

### Tarea N°3 (20%)

Prof: Ana Aguilera Faraco  
Ayudante: Fernanda Fuentes  
[Fernanda.fuentes@estudiantes.uv.cl](mailto:Fernanda.fuentes@estudiantes.uv.cl)  
Octubre 2025

#### **Instrucciones:**

- La tarea N°3 es grupal (máximo 3 integrantes), y en caso de copia se aplicarán las sanciones correspondientes.
- El nombre del archivo debe ser “T3-NombreApellido.ipynb”, debe contener el nombre y apellido de todos los integrantes separado por un guión alto.
- [Portal de Datos Abiertos](#)
- Puntaje total: 100 puntos. Nota 4,0: 60 puntos.

#### **Introducción:**

En el análisis de bases de datos, las técnicas de agrupamiento o clustering permiten asignar, segmentar o agrupar registros en función de sus características, facilitando la descripción del conjunto de datos y la predicción en el caso de nuevos datos. En esta tarea, utilizaremos un dataset de Proceso de hospitalización para explorar cómo los algoritmos de clustering pueden ayudarnos a entender mucho mejor.

#### **Descripción del Dataset:**

El conjunto de datos utilizado corresponde a los establecimientos hospitalarios públicos de Chile, recopilado a partir de los registros oficiales del proceso de hospitalización en distintos hospitales del país. Cada registro representa la actividad mensual de un establecimiento, reflejando el comportamiento operativo y asistencial del sistema hospitalario.

Durante el período de análisis, se registran indicadores relacionados con la gestión y uso de los recursos hospitalarios, tales como el número de camas disponibles y ocupadas, la cantidad de egresos hospitalarios, los traslados de pacientes, los días totales de estadía, así como los egresos de pacientes fallecidos. Estos datos permiten analizar el funcionamiento interno de cada hospital y evaluar su desempeño dentro del sistema de salud nacional.

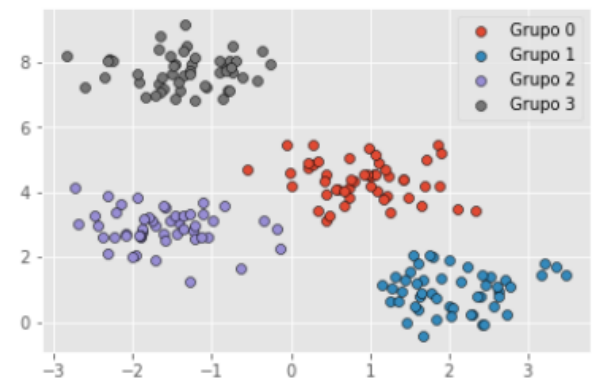
Este conjunto de datos resulta fundamental para el análisis del desempeño y eficiencia hospitalaria, ya que posibilita la aplicación de técnicas de clustering orientadas a agrupar los establecimientos según sus características operativas, niveles de ocupación y patrones de atención, con el propósito de identificar similitudes, diferencias y posibles tipologías de funcionamiento entre los hospitales del país.

#### **Objetivo de la tarea:**

Resolver un problema de agrupamiento (Clustering) es decir, dado un conjunto de datos con N elementos ser capaces de clasificarlos de manera que los datos pertenecientes a un grupo (clúster) sean similares entre sí y los distintos grupos (clústers) distantes como sea posible, una representación gráfica de esto se muestra a continuación como un ejemplo de

agrupamiento de datos.

**Debe considerar lo siguiente:**



1. Conjunto de datos:
  - Debe utilizar el dataset de la tarea de clasificación. Seleccione un conjunto de variables o características (columnas) a utilizar en su tarea y defina el objetivo de su investigación.
2. Pre-procesamiento de datos:
  - Limpiar el dataset según el tipo de dato a trabajar.
  - Grafique los datos para analizar su dispersión con un gráfico de variables aleatorias y dos más de su elección (concéntrico, luna, nubes de puntos, etc).
  - Normalizar los datos.
  - Mostrar el conjunto final de datos.
3. ¿Qué diferencia hay entre el conjunto inicial y el final? ¿Qué tan importante es la normalización en este proceso?
4. Análisis de clústers:
  - Debe realizar 3 análisis de clustering, 1 para cada grupo, es decir: 1 algoritmo para *Partitioning Clustering*, 1 para *Hierarchical Clustering* y 1 para *Density-Based Clustering*. Use este enlace como referencia <https://cienciadedatos.net/documentos/py20-clustering-con-python>.
  - Medidas de distancia, por ejemplo, Euclídea, Manhattan entre otras.
    - Elección del número óptimo de clústers, aplicando obligatoriamente el método Elbow y dos más de su elección, por ejemplo Average Silhouette, GAP entre otros. Debe presentar gráficas para el número óptimo.
  - Gráficas y resultados de los 3 algoritmos.
5. ¿Qué puede concluir respecto a los métodos utilizados?, ¿entregan resultados similares? Discuta los resultados del trabajo realizado.
6. Es su deber aplicar las técnicas aprendidas en clases y si es necesario innovar no dude en hacerlo.

## Rúbrica de Evaluación

Presenta	Aspectos a evaluar	No aplica (0%)	Deficiente (30%)	Regular (60%)	Bueno (80%)	Destacado (100%)	Puntaje máximo del ítem
Calidad en la presentación del cuaderno	Orden de código Buena ortografía Comentarios (buena redacción y entendibles) Documentación (si utiliza librerías debe especificar)	No incluye estos aspectos.	Solo incluye solo comentarios y el orden de código no es adecuado.	1 aspecto no queda del todo claro y el código no tiene buen orden.	1 aspecto no queda del todo claro.	Todos los aspectos están claros.	5
Pre-procesamiento	Limpiar el dataset según el tipo de dato a trabajar. Implementa los 3 tipos de gráficos. Justifica la normalización de datos. Normaliza los datos. Muestra el conjunto final de datos.	No incluye estos aspectos.	2 aspectos están presentes, pero no detallados en profundidad.	3 aspectos están presentes, pero no detallados en profundidad.	1 aspecto no queda del todo claro.	Todos los aspectos están claros, completos y consistentes.	10
Primer análisis de Clustering	Realiza el análisis Partitioning Clustering  Número óptimo de clusters  Clasificación con el modelo  Matriz de confusión: grupos originales vs clusters creados  Gráficos: grupos originales vs clusters creados  Uso de métricas de evaluación	No incluye estos aspectos.	No queda claro el análisis y solo incluye 1 aspecto	2 aspectos están presentes pero no detallados en profundidad	1 aspecto no queda claro o no está presente.	Todos los aspectos están claros, son completos y consistentes.	25

Segundo análisis de Clustering	Realiza el análisis Hierarchical Clustering Clasificación con el modelo Número óptimo de clusters Utiliza representación gráfica ejemplo "Dendrogramas" si se da el caso. Uso de métricas de evaluación	No incluye estos aspectos.	No queda claro el análisis y solo incluye 1 aspecto.	2 aspectos están presentes, pero no detallados en profundidad	1 aspecto no queda claro o no está presente.	Todos los aspectos están claros, son completos y consistentes.	25
Tercer análisis de Clustering	Realiza el análisis Density-Based Clustering Clasificación con el modelo Número óptimo de clusters Observaciones "outliers" Uso de métricas de evaluación	No incluye estos aspectos.	No queda claro el análisis y solo incluye 1 aspecto.	2 aspectos están presentes, pero no detallados en profundidad	1 aspecto no queda claro o no está presente.	Todos los aspectos están claros, son completos y consistentes.	25
Conclusiones	Responde a las preguntas del <u>punto 3 y 5.</u> Detalla de buena manera sus conclusiones. Domina el tema de clustering.	No incluye conclusiones y no domina el tema.	Responde 1 de las 2 preguntas y concluye que son deficientes o no aplican.	Responde 2 preguntas y obtiene conclusiones, pero no domina bien el tema.	1 aspecto no queda claro o no está presente.	Todos los aspectos están claros y detallados.	10