UNIVERSIDAD DEL VALLE DE GUATEMALA
Minería de datos, sección 30

Entrega 01 de Avances (EDA)
Proyecto 3

Sofia Garcia – 22210
Julio Garcia Salas – 22076
Sebastian Garcia – 22291
Jose Angel Morales – 22689

GUATEMALA, marzo de 2025

Índice

Índice 2
Introducción 2
Problema Científico 3
Objetivos 3
Objetivo general 3
Objetivos específicos 3
Descripción de variables 3
Clustering 10
Análisis exploratorio (tipo_medida) 13
Las mejores variables predictoras 17
Descripción de los datos para responder el problema planteado 17
Conclusiones 19

Introducción

La violencia intrafamiliar constituye una de las violaciones de derechos humanos más severas en Guatemala, impactando de forma desproporcionada a mujeres, niñas y niños. Sus consecuencias inmediatas incluyen daños físicos, trauma psicológico y desintegración familiar, mientras que a largo plazo incrementa la carga sobre el sistema de salud, sobrecarga al sistema judicial y perpetúa ciclos de violencia intergeneracional. Contar con evidencia cuantitativa precisa sobre los factores que influyen en las decisiones judiciales es crucial para diseñar políticas públicas y protocolos de atención más efectivos.

Este estudio aprovecha un repositorio único que agrupa más de diez años (2014–2023) de denuncias oficiales de violencia intrafamiliar y otros tipos de agresión en Guatemala, con más de 70 variables por caso —desde datos sociodemográficos hasta medidas judiciales aplicadas—. El objetivo central es desarrollar un modelo predictivo que, a partir de características observables de cada denuncia, estime el tipo de medida legal que se impondrá al presunto agresor. Esta aproximación permitirá anticipar patrones de decisión judicial y proporcionar información útil para optimizar recursos institucionales y fortalecer los mecanismos de protección

Situación Problemática

Las denuncias por violencia intrafamiliar en Guatemala han crecido un 25% entre 2019 y 2023, superando las 35 000 denuncias anuales, según registros oficiales. Las mujeres representan más del 85% de las víctimas, mientras que niñas y niños —aunque subregistrados— sufren impactos severos en su desarrollo integral. A pesar de esta realidad alarmante, la mayoría de los estudios existentes se centran en análisis descriptivos de prevalencia, sin explorar cómo variables como el tipo de agresión, antecedentes penales, ubicación geográfica, o el nivel socioeconómico inciden en la selección de medidas judiciales.

Esta brecha limita la capacidad de las instituciones para diseñar respuestas diferenciadas y basadas en evidencia. Las medidas legales disponibles en Guatemala —desde órdenes de alejamiento y medidas cautelares hasta procesos penales formales— se aplican de manera heterogénea, lo que puede generar inconsistencias, retrasos en la protección de las víctimas y revictimización.

Problema Científico

¿Es posible predecir el tipo de medida judicial que se impondrá a una persona denunciada por violencia intrafamiliar utilizando únicamente los datos administrativos de la denuncia? Resolver esta pregunta implica identificar cuáles variables sociodemográficas, contextuales e institucionales tienen mayor capacidad explicativa sobre la decisión judicial, y determinar si esta información aporta señal suficiente para construir un modelo predictivo robusto y generalizable.

Objetivos

Objetivo general

Desarrollar y validar un modelo de clasificación supervisada capaz de predecir el tipo de medida legal impuesta a presuntos agresores en casos de violencia intrafamiliar a partir de datos administrativos.

Objetivos específicos

1. Identificar y cuantificar la influencia de variables sociodemográficas, contextuales e institucionales sobre la probabilidad de cada tipo de medida judicial.
2. Evaluar el desempeño del modelo mediante métricas de clasificación (precisión, recall, F1-score) y validación cruzada para garantizar su robustez y capacidad de generalización.
3. Generar perfiles interpretables que permitan entender los factores determinantes en la asignación de medidas legales, facilitando la toma de decisiones institucionales y la priorización de recursos según contexto geográfico y nivel de riesgo.

Descripción de variables

El archivo todoscsvs.csv reúne 10 años (2014–2023) de denuncias por violencia — incluyendo violencia intrafamiliar— registradas ante instituciones oficiales en Guatemala.

Cada fila es una denuncia; cada columna, un atributo. Teniendo un total de (328959, 76) el dataset.

Alcance temporal y geográfico

- Periodo: Enero 2014 – Diciembre 2023
- Ubicación: Departamentos y municipios de Guatemala

Tamaño del dataset

- Filas (observaciones): Corresponde a cada denuncia registrada
- Columnas (variables): 70

Grupos de variables

Categoría: Fecha ocurrencia

- Variables: dia_ocurrencia, mes_ocurrencia, anio_ocurrencia
- Tipo de dato: Categórico
- Descripción breve: Día, mes y año en que ocurrió el hecho.

Categoría: Ubicación ocurrencia

- Variables: dep_municipio_ocurrencia
- Tipo de dato: Categórico
- Descripción breve: Código (XXYY) de departamento+municipio donde ocurrió el hecho.

Categoría: Tipo de agresión

- Variables: tipo_agresion
- Tipo de dato: Categórico
- Descripción breve: Código del tipo de agresión sufrida.

Categoría: Fecha registro

- Variables: dia_registro, mes_registro, anio_registro
- Tipo de dato: Categórico
- Descripción breve: Día, mes y año en que se registró la denuncia.

Categoría: Ubicación registro

- Variables: dep_municipio_registro, departamento_registro
- Tipo de dato: Categórico

- Descripción breve: Código municipio y departamento donde se registró la denuncia.

Categoría: Reportante

- Variables: quien_reporta
- Tipo de dato: Categórico
- Descripción breve: Código de la persona que reportó el hecho.

Categoría: Víctima

- Variables: sexo_victima, edad_victima, total_hijos_victima, hijos_hombres_victima, hijas_mujeres_victima, alfabeta_victima, escolaridad_victima, estado_civil_victima, pueblo_victima, nacionalidad_victima, trabaja_victima, ocupacion_victima, dedica_victima, discapacidad_victima, tipo_discapacidad_victima, relacion_victima_agresor, otras_victimas_total, otras_victimas_hombres, otras_victimas_mujeres, otras_victimas_ninos, otras_victimas_ninas
- Tipo de dato: Mixto
- Descripción breve: Datos demográficos y contexto de la víctima.

Categoría: Agresor

- Variables: sexo_agresor, edad_agresor, alfabeta_agresor, escolaridad_agresor, estado_civil_agresor, AGR_GURPET (pueblo_agresor), nacionalidad_agresor, trabaja_agresor, ocupacion_agresor, dedica_agresor, otros_agresores_total, otros_agresores_hombres, otros_agresores_mujeres, otros_agresores_ninos, otros_agresores_ninas
- Tipo de dato: Mixto
- Descripción breve: Datos demográficos y contexto del agresor.

Categoría: Instituciones

- Variables: institucion_denuncia_previa, institucion_registro, organismo_jurisdiccional, conducente, organismo_remite
- Tipo de dato: Categórico
- Descripción breve: Instituciones involucradas en la denuncia.

Categoría: Legislación y medidas

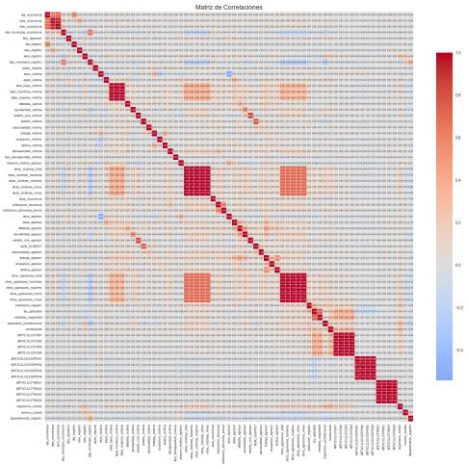
- Variables: ley_aplicable, medidas_seguridad, tipo_medida, ARTICULOVCM1–4, ARTICULOCODPEN1–4, ARTICULOTRAS1–4
- Tipo de dato: Categórico
- Descripción breve: Leyes aplicadas y medidas judiciales.

Categoría: Identificación

- Variables: numero_boleta
- Tipo de dato: Numérico
- Descripción breve: Folio interno de la denuncia.

Resumen de las variables Numéricas (Matriz de correlación)

Se realizo una matriz de correlación con las columnas numéricas, con el propósito de poder observar las relaciones que tienen las variables entre sí, para ver si se encontraban, tendencia o patrones en el conjunto de datos.



Luego se realizó una Prueba de normalidad (Shapiro–Wilk), este consiste en una prueba estadística que verifica si los datos cuentan con una distribución normal. La hipotesis nula de esta prueba indica que los datos siguen una distribución normal, si el valor p es mayor a 0.05, no se rechaza la hipótesis nula, indicando que, si siguen una distribución normal, si el valor de p es menor a 0.05 si se rechaza la hipótesis indicando que los datos no siguen una distribución normal.

Resultados de la prueba de Shapiro-Wilk:

- **dep_municipio_ocurrencia**: W=0.7391, p=0.0000 (No normal)
- **tipo_agresion**: W=0.6681, p=0.0000 (No normal)
- **dep_municipio_registro**: W=0.9226, p=0.0000 (No normal)
- **quien_reporta**: W=0.1207, p=0.0000 (No normal)
- **sexo_victima**: W=0.3876, p=0.0000 (No normal)
- **edad_victima**: W=0.8623, p=0.0000 (No normal)
- **total_hijos_victima**: W=0.5217, p=0.0000 (No normal)

- **hijos_hombres_victima**: W=0.5098, p=0.0000 (No normal)
- **hijas_mujeres_victima**: W=0.5093, p=0.0000 (No normal)
- **alfabeta_victima**: W=0.2628, p=0.0000 (No normal)
- **escolaridad_victima**: W=0.8461, p=0.0000 (No normal)
- **estado_civil_victima**: W=0.6971, p=0.0000 (No normal)
- **pueblo_victima**: W=0.5959, p=0.0000 (No normal)
- **nacionalidad_victima**: W=0.0564, p=0.0000 (No normal)
- **trabaja_victima**: W=0.4437, p=0.0000 (No normal)
- **ocupacion_victima**: W=0.6617, p=0.0000 (No normal)
- **dedica_victima**: W=0.1631, p=0.0000 (No normal)
- **discapacidad_victima**: W=0.2118, p=0.0000 (No normal)
- **tipo_discapacidad_victima**: W=0.0390, p=0.0000 (No normal)
- **relacion_victima_agresor**: W=0.7323, p=0.0000 (No normal)
- **otras_victimas_total**: W=0.6302, p=0.0000 (No normal)
- **otras_victimas_hombres**: W=0.6205, p=0.0000 (No normal)
- **otras_victimas_mujeres**: W=0.6214, p=0.0000 (No normal)
- **otras_victimas_ninos**: W=0.6237, p=0.0000 (No normal)
- **otras_victimas_ninas**: W=0.6234, p=0.0000 (No normal)
- **area_ocurrencia**: W=0.4005, p=0.0000 (No normal)
- **reiteracion_denuncia**: W=0.2946, p=0.0000 (No normal)
- **institucion_denuncia_previa**: W=0.3128, p=0.0000 (No normal)
- **sexo_agresor**: W=0.4354, p=0.0000 (No normal)
- **edad_agresor**: W=0.7663, p=0.0000 (No normal)
- **alfabeta_agresor**: W=0.2468, p=0.0000 (No normal)
- **escolaridad_agresor**: W=0.8023, p=0.0000 (No normal)
- **estado_civil_agresor**: W=0.7008, p=0.0000 (No normal)
- **AGR_GURPET**: W=0.6022, p=0.0000 (No normal)
- **nacionalidad_agresor**: W=0.1429, p=0.0000 (No normal)

- **trabaja_agresor**: W=0.3398, p=0.0000 (No normal)
- **ocupacion_agresor**: W=0.9392, p=0.0000 (No normal)
- **dedica_agresor**: W=0.3403, p=0.0000 (No normal)
- **otros_agresores_total**: W=0.6390, p=0.0000 (No normal)
- **otros_agresores_hombres**: W=0.6375, p=0.0000 (No normal)
- **otros_agresores_mujeres**: W=0.6378, p=0.0000 (No normal)
- **otros_agresores_ninos**: W=0.6368, p=0.0000 (No normal)
- **otros_agresores_ninas**: W=0.6367, p=0.0000 (No normal)
- **institucion_registro**: W=0.7027, p=0.0000 (No normal)
- **ley_aplicable**: W=0.4381, p=0.0000 (No normal)
- **medidas_seguridad**: W=0.1104, p=0.0000 (No normal)
- **organismo_jurisdiccional**: W=0.2996, p=0.0000 (No normal)
- **conducente**: W=0.2696, p=0.0000 (No normal)
- **ARTICULOVCM1**: W=0.0996, p=0.0000 (No normal)
- **ARTICULOVCM2**: W=0.0998, p=0.0... (No normal)

Como se observó, todos los valores p de las variables son menores a 0.05, por lo tanto, la hipótesis nula de que los datos siguen una distribución normal fue rechazada.

Analisis de frecuencia de variables Categoricas

dia_ocurrencia	frecuencia
1	13745
2	12169
5	11658
3	11568
15	11545
10	11475
4	11291
20	11182
8	10893
18	10881
12	10853
6	10832
17	10827
7	10786
9	10712
16	10542
25	10536
11	10482
14	10355
13	10296

mes_ocurrencia	frecuencia
5	28816
7	28770
8	28355
3	28201
6	27399
4	27306
1	26868
2	26619
9	26467
10	26191
11	25420
12	24807
99	3740

anio_ocurrencia	frecuencia
2022	36565
2021	36318
2023	35832
2014	33653
2019	31538
2015	31343
2016	30845
2017	29957
2018	29895
2020	28270
9999	3740
2013	667
2012	99
2011	62
2010	57
2009	28
2008	19
2001	15
2006	15
2004	12

dia_registro	frecuencia
3	12188
4	11883
5	11715
2	11704
18	11436
6	11420
8	11345
7	11191
11	11006
13	10990
17	10979
12	10921
9	10920
27	10905
10	10881
16	10828
19	10810
21	10777
20	10755
28	10737

mes_registro	frecuencia
7	29257
8	29088
5	29041
3	28644
6	27724
4	27552
9	26939
1	26898
2	26648
10	26549
11	25992
12	24627

anio_registro	frecuencia
2023	37348
2022	37194
2021	36435
2014	34330
2015	31929
2019	31898
2016	31190
2017	30384
2018	29992
2020	28259

tipo_medida	frecuencia
IJ	260141
AIJ	20239
Z	12343
I	6985
CIJ	5741
IJN	2652
ACIJ	2508
IJK	1334
BIJ	1144
ABC	788
AIJK	675
FIJ	657
AB	603
ABIJ	509
J	505
FGIJ	485
GIJ	483
BI	406
CIJN	386
BCI	378

cod_dep_ocurrencia

cod_dep_ocurrencia	frecuencia
Unknown	328959

cod_mun_ocurrencia

cod_mun_ocurrencia	frecuencia
Unknown	328959

cod_dep_registro

cod_dep_registro	frecuencia
Unknown	328959

cod_mun_registro

cod_mun_registro	frecuencia
Unknown	328959

Día, Mes y Año de Ocurrencia

Se observo que los eventos tienden a concentrarse en los primeros días del mes, con el día 1 registrando la mayor frecuencia (13,745 eventos). En cuanto a los meses, se aprecia un incremento en la frecuencia durante el periodo intermedio del año, destacando el mes de mayo con 28,816 registros. Los años más recientes presentan una mayor cantidad de registros, siendo 2022 el año con la cifra más elevada (36,565 eventos), lo que podría indicar un aumento en la incidencia de los eventos o una mejora en la recolección de datos.

Día, Mes y Año de Registro

El patrón de registro es similar al de ocurrencia, con una mayor frecuencia en los primeros días del mes, donde el día 3 presenta 12,188 registros. Los meses intermedios también destacan por un número elevado de registros, sobresaliendo julio con 29,257 registros. Los años más recientes concentran la mayor parte de los registros, con 2023 liderando con 37,348 registros, lo que sugiere una tendencia creciente en la documentación de los eventos.

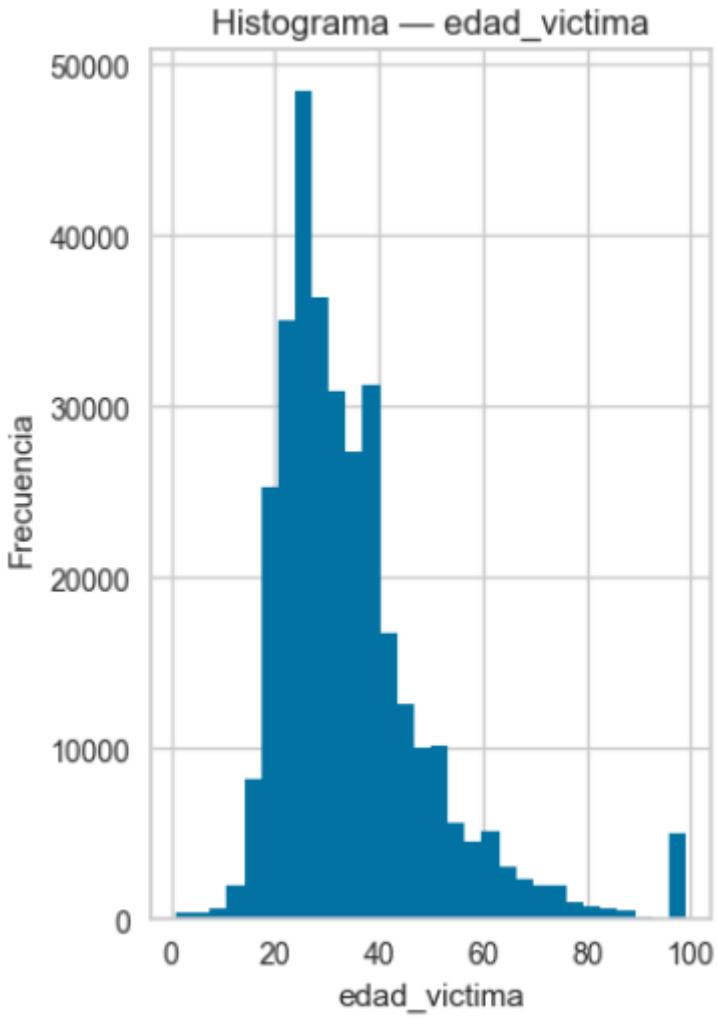
Calidad de los Datos Geográficos

Se detecta una ausencia total de información en las variables relacionadas con la ubicación geográfica (cod_dep_ocurrencia, cod_mun_ocurrencia, cod_dep_registro, cod_mun_registro), ya que el 100 % de los registros (328,959 eventos) están clasificados como "Unknown".

Distribución del Tipo de Medida

La variable relacionada con el tipo de medida muestra una concentración significativa en pocas categorías, con una amplia mayoría de registros (260,141 eventos) pertenecientes a la clasificación "IJ". Las demás categorías tienen una representación mucho menor, por ejemplo, "AIJ" con 20,239 registros y "Z" con 12,343 registros, lo que indica una distribución desigual.

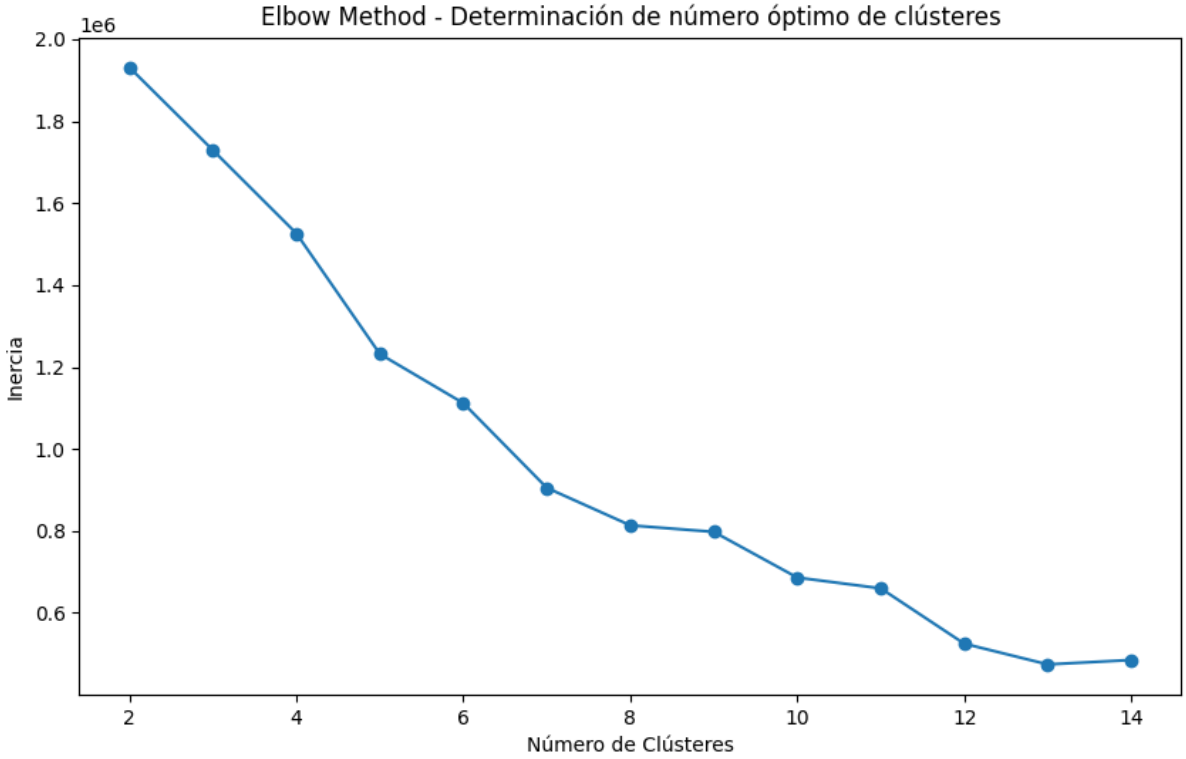
Gráficos exploratorios



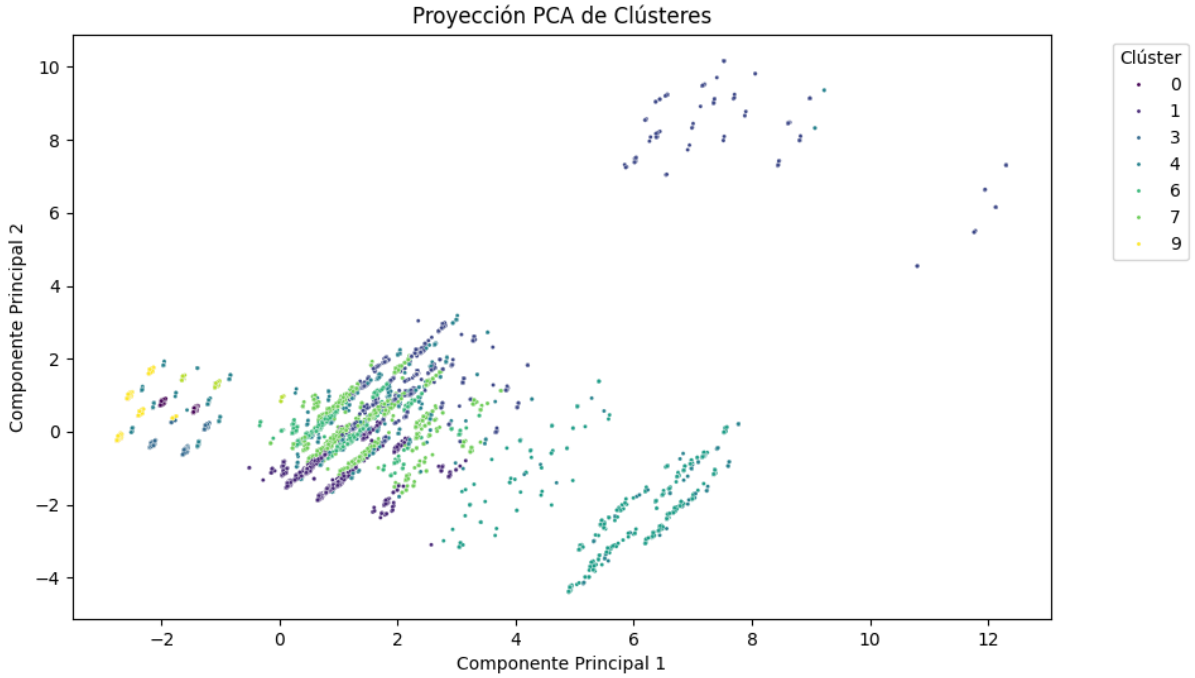
Histograma de la distribución de edades de las víctimas: El gráfico muestra la frecuencia de casos según la edad de las víctimas, indicando que la mayoría se encuentra en el rango de 20 a 40 años. Este análisis se llevó a cabo mediante el uso de herramientas de análisis de datos en el entorno Jupyter Notebook. Para más información sobre el proceso y los resultados completos, consulte el documento proyecto3.ipynb.

Clustering

El clustering o análisis de conglomerados es una técnica no supervisada de minería de datos que permite identificar grupos o clústeres en un conjunto de datos. Esta técnica es útil cuando se desea explorar patrones ocultos o estructuras internas en los datos sin tener una variable objetivo claramente definida para la predicción. En el contexto de este estudio sobre medidas aplicadas en casos de violencia intrafamiliar, se exploró el uso de clustering con el objetivo de agrupar incidentes según variables asociadas al registro institucional, leyes aplicables, organismos jurisdiccionales, entre otras, con el fin de identificar patrones relevantes que puedan ayudar en la predicción del tipo de medida aplicada.



El gráfico del método del codo muestra cómo la inercia disminuye continuamente al incrementar el número de clústeres. Sin embargo, no presenta un punto de inflexión muy definido, lo cual indica cierta dificultad para seleccionar el número óptimo de clústeres. A pesar de esto, se puede observar que aproximadamente a partir de 10 clústeres la disminución en la inercia comienza a estabilizarse. Este resultado sugiere que un número cercano a 10 clústeres podría ser adecuado para explorar la estructura subyacente del conjunto de datos.



La proyección realizada mediante Análisis de Componentes Principales (PCA) revela visualmente cómo se distribuyen los datos en función de los dos componentes principales. Se observa que existen ciertos clústeres claramente diferenciados y dispersos, especialmente los grupos en los extremos de la gráfica (por ejemplo, clústeres 0, 4 y 1). No obstante, la mayoría de los clústeres presentan una significativa superposición, especialmente en la zona central del gráfico, lo que sugiere una moderada calidad del agrupamiento. Esto coincide con el coeficiente de silueta obtenido (0.4743), indicando una separación moderada de los grupos.

Coeficiente de Silueta: 0.4743228460655563

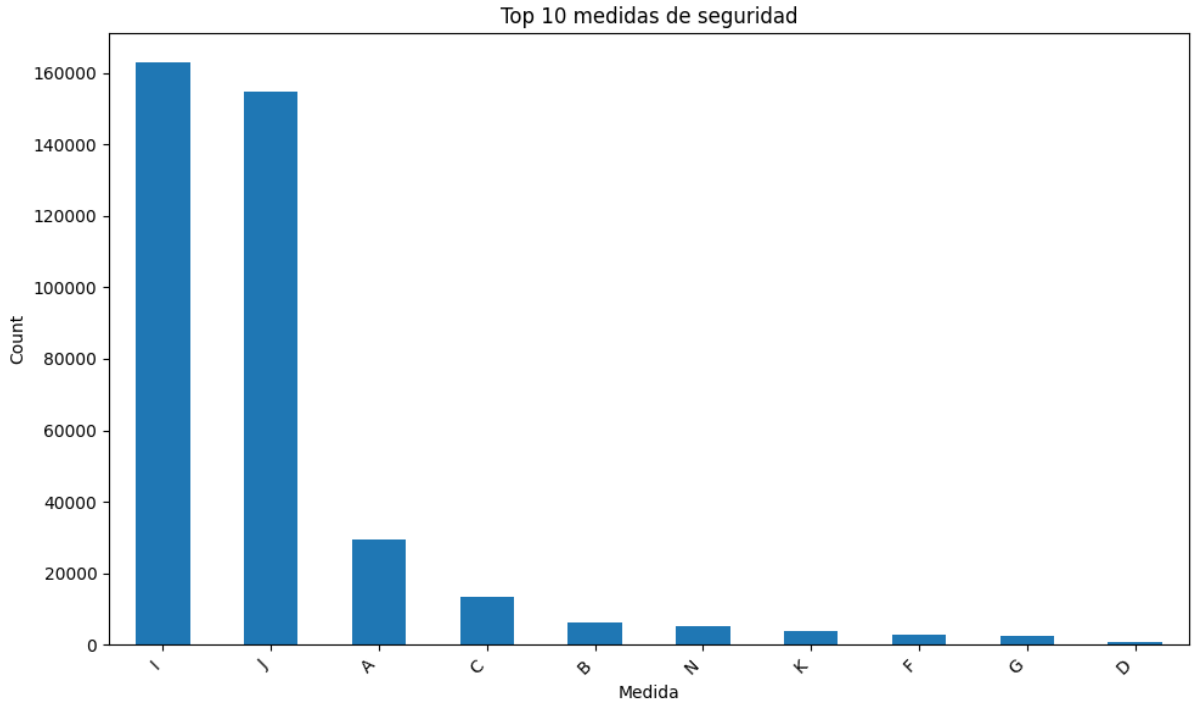
El coeficiente de silueta es una métrica que evalúa la calidad del agrupamiento, midiendo qué tan bien están separados los clústeres y qué tan cohesionados se encuentran los elementos dentro de cada clúster. El valor obtenido de 0.4743 indica una calidad moderada del clustering, sugiriendo que aunque existen algunos grupos definidos, también hay cierta ambigüedad en la asignación de datos a los clústeres, lo que limita su efectividad como método único para este análisis.

Conclusión

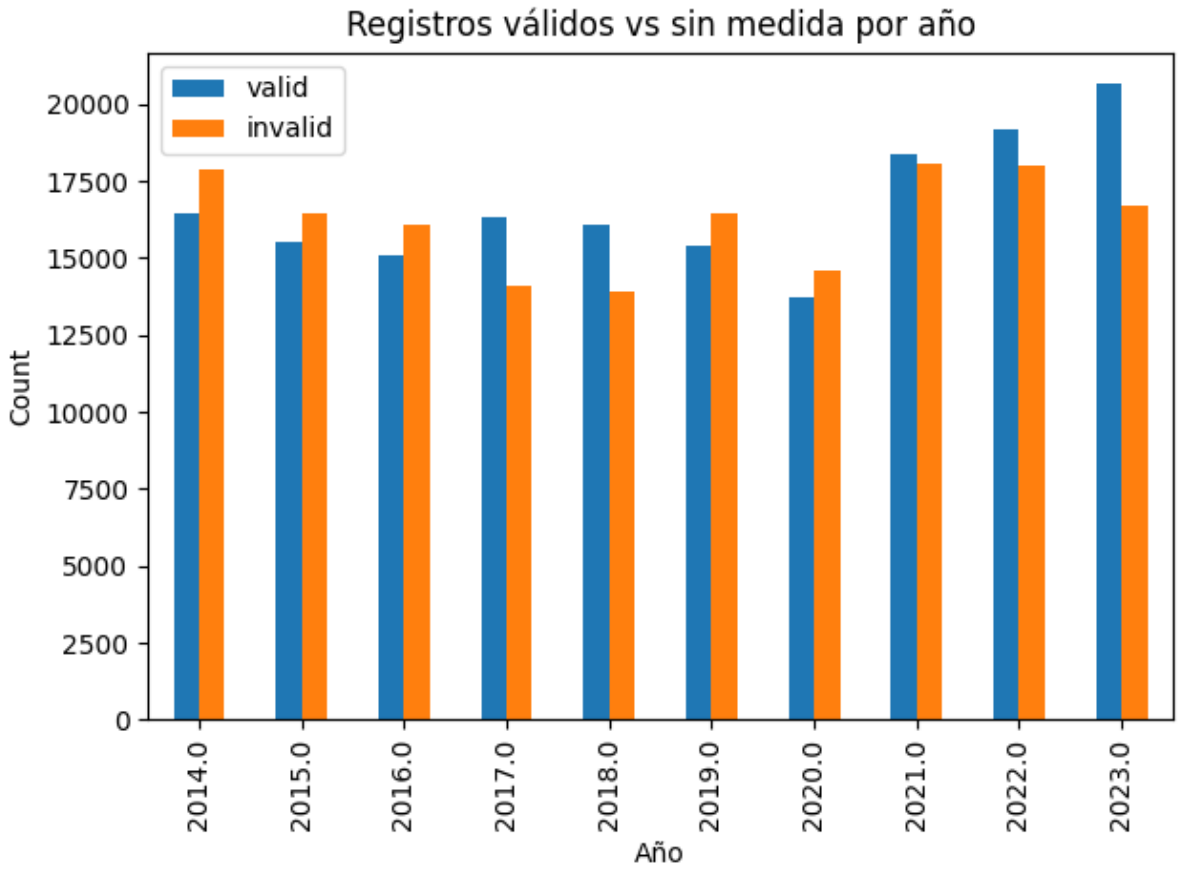
A partir del análisis realizado, se concluye que el clustering aplicado a este conjunto de datos tiene limitaciones significativas para la predicción del tipo de medida aplicada en incidentes de violencia intrafamiliar. El coeficiente de silueta obtenido (0.4743) indica que la calidad del agrupamiento es moderada, lo que refleja una separación y coherencia limitada de los grupos generados. Adicionalmente, la presencia de numerosos datos incompletos dentro de algunos clústeres disminuye considerablemente la utilidad práctica del método.

Por consiguiente, se recomienda considerar enfoques supervisados para la predicción del tipo de medida aplicada, ya que estos podrían aprovechar mejor la estructura de los datos y las relaciones específicas entre las variables estudiadas y la variable objetivo.

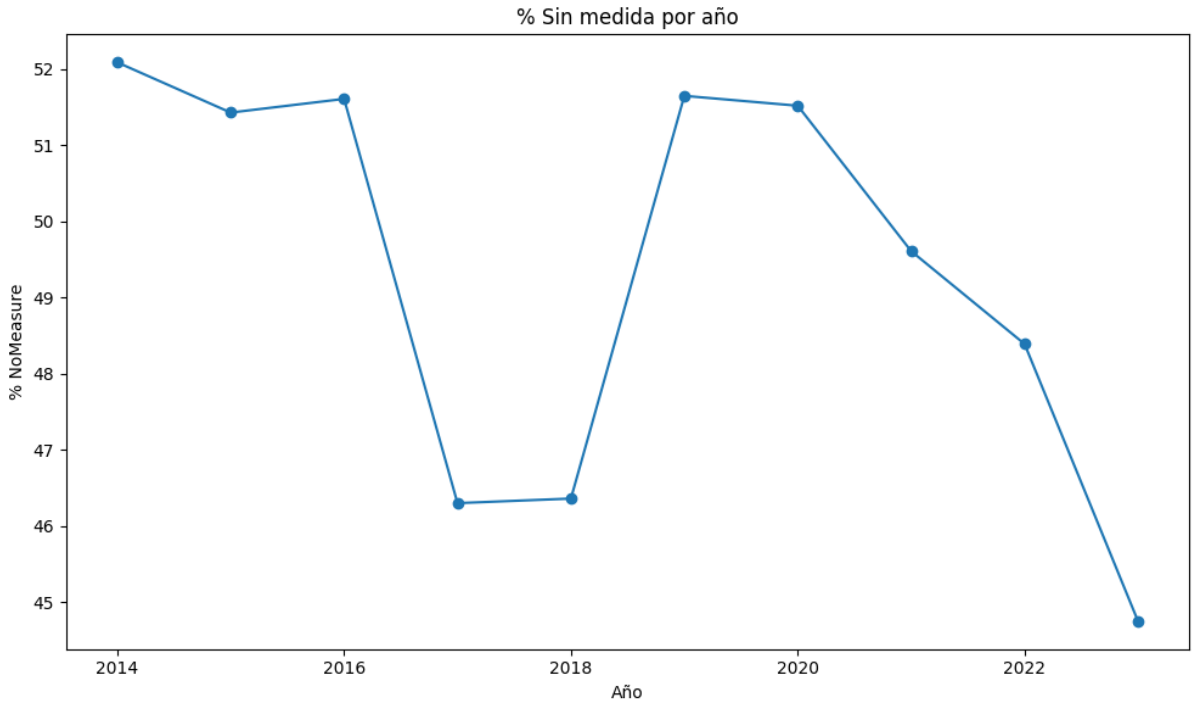
Análisis exploratorio (tipo_medida)



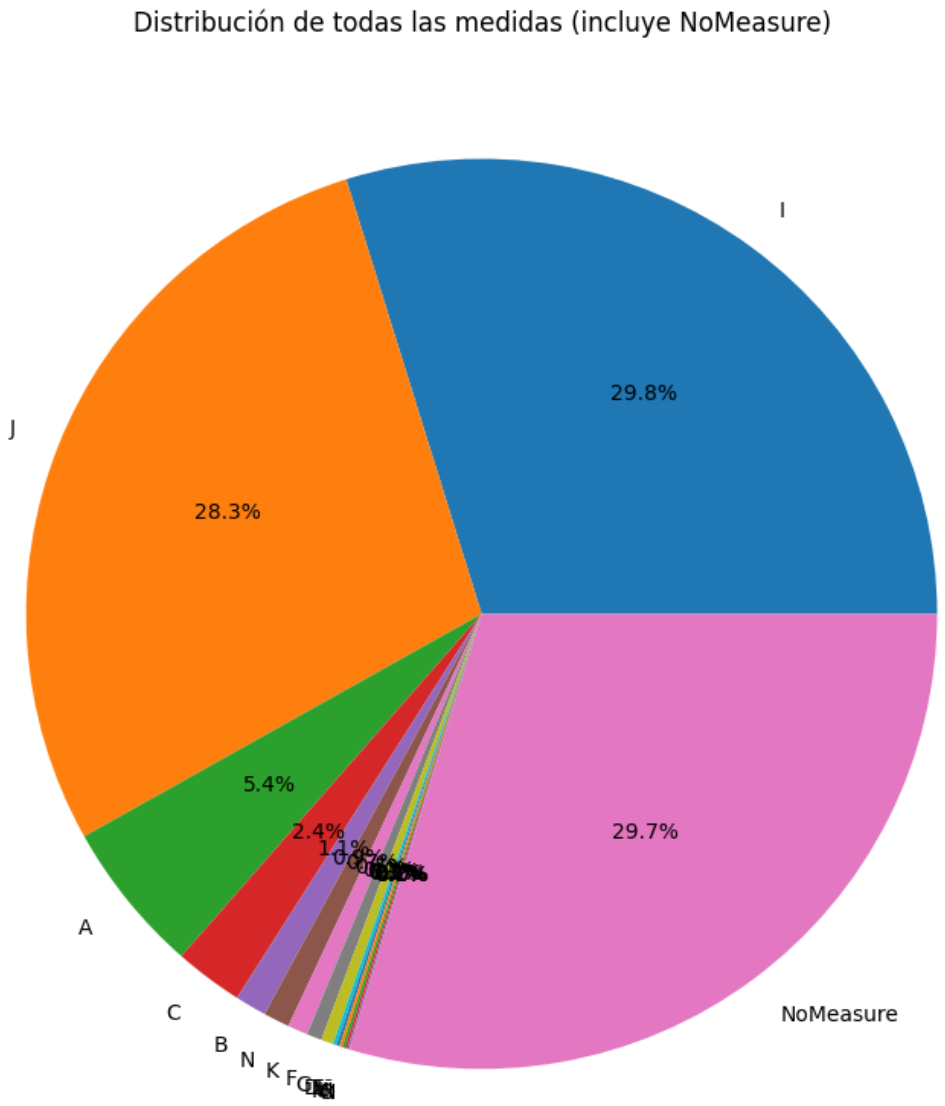
Al realizar un análisis de frecuencias sobre las distintas medidas de seguridad aplicadas en incidentes de violencia intrafamiliar, se identificó una notable concentración en pocas categorías. Destacan principalmente las medidas I (Prohibición al presunto agresor de perturbar o intimidar a cualquier integrante del grupo familiar) y J (Prohibición al agresor de acceder al domicilio o lugar de trabajo de la víctima), con un 49.52% y un 47.06% respectivamente. Es relevante destacar que prácticamente la mitad de los casos (49.31%) no cuentan con una medida asignada (NoMeasure). Otras medidas aparecen con una frecuencia significativamente menor, ninguna superando el 10% del total, siendo las más destacadas A (Orden para que el agresor abandone la residencia) con un 9.00% y C (Orden de allanamiento en caso de riesgo grave) con un 4.04%.



Al contrastar la cantidad de registros válidos (casos con medidas aplicadas) e inválidos (casos sin medidas aplicadas) a lo largo de los años, se observa cierta estabilidad en las cantidades absolutas, aunque destaca claramente que en años recientes como 2022 y 2023, la cantidad de casos con medidas aplicadas supera a aquellos sin medidas. Esta tendencia podría sugerir una mejora en la respuesta institucional frente a casos reportados de violencia intrafamiliar.



La evolución anual del porcentaje de incidentes sin medida aplicada (NoMeasure) muestra una tendencia decreciente desde el año 2014 (52.09%) hasta el 2023 (44.75%). Este descenso gradual podría indicar un fortalecimiento progresivo en la capacidad institucional para emitir medidas de seguridad, lo que representa una evolución positiva en la gestión y atención de incidentes de violencia intrafamiliar.



Principales Conclusiones

- Existe un desequilibrio evidente en la distribución de medidas de seguridad, con dos categorías dominantes (I y J) y una gran proporción de casos sin medida.
- Las categorías menos frecuentes requerirán técnicas específicas de tratamiento y balanceo de datos para asegurar que los modelos predictivos generados consideren estas etiquetas minoritarias.
- El descenso paulatino en los porcentajes de casos sin medida aplicada señala potenciales cambios positivos en los procedimientos o protocolos de actuación institucional frente a la violencia intrafamiliar.
- Es recomendable considerar variables temporales (como el año de registro) en el modelado predictivo, dado que se ha detectado una evolución clara en el comportamiento de las instituciones encargadas de emitir estas medidas.

Las mejores variables predictoras

La Información Mutua (Mutual Information, MI) es una métrica clave en la selección de variables para modelos predictivos, ya que mide la cantidad de información que una variable predictora proporciona acerca de una variable objetivo. Es especialmente útil en contextos donde las relaciones entre las variables pueden ser no lineales, complejas o categóricas, como ocurre con la predicción del tipo de medida aplicada en incidentes de violencia intrafamiliar.

En este estudio, se utilizó la Información Mutua para determinar la relevancia predictiva de diversas variables en relación con cada tipo de medida de seguridad aplicada. Los resultados destacan que las variables "institucion_registro", "ley_aplicable" y "medidas_seguridad" poseen los valores más altos de MI, oscilando entre 0.55 y 0.63. Este rango indica una relación fuerte entre estas variables y las medidas más frecuentes, específicamente las categorías I (prohibición de perturbar o intimidar), J (restricción de acceso al domicilio o lugar de trabajo de la víctima) y NoMeasure (sin medida aplicada).

Asimismo, las variables "organismo_jurisdiccional" y "organismo_remite" presentan también valores significativos de MI (entre 0.24 y 0.30), revelando una relevancia importante aunque secundaria en comparación con las primeras. Estas variables aportan información valiosa sobre el contexto institucional y jurisdiccional del caso, que puede ser determinante en la decisión sobre qué medida se aplicará.

Por otro lado, variables adicionales como "quien_reporta" y "otras_victimas_total" mostraron valores moderados de Información Mutua, especialmente útiles para identificar y predecir medidas menos frecuentes, donde la información proporcionada por estas variables es crucial debido a su especificidad.

La elección y justificación del uso de la métrica de Información Mutua en este estudio radica en su capacidad de captar y cuantificar relaciones tanto lineales como no lineales entre variables categóricas, haciendo posible identificar y seleccionar las variables más relevantes de manera objetiva y precisa. Esto permite optimizar los modelos predictivos, enfocándolos en aquellas variables que realmente tienen un impacto significativo sobre la variable objetivo.

En conclusión, la aplicación de Información Mutua proporciona una base sólida para el desarrollo de modelos predictivos robustos, al garantizar que se incluyan las variables con mayor poder informativo, mejorando así la precisión y la capacidad explicativa del modelo final.

Descripción de los datos para responder el problema planteado

A continuación, se presenta una descripción detallada de las variables consideradas más relevantes para la predicción del tipo de medida aplicada en incidentes de violencia intrafamiliar, indicando por qué cada una podría contribuir significativamente a responder el problema planteado.

- institucion_registro: Representa la institución que registra formalmente el incidente (por ejemplo, Ministerio Público o Policía Nacional Civil). Esta variable es crucial porque refleja el grado de formalización institucional del caso, pudiendo afectar directamente la probabilidad de que se apliquen ciertas medidas más severas o específicas.
- ley_aplicable: Indica la ley bajo la cual se evalúa y procesa cada incidente. Dado que diferentes leyes contemplan diferentes tipos y niveles de intervención, esta variable es fundamental para determinar las medidas que son legalmente pertinentes y por lo tanto más probables de aplicar.
- medidas_seguridad: Señala si en el caso se aplicaron o no medidas de seguridad. Esto es clave para poder determinar previamente que tipo de casos necesitaron medidas de seguridad.
- organismo_jurisdiccional: Especifica la autoridad judicial encargada de la evaluación y decisión del caso (por ejemplo, juzgado de paz o primera instancia). El tipo de organismo judicial involucrado es clave para entender la naturaleza de la intervención legal esperada y las posibles medidas que pueden ser decretadas.
- organismo_remite: Indica la institución u organismo que remitió inicialmente la denuncia o reporte del incidente. La procedencia de la denuncia puede influir notablemente en el seguimiento institucional y, en consecuencia, en las medidas específicas que se consideren necesarias y viables.
- quien_reporta: Identifica a la persona o entidad que realizó formalmente la denuncia. La identidad del denunciante puede proporcionar información crítica sobre el contexto del incidente, incluyendo posibles sesgos, relaciones familiares o institucionales, lo que puede afectar las decisiones sobre medidas aplicadas.
- otras_victimas_total: Indica el número total de víctimas adicionales involucradas en el incidente reportado. Esta variable proporciona una dimensión del alcance del daño y puede estar directamente relacionada con la severidad percibida del incidente, influyendo en la aplicación de medidas de protección más extensas.
- Tipo_agresión: La variable **tipo_agresion** representa la categoría o naturaleza específica de la agresión reportada en los casos de violencia intrafamiliar. Esta variable incluye distintas clasificaciones tales como violencia física, psicológica, sexual, patrimonial, entre otras. Es fundamental para el análisis y predicción del tipo de medida aplicada porque ofrece información directa sobre la gravedad y el contexto del incidente reportado.

Cada una de estas variables ayuda a comprender mejor el contexto específico de cada caso de violencia intrafamiliar y permite a las autoridades involucradas tomar decisiones informadas sobre qué medidas aplicar, lo que las hace indispensables para resolver eficazmente el problema planteado en el estudio.

Conclusiones

- El conjunto de datos comprende denuncias por violencia intrafamiliar registradas en Guatemala entre 2014 y 2023, abarcando diversas características demográficas, institucionales y contextuales relacionadas con víctimas y agresores.
- Todas las variables numéricas analizadas mostraron distribuciones no normales, confirmadas por la prueba de Shapiro-Wilk con valores p extremadamente bajos ($p < 0.05$), sugiriendo que no se cumplen supuestos de normalidad y reforzando el uso de técnicas robustas en análisis posteriores.
- Se identificaron correlaciones relevantes principalmente dentro de grupos específicos de variables, como atributos demográficos y relaciones familiares, reflejando coherencia interna del conjunto de datos y ofreciendo potencial para modelos predictivos.
- La mayoría de las denuncias registradas durante el periodo 2014-2023 no cuentan con una medida asignada (NoMeasure), aunque esta tendencia ha ido disminuyendo con los años, mostrando un incremento gradual en la aplicación de medidas de seguridad específicas.
- Las medidas más comúnmente aplicadas son I (prohibición al agresor de perturbar o intimidar) y J (restricción de acceso del agresor al domicilio o lugar de trabajo), representando aproximadamente el 50% cada una, mientras que otras medidas aparecen con frecuencias mucho más bajas, señalando un marcado desequilibrio en las categorías.
- Las variables institucionales (institucion_registro, ley_aplicable, medidas_seguridad, organismo_jurisdiccional, y organismo_remite) destacan como factores altamente relevantes para predecir el tipo de medida aplicada, debido a que proporcionan un contexto jurídico e institucional claro para cada caso.
- El tipo específico de agresión (tipo_agresion) también es relevante, ya que determina en gran medida la severidad y urgencia en la asignación de medidas de protección.
- Las variables quien_reporta y otras_victimas_total, aunque secundarias en términos generales, ofrecen información valiosa especialmente en la predicción de medidas menos frecuentes, destacando la importancia del contexto social y familiar en la aplicación de medidas de seguridad.
- El análisis de clustering mostró limitaciones claras, con una calidad moderada (coeficiente de silueta de 0.4743), indicando que este método no es el más adecuado para predecir el tipo de medida aplicada, debido a la complejidad y falta de claridad en la separación de los grupos obtenidos.
- Finalmente, para mejorar la precisión y relevancia del modelo predictivo, se recomienda utilizar técnicas de aprendizaje supervisado, incorporando métodos de balanceo para tratar el marcado desequilibrio de las categorías de tipo_medida. Además, considerar la variable temporal anio_registro puede ayudar a capturar tendencias en la aplicación de medidas a lo largo del tiempo.