



DETECCION DE DRIVERS QUE DEFINEN PATRONES DE VICTIMIZACIÓN DEL ENTORNO EDUCATIVO

Pedro Alberto Polanco Diaz.^a
pedro.polanco@usantotomas.edu.co

Director tesis: Alex Johhan Zambrano.^b
alex.zambrano@usantotomas.edu.co

Resumen

La Secretaria de Educación del Distrito de Bogotá lideró la Encuesta de Clima Escolar y Victimización Estudiantil, con el fin de conocer sobre el clima escolar y la seguridad de las escuelas. El presente documento tiene como objetivo determinar los drivers que definen patrones de victimización en los entornos educativos, seleccionando las variables que se relacionan directamente con estos hechos. Para ello, se realizó el procedimiento de imputación múltiple con el fin de completar la información faltante de la base, posteriormente se realizó un análisis de componentes principales y un análisis de clasificación, resultados que arrojaron tres grupos de colegios los cuales se definieron como Sanos, Afectados y críticos; para cada grupo se realiza una categorización de manera que se describan los aspectos o fenómenos que mejor representan cada grupo de colegios. Finalmente estos grupos se tomaron como variable respuesta para realizar un árbol de clasificación y su validación, de esta manera identificar los drivers de victimización se presentan en los entornos educativos.

Abstract

1. INTRODUCCIÓN

En los últimos años, se han realizado una serie de estudios que pretenden conocer sobre el entorno social de los centros educativos y los principales aspectos que relacionan el comportamiento que se percibe sobre los estudiantes, por esta razón la Secretaria de Educación del Distrito de Bogotá ha liderado una serie de encuestas la cuales tienen como objetivo establecer una línea base del estado sobre el clima escolar y la seguridad en las escuelas. Los resultados de las encuestas son de fácil acceso al público y han sido utilizados para realizar diferentes investigaciones bajo diversas metodologías, entre ellos se encuentran las publicaciones realizadas por la Secretaria de Educación, como el boletín de prensa (Boletín 2014) y la publicación denominada "Educación para la Ciudadanía y Convivencia". Estos resultados se muestran para diferentes análisis en aspectos como el sexo, tipo de colegio, maltrato de índole físico, psicológico, así como sucesos de victimización, haciendo énfasis en la evolución de dichos aspectos para los años 2006, 2011 y 2013. El presente documento tiene como objetivo determinar los drivers que definen patrones de victimización en los entornos educativos, para ello se tomará las variables que se relacionan directamente con estos hechos. Esto se realizará empleando técnicas conocidas en la literatura de la estadística multivariada y la minería de datos, estableciendo como unidad de observación los colegios, de manera que estos agrupen la información manifestada por cada estudiante y

^aEstudiante de estadística Universidad Santo Tomás Bogotá

^bProfesor de la facultad de estadística Universidad Santo Tomás

de esta manera apreciar un panorama de los colegios de Bogotá. En este sentido se hará un procedimiento de reducción de dimensiones de las variables seleccionadas utilizando el análisis de componentes principales, análisis clasificación jerárquica que agrupe los colegios que cuenten con características similares y finalmente estos resultados serán utilizados para realizar un árbol de clasificación que defina los patrones o conductores de victimización que se presentan en los entornos educativos.

Para referimos a victimización estudiantil, se hace claridad sobre algunos conceptos como violencia y victimización; la primera se puede clasificar en dos tipos: violencia verbal y no verbal. La violencia no verbal está directamente relacionada con los sucesos de victimización, muchos autores la han definido como un acto que expone la integridad física de una persona, donde el agresor utiliza armas letales y no letales, o en su defecto expone su cuerpo para ocasionar algún tipo de daño. Según la Organización de las Naciones Unidas (ONU 1986) menciona que una Víctima es "aquella persona que ha sufrido un perjuicio como consecuencia de una acción u omisión que constituya un delito con arreglo a la legislación nacional o del Derecho Internacional".

Uno de los pioneros en investigar los fenómenos que se relacionan con los casos de victimización estudiantil es Dan Olweus, quien ha hecho importantes investigaciones sobre maltrato escolar, como la elaboración de dos encuestas a estudiantes de Noruega, la primera fue en el año 1983 y la segunda en el año 2001. Para las dos encuestas se emplearon las mismas preguntas y sus resultados mostraron un incremento en aspectos sociales que afectan la convivencia estudiantil como son el acoso y la victimización escolar. Posteriormente sus estudios sobre violencia escolar fueron tomando como referencia en una gran variedad de países de todos los continentes como es el caso de Japón, Reino Unido, Francia, España, y Estados Unidos, entre otros, estos estudios fueron fundamentales en la estructuración de políticas nacionales. Para Colombia según (Valencia, "Conflicto y violencia escolar en Colombia" 2004), manifiesta que los estudios realizados fueron resultados de cuestionamientos académicos en universidades, cuyas publicaciones se hicieron en revistas indexadas y extranjeras. Así mismo, Valencia afirma que existen organizaciones como FECODE, (Federación Nacional de Trabajadores de la Educación) escritores e investigadores que han hecho planteamientos generales sobre los sucesos que ocasionan la victimización estudiantil. Sin embargo, la Secretaría de Educación de Bogotá ha venido liderando una serie de encuestas cuyo fin es conocer aspectos importantes que influyen en la victimización escolar. Estas encuestas tiene como objetivo la población de estudiantes de los grados 6° a 11° en colegios públicos y privados. La primer fue realizada en el año 2006 en cooperación de la Universidad de los Andes. Posteriormente, se realizó en el 2011 en conjunto con el Departamento Nacional de Estadística - DANE y finalmente en el 2013 se realizó la última encuesta en cooperación con la Universidad Distrital Francisco José de Caldas y su instituto de extensión y evaluación para el trabajo y el Desarrollo Humano (DEHUD). Estos instrumentos se encuentran alineados con las políticas gubernamentales ya que estas establecen como prioridad el desarrollo adecuado de la educación. Según la Secretaría de Educación de Bogotá los resultados de la encuesta 2013 fueron vitales para adoptar medidas en los programas como orientación escolar, acoso, matoneo, entre otros; sin embargo dada la necesidad de conocer las nuevas modalidades que afectan el desarrollo académico, los cuestionamientos realizados en cada encuesta han presentado ciertas modificaciones buscando identificar la evolución de los fenómenos que se presentan al interior y exterior de los colegios, por esta razón se plantearon preguntas en el año 2013 sobre la presencia de armas, alcohol, pandillas y drogas, así como acoso, matoneo y acoso sexual, tanto en el colegio como en sus entornos.

En la práctica y la literatura se encuentran varios investigaciones que muestran los resultados para este tipo de encuestas; una de ellas es el estudio **Violencia contra la Mujer, Análisis en el instituto materno infantil de Bogotá, Colombia 2015, Estudio de corte trasversal** (Milena López, Iván Sánchez e Íngrid Rodríguez 2015) Documento que muestra los resultados de un estudio que se realizó en el Instituto Materno Infantil de la ciudad de Bogotá, el cual tomó una muestra a 215 pacientes de sexo femenino mayores de 18 años, en los periodos de febrero de 2005, hasta el 31 de agosto de 2005. La recolección de la información excluida a los individuos si cumplían ciertas caracterizas o condiciones, así mismo la información objetivo se dividió en dos, la primera son los aspectos relacionados con síntomas de violencia y la segunda son aspectos demográficos. Para realizar el respectivo análisis de la información se emplearon técnicas del análisis descriptivo de los datos, y técnicas de la estadística multivariada, como el **Análisis de clasificación, Correspondencia y Factorial**, La técnica de correspondencias múltiples, el cual se empleó para resumir la información colectada en el cuestionario, posteriormente se combinó las metodologías de centros móviles y método de Ward, esa última para clasificar los pacientes con características diferentes, cuyos resultados permitieron minimizar la

inercia seleccionando un mínimo de clases, para ser tomadas como un centroide y aplicarlas en la metodología de k-means; Los resultados de los análisis realizados permitieron conocer ciertos patrones en la victimización de las mujeres, donde se resaltan aspectos como el maltrato psicológico, violencia física y violencia sexual, además de ello resaltan las principales causas que se asocian a estos fenómenos; este estudio conlleva realizar conclusiones sobre el panorama que presenta esta población, combinando aspectos de violencia física y sexual que fueron fundamentales para realizar un perfilamiento de las mujeres víctimas de violencia sexual entre las cuales se destacan nivel de estudio, extractos socioeconómico, principales causas que inciden en la afectación de la integridad de la mujer, entre otros. Así como adoptar programas de prevención.

Finalmente se hace necesario aclarar sobre la estructura que conlleva el presente documento, así: en la sección 2.) Se encuentran la descripción de los datos, aspectos generales sobre la Encuesta de Clima Escolar y Victimización. En la sección 3.) Se describe las técnicas de la estadística utilizadas en el presente análisis, Imputación de datos, análisis de componentes principales, clasificación jerárquica y finalmente árbol de clasificación. En la sección 4.) Se describe la selección de las variables y los resultados de la imputación de datos. En la sección 5.) Se muestran los resultados de la aplicación. En la sección 6.) se muestra la validación del modelo generado a partir de los árboles de clasificación, mostrando algunas conclusiones, recomendaciones y trabajos futuros.

2. Marco Referencial - Encuesta de Clima Escolar y Victimización Estudiantil

Para abordar el presente trabajo se cuenta con los resultados de la encuesta realizada por el Distrito de Bogotá denominada "Encuesta de clima escolar y victimización en Bogotá" (Encuesta - Secretaria de Educación Distrital 2013), la cual fue aplicada a estudiantes entre 6° y 11° en los colegios públicos y privados. Esta encuesta busca conocer la percepción que tienen los estudiantes sobre la convivencia dentro de los colegios, así como los factores que relacionan agresión, violencia y delincuencia; los cuales repercuten en el desarrollo óptimo de la educación para los jóvenes que viven en la capital de Bogotá.

Dicha encuesta se realizó en los colegios públicos y privados ubicados en 19 localidades de Bogotá, pertenecientes a las zonas urbanas; para ello se cuenta con un total de 623 colegios públicos, y los colegios privados de Bogotá. Para el levantamiento de la encuesta se tuvo en cuenta la siguiente metodología (Encuesta - Secretaria de Educación Distrital 2013)

- En los colegios públicos se seleccionaron al azar 22 estudiantes de cada grado entre 6º y 11º. En los privados 25 estudiantes de cada grado.
- Se encuestaron 118.000 estudiantes en total, que representaron a los 568.000 estudiantes, total de estudiantes de los colegios públicos y privados en esos grados.
- Del total de encuestados 86.500 son de colegios públicos y representaron a 377.000 estudiantes, y 31.400 de colegios privados, que representaron a 191.000 estudiantes.
- Las características de la muestra para cada situación se pueden leer en la tabla siguiente.

Con el fin de darnos una idea de como se encuentra actualmente los colegios de bogota conforme a la organización territorial, la figura 1 muestra la organización de los colegios conforme a los resultados de las pruebas saber.

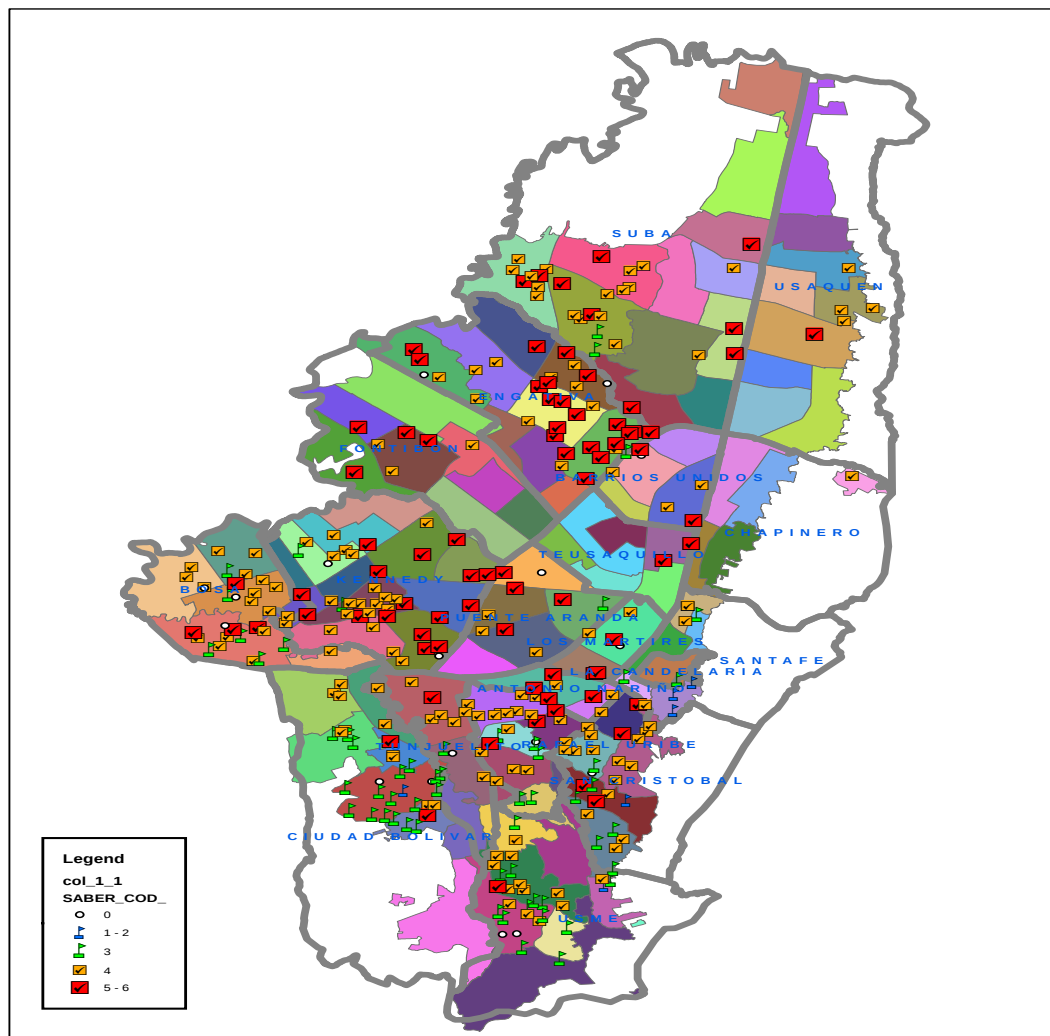


Figura 1. Mapa Ciudad de Bogotá con los resultados de la prueba saber pro, nformación suministrada por la Secretaria de Educación de Bogotá

3. Marco Teorico

3.1. Imputacion de datos Fatantes

La existencia de datos faltantes son un problema al momento de realizar tratamiento de la bases de datos, debido a que los métodos estadísticos están diseñados para el 100 % de los datos y al no contar con dicha disponibilidad tienen a ser susceptibles a interpretaciones erróneas, o en algunas ocasiones a levantar nuevamente la información lo que comúnmente conllevaría a aumentar los recursos utilizados para su levantamiento.

Como alternativa a estas necesidades, varios investigadores han dispuestos numerosos métodos que permiten realizar la imputación de datos faltantes para variables univariadas y multivariadas. Estos metodologias recomiendan realizar un análisis estructural de los datos a priori, realizando una descripción de la información faltante, siendo más útil al momento de establecer un criterio para la construcción del estimador, buscando afectar en lo más mínimo las medidas de tendencia central, ya que si se utiliza el método inapropiado podría cambiar las estructuras de las variables, sesgando la información o afectando la distribución de la misma.

La siguiente metodología sigue el libro "Imputación de Datos Teórico y Práctico", por esta razón si el lector desea profundizar sobre los metodos de Imputación se recomienda consultar en el libro escrito por Fernando Medina y Marco Galvan - 2007. Según (Medina - Galvan 2007) manifiesta que las bondades de los métodos de imputación no se deben valorar por el hecho que permiten completar información para ajustar modelos y probar hipótesis, sino que esta técnica debe considerarse como una prioridad en el proceso de la investigación; obteniendo conclusiones sustentables. El documento menciona que estimaciones como la media no es apropiada ya que las imputaciones pierden dispersión de los datos, lo que podría conllevar a sesgar la información a su media. Dicho autor recomienda que el método de estimación Hotdeck ha sido implementado exitosamente en censos y encuestas, ya que este método no tiene afectación sobre las distribuciones originales de las bases de datos, lo que hace que el metodo sea mas eficiente.

3.1.1. Método de Imputación simple Mediante una distribución no condicionada

Rubín fue uno de los pioneros en realizar clasificaciones a los modelos de perdida de datos, sus primeros métodos propuestos se basan en estimación simple (Medina - Galvan 2007), menciona que el método tiene como objetivo seleccionar valores aleatorios de los datos completos "donates" para imputar la información faltante "receptores", para ello identifica las características similares entre los receptores y donates evitando sesgos en el estimador de varianza.

El metodo utiliza un conjunto de variables correlacionadas entre los datos receptores y los donantes, donde se genera un criterio que identifique los valores adecuados para la imputación, este método tiene variantes, el método secuencial y el método aleatorio. El método secuencial, divide los datos en subgrupos donde cada uno es analizado de forma individual para encontrar los individuos donantes; las desventajas se presentan cuando un subgrupo cuenta con la mayoría de datos receptores este puede presentar sesgos en la información. El Método aleatorio, Puede elegir el donante a través de un procedimiento estocástico, en algunos casos cuando el donantes es un dato faltante, cambia la metodología eligiendo el vecino más cercano.

3.1.2. Método Imputación Multiple IM

Propuesto por Rubin (1987), parte de una simulación de Monte Carlo el cual combina distintos estimadores generados a partir de m imputaciones "se recomienda que m debe estar en un rango de 3 a 10"; donde cada imputación analizan la matriz de datos completamente para realizar los estimadores robustos, cuyo procedimiento se enuncia a continuación:

Considerando la notación de Schafer (1999), sea una variable $\hat{Q}$ y U su estimador de varianza. Una vez se genera las m imputaciones (con m puntos y estimaciones de varianzas del parametro) de los conjuntos mediante simulaciones, se tiene un estimador de $\hat{Q}$ y U .

$$\bar{Q} = \frac{1}{m} \sum_{i=1}^m \hat{Q}_i$$

Existen dos componentes de la varianza de $\bar{Q}$, donde la varianza de cada imputación es:

$$\bar{U} = \frac{1}{m} \sum_{i=1}^m \hat{U}_i$$

Y la varianza de los datos imputados es

$$B = \frac{1}{m-1} \sum_{i=1}^m (\hat{Q}_i - \bar{Q})^2$$

Por lo tanto la varianza total (T) es la varianza asociada con $\bar{Q}$ se obtiene sumando las dos expresiones anteriores, corrigiendo el numero finito de imputaciones por el estimador.

Valor $\frac{m+1}{m}$

$$T = \bar{U} + \frac{m+1}{m}B$$

Donde B/U indican cuanta información existe en los datos faltantes y se estima a partir de $\frac{\gamma}{1-\gamma}$, donde γ representa la relación de información que se pierde por falta de respuesta, cuando $B=0$ el intervalo de confianza se optiene por medio de:

$$\bar{Q} \pm t_{gl} \sqrt{T}$$

Y los grados de libertad de t se calculan como:

$$gl = (m-1)\left(1 + \frac{1}{r^2}\right)$$

$$r = \left(1 + \frac{1}{m}\right) \frac{B}{U}$$

El procedimiento de imputación Multiple se describe a continuación:

Considera una matriz de datos X la cual se puede descomponer $X = (X_{obs}, X_{mis})$, donde X_{obs} son los datos observados y X_{mis} son los datos sin información, ahora supongamos que tenemos una variable aleatoria Q y (X_{obs}, X_{mis}) es el valor estimado de Q que se Genera a partir de $U = U(X_{obs}, X_{mis})$ el error estandar $\hat{e}$.

donde,

$$\left(\frac{\hat{Q} - Q}{\sqrt{U}}\right)(0, 1)$$

Entonces, suponiendo que existen $m > 1$ simulaciones de los datos $X = (X_{obs}, X_{mis})$ permite calcular el estimador de $\hat{Q}^{(i)} = \hat{Q}(X_{obs}, X_{mis}^{(i)})$ y sus errores estandar $\hat{U}^{(i)} = \hat{U}(X_{obs}, X_{mis}^{(i)}), (i, \dots, m)$, el estimador el estimador de Q es $\bar{Q}$ ya se denoto con anterioridad.

El error estandar de $\bar{Q} = B = (m-1)^{-1} \sum (\hat{Q}^i - \bar{Q})^2$, la varianza de cada imputación es $\bar{U} = (m)^{-1} \sum U^1$ y el estimador de varianzas $T = (1 + m^{-1})B + \bar{U}$.

3.2. Análisis de Componentes Principales

El Análisis de componentes principales (ACP), es comúnmente utilizado en la reducción de dimensiones, al momento de contar con un conjunto de $p_1, \dots, p_i$ variables, busca describir la mejor precisión posible con un conjunto de $r < p$, según (cuadras) los componentes principales son unas variables compuestas incorrelacionadas, tales que, unas pocas expliquen la mayor variabilidad posible. Para hacer esto el ACP maximiza la

varianza multivariada y encuentra que la mejor dimensión son los ejes que tienen como dirección los valores propios de la matriz en forma ordenada. De esta forma el primer eje retiene la mayor variabilidad posible y de manera descendiente para cada eje. Así los ejes pueden ser una combinación lineal de X_i , facilitando el cálculo de las coordenadas con las variables e individuos.

Para hacer este procedimiento, existen dos métodos la matriz de correlaciones y la matriz de covarianzas. La matriz de correlaciones se obtiene calculando los valores y vectores propios de la matriz de correlaciones, lo cual es llamado análisis normado. La matriz de covarianza se utiliza cuando las unidades de medida de las variables son fácilmente comparables, es decir, varianza grande se asocia con valores propios grandes y de manera sucesiva. para abordar de mejor forma el análisis de componentes principales se recomienda consultar el libro de (Cuadras 2007).

3.2.1. Cálculo de número de componentes

No existe una única regla para la elección del número de componentes, sin embargo, se mencionan las más comunes:

- Gráfico de sedimentación: se grafican los valores propios frente a un número i de componentes y se observa en que punto de i los valores propios empiezan a comportarse de manera similar sin presentar cambios bruscos.
- Varianza total acumulada: se acumula la varianza que está siendo explicada por cada componente hasta que el valor de esta sea superior o igual a un valor determinado previamente.
- Criterio de Kaiser: Retener las primeras m componentes tales $\lambda_m > 1$, donde m , son los valores propios de la matriz de correlaciones R , es decir las varianzas de los componentes.

3.3. Análisis de Conglomerados

En análisis Clúster o Clasificación Jerárquico, es una técnica utilizada que permite conformar grupos de una población de acuerdo a algunas características de homogeneidad, denominados clusters, donde cada uno de estos debe ser los más heterogéneos posibles entre cada sí. Estos métodos también se denominan clasificación automática o no supervisada y se agrupan en dos categorías algoritmos de partición y algoritmos jerárquicos.

Para la realización de un análisis jerárquico algunos autores recomiendan, primero realizar un análisis exploratorio de los datos, según Fuente (Santiago, Fuentes (2011)) se deben realizar las siguientes etapas en un análisis conglomerados:

1. **Selección de las variables:** de acuerdo a las características propias de la variable puede ser cuantitativa o cualitativa
2. **Elección de la medida de asociación:** estas medidas permiten hacer la relación entre variables e individuos, las cuales se distinguen por ser una distancia o similaridad.
3. **Distancia:** por ejemplo la euclidiana, los grupos conformados concentran individuos parecidos, de tal forma que la distancia entre ellos es mínima.
4. **Similaridad:** por ejemplo el coeficiente de correlación donde los grupos formados tienen alta similaridad entre ellos.
5. **Elección de Cluster:** Estos métodos de clasificación se dividen en dos jerárquicos y no jerárquicos, así:

Metodos Jerarquicos

Formar un nuevo cluster para maximizar la medida de similaridad o minimizar la distancia. Para esta clasificación se utiliza dos metodos la asociación o aglomerativos y disociativos. Donde finalmente esto permite la construcción de un arbol de clasificación.

Entre los metodos aglomerativos de asociación encontramos

- Vecino mas cercano
- Vecino mas lejano
- Promedio entre grupos
- Metodo del centroide
- Metodo de la mediana
- Metodo de Ward

Para los metodos disociativos encontramos.

- Linkage simple
- Linkage completo
- Promedio entre grupos
- Método del centroide
- Método de la mediana
- Metodo de ward
- Analisis de asociación

Metodos no Jerarquicos

Esta diseñados para clasificar individuos en una clasificación de k cluster, donde k se define a priori o se determina en alguna parte del proceso, dentro de estos procedimientos encontramos:

- K- Means
- Nubes Dinamicas
- Analisis Modal
- Metodo de Woldf
- Metodos reductivos: Analisis factorial tipo Q

Validación de resultados: Finalmente se debe hacer una validación de los resultados, y de ser necesario eliminar los datos atipicos.

3.3.1. Método Jerárquico

Estos métodos son comúnmente utilizados en datos cuantitativos, según Peña (2002) si todas las distancias son continuas la distancia más utilizada es la euclidiana y la decisión de estandarizar o no depende del analista ya que si no se realiza la estandarización la distancia depende de los valores más extremos superiores y si se estandariza damos un peso semejante a la variable, con independencia de su variable original.

Según Peña (2002) para definir las distancias entre un grupo AB y el grupo C; donde el primero se crea a partir de la fusión del grupo A con n_a elementos y el grupo B con n_b elementos y el tercer con n_c elementos, habitualmente se emplean algunas de las siguientes 5 técnicas, así:

1. **Vecino más cercano:** La distancia entre la conformación de nuevos grupos, es la menor distancia entre grupos antes de la fusión, es decir:

$$d(C;AB) = \min(D_{CA}; D_{CB})$$

2. **Vecino más lejano:** La distancia entre la conformación de nuevos grupos, es la mayor distancia entre grupos antes de la fusión, es decir:

$$d(C;AB) = \max(D_{CA}; D_{CB})$$

3. **Media de Grupos:** La distancia entre la conformación de nuevos grupos, es el promedio ponderado distancia entre grupos antes de la fusión, es decir:

$$d(C;AB) = \frac{n_a}{n_a + n_b} d_{CA} + \frac{n_b}{n_a + n_b} d_{CB}$$

4. **Método del centroide:** método diseñado exclusivamente a variables continuas, la distancia entre grupos se hace igual que la euclidiana entre sus centros, tomando como centros los vectores de observaciones que pertenecen a cada grupo, de forma que al unir los grupos se puedan calcular nuevas distancias entre ellos sin utilizar las variables originales.

$$D^2(C;AB) = \frac{n_a}{n_a + n_b} d_{CA}^2 + \frac{n_b}{n_a + n_b} d_{CB}^2 - \frac{n_a n_b}{(n_a + n_b)^2} d_{AB}^2$$

5. **Método de Ward:** el propuesto diseñado por Ward y Wishart que se diferencia con los métodos anteriores porque no utiliza la matriz de distancias partiendo de cada elemento directamente definiendo una medida global de heterogeneidad agrupando observaciones en grupos, esta medida es la suma de elementos al cuadrado entre el elemento y la media de su grupo.

$$W = \sum_g \sum_{i \in g} (X_{ig} - \bar{X}_g)' (X_{ig} - X_g)$$

Donde $\bar{X}_g$ es la media del grupo g . El criterio comienza suponiendo que cada dato forma un grupo, $g = n$, por tanto W es cero. A continuación se unen los elementos que produzcan el incremento mínimo de W , tomando los más próximos con la distancia euclidiana. En la siguiente etapa $n - 1$ grupos, $n - 2$ grupos de un elemento y uno de dos elementos. Se decide de nuevo que grupo unir para que W crezca lo menos posible de esta manera se tiene la siguiente generalización:

$$\min \frac{n_a n_b}{n_a + n_b} (\bar{x}_a - \bar{b})' (\bar{x}_a - \bar{b})$$

Los métodos citados con anterioridad son útiles al momento de crear árboles de clasificación, llamados dendogramas o árbol jerárquico, estos métodos tienen la propiedad de que, si consideramos tres grupos A, B, C se verifica que.

$$d(A, C) \leq \max(d(A, B), d(B, C))$$

Una medida de distancia que tiene esta propiedad es ultramétrica. El dendograma es la representación de una ultramétrica y se constituye, así:

- En la parte inferior del gráfico se disponen los n elementos iniciales.
- Las uniones entre elementos se representan por tres líneas rectas. Dos dirigidas a los elementos que se unen y que son perpendiculares al eje de los elementos y una paralela a este eje que se sitúa al nivel en que se unen.
- el proceso se repite hasta que todos los elementos estén conectados por líneas rectas.

Si se corta el dendograma a un nivel a un nivel de distancia dado, se obtienen una clasificación del número de grupos existentes a dicho nivel y los elementos que lo conforman. Siendo útil cuando permita identificar el número de particiones que deben hacerse sobre la población. El número mismo se definirá cuando los niveles no presenten diferencias significativas en la ganancia de la varianza acumulada.

3.4. Árboles de Clasificación

Los árboles de clasificación hacen parte de una metodología desarrollada en los años 80 por Breiman (1984) y requiere, a priori, de una clase o una variable a explicar. Este modelo jerárquico divide el espacio en particiones, con una estructura fácil de comprender, se componen de nodos y hojas. Un nodo es una rama del árbol, se puede definir como un conjunto de decisiones y cada una de estas permite generar nuevas reglas que buscan realizar una clasificación, es decir, una condición que divide la muestra por unas características similares. Los nodos pueden ser padres o hijos, los padres están en la cabeza del árbol maximizando el criterio de división, seguidos por los nodos hijos que siguen un proceso de partición de manera independiente y finalmente los nodos (hojas) que contienen las predicciones dadas por un conjunto de características.

Esta metodología que divide rápidamente los datos muestrales, donde cada división es mutuamente excluyente, a través de una combinación entre las preguntas seleccionadas, busca encontrar la mejor división que a su vez debe contar con la mayor homogeneidad posible. Para realizar dicho análisis se (Timofeev, 2204) recomiendan construir el árbol máximo, podar el árbol y finalmente seleccionar el árbol óptimo mediante un proceso de validación cruzada.

▪ Construcción del árbol máximo

Para la construcción del árbol se utiliza un procedimiento de partición binario, se inicia con la raíz del árbol y se efectúa con los datos de entrenamiento que son los datos objeto de análisis, ajustando los niveles en cada uno de los nodos. El primer nodo se utiliza el criterio de optimización de un nodo, llamado función de impureza, esta función permite detectar la calidad de un nodo, supongamos que se calcula el nodo t , donde:

$$i(t) = \theta(p_1(t), \dots, p_j(t))$$

Donde $p_j(t)$ es la probabilidad condicional de que un elemento pertenezca a la clase j , dado que el nodo (t) y se estima como la proporción de elementos de la clase j en el nodo t , a continuación se denotan algunas de las funciones de impureza:

Indice de Ginni

$$iG(t) = 1 \sum_{j=1}^j (p_j(r))^2$$

Medida de Entropía

$$iE(t) = \sum_{j=1}^J P_j(t) \log(p_j(t))$$

Esta metodología permite calcular la variable explicativa con mayor impureza, la cual es usada como nodo padre, seguido de aquellas variables que maximicen el cambio de impureza, el cual se define como:

$$\triangle_i(t) = i(t_p) - E(I(t_c))$$

Donde, t_c son los nodos hijo de la izquierda y la derecha del nodo padre t_p , Ahora asumiendo que P_l y P_r son probabilidades de dichos nodos, con una función de impureza Bernoulli, donde solo toma los valores si o no, se obtiene:

$$\triangle_i(t) = i(t_p) - P_l i(t_l) - P_r i(t_r)$$

Por lo tanto, cada nodo CART resuelve el problema de maximización, así:

$$\operatorname{argmax}[i(t_p) - P_l i(t_l) - P_r i(t_r)]$$

El estimador del error es la proporción promedio de mala clasificación mediante validación cruzada:

$$R_1^{uc}(A_1) = \sum_{t \in A_1} r(t) p_j(t)$$

Además de las funciones citadas anteriormente, existen más funciones de impureza como la regla de Twoing, X^2 y la máxima desviación. Sin embargo se ha demostrado que la elección del árbol final es insensible de la regla de la división (Breiman (1984)).

■ Podar el arbol

Pensemos en un árbol con todas las hojas y nodos posibles, si tenemos un árbol sobreajustado donde las condiciones están a su máximo nivel (hojas), este conllevaría a un nivel de complejidad muy alta, siendo necesario realizar un proceso de parada y recorte, de manera que se optimice el arbol hasta encontrar el tamaño adecuado, si es necesario cortando los nodos insignificantes. Dicho proceso debe realizado antes de ser utilizado para la clasificación de nuevos datos.

Validación Cruzada

El método de validación cruzada consiste en tomar reiteradamente muestras y calcular con estos el error de predicción, encontrando el tamaño de árbol que maximiza el error de coste-complejidad definido como (Deconinck, 2006):

$$R_\alpha(T) = R(T) + \alpha(T)$$

Donde $R(T)$ es la tasa de mala clasificación, o el error de clasificación del árbol T ; $\alpha(T)$ es una medida de la complejidad que depende de T que es el número total de nodos terminales del árbol. El parámetro α se encuentra a través de la secuencia de las pruebas dentro de la muestra, donde una parte se utiliza para construir el modelo, la otra parte de los datos se toma como una muestra de prueba. Por cuanto es recomendable repetir m veces dicho proceso.

Cuando el árbol está en su máximo nivel, $R(T) = 0$ y a medida que se incrementa $R(T)$ se reduce el tamaño del árbol.

4. Metodología

Dentro de las técnicas que se relacionan en la literatura académica para realizar el tratamiento a los datos con características univariados, bivariados y multivariantemente, son los análisis multivalente y la Minería de Datos, los cuales tiene como objetivo analizar simultáneamente conjuntos de datos que relacionan en cierta medida características y objetos, así como el análisis de grandes volúmenes de datos bajo una estructura de inteligencia artificial, con aprendizaje automático y supervisado, que emplean diferentes metodologías para la dar respuestas a cuestionamientos en diferentes estudios.

El procedimiento que sigue el presente documento se puede visualizar en la figura 2, se fundamenta en dos estructuras clasificación supervisada y no supervisada, la cual busca realizar agrupaciones de individuos que se caractericen por ser homogéneos entre los individuos que conforman el grupo y heterogeneas entre cada grupo, con características únicas entre cada grupo de individuos, y la clasificación supervisada la discriminación toma estas características como referencia y se construye un modelo de clasificación. Para ello se realizaron los siguientes pasos:

- a) Tratamiento de la información (Imputación de datos faltantes)
- b) Selección de variables objeto de analisis.
- c) Agregación por colegio
- d) Elaboración del Análisis de Componentes Principales (ACP).
- e) Con el ACP se realizó el agrupamiento de individuos, utilizando la librería FactoMiner, la cual sigue lo propuesto por Lebart (1995).
- f) Se caracterizó cada uno de los clúster.
- g) Se tomó la clasificación como variable predictiva y se generó un árbol de desición, utilizando la librería Rpart, el cual sigue la metodología propuesta por Breiman, Friedman, Olshen and Stone (1984).
- h) Se hizo la validación cruzada y la matriz de confusión.
- i) Se plantearon las reglas y las estrategias para mejorar la percepción estudiantil y la victimización.

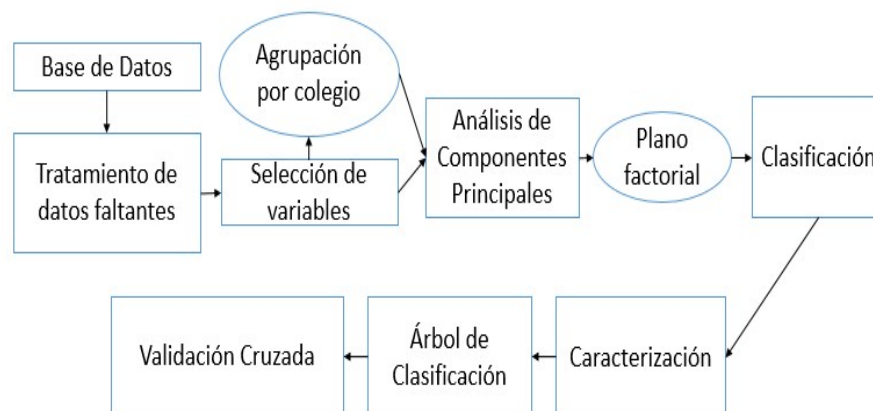


Figura 2. Metodología Utilizada

4.1. Base de datos

La encuesta cuenta con 115 preguntas y 117772 registros y esta dividida agrupaciones que tienen como objetivo:

- Realizar una descripción del estudiante y su entorno familiar.
- Conocer sobre el clima en el entorno educativo.
- Factores que provocan la victimización a los estudiantes (drogas, alcohol, pandillas, drogadicción, porte de armas seguridad)

Sin embargo, haremos referencia sobre el objetivo de la investigación, por este motivo se resalta la importancia de seleccionar adecuadamente las variables que permitan identificar los principales sucesos de victimización, en el presente documento no se hace un análisis descriptivo de las variables debido a que este trabajo fue realizado con anterioridad por la Secretaria de Educación, dicho análisis está en el boletín de prensa publicado en el año " Educación Para la Ciudadanía y Convivencia" (Encuesta - Secretaria de Educación Distrital 2013), bajo este contexto se tomaron en cuenta los siguientes módulos para la selección de variables que relacionan los principales aspectos como presencia de armas, alcohol, pandillas y drogas, así como acoso, matoneo y acoso sexual.

- **Como Eres**
- **Tu Colegio - Agresividad en tu colegio**
- **Acoso repetido**
- **Armas en el colegio**
- **Peleas en el Colegio**
- **Porte de armas**
- **Pandillas**
- **Alcohol y Drogas**

Antes de realizar el analisis para seleccionar las variables objeto de estudio, se encontraron datos de respuestas con codificaciones que no correspondían a alguna opción de la encuesta, estos datos fueron tomados como datos faltantes, lo que conllevo a utilizar un procedimiento estadístico que permitiera la imputación de estos datos; la relación de datos faltantes para las variables que se encuentran en cada uno de los módulos mencionados anteriormente se describe en la siguiente tabla:

Modulo	Porcentaje Datos faltantes
Como Eres	1,51 %
Tu Colegio, Agresividad en tu colegio, Acoso repetitivo	2 %
Armas y peleas en el Colegio	3,2 %
Pandillas, Alcohol Y Drogas	2,51 %

4.2. Tratamiento a los datos faltantes - Método de Imputación

Para el tratamiento de los datos faltantes, datos de Victimización, se emplearon numerosos procedimientos encontrados en la literatura de la estadística y con aplicación del Software R. Dichos procedimiento en algunos casos mostraron Imputación de datos de forma adecuada y en otros casos no presentaron los datos esperados:

- Paquete estadístico MICE y Amelia

Estos paquetes emplean un método de imputación múltiple, el algoritmo MICE, está diseñada para datos continuos, binarios y categóricos; donde cada variable implementa un modelo de imputación teniendo en cuenta su distribución y las variables que lo acompañan.

En cuanto al paquete estadístico Amelia este está diseñado para realizar imputación múltiple se ha comprobado que este paquete es mas potente que otros paquetes, en datos de series de tiempo o datos continuos o que provienen de una distribución normal.

Al momento de realizar la aplicación de estas imputaciones no se contó con los resultados esperados, ya que la cantidad de recursos que demandan para realizar la imputación en la base era demasiado grande en algunos casos se tardó casi dos días, como alternativa se dividió la base de datos en dos tomando como referencia las localidades del sur y norte. Posteriormente se volvió aplicar el método de imputación, sin embargo no fue posible contar con datos precisos, ya que las variables binarias las imputaba de forma continua, lo que generaba cambios en las distribuciones de las variables y en algunos casos generaba datos muy atípicos; por ejemplo la edad imputaba con datos superiores a 20 años.

Teniendo presente los resultados se descartó aplicar estas metodologías, siendo necesario emplear un nuevo método. Motivo por el cual se realizó la implementación de la metodología Imputación Múltiple propuesta por Rubín (1987), utilizando el paquete estadístico dispuesto en el Software R "HotDeackImputation", versión 1.1.

- Aplicación del Paquete estadístico Hotdeck

Una vez aplicada la Imputación de datos faltantes, se verificó que la estructura de los datos no se cambiaron las estadísticas descriptivas, conservando sus medidas de tendencia central. Para lo cual mostraremos algunos ejemplos de unas variables seleccionadas:

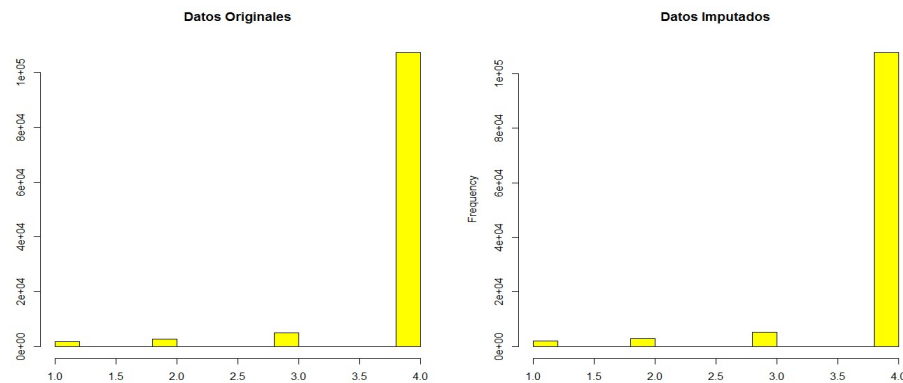


Figura 3. Dentro de tu colegio Alguien te amenazo con un arma datos imputados vs originales

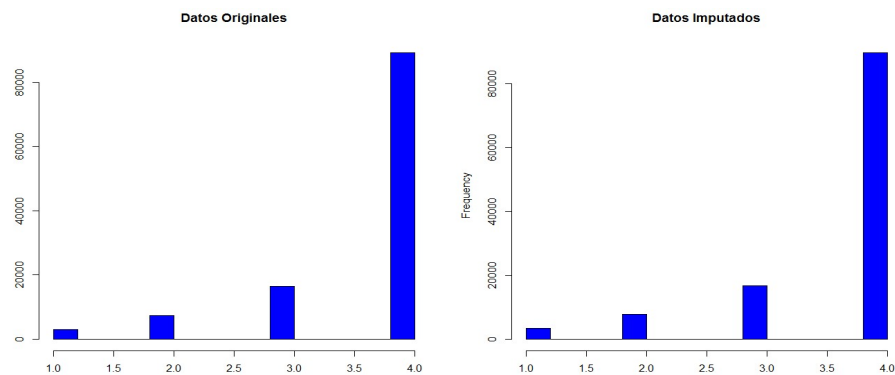


Figura 4. Peleas físicas dentro de tu colegio datos imputados vs originales

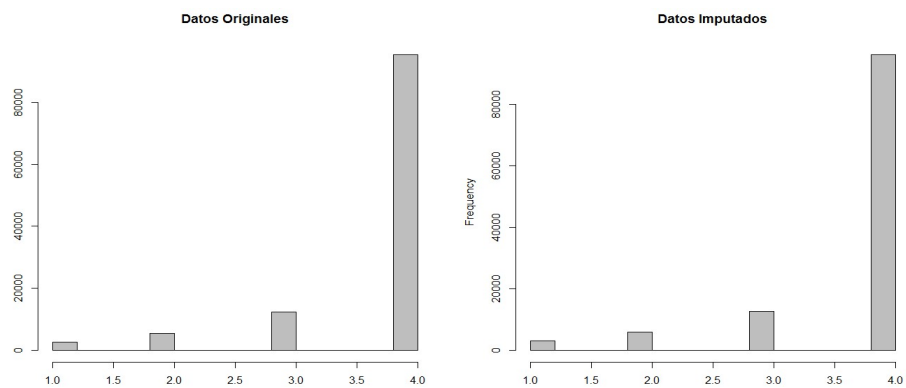


Figura 5. Peleas individuales de camino a casa o colegiodatos imputados vs originales

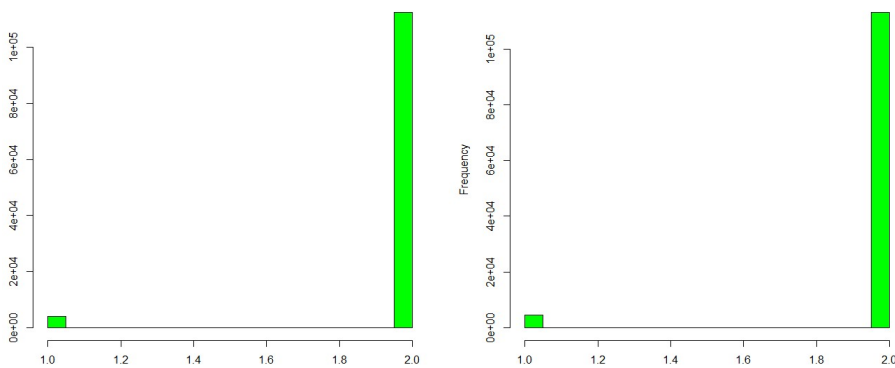


Figura 6. Camino de ida y vuelta a colegio alguien te forzo o intento tener relaciones sexuales datos originales vs imputados

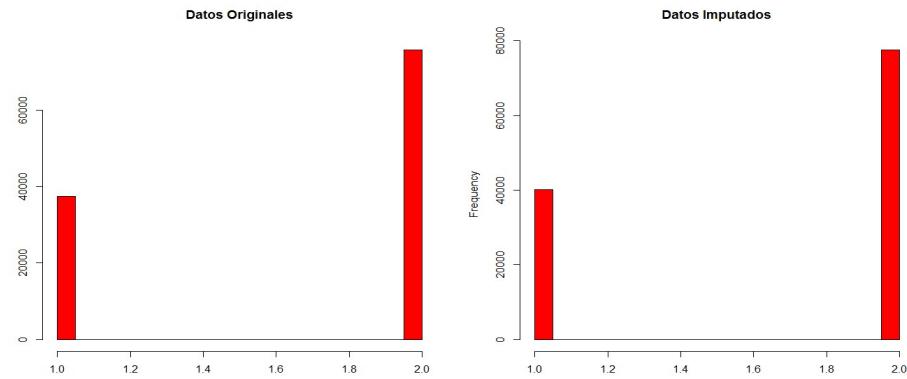


Figura 7. Peleas o ataques u otro tipo de violencia causada por pandillas datos originales vs imputados

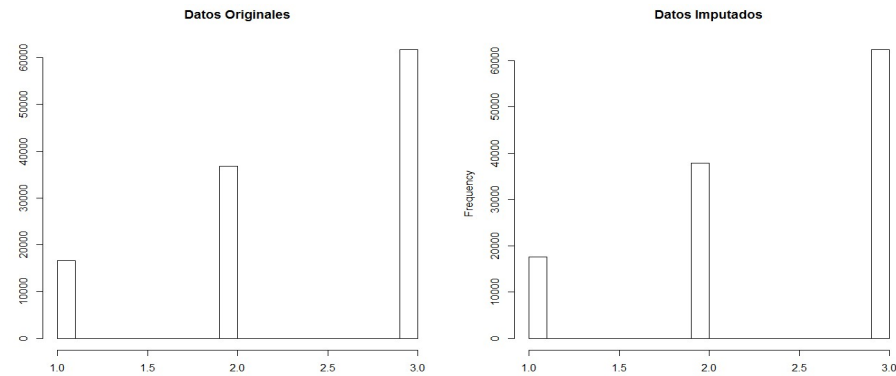


Figura 8. Se venden drogas dentro de tu colegio datos originales vs imputados

Una vez realizada la respectiva imputación de los datos faltantes se verificó que los datos no presentaran alteraciones, tal como se evidencia en los graficos; adicionalmente se realizo la prueba de Prueba de Kolmogorov, asi:

Prueba de Kolmogorov

El contraste de Kolmogorov-Smirnov, se utiliza para verificar si existe diferencia entre dos distribuciones, es decir verifica el nivel de concordancia existe entre en una distribución y otra distribución teórica especificada.

Estadístico de prueba es:

$$D_{n_1, n_2} = \text{Max} | F_1(x) - F_2(x) |$$

$$H_o : F_1(x) = F_2(x)$$

El criterio de aceptación o rechazo de la hipótesis nula es:

- si: P - Valor $> \alpha$ se acepta H_o
- si: P - Valor $\leq \alpha$ se rechaza H_o

Aplicada la prueba para las anteriores variables, se obtiene los siguientes p valor:

Variable	p valor
En el colegio amenazas con armas	5,845 %
Peleas física grupales en el colegio	19,8 %
Pelas de ida y vuelta peleas individual	8,8 %
Intento o relaciones sexuales a la fuerza	45 %
Ha presenciado peleas por pandillas	10,25 %
Venta de Drogas en el colegio	8,25 %

Para las pruebas se toma un nivel de confianza de 5%, y al momento de contrastar de las hipótesis, se comprueba que no hay evidencia estadística para rechazar las hipótesis nulas de las variables seleccionadas, garantizando que los datos no cambiaron las estructuras y que la distribución se mantiene; por cuanto se concluye que la imputación realizada no presento alteraciones sobre las variables. Teniendo certeza que la estructura de los datos no se ve afectada con la imputación aplicada, se procede a realizar la selección de variables bajo el siguiente argumento.

4.3. Selección de Variables

La selección de las variables utilizadas en la presente investigación, los cuales llamaremos "datos de victimización", obedecen a una estructura que se desea llevar de manera que permita identificar algunas acciones que sirvan para contrarrestar los fenómenos de violencia que se ven al interior y alrededor de los colegios, causando en gran medida los sucesos de victimización. Por esta razón los cuestionamientos realizados en la encuesta con preguntas como pertenencia pandillas, peleas o riñas, venta de drogas dentro y fuera del colegio, presencia de armas y consumo de alcohol, fundamentos que fueron importantes para determinar en qué medida afectan sobre la integridad estudiantil.

En los últimos años la puesta en marcha de las estrategias para la desarticulación de los principales expendidos de drogas o como se conocen comúnmente "ollas" en la ciudad de Bogotá, del cual se resalta el Cartucho, la L, o el Bronx, entre otras. Se ha generado la evolución de las estrategias empleadas por los delincuentes para la distribución, comercialización de las drogas ilícitas, una ella es la descentralización a parques, y lugares cercanos a los centros educativos, de manera que su comercialización sea de manera micro, lo que comúnmente se conoce como una de las modalidades del microtráfico; al igual que muchos estudios, el realizado en la Gobernación del Atlántico (Pandillas Juveniles), y el realizado en Bogotá "Informe ejecutivo sobre la situación de convivencia y seguridad escolar en la UPZ 47- Kennedy Central localidad OCHO" (Respuesta Integral y Orientación Ecolar) señalan que los grupos delincuenciales involucran a los jóvenes menores de edad para realizar en cierta medida el tráfico de sustancias psicoactivas, así como la venta de drogas, convirtiendo a estos jóvenes en jibaros, estos grupos se ven involucrados en conflictos para la disputa del control territorial.

Según el documento "Definición y Categorización de Pandillas" (Secretaría General de los Estados Unidos), menciona que en los casos en que las pandillas juveniles actúan en forma delictiva, participan en robos, tráfico de drogas, lesiones y asesinatos vulnerando los derechos de las personas. Una vez los jóvenes ingresan a una pandilla se generan en algunos casos, ciertas motivaciones que se orientan a continuar perteneciendo a estos grupos delincuenciales, el documento citado menciona que existen dos motivaciones; la primera es una motivación estructural y la segunda es una motivación personal, estas se fundamentan en que los jóvenes son débiles y en cierta edad están en búsqueda de una identidad, sobresaliendo una necesidad de ser reconocidos, lo que los convierte en personas débiles o sensibles a la atracción de grupos que comparten características que los motiven a desentenderse de las responsabilidades que existe la sociedad en común.

Por esta razón existe la necesidad de seleccionar las preguntas que se relacionan directamente con la presencia de pandillas, y la relación que hay entre el estudiante y estos grupos, además de ello conocer el impacto que se genera con la presencia de conflictos dentro y fuera del colegio y la presencia de drogas ilícitas. Ahora, si bien es cierto no se debe dejar a un lado el consumo de bebidas alcohólicas, ya que este en cierta medida esta ligado al consumo de drogas, según el estudio (La relación droga y delito en adolescentes infractores de la

ley), se logró comprobar que existe una alta relación entre las personas que consumen bebidas alcohólicas y las que consumen drogas especialmente "marihuana", y que a su vez la mayoría de las infracciones cometidas por los jóvenes objeto de estudio tienden a ser bajo estos efectos.

Es por ello que se realiza la selección de las variables que se relacionan a continuación, ya que se considera que estas se encuentra asociadas a las problemáticas que se evidencia día a día en los colegios, donde las principales víctimas de los fenómenos delictivos son los estudiantes. Permitiendo realizar el análisis directamente sobre las variables objeto de análisis de manera que la metodología utilizada en la presente investigación muestre reglas o drivers claros que permitan tomar decisiones que conlleven a contrarrestar los fenómenos que presenten síntomas de violencia estudiantil.

Variebles objeto de estudio

- a) **Dane:** Codigo del colegio
- b) **Localidad:** Donde se encuentra el colegio
- c) **Preguntas 59: El mes pasado un compañero te pego, o te amenazó con hacerte daño, haciéndote sentir muy mal**
- d) **Pregunta 62: En este año, dentro de tu colegio ¿cuantas veces alguien te amenazó con un arma?** Categoría de respuesta
- e) **Pregunta 65: En el mes pasado dentro de tu colegio, cuantas veces tuviste una pelea grupal física con otra persona** Categoría de respuesta
- f) **Pregunta 72: En este año, ¿trajiste ARMAS al colegio?**Categoría de respuesta
- g) **Pregunta 79: En el mes pasado de ida y vuelta a tu colegio, cuantas veces tuviste una pelea individual con otra persona** Categoría de respuesta

La amenazas y riñas o pelea, dentro y fuera de los centros educativos de Bogotá, en los últimos años ha presentado cierta evolución, según el (DANE 2012) realizo la publicación de la circular "Alarma en los colegios de Bogotá ", donde informa que los estudiantes encuestados han manifestado ser víctimas de amenazas y de peleas al interior y alrededor de los colegios. Este argumento se toma para la selección de estas variables.

Para las preguntas 65 y 79 tiene como opción múltiple 1) más de cinco veces, 2) más de cuatro veces, 3) una vez y 4) ninguna vez; por cuanto se realizará agrupaciones de dicha variable de manera que permita conocer si participo en alguna pelea o no, de esta forma se agruparon las opciones 1, 2, 3 como una única opción, de tal manera que agrupe los individuos que participaron en peleas grupales y los que no participan en dichas peleas al interior y exterior del colegio.

- h) **Pregunta 86: Este año, en el camino de ida y vuelta de tu colegio, ¿alguien te forzó o intento forzarte a tener una relación sexual sin que lo desearas** Categoría de respuesta
- i) **Pregunta 88: Perteneces a alguna pandilla** Categoría de respuesta

- j) **Pregunta 90: En este año: Ha visto dentro de tu colegio peleas, ataques, u otros tipos de violencia realizadas por pandillas** Categoría de respuesta
- k) **Pregunta 92: Se venden drogas en tu colegio** Categoría de respuesta
- l) **Pregunta 93: Se venden drogas cerca tu colegio** Categoría de respuesta
- m) **Pregunta 94: Ha presenciado a algun compañero de tu curso consumiendo drogas dentro del colegio?** Categoría de respuesta
- n) **Pregunta 97: Con qué frecuencia consumes bebidas alcohólicas dentro de tu colegio** Categoría de respuesta

Para el caso de estas preguntas se recodificaron con el fin de clasificar las variables categoriales, de esta manera las personas que afirman de manera positiva en la venta de drogas o consumo de bebidas alcohólicas, se encuentren en un solo rango.

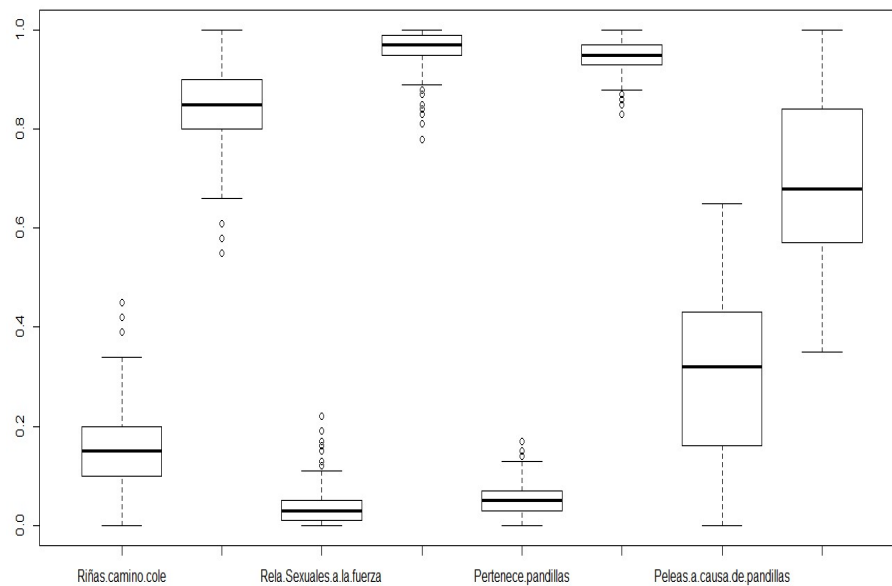
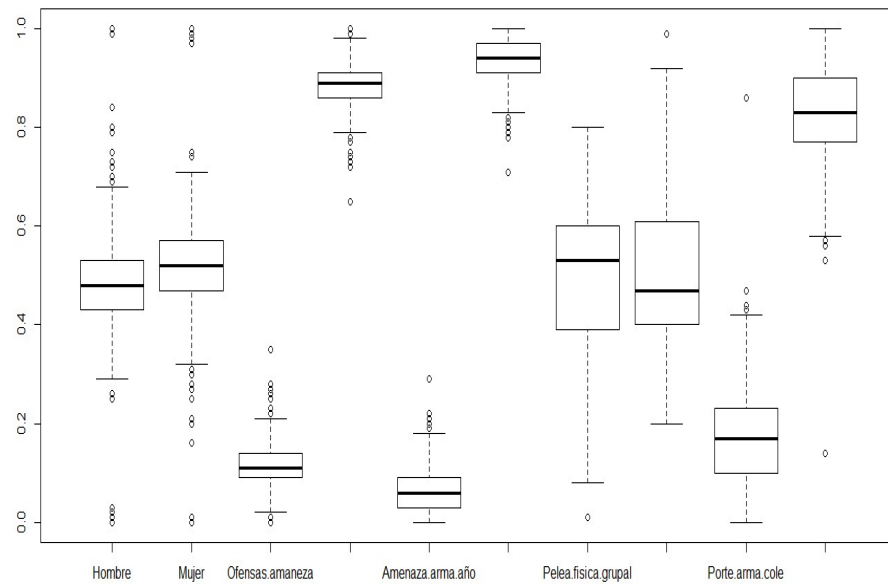
Posteriormente se procede a realizar los conteos por colegio, agrupando la cantidad de personas que contestaron en cada una de las opciones que permite las preguntas; debido a que los colegios serán las unidades de medida para el presente estudio. Con esto cada una de las opciones se convierte en una variable, objeto de estudio, luego se calculó la proporción por colegio debido a que esta se requiere tomar una unidad de observación para todos los colegios y de esta manera evitar sesgos de los colegios grandes en cuento a la cantidad de estudiantes.

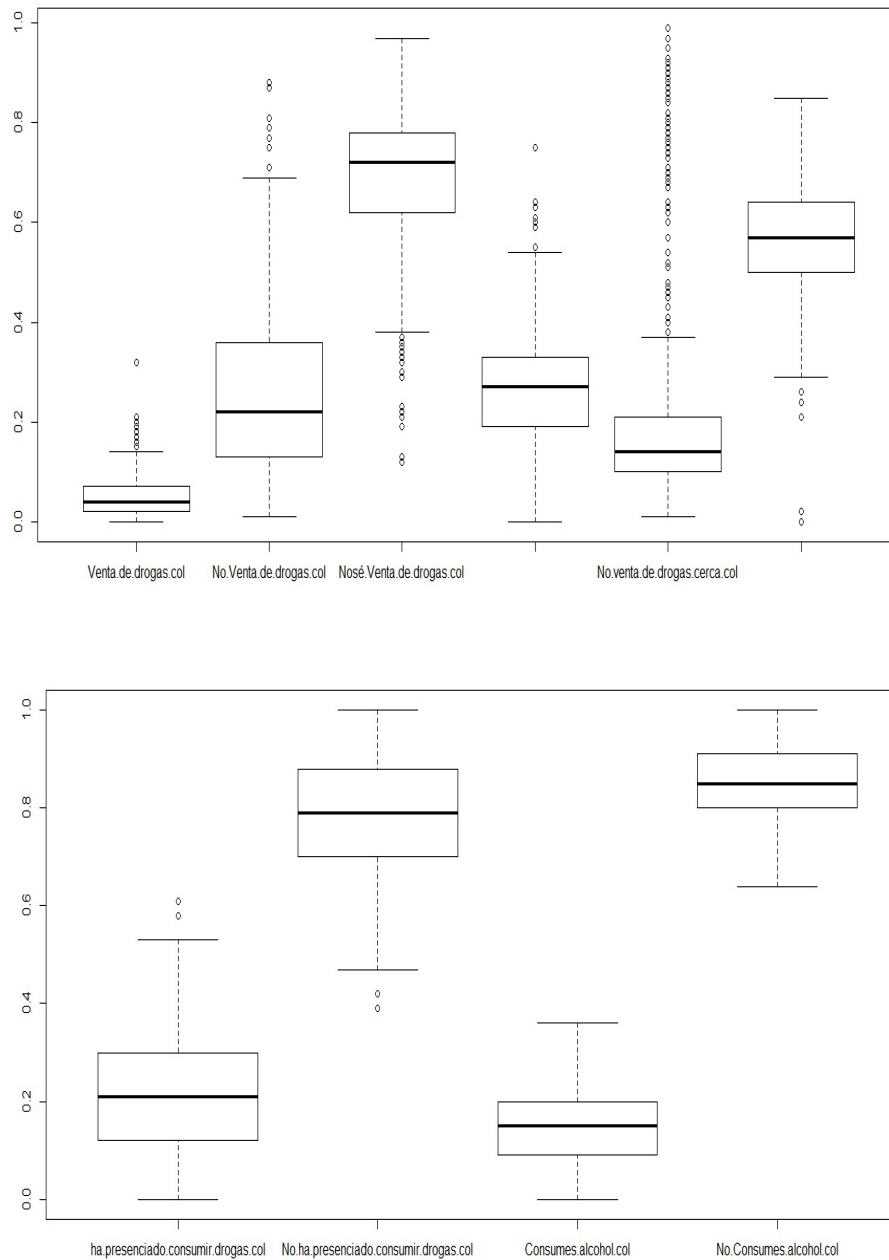
Finalmente, se realiza una agrupación por las localidades en cuatro zonas, donde cada una se compone de las localidades que se geográficamente se encuentran cerca, buscando así realizar un análisis más agregado de los colegios de Bogotá.

zona	localida
Central	La Candelaria Barrios Unidos Santa Fe Teusaquillo Antonio Nariño Los Mártires Chapinero
Norte Occidente	Usaquén Engativá Suba
Oriente	Tunjuelito Rafael Uribe Uribe San Cristóbal Usme Ciudad Bolívar
Sur occidente	Puente Aranda Kennedy Bosa Fontibón

5. Resultados de la aplicación

Los siguientes boxplot muestran la estructura de las variables respecto a su media, mediana y cuartiles, al igual se puede ver que las variables donde relacionan la proporción de individuos que niega "Sin" "Etiqueta" o desconoce la presencia de cada uno de los fenómenos, tiene el promedio más alto, por cuanto nos da indicios que para los colegios de Bogotá no se cuenta con altos fenómenos de victimización.





Para nuestro análisis se tomaran las variables que relacionan la proporción de estudiantes por colegio que afirmaron o manifestaron desconocer sobre cada uno de los cuestionamientos realizados. Esto se realiza con el fin de evitar problemas de datos composicionales ¹, ya seleccionadas las variables se realiza la reducción de las dimensiones, empleando la metodología de Análisis de Componentes Principales normado para estudiar estos datos que se conforman por 613 individuos (colegios) y 25 variables categóricas de las cuales, 4 son suplementarias.

¹Los datos composicionales son realizaciones de vectores aleatorios de suma constante

5.1. Reducción de dimensiones - Analisis de Componentes Principales

Ejes factoriales

El analisis de componentes principales se realizo, utilizando el paquete FactoClass, el cual sigue la metodologia de Lebart (1995), este desarrolla el metodo de correlaciones. se eligen los cuatro primeros componentes, debido a que la suma de ellos explica el 69.2% de la variabilidad total, estos cuatros componentes satisfacen lo indicado en el índice de Kaiser Meyer Olkin debido a que su valor propio es superior a uno (1), adicionalmente al tomar los cuatro primeros se presenta los mayores saltos "grafico de codo"; sin bien es cierto se pierde 30,8% de la variabilidad total, tomar un quinto o sexto componente no generan mayor ganancia en el análisis, y se corre el riesgo de no reducir de forma adecuada los datos para el análisis, lo cual no ayudaría a explicar los datos de la mejor forma.

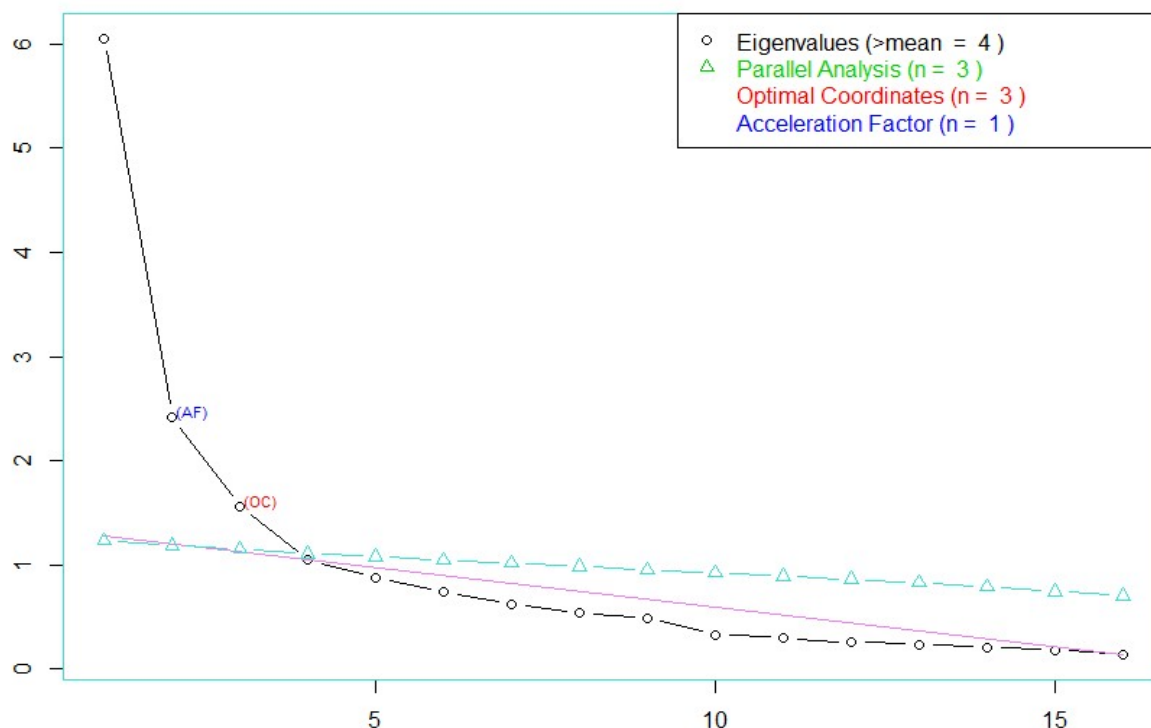


Figura 9. Eje Factorial

Circulo de correlaciones entre variables

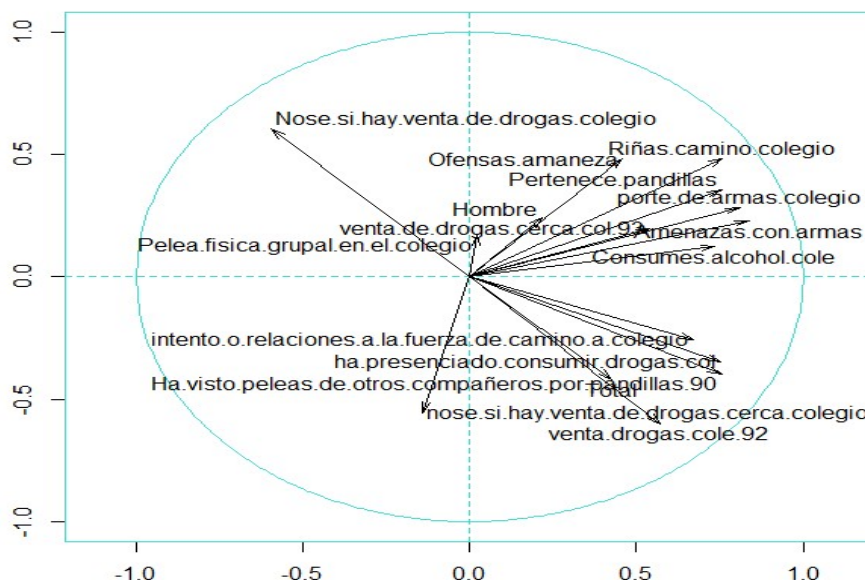


Figura 10. Grafico Factorial

Como se muestra en la figura 10, el Gráfico Factorial, se muestra el primer y segundo factor, donde se aprecia las variables que pertenecen al interior de los factores, y su direcciones se encuentran en el mismo sentido y normadas, la nube de variables se divide en dos grupos, el primer grupo se encuentra en el primer cuadrante de dicho mapa y son las variables que están correlacionadas como: la proporción de estudiantes que manifestaron presencia golpes o amenazas de parte de los compañeros, presencia de riñas de camino al colegio o a casa, pertenencia a pandillas, llevar armas al colegio, sexo femenino, venta de drogas cerca del colegio y consumo de bebidas alcohólicas.

Se observa un segundo grupo de variables, ubicados en el cuarto cuadrante, existen las variables relacionan la proporción de estudiantes que afirman intento de relaciones sexuales a la fuerza cuando se desplaza al colegio o la casa, estudiantes que han presenciado consumir compañeros drogas en el colegio, colegios grandes, y finalmente venta de drogas en el colegio.

Para las variables que se encuentran al interior de estos grupos se muestra alta correlación, donde el primer grupo de variables está altamente influenciado por la proporción de estudiantes, que manifiestan desconocer sobre venta de drogas cerca del colegio, así como pelea física o grupal de camino al colegio la cual está ubicada sobre el primer plano factorial (cercana a -1). Para el segundo grupo de variables está influenciado por que manifiestan desconocimiento de venta de drogas cerca del colegio.

5.2. Analisis de Clasificación

La clasificacion nos permiten hacer sencillos los análisis estadísticos, de esta manera se presenta la agrupación de los colegios cada uno de los grupos, así.

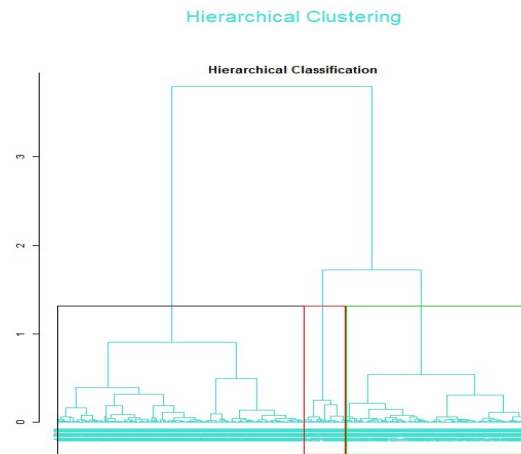


Figura 11. Clasificación

Usando el dendrograma de la figura 11, se observa que al seleccionar 3 grupos para la clasificación, presenta una buena partición de los datos, para lo cual procedemos a realizar una descripción de los tres grupos de seleccionados.

La figura 12 se muestran los grupos conformados sobre las dimensiones estructuradas por el ACP, estos grupos de colegios están bien representados en el plano ya que no se evidencia que se contraponga individuos, de manera que sus resultados son fácilmente comprensibles para el lector, y cuya interpretación se realiza a continuación. La grafica 13 muestra la distribución de colegios por cada uno de los grupos conformados y de los cuales se caracterizan en la siguiente sesión.

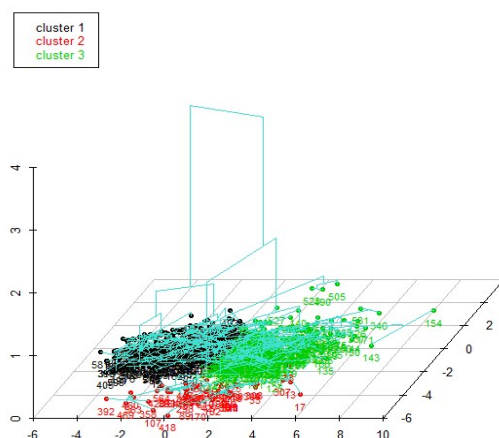


Figura 12. Agrupaciones Jerarquica de Componentes

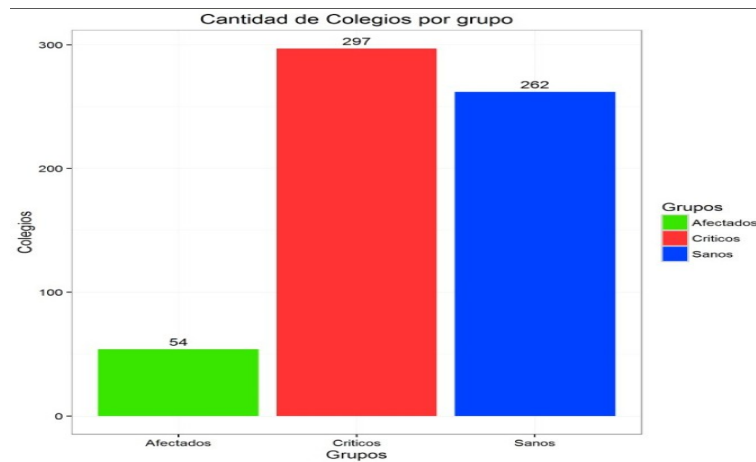


Figura 13. Colegios por Grupo

Caracterización de los componentes.

a) Primer Grupo

El primero grupo se caracteriza con la variables que relaciona los colegios donde la proporción de estudiantes que contestaron sobre desconocer si hay presencia de ventas de drogas en el colegio, siendo la variable que mejor caracteriza el grupo, se puede evidenciar que el promedio de la proporción de estudiantes es superior a la media superior a la global. Las dimensiones como: ha visto peleas en el colegio causadas por otros compañeros, ha presenciado consumir drogas por un compañero en el colegio, amenaza con arma de fuego, trajo arma al colegio, intento de relaciones sexuales en el colegio y venta de droga en el colegio, se contraponen de manera negativa a este grupo de colegios. Lo que nos da un indicio de que los jóvenes que estudian en estos colegios cuentan con niveles de percepción son bajos sobre hechos de victimización, ya que las variables que mejor caracterizan el Grupo, no se asocian a afirmaciones sobre los cuestionamientos realizados. Adicionalmente el promedio de las proporciones de las variables, está por debajo de la media global, dando siendo grupo de colegios con los que mejores niveles de percepción tienen, por lo que los denominaremos como grupo de colegios sin mayor afectación a sucesos de victimización "Sanos".

Variable	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
Nose si hay venta de drogas colegio	14,7	0,37	0,26	0,16	0,16	0,00
Hombre	-2,7	0,45	0,47	0,20	0,15	0,01
Ofensas amenaza	-6,6	0,10	0,12	0,04	0,04	0,00
Venta de drogas cerca colegio	-9,2	0,22	0,27	0,11	0,11	0,00
Total estudiantes	-11,1	657	933	358	531	0,00
Riñas camino colegio	-13,2	0,11	0,15	0,05	0,07	0,00
Pertenece pandillas	-13,3	0,03	0,06	0,02	0,03	0,00
Venta drogas colegio	-13,8	0,02	0,05	0,02	0,04	0,00
Consumes alcohol cole	-13,9	0,01	0,02	0,01	0,02	0,00
Intento o relaciones a la fuerza de camino a colegio	-14,6	0,01	0,03	0,01	0,03	0,00
Ha llevado armas al colegio	-15,7	0,10	0,17	0,05	0,09	0,00
Amenazas con armas	-15,8	0,03	0,07	0,02	0,05	0,00
Presencia de compañeros consumiendo drogas	-16,6	0,12	0,21	0,08	0,12	0,00
Ha visto peleas de otros compañeros por pandillas	-18,5	0,16	0,30	0,11	0,16	0,00

Debido a que en el análisis se tomaron tres componentes que se denominaran como "Sanos", "Afectados" y "Crítico", donde explicaremos los dos ultimos más adelante, se hace el análisis

de la variable "No sé si hay presencia de venta de drogas" para los tres componente, con el fin ver su comportamiento en cada Grupo.

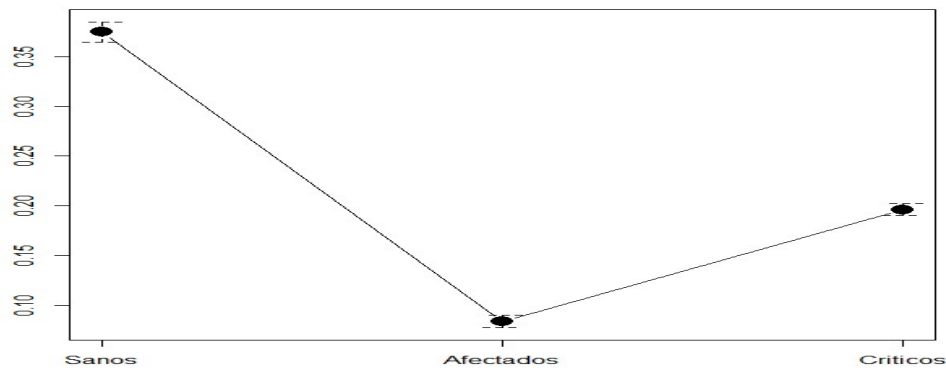


Figura 14. Nose si hay presencia de ventas de drogas en el colegio

Se realiza el Test LSD de Fisher, la cual tiene como objetivo determinar qué tipo de variables se parecen entre sí para cada Grupo o clúster, se basa en los intervalos de confianza y una prueba de comparación de medias, esta prueba busca detectar que las variables analizadas se parecen al interior de cada uno de los grupos analizados, descartando igualdad de medias. Los resultados se muestran con letras; donde las letras diferentes indican que existen suficientes argumentos estadísticos para determinar que el conjunto de individuos que conforma cada grupo se comporta de manera diferente, caso contrario pasa con las iguales.

Groups	Treatments and means
a	Sanos 0.3772587
b	crítico 0.1958
c	Afectados 0.08425926

Con estos resultados podemos concluir que la variable que relaciona los estudiantes que afirman sobre no saber si hay ventas de drogas en el colegio, tiene el promedio más grande en el clúster "sanos", adicionalmente el Test LSD de Fisher, indica que no hay suficientes pruebas estadísticas para determinar que los grupos de colegios son parecidos, deduciendo que no hay igualdad entre estos grupos.

b) Segundo Grupo

Las dimensiones que mejor caracterizan este grupo corresponden a la proporción de estudiantes que manifiestan desconocer sobre la venta de drogas en el colegio, seguido de intento de relaciones sexuales a la fuerza de camino al colegio, venta de drogas en el colegio, presencia de algún compañero consumiendo drogas, peleas ocasionadas por pandillas y finalmente colegios grandes en cuanto a la cantidad de estudiantes; esto nos indica que estos fenómenos de victimización son más comunes para este grupo de colegios, debido a la que la percepción de los estudiantes que afirmaron sobre estas preguntas son las más alta. Las dimensiones donde los estudiantes contestaron que desconocen si hay venta de drogas en el colegio, peleas u amenazas por compañeros y venta de drogas cerca, se asocian de manera negativa al componente. Debido a los resultados presentados en la clasificación y a que las variables que mejor describen el grupo se asocian a conductas que posiblemente que posiblemente generen afectación de la integridad de los estudiantes y teniendo presente que los niveles de percepción para estas variables se encuentran por encima de la media global de denominará este grupo de colegios con moderada afectación en relación a los sucesos de victimización analizados, como "Afectados".

Variable	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
No se si hay venta de drogas cerca colegio	22,1	0,8	0,21	0,11	0,2	0
Intento de relaciones a la fuerza de camino a colegio	13,2	0,1	0,03	0,04	0,03	0
Venta de drogas colegio	12,4	0,1	0,05	0,06	0,04	0
Presencia de peleas por pandillas	9,4	0,5	0,3	0,05	0,16	0
Presencia de compañeros consumiendo drogas	5,9	0,3	0,21	0,11	0,12	0
Total	2,4	1096,7	933	799	531	0,02
Ofensas amenaza	-2,4	0,1	0,12	0,04	0,04	0,02
Ha llevado armas al colegio	-2,4	0,1	0,17	0,07	0,09	0,02
Pertenece pandillas	-3,2	0	0,06	0,02	0,03	0
Venta de drogas cerca colegio	-3,9	0,2	0,27	0,11	0,11	0
Riñas de camino al colegio	-7	0,1	0,15	0,04	0,07	0
No se si hay venta de drogas en el colegio	-8,4	0,1	0,26	0,05	0,16	0

- No se si hay venta de drogas cerca al colegio?

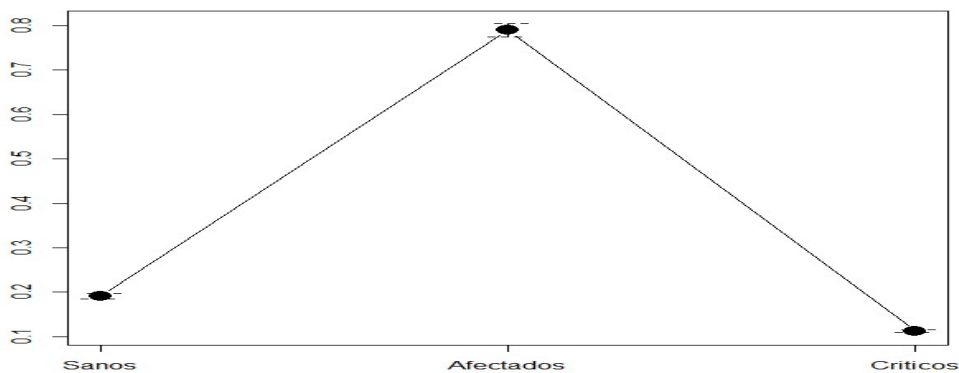


Figura 15. No se si hay venta de drogas cerca al colegio

Groups	Treatments and means
a	crítico 0.7896
b	Sanos 0.1913
c	Afectados 0.1123

La variable es muy parecida en los grupos "Sanos" y "Crítico", aparentemente no hay dispersión de los datos en cada uno de los componentes analizados, esta variable es la que mejor caracteriza el grupo de los "Afectados".

El resultado del Test LSD de Fisher ayuda a concluir que no hay suficientes pruebas estadísticas para decir que al menos un grupo de colegios de cada componente se parece a otro.

- Este año, en el camino de ida y vuelta de tu colegio, ¿alguien te forzó o intento forzarte a tener una relación sexual sin que lo desearas??

Groups	Treatments and means
a	Afectados 0.0875
b	críticos 0.0431
c	Sanos 0.01248

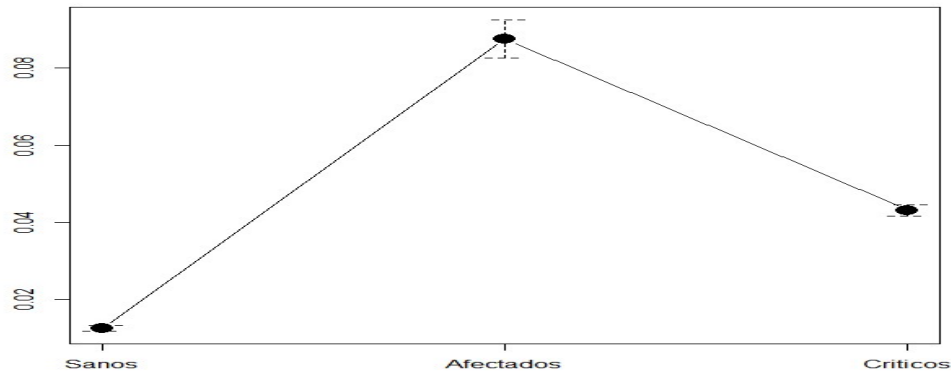


Figura 16. Intento de relaciones sexuales a la fuerza de camino a casa o colegio

Se evidencia que los grupos "Afectados" y "Críticos", presentan la mayor afectación respecto al promedio global de 3%. Para este grupo se evidencia que el promedio de la proporción de estudiantes que afirmaron presencia de este fenómeno es pequeño, se puede evidencia que es la segunda variable que mejor caracteriza este grupo ya que se encuentra por encima de la media global. Se verifico el resultado del Test LSD de Fisher ayuda a concluir que no hay suficientes pruebas estadísticas para decir que al menos un grupo de colegios de cada componente se parece a otro.

- Venta de Drogas en el colegio?



Figura 17. Venta de Drogas Colegio

Groups	Treatments and means
a	Afectados 0.1201
b	críticos 0.0611
c	Sanos 0.0207

Se evidencia que los grupos "Afectados" y "Críticos", presentan la mayor afectación respecto al promedio global de 5%. Donde el promedio de la proporción de estudiantes que afirmaron presencia de este fenómeno para este grupo es de 12%, se puede evidencia que es la tercera variable que mejor caracteriza este grupo ya que se encuentra por encima de la media global. Se verifico el resultado del Test LSD de Fisher ayuda a concluir que no hay suficientes pruebas estadísticas para decir que al menos un grupo de colegios de cada componente se parece a otro.

c) Tercer Grupo

Este grupo se caracteriza por proporción de estudiantes que manifestaron presencia de riñas de camino al colegio, porte de armas en el colegio, amenazas con armas, pertenece a pandillas, consumo de alcohol en el colegio, presencia de compañeros consumiendo drogas en el colegio, peleas ocasionadas por pandillas, venta de drogas cerca al colegio, y se contraponen con las dimensiones donde los estudiantes manifiestan desconocer si hay venta de drogas cerca de colegio y venta de drogas en el colegio. Dadas las condiciones se evidencia que este grupo asociado con pandillas y que a su vez esto puede generar afectación a la integridad de los estudiantes, ya que se relaciona con las riñas, porte de armas, y pandillas; debido a estas condiciones se llamará a este grupo de colegios como Colegios con alta afectación - "Críticos".

Variable	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
Riñas de camino al colegio	17,0	0,20	0,15	0,05	0,07	0,00
Ha llevado armas al colegio	16,9	0,24	0,17	0,08	0,09	0,00
Amenazas con armas	16,7	0,10	0,07	0,04	0,05	0,00
Pertenece a pandillas	15,0	0,08	0,06	0,03	0,03	0,00
Consumo de alcohol en el colegio	14,3	0,03	0,02	0,02	0,02	0,00
Presencia de compañeros consumiendo drogas	13,1	0,28	0,21	0,09	0,12	0,00
Presencia de peleas por pandillas	13,0	0,39	0,30	0,10	0,16	0,00
Venta de drogas cerca col	11,3	0,32	0,27	0,09	0,11	0,00
Total	9,7	1148	933	485	531	0
Ofensas amenaza	7,9	0,13	0,12	0,04	0,04	0,00
intento de relaciones a la fuerza de camino a colegio	7,0	0,04	0,03	0,03	0,03	0,00
Venta de drogas cerca colegio	6,6	0,06	0,05	0,04	0,04	0,00
Hombre	2,6	0,49	0,47	0,09	0,15	0,01
Venta drogas colegio	-9,8	0,20	0,26	0,10	0,16	0,00
No se si hay venta de drogas cerca colegio	-11,0	0,11	0,21	0,05	0,20	0,00

- En el mes pasado dentro de tu colegio, tuviste una pelea grupal física con otra persona?

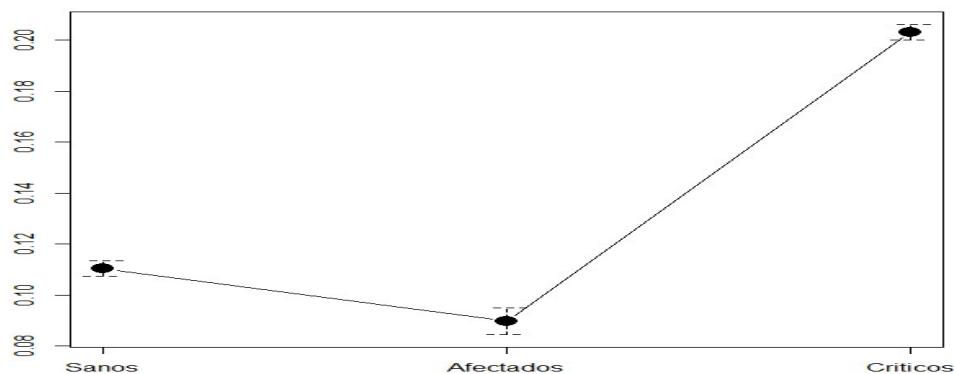


Figura 18. Peleas o riñas dentro de tu colegio

La variable es muy parecida en los grupos "Sanos" y "Afectados", aparentemente no hay dispersión de los datos en cada uno de los grupos analizados, esta variable es la que mejor caracteriza el grupo de los "Críticos", donde existe los mayores promedios de las proporciones de colegios donde los estudiantes afirmaron sobre la existencia peleas físicas con alguna persona en el colegio, así mismo las localidades como Chapinero, Santa fe, Rafael Uribe Uribe Antonio Nariño, y Usme, superan un promedio de 23%.

Groups	Treatments and means
a	crítico 0.2024333
b	Sanos 0.1103475
c	Afectados 0.08981481

El resultado del Test LSD de Fisher ayuda a concluir que no hay suficientes pruebas estadísticas para decir que al menos un grupo de colegios de cada componente se parece a otro.

- En este año, trajiste ARMAS al colegio

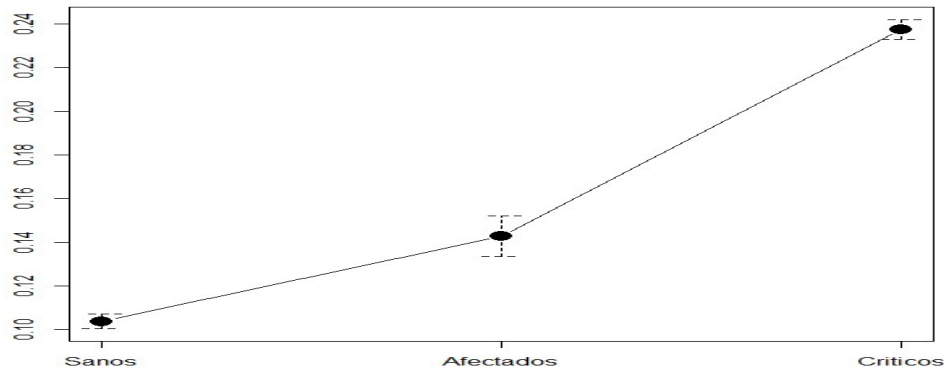


Figura 19. En este año, trajiste ARMAS al colegio

Groups	Treatments and means
a	crítico 0.2363333
b	Afectados 0.1427778
c	Sanos 0.1034749

La segunda variable que mejor caracteriza el componente "Críticos", donde el promedio de la proporción de estudiantes por colegios que manifestaron llevar arma es de 24%, valor que supera al promedio global. Los resultados nos muestran que las localidades donde con los promedios más elevados son Antonio Nariño, Barrios Unidos, Bosa, Chapinero y Ciudad Bolívar, donde la proporción de estudiantes que afirman que se presentó este fenómeno es superior al 25% valor que se encuentra por encima del promedio global y del grupo de colegios en estado "Críticos".

con los resultados del Test LSD de Fisher podemos concluir que al menos ningún grupo de colegios de cada componente es parecido a otro grupo.

- En este año, dentro de tu colegio ¿cuántas veces alguien te amenazó con un arma

Groups	Treatments and means
a	crítico 0.09756667
b	Afectados 0.05481481
c	Sanos 0.03223938

Los promedios para cada clúster son bajos sin embargo esta variable es la tercera en importancia al momento de caracterizar el componente, los colegios que se encuentran en este grupo cuentan con el promedio proporcional de estudiantes que manifiesta haber sido objeto de amenazas del 9%, explorando los datos se encontró que estos colegios se encuentran en las localidades de Los mártires, Antonio Nariño, Santa Fe, Rafael Uribe Uribe y Usme. Los resultados del Test LSD de Fisher podemos concluir que al menos ningún grupo de colegios de cada componente es parecido a otro grupo.

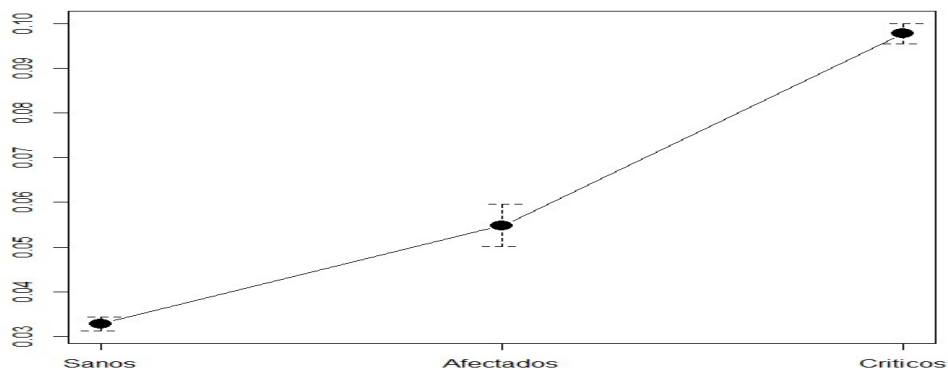


Figura 20. En este año, dentro de tu colegio ¿alguien te amenazó con un arma

- Pertenece a alguna pandilla

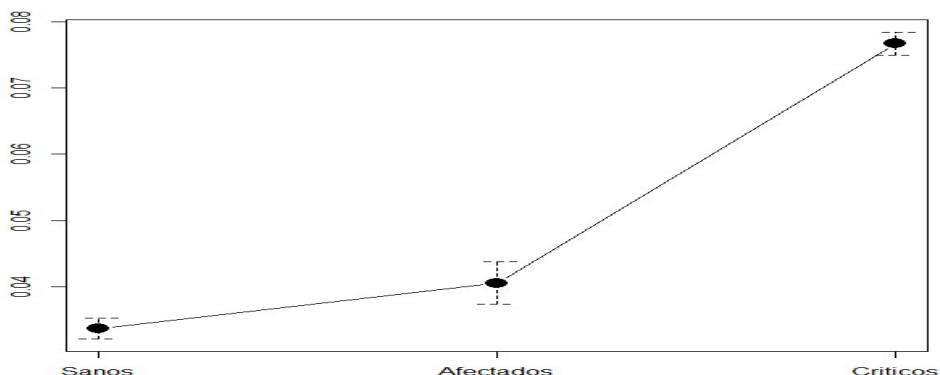


Figura 21. Pertenece a alguna pandilla

Groups	Treatments and means
a	críticos 0.0767
b	Afectados 0.0405
b	Sanos 0.0336

Para dicho el grupo "crítico", se encontró que el 7%, de los estudiantes afirmaron pertenecer a pandillas, las localidades donde mayor proporción de colegios son Chapinero, Antonio Nariño, Usme, Rafael Uribe Uribe, Ciudad Bolívar y San Cristóbal, estas localidades cuentan con un promedio superior al 8%. Adicionalmente se observa con los resultados del Test LSD de Fisher muestra que los colegios de los grupos Afectados y Sanos, presenta suficientes pruebas estadísticas para asegurar que son parecidos.

- Consumes bebidas alcohólicas dentro de tu colegio

Groups	Treatments and means
a	crítico 0.03256667
b	Afectados 0.01925926
c	Sanos 0.00988417

Para este componente se observa que los promedios del porcentaje de estudiantes que contestaron que si consume bebidas alcohólicas son bajos y adicionalmente que no hay suficientes pruebas estadísticas que permitan asegurar que los grupos colegios que pertenecen a cada grupo se parece al parezcan entre sí, se observó que las localidades con los promedios superiores al 5%, son Los mártires y Tunjuelito.

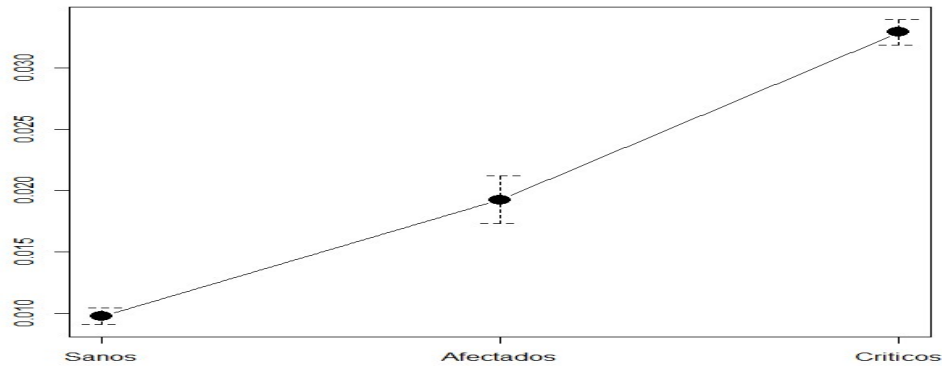


Figura 22. Consumes bebidas alcohólicas dentro de tu colegio

- Ha presenciado a algún compañero de tu curso consumiendo drogas dentro del colegio?

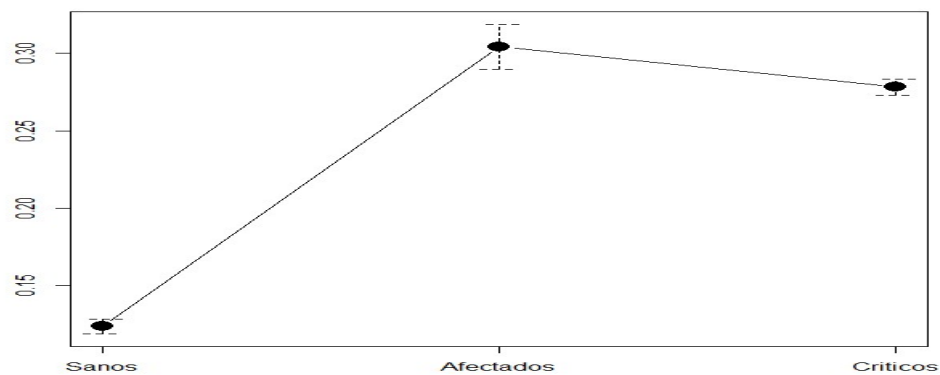


Figura 23. Ha presenciado a algún compañero de tu curso consumiendo drogas dentro del colegio?

Groups	Treatments and means
a	Afectados 0.3042593
a	crítico 0.2778
b	Sanos 0.122278

Para esta variable se evidencia que los grupos "Afectados" y "Críticos" tienen un promedio superior a 25 % del porcentaje de estudiantes que afirmaron ver algún compañero consumiendo drogas, se puede ver que los resultados del Test LSD de Fisher muestra que hay suficientes pruebas estadísticas para afirmar que el grupo de colegios que se encuentran en estos grupos son parecidos; aun así dicha variable tiene mayor influencia en el componente "Afectados". Se encontró que la localidades que supera el promedio del componente "Críticos" son Rafael Uribe Uribe, Santa Fé, La Candelaria, los cuales son mayores al 30 %

- En este año: Ha visto dentro de tu colegio peleas, ataques, u otros tipos de violencia realizadas por pandillas?

Groups	Treatments and means
a	Afectados 0.4998148
b	crítico 0.3884333
c	Sanos 0.160888

Se evidencia que dicha variable es influyente en los Grupos "Afectados" y "Críticos", aun así los datos no son dispersos al interior de cada componente y el Test LSD de

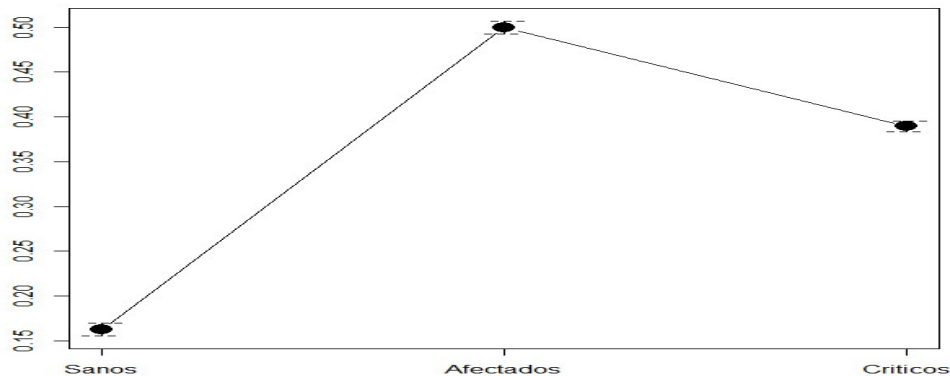


Figura 24. En este año: Ha visto dentro de tu colegio peleas, ataques, u otros tipos de violencia realizadas por pandillas?

Fisher nos ayuda a corroborar que no existen suficientes pruebas estadísticas para afirmar que el conjunto de colegios que hay en estos clúster sean parecidos. Se observa que las localidades de Usme, Ciudad Bolívar, Rafael Uribe Uribe, Bosa y La Candelaria, presentan los promedios más elevados, superior a 40%.

- Se venden drogas cerca tu colegio

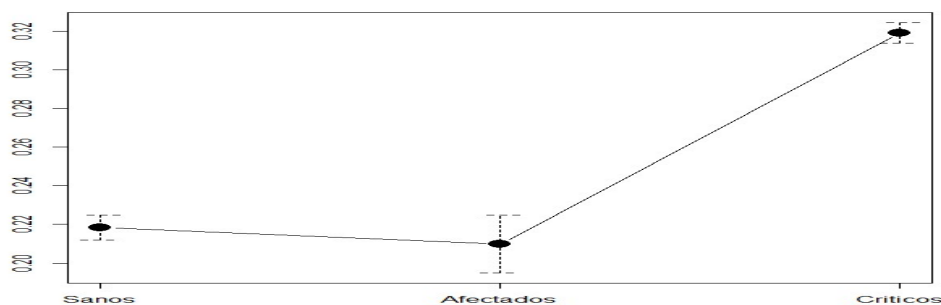


Figura 25. Se venden drogas cerca tu colegio

Groups	Treatments and means
a	crítico 0.3197667
b	Sanos 0.2164479
b	Afectados 0.2098148

La proporción de estudiantes por colegio que afirmaron se evidencia venta de drogas cerca del colegio presenta el promedio más elevado para el componente "crítico", se evidencia que existe suficiente prueba estadística para afirmar que el grupo de colegios que se encuentra en el clúster "Sanos" y "Afectados", son muy parecidos al interior. los colegios que se encuentran grupo pertenecen a las localidades de Santa Fe Ciudad Bolívar y Usme 44%.

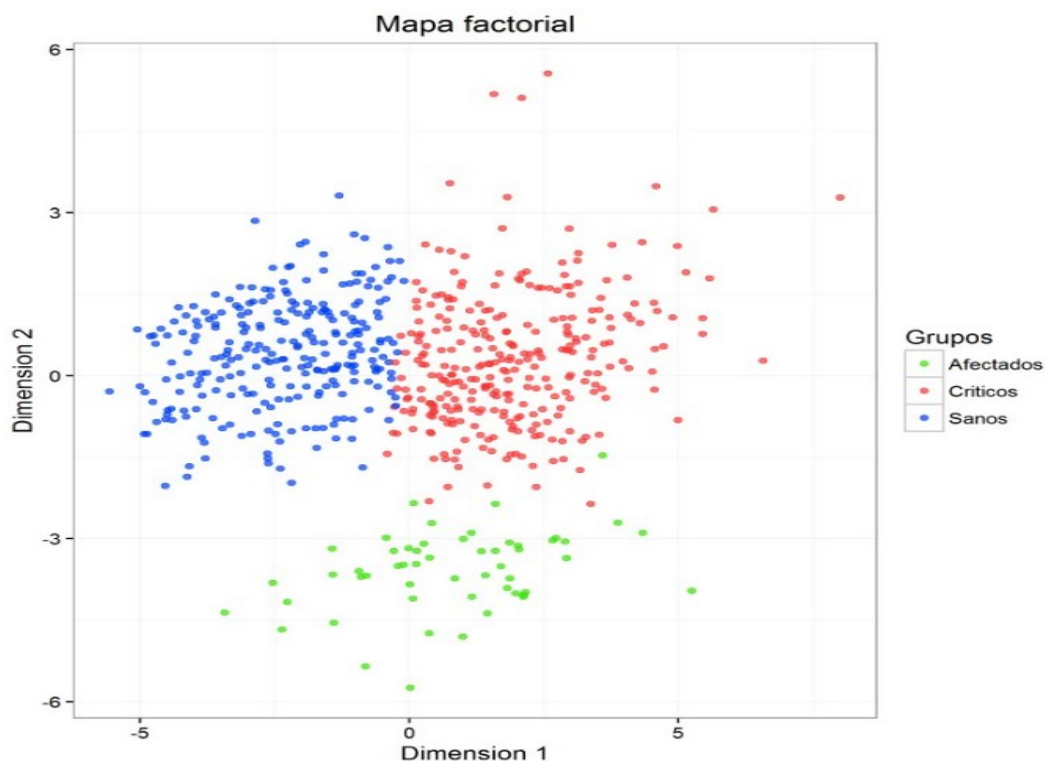


Figura 26. Agrupaciones de colegios

Las figuras 26 nos muestran un panorama de cómo se encuentran represados cada uno de los clúster "Críticos", "Sanos" y "Afectados", por los colegios de acuerdo a las zonas y sus localidades.

Retomando los resultados del plano factorial, se observa los colegios más cercanos a cada uno de los grupos, por ejemplo para el grupo uno se evidencia que el mejor representado es el colegio 577. Para el caso del clúster "Críticos", el colegio 505 y 154, ubicados en las localidades de Chapinero y Santa fe presenta la mayor distancia o mayor variabilidad respecto al centro de gravedad de clúster.

La figura 27 muestra cómo se encuentra como se encuentra los colegios ubicados en cada una de las cuatro zonas, tal como se aprecia existe buena concentración de los colegios diferenciando los colegios que se encuentran en cada uno de los clúster "Sanos" y "Críticos". Es de resaltar que la figura nos muestra que para el caso del clúster "Afectado" está ubicado en la zona Norte Occidente y que a su se pudo indagar que los colegio están ubicado especialmente en la localidad de Engativá.

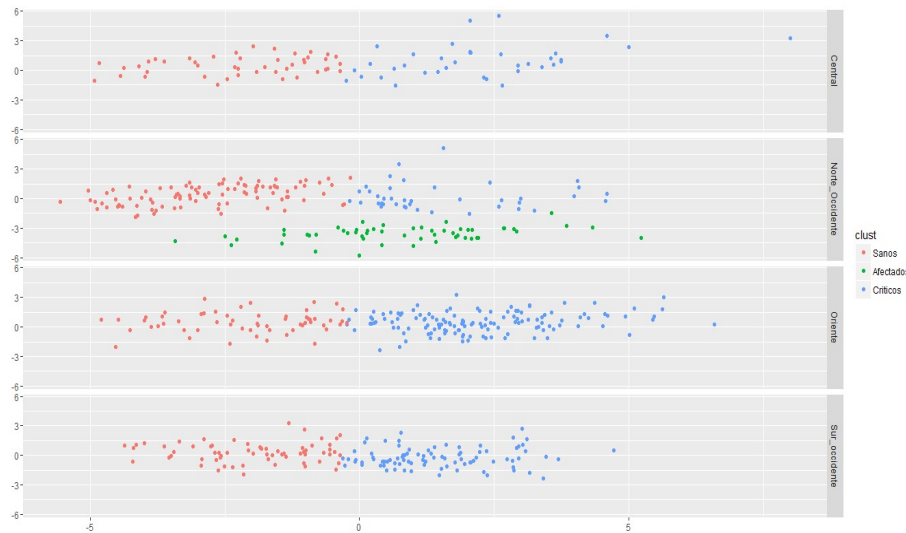


Figura 27. Agrupaciones de colegios por zonas de acuerdo al cluster

Las figuras 28 muestran la cantidad de colegios de cada localidad respecto a cada uno de los grupos, seleccionados.

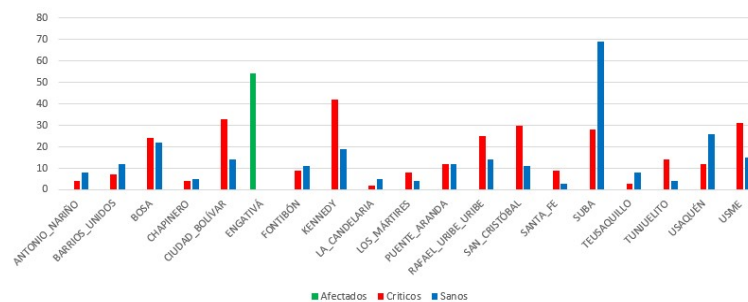


Figura 28. Cantidad de Colegios en las localidades por grupo

5.3. Modelamiento del Árbol de clasificación

Realizada la clasificación se encontraron los colegios que están en cada uno de los grupos definidos como "Sanos", "Afectados" y "Críticos", posteriormente se emplea un modelo que describa en qué condiciones o parámetros de victimización se define que un colegio pertenezca a uno de los grupos ya definidos, bajo este contexto se utilizará la metodología de Árboles de clasificación, explicada anteriormente. Para ello se utilizará una muestra del 70% y el 30% restante será utilizando como prueba para calcular el índice de validación cruzada.

La interpretación "Árbol de Clasificación" es muy fácil de comprender, las variables más altas son los nodos padres que dividen la población según la clase a tomar, es decir las variables predictivas, posteriormente siguen los nodos hijos, los cuales cumplen la misma condición del nodo padre, finalmente estos nodos conllevan a tomar una decisión, sobre una regla se debe tomar.

Los resultados arrojados en la figura 29 muestran como es la estructura del árbol, para las tres medidas o grupos definidos en la clasificación. El nodo padre fue origen de las variables que relacionan la proporción de estudiantes que contestaron afirmativamente sobre el siguiente cuestionamiento ¿En este año, ¿Has visto dentro de tu COLEGIO peleas, ataques u otros tipos de violencia realizados por pandillas?.

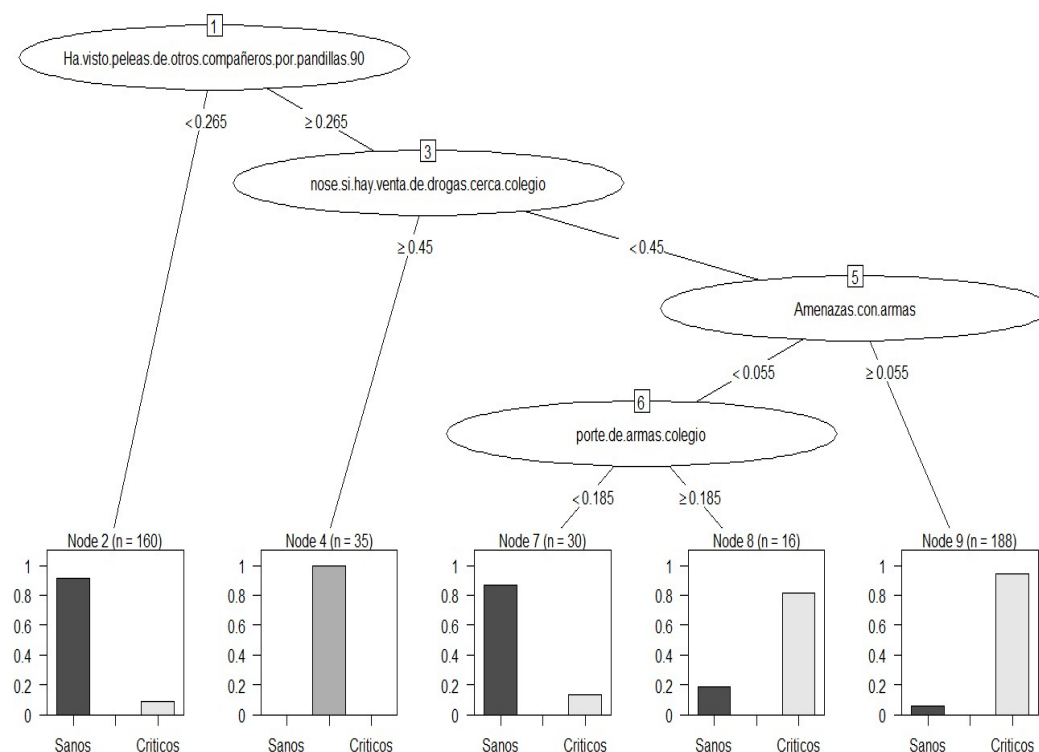


Figura 29. Árbol para Criterios de victimización

Para predecir si un colegio está en determinado grupo "Sano", "Afectado", "Crítico", se observa la tabla de validación cruzada la cual consiste en una tabla de igual número de filas y de columnas que cruza las clases predichas con las reales, tal como se muestra en la siguiente tabla, nos muestra una tasa de predicción del 89,13 %, con un intervalo de confianza de (83.7 %, 93.23 %).

Prediction	Sanos	crítico	Afectados
Sanos	71	0	15
Afectados	0	19	0
Críticos	5	0	14

Apoyandonos con la figura 29 y la tabla anterior podemos, podemos mencionar que el árbol solo presenta dos hojas para predecir que un colegio está en el grupo de los "Afectados", esta predicción se realiza con un porcentaje de 100 % y su interpretación es la siguiente:

- Si la proporción de estudiantes que afirma que ¿En este año, ¿Has visto dentro de tu COLEGIO peleas, ataques u otros tipos de violencia realizados por pandillas? es mayor o igual a 0,26 y la proporción de estudiantes por colegios que manifestaron sobre desconocimiento de la venta de drogas cerca del colegio es mayor o igual a 0,45, se puede decir que un colegio está en la categoría "Afectados".

Para predecir si un colegio se encuentra en el grupo "Sanos", se cuenta con una tasa de precisión 82.55 %, y para ello la metodología genera las siguientes dos condiciones:

- Si la proporción de estudiantes que afirma que ¿En este año, ¿Has visto dentro de tu colegio peleas, ataques u otros tipos de violencia realizados por pandillas? es menor 0.26, existe una probabilidad de pertenecer al grupo de los "Sanos" de 91 %.
- Si la proporción de estudiantes que afirma que ¿En este año, ¿Has visto dentro de tu colegio peleas, ataques u otros tipos de violencia realizados por pandillas? es mayor o igual 0,26, y si la

proporción de estudiantes que menciona desconocer sobre venta de drogas en el colegio es menor al 0.45 y la proporción de estudiantes que contestan afirmativamente la pregunta ¿En este año, dentro de tu colegio te amenazaron con un arma? es menor a 0.055 y la proporción de estudiantes que manifestaron que llevaron armas al colegio es mejor a 0.185. Un colegio que cumple estas condiciones tiene una probabilidad de pertenecer al grupo de los "Sanos" del 86 %.

Para predecir si un colegio se encuentra en el grupo "críticos", se cuenta con una tasa de precisión 73 %, y para ello la metodología genero las siguientes dos condiciones:

- Si la proporción de estudiantes que afirma que ¿En este año, ¿Has visto dentro de tu COLEGIO peleas, ataques u otros tipos de violencia realizados por pandillas? es mayor o igual 0.26, y la proporción de estudiantes que manifiesta desconocer si hay venta de drogas cerca es menor a 0,45. Y la proporción de estudiantes que manifiesta haber sido amenazados con arma es menor a 0.055 y la proporción de estudiantes que afirma haber llevado arma al colegio es mayor o igual a 0.18. La probabilidad de que un colegio pertenezca al grupo de los críticos es de 81,2 %.
- Si la proporción de estudiantes que afirma que ¿En este año, ¿Has visto dentro de tu COLEGIO peleas, ataques u otros tipos de violencia realizados por pandillas? es mayor o igual 0.26, y la proporción de estudiantes que manifiesta desconocer si hay venta de drogas cerca es menor a 0,45. Y la proporción de estudiantes que manifiesta haber sido amenazados con arma mayor o igual a 0.055. La probabilidad de que un colegio pertenezca al grupo de los críticos es de 94,14 %.

5.3.1. Validación del modelo

Se realizó simulación realizada con 10000 réplicas de Árbol de Clasificación, la figura 32 muestra el comportamiento del porcentaje de precisión del modelo de clasificación, donde podemos ver que dicho árbol tiene un promedio de predicción del 89.7 % y a su vez la dispersión no es alta y los valores mínimos no se encuentran por debajo del 80 %, esto nos da indicios que la tasa de precisión es buena.

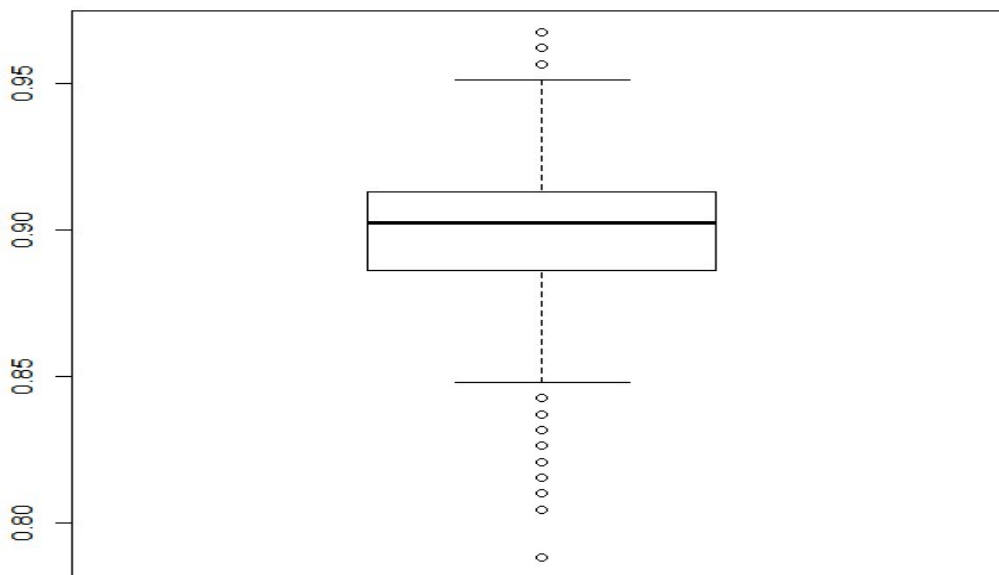


Figura 30. Boxplot de la matriz de Confusión

6. Conclusiones

- a) La estructuración de los grupos generados a partir de las técnicas aplicadas, permitieron identificar tres grupos de colegios Sanos, Afectados, Críticos, por cuanto dicha información puede ser útil momento de intervenir en los colegios que se encuentran en cada grupo, con el fin de realizar campañas de prevención.

- Los colegios que tienen características de sanos, tienen la particularidad que la proporción de estudiantes por colegio que manifiestan no tener conocimiento sobre venta de drogas cerca a los centros educativos es alta, en cuanto a las características donde la percepción de estudiantes por colegio que mencionan que existe presencia de peleas, porte y amenaza con arma, presencia de relaciones sexuales a la fuerza y presencia de drogas, tiene unas proporciones por debajo de la media de Bogotá.

Finalmente las reglas generadas a partir del árbol de clasificación, menciona que a medida que un colegio cuente con proporciones bajas en la pregunta que relaciones presencia de peleas causadas por pandillas, tiene un alta probabilidad de pertenecer a este grupo de colegios, en caso contrario que la proporción sea alta se tendría que verificar aspectos que relacionen baja proporción de porte de armas, amenazas con las mismas y desconocimiento de venta de drogas cerca al colegio.

- Las características de los colegios que se encuentran en el grupo de colegios Afectados, están ligados a las dimensiones donde los estudiantes manifiestan en una alta proporción por colegio que desconocen si hay venta de droga cerca al colegio, presencia de casos donde alguien intento de relaciones sexuales a la fuerza, venta de drogas en el colegio y alta proporción de presencia de peleas donde los participantes están asociados a pandillas.

De acuerdo a las reglas generadas por el árbol permitió deducir que un colegio que cuente con alta proporción de presencia de peleas donde los participantes son las pandillas y a su vez desconocer si hay venta de drogas cerca del colegio, tienen una alta probabilidad de pertenecer a este grupo de colegios.

- Los colegios que se encuentran en el grupo de críticos, se caracteriza por tener un porcentaje alto de estudiantes por colegio que manifiestan, presencia de peleas física o grupal en el colegio, asociado a porte de armas, amanezcas con armas, presencia de riñas causadas por pandillas, consumo de alcohol y drogas, así como la presencia de venta de drogas cerca y dentro del colegio, presencia peleas donde los participantes son las pandillas y finalmente intento de relaciones sexuales de forma forzosa.

Las reglas generadas por el Árbol de Clasificación (ver figura 29) nos ayuda a concluir que un colegio que cuente con altas proporciones de estudiantes por colegio que manifiesten presencia de amenazas con armas, peleas causadas por pandillas y a su vez bajas proporciones en los estudiantes que desconocen si hay venta de drogas cerca del colegio tiene una alta probabilidad de pertenecer al grupo de los críticos, si la proporción es baja, para el caso que el porcentaje de amenazas con armas sea bajo pero el porte de armas sea alto también genera alta probabilidad de pertenecer a pandillas.

- b) Las reglas generadas, a partir del árbol de decisión, pueden ser aplicadas en los colegios de Bogotá, al momento de hacer una encuesta que contengan las variables seleccionadas, de manera que determine en que condición está el colegio objeto de análisis.

6.1. apendices

Resumen de los resultados del analisis de componentes principales

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	6,1	37,8	37,8
comp 2	2,4	15,1	53,0
comp 3	1,6	9,8	62,7
comp 4	1,0	6,5	69,3
comp 5	0,9	5,5	74,8
comp 6	0,7	4,6	79,4
comp 7	0,6	3,9	83,3
comp 8	0,5	3,4	86,6
comp 9	0,5	3,0	89,7
comp 10	0,3	2,0	91,7
comp 11	0,3	1,9	93,6
comp 12	0,3	1,6	95,2
comp 13	0,2	1,5	96,7
comp 14	0,2	1,3	98,0
comp 15	0,2	1,1	99,1
comp 16	0,1	0,9	100

Figura 31. Inercia del ACP

7. Bibliografia

Referencias

- [1] La ONU *Organización Naciones Unidas*, (1986)
- [2] Fernando, Valencia, *Conflicto y violencia escolar en Colombia*, Lectura breve de algunos materiales escritos, Revista científica Guillermo de Ockham. Vol. 7 (1), (2004)
- [3] Dan, Áke Olweus, *profesor de psicología de la Universidad de Bergen, Noruega*, 18 de abril de 1931 en Kalmar (Suecia), especializado en la investigación de violencia escolar.
- [4] Schafer, J, L,(1999) *Multiple Imputation, a prime Statistical Methods in Medical Research 8*, Analysis of incomplete multivariate data, London, Chapman y Hall.
- [5] Breiman, L.,J.H, Firedman, R.A Olshen, and C.J. Stone (1984) *Classification And Regression Tress*. New York. Chapman and Hall.
- [6] Lebart, L., Piron, M. and Morineau, A.,*Statistique exploratoire multidimensionnelle. Visualisation et inférence en fouilles de données*, 4 ed, pg 256, (2006), Dunod, Paris.
- [7] Carles, M.,Cuadras, *Nuevos Metodos de Análisis Multivariante*. CMC Editions Manacor 30, Barcelona Spain, Febrero (2007)
- [8] Timofeev, Roman, *Classification and Regression Tress (CART) Theory and Applications*. Master Thesis, Center of Applied Statistics and Economics, Humboldt University, Berlin, Germany, December 20, 2004.
- [9] Sandra Milena López, Iván Gómez Sánchez, Ingrid Arévalo Rodríguez, *Violencia contra la Mujer, Análisis en el instituto materno infantil de Bogotá, Colombia 2015, Estudio de corte trasversal*, Federeación Colombiana de Asociaciones de obstetricia y Ginecología.

- [10] Peña, Daniel, *Análisis de datos multivariantes*, (2002)
- [11] Fernando, Medina y Marco, Galvan. *Imputación de datos Teoría y practica*, estudios estadísticos y prospectivos, (2007), pages 21-37.
- [12] Santiago Fuentes, F. *Análisis de Conglomerados*, Ciencias Economicas y Empresariales, Universidad Autonoma de Madrid, (2011).
- [13] *Encuesta Clima Escolar y Victimización Estudiantil*, Alcaldia Mayor de Bogota D.C. - Secretaria Educación de Bogotá.
- [14] *Educación para la Ciudadanía y Convivencia*, Boletin de prensa 2014, Alcaldia Mayor de Bogota D.C. - Secretaria Educación de Bogotá.
- [15] Santiago Fuentes, F. *Análisis de Conglomerados*, Ciencias Economicas y Empresariales, Universidad Autonoma de Madrid, (2011).
- [16] , Pandillas Juveniles, *Pandillas juveniles - factor de riesgo en la seguridad ciudadana*, Atlantico 2013.
- [17] , Rubin, D. B, *Informe ejecutivo sobre situación de convivencia y seguridad escolar en la upz 47-kennedy central localidad 8 - Kennedy*.
- [18] , *Definición y Categorización de Pandillas*,Secretaria General de la Organización de los Estados Americanos, Junio 2007.
- [19] , *Naciones Unidas, Oficina Contra la Droga y el Delito*, Quinto Informe Conjunto 2010.
- [20] Bogotá Humana, *Educación para la ciudadanía y la Convivencia*. Encuesta de Clima Escolar y victimización, Secretaria de Educación del Distrito.
- [21] DANE, *Alarma en colegios de Bogotá*, Comunicado de Prensa, Mayo 2012.