Bogotá, 5 de julio de 2023

ANÁLISIS DE FACTORES Y ALERTA TEMPRANA DEL RIESGO DE VBG EN COLOMBIA

DAVID SAMUEL BARRERA (ID 8971643)
ELIANA POVEDA AGUIRRE (ID 8972747)

Nota de Aceptación

Certificamos que el presente Trabajo de Grado Satisface, en alcances y calidad, todos los requisitos que demanda un Trabajo de Grado de Maestría.

David Arango Londoño

David Arango Londoño 1130586950 de Cali

Luis Eduardo Tobón

Valentina Corchuelo

Aprobado en cumplimiento de los requisitos exigidos por la Pontificia Universidad Javeriana Cali, para optar el título de Magister en ciencias de datos.

Camilo Rocha

HERNÁN CAMILO ROCHA NIÑO Ph. D.
Decano Facultad de Ingeniería y Ciencias

JUAN CARLOS MARTÍNEZ ARIAS
Director Posgrados de Ingeniería y Ciencias

Acta de Correcciones al Documento de Trabajo de Grado

Santiago de Cali, julio 6 de 2023

Autores: DAVID SAMUEL BARRERA (ID 8971643), ELIANA POVEDA AGUIRRE (ID 8972747)

Título del Trabajo de Grado: "ANÁLISIS DE FACTORES Y ALERTA TEMPRANA DEL RIESGO DE VBG EN COLOMBIA"

Director:

Como indica el artículo 2.13 de las Directrices para Trabajo de Grado de Maestría, he verificado que el estudiante indicado arriba ha implementado todas las correcciones que los Jurados del Proyecto de Trabajo de Grado definieron que se efectuaran, como consta en el Acta de Evaluación correspondiente.

David Arango Londoño

_____

Firma del Director del Trabajo de Grado

David Arango Londoño 1130586950 de Cali

PROYECTO DE GRADO MAESTRÍA EN CIENCIA DE DATOS.
ANÁLISIS DE FACTORES Y ALERTA TEMPRANA DEL RIESGO DE VBG EN COLOMBIA

DAVID SAMUEL BARRERA (ID 8971643)
ELIANA POVEDA AGUIRRE (ID 8972747)

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director
David Arango Londoño

FACULTAD DE INGENIERÍA Y CIENCIAS MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, MAYO DE 2023

SANTIAGO DE CALI, 29 DE MAYO DE 2023

Ingeniero:
Juan Carlos Martínez Arias
Director Posgrados de Ingeniería
Facultad de Ingeniería y Ciencias
Pontificia Universidad Javeriana – Cali

Con el fin de cumplir con los requisitos exigidos por la Universidad para llevar a cabo el Trabajo de Grado y posteriormente optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el proyecto de Trabajo de Grado denominado " Análisis de factores y alerta temprana del riesgo de Violencia Basada en Género en Colombia", el cual fue realizado por los estudiantes Eliana Poveda Aguirre (ID 8972747) y David Samuel Barrera (ID 8971643**)** con códigos, bajo la dirección del profesor David Arango.

El suscrito director del Trabajo de Grado autoriza para que se proceda a hacer la evaluación de este Proyecto ante el Tribunal que para el efecto se designe, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado oficialmente.

Atentamente,

Director David Arango Londoño
C.C. 1130586950 de Cali

Eliana Liney Poveda Aguirre
Estudiante
CC 1026563306 de Bogotá

David Samuel Barrera
Estudiante
CC 1070982847 de Facatativá

FICHA RESUMEN PROYECTO DE TRABAJO DE GRADO

-TÍTULO: Análisis de factores y alerta temprana del riesgo de Violencia Basada en Género en Colombia
-ÁREA DE TRABAJO: Aplicado al sector gubernamental
-TIPO DE PROYECTO (Aplicado, Innovación, Investigación): Aplicado
-ESTUDIANTE(S): David Samuel Barrera, Eliana Poveda Aguirre
-CORREO ELECTRÓNICO: eliana1990@javerianacali.edu.co; dsbarrerab@javerianacali.edu.co
-DIRECCIÓN Y TELÉFONO:
Calle 52 # 14-52 APTO 401- 3176483198
Calle 15 # 119a - 60 Torre 14, APT 102 - 3046430297
-DIRECTOR: David Arango Londoño
-VINCULACIÓN DEL DIRECTOR: Planta 8.
-**CORREO ELECTRÓNICO DEL DIRECTOR**: David.arango@javerianacali.edu.co
-**CO-DIRECTOR** (Si aplica): N/A
-**GRUPO O EMPRESA QUE LO AVALA** (Si aplica): N/A
-OTROS GRUPOS O EMPRESAS: N/A
-**PALABRAS CLAVE** (al menos 5): VBG, violencia de pareja, violencia, sexual, Sociodemográfico, Analítica, Factores, clasificación, K-means, agrupación.
FECHA DE INICIO: 11 de junio del 2022
-RESUMEN:
La violencia física, psicológica, sexual y económica contra las mujeres hacen parte de las distintas formas de violencia basada en género (VBG). En consecuencia, con el presente proyecto se creó, tentativamente, un modelo de aprendizaje no supervisado que permitió identificar los determinantes que inciden en la VBG y, con ello, visibilizar el uso de herramientas de machine learning para la comprensión de este fenómeno a nivel nacional. Conocer dónde se concentra, por qué, y en qué casos se incrementa la violencia de pareja y sexual es relevante para la prevención y, en particular, para la planificación de los recursos y servicios institucionales implicados en la lucha contra la VBG, especialmente de intervención temprana.

TABLA DE CONTENIDO

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE GRÁFICAS

ÍNDICE DE TABLAS

1.INTRODUCCIÓN

El aprendizaje no supervisado es una técnica en el campo del machine learning que se utiliza para analizar grandes conjuntos de datos sin la necesidad de etiquetas o información previa. A diferencia del aprendizaje supervisado, donde se proporcionan datos etiquetados para entrenar un modelo, el aprendizaje no supervisado busca descubrir patrones, tendencias y estructuras ocultas en los datos de manera automática.

En el contexto de este proyecto, el aprendizaje no supervisado se emplea para identificar los determinantes de la violencia basada en género (VBG) en dos de sus expresiones, la violencia sexual y de pareja, a partir de los datos del Instituto de Medicina Legal y Ciencias forenses durante los años 2019-2021 en Colombia. Lo anterior, debido a que el machine learning ha demostrado ser una herramienta útil para prevenir y conocer los determinantes de este tipo de violencia, analizar patrones de comportamiento o identificar las características de los agresores y las víctimas.

El proyecto se enfocó en explorar las técnicas de aprendizaje no supervisado a partir de las bases de datos de violencia sexual y de pareja, permitiendo identificar sus determinantes principales. Adicionalmente, mediante el proyecto se buscó identificar las variables más apropiadas de las mencionadas bases de datos para la creación óptima del modelo; establecer patrones, clasificaciones o agrupaciones de los datos relacionados con la ocurrencia de la violencia sexual y de pareja y, por último, generar conclusiones a partir de la implementación del modelo.

En este documento, se presenta el problema, un resumen detallado de los objetivos del proyecto y el problema identificado. Adicionalmente, en un primer capítulo se contextualizan las bases de datos, su transformación y selección de variables, además del análisis exploratorio de las mismas. En un segundo capítulo se exploran las distintas técnicas aplicadas a las bases de datos en mención. En un tercer capítulo se aplica la librería de Factoclass de R a las bases de datos, que agrupa mediante su aplicación el análisis factorial, la técnica de K-means y de clasificación jerárquica, permitiendo así la identificación de agrupaciones e identificación de tendencias de violencia sexual y de pareja. Finalmente, en un cuarto capítulo se presentan conclusiones frente a los patrones y determinantes de la violencia sexual y de pareja en Colombia en los años en estudio.

Este proyecto tiene un gran potencial para demostrar que el machine learning puede ser utilizado para analizar grandes cantidades de datos y detectar patrones que permitan identificar tempranamente situaciones de riesgo en los casos de violencia sexual y de pareja, como analizar patrones u/o características de los agresores o víctimas o identificar áreas geográficas donde hay una alta incidencia de la violencia de género.

2. DEFINICIÓN DEL PROBLEMA

La Violencia Basada en Género, en adelante (VBG), se define como cualquier acto con el que se busque dañar a una persona por su género, puede ser de tipo sexual, físico, psicológico o económico. La violencia de género puede tomar muchas formas: violencia de pareja, sexual, psicológica, matrimonio infantil, mutilación genital femenina, entre otras formas [1]. Según el Observatorio de Medicina Legal, en el año 2021 se registraron en Colombia 55.582 casos de violencia basada en género, representados en 106 feminicidios, 21.434 casos de violencia sexual y 34.042 de violencia de pareja. Estos casos de VBG representaron un incremento del 19% con relación a los casos del año 2020, donde se registraron 44.614 casos entre feminicidios (90), violencia sexual (18.054) y violencia de pareja (26.470) [2]. Si bien a nivel nacional se ha avanzado en la producción de estadísticas sobre la VBG, organizaciones como la CEPAL han advertido sobre la necesidad de fortalecer los sistemas de registro a nivel nacional y estandarizar la información para contar con mejores datos, analizar las características de las distintas formas de violencia a nivel nacional y mejorar la comparabilidad regional [3]. Bajo este panorama, es necesario contar con un análisis de datos, basado en técnicas de machine learning, que contribuya a entender mejor la ocurrencia de la VBG y sus determinantes (causas o factores contextuales amplios asociados a la VBG), en particular, en sus dos tipos de VBG más frecuente (la violencia de pareja y la violencia sexual contra las mujeres).

2.1 PLANTEAMIENTO DEL PROBLEMA

En el año 2021, según los registros de Medicina Legal, se reportaron 51.610 casos de violencia intrafamiliar, en 40.058 de los casos la víctima fue una mujer, lo que significó un aumento del 10 por ciento frente a las 36.399 mujeres víctimas que hubo en 2020 (de un total de 47.177 casos). Así mismo, a enero de 2022, 2.144 mujeres habían sido agredidas por su pareja [4]. Frente a la violencia sexual, según datos del Sistema de Información Estadístico, Delincuencial, Contravencional y Operativo de la Policía Nacional (SIEDCO), entre enero de 2019 y enero del 2022 se denunciaron 97.237 casos de violencia sexual. De ese total, el 85% (82.873) se cometieron contra mujeres, mientras que el 15% (14.364) contra hombres. [5]. Ante este panorama, los datos se han convertido en un insumo de valor para definir, categorizar, interpretar y hacer predicciones sobre la ocurrencia de la VBG, en particular, contra la violencia de pareja y la violencia sexual [6]. En Colombia, aunque distintas organizaciones de mujeres han unido esfuerzos para consolidar e interpretar los datos que proporcionan las organizaciones estatales sobre estas victimizaciones, no se cuenta aún con una herramienta que, por ejemplo, permita identificar los determinantes que inciden en la VBG (violencia de pareja y violencia sexual). Lo anterior es relevante para la prevención y, en particular, para la planificación de los recursos y servicios institucionales implicados en la lucha contra la VBG contra las mujeres, especialmente de intervención temprana.

2.2 FORMULACIÓN DEL PROBLEMA

¿Cómo utilizar el aprendizaje no supervisado para identificar los determinantes que inciden en la violencia basada en género (VBG), violencia sexual y de pareja a partir de los datos del Instituto de Medicina Legal y Ciencias forenses durante los años 2019-2021 en Colombia?

Subproblemas:

- ¿Qué patrones asociados a la VBG (violencia sexual y de pareja) es posible identificar por medio de las técnicas asociadas al aprendizaje no supervisado?
- ¿Cómo identificar las técnicas de aprendizaje no supervisado más idóneas para el análisis de las bases de datos de VBG (Violencia de pareja y violencia sexual)
- ¿Qué variables inciden mayormente en la ocurrencia de la VBG – violencia sexual y de pareja?
- ¿Existe alguna correlación entre el ciclo vital, escolaridad, pertenencia étnica, circunstancias del hecho y la violencia sexual y de pareja?
- ¿Existe alguna correlación entre el presunto agresor, las circunstancias y zona del hecho y la violencia sexual y de pareja?

3. OBJETIVOS DEL PROYECTO

-Crear un modelo de aprendizaje no supervisado que permita identificar los determinantes que inciden en la VBG a nivel nacional, en particular, en dos de sus expresiones más frecuentes (violencia de pareja y sexual), durante los años 2019-2021 a partir de los datos del Instituto de Medicina Legal y ciencias forenses.

3.1 OBJETIVOS ESPECÍFICOS

--Determinar las variables más apropiadas de las bases de datos de violencia sexual y de pareja para la creación óptima del modelo de aprendizaje no supervisado.

-Identificar las técnicas más apropiadas de aprendizaje no supervisado, tales como K-means, K-medoides, clusterización jerarquica o Density Based Scan Clustering (DBSCAN) que permitan la agrupación y clasificación de los datos de violencia de pareja y violencia sexual.

-Establecer patrones, clasificaciones o agrupaciones de los datos relacionados con la ocurrencia de la violencia sexual y de pareja a partir de técnicas de aprendizaje no supervisado.

-Generar conclusiones sobre los determinantes que inciden en la VBG en Colombia (violencia sexual y de pareja) a partir de la implementación del modelo de aprendizaje no supervisado propuesto.

4.MARCO TEORICO

Según la ONU, se entiende por violencia de género (VBG) a los actos perjudiciales dirigidos hacia una persona o grupo debido a su género. El propósito de este término es resaltar que las desigualdades de poder basadas en el género exponen a las mujeres y niñas a diversos tipos de violencia, colocándolas en situaciones de riesgo [7]. Según Tibada et al, la violencia de género se refiere a la violencia experimentada por las mujeres, la cual tiene sus fundamentos en la discriminación histórica y la falta de derechos que han padecido y siguen padeciendo en diversas regiones del mundo. Esta violencia se basa en una construcción cultural relacionada con el género [8].

En esta perspectiva, la violencia de género abarca cualquier acción o falta intencionalmente agresiva que se fundamenta en relaciones de poder desiguales, donde el hombre o la figura masculina adopta el control según el sistema patriarcal, afectando de manera directa e indirecta al género femenino, especialmente en el ámbito familiar. Esta violencia causa daños de naturaleza física, psicológica, sexual, emocional, vicaria, económica y patrimonial [9].

De acuerdo con el Ministerio de Salud, la violencia basada en el género puede manifestarse a través de diversas formas, como la violencia intrafamiliar o doméstica, la violencia en parejas o conyugal, el maltrato infantil y diversas formas de violencia sexual. Esta violencia puede presentarse de manera sutil, como comentarios irrespetuosos o chistes hacia las mujeres, así como maltrato psicológico y agresión por parte de autoridades, en entornos educativos o laborales, y otros espacios de socialización. También puede tomar formas más graves, como la violencia física, y llegar a casos de acoso sexual, explotación, trata de mujeres, violación sexual y la utilización del cuerpo femenino como un territorio de guerra en contextos de conflicto armado [10].

En Colombia la Ley 1257 de 2008 (por la cual se dictan normas de sensibilización, prevención y sanción de formas de violencia y discriminación contra las mujeres) define la violencia contra la mujer como: “cualquier acción u omisión, que le cause muerte, daño o sufrimiento físico, sexual, psicológico, económico o patrimonial por su condición de mujer […] en el ámbito público o en el privado.” [11].

Partiendo de estas definiciones, es necesario tener en consideración que la violencia de género continúa afectando a miles de mujeres y niñas cada año en América Latina y el Caribe. De acuerdo con datos de la CEPAL, a partir de encuestas nacionales de seis países de la región, entre el 60% y el 76% de las mujeres (alrededor de 2 de cada 3) ha sido víctima de violencia por razones de género en distintos ámbitos de su vida. Además, en promedio 1 de cada 3 mujeres ha sido víctima o vive violencia física, psicológica y/o sexual, por un perpetrador que era o es su pareja, lo que conlleva el riesgo de la violencia letal: el feminicidio o femicidio [12].

En particular, en Colombia, Según el Observatorio de Medicina Legal, en el año 2021 se registraron 55.582 casos de violencia basada en género, representados en 106 feminicidios, 21.434 casos de violencia sexual y 34.042 de violencia de pareja. Estos casos de VBG representan un incremento del 19% con relación a los casos del año 2020, donde se registraron 44.614 casos entre feminicidios (90), violencia sexual (18.054) y violencia de pareja (26.470) [13]. Adicionalmente, para abril de 2022, según el Observatorio colombiano de feminicidios, se habían registrado en el país 62 feminicidios, en 16 departamentos del país, es decir, en el 53% del territorio nacional [14].

Ahora bien, durante la pandemia del Covid-19, según Sisma Mujer, aumentó el riesgo de feminicidio de las mujeres en el país. De hecho, de enero a julio de 2020 se realizaron 2.072 valoraciones del riesgo de violencia mortal contra mujeres por parte de su pareja o expareja. En el 2020 con corte a julio, se presentó un incremento de 6 puntos porcentuales en las valoraciones clasificadas con riesgo extremo de feminicidio [15]. Lo anterior debido a que el confinamiento obligó a las mujeres a estar encerradas con sus maltratadores y el encierro hizo que se incremente el riesgo de violencia contra ellas en la medida en que aumentó el tiempo de convivencia con sus agresores.

Ante el complejo panorama de incremento en casos de violencia de género contra mujeres han surgido diversos esfuerzos desde la sociedad civil y del Estado para cuantificar el fenómeno y con ello contar con datos precisos que permiten medir la incidencia delictiva de feminicidios con base en datos públicos. Adicionalmente, mediante el uso del Big Data e inteligencia artificial se han buscado crear herramientas para prevenir la ocurrencia de feminicidios e, incluso, para focalizar acciones de prevención en materia de violencia de género.

Así, por ejemplo, organizaciones como la Red Latinoamericana contra la Violencia de Género creó una herramienta para monitorear, visibilizar y erradicar la violencia de género en América Latina, a través del mapa latinoamericano de feminicidios que condensa datos de feminicidios en Argentina, Chile, Colombia Ecuador Panamá, Puerto Rico y Uruguay [16].

A nivel latinoamericano y el Caribe, se han creado también otros esfuerzos para cuantificar el fenómeno y contar con sistemas de registro de VBG con información estandarizada. Así, por ejemplo, en México se conformó Data Cívica, un proyecto mediante el cual se utilizan modelos estadísticos para medir la VBG en México. Este proyecto tuvo la finalidad de conocer qué factores sociodemográficos están relacionados con la violencia de género en México y poder ofrecer posibles explicaciones. Para ello se estimó un modelo de Regresión Binomial Negativa (NB por sus siglas en inglés) con datos a nivel municipal durante los últimos tres sexenios en México (Felipe Calderón Hinojosa, Enrique Peña Nieto y Andrés Manuel López Obrador). El objeto de este modelo es analizar la relación existente entre los factores sociodemográficos de un municipio en determinado sexenio y la violencia letal contra las mujeres [17].

En Argentina, por ejemplo, se creó el sistema de Alertas de Correo de Datos Contra Feminicidio, consistente en la creación de un algoritmo de aprendizaje automatizado y alertas de correo para informar sobre posibles casos de feminicidio. El sistema busca noticias en la base de datos de MediaCloud según los términos de búsqueda y la región determinados. El sistema filtra los resultados a través del algoritmo que fue específicamente entrenado para calcular la probabilidad de que un artículo refiera a un caso de feminicidio y para agrupar artículos de distintas fuentes que refieren al mismo caso. Finalmente, el sistema envía un mensaje de correo con los artículos relevantes, según la frecuencia determinada por cada usuario [18].

También, en República Dominicana, por ejemplo, las autoridades desarrollaron una herramienta denominada “Eagle eyes”, una plataforma de análisis de datos e inteligencia artificial que genera un modelo en base en la ocurrencia de feminicidios. El modelo predice la ocurrencia de feminicidios generando un escenario predictivo con variables tan determinantes como el municipio, hora o factores económicos, entre otros. Así mismo, identifica los meses y municipios donde más feminicidios se podrían estar registrando [19].

En el caso de Colombia, la Ley 1761 de 2015, dispuso la creación y puesta en marcha del Sistema Integrado de Información de Violencias basadas en Género (SIVIGE) a cargo del DANE, Medicina Legal y el Ministerio de Justicia, por el cual se obliga a dichas entidades la adopción de un Sistema Nacional de Estadísticas sobre Violencia Basada en Género [20].

En el país, por ejemplo, la Secretaría de la Mujer de Bogotá cuenta con el Observatorio de Mujeres y Equidad de Género de Bogotá (OMEG), que recopila la información más relevante sobre la violencia contra las mujeres en Bogotá, entre ellas la violencia de pareja y la violencia sexual. En particular el sistema integra información proveniente de diversas fuentes: Instituto Nacional de Medicina Legal, SIEDCO (Policía Nacional), y SIMISIONAL (Secretaría Distrital de La Mujer) [21].

Aunque en la literatura no son abundantes las experiencias de la aplicación del machine learning en temas como la VBG y el feminicidio en Colombia, se encuentran algunas iniciativas como la desarrollada por la Secretaria Distrital de la Mujer y Fedesarrollo mediante la cual se realizó un análisis de Big Data, específicamente de conversaciones abiertas de Twitter, empleando técnicas de Natural Language Processing (NLP) mediante un modelo denominado Latent Dirichlet Allocation (LDA) para analizar la percepción sobre VGB antes y después de la cuarentena. De esta manera, se analizaron dos millones de conversaciones en el contexto de Colombia en Twitter entre el primero de enero y el 30 de julio de 2020 para ver los cambios entre antes y después de la cuarentena por el Covid-19 [22].

4.1. MACHINE LEARNING Y SU APLICACIÓN

En el contexto de la aplicación del machine learning, se define como una subcategoría de inteligencia artificial que se basa en permitir que los algoritmos descubran patrones recurrentes en conjuntos de datos, ya sean números, palabras, imágenes, estadísticas u otros tipos de información [23]. En resumen, los algoritmos de machine learning aprenden de manera autónoma a realizar tareas o hacer predicciones a partir de datos, y mejoran su rendimiento a medida que adquieren experiencia. Una vez entrenados, estos algoritmos pueden identificar patrones en nuevos conjuntos de datos [24].

Dentro del ámbito del machine learning, se emplean técnicas como el aprendizaje supervisado y no supervisado. En el aprendizaje supervisado, se trabaja con conjuntos de datos que ya están etiquetados, lo que significa que se conoce el valor del atributo objetivo para cada dato en el conjunto. Por otro lado, en el aprendizaje no supervisado, se utilizan datos que no han sido previamente etiquetados [25].

Para este proyecto en particular, como se mencionó anteriormente, se emplearon técnicas de aprendizaje no supervisado. Estas técnicas se basan en algoritmos que se entrenan utilizando conjuntos de datos sin etiquetas o clases predefinidas. En otras palabras, no se conoce de antemano ningún valor objetivo o clase, ya sea categórico o numérico. El aprendizaje no supervisado se enfoca en tareas de agrupamiento, también conocidas como clustering o segmentación, cuyo objetivo es encontrar grupos similares dentro del conjunto de datos [26]. Estas técnicas se dividen en dos categorías principales: (I) Métodos Jerárquicos: Estos algoritmos no requieren que el usuario especifique de antemano el número de clústeres y (ii) Métodos No Jerárquicos: estos algoritmos requieren que el usuario especifique previamente el número de clústeres que se crearán, como por ejemplo K-means, K-medoids y CLARA. Además, existen otros métodos que combinan o modifican las técnicas mencionadas anteriormente, como hierarchical K-means, fuzzy clustering, model-based clustering y density-based clustering [27].

4.2. MÉTODOS DE CLASIFICACIÓN NO JERÁRQUICOS

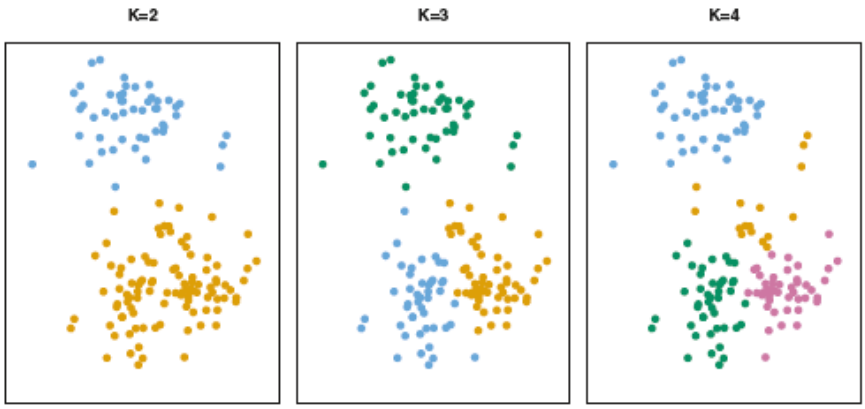
- **K-MEANS**

Este tipo de algoritmos, según la literatura revisada, tienen como objetivo principal dividir un conjunto de n observaciones en k grupos distintos. Cada grupo se representa mediante el promedio de los puntos que lo conforman, conocido como centroide. El número de grupos, k, se define de antemano como un parámetro. El proceso de clustering inicia con la ubicación aleatoria de k centroides y asigna cada observación al centroide más cercano. Luego de asignar las observaciones, los centroides se desplazan hacia la posición promedio de todos los datos asignados a ellos, y se vuelven a asignar los puntos según las nuevas ubicaciones de los centroides [28].

El objetivo del algoritmo K-means es agrupar las observaciones de tal manera que las que pertenezcan al mismo grupo sean lo más similares entre sí, mientras que las pertenecientes a grupos diferentes sean lo más diferentes posibles. Para lograr esto, se utilizan medidas de distancia, como la euclidiana, para evaluar la similitud y la diferencia. Una medida utilizada para evaluar qué tan bien los centroides representan a los miembros de su grupo es la suma de los errores al cuadrado. En cada iteración, el algoritmo K-means busca reducir el valor de la suma de los errores al cuadrado [29].

En suma, k-means funciona primero a) seleccionando k centroides iniciales, b) asignando cada muestra al centroide más cercano, c) recalculando el centroide para cada grupo de muestras asignadas, d) repitiendo los pasos b y c hasta que los centroides dejen de cambiar de manera significativa. En notación matemática, esto se puede escribir de la siguiente manera: Sea $X = \{x1, x2, \ldots, xn\}$ un conjunto de n observaciones y *k* el número de grupos que se quieren formar. Sea $C = \{c1, c2, \ldots, ck\}$ un conjunto de *k* centroides. El objetivo del algoritmo k-means es entonces encontrar una partición de *X* en *k* conjuntos: $\mathrm{S} = \{\mathrm{S1}, \mathrm{S2}, \ldots, \mathrm{Sk}\}$ de tal forma que se minimice la suma de los cuadrados de las distancias entre cada observación xi y su centroide k correspondiente: $\min \mathrm{S} \Sigma \mathrm{i} == 1\mathrm{k} \Sigma \mathrm{x} \in S\mathrm{i} \left|\left|\mathrm{x} - \mathrm{ck}\right|\right|^2$ [30].

Gráfica 1. Agrupación mediante técnica de K-means



Fuente: Facultad de Informática. UNLP [31].

En particular, la técnica de K-means ha sido utilizada, por ejemplo, para el análisis de la violencia intrafamiliar en el departamento del Atlántico. En este caso la técnica fue utilizada para comprender, por ejemplo, los grupos de municipios en los que se concentró mayormente este fenómeno[32], o el India donde, a través del análisis de conglomerados de k- means y el preprocesamiento del conjunto de datos sobre delitos contra las mujeres en ese país, fue posible identificar las regiones de Maharashtra, Tamil Nadu, Bengala Occidental, como aquellas con mayor número de casos de delitos cometidos contra mujeres en comparación con otros estados de la India. En este estudio, los clústeres agrupaban los delitos de violencia sexual, secuestro, trafico sexual, entre otros, en relación con las regiones de la India [33].

Otro análisis llevado a cabo en Perú identificó, a través de las técnicas K-means, que la celopatía y el deseo sexual patológico, son dos de las principales motivaciones para la ocurrencia de la violencia de género contra las mujeres. Para esta investigación se utilizaron datos provenientes del "Censo Nacional Anual de la Población Penitenciaria en el Perú", en 66 establecimientos penitenciarios a nivel nacional. En la experimentación de la búsqueda de patrones a través del algoritmo K-means, se reconocieron 2 grupos que definen las características cercanas involucradas con la motivación para perpetrar el delito: la celopatía y la motivación sexual [34].

- **K-MEDOIDES**

Según Correa, K-medoids es un método de agrupamiento que, al igual que K-means, organiza las observaciones en K clústeres, donde K es un valor predefinido por el usuario. En este método, cada clúster está representado por una observación en particular, conocida como "medoide", que es el elemento central del clúster y se considera el más representativo. El medoide se elige de manera que su distancia promedio con todos los demás elementos del mismo clúster sea mínima. A diferencia de K-means, en K-medoids se enfoca en la observación del medoide en lugar

de utilizar centroides. En K-means, cada clúster está representado por un centroide que corresponde al promedio de todas las observaciones del clúster, sin destacar ninguna observación en particular. La utilización de medoides en lugar de centroides hace que K-medoids sea un método más resistente a valores atípicos o ruido, lo que lo hace más robusto que K-means [35].

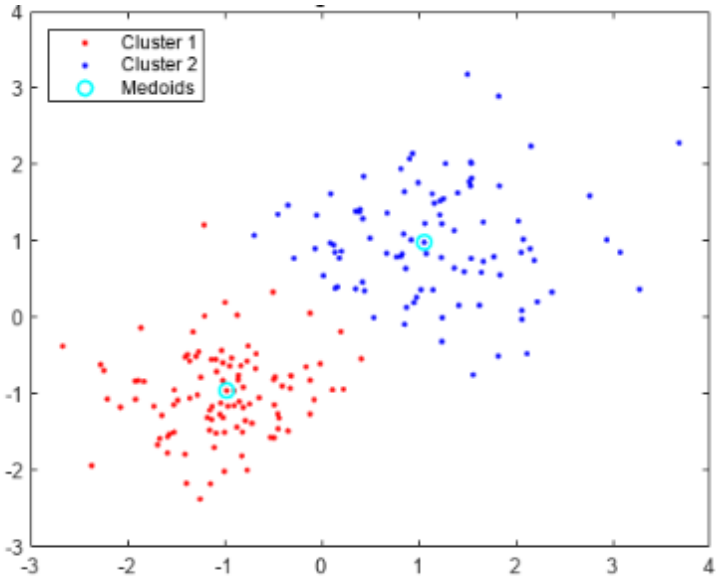
En suma, K-medoids funciona primero a) seleccionando k medoides iniciales, b) asignando cada muestra al medoide más cercano, c) recalculando el medoide para cada grupo de muestras asignadas, d) repitiendo los pasos b y c hasta que los medoides dejen de cambiar de manera significativa. En notación matemática, esto se puede escribir de la siguiente manera:

Sea $\mathrm{X} = \{x1,\ x2,\ \ldots,\ xn\}$ un conjunto de *n* observaciones y *k* el número de grupos que se quieren formar. Sea $M = \{m1, m2, \ldots, mk\}$ un conjunto de *k* medoides. El objetivo del algoritmo k-medoids es encontrar una partición de *X* en *k* conjuntos $S = \{S1, S2, \ldots, Sk\}$ de tal forma que se minimice la suma de las distancias entre cada observación *xi* y su medoide correspondiente:

$$\min \mathrm{S}\,\Sigma\mathrm{i} = 1\mathrm{k}\,\Sigma\mathrm{x} \in S\mathrm{i}\ \mathrm{dist}(\mathrm{x}, \mathrm{mi})$$

donde $dist(x, mi)$ es la distancia entre la observación *x* y el medoide *mi* [36].

Gráfica 2. Agrupación mediante técnica de K-medoides



Fuente: Mathworks [37]

- **CLARA**

El algoritmo CLARA nace ante la necesidad de superar las barreras de memoria y tiempo de cómputo del algoritmo k-medoides (también conocido como Partitioning Around Medoids PAM). El método consiste, en términos generales, de dos pasos. Primero, se obtiene una muestra de objetos de los cuales se generan k grupos usando el algoritmo k-medoides. Es decir, se tienen k objetos representativos (medoi- des) de cada grupo. Segundo, cada objeto que no pertenece a la muestra es asignado al objeto más cercano de los k representativos. Esto resulta en una partición de todo el conjunto de objetos [38].

4.3. MÉTODOS DE CLASIFICACIÓN JERÁRQUICA

En cuanto a la técnica de clustering jerárquico, esta ofrece una alternativa a los algoritmos de agrupación basados en prototipos. En el clustering jerárquico, los clusters se crean de tal manera que sigan un orden predefinido, es decir, una estructura jerárquica. Una de las principales ventajas de la agrupación jerárquica es que no es necesario especificar el número de clusters de antemano, ya que se determina de forma automática. Además, esta técnica permite la visualización de dendrogramas. Los dendrogramas son representaciones visuales de agrupaciones jerárquicas binarias [39].

- **MÉTODO DE WARD**

En el proceso de agrupación de elementos, uno de los desafíos más importantes es determinar el número adecuado de clusters. En este sentido, los métodos jerárquicos abordan esta dificultad construyendo una estructura en la cual los elementos se agrupan en subconjuntos cada vez más amplios hasta que todos pertenecen al mismo conjunto final. Esta estrategia no solo muestra los grupos en sí, sino también revela las relaciones de proximidad existentes entre los elementos [40]. El Metodo de Ward, por lo tanto, es un método que permite crear construir un árbol de clasificación o dendograma. Un dendograma es una representación gráfica con forma de árbol que sirve para resumir el proceso de agrupación del análisis de clusters. Este método se basa en la suma de cuadrados y permite crear grupos de tamaño similar, dando paso a realizar buenos análisis de varianza por la producción de clusters definidos [41].

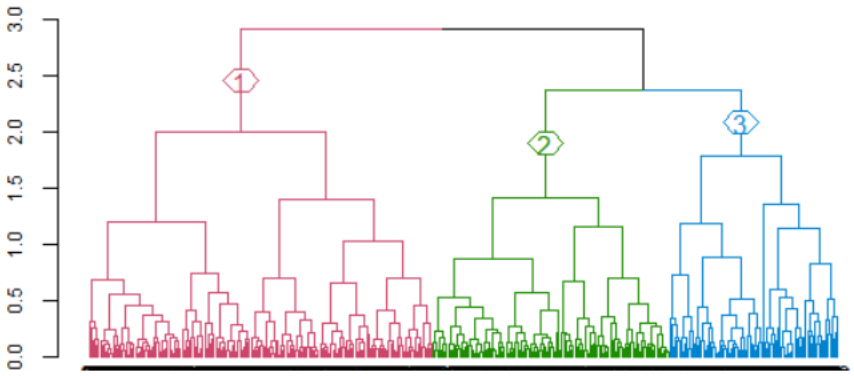
En suma, el objetivo del método de Ward es encontrar la partición óptima que minimice la suma de los cuadrados de las desviaciones de cada observación al centroide de su grupo. Para ello, se utiliza la siguiente expresión matemática:

$$\min \mathrm{C}\,\Sigma\mathrm{i} \;=\; 1\mathrm{k}\;\Sigma\mathrm{x} \in Cd(x, ci)^2$$

- $C = C1, C2, \ldots, Ck$ es una partición de *X* en *k* conjuntos.
- *Ci* es el conjunto de observaciones del grupo *i* en la partición *C*.
- x es una observación en el conjunto *X*.
- ci es el centroide del grupo i en la partición *C*.
- $\mathrm{d}(x, ci)$ es la distancia euclidiana entre la observación x y el centroide *ci*.

Así, el objetivo es encontrar la partición *C* que minimiza la suma de los cuadrados de las distancias entre cada observación *x* en su grupo *Ci* y su centroide *ci* [42].

Gráfica 3. Clasificación mediante método de Ward



Fuente: De la Hoz, Enrique [43].

- **K-VECINOS MÁS CERCANOS**

Es un algoritmo que se utiliza principalmente para clasificar valores al buscar los puntos de datos más similares. El proceso implica calcular la distancia entre el elemento a clasificar y el resto de los elementos en el conjunto de datos de entrenamiento. Luego, se seleccionan los "k" elementos más cercanos, determinados por la menor distancia según la función utilizada. El elemento no etiquetado se asigna a la clase que predomina entre los k vecinos más cercanos. La elección de un valor adecuado para k es crucial para el buen funcionamiento de este método, ya que determinará en gran medida a qué grupo pertenecerán los puntos, especialmente en las regiones limítrofes entre grupos [44].

El funcionamiento de K-NN se puede explicar sobre la base del siguiente algoritmo: (I)selecciona el número K de los vecinos; (ii) calcula la distancia euclidiana de K número de vecinos, (iii) toma los K vecinos más cercanos según la distancia euclidiana calculada, (iv) entre estos k vecinos, se cuenta el número de puntos de datos en cada categoría, (v) asigna los nuevos puntos de datos a esa categoría para la cual el número de vecinos es máximo [45].

También, el algoritmo se puede expresar matemáticamente de la siguiente manera [46]:
Sea $\mathrm{X} = (\mathrm{x1}, \mathrm{y1}), (\mathrm{x2}, \mathrm{y2}), \ldots, (\mathrm{xn}, \mathrm{yn})$ el conjunto de entrenamiento donde cada observación xi tiene una clase *yi*, y sea *x0* la nueva observación a clasificar, entonces la clasificación de *x0* se obtiene de la siguiente manera:

1. Se calcula la distancia euclidiana entre *x0* y cada observación *xi* del conjunto de entrenamiento *X*.
2. Se Seleccionan los *k* vecinos más cercanos de $x0$. Es decir, aquellos puntos en el conjunto de entrenamiento X cuya distancia euclidiana a $x0$ es la menor.
3. Se determina la clase de $x0$ mediante la mayoría de los votos de los *k* vecinos más cercanos.

La fórmula matemática para calcular la distancia euclidiana entre dos puntos *xi* y *xj* es la siguiente:

$$\mathrm{d}(\mathrm{xi}, \mathrm{xj}) = \sqrt{(\Sigma \mathrm{k} = 1\mathrm{p}\ (\mathrm{xik} - \mathrm{xjk})^2)}$$

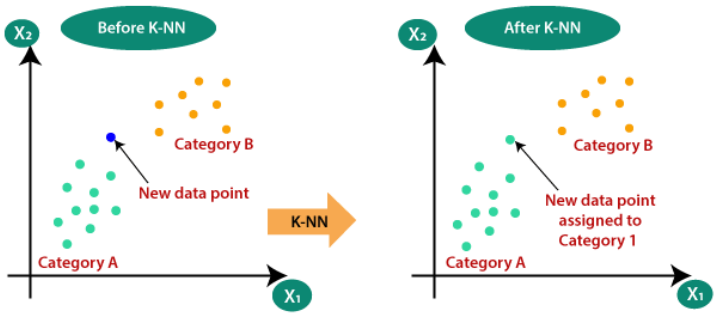
- p es el número de variables (dimensiones) de cada observación
- xik es el valor de la variable *k* de la observación xi
- xjk es el valor de la variable *k* de la observación xj

La fórmula matemática para determinar la clase de $x0$ mediante la mayoría de los votos de los k vecinos más cercanos es la siguiente:

$$\mathrm{y0} = \mathrm{argmax}\left(\Sigma \mathrm{i} = 1\mathrm{k}\, \mathrm{I}(\mathrm{yi} = \mathrm{c})\right)$$

- $y0$ es la clase asignada a la nueva observación $x0$.
- c es una de las clases posibles.
- I es una función indicadora que toma el valor 1 si $yi = c$ y *0* en caso contrario.

Gráfica 4. Clasificación mediante método K-vecino más cercano



Fuente: JavaTPoint [47].

- **DBSCAN**

Finalmente, la Agrupación Espacial Basada en Densidad de Aplicaciones con Ruido, o DBSCAN (Density-Based Spatial Clustering of Applications with Noise), "es otro algoritmo de agrupación especialmente útil para identificar correctamente el ruido en los datos" [48]. En suma, es un algoritmo de clúster o agrupamiento basado en la densidad que puede ser utilizado para identificar clústeres de cualquier forma en un conjunto de datos que contiene ruido y valores atípicos. Este algoritmo requiere de dos parámetros: Épsilon (eps): especifica lo cerca que deben estar los puntos entre sí para ser considerados parte de un clúster. Esto significa que, si la distancia entre dos puntos es menor o igual a este valor de épsilon, estos puntos se consideran vecinos y Puntos mínimos (minPts): el número mínimo de puntos para formar una región densa. A diferencia de K Means, DBSCAN no requiere que el usuario especifique el número de clústeres que se generarán [49].

El algoritmo DBSCAN se puede expresar matemáticamente de la siguiente manera: dado un conjunto de puntos P en un espacio euclidiano d-dimensional, se define la vecindad de un punto p como el conjunto de puntos que se encuentran a una distancia euclidiana menor o igual a un valor de ∈:

$$\mathrm{N}\epsilon(p) = \{q \epsilon P : dist(p,q) \leq \epsilon\}$$

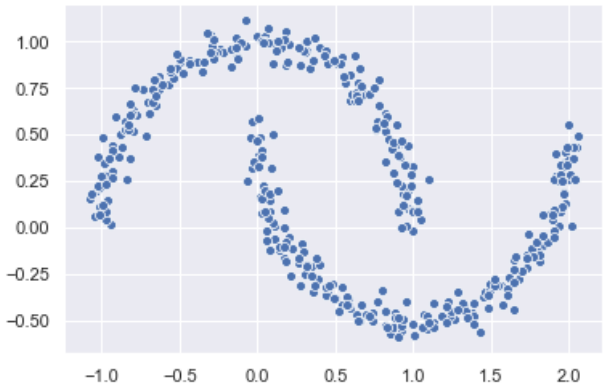
$dist(p,q)$ es la distancia euclidiana entre los puntos p y q. Se define la densidad de un punto p como el número de puntos en su vecindad:

$$\mathrm{p(q)} = |N\epsilon(p)|$$

Un punto p se considera un punto central si su densidad es mayor o igual a un valor mínimo de *MinPts*: p es un punto central si $\rho(\mathrm{p}) \geq m\mathrm{inPts}$

Se define la conectividad entre dos puntos *p* y *q* como la existencia de una cadena de puntos que conecte ambos puntos y cuyos puntos intermedios tienen una densidad mayor o igual a *MinPts*: *p* está conectado con *q* si existe una cadena de puntos $\{\mathrm{p1}, p2, \ldots, pn\}$ con $\mathrm{p1} = \mathrm{p}$, $\mathrm{pn} = \mathrm{q}$ y $\rho(p1) \geq m\mathrm{inPts}$ para $2 \leq i \leq n-1$ [46].

Gráfica 5. Agrupación mediante método DBSCAN



Fuente: interactivechaos [50]

5. IDENTIFICACIÓN DE VARIABLES Y ANÁLISIS EXPLORATORIO

Para llevar a cabo la construcción del modelo de aprendizaje no supervisado para la identificación de los determinantes de la VBG en Colombia se utilizaron las bases de datos del Instituto de Medicina Legal y Ciencias Forenses que, a través del Sistema de Información Red de Desaparecidos y Cadáveres – SIRDEC (para lesiones fatales) y en el Sistema de información de clínica y Odontología SICLICO (para lesiones no fatales), provee cifras estadísticas obtenidas de la práctica forense. Lo anterior con el objetivo de que estos datos puedan ser consultadas por el público en general y sean utilizadas como insumo para generar políticas públicas y toma de decisiones en cuanto a la mitigación y prevención de las lesiones de causa externa que ocurren en el territorio nacional colombiano [51].

Es importante anotar que el Instituto de Medicina Legal cuando clasifica un caso en la categoría solicitada, no tipifica el delito, tampoco hace juicios de responsabilidad y no determina la legalidad o ilegalidad del hecho, toda vez que estas actividades corresponden a las autoridades competentes. La entidad únicamente realiza una clasificación de carácter forense y genera hallazgos durante el procedimiento de necropsia y del relato de la víctima, dependiendo del caso.

Las bases de datos proporcionadas por Medicina Legal, tanto la base de datos de violencia sexual como la base de datos de violencia de pareja, incluyeron una amplia variedad de variables relevantes para el análisis y que son descritas a continuación:

- VARIABLE GRUPO DE EDAD: Se refiere a la clasificación de la población en diferentes grupos según su edad.
- VARIABLE GRUPO EDAD JUDICIAL: Se refiere a la clasificación de las personas que se encuentran en el sistema judicial según su edad.
- VARIABLE CICLO VITAL: se refiere a las diferentes etapas que experimenta una persona a lo largo de su vida.
- VARIABLE ESCOLARIDAD: Representa el nivel de educación formal que ha alcanzado una persona por el título o grado académico obtenido.
- VARIABLE ESTADO CIVIL: Indica la situación legal de una persona en relación con su matrimonio o convivencia con otra persona.
- VARIABLE FACTOR DE VULNERABILIDAD: indica una condición, situación o circunstancia que puede aumentar la susceptibilidad de una persona a ser víctima de algún tipo de daño o violencia.
- VARIABLE TIPO DE DISCAPACIDAD: Tipo de limitación física, sensorial, cognitiva o intelectual que una persona puede tener.
- VARIABLE PERTENENCIA ÉTNICA: se refiere a la identificación de la persona con un grupo étnico determinado, lo cual puede influir en diferentes aspectos de la vida, como la cultura, la religión, las tradiciones y la identidad.
- VARIABLE PRESUNTO AGRESOR: se refiere a la persona que se presume ha cometido la agresión o delito investigado en un proceso judicial.
- VARIABLE SEXO DEL PRESUNTO AGRESOR: se refiere al género del presunto agresor y puede ser relevante en el análisis de las motivaciones o patrones de comportamiento en casos de violencia.
- VARIABLE CIRCUNSTANCIA DEL HECHO: Se refiere a la situación en la que ocurrió el hecho que se está investigando.
- VARIABLE ACTIVIDAD DURANTE EL HECHO: se refiere a la actividad que realizaba la víctima o el presunto agresor en el momento en que ocurrió el hecho.
- VARIABLE PAÍS DE NACIMIENTO DE LA VÍCTIMA: se refiere al país en el que nació la víctima del hecho investigado.
- VARIABLE FACTOR DESENCADENANTE DE LA AGRESIÓN: se refiere al factor o razón que se presume que desencadenó la agresión o el delito.
- VARIABLE MECANISMO CAUSAL: se refiere al medio o la forma en que se cometió el hecho, por ejemplo, si fue mediante el uso de un arma de fuego o un objeto contundente.
- VARIABLE DIAGNÓSTICO TOPOGRÁFICO DE LA LESIÓN: se refiere a la localización y tipo de lesión que sufrió la víctima del hecho investigado.
- VARIABLE DÍAS DE INCAPACIDAD MÉDICO LEGAL: se refiere a la cantidad de días que la víctima se encuentra incapacitada debido a la lesión que sufrió.
- VARIABLE ESCENARIO DEL HECHO: se refiere al lugar donde ocurrió el hecho investigado

- VARIABLE ZONA DEL HECHO: se refiere a la ubicación geográfica del hecho investigado, como una ciudad, un barrio, una zona rural, etc.
- VARIABLE MES DEL HECHO: se refiere al mes en que ocurrió el hecho investigado.
- VARIABLE DÍA DEL HECHO: se refiere al día del mes en que ocurrió el hecho investigado.
- VARIABLE RANGO DE HORA DEL HECHO (X 3 HORAS): se refiere a la franja horaria en la que ocurrió el hecho investigado.

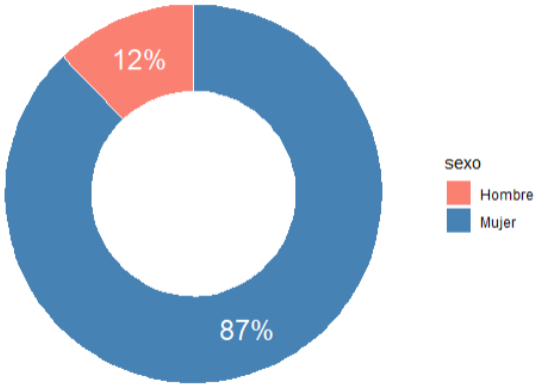
Ahora bien, para utilizar los datos proporcionados por Instituto de Medicina Legal fue necesario transformar y optimizar las bases de datos para su posterior análisis y manipulación. Estas bases de datos fueron construidas en un formato que dificultaba su análisis, pues las variables se encontraban organizadas verticalmente y, además, de una sola variable se desprendían numerosas subvariables. Así mismo, los datos sobre cada año (2019,2020,2021) se encontraban en tablas diferentes. De esta manera, fue necesario realizar la siguiente transformación:

1. Se pivoteó la base en un formato vertical para facilitar su manipulación.
2. Se unificaron en una sola base de datos los registros de la entidad de 2019,2020, 2021.
3. Se unificó en la misma base de datos los registros de la ciudad de Bogotá, que se encontraban divididos por localidades.
4. Se realizó una selección de las variables en las bases de datos de violencia sexual y de pareja que serían útiles para el análisis eliminando aquellas que no aportarían lo suficiente al análisis o que, en su defecto, sería muy difíciles de agrupar por su heterogeneidad o multiplicidad de opciones. Lo anterior debido a que la base de datos de presuntos delitos de violencia sexual y de pareja de Medicina Legal contaban con un elevado número de variables lo que eventualmente podría generar dificultades con la construcción y funcionamiento del modelo. En el **Anexo 1** se presenta la selección de variables de las bases de datos de violencia sexual y de pareja junto con las respectivas agrupaciones que se llevaron a cabo para el análisis.
5. Finalmente, para la creación del modelo de aprendizaje no supervisado y para conocer las tendencias reales de la VBG en Colombia fue necesario convertir los totales de las bases de datos en tasas por cada 100.000 habitantes, a partir del censo de DANE del año 2020. Esto, debido a que de no realizarse esta transformación de los datos aquellos hubiesen podido quedar sesgados, toda vez que las ciudades capitales tales como Bogotá, Medellín o Cali concentrarían siempre el mayor número de casos dado que su mayor número de población. Finalmente, se realizó un análisis exploratorio con el objetivo de conocer las tendencias de la violencia sexual y de pareja en Colombia durante los tres años de estudio. Este análisis exploratorio arrojó los siguientes resultados:

5.1 ANÁLISIS EXPLORATORIO-PRESUNTOS CASOS DE VIOLENCIA SEXUAL EN COLOMBIA

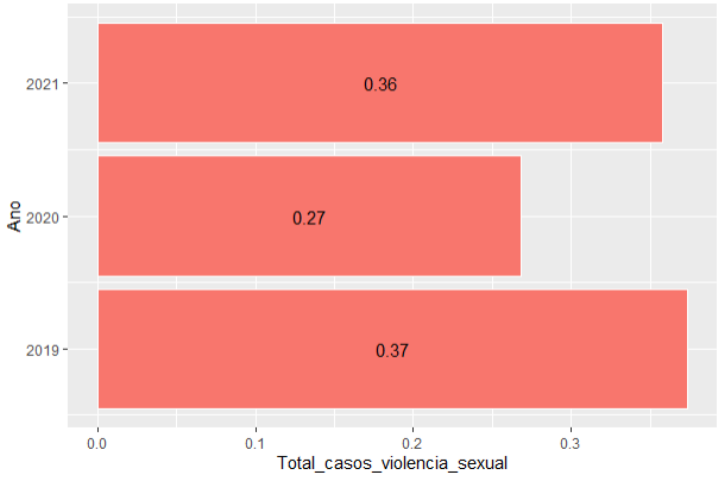
Según los datos del Instituto de Medicina Legal y Ciencias Forenses, a partir de la Base de la datos del Sistema de Información de Clínica y Odontología Forense – SICLICO, entre los años 2019-2021 se registraron en Colombia 66.733 presuntos casos de violencia sexual contra hombres y mujeres. Tal como muestra la gráfica 6, de estos, el 87% (57.786) correspondió a mujeres, mientras que el 12% (9.033) contra hombres.

Gráfica 6 Sexo víctima presuntos casos de violencia sexual 2019-2021



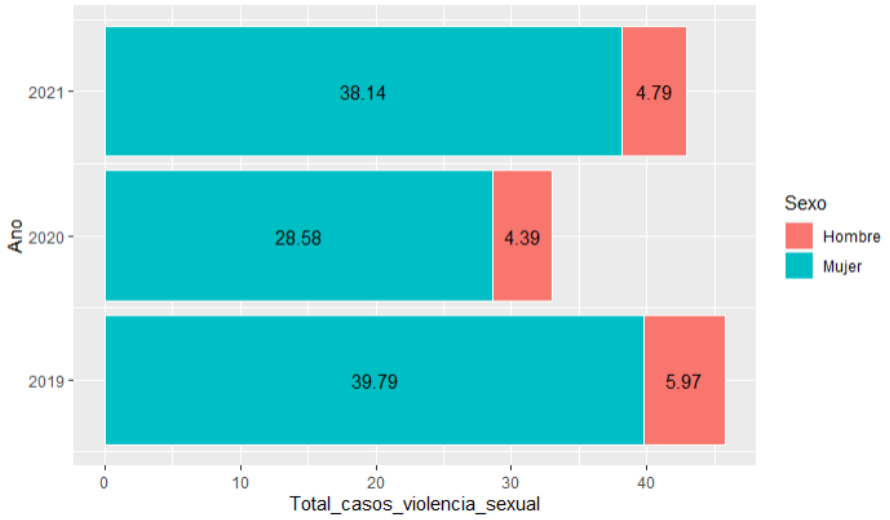
Adicionalmente, como muestra la gráfica 7, el año que concentró el mayor número de presuntos casos de violencia sexual contra mujeres en Colombia por cada 100.000 habitantes fue 2019, año en el que se registró el 37% total de casos, seguido de 2021 con el 36% y 2020 con el 27%.

Gráfica 7 Año presuntos casos de violencia sexual 2019-2021



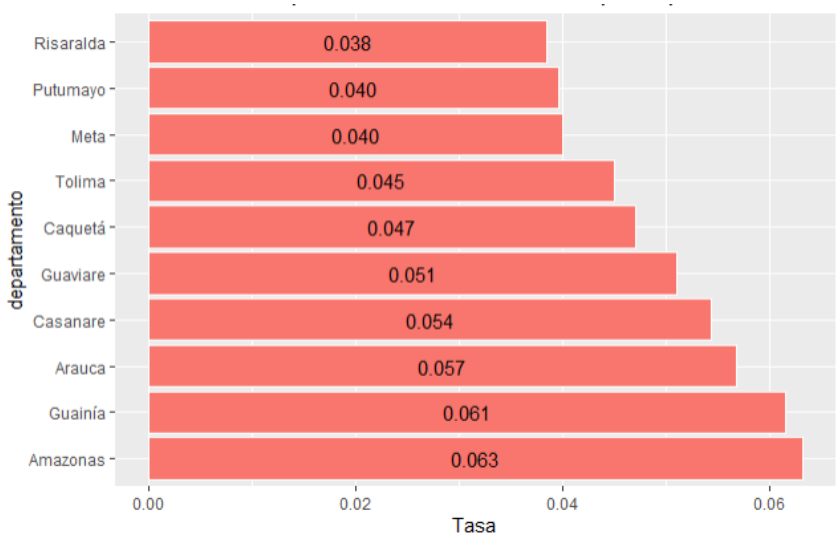
En relación con la ocurrencia de los presuntos casos de violencia sexual en los años de estudio, tal como muestra la gráfica 8, los tres años el mayor porcentaje de casos ocurrió contra mujeres, solo una pequeña proporción ocurrió contra hombres en los años de estudio.

Gráfica 8. Sexo vs año presuntos casos de violencia sexual 2019-2021



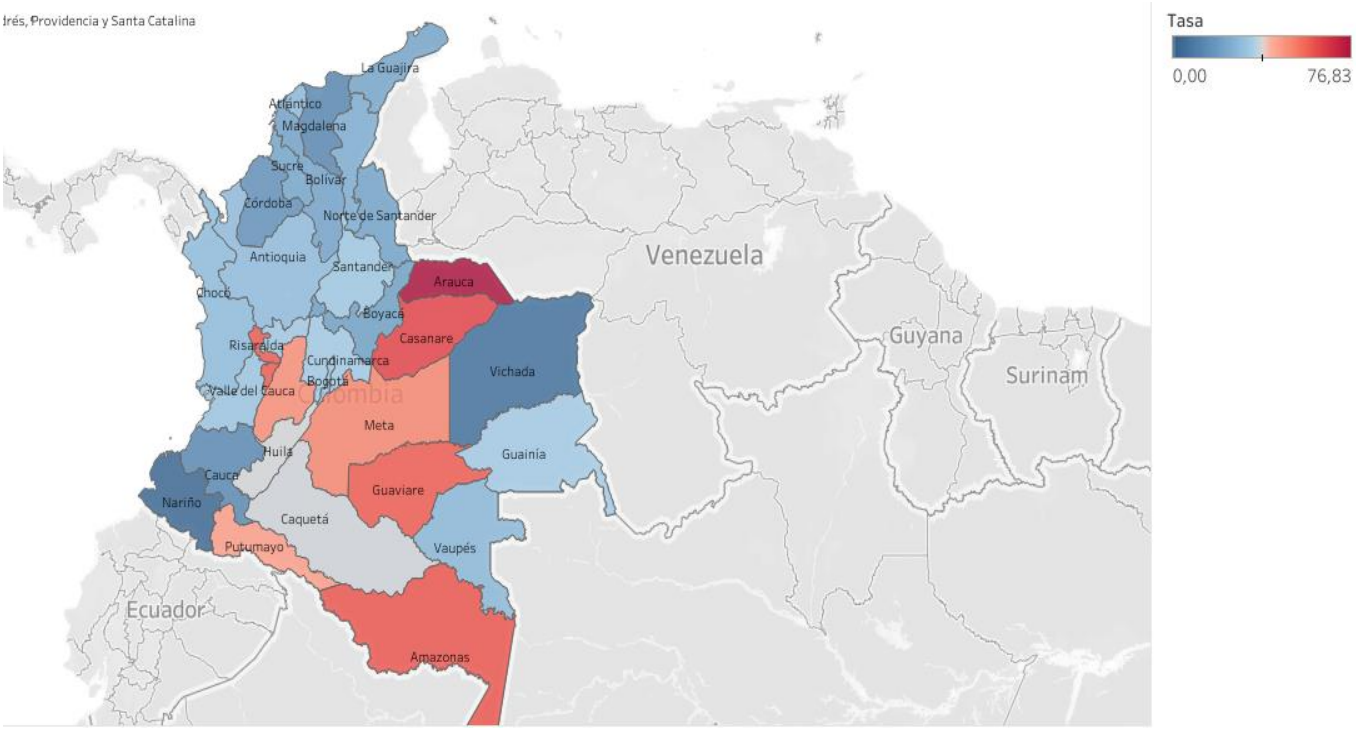
Frente a los departamentos que albergaron la mayor tasa frente al número de presuntos casos de violencia sexual contra mujeres por cada 100.000 habitantes en los tres años de estudio, tal como muestra la gráfica 9, se encuentran en el top diez Amazonas, Guainía, Arauca, Casanare, Guaviare, Caquetá, Tolima, Meta Putumayo y Risaralda.

Gráfica 9. Presuntos casos de violencia sexual por cada 100.000 habitantes por departamento



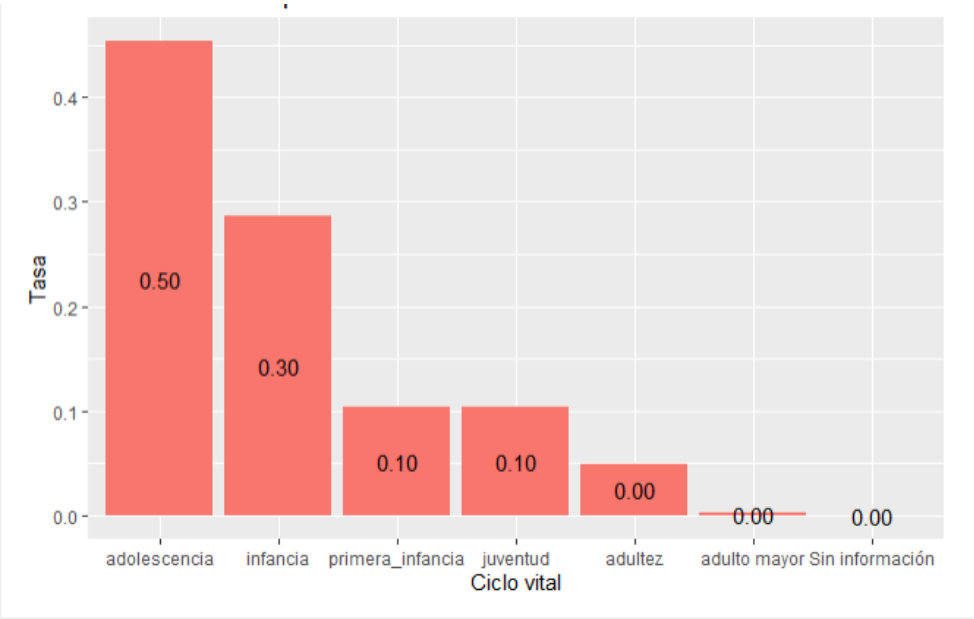
En el siguiente mapa (ilustración 1) es posible identificar que los departamentos que se encuentran cubiertos por tonalidades rojizas son quienes representan el mayor número de casos por cada 100.000 habitantes en los tres años de estudio.

Ilustración 1. Georreferenciación Presuntos casos de violencia sexual por cada 100.000 habitantes por departamento



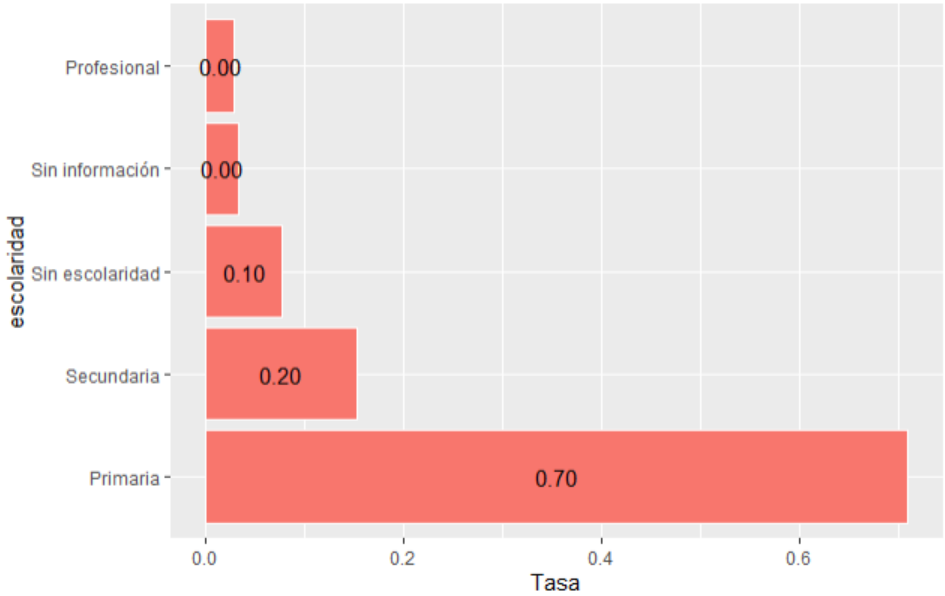
Frente al ciclo vital de las mujeres que fueron presuntamente víctimas de violencia sexual en Colombia, según los datos del Instituto de Medicina Legal, se tiene que en su mayoría se encontraban en su adolescencia (12-17 años) con el 50% de los casos, seguido de la infancia (6-11 años), con el 30%, y primera infancia (00-5 años) con el 10% (gráfica 10).

Gráfica 10. Presuntos casos de violencia sexual por cada 100.000 habitantes según ciclo vital



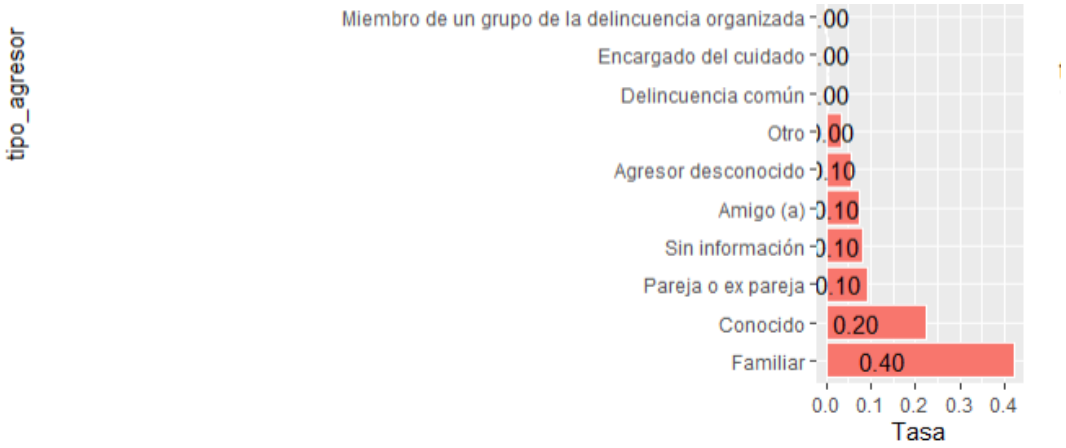
En relación a la escolaridad de las mujeres presuntamente víctimas de violencia sexual, según los datos del Instituto de Medicina Legal, la mayor tasa por cada 100.000 habitantes la ocupan quienes contaban con un nivel de escolaridad correspondiente a primaria, con un 70%, seguido de secundaria 20% y sin escolaridad 10% (gráfica 11).

Gráfica 11. Presuntos casos de violencia sexual por cada 100.000 habitantes según escolaridad



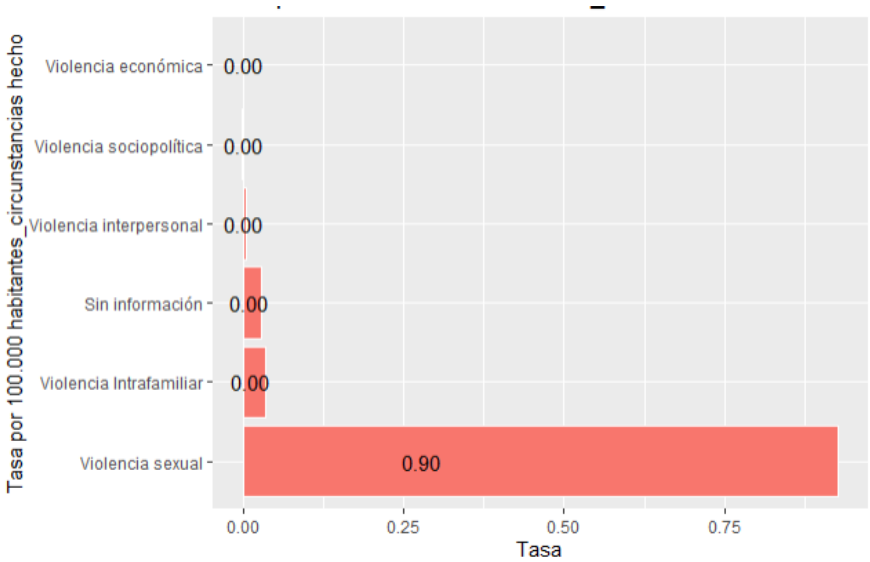
Adicionalmente, los datos del Instituto de Medicina Legal indican que, en los casos presuntos casos de violencia sexual contra las mujeres, los principales presuntos agresores corresponden a familiares (40%), seguido de conocidos (20%) y de pareja o exparejas (10%) (gráfica 12).

Gráfica 12. Presuntos casos de violencia sexual por cada 100.000 habitantes según presunto agresor



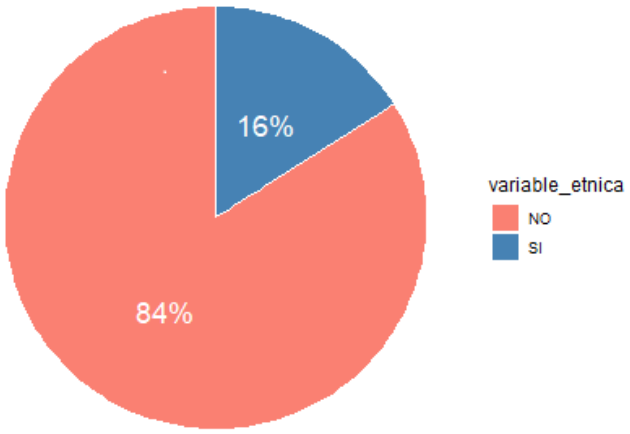
Frente a las circunstancias del hecho, los datos indican que la tasa más alta correspondió a la violencia sexual, con el 90%, en el que se circunscribe el abuso sexual, asalto sexual, acceso carnal violento/acto sexual violento con persona protegida), seguido de la violencia intrafamiliar 0,3%, y sin información (gráfica 13).

Gráfica 13. Presuntos casos de violencia sexual por cada 100.000 habitantes según circunstancias del hecho



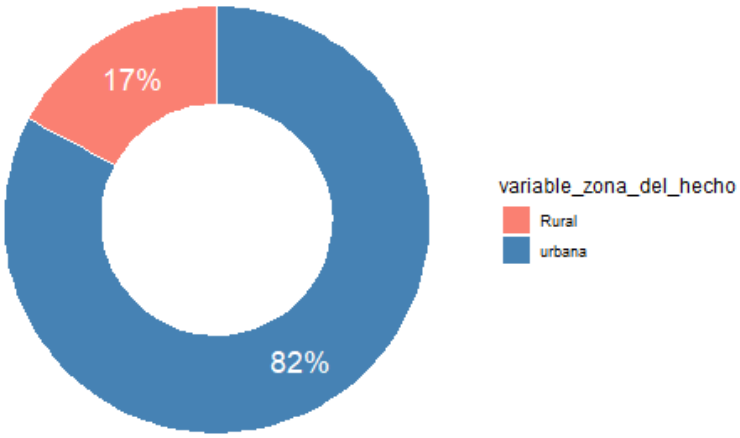
En relación con la pertenencia étnica de las mujeres que fueron presuntamente víctimas de violencia sexual, los datos indican que el 16% tenía algún tipo de pertenencia étnica (indígena, negro afrodescendiente, palenquero, raizal, rom), mientras que el 84% no se identificaba con algún tipo de pertenencia étnica (gráfica 14).

Gráfica 14. Presuntos casos de violencia sexual por cada 100.000 habitantes según pertenencia étnica



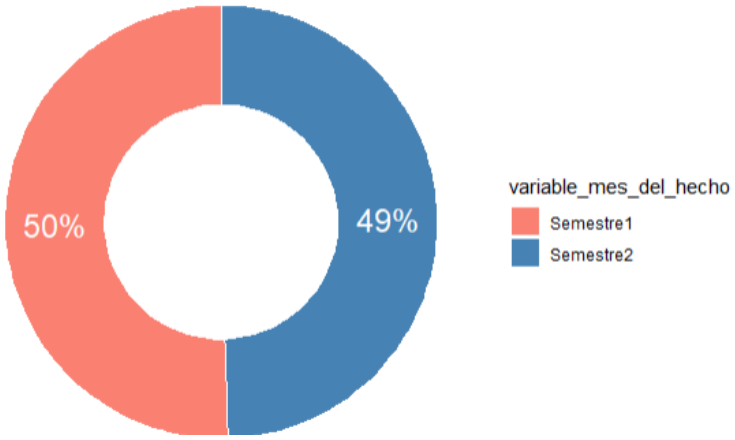
Frente a la zona del hecho, el 82% de los presuntos casos de violencia sexual contra mujeres en Colombia en el periodo de estudio ocurrieron en zona urbana y en una menor proporción, el 17%, tuvieron lugar en zonas rurales (gráfica 15).

Gráfica 15. Presuntos casos de violencia sexual por cada 100.000 habitantes según zona del hecho



Finalmente, el 50% de los presuntos casos de violencia sexual contra las mujeres en Colombia durante el periodo de estudio ocurrieron en el primer semestre del año, mientras que el 49% tuvo lugar en el segundo semestre (gráfica 16).

Gráfica 16. Presuntos casos de violencia sexual por cada 100.000 habitantes según semestre del año



Conclusiones:

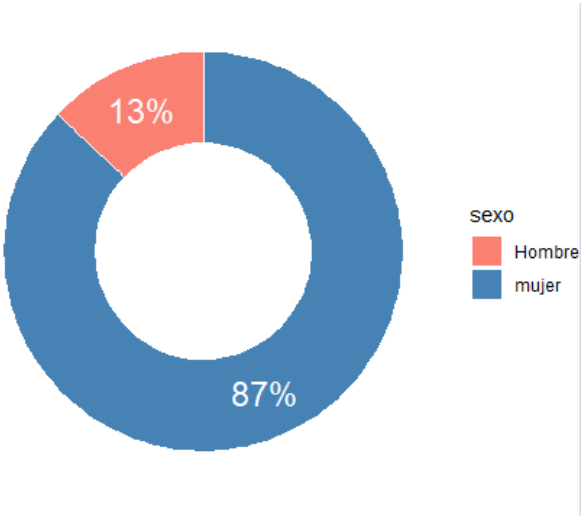
- El análisis exploratorio frente a los presuntos casos de violencia sexual en Colombia en los tres años de estudio muestra que la mayoría de casos ocurrió contra las mujeres (87%), a diferencia de los hombres con un (12%).
- El año que concentró el mayor número de presuntos casos de violencia sexual contra mujeres en Colombia fue 2019, año en el que se registró el 37% total de casos, seguido de 2021 con el 36% y 2020 con el 27%.
- Frente a los departamentos que albergaron la mayor tasa frente al número de presuntos casos de violencia sexual contra mujeres por cada 100.000 habitantes en los tres años de estudio se encuentran en el top diez Amazonas, Guainía, Arauca, Casanare, Guaviare, Caquetá, Tolima, Meta Putumayo y Risaralda.
- Frente al ciclo vital de las mujeres que fueron presuntamente víctimas de violencia sexual en Colombia se tiene que en su mayoría se encontraban en su adolescencia (12-17 años) con el 50% de los casos.

- En relación a la escolaridad de las mujeres presuntamente víctimas de violencia sexual, la mayor tasa por cada 100.000 habitantes la ocupan quienes contaban con un nivel de escolaridad correspondiente a primaria, con un 70%.
- En los presuntos casos de violencia sexual contra las mujeres, los principales presuntos agresores correspondieron a familiares (40%).
- Frente a las circunstancias del hecho, los datos indican que el 90% de los casos correspondieron a hechos de violencia sexual el que se circunscribe el abuso sexual, asalto sexual, acceso carnal violento/acto sexual violento con persona protegida.
- Frente a la zona del hecho, el 82% de los presuntos casos de violencia sexual contra mujeres en Colombia en el periodo de estudio ocurrieron en zona urbana.
- Finalmente, el 50% de los presuntos casos de violencia sexual contra las mujeres en Colombia durante el periodo de estudio ocurrieron en el primer semestre del año.

5.2. ANÁLISIS EXPLORATORIO VIOLENCIA DE PAREJA

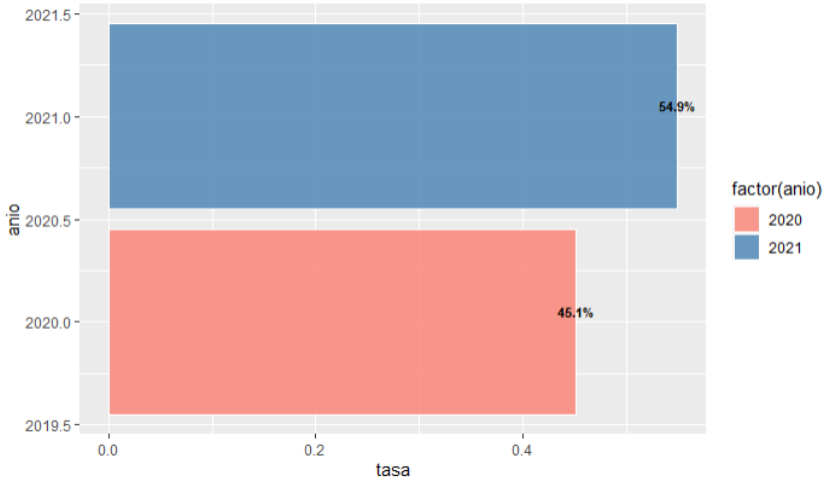
Según los datos del Instituto de Medicina Legal y Ciencias Forenses, a partir de la Base de datos del Sistema de Información de Clínica y Odontología Forense – SICLICO, entre los años 2020-2021 se registraron en Colombia 65.523 presuntos casos de violencia de pareja contra hombres y mujeres. De estos, el 87% (56.906) correspondió a mujeres, mientras que el 13% (8.617) contra hombres (gráfica 17).

Gráfica 17. Sexo víctima presuntos casos de violencia de pareja 2020-2021



Según los datos del Instituto de Medicina Legal, el año que concentró la mayor tasa de casos de violencia de Pareja contra mujeres en Colombia fue 2021 en el que se registró el 55% total de casos, seguido de 2020 con el 45% (gráfica 18).

Gráfica 18. Año presuntos casos de violencia de pareja 2020-2021



En cuanto a los departamentos con mayor tasa frente al número de casos de violencia de pareja contra las mujeres por cada 100.000 habitantes se encuentran en el top diez Amazonas, San Andrés, Casanare, Meta, Arauca, Tolima, Bogotá, Cundinamarca, Vaupés y Guainía (gráfica 19) (Ilustración 2).

Gráfica 19. Presuntos casos de violencia de pareja por cada 100.000 habitantes por departamento

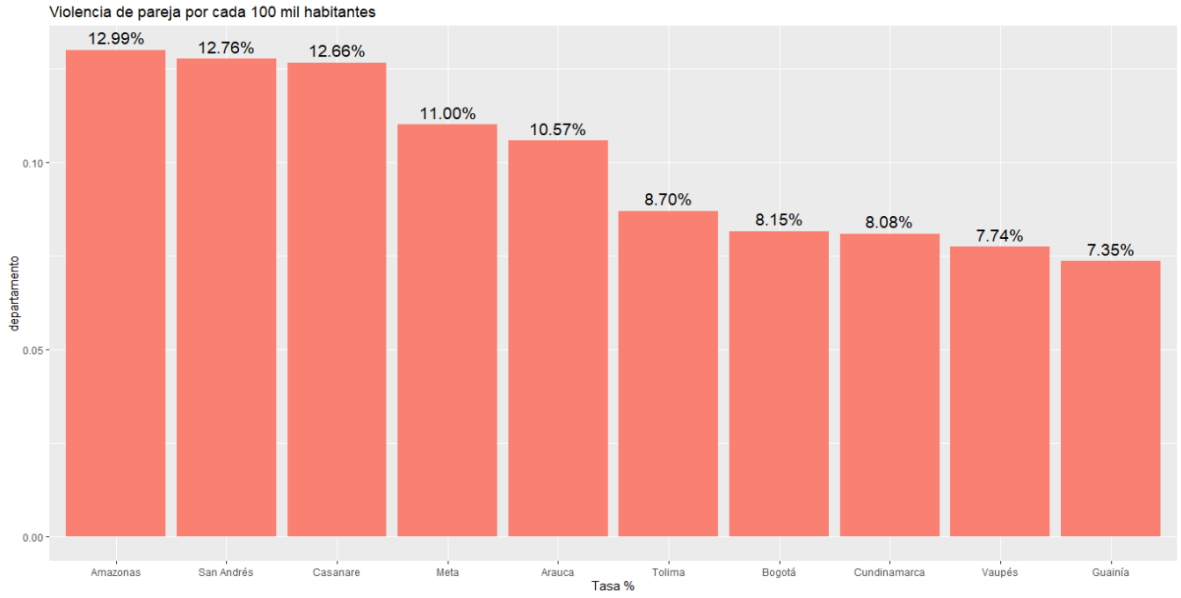
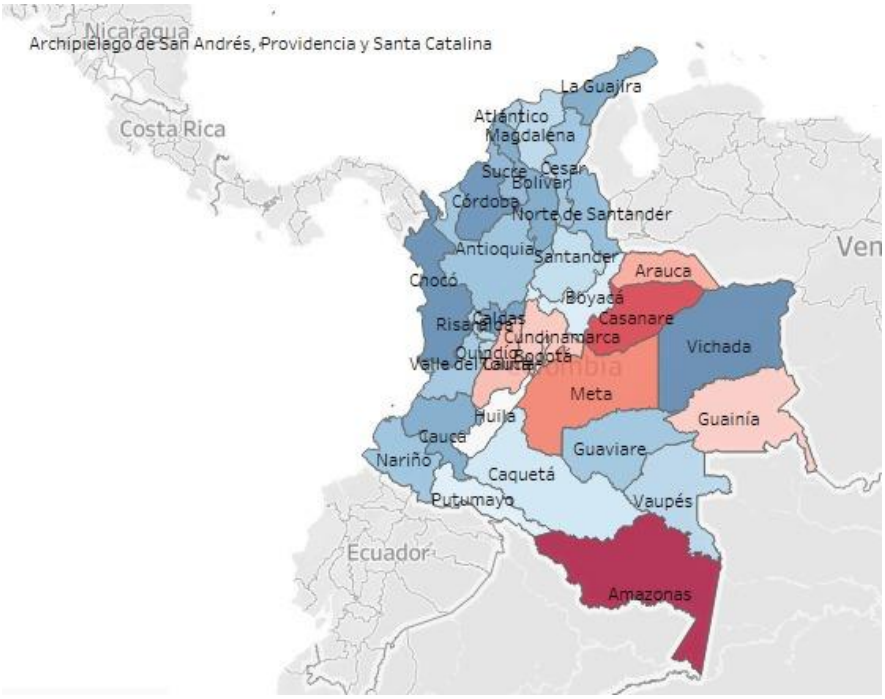
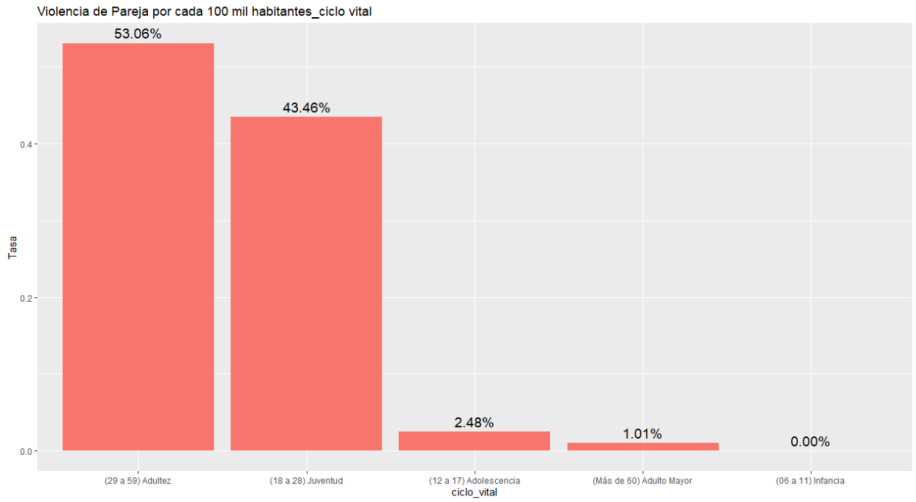


Ilustración 2. Georreferenciación Presuntos casos de violencia de pareja por cada 100.000 habitantes por departamento



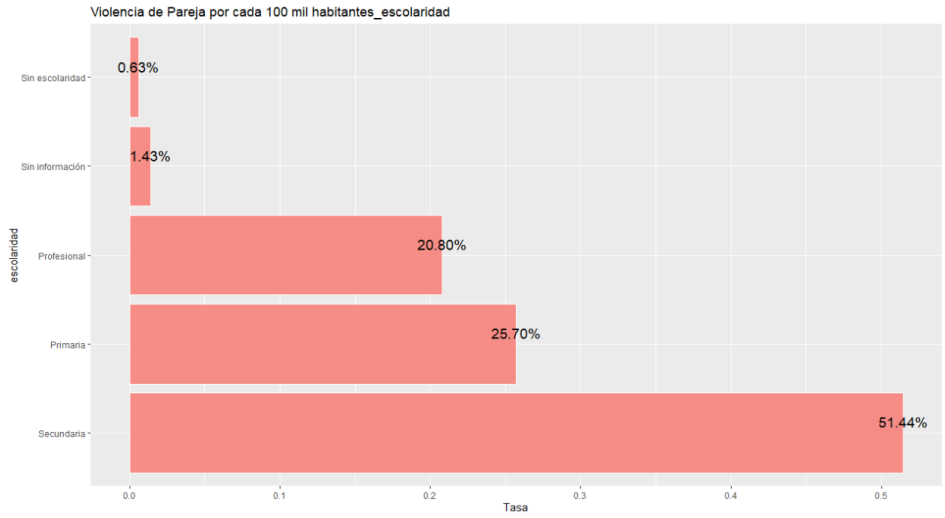
En cuanto al ciclo vital de las mujeres que fueron víctimas de violencia de pareja en Colombia, según los datos del Instituto de Medicina Legal, se tiene que en su mayoría se encontraban en su adolescencia (12-17 años), seguido de la juventud (18-28 años) (Gráfica 20).

Gráfica 20. Presuntos casos de violencia de pareja por cada 100.000 habitantes según ciclo vital



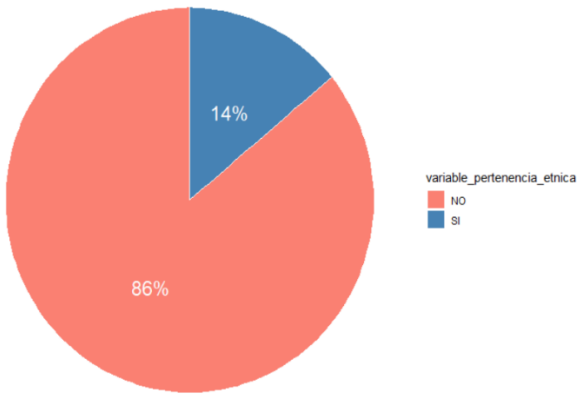
Con respecto al grado de escolaridad las mujeres que fueron víctimas de violencia de pareja en Colombia, según los datos del Instituto de Medicina Legal, en su mayoría cuentan con un rango de escolaridad de Secundaria con un porcentaje de 51.4%, seguido del grado de escolaridad primaria con un 25.7% (Gráfica 21*)*.

Gráfica 22. Presuntos casos de violencia de pareja por cada 100.000 habitantes según escolaridad



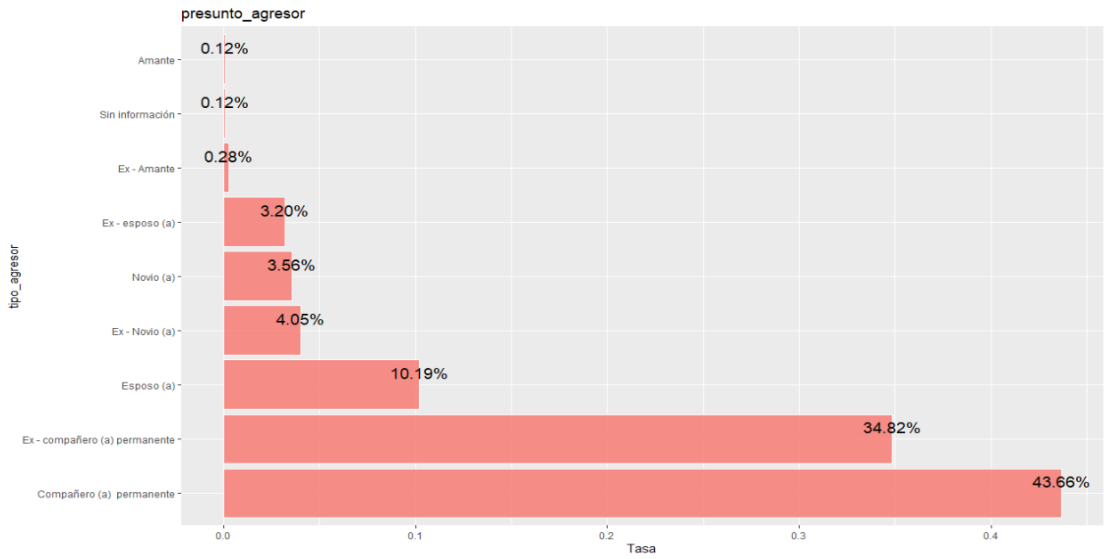
En relación con la pertenencia étnica de las mujeres víctimas de violencia de pareja, los datos indican que el 14% tenía algún tipo de pertenencia étnica (indígena, negro afrodescendiente, palenquero, raizal, rom), mientras que el 86% no se identificaba con algún tipo de pertenencia étnica (Gráfica 22).

Gráfica 23. Presuntos casos de violencia de pareja por cada 100.000 habitantes según pertenencia étnica



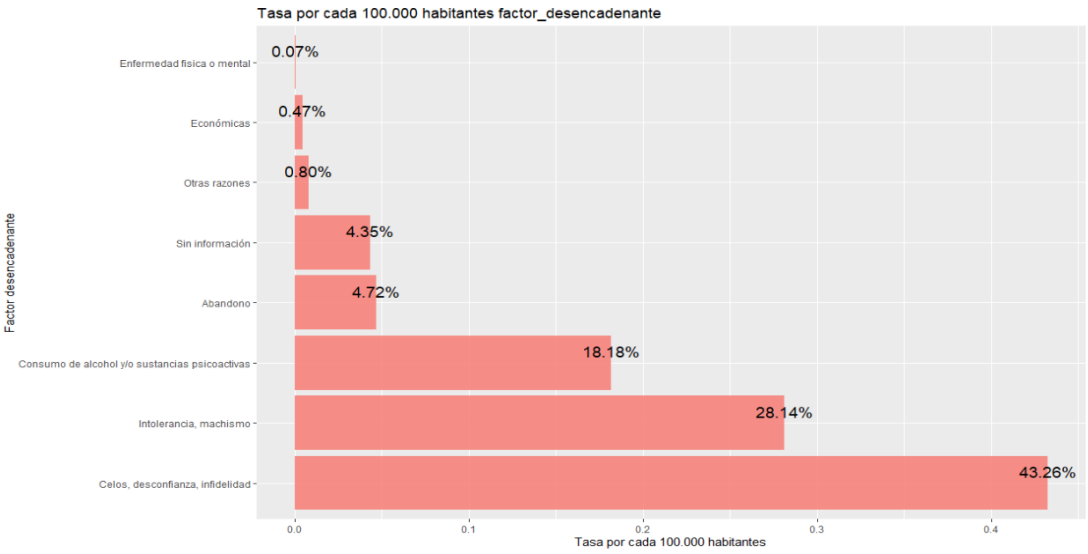
Con respecto a la variable presunto agresor las mujeres víctimas de violencia de pareja en su mayoría fueron agredidas por su compañero Permanente con un 44%, seguidas por un 35% por su Ex-compañero permanente (Gráfica 23).

Gráfica 24. Presuntos casos de violencia de pareja por cada 100.000 habitantes según presunto agresor



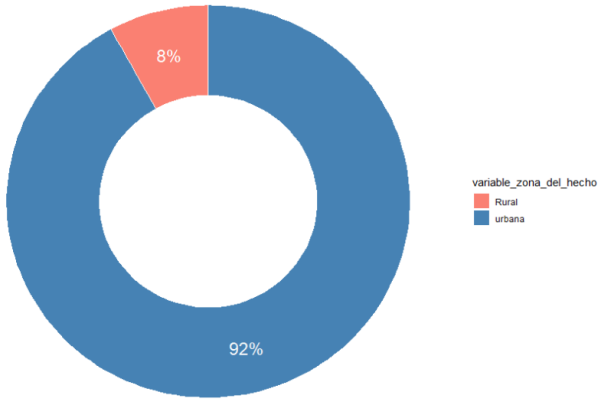
Según la variable factor desencadenante, la mayor tasa de casos de violencia de Pareja contra mujeres en Colombia fue dada por el factor de Celos, Desconfianza, Infidelidad con un 43%, seguido de la variable Intolerancia, Machismo con un 28% (Gráfica 24).

Gráfica 25. Presuntos casos de violencia de pareja por cada 100.000 habitantes según factor desencadenante



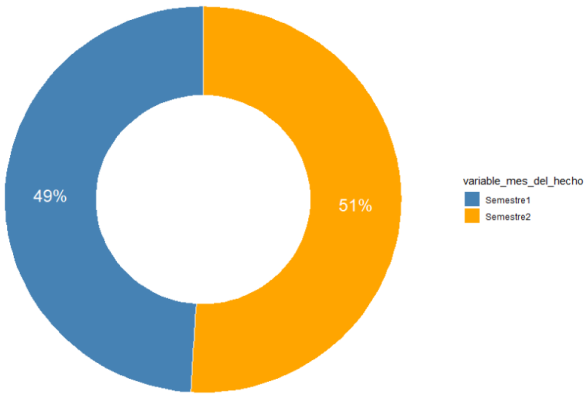
En cuanto a la zona de hecho el 92% de los casos de violencia de pareja contra mujeres en Colombia en el periodo de estudio ocurrieron en zona urbana y en una menor proporción, el 92%, tuvieron lugar en zonas rurales (Gráfica 25).

Gráfica 26. Presuntos casos de violencia de pareja por cada 100.000 habitantes según zona del hecho



Finalmente, en cuanto al mes del hecho el mayor porcentaje de los casos de violencia de pareja contra mujeres en Colombia fue para el segundo semestre de los años estudiados con un 51% (Gráfica 26).

Gráfica 27. Presuntos casos de violencia de pareja por cada 100.000 habitantes según semestre del hecho



6.EXPLORACIÓN DE TÉCNICAS DE APRENDIZAJE NO SUPERVISADO

En este capítulo se exploran algunas técnicas de aprendizaje no supervisado aplicadas a los datos de violencia de pareja y violencia sexual del Instituto de Medicina Legal y ciencias forenses. Se selecciona el aprendizaje no supervisado dado que los datos proporcionados por el Instituto de Medicina Legal no cuentan con clases previamente definidas, es decir, se encuentran totales y no cuentan con datos de salida que correspondan a un determinado input.

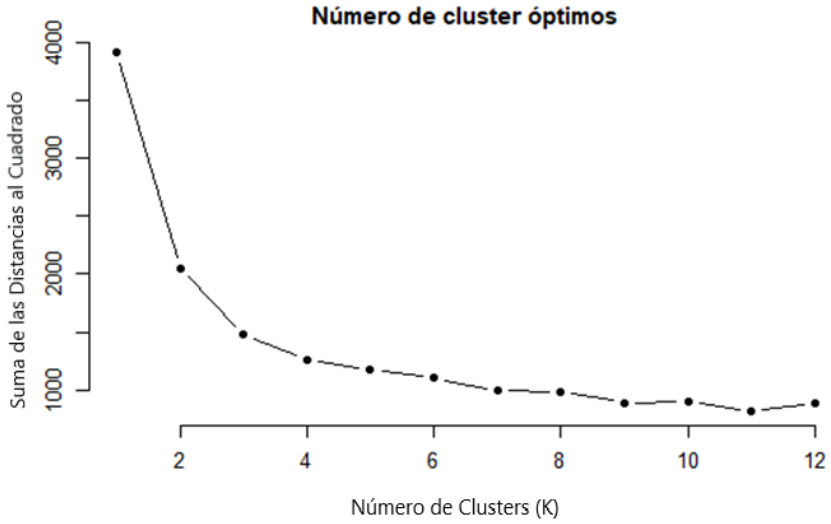
6.1. MODELO K MEANS

Uno de los objetivos de este proyecto consistió en identificar las técnicas de aprendizaje no supervisado óptimas para identificar los determinantes de la violencia sexual y de pareja en Colombia, por lo tanto, se inició con la exploración de la técnica de K-means, mediante la cual es posible agrupar las observaciones de forma tal que todas las que agrupen en el mismo grupo sean lo más semejantes entre sí y que las pertenecientes a grupos distintos sean lo más desemejantes entre sí. Tal como recoge Correa, uno de los problemas más comunes al aplicar alguno de los métodos de clustering como k-means y k-mediods, es el hecho de establecer el número de clústers, pues "si se definen pocos clústeres pueden causar problemas de heterogeneidad dentro de los mismos; y si, por el contrario, se definen muchos clústeres, pueden causar que datos muy similares sean divididos en más grupos de lo necesario" [52].

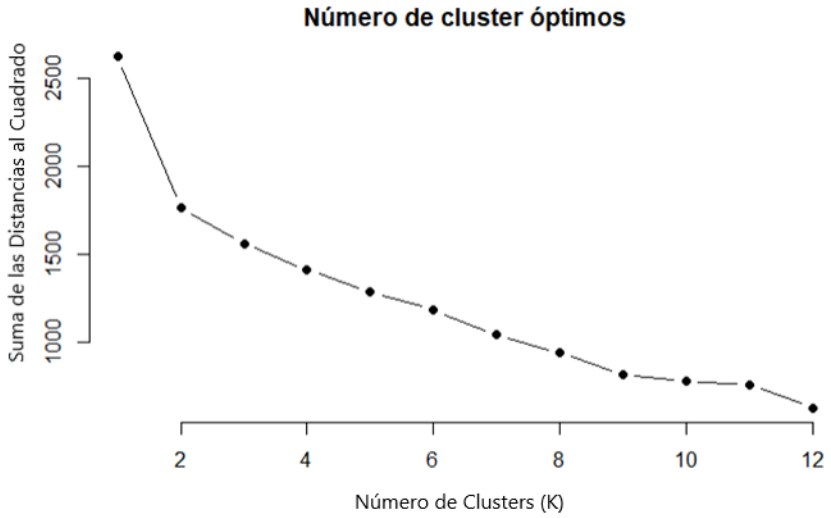
En ese sentido, existen pruebas de validación que permiten identificar el número óptimo de clústeres o centroides. Uno de estas pruebas de validación consiste en la prueba del codo que permite identificar la cantidad óptima de centroides *k* a utilizar. Básicamente, este método busca seleccionar la cantidad ideal de grupos a partir de la optimización de los WCSS (Within Clusters Summed Squares) [53].

Entonces, a fin de implementar el modelo de K-Medias a partir de la base de violencia sexual y de pareja, se inicia por determinar la cantidad óptima de centroides a utilizar a partir del *Método del Codo*. Para ello, se aplica la función kmeans de R al conjunto de datos, variando en cada caso el valor de *k*, y acumulando los valores de *WCSS* obtenidos. La función de R (fviz_nbclust) proporciona una solución conveniente para estimar la cantidad óptima de clústeres. Una vez calculados los valores de *WCSS* en función de la cantidad de centroides *k*, se grafican los resultados:

Gráfica 28. Validación técnica del codo base de datos presuntos casos de violencia sexual por cada 100.000 habitantes



Gráfica 29. Validación técnica del codo base de datos presuntos casos de violencia de pareja por cada 100.000 habitantes



En el análisis de las gráficas 27 y 28 se puede observar que a medida que incrementamos la cantidad de centroides (representados por el valor k) en ambas bases de datos, el valor de WCSS (Within-Cluster Sum of Squares) disminuye. Esto indica que los centroides están mejor ajustados a los datos y los grupos resultantes son más compactos. Sin embargo, hay un punto crítico en la gráfica donde el descenso en el valor de WCSS se vuelve menos pronunciado, y la curva adopta

una forma de codo. Este punto representa un equilibrio entre la cantidad de centroides utilizados y la calidad de la agrupación resultante.

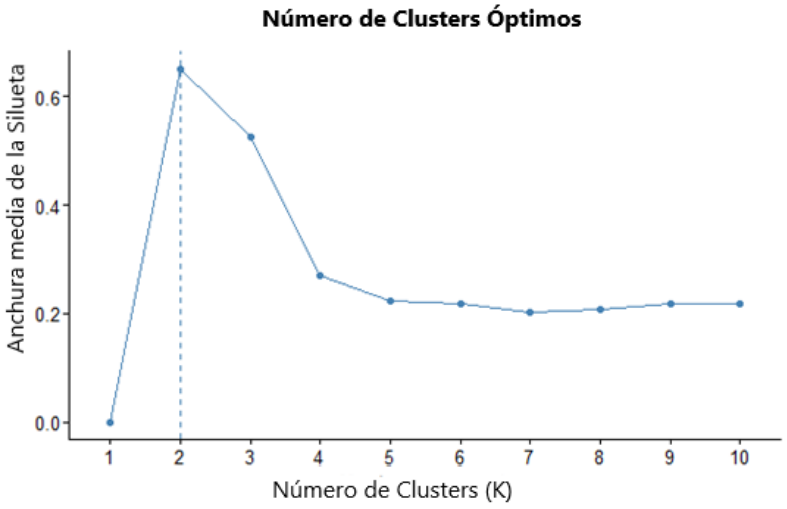
Para seleccionar el valor óptimo de k, se busca ese punto donde ya no se producen variaciones significativas en el valor de WCSS al aumentar k. En otras palabras, se busca el punto en el que agregar más centroides no mejora significativamente la calidad de la agrupación.

En el caso específico mencionado, este punto se encuentra a partir de k >= 2, lo que significa que agregar más centroides después de k = 2 no genera cambios importantes en el valor de WCSS. Por lo tanto, seleccionar k = 2 como valor óptimo en este caso sería razonable, ya que proporciona una buena agrupación con una cantidad mínima de centroides.

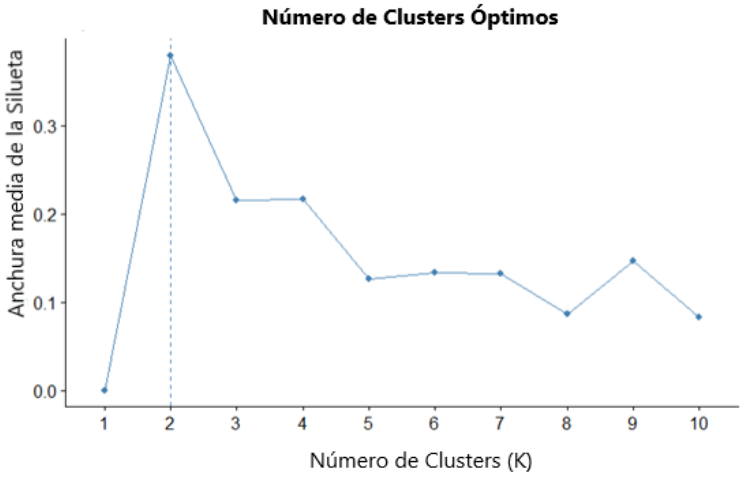
Es importante destacar que la elección del valor óptimo de k puede variar según el conjunto de datos y el contexto del problema, por lo que se deben considerar otros criterios y técnicas de validación para tomar una decisión final.

Otra prueba de validación utilizada es el método de la silueta (gráfica 29 y 30), el cual evalúa la calidad de una agrupación al determinar cómo se encuentra cada objeto dentro de su grupo. Una alta medida de silueta promedio indica un buen agrupamiento. El método de la silueta promedio calcula la medida de silueta promedio para diferentes valores de "k". El número óptimo de grupos "k" es aquel que maximiza la medida de silueta promedio dentro de un rango de valores posibles para "k"[54].

Gráfica 30. Validación método de la silueta base de datos presuntos casos de violencia sexual por cada 100.000 habitantes



Gráfica 31. Validación método de la silueta base de datos presuntos casos de violencia de pareja por cada 100.000 habitantes

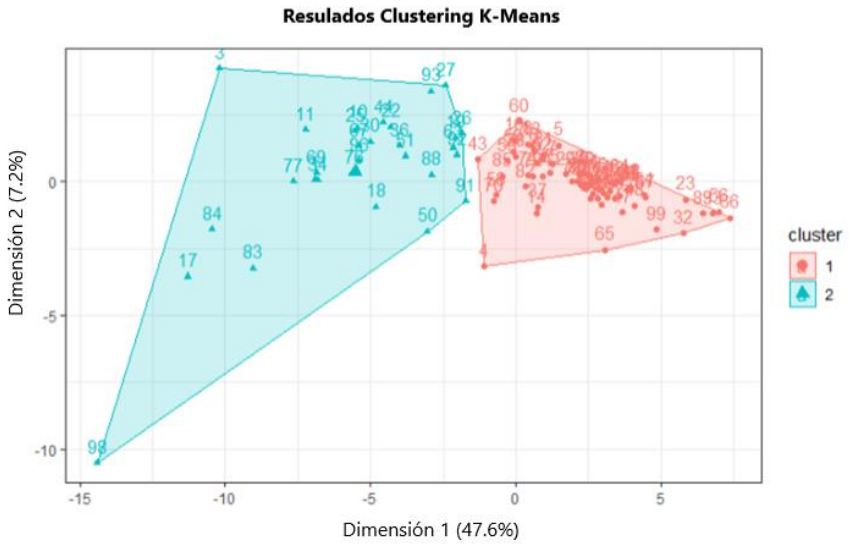


En este caso, se ve claramente en los Gráficos 29 – 30, para un valor de k=2, se obtiene el mejor promedio, por lo que este se convierte en el número de grupos óptimo. Ahora, a partir de la identificación del número de clúster optimo es posible aplicar el algoritmo con la cantidad de *k* seleccionada. La función de R *Fviz_cluster() [factoextrapackage*] se puede utilizar para visualizar fácilmente la agrupación de los clusters, que toma los resultados de k-medias y los datos originales como argumentos. En la gráfica resultante, las observaciones se representan por puntos, usando componentes principales si el número de variables es mayor que 2.

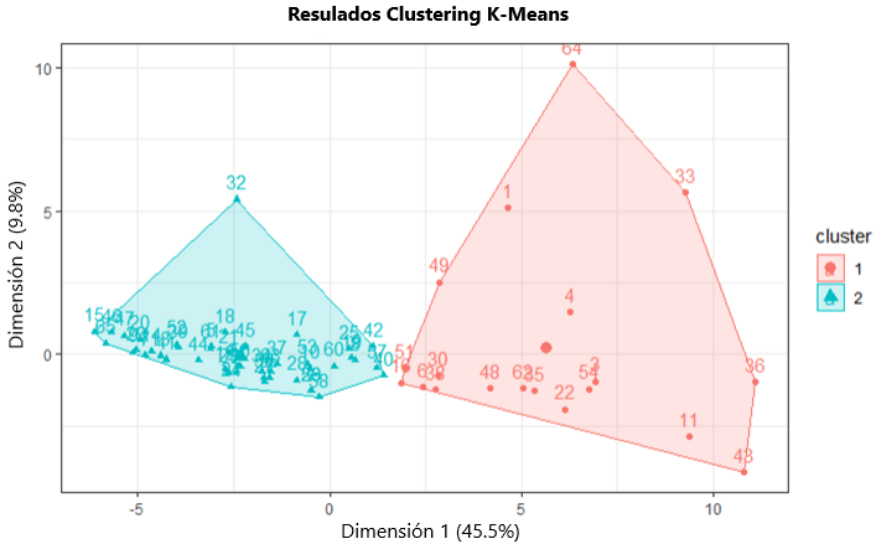
Es preciso mencionar que la función R *Fviz_cluster*, función del paquete factoextra, permite visualizar de manera fácil la agrupación de los clusters. Esta función toma como argumentos los resultados del algoritmo k-medias y los datos originales. Al aplicar la función *Fviz_cluster*(), se generará una gráfica que representa las observaciones en el espacio de las variables. En esta gráfica, cada observación se representa mediante un punto. Así, si el número de variables es mayor que 2, se utiliza una técnica llamada Análisis de Componentes Principales (PCA) para reducir la dimensionalidad de los datos. El PCA es una técnica estadística que permite sintetizar la información contenida en múltiples variables en un número menor de componentes principales que capturan la mayor parte de la variabilidad de los datos. Estos componentes principales se utilizan para representar las observaciones en un espacio bidimensional o tridimensional [55].

El uso de componentes principales en la visualización de la agrupación de clusters ayuda a comprender la estructura de los datos y la proximidad entre las observaciones en un espacio de menor dimensión, lo que facilita la interpretación de los resultados del clustering.

Gráfica 32. Segmentación K-means base de datos presuntos casos de violencia sexual por cada 100.000 habitantes



Gráfica 33. Segmentación K-means base de datos presuntos casos de violencia de pareja por cada 100.000 habitantes



Como resultado, en las gráficas anteriores se representa cada clúster con un color diferente y además se muestra la posición de cada centroide, en el grupo 1 con un círculo y en grupo 2 con un triángulo. Como puede verse, con k = 2 el modelo asigna clases consistentes a los datos de entrada, en especial al observar los agrupamientos.

- **RESULTADOS K MEANS VIOLENCIA SEXUAL**

Al entrar a identificar la diferencia entre los grupos es posible notar que en el grupo 1 se encuentran los departamentos con menor número de casos de violencia sexual por cada 100.000 habitantes en los tres años que contempla en análisis mientras que en el clúster 2 se encuentran los departamentos con mayor número de casos de violencia sexual por cada 100.000 habitantes. Se identifican que algunos departamentos se agruparon tanto en el clúster 1 como en clúster 2, dependiendo del año, entre ellos Caquetá, Huila, Meta, Quindío, Putumayo, Risaralda, Tolima y Vaupés.

Tabla 1. Clasificación clases K-means por departamento violencia sexual

Grupos K-means		
Año	**Clúster 1**	**Clúster 2**
2019	Antioquia, San Andrés, Providencia y Santa Catalina, Atlántico, Bogotá, Bolívar, Boyacá, Caldas, Cauca, Cesar, Chocó, Córdoba, Cundinamarca, La Guajira, Magdalena, Nariño, Norte de Santander, Santander, Sucre, Valle del Cauca, Vaupés, Vichada	Amazonas, Arauca, Caquetá, Casanare, Guainía, Guaviare, Huila, Meta, Putumayo, Quindío, Risaralda, Tolima
2020	Antioquia, San Andrés, Providencia y Santa Catalina, Atlántico, Bogotá, Bolívar, Boyacá, Caldas, Caquetá, Cauca, Cesar, Chocó, Córdoba, Cundinamarca, Huila, La Guajira, Magdalena, Meta, Nariño, Norte de Santander, Putumayo, Quindío, Risaralda, Santander, Sucre, Valle del Cauca, Vaupés, Vichada	Amazonas, Arauca, Casanare, Guainía, Guaviare, Tolima
2021	Antioquia, San Andrés, Providencia y Santa Catalina, Atlántico, Bogotá, Bolívar, Boyacá, Caldas, Cauca, Cesar, Chocó, Córdoba, Cundinamarca, Huila, La Guajira, Magdalena, Nariño, Norte de Santander, Santander, Sucre, Valle del Cauca, Vichada	Amazonas, Arauca, Caquetá, Casanare, Guainía, Guaviare, Meta, Putumayo, Quindío, Risaralda, Tolima, Vaupés

Basándonos en los resultados proporcionados sobre el análisis de K-means de la violencia sexual en Colombia, se pueden obtener las siguientes conclusiones:

-Diferenciación de grupos: El análisis de K-means permitió identificar dos grupos distintos. El clúster 1 contiene departamentos con un menor número de casos de violencia sexual por cada

100.000 habitantes durante los años 2019-2021, mientras que el clúster 2 agrupa a los departamentos con un mayor número de casos. Esta diferenciación resalta la variabilidad en la incidencia de violencia sexual en diferentes regiones del país.

-Características de las víctimas en el clúster 1: Dentro del clúster 1, se observa que las características de las víctimas de violencia sexual incluyen: estar en la adolescencia, tener un nivel educativo limitado (haber cursado únicamente la primaria), no tener una pertenencia étnica reconocida, ser agredidas por un miembro de su familia, ser víctimas de violencia de género (especialmente violencia sexual), sufrir la agresión en zonas urbanas, experimentarla en el primer semestre del año y ser agredidas durante la semana.

-Características de las víctimas en el clúster 2: En el clúster 2, donde se encuentran los departamentos con un mayor número de casos de violencia sexual, se observan características similares a las del clúster 1, pero con una incidencia aún más alta. Esto implica que en estos departamentos se presenta una mayor concentración de víctimas de violencia sexual, especialmente en la adolescencia, con niveles educativos limitados, sin una pertenencia étnica reconocida, siendo agredidas por miembros de su familia, siendo víctimas de violencia de género (en particular, violencia sexual), experimentando la agresión en zonas urbanas, durante el primer semestre del año y durante la semana.

- **RESULTADOS K MEANS VIOLENCIA DE PAREJA**

A continuación, se presentan los resultados del análisis de la violencia de pareja en dos años consecutivos: 2020 y 2021. Utilizando el método de K-means, se identificaron dos clústeres principales en cada año. El clúster 1 agrupó a los departamentos con mayor incidencia de casos por cada 100.000 habitantes, mientras que el clúster 2 incluyó a aquellos con menor incidencia. Por lo tanto, se analizan los cambios en la composición de estos clústeres entre los dos años, lo que proporciona información valiosa sobre la distribución de la violencia de pareja en Colombia.

Tabla 2. Clasificación clases K-means por departamento violencia de pareja

Grupos_Kmeans		
Año	**Clúster 1**	**Clúster 2**
2020	Amazonas, Arauca, San Andrés, Bogotá ,Casanare ,Cundinamarca ,Meta ,Tolima.	Antioquia, Atlántico, Bolívar, Boyacá, Caldas, Caquetá, Cauca, Cesar, Chocó, Córdoba, Guainía, Guaviare, Huila, La Guajira, Magdalena, Nariño, Norte de Santander, Putumayo, Quindío, Risaralda, Santander, Sucre, Valle del Cauca, Vaupés.
2021	Amazonas, Arauca, San Andrés, Bogotá, Casanare, Cundinamarca, Guainía, Huila, Meta, Tolima, Vaupés.	Antioquia, Atlántico, Bolívar, Boyacá, Caldas, Caquetá, Cauca, Cesar, Chocó, Córdoba, Guaviare, La Guajira, Magdalena, Nariño, Norte de Santander, Putumayo, Quindío, Risaralda, Santander, Sucre, Valle del Cauca, Vichada.

A partir de los resultados obtenidos mediante el método de K-medias para el análisis de la violencia de pareja en los años 2020 y 2021, se pueden extraer las siguientes conclusiones:

En el año 2020, se identificó que el clúster 1 agrupaba a los departamentos con la mayor cantidad de casos de violencia de pareja por cada 100.000 habitantes. Este clúster estaba compuesto por siete departamentos, incluyendo Amazonas, Arauca, San Andrés, Bogotá, Casanare, Meta y Tolima. Sin embargo, en el año 2021, el número de departamentos en este clúster aumentó a diez, añadiendo a Cundinamarca, Guainía y Vaupés. Este aumento sugiere un incremento en las tasas de víctimas en comparación con el año anterior.

Por otro lado, el clúster 2 englobaba a los departamentos con una menor cantidad de casos de violencia de pareja por cada 100.000 habitantes, sumando un total de 25 en el año 2020. Entre ellos, se encontraban Antioquia, Atlántico, Bolívar, Boyacá, Caldas, Caquetá, Cauca, Cesar, Chocó, Córdoba, Cundinamarca y Guainía, entre otros. Sin embargo, en el año 2021, el clúster 1 estuvo compuesto por solo 23 departamentos, ya que Cundinamarca, Guainía y Vaupés salieron de este grupo debido al aumento de sus tasas de víctimas. Además, el departamento de Vichada se añadió al Clúster 2, siendo el único con una tasa menor en comparación al año anterior.

En cuanto a las características del clúster 1, se observa que las mujeres adultas con educación secundaria y en estado civil de unión libre presentaron las mayores tasas de víctimas de violencia de pareja. En la mayoría de los casos, el agresor fue el compañero permanente de la víctima, y los celos se identificaron como el factor principal desencadenante. Además, se destaca que la mayoría de las víctimas no tenían ninguna pertenencia étnica específica, las agresiones ocurrieron principalmente en zonas urbanas durante el primer semestre del año y entre semana.

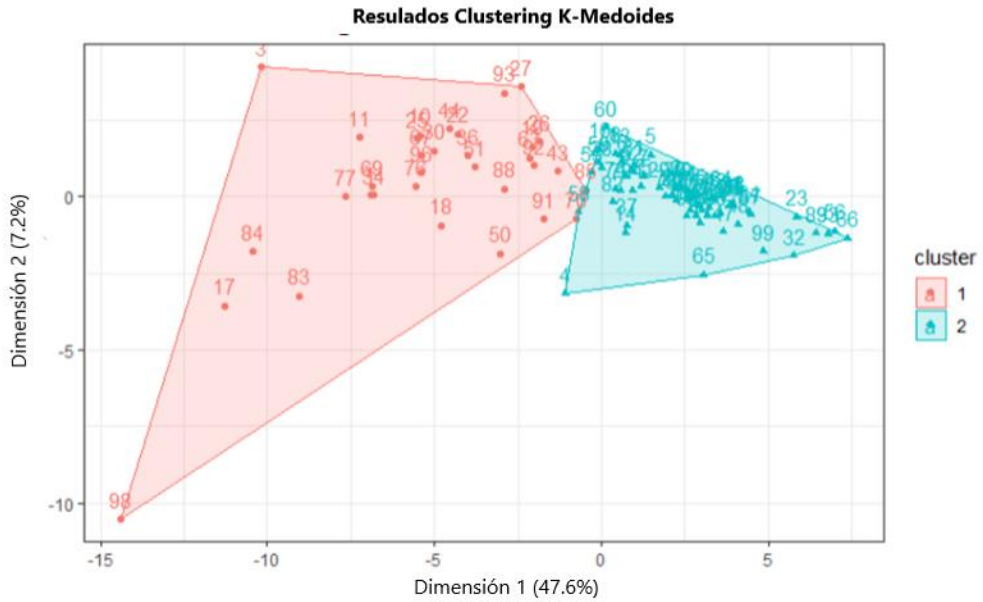
En el caso del clúster 2, se observa que las características son similares, con mujeres adultas, educación secundaria, estado civil de unión libre y agresor compañero permanente. Los celos también se identificaron como el principal factor en estos casos. Asimismo, la mayoría de las víctimas no tenían pertenencia étnica específica y las agresiones se registraron principalmente en zonas urbanas durante el primer semestre del año y entre semana.

6.2. K-MEDOIDES

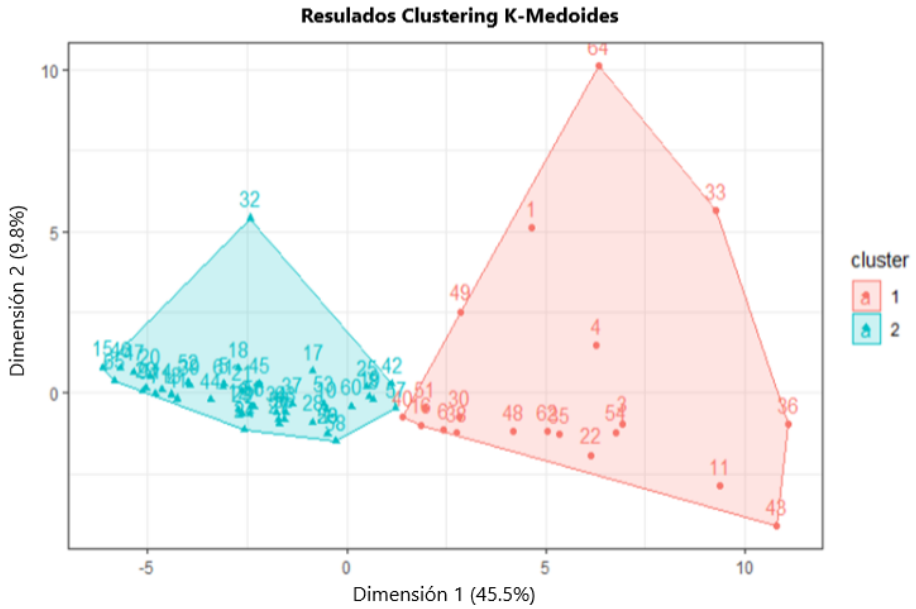
Como se explicó previamente en el marco teórico, K-medoides es una técnica clásica de segmentación de grupos que divide los datos conformados por n objetos en k grupos (con k conocido de antemano). En contraste con el algoritmo de k- medias, k -medoides elige puntos de datos reales como centros (medoides). Para la aplicación de este modelo, se utilizó la función pam() de R del del paquete [clusterpackage] y pamk() del paquete [fpcpackage]. Se seleccionó k=2, basándose previamente en la estimación óptima de clústeres a partir del método de la silueta.

Finalmente se utilizó como métrica la distancia euclidiana como distancia a utilizar. Las opciones disponibles son “euclidiana” y “manhattan” [56].

Gráfica 34. Segmentación K-medoides base de datos presuntos casos de violencia sexual por cada 100.000 habitantes



Gráfica 35. Segmentación K-medoides base de datos presuntos casos de violencia de pareja por cada 100.000 habitantes



- **RESULTADOS K MEDOIDES BASE DE DATOS VIOLENCIA SEXUAL**

Al entrar a identificar la diferencia entre los grupos es posible notar que, a diferencia de K-means, en el grupo 1 se concentraron los departamentos con mayor número de casos por cada 100.000 habitantes y en el clúster 2 los de menor número. En el clúster 1 se agruparon 33 observaciones y en el clúster 2, 66.

Tabla 3. Clasificación clases K-medoides por departamento violencia sexual

Grupos K-medoides violencia sexual		
Año	**Clúster 1**	**Clúster 2**
2019	Amazonas, Arauca, Caquetá, Casanare, Guainía, Guaviare, Huila, Meta, Putumayo, Quindío, Risaralda, Tolima	Antioquia, San Andrés, Providencia y Santa Catalina, Atlántico, Bogotá, Bolívar, Boyacá, Caldas, Cauca, Cesar, Chocó, Córdoba, Cundinamarca, La Guajira, Magdalena, Nariño, Norte de Santander, Sucre, Valle del Cauca, Vaupés, Vichada
2020	Amazonas, Arauca, Casanare, Guainía, Guaviare, Tolima, Caquetá	Antioquia, San Andrés, Providencia y Santa Catalina, Atlántico, Bogotá, Bolívar, Boyacá, Caldas, Cauca, Cesar, Chocó, Córdoba, Cundinamarca, Huila, La Guajira, Magdalena, Meta, Nariño, Norte de Santander, Putumayo, Quindío, Risaralda, Santander, Sucre, Valle del Cauca, Vaupés, Vichada
2021	Amazonas, Arauca, Caquetá, Casanare, Guainía, Guaviare, Meta, Putumayo, Quindío, Risaralda, Tolima, Vaupés	Antioquia, San Andrés, Providencia y Santa Catalina, Atlántico, Bogotá, Bolívar, Boyacá, Caldas, Cauca, Cesar, Chocó, Córdoba, Cundinamarca, Huila, La Guajira, Magdalena, Nariño, Norte de Santander, Santander, Sucre, Valle del Cauca, Vichada

A partir de los resultados anteriores, se observó que, a lo largo de los tres años (2019-2021) según K medoides, los departamentos se mantuvieron consistentes en su agrupación dentro de los dos clústeres. El clúster 1 incluyó consistentemente a los departamentos con un mayor número de casos de violencia sexual por cada 100.000 habitantes, mientras que el clúster 2 agrupó a los departamentos con un menor número de casos.

También, en los departamentos que conforman el clúster 1, con un mayor número de casos de violencia sexual, se encontraron características comunes en las víctimas. Estas características incluyen una alta proporción de mujeres abusadas sexualmente en la adolescencia, con niveles educativos limitados (únicamente primaria), ausencia de pertenencia étnica reconocida, agresiones sexuales perpetradas por miembros de la familia y una alta prevalencia de violencia de

género, especialmente violencia sexual. Además, las agresiones sexuales tendieron a ocurrir en zonas urbanas, durante el primer semestre del año y en días de semana.

Finalmente, en los departamentos que conforman el clúster 2, con un menor número de casos de violencia sexual, también se observaron características similares a las del clúster 1, pero con una incidencia más baja. Esto indica que estos departamentos presentan un menor nivel de violencia sexual, aunque las características de las víctimas, como la adolescencia, los niveles educativos limitados y la falta de pertenencia étnica reconocida, siguen siendo relevantes.

- **RESULTADOS K MEDOIDES VIOLENCIA DE PAREJA**

En cuanto a los resultados del análisis de la violencia de pareja en dos años consecutivos: 2020 y 2021. Mediante el método de k- medoides, se identificaron dos clústeres principales en cada año. El clúster 1 incluye los departamentos Amazonas, Arauca, San Andrés, Bogotá, Casanare, Cundinamarca, Meta, Tolima, Boyacá, Guainía, Huila y Vaupés, mientras que el clúster 2 comprende los departamentos Antioquia, Atlántico, Bolívar, Caldas, Caquetá, Cauca, Cesar, Chocó, Córdoba, Guaviare, La Guajira, Magdalena, Nariño, Norte de Santander, Putumayo, Quindío, Risaralda, Santander, Sucre, Valle del Cauca y Vichada. A continuación, examinaremos las características asociadas a cada uno de estos clústeres para obtener una visión más completa de la situación de la violencia de pareja en estos departamentos.

Tabla 4. Clasificación clases K-medoides por departamento violencia de pareja

Grupos_kmedoides		
Año	**clúster 1**	**clúster 2**
2020	**Amazonas, Arauca, San Andrés, Bogotá, Casanare, Cundinamarca, Meta, Tolima**.	Antioquia, Atlántico, Bolívar, Boyacá, Caldas, Caquetá, Cauca, Cesar, Chocó, Córdoba, Guainía, Guaviare, Huila, La Guajira, Magdalena, Nariño, Norte de Santander, Putumayo, Quindío, Risaralda, Santander, Sucre, Valle del Cauca, Vaupés.
2021	**Amazonas, Arauca, Boyacá, San Andrés, Bogotá, Casanare, Cundinamarca**, Guainía, Huila, **Meta**, **Tolima**, Vaupés.	Antioquia, Atlántico, Bolívar, Caldas, Caquetá, Cauca, Cesar, Chocó, Córdoba, Guaviare, La Guajira, Magdalena, Nariño, Norte de Santander, Putumayo, Quindío, Risaralda, Santander, Sucre, Valle del Cauca, Vichada.

En los departamentos que conforman el Clúster 1 (con mayor número de casos) se encontraron los siguientes hallazgos, según arrojó el modelo de K-medoides para los años 2020 y 2021:

Para el año 2020 se identificó que el clúster 1 incluía los departamentos con mayor cantidad de casos por cada 100.000 habitantes, sumando un total de ocho los cuales son Amazonas, Arauca, San Andrés, Bogotá, Casanare, Cundinamarca, Meta, Tolima. Sin embargo, para el año 2021, el clúster 1 aumento en 4 departamentos, conformado así por 12, ya que Boyacá, Guainía, Huila,

Vaupés se incluyeron en este clúster, lo que indica que hubo un incremento de sus tasas de víctimas.

Por otro lado, el clúster 2 agrupó los departamentos con menor cantidad de casos por cada 100.000 habitantes. En el año 2020 se conformó con 24 departamentos, incluyendo Boyacá, Guainía, Huila, Vaupés, Caldas, Caquetá, Cauca, Cesar, Chocó, Córdoba, Guainía, Guaviare. Sin embargo, para el año 2021, cabe destacar que el número de departamentos incluidos en este clúster disminuyó de un año a otro, lo que sugiere un posible aumento en las tasas de víctimas en los departamentos restantes.

En el clúster 1, se observa que las víctimas de violencia de pareja son predominantemente adultas, con educación secundaria y en estado civil de unión libre. En la mayoría de los casos, el agresor fue el compañero permanente de la víctima y los celos fueron el principal factor desencadenante. Además, se destaca que la mayoría de las víctimas no pertenecen a ningún grupo étnico específico, y las agresiones ocurrieron principalmente en zonas urbanas durante el segundo semestre del año y entre semana.

En el clúster 2, se observan características similares: las víctimas también son en su mayoría adultas, con educación secundaria y en estado civil de unión libre. El agresor suele ser el compañero permanente y los celos son el factor principal en estos casos. Asimismo, la mayoría de las víctimas no tienen ninguna pertenencia étnica específica, y las agresiones ocurrieron principalmente en zonas urbanas durante el segundo semestre del año.

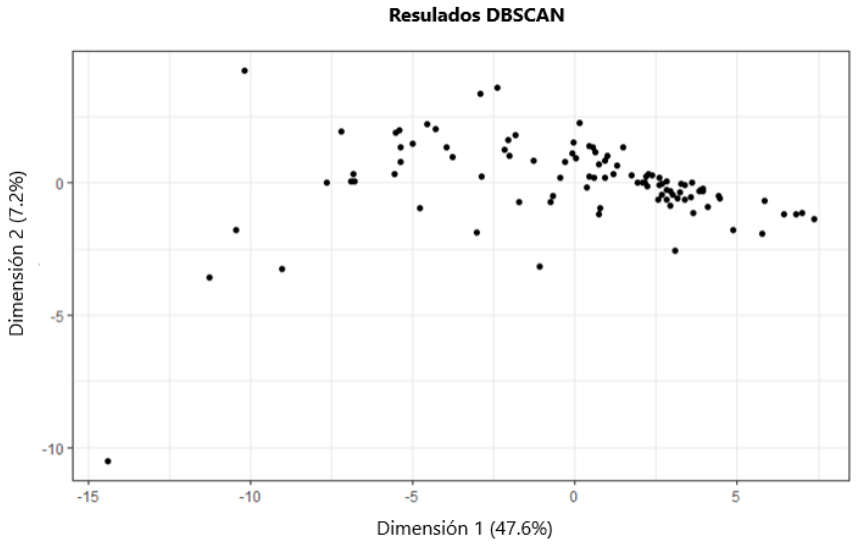
En resumen, tanto en el clúster 1 como en el clúster 2, se identifican similitudes en cuanto a las características de las víctimas de violencia de pareja. Sin embargo, es importante destacar que en el clúster 2 se observa una reducción en el número de departamentos, lo que indica que en los lugares restantes podría haber un aumento en las tasas de víctimas.

7.3. DBSCAN

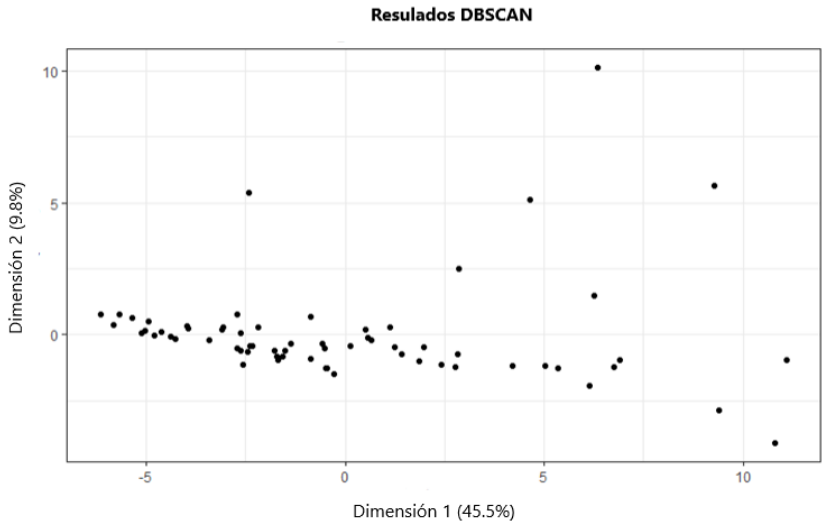
Este es un método de cauterización adecuado para buscar patrones de agrupación y es mayormente utilizado cuando los métodos jerárquicos no funcionan del todo bien debido al ruido y los valores atípicos. Este algoritmo agrupa las observaciones que están más cercanas respecto a alguna métrica, comúnmente se utiliza la distancia euclidiana, en el que cada clúster contendrá un mínimo de observaciones. Este método tiene sus ventajas y desventajas respecto otros métodos de clusterización como K-means y K-medoides, por ejemplo. Entre las ventajas se encuentra que DBSCAN no necesita que se especifique el número de clústeres, además que mediante este algoritmo se garantiza que cada clúster tendrá un mínimo número de observaciones. Por el lado de las desventajas, se tiene que DBSCAN no funciona bien en clústeres de diferentes densidades, además de que los parámetros deben ser elegidos con mayor cuidado o precisión [57].

Para su aplicación en R se utiliza la librería DBSCAN, posteriormente este método necesita sólo dos parámetros: eps, que es la distancia que se tomará como radio de los clústeres, y minPoints, que es el mínimo de números que deben estar en un grupo para que el algoritmo lo tome como un clúster. El parámetro Epsilon se definió en 0.15. Para obtener el eps, se emplea la función kNNdistplot:

Gráfica 36. Segmentación modelo DBSCAN base de datos violencia sexual por cada 100.000 habitantes



Gráfica 37. Segmentación modelo DBSCAN base de datos violencia de pareja por cada 100.000 habitantes



Al aplicar el algoritmo de agrupamiento DBSCAN a las bases de datos de violencia sexual y violencia de pareja, se observa que el modelo no logra segmentar adecuadamente las diferencias entre los grupos, ya que no se registran observaciones en ninguno de ellos. Esto puede deberse a que los clústeres tienen diferentes densidades, lo que hace que la técnica utilizada no sea apropiada en este caso. Es necesario explorar otras técnicas de agrupamiento para poder identificar patrones significativos en los datos.

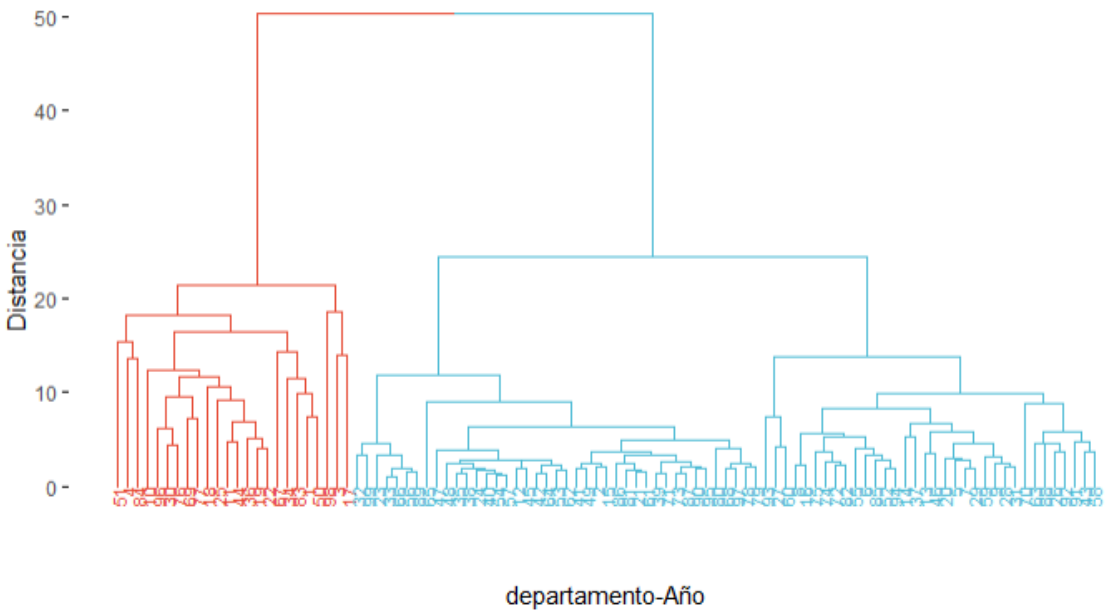
6.3. MÉTODO WARD

Una de las mayores dificultades al agrupar elementos es encontrar el número apropiado de clústeres, por lo que los métodos jerárquicos construyen una estructura en la que los elementos se agrupan en subconjuntos cada vez mayores hasta que todos pertenecen al mismo conjunto. De esta forma, no se muestra un agrupamiento sino las relaciones de proximidad que existen entre los elementos [58].

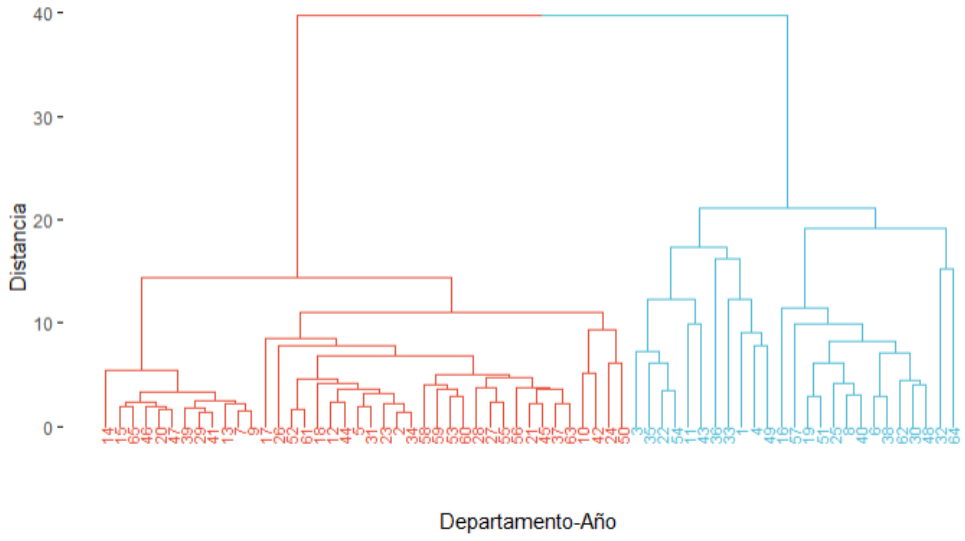
El Método de Ward, por lo tanto, es un método que permite construir un árbol de clasificación o dendograma. Un dendograma es una representación gráfica con forma de árbol que sirve para resumir el proceso de agrupación del análisis de clústeres. Este método se basa en la suma de cuadrados y permite crear grupos de tamaño similar, dando paso a realizar buenos análisis de varianza por la producción de clústeres definidos. [59].

Para la aplicación del método de Ward en R se utilizó la librería clúster. En esta última hay varias funciones disponibles en R para clústeres jerárquicos, en este caso, se utilizó la función 'hclust', que determina los valores de distancia que se pueden calcular en R utilizando la función 'dist'. La medida predeterminada para la función dist es 'Euclidiana', adicionalmente se definió el método de vinculación, en este caso, "ward.D2". Finalmente, se asignan los clústeres a los puntos de datos especificando el número deseado de grupos (k), en este caso k=2, basándose previamente en el método de la silueta y el codo. Como resultado, mediante la librería Factoextra, se grafica el dendograma:

Gráfica 38. Agrupamiento método de Ward base de datos violencia sexual por cada 100.000 habitantes



Gráfica 39. Agrupamiento método de Ward base de datos violencia de pareja por cada 100.000 habitantes



- **RESULTADOS AGRUPAMIENTO MÉTODOS DE WARD BASE DE DATOS DE VIOLENCIA SEXUAL**

Al aplicar este modelo a la base de datos de violencia sexual se observa que el dendograma muestra la agrupación de dos grupos, en el grupo 1 se encuentran 24 observaciones que, al igual que K-medoides corresponde a los departamentos con mayor número de casos de violencia sexual, mientras que en el grupo 2 se agrupan 75 observaciones que corresponde a los departamentos con menor número de casos.

Tabla 5. Clasificación clases método de Ward por departamento violencia sexual

Año	Clúster 1	Clúster 2
2019	**Amazonas**, **Arauca,** San Andrés, Providencia y Santa Catalina, Caquetá, Casanare, **Guainía**, **Guaviare**, Huila, Meta, Putumayo, Tolima	Antioquia, Atlántico, Bogotá, Bolívar, Boyacá, Caldas, Cauca, Cesar, Chocó, Córdoba, Cundinamarca, La Guajira, Magdalena, Nariño, Norte de Santander, Quindío, Risaralda, Santander, Sucre, Valle del Cauca, Vaupés, Vichada
2020	**Amazonas, Arauca,** Casanare**, Guainía, Guaviare,**	Antioquia, San Andrés, Providencia y Santa Catalina, Atlántico, Bogotá, Bolívar, Boyacá, Caldas, Caquetá, Cauca, Cesar, Chocó, Córdoba, Cundinamarca, Huila, La Guajira, Magdalena, Meta, Nariño, Norte de Santander, Putumayo, Quindío, Risaralda, Santander, Sucre, Tolima, Valle del Cauca, Vaupés, Vichada
2021	**Amazonas, Arauca,** Caquetá, Casanare, **Guainía**, **Guaviare,** Tolima, Vaupés.	Antioquia, San Andrés, Providencia y Santa Catalina, Atlántico, Bogotá, Bolívar, Boyacá, Caldas, Cauca, Cesar, Chocó, Córdoba, Cundinamarca, Huila, La Guajira, Magdalena, Meta, Nariño, Norte de Santander, Putumayo, Quindío, Risaralda, Santander, Sucre, Valle del Cauca, Vichada

De esta manera, vemos que mediante este método solo cuatro departamentos (Amazonas, Arauca, Guainía y Guaviare) son clasificados en los tres años de análisis en el clúster 1 correspondiente al clúster que agrupa el mayor número de casos por cada 100.000 habitantes, mientras departamentos como San Andrés, Providencia y Santa Catalina, Caquetá, Casanare y Tolima son agrupados en el clúster 1 o clúster 2, según su número de casos.

Al igual que en los resultados de K-means y K-medoides, los departamentos se agruparon en dos clústeres. Sin embargo, hubo variaciones en la composición de los grupos a lo largo de los años, lo que indica cambios en los patrones de violencia sexual en diferentes regiones.

En el clúster 1, conformado por departamentos con un mayor número de casos de violencia sexual, se observó una mayor proporción de mujeres abusadas sexualmente en la adolescencia y con bajo nivel educativo (únicamente primaria). Además, hubo una mayor incidencia de violencia sexual en mujeres sin ninguna pertenencia étnica. En el clúster 1, también se registró una proporción significativa de mujeres agredidas sexualmente por algún familiar y una mayor incidencia de violencia sexual en zonas urbanas.

- **RESULTADOS MÉTODO DE WARD BASE DE DATOS DE VIOLENCIA DE PAREJA**

En este estudio, se utilizó el método de Ward para analizar la violencia de pareja en Colombia durante los años 2020 y 2021. Se identificaron dos clústeres que agrupan los departamentos del país en función de la cantidad de casos por cada 100.000 habitantes.

El clúster 1 incluyó los departamentos con mayor incidencia de violencia de pareja, con un total de 12 departamentos en 2020 y 13 en 2021. Por otro lado, el clúster 2 agrupó los departamentos con menor cantidad de casos con un total de 24 para el año 2020 y 24 para el año 2021. Aunque el número de departamentos se mantuvo igual en ambos años, se añadió Vichada y se excluyó Guainía en 2021. Esto sugiere una disminución en las tasas de víctimas en Vichada en comparación con el año anterior.

Tabla 6. Clasificación clases método de Ward por departamento violencia de pareja

Método de Ward		
Año	**clúster 1**	**clúster 2**
2020	Amazonas, Arauca, Bogotá, Casanare, Cundinamarca, Meta, San Andrés, Tolima.	Antioquia, Atlántico, Bolívar, Boyacá, Caldas, Caquetá, Cauca, Cesar, Chocó, Córdoba, Guainía, Guaviare, Huila, La Guajira, Magdalena, Nariño, Norte de Santander, Putumayo, Quindío, Risaralda, Santander, Sucre, Valle del Cauca, Vaupés.
2021	Amazonas, Arauca, Bogotá, Boyacá, Casanare, Cundinamarca, Guainía, Huila, Meta, San Andrés, Tolima, Vaupés.	Antioquia, Atlántico, Bolívar, Caldas, Caquetá, Cauca, Cesar, Chocó, Córdoba, Guaviare, La Guajira, Magdalena, Nariño, Norte de Santander, Putumayo, Quindío, Risaralda, Santander, Sucre, Valle del Cauca, Vichada.

Al aplicar el método de Ward para analizar la violencia de pareja en los años 2020 y 2021, se identificaron dos clústeres distintos. El clúster 1 agrupó los departamentos con mayor cantidad de casos por cada 100.000 habitantes, y se observó un incremento de un departamento de un año a otro. En el año 2020, este clúster estaba conformado por 12 departamentos, incluyendo Vaupés, Putumayo, Amazonas, Boyacá, Huila, Cundinamarca, San Andrés, Tolima, Bogotá, Meta, Arauca y Casanare. Para el año 2021, Guainía fue incluido en este clúster, lo que indica un aumento en las tasas de víctimas en este departamento.

Por otro lado, el clúster 2 agrupó los departamentos con menor cantidad de casos por cada 100.000 habitantes. En el año 2020, este clúster estaba compuesto por 20 departamentos,

incluyendo Antioquia, Atlántico, Bolívar, Caldas, Caquetá, Cauca, Cesar, Chocó, Córdoba, Guainía, Guaviare, La Guajira, Magdalena, Nariño, Norte de Santander, Quindío, Risaralda, Santander y Sucre. En el año 2021, el número de departamentos en este clúster se mantuvo igual, pero se incluyó al departamento de Vichada y se excluyó a Guainía. Esto sugiere una disminución en la tasa de víctimas en el departamento de Vichada en comparación con el año anterior.

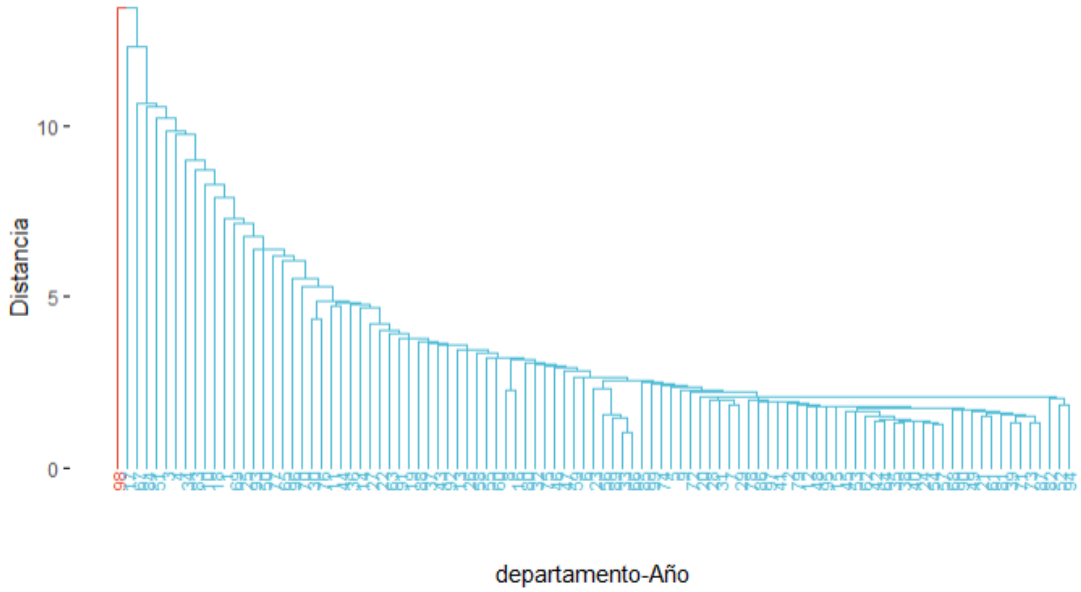
El clúster 1, compuesto por los departamentos con mayor cantidad de casos por cada 100.000 habitantes, muestra algunas características comunes. Se observa que las víctimas de violencia de pareja en este clúster tienden a ser mujeres adultas, con niveles educativos hasta secundaria, y se encuentran en unión libre como estado civil predominante. Además, el agresor más frecuente en este clúster es el compañero permanente de la víctima, y el factor principal asociado a la violencia son los celos. Por otro lado, la mayoría de las víctimas en este clúster no tienen pertenencia étnica, son agredidas en zonas urbanas y principalmente durante el segundo semestre del año, así como entre semana.

El clúster 2, que agrupa los departamentos con menor cantidad de casos por cada 100.000 habitantes, presenta características distintas a las del clúster 1. En este clúster, las víctimas de violencia de pareja también son mayormente mujeres adultas, pero con un nivel educativo hasta secundaria. El estado civil predominante es la unión libre, y el agresor principal es el compañero permanente. Sin embargo, en este clúster, el factor principal asociado a la violencia son también los celos. Al igual que en el clúster 1, la mayoría de las víctimas no tienen pertenencia étnica y son agredidas en zonas urbanas. Sin embargo, en este caso, las agresiones ocurroeron principalmente durante el primer semestre del año y entre semana.

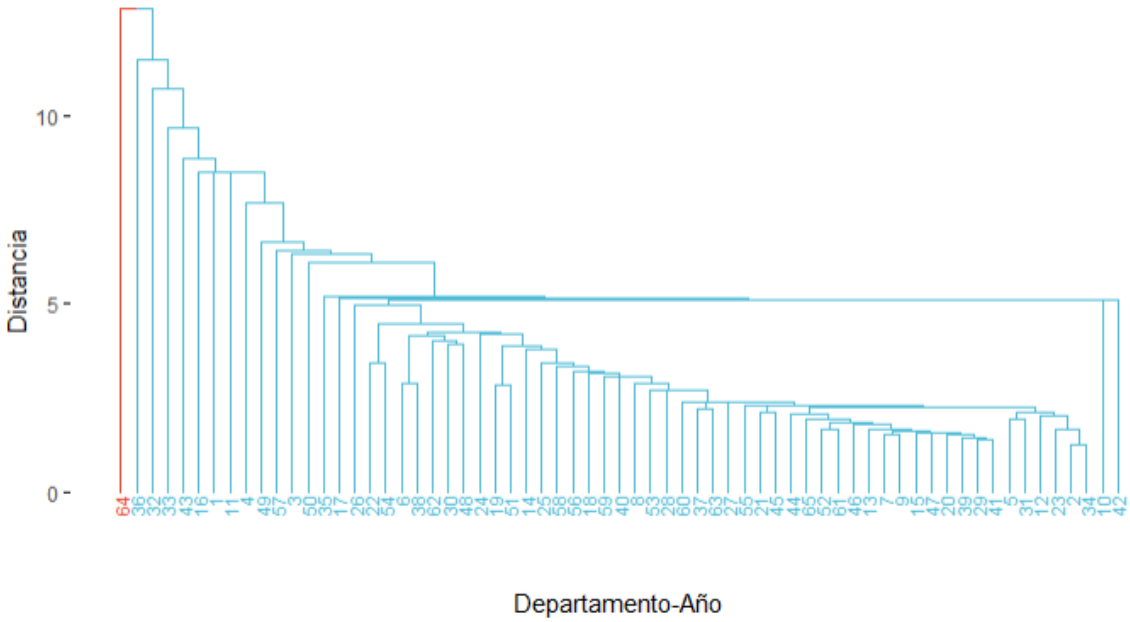
6.4. K-VECINOS MÁS CERCANOS

Finalmente, K-vecinos más cercanos es un algoritmo que esencialmente funciona para clasificar valores buscando los puntos de datos más similares. De esta manera, calcula la distancia entre el item a clasificar y el resto de items del dataset de entrenamiento y, posteriormente, selecciona los “k” elementos más cercanos (con menor distancia, según la función que se use. Al registro no etiquetado se le asigna la clase de la mayoría de los k vecinos más cercanos. Para que este método funcione bien es importante la selección de un k apropiado pues este terminará casi por definir a qué grupo pertenecerán los puntos, sobre todo en las “fronteras” entre grupos [60].

Gráfica 40. Agrupamiento método K-vecino más cercano base de datos violencia sexual por cada 100.000 habitantes



Gráfica 41. Agrupamiento método K-vecino más cercano base de datos violencia de pareja por cada 100.000 habitantes



Al aplicar el modelo de K-vecino más cercano a las bases de datos de violencia sexual y violencia de pareja, se ha observado que los resultados no son óptimos en términos de la segmentación adecuada de los grupos y su jerarquización. De hecho, se ha detectado que solo una observación en ambos casos pertenece a un grupo diferente, lo cual dificulta la identificación de distintos niveles de agrupamiento. Esto puede deberse a que las características de los datos no se ajustan adecuadamente a los supuestos del modelo, como, por ejemplo, la homogeneidad en la densidad de los clústeres. Por lo tanto, sería necesario explorar otras técnicas de análisis que se ajusten mejor a las características de estos datos.

6.5. CONCLUSIONES

Como se observó anteriormente existen distintos métodos de aprendizaje no supervisado que permiten encontrar grupos similares en el conjunto de datos sin etiquetas o clases previamente definidas. En el caso los datos de violencia sexual y violencia de pareja, los métodos de clasificación no jerárquica, como K-means, K-medoides, permiten identificar las distintas agrupaciones por departamento a partir de la agrupación de clústeres, entre aquellos con mayor y menor número de casos.

Sin embargo, para su funcionamiento estas dos técnicas requieren previamente definir el número de clústeres y esta decisión puede afectar seriamente los resultados. Además, como la ubicación de los centroides iniciales es aleatoria, los resultados pueden no ser comparables y mostrar una falta de consistencia.

A su vez, tras la aplicación del método Ward a la base de datos se observan dos grupos bien diferenciados, los departamentos con mayor y menos número de casos. Sin embargo, al igual que K-means y K-Medoides, este método depende también de la identificación exitosa del número de clústeres. De otro lado, se observó que el caso de la técnica de DBSCAN el modelo no segmentó adecuadamente la diferencia entre los grupos con k=2, mientras que con el modelo de K-vecino más cercano no se delimitaron tampoco los diferentes grupos.

En consecuencia, se buscó un modelo optimo que permitió identificar el número de clúster más adecuado, a la vez que segmente eficientemente los grupos para de esta forma observar las distintas tendencias de la violencia sexual y de pareja en Colombia.

7. PATRONES, CLASIFICACIONES O AGRUPACIONES DE VIOLENCIA SEXUAL Y DE PAREJA A PARTIR DE TÉCNICAS DE APRENDIZAJE NO SUPERVISADO- PAQUETE FACTOCLASS DE R.

Según Pardo y Campo, el paquete de R, FactoClass combina métodos factoriales con análisis de conglomerados o clasificación no supervisada. De esta manera, primero, realiza un análisis factorial según la naturaleza de los datos y luego una clasificación basada en un algoritmo mixto: clasificación jerárquica con el método de Ward y agregación alrededor de centros móviles (K-medias). Finalmente se obtiene una partición del conjunto de datos y la caracterización de cada una de las clases, según las variables activas e ilustrativas, ya sean cuantitativas o cualitativas [61]. La librería FactoClass en R es especialmente útil para el análisis de datos complejos y heterogéneos. También es útil para el análisis de datos en ciencias sociales y de marketing, así como en otros campos que requieren análisis de clasificación y ordenamiento de datos. En resumen, FactoClass es una librería en R que permite realizar análisis de clasificación y ordenamiento de datos, lo que es útil para identificar grupos homogéneos en una matriz de datos y analizar la estructura de relaciones entre variables[62].

En el caso concreto de este proyecto de investigación, FactoClass es una librería especialmente útil por su flexibilidad, dado que permite realizar diferentes tipos de análisis, incluyendo análisis factorial de correspondencia, análisis de componentes principales, análisis de correspondencia jerárquica y análisis de conglomerados. Esto la hace útil para analizar diferentes tipos de datos de violencia sexual y de pareja y responder nuestra pregunta de investigación. Así mismo, permite la visualización de los resultados de los análisis en gráficos y tablas de manera fácil y rápida lo que contribuye a una mejor comprensión de los resultados y facilita su interpretación.

- **RESULTADOS MÉTODO DE FACTOCLASS VIOLENCIA SEXUAL**

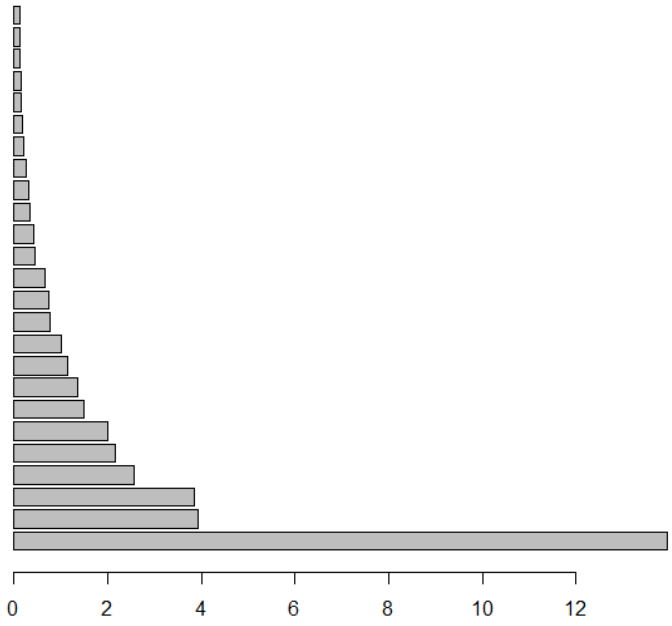
Según Pardo y Campo, el paquete FactoClass de R aprovecha las ventajas del método de Ward y las del K-medias, combinándolos de la manera siguiente:

1) *Clasificación inicial:* mediante esta clasificación inicial se busca obtener rápidamente y a bajo costo una partición de los individuos en clases homogéneas y se emplea el algoritmo de agregación alrededor de centros móviles (K-medias) [63].

En el caso concreto de la base de datos de violencia sexual, Factoclass realiza una clasificación inicial, en la que permite identificar el número de clases deseadas para la partición, decisión que se toma observando el diagrama de índices de nivel. En el gráfico de índices de nivel es más fácil observar los cambios de inercia más grandes (saltos) y decidir el número de clases K, en este caso se seleccionan 2 clusters. El primer análisis se lleva a cabo con los datos del año 2019, filtrando por

sexo – mujer, lo que permite filtrar únicamente los casos de violencia sexual contra las mujeres en Colombia.

Gráfica 42. gráfico de índices de nivel base de datos violencia sexual año 2019



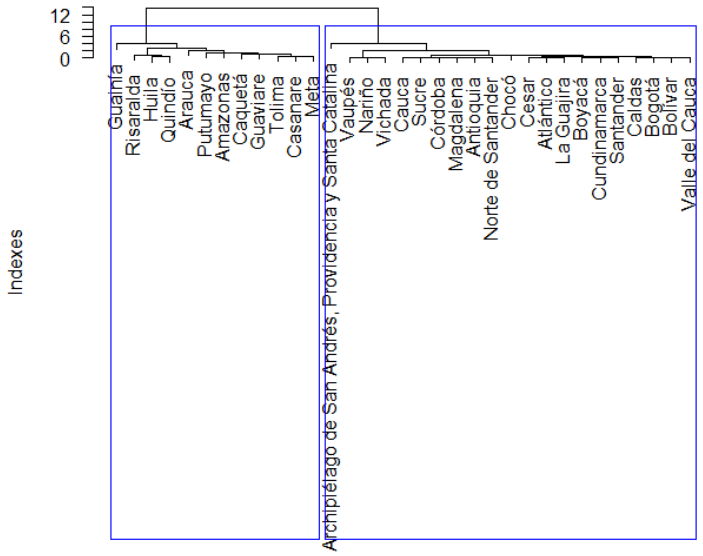
2) *Agregación jerárquica con el método de Ward*. En esta etapa, Factoclas efectúa una clasificación ascendente jerárquica donde los elementos terminales del árbol son las s clases de la partición inicial o los individuos directamente. En el caso de la base de datos de violencia sexual, se identifica la partición en dos clases a partir de los casos registrados por cada departamento.

Tabla 7. División de clases por departamento Factoclass base de datos violencia sexual -2019

Clase 1	Amazonas, Arauca, Caquetá, Casanare, Guaviare, Guainía, Meta, Putumayo, Tolima,
Clase 2	Antioquia, San Andrés y Providencia Atlántico, Atlántico, Bogotá, Bolívar, Boyacá, Caldas, Cauca, Cesar, Chocó, Córdoba, Cundinamarca, Huila, La Guajira, Magdalena, Nariño, Norte de Santander, Quindío, Risaralda, Santander, Sucre, Valle del Cauca, Vaupés, Vichada

4. *Corte del árbol*: El árbol o dendrograma que resume el procedimiento de clasificación permite ver la estructura de clases de los individuos que son objeto de análisis.

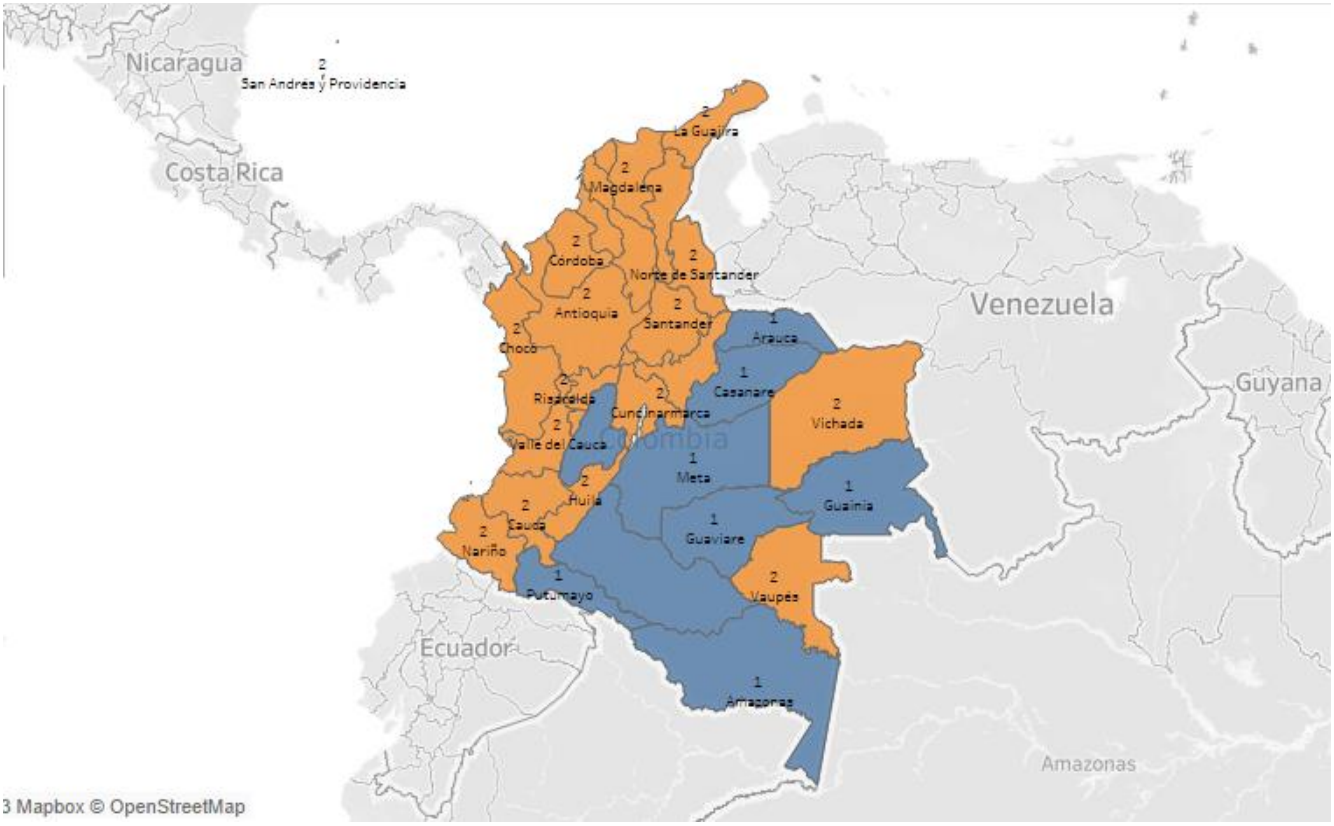
Gráfica 43. Agrupación departamentos mediante Factoclass base de datos violencia sexual 2019



4. Consolidación de la clasificación: De acuerdo con Pardo y Campo, la partición obtenida en el paso anterior no es óptima siempre, debido a la estructura de particiones anidadas del dendrograma obtenido. Para mejorarla se utiliza de nuevo un procedimiento de agregación alrededor de centros móviles (K-medias), utilizando los centros de gravedad de las clases obtenidas al cortar el árbol como centros iniciales [64]. En este caso, la consolidación de la clasificación arroja los resultados disponibles en el **Anexo 2**.

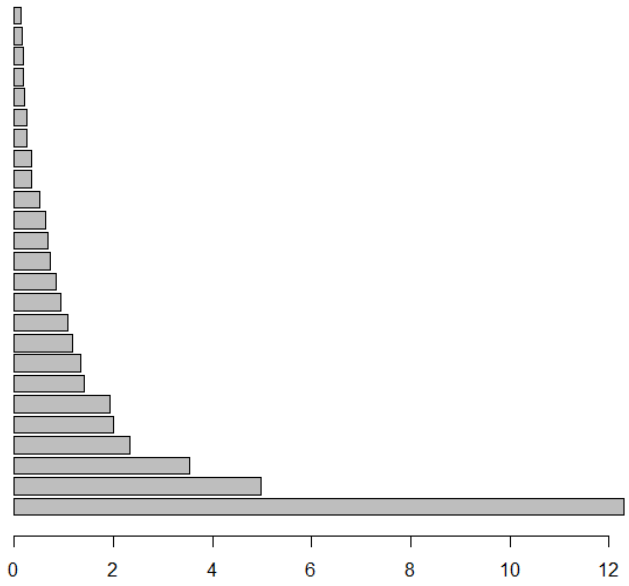
En el siguiente mapa es posible observar la división de clases a partir de la clasificación realizada por Factoclass. Se observa, por ejemplo, que los departamentos con mayor número de casos, los del clúster 1, se ubican sobre todo en la parte sur del país.

Mapa 1. Georreferenciación división de clases violencia sexual – 2019



Ahora, se continúa con el análisis utilizando la librería de Factoclass para los datos del año **2020** de violencia sexual en Colombia. Al respecto, Factoclass realiza una división inicial y mediante el gráfico de índices se decide el número de clases K, en este caso se seleccionan 3 clases.

Gráfica 44. Índices de nivel base de datos violencia sexual año 2020



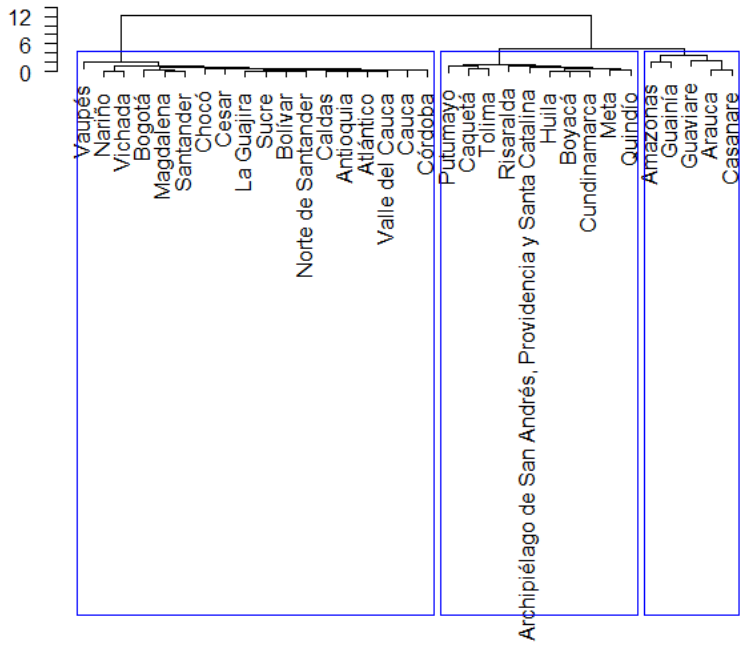
Factoclass continúa con una agregación jerárquica mediante el método de Ward. En el caso de la base de datos de violencia sexual para el año 2020, se identifica la partición en tres clases a partir de los casos registrados por cada departamento.

Tabla 8. División de clases por departamento factoclass base de datos violencia sexual -2020

Clase 1	Amazonas, Arauca, Casanare, Guainía Guaviare
Clase 2	Antioquia, Atlántico, Bogotá, Bolívar, Boyacá Caldas, Cauca, Cesar, Chocó, Córdoba, Cundinamarca, La Guajira, Magdalena Nariño, Norte de Santander, Santander Sucre, Valle del Cauca, Vaupés, Vichada.
Clase 3	San Andrés, Providencia y Santa Catalina, Caquetá, Huila, Meta, Putumayo, Quindío, Risaralda, Tolima.

Factoclass ilustra el dendrograma que resume el procedimiento de clasificación y permite ver la estructura de clases de los individuos que son objeto de análisis para el año 2020.

Gráfica 45. Agrupación departamentos mediante Factoclass base de datos violencia sexual 2020



Consolidación de la clasificación: la partición obtenida en el paso anterior arroja los siguientes resultados para el año 2020. Los departamentos de la clase 1 (Amazonas, Arauca, Casanare, Guainía Guaviare) se destacan, en su mayoría, por tener el mayor número de casos en 2020 en comparación de los departamentos de la clase 2 y 3. En el **anexo 3** se incluyen los resultados de la consolidación de la clasificación para el clúster 1 en el año 2020.

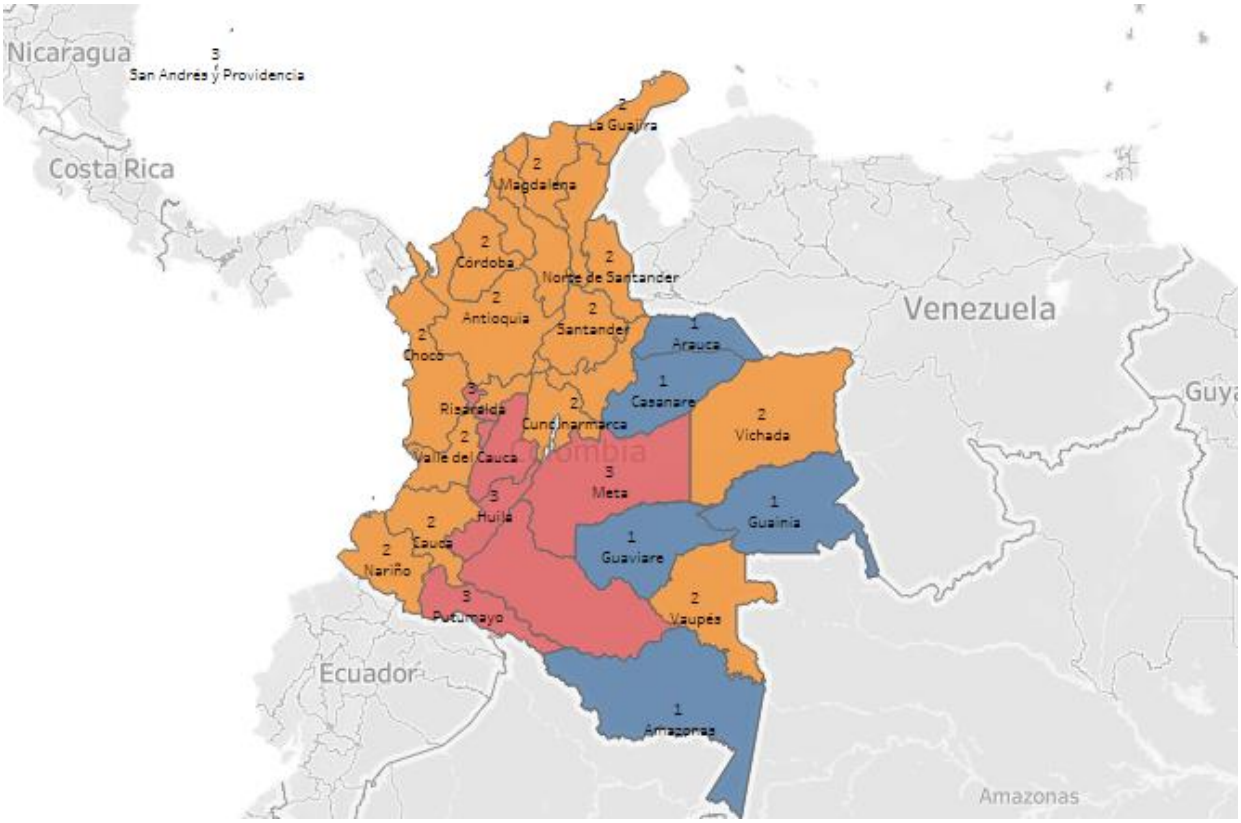
Los departamentos de la clase 2 (Antioquia, Atlántico, Bogotá, Bolívar, Boyacá, Caldas, Cauca, Cesar, Chocó, Córdoba, Cundinamarca, La Guajira, Magdalena, Nariño, Norte de Santander, Santander, Sucre, Valle del Cauca, Vaupés, Vichada) se destacan, en su mayoría, por tener un menor número de casos en 2020 en comparación de los departamentos de la clase 2 y 3. En el **anexo 3** se incluyen los resultados de la consolidación de la clasificación para el clúster 2 en el año 2020

Los departamentos de la clase 3 (San Andrés, Providencia y Santa Catalina, Caquetá, Huila, Meta, Putumayo, Quindío, Risaralda, Tolima) se destacan, en su mayoría, por tener un menor número de casos en 2020 en comparación de los departamentos de la clase 1 y 2. En el **anexo 3** se incluyen los resultados de la consolidación de la clasificación para el clúster 3 en el año 2020.

En el siguiente mapa es posible observar la división de clases a partir de la clasificación realizada por Factoclass en el año 2020. Se observa, por ejemplo, que los departamentos del clúster 1 con mayor número de casos por cada 100.000 habitantes, al igual que en el año 2019, correspondieron a Arauca, Casanare, Guaviare, Guainía y Amazonas. A diferencia, del año 2019, en el año 2020

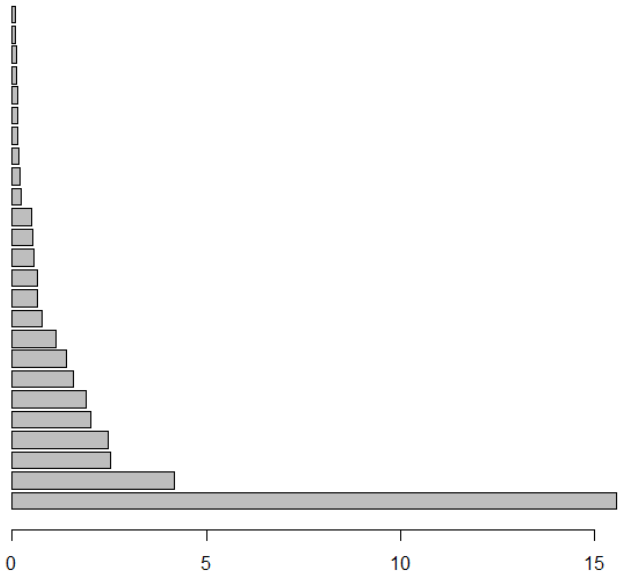
vemos que Factoclass realiza la clasificación en el clúster 3 de los departamentos de San Andrés, Providencia y Santa Catalina, Caquetá, Huila, Meta, Putumayo, Quindío, Risaralda y Tolima, seleccionando ocho variables que no tienen una relevancia significativa en comparación con las tendencias que muestran la clase 1 y 2.

Mapa 2. Georreferenciación división de clases violencia sexual – 2020



Finalmente, se continúa con el análisis para los datos del año 2021 de violencia sexual en Colombia. Al respecto, Factoclass realiza una división inicial y mediante el gráfico de índices se decide el número de clases K, en este caso se seleccionan 2 clases.

Gráfica 46. Gráfico de índices de nivel base de datos violencia sexual año 2021



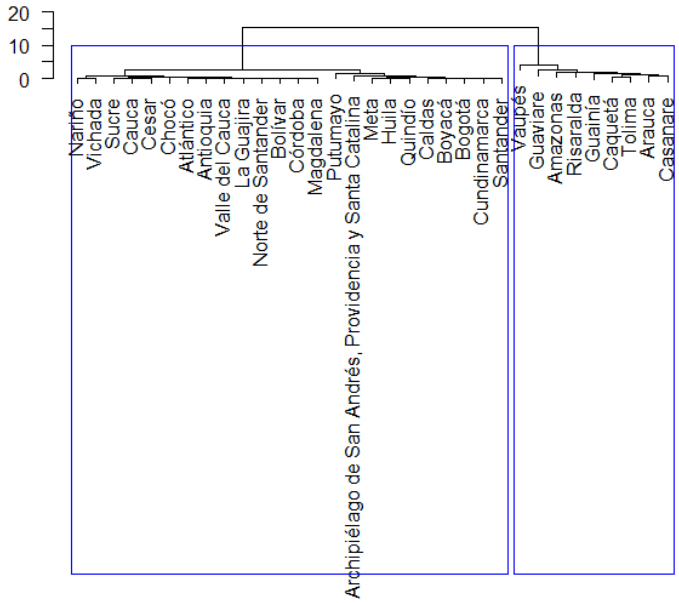
Factoclass continúa con una agregación jerárquica mediante el método de Ward. En el caso de la base de datos de violencia sexual para el año 2021, se identifica la partición en tres clases a partir de los casos registrados por cada departamento.

Tabla 9. División de clases por departamento factoclass base de datos violencia sexual -2021

Clase 1	Amazonas, Arauca, Caquetá, Casanare, Guainía, Guaviare, Huila, Meta, Putumayo, Quindío, Risaralda, Tolima
Clase 2	Antioquia, Caldas, Cundinamarca, Santander, Atlántico, Cauca, La Guajira, Sucre, Bogotá, Cesar, Magdalena, Valle del Cauca, Bolívar, Chocó, Nariño, Norte de Santander, Vaupés, Boyacá, Córdoba, Vichada.

Factoclass ilustra el dendrograma que resume el procedimiento de clasificación y permite ver la estructura de clases de los individuos que son objeto de análisis para el año 2021.

Gráfica 47. Agrupación departamentos mediante Factoclass base de datos violencia sexual 2021

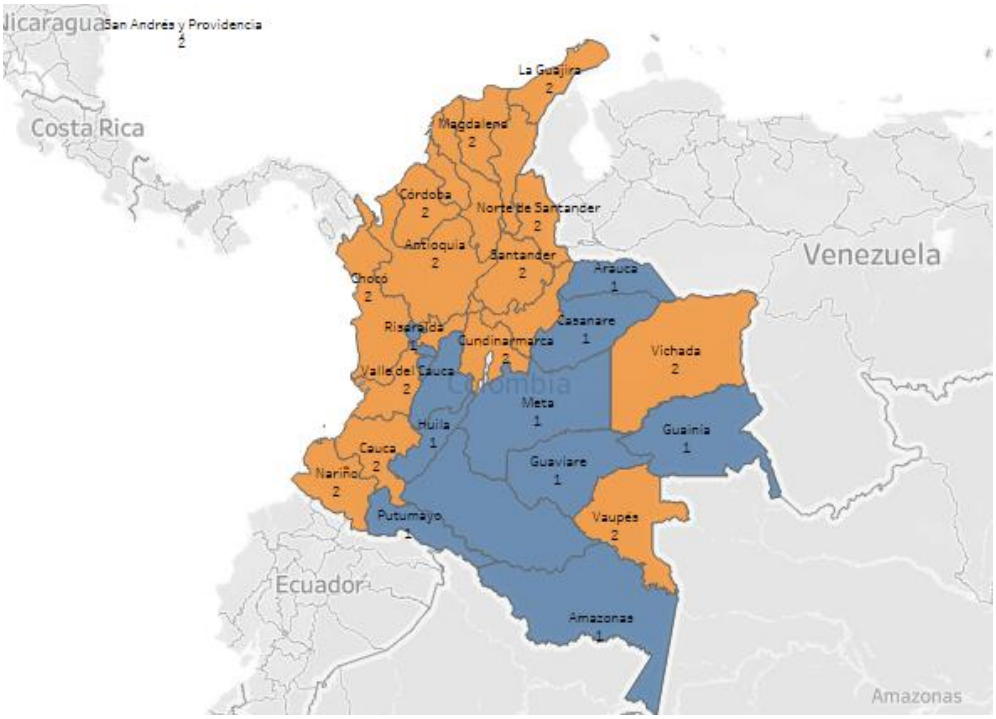


Los departamentos de la clase 1 (Amazonas, Arauca, Caquetá, Casanare, Guainía, Guaviar , Huila, Meta, Putumayo, Quindío, Risaralda, Tolima) se destacan, en su mayoría, por tener el mayor número de casos en 2021 en comparación de los departamentos de la clase 2. En el **anexo 4** se incluyen los resultados de la consolidación de la clasificación para el clúster 1 en el año 2021

Los departamentos de la clase 2 (Antioquia, Caldas, Cundinamarca, Santander, Atlantico, Cauca, La Guajira, Sucre, Bogotá, Cesar, Magdalena, Valle del Cauca, Bolívar, Chocó, Nariño, Norte de Santander, Vaupés, Boyacá, Córdoba, Vichada) se destacan, en su mayoría, por tener el mayor número de casos en 2021 en comparación de los departamentos de la clase 1. En el **anexo 4** se incluyen los resultados de la consolidación de la clasificación para el clúster 2 en el año 2021.

En el siguiente mapa se observa la división de clases que realiza Factoclass para el año 2021. Vemos, por ejemplo, que al igual que 2019 y 2020 se clasificó a los departamentos de Amazonas, Arauca, Casanare, Guainía, Guaviare como aquellos con mayor número de casos de violencia sexual por cada 100.000 habitantes. Sin embargo, ese año los departamentos de Huila, meta, Putumayo, Casanare, Huila también fueron también clasificados en ese mismo clúster. Nuevamente, los departamentos con mayor número de casos se ubicaron, en su mayoría, en el sur del país.

Mapa 3. Georreferenciación división de clases violencia sexual – 2021



De acuerdo con los resultados obtenidos por la librería de Factoclass, existe una diferencia notable en la distribución de casos de violencia sexual entre los departamentos de Colombia. Los departamentos clasificados en la clase 1 (Amazonas, Arauca, Caquetá, Casanare, Guainía, Guaviare, Huila, Meta, Putumayo, Quindío, Risaralda, Tolima) tienden a registrar un mayor número de casos en comparación con los departamentos de la clase 2 (Antioquia, Caldas, Cundinamarca, Santander, Atlántico, Cauca, La Guajira, Sucre, Bogotá, Cesar, Magdalena, Valle del Cauca, Bolívar, Chocó, Nariño, Norte de Santander, Vaupés, Boyacá, Córdoba, Vichada).

También, es posible observar que, en los departamentos con mayor incidencia de violencia sexual, la mayoría de las víctimas no se identificaban con ninguna pertenencia étnica y que las mujeres agredidas sexualmente tienden a residir en zonas urbanas, tener un nivel educativo bajo (haber cursado únicamente la primaria) y pertenecer a grupos de edad más jóvenes (como la adolescencia).

También, se destaca que los casos de violencia sexual son más frecuentes durante los días de la semana o en ciertos períodos del año (como el primer semestre). Estos patrones temporales pueden indicar posibles influencias o contextos específicos en los que ocurre la violencia sexual.

Ahora bien, al comparar los resultados de FactoClass con el uso particular de los algoritmos de K-means, K-medoides y método de Ward, se observa que la ventaja de este método es que combina las diferentes técnicas, proporciona una visión más detallada de la estructura del conjunto de datos que simplemente usar un solo algoritmo. Adicionalmente, nos ofrece una variedad de visualizaciones útiles para ayudar a interpretar los resultados del análisis, por medio del dendrograma y el diagrama de luces, que nos ilustra el número de clúster optimo.

En ese sentido, se identifica que los resultados de FactoClass en comparación con las distintas técnicas (K-means, K-medoides y Metodo de Ward) son coherentes. Así, por ejemplo, para el año 2019, Factoclass clasificó en el clúster 1, al igual que el resto de los algoritmos mencionados, a los departamentos de Amazonas, Arauca, Caquetá, Casanare Guainía, Meta, Putumayo y Tolima, equivalente a aquellos departamentos con mayor número de casos por cada 100.000 habitantes.

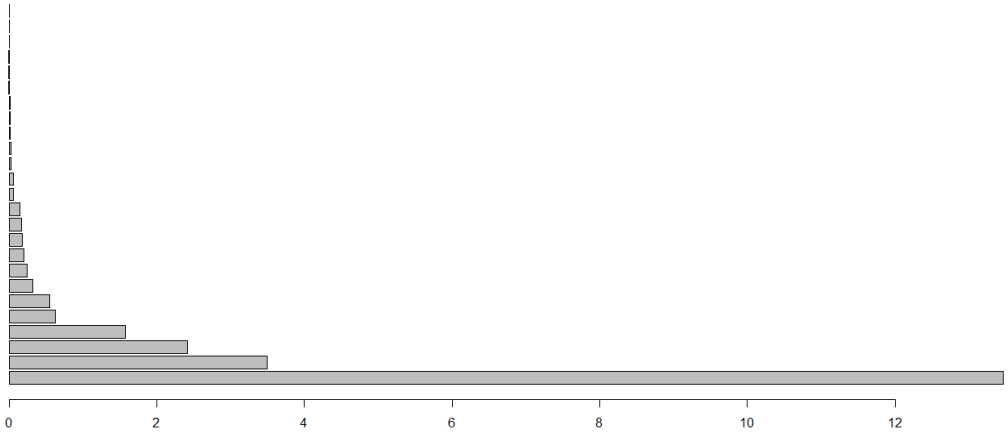
En el año 2020 también clasificó en el Clúster 1 a Amazonas, Arauca, Casanare, Guainía como aquellos departamentos con mayor número de casos por cada 100.000 habitantes, al igual que K-means, K-medoides y método de Ward. Ese año, sin embargo, FactoClass clasificó en un clúster diferente (clúster 3) a los departamentos de San Andrés y Providencia, Caquetá, Meta, huila, Putumayo, Quindío, Risaralda y Tolima, que se ubicaron en un nivel intermedio entre aquellos departamentos con mayor número de casos y los de menor.

Finalmente, en el año 2021, los resultados de FactoClass también fueron coherentes con los de los demás algoritmos aplicados al agrupar en el clúster 1 los departamentos de Amazonas, Arauca, Casanare, Guainía, Guaviare, Tolima como aquellos con mayor número de casos de violencia sexual por cada 100.000 habitantes. En el **anexo 5** es posible contrastar la clasificación de departamental según los distintos modelos y sus respectivas clases durante los años 2019,2020 y 2021.

- **RESULTADOS MÉTODOS DE FACTOCLASS VIOLENCIA DE PAREJA**

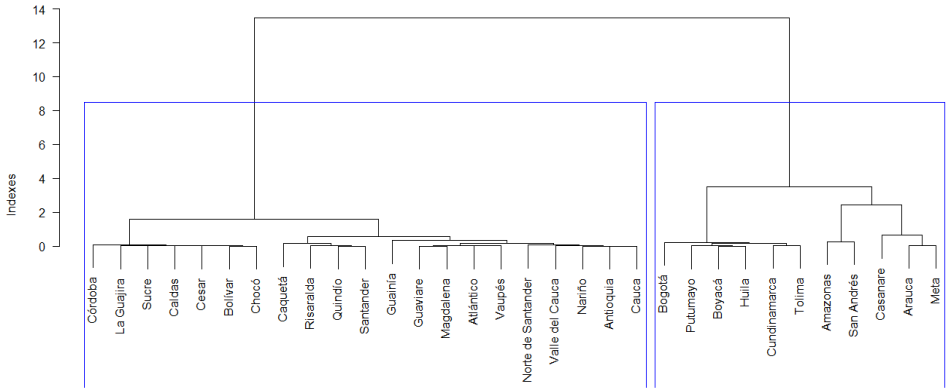
En el caso específico de la base de datos de violencia de pareja, FactoClass realiza una clasificación inicial para identificar el número de clases deseadas para la partición. Esta decisión se toma observando el diagrama de índices de nivel, donde es más fácil identificar los cambios de inercia más significativos y decidir el número de clústeres K. En este caso, se seleccionaron 2 clústeres. El primer análisis se llevó a cabo utilizando los datos del año 2020 y filtrando por sexo - mujer, lo que permitió enfocarse únicamente en los casos de violencia de pareja contra mujeres en Colombia.

Gráfica 48. Gráfico de índices de nivel base de datos violencia de pareja año 2020



A partir de la identificación del número de clústeres, aplicamos FactoClass, el cual nos solicita la cantidad de ejes sobre los que se trabajará. En este caso, se utilizarán las 45 variables de la base de datos. Luego se nos pedirá la cantidad de ejes agrupados, que será 2, y finalmente la cantidad de clústeres determinados, que en este caso son 2. FactoClass realizará una partición y, a través de un gráfico jerárquico, muestra los agrupamientos realizados por los clústeres determinados.

Gráfica 49. Agrupación departamentos mediante FactoClass base de datos violencia de pareja 2020



Teniendo en cuenta la partición, se observa que los departamentos para el año 2020 se dividieron en los siguientes clústeres

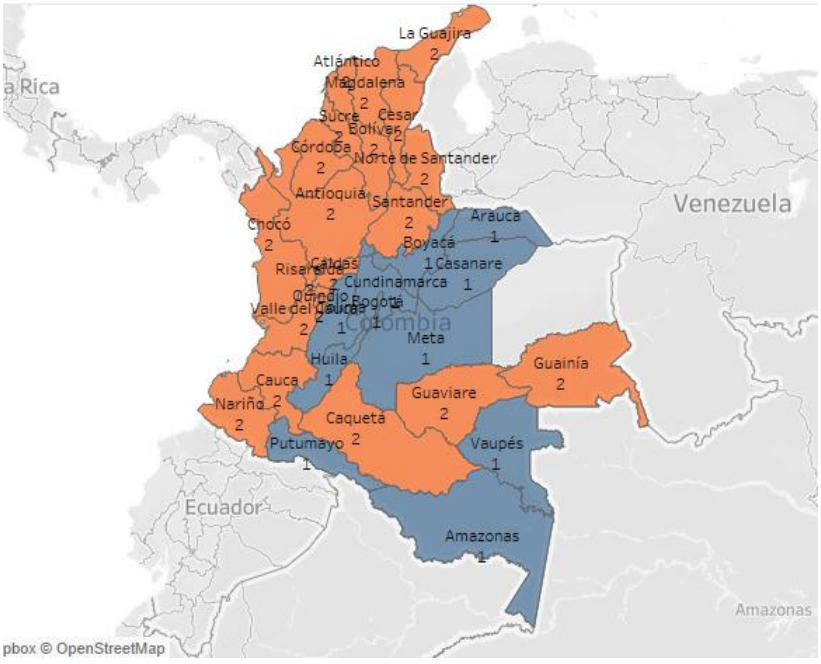
Tabla 10. División de clases por departamento FactoClass base de datos violencia de pareja -2020

Class 1	Amazonas, Arauca, San, Andrés, Bogotá, Casanare, Cundinamarca, Meta, Tolima
Class 2	Antioquia, Atlántico, Bolívar, Boyacá, Caldas, Caquetá, Cauca, Cesar, Chocó, Córdoba, Guainía, Guaviare, Huila, La Guajira, Magdalena, Nariño, Norte de Santander, Putumayo, Quindío, Risaralda, Santander, Sucre, Valle del Cauca, Vaupés

Agrupados los clústeres determinados, se observan los resultados que se determinan para cada uno de los clúster, junto con las variables y sus tasas, para el caso el clúster con mayor cantidad de índice de casos por cada 100.000 habitantes es el clúster 1. En el **anexo 6** se incluyen los resultados de la consolidación de la clasificación para el clúster 1 y 2 en el año 2020.

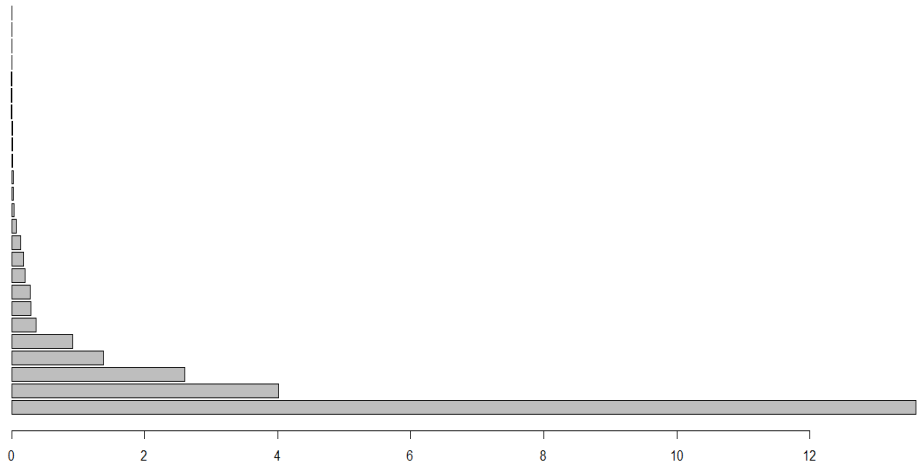
A continuacion, se realizó un análisis geoespacial con los resultados obtenidos del método FactoClass, el cual proporciona una visión visual de la distribución de la violencia de pareja en el territorio nacional durante el período analizado. En el año 2020, se observa que la Clase 1, que comprende los departamentos con la mayor cantidad de casos de violencia de pareja por cada 100.000 habitantes, presenta una concentración en el centro y sur del país. Por otro lado, la Clase 2 agrupa al resto de los departamentos, los cuales se concentran en el norte del país. Es importante destacar que en este año, el departamento de Vichada no registró casos de violencia de pareja, según los datos recopilados.

Mapa 4. Georreferenciación división de clases violencia de pareja – 2020



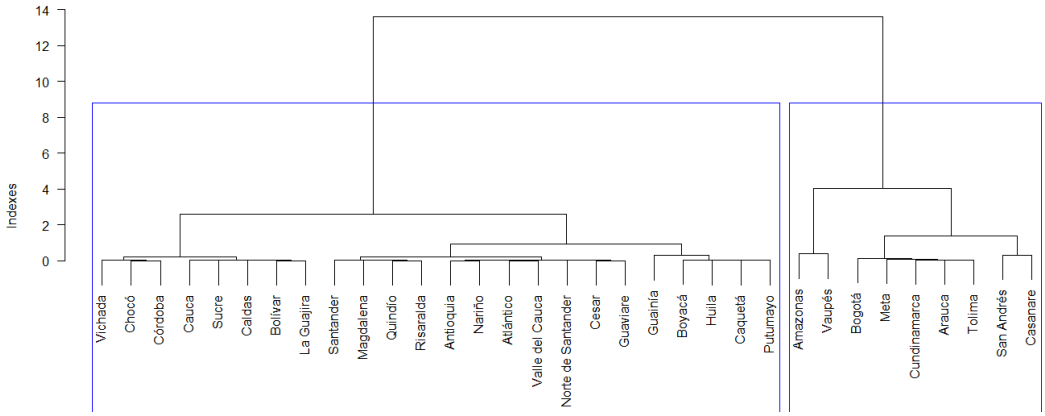
Luego de aplicar el analisis para el año 2020, se realizaron el mismo ejercicio para el año 2021, el cual genera graficos parecidos, pero según se observa en las agrupaciones un aumento de departamentos en el cluster 1.

Ilustración 3. gráfico de índices de nivel base de datos violencia de pareja año 2021



El aumento de departamentos en el clúster 1 indica que para este año se presentó una mayor tasa de violencia de pareja.

Gráfica 50. Agrupación departamentos mediante FactoClass base de datos violencia de pareja 2021



Luego de la agrupación los clústeres se conformaron de los siguientes departamentos:

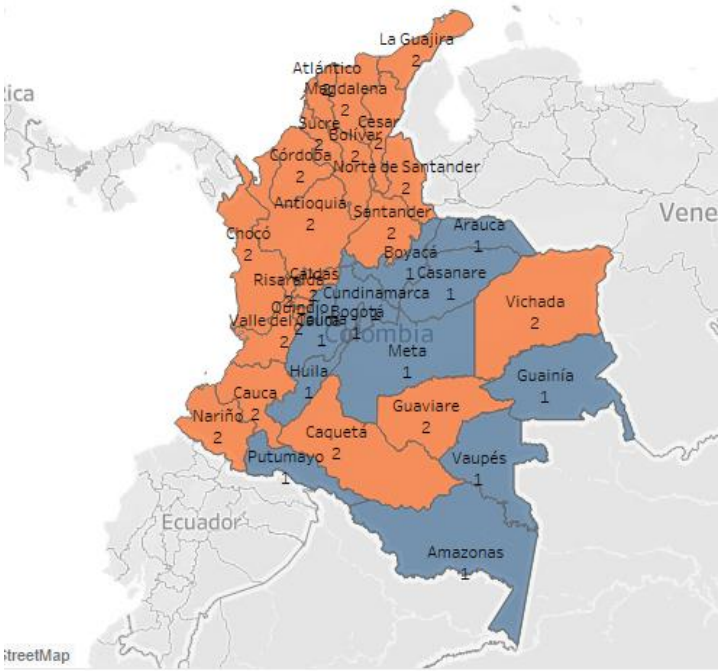
Tabla 11. División de clases por departamento FactoClass base de datos violencia de pareja -2021

Class 1	Amazonas, Arauca, San, Andrés, Bogotá, Casanare, Cundinamarca, Guainía, Meta, Tolima, Vaupés
Class 2	Antioquia, Atlántico, Bolívar, Boyacá, Caldas, Caquetá, Cauca, Cesar, Chocó, Córdoba, Guaviare, Huila, La Guajira, Magdalena, Nariño, Norte de Santander, Putumayo, Quindío, Risaralda, Santander, Sucre, Valle del Cauca, Vichada

Al igual que para el año 2020, la clase 1 corresponde a los departamentos con mayor tasa de violencia de pareja contra las mujeres. En el **anexo 7** se incluyen los resultados de la consolidación de la clasificación para el clúster 1y 2 en el año 2021.

El siguiente análisis permite observar la clasificación en un mapa de las clases resultante de aplicar el método FactoClass, lo que permite visualizar de manera clara y concisa las áreas geográficas más afectadas por la violencia de pareja en el país, En la Clase 1, que engloba los departamentos con la mayor cantidad de casos de violencia de pareja por cada 100.000 habitantes, se evidencia que en su mayoría se concentran en el centro y sur del país. Por otro lado, en la Clase 2 se agrupan los demás departamentos los cuales se concentran en el occidente y norte del país.

Mapa 5. Georreferenciación división de clases violencia de pareja– 2021



El análisis mediante el método FactoClass reveló que la violencia de pareja en Colombia durante los años 2020 y 2021 se concentró principalmente en dos clases geográficas. La clase 1 abarcó departamentos del centro y sur del país, mientras que la clase 2 agrupó a los demás departamentos, principalmente en el occidente y norte. Se observó que la mayoría de las mujeres víctimas de violencia de pareja en ambas clases tenían hasta educación secundaria, fueron agredidas en zonas urbanas y no tenían pertenencia étnica. Las agresiones ocurrieron principalmente entre semana en la clase 1 y se distribuyeron de manera más equitativa en la clase 2. Además, el factor principal identificado en ambos casos fueron los celos. Estos hallazgos destacan la importancia de implementar políticas y acciones dirigidas a prevenir y abordar la violencia de pareja en las áreas geográficas compuestas por los departamentos con mayor tasa de víctimas.

Ahora bien, los resultados obtenidos al aplicar los diversos métodos de análisis (K-means. K-medoides, método de Ward y FactoClass) a la base de datos de violencia de pareja en Colombia durante los años 2020 y 2021, revelan patrones y agrupaciones en los departamentos del país. A continuación, se presentan los resultados obtenidos:

En el año 2020, se identificaron dos agrupaciones principales mediante los métodos de grupos k-medias, grupos k-medoides, el método de Ward y FactoClass. El clúster 1 comprende los departamentos de Amazonas, Arauca, San Andrés, Bogotá, Casanare, Cundinamarca, Guainía, Meta y Tolima, mientras que el clúster 2 incluye los demás departamentos. Es importante mencionar que, para los métodos k-medias, grupos k-medoides y el método de Ward, se evidencia la aparición de los departamentos Boyacá, Huila, Putumayo y Vaupés en el clúster 1, mientras que, en FactoClass, estos departamentos corresponden al clúster 2.

En el año 2021, también se observaron dos agrupaciones predominantes utilizando los mismos métodos de análisis. El clúster 1 está conformado por los departamentos de Amazonas, Arauca, San Andrés, Bogotá, Casanare, Cundinamarca, Guainía, Huila, Meta, Nariño, Tolima, Vaupés y Vichada. Por otro lado, el clúster 2 engloba los restantes departamentos. En este caso, se destaca que el departamento de Vaupés experimentó un aumento en la cantidad de casos de violencia en comparación con el año anterior, mientras que Vichada, aunque no fue considerado en el análisis del año anterior, se encuentra entre los departamentos con mayor cantidad de casos registrados en el año 2021.En el **anexo 8** se muestra una tabla que resume la agrupación de los departamentos según la aplicación de cada método, destacando aquellos departamentos que se mantuvieron en el clúster 1, el cual representa el grupo con la mayor cantidad de casos de violencia registrados durante los años analizados.

8. CONCLUSIONES SOBRE LOS DETERMINANTES QUE INCIDEN EN LA VBG EN COLOMBIA (VIOLENCIA SEXUAL Y DE PAREJA)

Conclusiones determinantes que inciden en la violencia sexual en Colombia 2019-2021:

- Características de las víctimas de violencia sexual: A partir del análisis realizado con la librería FactoClass, se identificaron ciertas características sobresalientes de las víctimas de violencia sexual en Colombia durante los años 2019, 2020 y 2021. Estas características incluyen: no reconocerse bajo ninguna pertenencia étnica, ser agredidas mayormente entre semana y durante el primer semestre del año, tener al agresor como un miembro de su familia, estar en la adolescencia y tener un nivel de escolaridad correspondiente a la primaria. Además, el tipo de violencia de género más prevalente fue la violencia sexual.
- Departamentos con alto número de casos de violencia sexual: Los departamentos de Amazonas, Arauca, Guainía y Guaviare se destacaron por tener el mayor número de casos de violencia sexual por cada 100,000 habitantes durante los tres años estudiados. Estos departamentos comparten similitudes socioeconómicas, como su baja situación de desarrollo humano, elevada tasa de pobreza, dependencia de la economía extractiva y presencia de poblaciones vulnerables. Los índices de desarrollo humano y de pobreza en estos departamentos son más altos en comparación con el promedio nacional.
- Otros departamentos relevantes en el modelo: Además de los cuatro departamentos mencionados anteriormente, los departamentos de Meta, Putumayo, Casanare, Tolima, Huila y Caquetá también fueron destacados en el modelo durante los tres años de estudio. En contraste, hubo una menor incidencia de casos de violencia sexual por cada 100,000 habitantes en departamentos como Antioquia, San Andrés y Providencia, Atlántico, Bogotá, Bolívar, Boyacá, Caldas, Cauca, Cesar, Chocó, Córdoba, Cundinamarca, Huila, La Guajira, Magdalena, Nariño, Norte de Santander, Quindío, Risaralda, Santander, Sucre, Valle del Cauca, Vaupés y Vichada, los cuales se agruparon principalmente en el clúster 2.
- Vulnerabilidad en departamentos con bajos índices de desarrollo humano: Los departamentos de Amazonas, Arauca, Guainía y Guaviare, junto con otros mencionados previamente, presentan índices de desarrollo humano y pobreza por encima del promedio nacional. Estos resultados sugieren que existe una asociación entre la vulnerabilidad socioeconómica y la incidencia de violencia sexual. Es fundamental abordar las desigualdades y mejorar las condiciones socioeconómicas en estos departamentos para reducir la violencia de género.
- Importancia de la familia como agresor: Según los resultados, se observa que los agresores de las víctimas de violencia sexual en Colombia, en su mayoría, son miembros de su propia familia. Este hallazgo resalta la necesidad de trabajar en la prevención y la educación en el seno familiar, así como fortalecer los mecanismos de protección de los derechos de las mujeres y las niñas dentro de los hogares.

- Variaciones temporales en la incidencia de la violencia sexual: Se identificó que la violencia sexual contra las mujeres en Colombia tiende a aumentar durante el primer semestre del año y ocurre principalmente entre semana. Estas variaciones temporales pueden proporcionar información útil para la planificación de intervenciones y la asignación de recursos en momentos específicos del año, así como para la implementación de medidas de prevención y apoyo en días específicos de la semana.
- Impacto desproporcionado en ciertos grupos demográficos: Los resultados revelan que las mujeres en la adolescencia y aquellas con bajo nivel educativo (únicamente primaria) son más vulnerables a la violencia sexual en ambos clústeres. Estos grupos demográficos requieren una atención especial en términos de prevención, protección y acceso a servicios de apoyo.
- Persistencia de la violencia sexual: Existe una alta prevalencia de violencia sexual en Colombia, con un aumento en los casos reportados en los años 2019 y 2020. Este problema representa una preocupación importante en el país.
- Estas conclusiones resaltan la importancia de considerar factores socioeconómicos, como el desarrollo humano y la pobreza, al analizar la incidencia de la violencia sexual en diferentes departamentos de Colombia. Además, se resaltan las características específicas de las víctimas de violencia sexual, lo cual puede ayudar a comprender mejor los perfiles de las personas afectadas y desarrollar estrategias de prevención y apoyo adecuadas.

- **Conclusiones determinantes que inciden en la violencia de pareja en Colombia 2020-2021:**

- Tras analizar las tasas de violencia de pareja en varios departamentos de Colombia, se observa que algunos presentan tasas más altas que otros. En particular, se destaca la presencia de departamentos como Amazonas, Arauca, San Andrés, Casanare, Guainía, Meta, Tolima y Vaupés, donde las tasas de violencia de pareja son significativamente mayores. Estos resultados sugieren que la violencia de pareja es un problema grave en estos departamentos y que se deben tomar medidas para prevenir y combatir esta problemática.
- Los resultados permiten identificar patrones de distribución de la violencia de pareja en Colombia durante los años analizados. Las agrupaciones obtenidas a través de los diferentes métodos de análisis proporcionan información valiosa para comprender la incidencia y características de este problema en cada región del país. En cuanto a las características y patrones identificados:

- De acuerdo con la aplicación de los modelos en la base de violencia de pareja para los años 2021 y 2021 según los datos obtenidos, muestran una serie de características comunes entre las mujeres víctimas de violencia de pareja en Colombia. La mayoría de las víctimas tenían niveles educativos hasta secundaria, fueron agredidas en zonas urbanas y no pertenecían a ninguna etnia. Además, la mayoría de las agresiones ocurrieron entre semana, por su pareja permanente y en su etapa de adultez, en unión libre. Los celos

fueron el principal factor que desencadenó la violencia en estos casos. También es importante destacar que la mayoría de las agresiones ocurrieron en el segundo semestre del año.

- Estos datos son preocupantes y deben ser tomados en cuenta por las autoridades y la sociedad en general. Se necesitan medidas preventivas y de protección para las mujeres que sufren violencia de pareja, así como programas de educación y sensibilización sobre la importancia del respeto y la igualdad de género. Es fundamental trabajar juntos para erradicar la violencia de pareja en Colombia y crear un entorno seguro para todas las mujeres.
- Es importante tener en cuenta que la violencia de pareja es un fenómeno complejo que puede estar influenciado por diversos factores socioeconómicos, culturales y psicológicos, por lo que se hace necesario abordar la problemática desde múltiples perspectivas y con enfoques interdisciplinarios. En definitiva, estos hallazgos destacan la necesidad de seguir investigando y generando políticas públicas que permitan reducir y prevenir la violencia de pareja en todo el país, y en especial en los departamentos donde se han encontrado tasas más altas.

9. BIBLIOGRAFÍA

[1] ACNUR. Violencia de género. [En Línea]. Disponible en

[2] OCHA. "Colombia: Situación de la Violencia Basada en Género (VBG), Comparativo 2020 - 2021 (abril 2022). Disponible en: https://reliefweb.int/report/colombia/colombia-situacion-de-la-violencia-basada-en-g-nero-vbg-comparativo-2020-2021-abril

[3]CEPAL. Al menos 4.091 mujeres fueron víctimas de feminicidio en 2020 en América Latina y el Caribe, 2021. [En Línea]. Disponible en: https://www.cepal.org/es/comunicados/cepal-al-menos-4091-mujeres-fueron-victimas-feminicidio-2020-america-latina-caribe-pese

[4] El Tiempo. Este año más de 2.100 mujeres han sido víctimas de violencia intrafamiliar. 7 de marzo de 2002. [En línea] Disponible en: https://www.eltiempo.com/justicia/investigacion/violencia-intrafamiliar-y-feminicidios-subieron-en-2021-656374#:~:text=En%20todo%202021%2C%20seg%C3%BAn%20los,un%20total%20de%2047.177%20casos).

[5]Pietro, Laura. "Balance sobre la violencia basada en género en Colombia". Fundación Paz y Reconciliación (PARES). Disponible en https://www.pares.com.co/post/balance-sobre-la-violencia-basada-en-g%C3%A9nero-en-colombia

[6] Escuela de datos. La importancia de los datos en la lucha de las mujeres. 2021. [En Línea]. Disponible https://escueladedatos.online/la-importancia-de-los-datos-en-la-lucha-de-lashttps://escueladedatos.online/la-importancia-de-los-datos-en-la-lucha-de-las-mujeres/mujeres/

[7] Onu Mujeres. "Preguntas frecuentes: Tipos de violencia contra las mujeres y lasniñas". [En Línea]. Disponible en: https://www.unwomen.org/es/what-we-do/ending-violence-against-women/faqs/types-of-violence

[8] Tibadá Río, Carolina et al. Tipos y manifestaciones de la violencia de género: una visibilización a partir de relatos de mujeres víctimas en Soacha, Colombia. Prospectiva no.30 Cali July/Dec. 2020 Epub Dec 30, 2020. http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0122-12132020000200006

[9] Profamilia, Defensoría del Pueblo, Organización Internacional para las Migraciones. Guía para la formación en derechos sexuales y reproductivos para población en situación de desplazamiento con énfasis en violencia intrafamiliar y violencia sexual. Bogotá, 2007, Pag. 31-39

[10] Ministerio de Salud. "Nada justifica la violencia contra las mujeres. Trazando una ruta para motivar reflexiones en torno a las violencias basadas en género. Sin fecha. [En Línea]. Disponible en https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/PP/cartilla-nada-justifica-la-vcm.pdf

[11] Tibadá Río, Carolina et al. Óp. Cit.

[12]. CEPAL. "CEPAL: Preocupa la persistencia de la violencia contra las mujeres y las niñas en la región y su máxima expresión, el feminicidio o femicidio". [En Línea]. Disponible en https://www.cepal.org/es/comunicados/cepal-preocupa-la-persistencia-la-violencia-mujeres-ninas-la-region-su-maxima-expresion

[13] ACNUR. Violencia de género. [En Línea]. Disponible en https://www.acnur.org/violencia-sexual-y-de-genero.html

[14] El Tiempo. Este año más de 2.100 mujeres han sido víctimas de violencia intrafamiliar. 7 de marzo de 2002. [En línea] Disponible en: https://www.eltiempo.com/justicia/investigacion/violencia-intrafamiliar-y-feminicidios-subieron-en-2021-656374#:~:text=En%20todo%202021%2C%20seg%C3%BAn%20los,un%20total%20de%2047.177%20casos).

[15] Sisma Mujer. Día Internacional de la Mujer 2022 Violencias contra las mujeres y participación en el mercado laboral. Boletín No. 29, 2022. [En Línea]. Disponible en: https://www.sismamujer.org/wp-content/uploads/2022/03/VF-Boletin-8M-2022-1.pdf

[16] Observatorio Colombiano de feminicidios. Red feminista antimilitarista. [En Línea]. Disponible https://observatoriofeminicidioscolombia.org/index.php

[17] Datacívica y USAID. "Mi experiencia puede servir para que otras no tengan miedo". Señales y estrategias para prevenir la violencia feminicida. Agosto, 2021

[18] Datos contra el femicidio. Herramientas para apoyar la recopilación de datos sobre feminicidios en los medios de comunicación Alerta de correo datos contra feminicidio. [En Línea]. Disponible https://datoscontrafeminicidio.net/herramientas-para-apoyar-lahttps://datoscontrafeminicidio.net/herramientas-para-apoyar-la-recopilacion-de-datos-sobre-feminicidios-de-los-medios-de-comunicacion/recopilacion-de-datos-sobre-feminicidios-de-los-medios-de-comunicacion/

[19] El Sol de México. Usarán inteligencia artificial para prevenir feminicidios. Martes 11 de mayo de 2021. . [En Línea]. Disponible https://www.elsoldemexico.com.mx/mundo/usaranhttps://www.elsoldemexico.com.mx/mundo/usaran-inteligencia-artificial-para-prevenir-feminicidios-eagle-eye-plataforma-violencia-mujeres-seguridad-6700567.htmlinteligencia-artificial-para-prevenir-feminicidios-eagle-eye-plataforma-violencia-mujereshttps://www.elsoldemexico.com.mx/mundo/usaran-inteligencia-artificial-para-prevenir-feminicidios-eagle-eye-plataforma-violencia-mujeres-seguridad-6700567.htmlseguridad-6700567.html

[20] Agaton, Isabel. Balance de la Ley Rosa Elvira Cely contra el feminicidio, a dos años de su vigencia, 2017. [En Línea]. Disponible https://www.ambitojuridico.com/noticias/constitucionalhttps://www.ambitojuridico.com

/noticias/constitucional-y-derechos-humanos/balance-de-la-ley-rosa-elvira-cely-contra-el-feminicidioy-derechos-humanos/balance-de-la-ley-rosa-elvira-cely-contra-el-feminicidio

[21] Secretaría Distrital de La Mujer. el Observatorio de Mujeres y Equidad de Género de Bogotá (OMEG). Visualizador de datos. Bogotá. 2021.

[22] Martínez, Susana et, al. Violencias basadas en género en tiempos de Covid-19. Género y Covid-19. Brief 5. Secretaría Distrital de la Mujer y Fedesarrollo, 2021 [En Línea] https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/4012/Repor_Septiembre_2020_Mart%C3%ADnez_y_et_al.pdf?sequence=1&isAllowed=y

[23] Datascientest.com. "Machine Learning: definición, funcionamiento, usos". [En Línea]. Disponible en https://datascientest.com/es/machine-learning-definicion-funcionamiento-usos

[24] Ibid.

[25] Vallanta, Juan Francisco. "Aprendizaje supervisado y no supervisado". Health Data Miner. [En línea]. Disponible en Escuela de formación en inteligencia artificial en salud

[26] Román, Víctor. "Aprendizaje no supervisado en machine learning: agrupación". [En línea]. Disponible en https://medium.com/datos-y-ciencia/aprendizaje-no-supervisado-en-machine-learning-agrupaci%C3%B3n-bb8f25813edc

[27] Correa Henao, Marisol. "Análisis de clúster automático". Tesis. Universidad Nacional de Colombia Facultad de Minas, Departamento de Ciencias de la Computación y la Decisión Medellín, Colombia 2021. [En línea]. Disponible en https://repositorio.unal.edu.co/bitstream/handle/unal/80784/1017230592.2021.pdf?sequence=2&isAllowed=y

[28] K-medias. Facultad de Informática. UNLP. 2016. [En línea]. Disponible en : http://163.10.22.82/OAS/Agrupamiento_Kmedias/definicin.html

[29] Ibid.

[30] Scikit-learn documentation: K-means Clustering. [En línea]. Disponible en: https://scikit-learn.org/stable/modules/clustering.html#k-means

[31] K medias. Facultad de Informática. UNLP. Óp. Cit.

[32] K. Chamorro, N. Laza, H. Noriega, R. Rojano, J. Vega & D. Heredia, "Aplicación de Machine Learning para análisis de los fenómenos de violencia intrafamiliar en el departamento del Atlántico", Investigación y Desarrollo en TIC, vol. 12, no. 1, pp. 1-12 2021.

[33] Rishabh Singh et al. "K-means Clustering Analysis of Crimes on Indian Women". Journal of Cybersecurity and Information Management (JCIM). Vol. 4, No. 1, PP. 5-25, 2020

[34] Hidalgo Pilar. "Celopatía y motivación sexual: Un análisis a través de K-Means en los establecimientos penitenciarios del Perú". Vol. 6 Núm. 01 (2017): YACHAY; Enero - Diciembre. [En Línea] Disponible en https://revistas.uandina.edu.pe/index.php/YACHAY/article/view/44

[35] Correa Henao, Marisol. "Análisis de clúster automático". Tesis. Universidad Nacional de Colombia Facultad de Minas, Departamento de Ciencias de la Computación y la Decisión Medellín, Colombia 2021. [En línea]. Disponible en https://repositorio.unal.edu.co/bitstream/handle/unal/80784/1017230592.2021.pdf?sequence=2&isAllowed=y

[36] Mathworks. "K-medoids". [En Línea]. Disponible en https://www.mathworks.com/help/stats/kmedoids.html

[37] Ibid.

[38] Morales-Oñate, Víctor y Morales-Oñate, Bolívar. "Una técnica de agrupación robusta para un enfoque Big Data: Clarabd para tipos de datos mixtos". Instituto de Estadística, Universidad de Valparaíso, Valparaíso, Chile, 2019. [En Línea] Disponible en http://ceaa.espoch.edu.ec:8080/revista.perfiles/faces/Articulos/Perfiles22Art12.pdf

[39] Román, Víctor. "Aprendizaje no supervisado en machine learning: agrupación". [En línea]. Disponible en https://medium.com/datos-y-ciencia/aprendizaje-no-supervisado-en-machine-learning-agrupaci%C3%B3n-bb8f25813edc

[40] Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58(301), 236-244.

[41] Fuente, Laura. "Análisis de clústeres no jerárquicos". [En línea]. Disponible en https://www.fuenterrebollo.com/Master-Econometria/Analisis_Cluster.pdf

[42] Ibid.

[43] De la Hoz, Enrique. "Metodología de Aprendizaje Automático para la Clasificación y Predicción de Usuarios en Ambientes Virtuales de Educación", 2019. [En Línea]. Disponible en https://www.researchgate.net/figure/Dendograma-para-3-grupos-con-el-metodo-Ward_fig1_331462041

[44] Aprendemachinelearning. "Clasificar con K-Nearest-Neighbor ejemplo en Python". [En Línea]. Disponible en https://www.aprendemachinelearning.com/clasificar-con-k-nearest-neighbor-ejemplo-en-python/

[45] JavaTPoint. "K-Nearest Neighbor(KNN) Algorithm for Machine Learning". [En Línea]. Disponible en https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

[46] Ibid.

[47] Ibid.

[48] Aprende IA. "DBSCAN". [En línea]. Disponible en https://aprendeia.com/dbscan-teoria/#:~:text=A%20diferencia%20de%20K%20Means,ruido%20y%20los%20valores%20at%C3%ADpicos.

[49] Ibid.

[50] Interactivechaos. "Ejemplo con DBSCAN". [En Línea]. Disponible en https://interactivechaos.com/es/manual/tutorial-de-machine-learning/ejemplo-con-dbscan

[51] Instituto de Medicina Legal y Ciencias Forenses. Oficio No. 123 -GCERN-SSF-2022 Bogotá, 2022-08-30. Asunto: Respuesta a solicitud de información estadística de riesgo de feminicidio Referencia: RADICADO N° 2022DG-GNSC01792 del 11 de agosto de 2022

[52] Correa Henao, Marisol. "Análisis de clúster automático". Tesis. Universidad Nacional de Colombia Facultad de Minas, Departamento de Ciencias de la Computación y la Decisión Medellín, Colombia 2021. [En línea]. Disponible en https://repositorio.unal.edu.co/bitstream/handle/unal/80784/1017230592.2021.pdf?sequence=2&isAllowed=y

[53] Delgado, Ronald. "introducción a los modelos de agrupamiento (clustering) en R". [En Línea] Disponible en https://rpubs.com/rdelgado/399475

[54] Paredes, Daniel. "Análisis de datos y algoritmos de predicción en R. [En línea]. Disponible en https://bookdown.org/dparedesi/data-science-con-r/

[55] Kassambara, A. (2020). Factoextra: Extract and visualize the results of multivariate data analyses. R package version 1.0.7. Retrieved from https://cran.r-project.org/package=factoextra

[56] Acero, Andrés. "Clustering basado en la media". [En Línea]. Disponible en: https://rpubs.com/andresacerol/C13ITAU1

[57] Soto, Jurgen. "DBSCAN". [En Línea]. Disponible en: https://rstudio-pubs-static.s3.amazonaws.com/605966_6764f95937ed4c37934bd63c7bb99c98.html

[58] Arita, Brian et. al. "Laboratorio 2, análisis de clusters". Universidad de El Salvador, Facultad de ciencias económicas . [En Línea]. Disponible en https://rpubs.com/Eunice_Ramirez/analisis_cluster

[59] Ibid.

[60] Aprendemachinelearning. "Clasificar con K-Nearest-Neighbor ejemplo en Python". [En Línea]. Disponible en https://www.aprendemachinelearning.com/clasificar-con-k-nearest-neighbor-ejemplo-en-python/

[61] Pardo, Campo Elías y Del Campo, Pedro Cesar. "Combinación de métodos factoriales y de análisis de conglomerados en *R*: el paquete *FactoClass.* Rev.Colomb.Estad. vol.30 no.2 Bogotá July/Dec. 2007

[62] Ibid.

[63] Ibid.

[64] Ibid.

10.ANEXOS

- **ANEXO 1**

Tabla 12. Selección de variables base de datos presuntos casos de violencia sexual

Variables base de datos original violencia sexual	Variables seleccionadas o agrupadas violencia sexual
Variable grupo de edad: Se presenta el rango de edad de las víctimas de los 00-04 años hasta los 80 años y más.	Se eliminó esta variable y se dejó únicamente la variable ciclo vital que agrupa los rangos de edad en primera infancia (00-05 años), (6-11 años) infancia, (12-17 años) adolescencia, (18 a 28) Juventud, (29 a 59) Adultez, (Más de 60) Adulto Mayor.
Variable grupo de edad judicial: Se presenta el rango de edad de las víctimas de los 00-04 años hasta los 80 años y más.	Se eliminó esta variable y se dejó únicamente la variable ciclo vital que facilita el análisis por edad.
Variable escolaridad: se clasifica a las presuntas víctimas de violencia sexual según su rango de escolaridad en educación inicial y educación preescolar, educación básica primaria, educación básica secundaria o secundaria baja, educación media o secundaria alta, educación técnica profesional y tecnológica, universitario, especialización, maestría o equivalente, sin escolaridad y sin información.	Se realizó una agrupación de esta variable dejando únicamente los siguientes rangos primaria, secundaria, profesional, sin escolaridad, sin información.
Variable estado civil: se clasifica a las presuntas víctimas de violencia sexual según su estado civil (a), Unión libre, Casado (a), Separado (a), Divorciado (a), Viudo (a), No aplica, Sin información	No se tuvo en cuenta esta variable para el análisis, dado que se seleccionaron otras variables que podrían ser más relevantes para el análisis.
Variable factor de vulnerabilidad: Campesinos (as) y/o trabajadores (as) del campo, Comunidad LGBT, Conductores de vehículos de servicio público, Defensores de los Derechos Humanos, funcionarios judiciales, Grupos étnicos, Niños, niñas, adolescentes en condición de abandono, entre otras	No se tuvo en cuenta esta variable para el análisis, pues las múltiples opciones que daban cuenta de los factores de vulnerabilidad dificultaban su agrupación.
Variable Tipo de discapacidad: Auditiva, Física, Mental, Psíquica, Visual, Discapacidad múltiple, ninguna.	No se tuvo en cuenta esta variable para el análisis, dado que se seleccionaros otras variables que podrían ser más relevantes para el análisis.
Variable pertenencia étnica: Indígena, Negro/Afrodescendiente, Palenquero, Raizal, Rom (Gitano), Sin Pertenencia Étnica	Se agrupó esta variable dejando únicamente la opción Sí o No.

Variable presunto agresor: agresor desconocido; Amigo (a); Conocido; Delincuencia común; Desmovilizados/Reinsertados; Encargado del cuidado; Familiar (Abuelo (a), Cuñado (a), Hermano (a), Hijo (a), Madrastra, Madre, entre otros); Miembro de grupos alzados al margen de la ley(FARC, ELN, EPL); Miembro de un grupo de la delincuencia organizada (Bandas criminales, Miembro de un grupo de la delincuencia organizada, Paramilitares); Miembros de las fuerzas armadas, de policía, policía judicial y servicios de inteligencia; Pareja o ex pareja (Amante, Compañero (a) permanente, Esposo (a), Ex – Amante, entre otros); Personal de custodia; sin información.	Se agrupó esta variable dejando únicamente agresor desconocido, amigo (a), conocido, delincuencia común, Desmovilizados/Reinsertados; Encargado del cuidado; Familiar; Miembro de grupos alzados al margen de la ley; Miembro de un grupo de la delincuencia organizada; Miembros de las fuerzas armadas, de policía, policía judicial y servicios de inteligencia; Pareja o ex pareja; Personal de custodia; sin información
Variable sexo el presunto agresor: Hombre, mujer, sin información.	No se tuvo en cuenta esta variable para el análisis, dado que se seleccionaron otras variables que podrían ser más relevantes para el análisis.
Variable circunstancia del hecho: Violencia económica (Atraco callejero o intento de hurto); Violencia interpersonal (Ajuste de cuentas , Contacto engañoso vía internet, Ejercicio de actividades ilícitas, entre otras); Violencia Intrafamiliar (Violencia a niños, niñas y adolescentes, Violencia al adulto mayor, entre otras); Violencia sexual (Abuso dentro de establecimiento prestador de servicios de salud, abuso sexual, Acceso carnal violento/acto sexual violento con persona protegida, Asalto sexual, Pornografía, entre otras); Violencia sociopolítica (Acción bandas criminales, Acción grupos alzados al margen de la ley, retención ilegal, entre otras).	Se agrupó esta variable dejando únicamente las opciones de Violencia económica, Violencia interpersonal, Violencia Intrafamiliar, Violencia sexual, Violencia sociopolítica, sin información.
Variable actividad durante el hecho: Actividades de desplazamiento de un lugar a otro; Actividades de trabajo doméstico no pagado para el uso del propio hogar; Actividades ilícitas o delictivas; Actividades relacionadas con el aprendizaje; Actividades relacionadas con el cuidado no pagado de miembros del hogar, entre otras.	No se tuvo en cuenta esta variable para el análisis, pues las múltiples opciones que daban cuenta de las actividades durante el hecho dificultaban su agrupación
Variable país de nacimiento de la víctima: Argentina, Aruba, Bélgica, Colombia, entre otros.	No se tuvo en cuenta esta variable para el análisis, dado que se seleccionaron otras variables que podrían ser más relevantes para el análisis.
Variable escenario del hecho: Ambulancia - transporte sanitario; Áreas deportivas y/o recreativas; Calle (autopista, avenida, dentro de la ciudad); Centros de reclusión; Centros educativos, entre otras.	No se tuvo en cuenta esta variable para el análisis, pues las múltiples opciones que daban cuenta del escenario del hecho dificultaban su agrupación
Variable zona del hecho: Cabecera municipal; Centro poblado (corregimiento, inspección de policía y caserío); Parte rural (vereda y campo); Sin información.	Se agrupó esta variable dejando únicamente las opciones de rural o urbano.

Variable mes del año: enero, febrero, marzo, abril hasta diciembre.	Se agrupó esta variable dejando únicamente las opciones Semestre I o Semestre II.
Variable día del hecho: lunes a Domingo	Se agrupó esta variable dejando únicamente las opciones entre semana o fines de semana.

Tabla 13. Selección de variables base de datos presuntos casos violencia de pareja

Variables base de datos original violencia de Pareja	**Variables seleccionadas o agrupadas violencia de Pareja**
Variable grupo de edad: Representa el rango de edad de las víctimas la cual comprende entre los 0 y +60, agrupadas en rangos cada 4 años	Se eliminó esta variable y se dejó únicamente la variable ciclo vital que facilita el análisis por edad
Variable grupo edad judicial: Representa el rango de edad de las víctimas la cual comprende entre los 0 y +60, agrupadas en rangos cada 3 años	Se eliminó esta variable y se dejó únicamente la variable ciclo vital que facilita el análisis por edad
Variable ciclo vital: (06 a 11) Infancia - (12 a 17) Adolescencia - (18 a 28) Juventud - (29 a 59) Adultez - (Más de 60) Adulto Mayor	Esta variable se mantuvo ya que permitía analizar mejor los grupos poblacionales por rango de edades los cuales son (06 a 11) Infancia - (12 a 17) Adolescencia - (18 a 28) Juventud - (29 a 59) Adultez - (Más de 60) Adulto Mayor
Variable escolaridad: - Educación inicial y educación preescolar, - Educación básica primaria, - Educación básica secundaria o secundaria baja,- Educación media o secundaria alta,- Educación técnica profesional y tecnológica,- Universitario,- Especialización, Maestría o equivalente,- Doctorado o equivalente,- Sin escolaridad,- Sin información,	La variable se mantuvo y se redujo a 5 grandes grupos los cuales representaban mejor a la población. Primaria - Secundaria - Profesional - Sin escolaridad - Sin información
variable estado civil: Soltero (a) - Unión libre - Casado (a) - Separado (a), Divorciado (a) - Viudo (a) - Sin información	Esta variable se tuvo en cuenta ya que es una variable importante en el análisis de violencia de pareja al determinar la situación legal de la víctima, de esta variable se obtuvieron los siguientes grupos. Soltero (a) - Casado (a) - Sin información - Viudo (a) - Separado (a), Divorciado (a) - Unión libre
Variable factor de vulnerabilidad: - Campesinos (as) y/o trabajadores (as) del campo,- Comunidad LGBT,- Conductores de vehículos de servicio público,- Grupos étnicos,- Mujer cabeza de hogar o de familia ,- Persona adicta a una droga natural o sintética,- Persona en condición de desplazamiento ,- Persona en situación de calle,- Persona en situación de prostitución,- Persona recluida en establecimiento de rehabilitación y pabellones psiquiátricos,- Personas bajo custodia,- Personas desmovilizadas o reinsertadas,- Personas mayores en hogares de cuidado,- Tribus urbanas,- Ninguno ,- Sin información,	La variable no se tuvo en cuenta para el análisis, dado que se seleccionaros otras variables que podrían ser más relevantes.

Variable tipo de discapacidad: Auditiva, - Física,- Mental, - Siquia, - Visual,- Discapacidad múltiple,- Ninguna	La variable no se tuvo en cuenta para el análisis, dado que se seleccionaros otras variables que podrían ser más relevantes.
Variable pertenencia étnica: Indígena, - Negro/Afrodescendiente,- Palenquero,- Raizal,- Rom (Gitano),- Sin Pertenencia Étnica, - Sin Información	Esta variable se agrupo teniendo en cuenta SI la población pertenecía a un grupo étnico o NO, por lo que los grupos resultantes fueron SI, NO
Variable presunto agresor: - Pareja o ex pareja, - Amante,- Compañero (a) permanente,- Esposo (a),- Ex - Amante,- Ex - compañero (a) permanente,- Ex - esposo (a),- Ex - Novio (a),- Novio (a),- Sin información	La variable se agregó debido al tipo de violencia que se está analizando pata esta base, permaneció en los grupos establecidos , Pareja o ex pareja,- Amante,- Compañero (a) permanente,- Esposo (a),- Ex - Amante,- Ex - compañero (a) permanente,- Ex - esposo (a),- Ex - Novio (a),- Novio (a),- Sin información
Variable sexo del presunto agresor: Hombre, - Mujer,- Sin información	La variable no se tuvo en cuenta para el análisis, dado que se seleccionaros otras variables que podrían ser más relevantes.
Variable actividad durante el hecho: - Actividades de desplazamiento de un lugar a otro,- Actividades de trabajo doméstico no pagado para el uso del propio hogar,- Actividades ilícitas o delictivas,- Actividades relacionadas con el aprendizaje,- Actividades relacionadas con el cuidado no pagado de miembros del hogar,- Actividades relacionadas con el trabajo remunerado,- Actividades relacionadas con la asistencia a eventos culturales, de entretenimiento y/o deportivos,- Actividades relacionadas con los deportes y el ejercicio físico,- Actividades relacionadas con manifestaciones públicas (marchas, protestas, etc.),- Actividades vitales o relacionadas con el cuidado personal,- Actividades relacionadas con enfrentamientos armados,- Sin información,- Otra	La variable no se tuvo en cuenta para el análisis, dado que se las múltiples actividades durante el hecho impedían un agrupamiento.
Variable factor desencadenante de la agresión: - Abandono, - Celos, desconfianza, infidelidad, - Consumo de alcohol y/o sustancias psicoactivas,- Económicas,- Enfermedad física o mental,- Intolerancia, machismo,- Otras razones,- Sin información,	El factor desencadénate es una variable vital para determinar por qué se generó el tipo de violencia, los grupos permanecieron iguales: - Abandono,- Celos, desconfianza, infidelidad,- Consumo de alcohol y/o sustancias psicoactivas,- Económicas,- Enfermedad física o mental,- Intolerancia, machismo,- Otras razones,- Sin información,
Variable mecanismo causal: - Abrasivo, - Agente químico, - Agentes y mecanismo explosivo,- Agentes y mecanismos biológicos,- Biodinámico,- Cáustico ,- Contundente,- Cortante,- Corto contundente,- Corto punzante,- Eléctrico,- Generadores de asfixia,- Mecanismo múltiple,- Proyectil de arma de fuego,- Punzante,- Térmico,- Tóxico,- Por determinar,	La variable no se tuvo en cuenta para el análisis, dado que se los múltiples factores impedían un agrupamiento.

Variable diagnóstico topográfico de la lesión: - Piel y faneras, - Politraumatismo, - Trauma área Pélvica,- Trauma craneano,- Trauma de abdomen,- Trauma de cuello,- Trauma de miembros,- Trauma de tórax,- Trauma facial,- Por determinar	La variable no se tuvo en cuenta para el análisis, dado que se seleccionaron otras variables que podrían ser más relevantes.
Variable días de incapacidad médico legal: 1 a 30 31 a 90 Más de 90 Cero días y Sin información	La variable no se tuvo en cuenta para el análisis, dado que se seleccionaros otras variables que podrían ser más relevantes.
Variable país de nacimiento de la víctima: - Afganistán, - Argentina,- Aruba,- Bolivia,- Brasil,- Canadá,- Chile,- Colombia,- Costa Rica,- Cuba,- Ecuador,- España,- Estados Unidos,- Filipinas,- Francia,- India,- Israel,- Italia,- Líbano,- Letonia,- Madagascar,- México,- Nicaragua,- Noruega,- Países Bajos,- Panamá,- Paraguay,- Perú,- Rumania,- Venezuela,- Sin información,	La variable no se tuvo en cuenta para el análisis, dado que se seleccionaron otras variables que podrían ser más relevantes.
Variable escenario del hecho: Áreas deportivas y/o recreativas Calle (autopista, avenida, dentro de la ciudad) Carretera (fuera de la ciudad) Centro de atención médica (hospital, clínica, consultorio, etc.) Centros de reclusión Centros educativos	La variable no se tuvo en cuenta para el análisis, dado que se seleccionaron otras variables que podrían ser más relevantes.
Variable zona del hecho: - Cabecera municipal, - Centro poblado (corregimiento, inspección de policía y caserío) ,- Parte rural (vereda y campo),- Sin información,	La variable es importante con el fin de determinar en qué zonas se presentan más casos de violencia, los grupos obtenidos fueron: Rural, Urbano
Variable mes del hecho: Enero a Diciembre	La variable es importante pues permite determinar en qué semestre del año se presentan más casos de violencia, los grupos obtenidos fueron: Semestre 1, Semestre 2
Variable día del hecho: lunes a Domingo	La variable es importante pues permite determinar si los casos de violencia se dan más entre semana o en fines de semana, los grupos obtenidos fueron: Entre Semana, Fin de semana
Variable rango de hora del hecho (X 3 HORAS)	La variable no se tuvo en cuenta para el análisis, dado que se seleccionaron otras variables que podrían ser más relevantes.

- **ANEXO 2.**

Ilustración 4. Consolidación de la clasificación Factoclass base de datos violencia sexual 2019-cluster 1

```
class: 1
                    Test.Value Class.Mean Frequency Global.Mean
sin_informacion_2        5.145      8.232         9       3.181
v_intrafamiliar          4.800      4.726         9       2.047
entre_semana             4.738    164.964         9      97.410
semestre_1               4.663    105.606         9      61.598
conocido                 4.660     50.371         9      28.212
v_sexual                 4.597    192.165         9     114.632
rural                    4.570     42.216         9      20.546
infancia                 4.565     65.449         9      37.426
primaria                 4.549    160.606         9      91.902
semestre_2               4.520    100.245         9      58.973
adolescencia             4.488     94.228         9      53.106
urbana                   4.358    163.634         9     100.025
pareja_expareja          4.186     17.468         9       9.628
per_etnica_no            4.155    170.044         9     103.327
juventud                 4.133     18.623         9      11.063
fin_semana               4.090     40.886         9      23.160
otro                     4.023     13.603         9       6.336
familiar                 4.019     82.677         9      50.853
sin_informacion          3.401     14.947         9       7.589
sin_escolaridad          3.353     15.011         9       9.525
grupos_armados           3.300      1.525         9       0.495
sin_informacion_16       3.261      9.171         9       3.852
Primera_infancia         2.891     19.401         9      13.200
profesional              2.702      4.263         9       2.892
secundaria               2.635     16.799         9      12.398
agre_desconocido         2.448     10.825         9       7.238
encargado_del_cuidado    2.438      2.001         9       0.699
adultez                  2.375      7.819         9       5.447
amigo                    2.373     10.982         9       8.069
```

Ilustración 5. Consolidación de la clasificación Factoclass base de datos violencia sexual 2019-cluster 2

```
class: 2
                    Test.Value Class.Mean Frequency Global.Mean
amigo                   -2.373      6.976        24       8.069
adultez                 -2.375      4.557        24       5.447
encargado_del_cuidado   -2.438      0.210        24       0.699
agre_desconocido        -2.448      5.893        24       7.238
secundaria              -2.635     10.747        24      12.398
profesional             -2.702      2.378        24       2.892
Primera_infancia        -2.891     10.875        24      13.200
sin_informacion_16      -3.261      1.858        24       3.852
grupos_armados          -3.300      0.108        24       0.495
sin_escolaridad         -3.353      7.468        24       9.525
sin_informacion         -3.401      4.830        24       7.589
familiar                -4.019     38.919        24      50.853
otro                    -4.023      3.611        24       6.336
fin_semana              -4.090     16.513        24      23.160
juventud                -4.133      8.228        24      11.063
per_etnica_no           -4.155     78.308        24     103.327
pareja_expareja         -4.186      6.687        24       9.628
urbana                  -4.358     76.171        24     100.025
adolescencia            -4.488     37.685        24      53.106
semestre_2              -4.520     43.496        24      58.973
primaria                -4.549     66.139        24      91.902
infancia                -4.565     26.918        24      37.426
rural                   -4.570     12.419        24      20.546
v_sexual                -4.597     85.558        24     114.632
conocido                -4.660     19.902        24      28.212
semestre_1              -4.663     45.095        24      61.598
entre_semana            -4.738     72.078        24      97.410
v_intrafamiliar         -4.800      1.042        24       2.047
sin_informacion_2       -5.145      1.287        24       3.181
>
```

- **ANEXO 3**

Ilustración 6. Consolidación de la clasificación Factoclass base de datos violencia sexual 2020-cluster 1

class: 1

	Test.Value	Class.Mean	Frequency	Global.Mean
semestre_2	4.428	90.070	5	42.492
primaria	4.326	129.907	5	60.620
urbana	4.305	147.837	5	71.808
entre_semana	4.245	138.540	5	68.736
infancia	4.242	55.144	5	25.504
conocido	4.147	42.494	5	19.333
adolescencia	4.137	76.779	5	38.571
otro	4.135	12.749	5	4.405
v_sexual	4.129	159.156	5	81.395
semestre_1	3.987	83.420	5	44.126
v_intrafamiliar	3.880	7.044	5	1.999
fin_semana	3.792	34.949	5	17.883
familiar	3.649	67.211	5	37.068
Primera_infancia	3.563	17.654	5	9.537
fuerzas_armadas	3.534	1.139	5	0.284
juventud	3.473	17.519	5	9.243
sin_informacion	3.454	15.856	5	5.551
pe_etnica_si	3.388	46.665	5	13.283
adultez	3.343	6.394	5	3.579
sin_informacion_16	3.240	7.928	5	3.051
per_etnica_no	3.155	126.824	5	73.336
secundaria	2.906	21.545	5	14.146
pareja_expareja	2.874	13.683	5	7.601
sin_informacion_2	2.661	7.076	5	2.720
amigo	2.514	10.965	5	7.025
delicuencia_organizada	2.225	1.264	5	0.261
encargado_del_cuidado	2.210	0.488	5	0.178
rural	2.197	25.653	5	14.810
sin_escolaridad	2.152	10.661	5	6.507
personal_custodia	2.095	0.588	5	0.140
agre_desconocido	2.079	6.946	5	4.338
adulto_mayor	-2.019	0.000	5	0.184

Ilustración 7. Consolidación de la clasificación Factoclass base de datos violencia sexual 2020-cluster 2.

class: 2

	Test.Value	Class.Mean	Frequency	Global.Mean
pe_etnica_si	-2.019	6.506	20	13.283
encargado_del_cuidado	-2.024	0.081	20	0.178
fuerzas_armadas	-2.063	0.114	20	0.284
sin_informacion	-2.485	3.025	20	5.551
sin_informacion_16	-2.740	1.646	20	3.051
profesional	-2.832	1.520	20	2.295
agre_desconocido	-2.925	3.088	20	4.338
v_intrafamiliar	-2.998	0.671	20	1.999
otro	-3.239	2.178	20	4.405
sin_escolaridad	-3.325	4.320	20	6.507
rural	-3.463	8.989	20	14.810
adultez	-3.679	2.524	20	3.579
amigo	-3.857	4.966	20	7.025
fin_semana	-3.928	11.859	20	17.883
juventud	-3.934	6.048	20	9.243
conocido	-3.935	11.847	20	19.333
pareja_expareja	-4.198	4.575	20	7.601
secundaria	-4.226	10.480	20	14.146
adolescencia	-4.392	24.751	20	38.571
urbana	-4.422	45.203	20	71.808
per_etnica_no	-4.441	47.685	20	73.336
infancia	-4.458	14.891	20	25.504
primaria	-4.471	36.224	20	60.620
semestre_2	-4.566	25.777	20	42.492
semestre_1	-4.679	28.415	20	44.126
v_sexual	-4.687	51.324	20	81.395
familiar	-4.708	23.817	20	37.068
entre_semana	-4.713	42.333	20	68.736
Primera_infancia	-4.847	5.776	20	9.537

Ilustración 8. Consolidación de la clasificación Factoclass base de datos violencia sexual 2020 cluster 3

class: 3

	Test.Value	Class.Mean	Frequency	Global.Mean
Primera_infancia	2.545	13.868	8	9.537
per_etnica_no	2.424	104.032	8	73.336
secundaria	2.387	18.686	8	14.146
pareja_expareja	2.381	11.365	8	7.601
familiar	2.315	51.356	8	37.068
amigo	2.294	9.711	8	7.025
rural	2.110	22.588	8	14.810
profesional	2.027	3.510	8	2.295

- **ANEXO 4.**

Ilustración 9.Consolidación de la clasificación Factoclass base de datos violencia sexual 2021-cluster 1

class: 1

	Test.Value	Class.Mean	Frequency	Global.Mean
semestre_2	4.952	99.232	10	57.646
primaria	4.902	131.580	10	76.437
entre_semana	4.874	151.855	10	90.800
infancia	4.809	53.550	10	29.747
v_sexual	4.809	168.295	10	103.170
adolescencia	4.798	93.161	10	54.690
urbana	4.779	153.857	10	93.783
semestre_1	4.756	97.270	10	57.933
familiar	4.719	79.738	10	47.927
fin_semana	4.635	44.647	10	24.779
sin_escolaridad	4.587	15.953	10	8.796
conocido	4.567	41.866	10	24.521
rural	4.253	42.645	10	21.796
Primera_infancia	4.209	18.194	10	10.895
secundaria	4.206	34.275	10	22.458
sin_informacion	4.202	25.716	10	13.224
pareja_expareja	4.172	20.282	10	12.737
juventud	4.168	21.703	10	13.173
v_intrafamiliar	3.919	15.775	10	7.405
per_etnica_no	3.834	143.125	10	93.180
sin_informacion_2	3.450	9.254	10	3.680
amigo	3.412	13.471	10	8.864
agre_desconocido	2.964	12.077	10	6.722
sin_informacion_16	2.879	9.508	10	3.817
adultez	2.867	9.554	10	6.794
v_interpesonal	2.792	2.184	10	0.903
encargado_del_cuidado	2.738	1.205	10	0.444
pe_etnica_si	2.597	53.377	10	22.399
desmovi_reinsertados	2.146	0.039	10	0.012

Ilustración 10. Consolidación de la clasificación Factoclass base de datos violencia sexual 2021-cluster 2

```
class: 2
                      Test.value Class.Mean Frequency Global.Mean
desmovi_reinsertados      -2.146      0.000        23       0.012
pe_etnica_si              -2.597      8.931        23      22.399
encargado_del_cuidado     -2.738      0.113        23       0.444
v_interpesonal            -2.792      0.346        23       0.903
adultez                   -2.867      5.595        23       6.794
sin_informacion_16        -2.879      1.343        23       3.817
agre_desconocido          -2.964      4.394        23       6.722
amigo                     -3.412      6.862        23       8.864
sin_informacion_2         -3.450      1.257        23       3.680
per_etnica_no             -3.834     71.464        23      93.180
v_intrafamiliar           -3.919      3.766        23       7.405
juventud                  -4.168      9.464        23      13.173
pareja_expareja           -4.172      9.456        23      12.737
sin_informacion           -4.202      7.793        23      13.224
secundaria                -4.206     17.320        23      22.458
Primera_infancia          -4.209      7.722        23      10.895
rural                     -4.253     12.731        23      21.796
conocido                  -4.567     16.980        23      24.521
sin_escolaridad           -4.587      5.685        23       8.796
fin_semana                -4.635     16.140        23      24.779
familiar                  -4.719     34.096        23      47.927
semestre_1                -4.756     40.831        23      57.933
urbana                    -4.779     67.664        23      93.783
adolescencia              -4.798     37.963        23      54.690
infancia                  -4.809     19.398        23      29.747
v_sexual                  -4.809     74.855        23     103.170
entre_semana              -4.874     64.255        23      90.800
primaria                  -4.902     52.461        23      76.437
semestre_2                -4.952     39.565        23      57.646
```

- **ANEXO 5.**

Tabla 14. Contrastación departamental según resultados de modelos violencia sexual 2019-2021

Año	Departamentos	K-means	K-medoides	Metodo de Ward	Factoclass
2019	Amazonas	1	1	1	1
	Antioquia	2	2	2	2
	Arauca	1	1	1	1
	Atlántico	2	2	2	2
	Bolívar	2	2	2	2
	Boyacá	2	2	2	2
	Caldas	2	2	2	2
	Caquetá	1	1	1	1
	Casanare	1	1	1	1
	Cauca	2	2	2	2
	Cesar	2	2	2	2
	Chocó	2	2	2	2

	Córdoba	2	2	2	2
	Cundinamarca	2	2	2	2
	Guainía	1	1	1	1
	Guaviare	2	1	1	1
	Huila	2	1	1	2
	La Guajira	2	2	2	2
	Magdalena	2	2	2	2
	Meta	1	1	1	1
	Nariño	2	2	2	2
	Norte de Santander	2	2	2	2
	Putumayo	1	1	1	1
	Quindío	1	1	2	2
	Risaralda	1	1	2	2
	San Andrés y Providencia	2	2	1	2
	Santander	2	2	2	2
	Sucre	2	2	2	2
	Tolima	1	1	1	1
	Valle del Cauca	2	2	2	2
	Vaupés	2	2	2	2
	Vichada	2	2	2	2
2020	Amazonas	1	1	1	1
	Antioquia	2	2	2	2
	Arauca	1	1	1	1
	Atlántico	2	2	2	2
	Bolívar	2	2	2	2
	Boyacá	2	2	2	2
	Caldas	2	2	2	2
	Caquetá	2	1	2	3
	Casanare	1	1	1	1
	Cauca	2	2	2	2
	Cesar	2	2	2	2
	Chocó	2	2	2	2
	Córdoba	2	2	2	2
	Cundinamarca	2	2	1	2
	Guainía	1	1	1	1
	Guaviare	1	1	2	1
	Huila	2	2	2	3
	La Guajira	2	2	2	2
	Magdalena	2	2	2	2

	Meta	2	2	2	3
	Nariño	2	2	2	2
	Norte de Santander	2	2	2	2
	Putumayo	2	2	2	3
	Quindío	2	2	2	3
	Risaralda	2	2	2	3
	San Andrés y Providencia	2	2	2	3
	Santander	2	2	2	2
	Sucre	2	2	2	2
	Tolima	1	1	2	3
	Valle del Cauca	2	2	2	2
	Vaupés	2	2	2	2
	Vichada	2	2	2	2
2021	Amazonas	1	1	1	1
	Antioquia	2	2	2	2
	Arauca	1	1	1	1
	Atlántico	2	2	2	2
	Bolívar	2	2	2	2
	Boyacá	2	2	2	2
	Caldas	2	2	1	2
	Caquetá	1	1	2	1
	Casanare	1	1	1	1
	Cauca	2	2	2	2
	Cesar	2	2	2	2
	Chocó	2	2	2	2
	Córdoba	2	2	2	2
	Cundinamarca	2	2	2	2
	Guainía	1	1	1	1
	Guaviare	1	1	1	1
	Huila	2	2	2	1
	La Guajira	2	2	2	2
	Magdalena	2	2	2	2
	Meta	1	1	2	1
	Nariño	2	2	2	2
	Norte de Santander	2	2	2	2
	Putumayo	1	1	2	1
	Quindío	1	1	2	1
	Risaralda	1	1	2	1
	San Andrés y Providencia	2	2	2	2

	Santander	2	2	2	2
	Sucre	2	2	2	2
	Tolima	1	1	1	1
	Valle del Cauca	2	2	2	2
	Vaupés	1	1	1	2
	Vichada	2	2	2	2

- **ANEXO 6.**

Ilustración 11. Consolidación de la clasificación Factoclass base de datos violencia de pareja 2020- cluster 1

class: 1

	Test.Value	Class.Mean	Frequency	Global.Mean
Secundaria	4.840	121.675	8	66.385
Urbana	4.765	208.621	8	116.036
Juventud	4.740	100.808	8	55.244
Fin_de_semana	4.737	83.178	8	47.161
Semestre1	4.721	119.919	8	67.849
Adultez	4.704	118.745	8	66.193
Entre_Semana	4.642	143.317	8	79.072
Profesional	4.556	48.425	8	25.132
Compañero_apermanente	4.550	102.465	8	56.071
Ex_compañero_apermanente	4.461	74.347	8	42.712
Semestre2	4.410	106.576	8	58.385
Per_Etnia_NO	4.378	196.620	8	109.121
Union_libre	4.223	116.475	8	62.450
Intolerancia_machismo	4.075	65.866	8	37.608
Casado_a	3.821	23.216	8	12.848
Soltero_a	3.790	71.088	8	37.633
Sin_informacion_Factor	3.676	22.423	8	10.834
Primaria	3.650	53.840	8	32.530
Novio_a	3.648	8.058	8	3.946
Celos	3.514	90.760	8	53.428
Consumodealcohol	3.412	42.226	8	21.540
Esposo_a	3.152	24.518	8	13.996
Ex_Novio_a	2.883	8.046	8	4.555
Rural	2.491	17.874	8	10.197
Otras_razones	2.454	3.711	8	1.728
Ex_esposo_a	2.380	8.350	8	4.255
Adolescencia	2.023	5.329	8	3.501

Ilustración 12. Consolidación de la clasificación Factoclass base de datos violencia de pareja 2020- cluster 2

```
class: 2
                          Test.Value Class.Mean Frequency Global.Mean
Adolescencia                  -2.023      2.892        24       3.501
Ex_esposo_a                   -2.380      2.889        24       4.255
Otras_razones                 -2.454      1.067        24       1.728
Rural                         -2.491      7.639        24      10.197
Ex_Novio_a                    -2.883      3.391        24       4.555
Esposo_a                      -3.152     10.489        24      13.996
Consumodealcohol              -3.412     14.644        24      21.540
Celos                         -3.514     40.983        24      53.428
Novio_a                       -3.648      2.575        24       3.946
Primaria                      -3.650     25.427        24      32.530
Sin_informacion_Factor        -3.676      6.972        24      10.834
Soltero_a                     -3.790     26.482        24      37.633
Casado_a                      -3.821      9.393        24      12.848
Intolerancia_machismo         -4.075     28.189        24      37.608
Union_libre                   -4.223     44.441        24      62.450
Per_Etnia_NO                  -4.378     79.955        24     109.121
Semestre2                     -4.410     42.321        24      58.385
Ex_compañero_apermanente      -4.461     32.167        24      42.712
Compañero_apermanente         -4.550     40.606        24      56.071
Profesional                   -4.556     17.367        24      25.132
Entre_Semana                  -4.642     57.657        24      79.072
Adultez                       -4.704     48.676        24      66.193
Semestre1                     -4.721     50.492        24      67.849
Fin_de_semana                 -4.737     35.156        24      47.161
Juventud                      -4.740     40.056        24      55.244
Urbana                        -4.765     85.174        24     116.036
Secundaria                    -4.840     47.955        24      66.385
>
```

- **ANEXO 7**

Ilustración 13. Consolidación de la clasificación Factoclass base de datos violencia de pareja 2021- cluster 1

```
class: 1
                          Test.Value Class.Mean Frequency Global.Mean
Juventud                       4.887    113.390        10      64.380
Semestre2                      4.857    137.984        10      81.701
Entre_Semana                   4.833    154.806        10      91.188
Secundaria                     4.803    126.170        10      75.237
Fin_de_semana                  4.788     96.395        10      57.812
Urbana                         4.780    231.662        10     137.932
Compañero_apermanente          4.749    117.926        10      64.126
Semestre1                      4.720    113.217        10      67.300
Adultez                        4.683    131.440        10      79.809
Union_libre                    4.213    126.694        10      70.394
Primaria                       4.097     66.190        10      38.221
Profesional                    4.081     53.094        10      32.092
Celos                          3.956    102.215        10      65.614
Per_Etnia_NO                   3.760    194.754        10     126.812
Consumodealcohol               3.735     58.062        10      28.466
Ex_compañero_apermanente       3.670     78.299        10      53.096
Novio_a                        3.640     11.874        10       5.825
Soltero_a                      3.593     83.144        10      49.530
Intolerancia_machismo          3.280     59.588        10      39.899
Sin_informacion_Factor         3.280      3.010        10       1.310
Abandono                       3.203     26.343        10      12.581
Casado_a                       3.196     22.578        10      14.150
Esposo_a                       3.153     24.104        10      14.097
Rural                          2.876     19.539        10      11.069
Per_Etnia_SI                   2.744     56.447        10      22.189
Ex_Novio_a                     2.351     10.768        10       6.571
Otras_razones                  2.323      1.044        10       0.507
Viudo_a                        2.192      0.340        10       0.192
Adulto_Mayor                   2.033      2.277        10       1.485
```

Ilustración 14. Consolidación de la clasificación Factoclass base de datos violencia de pareja 2021-cluster 2

```
class: 2
                        Test.Value Class.Mean Frequency Global.Mean
Adulto_Mayor                -2.033      1.141        23       1.485
Viudo_a                     -2.192      0.128        23       0.192
Otras_razones               -2.323      0.274        23       0.507
Ex_Novio_a                  -2.351      4.746        23       6.571
Per_Etnia_SI                -2.744      7.294        23      22.189
Rural                       -2.876      7.386        23      11.069
Esposo_a                    -3.153      9.747        23      14.097
Casado_a                    -3.196     10.485        23      14.150
Abandono                    -3.203      6.598        23      12.581
Intolerancia_machismo       -3.280     31.338        23      39.899
Sin_informacion_Factor      -3.280      0.571        23       1.310
Soltero_a                   -3.593     34.916        23      49.530
Novio_a                     -3.640      3.196        23       5.825
Ex_compañero_apermanente    -3.670     42.138        23      53.096
Consumodealcohol            -3.735     15.598        23      28.466
Per_Etnia_NO                -3.760     97.272        23     126.812
Celos                       -3.956     49.701        23      65.614
Profesional                 -4.081     22.961        23      32.092
Primaria                    -4.097     26.061        23      38.221
Union_libre                 -4.213     45.915        23      70.394
Adultez                     -4.683     57.360        23      79.809
Semestre1                   -4.720     47.335        23      67.300
Compañero_apermanente       -4.749     40.734        23      64.126
Urbana                      -4.780     97.179        23     137.932
Fin_de_semana               -4.788     41.037        23      57.812
Secundaria                  -4.803     53.092        23      75.237
Entre_Semana                -4.833     63.528        23      91.188
Semestre2                   -4.857     57.230        23      81.701
Juventud                    -4.887     43.071        23      64.380
>
```

- **ANEXO 8.**

Tabla 15. Contrastación departamental según resultados de modelos violencia de pareja 2020-2021

Año	Departamento	k-medias	k-medoides	Metodo de Ward	Factoclass
2020	Amazonas	1	1	1	1
	Antioquia	2	2	2	2
	Arauca	1	1	1	1
	San Andrés	1	1	1	1
	Atlántico	2	2	2	2
	Bogotá	1	1	1	1
	Bolívar	2	2	2	2
	Boyacá	2	2	2	1
	Caldas	2	2	2	2
	Caquetá	2	2	2	2
	Casanare	1	1	1	1
	Cauca	2	2	2	2
	Cesar	2	2	2	2
	Chocó	2	2	2	2
	Córdoba	2	2	2	2

	Cundinamarca	1	1	1	1
	Guainía	2	2	2	2
	Guaviare	2	2	2	2
	Huila	2	2	2	1
	La Guajira	2	2	2	2
	Magdalena	2	2	2	2
	Meta	1	1	1	1
	Nariño	2	2	2	2
	Norte de Santander	2	2	2	2
	Putumayo	2	2	2	1
	Quindío	2	2	2	2
	Risaralda	2	2	2	2
	Santander	2	2	2	2
	Sucre	2	2	2	2
	Tolima	1	1	1	1
	Valle del Cauca	2	2	2	2
	Vaupés	2	2	2	1
2021	Amazonas	1	1	1	1
	Antioquia	2	2	2	2
	Arauca	1	1	1	1
	San Andrés	1	1	1	1
	Atlántico	2	2	2	2
	Bogotá	1	1	1	1
	Bolívar	2	2	2	2
	Boyacá	2	1	1	1
	Caldas	2	2	2	2
	Caquetá	2	2	2	2
	Casanare	1	1	1	1
	Cauca	2	2	2	2
	Cesar	2	2	2	2
	Chocó	2	2	2	2
	Córdoba	2	2	2	2
	Cundinamarca	1	1	1	1
	Guainía	1	1	1	1
	Guaviare	2	2	2	2
	Huila	1	1	1	1
	La Guajira	2	2	2	2
	Magdalena	2	2	2	2
	Meta	1	1	1	1
	Nariño	2	2	2	2
	Norte de Santander	2	2	2	2
	Putumayo	2	2	2	1
	Quindío	2	2	2	2
	Risaralda	2	2	2	2
	Santander	2	2	2	2

	Sucre	2	2	2	2
	Tolima	1	1	1	1
	Valle del Cauca	2	2	2	2
	Vaupés	1	1	1	1
	Vichada	2	2	2	2