CMA IADCOM - UBA | CENTRO DE INVESTIGACIÓN EN METODOS CUANTITATIVOS APLICADOS A LA ECONOMÍA Y LA GESTIÓN

Universidad de Buenos Aires
Facultad de Ciencias Económicas

http://www.economicas.uba.ar/institutos_y_centros/revista-modelos-matematicos/

ANÁLISIS DISCRIMINANTE Y TEOREMA DE BAYES PARA LA CLASIFICACIÓN EN GRUPOS*

Vitale, Blanca Rosa

Universidad de Buenos Aires. Facultad de Ciencias Económicas. Av. Córdoba 2122 – 1120AAQ. Ciudad Autónoma de Buenos Aires, República Argentina

blancavitale11@gmail.com

Resumen

Recibido: 16-10-2023

Aceptado: 15-12-2023

Palabras clave

Clasificación en grupos,
Función discriminante,
Teorema de Bayes.

El Análisis Discriminante es una técnica multivariada de dependencia, de suma utilidad para la clasificación en grupos. Junto con el Análisis de Conglomerados y el Análisis Multivariado de la Varianza, se utiliza para clasificar a individuos en distintos grupos, a partir del empleo de variables categóricas, con fines explicativos y predictivos. Se desarrolla en el ámbito de la ciencia de datos, y resulta una sinergia entre la teoría y la evidencia empírica, a la luz de los grandes volúmenes de datos que permite gestionar la tecnología.

Por un lado, explica la pertenencia de cada individuo a un grupo determinado, de acuerdo con las variables clasificadoras elegidas. Por otro lado, predice a qué grupo podría pertenecer el nuevo individuo en función de la información que se posee sobre las variables clasificadoras. Se muestra en el desarrollo del trabajo cómo las variables predictoras pueden tomarse tanto en forma individual como en forma conjunta, mediante la aplicación de un recurso matemático basado en la aplicación de la Función Discriminante de Fisher.

En este trabajo, se aborda la técnica multivariada del análisis discriminante considerando dos variables clasificadoras. El análisis se lleva a cabo no sólo a través del uso de puntuaciones discriminantes sino mediante el cálculo de la probabilidad de pertenencia a un grupo determinado. Dicho cálculo se realiza en el contexto de la Teoría de la decisión bayesiana. Además de clasificar a cada individuo en cada uno de los grupos preestablecidos, interesa disponer de información sobre la probabilidad de pertenencia. Para lograr este objetivo se utiliza el Teorema de Bayes, tomando como información muestral las puntuaciones discriminantes aportadas por la Función discriminante de Fisher.

ISSN (En línea) 2362 3225

* Este artículo se basa en el trabajo "*Aplicación del análisis discriminante y de la teoría de la decisión Bayesiana para la clasificación en grupos"* presentado en las *XXII Jornadas de Tecnología Aplicada a la Educación Matemática Universitaria* realizadas en la Facultad de Ciencias Económicas de la Universidad de Buenos Aires.

DISCRIMINANT ANALYSIS AND BAYES THEOREM FOR CLASSIFICATION GROUP

Abstract

KEYWORDS

Classification group
Discriminant function
Bayes Theorem

Discriminant Analysis is a multivariate dependency technique, extremely useful for classification into groups. Together with Cluster Analysis and Multivariate Analysis of Variance, it is used to classify individuals into different groups, based on the use of categorical variables, for explanatory and predictive purposes. Discriminant Analysis is developed in the field of data science, and is a synergy between theory and empirical evidence, in light of the large volumes of data that technology allows to manage.

On the one hand, it explains the membership of each individual to a specific group, according to the chosen classifying variables. On the other hand, it predicts which group the new individual could belong to based on the information we have about the classifying variables. It is shown in the development of the work how the predictor variables can be taken both individually and jointly, through the application of a mathematical resource based on the application of the Fisher Discriminant Function.

In this work, the multivariate technique of discriminant analysis is addressed considering two classifying variables. The analysis is carried out not only through the use of discriminant scores but also by calculating the probability of belonging to a given group. This calculation is carried out in the context of Bayesian Decision Theory. In addition to classifying each individual into each of the pre-established groups, it is interesting to have information on the probability of membership. To achieve this objective, Bayes' Theorem is used, taking as sample information the discriminant scores provided by Fisher's Discriminant Function.

ISSN (En línea) 2362 3225

INTRODUCCIÓN

El análisis multivariante es una rama de la estadística que consiste en un conjunto de técnicas que pueden ser utilizadas para analizar múltiples variables simultáneamente. El principal objetivo de la aplicación de técnicas multivariantes consiste en la reducción de la dimensionalidad de un conjunto de datos multivariantes. Su finalidad es encontrar patrones, relaciones y asociaciones entre las variables. Así como el análisis univariante se enfoca en una sola variable, el análisis multivariante examina la interacción entre múltiples variables.

Un problema que surge en la ciencia de datos consiste en contar con un *dataset* que contiene una alta dimensionalidad. Se lo conoce como la *maldición de la dimensionalidad*, frase que se le atribuye a Richard Bellman (1957), también conocida como *efecto Hughes*.

La *maldición de la dimensión* se refiere a la problemática que se presenta al analizar y organizar datos de espacios de múltiples dimensiones. Estos fenómenos se presentan en la minería de datos, el aprendizaje automático y el muestreo, entre otros.

Al aumentar la dimensión, crece exponencialmente la cantidad de datos, lo cual dificulta tanto su organización como su búsqueda. Las técnicas multivariantes se clasifican en técnicas de análisis de dependencia y técnicas de análisis de interdependencia. En todos los casos es importante reconocer como están medidas las variables involucradas, y por lo tanto, cuál debe ser el tipo de escala de medición a utilizar. Las escalas pueden ser métricas o no métricas.

Las técnicas de análisis de dependencia se utilizan para buscar la existencia o ausencia de relaciones entre grupos de variables. Tienen por objetivo establecer si el conjunto de variables independientes afecta al conjunto de variables dependientes individualmente o de manera conjunta. Algunas técnicas multivariantes de análisis de dependencia se mencionan a continuación:

- Ecuaciones estructurales
- Regresión lineal múltiple
- Regresión de variable dependiente limitada
- Análisis discriminante
- Análisis de la varianza
- Análisis multivariado de la varianza
- Correlación canónica

Puede ocurrir que no sea posible distinguir entre variables dependientes e independientes y sea de interés conocer cómo se relacionan entre sí todas las variables de un problema. En esos casos conviene utilizar las técnicas de análisis de interdependencia. Un análisis de interdependencia es aquel en el que ninguna variable o grupo puede distinguirse como dependiente o independiente y deben analizarse todas las variables simultáneamente. Las técnicas de análisis de interdependencia más utilizadas son las siguientes:

- Análisis factorial
- Componentes principales
- Análisis de conglomerados
- Análisis de correspondencias

- Escalamiento multidimensional

A continuación, se presentará una breve descripción de las técnicas multivariantes más utilizadas para la clasificación en grupos, a saber: análisis de conglomerados, análisis multivariado de la varianza, y en particular, análisis discriminante.

1. TÉCNICAS MULTIVARIANTES PARA LA CLASIFICACIÓN EN GRUPOS HOMOGÉNEOS

Si se tiene como objetivo clasificar a individuos en grupos de cierta homogeneidad que presenten características comunes, los métodos multivariantes más utilizados son el análisis de conglomerados o análisis clúster, el análisis multivariado de la varianza y el análisis discriminante.

En este apartado se realizar una breve descripción de los tres métodos. En el apartado siguiente se desarrollará de modo detallado la técnica de análisis discriminante, objeto de este trabajo.

1.1. TÉCNICA ANÁLISIS MULTIVARIANTE DE LA VARIANZA

El análisis multivariante de la varianza de la varianza o MANOVA (en inglés *Multivariante Análisis of Variance*) es una extensión del análisis de la varianza o ANOVA (*Análisis of Variance*) para aquellos casos en los que hay más de una variable dependiente. Es una técnica multivariante de dependencia que pretende identificar si los cambios en las variables independientes tienen efectos significativos en las variables dependientes, pero también trata de detectar las interacciones entre las variables independientes y su grado de asociación con las variables dependientes. El conjunto de variables dependientes es un vector en el que cada uno de los elementos es una variable dependiente. Las variables dependientes son métricas y las variables independientes (también llamadas factores) son no métricas. Se utilizan herramientas de la estadística inferencial para tomar decisiones acerca de la existencia de diferencias significativas entre las poblaciones.

1.2. TÉCNICA ANÁLISIS DE CONGLOMERADOS

El análisis *cluster* o análisis de conglomerados es una técnica multivariante de interdependencia que se aplica sobre los individuos. No es de carácter inferencial, sino que es una técnica de análisis exploratorio que aplica algoritmos en etapas sucesivas, con el objeto de agrupar individuos que tengan características similares. Se trata de formar grupos lo más homogéneos en sí y lo más heterogéneos entre sí. La composición de los grupos es *desconocida a priori* y es necesario derivarlas a partir de las observaciones. Para determinar la proximidad entre cada par de observaciones se utilizan medidas de similaridad. Si las variables que se utilizan para caracterizar las observaciones son métricas, es decir, admiten una escala de intervalo o de razón, la medida más utilizada es la distancia euclídea, o la distancia euclídea la cuadrado.

Si se consideran dos observaciones *i* y *j* de las *n* posibles, la distancia euclídea está dada por (1)

$$D_{i,j} = \sqrt{\sum_{p=1}^{k} \left(x_{i,p} - x_{j,p}\right)^2} \qquad (1)$$

donde:

$$x_{i,p} : valor\ que\ toma\ la\ observacion\ i\ de\ la\ variable\ x_p$$

$$x_{j,p} : valor\ que\ toma\ la\ observacion\ j\ de\ la\ variable\ x_p$$

$$p = 1,2, \ldots\ldots\ldots\ldots, k$$

Los métodos de análisis de conglomerados se clasifican en dos grupos: métodos jerárquicos y métodos no jerárquicos. En los métodos jerárquicos no se conoce a priori el número de conglomerados que se pueden formar, mientras que en los métodos no jerárquicos se conoce a priori el número de grupos. En los métodos jerárquicos inicialmente cada individuo es un grupo en sí mismo. Sucesivamente, se van formando grupos de mayor tamaño fusionando elementos cercanos entre sí. En los métodos no jerárquicos los grupos no se forman en un proceso de fusión de grupos de menor tamaño. Se establece un número de grupos a priori y los individuos se van clasificando en cada uno de esos grupos. Una solución es aquella donde los miembros de cada grupo son lo más homogéneos en sí y lo más heterogéneos entre sí. En cuanto al número ideal de grupos que deben constituirse, autores como Hair *et al* (2014) señalan que no existe un criterio objetivo ya que no se ha construido un estadístico de prueba que permita establecer un criterio de decisión sobre la base de la inferencia estadística (Hartigan, 1985; Bock, 1985) y proponen métodos alternativos.

1.3. TÉCNICA ANÁLISIS DISCRIMINANTE

El análisis discriminante es una técnica multivariada de dependencia mediante la cual la pertenencia a un grupo se introduce a través de una variable categórica que toma tantos valores como grupos existentes. Para llevar a cabo dicha clasificación se tienen en cuenta variables explicativas, que también reciben el nombre de predictores. A diferencia del análisis multivariado de la varianza, en el análisis discriminante la variable dependiente es categórica y las variables explicativas son métricas. Se utiliza con fines tanto explicativos como predictivos. En cuanto a los fines explicativos, se utiliza para determinar el aporte de cada predictor a la clasificación acertada de un individuo a un grupo. Con respecto a los fines predictivos, sirve para determinar el grupo al que podría pertenecer un individuo con probabilidad más alta.

Dado que este trabajo se focaliza esencialmente en la aplicación de la técnica multivariante del análisis discriminante, la misma será desarrollada con mayor nivel de detalle en el próximo apartado.

2. ANÁLISIS DISCRIMINANTE

La aplicación de la técnica multivariante denominada análisis discriminante que se presenta en este trabajo se focaliza en detección de la capacidad de pago de tomadores de crédito, mediante una herramienta eficiente conduce a la toma de decisiones apropiadas al clasificar a los individuos en grupos de confiabilidad, y facilita los procesos de selección a la hora de gestionar la oferta crediticia, conduciendo a una mayor competitividad en la aplicación de métodos y procedimientos.

El análisis discriminante se desarrolla en el ámbito de la ciencia de datos, y resulta una sinergia entre la teoría y la evidencia empírica, a la luz de los grandes volúmenes de datos que permite gestionar la tecnología. Por un lado, explica la pertenencia de cada individuo a un grupo determinado, de acuerdo con las variables clasificadoras elegidas. Por otro lado, predice a qué grupo podría pertenecer el nuevo individuo, del modo más fiable, en función de la información que se posee sobre las variables clasificadoras o predictoras. Se señalan a continuación algunas características de la técnica de análisis discriminante.

2.1. CARACTERÍSTICAS Y SUPUESTOS DEL ANÁLISIS DISCRIMINANTE

Se utiliza para explicar la pertenencia de individuos a grupos a partir de los valores de un conjunto de variables. Cada individuo puede pertenecer a un solo grupo.

2.1.1. CARACTERÍSTICAS

- Es una técnica de análisis de dependencia
- La pertenencia a un grupo se introduce mediante una variable categórica que toma tantos valores como grupos existentes
- Variable categórica es la variable dependiente del problema
- Variables clasificadoras son las variables independientes del problema y son las que se utilizan para realizar la clasificación de los individuos (también llamadas variables criterio, variables predictoras o variables explicativas)

2.1.2. SUPUESTOS

El método de análisis discriminante se sustenta en los siguientes supuestos:

- Las variables clasificadoras son independientes y se distribuyen normalmente
- Las matrices de varianzas y covarianzas son iguales en todos los grupos
- No hay multicolinealidad entre las variables clasificadoras
- Se requiere una muestra aleatoria multivariante independiente en cada grupo

3. REGLAS DE DECISIÓN PARA LA CLASIFICACIÓN EN DOS GRUPOS

A fin de brindar un fundamento teórico para el caso de aplicación que se presentará más adelante, se describirá en este apartado la metodología para la clasificación en grupos para el caso de dos grupos y dos variables predictoras.

La asignación de un individuo a un grupo puede efectuarse de tres maneras posibles: tomando cada variable clasificadora por separado (punto de corte), tomando una combinación lineal de las variables (función discriminante de Fisher) o estimando la probabilidad de pertenencia a un grupo determinado (Teorema de Bayes).

3.1. PUNTO DE CORTE PARA DOS VARIABLES CLASIFICADORAS TOMADAS POR SEPARADO

Considérense dos grupos, a los que llamaremos I y II , y dos variables clasificadoras X_1 y X_2, que cumplen con los supuestos señalados en **2.1.2.**

Sean $\bar{X}_{1,I}$ y $\bar{X}_{1,II}$ y las medias muestrales de la variable X_1 para los grupos I y II respectivamente, y $\bar{X}_{2,I}$ y $\bar{X}_{2,II}$ las medias muestrales de la variable X_2 para los grupos I y II, respectivamente.

Si se toman las variables por separado se calcula el *punto de corte*, que no es otra cosa más que el promedio de las medias muestrales de los grupos I y II para cada una de las variables predictoras y se establece el siguiente criterio de decisión: Para la variable X_1 , C_1 toma el valor indicado en (2)

$$C_1 = \frac{\overline{X}_{1,I} + \overline{X}_{1,II}}{2} \tag{2}$$

Para la variable X_2 , C_2 toma el valor indicado en (3)

$$C_2 = \frac{\overline{X}_{2,I} + \overline{X}_{2,II}}{2} \tag{3}$$

Si la observación $X_i < C_1$, se clasifica al individuo en el grupo I

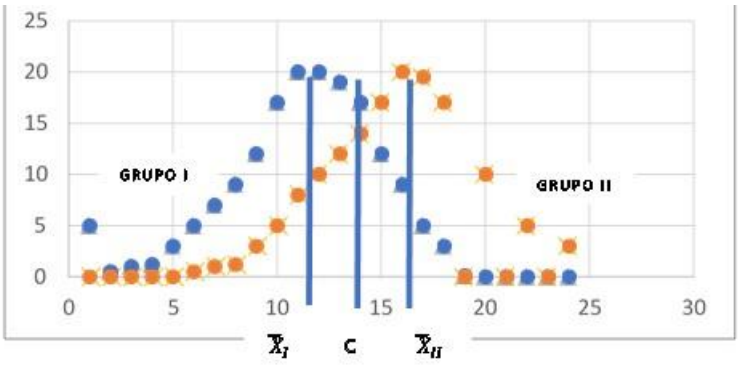
Si la observación $X_i > C_1$, se clasifica al individuo en el grupo II

Análogamente, para la variable X_2 .

De acuerdo con el tipo de variable utilizada, el sentido de la desigualdad puede invertirse.

Si se representaran gráficamente las distribuciones de frecuencias de una misma variable para caca uno de los grupos, el número de aciertos será mayor cuanto menor sea el área de solapamiento de ambas distribuciones.

Figura 1. Representación ilustrativa de las distribuciones de frecuencias para ambos grupos correspondientes a una única variable



Fuente: Elaboración propia

En el caso de que se tome cada una de las variables por separado, el número de aciertos será menor que en el caso de que se tomen las variables combinadas. Para obtener mayor precisión en la clasificación se utilizará la función discriminante de Fisher.

3.2. FUNCIÓN DISCRIMINANTE DE FISHER

La función discriminante de Fisher se construye como una función lineal de k variables clasificadoras, cuyos coeficientes se obtienen mediante la resolución de un problema de optimización planteado por Fisher. Fisher utilizó un criterio racional, en el que postula que la función discriminante será tanto mejor si tiene en cuenta que los grupos resultantes sean los más homogéneos en sí y los más heterogéneos entre sí. Los coeficientes de la función discriminante

de Fisher se obtienen buscando aquellos que maximicen el cociente entre la variabilidad entre los grupos y la variabilidad dentro de los grupos.

Una vez obtenidos los coeficientes $u_1,u_2,\ldots\ldots u_k$ de las k variables explicativas, se construye la función discriminante de Fisher D, como combinación lineal de las k variables clasificadoras que se muestran en (4):

$$D= \mu_1.X_1 + \mu_2.X_2 + \cdots + \mu_k.X_k \qquad (4)$$

$$D_i= \mu_1.X_{1,i} + \mu_2.X_{2,i} + \cdots + \mu_k.X_{k,i} \qquad (5)$$

donde D_i es la *puntuación discriminante* correspondiente a la observación i .

Si se consideran los vectores cuyas componentes son las medias muestrales de las variables explicativas, (5) y (6) para cada uno de los grupos, también llamados *centroides*

$$\overline{X_I}= \begin{bmatrix} \bar{X}_{1,I} \\ \bar{X}_{2,I} \\ \vdots \\ \overline{X}_{2,I} \end{bmatrix} \qquad (6)$$

$$\bar{X}_{II}= \begin{bmatrix} \bar{X}_{1,II} \\ \bar{X}_{2,II} \\ \vdots \\ \overline{X}_{2,II} \end{bmatrix} \qquad (7)$$

sustituyendo las componentes de los centroides en la función discriminante de Fisher , se obtiene

$$\overline{D}_I = \mu_1.\bar{X}_{1,I} + \mu_2.\bar{X}_{2,I} + \cdots + \mu_P.\bar{X}_{K,I} \qquad (8)$$

$$\overline{D}_{II} = \mu_1.\bar{X}_{1,II} + \mu_2.\bar{X}_{2,II} + \cdots + \mu_P.\bar{X}_{K,II} \qquad (9)$$

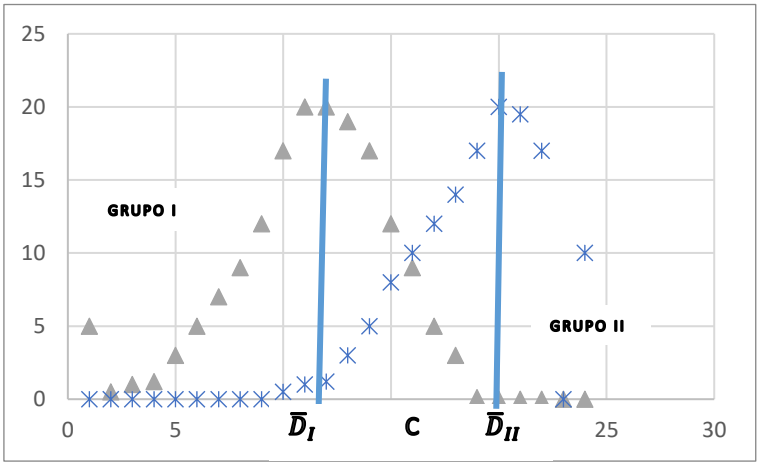
Con las ecuaciones (8) Y (9) se obtiene el *punto de corte discriminante* (10) y se postula el criterio de clasificación para el i-ésimo individuo :

$$C = \frac{\overline{D}_I + \overline{D}_{II}}{2} \qquad (10)$$

Si $D_i < C$, el individuo pertenece al grupo I y si $D_i > C$, el individuo pertenece al grupo II.

Dependiendo de los tipos de variables utilizadas, las desigualdades pueden invertirse.

Figura 2. Representación ilustrativa de las distribuciones de frecuencias para ambos grupos si se utiliza la Función discriminante Fisher



Fuente: Elaboración propia

Puede verificarse que si se utiliza como criterio de clasificacion a la función discriminante de Fisher se tendrán más clasificaciones correctas que si se utilizan las variables clasificadoras por separado, y por lo tanto el área de solapamiento de ambas distribuciones será menor.

4. TEOREMA DE BAYES Y PROBABILIDAD DE PERTENENCIA A UN GRUPO

Los criterios de decisión vistos hasta el momento sirven para asignar individuos a un grupo determinado, pero sería deseable contar con informacion adicional acerca de la pertenencia de un individuo a un grupo. Por ejemplo, puede ser de suma utilidad estimar la probabilidad de pertenencia de un individuo a un grupo determinado.

Para lograr este propósito se utiliza el Teorema de Bayes, desarrollado por el reverendo Thomas Bayes (1702-1761). El Teorema de Bayes ocupa un lugar destacado en la Teoría de la Probabilidad. A partir de este teorema se pueden calcular las *probabilidades a posteriori* a partir de las *probabilidades a priori* y de la informacion muestral que proporcionan las puntuaciones discriminantes.

Si consideramos la existencia de *G* grupos (entendidos como eventos mutuamente excluyentes y colectivamete exhaustivos en el contexto de Bayes), el teorema establece que la probabilidad a posteriori de pertenecer a un grupo *g* si se sabe que se obtuvo una puntuacion discriminate *D*, es decir la probabilidad condicional P (g/D) se calcula con la expresion

$$\Pr\,(g/D)=\frac{\pi g.\ \Pr\,(D/g)}{\sum_{i=1}^{G}\pi_i.\Pr\,(D/g)} \qquad (11)$$

$$\sum_{i=1}^{G}\pi_i = 1 \qquad (12)$$

donde π_g representa las probabilidades *a priori* y Prob(D/g) indican las probabilidades condicionadas. Prob(D/g) se obtiene calculando la probabilidad de la puntuación observada suponiendo la pertenencia a un grupo g. Se asignará un individuo al grupo para el cual sea mayor su probabilidad *a posteriori*.

El cálculo de probabilidades puede realizarse bajo tres supuestos diferentes: sin información a priori, en cuyo caso se supone que la probabilidad de pertenencia *a priori* a cada grupo es la misma, con información a priori y por último, con información a priori y consideración de costes.

5. CASO DE APLICACIÓN PARA DOS GRUPOS Y DOS VARIABLES CLASIFICADORAS

En este apartado se presentará un caso de aplicación con el fin de ilustrar lo expuesto hasta ahora.

CASO DE APLICACIÓN PARA EL OTORGAMIENTO DE UN CONTRATO DE LEASING EN EL SECTOR AGROPECUARIO

Una empresa de maquinarias para la agroindustria está evaluando la posibilidad de otorgar un vehículo mediante un contrato de leasing. La empresa cuenta con información sobre dieciseis productores agropecuarios a los que se les concedió

vehículo hace algunos años. Ocho de ellos cumplieron con el contrato y los ocho restantes no cumplieron. De cada locatario se conoce información sobre su capital de trabajo (KT) y su pasivo corriente (PC) al momento de la solicitud. Se presentan dos nuevos interesados. Uno de ellos posee un KT de 10,1 unidades monetarias y un PC de 6,8 unidaes monetarias. El segundo posee un KT de 9,7 unidades monetarias y un PC de 2,2 unidades monetarias (um). La información disponible se muestra en la **Tabla 1**.

Tabla 1: Datos

NO CUMPLIDORES	VARIABLE 1	VARIABLE 2	CUMPLIDORES	VARIABLE 1	VARIABLE 2
LOCATARIO	CAPITAL DE TRABAJO	PASIVO CORRIENTE	LOCATARIO	CAPITAL DE TRABAJO	PASIVO CORRIENTE
1	1,3	4,1	9	5,2	1
2	3,7	6,9	10	9,8	4,2
3	5	3	11	9	4,8
4	5,9	6,5	12	12	2
5	7,1	5,4	13	6,3	5,2
6	4	2,7	14	8,7	1,1
7	7,9	7,6	15	11,1	4,1
8	5,1	3,8	16	9,9	1,6
TOTAL	40	40		72	24
MEDIA	5	5		9	3

Se estudiará la pertenencia a cada grupo de tres maneras distintas: tomando cada variable clasificadora por separado, utilizando una combinación lineal de las mismas mediante la función

discriminante de Fisher, y por último se comentará el modo de aplicación del Teorema de Bayes para el cálculo de la probabilidad de pertenencia a un grupo.

5.1. Variable Clasificadora Capital de Trabajo

Con los valores de la media muestrales de la variable capital de trabajo se calcula el punto de corte discriminante:

$$C_1 = 7$$

Si $X_i < C = 7$ se clasifica al individuo i en el grupo I

Si $X_i > C = 7$ se clasifica al individuo i en el grupo II

Con esta regla de decisión se obtiene la matriz de confusión que se presenta en la **Tabla 2**.

Tabla 2: Matriz de confusión de la Variable Capital de Trabajo

	CLASIFICADOS	COMO	
SITUACION REAL	NO CUMPLIDORES	CUMPLIDORES	TOTAL
NO CUMPLIDORES	6	2	8
CUMPLIDORES	2	6	8
NO CUMPLIDORES	75%	25%	100%
CUMPLIDORES	5%	75%	100%

5.2. Variable Clasificadora Pasivo Corriente

Del mismo modo, se calcula el punto de corte discriminante de la variable Pasivo Corriente

$C_2 = 4$

Si $X_i < C = 4$ se clasifica al individuo i en el grupo II

Si $X_i > C = 4$ se clasifica al individuo i en el grupo I

Con esta regla de decisión se obtiene la matriz de confusión que se presenta en la Tabla **3**:

Tabla 3: Matriz de confusión de la Variable Pasivo Corriente

	CLASIFICADOS	COMO	
SITUACION REAL	NO CUMPLIDORES	CUMPLIDORES	TOTAL
NO CUMPLIDORES	5	3	8
CUMPLIDORES	4	4	8
NO CUMPLIDORES	62.5%	37.5%	100%
CUMPLIDORES	50%	50%	100%

5.3. Función discriminante de Fisher

Con el uso del sofware R Studio se calculan los coeficientes de la función discriminante de Fisher y se obtiene

$$\overline{D_{=}} 0.422\ X_1(capital\ trabajo) - 0.38\ .\ X_{2.}(pasivo\ corriente) \qquad (12)$$

El punto de corte discriminante obtenido con la función discriminante de Fisher resulta

$$C = (0.21134 \text{ um} + 2.066176 \text{ um}) / 2 = 1.4366 \text{ um} \qquad (13)$$

La regla de decisión para el i-ésimo individuo estará dada en este caso por :

Si $D_i < C$ o $D_i - C < 0$ ➔ se clasifica al individuo i en el grupo I

Si $Di < C$ o $D_i - C > 0$ ➔ se clasifica al individuo i en el grupo II

donde D_i es la puntuación discriminante para el individuo i i = 1,…..,16

Con esta regla de decisión se obtiene la matriz de confusión que se presenta en la Tabla **4**.

Tabla 4: Matriz de confusión Función Discriminante de Fisher

	GRUPO POR FDF PREDICHO		
GRUPO VERDADERO	1	2	TOTAL
1	8	0	8
	100%	0%	50%
2	1	7	8
	12.5%	87.5%	50%
TOTAL	9	7	16

Se observa un mayor porcentaje de clasificaciones correctas que el que se obtuvo utilizando las variables capital de trabajo y pasivo corriente por separado.

Si se calculan las puntuaciones discriminantes para cada uno de los nuevos interesados, se obtienen los siguientes resultados:

Locatario 1: D_1 = 1,67 um > 1,4366 um

Locatario 2: D_1 = 3,2574 um > 1,4366 um

De acuerdo con este criterio, y con fines predictivos, los nuevos interesados en celebrar el contrato de leasing se consideran cumplidores en el futuro.

5.4. TEOREMA DE BAYES

Si las probabilidades a priori de pertenecer a alguno de los dos grupos son desconocidas y se consideran ambas iguales al 0.5, se puede verificar que se obtienen las mismas conclusiones que con la función discriminante de Fisher.

Si las probabilidades a priori son conocidas, ya sea porque se les asigna un valor histórico o estimado, se puede comprobar que los resultados de la clasificación mejoran y disminuye el número de clasificaciones erróneas.

En este caso el punto de corte estará dado por la expresión (14)

$$C_p = \frac{\overline{D_I}+\overline{D_{II}}}{2} - \ln\frac{\pi_I}{\pi_{II}} \qquad (14)$$

donde π_I y π_{II} son las probabilidades a priori de pertenecer a los grupos I y II respectivamente. Puede verificarse, por ejemplo, que si se asumiera que las probabilidades a priori de pertenecer a los grupos I y II fueran respectivamente iguales a 0,10 y 0,90, el individuo número 13 clasificado erróneamente como no cumplidor por los otros criterios, resultaría correctamente clasificado como cumplidor, ya que las probabilidades a posteriori de ser cumplidor y no cumplidor serían respectivamente iguales a 0,58 y 0, 42para ese individuo. También con el sofware R Studio se pueden obtener las matrices de confusión y las probabilidades resultantes de la aplicación del Teorema de Bayes.

CONCLUSIÓN

El análisis discriminante es una técnica multivariante sumamente útil para clasificación en grupos. Entre otras aplicaciones, resulta una herramienta poderosa para el análisis de riesgo crediticio. La detección de la capacidad de pago de tomadores de crédito mediante una herramienta eficiente conduce a la toma de decisiones apropiadas al clasificar a los individuos en grupos de confiabilidad, y facilita los procesos de selección a la hora de gestionar la oferta crediticia.

La asignación a un grupo determinado puede realizase utilizando cada variable clasificadora por separado (variables métricas) o mediante una función que resulta de una combinación lineal de todas las variables clasificados elegidas. Dicha función recibe el nombre de función discriminante de Fisher. Se observa que el porcentaje de casos clasificados correctamente mejora utilizando las variables combinadas respecto de la utilización por separado de las variables predictoras, por lo que resulta más conveniente usar la función discriminante de Fisher para lograr un mayor número de asignaciones correctas que permitan predecir casos futuros con mayor grado de confiabilidad para la toma de decisiones.

También se puede predecir la pertenencia a un grupo mediante el cálculo de probabilidades a posteriori con la utilización del Teorema de Bayes. El individuo será asignado a aquel grupo cuya probabilidad de pertenencia a posteriori calculada con la fórmula de Bayes sea mayor. Las probabilidades a priori pueden ser conocidas o desconocidas. En el caso de que sean desconocidas, se asume equiprobabilidad y se obtienen las mismas conclusiones que con la Función Discriminante de Fisher. El análisis se puede completar considerando los costos de la clasificación errónea.

La técnica de análisis discriminante puede extenderse a más de dos grupos con más de dos variables clasificadoras, mediante la generalización de los resultados expuestos en este trabajo.

REFERENCIAS BIBLIOGRÁFICAS

Anderson, T. (2003). *An Introduction to Multivariate Statistical Analysis*. New Jersey. Willey Series in Probability and Statistics.

Bellman, R.E. (1957). *Dynamic Programming*. Princeton University Press, Princeton, NJ. Republished (2003): DOVER, ISBN 0486428095.

Bock. H.H. (1985). *On Some Significance Test in Cluster Analysis. Journal of Classification*, 2(1): 77-108

Cuadras, C. M. (2019). *Nuevos Métodos de Análisis Multivariante*. Barcelona. CMC Editions.

Hair, J. F., Anderson, R. E., Tatham, R. L. y Black, W. (1998). *Multivariate Data Analysis*. N.J. Prentice Hall College Division.

Hair, J.F., Black, W.C., Babin, B.J. y Anderson, R.E. (2014a). *Multivariate Data Analysis*. Pearson, Harlow, U.K., 7ma edición

Hartigan, J. (1985). *Statistical Theory in Clustering. Journal of Classification*, 2(1): 63-76

Johnson A., Wichern D.W. (2007). *Applied Multivariate Statistical Analysis*. N.J.Pearson.

Perez, C. (2008) . *Técnicas de Análisis Multivariante de Datos*. Pearson Prentice Hall. Madrid.