# UNIVERSIDAD NACIONAL DEL ALTIPLANO

## FACULTAD DE INGENIERÍA ESTADÍSTICA E INFORMÁTICA

## ESCUELA PROFESIONAL DE INGENIERÍA ESTADÍSTICA E INFORMÁTICA

# MINERÍA DE DATOS PARA EXPLORAR INFORMACIÓN QUE NOS PERMITA ENCONTRAR QUÉ RELACIÓN TIENEN LOS REPORTES ATENDIDOS POR EL PROGRAMA NACIONAL CONTRA LA VIOLENCIA FAMILIAR DEL AÑO 2020.

**TESIS**

**PRESENTADA POR:**

**Bach. LEHIDY YENIFER BERRIOS MAMANI**

**PARA OPTAR EL TÍTULO PROFESIONAL DE:**

**INGENIERO ESTADÍSTICO E INFORMÁTICO**

**PUNO – PERÚ**

**2023**

# DEDICATORIA

*A Dios todo poderoso por guiarme, por fortalecerme y acompañarme todos los días de mi vida.*

*Con mucha gratitud y amor a mis padres, Irma Mamani Mamani y Teófilo Berrios Chura, por su esfuerzo, apoyo y compresión.*

*A mis hermanitos Anthony y Robinho por su apoyo incondicional y también a mis queridos tíos Zenovia y Adolfo por haber confiado en mí a lo largo de mi formación académica.*

***Lehidy Yénifer Berrios Mamani***

# AGRADECIMIENTOS

*Agradezco a Dios, por haberme guiado y acompañado a lo largo de mi formación académica, por ser mi luz en mi camino y por darme sabiduría, fortaleza para alcanzar mis objetivos.*

*Agradecer también a mi asesor al Dr. LEONEL COYLA IDME por su gran apoyo en compartir sus conocimientos al desarrollo de este proyecto.*

***Lehidy Yénifer Berrios Mamani***

# ÍNDICE GENERAL

## CAPÍTULO IV

## RESULTADOS Y DISCUSIÓN



**Área**: Estadística e informática

**Línea:** Análisis multivariado, big data, minería de datos e investigación de mercados

**FECHA DE SUSTENTACIÓN:** 4 de enero del 2023

# ÍNDICE DE TABLAS

# ÍNDICE DE FIGURAS

# ÍNDICE DE ANEXOS

# ÍNDICE DE ACRÓNIMOS

**ACP**: Análisis de Componentes Principales

**CP**: Componente Principal

**K:** Número de Clúster generados por el algoritmo “k-means”

**KMO**: Estadístico de medida de adecuación muestral de Kaiser-Mayer-Olikin

**MD**: Minería de datos

**NCT:** Número de consultas total

# RESUMEN

Los reportes del Programa Nacional para la Prevención y Erradicación de la Violencia contra las Mujeres e Integrantes del Grupo Familiar, son generados por los trabajadores de dicha institución para almacenar registros sobre, el número de consultas de violencia física, violencia psicológica, violencia sexual, violencia económica, así como el género de las personas, entre otras variables. Así El presente trabajo de investigación se realiza con el objeto de aplicar minería de datos para explorar información que nos permita encontrar qué relación tienen los reportes atendidos por el Programa Nacional para la Prevención y Erradicación de la Violencia contra las Mujeres e Integrantes del Grupo Familiar. La metodología empleada es de enfoque cuantitativo, descriptivo, aplicado. La muestra es igual a la población conformada por 230 reportes, se utilizó la metodología CRISP-DM para el desarrollo de la minería de datos, concluyendo que se logró desarrollar la minería de datos utilizando técnicas de Clusterización para los grupos de individuos mediante el algoritmo de "K means" con 2 Clústers integrados por el 47.83% de los reportes en el primero y 52.17% de los reportes en el segundo que tienen una suma total de cuadrados de 29971294 de distancia entre todos los puntos a sus centroides, para agrupar las variables del Análisis de Componentes Principales agrupando las variables en 2 Clústeres o Componentes Principales, el primero con 85.47% de representación denominado "Clúster en el que no incluye la violencia económica", el segundo con un 76.28% de representación denominado "Clúster en el que la violencia económica es muy importante".

**PALABRAS CLAVE:** Minería de datos, violencia familiar, Clústeres, CRISP-DM, K-means.

# ABSTRACT

The reports of the National Program for the Prevention and Eradication of Violence against Women and Family Members are generated by the workers of this institution to store records on the number of consultations of physical violence, psychological violence, sexual violence, economic violence, as well as the gender of the persons, among other variables. Thus, the present research work is carried out with the purpose of applying data mining to explore information that will allow us to find the relationship between the reports attended by the National Program for the Prevention and Eradication of Violence against Women and Family Members. The methodology used is quantitative, descriptive and applied. The sample is equal to the population formed by 230 reports, the CRISP-DM methodology was used for the development of data mining, concluding that it was possible to develop data mining using Clustering techniques for the groups of individuals by means of the "K means" algorithm with 2 Clusters integrated by 47.83% of the reports in the first one and 52. 17% of the reports in the second one that have a total sum of squares of 29971294 of distance between all the points to their centroids, to group the variables of the Principal Component Analysis grouping the variables in 2 Clusters or Principal Components, the first one with 85.47% of representation called "Cluster in which economic violence is not included", the second one with 76.28% of representation called "Cluster in which economic violence is very important

**KEY WORDS:** Data mining, family violence, Clusteres, CRISP-DM, K-means.

# CAPÍTULO I

## INTRODUCCIÓN

La violencia familiar es un problema que aqueja en todas las sociedades, especialmente la violencia contra la mujer, causando estragos en la convivencia, el desarrollo de los niños, degradando la autoestima de las personas involucradas, entre muchas otras consecuencias, por ello se busca alternativas de solución.

Según Muñiz et al. (1998) la complejidad de la violencia no debe utilizarse como excusa para la complacencia hacia ella; el mero hecho de tratar a los integrantes de una familia con falta de respeto, reprenderlos con dureza, etc., imprime en ellos una dosis diaria de violencia que los convertirá a los niños en adultos agresivos en el futuro.

El concepto de DM ha surgido y sigue desarrollándose a través de la intersección de la investigación en campos como las bases de datos, el aprendizaje automático, el reconocimiento de patrones, la estadística, la teoría de la información, la inteligencia artificial, el razonamiento con incertidumbre, la visualización de datos y la informática de alto rendimiento (Riquelme, Ruiz & Gilbert, 2006).

La minería de datos hoy en día es utilizada por diversos sectores desde la agricultura, telecomunicaciones, banca, así como las ciencias sociales en este sentido, la minería de datos permite comprender información que a simple vista el ser humano no puede abstraer por sí mismo, así la minería de datos brinda herramientas para descubrir patrones en los datos que son previamente desconocidos (Carrascal &Jiménez, 2018).

Por ello en este trabajo de investigación se pretende utilizar la minería de datos para dar solución a la comprensión de los reportes de la violencia familiar.

La presente tesis desarrollada cuenta con la siguiente estructura:

**Capítulo I:** Se presenta la descripción del problema, formulación del problema, justificación de esta, como también se muestran los objetivos e hipótesis de investigación respecto a la violencia familiar.

**Capítulo II:** Se detallan la bibliografía utilizada, conceptos utilizados y antecedentes referentes a la minería de datos y la violencia familiar.

**Capítulo III:** Aquí presenta materiales y métodos, donde se detalla la población y muestra, diseño, alcance de la investigación, técnicas e instrumentos de recolección de datos, metodología estadística utilizada.

**Capítulo IV:** Se exponen los resultados en base a los objetivos propuestos la discusión de estos, constituidos de cada objetivo planteado en la investigación, demostrando los resultados mediante la minería de datos con la metodología propuesta, encontrándose clústeres.

Finalmente se comprende la bibliografía utilizada y anexos.

## 1.1 DESCRIPCIÓN DEL PROBLEMA

El Programa Nacional para la Prevención y Erradicación de la Violencia contra las Mujeres e Integrantes del Grupo Familiar es una entidad del estado peruano y tiene como propósito diseñar, ejecutar acciones, crear políticas de atención, prevención y apoyo a las personas involucradas en hechos de violencia familiar y/o sexual, contribuyendo así a mejorar la calidad de vida de la población.

La violencia siempre ha sido una parte desagradable de la experiencia humana, constituyendo un desafío en la salud pública situándose como una fuente de condiciones cuyo desenlace es la muerte.

La violencia familiar definida como el daño físico, sexual o psicológico que sufre un integrante causado por uno o más integrantes de dicha familia, induce un patrón de conductas, es también uno de los que mayores secuelas deja a la víctima como baja autoestima, retraso en las habilidades verbales, depresión, entre muchos otros.

Los reportes del Programa Nacional para la Prevención y Erradicación de la Violencia contra las Mujeres e Integrantes del Grupo Familiar, son generados por los trabajadores de dicha institución para almacenar registros sobre, el número de consultas de violencia física, violencia psicológica, violencia sexual, violencia económica, así como el género de las personas, entre otras variables.

Saber con anticipación el comportamiento de la violencia familiar, nos permite tomar decisiones, así ser efectivo en las medidas prevención de la violencia familiar.

La minería de datos es una herramienta informática que nos permite desarrollar técnicas tienen como objetivo segmentar y/o predecir un comportamiento futuro, así como modelos de riesgo.

Los algoritmos son desarrollados a medida y consideraremos variables estadísticas acorde a problema en específico mediante herramientas matemáticas y computacionales, por ello se ha ido propagando rápidamente en los últimos años, ya que demostró ser muy eficaz en campos como ingenierías, ciencias sociales, bioinformática, secuenciamiento del ADN, datos espaciales y entre otros campos ya sean de carácter científico o simplemente para dar solución a problemas.

El presente trabajo de investigación se enfoca en la aplicación de minería de datos para explorar información permite hallar relaciones en datos de carácter de las ciencias sociales, así como en los reportes atendidos por el Programa Nacional para la Prevención y Erradicación de la Violencia contra las Mujeres e Integrantes del Grupo Familiar. Así Ello nos permite comprender el caso de estudio y hallar patrones de la información respecto a la violencia de la mujer.

## 1.2 FORMULACIÓN DEL PROBLEMA

### 1.2.1 Problema general

¿Al aplicar minería de datos esta nos permitirá encontrar la relación en los reportes atendidos por el Programa Nacional para la Prevención y Erradicación de la Violencia contra las Mujeres e Integrantes del Grupo Familiar?

### 1.2.2 Problemas específicos

¿La minería de datos nos permitirá realizar la comprensión del caso de estudio y la información de los reportes atendidos por el Programa Nacional para la Prevención y Erradicación de la Violencia contra las Mujeres e Integrantes del Grupo Familiar?

¿El análisis de minería de datos aportará información descriptiva de los reportes atendidos por el Programa Nacional para la Prevención y Erradicación de la Violencia contra las Mujeres e Integrantes del Grupo Familiar?

## 1.3 JUSTIFICACIÓN DE LA INVESTIGACIÓN

### Teórica

Este trabajo de investigación se desarrolla con el propósito de aportar conocimientos científicos sobre el uso de la minería de datos en las ciencias sociales, así como incorporarse como un antecedente a trabajos de investigación futuros, también exponer la teoría de sus algoritmos de agrupación más conocidos como el “K means” y “ACP”.

### Práctica

La minería de datos aplicado a las ciencias sociales es una disciplina emergente que desarrolla y aplica métodos sobre datos que vienen de entornos sensibles, y se usa para entender la complejidad de sus variables, tiene una alta relevancia en las organizaciones gubernamentales para la resolución de problemas concernientes a la generación de información y la toma de decisiones, ya que la administración de la información a medida que se inducen a las nuevas tecnologías, esta se vuelve más compleja; esta nos sirve como una herramienta que nos permite abordar variables de violencia familiar, y al ser analizadas brota información relevante.

### Metodológica

Se desarrolla esta investigación para aportar, exhibir y desarrollar una metodología propuesta denominada por “CRISP-DM” de la minería de datos, así como dentro de la violencia familiar que demanda servicios de análisis de datos con capacidades

tales que permitan evidenciar patrones de comportamiento útiles para la resolución de conflictos.

Por ello el presente trabajo de investigación emerge a la necesidad de utilizar las técnicas de minería de datos en problemas de las ciencias sociales.

## 1.4 OBJETIVOS DE LA INVESTIGACIÓN

### 1.4.1 Objetivo General

Aplicar minería de datos para explorar información que nos permita encontrar la relación de los reportes atendidos por el Programa Nacional para la Prevención y Erradicación de la Violencia contra las Mujeres e Integrantes del Grupo Familiar.

### 1.4.2 Objetivos Específicos

Realizar la comprensión del caso de estudio y la información de los reportes atendidos por el Programa Nacional para la Prevención y Erradicación de la Violencia contra las Mujeres e Integrantes del Grupo Familiar.

Realizar el análisis de minería de datos descriptiva de los reportes atendidos por el Programa Nacional para la Prevención y Erradicación de la Violencia contra las Mujeres e Integrantes del Grupo Familiar.

## 1.5 HIPÓTESIS DE LA INVESTIGACIÓN

### 1.5.1 Hipótesis General

La minería de datos nos permite explorar información previamente desconocida y la relación que tienen los reportes atendidos por el Programa Nacional para la Prevención y Erradicación de la Violencia contra las Mujeres e Integrantes del Grupo Familiar.

# CAPÍTULO II

# REVISIÓN DE LITERATURA

## 2.1 ANTECEDENTES

### Internacional

Carrascal, Carrascal & Saldarriaga (2015) concluyen que

- En el análisis predictivo se utilizan con mayor frecuencia las redes neuronales y los árboles de decisión.
- En el análisis descriptivo se utiliza frecuentemente el método particional k-means.
- Como metodología de minería de datos para aplicaciones de salud se utiliza con mayor frecuencia CRISP-DM.
- Finalmente, en la etapa de desarrollo WEKA se desataca como la herramienta preferida en aplicaciones de salud.

Echeverry & Salazar (2017) concluyen que: “la técnica de MD implica, en sus primeras fases, limpieza, ajuste y transformación, de acuerdo a las necesidades del estudio. Estos procesos son necesarios porque, a pesar de todos los controles para la gestión de la información, los usuarios incurren de modo permanente en errores de digitación que hacen difícil la coincidencia de los datos y la futura gestión del conocimiento”.

Regalado (2019) concluyó que: “ la técnica predictiva de minería de datos implementada en MATLAB utilizando como clasificador una red neuronal, obtuvo un excelente desempeño en el entrenamiento supervisado, los parámetros configurados fueron 3 capas y 100 neuronas haciendo la ligera y con exactitud del 86,86% en entrenamiento”.

Abad-Vich (2016) “afirma que la aplicación desarrollada para la extracción de reglas descriptivas cumple su función de proporcionar una interfaz sencilla e intuitiva para la realización de experimentos y obtención de resultados que muestren que se ha producido una extracción del conocimiento de los datos utilizados”.

Cordovilla (2019) “determinó que la aplicación de la Minería de Datos en la determinación de los perfiles de estudiantes en la Unidad de Titulación permite considerar la orientación académica del estudiante para la asignación de los temas para Trabajo de Titulación”.

Padilla (2019) concluye que:

- La obtención de patrones en cada uno de los procesos se realiza basado en características de los datos, así para generar las reglas de asociación es necesario una estructura de archivo que identifique si un producto es parte de una transacción, lo que se consigue resumiendo la información de venta de productos.
- La parte predictiva utiliza modelos de regresión que son aplicados dependiendo de la dimensionalidad buscada: tiempo o espacio.
- La estimación a futuro utilizando una serie de tiempo, se realiza en base a los atributos de tiempo y valor de venta.

Acosta, Medina & González (2020) concluye: “La minería de datos con el método y las técnicas aplicadas son herramientas esenciales y de alto valor para analizar oportunamente el comportamiento y el mejor uso del espacio público”.

Orea, Vargas & Alonso (2005) demostraron que: “las técnicas de minería de datos que usamos proporcionan una manera que permite determinar aquellos alumnos que son candidatos a desertar”.

Lagla et al. (2019) afirman que “la minería de datos se ha convertido en una herramienta estratégica especialmente en las empresas porque por medio de ella se podrá realizar una extracción de la información, un análisis de datos y a su vez estos resultados permitirán tomar mejores decisiones sobre la empresa”.

**Nacional**

Carrascal & Jiménez (2018) concluyen:

- “Los factores que intervienen en la identificación de los estilos de aprendizaje según el cuestionario del modelo de estilos de aprendizaje de VARK y el modelo de estilos de aprendizaje de CHAEA. Asociaciones existentes entre los dos modelos de estilos de aprendizaje aplicados al estudio. A igual que las relaciones existentes entre las áreas de conocimiento y cada uno de los modelos”.
- “De la repitencia se identifican las características de los estudiantes con riesgo de repitencia. Queda abierta la gama de opciones de estudios aplicando técnicas de minería de datos a nivel de las instituciones educativas de básica secundaria, debido a que la mayoría de referentes se remiten a estudios de educación superior. Para ello es necesario promover la cultura del valor de la calidad de datos en las organizaciones y su gestión para utilizar en futuros estudios”.

Neyra & Bazán (2020) “se estimó la energía empleando integración numérica. Se ha utilizado el software Weka para elaborar un modelo matemático mediante una red neuronal el cual permite generalizar el comportamiento y su validación de datos que se consigue mediante estadísticos (coeficiente de correlación, error medio, error cuadrático, error relativo absoluto y error cuadrático relativo). Según la máxima demanda en todo el sector, se estima que requiere de 21 paneles fotovoltaicos, para cubrir la demanda de corriente que debe entregar el sistema para cubrir la demanda eléctrica, considerando un

ángulo 7.6° haciendo alusión al mes de junio por ser el mes más crítico del año con un valor de (4,39 kWh/m2) /día “.

Rioja (2020) “se segmentó el total de leads a través de la aplicación de un algoritmo basado en la segmentación de k-means y las políticas internas de la organización, resultando en veinte (20) perfiles de consumidor y seis (06) subgrupos por cada uno de ellos”.

Ávalos, Torreblanca & Mamani (2020) “se utilizó la técnica de minería de datos y el método de K medias para hallar asociación entre transferencias económicas, desarrollo humano y pobreza, usando datos pertenecientes a organismos oficiales del Estado peruano. Se encontró una asociación significativa entre exportaciones con inversiones mineras y transferencias económicas y baja asociación de las inversiones con las transferencias. Además, no todos los distritos que recibieron fondos de la minería fueron exitosos respecto a la disminución de la pobreza”.

**Local**

Choque (2019) demostró que:

- Es posible aplicar la técnica Clustering con el algoritmo K means pertenecientes a la Minería de Datos para la identificación de factores de deserción universitaria en programas de pre grado.
- La técnica Clustering se escogió debido a que es una técnica descriptiva de Minería de datos. La investigación aplicó esta técnica de modo que se analizó el estado de los datos históricos de la escuela para así obtener una visión más amplia de la situación de deserción.
- Se logró preparar y depurar la información obtenida de la institución. Para lo cual paso por procesos de extracción, transformación y carga hasta lograr la base de

datos necesaria para la construcción del modelo de deserción y posterior desarrollo de las fases de Minería de datos.

Melo (2018) “el modelo CRISP – DM es adecuado para el uso de minería de datos y sumamente importante en la realización del proceso de extracción, normalización, limpieza y carga de datos existentes, como también para eliminar información innecesaria, inconsistente, redundante o errónea en el diseño del Data Mart”.

Ticona (2020) se concluye que “los patrones de árboles de decisiones aplicados a la resiliencia y el optimismo en estudiantes de universidades licenciadas del Departamento de Puno – Perú, 2019 se asocian positivamente, es decir que existe una relación significativa entre las variables y que una depende la otra. En los patrones encontrados existen una mayor cantidad de estudiantes optimistas que son de la Universidad Nacional del Altiplano Puno, las estudiantes mujeres son más resilientes que los hombres, esto puede deberse a que las mujeres sean más conscientes de sus potencialidades y limitaciones, lo que hace que puedan enfrentar las adversidades y los retos, conocen sus fortalezas y habilidades, así como sus limitaciones y defectos”.

Rosas (2020) “dentro del análisis de minería de datos se pudo validar el proceso, llegando a determinar las siguientes asociaciones: La escuela de Ingeniería de Minas al presentarla mayor actividad obtiene una mayor asociación en los módulos: Actividad a las 00 horas FORO a horas 6:00 a.m., Asignación de Tareas a horas 20:00 p.m., examen y archivo en distintas horas, pero solo el módulo Folder aparece en esta escuela. Con la metodología KDD, implementando Minería de Datos en el programa WEKA y aplicando la técnica de asociación A PRIORI, se pudo encontrar los siguientes patrones desconocidos: En el módulo Vista la escuela de Ingeniería de Minas e Ingeniería Topográfica y Agrimensura presentan un mismo comportamiento con una confiabilidad del 97%”.

## 2.2 MARCO TEÓRICO

### 2.2.1 Violencia

La violencia es un problema importante de salud pública que desarrolla consecuencias muy negativas en todos los miembros una familia, es considerada como un grave obstáculo para el desarrollo y la paz. En estudios realizados a escala mundial se han identificado múltiples, dañinas y dolorosas consecuencias físicas y psicológicas para la persona que sufre violencia (Walton & Pérez, 2019).

### 2.2.2 Violencia física

Arenas (2015) la más clara y evidente del maltrato por constituir una invasión al espacio físico de la víctima, esta suele ser de dos maneras; por contacto directo con el cuerpo mediante: patadas, empujones, pellizcos, jalón de pelo, golpes, cualquier tipo de contacto físico no deseado y la otra manera es limitar sus movimientos, privando a la víctima en un espacio limitado.

### 2.2.3 Violencia sexual

“La violencia Sexual es uno de los mayores atentados cometidos contra los derechos humanos en nuestros tiempos. Es un problema que obedece a las estructuras jerárquicas patriarcales que reproducen una cultura donde las mujeres son vistas como objetos desechables y maltratable” (Toro, 2013, p. 1).

Reconocer las situaciones de Violencia Sexual implica trabajar sobre nuestros prejuicios, identificar los aspectos de la cultura y repensar nuestras propias experiencias de abuso, nuestro concepto de sexualidad y nuestra idea de consentimiento. (Hermosa & Polo, 2018).

### 2.2.4 Violencia en la familia

Vaca & Díaz (2009) señalan que la violencia es uno de esos comportamientos con alta capacidad para reproducirse, especialmente sus consecuencias. La familia puede convertirse en un agente que propicia este tipo de conductas debido a que está comprobado que la misma es una pequeña sociedad de importancia y constituye un ambiente constante de aprendizaje grupal e individual de normas de convivencia.

"El reconocimiento de la realidad de la violencia como una construcción que se presenta y legítima en la práctica familiar cotidiana, hace necesario generar procesos de reflexión frente a las prácticas de socialización que ocurren en la vida familiar, pues en los procesos de socialización se expresan las concepciones que una cultura tiene acerca del desarrollo de sus miembros" (Vaca & Díaz, 2009).

Es la perpetrada en el hogar o unidad doméstica, generalmente por un integrante de la familia que vive con la víctima, que puede de sexo masculino o femenino, infantil, adolescente o adulto, con el empleo deliberado de la fuerza o maltrato psicológico (Rodríguez et al., 2018).

Según Corsi & Bonino (2003) la violencia en la familia incluye a todos los tipos de abusos cotidianos que involucran a los miembros de una familia. Se refieren a las distintas formas de relación abusiva que caracterizan de modo permanente o cíclico al vínculo familiar. Enfatizan que un miembro de la familia, independientemente de su raza, sexo y edad, puede ser víctima o victimario. Los integrantes de una familia tienen misma probabilidad todos los miembros de ser tanto víctimas como victimarios, por desarrollo de las dinámicas familiares.

"La violencia contra la mujer constituye una de las violaciones más grandes a los derechos humanos, dado que menoscaba el desarrollo personal y limita el ejercicio de los

derechos civiles, económicos, sociales y culturales propios del ser humano" (Parra & Villalobos, 2017, p.12).

### 2.2.5 La nueva victimología

En la actualidad no solo tiene la función describir fenomenológicamente, la vinculación de acciones entre la víctima y el victimario también definir y caracterizar, además de la condiciones en las que sucede, bajo la premisa del descubrimiento de las hipótesis planteadas de descubrir las acciones y las condiciones en las que el victimario asecha a la víctima, la percepción del mismo referente a ella, o las justificaciones en las que se excusa el victimario posterior a la acción delictiva (Dumont, 2015).

### 2.2.6 Minería de datos

Gorbea (2013) propone que la informática y la minería de datos son herramientas útiles para abordar procesos esenciales donde se aplican una serie de métodos inteligentes para poder extraer y descubrir patrones mediante el uso de algoritmos como regresión lineal de variable continua, regresión logística de variable discreta, algoritmos de asociación (clustering), entre otros para una toma de decisiones en el área de estudio o de conocimiento.

Los algoritmos en la minería de datos, independientemente del sector en el que va ser empleado, o del modelo, de la complejidad de los datos; se seleccionan procediendo a su aplicación para encontrar la información que no era visible o que está incompleta que nos permitan plantear soluciones a los problemas abordados (Witten & Frank, 2002).

La minera de datos es aplicada en una variedad de ciencias, disciplinas y saberes como la educación, el comercio, finanzas, medicina, en las actividades agropecuarias, agroforestales, en las ciencias sociales y la gestión gubernamental. En general puede ser utilizada en todas las actividades que realiza y procesa el ser humano que presenten la

generación de datos que requieran ser analizados para proporcionar un conocimiento (Rojas & Gomez 2014).

### 2.2.7 Minería de datos descriptiva

El análisis descriptivo permite descubrir conocimiento oculto en un conjunto de datos mediante agrupaciones, reglas de asociación y selección de actores (Brusil, 2020).

Se tiene un conjunto de datos de entrada y se busca establecer patrones para realizar el etiquetado de grupos o patrones previamente desconocidos (Cardoso et al., 2020.

El análisis por de minería de datos descriptiva se realiza a través de los algoritmos, que pretenden realizar la agrupación de observaciones en subgrupos *"clusters"* para que las observaciones en cada grupo se asemejen entre sí (Cambronero & Moreno, 2006).

Dentro de los algoritmos de minería de datos descriptiva se encuentran:

- K-means
- Clúster jerárquico
- Análisis de componentes principales
- Análisis de correspondencias
- Escalamiento multidimensional
- Otros

### 2.2.8 K –means

El algoritmo de "K-means" es uno de los métodos de clusterización más conocidos. En el que se agrupa las observaciones en k conjuntos, el número total de miembros del grupo es determinado por el cálculo del centro para cada conjunto y

asignando cada observación al conjunto con el centro más próximo a este (Piñerez, Ramírez & Escobar, 2017).

Según Sánchez (2021) este es el algoritmo de la minería de datos más popular, debido a su sencillo entendimiento, ya que solo se tiene que elegir el correcto parámetro "K". También es denominado el algoritmo de las k medias móviles porque en cada iteración se recalculan los centros de los agrupamientos y cambiando su valor.

Plaza & Cardozo (2018) el algoritmo requiere de 3 parámetros, X, que es el conjunto de observaciones de entrada, K, que es el número de agrupamientos que debe encontrar el N_iter, el número máximo de iteraciones.

Según Plaza & Cardozo (2018) el algoritmo de las K-medias se puede escribir en pseudo código de la siguiente manera:

- K-medias (X, K, N_iter)
- Entrada: X: un conjunto de N variables $\{X_1, X_2, \ldots, X_p\}$
- K: número de agrupamientos
- N_iter: número máximo de iteraciones
- Salida: $S_1, S_2, \ldots, S_k$: K conjuntos de patrones $Z_1, Z_2, \ldots, Z_k$, de los k grupos

**Algoritmo:**

***Paso 1**: Inicializar centros aleatoriamente de los k grupos*

***Paso 2:** Asignación de Observaciones*

***Si t ≤ iter***

*Hacer para todas las observaciones:*

*j=1:p k = 1:K*

*Calcular las distancias de la observación a cada uno de los centros*

$$d_{jk} = \sqrt{(X_{j1} - Z_{k1})^2 + (X_{j2} - Z_{k2})^2 + \cdots + (X_{jN} - Z_{kn})^2}$$

$$d_{jk} = \sqrt{\sum_{i=1}^{n} (X_{ji} - Z_{ki})^2}$$

*Determinar a qué grupo pertenece la observación*

*Min = min ($d_{j1}$, $d_{j2}$, …, $d_{jk}$)*

*Fin hacer*

***Fin si***

***Paso 3:***

*"Actualización de centros, recalculando los centros con el promedio aritmético de los elementos que pertenecen al grupo"* (Plaza & Cardozo, 2018).

### 2.2.9 Número óptimo de Clústers

Como se observa el minero de datos es quien especifica el número de agrupaciones o clúster a crear; preferiblemente, debe utilizarse el número óptimo de conglomerados, se puede utilizar para estudiar la distancia de separación entre los grupos resultantes (Plaza & Cardozo, 2018).

### 2.2.10 Método del codo

La idea básica de los métodos de partición de clúster, como el clustering de k-means, es definir los Clústers de forma que se minimice la variación total intra-Clúster (conocida como variación intra-cluster total o suma cuadrada total intra-cluster) como se muestra a continuación (Bholowalia & Kumar, 2014).

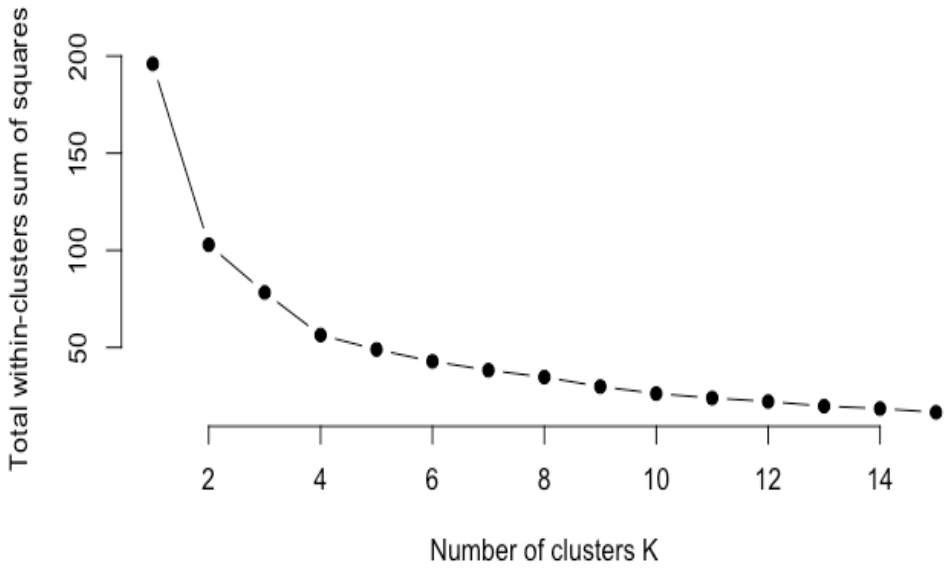



Figura 01.Método del codo

**Fuente:** https://afit-r.github.io/kmeans_clustering#fn:kauf

1. Calcule el algoritmo de agrupamiento (p. ej., K means) para diferentes valores. Por ejemplo, de 1 a 10 grupos
2. Para cada *k*, calcule la suma total del cuadrado dentro del grupo (wss)
3. Trace la curva de wss según el número de grupos *k*.
4. La ubicación de una curva (codo) en la parcela generalmente se considera como un indicador del número apropiado de conglomerados.

### 2.2.11 Método de Average Silhouette

Kaufman & Rousseeuw (2009) el enfoque de average Silhouette mide la calidad de agrupamiento. Es decir, determina qué tan bien se encuentra cada objeto dentro de su grupo. Un ancho de silueta promedio alto indica una buena agrupación. El método de la silueta promedio calcula la silueta promedio de las observaciones para diferentes valores

de k. El número óptimo de conglomerados k es el que maximiza la silueta promedio sobre un rango de valores posibles para k.

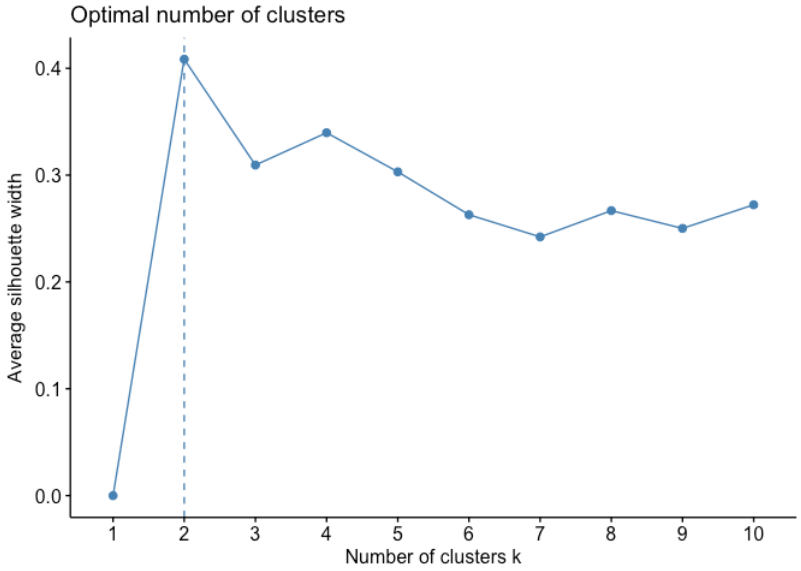



Figura 02.Método de Average Silhouette

**Fuente:** https://afit-r.github.io/kmeans_clustering#fn:kauf

### 2.2.12 Análisis de componentes principales (ACP)

El análisis en componentes principales es una técnica de interdependencia entre variables que nos permite estudiar el tipo de relaciones lineales que existen entre las variables cuantitativas (numéricas), sin considerar a priori, ninguna estructura, ni de variables, ni de individuos (Martínez, 2016).

"El análisis de componentes principales se crean factores que resultan de la combinación de las variables observables y cuyo cálculo se basa en aspectos matemáticos sin tener en cuenta su interpretabilidad teórica o aplicada" (López & Gutiérrez, 2019, p. 3).

Pérez (2008) el análisis de componentes principales busca descubrir variables latentes no observables directamente. Tiene como objetivo descubrir la estructura

subyacente de un conjunto de datos cuantitativos definiendo un pequeño número de agrupaciones comunes que expliquen la mayor parte de la varianza observada.

### 2.2.13 Círculo de Correlación

Abdi & Williams (2010) se conoce como gráficos de correlación de variables. Muestra las relaciones entre todas las variables. Se puede interpretar de la siguiente manera:

- Las variables correlacionadas positivamente se agrupan.
- Las variables correlacionadas negativamente se colocan en lados opuestos del origen del gráfico (cuadrantes opuestos).
- La distancia entre las variables y el origen mide la calidad de las variables en el mapa de factores. Las variables que están alejadas del origen están bien representadas en el mapa de factores.

### 2.2.14 Calidad de la representación

La calidad de representación evalúa en qué medida es representada una variable en un componente o Clúster Abdi & Williams (2010) así la calidad de representación de las variables en el mapa de factores se llama cos2 (coseno cuadrado).

- Un cos2 alto indica una buena representación de la variable en el componente principal. En este caso la variable se posiciona cerca de la circunferencia del círculo de correlación.
- Un cos2 bajo indica que la variable no está perfectamente representada por las PC. En este caso la variable está cerca del centro del círculo.

### 2.2.15 Contribución de las variables

Abdi & Williams (2010) las contribuciones de las variables para explicar la variabilidad en un componente principal dado se expresan en porcentaje.

- Las variables que están correlacionadas con CP 1 (es decir, Dim.1) y CP 2 (es decir, Dim.2) son las más importantes para explicar la variabilidad en el conjunto de datos.
- Las variables que no se correlacionan con ningún CP o se correlacionan con las últimas dimensiones son variables de baja contribución y podrían eliminarse para simplificar el análisis general

### 2.2.16 Minería de datos predictiva

Según Torres & Cardenas (2021) "nos permite diferenciar entre distintas clases de una variable, crear una nueva clasificación conceptual, seleccionar los atributos más representativos en la variable a predecir, y ser capaces de predecir secuencias lógicas Es decir que un algoritmo de minería de datos es un conjunto de cálculos y reglas heurísticas que permite crear un modelo de minería de datos a partir de los datos, el algoritmo analiza primero los datos proporcionados, en busca de patrones específicos".

El análisis predictivo permite predecir acciones futuras mediante el análisis de un histórico de datos, algunos ejemplos son:

- Arboles de decisión ID3
- Arboles de decisión C4.5
- Regresión lineal
- Regresión logística

- Otros

### 2.2.17 Minería de datos en la Industria 4.0

El término industria 4.0 se refiere a un nuevo modelo de organización y de control de la cadena de valor a través del ciclo de vida del producto y a lo largo de los sistemas de fabricación apoyado y hecho posible por las tecnologías de la información (Del Val, 2016).

Con un número creciente de productos y sistemas inteligentes en las fábricas, el mercado y en los sistemas de gestión pública, la cantidad de datos de que se disponen permitirá identificar patrones e interdependencias, analizar los procesos y descubrir ineficiencias e incluso predecir eventos futuros. Con ello se abrirán nuevas oportunidades, no sólo de mejora de la eficiencia, sino de descubrimiento de servicios para el cliente y/o usuario, al que se conocerá mucho mejor (Del Val, 2016)

Prada (2021) la Industria 4.0 beneficia al Perú desde distintas perspectivas. La digitalización de los procesos operativos hará que disminuyamos los costos, además de propiciar el uso eficiente de los recursos, esto fomentará que tanto organizaciones públicas como privadas enfrenten los retos que toda revolución implica

Se propone constantemente el desarrollo de unas nuevas herramientas de monitorización y control inteligente para la recopilación y análisis de todos los datos producidos, desde el punto de vista económico, el hecho de capturar y analizar eficientemente grandes cantidades de datos tiene como consecuencia directa la mejora de calidad y productividad respecto a los sistemas tradicionales (Escobar & Morales, 2018).

## 2.3 TÉRMINOS UTILIZADOS

**Reportes atendidos:** Son informes creados en un lapso de tiempo determinado, normalmente se registra con la fecha que es publicado y que a su vez contiene información sobre las consultas telefónicas de las personas que acuden

**Clúster:** Agrupación creada artificialmente mediante alguna técnica de minería de datos

**Componente principal:** Clúster al que está asociado una relación de variables correlacionadas entre sí.

# CAPÍTULO III

# MATERIALES Y MÉTODOS

## 3.1 POBLACIÓN

La población está constituida por los reportes atendidos por el programa nacional contra la violencia familiar desde el 5 de enero del 2014 hasta el 14 de febrero del 2020 un total de 230 reportes.

## 3.2 MUESTRA

La muestra es igual a la población ya que se cuenta con todos los reportes del programa nacional contra la violencia familiar registrados en su base de datos.

## 3.3 DISEÑO DE INVESTIGACIÓN

Según Hernández (2018) de enfoque cuantitativo porque "las variables incluidas en el estudio son medidas y sus resultados dan a conocer una magnitud". Ya que las variables analizadas proceden de una base de datos estructurada respecto a la suma total del tipo de violencia son de naturaleza cuantitativa.

## 3.4 ALCANCE DE INVESTIGACIÓN

De alcance descriptivo porque se busca representar algún parámetro de la población el cual suele ser representado por un valor numérico (promedio, proporción u otro)" así como de tipo aplicada, al aplicar la minería de datos para abordar la problemática de la violencia familiar y describir los resultados encontrados.

### Variables en estudio

1.- Número de consultas telefónicas hombres

Valores positivos > = 0

2.- Número de consultas telefónicas mujeres.

Valores positivos > = 0

3.- Número de consultas telefónicas sin sexo

Valores positivos > = 0

4.- Número de consultas telefónicas

Valores positivos > = 0

5.- Número de consultas telefónicas de violencia psicológica

Valores positivos > = 0

6.- Número de consultas telefónicas de violencia física

Valores positivos > = 0

7.- Número de consultas telefónicas de violencia sexual

Valores positivos > = 0

8.- Número de consultas telefónicas de violencia económica

Valores positivos > = 0

9.- Número de consultas de otro tipo

### 3.5 TÉCNICA E INSTRUMENTOS DE RECOLECCIÓN DE DATOS

**Tipo**: documental, mediante la base de datos estructurada línea 100 (ver anexo 02)

**Metodología utilizada**

CRISP-DM (Cross Industry Standard Process for Data Mining) según Rodríguez (2010).

El análisis de datos ha sido históricamente un procedimiento manual en muchas disciplinas. Con la ayuda de enfoques estadísticos, uno o varios analistas familiarizados con los datos presentaban resúmenes y preparaban informes. Sin embargo, este método evolucionó como consecuencia del aumento de la cantidad de datos, por lo que esta estrategia cambió (Riquelme et al., 2006).

La minería de datos es necesaria cuando la manipulación, el alcance del procesamiento, la exploración e inferencia de los datos se supera la capacidad que puede manejar el ser humano. Usualmente a través de softwares especializados y se considera el costo computacional de los algoritmos a partir de $10^3$ variables (Riquelme et al., 2006).

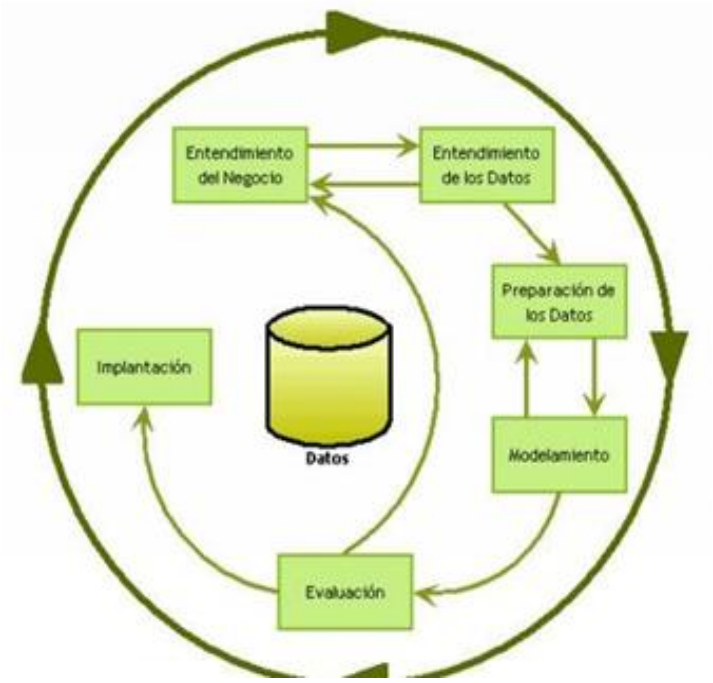



Figura 03.CRISP-DM

Fuente: Rodríguez (2010)

**Entendimiento del negocio**

Según Galán (2016). Esta primera fase es probablemente la más importante y aglutina las tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva de negocio, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto. Sin lograr comprender dichos objetivos, ningún algoritmo por muy sofisticado que sea, permitirá obtener resultados fiables. Para obtener el mejor provecho de la minería de datos, es necesario entender de la manera más completa el problema que se desea resolver, esto permitirá recolectar los datos correctos e interpretar correctamente los

resultados. En esta fase, es muy importante la capacidad de poder convertir el conocimiento adquirido del negocio en un problema de minería de datos y en un plan preliminar cuya meta sea el alcanzar los objetivos del negocio. A continuación, vemos una descripción de cada una de las principales tareas que componen esta fase.

**Comprensión de los datos**

Según Galán (2016) "esta segunda fase comprende la recolección inicial de los datos con el objetivo de establecer un primer contacto con el problema, familiarizarse con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis. Esta fase junto a las dos siguientes fases son las que demandan el mayor esfuerzo y tiempo en un proyecto de minería de datos. Por lo general si la organización cuenta con una base de datos corporativa, es deseable crear una nueva base de datos específica para el proyecto de DM (Data Mining), ya que durante el desarrollo del proyecto es posible que se generen frecuentes y abundantes accesos a la base de datos con el fin de realizar consultas y probablemente se produzcan modificaciones, lo cual podría generar muchos problemas. Vemos las tareas que componen esta fase".

**Preparación de los datos**

Según Galán (2016) "en esta fase y una vez efectuada la recolección inicial de los datos, se procede a su preparación para adaptarlos a las técnicas de minería de datos que se van a utilizar posteriormente, éstas pueden ser técnicas de visualización de datos, de búsqueda de relaciones entre variables u otras medidas para explotación de los datos. La preparación de los datos incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica".

**Modelado**

Según Galán (2016) en esta fase de CRISP-DM se seleccionan las técnicas de modelado más apropiadas para el proyecto de minería de datos específico. Las técnicas a utilizar en esta fase se eligen en función de los siguientes criterios:

- Ser apropiada para el problema.
- Disponer de los datos adecuados.
- Cumplir los requisitos del problema.
- Tiempo adecuado para obtener un modelo.
- Conocimiento de la técnica.

**Evaluación**

Según Galán (2016) en esta fase se evalúa el modelo, " teniendo en cuenta el cumplimiento de los criterios de éxito del problema. Debe considerarse además que la fiabilidad calculada para el modelo se aplica solamente para los datos sobre los que se realizó el análisis".

**Implantación**

El conocimiento obtenido debe organizarse y presentarse al usuario final de modo que pueda comprenderlo, puede ser tan simple como la generación de un informe o tan complejo como de automatizar todo un extenso proceso de análisis de datos mediante otras tecnologías en la organización.

# CAPÍTULO IV

## RESULTADOS Y DISCUSIÓN

**Objetivo específico 1.** Realizar la comprensión del caso de estudio y la información de los reportes atendidos por el programa nacional contra la violencia familiar

Para cumplir con los objetivos propuestos se utiliza la metodología CRISP-DM y para el cumplimiento de este objetivo específico implica las fases entendimiento del negocio y comprensión de los datos.

### 4.1 ENTENDIMIENTO DEL NEGOCIO

El Programa Nacional para la Prevención y Erradicación de la Violencia contra las Mujeres e Integrantes del Grupo Familiar (AURORA) es una institución que pertenece al Ministerio de la Mujer y Poblaciones Vulnerables creado con Resolución Ministerial Nº 088-2016-MIMP

### 4.2 COMPRENSIÓN DE LOS DATOS

**1.- Número de consultas telefónicas hombres**

**Tipo de variable:** Cuantitativo – discreto.

Valores positivos > = 0

**2.- Número de consultas telefónicas mujeres.**

**Tipo de variable:** Cuantitativo – discreto.

Valores positivos > = 0

**3.- Número de consultas telefónicas sin sexo**

**Tipo de variable:** Cuantitativo – discreto.

Valores positivos > = 0

**4.- Número de consultas telefónicas**

**Tipo de variable:** Cuantitativo – discreto

Valores positivos > = 0

**5.- Número de consultas telefónicas de violencia psicológica**

**Tipo de variable:** Cuantitativo – discreto

Valores positivos > = 0

**6.- Número de consultas telefónicas de violencia física**

**Tipo de variable:** Cuantitativo – discreto

Valores positivos > = 0

**7.- Número de consultas telefónicas de violencia sexual**

**Tipo de variable:** Cuantitativo – discreto

Valores positivos > = 0

**8.- Número de consultas telefónicas de violencia económica**

**Tipo de variable:** Cuantitativo – discreto

Valores positivos > = 0

**9.- Número de consultas de otro tipo**

**Tipo de variable:** Cuantitativo – discreto

Valores positivos > = 0

Observándose que todas las variables en estudio son de tipo cuantitativo discreto, por lo que requiere de métodos específicos para el análisis de estos datos.

Padilla (2019) propone que la obtención de patrones en cada uno de los procesos se realiza basado en características de los datos, así para generar las reglas de asociación es necesario una estructura de los datos.

Al ser datos de ciencias sociales de acuerdo a Rojas & Gomez (2014) la minera de datos es aplicada en una variedad de ciencias, puede ser utilizada en todas las actividades que realiza y procesa el ser humano que presenten la generación de datos que requieran ser analizados para proporcionar un conocimiento.

**Objetivo específico 2.** Realizar el análisis de minería de datos descriptiva de los reportes atendidos por el programa nacional contra la violencia familiar.

## 4.3 PREPARACIÓN DE LOS DATOS

En esta etapa de preparación de datos se verifica la cantidad de datos vacíos (faltantes) en cada variable, cuantificándose y presentándolo en el software R 4.2.X

Tabla 01.Porcentaje de datos vacíos por variable

| Variable | Porcentaje de vacíos |
|---|---|
| NCT Violencia sexual | 3.91% |
| Otras Consultas | 3.91% |
| NCT Hombres | 4.35% |
| NCT Violencia Psicológica | 7.39% |
| NCT Violencia Física | 18.70% |
| NCT Mujeres | 34.35% |
| NCT Total | 41.74% |
| NCT Violencia Económica | 66.09% |
| NCT Sin Sexo | 66.52% |

A partir de la tabla 01 se genera la siguiente figura

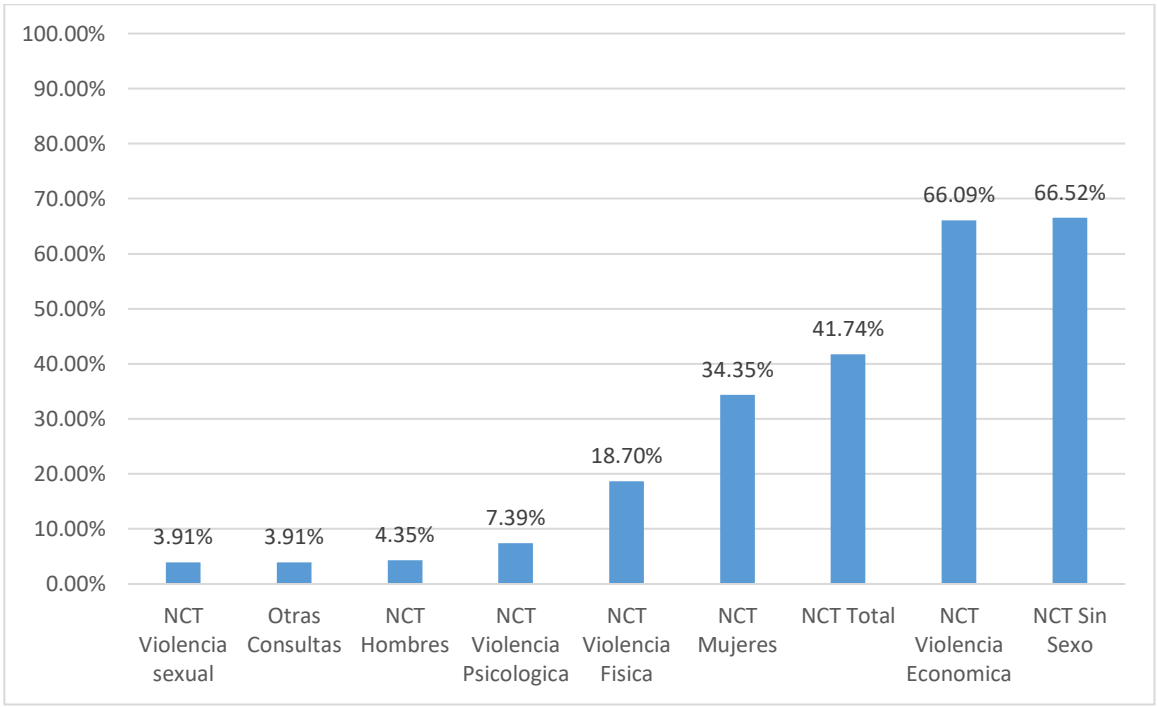



Figura 04.Porcentaje de datos vacíos por variable

Según la figura 04 el porcentaje de vacíos es de 66.52% en la variable Número de Consultas Totales Sin Sexo, 66.09% en el Número de Consultas Totales por Violencia Económica, 41.74% en el Número de Consultas Totales, 34.35% Número de Consultas Totales de Sexo Mujer, 18.70% de Número de Consultas Totales por Violencia Física, 7.39% Número de Consultas Total por violencia Psicológica, 4.35% Número de Consultas Total de Sexo Hombre, 3.91% Numero de Consulta Total por Otras consultas y el 3.91% Numero de Consulta Total por Violencia Sexual.

Imputando los datos vacíos mediante la librería MICE 3.13.0 usando el algoritmo 'CART'

Tabla 02.Porcentaje de datos vacíos luego de ser imputados

| Variable | Porcentaje de vacíos |
|---|---|
| NCT Violencia sexual | 0.00% |
| Otras Consultas | 0.00% |
| NCT Hombres | 0.00% |
| NCT Violencia Psicológica | 0.00% |
| NCT Violencia Física | 0.00% |
| NCT Mujeres | 0.00% |
| NCT Total | 0.00% |
| NCT Violencia Económica | 0.00% |
| NCT Sin Sexo | 0.00% |

Luego de ser imputados según la tabla 02 se aprecia que todas las variables contienen 0.00% de datos vacíos, por ello quedan listas para ser analizadas. Según Echeverry & Salazar (2017) estos procesos son necesarios porque, a pesar de todos los

controles para la gestión de la información, los usuarios incurren de modo permanente en errores de digitación que hacen difícil la coincidencia de los datos y la futura gestión del conocimiento.

Tabla 03.Resumen descriptivo

| Variable | n | Media | Desv.Est | Mediana | Mínimo | Máximo | Rango |
|---|---|---|---|---|---|---|---|
| NCT Total | 230 | 631.0 | 313.0 | 703.5 | 46.0 | 962.0 | 916.0 |
| NCT Hombres | 230 | 158.5 | 157.5 | 107.0 | 5.0 | 864.0 | 859.0 |
| NCT Mujeres | 230 | 587.4 | 329.3 | 554.5 | 38.0 | 975.0 | 937.0 |
| NCT Sin Sexo | 230 | 87.3 | 146.3 | 0.0 | 0.0 | 672.0 | 672.0 |
| NCT Violencia Psicológica | 230 | 299.0 | 241.5 | 211.5 | 10.0 | 964.0 | 954.0 |
| NCT Violencia Física | 230 | 447.5 | 305.9 | 381.5 | 17.0 | 990.0 | 973.0 |
| NCT Violencia Sexual | 230 | 88.1 | 97.5 | 55.5 | 3.0 | 548.0 | 545.0 |
| NCT Violencia Económica | 230 | 21.1 | 54.5 | 2.0 | 0.0 | 210.0 | 210.0 |
| Otras Consultas | 230 | 106.0 | 163.3 | 44.0 | 1.0 | 896.0 | 895.0 |

Según la tabla 03 en el resumen descriptivo para cada variable en la primera columna y sus respectivos valores donde observamos la cantidad de datos (n), el promedio aritmético por variable (Media), la desviación estándar (Desv.Est), la mediana (Mediana), el mínimo valor registrado (Mínimo), el máximo valor registrado (Máximo) y el Rango que es calculado por el máximo valor menos el mínimo valor.

Para la realización del análisis descriptivo se procede a usar el paquete ggpubr mediante la funcion gghistogram

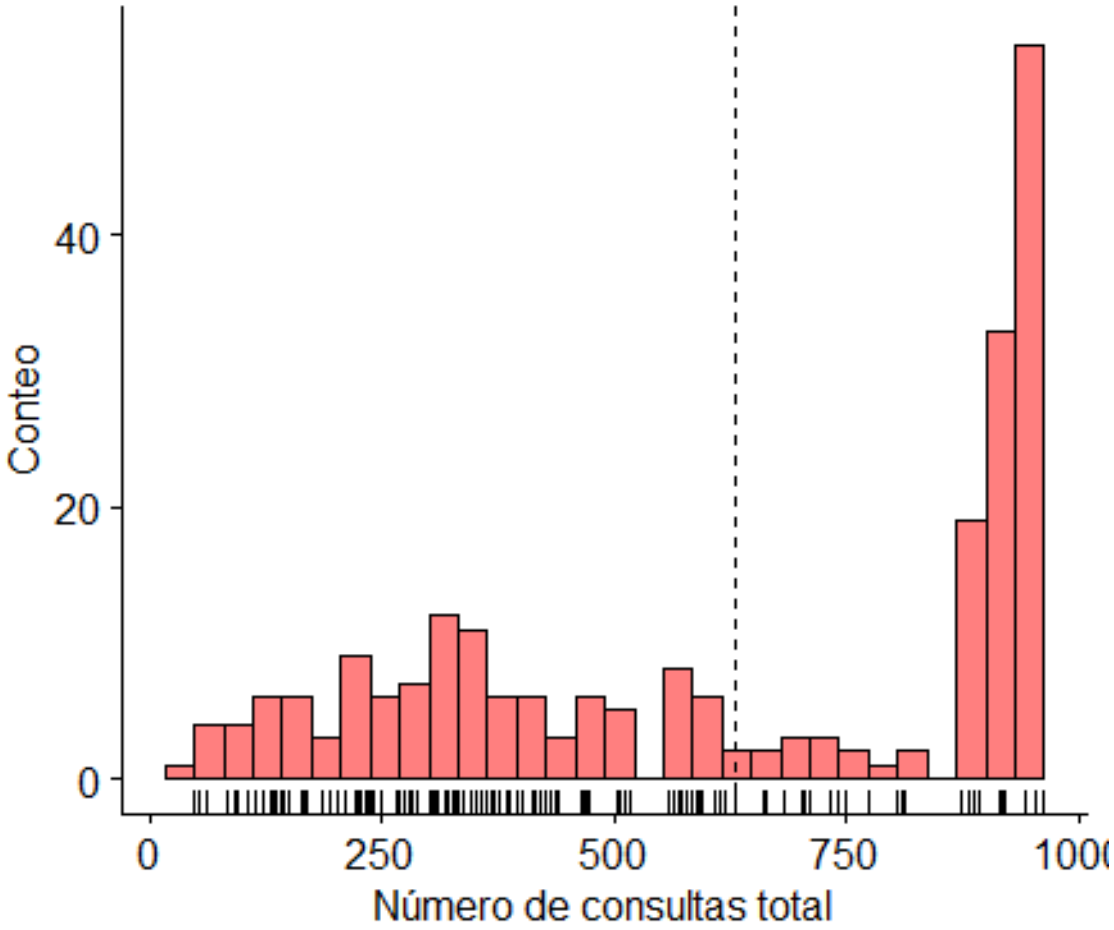



Figura 05. Histograma del Número de Consultas Total

En la figura 05 se observa el histograma del Número de Consultas Total en el que graficamente se observa una mayor concentración de datos alrededor de las 1000 consultas por reporte, y en la linea discontinua representando el promedio aritmetico de 631.0 consultas.

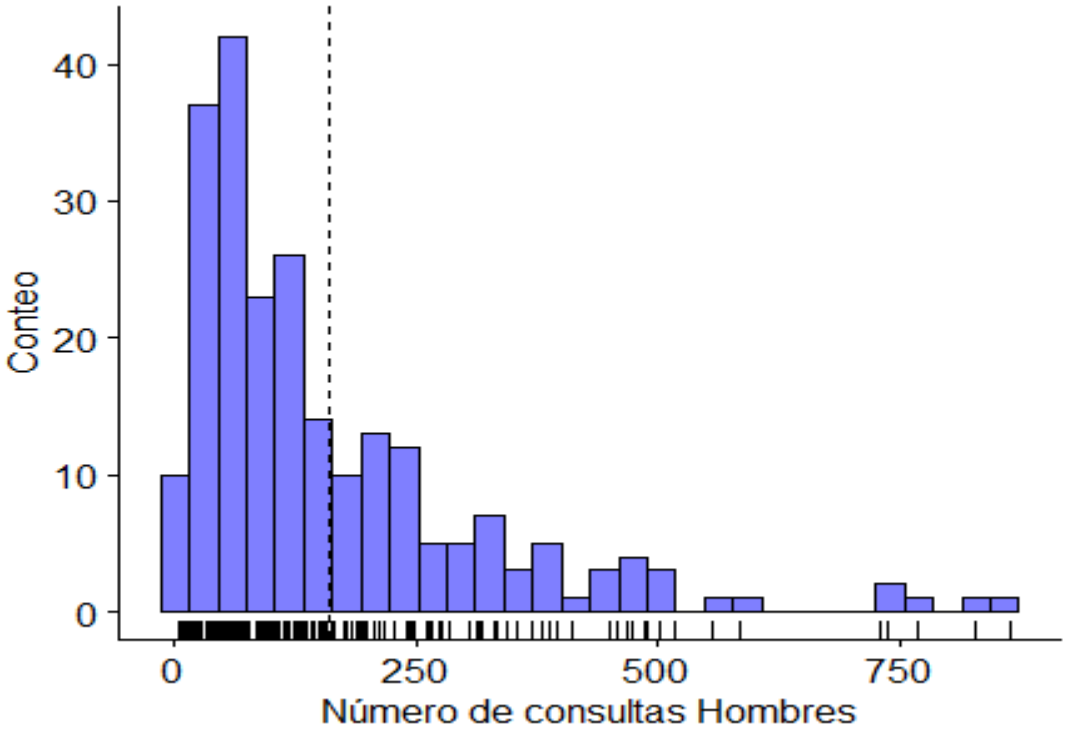



Figura 06. Histograma del Número de Consultas Hombres

En la figura 06 observamos el histograma del Número de Consultas Hombres en el que la mayoria de las barras oscilan alrededor del promedio aritmetico de 158.5 consultas por reporte.

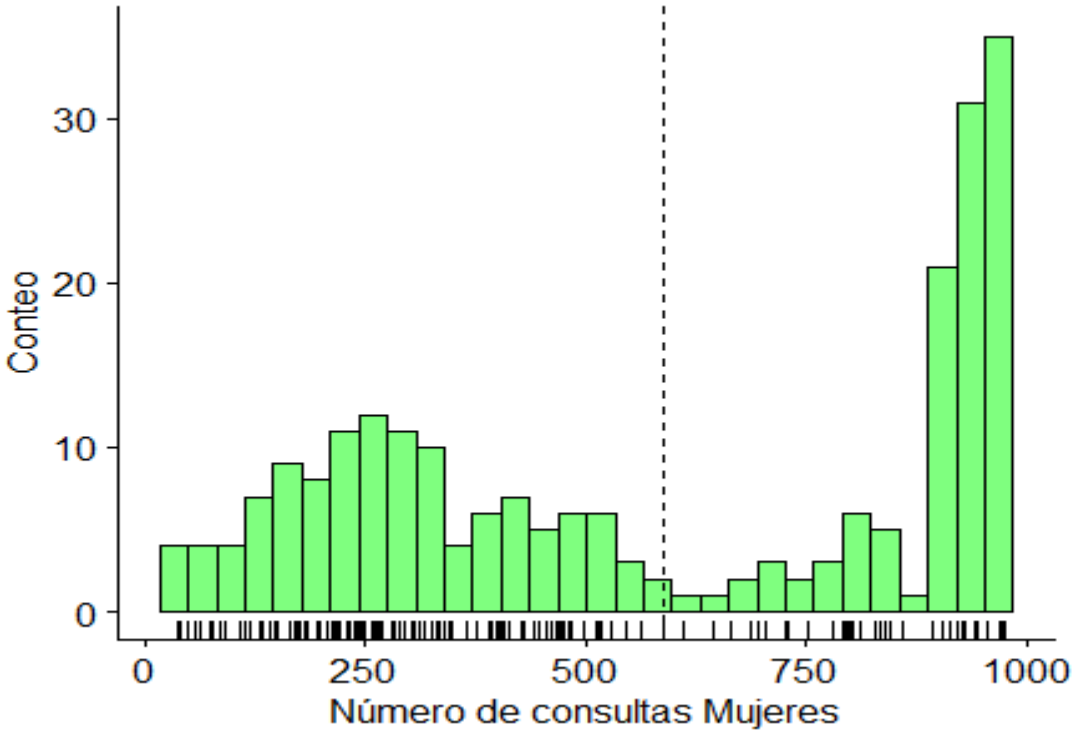



Figura 07. Histograma del Número de Consultas Mujeres

En la figura 07 observamos el histograma del Número de Consultas Mujeres donde se observa una alta concentración en Numero de Consultas de Sexo Mujer que se aproximan a las 1000 consultas por reporte , su promedio es de 587.4 consultas por reporte.

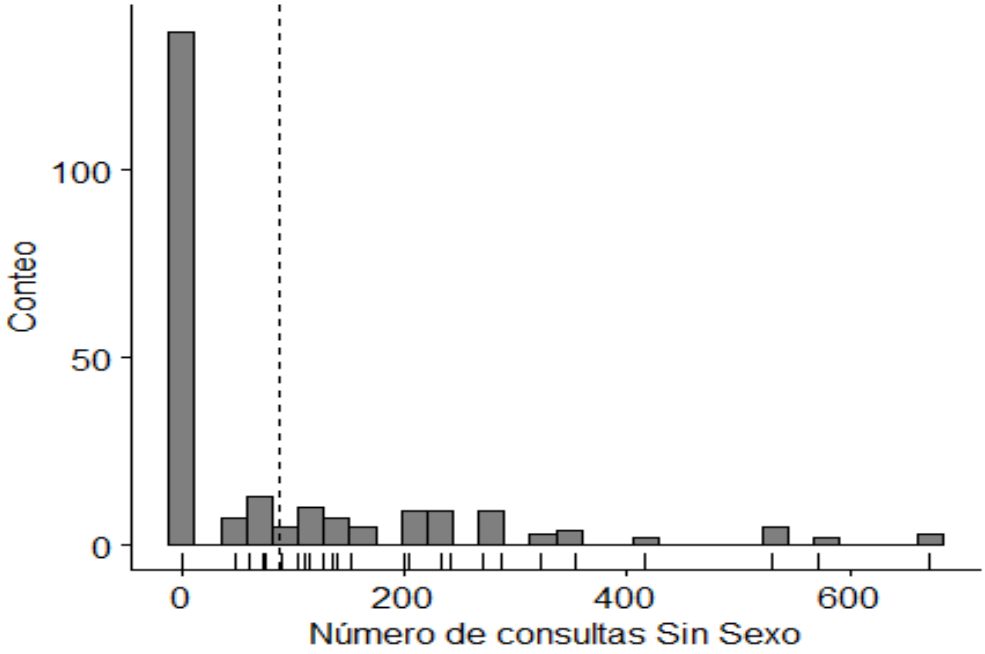



Figura 08.Histograma del Número de Consultas Sin Sexo

En la figura 08 se presenta el histograma del Número de Consultas Sin Sexo y se observa una alta concentración en Número de Consultas de no proporcionaron informacion sobre su Sexo de 0 por reporte consultas por reporte, en promedio 87.3 Consultas Sin Sexo por reporte.

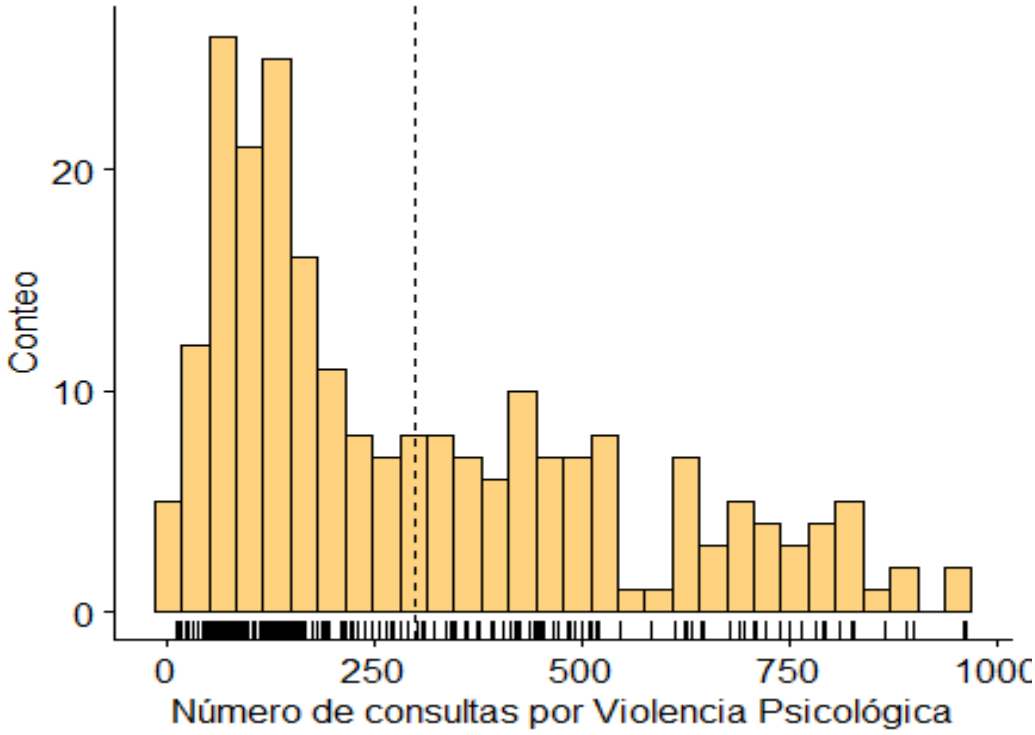



Figura 09. Histograma del Número de Consultas de Violencia Psicológica

En la figura 9 se presenta el histograma del Número de Consultas de Violencia Psicológica una mayor concentración inferior a la media de 299.0 consultas por reporte.

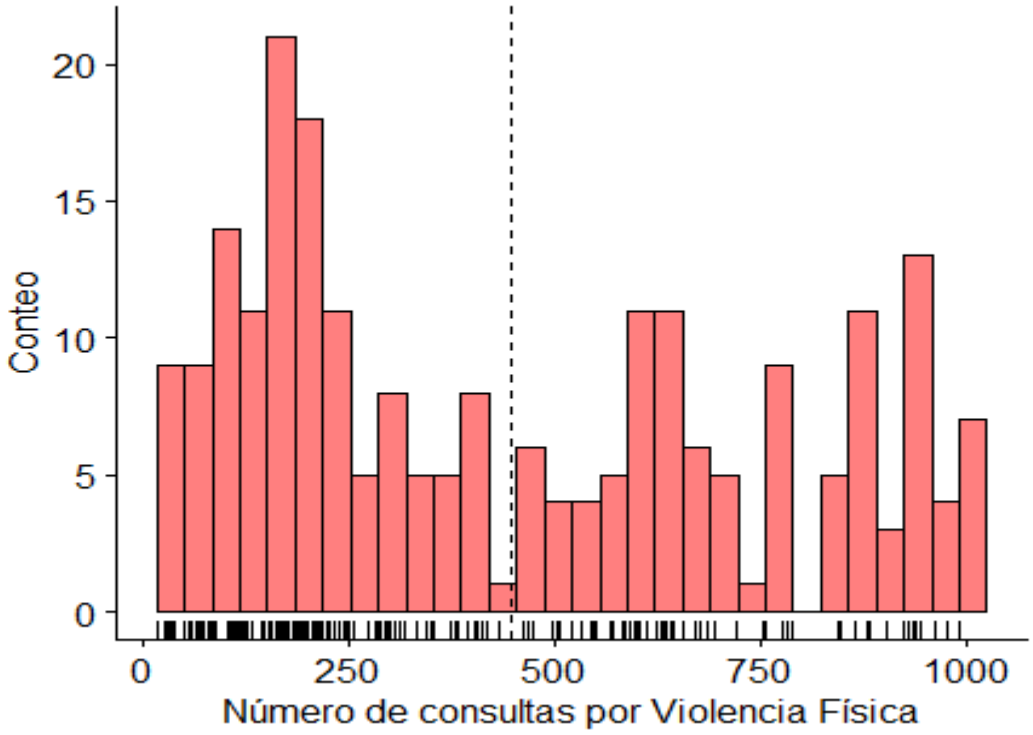



Figura 10. Histograma del Número de Consultas de Violencia Física

En la figura 10 histograma del Número de Consultas de Violencia Física una fuerte variación con respecto a su promedio 447.5 ya que las barras se encuentran distribuidas inferior y superiormente con respecto al promedio. Arenas (2015) esta violencia es la más clara y evidente del maltrato.

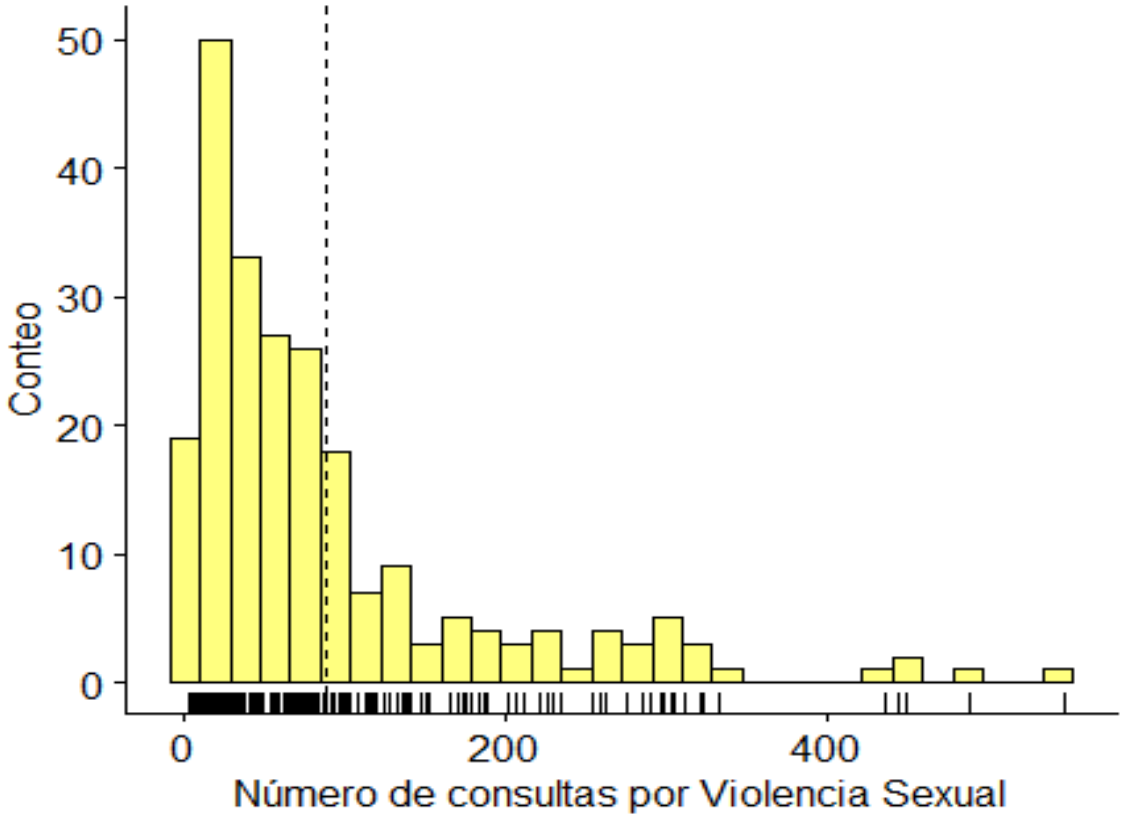



Figura 11. Histograma del Número de Consultas de Violencia Sexual

En la figura 11 se presenta el histograma del Número de Consultas de Violencia Sexual que presenta una media de 88.1 consultas por reporte y la mayoría de los registros son menores a 400 consultas de Violencia Sexual por reporte.

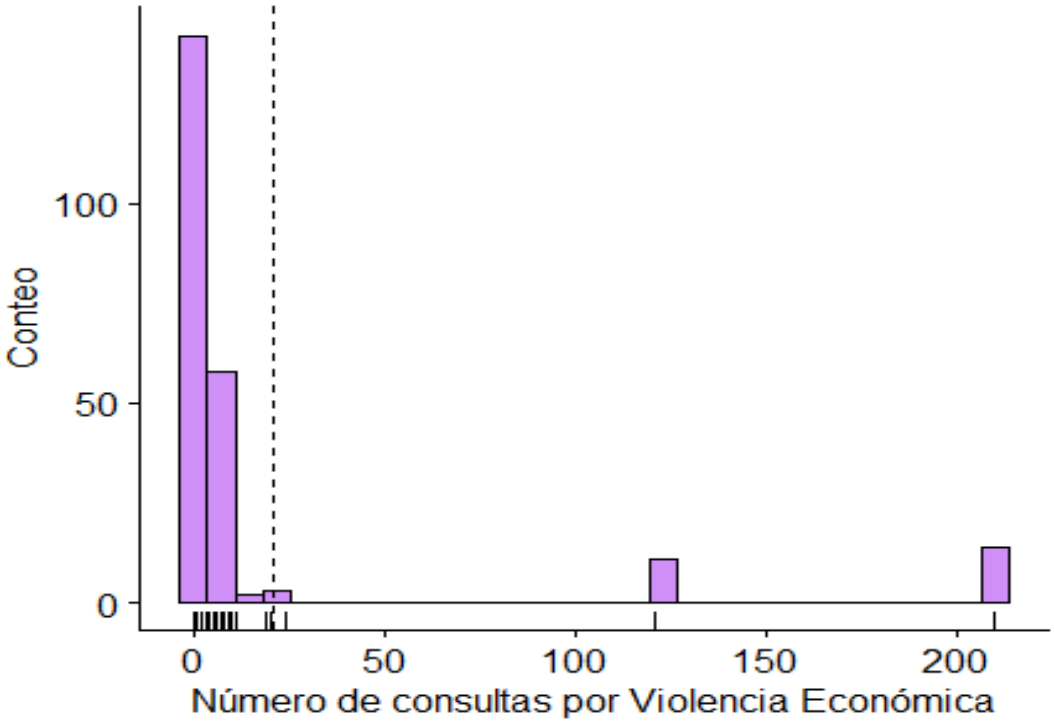



Figura 12. Histograma del Número de Consultas de Violencia Económica

En la figura 12 se presenta el histograma del Número de Consultas de Violencia Económica siendo el que menos Consultas tiene  siendo el promedio de 21.1 consultas por reporte, asimismo se observa una acumulación cercana a las 0 consultas por Violencia Económica en los reportes.

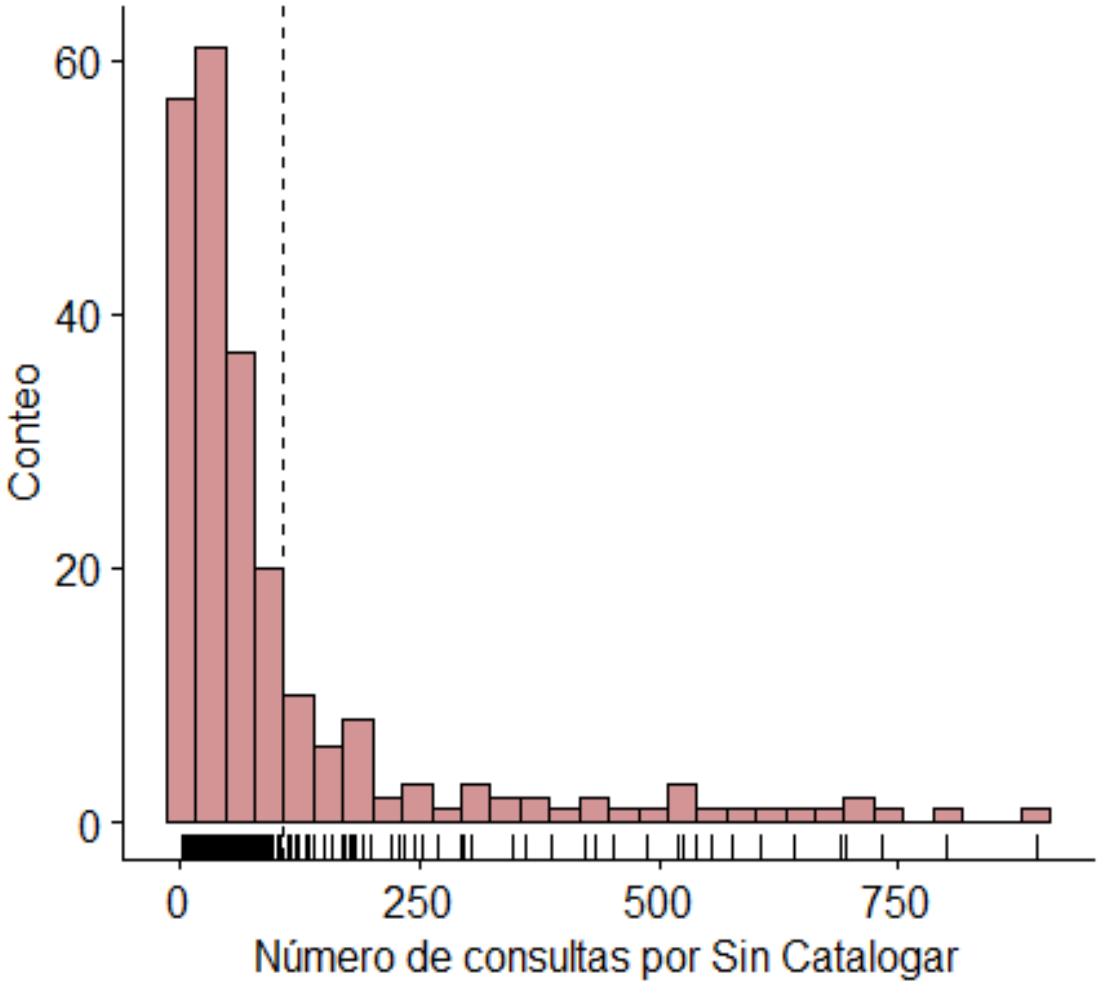



Figura 13. Histograma del Número de Consultas de Otro Tipo

En la figura 13 se presenta el histograma del Número de Consultas de Otro Tipo observándose que se acumulan este tipo de consultas alrededor de su promedio aritmético de 106.0 por reporte.

Para el análisis descriptivo y en busca de encontrar asociaciones para generar grupos se utiliza el coeficiente de correlación

De Pearson usando el paquete corrplot.

Tabla 04.Matriz de correlación entre las variables

| Variable | NCT Total | NCT Hombres | NCT Mujeres | NCT Sin Sexo | NCT Violencia Psicológica | NCT Violencia Física | NCT Violencia Sexual | NCT Violencia Económica | Otras Consultas |
|---|---|---|---|---|---|---|---|---|---|
| NCT Total | 1.00 | 0.67 | 0.98 | 0.47 | 0.81 | 0.91 | 0.63 | 0.24 | 0.44 |
| NCT Hombres | 0.67 | 1.00 | 0.68 | 0.32 | 0.73 | 0.72 | 0.93 | -0.05 | 0.89 |
| NCT Mujeres | 0.98 | 0.68 | 1.00 | 0.49 | 0.86 | 0.94 | 0.65 | 0.24 | 0.46 |
| NCT Sin Sexo | 0.47 | 0.32 | 0.49 | 1.00 | 0.54 | 0.65 | 0.33 | -0.02 | 0.16 |
| NCT Violencia Psicológica | 0.81 | 0.73 | 0.86 | 0.54 | 1.00 | 0.89 | 0.70 | 0.12 | 0.53 |
| NCT Violencia Física | 0.91 | 0.72 | 0.94 | 0.65 | 0.89 | 1.00 | 0.70 | 0.12 | 0.50 |
| NCT Violencia Sexual | 0.63 | 0.93 | 0.65 | 0.33 | 0.70 | 0.70 | 1.00 | -0.06 | 0.84 |
| NCT Violencia Económica | 0.24 | -0.05 | 0.24 | -0.02 | 0.12 | 0.12 | -0.06 | 1.00 | -0.11 |
| Otras Consultas | 0.44 | 0.89 | 0.46 | 0.16 | 0.53 | 0.50 | 0.84 | -0.11 | 1.00 |

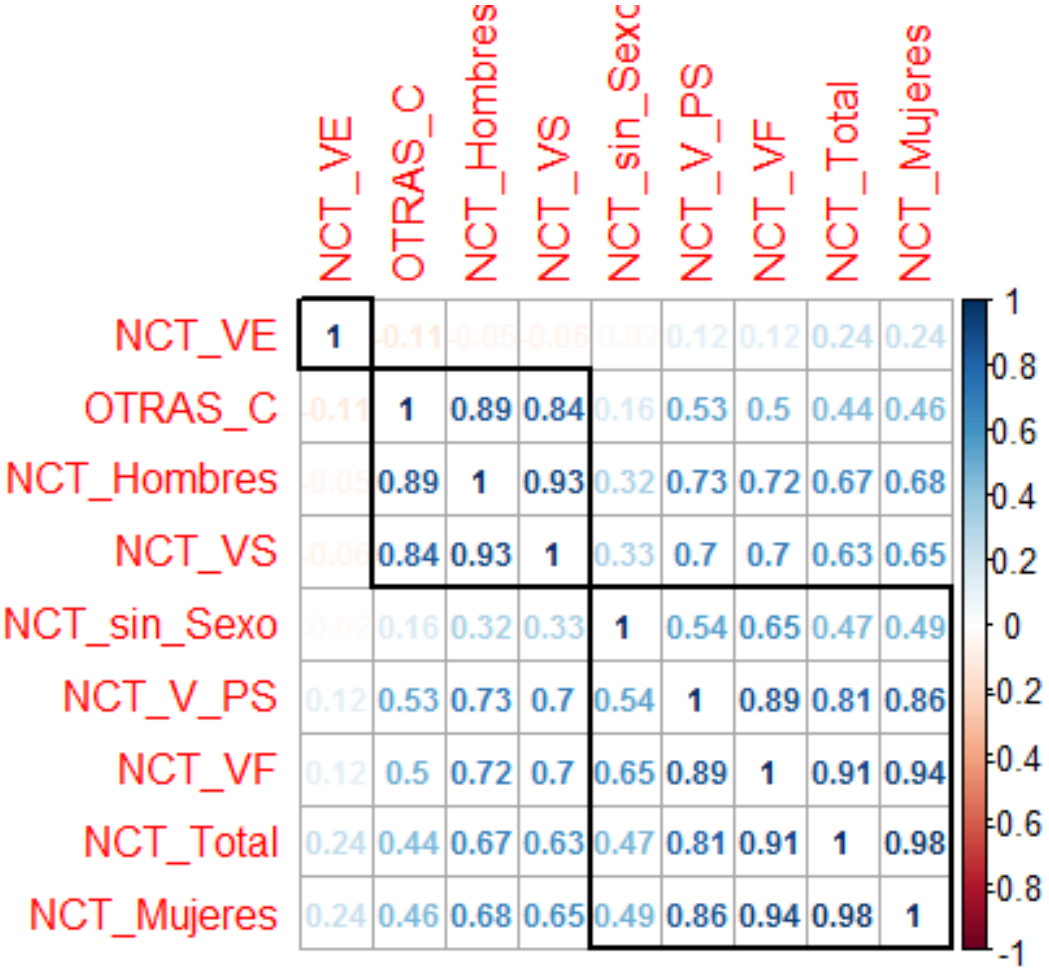



Figura 14. Matriz de correlaciones

En la figura 14 y tabla 04 observamos una alta correlación entre diversas variables que aparentemente estarían creando 2 clústeres encasillados en el que la correlación es mayor a 0.4.

## 4.4 MODELADO

Para la creación de los Clústeres se utiliza el algoritmo de "K means" para los Clústeres individuales y el Análisis de Componentes Principales para realizar los Clústers por variables.

### K medias

Clústers con diversos valores de "K" para observar las posibles agrupaciones y su comportamiento.

Usando los paquetes factoextra 1.0.7 y ggpubr 0.2.5

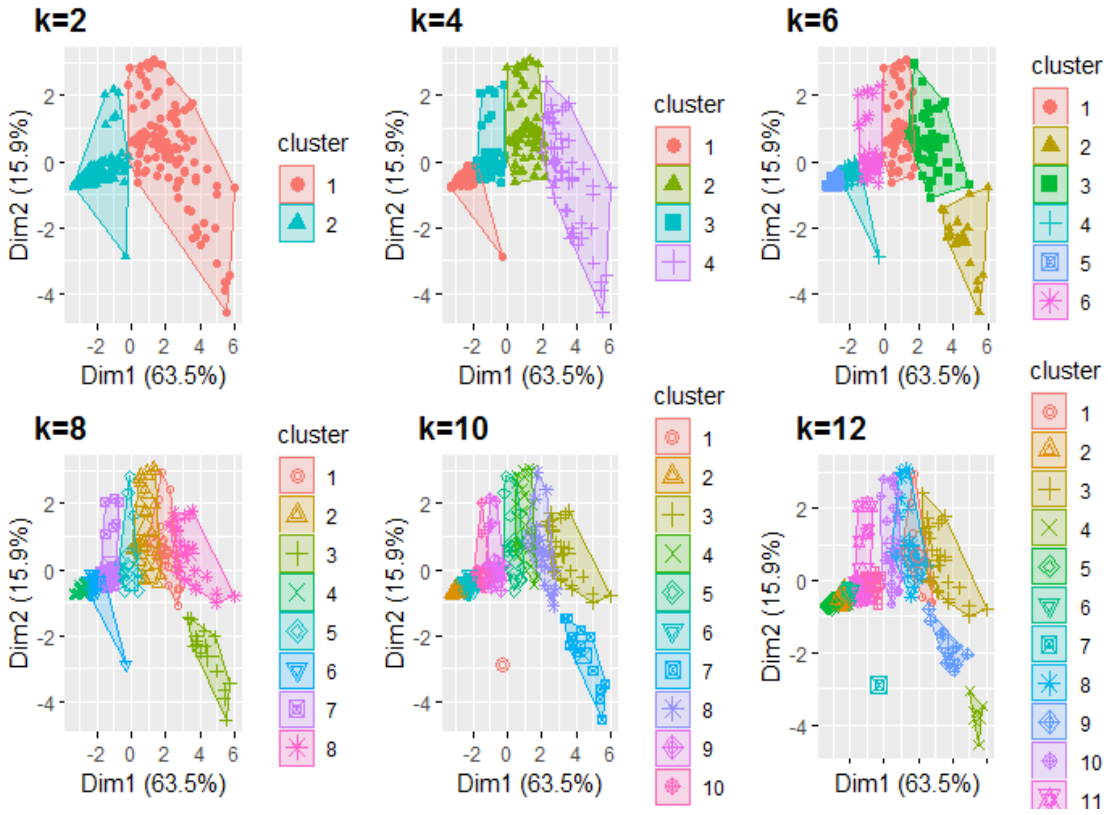



Figura 15.Clústeres con diversos valores de "K"

En la figura 15 se observa el comportamiento de las diferentes agrupaciones con valores de k desde 2 hasta 12 grupos o Clústeres.

**Elegir el número óptimo de k según el enfoque de average Silhouette**

Que mide la calidad de agrupación según Kaufman & Rousseeuw (2009) determina qué tan bien se encuentra cada objeto dentro de su grupo.

Utilizando el paquete factoextra 1.0.7

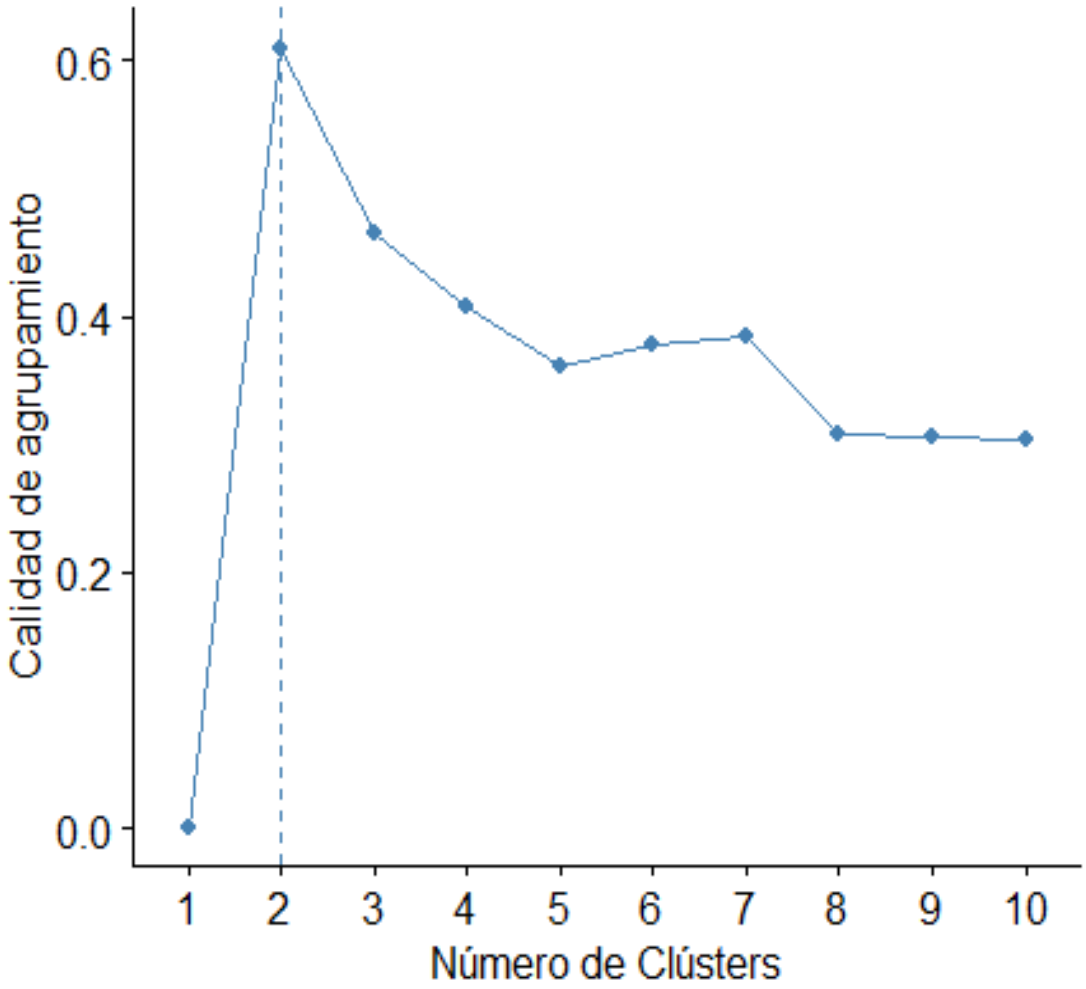



Figura 16. Calidad de agrupación según el enfoque de average Silhouette

En la figura 16 según este criterio de average Silhouette el número optimo de clusters (k) es 2 teniendo una calidad del 60% de agrupación.

Kaufman & Rousseeuw (2009) el enfoque de average Silhouette mide la calidad de un agrupamiento. En otras palabras, determina qué tan bien se encuentra cada objeto dentro de su grupo.

**Análisis de componentes principales**

Usando el paquete factominer 1.34 y factoextra 1.0.7

Tabla 05.Porcentaje de varianza explicando por Clústeres/Componentes

| Componente | Porcentaje de varianza | Porcentaje de varianza acumulada |
|---|---|---|
| comp 1 | 63.52% | 63.52% |
| comp 2 | 15.85% | 79.37% |
| comp 3 | 10.68% | 90.05% |
| comp 4 | 5.23% | 95.28% |
| comp 5 | 2.04% | 97.32% |
| comp 6 | 1.39% | 98.71% |
| comp 7 | 0.62% | 99.34% |
| comp 8 | 0.53% | 99.86% |
| comp 9 | 0.14% | 100.00% |

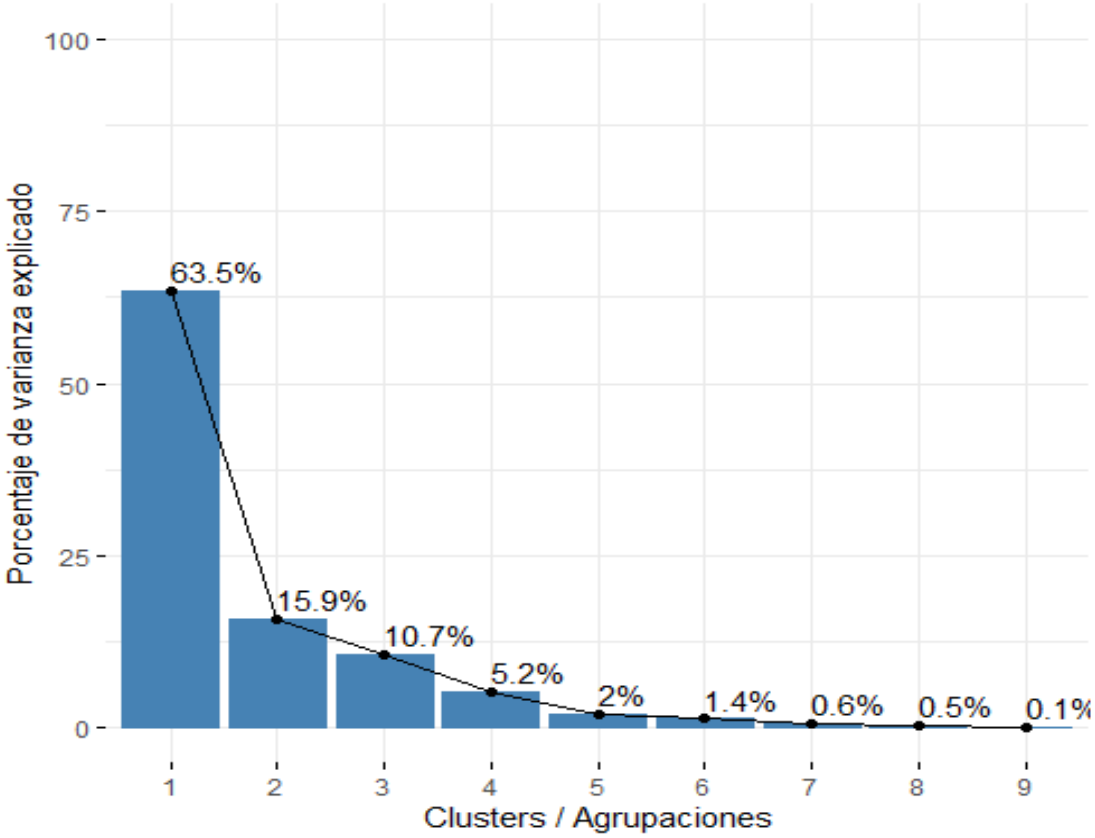



Figura 17.Porcentaje de varianza explicado por Clústeres

Según la figura 17 y tabla 05 por lo expuesto por Plaza & Cardozo (2018) la ubicación de una curva (codo) en la parcela/Clúster generalmente se considera como un indicador del número apropiado de conglomerados, y según el porcentaje de varianza explicado ya que utilizando 2 dimensiones (Clústers) se obtiene un porcentaje de varianza acumulado de 79.37% de representación.

Utilizando el criterio de % de varianza acumulada y el criterio de average Silhouette procede a realizar 2 agrupaciones.

## 4.5 EVALUACIÓN

Wss significa la suma total de cuadrados de las distancias entre los puntos y los centroides wss correspondientes para los k Clústers

Una medida es la suma de cuadrados dentro del clúster (WSS), que mide la distancia promedio al cuadrado de todos los puntos dentro de un clúster al centroide del clúster. Para calcular WSS, primero encuentra la distancia euclidiana (ver la figura a continuación) entre un punto dado y el centroide al que está asignado. Luego itera este proceso para todos los puntos en el grupo y luego suma los valores para el grupo y divide por el número de puntos. Finalmente, calcula el promedio en todos los grupos. Esto le dará el WSS promedio (Plaza & Cardozo, 2018).

Tabla 06. Distancias entre los puntos y los centroides correspondientes para los k Clústeres

| k2 | k4 | k6 | k8 | k10 |
|---|---|---|---|---|
| 29971294.41 | 17377601.25 | 11301514.06 | 10270356.59 | 7817624.459 |

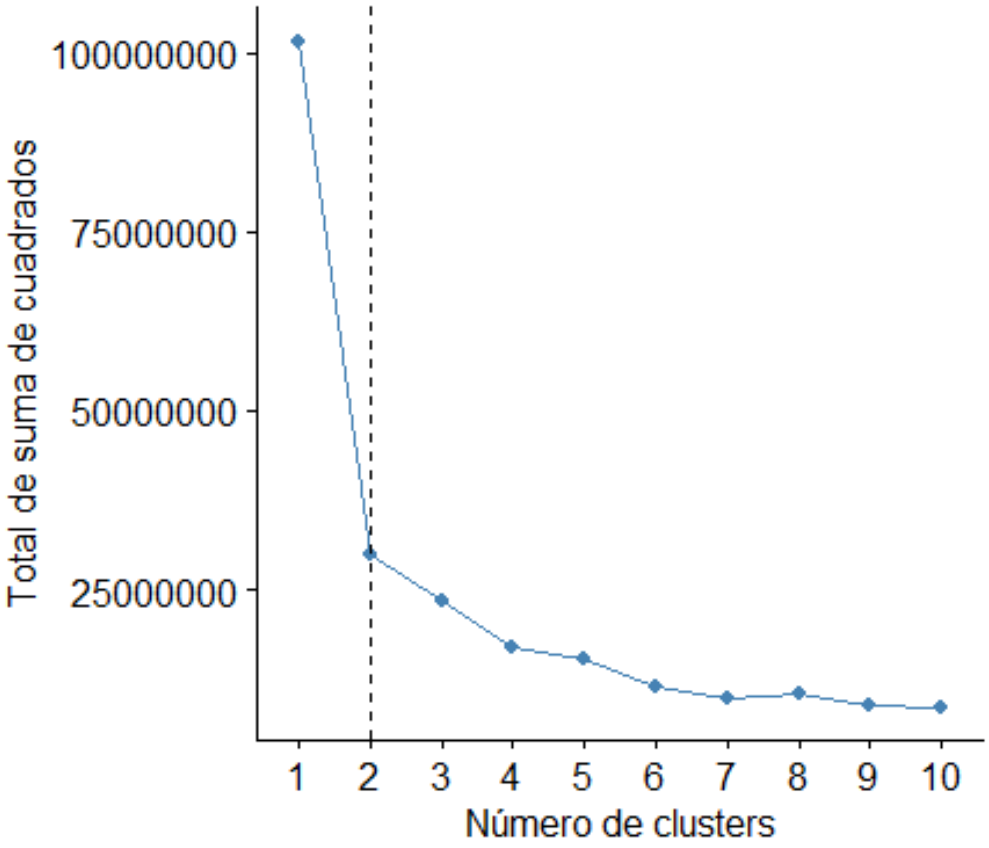



Figura 18.Suma total de cuadrados

De acuerdo a la figura 18 y tabla 06 la suma total de cuadrados utilizando 2 Clúster es de 29971294 de distancia entre todos los puntos a sus centroides, a pesar de ser un número elevado, estos resultados son apoyados por el análisis de componentes principales ya que asegura el 79.37% de representación.

## 4.6 IMPLANTACIÓN

Se aplica la Cauterización usando 2 agrupaciones.

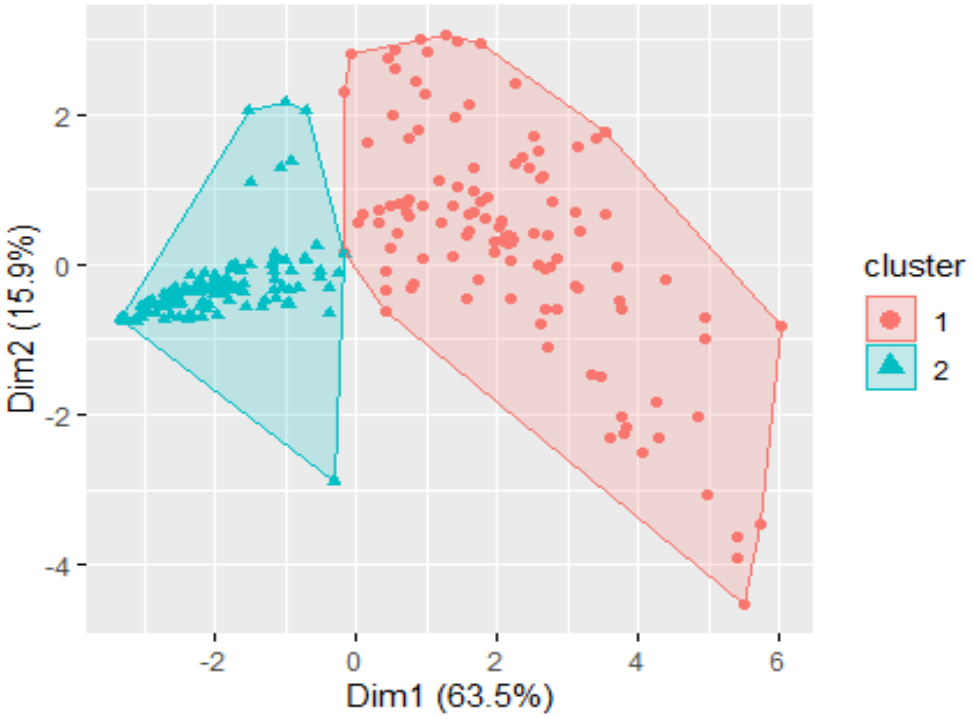



Figura 19.Clúster generados con k = 2

Tabla 07.Clúster generados con k = 2

| K means | N | % |
|---|---|---|
| Clúster 1 | 110 | 47.83% |
| Clúster 2 | 120 | 52.17% |
| Total | 230 | 100.00% |

Como se observa en la figura 19 y tabla 07 se aprecia una clara diferencia entre los Clúster 1 y 2 donde el 47.83% de los reportes pertenecen al Clúster 1, el 52.17% pertenecen al Clúster 2 (Ver anexo 5).

Tabla 08.Centroides del algoritmo k-means por variable y Clúster

| Clúster | NCT Total | NCT Hombres | NCT_Mujeres | NCT Sin Sexo | NCT Violencia Psico. | NCT Violencia Física | NCT Violencia Sexual | NCT Violencia Económica | Otras Consultas |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 926.481818 | 265.8636364 | 908.4454545 | 159.1181818 | 505.2636364 | 730.7545455 | 151.0818182 | 33.98181818 | 181.2818182 |
| 2 | 360.05 | 60.03333333 | 293.1166667 | 21.45833333 | 109.9916667 | 187.8833333 | 30.35 | 9.308333333 | 36.91666667 |

En la tabla 08 se observa los centroides calculados para cada variable y en cada Clúster

Determinando que variables están asociadas a los Clústeres mediante análisis de componentes principales y la calidad de representación

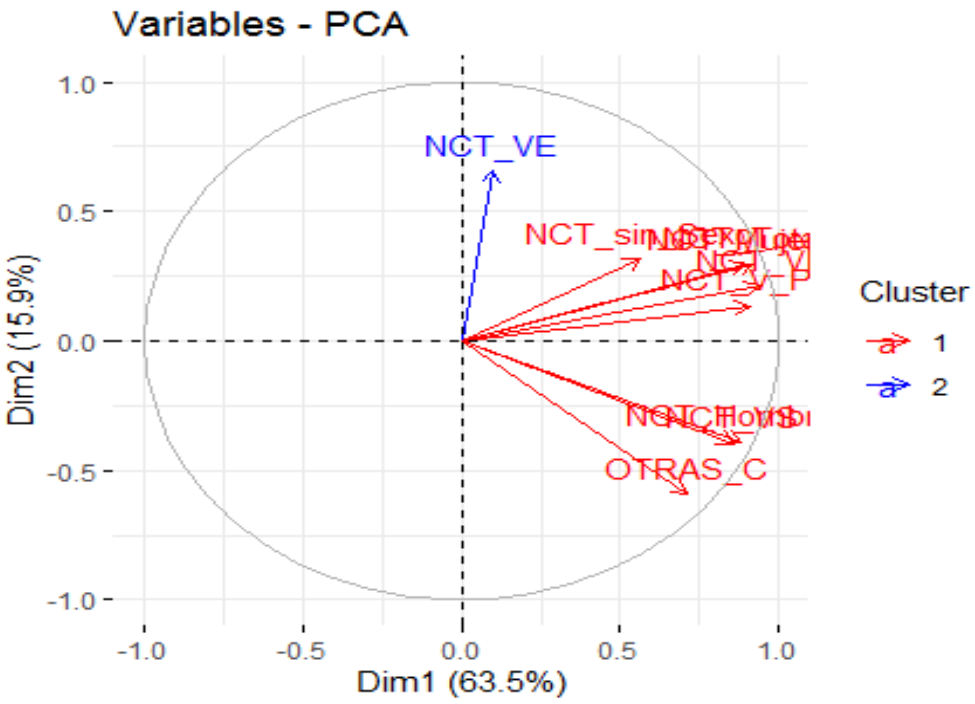



Figura 20. Clústeres y Componentes Principales

Según la figura 20 el círculo de correlación usando el de 79.37% de representación describe que todas las variables pertenecen al Clúster 1; excepto la variable de Número Total de Consultas Por Violencia Económica.

Abdi & Williams (2010) definen que las características del círculo de correlaciones son

- Las variables correlacionadas positivamente se agrupan.
- Las variables correlacionadas negativamente se colocan en lados opuestos del origen del gráfico (cuadrantes opuestos).
- La distancia entre las variables y el origen mide la calidad de las variables en el mapa de factores. Las variables que están alejadas del origen están bien representadas en el mapa de factores.

La correlación entre una variable y un componente principal (CP) se utiliza como las coordenadas de la variable en el CP. La representación de las variables difiere del gráfico de las observaciones: las observaciones están representadas por sus proyecciones, pero las variables están representadas por sus correlaciones (Abdi & Williams, 2010).

De ello se deduce que las flechas observadas de color rojo están relacionadas entre sí y que la variable número de consultas total de violencia económica no se asocia con las variables por ello se sitúa cercano al cuadrante horizontal superior.

**Contribución de las variables a los Clústeres mediante Análisis de Componentes Principales**

Tabla 09.Contribución de las variables a los Clústeres de ACP

| Variable | Clúster 1 | | Clúster 2 | |
|---|---|---|---|---|
| | % | Pertenece | % | Pertenece |
| NCT_Total | 14.02% | Si | **6.23%** | **no** |
| NCT_Hombres | 13.68% | Si | 10.64% | si |
| NCT_Mujeres | 14.76% | Si | **6.05%** | **no** |
| **NCT_sin_Sexo** | 5.49% | **No** | **6.91%** | **no** |
| NCT_V_PS | 14.50% | Si | **1.30%** | **no** |
| NCT_VF | 15.60% | Si | **3.24%** | **No** |
| NCT_VS | 12.91% | Si | 10.96% | Si |
| **NCT_VE** | 0.17% | **No** | 30.19% | Si |
| **OTRAS_C** | 8.86% | **No** | 24.49% | Si |
| Total de representación | 85.47% | - | 76.28% | - |

En la tabla 09 se presentan las contribuciones de las variables en estudio tanto para clúster 1 y para el clúster 2, se considera que una variable pertenece a un clúster cuando supera el 10% de contribución, valor planteado por el investigador.

Las variables que no se correlacionan con ningún Componente Principal o Clúster o de baja contribución podrían eliminarse para simplificar el análisis general.

Según Pérez (2008) el análisis de componentes principales busca descubrir variables latentes no observables directamente. Tiene como objetivo descubrir la estructura subyacente de un conjunto de datos cuantitativos definiendo un pequeño número de agrupaciones comunes que expliquen la mayor parte de la varianza observada.

Aquellas variables que pertenecen a un componente o clúster, según la teoría del análisis de componentes principales, esta agrupación se realiza entre variables que están correlacionadas altamente, en tal sentido según el objetivo general a continuación.

**Cumpliendo el objetivo general**

Aplicar minería de datos para explorar información que nos permita encontrar qué relación tienen los reportes atendidos por el Programa Nacional para la Prevención y Erradicación de la Violencia contra las Mujeres e Integrantes del Grupo Familiar.

Para ello utilizaremos el estadístico de medida de adecuación muestral de Kaiser-Mayer-Olikin (KMO)

Pérez & Medrano (2010) argumenta que “La lógica del índice KMO es que, si las variables comparten factores comunes, los coeficientes de correlación parcial deben ser pequeños y por ende los valores de la diagonal de la matriz deben ser elevados, es decir, si es elevada la proporción de coeficientes grandes en la matriz existe mayor interrelación entre las variables). El KMO se interpreta de manera semejante a los coeficientes de

confiabilidad, vale decir, con un rango de 0 a 1 y considerando como adecuado un valor igual o superior a 0.50, el cual sugiere una interrelación satisfactoria entre los ítems".

En otras palabras, esto quiere decir que si el coeficiente KMO supera el valor 0.5 según este autor existe una interrelación de las variables de los reportes del programa nacional contra la violencia familiar.

El test de esfericidad de Bartlett contrasta la hipótesis nula que supone que la matriz de correlaciones es una matriz identidad (1 en la diagonal principal el resto 0), de darse esta situación no existirían correlación significativa entre los reportes atendidos por el Programa Nacional para la Prevención y Erradicación de la Violencia contra las Mujeres e Integrantes del Grupo Familiar.

Tabla 10.Valores del índice KMO

| Variables | KMO |
|---|---|
| NCT_Total | 0.803725375 |
| NCT_Hombres | 0.837410751 |
| NCT_Mujeres | 0.762105261 |
| NCT_sin_Sexo | 0.744733751 |
| NCT_V_PS | 0.91809141 |
| NCT_VF | 0.855311502 |
| NCT_VS | 0.915538687 |
| NCT_VE | 0.739874614 |
| OTRAS_C | 0.845302296 |
| Total | 0.8372681 |

En la tabla 10 se observa que los índices KMO en total es de 0.8372681 que supera el 0.5 según Pérez & Medrano (2010) existe interrelación satisfactoria entre los ítems.

Test de esfericidad de Bartlett cuya hipótesis estadística es:

$H_0$: la matriz de correlaciones es igual a la matriz identidad, cuyo caso no existe relación entre los reportes atendidos por el Programa Nacional para la Prevención y Erradicación de la Violencia contra las Mujeres e Integrantes del Grupo Familiar.

$H_1$: la matriz de correlaciones es diferente a la matriz identidad, en tal medida existe relación entre los reportes atendidos por el Programa Nacional para la Prevención y Erradicación de la Violencia contra las Mujeres e Integrantes del Grupo Familiar.

Utilizando el paquete psych 2.2.9

Cuyo valor p es igual a 0 lo que implica que al nivel de confianza del 95% se acepta $H_1$: la matriz de correlaciones es diferente a la matriz identidad, existen correlaciones significativas de los reportes atendidos por el Programa Nacional para la Prevención y Erradicación de la Violencia contra las Mujeres e Integrantes del Grupo Familiar.

Lo que implica relación entre las variables en estudio.

De la tabla 09 resume la contribución de las variables para el clúster 1 en la siguiente figura

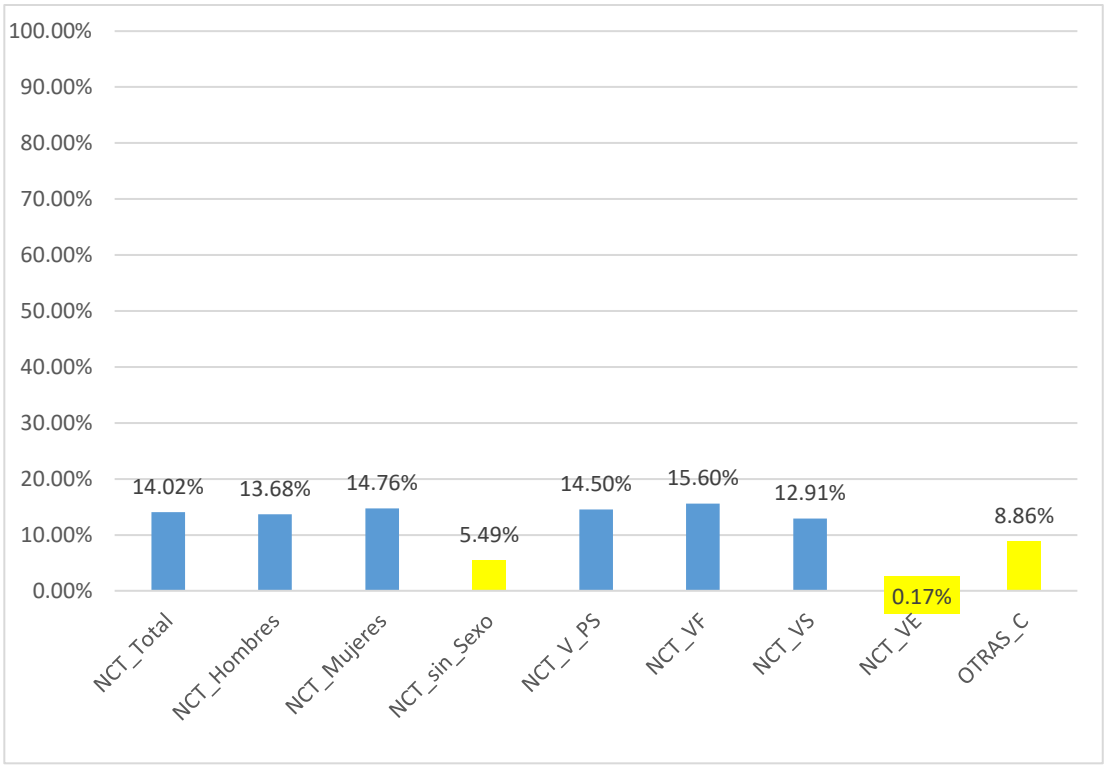



Figura 21. Contribución de las variables para el clúster 1

En la figura 21 se observa que de color azul son las variables que contribuyen al clúster 1 que superan el 10% de contribución y de color amarillo aquellas que no superan el 10% de contribución, las variables que contribuyen al clúster 1 son:

- Número de Consultas Telefónicas Total
- Número de Consultas Telefónicas Hombres
- Número de Consultas Telefónicas Mujeres
- Número de Consultas Telefónicas Por Violencia Psicológica
- Número de Consultas Telefónicas Por Violencia Física
- Número de Consultas Telefónicas Por Violencia Sexual

De la tabla 09 resume la contribución para el clúster 2 en la siguiente figura

Contribución de las variables al clúster 2

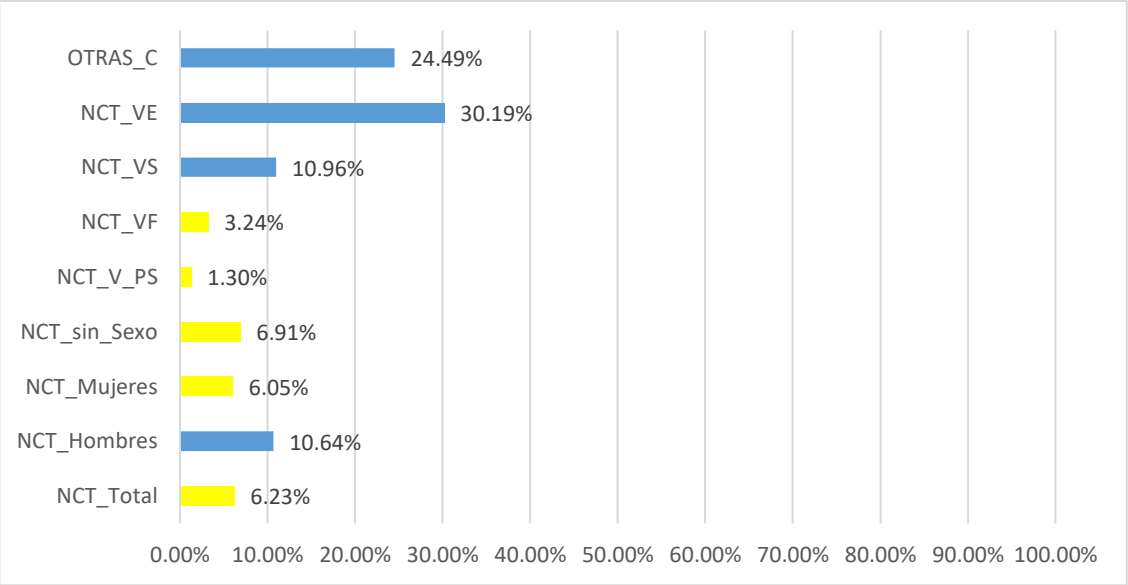



Figura 22. Contribución de las variables para el clúster 2

En la figura 22 se observa que de color amarillo son las variables que no contribuyen al clúster 2 y no superan el 10%, de color azul aquellas que superan el 10% de contribución, las variables que contribuyen al clúster 2 son:

Contribución de las variables al clúster 2

- Número de Consultas Telefónicas Por Violencia Sexual
- Número de Consultas Telefónicas Por Violencia Económica
- Número de Consultas Telefónicas Por Otras Consultas
- Número de Consultas Telefónicas Hombres

Estas variables estadisticamente alttamente correlacionadas entre si,

En la tabla 09 y figura 21 observamos que en el Clúster 1 pertenecen las siguientes variables:

- Número de Consultas Telefónicas Total
- Número de Consultas Telefónicas Hombres
- Número de Consultas Telefónicas Mujeres

- Número de Consultas Telefónicas Por Violencia Psicológica
- Número de Consultas Telefónicas Por Violencia Física
- Número de Consultas Telefónicas Por Violencia Sexual

Que conjuntamente acumulan un 85.47% de representación en el Clúster 1 denominado como “Clúster en el que no incluye la violencia económica”.

En la tabla 09 y figura 22 observamos que en el Clúster 2 pertenecen las siguientes variables

- Número de Consultas Telefónicas Por Violencia Sexual
- Número de Consultas Telefónicas Por Violencia Económica
- Número de Consultas Telefónicas Por Otras Consultas
- Número de Consultas Telefónicas Hombres

En conjunto acumulan un 76.28% de representación en el Clúster 2 denominado “Factor en el que la violencia económica es muy importante”.

# V. CONCLUSIONES

**PRIMERA**

Se logró desarrollar la minería de datos utilizando técnicas de Clusterización para los grupos de individuos mediante el algoritmo de "K means" con 2 Clústers integrados por el 47.83% de los reportes en el primero y 52.17% de los reportes en el segundo; con una suma total de cuadrados de 29971294 de distancia entre todos los puntos a sus centroides,, para agrupar las variables el Análisis de Componentes Principales agrupando las variables en 2 Clústers o Componentes Principales, el primero con 85.47% de representación denominado "Clúster en el que no incluye la violencia económica", el segundo con un 76.28% de representación denominado "Clúster en el que la violencia económica es muy importante". Siendo el Test de esfericidad de Bartlett significativo lo que implica que, la matriz de correlaciones es diferente a la matriz identidad, en tal medida existe relación entre los reportes atendidos por el Programa Nacional para la Prevención y Erradicación de la Violencia contra las Mujeres e Integrantes del Grupo Familiar confirmado por el índice KMO de 0.837 que se interpreta que existe interrelación entre los ítems analizados.

**SEGUNDA**

El Programa Nacional para la Prevención y Erradicación de la Violencia contra las Mujeres e Integrantes del Grupo Familiar (AURORA) es una institución que pertenece al Ministerio de la Mujer y Poblaciones Vulnerables que en el que el promedio de Consultas Hombres es de 159 por reporte, 587 en promedio de Llamadas Telefónicas Mujeres por reporte.

**TERCERA**

El Número de Consultas por Violencia Física en promedio comprende a 448 por reporte, el Número de Consultas por Violencia Psicológica en promedio son 299 consultas por reporte, el Número de Consultas por Violencia Sexual presenta una media de 88 consultas por reporte, el Numero de Consultas por Violencia Económica ese de 21 consultas por reporte. Lo que da a relucir que hay muchas más Consultas por Violencia Física.

# VI. RECOMENDACIONES

**PRIMERA**

Al Estado Peruano, a tomar medidas basadas en análisis de datos, así como es importante interconectar los sistemas informáticos y así aplicar sistemas automatizados en tiempo real con los algoritmos resultantes de la minería de datos brindando servicios más eficientes, así encaminarnos a un país de industria 4.0.

**SEGUNDA**

A tesistas interesados en la minería de datos considerar es el tema de la recopilación de información que no fue considerado dentro de los alcances de este trabajo, mediante herramientas de ingesta de datos de redes sociales como Flume.

**TERCERA**

A las instituciones públicas no considerar a la Minería de Datos como gastos innecesarios, se recomienda considerar como una inversión que proporciona ventajas competitivas implicando nuevas oportunidades de resolución de conflictos y de desarrollo nacional.

# VII. REFERENCIAS BIBLIOGRÁFICAS

Abad-Vich, D. (2016). Aplicación web para la extracción de reglas descriptivas en Minería de datos orientada al análisis de datos médicos.

Acosta Vargas, P., Medina, A., & González, M. (2020). Modelo de conglomerados para el uso del espacio público utilizando técnicas de minería de datos. Revista Ibérica de Sistemas e Tecnologias de Informação, (E27), 528-539.

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, *2*(4), 433-459.

Ávalos, A. A., Torreblanca, F. A., & Mamani, F. H. (2020). Impacto de las transferencias por canon-regalías en el índice de desarrollo humano y la pobreza de los distritos del Perú: aplicación de la técnica de minería de datos. Estudios del Desarrollo Social: Cuba y América Latina, 8(2), 245-258.

Arenas Conejo, M. (2015). Una Mirada Interseccional a la Violencia contra las Mujeres con Diversidad Funcional (An Intersectional Glance at Violence against Women with Functional Diversity). *Oñati socio-legal series*, *5*(2).

Bholowalia, P., & Kumar, A. (2014). EBK-means: A clustering technique based on elbow method and k-means in WSN. *International Journal of Computer Applications*, *105*(9).

Brusil Cruz, C. A. (2020). Análisis comparativo entre aprendizaje supervisado y aprendizaje semi-supervisado para la clasificación de señales sísmicas vulcanológicas del volcán Cotopaxi (Bachelor's thesis, Quito, 2020.).

Cambronero, C. G., & Moreno, I. G. (2006). Algoritmos de aprendizaje: knn & kmeans. Intelgencia en Redes de Comunicación, *Universidad Carlos III de Madrid*, *23*.

Cardoso, A. C., Talamé, M. L., Amor, M. N., & Monge, A. (2020). Creación de un corpus de opiniones con emociones usando aprendizaje automático. *Revista Tecnología y Ciencia*, (37), 11-23.

Carrascal, E. A. O., Carrascal, A. I. O., & Saldarriaga, G. L. V. (2015). Minería de datos: aportes y tendencias en el servicio de salud de ciudades inteligentes. Revista politécnica, 11(20), 111-120.

Carrascal, A. I. O., & Jiménez, G. A. (2018). Estudio sobre Estilos de Aprendizaje mediante Minería de Datos como apoyo a la Gestión Académica en Instituciones Educativas. RISTI-Revista Ibérica de Sistemas e Tecnologias de Informação, (29), 1-13.

Cordovilla Cordovilla, J. A. (2019). Diseño de un modelo predictivo, mediante la técnica de minería de datos, para identificar el perfil de éxito del estudiante en la unidad de titulación en la carrera de Ingeniería en Sistemas Computacionales de la Facultad de Ingeniería de la UCSG.

Corsi, J., & Bonino, L. (2003). Violencia y género: la construcción de la masculinidad como factor de riesgo. *Violencias Sociales Estudios sobre violencia. Barcelona, España: Editorial Ariel*.

Choque Soto, V. M. (2019). Minería de datos aplicada a la identificación de factores de deserción universitaria en programas de pre grado.

Del Val Román, J. L. (2016, March). Industria 4.0: la transformación digital de la industria. In *Valencia: Conferencia de Directores y Decanos de Ingeniería Informática, Informes CODDII*.

Dumont, J. R. D. (2015). Políticas públicas contra la violencia de género y evolución de las estadísticas de casos en el Perú. GRIN Verlag.

Escobar, C. A., & Morales-Menendez, R. (2018). Machine learning techniques for quality control in high conformance manufacturing environment. *Advances in Mechanical Engineering*, *10*(2), 1687814018755519.

Echeverry, C. E. M., Trujillo, M. L., & Salazar, M. H. M. (2017). Minería de datos en gestión del conocimiento de pymes de Colombia. Revista Virtual Universidad Católica del Norte, (50), 224-237.

Galán Cortina, V. (2016). Aplicación de la metodología CRISP-DM a un proyecto de minería de datos en el entorno universitario (Bachelor's tesis).

Gorbea Portal, S. (2013). Tendencias transdisciplinarias en los estudios métricos de la información y su relación con la gestión de la información y del conocimiento. *Perspectivas em Gestão & Conhecimento*, *3*(1), 13-27.

Hermosa, M. D. L., & Polo Usaola, C. (2018). Sexualidad, violencia sexual y salud mental. *Revista de la Asociación Española de Neuropsiquiatría*, *38*(134), 349-356.

Hernández-Sampieri, R., & Torres, C. P. M. (2018). Metodología de la investigación (Vol. 4). México eD. F DF: *McGraw-Hill Interamericana.*

Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis. *John Wiley & Sons*.

Khan, K., Rehman, S. U., Aziz, K., Fong, S., & Sarasvady, S. (2014, February). DBSCAN: Past, present and future. In *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)* (pp. 232-238). IEEE.

López Aguado, M., & Gutiérrez Provecho, L. (2019). Cómo realizar e interpretar un análisis factorial exploratorio utilizando SPSS. *REIRE Revista d'Innovació i Recerca en Educació*, *12*(2), 1-14.

Melo Chura, Alcides D. (2018). Patrones para la estimación de consumo de medicamentos con minería de datos redes Puno. Universidad Nacional del Altiplano.

Muñiz Ferrer, M. C., Jiménez García, Y., Ferrer Marrero, D., & González Pérez, J. (1998). La violencia familiar,¿ un problema de salud?. Revista Cubana de Medicina General Integral, 14(6), 538-541.

Neyra Carrasco, L., & Bazán Pérez, E. (2020). Estimación del Potencial de Energía Empleando Minería de Datos para el Diseño de un Sistema Fotovoltaico para el Sector San Isidro, Jaén–Perú.

Lagla, G. A. F., Moreano, J. A. C., Arequipa, E. E. Q., & Quishpe, M. W. V. (2019). Minería de datos como herramienta estratégica. RECIMUNDO: Revista Científica de la Investigación y el Conocimiento, 3(1), 955-970.

Orea, S. V., Vargas, A. S., & Alonso, M. G. (2005). Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. Ene, 779(73), 33.

Padilla Arias, W. R. (2019). Aplicación de técnicas de minería de datos geo- referenciados en los circuitos de comercialización alternativa de productos agrícolas 31 en Ecuador.

Parra Ugaz, M. G., & Villalobos Galbani, V. M. (2017). Análisis del servicio de atención urgente del Programa Nacional Contra la Violencia Familiar y Sexual. Universidad del Pacífico.

Pérez, C. (2008). Técnicas de análisis multivariante de datos. Aplicaciones con SPSS. Madrid: *Pearson, Prentice Hal.*

Pérez, E. R., & Medrano, L. A. (2010). Análisis factorial exploratorio: bases conceptuales y metodológicas. *Revista Argentina de Ciencias del Comportamiento (RACC), 2(1), 58-66.*

Piñerez, W. J. R., Ramírez, A. C., & Escobar, O. G. (2017). Análisis de datos funcionales aplicado en electroencefalogramas: agrupamiento por k-medias funcional. *Comunicaciones en Estadística*, *10*(1), 129-144.

Plaza, M. M. R., & Cardozo, D. D. Q. (2018, September). Implementación de algoritmo K-medias y modelo RGB para la Clasificación de Café Cereza. In *[2019-LISBOA] Congreso Internacional de Tecnología, Ciencia y Sociedad*.

Prada, Á. F. (2021). Perspectivas de la evolución dirigida en la Cuarta Revolución Industrial. *Futuro Hoy*, 8.

Regalado López, D. J. (2019). Reconocimiento de imágenes con técnicas de minería de datos (Bachelor's thesis, Quito: UCE).

Rioja Curo, W. M. (2020). Aplicación web para la elaboración de perfiles de consumidor basada en minería de datos y arquitectura cloud para el apoyo al proceso de conversión de leads en la asociación AIESEC en Perú.

Riquelme Santos, J. C., Ruiz, R., & Gilbert, K. (2006). Minería de datos: Conceptos y tendencias. Inteligencia Artificial: Revista Iberoamericana de Inteligencia Artificial, 10 (29), 11-18.

Rodríguez, O. (2010). Metodología para el desarrollo de proyectos en minería de datos CRISP-DM.

Rodríguez Calvo, M. D., Gómez Mendoza, C., Guevara de León, T., Arribas Llopis, A., Duarte Duran, Y., & Ruiz Álvarez, P. (2018). Violencia intrafamiliar en el adulto mayor. *Revista Archivo Médico de Camagüey*, *22*(2), 204-213

Rojas, F. M., & Gomez, C. (2014). Funcionalidades de la minería de datos. *Ingeniería y Región*, *12*, 31-40.

Rozas, T. P. A. (2020). Patrones de comportamiento en el uso de las aulas virtuales de la universidad nacional del altiplano área de ingenierías utilizando técnicas de minería de datos. *Revista de Investigaciones de la Escuela de Posgrado de la UNA PUNO, 9(4), 1833-1847*

Sánchez Álvarez, R. (2021). Clasificación no supervisada de imágenes médicas y minería de datos. Algoritmo S3 vs K-medias. *Revista Cubana de Investigaciones Biomédicas, 40*.

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)*, *42*(3), 1-21.

Ticona Condori, M. N. (2020). Árbol de decisiones para el análisis de la felicidad, resiliencia y optimismo en estudiantes de universidades licenciadas en el departamento de Puno–Perú 2019.

Torres, J. I. S., & Cardenas, E. G. (2021). Análisis y aplicación de algoritmos de minería de datos. *Revista Perspectivas*, *6*(21), 71-88.

Toro Merlo, J. J. (2013). Violencia sexual. *Revista de Obstetricia y Ginecología de Venezuela*, *73*(4), 217-220.

Vaca, P. V., & Díaz, M. C. R. (2009). Responsabilidad social de la Psicología frente a la violencia. *Pensamiento Psicológico*, *6*(13), 87-96.

Walton, S. M., & Pérez, C. A. S. (2019). La violencia intrafamiliar. Un problema de salud actual. *Gaceta médica espirituana*, *21*(1), 96-105.

Witten, I. H., & Frank, E. (2002). Data mining: practical machine learning tools and techniques with Java implementations. *Acm Sigmod Record*, *31*(1), 76-77.

# ANEXOS

Anexo 01.Datos recopilados

| Reporte | NCT - Total | NCT Hombres | NCT Mujeres | STC Sin sexo | NTC Violencia Psicológica | VTC Violencia Física | NTC Violencia Sexual | NTC Violencia Eco. | OTRAS CONSULTAS | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 363 | 37 | 326 | | 135 | 200 | 18 | | 10 | 2 |
| 2 | 325 | 36 | 289 | | 128 | 164 | 26 | | 7 | 2 |
| 3 | 241 | 34 | 207 | | 74 | 148 | 9 | | 10 | 2 |
| 4 | 1 645 | 109 | 1 536 | | 677 | 878 | 56 | | 34 | 1 |
| 5 | 571 | 91 | 480 | | 187 | 332 | 36 | | 16 | 2 |
| 6 | 1 091 | 119 | 972 | | 413 | 583 | 65 | | 30 | 1 |
| 7 | 2 510 | 226 | 2 284 | | 1 062 | 1 309 | 87 | | 52 | 1 |
| 8 | 918 | 107 | 811 | | 343 | 497 | 56 | | 22 | 1 |
| 9 | 168 | 21 | 147 | | 45 | 107 | 11 | | 5 | 2 |
| 10 | 425 | 49 | 376 | | 154 | 237 | 21 | | 13 | 2 |
| 11 | 1 139 | 101 | 1 038 | | 472 | 598 | 54 | | 15 | 1 |
| 12 | 1 108 | 97 | 1 011 | | 394 | 632 | 69 | | 13 | 1 |
| 13 | 1 851 | 194 | 1 657 | | 721 | 1 016 | 68 | | 46 | 1 |
| 14 | 502 | 54 | 448 | | 210 | 255 | 26 | | 11 | 2 |
| 15 | 23 576 | 2 137 | 21 439 | | 10 000 | 11 858 | 1 163 | | 555 | 1 |
| 16 | 412 | 48 | 364 | | 145 | 238 | 17 | | 12 | 2 |
| 17 | 149 | 19 | 130 | | 43 | 102 | 3 | | 1 | 2 |
| 18 | 162 | 12 | 150 | | 81 | 71 | 6 | | 4 | 2 |
| 19 | 249 | 16 | 233 | | 86 | 153 | 5 | | 5 | 2 |
| 20 | 1 775 | 144 | 1 631 | | 627 | 1 033 | 71 | | 44 | 1 |
| 21 | 1 040 | 109 | 931 | | 360 | 610 | 41 | | 29 | 1 |
| 22 | 773 | 129 | 644 | | 280 | 418 | 53 | | 22 | 1 |
| 23 | 309 | 22 | 287 | | 124 | 153 | 25 | | 7 | 2 |
| 24 | 269 | 22 | 247 | | 88 | 171 | 5 | | 5 | 2 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 25 | 473 | 47 | 426 | | | | | | 7 | 2 |
| 26 | 361 | 55 | 306 | | 130 | 189 | 31 | | 11 | 2 |
| 27 | 559 | 76 | 483 | | 210 | 295 | 41 | | 13 | 2 |
| 28 | 275 | 37 | 238 | | 96 | 160 | 14 | | 5 | 2 |
| 29 | 1 532 | 112 | 1 420 | | 633 | 782 | 67 | | 50 | 1 |
| 30 | 593 | 77 | 516 | | 195 | 343 | 37 | | 18 | 2 |
| 31 | 1 155 | 155 | 1 000 | | 377 | 641 | 99 | | 38 | 1 |
| 32 | 2 000 | 212 | 1 788 | | 825 | 990 | 114 | | 71 | 1 |
| 33 | 961 | 127 | 834 | | 301 | 568 | 63 | | 29 | 1 |
| 34 | 220 | 56 | 164 | | 79 | 120 | 16 | | 5 | 2 |
| 35 | 589 | 73 | 516 | | 215 | 316 | 43 | | 15 | 2 |
| 36 | 1 175 | 153 | 1 022 | | 442 | 635 | 63 | | 35 | 1 |
| 37 | 1 085 | 142 | 943 | | 389 | 582 | 82 | | 32 | 1 |
| 38 | 1 932 | 212 | 1 720 | | 708 | 1 036 | 118 | | 70 | 1 |
| 39 | 468 | 61 | 407 | | 185 | 245 | 28 | | 10 | 2 |
| 40 | 21 412 | 2 159 | 19 253 | | 9 155 | 10 288 | 1 235 | | 734 | 2 |
| 41 | 370 | 58 | 312 | | 121 | 198 | 32 | | 19 | 2 |
| 42 | 143 | 10 | 133 | | 50 | 84 | 8 | | 1 | 2 |
| 43 | 132 | 14 | 118 | | 54 | 59 | 13 | | 6 | 2 |
| 44 | 280 | 17 | 263 | | 92 | 166 | 18 | | 4 | 2 |
| 45 | 1 759 | 174 | 1 585 | | 646 | 990 | 80 | | 43 | 1 |
| 46 | 1 109 | 134 | 975 | | 348 | 654 | 66 | | 41 | 1 |
| 47 | 874 | 148 | 726 | | 308 | 468 | 62 | | 36 | 1 |
| 48 | 332 | 28 | 304 | | 131 | 176 | 20 | | 5 | 2 |
| 49 | 231 | 20 | 211 | | 86 | 133 | 10 | | 2 | 2 |
| 50 | 439 | 46 | 393 | | 180 | 231 | 21 | | 7 | 2 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 51 | 337 | 44 | 293 | | 135 | 171 | 10 | | 21 | 2 |
| 52 | 684 | 96 | 588 | | 264 | 331 | 48 | | 41 | 2 |
| 53 | 222 | 47 | 175 | | 74 | 122 | 15 | | 11 | 2 |
| 54 | 1 372 | 107 | 1 265 | | 584 | 684 | 42 | | 62 | 1 |
| 55 | 583 | 85 | 498 | | 190 | 348 | 33 | | 12 | 2 |
| 56 | 883 | 91 | 792 | | 340 | 473 | 46 | | 24 | 1 |
| 57 | 2 734 | 242 | 2 492 | | 1 222 | 1 295 | 103 | | 114 | 1 |
| 58 | 887 | 84 | 803 | | 340 | 468 | 35 | | 44 | 1 |
| 59 | 134 | 27 | 107 | | 36 | 79 | 14 | | 5 | 2 |
| 60 | 375 | 40 | 335 | | 148 | 185 | 21 | | 21 | 2 |
| 61 | 1 191 | 86 | 1 105 | | 510 | 601 | 47 | | 33 | 1 |
| 62 | 1 079 | 109 | 970 | | 413 | 570 | 55 | | 41 | 1 |
| 63 | 1 785 | 144 | 1 641 | | 695 | 933 | 72 | | 85 | 1 |
| 64 | 805 | 52 | 753 | | 311 | 432 | 23 | | 39 | 1 |
| 65 | 26 704 | 2 395 | 24 309 | | 11 489 | 13 034 | 1 110 | | 1 071 | 1 |
| 66 | 387 | 52 | 335 | | 157 | 184 | 20 | | 26 | 2 |
| 67 | 122 | 8 | 114 | | 49 | 67 | 4 | | 2 | 2 |
| 68 | 167 | 17 | 150 | | 71 | 85 | 5 | | 6 | 2 |
| 69 | 223 | 25 | 198 | | 86 | 117 | 8 | | 12 | 2 |
| 70 | 1 690 | 151 | 1 539 | | 625 | 935 | 55 | | 75 | 1 |
| 71 | 1 031 | 86 | 945 | | 360 | 597 | 41 | | 33 | 1 |
| 72 | 593 | 63 | 530 | | 238 | 306 | 30 | | 19 | 2 |
| 73 | 307 | 23 | 284 | | 116 | 163 | 22 | | 6 | 2 |
| 74 | 210 | 13 | 197 | | 92 | 109 | 3 | | 6 | 2 |
| 75 | 327 | 22 | 305 | | 114 | 181 | 19 | | 13 | 2 |
| 76 | 367 | 64 | 303 | | 140 | 172 | 34 | | 21 | 2 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 77 | 345 | 65 | 280 | | 124 | 186 | 14 | | 21 | 2 |
| 78 | 322 | 63 | 259 | | 90 | 210 | 9 | | 13 | 2 |
| 79 | 1 445 | 130 | 1 315 | | 678 | 629 | 70 | | 68 | 1 |
| 80 | 620 | 107 | 513 | | 213 | 353 | 32 | | 22 | 2 |
| 81 | 1 258 | 182 | 1 076 | | 421 | 694 | 82 | | 61 | 1 |
| 82 | 2 060 | 249 | 1 811 | | 901 | 961 | 102 | | 96 | 1 |
| 83 | 1 098 | 142 | 956 | | 392 | 591 | 56 | | 59 | 1 |
| 84 | 249 | 64 | 185 | | 78 | 143 | 19 | | 9 | 2 |
| 85 | 664 | 101 | 563 | | 193 | 381 | 61 | | 29 | 2 |
| 86 | 1 157 | 113 | 1 044 | | 501 | 542 | 71 | | 43 | 1 |
| 87 | 962 | 103 | 859 | | 340 | 504 | 76 | | 42 | 1 |
| 88 | 1 877 | 261 | 1 616 | | 750 | 922 | 108 | | 97 | 1 |
| 89 | 416 | 51 | 365 | | 164 | 209 | 23 | | 20 | 2 |
| 90 | 20 602 | 2 473 | 18 129 | | 8 859 | 9 413 | 1 207 | | 1 123 | 1 |
| 91 | 401 | 69 | 332 | | 163 | 184 | 40 | | 14 | 2 |
| 92 | 131 | 23 | 108 | | 42 | 73 | 12 | | 4 | 2 |
| 93 | 142 | 9 | 133 | | 66 | 64 | 7 | | 5 | 2 |
| 94 | 227 | 33 | 194 | | 89 | 112 | 14 | | 12 | 2 |
| 95 | 1 706 | 191 | 1 515 | | 632 | 936 | 76 | | 62 | 1 |
| 96 | 921 | 126 | 795 | | 290 | 548 | 46 | | 37 | 1 |
| 97 | 944 | 163 | 781 | | 335 | 473 | 91 | | 45 | 1 |
| 98 | 287 | 28 | 259 | | 110 | 144 | 20 | | 13 | 2 |
| 99 | 202 | 28 | 174 | | 76 | 112 | 10 | | 4 | 2 |
| 100 | 396 | 57 | 339 | | 149 | 201 | 29 | | 17 | 2 |
| 101 | 432 | 85 | 347 | | 159 | 207 | 37 | | 29 | 2 |
| 102 | 661 | 97 | 564 | | 222 | 344 | 48 | | 47 | 2 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 103 | 385 | 74 | 311 | | 115 | 226 | 25 | | 19 | 2 |
| 104 | 1 505 | 164 | 1 341 | | 646 | 676 | 99 | | 84 | 1 |
| 105 | 812 | 148 | 664 | | 247 | 461 | 57 | | 47 | 1 |
| 106 | 1 452 | 273 | 1 179 | | 437 | 787 | 127 | | 101 | 1 |
| 107 | 2 397 | 332 | 2 065 | | 964 | 1 139 | 170 | | 124 | 1 |
| 108 | 1 455 | 199 | 1 256 | | 517 | 776 | 98 | | 64 | 1 |
| 109 | 303 | 84 | 219 | | 98 | 152 | 34 | | 19 | 2 |
| 110 | 742 | 154 | 588 | | 237 | 382 | 74 | | 49 | 2 |
| 111 | 1 069 | 165 | 904 | | 426 | 502 | 77 | | 64 | 1 |
| 112 | 1 237 | 197 | 1 040 | | 443 | 598 | 137 | | 59 | 1 |
| 113 | 2 080 | 314 | 1 766 | | 810 | 1 016 | 150 | | 104 | 1 |
| 114 | 811 | 114 | 697 | | 310 | 393 | 74 | | 34 | 1 |
| 115 | 23 804 | 3 176 | 20 628 | | 10 050 | 10 619 | 1 613 | | 1 522 | 1 |
| 116 | 1 039 | 124 | 915 | | 375 | 520 | 88 | | 56 | 1 |
| 117 | 571 | 130 | 441 | | 185 | 294 | 64 | | 28 | 2 |
| 118 | 237 | 42 | 195 | | 52 | 146 | 29 | | 10 | 2 |
| 119 | 193 | 47 | 146 | | 58 | 115 | 10 | | 10 | 2 |
| 120 | 223 | 43 | 180 | | 79 | 117 | 16 | | 11 | 2 |
| 121 | 2 015 | 284 | 1 731 | | 690 | 1 110 | 132 | | 83 | 1 |
| 122 | 1 348 | 194 | 1 154 | | 445 | 751 | 70 | | 82 | 1 |
| 123 | 914 | 211 | 703 | | 270 | 474 | 103 | | 67 | 1 |
| 124 | 300 | 37 | 263 | | 122 | 147 | 17 | | 14 | 2 |
| 125 | 241 | 47 | 194 | | 59 | 152 | 18 | | 12 | 2 |
| 126 | 419 | 74 | 345 | | 155 | 205 | 42 | | 17 | 2 |
| 127 | 506 | 115 | 391 | | 120 | 283 | 71 | | 32 | 2 |
| 128 | 1 082 | 241 | 841 | | 310 | 622 | 92 | | 58 | 1 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 129 | 631 | 113 | 518 | | 139 | 406 | 53 | | 33 | 2 |
| 130 | 2 414 | 319 | 2 095 | | 811 | 1 163 | 260 | | 180 | 1 |
| 131 | 954 | 153 | 801 | | 254 | 546 | 102 | | 52 | 1 |
| 132 | 1 719 | 304 | 1 415 | | 464 | 936 | 179 | | 140 | 1 |
| 133 | 3 424 | 488 | 2 936 | | 1 072 | 1 801 | 306 | | 245 | 1 |
| 134 | 2 137 | 395 | 1 742 | | 545 | 1 273 | 184 | | 135 | 1 |
| 135 | 358 | 115 | 243 | | 93 | 192 | 53 | | 20 | 2 |
| 136 | 1 189 | 267 | 922 | | 321 | 670 | 112 | | 86 | 1 |
| 137 | 1 553 | 196 | 1 357 | | 455 | 842 | 153 | | 103 | 1 |
| 138 | 1 957 | 333 | 1 624 | | 521 | 1 095 | 230 | | 111 | 1 |
| 139 | 2 756 | 411 | 2 345 | | 794 | 1 554 | 235 | | 173 | 1 |
| 140 | 1 402 | 205 | 1 197 | | 420 | 755 | 139 | | 88 | 1 |
| 141 | 32 487 | 4 642 | 27 845 | | 10 475 | 16 587 | 2 888 | | 2 537 | 1 |
| 142 | 1 548 | 226 | 1 322 | | 450 | 847 | 139 | | 112 | 1 |
| 143 | 704 | 158 | 546 | | 164 | 412 | 79 | | 49 | 2 |
| 144 | 319 | 70 | 249 | | 48 | 195 | 58 | | 18 | 2 |
| 145 | 319 | 51 | 268 | | 105 | 175 | 16 | | 23 | 2 |
| 146 | 387 | 69 | 318 | | 102 | 222 | 37 | | 26 | 2 |
| 147 | 2 692 | 389 | 2 303 | | 792 | 1 525 | 226 | | 149 | 1 |
| 148 | 1 713 | 247 | 1 466 | | 509 | 974 | 136 | | 94 | 1 |
| 149 | 1 531 | 306 | 1 225 | | 372 | 880 | 186 | | 93 | 1 |
| 150 | 465 | 63 | 402 | | 159 | 226 | 45 | | 35 | 2 |
| 151 | 303 | 45 | 258 | | 84 | 185 | 19 | | 15 | 2 |
| 152 | 518 | 89 | 429 | | 129 | 297 | 66 | | 26 | 2 |
| 153 | 579 | 113 | 339 | 127 | 131 | 284 | 83 | 1 | 80 | 2 |
| 154 | 1 211 | 179 | 829 | 203 | 282 | 612 | 125 | 1 | 191 | 1 |

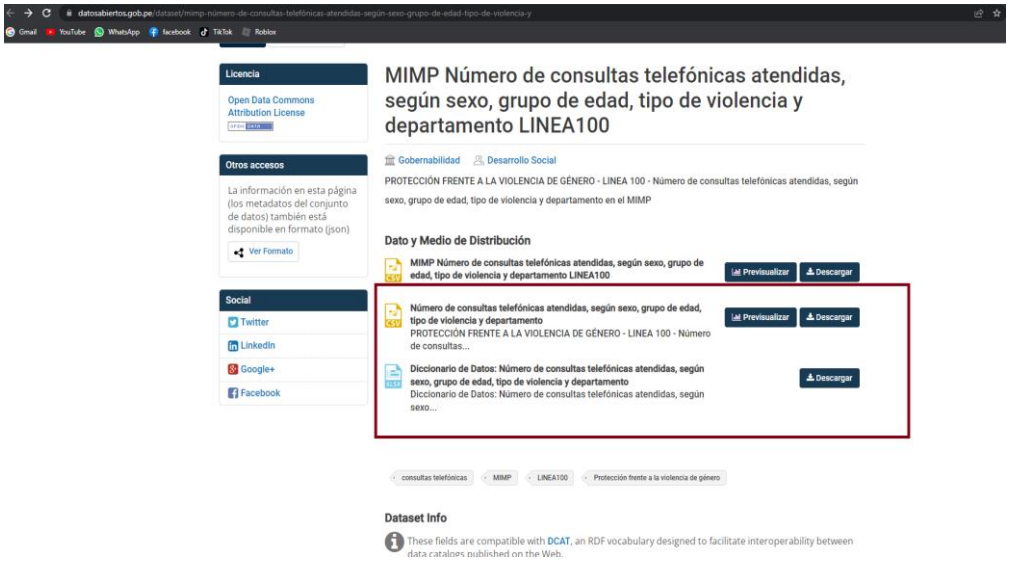
| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 155 | 703 | 131 | 462 | 110 | 147 | 402 | 76 | 3 | 75 | 2 |
| 156 | 3 127 | 379 | 2 332 | 416 | 892 | 1 381 | 324 | 10 | 520 | 1 |
| 157 | 1 099 | 183 | 781 | 135 | 230 | 602 | 114 | 4 | 149 | 1 |
| 158 | 1 865 | 285 | 1 294 | 286 | 405 | 901 | 254 | 8 | 297 | 1 |
| 159 | 3 809 | 486 | 2 651 | 672 | 961 | 1 901 | 322 | 19 | 606 | 1 |
| 160 | 2 150 | 332 | 1 548 | 270 | 490 | 1 151 | 206 | 9 | 294 | 1 |
| 161 | 503 | 118 | 324 | 61 | 86 | 282 | 75 | 1 | 59 | 2 |
| 162 | 1 293 | 246 | 895 | 152 | 271 | 655 | 165 | 3 | 199 | 1 |
| 163 | 1 845 | 263 | 1 294 | 288 | 519 | 880 | 173 | 5 | 268 | 1 |
| 164 | 2 290 | 343 | 1 624 | 323 | 490 | 1 163 | 286 | 4 | 347 | 1 |
| 165 | 3 180 | 473 | 2 136 | 571 | 781 | 1 646 | 312 | 8 | 433 | 1 |
| 166 | 1 473 | 199 | 1 041 | 233 | 390 | 721 | 140 | 2 | 220 | 1 |
| 167 | 39 363 | 5 335 | 27 895 | 6 133 | 10 371 | 17 907 | 3 644 | 121 | 7 320 | 2 |
| 168 | 1 561 | 217 | 1 103 | 241 | 362 | 788 | 175 | 7 | 229 | 1 |
| 169 | 750 | 135 | 476 | 139 | 159 | 402 | 96 | 2 | 91 | 2 |
| 170 | 436 | 56 | 265 | 115 | 70 | 249 | 58 | 0 | 59 | 2 |
| 171 | 351 | 48 | 214 | 89 | 102 | 160 | 36 | 1 | 52 | 2 |
| 172 | 333 | 64 | 220 | 49 | 72 | 161 | 37 | 2 | 61 | 2 |
| 173 | 2 928 | 353 | 2 044 | 531 | 709 | 1 565 | 262 | 5 | 387 | 1 |
| 174 | 1 903 | 285 | 1 418 | 200 | 443 | 1 063 | 140 | 3 | 254 | 1 |
| 175 | 1 716 | 275 | 1 088 | 353 | 341 | 929 | 211 | 1 | 234 | 1 |
| 176 | 579 | 91 | 414 | 74 | 151 | 273 | 61 | 1 | 93 | 2 |
| 177 | 326 | 35 | 219 | 72 | 69 | 193 | 28 | 1 | 35 | 2 |
| 178 | 615 | 102 | 408 | 105 | 132 | 310 | 96 | 0 | 77 | 2 |
| 179 | 894 | 206 | 688 | 0 | 174 | 404 | 116 | 1 | 199 | 1 |
| 180 | 1 824 | 388 | 1 436 | 0 | 406 | 864 | 188 | 5 | 361 | 1 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 181 | 1 039 | 240 | 799 | 0 | 221 | 567 | 93 | 0 | 158 | 1 |
| 182 | 4 867 | 864 | 4 003 | 0 | 1 241 | 2 167 | 437 | 9 | 1 013 | 1 |
| 183 | 1 793 | 370 | 1 423 | 0 | 448 | 864 | 173 | 5 | 303 | 1 |
| 184 | 2 804 | 584 | 2 220 | 0 | 644 | 1 274 | 299 | 10 | 577 | 1 |
| 185 | 6 163 | 1 005 | 5 158 | 0 | 1 587 | 2 574 | 490 | 20 | 1 492 | 1 |
| 186 | 3 611 | 737 | 2 874 | 0 | 867 | 1 742 | 297 | 10 | 695 | 1 |
| 187 | 711 | 239 | 472 | 0 | 142 | 374 | 77 | 2 | 116 | 2 |
| 188 | 2 077 | 449 | 1 628 | 0 | 486 | 944 | 222 | 2 | 423 | 1 |
| 189 | 3 030 | 517 | 2 513 | 0 | 764 | 1 364 | 254 | 5 | 643 | 1 |
| 190 | 3 828 | 731 | 3 097 | 0 | 828 | 1 849 | 450 | 11 | 690 | 1 |
| 191 | 4 660 | 829 | 3 831 | 0 | 1 137 | 2 171 | 445 | 11 | 896 | 1 |
| 192 | 2 418 | 469 | 1 949 | 0 | 627 | 1 132 | 201 | 4 | 454 | 1 |
| 193 | 61 971 | 11 038 | 50 933 | 0 | 15 980 | 24 624 | 5 222 | 210 | 15 935 | 1 |
| 194 | 2 611 | 458 | 2 153 | 0 | 613 | 1 180 | 275 | 5 | 538 | 1 |
| 195 | 1 240 | 313 | 927 | 0 | 270 | 642 | 148 | 4 | 176 | 1 |
| 196 | 734 | 122 | 612 | 0 | 118 | 401 | 100 | 4 | 111 | 2 |
| 197 | 511 | 106 | 405 | 0 | 136 | 213 | 48 | 3 | 111 | 2 |
| 198 | 563 | 109 | 454 | 0 | 127 | 288 | 57 | 0 | 91 | 2 |
| 199 | 4 515 | 769 | 3 746 | 0 | 1 142 | 2 234 | 333 | 4 | 802 | 1 |
| 200 | 2 974 | 501 | 2 473 | 0 | 740 | 1 408 | 291 | 8 | 527 | 1 |
| 201 | 2 414 | 557 | 1 857 | 0 | 483 | 1 132 | 304 | 6 | 489 | 1 |
| 202 | 919 | 188 | 731 | 0 | 273 | 380 | 82 | 2 | 182 | 1 |
| 203 | 569 | 101 | 468 | 0 | 140 | 300 | 36 | 0 | 93 | 2 |
| 204 | 1 046 | 199 | 847 | 0 | 222 | 531 | 120 | 3 | 170 | 1 |
| 205 | 83 | 20 | 63 | 0 | 16 | 36 | 8 | 0 | 23 | 2 |
| 206 | 184 | 44 | 140 | 0 | 51 | 68 | 22 | 1 | 42 | 2 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 207 | 105 | 20 | 85 | 0 | 29 | 50 | 13 | 0 | 13 | 2 |
| 208 | 466 | 76 | 390 | 0 | 157 | 157 | 32 | 1 | 119 | 2 |
| 209 | 165 | 32 | 133 | 0 | 48 | 83 | 10 | 0 | 24 | 2 |
| 210 | 266 | 53 | 213 | 0 | 51 | 116 | 27 | 0 | 72 | 2 |
| 211 | 609 | 96 | 513 | 0 | 153 | 249 | 48 | 0 | 159 | 2 |
| 212 | 309 | 70 | 239 | 0 | 71 | 131 | 31 | 0 | 76 | 2 |
| 213 | 61 | 22 | 39 | 0 | 10 | 33 | 3 | 0 | 15 | 2 |
| 214 | 209 | 38 | 171 | 0 | 52 | 89 | 25 | 0 | 43 | 2 |
| 215 | 329 | 41 | 288 | 0 | 79 | 160 | 23 | 1 | 66 | 2 |
| 216 | 356 | 72 | 284 | 0 | 83 | 146 | 46 | 2 | 79 | 2 |
| 217 | 594 | 124 | 470 | 0 | 166 | 248 | 47 | 1 | 132 | 2 |
| 218 | 282 | 71 | 211 | 0 | 62 | 125 | 25 | 1 | 69 | 2 |
| 219 | 7 051 | 1 335 | 5 716 | 0 | 1 839 | 2 590 | 548 | 24 | 2 050 | 1 |
| 220 | 332 | 64 | 268 | 0 | 84 | 156 | 33 | 0 | 59 | 2 |
| 221 | 112 | 21 | 91 | 0 | 23 | 57 | 11 | 1 | 20 | 2 |
| 222 | 62 | 6 | 56 | 0 | 17 | 30 | 5 | 0 | 10 | 2 |
| 223 | 46 | 8 | 38 | 0 | 12 | 25 | 3 | 0 | 6 | 2 |
| 224 | 53 | 13 | 40 | 0 | 16 | 17 | 5 | 0 | 15 | 2 |
| 225 | 471 | 73 | 398 | 0 | 137 | 216 | 24 | 3 | 91 | 2 |
| 226 | 275 | 46 | 229 | 0 | 75 | 103 | 27 | 1 | 69 | 2 |
| 227 | 234 | 62 | 172 | 0 | 55 | 110 | 21 | 1 | 47 | 2 |
| 228 | 95 | 22 | 73 | 0 | 24 | 38 | 9 | 0 | 24 | 2 |
| 229 | 52 | 5 | 47 | 0 | 10 | 29 | 4 | 0 | 9 | 2 |
| 230 | 92 | 17 | 75 | 0 | 21 | 39 | 14 | 0 | 18 | 2 |

Anexo 02.Acceso a los datos línea 100

https://www.datosabiertos.gob.pe/dataset/mimp-n%C3%BAmero-de-consultas-telef%C3%B3nicas-atendidas-seg%C3%BAn-sexo-grupo-de-edad-tipo-de-violencia-y



Anexo 03.Código utilizado

```
library(mice)
library(factoextra)
library(ggpubr)
library(psych)
library(FactoMineR)
library(corrplot)
gggdatos <- read.csv("Datos.csv",stringsAsFactors = F)
datos <- as.data.frame(apply(datos, 2, as.numeric))
na_count <-function (x) sapply(x, function(y) sum(is.na(y)))
vacios<-na_count(datos)
vacios<-sort(vacios/length(datos$NCT_Total)*100)
imputed_Data <- mice(datos, m=5, maxit = 20, method = 'cart', seed = 500)
datos <- complete(imputed_Data)
    colnames(datos) <-
    c('NCT_Total','NCT_Hombres','NCT_Mujeres','NCT_sin_Sexo','NCT_V_PS','N
    CT_VF','NCT_VS','NCT_VE','OTRAS_C')


#histogramas
gghistogram(datos, x = "NCT_Total", fill = "red",
    add = "mean", rug = TRUE)
gghistogram(datos, x = "NCT_Hombres", fill = "blue",
    add = "mean", rug = TRUE)
gghistogram(datos, x = "NCT_Mujeres", fill = "green",
    add = "mean", rug = TRUE)
gghistogram(datos, x = "NCT_sin_Sexo", fill = "black",
    add = "mean", rug = TRUE)
gghistogram(datos, x = "NCT_V_PS", fill = "orange",
    add = "mean", rug = TRUE)
gghistogram(datos, x = "NCT_VF", fill = "red",
    add = "mean", rug = TRUE)
gghistogram(datos, x = "NCT_VS", fill = "yellow",
    add = "mean", rug = TRUE)
```

```
gghistogram(datos, x = "NCT_VE", fill = "lightgray",
          add = "mean", rug = TRUE)
gghistogram(datos, x = "OTRAS_C", fill = "lightgray",
          add = "mean", rug = TRUE)
cor1<-cor(datos)
corrplot(cor1,method='number',order = 'hclust',addrect = 3,addgrid.col = "darkgray",number.cex = 0.75)
kmo<- KMO(datos)
print(kmo)
k2 <- kmeans(datos, centers = 2)
k4 <- kmeans(datos, centers = 4)
k6 <- kmeans(datos, centers = 6)
k8 <- kmeans(datos, centers = 8)
k10 <- kmeans(datos, centers = 10)
k12 <- kmeans(datos, centers = 12)
k2 <-fviz_cluster(k2,datos,geom ="point",main="")
k4 <-fviz_cluster(k4,datos,geom='point',main="")
k6 <-fviz_cluster(k6,datos,geom='point',main="")
k8 <-fviz_cluster(k8,datos,geom='point',main="")
k10<-fviz_cluster(k10,datos,geom='point',main="")
k12<-fviz_cluster(k12,datos,geom='point',main="")
ggarrange(k2, k4,k6,k8,k10,k12 + rremove("x.text"),
        labels = c(" k=2", " k=4", " k=6"," k=8"," k=10",' k=12'))
fviz_screeplot(res.pca, addlabels = TRUE, ylim = c(0, 100))+
          labs(x="Clusters / Agrupaciones",y="Porcentaje de varianza explicado",main=NULL)
res.pca <- PCA(datos,  graph = FALSE)
var <- get_pca_var(res.pca)
grp <- as.factor(k2$cluster)
res.km <- kmeans(var$coord, centers = 2, nstart = 25)
grp <- as.factor(res.km$cluster)
fviz_pca_var(res.pca, col.var = grp,
           palette = c("red", "blue"),
```

```
        legend.title = "Cluster")
        #numero optimo
set.seed(123)
fviz_nbclust(datos, kmeans, method = "wss")
set.seed(123)
fviz_nbclust(datos, kmeans, method = "silhouette")+
      labs(x = 'Número de Clústers', y = 'Calidad de agrupamiento',title=NULL)
```