МІНІСТЕРСТВО ОСВІТИ I НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ   
ІМЕНІ ІГОРЯ СІКОРСЬКОГО»

ФАКУЛЬТЕТ БІОМЕДИЧНОЇ ІНЖЕНЕРІЇ

КАФЕДРА БІОМЕДИЧНОЇ КІБЕРНЕТИКИ

**Комп’ютерний практикум №4**

з дисципліни «Інтелектуальний аналіз даних»

на тему: «Ієрархічна кластеризація»

Варіант №1

**Виконав:**

Ст. 3 курсу гр. БС-12

Адамов О.В.

**Перевірила:**

ст. вик. БМК

Настенка Є. А.

Зараховано від \_\_\_.\_\_\_.\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(підпис викладача)

Київ-2023

**Завдання:**

1. Реалізувати ієрархічну кластеризацію у вигляді функції. В якості метрики відстані використовувати квадрат відстані Евкліда.
2. Функція має повертати матрицю класифікації для варіантів розділення на 3-7 кластерів.
3. Провести кластерний аналіз з використанням створеної функції.
4. Зробити висновки про теоретичну оптимальну кількісті кластерів, базуючись на матриці відстані та на відстані між кластерами.

**Вхідні дані:**За варіантами. Див. вкладений файл. УВАГА! Для ієрархічної кластеризації необхідно взяти 20 випадкових об'єктів.

**Вихідні дані:**

1. Матриця класифікації (приклад можна знайти в pdf файлі).

**Виконання:**

***Теоретичні дані:***

Кластеризація є статистичним методом, який має на меті розділення об'єктів на групи, що називаються кластерами. В межах одного кластера об'єкти схожі між собою, тоді як об'єкти з різних кластерів мають істотні розбіжності. Ця задача належить до області навчання без вчителя. При кластеризації групування може базуватися на відстані між об'єктами, їхньої щільності у даних чи інших статистичних характеристиках.

Кластерний аналіз — це не завжди автоматичний процес. Часто потрібно коригувати параметри та методи аналізу, щоб досягти бажаного результату. Визначення оптимальної кількості кластерів та вибір правильної метрики часто є суб'єктивними і залежать від досвіду спеціаліста.

Об'єкти в межах кластера мають бути подібними між собою, відмінно від об'єктів інших кластерів. Кластеризація знаходить застосування в ряді сфер, включаючи машинне навчання, аналіз даних та інше. Є дві основні форми кластеризації: жорстка та м'яка, де в жорсткій кожен об'єкт належить лише одному кластеру, а в м'якій — може належати кільком кластерам з певною ймовірністю.

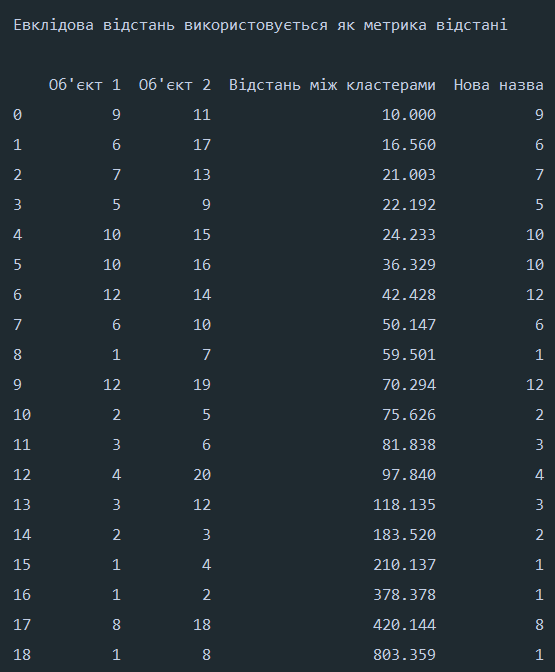
Дані для кластерного аналізу можуть бути різних типів, що може ускладнити процес аналізу.

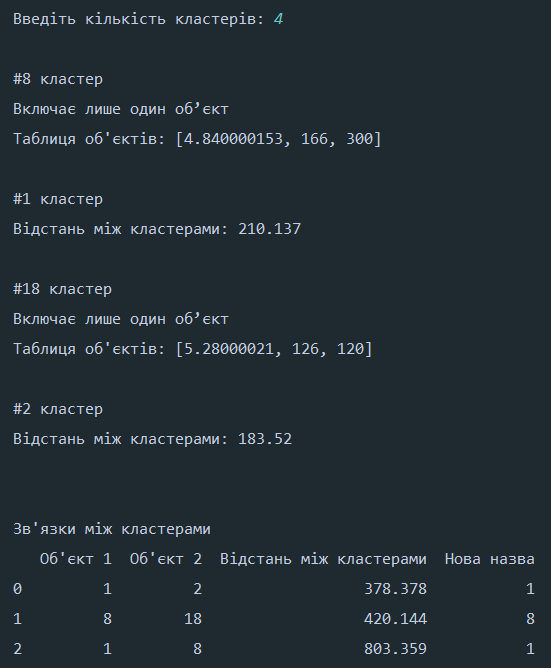
Ієрархічна кластеризація включає в себе підходи, які створюють систему вкладених груп. Тут існують агломератні методи, де малі кластери об'єднуються, та дивізійні, де великі кластери розділяються. Алгоритм такої кластеризації базується на обчисленні відстаней між об'єктами і їх подальшому об'єднанні або розділенні до досягнення заданої умови.

***Набір даних, що був використаний:***

| Об'єкт | Гемоглобін | Вага | Протромбіновий час |
| --- | --- | --- | --- |
|  |
| 1 | 4.309999943 | 121 | 720 |  |
| 2 | 4.440000057 | 218 | 360 |  |
| 3 | 4.639999866 | 256 | 180 |  |
| 4 | 4.630000114 | 119 | 480 |  |
| 5 | 7.71999979 | 183 | 300 |  |
| 6 | 4.809999943 | 190 | 195 |  |
| 7 | 4.840000153 | 191 | 720 |  |
| 8 | 3.940000057 | 170 | 1020 |  |
| 9 | 4.840000153 | 166 | 300 |  |
| 10 | 5.78000021 | 202 | 240 |  |
| 11 | 4.769999981 | 156 | 300 |  |
| 12 | 4.96999979 | 195 | 90 |  |
| 13 | 4.46999979 | 170 | 720 |  |
| 14 | 4.590000153 | 165 | 60 |  |
| 15 | 4.670000076 | 183 | 225 |  |
| 16 | 3.410000086 | 157 | 240 |  |
| 17 | 4.329999924 | 183 | 180 |  |
| 18 | 4.53000021 | 159 | 1440 |  |
| 19 | 5.28000021 | 126 | 120 |  |
| 20 | 11.239999779 | 196 | 540 |  |

***Отримані результати:***





**Висновки про теоретичну оптимальну кількісті кластерів, базуючись на матриці відстані:**

Будемо визначати за допомогою "методу ліктя" на дендрограмі. Шукаємо "лікоть" на графіку відстаней між об'єднаними кластерами, якщо розглядати їх у порядку збільшення. "Лікоть" відповідає точці, де зростання відстані між об'єднаними кластерами стає більш стрімким. Тобто це місце, де об'єднання кластерів призводить до значного збільшення відстані між ними. Шукатимемо місце, де різниця в відстанях між послідовними об'єднаннями стає значною. Отримаємо наступне значення:



**Висновки до роботи:**

У даній роботі спочатку було проведено ознайомлення з основами кластеризації, включаючи ієрархічну кластеризацію та її агломеративний підхід. На наступному етапі була здійснена програмна реалізація ієрархічної кластеризації із застосуванням евклідової відстані як основної метрики. У результаті була отримана класифікаційна матриця, на підставі якої було зроблено висновки щодо теоретичної оптимальної кількості кластерів для представлених даних.