UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE ESTUDIOS SUPERIORES ACATLÁN

SEGMENTACIÓN NO SUPERVISADA DE TARJETA HABIENTES

TRABAJO PROFESIONAL

QUE PARA OBTENER EL TITULO DE:

LICENCIADA EN MATEMÁTICAS APLICADAS Y COMPUTACIÓN

PRESENTA

PATRICIA CASTRO AYALA

ASESOR: ING. RUBÉN ROMERO RUÍZ

DICIEMBRE 2006

Universidad Nacional Autónoma de México
UNAM
Biblioteca Central
Dirección General de Bibliotecas de la UNAM

UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

CONTENIDO

RESUMEN

Actualmente en todas las empresas se generan millones de datos diariamente; específicamente en las institución financiera existen millones de clientes con tarjeta de crédito y por cada uno de ellos se genera información de los cargos a la tarjeta de crédito, disposiciones, intereses generados, promedio de compras al mes, etc. A partir de todos estos datos; la minería de datos nos permite encontrar patrones en los clientes y de esta manera clasificarlos en grupos, para tomar decisiones sobre estos; por ejemplo: para realizar campañas dirigidas y no masivas; a su vez, estas empresas dependen de sus datos para retener a sus clientes, expandir su cuota de mercado y diferenciarse de sus rivales; es por ello que la necesidad de explorar estas bases de datos y extraer información y conocimiento que sea de interés para los propietarios de las mismas, se ha incrementado.

En este contexto, el propósito de este proyecto es clasificar una base de tarjetahabientes de una Institución Financiera utilizando como herramienta de desarrollo SAS Enterprise Miner y como técnica de solución análisis de clusters; guiándonos en la metodología SEMMA desde la etapa de preparación y limpieza de datos hasta la realización del modelo y análisis de resultados que nos permiten llevarlos a reglas de negocio para una toma de decisiones.

INTRODUCCIÓN

Hoy en día, la cantidad de datos que ha sido almacenada en las bases de datos excede nuestra habilidad para reducir y analizar los datos sin el uso de técnicas de análisis automatizadas. Muchas bases de datos comerciales transaccionales y científicas crecen a una proporción fenomenal.

Ha llegado un momento en el que disponemos de tanta información que nos vemos incapaces de sacarle provecho. Los datos tal cual se almacenan no proporcionan beneficios directos. Su valor real reside en la información que podamos extraer de ellos, información que nos ayude a tomar decisiones o a mejorar nuestra comprensión de los fenómenos que nos rodean.

Se requiere de grandes cantidades de datos que proporcionen información suficiente para derivar un conocimiento adicional.

La asimilación de hechos pasados permite enfrentar al futuro con más posibilidades de éxito, sin tener que recordar todos los detalles del pasado. Esto es claro en personas, pero, ¿cómo aplicar esto a las empresas?.

Los datos recogen un conjunto de hechos (una base de datos) y los patrones son expresiones que describen un subconjunto de los datos (un modelo aplicable a ese subconjunto).

Se ha llamado Minería de Datos (Minería datos) al análisis de archivos y registros de transacciones con el fin de descubrir patrones, relaciones, reglas, asociaciones o incluso excepciones que sean útiles para la toma de decisiones.

En particular, el objetivo fundamental es encontrar conocimiento útil, válido, relevante y nuevo sobre un fenómeno o actividad mediante algoritmos eficientes, dadas las

crecientes órdenes de magnitud en los datos. Al mismo tiempo hay un profundo interés por presentar los resultados de manera visual o al menos de manera que su interpretación sea muy clara.

1. MINERIA DE DATOS

1.1. DEFINICIÓN

Es un proceso de descubrimiento de los patrones, perfiles y tendencias presentes y significativas a través del análisis de los datos utilizando tecnologías de reconocimientos de patrones, como las redes neuronales, árboles de decisión. Es un proceso iterativo de extracción de patrones procedentes de las transacciones de negocios.

La Minería de Datos ha surgido con el fin de obtener conocimiento que apoye la toma de decisiones y que pueda construir una experiencia a partir de los millones de transacciones detalladas que registra una corporación en sus sistemas informáticos.

En la actualidad, las empresas, inundadas con datos generados diariamente por las transacciones con cada cliente (las visitas a las web, códigos de barras, cargos de las tarjetas de crédito y las llamadas telefónicas), afrontan el mismo reto del reconocimiento de patrones de oportunidad para su supervivencia.

A través del reconocimiento de los clientes mas rentables de una empresa, ésta puede empezar a establecer una relación más estrecha con ellos, lo que a largo plazo reforzará su capacidad de retenerlos a lo largo de su vida de cliente.

1.2. PROBLEMAS DE MINERÍA DE DATOS

La modelación en computadora del tiempo y el espacio son problemas complejos, especialmente para hacer inferencias. Esto hace que las técnicas de Aprendizaje Automático enfrenten mayores dificultades cuando abordan los temas que parecen más interesantes, de descubrimiento de patrones.

A esto se añaden dos tipos de problemas:

1. El primero, es el problema de Minería de Datos con relaciones en el tiempo. Es muy posible que se deseen hacer inferencias y análisis de datos sobre un periodo determinado, pero que durante dicho periodo no se haya registrado el mismo número de variables, o que éstas no tengan la misma precisión, o carezcan de la misma interpretación. En ciertos casos, puede que se haya hecho un ejercicio de Minería de Datos en el pasado y que los datos se hayan descartado o destruido, pero que se desee hacer una comparación con datos más recientes. Nótese que un ejercicio de Minería de Datos puede traer a la luz relevancia de variables y factores, pero que sea imposible recopilar estas variables y completar adecuadamente conjuntos de datos del pasado. Otros problemas de análisis de datos es que no se conocen todos en un tiempo continuo. Por ejemplo, si se hacen recopilaciones mensuales, es imposible hacer una predicción semanal.

2. Desde el punto de vista de la privacidad. Cuando la Minería de Datos era aún emergente, se pensó que no presentaba ningún peligro para la privacidad de los clientes. Hoy en día, se piensa todo lo contrario. Sin embargo, no existe un marco jurídico que haya mantenido el paso con el avance tecnológico; esto es, hoy en día, las corporaciones comercializan con millones de perfiles personales, sin que aquellos a quienes se refieren dichos datos estén en posibilidad de intervenir. Cada llamada telefónica, cada transacción bancaria, cada compra, es registrada en una computadora, y si la compañía de teléfonos, el banco y otros combinan sus bases de datos, están en condiciones de elaborar un perfil muy completo. Este perfil definiría a más de una persona. Si a esto añadimos qué sitios web visita, qué y dónde compró con la tarjeta de crédito, etc., no existe ninguna privacidad. El problema va desde las definiciones de qué constituye privacidad y quién es el propietario de los datos, para poder hacer uso de estos.

1.3. APLICACIONES DE MINERÍA DE DATOS

La Minería de datos ha tenido un crecimiento en el área de marketing en bases de datos. En aplicaciones como:

- Retención del cliente y gestión de los abandonos;
- Venta cruzada (cross-selling);
- Gestión de campañas;
- Análisis del mercado, canal y precio;
- Análisis de segmentación del cliente.

La tecnología de la Minería de datos se está convirtiendo rápidamente en algo popular debido a la necesidad del marketing centrado en el cliente (Customer Relationship Management - CRM). Los departamentos de Tecnología de la Información y de Marketing que no siguen ésta tendencia se despertarán para descubrir que sus competidores les han desplazado.

La Minería de Datos proporciona a las empresas una visión muy valiosa del negocio que, hasta la actualidad, ha permanecido guardada celosamente como tecnología de secreto corporativo para proteger la ventaja competitiva.

Tradicionalmente, las tecnologías de Minería de Datos se han aplicado para:

1.3.1. Banca y empresas aseguradoras

Creación de modelos financieros, Detección de fraudes, Creación de perfiles del mercado y de la industria, Marketing de base de datos, Segmentación de Mercados. En el sector bancario la información que puede almacenarse es, además de las cuentas de los clientes, la relativa a la utilización de las tarjetas de crédito, que puede permitir conocer hábitos y patrones de comportamiento de los usuarios.

Esta información puede aplicarse para:

- Detectar patrones de uso fraudulento de tarjetas de crédito.
- Identificar clientes leales: Es importante para las compañías de cualquier sector mantener los clientes. Ya que hay estudios que demuestran que es cuatro veces más caro obtener nuevos clientes que mantener los existentes.
- Predecir clientes con probabilidad de cambiar su afiliación.
- Determinar gasto en tarjeta de crédito por grupos.
- Encontrar correlaciones entre indicadores financieros.
- Identificar reglas de mercado de valores a partir de históricos.

1.3.2. Compañías de Seguros

En el sector de las compañías de seguros y la salud privada, se pueden emplear las técnicas de minería de datos, por ejemplo para:

- Análisis de procedimientos médicos solicitados conjuntamente.
- Predecir qué clientes compran nuevas pólizas.
- Identificar patrones de comportamiento para clientes con riesgo.
- Identificar comportamiento fraudulento

1.3.3. Marketing

Marketing de base de datos, Gestión de la eficacia de la publicidad, del inventario y de las categorías. Actualmente con la generación de los puntos de ventas informatizados y conectados a un ordenador central, y el constante uso de las tarjetas de créditos se genera gran cantidad de información que hay que analizar. Con ello se puede emplear la minería de datos para:

- Identificar patrones de compra de los clientes: Determinar cómo compran, a partir de sus principales características, conocer el grado de interés sobre tipos de productos, si compran determinados productos en determinados momentos.
- Segmentación de clientes: Consiste en la agrupación de los clientes con características similares, por ejemplo demográficas. Es una importante herramienta en la estrategia de marketing que permite realizar ofertas acordes a diferentes tipos de comportamiento de los consumidores.
- Predecir respuestas a campañas de mailing: Estas campañas son caras y pueden llegar a ser molestas para los clientes a los que no le interesan el tipo de producto promocionado por lo que es importante limitarlas a los individuos con una alta probabilidad de interesarse por el producto.
- Está por ello muy relacionada con la segmentación de clientes. Análisis de cestas de la compra [market-basket analysis]: Consiste en descubrir relaciones entre productos, esto es, determinar qué productos suelen comprarse junto con otros, con el fin de distribuirlos adecuadamente.

1.3.4. Telecomunicaciones

Análisis de los registros de las llamadas detalladas, Utilización óptima del equipamiento capital, Marketing dirigido

En el sector de las telecomunicaciones se puede almacenar información interesante sobre las llamadas realizadas, tal como el destino, la duración, la fecha, en que se realiza la llamada, por ejemplo para:

- Detección de fraude telefónico: Mediante el agrupamiento o clustering se pueden detectar patrones en los datos que permitan detectar fraudes.

1.3.5. Gobierno

- Recolecciones, Selección de puestos de trabajo, Detección de fraudes, Logística, y Estrategia inteligente.

1.3.6. Manufacturación

- Control de Calidad de la producción, Control de la Cadena de Montaje y del Inventario

1.3.7. Medicina

También en el campo médico se almacena gran cantidad de información, sobre los pacientes, tal como enfermedades pasadas, tratamientos impuestos, pruebas realizadas, evolución. Se pueden emplear técnicas de minería de datos con esta información, por ejemplo, para:

- Identificación de terapias médicas satisfactorias para diferentes enfermedades.
- Asociación de síntomas y clasificación diferencial de patologías.
- Estudio de factores (genéticos, precedentes, hábitos alimenticios) de riesgo para la salud en distintas patologías.
- Segmentación de pacientes para una atención más inteligente según su grupo.
- Estudios epidemiológicos, análisis de rendimientos de campañas de información, prevención, sustitución de fármacos.
- Identificación de terapias médicas y tratamientos erróneos para determinadas enfermedades.

1.3.8. Industria farmacéutica

En el sector químico y farmacéutico se almacenan gran cantidad de información:
Bases de datos de dominio público conteniendo información sobre estructuras y propiedades de componentes químicos:

- Resultados de universidades y laboratorios publicadas en revistas técnicas.
- Datos generados en la realización de los experimentos.
- Datos propios de la empresa.
- Los datos son almacenados en diferentes categorías y a cada categoría se le aplica un diferente trato. Se podrían realizar, entre otras, las siguientes operaciones con la información obtenida:
- Clustering de moléculas: Consiste en el agrupamiento de moléculas que presentan un cierto nivel de similitud, con lo que se pueden descubrir importantes propiedades químicas.
- Búsqueda de todas las moléculas que contienen un patrón específico:
- Se podría introducir una subestructura (un patrón), devolviendo el sistema todas las moléculas que son similares a dicha estructura.
- Búsqueda de todas las moléculas que vincula un camino específico hacia una molécula objetivo: Realizar una búsqueda exhaustiva puede ser impracticable, por lo que se pueden usar restricciones en el espacio de búsqueda.
- Predicción de resultado de experimentos de una nueva molécula a partir de los datos almacenados: a través de determinadas técnicas de inteligencia artificial es posible predecir los resultados a nuevos experimentos a partir de los datos, con el consiguiente ahorro de tiempo y dinero.

1.4. NUEVAS APLICACIONES DE MINERÍA DE DATOS

En Internet, la Minería de Datos aporta básicamente dos nuevas aplicaciones en apoyo a la toma de decisiones:

- La Minería de datos se puede utilizar para analizar los datos generados por un website en la red.
- La Minería de Datos se puede utilizar para monitorear y detectar la aparición de anomalías y de problemas potenciales en la red antes de que ocurran.
- La Minería de datos Web es una tecnología usada para descubrir conocimiento interesante en todos los aspectos relacionados a la Web. Es uno de los mayores retos. El enorme volumen de datos en la Web generado por la explosión de usuarios y el desarrollo de librerías digitales hace que la extracción de la información útil sea un gran problema. Cuando el usuario navega se encuentra frecuentemente saturado por los datos. La integración de herramientas de minería de datos puede ayudar a la extracción de la información útil.

La Minería de datos Web se puede clasificar en tres grupos distintos no disjuntos, dependiendo del tipo de información que se quiera extraer, o de los objetivos

1. Minería del Contenido de la Web [Web Content Mining]: Extraer información del contenido de los documentos en la web. Se puede clasificar a su vez en:

    - Text Mining: Si los documentos son textuales (planos).
    - Hypertext Mining: Si los documentos contienen enlaces a sí mismos o a otros documentos
    - Markup Mining: Si los documentos son semiestructurados (con marcas).
    - Multimedia Mining: Para imágenes, audio, vídeo.

2. Minería de la Estructura de la Web [Web Structure Mining]: Se intenta descubrir un modelo a partir de la tipología de enlaces de la red. Este modelo puede ser útil para clasificar o agrupar documentos.

3. Minería del Uso de la Web [Web Usage Mining]: Se intenta extraer información (hábitos, preferencias, etc. de los usuarios o contenidos y relevancia de documentos) a partir de las sesiones y comportamiento de los usuarios navegantes.

2. METODOLOGÍAS PARA LA REALIZACIÓN DE PROYECTOS DE MINERÍA DATOS

Los proyectos de Minería datos tienen por objetivo extraer información útil a partir de grandes cantidades de datos y se aplican a todos los sectores y en todos los campos. Así existen proyectos de este tipo en sectores tan dispares como el comercio electrónico, la banca, las empresas industriales o la exploración petrolífera.

La extracción de esta información útil es un proceso complejo, que requiere la aplicación de una metodología estructurada para la utilización ordenada y eficiente de las técnicas y herramientas disponibles.

2.1. METODOLOGÍA SAS

Ante la necesidad existente en el mercado de una aproximación sistemática para la realización de los proyectos de Minería datos, diversas empresas y consultorías han especificado un proceso de modelado diseñado para guiar al usuario a través de una sucesión de pasos que le dirijan a obtener buenos resultados.

SAS propone la utilización de la metodología SEMMA (Sample, Explore, Modify, Model, Assess). En 1999 un importante consorcio de empresas europeas, NCR (Dinamarca), AG(Alemania), SPSS (Inglaterra) y OHRA (Holanda), unieron sus recursos para el desarrollo de la metodología de libre distribución CRISP-DM (Cross-Industry Standard Process for Minería datos). Esta metodología, junto con la metodología SEMMA, son las mas utilizadas por los analistas en los proyectos de Minería datos.

La metodología SAS consiste de seis actividades (Figura 1):

2.1.1. Definir el problema

Es importante contar con un equipo de trabajo en el que se involucre a un experto del negocio que podrá definir la situación actual, haciendo algunos cuestionamientos que llevarán a un planteamiento de problema que finalmente se transmitirán al experto en minería de datos; quien también deberá involucrarse y entender sobre las reglas del negocio para poder llegar a un resultado exitoso.

2.1.2. Evaluar el medio

En esta parte se evalúan las herramientas con las que se cuentan para realizar la minería; así como la fuente de datos desde sistemas implementados hasta archivos. También se encuentran las consideraciones de negocio sobre los datos que se van a poder acceder y en que formato se encuentran y con que periodicidad se generan, para conocer su validez.

2.1.3. Obtener los datos

Se deben obtener los insumos necesarios con los que se trabajará para crear un ambiente de minería de datos básicamente se definen los datos que se necesitarán así como la historia de estos y también se realiza la validez y preparación para la aplicación de la metodología SEMMA.

2.1.4. Minería de Datos en ciclos

Después de clarificar las preguntas del negocio y de contar con una fuente de datos; se puede empezar el proceso de minería siguiendo la metodología SEMMA (Sample, Explore, Modify, Model y Asses). Como su nombre lo dice se empieza obteniendo una muestra (Sample) de la población que se estudiará; después se hace un análisis exploratorio (Explore) apoyándonos de la estadística y si es necesario se aplican

algunas reglas para hacer cambios y transformaciones (Modify.) a partir de esto podemos realizar varios modelos (Model) que finalmente tendremos que evaluar (Assess) para seleccionar el mejor.

2.1.5. Implementación

Después de elegir el mejor modelo es necesario crear reportes de resultados que serán de gran utilidad al negocio así como implementarlo en producción esto es desde definir la arquitectura para su integración hasta el mantenimiento de este (periodicidad con que se tendrá que actualizar y tiempo de vida del mismo).

2.1.6. Revisión

Es importante saber que tan confiables son los modelos con los que se trabajan para esto es necesario dar seguimiento de los resultados obtenidos a partir de la implementación de los mismos; esto nos dará idea de su buen desempeño y beneficios obtenidos por las empresas.

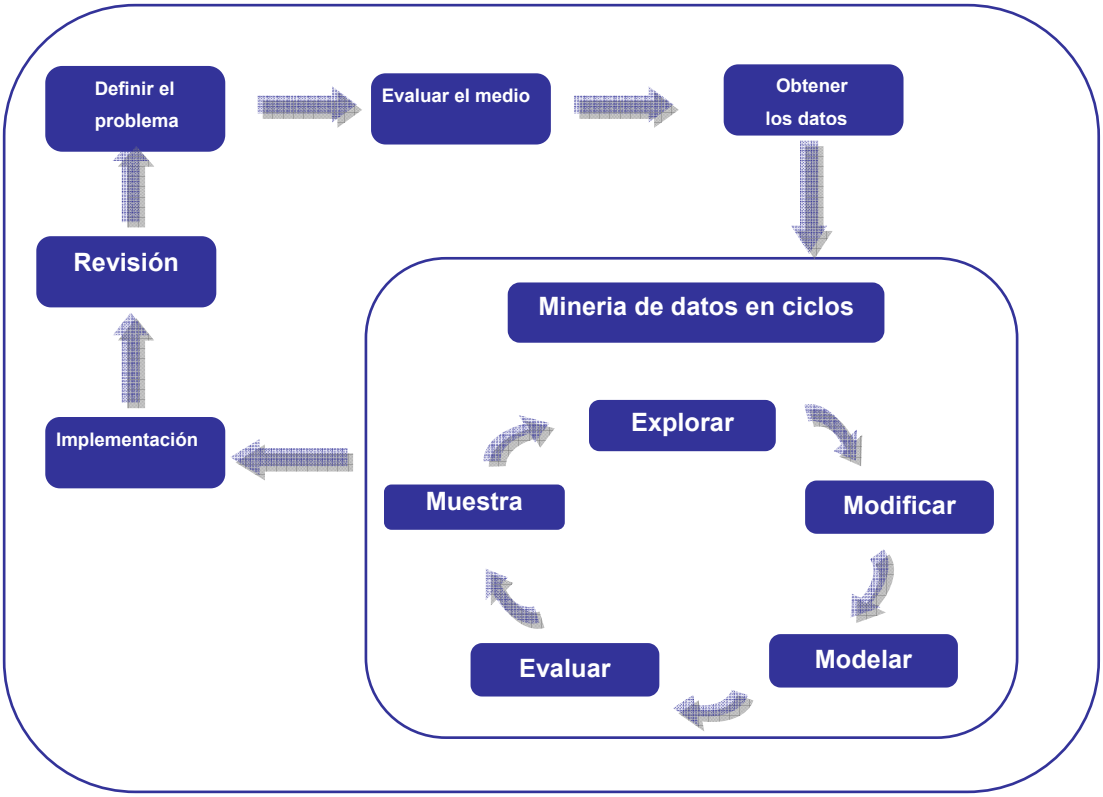


Figura 1.1 **Esquema de los componentes que integran la metodología SAS para minería de datos**

2.2. METODOLOGÍA SEMMA

SAS Institute desarrollador de esta metodología, la define como el proceso de selección, exploración, modelado de grandes cantidades de datos para descubrir patrones de negocio desconocidos. El nombre de esta terminología es el acrónimo correspondiente a las cinco fases básicas del proceso.

2.2.1. Obtener una muestra de los datos

El proceso se inicia con la extracción de la población maestral sobre la que se va a aplicar el análisis. El objetivo de esta fase consiste en seleccionar una muestra representativa del problema en estudio. La representatividad de la muestra es indispensable ya que de no cumplirse invalida todo el modelo y los resultados dejan

de ser admisibles. La forma más común es mediante el método de muestreo se denomina muestreo aleatorio simple.

- Muestreo simple
- Muestreo estratificado
- Muestreo sistemático

2.2.2. Explorar los datos

Una vez determinada una muestra o conjunto de muestras representativas de la población en estudio, la metodología SEMMA indica que se debe proceder a una exploración de la información disponible con el fin de simplificar en lo posible el problema para optimizar la eficiencia del modelo. Para lograr este objetivo se propone la utilización de herramientas de visualización o de técnicas estadísticas que ayuden a poner de manifiesto relaciones entre variables. De esta forma se pretende determinar cuáles son las variables explicativas que van a servir como entradas al modelo.
Esta fase consiste en familiarizarse con los datos, se obtendrá conocimiento a cerca de:

- Su tendencia
- Sus rangos
- Frecuencias
- Correlaciones

2.2.3. Modificar los datos

La tercera fase de la metodología consiste en la manipulación de los datos, en base a la exploración realizada, de forma que se definan y tengan el formato adecuado los datos que serán introducidos en el modelo.

- Algunos métodos requieren variables en una escala específica
- Conversión de variables nominal a entera
- Discretización de variables
- Sustituir valores perdidos por la media
- Las transformaciones de las variables usando operaciones logarítmicas, exponencial, inversa, etc.
- Quitar outliers
- Creación de nuevas variables

Los datos se pueden dividir en tres conjuntos que servirán en la creación y validación del modelo:

- Conjunto de entrenamiento, con el cual se crea el modelo
- Conjunto de validación, que sirve para evaluar el desempeño del modelo
- Conjunto de prueba, asegura el desempeño de los modelos ante datos nuevos

2.2.4. Construir el modelo

Una vez que se han definido las entradas del modelo, con el formato adecuado para la aplicación de la técnica de modelado, se procede al análisis y modelado de los datos. El objetivo de esta fase consiste en establecer una relación entre las variables explicativas y las variables objeto del estudio, que posibiliten inferir el valor de las mismas con un nivel de confianza determinado. Las predicciones se utilizan para prever el comportamiento futuro de algún tipo de entidad mientras que una descripción puede ayudar a su comprensión. De hecho, los modelos predictivos pueden ser descriptivos (hasta donde sean comprensibles por personas) y los modelos descriptivos pueden emplearse para realizar predicciones.

Una técnica constituye el enfoque conceptual para extraer la información de los datos y en general es implementada por varios algoritmos. Cada algoritmo representa, en la práctica, la manera de desarrollar una determinada técnica paso a paso, de forma que es preciso un entendimiento de alto nivel de los algoritmos para saber cual es la técnica más apropiada para cada problema. Asimismo es preciso entender los parámetros y las características de los algoritmos para preparar los datos a analizar.

De esta forma, hay algoritmos o técnicas que pueden servir para distintos propósitos, por lo que en la tabla1.1 se representan para qué propósito son más utilizadas las técnicas. Por ejemplo, las redes de neuronas pueden servir para predicción, clasificación e incluso para aprendizaje no supervisado.

Tabla 1.1

	Análisis supervisado	**Análisis no supervisado**
	Técnica de minería	
Modelos de predicción	Árboles de decisión	No factible
	Redes neuronales	
	Regresión	
Segmentación	Árboles de decisión	Análisis de clusters
	Redes neuronales	

2.2.5. Evaluar

Finalmente, la última fase del proceso consiste en la valoración de los resultados mediante el análisis de bondad del modelo o modelos, contrastado con otros métodos estadísticos o con nuevas poblaciones muéstrales.

3. COMPRENSIÓN Y PREPARACIÓN DE LOS DATOS

Este paso ocupa la mayor parte del tiempo dedicado a los proyectos de minería de datos. La tarea de capturar datos para análisis se puede complicar ya que para muchos proyectos, los datos deben ser recuperados de varias fuentes combinadas y limpias, antes de realizar algún análisis.

Algunos proyectos pueden tener los datos requeridos; almacenados en un data warehouse o data mart. Esto reduce dramáticamente los tiempos de acceso y esfuerzos, y resulta uno de los motivos a considerar al evaluar su construcción.

En pocas palabras, a menos que ya exista en funcionamiento un data warehouse o mart relevante, el tiempo dedicado al proyecto en tareas de acceso, transformación y limpieza de datos, será entre un 50% y un 80%.

La mayoría de los sistemas de minería datos requieren archivos denormalizados (rectangulares) para su análisis

- Un archivo de datos rectangular o regular tiene las siguientes características:
- Todos los registros (casos) contienen los mismos campos de datos (variables o atributos)
- Todos los registros (casos) contienen los mismos campos de datos (variables o atributos) en el mismo orden.
- Todos los registros (casos) contienen información completa en todos sus campos de datos (variables) [pero rara vez sucede en la práctica]

En el rubro de la minería datos, un archivo de datos de estas características se conoce como “denormalizado.” Si bien archivo puede no poseer esta estructura

inicialmente, generalmente deberá ser convertido a la misma antes de comenzar el proceso analítico.

3.1. TAMAÑO DEL ARCHIVO DE DATOS

Los archivos de datos poseen dos dimensiones: registros (casos) y campos (variables o atributos) Los archivos de datos podrán ser de tamaño considerable debido a que alguna o ambas dimensiones poseen esa característica. Por ejemplo, un archivo transaccional de llamados telefónicos realizados en un área tendría un número limitado de campos (tal vez los números de origen y destino del llamado, duración del mismo, e información de facturación, entre otros), pero posiblemente contenga decenas de millones de registros (o llamadas telefónicas). Un banco podrá compilar un registro para cada cliente conteniendo información de actividad para cada cuenta con una frecuencia mensual.

Dependiendo de la ventana temporal y del número de cuentas, es posible que implique cientos y hasta miles de variables.

También es importante denotar la diferencia entre el tamaño del archivo de datos utilizado para la generación del modelo y el tamaño del archivo al cual será aplicado el modelo. En términos generales no hay limites prácticos al tamaño de este último, por lo que el modelo podrá ser ejecutado con grandes bases de datos.

3.1.1. Selección de datos

Es muy raro que sean utilizados absolutamente todos los datos (todos los registros y campos) de una base de datos, data warehouse, data mart en el proceso analítico del minería datos. A continuación mencionamos algunas consideraciones a tomar en cuenta al seleccionar datos o una muestra de los datos para su análisis.

La frase en inglés “garbage in, garbage out” (basura de entrada, basura a la salida) es relevante en el contexto de la minería datos. Debe dedicarse algún esfuerzo al decidir que campos de datos (variables) serán incluidos en el análisis. Las variables que sean redundantes (parciales facturados y facturación total) y las que no tienen relevancia lógica (ID o código de identificación) deberán ser excluidas. De la misma manera, al construir modelos predictivos, deberán ser utilizadas como variables predictoras aquellas que sean resultantes de los mismos. Por ejemplo, todos aquellos que actualizan un sistema informático deben, inicialmente, haber adquirido el producto. Esto puede ocurrir cuando son extraídos de una base de datos campos que están relacionados entre sí.

Estas decisiones por lo general se basan en un conocimiento general del negocio que sirve de contexto para el análisis.
Resumiendo, los campos de datos utilizados en el análisis deberán tener sentido en el contexto del negocio.

Si los registros poseen un componente temporal, entonces habrá que tomar decisiones concerniendo el rango temporal de los datos a analizar. Muchas veces esto surge directamente de la lógica del negocio. Por ejemplo, algunas industrias sufrieron hace pocos años un proceso desregulatorio, y solo resulta relevante incluir los datos que han sido capturados a partir de entonces al realizar procesos analíticos. O, si estuviésemos estudiando la tasa de renovación de contratos para algún servicio en particular, no sería correcto incluir aquellos contratos que han sido firmados recientemente en el cálculo debido a que aún no les corresponde renovación alguna.

Por otro lado el sobre muestreo de casos poco frecuentes (por ejemplo, respuestas positivas a una campaña de marketing directo o casos de enfermedades poco frecuentes), suele hacerse con el propósito de obtener una mayor precisión al identificar estos casos en particular. Al muestrear de esta manera resulta importante expresar dichos resultados que dependen en el mecanismo de muestreo, por

ejemplo una tabla de clasificación errónea, en términos de la distribución real y no de la distribución sobre muestreada. Esto se debe a que el modelo será aplicado a la población original y no a una población en la que un caso raro ocurra con frecuencia.

3.1.2. Revisando y transformando datos

Para alcanzar buenos resultados es necesario comprender que la minería de datos no se basa en una metodología estándar y genérica que resuelve todo tipo de problemas, sino que consiste en una metodología dinámica e iterativa que va a depender del problema planteado, de la disponibilidad de la fuente de datos, del conocimiento de las herramientas necesarias, de la metodología desarrollada, y de los requerimientos y recursos de la empresa.

El procedimiento para resolver un problema a través de la minería de datos se divide en dos grandes etapas: la preparación de los datos y la minería de datos propiamente dicha.

La preparación de los datos consume de un 60% a 90% del tiempo en un proyecto de minería de datos y contribuye de un 75% a 90% al éxito del proyecto; es muy importante ya que para cada algoritmo de minería se deben cumplir ciertas condiciones; y así encontrar un patrón en ellos.

La estadística juega un importante papel en el análisis de los datos; algunos paquetes estadísticos que han sido utilizados durante mucho tiempo (para generar promedios, sumas, y diferentes distribuciones para diferentes aplicaciones), se han integrado con las diferentes bases de datos, y se están comercializándose en la actualidad como productos para la Minería de Datos.
Los datos del mundo real son, incompletos y en muchos casos sin atributos de interés, o sólo se dispone de datos que tienen:

- Ruido: errores y outliers
- Inconsistencia: contienen discrepancias en esquema o en datos por actualizaciones inconsistentes

Sin datos de calidad (datos limpios), no hay buenos resultados; los datos deben ser consistentes; por eso es tan importante la fase de preparación de datos, los principales pasos son:

- Planteamiento del problema: Definir de manera objetiva cuál es el problema a resolver, determinar con qué recursos humanos y tecnológicos se cuenta, cuáles son las fuentes de información y cuál es la disponibilidad de la información.

- Selección de los datos: De todas las fuentes de información disponibles se debe establecer cuáles son las que se van a considerar. Es decir, se decide sobre qué datos se va a trabajar, tanto desde el punto de vista físico, como desde el punto de vista lógico. Se debe realizar un tratamiento y estructuración de la información con el objetivo de presentarla de la mejor manera posible para posteriores análisis.

- Limpieza y preprocesamiento de los datos: En esta fase se analizan los datos con la finalidad de reorganizar la información eliminando aquella que es poco útil o completando la que nos falta. Se eliminan los datos irrelevantes, se unifican los criterios de representación que pueden no ser los mismos en todas las fuentes de datos y se eliminan redundancias y duplicados.

- Reducción y proyección de datos: Consiste en encontrar las características útiles que representan las dependencias de los datos en el objetivo del proceso.

El punto es que la limpieza de los datos y las soluciones de problemas deberán ser completadas antes de comenzar la fase analítica.

3.2. TIPOS DE DATOS

Antes de empezar la limpieza de datos; es importante identificar los tipos de datos y niveles de medición y de esta manera dar el conjunto de sus valores permisibles.

Se distinguen dos tipos de variables: las discretas y las continuas(ver cuadro 3.1). Como consecuencia de la diferencias que existen entre las variables continuas y discretas, en lo que a propiedades se refiere, es necesario tener en cuenta las escalas de medición, lo que a su vez tiene implicación en el tipo de prueba estadística a seleccionar.

Cuadro 3.1. Tipo de datos

TIPO DE DATOS		
Cualitativos o de atributo	Cuantitativos o Numéricos	
	Discretos	Continuos

Los datos se clasifican de acuerdo a su nivel de medición. El nivel de medición de un conjunto de datos determina los cálculos que se pueden realizar para resumir y presentar la información correspondiente a estos datos, así como a las pruebas estadísticas que se pueden aplicar.

Existen cuatro niveles de medición:

1. Nivel nominal
2. Nivel ordinal
3. Nivel de intervalo
4. Nivel de razón

3.2.1. Datos de nivel nominal

En el nivel de medición nominal no se define un orden o relación lógica para las variables, las observaciones solamente pueden clasificarse o contarse. Ejemplos de variables con nivel de medición nominal son las siguientes:

- Color del objeto
- GENERO de la persona
- Tipo de alimento
- Tipo de escuela
- Afiliación política
- Tipo de religión
- Tipo de tarjeta de crédito
- Grado de acuerdo o desacuerdo frente a una afirmación

3.2.2. Datos de nivel ordinal

En el nivel de medición ordinal los datos se clasifican en categorías, las cuales mantienen entre las mismas un orden lógico. Los diferentes valores que puede tomar la variable vienen jerarquizados de acuerdo a un rango.

Las categorías en las que se clasifican los datos son excluyentes y exhaustivas. Estas categorías se clasifican/ordenan de acuerdo con las características particulares que poseen. Ejemplos de variables con nivel de medición ordinal son las siguientes:

- Calificación en un examen (aprobado, regular, bien, muy bien, excelente)
- Nivel de escolaridad (primaria, secundaria, preparatoria, licenciatura, maestría, doctorado)

- Categoría de una plaza de profesor-investigador (asociado A, asociado B, asociado C, asociado D, titular A, titular B, titular C)

3.2.3. Datos de nivel de intervalo

El nivel de medición de intervalo incluye todas las características del nivel ordinal, añadiendo además la característica de que la distancia entre los intervalos está claramente determinada y que éstos son iguales entre sí.

Un ejemplo de escala de intervalos iguales está dado por la escala temperatura. Por ejemplo, entre 18 y 19 grados centígrados existe la misma diferencia que hay entre 31 y 32 grados centígrados. Otra característica de los datos de nivel ordinal es que las diferencias iguales en la característica son representadas por diferencias iguales en los números asignados a las categorías.

Otros ejemplos de variables con nivel de medición de intervalo son: número de hijos, peso, estatura y presión arterial.

3.2.4. Datos de nivel de razón

El nivel de medición de razón posee todas las características del nivel de intervalo, añadiendo además otras dos características: (1) la existencia de un valor cero (0) real, el cual representa la ausencia de la característica, y (2) la importancia de la razón o cociente entre dos números. Ejemplos de variables con nivel de medición de razón son las siguientes.

- Variables del mundo físico (longitud, área, masa, intensidad)
- Salarios
- Unidades de producción
- Peso y estatura

3.3. LIMPIEZA DE DATOS

Una vez que se genera el archivo de datos para el análisis, el próximo paso generalmente resulta ser la exploración de los datos de diversas maneras. Es posible generar un resumen de estadísticos para cada campo. Es importante revisar la existencia de que ciertos campos posean relaciones lógicas entre sí (por ejemplo, ventas totales debe ser igual o superior a cualquiera de sus ventas parciales). Valores de variables extraños o extremos deberán ser identificados y deberá decidirse de qué manera tratarlos.

3.3.1. Valores perdidos

Los valores perdidos deberán ser identificados y marcados en el archivo de análisis. Si bien los códigos para los valores perdidos de cada campo (atributo, variable) deberán ser registrados como parte de la auditoria de datos será necesario obtener estadísticos descriptivos o realizar un análisis exploratorio complementario para asegurar la consistencia de los datos. Deberá verificar que los valores perdidos están siendo identificados por el sistema analítico así como examinar el volumen de datos faltantes. Específicamente entender, dados la fuente de datos y su método de captura, si la cantidad de datos faltantes para cada campo es razonable.

Por ejemplo, no sería sorprendente encontrar un faltante del 20% de los datos demográficos frente a la variable “edad al casarse por primera vez” (dado que muchos de los encuestados aún deben ser solteros), pero seria extremadamente sospechoso descubrir un faltante del 20% de los totales de facturas. Aquí es donde el conocimiento del negocio y de las fuentes de datos es fundamental para evaluar la situación.

Los valores perdidos también podrán ser examinados un registro a la vez. En este caso el interés estaría centrado en identificar aquellos registros que contienen

muchos valores perdidos, para luego evaluar si vale la pena que tales registros sean considerados en el proceso analítico.

3.3.2. Valores incorrectos

El archivo de datos podrá contener valores que son incorrectos pero que no han sido codificados como perdidos. Si los valores registrados se encuentran dentro del rango de valores normales, resulta complicado detectar los errores mencionados. En algunos casos puede ser utilizada una fuente de datos alternativa para validar los valores de ciertas variables o campos.

Los valores que se encuentran fuera del rango norma! pueden ser marcados para ser revisados con posterioridad.

Por ejemplo, la edad al casarse por primera vez de un cliente, codificado como 5 años, indica un error de datos en muchas sociedades. Siempre que sean conservados los documentos originales es posible extraer una pequeña muestra y revisarla. De todas maneras esto resulta ser una labor intensiva y tiende a identificar mayormente errores de carácter sistémico.

3.3.3. Valores consistentes entre fuentes de datos

Cuando se utiliza más de una base de datos por lo general algunos de los campos de datos (variables, atributos) se encontrarán duplicados en las diferentes fuentes. Dichos campos podrán ser comparados como método de validación y podrán ser definidas algunas reglas que determinen el valor a tomar en cuenta en caso de discrepancias. Este es uno de los pasos al construir un data warehouse y se detalla en las referencias.

3.3.4. Confirmaciones lógicas a través de los campos

Las relaciones que deberían existir entre los valores de diferentes campos podrán ser examinados, identificando con seguridad algunas excepciones. Por ejemplo, en una encuesta es necesario que el valor correspondiente a la edad del encuestado sea superior o igual a su edad al casarse por primera vez. A menos que existan descuentos, el total de una factura no podrá ser inferior al monto asociado a cualquiera de los artículos que la componen.

3.3.5. Valores extremos

Valores extremos son valores de datos que se encuentran a una distancia considerable del centro de la distribución de los datos en su conjunto. Si bien son poco frecuentes, podrán influir desproporcionadamente en los procesos de modelación, por lo que deberán ser identificados. En algunos casos los valores extremos son simplemente errores en los datos, pero en otros se tratan de casos inusuales válidos (por ejemplo, gran cantidad de pasajeros vuelan diariamente desde el aeropuerto de Chicago, pero muy pocos lo hicieron el primero de enero de 1999, debido a una tormenta de nieve que afectó el tránsito normal)

La decisión de si los valores extremos deberán ser incluidos, excluidos o modificados para el análisis dependerá mayormente del propósito de dicho análisis. Por ejemplo, si quisiéramos construir un modelo predictivo simple de transito de pasajeros de aerolíneas basado en los días de la semana, entonces el día de la tormenta de nieve podría ser omitido. Esto se debe a que una tormenta de nieve es un evento devastador de rara ocurrencia que no está relacionado naturalmente con día alguno de la semana. Ahora, si lo que queremos es construir un modelo predictivo que relacione las condiciones climáticas con la facturación de la aerolínea, entonces los datos del día en que ocurrió la tormenta de nieve bien podrían ser incluidos.

3.3.6. Valores extremos estadísticos

Los valores extremos estadísticos son puntos ubicados lejos del centro de los datos, basándonos en una medida de distancia. Comúnmente los valores extremos son definidos como dos (estudios sociales) o tres (ingeniería) desviaciones estándar de la media. También están disponibles otras medidas (por ejemplo, rangos intercuartiles) Los sistemas para análisis estadístico cuentan, por lo general, con uno o más de estos métodos para la detección de valores extremos.

3.3.7. Valores extremos culturales

A falta de una definición mejor, los valores extremos culturales son valores inusuales en el entorno cultural o de negocios. Podrán ser también valores extremos estadísticos, pero no es imperativo que lo sean. Como ejemplo, tomando los datos de consumo en Estados Unidos, consideremos a alguien con una edad al casarse por primera vez de 13 años. Este valor podrá no ser un valor extremo estadístico, dado que la edad promedio al casarse por primera vez es de cerca de 20 años y muchos se casan a los 18 y 19 años. De todas maneras, es muy raro desde el punto de vista social que alguien se case a los 13 años en los Estados Unidos.

Otro ejemplo podría ser el estudio de visitas a mini mercados en los Estados Unidos (en los que se venden leche, gaseosas, café, sandwiches, golosinas y artículos de limpieza para el hogar). Mientras que el número de visitas por semana promediaba en 2, un subgrupo reveló visitar el mini mercado cerca de 20 veces por semana. Luego de examinar el caso con cuidado, se descubrió que el subgrupo estaba compuesto por adultos mayores, sin pareja, que recurrían al lugar varias veces al día para obtener casi todos sus alimentos. Este subgrupo constituyó un valor extremo interesante desde muchos puntos de vista.

3.4. MÉTODOS DE REDUCCIÓN DE DATOS

3.4.1. Componentes principales

El análisis de componentes principales es una técnica estadística que crea variables nuevas, compuestas mediante la aplicación de pesos diferenciales al combinar las variables originales. Los pesos son derivados de manera que cada nueva variable compuesta maximice la variación restante de las variables originales. De esta manera las variables compuestas captura eficientemente la variación entre las variables iniciales. Las variables con alto grado de correlación entre sí tienden a pesados en el mismo componente y se ven muy representadas por esa variable compuesta. Dado que los componentes principales se basan en la covarianza entre variables, las variables se presumen como continuas (no categóricas).

Luego de ejecutar el análisis de componentes principales, las variables originales se ven reemplazadas por un número muy inferior de variables compuestas en el análisis de minería datos.

Existe un posible ángulo débil del análisis de componente principal cuando es llevado a cabo antes de construir un modelo predictivo. Los componentes tienden a ser construidos a partir de las relaciones entre las variables en el análisis de componentes, y no en una forma que maximice la precisión predictiva posterior.

3.4.2. Selección de datos antes de la transformación

Es importante destacar que cuando corresponda seleccionar datos o efectuar un muestreo, de ser posible deberá realizarse antes de efectuar las transformaciones. Esto acelerará el procesamiento ya que las transformaciones no serán aplicadas a aquellos registros que eventualmente serán abandonados por el proceso analítico. Probablemente sea más eficiente efectuar la selección de los datos a nivel de la base de datos o del warehouse, siempre que su lenguaje SQL soporte las operaciones (tal

vez la generación de números seudo-aleatorios para el muestreo aleatorio) necesarias para la selección de datos.

4. TÉCNICAS DE MINERÍA DE DATOS

Las técnicas de Minería de Datos intentan obtener patrones o modelos a partir de los datos recopilados. Decidir si los modelos obtenidos son útiles o no suele requerir una valoración subjetiva por parte del usuario. Las técnicas de Minería de Datos se clasifican en dos grandes categorías: supervisadas o predictivas y no supervisadas o descriptivas.

4.1. APRENDIZAJE SUPERVISADO Y NO SUPERVISADO

En el aprendizaje supervisado existe un atributo especial, normalmente denominado clase, presente en todos los ejemplos que especifica si el ejemplo pertenece o no a un cierto concepto, que será el objetivo del aprendizaje. El atributo clase normalmente toma los valores + y -, que significan la pertenencia o no del ejemplo al concepto que se trata de aprender; es decir, que el ejemplo ejemplifica positivamente al concepto -pertenece al concepto- o bien lo ejemplifica negativamente -que no pertenece al concepto. Mediante una generalización del papel del atributo clase, cualquier atributo puede desempeñar ese papel, convirtiéndose la clasificación de los ejemplos según los valores del atributo en cuestión, en el objeto del aprendizaje. Expresado en una forma breve, el objetivo del aprendizaje supervisado es: a partir de un conjunto de ejemplos, denominados de entrenamiento, de un cierto dominio D de ellos, construir criterios para determinar el valor del atributo clase en un ejemplo cualquiera del dominio. Esos criterios están basados en los valores de uno o varios de los otros pares (atributo; valor) que intervienen en la definición de los ejemplos. Es sencillo transmitir esa idea al caso en el que el atributo que juega el papel de la clase sea uno cualquiera o con más de dos valores. Dentro de este tipo de aprendizaje se pueden distinguir dos grandes grupos de técnicas: la predicción y la clasificación.

Tecnicas de Análisis supervisado para modelos de predicción:

- Árboles de decisión
- Redes neuronales
- Regresión

El aprendizaje no supervisado estudia el aprendizaje sin la ayuda del maestro; es decir, se aborda el aprendizaje sin supervisión, que trata de ordenar los ejemplos en una jerarquía según las regularidades en la distribución de los pares atributo-valor sin la guía del atributo especial clase. Éste es el proceder de los sistemas que realizan clustering conceptual y de los que se dice también que adquieren nuevos conceptos. Otra posibilidad contemplada para estos sistemas es la de sintetizar conocimiento cualitativo o cuantitativo, objetivo de los sistemas que llevan a cabo tareas de descubrimiento.

Técnicas de Análisis no supervisado para modelos de segmentación:

- Análisis de cluster
- Redes neuronales

4.2. ANÁLISIS DE CLUSTER (SEGMENTACIÓN)

También llamada agrupamiento, permite la identificación de tipologías o grupos donde los elementos guardan gran similitud entre sí y muchas diferencias con los de otros grupos. Así se puede segmentar el colectivo de clientes, el conjunto de valores e índices financieros, el espectro de observaciones astronómicas, el conjunto de zonas forestales, el conjunto de empleados y de sucursales u oficinas, etc. Este análisis se conoce también como análisis de clasificación o taxonomía numérica. Nos ocupamos de los procedimientos de conglomerados que asignan cada objeto a un solo grupo. La Figura 4.1 muestra un caso de conglomerado ideal en el que los grupos se separan en dos variables: conciencia de calidad (EJE X) y susceptibilidad

al precio (EJE Y). Nótese que cada consumidor pertenece a un grupo y no existen áreas que se superpongan. Por otra parte, la Figura 4.1 presenta el caso de una agrupación que puede encontrarse en la realidad. Las fronteras de algunos de los grupos no están definidas con claridad y la clasificación de algunos consumidores no es obvia porque muchos de ellos podrían agruparse en un grupo u otro.

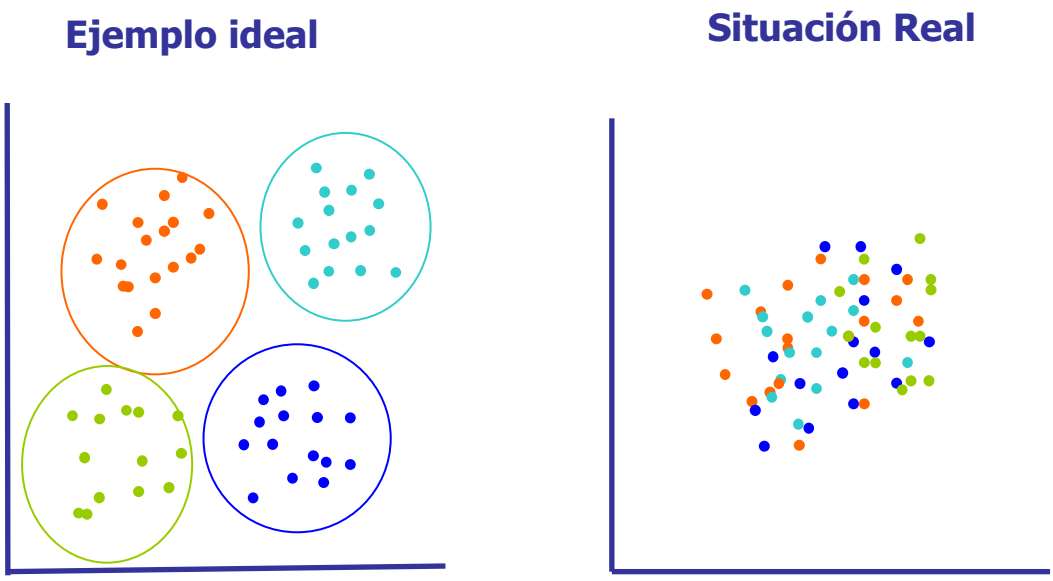


Figura 4.1. Conglomerado Ideal / Conglomerado Real

La segmentación está teniendo mucho interés desde hace ya tiempo dadas las importantes ventajas que aporta al permitir el tratamiento de grandes colectivos de forma pseudoparticularizada, en el más idóneo punto de equilibrio entre el tratamiento individualizado y aquel totalmente masificado. Las herramientas de segmentación se basan en técnicas de carácter estadístico, de empleo de algoritmos matemáticos, de generación de reglas y de redes neuronales para el tratamiento de registros.

Para otro tipo de elementos a agrupar o segmentar, como texto y documentos, se usan técnicas de reconocimiento de conceptos. Esta técnica suele servir de punto de

partida para después hacer un análisis de clasificación sobre los clusters. La principal característica de esta técnica es la utilización de una medida de similaridad que, en general, está basada en los atributos que describen a los objetos, y se define usualmente por proximidad en un espacio multidimensional. Exiten varios algoritmos de clustering, a continuación se exponen los mas conocidos:

4.3. MÉTODOS DE CLUSTER JERÁRQUICOS

En la práctica, no se pueden examinar todas las posibilidades de agrupar los elementos, incluso con los ordenadores más rápidos. Una solución se encuentra en los llamados métodos jerárquicos. Se tienen dos posibles formas de actuar, a continuación se muestra una tabla con los metodos de cluster jerarquicos y de partición (tabla 4.1):

Tabla 4.1 Métodos de clusters

<table>
<tr><th colspan="2">Métodos de clusters</th><th>Métodos de enlace</th></tr>
<tr><td rowspan="2">Métodos de cluster jerarquico</td><td>Método jerarquico aglomerativo</td><td rowspan="2">Sencillo
Completo
Promedio
Centroide</td></tr>
<tr><td>Método dividido</td></tr>
<tr><td>Método de partición</td><td>KMEDIAS</td><td>Algoritmo kmedias</td></tr>
</table>

Métodos jerárquicos aglomerativos: se comienza con los objetos o individuos de modo individual; de este modo, se tienen tantos clusters iniciales como objetos. Luego se van agrupando de modo que los primeros en hacerlo son los más similares y al final, todos los subgrupos se unen en un único cluster (ver figura 4.2).

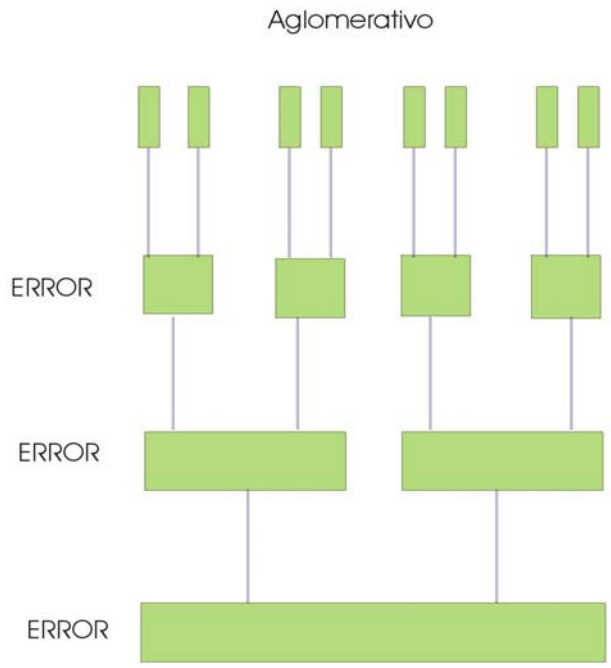


Figura 4.2. Metodo Aglomerativo

Métodos jerárquicos divididos: se actúa al contrario. Se parte de un grupo único con todas las observaciones y se van dividiendo según lo lejanos que estén(ver figura 4.3).

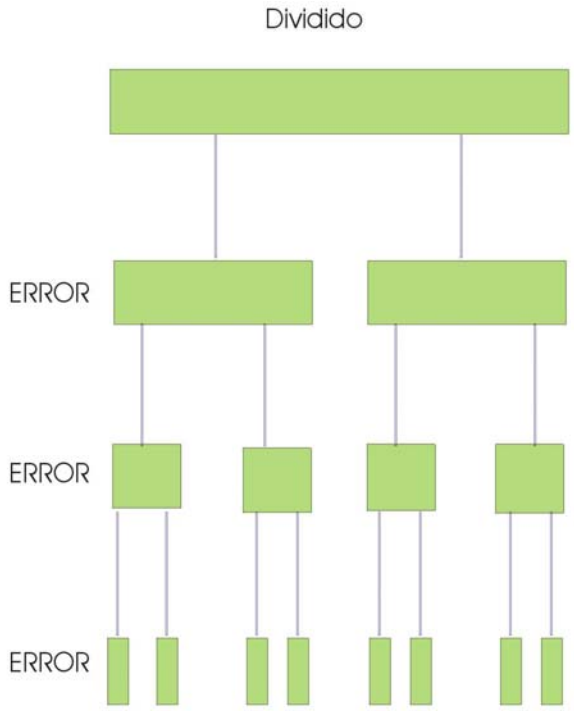


Figura 4.3 Metodo Dividido

Consideramos aquí los métodos aglomerativos con diferentes métodos de unión (linkage methods). Los más importantes son:

El método de enlace sencillo se basa en la distancia mínima o la regla del vecino más próximo. Los primeros dos objetos conglomerados son aquellos que tienen la menor distancia entre sí. La siguiente distancia más corta se identifica, ya sea que el tercer objeto se agrupe con los dos primeros o que se forme un nuevo conglomerado de dos objetos. En cada etapa, la distancia entre dos conglomerados es la distancia entre sus dos puntos más próximos (véase Figura 4.4). En cualquier etapa, dos conglomerados surgen por el enlace sencillo más corto entre éstos. Este proceso continúa hasta que todos los objetos se encuentren en un conglomerado. El método del enlace sencillo no funciona adecuadamente cuando los conglomerados no están bien definidos.

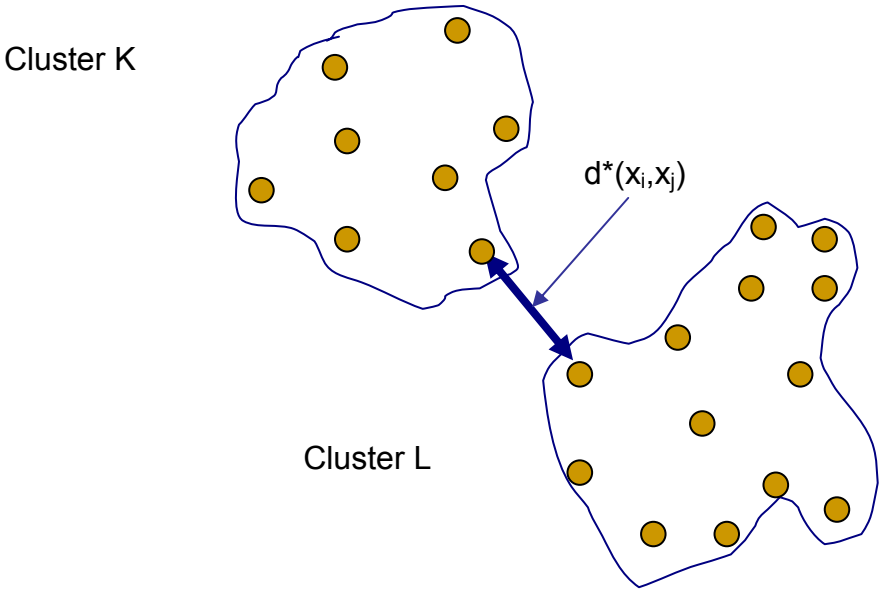


Figura 4.4. Enlace Sencillo

El método del enlace completo es similar al enlace sencillo, excepto que se basa en la distancia máxima o la estrategia del vecino más lejano. En el enlace completo, la distancia entre dos conglomerados se calcula como la distancia entre sus puntos más lejanos (ver figura 4.5).

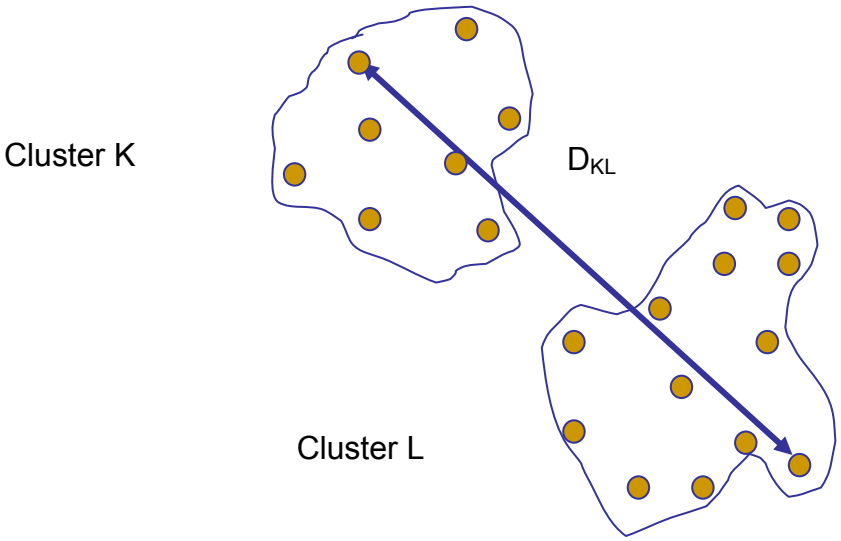


Figura 4.5. Enlace Completo

El método del enlace promedio funciona de manera similar, pero en este método, la distancia entre dos conglomerados se define como el promedio de las distancias

entre todos los pares de objetos, donde se encuentra un miembro del par de cada uno de los conglomerados (Figura 4.6). Como puede observarse, el método del enlace promedio emplea la información sobre todos los pares de distancias, no sólo las mínimas o máximas. Por esta razón, generalmente se prefiere a los métodos de enlace sencillo y completo.

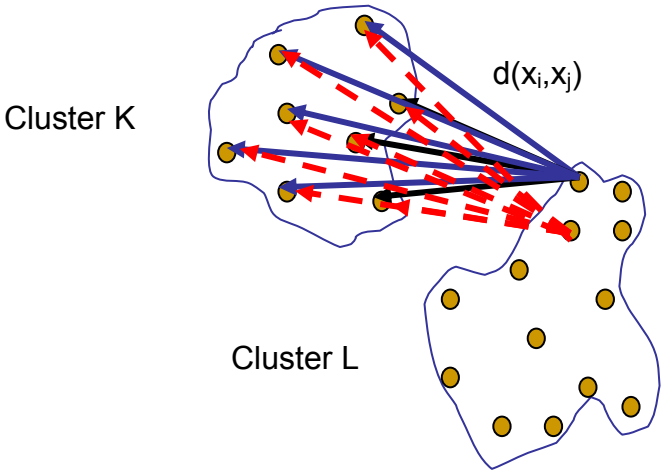


Figura 4.6. Enlace Promedio

En el método centroide, la distancia entre dos grupos es la distancia entre sus centroides (medias para todas las variables), como se muestra en la Figura 4.7. Cada vez que se agrupan los objetos, se calcula un centroide nuevo.

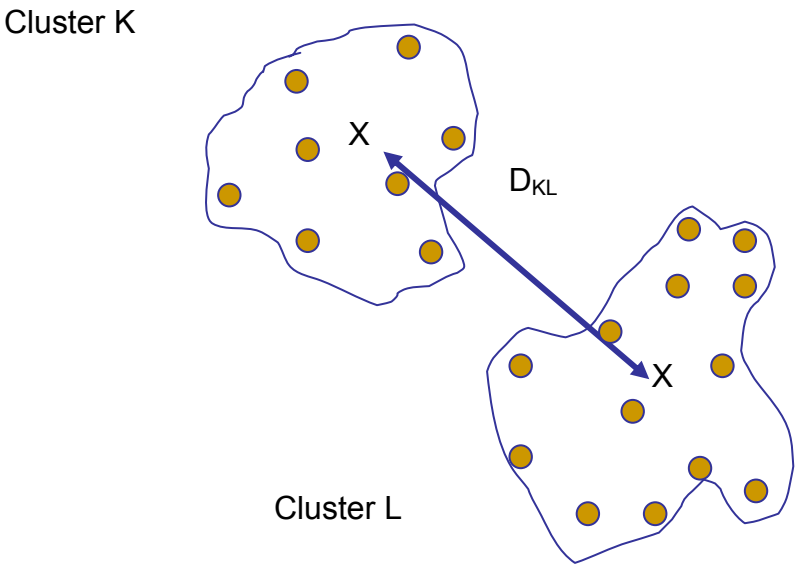


Figura 4.7. Método Centroide

4.4. MÉTODOS DE PARTICIÓN

El objetivo de estos clusters es minimizar la distancia de objetos ENTRE cluster y maximizar la distancia entre clusters (ver figura 4.8).

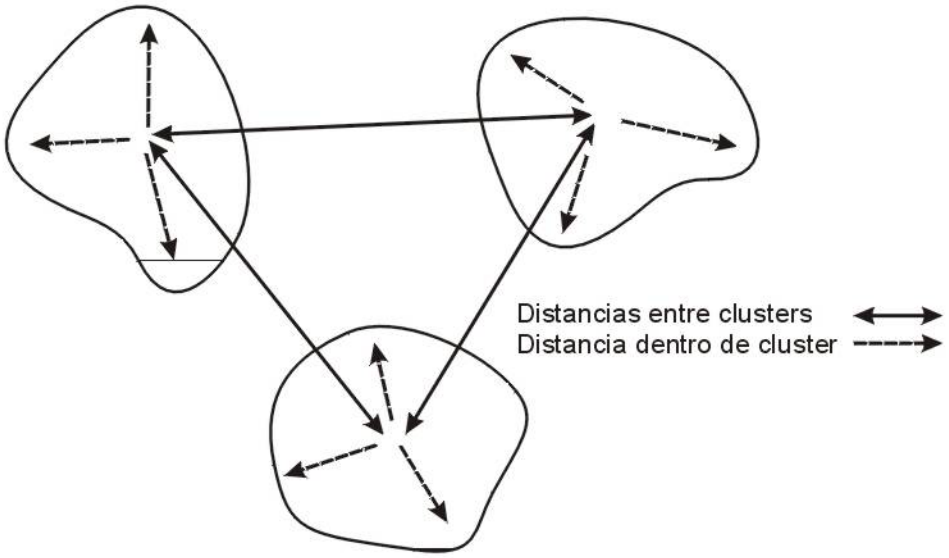


Figura 4.8. Distancias entre clusters

4.5. K MEDIAS

El algoritmo k-Medias es uno de los mas usados; el "K" de su nombre se refiere al hecho de que el algoritmo busca un número fijo de clusters que son definidos en términos de proximidad entre datos.

La versión descrita fue publicada por primera vez por la J. B. MacQueen en 1967. Para facilitar la explicación, la técnica se ilustra usando diagramas bidimensionales. Tenga en cuenta que en la práctica el algoritmo por lo general maneja mucho más de dos variables independientes. Esto significa que en vez de puntos correspondiente a vectores de dos elementos (x1, x2), los puntos corresponden al vector de n elementos (x1, x2..., xn). El procedimiento en sí mismo es inalterado.

Pasos del Algoritmo k medias

1. Seleccionar al azar k semillas. El algoritmo de MacQueen simplemente toma los primeros k registros aunque puede ser deseable escoger registros espaciados, o una selección hecha al azar de ellos. Cada una de las semillas es un cluster embrionario con un sólo elemento. Este ejemplo pone el número de clusters a 3.

2. Asigna cada registro a la semilla más cercana. Un modo de hacer esto es encontrando las fronteras entre los clusters, como se muestra en la figura 4.9. Las fronteras entre dos clusters son los puntos que están igualmente cerca a cada cluster. Considerando dos punto cualquiera, A y B, todos los puntos que son equidistantes a A y B a lo largo de una línea que es perpendicular a la conexión de A y B y a mitad de camino entre ellos. En la Figura 4.9, las líneas punteadas unen las semillas iniciales; las fronteras de clusters son mostradas con líneas sólidas que son perpendiculares las líneas punteadas. Usando estas líneas como guías, es obvio cuales registros son los más cercanos a cierta semilla. En tres dimensiones, estas

fronteras serían planos y en N dimensiones serían hiperplanos de dimensión N - 1. Los algoritmos de ordenador fácilmente manejan estas situaciones. Definir las fronteras entre cluster es útil para mostrar el proceso geométricamente. Aunque en la práctica, el algoritmo mide la distancia de cada registro a cada semilla y escoge la distancia mínima para este paso. Por ejemplo, considere el registro con la caja dibujada alrededor de ello. Sobre la base de las semillas iniciales, este registro es asignado al cluster controlado por la semilla 2 porque es más cercano a aquella semilla que a cualquiera de otros dos. En este paso, cada registro ha sido asignado a uno de los tres clusters centrados alrededor de las semillas originales.

3. El tercer paso debe calcular el centroides de los clusters; estos ahora hacen un mejor trabajo de caracterizar los clusters que las semillas iniciales. Encuentrar los centroides es simplemente una manera de tomar el valor medio de cada dimensión para todos los registros en el cluster. En la Figura 4.10, los nuevos centroides son marcados con una cruz. Las flechas muestran el movimiento de la posición de las semillas originales a nuevo centroide de los clusters.

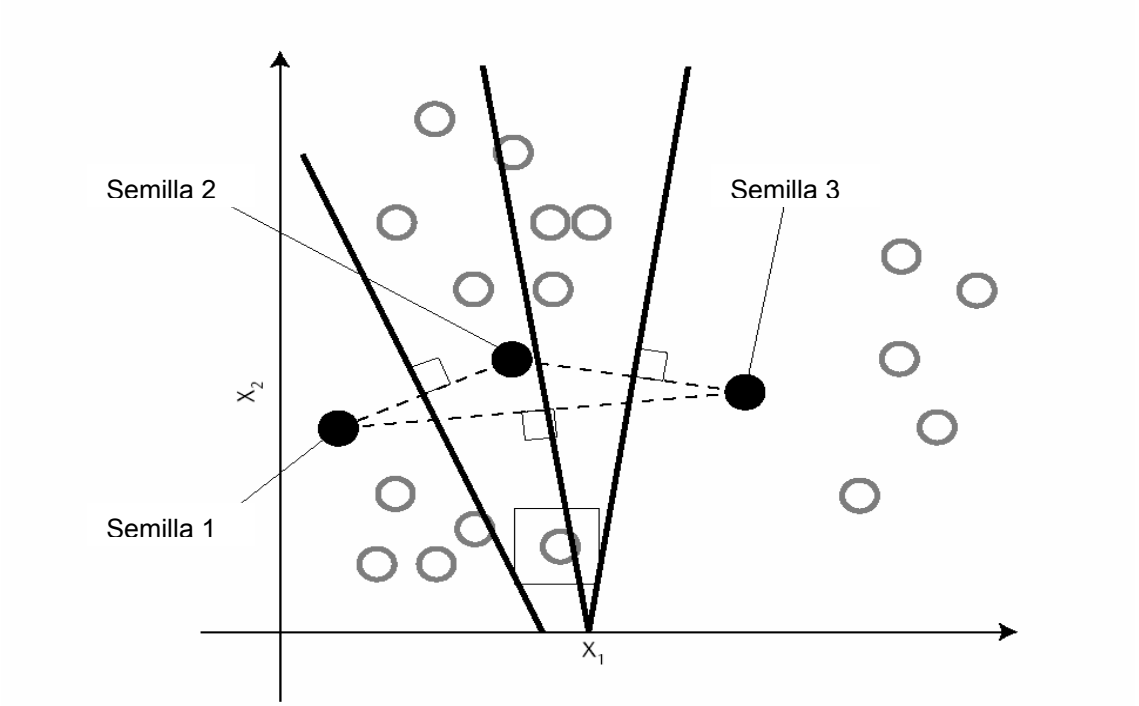


Figura 4.9. Las semillas iniciales determinan las fronteras de los clusters

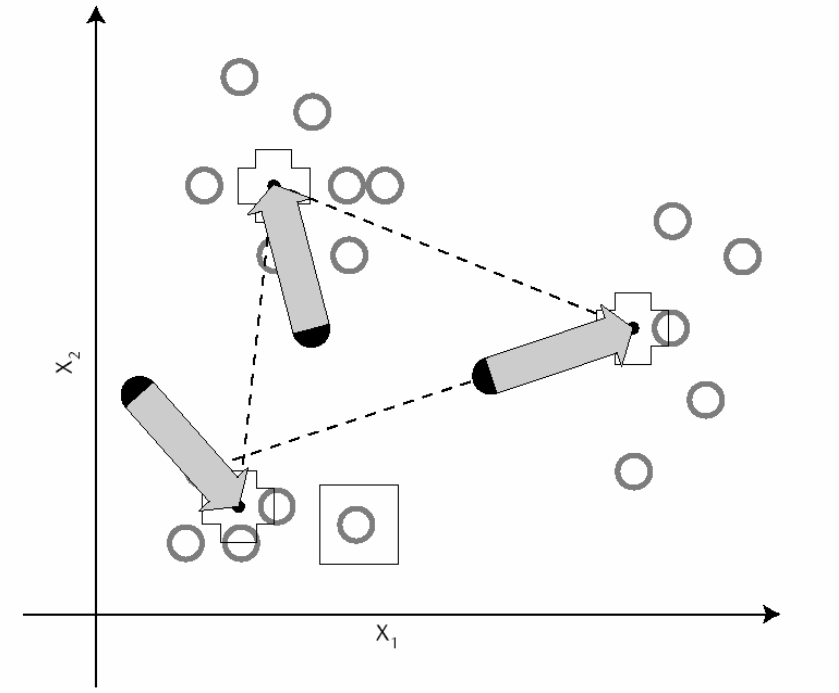


Figura 4.10. El centroide es calculado de los puntos que son asignados a cada cluster

Los centroides se convierten en las semillas en la siguiente iteración del algoritmo. El paso 2 es repetido, y cada punto otra vez es asignado al cluster con centroide más cercano. La figura 4.11 muestra el nuevo cluster formado por fronteras, como antes, dibujando líneas equidistantes entre cada par de centroides. Note que el punto con la caja alrededor de el, que al principio fue asignado al cluster 2, ahora ha sido asignado al cluster 1. El proceso de asignación indica el cluster y luego recalcula los centroides hasta que las fronteras de los clusters dejen de cambiarse. En la práctica, el algoritmo de k medias por lo general encuentra un juego de clusters estables después de unas docenas de iteraciones.

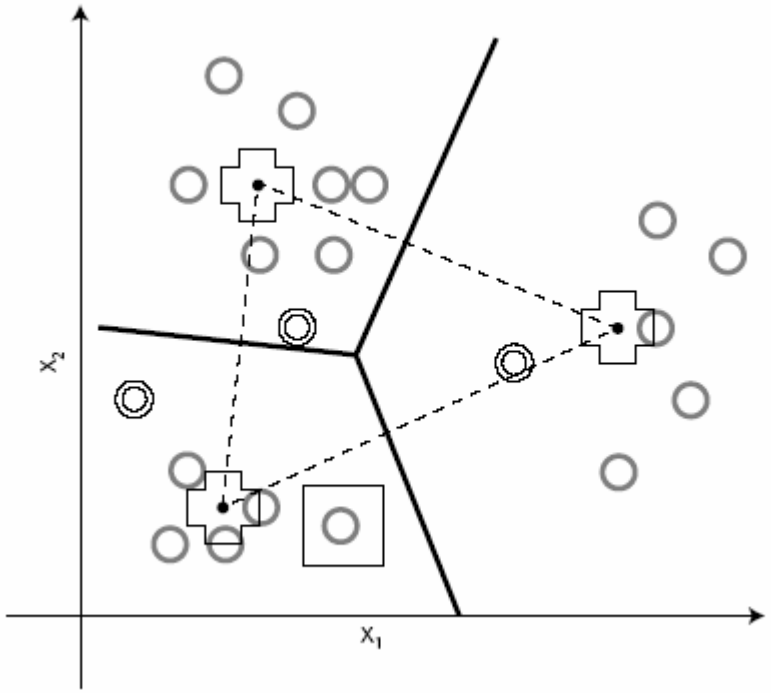


Figura 4.11. En cada iteración, todas las asignaciones de clusters son reconsideradas

Dos desventajas importantes de k medias son que el número de grupos debe especificarse previamente y que la selección de los centros de grupo es arbitraria. Además, los resultados del conglomerado pueden depender de la forma en que se seleccionan los centros. Muchos programas eligen los primeros k (k= número de

grupos) casos sin valores faltantes como los centros de grupo iniciales. De manera que, los resultados del conglomerado pueden depender del orden de las observaciones en los datos. No obstante, el cluster partición es más rápido que los métodos jerárquicos y es apropiado cuando el número de objetos u observaciones es alto.

Un aspecto importante en el análisis de conglomerados es decidir el número de éstos. A pesar de que no existe ninguna regla general y rápida, están disponibles algunos lineamientos.

Las consideraciones teóricas, conceptuales o prácticas pueden sugerir un número determinado de grupos. Por ejemplo, si el propósito de la agrupación es identificar los segmentos del mercado, es probable que la gerencia quiera un número de grupos en particular.

4.6. CÓMO REALIZAR EL ANÁLISIS DE CONGLOMERADOS

Los pasos que comprende la realización del análisis de conglomerados se mencionan en la Figura 4.12. El primer paso consiste en formular el problema de agrupación al definir las variables en las que se basa ésta. Después, debe seleccionarse una medida de distancia apropiada. La medida de distancia determina qué tan similares o diferentes son los objetos que se agrupan. Se han desarrollado varios procedimientos de agrupación. La decisión del número de conglomerados requiere del criterio del experto del negocio. Los conglomerados derivados deben interpretarse en términos de las variables utilizadas para formarlos, y deben perfilarse en términos de las variables sobresalientes adicionales. Por último, es preciso que se evalúe la validez del proceso de conglomerados.

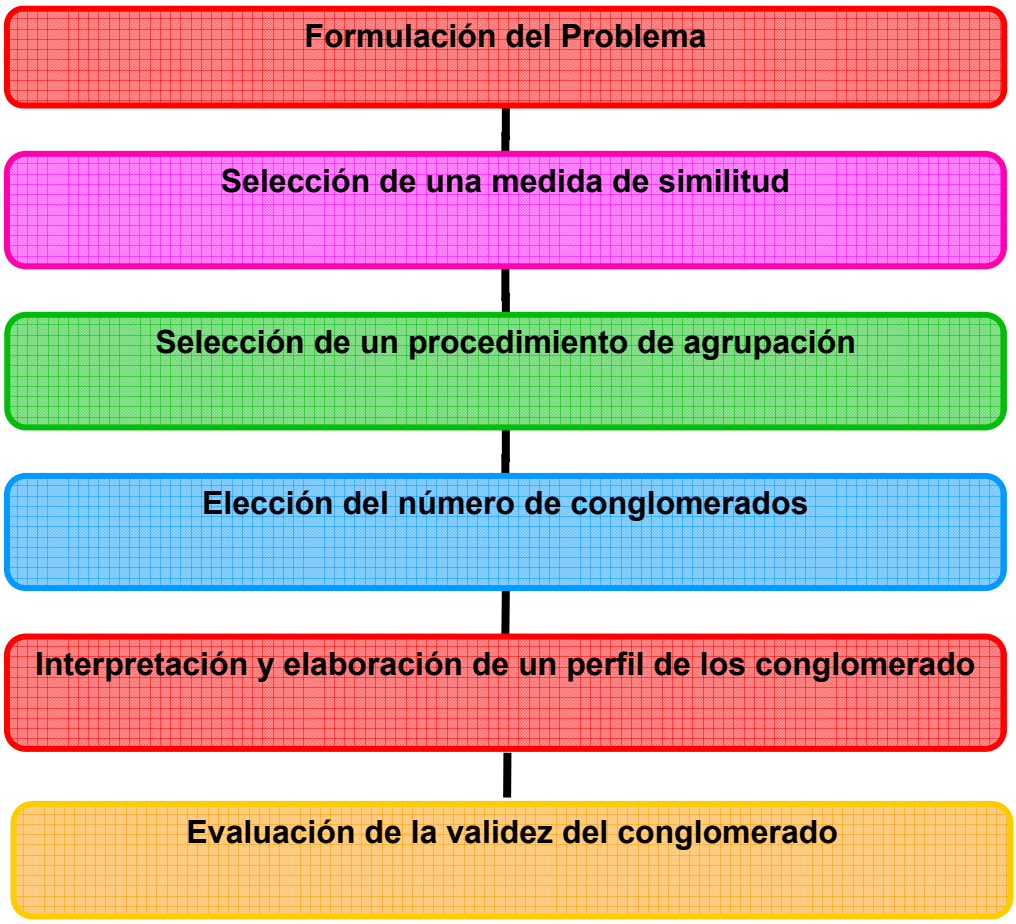


Figura 4.12. Pasos para realizar el análisis de conglomerados

4.6.1. Formulación del Problema

Quizá la parte más importante de la formulación del problema de conglomerados es la selección de las variables en las que se basa la agrupación. La inclusión de una o más variables irrelevantes puede distorsionar una solución de agrupación que de otra forma podría ser útil. Básicamente, el conjunto de variables seleccionado debe describir la similitud entre los objetos en términos relevantes para el problema de investigación de mercados. Las variables deben seleccionarse con base en la investigación previa, la teoría o una consideración de las hipótesis que se prueban. En la investigación exploratoria, se debe poner en práctica el criterio y la intuición.

4.6.2. Selección de la Medida de Distancia o Similitud

Ya que el objeto del conglomerado es agrupar objetos similares, se necesita alguna medida para evaluar las diferencias y similitudes entre objetos. La estrategia más común consiste en medir la equivalencia en términos de la distancia entre los pares de objetos. Los objetos con distancias reducidas entre ellos son más parecidos entre sí que aquellos que tienen distancias mayores. Existen varias formas de calcular las distancias entre dos objetos.

El uso de distintas medidas de distancia puede llevar a diversos resultados de conglomerado. Por consiguiente, se recomienda utilizar medidas diferentes y comparar los resultados. Después de seleccionar una medida de distancia o similitud, podemos elegir un procedimiento de agrupación.

4.6.3. Selección del Procedimiento de Aglomeración

Estos pueden ser jerárquicos o no. El conglomerado jerárquico se caracteriza por el desarrollo de una jerarquía o estructura en forma de árbol. A su vez, los métodos jerárquicos pueden ser:

- Análisis de Conglomerados por Aglomeración

El conglomerado por aglomeración empieza con cada objeto en un grupo separado. Los conglomerados se forman al agrupar los objetos en conjuntos cada vez más grandes. Este proceso continúa hasta que todos los objetos forman parte de un solo grupo.

- Análisis de Conglomerados por División

El conglomerado por división comienza con todos los objetos agrupados en un solo conjunto. Los conglomerados se dividen hasta que cada objeto sea un grupo independiente.

4.6.4. Elección del número de grupos

Un aspecto importante en el análisis de conglomerados es decidir el número de éstos. A pesar de que no existe ninguna regla general y rápida, están disponibles algunos lineamientos.

Las consideraciones teóricas, conceptuales o prácticas pueden sugerir un número determinado de grupos. Por ejemplo, si el propósito de la agrupación es identificar los segmentos del mercado, es probable que la gerencia quiera un número de grupos en particular.

En el conglomerado jerárquico, las distancias en las que los grupos se combinan pueden utilizarse como criterios. Esta información puede obtenerse del programa de aglomeración.

4.6.5. Interpretación y perfil de los grupos

La interpretación y el perfil de los grupos comprende el análisis de los centroides de grupo. Los centroides representan los valores medios de los objetos que contiene el grupo en cada una de las variables. Los centroides nos permiten describir cada grupo al asignarle un nombre o etiqueta.

Resulta útil elaborar el perfil de los grupos en términos de las variables utilizadas para el conglomerado, como los datos demográficos, los psicográficos, uso del producto, uso de los medios u otras variables. Por ejemplo, los grupos pueden haberse derivado con base en los beneficios que se buscan. Puede realizarse un

perfil más detallado, en términos de las variables demográficas y psicográficas para dirigir los esfuerzos de mercadotecnia hacia cada grupo.

4.6.6. Determinación de la confianza y validez

Dados los criterios generales que comprende el análisis de conglomerados, no debe aceptarse ninguna solución de agrupación sin una evaluación de su confianza y validez. Los siguientes procedimientos ofrecen revisiones adecuadas de la calidad de los resultados de la agrupación.

- Realice el análisis de conglomerados con los mismos datos y utilice distintas medidas de distancia. Compare los resultados con todas las medidas a fin de determinar la estabilidad de las soluciones.
- Utilice diversos métodos de conglomerado y compare los resultados.
- Divida los datos a la mitad en forma aleatoria. Realice el conglomerado por separado en cada mitad. Compare los centroides de grupo en las dos submuestras.
- Elimine las variables en forma aleatoria. Realice la agrupación con base en el conjunto reducido de variables. Compare los resultados basados en el conjunto completo con los que obtuvo al realizar el conglomerado.
- En el conglomerado, no jerárquico, la solución puede depender del orden de los casos en el conjunto de datos. Lleve a cabo corridas múltiples y utilice distintos órdenes de los casos hasta que la solución se estabilice.

4.6.7. Variables conglomeradas

En ocasiones, el análisis de conglomerados se utiliza también para identificar grupos homogéneos. En este caso, las unidades que se utilizan para el análisis son las variables y las medidas de distancia que se calculan para todos los pares de variables. Por ejemplo, el coeficiente de correlación, ya sea el valor absoluto o con el

signo, puede usarse como medida de similitud (la opuesta a la distancia) entre las variables.

El conglomerado jerárquico de las variables puede ayudar en la identificación de variables únicas, o variables que hacen una contribución única a los datos. El conglomerado puede emplearse también para reducir el número de variables. Una combinación de variables en el conglomerado, que se conoce como componentes de conglomerado, se encuentra asociado con cada conglomerado. Frecuentemente, un conjunto grande de variables puede reemplazarse con el conjunto de componentes de conglomerado con poca pérdida de información.

5. SEGMENTACIÓN DE UNA BASE DE TARJETA HABIENTES

5.1. PREPARACIÓN DE DATOS, COMPRENSIÓN DE VARIABLES Y REVISIÓN DE ESCALAS

La segmentación se llevará a cabo con una muestra de una base de datos de clientes de tarjeta de crédito. Esta base de datos, contiene diferentes variables que miden los hábitos de consumo, el comportamiento transaccional, el comportamiento de pagos, y datos demográficos básicos, como la edad, el genero, estado civil, así como el producto que maneja el cliente.

De acuerdo a la metodología de SAS, empezaremos con la revisión de escalas y medidas de los datos, así como identificación de las variables que no aportan información y que podemos excluir, sin perder certeza en el resultado final. Las variables iniciales que componen la base de datos, son las siguientes(ver tabla 5.1):

Tabla 5.1

VARIABLE	DESCRIPCIÓN	ESCALA
CVE_EDO_CIVIL	Estado Civil	Nominal
CVE_SEXO	Género	Nominal
DESCRIPCION_PRODUCTO	Tipo de tarjeta de crédito	Nominal
EDAD_EN_MESES	Edad del cliente en Años	Intervalo
INDICADOR_TARJETAS_ADICIONAL	Indicador de tarjetas adicionales	Binaria
MARG_BAS_NETO_ANUAL	Margen básico neto anual (indica rentabilidad del cliente)	Intervalo
MONTO_ANUAL_INTERESES	Monto anual de intereses	Intervalo
MONTO_COMPRAS_CLINICAS_SALUD	Monto de compras en clínicas de salud en el año	Intervalo
MONTO_COMPRAS_CTAS_DIVERS	Monto de compras en cuentas diversas	Intervalo
MONTO_COMPRAS_ROPA	Monto de compras en ropa	Intervalo
MONTO_COMPRAS_TIENDAS_DEP	Monto de compras en tiendas departamentales	Intervalo
MONTO_CUOTA_ANUAL	Monto cuota anual	Intervalo
MONTO_DISP_CAJEROS	Montos de disposiciones en cajeros	Intervalo
MONTO_DISP_VENTANILLA	Monto disposiciones en ventanilla	Intervalo
MONTO_ULT_MOROSIDAD_30DIAS_ANUAL	El monto de la ultima morosidad de 30 días	Intervalo
NUM_COBROS_ANUAL	Número de cobros anuales	Intervalo
NUM_GIROS_COMPRAS	Número de giros en que compro durante el año	Intervalo
NUM_MESES_ANTIGUEDAD	Antigüedad del cliente en meses	Intervalo
NUM_MESES_CON_INTERESES_ANUAL	Número de meses que se cobró intereses en el año	Intervalo
NUM_MESES_CON_SALDO_ANUAL	Número de meses con saldo en el año.	Intervalo
NUM_MESES_QUE_FACTURO_ANUAL	Número de meses en que facturo en el año	Intervalo
NUM_VECES_MOROSIDAD_30DIAS_ANUAL	Número de veces que tuvo morosidad de 30 dias en el año	Intervalo
NUMERO_CUENTA	Numero de cuenta	Nominal
PROM_ANUAL_INTERESES	Monto promedio de inereses en el año	Intervalo
PROM_ANUAL_LINEA_CREDITO	Monto promedio de linea de crédito anual	Intervalo
PROM_ANUAL_PAGOS	Monto promedio de pagos anuales	Intervalo
PROM_SALDO_CORTE	Promedio del saldo al corte	Intervalo
PUNT_CREDITO	Punto de crédito	Intervalo
SCORE_ACTIVIDAD	Score de actividad del cliente	Intervalo
TXS__DISP_CAJEROS	Transacciones de disposiciones en cajeros	Intervalo
TXS_COMPRAS_AUTOS_CAMIONES	Transacciones de compras en autos y camiones	Intervalo
TXS_COMPRAS_CLUB_MAYORISTAS	Transacciones de compras en club mayoristas	Intervalo
TXS_COMPRAS_MARITIMOS	Transacciones de compras en marítimos	Intervalo
TXS_COMPRAS_NEGOCIOS	Transacciones de compras en negocios	Intervalo
TXS_COMPRAS_ROPA	Transacciones de compra en ropa	Intervalo
TXS_COMPRAS_TIENDAS_DEP	Transacciones de compras en tiendas departamentales	Intervalo
TXS_COMPRAS_VTAS_DIVERS	Transacciones de compras en diversos	Intervalo

5.2. LA BASE DE DATOS

El modelo siguiente (ver figura 5.1) nos sirvió para llevar a cabo el análisis estadístico inicial, para determinar que variables están correlacionadas, y la identificación de que variables nos aportan el 70% de la varianza de la población:

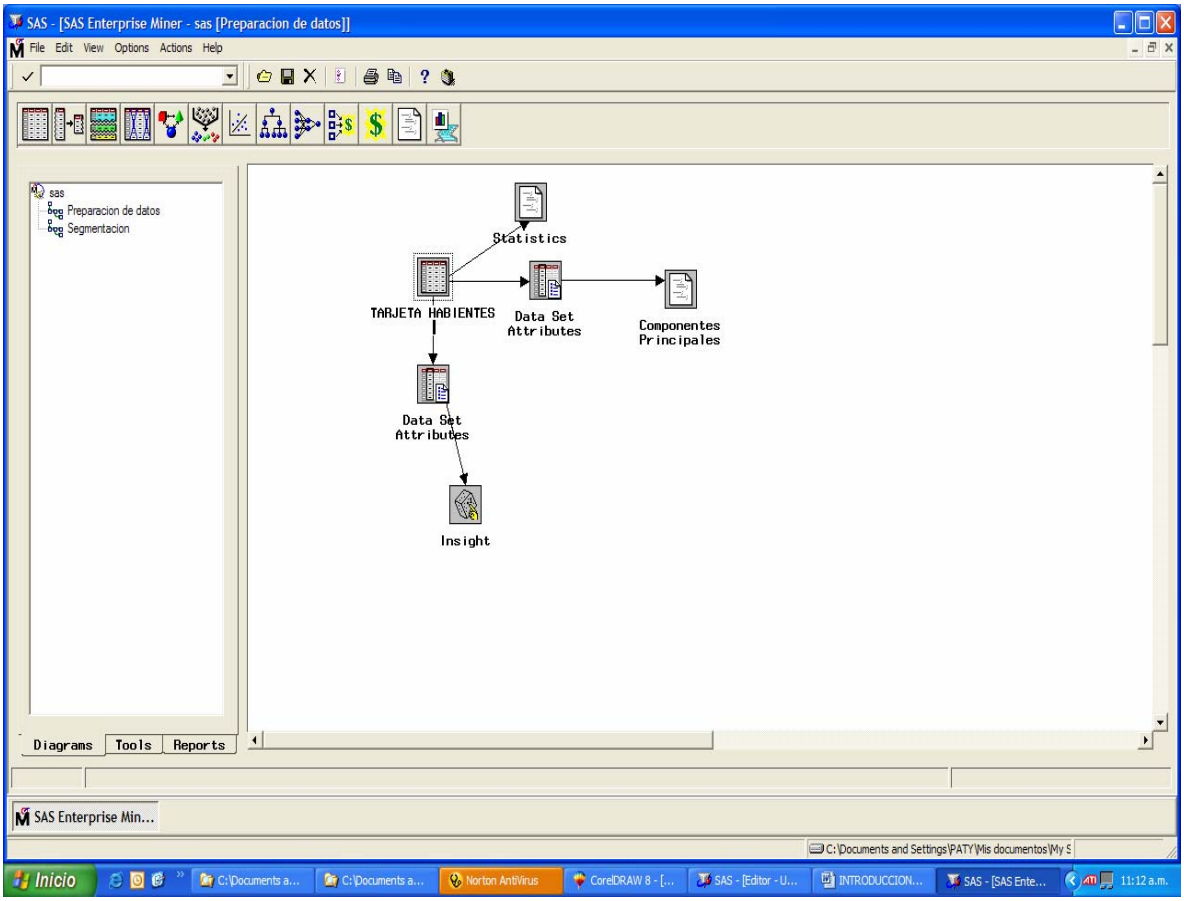


Figura 5.1 Modelo para análisis estadistico.

El diagrama tiene los siguientes componentes:

- 1 Nodo de Input Data Source: Que contiene la base de datos que vamos a analizar.(TARJETA HABIENTES)

- 2 Nodos de SAS Code: que contiene el Análisis Estadístico Univariado(Summary Statistics), y Análisis de Componentes Principales. (Componentes Principales)
- 2 Data Set Attributes: Que nos sirven para decirle al modelo que variables no deben de ser incluidas en el análisis.
- 1 Nodo Insight que nos permine dejar la nueva base en otro archivo.

El nodo de SAS Code: que contiene el Análisis Estadístico nos da los siguientes resultados (figura 5.2):

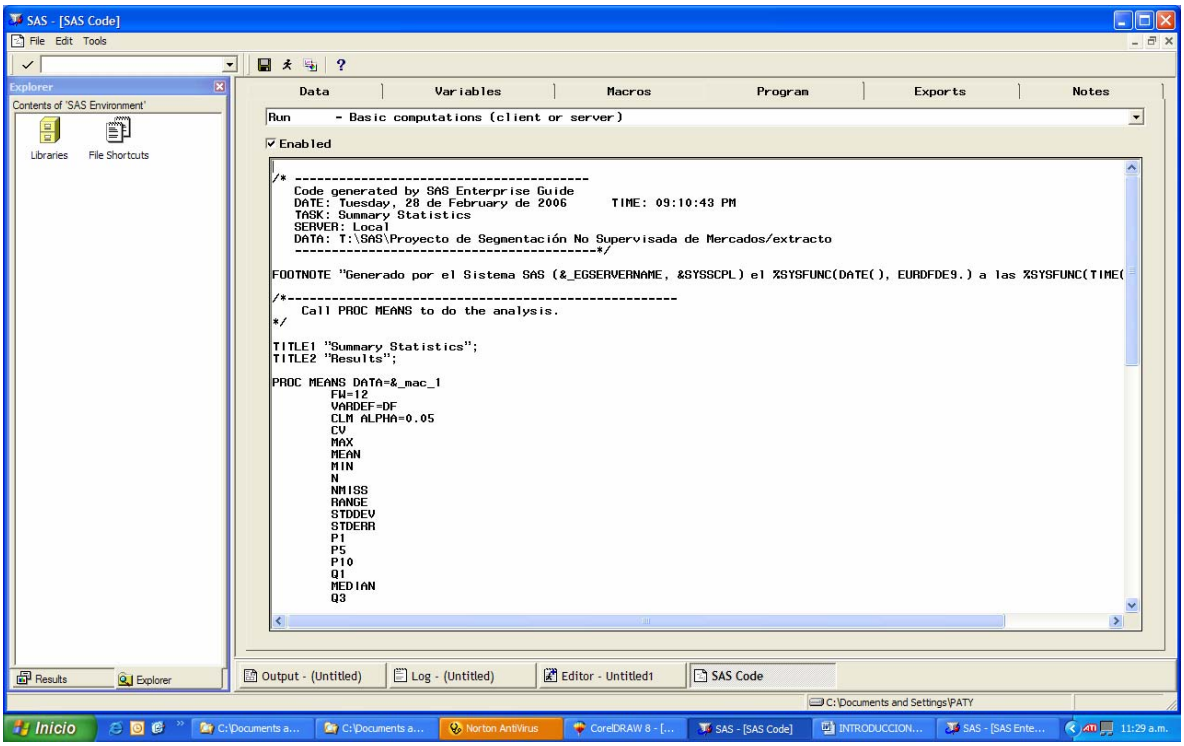


Figura 5.2 Código para obtener análisis estadístico

A continuación podemos ver los resultados estadisticos (tabla5.2).

Tabla 5.2 Resultados estadísticos.

Variable	N	Minimum	Maximum	Mean	Range	Median	Std Dev	Std Error	Sum	Variance
EDAD_EN_MESES	20000	4	99	43.7455	95	43	11.0238109	0.0779501	874910	121.524406
INDICADOR_TARJETAS_ADICIONAL	20000	0	1	0.1588	1	0	0.3654986	0.0025845	3176	0.1335892
MARG_BAS_NETO_ANUAL	20000	0	53581.17	1923	53581.17	1430.78	1896.29	13.4088151	38459943.33	3595926.47
MONTO_ANUAL_INTERESES	20000	0	49791.3	1570.66	49791.3	929.68	2307.76	16.3183486	31413256.09	5325770.04
MONTO_COMPRAS_CLINICAS_SALUD	20000	0	25630.22	37.3351715	25630.22	0	431.2163082	3.0491598	746703.43	185947.5
MONTO_COMPRAS_CTAS_DIVERS	20000	0	423572.2	1272.48	423572.2	0	6410.59	45.3297461	25449687.92	41095717.55
MONTO_COMPRAS_ROPA	20000	0	845279.28	610.312087	845279.28	0	7201.59	50.9228979	12206241.74	51862830.52
MONTO_COMPRAS_TIENDAS_DEP	20000	0	276305.8	2562.05	276305.8	0	6975.29	49.3227535	51240983.95	48654680.25
MONTO_CUOTA_ANUAL	20000	0	3360	311.143515	3360	270	168.2221012	1.1895099	6222870.3	28298.68
MONTO_DISP_CAJEROS	20000	0	475901	1168.9	475901	0	7236.26	51.1680634	23378083.74	52363414.31
MONTO_DISP_VENTANILLA	20000	0	1051200	1033	1051200	0	11823.11	83.6020257	20660079.76	139785974
MONTO_ULT_MOROSIDAD_30DIAS_ANUAL	20000	0	8991.36	132.049513	8991.36	0	296.7547386	2.0983729	2640990.26	88063.37
NUM_COBROS_ANUAL	20000	0	188	3.2106	188	0	9.0421477	0.0639376	64212	81.7604357
NUM_GIROS_COMPRAS	20000	0	23	2.7789	23	0	3.6744782	0.0259825	55578	13.5017899
NUM_MESES_ANTIGUEDAD	20000	11	425	53.5471	414	31	59.6577307	0.4218439	1070942	3559.04
NUM_MESES_CON_INTERESES_ANUAL	20000	0	12	7.6121	12	10	4.1409771	0.0292811	152242	17.147691
NUM_MESES_CON_SALDO_ANUAL	20000	0	12	10.49545	12	12	2.9873442	0.0211237	209909	8.9242255
NUM_MESES_QUE_FACTURO_ANUAL	20000	0	12	7.25475	12	8	4.0949836	0.0289559	145095	16.7688909
NUM_VECES_MOROSIDAD_30DIAS_ANUAL	20000	0	89	2.45205	89	0	5.9732725	0.0422374	49041	35.6799848
PROM_ANUAL_INTERESES	20000	0	4149.28	130.888931	4149.28	77.47	192.3136169	1.3598626	2617778.62	36984.53
PROM_ANUAL_LINEA_CREDITO	20000	1	812500	17908.18	812499	11000	20322.87	143.7043574	358163654	413018847
PROM_ANUAL_PAGOS	20000	0	249096.15	1440.58	249096.15	760.46	3260.06	23.0520991	28811559.68	10627985.42
PROM_SALDO_CORTE	20000	0	385648.16	6735.28	385648.16	4481.08	8905.08	62.9684223	134705588	79300444.08
PUNT_CREDITO	20000	0	742	639.2908	742	677	158.345346	1.1196707	12785816	25073.25
SCORE_ACTIVIDAD	20000	0.1430556	1	0.6717514	0.8569444	0.7569444	0.3274779	0.0023156	13435.03	0.1072418
TXS__DISP_CAJEROS	20000	0	120	1.01615	120	0	4.5731974	0.0323374	20323	20.9141349
TXS_COMPRAS_AUTOS_CAMIONES	20000	0	483904.61	346.268443	483904.61	0	4192.49	29.6453478	6925368.86	17576932.93
TXS_COMPRAS_CLUB_MAYORISTAS	20000	0	69	0.468	69	0	2.0616228	0.0145779	9360	4.2502885
TXS_COMPRAS_MARITIMOS	20000	0	224	1.4762	224	0	5.3866673	0.0380895	29524	29.0161844
TXS_COMPRAS_NEGOCIOS	20000	0	95	1.6651	95	0	4.0546719	0.0286709	33302	16.440364
TXS_COMPRAS_ROPA	20000	0	51	0.62945	51	0	2.0730793	0.0146589	12589	4.2976576
TXS_COMPRAS_TIENDAS_DEP	20000	0	343	5.49335	343	0	14.0812631	0.0995696	109867	198.2819699
TXS_COMPRAS_VTAS_DIVERS	20000	0	211	2.61265	211	0	6.9155478	0.0489003	52253	47.8248012

De estos estadísticos descriptivos, podemos observar que ya tenemos variables que podríamos descartar, debido a que no proporcionan información suficiente que nos pueda ayudar a segmentar la información, como puede ser:

El campo de MONTO_COMPRAS_TIENDAS_DEP, esto es debido a que mas del 75% de los valores son 0 [cero], para visualizarlo mejor, a continuación tenemos su histograma (figura5.3):

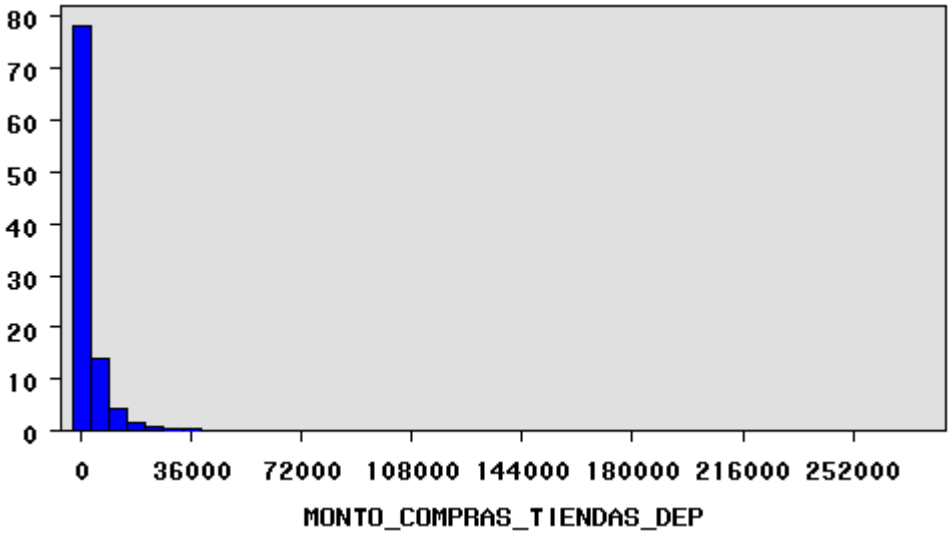


Figura 5.3 Histograma monto compras

Otra variable que podriamos eliminar, seria TXS_COMPRAS_CLUB_MAYORISTAS, esto debido a que mas del 80% de sus valores son 0 [cero], a continuación su histograma (figura 5.4):

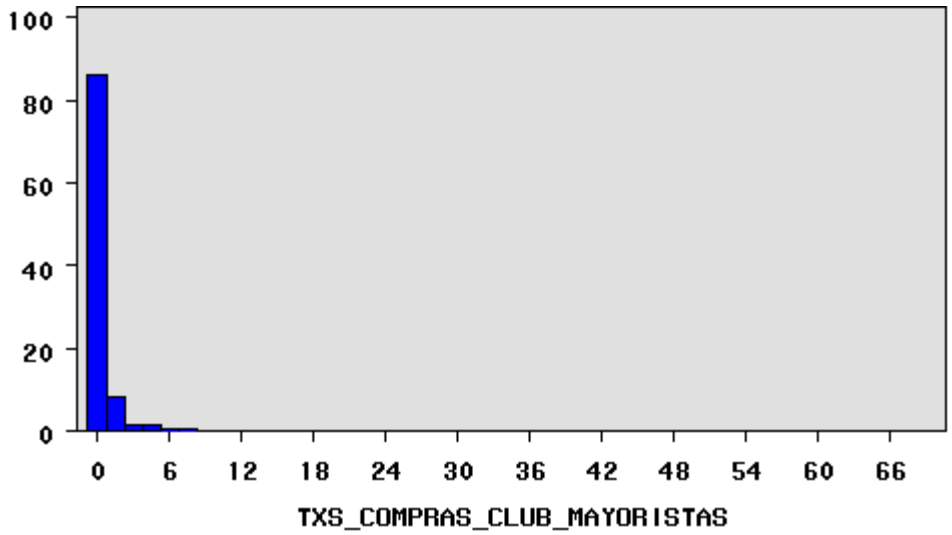


Figura 5.4. Histograma de compras club mayoristas

Estos supuestos, debemos revisarlos con el análisis multivariado de componentes principales.

Interesados en estas variables, procedemos a incluir un DataSet Attributes, para excluir las variables redundantes(ver figura 5.5) [PROM_SALDO_CORTE, MARG_BAS_NETO_ANUAL,PROM_ANUAL_INTERESES,NUM_MESES_QUE_FACTURO_ANUAL],

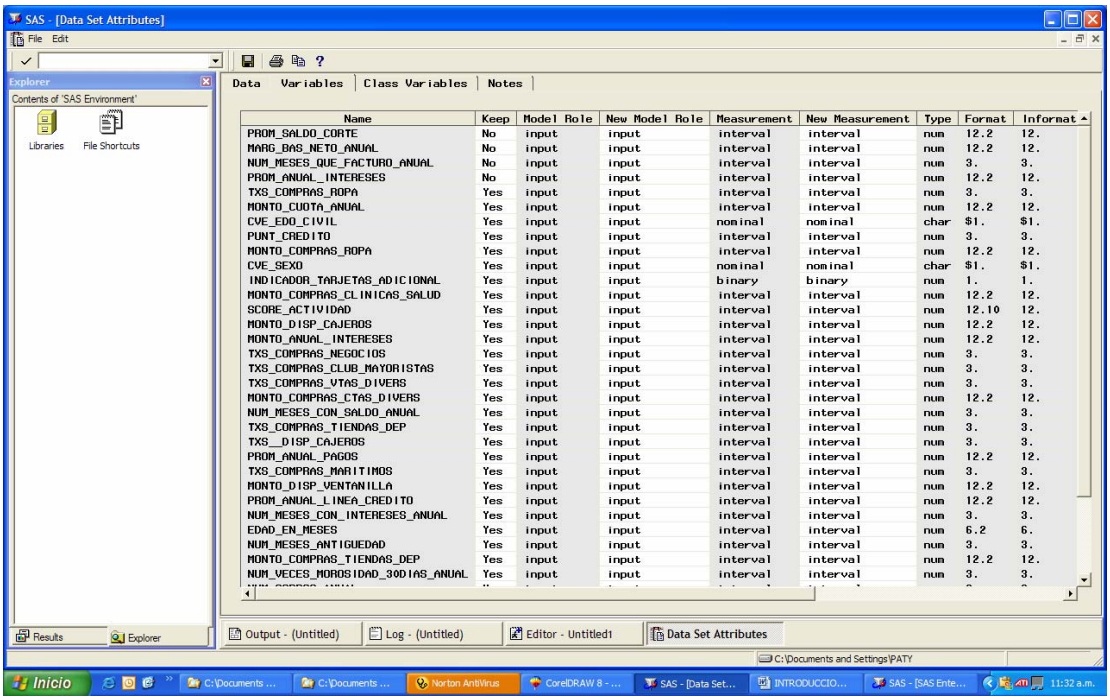


Figura 5.5 Variables redundantes

.

Esto con el fin de correr los componentes principales (ver figura 5.6) para poder observar que componentes nos proporcionan el 70% de la varianza.

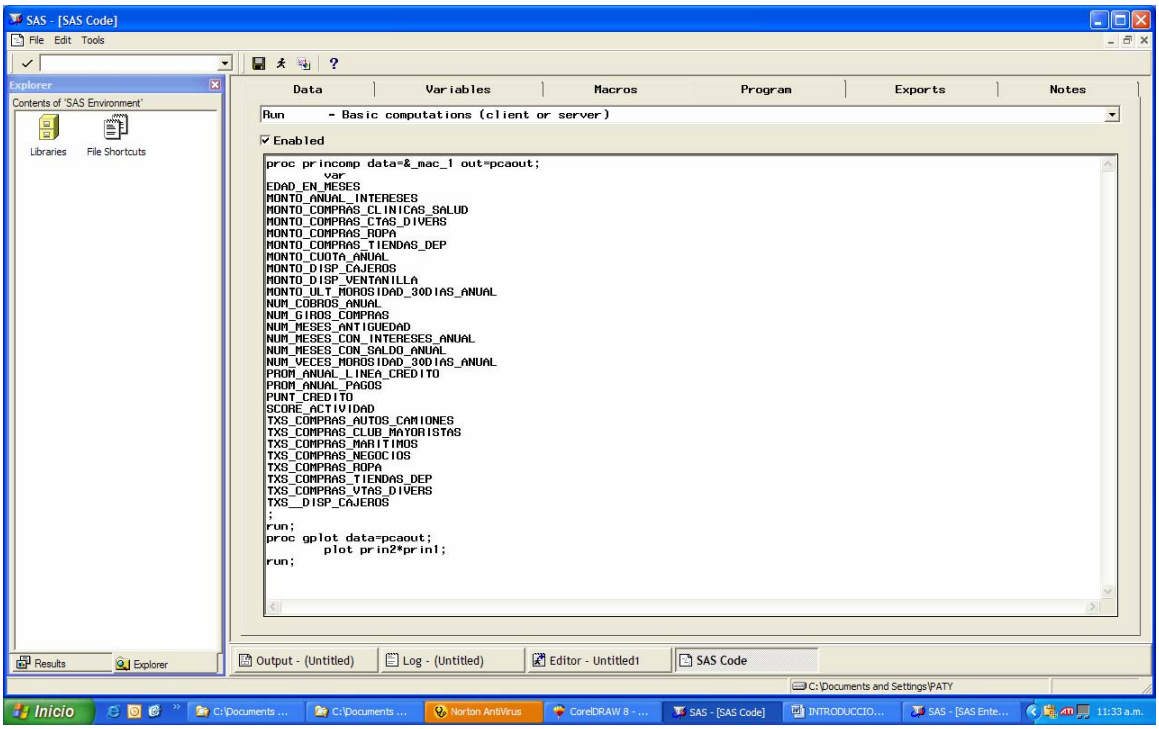


Figura 5.6 Código para correr componentes principales.

De acuerdo a la salida siguiente:

The PRINCOMP Procedure

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	5.41121818	2.84251983	0.1933	0.1933
2	2.56869835	0.52215217	0.0917	0.2850
3	2.04654618	0.17407786	0.0731	0.3581
4	1.87246832	0.33221003	0.0669	0.4250
5	1.54025829	0.13463294	0.0550	0.4800
6	1.40562535	0.19586731	0.0502	0.5302
7	1.20975804	0.09467391	0.0432	0.5734
8	1.11508413	0.11273231	0.0398	0.6132
9	1.00235182	0.05086026	0.0358	0.6490
10	0.95149156	0.03136802	0.0340	0.6830
11	0.92012353	0.12271058	0.0329	0.7158
12	0.79741296	0.04699131	0.0285	0.7443
13	0.75042164	0.01533865	0.0268	0.7711
14	0.73508299	0.03664636	0.0263	0.7974
15	0.69843663	0.01803681	0.0249	0.8223
16	0.68039982	0.05295126	0.0243	0.8466
17	0.62744855	0.05212134	0.0224	0.8690
18	0.57532721	0.08230959	0.0205	0.8896
19	0.49301762	0.01028869	0.0176	0.9072
20	0.48272893	0.03938355	0.0172	0.9244
21	0.44334538	0.04265632	0.0158	0.9403
22	0.40068906	0.09090628	0.0143	0.9546
23	0.30978278	0.03499944	0.0111	0.9656
24	0.27478334	0.06164782	0.0098	0.9754
25	0.21313551	0.03573224	0.0076	0.9831
26	0.17740328	0.01000267	0.0063	0.9894
27	0.16740061	0.03784065	0.0060	0.9954
28	0.12955996		0.0046	1.0000

Podemos observar que en el componente 11, tenemos explicada el 70% de la varianza, por lo que ahora procedemos a revisar que variables son las que forman este 70% y aquellas que no aporten nada, serán excluidas del modelo.

A continuación se hace el analisis de la tabla de componentes principales (tabla 5.3). Podemos observar que las varibles:

- SCORE_ACTIVIDAD
- TXT_COMPRAS_CLUB_MAYORISTA
- TXT_COMPRAS_ROPA
- TXS_COMPRAS:TIENDAS _DEP

no aportan información en alguno de los 11 componente.

VARIABLE	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11
EDAD_EN_MESES	0.03564	-0.04058	0.11676	0.16042	0.39135	-0.00513	-0.47102	-0.13119	-0.08556	0.01875	0.03803
MONTO_ANUAL_INTERESES	0.09960	0.33261	0.08938	0.34295	-0.12822	0.11890	-0.24308	-0.22096	-0.14633	0.00425	0.03662
MONTO_COMPRAS_CLINICAS_SALUD	0.11526	-0.04592	0.06251	-0.01267	-0.06970	0.11412	0.06857	0.18783	-0.22021	0.45470	0.78565
MONTO_COMPRAS_CTAS_DIVERS	0.21007	-0.02828	0.05715	0.13646	0.03077	-0.26753	0.42190	-0.28671	-0.20206	-0.04780	0.07299
MONTO_COMPRAS_ROPA	0.19387	-0.05049	0.07004	0.30248	-0.18385	-0.14710	0.03291	0.49199	-0.14232	0.05708	-0.30257
MONTO_COMPRAS_TIENDAS_DEP	0.33149	-0.09132	0.04020	-0.11302	-0.05683	0.19488	-0.08282	0.02961	0.01561	0.05361	-0.02466
MONTO_CUOTA_ANUAL	0.07406	0.03707	0.24098	0.07590	0.43988	0.18132	0.16825	0.15570	0.06110	0.05167	-0.00168
MONTO_DISP_CAJEROS	0.09945	0.17762	-0.50778	0.09822	0.16554	0.09229	0.03535	0.01000	0.06065	0.05100	0.00342
MONTO_DISP_VENTANILLA	0.08017	0.03665	-0.14699	0.20213	-0.01593	-0.15994	0.08608	-0.22780	0.65262	0.47448	-0.00231
MONTO_ULT_MOROSIDAD_30DIAS_ANUAL	0.00210	0.26929	0.08395	0.27785	-0.19502	0.34876	0.04122	-0.25381	-0.06117	-0.13749	0.06554
NUM_COBROS_ANUAL	0.01080	0.22037	-0.34887	-0.02506	0.14211	0.05454	0.01660	0.24985	-0.22669	-0.07669	0.00978
NUM_GIROS_COMPRAS	0.33616	0.00140	-0.05980	-0.16287	-0.02063	0.04025	0.06156	-0.15827	0.01262	-0.10037	0.04335
NUM_MESES_ANTIGUEDAD	0.04350	-0.01354	0.24439	0.10324	0.56567	0.15624	0.06355	0.09775	0.07674	0.02159	-0.03526
NUM_MESES_CON_INTERESES_ANUAL	-0.06795	0.48949	0.14446	-0.03270	-0.18035	0.01257	-0.00610	0.03149	-0.02374	0.04898	-0.01209
NUM_MESES_CON_SALDO_ANUAL	0.08986	0.48616	0.19802	-0.20595	-0.05207	-0.12891	-0.02752	0.08640	0.05258	0.02965	-0.01412
NUM_VECES_MOROSIDAD_30DIAS_ANUAL	-0.06162	0.16726	0.19184	0.13601	0.04431	0.39592	0.37799	0.10108	0.19205	-0.07351	-0.07824
PROM_ANUAL_LINEA_CREDITO	0.20700	-0.05937	0.03616	0.38463	0.02957	-0.09695	-0.34693	-0.07690	-0.12753	-0.00416	0.01649
PROM_ANUAL_PAGOS	0.30924	0.00118	-0.04506	0.33304	-0.07768	-0.16993	0.03583	0.19247	0.16291	0.07419	-0.08077
PUNT_CREDITO	0.11612	0.33790	0.15835	-0.24889	0.08262	-0.26621	-0.10676	0.11863	0.12708	0.07384	-0.00730
SCORE_ACTIVIDAD	0.19700	0.17061	0.01809	-0.22871	0.22179	-0.29607	-0.11677	-0.04128	-0.02768	-0.02824	0.02718
TXS_COMPRAS_AUTOS_CAMIONES	0.09381	-0.02825	0.00984	0.12724	-0.06550	-0.12435	-0.05397	0.27255	0.40517	-0.64251	0.47178
TXS_COMPRAS_CLUB_MAYORISTAS	0.20086	-0.05141	0.00385	-0.16130	-0.05131	0.19964	-0.07259	-0.21402	0.14321	-0.04276	0.02308
TXS_COMPRAS_MARITIMOS	0.25031	-0.06066	0.01792	-0.05558	-0.07067	0.11809	-0.05725	-0.03713	0.04353	-0.16407	-0.07831
TXS_COMPRAS_NEGOCIOS	0.31720	-0.06622	0.03649	-0.10933	-0.07425	0.14370	0.00246	0.02539	-0.02619	-0.02152	-0.09723
TXS_COMPRAS_ROPA	0.26629	-0.08820	0.06629	-0.08912	-0.14177	0.12297	0.04060	0.23385	-0.09433	0.17429	-0.14036
TXS_COMPRAS_TIENDAS_DEP	0.29597	-0.06654	-0.01883	-0.24520	-0.01936	0.25492	-0.13059	-0.07692	0.10721	-0.00728	-0.04255
TXS_COMPRAS_VTAS_DIVERS	0.24239	0.01507	0.04994	0.03895	0.12688	-0.25768	0.39851	-0.26527	-0.23145	-0.15560	0.02483
TXS__DISP_CAJEROS	0.10084	0.20911	-0.54815	0.02009	0.18346	0.13143	0.04820	0.07334	-0.07615	-0.04021	0.00809

Con el nodo data set attributes excluimos estas variables que no aportan información (figura 5.7):

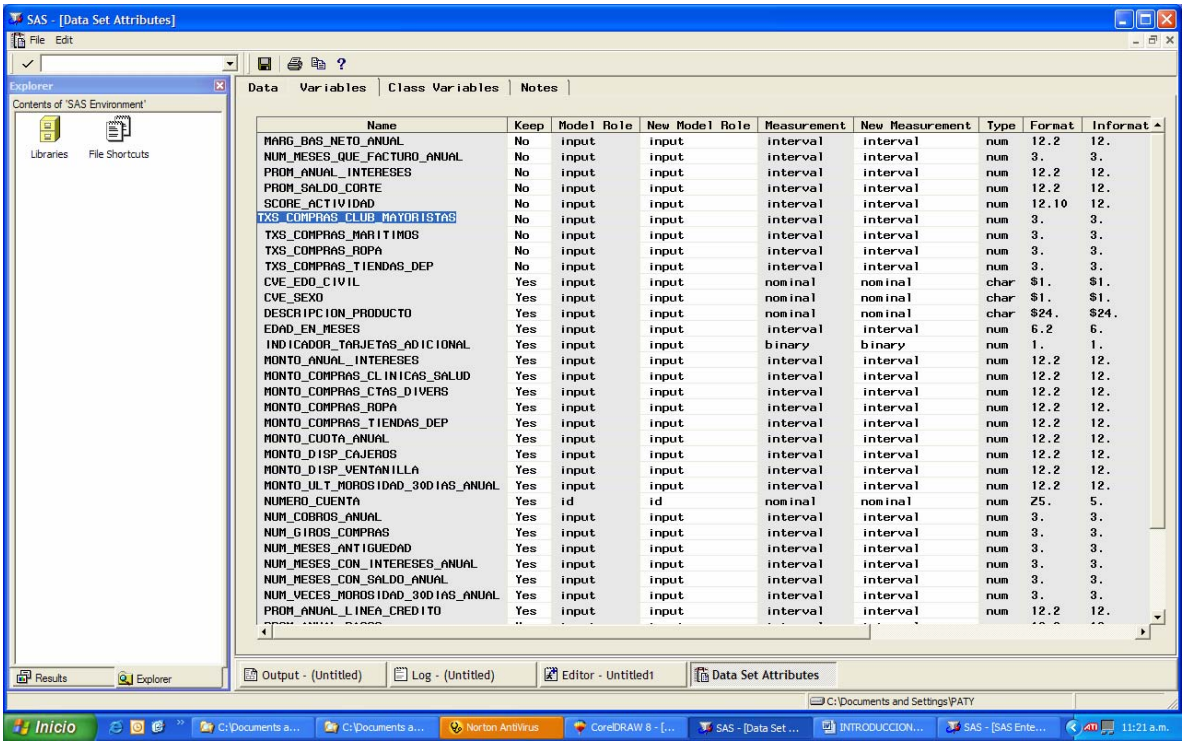


Figura 5.7 variables sin información

Ya que tenemos las variables con las que nos vamos a quedar, empezaremos el estudio inicial de Cluster que nos ayude a resolver las dudas comerciales y de negocio para así poder determinar las acciones necesarias para cada segmento de los clientes.

Con esta información, procedemos a extraer la información por medio del nodo Insight (figura 5.8):

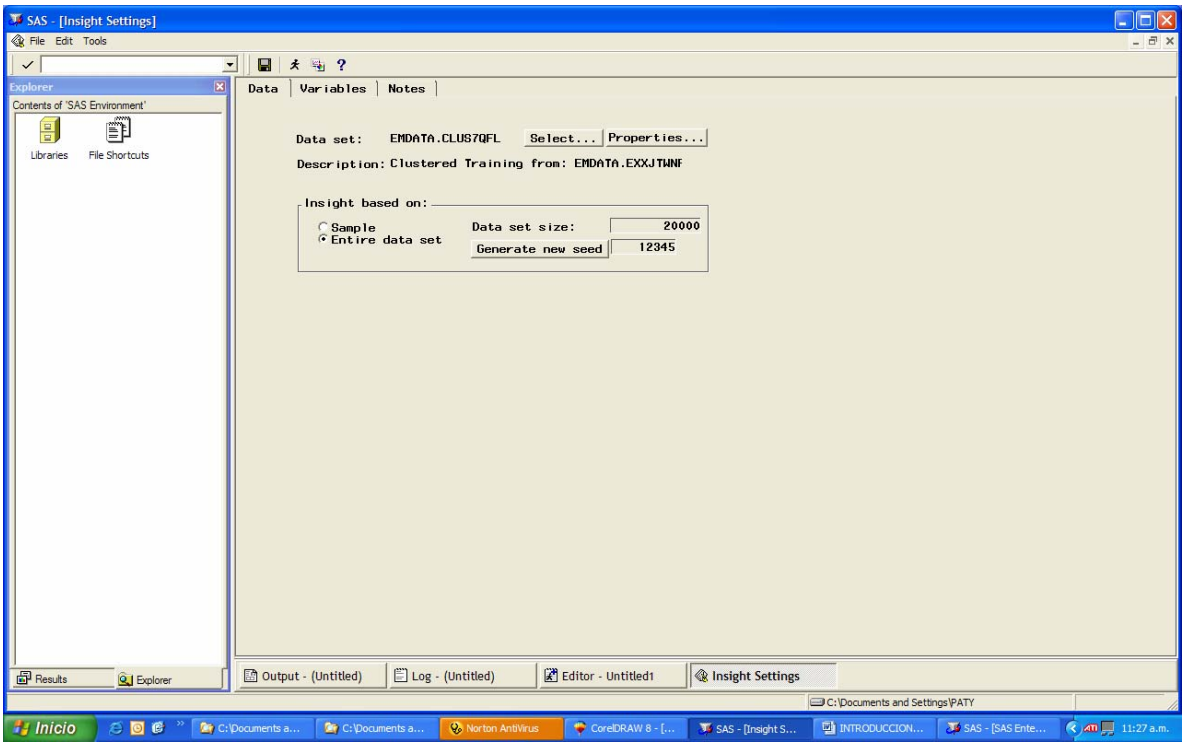


Figura 5.8. Obtener información

5.3. MODELOS ANALIZADOS

Se genera un nuevo diagrama para empezar la segmentación no supervisada, el siguiente diagrama nos muestra los métodos que se ocuparon (figura 5.9):

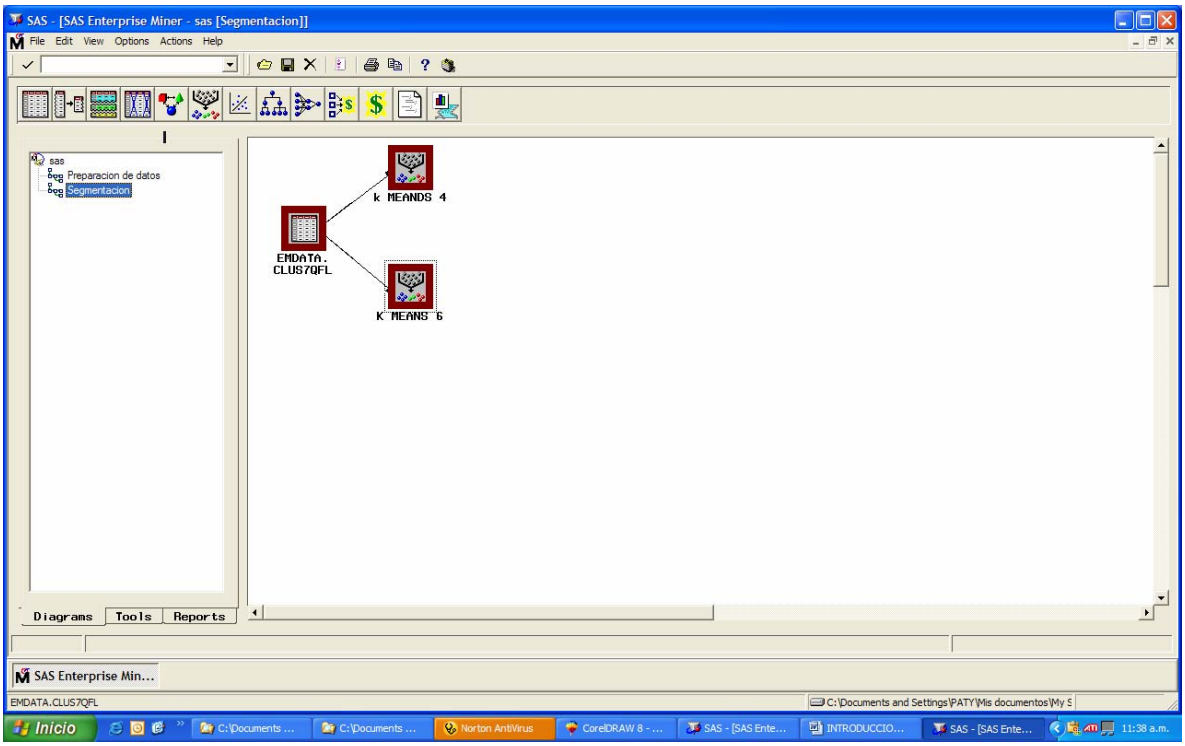


Figura 5.9 Modelos utilizados

El diagrama consta de los siguientes elementos:

Un Nodo de Input Data Source para incluir el archivo que nos quedo del análisis inicial [EMDATA.CLUS7QFL].

Dos nodos de clustering.de 6 y 4 clusters

Se probaron dos métodos de K-Means para determinar cual de ellos nos daba resultados significativos, el de 6 clusters no dio buenos resultados y esto lo podemos corroborar con la explicación de la siguiente tabla que muestra los resultados (tabla 5.4):

Tabla 5.4. resultado de los clusters

CLUSTER	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster
1	1,549	1.254766013	22.38798091	5	3.527416226
2	2	3.190176865	13.3454673	6	54.65844984
3	13	2.006253902	19.42710394	5	21.32417798
4	1	.	0	1	133.1665461
5	15,946	0.901508293	23.48772738	1	3.527416226
6	188	1.573065454	24.09228133	5	8.909440799

La frecuencia por cluster, nos muestra que en uno de ellos se encuentra más del 50% de los casos (Cluster 5), y otro sólo tiene 1 caso (Cluster 4), esto ya es un indicio de una mala segmentación.

Adicional a lo anterior, podemos observar que los clusters 1, 5 y 6 están muy cercanos mientras que el 2, 3 y 4 (con menor número de casos) están mas alejados.

El diagrama de proximidades nos muestra como están distribuidos (figura 5.10):

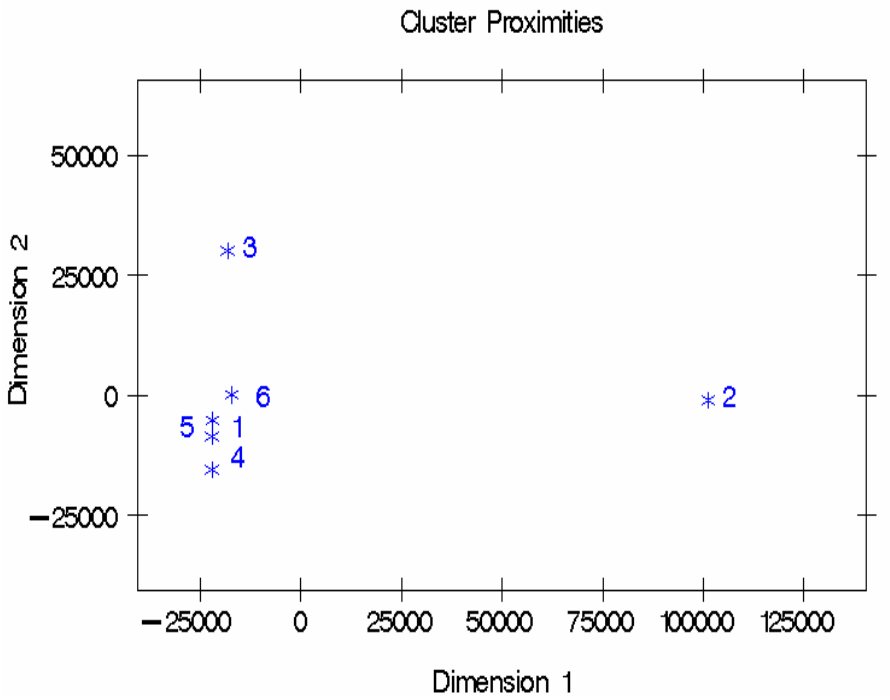


Figura 5.10 diagrama de proximidades

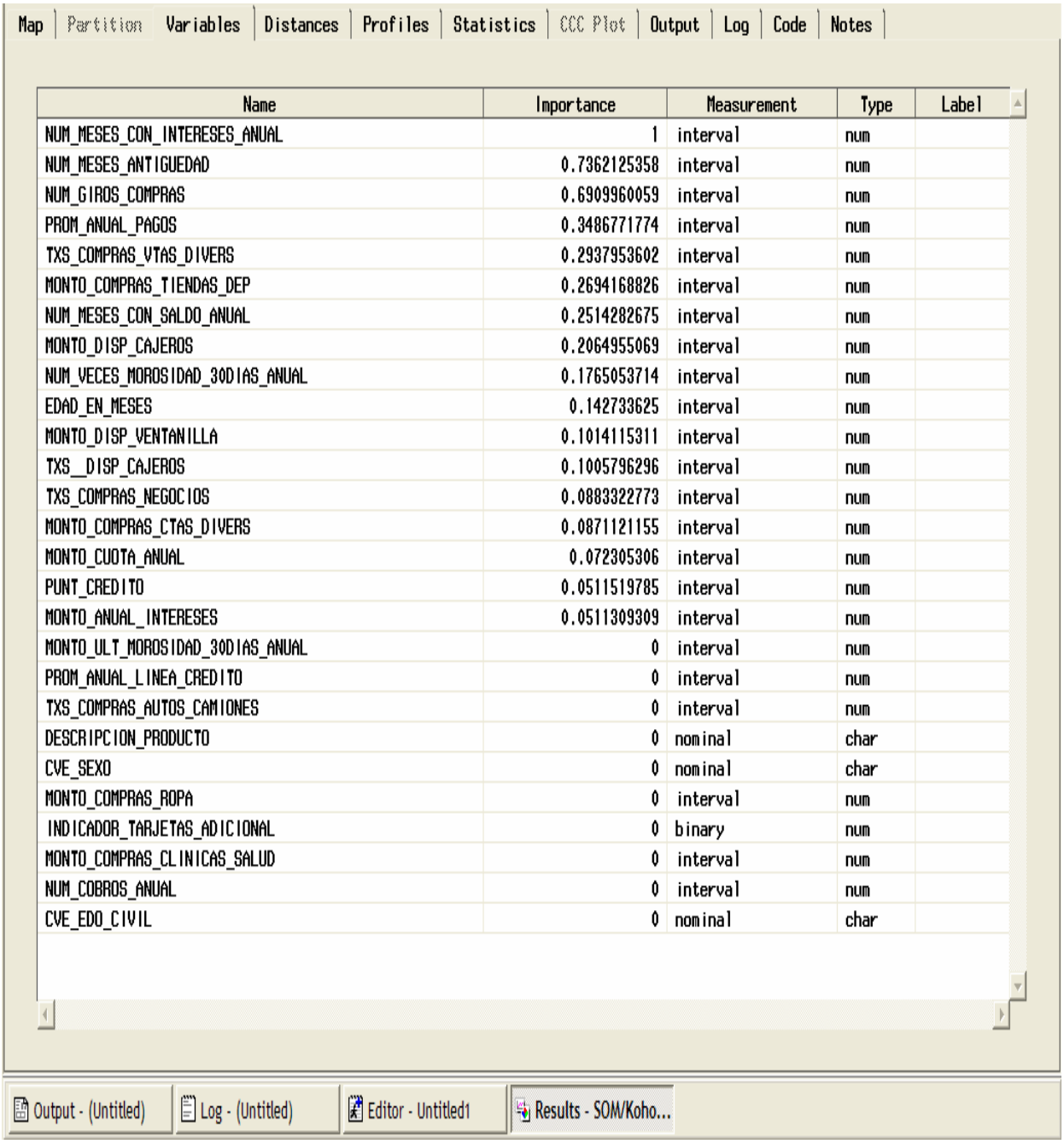
Como no nos proporciona una buena segmentación, procedemos a utilizar el método de k medias con 4 clusters.

Se obtuvieron los siguientes resultados (tabla 5.5): la distribución de los casos es más uniforme:

Tabla 5.5 resultado de los clusters

CLUSTER	1	2	3	4
Frequency of Cluster	2,147	3,273	9,199	3,080
Root-Mean-Square Standard Deviation	0.7703	0.6859	0.5887	1.2515
Maximum Distance from Cluster Seed	59.6737	18.8186	21.0019	68.1392
Distance to Nearest Cluster	3.1819	3.3487	3.1819	4.3178
Nearest Cluster	3	3	1	2

Tmbién las variables que modelan estos clusters, están relacionadas con la preguntas de negocio que queremos contestar (figura 5.11).



Name	Importance	Measurement	Type	Label
NUM_MESES_CON_INTERESES_ANUAL	1	interval	num	
NUM_MESES_ANTIGUEDAD	0.7362125358	interval	num	
NUM_GIROS_COMPRAS	0.6909960059	interval	num	
PROM_ANUAL_PAGOS	0.3486771774	interval	num	
TXS_COMPRAS_VTAS_DIVERS	0.2937953602	interval	num	
MONTO_COMPRAS_TIENDAS_DEP	0.2694168826	interval	num	
NUM_MESES_CON_SALDO_ANUAL	0.2514282675	interval	num	
MONTO_DISP_CAJEROS	0.2064955069	interval	num	
NUM_VECES_MOROSIDAD_30DIAS_ANUAL	0.1765053714	interval	num	
EDAD_EN_MESES	0.142733625	interval	num	
MONTO_DISP_VENTANILLA	0.1014115311	interval	num	
TXS__DISP_CAJEROS	0.1005796296	interval	num	
TXS_COMPRAS_NEGOCIOS	0.0883322773	interval	num	
MONTO_COMPRAS_CTAS_DIVERS	0.0871121155	interval	num	
MONTO_CUOTA_ANUAL	0.072305306	interval	num	
PUNT_CREDITO	0.0511519785	interval	num	
MONTO_ANUAL_INTERESES	0.0511309309	interval	num	
MONTO_ULT_MOROSIDAD_30DIAS_ANUAL	0	interval	num	
PROM_ANUAL_LINEA_CREDITO	0	interval	num	
TXS_COMPRAS_AUTOS_CAMIONES	0	interval	num	
DESCRIPCION_PRODUCTO	0	nominal	char	
CVE_SEXO	0	nominal	char	
MONTO_COMPRAS_ROPA	0	interval	num	
INDICADOR_TARJETAS_ADICIONAL	0	binary	num	
MONTO_COMPRAS_CLINICAS_SALUD	0	interval	num	
NUM_COBROS_ANUAL	0	interval	num	
CVE_EDO_CIVIL	0	nominal	char	

Figura 5.11 variables que modelan el cluster

Por ultimo se puede observar que los clusters están lo suficientemente separados como para poder determinar que están bien clasificados (figura 5.12).

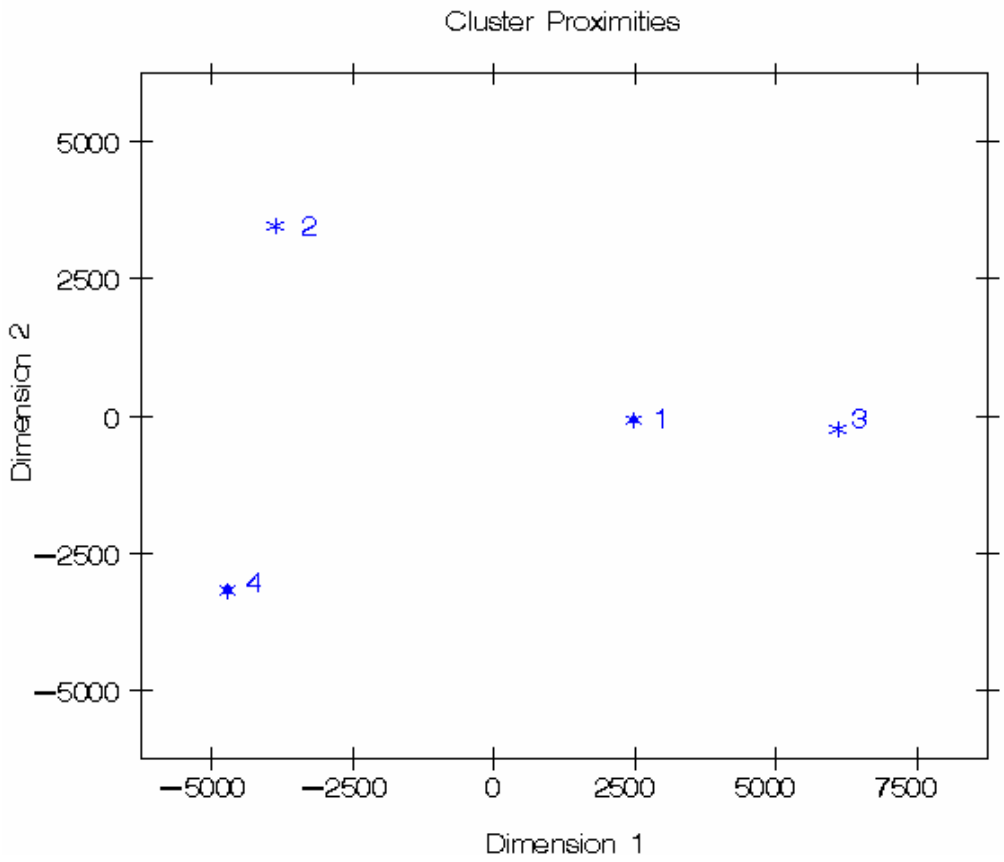


Figura 5.12 distribución de los clusters

A continuación veremos la distribución de las variables de cada cluster para ir observando la composición de cada uno.

GENERO vs. Tarjetas adicionales (figura 5.13)

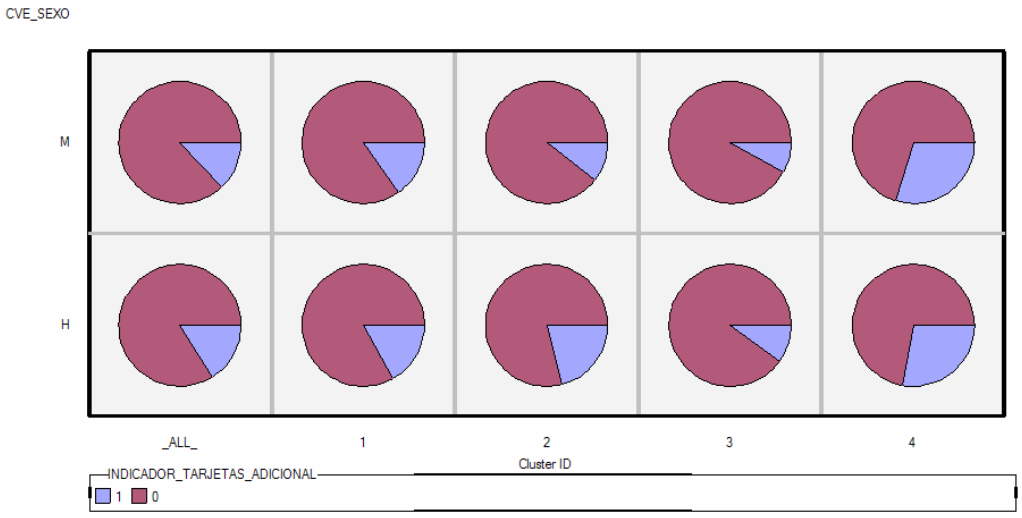


Figura 5.13Tarjetas adicionales

Se puede observar que el GENERO no es un factor decisivo para utilizar una tarjeta adicional, ya que el promedio de hombres vs. el promedio de mujeres es casi el mismo en todos los clusters.

GENERO vs. Producto

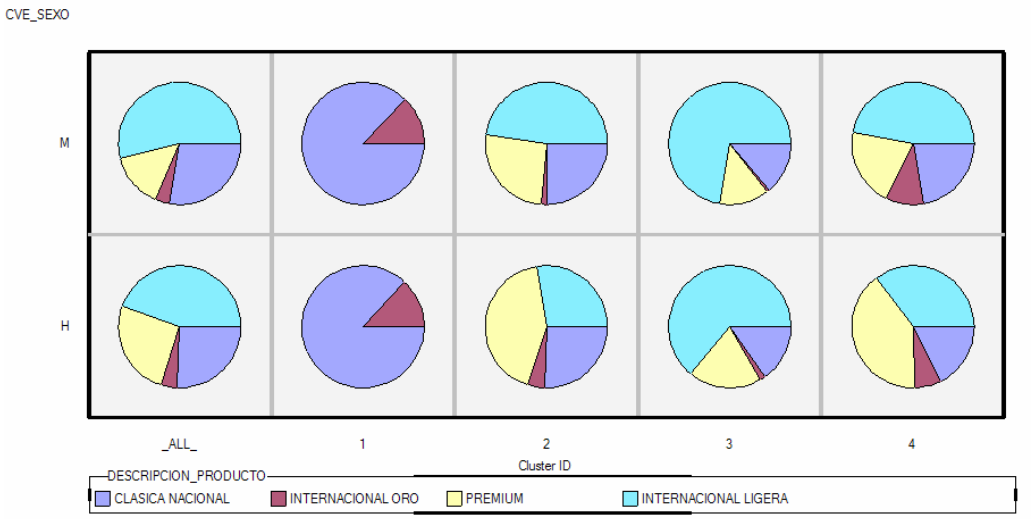


Figura 5.14 genero vs producto

La tarjeta internacional Oro es la que menor número de clientes tiene(figura 5.14). Y prácticamente en el cluster 3 no existen clientes de este tipo de producto.

En el cluster 2, la tarjeta Premium la prefieren más hombres que mujeres, mientras que la tarjeta Internacional Ligera es preferida entre las mujeres.

En el cluster 3, la preferencia de productos es mas uniforme y prácticamente existe el mismo numero de clientes hombres y mujeres.

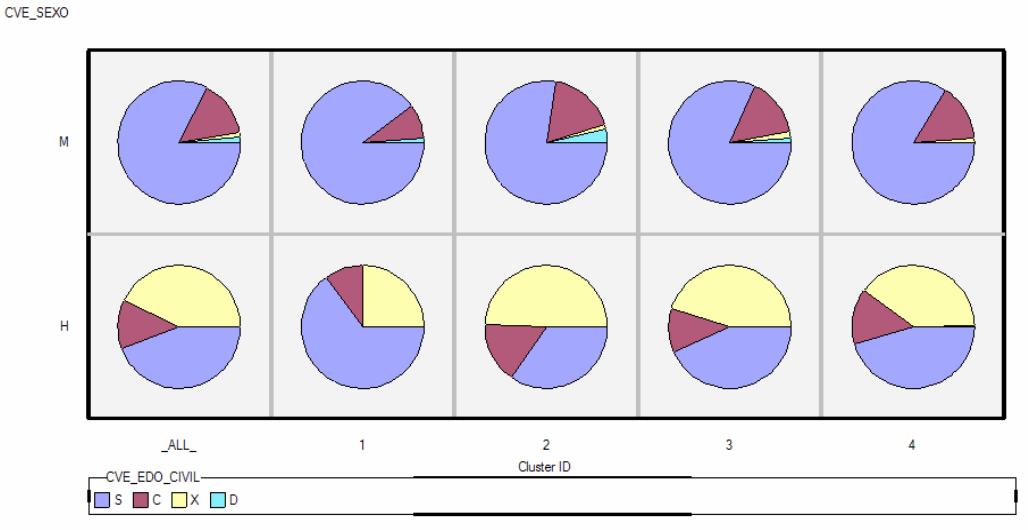


Figura 5.15 GENERO vs. Estado Civil

Se observa que la población de mujeres en su gran mayoría son solteras (figura 5.15).

Y en el caso de hombres más del 40% no informa su estado civil.

Tarjetas Adicionales vs. Producto

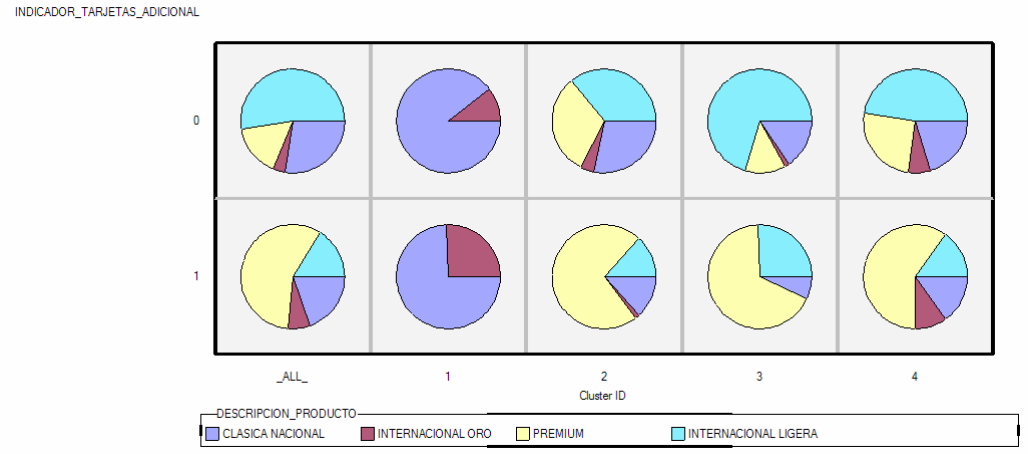


Figura 5.16 tarjetas adicionales vs producto

La tarjeta Premium es la que más clientes tiene con tarjetas adicionales. Mientras que la tarjeta Internacional Ligera es la que menos tiene tarjetas adicionales (figura 5.16).

Observemos ahora la composición de las variables de clase y categóricas que componen cada cluster.

CLUSTER 1

Figura 5.17 cluster 1

Los clientes de este segmento usan poco su tarjeta, son los que mas antigüedad tienen con la tarjeta, y también son los mas grandes en edad, se retrasan demasiado en sus pagos, en su gran mayoría son tarjetas Clásicas Nacional y su línea de

crédito está por debajo del promedio de la población y en su gran mayoría son solteros(figura 5.17)..

CLUSTER 2

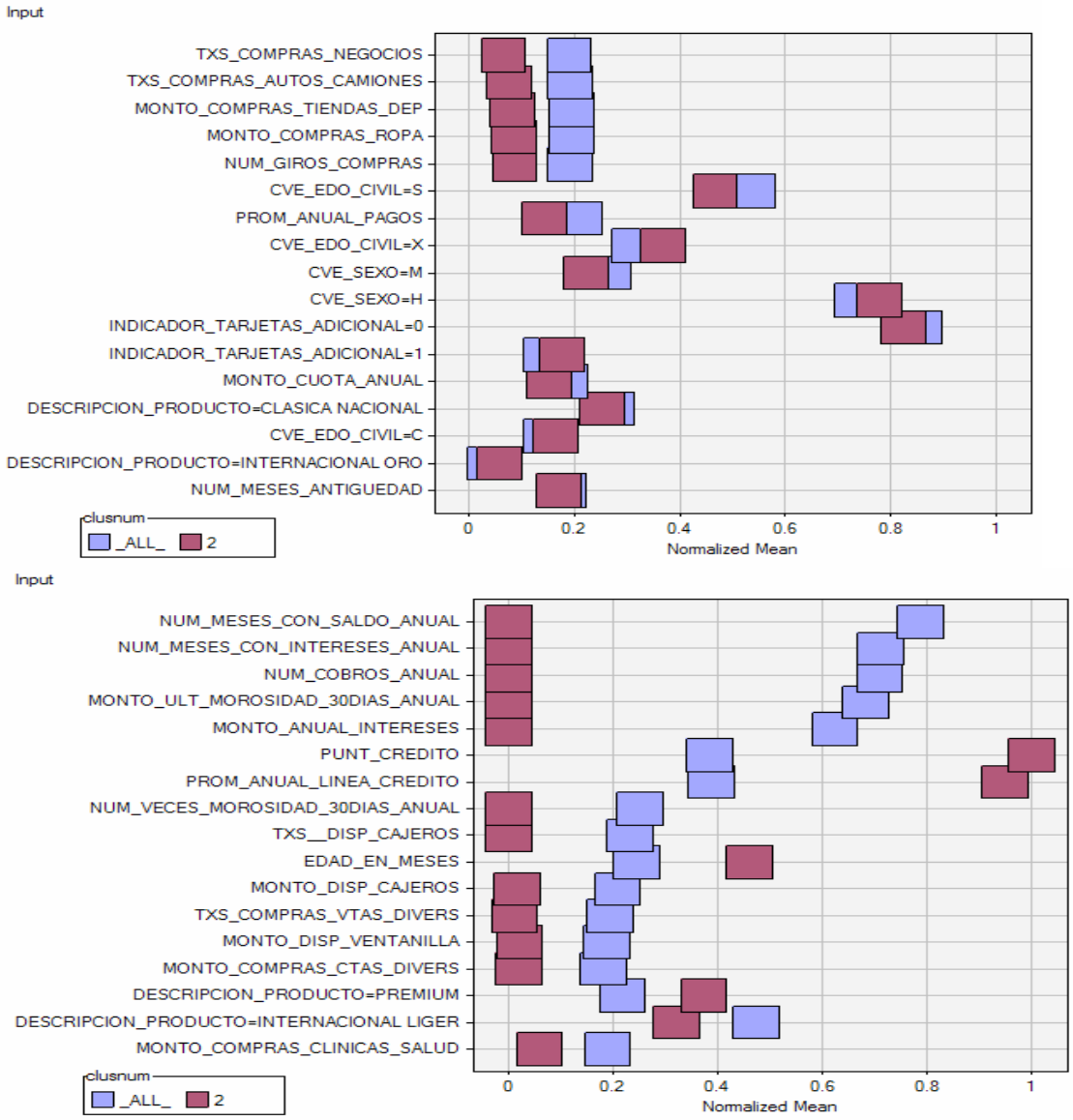


Figura 5.18 cluster 2

Los clientes de este segmento tienen la línea de crédito más alta, en su gran mayoría son hombres, tienen un uso moderado de la tarjeta ligeramente menor al promedio de la población, prácticamente no generan intereses, en su gran mayoría son tarjetas Premium (figura 5.18).

CLUSTER 3

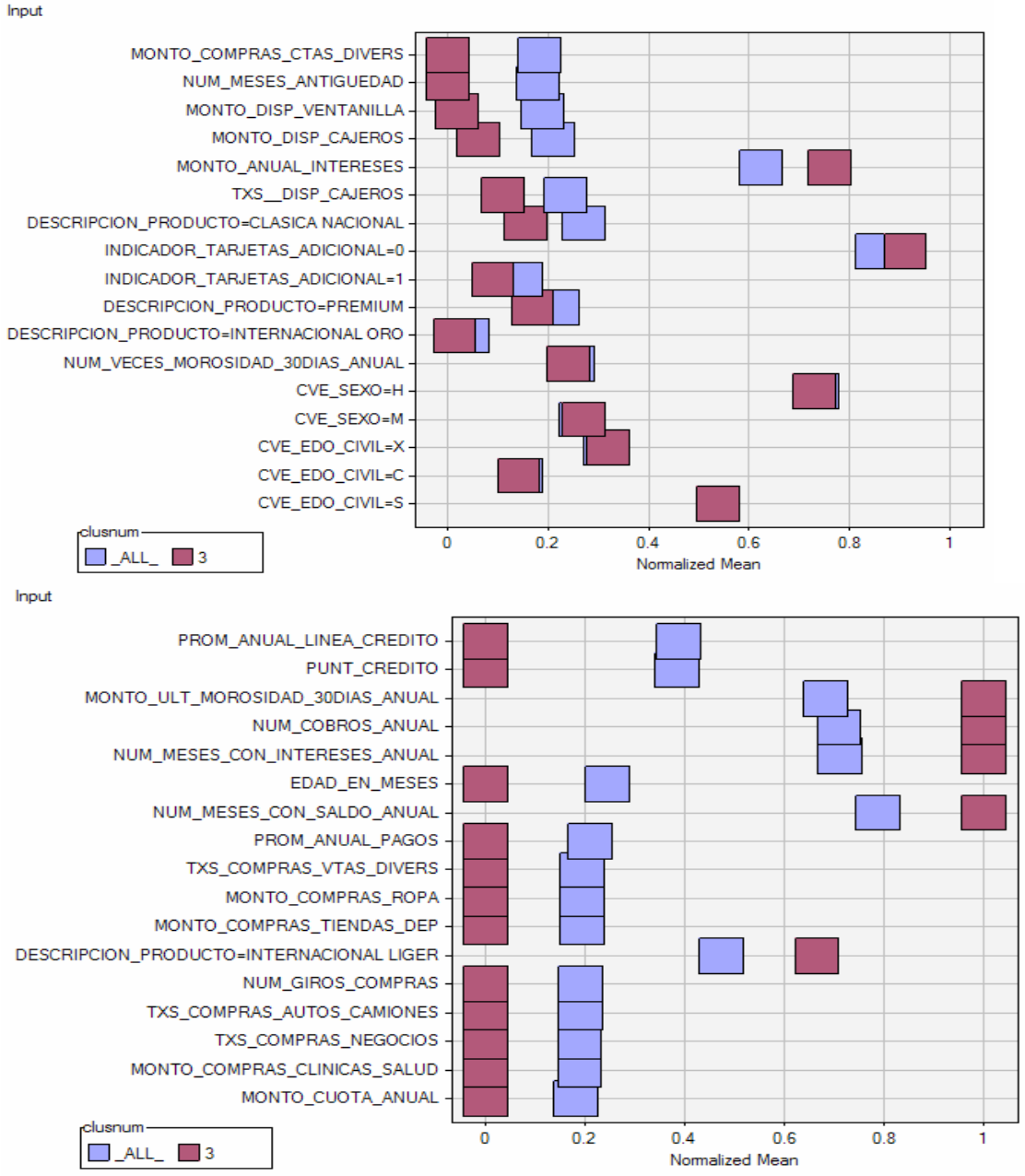


Figura 5.19 cluster 3

Los clientes de este segmento tienen la menor línea de crédito, generan muchos intereses, tienen el mayor saldo revolvente de la población, son los más jóvenes en cuanto a edad y antigüedad con la tarjeta, en su gran mayoría son tarjetas Internacional Ligera, en general sus transacciones son de disposiciones, más que de compras y en su gran mayoría son hombres solteros (figura 5.19).

CLUSTER 4

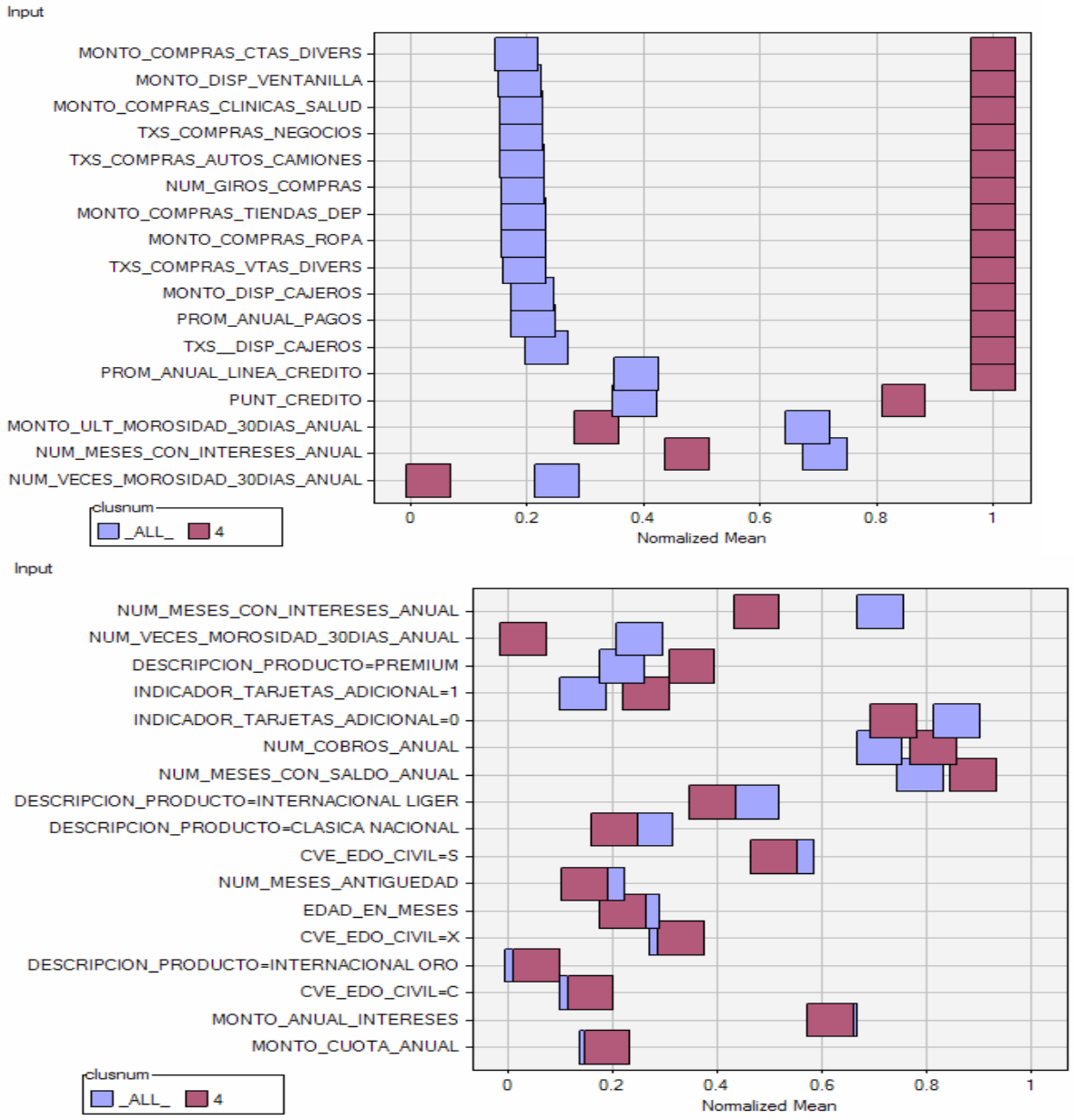


Figura 5.20 cluster 4

Los clientes de este segmento tienen alta transaccionalidad, tienen la mayor línea de crédito, son puntuales en sus pagos, normalmente no caen en mora y tienen en promedio mas tarjetas adicionales, generan en promedio intereses iguales al de la población (figura 5.20).

5.4. ANÁLISIS DE LOS RESULTADOS

De acuerdo al análisis anterior, los segmentos quedarían de la siguiente manera:

Los tarjeta habientes del Cluster 1 tienen poca línea de crédito, se retrasan en sus pagos y son los que tienen más antigüedad.

Los tarjeta habientes del Cluster 2 tienen la línea de crédito más alta, en su mayoría son hombres y tienen un uso moderado de la tarjeta.

Los tarjeta habientes del Cluster 3 tienen la línea de crédito mas baja, son los más jóvenes y utilizan la tarjeta prácticamente para disposiciones, son altos revolventes.

Los tarjeta habientes del Cluster 4 usan mucho la tarjeta, tienen la línea de crédito más alta y son muy puntuales en sus pagos.

5.4.1. Específicas según los resultados

Después de hacer el análisis de clusters e interpretar los resultados que finalmente llevarlos a reglas de negocio; tenemos suficientes elementos para hacer la toma de decisiones. Estos estudios son muy útiles para realizar campañas dirigidas en lugar de masivas es mucho mas probable que el cliente acepte la oferta si ya se estudiaron sus características permitiéndonos conocer sus tendencias y preferencias; además evita un gasto innecesario al enviar una oferta al total de la población.

- Decisiones tomadas sobre el cluster1

Derivado de que en la gran mayoría son clientes solteros y se retrasan demasiado en sus pagos, podríamos llevar a cabo las siguientes acciones:

Ofrecer un plan de pagos fijos.

Por pago puntual y llegando a un 50% de utilización, incrementar la Línea de Crédito en un 20%.

Inhibir la disposición de efectivo hasta que el porcentaje de utilización llegue al 20%.

- Decisiones tomadas sobre el cluster 2

Los clientes de este segmento, son en su gran mayoría Totaleros, por lo que nos interesa que revuelvan y generen mas intereses, para ello se proponen las siguientes acciones:

Ofrecer préstamos preaprobados que afecten directamente a la línea de crédito, con tasa de interés preferencial.

Ofrecer Meses Sin Intereses con establecimientos y tiendas departamentales.
Ofrecer domiciliaciones de servicios (Luz, Televisión de Paga, Escuelas, telefonía, etc.)

- Decisiones tomadas sobre el cluster 3

Como son los clientes más jóvenes, con líneas de crédito bajas y que prácticamente la usan para disposiciones, se proponen las siguientes acciones para incentivar las compras:

Ofrecer como incremento de línea, el promedio de las compras del mes, esto es, Línea de Crédito + promedio de compras.
Sistema de descuento de tasa por compras.
Ofrecer meses sin intereses en viajes.

- Decisiones tomadas sobre el cluster 4

Estos clientes son los mas rentables, por lo que hay que cuidar su comportamiento y ofrecer tasas bajas con respecto al mercado para que se vuelvan clientes promotores, por lo que podríamos sugerir:

Ofrecer tarjetas adicionales sin cuota anual por el primer año.
Ofrecer tasas preferenciales.
Una segunda tarjeta.

CONCLUSIONES

La implementación de este modelo (en SAS Enterprise Miner y la población a clasificar se encuentra en oracle) fue muy importante para el “Banco” ya que nos ayudo a identificar los patrones de comportamiento entre los clientes de tarjeta de crédito y de esta manera realizar a partir de su culminación en febrero de 2006 tres campañas dirigidas, con gran éxito. Permitiéndonos de esta manera fortalecer relaciones e incrementar ganancias.

Cabe mencionar que entre las mejoras que se le pueden hacer al modelo se encuentra la integración de un nodo mas que nos permita eliminar outliers (valores extremos) de la población y de esta manera evitar que se integren como otro cluster.

Para poder lograr la realización del modelo fueron necesarios varios elementos que a continuación se mencionan:

Un equipo de trabajo completo se requiere tanto de expertos en minería de datos como en el negocio ya que es importante conocer la situación actual de las organizaciones y definir claramente las necesidades ya que a partir de estas se establecerán objetivos y metas que deberán cumplirse a partir de la implementación de un modelo de minería de datos.

Una Estrategia de Datos: Las empresas invierten millones de dólares en sistemas que extraen datos de todas las fuentes imaginables. La planificación de recursos empresariales, la gestión de relaciones con el cliente, el punto de venta y otros sistemas garantizan que todas las transacciones y demás intercambios relevantes dejen su marca. Pero para competir sobre la base de esta información, las empresas deben presentarla en formatos uniformes, integrarla, almacenarla y volverla fácilmente accesible a todos y cada uno de sus empleados. Y necesitan enormes cantidades de datos. Por ejemplo una empresa puede pasar varios años acumulando

datos sobre distintos enfoques de marketing hasta que alcanza la cantidad que necesita para analizar de manera confiable la efectividad de una campaña publicitaria.

Software de Inteligencia de Negocios: Abarca un amplio abanico de procesos y software utilizados para recabar, analizar y divulgar información con el propósito de mejorar la toma de decisiones. Las herramientas de inteligencia de negocios permiten a los empleados extraer, transformar y cargar datos para analizar y, luego, volver esos análisis accesibles en informes, alertas y tableros de control. La popularidad de la competencia analítica es, en parte, el resultado de la aparición de paquetes integrados de estas herramientas.

La mayoría de las empresas en la mayoría de las industrias tienen excelentes motivos para perseguir estrategias impulsadas por el análisis. Casi todas las organizaciones que identificamos como competidores analíticos agresivos son líderes indiscutibles.

BIBLIOGRAFÍA

Berry, Michael E., Data Minino Techniques for Marketing, Sales and Customer Support, Ed. John Wiley & Sons, Inc, E.E. U.U., 1997.

SAS INSTITUTE, SAS Institute Methology Data Mining Projects"SAS, 1998.

Paul Smolenski, THE HANDBOOK OF DATA MINING, Ed. Nong Ye Arizona State University, 2003.

Michael J. A. Berry, Gordon S Linoff, Data Mining Techiniques for marketing sales and custumer Relationship , Ed. Wiley Publishing, 2004

Ian H. Witten, Eiben Frank, Data Mining Practical Machine Learning Tools and Techniques, Ed ELSEVIER, 2005.