UNIVERSIDAD DE GRANADA

Facultad de Ciencias

Influencia de diversos indicadores económicos en el beneficio de una empresa

José Juan García Rodríguez, María José Mora Díaz, Daniel Vallejo España

Estadística Multivariante, Curso 2022-2023

15 de diciembre de 2022

1. INTRODUCCIÓN

Hoy en día, existen muchos factores que influyen en la economía de una empresa, desde el volumen de facturación, el nivel de nueva contratación y el total de clientes hasta los niveles de organización empresarial y la relación con otras empresas. Un problema bastante común es el de, a partir de algunos de estos indicadores económicos, predecir el beneficio de la empresa, con el objetivo de tomar decisiones sobre la empresa y adoptar las medidas necesarias.

Las herramientas de aprendizaje supervisado (*Machine learning*), como el Análisis Discriminante (Lineal y Cuadrático), han demostrado su efectividad para establecer métodos de clasificación en una variable categórica usando una serie de variables predictoras.

El objetivo de este trabajo es el estudio de un conjunto de datos sobre varios indicadores económicos recogidos en varias empresas. Se pretende hacer una reducción de la dimensión, obteniendo los indicadores que mejor explican los datos y las variables no observables de mayor importancia; y obtener un modelo de clasificación del nivel de beneficios de una empresa atendiendo a estos indicadores.

2. MATERIALES Y MÉTODOS

2.1. MATERIALES

Para realizar este estudio se ha tomado una base de datos. Esta contiene información sobre **8 indicadores económicos** medidos en **14 empresas**. Estos indicadores son los siguientes:

- Indicador del volumen de facturación de la empresa (x_1).
- Indicador del nivel de nueva contratación (x_2).
- Indicador del total de clientes (x_3).
- Indicador de beneficios de la empresa (x_4).
- Indicador del nivel de retribución salarial de los empleados (x_5).
- Indicador del nivel de organización empresarial dentro de la empresa (x_6).
- Indicador del nivel de relaciones con otras empresas (x_7).
- Indicador del nivel de equipamiento (ordenadores, maquinaria, etc.) (x_8).

A continuación se da una tabla que incluye las principales medidas de posición, dispersión y forma más importantes en estos datos:

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Mínimo	0.13	0.94	2.17	0.03	2.56	6.13	1.06	3.95
Primer Cuartil	3.360	9.3298	9.208	4.2718	3.678	10.042	4.009	6.958
Mediana	7.80	16.42	15.58	5.60	5.41	11.24	6.09	10.59
Tercer Cuartil	11.321	21.1861	21.125	7.1857	5.908	31.449	10.070	19.584
Máximo	20023.00	25.43	28.17	11.40	10.04	43.28	12.37	26.57
Rango	20022.87	24.48	26.00	11.37	7.48	37.14	11.31	22.62
Media	1436.67	15.13	15.23	5.58	5.31	20.22	6.60	13.07
Varianza	28617304.20	62.18	65.77	7.84	4.26	183.93	15.00	62.82
Desviación típica	5349.51	7.89	8.11	2.80	2.06	13.56	3.87	7.93
Coeficiente de variación	3.72	0.52	0.53	0.50	0.39	0.67	0.59	0.61
Coeficiente de asimetría	2.98	-0.27	-0.02	-0.06	0.62	0.49	0.09	0.45
Coeficiente de curtosis	7.41	-1.34	-1.34	-0.17	-0.31	-1.56	-1.52	-1.41

2.2. MÉTODOS ESTADÍSTICOS

En primer lugar, se ha realizado un **análisis exploratorio** previo de los datos para identificar posibles **valores perdidos** y **valores extremos** o **"outliers"** y se han tomado las decisiones correspondientes para tratarlos.

- En las variables con **más de un** 5 % de valores perdidos se ha analizado el patrón aleatorio de los mismos estudiando la **homogeneidad** según grupos con otras variables sin datos perdidos. Como las variables son cuantitativas, se ha utilizado un *test de Student*.
- En las variables con **valores extremos**, estos se han identificado con gráficos *boxplot* y, como la base de datos no tiene suficientes registros, se ha tomado la decisión de sustituirlos por la media, pues las variables son cuantitativas.

En segundo lugar, se ha realizado un **análisis descriptivo numérico clásico**, dando las principales medidas de posición, dispersión y forma para poder entender mejor los datos.

En tercer lugar, se ha procedido a aplicar diferentes técnicas estadísticas de Análisis Multivariante, comprobando los supuestos necesarios para cada una de ellas:

- Comprobación de la **correlación** entre los datos: A nivel poblacional se ha justificado utilizando *el contraste de esfericidad de Bartlett*, que permite comprobar si las correlaciones son distintas de 0 de modo significativo; y a nivel muestral se ha justificado observando la matriz de correlaciones, la matriz de correlaciones policrítica y otras representaciones gráficas.
- Comprobación de la **normalidad univariante**: Se ha hecho una exploración gráfica de la normalidad de las distribuciones individuales mediante histogramas y gráficos *qqplots*. Esto podría dar una idea de la posible distribución normal de las variables unidimensionales, pero la conclusión definitiva se ha obtenido mediante un test de hipótesis, que en este caso ha sido el de *Shapiro-Wilks*.
- Comprobación de la **normalidad multivariante**: Se ha estudiado mediante tests de hipótesis como el de *Royston* y el de *Henze-Zirkler*. Hay que tener en cuenta que la normalidad multivariante puede verse afectada por la presencia de outliers multivariantes, por lo que se ha hecho un análisis de los mismos.

Una vez comprobados los supuestos necesarios, se han aplicado las diversas técnicas del Análisis Multivariante. En este caso, se han realizado un **Análisis de Componentes Principales** para reducir la dimensión por medio de variables observables; un **Análisis Factorial** para identificar *variables latentes* (no observables) que tengan una alta correlación con un grupo de variables observables y correlación prácticamente nula con el resto; y un **Análisis Discriminante** (tanto **Lineal** como **Cuadrático**) para establecer un método de clasificación de nuevas observaciones de una variable cualitativa según sus características (variables explicativas o predictores).

3. RESULTADOS

En primer lugar, se ha observado que solo **dos variables** presentan **valores perdidos**. Como ambas tienen más del 5 % de valores perdidos, se ha analizado el patrón aleatorio de los mismos, y **no** se han obtenido **evidencias** para **rechazar la hipótesis nula de homogeneidad**, por lo que se ha concluido que el **patrón** es **aleatorio**. Teniendo en cuenta esto, se ha decidido sustituir los valores perdidos por la media de los valores no perdidos.

En segundo lugar, se ha observado la **presencia** de algunos **outliers**. Para tratarlos se ha decidido sustituirlos por la media de los demás valores.

En tercer lugar, el *test de Bartlett* ha dado evidencia para **rechazar** la hipótesis de **independencia** de los datos. De este modo, se ha concluido que **existe correlación** entre los datos y, por tanto, se puede plantear una **reducción de la dimensión** mediante un **Análisis de Componentes Principales** o un **Análisis Factorial**. Además, en las representaciones gráficas de la correlación se ha observado que hay entre 2 y 3 grupos de variables latentes con alta correlación con un grupo de las variables observables y poca correlación con el resto.

En cuarto lugar, mediante los tests correspondientes se ha concluido que **hay normalidad univariante** para la mayoría de las variables condicionadas a cada modalidad de la variable cualitativa estudiada en el **Análisis Discriminante**. A pesar de que para dos de ellas se ha rechazado la hipótesis nula de normalidad, se continúa con el análisis. Además, se ha concluido la **presencia de normalidad multivariante**.

En quinto lugar, gracias el **Análisis de Componentes Principales**, de las 8 variables de partida se ha pasado a un conjunto de **3 variables observables**, que son combinaciones lineales de las originales y acumulan casi un 90 % de la varianza explicada. Las **componentes principales** obtenidas son las siguientes:

- $0.3807652 \cdot x_1 + 0.3944434 \cdot x_2 + 0.3966045 \cdot x_3 + 0.1677333 \cdot x_4 - 0.2179417 \cdot x_5 + 0.3922894 \cdot x_6 + 0.3850908 \cdot x_7 + 0.4053617 \cdot x_8$
- $0.43496807 \cdot x_1 + 0.40873039 \cdot x_2 + 0.41399128 \cdot x_3 - 0.11769717 \cdot x_4 + 0.05007125 \cdot x_5 - 0.39514018 \cdot x_6 - 0.40485150 \cdot x_7 - 0.36871834 \cdot x_8$
- $-0.04804318 \cdot x_1 - 0.08297130 \cdot x_2 - 0.04198928 \cdot x_3 + 0.70489967 \cdot x_4 - 0.643082395 \cdot x_5 - 0.14505034 \cdot x_6 - 0.15255126 \cdot x_7 - 0.18518781 \cdot x_8$

En el siguiente gráfico se muestra la varianza explicada por cada una de las componentes principales. Así, se aprecia que tres de ellas acuulan casi el 90 % de la varianza de los datos.

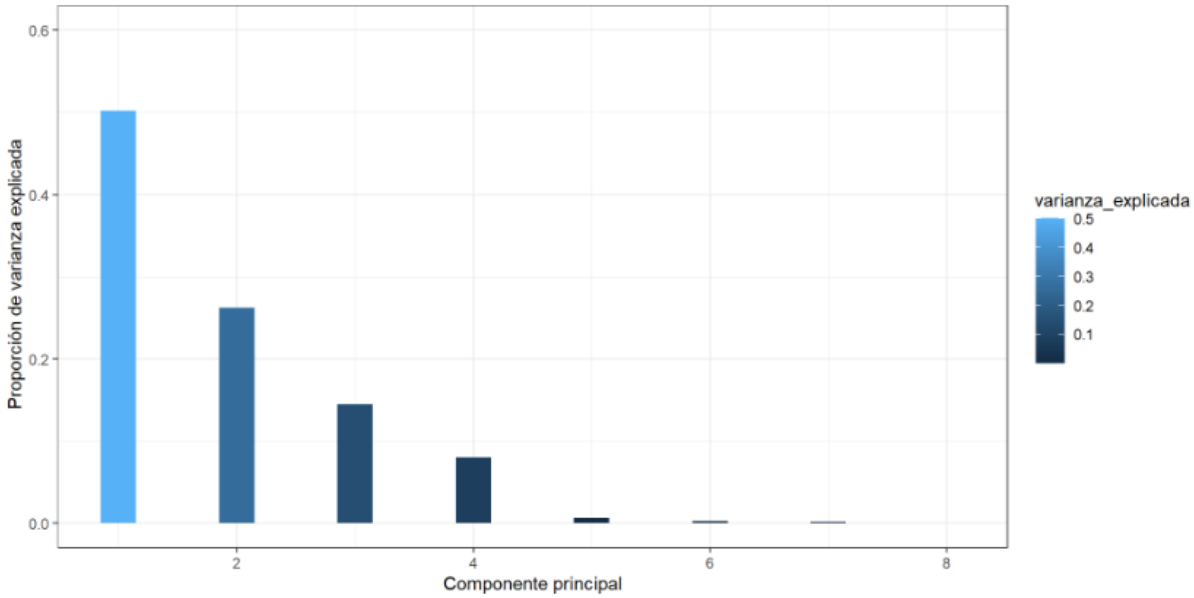


Figura 1:
Varianza explicada por las Componentes Principales.

En sexto lugar, en el **Análisis Factorial** se ha concluido la existencia de **tres factores latentes** que son suficientes para explicar los datos. Estos factores correlacionan mucho con un grupo de los datos y poco con el resto. En este caso, se ha observado que el primer factor latente correlaciona con las variables x_1, x_2 y x_3; el segundo factor latente correlaciona con las variables x_6, x_7 y x_8; y el tercer factor latente correlaciona con las variables x_4 y x_5.

Finalmente, se ha definido una variable categórica y se ha realizado un **Análisis Discriminante**, tanto **Lineal** (ADL) como **Cuadrático** (ADC). Se ha definido la variable respuesta a partir de x_4 y se han usado las demás variables como predictores.

El modelo de clasificación discriminante lineal obtenido viene dado por la siguiente expresión:

$$\begin{aligned} Odds\_ratio =& 0.8719170 \cdot x_1 + 1.0148619 \cdot x_2 - 1.3460782 \cdot x_3 \\ &+ 0.7918076 \cdot x_5 + 0.4741444 \cdot x_6 + 0.1987455 \cdot x_7 - 0.9608946 \cdot x_8 \end{aligned}$$

Por ejemplo, se ha introducido una nueva empresa con los siguientes indicadores:

$$x_1 = 1,\ x2 = 0,\ x3 = 9,\ x5 = 2,\ x6 = 8,\ x7 = 1,\ x8 = 14$$

y se ha obtenido que una empresa con tales valores para los indicadores tendrá un beneficio alto.

En el caso del ADL, se ha obtenido un método de clasificación con un porcentaje de error del $7'14\,\%$; y en el caso del ADC, se ha obtenido un método de clasificación con un porcentaje de error del $0\,\%$. Cabe destacar que para este último se ha quitado el predictor x_7 ya que se tienen pocos datos y muchos predictores.

En las siguientes gráficas se observa la exploración gráfica de los datos mediante una nube de puntos (observando x_5 y x_6 para ver si separan bien las modalidades de la variable categórica), un modelo de discriminante lineal con dos regresores (x_5 y x_6) y un modelo de discriminante cuadrático también con dos regresores (x_5 y x_6):

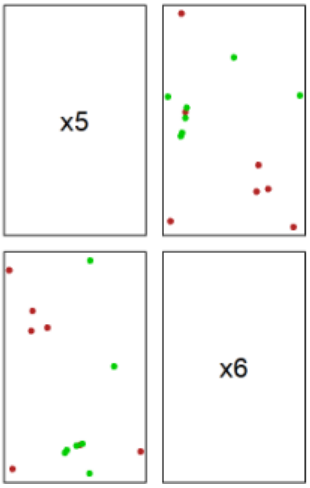
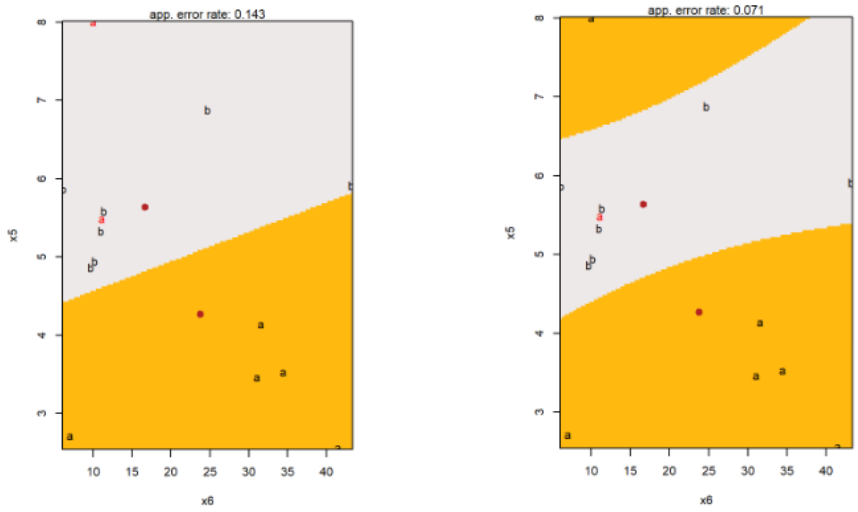


Figura 2:
Exploración gráfica de los datos (x_5 y x_6).



(a)
ADL (x_5 y x_6).

(b)
ADC (x_5 y x_6).

Figura 3:
Comparación ADL y ADC.

En estas representaciones gráficas se observa que los modelos de clasificación lineal y cuadrático con dos regresores también tienen un porcentaje de error bajo, por lo que podría interesar utilizar menos predictores (menor costo en recogida y computación de los datos), asumiendo un error un poco mayor.

4. DISCUSIÓN

Los datos objeto de estudio eran referentes a 8 indicadores económicos medidos en 14 empresas distintas. El objetivo del estudio era hacer una reducción de la dimensión y establecer un método de clasificación del nivel de beneficios de la empresa atendiendo a estos indicadores económicos.

A la vista de las componentes principales obtenidas, se puede observar que los indicadores que aportan más información sobre la economía de la empresa son el volumen de facturación de la empresa, el

nivel de nueva contratación, el total de clientes, la organización, las relaciones con otras empresas y el equipamiento.

El Análisis Factorial ha mostrado la existencia de tres factores latentes que son suficientes para explicar los datos:

- El primero correlaciona mucho con las variables que indican el volumen de facturación de la empresa, el nivel de nueva contratación y el total de clientes, de forma que se puede interpretar como el **tamaño de la empresa**.
- El segundo correlaciona mucho con los niveles de organización empresarial dentro de la empresa, relaciones con otras empresas y equipamiento, de forma que se puede interpretar como **política de organización interna y externa** de la empresa.
- El tercero correlaciona mucho con los indicadores de beneficios y nivel de retribución salarial a los empleados, de forma que se puede interpretar como **ganancias y gastos** de la empresa.

Finalmente, se han querido clasificar los datos en función del nivel de beneficio de la empresa, teniendo en cuenta los indicadores económicos como variables predictoras y con el objetivo de, dados los datos de una nueva empresa, poder clasificarla según sean sus beneficios.

Para las empresas de la base de datos, se ha dicho que tienen beneficio "alto" si el indicador de beneficio x_4 es superior a la media y que tienen beneficio "bajo" en caso contrario. Entonces, mediante una Análisis Discriminante (tanto Lineal como Cuadrático) se han obtenido métodos de clasificación para nuevas observaciones.

5. CONCLUSIÓN

En este trabajo se han estudiado una serie de datos sobre indicadores económicos en varias empresas. Se han observado los indicadores que tienen más peso para explicar los datos y tres características importantes que determinan la economía de la empresa, que son su **tamaño**, su **política de organización interna y externa** y las **ganancias** de la misma. Además, se ha obtenido un método de clasificación del nivel de beneficio según estos indicadores, lo que puede permitir a las empresas mejorar este nivel de beneficio teniendo en cuenta la influencia de cada uno de estos indicadores.

Contribuciones de los autores: Conceptualización e investigación: J.J.G.R., M.J.M.D. y D.V.E. ; Pre-procesamiento de los datos: J.J.G.R. ; ACP y AF: M.J.M.D. ; ADL y ADC: D.V.E. ; Redacción del informe: J.J.G.R. ; Repaso y edición del informe: M.J.M.D. y D.V.E.