Міністерство освіти і науки України  
НТУУ «КПІ ім. Ігоря Сікорського»

Навчально-науковий інститут атомної та теплової енергетики

Кафедра цифрових технологій в енергетиці

Лабораторна робота №5

з дисципліни «Вступ до інтелектуального аналізу даних»

Тема «Ієрархічна кластеризація та дендрограми в Python»

Варіант №16

Виконав:

студент 3-го курсу, НН ІАТЕ гр. ТР-23

Ровний Григорій

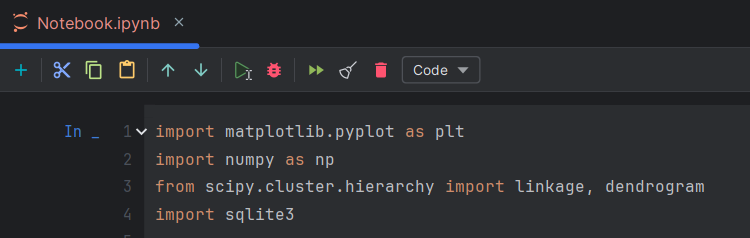
Перевірив: проф. Путренко В.В.

КИЇВ 2025

**Мета:** Опрацювати приклад роботи з кластеризацією та дендрограмами, використовуючи Jupyter Notebook. Виконати поставлене завдання.

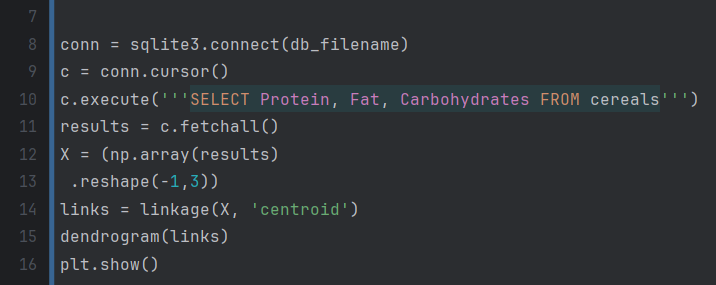
**Хід роботи**

Підключаємо потрібні бібліотеки:



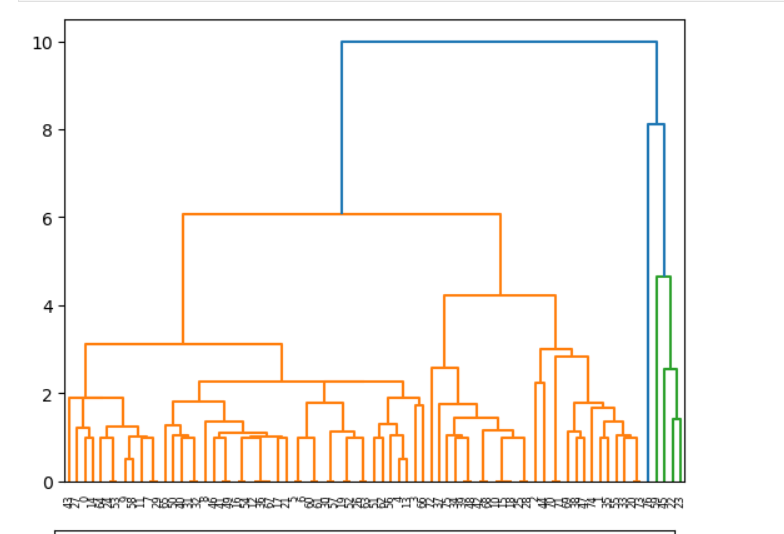
Ці рядки імпортують необхідні бібліотеки. Бібліотека matplotlib потрібна для відображення графіка в середовищі Jupyter Notebook, numpy використовується для обробки даних. Бібліотека scipy побудована на numpy та надає функції для створення та відображення дендрограми.

Створення дендрограми:

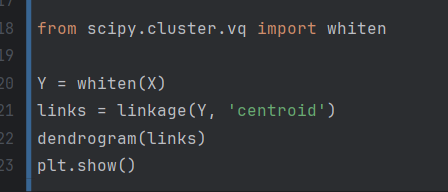


Ці три рядки за збір даних із бази даних і упаковку їх у відповідний формат для scipy. Перший рядок виконує запит і збирає вміст білка, жиру та вуглеводів для кожної крупи. Другий — перетворює отримані дані з формату кортежу, повернутого sqlite3, на матрицю numpy, а потім транспонує матрицю в правильний формат. Четвертий відповідає за обчислення як кластеризації, яка інформує дендрограму, так і зв’язків дендрограми. Функція приймає дані, X і специфікатор метрики відстані для використання під час обчислення відстаней кластера. У цьому випадку ми використовуємо «центроїд», який обчислює центроїд кожного кластера, а потім відстань між ними під час об’єднання кластерів.

Як бачимо наступний результат:

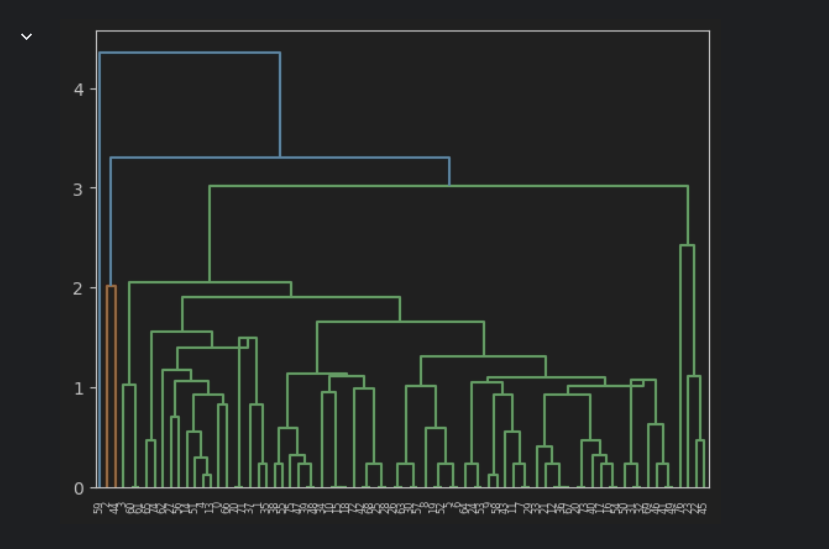


Далі ми нормалізуємо ті самі дані, а потім знову намалюємо дендрограму:



Щоб нормалізувати дані, ми імпортуємо функцію «whiten» із scipy, а потім застосовуємо її до транспонованої матриці даних. Це нормалізує значення ознак до діапазону 0 —> 1.

Отримуємо наступний результат:



**Висновок**: Під час виконання лабораторної роботи були набуті практичні навички роботи з дендрограмами та кластеризацією. Було побудовано дві дендрограми.

**Програмний код:**

import matplotlib.pyplot as plt  
import numpy as np  
from scipy.cluster.hierarchy import linkage, dendrogram  
import sqlite3  
  
db\_filename = 'data/cereals-1.db'  
  
conn = sqlite3.connect(db\_filename)  
c = conn.cursor()  
c.execute('''SELECT Protein, Fat, Carbohydrates FROM cereals''')  
results = c.fetchall()  
X = (np.array(results)  
 .reshape(-1,3))  
links = linkage(X, 'centroid')  
dendrogram(links)  
plt.show()  
  
from scipy.cluster.vq import whiten  
  
Y = whiten(X)  
links = linkage(Y, 'centroid')  
dendrogram(links)  
plt.show()