia y las características sociodemográficas más relevantes, aunque otros aspectos, como las consecuencias de la violencia, no fueron abordados en profundidad. Durante el preprocesamiento, se prestó especial atención a la limpieza de datos y la imputación de valores faltantes de manera que no se alteraran los valores verdaderos de los datos, asegurando que los

resultados fueran fieles a la realidad observada.

Parte crucial de la investigación se basa en los modelos predictivos, los mismos son herramientas fundamentales en el análisis de datos, especialmente en contextos complejos como la predicción de la violencia basada en género (VBG). En el contexto de la VBG, los modelos predictivos ayudan a detectar casos potenciales de violencia y a comprender las dinámicas subyacentes que perpetúan estas conductas, facilitando la implementación de estrategias de prevención más efectivas.

Los modelos predictivos pueden clasificarse en dos categorías principales: modelos basados en reglas y modelos de aprendizaje automático. Los modelos basados en reglas son más sencillos y se construyen a partir de lógica predefinida de decisiones, lo que los hace menos flexibles, pero más interpretables. En contraste, los modelos de aprendizaje automático emplean algoritmos avanzados para detectar patrones complejos en grandes volúmenes de datos sin requerir intervención humana constante. Estos modelos han demostrado ser particularmente útiles en la predicción de fenómenos sociales complejos, como la VBG, debido a su capacidad para adaptarse y mejorar con la incorporación de nuevos datos (González-Prieto et al., 2023).

Los modelos a ser analizados son: Regresión logística, Random Forest y XGBoost.

La Regresión Logística es ampliamente reconocida como uno de los modelos más utilizados para problemas de clasificación

binaria, especialmente en contextos donde se busca predecir la presencia o ausencia de un evento específico, como la violencia de género. Este modelo estadístico estima la probabilidad de que ocurra un evento basado en una o más variables independientes, lo que permite evaluar cómo diferentes factores afectan la probabilidad de que se produzca dicho evento (Hosmer, Lemeshow, & Sturdivant, 2013).

El modelo de Random Forest es una extensión de los árboles de decisión que mejora la precisión y la robustez al combinar los resultados de múltiples árboles de decisión entrenados en diferentes subconjuntos del mismo conjunto de datos. Este enfoque de ensamble, donde cada árbol en el bosque emite un voto y la clase con más votos se selecciona como la predicción final, reduce significativamente el riesgo de sobreajuste que a menudo afecta a los modelos de árboles de decisión individuales (Breiman, 2001).

XGBoost es un algoritmo de boosting basado en árboles de decisión que ha ganado popularidad por su alto rendimiento y eficiencia computacional. Este modelo de ensamble mejora iterativamente los resultados de los árboles de decisión individuales al enfocarse en los errores de predicción de los modelos anteriores, ajustando las predicciones y mejorando la precisión general del modelo (Chen & Guestrin, 2016). Estudios recientes han mostrado su efectividad en la identificación de factores de riesgo y la predicción de violencia de género, convirtiéndolo en una

herramienta valiosa para los investigadores y formuladores de políticas que buscan aplicar enfoques de análisis de datos avanzados para abordar problemas sociales complejos (González-Prieto et al., 2023).

La evaluación del rendimiento es una etapa crucial en el desarrollo de modelos predictivos, ya que proporciona una medida de la efectividad del modelo en la tarea para la cual fue diseñado. En el contexto de la violencia basada en género, la evaluación rigurosa del rendimiento de los modelos es esencial para garantizar que las predicciones sean precisas y útiles para la toma de decisiones políticas. La selección de métricas de evaluación adecuadas permite a los investigadores comparar diferentes modelos y elegir el más apropiado para su propósito específico.

La precisión es una métrica básica que indica la proporción de predicciones correctas realizadas por el modelo en relación con el total de predicciones. En el contexto de la VBG, la precisión muestra qué tan bien el modelo puede identificar tanto los casos de violencia (positivos) como los casos de no violencia (negativos). Sin embargo, en problemas de clasificación desbalanceados, como la VBG, la precisión puede ser engañosa, ya que un modelo que predice predominantemente la clase mayoritaria puede tener una alta precisión sin ser efectivo en identificar la clase minoritaria (Goutte & Gaussier, 2005).

El recall mide la capacidad del modelo para identificar correctamente todos los casos positivos. En el contexto de la violencia de género, un alto recall significa que el

modelo es efectivo para detectar la mayoría de los casos de violencia, lo cual es crucial para prevenir y mitigar los efectos de la violencia. Un modelo con bajo recall podría pasar por alto casos de violencia, lo que sería inaceptable en aplicaciones donde la seguridad y el bienestar de las personas están en juego (Powers, 2011).

El F1-score es la media armónica de la precisión y el recall, proporcionando una métrica equilibrada que es útil cuando se busca un compromiso entre ambas. Un F1-score alto indica que el modelo tiene un buen equilibrio entre la precisión y la capacidad de detectar todos los casos positivos. Esta métrica es particularmente valiosa en contextos donde las clases están desbalanceadas y se desea minimizar tanto los falsos positivos como los falsos negativos (Sasaki, 2007).

El AUC (Área Bajo la Curva) es una métrica que evalúa la capacidad de un modelo para distinguir entre clases positivas y negativas a través de diferentes umbrales de clasificación. Un AUC cercano a 1 indica que el modelo tiene una excelente capacidad de discriminación, mientras que un AUC cercano a 0.5 sugiere que el modelo no es mejor que una predicción aleatoria. En la evaluación de modelos predictivos para la VBG, un alto AUC es deseable ya que indica que el modelo puede diferenciar eficazmente entre situaciones de violencia y no violencia, independientemente del umbral utilizado para la clasificación (Bradley, 1997).

AP (Promedio de Precisión) es una métrica que resume la precisión en

diferentes niveles de recall, proporcionando una visión completa del rendimiento del modelo en la clasificación positiva. Esta métrica es útil para evaluar cómo cambia la precisión del modelo cuando se ajustan los umbrales para maximizar el recall, especialmente en problemas de clasificación desbalanceada.

La evaluación del rendimiento de los modelos predictivos a través de múltiples métricas permite obtener una visión integral de su efectividad y adecuación para tareas específicas. En el caso de la violencia basada en género, donde las consecuencias de los errores de predicción pueden ser significativas y potencialmente perjudiciales, la utilización de varias métricas es esencial para asegurar que los modelos no solo son precisos, sino también sensibles a los casos de interés (Fawcett, 2006). La precisión y la capacidad de identificación precisa son factores críticos, ya que un error en la clasificación podría significar la falta de intervención en casos de violencia o, inversamente, la movilización innecesaria de recursos para falsos positivos. El uso de modelos predictivos también facilita una orientación preventiva, ayudando a identificar patrones tempranos de violencia que pueden ser abordados antes de que se desarrollen en situaciones más graves. La implementación de políticas basadas en predicciones permite actuar con anticipación, desarrollando programas de educación, apoyo y prevención dirigidos a las comunidades más vulnerables (Yuliawan et al., 2022). La utilización de modelos predictivos avanzados, como los métodos

híbridos de machine learning, ha demostrado ser una herramienta valiosa para la evaluación del riesgo de crímenes de género. Estos modelos permiten la identificación de patrones complejos y la predicción de futuros incidentes de violencia, mejorando la capacidad de respuesta de las políticas públicas. González-Prieto et al. (2023) destacan que no solo permite la detección temprana de situaciones de violencia, sino que también ayuda a priorizar los casos más críticos para una intervención efectiva. Esto se alinea con los Objetivos de Desarrollo Sostenible (ODS), particularmente con el ODS 5, que busca lograr la igualdad de género y empoderar a todas las mujeres y niñas, y el ODS 16, que promueve sociedades pacíficas e inclusivas mediante la justicia y la protección. Este enfoque proactivo no solo protege a las potenciales víctimas, sino que también contribuye a la creación de una sociedad más segura y equitativa.

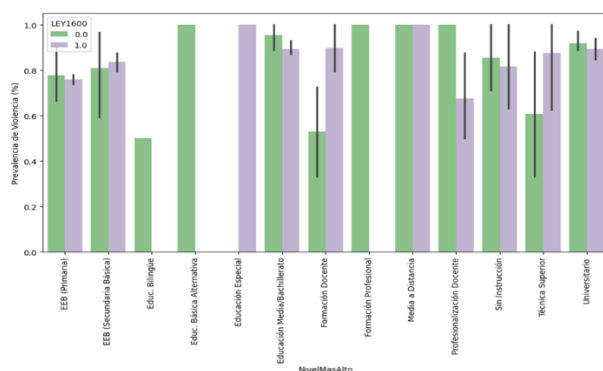
## RESULTADOS

Las características identificadas reflejan los factores que más influyen en la ocurrencia de violencia en función de los datos recogidos. En este análisis se abordaron varias características demográficas, socioeconómicas y educativas que podrían influir en la prevalencia de violencia entre las mujeres, dentro de estas fue evaluado el conocimiento de las leyes, de cara a poner a prueba la hipótesis dada de la investigación. A continuación, se presentan los gráficos que exploran la relación entre el nivel educativo, el conocimiento de las leyes

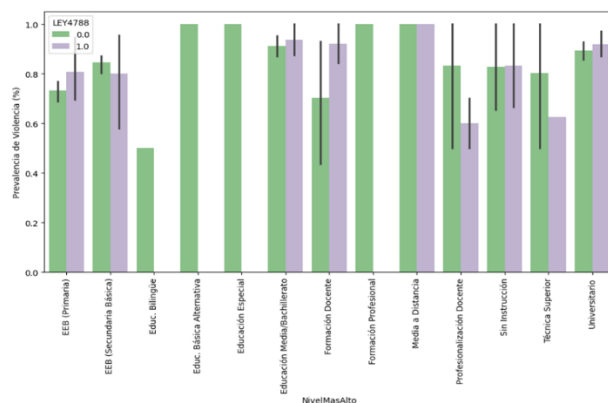


de protección contra la violencia, y la prevalencia de violencia experimentada. Estos gráficos buscan analizar si un mayor nivel educativo y el conocimiento de las leyes están asociados con una menor prevalencia de violencia entre las mujeres, proporcionando una visión crítica de los factores que pueden influir en su protección frente a la violencia

**Figura N° 1.** Prevalencia de Violencia por Nivel Educativo y Conocimiento de la Ley N° 1600.



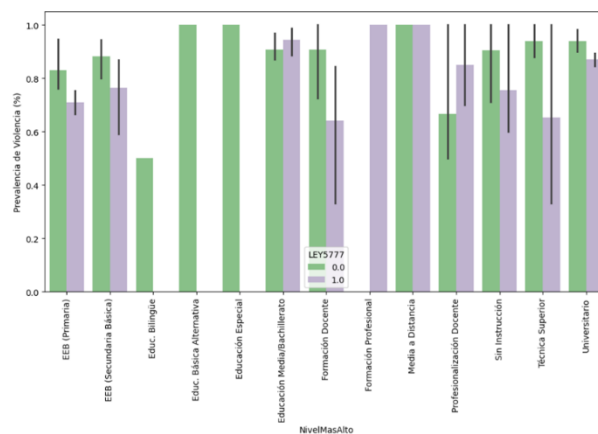
**Figura N° 2.** Prevalencia de Violencia por Nivel Educativo y Conocimiento de la Ley N° 4788.



La hipótesis plantea que las mujeres con mayor nivel educativo y conocimiento de las leyes tienen menor prevalencia de violencia. Con respecto al conocimiento de las leyes: Solo la Ley 5777 muestra tener un impacto significativo en la reducción de la violencia

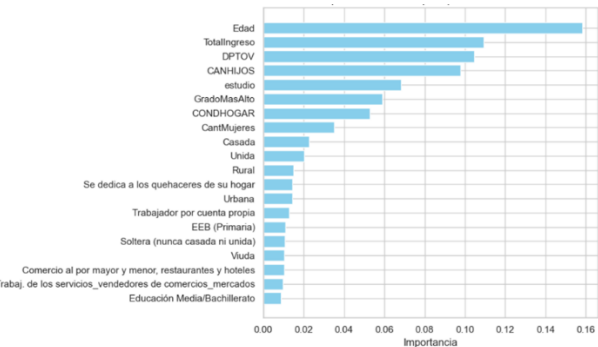
entre quienes la conocen, lo que respalda parcialmente la hipótesis. Sin embargo, el conocimiento de la Ley 1600 y la Ley 4788 no demuestra estar relacionado con una disminución significativa en la prevalencia de violencia.

**Figura N° 3.** Prevalencia de Violencia por Nivel Educativo y Conocimiento de la Ley N° 5777.



La Ley 5777 muestra un impacto más claro en mujeres con niveles educativos superiores (técnica superior y formación docente), lo que podría sugerir que este grupo tiene más acceso a recursos y apoyo para usar los recursos legales en su defensa. A pesar de que la educación superior muestra una ligera reducción en la prevalencia de violencia (especialmente con la Ley 5777), la diferencia no es tan marcada. Esto indica que, aunque la educación puede ser un factor mitigante, no es suficiente por sí sola para garantizar la reducción de la violencia. Además, las mujeres con educación superior también parecen enfrentar barreras para usar eficazmente las leyes de protección.

**Figura 4.** Importancia de las Características en la Predicción de Violencia Basada en Género.



Las 20 características más importantes identificadas por el mejor modelo, Random Forest, proporcionan una visión clara de los principales factores que influyen en la predicción de la violencia. La edad resultó ser la variable más importante, lo que sugiere que este factor está estrechamente relacionado con la prevalencia de la violencia, sugiriendo que en diferentes etapas de la vida las mujeres pueden ser más vulnerables a la violencia. El ingreso total también se posiciona como un predictor clave, indicando que los factores socioeconómicos juegan un papel significativo en la exposición a la violencia. El número de hijos y el año de estudios completados destacan la relevancia de la estructura familiar y el nivel educativo, mientras que el grado educativo más alto alcanzado refuerza la importancia de la educación como factor protector.

Variables como el estado civil, las dinámicas del hogar y las condiciones geográficas juegan un papel clave, lo que indica que las situaciones maritales y el lugar de residencia pueden influir en el

riesgo de violencia. Aunque algunas características, como el tipo de ocupación o el nivel educativo medio, tienen un impacto menor, siguen siendo relevantes en la comprensión global del fenómeno. En particular, las mujeres dedicadas a las labores del hogar, que dependen de la economía doméstica o trabajan por cuenta propia, pueden estar en una situación de mayor vulnerabilidad. Estos hallazgos proporcionan un marco integral que combina factores demográficos, educativos, económicos y laborales, ofreciendo una visión más completa para entender los predictores de la violencia.

Desempeño de Modelos de Clasificación

En este estudio, se evaluó el desempeño de varios modelos de clasificación para predecir la violencia basada en género. La Tabla 1 presenta un resumen de las métricas de desempeño utilizadas para comparar estos modelos: precisión (accuracy), precisión de clasificación positiva (precision), sensibilidad (recall), F1 Score, área bajo la curva ROC (AUC), y promedio de precisión (AP).

**Tabla 1.** Desempeño de los modelos de clasificación

Modelo	Accuracy	Precision	Recall	F1 Score	AUC	AP
Regresión Logística	0.82	0.82	1.00	0.90	0.67	0.90
Random Forest	0.82	0.82	0.99	0.90	0.66	0.90
Gradient Boosting	0.80	0.82	0.96	0.89	0.64	0.89

Análisis de Modelos de Machine Learning etiquetados para la Predicción de la Violencia Contra la Mujer

Este análisis compara tres modelos de machine learning utilizados para predecir la violencia contra la mujer, utilizando las siguientes métricas: Accuracy (precisión general), Precision (precisión de clasificación positiva), Recall (sensibilidad), F1 Score, AUC (Área bajo la curva ROC) y AP (Promedio de Precisión). Estas métricas proporcionan una visión integral del rendimiento de cada modelo en la tarea de predicción.

En el análisis comparativo de los modelos de Regresión Logística, Random Forest y Gradient Boosting, se observan rendimientos muy similares en precisión, recall y F1 score, aunque cada modelo tiene ligeras diferencias en métricas clave como el AUC y el Average Precision (AP).

Aunque los tres modelos tienen un rendimiento similar, Random Forest se destaca como la mejor opción debido a su equilibrio general en precisión, recall y F1 score, y su AP de 0.90, lo que lo convierte en un modelo robusto para predecir casos de violencia. Aunque su AUC es ligeramente inferior al de la regresión logística, su alta capacidad para detectar casi todos los casos de violencia y su precisión constante lo hacen la mejor opción entre los modelos evaluados.

## DISCUSION

Los resultados de este estudio proporcionan una perspectiva profunda sobre los determinantes de la violencia de género en Paraguay. El análisis de los datos provenientes de la Encuesta Nacional sobre

Situación de las Mujeres en Paraguay (ENSIMUP) mediante modelos de machine learning como el Random Forest ha permitido identificar las características más relevantes que influyen en la ocurrencia de la violencia, confirmando y ampliando hallazgos de estudios previos.

### Factores determinantes

La edad se destacó como la variable más importante en la predicción de la violencia, lo cual está en consonancia con los hallazgos de McDougal et al. (2021), quienes identificaron que las mujeres en diferentes etapas de la vida son vulnerables a distintas formas de violencia debido a factores como la dependencia económica, el desequilibrio de poder en las relaciones y la falta de acceso a recursos. Este estudio refuerza la idea de que las mujeres más jóvenes, en particular, son especialmente susceptibles a la violencia debido a su limitada autonomía financiera y social. Sin embargo, la vulnerabilidad no se limita a las jóvenes; las mujeres de mayor edad también enfrentan riesgos debido a su dependencia económica o emocional, factores que a menudo están presentes en relaciones de largo plazo o en matrimonios donde las dinámicas de poder se agravan con el tiempo. Esta evidencia subraya la necesidad de políticas de intervención que enfoquen a grupos de mujeres en función de sus ciclos de vida y su contexto socioeconómico.

El nivel de ingresos y las condiciones de la vivienda también emergen como predictores claves. Investigaciones previas como las de González-Prieto et al. (2023) y

Vyas & Watts (2009) han demostrado que las mujeres en condiciones económicas precarias tienen menos oportunidades para escapar de situaciones de violencia debido a su dependencia financiera. Este análisis refuerza esos hallazgos, destacando que las mujeres en situaciones de mayor vulnerabilidad económica, especialmente aquellas dedicadas a los quehaceres del hogar o que dependen de ingresos irregulares o trabajos informales, enfrentan mayores barreras para salir de relaciones abusivas. La precariedad económica, combinada con dinámicas de poder desequilibradas dentro del hogar, crea un ciclo que perpetúa la violencia, lo que sugiere la urgencia de implementar programas que fortalezcan la independencia económica de las mujeres, proporcionando redes de apoyo financiero y oportunidades laborales más estables.

Además, el estudio muestra que la cantidad de mujeres en el hogar, así como las responsabilidades familiares, como el número de hijos (CANHIJOS), tienen un impacto relevante en las dinámicas que perpetúan la violencia. Un hogar con más mujeres puede generar una sobrecarga en los roles asignados y aumentar las tensiones relacionadas con el cuidado, la crianza o el soporte emocional, elementos que también se asocian con situaciones de violencia. Esta perspectiva se alinea con estudios que resaltan cómo las dinámicas familiares y los roles tradicionales de género son factores clave en la exposición a la violencia de género. Así, las dinámicas del hogar, que abarcan la distribución de responsabilidades

y el poder dentro de la familia, influyen en la capacidad de las mujeres para negociar situaciones de conflicto o buscar apoyo.

Finalmente, las diferencias geográficas entre áreas rurales y urbanas también juegan un rol significativo. Las mujeres en zonas rurales pueden tener menos acceso a recursos de apoyo, redes de seguridad y oportunidades de educación y empleo, lo que exacerba la violencia. Este aspecto está respaldado por investigaciones que destacan la brecha en acceso a servicios entre áreas rurales y urbanas, lo que lleva a una mayor exposición a la violencia en comunidades más aisladas o con menos recursos (McDougal et al., 2021). En este sentido, las políticas públicas deben enfocarse no solo en las dinámicas económicas y familiares, sino también en superar las barreras geográficas para asegurar que las mujeres de todas las regiones tengan acceso a los recursos necesarios para su protección y empoderamiento.

En conjunto, este análisis de los factores demográficos, económicos, educativos y laborales no solo proporciona un marco integral para entender mejor los predictores de la violencia, sino que también subraya la urgencia de diseñar políticas basadas en evidencia que puedan intervenir en las causas estructurales de la violencia de género, tanto a nivel individual como colectivo.

### **Comparación con Estudios Previos**

Los resultados del presente estudio subrayan la importancia de las

características sociodemográficas y socioeconómicas en la predicción de la violencia basada en género, lo cual es consistente con hallazgos de investigaciones anteriores. La identificación de la Edad como el factor más influyente en la predicción de violencia se alinea con los estudios de McDougal et al. (2021), quienes también encontraron que las mujeres jóvenes son más vulnerables a la violencia de género. Este patrón podría deberse a la mayor exposición de las mujeres jóvenes a dinámicas de poder desiguales y a entornos de trabajo y sociales menos estables.

La relevancia del Total de Ingreso y Condición del hogar en la predicción de la violencia destaca la relación entre la estabilidad económica y las condiciones de vida, y la vulnerabilidad a la violencia. González-Prieto et al. (2023) también identificaron que las condiciones socioeconómicas desempeñan un papel crucial en la evaluación del riesgo de crímenes de género, sugiriendo que mujeres en situaciones de pobreza o con inestabilidad económica tienen menos recursos para escapar o denunciar situaciones de violencia. Los resultados de este estudio sugieren que las políticas enfocadas en mejorar el nivel de ingresos, la educación, las condiciones de vivienda y la promoción de relaciones familiares saludables pueden ser eficaces en la prevención de la violencia de género. Las mujeres con menor nivel educativo y menores ingresos son más vulnerables a la violencia debido a la dependencia económica y la falta de autonomía (Becker,

1964; Vyas & Watts, 2009). Mejorar las condiciones de vivienda puede reducir el estrés y los conflictos, lo que disminuye el riesgo de violencia (Sadownik, 2023).

Los resultados de este estudio coinciden con investigaciones realizadas en otros contextos. Por ejemplo, Rodríguez-Rodríguez et al. (2020) encontraron que las mujeres con niveles educativos más altos eran menos propensas a sufrir violencia, un hallazgo respaldado por los datos de este estudio. Sin embargo, este estudio también muestra que, incluso entre las mujeres con mayor educación, persisten barreras significativas para acceder y utilizar las herramientas legales disponibles.

Los modelos predictivos como el Random Forest y la Regresión Logística se destacaron por su capacidad para manejar grandes volúmenes de datos y proporcionar predicciones precisas, lo que apoya la creciente literatura que resalta la utilidad de los enfoques de machine learning en el análisis de fenómenos sociales complejos como la violencia de género (González-Prieto et al., 2023). Como sugieren Yuliawan et al. (2022), el uso de técnicas avanzadas de Machine Learning permite a los responsables de la formulación de políticas adaptar sus estrategias basadas en datos reales y en la identificación de grupos de alto riesgo. Esto permite una asignación más efectiva de los recursos y el diseño de intervenciones específicas que pueden reducir la incidencia de violencia de manera más eficaz.

## Modelos predictivos contra la Violencia Basada en Género

Se ha demostrado que los modelos predictivos pueden ser aplicados para predecir la violencia de género utilizando datos históricos. Estudios realizados en España utilizando algoritmos como Random Forest han mostrado cómo los datos pueden ser utilizados para prever la ocurrencia de violencia basada en género, facilitando la planificación de recursos y la implementación de políticas públicas efectivas (Rodríguez-Rodríguez et al., 2020). El uso de modelos predictivos, como Random Forest, representa un avance significativo en la lucha contra la violencia basada en género (VBG). Estos modelos permiten identificar patrones ocultos y factores clave en grandes volúmenes de datos que los métodos tradicionales no pueden detectar con la misma precisión. Al predecir el riesgo de violencia, es posible implementar intervenciones preventivas tempranas, lo que optimiza la asignación de recursos y mejora la efectividad de las políticas públicas.

**Conflictos de interés:** Los autores declaran no tener conflicto de interés.

## REFERENCIAS BIBLIOGRÁFICAS

- Arias, Eugenia Fabián, Vilcas Baldeón, L. M., & Alberto Bueno, Y. (2019). Factores de riesgo de violencia a la mujer de parte del cónyuge. *Socialium*, 3(1), 69–96.  
<https://doi.org/10.31876/sl.v3i1.67>
- Becker, Gary Stanley. (1975). Human capital: A theoretical and empirical analysis, with special reference to education. Midway Reprint.
- Bradley, Andrew P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7), 1145–1159.  
[https://doi.org/10.1016/s0031-3203\(96\)00142-2](https://doi.org/10.1016/s0031-3203(96)00142-2)
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
- Dahlberg, L. L., & Krug, E. G. (2002). Violence—a global public health problem. En E. G. Krug, L. L. Dahlberg, J. A. Mercy, A. B. Zwi, & R. Lozano (Eds.), *World report on violence and health* (pp. 1–56). World Health Organization.
- Escobar Manero, E. et al. (2024). Salud y Género. Observatorio de Salud Estudio de Comunicación.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874.
- González-Prieto, Ángel, Brú, A., Nuño, J. C., & González-Álvarez, J. L. (2023). Hybrid machine learning methods for risk assessment in gender-based crime. *Knowledge-Based Systems*, 260(110130), 110130.  
<https://doi.org/10.1016/j.knosys.2022.110130>

- Goode, William J. (1971). Force and Violence in the Family. *Journal of marriage and the family*, 33(4), 624. <https://doi.org/10.2307/349435>
- Goutte, Cyril, & Gaussier, E. (2005). A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. En *Lecture Notes in Computer Science* (pp. 345–359). Springer Berlin Heidelberg.
- Heise, L. L. (1998). Gender-based violence: An overview of the issues. *Reproductive Health Matters*, 6(11), 6–13.
- Kohavi, Ron. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, 1137–1143.
- Marbn, Scar, Mariscal, G., & Segovi, J. (2009). A data mining & Knowledge discovery process model. En *Data Mining and Knowledge Discovery in Real Life Applications*. I-Tech Education and Publishing.
- McDougal, Lotus, Dehingia, N., Bhan, N., Singh, A., McAuley, J., & Raj, A. (2021). Opening closed doors: using machine learning to explore factors associated with marital sexual violence in a cross-sectional study from India. *BMJ Open*, 11(12), e053603. <https://doi.org/10.1136/bmjopen-2021-053603>
- Mujeres, O. N. U. (2017a). EL EMPODERAMIENTO ECONÓMICO DE LA MUJER EN EL CAMBIANTE MUNDO DEL TRABAJO. Unwomen.org. <https://www.unwomen.org/sites/default/files/Headquarters/Attachments/Sections/CSW/61/CSW-Conclusions-61-SP-WEB.pdf>
- Mujeres, O. N. U. (2017b). El empoderamiento económico de la mujer en el cambiante mundo del trabajo. <https://www.unwomen.org/sites/default/files/Headquarters/Attachments/Sections/CSW/61/CSW-Conclusions-61-SP-WEB.pdf>
- Organization, W. H. (2014). *Global Status Report on Violence Prevention 2014*.
- Powers, David M. W. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. <https://doi.org/10.48550/ARXIV.2010.16061>
- Rodríguez-Rodríguez, Ignacio, Rodríguez, J.-V., Pardo-Quiles, D.-J., Heras-González, P., & Chatzigiannakis, I. (2020). Modeling and forecasting gender-Based Violence through Machine Learning techniques. *Applied Sciences* (Basel, Switzerland), 10(22), 8244. <https://doi.org/10.3390/app10228244>
- Sasaki, Yutaka. (2007). The truth of the F-measure. *Teach Tutor Mater*.
- Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to*

Data Mining (2a ed.). Pearson Education.

Vyas, Seema, & Watts, C. (2009). How does economic empowerment affect women's risk of intimate partner violence in low and middle income countries? A systematic review of published evidence. *Journal of International Development*, 21(5), 577–602.

<https://doi.org/10.1002/jid.1500>

Yuliawan, D., Hakim, D. B., Juanda, B., & Fauzi, A. (2022). Classification and prediction of rural socio-economic vulnerability (IRSV) integrated with social-ecological system (SES). *Decision Science Letters*, 11(2), 223–234.