REVISTA INCAING
ISSN 2448 9131

Aplicación del análisis multivariado de datos en un sistema basado en la web para la medición del índice de inclusión social en una institución de educación superior

Joseph Leandro Trejos Hilarión[1], Ing. César Yesid Barahona Rodríguez[2]
[1]Estudiante de ingeniería de sistemas Universidad de Cundinamarca, jltrejos@ucundinamarca.edu.co
[2]Maestría en sistemas computacionales de la Universidad Umecit Panamá, Especialista en gestión de proyectos Unad Colombia, Ingeniero en Telecomunicaciones UMNG Colombia, Docente Investigador grupo GISTFA Universidad de Cundinamarca, cbarahona@ucundinamarca.edu.co

Resumen – A través del análisis de los factores involucrados en el aspecto de la educación de calidad por parte del ministerio de educación nacional (MEN), partiendo de las políticas establecidas por la UNESCO sobre la educación inclusiva, y buscando un aporte encaminado a los objetivos de desarrollo sostenible más específico a los objetivos 4 y 6, se desarrolló un algoritmo con la capacidad de aplicar el análisis multivariado de datos por medio del análisis de correspondencias múltiples y análisis de componentes principales, el cual aplicado a una herramienta de desarrollo web obtiene la capacidad de medir el índice de inclusión social que posee la institución de educación superior mediante la aplicación de una encuesta.

Índice de Términos – Análisis multivariado, análisis de correspondencias múltiples, análisis de componentes principales, índice de inclusión social.

I. INTRODUCCION

El análisis multivariante se puede conceptualizar como una serie o un conjunto de métodos matemáticos y estadísticos, los cuales tienen como objetivo interpretar y describir los datos provenientes de la observación de distintas variables estadísticas ya sean cuantitativas o cualitativas (Cuadras, 2007).

Dentro de este tipo de análisis se pueden observar los modelos utilizados para el desarrollo del documento, los cuales son el análisis de correspondencias múltiples y el análisis de componentes principales.

Análisis de correspondencias

Previo a la conceptualización del análisis de correspondencias múltiple, se debe hablar del análisis de correspondencias, o también conocido como análisis de correspondencias simple.

Contando como otro de los métodos factoriales para el análisis exploratorio de datos multidimensionales junto al análisis de correspondencias múltiples y el análisis de componentes principales, el análisis de correspondencias simple es una técnica estadística útil para quienes trabajan con datos categóricos, como ejemplo y en este caso, los datos obtenidos en encuestas sociales. Este método presenta como punto fuerte su eficacia para analizar tablas de contingencia con datos de frecuencias numéricas, ya que proporciona una representación gráfica elegante y simple que permite una rápida interpretación y comprensión de los datos (Michael, 2002).

Análisis de correspondencias múltiples

En esencia, esta técnica es similar al análisis de correspondencias simples, su diferencia se da en que esta busca describir, en un espacio de pocas dimensiones o factores, la estructura de asociaciones entre un grupo de hasta más de dos variables categóricas, así como sus similitudes y diferencias entre los individuos a los cuales estas variables aplican (Ledesma, 2008).

Ya que se esta técnica se puede adaptar para estudiar más de dos variables categóricas, a continuación, se muestra primero el procedimiento para dos y luego se podrá generalizar.

Primero se escribe la matriz $n \times (I + J)$ de datos binarios como una matriz $n \times (J1 + J2)$ generando de esta forma (para efectos prácticos se cambian los nombres de los datos).

$$Z = [Z_1, Z_2].$$

Entonces se tiene

$$B_U = Z'Z = \begin{bmatrix} Z'_1 Z_1 & Z'_1 Z_2 \\ Z'_2 Z_1 & Z'_2 Z_2 \end{bmatrix} = n \begin{bmatrix} D_r & P \\ P' & D_c \end{bmatrix}$$

Luego, la matriz de frecuencias, donde F y C contienen las marginales de filas y columnas, $B_U = \begin{bmatrix} F & N \\ N' & C \end{bmatrix}$ se identifica como la matriz de Burt. A lo cual ya se pueden realizar el análisis de correspondencias diferentes sobre las matrices:

- N
- Z_1, Z_2
- B_U

Se considera por siguiente Q variables categóricas con J1, ... ,JQ estados, respectivamente, sobre n individuos. Si J = J1 + ... + JQ. La tabla de datos, de orden n × J es la super-matriz de indicadores $Z = [Z1, \ldots, Zj, \ldots, Zq]$, donde Zj es n × Jj y contiene los datos binarios de la variable j. La tabla de contingencia que tabula la combinación de las variables i, j es $N_{ij} = Z'_i Z_j$. La matriz de Burt de orden J × J es:

$$B_U = Z'Z = \begin{bmatrix} Z'_1 Z_1 & \ldots & Z'_1 Z_q \\ \ldots & \ldots & \ldots \\ Z'_q Z_1 & \ldots & Z'_q Z_q \end{bmatrix}$$

En donde las matrices Z'_j, Z_j son diagonales (Cuadras, 2007).

Análisis de componentes principales

El Análisis de Componentes Principales al igual que los anteriormente tratados, es una técnica de cálculo multivariante que tiene como objetivo reducir el tamaño de una matriz de datos ya sea de covarianza o correlacional demasiado grande debido a la gran cantidad de variables x1, x2, ..., xn que contiene, y conserva algunas de las variables C1, C2, ..., Cp combina las siglas (componentes principales) con gran poder de cómputo y agrega mucha información comprimida en sus datos. En primer lugar, hay tantos elementos como variables:

$$\begin{aligned} C_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n) \\ &\vdots \\ C_n &= a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n) \end{aligned}$$

Fig 1 Componentes a reducir a través del análisis de componentes principales, fuente (César, 2004)

Pero solo se conservan las componentes p (componentes principales), lo que explica la alta tasa de variación de las variables originales (C1, C2, ..., Cp).
Deberá considerar el tipo de variable que se está manejando. En el análisis de componentes principales, las variables deben ser cuantitativas. Los componentes deben ser suficientes para resumir la mayor parte de la información contenida en las variables originales. Asimismo, cada variable de entrada se puede representar como una función de los componentes principales, de modo que la varianza de cada variable de entrada se explica completamente por los componentes para los que la combinación lineal la define.

$$\begin{aligned} x_1 &= r_{11}C_1 + r_{12}x_2 + \cdots + r_{1p}C_p \\ &\vdots \\ x_n &= r_{n1}C_1 + r_{n2}C_2 + \cdots + r_{np}C_p \end{aligned} \qquad r_{ij} = \sqrt{\lambda_i}\, a_{ij}$$

Fig 2 Coeficiente de relación ACP ***Coeficiente de relación ACP***

Se puede demostrar que r_{ij} actúa como el coeficiente de correlación entre la componente C_i y la variable x_j y se obtiene el cálculo a través de la multiplicación del peso a_{ij} de la variable en esa componente y la raíz cuadrada de su valor propio λi (cada componente principal C_i se asocia con el valor propio i-ésimo (en magnitud) de la matriz (a_{ij}).

Aplicación de librería PRINCE

Una vez entendido el concepto de los tipos de análisis multivariado, se propuso la aplicación de la librería prince del lenguaje de programación Python. Esta librería utiliza en esencia las funciones de scikit learn realizando una implementación más eficiente y variando sus parámetros se describe como una librería de análisis factorial que posee entre sus funciones el análisis de correspondencias múltiples y el análisis de componentes principales (*Prince · PyPI*, n.d.).

A continuación, se muestra la aplicación de esta librería sobre un Dataframe de la librería pandas, en este ejemplo se utilizó un set de datos en donde cada fila representa a una persona y cada columna la respuesta elegida por el mismo sobre los primeros tres factores a evaluar en el índice de inclusión social.

En el primer paso, además de la librería prince, se importaron las librerías necesarias para contener los datos y graficarlos, de forma que sea más fácil el tratamiento de la información y la representación de la misma.

```
import prince
import pandas as pd
import matplotlib.pyplot as plt
```

Fig 3 Importación de librerías, elaboración propia

Luego se establecieron los Dataframes a través de la librería pandas para los tres factores.

```
pd.set_option('display.float_format', lambda x: '{:.6f}'.format(x))

Factor1 = pd.DataFrame(
    data=[
        ["Existe y se implementa"],
        ["No existe"],
        ["Existe y no se implementa"],
        ["Existe y se implementa"],
        ["No sabe"],
        ["Existe y se implementa"],
        ["Existe y no se implementa"],
        ["Existe y no se implementa"],
        ["No existe"],
        ["No existe"],
        ["Existe y se implementa"],
        ["No existe"],
        ["Existe y no se implementa"],
        ["Existe y se implementa"],
        ["Existe y se implementa"]
    ],
    columns=pd.Series(['A1']),
)
```

Fig 4 Dataframe factor 1, elaboración propia

```
Factor2 = pd.DataFrame(
    data=[
        ["Existe y no se implementa", "Nunca"],
        ["Existe y se implementa", "Algunas veces"],
        ["Existe y no se implementa", "Nunca"],
        ["Existe y se implementa", "Siempre"],
        ["No existe", "Algunas veces"],
        ["Existe y se implementa", "Algunas veces"],
        ["No existe", "Algunas veces"],
        ["Existe y no se implementa", "Nunca"],
        ["No sabe", "Nunca"],
        ["No existe", "Algunas veces"],
        ["No existe", "Nunca"],
        ["Existe y se implementa", "Algunas veces"],
        ["Existe y no se implementa", "No sabe"],
        ["No sabe", "Siempre"],
        ["Existe y se implementa", "No sabe"]
    ],
    columns=pd.Series(['B1', 'B2']),
    #index=pd.Series(['Blue', 'Light', 'Medium', 'Dark'])
)
```

Fig 5 Dataframe factor 2, elaboración propia

```
Factor3 = pd.DataFrame(
    data=[
        ["Existe y se implementa", "siempre"],
        ["Existe y no se implementa", "Algunas veces"],
        ["No existe", "Nunca"],
        ["No existe", "Siempre"],
        ["Existe y se implementa", "No sabe"],
        ["Existe y se implementa", "Algunas veces"],
        ["Existe y no se implementa", "Algunas veces"],
        ["Existe y no se implementa", "Nunca"],
        ["No existe", "Siempre"],
        ["No sabe", "Algunas veces"],
        ["Existe y se implementa", "No sabe"],
        ["Existe y no se implementa", "Algunas veces"],
        ["No sabe", "Nunca"],
        ["Existe y se implementa", "Siempre"],
        ["No existe", "Algunas veces"]
    ],
    columns=pd.Series(['C1', 'C2']),
    #index=pd.Series(['Blue', 'Light', 'Medium', 'Dark'])
)
```

Fig 6 Dataframe factor 3, elaboración propia

Después de la declaración de los dataframes, se realizó la aplicación del análisis de correspondencias múltiples para cada factor, se ajustó la matriz resultante a los datos

mediante el método fit, se estandarizaron a través de la función transform, y finalmente se unificaron estos resultados.

```
]: mca1 = prince.MCA(
        n_components=2,
        n_iter=3,
        copy=True,
        check_input=True,
        engine='auto',
        random_state=101
    )
mca1 = mca1.fit(Factor1)
tips_mca1 = mca1.transform(Factor1)
tips_mca1.head()
```

Fig 7 Aplicación del MCA de prince, elaboración propia

En la siguiente figura se muestran los datos resultantes de cada factor aplicando el análisis de correspondencias múltiples y unificando las matrices finales, de esta manera, las primeras dos columnas corresponden a los componentes del factor 1, las dos siguientes al factor 2, y las últimas dos al factor 3.

0	-0.444414	-0.968948	1.292708	-0.588348	1.581322	-0.543589
1	-0.253880	1.523781	-0.980589	-0.000000	-0.743941	-0.779754
2	0.000000	0.000000	1.292708	-0.588348	-0.631284	0.875587
3	-0.444414	-0.968948	-0.418577	1.470871	0.031302	1.760262
4	3.682004	-0.281438	-0.944310	-0.588348	1.581322	-0.543589
5	-0.444414	-0.968948	-0.980589	-0.000000	0.388717	-0.509130
6	0.000000	0.000000	-0.944310	-0.588348	-0.743941	-0.779754
7	0.000000	0.000000	1.292708	-0.588348	-0.939559	-0.388059
8	-0.253880	1.523781	1.015054	0.882523	0.031302	1.760262
9	-0.253880	1.523781	-0.944310	-0.588348	-0.808511	-0.585859
10	-0.444414	-0.968948	0.214180	-0.588348	1.581322	-0.543589
11	-0.253880	1.523781	-0.980589	-0.000000	-0.743941	-0.779754
12	0.000000	0.000000	0.891073	-0.588348	-1.004130	-0.194164
13	-0.444414	-0.968948	0.418577	2.353394	0.855685	0.767240
14	-0.444414	-0.968948	-0.223734	-0.000000	-0.435666	0.483892

Fig 8 Representación de matriz resultante, elaboración propia

Teniendo la matriz correlacional, se pueden mostrar los resultados a través de la librería matplotlib, en este caso se muestra el factor 1 después del procesamiento de ACM de la librería prince.

```
ax = mca1.plot_coordinates(
    X=Factor1,
    ax=None,
    figsize=(6, 6),
    show_row_points=True,
    row_points_size=10,
    show_row_labels=False,
    show_column_points=True,
    column_points_size=30,
    show_column_labels=True,
    legend_n_cols=1)
```

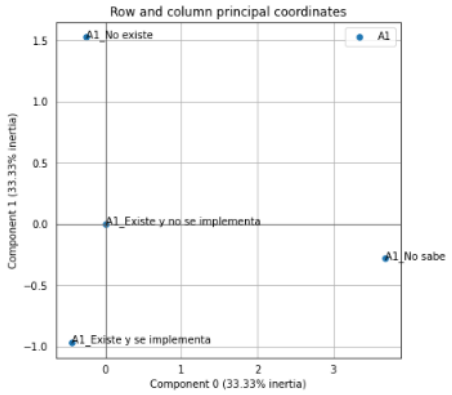

Fig 9 Representación MCA, elaboración propia

Por último, se aplicó la función PCA de prince sobre la matriz de análisis de correspondencias múltiples unificada de cada factor, graficando los resultados a través de un diagrama de cajas.

```
fig = plt.figure(figsize =(10, 7))
ax = fig.add_axes([0, 0, 1, 1])
bp = ax.boxplot(pca)
plt.show()
```

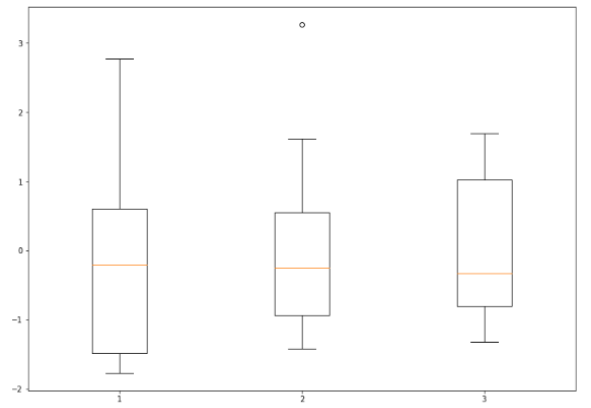

Fig 10 Diagrama de cajas PCA, elaboración propia

De esta forma, se observa cómo la aplicación del análisis multivariado puede representar el comportamiento de los datos a través de la reducción de la información de origen ingresada manteniendo los componentes más importantes que definen la tendencia de los datos.

Resultados

Una vez entendido el funcionamiento y aplicación del análisis multivariado sobre el lenguaje de programación Python y viendo las ventajas brindadas al aplicar estas funciones en cuanto a resultados, se puede implementar este algoritmo a más entornos de programación que utilicen Python como lenguaje, cabe recalcar que el lenguaje de programación no aplica como impedimento al momento de realizar el procesamiento y cálculo de datos con las técnicas vistas en este apartado.

Finalmente se implementó el algoritmo con la librería prince a un framework para Python, desplegando el código desarrollado en este framework se puede obtener el número y el tipo de respuestas seleccionadas por el usuario por medio de una encuesta basada en un documento suministrado por el ministerio de educación nacional conocido como Índice de inclusión para la educación superior en Colombia o "INES" el cuál está abierto para que las universidades puedan seguir lineamientos de una educación de calidad en el aspecto de inclusión social y de esta manera realizar un proceso de autoevaluación de forma independiente. Las preguntas en esta encuesta fueron clasificadas en distintos factores los cuales son:

FACTOR

1. Misión y Proyecto Institucional:

   1.1. Barreras para el aprendizaje y la participación
   1.2. Identificación y caracterización de estudiantes desde la educación inclusiva

2. Estudiantes:

   2.1. Participación de estudiantes
   2.2. Admisión, permanencia y sistemas de estímulos y créditos para estudiantes

3. Profesores

   3.1. Participación de docentes
   3.2. Docentes inclusivos

4. Procesos académicos:

   4.1. Interdisciplinariedad y flexibilidad curricular
   4.2. Evaluación flexible

5. Visibilidad nacional e internacional

   5.1. Inserción de la institución en contextos académicos nacionales e internacionales
   5.2. Relaciones externas de profesores y estudiantes

6. Investigación y creación artística y cultural

   6.1. Investigación, innovación y creación artística y cultural en educación inclusiva
   6.2. Articulación de la educación inclusiva con los procesos de investigación, innovación y creación artística y cultural

7. Pertinencia e impacto social

   7.1. Extensión, proyección social y contexto regional
   7.2. Seguimiento y apoyo a vinculación laboral

8. Procesos de autoevaluación y autorregulación

   8.1. Procesos de autoevaluación y autorregulación con enfoque de educación inclusiva
   8.2. Estrategias de mejoramiento
   8.3. Sistema de información inclusivo

9. Organización, administración y gestión

   9.1. Procesos administrativos y de gestión flexibles
   9.2. Estructura organizacional

10. Planta física y recursos de apoyo académico

    10.1. Recursos, equipos y espacios de práctica

    10.2. Instalaciones e infraestructura

11. Bienestar institucional

11.1. Programas de bienestar universitario
11.2. Permanencia estudiantil

12. Recursos financieros

12.1. Programas de educación inclusiva sostenibles
12.2. Apoyo financiero a estudiantes

Cada uno de los sub-factores descritos representa un aspecto a evaluar dentro del índice de inclusión educativa y está asignado a una pregunta en la encuesta, dicha pregunta, tiene a su vez indicadores que permitirán establecer las variables categóricas dentro de la recolección de los datos, estas variables son:

Indicadores de existencia: Existe y se implementa, existe y no se implementa, no existe, no sabe.

Indicadores de reconocimiento: Si, no, no sabe.

Indicadores de frecuencia: Siempre, algunas veces, nunca, no sabe.

La información recolectada de la encuesta aplicada en la Universidad de Cundinamarca en su extensión Facatativá permite entre otros rasgos, identificar la persona a la cuál se le asociará con una serie de respuestas, de esta manera se tiene armado el Dataframe visto en las figuras 4,5 y 6 pero en esta ocasión analizando todos los demás factores, de esta forma, aplicando el análisis multivariado se obtuvo la serie de datos normalizada que mejor representa las respuestas de los participantes para posteriormente graficarla mediante un diagrama de cajas o bigotes:

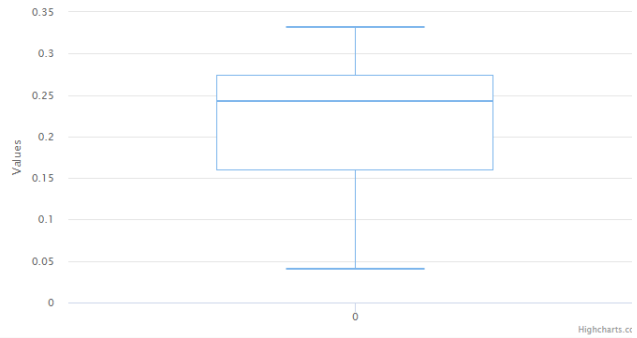

Figura 11 Índice de inclusión social educative, elaboración propia

Conclusión

El uso del análisis multivariado permite procesar datos cualitativos, de esta manera es posible aplicarlo no solo a campos inclusivos sino también en más aspectos sociales. Estas técnicas permiten reducir los datos a los elementos de más importancia y de los cuales se puede obtener una perspectiva clara sobre el comportamiento de algún fenómeno. Su sencilla implementación por medio del lenguaje Python, otorga la capacidad de tener una implementación sencilla en cualquier framework para desarrollo web que acepte este lenguaje de programación, de esta forma consiguiendo ventajas en la recolección, precisión e integridad de los datos a ser analizados posteriormente.

Bibliografía

César, P. (2004). *Técnicas de análisis multivariante de datos*.

Cuadras, C. M. (2007). Nuevos Metodos de Analisis Multivariante. *Revista Española de Quimioterapia : Publicación Oficial de La Sociedad Española de Quimioterapia*, *20*(3), 249. http://www.ncbi.nlm.nih.gov/pubmed/19406528

Ledesma, R. (2008). Software De Análisis De Correspondencias Múltiples : Una. *CONCIT*, *7600*, 59–75.

Michael, G. (2002). *La práctica del análisis de correspondencias*. Fundación BBVA.

prince · PyPI. (n.d.). Retrieved May 15, 2022, from https://pypi.org/project/prince/