5.3. Análisis discriminante

5.3.1. Cuándo tenemos que utilizar el análisis discriminante

Para resolver muchos problemas de *marketing*, es preciso investigar las diferencias entre grupos para conocer las características distintivas de los individuos de cada grupo, identificarlas y utilizarlas para asignar a otros individuos cuya pertenencia a alguno de estos grupos desconocemos.

El análisis multivariable nos ofrece una serie de técnicas, tanto explicativas como descriptivas, para investigar las diferencias entre grupos. Entre las técnicas explicativas, se encuentra el análisis discriminante. El análisis discriminante permite conseguir dos objetivos:

1) Determinar qué variables, de entre las seleccionadas previamente, explican mejor la pertenencia de un individuo a un grupo determinado. Por ejemplo, aporta respuestas a las preguntas siguientes:

- ¿Cuáles son las características demográficas diferenciadoras entre los clientes habituales y los ocasionales de una cadena de supermercados?

- ¿Es diferente el estilo de vida de los compradores de productos de alimentación sensibles al precio del estilo de vida de los que son sensibles a la marca?

- ¿En qué se diferencian los consumidores que han respondido de una manera positiva a una campaña de *marketing* directo de los que no lo han hecho?

2) Determinar el grupo al cual pertenece un individuo pendiente de clasificación a partir de la respuesta/valor que toma en la serie de variables seleccionadas previamente. Por ejemplo:

- ¿Qué marca de coches comprará un nuevo comprador?

- ¿En qué grupo de consumo de un producto (elevado, medio o bajo) se sitúan los individuos que se acaban de incorporar al mercado?

- ¿Cuál es el riesgo (el límite de crédito) que puede darse a un cliente bancario?

Ejemplo

En un estudio cuyo objetivo consistía en determinar las características diferenciadoras entre las familias que suelen ir de vacaciones y las familias que suelen hacerlo poco o no van nunca de vacaciones, se obtuvo información sobre 300 familias. En la variable V1 se clasificaron las familias entrevistadas en función de si fueron de vacaciones en los últi-

mos dos años (valor 1) o no lo hicieron en los últimos dos años (valor 2). El resto de las variables del estudio fueron las siguientes:

V2 Ingresos anuales del hogar (en millones de u.m.).
V3 Actitud hacia los viajes (en una escala de nueve puntos en la cual 1 significaba una actitud muy negativa con respecto a los viajes, y 9, una actitud muy positiva).
V4 Importancia dada al hecho de pasar las vacaciones con la familia (en una escala de nueve puntos en la cual 1 significaba poco importante, y 9, muy importante).
V5 Tamaño del hogar (en número de individuos).
V6 Edad del principal responsable del hogar (en años).

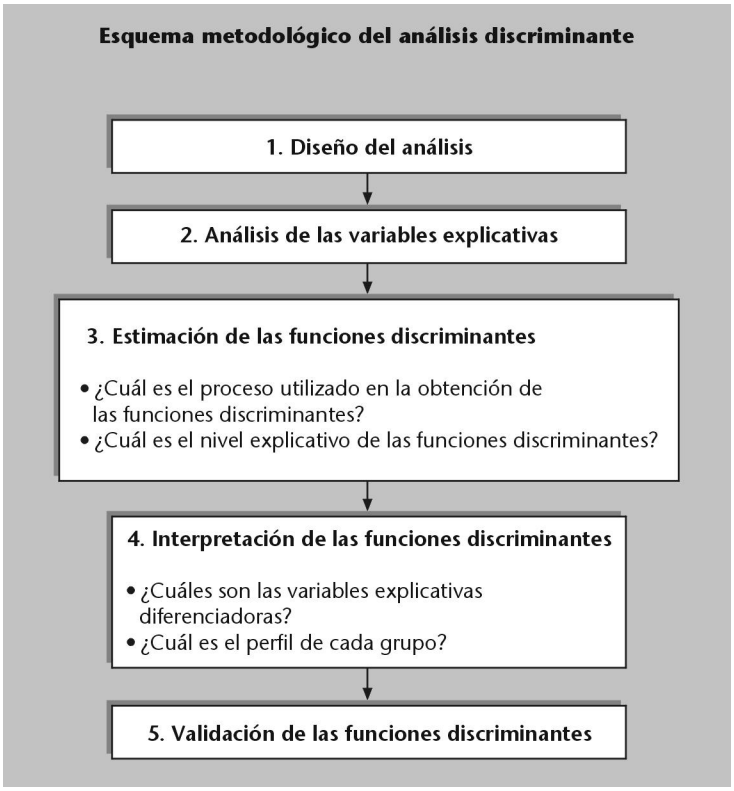
La aplicación de un análisis discriminante a este ejemplo permite conocer si las familias que fueron de vacaciones tienen un perfil diferente de las familias que no lo hicieron y cuáles de las variables V2 a V6 diferencian más a los dos grupos.

Núm.	V1	V2	V3	V4	V5	V6
1	1	5,02	5	8	3	43
2	1	7,03	6	7	4	61
3	2	6,29	7	5	6	52
4	1	4,85	7	5	5	36
5	2	5,27	6	6	4	55
7	2	4,62	5	3	3	62
8	1	5,70	2	4	6	51
9	1	6,41	7	5	4	57
10	2	6,81	7	6	5	45
11	1	7,30	6	7	5	44
.	.	.	.	.	.	
295	1	3,73	2	7	4	54
296	2	4,18	5	1	3	56
297	1	5,70	8	3	2	36
298	2	3,34	6	8	2	50
299	2	3,75	3	2	3	48
300		4,13	3	3	2	42

Matriz de datos.

5.3.2. Metodología del análisis discriminante

El proceso metodológico del análisis discriminante consta de cinco etapas:



Etapa 1: diseño del análisis

El análisis discriminante es un método de análisis explicativo que permite estudiar las relaciones entre una variable dependiente nominal (pertenencia a uno de los grupos) y un conjunto de variables independientes cuantitativas, que en investigación comercial suelen ser características socioeconómicas o sociodemográficas, hábitos de compra o de consumo, actitudes, etc. de los individuos analizados.

Ejemplo

En nuestro ejemplo, la variable dependiente del modelo discriminante es el grupo de pertenencia de las familias, es decir, el grupo 1 si fueron de vacaciones en los últimos años o el grupo 2 si no fueron de vacaciones en los últimos años. Las variables independientes del modelo son las variables V2 a V6.

Para contestar al conjunto de preguntas que se han expuesto más arriba, el análisis discriminante se centra en cubrir los aspectos que vienen a continuación:

1) Obtener unas funciones discriminantes que discriminen entre los grupos, es decir, entre las categorías de la variable dependiente (etapa 3).

2) Detectar, a partir de las variables independientes utilizadas, si hay diferencias significativas entre los grupos de la variable dependiente (etapa 4).

3) Clasificar a los individuos en uno de los grupos basándonos en los valores tomados en las variables independientes y en las funciones discriminantes obtenidas, y evaluar la precisión de la clasificación (etapa 5).

Etapa 2: análisis de las variables explicativas

Aunque no forme parte del procedimiento específico del cálculo del análisis discriminante, antes de iniciar la estimación de las funciones discriminantes conviene analizar con detalle las variables explicativas que intervienen en el modelo. Con esta finalidad obtenemos dos tipos de información:

1) Obtenemos para cada variable sus valores medios y sus desviaciones típicas dentro de cada grupo.

Ejemplo

En nuestro ejemplo, los dos grupos se diferencian mucho más en cuanto a ingresos anuales (V2) que en el resto de las variables. El grupo de las familias que han ido de vacaciones (grupo 1) es el que tiene los ingresos anuales del hogar más elevados. También se observan unas diferencias determinadas entre los grupos en la importancia concedida a las vacaciones en familia (V4). Si bien las diferencias en la edad media del responsable principal del

hogar (V6) pueden parecer considerables respecto del resto de las variables, la desviación típica elevada de esta variable hace que sea poco determinante.

```
Group means

        V1         V2         V3         V4         V5         V6

         1    6,05200    5,40000    5,80000    4,33333   53,73333
         2    4,19133    4,33333    4,06667    2,80000   50,13333

     Total    5,12167    4,86667    4,93333    3,56667   51,93333

Group standard desviations

        V1         V2         V3         V4         V5         V6

         1     ,98307    1,91982    1,82052    1,23443    8,77062
         2     ,75511    1,95180    2,05171     ,94112    8,27101

     Total    1,27952    1,97804    2,09981               8,57395
```

Valores medios y desviaciones típicas.

2) Hay un conjunto de parámetros estadísticos que nos permite determinar si cada una de las variables explicativas, de manera aislada, diferencia de forma significativa los grupos de la variable que hay que explicar:

a) La lambda de Wilks. Indica en qué medida los valores tomados por una variable explicativa son diferentes en cada uno de los grupos de la variable que hay que explicar. Su rango de variación va de 0 a 1. Valores altos de λ (próximos a 1) indican que la medida de la variable explicativa correspondiente es igual en cada grupo. En cambio, valores bajos de λ (próximos a 0) indican que la media es diferente.

Ejemplo

En nuestro caso, sólo las variables "Nivel de ingresos en el hogar" (V2) y "Tamaño del hogar" obtienen valores muy diferentes en los dos grupos.

Podéis consultar el ejemplo del subapartado 5.2.3 de este módulo.

b) El estadístico *F*. Se calcula a partir de un Anova, en el cual la variable dependiente del modelo discriminante es la variable categórica independiente del modelo Anova. Cada variable independiente del modelo discriminante se utiliza como una variable dependiente en dicho modelo. El estadístico *F* indica también el grado de influencia de cada variable explicativa por separado sobre la variable que hay que explicar.

Para el estudio del Anova, podéis consultar el subapartado 3.2.2 de este módulo.

Ejemplo

En nuestro ejemplo, las variables con más poder diferenciador son nuevamente el nivel de ingresos del hogar (V2), el tamaño del hogar (V5) y en menor medida la importancia concedida a las vacaciones en familia (V4). Los estadísticos *F* asociados a estas variables tienen un nivel de significación inferior al 5%. En cambio, la actitud hacia los viajes (V3) y la edad del responsable principal del hogar (V6) no son diferentes en los dos grupos.

Los estadísticos expuestos indican que hay unas determinadas diferencias entre los dos grupos de familias en algunas de las variables explicativas; pero debemos preguntarnos:

- ¿Se diferencian realmente las familias que han ido de vacaciones en los últimos dos años de las que no lo han hecho?
- ¿Cuáles son las variables que mejor diferencian a los dos tipos de familias?
- ¿Cuál es el perfil de cada tipo de familia?

La utilización del análisis discriminante permite contestar a estas preguntas.

```
Wilks' Lambda (U-statistic) and univariate F-ratio
with 1 and 28 degrees of freedom

Variable    Wilks' Lambda        F        Significance

V2              ,45310        33,7958          ,0000
V3              ,92479         2,2770          ,1425
V4              ,82377         5,9899          ,0209
V5              ,65672        14,6364          ,0007
V6              ,95441         1,3376          ,2572
```

Lambda de Wilks y estadístico *F*.

Etapa 3: estimación de las funciones discriminantes

El análisis discriminante estima unas funciones discriminantes en dos fases:

1) Obtención de la ecuación asociada a cada función discriminante. En general, si la variable que hay que explicar es de *m* grupos, el análisis discriminante calcula *m*-1 funciones discriminantes.

Ejemplo

Dado que en nuestro ejemplo la variable que hay que explicar es de dos grupos, obtenemos sólo una función discriminante.

La estimación de las funciones discriminantes se lleva a cabo reduciendo las variables explicativas iniciales a unas cuantas variables nuevas, combinaciones lineales de las primeras. Los valores tomados por estas variables nuevas se llaman **puntuaciones discriminantes**. Cada individuo obtiene una puntuación discriminante en cada una de las funciones discriminantes.

Si llamamos Z_i a la puntuación discriminante asociada al individuo i ($i = 1... n$) en una función discriminante cualquiera, Z_i será una combinación lineal de las variables explicativas iniciales X_p ($p = 1... P$):

$$Z_i = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + \ldots + b_pX_{pi}\ , \qquad \forall i = 1,\ldots n$$

donde b_p es el coeficiente discriminante o peso asociado a la variable X_p.

Los coeficientes discriminados o pesos b_p se estiman teniendo en cuenta que las puntuaciones discriminantes de los individuos de un grupo sean tan diferentes como sea posible de las puntuaciones discriminantes entre grupos. Esto ocurre cuando las variaciones de las puntuaciones discriminantes entre grupos, es decir, la suma de cuadrados intergrupos (SQ_{interg}), son superiores a las variaciones de las puntuaciones discriminantes dentro de cada grupo, es decir, la suma de cuadrados intragrupos (SQ_{intrag}), o, lo que es lo mismo, cuando el ratio SQ_{interg}/SQ_{intrag} sea el máximo.

La estimación de los coeficientes b_p se lleva a cabo maximizando el ratio SQ_{interg}/SQ_{intrag}.

Ejemplo

A continuación obtenemos la estimación de los coeficientes b_p asociados a la función discriminante de nuestro ejemplo:

```
Unstandardized canonical discriminant functions coefficients

                    Func 1

V2                  ,8476710
V3                  ,0496446
V4                  ,1202813
V5                  ,4273893
V6                  ,0245438
(Constant)          -7,9754761
```

Estimación de los coeficientes b_p de la función discriminante.

Así, la ecuación lineal asociada a la función discriminante es la siguiente:

$$Z_i = -7{,}98 + 0{,}85.\mathrm{V}_{2i} + 0{,}05.\mathrm{V}_{3i} + 0{,}12.\mathrm{V}_{4i} + 0{,}43.\mathrm{V}_{5i} + 0{,}02.\mathrm{V}_{6i}$$

$$\forall i = 1, \ldots, 300$$

2) Determinación del nivel explicativo de cada función discriminante. Antes de interpretar las funciones discriminantes, tenemos que asegurarnos de que su nivel explicativo es elevado, es decir, de que ayudan realmente a diferenciar los grupos de población analizados. Con esta finalidad, se utilizan los parámetros estadísticos siguientes:

a) El valor propio μ (***eigenvalue***) asociado a cada función discriminante.

$$\mu = \frac{SQ_{interg}}{SQ_{intrag}}$$

No olvidemos que las funciones discriminantes se obtienen maximizando este ratio; así, valores propios elevados implican funciones discriminantes con un poder explicativo elevado.

b) El porcentaje de varianza entre grupos explicada por cada función discriminante. Se calcula en función del valor propio asociado a cada función discriminante. Si μ_k es el valor propio asociado a la función discriminante D_k, el porcentaje de varianza entre grupos (SQ_{interg}) explicada por D_k es el siguiente:

$$V(D_k) = \frac{\mu_k}{\sum_{k=1}^{K} \mu_k}$$

donde K es el número total de funciones discriminantes.

c) La correlación canónica. Es una medida de la asociación entre cada función discriminante y la variable que hay que explicar. El cuadrado de la correlación canónica indica el porcentaje de la varianza total de la variable dependiente (SQ_T), que se explica por la función discriminante correspondiente, donde $SQ_T = SQ_{interg} + SQ_{intrag}$.

Ejemplo

En nuestro ejemplo, el cálculo de los tres primeros parámetros se presenta en el cuadro siguiente:

Fcn	Eigenvalue	Pct of Variance	Cum Pct	Canonical Corr.
1	1,7862	100,00	100,0	,8007

Parámetros determinados del nivel explicativo de cada función discriminante.

El valor propio asociado a la función discriminante es de 1,7862. Por el hecho de ser la única función discriminante, explica el 100% de la varianza entre grupos (SQ_{interg}). La correlación canónica es de 0,8007. El cuadrado de esta correlación, $(0,8007)^2 = 0,64$, indica que el 64% de la varianza total de la variable dependiente (SQ_T) se explica por la función discriminante.

Etapa 4: interpretación de las funciones discriminantes

En primer lugar, debemos **analizar la importancia relativa de cada variable explicativa** en la diferenciación de los grupos. Si todas las variables explicativas tienen el mismo rango de variación, los coeficientes iniciales b_p indican el peso de cada variable explicativa en la diferenciación de los grupos. En cambio, si los rangos de variación de las variables explicativas son diferentes, caso habitual en investigación comercial, hay que utilizar los coeficientes b_p normalizados, es decir, estimados a partir de las variables iniciales normalizadas.

Ejemplo

En nuestro ejemplo, nos encontramos en la segunda situación, dado que las variables independientes se han medido en millones de u.m. (V2), en escalas de intervalo (V3, V4), en número de individuos (V5) y en años (V6).

Las variables con coeficientes elevados, tanto positivos como negativos, son las que contribuyen más al poder discriminador de las funciones.

Otra manera de determinar la importancia relativa de cada variable explicativa consiste en analizar las correlaciones entre cada variable y las funciones discriminantes. Estas correlaciones representan el porcentaje de la varianza de cada variable que está explicada por cada función discriminante. Correlaciones elevadas, tanto positivas como negativas, indican niveles explicativos elevados para las variables explicativas correspondientes. Un coeficiente b_p elevado indica una correlación elevada y viceversa.

Ejemplo

A continuación presentamos los coeficientes b_p normalizados, y las correlaciones entre las variables explicativas y la función discriminante en nuestro caso. En los dos indicadores, la variable "Ingresos anuales del hogar" (V2) es la más importante a la hora de discriminar entre los dos grupos de familias, seguida del "Tamaño del hogar" (V5) y la "Importancia dada a las vacaciones en familia" (V4).

Standardized canonical discriminant function coefficients

	Func 1
V2	,74301
V3	,09611
V4	,23329
V5	,46911
V6	,20922

Pooled within-groups correlations between discriminating variables and canonical discriminant functions (Variables ordered by size of correlation within function)

	Func 1
V2	,82202
V3	,54096
V4	,34607
V5	,21337
V6	,16354

Coeficientes b_p estandarizados y correlaciones entre cada variable y la función discriminante.

A partir de estos resultados, podemos representar de forma gráfica la función discriminante obtenida, teniendo en cuenta los aspectos siguientes:

1) Sólo se posicionan en la función discriminante las variables con correlaciones o coeficientes normalizados elevados.

2) Si la correlación de una variable explicativa con la función discriminante es positiva, valores altos de la variable en cuestión implican puntuaciones discriminantes (Z_i) altas en la función discriminante. En este caso, situamos los valores altos de la variable en la parte positiva de la función, es decir, a la derecha, y los valores bajos en la parte negativa, es decir, a la izquierda. En cambio, si la correlación es negativa, unos valores altos de la variable implican unas puntuaciones discriminantes bajas y unos valores bajos implican unas puntuaciones elevadas. En este caso, los valores altos de la variable se sitúan en la parte negativa de la función, y los valores bajos, en la parte positiva.

Ejemplo

En nuestro ejemplo, las correlaciones más altas (V2 y V5) son positivas, y obtenemos la representación siguiente:

Representación gráfica de la función discriminante

- ingresos anuales del hogar bajos
- tamaño pequeño del hogar

⟷

- ingresos anuales del hogar elevados
- tamaño grande del hogar

El paso siguiente consiste en determinar las características diferenciadoras de cada grupo. Con esta finalidad, el análisis calcula la puntuación discriminante media de cada grupo. Se obtienen sustituyendo en la función discriminante cada variable explicativa por su valor medio dentro del grupo:

La puntuación discriminante media...

... también se denomina *centroide del grupo*. En el caso de nuestro ejemplo, Z_1 y Z_2.

```
   Canonical discriminant functions evaluated at group means
                    (group centroids)

Group      Func 1

    1     1,29118
    2    -1,29118
```

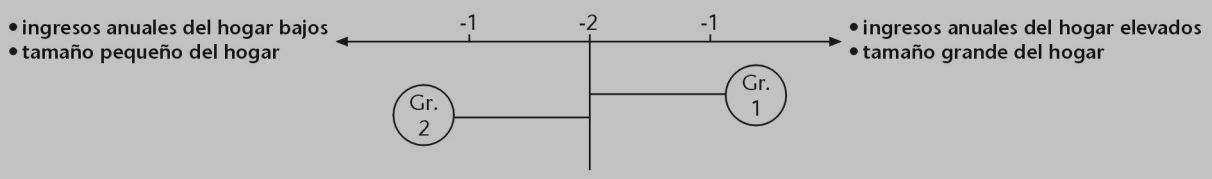
Centroides de cada grupo.

Posicionando estos dos valores en la función discriminante, podemos definir el perfil medio de cada grupo.

Ejemplo

En nuestro ejemplo, las familias del grupo 1 (familias que han ido de vacaciones en los dos últimos años) disponen de unos ingresos anuales más elevados y son más numerosas. En cambio, las del grupo 2 (familias que no han ido de vacaciones en los dos últimos años) tienen ingresos anuales más bajos y son menos numerosas.

Posicionamiento de los grupos en la función discriminante



Etapa 5: validación de las funciones discriminantes

El proceso se lleva a cabo de la manera siguiente:

1) Cada individuo obtiene una puntuación discriminante Z_i al multiplicar los coeficientes no estandarizados b_p de la función discriminante por los valores tomados en las variables explicativas correspondientes.

2) Cada individuo está asignado a uno de los grupos basándose en su puntuación discriminante y en una regla de asignación adecuada. En el caso de una variable que hay que explicar de dos grupos, un individuo está asignado al gru-

po cuya puntuación discriminante media (centroide) sea más próxima a su puntuación discriminante.

3) Si, una vez asignados, todos los individuos se vuelven a clasificar en su grupo inicial de pertenencia, obtenemos el 100% de individuos bien clasificados y podemos concluir que la función discriminante encontrada explica la totalidad de las diferencias entre grupos. En la práctica, difícilmente suele ser así; podría considerarse un porcentaje razonable de individuos bien clasificados aquel que es superior en un 25% al que se obtendría clasificando de forma correcta a los individuos al azar. Por ejemplo, cuando los grupos analizados son del mismo tamaño, el porcentaje de individuos correctamente clasificados al azar es de un individuo por número de grupos. En el caso de dos grupos, el porcentaje de individuos correctamente clasificados tendría que ser superior al 62,5% (50% = 50% × 0,25). Los grupos nuevos resultantes del proceso de asignación suelen llamarse **grupos predichos.**

4) Los resultados del proceso de asignación se presentan en una matriz de clasificación obtenida a partir del cruce entre los grupos iniciales de pertenencia de los individuos y los grupos resultantes de la aplicación de las funciones discriminantes (grupos predichos). Esta matriz indica el porcentaje de individuos clasificados correctamente una vez efectuada la asignación.

Ejemplo

En nuestro ejemplo, todos los individuos del grupo 2 vuelven a su grupo inicial y se reasignan a partir de la función discriminante. En cambio, 30 individuos del grupo 1 se clasifican en el grupo 2, con lo que el porcentaje de individuos clasificados correctamente en este grupo es del 80%. El porcentaje total de individuos correctamente clasificados es del 90%. Este porcentaje se obtiene sumando los casos bien clasificados y dividiendo por el número total de casos.

$$\left(\frac{120+150}{300}=0,9\right)$$

Podemos considerar válida la función discriminante.

Classification results

	No. of cases	Predicted Group Membership	
Actual group		1	2
Group 1	150	120	30
		80,0%	20,0%
Group 2	150	0	150
		,0%	100,0%

Percent of "grouped" casrs correctly classified: 90,00%

Resultados del proceso de asignación.

La extensión del análisis discriminante a una variable que hay que explicar de más de dos grupos incluye las mismas etapas.

5.3.3. Aplicaciones del análisis discriminante

Las aplicaciones habituales del análisis discriminante en investigación comercial son las siguientes:

1) Determinar cuáles son las variables que explican mejor la pertenencia de un individuo a un grupo determinado *a priori*. Por ejemplo, permite contestar a las preguntas siguientes:

- ¿Cuáles son las variables fundamentales que explican el consumo de una marca o de otra?

- ¿Hay diferencias entre innovadores y tradicionales de acuerdo con sus perfiles?

2) Determinar con una finalidad predictiva el grupo al cual pertenece un individuo pendiente de clasificación, ya sea porque durante la entrevista no ha manifestado su grupo de pertenencia o porque es un individuo que no forma parte de la muestra analizada: comprador nuevo de un producto, consumidor nuevo, etc.

Ejemplo

Consideremos a un grupo de consumidores que consumen exclusivamente una de las tres marcas siguientes: A, B o C Para cada uno de estos consumidores se dispone de una información sobre determinadas características (variables de actitud, socioeconómicas, etc.). Lo que nos permite el análisis discriminante es:

1. Encontrar, en una primera fase, cuáles son las variables fundamentales para explicar el consumo de una marca o de otra.

2. Con posterioridad, mediante el estudio de los valores que toman estas variables para un consumidor nuevo, el análisis discriminante predecirá, por medio de un proceso de asignación idéntico al que se ha utilizado en la validación de las funciones discriminantes (podéis consultar la etapa 5), la marca que comprará o bien la marca que tiene más probabilidades de comprar.

Los ejemplos de aplicación con finalidad predictiva del análisis discriminante abundan en investigación comercial, y pueden señalarse los siguientes: predecir el riesgo (límite de crédito) que puede darse a un cliente en función de su perfil socioeconómico, predecir la marca que comprará un comprador nuevo, etc.