

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



Relación entre autonomía económica y violencia de género en el estado de Baja California: caracterización y modelado predictivo

TRABAJO RECEPCIONAL que para obtener el **GRADO** de
Maestra en Ciencia de Datos

Presenta:
Alexandra Arciniega Chavez

Director:
Maestro Byron Michael Motta Bonilla

Tlaquepaque, Jalisco, 26 de mayo de 2026

Relación entre autonomía económica y violencia de género en el estado de Baja California: caracterización y modelado predictivo

Alexandra Arciniega Chavez

Resumen

El presente trabajo analiza la relación entre la autonomía económica y la violencia de género en el estado de Baja California utilizando datos de la ENDIREH 2021. A través de un Análisis Exploratorio de Datos (EDA), se caracterizó el entorno sociodemográfico y la autonomía económica de las mujeres; los resultados sugieren que, al no presentarse patrones lineales marcados, la violencia de género debe abordarse como un fenómeno multicausal y multidimensional. Posteriormente, se realizó un modelado predictivo comparando seis algoritmos de aprendizaje automático, desde una Regresión Logística hasta modelos de mayor complejidad. Se determinó que el Random Forest obtuvo el mejor rendimiento, evaluado primordialmente a través del F1-Score por su capacidad para balancear la precisión y la sensibilidad en la detección de casos positivos. Tras un proceso de selección de importancia de variables para reducir la dimensionalidad y el ruido, se identificó que la edad y el ingreso mensual son los factores más determinantes. Estos hallazgos sugieren que las mujeres jóvenes que inician su vida laboral enfrentan un riesgo mayor debido a alteraciones en las dinámicas tradicionales de poder. El estudio destaca la eficacia de la ciencia de datos para facilitar la identificación de perfiles vulnerables y apoyar a las instituciones gubernamentales en la implementación de medidas de prevención y mitigación de riesgos.

Tabla de Contenidos

	Página
1 Introducción	13
1.1. Contexto	14
1.2. Justificación	17
1.3. Problema	20
1.4. Objetivos	20
1.4.1. Objetivo general	20
1.4.2. Objetivos específicos	20
2 Metodología	23
2.1. Descripción de los datos	23
2.1.1. Fuente de datos	23
2.1.2. Definición del conjunto de datos y selección de variables	25
2.1.3. Carga de datos	29
2.2. Preprocesamiento	29
2.2.1. Preparación y estandarización de los módulos	30
2.2.2. Recodificación de variables	32
2.2.3. Ingeniería de características e indicadores sintéticos	34
2.2.4. Tratamiento de datos faltantes (imputación) .	35
2.2.5. Integración y depuración final	36
2.3. Análisis exploratorio	36
2.3.1. Perfil sociodemográfico	36
2.3.2. Caracterización de la autonomía económica .	39
2.3.3. Análisis comparativo de los factores sociode- mográficos y de autonomía frente a la violencia	44
2.4. Descripción y justificación de los modelos	52
2.4.1. Regresión Logística	52
2.4.2. Análisis Discriminante Lineal (LDA)	53
2.4.3. Máquinas de Vector de Soporte (SVM)	54
2.4.4. Decision Trees	55
2.4.5. Bagging Trees	56
2.4.6. Random Forest	57
2.5. Descripción y justificación de las métricas de evaluación	58
2.5.1. Matriz de confusión	58

2.5.2.	Accuracy	59
2.5.3.	Precision	59
2.5.4.	Recall	59
2.5.5.	F1-Score	60
2.5.6.	AUC-ROC	60
2.6.	Descripción de los experimentos	62
2.6.1.	División del conjunto de datos	62
2.6.2.	Tratamiento de valores atípicos y sesgos en variables continuas	62
2.6.3.	Escalado de variables continuas	62
2.6.4.	Grid Search y Cross-Validation	63
2.6.5.	Balanceo de pesos	63
2.6.6.	Ajuste del umbral	64
2.6.7.	Evaluación de los modelos	64
3	Resultados y discusión.	67
3.1.	Resultados	67
3.1.1.	Comparación entre los modelos	67
3.1.2.	Refinamiento del mejor modelo	69
3.1.3.	Resultados en el conjunto de prueba	71
3.2.	Discusión de los resultados	72
3.2.1.	Comparativa de modelos y análisis de desempeño	72
3.2.2.	Optimización y refinamiento del modelo final	73
3.2.3.	Evaluación final y capacidad de generalización	74
3.2.4.	Interpretación de resultados en el contexto del estudio	74
4	Conclusiones y trabajo futuro.	77
4.1.	Conclusiones	77
4.2.	Trabajo futuro	78
4.2.1.	Inclusión de variables multidimensionales . .	78
4.2.2.	Exploración de algoritmos de mayor comple- jidad	79
4.2.3.	Replicabilidad en otros contextos regionales .	79
4.2.4.	Incorporación de la ENDIREH 2026	79

Índice de figuras

	Página
1.1. Proporción global de mujeres víctimas de violencia (1 de cada 3). Fuente: Elaboración propia con datos de la OMS [2].	13
1.2. Representación geográfica del estado de Baja California, región donde se centra el estudio de autonomía económica y violencia de género. Fuente: Elaboración propia con datos del INEGI [8].	16
1.3. Comparativa de la Población Económicamente Activa por género a nivel nacional y estatal (Baja California). Fuente: Elaboración propia con datos de la ENOE [12], [13].	17
1.4. Comparativa nacional y estatal de la prevalencia de violencia de género según la ENDIREH 2021. Fuente: Elaboración propia con datos del INEGI [15].	18
1.5. Arquitectura conceptual de las variables predictoras e indicadores de salida del modelo. Fuente: Elaboración propia.	19
2.1. Distribución por edades de las mujeres en la muestra. . .	37
2.2. Distribución del nivel educativo alcanzado según grupo de edad.	37
2.3. Distribución original del estado conyugal y su correspondiente recodificación binaria.	38
2.4. Distribución de la muestra por tipo de dominio geográfico. .	38
2.5. Distribución de la muestra según condición de actividad económica.	39
2.6. Distribución de la muestra según posición ocupacional. .	40
2.7. Distribución general de los ingresos mensuales totales (A) y detalle excluyendo valores en cero (B).	40
2.8. Distribución de la muestra según la capacidad de decisión sobre el uso del dinero.	41
2.9. Perfil promedio de las dimensiones sintéticas de autonomía.	42

2.10. Distribución de la muestra según las actitudes de la pareja hacia la autonomía económica de la mujer.	43
2.11. Prevalencia de violencia en los últimos 12 meses.	44
2.12. Distribución de la muestra por edad según la prevalencia de violencia.	45
2.13. Distribución de los ingresos mensuales totales según la prevalencia de violencia.	45
2.14. Relación entre edad e ingreso mensual total según la prevalencia de violencia.	46
2.15. Prevalencia de violencia según nivel educativo.	47
2.16. Prevalencia de violencia según condición de actividad económica.	47
2.17. Prevalencia de violencia según posición ocupacional. . .	48
2.18. Prevalencia de violencia según la capacidad de decisión sobre el uso de dinero.	48
2.19. Prevalencia de violencia según autonomía patrimonial. .	49
2.20. Densidad del índice de actitud según la prevalencia de violencia.	49
2.21. Comparativa de prevalencia de violencia por dimensiones de autonomía.	50
2.22. Matriz de correlación (Coeficiente de Spearman).	51
2.23. Ejemplo de curva ROC [38].	61
3.1. Comparativa de matrices de confusión de los modelos. .	68
3.2. Comparativa de las curvas ROC de los modelos.	69
3.3. Importancia de las variables.	69
3.4. Matriz de confusión del modelo Random Forest optimizado.	70
3.5. Curva ROC del modelo Random Forest optimizado. . .	71
3.6. Matriz de confusión del modelo Random Forest optimizado en el conjunto de prueba.	72
3.7. Curva ROC del modelo Random Forest optimizado en el conjunto de prueba.	72

Índice de tablas

	Página
1.1. Clasificación y definiciones de los tipos de violencia de género según la LGAMVLV [5].	14
1.2. Clasificación y definiciones de las modalidades de ocurrencia de la violencia de género según la LGAMVLV [5].	15
2.1. Variables seleccionadas del módulo TSDem (variables sociodemográficas) [20].	26
2.2. Variables seleccionadas del módulo TB_SEC_IV (variables sobre autonomía económica: ingresos) [20].	27
2.3. Variables seleccionadas del módulo TB_SEC_XV (variables de autonomía económica: toma de decisiones) [20].	28
2.4. Definición de la variable objetivo proveniente del módulo TB_VD [20].	29
2.5. Descripción de los filtros por módulo.	30
2.6. Mapeo y normalización de identificadores: Módulo TSDem.	30
2.7. Mapeo y normalización de identificadores: Módulo TB_SEC_IV.	31
2.8. Mapeo y normalización de identificadores: Módulo TB_SEC_XV.	31
2.9. Mapeo y normalización de identificadores: Módulo TB_VD.	32
2.10. Identificadores de variables con codificación binaria.	32
2.11. Identificadores de variables categóricas recodificadas.	33
2.12. Grado de autonomía patrimonial según la titularidad de bienes y activos.	43
2.13. Ejemplo de matriz de confusión.	58
3.1. Mejores hiperparámetros para cada modelo de aprendizaje automático.	67
3.2. Métricas de desempeño de los mejores modelos de aprendizaje automático.	68
3.3. Mejores hiperparámetros del modelo Random Forest optimizado.	70

3.4. Métricas de desempeño del modelo Random Forest optimizado.	70
3.5. Métricas de desempeño del modelo Random Forest optimizado en el conjunto de prueba.	71

Dedicado a...

Mi papá Juan Manuel Arciniega, que aunque ya no está presente, me acompaña en este logro. Gracias por inculcarme disciplina y el amor por el aprendizaje continuo y la ciencia. Me inspiras a querer cambiar el mundo.

Mi mamá Cecilia Chavez, por creer siempre en mí y cuidar de mí incansablemente. Gracias por tu apoyo incondicional y por el gran esfuerzo que hiciste, y sigues haciendo, para que yo pueda lograr todo lo que me propongo. Este logro es tanto mío como tuyo.

Mis hermanas Elizabeth y Rocio, por estar conmigo y apoyarme desde el día que nací. A mis sobrinas Lizeth y Ximena, por su chispa con la que alegran cada uno de mis días. Que este trabajo sirva para que sepan que no hay límites en sus sueños, y que sea un granito de arena para construir un lugar más seguro para ustedes.

Mi novio José Luis, por ser mi mayor refugio y por su paciencia y comprensión infinita. Gracias por creer en mi capacidad incluso cuando yo misma dudaba, y por celebrar cada pequeño logro de esta maestría. No pude haber pedido un mejor compañero.

Finalmente, dedico este trabajo a las mujeres que, a través de sus datos, nos permiten visibilizar su realidad, y a mis colegas en las áreas de STEM. Que la tecnología sea siempre una herramienta al servicio de la equidad y un espacio donde nuestras voces sigan cobrando fuerza.

1 Introducción

La Organización de las Naciones Unidas (ONU) define la violencia de género como aquellos actos dañinos dirigidos hacia una persona o grupo de personas en razón de su género. Si bien tanto hombres como mujeres pueden ser víctimas, suele asociarse a la violencia ejercida en contra de las mujeres debido a la manera desproporcionada en la que las afecta [1].

A nivel global, se estima que una de cada tres mujeres ha sufrido violencia física o sexual al menos una vez en su vida, cifra que ha permanecido prácticamente constante durante los últimos 20 años [2]. Debido a este impacto, se considera una de las violaciones a los derechos humanos más generalizadas en el mundo [1].



Figura 1.1: Proporción global de mujeres víctimas de violencia (1 de cada 3). Fuente: Elaboración propia con datos de la OMS [2].

Como estrategia frente a esta problemática, Lourdes Colinas, Oficial Nacional y Especialista en Empoderamiento en ONU Mujeres México, afirma que las mujeres que gozan de autonomía económica presentan una menor exposición a situaciones de violencia. De acuerdo con la CEPAL, esta autonomía se define como la capacidad de las mujeres para generar, acceder y controlar ingresos propios, además de disponer de su tiempo y de la propiedad [3], [4].

Colinas recalca la importancia de afrontar las barreras del mercado laboral para facilitar la participación de las mujeres en la economía, contribuyendo así al bienestar tanto de las mujeres como de la sociedad en su conjunto [3].

El presente trabajo busca analizar si existe una relación entre la autonomía económica y la violencia de género para, en caso de confirmarse, proponer posibles estrategias de prevención.

1.1 Contexto

El fenómeno de la violencia de género en México se rige por la Ley General de Acceso de las Mujeres a una Vida Libre de Violencia (LGAMVLV), publicada en 2007. Esta normativa establece la coordinación entre la Federación, las entidades estatales y los municipios para prevenir, sancionar y erradicar la violencia, garantizando el bienestar y el desarrollo integral de las mujeres [5]. La LGAMVLV clasifica la violencia en las siguientes siete categorías:

Tipo de violencia	Definición
Violencia psicológica	Es cualquier acto u omisión que dañe la estabilidad psicológica; puede consistir en negligencia, abandono, celotipia, insultos, humillaciones, marginación o amenazas. Estas conductas pueden conducir a la víctima a la depresión, el aislamiento, la pérdida de autoestima o incluso al suicidio.
Violencia física	Cualquier acto que infrinja daño no accidental mediante el uso de la fuerza física, armas, objetos o sustancias corrosivas, tóxicas e inflamables que puedan provocar lesiones internas, externas o ambas.
Violencia patrimonial	Acto u omisión que afecta la supervivencia de la víctima mediante la transformación, sustracción, destrucción o retención de objetos, documentos personales, bienes, valores o recursos económicos destinados a satisfacer sus necesidades.
Violencia económica	Acción u omisión del agresor que afecta la supervivencia financiera de la víctima a través de limitaciones para controlar sus ingresos o la percepción de un salario menor por un trabajo igual dentro de un mismo centro laboral.
Violencia sexual	Cualquier acto que degrada o daña el cuerpo y la sexualidad de la víctima, atentando contra su libertad, dignidad e integridad física. Es una expresión de abuso de poder que implica la supremacía masculina al denigrar a la mujer y concebirla como objeto.
Violencia a través de interpósita persona	Acto u omisión que, para dañar a la mujer, se dirige contra sus hijas, hijos, familiares o personas allegadas, independientemente de si existe cohabitación o una relación vigente con el agresor.
Otras formas	Cualesquiera otras conductas análogas que lesionen la dignidad, integridad o libertad de las mujeres.

Tabla 1.1: Clasificación y definiciones de los tipos de violencia de género según la LGAMVLV [5].

Asimismo, la LGAMVLV define las modalidades de violencia de género como las formas, manifestaciones o ámbitos de ocurrencia en que se presentan estas agresiones. La Ley reconoce las siguientes siete modalidades [5]:

Modalidad de violencia	Definición
Violencia familiar	Acto abusivo de poder u omisión intencional dirigido a dominar, someter o agredir a las mujeres dentro o fuera del domicilio familiar, por parte de personas con quienes tengan o hayan tenido un vínculo sentimental, de parentesco o responsabilidades de cuidado.
Violencia laboral y docente	Ejercida por personas con un vínculo laboral o educativo con la víctima, independientemente de la jerarquía. Consiste en abusos de poder que dañan la autoestima, salud e integridad de la mujer, impidiendo su desarrollo y atentando contra la igualdad. Incluye el hostigamiento y acoso sexual.
Violencia en la comunidad	Actos individuales o colectivos que transgreden los derechos fundamentales de las mujeres en el ámbito público, propiciando su denigración, discriminación o exclusión.
Violencia institucional	Actos u omisiones de servidores públicos que, mediante prejuicios o estereotipos de género, obstaculicen el goce de los derechos humanos de las mujeres o el acceso a políticas públicas de atención y justicia.
Violencia política	Acción u omisión basada en elementos de género que tenga por objeto limitar, anular o menoscabar el ejercicio de los derechos políticos y electorales de las mujeres, así como el desempeño de sus funciones o cargos públicos.
Violencia digital y mediática	Acción dolosa realizada mediante tecnologías de la información para exponer, distribuir o comercializar contenido íntimo sexual de una persona sin su consentimiento, o actos que causen daño a la intimidad y dignidad de las mujeres en el entorno digital.
Violencia feminicida	Forma extrema de violencia de género producto de la violación de derechos humanos y el abuso de poder. Se manifiesta a través de conductas de odio que pueden culminar en feminicidio u otras formas de muertes evitables.

Tabla 1.2: Clasificación y definiciones de las modalidades de ocurrencia de la violencia de género según la LGAMVLV [5].

Para este proyecto, el área de estudio se delimitó al estado de Baja California, como se muestra en la Figura 1.2. Esta entidad ocupa una posición geográfica estratégica al localizarse en la frontera norte de México y colindar con Estados Unidos, la mayor potencia económica a nivel mundial. Entre 2019 y 2021, se registraron 3.3 millones de entradas internacionales a través del cruce fronterizo de Tijuana-San Diego, el cual representa el principal punto de internación terrestre al país [6]. El constante flujo de personas entre ambas ciudades consolida una región transfronteriza y binacional que integra dinámicas económicas y culturales compartidas [7]. Asimismo, el estado destaca por su alta densidad de población inmigrante; según el Censo 2020, residían en la entidad 152,377 personas nacidas en el extranjero, lo que posiciona a Baja California como el estado con mayor número de población extranjera en el país [6].

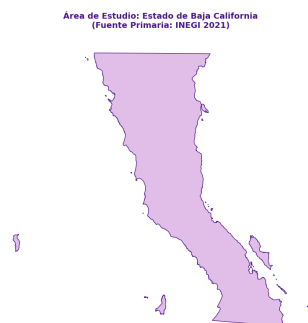


Figura 1.2: Representación geográfica del estado de Baja California, región donde se centra el estudio de autonomía económica y violencia de género. Fuente: Elaboración propia con datos del INEGI [8].

En 2021, se declaró la Alerta de Violencia de Género contra las Mujeres (AVGM) por violencia feminicida en los municipios de Ensenada, Mexicali, Playas de Rosarito, San Quintín, Tecate y Tijuana, extendiéndose a la totalidad del estado de Baja California [9]. La AVGM es un mecanismo de protección legal que se activa ante un incremento crítico de la violencia contra las mujeres en una zona específica, considerando actos como agresiones, desapariciones y feminicidios. El objetivo primordial de esta alerta es exigir a las autoridades la implementación de acciones inmediatas que garanticen la protección de las mujeres, el acceso a la justicia y la prevención de futuros actos violentos [10]. Este panorama crítico en la entidad subraya la urgencia de desarrollar estrategias de prevención.

La autonomía es considerada un factor fundamental para garantizar el ejercicio de los derechos humanos en igualdad de condiciones para hombres y mujeres. La CEPAL estructura la autonomía de las mujeres en tres dimensiones principales: autonomía económica, física y en la toma de decisiones [11]. Este proyecto tiene como objetivo analizar la relación entre la autonomía económica y la violencia de género, por lo que se centrará específicamente en este concepto, complementándolo con el análisis de la toma de decisiones en asuntos financieros.

La autonomía económica se vincula a la posibilidad de controlar activos y recursos. En este sentido, el trabajo remunerado constituye la principal fuente para la obtención de ingresos propios, lo cual no solo influye directamente en la independencia económica de las personas, sino también en su desarrollo personal, la toma de decisiones y la capacidad de enfrentar situaciones de riesgo, como la violencia de pareja [11].

En este aspecto, las mujeres se encuentran en desventaja. Según información de la Encuesta Nacional de Ocupación y Empleo (ENOE) del segundo semestre de 2021, realizada por el Instituto Nacional de Estadística y Geografía (INEGI), el 43.9 % de las mujeres de 15 años o más en México eran económicamente activas, en contraste con el 75.9 % de los hombres [12]. Por otro lado, en Baja California, el 47 % de las mujeres y el 73.7 % de los hombres fueron económicamente activos durante el mismo periodo [13].

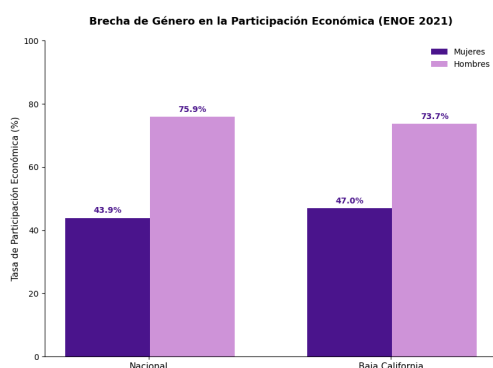


Figura 1.3: Comparativa de la Población Económicamente Activa por género a nivel nacional y estatal (Baja California). Fuente: Elaboración propia con datos de la ENOE [12], [13].

Dos de las principales limitantes para la inserción de las mujeres en el trabajo remunerado son la sobrecarga de tareas de cuidado a terceros y la desigual distribución del trabajo en el hogar. Estos obstáculos, que impiden a las mujeres generar ingresos propios, sumados a la falta de un manejo libre de los recursos, afectan su autonomía económica y limitan la posibilidad de tomar decisiones en otras áreas de su vida [11].

1.2 Justificación

La violencia de género conlleva consecuencias críticas para la salud de las víctimas. Las tasas de depresión, trastornos de ansiedad, embarazos no deseados e infecciones de transmisión sexual son superiores en mujeres que han sufrido violencia en comparación con aquellas que no [2]. Este fenómeno representa un problema de magnitud social cuyas manifestaciones extremas derivan en homicidio o suicidio.

Aunque las mujeres son las principales afectadas, el impacto de la violencia se extiende a su entorno inmediato. Los niños expuestos a agresiones en

el ámbito familiar son vulnerables a desarrollar trastornos conductuales y emocionales. Esta exposición incrementa la probabilidad de que, en la etapa adulta, se reproduzcan dichos patrones, ya sea como perpetradores o como víctimas. Asimismo, la violencia de pareja se asocia con un incremento en las tasas de morbilidad y mortalidad durante la infancia y el periodo lactante [14].

Adicionalmente, la violencia de género genera impactos económicos significativos que afectan al Estado, a las víctimas y a la sociedad. Estos costos abarcan tanto el gasto directo en programas de prevención y refugios, como efectos indirectos traducidos en la pérdida de productividad laboral y la saturación de los sistemas de salud y justicia [2].

Con el objetivo de contribuir al conocimiento sobre este fenómeno, el INEGI realiza la Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares (ENDIREH). Esta encuesta especializada constituye un elemento esencial para conocer la magnitud y prevalencia de la violencia contra las mujeres en México [15].

A nivel nacional, la ENDIREH 2021 estima que el 70.1 % de las mujeres de 15 años o más han experimentado al menos un incidente de violencia a lo largo de su vida, mientras que el 42.8 % reportó violencia durante los últimos 12 meses previos al levantamiento de la encuesta. En Baja California, la prevalencia acumulada alcanza el 69.2 %, mientras que el 37.2 % de la población femenina en el mismo rango de edad sufrió situaciones de violencia durante los últimos 12 meses [15]. Dicha comparativa se visualiza detalladamente en la Figura 1.4.

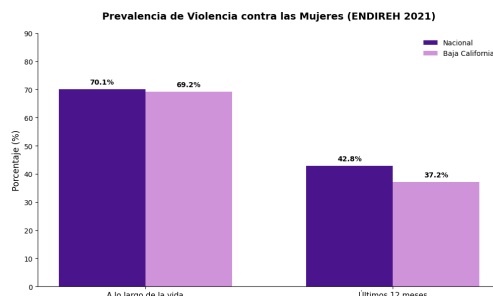


Figura 1.4: Comparativa nacional y estatal de la prevalencia de violencia de género según la ENDIREH 2021. Fuente: Elaboración propia con datos del INEGI [15].

La magnitud de estas cifras y la severidad de sus consecuencias exigen el desarrollo de estrategias de prevención dinámicas. Si bien las instituciones gubernamentales han implementado mecanismos de atención, la integración de tecnologías innovadoras como el aprendizaje automático (*machine learning*) permite abordar el fenómeno desde una perspectiva analítica avanzada. A diferencia de los métodos estadísticos tradicionales, estos modelos poseen la

capacidad de identificar patrones latentes y relaciones no lineales entre variables que podrían resultar imperceptibles mediante análisis convencionales.

Asimismo, la carencia de modelos predictivos orientados a la identificación de riesgos individuales subraya la relevancia de este trabajo. El desarrollo de este modelo es fundamental para facilitar la transición hacia protocolos de detección temprana y permitir la futura implementación de sistemas de alerta. Este enfoque proactivo representa una evolución necesaria frente a los mecanismos de atención correctivos predominantes en la actualidad.

El trabajo profundiza en la autonomía económica como un factor determinante de la vulnerabilidad o protección frente a la violencia. Este concepto trasciende el acceso a recursos financieros; implica el control efectivo y la capacidad de agencia en la toma de decisiones sobre los mismos. Por tal motivo, el presente estudio integra variables vinculadas a la percepción de ingresos, la propiedad de bienes y la autonomía en la gestión económica del hogar.

La selección de la ENDIREH 2021 como fuente primaria se fundamenta en su carácter oficial y en el respaldo técnico del INEGI, lo cual otorga validez y confiabilidad al proyecto. Esta base de datos proporciona un alto nivel de detalle sobre los diferentes tipos y modalidades de violencia en cada entidad federativa, integrando además variables sociodemográficas y de gestión económica. Dichos indicadores permiten realizar un análisis exhaustivo de la relación entre la autonomía financiera y la violencia de género en Baja California, facilitando el desarrollo de un modelo predictivo con alto grado de precisión. Asimismo, el uso de la edición 2021 asegura que el análisis refleje la dinámica reciente del fenómeno en la región.

Variables Predictoras Consideradas en el Estudio

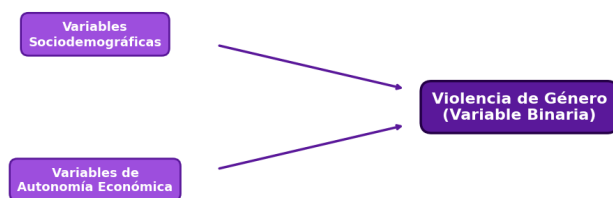


Figura 1.5: Arquitectura conceptual de las variables predictoras e indicadores de salida del modelo.
Fuente: Elaboración propia.

1.3 Problema

Aunque encuestas como la ENDIREH cierran brechas críticas en la disponibilidad de datos sobre violencia de género en México, su naturaleza es predominantemente descriptiva y retrospectiva. Esto plantea la oportunidad de implementar modelos predictivos de aprendizaje automático que integren la autonomía económica como indicador clave para anticipar situaciones de vulnerabilidad. Una anticipación efectiva requiere profundizar en la correlación entre la independencia financiera y el riesgo de violencia; al identificar estos patrones, es posible transitar hacia métodos preventivos oportunos que mitiguen el fenómeno antes de su incidencia.

Ante la escasez de herramientas de ciencia de datos y aprendizaje automático aplicadas al análisis del contexto específico de Baja California, se proponen los siguientes elementos de investigación:

Pregunta de investigación: ¿En qué medida los indicadores de autonomía económica y la violencia de género permiten crear un modelo predictivo, basado en aprendizaje automático, capaz de identificar a mujeres en riesgo en el estado de Baja California?

Hipótesis: Existe una relación significativa entre los indicadores de autonomía económica y la violencia de género, permitiendo que un modelo predictivo de ciencia de datos identifique patrones de perfiles de vulnerabilidad en mujeres del estado.

1.4 Objetivos

1.4.1 Objetivo general

Caracterizar y desarrollar un modelo de clasificación binaria mediante aprendizaje automático para predecir el riesgo de sufrir violencia de género en Baja California, analizando la autonomía económica y factores sociodemográficos como variables predictoras.

1.4.2 Objetivos específicos

1. Seleccionar y codificar las variables de la ENDIREH 2021 que permitan medir las dimensiones de autonomía económica y violencia de género en mujeres de 15 años o más en el estado de Baja California.
2. Realizar el preprocesamiento e ingeniería de características sobre el conjunto de datos, con el fin de construir indicadores de autonomía económica que optimicen la capacidad predictiva del modelo.

3. *Efectuar un análisis exploratorio de datos (EDA) para examinar las distribuciones, identificar patrones y analizar las correlaciones entre variables sociodemográficas, indicadores de autonomía económica y la violencia de género.*
4. *Diseñar y entrenar varios modelos de clasificación para estimar la probabilidad de riesgo de sufrir violencia de género en mujeres de Baja California, basándose en la autonomía económica y factores sociodemográficos.*
5. *Evaluar el rendimiento y comparar los modelos propuestos utilizando métricas de clasificación (F1-Score, Recall, AUC-ROC, Precision, y Accuracy) para seleccionar el modelo más robusto para la predicción del riesgo en la región.*

2 Metodología

2.1 Descripción de los datos

2.1.1 Fuente de datos

Los datos utilizados en esta investigación provienen del Instituto Nacional de Estadística y Geografía (INEGI), organismo público autónomo responsable de normar y coordinar el Sistema Nacional de Información Estadística y Geográfica (SNIEG) en México. El INEGI es la fuente oficial encargada de captar y difundir información estratégica sobre el territorio, la población y la economía del país [16], garantizando la validez institucional de los microdatos empleados para el entrenamiento de los modelos predictivos.

El INEGI produce la Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares (ENDIREH), la cual representa la principal fuente de información sobre la violencia contra las mujeres en México y es un referente internacional por sus estándares éticos y metodológicos. A la fecha, cuenta con cinco ediciones (2003, 2006, 2011, 2016 y 2021) [17], utilizándose la versión más reciente de 2021 como la fuente primaria de datos para el desarrollo del presente modelo predictivo.

Desde 2015, la ENDIREH está clasificada como Información de Interés Nacional (IIN). Esta distinción garantiza que la encuesta emplea una metodología científicamente validada y procedimientos estadísticos alineados con estándares internacionales. Asimismo, su periodicidad de cinco años asegura la comparabilidad temporal y espacial de los datos, otorgándoles un carácter oficial y obligatorio para el diseño, implementación y evaluación de políticas públicas en México [17].

La ENDIREH proporciona datos detallados sobre incidentes de violencia física, económica, sexual, psicológica y patrimonial experimentados por mujeres de 15 años o más. La encuesta segmenta estas experiencias en cinco ámbitos principales: familiar, escolar, laboral, comunitario y de pareja. Asimismo, incluye atributos específicos sobre las víctimas y los agresores, junto con variables multidimensionales que permiten un análisis profundo de la violencia de género en el contexto nacional [18].

Las dimensiones temáticas que integran la encuesta y que sirven de base para la selección de variables predictoras son [17]:

- **Autonomía y recursos:** Situación de la pareja, gestión de ingresos y acceso a recursos económicos.
- **Dinámica social y género:** Opinión sobre roles de género, toma de decisiones y libertad personal.
- **Manifestaciones de violencia:** Experiencias de violencia desglosadas por ámbitos (escolar, laboral, comunitario, familiar y de pareja).
- **Contexto individual y familiar:** Discapacidad, familia de origen, vida en pareja y percepción de conflictos.
- **Indicadores específicos:** Atención obstétrica, violencia sexual en la infancia y situación de mujeres de 60 años o más.

Con el objetivo de garantizar la fiabilidad y el rigor ético de la información, el INEGI implementa protocolos metodológicos basados en recomendaciones internacionales de la Organización Mundial de la Salud (OMS) y la Organización de las Naciones Unidas (ONU). Estos protocolos priorizan los principios de privacidad, consentimiento informado y seguridad, tanto para las entrevistadas como para el personal operativo. La estricta aplicación de estas medidas es fundamental para mitigar el sesgo de reporte en incidentes sensibles y asegurar que los microdatos reflejen con precisión la realidad de las participantes, salvaguardando en todo momento la integridad de las mujeres involucradas [17]. Como parte de este marco de seguridad, el INEGI aplica los siguientes lineamientos:

1. **Homogeneidad de género:** El levantamiento de datos es realizado exclusivamente por personal femenino.
2. **Consentimiento informado:** Es obligatorio obtener la autorización explícita de las participantes antes del inicio de la entrevista.
3. **Privacidad absoluta:** Se garantizan entornos de confidencialidad total durante la aplicación del instrumento para mitigar sesgos de respuesta.
4. **Autocuidado operativo:** Implementación de protocolos de seguridad para el personal de campo.

El fortalecimiento de estas pautas éticas resultó en un incremento del 4.6 % en los niveles de privacidad reportados durante la entrevista en comparación con la edición de 2016 [17].

La ENDIREH emplea un diseño muestral probabilístico, trietápico, estratificado y por conglomerados. Esta estructura garantiza la representatividad de

los datos a nivel nacional y estatal, permitiendo una focalización precisa en el contexto de Baja California. Asimismo, la cobertura incluye estratos urbanos y rurales, lo cual enriquece el análisis de los perfiles sociodemográficos integrados en el modelo. El levantamiento se realizó del 4 de octubre al 30 de noviembre de 2021, alcanzando una muestra nacional de 140,784 viviendas donde se entrevistó a mujeres de 15 años o más [17].

Asimismo, la ENDIREH 2021 se instrumenta mediante cuatro cuestionarios diferenciados según la situación conyugal de las participantes. Esta estructura es determinante para precisar la exposición al riesgo en el ámbito de pareja [19].

1. **Cuestionario general:** Instrumento base para la recolección de características sociodemográficas y del entorno.
2. **Cuestionario A:** Aplicado a mujeres actualmente casadas o en unión libre.
3. **Cuestionario B:** Dirigido a mujeres con situación conyugal de separación, divorcio o viudez.
4. **Cuestionario C:** Diseñado para mujeres solteras (nunca unidas).

Para el presente estudio, se integraron los indicadores de autonomía económica provenientes de la totalidad de los cuestionarios. Esto permitió consolidar una base de datos que abarca tanto a mujeres con antecedentes de vida en pareja (pasada o presente) como a mujeres solteras, asegurando una muestra representativa de las diversas dinámicas de independencia financiera y su relación con el riesgo de violencia.

2.1.2 Definición del conjunto de datos y selección de variables

El conjunto de datos del presente trabajo se restringe a la submuestra de mujeres de 15 años o más residentes en el estado de Baja California. A partir de esta delimitación geográfica, se realizó un proceso de selección de variables extraídas de la ENDIREH 2021, las cuales fueron categorizadas de acuerdo con la dimensión analítica que representan para el modelo:

- **Variables sociodemográficas:** Incluyen factores como la edad, nivel educativo y estado conyugal. Sirven para establecer el perfil de vulnerabilidad de las mujeres.
- **Variables de autonomía económica:** Incluyen indicadores sobre ingresos propios, condición de actividad económica, autonomía en la toma de decisiones sobre el uso del dinero, control de bienes y arreglos sobre temas económicos con la pareja o el hogar. Estas variables permiten medir el empoderamiento financiero.

- **Variable dependiente (objetivo):** *Incidencia de haber sufrido algún tipo de violencia (psicológica, física, sexual, económica o patrimonial) en los últimos 12 meses. Utilizar esta variable como objetivo representa la forma más directa de medir el riesgo reciente de sufrir violencia, tomando en cuenta las cuestiones actuales de autonomía económica de cada mujer.*

La arquitectura de datos de la ENDIREH 2021 se compone de 28 tablas que segmentan las diversas temáticas de la encuesta. Para la construcción del conjunto de datos, se ejecutó una extracción selectiva de variables provenientes de los módulos TSDem, TB_SEC_IV, TB_SEC_XV y TB_VD. El proceso de selección se basó en la relevancia estadística y teórica de cada atributo, categorizándolos de la siguiente manera:

Variables sociodemográficas (módulo TSDem)

TSDem: *Contiene las características sociodemográficas y culturales de cada una de las personas residentes de la vivienda [20].*

Variable	Definición
ID_PER	Número de identificación único de la mujer de 15 años y más.
EDAD	Edad del informante, expresado en años cumplidos.
NIV	Nivel máximo de estudios acreditados.
P2_9	Asistencia a una escuela al momento de la aplicación del cuestionario.
P2_13	Condición de actividad económica en la semana de referencia.
P2_14	Condición de actividad (identifica si realiza alguna actividad informal).
P2_15	Relación que estableció con su empleo o lugar de trabajo en la semana de referencia (jornalero(a), peón(a), empleado(a), patrón(a), etc.).
P2_16	Situación conyugal.
DOMINIO	Agrupaciones de viviendas con características diferenciadas: urbano, complemento urbano y rural.

Tabla 2.1: Variables seleccionadas del módulo TSDem (variables sociodemográficas) [20].

Variables de autonomía económica (módulos TB_SEC_IV y TB_SEC_XV)

Estas variables son el núcleo de la hipótesis, y se utilizarán para construir indicadores de autonomía económica considerando aspectos como los ingresos,

toma de decisiones/arreglos sobre dinero y posesión de bienes.

TB_SEC_IV: Este módulo consolida la información de la pareja o ex-pareja y detalla la estructura de ingresos de la mujer y su contraparte. Asimismo, cuantifica otras fuentes de financiamiento y la titularidad de activos o propiedades [20].

Variable	Definición
ID_PER	Identificador de la mujer.
P4_1	Situación laboral actual de la informante.
P4_2	Ingreso económico mensual aproximado por concepto de trabajo.
P4_2_1	Periodicidad del ingreso laboral (temporalidad del salario).
P4_8_1	Recepción de ingresos por jubilación o pensión.
P4_8_7	Recepción de ingresos por rentas, intereses bancarios u otros medios.
P4_9_1	Monto mensual aproximado recibido por jubilación o pensión.
P4_9_7	Monto mensual aproximado por ingresos de fuentes diversas.
P4_11	Grado de autonomía en la toma de decisiones sobre el uso del dinero.
P4_13_1	Titularidad de terrenos o tierras de cultivo.
P4_13_2	Titularidad de vehículos (automóviles o camionetas).
P4_13_3	Titularidad y control de ahorros personales o familiares.
P4_13_4	Titularidad de la vivienda habitada por la informante.
P4_13_5	Titularidad de inmuebles comerciales (locales, bodegas u oficinas).
P4_13_6	Titularidad de activos comerciales fijos (puestos).
P4_13_7	Titularidad de propiedades residenciales adicionales (casa o departamento).

Tabla 2.2: Variables seleccionadas del módulo TB_SEC_IV (variables sobre autonomía económica: ingresos) [20].

TB_SEC_XV: El módulo incluye información sobre la toma de decisiones en el hogar o en la relación de pareja de las mujeres de 15 años o más [20].

Variable	Definición
ID_PER	Identificador de la mujer.
P15_1AB_1	Autonomía en la toma de decisiones sobre trabajar o estudiar.
P15_1AB_3	Capacidad de decisión sobre el uso y destino del dinero propio.
P15_1AB_4	Autonomía para realizar gastos personales (compras para sí misma).
P15_1AB_7	Capacidad de decisión sobre la administración y ahorro del hogar.
P15_1AB_8	Poder de decisión sobre el destino de los ingresos de la pareja.
P15_2AB_1	Actitud de la pareja ante la decisión de la mujer de trabajar o estudiar.
P15_2AB_3	Actitud de la pareja ante el uso del dinero que la mujer genera o posee.
P15_2AB_4	Actitud de la pareja ante el deseo de la mujer de realizar gastos personales.
P15_2AB_7	Actitud de la pareja ante la administración financiera de la mujer.
P15_2AB_8	Actitud de la pareja ante la intervención de la mujer en el gasto de él.
P15_3AB_1	Arreglos y acuerdos con la pareja para laborar de forma remunerada.
P15_3AB_2	Arreglos y acuerdos con la pareja para realizar compras.
P15_1C_1	Identificación del tomador de decisiones sobre el estudio de la informante.
P15_1C_2	Identificación del tomador de decisiones sobre el trabajo de la informante.
P15_1C_6	Identificación del tomador de decisiones sobre el dinero de la informante.
P15_1C_7	Identificación del tomador de decisiones sobre gastos personales de ella.
P15_2C_1	Tipo de acuerdos y actores involucrados para permitir el estudio de la informante.
P15_2C_2	Tipo de acuerdos y actores involucrados para permitir el acceso al trabajo.
P15_2C_6	Tipo de acuerdos y actores involucrados para el uso del dinero propio.
P15_2C_7	Tipo de acuerdos y actores involucrados para realizar gastos personales.

Tabla 2.3: Variables seleccionadas del módulo TB_SEC_XV (variables de autonomía económica: toma de decisiones) [20].

Variable objetivo: prevalencia de violencia en los últimos 12 meses (módulo TB_VD)

TB_VD: Módulo que sistematiza la incidencia de violencia total, categorizada por ámbito y tipo, tanto en periodos históricos como en los últimos 12 meses anteriores al levantamiento [20].

Variable	Definición
ID_PER	Identificador de la mujer.
VTOT_12M	Indicador binario de prevalencia de violencia. Identifica si la informante experimentó al menos un incidente de violencia (escolar, laboral, comunitario, familiar o de pareja) en el periodo de octubre de 2020 a la fecha de la entrevista.

Tabla 2.4: Definición de la variable objetivo proveniente del módulo TB_VD [20].

2.1.3 Carga de datos

El conjunto de datos utilizado en esta investigación se obtuvo a partir del repositorio oficial de microdatos de la ENDIREH 2021 administrado por el INEGI [21]. El acceso a la información se realizó mediante la descarga de los activos digitales en formato comprimido (ZIP), los cuales integran 28 archivos de valores separados por comas (CSV) que estructuran las diversas dimensiones de la encuesta.

Para la implementación del modelo, se realizó la carga de los cuatro módulos de interés en estructuras de datos tipo DataFrame. Durante esta fase de ingesta, se ejecutó un filtrado selectivo para extraer únicamente las variables (features) previamente especificadas en el diseño metodológico, optimizando así el uso de memoria y garantizando la integridad de los datos de entrada para las etapas de procesamiento posterior.

2.2 Preprocesamiento

El preprocesamiento se orientó a la transformación de las variables originales en un conjunto de datos depurado y estructurado, listo para poder realizar de forma más eficiente su análisis exploratorio y la posterior implementación de diferentes modelos de aprendizaje automático. Las operaciones de limpieza y transformación se realizaron de manera individual en cada módulo para garantizar la consistencia de la clave de unión (ID_PER) previa a la unión de las tablas, para finalizar con un último tratamiento enfocado en consolidar un conjunto de datos libre de inconsistencias.

2.2.1 Preparación y estandarización de los módulos

La fase inicial de procesamiento comprendió la adecuación individual de los módulos TSDem, TB_SEC_IV, TB_SEC_XV y TB_VD. En cada una de estas tablas, se aplicó un protocolo de filtrado basado en dos criterios de inclusión críticos: el geográfico, para restringir las instancias al estado de Baja California, y el demográfico, para seleccionar exclusivamente los registros correspondientes a mujeres de 15 años o más. Este procedimiento garantiza la homogeneidad de los datos previos a la etapa de integración (merge) de los módulos.

Módulo	Condiciones
TSDem	NOM_ENT = 'BAJA CALIFORNIA'
	COD_M15 = 1 - Mujeres con 15 años o más
	SEXO = 2 - Mujer
TB_SEC_IV	NOM_ENT = 'BAJA CALIFORNIA'
TB_SEC_XV	NOM_ENT = 'BAJA CALIFORNIA'
TB_VD	NOM_ENT = 'BAJA CALIFORNIA'
	POBT = 1 - Mujeres con 15 años o más

Tabla 2.5: Descripción de los filtros por módulo.

Posteriormente, se implementó un proceso de normalización mediante el renombrado de variables. Los códigos alfanuméricos técnicos del INEGI fueron sustituidos por identificadores descriptivos bajo una convención de nomenclatura estandarizada. Este procedimiento no solo optimiza la legibilidad del código fuente, sino que reduce la probabilidad de errores durante las etapas de ingeniería de características (feature engineering) y validación del modelo.

El proceso de normalización de identificadores para el módulo TSDem utilizó el mapeo de los códigos originales a nombres descriptivos, detallando en la Tabla 2.6 únicamente aquellas variables que requirieron ser renombradas:

Módulo	Código INEGI	Nombre actualizado
TSDem	P2_9	asiste_escuela
	P2_13	condicion_actividad
	P2_14	verificacion_actividad
	P2_15	posicion_ocupacion
	P2_16	estado_conyugal

Tabla 2.6: Mapeo y normalización de identificadores: Módulo TSDem.

El mapeo de variables para el módulo de ingresos y actividad laboral (TB_SEC_IV) se detalla en la Tabla 2.7:

Módulo	Código INEGI	Nombre actualizado
TB_SEC_IV	P4_1	situacion_laboral
	P4_2	monto_ingreso_trabajo
	P4_2_1	frecuencia_ingreso_trabajo
	P4_8_1	ingreso_jubilacion_pension
	P4_8_7	ingreso_otro_medio
	P4_9_1	monto_jubilacion_pension
	P4_9_7	monto_ingreso_otro
	P4_11	autonomia_uso_dinero
	P4_13_1	titular_inmuebles
	P4_13_2	titular_automoviles
	P4_13_3	titular_ahorros
	P4_13_4	titular_vivienda
	P4_13_5	titular_locales
	P4_13_6	titular_puestos_fijos
	P4_13_7	titular_otra_vivienda

Tabla 2.7: Mapeo y normalización de identificadores: Módulo TB_SEC_IV.

La normalización de los identificadores del módulo de autonomía y toma de decisiones (TB_SEC_XV) se presenta en la Tabla 2.8:

Módulo	Código INEGI	Nombre actualizado
TB_SEC_XV	P15_1AB_1	decision_trabajar_estudiar
	P15_1AB_3	decision_uso_dinero_propio
	P15_1AB_4	decision_compras_personales
	P15_1AB_7	decision_gasto_economia_hogar
	P15_1AB_8	decision_dinero_pareja
	P15_2AB_1	actitud_pareja_trabajar_estudiar
	P15_2AB_3	actitud_pareja_uso_dinero_propio
	P15_2AB_4	actitud_pareja_compras_personales
	P15_2AB_7	actitud_pareja_gasto_economia
	P15_2AB_8	actitud_pareja_uso_dinero_propio_el
	P15_3AB_1	arreglos_trabajar_remuneracion
	P15_3AB_2	arreglos_ir_compras
	P15_3AB_4	arreglos_compras_personales
	P15_1C_1	decision_estudiar_C
	P15_1C_2	decision_trabajar_C
	P15_1C_6	decision_uso_dinero_propio_C
	P15_1C_7	decision_compras_personales_C
	P15_2C_1	arreglos_estudiar_C
	P15_2C_2	arreglos_trabajar_C
	P15_2C_6	arreglos_uso_dinero_propio_C
	P15_2C_7	arreglos_compras_personales_C

Tabla 2.8: Mapeo y normalización de identificadores: Módulo TB_SEC_XV.

Finalmente, el renombrado de la variable objetivo proveniente del módulo TB_VD se muestra en la Tabla 2.9:

Módulo	Código INEGI	Nombre actualizado
TB_VD	VTOT_12M	target_violencia

Tabla 2.9: Mapeo y normalización de identificadores: Módulo TB_VD.

2.2.2 Recodificación de variables

Con el objetivo de garantizar la consistencia analítica, se ejecutó una fase de homogeneización en el conjunto de datos. Las variables de naturaleza binaria fueron transformadas a un formato numérico discreto mediante un mapeo booleano. Este procedimiento consistió en asignar de manera estandarizada el valor numérico de 1 para la presencia del atributo o respuesta afirmativa (“Sí”) y el valor 0 para la ausencia o respuesta negativa (“No”).

Módulo	Variable
TSDem	asiste_escuela
	condicion_actividad
TB_SEC_IV	situacion_laboral
	ingreso_jubilacion_pension
	ingreso_otro_medio
	autonomia_uso_dinero
TB_VD	target_violencia

Tabla 2.10: Identificadores de variables con codificación binaria.

De manera complementaria, ciertas variables sociodemográficas que originalmente presentaban múltiples categorías dispersas fueron sometidas a un proceso de recodificación. Esta transformación se ejecutó mediante una asignación numérica con el propósito de simplificar la granularidad de la encuesta y establecer una estructura explícita en el análisis; un ejemplo de ello fue la construcción de la variable “estatus_conyugal_bin” a partir del indicador original “estado_conyugal”, logrando así representar de mejor manera su nueva naturaleza binaria.

Variable	Rango	Descripción
NIV	0	Ninguno o preescolar
	1	Primaria
	2	Secundaria
	3	Preparatoria o bachillerato
	4	Estudios técnicos, comerciales o normales
	5	Licenciatura o profesional (incluye normal)
	6	Posgrado (Especialidad, maestría o doctorado)
DOMINIO	1	R: Rural
	2	C: Complemento urbano
	3	U: Urbano
estatus_conyugal_bin	0	Sin pareja (Separada, divorciada, viuda o soltera)
	1	Con pareja (Casada o en unión libre)
posicion_ocupacion	0	Persona No Económicamente Activa
	1	Trabajadora sin pago
	2	Jornalera o peona
	3	Obrera
	4	Empleada
	5	Trabajadora por su cuenta (no contrata trabajadores)
	6	Patrona

Tabla 2.11: Identificadores de variables categóricas recodificadas.

Para las variables de autonomía económica y toma de decisiones, se diseñaron escalas ordinales de tres niveles (0-2), donde el valor 0 representa autonomía limitada o nula, el 1 autonomía compartida o negociada, y el 2 autonomía plena o independiente. En el caso de las variables sobre posesión de bienes, los identificadores originales se renombraron bajo el prefijo "autonomia_XX" (en sustitución de "titular_XX") con la finalidad de reflejar con mayor precisión la dimensión de agencia que se pretende medir.

- *autonomia_inmuebles*
- *autonomia_automoviles*
- *autonomia_ahorros*
- *autonomia_vivienda*
- *autonomia_locales*
- *autonomia_puestos_fijos*
- *autonomia_otra_vivienda*
- *decision_trabajar_estudiar*
- *decision_uso_dinero_propio*
- *decision_compras_personales*
- *decision_gasto_economia_hogar*

- *decision_dinero_pareja*
- *actitud_pareja_trabajar_estudiar*
- *actitud_pareja_uso_dinero_propio*
- *actitud_pareja_compras_personales*
- *actitud_pareja_gasto_economia*
- *actitud_pareja_uso_dinero_el*
- *arreglos_trabajar_remuneracion*
- *arreglos_ir_compras*
- *arreglos_compras_ella*
- *decision_estudiar_C*
- *decision_trabajar_C*
- *decision_uso_dinero_propio_C*
- *decision_compras_personales_C*
- *arreglos_estudiar_C*
- *arreglos_trabajar_C*
- *arreglos_uso_dinero_propio_C*
- *arreglos_compras_personales_C*

2.2.3 *Ingeniería de características e indicadores sintéticos*

Tratamiento de ingresos: Se normalizaron los montos de ingresos a una frecuencia mensual unificada. Para mitigar el sesgo por registros donde no hubo una respuesta definida o un posible error de captura (999,998 - No sabe y 999,999 - error), estos valores se sustituyeron por NaN, para posteriormente ser imputados mediante la mediana de la población económicamente activa. De la misma manera, se realizó una imputación por o para la población no económicamente activa. Finalmente, se consolidó la variable “*ingreso_mensual_total*” sumando todas las fuentes que representan un ingreso mensual para la entrevistada, ya sea que deriven de un trabajo, pensión u otros medios.

Con la finalidad de reducir la dimensionalidad y sintetizar la información de las diferentes dimensiones sobre autonomía económica analizadas en este

trabajo, se construyeron variables compuestas e indicadores sintéticos mediante técnicas de ingeniería de características:

Autonomía laboral y patrimonial: Se consolidaron las variables “economicamente_activa” y “nivel_max_autonomia_patrimonial”, integrando variables sobre condición y verificación de actividad, así como autonomía de bienes.

Dimensiones de autonomía: Se unieron variables equivalentes de los cuestionarios de secciones de pareja (A y B) y familia (C) mediante la agregación por valor máximo y promedio, resultando en seis indicadores clave. A través del valor máximo se obtuvieron: “ind_dec_desarrollo” (decisiones sobre desarrollo personal), “ind_dec_recursos_individuales” (decisiones sobre finanzas personales), “ind_dec_recursos_hogar” (decisiones en gestión del hogar), “ind_arr_desarrollo” (arreglos sobre desarrollo personal) e “ind_arr_recursos” (arreglos sobre consumo). Por otro lado, mediante agregación por promedio, se construyó el indicador “ind_actitud_pareja” para representar de forma sintética las actitudes de la pareja.

2.2.4 Tratamiento de datos faltantes (imputación)

Se optó por dos alternativas diferentes para el tratamiento de valores nulos en los indicadores creados:

Indicador de “Actitud de pareja”: Debido a la presencia de 1,162 valores faltantes (alrededor del 36 % de los datos) se aplicó una estrategia de Imputación por Indicador de Ausencia (Missing Indicator Method). Esta técnica consistió en:

- Crear la variable binaria “actitud_pareja_es_nulo” para capturar explícitamente la falta de respuesta.
- Imputar los valores faltantes en la variable original utilizando la mediana de la muestra.

Se decidió utilizar este enfoque para evitar el sesgo de asignar arbitrariamente un valor de autonomía nula (0) o autonomía total (2), permitiendo así que el modelo identifique si la omisión de respuesta constituye un factor de riesgo.

Resto de los indicadores: Los indicadores restantes no presentaron un porcentaje elevado de valores faltantes. Para estas variables se optó por una imputación con valor cero bajo el supuesto de que si una mujer no reportó información en ninguna de las preguntas del bloque, experimenta una ausencia de libertad o autonomía nula en esa dimensión específica. Este criterio permite conservar los registros sin sobreestimar los niveles de agencia en la muestra.

2.2.5 Integración y depuración final

Una vez concluida la limpieza particular de cada módulo, se ejecutó la consolidación de las tablas mediante una unión interna (inner join) utilizando la clave única "ID_PER". Posteriormente, se llevó a cabo un filtrado definitivo que consistió en:

- Eliminación de la variable auxiliar "ID_PER".
- Eliminación de registros con faltantes en la variable objetivo (total de 13 registros eliminados).
- Conversión de tipos de datos a formato entero (int) para todas las variables binarias y categóricas.

El universo inicial del estudio para el estado de Baja California constó de 4,707 registros demográficos generales y 3,226 registros correspondientes a los módulos vinculados con la autonomía económica de las mujeres. Tras aplicar las técnicas de preprocesamiento, ingeniería de características y la exclusión de los 13 casos sin respuesta en la variable objetivo, la muestra final se delimitó a 3,213 observaciones. Este conjunto de datos consolidado cuenta con información íntegra y validada en las dimensiones sociodemográfica, económica, de toma de decisiones y violencia, resultando óptimo para el entrenamiento de los modelos predictivos.

2.3 Análisis exploratorio

El análisis exploratorio se centró en caracterizar el perfil sociodemográfico de la muestra, así como en examinar las dimensiones de autonomía económica de las mujeres en Baja California. Además, se buscó identificar patrones de correlación entre estos indicadores sintéticos y la prevalencia de violencia reportada en los últimos 12 meses, con el fin de establecer una base comparativa que permita validar e interpretar los resultados obtenidos posteriormente mediante el modelado predictivo.

2.3.1 Perfil sociodemográfico

Los análisis presentados en este apartado permiten caracterizar el perfil sociodemográfico de la muestra en el estudio, proporcionando una visión integral de variables clave como la edad, el nivel educativo y el entorno geográfico de las mujeres en Baja California. Estos factores constituyen el contexto estructural antes de evaluar sus dimensiones de autonomía.

La distribución de edades de las mujeres en Baja California (Figura 2.1) abarca un espectro amplio que comprende desde los 15 hasta superar los 90

años. El comportamiento de la variable exhibe un sesgo positivo, evidenciando una mayor densidad poblacional en etapas de vida joven y adulta, con una concentración predominante entre los 25 y 45 años. Esta composición demográfica resulta fundamental para el trabajo, ya que sitúa el análisis prioritariamente en los periodos de mayor actividad económica, agencia laboral y participación en la toma de decisiones al interior del hogar.

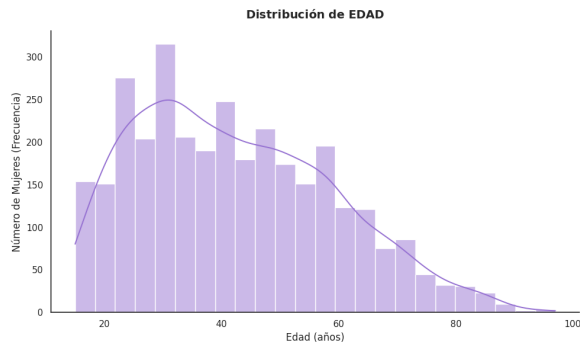


Figura 2.1: Distribución por edades de las mujeres en la muestra.

La relación entre el nivel educativo y la edad de las mujeres en la muestra permite detallar su distribución e interacciones (Figura 2.2). Se observa que la ausencia de estudios (nivel 0) es mínima en el grupo de mujeres jóvenes, concentrándose principalmente en el grupo de 65 años o más. Este patrón de brecha generacional es persistente: a menor edad, se evidencia un incremento en el nivel de estudios alcanzado. Mientras que las mujeres con estudios técnicos, comerciales o normales sin licenciatura son una minoría con mayor presencia en edades avanzadas, la formación profesional (licenciatura) alcanza su máxima representatividad en los rangos de 15 a 44 años. Asimismo, el acceso a posgrados sigue siendo limitado a través de todos los grupos de edad. Estos hallazgos refuerzan la caracterización del segmento predominante del estudio (24 a 45 años) como un grupo que no solo se encuentra en su etapa de mayor actividad económica, sino que posee una sólida preparación académica, factor que resulta determinante al analizar sus capacidades de autonomía y toma de decisiones.

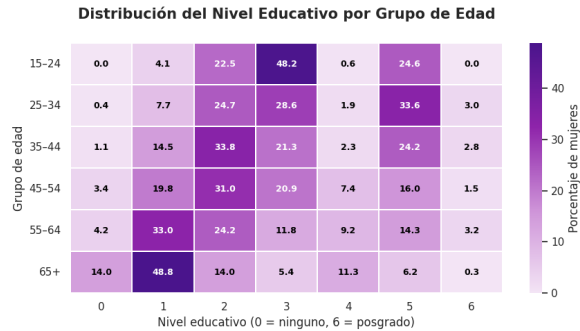


Figura 2.2: Distribución del nivel educativo alcanzado según grupo de edad.

Originalmente, la variable de estado conyugal capturaba 6 categorías en la ENDIREH. Sin embargo, se realizó una recodificación binaria en “presencia” o “ausencia” de vínculo de pareja para optimizar el rendimiento del modelo al reducir la dispersión en categorías con baja representatividad (como viudez o divorcio), y concentrar el análisis en la convivencia actual, factor crítico para el estudio de violencia en la pareja.

Esta evolución estructural permite contrastar sus cambios (Figura 2.3). La gráfica de la izquierda refleja la distribución original en Baja California antes de la integración y el filtrado de casos, mientras que la derecha muestra la variable ya recodificada. En ambas se observa que la mayoría de las mujeres se encuentran en una relación de pareja, hallazgo determinante para contextualizar el riesgo y la vulnerabilidad en el ámbito doméstico.

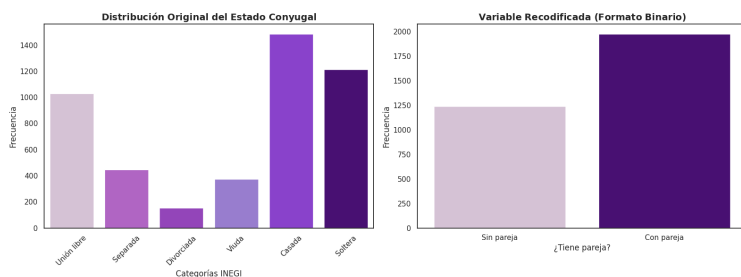


Figura 2.3: Distribución original del estado conyugal y su correspondiente recodificación binaria.

La variable de Dominio permite clasificar el entorno de decisión de las mujeres de Baja California en tres niveles: Rural, Complemento Urbano y Urbano. Como se aprecia en la caracterización geográfica de la muestra (Figura 2.4), se presenta una fuerte concentración en el ámbito Urbano, lo cual coincide con la alta tasa de urbanización de la entidad. No obstante, el incluir las categorías de Rural y Complemento Urbano es fundamental para este estudio, ya que identifica a mujeres que viven en áreas que a menudo presentan problemas particulares en el acceso a servicios y a redes de apoyo, factores que pueden influir tanto en la capacidad de ejercer la autonomía económica como en la exposición a situaciones de violencia.

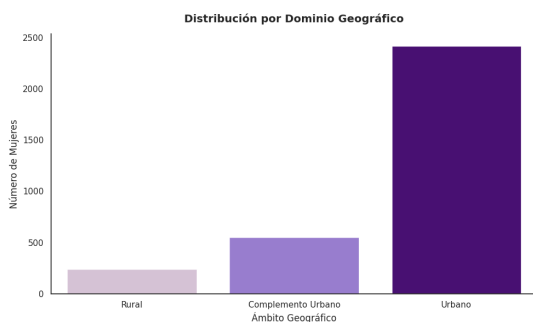


Figura 2.4: Distribución de la muestra por tipo de dominio geográfico.

2.3.2 Caracterización de la autonomía económica

La distribución de la muestra entre mujeres económicamente activas y no activas permite caracterizar su participación (Figura 2.5). Esta variable es el resultado de un proceso de ingeniería de datos que integra la condición de actividad y la verificación de tareas remuneradas, con el objetivo de mitigar la omisión de actividades económicas no convencionales. Al refinar este criterio, se garantiza una clasificación más precisa de la participación económica, evitando que labores informales sean excluidas del análisis de autonomía.

Los resultados muestran una distribución equilibrada, con una proporción ligeramente superior de mujeres económicamente activas, aunque la diferencia respecto al grupo no activo es mínima. Esta equivalencia permite analizar y contrastar la prevalencia de violencia en dos contextos financieramente distintos, pero igualmente representados: el de las mujeres que generan ingresos propios y el de aquellas cuya subsistencia depende de arreglos económicos externos.

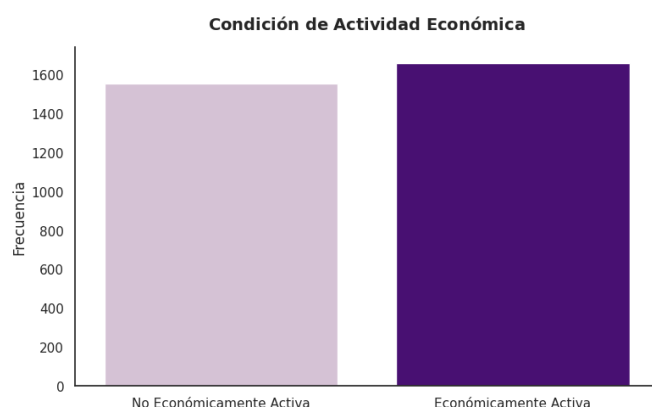


Figura 2.5: Distribución de la muestra según condición de actividad económica.

La composición de la muestra según la posición laboral de las mujeres se encuentra estructurada mediante una jerarquía ordinal diseñada para este estudio (Figura 2.6). A diferencia de la codificación original del INEGI, se optó por un mapeo que refleja niveles crecientes de autonomía en el entorno de trabajo, integrando además la categoría “o” para representar a las mujeres No Económicamente Activas (PNEA).

Como se observa, la categoría con mayor frecuencia es la PNEA, resultado que mantiene consistencia con la distribución de actividad económica presentada anteriormente. Dentro del grupo de mujeres con participación laboral activa, destaca la posición “empleada” como la más representativa, seguida por la “trabajadora por cuenta propia”. Esta caracterización permite contrastar si el grado de autoridad y control sobre los recursos laborales actúa como un factor diferenciador en la exposición a dinámicas de violencia.

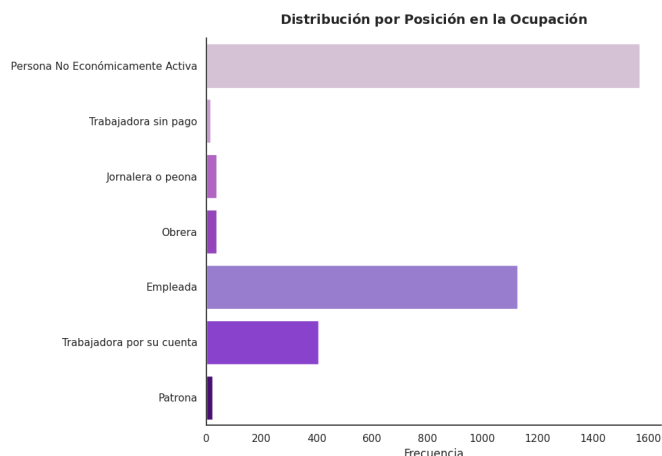


Figura 2.6: Distribución de la muestra según posición ocupacional.

La distribución de los ingresos mensuales totales de la muestra integra percepciones por concepto de trabajo remunerado, pensiones y otros apoyos económicos (Figura 2.7). Como se detalló en el apartado de preprocesamiento, se realizó un proceso de limpieza de datos en el que las respuestas de desconocimiento (999,998) y los posibles errores de captura (999,999) fueron sustituidos mediante la imputación por la mediana, con el fin de evitar distorsiones en la tendencia central.

En el panel izquierdo (A), destaca una frecuencia elevada cercana a los 4,000 registros, cifra que tiene consistencia con el volumen de mujeres clasificadas como económicamente no activas; sin embargo, es importante notar que un segmento de este grupo reporta ingresos por pensión o por otros medios. Por su parte, el panel derecho (B) excluye los valores de cero para permitir una visualización más clara de los ingresos. Aunque la mayoría de las observaciones se sitúan por debajo de los \$50,000 pesos, la presencia de valores atípicos superiores explica la asimetría positiva, fenómeno característico del comportamiento de los ingresos en la región.

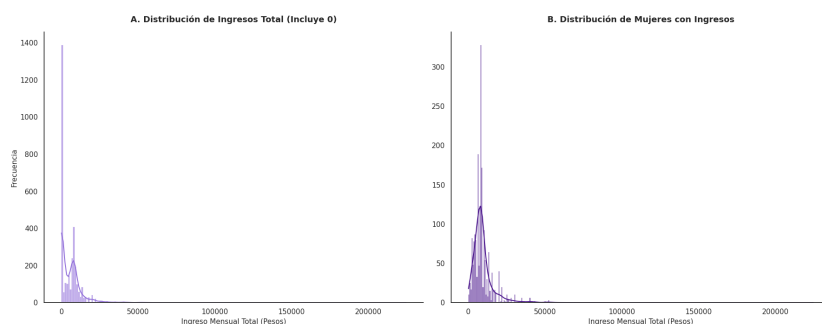


Figura 2.7: Distribución general de los ingresos mensuales totales (A) y detalle excluyendo valores en cero (B).

La autonomía de las mujeres de la muestra respecto al uso del dinero permite ilustrar su capacidad de agencia en el hogar (Figura 2.8). Si bien la mayoría manifiesta poseer poder de decisión en este ámbito, persiste una proporción considerable que carece de ello. El contraste entre quienes generan ingresos propios y quienes efectivamente pueden decidir sobre su uso —analizado en la sección previa—, permitirá identificar en etapas posteriores del modelo si la autonomía actúa como factor que ayude a prevenir dinámicas de abuso.

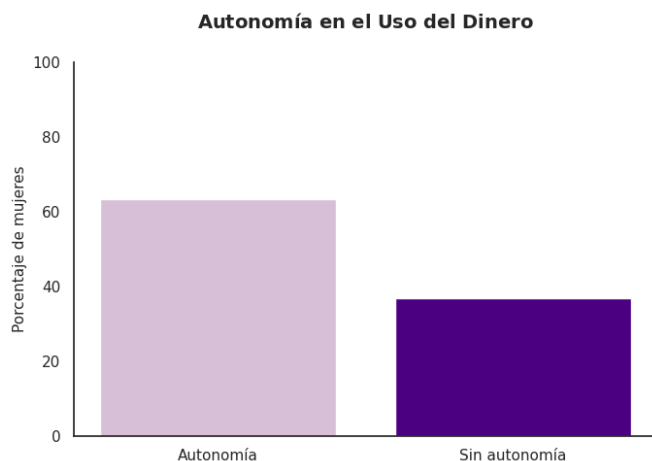


Figura 2.8: Distribución de la muestra según la capacidad de decisión sobre el uso del dinero.

La comparación de las dimensiones sintéticas de autonomía construidas durante la fase de ingeniería de características permite analizar sus interacciones mediante un perfil radial (Figura 2.9). Estos indicadores integran las respuestas de mujeres con pareja (actual o anterior) y, en casos de ausencia de pareja, las dinámicas reportadas en el entorno familiar. El perfil se compone de cinco dimensiones: Toma de decisiones de desarrollo, en recursos personales y recursos del hogar, así como arreglos de recursos y de desarrollo.

Como se observa en este perfil, la dimensión con mayor nivel de autonomía promedio es la Toma de decisiones sobre recursos personales (aprox. 1.75), seguida por la Decisión de desarrollo (aprox. 1.5). Estos valores sugieren que una proporción significativa de la muestra ejerce un poder de decisión autónomo o mayoritariamente propio sobre sus ingresos y sus actividades laborales o académicas.

En cuanto a la dimensión de Arreglos de consumo (aprox. 1.25), el valor se sitúa por encima del umbral medio, lo que indica una tendencia a ejercer el gasto sin necesidad de supervisión constante. No obstante, este puntaje también revela la persistencia de un segmento de mujeres subordinadas a la consulta o la solicitud de permiso para realizar compras, o incluso, de ser acompañadas como medida de control.

Finalmente, las dimensiones con menores niveles de autonomía son la Decisión en gestión del hogar y los Arreglos de desarrollo, ambos con promedios cercanos a 1, siendo este último el indicador con menor valor de todo el perfil. Es relevante destacar que, si bien las mujeres reportan decidir individualmente sobre su formación educativa o empleo (Decisión de desarrollo), el bajo puntaje en Arreglos de Desarrollo revela una autonomía condicionada: en la práctica cotidiana, estas acciones suelen estar restringidas a la solicitud de permiso, la vigilancia o el acompañamiento por parte de la pareja o familiares, evidenciando mecanismos de control que limitan la ejecución real de su autonomía.

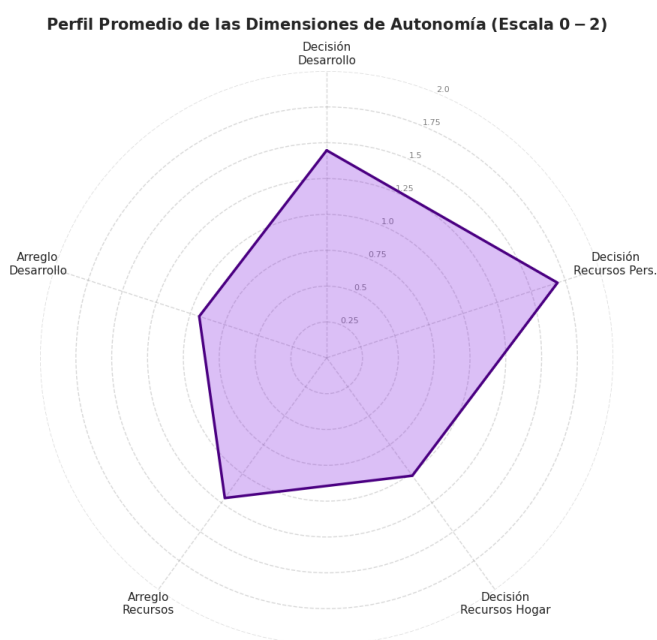


Figura 2.9: Perfil promedio de las dimensiones sintéticas de autonomía.

La distribución del grado de autonomía patrimonial según la titularidad de los bienes y activos permite analizar la estructura de propiedad de la muestra (Tabla 2.12). Los datos revelan que la mayoría de las mujeres presentan una autonomía nula en casi todas las categorías analizadas, especialmente en activos de alto valor como inmuebles y vehículos. Este hallazgo es crítico, ya que la carencia de propiedad constituye un factor de vulnerabilidad que limita la autonomía económica. Esta situación puede atribuirse a dos escenarios: que la titularidad recaiga exclusivamente en la pareja o familiares, o que el hogar en general carezca de bienes.

No obstante, al contrastar los niveles de propiedad, surge un descubrimiento relevante para el estudio: entre las mujeres que sí poseen activos, la autonomía total es más frecuente que la autonomía compartida. Esto indica una polarización en la muestra: existe un grupo mayoritario que no posee activos, pero aquellas que logran acceder a la propiedad tienden a hacerlo de forma individual y no conjunta.

Activo patrimonial	Autonomía nula	Autonomía compartida	Autonomía total
Inmuebles	55.26 %	11.84 %	32.89 %
Automoviles	56.06 %	16.59 %	27.35 %
Ahorros	24.44 %	32.92 %	42.64 %
Vivienda	49.79 %	15.36 %	34.85 %
Locales	64.41 %	11.86	23.73 %
Puestos fijos	53.85 %	19.23 %	26.92 %
Otra vivienda	50.50 %	8.91 %	40.59 %

Tabla 2.12: Grado de autonomía patrimonial según la titularidad de bienes y activos.

La distribución del indicador sintético sobre las actitudes de la pareja hacia la participación laboral o académica de la mujer y su gestión del gasto permite caracterizar este entorno (Figura 2.10). Para el tratamiento de esta variable, los registros faltantes se sustituyeron mediante imputación por la mediana y se generó una variable indicadora (dummy) para identificarlos, dado que la ausencia de respuesta suele vincularse a la falta de pareja actual.

La tendencia muestra una marcada asimetría negativa, donde la mayoría reporta una actitud favorable de su pareja hacia su desarrollo y autonomía financiera. No obstante, la concentración extrema en el puntaje máximo sugiere una posible normalización de conductas, considerando que, en contextos de control sutil, las actitudes restrictivas pueden no ser percibidas o reportadas como negativas en primera instancia.

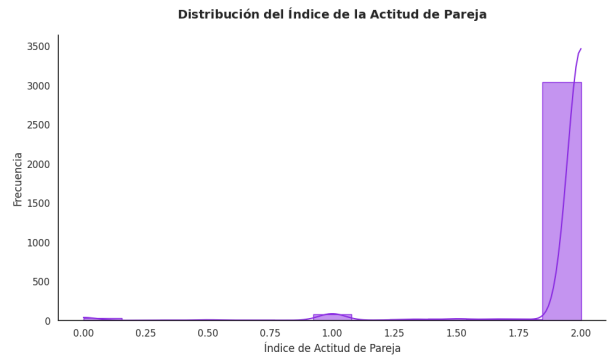


Figura 2.10: Distribución de la muestra según las actitudes de la pareja hacia la autonomía económica de la mujer.

2.3.3 *Análisis comparativo de los factores sociodemográficos y de autonomía frente a la violencia*

A partir de la caracterización del perfil sociodemográfico de las mujeres y de las dimensiones que integran su autonomía económica, esta sección analiza la relación entre dichas características y la prevalencia de violencia ocurrida en los últimos 12 meses. El objetivo es comprender el contexto en el que se desenvuelven las mujeres e identificar relaciones y posibles factores de riesgo o protección que permitan sustentar el modelo predictivo posterior.

La distribución de la prevalencia de violencia contra las mujeres en Baja California durante el último año permite dimensionar la magnitud de esta problemática (Figura 2.11). Se observa que el 35.7 % de la muestra reportó haber experimentado al menos una situación de violencia, frente a un 64.3 % que no la reportó. Si bien la mayoría de las mujeres no se encuentran en esta situación, la cifra del 35.7 % resulta alarmante ya que indica que aproximadamente una de cada tres mujeres en la entidad fue víctima de violencia en un periodo de tan solo doce meses. Esta magnitud enfatiza la urgencia de analizar los factores determinantes de autonomía económica que podrían estar influyendo en dicho fenómeno.

Prevalencia de Violencia (Últimos 12 Meses)

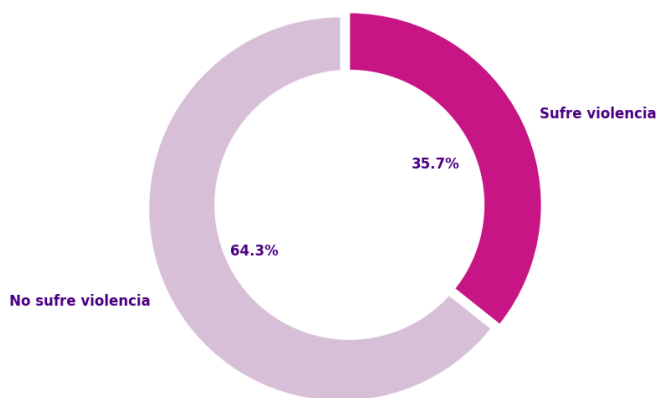


Figura 2.11: Prevalencia de violencia en los últimos 12 meses.

La distribución de la edad de las mujeres en relación con la prevalencia de violencia permite contrastar el comportamiento de esta variable (Figura 2.12). Los datos revelan una tendencia clara: la violencia de género se concentra en mujeres más jóvenes, particularmente en etapas de mayor actividad productiva

y reproductiva. La mediana de edad para quienes reportan violencia se sitúa cerca de los 34 años, notablemente inferior a la mediana de alrededor de 44 años observada en el grupo que no la reporta. Asimismo, el rango intercuartílico refuerza este hallazgo, con el 75 % de las mujeres violentadas ubicándose por debajo de los 47 años. No obstante, la presencia de valores atípicos en edades avanzadas (superiores a los 80 años) es un descubrimiento crítico que evidencia que, si bien la juventud es un factor de mayor vulnerabilidad, el riesgo de sufrir violencia persiste a lo largo de toda la vida.

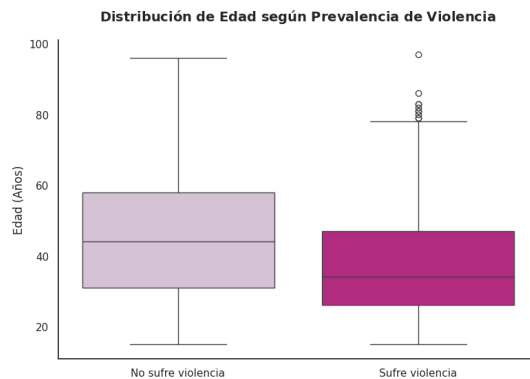


Figura 2.12: Distribución de la muestra por edad según la prevalencia de violencia.

La distribución de los ingresos mensuales en relación con la prevalencia de violencia de género permite contrastar el comportamiento económico entre ambos grupos (Figura 2.13). Si bien las distribuciones generales muestran una estructura similar, se identifica que el grupo de mujeres que experimenta violencia presenta ingresos ligeramente superiores. Esta diferencia se hace más evidente al analizar la mediana, la cual se sitúa por encima de la del grupo sin violencia. Asimismo, destaca la presencia de valores atípicos significativos en ambos grupos, resaltando un caso extremo en el segmento de mujeres que sufren violencia con un ingreso superior a los 200,000 MXN. Estos hallazgos refuerzan la idea de que la violencia no es exclusiva de sectores con bajos recursos económicos y sugieren que el aumento en la percepción de ingresos podría estar asociado a cambios en las dinámicas de poder dentro del hogar.

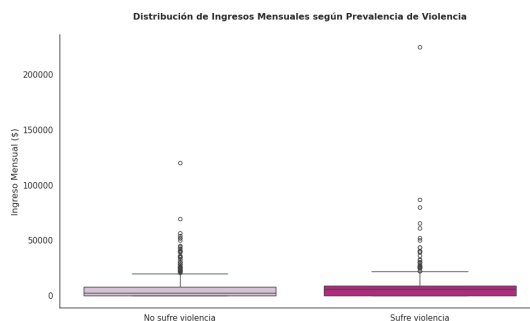


Figura 2.13: Distribución de los ingresos mensuales totales según la prevalencia de violencia.

La intersección entre la edad, el ingreso mensual y la prevalencia de violencia permite contrastar la distribución de estas variables en la muestra (Figura 2.14). Se observa una alta densidad de población con ingresos inferiores a los \$25,000 MXN; no obstante, destaca un segmento de mujeres entre los 20 y 40 años que concentra los niveles de ingreso más elevados. Como se identificó en etapas previas del estudio, la incidencia de violencia es notablemente mayor en rangos de edad jóvenes. Un hallazgo crítico en esta visualización es el valor atípico más extremo de la muestra: una mujer cercana a los 30 años con el ingreso más alto reportado (superior a los \$200,000 MXN), quien declara haber sufrido violencia en los últimos 12 meses. Este fenómeno refuerza la idea de que el riesgo de violencia no disminuye necesariamente ante una mayor autonomía económica, sino que se manifiesta de manera persistente incluso en los niveles de ingreso más altos dentro de los grupos de edad más jóvenes.

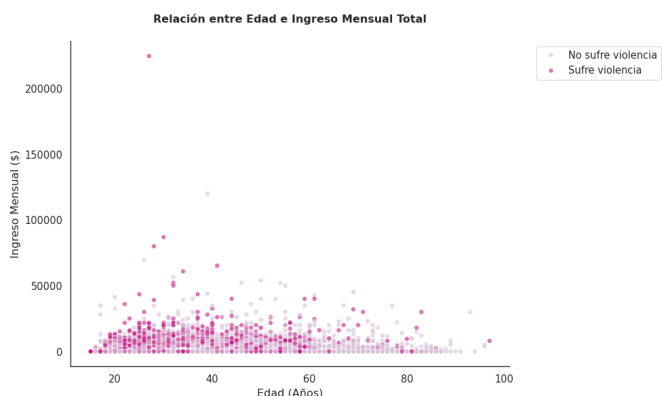


Figura 2.14: Relación entre edad e ingreso mensual total según la prevalencia de violencia.

La relación entre la educación académica y la prevalencia de violencia revela un comportamiento contraintuitivo en la muestra (Figura 2.15). Se observa que los niveles educativos bajos reportan las menores incidencias: el nivel 0 (sin estudios) presenta un 18 %, seguido del nivel 4 (técnico/comercial/normal) con un 24 % y el nivel 1 (primaria) con un 29 %. En contraste, a partir de la educación secundaria y hasta el posgrado (niveles 2, 3, 5 y 6), la prevalencia se eleva significativamente, hasta alrededor de un 40 %. Estos hallazgos sugieren que un mayor grado de estudios no constituye necesariamente un factor reductor de riesgo, sino por el contrario, el incremento en la autonomía potencial de la mujer puede interpretarse como un desafío a las jerarquías de género tradicionales. Bajo esta lógica, el éxito de la mujer puede percibirse como una amenaza hacia el estatus del hombre, provocando el riesgo de sufrir conductas agresivas como mecanismo de control.

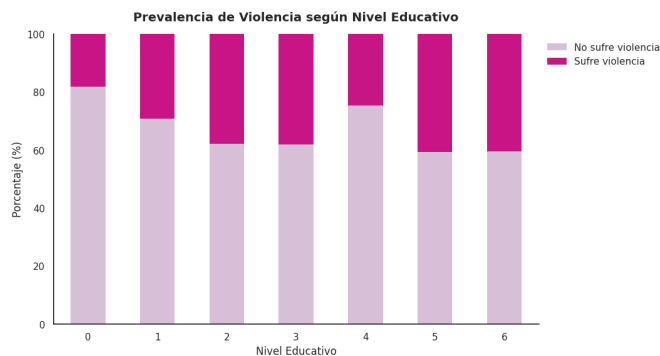


Figura 2.15: Prevalencia de violencia según nivel educativo.

La comparación de la prevalencia de violencia entre mujeres económicamente activas y no activas permite contrastar el comportamiento de esta problemática según su inserción laboral (Figura 2.16). Los datos revelan una tendencia significativa: las mujeres que participan en el mercado laboral reportan una mayor incidencia de violencia en comparación con aquellas que no lo hacen. Este descubrimiento complementa el análisis de nivel educativo, en el que la autonomía financiera podría percibirse como una amenaza al control del agresor, quien puede recurrir a la violencia física o psicológica como mecanismo de represalia, o incluso ejercer violencia económica para intentar contrarrestar la independencia de la mujer.

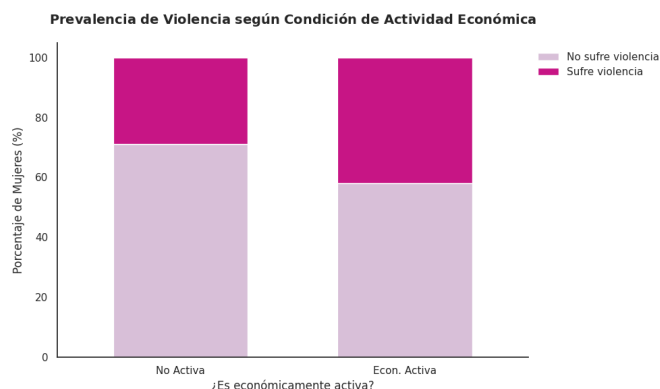


Figura 2.16: Prevalencia de violencia según condición de actividad económica.

La prevalencia de violencia desglosada por la posición jerárquica en el entorno laboral permite analizar las variaciones en las tasas de incidencia (Figura 2.17). Se observa que las incidencias más elevadas, cercanas al 50 %, se concentran en los niveles 3 y 4, correspondientes a obreras y empleadas. A estos niveles les siguen las trabajadoras sin pago y las trabajadoras por cuenta propia (niveles 1 y 5). Por el contrario, las menores prevalencias se encuentran en los extremos: mujeres no económicamente activas y patronas (niveles 0 y 6). Estos hallazgos revelan una dinámica compleja, ya que no hay un patrón interpretable a simple vista que asegure que cierto nivel de posición laboral sea

un factor protector ante la violencia de género.

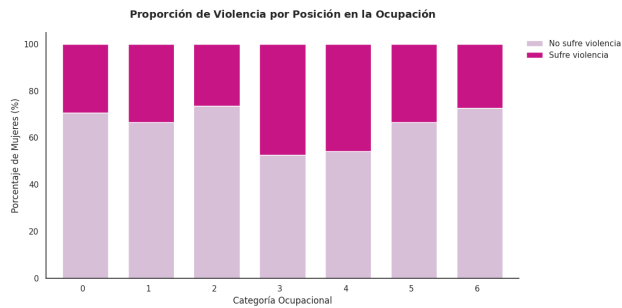


Figura 2.17: Prevalencia de violencia según posición ocupacional.

La distribución de la violencia en función de la autonomía de la mujer para gestionar recursos monetarios permite analizar el impacto directo de esta variable (Figura 2.18). A diferencia de otros indicadores de autonomía, en esta dimensión se observa una distribución prácticamente simétrica: tanto las mujeres con autonomía en el uso del dinero como aquellas que carecen de ella reportan porcentajes de violencia casi idénticos. Esto sugiere que, en la muestra de Baja California, la capacidad individual de gestión de dinero no actúa de manera aislada como un factor determinante para mitigar o incrementar el riesgo.

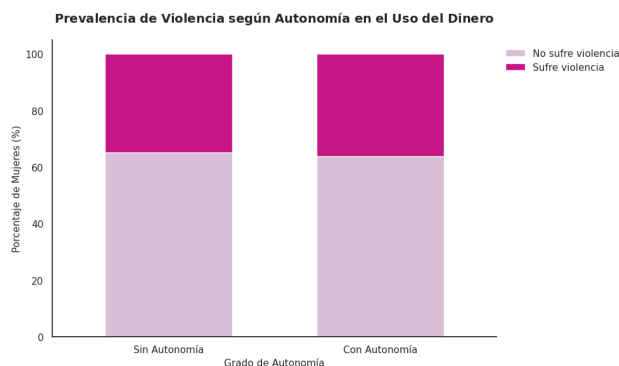


Figura 2.18: Prevalencia de violencia según la capacidad de decisión sobre el uso de dinero.

La relación entre la propiedad de bienes y la prevalencia de violencia permite analizar el impacto de los activos en el entorno familiar (Figura 2.19). Se observa que el mayor índice de violencia se concentra en el nivel 0, correspondiente a mujeres sin autonomía patrimonial. Por el contrario, el nivel 1 (autonomía compartida) presenta la menor incidencia, seguido del nivel 2 (autonomía total). Estos datos sugieren que la propiedad compartida o posesión de activos puede actuar como un factor mitigante de la vulnerabilidad a la violencia al proveer a la mujer de mayores recursos de negociación en el hogar. No obstante, las diferencias porcentuales entre categorías son moderadas, lo que indica que, si bien el patrimonio es un factor de apoyo, no constituye por sí mismo un determinante absoluto para eliminar el riesgo de violencia.

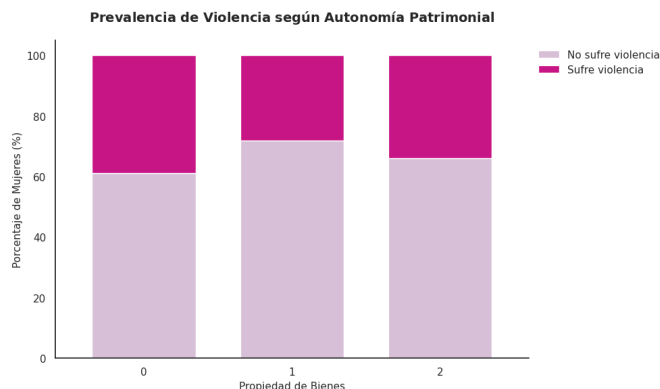


Figura 2.19: Prevalencia de violencia según autonomía patrimonial.

La distribución de densidad del Índice de Actitud de la Pareja permite contrastar el comportamiento de esta variable entre ambos grupos (Figura 2.20). Se observa que, si bien ambos grupos muestran una alta concentración de datos en el nivel 2 (indicando una percepción de actitudes positivas), el grupo de mujeres que reportaron haber vivido violencia presenta una variabilidad significativamente mayor. Esto implica una mayor frecuencia de puntajes bajos en comparación con el grupo sin violencia. Un hallazgo crítico es que en los niveles cercanos a 0, aunque la densidad es baja, la presencia de respuestas pertenece exclusivamente al grupo que sufre violencia. Esto sugiere que, si bien la mayoría de las mujeres perciben actitudes favorables en sus parejas, la manifestación de rasgos negativos en la conducta del compañero sentimental es un indicador altamente asociado con la presencia de dinámicas violentas.

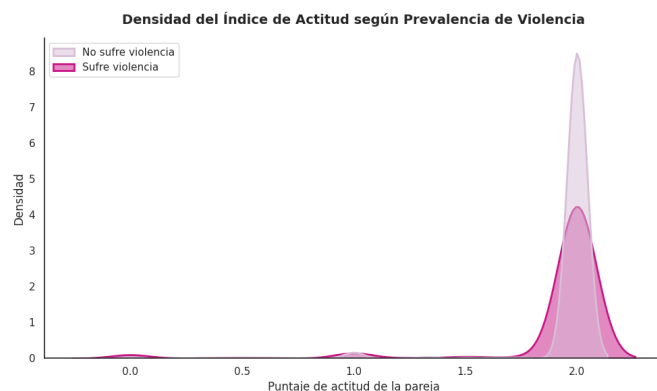


Figura 2.20: Densidad del índice de actitud según la prevalencia de violencia.

El desglose de la prevalencia de violencia en las cinco dimensiones sintéticas de autonomía permite contrastar el comportamiento de cada indicador (Figura 2.21). En todos ellos se confirma que el nivel 0 (autonomía nula) presenta la mayor incidencia, lo que corrobora la relación entre la subordinación en la toma de decisiones y el riesgo de violencia. Sin embargo, surge un descubrimiento diferenciado entre las dimensiones de Decisiones y Arreglos. En las primeras, el nivel 1 (toma de decisiones compartidas) reporta la menor

prevalencia, sugiriendo que los entornos basados en la negociación y el consenso actúan como un factor de protección más eficaz que la autonomía puramente individual (nivel 2). Por el contrario, en las dimensiones de Arreglos, se observa un patrón lineal descendente: el riesgo disminuye conforme la mujer pasa de la necesidad de pedir permiso (nivel 0) o avisar (nivel 1) hacia la total independencia (nivel 2). Esto sugiere que, mientras que en la toma de decisiones el consenso protege, en los arreglos, la libertad plena de acción es el principal reductor del riesgo.

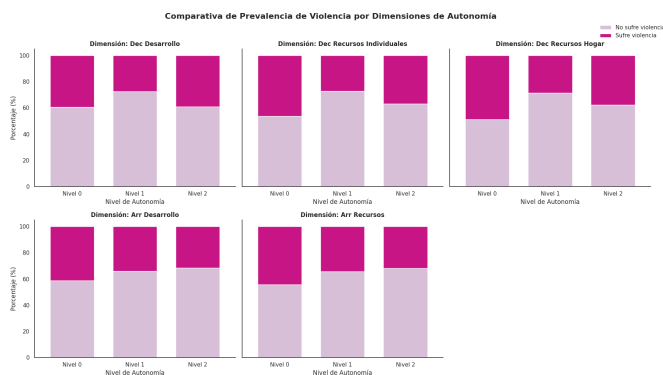


Figura 2.21: Comparativa de prevalencia de violencia por dimensiones de autonomía.

La matriz de correlación calculada para finalizar el análisis exploratorio permite identificar visualmente el grado de asociación entre las variables predictoras, así como su relación con la variable objetivo (Figura 2.22). Para este cálculo se utilizó el Coeficiente de Spearman, el cual evalúa la relación monótona entre dos variables continuas u ordinales. En una relación monótona, las variables tienden a cambiar simultáneamente, aunque no necesariamente a un ritmo constante [22]. Esta elección metodológica responde a dos factores clave: la naturaleza ordinal de los indicadores sintéticos de autonomía y la complejidad característica de las variables sociales analizadas, las cuales no siempre presentan comportamientos lineales.

Al analizar la matriz, se observa que no existen correlaciones lineales fuertes (superiores a 0.5) con la variable objetivo, lo que confirma la naturaleza compleja de la violencia de género. No obstante, el coeficiente de Spearman permite identificar tendencias significativas que se describen a continuación:

1. **Factores de riesgo (correlación positiva):** Las variables con mayor asociación positiva con la violencia son aquellas vinculadas a la situación laboral y el nivel educativo, tales como “situacion_laboral” (0.14), “economicamente_activa” (0.13) e “ingreso_mensual_total” (0.12). Esto sugiere que el acceso a recursos económicos y el desempeño de un rol productivo no actúan como factores protectores por sí solos; por el contrario, en contextos de estructuras tradicionales, pueden exacerbar el riesgo de sufrir violencia como una reacción al cambio en la dinámica de poder.

2. **Factores protectores (correlación negativa):** Las variables que muestran una relación inversa con el riesgo son principalmente los indicadores sintéticos de autonomía y la actitud de la pareja. Destaca la variable EDAD con la correlación negativa más alta (-0.21), confirmando que la juventud es el factor de vulnerabilidad más crítico en la muestra. Asimismo, los índices de autonomía en Arreglos de Recursos (-0.09) y la Actitud de la Pareja (-0.08) refuerzan la idea de que la capacidad de negociar y la calidad del vínculo en las relaciones son elementos que mitigan la probabilidad de victimización.
3. **Multicolinealidad entre variables predictoras:** En cuanto a la relación entre variables independientes, se identificaron correlaciones elevadas entre “economicamente_activa”, “posicion_ocupacion” e “ingreso_mensual_total”. Debido a que estas variables capturan dimensiones similares del fenómeno laboral, se determinó eliminar “situacion_laboral” para prevenir problemas de redundancia y mejorar la eficiencia del modelo predictivo.

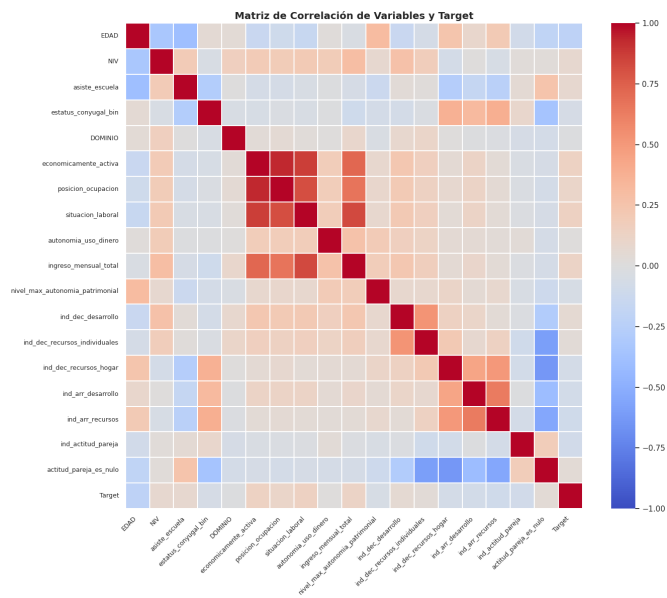


Figura 2.22: Matriz de correlación (Coeficiente de Spearman).

2.4 Descripción y justificación de los modelos

El presente proyecto aborda un problema de clasificación binaria, cuyo objetivo es categorizar la muestra en dos grupos diferenciados: mujeres que experimentan violencia y aquellas que no. Dado que la naturaleza de la variable objetivo es categórica, se seleccionaron algoritmos de aprendizaje supervisado especializados en tareas de clasificación. A continuación, se describen los modelos elegidos y la justificación para su uso en el análisis de este fenómeno social:

2.4.1 Regresión Logística

Orientada a la predicción de resultados categóricos, la regresión logística constituye un modelo lineal que examina la relación entre las variables predictoras y una variable objetivo. Al igual que la regresión lineal, el algoritmo estima la probabilidad de ocurrencia de un evento específico para, posteriormente, clasificar el resultado como 0 o 1 en función de un umbral de decisión preestablecido —generalmente de 0.5— [23].

En la regresión logística se utiliza la función logística para modelar dicha probabilidad [24]:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k}} \quad (2.1)$$

Donde:

- $X = (X_1, X_2, \dots, X_k)$ representa el vector de variables predictoras (los factores sociodemográficos e indicadores de autonomía).
- β_0 representa el intercepto o término constante del modelo.
- $\beta_1, \dots, \beta_k$ representan las pendientes o coeficientes asociados a cada variable predictora, los cuales determinan la dirección y magnitud de la asociación.

Este modelo destaca por su moderada complejidad matemática, facilidad de implementación e interpretabilidad, ya que no solo permite cuantificar la relevancia de un predictor, sino también identificar su dirección de asociación. Asimismo, su eficiencia computacional facilita el procesamiento ágil de grandes volúmenes de datos [25], [26].

Debido a su simplicidad, se seleccionó como el modelo base (baseline) de este trabajo. Su aplicación permitirá determinar si existe una relación lineal entre las variables sociodemográficas y los indicadores de autonomía económica, facilitando la identificación de mujeres en riesgo de sufrir violencia en el estado de Baja California.

2.4.2 *Análisis Discriminante Lineal (LDA)*

Fundamentado en el teorema de Bayes, el *Análisis Discriminante Lineal* (LDA) constituye un método probabilístico y de reducción de dimensionalidad orientado a la clasificación supervisada. Este modelo calcula la probabilidad de que un conjunto de datos pertenezca a una categoría específica mediante la identificación de una combinación lineal de características que maximice la separación interclase. Dicho proceso se realiza proyectando los datos multidimensionales en un espacio de menor dimensión, optimizando de este modo la diferenciación entre los grupos analizados [27].

Este modelo asume que los datos siguen una distribución normal o gaussiana. Su carácter lineal radica en que las funciones discriminantes $\delta_k(x)$ se estructuran como combinaciones lineales de los predictores x . Esta función tiene como propósito computar la frontera de decisión, asignando cada observación a la clase para la cual el score discriminante resulta máximo [24]:

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k) \quad (2.2)$$

Donde:

- x es el vector de características de la observación analizada.
- μ_k representa el vector de medias de las variables predictoras para la clase k (mujeres con o sin prevalencia de violencia).
- Σ^{-1} es la inversa de la matriz de covarianza común, la cual captura la variabilidad compartida entre los predictores.
- π_k es la probabilidad a priori de que una observación pertenezca a la clase k .

El LDA ofrece la ventaja de cumplir un doble propósito: reducir la dimensionalidad del problema mientras actúa simultáneamente como clasificador. A diferencia del *Análisis de Componentes Principales* (PCA), que es un método no supervisado y, por tanto, ajeno a los objetivos de este estudio, el LDA aprovecha las etiquetas de clase para mejorar la separación. Al igual que la regresión logística, este algoritmo destaca por su eficiencia computacional y su alta capacidad de interpretación [28].

Se seleccionó este modelo como una extensión del análisis lineal, permitiendo evaluar la estructura de los datos tras el preprocesamiento de normalización realizado. Dado que el modelo procesa un espacio dimensional de 17 variables predictoras, la capacidad del LDA para proyectarlas en un subespacio menor maximizando la separación interclase resulta fundamental para identificar perfiles de riesgo. Su implementación constituye un punto de contraste técnico que incrementa gradualmente la complejidad del estudio.

2.4.3 Máquinas de Vector de Soporte (SVM)

La Máquina de Vectores de Soporte (SVM) es un algoritmo de aprendizaje supervisado diseñado para clasificar datos mediante la identificación de un hiperplano óptimo en un espacio N-dimensional. Este modelo busca maximizar el margen de separación entre los puntos más cercanos de clases opuestas, conocidos como vectores de soporte. Al maximizar esta distancia, el algoritmo garantiza la selección de la frontera de decisión con mayor capacidad de generalización entre las múltiples opciones posibles [29].

Analíticamente, el hiperplano de decisión se define mediante una combinación lineal de las características de entrada. Para un problema de clasificación binaria con etiquetas $y_i \in \{-1, 1\}$, la función de clasificación predictiva para una nueva observación x se expresa como [24]:

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i y_i K(x, x_i) \quad (2.3)$$

Donde:

- S representa el conjunto de índices de los vectores de soporte, los cuales son los únicos puntos de la muestra que definen la frontera de decisión.
- α_i son los multiplicadores de Lagrange obtenidos mediante la resolución del problema de optimización cuadrática bajo restricciones.
- β_0 es el intercepto o término de sesgo (bias) del hiperplano.
- $K(x, x_i)$ representa la función kernel, la cual calcula la similitud o producto interno entre la observación x y el vector de soporte x_i .

Las SVM son particularmente efectivas en espacios de alta dimensionalidad y poseen la capacidad de capturar relaciones no lineales a través de la implementación de funciones kernel, tales como el Kernel de Función de Base Radial (RBF) o el polinomial, los cuales proyectan los datos implícitamente a un espacio de características de mayor dimensión donde las clases se vuelven linealmente separables. Si bien destacan por su eficiencia en el uso de memoria, su interpretabilidad disminuye a medida que aumenta la complejidad y dimensionalidad del espacio de características [30].

Se seleccionó este modelo como un puente metodológico entre los modelos lineales y los algoritmos basados en árboles. Al combinar la capacidad de identificar fronteras tanto lineales como no lineales a través de sus componentes matemáticos, las Máquinas de Vectores de Soporte (SVM) constituyen un punto de comparación clave para evaluar el desempeño de ambos enfoques en la detección de riesgo de violencia. A las SVM enfocadas en tareas de clasificación se les conoce formalmente como SVC (Support Vector Classifier), término bajo el cual se reportará este algoritmo en las secciones subsecuentes de resultados.

2.4.4 Decision Trees

Un Decision Tree es un algoritmo de aprendizaje supervisado no paramétrico con una estructura jerárquica compuesta por un nodo raíz, ramas, nodos internos y nodos hoja. El proceso de aprendizaje emplea una estrategia de búsqueda codiciosa (greedy) para realizar particiones binarias y recursivas de forma descendente en el espacio de características, deteniendo el procedimiento una vez que los registros han sido asignados a etiquetas de clase específicas [31].

En tareas de clasificación binaria, el criterio matemático óptimo para determinar el punto de división (split) en un nodo se calcula evaluando la pureza de las ramificaciones resultantes. Para ello, el algoritmo maximiza la ganancia de información utilizando la métrica del Índice Gini [24]:

$$G = \sum_{k=1}^K \hat{p}_{mk}(1 - \hat{p}_{mk}) \quad (2.4)$$

Donde $\hat{p}_{mk}$ representa la proporción de observaciones en el nodo m que pertenecen a la clase k . Un índice cercano a cero indica que el nodo es puramente homogéneo, garantizando que la división separa eficazmente a los grupos de la muestra.

Asimismo, otra métrica clásica utilizada para evaluar la impureza de las divisiones es la Entropía, la cual se define como [24]:

$$D = - \sum_{k=1}^K \hat{p}_{mk} \log \hat{p}_{mk} \quad (2.5)$$

Entre sus principales ventajas destacan su interpretabilidad y transparencia, ya que el flujo lógico de decisiones desde la raíz hasta las hojas es visualmente directo. Además, requieren un preprocesamiento mínimo y presentan robustez ante valores atípicos, permitiendo el manejo simultáneo de datos numéricos y categóricos. No obstante, son propensos al sobreajuste (overfitting), por lo que suelen requerir técnicas como la poda (pruning) para eliminar nodos con bajo poder predictivo. Igualmente, pueden presentar sesgos hacia las clases dominantes en conjuntos de datos desbalanceados [32].

Este algoritmo sirve como punto de partida para el análisis basado en árboles. Su uso justifica la transición hacia modelos más robustos que puedan modelar con mayor precisión la naturaleza multidimensional de este fenómeno social.

2.4.5 Bagging Trees

El Bagging (Bootstrap Aggregating) es un método de aprendizaje supervisado por conjuntos que consiste en extraer múltiples muestras aleatorias de un conjunto de entrenamiento mediante reemplazo. Esto implica que un mismo dato puede ser seleccionado en más de una ocasión. En este sentido, se entrenan diversos modelos de manera independiente para que, mediante la combinación de sus predicciones, se obtenga una estimación más robusta [33].

El proceso se estructura en tres etapas fundamentales [24], [33]:

1. **Bootstrapping:** Técnica de remuestreo que genera B diversos subconjuntos de entrenamiento seleccionando puntos de datos al azar y con reemplazo.
2. **Entrenamiento paralelo:** Cada muestra generada se utiliza para entrenar un Decision Tree individual de forma independiente y simultánea.
3. **Agregación:** Fase donde se combinan las predicciones de los B árboles para emitir el diagnóstico final.

Originalmente, la formulación matemática del Bagging se define para tareas de regresión mediante el promedio aritmético de las estimaciones individuales [24]:

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}^{*b}(x) \quad (2.6)$$

Donde:

- $\hat{f}_{bag}(x)$ representa la predicción final agregada del ensamble para la observación x .
- B es el número total de muestras bootstrap y, por consiguiente, el número de árboles entrenados.
- $\hat{f}^{*b}(x)$ corresponde a la predicción numérica emitida por el b -ésimo Decision Tree individual.

El algoritmo de Bagging puede ser extendido a problemas de clasificación de diversas maneras, siendo la más simple e intuitiva la implementación de un sistema de votación mayoritaria. Bajo este enfoque, la predicción final del ensamble corresponderá a la etiqueta de clase que haya sido seleccionada con mayor frecuencia a lo largo de las B predicciones individuales [24].

Entre las ventajas del Bagging destaca su capacidad para mitigar el sobreajuste (overfitting) propio de los Decision Trees individuales. Asimismo, reduce el impacto de valores atípicos y el ruido, disminuyendo la varianza y fortaleciendo la capacidad de generalización del modelo [34].

Se seleccionó este método como una extensión natural de los Decision Trees, con el fin de evaluar cómo la agregación de modelos mejora la estabilidad predictiva y la precisión en la identificación del riesgo de violencia.

2.4.6 Random Forest

El Random Forest es un algoritmo de aprendizaje supervisado que combina los resultados de múltiples Decision Trees para generar una predicción única. Se considera una extensión del método bagging, ya que integra tanto el remuestreo de datos como la aleatoriedad de características para construir un bosque de árboles no correlacionados. A diferencia de un Decision Tree individual o del bagging convencional, que evalúan la totalidad de los predictores en cada división, este método restringe la búsqueda a un subconjunto aleatorio de variables, garantizando una menor correlación y mayor diversidad entre los modelos que integran el ensamble [35].

Esta modificación representa una optimización técnica respecto al bagging, pues aunque ambos construyen múltiples árboles a partir de conjuntos de entrenamiento remuestreados, Random Forest introduce una restricción en la selección de características. Si el espacio total de características está definido por p variables predictoras, el algoritmo selecciona aleatoriamente un número m de variables en cada nodo para competir por la división (split). Típicamente, para tareas de clasificación, dicha selección se define como [24]:

$$m \approx \sqrt{p} \quad (2.7)$$

Donde:

- p representa el número total de variables independientes analizadas en el conjunto de datos (en este estudio, $p = 17$).
- m es la cantidad aproximada de variables permitidas para competir en cada punto de división, lo que fuerza al algoritmo a explorar predictores alternativos y reduce la varianza global del ensamble.

Entre sus principales ventajas destaca su alta precisión, derivada de la agregación de múltiples predicciones. Asimismo, es un modelo robusto ante el ruido y de naturaleza no paramétrica, lo que significa que no requiere supuestos sobre la distribución de los datos. Una característica fundamental es su capacidad para calcular la importancia de las variables predictoras, basándose en la reducción de la impureza que aporta cada variable en las tareas de clasificación. No obstante, este algoritmo presenta una mayor complejidad computacional, un consumo superior de memoria y una pérdida de interpretabilidad en comparación con los modelos lineales [36].

Se seleccionó este algoritmo para cerrar el bloque de modelos de árboles, añadiendo el componente de aleatoriedad al bagging. Al ser el modelo de mayor complejidad técnica en este estudio, se espera que logre capturar las interacciones intrincadas entre los perfiles sociodemográficos de las mujeres, sus indicadores de autonomía económica y el riesgo de experimentar violencia de género.

2.5 Descripción y justificación de las métricas de evaluación

A continuación, se describen las métricas seleccionadas para evaluar el desempeño de los modelos de clasificación:

2.5.1 Matriz de confusión

La matriz de confusión, también conocida como tabla de contingencia, es una herramienta fundamental para la evaluación del desempeño de un modelo de clasificación. Consiste en una tabla de 2×2 que provee un análisis detallado de los resultados de la clasificación binaria. Se compone de los siguientes cuatro elementos [37]:

- **Verdaderos Positivos (TP):** El número de instancias positivas que fueron clasificadas correctamente como positivas por el modelo.
- **Falsos Positivos (FP):** El número de instancias negativas que fueron clasificadas incorrectamente como positivas por el modelo.
- **Verdaderos Negativos (TN):** El número de instancias negativas que fueron clasificadas correctamente como negativas por el modelo.
- **Falsos Negativos (FN):** El número de instancias positivas que fueron clasificadas incorrectamente como negativas por el modelo.

Se representa de la siguiente manera en la Tabla 2.13:

	Positivo predicho	Negativo predicho
Positivo real	Verdadero Positivo (TP)	Falso Negativo (FN)
Negativo real	Falso Positivo (FP)	Verdadero Negativo (TN)

Tabla 2.13: Ejemplo de matriz de confusión.

La matriz de confusión es un instrumento útil para apreciar visualmente de qué manera se distribuyen los resultados de los modelos y así realizar un análisis final más completo.

2.5.2 Accuracy

La exactitud (Accuracy) es la métrica más utilizada para evaluar modelos de clasificación binaria. Esta mide la proporción de predicciones correctas hechas por el modelo sobre el total de las estimaciones realizadas [37]:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.8)$$

Donde los componentes de la ecuación corresponden a los elementos analíticos de la matriz de confusión descritos previamente.

Si bien el accuracy permite realizar un análisis complementario sobre el desempeño global de los algoritmos, no se considera la métrica más adecuada para fundamentar la decisión final. Debido a la naturaleza compleja del fenómeno de estudio y al desbalance de clases, un alto porcentaje de aciertos totales no garantiza que el modelo tenga la capacidad suficiente para identificar correctamente a la clase de interés (mujeres en riesgo de violencia). Por lo tanto, el accuracy se utiliza únicamente como una referencia general.

2.5.3 Precision

La precisión (Precision) es la métrica que mide la proporción de instancias verdaderas positivas entre el total de las observaciones que fueron predichas como positivas por el modelo. Es decir, evalúa la certeza de las clasificaciones positivas emitidas por el algoritmo [37]:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2.9)$$

Donde los componentes de la ecuación corresponden a los elementos analíticos de la matriz de confusión descritos previamente.

El precision es una métrica relevante para este análisis, ya que asegura que las identificaciones de riesgo sean certeras. Un valor alto en esta métrica indica que el modelo es confiable al clasificar una instancia como positiva, lo cual es de suma importancia para mitigar la presencia de falsas alarmas en el diagnóstico.

2.5.4 Recall

La sensibilidad (Recall) es la métrica que mide la proporción de instancias positivas que son identificadas correctamente por el modelo de clasificación de

entre todas las observaciones que son realmente positivas. En otras palabras, evalúa la capacidad intrínseca del algoritmo para detectar de manera efectiva los casos pertenecientes a la clase de interés [37]:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2.10)$$

Donde los componentes de la ecuación corresponden a los elementos analíticos de la matriz de confusión descritos previamente.

El uso de recall en este proyecto es fundamental, ya que indica la capacidad del modelo para identificar correctamente a las mujeres que efectivamente se encuentran en situación de riesgo. En un fenómeno complejo como la violencia de género, se prioriza minimizar los falsos negativos con el propósito de asegurar que la mayor cantidad posible de mujeres que estén en peligro reciban la atención y el respaldo necesario.

2.5.5 F1-Score

El F1-Score es una métrica que combina la precisión y la sensibilidad para proporcionar una evaluación integral de un modelo de clasificación binaria. Matemáticamente, se define como la media armónica de ambas métricas, otorgándoles una importancia equivalente en el cálculo del desempeño global [37]:

$$F1 - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2.11)$$

En este estudio, el F1-Score se establece como la métrica de mayor relevancia para la toma de decisiones, ya que se busca alcanzar un equilibrio óptimo entre precisión y recall. Un valor alto en esta métrica indica que el modelo logra detectar eficazmente a las mujeres en situación de riesgo (clase positiva) sin generar un volumen excesivo de falsas alarmas, garantizando así la confiabilidad de los resultados en la identificación de la violencia de género.

2.5.6 AUC-ROC

Una de las métricas más utilizadas para evaluar el desempeño de los modelos es la curva AUC-ROC (Área Bajo la Curva de la Característica Operativa del Receptor). La curva ROC es una representación gráfica del desempeño de un modelo de clasificación que muestra la relación entre la Tasa de Verdaderos Positivos (TPR) y la Tasa de Falsos Positivos (FPR) a través de distintos umbrales de decisión [37].

La Tasa de Verdaderos Positivos (True Positive Rate), que corresponde analíticamente al concepto de recall, se calcula mediante la siguiente expresión matemático-estadística:

$$TPR = \frac{TP}{TP + FN} \quad (2.12)$$

Por su parte, la Tasa de Falsos Positivos (False Positive Rate) se define formalmente como:

$$FPR = \frac{FP}{FP + TN} \quad (2.13)$$

Donde los componentes de ambas ecuaciones corresponden a los elementos analíticos de la matriz de confusión descritos previamente.

El valor AUC representa el área total bajo dicha curva y oscila en un rango de 0 a 1. Un valor de AUC más elevado indica una mayor capacidad del modelo para distinguir correctamente entre las clases (en este caso, mujeres en riesgo y mujeres que no lo están). Un AUC de 0.5 sugiere que el modelo no tiene poder discriminativo (equivalente al azar), mientras que un valor cercano a 1 refleja un desempeño excepcional en la separación de las categorías.

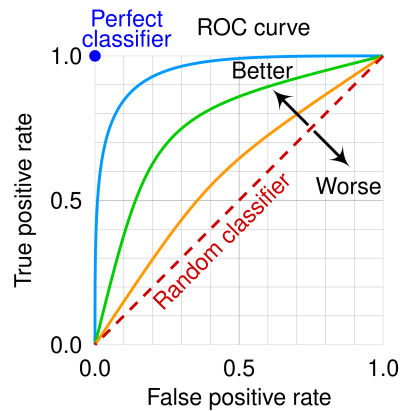


Figura 2.23: Ejemplo de curva ROC [38].

2.6 Descripción de los experimentos

2.6.1 División del conjunto de datos

Para garantizar la capacidad de generalización de los algoritmos y evitar el sobreajuste (*overfitting*), se implementó una estrategia de división de datos en tres subconjuntos independientes, manteniendo la proporcionalidad de las clases mediante un muestreo estratificado:

- **Entrenamiento (60 %):** Muestra de datos utilizada para ajustar el modelo. El modelo observa y aprende de estos datos [39].
- **Validación (20 %):** Muestra de datos utilizada para proporcionar una evaluación imparcial del ajuste de un modelo en el conjunto de entrenamiento mientras se ajustan los hiperparámetros [39].
- **Prueba (20 %):** Muestra de datos utilizada para proporcionar una evaluación imparcial del ajuste del modelo final al conjunto de datos de entrenamiento [39].

2.6.2 Tratamiento de valores atípicos y sesgos en variables continuas

Una vez analizadas las relaciones entre variables y los patrones visuales, se implementaron pasos adicionales para optimizar el conjunto de datos para el modelado predictivo. Con el objetivo de asegurar la estabilidad y convergencia de los algoritmos, se realizó un preprocesamiento de las variables continuas. Específicamente, la variable de ingreso mensual total fue sometida a un proceso de winsorización, un método estadístico que limita el impacto de los valores extremos sin eliminarlos, ajustándolos a los límites del Rango Intercuartílico (IQR) [40]. Posteriormente, se aplicó una transformación logarítmica para reducir la asimetría en dicha variable. Asimismo, se implementó una transformación de raíz cuadrada a la variable de edad para suavizar su distribución. Este tratamiento permite que los modelos capturen la variabilidad real de la muestra, evitando que los valores atípicos distorsionen los coeficientes o las fronteras de decisión.

2.6.3 Escalado de variables continuas

Como proceso final previo a la implementación del modelado predictivo, se realizó el escalado de las variables continuas. Debido a la presencia de valores atípicos, particularmente en la variable de ingreso mensual total, se optó por utilizar un escalado robusto (*RobustScaler*). A diferencia de la estandarización convencional, este método mitiga el impacto de los valores extremos al utilizar

la mediana y el rango intercuartílico (IQR) para el ajuste de la escala [41]. Este paso asegura que las magnitudes de variables como la edad y el ingreso no distorsionen el entrenamiento de los algoritmos sensibles a la escala.

Cabe destacar que el escalado se ajustó exclusivamente a partir de los datos de entrenamiento para ser aplicado posteriormente a los conjuntos de validación o prueba. Este procedimiento evita la fuga de datos (data leakage) al no permitir que información externa influya en el modelo. Asimismo, el escalado se aplicó únicamente a las variables independientes continuas conservando el formato original de las variables categóricas y la variable objetivo, garantizando la integridad del análisis predictivo.

2.6.4 Grid Search y Cross-Validation

Para identificar la configuración óptima de cada algoritmo, se implementó la técnica de Grid Search. Este procedimiento consiste en la evaluación de múltiples combinaciones de hiperparámetros, seleccionando finalmente aquella que maximice el rendimiento del modelo [42]. El espacio de búsqueda se personalizó para cada modelo, considerando los parámetros cruciales que influyen en su capacidad de aprendizaje y regularización, y se ajustó sobre los datos de entrenamiento.

La métrica seleccionada para la evaluación de las combinaciones en el Grid Search fue AUC-ROC. Se optó por esta métrica debido a su capacidad para medir el rendimiento de los modelos de manera independiente del umbral de decisión, el cual fue optimizado en una etapa posterior [43]. Este enfoque asegura que el modelo posea una capacidad discriminatoria robusta antes de ajustar su umbral.

Asimismo, con el fin de garantizar la estabilidad de los resultados y mitigar el riesgo de sobreajuste (overfitting), se integró una estrategia de k-fold Cross-Validation con $k = 5$. Considerando esto, el conjunto de entrenamiento se divide en k subconjuntos, donde el modelo se entrena en $k - 1$ pliegues y se valida en el restante, repitiendo el proceso de forma iterativa. El desempeño final considerado es el promedio de las evaluaciones obtenidas en cada ciclo, asegurando que los parámetros aprendidos no dependan de un subconjunto específico de los datos y permitiendo una valoración más imparcial [44].

2.6.5 Balanceo de pesos

La variable objetivo del presente estudio muestra una distribución de 64.3 % para los casos negativos (ausencia de violencia) y 35.7 % para los casos positivos (presencia de violencia). Aunque esta proporción no es un desbalance extremo, representa un sesgo que podría inclinar las predicciones del modelo

hacia la clase mayoritaria, comprometiendo la sensibilidad en la detección de la clase de interés.

Para mitigar este efecto y asignar una importancia proporcional a la clase minoritaria, se implementó el parámetro “class_weight=‘balanced’” de la librería Scikit-Learn en los algoritmos compatibles (Regresión Logística, SVM, Decision Trees y Bagging Trees y Random Forest). Este ajuste recalcula automáticamente los pesos de las muestras de forma inversamente proporcional a las frecuencias de clase en los datos de entrada [45]. Esta configuración es fundamental para penalizar con mayor severidad los errores cometidos sobre la clase positiva, obligando al modelo a aprender con mayor eficacia los patrones asociados a las mujeres en situación de violencia y optimizando así el F1-Score y el Recall.

2.6.6 Ajuste del umbral

De forma complementaria al Grid Search, se implementó un procedimiento de optimización para el umbral de decisión de cada modelo. En lugar de adoptar el valor predeterminado de 0.5, se evaluaron 50 umbrales equidistantes en un rango de [0.1, 0.9] mediante la función linspace.

El objetivo de esta fase fue identificar el umbral que maximizara el F1-Score sobre el conjunto de validación. Este enfoque permite un ajuste preciso de los modelos, encontrando el balance óptimo entre los valores de precisión y la sensibilidad (recall) que mejor se adapten al desbalance de datos del estudio. Al realizar esto, se garantiza que el modelo final opere de la forma más eficiente para detectar casos de riesgo.

2.6.7 Evaluación de los modelos

Como se mencionó anteriormente, se empleó la métrica AUC-ROC para la optimización mediante Grid Search debido a su independencia del umbral de decisión, y el F1-Score para el ajuste posterior de dicho umbral. Tras obtener los resultados óptimos de cada arquitectura, se realizó una comparativa global de los modelos sobre el conjunto de validación. El criterio de selección primario fue el F1-Score, priorizando un equilibrio entre recall y precision. En casos donde los modelos pudieran tener resultados iguales, se consideró el recall como segundo criterio de decisión y el AUC-ROC como tercero. Adicionalmente, se calcularon accuracy y precision como métricas de soporte, mientras que la matriz de confusión y la curva ROC permitieron la validación visual de los resultados.

Tras este análisis, se determinó que el modelo Random Forest presentó el mejor desempeño. Una de las ventajas de este algoritmo es su capacidad para

determinar la importancia de las variables predictoras. Este cálculo se basa en la importancia de Gini, la cual mide la reducción total de la impureza de los datos lograda al realizar particiones sobre una variable específica [46]. A mayor reducción de la impureza, mayor es la relevancia de la variable en la capacidad predictiva del modelo.

Con el objetivo de simplificar el modelo y mejorar su interpretabilidad, se extrajeron las importancias de todas las características y se seleccionaron únicamente aquellas que acumulaban el 80 % de la importancia total. Finalmente, este modelo refinado fue sometido a una evaluación definitiva utilizando el conjunto de datos de prueba, garantizando así una estimación óptima e íntegra.

3 Resultados y discusión

3.1 Resultados

3.1.1 Comparación entre los modelos

En la Tabla 3.1 se detallan las configuraciones óptimas de hiperparámetros obtenidas mediante Grid Search para cada uno de los modelos analizados.

Modelo	Mejores parámetros
Random Forest	max_depth: 10 max_features: sqrt min_samples_leaf: 3 min_samples_split: 2 n_estimators: 500
Logistic Regression	C: 1.0 l1_ratio: 0.2 penalty: elasticnet solver: saga
LDA	shrinkage: None solver: svd
Bagging Tree	estimator__max_depth: 10 estimator__min_samples_leaf: 3 estimator__min_samples_split: 2 max_features: 0.7 max_samples: 0.7 n_estimators: 500
SVC	C: 10 gamma: scale kernel: rbf
Decision Tree	criterion: gini max_depth: 10 max_features: sqrt min_samples_leaf: 1 min_samples_split: 10

Tabla 3.1: Mejores hiperparámetros para cada modelo de aprendizaje automático.

La Tabla 3.2 expone la comparativa de desempeño con dichos parámetros, incluyendo el umbral óptimo y el tiempo de procesamiento de cada modelo.

Modelo	Umbral	F1-Score	Recall	AUC-ROC	Precision	Accuracy	Tiempo de procesamiento
Random Forest	0.34	0.578	0.865	0.671	0.434	0.547	26min 41s
Logistic Regression	0.39	0.574	0.878	0.665	0.426	0.533	11.4s
LDA	0.26	0.572	0.874	0.664	0.425	0.532	2.48s
Bagging Tree	0.28	0.568	0.830	0.664	0.431	0.547	18min 21s
SVC	0.28	0.561	0.835	0.648	0.423	0.533	2min 2s
Decision Tree	0.10	0.535	0.900	0.604	0.381	0.440	12.5s

Tabla 3.2: Métricas de desempeño de los mejores modelos de aprendizaje automático.

De forma complementaria, las Figuras 3.1 y 3.2 proporcionan el soporte visual mediante las matrices de confusión y las curvas ROC correspondientes.

Comparativa de Matrices de Confusión - Modelos Iniciales (Validación)

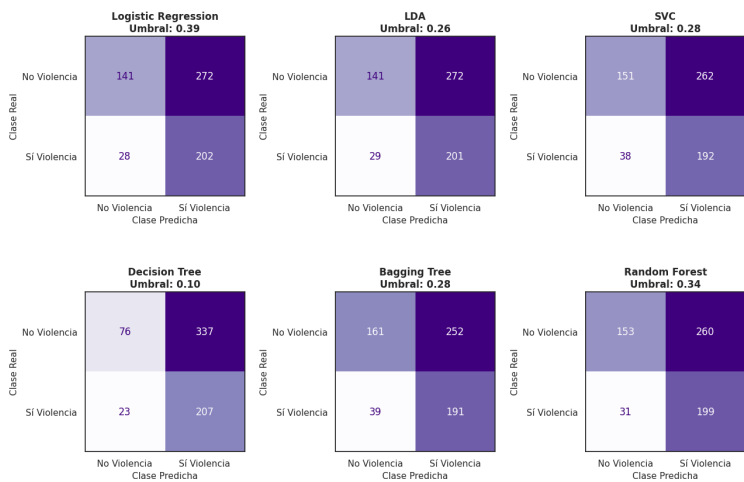


Figura 3.1: Comparativa de matrices de confusión de los modelos.

Comparativa de Curvas ROC - Modelos Iniciales (Validación)

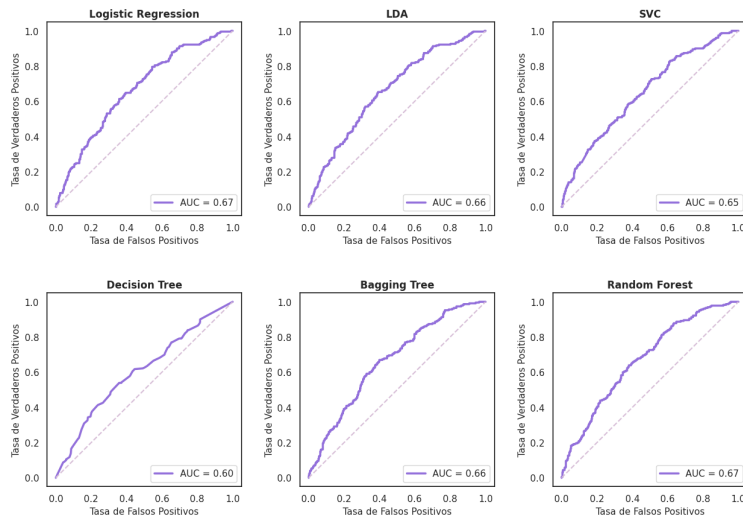


Figura 3.2: Comparativa de las curvas ROC de los modelos.

3.1.2 Refinamiento del mejor modelo

Tras evaluar las métricas de desempeño, se determinó que el modelo Random Forest presentó los mejores resultados, alcanzando el F1-Score más elevado. Aprovechando la capacidad de este algoritmo para identificar la relevancia de los predictores, se obtuvo la importancia de cada variable. Con el objetivo de optimizar el modelo, se aplicó el Principio de Pareto; si bien este sugiere que el 80 % de los efectos provienen del 20 % de las causas, se adaptó el criterio para seleccionar el conjunto de variables que explicaran el 80 % de la importancia acumulada [47]. Siguiendo este enfoque, se identificaron 9 variables con mayor importancia, las cuales se visualizan en la Figura 3.3. Estas variables se utilizaron posteriormente para re-entrenar el modelo, logrando una estructura más simplificada y eficiente.

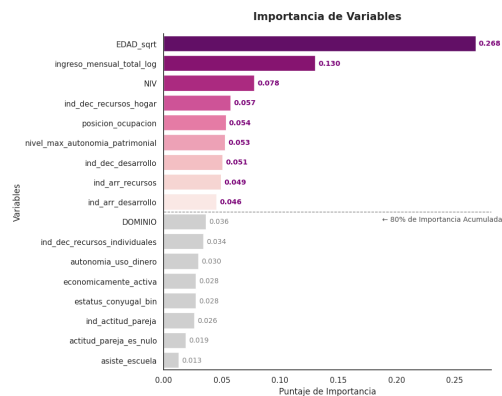


Figura 3.3: Importancia de las variables.

Luego de re-entrenar del modelo Random Forest utilizando únicamente el subconjunto de variables seleccionadas, se evaluó su desempeño en el conjunto de validación. Los resultados demostraron que la simplificación del modelo no impactó negativamente en su capacidad predictiva; por el contrario, se obtuvo una estructura más eficiente y con menor ruido.

Modelo	Mejores parámetros
Random Forest optimizado	max_depth: 10
	max_features: sqrt
	min_samples_leaf: 5
	min_samples_split: 2
	n_estimators: 500

Tabla 3.3: Mejores hiperparámetros del modelo Random Forest optimizado.

Modelo	F1-Score	Recall	AUC-ROC	Precision	Accuracy
Random Forest optimizado	0.583	0.848	0.664	0.444	0.566

Tabla 3.4: Métricas de desempeño del modelo Random Forest optimizado.

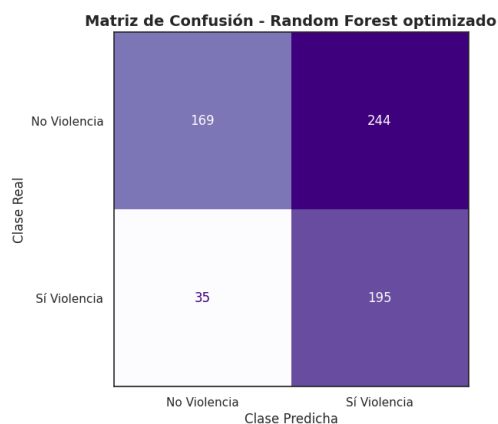


Figura 3.4: Matriz de confusión del modelo Random Forest optimizado.

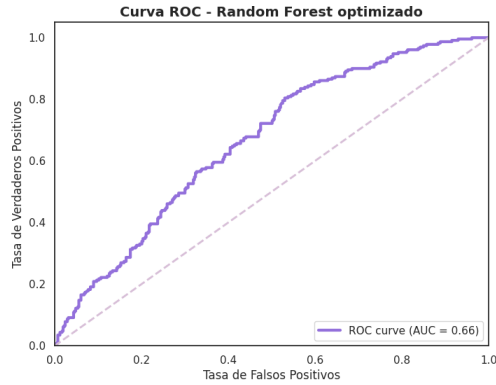


Figura 3.5: Curva ROC del modelo Random Forest optimizado.

Cabe destacar que, tras este proceso de refinamiento, las configuraciones óptimas de hiperparámetros se mantuvieron consistentes con las del modelo inicial, únicamente cambiando el parámetro “min_samples_leaf” de 3 a 5, lo que demuestra que la estructura del modelo es robusta y mantiene su eficacia incluso con menos variables.

3.1.3 Resultados en el conjunto de prueba

Finalmente, se realizó la evaluación definitiva del modelo Random Forest optimizado (utilizando el subconjunto de 9 variables) sobre el conjunto de prueba. Es fundamental destacar que estos datos se mantuvieron independientes durante todo el proceso de entrenamiento y ajuste de hiperparámetros, lo que permite verificar la efectividad real del modelo al enfrentarse a información completamente nueva y externa al proceso de entrenamiento.

En la Tabla 3.5 se presentan las métricas finales de desempeño, mientras que las Figuras 3.6 y 3.7 ilustran la matriz de confusión y la curva ROC resultantes, respectivamente. Estos valores representan la efectividad esperada del modelo ante nuevos datos de la ENDIREH para la región de Baja California.

Modelo	F1-Score	Recall	AUC-ROC	Precision	Accuracy
Random Forest optimizado (prueba)	0.553	0.769	0.637	0.432	0.558

Tabla 3.5: Métricas de desempeño del modelo Random Forest optimizado en el conjunto de prueba.

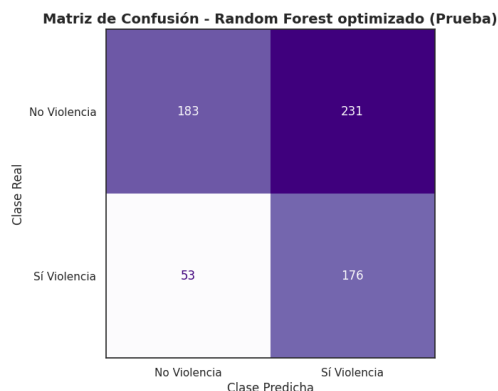


Figura 3.6: Matriz de confusión del modelo Random Forest optimizado en el conjunto de prueba.

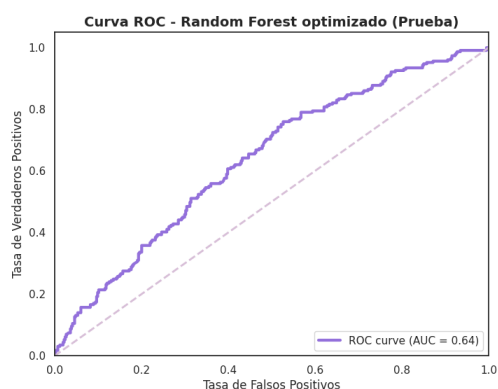


Figura 3.7: Curva ROC del modelo Random Forest optimizado en el conjunto de prueba.

3.2 Discusión de los resultados

3.2.1 Comparativa de modelos y análisis de desempeño

Se observa una consistencia notable en el aprendizaje de los modelos evaluados, con una variación de aproximadamente el 5 % entre los valores extremos de las métricas. No obstante, el algoritmo Random Forest demostró un desempeño superior, alcanzando el F1-Score más elevado (0.578). La selección de esta métrica como criterio principal se debe a su capacidad para balancear precisión y recall, factor crítico ante el desbalance de clases observado. Resulta relevante destacar que, tras el Random Forest, los modelos de Regresión Logística y Análisis Discriminante Lineal (LDA) obtuvieron los mejores resultados, lo que sugiere la presencia de componentes lineales significativos en la relación entre autonomía económica y violencia.

En cuanto a las arquitecturas basadas en árboles, los resultados validan la jerarquía teórica esperada: el Random Forest superó al Bagging Tree, y este a su vez al Decision Tree simple. Esto confirma que las técnicas de bagging y la aleatorización de variables reducen efectivamente la varianza del modelo base,

mejorando la robustez general de la predicción.

Un hallazgo fundamental es el predominio de la métrica de recall sobre las demás. Este fenómeno es consecuencia directa de la optimización del umbral de decisión, el cual se ajustó por debajo del estándar (predominantemente < 0.40). En el contexto de la violencia de género, el costo social de un falso negativo —es decir, no identificar a una mujer en situación de riesgo— es considerablemente superior al de un falso positivo. Sin embargo, como se visualiza en las matrices de confusión, esta priorización de la sensibilidad conlleva un sacrificio en la precisión, generando un volumen elevado de casos clasificados como positivos de manera errónea.

Finalmente, el análisis de eficiencia computacional revela que el Random Forest es el modelo más costoso, con un tiempo de procesamiento cercano a los 27 minutos, frente a los escasos segundos que requiere la Regresión Logística. Esta diferencia es un factor clave para la escalabilidad del proyecto; no obstante, para los fines de este estudio, la complejidad del Random Forest se justifica por su capacidad para capturar interacciones no lineales que los modelos simplificados podrían omitir.

3.2.2 Optimización y refinamiento del modelo final

Tras identificar al Random Forest como el algoritmo con mejor desempeño, se procedió a un proceso de refinamiento basado en la importancia de las características (feature importance). Mediante este análisis, se seleccionaron las variables que acumulaban el 80 % de la importancia predictiva, logrando reducir el set original de 17 variables a un subconjunto óptimo de 9.

Posteriormente, se reentrenó el algoritmo utilizando el mismo espacio de búsqueda (Grid Search) para asegurar una comparativa justa. Es notable que los hiperparámetros óptimos resultantes fueron prácticamente idénticos a los de la etapa inicial, con la única excepción de “min_samples_leaf”. Este ajuste sugiere que, al reducir la dimensionalidad del problema, el modelo adaptó su capacidad de ramificación para enfocarse en patrones más robustos y menos específicos.

Se hace evidente que el modelo mantiene su estabilidad y robustez incluso con una menor carga dimensional. La eliminación de estas 8 variables no solo simplifica la arquitectura del modelo, sino que descarta el “ruido” estadístico de predictores con baja capacidad predictiva. Al contrastar las matrices de confusión de ambas etapas, se observa una mejora cualitativa: mientras que la capacidad de detección de casos positivos (recall) se mantuvo prácticamente constante, se registró una ligera disminución en los falsos positivos. Esto se traduce en un incremento en la precisión del modelo refinado, logrando un

equilibrio más eficiente entre la sensibilidad necesaria para detectar la violencia y la reducción de clasificaciones erróneas.

3.2.3 *Evaluación final y capacidad de generalización*

Al contrastar los resultados obtenidos en el conjunto de prueba frente a los de validación, se observa que las métricas mantienen una similitud considerable. No obstante, se registra un ligero descenso en el desempeño general, lo cual es un indicador de un sobreajuste leve (overfitting). Es importante destacar que la métrica más afectada fue el Recall, lo que indica un incremento en la tasa de falsos negativos. Este comportamiento se confirma al analizar las matrices de confusión de la fase de prueba, donde el modelo presenta una mayor dificultad para identificar la totalidad de los casos positivos en datos no vistos previamente.

Por otro lado, se registró un aumento en los Verdaderos Negativos, incrementando la capacidad del modelo para identificar correctamente a las mujeres que no reportan violencia. Si bien este incremento mejora la precisión global, el interés primordial radica en la capacidad de detección de situaciones de riesgo. Por lo tanto, aunque el modelo demuestra ser robusto y capaz de generalizar los patrones de autonomía económica, este ligero aumento en las omisiones resalta la complejidad de predecir un fenómeno social tan dinámico mediante variables puramente económicas.

3.2.4 *Interpretación de resultados en el contexto del estudio*

Los resultados obtenidos en la fase de prueba muestran un recall considerablemente alto, un logro relevante dada la complejidad del concepto de autonomía económica y su relación con la violencia de género. El modelo logra identificar correctamente la mayoría de los casos positivos; sin embargo, esta sensibilidad conlleva un costo elevado en términos de falsos positivos. Esto es producto de la optimización del umbral de decisión, una estrategia diseñada para maximizar la detección de casos de riesgo. No obstante, en un escenario de implementación real, una baja precisión podría comprometer la optimización de recursos, distribuyendo el presupuesto institucional limitado en casos de mujeres que no presentan el mismo riesgo y limitando la pronta atención a mujeres en situaciones de alta prioridad.

Es necesario reconocer que el poder predictivo global del modelo es moderado. Si bien se logra la clasificación, la precisión sugiere que las variables económicas, por sí solas, no capturan la totalidad de la problemática. Este resultado es congruente con el Análisis Exploratorio de Datos (EDA) inicial, donde se observaron correlaciones débiles y la ausencia de patrones lineales

concluyentes. La matriz de correlación ya anticipaba que la variable objetivo está influenciada por factores más allá de lo económico.

A pesar de estas limitaciones, el análisis de importancia de variables del Random Forest aporta hallazgos sociológicos significativos. El hecho de que la edad sea el predictor principal (26.8%), seguido del ingreso mensual total, describe un patrón de riesgo específico: las mujeres jóvenes en edad laboral activa presentan un mayor riesgo de sufrir violencia. En consecuencia, el empoderamiento financiero temprano no garantiza una inmunidad frente a la agresión, sino que puede representar una etapa crítica donde la reconfiguración de roles económicos aumenta la exposición al riesgo.

4 Conclusiones y trabajo futuro

4.1 Conclusiones

Tras realizar los procedimientos descritos en este trabajo, se concluye que se cumplieron los objetivos propuestos al caracterizar el entorno sociodemográfico y la autonomía económica de las mujeres del estudio, logrando desarrollar un modelo predictivo capaz de identificar situaciones de riesgo. Este análisis permitió entender mejor cómo estas variables influyen en el fenómeno de la violencia. Asimismo, al evaluar los seis algoritmos de aprendizaje automático, se determinó que Random Forest fue el más efectivo. Su capacidad para capturar las interacciones complejas entre los datos lo posicionó por encima del resto de las propuestas, confirmando que un problema social tan multidimensional requiere herramientas con un alto nivel de robustez. No obstante, el modelo presenta un margen de mejora considerable en su precisión, ya que, incluso buscando el mejor equilibrio mediante el F1-Score, se observa un volumen elevado de falsos positivos.

Posteriormente a la selección de Random Forest, se realizó un proceso de filtrado basado en la importancia de las variables (Feature Importance). Este paso fue fundamental, no con el objetivo de elevar las métricas, sino para simplificar el modelo mediante la identificación de variables con bajo poder predictivo. Al eliminar este "ruido", se logró una arquitectura más robusta, simple y estable que mantuvo resultados muy similares a la versión original. Este proceso de refinamiento demuestra que es posible explicar el fenómeno utilizando un conjunto más compacto de variables, lo que facilita la interpretación de los factores que realmente inciden en la autonomía económica y el riesgo de violencia.

A partir del ranking de importancia de variables, se identificó que los factores más determinantes son la edad y el ingreso mensual total. Al contrastar esto con el análisis exploratorio, donde se observó que las mujeres más jóvenes tienden a sufrir más violencia, se deduce que quienes se incorporan al mercado laboral atraviesan una etapa de mayor vulnerabilidad. Debido a los roles de género tan arraigados en nuestra sociedad, el incremento en los ingresos de la mujer puede generar tensiones en las dinámicas de poder, lo que en ocasiones provoca reacciones violentas por parte de la pareja al sentir amenazado su rol

tradicional. Esto sugiere que la autonomía económica real no solo depende de recibir un salario, sino de poseer poder de decisión sobre el dinero y acceso a otros activos, como propiedades o posiciones de liderazgo. Finalmente, el estudio confirma que la autonomía económica no es el único detonante de la violencia, ya que este fenómeno es un problema multicausal influenciado por una amplia red de factores sociales y culturales.

Para finalizar, este trabajo demuestra que la ciencia de datos es una herramienta útil y eficaz para abordar problemáticas sociales críticas como la violencia de género. La existencia de encuestas como la ENDIREH es fundamental para dimensionar estadísticamente la gravedad de una situación que suele verse como casos aislados en los medios de comunicación. En este sentido, integrar el uso de la inteligencia artificial permite transformar esos datos en evidencia para la toma de decisiones que apoye a las instituciones. Resulta vital continuar realizando estudios que caractericen los entornos sociodemográficos de las mujeres, ya que solo mediante la identificación de patrones claros será posible diseñar medidas de prevención y protección más efectivas. El uso ético y constante de la tecnología aplicada a los datos abre una vía de esperanza para mitigar este fenómeno y mejorar la seguridad de las mujeres en nuestra sociedad.

4.2 Trabajo futuro

A pesar de haber desarrollado un modelo sólido y funcional, tanto la ciencia de datos como los fenómenos sociales se encuentran en constante evolución. La complejidad de la violencia de género y la naturaleza iterativa de los procesos analíticos implican que siempre existan oportunidades para refinar los hallazgos y profundizar en el estudio. En este sentido, y con el fin de robustecer las herramientas de detección desarrolladas en este proyecto, se presentan las siguientes propuestas de trabajo a futuro:

4.2.1 Inclusión de variables multidimensionales

Dado que se comprobó que la violencia de género es un fenómeno multicausal que trasciende el ámbito financiero, se propone la integración de variables pertenecientes a otras dimensiones sociales y relacionales. Incorporar indicadores sobre dinámicas específicas de pareja, procesos de toma de decisiones en el hogar o antecedentes del entorno familiar permitiría construir un modelo mucho más robusto. Este enfoque multidimensional facilitaría una predicción más certera del riesgo, capturando la complejidad del fenómeno de manera más integral.

4.2.2 Exploración de algoritmos de mayor complejidad

Si bien los modelos implementados en este estudio, especialmente Random Forest, representan un punto de partida robusto, existen arquitecturas más avanzadas que podrían capturar relaciones aún más intrincadas entre las variables. Se propone para futuras etapas la implementación de algoritmos basados en técnicas de Boosting, tales como AdaBoost y XGBoost. Estos modelos, al aprender de manera secuencial de los errores de clasificación previos, podrían adecuarse de forma óptima a la naturaleza de los datos de la ENDIREH o a la integración de nuevas variables, buscando mejorar la precisión del sistema y el equilibrio del F1-Score.

4.2.3 Replicabilidad en otros contextos regionales

El presente estudio se desarrolló específicamente en el contexto de Baja California. Como trabajo a futuro, se propone replicar esta metodología en otras entidades de la República, integrando estados de las regiones centro y sur del país. Esto permitiría contrastar las dinámicas entre la autonomía económica y la violencia de género en contextos geográficos y culturales distintos, ayudando a identificar si los patrones de riesgo detectados en el norte se mantienen o si varían significativamente según el entorno regional.

4.2.4 Incorporación de la ENDIREH 2026

Dado que se prevé la aplicación de la sexta edición de la ENDIREH durante el año 2026, la integración de estos nuevos datos representa una gran oportunidad para actualizar el estudio. Si bien la edición de 2021 proporciona una base sólida, sus resultados reflejan un contexto post-pandemia inmediato que pudo haber influido en las dinámicas de convivencia y economía familiar. El uso de la información más reciente permitirá validar si los patrones de riesgo identificados persisten en la actualidad o si han surgido nuevas tendencias en la relación entre la autonomía de las mujeres y la violencia de género, asegurando que las herramientas de detección se mantengan vigentes y precisas.

Bibliografía

- [1] O. N. U. España. «La violencia de género es una de las violaciones más generalizadas de los derechos humanos en el mundo», visitado 31 de mar. de 2026. dirección: <https://unric.org/es/la-violencia-de-genero-segun-la-onu/>.
- [2] S. Left. «Datos y cifras: violencia contra las mujeres», visitado 31 de mar. de 2026. dirección: <https://www.unwomen.org/es/articulos/datos-y-cifras/datos-y-cifras-violencia-contra-las-mujeres>.
- [3] «Autonomía económica para salir de una situación de violencia», visitado 31 de mar. de 2026. dirección: <https://mexico.unwomen.org/es/noticias-y-eventos/articulos/2021/11/autonomia-economica-para-salir-de-una-situacion-de-violencia>.
- [4] «Autonomía económica | Observatorio de Igualdad de Género», visitado 31 de mar. de 2026. dirección: <https://oig.cepal.org/es/autonomias/autonomia-economica>.
- [5] *Ley General de Acceso de las Mujeres a una Vida Libre de Violencias*. dirección: <https://www.diputados.gob.mx/LeyesBiblio/pdf/LGAMVLV.pdf>.
- [6] «Movilidad Humana Baja California», Gobierno de México, 2021. dirección: https://portales.segob.gob.mx/work/models/PoliticaMigratoria/CPM/foros_regionales/estados/norte/info_diag_F_norte/info_BajaCalif.pdf.
- [7] «Cali baja, la dinámica megaregión en la frontera entre estados unidos y méxico que genera millones de dólares y empleos», *BBC News Mundo*, visitado 31 de mar. de 2026. dirección: <https://www.bbc.com/mundo/noticias-60891469>.
- [8] «Marco Geoestadístico», visitado 14 de mayo de 2026. dirección: <https://www.inegi.org.mx/temas/mg/>.
- [9] Secretaría de Gobernación, *Declaratoria de Alerta de Violencia de Género contra las Mujeres en el Estado de Baja California*, 25 de jun. de 2021. dirección: https://www.gob.mx/cms/uploads/attachment/file/753809/24._Declaratoria_AVGM_BC_25-06-21.pdf.

- [10] «Congreso del Estado de Baja California», visitado 31 de mar. de 2026. dirección: <https://www.congresobc.gob.mx/Informacion/BCLibreViolencia>.
- [11] Instituto Nacional de las Mujeres, «Las mujeres y la autonomía económica. Desigualdad, Año 9, Boletín No. 1», INMUJERES, 2023. dirección: http://cedoc.inmujeres.gob.mx/documentos_download/BA9N01.pdf.
- [12] Instituto Nacional de Estadística y Geografía, «Encuesta Nacional de Ocupación y Empleo (Nueva Edición). ENOEN: Resultados del segundo trimestre de 2021», INEGI, 2021. dirección: https://www.inegi.org.mx/contenidos/programas/enoe/15ymas/doc/enoe_n_presentacion_ejecutiva_trim2_2021.pdf.
- [13] «RESULTADOS DE LA ENCUESTA NACIONAL DE OCUPACIÓN Y EMPLEO. NUEVA EDICIÓN (ENOEN)», 2021. dirección: https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2021/enoe_ie/enoe_ie2021_08_BC.pdf.
- [14] «Violencia contra la mujer», visitado 1 de abr. de 2026. dirección: <https://www.who.int/es/news-room/fact-sheets/detail/violence-against-women>.
- [15] «Violencia contra las mujeres en México», visitado 31 de mar. de 2026. dirección: <https://www.inegi.org.mx/tablerosestadisticos/vcmm/>.
- [16] Instituto Nacional de Estadística y Geografía (INEGI). «Quiénes somos», visitado 1 de abr. de 2026. dirección: https://www.inegi.org.mx/inegi/quienes_somos.html.
- [17] Instituto Nacional de Estadística y Geografía, «Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares (ENDIREH) 2021. Principales resultados.», INEGI. dirección: https://www.inegi.org.mx/contenidos/programas/endireh/2021/doc/endireh2021_presentacion_ejecutiva.pdf.
- [18] «Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares (ENDIREH) 2021», visitado 1 de abr. de 2026. dirección: <https://www.inegi.org.mx/programas/endireh/2021/>.
- [19] «Mexico - Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares (ENDIREH) 2021», visitado 1 de abr. de 2026. dirección: <https://www.inegi.org.mx/rnm/index.php/catalog/801/study-description>.
- [20] «Mexico - Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares (ENDIREH) 2021», visitado 2 de abr. de 2026. dirección: <https://www.inegi.org.mx/rnm/index.php/catalog/801/data-dictionary>.

- [21] «Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares (ENDIREH) 2021», visitado 15 de mayo de 2026. dirección: <https://www.inegi.org.mx/programas/endireh/2021/#microdatos>.
- [22] «Una comparación de los métodos de correlación de Pearson y Spearman», visitado 18 de abr. de 2026. dirección: <https://support.minitab.com/es-mx/minitab/help-and-how-to/statistics/basic-statistics/supporting-topics/correlation-and-covariance/a-comparison-of-the-pearson-and-spearman-correlation-methods/>.
- [23] F. Lee. «¿Qué es la regresión logística? | IBM», visitado 21 de abr. de 2026. dirección: <https://www.ibm.com/mx-es/think/topics/logistic-regression>.
- [24] G. James, D. Witten, T. Hastie y R. Tibshirani, *An Introduction to Statistical Learning: with Applications in Python* (Springer), 2nd. New York: Springer, 2023, ISBN: 978-1-0716-2916-1. dirección: <https://www.statlearning.com/>.
- [25] «¿Qué es la regresión logística? - Explicación del modelo de regresión logística - AWS», visitado 21 de abr. de 2026. dirección: <https://aws.amazon.com/es/what-is/logistic-regression/>.
- [26] «Advantages and Disadvantages of Logistic Regression», visitado 21 de abr. de 2026. dirección: <https://www.geeksforgeeks.org/data-science/advantages-and-disadvantages-of-logistic-regression/>.
- [27] «¿Qué es el análisis discriminatorio lineal? | IBM», visitado 21 de abr. de 2026. dirección: <https://www.ibm.com/mx-es/think/topics/linear-discriminant-analysis>.
- [28] «Linear Discriminant Analysis: Beyond Dimension Reduction», visitado 21 de abr. de 2026. dirección: <https://www.datacamp.com/tutorial/linear-discriminant-analysis>.
- [29] E. Kavlakoglu. «¿Qué es Support Vector Machine? | IBM», visitado 21 de abr. de 2026. dirección: <https://www.ibm.com/mx-es/think/topics/support-vector-machine>.
- [30] «Support Vector Machine (SVM) Algorithm», visitado 21 de abr. de 2026. dirección: <https://www.geeksforgeeks.org/machine-learning/support-vector-machine-algorithm/>.
- [31] E. Kavlakoglu. «¿Qué es un árbol de decisión? | IBM», visitado 22 de abr. de 2026. dirección: <https://www.ibm.com/mx-es/think/topics/decision-trees>.

- [32] M. P.-D. 3. M. Brief. «Unlocking the Power of Decision Trees: Advantages and Applications in Machine Learning», Medium, visitado 22 de abr. de 2026. dirección: <https://medium.com/@jangdaehan1/unlocking-the-power-of-decision-trees-advantages-and-applications-in-machine-learning-9d560cf9c4bb>.
- [33] «¿Qué es el bagging? | IBM», visitado 22 de abr. de 2026. dirección: <https://www.ibm.com/mx-es/think/topics/bagging>.
- [34] «What is Bagging Classifier», visitado 22 de abr. de 2026. dirección: <https://www.geeksforgeeks.org/machine-learning/what-is-bagging-classifier/>.
- [35] E. Kavlakoglu. «¿Qué es random forest? | IBM», visitado 22 de abr. de 2026. dirección: <https://www.ibm.com/mx-es/think/topics/random-forest>.
- [36] «What are the Advantages and Disadvantages of Random Forest?», visitado 22 de abr. de 2026. dirección: <https://www.geeksforgeeks.org/machine-learning/what-are-the-advantages-and-disadvantages-of-random-forest/>.
- [37] P. Programmer. «Evaluation Metrics for Classification», Medium, visitado 22 de abr. de 2026. dirección: <https://medium.com/@mlmind/evaluation-metrics-for-classification-fc770511052d>.
- [38] *Receiver operating characteristic*, en *Wikipedia*, Page Version ID: 1339379744, 20 de feb. de 2026. visitado 21 de mayo de 2026. dirección: https://en.wikipedia.org/w/index.php?title=Receiver_operating_characteristic&oldid=1339379744.
- [39] T. Shah. «About Train, Validation and Test Sets in Machine Learning», TDS Archive, visitado 19 de abr. de 2026. dirección: <https://medium.com/data-science/train-validation-and-test-sets-72cb40cba9e7>.
- [40] *Winsorizing*, en *Wikipedia*, Page Version ID: 1314972465, 4 de oct. de 2025. visitado 18 de abr. de 2026. dirección: <https://en.wikipedia.org/w/index.php?title=Winsorizing&oldid=1314972465>.
- [41] «StandardScaler, MinMaxScaler and RobustScaler techniques - ML», visitado 19 de abr. de 2026. dirección: <https://www.geeksforgeeks.org/machine-learning/standardscaler-minmaxscaler-and-robustscaler-techniques-ml/>.
- [42] «Python Machine Learning - Grid Search», visitado 23 de abr. de 2026. dirección: https://www.w3schools.com/PYTHON/python_ml_grid_search.asp.

- [43] «AUC and the ROC Curve in Machine Learning», visitado 23 de abr. de 2026. dirección: <https://www.datacamp.com/tutorial/auc>.
- [44] «3.1. Cross-validation: evaluating estimator performance», visitado 23 de abr. de 2026. dirección: https://scikit-learn/stable/modules/cross_validation.html.
- [45] «How Does the class_weight Parameter in Scikit-Learn Work?», visitado 23 de abr. de 2026. dirección: <https://www.geeksforgeeks.org/machine-learning/how-does-the-classweight-parameter-in-scikit-learn-work/>.
- [46] P. Ghattikar. «Using Random Forest For Feature Importance And Feature Selection», Medium, visitado 23 de abr. de 2026. dirección: <https://medium.com/@prasannarghattikar/using-random-forest-for-feature-importance-118462c40189>.
- [47] *Principio de Pareto*, en *Wikipedia, la enciclopedia libre*, Page Version ID: 172064177, 14 de feb. de 2026. visitado 27 de abr. de 2026. dirección: https://es.wikipedia.org/w/index.php?title=Principio_de_Pareto&oldid=172064177.