UNIVERSIDAD NACIONAL DE CATAMARCA

FACULTAD DE TECNOLOGÍA Y CIENCIAS APLICADAS

INGENIERÍA EN INFORMÁTICA

TRABAJO FINAL

"Minería de Datos aplicada a Datos del Gran Catamarca de las Encuestas Permanentes de Hogar del año 2017"

Autor:

Ramos, César Alejandro (MU: 779)

Director:

Dr. Hernán Ahumada

Co-Director:

Ing. Carlos Herrera

Catamarca, Diciembre 2018

AGRADECIMIENTOS

Agradezco y dedico este Trabajo en primer lugar a Dios; a Dios Padre, a mi Amado Jesús y a mi Precioso Espíritu Santo, mi Mejor Amigo. Sin Dios no habría podido terminar esta carrera; si he llegado hasta aquí es porque El me dio las fuerzas para continuar y no abandonar; cuando todo se ponía cuesta arriba y deseaba “tirar la toalla”, El se bajó de la tribuna para ayudarme a llegar a la meta. Aquella tarde cuando escuché al Pastor Ricardo Rodríguez del Centro Mundial de Avivamiento contar la historia de “Eric Lidell”, aquel atleta que decidió Honrar a Dios un domingo y abandonar la carrera de los 100 metros en aquellas Olimpiadas de París en 1924, por la que se preparó durante cuatro años; ese atleta que se preparó para una carrera de 100mts y no solo terminó corriendo la de los 400mts sino que obtuvo el primer lugar, y cuando le preguntaron cómo logró ganar el respondió: ***“corro la mitad de mi carrera con todas mis fuerzas, y cuando llego a la mitad levanto mis manos y le pido a Dios que corra por mi”***. Esa historia marcó mi vida, porque estaba estancado en mi carrera, justo con la mitad de materias aprobadas y no podía avanzar; ese día entre lágrimas le dije a Dios: “Señor, he corrido con mis fuerzas esta carrera, y aquí estoy estancado en la mitad, ahora ayúdame, corre por mi”. Y es a partir de ahí que los horarios de trabajo y universidad encajaron como rompecabezas, y terminé aprobando un promedio de 8 materias por año en los siguientes dos años. Esto no lo podría haber hecho en mis fuerzas ni mi capacidad, es mi Señor quien merece todos los méritos. A Él le doy las Gracias, la Gloria y la Honra, porque decidí darle el Primer Lugar de mi vida, y El fue Fiel a su Palabra que dice ***“honraré a los que me honran”*** (1 Samuel 2:30; NTV) y ***“Todo lo puedo en Cristo que me fortalece”*** (Filipenses 4.13; RV).

Agradezco a mi Familia, en especial a mi mamá Noemí, mi papá Julio, mi abuelo y papá de crianza Cayetano, mi abuela Juana que no podrá ver a su Nieto recibirse, pero estará presente en mi corazón, a mis abuelos Julia y Ramón, a mi Tía Alejandra que me ayudó a tomar la decisión de elegir esta carrera, a mis Tías Alicia, Mónica y Mara por su apoyo incondicional, a todos ellos porque estuvieron siempre a mi lado y no me dejaron abandonar la carrera cuando pensé que no podría avanzar. Gracias a ellos que soportaron mis estados de ánimo, buenos y malos, que se alegraron con cada materia aprobada, que me consolaron en algunas derrotas, y me contuvieron en alguna injusticia que me tocó vivir.

Agradezco a los hermanos en la Fe que han orado por mí, para poder terminar esta carrera. A los Hermanos de la Iglesia que Pastoreo, Iglesia Avivamiento Ríos de Dios, y de otras congregaciones. A los Pastores de la “Unión de Pastores Evangélicos de Catamarca” (UPEC), en especial a la Pastora Norma de Conti, Pastores Antonio Trejo, Pablo Díaz, Gustavo Rodríguez, Jimmy Gonzaga, Antonio Suarez, Adriana Mosca y demás consiervos, por sus oraciones. Al Pastor Gustavo Ortega de la Iglesia Avivamiento de Olavarría, y al Pastor Pedro Rodríguez de Santiago del Estero, por su continuo apoyo y oraciones. Al Pastor Robert Acosta del Centro Cristiano de Avivamiento de Chaco por sus oraciones. Hermanos en Cristo no solo de Catamarca, sino de otros lugares del globo, como lo es Colombia; a mis hermanos del Centro Mundial de Avivamiento, Yazmín, Dora, Ana y Karol, gracias por sus oraciones.

Agradezco a los amigos y compañeros que me acompañaron a lo largo de este caminar; aquellos con los que realizamos trabajos juntos, con quienes compartimos momentos no solo de estudio sino de dispersión y bromas. En especial debo agradecer por su apoyo y amistad a Valeria y a Ximena. Gracias por el apoyo en todos estos años.

Agradezco a todos los profesores que me alentaron y ayudaron en todos estos años, en especial a las Lic. Elda Zoto, Natalia Fernandez, Ana Gabriel Buenader, Daniela Lobos Anfuso y Vélia López. Al Dr. Ahumada y el Ing. Carlos Herrera, Director y Co-Director en este trabajo, por el apoyo brindado a lo largo del desarrollo de esta Tesis. Agradezco también al Lic. Luis Alfonso Cutro, de la “Universidad Nacional del Nordeste” por su colaboración y guía en asuntos pertinentes a este trabajo.

Agradezco al Sr. Carlos Alberto Romero y a la Profesora Susana Ogas de la Dirección de Estadística y Censo, por su predisposición y por proveerme del material necesario. También dedico este trabajo al Dr. Juan Carlos Juárez por todo lo que hizo por mí de niño, pues gracias a él y a Dios soy un joven sano y he logrado alcanzar esta meta.

Si tuviera que nombrar uno por uno a quien agradecer, probablemente no alcanzarían las páginas. Pero quiero hacer una mención especial a mis mascotas: Panchito, Pollito y ToTo; Kofi, Vaquita, Esquinsy, Sofi y Pandi. A la mascota que quizás más amé; un Cobayo a quien llame “Panchito”, quien estuvo a mi lado mientras yo estudiaba, que me miraba fijo cuando me veía hablar frente a un pizarrón preparando alguna materia. Mi gordito ya no está, hace un año me dejó; pero él era el público ante quien practicaba mis exámenes. Panchito, mi gordo, siempre estarás en mis recuerdos preparando las materias para rendir. Para él también este trabajo. Y también mencionar a mi perro ToTo que murió este año; el cual me robaba las gomas y lapiceras mientras estudiaba.

Hoy puedo decir, como escribieron en aquel Telegrama los hermanos “Wright” al tener la primera prueba exitosa del vuelo de una aeronave: “***Conseguido***”.

Finalizo mis agradecimientos, dándole una vez más la Gloria al Dios a quien sirvo, al Primero y al Último, al Dios de Israel, con las cinco “solas” que marcaron la Reforma Protestante, impulsada principalmente por el Teólogo “Martin Luther” en el siglo XVI:

“Sola Scriptura”

“Sola Fide”

“Sola Gratia”

“Solo Christo”

“SOLI DEO GLORIA”

“¡Pues el SEÑOR concede sabiduría! De su boca provienen el saber y el entendimiento.”

(Proverbios 2.6; NTV)

Tabla de contenido

FIGURAS

TABLAS

RESUMEN

El presente Trabajo Final tiene como objetivo aportar y complementar a los Análisis Clásicos de Estadística, con Técnicas novedosas de “Minería de Datos” que permitan detectar características comunes entre Hogares e individuos del Gran “Catamarca” en base a los Datos de la “Encuesta Permanente de Hogar”, realizada por el INDEC, correspondiente al primer Trimestre del año 2017.

En este trabajo se plantea el conocimiento existente sobre Minería de Datos, el Aprendizaje Automatizado Supervisado y No Supervisado, y sus Técnicas: “Árboles de Decisión” y “Clustering” respectivamente. Las Tecnologías básicas empleadas fueron: el Lenguaje de Programación “R”, y la herramienta “RStudio”.

La Metodología que se implementó en este trabajo es “CRIDP-DM”, una de las mas empleadas por los Analistas de Negocio, principalmente en el Proceso de Minería de Datos. Es una Metodología sujeta a estándares internacionales, además de ser confiable y amigable para el usuario.

Se han aplicado las Técnicas de “Árboles de Decisión” y “Clustering”, obteniéndose, mediante estos análisis, descripciones gráficas en base a la situación socio-económica de la muestra poblacional. Mediante Clustering se pudo agrupar la población con características similares y se profundizó en el estudio de los grupos familiares de los centroides de los respectivos clusters. Con la Árboles de decisión se pudo determinar jerárquicamente la influencia de las variables objetivo “Nivel de Estudio” y “Tipo de Cobertura Social”, en función de un selecto grupo de variables predictoras; analizando las distintas situaciones socio-económicas.

Palabras Claves: Minería de Datos, Encuesta Permanente de Hogar (EPH), Instituto Nacional de Estadística y Censo (INDEC), Aprendizaje Automatizado, Árboles de Decisión, Clustering, Lenguaje R, RStudio, CRISP-DM.

Capitulo 1: Introducción.

1.1 Introducción

La demografía es una disciplina que estudia las características de una población. Funda sus métodos en el manejo de información estadística, por tanto resulta de interés tanto a urbanistas, sociólogos, antropólogos, arquitectos y otros profesionales de las ciencias sociales, como también a estadísticos.

En la República Argentina, el relevamiento de la evolución de la población se lleva a cabo mediante la comparación de los datos del INDEC (Instituto Nacional de Estadística y Censos), a través de los censos que se realizan cada diez años, y también mediante las Encuesta Permanente de Hogar (EPH) que se realizan en forma trimestral sobre una muestra poblacional. Estos datos sirven de fundamento para inferir sobre las dinámicas poblacionales reales.

1.1.1 Sobre la EPH

“La Encuesta Permanente de Hogares (EPH) es un programa nacional de producción sistemática y permanente de indicadores sociales que lleva a cabo el Instituto Nacional de Estadística y Censos (INDEC), que permite conocer las características socio-demográficas y socioeconómicas de la población” (INDEC, 2003). “En su modalidad original, se ha venido aplicando en Argentina desde 1973, dos veces al año (mayo y octubre)” (Gentile, 2015). Actualmente se realiza de forma trimestral y a diferencia de los censos, se realiza sobre una muestra poblacional.

“La EPH es una encuesta por muestreo, esto significa que para conocer las diversas características del total de los hogares, se encuentra una pequeña fracción representativa de los mismos.” (Gentile, 2015)

La EPH del primer Trimestre del año 2017 incluye los datos de 18.478 Hogares y 28.595 Individuos de 32 aglomerados urbanos de la República Argentina. Los relevamientos son realizados por las Direcciones Provinciales de Estadística bajo las normas técnicas y metodológicas fijadas y monitoreadas por el equipo central de la EPH en el INDEC. En el presente trabajo solo se trabajará con los registros de datos pertenecientes al Aglomerado “Gran Catamarca”, siendo estos 526 Hogares y 1907 Individuos encuestados.

Los hogares a encuestar se seleccionan de forma aleatoria, en dos etapas:

1) Dentro de cada Aglomerado se selecciona una cantidad de radios censales o subdivisiones de las mismas (áreas).
2) Se listan las viviendas particulares de las áreas seleccionadas, y a partir de ese listado se realiza una selección aleatoria de las viviendas. Los hogares que habitan esas viviendas son los hogares que se encuestarán.

“Cabe señalar que en base a esta encuesta se proporcionan regularmente, entre otros resultados, las tasas oficiales de empleo, desocupación, subocupación y pobreza. La difusión de esos resultados se complementa con la producción habitual de una gran cantidad de tabulados (para cada uno de los aglomerados, para las regiones estadísticas y para el total de los aglomerados), bases de datos y publicaciones”. (INDEC, 2003)

Las Técnicas Estadísticas de Análisis de Datos, con las que en general se realiza este análisis proporcionan medidas de tendencia central y de dispersión acerca del comportamiento de las variables en forma individual. El Análisis con Minería de Datos permitirá encontrar interrelaciones entre diferentes variables sin necesidad hacer conjeturas al respecto. De esta forma, aplicando las Técnicas de Minería de Datos, podremos corroborar resultados de las Técnicas Estadísticas tradicionales, o encontrar nuevos parámetros.

1.1.2 Objetivos de las EPH:

“El propósito central de la investigación que sustenta la EPH consiste en caracterizar a la población en términos de su inserción socioeconómica teniendo peso significativo para su determinación los aspectos socio-laborales. En este sentido, pretende conocer la situación de la población en la estructura social a través de la posición que tienen los individuos y hogares, núcleos básicos de convivencia de los cuales las personas se asocian.

En función de esos objetivos generales, la EPH rescata un conjunto de dimensiones básicas que responden a los siguientes ejes conceptuales:

Caracterizar a la población en términos de:

a) Sus características demográficas.
b) Su inserción en la producción social de bienes y servicios.
c) Su participación en la distribución del producto social.” (INDEC, 2003)

De estos tres ejes fundamentales se derivan las temáticas centrales de la Encuesta:

1. Características Demográficas: La Encuesta mide características demográficas básicas.
2. Inserción en la producción Social de bienes y servicios: Se miden las características ocupacionales y de migraciones.
3. Participación en la distribución del producto social: Mide características habitacionales, educacionales y de ingresos.

1.1.3 Instrumentos de Medición.

El cuestionario consiste en planillas con preguntas breves, las cuales son leídas textualmente. Los encuestadores son capacitados en las categorías y las variables sujetas a medición, para la correcta aplicación del cuestionario.

La EPH consta de dos cuestionarios: uno para el Hogar y otro individual para cada una de las personas que lo habitan.

“Esta encuesta se basa en un cuestionario de respuestas múltiples conformado por secciones que recopilan información que incluye, entre otros aspectos, la siguiente información” (Torres, D. L., et.al, 2011):

1.1.4. Principales Temáticas de la EPH.

1.1.4.1 Condición de Residencia

Esta temática es de principal importancia, dado que permite conformar la población objetivo por medio de la identificación de los miembros del hogar. “En términos de la EPH un hogar se define como una persona o grupo de personas, parientes o no, que habitan bajo un mismo techo en un régimen de tipo familiar; es decir, comparten sus gastos en alimentación u otros esenciales para vivir”. (INDEC, 2003)

1.1.4.2. Características habitacionales y del hábitat.

En esta sección se indaga sobre las características habitacionales del hogar que habita en dicha vivienda; indicadores con capacidad de discriminar situaciones de precariedad habitacional y de hábitat inadecuado; referencias al material predominante de los pisos interiores, material de cubierta exterior del techo, existencia de cielorraso, la fuente de provisión del agua, existencia de baños o letrinas, conexiones de electricidad y agua, existencia de basurales, o si la casa está en zona de inundación o zonas de emergencia u otras.

Estas características se relevan solo en la primera entrevista de la vivienda seleccionada, debido a que los cambios estructurales por lo general cambian en periodos largos de tiempo.

1.1.4.3. Características Socio-demográficas.

Las variables captadas son: sexo, edad, relación de parentesco, situación conyugal y educación, existencia de cobertura médica, lugar de nacimiento y último lugar de residencia (específicamente para casos de migraciones), entre otras.

1.1.4.4. Organización del hogar.

“La organización familiar del trabajo doméstico puede asociarse con otras temáticas no menos significativas, tales como: la transformación de los roles al interior del hogar por distintas causas sociales y económicas, la compatibilización de las obligaciones familiares y laborales, la carga de trabajo, la existencia de tiempo libre para el descanso, etc.” (INDEC, 2003)

También hay bloques de preguntas sobre la división familiar de las tareas domésticas e indagando sobre la/s persona/s responsable/s de realizarlas y la/s que colabora/n con dichas tareas, y sobre la presencia de personas con discapacidades en el hogar.

1.1.4.5. Estrategias de manutención de los hogares.

“Esta temática tiene por objeto indagar las diversas modalidades de obtención de recursos que utilizan los hogares para su manutención.”(INDEC, 2003)

Se tienen en cuenta los ingresos tradicionales y otras estrategias de manutención. Las distintas estrategias se pueden caracterizar según la pertenencia socioeconómica de las unidades domésticas.

1.1.5. Características de la EPH

1.1.5.1. Características del diseño de la muestra.

La EPH es una encuesta por muestreo. Es decir, que tomando una pequeña fracción que represente a la población, se puede conocer una tendencia, o características del total de los hogares. La precisión de los datos obtenidos son garantizados por las técnicas estadísticas utilizadas.

Los hogares encuestados son seleccionados de forma aleatoria en dos etapas:

- 1ra etapa: Dentro de cada Aglomerado se seleccionan una cantidad de áreas o subdivisiones del mismo.
- 2da etapa: Se listan todas las viviendas particulares de las áreas seleccionadas. Con ese listado se realiza una selección aleatoria de las viviendas. Los hogares a encuestar son los hogares que habitan en esas viviendas.

1.1.5.2. Periodicidad.

La Encuesta tiene una periodicidad trimestral, definidos de la siguiente forma:

1er Trimestre: Enero, Febrero, Marzo.

2do Trimestre: Abril, Mayo, Junio.

3er Trimestre: Julio, Agosto, Septiembre.

4to Trimestre: Octubre, Noviembre, Diciembre.

De esta forma, se obtienen cuatro estimaciones por año.

1.1.5.3. Ventana de observación.

“El periodo para el cual se brinda información se denomina ‘ventana de observación’”. (INDEC, 2003)

La EPH plantea como “ventana de observación” el trimestre. De esta forma se brinda información con mayor frecuencia, y se observa el comportamiento de las distintas variables a lo largo del año. Además se evita “el riesgo de observar una semana atípica y considerarla como representativa de la situación laboral, que puede cambiar en un periodo más largo.” (INDEC, 2003)

1.1.5.4. Tamaño de la Muestra.

“Pensando en la extensión de la muestra a nivel nacional se propuso un total de 25.000 hogares por trimestre y 100.000 hogares por año, bajo los siguientes supuestos:

- Detectar diferencias significativas de al menos un 0,5% entre dos estimaciones de la tasa de desempleo de dos trimestres consecutivos a nivel nacional y de por lo menos un 1% a nivel regional.

- Estimar la tasa nacional de desempleo con un coeficiente de variación al 2% y las tasas por región con un coeficiente de variación del 5%.” (INDEC, 2003)

Se llama “dominio” a “cualquier subdivisión de la población acerca de la cual se puede dar información numérica de precisión conocida”. (INDEC, 2003)

Los datos a Analizar de la EPH, corresponden al primer Trimestre del año 2017 incluye los datos de 18.478 Hogares y 28.595 Individuos de 32 aglomerados urbanos de la República Argentina. En el presente trabajo solo se trabajará con los registros de datos pertenecientes al Aglomerado “Gran Catamarca”, siendo estos 526 Hogares y 1907 Individuos encuestados.

1.2 Planteamiento del problema

Mediante la Estadística tradicional, es posible obtener resultados en torno a dos variables, y expresarlos en porcentajes, tablas y gráficos. En el presente Trabajo Final se propone realizar un proceso de Minería de Datos, para así obtener nuevos patrones o comportamientos, como resultado de analizar más de dos variables en forma simultánea, mediante los métodos y algoritmos del Aprendizaje Automatizado (Machine Learning) utilizando el lenguaje de programación “R”, con la finalidad complementar al análisis de datos de la Estadística tradicional.

Se busca obtener patrones o comportamientos de los Datos provistos por las EPH a través de Procesos de Minería de Datos, mediante algoritmos de Aprendizaje Automatizado Supervisados (Árboles de Decisión) y No Supervisados (Clustering)

Actualmente, la Minería de Datos mediante técnicas de Aprendizaje Automatizado, permite detectar la existencia de interrelaciones entre los datos, y así descubrir patrones o comportamientos que sirvan para complementar los resultados obtenidos por la Estadística clásica.

A los fines de encontrar características Socioeconómicas comunes o frecuentes, tanto en Hogares como en Individuos, en el Aglomerado “Gran Catamarca” perteneciente a la Región NOA, es que se analizarán los datos publicados por el INDEC (http://www.indec.gob.ar/), correspondientes a la EPH del primer Trimestre del Año 2017.

Es necesario aplicar métodos No Estadísticos para complementar y ampliar la información obtenida mediante Técnicas Estadísticas. Estos Métodos permiten, mediante algoritmos, de una forma rápida y eficiente encontrar tendencias o características que se podrán contrastar con los Resultados Estadísticos. Mediante la Técnica de Árboles de Decisión se podrán encontrar en los Datos Analizados en el presente trabajo, características Socio-Económicas que nos permitirán analizar la Población. Esta Técnica permitirá estudiar una Variable Objetivo en función de la cantidad deseada de variables predictoras. Por otra parte, la Técnica de Clustering, permitirá discriminar la Población en grupos con

características Socio-Económicas similares. Los resultados de aplicar estas Técnicas se podrán contrastar con los Resultados Estadísticos, como así también podrán aportar Resultados Innovadores que los Métodos Estadísticos tarden tiempo en analizar.

1.3 Antecedentes

Si bien las investigaciones relacionadas a la Minería de Datos aplicadas a datos socio-demográficos son numerosas, los trabajos aplicados a datos del INDEC, específicamente a las EPH, no abundan.

Entre los trabajos enfocados en la temática se pueden citar (Cutro, 2008), donde “se propone desarrollar un proceso de extracción de conocimiento a partir de los datos de la Encuesta Permanente de Hogares (EPH) suministradas por el Instituto Nacional de Estadística y Censo (http://www.indec.com.ar/)”. En (Torres, et.al, 2011) se utilizaron algoritmos de clasificación con enfoques descriptivos para obtener nuevos conocimientos sobre familias de Santa Fe. Para ello se utilizaron los datos del relevamiento del año 2009, y se trabajó con el software de Minería de Datos Weka 3.6.2.

Podemos enunciar trabajos de otros países, dedicados a abordar la Minería de Datos sobre datos socio-económicos, como (Vásquez, E., 2013) en el Perú. Este libro “aspira a brindar un conjunto de herramientas técnicas que permitan al gestor social mejorar su competitividad profesional. Para ello, las contribuciones son ensayos técnicos que buscan presentar metodologías de cálculos para abordar temas claves en el quehacer de las políticas y proyectos sociales. Específicamente, aquí se abordan mediciones de una gama de indicadores de pobreza, metodologías de manejo de bases de datos a partir de encuestas a hogares, así como la formulación de iniciativas de inversión en salud, educación, alimentación y protección social, entre otros.” “Para ello, se comienza con el tratamiento de las encuestas de hogares, se estudian los alcances y limitaciones de la “minería de datos”, se escudriñan las bases de datos observables en temas educativos y se termina con soluciones posibles cuando un proyecto no tiene línea de base”. (Vásquez, E., 2013). Otro trabajo que aplica Minería de Datos a datos sobre Rendimiento Académico, como (Ahumada, et.al 2016), donde se utilizó la Técnica de Reglas de Asociación para evaluar el rendimiento académico de los Estudiantes de Ingeniería. Esta técnica permitió realizar un análisis novedoso de los resultados obtenidos. Los resultados evidencian el alto porcentaje de alumnos que no logran regularizar ninguna materia del primer año, y otro porcentaje importante no logra regularizar el número de asignaturas necesarias para mantenerse como alumno activo de la facultad. También, por la aplicación de las mismas herramientas a utilizar en este trabajo, citamos (Martinez, 2016), donde se hizo uso de Minería de Datos y el Lenguaje de programación “R” con la herramienta RStudio para desarrollar un módulo de pronóstico de la demanda eléctrica en una zona geográfica determinada; utilizando para ello como datos, la temperatura en grados centígrados y de la demanda en MW (Megavatios).

1.4 Objetivos

Objetivo General

- Aportar y complementar los Análisis Clásicos de Estadística, con técnicas novedosas de Minería de Datos que permitan detectar características comunes entre hogares e individuos del Gran Catamarca en base a Datos de la EPH.

Objetivos Específicos

- Obtener descripciones gráficas de diferentes tipos de hogares e individuos en base a su situación socio-económica.
- Determinar agrupamientos de hogares e individuos que reúnan características similares y que a su vez difieren con respecto a elementos de otros grupos
- Determinar jerárquicamente la influencia de las diferentes variables demográficas con relación a distintas situaciones socioeconómicas.

1.5 Importancia y Alcance del estudio

El alcance de la informática y las comunicaciones ha producido una sociedad globalizada que se alimenta de la información, sin embargo la mayoría de esta información se encuentra en un formato crudo. Existe una gran cantidad de conocimiento atrapado en Bases de Datos, el cual es importante para cada negocio, pero que aun no ha sido descubierto.

Hoy existe el Hardware necesario para almacenar y poder acceder a volúmenes de datos impensados hace décadas atrás. Pero esto hace que sea humanamente imposible analizar toda esta información o encontrar patrones escondidos en estas Bases de Datos. Es aquí donde reside la importancia de la “Minería de Datos”.

La Minería de Datos forma parte de un proceso llamado KDD (Descubrimiento de Conocimiento de Base de Datos por sus siglas en Ingles), y es aquí donde se produce la extracción del conocimiento. Mediante las Técnicas de Minería de Datos se puede acceder a grandes volúmenes de Datos y es posible encontrar patrones, realizar predicciones, y aplicaciones tan variadas como el reconcomiendo de voz o de rostros, análisis de ventas y marketing, análisis poblaciones, entre otros.

En el presente Trabajo se aplicarán Técnicas de Minería de Datos para encontrar Patrones Socio-Económicos en la “Encuesta Permanente de Hogar” correspondiente al primer Trimestre del año 2017 en la Región denominada “Gran Catamarca”. Como ya se ha expuesto anteriormente, la EPH es una encuesta realizada por el INDEC en forma trimestral en todo el Territorio Argentino, seleccionando muestras Poblacionales que marcan una tendencia en la Población en general. Esta encuesta es de suma importancia, dado que en base a sus resultados se publican, por ejemplo, los índices de empleo y desempleo. Este trabajo se realiza mediante Técnicas Estadísticas, y es mediante la Minería de Datos que se busca hacer un aporte utilizando Análisis No Estadísticos.

1.6 Estructura del trabajo

El presente Trabajo se divide en capítulos, los cuales se describen a continuación:

- **Capítulo 1. “Introducción”:** Se presenta el trabajo propuesto, el problema planteado, los antecedentes existentes, justificación y objetivos.
- **Capítulo 2. “Minería de Datos”:** En este capítulo se presenta el Marco Teórico sobre Minería de Datos, su importancia, historia, aplicaciones, tipos de datos que se pueden tratar y los análisis correspondientes.
- **Capítulo 3. “Análisis Exploratorio de los Datos”:** En el presente capítulo se presenta el enfoque metodológico. Se procede a explicar las etapas de la Metodología CRIP-DM (una de las principales en el proceso de Minería de Datos) y se desarrollan las primeras, exponiendo de esta manera las Técnicas a utilizar.
- **Capítulo 4. “Evaluación de los Datos”:** En este capítulo se introduce el Aprendizaje Automatizado, y se aplican las Técnicas correspondientes sobre los Datos de la EPH: Árboles de Decisión en el Aprendizaje Supervisado y Clustering en el Aprendizaje No Supervisado. Esto en el marco de la Metodología CRISP-DM. Luego de aplicar estas técnicas se describen los resultados obtenidos.
- **Capítulo 5. “Conclusión”:** Se describen las conclusiones del presente trabajo, poniendo en manifiesto si se cumplieron los objetivos iniciales.

.

Capítulo 2: Minería de Datos

2.1 Introducción

“La importancia de los datos está en su capacidad de asociarse dentro de un contexto para convertirse en información. Por sí mismo los datos no tienen capacidad de comunicar un significado y por lo tanto no pueden afectar el comportamiento. En cambio la información reduce nuestra incertidumbre (sobre algún aspecto de la realidad) y, por tanto, nos permite tomar mejores decisiones.” (Cutro, 2008)

Nos encontramos en un periodo de la historia, donde diariamente se almacenan grandes cantidades de datos; y el análisis de dichos datos es una “Necesidad”.

Esta “Necesidad” puede ser suplida por un campo de la Informática llamado “Minería de Datos”. Esta disciplina proporciona las herramientas necesarias para obtener conocimiento a partir de los datos.

2.2 La era de los datos

Hay mucha información digitalizada y almacenada en Bases de datos, se podría decir que estamos “ahogados” de información, pero “sedientos” de conocimiento. Estos datos podemos catalogarlos de “baratos”, y que lo realmente es “valioso” es el conocimiento. Por lo que podemos denominar a este periodo de tiempo, “La era de los datos”, debido a que 2.5 Exabytes (2.500.000 Tb) de datos se vierten en nuestras redes informáticas, la World Wide Web (www) y distintos medios de almacenamiento, diariamente desde los distintos ámbitos de desarrollo de la sociedad, comercios, medios de comunicación, redes sociales, películas, libros, sistemas de salud, radiografías, etc. Un crecimiento que resulta de la informatización de nuestra sociedad y el desarrollo de herramientas de almacenamiento de datos. “Esta explosión de datos no supone un aumento de nuestro conocimiento, puesto que resulta imposible procesarlos con los métodos clásicos.” (Cutro, 2008)

Grandes volúmenes de datos son generados por las empresas de todo el mundo; esto incluye transacciones de ventas, descripciones de productos, promociones de venta, el desempeño de la empresa, comentarios de los clientes, etc. Un ejemplo es la cadena Wal-Mart, la cual tiene miles de sucursales en el mundo, y deben administrar todos sus datos. La industria Médica y de Salud también genera grandes cantidades de datos de registros médicos, monitoreo de pacientes e imágenes médicas, etc. Cabe mencionar los protocolos de intercambio de datos en salud como HL7 (Health Level Seven), los cuales permiten compartir los datos de los pacientes entre distintas instituciones médicas y sistemas de

salud (HL7, 2007). No podemos pasar por alto las comunidades y redes sociales, blogs y demás, los cuales producen miles de datos, imágenes y videos digitales. Y podríamos seguir con una lista interminable de fuentes que generan estas enormes cantidades de datos.

Es debido a este crecimiento que llamamos a nuestro tiempo “La era de los datos”. Para descubrir información valiosa a partir de estos grandes volúmenes de datos, y transformarlos en conocimiento organizado, es que existen herramientas potentes para lograrlo. La necesidad de poder interpretar estos datos, organizarlos y obtener conocimiento, es lo que ha llevado al nacimiento de la “Minería de Datos”.

“Un motor de búsqueda como Google recibe cientos de millones de consultas cada día. Cada consulta se puede ver como una transacción en la que el usuario describe su ***necesidad de información***. Algunos patrones encontrados en las consultas de búsqueda de usuarios pueden revelar un conocimiento invaluable que no puede obtenerse leyendo solo datos individuales. Por ejemplo, Google Flu Trends utiliza términos de búsqueda específicos como indicadores de la actividad de la gripe. Utilizando datos agregados de búsqueda de Google, Flu Trends puede estimar la actividad de la gripe hasta dos semanas más rápido que los sistemas tradicionales. Este ejemplo muestra cómo la minería de datos puede convertir una gran colección de datos en conocimiento que puede ayudar a enfrentar un desafío global actual.” (Han et al. 2011)

2.3 Un poco de Historia

La minería de datos se puede ver como resultado de la evolución natural de la tecnología de la información.

Desde la **década de 1960**, las bases de datos y la tecnología de la información ha evolucionado sistemáticamente desde los sistemas de procesamiento de archivos primitivos hasta sofisticados y poderosos sistemas de bases de datos.

En la **década de 1970**, la investigación y el desarrollo de los sistemas de bases de datos pasaron de los primeros sistemas de bases de datos jerárquicos y de red a sistemas de bases de datos relacionales. Los usuarios obtuvieron de esta forma, un acceso conveniente y flexible a los datos a través de los lenguajes de consulta, interfaces de usuario, optimización de consultas y administración de transacciones. Los métodos eficientes para el “Procesamiento de Transacciones en Línea” (OLTP), donde una consulta se ve como una transacción de sólo lectura, contribuyeron sustancialmente a la evolución y amplia aceptación de la tecnología relacional como una herramienta importante para el almacenamiento, recuperación y administración eficientes de grandes cantidades de datos.

Tras el establecimiento de sistemas de gestión de bases de datos, la tecnología de bases de datos se orientó hacia el desarrollo de sistemas avanzados de bases de datos, almacenamiento de datos y minería de datos para el análisis avanzado de datos y bases de datos basadas en web.

El análisis avanzado de los datos se originó a partir de finales de los **años ochenta**. Esta tecnología proporciona un gran impulso a la base de datos y a la industria de la información, y permite disponer de un gran número de bases de datos y repositorios de información para la gestión de transacciones, la recuperación de información y el análisis de datos. Los datos ahora se pueden almacenar en muchos tipos diferentes de bases de datos y repositorios de información.

Una arquitectura emergente de repositorio de datos es el **Almacén de Datos** (**Data Warehouse**). Este es un repositorio de múltiples orígenes de datos heterogéneos organizados bajo un esquema unificado en un único sitio para facilitar la toma de decisiones de gestión. La tecnología de almacén de datos incluye la limpieza de datos, la integración de datos y el “Procesamiento Analítico en Línea” (OLAP). Las herramientas OLAP soportan el análisis multidimensional y la toma de decisiones, pero requieren herramientas de análisis de datos adicionales para el análisis en profundidad, como herramientas de minería de datos que proporcionan clasificación de datos, clustering, detección de anomalías y la caracterización de cambios en los datos a lo largo del tiempo.

Se han acumulado enormes volúmenes de datos más allá de bases de datos y almacenes de datos. Durante los **años noventa**, comenzaron a aparecer la World Wide Web y bases de datos basadas en web (por ejemplo, bases de datos XML). Las bases de información globales basadas en Internet, como la WWW y varios tipos de bases de datos interconectadas y heterogéneas, han surgido y desempeñan un papel vital en la industria de la información.

”En resumen, la abundancia de datos, junto con la necesidad de poderosas herramientas de análisis de datos, ha sido descrita como una situación rica en datos pero pobre en información. La enorme cantidad de datos, que se acumulan rápidamente y se almacenan en grandes y numerosos repositorios de datos, ha superado ampliamente nuestra capacidad humana de comprensión sin el uso de herramientas. Como resultado, los datos recogidos en grandes repositorios de datos se convierten en "tumbas de datos", archivos de datos que rara vez se visitan. Como consecuencia, las decisiones importantes se toman a menudo basadas no en los datos ricos en información almacenados en los repositorios de datos, sino más bien en la intuición del decisor, simplemente porque el tomador de decisiones no tiene las herramientas para extraer el valioso conocimiento incrustado en la gran cantidad de datos. Se han hecho esfuerzos para desarrollar tecnologías de conocimiento y sistemas basados en el conocimiento, que normalmente se basan en los usuarios o expertos en el dominio para introducir manualmente los conocimientos en bases de conocimientos. Desafortunadamente, sin embargo, el procedimiento manual de introducción de conocimientos es propenso a sesgos y errores y es extremadamente costoso y requiere mucho tiempo. La ampliación de la brecha entre los datos y la información requiere el desarrollo sistemático de herramientas de minería de datos que pueden convertir las tumbas de datos en "nuggets de oro" del conocimiento.” (Han et al, 2011)

2.4 Que es la Minería de Datos

En (Witten et al. 2000) citado por (Torres, D. L., et.al, 2011) “conceptualizan a la Minería de Datos como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos.”

La Minería de Datos es un tema interdisciplinario y puede definirse de muchas maneras diferentes. Incluso el término Minería de Datos no presenta realmente todos los componentes principales. Para referirse a la extracción de oro de los arrecifes o la arena, decimos minería de oro en lugar de minería de roca o arena. Análogamente, la Minería de Datos debería haberse denominado más apropiadamente "Minería del Conocimiento a Partir de Datos". Sin embargo, a corto plazo, la minería del conocimiento puede no reflejar el énfasis en la minería de grandes cantidades de datos. Podemos decir que la Minería es un término vívido que caracteriza el proceso que encuentra un pequeño conjunto de preciosas pepitas de una gran cantidad de materia prima. Por lo tanto, tal denominación errónea que llevaba tanto "Datos" como "Minería" se convirtió en una opción popular.

“El KDD (Knowledge Discovery from Database) es el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia, comprensibles a partir de los datos” (Cutro, 2008). Muchas personas tratan la Minería de Datos como un sinónimo del KDD (“Descubrimiento de Conocimiento a Partir de Datos), mientras que otros consideran que la minería de datos es simplemente un paso esencial en el proceso de descubrimiento de conocimiento. El proceso de “Descubrimiento de Conocimiento” se muestra en la Figura 2.1 como una secuencia iterativa de los siguientes pasos:

1. Limpieza de datos (para eliminar el ruido y datos inconsistentes).

2. Integración de datos (donde se pueden combinar varias fuentes de datos).

3. Selección de datos (donde los datos relevantes para la tarea de análisis se recuperan de la base de datos).

4. Transformación de datos (donde los datos son transformados y consolidados en formas apropiadas para la minería realizando operaciones de resumen o agregación).

5. La Minería de Datos (un proceso esencial en el que se aplican métodos inteligentes para extraer patrones de datos).

6. Evaluación de patrones (para identificar los patrones verdaderamente interesantes que representan el conocimiento basado en medidas de interés).

7. Presentación del conocimiento (donde la visualización y las técnicas de representación del conocimiento se utilizan para presentar el conocimiento minado a los usuarios). (Han et al. 2011)

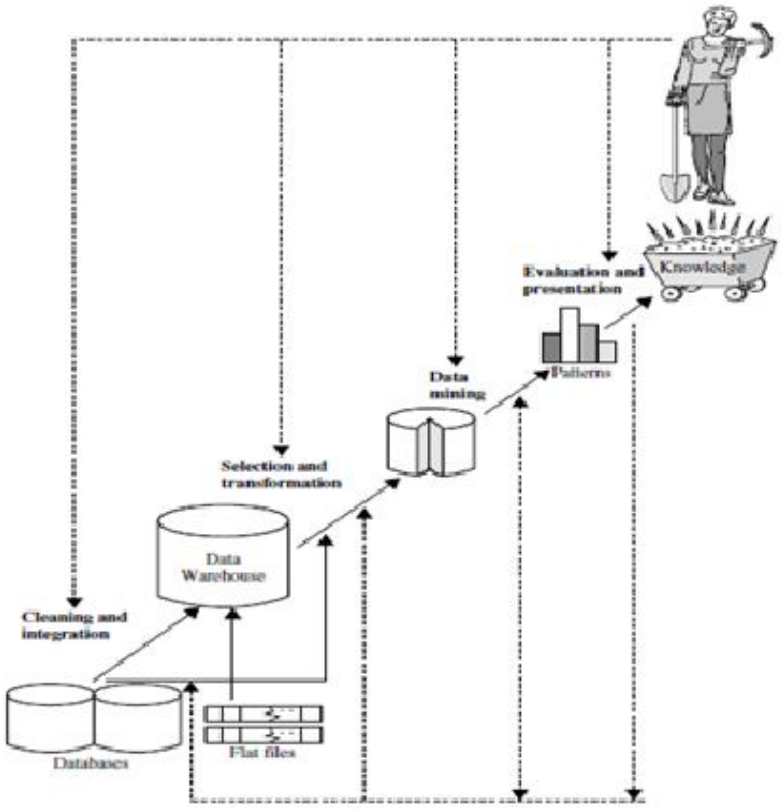


Figura 2. 1 Minería de Datos como uno de los pasos en el proceso de descubrimiento de conocimiento. (Han et al. 2011)

Los pasos 1 a 4 son las diferentes formas de pre-procesamiento de datos, donde los datos se preparan para la Minería. El paso de minería de datos puede interactuar con el usuario o una base de conocimientos. Los patrones interesantes se presentan al usuario y se pueden almacenar como nuevos conocimientos en la base de conocimientos.

La vista anterior muestra la Minería de Datos como un paso en el proceso de Descubrimiento de Conocimiento, aunque es esencial porque descubre patrones ocultos para la evaluación. Sin embargo, en la industria, en los medios de comunicación y en el ámbito de la investigación, el término “Minería de Datos” se utiliza a menudo para referirse a todo el proceso de descubrimiento de conocimiento, quizás por su simplicidad. Por lo tanto, adoptamos una visión amplia de la funcionalidad de Minería de Datos, y podemos decir que:

“La minería de datos es el proceso de descubrir patrones y conocimientos interesantes a partir de grandes cantidades de datos” (Han et al. 2011).

Los orígenes de los datos pueden incluir bases de datos, almacenes de datos, Web, otros repositorios de información o datos que se transmiten dinámicamente al sistema.

“La Minería de Datos es la etapa de descubrimiento en el proceso de KDD (Knowledge Discovery from Databases): paso consistente en el uso de algoritmos concretos que generan una enumeración de patrones a partir de los datos preprocesados. (G.; Smith P. et al, 2006). Para conseguirlo hace uso de diferentes tecnologías que resuelven problemas típicos de agrupamiento automático, clasificación y asociación de atributos, etc.” (Cutro, 2008)

Entonces, podemos decir que la Minería de Datos es un conjunto de Técnicas de Análisis de Datos que permiten:

- Extraer Patrones, Tendencias y Regularidades para describir y Comprender mejor los Datos.
- Extraer Patrones y Tendencias para predecir comportamientos futuros.

Y debido al gran volumen de datos, este análisis ya no puede ser manual (ni siquiera facilitado por herramientas de Almacenes de Datos y OLAP), sino que ha de ser semi-automático.

Desafío Actual de las Minería de Datos

El desafío actual es extraer conocimiento a partir de los Datos, y tratar que esto sea automático o semiautomático, con la menor participación del hombre.

Con este proceso de Análisis Automático o semiautomático de grandes cantidades de datos se podrán extraer Patrones interesantes hasta ahora desconocidos. Estos Patrones no se podrían detectar mediante la exploración tradicional de los Datos como SQL.

Uso de la Minería de Datos:

Las Técnicas y Análisis de datos en Data Mining, buscan tendencias para predecir comportamientos (casos futuros), pronósticos, riesgos y probabilidades, recomendaciones, búsqueda de secuencias, agrupación, etc.

Aplicaciones:

Algunas aplicaciones de la Minería de Datos:

- Reconocimiento del habla (aplicaciones de google).

- Reconocimiento de rostros.
- Reconocimiento de Texto.
- Detección de Fraudes.
- Diagnóstico Médico.
- Filtro de Anti-Spam.
- Sistemas de recomendación.
- Etc.

Ejemplo de Aplicaciones:

“En el ámbito web:

- Reglas de Asociación: Ejemplo, El 60% de las personas que esquían viajan frecuentemente a Europa.
- Clustering: Ejemplo, Los usuarios A y B tienen gustos parecidos (acceden a URLs similares).” (Cutro, 2008)

2.5 Tipos de datos que se pueden extraer

La Minería de Datos se puede aplicar a cualquier tipo de datos, siempre y cuando estos sean significativos para una aplicación de destino. Las formas más básicas de datos para aplicaciones de minería son datos de Bases de Datos, datos de Almacenes de Datos y datos Transaccionales. La Minería de Datos también puede aplicarse a otras formas de datos (por ejemplo, Flujos de Datos, Datos Ordenados/Secuenciales, Datos en Grafo o en Red, Datos Espaciales, Datos de Texto, Datos Multimedia y de la World Wide Web). El Tratamiento Integral se considera un tema avanzado. Sin duda, la minería de datos seguirá abarcando nuevos tipos de datos a medida que surjan. (Han et al. 2011)

2.6 Tipos de patrones que se pueden extraer.

Se han enunciado varios tipos de repositorios de datos e información en los que se puede realizar Minería de Datos. Ahora se enuncian los tipos de patrones que se pueden extraer.

Hay una serie de funcionalidades de Minería de Datos. Estos incluyen:

- Caracterización y Discriminación;
- La extracción de Patrones, Asociaciones y Correlaciones Frecuentes;
- Clasificación y regresión;
- Análisis de agrupamiento;
- Y análisis de Valores Anómalos.

Las funcionalidades de minería de datos se utilizan para especificar los tipos de patrones que se encuentran en las tareas de minería de datos. En general, estas tareas pueden clasificarse en dos categorías: Descriptiva y Predictiva.

“Las tareas Descriptivas caracterizan las propiedades de los datos en un conjunto de datos de destino. Las tareas Predictivas realizan inducción en los datos actuales con el fin de hacer predicciones.” (Han et al. 2011)

2.6.1 Clasificación y Regresión para el Análisis Predictivo

Clasificación es el proceso de encontrar un modelo (o función) que describe y distingue clases de Datos o Conceptos. El modelo se deriva basándose en el análisis de un conjunto de Datos de Entrenamiento (Objetos de Datos para los que se conocen las etiquetas de clase). El modelo se utiliza para predecir el rótulo de clase de los objetos para los cuales el rótulo de la clase es desconocido.

El Modelo Derivado puede representarse en varias formas:

- Reglas de clasificación (reglas IF-THEN).
- Árboles de decisión.
- Fórmulas matemáticas.
- Redes Neuronales.

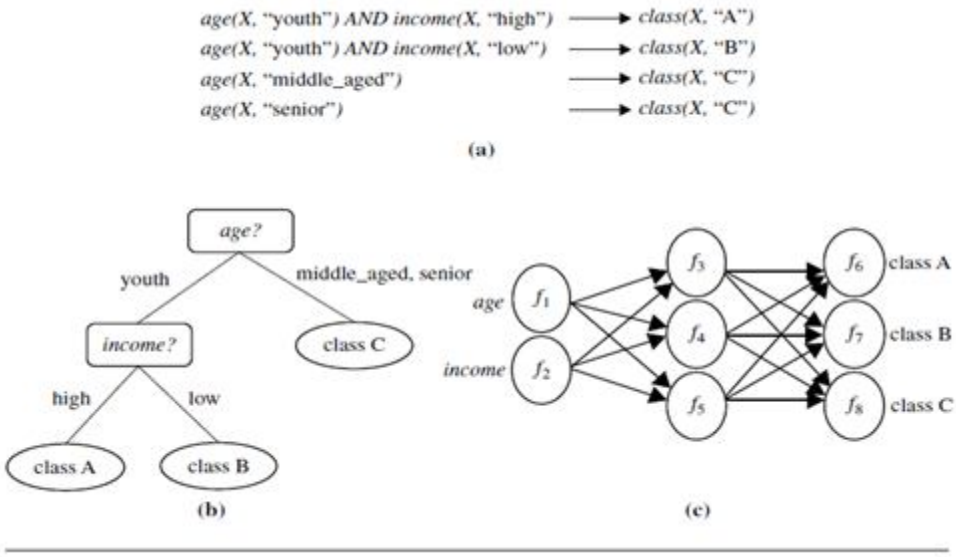


Figura 2. 2 Un modelo de clasificación se puede representar de varias formas. a) Regla If-Then, b) Árboles de Decisión, o c) Redes Neuronales. (Han et al. 2011)

Los Árboles de decisión son Diagramas de flujo, donde cada:

- Nodo: denota una prueba en un valor de atributo.
- Rama: representa un resultado de la prueba.
- Hojas de árbol: Representan clases o distribuciones de clase.

Los árboles de decisión se pueden convertir fácilmente en reglas de clasificación. Es una colección de unidades de procesamiento neuronales con conexiones ponderadas entre las unidades.

Existen muchos métodos para construir modelos de clasificación, como la Clasificación Bayesiana Naïev, las Máquinas de Vectores de Soporte y la Clasificación de k-vecino más cercano.

Mientras que la Clasificación predice Etiquetas Categóricas (discretas, desordenadas), la función de valores continuos produce Modelos de Regresión. Es decir, la Regresión se usa para predecir valores de datos numéricos faltantes o no disponibles en lugar de Etiquetas de clase (Discretas). El término predicción se refiere tanto a la predicción numérica como a la de la etiqueta de clase.

El Análisis de Regresión es una Metodología Estadística que se utiliza mayormente para la predicción numérica. También abarca la identificación de las tendencias de distribución basadas en los datos disponibles.

La Clasificación y la Regresión pueden necesitar ser precedidas por un análisis de relevancia, que intenta identificar los atributos que son relevantes para el proceso de

Clasificación y Regresión. Estos atributos serán seleccionados para este proceso. Otros atributos, que son irrelevantes, pueden ser excluidos de la consideración.

Ejemplo: Un gerente de ventas desea clasificar un gran conjunto de artículos en la tienda, en función de tres tipos de respuestas a una campaña de ventas: buena respuesta, respuesta moderada y sin respuesta. Desea derivar un modelo para cada una de estas tres clases basándose en las características descriptivas de los elementos: precio, marca, lugar realizado, tipo y categoría. La clasificación resultante debe distinguir al máximo cada clase de las demás, presentando una imagen organizada del conjunto de datos.

Si la clasificación resultante se expresa como un árbol de decisión, se puede identificar el precio como el único factor que mejor distingue las tres clases. El árbol puede revelar otras características que ayudan a distinguir aún más los objetos de cada clase incluyen la marca y el lugar. Tal árbol de decisión puede ayudarle a entender el impacto de la campaña de ventas dada y diseñar una campaña más efectiva en el futuro.

En lugar de predecir etiquetas de respuesta categóricas para cada elemento de tienda, se podría predecir la cantidad de ingresos que cada elemento generará durante una próxima venta, basado en los datos de ventas anteriores. (Han et al. 2011)

2.6.2 Análisis de Cluster (Agrupamiento).

A diferencia de la Clasificación y la Regresión, que analizan las Clases etiquetadas y Data Sets, Clustering analiza los objetos de datos sin consultar las etiquetas de clase. En muchos casos, los datos etiquetados pueden no existir al principio. La agrupación en clúster se puede usar para generar etiquetas de clase para un grupo de datos. “Los objetos se agrupan según el principio de maximizar la similitud intraclase y minimizar la similitud interclase” (Han et al. 2011). Es decir, los clusters (grupos) de objetos se forman para que los objetos dentro de un grupo tengan una gran similitud en comparación entre sí, pero son bastante diferentes a los objetos en otros grupos. Cada grupo así formado se puede ver como una clase de objetos, de los cuales se pueden derivar reglas. La agrupación también puede facilitar la formación de taxonomía, es decir, la organización de observaciones en una jerarquía de clases que agrupan eventos similares.

Ejemplo: Análisis de conglomerados. Se puede realizar en los datos de clientes de una empresa de ventas para identificar subpoblaciones homogéneas de clientes. Estos grupos pueden representar grupos específicos de marketing. La figura 2.3 muestra un diagrama en 2-D de los clientes con respecto a las ubicaciones de los clientes en una ciudad. Tres grupos de puntos de datos son evidentes.

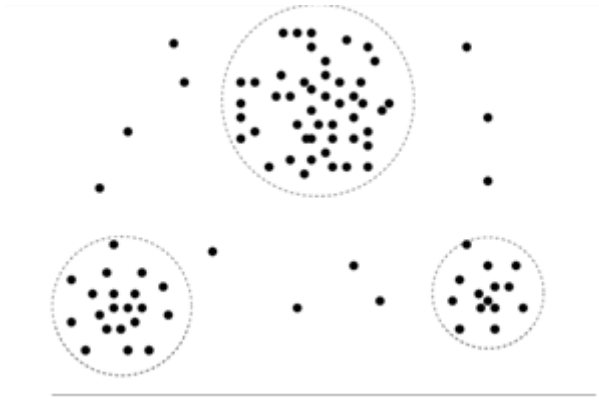


Figura 2. 3 Un diagrama en 2D de los datos del cliente con respecto a las ubicaciones de los clientes en una ciudad, que muestra tres grupos de datos. (Han et al. 2011)

2.6.3 Análisis atípico.

“Un Conjunto de Datos puede contener objetos que no cumplen con el comportamiento general; estos datos se llaman Valores Atípicos” (Han et al. 2011). Muchos de los métodos de extracción de datos descartan valores atípicos como ruido o excepciones. Sin embargo, en algunas aplicaciones, los eventos raros son más interesantes que los frecuentes. El Análisis de Datos atípicos también se denomina Análisis Atípico o Minería de Anomalías.

Los Valores Atípicos pueden detectarse utilizando pruebas estadísticas que asumen un modelo de distribución o probabilidad para los datos, o el uso de medidas de distancia donde los objetos que están alejados de cualquier otro grupo se consideran Valores Atípicos. En lugar de utilizar medidas estadísticas o de distancia, los métodos basados en Densidad pueden identificar Valores Atípicos en una región local, aunque se ven normales desde una vista de distribución estadística global.

Ejemplo: El Análisis Atípico puede revelar el uso fraudulento de las tarjetas de crédito mediante la detección de compras de montos extraordinariamente grandes para un número de cuenta dado en comparación con los cargos regulares incurridos por la misma cuenta. Los valores atípicos también pueden detectarse con respecto a las ubicaciones y los tipos de compra, o la frecuencia de compra. (Han et al. 2011)

Aquí se concluye el Marco Teórico referido a la Minería de Datos. Se presentó su importancia, historia, aplicaciones, tipos de datos que se pueden tratar y los análisis correspondientes. En el siguiente capítulo se abordará la Metodología a utilizar en el presente trabajo, la Metodología CRISP-DM.

Capítulo 3: Análisis Exploratorio de los Datos.

En el presente capítulo se aborda el enfoque metodológico. Se procede a explicar las etapas de la Metodología CRIP-DM, se desarrollan las primeras, y se exponen de esta manera las Técnicas a utilizar.

3.1 Metodología.

CRISP-DM (Cross-Industry Standard Processor - Data Mining / Proceso estándar entre industrias - Minería de Datos).

CRISP-DM es una de las Principales Metodologías empleadas por los Analistas en la Inteligencia de Negocios, y principalmente en el Proceso de Minería de Datos. Esta Metodología tiene sustento en Estándares Internacionales que reflejan la robustez de sus procesos y que facilitan la unificación de sus fases en una estructura confiable y amigable para el usuario. Además de ello, esta tecnología interrelaciona las diferentes fases del proceso entre sí, de tal manera que se consolida un proceso iterativo y recíproco. (The Modeling Agency, 1999-2000)

Se escogió esta metodología por su aprobación de mercado y su elevado nivel de estandarización, siendo esta utilizada por empresas líderes como ser ORACLE, para el desarrollo de la Minería de Datos y proyectos de descubrimiento de conocimiento. Además su flexibilidad hace posible adaptarla a proyectos de pequeña, mediana y gran envergadura, permitiendo crear un modelo de Minería de Datos que se adapte a las necesidades de cada proyecto. (Oracle, 2018)

El Ciclo de Vida de este modelo consiste en seis fases (Fig. 3.1). Se pueden ver flechas indicando las dependencias más importantes y frecuentes entre las fases.

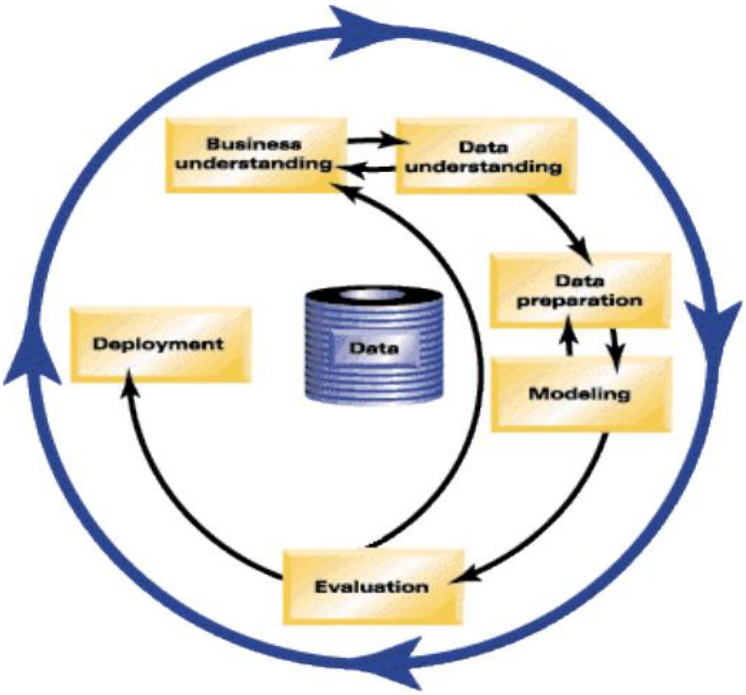


Figura 3. 1 Ciclo de Vida de CRISP-DM (IBM, 1994-2000)

Esta secuencia de fases no es estricta. De hecho, muchos proyectos pueden estar en una determinada fase pero no quita que se puede ir hacia una fase anterior o una fase posterior. Este pasaje de fases se realiza según las necesidades de cada proyecto. (IBM, 1994-2000)

3.2 Las Fases de CRISP-DM

En (The Modeling Agency, 1999-2000) y (Arancibia, J. A. G, 2010) se mencionan las fases de este proceso de Minería de Datos, de la siguiente forma:

3.2.1 Comprensión del Negocio.

Esta Fase Inicial se centra en la comprensión de los objetivos y requisitos del proyecto. Convirtiendo así, la definición del problema de Minería de Datos, en un plan preliminar diseñado para alcanzar los objetivos. Esta fase se subdivide a su vez en las siguientes categorías:

- Definición de los objetivos de negocio (inicial, objetivos de negocio y criterios de éxito del negocio).

- Evaluación de la situación (inventario de recursos, requisitos supuestos y requerimientos, riesgos y contingencias, terminología, costes y beneficios).
- Definición de los objetivos del Data Mining (objetivos y criterios de éxito).
- Realización del plan del proyecto (plan del proyecto y valoración inicial de herramientas y técnicas).

3.2.2 Comprensión de Datos.

El objetivo de esta fase es familiarizarse con los datos. Donde se comienza con una colección inicial de Datos para luego pasar a los procesos relacionados a calidad de datos. Sus subdivisiones son:

- Recolección inicial (Informe de recolección).
- Descubrimiento (Informe descriptivo de los datos).
- Exploración (Informe de exploración de los datos).
- Verificación de calidad (Informe de calidad de los datos).

3.2.3 Preparación de los Datos.

En esta instancia se deberá construir un conjunto de datos. Estas tareas incluyen selección y transformación de tablas, registros y atributos y limpiezas de datos para las herramientas de modelado. Sus subdivisiones son:

- Selección (Motivos para incluir o excluir los datos).
- Depuración (reporte de depuración).
- Estructuración (Generación de atributos y registros).
- Integración (Agrupar los datos).
- Formateo (Informe de la calidad de datos formateados).

3.2.4 Modelado.

En esta etapa del proyecto ya se tiene uno o varios modelos, que pueden ser evaluados. Sus subdivisiones son:

- Selección de la técnica de modelado (Técnica y sus supuestos).
- Generar el plan de pruebas (Plan de pruebas).
- Construcción del Modelo (Parámetros escogidos, Modelos, Descripción de los modelos).
- Evaluación del Modelo (Evaluar el Modelo, Revisión de los Parámetros elegidos).

3.2.5 Evaluación.

En esta etapa del proyecto ya se tiene uno o varios modelos, que pueden ser evaluados. Sus subdivisiones son:

- Evaluar Resultados (Valoración de los resultados respecto al éxito del negocio, Modelos aprobados).
- Proceso de revisión (Revisar el proceso).
- Determinación de los pasos siguientes (Listado de posibles acciones, técnica Modelada).

3.2.6 Despliegue (Implementación).

Esta fase depende de los requerimientos, pudiendo ser simple, como la generación de un reporte o compleja como la implementación de un proceso de explotación de información que atraviese a toda la organización. Sus subdivisiones son:

- Plan de Divulgación o Implementación (Plan de Implementación).
- Plan de Monitoreo y Mantenimiento (Plan de Monitoreo y Mantenimiento).
- Presentación del Informe Final (Informe Final, Presentación Final).
- Revisión del Proyecto (Documentación de la experiencia).

3.3 Exploración de los Datos.

Los Pasos de la Metodología anteriormente descrita, fueron implementados para el desarrollo de éste Módulo. En este Capítulo se desarrollan los pasos 1 al 3, correspondiente al Análisis Exploratorio de los Datos.

3.3.1 Fase de Comprensión del Problema.

Para el caso que trata este trabajo se entrevistaron a las siguientes personas:

- Sr. Carlos Alberto Romero. Dir. de Producción y Estadísticas.
- Prof. Susana Ogas. Empleada Administrativa.

de la “Dirección Provincial de Estadística y Censo” de la Provincia de Catamarca.

Luego de mantener entrevistas con ellos, cada profesional aportó la información necesaria sobre las “Encuestas Permanentes de Hogar” (EPH).

Definición de los objetivos de Negocio:

En base a los datos provistos, y como indica la Metodología CRISP-DM, se enuncian los siguientes objetivos del proceso de minería de datos:

Objetivo General

- Aportar y complementar los Análisis Clásicos de Estadística, con técnicas novedosas de Minería de Datos que permitan detectar características comunes entre hogares e individuos del Gran Catamarca en base a Datos de la EPH.

Objetivos Específicos

- Obtener descripciones gráficas de diferentes tipos de hogares e individuos en base a su situación socio-económica.

- Determinar agrupamientos de hogares e individuos que reúnan características similares y que a su vez difieren con respecto a elementos de otros grupos
- Determinar jerárquicamente la influencia de las diferentes variables demográficas con relación a distintas situaciones socioeconómicas.

Evaluación de la Situación:

Para comprender el problema, tenemos que tener en cuenta que “La EPH es un programa nacional de producción permanente de indicadores sociales cuyo objetivo es conocer las características socio-económicas de las población. Es realizada en forma conjunta por el Instituto Nacional de Estadística y Censo (INDEC) y las Direcciones Provinciales de Estadística (DPE)”. (INDEC 2, 2003). Las Encuestas se realizan en forma trimestral, en todo el Territorio Argentino, el cual se divide en Regiones, y estas a su vez en Conglomerados. Para el presente trabajo, se han filtrado los datos para investigar solamente los pertenecientes al Conglomerado “Gran Catamarca”, perteneciente a la Región “NOA”.

En cuanto a la demografía, es una disciplina que estudia las características de una población. Funda sus métodos en el manejo de información estadística, por tanto resulta de interés tanto a urbanistas, sociólogos, antropólogos, arquitectos y otros profesionales de las ciencias sociales, como también a estadísticos, quienes desde su formación en el manejo de datos pueden aportar a los análisis demográficos la obtención de los resultados mediante el uso de paquetes y herramientas computacionales, como lo es el Lenguaje R, aplicado en la Minería de Datos, que facilitan la comprensión de grandes cantidades de datos.

El relevamiento de la evolución de la población se lleva a cabo mediante la comparación de los datos del INDEC a través de los censos que se realizan cada diez años, como así también las Encuestas Permanentes de Hogar que se realizan en forma trimestral sobre una muestra poblacional. Estos datos sirven de fundamento para inferir sobre las dinámicas poblacionales reales.

Definición de los objetivos del Data Mining:

Mediante la Estadística tradicional, es posible obtener resultados en torno a dos variables, y expresarlos en porcentajes, tablas y gráficos. En el presente Trabajo Final se propone realizar un proceso de Minería de Datos, para así obtener nuevos patrones o comportamientos, como resultado de analizar más de dos variables en forma simultánea, mediante los métodos y algoritmos de Aprendizaje Automatizado utilizando el lenguaje de programación “R”, con la finalidad de complementar al análisis de datos de la Estadística tradicional.

Realización del plan del proyecto:

Se busca obtener patrones o comportamiento de los Datos provistos por las EPH a través de Procesos de Minería de Datos, mediante algoritmos de Aprendizaje Automatizado Supervisados (Árboles de Decisión) y No Supervisados (Clustering y Reglas de Asociación)

Los datos de las “EPH” correspondientes al año 2017 fueron compartidos por los profesionales nombrados anteriormente, en formato excel (.xlsx), lo cual garantiza la calidad de los mismos.

Para lograr los objetivos propuestos se hará uso de la Herramienta Informática “RStudio”, y el Lenguaje de Programación “R”.

3.3.2 Fase de Comprensión de los Datos

Recolección de Datos Iniciales:

Los Datos que son objeto de estudio son, las Bases de Datos de los relevamientos de las EPH del 1er Trimestre del año 2017.

“Las Bases de Datos contienen una significativa cantidad de variables de hogar y personas para posibilitar el análisis de las principales características demográficas y socioeconómicas de la población. “ (INDEC, 2016).

La Encuesta consiste en unas planillas a completar, una para cada Individuo, y otra para los datos del hogar. En ellas cada registro tiene un número de Identificación (CODUSU), que permite relacionar una vivienda con los hogares y personas que la componen. Estos resultados luego son cargados en una Base de Datos, organizados en formato tabular. Para un mejor tratamiento de los Datos con el Lenguaje “R”, las Bases de Datos fueron provistas en formato Excel (.xlsx).

“En la Base Hogar, todos los hogares que pertenecen a una misma vivienda poseen el mismo CODUSU. Para identificar los hogares se deben utilizar CODUSU y Nro_Hogar. En la de personas, todos los miembros del hogar tienen el mismo CODUSU y Nro_Hogar, pero se diferencian por el número de COMPONENTE.” (INDEC,2016)

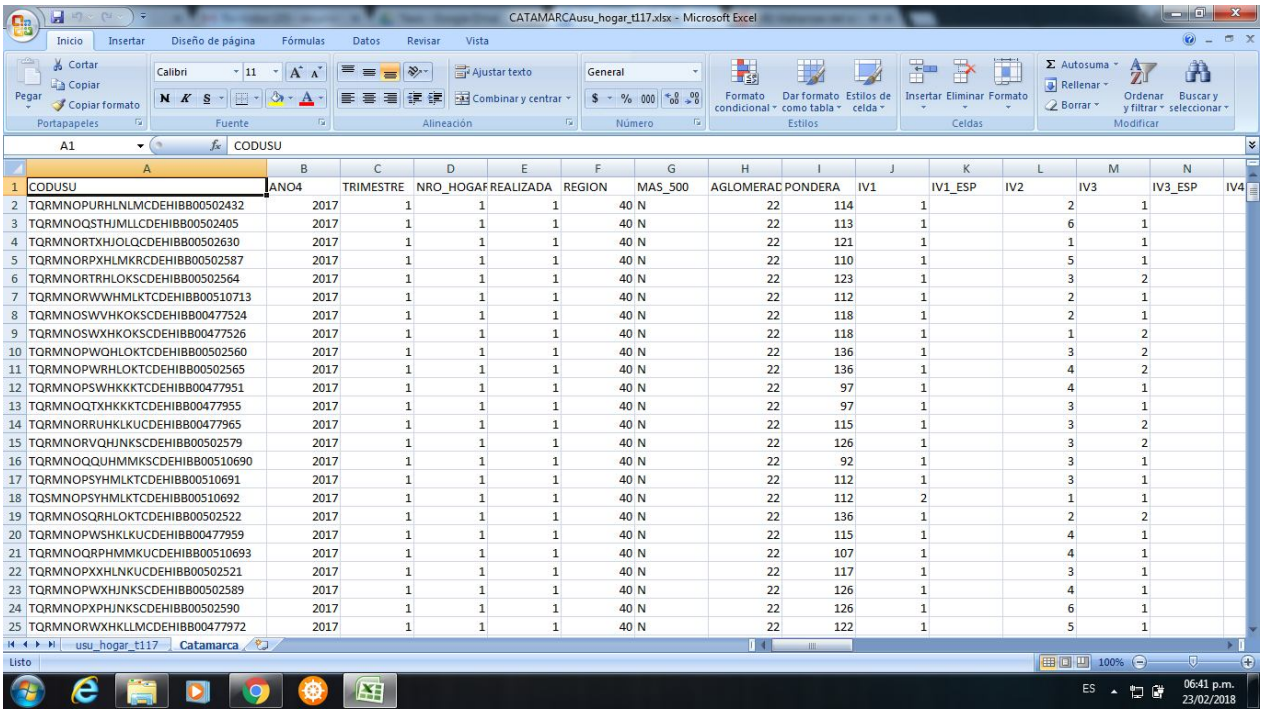
Descripción de los datos:

Para la rápida comprensión de los Datos presentes en las tablas, se describirán los grupos de datos de ambas tablas.

Base Hogar (INDEC, 2016):

- Identificación (Campos claves CODUSU, Nro_Hogar, año, trimestre, región, conglomerado, etc)
- Características de las Vivienda (Tipo de vivienda, tipo de conexión de agua, características del baño, etc.).
- Características habitacionales del Hogar (Cantidad de habitaciones, usos de las habitaciones, lavadero, garaje, etc).
- Estrategias del Hogar (Origen de los ingresos, trabajo, pensión, jubilación; beneficios del gobierno, etc).
- Resumen del Hogar (Cantidad de miembros discriminados por edad)
- Ingreso Total Familiar (Monto total familiar)
- Ingresos Per-cápita Familiar (Monto per-cápita familiar).
- Organización del Hogar (Realización de las tareas de la casa, personas que ayudan en las tareas de la casa, etc).

Esta Base tiene 88 campos, organizados por los grupos antes mencionados; y se cuenta con 526 Registros correspondientes al “Gran Catamarca”.



CODUSU	ANO4	TRIMESTRE	NRO_HOGAR	REALIZADA	REGION	MAS_500	AGLOMERAD	PONDERA	IV1	IV1_ESP	IV2	IV3	IV3_ESP	IV4
TQRMNOPURHLNLMCDEHIBB00502432	2017	1	1	1	40	N	22	114	1		2	1		
TQRMNOQSTHJMLLCDEHIBB00502405	2017	1	1	1	40	N	22	113	1		6	1		
TQRMNORTXHJOLQCDEHIBB00502630	2017	1	1	1	40	N	22	121	1		1	1		
TQRMNORPXHLMKRCDEHIBB00502587	2017	1	1	1	40	N	22	110	1		5	1		
TQRMNORTRHLOKSCDEHIBB00502564	2017	1	1	1	40	N	22	123	1		3	2		
TQRMNORWWHMLKTCDEHIBB00510713	2017	1	1	1	40	N	22	112	1		2	1		
TQRMNOSWVHKOKSCDEHIBB00477524	2017	1	1	1	40	N	22	118	1		2	1		
TQRMNOSWXHKOKSCDEHIBB00477526	2017	1	1	1	40	N	22	118	1		1	2		
TQRMNOPWQHLOKTCDEHIBB00502560	2017	1	1	1	40	N	22	136	1		3	2		
TQRMNOPWRHLOKTCDEHIBB00502565	2017	1	1	1	40	N	22	136	1		4	2		
TQRMNOPSWHKKKTCDEHIBB00477951	2017	1	1	1	40	N	22	97	1		4	1		
TQRMNOQTXHKKKTCDEHIBB00477955	2017	1	1	1	40	N	22	97	1		3	1		
TQRMNORRUHKLKUCDEHIBB00477965	2017	1	1	1	40	N	22	115	1		3	2		
TQRMNORVQHJNKSCDEHIBB00502579	2017	1	1	1	40	N	22	126	1		3	2		
TQRMNOQQUHMMKSCDEHIBB00510690	2017	1	1	1	40	N	22	92	1		3	1		
TQRMNOPSYHMLKTCDEHIBB00510691	2017	1	1	1	40	N	22	112	1		3	1		
TQSMNOPSYHMLKTCDEHIBB00510692	2017	1	1	1	40	N	22	112	2		1	1		
TQRMNOSQRHLOKTCDEHIBB00502522	2017	1	1	1	40	N	22	136	1		2	2		
TQRMNOPWSHKLKUCDEHIBB00477959	2017	1	1	1	40	N	22	115	1		4	1		
TQRMNOQRPHMMKUCDEHIBB00510693	2017	1	1	1	40	N	22	107	1		4	1		
TQRMNOPXXHLNKUCDEHIBB00502521	2017	1	1	1	40	N	22	117	1		3	1		
TQRMNOPWXHJNKSCDEHIBB00502589	2017	1	1	1	40	N	22	126	1		4	1		
TQRMNOPXPHJNKSCDEHIBB00502590	2017	1	1	1	40	N	22	126	1		6	1		
TQRMNORWXHKLLMCDEHIBB00477972	2017	1	1	1	40	N	22	122	1		5	1		

Figura 3. 2 Captura de Excel de la Tabla Hogares de la EPH

Base Personas (Individuos) (INDEC, 2016)

- Identificación (Campos claves CODUSU, Nro_Hogar, año, trimestre, región, conglomerado, etc)
- Características de los miembros del hogar (Cuestionario del hogar: Relación de parentesco, sexo, fecha de nacimiento, edad, cobertura médica, nivel educativo, lugar de nacimiento y residencia, etc).
- Ocupados que trabajaron en la semana de referencia (Donde trabajó la semana anterior a la encuesta, cantidad de ocupaciones, horas trabajadas, etc).
- Para todos los ocupados (Horas que buscó trabajar en el último mes, etc).
- Ocupación Principal (Si es estatal, privada o de otro tipo, si presta servicios domésticos, etc).
- ¿Cuánto tiempo hace que trabaja allí? (en la casa que más hs tiene; cuantas personas trabajan allí; donde realizan principalmente sus tareas, etc).
- Ocupación Principal de los trabajadores Independientes (Cuanto tiempo trabajan en ese empleo de forma continua; si el negocio/empresa/actividad tiene maquinaria/equipos, local, vehículo; Ingresos de la ocupación principal de los trabajadores independientes; Ocupación principales de los asalariados; Ingresos de la ocupación principal de los asalariados; Movimientos Interurbanos).
- Desocupados (Cuánto tiempo hace que busca trabajo, si en ese tiempo ha realizado algún trabajo, etc).

- Desocupados con empleo anterior: Última ocupación/changa (finalizada hace tres años o menos; ¿Cuánto tiempo trabajó allí?; ¿Cuánto tiempo seguido trabajó allí?).
- Ingresos de la ocupación principal (Monto de Ingreso de la ocupación Principal).
- Ingresos de otras ocupaciones (Monto de Ingresos de Otras Ocupaciones)..
- Ingreso Total Individual (Monto de Ingreso Total Individual).
- Ingresos No Laborales (Monto de Ingresos por Jubilación o Pensión, Indemnización o despido, seguro de desempleo, subsidio o ayuda social del gobierno o iglesias, por alquiler, rentas, aportes de personas que no viven en el hogar, Monto total de ingresos no laborales, etc).
- Ingreso Total Familiar (Monto del Ingreso Total Familiar).
- Ingreso Per-Cápita Familiar (Monto del Ingreso Per-Cápita Familiar).

Esta Base tiene 177 campos, organizados por los grupos antes mencionados, y cuenta con 1907 Registros.

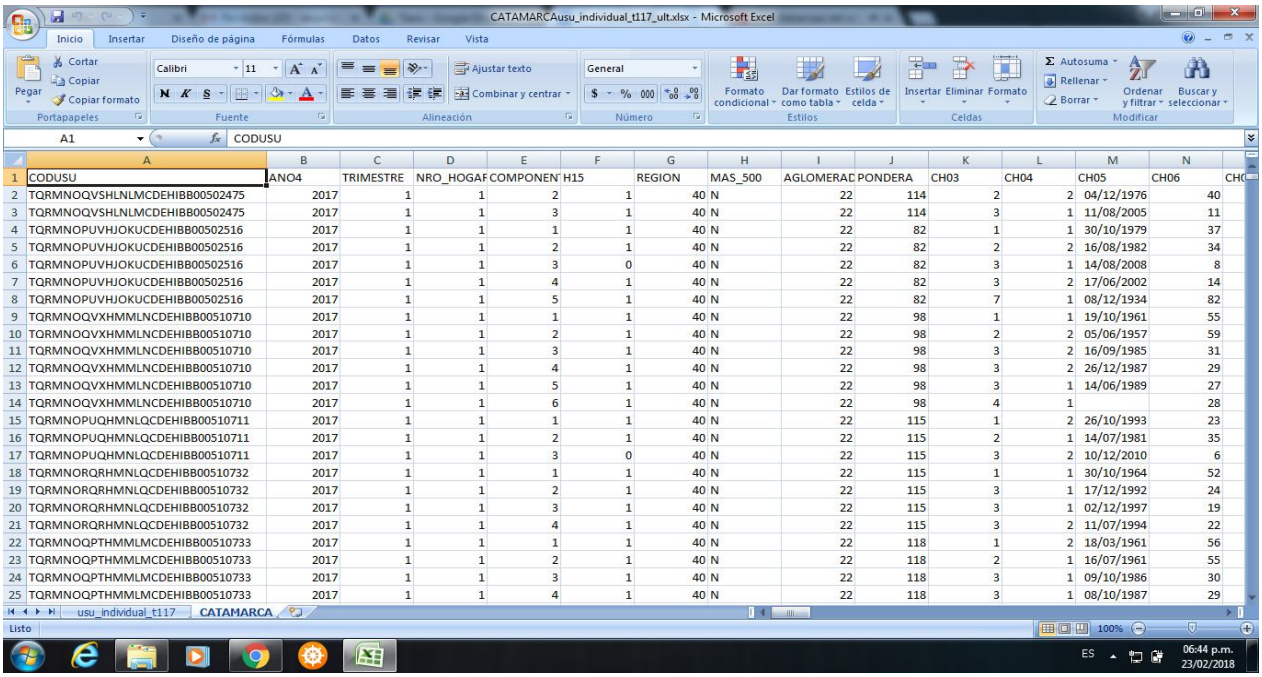


Figura 3. 3 Captura de Excel de la Tabla Individuos de la EPH

3.3.3 Fase de Preparación de los Datos.

Selección de Datos:

Los nombres originales de los archivos son: “usu_hogar_t117.xlsx” y “usu_individual_t117_ult.xlsx”. En ambas tablas, se encuentran datos de las EPH de todo el país, pertenecientes al primer Trimestre del año 2017.

Se procedió a filtrar las dos tablas, para así visualizar solo los Registros correspondientes al Aglomerado “Gran Catamarca”, de la Región “NOA”. Para esto, haciendo uso de la herramienta informática Excel, se filtró el Campo “Aglomerado”, e ingresó el valor 22, que corresponde a Nuestra Ciudad. Trabajando, de esta forma, solo con las tuplas de los Hogares e Individuos del Gran Catamarca.

Limpieza de Datos:

Los Datos brindados por los profesionales del Departamento Provincial de Estadística y Censo, fueron provistos en formato .xlsx. Para el uso de dichos datos en RStudio, se procedió a guardar el archivo en formato “.csv” (con delimitador de comas).

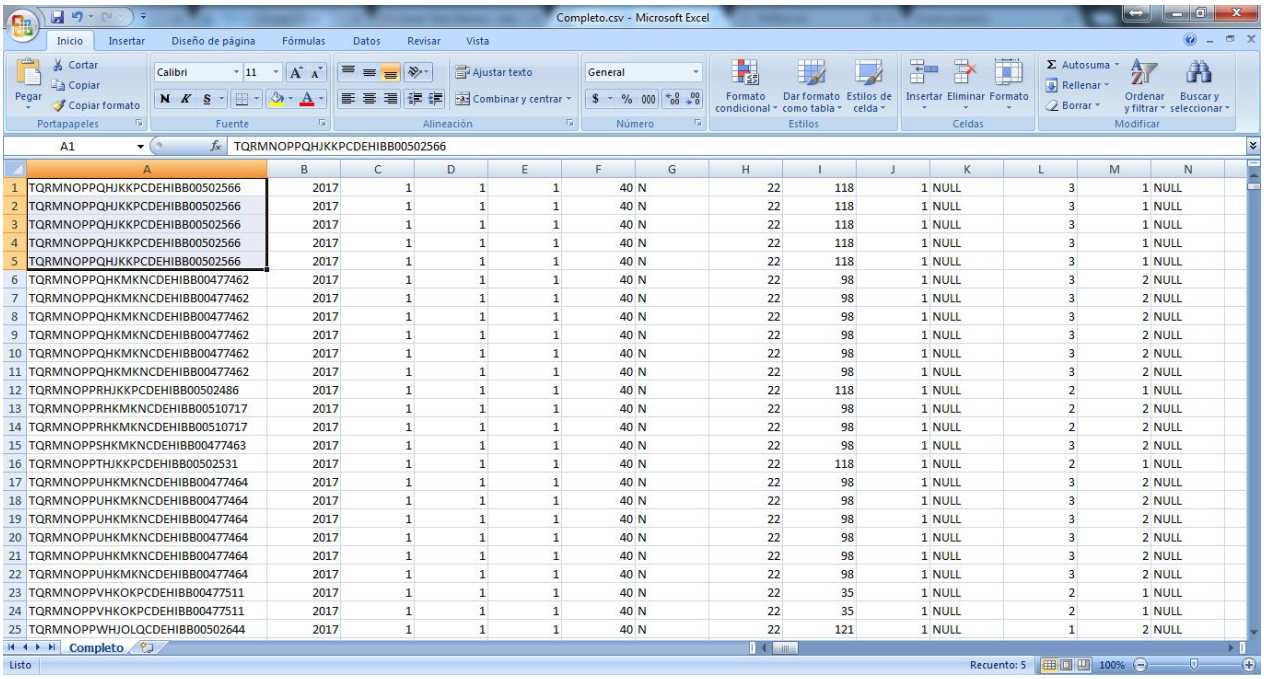
Los nombres originales de los archivos son: “usu_hogar_t117.xlsx” y “usu_individual_t117_ult.xlsx”. A los cuales, a fines de facilitar su tratamiento, luego de cambiarles el formato, se les modificó el nombre: “Hogares.csv” e “Individuos.csv”, respectivamente.

Estructuración e Integración de los Datos:

También se procedió a unificar las tablas para comparar con los resultados de las tablas individuales. Para esto se utilizó el motor de base de datos “SQLServer” y se exportaron los archivos Hogares (H) e Individuos (I), en formato “.xsls”. Creadas las tablas y exportados los datos, se procedió a hacer la siguiente sentencia:

```
SELECT H.*, I.*
  FROM [dbo].[Hogares] as H, [dbo].[Individuos] as I
where H.CODUSU = I.CODUSU
```

Como resultado se obtuvo una tabla con todas las filas y columnas de ambas tablas. De esta manera, ordenando las filas por la clave CODUSU, se pueden agrupar los registros de cada hogar con los individuos que lo habitan.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	TQRMNOPPQHJKKPCDEHIBB00502566	2017	1	1	1	40	N	22	118	1	NULL	3	1	NULL
2	TQRMNOPPQHJKKPCDEHIBB00502566	2017	1	1	1	40	N	22	118	1	NULL	3	1	NULL
3	TQRMNOPPQHJKKPCDEHIBB00502566	2017	1	1	1	40	N	22	118	1	NULL	3	1	NULL
4	TQRMNOPPQHJKKPCDEHIBB00502566	2017	1	1	1	40	N	22	118	1	NULL	3	1	NULL
5	TQRMNOPPQHJKKPCDEHIBB00502566	2017	1	1	1	40	N	22	118	1	NULL	3	1	NULL
6	TQRMNOPPQHKMKNCDEHIBB00477462	2017	1	1	1	40	N	22	98	1	NULL	3	2	NULL
7	TQRMNOPPQHKMKNCDEHIBB00477462	2017	1	1	1	40	N	22	98	1	NULL	3	2	NULL
8	TQRMNOPPQHKMKNCDEHIBB00477462	2017	1	1	1	40	N	22	98	1	NULL	3	2	NULL
9	TQRMNOPPQHKMKNCDEHIBB00477462	2017	1	1	1	40	N	22	98	1	NULL	3	2	NULL
10	TQRMNOPPQHKMKNCDEHIBB00477462	2017	1	1	1	40	N	22	98	1	NULL	3	2	NULL
11	TQRMNOPPQHKMKNCDEHIBB00477462	2017	1	1	1	40	N	22	98	1	NULL	3	2	NULL
12	TQRMNOPPRHJKKPCDEHIBB00502486	2017	1	1	1	40	N	22	118	1	NULL	2	1	NULL
13	TQRMNOPPRHKMKNCDEHIBB00510717	2017	1	1	1	40	N	22	98	1	NULL	2	2	NULL
14	TQRMNOPPRHKMKNCDEHIBB00510717	2017	1	1	1	40	N	22	98	1	NULL	2	2	NULL
15	TQRMNOPPSHKMKNCDEHIBB00477463	2017	1	1	1	40	N	22	98	1	NULL	3	2	NULL
16	TQRMNOPPTHJKKPCDEHIBB00502531	2017	1	1	1	40	N	22	118	1	NULL	2	1	NULL
17	TQRMNOPPUHKMKNCDEHIBB00477464	2017	1	1	1	40	N	22	98	1	NULL	3	2	NULL
18	TQRMNOPPUHKMKNCDEHIBB00477464	2017	1	1	1	40	N	22	98	1	NULL	3	2	NULL
19	TQRMNOPPUHKMKNCDEHIBB00477464	2017	1	1	1	40	N	22	98	1	NULL	3	2	NULL
20	TQRMNOPPUHKMKNCDEHIBB00477464	2017	1	1	1	40	N	22	98	1	NULL	3	2	NULL
21	TQRMNOPPUHKMKNCDEHIBB00477464	2017	1	1	1	40	N	22	98	1	NULL	3	2	NULL
22	TQRMNOPPUHKMKNCDEHIBB00477464	2017	1	1	1	40	N	22	98	1	NULL	3	2	NULL
23	TQRMNOPPVHKOKPCDEHIBB00477511	2017	1	1	1	40	N	22	35	1	NULL	2	1	NULL
24	TQRMNOPPVHKOKPCDEHIBB00477511	2017	1	1	1	40	N	22	35	1	NULL	2	1	NULL
25	TQRMNOPPWHJOLQCDEHIBB00502644	2017	1	1	1	40	N	22	121	1	NULL	1	2	NULL

Figura 3. 4 Captura de Excel de la Tabla unificada EPH

Formateo de los Datos:

En cuanto al formateo de los datos, se procedió a cambiar los nombres de las bases de datos para facilitar el trabajo. Como anteriormente se enunció, los nombres originales de los archivos son: “usu_hogar_t117.xlsx” y “usu_individual_t117_ult.xlsx”. A los cuales, luego de cambiarles el formato, se les modificó el nombre a “Hogares.csv” e “Individuos.csv”, respectivamente.

Posteriormente, para el tratamiento de unificar las tablas, mediante SQL, a la Tabla Hogares se le llamó “H”, y a la tabla Individuos “I”.

3.3.4 Fase de Modelado

3.3.4.1 Selección de Herramientas

Sobre R

R es un lenguaje y entorno de programación, creado en 1993 por Ross Ihaka y Robert Gentleman del Departamento de Estadística de la Universidad de Auckland, cuya

característica principal es formar un entorno de análisis estadístico para la manipulación de datos, su cálculo y la creación de gráficos. En su aspecto, R puede considerarse como otra implementación del lenguaje de programación S (creado por los Laboratorios AT&T Bell), con la particularidad de que es un software GNU, General Public Licence (conjunto de programas desarrollados por la Free Software Foundation), es decir, de uso libre.

“Una de las fortalezas de este lenguaje es su capacidad multiplataforma, debido a que su código puede ser compilado y ejecutado en una gran variedad de plataformas (Unix o sistemas similares incluyendo FreeBSD y Linux, también en Windows y MascOS)” (Martinez, 2016). “R está disponible como código fuente (el cual está escrito principalmente en C y algunas rutinas en Fortran), esencialmente para máquinas Unix y Linux, o como archivos binarios pre-compilados para Windows, Linux, Macintosh, y Alpha Unix. Los archivos necesarios y las instrucciones para la instalación, se distribuyen desde el sitio de internet de CRAN (Comprehensive R Archive Network).” (Ahumada, 2003).

R ofrece una variedad de funciones para el Análisis Estadístico y Gráfico. Entre las técnicas estadísticas encontramos: Modelado Lineal y No Lineal, Test Estadísticos Clásicos, Análisis de series de tiempo, Clasificación, Clustering, entre otras. Los gráficos son visualizados de forma inmediata y pueden ser guardados en distintos formatos (jpg, png, bmp, ps, pdf, emf, pictex, xfig), dependiendo del sistema operativo.

El lenguaje permite al usuario programar bucles para analizar conjuntos de datos, como así también combinar en un solo programa diferentes funciones estadísticas para realizar análisis complejos. También es posible utilizar directamente en R, algunos programas de S.

“Una característica es su gran **flexibilidad**. R guarda los resultados como “objetos”, de tal forma que se puede hacer un análisis sin necesidad de mostrar su resultado inmediatamente. El usuario puede, si lo desea, extraer solo aquellos resultados que le interesan. Por ejemplo, si uno corre una serie de 20 regresiones y quiere comparar los coeficientes de regresión, R puede mostrar solamente los coeficientes estimados: de esta manera los resultados se pueden resumir en una sola línea, mientras que un programa clásico podría abrir 20 ventanas de resultados.” (Ahumada, 2003)

Como funciona R

R es un lenguaje Orientado a Objetos. “Es decir, que las variables, datos, funciones, resultados, etc., se guardan en la memoria activa del computador en forma de objetos con un nombre específico. El usuario puede modificar o manipular estos objetos con operadores (aritméticos, lógicos y comparativos) y funciones (que a su vez son objetos).” (Ahumada, 2003)

Es un lenguaje interpretado (como Java), y no compilado (como C, C++, Fortran, Pascal, etc), lo que permite que los comandos escritos sean ejecutados directamente sin necesidad de construir ejecutables.

Su sintaxis es simple e intuitiva. Para que una función sea ejecutada, debe ir siempre acompañada de paréntesis. Si se escribe sin paréntesis, se mostrará el código de la función.

Las funciones disponibles están guardadas en una librería localizada en el directorio R_HOME/library (R_HOME es el directorio donde R está instalado). Este directorio contiene paquetes de funciones, las cuales a su vez están estructuradas en directorios. El paquete denominado “Base” constituye el núcleo de R y contiene las funciones básicas del lenguaje para leer y manipular datos, algunas funciones gráficas y algunas funciones estadísticas (regresión lineal y análisis de varianza). Cada paquete contiene un directorio denominado R con un archivo con el mismo nombre del paquete (por ejemplo, para el paquete “base”, existe el archivo R_HOME/library/base/R/base), Este archivo está en formato ASCII y contiene todas las funciones del paquete. (Ahumada, 2003).

R-Studio

R-Studio es un IDE (Entorno de Desarrollo Integrado) gratuito para R. Está disponible en dos ediciones: R-Studio para escritorio, donde el IDE corre de manera local en una PC y R-Studio Server, el cual permite acceder a R-Studio por medio de un navegador web mientras es ejecutado de forma remota en un servidor Linux (Martínez, 2016). Incluye una consola, editor de resaltado de sintaxis que admite la ejecución directa del código, así como herramientas para trazado, historial, depuración de trabajo.

RStudio está disponible en ediciones comerciales y de código abierto y se ejecuta en el escritorio (Windows, Mac y Linux) o en un navegador conectado a RStudio Server o RStudio Server Pro (Debian / Ubuntu / RedHat / CentOS y SUSE Linux). (https://www.rstudio.com/products/rstudio/) (RStudio, 2018).

Fue escrito en el lenguaje de programación C++ y usa el framework Qt para su interfaz gráfica. Además, puede ser corrido en la mayoría de las plataformas de escritorio (Windows, Mac OS X y Linux). (Martínez, 2016)

3.3.4.2 Selección de la Técnica de Modelado:

3.3.4.2.1 ÁRBOLES DE DECISIÓN

Un árbol de decisión construye un modelo en forma de estructura de árbol, que comprende una serie de decisiones lógicas, similares a un diagrama de flujo, con nodos de decisión que indican una decisión sobre un atributo. Estos se dividen en ramas que indican el flujo de las decisiones. El árbol termina con nodos de hoja (nodos terminales) que denotan el resultado de seguir una combinación de decisiones. (Lantz, B, 2013)

“Tienen la función de clasificar los datos y predecir el comportamiento de manera estadística, a partir de construir diagramas lógicos que categorizan y representan una serie de condiciones de forma sucesiva.” (Torres, D. L., et.al, 2011)

Desde otro punto de vista, los Árboles de Decisión aproximan funciones discretas, y lo hacen de la siguiente forma:

- Clasifica las instancias yendo desde la raíz hacia las hojas.
- En cada nodo se testea un atributo y se baja por la rama asociada al valor de la instancia.
- El proceso se repite hasta llegar a una hoja, en donde está el resultado.

“Un Árbol de Decisión es un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas.” (Torres, D. L., et.al, 2011). Las instancias se representan como un conjunto de parejas atributo-valor. La función objetivo toma valores discretos y las descripciones requieren disyunciones. El conjunto de entrenamiento puede contener errores y las instancias de entrenamiento pueden no tener todos los atributos.

“Los Árboles de Decisión se construyen usando un enfoque que generalmente se conoce como “Divide y Vencerás”, porque utiliza los valores de las características para dividir los datos en subconjuntos cada vez más pequeños de clases similares.” (Lantz, B, 2013)

“Los Árboles de decisión son un método usado en distintas disciplinas como modelo de predicción. Estos son similares a diagramas de flujo, en lo que llegamos a puntos en los que se toman decisiones de acuerdo a una regla.” (Juan Bosco Mendoza Vega, 2018)

En el campo del Aprendizaje Automatizado, hay distintas maneras de obtener árboles de decisión. Para ello, “tenemos una variable objetivo (dependiente) y nuestra meta es obtener una función que nos permita predecir, a partir de variables predictoras (independientes), el valor de la variable objetivo para casos desconocidos”. (Juan Bosco Mendoza Vega, 2018)

“Señalan (Hernández, Ramírez & Ferrari, 2004), una de las grandes ventajas de los árboles de decisión es que, en su forma más general, las operaciones posibles a partir de una determinada condición son excluyentes.” (Torres, D. L., et.al, 2011)

“Las principales ventajas de este método son su interpretabilidad, pues nos da un conjunto de reglas a partir de las cuales se pueden tomar decisiones. Este es un algoritmo que no es demandante en poder de cómputo comparado con procedimientos más sofisticados y, a pesar de ello, que tiende a dar buenos resultados de predicción para muchos tipos de datos.” (Juan Bosco Mendoza Vega, 2018)

“Sus principales desventajas son que en este tipo de clasificación es “débil”, pues sus resultados pueden variar mucho dependiendo de la muestra de datos usados para entrenar el modelo. Además es fácil sobre ajustar los modelos, esto es, hacerlos excelentes para clasificar datos que conocemos, pero deficientes para datos conocidos.” (Juan Bosco Mendoza Vega, 2018)

Aplicaciones

Un Árbol de decisión es prácticamente un diagrama de flujo, y se utiliza especialmente en situaciones donde el mecanismo de clasificación debe ser lo más trasparente posible por razones legales, o porque los resultados deben compartirse para facilitar la toma de decisiones. (Yu-Wei, C. D. C. ,2015)

Algunas aplicaciones prácticas son:

- Detección de fraudes con tarjetas de crédito.
- Modelos de clasificación crediticia, donde los criterios para que un cliente sea rechazado o no, deben estar bien fundamentados.
- Toma de decisiones médicas.
- Estudios de Marketing sobre la satisfacción o el abandono de clientes que se comparten con las agencias de gestión.

Árboles de Decisión es una de las técnicas más utilizadas del Aprendizaje Automatizado, y se puede aplicar para modelar casi cualquier tipo de datos.

Pero a pesar de las amplias aplicaciones, hay situaciones donde los Árboles no son el modelo ideal. Por ejemplo, una tarea donde los datos tienen una gran cantidad de características nominales con muchos niveles, o si los datos tienen una gran cantidad de características numéricas. Estos casos pueden generar un árbol demasiado complejo, con una gran cantidad de decisiones.

Formas de Inferir el Árbol

- Trivial: Se debe crear una ruta del Árbol por cada instancia de entrenamiento. Pero son Árboles excesivamente grandes, y no funcionan bien con instancias nuevas.
- Óptimo: Es el Árbol más pequeño posible compatible con todas las instancias. Como desventaja es Inviable computacionalmente.

- Pseudo-Optimo (Heurístico): La selección del atributo en cada nivel del árbol es en función de la calidad de la división que produce. Los principales programas de generación de árboles utilizan procedimientos similares (ID3, C4.5, CART, etc).

Entropía

Lo primero que debe definir un árbol de decisión es determinar qué características dividir. Para ello, se buscan valores de estas características que dividan los datos, de manera que las particiones contengan ejemplos de una sola clase (de ser así se consideran puros). Hay muchas formas de medir la pureza para identificar los criterios de división, uno de ellas es la Entropía,

La Entropía mide el nivel de incertidumbre, desorden o impureza que hay en un conjunto de datos.

“La Entropía de una muestra de datos indica que tan mezclados están los valores de clase; el valor mínimo de 0 indica que la muestra es completamente homogénea, mientras que 1 indica la cantidad máxima de desorden. La definición de Entropía está especificada por:

$$Entropia(S) = \sum_{i=1}^{c} - p_i \ log_2 \ (p_i)$$

En la fórmula de la Entropía, para un segmento de datos (S), el término “c” se refiere al número de diferentes niveles de clase, y p_i se refiere a la proporción de valores que caen en el nivel de clase i” (Lantz, B., 2013).

Ejemplo: Suponiendo que tenemos una partición de datos con dos clases: Trabajador con Relación de Dependencia (60%) y Monotributista (40%). Calculamos la Entropía:

```
> -0.60*log2(0.60) – 0.40*log2(0.40)
```

Da como resultado

```
[1] 0.9709506
```

Usando la función curve() es posible obtener la gráfica de la Entropía:

```
> curve(-x*log2(x)-(1-x)*log2(1-x), col=”blue”, xlab = “x”, ylab = “Entropia”, lwd = 4)
```

Y se obtiene:

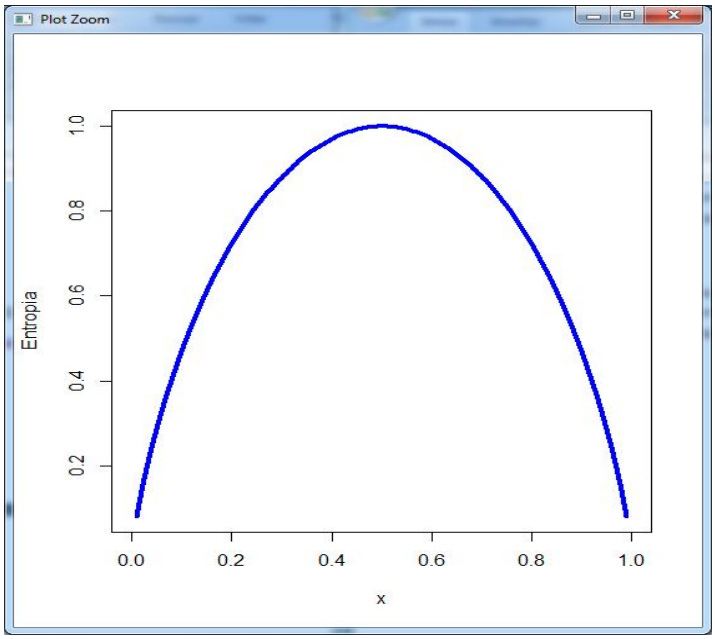


Figura 3. 5 Gráfico de Entropía realizado en Rstudio con el Lenguaje R.

Luego de obtener la medida de pureza, se debe decidir que característica dividir. Para ello, se usa la Entropía para calcular el cambio en la homogeneidad resultante de una división en cada característica posible. El cálculo se conoce como ganancia de información.

La ganancia de información para una característica F se calcula como diferencia entre la Entropía en el segmento antes de la división (S1) y las particiones resultantes de la división (S2):

$$Ganancia\ (F) = Entropía\ (S1) - Entropía\ (S2)$$ (Lantz, B., 2013)

El único problema es que luego de una división, los datos se dividen en más de una partición. Entonces, es necesario que la función Entropía(S2) considere la Entropía total en todas las particiones. Esto se puede establecer como la siguiente fórmula:

$$Entropia(S) = \sum_{i=1}^{n} - w_i\ Entropia\ (P_i)$$

“La Entropía total resultante de una división es la suma de la Entropía de cada una de las n particiones ponderadas por la proporción de ejemplos que caen en esa partición (w_i).” (Lantz, B., 2013)

La Entropía resultante de una división es la suma de la Entropía de cada una de las n-particiones ponderadas por la proporción de ejemplos que hay en esa partición (w_i).

En cuanto a la Ganancia de la Información:

- Cuanto mayor sea, mejor será la función para crear grupos homogéneos después de la división de esa función.
- Si es cero, no hay reducción en la Entropía para dividir esa característica.

- Es Máxima, si es igual a la Entropía anterior a la división. Lo cual implica que la Entropía después de la división sea cero, es decir que la decisión da como resultado grupos completamente homogéneos.

La Ganancia de Información no es el único criterio de división que se puede utilizar para construir árboles de decisión. También están el Índice de Gini, la Estadística Chi-Cuadrado y la relación de Ganancia.

Algoritmos de Construcción de Árboles de Decisión

Algunos de los algoritmos que utilizan para la creación de los Árboles de Decisión son:

- ID3: Es el más básico, y construye los árboles de manera recursiva, desde la raíz hacia las hojas, seleccionando cada momento el mejor nodo para poner en el árbol.
- C4.5 (o J48): Este algoritmo trata con valores continuos y usa criterios estadísticos para impedir que el árbol se sobreadapte (es decir, que crezca demasiado, que se aprenda los datos en lugar de generalizar).

Algoritmo ID3

Pasos a seguir:

1- Detener la construcción del árbol si:
   1) Todos los ejemplos pertenecen a la misma clase.
   2) Si no quedan ejemplos o atributos.

2- Si no, elegir el mejor atributo para poner en ese nodo (el que minimice la entropía media).
3- Crear de manera recursiva tantos sub-árboles como posibles valores tenga el atributo seleccionado.

ID3 – Pseudo-Código

Id3(Ejemplos, Atributo-objetivo, Atributos)

Si todos los ejemplos son positivos devolver un nodo positivo.

Si todos los ejemplos son negativos devolver un nodo negativo.

Si Atributos está vacío devolver el voto mayoritario del valor del atributo objetivo en Ejemplos.

En otro caso

Sea A Atributo el MEJOR de atributos

Para cada v valor del atributo hacer

Sea Ejemplos(v) el subconjunto de ejemplos cuyo valor de atributo A es v

Si Ejemplos(v) esta vacío devolver un nodo con el voto mayoritario del Atributo objetivo de Ejemplos

Sino Devolver **Id3**(Ejemplos(v), Atributo-objetivo, Atributos/{A})

Algoritmo C4.5

Pasos a seguir:

1- Detener la construcción del árbol si:
   1) Todos los ejemplos pertenecen a la misma clase.
   2) Si no quedan ejemplos o atributos.
   3) Si no se espera que se produzcan mejoras continuando la subdivisión.

2- Si no, elegir el mejor atributo para poner en ese nodo (el que minimice la Entropía media).
3- Crear de manera recursiva tantos sub-árboles como posibles valores tenga el atributo seleccionado.

3.3.4.2.2 CLUSTERING:

Según (Jain, A. K., et al; 1999), el “Clustering es una Clasificación No Supervisada de Patrones (observaciones, elementos de Datos o Vectores de Características) en grupos (clusters).”

El Clustering ha sido abordado en muchos contextos y por investigadores en muchas disciplinas; lo cual refleja su amplio atractivo y utilidad como uno de los pasos en el Análisis de Datos Exploratorios y se pueden estudiar los métodos de agrupamiento de patrones desde una perspectiva de reconocimiento estadístico de patrones.

Algunas aplicaciones importantes de los algoritmos de agrupamiento son la segmentación de imágenes, el reconocimiento de objetos y la recuperación de información.

Sobre Clustering

El análisis de datos subyace a muchas aplicaciones informáticas, ya sea en una fase de diseño o como parte de sus operaciones en línea. Los procedimientos de análisis de datos pueden ser dicotomizados como exploratorios o confirmatorios, basados en la disponibilidad de modelos apropiados para la fuente de datos, pero un elemento clave en ambos tipos de procedimientos es la agrupación o clasificación de mediciones. “El análisis de clusters es la organización de una colección de patrones (generalmente representados como un vector de medidas, o un punto en un espacio multidimensional) en grupos basados en la similitud.” (Jain, A. K., et al; 1999)

Los patrones dentro de un clúster son más similares entre sí que lo que son para un patrón perteneciente a un clúster diferente. Un ejemplo de agrupamiento se muestra en la Figura 3.6. Los patrones de entrada se muestran en la Figura 3.6 (a), y los grupos deseados se muestran en la Figura 3.6 (b). Aquí, los puntos que pertenecen al mismo clúster reciben la misma etiqueta. (Jain, A. K., et al; 1999)

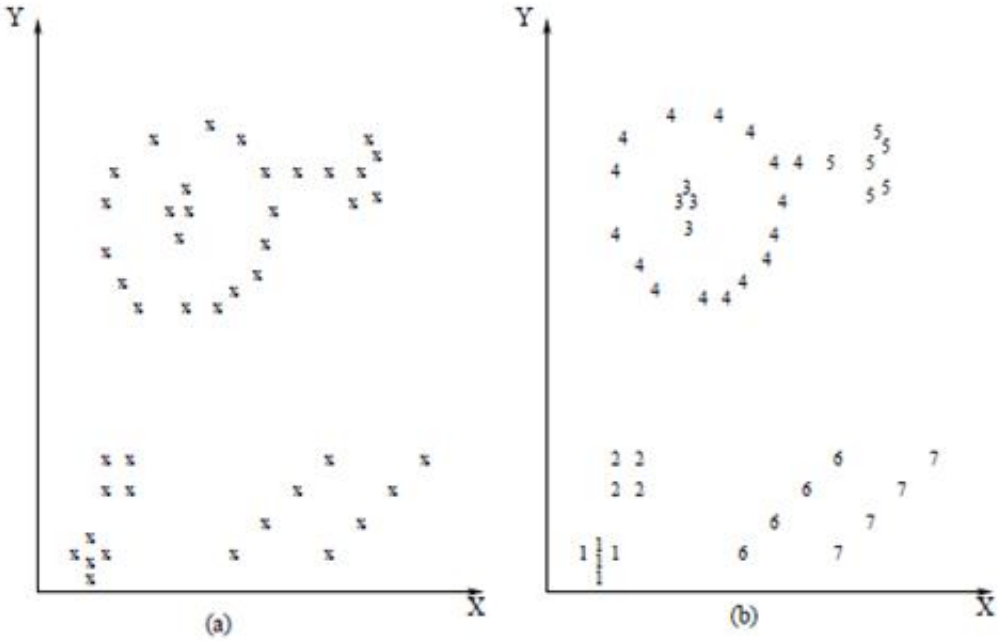


Figura 3. 6 Data Clustering (Jain, A. K., et al; 1999)

La existencia de numerosas técnicas para representar datos, medir la proximidad (similitud) entre elementos de datos y agrupar elementos de datos, ha producido una gran variedad de métodos de agrupamiento.

El Clustering es útil en varias situaciones exploratorias de Análisis de patrones, Agrupamiento, Toma de decisiones y Aprendizaje Automatizado, que incluyen Minería de Datos, Recuperación de documentos, Segmentación de imágenes y Clasificación de

patrones. Sin embargo, en muchos de estos problemas, hay poca información previa sobre los datos (por ejemplo, modelos estadísticos), y el responsable de la toma de decisiones debe hacer suposiciones sobre los datos. Es bajo estas restricciones que la Metodología de Clustering es particularmente apropiada para la Exploración de interrelaciones entre los puntos de datos para hacer una evaluación (quizás preliminar) de su estructura.

Componentes de una tarea de agrupamiento

La actividad típica de agrupamiento de patrones implica los siguientes pasos (Jain y Dubes; 1988):

(1) Representación de patrones (que incluye opcionalmente extracción y/o selección de características),

(2) Definición de una Medida de Proximidad de patrones apropiada para el dominio de datos,

(3) Agrupamiento o Agrupación,

(4) Abstracción de datos (si es necesario), y

(5) Evaluación del producto (si es necesario).

La Figura 3.7 muestra una secuencia típica de los primeros tres de estos pasos, incluida una ruta de retroalimentación donde la salida del proceso de agrupamiento podría afectar la extracción de características posteriores y los cálculos de similitud.

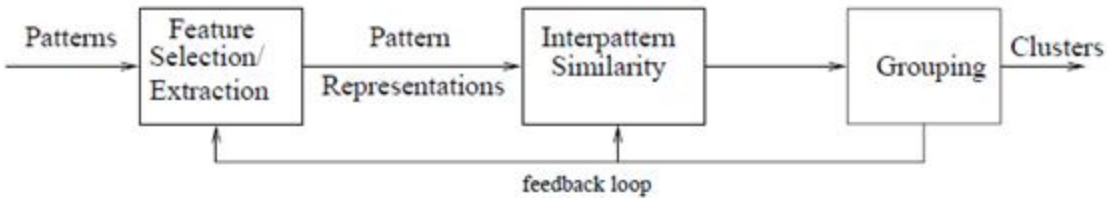


Figura 3. 7 Etapas del Clustering (Jain, A. K., et al; 1999)

La **Representación de patrones** se refiere a la cantidad de clases, el número de patrones disponibles y el número, tipo y escala de las características disponibles para el algoritmo de agrupamiento. La selección de características es el proceso de identificación del subconjunto más efectivo de las características originales para usar en la agrupación. La extracción de características es el uso de una o más transformaciones de las características de entrada para producir nuevas características destacadas. Cualquiera de estas técnicas o ambas se pueden usar para obtener un conjunto apropiado de características para usar en clustering.

La **Proximidad** del patrón generalmente se mide mediante una función de distancia definida en pares de patrones. La Distancia Euclidiana a menudo se puede utilizar para reflejar la disimilitud entre dos patrones, mientras que otras medidas de similitud se pueden utilizar para caracterizar la similitud conceptual entre patrones (Michalski & Stepp; 1983).

El paso de **Agrupación** se puede realizar de varias maneras. Los algoritmos de agrupamiento jerárquico producen una serie anidada de particiones basadas en un criterio para fusionar o dividir clústeres según la similitud. Los algoritmos de agrupamiento particional identifican la partición que optimiza (generalmente localmente) un criterio de agrupamiento. Las técnicas adicionales para la operación de agrupamiento incluyen métodos de agrupamiento probabilísticos (Brailovski, 1991) y teórico de gráficos (Zahn, 1971).

La **Abstracción** de datos es el proceso de extracción de una representación simple y compacta de un conjunto de datos. En el contexto de agrupamiento, una abstracción de datos típica es una descripción compacta de cada grupo, generalmente en términos de prototipos de clúster o patrones representativos como el centroide (Diday & Simon, 1976).

Para **Evaluar** el resultado de un algoritmo de agrupamiento, y caracterizarlo como un resultado 'bueno' o 'malo', los algoritmos cuando se presentan con datos, producen clústeres, independientemente de si los datos contienen o no clusters. Si los datos contienen clústeres, algunos algoritmos de agrupamiento pueden obtener clústeres "mejores" que otros.

No existe una técnica de agrupación que sea universalmente aplicable para descubrir la variedad de estructuras presentes en conjuntos de datos multidimensionales. Por ejemplo, considere el conjunto de datos bidimensionales que se muestra en la Figura 3.6 (a). No todas las técnicas de agrupamiento pueden descubrir todos los clusters presentes aquí con la misma facilidad, porque los algoritmos de clúster a menudo contienen suposiciones implícitas sobre la forma del clúster o configuraciones de clúster múltiple basadas en las medidas de similitud y los criterios de agrupación utilizados.

TÉCNICAS DE AGRUPAMIENTO (CLUSTERING)

Se pueden describir diferentes enfoques a la agrupación de datos con la ayuda de la jerarquía que se muestra en la Figura 3.8 (Jain, A. K., et al; 1999):

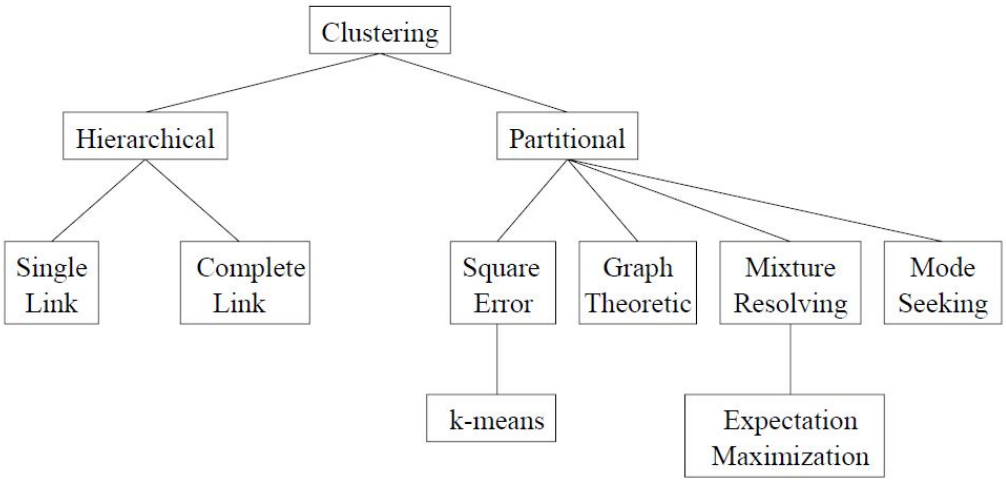


Figure 7. A taxonomy of clustering approaches.

Figura 3. 8 Taxonomía de Enfoques de Clustering (Jain, A. K., et al; 1999)

En el nivel superior, hay una distinción entre los enfoques: Jerárquico y Particional. Los métodos jerárquicos producen una serie anidada de particiones, mientras que los métodos particionales producen solo uno.

1) Algoritmos de agrupamiento jerárquico

El funcionamiento de un algoritmo de agrupamiento jerárquico se ilustra utilizando el conjunto de datos bidimensionales de la Figura 3.9.

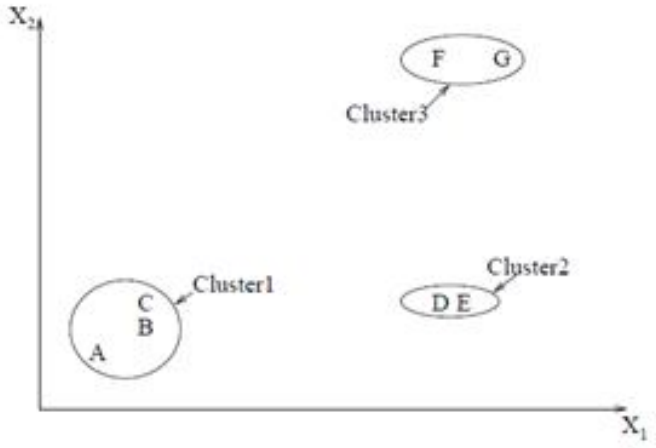


Figura 3. 9 Puntos que caen en Tres Clusters (Jain, A. K., et al; 1999)

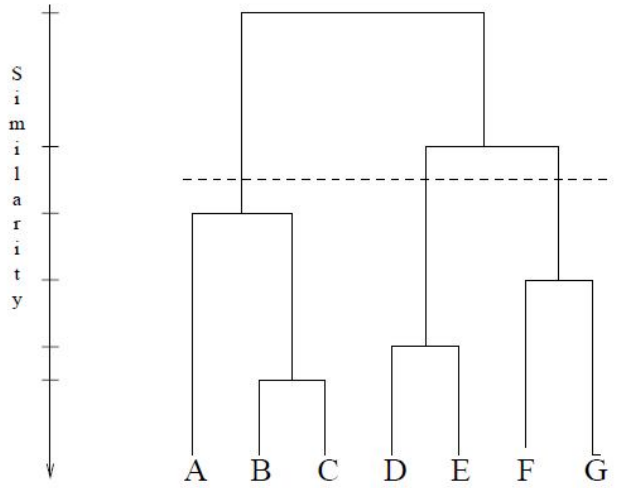


Figure 10. The dendrogram obtained using the single-link algorithm.

Figura 3. 10 Dendograma obtenido usando el Algoritmo de Enlace Simple. (Jain, A. K., et al; 1999)

Esta figura muestra siete patrones etiquetados como A, B, C, D, E, F y G en tres grupos. Un algoritmo jerárquico produce un dendrograma que representa la agrupación anidada de patrones y niveles de similitud a los que cambian las agrupaciones. En la Figura 3.10 se muestra un dendrograma correspondiente a los siete puntos de la Figura 3.9 (Jain & Dubes 1988). El dendograma se puede dividir en diferentes niveles para producir agrupamientos diferentes de los datos.

Algoritmo de Clustering de Agrupamiento Jerárquico

(1) Calcule la matriz de proximidad que contiene la distancia entre cada par de patrones. Trate cada patrón como un grupo.

(2) Encuentre el par de conglomerados más similar utilizando la matriz de proximidad. Combina estos dos grupos en un grupo. Actualice la matriz de proximidad para reflejar esta operación de fusión.

(3) Si todos los patrones están en un clúster, deténgalo. De lo contrario, vaya al paso 2.

Según la forma en que se actualiza la matriz de proximidad en el paso 2, se pueden diseñar una variedad de algoritmos de aglomeración. Los algoritmos de división jerárquica comienzan con un único grupo de todos los objetos dados y siguen dividiendo los clústeres en función de algún criterio para obtener una partición de clústeres.

2) Algoritmos de partición

Un algoritmo de agrupamiento particional obtiene una única partición de los datos en lugar de una estructura de agrupamiento, como el dendograma producido por una técnica jerárquica. Los métodos de partición tienen ventajas en aplicaciones que implican grandes conjuntos de datos para los cuales la construcción de un dendograma es computacionalmente prohibitiva. Un problema que acompaña al uso de un algoritmo de partición es la elección del número de clústeres de salida deseados. Las técnicas de partición usualmente producen clústers al optimizar una función de criterio definida localmente (en un subconjunto de los patrones) o globalmente (definida en todos los patrones). La búsqueda combinatoria del conjunto de posibles etiquetas para un valor óptimo de un criterio es claramente computacionalmente prohibitiva. En la práctica, por lo tanto, el algoritmo generalmente se ejecuta varias veces con diferentes estados de inicio, y la mejor configuración obtenida de todas las ejecuciones se utiliza como clúster de salida.

Algoritmos de error al cuadrado

La función de criterio más intuitiva y de uso frecuente en las técnicas de agrupamiento particional es el criterio de error al cuadrado, que tiende a funcionar bien con clústeres aislados y compactos. El error al cuadrado para una clustering “L” de un conjunto de patrones “H” (que contiene K clusters) es

$$e^2(H,L) = \sum_{j=1}^{K} \sum_{i=1}^{n_j} \left\| \boldsymbol{X}_i^{(j)} - \boldsymbol{c}_j \right\|^2$$

Donde $\boldsymbol{X}_i^{(j)}$ es el i-ésimo patrón que pertenece al j-ésimo clúster y c_j es el centroide del j-ésimo clúster.

El k-means es el algoritmo más simple y más comúnmente usado que emplea un criterio de error al cuadrado (McQueen, 1967). Comienza con una partición inicial aleatoria y sigue reasignando los patrones a clusters en función de la similitud entre el patrón y los centros del clúster hasta que se cumple un criterio de convergencia (por ejemplo, no hay reasignación de ningún patrón de un clúster a otro, o el cuadrado el error deja de disminuir significativamente después de cierto número de iteraciones).

El algoritmo k-means es popular porque es fácil de implementar, y su complejidad de tiempo es O(n), donde n es el número de patrones. Un problema importante con este algoritmo es que es sensible a la selección de la partición inicial y puede converger a un mínimo local del valor de función criterio si la partición inicial no se elige correctamente.

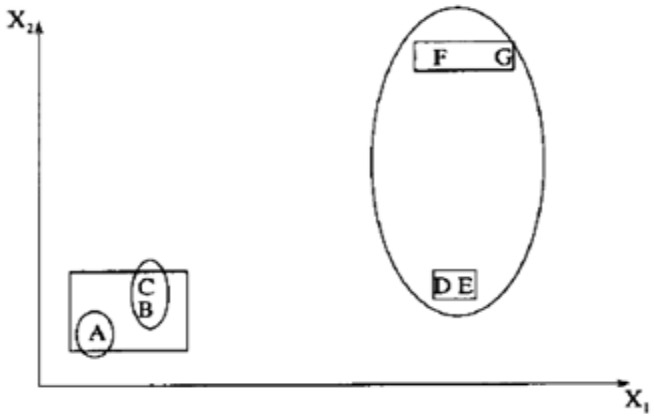


Figura 3. 11 El Algoritmo k-means es sensible a la partición inicial. (Jain, A. K., et al; 1999)

La figura 3.11 muestra siete patrones bidimensionales. Si comenzamos con los patrones A, B y C como el medio inicial alrededor del cual se construyen los tres clusters, entonces terminamos con la partición {{A}, {B, C}, {D, E, F, G} } mostrado por elipsis. El valor del criterio de error al cuadrado es mucho más grande para esta partición que para la mejor partición {{A, B, C}, {D, E}, {F, G}} mostrada por rectángulos, que arroja el valor mínimo global del cuadrado función de criterio de error para un clúster que contiene tres clústeres. La solución correcta de tres clústeres se obtiene eligiendo, por ejemplo, A, D y F.

Algoritmo de agrupamiento k-medias

(1) Elija k centros de clúster para que coincidan con k patrones elegidos al azar o k puntos aleatoriamente definidos dentro del hipervolumen que contiene el conjunto de patrones.
(2) Asigne cada patrón al centro del clúster más cercano.
(3) Volver a calcular los centros de clusters utilizando las membresías actuales del clúster.
(4) Si no se cumple un criterio de convergencia, vaya al paso 2. Los criterios de convergencia típicos son: no reasignación (o mínima) de patrones a nuevos centros de clúster o disminución mínima en el error al cuadrado.

Hay muchas variantes del algoritmo k-means (Anderberg, 1973). Algunos de ellos intentan seleccionar una buena partición inicial para que el algoritmo tenga más probabilidades de encontrar el valor mínimo global.

Otra variación es permitir la división y fusión de los clusters resultantes. Normalmente, un clúster se divide cuando su varianza está por encima de un umbral pre-especificado, y dos clústeres se fusionan cuando la distancia entre sus centroides está por debajo de otro umbral pre-especificado. Con esta variante, es posible obtener la partición óptima a partir de cualquier partición inicial arbitraria, siempre que se especifiquen los valores de umbral adecuados. El conocido algoritmo ISODATA (Ball & Hall, 1965) emplea esta técnica de fusión y división de clusters. Si a ISODATA se le da el particionamiento de "elipse" que se muestra en la Figura 3.11 como una partición inicial, producirá la partición óptima de tres clústeres. ISODATA fusionará primero los clusters {A} y {B, C} en un clúster porque la distancia entre sus centroides es pequeña y luego dividirá el clúster {D, E, F, G}, que tiene una gran varianza, en dos clusters {D, E} y {F, G}. Otra variación del algoritmo k-means implica seleccionar una función de criterio diferente por completo. El algoritmo de agrupamiento dinámico (que permite representaciones distintas del centroide para cada grupo), describe un enfoque de agrupamiento dinámico obtenido al formular el problema de agrupamiento en el marco de la estimación de máxima verosimilitud (Diday, 1973) (Symon, 1977).

En este Capítulo se ha presentado la Metodología CRISP-DM con sus etapas correspondientes. Además se han desarrollado la primeras etapas: Comprensión del Negocio, Comprensión, Preparación de los Datos, y Modelado. Siendo que en esta última, se han abordado las Técnicas a usar en este Trabajo Final. De esta manera se han abordado las Técnicas de Árboles de Decisión y Clustering con sus principales características.

En el siguiente Capítulo se continuará con las fases restantes de la Metodología CRISP-DM, discriminadas según la Técnica a utilizar y lo concerniente al Aprendizaje Automatizado.

Capítulo 4: Evaluación de los Datos

En el presente capítulo se introduce el Aprendizaje Automatizado, y se aplican las Técnicas correspondientes sobre los Datos de la EPH: Árboles de Decisión en el Aprendizaje Supervisado y Clustering en el Aprendizaje No Supervisado. Esto dentro del marco de la Metodología CRISP-DM. Luego de aplicar estas técnicas se describen los resultados obtenidos.

4.1 Aprendizaje Automatizado

El Aprendizaje Automatizado (AA) o Machine Learning (ML) es una disciplina científica del ámbito de la Inteligencia Artificial que crea sistemas que aprenden automáticamente. En este entorno, *Aprender* es reconocer patrones complejos en millones de datos. La “máquina” que realmente aprende es un “algoritmo”**,** el cual examina los datos para así predecir comportamientos futuros. *Automáticamente* estos sistemas podrán mejorar de forma autónoma con el tiempo, sin intervención humana.

El **Aprendizaje Automatizado** investiga como las computadoras pueden aprender (o mejorar su rendimiento) basado en datos. La principal área de investigación está en que los programas de computadora aprendan automáticamente a reconocer patrones complejos y tomen decisiones inteligentes basadas en datos. Por ejemplo, un problema típico de Machine Learning es programar una computadora para que pueda reconocer automáticamente los códigos postales escritos a mano en el correo después de aprender de un conjunto de ejemplos (Han, J. et al, 2011).

El **Aprendizaje Automatizado** puede ser Supervisado o No Supervisado y el objetivo es generar un modelo general a partir de ejemplos específicos. El modelo a generar se puede expresar de diversas formas: Árboles de Decisión, Lista de Reglas, Redes Neuronales, Modelos Bayesianos o Probabilísticos, etc.

4.1.1 Aprendizaje Supervisado

Según (Alejandro Cassis, 2015), se puede decir de forma sencilla que en el Aprendizaje Supervisado “se cuenta con un conjunto de ejemplos de los cuales conocemos la respuesta. Lo que deseamos es formular algún tipo de regla o correspondencia que nos permita dar (o aproximar) la respuesta para todos los objetos que se nos presenten.”

El **Aprendizaje Supervisado** es básicamente sinónimo de Clasificación. La supervisión en el aprendizaje proviene de los ejemplos etiquetados en el conjunto de datos. Por ejemplo, en el problema de reconocimiento de código postal, se utilizan un conjunto de imágenes de códigos postales escritos a mano y sus correspondientes traducciones legibles por máquina como ejemplos de entrenamiento, que supervisan el aprendizaje del modelo de clasificación. (Han, J. et al, 2011)

El **Aprendizaje Supervisado**, busca una función de correspondencia entre las entradas y las salidas deseadas del sistema; y la base de conocimiento del sistema está formada por ejemplos etiquetados anteriores.

Esto tiene su base en que “el Aprendizaje Estadístico Supervisado implica la construcción de un modelo estadístico para predecir o estimar una producción basada en una o más entradas.” (James, G., et al, 2013)

A partir de un conjunto de datos ya clasificados se busca una manera de derivar un modelo que permita clasificar ejemplos no vistos. Ejemplo: dados unos datos sobre un cliente, determinar si devolverá el crédito o no.

Hay muchos Algoritmos de Aprendizaje Supervisado, pero entre los más importantes se encuentran:

- **Algoritmo de Clasificación Naïve Bayes:** el cual, dado un ejemplo, permite encontrar la hipótesis que mejor lo describe. Para llegar a una conclusión, es necesario que el sistema se nutra de datos suficientes; pero si el volumen de información es muy grande, hay que recurrir a la hipótesis de la independencia condicional, que permite simplificar la expresión del Teorema de Bayes, factorizando la probabilidad. Entre los usos más comunes se encuentran el reconocimiento facial, y la detección de correo electrónico como spam o no spam.
- **Árboles de Decisión:** En base a un gráfico, se logra servir como apoyo a una toma de decisiones informada, al exponer las distintas opciones y sus posibles consecuencias. Como ventaja, los árboles de decisión, permiten abordar el problema de una manera estructurada y sistemática para llegar a una conclusión lógica. Entre sus usos, están el de predecir la respuesta del público ante el lanzamiento de un nuevo producto, o averiguar la idoneidad de una campaña de marketing.
- **Modelos de Regresión Lineal:** Basada en el Método de los Mínimos Cuadrados, permite realizar la regresión Lineal que puede aplicarse al análisis de relaciones entre variables financieras. Permite desde desarrollar previsiones de futuro, hasta identificar los factores que mayor incidencia tienen en la generación de beneficios de una corporación o determinar cuánto afectará un cambio en las tasas de interés a una cartera de bonos.
- **K-Vecinos más cercanos (k-NN Nearest Neighbour):** Es un sistema de clasificación supervisado basado en criterios de vecindad. K-NN se basa en

> la idea de que los nuevos ejemplos serán clasificados a la clase a la cual pertenezca la mayor cantidad de vecinos más cercanos de entrenamiento más cercanos a él. Este algoritmo explora todo el conocimiento almacenado en el conjunto de entrenamiento para determinar cuál será la clase a la que pertenece la muestra, teniendo únicamente en cuenta al vecino más cercano.

(Logicalis, 2017) (Fernando Sancho Caparrini, 2017)

La técnica que usaremos dentro de esta taxonomía es: ***Árboles de Decisión***.

4.1.2 Aprendizaje No Supervisado

Contrariamente a Aprendizaje Supervisado, los sistemas de clasificación No Supervisados son aquellos en los que no disponemos de una batería de ejemplos previamente clasificados, sino que a partir de las propiedades de los ejemplos intentamos dar una agrupación (clasificación, clustering) de los ejemplos según su similaridad. (Fernando Sancho Caparrini, 2017)

El **Aprendizaje No Supervisado** es esencialmente un sinónimo de agrupamiento. El proceso de aprendizaje no está supervisado ya que los ejemplos de entrada no están etiquetados por clase. Por lo general, podemos usar el agrupamiento para descubrir clases dentro de los datos. Por ejemplo, un método de aprendizaje no supervisado puede tomar, como entrada, un conjunto de imágenes de dígitos escritos a mano. Supongamos que encuentra 10 grupos de datos. Estos clústeres pueden corresponder a los 10 dígitos distintos de 0 a 9, respectivamente. Sin embargo, dado que los datos de entrenamiento no están etiquetados, el modelo aprendido no puede decirnos el significado semántico de los grupos encontrados (Han, J. et al, 2011).

En el **Aprendizaje No Supervisado**, tenemos un conjunto de ejemplos formado solamente por entradas al sistema. No se tiene información sobre las categorías de esos ejemplos. Por lo tanto, el sistema tiene que ser capaz de reconocer patrones para poder etiquetar las nuevas entradas. Tiene como objetivo descubrir grupos de ejemplos similares (llamados agrupamientos o clusters), o determinar la distribución de los datos dentro del espacio de entrada (density estimation), o bien proyectar óptimamente los datos de un espacio de alta dimensionalidad en un espacio de dos o tres dimensiones que haga posible la visualización de los mismos.

“Con el Aprendizaje Estadístico No Supervisado, hay insumos (inputs) pero no se supervisa la producción (outputs); sin embargo podemos aprender las relaciones y la estructura de tales datos” (James, G., et al, 2013).

La técnica que usaremos dentro de esta taxonomía es: ***Clustering.***

Es importante entender la diferencia entre la Clasificación No Supervisada y la Clasificación Supervisada. En la clasificación supervisada, se nos proporciona una colección

de patrones etiquetados (preclasificados); el problema es etiquetar un patrón recién encontrado, pero sin etiqueta. Normalmente, los patrones etiquetados (de entrenamiento) se utilizan para aprender las descripciones de las clases, que a su vez se utilizan para etiquetar un nuevo patrón. En el caso de la agrupación en clústeres, el problema es agrupar una determinada colección de patrones sin etiqueta en clústeres significativos. En cierto sentido, las etiquetas también están asociadas a clústeres, pero estas etiquetas de categoría están basadas en datos; es decir, se obtienen únicamente a partir de los datos (Jain, A. K., et al; 1999).

4.2 Aprendizaje Supervisado – Árboles de Decisión:

4.2.1 Fase de Evaluación

Construcción y evaluación del Modelo:

Seleccionada la técnica, la ejecutamos sobre los datos que se prepararon para generar los modelos y realizamos algunas pruebas.

Para ello, cargamos las tablas, haciendo uso de la herramienta RStudio con las siguientes sentencias en lenguaje R:

```
hogares=read.csv("C:/EPH/Hogares.csv", header=T, sep = ";")
individuos=read.csv("C:/EPH/Individuos.csv", header=T, sep = ";")
```

Con la sentencia “attach”, podemos trabajar individualmente con cada columna de la tabla:

```
attach(hogares)
attach(individuos)
```

Seguidamente, cargamos la librería “party” para hacer uso del algoritmo de Árboles de Decisión:

```
library(party)
```

Pruebas:

Aquí plasmamos algunas de las pruebas con resultados más interesantes que se obtuvieron. En la etapa de Evaluación, se mejoran los modelos y analizarán las pruebas:

Tabla Hogares:

“**Ej1.1** - Cantidad de hogares con 1, 2, 3… habitantes:

Para esto, utilizamos la función “summary (objeto)” que “imprime un resumen estadístico completo del análisis de regresión” (R Development Core Team, 1999). El parámetro de dicha función es “as.factor(IX_TOT)”, siendo “IX_TOT” la cantidad de habitantes por hogar. Usamos “as.factor” dado que “un factor es una variable categórica con un número finito de valores o niveles. En R los factores se utilizan habitualmente para realizar clasificaciones de datos, estableciendo su pertenencia a los grupos o categorías determinados por los niveles del factor.” (Santana, Hernández; 2016). De esta forma, podemos obtener una lista con la cantidad de hogares de acuerdo a la cantidad de personas que lo habitan.

```
> summary(as.factor(IX_TOT))
  1   2   3   4   5   6   7   8   9  10  11  12  13  14
 82 105  93  88  68  45  18  16   2   3   3   1   1   1
```

"**Ej2** – Variable Objetivo “Tipo de vivienda (**IV1**)” en función del tipo de conexión de agua (**IV6**), tipo de baño (**IV8**), si viven de lo que ganan en el trabajo (**V1**), o de jubilación (**V2**), o aguinaldo o pensión (**V21**), o subsidios (**V5**), o ganancia de otro negocio (**V9**), o beca de estudio (**V11**), si menores de 10 años ayudan con dinero trabajando (**V19_A**), o pidiendo dinero (**V19_B**), cantidad de miembros del hogar (**IX_Tot**), cantidad de miembros menores de 10 años (**IX_Men10**), cantidad de miembros mayores de 10 (**IX_Mayeq10**), Ingreso total familiar (**ITF**) e ingreso per-cápita familiar (**IPCF**)"

Primeramente se asigna la tabla hogares a una tabla auxiliar hogares.00:

hogares.00 <- hogares

Luego escribimos la fórmula para realizar el árbol, utilizando la sentencia “as.factor” para la “variable objetivo” (vble a predecir) que es numérica, de modo que sea posible realizar una categorización y estará en función (~) de las variables predictoras. Esta sentencia, a su vez es asignada a la variable hogares.formula.

hogares.formula <- as.factor(IV1) ~ IV6 + IV8 + + V1 +V2 + V21 + V5 + V9 + V11 + V19_A + V19_B + IX_TOT + IX_MEN10 + IX_MAYEQ10 + ITF + IPCF

Paso siguiente se crea el modelo. Para ello usamos la función ctree, la cual es “una clase no paramétrica de árboles de regresión que integran modelos de árboles regresión estructurados en una teoría bien definida de los procedimientos de inferencia condicional.” (Hothorn, Hornik, Zeileis; 2006). Los parámetros de ctree son la fórmula (hogares.formula) y la tabla auxiliar hogares.00 asignada a la variable data.

hogares.modelo <- ctree(hogares.formula, data=hogares.00)

Por último se grafica el árbol:

plot(hogares.modelo)

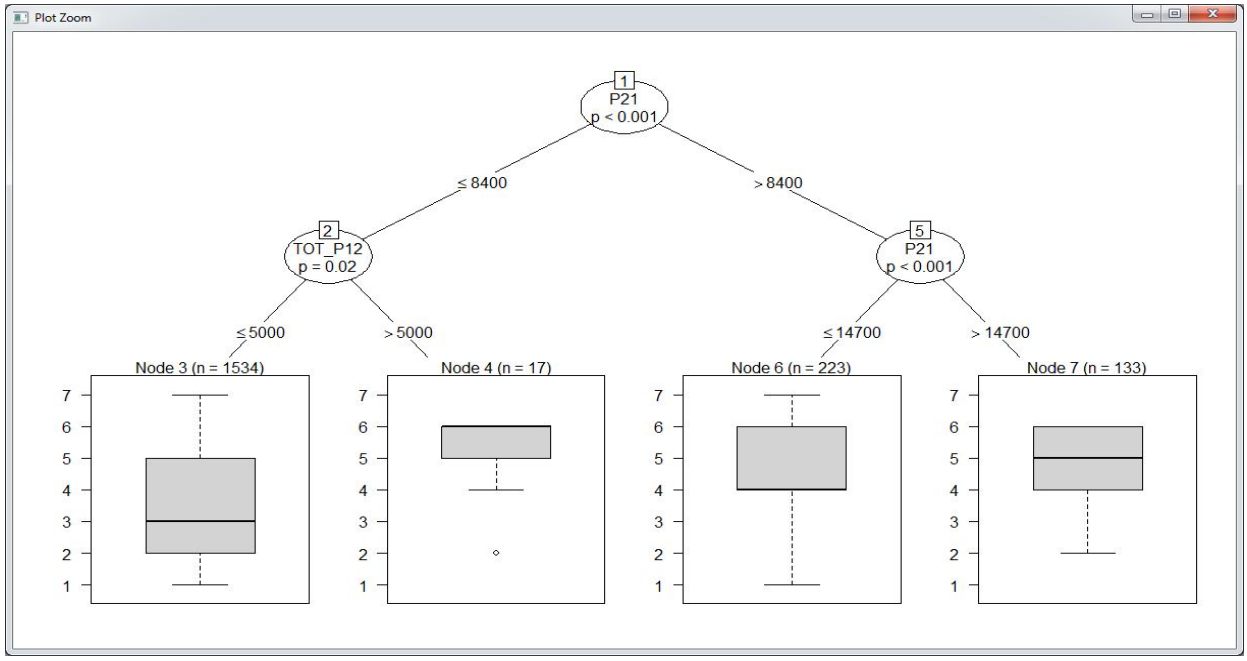


Figura 4. 1 Ejemplo de Árbol de Decisión.

En los nodos terminales observados la distribución según los tipos de hogares. Resultados que analizaremos en la etapa de Evaluación de la Metodología CRISP-DM.

Tabla Individuos:

"**Ejemplo 1**. Variable Objetivo: Nivel educativo (**NIVEL_ED**) en función del sexo (**ch04**), edad (**ch06**), Cobertura social (**ch08**) y dependencia laboral (**PP04A**), Ingreso ocupación principal (**P21**), Ingreso de otras ocupaciones (**Tot_p12**)".

Se establece un límite inferior de edad, es decir, mayores de 18 años:

mayor <- as.factor(CH06 >= 18)

Fórmula:

individuos.formula <- as.factor(NIVEL_ED) ~ CH04 + mayor + CH08 + PP04A + P21 + TOT_P12

Modelo:

individuos.modelo <- ctree(individuos.formula, data=cbind(mayor,individuos.00))

Gráfica:

plot(individuos.modelo, type="simple")

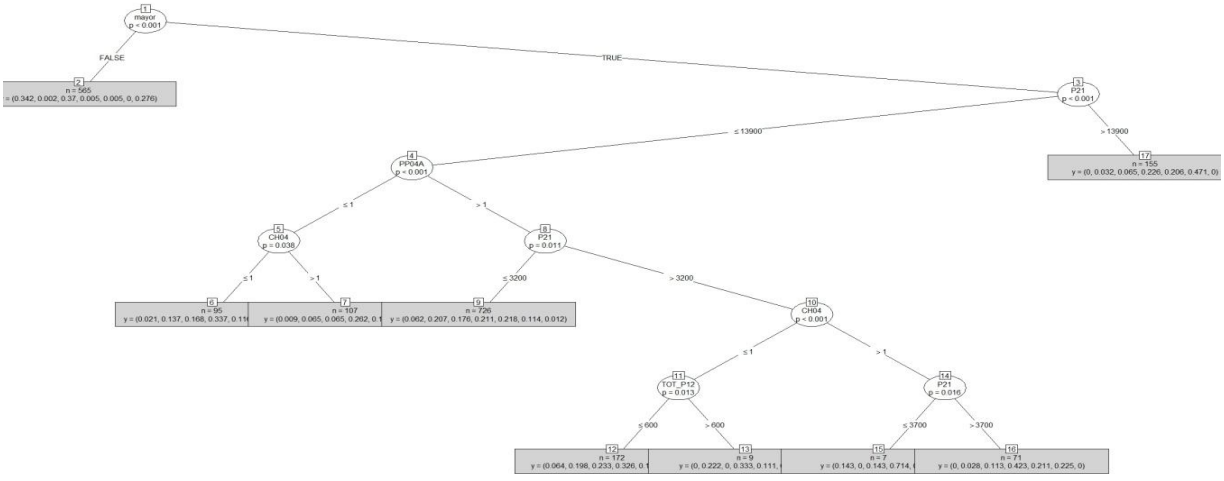


Figura 4. 2 Árbol de Decisión del Nivel educativo en función del sexo, edad, Cobertura social, Dependencia laboral, Ingreso ocupación principal, e Ingresos de otras ocupaciones.

Este gráfico es solo a modo de prueba. Los que se analizarán son los que vienen a continuación

4.2.2 Resultados

4.2.2.1 Caso 1 - Tabla individuos: Análisis del Nivel de Estudio

Caso 1: **Variable Objetivo:** Nivel educativo (**NIVEL_ED**) en función del sexo (**ch04**), edad (**ch06**), Cobertura social (**ch08**) y dependencia laboral (**PP04A**), Ingreso ocupación principal (**P21**), Ingreso de otras ocupaciones (**Tot_p12**)

Primeramente creamos una tabla auxiliar, a la cual se le asigna la tabla original de Individuos:

```
individuos.00 <- individuos
```

Para una mejor discriminación de las variables, les asignamos un nombre y algunas serán categorizadas con la función as.factor(). La función as.factor(), al ser aplicada a distintas variables, nos generará diversos árboles del mismo caso; por lo cual iremos analizando diferentes propuestas. La variable objetivo siempre estará categorizada para poder discriminar los distintos casos y sus porcentajes, en este caso el Nivel Educativo. La única variable predictora que siempre será categorizada es la edad, para poder diferenciar la mayoría de edad. En este primer caso, ninguna de las otras variables llevará ese tratamiento.

Variable Objetivo

```
nivelEducativo <- as.factor(NIVEL_ED)
```

Variables predictoras

```
sexo <-  CH04
mayor <- as.factor(CH06 >= 18 )
coberturaSocial <-  CH08
dependenciaLaboral <- PP04A

ingresoDeOcupacionPrincipal <- P21
ingresosDeOtrasOcupaciones <- TOT_P12
```

En cuanto a las variables de ingresos económicos no las categorizamos, de esta forma el árbol bifurcará en ganancias “menores a” y “mayores a”, simplificando la legibilidad del árbol. Si se aplicara “as.factor()” a estas variables mostrarían vectores con todos los montos posibles de ingresos, lo cual dificulta la interpretación del árbol.

Generamos la fórmula:

```
individuos.formula <- nivelEducativo ~ sexo + mayor + coberturaSocial + dependenciaLaboral + ingresoDeOcupacionPrincipal + ingresosDeOtrasOcupaciones
```

Creamos el modelo del árbol, donde la función “ctree” tiene como primer atributo la fórmula, y como segundo atributo la variable que indica la mayoría de edad y la tabla auxiliar individuos.00:

```
individuos.modelo <- ctree(individuos.formula, data=cbind(mayor,individuos.00))
```

Ahora se da origen a un primer gráfico:

```
plot(individuos.modelo)
```

El modelo resultante se observa en la Fig. 4.3:

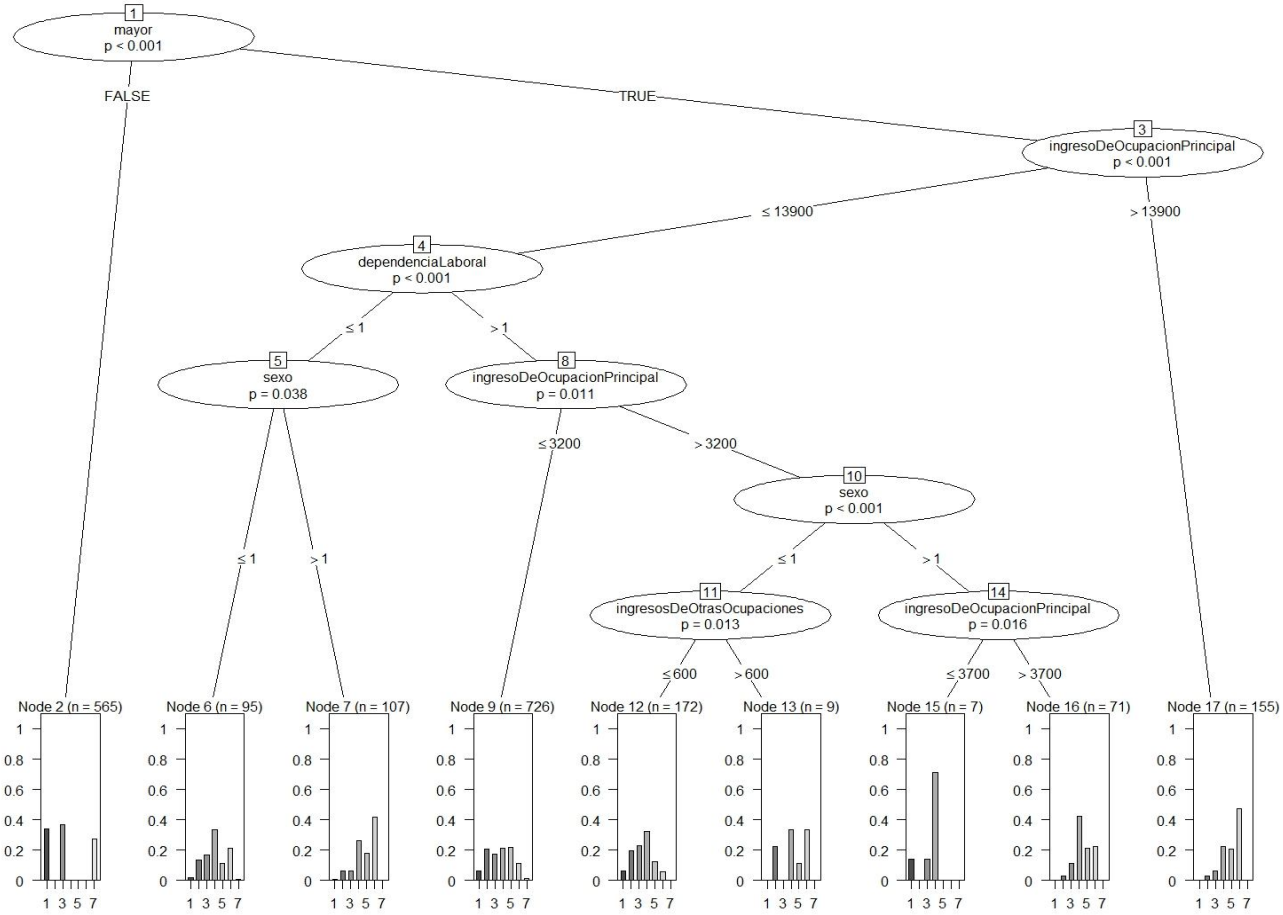


Figura 4. 3 Árbol de Decisión del Nivel educativo en función del sexo, edad, Cobertura social y dependencia laboral, Ingreso ocupación principal, e Ingreso de otras ocupaciones.

Ahora generamos un segundo árbol, que presenta los resultados en porcentajes, como lo vemos en la Fig.4.4:

plot(individuos.modelo, type="simple")

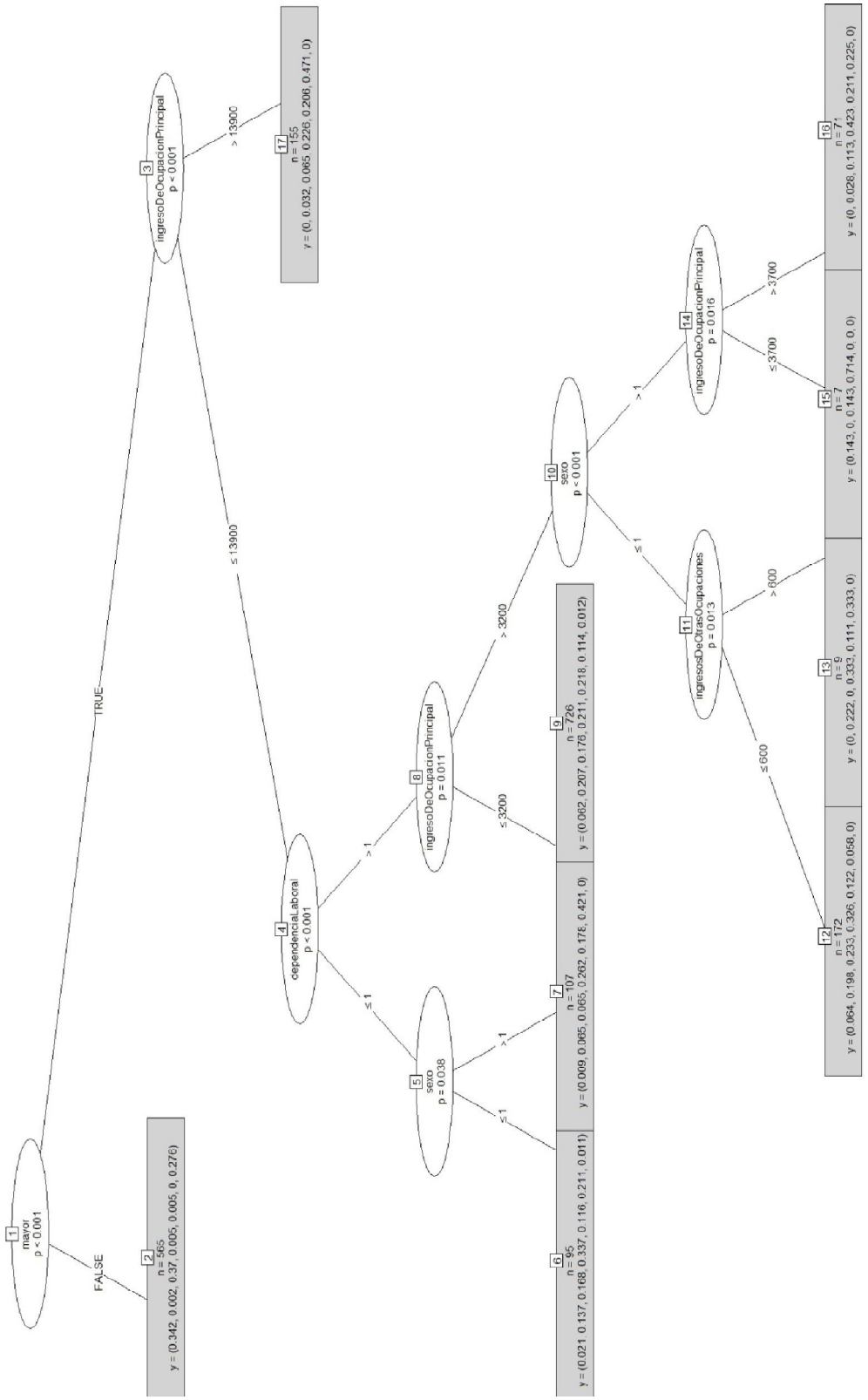


Figura 4. 4 Árbol de Decisión del Nivel educativo en función del sexo, edad, Cobertura social y dependencia laboral, Ingreso ocupación principal, e Ingreso de otras ocupaciones, con porcentajes de los niveles de Estudio en los Nodos terminales.

A continuación los significados de los atributos presentes en el Árbol:

Mayor = False → mayor < 18

Mayor = True → mayor >= 18

Sexo <= 1 → varón

Sexo > 1 → mujer

Dependencia Laboral <= 1 → Estatal

Dependencia Laboral > 1 → Privado

NivelEducativo:

= 1 → Primaria Incompleta (Incluye Educación Especial)

= 2 → Primaria Completa

= 3 → Secundaria Incompleta

= 4 → Secundaria Completa

= 5 → Superior Universitaria Incompleta

= 6 → Superior Universitaria Completa

= 7 → Sin instrucción

= 9 → Ns/Nr.

A continuación se procede a analizar los nodos de mayor importancia. Se analizan los que presentan características más sobresalientes. Descartándose de esta forma, nodos con muestras muy pequeñas que no influyen en la población; tampoco se tienen en cuenta los nodos correspondientes a menores de edad, dado que se busca encontrar características socio-económicas que correspondan principalmente a los Ingresos, y como esto influye en el Nivel de Educación, Cobertura Social, entre otros.

La descripción del Nodo corresponde a conjunciones lógicas de acuerdo a las bifurcaciones del Árbol. El camino desde el nodo raíz al nodo terminal, será expresado mediante estas conjunciones lógicas. Por empleo: (atributoA = 1 ^ atributoB > 2 ^ atributoC < 3). Seguidamente se expresa la cantidad de individuos mediante la conjunción: n = xx; por ejemplo, n = 22, quiere decir que ese Nodo representa a un grupo de 22 personas. Teniendo en cuenta esto, podemos interpretar esto como: Un grupo de 22 personas donde el Atributo A es igual 1, el Atributo B es mayor a 2 y el atributo C es menor a 3. Y de acuerdo a los valores de los atributos y sus significados se procede a interpretar las características socio-económicas de cada Nodo. Gracias al Árbol de la figura 4.4, podemos acceder a los porcentajes concernientes, y con estos valores, mediante unas sentencias en R, se obtienen gráficas circulares para una mayor apreciación de los datos encontrados.

Ahora, se procede a analizar los Nodos:

Nodo 6: Nodo 6 del Árbol correspondiente a las Figuras 4.3 y 4.4:

(mayor = TRUE ∧ ingresoDeOcupacionPrincipal <= 13900 ∧ dependeciaLaboral <= 1 ∧ sexo <= 1) → n = 95.

Grupo de 95 Hombres mayores a 18 años, con Ingresos de la Ocupación Principal menores a $13.900, y trabajan en el Estado:

= 1 → 2,1% Primaria Incompleta (Incluye Educación Especial)

= 2 → 13,7% Primaria Completa

= 3 → 16,8% Secundario Incompleta

= 4 → 33,7% Secundario Completa

= 5 → 11,6% Superior Universitario Incompleta

= 6 → 21,1% Superior Universitario Completa

= 7 → 1,1% Sin instrucción

Entre 95 hombres que trabajan en el Estado, con sueldos menores a $13.900, el 21,1% de los encuestados tiene Nivel superior Universitario Completo, y un 11,6% Incompleto (puede ser que este cursando o haya abandonado); un 33,7% tiene Secundario Completo y un 16,8% incompleto; por último hay un 13,7% que solo tiene nivel de Primaria Completo, un 2,1% que no terminó la Primaria y un 1,1% sin instrucción.

Ahora se obtiene un gráfico circular mediante las siguientes sentencias en R:

Se cargan los porcentajes:

```
ej1nodo6Datos <- matrix(c(0.021, 0.137, 0.168, 0.337, 0.116, 0.211, 0.011), ncol = 7)
```

Se nombran los atributos:

```
ej1nodo6Name<- colnames(ej1nodo6Datos) <- c("Primaria Incompleta", "Primaria Completa", "Secundaria Incompleta", "Secundaria Completa", "Superior Universitaria Incompleta", "Superior Universitaria Completa", "Sin Instruccion")
```

Por último, se grafica:

```
pie(ej1nodo6Datos, labels = ej1nodo6Name, radius = 1, col=rainbow(14), main = "Ej 1 - Nodo 6")
```

De igual forma se procederá en los demás nodos.

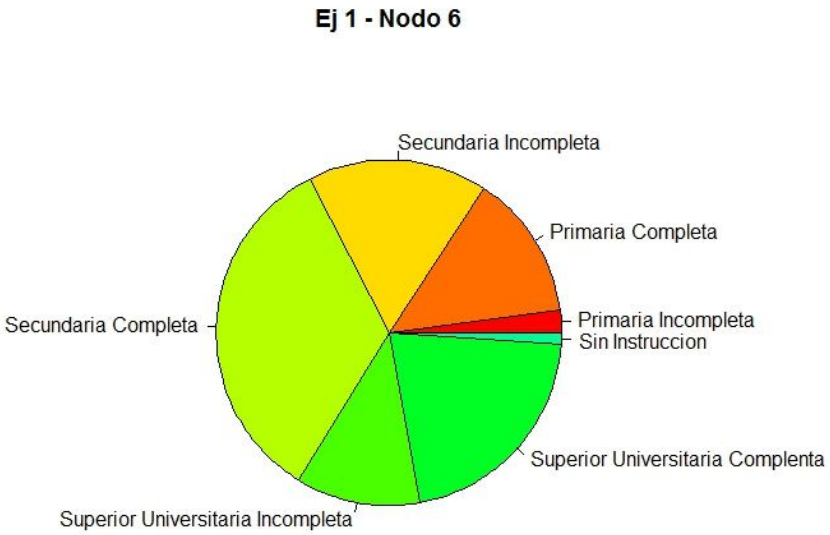


Figura 4. 5 Nodo 6 del Árbol con Nivel educativo en función del sexo, edad, Cobertura social y dependencia laboral, Ingreso ocupación principal, e Ingreso de otras ocupaciones.

Nodo 7: Nodo 7 del Árbol correspondiente a las Figuras 4.3 y 4.4:

(mayor = TRUE ∧ ingresoDeOcupacionPrincipal <= 13900 ∧ dependeciaLaboral <= 1 ∧ sexo > 1) → n = 107.

Grupo de 107 Mujeres mayores a 18 años, con Ingresos de la Ocupación Principal menores a $13.900, trabajan en el Estado:

= 1 → 0,9% Primaria Incompleta (Incluye Educación Especial)

= 2 → 6,5% Primaria Completa

= 3 → 6,5% Secundaria Incompleta

= 4 → 26,2% Secundaria Completa

= 5 → 17,8% Superior Universitaria Incompleta

= 6 → 42,1% Superior Universitaria Completa

= 7 → 0% Sin instrucción

Tenemos 107 casos de mujeres mayores de 18 años que trabajan en el Estado con sueldos menores a $13.900, donde predomina el nivel Superior Universitario, completo (42,1%) e incompleto (17,8%), y el nivel Secundario Completo con un 26,2%. Dado que los resultados vienen denotando que a mayor nivel de instrucción mayor es el poder adquisitivo, se puede suponer que en este grupo de universitarios recibidos, con sueldos menores a $13.900, sean profesionales con pocos años de antigüedad.

```
ej1nodo7Datos <- matrix(c(0.009, 0.065, 0.065, 0.262, 0.178, 0.421, 0), ncol = 7)

ej1nodo7Name<- colnames(ej1nodo7Datos) <- c("Primaria Incompleta", "Primaria Completa", "Secundaria Incompleta", "Secundaria Completa", "Superior Universitaria Incompleta", "Superior Universitaria Completa", "Sin Instruccion")

pie(ej1nodo7Datos, labels = ej1nodo7Name, radius = 1, col=rainbow(20), main = "Ej 1 - Nodo 7")
```

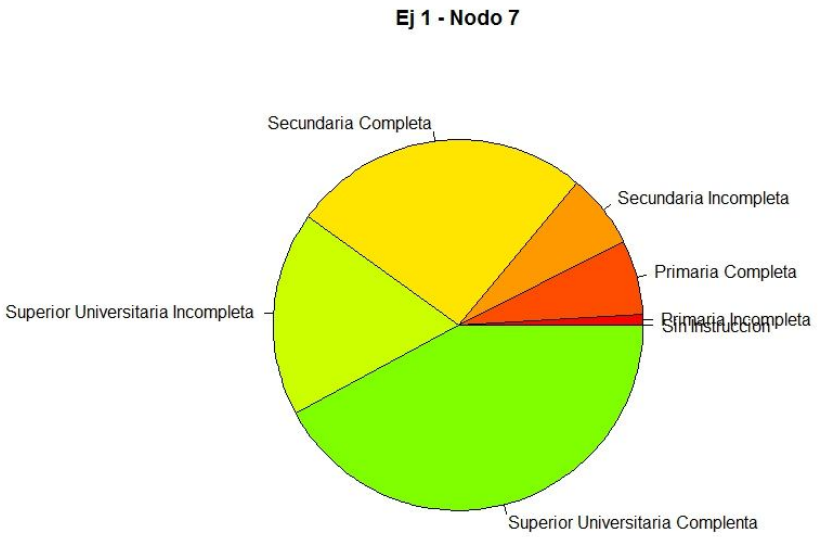


Figura 4. 6 Nodo 7 del Árbol con Nivel educativo en función del sexo, edad, Cobertura social y dependencia laboral, Ingreso ocupación principal, e Ingreso de otras ocupaciones.

Comparando los Nodo 6 y 7, de las Figuras 4.5 y 4.6, se puede observar que en el grupo de 95 Hombres, hay un 21,1% individuos que tienen Nivel Superior Universitario Completo, esto es 20 de los 95; mientras que en el grupo de 107 Mujeres, el 42,1% llegan a este nivel, esto es 45 de las 107 Mujeres.

En este sector de la población, que trabaja para el Estado con sueldos menores a $13.900, se observa que entre Mujeres hay un mayor porcentaje con Nivel Universitario Completo, que entre los hombres, como también que es mayor la cantidad de mujeres con este Nivel Educativo que de varones.

También se puede observar que a pesar de que los Ingresos Principales son bajos, el Nivel de Estudios prominente es el Superior Universitario Completo.

Nodo 9: Nodo 9 del Árbol correspondiente a las Figuras 4.3 y 4.4:

(mayor = TRUE ^ ingresoDeOcupacionPrincipal <= 13900 ^ dependeciaLaboral > 1 ^ ingresoDeOcupacionPrincipal <= 3200) → n = 726.

Mayores a 18 años, que trabajan en el Sector Privado u otras, con Ingresos de la Ocupación Principal menores a $3.200,:

= 1 → 6,2% Primaria Incompleta (Incluye Educación Especial)

= 2 → 20,7% Primaria Completa

= 3 → 17,6% Secundaria Incompleta

= 4 → 21,1% Secundaria Completa

= 5 → 21,8% Superior Universitaria Incompleta

= 6 → 11,4% Superior Universitaria Completa

= 7 → 1,2% Sin instrucción

Este grupo presenta una muestra importante, dado que son 726 personas de 1.907 encuestados (esto es un 38,07% de la muestra general), y se observa que el 21,8% de 726 casos encuestados, mayores de 18 años con ingresos inferiores a $3.200 y trabajan en el área privada u otras, tienen Estudio Superiores Universitarios Incompletos, y el 11,4% tiene el nivel completo; un 21,1% tiene Secundario completo, y un 17,6% incompleto; en este grupo un 20,7% solo tiene nivel Primario completo, y un 6,2% incompleto. La distribución de los porcentajes en esta muestra es bastante equilibrada.

```
ej1nodo9Datos <- matrix(c(0.062, 0.207, 0.176, 0.211, 0.218, 0.114, 0.012), ncol = 7)
ej1nodo9Name<- colnames(ej1nodo9Datos) <- c("Primaria Incompleta", "Primaria Completa", "Secundaria Incompleta", "Secundaria Completa", "Superior Universitaria Incompleta", "Superior Universitaria Completa", "Sin Instruccion")
pie(ej1nodo9Datos, labels = ej1nodo9Name, radius = 1.1, col=rainbow(20), main = "Ej 1 - Nodo 9")
```

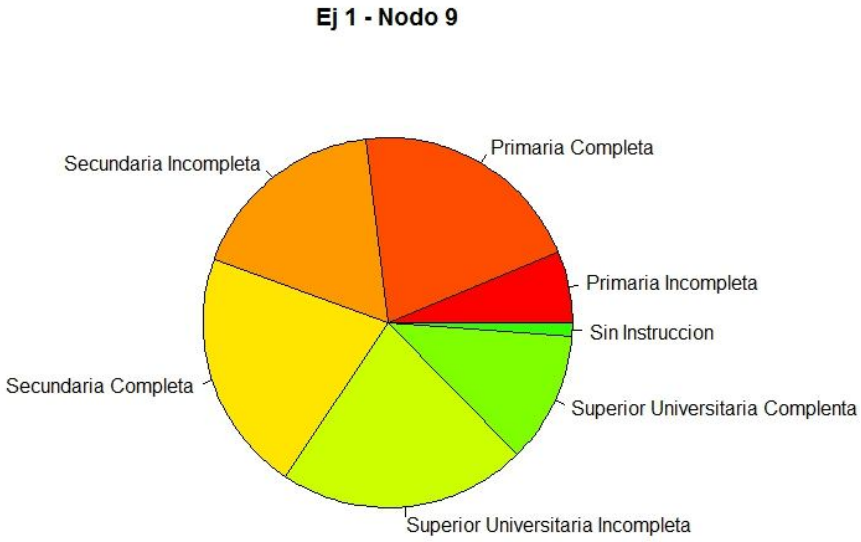


Figura 4. 7 Nodo 9 del Árbol con Nivel educativo en función del sexo, edad, Cobertura social y dependencia laboral, Ingreso ocupación principal, e Ingreso de otras ocupaciones.

Nodo 12: Nodo 12 del Árbol correspondiente a las Figuras 4.3 y 4.4:

(mayor = TRUE ∧ ingresoDeOcupacionPrincipal <= 13900 ∧ dependeciaLaboral > 1 ∧ ingresoDeOcupacionPrincipal > 3200 ∧ sexo <= 1 ∧ ingresosDeOtrasOcupaciones <= 600) → n = 172.

Hombres mayores a 18 años, con ingresos de la ocupación principal entre $3.200 y $13.900, trabajan en el Sector Privado u otros, y con ganancias de otras ocupaciones menores a $600:

= 1 → 6,4% Primaria Incompleta (Incluye Educación Especial)

= 2 → 19,8% Primaria Completa

= 3 → 23,3% Secundaria Incompleta

= 4 → 32,6% Secundaria Completa

= 5 → 12,2% Superior Universitaria Incompleta

= 6 → 5,8% Superior Universitaria Completa

= 7 → 0% Sin instrucción

Un gran porcentaje de este grupo de 172 Hombres, que trabaja en el área privada, con sueldos de entre $3.200 y $13.900, y otras ganancias menores a $600, tiene

Secundario Completa (32,6%) e incompleta (23,3%); el porcentaje de universitarios recibidos es bajo (solo un 5,8%).

ej1nodo12Datos <- matrix(c(0.064, 0.198, 0.233, 0.326, 0.122, 0.058, 0), ncol = 7)

ej1nodo12Name<- colnames(ej1nodo12Datos) <- c("Primaria Incompleta", "Primaria Completa", "Secundaria Incompleta", "Secundaria Completa", "Superior Universitaria Incompleta", "Superior Universitaria Completa", "Sin Instruccion")

pie(ej1nodo12Datos, labels = ej1nodo12Name, radius = 1.1, col=rainbow(20), main = "Ej 1 - Nodo 12")

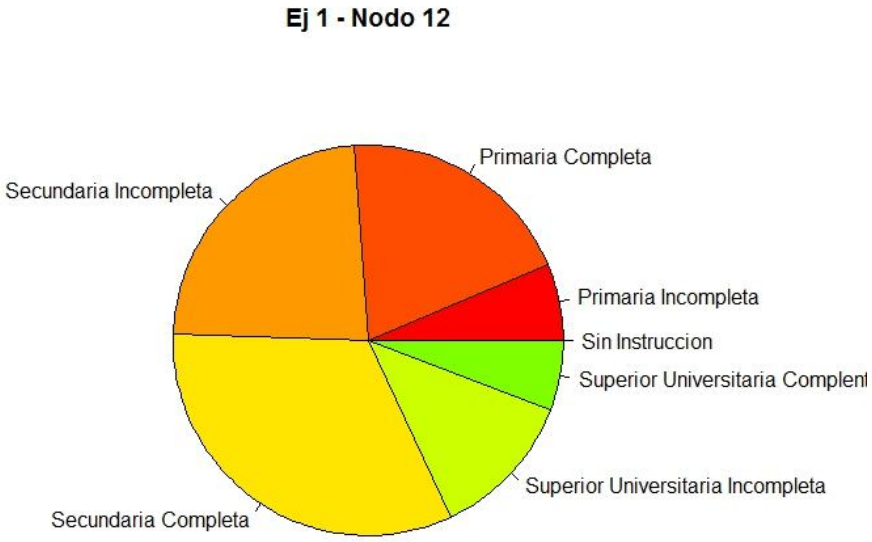


Figura 4. 8 Nodo 12 del Árbol con Nivel educativo en función del sexo, edad, Cobertura social y dependencia laboral, Ingreso ocupación principal, e Ingreso de otras ocupaciones.

Nodo 13: Nodo 13 del Árbol correspondiente a las Figuras 4.3 y 4.4:

(mayor = TRUE ^ ingresoDeOcupacionPrincipal <= 13900 ^ dependeciaLaboral > 1 ^ ingresoDeOcupacionPrincipal > 3200 ^ sexo <= 1 ^ ingresosDeOtrasOcupaciones > 600) → n = 9.

Hombres mayores a 18 años, con ingresos de la ocupación principal entre $3.200 y $13.900, trabajan en el área privada u otras, y con ganancias de otras ocupaciones mayores a $600:

= 1 → 0% Primaria Incompleta (Incluye Educación Especial)

= 2 → 22,2% Primaria Completa

= 3 → 0% Secundaria Incompleta

= 4 → 33,3% Secundaria Completa

= 5 → 11,1% Superior Universitaria Incompleta

= 6 → 33,3% Superior Universitaria Completa

= 7 → 0% Sin instrucción

En este grupo de hombres mayores de 18 años, que trabajan en el área privada, con sueldos entre $3.200 y $13.900, y con otras ganancias mayores a $600, hay un 33,3% con nivel Universitario Completo, el mismo porcentaje con Secundario Completo y un 22,2% con Primaria Completa.

```
ej1nodo13Datos <- matrix(c(0, 0.222, 0, 0.333, 0.111, 0.333, 0), ncol = 7)
ej1nodo13Name<- colnames(ej1nodo13Datos) <- c("Primaria Incompleta", "Primaria Completa", "Secundaria Incompleta", "Secundaria Completa", "Superior Universitaria Incompleta", "Superior Universitaria Completa", "Sin Instruccion")
pie(ej1nodo13Datos, labels = ej1nodo13Name, radius = 1.1, col=rainbow(20), main = "Ej 1 - Nodo 13")
```

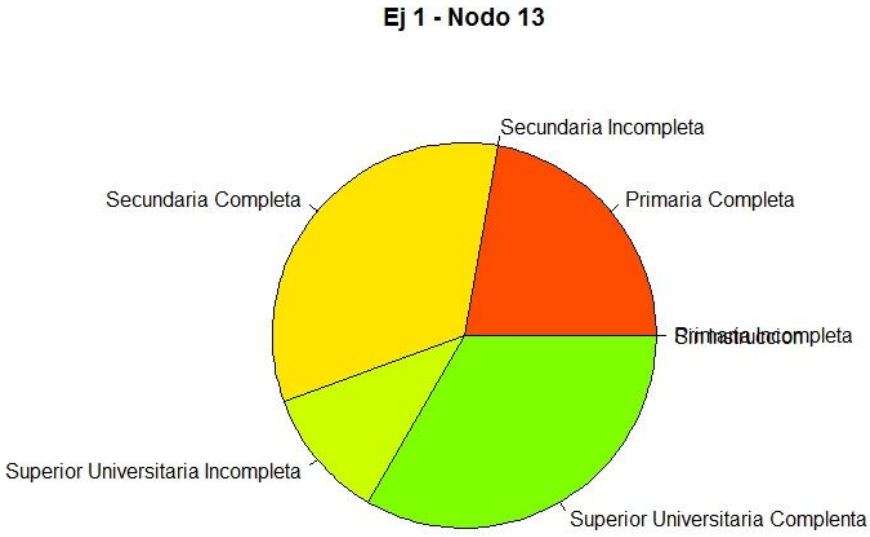


Figura 4. 9 . Nodo 13 del Árbol con Nivel educativo en función del sexo, edad, Cobertura social y dependencia laboral, Ingreso ocupación principal, e Ingreso de otras ocupaciones.

En los Nodos 12 y 13, correspondientes a las Figuras 4.8 y 4.9, tenemos Hombres mayores de 18 años que trabajan en el sector Privado con Ingresos de su Ocupación principal de entre $3.200 y $13.900. En el Nodo 12 se presentan 172 individuos que tienen Ingresos de otras actividades menores a $600, donde predomina el Nivel Secundario Completo, mientras que en el Nodo 13 hay un pequeño grupo de 9 hombres que tienen

ingresos de otras actividades mayores a $600, donde no solo predomina el Nivel Secundario Completo, sino también el Nivel Superior Universitario Completo, que en el Nodo 12 es muy bajo.

Nodo 17: Nodo 17 del Árbol correspondiente a las Figuras 4.3 y 4.4:

(mayor = TRUE ^ ingresoDeOcupacionPrincipal > 13900) → n = 155

Mayores a 18 años con ingresos de la ocupación principal mayores a $13.900:

= 1 → 0% Primaria Incompleta (Incluye Educación Especial)

= 2 → 3,2% Primaria Completa

= 3 → 6,5% Secundaria Incompleta

= 4 → 22,6% Secundaria Completa

= 5 → 20,6% Superior Universitaria Incompleta

= 6 → 47,1% Superior Universitaria Completa

= 7 → 0% Sin instrucción

En este grupo de 155 casos, con personas mayores de 18 años, sin distinguir sexo o dependida laboral, con sueldos mayores a $13.900, un 47,1% tiene Nivel Superior Universitario Completo y un 20,6% Incompleto; y se observa un 22,6 con Secundario Completo.

```
ej1nodo17Datos <- matrix(c(0, 0.032, 0.065, 0.226, 0.206, 0.471, 0), ncol = 7)

ej1nodo17Name<- colnames(ej1nodo17Datos) <- c("Primaria Incompleta", "Primaria Completa", "Secundaria Incompleta", "Secundaria Completa", "Superior Universitaria Incompleta", "Superior Universitaria Completa", "Sin Instrucción")

pie(ej1nodo17Datos, labels = ej1nodo17Name, radius = 1.1, col=rainbow(20), main = "Ej 1 - Nodo 17")
```

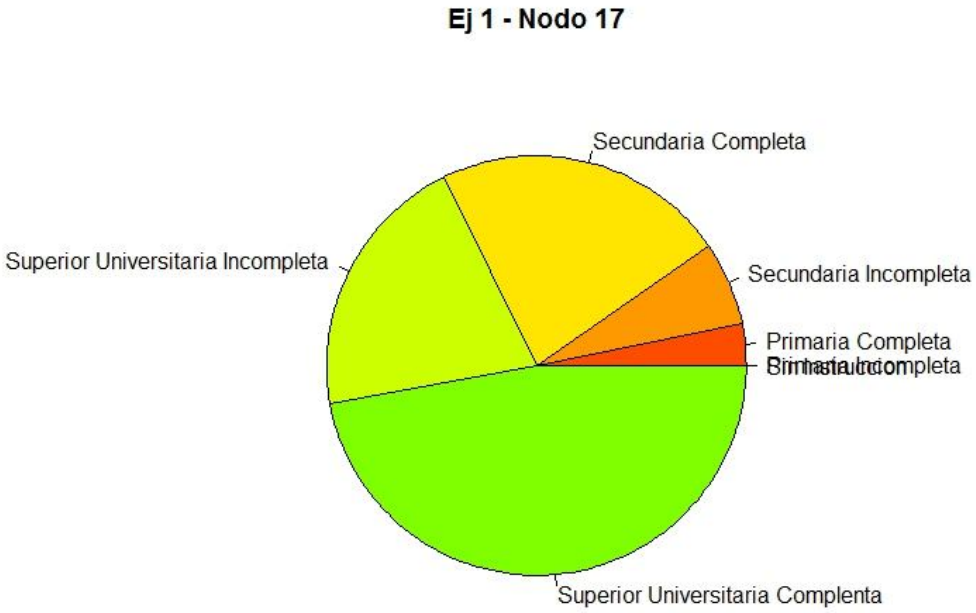


Figura 4. 10 Nodo 17 del Árbol con Nivel educativo en función del sexo, edad, Cobertura social y dependencia laboral, Ingreso ocupación principal, e Ingreso de otras ocupaciones.

Conclusión:

De acuerdo al Estudio realizado con el Método de Árboles de Decisión, se arriba a la siguiente conclusión:

a) Los Ingresos responden al Nivel Educativo, independientemente de si trabajan para el Estado o en el área Privada u otros. Los mayores sueldos se ven reflejados en el Nivel de Instrucción, en especial los estudios Superiores Universitarios, sean Completos o Incompletos. Aunque siempre existen excepciones a la regla.
b) Por otra parte, hay quienes realizan otras actividades fuera de la actividad principal, en donde también influye el Nivel Educativo. Ya que se ve reflejado que quienes mayor instrucción tienen, mas son las ganancias de estas actividades adicionales. Esto se refleja en los Ingresos mensuales, el poder adquisitivo y su lugar en la comunidad, esto es debido, a que cuanto mayor sea su nivel de instrucción, el círculo de personas con quienes tiene contacto tiende a ser del mismo nivel, y esto lleva a obtener mayores ganancias de sus actividades, sean changas u otros trabajos independientes.
c) También es necesario resaltar que entre los Empleados Estatales con Ingresos menores a $13.900, las Mujeres tienen un mayor Nivel de Instrucción que los Hombres. Se puede concluir que 2 de cada 10 varones tienen Nivel Superior Universitario Completo, mientras que 4 de 10 Mujeres llegan a este Nivel de Instrucción.

4.2.2.2 Caso 1.1 - Tabla individuos: Análisis del Nivel de Estudio

"**Caso 1.1- Variable Objetivo:** Nivel educativo (**NIVEL_ED**) en función del sexo (**ch04**), edad (**ch06**), Cobertura social (**ch08**) y dependencia laboral (**PP04A**), Ingreso ocupación principal (**P21**), Ingreso de otras ocupaciones (**Tot_p12**)"

Variable Objetivo:

nivelEducativo <- as.factor(NIVEL_ED)

Variables predictoras:

Hacemos una modificación, **categorizamos la variable “coberturaSocial”** con la función “as.factor()” :

sexo <- CH04

mayor <- as.factor(CH06 >= 18)

coberturaSocial <- as.factor(CH08)

dependenciaLaboral <- PP04A

ingresoDeOcupacionPrincipal <- P21

ingresosDeOtrasOcupaciones <- TOT_P12

Fórmula:

individuos.formula <- nivelEducativo ~ sexo + mayor + coberturaSocial + dependenciaLaboral + ingresoDeOcupacionPrincipal + ingresosDeOtrasOcupaciones

Modelo:

individuos.modelo <- ctree(individuos.formula, data=cbind(mayor,individuos.00))

Gráfico:

plot(individuos.modelo)

A continuación, en la Fig. 4.11, el modelo resultante:

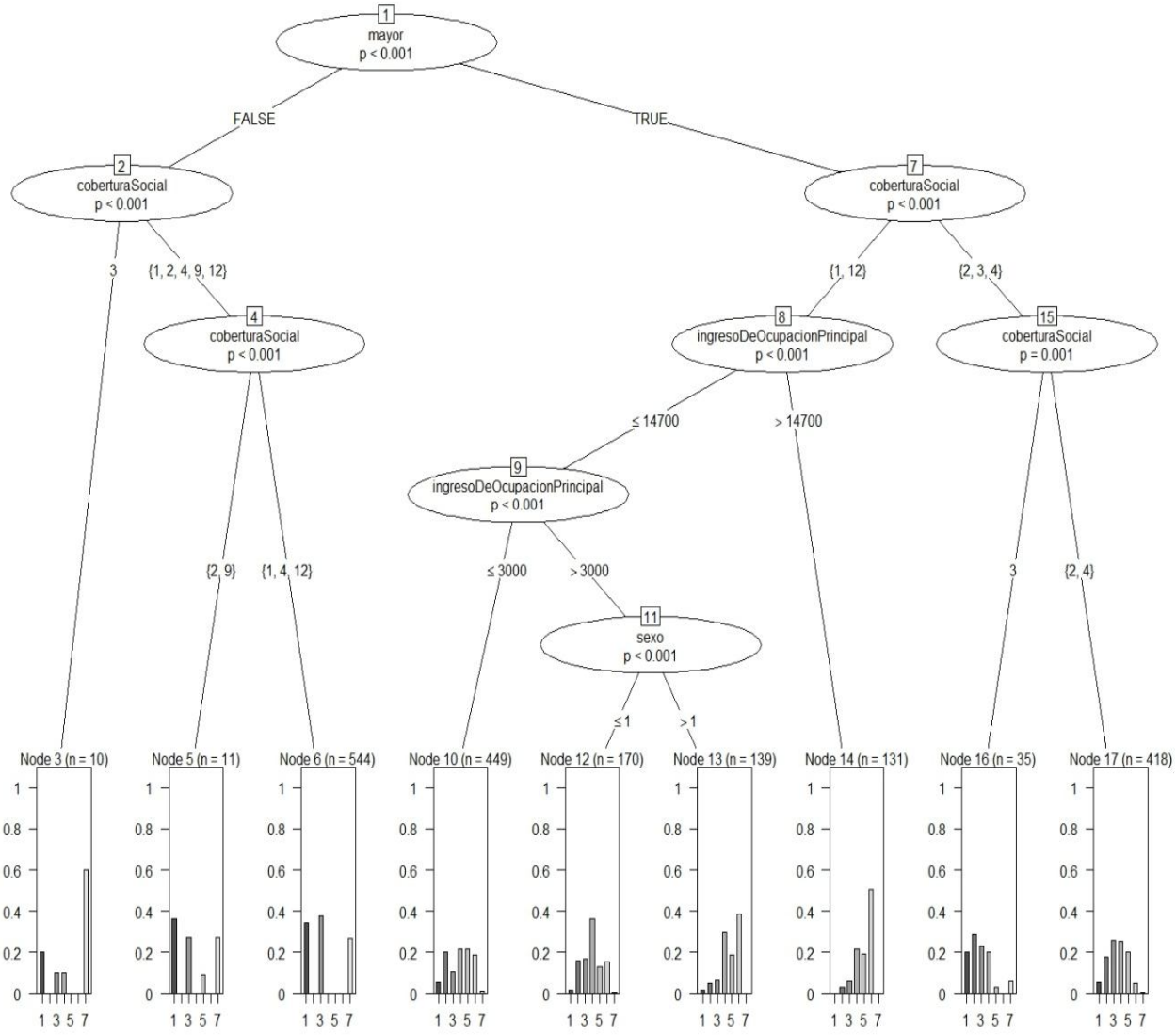


Figura 4. 11 Árbol de Decisión del Nivel educativo en función del sexo, edad, Cobertura social y dependencia laboral, Ingreso ocupación principal, e Ingreso de otras ocupaciones. Variable Cobertura Social Categorizada.

En la Figura 4.12 se obtiene el mismo modelo, presentando la variante que en los nodos terminales se presentan los porcentajes obtenidos. Esto mediante la siguiente sentencia:

plot(individuos.modelo, type="simple")

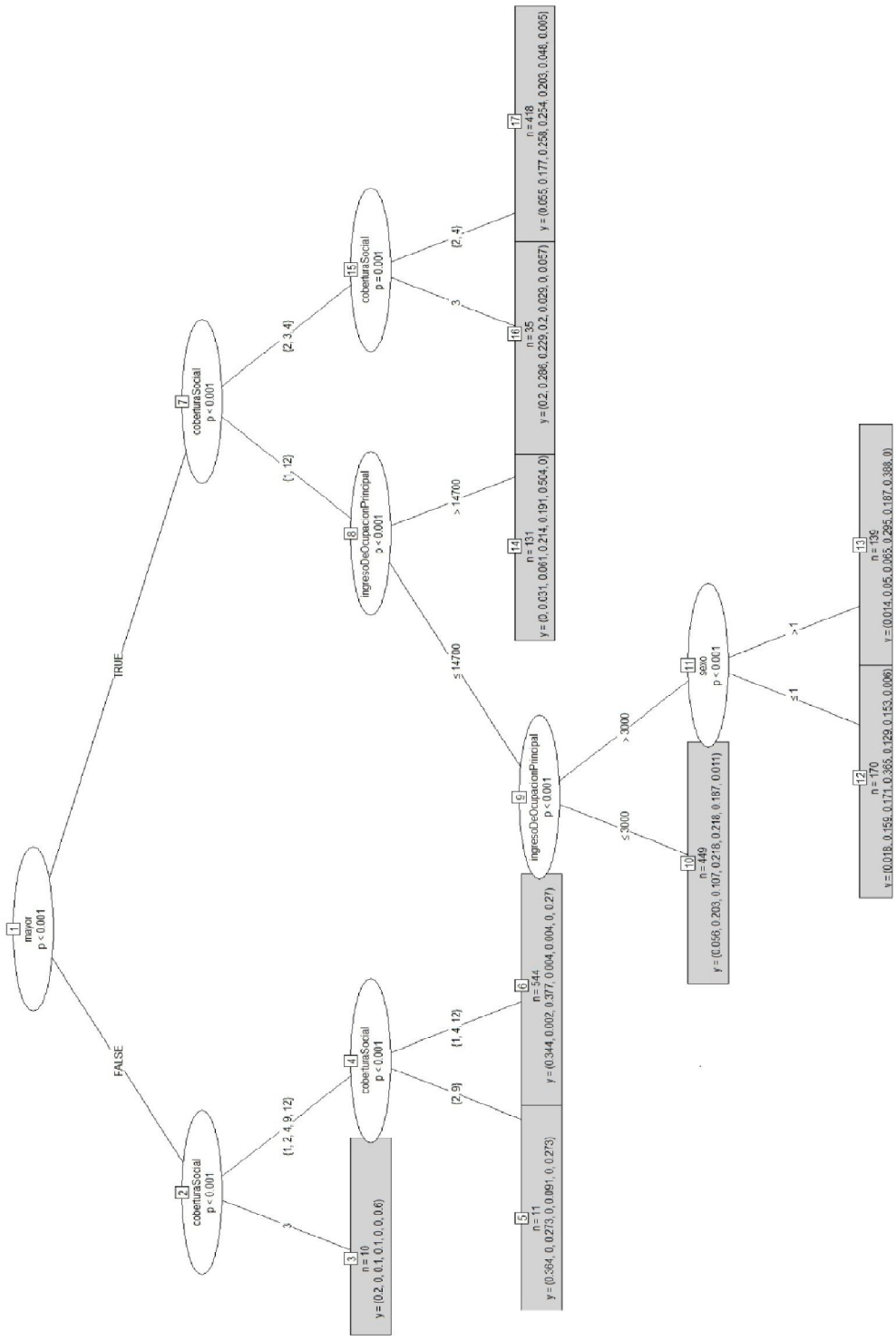


Figura 4. 12 Árbol de Decisión del Nivel educativo, con porcentajes de los niveles de Estudio en los Nodos terminales y Variable Cobertura Social Categorizada.

A continuación los significados de los atributos presentes en el Árbol:

Mayor = False → menor a 18 años

Mayor = True → mayor de 18 años

Sexo <= 1 → varón

Sexo > 1 → mujer

Cobertura Social:

= 1 → Obra Social (Incluye PAMI)

= 2 → Mutual / Prepaga / Servicio de emergencia

= 3 → Planes y Seguros públicos

= 4 → No paga ni le descuentan

= 9 → Ns. / Nr.

= 12 → Obra Social y Mutual / prepaga / servicio de emergencia.

= 13 → Obra Social y Planes y Seguros Públicos

= 23 → Mutual / prepaga / servicio de emergencia / Planes y Seguros Públicos

= 123 → Obra Social, mutual / prepaga / servicio de emergencia y Planes y Seguros Públicos.

A continuación se procede a analizar los nodos de mayor importancia. Se analizan los que presentan características más sobresalientes. Descartándose de esta forma, nodos con muestras muy pequeñas que no influyen en la población; tampoco se tienen en cuenta los nodos correspondientes a menores de edad, dado que se busca encontrar características socio-económicas que correspondan principalmente a los Ingresos, y como esto influye en el Nivel de Educación, Cobertura Social, entre otros.

La descripción del Nodo corresponde a conjunciones lógicas de acuerdo a las bifurcaciones del Árbol. El camino desde el nodo raíz al nodo terminal, será expresado mediante estas conjunciones lógicas. Por empleo: (atributoA = 1 ∧ atributoB > 2 ∧ atributoC < 3). Seguidamente se expresa la cantidad de individuos mediante la conjunción: n = xx; por ejemplo, n = 22, quiere decir que ese Nodo representa a un grupo de 22 personas. Teniendo en cuenta esto, podemos interpretar esto como: Un grupo de 22 personas donde el Atributo A es igual 1, el Atributo B es mayor a 2 y el atributo C es menor a 3. Y de acuerdo a los valores de los atributos y sus significados se procede a interpretar las características socio-económicas de cada Nodo. Gracias al Árbol de la figura 4.12, podemos acceder a los porcentajes concernientes, y con estos valores, mediante unas sentencias en R, se obtienen gráficas circulares para una mayor apreciación de los datos encontrados.

Ahora, se procede a analizar los Nodos:

Nodo 12: Nodo 12 del Árbol correspondiente a las Figuras 4.11 y 4.12:

(mayor = TRUE ∧ coberturaSocial = {1,12} ∧ ingresoDeOcupacionPrincipal <= 14700 ∧ ingresoDeOcupacionPrincipal > 3000 ∧ sexo <= 1) → n = 170

Hombres mayores de 18 años con Obra Social o bien Obra Social y Mutual o Prepaga o Servicios de Emergencia, con Ingresos de la Ocupación Principal de entre $3.000 y $14.700:

= 1 → 1,8% Primaria Incompleta (Incluye Educación Especial)

= 2 → 15,9% Primaria Completa

= 3 → 17,1% Secundaria Incompleta

= 4 → 36,5% Secundaria Completa

= 5 → 12,9% Superior Universitaria Incompleta

= 6 → 15,3% Superior Universitaria Completa

= 7 → 0,6% Sin instrucción

En este grupo predomina el Secundario Completo (36,5%) e incompleto (17,1%), y le sucede los niveles Universitarios, Completos (15,3%) e Incompletos (12,9%).

```
ej1.1nodo12Datos <- matrix(c(0.018, 0.159, 0.171, 0.365, 0.129, 0.153, 0.006), ncol = 7)
ej1.1nodo12Name<- colnames(ej1.1nodo12Datos) <- c("Primaria Incompleta", "Primaria Completa", "Secundaria Incompleta", "Secundaria Completa", "Superior Universitaria Incompleta", "Superior Universitaria Completa", "Sin Instruccion")
pie(ej1.1nodo12Datos, labels = ej1.1nodo12Name, radius = 1.1, col=rainbow(20), main = "Ej 1.1 - Nodo 12")
```

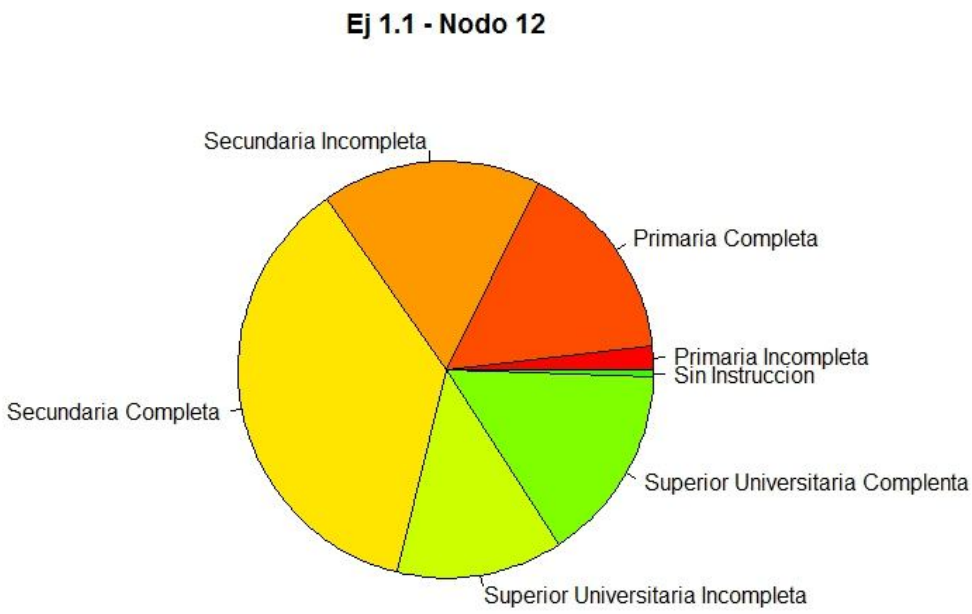


Figura 4. 13 Nodo 12 del Árbol de Decisión del Nivel educativo, con porcentajes de los niveles de Estudio en los Nodos terminales y Variable Cobertura Social Categorizada.

Nodo 13: Nodo 13 del Árbol correspondiente a las Figuras 4.11 y 4.12:

(mayor = TRUE ∧ coberturaSocial = {1, 12} ingresoDeOcupacionPrincipal <= 14700 ∧ ingresoDeOcupacionPrincipal > 3000 ∧ sexo > 1) → n= 139

Mujeres mayores 18 años con Obra Social o bien Obra Social y Mutual o Prepaga o Servicios de Emergencia, con Ingresos de entre $3.000 y $14.700:

= 1 → 1,4% Primaria Incompleta (Incluye Educación Especial)

= 2 → 5% Primaria Completa

= 3 → 6,5% Secundaria Incompleta

= 4 → 29,5% Secundaria Completa

= 5 → 18,7% Superior Universitaria Incompleta

= 6 → 38,8% Superior Universitaria Completa

= 7 → 0% Sin instrucción

De Mujeres que tiene Obra Social, u Obra Social y Mutual o prepaga, o Servicios de emergencia, y que tiene un sueldo de entre $3.000 y $14.700, predomina el Nivel Universitario Completo (38,8%) y el Secundario Completo (29,5%).

```
ej1.1nodo13Datos <- matrix(c(0.014, 0.05, 0.065, 0.295, 0.187, 0.388, 0), ncol = 7)

ej1.1nodo13Name<- colnames(ej1.1nodo13Datos) <- c("Primaria Incompleta", "Primaria Completa", "Secundaria Incompleta", "Secundaria Completa", "Superior Universitaria Incompleta", "Superior Universitaria Completa", "Sin Instruccion")

pie(ej1.1nodo13Datos, labels = ej1.1nodo13Name, radius = 1.1, col=rainbow(20), main = "Ej 1.1 - Nodo 13")
```

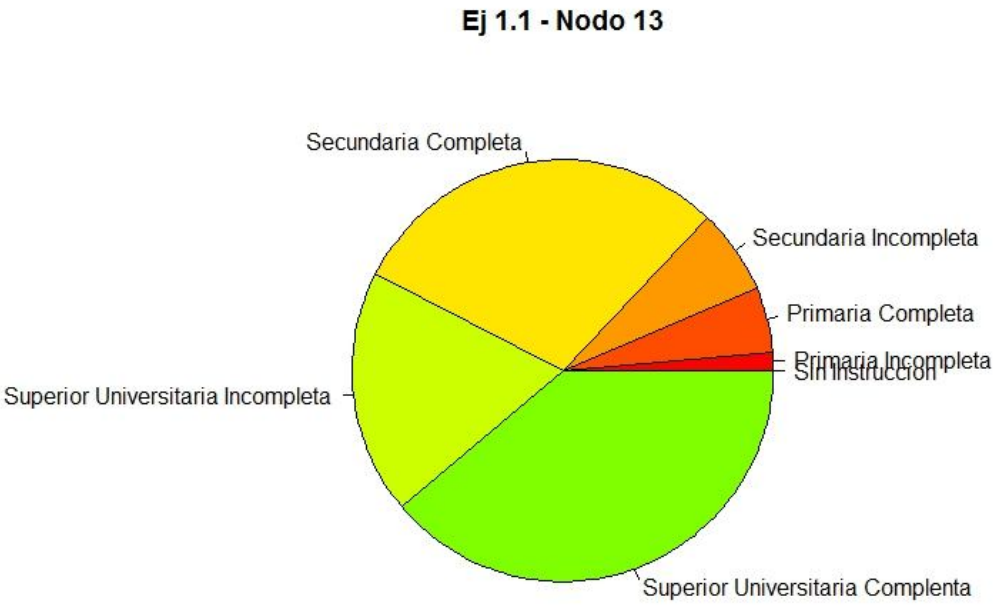


Figura 4. 14 Nodo 13 del Árbol de Decisión del Nivel educativo, con porcentajes de los niveles de Estudio en los Nodos terminales y Variable Cobertura Social Categorizada.

En los Nodos 12 y 13, de las Figuras 4.13 y 4.14,, se pueden corroborar los datos encontrados en el Árbol anterior, respecto al Nivel de Instrucción entre Hombres y Mujeres de iguales características socio-económicas. En este caso se agrega el Dato de la Cobertura Médica-Social.

Se observa un grupo de 170 hombres y 139 mujeres, mayores de 18 años, con sueldos de entre $3.000 y $14.700, que tienen Obra Social, o bien además de la Obra Social tienen acceso a una Prepaga, Mutual o Servicio de Emergencia, pero sin especificar su Dependencia Laboral.

Entre los hombres de este grupo, predomina el Nivel Secundario Completo en un 36,5% (esto es que 62 de los 170); mientras que solo el 15,3% tiene Nivel Superior Universitario Completo (esto es 26 de los 170). Mientras que en el grupo de 139 Mujeres, el 38,8% tiene Nivel Universitario Completo (esto es 54 de las 139).

Se verifican los datos encontrados anteriormente, siendo que en este grupo de 309 personas, con iguales condiciones socio-económicas, el Nivel de Instrucción en Mujeres es

mayor que en Hombres. No solo que entre Mujeres predomina el Nivel Superior Universitario Completo, sino que es superior en cantidad a Hombres. Ya que en este grupo de 309 personas, hay 54 Mujeres y 26 Hombres con este Nivel.

Nodo 14: Nodo 14 del Árbol correspondiente a las Figuras 4.11 y 4.12:

(mayor = TRUE ^ coberturaSocial = {1, 12} ^ ingresoDeOcupacionPrincipal > 14700) → n = 131

Mayores de 18 años con obra social (incluye PAMI), o bien Obra Social y Mutual o Prepaga o Servicio de Emergencia y con ingresos de su ocupación principal mayores a $14.700:

= 1 → 0% Primaria Incompleta (Incluye Educación Especial)

= 2 → 3,1% Primaria Completa

= 3 → 6,1% Secundaria Incompleta

= 4 → 21,4% Secundaria Completa

= 5 → 19,1% Superior Universitaria Incompleta

= 6 → 50,4% Superior Universitaria Completa

= 7 → 0% Sin instrucción

Observamos en este grupo de 131 Hombres mayores de 18 años, cuyos ingresos de su ocupación principal supera los $14.700, pueden pagar su Obra Social, Mutual, Prepaga o Servicio de Emergencia, y su Nivel de Estudio es un 50,4% Superior Universitario Completo; a esto le siguen un 21,4% Secundario Completo, y 19,1% Superior Universitario Incompleto.

```
ej1.1nodo14Datos <- matrix(c(0, 0.031, 0.061, 0.214, 0.191, 0.504, 0), ncol = 7)
ej1.1nodo14Name<- colnames(ej1.1nodo14Datos) <- c("Primaria Incompleta", "Primaria Completa", "Secundaria Incompleta", "Secundaria Completa", "Superior Universitaria Incompleta", "Superior Universitaria Completa", "Sin Instruccion")
pie(ej1.1nodo14Datos, labels = ej1.1nodo14Name, radius = 1.1, col=rainbow(20), main = "Ej 1.1 - Nodo 14")
```

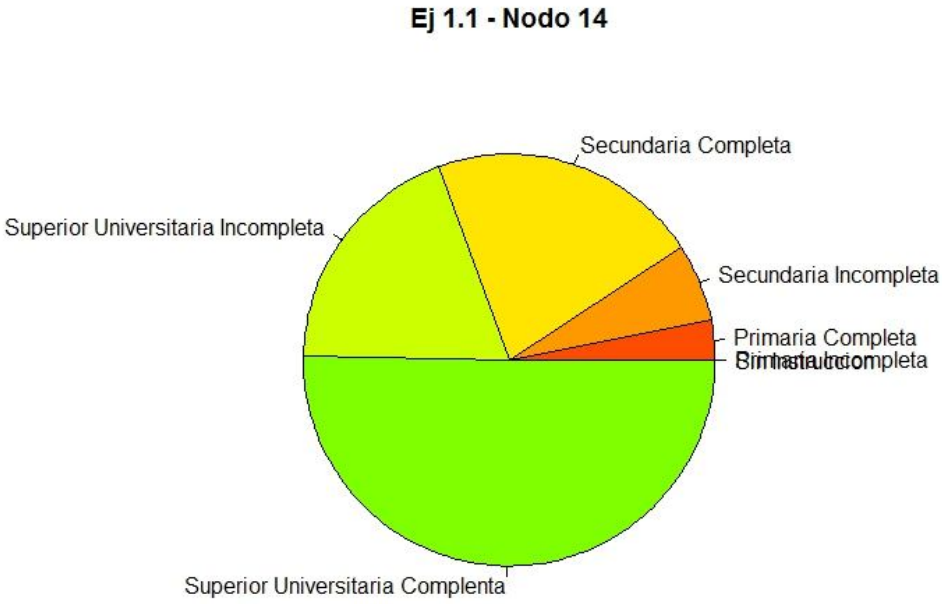


Figura 4. 15 Nodo 14 del Árbol de Decisión del Nivel educativo, con porcentajes de los niveles de Estudio en los Nodos terminales y Variable Cobertura Social Categorizada.

Nodo 16: Nodo 16 del Árbol correspondiente a las Figuras 4.11 y 4.12:

(mayor = TRUE ∧ coberturaSocial = {2, 3, 4} ∧ coberturaSocial = {3}) → n = 35

Mayores de 18 años con Planes y Seguros Públicos:

= 1 → 20% Primaria Incompleta (Incluye Educación Especial)

= 2 → 28,6% Primaria Completa

= 3 → 22,9% Secundaria Incompleta

= 4 → 20% Secundaria Completa

= 5 → 2,9% Superior Universitaria Incompleta

= 6 → 0% Superior Universitaria Completa

= 7 → 5,7% Sin instrucción

En este grupo de 35 personas encuestadas, tienen Planes y Seguros Públicos, y su nivel de Instrucción predominante es Primario y Secundario; 22,9% Secundario Incompleto, 28,6% Primaria Completa, 20% Secundaria Completa y 20% Primaria incompleta. El nivel Universitario Incompleto es excesivamente baja, y no existe Nivel Universitario Completo.

```
ej1.1nodo16Datos <- matrix(c(0.2, 0.286, 0.229, 0.2, 0.029, 0, 0.057), ncol = 7)

ej1.1nodo16Name<- colnames(ej1.1nodo16Datos) <- c("Primaria Incompleta", "Primaria Completa", "Secundaria Incompleta", "Secundaria Completa", "Superior Universitaria Incompleta", "Superior Universitaria Completa", "Sin Instrucción")

pie(ej1.1nodo16Datos, labels = ej1.1nodo16Name, radius = 1.1, col=rainbow(20), main = "Ej 1.1 - Nodo 16")
```

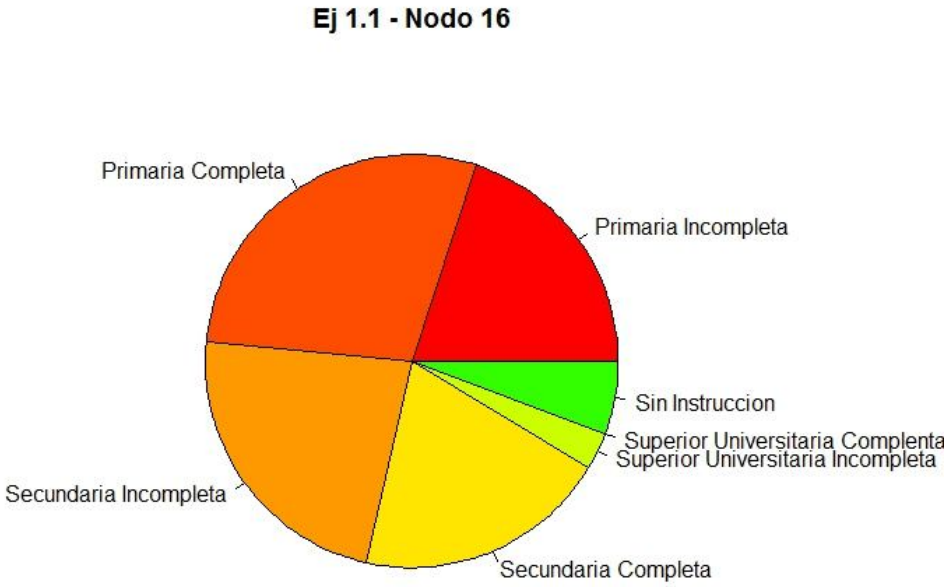


Figura 4. 16 Nodo 16 del Árbol de Decisión del Nivel educativo, con porcentajes de los niveles de Estudio en los Nodos terminales y Variable Cobertura Social Categorizada.

Conclusión:

Se puede concluir que:

a) En un determinado sector social, predominan coberturas sociales a través de Emergencias Médicas o Planes o Seguros Sociales.
b) A mayor Nivel de Instrucción, los recursos económicos son más elevados que un sueldo básico, y tienen a su vez mayor posibilidad de acceder a una mejor Cobertura Médico-Social.
c) En este Árbol se verifica la información encontrada en el Árbol anterior, sobre el Nivel de Instrucción; siendo que entre personas de iguales características Socio-Económicas, se ha encontrado que es mayor el número de Mujeres con Nivel Superior Universitario Completo que entre Hombres.

4.2.2.3 Tabla individuos: Análisis del tipo de Cobertura Médico-Social (Tabla Individuos):

"Caso 2 – Variable Objetivo: Cobertura social (**ch08**) en función del sexo (**ch04**), edad (**ch06**), Nivel Educativo (**NIVEL_ED**), Dependencia Laboral (**PP04A**), Ingreso de Ocupación Principal (**P21**), Ingreso de otras ocupaciones (**Tot_p12**)"

Factorizamos las variables “nivelEducativo”, “sexo”, y “dependenciaLaboral”.

Variable Objetivo:

coberturaSocial <- as.factor(CH08)

Variables predictoras:

sexo <- as.factor(CH04)

mayor <- as.factor(CH06 >= 18)

nivelEducativo <- as.factor(NIVEL_ED)

dependenciaLaboral <- as.factor(PP04A)

ingresoDeOcupacionPrincipal <- P21

ingresosDeOtrasOcupaciones <- TOT_P12

Fórmula:

individuos.formula <- coberturaSocial ~ sexo + mayor + nivelEducativo + dependenciaLaboral + ingresoDeOcupacionPrincipal + ingresosDeOtrasOcupaciones

Modelo:

individuos.modelo <- ctree(individuos.formula, data=cbind(mayor,individuos.00))

Gráfico:

plot(individuos.modelo)

El modelo resultante se observa en la Figura 4.17:

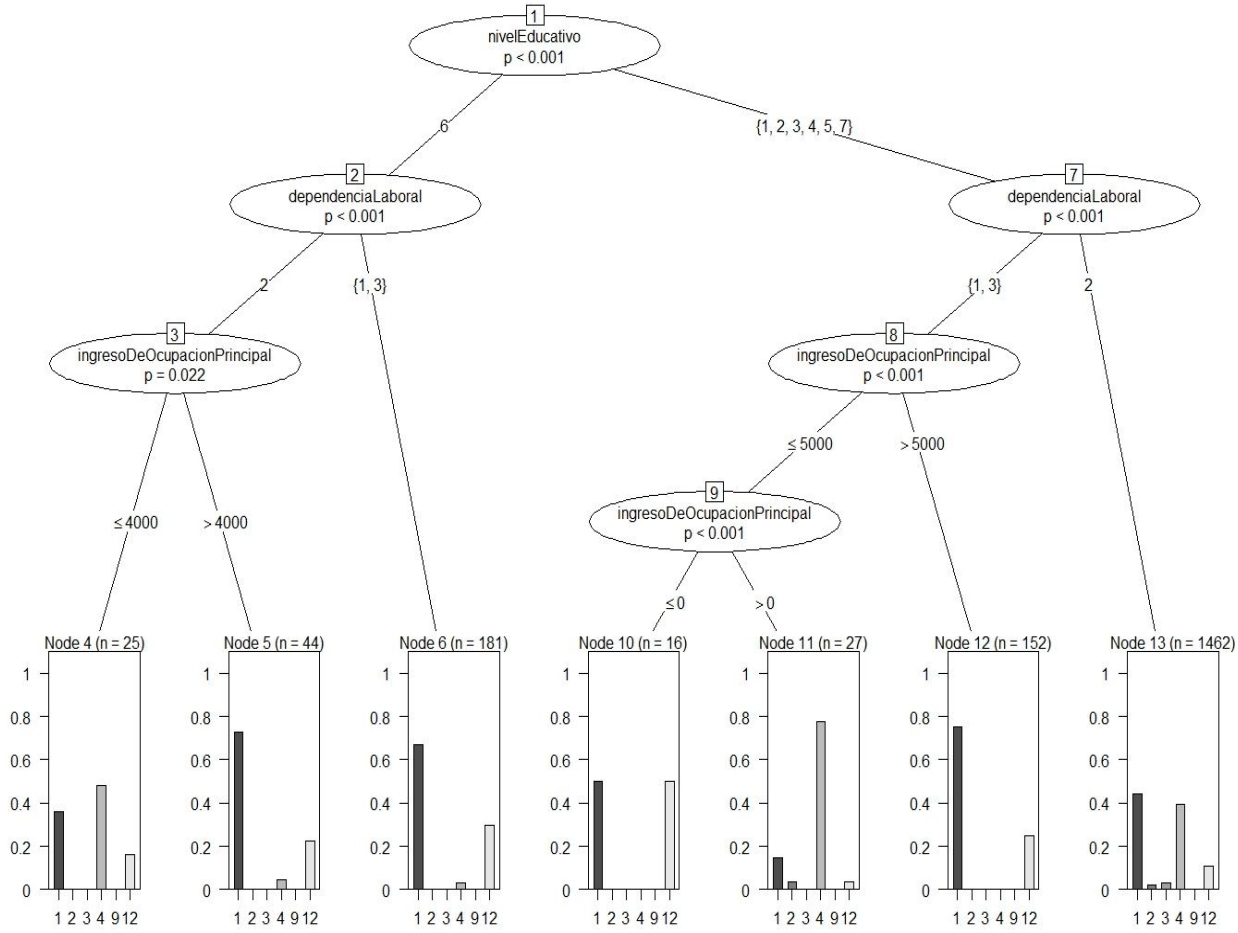


Figura 4. 17 Árbol de Decisión del Tipo de Cobertura social en función del sexo, edad, Nivel Educativo, Dependencia Laboral, Ingreso de Ocupación Principal, Ingreso de otras ocupaciones. Variables nivelEducativo, sexo y dependenciaLaboral categorizadas.

Una variante del árbol se observa en la Figura 4.18. En los nodos terminales presenta los porcentajes en números en lugar de un gráfico estadístico. Esto, mediante la siguiente sentencia:

plot(individuos.modelo, type="simple")

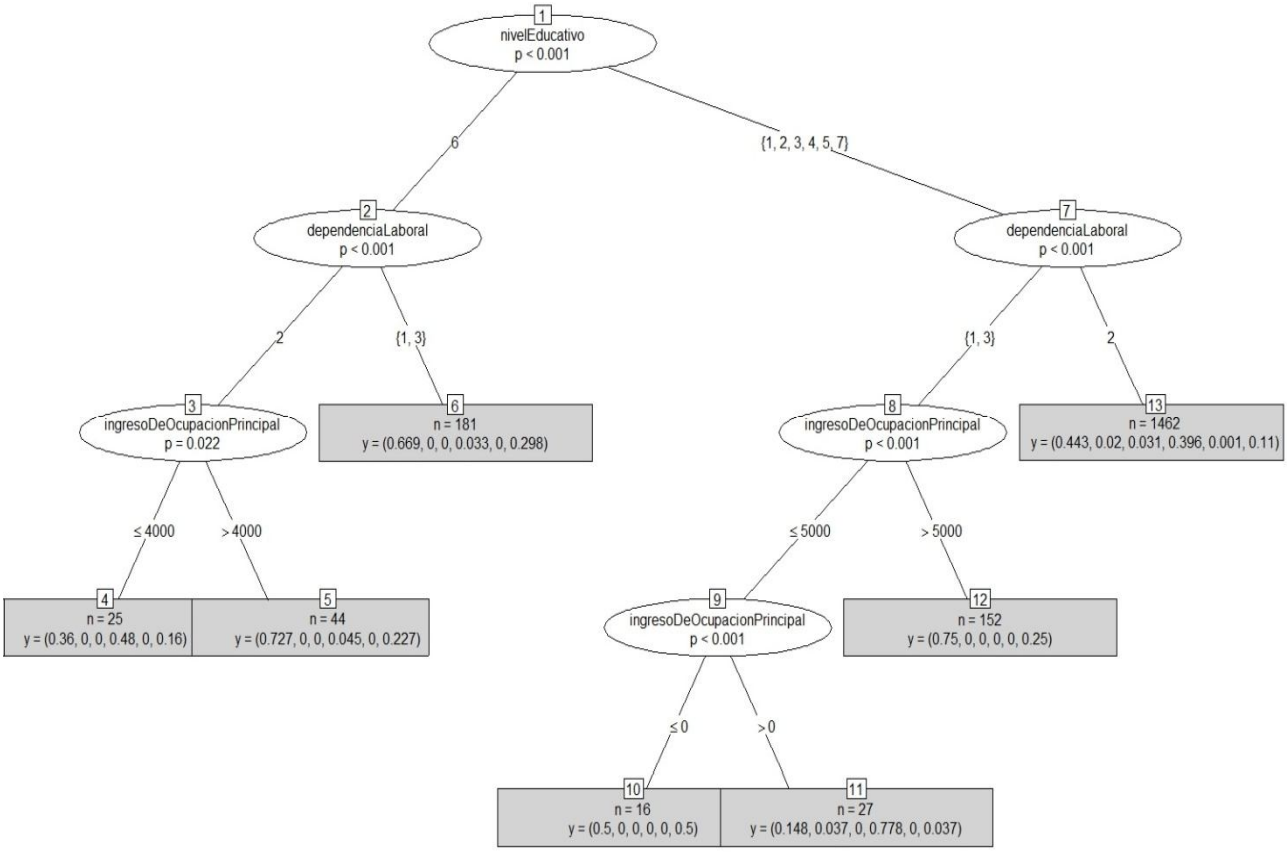


Figura 4. 18 Árbol de Decisión del Tipo de Cobertura social con porcentajes de los niveles de Estudio en los Nodos terminales. Variables nivelEducativo, sexo y dependenciaLaboral categorizadas.

A continuación los significados de los atributos presentes en el Árbol:

nivelEducativo:

1= Primaria Completa (incluye educación especial).

2= Primaria Incompleta.

3= Secundaria Incompleta.

4= Secundaria Completa.

5= Superior Universitaria Incompleta.

6= Superior Universitaria Completa.

7= Sin Instrucción.

9= Ns. / Nr.

dependenciaLaboral

= 1 → Estatal

= 2 → Privado

= 3 → Otro tipo

CoberturaSocial:

1 = Obra Social (Incluye PAMI).

2= Mutual / Prepaga / Servicio de Emergencia.

3= Planes y Seguros Sociales.

4= No paga ni le descuentan.

9= Ns. / Nr.

12= Obra Social y mutual / Prepaga / Servicio de Emergencia.

A continuación se procede a analizar los nodos de mayor importancia. Se analizan los que presentan características más sobresalientes. Descartándose de esta forma, nodos con muestras muy pequeñas que no influyen en la población; tampoco se tienen en cuenta los nodos correspondientes a menores de edad, dado que se busca encontrar características socio-económicas que correspondan principalmente a los Ingresos, y como esto influye en el Nivel de Educación, Cobertura Social, entre otros.

La descripción del Nodo corresponde a conjunciones lógicas de acuerdo a las bifurcaciones del Árbol. El camino desde el nodo raíz al nodo terminal, será expresado

mediante estas conjunciones lógicas. Por empleo: (atributoA = 1 ^ atributoB > 2 ^ atributoC < 3). Seguidamente se expresa la cantidad de individuos mediante la conjunción: n = xx; por ejemplo, n = 22, quiere decir que ese Nodo representa a un grupo de 22 personas. Teniendo en cuenta esto, podemos interpretar esto como: Un grupo de 22 personas donde el Atributo A es igual 1, el Atributo B es mayor a 2 y el atributo C es menor a 3. Y de acuerdo a los valores de los atributos y sus significados se procede a interpretar las características socio-económicas de cada Nodo. Gracias al Árbol de la figura 4.18, podemos acceder a los porcentajes concernientes, y con estos valores, mediante unas sentencias en R, se obtienen gráficas circulares para una mayor apreciación de los datos encontrados.

Ahora, se procede a analizar los Nodos:

Nodo 4: Nodo 4 del Árbol correspondiente a las Figuras 4.17 y 4.18:

(nivelEducativo = 6 ^ dependenciaLaboral = 2 ^ ingresoDeOcupacionPrincipal <= 4000) → n = 25

Personas con nivel de Estudio Superior Universitario Completo y que trabajan en el sector Privado, con sueldos inferiores a los $4.000:

1 = 36% Obra Social (Incluye PAMI).

2= 0% Mutual / Prepaga / Servicio de Emergencia.

3= 0% Planes y Seguros Sociales.

4= 48% No paga ni le descuentan.

9= 0% Ns. / Nr.

12= 16% Obra Social y mutual / Prepaga / Servicio de Emergencia.

En este grupo de 25 personas, sin distinguir sexo o edad, que trabajan en el sector Privado, con sueldos inferiores a los $4.000, y que tienen Estudios Terciarios o Universitarios Completos, el 48% no paga ni le descuentan por Cobertura Médico-Social, mientras que el 36% tiene solo Obra Social, incluyendo PAMI, y solo el 16% restante tiene acceso a Obra Social más Mutual o Prepaga o Servicio de Emergencias.

```
ej2.1nodo4Datos <- matrix(c(0.36, 0, 0, 0.48, 0, 0.16), ncol = 6)
```

```
ej2.1nodo4Name<- colnames(ej2.1nodo4Datos) <- c("Obra Social (Incluye PAMI)", "Mutual / Prepaga / Servicio de Emergencia.", "Planes y Seguros Sociales.", "No paga ni le descuentan.", "Ns. / Nr.", "Obra Social y mutual / Prepaga / Servicio de Emergencia.")
```

pie(ej2.1nodo4Datos, labels = ej2.1nodo4Name, radius = 1.1, col=rainbow(20), main = "Ej 2.1 - Nodo 4")

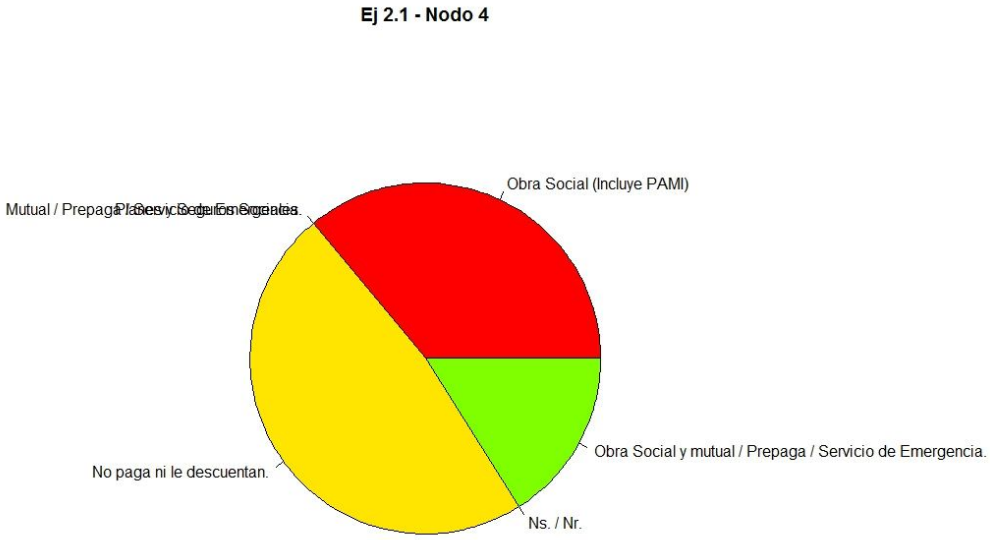


Figura 4. 19 Nodo 4 del Árbol de Decisión del Tipo de Cobertura social con porcentajes de los niveles de Estudio en los Nodos terminales. Variables nivelEducativo, sexo y dependenciaLaboral categorizadas.

Nodo 5: Nodo 5 del Árbol correspondiente a las Figuras 4.17 y 4.18:

(nivelEducativo = 6 ∧ dependenciaLaboral = 2 ∧ ingresoDeOcupacionPrincipal > 4000) → n = 44

Personas cuyo nivel de Estudio es Superior Universitario Completo, que trabajan en el sector Privado, con ingresos superiores a los $4.000, y:

1 = 72,7% Obra Social (Incluye PAMI).

2= 0% Mutual / Prepaga / Servicio de Emergencia.

3= 0% Planes y Seguros Sociales.

4= 4,5% No paga ni le descuentan.

9= 0% Ns. / Nr.

12= 22,7% Obra Social y mutual / Prepaga / Servicio de Emergencia.

En este grupo de 44 personas que trabajan en el sector privado, con sueldos superiores a los $4.000, con nivel Educativo Superior Universitario Completo, el 72,7% solo tiene Obra Social, incluyendo PAMI, mientras que un 22,7% además de Obra Social tiene

acceso a una Mutual o Prepaga o Servicio d Emergencia, y el 4,5% restante no paga ni le descuentan por cobertura Médico-Social.

ej2.1nodo5Datos <- matrix(c(0.727, 0, 0, 0.045, 0, 0.227), ncol = 6)

ej2.1nodo5Name<- colnames(ej2.1nodo5Datos) <- c("Obra Social (Incluye PAMI)", "Mutual / Prepaga / Servicio de Emergencia.", "Planes y Seguros Sociales.", "No paga ni le descuentan.", "Ns. / Nr.", "Obra Social y mutual / Prepaga / Servicio de Emergencia.")

pie(ej2.1nodo5Datos, labels = ej2.1nodo5Name, radius = 1.1, col=rainbow(20), main = "Ej 2.1 - Nodo 5")

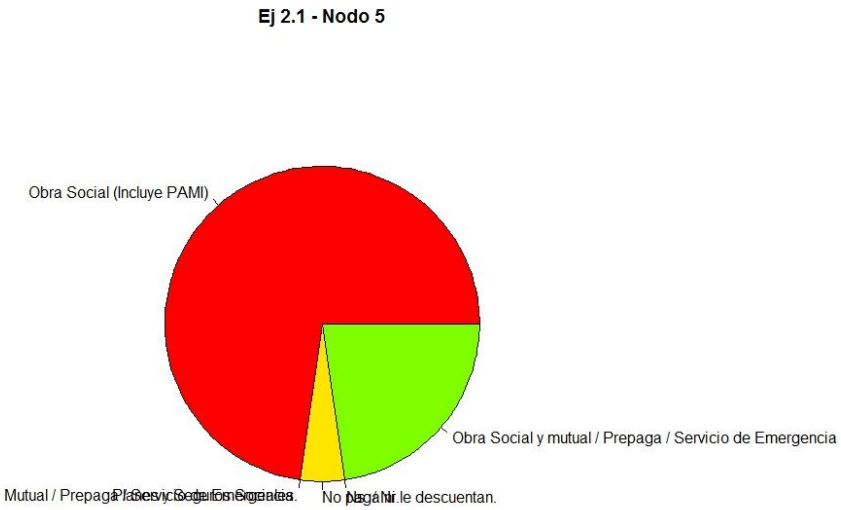


Figura 4. 20 Nodo 5 del Árbol de Decisión del Tipo de Cobertura social con porcentajes de los niveles de Estudio en los Nodos terminales. Variables nivelEducativo, sexo y dependenciaLaboral categorizadas.

Nodo 6: Nodo 6 del Árbol correspondiente a las Figuras 4.17 y 4.18:

(nivelEducativo = 6 ∧ dependenciaLaboral = {1, 3}) → n = 181

Personas con nivel de Estudios Terciarios o Universitarios Completos, que trabajan en el Estado (1), u otro sector (3):

1 = 66,9% Obra Social (Incluye PAMI).

2= 0% Mutual / Prepaga / Servicio de Emergencia.

3= 0% Planes y Seguros Sociales.

4= 3,3% No paga ni le descuentan.

9= 0% Ns. / Nr.

12= 29,8% Obra Social y mutual / Prepaga / Servicio de Emergencia.

En este grupo de 181 personas, con nivel de Estudios Superior Universitario Completo, que trabajan en el Estado u otro sector, el 66,9% solo tiene Obra Social, incluyendo PAMI, mientras que el 29,8% además de Obra Social, tiene acceso a Mutual, o Prepaga, o Servicio de Emergencia; y el 3,3% restante no paga ni le descuentan por Cobertura Médico-Social.

```
ej2.1nodo6Datos <- matrix(c(0.669, 0, 0, 0.033, 0, 0.298), ncol = 6)
ej2.1nodo6Name<- colnames(ej2.1nodo6Datos) <- c("Obra Social (Incluye PAMI)", "Mutual / Prepaga / Servicio de Emergencia.", "Planes y Seguros Sociales.", "No paga ni le descuentan.", "Ns. / Nr.", "Obra Social y mutual / Prepaga / Servicio de Emergencia.")

pie(ej2.1nodo6Datos, labels = ej2.1nodo6Name, radius = 1.1, col=rainbow(20), main = "Ej 2.1 - Nodo 6")
```

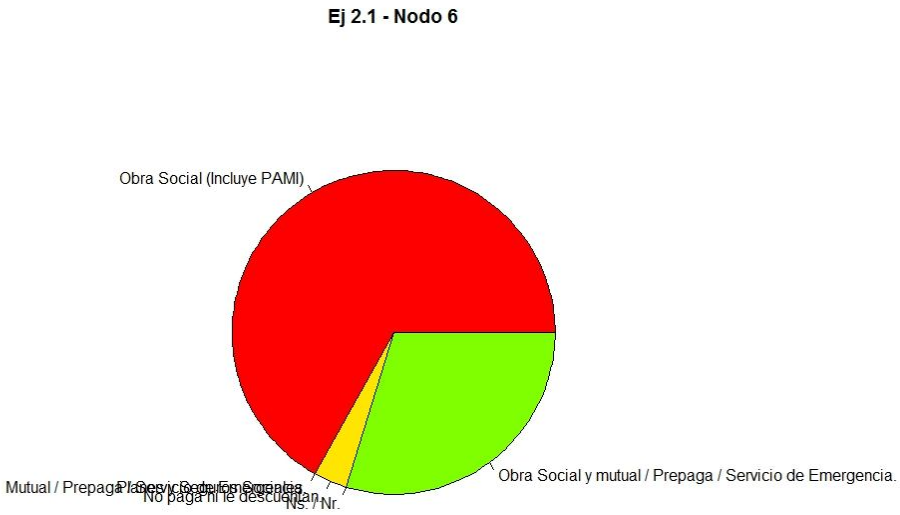


Figura 4. 21 Nodo 6 del Árbol de Decisión del Tipo de Cobertura social con porcentajes de los niveles de Estudio en los Nodos terminales. Variables nivelEducativo, sexo y dependenciaLaboral categorizadas.

Nodo 12: Nodo 12 del Árbol correspondiente a las Figuras 4.17 y 4.18:

(nivelEducativo = {1, 2, 3, 4, 5, 7} ^ dependenciaLaboral = {1, 3} ^ ingresoDeOcupacionPrincipal > 5000^) → n = 152

Personas con nivel de Estudio variado (solo el nivel Superior Universitario está ausente), que trabajan en el Estado u otros sectores (excluyendo el sector Privado), con sueldo superiores a los $5.000,:

1 = 75% Obra Social (Incluye PAMI).

2= 0% Mutual / Prepaga / Servicio de Emergencia.

3= 0% Planes y Seguros Sociales.

4= 0% No paga ni le descuentan.

9= 0% Ns. / Nr.

12= 25% Obra Social y mutual / Prepaga / Servicio de Emergencia.

En este grupo de 152 personas que trabajan en el Estado u otros sectores, a excepción del sector Privado, con ingresos de su ocupación principal superiores a los $5.000, y con un nivel de Estudio variado (excluyendo al nivel Superior Universitario Completo), el 75% solo tiene Obra Social, incluyendo PAMI, y el 25% restante, además d tener acceso además de su Obra Social, a una Mutual o Prepaga o Servicio de Emergencia.

```
ej2.1nodo12Datos <- matrix(c(0.75, 0, 0, 0, 0, 0.25), ncol = 6)
ej2.1nodo12Name<- colnames(ej2.1nodo12Datos) <- c("Obra Social (Incluye PAMI)", "Mutual / Prepaga / Servicio de Emergencia.", "Planes y Seguros Sociales.", "No paga ni le descuentan.", "Ns. / Nr.", "Obra Social y mutual / Prepaga / Servicio de Emergencia.")
pie(ej2.1nodo12Datos, labels = ej2.1nodo12Name, radius = 1.1, col=rainbow(20), main = "Ej 2.1 - Nodo 12")
```

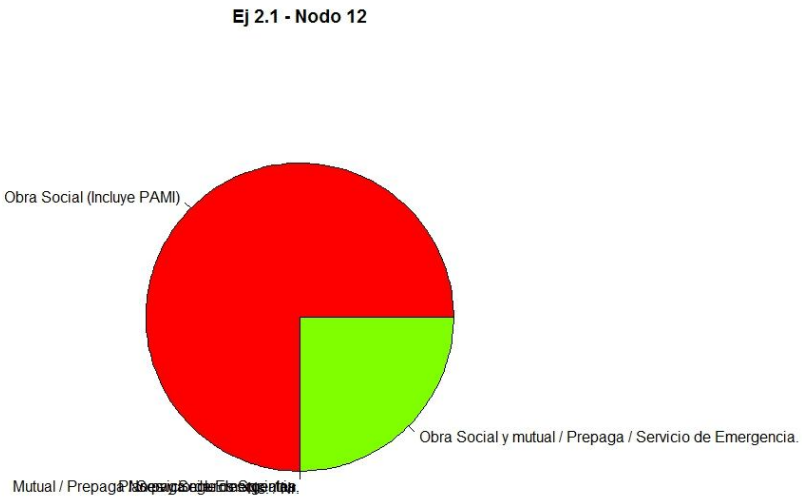


Figura 4. 22 Nodoo 12 del Árbol de Decisión del Tipo de Cobertura social con porcentajes de los niveles de Estudio en los Nodos terminales. Variables nivelEducativo, sexo y dependenciaLaboral categorizadas.

Nodo 13: Nodo 13 del Árbol correspondiente a las Figuras 4.17 y 4.18:

(nivelEducativo = {1, 2, 3, 4, 5, 7} ∧ dependenciaLaboral = 2) → n = 1462

Personas con un nivel de Estudio variado, sin contar con el nivel Superior Universitario Completo y que traban en el Sector Privado,:

1 = 44,3% Obra Social (Incluye PAMI).

2= 2% Mutual / Prepaga / Servicio de Emergencia.

3= 3,1% Planes y Seguros Sociales.

4= 39,6% No paga ni le descuentan.

9= 0,1% Ns. / Nr.

12= 11% Obra Social y mutual / Prepaga / Servicio de Emergencia.

En este grupo de 1.462 personas, del total de 1.907 encuestados, observamos que trabajan en el sector privado, y hay un nivel de estudio muy variado, pero siendo que no hay nivel Superior Universitario Completo; el 44,3% solo tiene Obra Social, incluyendo PAMI, el 39,6% no paga ni le descuentan por cobertura Médico-Social; el 3,1% tiene Planes y Seguros Sociales, el 11% tiene acceso, además de su Obra Social a una Mutual, Prepaga o Servicio de Emergencia y solo el 2% solo tiene acceso a Mutual, o Prepaga o Servicio de Emergencia.

```
ej2.1nodo13Datos <- matrix(c(0.443, 0.02, 0.031, 0.396, 0.001, 0.11), ncol = 6)

ej2.1nodo13Name<- colnames(ej2.1nodo13Datos) <- c("Obra Social (Incluye PAMI)", "Mutual / Prepaga / Servicio de Emergencia.", "Planes y Seguros Sociales.", "No paga ni le descuentan.", "Ns. / Nr.", "Obra Social y mutual / Prepaga / Servicio de Emergencia.")

pie(ej2.1nodo13Datos, labels = ej2.1nodo13Name, radius = 1.1, col=rainbow(20), main = "Ej 2.1 - Nodo 13")
```

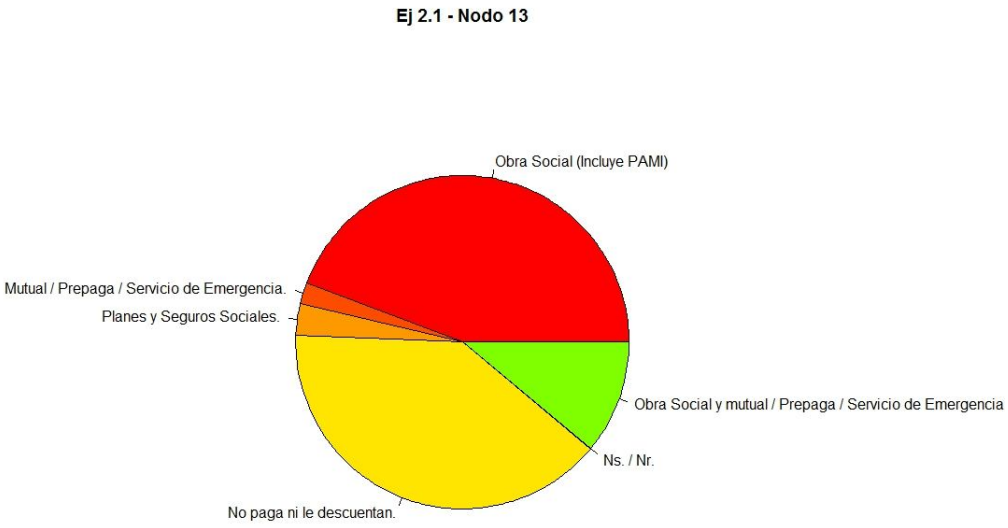


Figura 4. 23 Nodo 13 del Árbol de Decisión del Tipo de Cobertura social con porcentajes de los niveles de Estudio en los Nodos terminales. Variables nivelEducativo, sexo y dependenciaLaboral categorizadas.

Conclusión:

En cuanto a la Cobertura Médico-Social, se observa que:

- En personas que trabajan en el Sector Privado, con sueldos inferiores a $4.000 y Nivel de Estudios Superior Universitario Completo, no hay ´porcentajes referente a Planes Sociales, sino que la mitad de este grupo poblacional tiene algún tipo de Cobertura y la otra mitad no paga ni le descuentan.
- Se observa que en un grupo de personas que trabajan en el sector privado con sueldos mayores a $4.000 y Nivel de Estudios Superior Universitario Completo,, aproximadamente el 95% tiene algún tipo de cobertura, mientras que solo el 5% no paga ni le descuentan.
- En personas que trabajan en el Estado u otro sector que o sea el Privado, con Nivel Superior Universitario Completo, un 97% aproximadamente tiene algún tipo de cobertura y el resto no paga ni le descuentan.
- Se puede observar que en un grupo poblacional que trabaja en el Estado u otro sector que no sea el privado, con sueldos superiores a los $5.000 y con un nivel estudio variado (excluyendo al Nivel Superior Universitario Completo), el 100% tiene algún tipo de Cobertura Médico-Social.
- En el grupo más numeroso, se observa que trabaja en el sector Privado, y con un nivel de Estudio variado (excluyendo al Nivel Superior Universitario Completo), el porcentaje de Planes y Seguros Sociales es bajo, cerca de un

40% no paga ni le descuentan y el resto tiene algún tipo de Cobertura.

- Podemos concluir que a Mayor Nivel de Estudios, mayor es la posibilidad de acceder a trabajos con sueldos sustentables (teniendo en cuenta que estos números corresponden al primer trimestre del año 2017), y esto permite tener acceso a una Cobertura Médico Social; la mayoría de los casos son porque en el mismo lugar de trabajo les realiza los descuentos por cobertura Social, pero también hay quienes tienen la posibilidad de acceder a una pre-paga, servicios de Emergencia u otros.

4.3 Aprendizaje No Supervisado – Clustering:

En el Capítulo 3 se abordó en forma Teórica esta Técnica del Aprendizaje Automatizado No Supervisado, llamada Clustering. Con esta Técnica lo que buscamos es agrupar sectores de la muestra Poblacional en Clusters. De esta forma se obtendrán grupos (clusters) de acuerdo a las condiciones socio-económicas y se podrá clasificar o discriminar la Población de acuerdo a estas características.

En este estudio se trabajará con la Base de Datos “EPH” (tabla unificada), de modo que se cuenta con los datos tanto de Hogares como de Individuos. Luego de cargar la Base de Datos en el RStudio se escalan las variables (columnas) con las que se hará el análisis; paso siguiente se aplicará el Algoritmo “K-medois” y luego se elije la cantidad de Clusters con los que se desea trabajar y el tipo de Métrica, en este caso Euclídea.

Para continuar con el análisis, se procederá a estudiar los centroides de los Clusters. Mediante sentencias en el lenguaje R, obtendremos la fila del individuo (centroide), y los valores de las variables de estudio. El centroide lo que nos indicará son las características socio-económicas medias de cada Cluster. Luego también se podrá obtener la cantidad de individuos que hay en cada grupo.

Con esta información, se podrá describir cada Cluster, y también se procederá a analizar el grupo familiar de cada Centroide. Por último también hará un análisis de algunos gráficos de dos variables obtenidos en RStudio luego de este estudio.

4.3.1 Fase de Evaluación

Construcción y evaluación del Modelo:

Seleccionada la técnica, cargamos la tabla EPH (tabla unificada), haciendo uso de la herramienta RStudio con las siguientes sentencias en lenguaje R:

```
eph = read.csv("C:/EPH/Completo2.csv", header = T, sep = ";")
```

Con la sentencia “attach”, podemos trabajar individualmente con cada columna de la tabla:

```
attach(eph)
```

Seguidamente, cargamos la librería “cluster” para hacer uso del algoritmo:

```
library(cluster)
```

4.3.2 Resultados

4.3.2.1 Tabla individuos: Análisis del Nivel de Estudio

Prueba 1:

Variables de estudio:

"10= IV1= tipo de hogar; 65=IX_TOT=Cant de habitantes por hogar; 68=ITF = Ingreso Total Familiar; 104=CH08= Cobertura Medico Social; 115=Nivel de educación; 140= Dependencia Laboral; 221=P21=Ingreso actividad principal; 229=Tot_12=ingresos de otras ocupaciones"

Se escalan las variables:

```
> eph.scale  <-  scale(eph[,c(10, 65, 68,104, 115, 140,221,229)])
```

Se aplica el algoritmo k-medois:

```
> eph.kmedois <- rbind(cbind(eph[,c(10, 65, 68, 104, 115, 140,221,229)]))
```

Paso siguiente se elige la cantidad de Clusters y la Métrica Euclídea:

```
> pamx <- pam(eph.kmedois, 4, metric = "euclidean", stand = TRUE):
```

Ahora se grafican todas las combinaciones posibles de a 2 variables, en eje horizontal y vertical. Cada color de los puntos representan los clusters al que ese ejemplo pertenece. Más adelante, se visualizará por separado un caso de todos ellos y se interpretará el significado del mismo.

```
> plot(eph.kmedois, col=pamx$clustering)
```

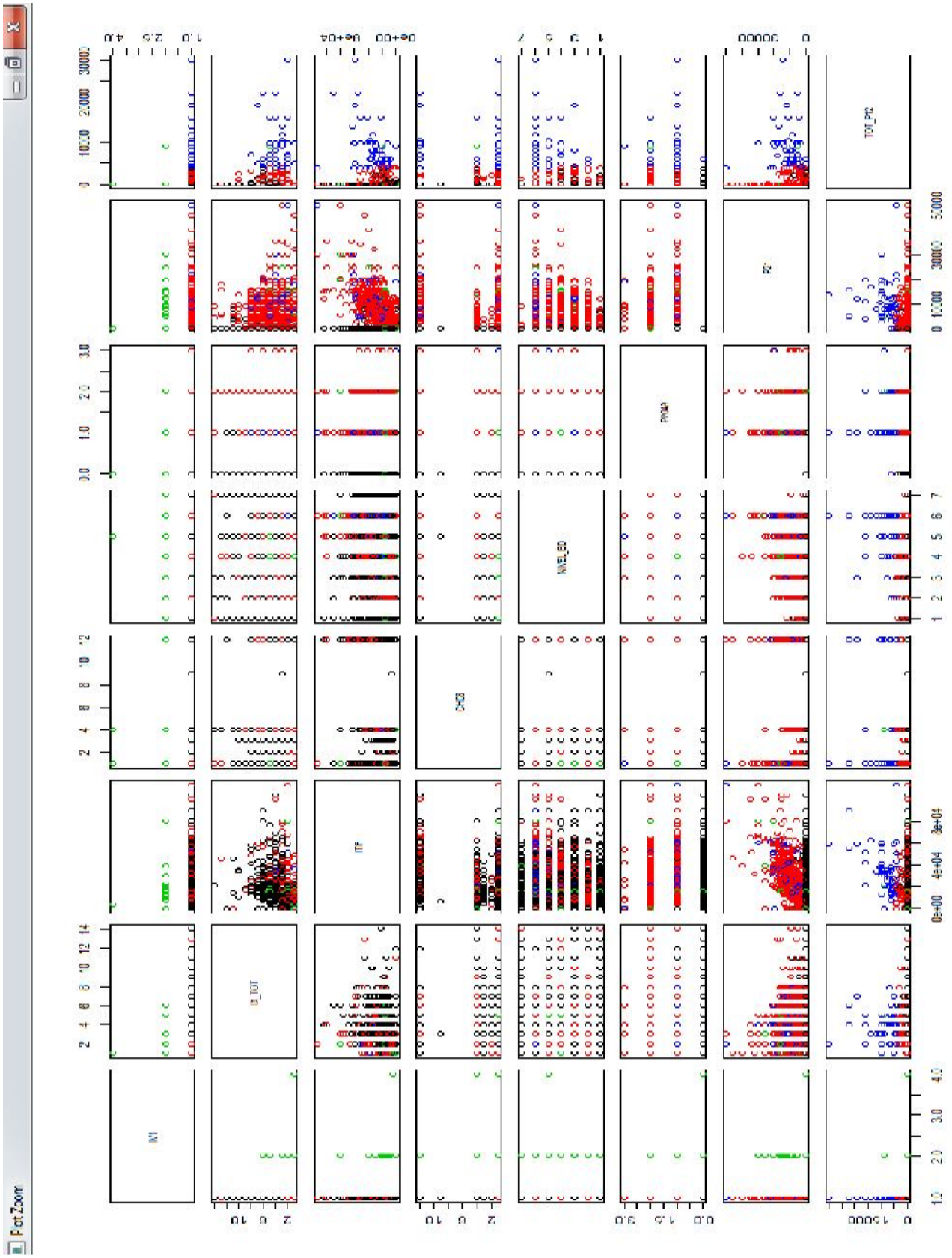


Figura 4. 24 Gráfica de Clustering con todas las variables en estudio.

Procedemos a visualizar los Centroides de los Clusters con los valores de las Variables en estudio:

```
> pamx$medoids
```

	IV1	IX_TOT	ITF	CH08	NIVEL_ED	PP04A	P21	TOT_P21
474	1	5	15910	3	3	0	0	0
1502	1	4	21000	4	4	2	6000	0
1915	2	5	15720	1	4	0	0	0
687	1	4	28000	1	6	1	8000	8000

Tabla 4. 1 Centroides de los Clusters con sus características.

Ahora se obtiene la cantidad de Individuos por Cluster:

```
> summary(as.factor(pamx$clustering))
```

1	2	3	4
1127	703	58	52

Tabla 4. 2 Cantidad de Individuos por Cluster

Análisis de los Clusters:

De acuerdo a los Datos Obtenidos Anteriormente, se pueden determinar las siguientes características de cada Cluster:

Cluster 1: 1.127 individuos.

Características medias:

Tipo de Vivienda: Casa

Cantidad de habitantes del hogar: 5

Ingreso Total Familiar: $15.910

Cobertura Médico-Social: Planes y Seguros Públicos.

Nivel Educativo: Secundario Incompleto.

Dependencia Laboral: Ninguna.

Ingreso de Actividad Principal: $0

Ingresos de Otras Actividades: $0

Cluster 2: 703 individuos.

Características medias:

Tipo de Vivienda: Casa

Cantidad de habitantes del hogar: 4

Ingreso Total Familiar: $21.000

Cobertura Médico-Social: No paga ni le descuentan.

Nivel Educativo: Secundario Completo.

Dependencia Laboral: Privada.

Ingreso de Actividad Principal: $6.000

Ingresos de Otras Actividades: $0

Cluster 3: 58 individuos.

Características medias:

Tipo de Vivienda: Departamento

Cantidad de habitantes del hogar: 5

Ingreso Total Familiar: $15.720

Cobertura Médico-Social: Obra Social (Incluye PAMI).

Nivel Educativo: Secundario Completo.

Dependencia Laboral: Ninguna.

Ingreso de Actividad Principal: $0

Ingresos de Otras Actividades: $0

Cluster 4: 52 individuos.

Características medias:

Tipo de Vivienda: Casa

Cantidad de habitantes del hogar: 4

Ingreso Total Familiar: $28.000

Cobertura Médico-Social: Obra Social (Incluye PAMI).

Nivel Educativo: Universitario Completo.

Dependencia Laboral: Estatal.

Ingreso de Actividad Principal: $8.000

Ingresos de Otras Actividades: $8.000

Paso siguiente, visualizamos los resultados:

```
> resultado <- cbind(eph.kmedois, pamx$clustering)
> resultado
```

Estas sentencias visualizará toda la tabla, con las variables elegidas y el cluster al que pertenece cada individuo. Al trabajar con la Tabla unificada (EPH), ordenando los resultados por el CODUSU, en la lista se mostraran los grupos Familiares. Para facilitar el análisis de los clusters, se estudiarán los grupos Familiares de os Centroides de cada Cluster.

(En las tablas siguientes, los centroides están en Negrita y pintados de naranja)

	IV1	IX_TOT	ITF	CH08	NIVEL_ED	PP04A	P21	TOT_P12	pamx$clustering
473	1	5	15910	4	3	2	800	0	2
474	**1**	**5**	**15910**	**3**	**3**	**0**	**0**	**0**	**1**
475	1	5	15910	1	4	2	10000	0	2
476	1	5	15910	4	6	2	800	0	2
477	1	5	15910	4	5	0	0	0	1
687	**1**	**4**	**28000**	**1**	**6**	**1**	**8000**	**8000**	**4**
688	1	4	28000	1	4	2	12000	0	2
689	1	4	28000	1	3	0	0	0	1
690	1	4	28000	1	3	0	0	0	1
1502	**1**	**4**	**21000**	**4**	**4**	**2**	**6000**	**0**	**2**
1503	1	4	21000	1	6	1	10000	0	2
1504	1	4	21000	1	1	0	0	0	1
1505	1	4	21000	1	7	0	0	0	1
1914	2	5	15720	1	4	1	15000	0	3
1915	**2**	**5**	**15720**	**1**	**4**	**0**	**0**	**0**	**3**
1916	2	5	15720	1	5	0	0	0	3
1917	2	5	15720	1	3	0	0	0	3
1918	2	5	15720	1	1	0	0	0	3

Tabla 4. 3 Grupos Familiares de los Centroides de cada Cluster.

En la tabla 4.3 se observan los grupos familiares de los centroides y las variables de estudio con sus respectivos valores. Esto es: "IV1= tipo de hogar; IX_TOT=Cantidad de habitantes por hogar; ITF = Ingreso Total Familiar; CH08= Cobertura Medico Social; NIVEL_ED = Nivel de educación; PP04A = Dependencia Laboral; P21=Ingreso actividad principal; Tot_12=ingresos de otras ocupaciones y pamx$clustering = Cluster al que pertenece el individuo.

El primer grupo familiar corresponde al centroide del Cluster 1 (fila 474) que se muestra en la tabla 4.4, el segundo grupo familiar, que se muestra en la tabla 4.7, pertenece al del centroide del Cluster 4, el tercer grupo, que se muestra en la tabla 4.5, corresponde al centroide del Cluster 2 y el último grupo, que se observa en la tabla 4.6, al del centroide del Cluster 3.

Paso siguiente, se analizarán cada uno de estos grupos Familiares.

Análisis de los grupos Familiares de los Centroides:

Grupo Familiar del Centroide del Cluster 1:

	IV1	IX_TOT	ITF	CH08	NIVEL_ED	PP04A	P21	TOT_P12	pamx$clustering
473	1	5	15910	4	3	2	800	0	2
474	**1**	**5**	**15910**	**3**	**3**	**0**	**0**	**0**	**1**
475	1	5	15910	1	4	2	10000	0	2
476	1	5	15910	4	6	2	800	0	2
477	1	5	15910	4	5	0	0	0	1

Tabla 4. 4 Grupo Familiar del Centroide del Cluster 1

En la tabla 4.4 se visualiza un grupo familiar, la fila resaltada representa al Centroide del Cluster 1. Se puede observar que este grupo familiar consta de 5 individuos, el tipo de Vivienda es Casa y el Ingreso Total Familiar es de $15.900.

En el caso del centroide (fila 474), el tipo de Cobertura Médico-Social es de planes y Seguros Sociales; su Nivel de Estudio es de Secundaria Incompleta, y no presenta dependencia Laboral ni tipo de Ingresos algunos. Además analizando otros datos de la Tabla, se observa que el Individuo es un Hombre de 47 años de edad al momento de la Encuesta. Este centroide agrupa 1127 individuos con condiciones Socio-Económicas similares. Teniendo en cuenta de los menores de edad son alrededor de 600 de los casi 2.000 encuestados.

Si continuamos con el Análisis de grupo Familiar, el ingreso Principal lo provee el Individuo de la fila 475; el cual tiene Cobertura Social, Nivel de Instrucción de Secundario Completo, trabaja en sector privado, y el Ingreso de su Ocupación Principal es de $10.000, y no posee ingresos de otras actividades. Este individuo también es un Hombre de 47 años al momento de ser encuestado, y pertenece al Cluster 2.

El individuo de la fila 473 es una Mujer de 53 años, no paga ni le descuentan por cobertura Médico Social, su Nivel de Instrucción es de Secundario Incompleto, trabaja en el Sector Privado u otros, y el Ingreso de su Actividad Principal es de $800. Pertenece al Cluster 2.

Continuando con la fila 476, es un Hombre de 31 años con Nivel de Educación Universitario Superior Completo, trabaja en el Sector Privado, pero no paga ni le descuentan por cobertura Médico Social. Los ingresos de su Actividad Principal es de $800, y pertenece al Cluster 2.

El individuo restante (Fila 477), es un joven de 18 años de sexo masculino. No tiene dependencia Laboral ni Ingresos de tipo alguno, y no paga por Cobertura Médico-Social. Su Nivel de Estudio es Superior Universitario Incompleto, probablemente cursando, y pertenece al Cluster 1.

Grupo Familiar del Centroide del Cluster 2:

	IV1	IX_TOT	ITF	CH08	NIVEL_ED	PP04A	P21	TOT_P12	pamx$clustering
1502	**1**	**4**	**21000**	**4**	**4**	**2**	**6000**	**0**	**2**
1503	1	4	21000	1	6	1	10000	0	2
1504	1	4	21000	1	1	0	0	0	1
1505	1	4	21000	1	7	0	0	0	1

Tabla 4. 5 Grupo Familiar del Centroide del Cluster 2

En el grupo Familiar de la Tabla 4,5, se visualiza en la fila resaltada al Centroide del Cluster 2. El grupo Familiar del Centroide del Cluster 2, tiene Casa como tipo de Vivienda y consta de 4 integrantes, con un Ingreso Total Familiar de $21.000.

En el caso particular del centroide (fila 1.502), es un Varón de 30 años, su Nivel de estudio es Secundario Completo, trabaja en el Sector Privado, con Ingresos de su Actividad Principal de $6.000, y sin Ingresos de otras actividades; no paga ni le descuentan por Cobertura Médico Social. El Cluster 2, agrupará 703 individuos con condiciones Socio-Económicas similares.

El otro individuo que aporta con sus ingresos al Hogar, es una Mujer de 28 años (Fila 1.503), que trabaja en el Estado, con Ingresos de su Actividad Principal de $10.000 y posee Obra Social. Su nivel de estudio es Superior Universitario Completo, no posee Ingresos de Otras Actividades, y pertenece al Cluster 2.

En la fila 1.504 se puede observar a un niño de 5 años que está cursando la Primaria (Primaria Incompleta); tiene Obra Social por extensión de su Madre (Fila 1.503), y

por obvias razones no tiene Dependencia Laboral ni ingresos algunos. Pertenece al Cluster 1. En la Fila 1.505 tenemos otro niño, el cual tiene 3 años, y no aún no cursa ningún nivel Educativo. También tiene Obra Social a cargo de su Madre y pertenece al Cluster 1.

Grupo Familiar del Centroide del Cluster 3:

	IV1	IX_TOT	ITF	CH08	NIVEL_ED	PP04A	P21	TOT_P12	pamx$clustering
1914	2	5	15720	1	4	1	15000	0	3
1915	**2**	**5**	**15720**	**1**	**4**	**0**	**0**	**0**	**3**
1916	2	5	15720	1	5	0	0	0	3
1917	2	5	15720	1	3	0	0	0	3
1918	2	5	15720	1	1	0	0	0	3

Tabla 4. 6 Grupo Familiar del Centroide del Cluster 3

En la tabla 4.6 se observa el grupo Familiar del Centroide del Cluster 3 (fila resaltada). El grupo Familiar del Centroide del Cluster 3 consta de 5 integrantes, tiene como Tipo de Hogar “Departamento” y un Ingreso Total Familiar de $15.720.

El caso del Centroide del Cluster 3 (Fila 1.915) corresponde a una Mujer de 42 años, la cual tiene Obra Social por extensión de su pareja, su Nivel de Estudios es Secundario Completo y no tiene dependencia Laboral ni tipo de Ingreso alguno. Estas son las características Socio-Económicas medias de 58 individuos del Cluster 3.

En la Fila 1.914 se observa un Hombre de 49 años, con Obra Social, que trabaja para el Estado, cuyo Ingreso de su Actividad Principal es de $15.000. Su nivel de Instrucción es Secundario Completo, y también pertenece al Cluster 3.

El individuo de la fila 1.916, es una Mujer de 18 años, cuyo Nivel Educativo es Superior Universitario Incompleto, probablemente cursando. Tiene Obra Social por extensión de su Padre; y se observa que no realiza ningún tipo de de trabajo por no presentar dependencia laboral ni Ingresos de tipo Alguno. Pertenece al Cluster 3.

La Fila 1.917 corresponde a un varón de 12 años, el cual tiene Obra Social por extensión de su Padre, está cursando el Secundario, y por obvias razones no presenta dependencia Laboral ni Ingresos. Pertenece al Cluster 3.

Por último, el individuo de la fila 1.918 es también un niño de 9 años al momento de la encuesta. Tiene Obra Social por extensión de su Padre, está cursando la Primaria y pertenece al Cluster 3.

Grupo Familiar del Centroide del Cluster 4:

	IV1	IX_TOT	ITF	CH08	NIVEL_ED	PP04A	P21	TOT_P12	pamx$clustering
687	**1**	**4**	**28000**	**1**	**6**	**1**	**8000**	**8000**	**4**
688	1	4	28000	1	4	2	12000	0	2
689	1	4	28000	1	3	0	0	0	1
690	1	4	28000	1	3	0	0	0	1

Tabla 4. 7 Grupo Familiar del Centroide del Cluster 4

En la tabla 4.7 se observa el grupo familiar del Centroide del Cluster 4, siendo el centroide la fila resaltada. El grupo Familiar del Centroide del Cluster 4 consta de 4 integrantes, el Tipo de Hogar es Casa, y el Ingreso Total Familiar es de $28.000.

El caso particular del Centroide (fila 687), se trata de una Mujer de 50 años al momento de realizar la Encuesta. Su Nivel de Instrucción es Superior Universitario Completo, trabaja en el Estado y posee Obra Social; el Ingreso de su Actividad Principal es de $8.000, y aparte posee Ingresos de otras actividades también de $8.000. Estas son las características Socio-Económicas medias de 52 individuos del Cluster 4.

En la fila 688 se presenta a un Hombre de 54 años. Su Nivel de Estudio es Secundario Completo, trabaja en el Sector Privado, y el Ingreso de su Actividad Principal es de $12.000. Posee Obra Social, y recibe Ingresos de otras Actividades. Pertenece al Cluster 2.

El individuo de la fila 689 es un varón de 19 años, el cual no tiene dependencia Laboral ni ingresos de tipo alguno. Su nivel de Estudios es Secundario Incompleto, y posee Obra social por extensión de alguno de sus Padres. Pertenece al Cluster 1.

Por último, el individuo de la fila 690, corresponde a un niño de 13 años, con Nivel de Estudios Secundario Incompleto, seguramente cursando. Por obvias razones no posee

dependencia Laboral ni ingresos de tipo alguno. Tiene Obra Social por extensión de alguno de sus Padres, y pertenece al Cluster 1 al igual que su hermano.

Análisis de Dos Variables

En la Figura 4.24, se visualizan todas las combinaciones posibles en dos variables. De todas ellas se ha seleccionado la gráfica correspondiente al Ingreso de la Actividad Principal y el Tipo de Hogar (Fig 4.25). Siendo de interés porque tratamos con una variable de la Tabla Individuos (Ingreso de la Actividad Principal), y con una variable de la Tabla Hogares (Tipo de Hogar). A continuación, la gráfica con su correspondiente Análisis:

Ingreso de Actividad Principal / Tipo de Hogar:

Con la siguiente sentencia, obtenemos la gráfica dela Figura 4.25:

```
plot(jitter(eph.kmedois[,7]), jitter(eph.kmedois[,1]), col=pamx$clustering)
```

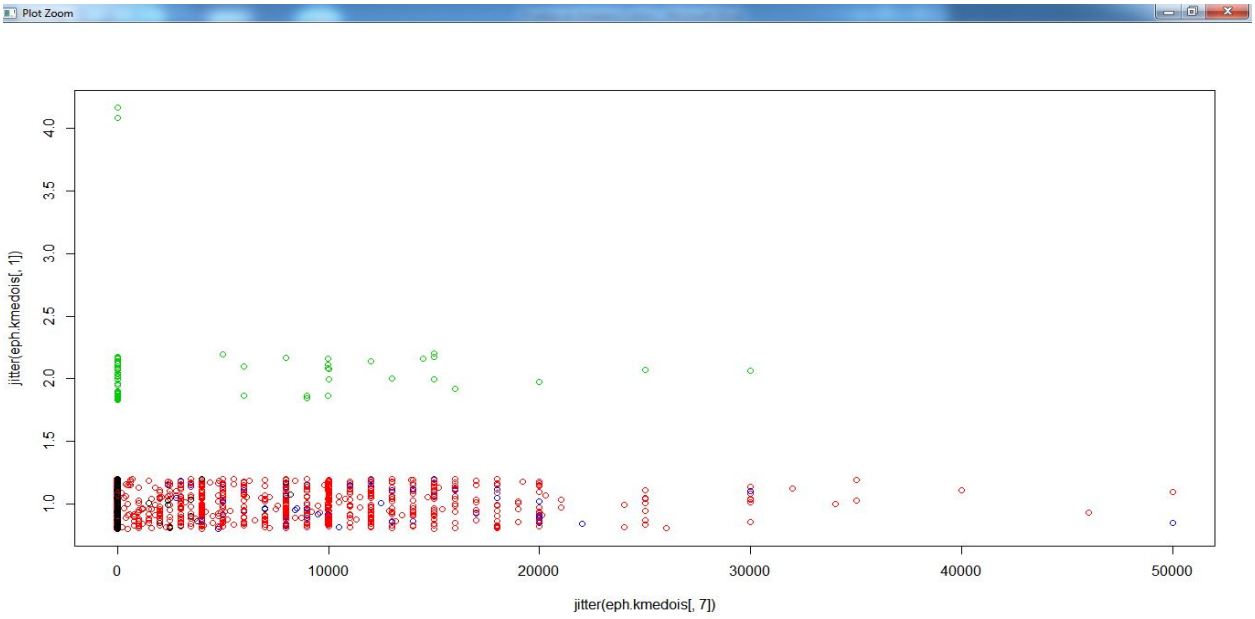


Figura 4. 25 Gráfico de dos variables de Ingreso de Actividad Principal (eje x) y el Tipo de Hogar (eje y).

En la Figura 4.25 se pueden observar en dos Variables los Clusters del Tipo de Hogar respecto a los Ingreso de la Actividad Principal. El eje “x” se refiere al Ingreso de la Actividad Principal, y en el eje de las “y” el Tipo de Hogar.

Los puntos que se ubican para el valor 1 del tipo de Hogar, esto es tipo de Hogar “Casa”, la nube de puntos es muy variada; hay puntos de color negro, rojo y azul. Estos

puntos corresponden a los Clusters 1, 2 y 4, que como hemos visto en la descripción de cada Cluster, los individuos que pertenecen a estos grupos viven en Casas. Los puntos de color negro, corresponderían al Cluster 1, dado que los ingresos de las Actividades principales de este grupo son nulos. Los puntos de color rojo corresponderían al Cluster 2, tanto por el nivel de ingresos como también por ser el grupo mayoritario. Y los puntos de color azul, pertenecerían al Cluster 4.

Mientras, los puntos de color verde, corresponden al Cluster 3, cuyo tipo de vivienda es “Departamento”; y cabe resaltar que quienes viven en Departamentos, en su mayoría tienen Ingresos menores a $20.000.

En cuanto a quienes tienes tienen Ingresos de actividad Principal de $0, se encuentran aproximadamente entre 500 y 600 menores de edad. El resto presumiblemente son personas desocupadas y también se presenta el caso de personas que están bajo dependencia Laboral pero por alguna razón no están recibiendo el cobro correspondiente.

Conclusión:

La Técnica de Clustering ha permitido agrupar la población de acuerdo a características socio-económicas similares. Es oportuno aclarar, que no todo el grupo Familiar necesariamente pertenece a un mismo Cluster, dado que el Nivel de Estudio, los ingresos y el tipo de Dependencia Laboral influyen. Siendo de esta forma que, por ejemplo, los menores de edad de un grupo familiar pertenezcan a un Cluster distinto al que los mayores de ese mismo grupo familiar.

En el Cluster 1 tenemos un grupo cuyo tipo de vivienda es “Casa”, con una cantidad promedio de 5 habitantes y con Ingresos Familiares que ronda en los $15.000, provenientes de Planes y Seguros Sociales. La desocupación es evidente al no existir Dependencia Laboral ni mucho menos Ingresos de una Actividad principal. También es de notar que el Nivel medio de Estudio es Secundario Incompleto. Si bien muchos de ellos son menores de edad, pero como se ha profundizado en el estudio del grupo Familiar este Cluster, el centroide de este grupo, es un Hombre de 47 años al momento de la Encuesta. Profundizar en estas características sería un trabajo a realizar a futuro.

El Cluster 3 tiene características similares al Cluster 1. Dado a que los Ingresos Familiares totales rondan también en los $15.000, no hay dependencia Laboral ni Ingresos de Actividades principales, lo cual indica desocupación. El nivel de estudios medio es similar al del Cluster 1, Secundario Completo, y si bien, por las características, algunos de ellos podrían ser menores de edad, como se ha analizado anteriormente el centroide de este Cluster es una Mujer de 42 años; de igual forma, sería una temática a profundizar en un trabajo futuro. Lo que diferencia a este Cluster del primero es el tipo de Hogar, que este caso es “Departamento”, y como se ha visto en el gráfico de dos variables de la figura 4.25, los ingresos medios concuerdan, siendo estos inferiores a los $20.000. Por último, lo que diferencia a este Cluster es que si tienen Cobertura Social, más específicamente Obra Social incluyendo PAMI. Por lo cual se puede concluir que este grupo de personas está compuesto por menores de edad, y entre los mayores de edad probablemente jubilados y

desocupados. Una característica especial del Grupo Familiar del Centroide del Cluster 3, es que todo este grupo Familiar pertenece al mismo Cluster.

El Cluster 2 tiene como tipo de vivienda “Casas”, con una cantidad promedio de 4 habitantes, y con un Ingreso Familiar de $21.000. Trabajan en el Sector Privado, pero no paga no le descuentan por cobertura Médico-Social. Los Ingresos de la Actividad principal rondan en los $6.000, y el nivel de estudios promedio Secundario Completo. Este Cluster tiene algunas similitudes con el Cluster 4, pero tienen diferencias muy marcadas. En el Cluster 4, también el tipo de hogar es “Casa” y la habitan un promedio de 4 personas. Pero los Ingresos Familiares son las elevados ($28.000), y esto se debe a que el Nivel de estudios es más alto (Superior Universitario Completo), y también quizás porque este grupo trabaja en el Estado. Además en Cluster 4, se observa que los individuos pueden acceder a otras actividades y tener ganancias casi iguales a las del ingreso principal. En este último Cluster se puede observar el grupo de individuos con mayor nivel de Estudio, mayores ingresos Familiares ($28.000), de la Actividad Principal ($8.000) y de otras actividades ($8.000). Poseen Cobertura Socia la cual incluye PAMI, pero sin duda alguna, y sería trabajo a profundizar a futuro, a este grupo pertenecen las personas que además de tener Obra Social, tienen acceso a Pre-pagas y Servicios de Emergencias.

Con este estudio se pueden ver claramente 4 grupos con características socio-económicas deferentes. Y aunque el Cluster 1 tenga características similares al Cluster 3, y lo mismo ocurra con los Cluster 2 y 4, existen marcadas diferencias que nos exponen la variedad que existe en el Entorno Social y Económico del Aglomerado “Gran Catamarca”, correspondiente al Primer Trimestre del año 2017.

4.2.3 Despliegue (Implementación).

En esta etapa corresponde hacer un informe final con el trabajo realizado y los resultados obtenidos. El presente trabajo forma parte de esta etapa, siendo este el Informe Final.

Capítulo 5: Conclusión

En el presente Trabajo Final se ha aplicado la Metodología CRISP-DM y se han utilizado algoritmos del Aprendizaje Automatizado. Más precisamente la Técnica de Árboles de Decisión para el Aprendizaje Automatizado Supervisado, y Clustering para el Aprendizaje Automatizado No Supervisado. Estas herramientas permitieron alcanzar el Objetivo General y los Objetivos Específicos como se detalla a continuación:

Objetivo General

- Aportar y complementar los Análisis Clásicos de Estadística, con técnicas novedosas de Minería de Datos que permitan detectar características comunes entre hogares e individuos del Gran Catamarca en base a Datos de la EPH.

Mediante las Técnicas de Árboles de Decisión y Clustering se pudo detectar características Socio-económicas entre hogares e Individuos del Aglomerado “Gran Catamarca”, correspondientes a la EPH del primer trimestre del año 2017.

Con Árboles de Decisión se obtuvo un dato interesante e innovador en cuanto nivel de Estudio entre Hombres y Mujeres con sueldos promedios de $10.000, independientemente de si trabajan en el Estado o en el Sector Privado. Analizando los datos de ese sector poblacional se encontró que es mayor el número de Mujeres con Nivel Superior Universitario Completo que de Hombres; siendo que aproximadamente 2 de cada 10 hombres tienen ese Nivel de Estudios, mientras que 4 de cada 10 mujeres llegan a este Nivel de Instrucción. Esto siempre y cuando cumpliendo con las características Socio-Económicas que se estudiaron en el Capítulo 4. También se corroboraron Resultados ya conocidos; esto es que a Mayor Nivel de Instrucción, mayores son las posibilidades de acceder a un trabajo y que este sea bien remunerado, esto a su vez permitirá que el individuo pueda tener acceso a una mejor Cobertura Medico-Social. Este resultado nos permite Complementar el Análisis Estadístico.

En cuanto a Clustering se agrupó la Población en 4 Clusters, cada uno con sus características particulares y marcadas. También, en base al mismo estudio de Clustering se analizó uno de los tantos gráficos de dos variables generados; el caso que se estudió involucró a una variable de la Tabla Individuos (Ingreso de la Actividad Principal) y una variable de la Tabla Hogares (Tipo de Hogar). Las Conclusiones particulares de caso de Estudio se han expuesto en el Capítulo 4.

Objetivos Específicos

- Obtener descripciones gráficas de diferentes tipos de hogares e individuos en base a su situación socio-económica.

Se han obtenido las gráficas necesarias para cada estudio de acuerdo a la Técnica implementada. En Árboles de Decisión, además de los Gráficos de Árboles, y los gráficos estadísticos presentes en los Nodos Terminales, mediante los porcentajes obtenidos se logró realizar gráficos Circulares que ayudan a visualizar e interpretar los resultados Obtenidos. Mientras que en Clustering se obtuvo una gráfica con todas las combinaciones posibles de las variables estudiadas, con gráficos de dos variables. De todos estos gráficos se eligió uno, el cual relaciona una variable de la Tabla Individuos (Ingreso de Actividad principal) y una variable de la tabla Hogares (Tipo de Hogar), permitiendo distinguir los distintos Clusters de acuerdo al Ingreso de la Actividad Principal y el Tipo de Hogar en que viven.

- Determinar agrupamientos de hogares e individuos que reúnan características similares y que a su vez difieren con respecto a elementos de otros grupos

Mediante la Técnica de Clustering se logró determinar agrupamientos de Hogares e Individuos que reúnen características similares; este análisis se hizo sobre la tabla unificada (EPH). De esta forma se logró obtener 4 grupos con características bien marcadas, diferenciándose cada uno del resto, pero además se observó similitudes, siendo que el Cluster 1 tiene características muy parecidas al Cluster 3, y lo mismo ocurre con los Cluster 2 y 4.

El Cluster 1 se relaciona con el Cluster 3, porque en ambos se observa un bajo Nivel de Estudios, como así también la carencia de empleo e Ingresos de una Actividad Principal, sin embargo ambos grupos tienen ingresos Familiares similares. La diferencia entre estos grupos está en que el primero tiene como cobertura, Planes y Seguros Públicos, mientras que el otro grupo tiene Obra Social, además en el Cluster 1 el tipo de vivienda es “Casa”, mientras que en el Cluster 3, es Departamento.

En los Cluster 2 y 4, las similitudes vienen dadas por tener el mismo tipo de vivienda (Casa), Ingresos Familiares similares (entre los $21.000 y los $28.000), y que en ambos grupos tienen Cobertura Medico-Social. Si bien los ingresos Familiares son similares, al profundizar, se observa que la diferencia está en que en el Cluster 4 los ingresos Familiares son mayores, porque los Ingresos de las actividades principales de cada individuos son mayores, pero además tienen ingresos de otras actividades; esto está marcado seguramente por el Nivel de Estudio, dado que en el Cluster 4, el nivel es Superior Universitario Completo, mientras que en el Cluster 1, es Secundario completo.

De modo que este objetivo se cumplió en desarrollo del presente trabajo, como se puede ver plasmado en el Capítulo 4.

- Determinar jerárquicamente la influencia de las diferentes variables demográficas con relación a distintas situaciones socioeconómicas.

Se determinó Jerárquicamente la influencia de las variables demográficas “Nivel de Estudio” y “Tipo de Cobertura Social” con relación a una selección de variables predictoras mediante la Técnica de “Árboles de Decisión”. El “Nivel de Estudio” se analizó en función del “sexo”, “edad”, “Tipo de Cobertura Social”, “Dependencia Laboral”, “Ingreso de Ocupación Principal” e “Ingreso de otras ocupaciones”; lo propio se hizo con el “Tipo de Cobertura Social” en función del “sexo”, “edad”, “Nivel de Estudio”, “Dependencia Laboral”, “Ingreso de Ocupación Principal” e “Ingreso de otras ocupaciones”.

De esta forma se encontraron resultado innovadores y también se corroboraron otros con los cuales se han complementando los resultados obtenidos mediante Técnicas Estadísticas. Como se enunció anteriormente, como resultado innovador, se encontró que en un cierto grupo de la muestra, la cantidad de mujeres con Nivel de Estudio Superior Universitario Completo es mayor al de Hombres. Y se complementa con Técnicas de Minería de Datos los resultados obtenidos con Técnicas Estadísticas, corroborando de esta forma que a Mayor Nivel de Instrucción, mayores son las posibilidades de acceder a un trabajo y que este sea bien remunerado, esto a su vez permitirá que el individuo pueda tener acceso a una mejor Cobertura Medico-Social.

Por lo tanto, en base a los resultados expuestos en este capítulo, se concluye que los Objetivos planteados en el presente trabajo se han cumplido.

Bibliografía

- Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.
- Ahumada, H. C., Dip, H., Herrera, C. G., & Leguizamón Almendra, J. C. (2016). Utilización de Minería de datos en el Análisis de Rendimiento Académico de Alumnos de Primer Año de Ingeniería. In *III Congreso Argentino de Ingeniería – IX Congreso de Enseñanza de la Ingeniería (Resistencia, 2016)*.
- Arancibia, J. A. G. (2010). Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM. *Recuperado de http://oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM*, *2385037*.
- Argentina, I. N. D. E. C. La nueva encuesta permanente de hogares de Argentina 2003.
- Garcıa, J. M. C., Portillo, E. M., & Cezón, P. A. (2010). Introducción a la programación estadística con R para Profesores.
- Luis Alfonso Cutro (2008). Minería de Datos Aplicada a la Encuesta Permanente de Hogares (**Tesis** de grado). Universidad Nacional del Nordeste, Corrientes, Argentina.
- Daniel Germán Martinez (2016). Módulo para pronóstico de Consumo Eléctrico (Tesis de Grado). Universidad Nacional del Nordeste, Corrientes, Argentina.
- Torres, D. L., Meyer, R. D., & Cárdenas, V. T. (2011). Minería de datos en la encuesta permanente de hogares 2009, universidad nacional del litoral, Santa Fe, Argentina. *Revista Ingeniería Industrial*, *10*(1), 19-28.
- Elizalde, M., Pok, C., Botta, A., & Villareal, J. (1974). Encuesta Permanente de Hogares: marco teóricometodológico de la investigación temática.
- Ahumada, J. A. (2003). R para Principiantes. *University of Hawaii*.
- Tusell, F. (2008). Lectura, manipulación y análisis de datos en R.
- Vásquez, E. (2013). Inversión social: indicadores, bases de datos e iniciativas.
- Lantz, B. (2013). *Machine learning with R*. Packt Publishing Ltd.
- Yu-Wei, C. D. C. (2015). *Machine learning with R cookbook*. Packt Publishing Ltd.
- Witten, I.H. & Frank, E.(2000). Data Mining. Practical Machine Learning Tools and Techniques with Java Implementation. Ed. Morgan Kaufmann Publisher. San Francisco CA, USA p. 7.
- Hernández, J., Ramírez, Ma.J. & Ferri, C. (2004)**.** ¿Qué es la minería de datos? In: Introducción a la Minería de Datos. (1ra. Ed. pp. 25-283). Madrid, España.
- G.; Smith P.; Ramasasmy U. Fayyad, U.M.; Piatestskiy-Shapiro. Advances in Knowledge Discovery and Data Mining. AAAI Press / MIT Press, 2006.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). New York: springer.

- Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, *31*(3), 264-323.
- JAIN, A. K. AND DUBES, R. C. 1988. *Algorithms for Clustering Data*. Prentice-Hall advanced reference series. Prentice-Hall, Inc., Upper Saddle River, NJ.
- ANDERBERG, M. R. 1973. *Cluster Analysis for Applications*. Academic Press, Inc., New York, NY.
- DIDAY, E. AND SIMON, J. C. 1976. Clustering analysis. In *Digital Pattern Recognition*, K. S. Fu, Ed. Springer-Verlag, Secaucus, NJ, 47–94.
- MICHALSKI, R., STEPP, R. E., AND DIDAY, E. 1983. Automated construction of classifications: conceptual clustering versus numerical taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell. PAMI-5*, 5 (Sept.), 396–409.
- BRAILOVSKY, V. L. 1991. A probabilistic approach to clustering. *Pattern Recogn. Lett. 12*, 4 (Apr. 1991), 193–198.
- ZAHN, C. T. 1971. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Trans. Comput. C-20* (Apr.), 68–86.
- MCQUEEN, J. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 281–297.
- BALL, G. H. AND HALL, D. J. 1965. ISODATA, a novel method of data analysis and classification. Tech. Rep.. Stanford University, Stanford, CA.
- DIDAY, E. 1973. The dynamic cluster method in non-hierarchical clustering. *J. Comput. Inf. Sci. 2*, 61–88.
- SYMON, M. J. 1977. Clustering criterion and multi-variate normal mixture. *Biometrics 77*, 35–43.

Sitios Web:

- INDEC. (2003). La nueva Encuesta Permanente de Hogares de Argentina. 2003. 01/11/2017, de INDEC Sitio web:

  https://www.indec.gob.ar/ftp/cuadros/sociedad/Metodologia_EPHContinua.pdf

- INDEC. (2016). Encuesta Permanente de Hogares Diseño de Registro y Estructura para las bases preliminares Hogar y Personas. Tercer trimestre de 2016. 22/02/2018, de INDEC Sitio web:

  https://www.indec.gob.ar/ftp/cuadros/menusuperior/eph/EPH_registro_3_trim_2016.pdf

- The Modeling Agency. (1999-2000). “CRISP-DM 1.0 Step-by-step data mining guide”. 20/02/2018, de The Modeling Agency Sitio web: https://www.the-modeling-agency.com/crisp-dm.pdf

- Oracle. (2018). “Data Mining Process”. 20/02/20218, de Oracle Sitio web: https://docs.oracle.com/cd/B19306_01/datamine.102/b14339/5dmtasks.htm

- IBM. (1994-2011). IBM SPSS Modeler CRISP-DM Guide. 21/02/2018, de IBM Sitio web: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf
- INDEC 2. (2003). ENCUESTA PERMANENTE DE HOGARES (EPH). CAMBIOS METODOLÓGICOS. 23/02/2018, de INDEC Sitio web: https://www.indec.gob.ar/ftp/cuadros/sociedad/Gacetilla_EPHContinua.pdf
- RStudio. (2018). RStudio IDE features. 11/05/2018, de RStudio Sitio web: https://www.rstudio.com/products/rstudio/features/
- Carlos J. Gil Bellosta. (2018). R para profesionales de los datos: una introducción.. 28/05/2018, de Data Analytics Sitio web: https://www.datanalytics.com/libro_r/arboles-de-decision.html
- Juan Bosco Mendoza Vega. (2018). Arboles de decisión con R - Clasificación. 29/05/2018, de RPubs Sitio web: https://rpubs.com/jboscomendoza/arboles_decision_clasificacion
- Patricio Almeida Gentile. (2015). Encuesta Permanente de Hogares (EPH). 05/06/2018, de Observatorio Económico Social UNR Sitio web: http://www.observatorio.unr.edu.ar/encuesta-permanente-de-hogares-eph/
- HL7. (2007). About HL7. 11/06/2018, de HL7 Sitio web: http://www.hl7.org/about/index.cfm?ref=nav
- Angelo Santana, Carmen N. Hernández. (2016). Objetos en R: Factores. 16/08/2018, de Departamento de Matemáticas, ULPGC Sitio web: http://www.dma.ulpgc.es/profesores/personal/stat/cursoR4ULPGC/6d-Factores.html
- R Development Core Team. (1999). Introducción a R. 16/08/2018, de R and CRAN Sitio web: https://cran.r-project.org/doc/contrib/R-intro-1.1.0-espanol.1.pdf
- Torsten Hothorn, Kurt Hornik, Achim Zeileis. (2006). ctree: Conditional Inference Trees. 16/08/2018, de R and CRAN Sitio web: https://cran.r-project.org/web/packages/partykit/vignettes/ctree.pdf
- Alejandro Cassis. (2015). Aprendizaje Supervisado. 19/09/2018, de Wordpress Sitio web: https://inteligenciaartificial101.wordpress.com/2015/10/20/aprendizaje-supervisado/
- Logicalis. (2017). Learning machine, los usos del aprendizaje supervisado. 19/09/2018, de Logicalis.com Sitio web: https://blog.es.logicalis.com/analytics/learning-machine-los-usos-del-aprendizaje-supervisado
- Fernando Sancho Caparrini. (2017). Clasificación Supervisada y No Supervisada. 19/09/2018, de http://www.cs.us.es Sitio web: http://www.cs.us.es/~fsancho/?e=77
- Laura Aspirot, Sebastián Castro. (2013). Introducción a las técnicas estadísticas de clasificación y regresión. Aprendizaje no supervisado - Clustering. 19/09/2018, de Universidad de la Republica (UdelaR) Sitio web: http://www.iesta.edu.uy/wp-

content/uploads/2014/05/Escueladeverano_RegionalNorteSalto_2013_PresentacionNoSupervisado_Aspirot_Castro1.pdf