

# ***PROYECTO***

## **1. Conjunto de Datos**

La matriz de datos se deja como una elección personal, inclusive se pueden utilizar diferentes bases para desarrollar cada una de las etapas del proyecto. No olvide que las variables seleccionadas deben responder a un objetivo en particular y a las características de la técnica multivariante que se vaya aplicar.

En lo que siguen se dan las directrices para aplicar las técnicas multivariantes vistas en clases utilizando el Software Estadístico de su preferencia, una vez desarrollado el trabajo resumir los resultados obtenidos en un informe. Si algún tipo de código fuese necesario, éste debe ser incluido en un apéndice, si el código fuese una o dos líneas incluirlo en el informe, ojo que no se aceptarán informes que sean solo corridas o salidas que proporcionan los softwares estadísticos. Finalmente, es bueno aprender a seleccionar aquello que realmente es importante, redactarlo, darle un formato personal, por otro lado no olvidar que la interpretación personal de los resultados, todos estos elementos son sumamente importantes en todo el desarrollo del trabajo.

## **2. PRIMERA FASE:** **Análisis de Componentes Principales**

### **2.1. Poblaciones Bivariantes**

En este apartado utilice una base de datos conformada por dos variables altamente correlacionadas

1. Construir los Estadígrafos Multivariantes e interpretar los mismos.
2. Realizar un análisis gráfico de los datos
3. Determinar si los datos provienen de una distribución normal bi-variante.

4. Construir la región elipsoidal con un 95 % de confianza, además determine el porcentaje de vectores de observación que se hallan dentro de la región elipsoidal.
5. Aplicar las pruebas de independencia y equicorrelación
6. Construir las componentes Principales
7. Estimar las componentes Principales y graficar las mismas.
8. Utilizando las estimaciones verificar si los supuestos del modelo de Componentes Principales se cumplen

## 2.2. Poblaciones Multivariantes

1. Un análisis descriptivo de los datos, se pueden utilizar elementos de la estadística univariante además de algunas representaciones gráficas propias del análisis multivariante.
2. Determinar si los datos provienen de una población normal multivariante
3. Aplicar la prueba de independencia
4. Construir las componentes asociadas al conjunto de datos
5. Calcular e interpretar  $\rho_{Y_i, X_j}$ , para  $i, j = 1, 2, \dots, p$ .
6. Seleccionar las componentes necesarias para representar los datos, para tal efecto aplicar los criterios:
  - Porcentaje de la Varianza
  - Promedio de los eigen valores
  - Scree Plot
  - Prueba de raíces retenidas (Suponga que los datos provienen de una población normal multivariante).

Emitir una conclusión sobre el número de componentes  $m$  que deben ser retenidas.

7. Para detectar observaciones sospechosas y evidenciar la normalidad de los datos, graficar las componentes estimadas de la muestra, es decir que se pueden construir los Q-Q plots (permiten identificar puntos atípicos y un comportamiento lineal indica normalidad en la componente), además de los diagramas de dispersión entre pares de componentes que permiten identificar estructuras elipsoidales. Aplicar estas ideas en el conjunto de datos.
8. Monitorear la calidad de los datos utilizando la relación:

$$\frac{\hat{y}_1^2}{\hat{\lambda}_1} + \frac{\hat{y}_2^2}{\hat{\lambda}_2} + \dots + \frac{\hat{y}_m^2}{\hat{\lambda}_m} \leq \chi_{m, \alpha}^2, \quad m : \text{nro. raíces retenidas}$$

9. Utilizar el gráfico  $T^2$  para determinar si no existe información significativa en las componentes descartadas.

$$T^2 = \frac{\hat{y}_{m+1}^2}{\hat{\lambda}_{m+1}} + \frac{\hat{y}_{m+2}^2}{\hat{\lambda}_{m+2}} + \dots + \frac{\hat{y}_p^2}{\hat{\lambda}_p}$$

### 3. TERCER ETAPA - Análisis Factorial

#### 3.1. Matriz de Correlaciones de Pearson

En este apartado se debe seleccionar de su base de datos un conjunto de **variables continuas** altamente correlacionadas.

1. Estimar la matriz de correlaciones.
2. Aplicar la prueba de independencia.
3. Estimar los parámetros del Modelo Factorial utilizando:
  - a) Componentes Principales
  - b) Factor Principal
  - c) Máxima Verosimilitud

4. Seleccionar el número de factores óptimo que cubra un porcentaje de la varianza total explicada. Utilice los criterios de componentes principales además de la prueba de Bondad de Ajuste.
5. Rotar los ejes para encontrar una estructura factorial más simple, para tal efecto realizar lo siguiente:
  - a) Utilizar el método cuartimax, además indique la matriz de rotación  $T$  utilizada.
  - b) Utilizar el método varimax, además indique la matriz de rotación  $T$  utilizada.

***Nota.** Debe aplicarse en los resultados de los tres métodos estimación de cargas factoriales utilizados.*

6. Estimar las varianzas comunes y las varianzas específicas, además de la proporción de variabilidad explicada por cada factor.
 

***Nota.** Debe aplicarse en los resultados de los tres métodos estimación de cargas factoriales.*
7. Estimar los puntajes factoriales utilizando cualquiera de los métodos estudiados.
8. Utilizando las estimaciones verificar que los supuestos del modelo factorial se cumplen.

### 3.2. Matriz de Correlaciones Policóricas o tetracóricas

En este apartado se debe seleccionar de su base de datos un conjunto de **variables cualitativas nominales u ordinales** altamente correlacionadas.

1. Estimar la matriz de correlaciones policórica o tetracórica.
2. Estimar los parámetros del Modelo Factorial utilizando Componentes Principales.
3. Seleccionar el número de factores óptimo que cubra un porcentaje de la varianza total explicada. Utilice los criterios de componentes principales.
4. Rotar los ejes para encontrar una estructura factorial más simple, para tal efecto realizar lo siguiente:
  - Utilizar el método cuartimax, además indique la matriz de rotación  $T$  utilizada

- Utilizar el método varimax, además indique la matriz de rotación  $T$  utilizada
5. Estimar:
    - Varianzas comunes y las varianzas específicas
    - La proporción de variabilidad explicada por cada factor.
  6. Estimar los puntajes factoriales utilizando el método de mínimos cuadrados.
  7. Utilizando las estimaciones verificar que los supuestos del modelo factorial se cumplen.

## **4. CUARTA ETAPA - Análisis de Conglomerados**

### **4.1. Análisis de Conglomerados con variables cuantitativas**

1. Seleccionar del conjunto de datos aquellas variables que son cuantitativas, para luego realizar un análisis exploratorio gráfico de los datos, analizando los resultados provistos por el: Diagrama de líneas, Diagrama de Estrellas y las caras de Chernoff.
2. Construir la matriz de distancias utilizando la distancia Euclidea
3. Aplicar los algoritmos jerárquicos aglomerativos referenciados en clases, comparar cual o cuales de ellos parecen razonables tomando como base los dendogramas resultantes.
4. Aplicar el algoritmo jerárquico divisivo visto en clase, según el árbol binario obtenido que conglomerados serían los seleccionados.
5. Aplicar los algoritmos de particionamiento de las k-medias tomando como valor de entrada  $k$ , como aquel número que parece más adecuado de acuerdo a lo visto en los algoritmos jerárquicos.

## **4.2. Análisis de Conglomerados con solo variables cualitativas**

1. Seleccionar del conjunto de datos aquellas variables que son cualitativas, en caso de que algunas no fuesen dicotómicas transformarlas de acuerdo al criterio dado en clases.
2. Construir la matriz de similaridades utilizando la similaridad de Jaccard, para posteriormente construir la matriz de distancias utilizando algún tipo de transformación
3. Aplicar los algoritmos jerárquicos aglomerativos referenciados en clases, comparar cual o cuales de ellos parecen razonables tomando como base los dendogramas resultantes.

## **4.3. Análisis de Conglomerados con variables de distinta naturaleza**

1. Seleccionar del conjunto de datos aquellas variables cuantitativas y cualitativas.
2. Construir la matriz de distancias utilizando la distancia de Gower
3. Aplicar los algoritmos jerárquicos aglomerativos referenciados en clases, comparar cual o cuales de ellos parecen razonables tomando como base los dendogramas resultantes.