



Universidad Nacional de Ingeniería  
Facultad de Ciencias  
Escuela Profesional de Ciencia de la Computación

## Práctica Calificada IV - II

### Minería de Datos

13/02/2026

CC442

Ciclo: 2025-III

Puntaje: 12 puntos

Duración: 60 minutos

## Caso: Análisis de Co-expresión Génica mediante Clustering Jerárquico Avanzado

### Contexto

En bioinformática, el análisis de microarrays y secuenciación de ARN permite medir la expresión de miles de genes simultáneamente. Un objetivo común es identificar grupos de genes o muestras que se comporten de manera similar (co-expresión), lo que puede indicar que participan en la misma ruta biológica o que pertenecen a un mismo tipo de tejido tumoral.

En este ejercicio, trabajarás con un dataset sintético que simula la expresión de **500 genes** en **100 muestras** biológicas. El reto consiste en encontrar la estructura oculta de los datos utilizando técnicas de clustering jerárquico, validando estadísticamente la calidad de los grupos formados.

### Objetivos

1. **Preprocesamiento de datos de alta dimensionalidad:** Escalar y normalizar datos donde el número de variables (genes) supera al número de observaciones (muestras).
2. **Selección de métricas de distancia:** Comparar y justificar el uso de la *Distancia de Correlación* frente a la *Distancia Euclídea* en contextos biológicos.
3. **Validación Jerárquica:** Utilizar el *Coefficiente de Correlación Cofenética* para evaluar qué método de enlace (Linkage) preserva mejor las distancias originales.
4. **Optimización de Hiperparámetros:** Determinar el número óptimo de clusters ( $K$ ) mediante el análisis de la *Silüeta*.
5. **Visualización Multidimensional:** Generar representaciones gráficas que permitan interpretar la relación entre muestras y genes simultáneamente.

### Tareas a realizar

#### 1. Preparación y Exploración

- Generar un dataset de 100 muestras y 500 genes con ruido gaussiano.
- Inyectar artificialmente tres grupos con comportamientos distintos para simular diferentes condiciones biológicas.
- Estandarizar los datos para asegurar que ningún gen domine el análisis debido a su escala de medida.

#### 2. Construcción de la Jerarquía

- Calcular la matriz de distancias utilizando la **métrica de correlación de Pearson**.
- Ejecutar el algoritmo de clustering jerárquico utilizando al menos cuatro métodos de enlace: `single`, `complete`, `average` y `ward`.

### 3. Validación Estadística

- Para cada método de enlace, calcula el **Coficiente Cofenético**. Explica cuál de los métodos de enlace genera un dendrograma que representa con mayor fidelidad las distancias originales de la matriz de entrada.
- Selecciona el mejor método para los pasos siguientes.

### 4. Determinación del Punto de Corte

- Realiza un barrido del número de clusters ( $k$ ) de 2 a 10.
- Calcula el **Silhouette Score** para cada  $k$ .
- Identifica el  $k$  óptimo y justifica tu elección basándote en la gráfica de silueta.

### 5. Interpretación Visual

- Dibuja el **dendrograma final** marcando con una línea horizontal el umbral de corte que corresponde al  $k$  óptimo.
- Genera un **Clustermap** (mapa de calor jerárquico) que muestre el clustering bidimensional: las muestras en las columnas y los genes en las filas.

### Entregables

- **Código en Python** limpio y comentado.
- **Gráficas:** Dendrograma, Gráfica de Silueta y Clustermap.
- **Breve conclusión:** ¿Cuántos grupos biológicos se identificaron y qué tan confiable es la separación según el coeficiente cofenético?