Universidad de Valparaíso
Facultad de Ingeniería
Escuela de Ingeniería Informática

Bases de Datos III

Tarea Nº3 (20%)

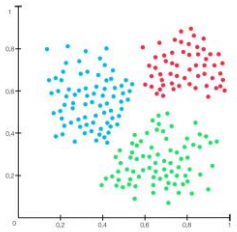
Prof: Ana Aguilera Faraco
Ayudante: Camila Araya

Octubre 2023

Instrucciones:

- La tarea Nº3 es grupal (máximo 3 integrantes, si su grupo de proyecto es de 4 integrantes pueden dividirse y formar grupos de 2 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 ("-"). En caso de no cumplir este formato se descontarán 1 décimas en la nota final.
- Puntaje total: 100 puntos. Nota 4,0: 60 puntos.

Debe 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:
    a. Debe utilizar el siguiente archivo como base ya que contiene el dataset el cual pertenece a datos simulados. Debe cambiar el nombre al momento de entregar.

2. Pre-procesamiento de datos:
    a. Limpiar el dataset según el tipo de dato a trabajar.
    b. 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).
    c. Normalizar los datos.
    d. 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:
    a. 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.* Es deber suyo investigar y aplicar los métodos correspondientes indicando el por qué trabajará ese método y la fuente de donde investigó mediante un comentario. Puede ser de ayuda Referencia Clustering.
    b. Medidas de distancia, por ejemplo, Euclídea, Manhattan entre otras.

c. Elección del número óptimo de clústers, por ejemplo, método Elbow, Average Silhouette, GAP entre otros. Debe presentar gráficas para el número óptimo.
d. Gráficas y resultados de los 3 algoritmos y si necesita mayor detalle debe hacerlo.

5. ¿Qué puede concluir respecto de ambos métodos utilizados?, ¿entregan resultados similares? saque conclusiones del trabajo realizado.

6. Es su deber aplicar las técnicas aprendidas en clases y si es necesario innovar no dude en hacerlo.

Universidad de Valparaíso
Facultad de Ingeniería
Escuela de Ingeniería Informática

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	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.	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"	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 punto 3 y 5. • 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