UNIVERSIDAD DE LA REPÚBLICA

Facultad de Ciencias Económicas y de Administración

Licenciatura en Estadística

Informe final de pasantía

UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

Caracterización y clasificación de jóvenes en función de características sociodemográficas y sus distintas conductas

Leonardo Brito

Tutora: Laura Nalbarte

Montevideo, Uruguay -2018-

UNIVERSIDAD DE LA REPÚBLICA

FACULTAD DE CIENCIAS ECONÓMICAS Y ADMINISTRACIÓN

El tribunal integrado por los abajo firmantes aprueba el trabajo de Pasantía:

Caracterización y clasificación de jóvenes en función de características sociodemográficas y sus distintas conductas

Leonardo Brito.

Tutor académico: Laura Nalbarte.

Cátedra:

Puntaje:

Tribunal:

Profesor: .... Nalbarte, Laura

Profesor: .... Bourel, Mathías

Profesor: .... Alvarez,Ramón

Fecha:

Agradecimientos

Quiero agradecer a la Universidad de la República, por dar la oportunidad de estudiar y ser profesional, a todos los profesores durante toda la carrera han aportado con la formación y han compartido sus conocimientos. Principalmente agradezco a la Profesora Laura Nalbarte, quien me acompañó en calidad de tutora, por todo el apoyo brindado, sus visiones críticas en muchos aspectos y las recomendaciones brindadas, que ayudan a formarse como un mejor investigador. Y en general a todas las personas que apoyaron y ayudaron brindando su tiempo para permitir la culminación de una etapa importante en mi vida.

A todos, muchas gracias.

Resumen

El estudio se basa en la Encuesta Nacional de Adolescencia y Juventud (ENAJ) realizada en el 2013. En la misma se encuesta a 3.818 jóvenes de 12 a 29 años residentes en las localidades de 5.000 y más habitantes del País.

El objetivo del trabajo es la construcción de tipologías de jóvenes en función de sus distintas conductas y determinar en que medida se asocian las características sociodemográficas y ciertos pensamientos de los jóvenes a las mismas.

Se emplearon técnicas de análisis factorial, concretamente Análisis de correspondencia múltiple (ACM) para estudiar la asociación del conjunto de modalidades y variables.

Para la obtención de grupos se emplean 2 estrategias, por un lado se considera únicamente la información de las variables de conducta y por otro se trata la totalidad de las variables originales donde se utilizan métodos jerárquicos y no jerárquicos que posteriormente son comparados, con el fin de seleccionar la estructura más adecuada.

Para la modelización del grupo de pertenencia de cada joven se utilizan técnicas paramétricas y no paramétricas como la Regresión logística y los modelos de decisión del tipo CART (Classification and Regression Tree) respectivamente. Para esto se adopta como estrategia particionar la muestra en muestra de entrenamiento para aplicar las técnicas y muestra de prueba para evaluar su poder predictivo.

Como resultados se encontraron distintos grupos entre los cuales se identificó que los actos indebidos más frecuentes en los jóvenes son, manejar sin libreta y consumir sustancias ilegales.

Mediante ACM se identifica un comportamiento diferenciado en cuanto a los pensamientos de los jóvenes encontrando una fuerte asociación de quienes están insatisfechos con la seguridad en general, sufrieron discriminación y toman algún comportamiento indebido, respecto quienes están satisfechos en términos de seguridad, no sienten discriminación y no toman comportamientos inadecuados.

Los resultados obtenidos determinan que si bien se encuentran distintas tipologías de jóvenes en función de sus conductas que permiten observar si el joven es más o menos violento, no se encuentran características sociodemográficas que contribuyan a explicar las mismas.

Palabras claves: ACM, clustering, modelos de clasificación, conductas indebidas.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

Las personas realizan infinidad de actuaciones a lo largo de su vida. Si bien muchos comportamientos no deseados, no suelen considerarse como un problema a pesar de los daños que causan, existe una preocupación por parte de la sociedad cuando se irrumpe o rompe una norma que viola principios legales y/o morales. Para ello existen varias organizaciones/personas que cuentan con medios que luchan contra estos actos cuando los mismos representan un problema para la sociedad. Esto motiva la necesidad de estudiar los distintos fenómenos, analizar entre otras cosas si existen factores que influyen o determinan ciertos comportamientos indebidos así como tratar de identificar las posibles relaciones entre características de los individuos y sus comportamientos (indebidos o no). A su vez, se busca construir tipologías de individuos que aporten información al estudio de la problemática.

Los datos analizados en el presente trabajo provienen de la Encuesta Nacional de Adolescentes y Jóvenes (ENAJ) realizada por el Instituto Nacional de Estadística (INE), Ministerio de Desarrollo Social (MIDES), el Instituto Nacional de Evaluación Educativa (INEED) y la Facultad de Ciencias Sociales de la Universidad de la República a través del departamento de Sociología en el año 2013. Esta encuesta fue aplicada a 3.818 jóvenes de 12 a 29 años, residentes en localidades del país de 5.000 y más habitantes. Según el Censo de Población, Hogares y Vivienda del año 2011, la cantidad de personas de 12 a 29 años representan el 27 % de la población. (ENAJ 2013).

El trabajo está estructurado en 5 capítulos más un anexo. En el presente capítulo se plantean objetivos y se presentan antecedentes a nivel nacional e internacional. En el capítulo siguiente se presentan los aspectos principales de las metodologías utilizadas que refieren a técnicas de reducción de datos como el análisis factorial y técnicas que permiten la modelización de variables de respuesta discreta.

Luego se dedica un capítulo al análisis descriptivo de los datos, donde se expone el tratamiento de las variables, su vinculación y el diseño muestral. A continuación, se presentan los resultados obtenidos según las distintas metodologías aplicadas.

En el último capítulo se realiza una síntesis y resumen de las conclusiones del trabajo donde se comparan los resultados alcanzados con los objetivos planteados y se describen alcances y limitaciones así como consideraciones para trabajos futuros. En el capítulo de anexos se adjuntan por un lado metodologías utilizadas y por otro se determinan resultados alcanzados mediante los análisis que se mencionan en el correr de los capítulos tres y cuatro.

1.1. Objetivos

A continuación se presentan los objetivos del presente estudio, detallando el objetivo general y los objetivos específicos del mismo.

1.1.1. Objetivos generales

El objetivo general del presente trabajo es la construcción de tipologías de jóvenes en función de sus distintas conductas y determinar en que medida contribuyen a las mismas las características sociodemográficas así como las opiniones o pensamientos que tienen dichos jóvenes.

1.1.2. Objetivos específicos

- Analizar las posibles asociaciones entre los factores sociodemográficos y opiniones de los jóvenes distinguiendo entre quienes manifiestan haber tomado conductas "indebidas" frente a quienes manifiestan no haber tomado dichas conductas.

- Determinar si existen grupos bien definidos de mayor vulnerabilidad en cuanto a sus conductas y características sociodemográficas.
- Estimar la probabilidad de que un joven se encuentre en un grupo más propenso a tener conductas ”indebidas”.
- Buscar posibles ”trayectorias” o características que pueden determinar que un joven sea clasificado en los distintos grupos de conductas ”indebidas”.

Teniendo en cuenta la información que surge de los antecedentes estudiados, se entiende que las características sociodemográficas de los jóvenes pueden estar asociadas a las conductas indebidas que realizan los mismos. En ese sentido se plantean las siguientes hipótesis de trabajo:

- Existe diferenciación de las conductas según sexo. Los hombres son más propensos a cometer infracciones.
- El hogar en que viven los jóvenes es un factor importante vinculado a sus conductas. Los jóvenes que no viven con sus padres son más propensos a desarrollar conductas indebidas.
- Los jóvenes que pertenecen al quintil más alto de ingreso, no se asocian a las conductas indebidas más violentas, en el entendido que mayores niveles de bienestar aumentarían al incrementar la riqueza del hogar.
- La región donde reside el joven se asocia con ciertas conductas indebidas. Montevideo es la región donde se registran mayor cantidad de conductas indebidas más violentas.

A los efectos de cumplir con los objetivos antes mencionados se llevan adelante distintas estrategias metodológicas. Se aplican técnicas de análisis factorial para reducir dimensiones, técnicas de clasificación no supervisada para determinar tipologías y técnicas de clasificación supervisada para construir reglas de clasificación.

1.2. Antecedentes

Desde el punto de vista del derecho se puede encontrar mucha información en investigaciones a nivel judiciales y/o penal donde el centro de estudio son los comportamientos de los jóvenes. Estas se llevan a cabo con el propósito de reunir información que permita una mayor adecuación a los estándares nacionales e internacionales en materia de política penal juvenil. Existe bibliografía que se aproxima al estudio, es el caso del trabajo *'Adolescentes en conflicto con la Ley'* (Argentina), impulsado por Unicef del (MIDES, (2015)).

No existe un factor que, por sí solo, explique por qué una persona se comporta de cierta manera y otra no lo hace. El estudio del comportamiento humano es un problema complejo, enraizado en la interacción de muchos factores biológicos, sociales, culturales, económicos y políticos.

A nivel nacional si bien existen escasos trabajos que provienen de psicología, sociología, derecho, etc. referentes a los distintos comportamientos de los jóvenes, estos no abordan todas las técnicas empleadas en este trabajo, sino a una definición y estudio más general en esta rama. Básicamente se realizan estudios descriptivos donde solamente se menciona el peso de los individuos en forma general.

A continuación se presenta una breve descripción de algunos trabajos que comparten el tema de estudio ó algunas técnicas/procedimientos vinculadas con el presente estudio.

1.2.1. Caracterización de jóvenes uruguayos que no asisten al sistema educativo

Este estudio fue el trabajo final de pasantía "*Caracterización de jóvenes uruguayos que no asistieron al sistema educativo*" de Caballero, N y Jadra, G (2013). El mismo busca determinar las características que impactan en la asistencia a un centro educativo de los jóvenes uruguayos de 14 a 17 años de edad, tomando como fuente de datos los provenientes de la ECH2011 que elabora el INE.

En el trabajo se aplicaron distintas metodologías: modelos lineales generalizados, análisis factorial y árboles de decisión.

Los resultados obtenidos para los distintos modelos reflejan qué características del joven, como la edad, el ser estudiante activo, tener hijos a cargo y ser jefe del hogar tienen un impacto negativo en la asistencia, características que lo cual disminuyen la probabilidad de asistencia del joven. En cambio características como los años de educación y el clima educativo tienen un impacto positivo. Es decir, que cuanto mayor sean los años de educación acumulados por el joven y el clima educativo del hogar más probable es que asista. También surge un comportamiento diferenciado en cuanto al género, encontrándose una mayor probabilidad de asistencia en las mujeres respecto a los varones.

En lo que refiere a las características del hogar se constata una mayor probabilidad de asistencia en hogares donde la calidad de la vivienda es buena frente a aquellos donde es deficitaria. La condición de hacinamiento tiene un impacto negativo, disminuyendo la probabilidad de asistir. Respecto al lugar de residencia, el hecho de residir en la capital aumenta la probabilidad de asistir frente a los que residen en el interior del país. En lo que respecta a los árboles de clasificación, se aprecia que dicha técnica aporta a la descripción de la variable asistencia y a determinar cuáles son las variables que discriminan entre los que asisten y los que no a un centro educativo.

1.2.2. Informe mundial sobre la violencia y la salud

Otro trabajo que se aproxima al tema de estudio es el (Organización Panamericana de la Salud, (2002)) "*Informe mundial sobre la violencia y la salud*" [1] de la Organización Panamericana de la Salud, (2002), este recurre a un modelo ecológico para intentar comprender la naturaleza de la violencia. Dicho modelo, todavía está en fase de desarrollo y perfeccionamiento como instrumento conceptual. Su principal utilidad estriba en que ayuda a distinguir entre los inumerables factores que influyen en la violencia y los comportamientos indebidos. El modelo permite analizar los factores que influyen en el comportamiento (o que aumentan el riesgo de cometer o padecer actos violentos) clasificándolos en cuatro niveles.

1. En el primer nivel se identifican los factores biológicos y de la historia personal que influyen en el comportamiento de los individuos y aumentan sus probabilidades de convertirse en víctimas o perpetradores de actos violentos. Entre los factores que pueden medirse o rastrearse se encuentran las características demográficas (edad, educación, ingresos), los trastornos psíquicos o de personalidad, las toxicomanías y los antecedentes de comportamientos agresivos o de haber sufrido maltrato.

2. En el segundo nivel se abordan las relaciones más cercanas, como las mantenidas con la familia, los amigos, las parejas y los compañeros, y se investiga cómo aumentan éstas el riesgo de sufrir o perpetrar actos violentos. En la violencia juvenil, por ejemplo, tener amigos que cometan o alienten actos violentos

[1] Informe disponible en: http://www.who.int/violence.injury.prevention/violence/world.report/es/summary.es.pd

puede elevar el riesgo de que un joven los sufra o los perpetre.

3. En el tercer nivel se exploran los contextos comunitarios en los que se desarrollan las relaciones sociales, como las escuelas, los lugares de trabajo y el vecindario, y se intenta identificar las características de estos ámbitos que aumentan el riesgo de actos violentos. A este nivel, dicho riesgo puede estar influido por factores como la movilidad de residencia (por ejemplo, el hecho de que las personas de un vecindario tiendan a permanecer en él durante largo tiempo o se trasladen con frecuencia), la densidad de población, unos niveles altos de desempleo o la existencia de tráfico de drogas en la zona.

4. El cuarto nivel se interesa por los factores de carácter general relativos a la estructura de la sociedad que contribuyen a crear un clima en el que se alienta o se inhibe la violencia, como la posibilidad de conseguir armas y las normas sociales y culturales. Entre éstas se incluyen las que conceden prioridad a los derechos de los padres sobre el bienestar de los hijos, reafirman la dominación masculina sobre las mujeres y los niños, respaldan el uso excesivo de la fuerza policial contra los ciudadanos o apoyan los conflictos políticos. En este nivel, otros factores más generales son las políticas sanitarias, económicas, educativas y sociales que contribuyen a mantener las desigualdades económicas o sociales entre los grupos de la sociedad. Los niveles de bienestar personal, satisfacción individual y felicidad son mayores a medida que aumenta la riqueza del hogar. La ansiedad se reduce cuando crecen los ingresos familiares.

La figura que se presenta a continuación es extraída del *informe mundial sobre la violencia y salud* (OPS,2002) y hace referencia al modelo que usan en el informe para medir la violencia.

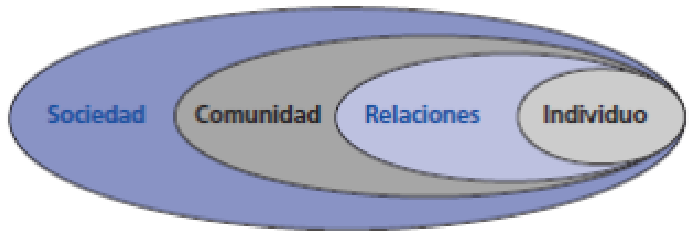

Figura 1.1: Modelo ecológico para comprender el comportamiento

El solapamiento de los anillos ilustra cómo los factores de cada nivel refuerzan o modifican los de otro. Así, por ejemplo, un individuo de personalidad agresiva tiene más probabilidades de actuar violentamente en el seno de una familia o una comunidad que acostumbra a resolver los conflictos mediante la violencia que si se encuentra en un entorno más pacífico.

Capítulo 2

Metodología

Este capítulo está estructurado en 3 secciones donde se detallan los distintos procedimientos que se llevan a cabo en la presente investigación para cumplir los objetivos planteados.

2.1. Análisis Factorial

El Análisis Factorial es una técnica de análisis multivariado cuyo propósito, entre otros, consiste en buscar el mínimo número de dimensiones o factores (variables no observadas) capaces de simplificar y explicar el máximo de información contenida en los datos, eliminando la información redundante. Se buscan nuevas variables (factores) que son combinación lineal de las variables originales.

El nombre factorial se debe a que la descomposición que se pretende de la matriz de datos se realiza en factores o también llamados ejes de inercia sobre los cuales se realiza la proyección de la nube de datos.

Los objetivos del análisis factorial son:

1. Eliminar información redundante. Se procuran nuevas variables, combinación lineal de las variables originales, incorrelacionadas entre si.
2. Simplificar, pasando de espacios de muchas dimensiones a otros de menores dimensiones. Se busca que la reducción de dimensiones impliquen la menor

pérdida posible de información (se procura que la nube de puntos se deforme lo menos posible).

3. Estudiar variables e individuos, asociaciones de las variables, caracterización de los individuos.

Existen distintos métodos de análisis factorial:

1. Análisis de Componentes Principales (ACP), aplicado a variables cuantitativas.
2. Análisis de Correspondencia Simple (ACS) aplicado a tablas de contingencia. Se analiza asociaciones de 2 variables cualitativas.
3. Análisis de Correspondencia Múltiple (ACM) aplicado a variables cualitativas. Se utilizan tablas disyuntivas.
4. Análisis discriminante Factorial, se trabaja con variables cuantitativas y subpoblaciones, buscando reglas de clasificación que discriminen lo mejor posible las subpoblaciones.

A continuación se presenta la técnica de ACM, para ello se sigue fundamentalmente el texto *Introducción al Análisis Multivariado*, (Blanco, J.(2006)).

2.1.1. Análisis de correspondencia múltiple (ACM)

El ACM es una técnica de análisis multivariado enmarcada dentro de las técnicas de análisis factorial aplicado cuando se busca resumir información y la misma es de tipo cualitativa. Cada variable tiene k modalidades, las que deben ser exhaustivas y excluyentes, es decir todos los individuos deben tener una modalidad y en las J variables.

Los datos pueden provenir de una tabla disyuntiva completa (TDC), Z_{IK} en donde las filas corresponden a los individuos y las columnas a las modalidades de cada variable (para cada individuo se registra 1 si posee la modalidad k y 0 si no la posee) o bien en una tabla de Burt = Z*Z (en la cual las filas y columnas representan al mismo tiempo todas las modalidades). En caso de que la información provenga de la matriz de datos de tipo individuos por variable (X_{IJ}) la misma se transforma en una disyuntiva completa.

Los objetivos de ACM son:

- Reducir la dimensión de la matriz de datos con el fin de eliminar información redundante y manifestar las relaciones existentes entre las variables a través de sus modalidades. Se obtienen nuevas variables o factores que resumen la información esencial que surge de la nube de puntos.
- Caracterización de individuos, lo mismo se realiza usando nociones de similitudes. Dos individuos serán más próximos cuanto mayor sea el número de modalidades en común.
- Asociación de Variables. Dos variables serán más cercanas cuanto más individuos la compartan.
- Estudio de las modalidades. Analizar las asociaciones entre modalidades de una misma variable y entre modalidades de distintas variables.

En la práctica, es de interés, obtener un primer eje factorial en el cual su dirección haga máxima la inercia respecto al baricentro. Una vez encontrado el primer eje, al segundo se le impone la condición de ser ortogonal al primero y así sucesivamente se van encontrando el resto de los ejes.

En el análisis factorial la inercia de la nube proyectada sobre un eje u_s es igual a: $\sum_i p_i[F_s(i)]^2$

Siendo p_i el peso de la fila i, $F_s(i)_i$ la proyección en un nuevo eje de la coordenada i y F_s la nueva variable combinación lineal de las x. Matricialmente se puede escribir en función de la matriz diagonal de los pesos D y del vector de las coordenadas de las proyecciones sobre u_sF_s como $F_s'DF_s$. Como $F_s = XMu_s$, resulta:

$Inercia = u_sMX'DXMu_s$

Siendo X la matriz de datos, M la métrica y u_s la dirección, eje de inercia o eje factorial.

En ACM la inercia total de la nube es $\frac{K}{J} - 1$ donde K es la cantidad de modalidades y J la cantidad de variables.

Se trabaja en dos nubes de puntos, en R^I y R^J transformando en perfiles (filas y columnas) las tablas de datos.

Los elementos con los que se trabajan en un ACM son:

- X_{IJ} matriz de datos de I x J
- Z_{IK} tabla disyuntiva completa de I x K
- D_f perfil filas
- D_c perfil columnas
- $F = \frac{Z}{IJ}$, $f_{ij} = \frac{Z_{ij}}{IJ}$

2.2. Análisis de Grupos

La técnica de análisis de grupos (análisis de cluster o conglomerados) consiste en clasificar a las observaciones/individuos en grupos (clusters), tales que los individuos dentro de cada cluster presenten cierto grado de homogeneidad en base a los valores adoptados sobre el conjunto de características que se miden. Estas son técnicas de clasificación no supervisada, en la medida que se desconoce a que grupo pertenece cada observación.

Se trata de clasificar las observaciones en grupos de modo que los individuos de un mismo grupo sean lo más similares posible y que los grupos entre sí sean muy distintos.

Se realizan técnicas de clasificación de tipo jerárquico y no jerárquico. Se proponen distintos procedimientos y posteriormente la comparación entre ellos.

2.2.1. Disimilaridades

Para la formulación matemática de los algoritmos que se plantean en las siguientes secciones es necesario identificar ciertos conceptos y definiciones relacionados con las distancias, similitudes y sus propiedades.

Se debe entender a una distancia como una medida de proximidad entre los objetos o instancias, que puede ser usado tanto para variables cuantitativas, como cualitativas.

Al trabajar con variables cuantitativas se suele utilizar las distancia euclídea. Sin embargo, es importante analizar si su uso es apropiado teniendo en cuenta las características de los datos. Por ejemplo, si las variables que se utilizan en el estudio están altamente correlacionadas la distancia a utilizar más apropiado debería ser la de Manhatan.

Cuando las variables son categóricas, a los efectos de definir que medida de similaridad utilizar, se debe tener en cuenta si las mismas son dicotómicas, nominales u ordinales.

Dismiliaridades utilizadas en la investigación

Una de las estrategias llevadas adelante en este trabajo implica la construcción de cluster con variables originales, todas de tipo categórico. En ese sentido se deben tener en cuenta que medida de similaridad utilizar. A los efectos de hacer los grupos se parte de la matriz de disimilaridades, para ello se recurre a la función daisy del software R. Esta función calcula una matriz de disimilaridad, a partir de una matriz de datos siguiendo un algoritmo diseñado por Kaufman y Rousseeuw (1990). La principal característica de esta función es su capacidad para manejar distintos tipos de variables, por ejemplo, nominal, ordinal, binaria, incluso cuando se quieren formar grupos usando variables de distinta naturaleza (cualitativas y cuantitativas).

Si las variables del conjunto de datos son del tipo cuantitativo, permite estandarizar los datos y elegir entre las distancias Euclídea y de Manhattan. Cuando alguna de las variables no es numérica se utiliza el coeficiente de disimilaridad de Gower (1971).

Esta función utiliza la siguiente medida:

$d(i,j) = \frac{\sum_{k=1} \delta_{ij}^k d_{ij}^k}{\sum_{k=1} \delta_{ij}^k}$

donde:

- δ_{ij}^k vale uno cuando las medidas x_{ij} y x_{ik} no son valores faltantes y cero en otro caso.
- δ_{ij}^k vale 0 cuando la variable k es binaria asimétrica y tenemos entre los individuos i y j un acoplamiento 0-0.
- El valor d_{ij}^k es lo que contribuye a la disimilaridad entre i y j la variable k.

R_k =maxx_{hk}−minx_{hk} donde h varía entre todos los individuos con valor de la variable k.

Si todas las variables son categóricas entonces $d(i, j)$ (la distancia entre los individuos i y j) da el número de acoplamientos del total de pares disponibles, es el coeficiente de acoplamiento simple, si las variables son binarias es el coeficiente de Jaccard.

Dado el conjunto de datos con $i = 1, ..., n$, utilizando las medidas de disimilaridades comentadas, se obtiene una matriz de dimensión $n \times n$ que tiene en la posición (i, j) la disimilaridad entre el elemento i y el j. Esta matriz resume la disimilaridad entre los individuos y es el insumo para la construcción de las tipologías.

A continuación se detallan los distintos procedimientos de agrupamiento que se van a utilizar, tanto jerárquico como no jerárquico.

2.2.2. Clasificación Jerárquica

Genera una serie de particiones encajadas. Los grupos que se forman a un nivel de distancia comprenden grupos obtenidos a un nivel de distancia inferior, (Blanco, J.,(2006)).

Una secuencia ascendente de particiones de I forman una jerarquía si y solo si para cada partición P_q y P_s con s > q todo grupo de P_s está contenido en P_q (Blanco, J.,(2006)).

En la clasificación jerárquica se pueden tener métodos que agregan y método que dividen. En un método agregativo se comienza con I clusters, uno para cada observación, y se finaliza con un único cluster que contiene las I observaciones. En cada paso una observación o cluster de ellas es absorbida en otro cluster. Los métodos jerárquicos divisivos comienzan con un cluster con I observaciones y se divide un cluster en cada paso.

2.2.2.1. Clasificación Jerárquica

Agregativa

En los métodos jerárquicos agregativos la agrupación se realiza mediante un proceso sucesivo cuyo resultado final es una jerarquía indexada, las observaciones que se unen en el paso k siguen juntas hasta que los individuos se unen en un solo grupo.

1. Al comienzo cada individuo es un grupo. Existen I grupos.

2. En cada paso el par de objetos más parecido se une para formar un nuevo grupo. En cada paso el número de grupos decrece en uno.

3. En cualquier paso los grupos que se forman son la unión de grupos formados en pasos previos. Si en una etapa dos individuos están juntos en un grupo estarán juntos en las etapas sucesivas.

4. En la última etapa se tiene un único grupo con los I individuos.

A efectos de seleccionar la cantidad final de grupos empleando métodos jerárquicos se utilizan reglas de detención. Se considera el aporte de los dendrogramas y los indicadores R^2, pseudo F y pseudo t^2. Al implementar métodos no jerárquicos para seleccionar la cantidad de grupos se trabajo con gráficos de siluetas propuestos por Kaufman y Rousseeuw, (1990).

En los métodos jerárquicos:

- Es necesario definir distancia entre los individuos, distancia entre los objetos y grupos (de más de un individuo) y distancia entre los grupos (de más de un individuo).
- Para definir las distancias entre grupos o entre individuos y grupos se definen distintos algoritmos de agregación: Método del vecino más cercano, Método del vecino más lejano, Método de Ward, Método del centroide.
- Si las distancias originales se transforman en una distancia ultramétrica, modificando lo menos posible las distancias originales entre objetos, se construye una jerarquía indexada. Se puede construir un "árbol de clasificación" o dendrograma, donde se representan las distintas particiones en cada paso y la distancia en que se fueron uniendo cada una de las mismas.
- El índice de agregación satisface las propiedades de una distancia ultramétrica. El índice de agregación de una clase queda definido por la distancia que determina el agrupamiento de los objetos en la clase.

- A toda jerarquía indexada le corresponde una distancia ultramétrica d.
- A todo distancia ultramétrica d le corresponde una jerarquía indexada.

2.2.3. Clasificación No Jerárquica

La clasificación no jerárquica, a diferencia de la jerárquica, necesita a priori tener información de la cantidad de grupos k. Las observaciones se agrupan, dependiendo del criterio de agrupación utilizado, en los k grupos definidos.

Estos métodos dividen a los datos en k conjuntos disjuntos de manera simultánea y como resultado de eso producen una clasificación en k clases sin relación entre ellas. La optimización de la función objetivo se realiza mediante un proceso iterativo. A continuación se detallan algunos de los algoritmos particionales más comunes.

2.2.3.1. Método de las k-medias

El algoritmo de k-medias(k-means) propuesto por MacQueen, J (1967) es la versión del algoritmo más conocida y ampliamente usada, ya que guarda un equilibrio entre la facilidad de implementación, velocidad y eficacia. Se lo denomina K-Medias porque representa a cada uno de los grupos por la media (o media ponderada) de sus puntos, denominado centroide. Estos puntos también se les suele llamar centros de gravedad, centro geométrico o puntos medios del cluster. En el algoritmo K-Means es necesario especificar a priori, el número k de grupos que desea formar para que el proceso se inicie.

El algoritmo es un procedimiento iterativo que sigue los siguientes pasos:

1 Se definen k cantidad de grupos.

2 Se escogen k puntos iniciales.

3 Se calcula la distancia que hay entre los centroides (puntos iniciales) $c^0 = c_1, c_2, ..., c_I$ y el resto de los puntos.

4 Se asignan las observaciones a los grupos en función de cuan próximo estén a los centro $c^0 = c_1, c_2, ..., c_k$

5 Una vez que se realiza la asignación, se vuelven a calcular los centroides de cada uno de los clusters $c^i = c_1, c_2, ..., c_k$

6 Con los nuevos centros de gravedad, se repite el proceso de reasignación hasta que no se presente un nuevo cambio de los k centroides.

El proceso se detiene si llega al número máximo de interacciones o se alcanza el umbral prefijado de reducción de la variabilidad intragrupos.

En términos generales, se puede decir que el objetivo del algoritmo K-Means es reducir al mínimo las distancia entre los puntos de cada grupo y su centroide.

Para ello se utiliza la siguiente función objetivo, que minimiza la suma de los cuadrados SSE entre cada observación y el centroide de su cluster.

$SSE = \sum_{i=1}^{k} \sum_{x \in c_k} ||x - u_k||^2$

donde el centroide u_k del cluster c_k es:

$u_i = \frac{1}{n_i} \sum_{x \in c_i}^{n} d(x_i, u_i)$

2.2.3.2. Particionamiento alrededor de los medoides (PAM)

Este algoritmo particional, fue propuesto en el año 1987 por Kaufman y Rousseeuw. Se considera como una variación del K-Means ya que usa la misma función objetivo con la restricción de determinar el mejor representante del centro de cada cluster (medoide). La función objetivo particularizada para el caso de la distancia entre un objeto x y su medoide u_i en un cluster c_i, es:

$min \sum_{i=1}^{k} \sum_{x \in c_i} d^2(x_i, u_i)$

Un medoide, de manera más formal, puede definirse como aquel objeto de un cluster cuyo promedio de disimilitud a todos los objetos en el cluster es mínimo.

El algoritmo PAM teóricamente sigue los siguientes pasos:

1 Calcula k objetos representativos (medoides).

2 Luego cada punto que no es un centro se agrupa a su medoide más cercano.

3 PAM intercambia los medoides con otros puntos candidatos, hasta obtener una configuración que minimice la función objetivo.

4 El proceso continua hasta que no se produzcan más intercambios de medoides.

Una de las mayores ventajas de este algoritmo es que puede tomar como entrada la matriz de disimilaridades. Estas disimilaridades, pueden ser por ejemplo, la evaluación subjetiva de relación entre objetos, en donde las medidas no son precisamente distancias. Esta importante condición permite que el algoritmo trabaje con diferentes métricas. Además, el algoritmo K-Medoids es menos sensible a la presencia de valores atípicos (outliers). Se considera que PAM es un algoritmo costoso en cuanto a la búsqueda de los medoides, ya que se compara un objeto con todo el conjunto de datos. Por lo tanto, es computacionalmente ineficiente para valores grandes de n y k.

2.2.4. Gráfico de silueta

El índice de silueta es una métrica interna que permite evaluar el buen funcionamiento de los algoritmos de aprendizaje no supervisado. El objetivo de este índice es identificar el número óptimo de agrupamientos. Para obtener el valor de $S(i)$ sólo es necesario tener dos cosas, los grupos obtenidos por la aplicación de un algoritmo de clustering y la colección de todas las proximidades entre los objetos. Posteriormente, se calcularán los valores:

- $a(i)$ que es la distancia media entre el objeto y todos los otros objetos de la misma clase y,
- $b(i)$ que es la distancia media entre el objeto y todos los otros objetos del cluster más cercano

El valor de $S(i)$ puede ser obtenido mediante la combinación de los valores de $a(i)$ y $b(i)$:

$$S(i) = \begin{cases} 1 - \dfrac{a_i}{b_i} \quad si \quad a_i < b_i \\ 0 \quad\quad si \quad a_i = b_i \\ \dfrac{b_i}{a_i} - 1 \quad si \quad a_i > b_i \end{cases}$$

De esta manera es posible expresar de forma genérica el valor del coeficiente de silueta bajo la siguiente ecuación:

$$S(i) = \frac{b(i) - a(i)}{max(a_i, b_i)}$$

Un valor más alto de este índice indica un mejor rendimiento del agrupamiento, ya que se está garantizando que la distancia inter-cluster es pequeña y la distancia intra-cluster es grande.

Este índice puede ser representado de manera gráfica, calculando el coeficiente de silueta para cada uno de los objetos. La silueta de cada cluster es graficada en orden decreciente para todos los objetos que conforman dicho agrupamiento. Con el fin de obtener una visión general. De esta manera toda la agrupación se puede mostrar por medio de una sola gráfica, que permite distinguir las agrupaciones "fuertes" de las "débiles". Un ancho de la silueta amplio, indica valores altos de $S(i)$ y por lo tanto un cluster más compacto, mientras que la otra dimensión de la silueta, la altura, indica simplemente el número de objetos en un determinado cluster. Así, el diagrama de silueta muestra cuáles objetos se encuentran bien agrupados dentro de su cluster, y cuáles están colocados de manera forzada o artificial.

2.3. Modelos de respuesta discreta

En esta sección se explica la metodología que será utilizada para alcanzar uno de los objetivos planteados. Se denomina modelos de respuesta discreta a aquellos modelos en los que la variable dependiente toma un conjunto discreto y finito de valores: 0, 1, 2,... Estos modelos reflejan las diferentes opciones cualitativas, excluyentes entre si, que pueden darse en una variable.

Cuando la variable de respuesta toma más de dos categorías, el modelo puede recibir el nombre de modelo de respuesta multinomial. Estos modelos tienen como objetivo pronosticar la pertenencia a un grupo a partir de una serie de variables independientes. La variable dependiente toma tantos valores como categorías haya, en este caso, la variable que se desea estudiar refiere a la clasificación de los individuos a los distintos grupos ya definidos, de esta manera Y indica a que grupo pertenece cada individuo.

2.3.1. Modelos de regresión logística multinomial

La Regresión logística multinomial es una técnica que pretende explicar el comportamiento de una variable cualitativa con más de dos categorías (puede ser de tipo nominal u ordinal) a partir de un conjunto de variables explicativas. Este tipo de regresión asume que los recuentos de las categorías de Y tienen una distribución multinomial. Esta temática se presenta siguiendo los fundamentos de Hosmer D.W. y Lemeshow S,(2013).

2.3.1.1. Modelo logístico para respuesta nominal

Estos modelos se aplican cuando la variable de respuesta es de tipo nominal, es decir, el orden entre las categorías es irrelevante. Se toma una de las categorías como referencia para comparar cómo cambian las probabilidades de las demás categorías respecto de ésta. Si la variable de respuesta Y presenta, por ejemplo, J categorías, se denota con $\pi_1, ..., \pi_J$ las probabilidades de cada categoría, es decir, con $\pi_j = \pi_j(x) = P(Y = j|x)\forall j = 1, ..., J$que satisfacen $\Sigma_{j=1}^{J}\pi_j = 1$ donde x representa el vector de las variables explicativas.

Si se toma como categoría de referencia la categoría (J) y se tienen K variables explicativas,$x = (x_1, ..., x_k)$, el *modelo logit* con respecto a ella se define como:

$$log(\frac{\pi_j}{\pi_J}) = \alpha_j + \beta_{j1}x_1 + \beta_{jK}x_K \quad donde \quad j = 1, ..., J - 1. \tag{2.1}$$

Además, mediante la ecuación general anterior se determinan los logits para cualquier pareja de categorías siempre tomando la misma categoría de referencia, en este caso la categoría J. Considerando dos categorías cualesquiera, por ejemplo 1 y 2, y aplicando propiedades de logaritmo se obtiene:

$$log(\frac{\pi_1}{\pi_2}) = log(\frac{\pi_1}{\pi_J}) - log(\frac{\pi_2}{\pi_J}) \tag{2.2}$$

La ecuación que expresa el modelo en términos de probabilidades de respuesta π_j es:

$$\pi_J = \frac{e^{(\alpha_j + x'\beta_j)}}{1 + \Sigma_{h=1}^{J-1} e^{(\alpha_h + x'\beta_j)}} \tag{2.3}$$

2.3.2. Métodos de selección del modelo

Un paso importante en la construcción de un modelo de regresión es la elección de variables a incluir. El modelo debe contener el menor número de variables explicativas posible que expliquen los datos (principio de parsimonia), y que además sea coherente e interpretable.

Para evaluar si las variables regresoras que se introducen en el modelo son significativas se analizan los contrastes de hipótesis de Wald y/ó el contraste condicional de razón de verosimilitudes.

Tanto los distintos contrastes del modelo que se utilizan como los procedimientos de selección de variables se encuentran en el Anexo Metodológico.

2.3.3. Bondad de ajuste

Para cuantificar la bondad del ajuste del modelo se analiza la tasa de clasificación global y la de los distintos grupos. Es decir, a partir del modelo ajustado, se clasifica cada observación en las distintas subpoblaciones y se construye una matriz de datos observados versus predichos. El porcentaje de clasificación correcta se utiliza como una medida de calidad de predicción.

Se construye la siguiente matriz:

	pred			
obs.	1	2	..	k
1	n_{11}			
2		n_{22}		
..			..	
k				n_{kk}

En la diagonal se encuentran la cantidad de observaciones donde el modelo acertó y fuera de la diagonal se encuentran los distintos errores.

La tasa de aciertos global se calcula como $\frac{n_{11}+n_{22}+..+n_{kk}}{n}$, el acierto en cada grupo es $\frac{n_{kk}}{n_K}$ $\forall k = 1,..,K$.

La tasa de error global se calcula como 1- tasa de acierto. Si el error de clasificación se calcula con la totalidad de las observaciones, dicho error presenta sesgo y subestima la verdadera tasa de error, a tales efectos se recomienda trabajar particionando la muestra, en muestra de entrenamiento y muestra de prueba.

2.4. Árboles de decisión – CART –

Para el desarrollo de la presente sección se siguió el libro *Classification and Regression Trees* (Breiman, L., Friedman, J., Olshen, R. A. & Stone, C. J. , 1984). Los árboles de decisión son técnicas multivariadas no paramétricas que se utilizan para encontrar reglas de clasificación (predicción), así como para realizar análisis descriptivos de los datos. El problema se puede presentar de la siguiente forma, dado un conjunto de datos $D = (X, Y)$, donde Y es la variable a explicar y $X = (X_1, ..., X_k)$ es un conjunto de k características que se miden a los individuos, el objetivo es predecir el valor de Y a partir de los valores observados de las variables X. Es decir, dado D se construye un predictor $\varnothing(x, D)$ que permita obtener estimaciones para valores desconocidos de Y.

Los árboles de decisión se pueden clasificar tomando en cuenta la naturaleza de la variable a explicar en:

- Árboles de clasificación: la variable dependiente es cualitativa.
- Árboles de regresión: la variable dependiente es cuantitativa.

En el caso de los árboles de clasificación, $Y?(1, 2, ..., J)$ y lo que se busca es clasificar a los individuos en alguno de los J grupos predeterminados usando k caracter?sticas $(X_1, ..., X_k)$

En los árboles de regresión, Y R y el objetivo es, al igual que en un modelo de regresión, obtener una estimación del valor de E(Y).

Esta técnica, también puede ser utilizada para seleccionar variables, es decir determinar cuáles características son las que mejor definen a las J clases. La relación entre Y y las X consiste en una función constante por conjuntos, los valores predichos de Y pueden expresarse como:

$E(Y|X=x) = \sum_{s=1}^{S} c_s I_{N_s}(x)$

Donde N_s representa una partición del espacio de las variables explicativas, I_{N_s} es la indicatriz que vale 1 cuando x $\in N_s$ y 0 en otro caso. La partición es tal que $N_s \cap N_i = 0$, por lo que c_s es la predicción de Y cuando x $\in N_s$. La determinación de la predicción de Y difiere según si el árbol es de clasificación o de regresión.

Árbol de clasificación $c_s = max_l\{\frac{card(Y=l)}{cardN_s}\}$

Árbol de regresión $c_s = \frac{1}{cardN_s}\sum_{i|X_i \in N_s} Y_i$

En particular los árboles de decisión que se utilizaron en este trabajo son del tipo CART (Classification and Regression Trees), que son técnicas de clasificación y regresión del tipo binarias. Estos árboles van generando una partición recursiva del espacio de representación a partir de un conjunto de reglas de decisión. Se parte de un nodo inicial que contiene a todos los datos, luego el nodo se particiona en dos nodos hijos de acuerdo a una regla de decisión, lo que se pretende es que los dos grupos resultantes sean lo más homogéneos posibles en su interior. Esta regla de decisión está basada en una única variable, y la misma se escoge de modo que la partición se haga en dos conjuntos lo más homogéneos posible.

2.4.1. Elementos básicos necesarios en el proceso de construcción del árbol.

El proceso de construcción del árbol es recursivo y se comienza con todo el conjunto de datos de entrenamiento *D = (Y,X)*. Luego se consideran un conjunto de particiones s, representadas por preguntas binarias del tipo x $\in$ Q , donde Q es un subconjunto de la muestra y se crea de acuerdo a una regla que toma en consideración una única variable. Se utiliza un criterio de bondad de ajuste para evaluar y determinar la mejor partición s, para cual es necesario contar con una medida de la impureza del nodo resultante. Una vez elegida la mejor partición, el conjunto de los datos es dividido en dos subconjuntos, luego en cada uno de los subconjuntos resultante se repite el proceso. El proceso continúa de este modo hasta que se verifica determinada regla de detención previamente definida o hasta que se obtienen nodos puros. A los nodos finales, resultantes de todo el proceso se los denomina terminales. Esta breve descripción da cuenta que en el proceso de partición recursiva se deben tener presente los siguientes elementos:

1. Conjunto de preguntas con respuesta binaria.

2. Criterio de bondad de ajuste de la partición que evalúa en cada nodo t la bondad de la partición s.

3. Regla de detención.

4. Regla para asignar cada nodo terminal a una clase.

2.4.2. Proceso de construcción del árbol óptimo.

La construcción del árbol se realiza en dos pasos, primero se construye un árbol maximal y luego se procede a la poda para obtener el árbol óptimo. Este procedimiento es válido tanto para árboles de regresión como árboles de clasificación. Si el proceso de partición se continúa hasta el final se obtienen nodos puros. El inconveniente es que el árbol resultante puede ser muy complejo, difícil su interpretación y no tener una buena performance en lo que a predicción se refiere ya que el mismo está muy ajustado a los datos de entrenamiento. Por tal motivo el árbol resultante es podado, cortando sucesivas ramas o nodos terminales hasta encontrar el tamaño adecuado del árbol. Una forma es considerar una secuencia de árboles anidados con el maximal de tamaño decreciente, luego estos son comparados para determinar el óptimo. Esta comparación está basada en una función de costo-complejidad, $R(T)$. Para cada árbol T, la función de costo$-$complejidad se define como:

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

donde R(T) puede ser la tasa de error global o la suma de cuadrados residuales total dependiendo del tipo de árbol, T es la complejidad del árbol, entendida como el número de nodos del subárbol y α es el parámetro de complejidad. De la secuencia de árboles anidados es necesario seleccionar el árbol óptimo para ello se evalúa tanto la complejidad como el poder predictivo del árbol resultante. Para evaluar el poder predictivo de la secuencia de árboles anidados se emplea un procedimiento de partición de la muestra o de validación cruzada. En el mismo se emplea una parte de la muestra para construir la regla de clasificación y la otra parte se emplea para predecir la variable de respuesta. Luego se obtienen los errores de predicción y se selecciona el árbol con el menor error de predicción. Breiman et al (1984) recomienda seleccionar el árbol donde se hace mínimo el error predictivo más un desvío estadar.

El criterio de bondad y ajuste de la partición que evalúa en cada nodo t, la bondad de la partición s, las reglas de detención y las reglas para asignar cada nodo terminal a una clase se encuentran en el Anexo Metodológico.

Capítulo 3

Análisis descriptivo

En este capítulo se realiza un análisis descriptivo de las principales variables utilizadas. En un primer apartado se presentan los datos, diseño muestral y el tratamiento realizado a los mismos, y en el segundo se describen algunas características sobresalientes de las variables a utilizar.

3.1. Datos utilizados

Se trabajó con datos correspondientes a la ENAJ 2013. La encuesta consta de 3.818 casos que representan aproximadamente 770.000 jóvenes entre 12 y 29 años de edad, residentes en localidades del país de 5.000 y más habitantes. A cada joven se le realizaron distintas preguntas que son resumidas en 753 variables.

El cuestionario se divide en 14 bloques, con información recabada relativa a características personales como: composición del hogar, educación, migración y predisposición migratoria, empleo, salud, uso de sustancias, opiniones y participación de la juventud, relaciones afectivas y sexualidad, victimización, conflictos con la ley y discriminación, tiempo libre e intereses, actividad física y cuidados[1].

En el presente trabajo, se toman en cuenta únicamente 6 módulos relativos a Composición del hogar, Educación, Trabajo, Opiniones de la Juventud, Sustancias

[1]Ver informe de ENAJ 2013 realizado por el instituto Nacional de la Juventud (INJU) disponible en http://www.inju.gub.uy/innovaportal/file/45835/1/informe-tercera-enaj-final.pdf

y Victimización, Conflictos con la Ley y Discriminación, a partir de los cuales se construyen 22 variables. En cuanto a la muestra, hay 2 observaciones que se desechan, una de ellas por tener la mayoría de sus valores NAs (con un peso igual a 2) y la otra por registrar un peso de 0. Finalmente la muestra consta de 3816 casos.

3.1.1. Diseño muestral

El detalle del diseño muestral se toma del informe realizado por el INE [2] .

Debido a que no existe un marco muestral actualizado de personas jóvenes para seleccionar la muestra de la ENAJ, se recurrió a seleccionar una submuestra aleatoria de la Encuesta Continua de Hogares (ECH) para los meses de enero a julio del año 2013.

El diseño muestral es aleatorio en dos fases de selección. Cada una de las dos fases implica varias etapas de selección [3]. El tamaño de muestra efectivo es 3.818 casos, obteniéndose así una tasa de respuesta del 91 % aproximadamente.

En el presente estudio se trabajó con 3816 dado que 2 observaciones presentaron datos faltantes.

3.1.2. Tratamiento de variables

Para realizar el análisis, de los 6 módulos tomados inicialmente, se construyeron 3 grupos que engloban las 22 variables a tratar. Un primer grupo refiere a las características personales, el segundo a pensamientos u opiniones y el tercero a los distintos comportamientos o conductas que toma el joven.

Características personales

Para caracterizar la población de jóvenes se cuenta con información sociodemográfica. Se dispone de datos respecto a sexo, edad, nivel educativo, si trabaja y características del hogar (donde reside, nivel de ingreso y composición).

[2]Informe disponible en http://www.ine.gub.uy/web/guest/encuesta-nacional-de-adolescencia-y-juventud

[3]Ver en Anexo Metodológico

La variable *Edad* fue recategorizada en 2 grupos, 12-19 y 20-29, fusionando los 4 tramos existentes originalmente.

La variable *Región* indica si el joven es de Montevideo o Interior

La información respecto al ingreso se refleja a través de la variable *Quintil de Ingreso*, para cada joven se tiene la información a que quintil de ingreso pertenece su hogar. Esta variable se recategorizó en tres tramos, fusionando las categorías *bajo−medio bajo* y *alto-medio alto*, de esta forma las modalidades de esta variable son *bajo−mediobajo*, *medio*, *alto−medioalto*.

A continuación se presenta en el siguiente cuadro un resumen de las variables, sus categorías y una breve descripción de las mismas.

Variable	**Categoría de la Variable**	**Descripción**
Sexo	M	Hombre
	F	Mujer
Edad	12-19	Jóvenes entre 12 y 19 años
	19-29	Jóvenes entre 20 y 29 años
Nivel.Ed	Primaria	Primaria
	Secundaria	Secundaria incompleta
	Secundaria Completa	Secundaria completa
	Terciaria	Terciario/Universitario, completo/incompleto
Región	Int	Interior
	Mdeo	Montevideo
Trab	Si	Trabaja
	No	No trabaja
Const.Hog	C.Padres	Vive con los padres
	S.Padres	Vive sin los padres
Q.Ingreso	Bajo.Medio.B	Quintil de ingreso bajo y medio bajo
	Medio	Quintil de ingreso medio
	Alto.Medio.A	Quintil de ingreso alto y medio alto

Cuadro 3.1: Características personales

Pensamientos/Opiniones

A continuación se detallan las variables utilizadas respecto a pensamientos y/u opiniones de los jóvenes en relación a la discriminación y a la seguridad.

La variable *Discrim* refleja si el individuo alguna vez se sintió discriminado.

Las variables *Seguridad* y *Seguridad.Centro.Ed* indican el estado de conformidad del joven frente al respeto que percibe en centros educativos y en cuanto a la seguridad respectivamente. Estas 3 variables son codificadas como binarias.

Variable	**Valores posibles**
Discrim	Vale 1 si alguna vez sintió discriminación y 0 en otro caso
Seguridad	Vale 1 si se siente satisfecho con la seguridad y 0 en otro caso
Seguridad.Centro.Ed	Vale 1 si se siente satisfecho con la seguridad en los centros.Ed y 0 en otro caso

Cuadro 3.2: Pensamientos/Opiniones

Actos/Comportamientos

En este grupo se toman en cuenta todas las variables que implican o se vinculan a distintas conductas del joven. Las variables son dicotómicas y reflejan si se registra o no algunas acciones 'indebidas'. A continuación se presenta una tabla resumen con las variables que se utilizaron en este módulo.

Variable	**Valores posibles**
Detenido	Vale 1 si alguna vez estuvo detenido y 0 en otro caso
Fuga.C	Vale 1 si alguna vez se fugo de la casa y 0 en otro caso
Robo.C	Vale 1 si alguna vez robo un comercio y 0 en otro caso
Robo.Centro.Ed	Vale 1 si alguna vez robo algo en un centro ed. y 0 en otro caso
Manejo.Sl	Vale 1 si alguna vez manejo sin libreta y 0 en otro caso
Porta.Arma	Vale 1 si alguna vez porto arma y 0 en otro caso
Golpea.Ap	Vale 1 si alguna vez golpeo apropósito y 0 en otro caso
Sust	Vale 1 si alguna vez estuvo involucrado sust. ilegales y 0 en otro caso
Daño	Vale 1 si alguna vez causo daño apropósito y 0 en otro caso
Pelea	Vale 1 si alguna vez participo en una riña o pelea y 0 en otro caso

Cuadro 3.3: Tipos de Conductas Delictivas

Tomando estas variables como referencia se construye la variable ***Indebido***, la cual indica si el individuo tomó alguno de estos comportamientos. Si el joven incurrió en alguna de las conductas antes detallados se etiqueta como 1, de lo contrario como 0.

3.2. Análisis

A continuación se presenta un análisis descriptivo de las variables utilizadas en el presente trabajo, se analizan la asociación de los distintos comportamientos con las características personales y las opiniones/pensamientos. El análisis que se lleva a cabo en esta sección tiene en cuenta los datos expandidos a toda la población de jóvenes.

Para realizar los análisis descriptivos del presente capítulo así como la modelización que se presenta en el capítulo de resultados se utilizó el software libre R. Se utilizaron bibliotecas existentes en el software y funciones específicas elaboradas por docentes del curso de Análisis Multivariado I de la Licenciatura en Estadística.

El siguiente cuadro presenta la distribución porcentual de los jóvenes que cometieron al menos un comportamiento indebido en relación a quienes no tomaron conductas indebidas.

Comportamiento del Joven	porcentaje
Al menos un comportamiento indebido	69.3
Ningún comportamiento indebido	32.7
Total	100

Cuadro 3.4: Distribución porcentual de los distintos comportamientos

Como se aprecia en el cuadro anterior, 7 de cada 10 jóvenes declara haber tenido alguna reacción indebida en algún momento. Se debe tener en cuenta que para ser considerado un comportamiento indebido se consideró que el joven registre alguna de las conductas señaladas en el cuadro 3.3, puede haber participado en una pelea, haber consumido alguna sustancia ilegal, haberse fugado de la casa o haber cometido un robo. Las distintas conductas tienen distintos niveles de gravedad, por lo que resumir las conductas en esa única variable no parece ser lo más adecuado para caracterizar a los jóvenes.

Tomando la distribución porcentual de las conductas indebidas por separado se destaca que la conducta indebida más frecuente es manejar sin libreta, alcanzando el 43.23 % de la población total, seguido de quienes consumieron alguna sustancia ilegal (39.3 %), por el contrario los comportamientos menos frecuentes refieren a quienes roban en comercios y roban en centros educativos, alcanzando el 4.81 % y 3.14 % respectivamente. La distribución porcentual del resto de las conductas varían entre 5.42 % para quienes declaran haber portado arma y 12.53 % para quienes manifestaron una pelea (11.01 % Detenido, 7.45 % Daño, 7.92 % Fuga de la casa, 7.16 % Golpea a propósito).

3.2.1. Características Personales y Conductas Indebidas

A continuación se presenta un análisis bivariado donde se analizan las características sociodemográficas y tipos de pensamientos del joven ante los distintos comportamientos. Se analizan los comportamientos según sexo, edad, nivel educativo, composición del hogar, ingreso, trabajo y región.

Comportamientos según Sexo

Al analizar los comportamientos indebidos de los jóvenes según su **sexo** se observa que del total de hombres, el 75.7 % manifiestan haber tenido conductas indebidas, en tanto en las mujeres, esta cifra alcanza el 58.5 %. Si bien se encuentra un comportamiento diferencial, ambos superan el 50 %. En relación a estos comportamientos, se destaca que las mujeres registran guarismos insignificantes por robo (comercio o en centros ed.) y/o porte de arma. En ambos sexos, los registros más altos se verifican en manejar sin libreta y consumir sustancias ilegales.

Comportamientos según tramo etario

Teniendo en cuenta las conductas según la **edad** se observan comportamientos claramente diferenciados, se puede apreciar que en el tramo de mayor edad (los mayores de 19) el 75.0 % declara haber tenido algún comportamiento indebidos, mientras que en los menores de 20 esa cifra es del 57.8 % .

Para ambos tramos de edad, el consumo de sustancias ilegales y manejar sin libreta son las conductas más registradas. Por su parte en el tramo de mayor edad

Variable	Porcentaje
Sexo	
Hombres	75.7
Mujeres	58.5
Grupo Etario	
Menores de 20 años	57.7
Entre 20 y 29 años	75.0
Nivel Educativo	
Primaria	58.9
Secundaria	66.2
Secundaria completa	64.7
Terciaria	75.2
Región	
Montevideo	66.1
Interior	68.4
Composición del Hogar	
Viven con padres	63.9
Viven sin padres	74.6
Trabajo	
Trabaja	77.6
No trabaja	57.6
Ingreso	
Bajo – medio bajo	70.7
Medio	70.6
Alto – medio alto	64.7

Cuadro 3.5: Porcentaje de conductas indebidas según las distintas variables socio-demográfica

se observa con mayor frecuencia el haber estado detenido, mientras en que en los más jóvenes se observan conductas más "físicas" como una pelea y hacer daño apropósito.

Comportamientos según nivel educativo

Al estudiar las conductas indebidas y la **educación**, se aprecia que al aumentar el nivel educativo, aumenta el porcentaje de jóvenes que manifiestan alguno de estos comportamientos, a excepción de quienes declaran tener secundaria completa, donde esta brecha en términos de porcentaje presenta una leve baja de 66.2 % para los que no terminaron secundaria a 64.7 % para quienes han finalizado educación secundaria. Esta asociación del nivel educativo con los distintos comportamientos indebidos puede deberse a la edad.

En aquellos jóvenes que tienen Primaria, la mayores conductas indebidas registradas son: el consumo de sustancias ilegales, manejar sin libreta y haber sufrido peleas (en ese orden), en cambio en los jóvenes de Educación Terciaria se mantienen el consumo de sustancias, en manejar sin libreta y se suma golpea apropósito.

Comportamientos según región

Si se analizan los comportamientos indebidos según **región** se observa que no existen diferencias significativas según la residencia. El 66.1 % de los jóvenes de Montevideo y el 68.4 % de los del interior dicen haber tenido alguna conducta indebida. Sin embargo algunos actos delictivos registran mayor frecuencia en el Interior que en Montevideo, como es el caso de consumir sustancias ilegales y portar arma.

Comportamientos según trabajo

Estudiando la **situación laboral** de los jóvenes se observa que en aquellos que declaran estar trabajando es donde se registra mayores comportamientos indebidos, 8 de cada 10 jóvenes que trabajan manifestaron tener conductas indebidas mientras que en los que no trabajan esa cifra es del 57.5 % (casi 6 de cada 10 jóvenes). Algunos actos delictivos registran mayor frecuencia en quienes trabajan frente a quienes no trabajan, como es el caso de robar algún comercio y/o estar detenido. Nuevamente, al igual que se mencionó en Educación la variable que podría estar determinando la asociación es la edad y no el trabajo, al ser más grande es más probable que trabaje.

Comportamientos según composición del hogar

Analizando los comportamientos de los jóvenes según la **composición del hogar**, se aprecia que quienes no viven con sus padres presentan un porcentaje mayor de comportamientos indebidos, la mayoría de estos jóvenes están comprendidos en la franja de mayor edad y con un nivel de estudios medio. En tanto, se destaca el mayor consumo de sustancias ilegales y manejo sin libreta se dan en mayor parte por jóvenes que viven con los padres. A su vez otras conductas de riesgo como manifestar pelea o robar comercio también es registrada mayormente por jóvenes que viven con los padres.

Comportamientos según ingreso

Al analizar las conductas indebidas según el **ingreso** de los hogares se observa que los hogares pertenecientes a la franja de ingreso bajo − medio bajo tienen un porcentaje más bajo de jóvenes que declaran haber tomado conductas indebidas. En estos hogares, el 64.7 % registra conductas indebidas, mientras que en el resto de los hogares ese porcentaje es del 70.6 %. Para aclarar esta situación, será necesario desglosar que conductas indebidas se cometen en mayor parte en cada una de las franjas porque estos registros pueden estar asociados al tipo de conducta.

Analizando los tipos de conductas que se registraron, se observa que en los jóvenes con ingresos bajos − medio bajos se encuentra la mayor cantidad de jóvenes detenidos y que también manifestaron haber participado en alguna pelea. Por otro lado, en el tramo de ingresos altos − medio altos se encuentran la mayor parte de jóvenes que consumieron sustancias ilegales.

3.2.2. Pensamientos y Conductas Indebidas

A continuación se analizan los comportamientos según pensamientos de los jóvenes respecto a la discriminación y a la seguridad. Al igual que el caso de las características personales los resultados mencionados son expresados en función de los cuadros adjuntos en Anexo Resultados (sección 6.2.1).

Comportamientos según Discriminación

Al analizar los comportamientos indebidos según lo que sintieron los jóvenes frente a la discriminación se observó que en aquellos jóvenes que se sintieron discriminados, el porcentaje de conductas indebidas es mayor frente a los que no fueron discriminados, 8 de cada 10 jóvenes que se sintieron discriminados registran conductas indebidas mientras que en los que no sintieron discriminación esa cifra equivale a 6 de cada 10 jóvenes.

Comportamientos según Seguridad

En cuanto a la seguridad, aquellos jóvenes que están insatisfechos, son los que registran el mayor porcentaje de comportamientos indebidos 76.5 %, mientras que en quienes dicen estar satisfechos esta cifra es del 64.2 %. Al analizar los comportamientos indebidos según el nivel de conformidad de los jóvenes con la seguridad se destaca que presentan el porcentaje más alto de insatisfacción con la seguridad quienes toman alguno de estos comportamientos.

Comportamientos según Seguridad en Centros Educativos

Al igual que el caso de seguridad en general, se observa que aquellos jóvenes que están insatisfechos con la seguridad en los centro educativos son los que tienen un mayor porcentaje de comportamientos indebidos 74.3 %, mientras que en los que dicen estar satisfechos esta cifra es del 64.9 %.

Teniendo en cuenta lo analizado anteriormente, se detecta que los valores más frecuentes para jóvenes que toma comportamientos indebidos son alcanzados por jóvenes de sexo masculino, comprendidos en el tramo de edad de 20-29 años que poseen estudios universitarios, están insatisfechos con la seguridad (en la calle y/o en centros educativos) y de alguna manera sintieron discriminación. Es importante tener en cuenta que tipos de comportamientos indebidos se registran, ya que estos tienen distintos nivel de gravedad, no es lo mismo manejar sin libreta que robar un comercio, en un caso es una falta y en el otro un delito.

Teniendo en cuenta que la variable *Indebido* resume conductas con distintos niveles de gravedad, es importante diferenciar estas distintas conductas y las posibles asociaciones, no solo considerar la variable de resumen. Por tal motivo en el siguiente capítulo se realiza un Análisis de Correspondencia Múltiple, en el se pueden estudiar las asociaciones entre modalidades y variables. Se analiza la asociación de las variables de conducta así como la asociación de estas en las características personales y pensamientos/opiniones.

Capítulo 4

Resultados

A lo efectos de construir tipologías de jóvenes en función de sus características, analizar asociaciones entre las distintas conductas y construir una posible regla de clasificación que discrimine entre grupos más o menos violentos, se llevaron adelante distintas modelizaciones, análisis de correspondencia múltiple, análisis de cluster, discriminante logístico y árboles de clasificación.

Se trabajó en 2 escenarios, el que sólo considera las variables de conducta y el que considera todas las variables (conductas, pensamientos y sociodemográficas) para la construcción de tipologías y reglas de clasificación.

En el primer escenario se manejaron 2 opciones: construcción de grupos con las variables originales o realizar el análisis de cluster con las variables que surgen del análisis de correspondencia.

A los efectos de una mejor visualización de los distintos escenarios, a continuación se presenta un esquema de los mismos:

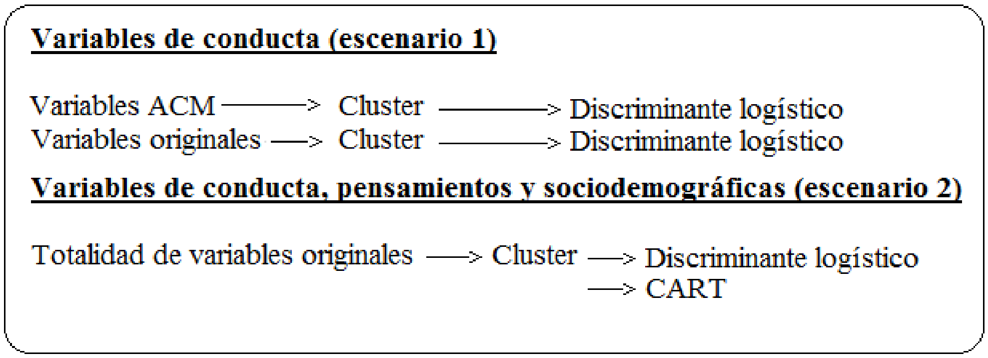

Figura 4.1: Distintos escenarios

Este capítulo se estructura en tres secciones que siguen el orden de las etapas planteadas en la metodología. Se debe tener en cuenta que los 3.818 jóvenes encuestados representan a la población de jóvenes entre 12 y 29 años residentes en localidades de 5.000 y más habitantes en Uruguay (2013). En ese sentido, cada joven tiene un peso diferente y considerar los mismos podría cambiar los resultados. Por tanto se evaluaron los resultados con datos expandidos y sin expandir y los mismos no cambian sustancialmente. Si existen asociaciones se verifican en ambos casos y cuando no existen se ven las distintas situaciones. Por esta razón se decide trabajar con los datos sin expandir. En Anexo Resultados (sección 6.2.1) se presenta un gráfico comparativo de ACM con datos expandidos y sin expandir.

Se realizaron varios ACMs con el fín de encontrar asociaciones entre las variables vinculadas a comportamientos, características personales y pensamientos. A suvez se analizan asociaciones entre las modalidades de las distintas variables. LosACMs se realizaron teniendo en cuenta la totalidad de las variables de cada grupo, alternando variables activas y suplementarias. Para el ACM descrito en la primer sección de este capítulo únicamente se toman las variables que reflejan los distintos comportamientos.

En la segunda sección se presentan los resultados de las distintas técnicas de clasificación no supervisada utilizadas, tanto jerárquicas como no jerárquicas. Se busca encontrar grupos de comportamiento en función de las características personales de los jóvenes y sus pensamientos. A los efectos de decidir cual es la mejor estructura de grupos se realizaron distintas aproximaciones y se analizaron las diversas conformaciones. Un elemento que se tomó en cuenta es la información proporcionada por los gráficos de silueta propuestos por (Kaufman y Rousseeuw (1990)).

Por último, en la tercera sección se analizan los resultados de distintas técnicas de clasificación supervisado utilizadas: discriminante logístico y árboles de clasificación. El objetivo es encontrar reglas de clasificación que permitan discriminar entre los grupos de los jóvenes más ó menos violentos y analizar que variables discriminan más entre uno y otro grupo. Los grupos son los que surgen del análisis de cluster y las variables explicativas son las características sociodemográficas y los pensamientos.

4.1. Análisis de Correspondencia Múltiple

Se realizaron varios ACMs teniendo en cuenta la totalidad de las variables, considerando distintas combinaciones de los tres grupos de variables como activas y suplementarias. Esta selección profundiza el ACM tomando únicamente las variables de comportamientos (10 variables). Se busca analizar asociación entre variables, modalidades, así como posibles tipologías de individuos.

4.1.1. Análisis de la Inercia

El ACM que se presenta aquí se realiza con 10 variables dicotómicas.

Para determinar cual es la dimensión del nuevo espacio, cuantos factores retener, se debe analizar la inercia total y la inercia acumulada por cada factor; Se tiene un porcentaje de inercia acumulada asociado a la cantidad de ejes elegidos. La cantidad de información conservada en cada factor se puede observar mediante la inercia asociada al mismo.

En este caso la inercia total es 1 $(\frac{Cant.Modalidades}{Cant.Variables} - 1 = \frac{20}{10} - 1)$

Como se puede apreciar en la tabla que se presenta a continuación, los ejes se presentan ordenados en función de sus valores propios. La primer columna muestra la cantidad de inercia que explica cada eje, la segunda va acumulando estos valores y la tercera representa la cantidad de inercia acumulada en porcentaje.

En este caso, para retener una parte importante de la inercia, parece necesario considerar un número elevado de ejes, ya que con 6 ejes solo se estaría conservando el 71,51 % de la información.

	inercia	**inercia acumulada**	**porcentual(%)**
Eje 1	0.258	0.258	25.85
Eje 2	0.105	0.363	36.39
Eje 3	0.095	0.459	45.94
Eje 4	0.091	0.550	55.07
Eje 5	0.083	0.634	63.45
Eje 6	0.080	0.715	**71.51**
Eje 7	0.077	0.792	79.28
Eje 8	0.070	0.863	86.37
Eje 9	0.070	0.933	93.38
Eje 10	0.066	1.000	100.00

Cuadro 4.1: Decomposición de la inercia

Sin embargo, podemos considerar que en el caso particular de ACM, los valores propios generalmente sobre presentan la capacidad de los ejes para retener la inercia. Por lo tanto, se utiliza el ajuste de los mismos mediante la fórmula propuesta por Benzecri en 1977 (Blanco, J., 2006). En concreto, se trata de un índice que permite ponderar la inercia explicada.

De esta forma, recomienda no utilizar aquellos ejes con una inercia inferior al inverso del número de variables (en este caso 1/10) para luego aplicar la función $(10/9)^2 * (i_s - 1/10)^2$, siendo i el auto valor de orden s, sobre cada uno de los valores propios restantes.

	inercia	**inercia aj.**	**porcentaje**	**porcentaje(acumulado)**
$[Eje1]$	0.25847	0.03100	0.99883	0.99883
$[Eje2]$	0.10545	0.00003	0.00116	1.00000

Cuadro 4.2: Inercia ajustada comportamientos indebidos

En este cuadro, se puede observar que el primer componente explica el 99,88 % de la inercia total, y que el primer plano contiene el 100 % de la misma. Por esta razón, al considerar únicamente este criterio la elección de dos componentes sería optima.

4.1.2. Representación de las modalidades

A continuación se observa que en el primer eje factorial no hay modalidades que conserven más del 45 % de su inercia original y por ende considerando este criterio no se encuentran bien representadas. A su vez, se destaca que con 2 ejes este número asciende a 6. En contraste, al incorporar un tercer componente, la cantidad de modalidades bien representadas aumenta a 9, igualmente se considera trabajar en el primer plano ya que no es conveniente agregar un componente adicional.

En base a este criterio, en el primer plano solo están bien representados: *Detenido.No*, *Daño.No*, *Robo.Centro.Ed* en su totalidad, *Fuga.casa.No* y *Golpea.Ap.No.* Se observa que no llegan a abarcar la mitad de las modalidades lo cual es una deducción esperable en ACM, ya que por lo general la mayoría de las modalidades no quedan bien representadas. En particular, esto implica que los resultados obtenidos respecto a las variables serían preliminares.

Modalidad	Cos^2**.1**	Cos^2**.2**	Cos^2**3**	Cos^2**.4**	**suma**Cos^2**.1.2**	**suma**Cos^2**.1.2.3**
Detenido.No	0.302	0.170	0.014	0.001	0.472	0.486
Detenido.Si	0.257	0.144	0.012	0.000	0.401	0.413
Daño.No	0.356	0.111	0.010	0.068	0.467	0.476
Daño.Si	0.313	0.098	0.008	0.060	0.411	0.419
Robo.centro.ed.No	0.230	0.356	0.092	0.061	0.585	0.677
Robo.centro.ed.Si	0.188	0.291	0.075	0.050	0.480	0.555
Pelea.No	0.228	0.009	0.365	0.090	0.237	0.602
Pelea.Si	0.202	0.008	0.324	0.080	0.210	0.535
Fuga.c.No	0.186	0.297	0.003	0.185	0.482	0.485
Fuga.c.Si	0.171	0.273	0.003	0.171	0.444	0.446
Robo.c.No	0.311	0.041	0.113	0.125	0.353	0.466
Robo.c.Si	0.258	0.034	0.094	0.103	0.293	0.386
Manejo.sl.No	0.261	0.052	0.019	0.184	0.314	0.333
Manejo.sl.Si	0.224	0.045	0.016	0.157	0.268	0.285
Porta.arma.No	0.280	0.032	0.057	0.084	0.312	0.369
Porta.arma.Si	0.252	0.029	0.052	0.075	0.281	0.333
Golpea.Ap.No	0.416	0.039	0.098	0.013	0.455	0.553
Golpea.Ap.Si	0.337	0.032	0.079	0.010	0.369	0.448
Sust.No	0.218	0.028	0.254	0.165	0.246	0.500
Sust.Si	0.196	0.025	0.228	0.148	0.221	0.448

Cuadro 4.3: Cosenos cuadrados de ACM con comportamientos indebidos

Por último, a través del cuadro de contribuciones presentado a continuación, se observan, 4 modalidades raras en el primer eje [*Robo.centro.Ed.No*, *Robo.c.No*, *Fuga.casa.No*, *Porta.Arma.No*] y 6 en el segundo [*Daño.No*, *Pelea* en su totalidad, *Robo.C.No*, *Porta.Arma.No* y *Golpea.Ap.No*]. Los cambios obtenidos al realizar el análisis sin estas modalidades, no resultaron relevantes, razón por la cual se decidió continuar el análisis incluyéndolas en el mismo.

Modalidad	**Contr.1**	**Contr.2**
Detenido.No	0.012	0.016
Detenido.Si	0.096	0.132
Daño.No	0.010	0.007
Daño.Si	0.119	0.092
Robo.centro.ed.No	0.003	0.010
Robo.centro.ed.Si	0.078	0.296
Pelea.No	0.010	0.001
Pelea.Si	0.073	0.007
Fuga.c.No	0.005	0.021
Fuga.c.Si	0.063	0.249
Robo.c.No	0.005	0.002
Robo.c.Si	0.104	0.034
Manejo.sl.No	0.040	0.020
Manejo.sl.Si	0.053	0.026
Porta.arma.No	0.006	0.002
Porta.arma.Si	0.097	0.027
Golpea.Ap.No	0.010	0.002
Golpea.Ap.Si	0.134	0.031
Sust.No	0.031	0.010
Sust.Si	0.048	0.015

Cuadro 4.4: Contribución de las modalidades a los ejes

Variable	Contr.1 (%)	Contr.2 (%)
Detenido	10.8	14.8
Daño	12.9	9.9
Robo.centro.ed	8.1	30.6
Pelea	8.3	0.8
Fuga.c	6.8	27
Robo.c	11.1	3.6
Manejo.sl	9.3	4.6
Porta.arma	10.3	2.9
Golpea.Ap.	14.4	3.3
Sust	7.9	2.5

Cuadro 4.5: Contribución de cada variable a cada eje factorial de ACM con comportamientos indebidos

En este cuadro se puede ver como se ven representadas las variables en cada factor. En particular, la contribución de cada variable a la inercia del factor es la suma de las contribuciones de sus modalidades. Aquí se puede apreciar, como el factor 1 y 2 se encuentra ligados mayormente a las modalidades Golpea.Ap y Robo.Centro.Ed respectivamente. A su vez, la contribución de cada variable a los ejes, da una idea sobre que variables es conveniente centrar el análisis de cada factor. En este sentido, el primer componente explicaría en forma más adecuada a las variables Daño, Robo.Comercio y Golpea.Ap, mientras que el segundo Fuga.c, Robo.Centro.Ed y Detenido.

A continuación se presenta la representación gráfica asociada a este análisis.

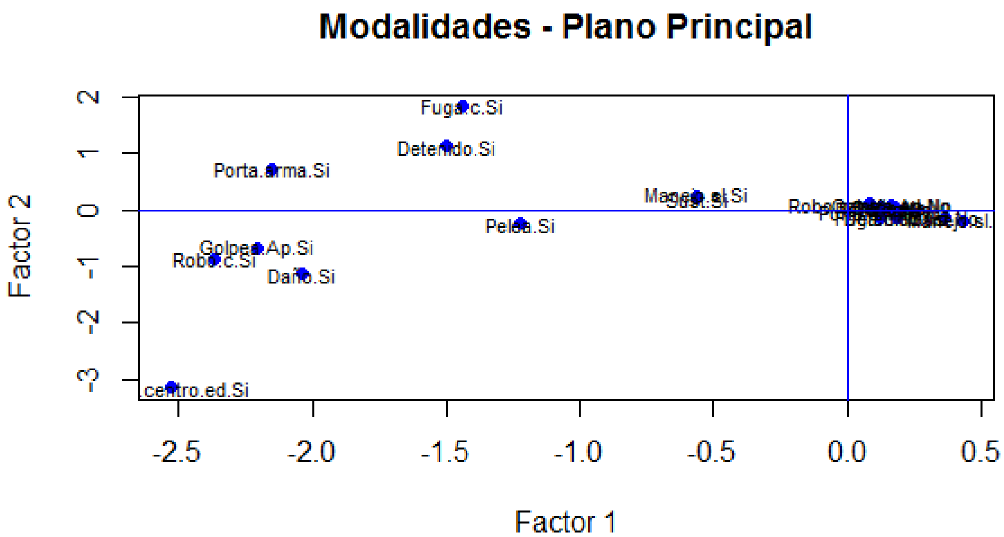

Figura 4.2: Plano factorial principal de ACM sobre comportamientos indebidos

De acuerdo a la figúra 4.2 se destaca que el primer eje factorial separa a la derecha a todas las modalidades que corresponden a no haber tomado conductas indebidas (coordenadas positivas). A su vez en el segundo eje se da una vista de la relación entre quienes golpean apropósito, hacen da no por gusto y roban comercio entre otros actos indebidos (coordenadas negativas) y tomando de esta manera contacto con terceros. Por otro lado se aprecia la asociación entre quienes manejaron sin libreta y consumieron alguna sustancia ilegal entre otros actos (coordenadas positivas próximas a 0).

Se podría decir que existen 3 grupos de individuos, los que:

[1] no tienen comportamientos indebidos o tienen manejo sin libreta y/o consumo de sustancias ilegales

[2] se fugan de la casa, portan arma y están detenidos

[3] golpearon, causaron da no y robaron

Se realizaron ACM con las variables pertenecientes a este módulo tomadas como activas y las variables de los otros módulos tomadas como suplementarias con el fín de ver si se cumple algún tipo de relación. Cabe destacar que las variables tomadas como suplementarias no participan a la hora de tomar las coordenadas del eje.

Obsrvando los gráficos, se apreció que las modalidades suplementarias aparecen muy baricentricas, aportando poca información en la contribución a estos ejes y dificultando el análisis. Por ello no se concluye nada en particular.

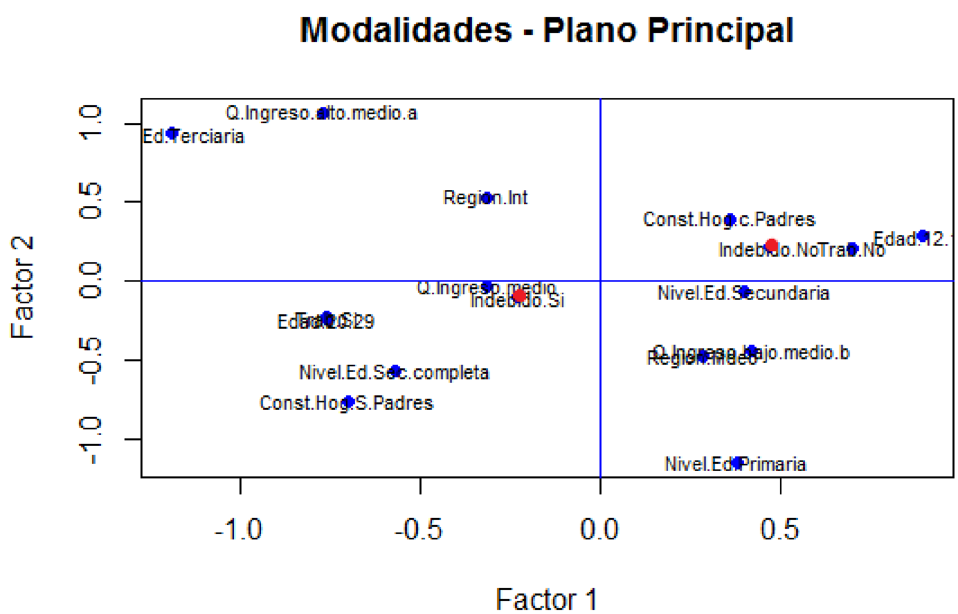

Figura 4.3: Plano factorial principal de ACM sobre características personales

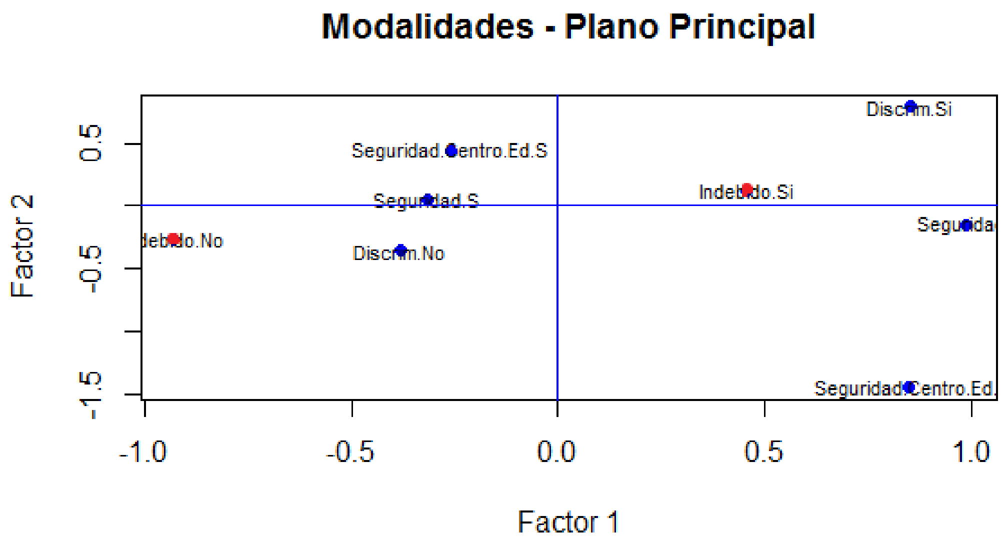

Figura 4.4: Plano factorial principal de ACM sobre pensamientos u opiniones

4.2. Análisis de Grupos

Acorde al objetivo planteado, en este apartado se busca la agrupacioón de los jóvenes en base a las distintas conductas. En el mismo se exploraron procedimientos jerárquicos y no jerárquicos, optando por el más adecuado para llevar a cabo la clasificación. Para seleccionar el procedimiento más eficiente y determinar la cantidad de grupos se determinaron medidas de similitud, distintos algoritmos de clasificación y se evalúan gráficos de silueta.

La formación de dichos grupos, permite ver que características los determinan, de forma que los elementos del grupo sean lo más parecidas entre sí, al tiempo, que se diferencien lo más posible de las observaciones de otros grupos.

Para la construcción de grupos se plantean 2 escenarios, en primer lugar únicamente se trata el set de variables que tienen que ver con los distintos comportamientos y se emplean 2 estrategias, por un lado se forman grupos teniendo en cuenta las nuevas variables que surgen del ACM y por otro se considera trabajar con las variables de conducta originales.

En el segundo escenario como alternativa adicional se decide formar grupos con la totalidad de variables originales.

4.2.1. Grupos en el primer escenario

4.2.1.1. Grupos con factores que surgen de ACM

Tomando la información que surge de los factores de ACM para la formación de dichos grupos, se realizan distintos procedimientos de tipo jerárquicos agregativos (Ward, Vecino más cercano y Vecino más lejano) para ver como se agrupan los jóvenes. Tras la observación de los resultados con los tres algoritmos se decidió continuar el análisis con el método de Ward, ya que este método forma grupos más esféricos y la estructura que alcanza no resulta compleja. Sin embargo, al tratar estos procedimientos formando 3,4 o 5 grupos (3 o 5 grupos son sugeridos mediante indicadores), se aprecia que en estos casos, en uno de los grupos se engloban más del 75 % de los jóvenes, por lo cual se descarta la aplicación de métodos jerárquicos tomando como referencia los factores.

A continuación se presenta el dendrograma obtenido.

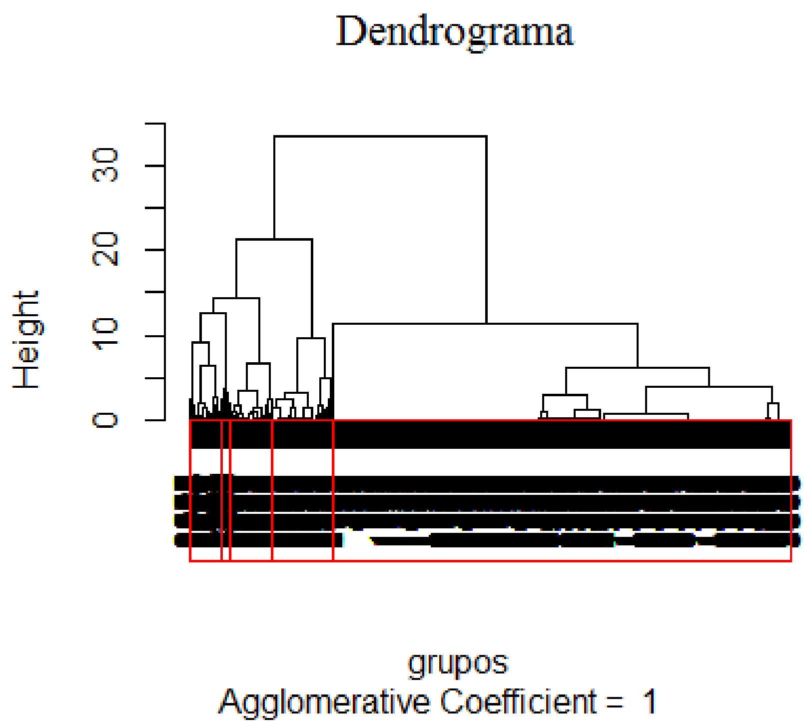

Figura 4.5: Dendrograma con el algoritmo de Ward

Cantidad de grupos	**Grupo 1**	**Grupo 2**	**Grupo 3**	**Grupo 4**	**Grupo 5**
3	530	0	0	382	2904
4	270	0	260	382	2904
5	213	57	260	382	2904

Cuadro 4.6: Distribución de jóvenes según grupos jerárquico

Los resultados que permiten decidir tanto el número de grupos sugerido como evaluar las distintas estructuras de agrupación se encuentran en el Anexo Resultados (sección 6.2.4).

4.2.1.2. Grupos con variables originales

En este caso, el análisis de grupos se emplea considerando las variables originales de conducta, este determina que el mejor ajuste se logra utilizando el procedimiento PAM tras formar 4 grupos, en este sentido se alcanza un índice valor promedio de silueta de 0.54.

A continuación se da vista que la mejor estructura de grupos se da a través del método PAM.

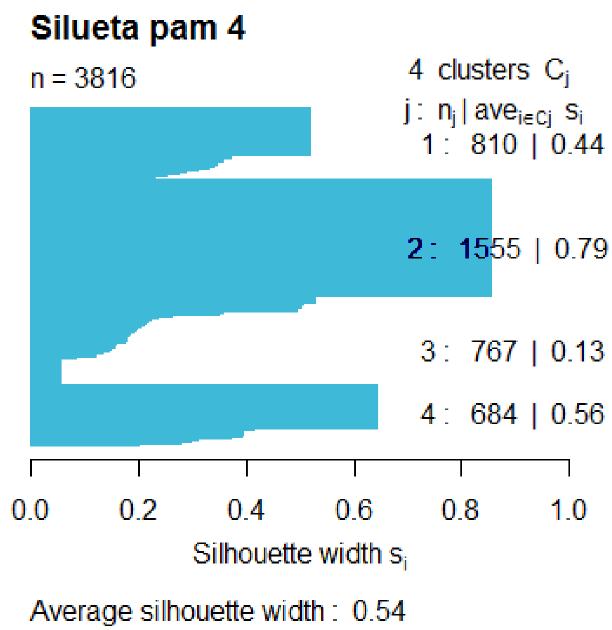

Figura 4.6: Gráfico de silueta del algoritmo PAM con las variables originales de conducta

Con los resultados alcanzados al desarrollar las distintas estrategias bajo el escenario 1, se observa que hay diferencia de trabajar con el set de variables originales de conducta y trabajar con la información de ACM. Tomando las variables originales se alcanzan grupos más homogéneos en cuanto a las dimensiones que presentan y se observa una estructura de silueta más fuerte, por ello al contrastar estas estrategias se consideran mejores los resultados que derivan del análisis de grupos con las variables de conductas originales.

4.2.2. Grupos en el segundo escenario

En este escenario se realizan tipologías utilizando todas las variables originales, las de conducta, pensamientos y sociodemográficas. Se implementaron distintos procedimientos Jerárquicos y no Jerárquicos y la mejor performance se encontró con el algoritmo de k−medoides (PAM). Se observó una disminución del valor del índice de silueta promedio al pasar de 4 a 5 grupos (de 0.19 a 0.16), por su parte no se registro variación al incrementar k, de 3 a 4 grupos. A continuación se presenta estas figuras.

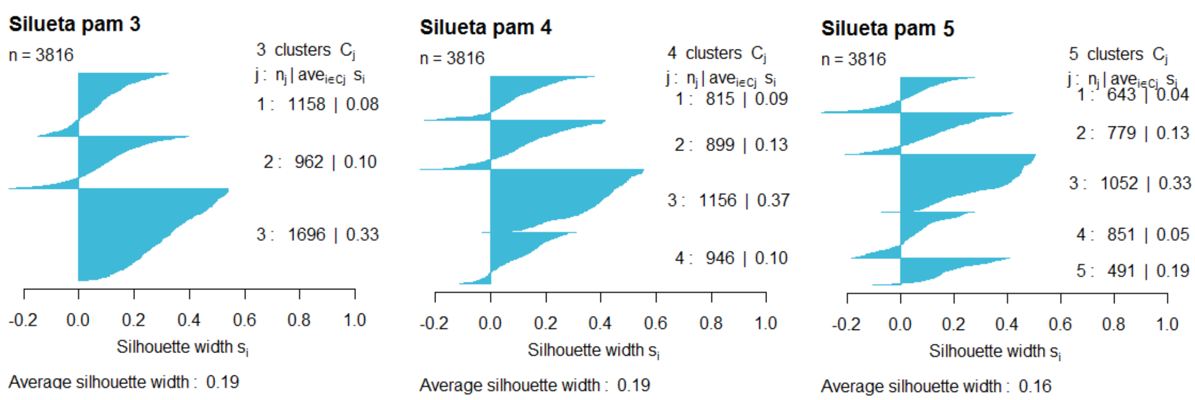

Figura 4.7: Siluetas PAM con todas las variables originales

Los resultados de implementar los distintos procedimientos variando el número de grupos se encuentran en el Anexo Resultados (sección 6.2.2).

La decisión de considerar 3 ó 4 grupos se apoyo por un lado en la información que aportó el gráfico de silueta y por otro en el análisis que se realizó en ACM donde podrán verse 3 grupos en función de comportamientos. En este sentido se decide trabajar con 3 grupos y caracterizar los mismos. La dimensiones de estos grupos son de 1.158, 962 y 1.696 jóvenes para los grupos 1,2 y 3 respectivamente.

Para determinar que características de conductas determinan los grupos construidos en este escenario, en la siguiente sección se presenta la distribución porcentual de la estructura de 3 grupos.

4.2.3. Distribución porcentual entre grupos según las distintas conductas

Como se vió en el cuadro 3.4 del capítulo anterior, más de la mitad de los jóvenes (69.2 %) declara haber tenido alguna conducta indebida en algún momento.

En particular, se busca encontrar diferentes grupos de jóvenes, determinados a raíz de las variables que los caracterizan. La formación de dichos grupos, permitirá ver que características los determinan, y así observar cómo se relacionan las variables de los comportamientos en estudio.

Con el fín de caracterizar los grupos construidos en el escenario 2 (totalidad de las variables), a continuación se presenta las frecuencias relativas de cada variable en cada grupo.

Conductas	**Grupo 1**	**Grupo 2**	**Grupo 3**
Detenido	16.2	17.5	2.3
Fuga de la casa	12.9	9.6	3.0
Maneja sin libreta	74.6	41.6	21.0
Porta arma	7.3	8.7	1.1
Sustancias	37.3	79.9	12.3
Daño a propósito	7.7	13.1	3.9
Golpea a propósito	9.3	9.9	3.2
Pelea	10.9	18.1	10.4
Robo	4.9	9.3	1.4
Robo centro educativo	4.9	4.6	1.4

Cuadro 4.7: Proporción de jóvenes que declara tener comportamientos indebidos por grupos

Teniendo en cuenta la información presentada, se puede caracterizar los grupos de la siguiente forma:

El grupo 1 en comparación con el resto de los grupos es un grupo en el que mayormente, las infracciones cometidas no intervienen terceros, siendo estas en su mayoría manejar sin libreta (74.6 %) o fugarse de la casa (12.9 %), también se registra un porcentaje importante en jóvenes que consumen alguna sustancia ilegal (37.7 %) no siendo el grupo donde mayormente se da esta conducta. Se destaca el valor registrado en el porcentaje de jóvenes que manejan sin libreta alcanzando casi el 75 % de los jóvenes del grupo. Este grupo contempla aproximadamente el mismo tamaño que el grupo 2, abarcando el 28.6 % de la población total.

El grupo 2 en comparación con los otros grupos es el grupo en que los jóvenes mayormente declaran haber golpeado apropósito (un 9.8 %), haber robado comercio/kiosco y en centros educativos (9.3 % y 4.6 % respectivamente), portado arma (8.7 %) y en su mayoría haber sufrido por lo menos una pelea/riña (18.0 %), se destaca el valor registrado en el porcentaje de jóvenes que consumieron sustancias ilegales abarcando el 79.9 % de los jóvenes de este grupo, por estas razones podría considerarse como el grupo más violento ya que en su mayoría estas infracciones involucran terceros y generalmente exponen a que el joven este ligado con la modalidad detenido, siendo el grupo que más individuos poseen esta modalidad (17.4 %).

Por lo contrario el grupo 3 contiene a los jóvenes que declaran haber manifestado la menor cantidad de comportamientos indebidos, de los cuales entre su mayoría se resalta haber conducido sin libreta (21.0 %) o haber consumido alguna sustancia ilegal (12.2 %), se aprecia que los registros de los comportamientos indebidos más altos en el grupo no son considerados relevantes, cabe destacar que es el grupo de mayor magnitud ya que abarca el 42.7 % de la población total.

4.3. Modelos de regresión logística multinomial

En esta sección se busca determinar si las variables sociodemográficas contribuyen a explicar los distintos grupos de conducta construidos mediante el análisis de cluster. Se trabaja siguiendo los escenarios vistos en cluster, grupos conformados únicamente con variables de conducta, escenario 1 (ACM y variables originales) y escenario 2 donde los grupos son conformados con las variables de conducta, sociodemográficas y de pensamiento.

A los efectos de estudiar si las características sociodemográficas contribuyen a explicar los grupos construidos con las variables de conducta, se realizó un modelo logístico. Se constató que trabajando con los grupos construidos únicamente con las variables de conducta originales (no ACM) la tasa de error global[1] fue de 62.7 % , por tanto se considera que con esta información no se obtienen buenos resultados. Al modelizar los grupos de comportamientos con las variables sociodemográficas se observa que las mismas no aporta en la clasificación de grupos, si bien los errores no son homogéneos en el grupo, el error global no es aceptable.

A continuación se decide modelizar los grupos que surgen del escenario 2, los construidos con variables de conducta y sociodemográficas. Se utilizan como variables explicativas las variables sociodemográficas. En la medida que como predictoras se utilizó un conjunto de variables que fueron utilizadas en la construcción de grupos, se espera que la performance del modelo mejore.

Para llevar a cabo este estudio como predictoras se seleccionan (inicialmente) todas las variables sociodemográficas y de pensamiento, Sexo, Edad, Nivel Educativo, Trabajo, Región, Quintiles de ingreso, Construcción del Hogar, Discriminación, Seguridad en Centro Educativos y Seguridad en General. La variable dependiente a modelizar son los 3 grupos que surgen del escenario 2 (guardada como *pam3*).

Y : "grupo de pertenencia", indica los grupos que surgen del análisis de cluster: 1, 2 y 3.

[1] Ver en Anexo resultado, sección 6.2.7

Para modelizar la probabilidad de pertenencia de cada individuo a los distintos grupos condicional a características sociodemográficas ($P(Y = J|X = x) = \pi_j$), se construyó un modelo discriminante logístico (logístico multinomial). El análisis trata de explicar la influencia de los factores sociodemográficos que llevan a clasificar los jóvenes a los distintos grupos. Se considera al grupo 2 ('más violento') como categoría de referencia. Teniendo en cuenta la formulación de los modelos realizados en el capítulo metodológico la ecuación (2.3), para el caso concreto de este trabajo queda formulada de la forma:

$$P(Y = 1|X = x) = \frac{e^{x'\beta_1}}{1+e^{x'\beta_1}+e^{x'\beta_3}} = \Pi_1 \tag{4.1}$$

$$P(Y = 3|X = x) = \frac{e^{x'\beta_3}}{1+e^{x'\beta_1}+e^{x'\beta_3}} = \Pi_3 \tag{4.2}$$

$$P(Y = 2|X = x) = \frac{1}{1+e^{x'\beta_1}+e^{x'\beta_3}} = \Pi_2 \tag{4.3}$$

Sujetos a la condición: $\sum_{k=1}^{3} \Pi_{ik} = 1 \forall i = 1, .., n$

Teniendo en cuenta la ecuación (2.1) del capítulo de metodología, en este caso los logits a estimar son:

- $log(\frac{\Pi_1}{\Pi_2}) = \alpha_1 + \beta_{11}x_1 + ... + \beta_{1k}x_p$
- $log(\frac{\Pi_3}{\Pi_2}) = \alpha_3 + \beta_{31}x_1 + ... + \beta_{3k}x_p$

donde $x_1, .., x_p$ son las características sociodemográficas de cada individuo.

4.3.1. Selección de la muestra de entrenamiento – prueba

Con el fín de seleccionar el mejor modelo que explique el grupo de correspondencia a cada jóvenes, se adoptó la estrategia de particionar la muestra en 2, muestra de entrenamiento y de prueba. Con la muestra de entrenamiento se estima el modelo y para evaluar su capacidad predictiva se usa la muestra de prueba.

Para obtener la muestra de entrenamiento se toma el 80 % de la muestra. Para ello se realiza un muestreo aleatorio simple estratificado, representando los estratos de cada grupo obtenido en la sección anterior. De los jóvenes pertenecientes a la muestra de entrenamiento un 28 % (874) pertenecen al grupo 1 y un 28 % (874) al grupo 2, mientras que un 43 % pertenecen al grupo 3 (1305).

Una vez seleccionada la muestra de entrenamiento se procede a estimar el modelo que mejor explique la clasificación de los jóvenes a sus respectivos grupos para luego evaluar la capacidad predictiva tomando la muestra de prueba.

La construcción del modelo más adecuado se va a obtener a partir de la selección de variables paso a paso. Este procedimiento permite realizar un stepwise de manera automática. Se toma el Akike Information criterious (AIC) asociado a los diferentes modelos que resultan en cada paso al añadir o quitar una variable y elige el modelo con menor valor de AIC. El procedimiento se detiene cuando no hay más variables que puedan incluirse o sacarse del modelo.

El modelo con menor valor AIC es el que se obtiene de incluir inicialmente la variable Edad. Esto es, de querer realizar un modelo con una única variable explicativa, el mismo sería constituido con la variable Edad. Este resultado confirma los resultados observados en el análisis descriptivo, donde se señaló que la edad es una característica que puede estar explicando los distintos comportamientos.

El modelo seleccionado finalmente incluye la totalidad de las variables sociodemográficas. A continuación se presenta información de los Logits del modelo seleccionado.

Variable	Logit 1	Logit 3
Intercepto	-5.06	-0.94
Edad 20−29	-0.85	-2.49
Q.Ingreso−Bajo.Medio.B	2.71	2.39
Q.Ingreso−Medio	2.10	1.35
Trab.si	0.25	-2.94
Const.Hog.S.Padres	3.16	0.07
Region.Mdeo	3.36	2.85
Sexo.F	1.96	2.74
Nivel.Ed.Sec.Completa	-0.34	-0.57
Nivel.Ed.Secundaria	0.61	0.12
Nivel.Ed.Terciaria	-1.91	-2.14
Discrim.No	0.49	1.17
Seguridad.I	-0.15	-0.72
Seguridad.C.Ed.I	0.15	-0.02

Cuadro 4.8: Coeficientes del modelo ajustado

4.3.1.1. Contraste y validación del modelo

Al realizar un contraste de razón de verosimilitud, para observar el efecto conjunto de las variables predictoras, se testea:

$H_0) \beta_1 = \beta_2 = ... = \beta_k = 0$

$H_1) algun \beta_k \neq 0$ con $k = 1, .., K$

Como resultado se obtiene el siguiente cuadro:

Model: Edad + Const.Hog + Region + Trab + Sexo + Q.Ingreso + Nivel.Ed + Discrim + Seguridad + Seguridad.C.Ed					
Resid. df	Resid. Dev	Test	Df	LR Stat.	Pr(chi)
1	6104	1337384.3	NA	NA	NA
2	6076	578710.7	28	758673.6	0

Para un nivel de significación se rechaza la hipótesis nula de que todos los coeficientes del modelo, a excepción de la constante, sean cero, por lo que se entiende que el modelo es significativo al 5 %. Cabe acotar que al construir un modelo paso a paso es de esperar que el modelo sea significativo y que en el test planteado anteriormente se rechace la hipótesis nula.

Se realiza una validación del modelo mediante los residuos de la devianza, considerando que los residuos que indican una falta de ajuste global son aquellos cuyo valor absoluto son mayores que 2 y se considera que la observación correspondiente es anormal. Al realizar un análisis descriptivo de los residuos se obtiene la siguiente información:

	media	desvío estándar	25 %	50 %	75 %	100 %
Grupo 1	-0.0090	0.3319	-0.16749	-0.09833	-0.04577	0.9779
Grupo 2	-0.0007	0.2905	-0.11372	-0.06057	-0.03195	0.9853
Grupo 3	0.0098	0.4013	0.05951	0.13981	0.25348	0.7576

Cuadro 4.9: Resumen de residuos del modelo

Como se puede observar en el cuadro 4.9, entre los máximos y mínimos de los valores que alcanzan los residuos, todos estos en valor absoluto son menores que 1, por lo que no hay ninguna observación que se considere 'atípica'.

Realizado el análisis de los contrastes para cada variable se observó que los parámetros son significativos, es decir cada variable es significativa ante la presencia de otra, los resultados se encuentran en Anexo Resultados (sección 6.2.8.2).

4.3.2. Bondad de ajuste

Como se mencionó en la metodología, una forma para comprobar si el modelo discrimina bien a los 3 grupos de jóvenes es observando la tasa de acierto global para las 3 categorías. Por tasa de acierto o de clasificación correcta se entiende el porcentaje de casos bien clasificados por el modelo, es decir le modelo clasifica igual a lo observado. Para evaluar el poder predictivo del modelo se adoptó como estrategia dividir la muestra en: muestra de entrenamiento y de prueba, 80 % y 20 % respectivamente. Con los datos de entrenamiento se construye la regla de clasificación y con los de prueba se evalúa el poder predictivo del modelo encontrado.

La tasa de clasificación correcta del modelo finalmente seleccionado (cuadro 4.10) alcanza un 82.0 % y en todos los grupos más del 75 % de los casos logran ser clasificados correctamente alcanzando un mejor nivel predictivo en el grupo 3.

A continuación se presenta la tabla de clasificaciones en términos porcentuales que se obtienen en la muestra de prueba:

	pronosticadas		
obs.	1	2	3
1	83.4	2.7	13.9
2	7.2	79.2	13.6
3	3.5	11.2	85.3

Cuadro 4.10: Bondad de ajuste del modelo

En el grupo 1 se han clasificado correctamente el 83.4 % de los individuos. En el caso del grupo 2 la tasa de clasificación alcanza una valor de 79.2 y en el grupo 3 se obtiene la mejor proporción de aciertos, un 85.3 % de individuos bien clasificados.

4.3.2.1. Odds ratio e intervalos de confianza

A continuación se presentan los odds ratios, para poder analizar el impacto de cada variable.

Para ver el impacto que tienen las variables explicativas en el modelo se decidió tomar como categoría de referencia el grupo más violento. El objetivo es expresar los resultados de estar en los distintos grupos frente a estar en el grupo más violento.

Las estimaciones de los intervalos de confianza al 95 % se encuentran en el Anexo Resultados (sección 6.2.8.1).

El siguiente cuadro muestra odds ratios estimados para las categorías de grupos 1 y 3 frente a pertenecer al grupo 2, siendo el grupo 2 el más violento.

Variable	Odds ratio 1/2	Odds ratio 3/2
(Intercept)	0.001	0.39
Edad 20-29	0.43	0.08
Q.Ingreso.bajo-medio.b	15.04	10.42
Q.Ingreso.medio	8.22	3.89
Trab.Si	1.30	0.05
Const.Hog.Sin.padres	23.69	1.04
Regio.Mdeo	28.86	17.39
Sexo.F	7.11	15.51
Nivel.Ed.Sec.Completa	0.71	0.56
Nivel.Ed.Secundario	1.86	1.14
Nivel.Ed.Terciario	0.15	0.12
Discrim.No	1.64	3.63
Seguridad.I	0.86	0.48
Seguridad.Centro.Ed.I	1.13	0.98

Cuadro 4.11: Odds ratio de los grupos 1 y 3 frente al grupo 2

Cocientes de probabilidades de estar en el grupo 1 y 3 frente a estar en el grupo 2

Tomando como grupo de referencia el grupo 2, se puede observar que no todos los cocientes de probabilidades son mayores que 1, por lo que no todas las variables personales actúan como factores predictores para los distintos grupos de conductas.

Los odds ratios menores a 1 ($\widehat{\beta}_k < 0$) indican que si la variable en cuestión es continua y se aumenta en una unidad de la misma dejando las demás constantes, disminuye la probabilidad de pertenecer al grupo 1 ó 3 respecto al 2.

En el caso de los quintiles de ingreso, se observa que quienes pertenecen al tramo de ingreso bajo $-$ medio bajo, dejando las demás variables constantes, el cociente de probabilidades de estar en el grupo 1 frente a estar en el grupo 2 es 15 veces mayor. A su vez el cociente de probabilidades de estar en el grupo 3 frente a estar en el grupo 2 es 10.4 veces mayor. Lo mismo pasa con los jóvenes que pertenecen al tramo de ingreso medio, dejando el resto de las variables constantes el cociente de probabilidades de estar en el grupo 1 frente a estar en el grupo 2 aumenta 8.22 veces más, mientras que el cociente de probabilidades de estar en el grupo 3 frente a estar en el 2 es 3.8 veces mayor. Los quintiles de ingreso discriminan más entre los grupos 1 y 2 que entre los grupos 3 y 2.

En el caso del trabajo, quienes se declaran activos trabajando, manteniendo el resto de las variables constantes, el cociente de probabilidades de estar en el grupo 1, frente a estar en el grupo 2 se incrementa en un 30 %.

Al analizar la constitución del hogar se observa que comparando quienes viven sin sus padres respecto a quienes viven con ellos, dejando las demás variables constantes el cociente de probabilidades de estar en el grupo 1 ante estar en el grupo 2 es 23.6 veces mayor.

En cuanto a la región, si el joven reside en Montevideo, dejando las demás variables constantes, el cociente de probabilidades de estar en el grupo 1 ante estar en el grupo 2 es 28.8 veces mayor, en cambio el cociente de probabilidades de estar en el grupo 3 ante estar en el grupo 2 es 17.3 veces mayor.

En el caso del sexo, si se es mujer, dejando las demás variables constantes, el cociente de probabilidades de estar en el grupo 1 frente a estar en el grupo 2 se incrementa es 7.1 veces mayor, en cambio el cociente de probabilidades de estar en el grupo 3 frente a estar en el grupo 2 es 15.5 veces mayor.

Analizando el nivel educativo (se sabe que esta variable tiene una asociaición muy grande con la edad), se aprecia que no todas las categorías tienen un impacto significativo, para quienes poseen un nivel educativo secundario incompleto se puede observar que, dejando las demás variables constantes el cociente de probabilidades de estar en el grupo 1 ante estar en el grupo 2 se incrementa en un 86 %, también se aprecia que en estos jóvenes (nivel educativo secundario incompleto) que el cociente de probabilidades de estar en el grupo 3 frente a estar en el grupo 2 se incrementa en un 14 %.

Respecto a los pensamientos de los jóvenes, en el caso de que el joven no haya sufrido discriminación, dejando las demás variables constantes el cociente de probabilidades de estar en el grupo 1 frente a estar en el grupo 2 se incrementa en un 64 %, en tanto, cociente de probabilidades de estar en el grupo 3 frente a estar en el grupo 2 es 3.6 veces mayor. El modelo presentado tiene en cuenta información sociodemográfica para explicar grupos conformados con variables de conducta.

4.4. Árboles de clasificación

Otro método para modelizar la pertenencia de los jóvenes a los distintos grupos de comportamientos, son técnicas de árboles de clasificación. Al igual que en el caso de los modelos logísticos y para comparar los resultados obtenidos con los mismos, se trabaja con los grupos que surgen del escenario 2. El árbol se construye con la muestra de entrenamiento y a través de la muestra de prueba se evalúa su poder predictivo.

Se trabaja con la muestra sin expandir ya que como se mencionó anteriormente los resultados no cambian sustancialmente.

Inicialmente las variables que participan como predictoras son los mismas que se tomaron para la construcción del modelo logístico: *Sexo*, *Edad*, *Nivel Educativo*, *Trabajo*, *Región*, *Quintiles de ingreso*, *Constitución del Hogar*, *Discriminación*, *Seguridad en Centro Educativos* y *Seguridad en General*.

La estrategia utilizada en la construcción del árbol es una de las sugeridas por Breiman et al (1984), la misma consiste en partir de un árbol maximal y luego proceder a su poda considerando la secuencia de árboles anidados que se generan. Para obtener el árbol óptimo se consideran en forma conjunta tanto la complejidad como el error global de clasificación.

Al construir un árbol maximal se obtiene la siguiente información: error de clasificación en el nodo raíz y los errores asociados a las distintas particiones. El error en el nodo raíz (*Root node error*) es igual a 0.5568 (1700/3053). A continuación se presenta la información para las distintas particiones.

obs	CP	particiones	error rel	xerror	xstd
1	0.3229	0	1.0000	1.0000	0.0161
2	0.1929	1	0.6770	0.6770	0.0157
3	0.0305	2	0.4841	0.4841	0.0144
4	0.0252	3	0.4535	0.4600	0.0141
5	0.0129	5	0.4029	0.4029	0.0135
6	0.0070	7	0.3770	0.3776	0.0132
7	0.0052	9	0.3629	0.3705	0.0131
8	0.0047	10	0.3576	0.3652	0.0130
9	0.0038	11	0.3529	0.3682	0.0131
10	0.0035	13	0.3452	0.3682	0.0131
11	0.0029	14	0.3417	0.3576	0.0129
12	0.0026	16	0.3358	0.3570	0.0129
13	0.0020	18	0.3305	0.3494	0.0128
14	0.0017	20	0.3264	0.3505	0.0128
15	0.0013	23	0.3205	0.3458	0.0128
16	0.0011	26	0.3164	0.3494	0.0128
17	0.0008	34	0.3064	0.3517	0.0128
18	0.0005	36	0.3047	0.3552	0.0129
19	0.0003	52	0.2947	0.3688	0.0131
20	0.0002	55	0.2935	0.3694	0.0131
21	0.0000	63	0.2911	0.3788	0.0132

Cuadro 4.12: Resumen árbol maximal

En el cuadro 4.12 se presenta información para distintos tamaños de árboles desde un único nodo al árbol maximal que en este caso tiene 64 nodos. El cuadro presenta el parámetro de complejidad (CP), la cantidad de particiones, el error de clasificación llamado error aparente o resubsititution error (*error rel*), el error de clasificación de crossvalidation[2] (*xerror*) y el desvío estándar (*xsdt*).

A los efectos de determinar el modelo final se considera el criterio propuesto por Breiman et al (1984) donde se sugiere considerar seleccionar el árbol que

[2]En el error aparente, se calcula los errores de clasificación con las mismas observaciones que se construye la regla de clasificación, en el error de crossvalidation dejando 1 observación fuera, se construyen las distintas reglas con $n-1$ observaciones. A los efectos de quitar la dependencia de los datos, la regla de clasificación se construye con $n-1$ observaciones y se clasifíca la observación que quedo afuera. El procedimiento se repite n veces. El error de crossvalidation es el error promedio de las observaciones clasificadas con las reglas construidas con $n-1$ observaciones.

presente el mínimo error, considerando los errores y sus desvíos (se selecciona aquel que donde el error + 1 desvío estándar sea minimo). En el cuadro 4.12 los errores están re−escalados llevando a 1 el error en el nodo raíz, en ese sentido a los efectos de calcular el error de cada árbol se debe multiplicar el error en el nodo raíz por el error de la fila correspondiente. Así en la partición 23 (árbol de tamaño 24) el error de crossvalidation sería igual a 0.3458*0.5568 = 0.1925.

Teniendo en cuenta la información aportada por el cuadro 4.12 y considerando la regla del mínimo error más un desvío se debería trabajar con el árbol de 24 nodos. Dado que la diferencia de los errores entre árboles de distinto tamaño, es muy chica se decide explorar la poda del árbol para distintos niveles del parámetro de complejidad.

A continuación se presenta el cuadro donde se observa la secuencia de árboles anidados podando el árbol maximal para diferentes parámetros de complejidad, las respectivas tasas de error y la complejidad del árbol resultante (nodos terminales). La tasa de error que se presenta en este cuadro es la tasa de error aparente.

Parámetro de complejidad	Tasa de error global (%)	Complejidad del árbol (nodos)
0.004	19.7	12
0.003	19.0	15
0.002	18.2	21
0.001	17.9	24
0.01	21.0	8

Cuadro 4.13: Secuencia de árboles anidados

Se busca encontrar la mejor relación entre la tasa de clasificación errónea y la complejidad del árbol, siendo la tasa de error de clasificación el cociente entre las observaciones mal clasificadas y el número total de observaciones y la complejidad del árbol el número de nodos terminales.

Al analizar los resultados presentados en el cuadro 4.13 se observa que la tasa de error de clasificación global es similar para árboles de tamaños que varían entre 24 y 12 nodos terminales (1.8 de diferencia entre un árbol de tamaño 24 y el de tamaño 12), en ese sentido se prioriza la construcción de un árbol más simple y se decide trabajar con un árbol de 12 nodos terminales.

Previo a la selección final del árbol se analizaron distintas muestras de entrenamiento y prueba, en cada caso se estudiaron las variables que intervienen en las particiones, los tamaños de los árboles y los errores de clasificación. El análisis se hace teniendo en cuenta que los árboles de clasificación son dependientes de los datos y se busca de esta forma descartar que la muestra seleccionada (tanto la de entrenamiento como la de prueba) no sea una muestra 'rara o atítipa'. Se seleccionaron varias muestras de entrenamiento y prueba y se construyeron árboles podando los mismos para distintos parámetros de complejidad (cp=0.004 y cp=0.003). En Anexo Resultados (sección 6.2.10) se presenta un cuadro de resumen y gráficos para 15 muestras.

Se puede observar que en la mayoría de los casos las variables que son utilizadas para la construcción de los árboles son las mismas, cambia en algunos casos quien realiza la primera partición (la variable *edad* o *trabaja*). En la mayoría de los casos la primera partición la realiza la variable *edad* y los nodos 1 y 2 se dividen por las variables *quintil de ingreso* y *trabajo*, respectivamente. Los tamaños de los árboles varían entre 11 y 13 nodos, los errores de clasificación globales y en cada grupo son similares.

El árbol finalmente seleccionado tiene 12 nodos terminales y se corresponde con una tasa de error global de 19.7 % y un cp de 0.004. A continuación se observa la figura que adopta el árbol seleccionado.

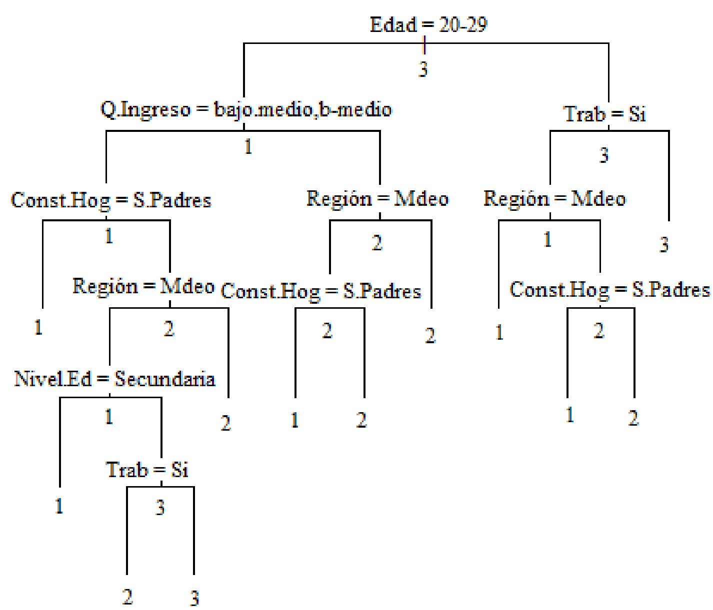

Figura 4.8: Árbol seleccionado

Las variables que participan a clasificar a los jóvenes en los distintos grupos son: *Edad*, *Trabajo*, *Quintil de Ingreso*, *Nivel Educativo*, *Región* y *Constitución del Hogar*, siendo la *Edad*, el *Ingreso* y el *Trabajo* las que generan las particiones. Si se considera un árbol de mayor complejidad se aprecia que el algoritmo recurre inicialmente a las mismas variables (ver en Anexo Resultados, sección 6.2.9).

Si se analiza el error de clasificación de la muestra de entrenamiento, se aprecia que el árbol con 12 nodos presenta una tasa de error global del 19.7 %.

El siguiente cuadro presenta una comparación entre lo observado y lo predicho para cada uno de los 3 grupos, para el caso de la muestra de entrenamiento.

	predichas		
obs.	1	2	3
1	84.2	9.6	6.2
2	13.4	74.6	12.0
3	12.6	6.4	81.0

Cuadro 4.14: Poder predictivo del árbol seleccionado (muestra de entrenamiento)

Las tasas de error para los distintos grupos son del 15.8 %, 25.4 % y 19.0 % para los grupos 1,2 y 3 respectivamente. Se observa que el mayor error se da en la clasificación al grupo 2 (25.4 %).

Para evaluar la perfomance del modelo para los casos en que se aplicará a nuevas observaciones, se utilizaron las observaciones de la muestra de prueba.

Para analizar la performance del modelo es necesario evaluar el comportamiento de la muestra de prueba, ya que la evaluación de los errores en la muestra de entrenamiento no da una idea de lo que puede suceder cuando se aplique el modelo a nuevas observaciones. Se tomó la misma muestra que se empleó para evaluar el poder predictivo del modelo logístico como forma de asegurar resultados comparables. El cuadro que se presenta a continuación muestra el comportamiento del modelo para el caso de las observaciones de la muestra de prueba.

	predichas		
obs.	1	2	3
1	83.4	10.6	6.0
2	16.7	74.5	8.8
3	13.8	5.8	80.4

Cuadro 4.15: Poder predictivo del árbol seleccionado (muestra de prueba)

Se aprecia en este caso que la tasa de error global es de 20.2 %. Las tasas de error para los distintos grupos son del 16.6 %, 25.5 % y 19.6 % para los grupos 1, 2 y 3 respectivamente.

Si se comparan estos resultados con el poder predictivo del modelo logístico se aprecia que ambos tiene una performance de predicción similar, el árbol alcanza una tasa de clasificación correcta de 79.8 % mientras que el modelo logístico llega a 82.5 %.

Por otro lado si se analiza el comportamiento de la clasificación del modelo logístico y el árbol de clasificación en cada uno de los grupos, tomando la muestra de prueba en ambos casos se observa que el primero presenta mayor nivel de acierto en el grupo 2 (los más 'violentos') siendo 79.2 % mientras que en el árbol de clasificación el nivel de acierto para el grupo 2 es de 74.5 %.

Si bien en los árboles de clasificación no se pueden calcular los aportes marginales de cada variable, se pueden visualizar claramente las distintas trayectorias que pueden llevar a clasificar a un joven a un grupo más o menos 'violento'.

A modo de ejemplo se puede ver que quienes están comprendidos en el tramo de mayor edad, pertenecen al quintil de ingreso, alto medio alto y residen en el interior, pertenecen al grupo más violento. A su vez, quienes son mayores de 20 años, pertenecen al quintil de ingreso bajo medio bajo o medio y viven sin los padres se clasifican en el grupo 1. Por otro lado, los más jóvenes que se encuentran en el tramo de edad más baja ($12 - 19$ años) y no están trabajando se clasifican en el grupo 3.

Capítulo 5

Conclusiones

En un principio se consideró que puede haber diferencia en cuanto a las distintas conductas según el género, los niveles de estudios, ingresos y características del hogar donde reside el joven ya que siguiendo la literatura hay autores que reafirman la predominación masculina sobre la femenina en el caso de actividades delictivas, como es el caso del *'Informe mundial sobre la violencia y salud'* (*OPS*, 2002), donde también son mencionadas las políticas que contribuyen a mantener desigualdades económicas y nivel educativo.

Si bien el análisis realizado en este trabajo arroja asociaciones entre ciertas conductas indebidas y la edad, el sexo, la composición del hogar y los quintiles de ingreso, el mismo no es concluyente.

Mediante ACM se identificaron asociaciones entre distintas modalidades. Por un lado los jóvenes comprendidos en el tramo de menor edad (12−19), viven con los padres y no trabajan, son quienes menos infracciones cometen y quienes están más a gusto en términos de seguridad. Al estudiar las variables de conductas y sus modalidades se pueden ver asociaciones entre las modalidades que implican no cometer faltas o delitos (coordenadas > 0, eje 1) y las "faltas o delitos más leves" manejar sin libreta y consumir alguna sustancia ilegal. El eje 2 separa las faltas más violentas o menos violentas.

Al analizar la relación de tener o no tener comportamientos indebidos (*Indebido*) y las variables de pensamientos, se identificaron claramente 2 tipologías de jóvenes, quienes se sienten satisfechos con el respeto de la seguridad en general y por ende que no se sienten discriminados, asociados a quienes declaran no haber cometido infracciones (no comportamientos indebidos) y, quienes no están satisfechos con el respeto de la seguridad, se sienten discriminados y han tenido conductas indebidas.

Para la construcción de tipologías, se plantearon 2 escenarios, inicialmente teniendo en cuenta únicamente la información de las variables de conducta (escenario 1) y en el segundo escenario considerando la totalidad de variables de conducta y sociodemográficas (escenario 2).

Se observó que existen diferencias entre conductas de acuerdo al grado de infracciones que toman los jóvenes.

Al implementar el análisis de grupos bajo el escenario 1 (incluyendo variables de conducta), bajo el procedimiento PAM se determinaron 4 grupos. Se consideró el grupo 1 como el grupo "más violento", aquel en que cual están integrados la mayoría de los jóvenes que entre sus distintos comportamientos involucran/dañan a terceros. El grupo "menos violento" se consideró aquel que contiene jóvenes con menor frecuencia de actos indebidos y los actos cometidos no involucran terceros, salvo en conducir sin libreta, esta modalidad mayormente esta ligada al grupo 3, y se consideró el grupo 2 como el "menos violento" por contener la menor población de jóvenes que cometieron infracciones, en tanto estas infracciones no involucran terceros. Cabe destacar que la mayoría de los comportamientos indebidos son realizados por hombres, lo cual concuerda con los antecedentes estudiados.

Por otro lado los resultados alcanzados bajo el escenario 2 (construcción de grupos con la totalidad de las variables originales), determinaron 3 grupos, menos violento, moderado y más violento. El grupo 2 se lo etiquetó como el grupo "más violento" ya que contiene la mayoría de los jóvenes que entre sus distintos comportamientos involucran daños a terceros y a su vez se registra las mayores frecuencias en todas las infracciones. El grupo 3 es considerado como el grupo "menos violento" por ser el grupo con menos infracciones y ser además infracciones que no involucran a terceros.

El grupo 3, el menos violento, es el de mayor tamaño, compuesto en su mayoría por mujeres que pertenecen a los quintiles de ingreso bajo $-$ medio bajo, no trabajan, se encuentran en el tramo de edad de 12-19 y viven con los padres. A su vez, en cuanto a los pensamientos, estos jóvenes son quienes están en su mayoría satisfechos con la seguridad en general y no sintieron discriminación. El grupo 1, considerado como grupo moderado, contiene la mayoría de jóvenes que manejaron sin libreta y una porción significativa de jóvenes que consumieron alguna sustancia ilegal. Los grupos 1 y 3 son grupos de comportamientos indebidos que no involucran terceros.

Se construyeron modelos de clasificación a los efectos de, por un lado estimar la probabilidad de que un joven se encuentre en un grupo más violento (modelos logísticos) y por otros analizar las trayectorias que puedan determinar que un joven sea más violento (árboles de clasificación).

En los modelos donde se buscó explicar los grupos de comportamiento del escenario 1, con las variables sociodemográficas, no se obtuvo un buen poder predictivo.

Si se tienen en cuenta los grupos conformados con variables de comportamiento y sociodemográficas y se modeliza considerando las características sociodemográficas como explicativas se encuentran modelos con mayor poder predictivo.

En este contexto se puede observar que, ser mujer aumenta la probabilidad de pertenecer a los grupos menos violentos. Así como también se ve que los cocientes de probabilidades de pertenecer a los grupos 1 y 3 frente a estar en el grupo 2 son mayores para quienes residen en Montevideo y pertenecen a los quintiles de ingreso bajo y medio bajo. También se aprecia un impacto positivo de pertenecer al grupo 1 frente a estar en el grupo 2 (más violento) en aquellos jóvenes que viven sin los padres.

La composición del hogar discrimina más claramente entre los grupos 1 y 2 que entre los grupos 3 y 2.

Por otro lado, mediante el proceso de construcción árbol de clasificación estimado se observó que las variables incluidas se encuentran en concordancia con las consideradas para el modelo logístico. Las variables que contribuyen más a discriminar entre los grupos, son la edad del joven, el quintil de ingreso al que pertenece, su condición laboral, lugar en donde reside, la constitución del hogar y el nivel educativo del joven, coincidiendo estas (en el mismo orden) con la incorporación de las variables al construir el modelo de regresión logística.

Tanto los modelo logísticos como los árboles de clasificación tienen un poder predictivo del orden del 80 %.

De acuerdo a estas observaciones se concluye que si bien se pueden encontrar alguna característica personal como el caso del sexo, no se encuentran variables sociodemográficas que contribuyan a explicar las tipologías.

A los efectos de poder determinar claramente la asociación de ciertos comportamientos indebidos con características sociodemográficas se podrían utilizar a futuro, otras técnicas de clasificación supervisada que podrían tener una mejor performance que los modelos aquí utilizados. A su vez, se podrían trabajar los modelos aquí estudiados incluyendo interacciones (edad$*$trabajo, edad$*$sexo, entre otras) y profundizando el estudio dentro de cada grupo de edad, realizando análisis por separado para los menores de 20 y para los mayores de 20 años.

Bibliografía

1. Blanco, J. (2006). *Introducción al Análisis Multivariado*. Facultad de Ciencias Económicas y de Administración, Universidad de la República, Montevideo.

2. Breiman, L., Friedman, J., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*. Chapman and Hall.

3. Caballero, N. y Jadra, G. (2013). *Caracterización de los jóvenes Uruguayos que no asisten al sistéma educativo*. Informe final de pasantía. Facultad de Ciencias Económicas y de Administración. Universidad de la República, Montevideo.

4. Castrillejo, A. y Nalbarte, L.(2014). Apuntes del curso: *Análisis Multivariado*. Facultad de Ciencias Económicas y de Administración, Universidad de la República, Montevideo.

5. Escoffier, A Pages, J. (2008).*Analyses Factorielles Simples et Multiples 4ta edición*. DUNOD.

6. Farraway, J. (2004). *Linear Models with R*. Tylor & Francis.

7. Fernandez, V. and San Martín, R. (2004).*Regresión logística multinomial, cuadernos de la Sociedad Española de Ciencias forestales*. pp. 323−327.

8. García Álvarez, A. (2013). *Patrones de multimorbilidad mediante Análisis Clúster con R*. Universidad de Granada (UGR), España.

9. García, T. y Montero, C. (2008). *Aplicación de la regresión logística multinomial en la detección de factores económicos que influyen la productividad de los sectores industriales*. Ingeniería UC , pp. 19 − 24.

10. Hosmer, D.W. and Lemeshow, S. (2000). *Applied logistic regression*. Second edition. New York: John Wiley and Sons.

11. Instituto Nacional de la Juventud (INJU). (2013). *Informe tercera Encuesta Nacional de Adolescencia y Juventud*, Montevideo.

12. Lumley, T. (2016). survey: analysis of complex survey samples. R package version 3.31-5.

13. Organización Panamericana de la Salud.(2002). *Informe Mundial sobre la Violencia y Salud*. Movi Mundy. Washington, D.C.

14. R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

15. Rosseeuw, P.and Kaufman, L. (1990). Finding Groups in Data. Beljium. Wiley − Interscience.

16. Salazar, A. (2008). *Modelos de respuesta discreta en R y aplicación on datos reales* . Universidad de Granada (UGR), España.

17. Therneau, T., Atkinson, B. and Brian Ripley (2015). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-10. https://CRAN.R-project.org/package=rpart .

18. Tillé, Y. and Matei, A. (2016). Sampling: Survey Sampling. R package version 2.8. https://CRAN.R-project.org/package=sampling .

19. Urrestarazu, I. (2014). Apuntes del curso: *Modelos Lineales.* Facultad de Ciencias Económicas y de Administración, Universidad de la República, Montevideo.

20. Venables, W. N. and Ripley, B. D.(2002). Modern Applied Statistics with S. Fourth Edition. New York: Springer.

Capítulo 6

Anexos

6.1. Anexo Metodológico

Fases del diseño muestral

El diseño muestral de la ECH (primera fase) es aleatorio, estratificado en dos o tres etapas de selección.

Luego, el diseño de la segunda fase (muestra ENAJ) es en tres etapas de selección. En la primera etapa, se conformaron 6 regiones geográficas:

Región	**Departamentos**
Metropolitana	Montevideo - Canelones - San José
Frontera con Brasil	Artigas - Cerro Largo - Rivera - Treinta y Tres
Costa Este	Canelones - Maldonado - Rocha
Litoral Sur	Colonia - Soriano - San José
Centro y centro Sur	Durazno - Flores - Florida - Lavalleja - Tacuarembó
Litoral Norte	Paysandú - Río Negro - Salto

Cuadro 6.1: Regiones geográficas para el diseño

Dentro de cada una de las regiones (a excepción de la región Metropolitana) se sortearon departamentos con probabilidad proporcional al tamaño en términos de la cantidad de jóvenes según estimaciones provenientes de la ECH.

Región	**Departamentos**
Metropolitana	Montevideo - Canelones - San José
Frontera con Brasil	Artigas - Rivera
Costa Este	Canelones - Maldonado - Rocha
Litoral Sur	Colonia - San José
Centro y centro Sur	Florida - Tacuarembó
Litoral Norte	Paysandú - Río Negro - Salto

Cuadro 6.2: Departamentos sorteados

En la segunda etapa, para cada uno de los departamentos seleccionados (a excepción de la región Metropolitana), se sortean bajo un muestreo aleatorio simple hogares en donde reside al menos una persona entre 12 y 29 años de edad. Para la región metropolitana se realiza un muestreo aleatorio simple estratificado, donde los estratos corresponden a los de la ECH.

Finalmente, en la tercera etapa se sortea una persona entre 12 y 29 años en cada uno de los hogares seleccionados.

El tamaño de muestra teórico es de 4200 personas y se distribuye por región de la siguiente forma:

Región		Tamaño de Muestra
	Todo el país	4200
	Total	1722
Montevideo	Bajo	342
	Medio Bajo	368
	Medio	453
	Medio Alto	387
	Alto	171
	Total	2478
Resto del País	Metropolitana	506
	Frontera con Brasil	388
	Costa Este	504
	Litoral Sur	321
	Centro y centro Sur	365
	Litoral Norte	393

Cuadro 6.3: Distribución por Región

6.1.1. Contrastes del modelo

6.1.1.1. Contraste de Wald

Significación de un parámetro en particular:

$H_0)\beta_k = 0 \; \forall k = 1, .., p$

$H_1)\beta_k \neq 0$

$W = \frac{\widehat{\beta}_k}{sd(\widehat{\beta}_k)}$

Se distribuye aproximadamente normal.

El nivel de significación de un test es un concepto estadístico asociado a la verificación de una hipótesis. Se define como la probabilidad de tomar la decisión de rechazar la hipótesis nula (H_0)) cuando esta es verdadera (decisión conocida como .Error de tipo I", o "falsos positivos"). La decisión se toma a menudo utilizando el p−valor: si el valor p es inferior a nivel de significación, entonces la hipótesis nula es rechazada (Blanco, 2006).

Una vez que se obtiene un modelo en donde tanto los parámetros como el modelo en su conjunto son significativos, se procede a elegir el punto de corte más apropiado y a comprobar cuán bueno fue el ajuste de los valores predichos por el modelo utilizando otras herramientas.

6.1.1.2. Contraste Condicional de Razón de Verosimilitudes

La razón de verosimilitud del modelo es una prueba para testear la significación del modelo. Se define $\lambda = \frac{L_R}{L_M}$ donde L_M es la verosimilitud del modelo completo $\frac{e^{\beta_0+\beta_1 x_1+...+\beta_k x_q+...+\beta_k x_p}}{1+e^{\beta_0+\beta_1 x_1+...+\beta_k x_q+...+\beta_k x_p}}$ y L_R la del modelo reducido, $\frac{e^{\beta_0+\beta_1 x_1+...+\beta_k x_q}}{1+e^{\beta_0+\beta_1 x_1+...+\beta_k x_q}}$, q < p.

$H_0)\beta_1 = \beta_2 = ... = \beta_p = 0$

$H_1) algun \beta_k \neq 0$ con $k = 1, .., p$

$-2ln(\lambda)$ se distribuye $\chi^2_{(p+1-q),\alpha}$ siendo (*p+1*) y q la cantidad de parámetros incluidos en el modelo completo y el modelo reducido respectivamente.

Test de razón de verosimilitud:

$$-2ln(\lambda) = -2ln\frac{L_R}{L_M} = -2(ln(L_R) - ln(L_M))$$

La hipótesis nula será rechazada para el nivel de significación α cuando $-2ln(\lambda) > \chi^2_{(p+1-q),\alpha}$. Esto es equivalente a que el p $-$ valor de contraste sea menor que el nivel de significación fijado (Blanco, 2006).

6.1.1.3. Stepwise

El procedimiento de selección stepwise, paso a paso, esta basado en estos contrastes condicionales de razón de verosimilitudes.

Una cuestión importante a tener en cuenta es el correcto manejo de las variables categóricas transformadas en varias variables ficticias. Siempre que se decida incluir (o excluir) una de estas variables, todas sus correspondientes variables ficticias deben ser incluidas (o excluidas) en bloque. No hacerlo así implicaría que se habría recodificado la variable, y por tanto la interpretación de la misma no seria igual.

Existen diferentes estrategias para la elección de variables a incluir en los modelos que se van a evaluar. A continuación se presenta en forma resumida las formas genéricas de selección a pasos.

Hacia atrás

Parte de un modelo con todas las variables y paso a paso va eliminando variables. La variable a ser eliminada es la que aporta menos a explicar la probabilidad de pertenecer al grupo J y cumple con los criterios de salida. La variable que menos aporta es aquella que su presencia no mejora la calidad del modelo, según los criterios de salida especificados.

El proceso se detiene cuando ya no hay más variables candidatas a salir.

Hacia delante

Parte de un modelo sin ninguna variables, solo con la constante y va ingresando variables y paso a paso. La primera variable candidata a entrar es la que aporta más a explicar la probabilidad de pertenecer al grupo J y cumple con los criterios de entrada. La siguiente candidata a entrar es la que aporta más a explicar a Y dado que el modelo contiene la variable ingresada en el primer paso. Si la variable que más aporta en este paso cumple con el criterio de entrada, la variable se incorpora al modelo. Una vez que la variable ingresa al modelo no se elimina.

El proceso se detiene cuando ya no hay más variables candidatas a entrar acorde a los criterios de entrada fijados.

Stepwise (paso a paso)

La selección 'stepwise', es un proceso de selección a pasos que combinan los dos criterios antes mencionados, comienza como un procedimiento hacia adelante pero pueden eliminarse variables i la significación de la misma cambia. Se deben definir criterios de entrada y de salida de variables. El procedimiento termina cuando se cumplen ambos criterios, no hay candidatas a entrar y tampoco a salir.

La selección "stepwise", o por pasos, es una versión modificada del proceso de regresión hacia adelante y hacia atrás, en la que en cada nuevo paso, cuando se incluye una nueva variable, además se reconsidera el mantener las que ya se habían añadido previamente, es decir que no solo puede entrar una nueva variable en cada paso sino que puede salir alguna de las que ya estaban en el modelo. El proceso finaliza cuando ninguna variable cumple la condición para entrar y, de las variables incluidas en la ecuación, ninguna cumple la condición para salir.

La forma que instrumentan estos procedimientos en el R es a través de la función *step*. Y el procedimiento de selección de modelos, tanto hacia atrás, adelante o stepwise se realiza con el indicador llamado Criterio de Información de Akaike (AIC),

$AIC = -2L_{Modelo} + k(\text{número de parámetros estimados})$

donde L_{Modelo} es el log verosimilitud del modelo ajustado en cada paso y k el número de parámetros estimados.

El criterio selecciona el modelo con el menor valor AIC.

6.1.2. Consideraciones para la construcción de un árbol

6.1.3. Conjunto de preguntas con respuesta binaria.

Para cada nodo t se tiene un conjunto de reglas de decisión s. Las particiones se seleccionan según si la variable interviniente es continua o categórica. Si la variable es cuantitativa las reglas son del tipo s: $X_i \leq m$ con $m \in$ R. Existen N posibles divisiones, que consisten en igualar m con cada uno de los valores observados de X_i. Si X_i es de tipo cualitativa con modalidades c_1, ..., c_L las preguntas serán del estilo: $X_i \in$ C , siendo C subconjunto de $C = c1, c2, ..., cl$. Con L categorías se definen 2^{L-1} -1 reglas para particionar el nodo t. Es decir que para el caso de que la variable interviniente en la decisión sea continua, se verifica si el valor de dicha variable es mayor que cierto valor específico. Si es mayor se sigue el camino de la derecha y si es menor el de la izquierda. Luego este procedimiento se vuelve a repetir en cada nodo, es decir se selecciona una variable y un punto de corte para dividir la muestra en dos partes más homogéneas.

6.1.4. Criterio de bondad de ajuste de la partición que evalúa en cada nodo t la bondad de la partición s.

En cada nodo se evalúa la bondad de la partición y se selecciona la mejor de ellas, para ello es necesario tener un criterio que permita evaluar que tan buena es la partición que se genera, este criterio está basado en la medida de impureza del nodo t (I(t)). La impureza de un nodo está asociado a la heterogeneidad presente en la variable dependiente en dicho nodo. Las formas de medir la impureza variará según el tipo de árbol con el que se trabaje, es decir si el mismo es de clasificación o de regresión. Para cada regla s se calcula la caída que se produce en la impureza al utilizar la regla s para dividir t, es decir se considera $\Theta(s,t) = I(t) - I(t_i) - I(t_d)$, donde t_i y t_d representan los nodos izquierdo y derecho resultantes de la partición del nodo t. La regla elegida es aquella que maximice $\Theta(s,t)$. A modo de ejemplo se presenta una de las formas de definir la impureza de un nodo cuando el árbol es de clasificación:

$I(t) = -\sum_{g=1}^{G} p(g|t) logp(g|t) \, donde \, p(g|t) = \frac{N_s(t)}{N(t)}$

En este caso $p(g|t)$ representa la probabilidad de que las observaciones que llegan al nodo t pertenezcan a cada una de las clases. Esta es máxima cuando $p(g|t) = \dfrac{1}{G}$. Por lo tanto la variable a ser seleccionada será la que minimice la heterogeneidad o impureza que resulta de la división del nodo.

6.1.5. Regla de detención.

Las reglas de detención son las que permiten declarar a un nodo como terminal y que el proceso de partición se detenga. Por lo general el analista adopta un criterio para decidir cuándo un nodo es lo suficientemente homogéneo y que el proceso se detenga. Si dicho criterio se verifica el nodo será un nodo terminal, en caso contrario será un nodo intermedio. Es habitual el utilizar como criterio de parada el grado de impureza del nodo, y establecer que el proceso se detenga cuando el decrecimiento en el nivel de impureza alcanza determinado umbral, establecido como crítico por el analista. También pueden adoptarse criterios relacionados con la cantidad de elementos o el grado de impureza del nodo.

6.1.6. Regla para asignar cada nodo terminal a una clase.

Para asignar cada nodo terminal a una clase nos fijaremos en la frecuencia de las observaciones contenidas en dicho nodo. Es decir que asignaremos todas las observaciones al grupo más probable en dicho nodo, lo que equivale a decir, el grupo con máxima $p(g|t)$.

6.2. Anexo Resultados

6.2.1. ACM datos con y sin pesos

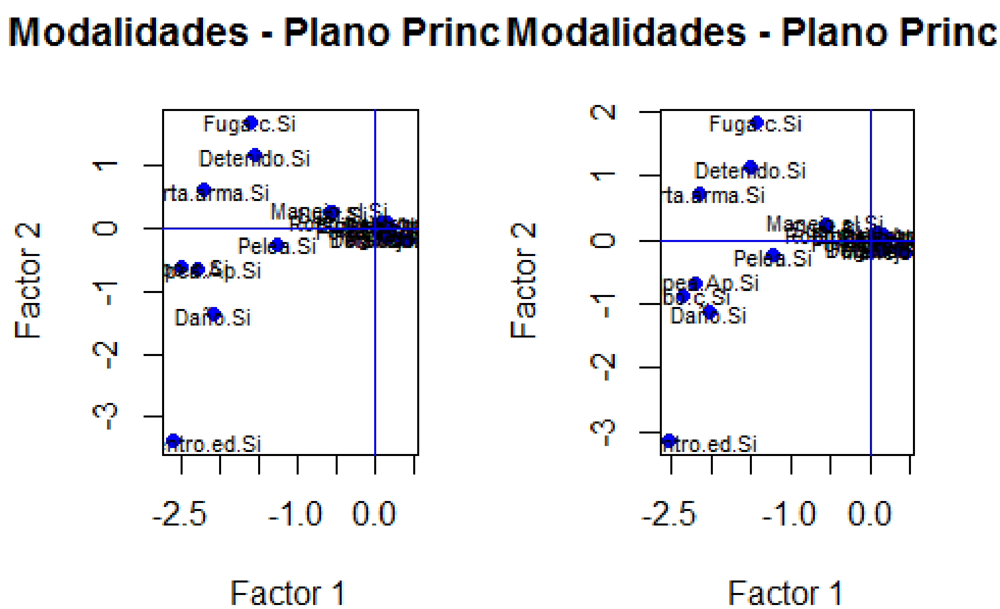

Figura 6.1: ACM distintas conductas con y sin pesos

6.2.2. Resultados gráficos de siluetas en el escenario 1

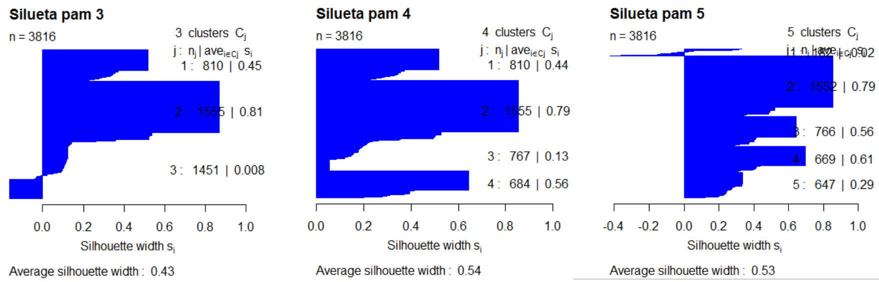

Figura 6.2: Silueta en PAM

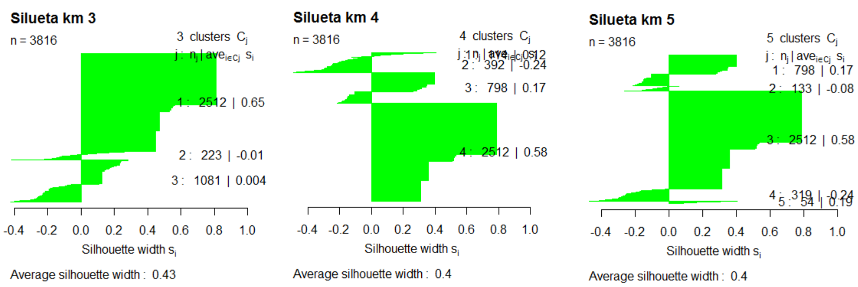

Figura 6.3: Silueta en K-MEANS

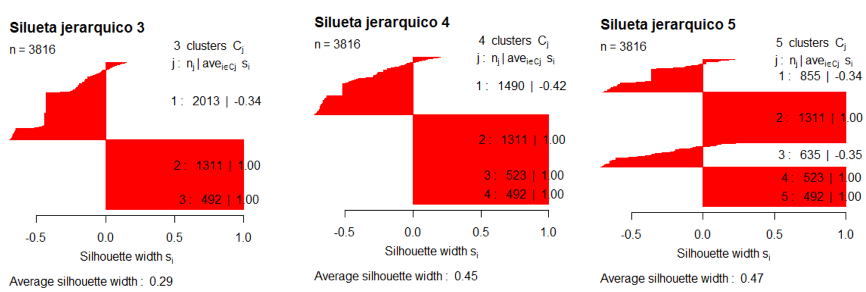

Figura 6.4: Silueta en JERÁRQUICO

6.2.3. Resultados figuras siluetas en el escenario 2

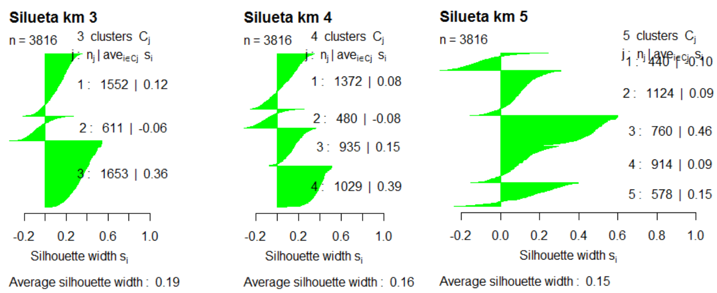

Figura 6.5: Gráficos de silueta en K-means

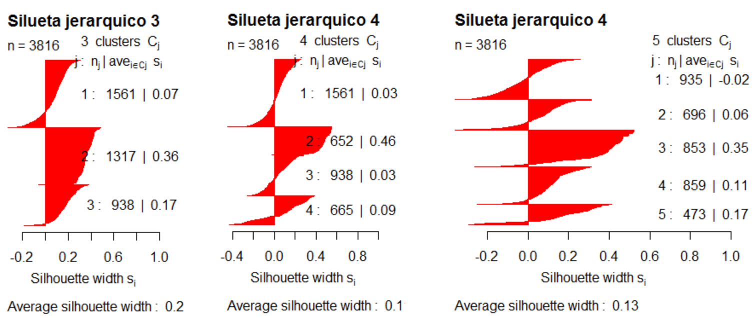

Figura 6.6: Gráficos de silueta en Jerárquico

6.2.4. Resultados con los distintos métodos de agrupación tratando la información que surge de los factores de ACM

Indicadores para determinar el número de grupos

A continuación se expone los gráficos de la información que aportan los indicadores R^2, pseudo F y pseudo t^2 obtenidos en base al método de Ward.

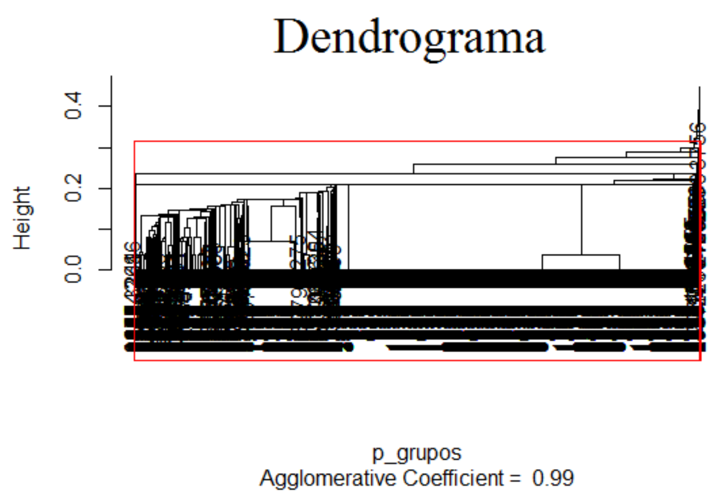

Figura 6.7: Dendrograma para el método del vecino más cercano

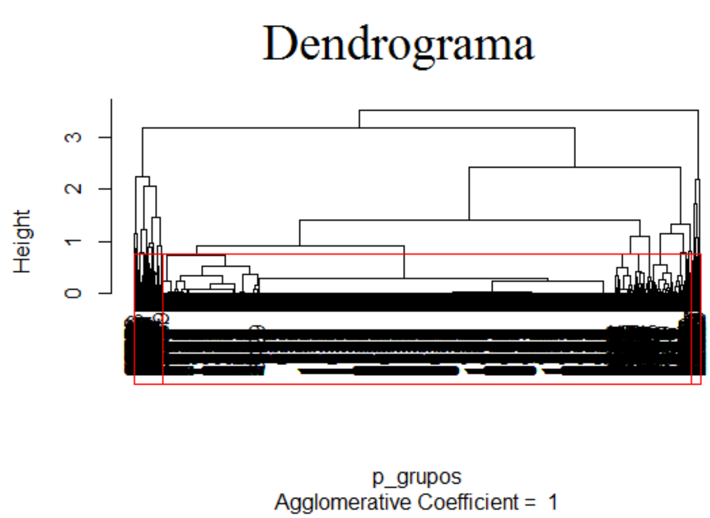

Figura 6.8: Dendrograma para el método del vecino más lejano

	.	history	Freq	Rcuad	psF	psT
3805	3789	3797	109	0.9191388	4325.096	128.6144
3806	3799	3804	1593	0.9040511	3984.550	1948.8307
3807	3791	3798	169	0.8878406	3766.972	311.4475
3808	3802	3793	270	0.8705260	3657.616	358.5017
3809	3796	3807	203	0.8383850	3293.226	215.6254
3810	3800	3805	382	0.8017701	3082.022	554.2386
3811	3542	3806	2904	0.7507701	2870.026	5400.5151
3812	3809	3803	260	0.6892528	2818.402	201.2935
3813	3812	3808	530	0.6082746	2960.429	250.5903
3814	3813	3810	912	0.4324618	2906.252	510.6264
3815	3814	3811	3816	0.0000000	NaN	2906.2522

Figura 6.9: Indicadores

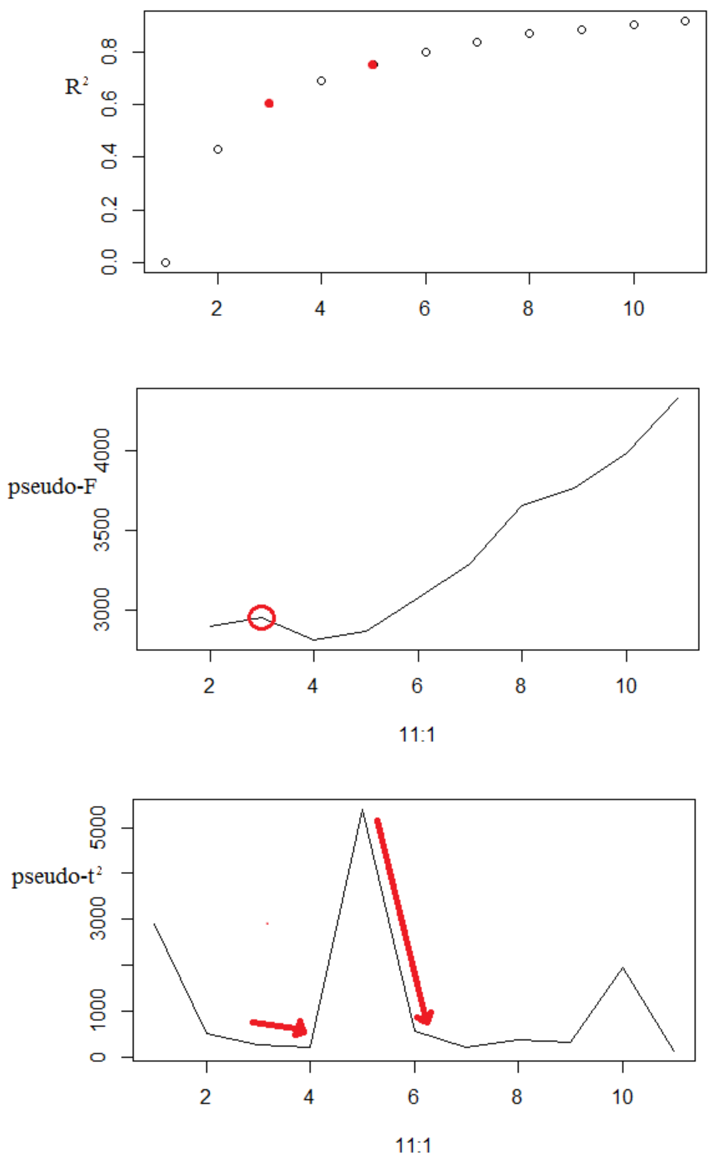

Figura 6.10: Indicadores R^2, pseudo F y pseudo t^2

6.2.5. Resultados cuadros ACM con variables originales

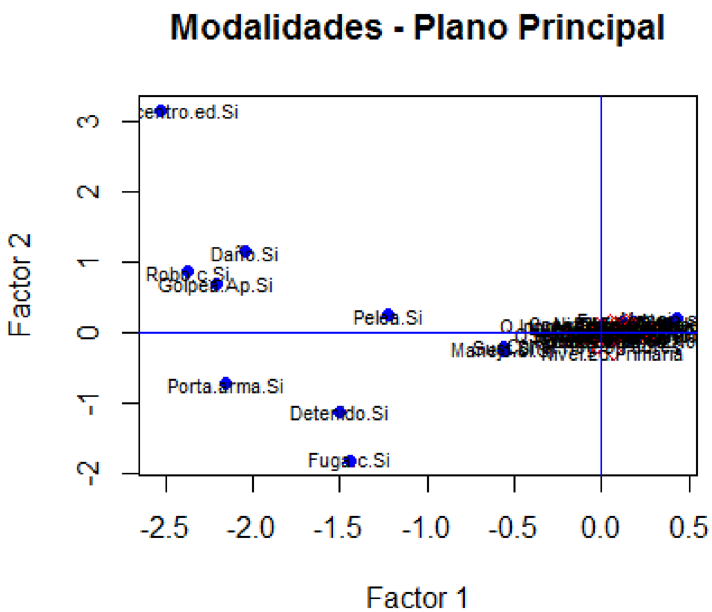

Figura 6.11: ACM caraterísticas personales - sup. conductas

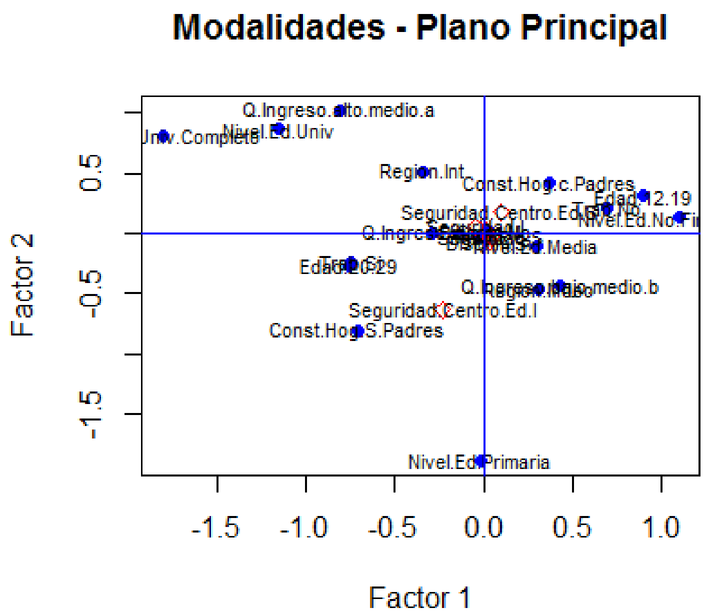

Figura 6.12: ACM caraterísticas personales - sup. pensamientos

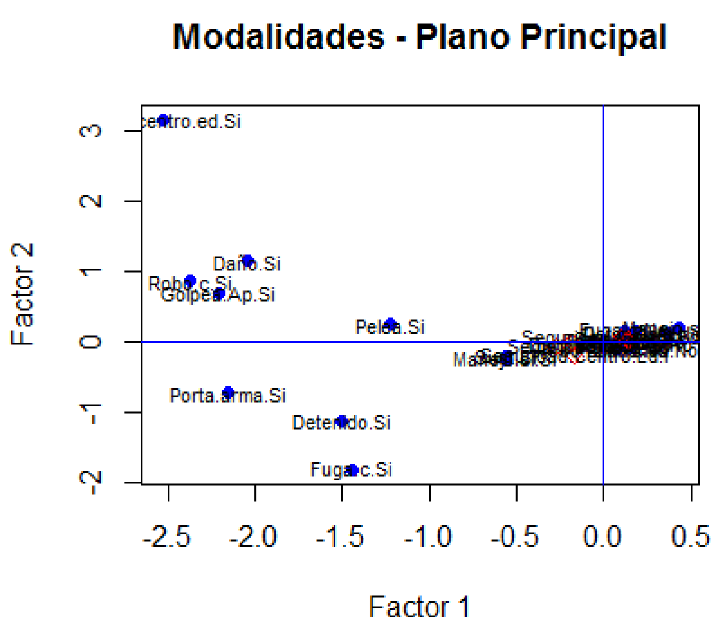

Figura 6.13: ACM conductas - sup. pensamientos

6.2.6. Caracterización de grupos en el escenario 1 con el procedimiento pam

	Detenido		**Fuga casa**		**Maneja sl**		**Porta arma**		**Sustancias**	
Grupo	No	Si	No	Si	No	Si	No	Si	No	Si
1	89.0	11.0	90.9	9.1	0	100	95.2	4.8	100	0
2	96.5	3.5	96.9	3.1	100	0	98.3	1.7	100	0
3	73.6	26.4	83.8	16.2	0	100	85.5	14.5	0	100
4	90.9	9.1	92.7	7.3	100	0	96.5	3.5	0	100

Cuadro 6.4: Indebidos 1 según grupos conducta

	Daño a p.		**Golpea a p.**		**Pelea**		**Robo c.**		**Robo c.Ed.**	
Grupo	No	Si	No	Si	No	Si	No	Si	No	Si
1	93.7	6.3	92.3	7.7	86.5	13.5	96.4	3.6	97.8	2.2
2	98.3	1.7	97.7	2.3	92.1	7.9	99.0	1.0	99.0	1.0
3	79.5	20.5	81.5	18.5	77.8	22.2	87.1	12.9	90.9	9.1
4	94.3	5.7	96.4	3.6	90.3	9.7	95.2	4.8	98.1	1.9

Cuadro 6.5: Indebidos 2 según grupos conducta

En base a los cuadros presentados se caracterizan los grupos:

El grupo 1 al igual que el grupo 3, son los grupos donde se concentran la mayoría de los jóvenes que cometieron infracciones. Una de las principales características de estos grupos es compartir que la totalidad de sus individuos en algún momento manejaron sin libreta. A su vez otro aspecto a destacar que discrimina estos grupos es que ningún joven del grupo 1 declara haber consumido sustancias ilegales (comparte esta característica con el grupo 2), en tanto el grupo 3 (al igual que el grupo 4) engloba en la totalidad de jóvenes que consumieron sustancias ilegales. Por su parte del resto de las infracciones que caracterizan al grupo 1, se basa en los registros de mayor frecuencia que son, haber manifestado una pelea (13.4 %) o haber estado detenido (10.9 %), en tanto los registros con menor frecuencia de este grupo son haber robado ya sea en un comercio o en centros educativos (3.5 % y 2.2 % respectivamente). Este grupo (grupo 1) contempla aproximadamente el mismo tamaño (21.2 %) que los grupos 3 y 4(21.0 % y 17.9 % respectivamente).

El grupo 2 en comparación con los restantes grupos, es el grupo que contiene a los jóvenes que declaran haber manifestado la menor cantidad de comportamientos indebidos, de los cuales entre su mayoría se resalta haber manifestado una pelea (7.9 %), los registros de los restantes comportamientos indebidos del grupo son irrelevantes, otro aspecto a tener en cuenta como se mencionó anteriormente, este grupo (grupo 2) engloba a los jóvenes que no manejaron sin libreta ni consumieron sustancias ilegales. Por estas razones podría considerarse como el grupo 'menos violento'. Cabe destacar que es el grupo de mayor magnitud ya que abarca el 40.7 % de la población total.

Por lo contrario como, el grupo 3 en comparación con los restantes grupos contiene la mayoría de jóvenes que declaran haber manifestado comportamientos indebidos, en este grupo resulta relevante la cantidad de jóvenes que declaran haber hecho daño apropósito (20.4 %), estado detenidos (26.3 %), fugado de la casa (16.1 %), golpead apropósito (18.4 %), portado arma (14.4 %), manifestado pelea (22.2 %), robado comercio y en centro educativo (12.9 % y 9.1 % respectivamente). Cabe destacar que en este grupo se engloba la totalidad de jóvenes que consumieron sustancias ilegales y manejaron sin libreta. Por estas razones podría considerarse como el grupo 'más violento' y exponen mayormente a los jóvenes estar ligados con la modalidad detenido, alcanzando una tasa de (26.3 %).

Los jóvenes clasificados en el grupo 4, al igual que los jóvenes del grupo 3 se caracterizan por haber consumido sustancias ilegales. Sin embargo estos jóvenes no declaran haber manejado sin libreta. En tanto otros aspectos a resaltar son haber estado detenido (9.0 %) haberse fugado de la casa (7.2 %) o haber manifestado una pelea (9.7 %), el resto de conductas registran frecuencias irrelevantes.

6.2.7. Bondad de ajuste del modelo en el escenario 1

Para evaluar la bondad de ajuste de un modelo logístico se utiliza como indicador el poder predictivo, allí se evalúa la performance del modelo, comparando lo observado con lo que es predicho por el modelo. Dicho poder predictivo se refleja a través de la siguiente tabla de clasificación:

6.2.8. Construcción del modelo

Start: AIC = 1337388 pam3 1

```
> (t=prop.table(table(obs,pre),m=1)*100)
   pre
obs         1         2         3         4
  1 12.962963 59.259259 22.222222  5.555556
  2  9.324759 76.527331  9.646302  4.501608
  3  9.803922 43.137255 39.869281  7.189542
  4  2.189781 53.284672 24.817518 19.708029
> sum(diag(t))/4
[1] 37.2669
```

Figura 6.14: Bondad de ajuste en el escenario 2

Step: AIC = 1068660 pam3 Edad

Step: AIC=871972.7 pam3 Edad + Q.Ingreso

Step: AIC=763769.2 pam3 Edad + Q.Ingreso + Trab

Step: AIC=685114.8 pam3 Edad + Q.Ingreso + Trab + Region

Step: AIC=591706 pam3 Edad + Q.Ingreso + Trab + Region + Const.Hog

Step: AIC=553024.5 pam3 Edad + Q.Ingreso + Trab + Region + Const.Hog + Sexo

Step: AIC=512269.7 pam3 Edad + Q.Ingreso + Trab + Region + Const.Hog + Sexo + Nivel.Ed

Step: AIC=502883.2 pam3 Edad + Q.Ingreso + Trab + Region + Const.Hog + Sexo + Nivel.Ed + Discrim

Step: AIC=500115.5 pam3 Edad + Q.Ingreso + Trab + Region + Const.Hog + Sexo + Nivel.Ed + Discrim + Seguridad

Step: AIC=499791.4 pam3 Edad + Q.Ingreso + Trab + Region + Const.Hog + Sexo + Nivel.Ed + Discrim + Seguridad + Seguridad.Centro.Ed

6.2.8.1. Intervalos de confianza de los odds ratio del grupo 1 y 3 ante el grupo 2

Grupo 1/grupo 2

	2.5 %	97.5 %
(Intercept)	0.006	0.007
Edad20-29	0.413	0.439
Q.Ingresobajo-medio.b	14.601	15.496
Q.Ingresomedio	7.965	8.474
TrabSi	1.250	1.324
Const.HogS.Padres	23.038	24.360
RegionMdeo	28.092	29.649
SexoF	6.925	7.294
Nivel.EdSec.completa	0.669	0.744
Nivel.EdSecundaria	1.780	1.939
Nivel.EdTerciaria	0.140	0.155
DiscrimNo	1.603	1.679
SeguridadI	0.839	0.880
Seguridad.Centro.EdI	1.139	1.197

Grupo 3/ grupo 2

	2.5 %	97.5 %
(Intercept)	0.370	0.411
Edad20-29	0.080	0.085
Q.Ingresobajo-medio.b	10.609	11.242
Q.Ingresomedio	3.766	4.018
TrabSi	0.051	0.054
Const.HogS.Padres	1.041	1.112
RegionMdeo	16.918	17.869
SexoF	15.080	15.943
Nivel.EdSec.completa	0.532	0.594
Nivel.EdSecundaria	1.091	1.187
Nivel.EdTerciaria	0.112	0.124
DiscrimNo	3.154	3.311
SeguridadI	0.473	0.497
Seguridad.Centro.EdI	0.948	1.005

Figura 6.15: Intervalos de confianza del modelo en escenario 1

6.2.8.2. Contraste de significación sobre los parámetros

```
> z <-summary(modelo.step)$coefficients/summary(modelo.step)$standard.errors
> p_valor <- (1 - pnorm(abs(z), 0, 1))*2
>
> p_valor
  (Intercept) Edad20-29 Q.Ingresobajo-medio.b Q.Ingresomedio      TrabSi Const.HogS.Padres RegionMdeo SexoF
1           0         0                     0              0 0.002015801      0.000000e+00          0     0
3           0         0                     0              0 0.000000000      1.072238e-05          0     0
  Nivel.EdSec.completa Nivel.EdSecundaria Nivel.EdTerciaria DiscrimNo SeguridadI Seguridad.Centro.EdI
1                    0       0.000000e+00                 0         0          0                    0
3                    0       4.993117e-11                 0         0          0                    0
```

Figura 6.16: Contraste sobre los parámetros

Tal y como presenta la figura, no existe ningún coeficiente que tenga asociado un p-valor mayor que 0.05, Por lo que a un nivel de significación del 5 % todos los coeficientes son significativamente distintos de cero.

6.2.9. Árboles de clasificación variando el parámetro cp

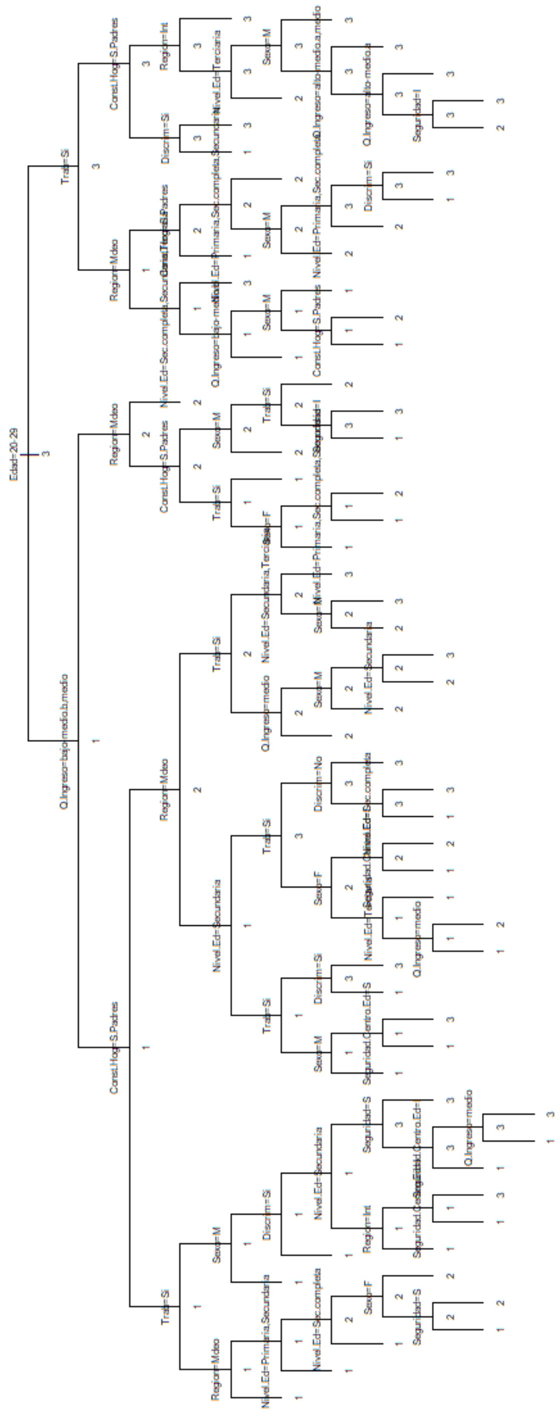

Figura 6.17: Árbol Máx

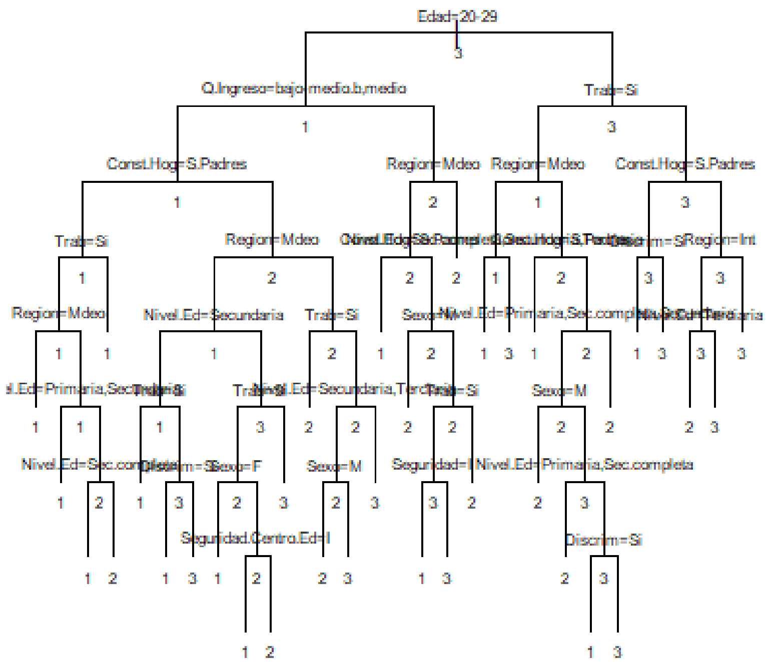

Figura 6.18: Árbol cp = 0.001

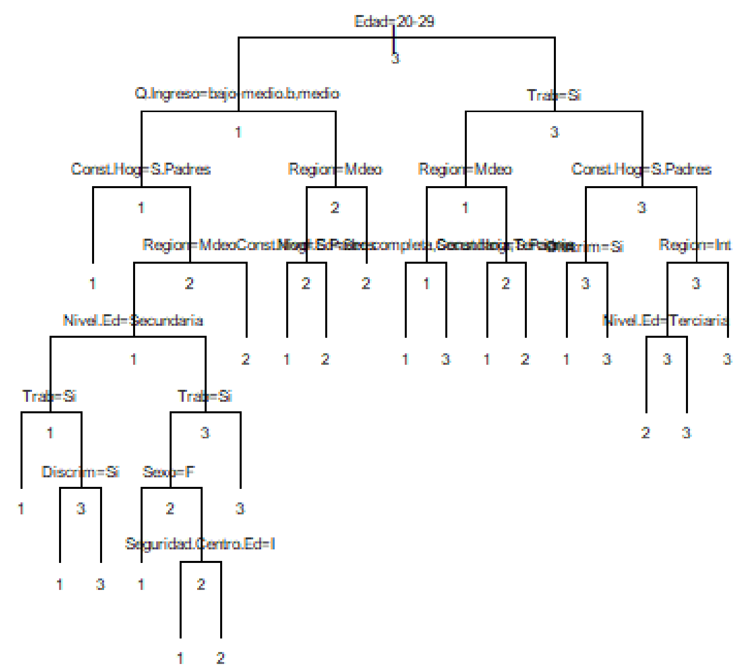

Figura 6.19: Árbol cp = 0.002

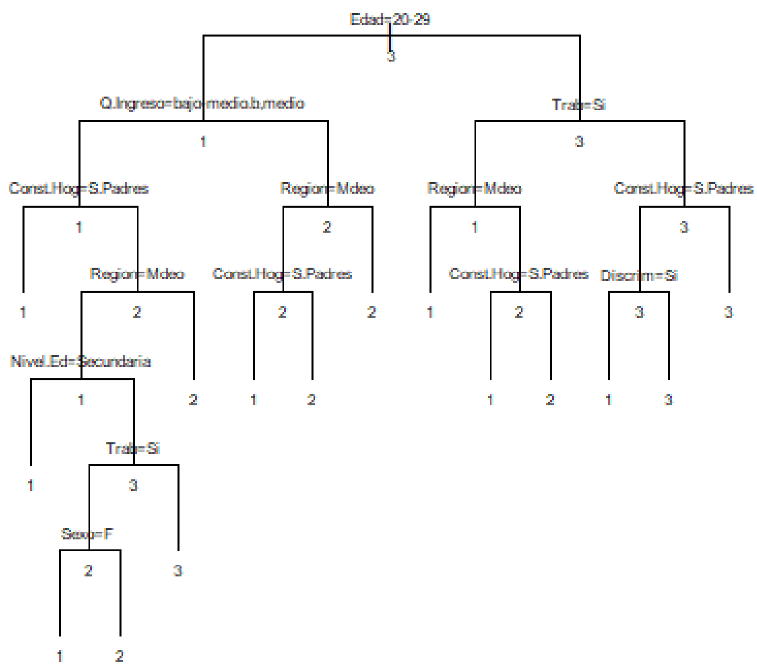

Figura 6.20: Árbol cp = 0.003

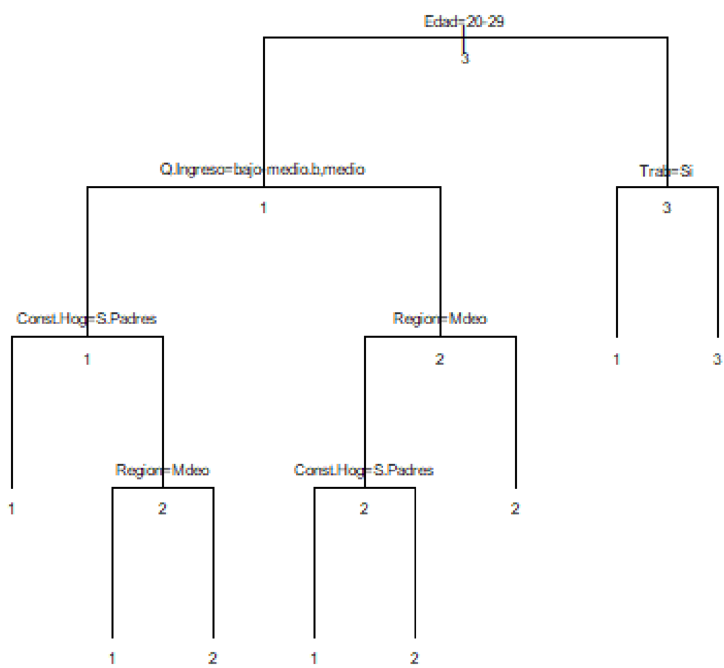

Figura 6.21: Árbol cp = 0.01

6.2.10. Distintas muestras con cp = 0.004

Resumen de distintas muestras:

Muestras	Cantidad Nodos	Variables que conforman las reglas de clasificación	Primer variable
1	13	Edad, Ingreso, Trabajo, Const. Hogar, Region, Discrimin.	Edad
2	13	Edad, Ingreso, Trabajo, Const. Hogar, Region, Discrimin.	Edad
3	13	Edad, Ingreso, Trabajo, Const. Hogar, Region, Sexo, Educación.	Edad
4	11	Edad, Ingreso, Trabajo, Const. Hogar, Region,	Edad
5	12	Edad, Ingreso, Trabajo, Const. Hogar, Region,	Edad
6	12	Trabajo, Region, Edad, Const.Hogar, Ingreso, Sexo, Educación.	Trabajo
7	13	Trabajo, Region, Edad, Const.Hogar, Ingreso, Educación.	Trabajo
8	12	Trabajo, Region, Edad, Const.Hogar, Ingreso, Discrim.	Trabajo
9	13	Trabajo, Region, Edad, Const.Hogar, Ingreso, Educación, Discrim.	Trabajo
10	12	Edad, Ingreso, Trabajo, Comp.. Hogar, Region,	Edad
11	12	Edad, Ingreso, Trabajo, Comp.. Hogar, Region,	Edad
12	12	Edad, Ingreso, Trabajo, Comp.. Hogar, Region,	Edad
13	13	Edad, Ingreso, Trabajo, Comp.. Hogar, Region,	Edad
14	11	Trabajo, Region, Edad, Const.Hogar, Ingreso, Educación.	Trabajo
15	12	Edad, Ingreso, Trabajo, Comp.. Hogar, Region, Educacion	Edad

Cuadro 6.6: Distintas muestras con cp=0.004

MUESTRA 1

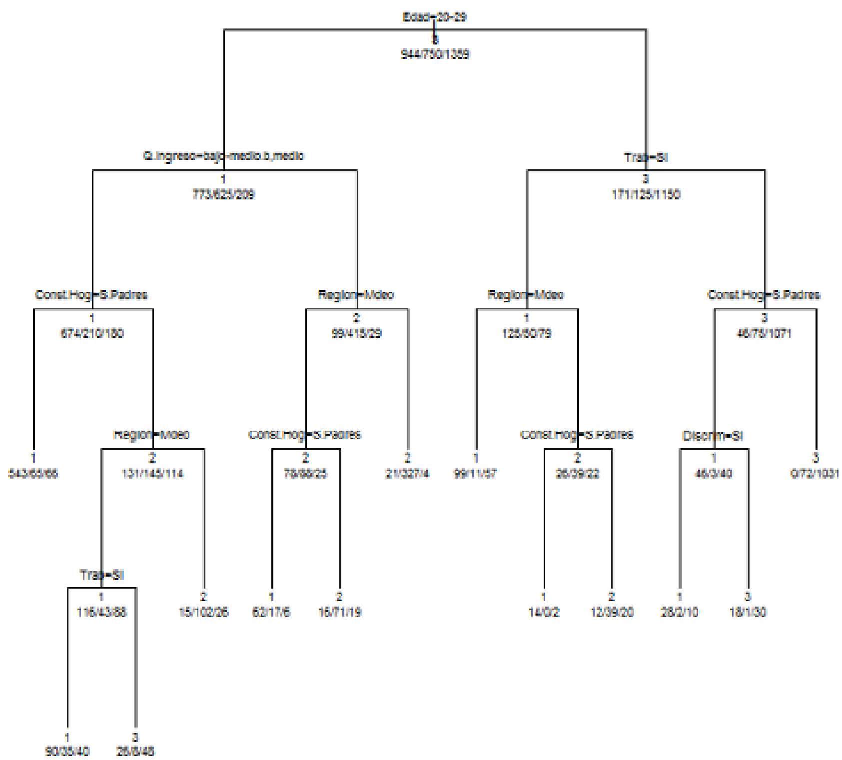

Figura 6.22: Muestra 1 cp = 0.004

MUESTRA 2

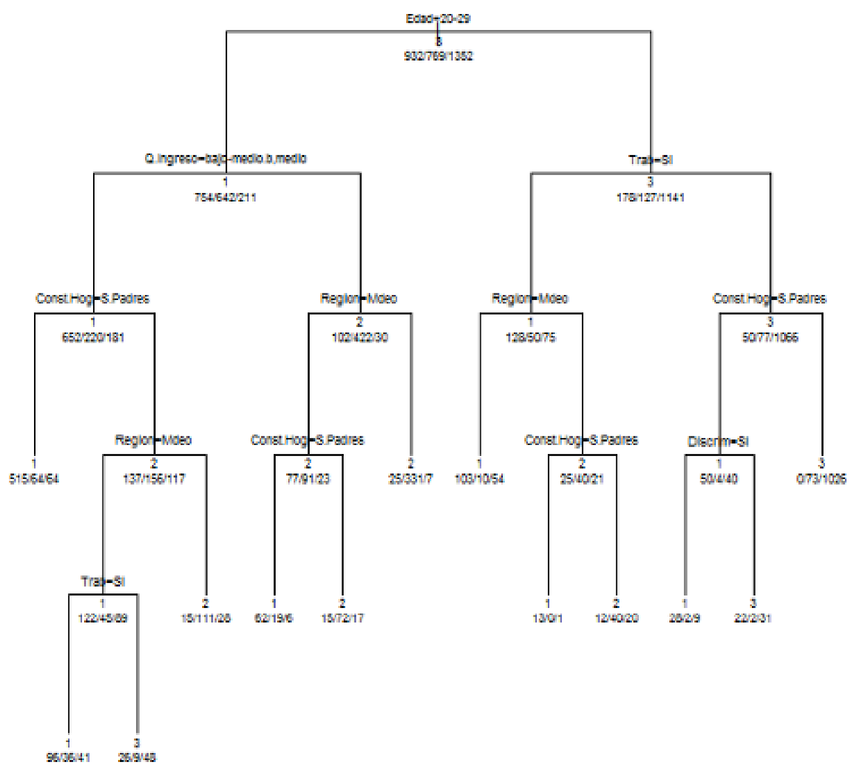

Figura 6.23: Muestra 2 cp = 0.004

MUESTRA 3

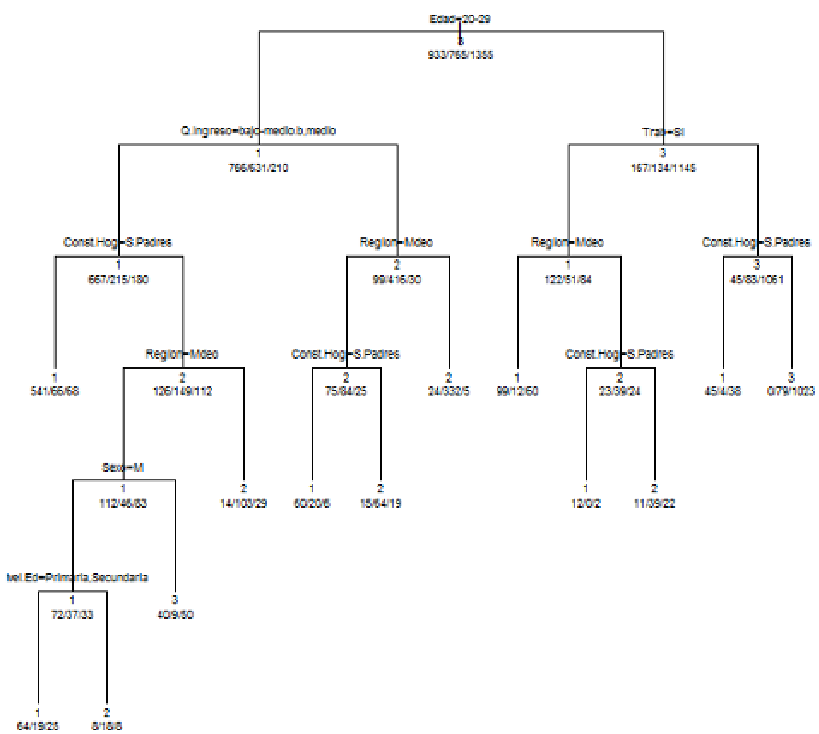

Figura 6.24: Muestra 3 cp = 0.004

MUESTRA 4

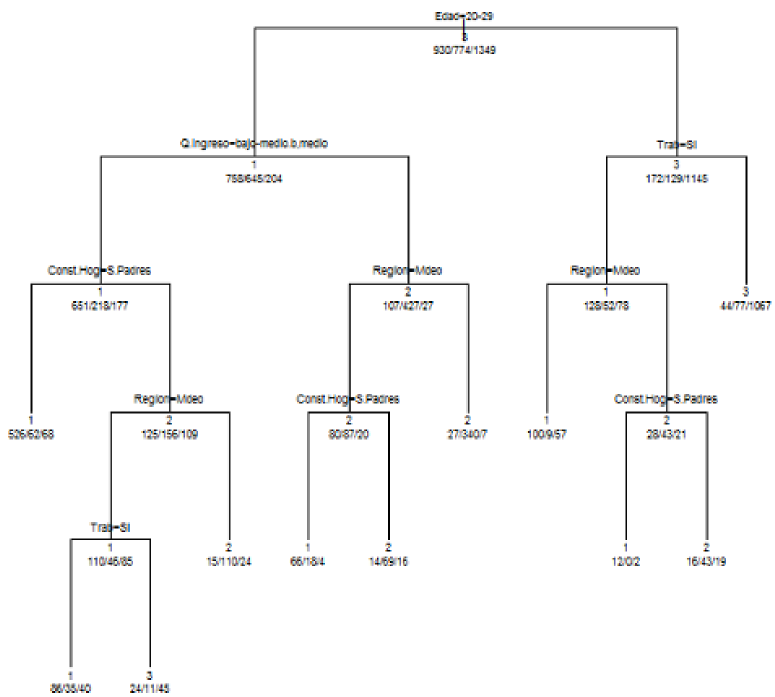

Figura 6.25: Muestra 4 cp = 0.004

MUESTRA 5

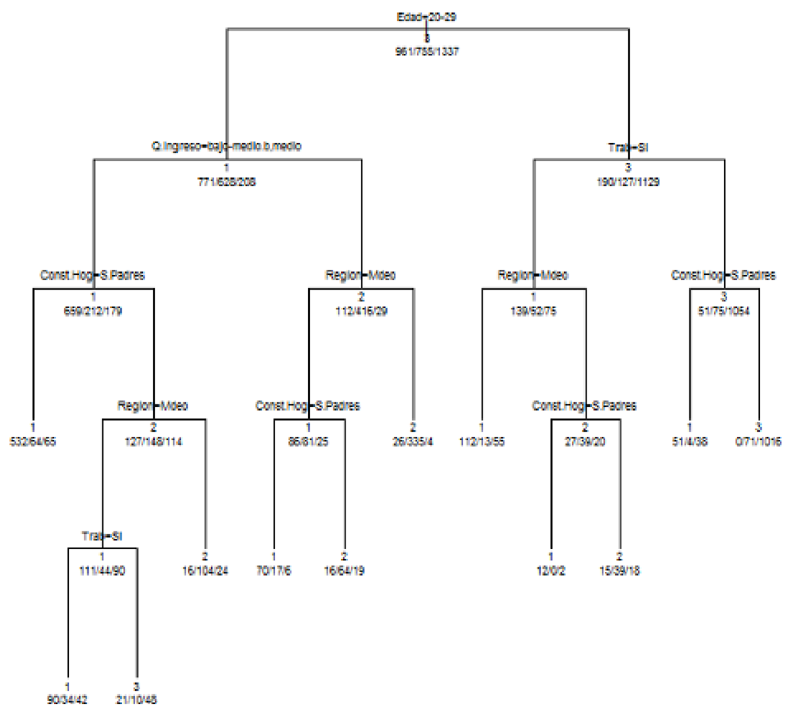

Figura 6.26: Muestra 5 cp = 0.004

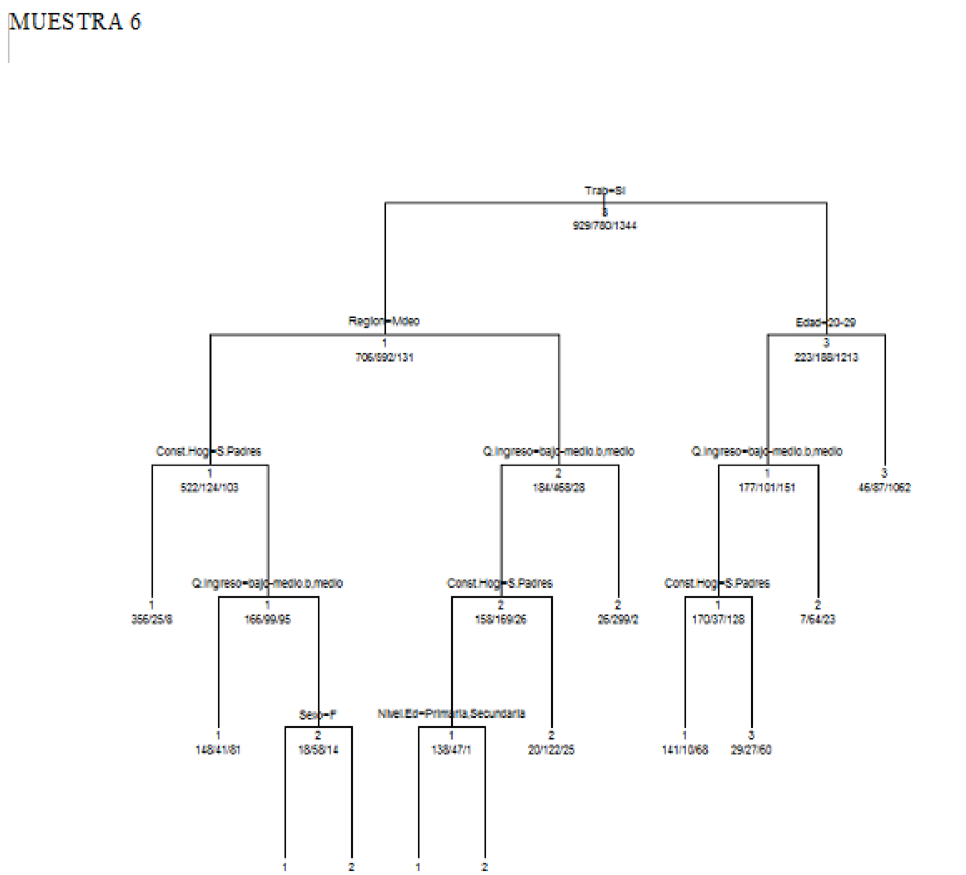

Figura 6.27: Muestra 6 cp = 0.004

MUESTRA 7

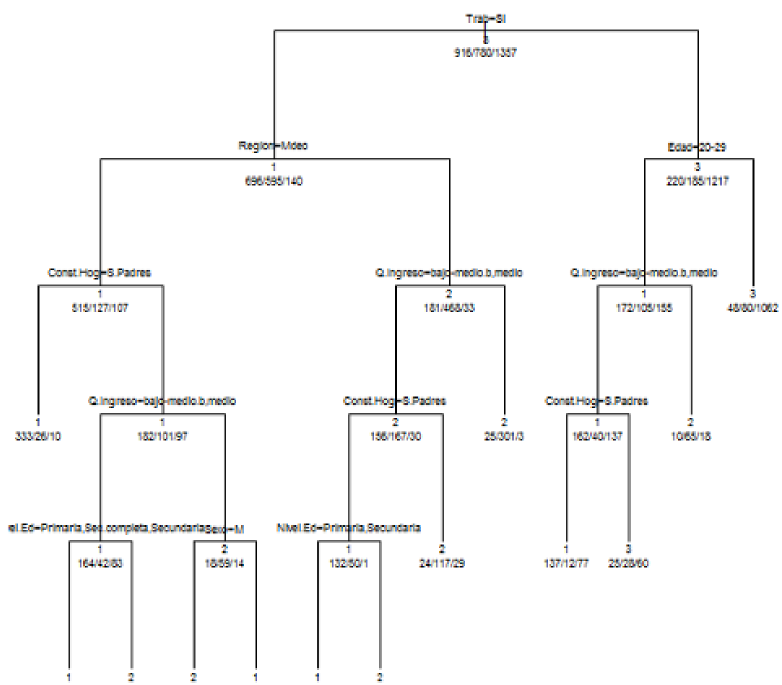

Figura 6.28: Muestra 7 cp = 0.004

MUESTRA 8

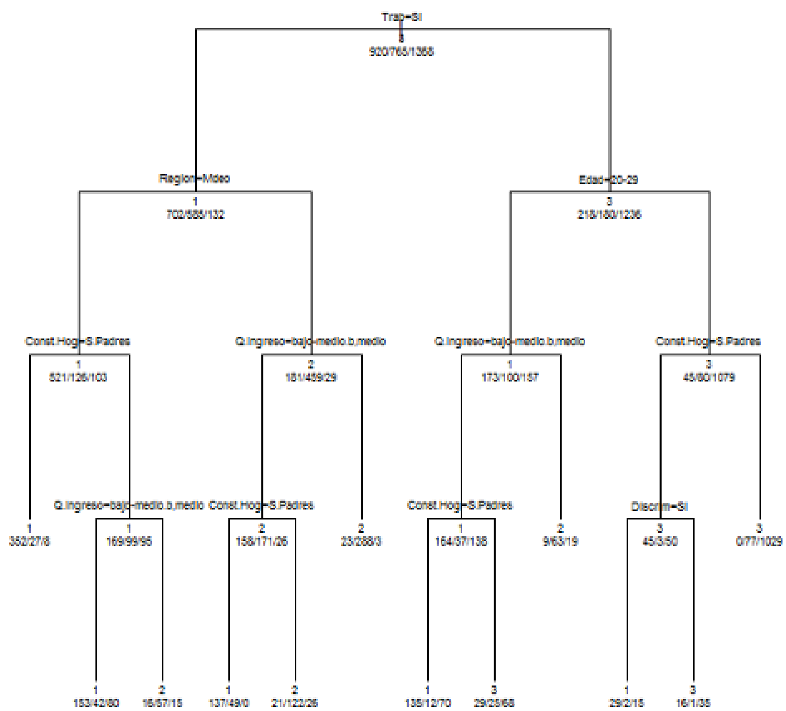

Figura 6.29: Muestra 8 cp = 0.004

MUESTRA 9

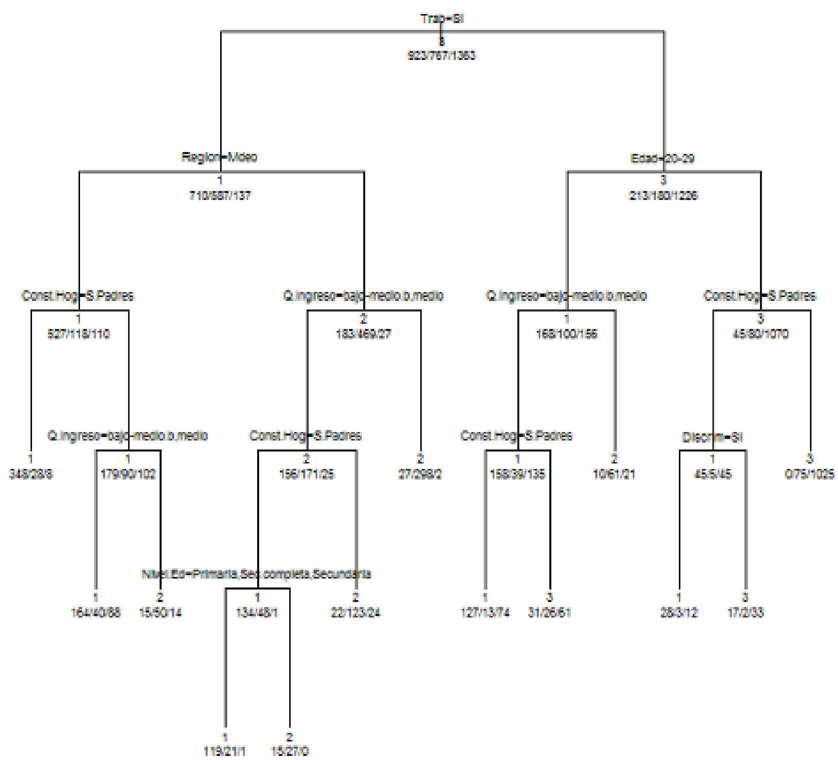

Figura 6.30: Muestra 9 cp = 0.004

MUESTRA 10

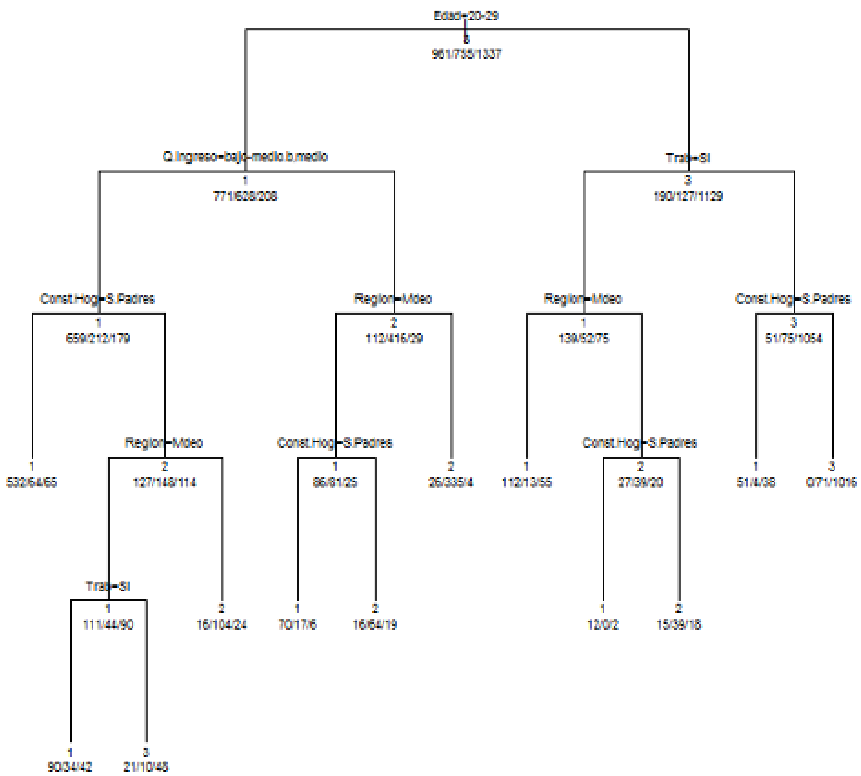

Figura 6.31: Muestra 10 cp = 0.004

MUESTRA 11

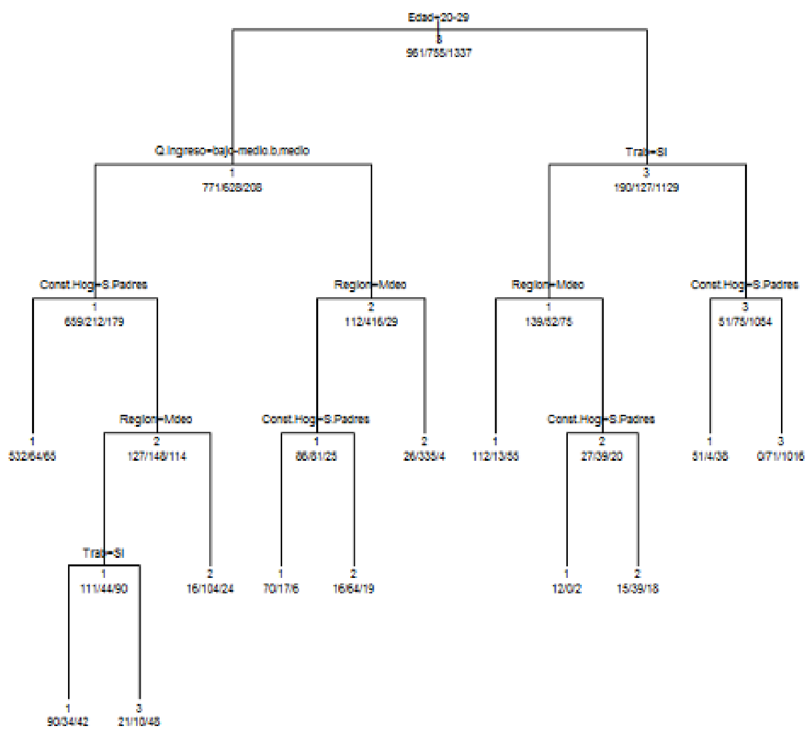

Figura 6.32: Muestra 11 cp = 0.004

MUESTRA 12

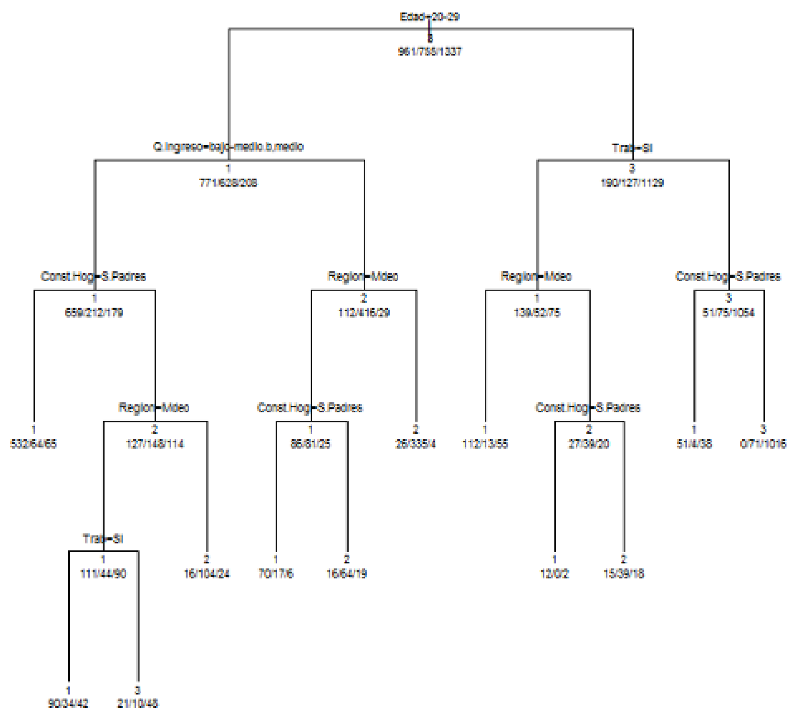

Figura 6.33: Muestra 12 cp = 0.004

MUESTRA 13

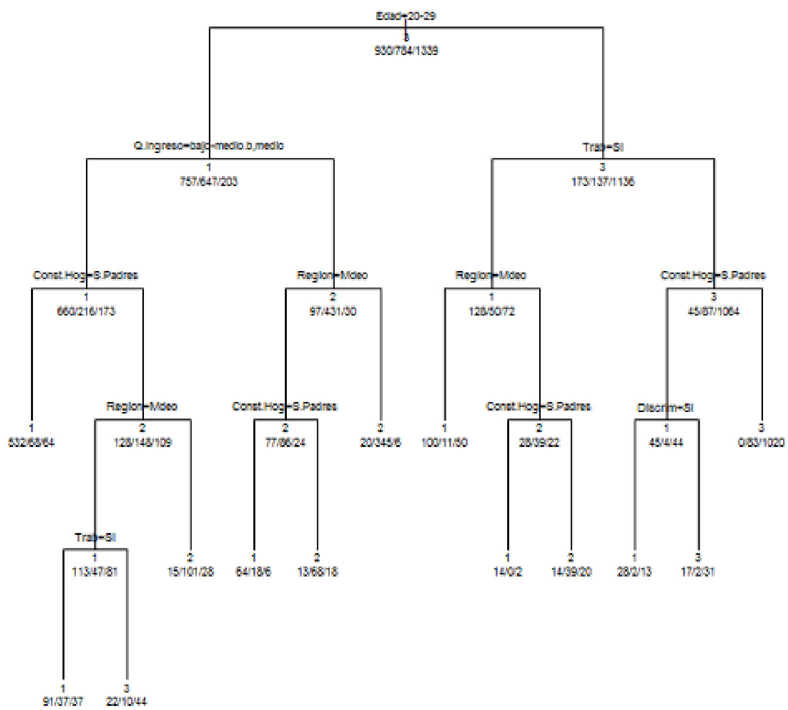

Figura 6.34: Muestra 13 cp = 0.004

MUESTRA 14

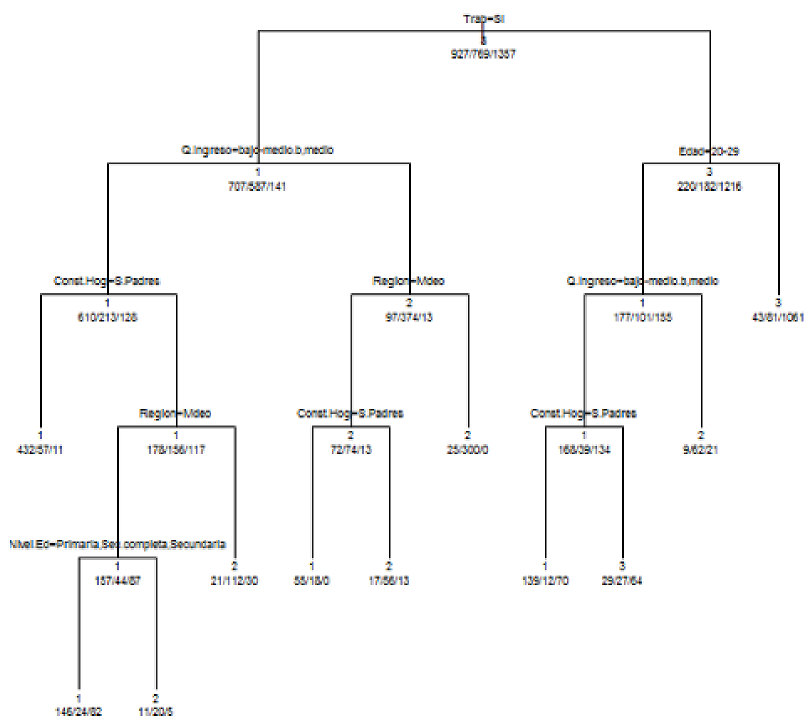

Figura 6.35: Muestra 14 cp = 0.004

MUESTRA 15

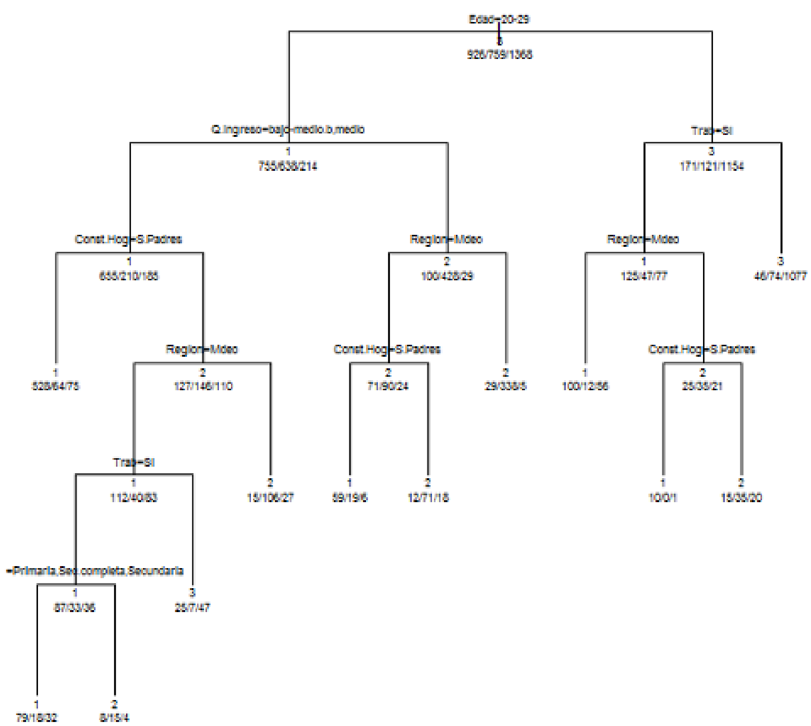

Figura 6.36: Muestra 15 cp = 0.004

6.3. Anexo Scripts utilizados

```
rm(list=ls(all=TRUE))
a = read.csv('base_ENAJ.csv',header = T,sep = ';')
B = a
#CONST. LA BASE
#Sexo
B$SexoEntrevistado = as.numeric(B$SexoEntrevistado)
B$Sexo = ifelse(a$SexoEntrevistado == 1,c(0),c(1))
B$Sexo = as.factor(B$Sexo)
levels(B$Sexo) = c('M','F')

#Edad
B$Edad = as.numeric(B$Edad)
B$Edad = ifelse(a$grupos_edad == 1 | a$grupos_edad == 2,c(0),c(1))
B$Edad = as.factor(B$Edad)
levels(B$Edad) = c("12-19","20-29")

#Region
B$region_enaj = as.numeric(B$region_enaj)
B$Region = ifelse(B$region_enaj == 1,c(0),c(1))
B$Region = as.factor(B$Region)
levels(B$Region) = c('Int','Mdeo')

#Quintil Ingreso
B$Quintiles_ingreso = as.numeric(B$Quintiles_ingreso)
B$Q.Ingreso = ifelse(B$Quintiles_ingreso == 1 |B$Quintiles_ingreso == 2 ,
                     'bajo-medio.b',ifelse(B$Quintiles_ingreso == 3,
                                           'medio',"alto-medio.a"))
B$Q.Ingreso = as.factor(B$Q.Ingreso)

#Constitucion Hogar
B$Const.Hog=ifelse (((a$A_1_B == 1 & a$A_1_C == 1)) &  a$A_1_K == 0 & a$A_1_H == 0 |
                      ((a$A_1_B == 1) & (a$A_1_C == 0)) &  a$A_1_K == 0 &
                      a$A_1_H == 0 | ((a$A_1_B == 0) & (a$A_1_C == 1)) &
                      a$A_1_K == 0 & a$A_1_H == 0,c(0),c(1))
B$Const.Hog = as.factor(B$Const.Hog)
levels(B$Const.Hog) = c("C.Padres","S.Padres")
#Nivel Ed.
#todos los juvenes asistieron a Ed.Primaria: table(a$B_4)
B$Nivel.Ed = ifelse ( a$B_8 == 2 | a$B_12 == 2 ,'Primaria',
             ifelse ( a$B_12 == 1 &  (a$B_23 == 2 |  a$B_23 == 3),'Secundaria',
             ifelse ( a$B_12 == 1 & a$B_23 == 1 & a$B_38 == 2 ,'Sec.completa',
             ifelse (( a$B_23 == 1 & a$B_38 == 1 & a$B_41 == 2) |
                       (a$B_38 == 1 & a$B_23 == 1 & a$B_41 != 1)| a$B_41 == 1 ,'Terciaria',"NC"))))
B$Nivel.Ed = as.factor(B$Nivel.Ed)

#Trabajo
B$Trab = ifelse(a$E_4 == 1,c(1),c(0))
B$Trab = as.factor(B$Trab)
levels(B$Trab) = c("No","Si")

#Discrim
B$Discrim = ifelse(a$K_5 == 1,c(0),c(1))
B$Discrim = as.factor(B$Discrim)
levels(B$Discrim) = c("Si","No")


# OPINIONES DE LA JUVENTUD
B$Seguridad = ifelse(a$F_19_4 == 3 | a$F_19_4 == 4 | a$F_19_4 == 5 ,c(0),c(1))
B$Seguridad = as.factor(B$Seguridad)
levels(B$Seguridad) = c("S","I")

B$Seguridad.Centro.Ed = ifelse(a$F_19_1 == 3 |
                                 a$F_19_1 == 4 | a$F_19_1 == 5 ,c(0),c(1))
B$Seguridad.Centro.Ed = as.factor(B$Seguridad.Centro.Ed)
levels(B$Seguridad.Centro.Ed) = c("S","I")


B=B[,c('Sexo','Edad','Nivel.Ed','Region','Trab','Const.Hog',
       'Q.Ingreso','Discrim','Seguridad','Seguridad.Centro.Ed','peso_enaj')]

poblacion = na.omit(B$peso_enaj)
(pob = sum(poblacion))# La población total es de 769.497 jovenes
dim(B)
```

```
### DEF. VARIABLES DE COMPORTAMIENTOS

B$Detenido = ifelse(a$K_3 == 1,c(1),c(0))
B$Detenido=as.factor(B$Detenido)
levels(B$Detenido)=c("No","Si")

B$Daño = ifelse(a$K_1_8 == 1,c(1),c(0))
B$Daño=as.factor(B$Daño)
levels(B$Daño)=c("No","Si")

B$Robo.centro.ed = ifelse(a$K_1_1 == 1,c(1),c(0))
B$Robo.centro.ed=as.factor(B$Robo.centro.ed)
levels(B$Robo.centro.ed)=c("No","Si")

B$Pelea = ifelse(a$K_11 != 0,c(1),c(0))
B$Pelea=as.factor(B$Pelea)
levels(B$Pelea)=c("No","Si")

B$Fuga.c = ifelse(a$K_1_2 == 1,c(1),c(0))
B$Fuga.c = as.factor(B$Fuga.c)
levels(B$Fuga.c)=c("No","Si")

B$Robo.c = ifelse(a$K_1_4 == 1,c(1),c(0))
B$Robo.c = as.factor(B$Robo.c)
levels(B$Robo.c) = c("No","Si")

B$Manejo.sl = ifelse(a$K_1_6 == 1,c(1),c(0))
B$Manejo.sl = as.factor(B$Manejo.sl)
levels(B$Manejo.sl) = c("No","Si")

B$Porta.arma = ifelse(a$K_1_9 == 1,c(1),c(0))
B$Porta.arma = as.factor(B$Porta.arma)
levels(B$Porta.arma) = c("No","Si")

B$Golpea.Ap = ifelse(a$K_1_13 == 1,c(1),c(0))
B$Golpea.Ap = as.factor(B$Golpea.Ap)
levels(B$Golpea.Ap) = c("No","Si")

B$Sust = ifelse( a$I_4_2 == 1 | a$I_4_3 == 1 | a$I_4_4 == 1 |
                    a$I_4_5 == 1 | a$I_4_6 == 1,c(1),c(0))
B$Sust = as.factor(B$Sust)
levels(B$Sust) = c("No","Si")

table(is.na(B$peso_enaj))#tiene 1 Na
(Obs.NA = which(is.na(B$peso_enaj)==TRUE))
#B[Obs.NA,] lo saco, tiene muchos NA
B = B[-Obs.NA,];
dim(B)

# Descriptivo indebidos
prop.table(xtabs(peso_enaj~B$Detenido,data=B))*100
prop.table(xtabs(peso_enaj~B$Daño,data=B))*100
prop.table(xtabs(peso_enaj~B$Robo.centro.ed,data=B))*100
prop.table(xtabs(peso_enaj~B$Pelea,data=B))*100
prop.table(xtabs(peso_enaj~B$Fuga.c,data=B))*100
prop.table(xtabs(peso_enaj~B$Robo.c,data=B))*100
prop.table(xtabs(peso_enaj~B$Manejo.sl,data=B))*100
prop.table(xtabs(peso_enaj~B$Porta.arma,data=B))*100
prop.table(xtabs(peso_enaj~B$Golpea.Ap,data=B))*100
prop.table(xtabs(peso_enaj~B$Sust,data=B))*100

#### ACM ####
source('acm.R') #par(mfrow=c(1,2))
source("acmPOND.R")#TRABAJO CON LOS PESOS
B$peso_enaj = as.integer(B$peso_enaj)

#Comportamientos Indebidos
acm_1 = acm(B[,c(12:21)],NF = 4,ByG = T,pesos = B$peso_enaj)
abline(h=0,v=0,col='blue')

#Caracteristicas personales
acm_2 = acm(B[,c(1:7)],NF = 4, ByG = T,pesos = B$peso_enaj)#caracteristias personales
abline(h=0,v=0,col='blue')

#Opinion y/o pensamientos
acm_3 = acm(B[,c(8:10)],NF = 2,ByG= T ,pesos = B$peso_enaj)
abline(h=0,v=0,col='blue')

######### ACM con Sup. ##################
#C.Personales - Sup. Actos Indebido
sup.acm.4 = B[,c( "Detenido","Pelea","Robo.centro.ed","Daño","Fuga.c","Robo.c"
                 ,"Manejo.sl","Porta.arma","Golpea.Ap","Sust","Sexo"
                 ,"Edad","Nivel.Ed","Region","Trab","Const.Hog","Q.Ingreso")]
acm_4 = acm(sup.acm.4, Csup = (11:17),NF= 4 , ByG = T ,pesos = B$peso_enaj)
abline(h=0,v=0,col='blue')

s = B[,c("Edad","Nivel.Ed","Region","Trab","Const.Hog","Q.Ingreso","Indebido")]
acms_4_ss =acm(s,NF = 4,ByG= T ,pesos = B$peso_enaj)
abline(h=0,v=0,col='blue')
```

6.3. ANEXO SCRIPTS UTILIZADOS

```
#C.Personales - Sup.Opinion
sup.acm.5 = B[,c("Sexo","Edad","Nivel.Ed","Region","Trab","Const.Hog","Q.Ingreso",
                 "Discrim","Seguridad","Seguridad.Centro.Ed")]
acm_5 = acm(sup.acm.5, Csup = (8:10),NF= 4, ByG = T,pesos = B$peso_enaj)
abline(h=0,v=0,col='blue')

s = B[,c("Sexo","Edad","Nivel.Ed","Region","Trab","Const.Hog","Q.Ingreso","Discrim",
         "Seguridad","Seguridad.Centro.Ed")]
acms_5_ss =acm(s,NF = 4,ByG= T ,pesos = B$peso_enaj)
abline(h=0,v=0,col='blue')

#Actos Indebido - Sup Opinion
sup.acm.6 = B[,c("Discrim","Seguridad","Seguridad.Centro.Ed","Detenido","Fuga.c",
                 "Robo.c","Daño","Robo.centro.ed","Pelea","Manejo.sl","Porta.arma","Golpea.Ap","Sust")]

#Actos Indebido - Sup Opinion
acm_6 = acm(sup.acm.6, Csup = c(1:3),NF= 4, ByG = T,pesos = B$peso_enaj)
abline(h=0,v=0,col='blue')

acm_6_ss = acm(B[,c(8:10,22)],NF = 4,ByG= T ,pesos = B$peso_enaj)
abline(h=0,v=0,col='blue')


#GRUPOS COMPORTAMIENTOS INDEBIDOS
library(cluster)
#C=B[,12:21] #Si considero únicamente variables de conducta, cambio B por C y considero 4 grupos

D = daisy(B)# se puede metric gower es un indice de similitud
F = as.matrix(D)

pam3 = pam(F,3,diss = TRUE)$clustering
B$pam3 = as.factor(pam3)
#Reordeno los niveles
B$pam3=relevel(B$pam3,ref="2")#B$pam3=order(B$pam3,levels=c("2","1","3"))
#aux=which(B$pam3==1);aux1=which(B$pam3==2);B$pam[aux]=2;B$pam[aux1]=1
plot(B$pam3,main = "Grupos de comportamientos PAM",xlab = "Grupos",ylab = "Frecuencia",col=rainbow(30, alpha=.6))
prop.table(table(B$pam3))*100

#Distintos procedimientos de agrupación
# pam
pam3 = pam(F,3,diss = TRUE)
pam4 = pam(F,4,diss = TRUE)
pam_3g = silhouette(pam3)
pam_4g = silhouette(pam4)
plot(pam_3g,main="Silueta pam 3",border="blue")
plot(pam_4g,main="Silueta pam 4",border="blue")
km3=kmeans(F,3)$cluster
B$km3=as.factor(km3)
plot(B$km3,main = "Grupos de comportamientos PAM",xlab = "Grupos",ylab = "Frecuencia",col=rainbow(30, alpha=.6))
prop.table(table(B$km3))*100

# kmeans
km3 = kmeans(F,3)
km4 = kmeans(F,4)
km_3g = silhouette(km3$cluster,F)
km_4g = silhouette(km4$cluster,F)
plot(km_3g,main="Silueta km 3",border="green")
plot(km_4g,main="Silueta km 4",border="green")

# hclust
cl=hclust(D,method = "ward.D")
plot(cl)
cl3=cutree(cl,3)
jerarquico_3g = silhouette(cl3,as.dist(F))
jerarquico_4g = silhouette(cutree(cl,4),as.dist(F))
plot(jerarquico_3g,main="Silueta jerarquico 3",border="red")
plot(jerarquico_4g,main="Silueta jerarquico 4",border="red")
B$cl3=cl3
B$cl3=as.data.frame(B$cl3)
### Grupos considerado factores ACM ###
library(cluster)
p_grupos=acm_1[7]
p_grupos=data.frame(p_grupos)
p_grupos=p_grupos[,1:2]
head(p_grupos)
grupos=agnes(p_grupos, metric = "euclidean", stand = FALSE, method = 'ward')
plot(grupos, main = "Dendograma",which=2)
rect.hclust(grupos, k=3)
k = 5
cl = cutree(grupos,k)
cl = factor(cl)
B$cl=cl
table(B$cl)

source('indicadores.R')

ind1=indicadores(grupos$merge,p_grupos,imprime=10)
plot(11:1,ind1$Rcuad)
plot(11:1,ind1$psF,type="l")
plot(11:1,ind1$psT,type="l")
```

```
###########################################################################################
#### Caracterización de conductas indebidas según de aspectos sociodemográficos #####
###########################################################################################
prop.table(xtabs(peso_enaj~B$Edad+Indebido,data=B),m=1)*100 #Edad
prop.table(xtabs(peso_enaj~B$Nivel.Ed+Indebido,data=B),m=1)*100 #Nivel Ed.
prop.table(xtabs(peso_enaj~B$Region+Indebido,data=B),m=1)*100 #Region
prop.table(xtabs(peso_enaj~B$Trab+Indebido,data=B),m=1)*100 #Trab
prop.table(xtabs(peso_enaj~B$Const.Hog+Indebido,data=B),m=1)*100 #Comp.Hog
prop.table(xtabs(peso_enaj~B$Q.Ingreso+Indebido,data=B),m=1)*100 #Quintil.Ingreso
prop.table(xtabs(peso_enaj~B$Discrim+Indebido,data=B),m=1)*100 #Discrim
prop.table(xtabs(peso_enaj~B$Seguridad+Indebido,data=B),m=1)*100 #Seguridad
prop.table(xtabs(peso_enaj~B$Seguridad.Centro.Ed+Indebido,data=B),m=1)*100 #Seguridad centro Ed.

#################################################################
# Caracterizo grupos en funcion de las variables de conductas #
#################################################################

barplot(table(B$pam3,B$Indebido),legend = TRUE,main = "Distribución de grupos según jóvenes Indebidos",
        col = heat.colors(3, alpha = .6),beside = TRUE )
(p = prop.table(xtabs  (peso_enaj ~ pam3+Indebido,data = B),m=1))*100

prop.table(xtabs(peso_enaj~B$pam3+Daño,data=B),m=1)*100 #Daño
prop.table(xtabs(peso_enaj~B$pam3+Detenido,data=B),m=1)*100 #Detenido
prop.table(xtabs(peso_enaj~B$pam3+Fuga.c,data=B),m=1)*100 #Fuga.c
prop.table(xtabs(peso_enaj~B$pam3+Golpea.Ap,data=B),m=1)*100 #Golpeo.ap
prop.table(xtabs(peso_enaj~B$pam3+Manejo.sl,data=B),m=1)*100 #Manejo.sl
prop.table(xtabs(peso_enaj~B$pam3+Porta.arma,data=B),m=1)*100 #Porta.arma
prop.table(xtabs(peso_enaj~B$pam3+Pelea,data=B),m=1)*100 #Pelea
prop.table(xtabs(peso_enaj~B$pam3+Robo.c,data=B),m=1)*100 #Robo.c
prop.table(xtabs(peso_enaj~B$pam3+Sust,data=B),m=1)*100 #Sust
prop.table(xtabs(peso_enaj~B$pam3+Robo.centro.ed,data=B),m=1)*100 #Robo centro Ed.


##############
## Muestreo ##
##############

library(survey)

#ntotal 3818 y nmuestra 3055
B$nper=seq(1,nrow(B))
(tabla.grupos=prop.table(table(B$pam3)))*100
nper=nrow(B)
muestra=round(nper*80/100)

g1=as.numeric(tabla.grupos[1])
g2=as.numeric(tabla.grupos[2])
g3=as.numeric(tabla.grupos[3])
#g4=as.numeric(tabla.grupos[4])

(gg1=round(muestra*g1))
(gg2=round(muestra*g2))
(gg3=round(muestra*g3))
#(gg4=round(muestra*g4))

install.packages("sampling")
library("sampling")

# MAS Estratificado
estrato=strata(B,stratanames=c("pam3"),size = c(gg1,gg2,gg3), method="srswor")#MAS sin remplazo
B.muestrado=getdata(B,estrato)

#jovenes muestrados
j.muestrados=subset(B.muestrado,select = c("nper","Stratum"))

j.muestra=merge(B,j.muestrados,by.x = c("nper"),by.y = c("nper"), all.x = T)
j.muestra$Stratum=as.integer(j.muestra$Stratum)
j.muestra$Stratum[is.na(j.muestra$Stratum)] = 0 #los que no selecciono como muestra le pone 0
table(j.muestra$Stratum)

#Tomando estos datos se realizan las predicciones
j.no.muestra=subset(j.muestra,(j.muestra$Stratum==0))
B.no.muestra=subset(B,j.muestra$Stratum==0)# fijarce si es B

################### DISEÑO ###############
library(survey)

diseño = svydesign(id=~nper,strata=~Stratum, weights=~peso_enaj, data=B.muestrado)

################# MODELOS #################
library(nnet)

modelo = multinom(pam3 ~ 1,weights = B.muestrado$peso_enaj, design=diseño,
                  subset = (Stratum!=0), data=B.muestrado)

modelo.step = step(modelo,scope = list(lower = pam3 ~ 1, upper = pam3 ~  Sexo + Edad +
                                         Nivel.Ed + Region + Trab + Const.Hog + Q.Ingreso +
                                         Seguridad.Centro.Ed + Seguridad + Discrim ),
                   direction = "both",design=diseño, subset = (Stratum!=0))

summary(modelo.step)
```

```
anova(modelo,modelo.step) #Contraste condicional de razon de verosimilitudes

############### PREDICCIONES ##############

pre=predict(modelo.step,newdata = j.no.muestra)
obs = j.no.muestra$pam3
(t=prop.table(table(obs,pre),m=1)*100)
cont = 0
for(i in 1:763){if(pre[i]==obs[i]) cont=cont+1 else cont=cont}
(tcc=cont/763)

#IC y OR
a=exp(confint(modelo.step))
round(a,3)
b=exp(coef(modelo.step))
round(b,2)
##############
#### ÁRBOL ####
##############
source("stratified.R")
B=cbind(B, pam3)
B$nper=seq(1,nrow(B))
set.seed(1234123488)
B.muestra=stratified(B, B$pam3,0.8)
j.muestra=merge(B,B.muestra,by.x = c("nper"),by.y = c("nper"), all.x=T)
j.muestra$pam3.y[is.na(j.muestra$pam3.y)] = 0
B.nomuestra=subset(j.muestra, j.muestra$pam3.y==0)
B.nomuestra=B.nomuestra[,-1]
B.nomuestra=B.nomuestra[,1:24]
B.muestra=B.muestra[,-25]
colnames(B.nomuestra)=colnames(B.muestra)
library(rpart)

#ÁRBOL MAX
arbol.max = rpart(pam3 ~ Sexo + Edad + Nivel.Ed + Region + Trab +
                  Const.Hog + Q.Ingreso + Seguridad.Centro.Ed + Seguridad + Discrim,
                  data=B.muestra,control =rpart.control(cp=0)
                  ,method = 'class')
plotcp(arbol.max)
printcp(arbol.max)#printcp muestra resultados, resumenes
plot(arbol.max,uniform = T,compress = T,margin = 0.0001)
text(arbol.max,use.n = F, all=T,pretty=0,cex=0.5)

pre=predict(arbol.max,newdata = B.muestra, type = "class" )# muestra de prueba
table(B.muestra$pam3,pre)
t=prop.table(table(B.muestra$pam3,pre),m=1)*100
round(t,1)
round(sum(diag(t)) / sum(t)*100,1)

pre0=predict(arbol.max,newdata = B.nomuestra, type = "class" )# muestra de prueba
table(B.nomuestra$pam3,pre0)
t1=prop.table(table(B.nomuestra$pam3,pre0),m=1)*100
round(t1,1)
round(sum(diag(t1)) / sum(t1)*100,1)

###########CP=0.004#############

fit4=prune(arbol.max,cp=0.004)
plot(fit4,uniform = T,compress = T,margin = 0.0001)
text(fit4,use.n = F,all = T,pretty = 0,cex=0.5)
printcp(fit4)

pre4=predict(fit4,newdata = B.muestra,type = "class")
TT=table(B.muestra$pam3,pre4)
t4=prop.table(table(B.muestra$pam3,pre4),m=1)*100
round(t4,1)
round(sum(diag(TT)) / sum(TT)*100,1)


pre41=predict(fit4,newdata = B.nomuestra,type = "class")
table(B.nomuestra$pam3,pre41)
t41=prop.table(table(B.nomuestra$pam3,pre41),m=1)*100
round(t41,1)
round(sum(diag(t41)) / sum(t41)*100,1)

###########CP=0.003#############

fit3=prune(arbol.max,cp=0.003)
plot(fit3,uniform = T,compress = T,margin = 0.0001)
text(fit3,use.n = F,all = T,pretty = 0,cex=0.5)
printcp(fit3)
pre3=predict(fit3,newdata = B.muestra,type = "class")
table(B.muestra$pam3,pre3)
t3=prop.table(table(B.muestra$pam3,pre3),m=1)*100
round(t3,1)
round(sum(diag(t3)) / sum(t3)*100,1)


pre31=predict(fit3,newdata = B.nomuestra,type = "class")
table(B.nomuestra$pam3,pre31)
t31=prop.table(table(B.nomuestra$pam3,pre31),m=1)*100
round(t31,1)
round(sum(diag(t31)) / sum(t31)*100,1)
```

```
##########CP=0.002############

fit2=prune(arbol.max,cp=0.002)
plot(fit2,uniform = T,compress = T,margin = 0.0001)
text(fit2,use.n = F,all = T,pretty = 0,cex=0.5)
printcp(fit2)

pre2=predict(fit2,newdata = B.muestra,type = "class")
table(B.muestra$pam3,pre2)
t2=prop.table(table(B.muestra$pam3,pre2),m=1)*100
round(t2,1)
round(sum(diag(t2)) / sum(t2)*100,1)


pre21=predict(fit2,newdata = B.nomuestra,type = "class")
table(B.nomuestra$pam3,pre21)
t21=prop.table(table(B.nomuestra$pam3,pre21),m=1)*100
round(t21,1)
round(sum(diag(t21)) / sum(t21)*100,1)

##########CP=0.001############

fit1=prune(arbol.max,cp= 0.00137255 )
plot(fit1,uniform = T,compress = T,margin = 0.0001)
text(fit1,use.n = F,all = T,pretty = 0,cex=0.5)
printcp(fit1)
pre1=predict(fit1,newdata = B.muestra,type = "class")
table(B.muestra$pam3,pre1)
t1=prop.table(table(B.muestra$pam3,pre1),m=1)*100
round(t1,1)
round(sum(diag(t1)) / sum(t1)*100,1)


pre11=predict(fit1,newdata = B.nomuestra,type = "class")
table(B.nomuestra$pam3,pre11)
t11=prop.table(table(B.nomuestra$pam3,pre11),m=1)*100
round(t11,1)
round(sum(diag(t11)) / sum(t11)*100,1)

##########CP=0.01############

fit01=prune(arbol.max,cp=0.01)
plot(fit01,uniform = T,compress = T,margin = 0.0001)
text(fit01,use.n = F,all = T,pretty = 0,cex=0.5)
printcp(fit01)

pre01=predict(fit01,newdata = B.muestra,type = "class")
table(B.muestra$pam3,pre01)
t01=prop.table(table(B.muestra$pam3,pre01),m=1)*100
round(t01,1)
round(sum(diag(t01)) / sum(t01)*100,1)
pre011=predict(fit01,newdata = B.nomuestra,type = "class")
table(B.nomuestra$pam3,pre011)
t011=prop.table(table(B.nomuestra$pam3,pre011),m=1)*100
round(t011,1)
round(sum(diag(t011)) / sum(t011)*100,1)


#errores
a=printcp(arbol.max)
write.table(a,"cp.txt",sep='\t')

TT4=table(B.muestra$pam3,pre4)
100-round(sum(diag(TT4)) / sum(TT4)*100,1)

TT41=table(B.nomuestra$pam3,pre41)
100-round(sum(diag(TT41)) / sum(TT41)*100,1)


##########CP=0.003############
TT3=table(B.muestra$pam3,pre3)
100-round(sum(diag(TT3)) / sum(TT3)*100,1)

TT31=table(B.nomuestra$pam3,pre31)
100-round(sum(diag(TT31)) / sum(TT31)*100,1)
round(t3,1)
round(sum(diag(t3)) / sum(t3)*100,1)


pre31=predict(fit3,newdata = B.nomuestra,type = "class")
table(B.nomuestra$pam3,pre31)
t31=prop.table(table(B.nomuestra$pam3,pre31),m=1)*100
round(t31,1)
round(sum(diag(t31)) / sum(t31)*100,1)
##########CP=0.002############
TT2=table(B.muestra$pam3,pre2)
100-round(sum(diag(TT2)) / sum(TT2)*100,1)

TT21=table(B.nomuestra$pam3,pre21)
100-round(sum(diag(TT21)) / sum(TT21)*100,1)
```

```
##########CP=0.001############

TT1=table(B.muestra$pam3,pre1)
100-round(sum(diag(TT1)) / sum(TT1)*100,1)

TT11=table(B.nomuestra$pam3,pre11)
100-round(sum(diag(TT11)) / sum(TT11)*100,1)
##########CP=0.01############

TT01=table(B.muestra$pam3,pre01)
100-round(sum(diag(TT01)) / sum(TT01)*100,1)

TT011=table(B.nomuestra$pam3,pre011)
100-round(sum(diag(TT011)) / sum(TT011)*100,1)

## MUESTRAS ÁRBOLES ##
source("stratified.R")

B=cbind(B, pam3)
B.muestra0=stratified(B, B$pam3,0.8)

set.seed(1234)
B.muestra1=stratified(B, B$pam3,0.8)

set.seed(12345678)
B.muestra2=stratified(B, B$pam3,0.8)

set.seed(8765)
B.muestra3=stratified(B, B$pam3,0.8)
set.seed(4321)
B.muestra4=stratified(B, B$pam3,0.8)

set.seed(1256)
B.muestra5=stratified(B, B$pam3,0.8)

set.seed(3478)
B.muestra6=stratified(B, B$pam3,0.8)

set.seed(123456)
B.muestra7=stratified(B, B$pam3,0.8)

set.seed(1234567)
B.muestra8=stratified(B, B$pam3,0.8)
set.seed(123456789)
B.muestra9=stratified(B, B$pam3,0.8)

set.seed(12341234)
B.muestra10=stratified(B, B$pam3,0.8)


set.seed(123412345)
B.muestra11=stratified(B, B$pam3,0.8)

set.seed(123412346)
B.muestra12=stratified(B, B$pam3,0.8)


set.seed(123412347)
B.muestra13=stratified(B, B$pam3,0.8)

set.seed(123412348)
B.muestra14=stratified(B, B$pam3,0.8)


set.seed(1234123489)
B.muestra15=stratified(B, B$pam3,0.8)


library(rpart)

#ÁRBOL MAX
arbol.max0 = rpart(pam3 ~ Sexo + Edad + Nivel.Ed + Region + Trab +
                Const.Hog + Q.Ingreso + Seguridad.Centro.Ed + Seguridad + Discrim,
                data=B.muestra0,control =rpart.control(cp=0)
                  ,method = 'class')
plotcp(arbol.max0)
printcp(arbol.max0)#printcp muestra resultados, resumenes
plot(arbol.max0,uniform = T,compress = T,margin = 0.0001)
text(arbol.max0,use.n = T, all=T,pretty=0,cex=0.5)


arbol.max1 = rpart(pam3 ~ Sexo + Edad + Nivel.Ed + Region + Trab +
                Const.Hog + Q.Ingreso + Seguridad.Centro.Ed + Seguridad + Discrim,
                data=B.muestra1,control =rpart.control(cp=0)
                  ,method = 'class')

plotcp(arbol.max1)
printcp(arbol.max1)#printcp muestra resultados, resumenes
plot(arbol.max1,uniform = T,compress = T,margin = 0.0001)
text(arbol.max1,use.n = T, all=T,pretty=0,cex=0.5)

arbol.max2 = rpart(pam3 ~ Sexo + Edad + Nivel.Ed + Region + Trab +
                Const.Hog + Q.Ingreso + Seguridad.Centro.Ed + Seguridad + Discrim,
                data=B.muestra2,control =rpart.control(cp=0)
                  ,method = 'class')
```

```
plotcp(arbol.max2)
printcp(arbol.max2)#printcp muestra resultados, resumenes

arbol.max3 = rpart(pam3 ~ Sexo + Edad + Nivel.Ed + Region + Trab +
                   Const.Hog + Q.Ingreso + Seguridad.Centro.Ed + Seguridad + Discrim,
                   data=B.muestra3,control =rpart.control(cp=0)
                     ,method = 'class')

plotcp(arbol.max3)
printcp(arbol.max3)#printcp muestra resultados, resumenes

arbol.max4 = rpart(pam3 ~ Sexo + Edad + Nivel.Ed + Region + Trab +
                   Const.Hog + Q.Ingreso + Seguridad.Centro.Ed + Seguridad + Discrim,
                   data=B.muestra4,control =rpart.control(cp=0)
                     ,method = 'class')

plotcp(arbol.max4)
printcp(arbol.max4)#printcp muestra resultados, resumenes

arbol.max5 = rpart(pam3 ~ Sexo + Edad + Nivel.Ed + Region + Trab +
                   Const.Hog + Q.Ingreso + Seguridad.Centro.Ed + Seguridad + Discrim,
                   data=B.muestra5,control =rpart.control(cp=0)
                     ,method = 'class')

plotcp(arbol.max5)
printcp(arbol.max5)#printcp muestra resultados, resumenes
plotcp(arbol.max5)
printcp(arbol.max5)#printcp muestra resultados, resumenes


arbol.max6 = rpart(pam3 ~ Sexo + Edad + Nivel.Ed + Region + Trab +
                   Const.Hog + Q.Ingreso + Seguridad.Centro.Ed + Seguridad + Discrim,
                   data=B.muestra6,control =rpart.control(cp=0)
                     ,method = 'class')

plotcp(arbol.max6)
printcp(arbol.max6)#printcp muestra resultados, resumenes

arbol.max7 = rpart(pam3 ~ Sexo + Edad + Nivel.Ed + Region + Trab +
                   Const.Hog + Q.Ingreso + Seguridad.Centro.Ed + Seguridad + Discrim,
                   data=B.muestra7,control =rpart.control(cp=0)
                     ,method = 'class')

plotcp(arbol.max7)
printcp(arbol.max7)#printcp muestra resultados, resumenes

arbol.max8 = rpart(pam3 ~ Sexo + Edad + Nivel.Ed + Region + Trab +
                   Const.Hog + Q.Ingreso + Seguridad.Centro.Ed + Seguridad + Discrim,
                   data=B.muestra8,control =rpart.control(cp=0)
                     ,method = 'class')

plotcp(arbol.max8)
printcp(arbol.max8)#printcp muestra resultados, resumenes

arbol.max9 = rpart(pam3 ~ Sexo + Edad + Nivel.Ed + Region + Trab +
                   Const.Hog + Q.Ingreso + Seguridad.Centro.Ed + Seguridad + Discrim,
                   data=B.muestra9,control =rpart.control(cp=0)
                     ,method = 'class')

plotcp(arbol.max9)
printcp(arbol.max9)#printcp muestra resultados, resumenes

arbol.max10 = rpart(pam3 ~ Sexo + Edad + Nivel.Ed + Region + Trab +
                   Const.Hog + Q.Ingreso + Seguridad.Centro.Ed + Seguridad + Discrim,
                   data=B.muestra10,control =rpart.control(cp=0)
                     ,method = 'class')

plotcp(arbol.max10)
printcp(arbol.max10)#printcp muestra resultados, resumenes


arbol.max11 = rpart(pam3 ~ Sexo + Edad + Nivel.Ed + Region + Trab +
                   Const.Hog + Q.Ingreso + Seguridad.Centro.Ed + Seguridad + Discrim,
                   data=B.muestra11,control =rpart.control(cp=0)
                     ,method = 'class')

plotcp(arbol.max11)
printcp(arbol.max11)#printcp muestra resultados, resumenes
```

```
arbol.max12 = rpart(pam3 ~ Sexo + Edad + Nivel.Ed + Region + Trab +
                Const.Hog + Q.Ingreso + Seguridad.Centro.Ed + Seguridad + Discrim,
                data=B.muestra12,control =rpart.control(cp=0)
                  ,method = 'class')

plotcp(arbol.max12)
printcp(arbol.max12)#printcp muestra resultados, resumenes


arbol.max13 = rpart(pam3 ~ Sexo + Edad + Nivel.Ed + Region + Trab +
                Const.Hog + Q.Ingreso + Seguridad.Centro.Ed + Seguridad + Discrim,
                data=B.muestra13,control =rpart.control(cp=0)
                  ,method = 'class')

plotcp(arbol.max13)
printcp(arbol.max13)#printcp muestra resultados, resumenes


arbol.max14 = rpart(pam3 ~ Sexo + Edad + Nivel.Ed + Region + Trab +
                Const.Hog + Q.Ingreso + Seguridad.Centro.Ed + Seguridad + Discrim,
                data=B.muestra14,control =rpart.control(cp=0)
                  ,method = 'class')

plotcp(arbol.max14)
printcp(arbol.max14)#printcp muestra resultados, resumenes


arbol.max15 = rpart(pam3 ~ Sexo + Edad + Nivel.Ed + Region + Trab +
                Const.Hog + Q.Ingreso + Seguridad.Centro.Ed + Seguridad + Discrim,
                data=B.muestra15,control =rpart.control(cp=0)
                  ,method = 'class')
plotcp(arbol.max15)
printcp(arbol.max15)#printcp muestra resultados, resumenes



##########CP=0.004############

fit0=prune(arbol.max0,cp=0.004)
plot(fit0,uniform = T,compress = T,margin = 0.0001)
text(fit0,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit0)


fit1=prune(arbol.max1,cp=0.004)
plot(fit1,uniform = T,compress = T,margin = 0.0001)
text(fit1,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit1)


fit2=prune(arbol.max2,cp=0.004)
plot(fit2,uniform = T,compress = T,margin = 0.0001)
text(fit2,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit2)


fit3=prune(arbol.max3,cp=0.004)
plot(fit3,uniform = T,compress = T,margin = 0.0001)
text(fit3,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit3)

fit4=prune(arbol.max4,cp=0.004)
plot(fit4,uniform = T,compress = T,margin = 0.0001)
text(fit4,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit4)


fit5=prune(arbol.max5,cp=0.004)
plot(fit5,uniform = T,compress = T,margin = 0.0001)
text(fit5,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit5)


fit6=prune(arbol.max6,cp=0.004)
plot(fit6,uniform = T,compress = T,margin = 0.0001)
text(fit6,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit6)

fit7=prune(arbol.max7,cp=0.004)
plot(fit7,uniform = T,compress = T,margin = 0.0001)
text(fit7,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit7)
```

```
fit8=prune(arbol.max8,cp=0.004)
plot(fit8,uniform = T,compress = T,margin = 0.0001)
text(fit8,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit8)


fit9=prune(arbol.max9,cp=0.004)
plot(fit9,uniform = T,compress = T,margin = 0.0001)
text(fit9,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit9)


fit10=prune(arbol.max10,cp=0.004)
plot(fit10,uniform = T,compress = T,margin = 0.0001)
text(fit10,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit10)


fit11=prune(arbol.max11,cp=0.004)
plot(fit11,uniform = T,compress = T,margin = 0.0001)
text(fit11,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit11)

fit12=prune(arbol.max12,cp=0.004)
plot(fit12,uniform = T,compress = T,margin = 0.0001)
text(fit12,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit12)


fit13=prune(arbol.max13,cp=0.004)
plot(fit13,uniform = T,compress = T,margin = 0.0001)
text(fit13,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit13)

fit14=prune(arbol.max14,cp=0.004)
plot(fit14,uniform = T,compress = T,margin = 0.0001)
text(fit14,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit14)


fit15=prune(arbol.max15,cp=0.004)
plot(fit15,uniform = T,compress = T,margin = 0.0001)
text(fit15,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit15)



###########CP=0.003#############

fit0=prune(arbol.max0,cp=0.003)
plot(fit0,uniform = T,compress = T,margin = 0.0001)
text(fit0,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit0)


fit1=prune(arbol.max1,cp=0.003)
plot(fit1,uniform = T,compress = T,margin = 0.0001)
text(fit1,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit1)


fit2=prune(arbol.max2,cp=0.003)
plot(fit2,uniform = T,compress = T,margin = 0.0001)
text(fit2,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit2)

fit3=prune(arbol.max3,cp=0.003)
plot(fit3,uniform = T,compress = T,margin = 0.0001)
text(fit3,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit3)


fit4=prune(arbol.max4,cp=0.003)
plot(fit4,uniform = T,compress = T,margin = 0.0001)
text(fit4,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit4)


fit5=prune(arbol.max5,cp=0.003)
plot(fit5,uniform = T,compress = T,margin = 0.0001)
text(fit5,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit5)


fit6=prune(arbol.max6,cp=0.003)
plot(fit6,uniform = T,compress = T,margin = 0.0001)
text(fit6,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit6)
```

```
ion + Trab +
o.Ed + Seguridad + Discrim,
(cp=0)

nes

ion + Trab +
o.Ed + Seguridad + Discrim,
(cp=0)

nes

ion + Trab +
o.Ed + Seguridad + Discrim,
(cp=0)

nes

ion + Trab +
o.Ed + Seguridad + Discrim,
(cp=0)

nenes
```

```
fit7=prune(arbol.max7,cp=0.003)
plot(fit7,uniform = T,compress = T,margin = 0.0001)
text(fit7,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit7)


fit8=prune(arbol.max8,cp=0.003)
plot(fit8,uniform = T,compress = T,margin = 0.0001)
text(fit8,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit8)


fit9=prune(arbol.max9,cp=0.003)
plot(fit9,uniform = T,compress = T,margin = 0.0001)
text(fit9,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit9)


fit10=prune(arbol.max10,cp=0.003)
plot(fit10,uniform = T,compress = T,margin = 0.0001)
text(fit10,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit10)


fit11=prune(arbol.max11,cp=0.003)
plot(fit11,uniform = T,compress = T,margin = 0.0001)
text(fit11,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit11)

fit12=prune(arbol.max12,cp=0.003)
plot(fit12,uniform = T,compress = T,margin = 0.0001)
text(fit12,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit12)

fit13=prune(arbol.max13,cp=0.0037)
plot(fit13,uniform = T,compress = T,margin = 0.0001)
text(fit13,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit13)


fit14=prune(arbol.max14,cp=0.003)
plot(fit14,uniform = T,compress = T,margin = 0.0001)
text(fit14,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit14)


fit15=prune(arbol.max15,cp=0.003)
plot(fit15,uniform = T,compress = T,margin = 0.0001)
text(fit15,use.n = T,all = T,pretty = 0,cex=0.5)
printcp(fit15)
```