ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

Minería de datos aplicada al manejo de la relación del Cliente en una Entidad Bancaria (CRM)

PROYECTO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO MATEMÁTICO

Raúl Andrés Muñoz Álvarez

raulandreuss@yahoo.com

DIRECTOR: MSC. Diego Maldonado

diego.maldonado6@gmail.com

CODIRECTOR: Dr. Diego Recalde

diego.recalde@epn.edu.ec

Quito, 2013

DECLARACIÓN

Yo, Raúl Andrés Muñoz Álvarez, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración cedo mis derechos de propiedad intelectual correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

Raúl Muñoz Álvarez

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por Raúl Andrés Muñoz Álvarez bajo mi supervisión.

_____

MSC. Diego Maldonado
DIRECTOR DE PROYECTO

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por Raúl Andrés Muñoz Álvarez bajo mi supervisión.

_____

Dr. Diego Recalde
CODIRECTOR DE PROYECTO

AGRADECIMIENTOS

Mi sincero agradecimiento:

Al MSC.Diego Maldonado por el apoyo brindado en la realización de este proyecto de titulación

A Paulina, Diana y Edgar por su sincera amistad y constante apoyo para cumplir con ésta meta.

DEDICATORIA

A mis padres,

por el apoyo incondicional brindado en el transcurso de la carrera

RESUMEN

Este documento analiza los patrones de comportamiento de un cliente inmersos en los datos recolectados en los canales transaccionales de una Entidad Bancaria, mediante la utilización de técnicas de Minería de Datos con la finalidad de determinar perfiles de clientes que permitan un mejor conocimiento de la relación existente entre cliente y su Entidad Bancaria.

En el capítulo 1 se estudia la relación entre el cliente y la Entidad Bancaria, desde el punto de vista del Marketing y CRM [1], además se describe el vínculo existente con la Minería de datos. Finalmente se detalla las investigaciones previas y trabajos relevantes.

En el capítulo 2 se realiza una introducción de la Minería de Datos, se analiza su relación con el Data warehousing y se describe la metodología a seguir para realizar un proceso de Minería de datos.

En el capítulo 3 se empieza con una introducción de lo que es un modelo de conglomerados y su aplicación en el Marketing, se hace hincapié en los diferentes métodos de segmentación y su utilidad para una Entidad Bancaria. Luego, se estudia un conjunto de métodos que permiten la detección de valores atípicos, finalmente se realiza una revisión de las diferentes técnicas de análisis de conglomerados.

En el capítulo 4 se analiza el vínculo que tiene los árboles de clasificación y la minería de datos, se describe las diferentes fases que realizan los árboles de clasificación y se examina el método CHAID.

En el capítulo 5 se desarrolla el modelo para encontrar agrupaciones de clientes con comportamiento transaccional similar. Para cada grupo, mediante árboles de decisión, se evalúa si éste tiene relación con sus características demográficas. Finalmente se detallan las conclusiones de los resultados obtenidos.

Palabras claves: Perfiles, Minería de datos, SEMMA, Conglomerados, Patrones, Árboles de clasificación.

[1] Customer Relationship Management

TABLA DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Capítulo 1

Relación Cliente - Entidad Bancaria

En la actualidad, el entorno financiero se ve marcado por el avance de la tecnología y los efectos que ésta proporciona en la relación entre una Entidad Bancaria y sus clientes. Negocios tales como la entrega de servicios y colocación de productos financieros, que tradicionalmente eran realizados en las oficinas o agencias de servicios, hoy en día son ofertados mediante los nuevos canales de contacto tecnológicos, llegando a convertirse en el principal vínculo entre el cliente y la entidad Bancaria. Esta nueva ola de cambios hace imperioso la necesidad de buscar una estrategia que permita a la Entidad Bancaria manejar adecuadamente el vínculo entre ésta y sus clientes.

La relación o vínculo Cliente-Entidad, ha pasado de ser presencial a semipresencial o virtual, esto se debe a que los canales transaccionales y de servicio se han transformado en los nuevos agentes o representantes que posee una Entidad Bancaria, permitiéndole tener una mayor cobertura de mercado, alcanzar nuevos segmentos de clientes, los cuales, en general no tiene la necesidad de asistir a las agencias para realizar una transacción, y la gestión comercial puede ser más rápida y eficaz.

Las Entidades Bancarias miran a los canales transaccionales y de servicios como oportunidades para captar, satisfacer y conocer a sus clientes, muchas de estas entidades envían comunicaciones referente a promociones y beneficios para conocimiento de sus cliente, sin embargo mucha de esta información es ignorada o no observada, y en algunos casos provoca malestar siendo esto perjudicial para el propósito de la entidad. Esta generalización de la comunicación puede venir acompañada por un desconocimiento del perfil del cliente al cuál va dirigida la promoción u oferta de valor, por ejemplo, el envío de prestamos por el canal internet pueden producir incomodidad a clientes que en general no están familiarizados con el uso de éste, pudiendo provocar desconfianza e incluso adversidad a utilizar este canal nuevamente.

Dentro de este aspecto el Marketing, tiene un papel importante tanto en el manejo actual de la entidad Bancaria, así como también en el solventar los nuevos desafíos que enfrenta la misma. De acuerdo a Roberto Dvoskin [4], el Marketing ”es una disciplina de la Ciencia Económica cuyo objetivo es potenciar las capacidades de las organizaciones y/o individuos oferentes de bienes y servicios que, insatisfechos con una situación competitiva dada, aspiran a pasar a otra más ventajosas”. Philip Koptler [20], lo define ”como un proceso social y administrativo por

el que individuos y los grupos, obtienen lo que necesitan y desean a través de la creación e intercambio de productos y su valoración con otros". la Asociación Americana de Marketing define al Marketing "como la actividad, conjunto de instituciones y proceso para crear, comunicar, entregar e intercambiar ofertas que satisfagan las necesidades de los clientes, socios y la sociedad en general". Ana Casado señala que "los responsables de marketing no solo deben de considerar las necesidades del cliente sino también todos los grupos claves que posibilitan su existencia, con la finalidad de desarrollar un intercambio que realice el bienestar a largo plazo de todo este grupo clave y de la sociedad en conjunto".

Tomando en cuenta las definiciones anteriores, el Marketing es el encargado de generar una estrategia que permita llevar de manera adecuada y sostenible la relación que vincula a los clientes hacia los productos y servicios de la Entidad Bancaria. Por tales motivos, surge la necesidad de entender y profundizar en su teoría e identificar cuáles son los componentes que efectivamente abarcan el vínculo entre el cliente y su Entidad. A continuación se presenta su división o sistematizado [21]:

1. Escuela de management.- Intenta conectar los elementos teóricos con los problemas reales. Incorpora los conceptos de segmentación y marketing mix entre otros.

2. Escuela del comportamiento.- Centra la atención en determinar las causas que inducen a la compra por parte del consumidor.

3. Escuela sistémica.- Considera al Marketing como un sistema que asocia diversas funciones.

4. Escuela del intercambio social.- El eje central lo asume el mercado donde se produce un intercambio entre compradores y vendedores.

5. Escuela funcional.- Destaca la importancia de las funciones que realiza el intermediario en el proceso de distribución.

6. Escuela geográfica.- Centra la atención en el concepto de distancia, que motiva un trasvase de bienes y servicios y resulta determinante en la decisión de compra.

7. Escuela institucional.- Nace con el propio Marketing como disciplina individualizada, estudia las organizaciones que operan en el Mercado.

8. Escuela de dinámica organizativa.- Es derivada de la anterior, pero da un mayor énfasis a los objetivos y necesidades del consumidor.

9. Escuela activista.- Pone su acento en los desequilibrios de poder que aparecen en el mercado entre compradores y vendedores.

10. Escuela del producto.- Éste considera al producto como el elemento fundamental cuyo estudio permitirá elegir las políticas y estrategias adecuadas.

11. Escuela macromarketing.- Estudia la interrelación entre las actividades comerciales y la sociedad.

1.1 Customer Relationship Management (CRM)

Dentro del área del Marketing, en la actualidad se ha ido direccionando a la aplicación del CRM (Customer Relationship Management), como la mejor manera de conseguir ganar, retener y aumentar la base de clientes, es decir, realizar un giro de análisis hacia el enfoque del cliente con la finalidad de diseñar estrategias de Marketing de acuerdo a sus necesidades.

Para Arjan Sundardas [23], el CRM es " el proceso de adquisición, retención y crecimiento de la rentabilidad de clientes para una empresa". Adryan Payne [22] lo define como " el acercamiento estratégico, interesado en crear un mejor valor para el accionista a través del desarrollo de relaciones apropiadas con los clientes importantes y segmentos de clientes". Para Tsiptsis y Chorianopoulos [24], "es la estrategia para la construcción, la gestión, el fortalecimiento de la lealtad y el relacionamiento a largo plazo de los clientes".

1.1.1 Tipo de CRM

El objetivo ahora es el identificar que tipos de CRM existen y hacia donde van aplicados [22].

- **CRM Operacional**.- Es el área que se encarga de la automatización de los procesos del negocio envueltos en los puntos de contacto del cliente. Estas areas incluyen las ventas automatizadas, marketing automatizado y servicios automáticos para el cliente.

- **CRM Analítico**.- Se encarga de la captura, almacenamiento, análisis, interpretación y uso de la data creada a partir de las operaciones del negocio, con la finalidad de crear información estratégica para la organización.

- **CRM colaborativo**.- Envuelve el uso de servicios colaborativos e infraestructura necesaria para realizar la iteración entre la compañía y sus multiples

canales . Esto permite el relacionamiento entre el cliente, la empresa y sus empleados.

Entonces para que un cliente tenga una experiencia adecuada que beneficie a la empresa, se requiere la integración de las tres componentes del CRM.

1.1.2 Objetivos estratégicos y tácticos del CRM

Una organización cuya finalidad es la de satisfacer a sus cliente y extender sus unidades de negocio, debe tener claridad en como su información es trasmitida y estructurada, pues una comunicación correcta le permitirá cumplir con sus metas y objetivos. Por lo que la meta del CRM es aumentar la oportunidad de mejorar los procesos de comunicación hacia el cliente correcto, proveyéndolo del producto o servicio correcto, a través del canal correcto y en el tiempo correcto [2].

El cliente correcto

- Manejar la relación del cliente a través de su ciclo de vida.
- Analizar el cliente potencial por medio del aumento de cartera.

La oferta correcta

- Introducir eficientemente al cliente y los nuevos prospectos a la empresa, a los productos y servicios.
- Adecuar las ofertas para cada uno de los clientes.

Canal correcto

- Coordinar la comunicación a través de cada punto de contacto que se tiene con el cliente.
- Habilidad para comunicar por el canal preferido del cliente.
- Capturar y analizar la información proveniente de los canales, para obtener un conocimiento continuo.

En el tiempo correcta

- Eficiente comunicación al cliente en un tiempo relevante.
- Habilidad para comunicarse en tiempo real.

1.1.3 Implementación del CRM

De cara a implementar un programa de CRM de forma eficiente se requiere de cinco elementos:

- Estrategia.
- Segmentación.
- Tecnología.
- Procesos.
- Organización.

Estrategia.- Hay seis tipo de estrategias que afectan a un programa de CRM: canal, segmentación, precio, marketing, marca y publicidad, de las cuales las tres primeras tienen un gran impacto. La segmentación identifica la estructura de los clientes en la interna de la empresa, la estrategia de precios es el principal diferenciador en un mercado comoditizado y determinará más de la mitad del valor de la oferta. La estrategia de canal determinará como debe llevar la oferta al cliente.

Segmentación.- Tradicionalmente, la segmentación se enfoca en un producto o mercado particular, pero la nueva tendencia es el evaluar el valor de cada cliente hacia la empresa. Esto significa que en la actualidad el objetivo es determinar las necesidades que poseen los clientes mediante la utilización de adecuados algoritmos y métodos matemáticos.

Tecnología.- El proceso del CRM depende de los datos. La consideración técnica más importante es concentrarse en crear una única, orientada a las operaciones, e integrada base de datos. Otros elementos esenciales a considerar son el software para la base de datos, la data mínima y las herramientas de gestión de campañas y de soporte a la decisión. El hardware, las bases de datos y el software de soporte a la toma de decisiones deben ser usados de forma coherente e integrada.

Proceso.- No es difícil identificar los proceso que se deben rehacer para implementar el CRM. Las dificultades aparecen en la venta del proyecto a la organización, así como los indicadores de medidas para analizar el progreso y poder así tomar decisiones de gestión. El proceso del CRM es la metodología mediante la cual las actividades de marketing directo son ejecutadas.

Organización.- La estructura organizacional es con frecuencia el componente más analizado de la implementación del CRM. La mayoria de empresas utilizan el marketing de medios básicamente.

1.1.4 Base de Datos de marketing

En la actualidad las Entidades Bancarias han comenzado a recolectar bases de datos relevantes de sus clientes (Database Marketing), con la finalidad de obtener información más real, completa, en línea y actualizada [10]. De manera global, permite la creación e implementación de modelos matemáticos descriptivos y predictivos con la finalidad de enviar mensajes deseados, en el momento oportuno, en la forma correcta y a las personas adecuadas. Todo esto con el objetivo de complacer a nuestros clientes, aumentar nuestra tasa de respuesta por dólar de comercialización, reduciendo el costo por el orden, fomentar nuestro negocio, e incrementando nuestros beneficios.

El Centro Nacional de bases de datos de Marketing define al Database Marketing, como la gestión de una base de datos relacional informatizada, en tiempo real, completa y actualizada, con datos relevantes sobre los clientes, consultas, prospectos y sospechosos, para identificar a nuestros más sensibles clientes con el propósito de desarrollar una alta calidad, larga relación y la regeneración de negocios mediante el desarrollo de modelos predictivos .

Como se puede observar la definición de database de Marketing está íntimamente relacionada con el concepto de CRM, pues es el input que los agentes de Marketing necesitan para realizar un correcto manejo de sus clientes, es claro que una empresa que contiene información relevante del cliente, almacenada en una base de datos, le permitirá analizar su comportamiento de compra, su perfil demografico y psicológico, las tendencias del mercado y sus potenciales prospectos.

La base de datos de marketing utiliza la información propia de los clientes, con la finalidad de servir de una manera más adecuada, detectar mejores servicios para ofrecer, hacer recomendaciones de productos, implementar promociones más efectivas haciéndole a la compañía potencialmente sostenible y con ventaja competitiva.

1.1.5 Relación del CRM y la Minería de Datos

Dado que el CRM se encarga de analizar el manejo de la relación del cliente con la empresa, ésta necesita analizar de una manera adecuada los datos almacenado en el Database de marketing con la finalidad de extraer conocimiento de la misma, para esto una de las herramientas estrategias para transformar los datos en conocimiento es la minería de datos, cuyo objetivo principal es el minar la montaña de datos de la empresa y encontrar las pepitas de oro (información relevante) del cliente.

Para una Entidad Bancaria sus clientes actuales, son la parte principal de la organización. Una entidad no puede realizar un nuevo negocio sin satisfacerlos, pues su deber es mantener la lealtad y una buena relación con los mismos. tal como se mencionó antes El CRM es la estrategia a aplicar para cumplir con el objetivo del manejo de la relación Cliente-Entidad.

Dentro del CRM, una de sus componentes principales es el analítico, éste es el encargado de analizar la información recolectada por la Entidad, con la finalidad de direccionar el mensaje correcto al cliente correcto. Entonces analizándolo de una manera más especifica El CRM analítico tiene como finalidad el detectar los patrones presentados en la información para optimizar el relacionamiento del consumidor. Pues de aquí surge la necesidad de buscar un método o metodología que permita cumplir los objetivos del CRM analítico. Como se puede observar, la minería de datos abarca la problemática presentada por el CRM, pues de acuerdo a su definición es la herramienta que permite extraer el conocimiento de un conjunto grande de datos usando técnicas de modelamientos estadísticos.

Los modelos estadísticos inmersos en la minería de datos, permiten analizar las diferentes operaciones que realiza un cliente en la organización haciendo más efectivo el manejo del cliente. Estos análisis permiten detectar los comportamientos inmersos en los datos y proveer conocimiento al analista, y de acuerdo a esto, establecer las estrategias para los clientes actuales y futuros.

Sin lugar a dudas CRM puede aportar datos que manifiesten la evolución de la actividad del negocio y su relación con el cliente, mientras que la minería de datos convierte estos datos en conocimiento para el negocio. Por eso es fundamental entender el proceso del Minería de datos y sus implicaciones para la obtención del conocimiento. En este proyecto de titulación se hará hincapié en este tema.

1.2 Investigaciones previas y Trabajos relevantes

Dentro de las problemáticas o desafíos que enfrenta una Entidad Bancaria, unos de los más importantes se refleja en el análisis de la relación que mantiene con sus clientes, es claro que los puntos de contactos actuales que posee una entidad son los canales de transacción y servicio, los cuales vienen a representar la imagen de la entidad frente a su cliente. Es importante destacar que los datos recolectados en los diferentes canales son de suma importancia para la estrategia a implementar por parte de la entidad, tal como lo menciona el CRM estos datos son la huella que deja el cliente al contactarse con su entidad, reflejando los productos y servicios que él necesita. Por otro lado la entidad Bancaria debe

almacenar y explotar estos datos con la finalidad de transformarlo en información útil para la aplicación de un CRM a sus clientes.

En la figura 1.1 se puede observar los diferentes canales de distribución y servicios de una entidad Bancaria.[1]

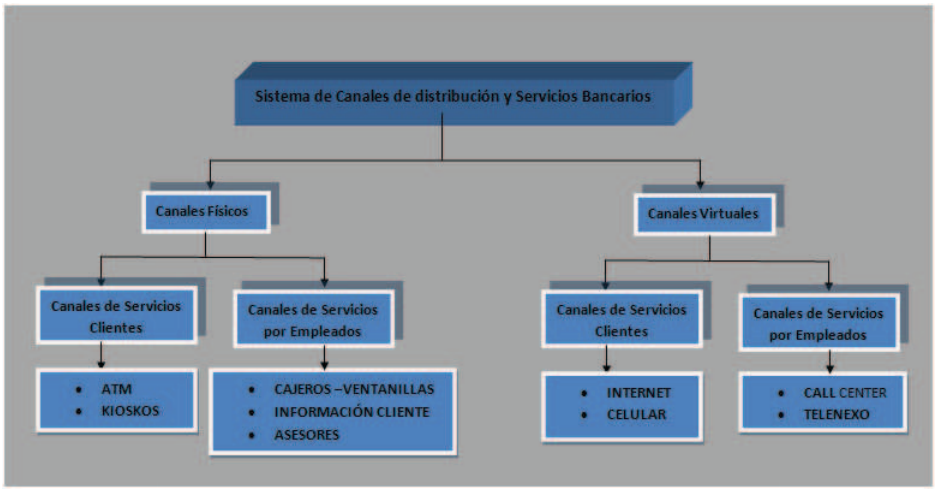


Figura 1.1: Canales de distribución y servicios

En la literatura de canales de servicio y transacción se ha encontrado que la interacción de un cliente y un servicio prestado por una entidad Bancaria puede ser clasificado como:

- Nivel de servicio de contacto alto.
- Nivel de servicio de contacto medio.
- Nivel de servicio de contacto bajo.

El nivel de servicio alto envuelve una significativa interacción con el cliente, éste visita las instalaciones de la entidad Bancaria con el objetivo de recibir o realizar un servicio de manera presencial. Ejemplo de ésto, es cuando un cliente realiza un depósito o retiro en el canal cajas o consultas en el balcón de servicios.

El nivel de servicio medio envuelve situaciones en las cuales el cliente visita servicios facilitados por la entidad Bancaria pero que no requieren de personal propio de la entidad para recibir o realizar el servicio. Ejemplo de estos son el canal ATM [2], kioskos y telenexos.

El nivel de servicio bajo envuelve situaciones en las cuales la participación del personal de la entidad Bancaria es mínima o casi nula, el cliente es el mayor participante, en esta categoría se encuentran los canales Internet y celular.

[1] Adaptado del documento Customer Efficiency, Channel Usage and Firm Performance in Retail Banking

[2] ATM: Automated Teller Machine

Este proyecto de titulación se enmarca en el análisis de los datos levantados en los canales transaccionales de una entidad Bancaria, con la finalidad de poder responder las siguientes preguntas:

1. ¿Cuál es la saturación de sus canales transaccionales.?
2. ¿Cuáles son los patrones de comportamiento de los clientes en el uso de canales?
3. ¿Existe algún patrón demográfico que determine el uso del canal?

Para esto, la documentación o literatura que abarque en conjunto todos estos objetivos es escaso, es destacable el documento ”La Estrategia Multicanal Desde el Punto De Vista del Cliente” realizado por Cortiñas, Chorraro y Villanueva, los cuales realizan un análisis empírico del comportamiento multicanal mediante la utilización del índice de entropía, el cual, les permite establecer una medida para identificar que clientes son multicanal. Luego observan la relación existente entre el índice obtenido y la rentabilidad Bancaria mediante la utilización de modelos de regresión de variables latente. Sus resultados determinan que un cliente que utiliza múltiples canales es más rentable, que aquel que utiliza solo un canal.

Otro documento interesante es ”Customer Efficiency, Channel Usage and Firm Performance in Retail Banking” realizado por Xue, Hitt, Harker, el cual se enfocan en un análisis empírico para determinar el uso de servicios automáticos, la asociación entre la elección de los canales por parte del cliente y sus afectación a los resultados de la entidad o empresa. Plantea la hipótesis de que los clientes eligen un canal basándose sobre el costo relativo y los beneficios que ellos reciben por cada elección. Utiliza un modelo que establece la utilidad obtenida por el uso del canal mediante el uso de regresión múltiple y transformaciones logarítmicas.

En este proyecto de titulación se realizará un análisis del comportamiento de los clientes en el uso de los diferentes canales transaccionales de una entidad Bancaria, uno de los objetivos a cumplir busca determinar grupos con patrones de comportamientos diferentes en los canales (Perfil transaccional), con la finalidad de permitir al área marketing aplicar una correcta estrategia de publicidad y oferta de valor a sus clientes. Con respecto a lo anterior se buscará también relacionar a estos grupos con sus datos demográficos, con el propósito de aportar de manera más concreta en la aplicación de la estrategia de marketing; dado que al identificar que un grupo de clientes con cierta características transaccionales en los canales, además tiene un patrón demográfico propio, permitirá a la entidad establecer campañas más acordes a estos perfiles identificados.

La parte teórica de este proyecto de titulación se enmarca en la aplicación de la minería de datos, la cual es un conjunto de técnicas que permiten el manejo de altos volúmenes de datos con mucha eficiencia. Estas técnicas en general permiten el descubrimiento de patrones, relaciones y tendencias intrínsecas en los datos, que mediante la aplicación de técnicas estadísticas, las transforman en información útil para la toma de decisiones. Dentro de las metodologías de minería de datos se analizará la presentada por el instituto SAS[3], la cual tiene las siglas SEMMA que significan Elección o muestreo, Exploración, Modificación, Modelización y Valoración, cada uno de estos ítems tiene inmersos un conjunto de técnicas de manipulación de datos y estadísticas que permiten obtener un modelo robusto y representativo del conjunto de datos.

Además se realizará un análisis teórico de los diferentes modelos de conglomerados existentes, con el objetivo de seleccionar uno de estos modelos para la detección de los grupos transaccionales. Dentro de este punto se realizará una clasificación de las técnicas de conglomerados y se mencionará en que software se encuentran implementados.

Se aplicará la técnica de conglomerados elegida a un conjunto de datos transaccionales en los canales ATM, Cajas, Internet, Telenexo, Kiosko y Celular de una entidad Bancaria, y se identificarán los diferentes perfiles transaccionales.

Se tomará como variable dependiente al identificador de grupo o conglomerado y mediante la aplicación de un árbol de clasificación, y utilizando a las variables demográficas como variables clasificadoras, se determinará si los grupos tienen algún patrón demográfico.

[3]SAS: Statistical Analysis Systems

Capítulo 2

Minería de Datos

2.1 Introducción

Hoy en día con el avance de la tecnología, las diferentes Entidades Bancarias han ido recolectando conjuntos de datos relevantes inherentes a la institución y sus clientes, estos datos han sido guardados en dispositivos con gran capacidad de almacenamiento (discos duros, servidores, etc.). Para una entidad Bancaria el transformar este conjunto de datos en información es de suma importancia, dado que al tener un conocimiento exacto de lo que sucede en su entorno, le permitirá tomar mejores decisiones permitiéndole ser eficiente y eficaz.

A finales de los 80, surge un nuevo campo de investigación que se denominó KDD (Knowledge Discovery in Databases) [25], y se lo identificó como ”el proceso no trivial de descubrimiento de patrones válidos, nuevos, potencialmente útiles y comprensibles en grandes volúmenes de datos”. Con el paso del tiempo el KDD ha ido tomando diferentes nombres para describir similares procesos. Algunos de ellos son: Knowledge Discovery, Data Discovery, Information Extraction, Data Extraction, Pattern Discovery, entre otros. En la actualidad el nombre que tiene o posee mayor aceptación es el Data Mining o Minería de Datos, este término proviene de la analogía entre enfrentarse a una gran cantidad de datos para descubrir patrones útiles y en la explotación de una montaña para encontrar un yacimiento de minerales preciosos.

Entonces a la minería de datos se la define como una herramienta tecnológica de manejo de información, cuya finalidad es la extracción de información desconocida de grandes bases de datos, esta información se caracteriza por tener una potencialidad para la realización de análisis y tomas de decisiones. La minería de datos automatiza el proceso de búsqueda de relaciones y patrones en los datos en bruto y proporciona resultados que pueden ser utilizados en un sistema de apoyo a las decisiones automatizadas o por quienes toman decisiones [26]. También a la minería de datos se la considera como el proceso de descubrir correlación significante, modelos, y tendencias mediante la excavación de cantidades grandes de datos guardados en los data Warehouse, usando estadísticos, máquina de aprendizaje, inteligencia artificial, y técnicas de visualización de los datos [27].

La minería de datos ayuda a los analistas a extraer información importante (no trivial) de cantidades grandes de datos, útiles para la toma de decisiones para la Entidad.

2.2 Relación entre la Minería de Datos y Data warehousing

En las actividades diarias de una Entidad Bancaria, tanto sus clientes como empleados, generan grandes cantidades de datos a través del uso de aplicaciones informáticas (ATM, depositarios, cajas, entre otros). Estos datos son conocidos como transacciones u operaciones, y son almacenados en bases de datos denominadas operacionales. Estas bases son diseñadas con la finalidad de soportar el volumen diario generado por los aplicativos por lo que en general no pueden almacenar datos históricos y tolerar consultas recurrentes.

El segundo tipo de base encontrada en la organización son los data warehouse, los cuales son diseñados para proveer soporte estratégico y un almacenamiento o acumulación de bases de datos operacionales. Un data warehouse es diseñado con la finalidad de recolectar datos que permitan responder a preguntas del negocio y apoyen a la toma de decisión; por consiguiente solo los datos que permitan cumplir con las dos premisas anteriormente planteadas, son extraidos desde las bases operacionales y guardardos en el data warehouse.

Otro aspecto a tomar en cuenta, es que debido a que el Data warehouse es diseñado para responder consultas inherentes al negocio, si la problemática o el tipo de decisión cambia por parte de la entidad, el diseño de cómo se almacenan los datos también cambia, es decir, el data warehouse debe ser implementado para transformaciones dinámicas de los datos.

Con respecto a la minería de datos, una de sus funciones es la de explotar y analizar las grandes masas de datos que posee la organización, ésto lo puede realizar de dos formas: Mediante la conexión directa hacia el data warehouse o por el uso de datos relevantes, que son de interés para el analista, que han sido extraídos y almacenados en un computador. Debido a esto es importante entender las demandas que tiene el usuario final, pues son estos requerimientos los que permitirán construir un adecuado data warehouse y por consiguiente realizar una buena minería de datos.

2.3 Proceso de la Minería de Datos

Como se describió anteriormente, existen varias metodologías enmarcadas en lo que es la minería de datos, sin embargo todas éstas siguen los siguientes procesos:

1. **Entendimiento del Negocio**.- El proyecto de minería de datos debe empezar entendiendo las necesidades, problemática u objetivos del negocios. Determinar los parámetros sobre los cuales se van a cumplir estos objetivos. Es claro que estas necesidades deben plasmarse en el modelo a desarrollar en la minería de datos.

2. **Entendimiento de los datos**.-Esta fase incluye analizar y considerar la data requerida para cumplir los objetivos planteados en el item anterior, esto envuelve la determininación de las fuentes que generan los datos, su disponibilidad y la estructura de los mismos. Aquí se puede definir el nivel mínimo al que puede llegar el proyecto. En muchos casos, los resultados obtenidos en esta fase, reformulan los objetivos planteados anteriormente.

3. **Extracción de los datos**.- Luego de haber determinado las fuentes de datos, entonces elementos relevantes son extraídos de bases de datos o de data Warehouse e incluidos dentro de una tabla, la cual contiene todas las variables necesarias para el modelamiento. Estas tablas son conocidas como tablas resumen ”rollup file” [26]. Cuando el conjunto de datos es demasiado grande, éste puede generar dificultades de rendimiento (procesamiento y almacenamiento) que se pueden evidenciar con tiempos extendidos en la ejecución de las técnicas de modelamiento o con la no finalización de las mismas, en estos casos generalmente se utilizan muestras de la base.

4. **Exploración de los datos**.- Un conjunto de pasos, de característica exploratoria, son realizados a los datos con la finalidad de revelar las relaciones existentes de la data y crear un resumen de estas propiedades. Este paso es conocido como EDA (Exploratory Data Analysis).

5. **Transformación de los datos**.- Basado sobre el resultado de la EDA, algunos procesos de transformación de los datos (categorización o normalización de variables, imputación de datos, entre otros) son realizados para destacar y tomar ventajas de las relaciones existentes entre las variables en el planteamiento del modelo.

6. **Modelamiento de los datos**.- Un conjunto de modelos de minería de datos son desarrollados mediante el uso de diferentes técnicas matemáticas, dependiendo del objetivo del análisis planteado y del tipo de variables envueltas. No todas las variables disponibles serán utilizadas para el modelo, sin embargo una reducción de los datos es a menudo necesario para seleccionar el conjunto de variables significativas para el modelo.

7. **Evaluación de los Modelamientos de los datos**.- Todas las técnicas matemáticas utilizadas en un modelo son evaluadas y la que presente un mejor resultado, de acuerdo a un criterio (test estadísticos), será seleccionada.

8. **Puntuación de los datos**.- El conjunto total de datos, al cual será aplicado el modelo seleccionado, es preparado en un proceso idéntico al que fue sometido el conjunto de datos de la tabla rollup file. Entonces el modelo óptimo seleccionado en el ítem anterior es aplicado a la totalidad de la base, y el resultado obtenido es almacenado en una tabla llamada scoring view.

La diferencia entre la minería de datos y el análisis que realizan otras disciplinas, radica en que los modelos tradicionales se basan en hipótesis y modelos previos; su finalidad es la verificación del modelo y el cumplimiento de dichas hipótesis. Mientras que la minería de datos permite establecer patrones de comportamiento e información relevante de forma inductiva mediante la utilización de técnicas de predicciones automáticas y descubrimiento autónomo de patrones.

Las técnicas de minería de datos puede clasificarse inicialmente en técnicas descriptivas y técnicas predictivas, tal y como se presenta en la tabla 2.1[3]:

Las técnicas predictivas especifican el modelo para los datos en base a un conocimiento previo. El modelo supuesto para los datos debe contrastarse después del proceso de minería de datos antes de aceptarlo como válido. Podemos incluir entre estas técnicas todos los tipos de regresión, series temporales, análisis de la varianza y covarianza, análisis discriminante, árboles de decisión y redes neuronales. Pero, tanto los árboles de decisión, como las redes neuronales y el análisis discriminantes son a su vez técnicas de clasificación que pueden extraer perfiles de comportamiento o clases, siendo el objetivo construir un modelo que permita clasificar los datos en grupos basados en los valores de las variables. El mecanismo de base consiste en elegir un atributo como raíz y desarrollar el árbol según las variables más significativas. Las técnicas de clasificación predictivas suelen denominarse técnicas de clasificación ad hoc ya que clasifican individuos u observaciones dentro de grupos previamente definidos.

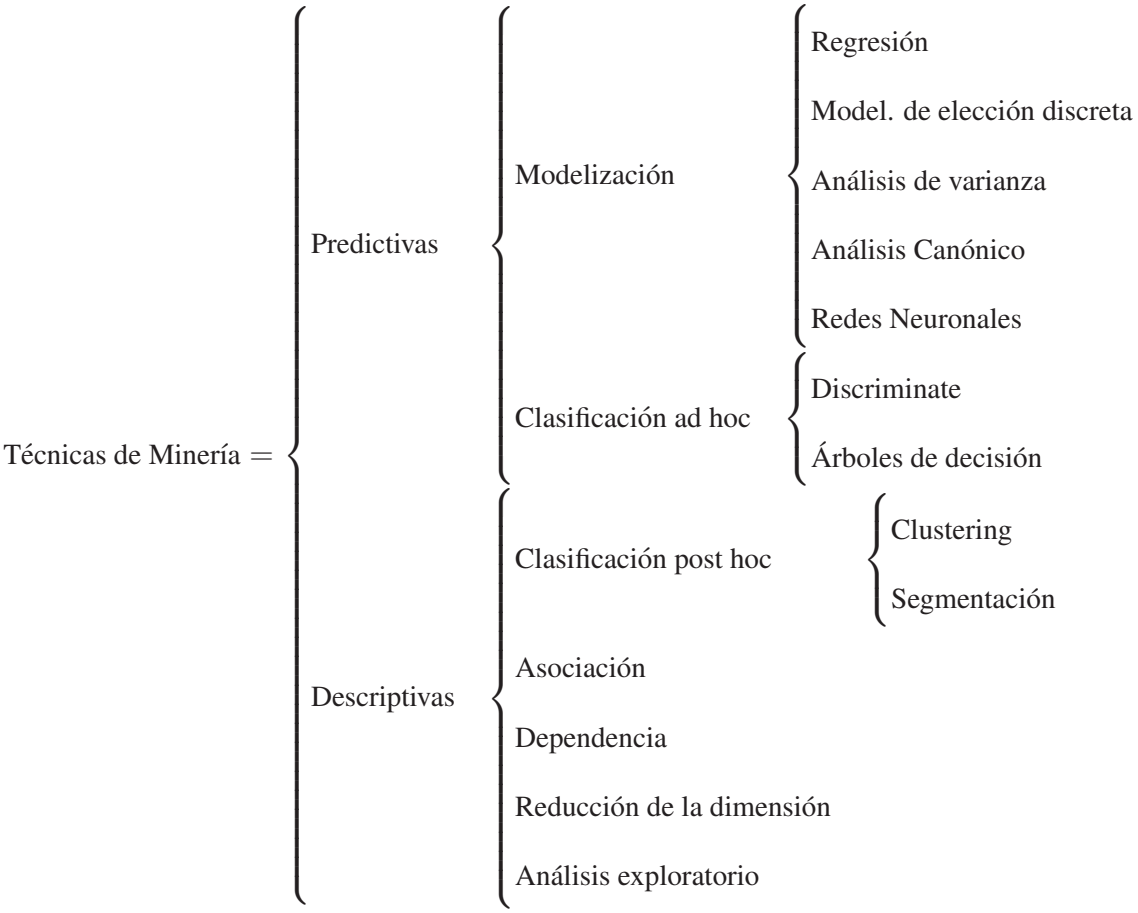


Tabla 2.1: Técnicas de Minería de datos

En las técnicas descriptivas no se asigna ningún papel determinado a las variables. No se supone la existencia de variables dependientes ni independientes y tampoco se supone la existencia de un modelo previo para los datos. Los modelos se crean automáticamente partiendo del reconocimiento de patrones. En este grupo se incluyen las técnicas de clustering y segmentación (que también son técnicas de clasificación en cierto modo), las técnicas de asociación y dependencia, las técnicas de análisis exploratorio de datos y las técnicas de reducción de la dimensión. Las técnicas de clasificación descriptivas se denominan técnicas de clasificación post hoc porque realizan la clasificación sin especificación previa de grupos.

Se observa que las técnicas de clasificación pueden pertenecer tanto al grupo de técnicas predictivas (discriminante, árboles de decisión y redes neuronales) como a las descriptivas (clustering y segmentación).

2.4 Aplicación de la minería de datos

Aunque la minería de datos recién empieza a tomar fuerza como una herramienta de análisis, las compañías en una gama amplia de industrias, tales como las finanzas, salud, fabricación, transportación, ya están usando las técnicas de minería de datos, para tomar ventaja de la historia de los datos. Usando tecnologías de reconocimiento de patrones y técnicas matemáticas y estadísticas se cierne información proveniente del data warehouse. La minería de los datos, ayuda a los analistas a reconocer hechos significantes, las relaciones, las tendencias, patrones, excepciones, y anomalías que podrían pasar inadvertidas. Para los negocios, el minéo de datos se usa para descubrir modelos y relaciones en los datos para ayudar a que las decisiones comerciales se las realicen bien. La minería de datos permite analizar las tendencias de ventas, desarrollar las campañas de mercadeo más inteligentes, y con precisión proyectar la lealtad del cliente. Los usos específicos de la mineria de datos incluyen:

- **Segmentación de mercados** Identificar las características comunes de los clientes, quienes compran los mismos productos de la organización.
- **Deserción de Clientes** Predecir cuáles clientes son probables a dejar la entidad e ir a un competidor.
- **Detección de Fraude** Identificar transacciones que son sospechosas o fraudulentas .
- **Marketing directo** Identificar prospectos más redituables para la organización, quienes pueden ser incluidos en campañas de fidelización y promoción, con la finalidad de obtener una alta tasa de respuesta.
- **Marketing iteractivo** Predecir qué información colocada en el sitio web es la más probable que un cliente visite.
- **Análisis de canastas de Mercados** Entender cuáles productos o servicios son comúnmente comprados en conjunto.
- **Análisis de tendencia** Revelar el comportamiento que tendrá un cliente a través del análisis histórico de sus datos

2.5 Metodología SEMMA

El instituto SAS define el concepto de minería de datos como el proceso de

Seleccionar, Explorar, Modificar, Modelizar y Valorar grandes cantidades de datos con el objetivo de descubrir patrones desconocidos que puedan ser utilizados como ventaja competitiva respecto a sus competidores. Cada uno de los proceso descritos anteriormente, vienen acompañados de un conjunto de técnicas estadísticas y de transformación de datos con la finalidad de crear una metodología de minería de datos.

En el proceso de seleccionar o tomar una muestra, se deben identificar el conjunto de datos de entrada (inputs), éstos pueden ser tablas de datos provenientes de Data Marts [1] o Data Warehouse, o ficheros de texto. En este proceso se debe determinar si se va a trabajar con el conjunto de Datos completos (población) o se va a tomar una muestra representativa del mismo. Se deben identificar los atributos de cada variable (tipo, longitud, formato), con la finalidad de que cuando se apliquen posteriores procesos de minería de datos, los resultados no se vean afectados o distorsionados por una mala estructuración de los datos.

Luego de haber identificado el conjunto de datos a utilizar viene el proceso de Explorar, el cual consiste en calcular estadísticos descriptivos y la utilización de herramientas de visualización. Para variables categóricas es de importancia el análisis de frecuencias y porcentaje con la finalidad de entender la estructura de los datos. Para las variables cuantitativas será necesario analizar medidas como el valor mínimo, máximo, media, moda y mediana. El uso de diagramas de Bigotes para la detección de valores anómalos dentro de la distribución de la variable. Todo esto para detectar valores nulos, inconsistencia y variables con ruidos.

El proceso de modificación tiene como objetivo la preparación de las variables, con la finalidad de prepararlas de forma adecuada para la aplicación de un modelo. En este proceso se pueden realizar transformaciones de variables, creación de nuevas variables, imputación de variables y reducción de variables.

Después de haber preparado las variables de la base de datos, el siguiente proceso es el del modelado, el cual es la parte central de la minería de datos, aquí se aplican algoritmos cuyo objetivo es la búsqueda de conocimiento de los datos preprocesados. La elección o aplicación de un algoritmo debe estar en concordancia al problema a resolver y a los tipos de datos con los que se está trabajando.

Como se observó en la tabla 2.1 los diferentes algoritmos se pueden clasificar en:

[1] Es un almacén de datos especializado, orientado a un tema, integrado, volátil y variante en el tiempo. Por especializado entiéndase que contiene datos para dar apoyo a un área específica de análisis de negocios; volátil hace referencia a que los usuarios pueden actualizar y crear datos para algún propósito.

- Descriptivos.
- Predictivos.

En el caso de los modelos descriptivos la meta es simplemente encontrar una descripción de los datos en estudio. Este tipos de modelos permiten resolver problemas como el de conocer cuáles son los clientes de una organización (características propias), o encontrar los productos que más frecuentemente se compran en conjuntos (canasta de productos) o los síntomas de enfermedades que se presentan juntos [25].

En el caso de los modelos predictivos la meta es obtener un modelo que en un futuro pueda ser aplicado para predecir comportamientos. Dentro de estos tipos de modelos se pueden resolver problemas como clasificaciones de clientes como buenos, malos o regulares; la probabilidad de que un cliente desierte o no de una entidad Bancaria.

Finalmente en el proceso de valoración, se realizan comparaciones entre los diferentes modelos estimados, ésto se realiza mediante la comparación de medidas de ajuste de los modelos, matriz de pérdidas y beneficios. Con ésto se elegirá el modelo más ajustado a los datos reales.

El instituto SAS ha implementado todo esta metodología en el software de minería de datos SAS Enterprise Miner, el cual presenta por cada proceso de la metodología SEMMA, diferentes técnicas de análisis estadísticos y de manipulación de datos, permitiendo transformar los datos en conocimiento.

Capítulo 3

Técnicas de Conglomerados

Para el área de Marketing de una Entidad Bancaria, uno de los objetivos fundamentales es el conocer el mercado en el cual ésta se desarrolla, entendiendo como mercado al conjunto de clientes que tienen distintos deseos, poder de compra, localización geográfica, actitudes y prácticas de compras [4]. Con la finalidad de entender el mercado, éste es dividido en partes más pequeñas conocidas como segmentos de mercados, los cuales están caracterizados por un conjunto de variables relevantes para el análisis. Para realizar una adecuada división del mercado se debe tomar en cuenta los siguientes pasos:

- **Definir el mercado relevante a ser analizado**.- Consiste en identificar el grupo general de clientes, para los cuales la entidad desea brindar sus productos y servicios. Es decir se debe definir el mercado objetivo para realizar una adecuada segmentación.

- **Analizar la viabilidad de realizar la segmentación**.- Aunque se haya determinado el mercado, es importante analizar los factores que influyen para realizar o no la segmentación, dentro de éstos se encuentran:

  - El segmento debe ser medible en tamaño y en característica.
  - El segmento debe ser significativo. Con la capacidad de generar suficientes beneficios a largo plazo, para merecer la atención de la comercialización por separado.
  - El segmento debe ser accesible dentro de los límites del presupuesto de la Entidad.
  - El segmento debe ser duradero en el tiempo con la finalidad de obtener beneficios a largo plazo, si la distinción entre segmentos provoca que disminuya el nivel de servicio, entonces no es adecuado un enfoque de segmentación.

- **Analizar la base de segmentación**.- Los mercados pueden ser segmentados en muchas formas, pero en general existen dos tipos de segmentaciones:

– **Segmentación de mercados Business-to-business.**

La segmentación por servicios

Este enfoque se refiere a cómo los clientes responden a servicios ofrecidos. Las empresas pueden ofrecer una gama de diferentes opciones de servicio y proporcionar distintos niveles de servicio dentro de las opciones, lo que permite grandes posibilidades para el diseño de paquetes de servicios adecuados a diferentes segmentos de mercado. Si un proveedor mide la importancia de los diferentes elementos de su servicio a través de la segmentación de mercados, éste puede responder a las necesidades identificadas de ese segmento y asignar una oferta de servicios adecuada a la misma.

La segmentación por valor buscado

Diferentes clientes pueden responder de manera diferente a los diferentes valores ofrecidos. Saber lo que los clientes valoran y qué peso se ponen en los elementos que diferencian una propuesta de valor puede ayudar a una empresa a desarrollar soluciones más específicas. Es fundamental tener una profunda comprensión de las motivaciones detrás de la decisión de compra.

segmentación del mercado Business-to-consumer

La segmentación geográfica

Este enfoque diferencia a los clientes sobre la base de donde se encuentran. Para que los clientes se pueden segmentar en áreas urbanas, suburbanas o los grupos rurales, éstos son generalmente segmentados por códigos postales, que también pueden representar los diferentes grupos en términos de riqueza relativa o los factores socioeconómicos.

La segmentación demográfica y socioeconómica

Esta segmentación se basa en una amplia gama de factores tales la como edad, sexo, tamaño de la familia, ingresos, educación, la clase social y origen étnico. Por lo tanto, es útil para indicar el perfil de personas que compran productos de una empresa o servicios.

Segmentación psicográfica

La segmentación psicográfica implica analizar las características de estilo de vida, las actitudes y personalidad. La investigación reciente en varios países sugiere que la población puede ser dividida entre diez y quince grupos, cada uno de los cuales tiene un conjunto identificable de estilo de vida, la actitud y características de la personalidad.

Segmentación por Beneficios

La segmentación de los clientes por beneficios, se basa en los bene-

ficios que se buscan a partir de un producto. Por ejemplo, los compradores de automóviles, buscan beneficios ampliamente variables, desde la economía en los combustibles, el tamaño y el espacio interior del vehículo, un buen rendimiento, fiabilidad o prestigio.

Segmentación por el uso

La Segmentación por uso es una variable muy importante para muchos productos. Por lo general, se divide a los consumidores en los grandes consumidores, los usuarios promedio, los usuarios ocasionales y los no usuarios del producto o servicio en cuestión. Los vendedores son a menudo relacionados con el segmento de usuarios grandes.

- **Elegir los segmentos importantes.**

  Las organizaciones están adoptando un enfoque más riguroso para realizar una segmentación . Por ejemplo, un banco analiza el número y el porcentaje de los jefes de hogares en cada categoría de estilo de vida en términos del perfil demográfico: entre ellos la edad del jefe del hogar, ocupación, la educación, la propiedad de la vivienda, los ingresos familiares anuales y saldos netos por valor y el promedio de penetración de los servicios por las cuentas transaccionales, cuentas de ahorro regulares y depósitos a plazo.

  El banco mide el beneficio producido en dólares, después de cada dólar de gastos al prestar un servicio a un segmento, con la finalidad de decidir si este segmento debe ser gestionado con todos sus esfuerzos.

Luego de haber realizado una breve introducción de la segmentación de Mercados, es indispensable entender la misma desde el enfoque de la minería de datos, para esto a continuación se presentan algunas definiciones de segmentación de acuerdo a diferentes autores y literatura en minería de datos. Para Roberto Dvoskin, la segmentación identifica grupos de clientes que presumiblemente se comporten de un modo similar ante un determinado producto o servicio. Para Tsiptsis y Chorianopoulos, la segmentación es el proceso de dividir a la base de clientes en distintos e internamente homogéneos grupos, con la finalidad de desarrollar diferentes estrategias de Marketing de acuerdo a la características que describen a cada conglomerado.

Dentro de la Minería de Datos, los modelos de conglomerados se encuentran dentro de los métodos multivariantes descriptivos, cuyo objetivo es el de determinar grupos naturales de observaciones, con la cualidad de que dentro de cada uno de los conglomerados los individuos sean semejantes o similares y entre grupos sean diferentes. Pues de aquí nace la conexión que existe entre la segmentación

de Mercados y los métodos de conglomerados, pues son estos, mediante ciertos algoritmos matemáticos, los que determinan los grupos o segmentos que describen de manera explícita el Mercado.

Entonces como se verá en la sección 3.2 uno de los objetivos a cumplir en este proyecto de titulación, es el describir las técnicas de agrupamiento de observaciones y así poder determinar los segmentos implícitos dentro de los datos que describen el Mercado, previamente se realizará un análisis del conjunto de datos, se determinará e identificará el número de variables y la población a ser analizada.

Con respecto al análisis de variables, las mismas deben estar enfocadas al objetivo del estudio, entonces el problema que se presenta es el de determinar cuáles y cuántas variables son necesarias para el análisis de conglomerados, Hand y Everitt (1987) encuentra que al ir incrementando el número de variables en el análisis de conglomerados, los resultados obtenidos identifican de mejor manera la estructura del conglomerado [29]. Price (1993), por otro lado encuentra que al aumentar el número de variables en el análisis de conglomerados los resultados identifican de manera pobre la estructura del conglomerado. Como se puede notar existe una contradicción en los dos análisis anteriores, Sin embargo Milligan (1980) determina que se debe analizar cuáles variables son relevantes para el análisis, puesto que los resultados obtenidos con variables no relevantes pueden ser muy distorsionadas. Entonces de aquí surge la necesidad de utilizar métodos multivariantes que permitan determinar cuáles variables son relevantes, dentro de estos métodos se destacan los métodos de correlación y componentes principales.[1]

Otro de los problemas a solucionar, es el tamaño adecuado de la muestra a ser analizada mediante una técnica de conglomerados. Se debe tener en cuenta que este problema no está relacionado a una estimación estadística (inferencia). La importancia del tamaño de la muestra radica en que ésta debe permitir representar a todos los grupos que se encuentren en la población, sobre todo los más pequeños. Es claro que si la muestra no es adecuada se pueda llegar a confundir entre grupos pequeños y observaciones anómalas (outliers).

Como se mencionó en el párrafo anterior, un problema que se presenta al determinar la estructura de los grupos en los datos, es la presencia de observaciones anómalas, las cuales hacen que las técnicas de conglomerados determine grupos no adecuados. Los valores anómalos pueden presentarse por las siguientes causas:

- Observaciones verdaderamente aberrantes que no son representantes de la

[1] Observar el apéndice para una mayor descripción teórica

población general.

- Observaciones que representan segmentos pequeños o insignificantes dentro de la población.
- Una muestra muy pequeña del grupo real en la población, lo que causa una representación pobre del grupo en la muestra.

En el primer caso, los valores anómalos distorsionan la estructura y derivan conglomerados que no representan la estructura real de la población. En el segundo caso, las observaciones anómalas son removidas hasta que los conglomerados resultantes representen los segmentos relevantes de la población. Sin embargo, en el tercer caso las observaciones anómalas deben ser incluidas en la solución, dado que éstas representan grupos válidos y relevantes. Por esta razón, una búsqueda preliminar de valores anómalos es siempre necesaria.

Dado que el análisis de conglomerados es una técnica multivariante, un conjunto de variables con diferente escala están inmiscuidas en el análisis, en general, variables con alta dispersión tienen un alto impacto al calcular el valor de semejanza entre observaciones, provocando una distorsión al agrupar las observaciones mediante una técnica de conglomerados. En este caso existen varias técnicas de estandarización de datos, que permiten homologar las escalas de las variables y encontrar grupos adecuados.[2]

3.1 Datos Atípicos

3.1.1 Introducción

Un supuesto básico de minería de datos es que los datos utilizados para la modelización, y más tarde para la puntuación, se obtiene a partir de (o generados por) un proceso que tiene un mecanismo dado (pero desconocido). Por lo tanto, las observaciones que no parecen seguir este mecanismo se define como los valores atípicos. Los valores atípicos puede ser consecuencia de:

1. Errores en los datos o que tengan observaciones que no se ajusten al mecanismo que genera los datos. Estos valores atípicos pueden ser identificados en las etapas de limpieza de la información. Sin embargo si estos valores se pasan por alto deben ser categorizados como valores perdidos.

[2]Observar el apéndice para una mayor descripción teórica

2. Eventos extraordinario los cuales generan las observaciones, lo que explica la singularidad de la observación. Por ejemplo si se analiza las transacciones de ventanilla y se produce una noticia de inestabilidad bancaria, los clientes acuden a la entidad Bancaria a retirar su fondos, produciendo un comportamiento diferente en el volumen de transacciones.

3. Observaciones que ocurren de un evento extraordinario para el cual el analista no tiene explicación.

4. Observaciones que pertenecen a un rango de valores sobre cada una de las variables analizadas, pero que en conjunto o combinadas son únicas.

Dado que el mecanismo que genera los datos es desconocido, la definición de los valores atípicos es subjetiva y especulativa. Por lo tanto, la exclusión de las observaciones que han sido etiquetadas como valores atípicos, debe hacerse después de un examen cuidadoso. En general, datos recogidos en condiciones de estrecho control, revelan que es frecuente que aparezcan entre 1 y 3 datos atípicos por cada 100 observaciones en la muestra. Cuando los datos se han recogido sin un cuidado especial, la proporción de datos atípicos puede llegar al 5 por 100 y ser incluso mayor [28].

3.1.2 Métodos de detección

Los valores atípicos pueden ser identificados desde una perspectiva univariante, bivariante o multivariante, basado sobre el número de variables consideradas. Hay tres enfoques principales para la identificación de valores atípicos.

Detección Univariante

La identificación univariante de valores atípicos examina la distribución de observaciones de cada variable utilizada para el análisis, selecciona como atípicos a aquellos valores que sobrepasan ciertos límites (rangos) establecidos en la distribución.

Los límites superior e inferior para la distribución, se determinan mediante una justificación estadística. Implementaciones comunes incluyen $\pm$ 3 desviaciones estándar del valor medio para las variables continuas. En el caso de las variables nominales y ordinales, un porcentaje mínimo del conteo de frecuencia de la categoría de 1 es de uso común.

Estos métodos tienen la ventaja de ser sencillo de aplicar y de interpretar. Sin embargo, no toma en cuenta la naturaleza multifactorial de los datos, es decir, se

basan en el examen de cada variable de manera independiente. Así, la interacción entre variables se ignora. Por otra parte, la aplicación de algunas transformaciones en variables continuas podrían reformar la distribución, y observaciones que se consideraron valores atípicos podrían llegar a ser normales otra vez. Por lo tanto, esta técnica solo debe utilizarse después de varios intentos para reformar la distribución de las variables y luego de examinar la interacción entre ellas.

Para evaluar un ejemplo de este método se simula un conjunto de datos con 1000 observaciones que siguen una distribución $N(100, 25)$ más 10 observaciones con $N(1, 16)$, su frecuencia es representada en el siguiente gráfico.

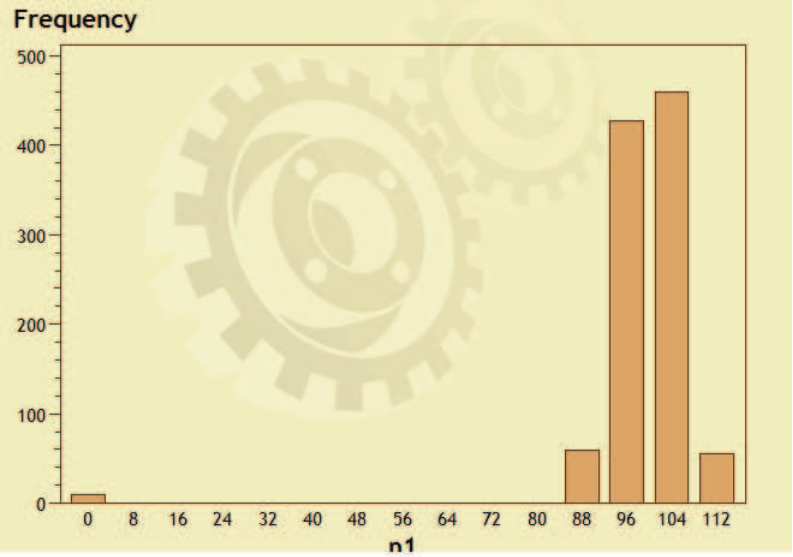


Figura 3.1: Frecuencia de Datos simulados

A continuación se calcula su media y desviación estándar:

estadístico	n1
Media	99.053
Desviación estándar	11.034

Tabla 3.1: Estadísticos

Luego se calculan los límites inferior=65.95 y superior =132.15 mediante la ecuación $\overline{x} \pm 3\sigma$, con estos valores se genera una nueva distribución de frecuencia, eliminando las observaciones que se encuentran fuera de este rango y que son de carácter atípicos:

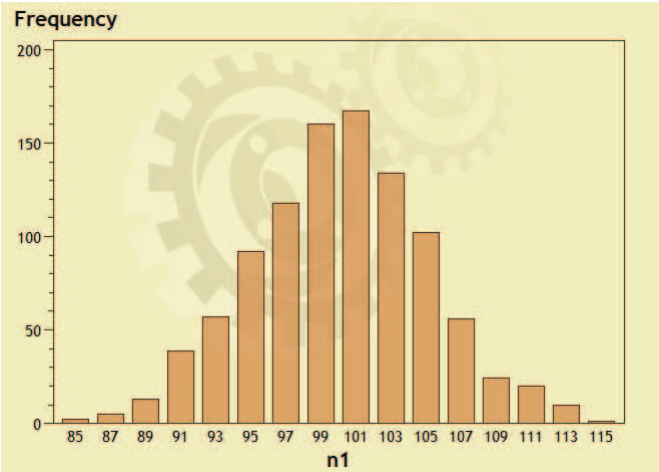


Figura 3.2: Frecuencia de Datos sin atípicos

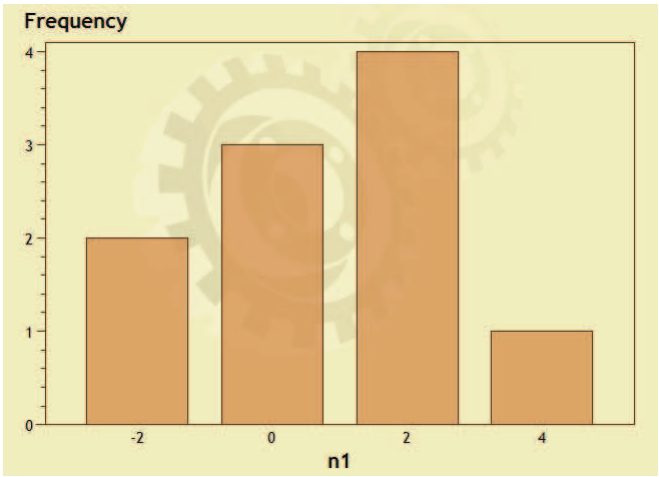


Figura 3.3: Frecuencia de Datos atípicos

En otros casos este mecanismo no es suficiente, puesto que el promedio es calculado con los valores atípicos, y ésto puede distorsionar los límites, sin embargo se puede utilizar en vez de la media, la mediana y en vez de sigma los valores del cuartil 1 y el cuartil 3, entonces se genera límites para la detección de valores anómalos.

Para el análisis mediante la mediana y los cuartiles, se utiliza el gráfico de caja con bigotes, Este gráfico permite analizar la simetría, detectar valores atípicos y descubrir el ajuste de distribuciones.

El gráfico divide los datos en cuatro áreas, una caja, que a la vez es dividida en dos áreas por una línea vertical (la mediana) y otras dos áreas representadas por los bigotes. La caja central encierra el 50% de los datos y si la línea no está en el centro no existe simetría. En los lados verticales de la caja se representa el primer y tercer cuartil. Se consideran valores atípicos a aquellos que se en-

cuentran fuera de la representación de la caja y los bigotes, los limites inferior y superior se calculan mediante las siguientes ecuaciones:

$C_1 - 1.5$ rango inter-cuartil y $C_3 + 1.5$ rango inter-cuartil respectivamente.

Con los mismo datos simulados, se analiza los datos atípicos mediante el gráfico de cajas y bigotes.

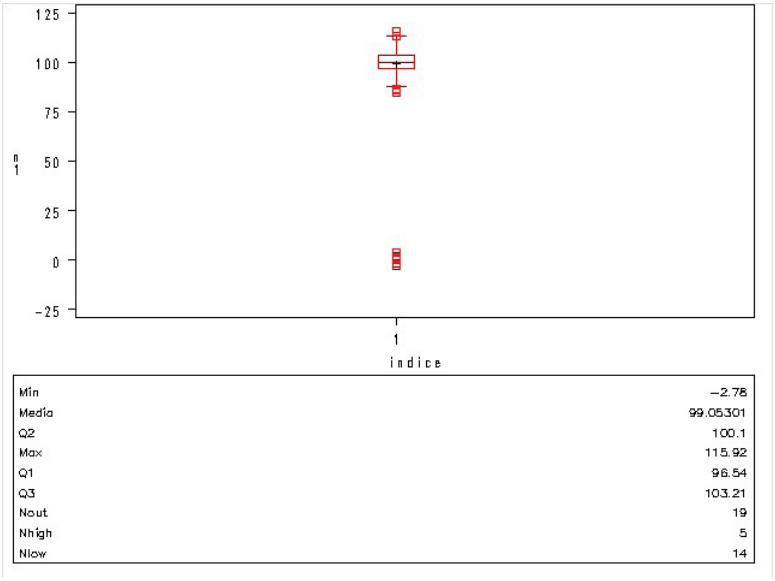


Figura 3.4: Gráfico de caja y bigote

Como se puede observar la mediana difiere de la media por lo cual no hay simetría, el límite inferior es 86.56 y el límite superior es 113.21. dando a notar una diferencia entre los limites calculados por este método y el que utiliza la media, esto se debe a que el método de caja y bigote no está afectado por la presencia de datos atípicos para su cálculo. Como se puede observar el gráfico nos presenta el número de observaciones etiquetadas como atípicos ”Nout” igual 19 y el número de observaciones sobre y bajo el límite ”Nhigh” y ”Nlow” respectivamente.

En la figura 3.5 se presenta los gráficos de caja y bigote para los datos sin valores anómalos.

Detección Bivariante

Con respecto al análisis de datos atípicos para un par de variables, o combinaciones de variables que pueden ser representadas por un gráfico de dispersión, los valores de observaciones que salen fuera del comportamiento común del conjunto de datos, pueden ser vistos como valores aislados en el gráfico de dispersión. Para determinar el rango esperado de observaciones anómalas, se trazan bandas de confianza que representan la distribución Normal bivariante con intervalos de

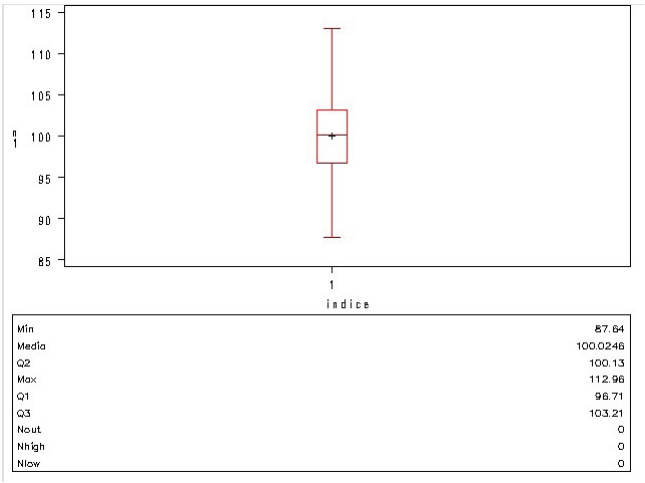


Figura 3.5: Gráfico de caja y bigote sin atípicos

confianza (nivel 90% o 95%) sobre el gráfico de dispersión, estas bandas permiten identificar fácilmente los valores atípicos.

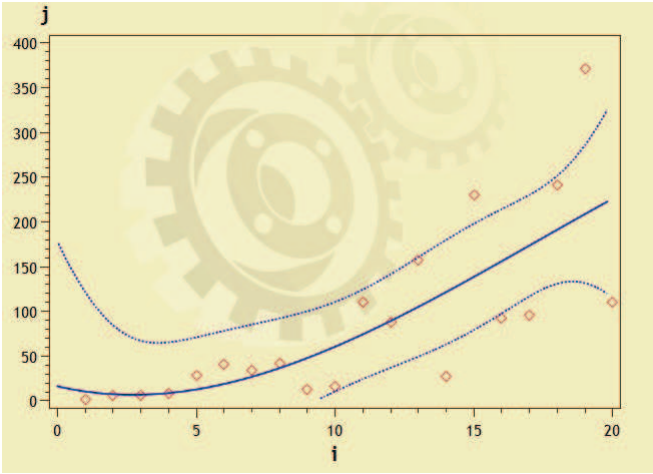


Figura 3.6: Gráfico de dispersión y detección de atípicos

Un método usado para la detección de valores atípicos, utiliza una variable dependiente y un conjunto de variables independientes para ajustar un modelo (mínimos cuadrados). Las observaciones que muestran una gran desviación del modelo ajustado son considerados como valores extremos. Estos métodos remedian de alguna manera las desventajas del método basado en rangos, es decir, toma en cuenta la naturaleza multifactorial de los datos. Sin embargo, se basan en asumir una cierta forma del modelo (generada por los datos). En aplicaciones típicas, el analista experimenta con diferentes técnicas de modelado y formas del modelo. Sin embargo, la utilización de uno de estos modelos para rechazar algunas observaciones porque no se adaptan al mismo puede ser engañoso. Por ejem-

plo, cuando uno utiliza la regresión lineal para encontrar los valores extremos, y luego utiliza un árbol de decisiones para desarrollar el modelo, los valores atípicos eliminados puede haber sido útil en la identificación de un nodo de importancia en el árbol de decisión.

Detección Multivariante

Dado que en muchos análisis se utilizan más de dos variables, la detección bivariante es un método no adecuado para el análisis de atípicos, Pues se requeriría de varios gráficos de dispersión, y está limitado a analizar dos variables a la vez. Entonces para realizar un análisis multivariante de datos atípicos la medida o distancia de Mahalanobis D^2 es utilizada. Este método calcula la distancia entre cada observación y el centro o media de todas las observaciones. Valores altos de D^2 representan valores alejados de la distribución de observaciones en un espacio muldimensional.

Una de las características estadísticas de la distancia de Mahalanobis es que al dividir su valor para el número de variables analizadas (D^2/df) se aproxima a una distribución t la misma que puede ser utilizada para analizar datos atípicos, valores de (D^2/df) exceden 2.5 en muestras pequeñas o que excedan en 3 o 4 en muestras grandes pueden ser considerados como atípicos. En el gráfico 3.7 se aplica el método de Mahalanobis para un conjunto de datos simulados, y se observa que los puntos identificados con la figura "+" son valores atípicos.

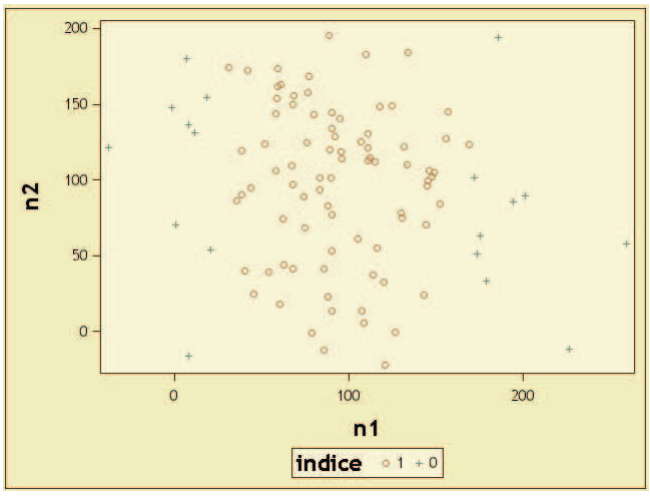


Figura 3.7: Gráfico de dispersión y detección de atípicos con distancia Mahalanobis

Otros tipos de métodos utilizan algoritmos de agrupamiento para encontrar subconjuntos de datos más pequeños (clusters). Las agrupaciones que contienen

un número muy pequeño de observaciones, idealmente una observación, se identifican como los valores extremos. Este es quizás el mejor método, que combina lo mejor de los métodos anteriormente analizados. Esto se justifica de la siguiente manera:

Los algoritmos de conglomerado agrupan observaciones similares en el mismo cluster basado en la distancia entre ellos. Esto es similar a la idea de especificar un rango aceptable , pero en este caso, el alcance se reduce a dentro de la agrupación o de un menor subconjunto de los datos.

La segunda característica es que los grupos con muy pocas observaciones se consideran como valores extremos. Esta idea es conceptualmente equivalente a rechazar las observaciones que no encajan en el modelo. El modelo aquí es definido como el conjunto de grupos con un número grande de observaciones, es decir, valores normales. Finalmente, debido a que los grupos se definen sobre la base de todas las variables, la naturaleza multifactorial de los datos ha sido atendida, sin asumir una forma específica o una estructura del modelo a utilizar.

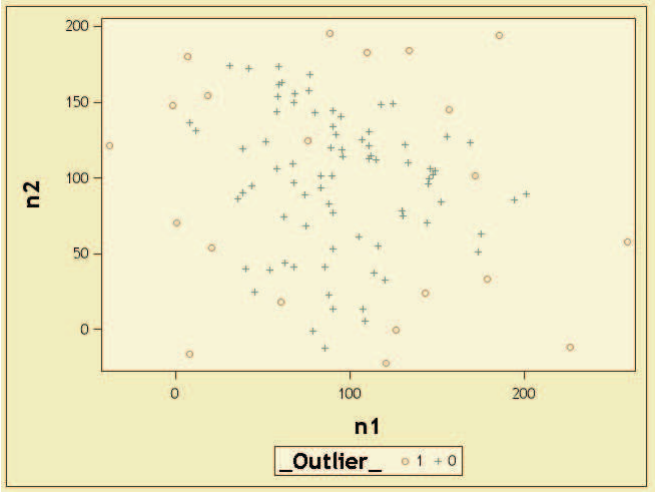


Figura 3.8: Gráfico de dispersión y detección de atípicos con método k-medias

En el gráfico 3.8 se realiza un análisis de datos atípicos con el método k-medias, lo interesante de este método, es que si se observa a los valores atípicos ”o”, existen varios de éstos dentro del conjunto de datos, a diferencia del método de Mahalanobis.

3.2 El análisis de conglomerados

El análisis de conglomerados se puede definir como el proceso de agrupar objetos o datos dentro de clases, grupos o nichos; con el fin de que los objetos dentro de

un grupo sean similares entre sí y que al mismo tiempo sean muy distintos a los objetos de los otros grupos. Estas técnicas, tienen sus raíces en diversos campos (estadística, química analítica, biología molecular, segmentación de mercados, etc.). A menudo se denomina aprendizaje no supervisado, porque no se tiene conocimiento a priori de una partición de los objetos. Esto está en contraste con el aprendizaje supervisado (por ejemplo la clasificación) para los cuales los objetos ya están etiquetados con clases conocidas [6].

Se debe entender que el realizar una definición formal de conglomerados es complicada y en mucho casos hasta engorrosa, esto es debido a que el agrupamiento de individuos u objetos se lo realiza bajo la percepción humana [30]. Por ejemplo al tomar una muestra de clientes de una entidad Bancaria, éstos poseen atributos (variables) que los caracterizan, dependiendo del análisis que se desee realizar, éstos se pueden agrupar en diferentes grupos o conglomerados. En efecto si se desea encontrar individuos que tengan estudios universitarios, entonces el atributo a analizar será el nivel de estudio, sin embargo si se desea analizar el grupo de clientes masculinos entonces la variable a analizar será el género. Como se puede ver en el primer ejemplo un grupo puede estar compuesto por hombres y mujeres los cuales tiene estudios universitarios, en cambio en el segundo ejemplo el grupo de hombres pueden tener educación primaria, secundaria, universitaria, etc. En los dos casos, el agrupamiento de los clientes es correcto, la elección del mejor depende del objetivo que se desea cumplir.

Es de vital importancia entender la diferencia entre conglomerados (clustering, clasificación no supervisada) y la clasificación supervisada. En la clasificación supervisada existe una variable objetivo (pre-clasificada), con la cual se determinan parámetros de clasificación, con la finalidad de poder etiquetar a nuevos individuos con una clase de la variable objetivo. En cambio, los algoritmos de conglomerados tratan de descubrir el agrupamiento natural del conjunto de datos, es decir, sin un conocimiento previo de los grupos.

A continuación se presentan las características del análisis de conglomerados.

- No hay distinción entre variables dependientes e independientes
- Se persigue establecer grupos homogéneos internamente y heterogéneos entre ellos.
- Se pueden agrupar casos o individuos pero también variables o características, a diferencia del análisis factorial que se centra en variables.
- Se trata de técnicas descriptivas, no de técnicas explicativas.

- Implícitamente se asume que la población a clasificar y las variables que se disponen, permiten realizar una posible clasificación, esto quiere decir que si existiera un conjunto de datos, los cuales producen una nube de puntos homogénea no tiene sentido realizar una clasificación con estas variables.

En la práctica las situaciones que se presentan son muy diferentes, se manejan grandes volúmenes de datos, y las variables a considerar son muchas. Por lo que la agrupación no resultaría tan evidente, así como también la existencia de grupos homogéneos y heterogéneos. Cabe señalar que existen varias formar de medir el parecido o la proximidad entre dos elementos, dando lugar a una gama de posibles resultados y haciendo mucho más difícil el análisis del problema planteado.

Para realizar la agrupación de los diferentes objetos dentro de los grupos, muchas técnicas empiezan calculando las similitudes por cada dos observaciones de toda la población. En muchos casos para medir las diferencia entre individuos se utiliza la definición de distancia. Otros métodos de conglomerados eligen aleatoriamente individuos como centros del conglomerados, y realizan una comparación de varianza dentro y fuera del grupo.

El análisis de conglomerados puede ser dividido en dos pasos fundamentales [31] :

- **Elección de una medida de proximidad**.- Una medida de similitud o proximidad es definida para medir la estrechez entre los objetos, mientras más estrechos se encuentren los objetos, más homogéneo es el grupo.

- **Elección de un algoritmo de agrupación**.- Sobre la elección de la medida de proximidad, los objetos son asignados a grupos, hasta que la diferencia entre grupos sea lo más grande posible y las observaciones dentro de cada grupo sean lo más homogéneas.

3.2.1 Proximidades

El punto de inicio del análisis de conglomerados es una matriz de datos X_{nxp}, cuyas columnas hacen referencia al conjunto de p variables y las filas contienen las medidas de las variable para cada uno de los n objetos o individuos. La proximidad entre objetos se representa mediante la matriz D_{nxn}.

$$D = \begin{pmatrix} d_{11} & d_{12} & \cdots & \cdots & d_{1n} \\ \vdots & d_{22} & \cdots & \cdots & \vdots \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & \cdots & d_{nn} \end{pmatrix}$$

La matriz D contiene las medidas de similitud o disimilitud entre los n objetos. Si los valores de d_{ij} son distancias, entonces esta evalúan la disimilitud, mientras más grande es la distancia, menos similares son los objetos. Si los valores d_{ij} son medidas de similitud entonces, mientras más grande es el valor calculado, más similares son los objetos. La distancia y la similitud tienen un comportamiento dual. Si d_{ij} es una distancia , entonces $d_{ij}^{'} = max_{ij}\{d_{ij}\} - d_{ij}$ es una medida de proximidad.

La naturaleza de las observaciones juega un papel importante en la elección de la medida de proximidad. Valores nominales conducen en general a medidas de proximidad, mientras que valores continuos llevan a evaluar matrices de distancias.

Medidas de similitud para variables discretas binarias

Las medidas de similitud son las más utilizadas para el análisis de la proximidad de individuos con características (variables) discretas, una de las variables discretas más comunes son las binarias, las cuales, acorde a Kaufman and Rousseeuw (1990), se pueden clasificar en variables simétricas o asimétricas, esta distinción se la realiza de acuerdo a si las características tienen igual importancia o no. Una variable binaria es simétrica si sus valores tienen igual significancia, mientras que es asimétrica si trata a uno de los valores, generalmente denotado por 1 como el más importante. Por ejemplo la variable genero tiene dos atributos hombre o mujer y pueden ser codificado como 1 y 0 sin afectar la evaluación de la similitud, clasificándose como variable binaria simétrica. Por otro lado si se realiza el análisis del uso de cajas de un banco por parte de los clientes, aquellos clientes que usan este canal se dicen que son similares, sin embargo el decir que no lo utilizan no quiere decir que los dos clientes necesariamente se parecen o son iguales, en este caso estamos hablando de variables binarias asimétricas.

Con la finalidad de medir la similitud entre objetos, siempre se empieza comparando pares de observaciones (x_i, x_j) donde $x_i^T = (x_{i1}, \ldots, x_{ip})$, $x_j^T = (x_{j1}, \ldots, x_{jp})$ y $x_{i,k}$, $x_{j,k} \in \{0, 1\}$, obviamente hay cuatro casos.

$$\begin{aligned} x_{ik} = 1 \quad , \quad x_{jk} = 1 \\ x_{ik} = 0 \quad , \quad x_{jk} = 1 \\ x_{ik} = 1 \quad , \quad x_{jk} = 0 \\ x_{ik} = 0 \quad , \quad x_{jk} = 0 \end{aligned}$$

el cual define:

$$\begin{aligned} a_1 &= \sum_{k=1}^{p} I(x_{ik} = 1, x_{jk} = 1) \\ a_2 &= \sum_{k=1}^{p} I(x_{ik} = 0, x_{jk} = 1) \\ a_3 &= \sum_{k=1}^{p} I(x_{ik} = 1, x_{jk} = 0) \\ a_4 &= \sum_{k=1}^{p} I(x_{ik} = 0, x_{jk} = 0) \end{aligned}$$

Se debe notar que los a_i dependen de los valores del par (x_i, x_k).

Las medidas de similitud que son usadas en la practica se derivan de la siguiente ecuación:

$$d_{ij} = \frac{a_1 + \delta a_4}{a_1 + \delta a_4 + \lambda(a_2 + a_3)} \tag{3.1}$$

donde los d_{ij} son elementos de la matríz de proximidad, δ y λ son conocidos como factores de peso. En la tabla se muestran algunas medidas de similitud de acuerdo a ciertos factores dados.

Nombre	δ	λ	definición
Jaccard	0	1	$\frac{a_1}{a_1+a_2+a_3}$
Tanimoto	1	2	$\frac{a_1+a_4}{a_1+2(a_2+a_3)+a_4}$
Unión simple	1	1	$\frac{a_1+a_4}{p}$
Russel y Rao	-	-	$\frac{a_1}{p}$
DIce	0	0.5	$\frac{2a_1}{2a_1+(a_2+a_3)}$
Kulezynski	-	-	$\frac{a_1}{a_2+a_3}$

Tabla 3.2: Coeficientes para medidas de Similitud

Donde: $p = a_1 + a_2 + a_3 + a_4$.

Ejemplo 3.2.1. *Se supone que se tiene tres observaciones,* $x_1 = (1, 1, 1, 0, 0, 0, 0, 0)$, $x_2 = (0, 0, 0, 1, 0, 0, 0, 1)$ *y* $x_3 = (0, 0, 1, 0, 0, 0, 0, 1)$. *Entonces la matriz de distancia para las similitudes [6] :*

Medida de similitud de Jaccard

$$D = \begin{pmatrix} 1 & 0 & 0.250 \\ & 1 & 0.333 \\ & & 1 \end{pmatrix}$$

Medida de similitud de Tanimoto

$$D = \begin{pmatrix} 1 & 0.231 & 0.454 \\ & 1 & 0.600 \\ & & 1 \end{pmatrix}$$

Medida de similitud de Unión simple

$$D = \begin{pmatrix} 1 & 0.375 & 0.625 \\ & 1 & 0.750 \\ & & 1 \end{pmatrix}$$

Propiedades de las medidas de Similitud

Sea X un conjunto y d es una medida de similitud sobre X si para todo $x_i, x_j, x_k \in X$ cumple con:

Propiedad 3.2.1. *Simetría*

$$d(x_i, x_j) = d(x_j, x_i)$$

Lo que implica que el cálculo de la medida de similitud no es afectada por el orden de las observaciones.

Propiedad 3.2.2. *No negativa*

$$d(x_i, x_j) \geq 0 \; \forall \, x_i \wedge x_j$$

Lo que implica que la medida de similitud no puede tomar valores menores a 0.
Si también cumple

Propiedad 3.2.3.

$$D(x_i, x_j)D(x_j, x_k) \leq [D(x_i, x_j) + D(x_j, x_k)]D(x_i, x_k) \; \forall \; x_i, \; x_j \wedge \; x_k$$

Lo que implica que el producto de dos similitudes es menor o igual a la suma de las similitudes entre las mismas observaciones multiplicado por la similitud de las dos observaciones extremas.

Propiedad 3.2.4.

$$D(x_i, x_j) = 1 \; \textit{Si} \; x_i = x_j$$

Esto significa que la similitud con respecto al mismo individuo es 1.
Entonces es llamada similitud métrica.

3.2.2 Similitud para variables discretas con más de dos valores

Para variables discretas con más de dos categorías, una estrategia simple es la de transformar a la variable en un conjunto de variables binarias. Por ejemplo si se toma la variable contacto con el cliente, teniendo como categorías alto, medio y bajo, puede ser codificada como tres nuevas variables binarias contacto-alto, contacto-medio y contacto-bajo, entonces si un individuo o registro pertenece a una de las categorías, por ejemplo alto entonces el tomará el valor de 1 en la variable contacto alto y cero en las otras variables. La desventaja de este método es que para introducir una variable con n categorías, se necesitan introducir n variables binarias.

Un método más utilizado para determinar la similitud entre dos observaciones, es aquel que se basa en el criterio de coincidencia, es decir, dado un par de observaciones x_i y $x_j \in \mathbb{R}^d$ la similitud se define como:

$$S(x_i, x_j) = \frac{1}{d} \sum_{l=1}^{d} S_{ijl} \tag{3.2}$$

donde:

$$S_{ijl} = \begin{pmatrix} 0 & Si\ x_i\ y\ x_j & \text{no coinciden en la l-ésima característica} \\ 1 & Si\ x_i\ y\ x_j & \text{coinciden en la l-ésima característica} \end{pmatrix}$$

$$d = \text{número de variables}$$

Ejemplo 3.2.2. *Suponga que se tiene la tabla 3.3 con información demográfica de 4 clientes:*

Obs	estado civil	nivel estudio	trabajo
1	1	1	1
2	3	3	0
3	1	2	1
4	3	3	0

Tabla 3.3: Datos para análisis de similitud

Aplicando la ecuación 3.2 se obtiene la siguiente matriz de similitud:

$$D = \begin{pmatrix} 1 & 0 & 0.66 & 0 \\ & 1 & 0 & 1 \\ & & 1 & 0 \\ & & & 1 \end{pmatrix}$$

Como se puede analizar las observaciones 2 y 4 son iguales por lo que reciben un valor de 1 en la matriz D, las observaciones 1 y 3 solo difieren en la variable nivel de estudio por lo que recibe el valor de 0.66.

3.2.3 Medidas de disimilitud o distancia para variables continuas

Las medidas de disimilitud o distancia son las que nos permite evaluar que tan lejanos se encuentran dos observaciones. Con respecto a esta definición, una extensa variedad de medidas de distancia han sido generadas, dentro de la cuales se destaca la distancia de Minkowski y se define como sigue:

$$d_{ij} = \|x_i - x_j\|_r = \left\{ \sum_{k=1}^{p} \mid x_{ik} - x_{jk} \mid^r \right\}^{1/r} \tag{3.3}$$

Aquí x_{ik} denota los valores de la k-ésima variable del individuo i. Es claro que $dii = 0$ para $i = 1, \ldots, n$. los diferentes tipos de distancia se generan al variar r dando lugar a medidas de disimilitud con diferente peso. Por ejemplo al tomar r el valor de 1 tenemos:

Distancia de Manhattan

$$d_{ij} = \sum_{k=1}^{p} \mid x_{ik} - x_{jk} \mid \tag{3.4}$$

Si r=2 tenemos:

Distancia Euclediana

$$d_{ij} = \left\{ \sum_{k=1}^{p} \mid x_{ik} - x_{jk} \mid^2 \right\}^{1/2} \tag{3.5}$$

Si r=∞ tenemos:

Distancia Máxima

$$d_{ij} = max_{1 \leq k \leq p} \mid x_{ik} - x_{jk} \mid \tag{3.6}$$

Ejemplo 3.2.3. *Se supone que se tiene tres observaciones,* $x_1 = (0, 0)$, $x_2 = (1, 0)$ *y* $x_3 = (5, 5)$ *[6]. Entonces la matriz de distancia para la norma* L_1 *es :*

$$D_1 = \begin{pmatrix} 0 & 1 & 10 \\ 1 & 0 & 9 \\ 10 & 9 & 0 \end{pmatrix}$$

Y para la norma L_2 o norma Euclidiana

$$D_1 = \begin{pmatrix} 0 & 1 & 50 \\ 1 & 0 & 41 \\ 50 & 41 & 0 \end{pmatrix}$$

Propiedades de las medidas de distancias

Sea x_i y x_j individuos de un conjunto de datos X, y sea D la medida de distancia entonces ésta debe cumplir las siguientes propiedades.

Propiedad 3.2.5. *Simetría*

$$D(x_i, x_j) = D(x_j, x_i)$$

Propiedad 3.2.6. *No negativa*

$$D(x_i, x_j) \geq 0 \; \forall \, x_i \; \wedge \; x_j$$

Si también cumple

Propiedad 3.2.7. *Desigualdad Triangular*

$$D(x_i, x_j) \leq D(x_i, x_k) + D(x_k, x_j) \; \forall \; x_i, \; x_j \; \wedge \; x_k$$

Propiedad 3.2.8. *Reflexividad*

$$D(x_i, x_j) = 0 \; \textit{Si} \; x_i = x_j$$

Entonces es llamada Métrica. Si no cumple con la desigualdad triangular es llamada semi métrica.

3.2.4 Medidas para variables mixtas

En mucha de las ocasiones, el conjunto de datos analizados puede tener más de un tipo de variable (continua, discreta). En este caso las medidas de similitud y de distancia analizadas anteriormente no pueden ser aplicadas directamente. Gower (1971) proponen algunas medidas generales

Un coeficiente de similitud general

El coeficiente de similitud general propuesto por Gower, ha sido implementado y usado para medir la similitud entre dos tipos de variables en el conjunto de datos.

Sea X y Y dos observaciones $\in \; \mathbb{R}^d$ entonces el coeficiente de similitud de Gower se define como:

$$S_{gower}(X,Y) = \frac{1}{\sum_{k=1}^{d} w(x_k, y_k)} \sum_{k=1}^{d} w(x_k, y_k) s(x_k, y_k) \tag{3.7}$$

donde $s(x_k, y_k)$ es un componente que mide la similitud para el atributo k-ésimo y $w(x_k, y_k)$ es un índice que toma los valores de 1 o 0 dependiendo si la comparación es válida o no para el atributo k-ésimo de las dos observaciones. Ellos son definidos respectivamente para diferentes tipos de atributos. Sean x_k y y_k, el atributo k-ésimo de X y Y,respectivamente.

Entonces $s(x_k, y_k)$ y $w(x_k, y_k)$ son definidas como sigue:

- Para atributos cuantitativos x_k y y_k, $s(x_k, y_k)$ se define como:

$$s(x_k, y_k) = 1 - \frac{\mid x_k - y_k \mid}{R_k}$$

  Donde $R_k = \max_m X_{mk} - \min_m X_{mk}$ es el rango del atributo k-ésimo ; $w(x_k, y_k) = 0$ si los puntos de datos x o y tienen valores perdidos para el atributo k-èsimo ; para los otros casos $w(x_k, y_k) = 1$;

- Para atributos binarios x_k y y_k, $s(x_k, y_k) = 1$ si ambos puntos de los datos x y y tiene el atributo k-ésimo ”Presente”; caso contrario $s(x_k, y_k) = 0$; $w(x_k, y_k) = 0$ si ambos puntos de datos x y y tiene en el atributo k-èsimo ”ausente”; en otros casos $w(x_k, y_k) = 1$

- Para atributos nominales o categóricos x_k y y_k,$s(x_k, y_k) = 1$ si $x_k = y_k$; en otro casos $s(x_k, y_k) = 0$; $w(x_k, y_k) = 0$ si los puntos de datos x o y tienen valores perdidos en el atributo k-ésimo; en otros caso $w(x_k, y_k) = 1$

Ejemplo 3.2.4. *Suponga que se tiene la tabla 3.4 con información demográfica de 4 clientes:*

Obs	edad	genero
1	22	1
2	35	0
3	40	0
4	32	1

Tabla 3.4: Datos para análisis de similitud variables mixtas

Utilizando la ecuación 3.7 se obtiene la siguiente matriz D

$$D_1 = \begin{pmatrix} 1 & 0.277 & 0 & 0.722 \\ & 1 & 0.86 & 0.916 \\ & & 1 & 0.555 \\ & & & 1 \end{pmatrix}$$

Un coeficiente de distancia general

Para medir la distancia entre dos puntos de datos cualesquiera x e y, la distancia general o coeficiente de Gower(1971) se define como:

$$d_{gower}(X,Y) = \left(\frac{1}{\sum_{k=1}^{d} w(x_k, y_k)} \sum_{k=1}^{d} w(x_k, y_k) d^2(x_k, y_k) \right)^{\frac{1}{2}} \tag{3.8}$$

Donde $d^2(x_k, y_k)$ hace referencia a la distancia al cuadrado para el atributo k-ésimo y $w(x_k, y_k)$ toma el mismo papel como en el coeficiente de similitud general, es decir, Si ambos x y y tienen observación para el k-ésimo atributo entonces $w(x_k, y_k) = 1$; en otros casos $w(x_k, y_k) = 0$. Para diferente tipos de atributos, $d^2(x_k, y_k)$ es definido diferente, como sigue:

- Para atributos ordinales $d(x_k, y_k)$ es definido como:

  $$d(x_k, y_k) = \frac{\mid x_k - y_k \mid}{R_k}$$

  Donde R_k es el rango del atributo k-ésimo

- Para atributos cuantitativos, $d(x_k, y_k)$ es definido como:

  $$d(x_k, y_k) = \mid x_k - y_k \mid$$

- Para atributos binarios, $d(x_k, y_k) = 0$ si ambos i y j tiene el atributo k-ésimo "Presente" o "Ausente": en otros casos $d(x_k, y_k) = 1$

- Para atributos nominales o categóricos, $d(x_k, y_k) = 0$ si ambos $x_k = y_k$; en otros casos $d(x_k, y_k) = 1$

Ejemplo 3.2.5. *Suponga que se tiene la tabla 3.5 con información demográfica de 4 clientes:*

Utilizando la ecuación 3.7 se obtiene la siguiente matriz D

$$D_1 = \begin{pmatrix} 0 & 0.72 & 1 & 0.80 \\ & 0 & 0.73 & 0.71 \\ & & 0 & 0.44 \\ & & & 0 \end{pmatrix}$$

Obs	edad	genero
1	22	1
2	35	0
3	40	0
4	32	1

Tabla 3.5: Datos para análisis de distancia variables mixtas

3.3 Métodos de conglomerados

En la actualidad, los métodos de conglomerados han tenido mucha diversificación y su desarrollo a ido creciendo de manera significativa, existen varias clasificaciones de estos métodos, dentro de la literatura analizada se destacan los trabajos realizados por Han y Kamber (2001), y Guojun Gan , Chaoqun Ma, y Jianhong Wu (2007), los cuales clasifican a los métodos de la siguiente manera:

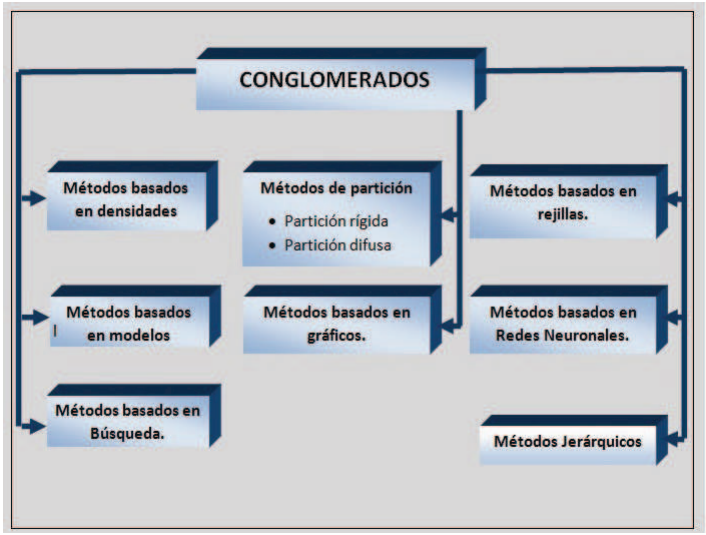


Figura 3.9: Clasificación de Conglomerados /Han y Kamber

3.3.1 Métodos de Partición

Dado un conjunto de datos con n observaciones, los métodos de partición obtienen k particiones de los datos originales, donde cada partición representa un conglomerado, y deben satisfacer las siguientes condiciones:

- Cada grupo debe contener al menos una observación.

- Cada objeto debe pertenecer a un solo grupo. (excepto en los métodos difusos).

Sea k el número de grupos a determinar, los algoritmo de partición empiezan definiendo grupos, y luego intercambia elementos entre ellos hasta llegar a un umbral de optimización. El criterio de una buena partición es la que los objetos en el mismo conglomerado deben ser ”cerrados” y objetos de los diferentes conglomerados deben ser diferentes.

Este proceso usualmente está acompañado por la optimización de una función. Específicamente, dado un conjunto de datos $x_j \in \mathbb{R}^d \, j = 1, \ldots, N$ el objetivo de los algoritmos de partición es organizar a los datos dentro de k grupos $(c_1, \ldots, c_k)$ mientras maximiza o minimiza una especifica función objetivo J.

Entonces de manera formal las técnicas de partición se define como:

Dado un conjunto de objetos $X_j \in \mathbb{R}^{d3} \, j = 1, \ldots, N$, el objetivo es organizarlos dentro de k conglomerados $C = \{C_1, \ldots, C_k\}$. Entonces se define una función J con el criterio de la suma de errores al cuadrado como sigue:

$$J(\Gamma, M) = \sum_{i=1}^{K} \sum_{j=1}^{N} \gamma_{ij} \|X_j - m_i\|^2 = \sum_{i=1}^{K} \sum_{j=1}^{N} \gamma_{ij} (X_j - m_i)^T (X_j - m_i) \tag{3.9}$$

donde:

$\Gamma = \{\gamma_{ij}\}$ es una matriz de partición,

$$\gamma_{ij} = \begin{cases} 1 \text{ si } X_j \in \text{ al conglomerado } i \\ 0 \text{ para otros casos} \end{cases}$$

con $\sum_{i=1}^{K} \gamma_{ij} = 1 \; \forall \; j$;

$M = [m_1, \ldots, m_k]$ es la matriz de centros o medias y

$m_i = \frac{1}{N_i} \sum_{j=1}^{N} \gamma_{ij} X_j$ es la media muestral para el i-ésimo conglomerado con N_i objetos.

La partición que minimice J es definida como óptima y es llamada la partición de mínima varianza. La suma de errores al cuadrado es apropiado para conglomerados que son compactos y bien separados. Sin embargo este criterio puede ser sensitivo cuando en los datos existen valores atípicos.

[3] Define el espacio de las variables

Algoritmo de k-medias

El método de K-medias es uno de los algoritmos de conglomerados más conocidos, desarrollado por Forgy en 1965 y Macqueen 1967, el método busca una óptima partición de los datos mediante la minimización de la suma de error al cuadrado con un método de optimización iterativa. El algoritmo básico del K-medias se resumen en los siguientes pasos:

1. Se determina aleatoriamente k partición. Calculando la matriz $M = [m_1, \ldots, m_k]$;
2. Entonces, se asigna cada observación del conjunto de datos al conglomerado C_l más cercano, es decir,

   $$x_j \in C_l,\ si\ \|x_j - m_l\| < \|x_j - m_i\|$$

   para

   $$j = 1, \ldots, N\ i \neq l,\ y\ i = 1, \ldots, k;$$

3. Luego, se recalcula la matriz M utilizando la partición actual.

   $$m_i = \frac{1}{N_i} \sum_{x_j \in C_i} X_j$$

4. Se repite el paso 2 y 3 hasta que no existan cambios significativos (criterio de parada) en cada conglomerado.

Como se observa el algoritmo de k-medias puede ser dividido en dos fases:

- La fase de inicialización que corresponde a los numerales 1, 2 y 3.
- La fase de iteración que corresponde al numeral 4.

Ventajas del Algoritmo de k-medias

- Es eficiente para conjuntos de datos grandes, su complejidad es linealmente proporcional al tamaño del conjunto de datos $O(nKt)$. donde n es el número de observaciones, k el número de conglomerados y t el número de iteraciones.
- A menudo termina en un mínimo local.
- Los conglomerados tienen formas convexas, como una pelota en el espacio tridimensional.
- Trabaja con datos numéricos.

Desventajas del algoritmo de k-medias

- El método de k-medias no es conveniente para descubrir conglomerados con formas no convexas o para conglomerados de tamaño muy diferente.
- Es sensitivo al ruído y datos anómalos, porque un número pequeño de tales datos puede substancialmente influir en el valor medio.
- El desempeño es dependiente de la inicialización de los centros.

Ejemplo 3.3.1. *Suponga que se tiene la tabla 3.6 con datos de dos variables* x_1 *y* x_2*:*

Obs	x_1	x_2
1	1	1
2	1.5	2
3	3	4
4	5	7
5	3.5	5
6	4.5	5
7	3.5	4.5

Tabla 3.6: Datos

A continuación se presenta su gráfico para analizar de manera visual los grupos existentes.

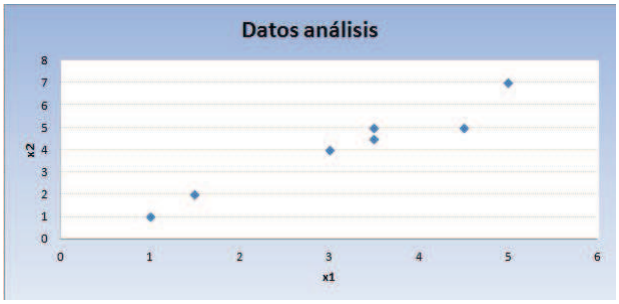


Figura 3.10: Datos para Análisis de Conglomerados

Como primer paso se toma k=2, esto quiere decir que se va agrupar los datos en dos conglomerados. Entonces se seleccionan las observaciones 1, 2 y 3 como el primer grupo y las observaciones 4,5,6 y 7 como el segundo grupo. Luego por cada conglomerado se calcula los promedios o centros de cada variable, obteniendo:

Conglomerado	x_1	x_2
1	1.83	2.33
2	4.12	5.37

Tabla 3.7: Centros de Conglomerados

Entonces, de acuerdo a la ecuación 3.9 se determina la distancia entre la observación i y los centros j con $j = 1, 2$, y así se determina si cada observación está adecuadamente clasificada.

cluster	distancia al cluster 1	distancia al cluster 2	Es correcta
1	1.572330189	5.376453292	si
1	0.471404521	4.275657844	si
1	2.034425936	1.7765838	no
2	5.639641439	1.845602883	si
2	3.144660377	0.728868987	si
2	3.771236166	0.530330086	si
2	2.733536578	1.075290658	si

Tabla 3.8: Distancias de las observaciones a los conglomerados

En la tabla 3.8 se visualiza que la observación 3 no está correctamente clasificada, por lo que se asigna al otro conglomerado, y entonces se vuelve a calcular los nuevos centros.

Conglomerado	x_1	x_2
1	1.25	1.5
2	3.9	5.1

Tabla 3.9: Nuevos centros de Conglomerados

En el gráfico 3.11 se presentan los datos y los centros calculados.

Con el software R el método de K medias se lo puede obtener mediante la función Kmeans que se encuetra dentro del paquete stats[4] el cual tiene las opciones siguiente:

kmeans(x, centers, iter.max = 10, nstart = 1, algorithm = c(”Hartigan-Wong”, ”Lloyd”, ”Forgy”,”MacQueen”))

[4]Package stats/http://stat.ethz.ch/R-manual/R-patched/library/stats/html/kmeans.html

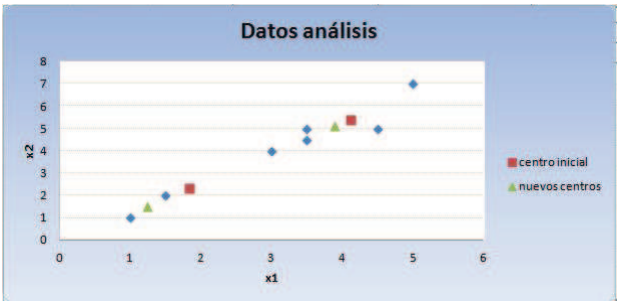


Figura 3.11: Gráfico de Conglomerados k-medias

donde x matríz de datos.

$Centers$ puede ser un vector con los centros iniciales, o tambien el número de grupos ha estimar (k).

$Iter.max$ es el máximo número de iteraciones a realizar.

Si la opción center es un número, nstar determina cuantos conjuntos aleatorios pueden ser elegidos.

$algorithm$: determina el algoritmo a utilizar.

En la figura 3.12 se presenta la sentencia utilizada para obtener los resultados del ejercicio anterior.

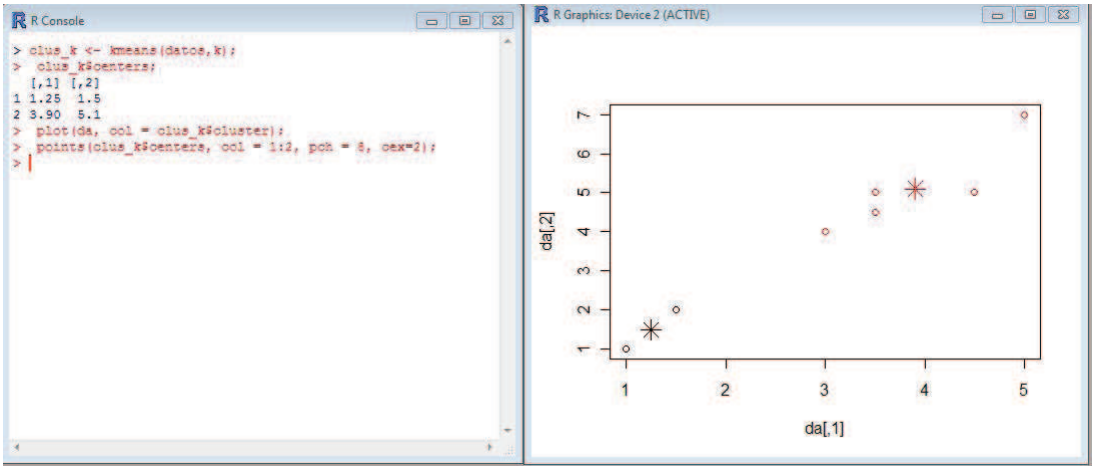


Figura 3.12: Conglomerados k-medias en R

Método Difuso C-Medias (FCM)

Introducción Lógica Difusa

La mayor parte del razonamiento humano se encuentra más cerca de lo aproximado que de lo preciso. De un modo bastante eficiente, los humanos somos capaces de tomar decisiones racionales con información imprecisa o incompleta, reconocer voces e imágenes distorsionadas, resumir y completar datos parcialmente desconocidos.

Desde esta perspectiva, la lógica difusa puede verse como un intento de construir un modelo del razonamiento humano que refleje su carácter aproximado o cualitativo. En este modelo, el razonamiento preciso debe verse como un caso límite que estará incluido en el anterior.

La utilidad de la lógica difusa, está en la posibilidad de tratar problemas demasiado complejos o muy mal definidos de tal forma que los métodos tradicionales no lo pueden manejar o admitir. El término difuso se debe a una connotación de incertidumbre, es decir, de imprecisión; lo difuso puede entenderse como la posibilidad de asignar más valores de verdad a los enunciados, diferentes a los clásicos ”falso” o ”verdadero” o en el término de clasifiación ”pertenece” o ”no pertenece” [32].

El desarrollo de la lógica difusa ha sido inspirado y guiado por Zadeh (1965), profesor de Ingeniería Electrónica de la Universidad de California en Berkeley, quién ahora es considerado como ’alma mater’ de esta teoría. La lógica difusa ha surgido como una herramienta para el control de subsistemas y procesos industriales complejos, así como para la electrónica de entretenimiento y hogar, sistemas de diagnóstico y otros sistemas expertos. Aunque la lógica difusa se formuló en Estados Unidos, el crecimiento rápido de esta tecnología ha comenzado desde Japón y ahora nuevamente ha alcanzado Estados Unidos y también Europa. En Japón y China todavía es un fenómeno a nivel científico, donde el número de patentes de aplicaciones aumenta exponencialmente.

La lógica difusa es cosiderada como una generalización de la teoría de conjuntos, debido a que ésta permite que los elementos de un universo, pertenezcan de manera parcial (grados de pertenecia, membresia o adhesión) a varios conjunto mediante una función caracterísitica. Veamos un ejemplo para explicar este concepto. En primer lugar consideremos un conjunto X con todos los números reales entre 0 y 10 que llamaremos universo. Se define un subconjunto A de X con todos los números reales en el rango entre 5 y 8. En virtud a la teoría clásica de conjuntos, la función característica o de pertenencia de A, asigna un valor de 1 ó 0 a cada uno de los elementos de un conjunto X, dependiendo de si el elemento

está o no en el subconjunto A. Otra forma de expresar ésto, es decir que el "grado de membresía " de un elemento particular con respecto a un determinado grupo es la unidad o cero.

$$x \in X \to \mu_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

En contraste con esto, en el caso de conjuntos difusos, el grado de pertenencia puede ser cualquier valor continuo entre cero y uno, y un particular elemento puede estar asociado con más de un grupo. En general, esta asociación involucra diferentes grados de adhesión con cada uno de los conjuntos difusos.

Conjuntos Difusos

El concepto de conjunto difuso generaliza la idea del conjunto clásico, debido a que permite describir conceptos en los que el límite entre tener una propiedad y no tenerla no es completamente nítido o claro.

Definición 3.3.1. *Sea X un conjunto clásico de referencia. Un conjunto difuso A sobre X queda definido por medio de una función característica [1].*

$$\mu_A : X \mapsto [0, 1]$$

Si $x \in X$, la expresión $\mu_A(x)$ se interpreta como el grado de membresía o pertenencia del elemento x en el conjunto difuso A. El conjunto X es llamado el dominio de A, y se denota $dom(A) = X$.

Para simplificar la notación, se puede escribir $A(x)$ en lugar de $\mu_A(x)$.

Además, si el conjunto $X = \{x_1 \dots x_n\}$ es finito, se puede caracterizar al conjunto difuso A de la siguiente manera:

$$\mu_A = \{\mu_1/x_1, \dots, \mu_n/x_n\}$$

donde el término μ_i/x_i expresa que $A(x_i) = \mu_i$. Esta notación puede emplearse también cuando el conjunto X sea infinito pero A posea un número finito de elementos con membresía no nula.

Ejemplo 3.3.2. *Un número difuso es un conjunto difuso sobre **R**. Aunque en principio cualquier función de membresía es valida, existen familias de números difusos con funciones de membresía parametrizadas. Algunos ejemplos de estas funciones de membresía son:*

- Funciones triangulares, con parámetros a, b, c, donde:

$$triang_{a,b,c}(x) = max[min(\frac{x-a}{b-a}, \frac{c-x}{c-b}), 0] \tag{3.10}$$

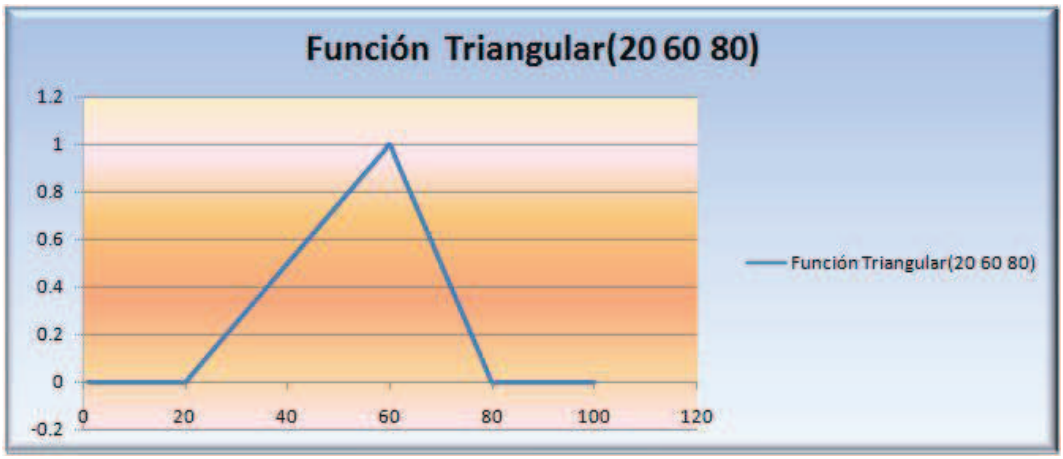


Figura 3.13: Función de membresía triangular

En la figura 3.13 se puede ver gráficamente la función de membresía triang(20 60 80).

- Funciones trapezoidales, con parámetros a, b, c, d, donde:

$$trap_{a,b,c,d}(x) = max[min(\frac{x-a}{b-a}, 1, \frac{d-x}{d-c}), 0] \tag{3.11}$$

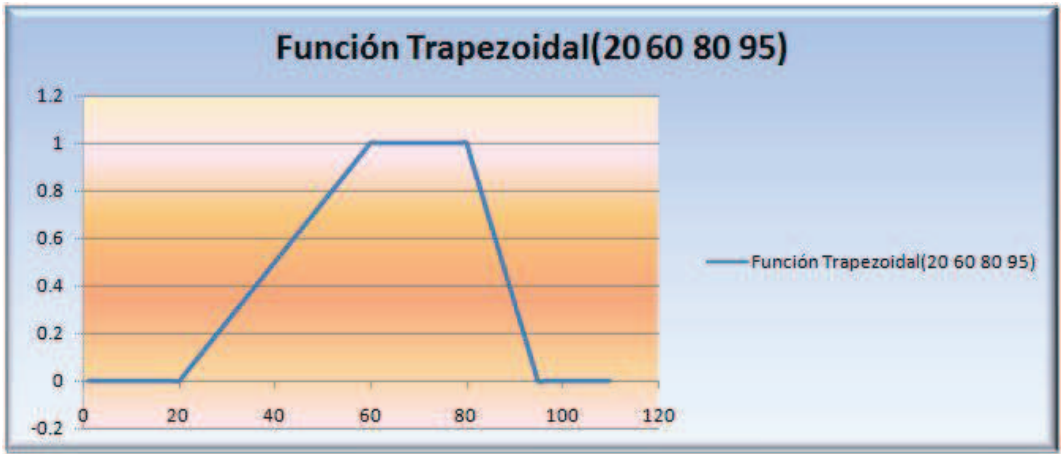


Figura 3.14: Función de membresía trapezoidal

En la figura 3.14 se puede ver gráficamente la función de membresía trap(10 20 60 95).

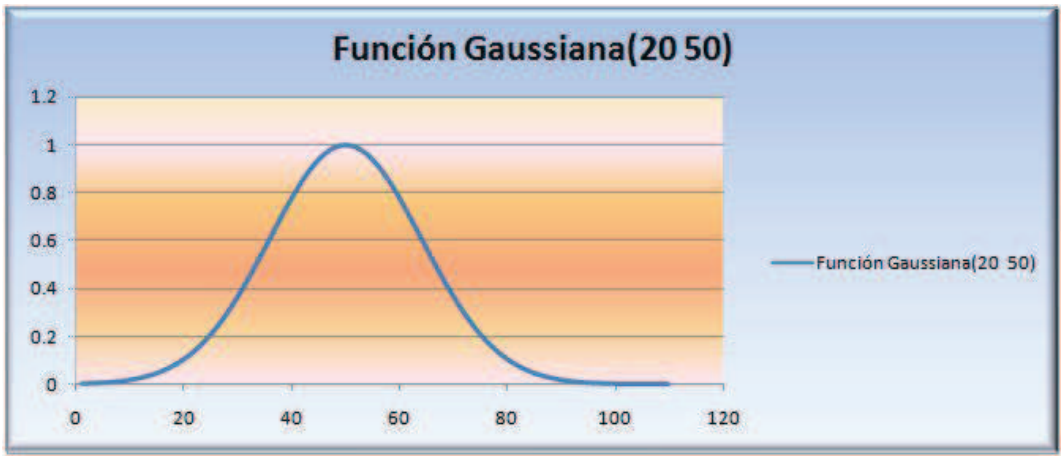


Figura 3.15: Función de membresía gaussiana

- Funciones gaussianas, con parámetros σ, c, donde:

$$gauss_{\sigma,c}(x) = e^{-(\frac{x-c}{\sigma})^2} \tag{3.12}$$

En la figura 3.15 se puede ver gráficamente la función de membresía gauss(20 50).

- Funciones campana, con parámetros a, b, c, donde:

$$bell_{a,b,c}(x) = \frac{1}{1 + |\frac{x-c}{a}|^{2b}} \tag{3.13}$$

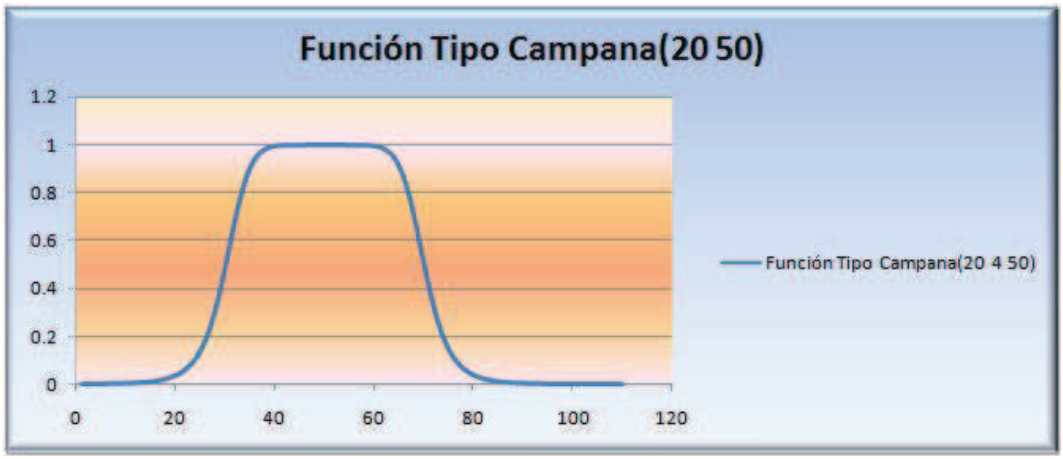


Figura 3.16: Función de membresía tipo campana

En la figura 3.16 se puede ver gráficamente la función de membresía bell(20 4 50).

Definición 3.3.2. *Un conjunto difuso A sobre X es normal si existe al menos un elemento* $x \in X$ *tal que* $A(x) = 1$*. De otra forma, A es subnormal.*

Definición 3.3.3. *La altura de un conjunto difuso A sobre X se define como*

$$altura(A) = \max_{x \in X}[A(x)]$$

Definición 3.3.4. *Sea A un conjunto difuso sobre X. El soporte de A, denotado Sop(A) es el subconjunto clásico de X cuyos elementos tienen membresía no nula en A.*

$$Sop(A) = \{x \in X : A(x) > 0\}$$

Definición 3.3.5. *Sea A un conjunto difuso sobre X. El núcleo de A, denotado Nu(A), es el subconjunto clásico de X cuyos elementos tienen membresía unitaria.*

$$N_u(A) = \{x \in X : A(x) = 1\}$$

Definición 3.3.6. *Sean A y B conjuntos difusos sobre X. Se dice que A es un subconjunto de B,*

$$A \subset B \Leftrightarrow A(x) \leq B(x)\ \forall x \in X$$

Definición 3.3.7. *Sean A y B subconjuntos difusos sobre X. A es igual a B,*

$$A = B \Leftrightarrow A \subset B \wedge B \subset A$$

Definición 3.3.8. *El conjunto difuso nulo sobre X, se denota* $\emptyset_x$ *o simplemente* $\emptyset$*. Su función de membresía es ;*$X(x) = \emptyset_X(x) = 0, \forall x \in X$*. Por otro lado, el conjunto difuso universal sobre X se denota* 1_X*, o bien X, y queda caracterizado por la función de membresía* $1_X(x) = 1, \forall x \in X$*.*

Conglomerado Difusos

Los conglomerados Difusos se caracterizan porque al estimar los grupos, éstos tienen bondades ”difusas”, en el sentido de que **cada valor de datos pertenece a cada grupo en un cierto grado o adhesión** . La pertenencia no es ”única, rígida o dura”. Después de haber decidido el número de agrupaciones a ser usadas, es necesario utilizar un procedimiento que permita localizar sus puntos medios (o de manera más general, sus centroides), determinar las funciones de pertenencia y el grados de pertenencia de los datos a cada grupo.

El algoritmo FCM es realmente una generalización del algoritmo c-medias ”duros”. Aparece desde Ruspini (1970), aunque algunos de los conceptos fueron explorados por MacQueen (1967). El algoritmo FCM está estrechamente relacionada con contribuciones de Bezdek (1973) y Dunn (1974, 1977), y se utiliza ampliamente en reconocimiento de patrones.

Este método se basa en el concepto de particiones difusas. Se define la matriz de datos X de dimensión mxn, y lo que se desea es dividir los ”n” datos en ”c” grupos difusos $G_k \ (k = 1, \ldots, c)$ donde $c < n$, y al mismo tiempo determinar la ubicación de estas agrupaciones en el espacio correspondiente.

Los datos pueden ser multidimensionales, y las métricas que constituyen la base para el FCM utilizan el ”Error al cuadrado de las distancia”. La base matemática para este procedimiento es el siguiente. Sea x_i el i-ésimo vector de puntos $(i = 1, 2, \ldots, n)$. Sea v_k el centro del k-ésimo grupo (difuso) $(k = 1, 2, \ldots, c)$. Sea $d_{ik} = ||x_i - v_k||$ la distancia entre x_i y v_k, y μ_{ik} el ”grado de adhesión o pertenencia” del vector ”i” en el grupo ”k”, donde:

$$\sum_{k=1}^{c}(\mu_{ik}) = 1$$

El objetivo es la partición de los datos en ”c” agrupaciones, y simultáneamente localizar las agrupaciones y determinar el correspondiente ”grados de pertenencia”, a fin de minimizar el funcional

$$J(U, V) = \sum_{k=1}^{c}\sum_{i=1}^{n}(\mu_{ik})^m(d_{ik})^2 \tag{3.14}$$

No hay forma prescrita para la elección del parámetro exponente ”m”, éste debe cumplir con la siguiente condición $1 < m < \infty$; y es muy importante que si éste toma valores cercanos a 1 la partición de los datos resultantes será rígida, mientras que valores más lejanos a 1 harán que la partición sea más difusa. En la práctica, m = 2 es una común elección [5].

En términos generales, el algoritmo FCM implica los siguientes pasos:

1. Seleccionar la ubicación inicial para la agrupación de centros.
2. Generar una (nueva) partición de los datos mediante la asignación de cada punto de datos a un grupo cuyo centro sea más cercano.
3. Calcular nuevo grupo de centros como los centroides de los grupos.
4. Si el grupo de partición es estable entonces se detiene. De lo contrario va al paso 2.

En el caso de membresías difusas, el multiplicador de Lagrange genera la siguientes expresión que se utilizarán en el paso 2:

$$\mu_{ik} = 1/\{\sum_{j=1}^{c}[(d_{ik}^2/d_{ij}^2)]^{1/(m-1)}\}$$

Demostración.- De una manera más general se desea resolver el problema:

$$\min J = \sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ik}^m \parallel x_i - v_k \parallel^2 \tag{3.15}$$

con respecto a $v_k \in V$ y $\mu_{ik} \in U$ [8] . La solución a este problema se la realiza en dos pasos, primero hay que tomar en cuenta que la norma está asociada a una distancia, luego utilizando los multiplicadores de Lagrange se calcula el gradiente de J con respecto a μ_{ik} y v_k.

Entonces, tomando d_{ik} como la distancia asociada a $\parallel x_i - v_k \parallel$ en la ecuación 3.15

se tiene:

$$J(U,V) = \sum_{k=1}^{c}\sum_{i=1}^{n}(\mu_{ik})^m(d_{ik})^2 \tag{3.16}$$

Sujeto a:

$$\sum_{k=1}^{c}(\mu_{ik}) = 1$$

Utilizando los multiplicadores de Lagrange se obtiene la siguiente expresión:

$$L = \sum_{k=1}^{c}\sum_{i=1}^{n}(\mu_{ik})^m(d_{ik})^2 - \lambda(\sum_{k=1}^{c}(\mu_{ik}) - 1) \tag{3.17}$$

Ahora, de la ecuación 3.17 se calcula la derivada de L con respecto a μ_{ik} y se obtiene:

$$\frac{\delta L}{\delta \mu_{ik}} = m\mu_{ik}^{m-1}d^2 - \lambda = 0$$

Entonces:

$$\mu_{ik} = (\frac{\lambda}{m})^{1/m-1} \frac{1}{(d_{ik})^{\frac{2}{m-1}}} \tag{3.18}$$

Ahora, utilizando $\sum \mu_{ik} = 1$ se consigue:

$$(\frac{\lambda}{m})^{1/m-1} \sum_{k=1}^{c} \frac{1}{(d_{ik})^{\frac{2}{m-1}}} = 1$$

Entonces despejando se llega a:

$$(\frac{\lambda}{m})^{1/m-1} = \frac{1}{\sum_{k=1}^{c} \frac{1}{(d_{ik})^{\frac{2}{m-1}}}} \tag{3.19}$$

reemplazando la ecuación 3.19 en la ecuación 3.18 se concluye:

$$\mu_{ik} = 1/\{\sum_{j=1}^{c} [(d_{ik}^2/d_{ij}^2)]^{1/(m-1)}\} \tag{3.20}$$

Nótese que una singularidad se producirá si $d_{ij} = "0"$ en la expresión anterior. Esto ocurre si, algún punto de los centros de los grupos coincida exactamente con un punto de los datos. Esto se puede evitar al inicio del algoritmo y, en general, no se produciría en la práctica debido a la precisión de la máquina. Si el conjunto de datos para las agrupaciones fuera del tipo clásico entonces

$$U_{ik} = 0; \quad \forall i \neq j$$

$$U_{ik} = 1; \quad \forall i = j$$

La actualización de los centros del conglomerado en el paso 3 anterior se obtiene a través de la expresión

$$v_k = [\sum_{i=1}^{n} (u_{ik})^m x_i] / [\sum_{i=1}^{n} (u_{ik})^m]; k = 1, 2, \ldots, c \tag{3.21}$$

Demostración De la ecuación 3.15 se obtiene el gradiente con respecto a v_k como sigue:

$$J = \sum_{k=1}^{c} \sum_{i=1}^{n} \mu_{ik}^m \parallel X_i - v_k \parallel^2$$

Sabiendo que $\parallel x_i - v_k \parallel^2$ es igual a $(x_i - v_k)^T (x_i - v_k)$ y éste es igual a $x_i^T x_i - 2x_i^T v_k + v_k v_k^T$ entonces, se tiene que:

$$\nabla_v J = 2 \sum_{i=1}^{n} \mu_{ik}^m (x_i - v_k) = 0$$

despejando v_k se obtiene la ecuación 3.21

La función de distancia (d_{ik}) utilizada en este caso es la distancia euclidiana. Debido a esto, el FCM solo detecta conglomerados con la misma forma (básicamente esféricos).

Ejemplo 3.3.3. *Suponga que se tiene la tabla 3.10 con datos de dos variables x_1 y x_2:*

Obs	x_1	x_2
1	0.11	0.44
2	0.47	0.81
3	0.24	0.83
4	0.09	0.18
5	0.09	0.63
6	0.58	0.33
7	0.9	0.11
8	0.68	0.17
9	0.82	0.11
10	0.65	0.5
11	0.98	0.24

Tabla 3.10: Datos

A continuación se presenta su gráfico para analizar de manera visual los grupos existentes.

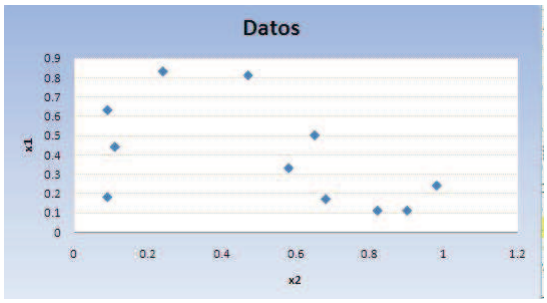


Figura 3.17: Datos para Análisis de Conglomerados

Como primer paso se toma k=2, esto quiere decir que se va agrupar los datos en dos conglomerados. Luego tal como se mencionó anteriormente el parametro m tiene el valor de 2. Además por cada conglomerado se asignan los centros iniciales para cada variable:

Conglomerado	x_1	x_2
1	0.2	0.5
2	0.8	0.5

Tabla 3.11: Centros de Conglomerados

Entonces, de acuerdo a la ecuación 3.20 se determina primero d_{ij} que representa la distancia entre la observación i y cada uno de los centros j. Luego para obtener μ_{ik} se toma como referencia el indice k, por ejemplo si $k = 1$ hacemos referencia al grado de pertenencia de la observación i con respecto al conglomerado 1, entonces la distancia d_ik es una de las d_{ij} de acuerdo al valor de k. Para la primera observación se tiene:
$d_{11} = (0.11 - 0.2)^2 + (0.44 - 0.5)^2 = 0.0117$
$d_{12} = (0.11 - 0.8)^2 + (0.44 - 0.5)^2 = 0.4797$
Para estimar μ_{11} se tiene:

$$\mu_{11} = \frac{1}{(\frac{d_{11}}{d_{11}+d_{12}})} = 0.976$$

$$\mu_{12} = \frac{1}{(\frac{d_{12}}{d_{11}+d_{12}})} = 0.023$$

En la tabla 3.12 se observa el grado de pertenencia μ_i que tiene cada observación con respecto a cada uno de los dos conglomerados. Con estos valores se obtiene el gráfico de membresía 3.18 donde se reflejan la función de pertenencia para cada conglomerado.

obs.	μ_1	μ_2
1	0.976190476	0.023809524
2	0.548128342	0.451871658
3	0.792682927	0.207317073
4	0.841192788	0.158807212
5	0.947272727	0.052727273
6	0.308459697	0.691540303
7	0.201566774	0.798433226
8	0.266536965	0.733463035
9	0.221335269	0.778664731
10	0.1	0.9
11	0.128865979	0.871134021

Tabla 3.12: Grados de pertenencia de las observaciones a los conglomerados

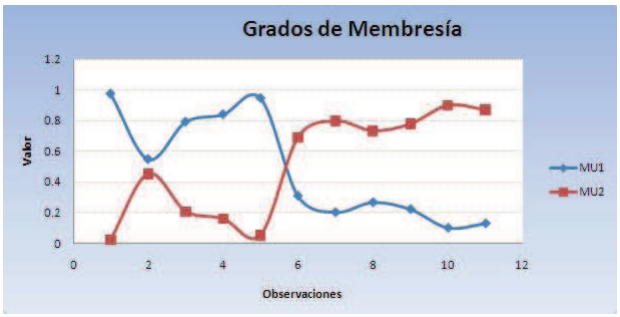


Figura 3.18: Grados de Membresía

La figura 3.18 nos permite analizar que las observaciones que se encuentran en el valor 0.5 son aquellas que se encuentran al borde o límite entre los conglomerados y el definir a cual conglomerado debe pertenecer es difuso.

Luego de haber estimado los grados de membresia, se procede a actualizar los centros de los conglomerados de acuerdo a la ecuación 3.21, como sigue:

Para calcular V_{11}, se debe sumarizar el producto $X_1 i * \mu_{1i}^2$ así como tambien se sumarizan los μ_{1i}^2

$$\sum_{i=1}^{11} X_1 i * \mu_{1i}^2 = 0.74$$

$$\sum_{i=1}^{11} \mu_{1i}^2 = 3.76$$

Entonces se tiene $v_1 = \frac{\sum_{i=1}^{11} X_1 i * \mu_{1i}^2}{\sum_{i=1}^{11} \mu_{1i}^2} = 0.19$.
En la tabla 3.13 se presentan los nuevos centros calculado.

Conglomerado	x_1	x_2
1	0.19	0.51
2	0.75	0.28

Tabla 3.13: Nuevos centros de Conglomerados Difusos

La figura 3.19 presenta los datos y los centros calculados.

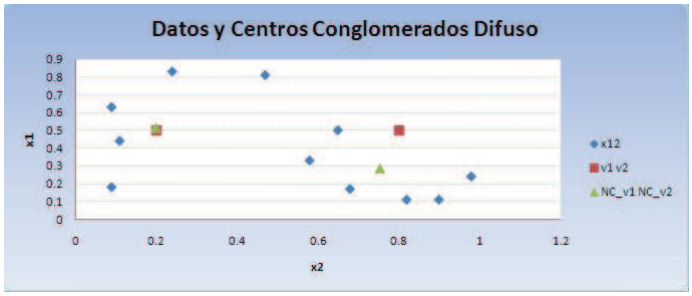


Figura 3.19: Gráfico de Conglomerados C-medias

Con el software R el método de c medias difuso se lo puede obtener mediante la función cmeans que se encuetra dentro del paquete e1071 [5] el cual tiene las opciones siguientes:

cmeans (x, centers, iter.max=100, verbose=FALSE, dist="euclidean", method="cmeans", m=2, rate.par = NULL)

donde:

x matríz de datos.
$Centers$: puede ser un vector con los centros iniciales.
$Iter.max$: es el máximo número de iteraciones a realizar.
$Verbose$: si toma el valor TRUE, presenta algunas salidas durante el aprendizaje.
$method$: determina el algoritmo ha utilizar.
m : es el grado de fusificación.
En la figura 3.20 se presenta la sentencia utilizada para obtener los datos del ejercicio anterior.

[5] http://www.stat.ucl.ac.be/ISdidactique/Rhelp/library/e1071/html/cmeans.html

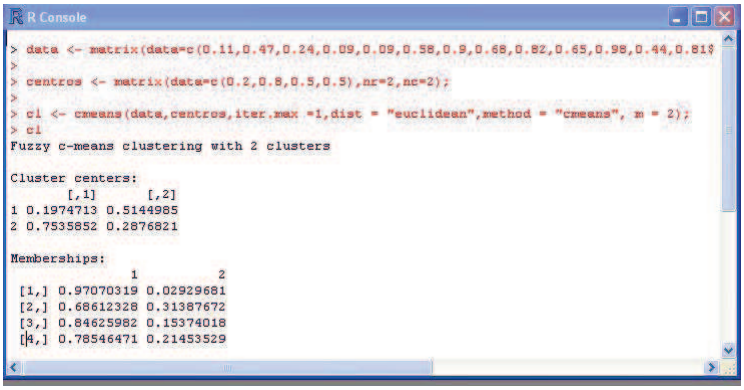


Figura 3.20: Conglomerados c-medias difuso en R

3.3.2 Métodos Jerárquicos

Los métodos jerárquicos parten de una matriz de distancias o similaridades calculada a partir de los elementos de la muestra. La idea base es crear una estructura jerárquica de los datos basada en estas distancias, bien sea aglomerándolos o separándolos consecutivamente entre sí. La información resultante se muestra en una especie de árbol, llamado dendograma, que puede ser muy útil y fácil de manejar si el conjunto de datos analizados no es muy grande (menos de 100-140 datos) pero si el conjunto de datos es grande (como es el caso de muchos procesos) se hace difícil descubrir la estructura agrupada de los datos. Los métodos jerárquicos se dividen a su vez en dos tipos de métodos [28]:

- **Aglomerativos**.- Parten de los elementos individuales y los van agregando en grupos.
- **Divisivos**.- Parten del conjunto de elementos y lo van dividiendo sucesivamente hasta llegar a los elementos individuales.

Métodos Jerárquicos aglomerativos

Los algoritmos aglomerativos tienen siempre la misma estructura y sólo se diferencian en la forma de calcular las distancia entre grupos. Dicha estructura se describe como sigue: [28]

1. Comenzar con tantas clases como elementos, n. Las distancias entre clases son las distancias entre los elementos originales.
2. Seleccionar los dos elementos más próximos en la matriz de distancia y formar con ellos una clase.

3. Sustituir los dos elementos utilizados en 2, por un nuevo elemento que represente la clase construida. Las distancias entre este nuevo elemento y los anteriores, se calculan con uno de los criterios que se explicará en los párrafos siguientes.

4. Volver a 2; y repetir 2 y 3, hasta que se tenga todos los elementos agrupados en una única clase.

La distancia utilizada para juntar los grupos, define el nombre del método aglomerativo, los principales métodos utilizados se muestran en la figura 3.21.

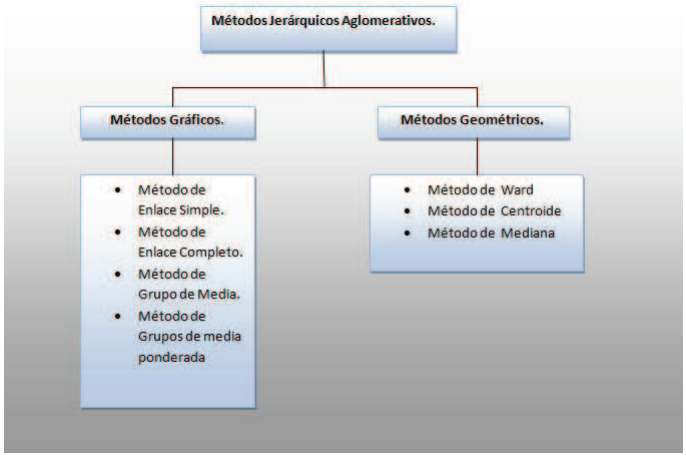


Figura 3.21: Métodos Jerárquicos aglomerativos

En 1983 Murtagh clasifica a los métodos jerárquicos como métodos gráficos (porque puede ser representado por un subconjunto de puntos o conexiones) y métodos geométricos (porque el grupo puede ser representado por un punto centro o promedio).

La idea general de los métodos, es que se tienen un grupo A con n_a elementos, y un grupo B con n_b elementos, y que ambos se fusionan para crear el grupo (AB) con $n_a + n_b$ elementos, mediante una distancia. A continuación se describen los métodos aglomerativos más utlizados.

El método de Enlace Simple

.

En el método de enlace Simple, la distancia entre dos observaciones A y B es definida como:

$$D(A, B) = min\{d(y_i, y_j)\ \forall y_i \in A\ y\ y_j \in B\} \tag{3.22}$$

Donde $d(y_i, y_j)$ representa la distancia Euclediana. Este método también es llamado el del vecino más cercano. En cada paso, el método calcula la distancia entre cada par de conglomerados, y los dos conglomerados con menor distancia son unidos. El valor asignado para la matríz de distancias del grupo nuevo, se la determina tomando el valor mínimo (distancia) entre los grupos unidos en la iteración anterior.

El método de Enlace Completo

Es también llamado el método del vecino más alejado, la distancia entre dos conglomerados A y B es definido como :

$$D(A, B) = max\{d(y_i, y_j) \ \forall y_i \in A \ y \ y_j \in B\} \tag{3.23}$$

En cada paso, el método calcula la distancia entre cada par de conglomerados, y los dos conglomerados con mayor distancia son unidos. El algoritmo se repite nuevamente y toma como distancia para el grupo nuevo la máxima entre los grupos unidos en la iteración anterior.

El método de Enlace Promedio

En este método la distancia entre dos conglomerados A y B es definida como el promedio de las distancias $n_A n_B$ entre los puntos n_A en A y n_B en B:

$$D(A, B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(y_i, y_j) \tag{3.24}$$

Donde la suma es sobre todos los y_i en A y todos los y_j en B. En cada paso, el método junta los conglomerados con menor distancia.

El método Centroide

En el método de centroide, la distancia entre A y B es definido como la distancia euclidiana entre los vectores de medias(centroides) de dos conglomerados.

$$D(A, B) = d(\overline{y_A}, \overline{y_B}) \tag{3.25}$$

Donde $\overline{y_A}$ y $\overline{y_B}$ son el vector de media para las observaciones en A y las observaciones en B, respectivamente, y $d(\overline{y_A}, \overline{y_B})$ es definida como:

$$d(\overline{y_A}, \overline{y_B}) = \sqrt{(\overline{y_A} - \overline{y_B})^T (\overline{y_A} - \overline{y_B})}$$

Y $\overline{y_A} = \sum_{i=1}^{n_A} \frac{y_i}{n_A}$. Los dos conglomerados con la distancia más pequeña entre los centroides son unidos en cada paso.

Después de que dos conglomerados A y B son unidos, el nuevo centroide del conglomerado AB está dado por la media ponderada siguiente:

$$Y_{AB} = \frac{n_A\overline{Y_A} + n_B\overline{Y_B}}{n_A + n_B}$$

El método de Mediana

Si dos conglomerados A y B son combinados por el método del centroide, y si A contiene un número mayor de observaciones que B , entonces los nuevos centroides $Y_{AB} = \frac{n_A\overline{Y_A}+n_B\overline{Y_B}}{n_A+n_B}$ pueden ser mucho más cercano a $\overline{Y_A}$ que ha $\overline{Y_B}$. Para evitar el peso en el vector de medias debido al tamaño del conglomerado, se utiliza la mediana de los dos puntos medios de A y B como los puntos para calcular la nueva distancia para los otros conglomerados:

$$m_{AB} = \frac{1}{2}(\overline{Y_A} + \overline{Y_B})$$

los dos conglomerados con más pequeñas distancia entre medianas son unidos cada paso. La terminología de media proviene de la mediana de un triángulo.

Método de Ward

El método de Ward o llamado también método suma de cuadrados incremental, define la proximidad entre dos conglomerados como el incremento en el error que resulta cuando dos grupos son unidos. Entonces si AB es el grupo obtenido por la combinación del grupo A y B, entonces la suma del error al cuadrado son:

$$\begin{aligned} SSE_A &= \sum_{i=1}^{n_A}(y_i - \overline{y}_A)^T(y_i - \overline{y}_A) \\ SSE_B &= \sum_{i=1}^{n_B}(y_i - \overline{y}_B)^T(y_i - \overline{y}_B) \\ SSE_{AB} &= \sum_{i=1}^{n_{AB}}(y_i - \overline{y}_{AB})^T(y_i - \overline{y}_{AB}) \end{aligned}$$

donde:

$$\overline{y}_{AB} = \frac{n_A y_A + n_B y_B}{n_A + n_B}; \quad n_{AB} = n_A + n_B$$

El método de Ward junta dos conglomerados A y B que minimiza el incremento en SSE definido como:

$$I_{AB} = SSE_{AB} - (SSE_A + SSE_B)$$
$$I_{AB} = \frac{n_A n_B}{n_A + n_B}(\overline{Y}_A - \overline{Y}_B)^T(\overline{Y}_A - \overline{Y}_B)$$

Ejemplo 3.3.4. *Suponga que se tiene la tabla 3.14 con datos de dos variables x_1 y x_2:*

Obs	x_1	x_2
1	1	1
2	1.5	2
3	3	4
4	5	7
5	3.5	5
6	4.5	5
7	3.5	4.5

Tabla 3.14: Datos

A continuación se presenta el gráfico 3.22 para analizar de manera visual los grupos existentes.

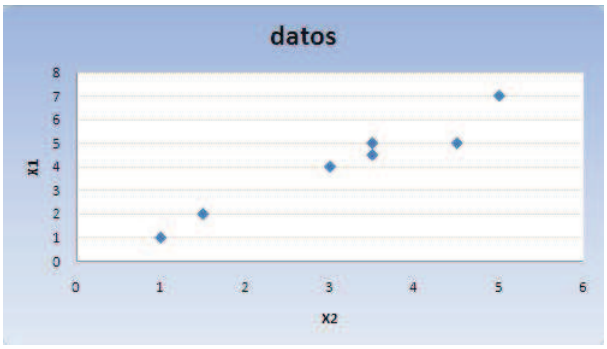


Figura 3.22: Datos para Análisis de Conglomerados

Como primer paso se calcula la matriz de distancia entre observaciones mediante la distancia Euclediana.

Obs	obs_1	obs_2	obs_3	obs_4	obs_5	obs_6	obs_7
obs_1	0						
obs_2	1.11	0					
obs_3	3.60	2.5	0				
obs_4	7.21	6.10	3.60	0			
obs_5	4.71	3.60	1.11	2.5	0		
obs_6	5.31	4.24	1.80	2.06	1	0	
obs_7	4.30	3.20	0.70	2.91	0.5	1.11	0

Tabla 3.15: Matríz de distancias

Entonces a esta matríz se le aplica el método de conglomerado jerárquico Enlace Simple con la ecuación 3.22, la cual dice que para agrupar a dos observaciones se debe elegir aquellas que tienen menor distancia entre si, en este caso es la observaciones 5 y.7 generando el Grupo 1.

Obs	obs_1	obs_2	obs_3	obs_4	obs_5	obs_6	obs_7
obs_1	0						
obs_2	1.11	0					
obs_3	3.60	2.5	0				
obs_4	7.21	6.10	3.60	0			
obs_5	4.71	3.60	1.11	2.5	0		
obs_6	5.31	4.24	1.80	2.06	1	0	
obs_7	4.30	3.20	0.70	2.91	**0.5**	1.11	0

Tabla 3.16: Formación del Primer Grupo

Luego para formar la matríz de distancias incluyendo el grupo 1 $(G_1 = (obs_5, obs_7))$ se debe comparar los dos valores y elegir el más pequeño de los dos. por ejemplo para determinar la distancia entre el grupo 1 y la obs_1, se deben comparar los valores 4.7 y 4.3 y elegir el valor menor, en este caso es 4.3, el cual es colocado en la intersección entre la observación 1 y el grupo G_1. Éste paso se lo realiza de manera similar con todas las demás observaciones. En la tabla 3.17 se aprecia la matríz obtenida.

Obs	obs_1	obs_2	obs_3	obs_4	obs_6	G_1
obs_1	0					
obs_2	1.11	0				
obs_3	3.60	2.5	0			
obs_4	7.21	6.10	3.60	0		
obs_6	5.31	4.24	1.80	2.06	0	
G_1	4.30	3.20	0.70	2.5	1	0

Tabla 3.17: Matríz de distancias nuevas

Luego, se realiza los mismos paso hasta llegar a formar un grupo único en este caso G_6. El resumen de los grupos y la distancia mínima, se presenta en la tabla 3.18.

Grupo	Observaciones	Mínima distancia
G_1	obs_5 obs_7	0.5
G_2	obs_3 G_1	0.7
G_3	G_2 obs_6	1
G_4	obs_1 obs_2	1.11
G_5	obs_4 G_3	2.06
G_6	G_4 G_5	2.5

Tabla 3.18: Tabla de agrupación

A continuación se presenta el gráfico de agrupación

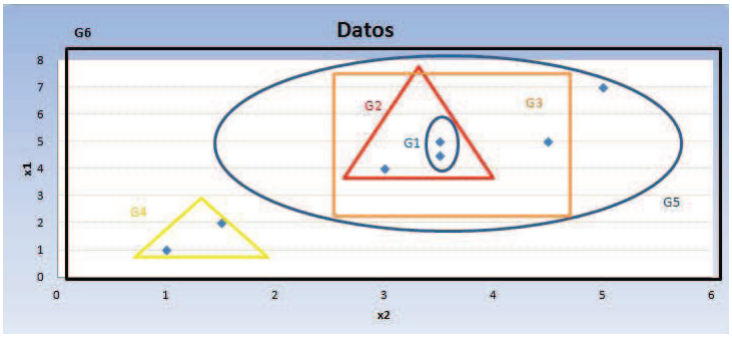


Figura 3.23: Modelo jerárquico

Con el software R los métodos jerárquicos se los puede obtener mediante la función hclust que se encuetra dentro del paquete stat [6] el cual tiene las opciones siguiente:
hclust(d, method = ”complete”) donde:
d es una matríz de distancias.
$method$ es el método de agrupamiento jerárquico. Teniendo las siguientes opciones ”ward”, ”single”, ”complete”, ”average”, ”mcquitty”, ”median” or ”centroid”.
En la figura 3.24 se presenta la sentencia a ejecuta en R para obtener los resultados del ejemplo:

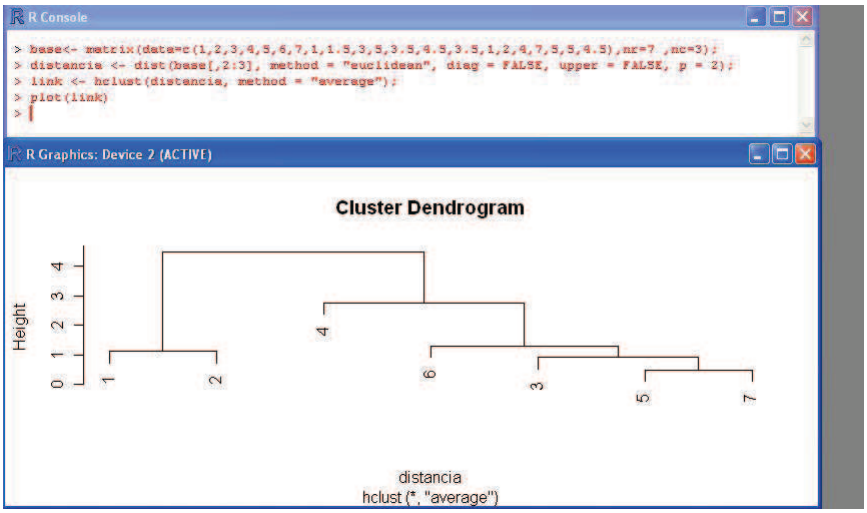


Figura 3.24: Modelo jerárquico R

[6] http://stat.ethz.ch/R-manual/R-devel/library/stats/html/hclust.html

3.3.3 Métodos basados en modelos

Los algoritmos de conglomerados basados en modelos, suponen que los datos se han generado por una mezcla de distribuciones de probabilidad desconocidas; y su objetivo es el de estimar los parámetros de dichas distribuciones.

Expectation-Maximization (EM)

En la práctica, cada grupo puede ser representado matemáticamente por una distribución de probabilidad paramétrica. El conjunto de datos en si es una mixtura o mezcla de esas distribuciones, donde cada distribución individual es típicamente referida como una componente de la distribución total. Por tanto se puede agrupar los datos usando un modelo de mezclas de k distribuciones de probabilidad, donde cada distribución representa un conglomerado. El problema es estimar los parámetros de la distribución de probabilidad que se ajuste de mejor manera a los datos.

El algoritmo EM es un método iterativo que puede ser usado para estimar los parámetros de la distribución de probabilidad, puede ser visto como una extensión del método de k-medias, su diferencia radica en que los objetos son asignados a un conglomerado de acuerdo a un peso, que representa la probabilidad de membrecía, es decir, no existe restricción de borde entre los conglomerados.

El algoritmo EM empieza con un estimación inicial de los parámetros del modelo de mezclas. Entonces iterativamente se clasifica los objetos y nuevamente la distribución de mezcla produce nuevos parámetros. A cada objeto se le asigna una probabilidad de que éste posea ciertas características de los atributos dado en cada conglomerado. El algoritmo se describe como sigue:

1. Hacer una estimación inicial del vector de parámetros: Esto envuelve una selección randómica de k objetos para representar la media o centro de los conglomerado, y la estimación de los parámetros adicionales.

2. Iterativamente se redefine los parámetros (o conglomerados) basados por los siguientes pasos:

   (a) Paso de expectación o esperanza: Asignamos a cada objeto x_i al conglomerado C_k con la probabilidad

   $$P(x_i \in C_k) = p(C_k/x_i) = \frac{p(C_k)p(x_i/C_k)}{p(x_i)}$$

   donde $p(x_i/C_k) \sim N(m_k, E_k(x_i))$ sigue una distribución normal con media m_k con esperanza, E_k, es decir, en este paso calcula la probabilidad

de membrecía para el objeto x_i para cada uno de los conglomerados. Esta probabilidad es el grado de membrecía esperada del objeto x_i a un conglomerado.

(b) paso de maximización: Usa las probabilidades estimadas anteriormente y re-estima el modelo de parámetros. Este paso es la maximización de la probabilidad de la distribución dada en los datos.

$$m_k = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i P(x_i \in C_k)}{\sum_j P(x_i \in C_j)}$$

En el software R el método EM se lo realiza mediante la función $Mclust$ como sigue:
Mclust(data,G=)[7]
donde:
$data$ es un vector o matríz de observaciones.
G es el número de mezclas o componentes de distribución.

Con los datos de la tabla 3.10 se tiene la siguiente salida:

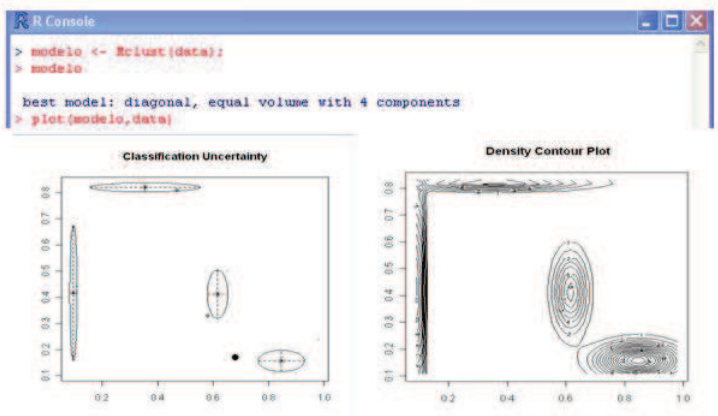


Figura 3.25: Modelo EM R

[7]http://www.stat.washington.edu/research/reports/2006/tr504.pdf

3.4 Validación de conglomerados

Debido a la variedad de algoritmos de conglomerados, y sus diferencias teorías, éstos pueden producir una partición diferente del conjunto de datos. Entonces para evaluar los resultados y la estructura de los conglomerados, en general, existen tres criterios fundamentales que cumplen con éste objetivo: [8]

1. Criterio Externo.
2. Criterio Interno.
3. Criterio relativo.

Para un conjunto de datos X y una estructura de conglomerados C derivada de la aplicación de un algoritmo de conglomerados sobre X, el criterio externo es utilizado para evaluar las coincidencias entre las asignación de los conglomerados, con las asignaciones de las categorías basadas sobre información a priori. En contraste al criterio externo, el criterio interno evalúa la estructura del conglomerado obtenido desde X sin alguna información externa. El criterio relativo compara C con otras estructuras de conglomerados, dichas estructuras son obtenidas de la ejecución de otros algoritmos de conglomerados o el mismo algoritmo pero con diferentes parámetros sobre X y entonces se determina cuál técnica representa mejor la estructura de X.

Los criterios tanto internos como externos están relacionados a métodos estadísticos y pruebas de hipótesis.[9]

En la validación de la estructura de un conglomerado, se realiza una prueba de hipótesis, con la finalidad de evaluar si existe una estructura sobre los datos X o si la estructura de X es aleatoria. Entonces existen tres hipótesis de nulidad definidas [10]

1. Hipótesis de posición aleatoria:
   H_0 :Todas los N puntos localizados en alguna región especifica de d-dimensiones son iguales.
2. Hipótesis de gráfico aleatorio :
   H_0 : Todas las matrices de proximidad de orden NxN son iguales.
3. Hipótesis de clasificación aleatoria
   H_0 : Todas las permutaciones de la clasificación de los N objetos son iguales.

[8] refiérase Jain and Dubes, 1988 ; Theodoridis and Koutroumbas, 2006

[9] refiérase Jain and Dubes, 1988

[10] refiérase Jain and Dubes, 1988

3.4.1 Criterio Externo

Si P es una partición perspectiva de un conjunto de datos X con N observaciones y es independiente de la estructura de la clasificación C (resultado de la aplicación de un algoritmo de conglomerado), entonces la evaluación de C por un criterio externo es realizado mediante la comparación entre C y P.

Para un par de puntos x_i y x_j tomados del conjunto de datos X, hay cuatro posibles casos que se pueden dar de acuerdo a como los dos puntos son clasificados en C y en P.

- Caso 1: x_i y x_j pertenecen al mismo conglomerado de C y a la misma categoría de P .
- Caso 2: x_i y x_j pertenecen al mismo conglomerado de C pero diferente categorías de P.
- Caso 3: x_i y x_j pertenecen a diferentes conglomerados de C pero a la misma categoría de P.
- Caso 4: x_i y x_j pertenecen a diferentes conglomerados de C y a diferentes categorías de P.

Correspondientemente, el número de puntos para las cuatro categorías son denotados como a, b, c , y d , y el número de total de puntos es $N(N-1)/2$, denotado como M , entonces se tiene que $a + b + c + d = M$.

De acuerdo a la notación, entonces se tienen los siguientes indices para evaluar las coincidencias de C y P :

Índice de Rand (Rand, 1971)

$R = (a + d)/M$;

Coeficiente de Jaccard

$J = a/(a + b + c)$;

Índice Fowlkes and Mallows (Fowlkes and Mallows, 1983)

$FM = \sqrt{\frac{a}{a+b} * \frac{a}{a+c}}$;

Estadístico Γ

$\Gamma = \frac{Ma - m_1 m_2}{\sqrt{m_1 m_2 (M - m_1)(M - m2)}}$

donde: $m_1 = a + b$ y $m_2 = a + c$.

Como se puede ver desde la definición, cuanto mayor sea el valor de los tres índices, más similares son C y P . Específicamente , los valores del estadístico Γ se encuentra entre -1 y 1, mientras que los valores del índice de Rand y el índice Jaccard están en el rango de [0, 1]. La mayor diferencia entre estos dos últimos estadísticos, es que el índice de Rand enfatiza en la situación de que los grupos pertenecen a un mismo grupo o a diferentes grupos en ambos C y P.

3.4.2 Criterio Interno

Un coeficiente de correlación cofenética (CPCC) (Jain and Dubes,1988 ; Rohlf and Fisher, 1968), es un índice usado para validar la estructura jerárquica de un conglomerado. Dada la matriz de proximidad $P = \{p_{ij}\}$ de X, entonces el CPCC mide el grado de similaridad entre P y la matriz cofénética $Q = \{qij\}$, cuyos elementos registran el nivel de proximidad de cada par de puntos x_i y x_j, cuando éstos fueron agrupados en un mismo conglomerado por primera vez. Sea μ_p y μ_q las medias de P y Q, es decir,

$$\mu_p = \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_{ij}$$

.

$$\mu_q = \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} q_{ij}$$

Donde $M = N(N-1)/2$, CPCC es definida como:

$$CPCC = \frac{\frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_{ij} q_{ij} - \mu_p \mu_Q}{\sqrt{(\frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_{ij}^2 - \mu_p^2)(\frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} q_{ij}^2 - \mu_Q^2)}}$$

El valor de $CPCC$ se encuentra en el rango de $[-1, 1]$, y un valor del índice cercano a 1 indica una significativa similaridad entre P y Q y un buen ajuste del método jerárquico en los datos.

Ejemplo 3.4.1. *Como se vió en el ejercicio del modelo de conglomerados jerárquico, se obtuvo la matriz de distancia 3.15, luego se agrupó las observaciones mediante el método*

de enlace simple y se obtuvo las distancias mínimas de agrupación, éstas son consideradas para formar la matriz cofenética Q, como sigue:

Obs	obs_1	obs_2	obs_3	obs_4	obs_5	obs_6	obs_7
obs_1	0						
obs_2	1.11	0					
obs_3	2.5	2.5	0				
obs_4	2.5	2.5	2.06	0			
obs_5	2.5	2.5	0.7	2.06	0		
obs_6	2.5	2.5	1	2.06	1	0	
obs_7	2.5	2.5	0.7	2.06	0.5	1	0

Tabla 3.19: Matríz Cofenética

Luego con la matriz de distancia y la matriz cofenética se calcula el índice de correlación, se puede formar dos columnas, con los valores por debajo de la matriz de distancia y la matriz cofenética respectivamente.

Valores Matriz Cofenética Q	valores Matriz Distancia P
1.11	1.11
2.5	3.60
2.5	7.21
2.5	4.71
2.5	5.31
2.5	4.30
2.5	2.5
2.5	6.1
2.5	3.6
2.5	4.24
2.5	3.20
2.0	3.60
0.7	1.11
1	1.80
0.7	0.70
2.0	2.5
2.0	2.06
2.0	2.91
1	1
0.5	0.5
1	1.11

Tabla 3.20: Valores entre la Matriz de distancia y la Matriz Cofenética

Con la matriz se calcula el coeficiente de correlación de Person, obteniendo el valor de 0.82, esto significa que existe un buen ajuste del método jerárquico con los datos.

En R se utiliza la función $cophenetic(A)$ para calcular la matriz cofenética, donde A es el resultado de aplicar el método jerárquico, luego simplemente se aplica la función $cor(cophenetic(A), P)$ para obtener el índice de correlación, donde P es la matriz de distancia.

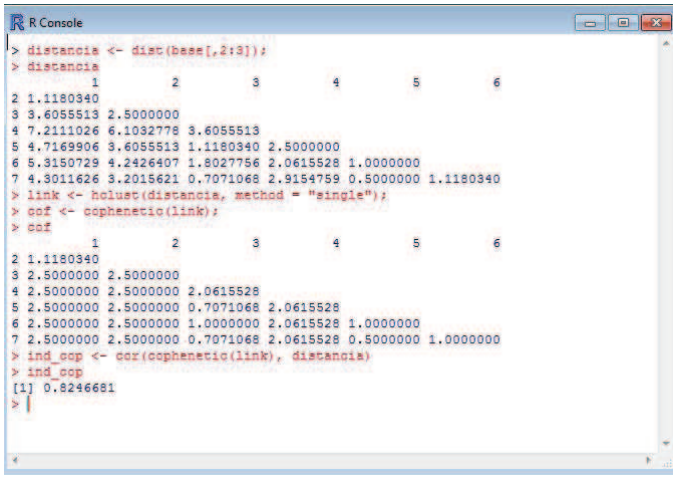


Figura 3.26: CPCC en R

3.4.3 Criterio Relativo

El criterio externo y el interno requieren pruebas estadísticas, las cuales pueden ser computacionalmente caras. Los criterios relativos eliminan tales requerimientos y se concentran sobre la comparación de los resultados generados por los diferentes algoritmos de conglomerados o sobre el mismo pero con diferentes parámetros. Como se sabe uno de los problemas importantes dentro de las técnicas de conglomerados es la determinación del número de conglomerados k. Para algoritmos de conglomerados jerárquicos un punto de corte debe ser determinado para cortar el dendograma en un cierto nivel, en orden a formar un conjunto de conglomerados. Para algoritmos de conglomerados particionales el valor de K es un parámetro definido por el usuario. Pero en general el número de conglomerados debe ser determinado desde los propios datos, puesto que una sobreestimación o una baja estimación del conglomerados, puede afectar la calidad de los resultados del conglomerados. Una partición con muchos conglomerados complica la verdadera estructura de los conglomerados, haciéndole difícil su interpretación y el análisis de resultado. En cambio, una partición con poco conglomerados puede causar perdida de información y equivocar una decisión final.

Visualización de los datos

El método más simple de determinar el número de conglomerados, es el de proyectar el conjunto de datos en un gráfico de dos o tres dimensiones en un espacio euclidiano, pues dichos gráficos pueden dar una alternativa apriori del número de los datos, sin embargo dada la complejidad de los datos, esta técnica puede resultar insuficiente.

Índices de validación y reglas de parada

Para un algoritmo de conglomerados que requiere como entrada el número de grupos k, una secuencia de estructuras de conglomerados puede ser obtenida por la corrida del algoritmo varias veces, desde el mínimo valor de k hasta el máximo valor de k. Estas estructuras son evaluadas en base a índices construídos, y la solución es determinada por la elección de uno, el cual tenga el mejor índice. En el caso de los métodos de estructura jerárquica, los índices son conocidos como reglas de parada, las cuales dicen donde es el mejor nivel de corte en el dendograma.

Como estándar para evaluar estos métodos, estos índices combinan la información de compacidad o aglutinamiento intra- conglomerado y la diferencia inter-conglomerados, así como también analizan funciones que estiman el error al cuadrado, la geométrica o estadística forma de los datos, las medidas de similaridad o distancia. Milligan y Cooper (1985) comparan y ordenan 30 indices de acuerdo a su efectividad sobre series simuladas de datos. Entre estos índices, uno de los de mejor efectividad es el índice de Calinski and Harabasz 1974, el cual se define como:

$$CH(k) = \frac{Tr(S_B)}{k-1} / \frac{Tr(S_W)}{N-k}$$

donde:

N es el número de objetos,

$$S_T = \sum_{j=1}^{N} (x_j - \overline{X})(x_j - \mu)^T$$

,

$$S_W = \sum_{i=1}^{K} \sum_{j=1}^{N} \gamma_{ij} (x_j - \overline{X}_i)(x_j - \overline{X}_i)^T$$

, y γ_{ij} es un indicador que toma valor de 1 si $x_j \in$ cluster i y 0 en otros casos. El valor de k que maximice $CH(k)$ sugiere la estimación de K.

El índice creado por Milligan y Cooper (1985), busca maximizar la distancia entre conglomerados y minimizar la distancia entre los centros del conglomerado y los otros puntos. Entonces se define un índice R_i por cada conglomerado como la máxima comparación entre el conglomerado i y los otros conglomerados.

$$R_i = \max(\frac{e_i + e_j}{D_{ij}})$$

Donde D_{ij} es la distancia entre los centros del conglomerado i y j, y e_i y e_j son el error promedio para el conglomerado i y j, respectivamente, el índice de Davies-Bouldin entonces se lo calcula como:

$$DB(K) = \frac{1}{K}\sum_{i=1}^{k} R_i$$

El mínimo $DB(k)$ indica el potencial número de conglomerados.

Índices de validación Difusos

En los conglomerados difusos, un punto x_j es asociado con cada uno de los conglomerados C_i de acuerdo a un grado de membresía U_{ij}, el cual define una matriz de partición $U_{K\times N}$. el algoritmo más utilizado para la realización de la partición difusa es el $c - medias$ de acuerdo a éste se analizan los índices de validación.

Tanto el coeficiente de partición (PC) (Bezdek, 1974) como el índice de partición Entropía (PE) (Bezdek, 1975) usan la información de la matriz de partición, sin considerar la data en si mismo. Especificamente, PC es definida como:

$$PC = \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{k} U_{ij}^2$$

y PE es determinado como:

$$PE = \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{k} U_{ij}\log_a U_{ij} \;\; a \in (1,\infty)$$

Los valores de PC se encuentra en el rango $[1/k, 1]$, mientras que los valores de PE están en el rango $[0, \log_a(k)]$. Si PC toma el máximo o PE toma el mínimo, entonces se obtiene una partición rígida. Si el caso contrario se cumple, PC toma el mínimo y PE toma el valor máximo entonces significa que el algoritmo no encuentra una estructura de conglomerados en los datos. Dado que los valores de PC y PE dependen del valor de k, k puede ser estimado mediante el análisis de la variación del valor tanto para PC (incremento) y PE(decremento).

3.4.4 Criterio para conglomerados basados en modelos

Para conglomerados que se basan en modelos de mezclas de probabilidad, el número de grupos k, se lo estima mediante el ajuste de un modelo con datos reales, y se lo optimiza bajo algún criterio dado. Usualmente, el algoritmo EM es

usado para determinar los parámetros del modelo para un k dado, los valores de k que maximicen o minimicen el criterio definido, es considerado óptimo.

Asumiendo que N es el número de registros de los datos, N_k es el número de parámetros para cada conglomerados, N_p es el número total de parámetros independientes a ser estimados por el modelo, y $l(\widehat{\theta})$ es el máximo logaritmo de la probabilidad, algunos criterios son definidos como sigue:

Criterio de información Akaike (AIC)

$$AIC(k) = \frac{-2(N-1-N_k-k/2)l(\widehat{\theta})}{N} + 3N_p$$

k es seleccionado con el mínimo valor de AIC(k).

Criterio de información Bayesiano (BIC)

$$BIC(k) = l(\widehat{\theta}) - (N_p/2)log(N)$$

k es seleccionado con el máximo valor de AIC(k).

Índice de Silueta

Kaufman y Rousseeuw [1990] presentan el índice de silueta como una forma de estimar el número de grupos en el conjunto de datos. Dado una observación i, se define la media de disimilitud para todos los otros puntos en el mismo conglomerado como a_i [33]. Para otro conglomerado C, $d(i,c)$ representa la media de i para todos los objetos en el conglomerado C. Finalmente, se define b_i como el mínimo de las disimilitudes promedios $\overline{d}(i,c)$. Entonces la anchura de la silueta se define como:

$$sw_i = \frac{b_i - a_i}{max(b_i, a_i)}$$

Finalmente la silueta promedio se calcula como:

$$\overline{sw} = \frac{1}{n}\sum_{i=1}^{n} sw_i$$

Entonces observaciones con un ancho de silueta grande se encuentran bien clasificadas, y aquellos que tienen valores pequeños son aquellos que se encuentran distribuidos entre los conglomerados. El índice de silueta toma valores de -1 a 1. Si una observación tiene un valor cercano a 1, entonces el punto está más compenetrado en su propio conglomerado que al de un vecino. Si el valor de la silueta es -1, entonces no está bien clasificado. Un anchura de silueta cerca a 0 indica que las observaciones pueden pertenecer a su conglomerado actual así como a un vecino. Kaufman and Rousseeuw usa la media de silueta para estimar el número de conglomerados. Valores mayores a 0.5 determinan una adecuada

partición, y valores menores a 0.2 indican que no existe estructura de conglomerados.

En el software R los índices de validación, se lo realiza mediante la función $Clvalid$ como sigue:

clValid(obj, nClust, clMethods=, validation=,maxitems=) [11]

donde:

obj es un vector o matriz de observaciones.

$nclust$ es el número de conglomerados a evaluar.

$validation$ es el método a evaluar, en este caso internal.

$maxitem$ es el número máximo de filas.

Con los datos de la tabla 3.10 se tiene la siguiente salida:

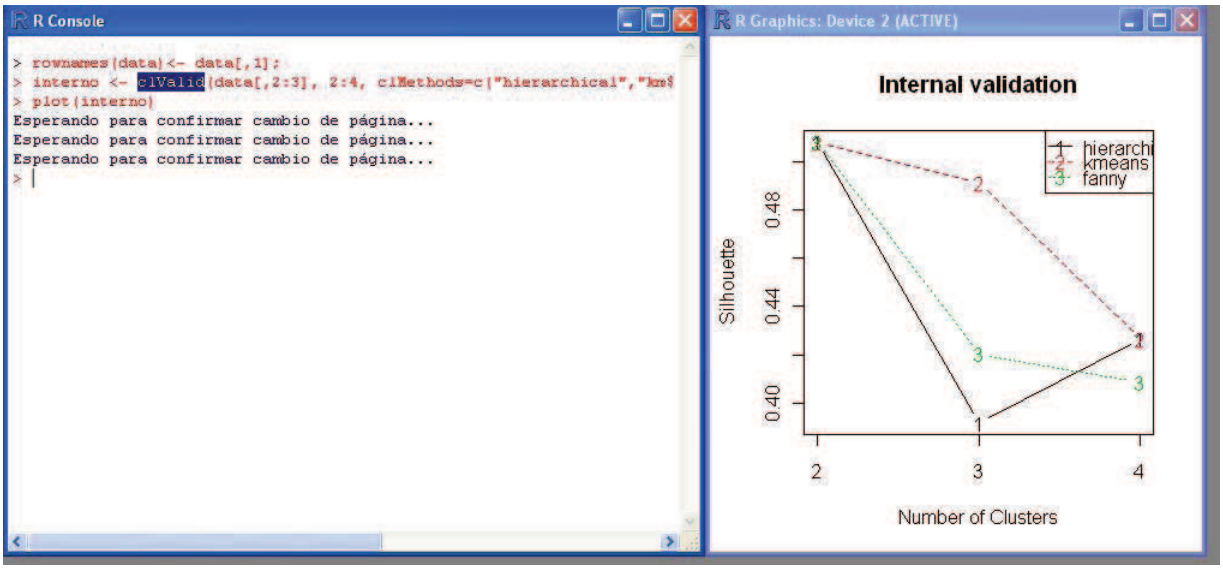


Figura 3.27: Indices de validación R

[11] http://guybrock.gpbrock.net/research

Capítulo 4

Arbol de Clasificación

4.1 Introducción

En la minería de datos, los árboles de clasificación constituyen o pertenecen a los métodos predictivos de segmentación. En la actualidad estos métodos son los más utilizados por los analistas, debido a que son métodos flexibles y transparentes, se encuentran dentro de los métodos no paramétricos, pueden manejar un gran volumen de variables e interacciones complicadas entre ellas, sus resultados son de fácil interpretación. Estos modelos necesitan dos tipos de variables, una llamada objetivo o dependiente y una o varias variables predictivas o independientes [3].

Los árboles de clasificación son particiones secuenciales del conjunto de datos cuya finalidad es maximizar las diferencias entre los valores de la variable dependiente, entonces conlleva por tanto, la división de las observaciones en grupos diferentes mediante el conjunto de variables independientes. Estos métodos se caracterizan por desarrollar un proceso de división de forma arborescente o jerárquica. El método empieza evaluando cada una de las variables explicativas mediante un indice o técnica estadistica, entonces se determina cual de estas variables realiza una mejor clasificación (pureza) de la variable objetivo. finalmente se genera una regla de corte o split y se segmenta al conjunto de datos.

De manera sucesiva se realizan nuevas segmentaciones de cada uno de los segmentos resultantes hasta que el proceso finaliza de acuerdo a un criterio de parada o norma estadística. Y finalmente se obtiene un perfil del segmento. El punto de inicio del árbol de clasificación se llama nodo raíz, el cual está compuesto del conjunto completo de variables de aprendizaje (L). Un nodo es un subconjunto del conjunto de variables, y puede tomar el nombre de terminal o no terminal. Un nodo no terminal o padre, es un nodo que se divide mediante un corte en dos nodos hijos (si es un corte binario). El corte es definido sobre el valor de una única variable. Si los datos cumplen la condición se lo envía a uno de los nodos hijo, caso contrario se lo envía al otro nodo.

Un nodo que no es cortado, es llamado Nodo terminal y se le asigna una etiqueta de la clase. Cada observación en L cae en dentro de uno de los nodos terminales. Cuando una observación de una clase no definida deja de partir al

árbol, entonces esta asigna la etiqueta al nodo terminal. Se puede dar el caso que varios nodos terminales tengan la misma etiqueta. El conjunto de todos los nodos terminales es llamado la partición de los datos.

Considere el siguiente ejemplo [34],se definen dos variables X_1 y X_2. Y las posibles etapas del árbol son:

(1) Es $X_2 \leq \theta_1?$. Si la respuesta es SI, entonces sigue por la rama izquierda; Si NO, continua por la rama derecha. (2) Si la respuesta a (1) es si, entonces se debe responder a la siguiente pregunta: Es $x_1 \leq \theta_2$ Una respuesta afirmativa termina en el nodo τ_1 con la respondiente región $R_1 = \{X_1 \leq \theta_2, X_2 \leq \theta_1\}$; y una respuesta negativa entonces termina en el nodo τ_2 con la correspondiente región $R_2 = \{X_1 > \theta_2, X_2 \leq \theta_2\}$. (3) Si la respuesta a (1) es NO, entonces se debe responder a la siguiente pregunta $X_2 \leq \theta_3?$. Si la respuesta a (3) es SI, entonces se debe responder a la pregunta $X_1 \leq \theta_4?$ una respuesta de SI termina en el nodo τ_3 con la correspondiente región $R_3 = \{X_1 \leq \theta_4, \theta_1 < X_2 \leq \theta_3\}$. Si NO, sigue por la rama derecha para terminar en la región τ_4 con la región correspondiente región $R_4 = \{X_1 > \theta_4, \theta_1 < X_2 \leq \theta_3\}$. (4) Si la respuesta a (3) es NO, entonces se llega hacia el nodo terminal τ_5 con su correspondiente región $R_5 = \{X_2 > \theta_3\}$. En este caso se asume que $\theta_2 < \theta_4$ y $\theta_1 < \theta_3$ Esto se puede observar en el siguiente gráfico:

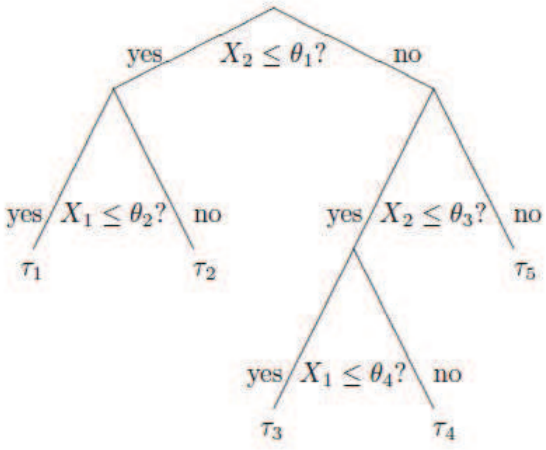


Figura 4.1: Ejemplo de árbol decisión

4.1.1 Ventajas de los árboles de clasificación

- Dado que es un modelo no paramétrico, los resultados son invariantes a una transformación monótona de las variables explicativas.

- La metodología se adapta fácilmente en situaciones donde aparecen datos missing, sin necesidad de eliminar la observación completa.
- Es útil en la valoración de la importancia de las variables y de como interactúan en ellas.
- Las reglas de asignación son simples y legibles, por tanto la interpretación de resultados es directa e intuitiva.

4.1.2 Desventajas

- El árbol final puede que no sea óptimo. La metodología que se aplica sólo asegura que cada subdivisión es óptima.
- Las interacciones de orden menor no preceden a las interacciones de orden mayor.
- Los árboles grandes tienen poco sentido intuitivo y las predicciones tienen, a veces, cierto aire de cajas negras.

4.2 Desarrollo de los árboles

Con la finalidad de desarrollar un árbol de clasificación, se deben responder a las siguientes 4 preguntas:

1. ¿Cómo elegir la condiciones de corte para cada nodo?
2. ¿ Cuál criterio puede ser usado para cortar un nodo padre en dos nodos hijos?
3. ¿Cómo decidir cuando un nodo es terminal?
4. ¿Cómo asignar la etiqueta al nodo terminal?

4.2.1 Estrategia de corte

Para cada nodo, el algoritmo que desarrolla un árbol tiene que decidir sobre cual variable se encuentra el mejor corte. Para esto se debe considerar cada posibilidad de corte sobre todas las variables presentes para ese nodo, entonces enumerando todos los posibles cortes, se evalúa cada uno, y se decide cual es el mejor de acuerdo a una medida.

Para una descripción de las reglas de corte, se debe distinguir entre variables continuas o ordinales y variables categóricas. Para variables continuas u ordinales, el número de posibles cortes para un nodo dado, es $n-1$ cortes, donde n representa los n valores distintos de la variable.

Para una variable categórica, la cual pose M valores distintos, $l_1, .., l_M$, en general hay $2^{M-1}-1$ cortes distintos en S.

Función de Impureza

Para elegir el mejor corte, sobre todas las variables, primero se necesita determinar el mejor corte por cada una de las variables. entonces de acuerdo a esto, se define una medida de bondad de corte.

Sea $\Pi_1, .., \Pi_k$ con $k \geq 2$ clases. Para el nodo τ se define la función de impureza $i(\tau)$ del nodo como:

$$i(\tau) = \phi(p(1/\tau), ..., p(k/\tau)) \tag{4.1}$$

Donde $p(k/\tau)$ es una estimación de $P(X \in \Pi_k/\tau)$, la probabilidad condicional de que una observación X sea de la clase Π_k dado que éste pertenece al nodo τ. En 4.1, se requiere que ϕ sea una función simétrica, definida sobre el conjunto de todas las posibles k-tuplas de probabilidad $(p_1, .., p_k)$, las cuales deben sumar 1, esta función se minimiza en los puntos $(1, 0, ..., 0)$, $(0, 1, ..., 0)$,.., $(0, 0, ..., 1)$ y se maximiza en el punto $(1/k, ..., 1/k)$. En el caso de dos clases $(k = 2)$, las condiciones de maximización de la función $\phi(p)$ se reducen a $p = 1/2$, con $\phi(0) = \phi(1) = 0$.

Algunas de las funciones ϕ son las siguientes:

Índice de Entropía.

$$i(\tau) = -\sum_{k=1}^{K} p(k/\tau) log p(k/\tau), \tag{4.2}$$

Índice de diversidad Gini.

$$i(\tau) = -\sum_{k \neq k'} p(k/\tau)p(k'/\tau) = 1 - \sum_{k} \{p(k/\tau)\}^2 \tag{4.3}$$

4.2.2 Eligiendo el mejor corte para una variable

Tomando en cuenta un nodo τ, al aplicar un corte s tal que una proporción p_L de observaciones son enviadas hacia el nodo hijo izquierdo τ_L y el resto p_R se dirige hacia el nodo hijo derecho τ_R , entonces la bondad del corte s para un τ, es dado

por la reducción de la impureza ganada al cortar el nodo padre τ en sus nodos hijos τ_R y τ_L,

$$\Delta_i(s,\tau) = i(\tau) - p_L i(\tau_L) - p_R i(\tau_R) \tag{4.4}$$

El mejor corte para una variable X_j, es aquel que tiene el valor mas grande de Δ_i sobre todos los $s \in S_j$, el conjunto de los distintos cortes para X_j.

4.2.3 Partición recursiva

Con el objetivo de desarrollar un árbol, éste se inicia en su nodo raíz L. Usando el criterio de mejor corte para una variable, el árbol encuentra el mejor corte para el nodo raíz para cada una de las variables $X_1, .., X_r$, El mejor corte s para el nodo raíz es aquel que tiene el valor más grande de acuerdo a 4.4 sobre todas las otras variables. Los próximos cortes para los nodos hijos, son calculados de similar forma.

4.2.4 Tamaño adecuado o problema de sobreajuste

Una de las características de los árboles de clasificación es que si no se estable un limite para el número de divisiones, se consigue siempre una clasificación pura, en la que en cada nodo contiene una única clase de objetos. Las clasificaciones puras presentan varios inconvenientes porque suelen ser poco realistas, se corre el riesgo de encontrar pocos términos en la clase y además llegan a extraer toda la información de los datos, incluída aquella información ruído, propia de la muestra que se está analizando, ésto se conoce como sobre ajuste, para combatirla se han planteado numerosas estrategias diferentes y en ocasiones complementarias. Dos de las principales son las reglas de parada y poda.

Reglas de paradas

Una primera estrategia consiste en detener la generación de nuevas divisiones cuando éstas supongan una mejora muy pequeña de la predicción. Las principales son:

- Extensión máxima del árbol, es decir, número de divisiones permitidos por debajo del nodo raíz.
- Mínimo número de casos en un nodo.

- Mínima fracción de objetos, que consiste en que los nodos no contengan más casos que una fracción determinada del tamaño de una o más clases.

Como se observa la regla se establece apriori por el analista de acuerdo a su experiencia o estudios previos .

Para validar el resultado del árbol después de haber detenido el proceso de división, se tienen las siguientes opciones:

- Validación cruzada en dos mitades, la cual consiste en dividir los datos en dos partes, la muestra de estimación y la muestra de validación, desarrollar un árbol a partir de la muestra de estimación y utilizarlo para predecir la clasificación de la muestra de validación

- Validación cruzada en v partes, De la muestra disponible, se extrae aleatoriamente v submuestras y se calcula el árbol de clasificación, cada vez dejando afuera una de las v muestras para validar el análisis, de tal manera que cada muestra s utiliza v-1 veces para obtener el árbol y una sola vez para validarlo. Útil para muestras pequeñas.

- Validación cruzada global, la cual replica el análisis completo un número determinado de veces apartando una fracción de los datos, para validar el árbol seleccionado.

Poda

Tras analizar los diferentes reglas de parada durante años, Breiman, Friedman, Losen y Stone (1984) llegan a la conclusión de que resulta imposible especificar una regla que sea totalmente fiable. Existe siempre el riesgo de no descubrir estructuras relevantes en los datos debido a una finalización prematura del análisis. Por ello un enfoque alternativo se lo define en dos fases. En la primera fase, se desarrolla un árbol enorme o completo con un sin número de nodos. En una segunda fase, el árbol es podado, eliminando las ramas innecesarias hasta dar con el tamaño adecuado del árbol.

Poda por coste-complejidad.

Esta técnica, intenta llegar a un acuerdo o estabilidad entre la precisión y el tamaño del árbol. La complejidad del árbol viene dada por el número de nodos terminales (hojas) que posee. Si T es el árbol de decisión usado para clasificar N casos de entrenamiento y se clasifican erróneamente M de ellos, la medida de coste-complejidad de T para un parámetro de complejidad α es:

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

Donde $L(T)$ es el número de hojas del árbol T y $R(T) = M/N$ es un estimador del error de T . Es decir, $R_\alpha(T)$ es una combinación lineal del coste del árbol y de su complejidad.

El árbol podado será el subárbol de mínimo error, aquel que minimice la medida de costecomplejidad $R_\alpha(T)$. Hay que resaltar que conforme el parámetro de complejidad crece el tamaño del árbol que minimiza $R_\alpha(T)$ decrece.

Ejemplo 4.2.1. *A continuación se presenta un ejemplo que ilustra el cálculo de una primera iteración, de un árbol de clasificación* [1]*:*

Obs	Género	Posesión de vehículo	Costo del vehículo	Nivel de Ingreso	Tipo de vehículo
obs_1	F	0	C	L	B
obs_2	M	0	C	L	B
obs_3	M	1	C	M	B
obs_4	M	1	C	M	B
obs_5	F	1	E	H	C
obs_6	F	2	E	H	C
obs_7	M	2	E	M	C
obs_8	F	1	C	M	T
obs_9	F	1	S	M	T
obs_{10}	M	0	C	M	T

Tabla 4.1: Tabla de datos para árbol de clasificación

[1] http://people.revoledu.com/kardi/tutorial/DecisionTree/how-decision-tree-algorithm-work.htm

En este caso la variable objetivo es tipo de vehículo, la cual será explicada por las variables género, posesión de vehículo, costo del vehículo y nivel de ingreso.

Como primer paso se calcula la frecuencia absoluta y relativa de la variable objetivo.

Categorías Vehículo	Frec.	Frec. relativa
Total B	4	0.4
Total C	3	0.3
Total T	3	0.3
Total	10	1

Tabla 4.2: Frecuencia variable objetivo

Luego con la ecuación 4.3 y las frecuencias relativas de la tabla 4.2 se calcula el índice de Gini:

$$i(\text{Tipo de vehículo}) = 1 - (0.4^2 + 0.3^2 + 0.3^2) = 0.66$$

Ahora por cada una de las variables se genera una tabla que contabiliza el número de repeticiones que tiene cada una de las categorías de la variable objetivo versus las categorías de la variable independiente.

Género	F	M	Fr. rel. F	Fr. rel. M
Total B	1	3	0.2	0.6
Total C	2	1	0.4	0.2
Total T	2	1	0.4	0.2
Total	5	5	1	1

Tabla 4.3: Género vs. Tipo vehículo

Entonces se calcula el índice de impureza de Gini.

$$i(\text{Género } F) = 1 - (0.2^2 + 0.4^2 + 0.4^2) = 0.64$$

$$i(\text{Género } M) = 1 - (0.6^2 + 0.2^2 + 0.2^2) = 0.56$$

Luego se calcula la ganancia de información que se esperaría si la variable Género es utilizada para dividir el nodo, mediante la ecuación 4.4.

$$\Delta(\text{género,Tipo vehículo}) = i(\text{Tipo de vehículo}) - (5/10)i(\text{género } F) - (5/10)i(\text{género } M)$$

$$\Delta(\text{género,Tipo vehículo}) = 0.66 - (5/10) * 0.64 - (5/10) * 0.56$$

$$\Delta(\text{género,Tipo vehículo}) = 0.06$$

El mismo procedimiento se lo realiza para las demás variables obteniendo:

$$\Delta(\text{género,Tipo vehículo}) = 0.06$$

$$\Delta(\text{posesión Veh.,Tipo vehículo}) = 0.20$$

$$\Delta(\text{Costo Veh.,Tipo vehículo}) = 0.5$$

$$\Delta(\text{Nivel de ngresos,Tipo vehículo}) = 0.29$$

De los resultados anteriores la variable costo del vehículo es la que aporta con una mejor ganancia de información para clasificar los datos, entonces se realiza el corte por esta variable, obteniendo dos nodos uno con la categoría C y el otro contiene las categorías $E\,y\,S$, luego con los datos de frecuencias relativas de la variable objetivo se calcula el índice de gini y luego se calcula la mejora de la impureza calculando la ganancia de información con el índice de gini del nodo padre, en este caso el nodo raíz y las dos índices gini de los nodos hijos.

Categorías Vehículo	Frec.	Frec. relativa
Total B	4	0.57
Total C	0	0
Total T	3	0.48
Total	7	1

Tabla 4.4: Frecuencia variable objetivo nodo 1

$$i(\text{Tipo de vehículo nodo 1}) = 1 - (0.57^2 + 0^2 + 0.48^2) = 0.48$$

$$i(\text{Tipo de vehículo nodo 2}) = 1 - (0^2 + 1^2 + 0^2) = 0$$

Categorías Vehículo	Frec.	Frec. relativa
Total B	0	0
Total C	3	1
Total T	0	0
Total	3	1

Tabla 4.5: Frecuencia variable objetivo nodo 2

Entonces la mejora de la impureza es:

Mejora de impureza $= i$(T. de veh. n_0)$-(7/10)*i$(T. de veh. n_1)$-(3/10)*i$(T. de veh. n_2)

Mejora de impureza $= 0.317$

En el programa R, para realizar un árbol de clasificación se usa la función $rpart$ con las siguientes opciones.

rpart(formula, data,method)

donde:

$Formula$: es la ecuación que describe la relación entre la variable objetivo y las variables predictoras.

$data$: es el conjunto de observaciones.

$method$ el método utilizado para calcular el árbol

A continuación en la figura 4.2 se presenta la sentencia realizada en R

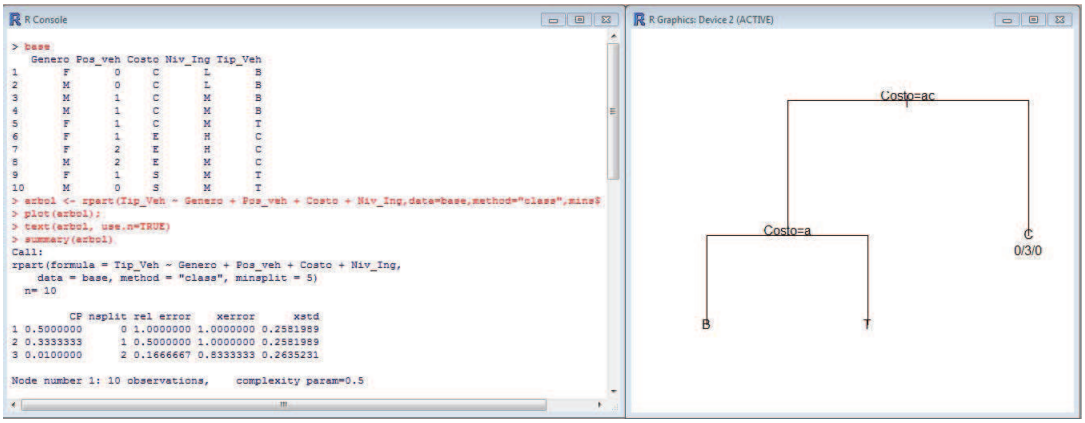


Figura 4.2: Arbol de clasificación en R

4.2.5 Árbol CHAID

CHAID es un método de árboles de decisión que fue presentado por KASS 1980, su siglas significan Chi-Square Automatic Interaction Detection, el método necesita dos tipos de variables, una variable dependiente u objetivos y un conjunto de variables predictivas o independientes, utiliza como principal herramienta el estadístico chi-cuadrado, con el cual CHAID particiona el conjunto de datos, en subconjuntos mutuamente exclusivos que describen de mejor manera la variable dependiente [35]. CHAID primero determina cuál variable predictiva es más efectiva para distinguir los niveles de la variable dependiente, esto lo hace mediante un análisis estadístico de significancia.

CHAID envuelve la aplicación del test de Chi-cuadrado [2] utilizado en las tablas de contingencia, este test es empleado en dos formas. Primero analiza si los diferentes niveles de la variable predictiva pueden ser unidos de manera significativa (test de independencia), entonces luego se determina que variable predictiva es la más importante para distinguir los diferentes niveles de la variable dependiente.

Para el método CHAID, la variable dependiente debe ser categórica, mientras que las variables predictivas pueden ser continuas, nominales y categorías. Cuando la variable es continua, ésta se debe discretizar con la finalidad de obtener una tabla de contingencia adecuada.

El algoritmo asume que la variable dependiente tiene d niveles, y un variable predictiva tiene c niveles. Estos datos puede ser resumidos en una tabla de contingencia con dimensiones cxd. Entonces como primer paso CHAID comprime las c filas de la tabla de contingencia, de tal manera que se incluya solo los niveles de la variable predictora, que realmente son significativos. Es decir, se reduce la tabla de contingencia de tamaño cxd a una tabla de contingencia de dimensión jxd, con niveles significativos. Luego entonces se elige la tabla de dimensión jxd que tiene el estadístico chi-cuadrado más significante. En general el algoritmo cuando dispone de varias variables explicativas realiza los siguientes pasos [25].

1. Para cada variable x se calculan las tablas de contingencia a partir de las categorías de x y de la variable dependiente.

2. Buscar el par de categorías de x cuya tabulación cruzada $2xd$ es menos significativa, Si esta significación no alcanza un determinado valor crítico, entonces ambas categorías se funden y son consideradas como una nueva categoría compuesta independiente.

[2] $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ donde $O_{ij} = n_{ij}$ y $E_{ij} = n_i * n_j/N$

3. Para cada categoría compuesta formada por 3 o más de las categorías originales, se busca la división binaria más significativa tomando en cuenta la corrección propuesta por Kass. Si alcanza un valor critico determinado se parte la categoría y se pasa al paso 2.

4. Cuando más de tres categorías originales, forman una categoría en la tabla de contingencia, Kass utiliza la corrección de Bonferroni para calcular el estadístico de significancia, en el caso de variables categóricas se tiene que el coeficiente es:

$$B = \sum_{i=0}^{r-1} (-1)^i \frac{(r-i)^c}{i!(r-i)!}$$

   donde c es el número de categorías originales y r es el número de grupos reducidos.

5. Calcular la significación de todas las variables explicativas compuestas por sus categorías óptimas y elegir la más significativa. Si este valor es mayor que un valor criterio determinado se subdividen los datos de acuerdo a las categorías de este predictor.

6. Para cada partición de los datos generada que no ha sido analizada, comenzar de nuevo con el paso 1. Este nuevo análisis termina al excluir particiones que no superen un mínimo de observaciones definidas por el analista o un p-valor crítico de división.

Ejemplo 4.2.2. *Se toma como punto de partida los datos de la tabla 4.1 a la cual se la aplicará el algoritmo de árboles CHAID como sigue:*

Como primer paso se realiza el cálculo de las frecuencias de la variable objetivo con la finalidad de analizar como están distribuidas las categorías de la variable.

Categorías Vehículo	Frec.	Frec. relativa
Total B	4	0.4
Total C	3	0.3
Total T	3	0.3
Total	10	1

Tabla 4.6: Frecuencia variable objetivo

Como segundo paso se debe realizar el análisis de tablas de contingencia de cada una de las variables con respecto de la variable objetivo, tomando en cuenta que si existen más de dos categorías en la variable independiente, se deben hacer combinaciones de las mismas.

Por ejemplo en el caso de la variable Costo se tienen 3 categorías $\{C, S, E\}$, por lo que existirían tres combinaciones $\{C, S\}$,$\{C, E\}$ y $\{S, E\}$. Entonces por cada una de las combinaciones, se genera las siguientes tablas de contingencia.

	Costo	Vehículo	
Vehículo	C	S	Frec. Marg.
Total B	4	0	4
Total C	0	0	0
Total T	1	2	3
Frec. Marg.	5	2	7
		Chi-cud.	P-value
		3.73	0.15

Tabla 4.7: Frecuencia variable objetivo vs Costo (C y S)

	Costo	Vehículo	
Vehículo	C	E	Frec. Marg.
Total B	4	0	4
Total C	0	3	3
Total T	1	0	1
Frec. Marg.	5	3	8
		Chi-cud.	P-value
		3	0.22

Tabla 4.8: Frecuencia variable objetivo vs Costo (C y E)

	Costo	Vehículo	
Vehículo	S	E	Frec. Marg.
Total B	0	0	0
Total C	0	3	3
Total T	2	0	2
Frec. Marg.	5	3	5
		Chi-cud.	P-value
		3	0.22

Tabla 4.9: Frecuencia variable objetivo vs Costo (S y E)

Con el objetivo de analizar cuál de las categorías son menos independientes, se analiza el P-valor, cuando éste es mayor a un $\alpha = 0.05$ entonces no se rechaza la hipótesis de independencia, en los tres casos se cumple este supuesto, sin embargo como hay que agrupar las categorías con la finalidad de obtener la más representativas, el valor más cercano 0.05 es la tabla con las categorías C y S, las cuales van a formar una única categoría para el siguiente paso.

Luego de haber realizado el mismo análisis a todas las variables, se debe calcular nuevamente las tablas de contingencia con las nuevas categorías agrupadas. En el caso que las categorías de una variable hayan sido agrupadas, el p-valor calculando debe ser multiplicado por la corrección de Bonferroni que en el caso de la variable género es 1 y para las demás es 3.

		Genero	
Vehículo	F	M	Frec. Marg.
Total B	1	3	4
Total C	2	1	3
Total T	2	1	3
Frec. Marg.	5	5	10
	Chi-cud.	P-value	P-value Bon.
	1.67	0.43	0.43

Tabla 4.10: Frecuencia variable objetivo vs Género

		Posesión Vehículo	
Vehículo	1	0 y 2	Frec. Marg.
Total B	2	2	4
Total C	1	2	3
Total T	2	1	3
Frec. Marg.	5	5	10
	Chi-cud.	P-value	P-value Bon.
	0.67	0.72	2.15

Tabla 4.11: Frecuencia variable objetivo vs Posesión de Vehículo

		Costo Vehículo	
Vehículo	C y S	E	Frec. Marg.
Total B	4	0	4
Total C	0	3	3
Total T	3	0	3
Frec. Marg.	7	3	10
	Chi-cud.	P-value	P-value Bon.
	10	0.01	0.02

Tabla 4.12: Frecuencia variable objetivo vs Costo del Vehículo

Entonces aquella variable que tiene el p-value más significativo es la que se toma para realizar el corte del árbol. en este caso la variable elegida para el corte es Costo de Vehículo obteniendo dos nodos hijos.

		Ingreso	
Vehículo	L y M	H	Frec. Marg.
Total B	4	0	4
Total C	1	2	3
Total T	3	0	3
Frec. Marg.	8	2	10
	Chi-cud.	P-value	P-value Bon.
	5.83	0.05	0.16

Tabla 4.13: Frecuencia variable objetivo vs Nivel de Ingreso

Categorías Vehículo	Frec.	Frec. relativa
Total B	0	0
Total C	3	1
Total T	0	0
Total	3	1

Tabla 4.14: Frecuencia variable objetivo nodo 1

4.3 El problema de la clase desbalanceada

El problema de la clase desbalanceada o desproporcionada dentro de un modelo de clasificación, consiste en encontrar dentro de la variable dependiente u objetivo, una categoría o clase en una proporción mayor en comparación a las otras, produciendo un desbalance entre las categorías. Otra forma de ver este problema es el de considerar que un conjunto de datos es desbalanceado si las categorías de clasificación no son aproximadamente iguales. [3].

Ejemplo sobre clases desbalanceada son observados al realizar análisis tales como:

- Detecciones de fraudes.
- Análisis de supervivencia de Clientes.
- La detección de derrames de petróleo en las imágenes de satélite.
- la clasificación de los píxeles en imágenes de mamografía con posible cáncer.

[3]Tomado de SMOTE: Synthetic Minority Over-sampling Technique chawla, Bowyer,Hall,Kegelmeyer

Categorías Vehículo	Frec.	Frec. relativa
Total B	4	0.57
Total C	0	0
Total T	3	0.43
Total	7	1

Tabla 4.15: Frecuencia variable objetivo nodo 2

Para afrontar el problema de la clase desbalanceada la comunidad de machine learning [4] ha tomado dos caminos:

- Asignar distintos costos para las muestras de entrenamiento
- Realizar un muestreo de los datos originales, ya sea por un sobre muestreo de la clase minoritaria y /o una sub muestra de las clase mayoritaria.

Para combatir el problema del desbalance de la clase, Japkowickz ha propuesto los siguientes métodos[5]:

- **Sobre muestreo**.- Dos métodos de sobre muestreo fueron considerados en esta categoría. El primero un sobre muestreo aleatorio, que consiste en sobre muestrear la clase más pequeña de manera aleatoria hasta obtener el mismo tamaño de registros que la otra clase. El segundo método, un muestreo enfocado, el cual consiste en muestrear la pequeña clase sólo con datos que se encuentran en el límite entre la categoría minoritaria y mayoritaria. Factores de $\alpha = 0.25$ son elegido para representar la igualdad en el límite.
- **Sub muestreo**.- Dos métodos de sub muestreo fueron considerados en esta categoría. El primero un muestreo aleatorio que elimina la clase de mayor tamaño de manera aleatoria hasta obtener el mismo tamaño de registro que la clase minoritaria clase. El segundo método es un muestreo focalizado, que consiste en eliminar elemento más lejanos del límite entre la categoría minoritaria y mayoritaria.
- **Modificación del Costo**.- El método de modificación de costo, consiste en modificar el costo relativo asociado a la mala clasificación de la clase positiva

[4]Tomado de SMOTE: Synthetic Minority Over-sampling Technique chawla, Bowyer,Hall,Kegelmeyer

[5]Japkowickz and Stephen the class Imbalance Problem

y negativa, hasta compensar el ratio de desbalanceado de las dos clases. Por ejemplo, si los datos presentan una proporción de 1 a 9 a favor de la clase negativa, el costo de mala clasificación de un ejemplo positivo representará 9 veces la mala clasificación de un negativo.

Japkowickz concluye que el sobre muestreo aleatorio es mucho más efectivo que el sub muestreo aleatorio, mientras que las técnicas de muestreo enfocado no tienen mayor relevancia. La modificación del costo aborda de manera efectiva el problema de la clase desbalanceada excepto en casos de alta complejidad.[6]

[6]Japkowickz and Stephen the class Imbalance Problem

Capítulo 5

Resultados y Análisis

5.1 Ejecución del Modelo

En este capítulo se aplicará la parte teórica descrita en los capítulos anteriores, con la finalidad de cumplir con los objetivos planteados en este proyecto de titulación.

La ejecución del modelo consta de las siguientes fases:

1. Elección de la muestra de datos y determinación de las variables para el modelo.
2. Análisis estadístico de los datos, con la finalidad de entender y conocer la base.
3. Transformación de los datos, para mejorar el aporte de las variables al modelo.
4. Determinación de conglomerados de acuerdo a variables transaccionales.
5. Identificación de la relación entre los grupos encontrados y sus características demográficas mediante la aplicación de árboles de decisión.

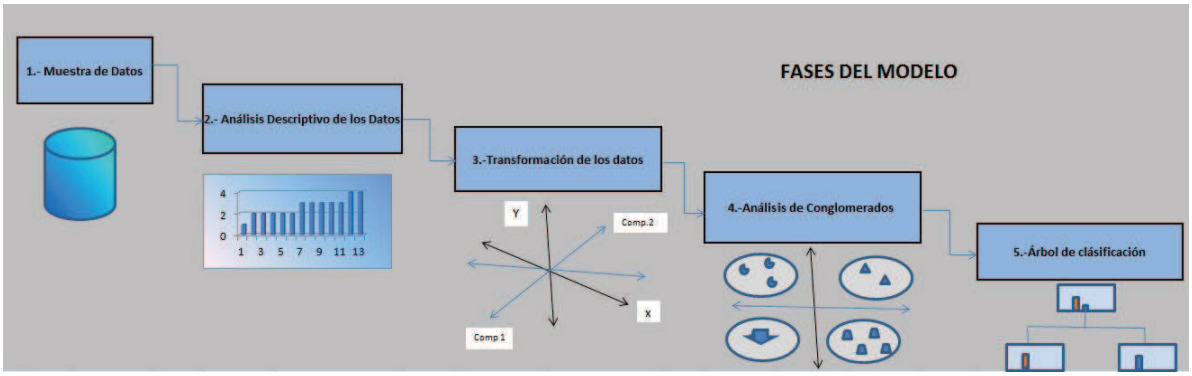


Figura 5.1: Fases del Modelo

Las fases descritas anteriormente, están enmarcadas en la metodología SEMMA analizada en el capítulo 2. su objetivo es transformar los datos en información valiosa para la toma de decisiones.

5.1.1 Base Inicial

Para la base inicial se tomó en cuenta al segmento de personas naturales, con un estado activo en la entidad Bancaria. El período de recolección de datos para la ejecución de este modelo fue de 12 meses, desde Enero 2010 a Diciembre del 2010. Se eligió una muestra de 10000 observaciones. En la tabla 5.1 se presentan las variables levantadas para el análisis.

Variable	Variable
Transacciones ATM	Transacciones Ventanilla
Transacciones Internexo	Transacciones Celular
Transacciones Kiosko	Transacciones Telenexo
Transacciones Balcones	Edad del cliente
Número Cargas	Estado civil
Genero	Vivienda
Nivel de Estudio	Provincia

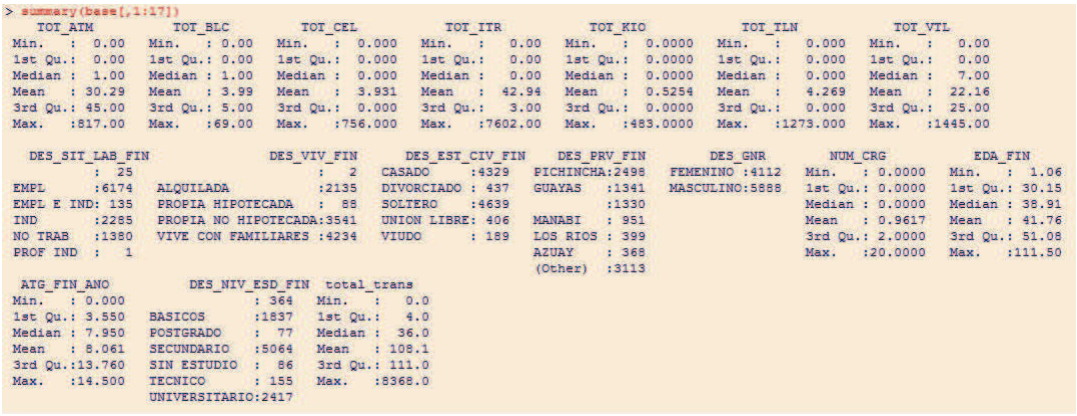
Tabla 5.1: Variables del Modelo

5.1.2 Análisis exploratorio de los Datos

Estadísticos y frecuencias

Con la finalidad de identificar la validez del conjunto de datos, a cada una de las variables se les aplicará estadísticos descriptivos y frecuencias mediante la función summary(base[,1:17]) de R.

En la figura 5.2 se observan los estadísticos media, mínimo, máximo, cuartil 1 y cuartil 3 de cada una de las variables continuas y las frecuencias de cada variable nominal. Con respecto a esta información se observa que:



```
> summary(base[,1:17])
    TOT_ATM          TOT_BLC          TOT_CEL          TOT_ITR          TOT_KIO          TOT_TLN          TOT_VTL
Min.   :  0.00   Min.   :  0.00   Min.   :  0.000   Min.   :   0.00   Min.   :  0.0000   Min.   :   0.000   Min.   :   0.00
1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.:   0.00   1st Qu.:  0.0000   1st Qu.:   0.000   1st Qu.:   0.00
Median :  1.00   Median :  1.00   Median :  0.000   Median :   0.00   Median :  0.0000   Median :   0.000   Median :   7.00
Mean   : 30.29   Mean   :  3.99   Mean   :  3.931   Mean   :  42.94   Mean   :  0.5254   Mean   :   4.269   Mean   :  22.16
3rd Qu.: 45.00   3rd Qu.:  5.00   3rd Qu.:  0.000   3rd Qu.:   3.00   3rd Qu.:  0.0000   3rd Qu.:   0.000   3rd Qu.:  25.00
Max.   :817.00   Max.   : 69.00   Max.   :756.000   Max.   :7602.00   Max.   :483.0000   Max.   :1273.000   Max.   :1445.00

  DES_SIT_LAB_FIN              DES_VIV_FIN      DES_EST_CIV_FIN    DES_PRV_FIN       DES_GNR        NUM_CRG          EDA_FIN
          :  25                       :   2   CASADO     :4329   PICHINCHA:2498   FEMENINO :4112   Min.   : 0.0000   Min.   :  1.06
EMPL      :6174   ALQUILADA           :2135   DIVORCIADO : 437   GUAYAS   :1341   MASCULINO:5888   1st Qu.: 0.0000   1st Qu.: 30.15
EMPL E IND: 135   PROPIA HIPOTECADA   :  88   SOLTERO    :4639                :1330                    Median : 0.0000   Median : 38.91
IND       :2285   PROPIA NO HIPOTECADA:3541   UNION LIBRE: 406   MANABI   : 951                    Mean   : 0.9617   Mean   : 41.76
NO TRAB   :1380   VIVE CON FAMILIARES :4234   VIUDO      : 189   LOS RIOS : 399                    3rd Qu.: 2.0000   3rd Qu.: 51.08
PROF IND  :   1                               AZUAY    : 368                                      Max.   :20.0000   Max.   :111.50
                                              (Other)  :3113

 ATG_FIN_ANO          DES_NIV_ESD_FIN  total_trans
Min.   :  0.000                : 364   Min.   :   0.0
1st Qu.:  3.550   BASICOS      :1837   1st Qu.:   4.0
Median :  7.950   POSTGRADO    :  77   Median :  36.0
Mean   :  8.061   SECUNDARIO   :5064   Mean   : 108.1
3rd Qu.: 13.760   SIN ESTUDIO  :  86   3rd Qu.: 111.0
Max.   : 14.500   TECNICO      : 155   Max.   :8368.0
                  UNIVERSITARIO:2417
```

Figura 5.2: Estadísticos Descriptivos

- El total de transacciones dentro de los cuartiles 1 y 3 de los canales telenexo, kiosko, celular es igual a 0, ésto nos da un indicio de que estos canales no son utilizados con frecuencia.
- La edad de los clientes en promedio es 41,76 años. Su máximo es 111 años y su mínimo es de 1 año [1].
- El promedio de cargas familiares por cliente es de 1, su máximo es 20 y su mínimo des 0.
- La variable total_trans tiene como valor mínimo 0, ésto significa que existen observaciones que no transaccionan por ninguno de los canales.
- En la variable situación laboral existen 25 valores perdidos o missing.
- En la variable vivienda existen 2 valores perdidos o missing.
- En la variable provincia constan 1330 valores perdidos o missing siendo éste un valor muy elevado.
- En la variable nivel de estudios existen 364 valores perdidos o missing.

Con respecto al número de cargas, se realizó la figura de distribución 5.3 y se observó que el mayor conjunto de datos se apilan en el valor 0, también se detecta la presencia de valores atípicos a partir del valor 6.

En la figura 5.4 se observa que las categorías más representativas de la variable Estado Civil son solteros y casados mientras que la menos representativas es Viudo.

[1] Se debe a cuentas de Ahorro e inversiones para niños

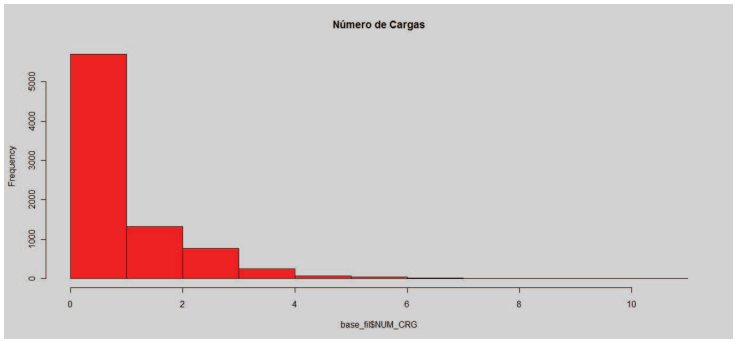


Figura 5.3: Frecuencias Número de Cargas

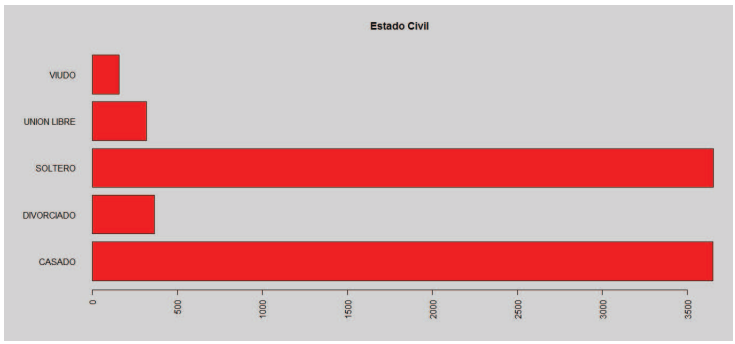


Figura 5.4: Frecuencias Estado Civil

En la figura 5.5 se visualiza que las categorías más representativas en la variable provincia son Pichincha, Guayas y Manabí, además se observa una alta frecuencia de valores ausentes 803.

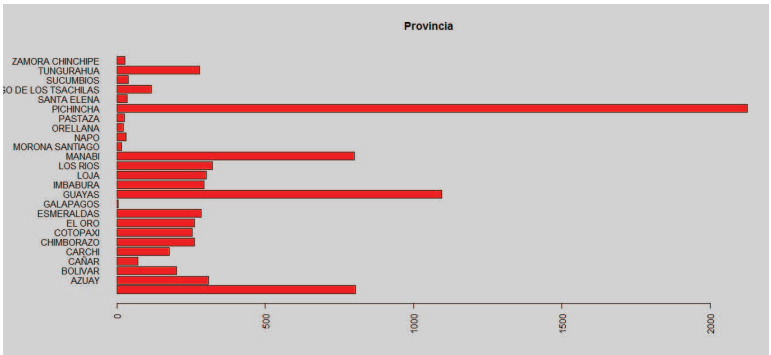


Figura 5.5: Frecuencias Provincia

En la figura 5.6 se observa que las categorías de la variable vivienda más representativas son vive con familiares, propias no hipotecaria y alquiler, mientras que la categoría propia hipotecada tiene una frecuencia pequeña.

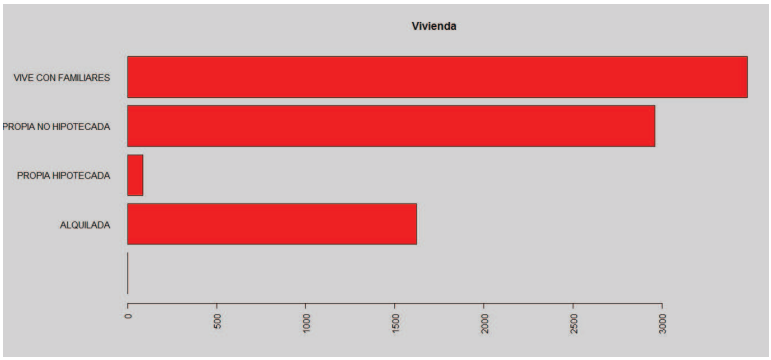


Figura 5.6: Frecuencias Vivienda

En la figura 5.7 se visualiza que la categoría masculino de la variable género es la más representativa.

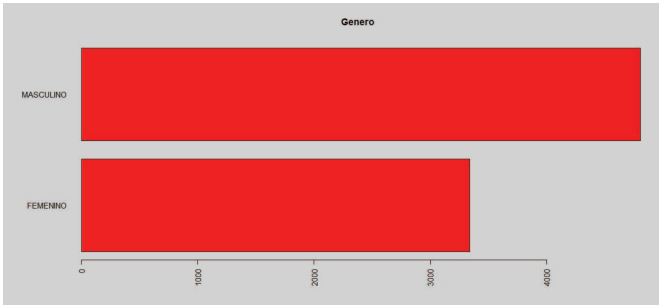


Figura 5.7: Frecuencias Género

En la figura 5.8 se observa que las categorías más representativas de la variable estudio, son estudios secundarios y universitarios.

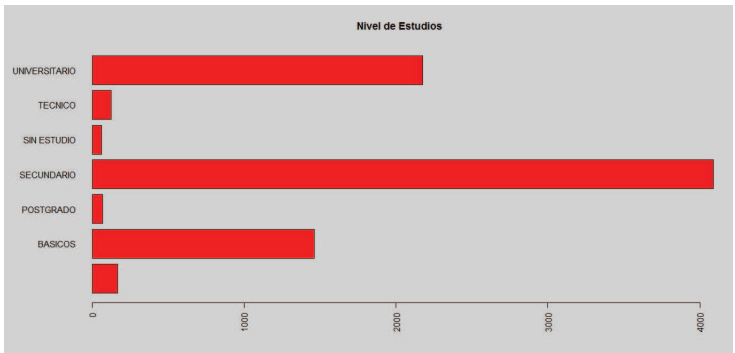


Figura 5.8: Frecuencias Nivel estudios

En la figura 5.9 se visualiza que las categorías más representativas son Empleados e Independientes. universitarios.

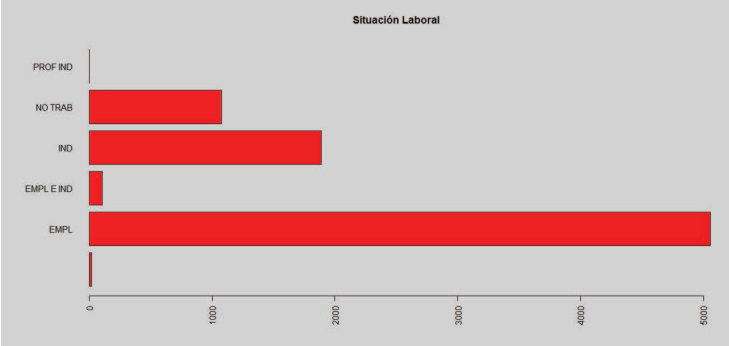


Figura 5.9: Frecuencias Situación Laboral

De acuerdo a los items anteriores, el uso de los estadísticos descriptivos nos permite obtener un resumen claro de cómo se encuentran los datos, en este caso, teniendo en mente que uno de los objetivo es analizar el perfil transaccional de los clientes, como primer paso se debe determinar cuales registros tienen el valor de 0 en todos los canales, pues estas observaciones no son relevantes para el estudio, aunque pueden dar un indicio de deserción de clientes.

Del conjunto de datos iniciales, introduciendo la condición de no transacción en todos los canales, en la tabla 5.2 se visualizan los siguientes resultados:

```
base_fil <- subset(base,total_trans >0);
```

Base	N. observaciones
Inicial	10,000
Base 1	8,147
Diferencia	1,853

Tabla 5.2: Modificación de la Base

Se observa que existen $1,853$ clientes que no transaccionan, los mismos que representa el 18% de la muestra.

Datos Atípicos

Al momento de observar los estadísticos de las transacciones por canal se observa que los Cuartiles 1 y 3 se encuentran lejanos del mínimo y máximo respectivamente, una de las razones que puede generar esta situación es la existencia de

valores atípicos, por otro lado dado que la primera fase del modelo es la extracción de grupos mediante un algoritmo de conglomerados, es importante detectar valores extraños que modifiquen o distorsionan los resultados obtenidos.

Para el análisis de atípicos se utilizó el algoritmo descrito en el capítulo 3, su implementación se encuentra en el apéndice.

Para alcanzar un total de valores atípicos cercanos al 3%, el porcentaje elegido para calificar un grupo como atípico fue 0.5%. Los resultados alcanzados se los presenta en la figura 5.10.

```
> summary(base_mod[,5:11]);
    TOT_ATM          TOT_BLC          TOT_CEL          TOT_ITR          TOT_KIO          TOT_TLN          TOT_VTL
 Min.   :  0.00   Min.   : 0.000   Min.   :  0.000   Min.   :   0.00   Min.   :  0.0000   Min.   :  0.000   Min.   :  0.00
 1st Qu.:  0.00   1st Qu.: 0.000   1st Qu.:  0.000   1st Qu.:   0.00   1st Qu.:  0.0000   1st Qu.:  0.000   1st Qu.:  3.00
 Median : 11.00   Median : 2.000   Median :  0.000   Median :   0.00   Median :  0.0000   Median :  0.000   Median : 12.00
 Mean   : 34.65   Mean   : 4.696   Mean   :  2.944   Mean   :  37.87   Mean   :  0.5413   Mean   :  3.105   Mean   : 21.92
 3rd Qu.: 56.00   3rd Qu.: 6.000   3rd Qu.:  0.000   3rd Qu.:   8.00   3rd Qu.:  0.0000   3rd Qu.:  0.000   3rd Qu.: 29.00
 Max.   :440.00   Max.   :67.000   Max.   :240.000   Max.   :1047.00   Max.   :483.0000   Max.   :355.000   Max.   :319.00
> summary(base_ati[,5:11]);
    TOT_ATM          TOT_BLC          TOT_CEL          TOT_ITR          TOT_KIO          TOT_TLN          TOT_VTL
 Min.   :  0.0   Min.   : 0.00   Min.   :  0.00   Min.   :   0.0   Min.   :  0.000   Min.   :   0.0   Min.   :   0.0
 1st Qu.:  1.5   1st Qu.: 4.00   1st Qu.:  0.00   1st Qu.:   0.0   1st Qu.:  0.000   1st Qu.:   0.0   1st Qu.:  33.0
 Median : 56.0   Median : 8.00   Median :  0.00   Median : 141.0   Median :  0.000   Median :   4.0   Median : 106.0
 Mean   :110.5   Mean   :10.75   Mean   : 59.48   Mean   : 484.0   Mean   :  3.657   Mean   :  67.3   Mean   : 180.7
 3rd Qu.:179.5   3rd Qu.:14.50   3rd Qu.:  2.00   3rd Qu.: 602.5   3rd Qu.:  0.000   3rd Qu.:  25.0   3rd Qu.: 262.5
 Max.   :817.0   Max.   :69.00   Max.   :756.00   Max.   :7602.0   Max.   :209.000   Max.   :1273.0   Max.   :1445.0
```

Figura 5.10: Resultados método multivariante Atípicos

Donde la base_mod con 7.876 registros, tiene los valores no atípicos y la base_ati con 271 registros, tiene los valores atípicos. Como se puede ver los promedios son muy diferentes para cada grupo, los datos atípicos presentan promedios de transacciones altas.

5.1.3 Transformación de los datos

Análisis de Componentes Principales

En esta etapa se realiza un análisis de componentes principales[2] con la finalidad de verificar si es necesario reducir el número de variables del análisis y así disminuir la variabilidad de los datos.

Para analizar la pertinencia de la aplicación del ACP, se realiza la prueba Bartlett [3], los resultados de la prueba se presentan en la tabla 5.3.

[2]Ver apéndice para la teoría

[3]descrita en el apéndice A.3.3

	Prueba de Bartlett
Chi Cuadrado	1828.131
Grados de Libertad	21
Significancia	0,000

Tabla 5.3: Resultados Prueba Bartlett

Dada la significancia=0, entonces se rechaza la hipótesis de ortogonalidad de las variables, y ésto nos indica que es adecuado realizar el análisis de componetes principales.

Para la ejecución del análisis de componentes principales se utiliza el procedimiento princomp de R y se introduce como variables de entrada para el análisis, al total de transacciones por canal. Los resultados se presentan en las figuras 5.11 y 5.12:

```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7
Standard deviation     1.2827103 1.0198496 0.9835341 0.9732652 0.9345687 0.9081154 0.8377852
Proportion of Variance 0.2350494 0.1485848 0.1381913 0.1353207 0.1247741 0.1178105 0.1002692
Cumulative Proportion  0.2350494 0.3836341 0.5218255 0.6571462 0.7819203 0.8997308 1.0000000
```

Figura 5.11: Resultados Componentes Principales

En la figura 5.11 se muestra la variabilidad recogida por cada una de las componente, como se puede observar el componente 1 representa el 23% de los datos, los demás componentes aportan casi de manera proporcional a la variabilidad, lo que significa que deben ser considerados para el modelo.

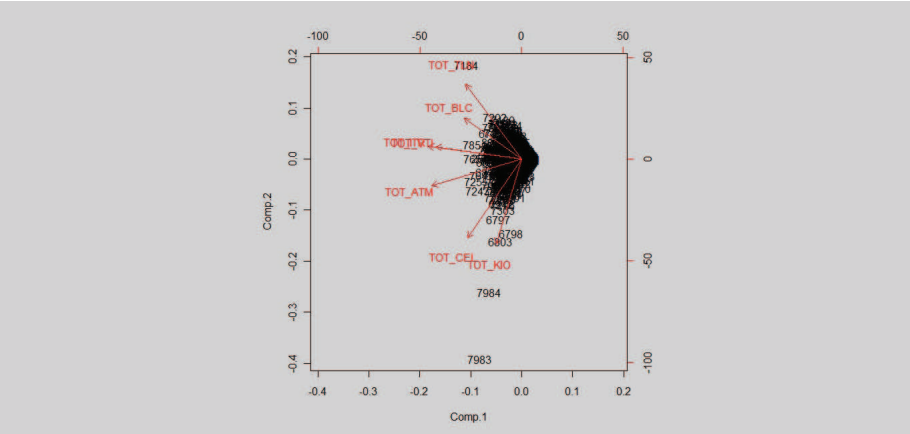


Figura 5.12: Figura de puntos en las nuevas coordenadas

En la figura 5.12 se presenta la gráfica de las variables en los dos primeros componentes, a simple vista se puede identificar 4 grupos: $\{Balcones, Telenexo\}$,

$\{Ventanilla, Internet\}$, $\{ATM\}$ y $\{Kiosko, Celular\}$

La figura 5.13 muestra el gráfico de sedimentación de los componentes principales, el mismo sugiere que son 6 los componentes a conservar (valores cercanos 1). sin embargo como ya se analizó anteriormente, la diferencia de contribución a la variabilidad de los datos entre los componentes no es significativa. Además como la intensión, es utilizar las puntuaciones de las observaciones en los nuevos componentes como entrada para la segmentación, se tomará en cuenta los resultados obtenidos para las 7 componentes.

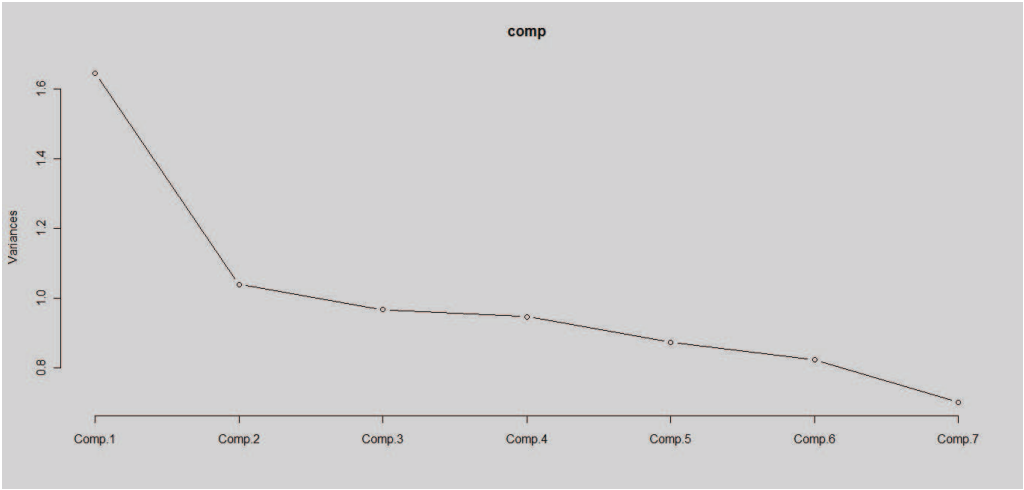


Figura 5.13: Gráfico de Sedimentación Componentes Principales

Estandarización de variables

Al aplicar un modelo multivariante una de las transformaciones fundamentales antes de ejecutar el mismo, es la estandarización de las variables, esto permite comparar en una misma escala todas las variables. La ecuación que se aplica es :

$$z_{ij} = \frac{x_{ij} - \overline{x_j}}{\sigma_{x_j}} \ \forall i = 1...n \ y \ \forall \ j = n...p$$

Se utiliza el proceso de estandarización de R llamado scale y se introducen como variables de análisis las componentes obtenidas del paso anterior y el resultado se presentan en la figura 5.14.

Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
Min. :-9.619e+00	Min. :-3.474e+01	Min. :-4.665e+01	Min. :-7.000e+00	Min. :-2.125e+01	Min. :-7.696e+00	Min. :-7.200e+00
1st Qu.:-3.234e-01	1st Qu.:-1.686e-01	1st Qu.:-1.361e-01	1st Qu.:-2.516e-01	1st Qu.:-2.438e-01	1st Qu.:-3.528e-01	1st Qu.:-3.910e-01
Median : 2.999e-01	Median :-6.244e-02	Median : 6.555e-03	Median : 2.063e-01	Median :-3.300e-02	Median : 3.747e-02	Median :-1.069e-01
Mean : 2.280e-19	Mean :-4.741e-18	Mean : 1.104e-17	Mean : 8.786e-18	Mean : 8.734e-18	Mean : 1.353e-17	Mean : 6.359e-18
3rd Qu.: 6.973e-01	3rd Qu.: 2.013e-01	3rd Qu.: 1.998e-01	3rd Qu.: 4.240e-01	3rd Qu.: 3.739e-01	3rd Qu.: 3.991e-01	3rd Qu.: 4.356e-01
Max. : 9.376e-01	Max. : 1.630e+01	Max. : 5.827e+00	Max. : 6.684e+00	Max. : 5.877e+00	Max. : 6.291e+00	Max. : 6.269e+00

Figura 5.14: Resultados Estandarización de variables

Como se puede observar cada una de las variables tiene media 0.

5.1.4 Modelo de Conglomerados

En esta etapa después de haber estandarizado los datos, y luego de ver que la reducción de la dimensión es conveniente, entonces se procede a buscar los grupos de cliente transaccionales.

Como se mencionó en el capítulo 3 para cumplir el objetivo de establecer los grupos se debe:

- Establecer la medida de proximidad.
- El algoritmo de agrupamiento.

Dentro de los algoritmos de conglomerados una de las partes fundamentales es la determinación del número de grupos, para esto se describió en la sección 3.4 un conjunto de índices que permiten determinar y validar el número de conglomerados.

Estimación del número de Conglomerados

Una forma de estimar el número de grupos es analizando la suma de cuadrados dentro de los grupos vs el número de grupo, el número de grupos serà elegido cuando la curva empiece a estabilizarse[18] .

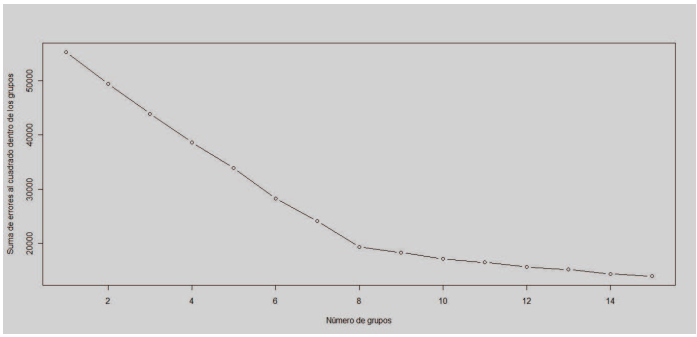


Figura 5.15: Números de grupos mediante el Índice de suma de cuadrados dentro de los grupos

En la figura 5.15 se observa que la curva se estabiliza a partir de n=8. Otro de los índices utilizados fueron los de Hartigan (1975) y Calinski y Harabasz (1974)[28].

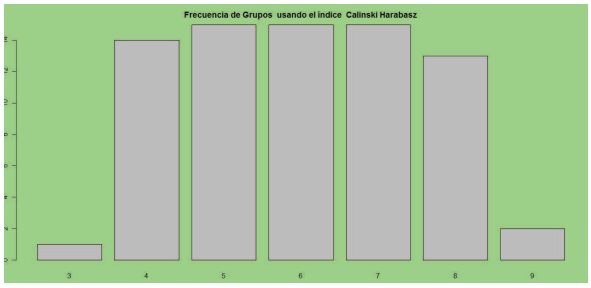


Figura 5.16: Números de grupos mediante el Índice de Hartigan

En la figura 5.16 se visualiza que las frecuencias obtenidas por la ejecución iterativa del índice de Hartigan nos otorga como número de grupos, valores entre 4 y 8.

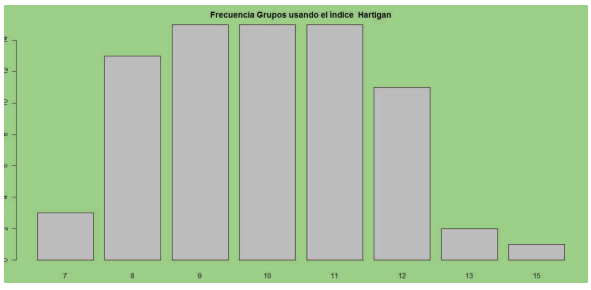


Figura 5.17: Números de grupos mediante el Índice de Calinski y Harabasz

En la figura 5.17 se observa que las frecuencias obtenidas por la ejecución iterativa del índice de Hartigan nos otorga como número de grupos, valores entre 8 y 11.

Analizando los 3 índices, el número de grupos coincidente entre éstos es n=8.

El algoritmo de Conglomerados

De acuerdo al análisis anterior, se encontró que el número de grupos estimados para el conjunto de variables transaccionales (ATM, internet, cajas, etc.), es 8, entonces se aplican dos algoritmos de conglomerados k- means y c-means [4] para determinar los grupos de clientes(perfiles transaccionales).

Algoritmo k-medias

Luego de aplicar el algoritmo de conglomerados k-medias se obtuvieron los siguientes resultados:

Tamaño de los conglomerados En la figura 5.18 se observa los porcentaje en tamaño de los ocho conglomerados estimados. El grupo más grande es el número 2 con 4012 observaciones, y el grupo más pequeño es el 8 con 96 observaciones.

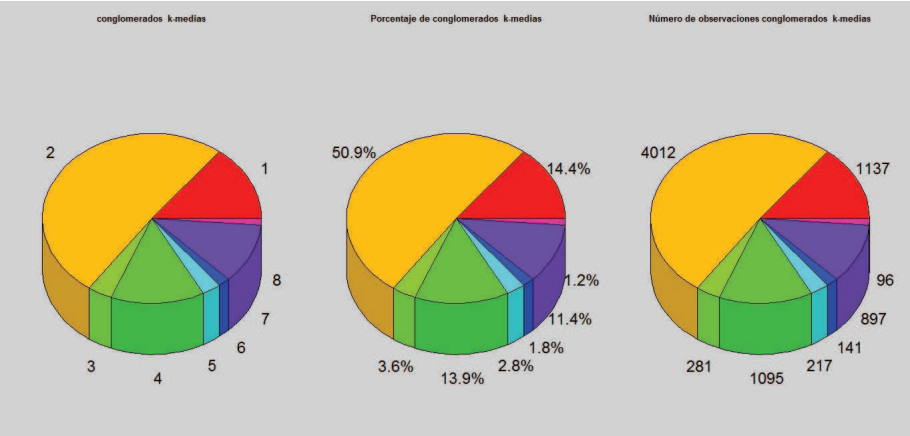


Figura 5.18: Tamaño de los grupos

A continuación se presenta un análisis descriptivo tanto de las variables transaccionales y demográficas, de cada uno de los conglomerados obtenidos con el método k-medias:

Conglomerado 1

group: 1

	var	n	mean	sd	median	trimmed	mad	min	max	range	se
TOT_ATM	1	1137	118.45	43.53	106	112.86	40.03	64	306	242	1.29
TOT_BLC	2	1137	4.13	4.82	3	3.26	4.45	0	29	29	0.14
TOT_CEL	3	1137	1.33	6.92	0	0.00	0.00	0	60	60	0.21
TOT_ITR	4	1137	37.17	66.98	0	20.27	0.00	0	337	337	1.99
TOT_KIO	5	1137	1.18	12.40	0	0.00	0.00	0	331	331	0.37
TOT_TLN	6	1137	3.19	7.26	0	1.26	0.00	0	49	49	0.22
TOT_VTL	7	1137	20.86	19.46	16	17.96	16.31	0	110	110	0.58

Figura 5.19: Centro de conglomerado 1 k-medias

[4] Observe el capítulo 3

En la figura 5.19 se observa que los clientes en este grupo, realizan un mayor volumen de transacciones en el canal ATM.

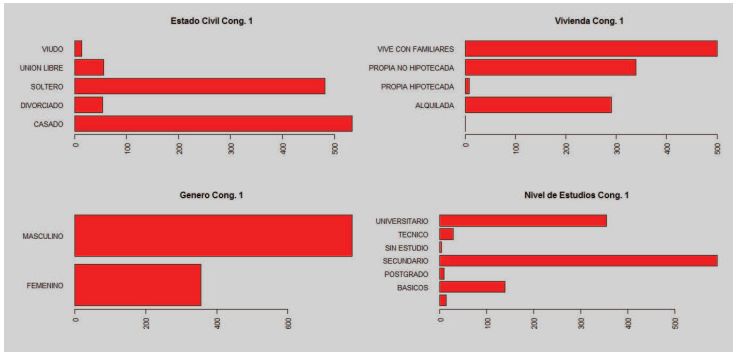


Figura 5.20: Variable Demográficas conglomerado 1 k-medias

En la figura 5.20 se visualiza que los clientes en su mayoría tienen un estado civil Casado y Soltero, tienen como vivienda la categoría Vive con Familiares y propia no Hipotecada, Son Hombres y en nivel de estudio predominan las categoria secundaria y universitario .

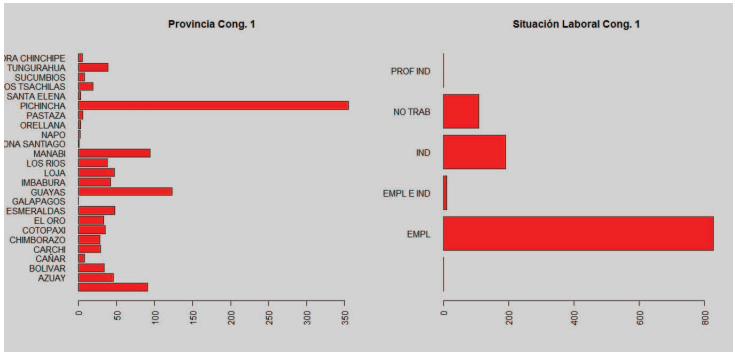


Figura 5.21: Variable Provincia y Situación laboral conglomerado 1 k-medias

La provincia en este grupo se encuentra representada por Pichincha y la situación laboral por la categoria Empleado.

group: 1

	var	n	mean	sd	median	trimmed	mad	min	max	range	se
NUM_CRG	1	878	0.94	1.22	0.00	0.74	0.00	0.00	7.00	7.00	0.04
EDA_FIN	2	878	39.15	13.54	36.61	37.98	13.85	2.95	92.07	89.12	0.46

Figura 5.22: Variable Edad y Cargas familiares conglomerado 1 k-medias

En la figura 5.22 se observa que en promedio el grupo tiene 1 carga familiar y una edad promedio de 39 años.

Conglomerado 2

En la figura 5.23 se visualiza que los clientes en este grupo, tienen un volumen bajo de transacciones en los canales.

```
group: 2
         var    n  mean    sd median trimmed  mad min max range   se
TOT_ATM    1 4012 13.49 19.17      1    9.73 1.48   0  92    92 0.30
TOT_BLC    2 4012  1.74  2.48      0    1.24 0.00   0  11    11 0.04
TOT_CEL    3 4012  0.64  4.53      0    0.00 0.00   0  55    55 0.07
TOT_ITR    4 4012 12.39 38.53      0    1.93 0.00   0 268   268 0.61
TOT_KIO    5 4012  0.18  3.37      0    0.00 0.00   0 189   189 0.05
TOT_TLN    6 4012  1.50  4.50      0    0.28 0.00   0  41    41 0.07
TOT_VTL    7 4012  8.25  8.38      6    7.10 8.90   0  40    40 0.13
```

Figura 5.23: Centro de conglomerado 2 k-medias

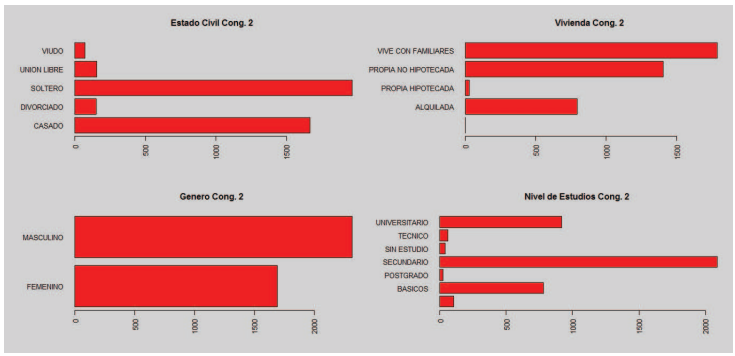


Figura 5.24: Variable Demográficas conglomerado 2 k-medias

En la figura 5.24 se observa que los clientes en su mayoría tienen un estado civil Casado y Soltero, tiene como vivienda la categoría Vive con Familiares y propia no Hipotecada. Son Hombres y Mujeres y en nivel de estudio predomina la categoría secundaria.

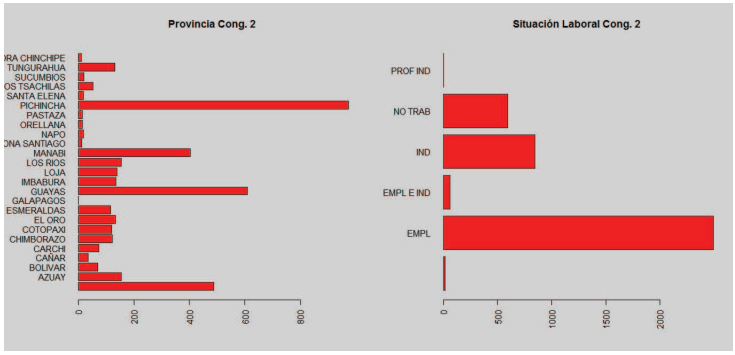


Figura 5.25: Variable Provincia y Situación laboral conglomerado 2 k-medias

La provincia en este grupo se encuentra representada por Pichincha y Guayas; y la situación laboral por la categoria Empleado.

```
group: 2
          var   n  mean    sd median trimmed   mad   min   max range   se
NUM_CRG     1 195  1.04  1.18    1.0    0.87  1.48  0.00  5.00  5.00 0.08
EDA_FIN     2 195 38.61 14.05   35.3   36.95 13.67 19.98 87.73 67.75 1.01
```

Figura 5.26: Variable Edad y Cargas familiares conglomerado 2 k-medias

En la figura 5.26 se visualiza que en promedio el grupo tiene 1 carga familiar y una edad promedio de 38 años.

Conglomerado 3

```
group: 3
         var   n   mean     sd median trimmed    mad min  max range   se
TOT_ATM    1 281  66.04  53.03     63   61.44  50.41   0  440   440 3.16
TOT_BLC    2 281   5.38   5.90      3    4.39   4.45   0   37    37 0.35
TOT_CEL    3 281   3.60  12.40      0    0.14   0.00   0   87    87 0.74
TOT_ITR    4 281 517.37 191.85    458  495.87 185.32 267 1047   780 11.44
TOT_KIO    5 281   1.37   6.94      0    0.04   0.00   0   76    76 0.41
TOT_TLN    6 281   5.99  10.15      1    3.77   1.48   0   78    78 0.61
TOT_VTL    7 281  41.25  38.68     32   34.96  28.17   0  257   257 2.31
```

Figura 5.27: Centro de conglomerado 3 k-medias

En la figura 5.27 se observa que los clientes en este grupo, realizan un mayor volumen de transacciones en el canal ATM, Internet y Ventanilla .

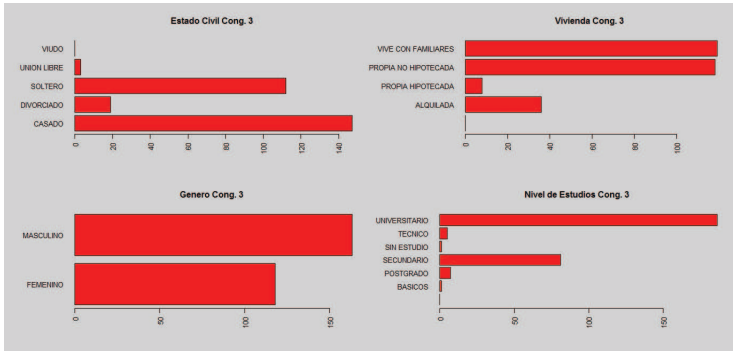


Figura 5.28: Variable Demográficas conglomerado 3 k-medias

En la figura 5.28 se visualiza que los clientes en su mayoría tienen un estado civil Casado y Soltero, tiene como vivienda la categoría Vive con Familiares y propia no Hipotecada, son Hombres y Mujeres; y en nivel de estudio predomina la categoría Universitario.

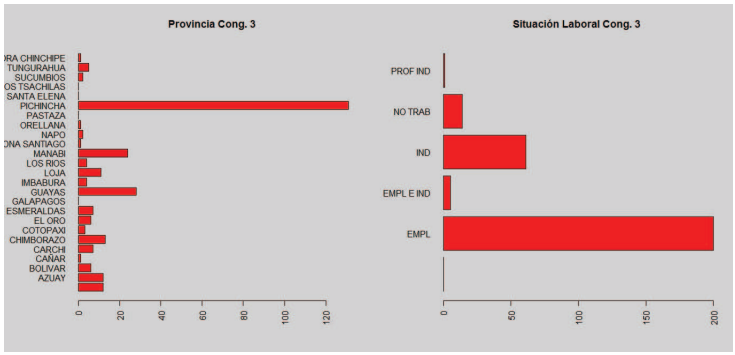


Figura 5.29: Variable Provincia y Situación laboral conglomerado 3 k-medias

La provincia en este grupo se encuentra representada por Pichincha y la situación laboral por la categoria Empleado.

```
group: 3
         var   n  mean    sd median trimmed   mad   min  max range   se
NUM_CRG    1 281  1.02  1.23   1.00    0.85  1.48  0.00  6.0  6.00 0.07
EDA_FIN    2 281 40.11 10.92  38.49   39.31 11.83 20.65 71.9 51.25 0.65
```

Figura 5.30: Variable Edad y Cargas familiares conglomerado 3 k-medias

En la figura 5.30 se observa que en promedio el grupo tiene 1 carga familiar y una edad promedio de 40 años.

Conglomerado 4

```
group: 4
         var    n  mean    sd median trimmed   mad min max range   se
TOT_ATM    1 1095 14.87 23.57      0   10.01  0.00   0 102   102 0.71
TOT_BLC    2 1095  2.69  3.45      1    2.05  1.48   0  17    17 0.10
TOT_CEL    3 1095  0.67  4.78      0    0.00  0.00   0  59    59 0.14
TOT_ITR    4 1095 12.93 41.87      0    1.41  0.00   0 286   286 1.27
TOT_KIO    5 1095  0.92 15.28      0    0.00  0.00   0 483   483 0.46
TOT_TLN    6 1095  1.85  5.89      0    0.29  0.00   0  42    42 0.18
TOT_VTL    7 1095 49.51 16.51     45   47.88 16.31  27  97    70 0.50
```

Figura 5.31: Centro de conglomerado 4 k-medias

En la figura 5.31 se visualiza que los clientes en este grupo, realizan un mayor volumen de transacciones en el canal ATM y Ventanilla .

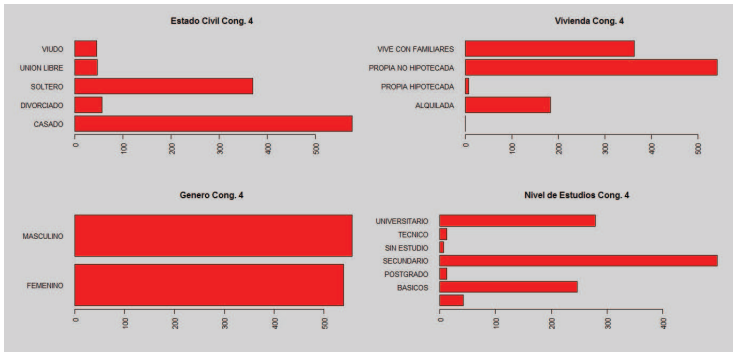


Figura 5.32: Variable Demográficas conglomerado 4 k-medias

En la figura 5.32 se observa que los clientes en su mayoría tienen un estado civil Casado y Soltero, tiene como vivienda la categoría Vive con Familiares y propia no Hipotecada, son Hombres y Mujeres; y en nivel de estudio predomina la categoría Secundario.

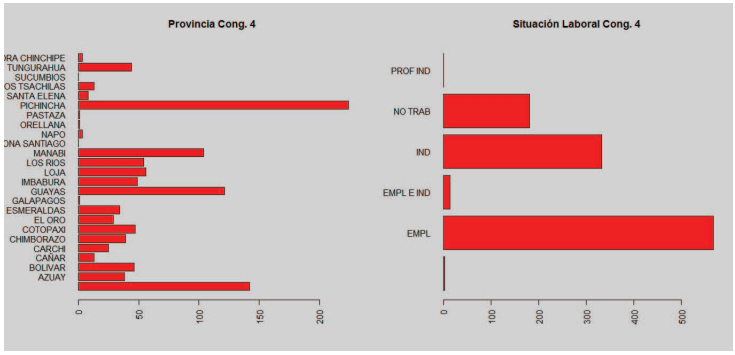


Figura 5.33: Variable Provincia y Situación laboral conglomerado 4 k-medias

La provincia en este grupo se encuentra representada por Pichincha y la situación laboral por las categorias Empleado e Independiente.

```
group: 4
        var   n  mean   sd median trimmed  mad  min   max range   se
NUM_CRG   1 207  1.28 1.41   1.00    1.11 1.48 0.00  7.00  7.00 0.10
EDA_FIN   2 207 47.55 13.91 47.28   47.37 12.59 6.43 87.76 81.33 0.97
```

Figura 5.34: Variable Edad y Cargas familiares conglomerado 4 k-medias

En la figura 5.34 se visualiza que en promedio el grupo tiene 1 carga familiar y una edad promedio de 47 años.

Conglomerado 5

En la figura 5.35 se observa que los clientes en este grupo, realizan un mayor volumen de transacciones en el canal ATM y Internet.

```
group: 5
         var   n   mean     sd median trimmed   mad min max range   se
TOT_ATM    1 217  51.17  61.87     23   41.40 34.10   0 252   252 4.20
TOT_BLC    2 217   7.90   8.16      5    6.58  5.93   0  43    43 0.55
TOT_CEL    3 217   1.99   9.89      0    0.00  0.00   0  61    61 0.64
TOT_ITR    4 217  72.30 145.12      0   37.76  0.00   0 975   975 9.85
TOT_KIO    5 217   0.63   2.95      0    0.00  0.00   0  30    30 0.20
TOT_TLN    6 217   3.45   8.24      0    1.23  0.00   0  58    58 0.56
TOT_VTL    7 217 136.42  38.82    126  130.87 29.65  84 302   218 2.64
```

Figura 5.35: Centro de conglomerado 5 k-medias

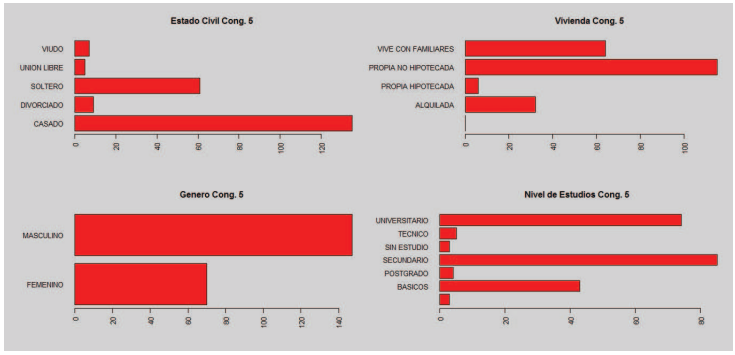


Figura 5.36: Variable Demográficas conglomerado 5 k-medias

En la figura 5.36 se visualiza que los clientes en su mayoría tienen un estado civil Casado, tiene como vivienda la categoría Vive con Familiares y propia Hipotecada, son Hombres y en nivel de estudio predominan las categoría Secundario y Universitario.

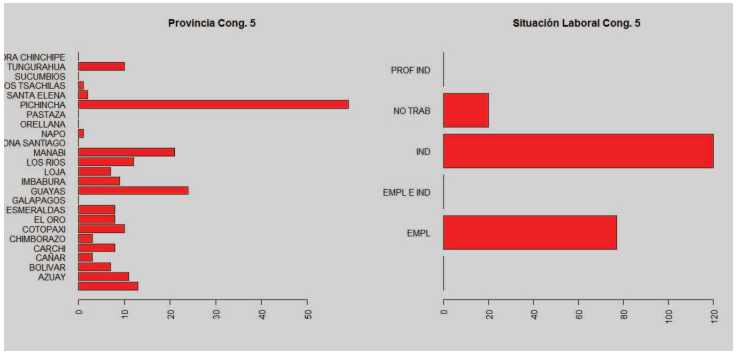


Figura 5.37: Variable Provincia y Situación laboral conglomerado 5 k-medias

La provincia en este grupo se encuentra representada por Pichincha y la situación laboral por las categorias Independiente y Empleado.

```
group: 5
         var    n  mean    sd median trimmed   mad  min   max range   se
NUM_CRG    1 3626  0.89  1.26   0.00    0.66  0.00 0.00  11.0 11.00 0.02
EDA_FIN    2 3626 40.85 15.21  37.55   39.33 13.69 1.99 111.5 109.51 0.25
```

Figura 5.38: Variable Edad y Cargas familiares conglomerado 5 k-medias

En la figura 5.38 se observa que en promedio el grupo tiene 1 carga familiar y una edad promedio de 40 años.

Conglomerado 6

```
group: 6
         var   n   mean     sd median trimmed   mad min max range    se
TOT_ATM    1 141  81.72  60.35     69   74.94 45.96   0 303   303  5.08
TOT_BLC    2 141   7.43   7.65      5    6.26  5.93   0  35    35  0.64
TOT_CEL    3 141 110.83  45.03     96  105.29 35.58  56 240   184  3.79
TOT_ITR    4 141  81.56 152.68     14   43.80 20.76   0 790   790 12.86
TOT_KIO    5 141   2.64  11.10      0    0.27  0.00   0  89    89  0.94
TOT_TLN    6 141   3.28   5.67      0    2.01  0.00   0  35    35  0.48
TOT_VTL    7 141  36.00  41.33     25   28.50 23.72   0 319   319  3.48
```

Figura 5.39: Centro de conglomerado 6 k-medias

En la figura 5.39 se observa que los clientes en este grupo, realizan un mayor volumen de transacciones en el canal Celular,ATM, Internet y kiosko.

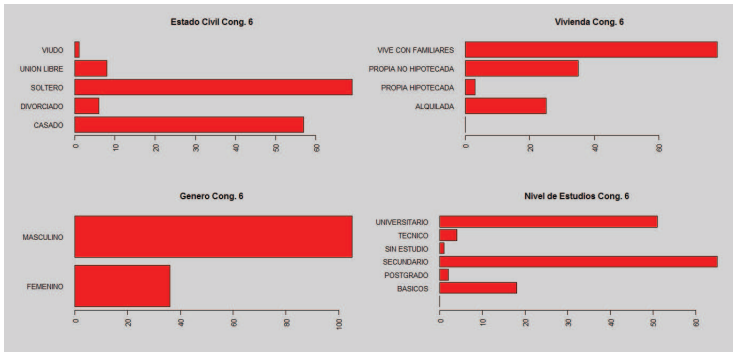


Figura 5.40: Variable Demográficas conglomerado 6 k-medias

En la figura 5.40 se visualiza que los clientes en su mayoría tienen un estado civil Soltero y Casado, tiene como vivienda la categoría Vive con Familiares, son Hombres y en nivel de estudio predominan las categoría Secundario y Universitario.

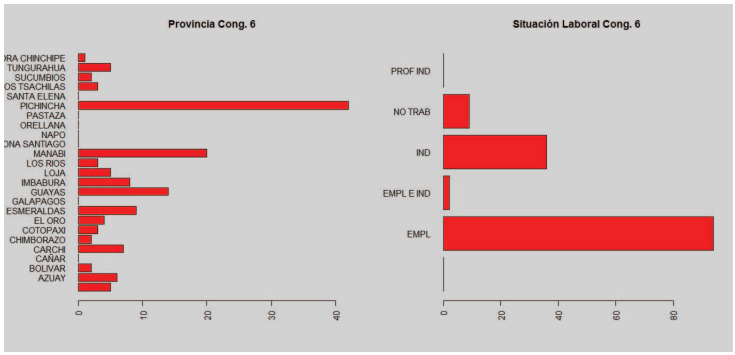


Figura 5.41: Variable Provincia y Situación laboral conglomerado 6 k-medias

La provincia en este grupo se encuentra representada por Pichincha y la situación laboral por la categoria Empleado.

```
group: 6
        var    n  mean    sd median trimmed   mad   min   max range   se
NUM_CRG   1 1028  1.15  1.40   1.00    0.94  1.48  0.00  9.00  9.00 0.04
EDA_FIN   2 1028 50.07 15.53  48.47   49.46 16.89 11.25 98.92 87.67 0.48
```

Figura 5.42: Variable Edad y Cargas familiares conglomerado 6 k-medias

En la figura 5.42 se observa que en promedio el grupo tiene 1 carga familiar y una edad promedio de 50 años.

Conglomerado 7

```
group: 7
        var   n  mean    sd median trimmed   mad min max range   se
TOT_ATM   1 897 25.73 31.74     12   20.14 17.79   0 177   177 1.06
TOT_BLC   2 897 19.40  7.60     19   18.27  5.93  11  67    56 0.25
TOT_CEL   3 897  1.23  6.06      0    0.00  0.00   0  57    57 0.20
TOT_ITR   4 897 15.91 43.38      0    4.43  0.00   0 380   380 1.45
TOT_KIO   5 897  0.27  2.60      0    0.00  0.00   0  57    57 0.09
TOT_TLN   6 897  2.37  6.58      0    0.58  0.00   0  44    44 0.22
TOT_VTL   7 897 12.65 15.85      7    9.53  8.90   0 120   120 0.53
```

Figura 5.43: Centro de conglomerado 7 k-medias

En la figura 5.43 se visualiza que los clientes en este grupo, realizan un mayor volumen de transacciones en el canal ATM y Balcones .

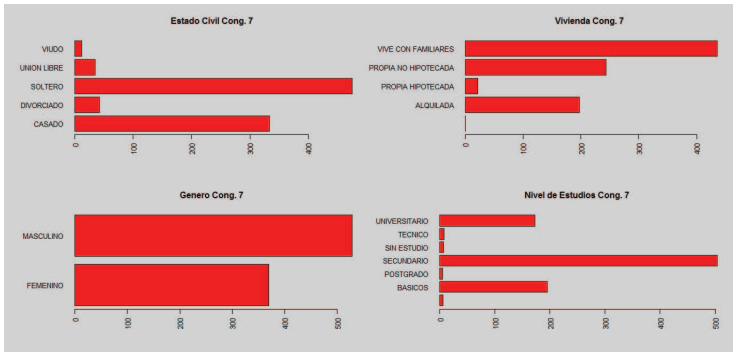


Figura 5.44: Variable Demográficas conglomerado 7 k-medias

En la figura 5.44 se observa que los clientes en su mayoría tienen un estado civil Soltero y Casado, tiene como vivienda la categoría Vive con Familiares y propia Hipotecada, son Hombres y Mujeres; y en nivel de estudio predomina la categoría Secundario.

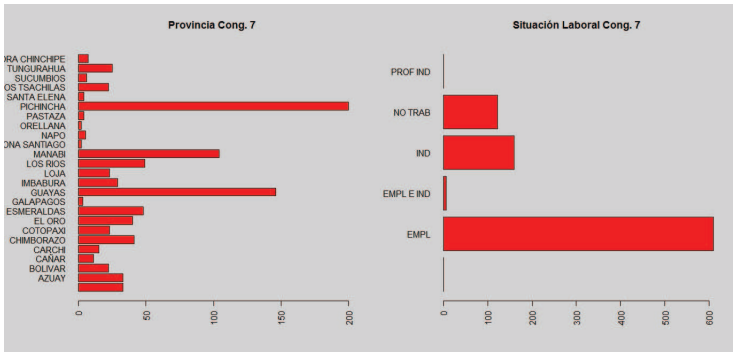


Figura 5.45: Variable Provincia y Situación Laboral conglomerado 7 k-medias

La provincia en este grupo se encuentra representada por Pichincha y Guayas; y la situación laboral por la categoria Empleado.

```
group: 7
        var    n  mean    sd median trimmed   mad   min   max range   se
NUM_CRG   1 1138  1.15  1.39   1.00    0.94  1.48  0.00 11.00 11.00 0.04
EDA_FIN   2 1138 39.22 12.79  35.94   38.00 12.34 11.73 85.37 73.64 0.38
```

Figura 5.46: Variable Edad y Cargas familiares conglomerado 7 k-medias

En la figura 5.46 se visualiza que en promedio el grupo tiene 1 carga familiar y una edad promedio de 39 años.

Conglomerado 8

En la figura 5.47 se observa que los clientes en este grupo, realizan un mayor volumen de transacciones en el canal ATM y Telenexo.

group: 8

	var	n	mean	sd	median	trimmed	mad	min	max	range	se
TOT_ATM	1	96	37.17	43.35	21	30.44	31.13	0	195	195	4.42
TOT_BLC	2	96	6.95	6.70	5	5.99	4.45	0	38	38	0.68
TOT_CEL	3	96	1.95	9.40	0	0.00	0.00	0	68	68	0.96
TOT_ITR	4	96	54.90	115.03	0	27.03	0.00	0	674	674	11.74
TOT_KIO	5	96	0.73	4.11	0	0.01	0.00	0	39	39	0.42
TOT_TLN	6	96	81.00	39.88	71	76.29	29.65	42	355	313	4.07
TOT_VTL	7	96	41.80	38.46	35	35.94	31.13	0	181	181	3.93

Figura 5.47: Centro de conglomerado 8 k-medias

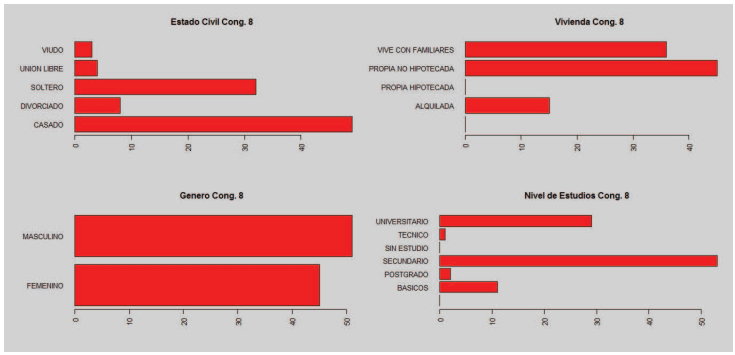


Figura 5.48: Variable Demográficas conglomerado 8 k-medias

En la figura 5.48 se visualiza que los clientes en su mayoría tienen un estado civil Soltero y Casado, tiene como vivienda la categoría Vive con Familiares y propia Hipotecada, son Hombres y Mujer; y en nivel de estudio predomina la categoría Secundario.

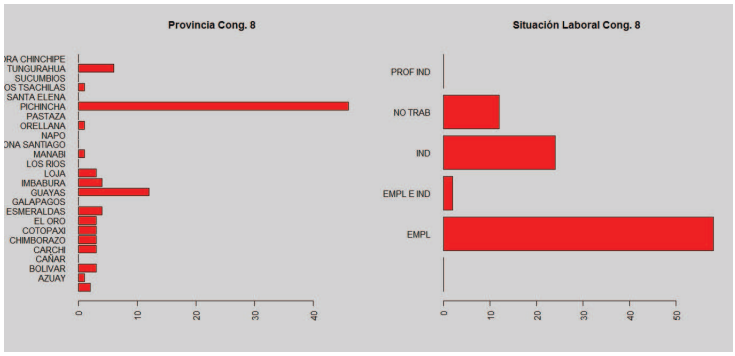


Figura 5.49: Variable Provincia y Situación Laboral conglomerado 8 k-medias

La provincia en este grupo se encuentra representada por Pichincha y la situación laboral por la categoría Empleado.

En la figura 5.50 se observa que en promedio el grupo tiene 1 carga familiar y una edad promedio de 34 años.

```
group: 8
         var   n  mean    sd median trimmed   mad  min   max range   se
NUM_CRG    1 544  0.74  1.20   0.00    0.48  0.00 0.00  6.00   6.0 0.05
EDA_FIN    2 544 34.38 13.51  31.11   33.23 12.34 1.06 82.86  81.8 0.58
```

Figura 5.50: Variable Edad y Cargas familiares conglomerado 8 k-medias

Algoritmo c-medias

Luego de aplicar el algoritmo de conglomerados c-medias se obtuvieron los siguientes resultados:

Tamaño de los conglomerados En la figura 5.51 se observa los porcentaje en tamaño de los ocho conglomerados estimados. El grupo más grande es el número 4 con 2515 observaciones, y el grupo más pequeño es el 3 con 144 observaciones.

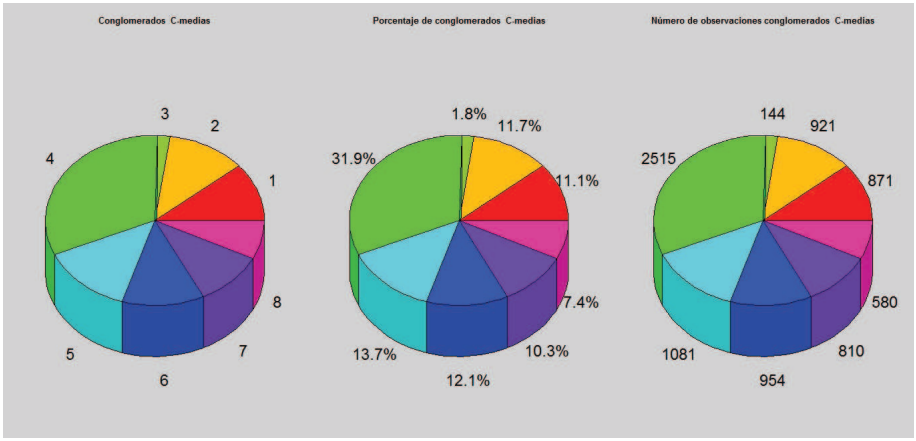


Figura 5.51: Tamaño de los grupos

A continuación se presenta un análisis descriptivo tanto de las variables transaccionales y demográficas, de cada uno de los conglomerados:

Conglomerado 1

```
group: 1
         var   n   mean    sd median trimmed   mad min  max range   se
TOT_ATM    1 871 135.88 43.63    128  130.65 41.51  81  440   359 1.48
TOT_BLC    2 871   4.53  5.22      3    3.57  4.45   0   29    29 0.18
TOT_CEL    3 871   2.54 13.37      0    0.00  0.00   0  144   144 0.45
TOT_ITR    4 871  39.13 80.95      0   19.97  0.00   0 1047  1047 2.74
TOT_KIO    5 871   0.64  3.74      0    0.01  0.00   0   52    52 0.13
TOT_TLN    6 871   3.35  8.07      0    1.29  0.00   0   84    84 0.27
TOT_VTL    7 871  24.51 26.97     17   19.68 17.79   0  248   248 0.91
```

Figura 5.52: Centro de conglomerado 1 C-medias

En la figura 5.52 se visualiza que los clientes en este grupo, realizan un mayor volumen de transacciones en el canal ATM y Internet.

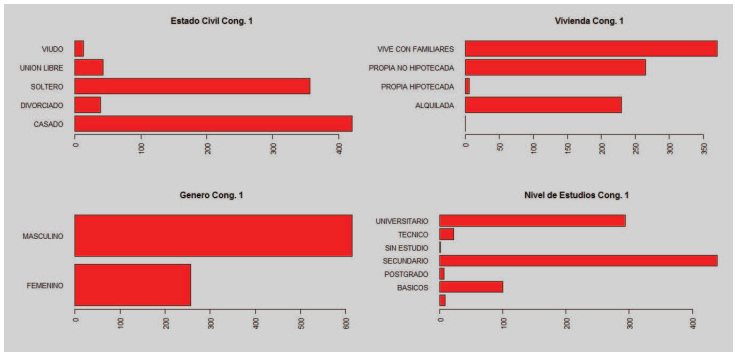


Figura 5.53: Variable Demográficas conglomerado 1 C-medias

En la figura 5.53 se observa que los clientes en su mayoría tienen un estado civil Soltero y Casado, tiene como vivienda la categoría Vive con Familiares y propia no Hipotecada, son Hombres y en nivel de estudio predomina la categoría Secundario.

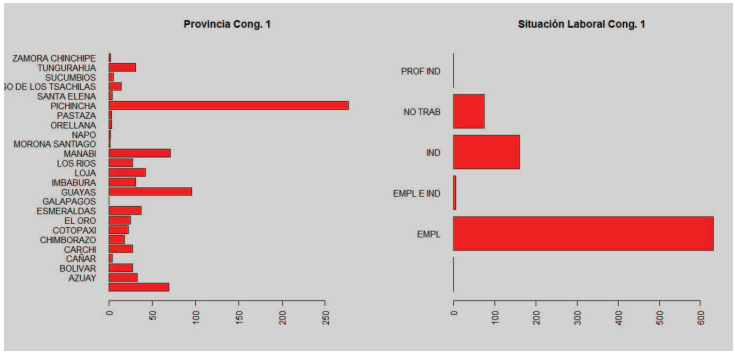


Figura 5.54: Variable Provincia y Situación laboral conglomerado 1 C-medias

La provincia en este grupo se encuentra representada por Pichincha y la situación laboral por la categoria Empleado.

```
group: 1
         var   n  mean    sd median trimmed   mad   min   max range   se
NUM_CRG    1 871  1.20  1.39   1.00    1.01  1.48  0.00 11.00 11.00 0.05
EDA_FIN    2 871 39.73 12.83  36.99   38.60 12.88 11.73 85.37 73.64 0.43
```

Figura 5.55: Variable Edad y Cargas familiares conglomerado 1 C-medias

En la figura 5.55 se observa que en promedio el grupo tiene 1 carga familiar y una edad promedio de 39 años.

Conglomerado 2

En la figura 5.56 se visualiza que los clientes en este grupo, realizan un mayor volumen de transacciones en el canal Balcones .

```
group: 2
         var   n  mean    sd median trimmed   mad min max range   se
TOT_ATM    1 921 11.59 12.82      7    9.80 10.38   0  51    51 0.42
TOT_BLC    2 921  5.75  2.74      5    5.60  2.97   0  13    13 0.09
TOT_CEL    3 921  1.33  7.28      0    0.00  0.00   0  87    87 0.24
TOT_ITR    4 921 10.49 29.77      0    2.48  0.00   0 233   233 0.98
TOT_KIO    5 921  0.28  3.12      0    0.00  0.00   0  72    72 0.10
TOT_TLN    6 921  2.23  5.91      0    0.62  0.00   0  44    44 0.19
TOT_VTL    7 921 11.61  9.92     10   10.63 11.86   0  43    43 0.33
```

Figura 5.56: Centro de conglomerado 2 C-medias

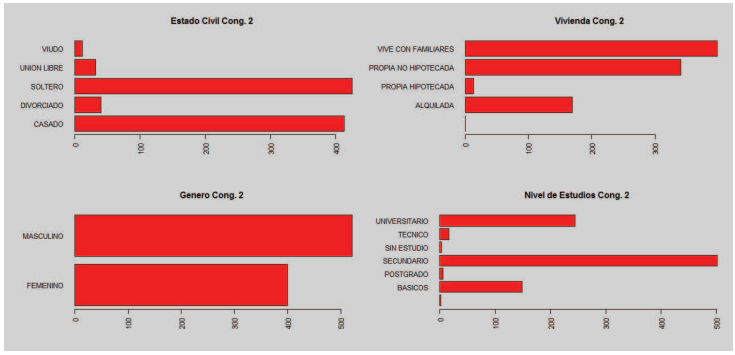


Figura 5.57: Variable Demográficas conglomerado 2 c-medias

En la figura 5.57 se observa que los clientes en su mayoría tienen un estado civil Soltero y Casado, tiene como vivienda la categoría Vive con Familiares y propia no Hipotecada, son Hombres y Mujer; y en nivel de estudio predomina la categoría Secundario.

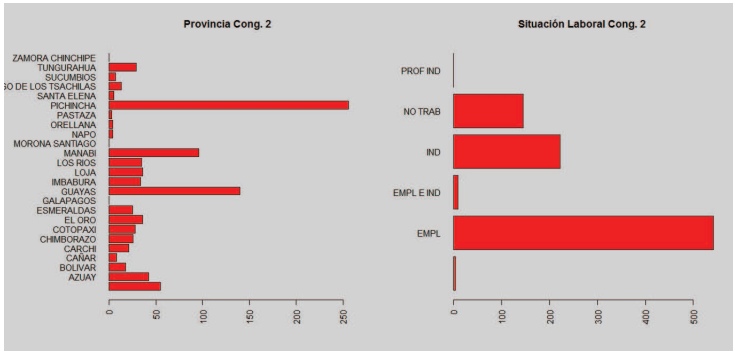


Figura 5.58: Variable Provincia y Situación laboral conglomerado 2 c-medias

La provincia en este grupo se encuentra representada por Pichincha y Guayas; y la situación laboral por la categoria Empleado.

```
group: 2
         var   n  mean    sd median trimmed   mad  min   max range   se
NUM_CRG    1 921  0.94  1.22   0.00    0.73  0.00 0.00  7.00  7.00 0.04
EDA_FIN    2 921 40.21 14.51  37.55   39.04 15.76 5.07 88.25 83.18 0.48
```

Figura 5.59: Variable Edad y Cargas familiares conglomerado 2 C-medias

En la figura 5.59 se visualiza que en promedio el grupo tiene 1 carga familiar y una edad promedio de 40 años.

Conglomerado 3

```
group: 3
         var   n  mean    sd median trimmed   mad min max range   se
TOT_ATM    1 144 52.31 21.22     53   53.27 20.02   0 103   103 1.77
TOT_BLC    2 144  8.35  4.08      9    8.50  4.45   0  16    16 0.34
TOT_CEL    3 144 15.45 31.96      0    7.41  0.00   0 137   137 2.66
TOT_ITR    4 144 14.21 24.94      0    8.14  0.00   0 108   108 2.08
TOT_KIO    5 144  0.33  2.02      0    0.00  0.00   0  22    22 0.17
TOT_TLN    6 144 13.15 14.69      8   11.10 11.86   0  51    51 1.22
TOT_VTL    7 144 30.18 17.25     29   29.31 16.31   0  92    92 1.44
```

Figura 5.60: Centro de conglomerado 3 C-medias

En la figura 5.60 se observa que los clientes en este grupo, realizan un mayor volumen de transacciones en el canal ATM, Internet, Celular, Telenexo y Ventanilla .

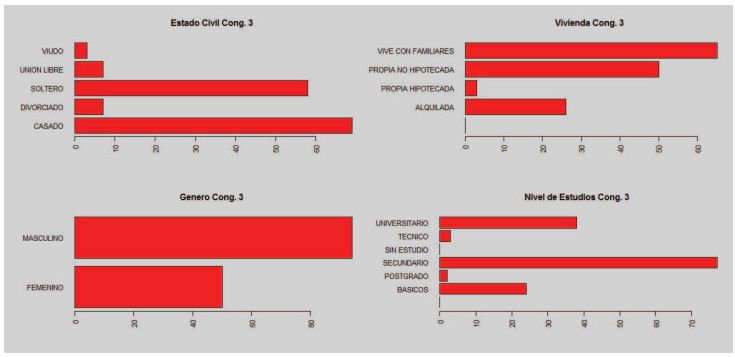


Figura 5.61: Variable Demográficas conglomerado 3 C-medias

En la figura 5.61 se visualiza que los clientes en su mayoría tienen un estado civil Soltero y Casado, tiene como vivienda la categoría Vive con Familiares y propia no Hipotecada, son Hombres y en nivel de estudio predomina la categoría Secundario.

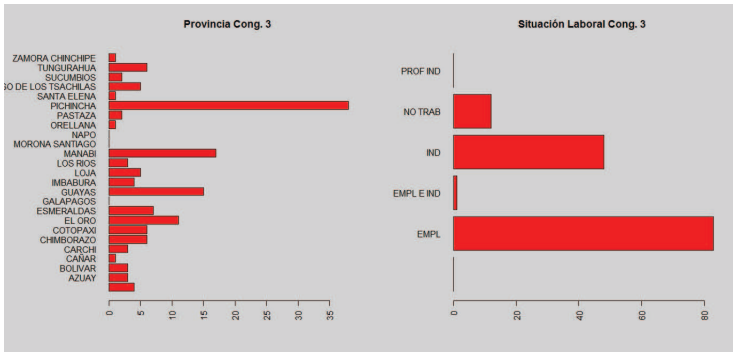


Figura 5.62: Variable Provincia y Situación Laboral conglomerado 3 C-medias

La provincia en este grupo se encuentra representada por Pichincha y Guayas; y la situación laboral por la categoria Empleado.

group: 3

	var	n	mean	sd	median	trimmed	mad	min	max	range	se
NUM_CRG	1	144	1.03	1.06	1.00	0.91	1.48	0.00	4.00	4.00	0.09
EDA_FIN	2	144	38.94	11.54	36.87	38.13	12.17	19.86	73.33	53.47	0.96

Figura 5.63: Variable Edad y Cargas familiares conglomerado 3 C-medias

En la figura 5.63 se observa que en promedio el grupo tiene 1 carga familiar y una edad promedio de 38 años.

Conglomerado 4

group: 4

	var	n	mean	sd	median	trimmed	mad	min	max	range	se
TOT_ATM	1	2515	4.11	7.89	0	2.10	0.00	0	50	50	0.16
TOT_BLC	2	2515	0.58	1.03	0	0.33	0.00	0	4	4	0.02
TOT_CEL	3	2515	0.45	4.58	0	0.00	0.00	0	115	115	0.09
TOT_ITR	4	2515	7.39	32.04	0	0.38	0.00	0	407	407	0.64
TOT_KIO	5	2515	0.16	4.04	0	0.00	0.00	0	189	189	0.08
TOT_TLN	6	2515	1.02	3.22	0	0.17	0.00	0	35	35	0.06
TOT_VTL	7	2515	6.39	7.32	3	5.19	4.45	0	31	31	0.15

Figura 5.64: Centro de conglomerado 4 C-medias

En la figura 5.64 se visualiza que los clientes en este grupo, tiende a no transaccionar por los canales.

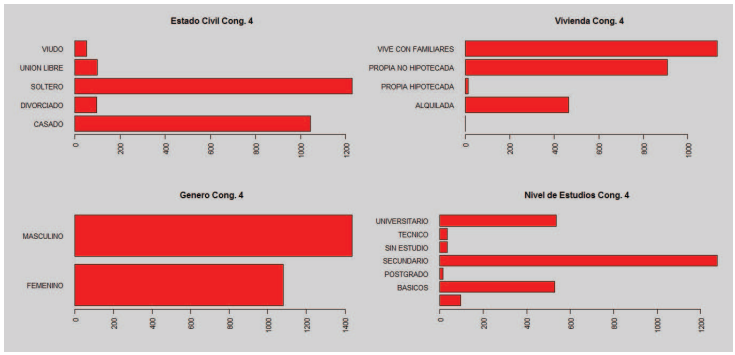


Figura 5.65: Variable Demográficas conglomerado 4 C-medias

En la figura 5.65 se observa que los clientes en su mayoría tienen un estado civil Soltero y Casado, tiene como vivienda la categoría Vive con Familiares y propia no Hipotecada, son Hombres y Mujer; y en nivel de estudio predomina la categoría Secundario.

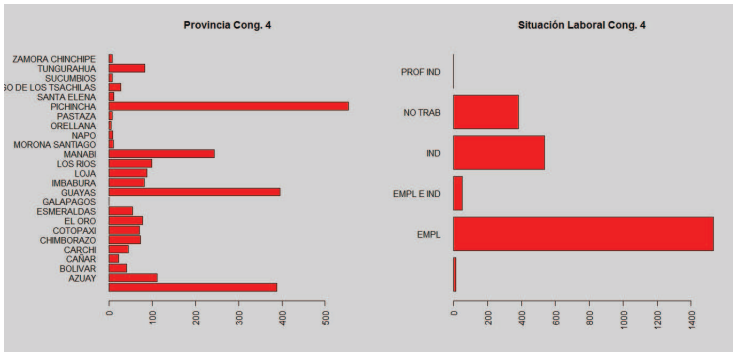


Figura 5.66: Variable Provincia conglomerado 4 C-medias

La provincia en este grupo se encuentra representada por Pichincha y Guayas; y la situación laboral por la categoria Empleado.

```
group: 4
        var    n  mean    sd median trimmed   mad  min   max  range   se
NUM_CRG   1 2515  0.88  1.27   0.00    0.65  0.00 0.00  10.0  10.00 0.03
EDA_FIN   2 2515 41.77 15.59  38.62   40.27 14.19 1.99 111.5 109.51 0.31
```

Figura 5.67: Variable Edad y Cargas familiares conglomerado 4 C-medias

En la figura 5.67 se observa que en promedio el grupo tiene 1 carga familiar y una edad promedio de 41 años.

Conglomerado 5

group: 5

	var	n	mean	sd	median	trimmed	mad	min	max	range	se
TOT_ATM	1	1081	12.56	26.32	0	5.76	0.00	0	233	233	0.80
TOT_BLC	2	1081	2.92	4.38	1	2.00	1.48	0	40	40	0.13
TOT_CEL	3	1081	2.07	13.45	0	0.00	0.00	0	195	195	0.41
TOT_ITR	4	1081	23.38	88.50	0	1.61	0.00	0	1004	1004	2.69
TOT_KIO	5	1081	0.40	3.77	0	0.00	0.00	0	98	98	0.11
TOT_TLN	6	1081	1.82	6.78	0	0.21	0.00	0	86	86	0.21
TOT_VTL	7	1081	67.74	41.15	54	60.09	25.20	28	319	291	1.25

Figura 5.68: Centro de conglomerado 5 C-medias

En la figura 5.68 se observa que los clientes en este grupo, realizan un mayor volumen de transacciones en el canal Internet y Ventanilla .

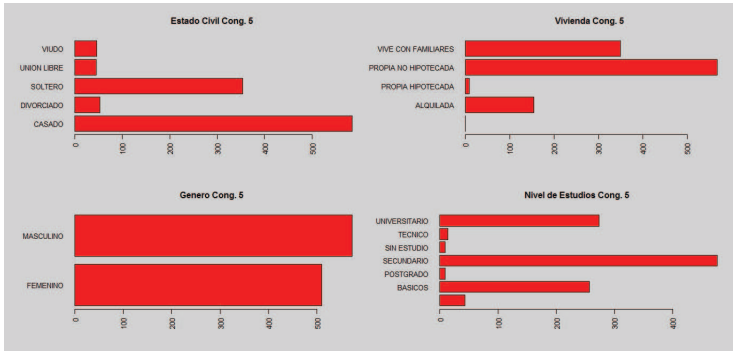


Figura 5.69: Variable Demográficas conglomerado 5 C-medias

En la figura 5.69 se observa que los clientes en su mayoría tienen un estado civil de Casado, tiene como vivienda la categoría propia no Hipotecada, son Hombres y Mujer; y en nivel de estudio predomina la categoría Secundario.

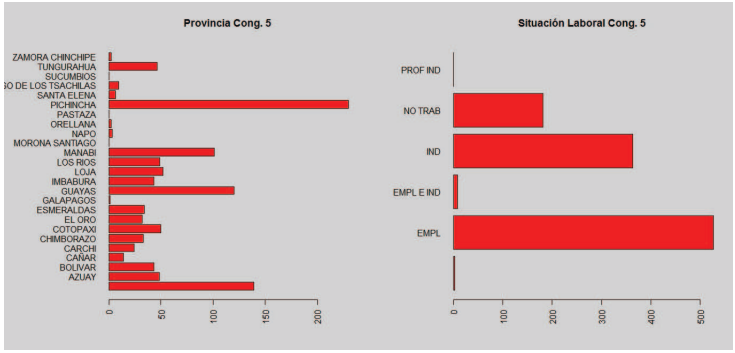


Figura 5.70: Variable Provincia y Situación Laboral conglomerado 5 C-medias

La provincia en este grupo se encuentra representada por Pichincha y Guayas; y la situación laboral por la categoría Empleado e Independiente.

```
group: 5
         var    n  mean    sd median trimmed   mad  min   max range   se
NUM_CRG    1 1081  1.20  1.44   1.00    0.99  1.48 0.00  9.00  9.00 0.04
EDA_FIN    2 1081 50.86 15.33  49.61   50.45 15.85 6.43 98.92 92.49 0.47
```

Figura 5.71: Variable Edad y Cargas familiares conglomerado 5 C-medias

En la figura 5.67 se visualiza que en promedio el grupo tiene 1 carga familiar y una edad promedio de 50 años.

Conglomerado 6

```
group: 6
         var   n  mean    sd median trimmed   mad min max range   se
TOT_ATM    1 954 56.42 15.20     56   56.15 17.79  26 104    78 0.49
TOT_BLC    2 954  1.79  2.39      1    1.32  1.48   0  13    13 0.08
TOT_CEL    3 954  1.21  6.56      0    0.00  0.00   0  90    90 0.21
TOT_ITR    4 954 17.58 40.21      0    6.91  0.00   0 260   260 1.30
TOT_KIO    5 954  0.38  3.23      0    0.00  0.00   0  55    55 0.10
TOT_TLN    6 954  0.97  3.04      0    0.08  0.00   0  28    28 0.10
TOT_VTL    7 954 14.29 11.52     12   13.19 11.86   0  58    58 0.37
```

Figura 5.72: Centro de conglomerado 6 C-medias

En la figura 5.72 se observa que los clientes en este grupo, realizan un mayor volumen de transacciones en el canal ATM.

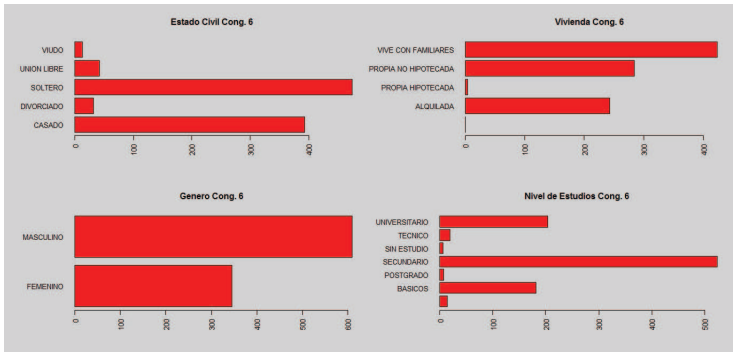


Figura 5.73: Variable Demográficas conglomerado 6 C-medias

En la figura 5.48 se visualiza que los clientes en su mayoría tienen un estado civil Soltero y Casado, tiene como vivienda la categoría Vive con Familiares y propia no Hipotecada, son Hombres y en nivel de estudio predomina la categoría Secundario.

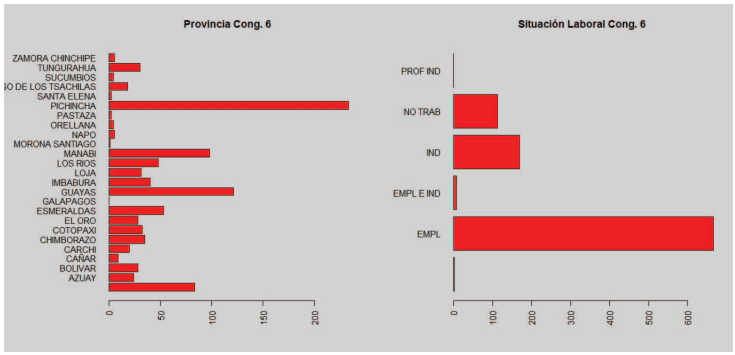


Figura 5.74: Variable Provincia y Situación Laboral conglomerado 6 C-medias

La provincia en este grupo se encuentra representada por Pichincha; y la situación laboral por la categoría Empleado.

```
group: 6
        var   n  mean    sd median trimmed  mad  min  max range   se
NUM_CRG   1 954  0.96  1.30   0.00    0.74  0.0 0.00 11.0 11.00 0.04
EDA_FIN   2 954 37.77 13.14  34.43   36.18 11.4 7.24 89.2 81.96 0.43
```

Figura 5.75: Variable Edad y Cargas familiares conglomerado 6 C-medias

En la figura 5.75 se observa que en promedio el grupo tiene 1 carga familiar y una edad promedio de 37 años.

Conglomerado 7

```
group: 7
        var   n  mean    sd median trimmed   mad min max range   se
TOT_ATM   1 810 25.13 31.81     11   19.20 16.31   0 177   177 1.12
TOT_BLC   2 810 20.60  7.67     20   19.49  5.93  11  67    56 0.27
TOT_CEL   3 810  2.06 11.38      0    0.00  0.00   0 164   164 0.40
TOT_ITR   4 810 18.43 64.57      0    3.90  0.00   0 823   823 2.27
TOT_KIO   5 810  0.29  2.72      0    0.00  0.00   0  57    57 0.10
TOT_TLN   6 810  2.19  6.62      0    0.47  0.00   0  76    76 0.23
TOT_VTL   7 810 13.89 21.18      6    9.43  7.41   0 200   200 0.74
```

Figura 5.76: Centro de conglomerado 7 C-medias

En la figura 5.76 se visualiza que los clientes en este grupo, realizan un mayor volumen de transacciones en el canal Balcones.

En la figura 5.77 se observa que los clientes en su mayoría tienen un estado civil Soltero y Casado, tiene como vivienda la categoría Vive con Familiares y propia no Hipotecada, son Hombres y Mujer; y en nivel de estudio predomina la categoría Secundario.

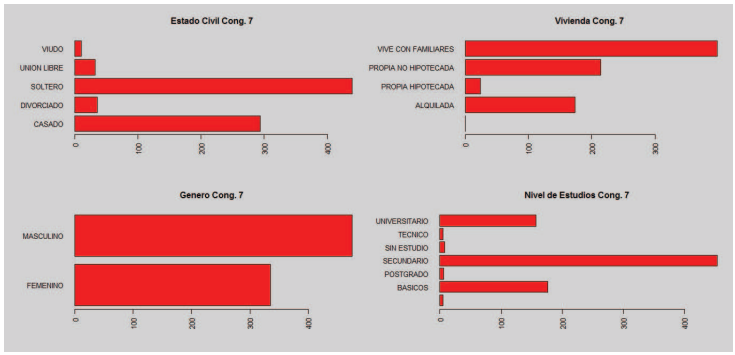


Figura 5.77: Variable Demográficas conglomerado 7 C-medias

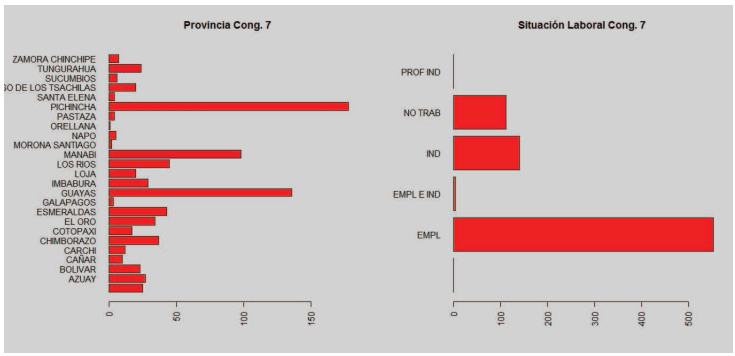


Figura 5.78: Variable Provincia conglomerado 7 C-medias

La provincia en este grupo se encuentra representada por Pichincha y Guayas; y la situación laboral por la categoria Empleado.

group: 7

	var	n	mean	sd	median	trimmed	mad	min	max	range	se
NUM_CRG	1	810	0.81	1.25	0.00	0.56	0.00	0.00	7.00	7.0	0.04
EDA_FIN	2	810	35.52	13.32	32.53	34.41	12.67	1.06	82.86	81.8	0.47

Figura 5.79: Variable Edad y Cargas familiares conglomerado 7 C-medias

En la figura 5.79 se observa que en promedio el grupo tiene 1 carga familiar y una edad promedio de 35 años.

Conglomerado 8

En la figura 5.80 se visualiza que los clientes en este grupo, realizan un mayor volumen de transacciones en el canal ATM, Celular, kiosko, Internet, ventanilla y Telenexo .

```
group: 8
          var   n   mean     sd median trimmed    mad min  max range   se
TOT_ATM     1 580  65.96  44.13   63.0   63.58  43.74   0  303   303 1.83
TOT_BLC     2 580   6.10   5.51    5.0    5.41   5.93   0   29    29 0.23
TOT_CEL     3 580  19.57  46.90    0.0    6.39   0.00   0  240   240 1.95
TOT_ITR     4 580 304.99 245.62  250.5  279.98 235.73   0 1015  1015 10.20
TOT_KIO     5 580   3.40  26.89    0.0    0.11   0.00   0  483   483 1.12
TOT_TLN     6 580  17.85  32.60    4.0    9.89   5.93   0  355   355 1.35
TOT_VTL     7 580  38.09  31.35   32.0   33.67  25.20   0  195   195 1.30
```

Figura 5.80: Centro de conglomerado 8 C-medias

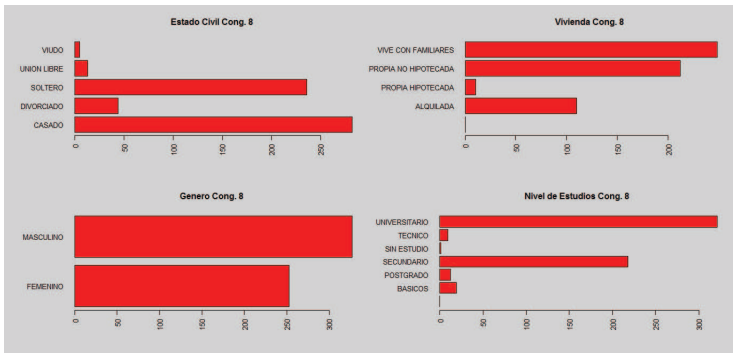


Figura 5.81: Variable Demográficas conglomerado 8 C-medias

En la figura 5.81 se observa que los clientes en su mayoría tienen un estado civil Soltero y Casado, tiene como vivienda la categoría Vive con Familiares y propia no Hipotecada, son Hombres y Mujer; y en nivel de estudio predomina la categoría Universitario.

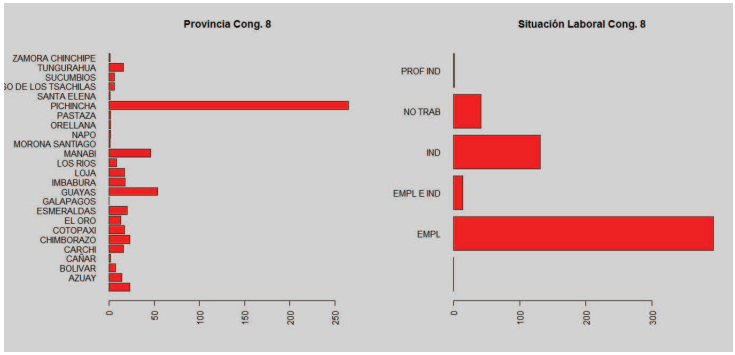


Figura 5.82: Variable Provincia y Situación Laboral conglomerado 8 C-medias

La provincia en este grupo se encuentra representada por Pichincha; y la situación laboral por la categoria Empleado.

En la figura 5.83 se visualiza que en promedio el grupo tiene 1 carga familiar y una edad promedio de 39 años.

```
group: 8
         var   n mean    sd median trimmed   mad   min   max range   se
NUM_CRG    1 580  0.9  1.07   0.00    0.73  0.00  0.00  4.00  4.00 0.04
EDA_FIN    2 580 39.3 12.19  36.67   38.10 11.78 19.98 92.07 72.09 0.51
```

Figura 5.83: Variable Edad y Cargas familiares conglomerado 8 C-medias

Migración Clientes

Con respecto a la estrategia que se quiera aplicar para la migración de clientes de un canal hacia otro, el método de C - medias nos presenta el grado de membrecía, el cual nos permite tener una medida de cuál es la posibilidad que un cliente pertenezca a otro grupo, para esto se va elegir los dos valores máximo del grado de membrecía de cada individuo y así conocer a cuál conglomerado pueda migrar. aplicando este método se obtiene la figura 5.84 donde se muestra la distribución de frecuencia de transición .

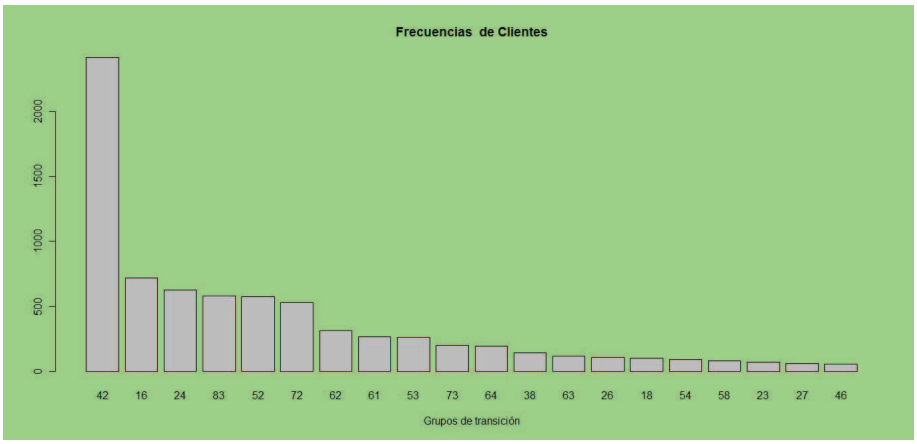


Figura 5.84: Transición de clientes

Donde por ejemplo, el valor 42 hace referencia a que el individuo pertenece al conglomerado 4 y podría trasladarse al conglomerado 2. Observando las características del grupo 4 y el grupo 2 se observa que la transición toma la dirección en aumento de transaccionalidad haciendo a los clientes mejor su relación con la entidad Bancaria. Para el valor 61 puede observarse que uno de los beneficios es que el conglomerado 1 realiza mayor transacción en el internet siendo esto beneficioso para la Entidad Bancaria.

Árbol de decisión

Después de haber determinado los perfiles de clientes mediante su comportamiento en los canales transaccionales, el siguiente objetivo a determinar es si existe una relación entre estos perfiles y sus características demográficas, para esto se ha

realizado un análisis de semejanza entre los grupos obtenidos, en la tabla 5.85 se observa la macro agrupación por contacto del cliente:

CONGLOMERADO	OBS-	GRUPO	TOT.OBS	CONTACTO
1	871	1	1825	BAJO
6	954	1		
5	1081	5	1081	ALTO
3	144	3	724	MULTICANAL
8	580	3		
4	2515	4	2515	NO TRANSACCIONA
2	921	2	1731	ALTO
7	810	2		

Figura 5.85: Comparación entre grupos

Tomando los macro grupos de dos en dos, se aplicó la técnica de árbol de decisión utilizando el método ctree del software R, el criterio de parada utilizado para el análisis fue que los nodos no tengan menos de 300 observaciones, que representa 20% de la base, además para el macro grupo 3 se utilizó la técnica de muestreo, con la finalidad de aumentar el número de observaciones a 1500, con estas condiciones se obtuvieron los siguientes resultados:

Macro grupo 1 Contacto Bajo y 2 Contacto Alto

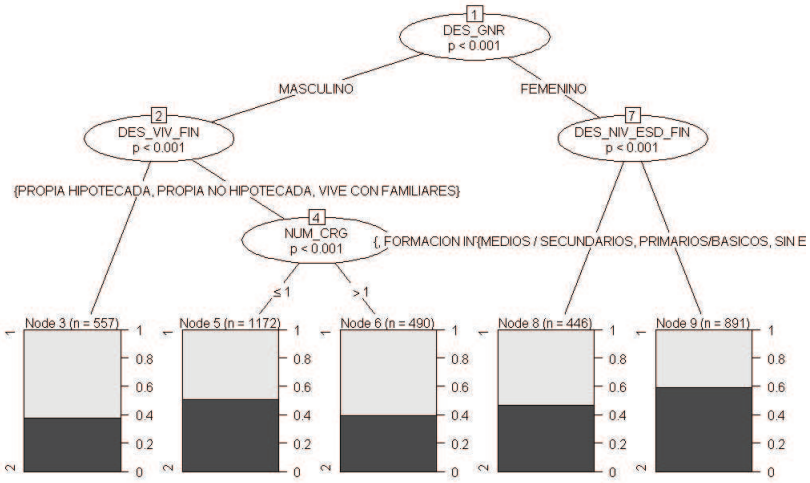


Figura 5.86: Árbol de decisión grupo 1 vs 2

De acuerdo a la figura 5.86, el método elige como primera variable al genero , con un p-valor < 0.001, y divide al nodo inicial en 2 nodos hijos 2 y 7, las observaciones con genero masculino se encuentran en el nodo 2 y las observaciones con genero femenino se encuentran en el nodo 7.

Al analizar el segundo nivel, la variable vivienda divide el nodo 2 en el nodo terminal 3 por la categoria Alquilada, y en el nodo 4 por el resto de categorias, el nodo terminal posee el 60% del macro grupo 1 contacto Bajo. Por otro lado, la

variable nivel de estudio divide al nodo 7 en dos nodos terminales, el nodo 8 con el 55% de la categoria 1 contacto bajo, mediante las categorias Intermedia, técnica, posgrado y Universitaria y el nodo 9 con el 60% del macro grupo 2 contacto alto, por la categoria Secundario, Primarios y sin estudios.

```
> chaid

        Conditional inference tree with 5 terminal nodes

Response:  as.character.mydata.grupo.
Inputs:  EDA_FIN, DES_NIV_ESD_FIN, DES_GNR, DES_SIT_LAB_FIN, DES_VIV_FIN, NUM_CRG
Number of observations:  3556

1) DES_GNR == {MASCULINO}; criterion = 1, statistic = 33.981
  2) DES_VIV_FIN == {ALQUILADA}; criterion = 1, statistic = 22.999
    3)*  weights = 557
  2) DES_VIV_FIN == {PROPIA HIPOTECADA, PROPIA NO HIPOTECADA, VIVE CON FAMILIARES}
    4) NUM_CRG <= 1; criterion = 1, statistic = 15.846
      5)*  weights = 1172
    4) NUM_CRG > 1
      6)*  weights = 490
1) DES_GNR == {FEMENINO}
  7) DES_NIV_ESD_FIN == {, FORMACION INTERMEDIA O TECNICA, POSTGRADO, UNIVERSITARIOS}; criterion = 1, statistic = 29.044
    8)*  weights = 446
  7) DES_NIV_ESD_FIN == {MEDIOS / SECUNDARIOS, PRIMARIOS/BASICOS, SIN ESTUDIOS}
    9)*  weights = 891
```

Figura 5.87: Estadístico del árbol de decisión Macro grupo 1 vs 2

Macro grupo 1 Contacto Bajo y 3 Multicanal

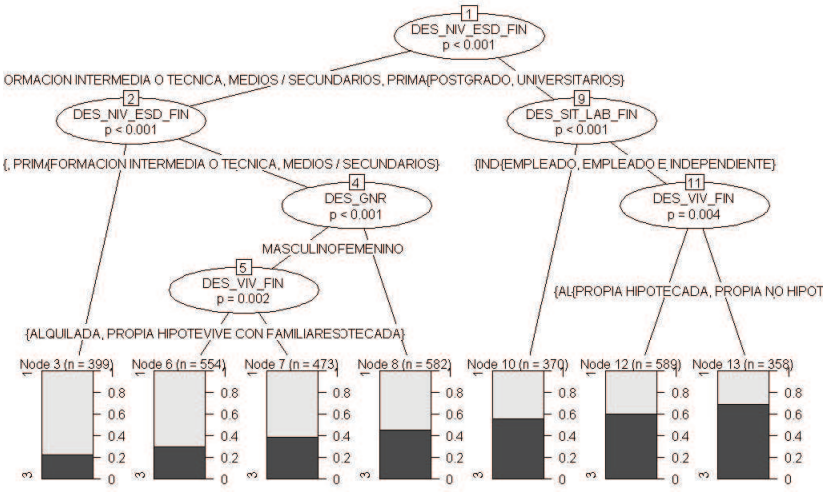


Figura 5.88: Árbol de decisión Macro grupo 1 vs 3

De acuerdo a la figura 5.88, el método elige como primera variable al nivel de estudio, con un p-valor < 0.001, y divide al nodo inicial en 2 nodos hijos 2 y 9, las observaciones con nivel de estudios sin estudios, básica, secundaria e intermedia se encuentran en el nodo 2 y las observaciones con nivel de estudios Posgrado y universitario se encuentran en el nodo 9.

Al analizar el segundo nivel, la variable nivel de estudios divide el nodo 2 en el nodo terminal 3 por las categorías Básico y Sin estudios, y en el nodo 4 por el resto de categorias, el nodo terminal posee el 80% del macro grupo 1 contacto Bajo. Por otro lado, la variable situación laboral divide al nodo 9 en dos nodos, el nodo terminal 10 con el 55% de la categoria 3 Multicanal, mediante las categorias Independiente y no trabajan; y el nodo 11 con el resto de categorías.

Al analizar el tercer nivel, el nodo 4 es dividido por la variable género formando el nodo terminal 8 con 55% de la categoría 1 contacto bajo con el valor femenino y el nodo 5 con el valor Masculino. El nodo 11 es dividido por la variable vivienda en dos nodos terminales 12 y 13 con el 60% y 65% de la categoría 3 Multicanal respectivamente.

En el cuarto nivel el nodo 5 es dividido por la variable vivienda en dos nodos terminales 6 y 7 con 75% y 60% de la categoría 1 contacto bajo.

```
Response:  as.character.mydata.grupo.
Inputs:  EDA_FIN, DES_NIV_ESD_FIN, DES_GNR, DES_SIT_LAB_FIN, DES_VIV_FIN, NUM_CRG
Number of observations:  3325

1) DES_NIV_ESD_FIN == {, FORMACION INTERMEDIA O TECNICA, MEDIOS / SECUNDARIOS, PRIMARIOS/BASICOS, SIN ESTUDIOS}; criterion $
  2) DES_NIV_ESD_FIN == {, PRIMARIOS/BASICOS, SIN ESTUDIOS}; criterion = 1, statistic = 38.835
    3)*  weights = 399
  2) DES_NIV_ESD_FIN == {FORMACION INTERMEDIA O TECNICA, MEDIOS / SECUNDARIOS}
    4) DES_GNR == {MASCULINO}; criterion = 1, statistic = 22.896
      5) DES_VIV_FIN == {ALQUILADA, PROPIA HIPOTECADA, PROPIA NO HIPOTECADA}; criterion = 0.998, statistic = 19.04
        6)*  weights = 554
      5) DES_VIV_FIN == {VIVE CON FAMILIARES}
        7)*  weights = 473
    4) DES_GNR == {FEMENINO}
      8)*  weights = 582
1) DES_NIV_ESD_FIN == {POSTGRADO, UNIVERSITARIOS}
  9) DES_SIT_LAB_FIN == {INDEPENDIENTE, NO TRABAJA}; criterion = 1, statistic = 34.91
    10)*  weights = 370
  9) DES_SIT_LAB_FIN == {EMPLEADO, EMPLEADO E INDEPENDIENTE}
    11) DES_VIV_FIN == {ALQUILADA, VIVE CON FAMILIARES}; criterion = 0.996, statistic = 17.014
      12)*  weights = 589
    11) DES_VIV_FIN == {PROPIA HIPOTECADA, PROPIA NO HIPOTECADA}
      13)*  weights = 358
```

Figura 5.89: Estadístico del árbol de decisión Macro grupo 1 vs 3

Macro grupo 1 Contacto Bajo y 5 Contacto Alto

De acuerdo a la figura 5.90, el método elige como primera variable a la edad, con un p-valor < 0.001, y divide al nodo inicial en 2 nodos hijos 2 y 9, las observaciones con edad $<= 41,66$ años se encuentran en el nodo 2 y las observaciones con edad $> 41,66$ se encuentran en el nodo 9.

Al analizar el segundo nivel, la variable situación laboral divide el nodo 2 en el nodo terminal 8 por las categorías Independiente y no trabaja, y en el nodo 3 por el resto de categorías, el nodo terminal posee el 65% del macro grupo 1 Contacto Bajo. Por otro lado, la variable edad divide al nodo 9 en dos nodos, el nodo terminal 13 con el 75% de la categoría 5 Contacto Alto, con edad $> 62,15$ años; y el nodo 10 con edad $<= 62,15$.

Al analizar el tercer nivel, el nodo 3 es dividido por la variable género formando el nodo terminal 4 con 80% de la categoría 1 contacto bajo con el valor femenino y el nodo 5 con el valor Masculino. El nodo 10 es dividido por la variable género en

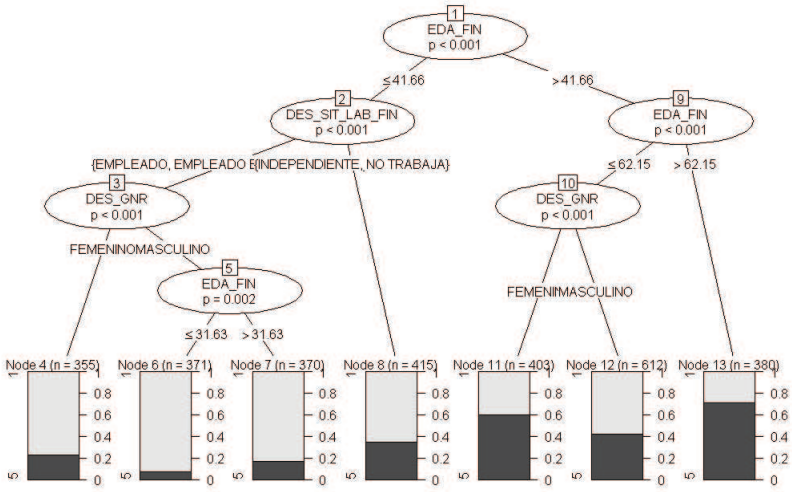


Figura 5.90: Árbol de decisión Macro grupo 1 vs 5

dos nodos terminales 11 y 12 con el 60% de la categoría 5 Contacto Alto y 60% de la categoría 1 Contacto Bajo respectivamente.

En el cuarto nivel el nodo 5 es dividido por la variable edad en dos nodos terminales 6 y 7 con 95% y 80% de la categoría 1 Contacto Bajo.

```
Response:  as.character.mydata.grupo.
Inputs:  EDA_FIN, DES_NIV_ESD_FIN, DES_GNR, DES_SIT_LAB_FIN, DES_VIV_FIN, NUM_CRG
Number of observations:  2906

1) EDA_FIN <= 41.66; criterion = 1, statistic = 438.612
  2) DES_SIT_LAB_FIN == (EMPLEADO, EMPLEADO E INDEPENDIENTE); criterion = 1, statistic = 91.327
    3) DES_GNR == (FEMENINO); criterion = 1, statistic = 22.666
      4)*  weights = 355
    3) DES_GNR == (MASCULINO)
      5) EDA_FIN <= 31.63; criterion = 0.998, statistic = 22.454
        6)*  weights = 371
      5) EDA_FIN > 31.63
        7)*  weights = 370
  2) DES_SIT_LAB_FIN == (INDEPENDIENTE, NO TRABAJA)
    8)*  weights = 415
1) EDA_FIN > 41.66
  9) EDA_FIN <= 62.15; criterion = 1, statistic = 53.326
    10) DES_GNR == (FEMENINO); criterion = 1, statistic = 30.527
      11)*  weights = 403
    10) DES_GNR == (MASCULINO)
      12)*  weights = 612
  9) EDA_FIN > 62.15
    13)*  weights = 380
```

Figura 5.91: Estadístico del árbol de decisión Macro grupo 1 vs 5

Macro grupo 2 Contacto Bajo y 3 Multicanal

De acuerdo a la figura 5.92, el método elige como primera variable al nivel de estudios, con un p-valor < 0.001, y divide al nodo inicial en 2 nodos hijos 2 y 7, las observaciones con categoría Sin estudio,Primaria, Básica, Secundario y Técnica se encuentran en el nodo 2 y las observaciones con categoría Postgrado y Universitario se encuentran en el nodo 7.

Al analizar el segundo nivel, la variable nivel de estudio divide el nodo 2 en el nodo terminal 6 por las categorías Primario y sin estudios, y en el nodo 3 por el resto de categorias, el nodo terminal posee el 80% del macro grupo 2 Contacto

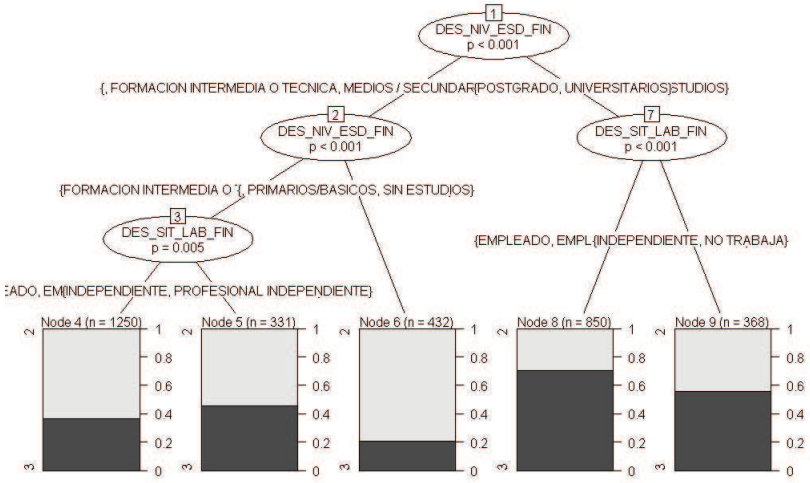


Figura 5.92: Árbol de decisión Macro grupo 2 vs 3

Bajo. Por otro lado, la variable edad divide al nodo 7 en dos nodos terminales, el nodo terminal 8 con el 75% de la categoria 3 Multicanal, con la categoría Empleado y Empleado e Independiente; y el nodo 9 con el 55% de la categoría 3 Multicanal con la categoría Independiente y no trabaja.

Al analizar el tercer nivel, el nodo 3 es dividido por la variable situación laboral formando el nodo terminal 4 con el 60% de la categoría 2 contacto bajo con el valor Empleado, Empleado e Independiente y no trabaja; y el nodo terminal 5 con el 55% de la categoría 2 contacto bajo con el valor Independiente y profesional independiente.

```
        Conditional inference tree with 5 terminal nodes

Response:  as.character.mydata.grupo.
Inputs:  EDA_FIN, DES_NIV_ESD_FIN, DES_GNR, DES_SIT_LAB_FIN, DES_VIV_FIN, NUM_CRG
Number of observations:  3231

1) DES_NIV_ESD_FIN == (, FORMACION INTERMEDIA O TECNICA, MEDIOS / SECUNDARIOS, PRIMARIOS/BASICOS, SIN ESTUDIOS); criterion $
  2) DES_NIV_ESD_FIN == (FORMACION INTERMEDIA O TECNICA, MEDIOS / SECUNDARIOS); criterion = 1, statistic = 51.266
    3) DES_SIT_LAB_FIN == (, EMPLEADO, EMPLEADO E INDEPENDIENTE, NO TRABAJA); criterion = 0.995, statistic = 21.065
      4)*  weights = 1250
    3) DES_SIT_LAB_FIN == (INDEPENDIENTE, PROFESIONAL INDEPENDIENTE)
      5)*  weights = 331
  2) DES_NIV_ESD_FIN == (, PRIMARIOS/BASICOS, SIN ESTUDIOS)
    6)*  weights = 432
1) DES_NIV_ESD_FIN == (POSTGRADO, UNIVERSITARIOS)
  7) DES_SIT_LAB_FIN == (EMPLEADO, EMPLEADO E INDEPENDIENTE); criterion = 1, statistic = 54.092
    8)*  weights = 850
  7) DES_SIT_LAB_FIN == (INDEPENDIENTE, NO TRABAJA)
    9)*  weights = 368
```

Figura 5.93: Estadístico del árbol de decisión Macro grupo 2 vs 3

Macro grupo 3 Multicanal y 4 No transaccionan

De acuerdo a la figura 5.94, el método elige como primera variable al nivel de estudios, con un p-valor < 0.001, y divide al nodo inicial en 2 nodos hijos 2 y 5, las observaciones con categoría Postgrado y Universitario se encuentran en el nodo 2 y las observaciones con categoría Sin estudio,Primaria, Básica, Secundario y Técnica se encuentran en el nodo 5.

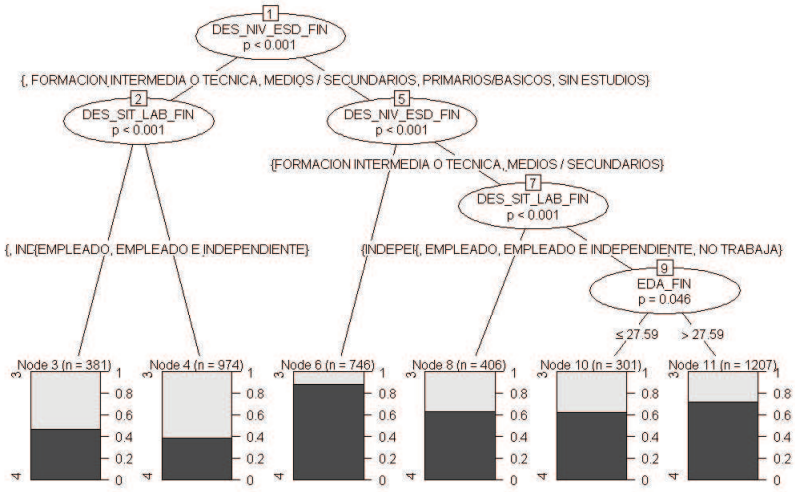


Figura 5.94: Árbol de decisión grupo 3 vs 4

Al analizar el segundo nivel, la variable situación laboral divide el nodo 2 en el nodo terminal 3 por las categorías Independiente y no trabaja y que tiene 55% de la categoría 3 Multicanal, y en el nodo terminal 4 por las categorías Empleado y Empleado e Independiente y que tiene 60% de la categoría 3 Multicanal. Con respecto al nodo 5 es dividido en dos nodos, el nodo terminal 6 por la categorias Primarios y sin estudios y tiene más del 80% de la categoría 4; y en el nodo 7 por la categoría Técnicas y Secundarias,

Al analizar el tercer nivel, el nodo 7 es dividido por la variable situación laboral formando el nodo terminal 8 con el 60% de la categoría 4 que no transacciona con el valor Independiente y profesional independiente; y el nodo 9 por los valor Empleado, Empleado e Independiente y No trabajo.

Al analizar el cuarto nivel, el nodo 9 es dividido en dos categorías por la variable edad, el nodo terminal 10 por el valor de edad $<= 27,59$ y que tiene el 60% de la categoría 4 no transacciona; y en nodo terminal 11 por el valor de edad $> 27,59$ y tiene el 70% de la categoria 4 no transacciona.

```
        Conditional inference tree with 6 terminal nodes

Response:  as.character.mydata.grupo.
Inputs:  EDA_FIN, DES_NIV_ESD_FIN, DES_GNR, DES_SIT_LAB_FIN, DES_VIV_FIN, NUM_CRG
Number of observations:  4015

1) DES_NIV_ESD_FIN == {POSTGRADO, UNIVERSITARIOS}; criterion = 1, statistic = 524.23
  2) DES_SIT_LAB_FIN == {, INDEPENDIENTE, NO TRABAJA}; criterion = 1, statistic = 25.464
    3)*  weights = 381
  2) DES_SIT_LAB_FIN == {EMPLEADO, EMPLEADO E INDEPENDIENTE}
    4)*  weights = 974
1) DES_NIV_ESD_FIN == {, FORMACION INTERMEDIA O TECNICA, MEDIOS / SECUNDARIOS, PRIMARIOS/BASICOS, SIN ESTUDIOS}
  5) DES_NIV_ESD_FIN == {, PRIMARIOS/BASICOS, SIN ESTUDIOS}; criterion = 1, statistic = 118.973
    6)*  weights = 746
  5) DES_NIV_ESD_FIN == {FORMACION INTERMEDIA O TECNICA, MEDIOS / SECUNDARIOS}
    7) DES_SIT_LAB_FIN == {INDEPENDIENTE, PROFESIONAL INDEPENDIENTE}; criterion = 1, statistic = 32.29
      8)*  weights = 406
    7) DES_SIT_LAB_FIN == {, EMPLEADO, EMPLEADO E INDEPENDIENTE, NO TRABAJA}
      9) EDA_FIN <= 27.59; criterion = 0.954, statistic = 23.804
        10)*  weights = 301
      9) EDA_FIN > 27.59
        11)*  weights = 1207
```

Figura 5.95: Estadístico del árbol de decisión grupo 3 vs 4

Macro grupo 3 Multicanal y 5 Contacto Alto

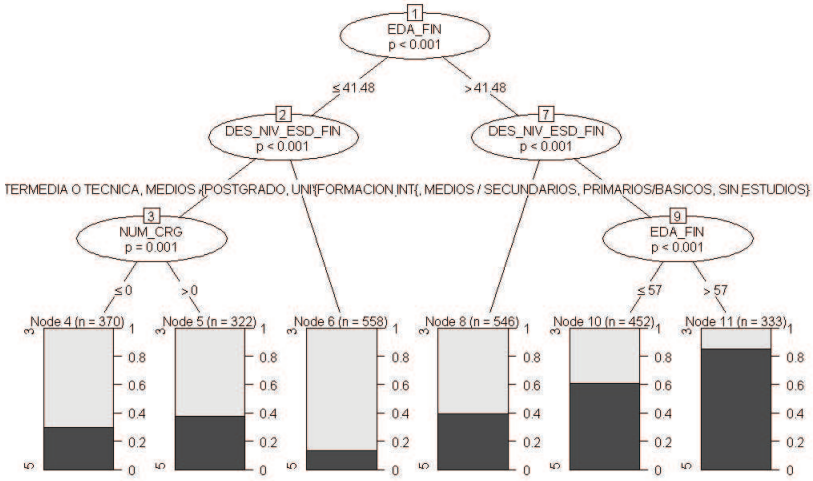


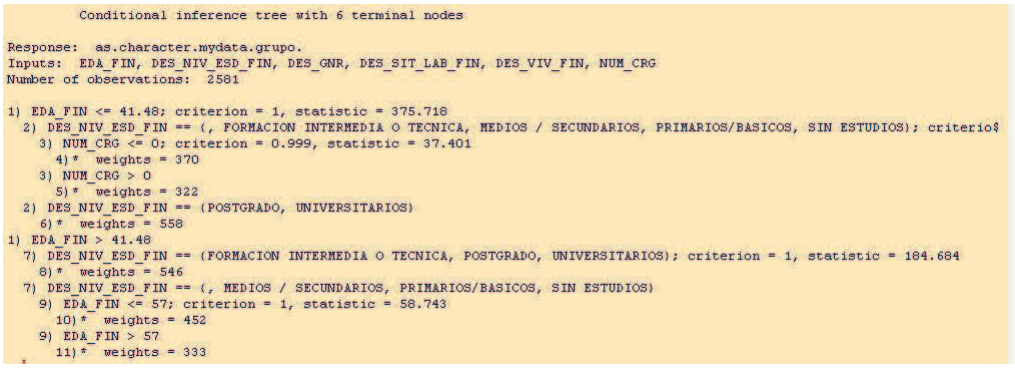
Figura 5.96: Árbol de decisión Macro grupo 3 vs 5

De acuerdo a la figura 5.96, el método elige como primera variable a la edad, con un p-valor < 0.001, y divide al nodo inicial en 2 nodos hijos 2 y 7, las observaciones con edad $<= 41.48$ se encuentran en el nodo 2 y las observaciones con edad > 41.48 se encuentran en el nodo 7.

Al analizar el segundo nivel, la variable nivel de estudios divide el nodo 2 en el nodo 3 por las categorías Sin estudio,Primaria, Básica, Secundario y Técnica; y en el nodo terminal 6 por las categorías Postgrado y Universitario y tiene 90% de la categoría 3 Multicanal. Con respecto al nodo 7 es dividido en dos nodos, el nodo terminal 8 por la categorias Técnica, Posgrado, Universitario y tiene más del 60% de la categoría 3 Multicanal; y en el nodo 9 por las categorías Primario, Secundario y Sin estudios.

Al analizar el tercer nivel, el nodo 3 es dividido por la variable número cargas

formando el nodo terminal 4 con el 75% de la categoría 3 Multicanal, y en el nodo terminal 5 con el 60% de la categoría 3 Multicanal. El nodo 9 es dividido por la variable edad en el nodo terminal 10 con el 60% de la categoría 5 contacto alto, y en el nodo terminal 11 con el 85% de la categoría Contacto alto.



```
        Conditional inference tree with 6 terminal nodes

Response:  as.character.mydata.grupo.
Inputs:  EDA_FIN, DES_NIV_ESD_FIN, DES_GNR, DES_SIT_LAB_FIN, DES_VIV_FIN, NUM_CRG
Number of observations:  2581

1) EDA_FIN <= 41.48; criterion = 1, statistic = 375.718
  2) DES_NIV_ESD_FIN == (, FORMACION INTERMEDIA O TECNICA, MEDIOS / SECUNDARIOS, PRIMARIOS/BASICOS, SIN ESTUDIOS); criterio$
    3) NUM_CRG <= 0; criterion = 0.999, statistic = 37.401
      4)*  weights = 370
    3) NUM_CRG > 0
      5)*  weights = 322
  2) DES_NIV_ESD_FIN == (POSTGRADO, UNIVERSITARIOS)
    6)*  weights = 558
1) EDA_FIN > 41.48
  7) DES_NIV_ESD_FIN == (FORMACION INTERMEDIA O TECNICA, POSTGRADO, UNIVERSITARIOS); criterion = 1, statistic = 184.684
    8)*  weights = 546
  7) DES_NIV_ESD_FIN == (, MEDIOS / SECUNDARIOS, PRIMARIOS/BASICOS, SIN ESTUDIOS)
    9) EDA_FIN <= 57; criterion = 1, statistic = 58.743
      10)*  weights = 452
    9) EDA_FIN > 57
      11)*  weights = 333
```

Figura 5.97: Estadístico del árbol de decisión Macro grupo 3 vs 5

De acuerdo a este análisis es clara la relación que existe entre el uso de canales y las caracteristicas demográficas de los clientes. Pues este resultado da una dirección a la entidad Bancaria con respecto a la factibilidad de migrar clientes a canales diferentes.

A continuación se presenta una tabla resumen de los modelos de árbol de decisión:

Comp. Macro Grupos	Nodo terminal	Características Demográficas	Num. de obs.	Porc. por Grupo			
1 Contacto Bajo y 2 Contacto Alto	3	Masculino/Vivienda Alquilada	557	1	60%	2	40%
	5	Masculino/Vivienda Propia Hipotecada y no Hipotecada,Vive con Familiares / cargas familiares < 1	1172	1	50%	2	50%
	6	Masculino/Vivienda Propia Hipotecada y no Hipotecada,Vive con Familiares / cargas familiares >1	490	1	60%	2	40%
	8	Femenino/Formación Técnica, Postgrado y Universitarios	446	1	55%	2	45%
	9	Femenino/ Nivel Estudios Secundaria,Primaria y Sin estudios	891	1	60%	2	40%
1 Contacto Bajo y 3 Multicanal	3	Sin estudios y básica	399	1	80%	3	20%
	6	Formación Técnica y Secundaria/Masculino/Vivienda Alquilada, Propia Hipotecada y no Hipotecada	554	1	75%	3	25%
	7	Formación Técnica y Secundaria/Masculino/Vive con Familiares	473	1	60%	3	40%
	8	Formación Técnica y Secundaria/Femenino	582	1	55%	3	45%
	10	Estudio Universitario y Postgrado/Independiente y no trabaja	370	1	45%	3	55%
	12	Estudio Universitario y Postgrado/Independiente y no trabaja/ Vivienda Alquilada y vive con Familiares	589	1	40%	3	60%
	13	Estudio Universitario y Postgrado/Independiente y no trabaja/Vivienda Propia Hipotecada y no hipotecada	358	1	35%	3	65%
1 Contacto Bajo y 5 Contacto Alto	4	Edad <=41.66/Empleado y Empleado e Independiente/Femenino	355	1	80%	5	20%
	6	Edad <=31.63/Empleado y Empleado e Independiente/Masculino	371	1	95%	5	5%
	7	Edad entre 31.63 y 41.66/Empleado y Empleado e Independiente/Masculino	370	1	80%	5	20%
	8	Edad <=41.66/Independiente y no trabajan	415	1	65%	5	35%
	11	Edad entre 41.66 y 62.15/Femenino	403	1	40%	5	60%
	12	Edad entre 41.66 y 62.15/Masculino	612	1	60%	5	40%
	13	Edad >62.15	380	1	25%	5	75%
2 Contacto Alto y 3 Multicanal	4	Formación Técnica, Secundaria/Empleado,Empleado e Independiente y No trabaja	1250	2	60%	3	40%
	5	Básica,Sin estudios/Independiente, Profesional Independiente	331	2	55%	3	45%
	6	Básica,Sin estudios	432	2	80%	3	20%
	8	Estudios Universitario y Posgrado/Empleado,Empleado e Independiente	850	2	25%	3	75%
	9	Estudios Universitario y Posgrado/Independiente,No trabaja	368	2	45%	3	55%
3 Multicanal y 4 No transaccionan	3	Estudios Universitario y Posgrado/Independiente,No trabaja	381	3	55%	4	45%
	4	Estudios Universitario y Posgrado/Empleado,Empleado e Independiente	974	3	60%	4	40%
	6	Primarios y Sin estudios	746	3	10%	4	90%
	8	Nivel Estudios Secundaria,Técnica /Independiente , Profesional Independiente	406	3	40%	4	60%
	10	Nivel Estudios Secundaria,Técnica /Empleado, Empleado e Independiente/ edad <=27.59	301	3	40%	4	60%
	11	Nivel Estudios Secundaria,Técnica /Empleado, Empleado e Independiente/ edad > 27.60	1027	3	30%	4	70%
3 Multicanal y 5 Contacto Alto	4	Edad <=41.48/Formación Técnica,Secundaria,Básica,Sin estudio/Cargas familiares<=0	370	3	75%	5	25%
	5	Edad <=41.48/Formación Técnica,Secundaria,Básica,Sin estudio/Cargas familiares>0	322	3	60%	5	40%
	6	Edad <=41.48/Estudios Universitario y Posgrado	558	3	90%	5	10%
	8	Edad > 41.48/Estudios Universitario,Posgrado y Formación Técnica	546	3	60%	5	40%
	10	Edad entre 41.48 y 57 /Secundaria,Básica,Sin estudio	452	3	40%	5	60%
	11	Edad > 57 /Secundaria,Básica,Sin estudio	333	3	15%	5	85%

Figura 5.98: Resumen Árbol de Decisión

5.2 Conclusiones y Recomendaciones

5.2.1 Conclusiones

1. Las Entidades Bancarias miran a los canales transaccionales y de servicios como el principal vínculo para relacionarse con sus clientes. Para las entidades los canales se han convertido en oportunidades para captar, satisfacer y conocer a sus clientes.

2. El Marketing dentro de una Entidad Bancaria, es el encargado de generar las estrategias necesarias para llevar de manera adecuada y sostenible la relación que vincula a los clientes hacia los productos y servicios. El CRM como parte del Marketing, tiene como meta el mejorar los procesos de comunicación hacia el cliente correcto, proveyéndolo del producto o servicio correcto, a través del canal correcto y en el tiempo correcto.

3. El CRM analítico es el encargado de analizar los datos recolectada por una Entidad, con la finalidad de detectar patrones presentados en la información y así poder optimizar el relacionamiento con el cliente.

4. La Minería de datos es una herramienta tecnológica de manejo de información, que permite la extracción de patrones relevante proveniente de grandes bases de datos, mediante la aplicación de modelos estadísticos, con la finalidad de proveer resultados que apoyen a la toma de decisiones de la Entidad.

5. EL proceso de Minería de datos cumple con los siguientes pasos:

   - Entendimiento del Negocio.
   - Entendimiento de los datos.
   - Extracción de los datos.
   - Exploración de los datos.
   - Transformación de los datos.
   - Modelamiento de los datos.
   - Evaluación de los Modelamientos de los datos.
   - Puntuación de los datos.

6. La segmentación de Mercados se vincula a los métodos de conglomerados, porque mediante estos, se determinan los grupos o segmentos que describen de manera explícitamente al Mercado.

7. Los métodos de análisis de datos atípicos, que utilizan algoritmos de agrupamiento para su detección, combinan las características principales de los métodos de detección univariante y bivariante.

8. Una parte importante antes de empezar un modelo matemático, es el análisis de calidad de los datos, y como se observó en este proyecto, las medidas de estadística descriptiva y frecuencias, nos permiten detectar facilmente la información errónea en los datos, y así levantar alertas para la depuración y mejoramiento de la información.

9. El análisis de datos atípicos, determinó un grupo de clientes, cuyo comportamiento transaccional, refleja un alto volumen de transacciones en todos los canales. De esto se puede concluir que en la Entidad Bancaria, el uso de todos los canales por un mismo clientes es mínima.

10. La aplicación de los componentes principales a las variables transaccionales, nos proporcionó una mejor distribución de los datos en los nuevos ejes (componentes), esto se vió reflejado en que se pudieron distinguir 4 grupos de variables transaccionales $\{Balcones, Telenexo\}$, $\{Ventanilla, Internet\}$, $\{ATM\}$ y $\{Kiosko, Celular\}$.

11. El análisis de conglomerados se dividido en dos pasos fundamentales:

    - **Elección de una medida de proximidad**.- Una medida de similitud o proximidad es definida para medir la estrechez entre los objetos, mientras más estrechos se encuentren los objetos, más homogéneo es el grupo.
    - **Elección de un algoritmo de agrupación**.- Sobre la elección de la medida de proximidad, los objetos son asignados a grupos, hasta que la diferencia entre grupos sea lo más grande posible y las observaciones dentro de cada grupo sean lo más homogéneas.

12. Una de los principales parámetros a detectar, es el número de grupos inmersos en los datos , como se observó existen varios métodos para la determinación el valor del mismo, y éstos dependen del algoritmo de conglomerados que se aplica. En este caso se analizó los índices de tipo interno, los cuales detectaron que el número de grupo es 8.

13. El método de conglomerados, nos permite detectar de manera simultánea, el comportamiento multicanal y monocanal de los clientes en los canales

transaccionales. Los grupos encontrados reflejan diferentes comportamientos o perfiles de clientes. Se detectaron clientes que utilizan canales virtuales para relacionarse con la Entidad Bancaria, así como clientes que prefieren estar presentes en las oficinas de la entidad para realizar sus actividades Bancarias.

14. El uso de los grados de pertenencia obtenidos en la aplicación del método C-medias, nos permitió evalúar cuáles clientes podrían cambiar de un perfil transaccional a otro, siendo ésto beneficioso para la Entidad Bancaria.

15. Al relacionar los grupos transaccionales con las características demográficas, mediante el uso de árboles de decisión, se pudo evidenciar que las variables determinantes son la edad y la educación. Se observa que los clientes con mayor edad y menos educación son los que más utilizan los canales de contacto alto. Lo que da un indicio de que la migración de un canal físico a un canal virtual es mucho más complicado para la Entidad.

5.2.2 Recomendaciones

1. Para realizar un modelo matemático para grandes volumen de información es una buena práctica para la entidad Bancaria el uso de la metodología de mineria de Datos.

2. Una de las partes críticas para el modelo matemático, es la calidad de información, por lo que es de suma importancia, el realizar una adecuada actualización y válidación de los datos.

3. Con respecto a los datos atípicos, es importante que la Entidad Bancaria analice este comportamiento anormal, con la finalidad de determinar cuáles son los factores que producen el mismo y de ser necesario aplicar las correcciones necesarias.

4. Se debe considerar que para realizar una campaña de migración de cliente de un canal hacia otro, los factores que influyen en la realización de dicha migración son la capacidad cognitiva y su ciclo de vida.

Apéndice A

A.1 Transformación de los Datos

En muchas aplicaciones reales, al realizar un modelo estadístico, el conjunto de variables no se encuentra en una forma conveniente para el modelamiento. La información puede necesitar ser modificada, con la finalidad de que los datos presenten no linealidad, asimetría, normalidad, entre otros. La transformación de los datos es resultado de la aplicación de una función matemática al conjunto de los datos en el caso de variables continuas, y un análisis de valores codificación conocida como dummy para variables categóricas.

A.1.1 Variables Continuas

Para variables intervalos:

1. Transformación simples
2. Transformación de variables continuas a discretas.
3. Transformación de variables mediante el método de Box Cox.

Transformaciones Simples

Dentro de las transformaciones simples se tienen las siguientes:

1. **Logaritmo** .- La variable es transformada tomando el log natural de la variable.
2. **Raíz Cuadrada**.- La variable es transformada tomando la raíz cuadrada de la variable.
3. **Inverso**.- La variable es transformada usando la inversa de la variable.
4. **Cuadrado**.- La variable es transformada usando el cuadrado de la variable.
5. **Exponencial**.- La variable es transformada usando el exponencial de la variable.

6. **Estandarización**.- La variable es transformada restando la media y dividiéndola para la desviación estándar.

Se debe tomar en cuenta que cuando se realiza una transformación, las nuevas variables pueden tomar valores perdidos por operaciones matemáticas ilegales (división por cero).

Transformaciones para categorizar variable continuas.

Esta transformación permite cambiar una variable continua en una variable ordinal. Hay tres tipos de transformaciones:

Intervalos.- Intervalos son creados dividiendo al conjunto de datos en intervalo de igual espacio, basados entre la diferencia del valor máximo y mínimo.

Cuantiles.- Los datos son divididos en grupos que tiene aproximadamente la misma frecuencia en cada grupo.

Intervalos óptimos para relacionarlo con una variable objetivo. Los datos son categorizados con la finalidad de optimizar la relación entre esta y la variable objetivo.

Conjunto de transformaciones Box- Cox

La familia de transformaciones más utilizada para resolver los problemas de falta de normalidad y de heterocedasticidad es la familia de Box-Cox.

Maxima Normailidad.-Este método elige la transformación que produce cuantiles de la muestra, que estén más cerca de los cuantiles teóricos de una distribución normal.

Maxima correlación con la variable objetivo.- Este método elige la transformación que tiene la mejor correlación al cuadrado con la variable objetivo..

Igualar propagación con los diferentes niveles de la variable objetivo-Este método elige la transformación que tiene la más pequeña varianza de entre los nivel de la variable objetivo.

Optimización de la máxima propagación con los diferentes niveles de la variable objetivo.-Este método elige la transformación que iguala la propagación de la variable y la relaciona con los niveles de la variable objetivo.

Todos los criterios mencionados evalúan las siguientes transformaciones:

- x
- $log(x)$

- $sqrt(x)$
- e^x
- $x^{1/4}$
- x^2
- x^4

Nota: Las variable son transformadas a una misma escala antes de aplicar cualquier transformación la escala de las variables es igual $(x-min)/(max-min)$.

A.1.2 Variables discretas

Para variables discretas o clases.

1. Transformación de niveles raros o atípicos a un grupo.
2. Transformación a indicadores dummy.

Transformaciones para agrupar niveles raros

Esta transformación es válida para variable tipo clase. Este método combina los niveles raros en un grupo separado que en general se los llaman ”otros” de acuerdo a un valor límite o corte representado en porcentaje.

Transformaciones para variables dummy

Este método crea indicadores dummy (0 o 1) para cada uno de los niveles de la variable.

A.2 Coeficiente de correlación entre variables

Es un coeficiente que nos permite medir el grado de dependencia lineal entre dos variables, es decir, que tan fuerte es la relación entre dos variables.

Entonces dado dos variables X y Y se define el coeficiente de correlación lineal o de Pearson como:

$$r_{xy} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{nS_xS_y} \tag{A.1}$$

$$r_{xy} = \frac{S_{xy}}{S_x.S_y} \tag{A.2}$$

Propiedades

- El índice de correlación de Pearson no puede valer menos de -1 ni más de +1.
  - Si r = -1, la correlación lineal es perfecta y inversa, o sea, la nube de puntos se sitúa sobre una linea recta decreciente.
  - Si r =1, la correlación lineal es perfecta y directa, o sea, la nube de puntos se sitúa sobre una línea recta creciente.
  - Un índice de correlación de Pearson de 0 indica ausencia de relación lineal. (Observar que un valor cercano a 0 del índice no implica que no haya algún tipo de relación no lineal: el índice de Pearson mide relación lineal.)
- El índice de correlación de Pearson (en valor absoluto) no varía cuando se transforman linealmente las variables.

Interpretación

Se debe tener en cuenta que el objetivo es medir cuán grande es la relación lineal existente entre dos variables en análisis. Para esto se deben considerar los siguientes pasos.

- En todo caso, es muy importante efectuar el diagrama de dispersión. Dado que un dato atípico introducido en los datos puede distorsionar el valor del coeficiente de Pearson.

- Es importante indicar que Correlación no implica causación. El que dos variables estén altamente correlaciones no implica que X causa Y ni que Y causa X.

- Es importante indicar que el coeficiente de correlación de Pearson puede verse afectado por la influencia de terceras variables.

  Por ejemplo, si fuéramos a un colegio y medimos la estatura y pasamos una prueba de habilidad verbal, saldrá que los más altos también tienen más habilidad verbal,pues esto puede ser debido simplemente a que en el colegio los niños más altos serán mayores en edad que los más bajos.

- Por otra parte, el valor del coeficiente de Pearson depende en parte de la variabilidad del grupo.

A.3 Análisis de Componentes Principales

A.3.1 Introducción

El análisis de componentes principales es una técnica que explica la estructura de varianza y covarianza de un conjunto de variables a través de una combinación lineal de estas variables. Sus principales objetivos son:

1. Reducción de los datos.

2. Interpretación de los datos

Aunque p componentes son requeridas para reproducir la variabilidad total del sistema, a menudo mucho de esta variabilidad puede ser acumulada por un pequeño grupo de k componentes principales. Entonces el conjunto original de p variables con n observaciones, puede ser reemplazada o reducida a n medidas sobre las k componentes principales.[1]

Un análisis de componente principales a menudo revela relaciones que no fueron previamente detectadas y así permite interpretaciones que no pueden ser resultados ordinarios.

A.3.2 Modelo

Algebraicamente, un análisis de componentes principales es una combinación lineal de p variable aleatorias $X_1, X_2, ..., X_P$. geométricamente, esta combinación lineal representa la elección de un nuevo sistema de coordenadas obtenida por la rotación del sistema originales con $X_1, X_2, ..., X_P$ como coordenadas de los ejes. Los nuevos ejes representan las direcciones con máxima variabilidad y provee una simple y más parsimoniosa descripción de la estructura de la covarianza.

Sea $X_T = [X_1, X_2, ..., X_P]$ con matriz de covarianza Σ cuyos valores propios son $\lambda_1 \geqslant \lambda_2 \geqslant ... \geqslant \lambda_p \geqslant 0$ y considerando las combinaciones lineales

$$\begin{array}{rcccccccccc} C_1 & = & a_1^T X & = & a_{11}X_1 & + & a_{12}X_2 & + & ... & + & a_{1p}X_p \\ C_2 & = & a_2^T X & = & a_{21}X_1 & + & a_{22}X_2 & + & ... & + & a_{2p}X_p \\ \vdots & & & & & & & & & & \vdots \\ C_p & = & a_P^T X & = & a_{p1}X_1 & + & a_{p2}X_2 & + & ... & + & a_{pp}X_p \end{array}$$

Entonces se tiene que:

$$\begin{array}{rcrcl} var(C_i) & = & a_i^T \Sigma a_i & i & = 1, 2, ..., p \\ cov(C_i) & = & a_i^T \Sigma a_k & i, k & = 1, 2, ..., p \end{array}$$

[1] Tomado el del libro Applied Multivariate Statistical /Richard Jhonson and Dean Wichern

Entonces los C_j son las componente principal definidas como la combinación lineal de los $a_{ij}X_j$.con la varianza $var(C_i)$ más grande posible.

Toda combinación lineal C_j de las variables originales puede expresarse de la siguiente manera:

$$C_i = a_{ii}X_i \qquad con \quad i = 1, \ldots, p$$

Entonces, la varianza de las componentes principales C se puede escribir como:

$$var(C_i) = a_i^t Var_x a_i$$

Así, la primera componente principal es la combinación lineal de las variables originales de **varianza máxima**. Por tanto, se busca a_1 de norma uno (con la finalidad de acotar en módulo y tener una única solución) de tal forma que la varianza de la primera componente principal C_1 sea máxima.

El planteamiento del problema se describe de la siguiente manera:

$$\begin{array}{ll} \max & a^T Var_x a \\ \text{sujeto a} & \|a\| = 1 \end{array}$$

Para resolver este problema se aplica el Lagrangiano:

$$L = a^T Var_x a - \lambda(a^T a - 1)$$

Derivando L con respecto a a e igualando a cero

$$\frac{\partial L}{\partial a} = 2Var_x a - 2\lambda a = 0 \quad \Rightarrow \quad Var_x a = \lambda a \quad \Rightarrow \quad (Var_x - \lambda I)a = 0$$

De donde se concluye que a es vector propio de la matriz de varianza-covarianza de los datos originales. Es decir, la primera componente principal c_1 se obtiene haciendo $c_1 = Xa_1$ donde a_1 es el vector propio de la matriz de varianza-covarianza con mayor valor propio asociado.

Puesto que la varianza es una medida de información, y sabiendo que:

$$var(C_i) = a_i^t Var_x a_i$$

Entonces la varianza es

$$var(C_i) = \lambda_i$$

Así, la información recogida por cada componente es el cociente entre la variabilidad de la componente y la varianza total

$$Infor_i = \frac{var(C_i)}{varianza \quad total} = \frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i}$$

A.3.3 Prueba de Bartlett

La prueba de Bartlett chequea si la matriz de correlación $R_{(pxp)}$ diverge significativamente de la matriz identidad. Realiza la prueba de hipótesis siguiente:

$$H_0 : lasvariablesnosonortogonales.$$

$$H_1 : lasvariablesnosonortogonales.$$

De acuerdo a esta prueba el ACP es correcto realizarlo si se rechaza la hipótesis nula. Para medir la relación entre las variables, se calcula el determinante de la matriz de correlación. Para H_0,$|R| = 1$; Si las variables son altamente correlacionadas, se tiene $|R| \approx 0$.La prueba estadística de Bartlett evalúa, que tan alejados estamos de la situación $|R| = 0$, mediante el estadístico χ^2.

$$\chi^2 = -(n - 1 - \frac{2p + 5}{6} * \ln |R|)$$

Donde n número de observaciones de la matriz de datos.

Bajo H_0, se sigue una distribución χ^2 con $[p * (p - 1)/2]$ grados de libertad.

Apéndice B

Código R

B.1 Implementación

```
#0.- Análisis Estadísticos

# Carga de la base
base <-read.table("RANDRANDOMSAMPLETEMP.txt",sep=" ",header=T);

#Resumen estadístico
resumen <- summary(base[,1:17]);

#Filtro de no transacciones
base_fil <- subset(base,total_trans>0);
summary(base_fil[,1:17]);

#número de cargas
par(bg="lightgray");
hist(base_fil$NUM_CRG, main="Número de Cargas", breaks=12, col="red") ;

#Estado civil
counts <- table(base_fil$DES_EST_CIV_FIN);
par(bg="lightgray",las=2,mar=c(5,8,4,2));
barplot(counts, main="Estado Civil",horiz=TRUE,col="red");

#Provincia
counts <- table(base_fil$DES_PRV_FIN);
par(bg="lightgray",las=2,mar=c(5,10,4,2));
barplot(counts, main="Provincia",horiz=TRUE,col="red");

#Vivienda
counts <- table(base_fil$ DES_VIV_FIN);
par(bg="lightgray",las=2,mar=c(5,11,4,2));
```

```
barplot(counts, main="Vivienda",horiz=TRUE,col="red");

#Genero
counts <- table(base_fil$ DES_GNR);
par(bg="lightgray",las=2,mar=c(5,8,4,2));
barplot(counts, main="Genero",horiz=TRUE,col="red");

#Nivel de Estudios
counts <- table(base_fil$ DES_NIV_ESD_FIN);
par(bg="lightgray",las=2,mar=c(5,8,4,2));
barplot(counts, main="Nivel de Estudios",horiz=TRUE,col="red");


#####################################################

#1.- Atipicos

#Entrada: base de datos=base_fil
#Salida: base_mod <- subset(temp3,atipico=0);
#        base_ati <- subset(temp3,atipico>0);

salida <- kmeans(base_fil[,1:7],30, iter.max = 30);

res_clus <- matrix(data=salida$cluster,nrow=nrow(base_fil),ncol=3);#id_cluster

porcentaje <- matrix(table(salida$cluster)*(1/nrow(base_fil))*100,nrow=30,ncol=3);

for (i in 1:30)

{
    if (porcentaje[i,1]< 1)
     {
    porcentaje[i,2]=1#atipico
    porcentaje[i,3]=i#id_cluster

   }
else
   {
```

```
   porcentaje[i,2]=0#no atipico
   porcentaje[i,3]=i#id_cluster
   }

}

temp_ind <- matrix(data=base_fil[,18],ncol=1,nrow=nrow(base_fil));#id_tabla

for (j in 1:nrow(base_fil))
{
res_clus[j,2]=j;#id_reg
res_clus[j,3]=temp_ind[j,1];#id_tabla
}

temp2 <-data.frame(id_clu=porcentaje[,3],atipico=porcentaje[,2]);
temp1 <-data.frame(id_clu=res_clus[,1], id_reg=res_clus[,2] ,  cod=res_clus[,3]);

atipicos <- merge(temp1,temp2,all.x=TRUE);

attach(atipicos);
atipicos <- atipicos[order(cod),];
detach(atipicos);

temp3 <- merge(atipicos,base_fil);

base_mod <- subset(temp3,atipico==0);
base_ati <- subset(temp3,atipico>0);

summary(base_mod[,5:11]);
summary(base_ati[,5:11]);

##############################################################################


##############################################################################

#2.- Componentes Principales
```

```
comp <- princomp(base_mod[,5:11], cor=TRUE)
summary(comp)
loadings(comp)
plot(comp,type="lines")
base_comp <- comp$scores
biplot(comp)

##############################################################################

##############################################################################
#3.-Estandarización
base_comp_std <- scale(base_comp)
summary(base_comp_std)

##############################################################################

##############################################################################

#4.- Número de grupos

SECG <- (nrow(base_comp_std)-1)*sum(apply(base_comp_std,2,var))
for (i in 2:15)
SECG[i] <- sum(kmeans(base_comp_std,centers=i,iter.max=20,)$withinss)
plot(1:15,SECG, type="b", xlab="Número de grupos",
ylab="Suma de errores al cuadrado dentro de los grupos")



F1 <- function(n,datos){
SCDG <- matrix(data=0,nrow=n-1,ncol=3);#W
SCEG <- matrix(data=0,nrow=n-1,ncol=3);#B


for (i in 2:n)

{
  res<-kmeans(datos,i,iter.max=20)
  #res<-pam(datos,i)
```

```
  SCDG[i-1,2]=res$tot.withinss;
  SCDG[i-1,1]=i;
  SCEG[i-1,2]=res$betweenss;
  SCEG[i-1,1]=i;
 }

a<- matrix(data=0,nrow=n-1,ncol=1);
b<- matrix(data=0,nrow=n-1,ncol=1);
F<- matrix(data=0,nrow=n-1,ncol=2);
CH<- matrix(data=0,nrow=n-1,ncol=2);

a[,1]=SCDG[,2];


for (i in 1:n-1)
{
   if (i==1)
   {
   b[i,1]=SCDG[i,2]
   }
   else
   {
   b[i,1] <- SCDG[i-1,2];
   }

F[i,2]<-  (b[i,1] - a[i,1])/ ( a[i,1]/  (length(datos[,1])-SCDG[i,1]-1));
F[i,1]<-  SCDG[i,1];

CH[i,1]<- SCEG[i,1];

CH[i,2]<- (  SCEG[i,2]/(SCDG[i,1]-1)  )   / (   SCDG[i,2] /(length(datos[,1])
- SCDG[i,1])   );
}

plot(F,xlab="n",ylab="Indice");
points(F,col=2);
points(CH,col=3);
leg.txt <-c("F","H");
```

```
#legend(n-4,80,leg.txt,col=2:3,pch=16:18);
legend("topright",leg.txt,col=2:3,pch=16:18);
return(list(SCDG,SCEG,F,CH))
}


resultado <- F1(25,base_comp_std);


#Nueva función para estimar los grupos


F2 <- function(k,datos){

t=15;

n_grupos1 <- matrix(data=0,nrow=5,ncol=k);
n_grupos2 <- matrix(data=0,nrow=5,ncol=k);

      for (i in 1:k)
      {

      resul<- F1(t,datos);

      sim1 <- matrix(data=unlist(resul[4]),nrow=t-1,ncol=2);

      dd <- sim1[order(sim1[,2]),];

     ini <- t-5;
     fin <- t-1;

      dd <- dd[ini:fin,];

      n_grupos1[,i] <- dd[,1];

      sim2 <- matrix(data=unlist(resul[3]),nrow=t-1,ncol=2);

      ee <- sim2[order(sim2[,2]),];
```

```
      ee <-  ee[ini:fin,];

      n_grupos2[,i] <- ee[,1];


      }

return(list(n_grupos1,n_grupos2))

}

resultado <- F2(15,base_comp_std);

contar_1<-table(unlist(resultado[1]));
contar_2<-table(unlist(resultado[2]));
barplot(contar_1, main="Frec. Grupos  índice  Hartigan",xlab="Número de  grupos");
barplot(contar_2, main="Frec. Grupos índice   Harabasz",xlab="Número de  grupos");

##############################################################################
##############################################################################
####### Algoritmo de  Conglomerados ##########################################

library(class);
library(e1071);
modelo_1 <- kmeans( base_comp_std,7);
modelo_2 <- cmeans( base_comp_std,7);

library(plotrix);
counts_1 <- table(modelo_1$cluster);
porc_1   <- round(counts_1/sum(counts_1) * 100, 1);
labels_1 <- paste(porc_1, "%", sep="");
pie3D(porc_1, main="Porcentaje de conglomerados
k-medias",labels=labels_1, cex=0.8)


library(psych);
base_resultado <- data.frame(base_mod, modelo_1$cluster);
```

```
resumen_1 <- describe.by(base_resultado[,5:11],modelo_1$cluster);

library(gplots)
attach(base_resultado)
par(mfrow=c(3,3))
plotmeans(TOT_ATM~modelo_1$cluster,data=base_resultado,col="red");
plotmeans(TOT_BLC~modelo_1$cluster,data=base_resultado,col="red");
plotmeans(TOT_CEL~modelo_1$cluster,data=base_resultado,col="red");
plotmeans(TOT_ITR~modelo_1$cluster,data=base_resultado,col="red");
plotmeans(TOT_KIO~modelo_1$cluster,data=base_resultado,col="red");
plotmeans(TOT_TLN~modelo_1$cluster,data=base_resultado,col="red");
plotmeans(TOT_VTL~modelo_1$cluster,data=base_resultado,col="red");

counts_2 <- table(modelo_2$cluster);
porc_2   <- round(counts_2/sum(counts_1) * 100, 1);
labels_2 <- paste(porc_2, "%", sep="");
pie3D(porc_2, main=" Porcentaje de conglomerados
C-medias",labels=labels_2,cex=0.8)

base_resultado2 <- data.frame(base_mod, modelo_2$cluster)
resumen_2 <- describe.by(base_resultado2[,5:11],modelo_2$cluster);

#library(gplots)
attach(base_resultado2)
par(mfrow=c(3,3))
plotmeans(TOT_ATM~modelo_2$cluster,data=base_resultado2,col="red");
plotmeans(TOT_BLC~modelo_2$cluster,data=base_resultado2,col="red");
plotmeans(TOT_CEL~modelo_2$cluster,data=base_resultado2,col="red");
plotmeans(TOT_ITR~modelo_2$cluster,data=base_resultado2,col="red");
plotmeans(TOT_KIO~modelo_2$cluster,data=base_resultado2,col="red");
plotmeans(TOT_TLN~modelo_2$cluster,data=base_resultado2,col="red");
plotmeans(TOT_VTL~modelo_2$cluster,data=base_resultado2,col="red");
```

BIBLIOGRAFÍA

[1] George J.klir and Bo Yuan (1995) *Fuzzy set and fuzzy logic theory and applications*, Prentice

[2] Ronald S. Swift (2001) *Accelerating Customer Relationships*, Prentice Hall

[3] César Pérez (2008) *Econometría Avanzada y herramientas* Pearson,Madrid-España

[4] Roberto Dvoskin (2004) *Fundamentos de Marketing*, Ediciones Granica S.A.,Argentina

[5] Witold Pedrycz (2005)*Knowledge-based clustering*, Wiley Interscience, Canada

[6] Alvin C. Rencher (2002) *Methods of Multivariate Analysis*,Wiley, Canada

[7] B.W. Silverman (2002)*Density estimation for statistics and data analysis*, paper

[8] Sadaaki Miyamoto and Katsuhiro Honda (2008)*Algorithms for Fuzzy Clustering Methods in c-Means Clustering with Applications*,Springer,Berlin

[9] Jussi Klemela (2009)*Smoothing of Multivariate Data*, Wiley,New Jersey

[10] Robert C. Blattberg,Byung-Do Kim and Scott A. Neslin (2008)*Database Marketing*,Springer

[11] Anil K. Jain, Richard C. Dubes (1988)*Algorithms for Clustering Data*, Prentice Hall

[12] Jiawei Han, Micheline Kamber (2006)*Data Mining:Concepts and Techniques*, Elsiver

[13] Guojun Gan, Chaoqun Ma, Jianhong Wu (2007)*Data Clustering Theory, Algorithms, and Applications*,SIAM

[14] Rui Xu, Donald C. Wunsch (2009)*Clustering*, Wiley

[15] Wolfgang Härdle, Léopold Simar (2007)*Applied Multivariate Statistical Analysis*, Springer

[16] Nong Ye (2003) *The Handbook of Data Mining*, Lawrence Erlbaum Associates, Inc.

[17] Ian H. Witten, Eibe Frank (2005) *Data Mining*, Elsiver.

[18] Brian S. Everitt and Ibrsten Hothorn (2010) *A Handbook of Statistical Analyses Using R*, Chapman and Hall.

[19] Luis Torgo (2003) *Data Mining with R*, LIACC-FEP.

[20] Philip Kotler y Gary Armstrong (2003) *Fundamentos de Marketing*, Pearson.

[21] Jaime Gil (1997) *Marketing para el Nuevo Milenio*, Piramide.

[22] Adrian Payne (2006) *Handbook of CRM*, Elsevier.

[23] Arjan Sundardas (2005) *Marketing Financiero*, MCGRAW-HILL.

[24] Konstantinos Tsiptsis, Antonios Chorianopoulos (2011) *Data Mining Techniques in CRM*, John Wiley and Sons.

[25] César Pérez (2006) *Data Mining*, Ra-Ma.

[26] Mamdouh Refaat(2010) *Data Preparation for Data Mining Using SAS*, Morgan Kaufmann.

[27] S. Sumathi, S.N. Sivanandam(2006) *Introduction to Data Mining and Its Applications*, Springer.

[28] Daniel Peña (2002) *Análisis de datos Multivariante*, MCGRAW-HILL.

[29] Howard E.A. Tinsley, Steven D. Brown (2000) *Handbook of Applied Multivariate Statistics and Mathematical Modeling*, Academic Press.

[30] Swagatam Das, Ajith Abraham, Amit Konar (2009) *Metaheuristic Clustering*, Springer.

[31] Wolfgang Karl. Härdle, Lâeopold Simar (2012) *Applied multivariate statistical analysis*, Springer.

[32] Iván Pérez y Betzabeth León (2007) *Logica difusa para principiantes*, Texto.

[33] Wendy L. Martinez, Angel R. Martinez, Jeffrey L. Solka (2011) *Exploratory Data Analysis with Matlab*, Taylor and Francis.

[34] Alan J. Izenman (2008) *Modern Multivariate Statistical Techniques*, Springer.

[35] K. P. Soman, Shyam Diwakar, V. Ajay (2006)*Insight Into Data Mining*, PHI Learning.