METODOLOGÍA PARA EL ANÁLISIS DE LA VIOLENCIA EN EL DEPARTAMENTO DE BOLÍVAR MEDIANTE TÉCNICAS DE MACHINE LEARNING

EVER FERNÁNDEZ CARABALLO

YAQUELINE GÓMEZ FRANCO

UNIVERSIDAD TECNOLÓGICA DE BOLÍVAR

FACULTAD DE INGENIERÍA

INGENIERÍA INDUSTRIAL

CARTAGENA

2018

METODOLOGÍA PARA EL ANÁLISIS DE LA VIOLENCIA EN EL DEPARTAMENTO DE BOLÍVAR MEDIANTE TÉCNICAS DE MACHINE LEARNING

EVER FERNÁNDEZ CARABALLO

YAQUELINE GÓMEZ FRANCO

TRABAJO DE GRADO PARA OPTAR AL TÍTULO DE INGENIERO INDUSTRIAL

DIRECTOR
ENRIQUE JOSÉ DE LA HOZ DOMÍNGUEZ
M.S.C INVESTIGACIÓN DE OPERACIONES

UNIVERSIDAD TECNOLÓGICA DE BOLÍVAR
FACULTAD DE INGENIERÍAS
PROGRAMA DE INGENIERÍA INDUSTRIAL
CARTAGENA DE INDIAS D.T Y C.

CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

INTRODUCCIÓN

A partir de la crisis del siglo xx, Colombia se vio afectada por una ola de violencia caracterizada por el aumento en los niveles de homicidios, agresiones, persecuciones, destrucción de la propiedad privada y terrorismo por la afiliación política. Esta situación fue más profunda en las principales ciudades, la cual estalló con el magnicidio de Jorge Eliecer Gaitán que tuvo como consecuencia el llamado "colombianazo" que dejó muchas víctimas en todo el país, que hoy, 70 años después de confrontaciones armadas en violencias sucesivas, ha cobrado la vida de cerca de un millón de personas, casi la mitad de la población de Medellín [1].

En este sentido, para poder prevenir, diseñar e implementar planes efectivos de contingencia contra estos delitos es fundamental el análisis de los registros de violencia. En Colombia este tipo de análisis se ha realizado históricamente mediante el uso de diversas herramientas tales como minería de datos, estadísticas descriptivas básicas, entre otras. Un ejemplo de estas se da con el estudio realizado en la ciudad de Bogotá en el cual implementaron técnicas de minería de datos para realizar exploración de la violencia sexual[2].

Así pues, para poder analizar grandes cantidades de datos es necesario el uso de programas que cuenten con una gran capacidad y eficiencia de almacenamiento, por esta razón, en muchos casos los registros almacenados son demasiado grandes o complejos como para analizar y superan el alcance de la estadística, por consiguiente, el machine learning es fundamental al ser una rama de la inteligencia artificial que busca que una máquina aprenda como un ser humano y que pueda desarrollar sus funciones mediante el uso de distintos algoritmos.

El uso de machine learning es de gran ayuda al existir una gran cantidad de información y de variables intervinientes que justifican el uso de herramientas más potentes que la estadística convencional. Así pues, la creación de una metodología para el análisis de la violencia mediante el uso de machine learning hace que se implementen técnicas avanzadas en nuevos campos.

En este contexto, el objetivo de este trabajo es realizar la metodología para el análisis de la violencia en el departamento de Bolívar mediante técnicas de machine learning evaluando la información de la violencia presente en la página datos.gov.co y comprobando su efectividad, replicabilidad y valor agregado.

1. Descripción del problema

1.1 Contexto global

La historia de la humanidad ha sido marcada por la presencia de la violencia en sus diferentes formas (personal, sexual, intrafamiliar, entre otras), las cuales se pueden evidenciar en todas las partes del mundo. La Cuadragésima Novena Asamblea Mundial de la Salud adoptó la resolución WHA49.25, que declara a la violencia como un importante y creciente problema de salud pública en todo el mundo [1]. Por lo que debe ser tomado muy en serio. Cada año, más de un millón de personas pierden la vida y muchas más sufren lesiones no mortales como resultado de la violencia autoinfligida, interpersonal o colectiva[3]. La violencia es uno de los fenómenos que más afectan la calidad de vida del hombre, así como su entorno social y cultural.

Para comprender mejor el tema debemos conocer en si cual es el concepto de violencia, La Organización Mundial de la Salud define violencia como: El uso intencional de fuerza física o poder, amenazado o real, contra uno mismo, contra otra persona o contra un grupo o comunidad, que resulta en o tiene una alta probabilidad de resultar en lesión, muerte, daño psicológico, mal desarrollo o privación [1]. Por otro lado, el termino violencia Según la Real Academia Española se puede definir de la siguiente forma: Cualidad de violento, acción y efecto de violentar o violentarse, acción violenta o contra el natural modo de proceder o acción de violar a una persona [4].

La violencia es algo que es muy difícil de medir, aunque se puede calcular cierta parte de la violencia a partir de algunas estadísticas que publica la policía como los datos con los que trabajaremos. Una parte difícil de medir es la menos marginal porque las declaraciones dependen de la confianza que se tenga en la policía y en la justicia y, en general, ésta no es muy alta en América del Sur. Otro problema para cuantificar este fenómeno es la existencia de varios grados de violencia. Estos, que van de los homicidios voluntarios a las infracciones en materia de droga pasando por infracciones sexuales, golpes y lesiones, robos a mano armada, estafas y falsificación de moneda, por ejemplo, dificultan la agregación de los hechos violentos[5].

Esta hace un daño inmenso a la sociedad afectando principalmente la integridad del hombre. Así pues, observamos como américa latina se ha convertido en uno de las regiones más violentas y más desiguales[6]. Cinco de los 10 países más desiguales del mundo están en América, entre ellos Brasil. El último quintil de ingreso tiene el 2.9% del ingreso en América Latina, mientras en Asia es el 8.7%, y en Europa el 6,6%.3. En América Latina el 20% más rico tiene el 57.8% del ingreso. Al mismo tiempo, tiene el 9% de la población del mundo y el 27% de los homicidios y 10 de los 20 países con mayores tasas de homicidios del mundo son Latinoamericanos [6].

Aproximadamente uno de cada tres latinoamericanos es pobre, significa que no tienen suficientes ingresos para satisfacer sus necesidades básicas y uno de cada 8 se encuentra en pobreza extrema definido a esta como la incapacidad de cubrir sus necesidades nutricionales básicas [7].

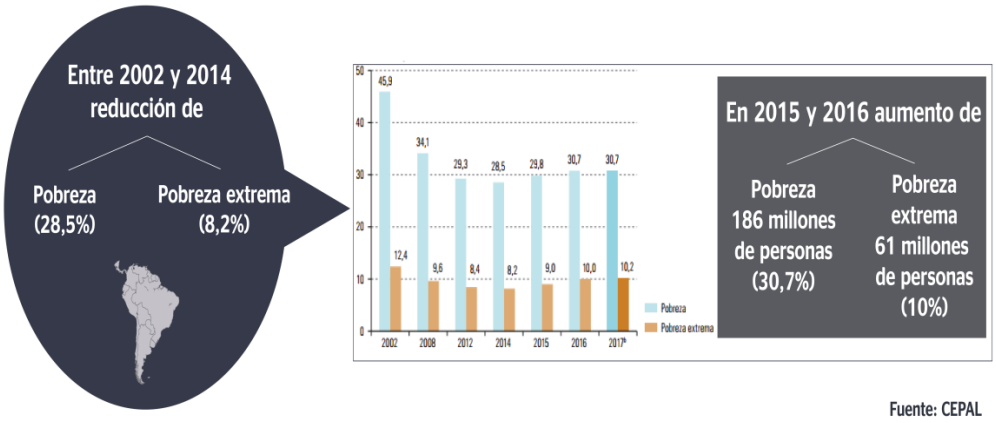
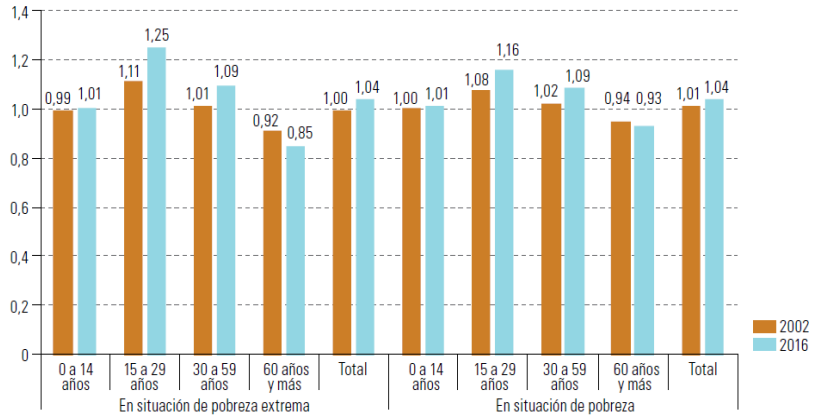


Figura 1. *Pobreza en América Latina 2002-2017*

Fuente: CEPAL, Panorama social de América Latina 2017, Grafico 11.1, p. 88 (edición en español)

De acuerdo a la gráfica anterior se refleja un incremento en la pobreza a partir de 2015, donde llegó a afectar al 29,8 % de la población, es decir, afectó a 178 millones de personas y la pobreza extrema llegó al 9 % de la población equivalente a 54 millones de personas. En el año 2017 la pobreza llegó a 187 millones de personas, es decir, el 30,7% de la población, mientras que la pobreza extrema afectó al 10,2% de la población, cifra equivalente a 62 millones de personas [8]

Por otro lado, las tasas de pobreza extrema en américa Latina son relativamente altas dado el nivel de PIB per capital, se encuentra por encima de aquellos del Medio Oriente y Norte de África, sus niveles de pobreza extrema son más altos [7].



Fuente: Comisión Económica para América Latina y el Caribe (CEPAL), sobre la base de Banco de Datos de Encuestas de Hogares (BADEHOG).
[a] Promedio ponderado de los siguientes países: Argentina, Bolivia (Estado Plurinacional de), Brasil, Chile, Colombia, Costa Rica, Ecuador, El Salvador, Guatemala, Honduras, México, Nicaragua, Panamá, Paraguay, Perú, República Dominicana, Uruguay y Venezuela (República Bolivariana de).

Figura 2. *Tasas de pobreza y pobreza extrema entre mujeres y hombres, por grupos de edad, 2002 y 2016*

Fuente: CEPAL, Panorama social de América Latina 2017; gráfico 11.10, P. 102 (edición en español);

De acuerdo a estas cifras podemos decir que la situación de pobreza y pobreza extrema ha aumentado, siendo las mujeres entre los 15 a 29 años las más afectadas. Así pues, el hecho de que la extrema pobreza en el grupo de edad de 60 años y más se mantenga y disminuya en los casos de pobreza extrema, podría reflejar el papel desempeñado por los sistemas de pensiones no contributivos, cuya cobertura aumentó en el período analizado hasta alcanzar una presencia relevante, sobre todo entre las mujeres [8].

Por otro lado, según el índice global de paz (IGP) o también conocido como IPG índice de paz Global del 2015, los países más violentos de América del sur son Brasil en la posición 103, donde ha aumentado la corrupción y la agitación civil, Venezuela en el puesto 142, donde hay manifestaciones violentas y el país está construyendo un arsenal militar, y Colombia en la posición 146 por la cantidad de refugiados resultados del conflicto con las fuerzas Armas Revolucionarias de Colombia (FARC). Entre Venezuela y Colombia esta México en el puesto 144 en Norteamérica, mientras que en la región centroamericana y el Caribe figura como país más violento El Salvador en el número 123 de la lista global. Como podemos ver Colombia es el país más violento de todo el continente americano, según el IPG el cual compara a 162 naciones y que mide entre otros 23 indicadores, la seguridad interna, la participación del país en conflictos y el grado de militarización que tiene [9].

Figura 3. *Niveles de violencia de los países*

Fuente de página: Runrun.es

Según la última edición del informe de Carga Global de la Violencia Armada un promedio anual de 60.000 mujeres en el mundo fueron víctimas de homicidios, lo que representa el 16% de los homicidios intencionales. Una de cada diez muertes violentas registradas en el mundo ocurre en situaciones de conflicto o ataques terroristas, mientras 396.000 homicidios intencionales ocurren cada año [10].

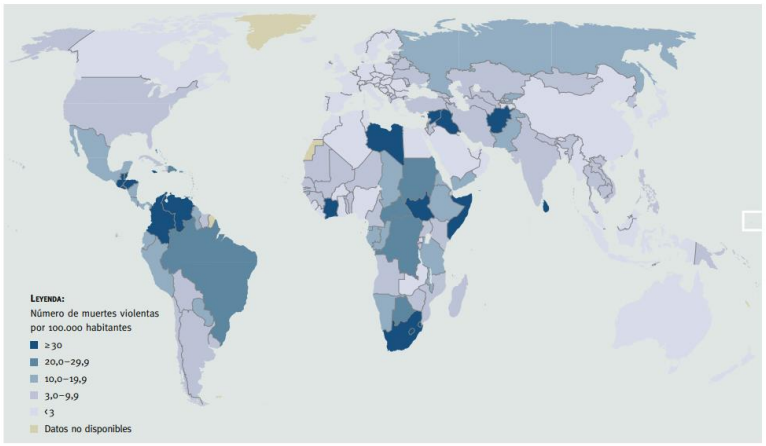


Figura 4. *Tasa promedio anual de muertes violentas por cada 100.000 habitantes 2007-2012*

Fuente: Base de datos CGVA 2015

Como podemos observar en el gráfico anterior Colombia en el 2012 era uno de los pocos países en el mundo que tenían una tasa promedio anual de muertes violentas por cada cien mil habitantes mayores a 30, un número bastante elevado tomando en cuenta los puntajes globales. Así mismo, Los 18 países con las tasas de muertes violentas más altas

albergan tan solo un 4% de la población mundial, pero son el escenario de aproximadamente el 24% de todas las muertes violentas en el mundo [11].

Un aspecto relevante que debe ser tomado en cuenta y que predomina mucho en la actualidad es la violencia contra la mujer, es uno de los grandes problemas que agobian al mundo. Según un estudio realizado por la Organización Mundial de la Salud (OMS) con datos de 2013 revela que los países del Sudeste Asiático son aquello en los que la violencia contra la mujer dentro de la pareja tiene una mayor prevalencia. El estudio dio como resultados que el 30% de las mujeres de todo el mundo han sufrido violencia física o sexual por parte de su pareja, un porcentaje que se eleva hasta el 35 % si se añade a personas distintas a la pareja [12]. En un estudio realizado en Ecatepec, Estado de México Se estimó que una de cada tres mujeres, reportaron eventos de violencia que es perpetrada por su pareja. Por tipología, la prevalencia de violencia psicológica fue de 32%, la violencia física se cuantificó en 19%, la violencia económica ocupó el tercer lugar con 14%, y, en menor proporción, se reportó la violencia sexual, con 8.5% [13].

Cada año, aproximadamente 66.000 mujeres son asesinadas en forma violenta en el mundo, lo que representa casi 17% del total de homicidios intencionales [10].

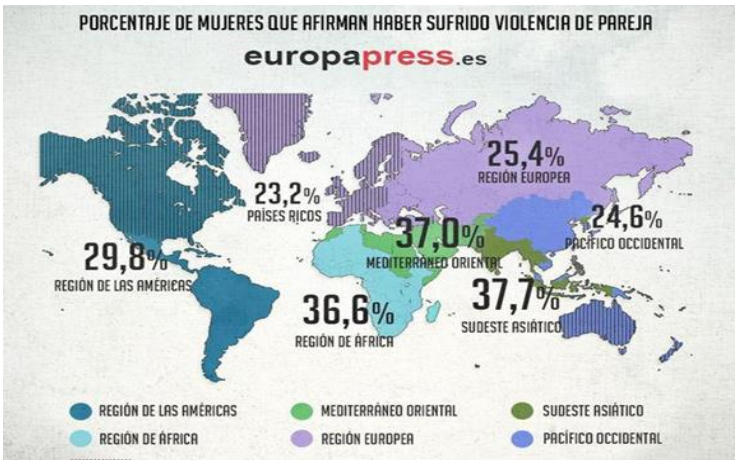


Figura 5. *Porcentaje de mujeres víctimas de violencia 2013*

Fuente: europapress.es

Podemos observar como la violencia ha hecho eco en nuestra sociedad, sin importar el país en donde te encuentres, la violencia se encuentra presente, de una mayor o menor forma. Aunque es difícil obtener estimaciones precisas, el costo de la violencia se traduce en billones de dólares estadounidenses en gastos anuales de atención médica a nivel mundial, y miles de millones en economías nacionales en términos de días perdidos en el trabajo, cumplimiento de la ley e inversión perdida [3].

1.2 Contexto Local

1.2.1 Colombia

Colombia, situada en América del Sur, tiene una superficie de 1.141.749 Km2, Siendo el vigésimo sexto país más grande del mundo. Con una población de 48.653.419 personas, se encuentra en la posición 27 de la tabla de población, compuesta por 195 países y presenta una moderada densidad de población, 43 habitantes por Km2. Su capital es Bogotá y su moneda Pesos colombianos.

La economía colombiana ocupa la posición 40 por volumen de PIB. Su deuda pública en enero de 2018 logró disminuir su endeudamiento externo con relación al Producto Interno Bruto (PIB), el monto incrementó en US$12 millones al pasar de US$124.363 millones en diciembre de 2017 hasta US$124.375 millones en enero de 2018 [14].

La violencia en Colombia como en otros países se puede presentar en cualquier lugar, ya sea en la calle, la casa, o el trabajo, etc. Puede ir desde una golpiza por parte del marido a su mujer hasta una masacre en la calle o simplemente una pelea entre vecinos.

Desde hace muchos años el país ha sido golpeado por la violencia generando pobreza, miedo, terror en las personas. Según el índice de paz global el cual es un indicador que mide el nivel de paz y la ausencia de violencia en un país y el cual es publicado por El Instituto para Economía y Paz nuestro país en el año 2017 ocupó el puesto 146 de 163 países evaluados.

Fecha	Índice de Paz Global	Ranking Paz Global
2017	2,777	146º
2016	2,764	147º
2015	2,720	146º
2014	2,701	150º
2013	2,634	147º
2012	2,640	143º
2011	2,697	140º
2010	2,791	136º
2009	2,625	129º
2008	2,614	126º

Figura 6. *Evolución de la posición de Colombia en el Índice de Paz Global*

Fuente: Expansión/ Datosmacro.com

De acuerdo a esto, podemos evidenciar como nos encontramos muy cerca de los últimos lugares, por lo que somos considerados un país violento. De igual forma, podemos ver cuánto ha empeorado en el tiempo, ya que en el año 2008 se encontraba en la posición 126 de 138 evaluados y en el año 2012 estaba en la posición 143 de 158 países evaluados,

lo cual indica un crecimiento significativo. En el año 2014 alcanzo la posición 150 de 162 países, la más alta en toda la historia del país.

La peor parte de los resultados de la medición, realizada desde el 2008 por el Instituto para la Economía y la Paz fundado por el empresario australiano Steve Killelea es que Colombia ocupa en la actualidad el noveno lugar entre los 10 países del mundo que más dinero de su PIB gastan en contener la violencia, según una fórmula que aplica el instituto [9].

La violencia, y de ella el homicidio, es la forma más extrema de resolución de los conflictos sociales entre las personas, en la ciudad de Medellín se reconoce como el principal problema social, económico, de salud pública y de seguridad ciudadana y afecta la calidad de vida de sus habitantes [15].

Figura 7. *Evolución de la tasa de mortalidad por homicidios en Medellín, 2000-2016*

Fuente: Medellín como vamos

En la ciudad de Medellín, el homicidio es la primera causa de mortalidad, siendo la principal causa de homicidios en los últimos 25 años el crimen organizado, pero con la diferencia de que en la actualidad las bandas son menos peligrosas que las de los años 90. Sin embargo, En dos décadas y media, Medellín logró salir de la lista de las 50 ciudades más violentas del mundo y pasó de ser la capital más peligrosa a ser modelo en reducción de homicidios [15].

Pero una de las partes más relevantes de todo esto es que este problema afecta en gran medida a la juventud, Se estima que 565 niños, adolescentes y jóvenes mueren diariamente en el mundo a causa de homicidio cometido por otros jóvenes o por adultos. En Colombia, en 2002, de las víctimas de homicidio en el país, el 34,2% eran jóvenes (edad=14-24 años). Para el 2004, la tasa de hombres jóvenes víctima de homicidio (edad=18-25 años) fue de 195 por 100.000 habitantes/ año, y dos terceras partes, tanto de las víctimas como de los detenidos por homicidio fueron hombres (edad=11-35 años), según datos de la Policía

Nacional. En Bogotá, el panorama no es muy diferente, en 2002 se presentaron 2.041 muertes a causa del homicidio. Los jóvenes (edad=15-25 años) fueron las víctimas en el 33,9% de los casos, y si se amplía a jóvenes entre 15 y 39 años, éstos representan el 75,9% de las víctimas [16].

En nuestro país hay un tipo de violencia que podemos ver cada día, esta es la violencia hacia la mujer. La ONU en 1994 define la violencia de pareja que se ejerce hacia las mujeres como cualquier acción, conducta u omisión que tenga la intención de menoscabar, o que ocasione daño físico, emocional o sexual e incluso la muerte, por parte del compañero íntimo [13].

La violencia sexual es una problemática frente a la cual cualquier niño, adolescente, o adulto puede estar expuesto. Es tan común y cercana que muchas veces para la sociedad es invisible, cotidiana o fuente de resignación e indignación [17]. Esta no solo afecta físicamente a la persona, sino también psicológicamente, Para la víctima siempre será una condición que la degrada y deshumaniza y que puede dejar consecuencias negativas que tendrá que afrontar de por vida [17]. Normalmente los casos denunciados no abarcar en lo más mínimo la cantidad real del fenómeno de la violencia sexual, y se estima que solo uno de cada 20 delitos sexuales es denunciado [17].

En nuestro país uno de los factores que aumentan los números de esta problemática es el conflicto armado, según el Registro Único Nacional de Víctimas, hay 12.740 mujeres afectadas por la violencia sexual en el conflicto, de estas 56º son niñas y adolescentes, y de las cuales 785 corresponden a cosos registrados durante los últimos dos años, lo cual equivale en promedio a una víctima cada día. Mientras en 2015 fueron atendidas 413 mujeres víctimas de este delito, hasta abril del año 2016 ya iban 139, relación que vista en el mismo contexto arroja también la conclusión de que cada día una mujer es agredida sexualmente señaló la Defensoría [18].

1.2.2 Cartagena

Cartagena de Indias, oficialmente Distrito Turístico y Cultural de Cartagena de Indias abreviado Cartagena de Indias, D. T. y C., es la capital del departamento de Bolívar, Colombia. Fue fundada el 1 de junio de 1533 por Pedro de Heredia y se encuentra localizada a orillas del mar Caribe.

A partir de su fundación en el siglo XVI y durante toda la época virreinal española, Cartagena de Indias fue uno de los puertos más importantes de América. De esta época procede la mayor parte de su patrimonio artístico y cultural. El 11 de noviembre de 1811, Cartagena se declaró independiente de España. Este día es fiesta nacional en Colombia y en la ciudad se celebra durante cuatro días conocidos como las "Fiestas de Independencia".

Desde sus inicios fue muy marcada por la violencia, ya que ha sido una ciudad principalmente asociada con la historia pirata, pues fue allí donde se presentaron numerosos ataques por parte de los piratas provenientes de Europa, que encontraron en la ciudad un

lugar adecuado para saquear en la época de administración española, lo que la hizo en su momento, convertirse en la ciudad más reforzada de América del Sur y el Caribe, llegando a estar casi tan reforzada como el mismo Golfo de México en su época. En la actualidad se mantiene su arquitectura virreinal.

Actualmente, la violencia en Cartagena ha hecho un eco enorme, ocasionando que muchas personas se encuentren padeciendo debido a este gran problema. Un claro ejemplo de esto se ve reflejado en la violencia de la que ha sido parte la mujer desde hace muchos años, ya que según cifras obtenidas, la violencia contra la mujer ha venido incrementando presentando diversas facetas que van desde la discriminación y el menosprecio hasta la agresión física, sexual, verbal o psicológica y el asesinato, manifestándose en diversos ámbitos de la vida social , laboral y política, entre los que se encuentran la propia familia, la escuela, la Iglesia, el Estado, entre muchas otras.

Al indagar por la frecuencia con que es empleada la violencia contra las mujeres y por el principal tipo de violencia que es usado en contra de ellas, se encuentra una percepción generalizada en todos los estratos, entre hombres y mujeres, en todas las localidades y en todos los rangos de edad, de que la violencia se presenta con una frecuencia muy alta. En una escala de 1 a 5, siendo 1 nada frecuente y 5 muy frecuente, el promedio de calificación fue 4,0 [19].

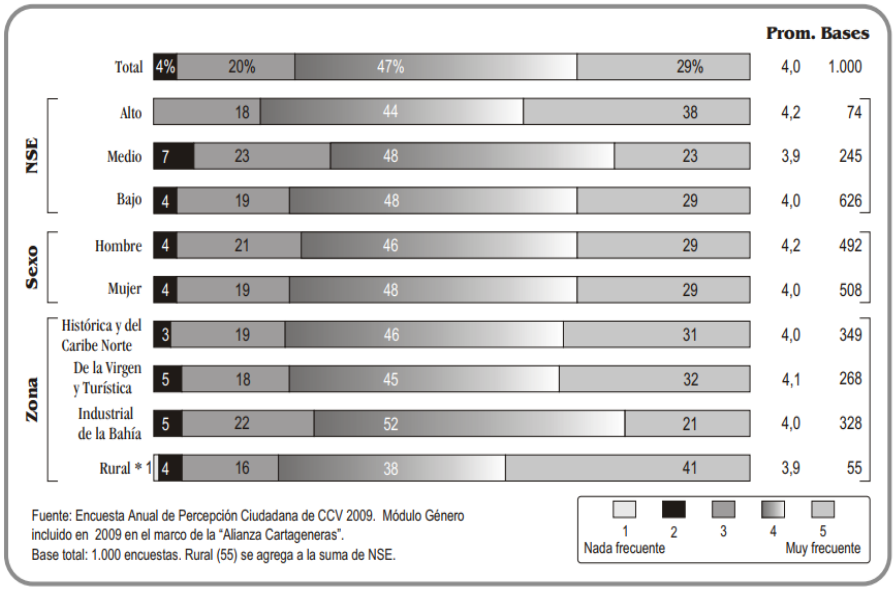


Figura 8. *Frecuencia con la que es empleada la violencia contra las mujeres*

De este modo, se puede evidenciar que el principal tipo de violencia que perciben las mujeres en Cartagena es la física (58.0%), seguida por la verbal.

Por otra parte, los casos de homicidios han aumentado en comparación con el año 2016 debido a que en este año había una tasa de 238 homicidios mientras que en el 2017 ocurrieron 252, así pues, esto se puede evidenciar en la siguiente gráfica.

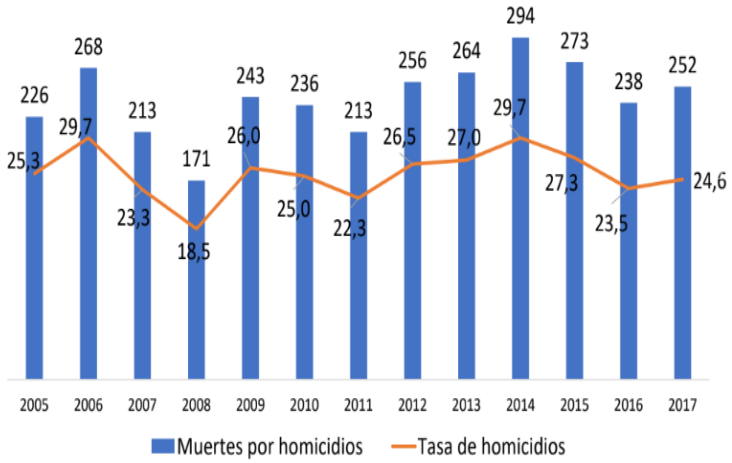


Figura 9. *Homicidios en Cartagena entre 2005-2017*

Por otro lado, hablamos de delito sexual del cual pueden existir múltiples definiciones, pero vale la pena referirnos a la que introduce el Dr. Reyes Echandía en su obra "Criminología", en la que expone acerca del surgimiento del "Delito Sexual" en los siguientes términos: "... La ulterior delictuosidad de esa conducta está vinculada a la existencia de normas de cultura que, en un momento dado y dentro de una determinada sociedad repudiaron ciertas manifestaciones eróticas por considerarlas contrarias a la moral pública o violatorias del derecho a disponer del propio cuerpo para fines sexuales. En la medida en que la sociedad valora negativamente tales hechos, estos se elevan a la categoría de prohibiciones, de tabúes, que van recibiendo el respaldo jurídico de la ley; he ahí el origen de la connotación jurídico penal de los llamados delitos sexuales". [20]

Así pues, en la siguiente gráfica podemos observar como durante el periodo comprendido entre 2008 y 2016 se han registrados 3836 casos con presunción de delito sexual, según registros del Instituto de Medicina legal y Ciencias forenses, un promedio de 426.2 casos por año. La cifra más baja se registró en 2010 con 339 dictámenes relacionados presumiblemente con delitos sexuales. Por el contrario, el año 2016 presenta la mayor cifra del periodo con 503. [21]

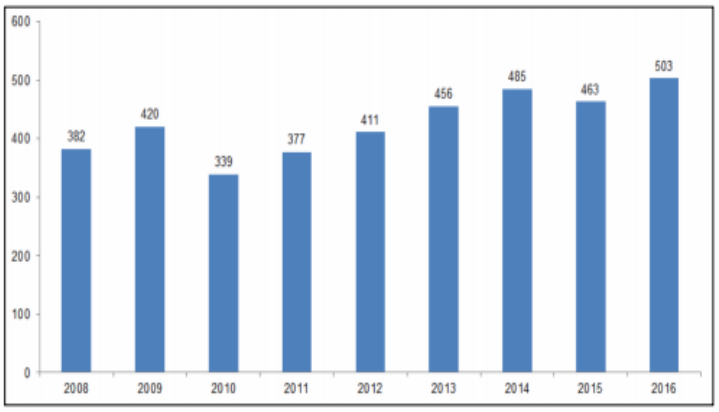


Figura 10. *Casos con presunción de delitos entre 2008-2016.*

Fuente: COSED

En este sentido, los 20 barrios que se registran como de mayor frecuencia de presuntos casos de delitos sexuales tienen perfiles socioeconómicos bastante homogéneos, su nivel de estratificación es mayormente de estrato 1. Tienen igualmente similitudes en cuanto a altas frecuencias de homicidios recurrentemente mes por mes tales como los barrios Olaya Herrera, El Pozón y Nelson Mandela. Los indicadores de violencia intrafamiliar son también más altos en los barrios mencionados. Los informes de la Secretaria de Educación y encuestas tales como las contratadas por CCV muestran a estos mismos barrios como zonas con altos niveles de jóvenes en edad de estudiar que están por fuera del sistema educativo distrital.

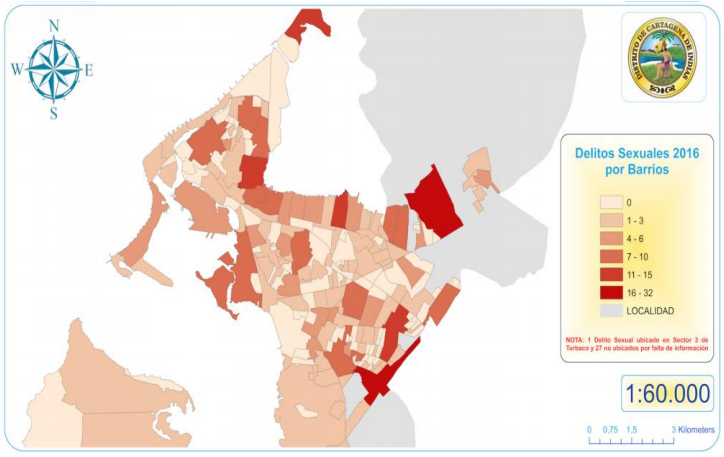


Figura 11. *Delitos sexuales por barrios en el año 2016*

Fuente: COSED

Podemos decir que al igual que los delitos sexuales, la violencia en Cartagena ha estado en aumento, causando un grave daño a la sociedad en general. Esto genera consecuencias directas tales como capacidad para generar miedo, incertidumbre, inseguridad e impotencia, la guerra y las violencias destruyen el tejido social, desconfiguran los valores sociales, desestructuran psicológica y afectivamente a las personas y a los grupos y crean un clima de desesperanza y tristeza.

En la actualidad la violencia tanto en Colombia como a nivel mundial ha generado grandes secuelas en la sociedad, dejando un legado que se reproduce a sí mismo a medida que las nuevas generaciones aprenden de la violencia de las anteriores, donde las víctimas aprenden de sus agresores y se permite que perduren las condiciones sociales que favorecen la violencia. Es así como a escala mundial y nacional existen datos, aunque muy escasos de los casos registrados de violencia, como lo son homicidios, suicidios, conflicto armado, entre otros. En este sentido, se estima que 1,6 millones de personas murieron en forma violenta en el 2000, es decir 28,8 por 100.000 habitantes. Casi la mitad de estas muertes fueron suicidios, cerca de la tercera parte fueron homicidios y una quinta parte por eventos relacionados con conflictos armados. Las cifras varían considerablemente entre y dentro de los países [22].

Dado que la violencia no se puede atribuir a una sola causa, pues éstas son complejas y ocurren en diferentes niveles, se creó una clasificación para representar dicha complejidad, la cual nos muestra una tipología de violencia, así como la naturaleza de la misma.

Así pues, el indicador de violencia con mayor magnitud en Colombia y por ello el más preocupante, es el de los homicidios. En los últimos 30 años del siglo pasado el país casi llego a un total de medio millón de homicidios. En la última década se registró un promedio anual de 25.000 homicidios. En el año 2000 el país supero dicho promedio y alcanzo un total de 25.655 homicidios, para una tasa de 61 homicidios por 100.000 habitantes, según los datos del centro nacional de referencia sobre violencia del instituto de medicina legal. En este sentido, desde hace tres décadas la violencia, ocupaba el noveno lugar entre las causas de muerte en Colombia. En los años setenta pasó a ocupar el cuarto lugar y desde los años ochenta se han ubicado en el primero, tomando cada vez mayor ventaja en relación a las demás causas de muerte de los colombianos. Pero los verdaderos indicadores del incremento de la violencia en Colombia son el número y la tasa de homicidios. En la última década ha habido en el país más de 230 000 homicidios, cifra que supera a los 200 000 estimados en Colombia en los años cuarenta y cincuenta, durante el período de la Violencia. En 1994 los homicidios constituyeron 70% de todas las muertes violentas registradas en el país, según datos suministrados por el Instituto Nacional de Medicina Legal y Ciencias Forenses, así pues, los últimos reportes hechos por esta entidad reportan 23.072 lesiones fatales, 255.161 no fatales y 6.712 personas desaparecidas. En cuanto a las primeras, en Colombia se registraron 10.870 víctimas de homicidios, 2.402 casos de suicidios, 6.499 muertes en accidentes de transporte y 3.301 muertes accidentales.

Por otra parte, en cuanto a las lesiones no fatales, 113.470 personas fueron víctimas de violencia interpersonal, 23.418 de presunto delito sexual, 76.785 de violencia intrafamiliar,

38.097 de lesiones en accidentes de transporte y 3.391 de lesiones accidentales. Cabe aclarar que la Entidad clasifica la violencia intrafamiliar en violencia a niños, niñas y adolescentes (10.425 víctimas), violencia de pareja (49.423), violencia entre otros familiares, (15.015) y violencia al adulto mayor (1.922) [23].

Por esta razón, la situación de violencia ha llevado, además, a modificaciones importante en nuestro estilo de vida. Hemos modificado los horarios de ciertas actividades sociales, hemos reducido los espacios de movilización y recreación, y un número cada vez mayor de personas se ha visto forzada a recurrir a mecanismos de seguridad por la pérdida de seguridad y el incremento de tenciones [24].

1.3 Preguntas Problemas

1.3.1 General

¿Se puede crear una metodología que permita analizar mediante el uso de machine learning la violencia en el departamento de Bolívar**?**

1.3.2 Especificas

De la pregunta problema surgen los siguientes interrogantes:

- ¿Existen diferencias en los niveles de violencia entre las zonas urbanas y rurales?
- ¿Cuál es el día de la semana donde se presenta la mayor cantidad de homicidios en el país y cuál es el sexo y grupo de edad que predomina?
- ¿Se puede encontrar diferentes perfiles de homicidios en el departamento de Bolívar?
- ¿Cómo definir perfiles de homicidio en el departamento de Bolívar?

1.4 Objetivos

1.4.1 Objetivo general

Crear una metodología que permita caracterizar la violencia en el departamento de Bolívar mediante la aplicación de machine learning.

1.4.2 Objetivos específicos

- Estimar si el nivel de violencia en el departamento difiere del entorno en el que se encuentre la población.
- Analizar Mediante el uso de machine learning la violencia presente en Cartagena.
- Caracterizar perfiles de homicidios en el departamento de Bolívar.
- Establecer una metodología para definir el perfil de homicidios en una ciudad o departamento.

1.5 Limitaciones

- Se contará solamente con datos recogidos en los años 2015.
- Solamente se utilizarán datos disponibles en la página datos.gov

1.6 Justificación

El incremento en los últimos años de homicidios en Colombia, especialmente en Cartagena ha hecho que muchos investigadores puedan buscar las causas que ha provocado esto, es por ello que ya se han realizado investigaciones con el fin de crear patrones de homicidios que ayuden posteriormente a crear estrategias para poder disminuir esta problemática, al igual que estudiar la causa de las mismas. Es por ello que el presente trabajo investigativo se enfocará en mostrar el uso de machine learning para crear dichos patrones, tendencias o hipótesis que ayuden a esclarecer que variables son las que influyen significativamente en la ocurrencia de estos hechos criminales. Así pues, es fundamental el uso de algún programa que nos ayude a analizar estos patrones, debido a la gran cantidad de datos que se maneja y a las relaciones complejas que existen entre ellas y que por medio de una exploración tradicional no se podría logara los resultados que estamos buscando.

En este sentido, la metodología busca el desarrollo que ayuden a interpretar mejor estos patrones criminales encaminados a reducir el número de homicidios en la Ciudad y los posibles efectos que esto produzca, así como la implementación de programas que ayuden a las personas afectadas por este flagelo. De igual forma, la metodología es de gran ayuda para poder investigar que variables son las que están más asociadas a las victimas tales como edad, estado civil, escolaridad, genero, los cuales tienen una gran incidencia en estos patrones criminales.

Finalmente, con este estudio se busca crear una metodología para investigar los homicidios en la ciudad de Cartagena, haciendo uso de la información recolectada, que permita identificar patrones y variables asociadas al comportamiento social de los ciudadanos, que pueden causar ser víctimas de estos actos criminales.

2. Marco teórico

Debido a que existen diversas investigaciones sobre el tema que se está desarrollando, se analizará la forma como diferentes actores han abordado el tema, que han realizado, que herramientas han utilizado, que problema han encontrado y como lo resolvieron.

Tabla 1. *Estado del arte*

Paper	Año	Autor	¿Qué hizo?	Técnica
Aplicación de la minería de datos para analizar los diferentes tipos de lesiones registrados en la división médico legal-distrito fiscal de Piura[25]	2015	Auriestela del pilar Vicente Llacsahuache	Mediante el desarrollo de modelos de Minería de datos se busca ayudar a las autoridades en la toma de decisiones sobre los hechos de violencia que causen lesiones físicas en las personas y se presenten en la región, el propósito de la implementación de dichos modelos es crear conocimiento que ayude a innovar los existentes. Así como mejorar la calidad de los servicios que presta la entidad.	Minería de datos utilizando los algoritmos de Bayes Naïve, Series de Tiempo y Arboles de Decisión
Modelo recursivo de reacción violenta en parejas[26]	2012	J. Moral de la Rubia y F. López	Se empleó una muestra no probabilística. Primero se realizó un cuestionario de violencia en la pareja Se contrastaron las diferencias de medias por la prueba t de Student y Las correlaciones se calcularon por el coeficiente producto-momento de Pearson. Para el contraste de modelos de relación entre la violencia recibida y ejercida se usó modelamiento de ecuaciones estructurales por el método de Mínimos Cuadrados Generalizados.	Cuestionarios
Violencia de pareja: tipo y riesgos en usuarias de atención	2017	Rovira Alcocer Gloria, Vital Hernández Omar,	Identificar la prevalencia y el tipo de violencia de pareja en mujeres usuarias de una unidad de atención	Mediciones principales Mediante

primaria de salud en Cancún, Quintana Roo, México[27]		Pat Espadas Fany Guadalupe, Sandoval-Jurado Luis y Jiménez-Báez María Valeria	primaria y estimar los riesgos para cada tipo de violencia.	escala
Violencia de Pareja en Mujeres: Prevalencia y Factores Asociados[13]	2015	Jaen Cortés Claudia Ivethe, Aragón Sofía Rivera, Amorin de Castro Elga Filipa y Rivera Rivera Leonor	Estimar la prevalencia y algunos factores asociados a la violencia de pareja en mujeres de Ecatepec, Estado de México.	Encuestas
Mujeres víctimas de violencia de género en centros de acogida: características sociodemográficas y del maltrato[28]	2017	Fernández-González Liria, Calvete Esther y Orue Izaskun	.Este estudio tuvo como objetivo principal describir las características sociodemográficas y del maltrato sufrido por este colectivo, así como analizar los cambios en las variables de estudio a lo largo de los 10 últimos años.	Estudio descriptivo mediante la técnica de análisis de documentos.
Análisis empírico de la relación entre la actividad económica y la violencia homicida en Colombia [29]	2011	Carranza Romero Juan Esteban, Dueñas Herrera Ximena, González Espitia Carlos Giovanni	El objetivo de este artículo es examinar la relación causal entre asesinatos y actividad económica en Colombia durante las décadas recientes. El análisis saca provecho de que la actividad económica en el resto de América Latina está altamente correlacionada con la actividad económica en Colombia, pero no es afectada directamente por la violencia homicida en Colombia.	Análisis de series de tiempo.
La violencia sexual como genocidio Memoria de las mujeres mayas sobrevivientes de violación sexual durante el conflicto armado en Guatemala [30]	2016	Fulchiron Amandine	Se analizó el uso sistemático y masivo de la violación sexual contra las mujeres mayas dentro del marco de la política contrainsurgente en Guatemala, nombrándolo y denunciándolo como feminicidio y genocidio; se evidenció cómo la violación sexual fue utilizada por el Estado para destruir la continuidad biológica, social y cultural del pueblo maya a través del cuerpo de las mujeres.	Epistemología feminista articulada con la de la cosmovisión maya.
Mecanismos de lesión en actos de		Arcaute-Velazquez Fernando Federico,	Los patrones lesionales que se encuentran en las víctimas de actos	Mecanismos de

violencia extrema [31]		García-Núñez Luis Manuel, Noyola-Villalobos Héctor Faustino, Espinoza-Mercado Fernando y Rodríguez-Vega Carlos Eynar	de violencia extrema son muy complejos y obedecen a distintos mecanismos de lesión de alta transmisión de energía. En este manuscrito exponemos los conceptos básicos en cinemática del trauma que rigen el abordaje clínico de las víctimas de actos de violencia extrema, en espera de que el clínico aumente su armamentario teórico, reflejándolo en la obtención de mejores índices pronósticos.	lesión
Violencia física marital en Barranquilla (Colombia): prevalencia y factores de riesgo [32]	2003	Tuesca R., Borda M.	Determinar la prevalencia de maltrato físico marital en mujeres en edad fértil que viven con su pareja, así como identificar factores personales, socioeconómicos y de función familiar que se relacionen con el maltrato.	Entrevista personal en el hogar a partir de un cuestionario estructurado.
Acción COST Femicide Across Europe, un espacio de cooperación trasnacional para el estudio y el abordaje del feminicidio en Europa [33]	2016	Sanz-Barbero Belén Otero-García Laura Boira Santiago Marcuello Chaime Vives Cases Carmen	En esta nota de campo se describen los principales objetivos de la Acción COST Femicide across Europe la cual fue implementada por el Programa de la Unión Europea «Redes de Cooperación Europea en Ciencia y Tecnología», los grupos de expertos y expertas que lo conforman, y los resultados obtenidos a medio plazo con dicha experiencia.	
La violencia en la escena del crimen en homicidios en la pareja [34]	2016	Company Alba, Soria Miguel Ángel.	La presente investigación aborda el tipo de violencia ejecutada, instrumental vs expresiva, en la EC (escena del crimen) de asesinatos, homicidios o tentativas y sus diferencias en función del sexo del homicida.	Protocolo de Análisis del Crimen Violento en Homicidios Familiares (PACVHF) Es un cuestionario
Analizando la violencia después del conflicto: el caso de Guatemala en un estudio sub-	2014	Tobón Aguirre Katherine	Conceptualiza la violencia después del conflicto, desarrolla su definición, los factores explicativos y construye una tipología de ocho formas de violencia. El estudio de	Se propone una tipología de ocho formas de violencia. "regionalizació

nacional [35]			caso de Guatemala permite identificar patrones relacionados con cada forma de violencia, en un análisis a nivel sub-nacional.	n" de la violencia después del conflicto.
La trinidad perversa de la que huyen las fugitivas centroamericanas [36]	2017	Varela Huerta Amarela	Se expone el contexto en el cual se desarrollan los mecanismos que las expulsan de sus países de origen. Se vierten argumentos que explican que su huida de Centroamérica es una "fuga" estratégica de resistencia contra una triple violencia: de Estado, de mercado y patriarcal.	sustentado en datos estadísticos, en lecturas de otras feministas, en los discursos publicados de organizaciones sociales y en las propias preguntas
Autoestima, violencia de pareja y conducta sexual en mujeres indígenas [37]	2017	Nava-Navarro V., Onofre-Rodríguez D., Báez-Hernández F.	Se busca conocer la relación de la autoestima, violencia de pareja y conducta sexual en mujeres indígenas.	investigación de tipo descriptivo-correlacional, con diseño transversal
Proyecto de atención integral a víctimas de violencia sexual en el departamento de Escuintla, Guatemala [38]	2012	Agustí Cristina, Sabidó Meritxell, Guzmán Karla, Pedroza María Isabel, Casabona Jordi.	Se implementó un proyecto de atención integral a las víctimas de violencia sexual en seis municipios del departamento de Escuintla, Guatemala. Estas víctimas recibieron atención médica y psicológica.	se entrevistaba con la víctima, mediante un cuestionario, y se le citaba para la próxima visita
La violencia interpersonal en España a través del Conjunto Mínimo Básico de Datos [39]	2018	Gil-Borrelli Christian Carlo, Latasa Zamalloa Pello, Martín Ríos María Dolores, Rodríguez Arenas M. Ángeles.	Busca Describir la epidemiología de la violencia interpersonal en España.	Estudio descriptivo de los casos de pacientes con diagnóstico secundario de agresión registrados.
Ámbitos y formas de violencia contra mujeres y niñas: Evidencias a partir de las encuestas [40]	2014	Frías Sonia M.	Se documenta algunas de las múltiples expresiones de violencia en contra de mujeres y niñas en México. La falta de análisis con muestras representativas sobre las distintas formas de violencia en contra de ellas dificulta la	Encuestas

			visibilidad de la problemática y, asimismo, dificulta que se incorpore el derecho de las mujeres a vivir una vida libre de violencia.	
La violencia contra las mujeres en la región occidente, México: Entre la inoperancia institucional y el conservadurismo social [41]	2014	Ochoa Ávalos María Candelaria, Reillo Fernando Calonge.	El presente artículo sintetiza los resultados más importantes de una investigación sobre la violencia contra las mujeres en la región occidente de México que forma parte del Estudio Nacional sobre las Fuentes, Orígenes y Factores que Producen y Reproducen la Violencia contra las Mujeres	estrategia metodológica basada en la triangulación de fuentes
Miedo, conformidad y silencio. La violencia en las relaciones de pareja en áreas rurales de Ecuador [42]	2016	Boira Santiago, Carbajosa Pablo, Méndez Raquel.	En este estudio se analizan los factores y dinámicas involucradas en la violencia dentro de la pareja en Latinoamérica en un contexto de pequeñas comunidades rurales.	Estudio cualitativo en comunidades rurales de la provincia de Imbabura, en Ecuador.
Identificación de patrones característicos de la población carcelaria mediante minería de datos [43]	2010	Gutiérrez Rüegg, P., Merlino, H., Rancan, C., Procopio, C., Rodríguez, D., Britos, P., García-Martínez, R.	Dado que en las entidades penitenciarias se estaba presentando problemas de hacinamiento, sobrepoblación, violencia, entre otros, Se realizó minería de datos con el fin de prevenir delitos.	Minería de datos

3. Diseño de la Metodología de la investigación

3.1 Marco metodológico

En el presente trabajo de investigación se puede observar como mediante el uso de machine learning se llega a una metodología para la caracterización de la violencia en el departamento de Bolívar. Para el desarrollo de este trabajo investigativo fue necesario principalmente información suministrada por el gobierno a través de la página datos.gov. El eje principal de la investigación se basa en modelos de aprendizaje supervisado y no supervisado de datos, con los cuales se busca determinar que variables son las más influyentes a la hora de poder predecir porque ocurre este tipo de violencia en Colombia. Así pues, la investigación se centró en el desarrollo de dos etapas: la reducción de todos los datos obtenidos mediante el uso de una técnica de aprendizaje no supervisado, dio como resultado dos grandes grupos significativos en los que se identificaron características

semejantes que no se podían ver con facilidad. Como valor agregado implementaron varias técnicas de aprendizaje supervisado tales como knn, arboles de decisión con los cuales se busca llegar al mejor resultado forest que dé respuesta a nuestros interrogantes.

3.2 Tipo de investigación

Se determinó que para este estudio investigativo se usaran modelos con los cuales se pueda predecir y explicar de manera objetiva la problemática que se presenta en Colombia debido a las diferentes formas de violencia que en ella existe. Así pues, dado que nuestros datos tienen un enfoque cuantitativo, se realizará la recolección de datos para probar una hipótesis, con base a la medición numérica y el análisis estadístico de datos, para establecer patrones de comportamientos y probar teorías.

3.3 Metodología de la investigación

Esta investigación tiene como finalidad analizar una problemática social desde un punto de vista cuantitativo, apoyado en el estudio y análisis de la base de datos suministrada por el gobierno. Así pues, se parte de una teoría o planteamiento del problema, basado en la observación de los datos relacionados con los actos de violencia presentes en Colombia, para posteriormente, establecer una hipótesis, a través de la inferencia obtenida de los datos, la cual finalmente será probada haciendo uso de las herramientas propuestas. Para esto, se implementará la técnica de machine learning, que nos permitirá extraer información sustancial del conjunto de datos con el fin de transformarla en una estructura comprensible. Así pues, El eje principal de la investigación se basa en modelos de aprendizaje supervisado y no supervisado de datos, los cuales fueron un clustering, la técnica de K vecinos más cercanos y árboles de decisión. Así pues, para la realización de este análisis se deben realizar los siguientes pasos:

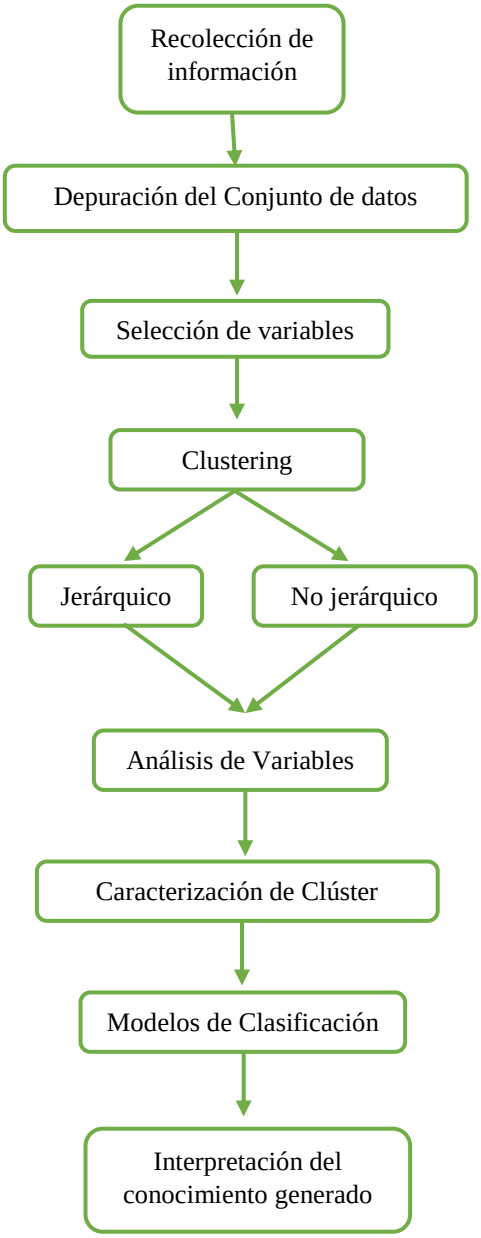


Figura 12. Metodología para el análisis de los homicidios basada en Machine Learning

El objetivo de esta metodología es lograr la obtención de perfiles característicos de una base de datos que trate sobre actos que generen violencia en un municipio, departamento, ciudad o incluso el país completo. Y con la creación de los diferentes perfiles permitirle al investigador entender un poco más el sistema y generar mejores estrategias para combatir este problema. En la figura 13 se identifican las fases propuestas de la metodología para el análisis de los homicidios basada en Machine Learning.

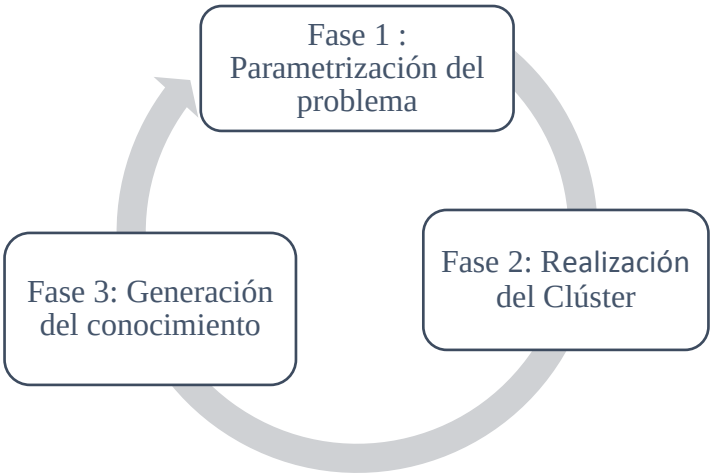


Figura 13. Fases de la metodología para el análisis de los homicidios basada en Machine Learning

1) Fase 1: Parametrización del problema

Cuando se hace referencia a la parametrizaciòn del problema nos enfocamos en los primeros pasos a seguir en el modelo, es por ello que para que ocurra esto se debe realizar lo siguiente:

I. **Recolección de información:** Consiste en la búsqueda y obtención de la información que se desea analizar, esta debe ser clara y real para poder realizar una investigación correcta. Lo recomendado es que la información se obtenga de páginas reconocidas que manejen datos de calidad, que se puedan descargar en archivos de Excel en CSV o CSV2 o incluso en archivos de textos. Dependiendo del formato en que sea descargado se utilizara un comando diferente para la lectura de los mismos. Por ejemplo, En R los comandos para CSV es read.csv y para los de texto es read.delim. Se recomienda realizar un resumen de la base de datos y así comprender un poco que hay en ella.

II. **Depuración del conjunto de datos**: Luego de tener toda la información recolectada y cargada en el software se procede a limpiar los datos, es decir, se deben eliminar aquellos datos que contengan valores faltantes como espacios en blancos, NA (Not Available), NAN (Not A Number), guiones o X (equis). El objetivo de esta etapa es detectar y eliminar todas las observaciones que contengan faltantes. De igual forma, se deben eliminar todas aquellas columnas que no le agregan nada al modelo, por ejemplo, cuando existen columnas con categorías que se encuentren muy repetidas, es decir, que estén en la mayor parte de los datos. Un ejemplo de esto se da cuando se tiene la columna nacionalidad, si el 95 % de las observaciones tienen la misma nacionalidad es recomendable eliminarla ya que no

le aporta nada significativo y diferenciador al algoritmo. Por otra parte, las variables que contengan demasiadas categorías como podrían ser la edad de una persona, se recomiendan eliminarlas o utilizarlas para crear nuevas variables.

III. **Selección de variables:** En esta etapa se seleccionan aquellas variables que aportan algo significativo al trabajo investigativo y se eliminan aquellas que no se consideren relevantes para su posterior uso en el algoritmo. Se recomiendan seleccionar una cantidad media de variables dependiendo del objetivo que se quiere alcanzar para obtener un modelo sin problemas y con estimaciones precisas, ya que si se escogen muchas será muy específico, se convertirá en un modelo sobre especificado y las estimaciones del modelo serán poco precisas, en cambio sí se seleccionan muy pocas variables se puede obtener un modelo mal detallado y producir estimaciones sesgadas. Se debe recordar que los modelos buscan simplificar la realidad, no reproducirla [44]. Por otro lado, las variables seleccionadas dependen mucho de la percepción del investigador y de cuales el considera que son más relevantes, sin embargo, se recomienda realizar el algoritmo principalmente con las variables seleccionadas y luego realizar diferentes pruebas añadiendo y quitando variables para así saber cuál de estas brinda un mejor resultado.

Fase 2: Clúster

Esta es la etapa principal o la base de todo el estudio. Así pues, El clustering es una técnica de agrupamiento que añade a todas las observaciones que contienen variables parecidas en un mismo grupo. Se enfoca principalmente en una alta similitud entre los miembros de un mismo grupo o clúster, y una baja entre un grupo con otro. Antes de iniciar, en el caso de que las variables sean numéricas se debe realizar una normalización de las escalas de las variables debido a que es probable que midan cosas distintas. Por ejemplo, pueden medir edad con número de siniestros ocurridos a un grupo de pasajeros, entre otros.

I. **Clustering jerárquico y no jerárquico:**
En este sentido, se inicia con el clúster jerárquico, el cual busca la unión de cada una de las observaciones basándose en distancias. Este algoritmo trabaja únicamente con valores numéricos, así que en el caso que se tenga valores categóricos es necesario realizar una dumificación, el cual es un proceso que convierte una variable categórica en numérica, al tener nuestra base de datos numérico se procede a realizar una matriz de distancia la cual se debe aplicar al algoritmo de clustering jerárquico. Al finalizar se recomienda crear un dendograma para apreciar mejor las uniones creadas por el algoritmo.
Por otro lado, El clúster no jerárquico a diferencia del anterior se le debe introducir el número de grupos que quieres que se formen, por ello es necesario que se realice primero el clúster jerárquico para poder conocer este valor, sin embargo, para estar

más seguro se recomienda realizar un análisis de silueta. Este análisis hace una evaluación de un número de k establecidos por el investigador e indica cual k da el mejor resultado de agrupamiento, maximizando las diferencias dentro de cada grupo y minimizando las diferencias internas dentro de cada clúster.

Así pues, se utilizara como algoritmo de clúster no jerárquico el kmeans. Los criterios de ajustes en el kmeans se basan en los conceptos de sumas de cuadrados entre grupos (betweens) y dentro de grupos (tot.withinss) [45]. Luego de obtener la gráfica, es necesario centrarse en los codos que se forman en ella para así seleccionar el k óptimo.

Luego de realizar la gráfica tot.withinss se procede con el kmeans con ambos números, en este el criterio de selección cambia ya que se escogen los que presentan la mejor división de clúster. Por último, se deben colocar las etiquetas obtenidas en el dataset inicial.

II. **Análisis de variables:** Al haber realizado el clustering o agrupamiento se procede a realizar un análisis de cada una de las variables de la base de datos para poder extraer información útil. Se extrae primeramente información general y después se debe extraer información particular de cada clúster obtenido. Esto será nuestra base para la caracterización de los clúster.

III. **Caracterización de clúster:** Esta etapa es el centro del modelo, debido a que se logra obtener perfiles característicos del tema abordado. Así pues, para la obtención de cada una de las características se basara en los valores que más se repiten dentro de cada variable en cada clúster. Esto arroja unas categorías dentro de cada una de las columnas del dataset para cada clúster, con las cuales se pueden identificar y lograr definir los perfiles.

En la figura 14 se puede evidenciar todo el proceso que se realiza para la realización del clustering, en el cual se realiza el clúster jerárquico y no jerárquico y termina con su caracterización.

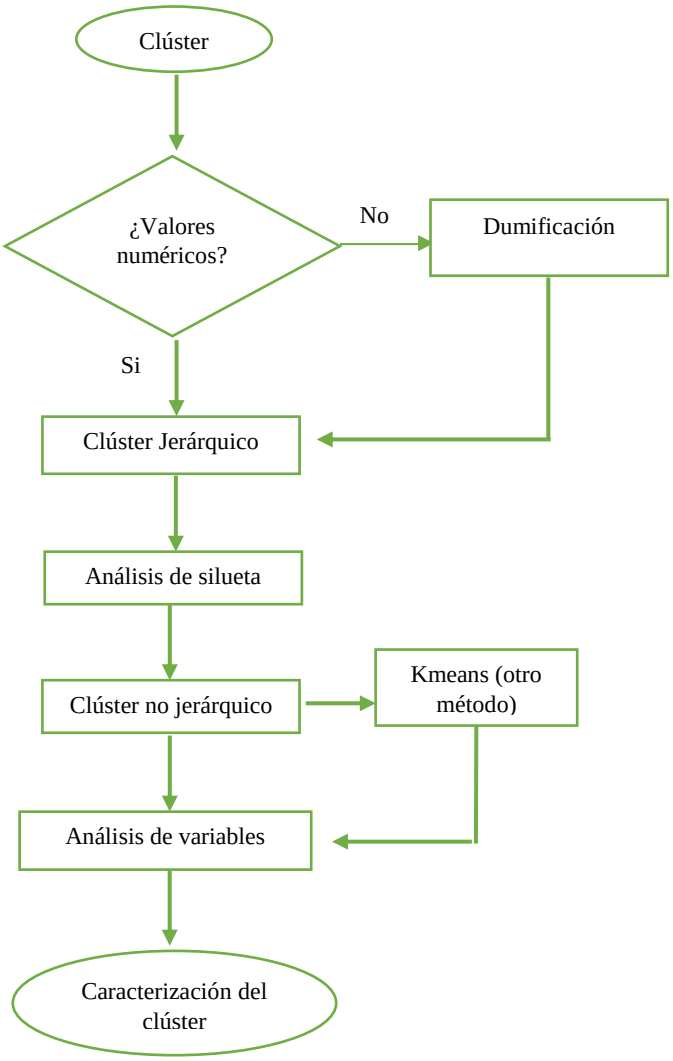


Figura 14. Metodología para la realización del clustering.

Fase 3: Generación del conocimiento

I. **Modelos de clasificación:** Una vez creado los perfiles se recomienda realizar un método de clasificación para saber si el algoritmo logra identificar un grupo de observaciones previamente etiquetadas dentro del grupo al que pertenecen, basándose en cada de una de las características de las variables del dataframe. Se recomienda usar dos métodos para probar su precisión de acierto, estos pueden ser el knn y el árbol de decisión. Así pues, ambos métodos nos ayudan a alcanzar nuestro objetivo, sin embargo, su principal diferencia radica en que el knn solo trabaja con valores numéricos mientras que el árbol de decisión puede trabajar tanto con valores numéricos como categóricos. Cabe resaltar que para trabajar con knn es necesario dumificar primero debido a que este no trabaja con datos categóricos.

Por otro lado, cuando se obtienen el número de aciertos de cada modelo para los grupos establecidos, entre mayor sea la cantidad de aciertos, mayor es el éxito del proceso de agrupamiento, lo cual me indica que el modelo de clasificación si logra diferenciar las observaciones de ambos clúster permitiendo que las diferencias entre observaciones de los grupos sean significativas.

I. **Interpretación del conocimiento generado:** Luego de haber realizado la caracterización esta se debe analizar y buscar relaciones entre cada una de las categorías seleccionadas para cada grupo, identificar el por qué se dan en cada situación y plantear un nombre que caracterice al perfil creado con el algoritmo.

4. Operacionalización de los objetivos

A continuación, en la tabla 2 se presenta la matriz de consistencia de este proyecto, con la cual se busca mostrar la operacionalización de los objetivos del trabajo.

Tabla 2. *Matriz de consistencia*

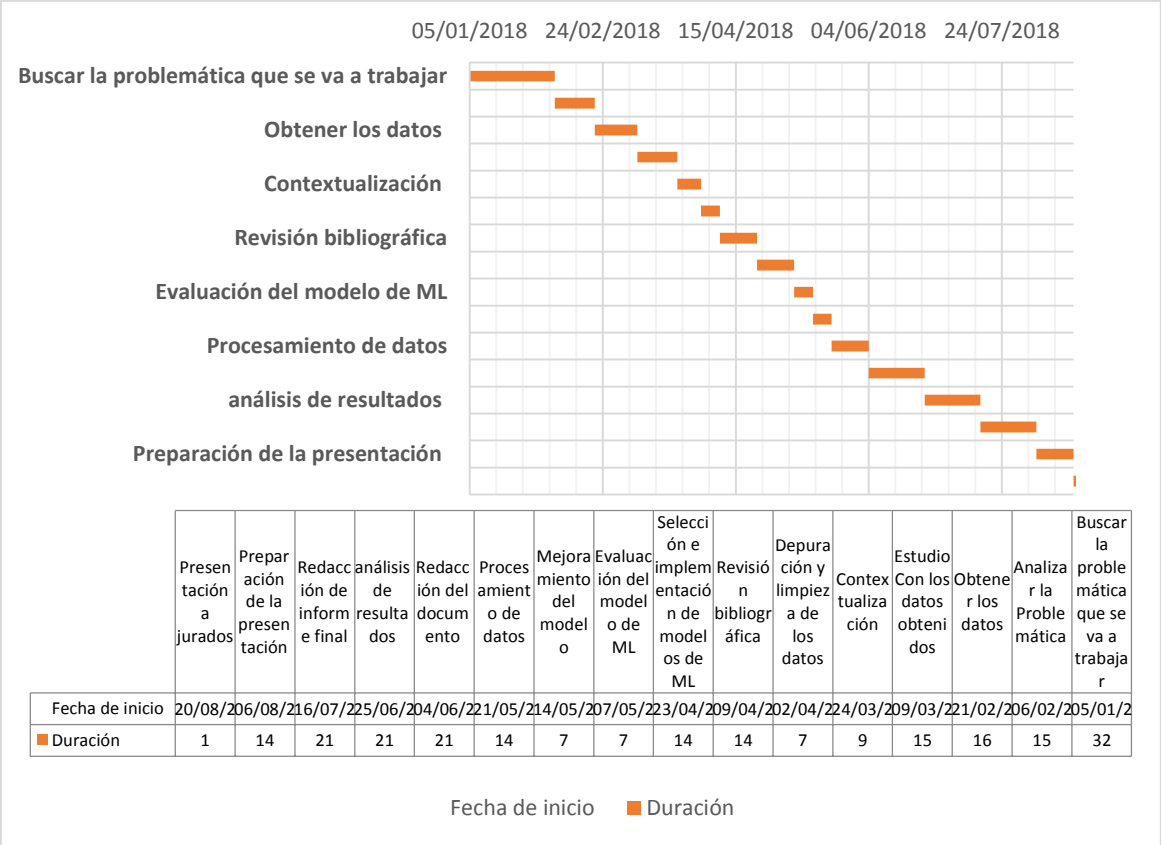
PROBLEMA	OBJETIVOS	HIPÓTESIS	MÉTODOS
Problema General	**Objetivo general**		
¿Se puede crear una metodología que permita analizar mediante el uso de machine learning la violencia en el departamento de Bolívar**?**	Crear una metodología que permita caracterizar la violencia en el departamento de Bolívar mediante la aplicación de minería de datos	Se puede crear una metodología que permita analizar la violencia presente en Colombia mediante minería de datos.	Modelo en R, arboles de decisión.
Problemas específicos	**Objetivos específicos**	**Hipótesis**	**Métodos**
¿Existen diferencias en los niveles de violencia entre las zonas urbanas y rurales?	Estimar si el nivel de violencia en el departamento difiere del entorno en el que se encuentre la población.	A priori se puede decir que el nivel de violencia es más alto en la zona urbana en comparación con la zona rural	Minería de datos

¿Cuál es el día de la semana donde se presenta la mayor cantidad de homicidios en el país y cuál es el sexo y grupo de edad que predomina?	Analizar Mediante el uso de machine learning la violencia presente en Cartagena.	Se espera que los días que se presentan la mayor cantidad de homicidios sean los sábados y domingos dado la ingesta del alcohol, al igual que el sexo más vulnerables sea la mujer.	Minería de datos
¿Se puede encontrar diferentes perfiles de homicidios en el departamento?	Caracterizar perfiles de homicidios en el departamento de Bolívar.	La metodología que se desarrollará permitirá determinar los diferentes perfiles de homicidio en el departamento de Bolívar.	Minería de datos con técnicas de machine learning.
¿Cómo definir perfiles de Homicidio en el departamento de Bolívar?	Establecer una metodología para definir el perfil de homicidios en una ciudad o departamento.	La metodología aplicada deberá establecer y definir el perfil de homicidios en el departamento de Bolívar.	Revisión bibliográfica, Minería de datos con técnicas de machine learning.

4.1 Cronograma de trabajo

En la tabla 3 se presenta el cronograma de trabajo con el cual se busca organizar el tiempo disponible para la realización del proyecto. En este sentido, se incluye las actividades que se realizaran con las fechas previstas de su comienzo y final, para realizarlo, se aplicó una herramienta conocida como el diagrama de Gantt, la cual es una herramienta que ayuda a programar todas las actividades a lo largo de un periodo determinado al igual que permite una mejor visualización.

Tabla 3. Cronograma de trabajo.



	Presentación a jurados	Preparación de la presentación	Redacción de informe final	análisis de resultados	Redacción del documento	Procesamiento de datos	Mejoramiento del modelo	Evaluación del modelo de ML	Selección e implementación de modelos de ML	Revisión bibliográfica	Depuración y limpieza de los datos	Contextualización	Estudio Con los datos obtenidos	Obtener los datos	Analizar la Problemática	Buscar la problemática que se va a trabajar
Fecha de inicio	20/08/2	06/08/2	16/07/2	25/06/2	04/06/2	21/05/2	14/05/2	07/05/2	23/04/2	09/04/2	02/04/2	24/03/2	09/03/2	21/02/2	06/02/2	05/01/2
Duración	1	14	21	21	21	14	7	7	14	14	7	9	15	16	15	32

4.2 Recursos a utilizar

Los recursos son aquellas herramientas que se usaron y que ayudaron a alcanzar los objetivos, es por ello que en la realización del proyecto se utilizaron:

- Laboratorio de simulación
- Base de datos de la página datos.gov
- Biblioteca
- Software R
- Profesor
- Antecedentes de la investigación

5. Entregables del proyecto

En la tabla 4 se detallan los entregables del proyecto, con el cual se busca mostrar todo lo que se genera como resultado del proyecto, así pues, se detallará objetivo por objetivo para dar cumplimiento y solución al objetivo planteado.

Tabla 4. *Entregables del proyecto.*

OBJETIVO	ENTREGABLE
Crear una metodología que permita caracterizar la violencia en el departamento de Bolívar mediante la aplicación de minería de datos	Documento donde quede señalado los requerimientos, características y especificaciones de la metodología.
Estimar si el nivel de violencia en el departamento difiere del entorno en el que se encuentre la población.	Documento descriptivo con los resultados de cada metodología estudiada (urbano o rural)
Analizar Mediante el uso de machine learning la violencia presente en Cartagena.	Tablas y gráficas donde se evidencien los resultados de cada metodología estudiada.
Caracterizar perfiles de homicidios en el departamento de Bolívar.	Tabla con la caracterización de homicidios, el código utilizado en R con todos los métodos implementados y gráficas.
Establecer una metodología para definir el perfil de homicidios en una ciudad o departamento.	Documento descriptivo con la metodología seleccionada.

6. Aprendizaje automático (machine learning)

El Machine Learning o Aprendizaje Automático se basa en el uso de distintos algoritmos que buscan crear sistemas que aprendan automáticamente. Por ello, resulta muy importante la elección del algoritmo más adecuado, así como también el hecho de disponer de un gran volumen de datos de suficiente calidad. En este sentido, existen dos definiciones muy populares, con las cuáles basta para tener una idea clara de la disciplina [46]:

- "El aprendizaje automático es el proceso que le da a las computadoras la habilidad de aprender sin ser explícitamente programadas".

Para poder hablar de machine Learning es necesario plantearse dos preguntas fundamentales, la primera es ¿Cuál es el objetivo del Machine Learning? Para resolver el problema es necesario preguntarse primero que es lo que se busca, puesto que existen muchos tipos de tareas que se pueden tratar, por ejemplo, de un problema de clasificación, como la detección de correo basura o "spam"; o de un problema de clustering, como recomendarle un libro a un cliente basándose en sus compras anteriores. Así pues, existen muchos problemas que pueden ser abordados usando diferentes algoritmos de machine Learning, es por ello que es fundamental saber qué se va a hacer con los datos que se tienen, ya que dependiendo de lo que se busca, el problema puede ser abordado desde diferentes enfoques.

En segundo lugar se pregunta ¿Cuáles son los insumos necesarios para desarrollar un modelo de Machine Learning?, Para esto es importante definir la pregunta objetivo a la cual se debe dar respuesta, así pues, en las fases iniciales del proceso de Data Science, es muy importante decidir si la estrategia que se utilizará será supervisada o no supervisada, y en éste último caso definir de forma precisa cuál va a ser la variable objetivo.

Por ejemplo, se presenta el conjunto de datos clientes.csv, el cual muestra información correspondiente a clientes de una compañía de telecomunicaciones. El objetivo es crear una modelo para predecir la pérdida de un cliente por parte de una compañía, este es un problema real para todas las empresas y se conoce como (Customer Churn). Como científicos de datos la misión es minimizar la tasa de Churn, es así como la labor consiste en analizar los datos y ayudar a la compañía a retener sus clientes. El dataset presenta información demográfica de los clientes, su comportamiento y como variable de decisión "Churn" la cual define si el cliente abandona la compañía o no. Se puede analizar todas las variables que se crean convenientes y sean aptas para el modelo que se piensa utilizar.

Esa cantidad ingente de datos son imposibles de analizar por una persona para sacar conclusiones y menos todavía para hacer predicciones. Los algoritmos en cambio sí pueden detectar patrones de comportamiento contando con las variables que se le proporcionen y descubrir cuáles son las que han llevado, en este caso, a darse de baja como cliente. Así pues, a continuación se dará respuesta a este problema usando una herramienta de Machine Learning:

Para desarrollar el código, se crea el modelo en R, el cual es:

```
clientes <- read.csv( "clientes.csv")
clientes

view(clientes)

# Para poder desarrollar el modelo decidimos que algoritmo utilizar

summary(clientes)

  library(rpart)

# Analizar y decidir las variables aptas para utilizar en el modelo

  sam <- sample(2, nrow(clientes), replace = TRUE, prob = c(0.75,0.25))
  train_data <- clientes [sam == 1,]
  test_data <- clientes[sam == 2,]

# Se utilizara el arbol de decisión para eso hacemos:

  Arbol <- rpart( Churn ~  ., method = "class", data = train_data)
  Prediccion <- predict(Arbol, test_data, type = "class")

#Con esto creamos la matriz de confusion

  table(Prediccion, test_data$Churn)

#Ahora procedemos a calcular la presición del modelo

  sum(Prediccion==test_data$Churn)/length(test_data$Churn)*100
```

```
Prediccion   No  Yes
       No  1285  369
       Yes   65   85
> sum(Prediccion==test_data$Churn)/length(test_data$Churn)*100
[1] 75.94235
>
```

Figura 15. *Código en R.*

Fuente: Elaboración Propia

En este sentido, se obtiene la tabla de confusión del modelo y así se obtiene un 75.94235 de precisión en el modelo.

De este modo, los algoritmos de aprendizaje automático se pueden dividir en tres grandes categorías: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje de refuerzo. El aprendizaje supervisado es útil en los casos en que una propiedad está disponible para un determinado conjunto de datos, pero debe predecirse para otras instancias. El aprendizaje no supervisado es útil en los casos que consiste en descubrir relaciones implícitas en un conjunto de datos en los cuales los elementos no han sido asignados previamente. El aprendizaje de refuerzo cae entre estos dos extremos: hay alguna forma de retroalimentación disponible para cada paso o acción predictiva, pero no hay etiqueta precisa o mensaje de error.

6.1 Aprendizaje supervisado

El aprendizaje supervisado hace referencia a problemas que ya han sido resueltos, pero que vuelven a surgir en un fututo. La supervisión no involucra intervención humana más bien el hecho de que los valores objetivos proporcionen al algoritmo de aprendizaje una manera en la que pueda optimizar una función. Así pues, el aprendizaje supervisado depende de datos que han sido etiquetados. Por ejemplo, que una computadora logré distinguir imágenes de coches, de aviones, entre otros. La idea es que las computadoras al ya tener los datos puedan aprender por si solas de los ejemplos que se les ha dado y partiendo de esto puedan desempeñarse individualmente sin necesidad de que se le vuelva a ingresar ninguna información. Uno de los usos más extendidos del aprendizaje supervisado consiste en hacer predicciones a futuro basadas en comportamientos o características que se han visto en los datos ya almacenados.

A su vez, existen muchos tipos de algoritmos que se desprenden del aprendizaje supervisado, los más habituales son:

- **Árboles de decisión:** Es una herramienta que ayuda en la toma de decisiones, al presentar las diferentes opciones que hay y sus posibles consecuencias. Pueden emplearse para predecir la respuesta del público ante el lanzamiento de un nuevo producto, entre otros [47].

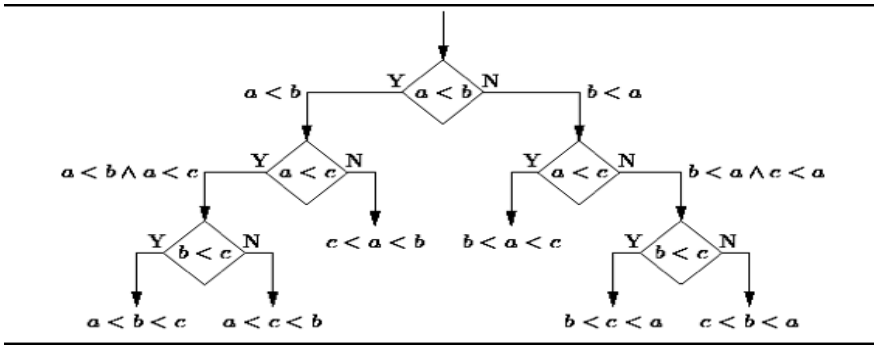


Figura 16. *Árboles de decisión*

- **Naïve Bayes Clasification:** Este tipo de algoritmo es usado para los software de reconocimiento facial, determinar en un texto si la emoción es negativa o positiva, puesto que dado un ejemplo permite encontrar la hipótesis que mejor se ajuste , para ellos es necesario que el sistema contenga datos suficientes.

- **Ordinary Least Squares Regression:** El método de los mínimos cuadrados ordinarios permite realizar la regresión lineal la cual se aplica al análisis de relaciones entre variables financieras. En este sentido, Linear se refiere al tipo de

modelo que está utilizando para ajustar los datos, mientras que los mínimos cuadrados se refieren al tipo de métrica de error que está minimizando [48].

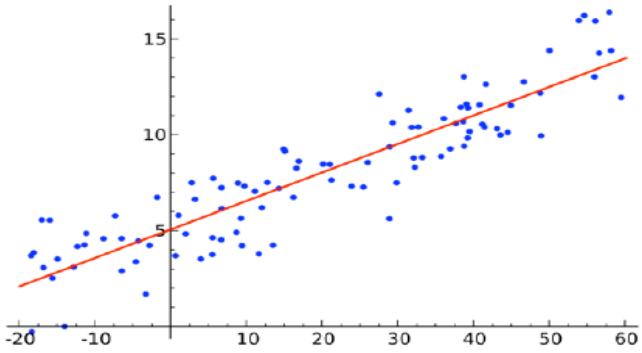


Figura 17. *Ordinary Least Squares Regression*

- **Logistic Regression:** La regresión logística es uno de los métodos más potentes cuando se tienen más de dos variables. Mide la relación entre la variable dependiente categórica y una o más variables independientes. Así pues, permite realizar clasificación binaria cuando se tienen dos categorías, dando como resultado valores entre 0 y 1.

6.2 Aprendizaje no supervisado

En el aprendizaje no supervisado el algoritmo no cuenta con etiqueta, de modo que no cuenta con ninguna indicación previa. En cambio, se le provee de una enorme cantidad de datos con las características propias de un objeto, para que pueda determinar qué es, a partir de la información recopilada, por tanto, tienen un carácter exploratorio.
Por ejemplo, las tareas de clustering, buscan agrupamientos basados en similitudes, pero nada garantiza que éstas tengan algún significado o utilidad. Así pues, existen muchos tipos de algoritmos que se desprenden del aprendizaje No supervisado, los más habituales son:

- **Algoritmos Clustering:** Clustering es la tarea de agrupar un conjunto de objetos tales que los objetos en el mismo grupo (clúster) son más similares entre sí que a los de otros grupos. Es usado mayormente cuando no se sabe que hay en los datos, por ello, él trata de crear grupos compactos. Cuando las variables son diferentes hay que estandarizarlos para poder distinguirlas y que estas no se confundan. Es importante tener claro que al momento de utilizar clustering no se está prediciendo, estos solo crean grupos.

- **Análisis de Componentes Independientes:** ICA es una técnica estadística que sirve para revelar los factores ocultos de los conjuntos de variables, mediciones o señales aleatorias. En el modelo, se supone que las variables de datos son mezclas lineales de algunas variables latentes desconocidas, y el sistema de mezcla también es desconocido.

5.3 Machine learning en la actualidad

En la actualidad el machine learning ha logrado que problemas que se creían imposibles de resolver hoy tengan solución, siempre y cuando el problema cuente con datos considerables y significativos. Así pues, aunque el machine learning es usado para tareas difíciles de darle solución, también es usado en tareas cotidianas como predicciones climáticas, servicios de traducción, entre otros. Dado esto, se puede aplicar a cualquier campo que cuente con bases de datos lo suficientemente grandes, tales como [49]:

- Clasificación de secuencias de DNA
- Predicciones económicas y fluctuaciones en el mercado bursátil
- Detección de fraudes
- Diagnósticos médicos
- Buscadores en Internet
- Sistemas de reconocimiento de voz
- Optimización e implementación de campañas digitales publicitarias, entre otros.

Podemos concluir que en un mundo donde la tecnología y la IA ha aumentado considerablemente, el machine learning ha tomado mucha fuerza, debido a la capacidad que tiene de trabajar con una enorme cantidad de datos y que al mismo tiempo obtengan de ellos conocimientos nuevos capaces de detectar tendencias más rápido de lo que cualquier humano podría hacerlo**.** Así pues, aunque se pueda trabajar con muchos datos y hacer más eficiente el proceso, es necesario la intervención del talento humano para poder perfeccionarse, ya que finalmente las computadoras no tienen un dominio elevado del lenguaje aplicado al razonamiento, es decir, que para que el machine learning se pueda desarrollar eficientemente, los expertos en cada campo de trabajo deben entrenar a las máquinas e irlas incorporando gradualmente a cada uno de los procesos que deseen afinar.

7. Análisis experimental del conjunto de datos

Fuentes de información para el análisis

La base datos fue obtenida de la página Datos abiertos gobierno digital de Colombia, la cual cuenta con una cantidad de 12439 observaciones de homicidios a nivel nacional [50] . El concepto de datos abiertos no es más que la información creada por la administración pública que pertenece a la sociedad, ya que han sido recopilados y financiados con dinero público, y por lo tanto debe estar disponible para cualquier ciudadano y para cualquier fin. Gracias a su publicación se promueve la transparencia de las entidades del estado, se ayuda a combatir la corrupción y se contribuye al empoderamiento y participación de los ciudadanos en la solución de problemas públicos. Todos estos datos deben ser publicado gracias a la existencia de una ley llamada la ley de transparencia y acceso a la información pública nacional (ley 1712 de 2014) [51] , todos estos datos son montados por el Ministerio de Tecnologías de la Información y las Comunicaciones. Este portal fue dispuesto por el gobierno colombiano para que las entidades dispongan sus datos abiertos a la ciudadanía.

Este trabajo se centrara principalmente en el departamento de Bolívar por lo que se utilizara solamente información perteneciente a este. Se contara con 404 observaciones del total antes mencionado.

7.1 Variables Seleccionadas

El dataset tenía una gran cantidad de variables, de las cuales algunas se consideraban relevantes y otras no, ya que no aportaban datos significativos para el desenvolvimiento del algoritmo. A continuación se mostraran las variables seleccionadas para la realización de nuestro estudio, se realizará una descripción de cada una de ellas y la frecuencia que tienen cada una de las categorías presentes en ellas.

7.1.1 Municipio

Esta variable indica el porcentaje de homicidios en cada municipio. En la siguiente figura se indica la frecuencia de ocurrencias por día.

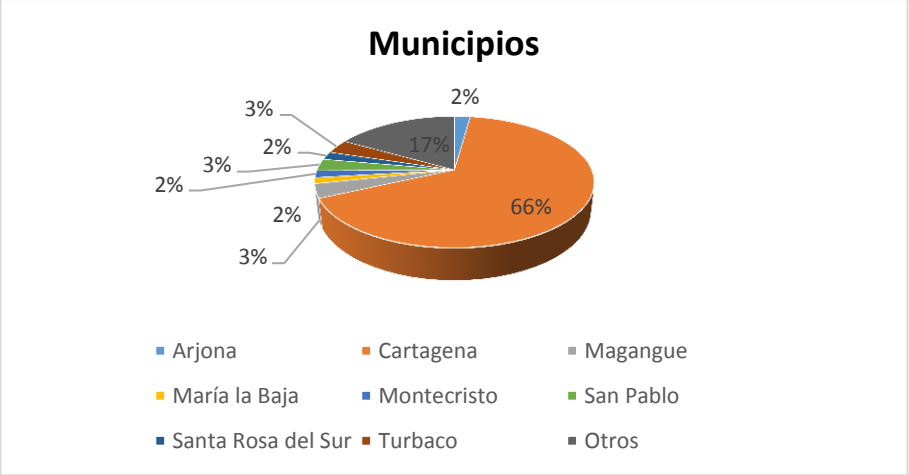


Figura 18. *Homicidios en municipios*

Como se puede observar en la Figura 22 la mayor parte de los homicidios se dan en la ciudad de Cartagena con un 66% de ocurrencia, seguido de Magangué con un 17%, otros municipios como Turbaco, Arjona, María la Baja tienen un porcentaje de homicidios relativamente bajos en comparación con estos dos.

7.1.2 Día

Esta variable indica el día en que ocurrió el homicidio. En la figura 23 se indica la frecuencia de ocurrencias por día.

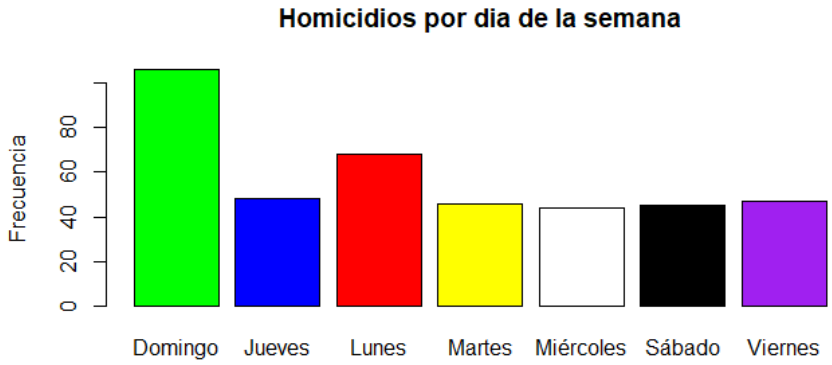


Figura 19. *Homicidios por día de la semana*

Como se puede observar en la Figura 23 la mayor parte de los homicidios se dan los días domingo y lunes, con 106 y 68 homicidios respectivamente. Teniendo el día domingo el doble de homicidios que el día de menores casos el cual es el miércoles con tan solo 44 observaciones. Así pues, se puede inferir que al ser un fin de semana, en el cual las personas normalmente salen, hay más consumo de alcohol y se pueden presentar más altercados y por ende desencadenarse en más homicidios. Según datos de medicina legal se reportaron alrededor de 31.766 entre 2007 y 2017, todo esto debido según la doctora Camila Quiñones al gran consumo de alcohol, el cual lleva a la persona a un estado en el que no puede razonar, convirtiéndola en una persona menos tolerante, provocando que se tomen decisiones equivocadas y que todo esto desenvuelva en pleitos o riñas, hasta el punto de terminar en asesinatos [52].

7.1.3 Barrio

Esta variable indica los barrios en los cuales se presentaron homicidios en el departamento de Bolívar, podemos observar a continuación las frecuencias más altas encontradas:

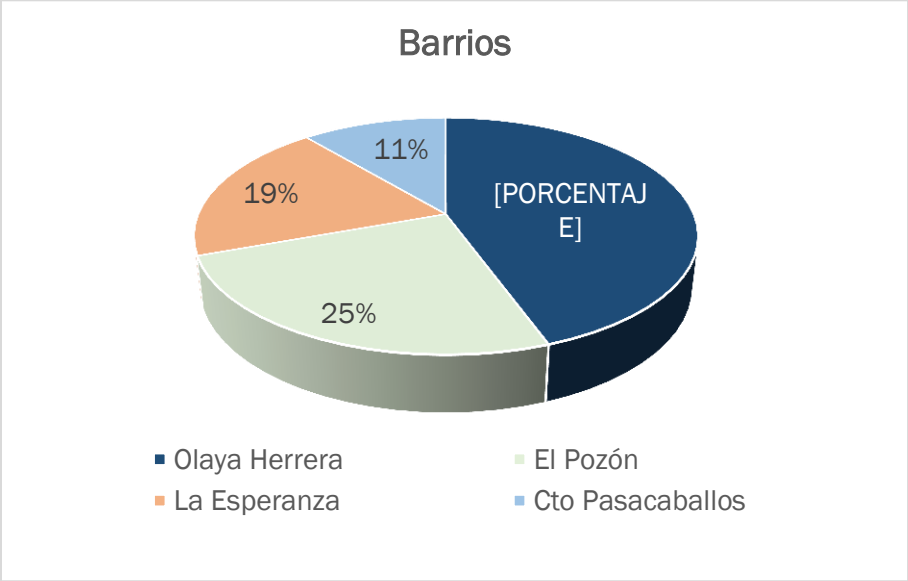


Figura 20. *Homicidios en barrios*

En la figura 24 se puede observar como las mayores tasas de homicidios se presenta en los barrios más vulnerables, la mayoría de estratos 1 y 2, en los cuales se puede observar grandes tasas de violencia, donde los niños desde pequeños crecen en medio de la guerra de pandillas, consumo de drogas entre otros problemas sociales. En este sentido, los jóvenes desde muy temprana edad son llevados a tomar caminos equivocados, convirtiéndose en

sicarios, ladrones, vendedores de droga, a que integren pandillas , ocasionando una guerra donde se matan unos a otros por la obtención de algún territorio [53].

7.1.4 Zona

Esta variable me indica la zona en la cual fue realizado el homicidio. Se pueden apreciar los datos en la siguiente ilustración:

Figura 21. *Zonas de homicidios.*

En la gráfica 25 se puede decir que el homicidio en la zona urbana domina, sobrepasando 4 veces el número de casos ocurridos en la zona rural, así pues, se puede inferir que al haber en la zona urbana muchas más personas y al ser la parte donde se encuentra la mayor parte de la población se van a presentar mucho más casos de homicidio. Haciendo una comparación entre Cartagena que está habitada por alrededor de 1.013.375 habitantes y malagana que tiene aproximadamente 8680 habitantes, se observa la gran diferencias de habitantes entre la zona rural y la zona urbana. En la actualidad, está aceptado que los impactos del delito y la delincuencia tienen, en el medio urbano, su principal centro de operaciones, y que es en las ciudades donde emergen los principales problemas de seguridad [54]. En un estudio realizado por el instituto nacional de medicina legal y ciencias forenses se dice que el homicidio es un fenómeno urbano aunque en las zonas rurales también se da en gran proporción debido a la violencia instrumental en la cual se busca obtener una ganancia por parte de la viíctima [55].

7.1.5 Clase de sitio

Esta variable me indica la clase de lugar en donde ocurre el homicidio. Se puede observar su frecuencia en la siguiente tabla:

Tabla 5. *Lugar de ocurrencia de los homicidios.*

Apartamento en conjunto cerrado	1
Bares, cantinas y similares	3
Billares	1
Carretera	12
Casas de habitación	5
Estadero	2
Fincas y similares	12
Hospitales	2
Hoteles, residencias y similares	1
Interior vehículo particular	1
Lagunas	2
Lote baldío	7
Parques	1
Pozo	1
Restaurante	1
Ríos	3
Tienda	1
Trocha	2
Vías publicas	346

En la tabla 5 se puede observar como la mayor parte de los homicidios ocurren en vías públicas, seguidas por carreteras y fincas y similares. Llevándose prácticamente todas las observaciones y dejando las otras categorías de la variable como casos muy particulares.

7.1.6 Arma o medio

Esta variable indica cual fue el arma utilizada para la realización del delito. A continuación se muestran los datos:

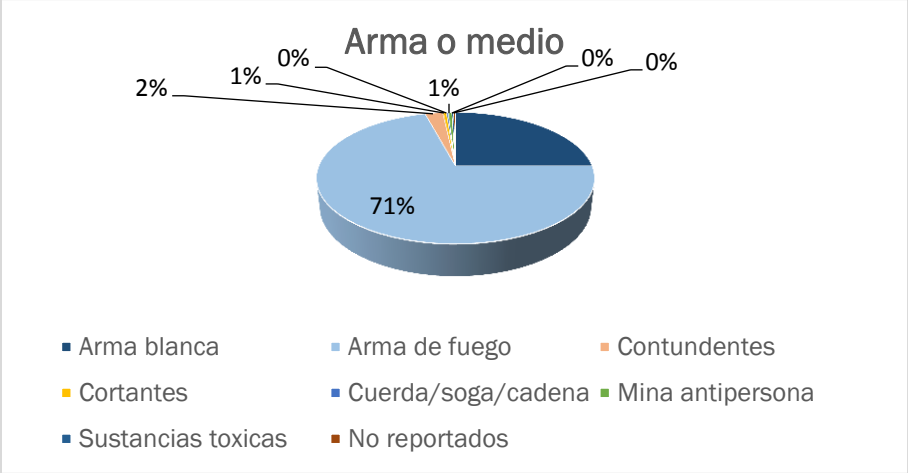


Figura 22. *Arma utilizada en los homicidios*

En la gráfica 26 se puede observar como más del 90% de los homicidios se dan por la utilización de armas de fuego o armas blanca, en este sentido, se puede decir que las armas de fuego son el medio más rápido, fácil y contundente para asesinar a una persona debido a que no necesitas de mucho esfuerzo para poder conseguir una, por ejemplo en Cartagena la venta ilegal de armas de fuego es preocupante ya que solo se necesita un poco de dinero para poder conseguirla , así pues, estas armas ilegales entran a la ciudad provenientes de Venezuela y Brasil. Investigaciones adelantadas por las autoridades también dan cuenta que son de fácil adquisición en varios municipios del Sur de Bolívar, donde están presente las banda criminales [56]. En este sentido, el uso de arma blanca queda de segunda posición con 101 en comparación con el arma de fuego la cual tuvo 286.

7.1.7 Móvil victima

Esta variable me representa el medio de transporte que estaba utilizando la víctima en el momento del asesinato. Los datos son los siguientes:

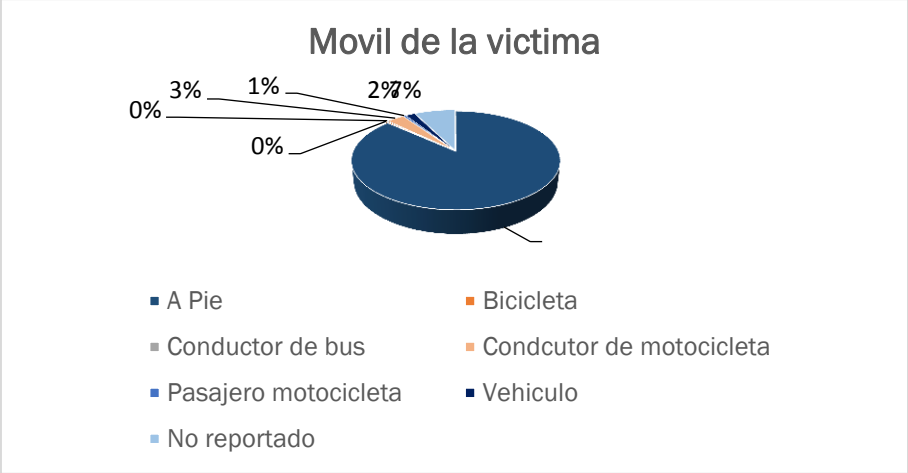


Figura 23. *Móvil de la victima de homicidio*

En la gráfica 27 se observa como la mayor parte de las víctimas de homicidio en bolívar se transportaban a pie en el momento en que se cometió el delito. En este sentido, Se puede decir que este medio de movilizarse es el más inseguro ya que es el transporte en el cual la víctima se encuentra más indefensa, desprevenida y sin ninguna oportunidad de defenderse. Así pues esto genera un problema de inseguridad muy grande, ya que hoy día está el miedo de que si sales pueden asesinarte, es así como la economía también se ve afectada. De acuerdo a un estudio que se realizó las decisiones de consumo y de inversión pueden cambiar ante la percepción que tengan los agentes sobre la duración de su vida [57].

7.1.8 Móvil agresor

Esta variable me indica el medio de transporte en el que iba la persona que cometió el homicidio. Se pueden ver los datos a continuación:

Tabla 6. *Móvil del agresor*

A pie	307
Conductor motocicleta	8
Pasajero motocicleta	60
Vehículo	3
No reportado	26

Como se puede observar en la tabla 6 la mayoría de los delitos son cometidos por personas que van a pie o se transportan como pasajeros de motocicletas. Y muy pocos casos se dan en vehículos.

7.1.9 Género

Esta variable nos indica el género de las personas que son asesinadas. Se puede apreciar en la gráfica 28 la diferencia porcentual en la cantidad de homicidios ocurridos en el departamento de bolívar dividiéndolos según su género con 385 y 19 para Hombres y mujeres respectivamente.

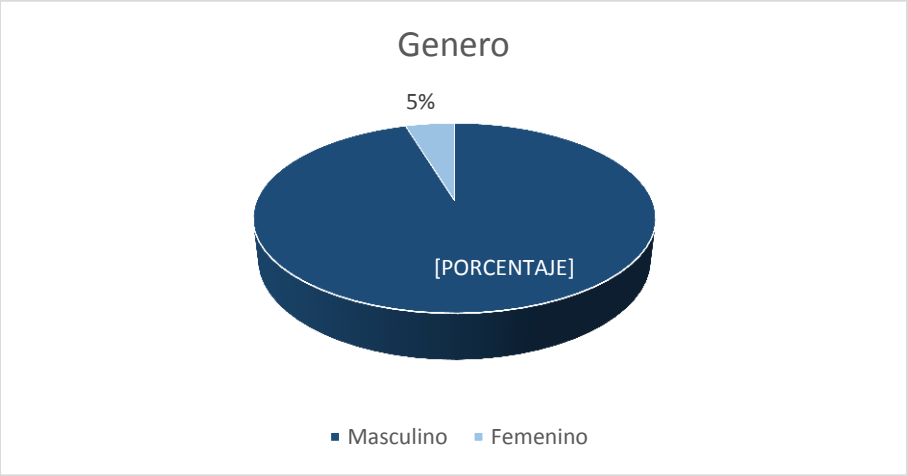


Figura 24. *Género de las personas víctimas de homicidio*

En la figura 28 se puede observar como la mayoría de las víctimas de homicidios tienden a ser hombres teniendo un 95.2% de los datos totales. En el artículo homicidio y lesiones infligidas intencionalmente por otra persona en los años 1973 a 1996 se obtuvo como resultado después de aplicar modelos de regresión logística que la tasa de homicidios oscilaba entre 33 y 150 por 100.000 habitantes, en cambio con las mujeres 3 y 12 por 100.000 habitantes, Una diferencia muy significativa [58]. Pero si se compara con datos obtenidos de un estudio que se realizó en la provincia de Sevilla en los años 2004 a 2007, Colombia se diferencia bastante, ya que las víctimas son en su mayoría varones presentando una relación hombre/mujer de 2:1 [59].

7.1.10 Estado civil

Esta variable describe cual es el estado civil de cada una de las víctimas. En el grafico 29 se puede ver la frecuencia de ocurrencia para cada una de las categorías.

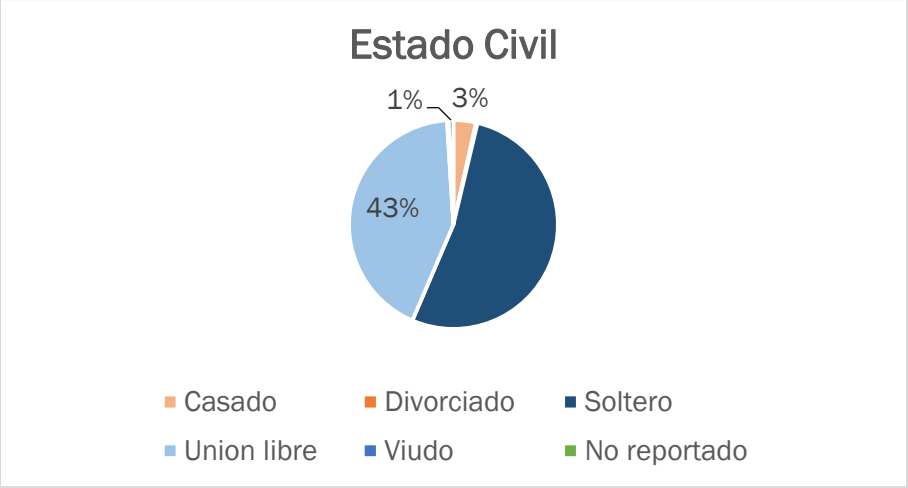


Figura 25. *Estado civil de las víctimas de homicidio*

Como se observa en la figura 29 la mayor parte de las victimas tienden a estar solteras o encontrarse en unión libre con 213 y 172 casos respectivamente. Y presentando un porcentaje casi nulo en aquellas personas divorciadas o viudas. En un estudio realizado por el instituto nacional de medicina legal y ciencias forenses con datos de Colombia en el año 2007 titulado Homicidios se obtuvo como resultado que los solteros son el grupo más afectado con el 28% del total de homicidios. Le siguen las parejas que estaban en unión libre con 21% de los casos, y los casados con 7,9%. Las categorías separadas, viudas y divorciado registran tan solo el 1,8%, 0,5% y 0,2% respectivamente [60]. Así pues, se puede decir que los solteros al no poseer un compromiso, exponen más sus vidas que aquellos que están casados o en unión libre al poseer ya un compromiso. Estos resultados concuerdan con los resultados que se obtuvieron para el año 2015.

7.1.11 Clase de empleado

Esta variable me indica que tipo de ocupación tenía la víctima. En la figura 30 se encuentra la frecuencia de las diferentes categorías dentro de la variable:

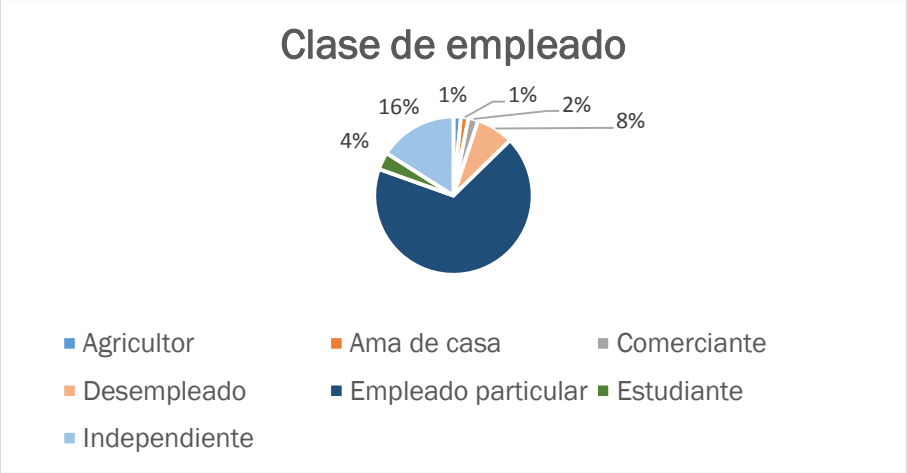


Figura 26. *Ocupación de la víctima de homicidio*

Como se puede observar en la figura 30 la mayor parte de las personas asesinadas eran empleados particulares, personas independientes, desempleados y estudiantes, teniendo la mayor frecuencia la categoría de empleado particular con 267 casos de los 404 que se encuentran en análisis y segundo la categoría independiente con 63 observaciones. En un estudio realizado para Colombia en el año 2008 se obtuvo como resultado que los mayores porcentajes se encontraban de igual manera en empleados particulares, aunque una gran parte de sus porcentajes se encontraban divididos en otras clases como labores del sector agropecuario, labores del sector de la construcción y en el sector de transporte, ocupando así de esta manera la mayor cantidad dentro del conjunto de datos. También se encontró que los bienes y servicios del sector económico informal y comerciantes y comercializadores también presentaban porcentajes significativos [60].

7.1.12 Escolaridad

Esta variable me representa el nivel de escolaridad que tenía la víctima. Se puede ver la frecuencia de cada categoría en la siguiente gráfica:

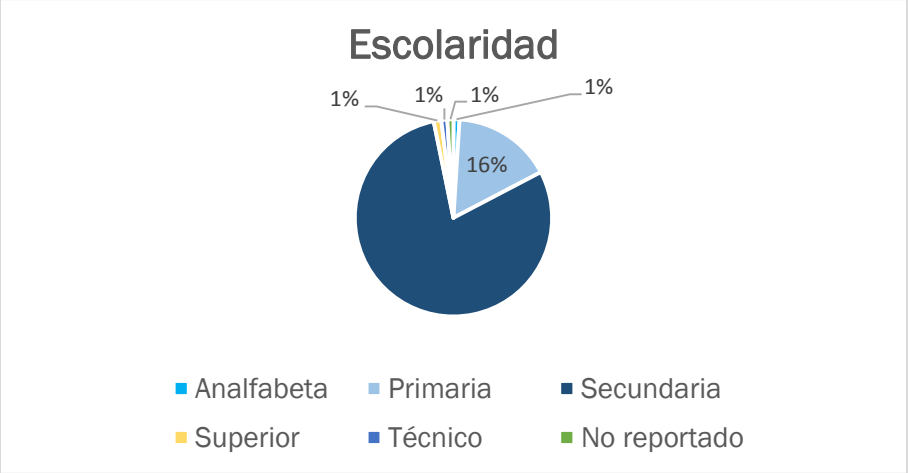


Figura 27. *Nivel de escolaridad de las víctimas de homicidio*

Fuente: Elaboración Propia

Como se puede observar en la figura 31 las mayores partes de las víctimas tenían una escolaridad de primaria o máximo secundaria. Presentando la mayor cantidad de casos la escolaridad secundaria con 321 datos de las 404 observaciones, siguiéndole la primaria con solo 66.

En el estudio realizado para Colombia en el año 2008 dieron como resultado que el 25% de las víctimas (3.515 casos) contaba con educación básica primaria completa; un 20% de las víctimas (2.559 casos) alcanzó la secundaria completa, y 2% se muestra sin ningún nivel educativo [60].

7.1.13 Intervalos de edades

Esta variable fue creada a partir de la columna de edades, en lugar de tener edades individuales se decidió crear diferentes grupos, los cuales cada categoría ocupa un intervalo de edades, permitiendo de esta forma lograr un trabajo mucho mejor y así obtener información más relevante. En el grafico 32 se pueden observar las frecuencias de cada uno de los intervalos:

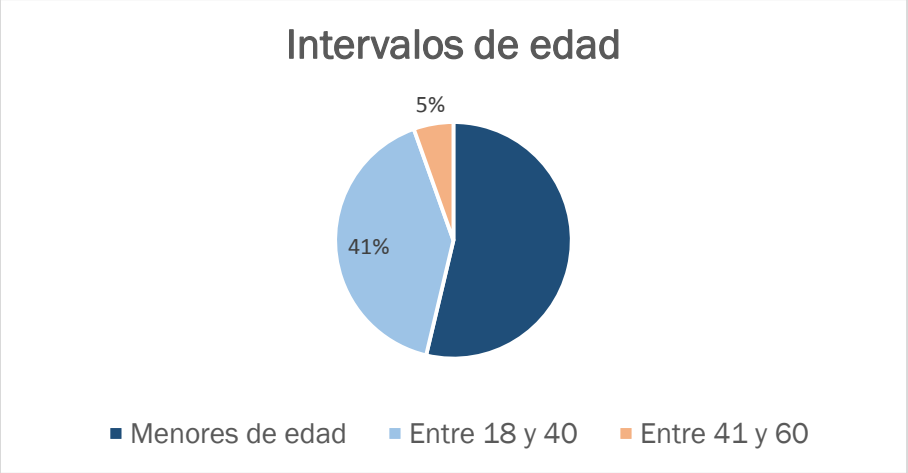


Figura 28. *Relación de homicidios con los intervalos de edad*

Fuente: Elaboración Propia

Como se puede observar en el grafico 32 existe una mayor presencia de menores de edad con 217 casos y de personas entre 18 y 40 años con 165. Es un dato preocupante que la mayoría de las victimas tienden a ser niños y jóvenes. En otro estudio realizado en Colombia en el año 2008 se encontró que, los rangos más afectados según sexo y edad son los hombres entre 20 y 44 años y las mujeres entre 20 y 29 años donde se encuentran las tasas más altas [60]. En un estudio realizado en Sevilla en los años 2004 a 2007 las víctimas son en su mayoría varones (relación hombre/mujer 2:1) con una edad media de 46 ± 21.2 años [59].

8. RESULTADOS DE LA METODOLOGÍA

Para iniciar con la aplicación de la minería de datos utilizamos todas las observaciones presentes en el departamento de Bolívar, y seleccionando cada una de las variables anteriormente explicadas.

Se procedió primero con la lectura de los datos y con la realización de su depuración, buscando valores faltantes y eliminándolos. Se eliminó la columna de departamentos ya que no aportaba nada a nuestro análisis.

Después se realizó una matriz de correlación para ver relación entre cada una de las variables, y como se puede ver en la gráfica 31 existe una correlación negativa entre el municipio y la zona de -0.44, el día presentaba una relación de 0.11 con el género y la escolaridad, el barrio tenía una correlación con escolaridad de 0.15, la zona con escolaridad presenta una fuerte correlación de 0.3 y con municipio de -0.44, la clase de sitio con municipio de -0.38, el móvil de la víctima con el móvil agresor también presenta una alta correlación de 0.41, la clase de empleado con municipio de 0.11, edad con el móvil agresor de 0.12 y de 0.1 con estado civil y por último el género con edad presento una correlación de -0.12.

	MUN	DIA	BARRIO	ZONA	C.SITIO	ARMA	M.VICT	M.AGR	EDAD	GEN	E.CIV	C.EMP	ESCOL	I.EDAD
MUN	1	0.06	0.07	-0.44	-0.38	0	-0.06	0.04	-0.02	0.01	-0.01	0.11	-0.29	0
DIA	0.06	1	-0.03	0	-0.06	0.09	0.04	0.04	0.03	0.11	0.04	-0.05	0.11	-0.01
BARRIO	0.07	-0.03	1	-0.02	0.03	-0.02	0.04	0.02	-0.01	0.07	-0.01	0.03	0.15	0
ZONA	-0.44	0	-0.02	1	0.41	-0.08	0.04	-0.01	-0.04	-0.01	-0.05	-0.24	0.3	-0.05
C.SITIO	-0.38	-0.06	0.03	0.41	1	-0.05	-0.04	0.02	0	0.03	0	-0.03	0.27	-0.07
ARMA	0	0.09	-0.02	-0.08	-0.05	1	0.04	0.11	-0.01	-0.2	-0.1	-0.03	0.04	0.01
M.VICT	-0.06	0.04	0.04	0.04	-0.04	0.04	1	0.41	0.05	0.03	-0.01	-0.03	0.12	0.05
M.AGR	0.04	0.04	0.02	-0.01	0.02	0.11	0.41	1	0.12	0.02	0.05	-0.05	0.11	0.09
EDAD	-0.02	0.03	-0.01	-0.04	0	-0.01	0.05	0.12	1	-0.12	0.1	-0.12	-0.14	0.89
GEN	0.01	0.11	0.07	-0.01	0.03	-0.2	0.03	0.02	-0.12	1	0.06	0.1	0.03	-0.08
E.CIV	-0.01	0.04	-0.01	-0.05	0	-0.1	-0.01	0.05	0.1	0.06	1	0.04	-0.03	0.1
C.EMP	0.11	-0.05	0.03	-0.24	-0.03	-0.03	-0.03	-0.05	-0.12	0.1	0.04	1	-0.1	-0.09
ESCOL	-0.29	0.11	0.15	0.3	0.27	0.04	0.12	0.11	-0.14	0.03	-0.03	-0.1	1	-0.11
I.EDAD	0	-0.01	0	-0.05	-0.07	0.01	0.05	0.09	0.89	-0.08	0.1	-0.09	-0.11	1

Figura 29. *Matriz de Correlación*

Después se procedió a la realización de un clúster jerárquico para poder tener una idea de cuál sería un número óptimo de clúster (K) a utilizar en el clúster no jerárquico utilizando el algoritmo kmeans. Pero como se está trabajando con variables del tipo categórico se procedió primeramente a realizar una dumificación, ya que se necesitan variables numéricas. La dumificación es un proceso utilizado para convertir variables categóricas en numéricas, seleccionando cada categoría de cada variable para convertirla en una columna o nueva variable. Por ejemplo tengo una variable que me indica si un carro es nuevo o viejo, estas dos últimas serían las categorías. Al realizar la dumificación nuevo y viejo serian dos nuevas columnas, y si la observación se encuentra en alguna de las dos colocara 1 y si no 0, ósea se convierte en una matriz de unos y ceros. Para poder realizarla se debe instalar el paquete "dummies".

Al tener el dataframe dumificado se procede a realizar una matriz de distancias utilizando la función "dist", ya que el clúster jerárquico es un algoritmo que trabaja con distancias.

Después se procedió a realizar el clúster jerárquico con la distancia euclidean, y se obtuvo como resultado el siguiente dendograma:

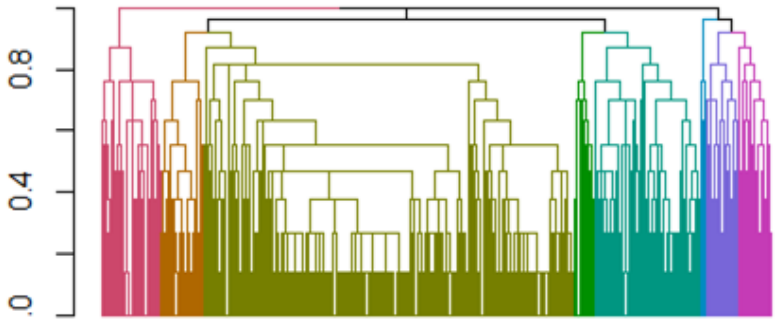


Figura 30*. Dendograma*

Como se puede apreciar en la figura 34 el dataset debe tener 8 clúster, de los cuales son 6 grandes o dominantes. Se probó con otro tipo de distancia para comparar resultados, se utilizó la distancia bray y se obtuvo la siguiente agrupación:

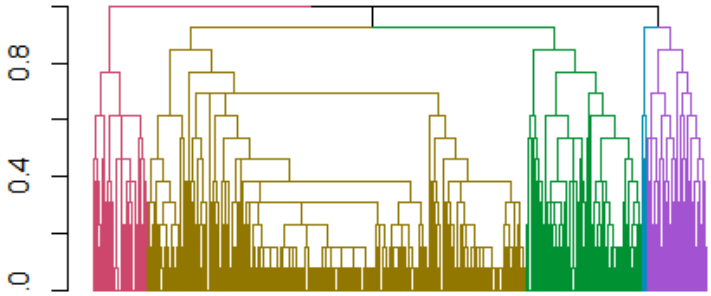


Figura 31. *Dendograma con distancia bray*

Como se puede apreciar en la figura 35 obtenida, ahora se cuenta con una agrupación de 5 clúster con 4 dominantes. Este último clustering obtenido dio mejores resultados así que probará con 4 clúster para la realización del agrupamiento usando Kmeans.

Pero antes de iniciar realicemos una gráfica de comparación entre cual es el mejor k para realizar un kmeans. Se busca el valor K para el cual la curva genera un quiebre en su pendiente, esto se conoce como el método del codo. Se basará en el tot withinss el cual me indica cual k proporciona una mejor agrupación para tener más parecidos intra clúster y más diferencias inter clúster. Podemos apreciar en la gráfica 36:

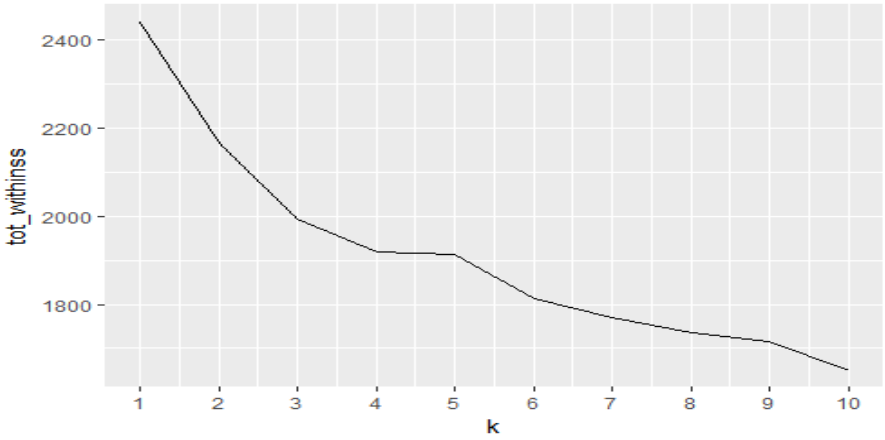


Figura 32. *Gráfica tot withinss*

Para la selección del mejor k se debe buscar el codo que se forma en el gráfico, al parecer en la figura 34 se encuentra en k = 2 o k = 3. Pero antes de realizar el kmeans se realizó otro método de evaluación para saber el mejor número de clúster. El otro método utilizado

fue el algoritmo PAM, primero se realizó un análisis de silueta para ver cuál es el k más óptimo a seleccionar y así definir el mejor número de clúster para trabajar.

En este tipo de gráficos se debe centrar principalmente en el punto más alto de la curva. En la figura 37se puede apreciar la imagen del análisis:

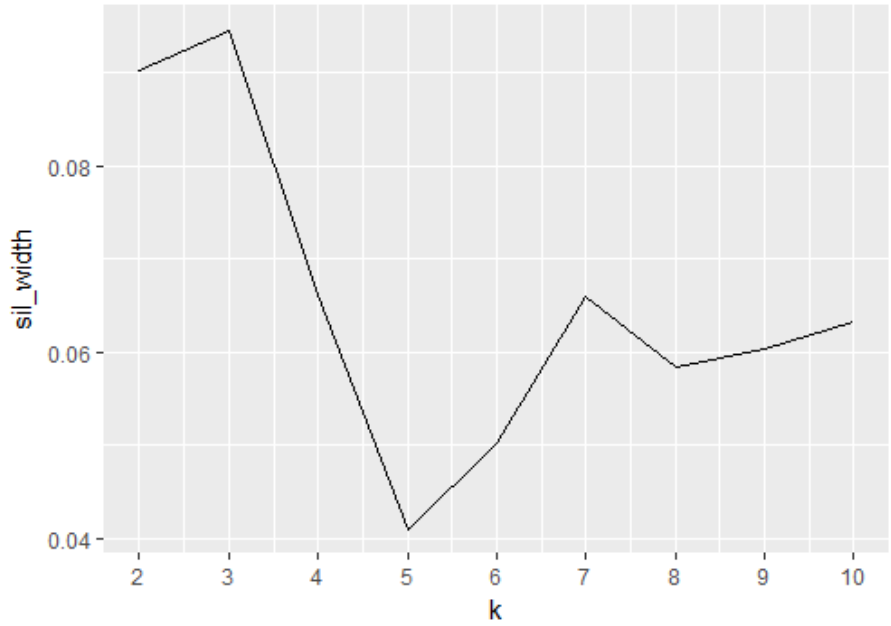


Figura 33. *Gráfica sil_width*

Como se puede ver en la gráfica 37 muestra que el mejor k es 3 ya que se encuentra en el punto más alto de la gráfica y después le sigue el número 2, e incluso también se tiene 4 y 7, pero esta vez el análisis se centrará en 2 y 3.

Se realizó el kmeans primeramente con 3 clúster para ver sus resultados. En este sentido, se realizó la implementación del PAM para k = 3 y comparar resultados posteriormente entre este y el kmeans y seleccionar el método de clustering no jerárquico más eficiente para el análisis.

Luego de haber realizado los clustering debemos ver gráficamente como quedaron cada uno, y así seleccionar el que presenta la mayor división entre los clúster creados, ya que este será el que me diferencie de mejor manera los clúster creados permitiendo obtener perfiles lo menos parecidos. Para hacerlo utilizaremos el paquete factominer y factoextra. Se aplica el fviz para todos los clúster, tanto kmeans como el PAM. Al aplicar el algoritmo al clúster creado con kmeans y k =3 se obtuvo el grafico 38:

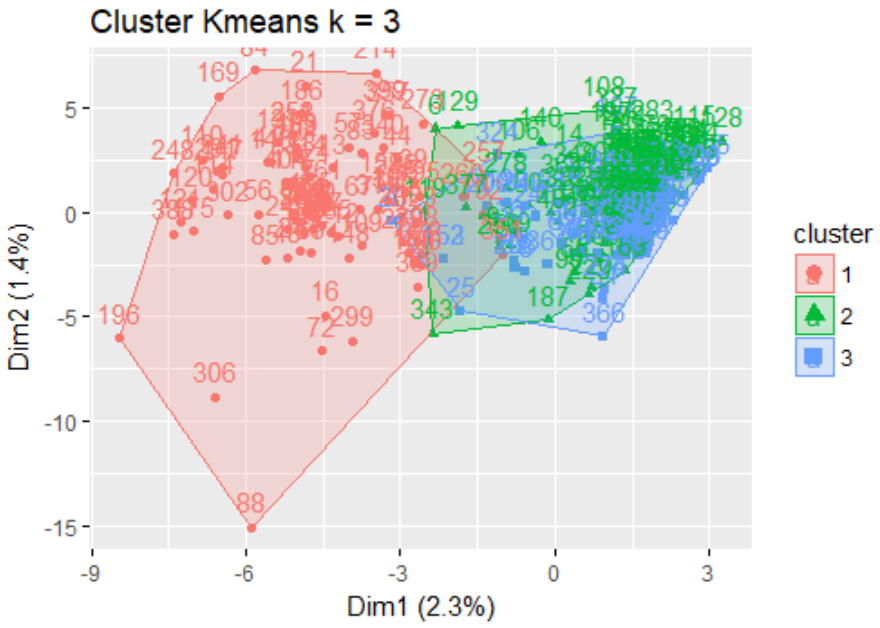


Figura 34. *Clúster kmeans*

Se puede observar en la figura 38 que hay un clúster que se separa casi completamente de los otros dos el cual es el 1, pero en el 2 y 3 se encuentran muy juntos por lo que podemos decir que ambos clúster presentan características muy similares, por lo que no difieren mucho las observaciones del clúster 2 y 3.

Al realizar el agrupamiento con el PAM se obtuvo el grafico 39:

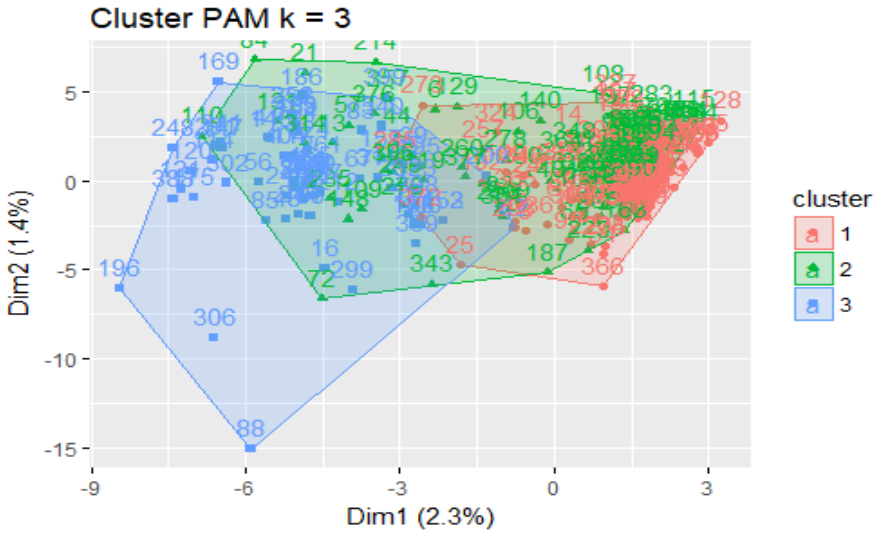


Figura 35. *Clúster Pam*

Como se puede observar en el grafico 39 los resultados esta vez fueron incluso peores que los obtenidos al aplicar el método kmeans con k = 3.

Después se procedió con la realización de un kmeans y un PAM pero esta vez se decidió utilizar el otro k optimo que teníamos el cual es 2.

Realizamos primero el clustering con la técnica PAM y se obtuvo el grafo 40:

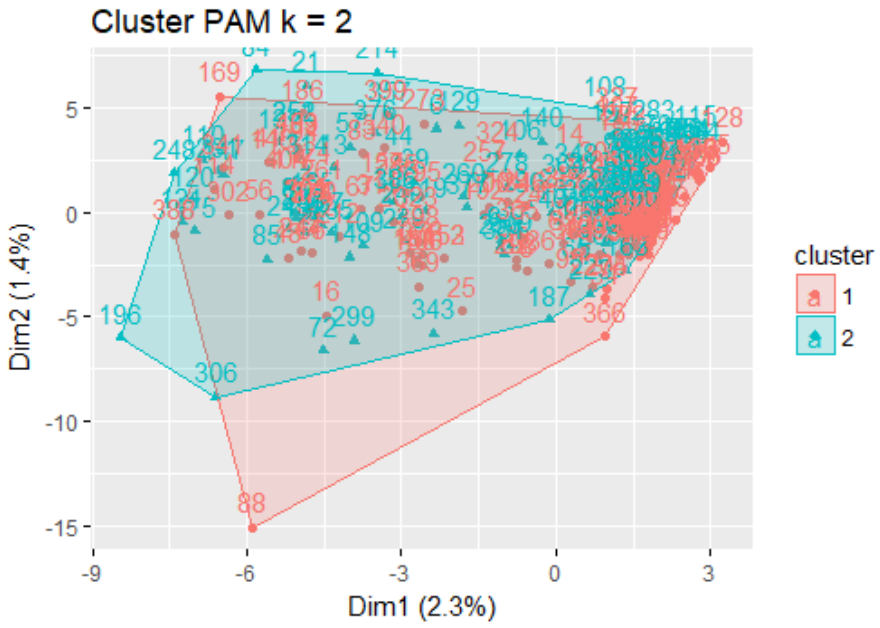


Figura 36. *Clúster Pam*

Como se puede observar en la figura 40 aun utilizando un k = 2 el PAM no logra dividir de la mejor manera los grupos creados, quedando así en ambos grupos observaciones muy parecidas. Después se realizó la técnica kmeans pero esta vez usando un k = 2. Y se obtuvo el grafico 41:

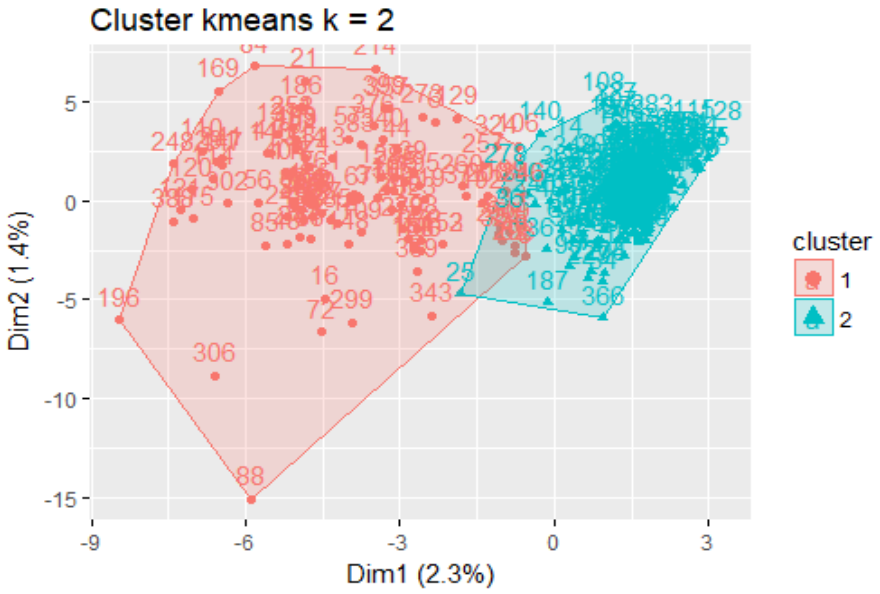


Figura 37. *Clúster kmeans*

Como se puede observar en la figura 41 al aplicar el kmeans con un k = 2 se logra diferenciar casi perfectamente a un clúster de otro, solo que algunas observaciones inter clúster comparten características similares, pero aun así se puede notar a simple vista la separación que existe entre ambos clúster.

Al tener lista la división por grupo, se asignó a cada observación del conjunto de datos el clúster que le fue dado por la aplicación del algoritmo. Seguido se realizó un análisis para descubrir cuáles eran las características que diferenciaban a cada clúster de acuerdo a cada una de las variables que se encontraban en estudio, y se obtuvo la caracterización representada en la tabla 7.

Tabla 7. *Observaciones representativas para el clúster 1 y 2*

Clúster	**Cant %**	**Municipio**	**Edad**	**Arma o medio**	**Día**	**Zona**	**Clase de sitio**	**Clase de empleado**	**Escolaridad**
1	27.5%	Magangué	Entre 18 y 40	Arma de fuego	Domingo y sábado	Rural	Vías públicas, carreteras y fincas	Independiente	Primaria
2	72.5%	Ctgna	Menores de edad	Arma de fuego	Domingo y lunes	Urbana	Vías públicas	Empleado particular	Secundaria

Como podemos observar en la tabla anterior existen múltiples diferencias entre un clúster con otro, permitiendo así tener dos grupos casi completamente separados. En la selección de las características o categorías de cada variable que representaría a cada clúster se basó principalmente en la moda de cada columna, ya que se está trabajando con variables categóricas. Ya con estos datos se pueden plantear estrategias para disminuir o atacar a este delito en el departamento de Bolívar de forma más eficiente.

La interpretación de la tabla 7 es la siguiente:

Clúster 1(27.5%): Se caracteriza principalmente por homicidios en su mayor parte hacia personas entre los 18 y 40 años, también el principal medio para cometer el delito es arma de fuego, los días que más ocurren son los fines de semana tanto sábados como domingos y se centran principalmente en las zonas rurales en vías públicas, carreteras y fincas. Se podría considerar como homicidios ocurridos por ajuste de cuentas.

Clúster 2(72.5%): se caracteriza por homicidios en menores de edad, mediante el uso de armas de fuego, ocurren en su mayor parte los días lunes y domingos en la zona urbana principalmente en las vías públicas. Se podría considerar a este tipo de homicidio como "homicidio ocurridos por riñas".

9. DISCUSIONES

Esta investigación tuvo como propósito la creación de una metodología que permitiera la caracterización de la violencia en el departamento de Bolívar mediante la aplicación de machine learning, más específicamente en el delito de homicidio. Es así como se realizó un análisis experimental del conjunto de datos estudiado, con el que se identificaron las variables asociadas a este tipo de actos delictivos tales como el arma usada, el género, la zona donde ocurrió, el municipio, entre otros. A continuación se realizará una discusión con los principales hallazgos de esta investigación.

Los resultados encontrados muestran una clara diferencia entre los dos clúster realizados por el algoritmo de agrupamiento, principalmente se puede ver que la zona en la que ocurre el mismo es una de las características más diferenciadoras. Otra característica que marcó la diferencia es el rango de edad que más prevalece en las víctimas de ambos grupos, en uno menores de edad, y en el otro en 18 y 40 años, dos rangos que pueden desprender un conjunto de hipótesis de las razones de homicidios y plantear estrategias para prevenirlo.

Por otro lado, de los datos expuestos se puede concluir que los hombres son las más vulnerables ante esta clase de delitos con un 95%, la mayor parte de estos eran trabajadores particulares, solteros, con educación básica secundaria. Por otro lado, el arma o medio con el cual realizaron el delito fue el arma de fuego debido a su rapidez y efectividad.

Con en el análisis realizado se buscó extraer conocimiento a partir de la base de datos del estado de los homicidios ocurridos en el departamento de bolívar. Para ello se aplicaron técnicas de minería de datos las cuales fueron implementadas en todo el desarrollo de la metodología permitiendo obtener resultados significativos.

En un estudio realizado en Argentina llamado "Aplicación de Minería de Datos Para la Exploración y Detección de Patrones Delictivos En Argentina" [61], tuvo como objetivo realizar una implementación de minería de datos en el análisis de información criminal en Argentina y comprobar su efectividad y valor agregado. Así pues, se utilizaron un conjunto de técnicas de minería de datos similares a las de esta investigación contando al igual con información suministrada por entidades del estado. Se aplicaron técnicas de minería de datos para el completo análisis de las variables, sin embargo, en este se empleó un programa diferente, el software propuesto fue Weka 3.5.5 que al igual que R poseen características similares, tales como el costo, ya que ambos son gratis, de igual forma, son amigables gráficamente en su interfaz y además son fáciles de usar. Así pues, una de las ventajas que tiene R es la gran gama de algoritmos que se pueden aplicar, ya que ofrece libertad a los usuarios de crear algoritmos y que ellos mismos puedan compartirlos con el resto de usuarios de manera gratuita.

De este modo, aunque se usaron software diferentes, se lograron los resultados esperados en ambos. Así pues, en el primero (Aplicación de Minería de Datos Para la Exploración y Detección de Patrones Delictivos En Argentina) se logró hacer un análisis completo de los clúster creados siendo el primer clúster clasificado como "homicidios en ocasión de riña o ajuste de cuentas" debido a que estos fueron cometidos en vías públicas y sin ninguna relación con otro tipo de delitos y el segundo clúster fue clasificado como "homicidios en ocasión de emoción violenta" debido a que estos fueron cometidos sin arma de fuego y en domicilios particulares. De esta forma, en esta investigación (metodología para el análisis de la violencia en el departamento de bolívar mediante técnicas de machine learning**)** se obtuvieron dos clúster correctamente diferenciados mediante el método de agrupamiento Kmeans, debido a que fue el que mejor resultados arrojó. En este sentido, el primer clúster se clasificó como "homicidios ocurridos por ajuste de cuentas" ya que fueron cometidos con arma de fuego en zonas rurales y carreteras y el segundo clúster se clasificó como "homicidio ocurridos por riñas" puesto que fueron cometidos con armas de fuego en vías públicas.

Por otro lado, de estos resultados se desprenden información relevante para futuras investigaciones. Es necesario que se investigue sobre otras clases de delitos que aquejan no solo nuestro departamento sino también nuestro país, es así, como se pueden crear bajo esta misma metodología patrones delictivos para hurto, delitos sexuales, entre otros. La exposición a la violencia a la que son expuestos los jóvenes hoy en día hace que este tipo de investigaciones sean significativas para que se estudian las variables que están influenciando en la ocurrencia de las mismas y de esta forma se puedan crear metodologías para la prevención de estos tipos de delitos.

10.CONCLUSIONES

En el presente trabajo investigativo se logró probar la utilidad de machine learning en el análisis de la violencia presente en Colombia especialmente en el ámbito de los homicidios, encontrando así información relevante que pueda generar nuevos conocimientos que permitan crear nuevos objetivos y estrategias para la disminución de los mismos.

En este sentido se logró dar respuesta a cada uno de los interrogantes planteados al inicio de la investigación por medio de la utilización de algoritmos en R:

- Se analizó mediante minería de datos la violencia presente en Bolívar, especialmente en la categoría de homicidios, esto se logró gracias a la aplicación de algoritmos de machine learning en el software de distribución libre y gratuita R, como pueden ser métodos supervisados y no supervisados.
- Se logró estimar que el nivel de homicidios en el departamento de Bolívar difiere de la zona en la que se encuentra la población, ya que al realizar el clustering con el algoritmo kmeans se obtuvieron resultados que permitían diferenciar dos patrones distintos de homicidios, especialmente debido a la zona en donde ocurrieron los hechos, ya fuera rural o urbana.

Ahora se procede a la utilización de un algoritmo de clasificación, para verificar si este logra identificar de acuerdo a unos datos recibidos el clúster al que pertenece cada observación.

Primeramente utilizaremos el algoritmo KNN para predecir nuestro dataset. Como se sabe el KNN solo trabaja con variables numéricas así que el dataframe datos2 en los cuales añadimos la columna de clúster la debemos dumificar como hicimos anteriormente.

Al haberla dumificado se debe proceder con la creación de un dataset de entrenamiento que es con el cual se prepara al algoritmo, y uno de test para probar si el algoritmo logra predecir cada uno de sus clúster, los cuales ya son conocidos. Para el dataset de entrenamiento seleccionaremos el 75% de los datos, y para el test el 25%. Ahora se procede con la aplicación del algoritmo, se utilizara el siguiente algoritmo:

Para su correcta aplicación se deben eliminar en el método las etiquetas de cada dataset, tanto en el de entrenamiento como en el test. Esta etiqueta no es más que el número de clúster asignado para cada observación, pero como lo estamos entrenando se debe añadir otro parámetro el cual es “cl”, en este parámetro se deben incluir los labels o etiquetas del train que se hizo anteriormente, y colocamos k = 2 ya que es el número de clúster en que queremos que prediga.

Listo, ahora se procede a ver cómo le fue al algoritmo, viendo en cuantas observaciones consiguió asignarle el número de clúster correcto.

Los resultados obtenidos son los siguientes:

Tabla 8. *Matriz de confusión*

	1	2
1	21	5
2	0	73

Por lo que la predicción del modelo fue de 94.94%

A continuación se implementará otro método de clasificación para comprobar cuál de los dos es más efectivo. Se utilizará Decisions trees o arboles de decisión. A diferencia del KNN el árbol de decisión si puede trabajar con valores categóricos tanto con numéricos, por lo que no hay necesidad de realizar una dumificación.

Primero se realiza la instalación de los paquetes necesarios, luego se procede a la creación de los dataset de entrenamiento y de test, igual a como hicimos anteriormente con el método KNN.

Ahora debemos aplicar el algoritmo del árbol de decisión, pero primero se crea por fuera una función donde se incluye una variable dependiente que será el clúster y algunas variables independientes.

funcion <- Cluster ~ MUNICIPIO + ZONA + CLASE.DE.SITIO + ARMA.O.MEDIO + MOVIL.VICTIMA + MOVIL.AGRESOR + ESCOLARIDAD + Intervalos_edades

Ahora se realiza la predicción del modelo con el algoritmo necesario.

Luego se debe verificar la cantidad de aciertos que tuvo el algoritmo con el data de test:

Al aplicar la verificación se obtuvo la tabla 9:

Tabla 9. Matriz de confusión

	1	2
1	23	0
2	3	73

Por lo que podemos ver tiene una precisión bastante alta, por lo que podríamos decir que logra diferenciar en gran manera ambos clúster, teniendo de 23 miembros del grupo 1 logro predecirlos todos, y de 76 del grupo dos solo fallo en 3. Su precisión fue del 96.9%, mejor qué la conseguida al aplicar el método KNN.

ANEXO

```
# Homicidos Bolivar 2015

# Lectura de los datos

datos <- read.csv("Homicidios_2015_solo_Bolivar.csv", header
= TRUE)

datos$EDAD <- as.numeric(datos$EDAD)

# Creamos los intervalos de edades

datos$Intervalos_edades <- cut(datos$EDAD, breaks = c(0, 17,
40, 60),
                               labels = c("Menores de edad",
"Entre 18 y 40", "Entre 41 y 60"),
                               include.lowest = FALSE ,
                               right = TRUE)

View(datos)

head(datos)

# Quitamos la columna de departamento ya que no aporta nada

datos <- datos[,-1]

# Verificamos si existen valores na en nuestros datos

sum(is.na(datos))

sapply(datos,function(x) sum(is.na(x)))

which(is.na(datos$Intervalos_edades))

# No encontramos valores faltantes



# Antes de proceder con el analisis, veamos una matriz de
correlacion de nuestros datos

# Convertir el dataframe en valores numericos para que sea
capaz de realizar un analisis de correlacion

datos_numericos = as.data.frame(lapply(datos, as.numeric))
```

```
#proceder con la correlacion

install.packages("corrplot")

library(corrplot)

M_correlacion=round(cor(datos_numericos,datos_numericos),
digits=2)

corrplot(M_correlacion, method = 'color', type = 'full',
addCoef.col="black",

         number.cex=0.55, title = "Matriz de
correlacion",mar=c(0,0,2,0))


# Como podemos observar Existe una correlación negativa entre
el municipio y la zona de -0.44.

# El día presenta una relación de 0.11 con el genero y la
escolaridad.

# El barrio presenta una correlación con escolaridad de 0.15

# Zona con escolaridad presenta una fuerte correlacion de 0.3
y con municipio de -0.44

# la clase de sitio con municipio de -0.38.

# El movil victima con el movil agresor tambien presenta una
alta correlacion de 0.41

# La clase de empleado con municipio de 0.11

# Edad con el movil agresor de 0.12 y de 0.1 con estado civil

# Genero con edad de -0.12


# Realizamos la dumificacion

install.packages("dummies")

library(dummies)


datos2 <- as.data.frame(datos[,-9]) # datos2 es datos con la
nueva columna de intervalos, pero sin la columna EDAD

View(datos2)

datos2_dumificados  <- dummy.data.frame(datos2)
```

```
View(datos2_dumificados)


# Realizamos una matriz de distancias
dist_datos2 <- dist(datos2_dumificados, method = "binary")


# Creacion del cluster Jerarquico


# Aplicamos el metodo jerarquico de cluster
hc_datos <- hclust(dist_datos2)
# Veamos como se fueron agrupando cada una de las
observaciones
hc_datos$merge
#Grafiquemos el dendograma e la union jerarquica
plot(as.dendrogram(hc_datos))


hc_datos$labels


# Grafiquemos con colores
library(dendextend)
plot(color_branches(as.dendrogram(hc_datos), h = 0.9))


# podemos ver 8 clusters


# Clustering utilizando la distancia bray
install.packages("vegan")
library(vegan)


dist_datos2_bray <- vegdist(datos2_dumificados, method =
"bray")
```

```
# Aplicamos el metodo jerarquico de cluster

hc_datos2 <- hclust(dist_datos2_bray)

# Veamos como se fueron agrupando cada una de las
observaciones

hc_datos2$merge

#Grafiquemos el dendograma e la union jerarquica

plot(as.dendrogram(hc_datos2))


#veamoslos mucho mas claramente

rect.hclust(hc_datos2,h = 0.9, border = "red")


# graficar con colores


plot(color_branches(as.dendrogram(hc_datos2), h = 0.9))


# Ahora que sabemos que el numero de cluster asignado por el
metodo jerarquico es 5 mucho mas claramente, aunque son

# 4 grandes grupos.

# podemos utilizar de mejor manera el kmeans o metodo de
clustering no jerarquico


library(purrr)

library(ggplot2)


tot_withinss <- map_dbl(1:10,  function(k){

  model1 <- kmeans(x = datos2_dumificados, centers = k)

  model1$tot.withinss

})


elbow_df <- data.frame(
```

```
  k = 1:10,
  tot_withinss = tot_withinss
)

# Plot the elbow plot
ggplot(elbow_df, aes(x = k, y = tot_withinss)) +
  geom_line() +
  scale_x_continuous(breaks = 1:10)



# Parece que los mejores k segun la prueba con el kmeans es 2
o 3. probemos con 3

kmeans_datos2_k3 <- kmeans(datos2_dumificados, centers = 3)

# Ahora realicemos la verificacion con el pam y su silieta o
analisis silueta para ver
# cual es el k mas optimo a seleccionar y asi definir el
mejor numero de cluster
# Para trabajar

install.packages("cluster")
library(dplyr)
library(cluster)

sil_width <- map_dbl(2:10,  function(k){
  model2 <- pam(x = datos2_dumificados, k = k)
  model2$silinfo$avg.width
})
```

```
# generamos un dataframe que contenga tanto los k como el
sil_width
sil_df <- data.frame(
  k = 2:10,
  sil_width = sil_width
)


# Plot the relationship between k and sil_width
ggplot(sil_df, aes(x = k, y = sil_width)) +
  geom_line() +
  scale_x_continuous(breaks = 2:10)


# Por lo que podemos ver segun el analisis de silueta el
mejor k es 3, probemoslo
pam_datos1 <- pam(datos2_dumificados, k=3)
# Estos serian los centroides de los 3 grupos
pam_datos1$id.med
View(datos2)


# Verifiquemos esto graficamente con el paquete factominer y
factoextra
install.packages("FactoMineR")
library(FactoMineR)
install.packages("factoextra")
library(factoextra)


# Aplico el fviz para todos los cluster, tanto kmeans, pam y
el jerarquico
fviz_cluster(kmeans_datos2_k3, data = datos2_dumificados,
main = "Cluster Kmeans k = 3")
```

```
# Por lo que podemos ver los grupos no estan claramente
separados

fviz_cluster(pam_datos1, data = datos2_dumificados, main =
"Cluster PAM k = 3")


# Aca en el pam estan mucho menos diferenciados


# Probemos ahora con 2 cluster


kmeans_datos2_k2 <- kmeans(datos2_dumificados, centers = 2) #
Mostrar este

# Ahora aplico el fviz_cluster para visualizar mucho mejor


fviz_cluster(kmeans_datos2_k2, data = datos2_dumificados,
main = "Cluster kmeans k = 2") # El mejor



# Apliquemos el fviz_cluster a un pam seleccionando k = 2

# Aca los grupos se ven claramente diferenciados , con estos
datos si se puede trabajar


# Probemos con dos cluster en el metodo pam

pam_datos <- pam(datos2_dumificados, k= 2)

fviz_cluster(pam_datos, data = datos2_dumificados, main =
"Cluster PAM k = 2") # No sirve


# Nos quedamos con la division realizada por el metodo no
jerarquico kmeans con k=2



# Colocacion de la variable cluster del kmeans al
kmeans_datos2_k2
```

```
kmeans_datos2_k2$cluster

datos2$Cluster <- kmeans_datos2_k2$cluster

table(datos2$CLASE.EMPLEADO, datos2$Cluster)

prop.table(table(datos2$MUNICIPIO, datos2$Cluster),2)*100


# saquemos los medioides o centroides de nuestro clustering

kmeans_datos2_k2$centers   #Duda con respecto a los barrios o
municipios, ¿Como se beria interpretar?


# Saquemos los municipios mas afectados

municipios <- prop.table(table(datos2$MUNICIPIO))*100

View(municipios)

# Realizamos un table para saber la cantidad de personas por
grupo dentro de los clsuter

table(datos$Intervalos_edades)

table(datos2$Intervalos_edades, datos2$Cluster)

prop.table(table(datos2$Intervalos_edades, datos2$Cluster))

prop.table(table(datos2$Intervalos_edades, datos2$Cluster),
2)

mean(datos$EDAD)

# Como podemos observar en ambos cluster se presenta una
semejanza en las proporciones

# de las edades, obteniendo asi un bajo porcentaje en ambos
para personas entre 41 y 60 años del 5.4%

# pero se obtiene en el cluster 2 una mayor presencia de
menores de edad con un 56,3 % mientras que en el cluster 1

# ocupan un 46.8%

# y en el cluster 1 la mayor proporcion la tiene las edades
entre 18 y 40 con un porcentaje bastante

# parecido al de menores de edad con 47.7%.
```

```
# Realizamos un table para saber el ARMA o MEdio

table(datos2$ARMA.O.MEDIO)


#En todo los datos obtenemos una presencia de armas de fuego ocupando un 70.7% y armas blancas con un 25%

# El resto de elementos utilizados estan en contundentes. cortantes, cuerda, soga o cadena, mina antipersona

# sutancias toxicas y algunos casos no reportados.


# Miremos por cluster

table(datos2$ARMA.O.MEDIO, datos2$Cluster)

prop.table(table(datos2$ARMA.O.MEDIO, datos2$Cluster),2)*100


# Segun la tabla se puede observar que tanto en el cluster 1 como en el 2 mas del 90% de

# las armas utilizadas para cometer los homicidos son arma blanca y arma de fuego,

# teniendo una mayor proporcion el arma de fuego en ambas con un 64.8 % y 73.03 % respetivamente.

# Algo interesante es que en el cluster 2 ocurren muchos mas homicidos utilizando armas de fuego sacandole casi un 10% mas.

#  y en el cluster 1 se presenta un 1.8% por minas antipersona, que aunque es muy poco me brinda una caracteristica

# especial delcluster


# Miremos por municipio

# En general

prop.table(table(datos2$MUNICIPIO))*100
```

```
# se puede observar que de todos los homicidios ocurrios en
el departamento de bolivar el 66.5%(269 casos) de estos
ocurren en el municipio de Cartagena

# el 2.22(9) en Arjona, 3.46(14) en magangue, el 3.46(14) en
Turbaco, y el 2.9(12) en san pablo.

# Por que no llega al 100%?

#Por cluster

table(datos2$MUNICIPIO, datos2$Cluster)


# Podemos ver que el cluster 2 contiene la mayor parte de los
datos por tener a cartagena con 269 casos.

# Tambien tiene a turbaco, mientras que en el cluster 1
contiene la mayor parte de los municipios incluyendo arjona
con 8 de las 9 muertes, magangue

# con 13 e 14 y san pablo con 11 de 12.

prop.table(table(datos2$MUNICIPIO, datos2$Cluster),2 )*100


# Segun los municipios podemos darnos cuenta que en el
cluster 2 se encuentra dominando

# cartagena con un 91.8 %, despues le sigue turbaco con un
4.09 % y el resto de los porcentaje

# se reparte entre el resto de municipios.



# Miremos por dia

table(datos$DIA)

barplot(table(datos$DIA), main = "Homicidios por dia de la
semana", ylab = "Frecuencia", col = c("green", "blue", "red",
"yellow", "white", "black", "purple"))

table(datos$DIA, datos2$Cluster)

prop.table(table(datos2$DIA, datos2$Cluster),2)*100
```

```
 # por lo que podemos ver la mayir porporcion de homicidios se da el dia domingo en ambos cluster, con un 29.7 % y 24.9 % respectivamente.

# aunque en el cluster 1 tambien ocurren los dias sabado con un 17.1 % y en el cluster 2 los lunes con un 18 %.



# Miremos por Barrio

# en general

table(datos2$BARRIO)


# podemos darnos cuenta que el barrior que presenta la mayor proporcion de aseinatos es olaya herrera con 32 asesinatos,

# despues le sigue el pozon con 18, y por ultimo la esperanza con 14 y pasacaballos con 8.


#Por cluster

table(datos2$BARRIO, datos2$Cluster)

prop.table(table(datos2$BARRIO, datos2$Cluster),2)*100


# En el cluster 1 se presenta 10 en el sector rural y presenta una minoria en otros barrios, mientras que en el cluster 2

# aparece pozon con 18, la esperanza con 14, nelson mandela con 12 y olaya con 32.


# Miremos por Zona

# En general

table(datos2$ZONA)
```

```
#se presentan 71 homicidios en la zona rural y 333 en la zona
urbana, por que ya sabemos que zona es la que deberia
reforzar su sistema de seguridad.


#Ahora miremos por cluster

table(datos2$ZONA, datos2$Cluster)

prop.table(table(datos2$ZONA , datos2$Cluster))

prop.table(table(datos2$ZONA, datos2$Cluster),2)*100


# Aca podemos encontrar una de las diferencias mas
significativas entre ambos cluster

# parece que los cluster se encuentran divididos por zonas, en
el cluster 2 podemos encontrar

# una dominacion total de la zona urbana con un 98.97 % de
los casos de homicidos,

# mientras que en el cluster 1 con una dominacion no tan
fuerte la zona rural ocupa un 61.26 % de los datos


# Miremos por clase de sitio

table(datos2$CLASE.DE.SITIO)

table(datos2$CLASE.DE.SITIO , datos2$Cluster)

prop.table(table(datos2$CLASE.DE.SITIO,
datos2$Cluster),2)*100


# segun estos datos se puede obserar que en el cluster numero
2 se presenta un  95.9 % de los homicidos

# en vias publicas, mientras que el cluster numero 1 ocupa un
58.5 % de los casos en vias publicas, pero

# tambien en fincas y similares con un 9.9 % y en carreteras
con un 10.8 %.


# Miremos por en que se movilizaba la victima
```

```
prop.table(table(datos2$MOVIL.VICTIMA))*100

table(datos$MOVIL.VICTIMA)

# Podemos ver que en forma general en nuestro dataset el 86.6
% de las victimas iban a pie, un 3.21% como conductor de
mtoocileta

# un 1.7 iban en vehiculos.

prop.table(table(datos2$MOVIL.VICTIMA ,
datos2$Cluster),2)*100

# hay prevalencia en ambos cluster de que las victimas se
movilizaban a pie con un 90.9 y un 84.9 para el cluster 1 y 2
respectuvamente



# Miremos por movil agresor

table(datos$MOVIL.AGRESOR)

barplot(table(datos$MOVIL.AGRESOR), main = "Movil Agresor",
ylab = "Frecuencia", col = c("green", "red", "blue",
"purple","orange"))

prop.table(table(datos2$MOVIL.AGRESOR , datos2$Cluster),2) *
100


# La mayor proporcon de los agresores van a pie o en
motocicletas 74.5 y 75.7 para el c1 y c2.



# Miremos por genero

table(datos$GENERO)

table(datos2$GENERO , datos2$Cluster)

prop.table(table(datos2$GENERO , datos2$Cluster))

prop.table(table(datos2$GENERO , datos2$Cluster), 2)*100

# En ambos cluster se presenta un alto indice de homicidos en
hombres con un 91.8 y 96.5 para el C1 y C2.
```

```
# Ahora miremos por estado civil

table(datos$ESTADO.CIVIL)

prop.table(table(datos2$ESTADO.CIVIL, datos2$Cluster),2)*100

# Miremos por clase de empleado

table(datos$CLASE.EMPLEADO)

table(datos2$CLASE.EMPLEADO , datos2$Cluster)

prop.table(table(datos2$CLASE.EMPLEADO , datos2$Cluster),
2)*100

# aca podemos observar que la mayor proporcion de personas
asesinadas en el cluster numero 2

# eran empleados particulares o desempleados con un 83.2 % y
9.21 % de los datos.

# mientras que en el cluster numero 1, independiente tienen
un 55.85 %, empleado particular un 20.72 %.

# y en oficios como agricultor, comerciante o ama de casa se
reparten casi un 13%.

# Miremos por escolaridad

table(datos$ESCOLARIDAD)

table(datos2$ESCOLARIDAD , datos2$Cluster)

prop.table(table(datos2$ESCOLARIDAD , datos2$Cluster), 2)*100

# en el cluster nuero 1 quein pose la mayor proporcion de
datos es la primaria con un 59.45 %,y le sigue

# la secundaria con un 31.5 %, con solo un 2.7% en la
educacion superior, pero en el cluster numero 2 existe

# un dominio total por parte de la secundaria con un 97.61 %
de los datos y solo un 0.68 % de nivel superior.


# Hay algo que podemos resaltar es que la mayor proporcion

#de los homicidos normalemmte se encuentran en el cluster 2
```

```
# Realicemos algunos metodos de clasificacion para ver si
logran

# colocar a cada dato en su cluster


# KNN

summary(datos2)

# Debemos  dumificar antes de aplicar el knn

library(class)

# Dumificamos nuestros datos2 en donde añadimos la columna
cluster

datos2_dum_final <- dummy.data.frame(datos2)

head(datos2_dum_final)

# Creacion del train y el test


sam <- sample(2, nrow(datos2_dum_final), replace = TRUE, prob
= c(0.75, 0.25))

train_knn <- datos2_dum_final[sam == 1,]

test_knn <- datos2_dum_final[sam == 2,]


# aplicar el knn

prediccion <- knn(train = train_knn[,-308], test =test_knn[,-
308], cl = train_knn[,308] , k = 2)


# listo, la prediccion esta en el objeto prediccion

print(prediccion)


# ahora aplicaremos el death test, para ver si nuestra
prediccion es correcta

table(test_knn[,308], prediccion)
```

```
# Por lo que podemos ver la prediccion nos dio bastante bien, nos detecto de 73 miembros

# del cluster numero 2, todos los 73 miembros, y de los 26 del cluster 1 fallo en 6.


# la  precision del modelo es

mean(prediccion == test_knn[,308])*100


# con una prediccion del 93,9 %


# Decision trees

# Leemos los paquetes y las librerias necesarias para hacer el arbol de decision

library(rpart)

library(rpart.plot)   # Libreria para graficar el arbol


# Ya que el arbol puede introudcirse datos tanto numericos como categoricos no hay necesidad de dumificar

# Sacamos los datos de estudio o entrenamiento y los de test


train_arbol <- datos2[sam == 1,]

test_arbol  <- datos2[sam == 2,]


# Aplicamos el arbol.....¿ Para la funcion se deberian cambiar otras variables?????

funcion <- Cluster ~  MUNICIPIO + ZONA + CLASE.DE.SITIO + ARMA.O.MEDIO + MOVIL.VICTIMA + MOVIL.AGRESOR + ESCOLARIDAD + Intervalos_edades

Arbol <- rpart(funcion , method = "class", data = train_arbol)


# Revisemos la informacion del arbol y grafiquemos
```

```
print(Arbol)

rpart.plot(Arbol, extra = 4)


# Realizar la prediccion

Prediccion_arbol <- predict(Arbol, test_arbol, type =
"class")



# Visualizamos la matriz de confusion

table(Prediccion_arbol, test_arbol$Cluster)


# Por lo que podemos ver tiene una precision bastante alta,
por lo que podriamos decir

# que logra diferenciar en gran manera ambos cluster,
teniendo de 23 miembros del cluster 1

# logro predecirlos todos, y de 76 del cluster dos solo fallo
en 3.


# su precision fue

mean(Prediccion_arbol==test_arbol$Cluster)*100

# 96.9 %

# miremos el error del arbol

install.packages("pROC")

library(pROC)

roc(as.numeric(Prediccion_arbol),
as.numeric(test_arbol$Cluster))

 # EL AREA bajo la curva fue de 0.98
```

BIBLIOGRAFÍA

[1] «70 años de una violencia que no termina», *www.elcolombiano.com*. [En línea]. Disponible en: http://www.elcolombiano.com/colombia/70-anos-de-una-violencia-que-no-termina-FF8513546. [Accedido: 06-sep-2018].

[2] N. J. Garnica, O. Murillo, y Á. Marcela, «Exploration of sexual violence in the city of Bogota: application of a data mining technique», *Rev. Crim.*, vol. 53, n.º 2, pp. 145-173, dic. 2011.

[3] E. G. Krug y Weltgesundheitsorganisation, Eds., *World report on violence and health*. Geneva, 2002.

[4] R.- ASALE y R.- ASALE, «Diccionario de la lengua española - Edición del Tricentenario», *Diccionario de la lengua española - Edición del Tricentenario*. [En línea]. Disponible en: http://dle.rae.es/?id=brdBvt6. [Accedido: 24-jul-2018].

[5] M. Camara y P. Salama, «HOMICIDES IN SOUTH AMERICA: ARE THE POOR DANGEROUS?», *Rev. Econ. Inst.*, vol. 6, n.º 10, pp. 159-181, jun. 2004.

[6] and, «Declining Inequality in Latin America», *Brookings*, 30-nov-2001. .

[7] «Politica Social Sintesis No 1 Pobreza y Desigualdad en America Latina.pdf». .

[8] «Panorama Social de América Latina 2017», p. 210, 2017.

[9] «América Latina es en 2015 algo menos pacífica que el año pasado, según índice de paz», *RunRun.es*, 18-jun-2015. .

[10] «GBAV2011-Ex-summary-SPA.pdf». .

[11] «GBAV-2015-ExecSum-SP.pdf». .

[12] E. P. L. S.L, «infoLibre – Información libre e independiente», *infoLibre.es*. [En línea]. Disponible en: https://www.infolibre.es/index.php/mod.global/mem.error404. [Accedido: 22-jun-2018].

[13] C. I. JaenCortés, S. R. Aragón, E. F. Amorin de Castro, y L. Rivera Rivera, «Violencia de Pareja en Mujeres: Prevalencia y Factores Asociados», *Acta Investig. Psicológica*, vol. 5, n.º 3, pp. 2224-2239, dic. 2015.

[14] «Deuda externa de Colombia en enero de 2018». [En línea]. Disponible en: https://www.dinero.com/economia/articulo/deuda-externa-de-colombia-en-enero-de-2018/257090. [Accedido: 07-sep-2018].

[15] «06.pdf». .

[16] Á. R. A. González, F. Escobar-Córdoba, y G. C. Castañeda, «Factores de riesgo para violencia y homicidio juvenil», *Rev. Colomb. Psiquiatr.*, n.º 1, p. 20, 2007.

[17] «laviolenciasexualencolombia.pdf». .

[18] C. E. E. Tiempo, «Más de 12 mil mujeres víctimas de violencia sexual en Colombia», *El Tiempo*. [En línea]. Disponible en: http://www.eltiempo.com/archivo/documento/CMS-16602558. [Accedido: 22-jun-2018].

[19] «Informe-de-Género-Cartageneras-Cifras-y-Reflexiones.pdf». .

[20] «DelitosSexuales.pdf». .

[21] «Estado - Instituto Nacional de Medicina Legal y Ciencias Forenses». [En línea]. Disponible en: http://www.medicinalegal.gov.co/documents/10180/33604/2+Delitosexual.pdf/7abb468a-ddd4-4f85-b2ef-bec607dba06b. [Accedido: 22-jun-2018].

[22] Naleja, «la violencia actual en colombia: HOMICIDIOS», *la violencia actual en colombia*, 24-ago-2009. .
[23] Y. S. Zapata-Bedoya, «Caracterización de las lesiones personales no fatales en la Regional Noroccidente (Antioquia) 1996-2002 y Medellín, 2003-2006», *Rev. SALUD PÚBLICA*, p. 13, 2011.
[24] «Etapas del conflicto armado en Colombia: hacia el posconflicto - ScienceDirect». [En línea]. Disponible en: https://www.sciencedirect.com/science/article/pii/S1665857416300102. [Accedido: 09-may-2018].
[25] «Aplicación de la minería de datos para analizar los diferentes tipos de lesiones registrados en la división medico legal Distrito Fiscal de Piura periodo 2005-2014». [En línea]. Disponible en: http://repositorio.unp.edu.pe/handle/UNP/704. [Accedido: 27-jun-2018].
[26] «Modelo recursivo de reacción violenta en parejas válido para ambos sexos», n.º 105, p. 14, 2012.
[27] G. Rovira Alcocer, O. Vital Hernandez, y F. G. Pat Espadas, «Violencia de pareja: tipo y riesgos en usuarias de atención primaria de salud en Cancún, Quintana Roo, México», *Aten. Primaria*, vol. 49, n.º 8, pp. 465-472, oct. 2017.
[28] L. Fernández-González, E. Calvete, y I. Orue, «Mujeres víctimas de violencia de género en centros de acogida: características sociodemográficas y del maltrato», *Psychosoc. Interv.*, vol. 26, n.º 1, pp. 9-17, abr. 2017.
[29] J. E. Carranza Romero, X. Dueñas Herrera, y C. G. González Espitia, «Análisis empírico de la relación entre la actividad económica y la violencia homicida en colombia1», *Estud. Gerenciales*, vol. 27, n.º 119, pp. 59-77, abr. 2011.
[30] A. Fulchiron, «La violencia sexual como genocidio Memoria de las mujeres mayas sobrevivientes de violación sexual durante el conflicto armado en Guatemala1», *Rev. Mex. Cienc. Políticas Soc.*, vol. 61, n.º 228, pp. 391-422, sep. 2016.
[31] F. F. Arcaute-Velazquez, L. M. García-Núñez, H. F. Noyola-Vilallobos, F. Espinoza-Mercado, y C. E. Rodríguez-Vega, «Mecanismos de lesión en actos de violencia extrema», *Cir. Cir.*, vol. 84, n.º 3, pp. 257-262, may 2016.
[32] R. Tuesca y M. Borda, «Violencia física marital en Barranquilla (Colombia): prevalencia y factores de riesgo», *Gac. Sanit.*, vol. 17, n.º 4, pp. 302-308, ene. 2003.
[33] B. Sanz-Barbero, L. Otero-García, S. Boira, C. Marcuello, y C. Vives Cases, «Acción COST Femicide Across Europe, un espacio de cooperación trasnacional para el estudio y el abordaje del feminicidio en Europa», *Gac. Sanit.*, vol. 30, n.º 5, pp. 393-396, sep. 2016.
[34] A. Company y M. Á. Soria, «La violencia en la escena del crimen en homicidios en la pareja», *Anu. Psicol. Juríd.*, vol. 26, n.º 1, pp. 13-18, ene. 2016.
[35] K. A. Tobón, «Analizando la violencia después del conflicto: el caso de Guatemala en un estudio sub-nacional», *Rev. Mex. Cienc. Políticas Soc.*, p. 43.
[36] A. Varela Huerta, «La trinidad perversa de la que huyen las fugitivas centroamericanas: violencia feminicida, violencia de estado y violencia de mercado», *Debate Fem.*, vol. 53, pp. 1-17, may 2017.
[37] V. Nava-Navarro, D. Onofre-Rodríguez, y F. Báez-Hernández, «Autoestima, violencia de pareja y conducta sexual en mujeres indígenas», *Enferm. Univ.*, vol. 14, n.º 3, pp. 162-169, jul. 2017.

[38] C. Agustí, M. Sabidó, K. Guzmán, M. I. Pedroza, y J. Casabona, «Proyecto de atención integral a víctimas de violencia sexual en el departamento de Escuintla, Guatemala», *Gac. Sanit.*, vol. 26, n.º 4, pp. 376-378, jul. 2012.
[39] C. C. Gil-Borrelli, P. Latasa Zamalloa, M. D. Martín Ríos, y M. Á. Rodríguez Arenas, «La violencia interpersonal en España a través del Conjunto Mínimo Básico de Datos», *Gac. Sanit.*, jun. 2018.
[40] S. M. Frías, «Ámbitos y formas de violencia contra mujeres y niñas: Evidencias a partir de las encuestas», *Acta Sociológica*, vol. 65, pp. 11-36, sep. 2014.
[41] M. C. Ochoa Ávalos y F. C. Reillo, «La violencia contra las mujeres en la región occidente, México: Entre la inoperancia institucional y el conservadurismo social», *Acta Sociológica*, vol. 65, pp. 121-150, sep. 2014.
[42] S. Boira, P. Carbajosa, y R. Méndez, «Miedo, conformidad y silencio. La violencia en las relaciones de pareja en áreas rurales de Ecuador», *Psychosoc. Interv.*, vol. 25, n.º 1, pp. 9-17, abr. 2016.
[43] P. V. Britos, *Minería de datos basada en sistemas inteligentes*. Buenos Aires: Nueva Librería, 2005.
[44] «¿Cómo seleccionar las variables adecuadas para tu modelo?», *Máxima Formación*, 03-oct-2017. .
[45] «Algoritmos de clustering y búsqueda de asociaciones». [En línea]. Disponible en: http://rstudio-pubs-static.s3.amazonaws.com/285462_c6f85d1b9c784c17b0b87dc7e018ae3c.html. [Accedido: 07-sep-2018].
[46] Alejandrocassis, «Aprendizaje Supervisado», *Inteligencia Artificial 101*, 20-oct-2015. .
[47] «Los 10 Algoritmos esenciales en Machine Learning», *Raona*, 31-may-2017. .
[48] «Machine Learning para dummies», *Paradigma*, 15-mar-2017. [En línea]. Disponible en: https://www.paradigmadigital.com/techbiz/machine-learning-dummies/. [Accedido: 27-jul-2018].
[49] «¿Qué es Machine Learning? -». [En línea]. Disponible en: https://blog.ubits.co/que-es-machine-learning-2/. [Accedido: 25-jul-2018].
[50] «Datos Abiertos Colombia | Datos Abiertos Colombia», *la plataforma de datos abiertos del gobierno colombiano*. [En línea]. Disponible en: https://www.datos.gov.co/. [Accedido: 07-jul-2018].
[51] «Preguntas Frecuentes - FAQ | Datos Abiertos Colombia», *la plataforma de datos abiertos del gobierno colombiano*. [En línea]. Disponible en: https://www.datos.gov.co/Ciencia-Tecnolog-a-e-Innovaci-n/Preguntas-Frecuentes-FAQ/antb-69gi/data. [Accedido: 29-jun-2018].
[52] R. A. Fernández, «Domingo, el día de más asesinatos en Colombia», *www.elcolombiano.com*. [En línea]. Disponible en: http://www.elcolombiano.com/colombia/domingo-el-dia-de-mas-asesinatos-en-colombia-CH7977004. [Accedido: 29-jun-2018].
[53] «El Pozón y Olaya Herrera, los barrios más golpeados por el homicidio», *El Universal Cartagena*, 02-ene-2014. [En línea]. Disponible en: http://www.eluniversal.com.co/sucesos/el-pozon-y-olaya-herrera-los-barrios-mas-golpeados-por-el-homicidio-147184. [Accedido: 29-jun-2018].
[54] «Delincuencia y ciudad. Hacia una reflexión geográfica comprometida». [En línea]. Disponible en: http://www.ub.edu/geocrit/b3w-349.htm. [Accedido: 29-jun-2018].

[55] «Homicidio.pdf». .
[56] «¿Cómo funciona el mercado ilícito de las armas de fuego en Cartagena? | EL UNIVERSAL - Cartagena». [En línea]. Disponible en: http://www.eluniversal.com.co/cartagena/como-funciona-el-mercado-ilicito-de-las-armas-de-fuego-en-cartagena-233322. [Accedido: 29-jun-2018].
[57] «ANÁLISIS ECONÓMICO DE LA VIOLENCIA EN COLOMBIA. UNA NOTA SOBRE LA LITERATURA». [En línea]. Disponible en: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0121-47722001000100009. [Accedido: 26-jul-2018].
[58] «Homicidio y lesiones infligidas intencionalmente por otra persona....» [En línea]. Disponible en: http://www.redalyc.org/html/806/80632301/. [Accedido: 29-jun-2018].
[59] «Estudio médico-legal del homicidio en la provincia de Sevilla (2004-2007): Especial referencia a los homicidios de mujeres en el contexto de violencia de género». [En línea]. Disponible en: http://scielo.isciii.es/scielo.php?pid=S1135-76062008000100005&script=sci_arttext&tlng=pt. [Accedido: 29-jun-2018].
[60] «Homicidios.pdf». .
[61] «PERVERSI-tesisdegradoeningenieria.pdf». .
[62] C. E. E. Tiempo, «Estas son las razones de la caída histórica de homicidios en Medellín», *El Tiempo*, 17-abr-2017. [En línea]. Disponible en: https://www.eltiempo.com/colombia/medellin/caida-historica-de-muertes-violentas-en-medellin-78542. [Accedido: 07-sep-2018].