

**Universidad del CEMA**

**Maestría en Dirección de Empresas**

**Utilización de clusters en la determinación de estados de  
segmentación**

**Autor: Act. Elizabeth Andrea Rasore**

**Profesor: Lic. Carlos A. Barbosa**

## Contenidos

1. Introducción	3
2. Análisis de cluster	5
2.1. Método de agrupación	6
2.2. Método de agregación	7
2.2.1. Método jerárquico	8
2.2.2. Método no jerárquico	10
3. Técnicas de segmentación orientadas	11
3.1. Análisis discriminante	11
3.2. Regresión logística	13
4. Aplicación práctica	14
4.1. Análisis de cluster	14
4.2. Análisis discriminante	15
Conclusión	19
Anexo	20
Bibliografía	23

## 1. Introducción

Todos los campos tienen la necesidad de conglomerar o juntar objetos similares. Cuando los investigadores tratan de volverse más científicos deben encontrar la necesidad de procedimientos que agrupen los objetos. Las aplicaciones prácticas para el análisis de agrupamiento son demasiadas numerosas.

A través de esta Tesina intento presentar una técnica básica de análisis de datos recolectados, cuyo objetivo es la agrupación de los mismos en grupos que sean homogéneos en sí y lo más heterogéneos posible entre sí. Expondré la utilización de métodos numéricos para el cálculo de la semejanza de los datos y su posterior agrupación según la proximidad entre cada uno. Esta práctica se denomina *análisis de cluster*.

Una vez identificados los grupos, plantearé la posibilidad de elegir uno/s y ser usado/s como mercado/s de prueba a través de un *análisis discriminante*.

Posteriormente, realizaré un caso práctico para que el lector tenga una idea más acabada de ambos métodos y para demostrar su alto grado de aplicación. Para el mismo usaré como herramienta informática el programa estadístico SPSS.

El objetivo del *análisis de cluster* es clasificar la población o conjunto de datos en un número menor de grupos mutuamente excluyentes. El objetivo del *análisis discriminante* es predecir el comportamiento de una persona con el menor error posible. Así pues, en primer lugar Ud. logra identificar un segmento de mercado y luego puede pronosticar la reacción de sus clientes.

El alcance o cobertura que el *análisis de cluster* tiene es vasto y diverso. Podría citar, a modo de ejemplo, la identificación de grupos de consumidores con comportamientos semejantes con el fin de probar diversas estrategias de marketing. Se pueden agrupar a los clientes mediante sus estilos de vida o los beneficios que buscan. Las marcas o productos también podrían agruparse con respecto al uso, para determinar qué marcas o productos son consideradas similares y, por lo tanto, competidoras. Si se planea un experimento de prueba de mercado, puede ser útil para identificar ciudades similares de

modo que diferentes programas de mercadotecnia puedan ser probados en distintas ciudades. Otro ejemplo podría ser la identificación de programas de televisión que atraen audiencias parecidas dentro de cada grupo.

## 2. Análisis de Cluster

El *análisis de cluster* es una técnica de exploración de datos observados con el fin de formar segmentos de mercado o grupos naturales y relativamente homogéneos, basados en su proximidad. Un *cluster* o conglomerado debe ser lo más homogéneo posible en sí y lo más heterogéneo posible con los demás. Los datos observados estarán en un mismo *cluster* o en otro, dependiendo de la homogeneidad entre ellos.

Debe haber una justificación teórica y una lógica que fundamentara y que guiara el análisis de agrupamiento. Teniendo en cuenta que es difícil evaluar la calidad de la agrupación, dado que no hay pruebas estadísticas para asegurar que el grupo obtenido no represente una aleatoriedad pura, es útil el análisis del analista sobre la medida del criterio y lo razonable de los resultados. Es importante preguntarse si tienen sentido los *clusters* formados.

No hay que creer que este método provee una solución perfecta, completa, única y definitiva, aunque sí ampliamente confiable. Debido a que es un método exploratorio, los resultados deben considerarse provisionales hasta que sean confirmados mediante otra muestra independiente.

La información puede estar expresada en forma de frecuencia o en términos binarios o ser de tipo continua. El escalamiento de las variables es un aspecto importante, ya que las diferencias en el mismo pueden afectar a las soluciones de conglomeración. Si las variables se caracterizan de diferentes maneras (por ejemplo, una variable se mide en dólares y la otra, en años), se recomienda la previa estandarización de las mismas, para que una no domine a las otras. Otro aspecto a tener en cuenta es evitar el uso de variables nominales (ej. 1 para casado, 2 para soltero y 3 para otros). Estas deberían convertirse en variables dummy o dicotómicas (ej. 0 para casado y 1 para otros). También debe evitarse la existencia de multicolinealidad<sup>1</sup> que se da cuando dos variables están altamente correlacionadas<sup>2</sup> porque una depende o es combinación de la otra.

---

<sup>1</sup> Existencia de una relación lineal entre las variables independientes de un modelo de regresión. Para ampliar sobre el tema, véase Gujarati, D en la Bibliografía.

<sup>2</sup> Relacionadas entre sí.

Deben incluirse todas las variables relevantes en el análisis. Si se omiten variables de interés, la solución obtenida puede ser equívoca.

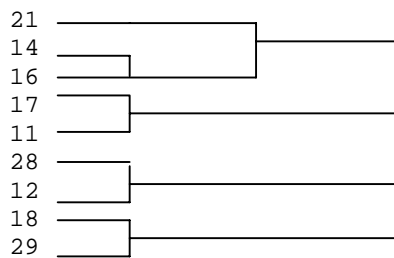
Durante el proceso, el analista tendrá que tomar dos decisiones: cuál será el *método de agrupación* y cuál, el de *agregación*. Con la primera, se observa de qué manera los individuos se unen en un mismo grupo. Con la segunda, se determina las semejanzas o diferencias entre *clusters*.

## 2.1 Método de agrupación

El investigador debe decidirse sobre un *criterio de distancia*, para saber lo próximos o parecidos entre sí que están los distintos datos (o individuos) a analizar. Esto es pues, la elección de cómo se calcularán las distancias entre cada observación. Una vez calculadas las distancias, se procede a juntar los datos más parecidos según un *método de agrupación*, llegando a grupos homogéneos en sí y diferentes grupo a grupo.

Gráficamente, resulta:

### Caso o individuo



Este es el resultado gráfico que muestra el proceso de agrupación de los casos. Vemos, por ejemplo, cómo el caso 14 se asemeja al 16 y ambos al 21.

Según sea el tipo de dato (de frecuencia, binario o continuo), existe una gran variedad de coeficientes que permiten obtener la distancia y/o similitud entre los datos observados. La explicación de cada criterio sobrepasa los límites de esta tesina<sup>3</sup>.

Cabe aclarar que no hay un cálculo de la distancia que sea el mejor para todas las situaciones posibles. Uno puede ser preferible a otro según las condiciones del análisis y

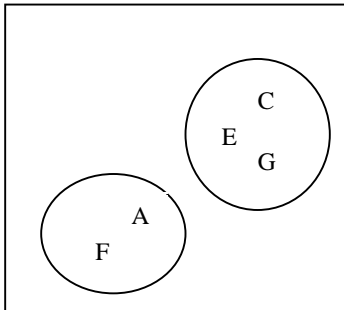
del criterio del analista. En general, el promedio de todas las distancias euclídeas al cuadrado<sup>4</sup> es el más usado. Si el lector desea profundizar sobre este tema, puede recurrir a la bibliografía recomendada<sup>5</sup>.

Es importante que quede claro que cualquiera sea la distancia que se use, Ud. debe considerar la calidad de la agrupación.

## 2.2 Método de agregación

Consiste en la decisión de cómo se van a separar los distintos *clusters*, a través de la elección del estadístico a usar sobre el cual se mide de distancia entre los grupos. Nótese la diferencia con la *agrupación*, donde se determina el tipo de distancia. Mientras que la *agregación* define sobre qué casos de cada grupo se va a aplicar esta distancia y distingue los grupos entre sí.

Gráficamente, resulta:



Vemos como los individuos A y F por su cercanía forman un *cluster* y como los individuos C, E y G forman otro conglomerado.

Existen dos enfoques alternativos: *jerárquico* y *no jerárquico*. Ambos permiten llegar a distintas soluciones a partir del mismo grupo de observaciones iniciales. Así es como el tipo de método de agregación a usar influye más que la decisión sobre cómo agrupar los datos, en el resultado final del análisis. Ud. puede realizar los dos tests (*jerárquico* y *no jerárquico*) y luego comparar los resultados usando una tabla de doble entrada.

---

<sup>3</sup> A modo de síntesis, véase la Tabla 1 en el Anexo 2.a.

<sup>4</sup> Véase Anexo 2.b.

<sup>5</sup> Véase Bisquerra, R en la Bibliografía.

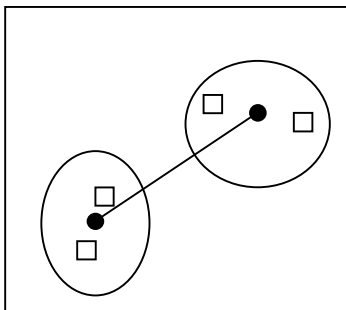
### 2.2.1 Método jerárquico

La *agrupación jerárquica* empieza con todos los objetos dentro de un grupo y los divide hasta que todos los objetos estén en su propio conjunto. Cada uno de los objetos es asociado con otro dentro de un agrupamiento. Esto hace que requiera de una gran cantidad de cálculos, lo que en ocasiones limita su posibilidad de aplicación en muestras muy grandes. La primera combinación o separación de objetos restringirá el resto del análisis, a raíz de que cuando dos datos se juntan no se separan más. Esto hace que el agrupamiento resulte ser relativamente inestable y poco confiable. Por todo lo mencionado anteriormente, no garantiza una solución óptima.

Dentro de este enfoque, existen distintas maneras de evaluar la distancia entre dos *clusters*. Ud. puede elegir entre:

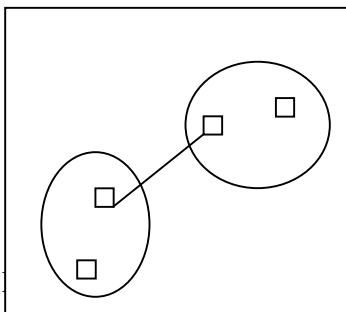
- la distancia entre las medias de cada grupo (*centroid method*)

Gráficamente,



- la distancia entre las medianas de cada grupo
- la distancia más corta entre un par de puntos, uno de cada grupo (*nearest-neighbor*)

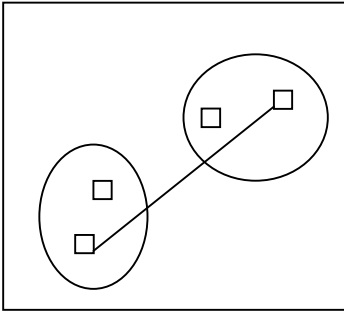
Gráficamente,





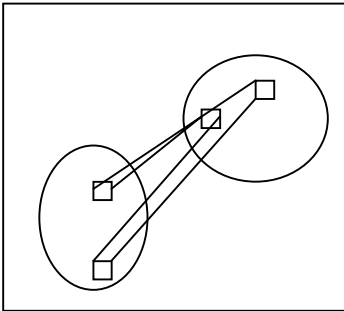
- la distancia más larga entre un par de puntos, uno de cada grupo (*furthest-neighbor*)

Gráficamente,



- el promedio de todos los posibles pares de puntos (*average linkage between groups*)

Gráficamente,



- la suma de las distancias al cuadrado (*método de Ward*).

Si el lector desea profundizar sobre este tema, puede recurrir a la bibliografía recomendada<sup>6</sup>.

Por ejemplo, para formar los conglomerados se puede medir la Distancia Euclídea (*método de agrupación*) entre las medias (*método de agregación*).

Según el criterio que elija, pueden obtenerse diferentes resultados. Para seleccionar la mejor solución, es importante el conocimiento que Ud. tenga de sus clientes y de los productos y/o servicios que quiera ofrecer.

---

<sup>6</sup> Véase Norusis, M en la Bibliografía.

### 2.2.2 Método no jerárquico

Un *agrupamiento no jerárquico* diferirá tan solo en que permitirá a los objetos, que dejen un grupo, que se unan a otro, a medida que se forman los conglomerados o *clusters*. De este modo el criterio de agrupamiento se va mejorando. Los objetos se mueven de grupo a grupo siguiendo un proceso de optimización. El proceso de asignación de las observaciones a los grupos es iterativo y no permanente entre las distintas iteraciones. Cada observación se asigna a distintos *clusters* hasta llegar a una solución óptima (lo que no se logra con los *métodos jerárquicos*).

El método más usado es el denominado K - medias. Como cada grupo queda representado por su media, no requiere de gran cantidad de cálculos, por lo cual es posible aplicarlo en bases grandes. Se utilizan especialmente variables continuas. Tiende a ser más confiable, las muestras divididas tenderán a verse más similares que en el método jerárquico.

Un problema central es determinar el número apropiado de conglomerados. Con este método, Ud. tiene la posibilidad de especificar la cantidad de grupos que desee formar, porque lo conoce o porque se deriva de un uso planeado de los mismos. En caso contrario, puede partir del patrón de grupos generados por el programa. Pero si ha seleccionado un número inapropiado de conglomerados o ha omitido variables relevantes, los resultados podrían ser equívocos.

Del mismo modo, puede definir el perfil de los mismos (si conoce de antemano esta información) y cada caso es comparado con dicho perfil.

Dicho de otro modo, Ud. elige la cantidad de *clusters* a formar. Puede especificar las medias iniciales de los conglomerados, si conoce de antemano dicha información. Si la desconoce, el centro de cada conglomerado se puede formar mediante la iteración de los casos. Durante el proceso los datos pueden cambiar de un grupo a otro (nótese la diferencia con el *enfoque jerárquico*). Se realizan tantas iteraciones (según la cantidad deseada) hasta que cada *cluster* sea lo más homogéneo posible y ningún punto cambie de lugar.

### 3. Técnicas de segmentación orientadas

El objetivo es determinar qué clientes, según ciertas características, seguirían una acción específica, discreta (comprar o no comprar, elegir producto A o B o C). Método predictivo cuyo resultado es una ecuación que permite estimar si el cliente seguirá o no la mencionada acción, dividiendo a los clientes en dos o más grupos. Esto significa, poder pronosticar el grupo de pertenencia de un caso a partir de las características observadas del mismo.

Existen dos técnicas: el *análisis discriminante* y la *regresión logística*. Ambos tienen el mismo propósito: construir un modelo predictivo. La diferencia entre ambos son las variables predictorias o independientes. El primer caso, requiere que esta variable sea continua. En el segundo caso, puede no serlo. Por supuesto, ambas técnicas coinciden en la necesidad de realizar previamente una selección de las variables a usar (especialmente para evitar multicolinealidad). Ninguna de estos dos métodos producen una lista de segmentos sino grupos de pertenencia. En la práctica, ambos arrojan resultados similares.

Para elegir entre ambos métodos debe considerarse cómo son las variables predictorias. Si estas son continuas, es recomendable el uso del *análisis discriminante*. Si son categóricas o dicotómicas, la regresión logística. De todos modos, en la práctica, ambos métodos proporcionan resultados muy similares.

#### 3.1 Análisis discriminante

Se usa para reducir distintos atributos a un número pequeño de dimensiones que discriminen o separen los objetos tanto como sea posible. El procedimiento genera una función discriminante basada en combinaciones lineales de las variables predictorias que, según su naturaleza, proporcionan la mejor discriminación posible entre los grupos. Los casos deben ser independientes. Estima los coeficientes de la función discriminante lineal que, en su parte derecha, se parece a una regresión múltiple.

Se requiere contar con un conjunto de variables que describan a cada uno de los datos o individuos (variables predictorias o independientes). Estas variables deben ser continuas. Además debe existir una variable que sea dependiente de las anteriores que permita una clasificación de los datos en grupos disjuntos entre sí<sup>7</sup>. Esta variable debe ser de tipo nominal y la única función que posee es la de indicar la pertenencia del dato o individuo a un grupo.

Es factible el uso de todas las variables que se tengan para predecir un comportamiento, sin embargo puede pasar que no todas sirvan. Para detectar las variables independientes más significativas de la variable dependiente, desechando las que no sirven, se usa el método paso a paso o *stepwise*. Este método se desarrolla en varias etapas o fases en las cuales se introducen y rechazan las distintas variables predictorias. A medida que se introducen nuevas variables puede darse la situación de que se eliminen otras variables consideradas anteriormente discriminantes. Debido a que como las funciones discriminantes son combinaciones lineales de las variables originales, las nuevas variables pueden estar correlacionadas con las anteriores, restando así su capacidad discriminante.

Existen diversos criterios de selección de las variables independientes para ser introducidas en la función discriminante. Por lo general, se recomienda la Lambda de Wilks<sup>8</sup> porque es el que menos supuestos hace sobre las variables. Si el lector desea profundizar sobre este tema, puede recurrir a la bibliografía recomendada<sup>9</sup>.

Se asume que la pertenencia al grupo es mutuamente exclusiva (es decir, ningún caso pertenece a más de un grupo) y exhaustiva de modo colectivo (es decir, todos los casos son miembros de un grupo). La variable de agrupación debe ser categórica y puede tener más de dos valores, siempre que éstos sean números enteros. Si la pertenencia al grupo se basa en los valores de una variable continua (por ejemplo, el salario), se recomienda utilizar una regresión lineal para aprovechar la información más rica ofrecida por la propia variable continua.

---

<sup>7</sup> Entiéndase por disjuntos a aquellos grupos que son mutuamente excluyentes.

<sup>8</sup> La selección de las variables introducidas se lleva a cabo mediante la maximización del cociente entre el determinante de la varianza inter-grupo y el determinante de las varianzas intragrupos.

<sup>9</sup> Véase Uriel, E en la Bibliografía.

Para una mayor comprensión, se cita el siguiente ejemplo. Un investigador supone que el clima, el PBI per capita y el índice de desempleo son factores que influyen en la discriminación que hace un individuo entre dos países (A o B) Las variables continuas predictorias son el clima, el PBI per capita y el índice de desempleo. La variable categórica de agrupación es el país.

### **3.2 Regresión logística**

Este método funciona como el anterior, con la diferencia de que acepta el uso de variables predictorias que no sean continuas (siempre que sean transformadas en variables dicotómicas)

## 4. Aplicación práctica

Se utiliza el programa estadístico SPSS.

### 4.1. Análisis de cluster

Estamos ante una empresa que vende cinco productos (A, B, C, D y E) y brinda cuatro servicios (F, G, H e I). El objetivo es determinar el uso relativo que los clientes hacen de los mismos<sup>10</sup>. Para ello, se ejecuta un *análisis de cluster* que encuentre la menor cantidad de grupos que sean homogéneos en sí y lo más heterogéneos posible entre sí. Se parte de una base de 310 clientes, donde cada producto y servicio fue convertido en una variable dicotómica<sup>11</sup>.

Se realiza un *análisis jerárquico*. El método de agrupación es el de Ward. La distancia seleccionada para proceder a observar las semejanzas entre las variables es la euclídea al cuadrado.

Del análisis surgen cuatros grupos de clientes bien diferenciados. En el siguiente cuadro, puede verse la cantidad de clientes que componen cada grupo.

	Frecuencia	Porcentaje	Porcentaje acumulado
Válidos 1	99	31.9	31.9
2	54	17.4	49.4
3	82	26.5	75.8
4	75	24.2	100.0
Total	310	100.0	

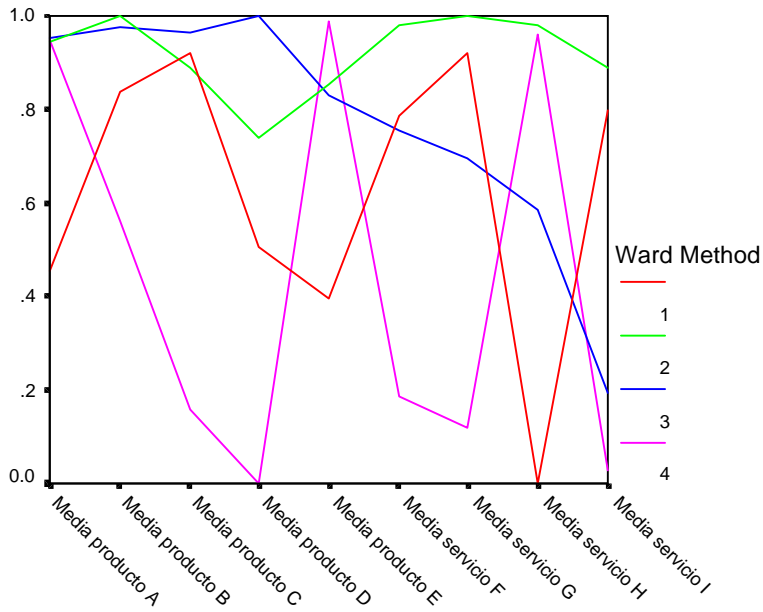
Vemos que en el *cluster* número uno hay noventa y nueve individuos que forman el 31.9% del total de casos. Nótese que los conglomerados formados son parecidos en su tamaño. Esta es una característica<sup>12</sup> del método de Ward.

<sup>10</sup> Para luego aplicar distintas estrategias de marketing según el perfil del grupo

<sup>11</sup> 0 si no lo usa, 1 si lo usa

<sup>12</sup> Formar *clusters* de tamaño parecido

El siguiente gráfico muestra el perfil medio de cada segmento.



Nótese como los cuatro conglomerados muestran diferentes perfiles. Los individuos del segundo grupo hacen un significativo uso de todos los productos y servicios de la empresa. Los clientes del grupo cuatro consumen los productos A, B y E y usan el servicio H; mientras que la utilidad que le dan al resto es muy baja.

Esta información podría emplearse, por ejemplo, para aplicar estrategias de marketing diferenciadas o selectivas.

El programa ofrece información adicional sobre el historial de aglomeración<sup>13</sup>, la matriz de distancias<sup>14</sup>, el conglomerado de pertenencia<sup>15</sup> y el gráfico del dendograma<sup>16</sup>.

## 4.2. Análisis discriminante

El objetivo es poder predecir qué cliente potencial, perteneciente a un segmento de mercado previamente identificado, se suscribiría a un nuevo servicio de cable.

<sup>13</sup> Detalla lo que ocurre en cada paso del análisis: qué individuos se agrupan y a qué distancia.

<sup>14</sup> Muestra la distancia calculada entre cada par de observaciones.

<sup>15</sup> Indica a qué *cluster* pertenece cada individuo.

Se tiene una base de datos demográficos correspondiente a los actuales clientes de la empresa de cable. En la misma, también aparecen aquellas personas que recibieron viejas promociones y rechazaron la suscripción. Cada individuo está identificado por su educación, sexo, edad, categoría de ingreso, cantidad de hijos, número de veces que cambió de compañía de cable y la cantidad de horas que mira TV por día. Estas serán las variables predictorias del análisis. Pude notarse que la mayoría son continuas, con excepción de sexo (dicotómica) y categoría de ingreso (ordinal). Recordemos que ante variables independientes continuas, es recomendable el uso del *análisis discriminante*. Esta base también tiene el resultado de anteriores promociones, donde se identifica si el potencial cliente se suscribió o no. Esta será la variable dependiente.

Se usa el método paso a paso o *stepwise* para la selección de variables independientes. Se aplica la Lambda de Wilks como criterio de entrada.

En tres pasos, el programa determinó las variables independientes significativas. Estas son edad, sexo y educación<sup>17</sup>. Las restantes no fueron incluidas<sup>18</sup>. Con las primeras, el programa genera una función discriminante basada en combinaciones lineales de las variables predictorias. El siguiente cuadro, muestra los coeficientes de cada variable.

**Coeficientes de la función de clasificación**

	Aceptó el nuevo servicio de cable	
	No	Sí
Edad en años	.325	.386
Educación en años	2.077	2.176
Sexo	1.985	1.265
(Constante)	-20.851	-24.200

Funciones discriminantes lineales de Fisher

Con estos coeficientes se construyen dos funciones que se usarán para clasificar a las nuevas observaciones. Estas son:

$$No = -20.851 + 0.325edad + 2.077educación + 1.985sexo$$

$$Sí = -24.200 + 0.386edad + 2.176educación + 1.265sexo$$

<sup>16</sup> Grafica los individuos que se juntan por su homogeneidad y a qué distancia

<sup>17</sup> Véase la Tablas 2 y 3 en el Anexo 4.c.



Supongamos que tenemos un potencial cliente con las siguientes características: 16 años de educación, sexo<sup>19</sup> femenino y 30 años de edad. Si se introduce cada valor en ambas funciones resultantes, se obtienen dos puntajes.

$$No = -20.851 + 0.325 \times 30 + 2.077 \times 16 + 1.985 \times 1 = 24.11664$$

$$Sí = -24.200 + 0.386 \times 30 + 2.176 \times 16 + 1.265 \times 1 = 23.46320$$

Visto de otra manera:

	Características del individuo	Aceptó el nuevo servicio de cable	
		No	Sí
Edad en años	30	9.75892	11.58701
Educación en años	16	33.22414	34.81134
Sexo	1	1.98503	1.26471
(Constante)		(20.85145)	(24.19986)
<b>Resultado</b>		24.11664	23.46320

El resultado más alto señala a qué grupo pertenece.

Vemos que ante una promoción, nuestro potencial cliente no aceptaría suscribirse a un nuevo servicio de cable. Así pues, se logró predecir la pertenencia de este individuo a uno de los dos grupos.

El siguiente cuadro muestra qué tan exitosa es la predicción.

**Resultados de la clasificación<sup>a</sup>**

Original	Recuento	Aceptó el nuevo servicio de cable	Grupo de pertenencia pronosticado		Total
			No	Sí	
		No	157	70	227
		Sí	72	142	214
	%	No	69.2	30.8	100.0
		Sí	33.6	66.4	100.0

a. Clasificados correctamente el 67.8% de los casos agrupados originales.

De las 227 personas que no aceptaron la suscripción, 157 fueron correctamente predichas por el modelo; esto es el 69.2 %. Con respecto a los 214 individuos que sí se

<sup>18</sup> Véase Tabla 4 en el Anexo 4.d.

<sup>19</sup> Es una variable dicotómica que toma valor 0 para masculino y 1 para femenino.

suscribieron, 142 fueron correctamente predichos; el 66.4%. Este modelo discriminante pronostica correctamente el 67.8% de los casos.

## **Conclusión**

Hemos visto una técnica matemática de análisis de datos, con el objetivo de agrupar los mismos en conglomerados que sean homogéneos en sí y lo más heterogéneos posible entre sí. Sin embargo, no hay que considerar que esta práctica provee una solución definitiva. Si bien es ampliamente confiable, es importante la interpretación del analista sobre la medida del criterio y lo razonable de los resultados.

Luego, analizamos cómo predecir el comportamiento de una persona.

## Anexo

### 2.a

Tabla 1 - Métodos de agrupación

Frecuencias
Chi - cuadrado ( $\chi^2$ )
Phi - cuadrado ( $\Phi^2$ )
VARIABLES BINARIAS
Coeficiente de Rusell y Rao
Coeficiente de parejas simples
Coeficiente de Catell
Coeficiente de SS2
Coeficiente de SS3
Coeficiente de Kendall y Jaccard
Coeficiente de Dice
Coeficiente de Roger y Tanimoto
Coeficiente de Sokal y Sneath (SS)
VARIABLES CONTINUAS
Distancia Euclídea
Distancia Euclídea Cuadrada
D <sup>2</sup> o Distancia de Mahalanobis
Distancia de Minkowsky
Cosenos de vectores
Coeficiente de correlación

### 2.b

Se denomina Distancia Euclídea a la línea recta formada por dos puntos. La Distancia Euclídea al cuadrado es la diferencia entre ambos puntos, elevada al cuadrado.

A modo de ejemplo, supongamos que existen tres variables x,y,z. La Distancia Euclídea entre dos observaciones de cada una de estas variables es

$\sqrt{(x1 - x2)^2 + (y1 - y2)^2 + (z1 - z2)^2}$  . Siguiendo con el mismo ejemplo, la Distancia Euclídea al cuadrado entre dos observaciones de cada una de estas variables es  $(x1 - x2)^2 + (y1 - y2)^2 + (z1 - z2)^2$

#### 4.c

Tabla 2

Variables introducidas/eliminadas <sup>a,b,c,d</sup>									
Paso	Introducidas	Lambda de Wilks				F exacta			
		Estadístico	gl1	gl2	gl3	Estadístico	gl1	gl2	Sig.
1	Edad en años	.907	1	1	391.000	39.947	1	391.000	.000
2	Sexo	.882	2	1	391.000	26.152	2	390.000	.000
3	Educación en años	.869	3	1	391.000	19.628	3	389.000	.000

En cada paso se introduce la variable que minimiza la lambda de Wilks global.

- El número máximo de pasos es 14.
- La F parcial mínima para entrar es 3.84.
- Maximum partial F to remove is 2.71.
- El nivel de F, la tolerancia o el VIN son insuficientes para continuar los cálculos.

Tabla 3

Variables en el análisis				
Paso		Tolerancia	F que eliminar	Lambda de Wilks
1	Edad en años	1.000	39.947	
2	Edad en años	.992	42.562	.978
	Sexo	.992	11.304	.907
3	Edad en años	.971	46.237	.972
	Sexo	.992	10.998	.893
	Educación en años	.978	5.920	.882

#### 4.d

Tabla 4

**Variables no incluidas en el análisis**

Paso		Tolerancia	Tolerancia mín.	F que introducir	Lambda de Wilks
0	Número de cambios de compañía de cable	1.000	1.000	.126	1.000
	Edad en años	1.000	1.000	39.947	.907
	Educación en años	1.000	1.000	2.738	.993
	Cantidad de hijos	1.000	1.000	4.029	.990
	Categoría de ingreso	1.000	1.000	8.029	.980
	Sexo	1.000	1.000	8.806	.978
	Horas diarias de TV	1.000	1.000	.122	1.000
1	Número de cambios de compañía de cable	1.000	1.000	.092	.907
	Educación en años	.978	.978	6.210	.893
	Cantidad de hijos	.727	.727	2.087	.902
	Categoría de ingreso	.964	.964	2.506	.902
	Sexo	.992	.992	11.304	.882
	Horas diarias de TV	.993	.993	.035	.907
2	Número de cambios de compañía de cable	.997	.989	.012	.882
	Educación en años	.978	.971	5.920	.869
	Cantidad de hijos	.718	.718	1.103	.879
	Categoría de ingreso	.851	.851	.206	.881
	Horas diarias de TV	.993	.985	.030	.882
3	Número de cambios de compañía de cable	.980	.962	.041	.868
	Cantidad de hijos	.681	.681	.259	.868
	Categoría de ingreso	.732	.732	.240	.868
	Horas diarias de TV	.909	.895	.830	.867

## **Bibliografía**

Aaker, D. *Investigación de mercados*. 3º edición. Madrid: Editorial McGraw-Hill, 1989

Bisquerra, R. *Introducción Conceptual al Análisis Multivariable. Un enfoque informático con los paquetes SPSS-X, BMDP, LISREL y SPAD*. Volumen I y II. Barcelona: Editorial Promociones y Publicaciones Universitarias S.A., 1989

Gujarati, Damodar. "Multicolinealidad". *Econometría*. 2º edición. México: Editorial McGraw-Hill, 1993, pp 213-214

Miquel, Salvador. "Aplicación del análisis de *cluster* y del análisis discriminante a la investigación comercial". *Investigación de mercados*. Madrid: Editorial McGraw-Hill, 1997, pp 269-302

Norusis, M. *SPSS for Windows. Profesional Statistics. Release 6.0*, SPSS Inc. Chicago, 1993

SPSS Manual. *Market Segmentation Using SPSS*, SPSS Inc. Chicago, 1994

Uriel, E. *Análisis de datos. Series temporales y Análisis multivariantes*. Madrid: Editorial AC, 1995