Прізвище: **КИРИЛЮК**
Ім'я: **Дмитро**
Група: **ПП-22**
Варіант: **08**
Дата захисту: **24.03.2025р**.

Кафедра: **САПР**
Дисципліна: **Системи інтелектуального аналізу та візуалізації даних**
Перевірив: **Андрій КЕРНИЦЬКИЙ**

ЗВІТ
до лабораторної роботи №05
на тему "**Алгоритми кластеризації k-внутрішніх середніх та ієрархічної кластеризації.**"

Мета роботи: вивчення та застосування трьох методів кластеризації в середовищі Weka: поділяючого методу кластеризації K-середніх (SimpleKMeans), ієрархічного методу кластеризації (HierarchicalClusterer) та ієрархічного методу кластеризації COBWEB (COBWEB). Студенти мають набути навичок роботи з цими алгоритмами, вміти прикладати їх для аналізу реальних даних та інтерпретувати отримані результати.

Тренувальне завдання

Вправа 1

Виконайте наступні завдання для набору даних 'bank.arff':
а. Запустіть алгоритм кластеризації SimpleKMeans, задаючи значення параметра K (кількість кластерів) від 1 до 12.
б. Запишіть в таблицю значення сум квадратичних помилок, одержуваних при різних значеннях K. Що означає цей параметр і як змінюються його значення?
в. Для значення K=5 вкажіть:

- скільки кластерів було створено;
- скільки примірників потрапило в кожен з кластерів (вказати кількість і відсоток);
- скільки ітерацій знадобилося для кластеризації даних; о складіть таблицю з характеристиками центроїдів.

г. Для значення K=5 візуалізуйте результати кластеризації (по осі абсцис відкласти назву (номер) кластера, по осі ординат - номер примірника в кластері) та дайте оцінку отриманим результатам:

- чи є значна відмінність у значеннях атрибуту «вік» (age) між кластерами?
- у яких кластерах домінують жінки (female), а в яких чоловіки (male)?
- що можна сказати про значення атрибута «регіон» (region) у кожному кластері?
- що можна сказати про розкид значень атрибуту «дохід» (income) між кластерами?
- у яких кластерах домінують сімейні люди (married), а в яких неодружені (unmarried)?
- у якій кластер потрапило найбільше людей з машинами?
- у яких кластерах переважають люди з ощадними рахунками (savings accounts)?
- що можна сказати про розкид значень атрибуту «поточний банківський рахунок» (current account) між кластерами?

- що можна сказати про розкид значень атрибуту «іпотека» (mortgage holdings) між кластерами?
- які кластери в основному складаються з людей, які придбали PEP (особистий план купівлі акцій), і які з людей, які не придбали його?

Вправа 2

1. Виконайте наступні завдання для набору даних ‘iris.arff’

а. Запустіть алгоритм кластеризації SimpleKMeans з K=3 та оцініть якість кластеризації, порівнюючи кластери з попередньо заданими класами:

- запишіть значення суми квадратичних помилок, кількість об'єктів в кластерах та характеристики кожного центроїду;
- проаналізуйте як співвідносяться кластери та значення цільового атрибуту, скільки екземплярів було віднесено до «невірних» кластерів, який клас виявився «складним» для виділення;
- візуалізуйте результати, використовуючи різні атрибути для осі ординат (при візуалізації екземпляри, позначені квадратами були віднесені до «невірного» кластеру);
- визначте, на що впливає параметр «seed» і чому він є важливим при кластеризації методом k-середніх; для цього проведіть експерименти з різними значеннями параметру і порівняйте отримані результати.

Вправа 3. Ієрархічна кластеризація

1. Завантажте набір даних ‘flagdata.arff. Цей файл представляє атрибути прапорів деяких європейських країн. Виконайте наступні дії:

- Запустіть алгоритм COBWEB з параметрами C=0,4 (0,35), saveInstanceData = True, cluster mode = Use training set;
- візуалізуйте отриману дендрограму та запишіть її, вкажіть, які країни потрапили в який кластер;
- вкажіть, що спільного у прапорів, що опинилися в одному кластері.

2. Завантажте набір даних ‘zoo.arff’ і виконайте наступні завдання:

- оберіть з вибірки частину тварин на власний розсуд (наприклад, ссавців);
- запустіть алгоритм Hierarchical Clusterer (тип тварини не використовувати в кластеризації, а назву за допомогою фільтру перетворити на рядковий тип – NominalToString);
- проекспериментуйте з налаштуванням алгоритму та візуалізуйте результати його роботи;
- оцініть, чи є логічний сенс в створюваних кластерах.

Індивідуальне завдання:

Моє завдання для цієї лабораторної роботи - оцінити алгоритми кластеризації за допомогою Weka.

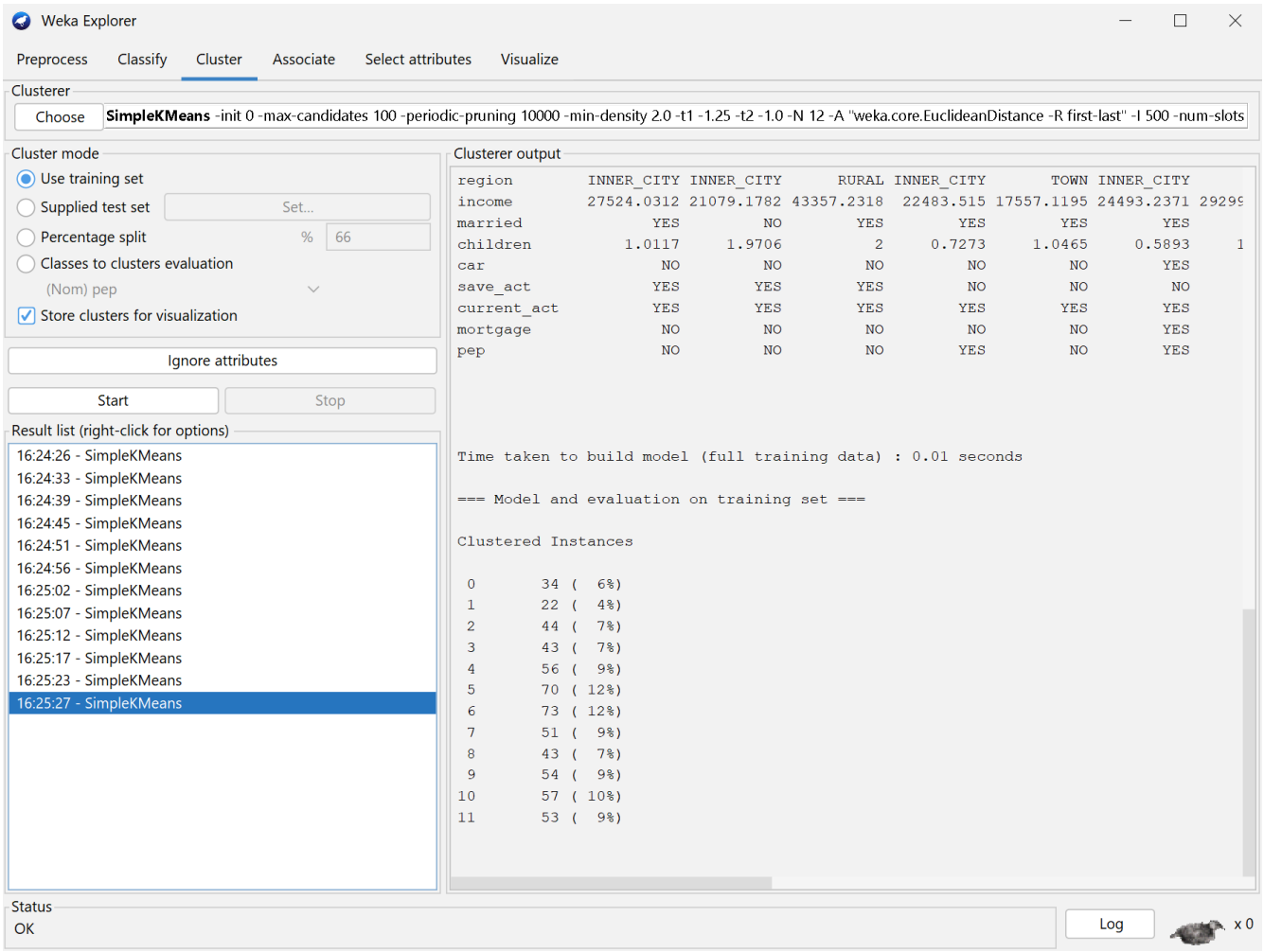
1. Використайте набір даних, який ви вибрали для лабораторної №1. Якщо він не підходить для задач кластеризації, то виберіть інший, який підходить.
2. Визначте, як ви будете вимірювати якість сформованих кластерів.

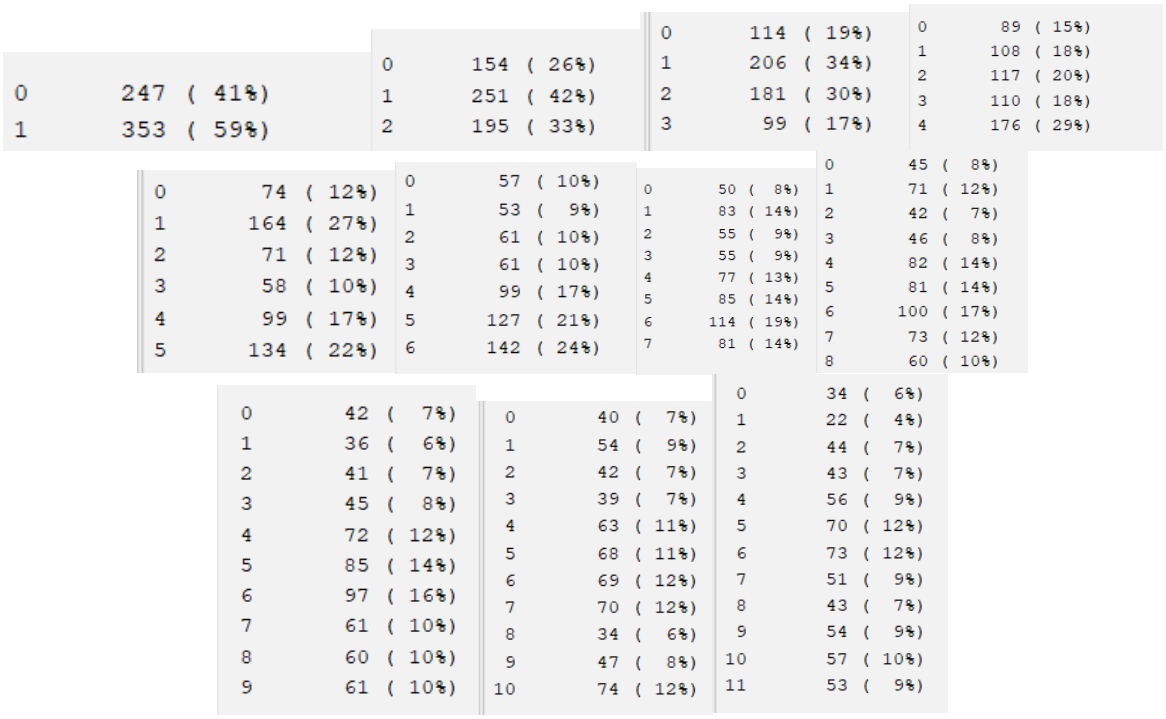
3. Для свого набору даних застосуйте три алгоритми кластеризації та порівняйте їх результати, використовуючи ваші показники якості.
4. Напишіть короткий звіт:
a. Опишіть набори даних та ваші показники якості.
b. Опишіть налаштування експерименту, наприклад, як ви попередньо обробили дані (якщо такі є), як вибрали параметри для вибраних алгоритмів (якщо такі є) та чому.
c. Представити результати експерименту. Вони не повинні бути простим копіюванням та вставкою з вихідних даних Weka, а скоріше представленими у вигляді таблиці або діаграми для зручності порівняння.
d. Запропонуйте ідеї та зробіть висновки зі своїх експериментів. Наприклад, чи різні методи кластеризації мають різницю щодо якості або продуктивності для певних наборів даних, які ви вибрали? І чому? Як може допомогти попередня обробка даних? Чи існують умови або загальні типи наборів даних, які роблять певні алгоритми більш придатними, ніж інші?

Результати виконання програми

Вправа 1

A)





Б)

K(кількість кластерів)	SSE (Сума квадратичних помилок)
0	2699.746484076493
1	2335.279199193316
2	2165.473046551613
3	2047.7296232073923
4	2047.391840488442
5	1955.4146634784236
6	1920.955483833581
7	1840.2291016426962
8	1778.1785075244788
9	1752.556839862609
10	1725.6457142351605
11	1667.1175568026724

SSE (Sum of Squared Errors) – це сума квадратів відстаней кожного об'єкта до центроїда його кластера. Це метрика, яка показує, наскільки добре об'єкти згруповані в кластери.

Як змінюється SSE при збільшенні K?

- Зменшення SSE при зростанні K:
  - При K=0 всі об'єкти належать одному кластеру, і помилка висока (2699.75).
  - Зі збільшенням K значення SSE зменшується (від 2699.75 → 1667.12 при K=11), оскільки більше кластерів означає кращу відповідність даних.
- Зменшення SSE стає менш значним:
  - Спочатку SSE зменшується різко, але після певного значення K швидкість спаду уповільнюється.
  - Це можна побачити в діапазоні K=4–11, де SSE зменшується не так швидко.

- Метод "лікоть" (Elbow Method):
  - На графіку SSE(K) зазвичай можна знайти точку "зламу" (elbow), яка показує оптимальне K.
  - Наприклад, якщо різке зниження SSE припиняється біля K=6 або K=7, це може бути оптимальним K.

В) Для значення K = 5:

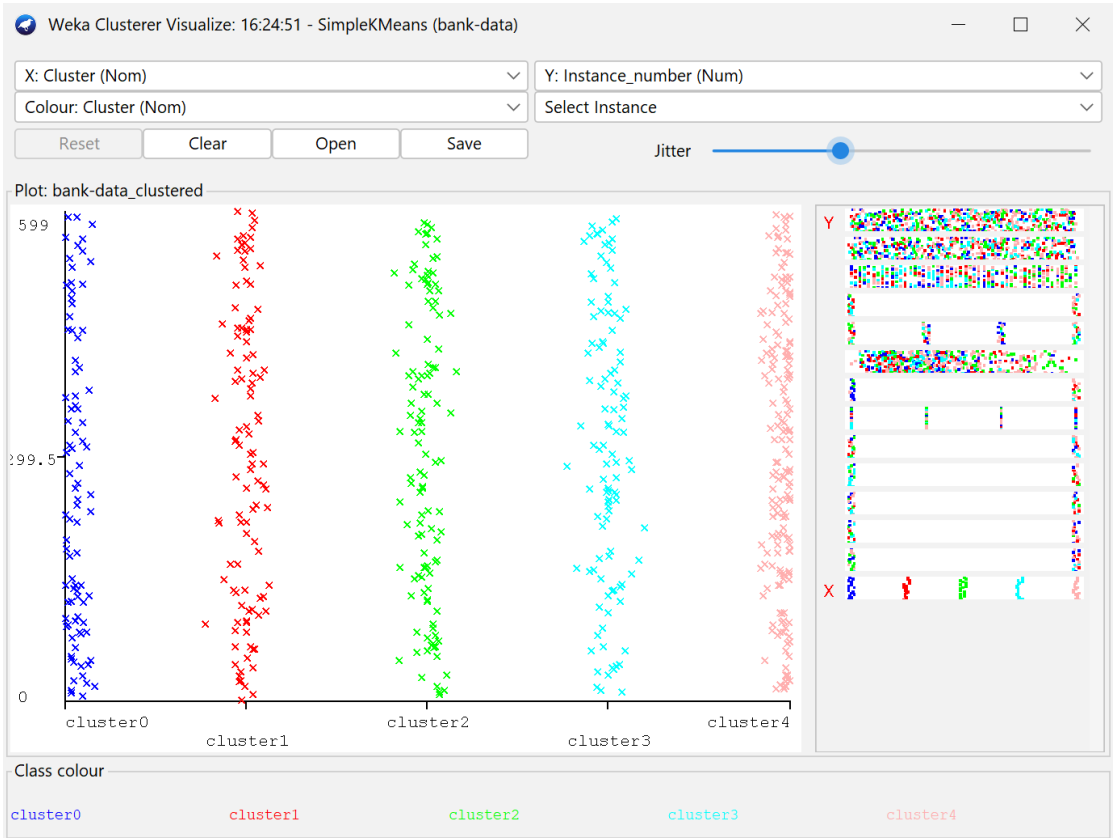
Було створено 5 кластерів

```
0        89 ( 15%)
1       108 ( 18%)
2       117 ( 20%)
3       110 ( 18%)
4       176 ( 29%)
```

```
Number of iterations: 6
```

```
Cluster 0: ID12614,25,FEMALE,RURAL,14505.3,NO,3,NO,YES,YES,NO,NO
Cluster 1: ID12131,61,FEMALE,RURAL,22942.9,YES,2,NO,YES,YES,NO,NO
Cluster 2: ID12190,54,FEMALE,INNER_CITY,31095.6,YES,2,NO,NO,YES,NO,YES
Cluster 3: ID12485,36,FEMALE,TOWN,26920.8,YES,0,NO,NO,YES,NO,NO
Cluster 4: ID12203,42,MALE,INNER_CITY,15499.9,YES,0,YES,NO,YES,YES,YES
```

Г)



Відповіді на запитання:

1) age 42.395 35.8824 55.7273 43.0227 29.8605 36.1964 43.6429 43.3562 51.8824 44.2558 41.3889 40.5965 46.5849

На основі таблиці кластерних центроїдів у Weka видно, що середній вік (age) значно відрізняється між кластерами:

- Найнижчий середній вік: 29.86
- Найвищий середній вік: 55.72

Також помітно, що в інших кластерах середній вік варіюється в діапазоні 35.88 - 51.88. Це свідчить про те, що вік є вагомим фактором у формуванні кластерів, оскільки деякі групи мають значно молодший або старший склад, що може вказувати на різні соціально-економічні характеристики цих груп.

2)

У кластерах 0,1,2,3,6,7,8,10 домінують жінки, а у 4,5,9,11 відповідно чоловіки.

Це може вказувати на певний гендерний вплив у кластеризації, що, ймовірно, пов’язано з іншими характеристиками, такими як дохід, сімейний статус чи регіон проживання.

3)

На основі таблиці значення атрибуту «регіон» (region) розподіляються між кластерами наступним чином:

- Найчастіше зустрічається значення INNER_CITY (внутрішнє місто), що вказує на значну частку мешканців із цієї зони.
- Деякі кластери мають значення RURAL (сільська місцевість) або TOWN (містечко), що означає, що ці групи можуть мати інший соціально-економічний профіль.
- Деякі кластери складаються лише з представників INNER_CITY, що може означати концентрацію певних характеристик серед міських мешканців.

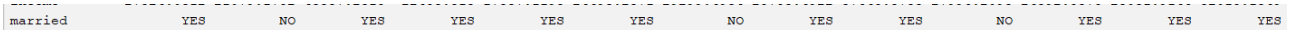
Таким чином, кластеризація враховує регіон проживання, і різниця в розподілі може впливати на інші атрибути, такі як дохід, сімейний стан та наявність автомобіля.

4)

Розглянемо розкид значень атрибуту «дохід» (income) між кластерами:

- Найнижчий середній дохід: 17,557.12
- Найвищий середній дохід: 43,357.23
- Загальний середній дохід по всіх даних: 27,524.03

Розкид доходів досить значний – різниця між найменшим і найбільшим середнім доходом складає приблизно 25,800, що вказує на суттєву фінансову неоднорідність між кластерами.

5) 

У кластерах 1,2,3,4,6,7,9,10,11, домінують одружені,а у 0,5,8 не одружені.

Це може вказувати на те, що певні кластери містять більше молодих або самостійних людей, тоді як інші — сімейних, можливо, з дітьми. Вплив цього атрибута може корелювати з іншими характеристиками, такими як дохід, регіон проживання та наявність іпотеки.

6)

car	NO	NO	NO	NO	NO	YES	YES	NO	NO	YES	NO	YES	NO

Згідно з таблицею, автомобілі мають представники кластерів:

- Кластер 4
- Кластер 5
- Кластер 8
- Кластер 10

Щоб визначити, в якому кластері найбільше людей з машинами, порівняємо розмір кластерів:

Attribute	Full Data	0	1	2	3	4	5	6	7	8	9	10	11
	(600.0)	(34.0)	(22.0)	(44.0)	(43.0)	(56.0)	(70.0)	(73.0)	(51.0)	(43.0)	(54.0)	(57.0)	(53.0)

- Кластер 4 – 56 осіб
- Кластер 5 – 70 осіб
- Кластер 8 – 43 особи
- Кластер 10 – 57 осіб

Отже, найбільше людей з машинами потрапило в кластер 5, який налічує 70 осіб.

7)

save_act	YES	YES	YES	NO	NO	NO	YES	YES	YES	YES	YES	YES	YES

- Кластери, де переважають люди з ощадними рахунками (YES):
  Кластери 0, 1, 5, 6, 7, 8, 9, 10, 11 – у цих групах більшість мають ощадні рахунки.
- Кластери, де переважають люди без ощадних рахунків (NO):
  Кластери 2 , 3, 4 – у цих групах більше людей без ощадних рахунків.

Більшість кластерів мають значну частку людей із ощадними рахунками, що може вказувати на їхню фінансову грамотність або стабільність. Однак у кластерах 2, 3, 4 люди, ймовірно, менш схильні до заощаджень, що може бути пов’язано з рівнем доходу чи способом життя.

8)

current_act	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	NO	YES	YES

Розкид значень для атрибуту «current_act» мінімальний, оскільки у всіх кластерах люди мають поточний банківський рахунок. Це може свідчити про високу фінансову залученість вибірки, тобто більшість людей користуються банківськими послугами.

9)

mortgage	NO	NO	NO	NO	NO	YES	NO	NO	NO	YES	NO	NO	YES

- Кластери, де більшість мають іпотеку (YES):
  Кластери 4, 8, 11 – у цих групах більше людей з іпотекою.
- Кластери, де більшість не мають іпотеки (NO):
  Кластери 0, 1, 2, 3, 5, 6, 7, 9, 10 – у цих групах переважають люди без іпотеки.

У більшості кластерів люди не мають іпотеки, що може вказувати на те, що вони або орендують житло, або вже його виплатили. Кластери 4, 8 і 11 містять більше людей з іпотекою, що може бути пов’язано з рівнем доходу або регіоном проживання. Розкид значень атрибута досить великий, оскільки є кластери з високою та низькою часткою іпотечних позичальників.

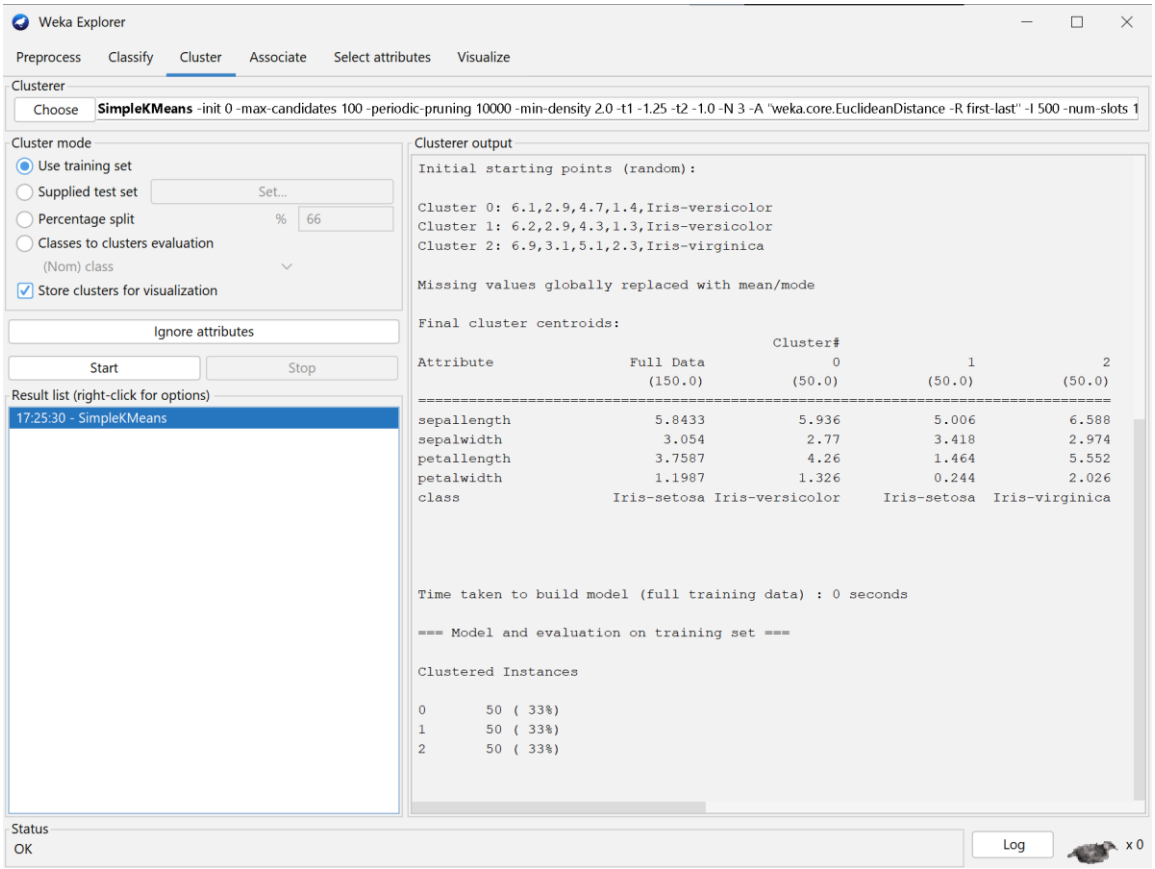
10)

pep	NO	NO	NO	YES	NO	YES	YES	NO	YES	NO	NO	NO	YES

У деяких випадках можна помітити, що певні групи мають схильність до купівлі PEP, наприклад, це можуть бути люди з вищими доходами або певного віку. Тому важливо подивитися й інші характеристики кластерів, щоб побачити закономірності.

Вправа 2

A)



Сума квадратичних помилок (SSE):

```
Within cluster sum of squared errors: 6.613823274690356
```

Кількість об'єктів у кожному кластері:

```
Cluster 0: 6.1,2.9,4.7,1.4,Iris-versicolor
Cluster 1: 6.2,2.9,4.3,1.3,Iris-versicolor
Cluster 2: 6.9,3.1,5.1,2.3,Iris-virginica
Cluster 3: 5.5,4.2,1.4,0.2,Iris-setosa
```

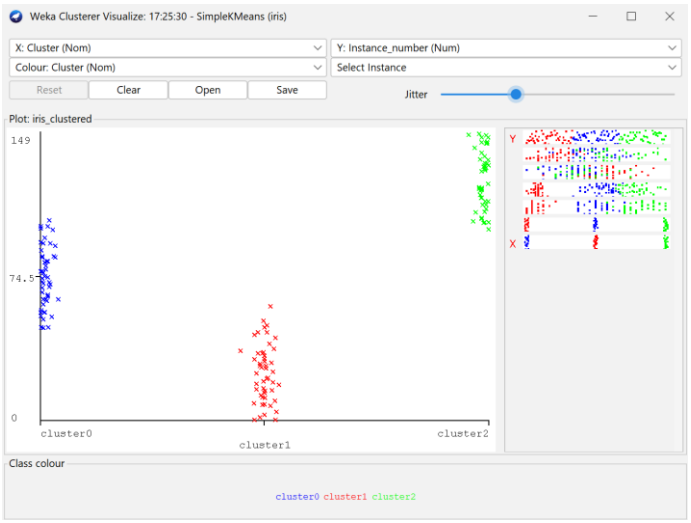
```
0        24 ( 16%)
1        26 ( 17%)
2        50 ( 33%)
3        50 ( 33%)
```

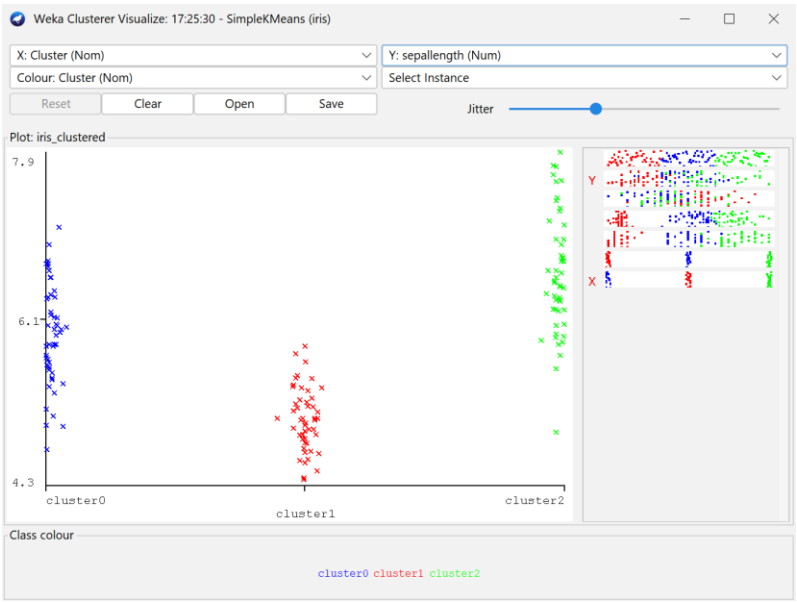
Характеристики центроїдів:

```
Final cluster centroids:
                                    Cluster#
Attribute               Full Data           0           1           2           3
                          (150.0)      (24.0)      (26.0)      (50.0)      (50.0)
=============================================================================================
sepallength               5.8433      6.3292      5.5731       6.588       5.006
sepalwidth                 3.054      2.9792      2.5769       2.974       3.418
petallength               3.7587         4.6      3.9462       5.552       1.464
petalwidth                1.1987      1.4625         1.2       2.026       0.244
class                Iris-setosa Iris-versicolor Iris-versicolor  Iris-virginica    Iris-setosa
```

- Кластер 2 (50 об’єктів) повністю відповідає класу Iris-virginica.
- Кластер 3 (50 об’єктів) добре відповідає класу Iris-setosa.
- Проблема виникає з кластером 0 та 1 (Iris-versicolor), оскільки вони частково змішані (24 і 26 об’єктів) і не ідеально розділені.

Це означає, що Iris-versicolor — "складний" клас для виділення, оскільки його характеристики перетинаються з іншими класами. Таким чином, частина екземплярів Iris-versicolor потрапили в "неправильний" кластер (кластеризація не співпала з реальними класами).

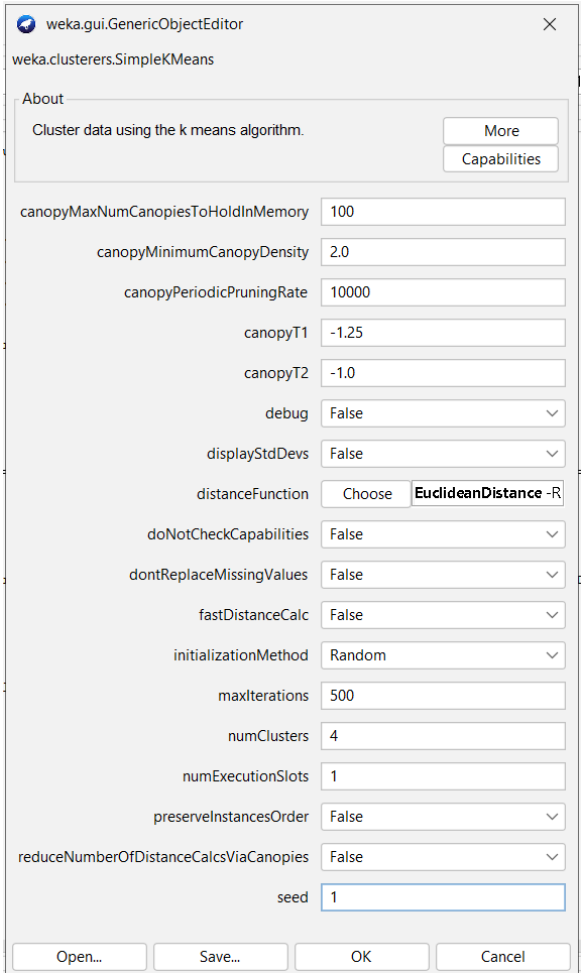




У алгоритмі k-середніх для кластеризації потрібно випадково вибрати початкові центроїди кластерів. Цей вибір залежить від генератора випадкових чисел. Параметр seed задає початкове значення генератора випадкових чисел.
Це значення гарантує, що при однаковому seed результат буде ідентичним (відтворюваність експерименту).

1. K-середніх може дати різні результати при різних seed:
   - Якщо обрати невдалі початкові центроїди, алгоритм може застрягти в локальному мінімумі (неоптимальне рішення).
   - Може вийти погана кластеризація з високою сумою квадратних помилок (SSE).
   - В інший раз при іншому seed алгоритм знайде набагато кращі центроїди.
2. Відтворюваність експериментів:
   - Щоб колеги/наукова спільнота могли повторити ваш експеримент, потрібно зазначити, з яким саме seed ви отримали результат.

Всі попередні дані нас були при seed = 10, давайте спробуєм ще 1 і 100.



Сума квадратичних помилок (SSE):

```
Within cluster sum of squared errors: 6.300610869554293
```

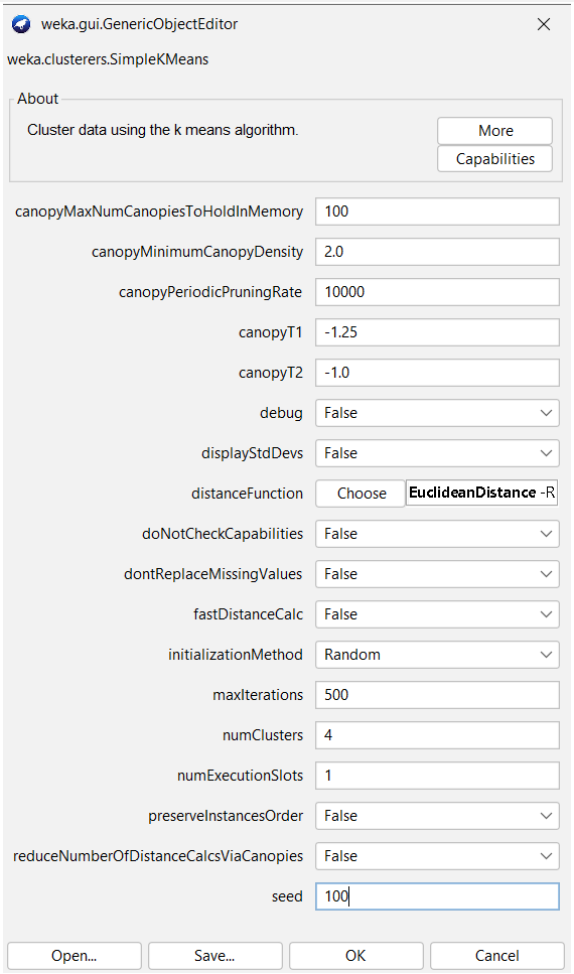
Кількість об'єктів у кожному кластері:

```
Cluster 0: 7.7,3,6.1,2.3,Iris-virginica
Cluster 1: 6.3,2.5,4.9,1.5,Iris-versicolor
Cluster 2: 6.4,2.7,5.3,1.9,Iris-virginica
Cluster 3: 5.1,3.5,1.4,0.2,Iris-setosa
```

```
0        27 ( 18%)
1        50 ( 33%)
2        23 ( 15%)
3        50 ( 33%)
```

Характеристики центроїдів:

```
Final cluster centroids:
                                Cluster#
Attribute               Full Data              0              1              2              3
                          (150.0)         (27.0)         (50.0)         (23.0)         (50.0)
=============================================================================================
sepallength                5.8433              7          5.936         6.1043          5.006
sepalwidth                  3.054         3.1593           2.77         2.7565          3.418
petallength                3.7587         5.8741           4.26         5.1739          1.464
petalwidth                 1.1987         2.1741          1.326         1.8522          0.244
class                  Iris-setosa  Iris-virginica Iris-versicolor  Iris-virginica     Iris-setosa
```



Сума квадратичних помилок (SSE):

```
Within cluster sum of squared errors: 6.856549502288228
```

Кількість об'єктів у кожному кластері:

```
Cluster 0: 6.4,3.2,5.3,2.3,Iris-virginica
Cluster 1: 5.4,3.4,1.5,0.4,Iris-setosa
Cluster 2: 4.4,3,1.3,0.2,Iris-setosa
Cluster 3: 6.6,3,4.4,1.4,Iris-versicolor
```

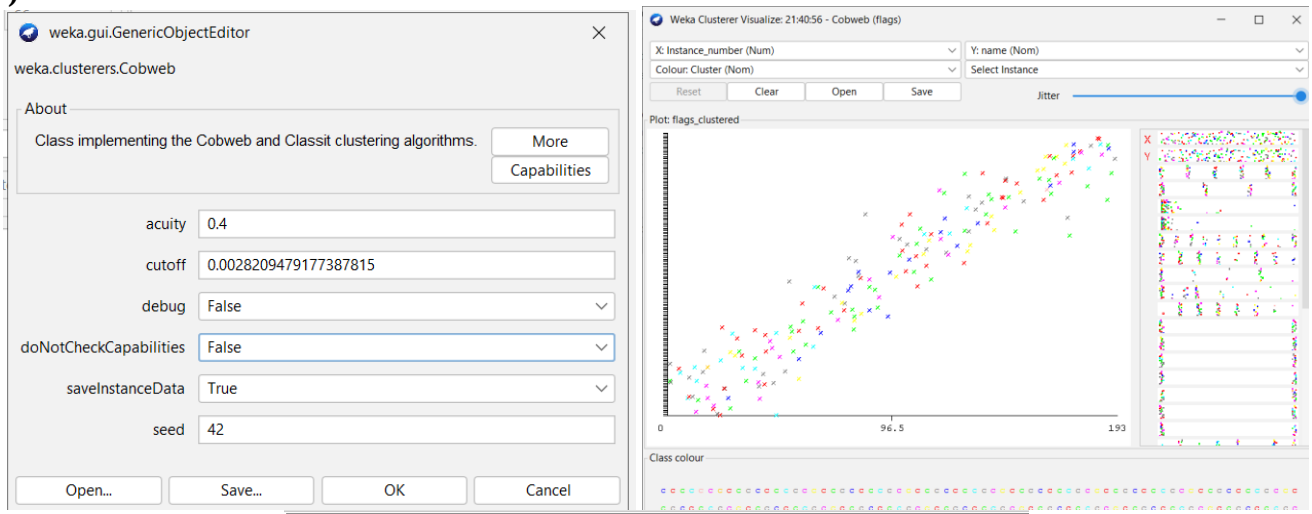
```
Clustered Instances

0       50 ( 33%)
1       28 ( 19%)
2       22 ( 15%)
3       50 ( 33%)
```

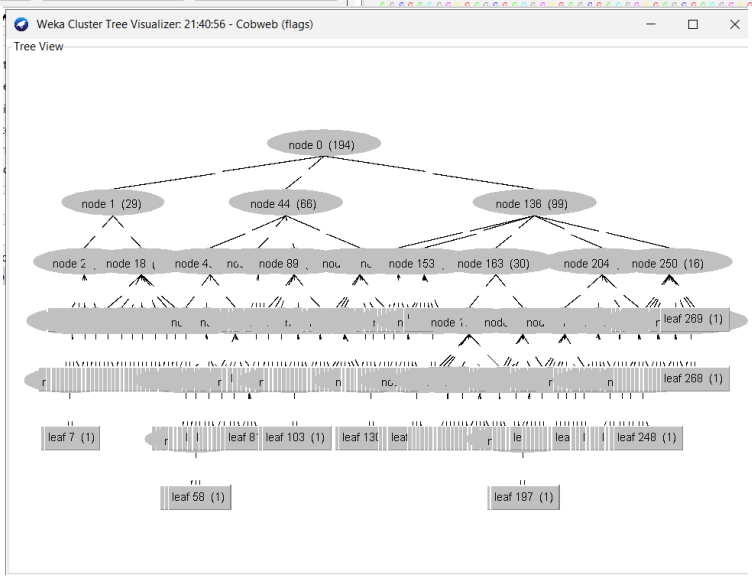
Характеристики центроїдів:

```
Final cluster centroids:
                                Cluster#
Attribute               Full Data           0           1           2           3
                          (150.0)      (50.0)      (28.0)      (22.0)      (50.0)
=============================================================================================
sepallength               5.8433       6.588      5.2321      4.7182       5.936
sepalwidth                 3.054       2.974      3.6679         3.1        2.77
petallength               3.7587       5.552      1.4857      1.4364        4.26
petalwidth                1.1987       2.026      0.2857      0.1909       1.326
class                 Iris-setosa  Iris-virginica  Iris-setosa  Iris-setosa Iris-versicolor
```

Вправа 3

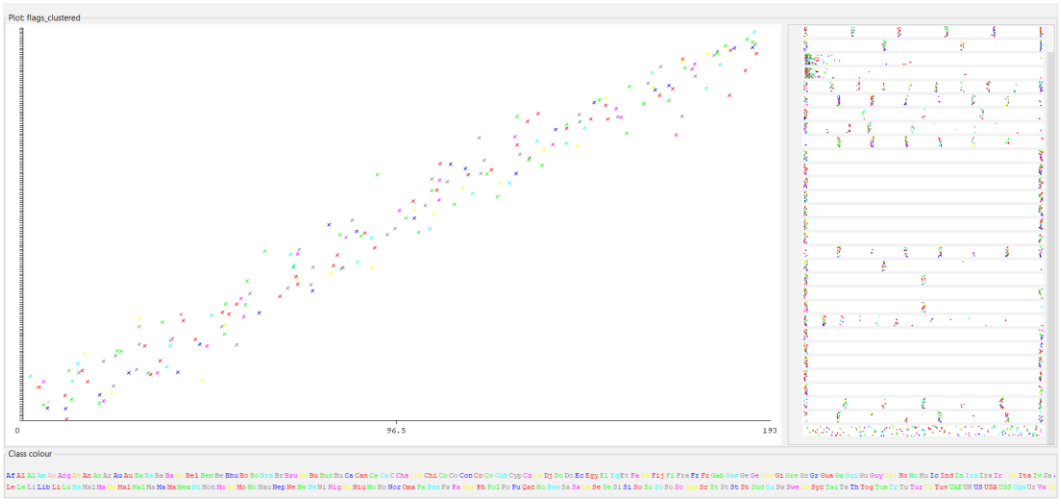
1)



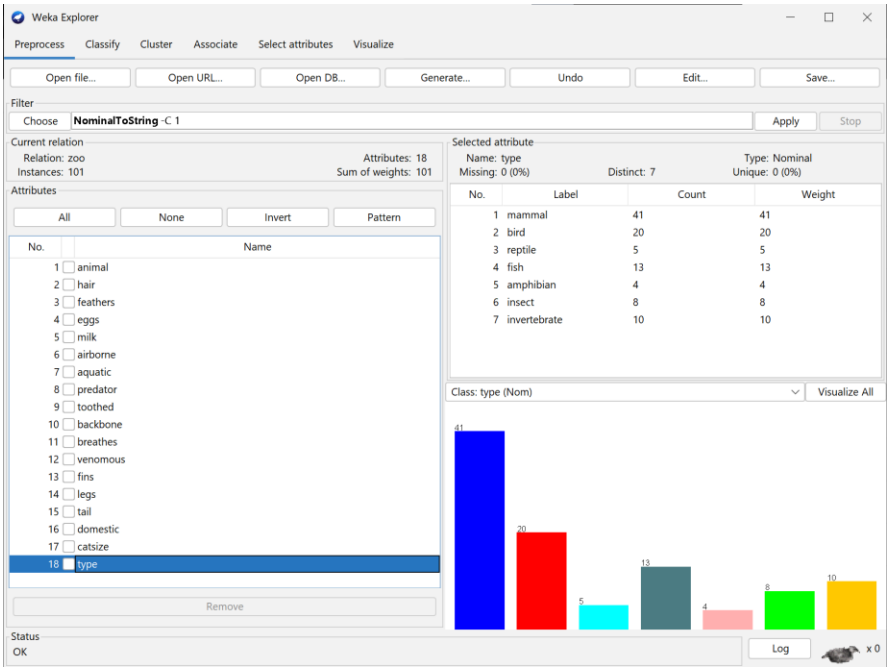


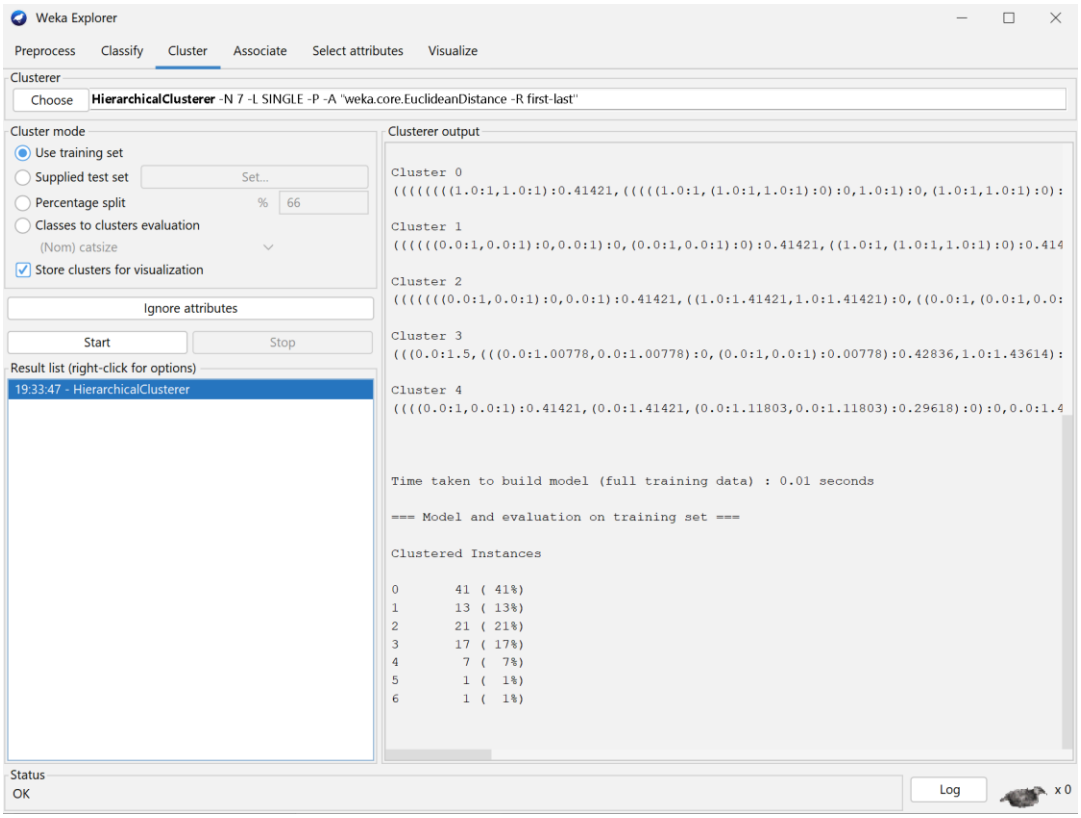
Датасет розбився на 145 кластерів, кількість об'єктів в кластерах розбиті наступним способом:

4	1 (1%)	40	1 (1%)	87	1 (1%)	120	1 (1%)	156	1 (1%)	203	1 (1%)	244	1 (1%)		
6	1 (1%)	42	1 (1%)	88	1 (1%)	121	1 (1%)	157	1 (1%)	206	1 (1%)	246	1 (1%)		
7	1 (1%)	43	1 (1%)	89	2 (1%)	122	1 (1%)	158	1 (1%)	207	1 (1%)	247	1 (1%)		
8	1 (1%)	51	1 (1%)	90	1 (1%)	123	1 (1%)	160	1 (1%)	208	1 (1%)	248	1 (1%)		
9	1 (1%)	52	1 (1%)	91	1 (1%)	125	1 (1%)	161	1 (1%)	210	1 (1%)	249	1 (1%)		
10	1 (1%)	56	1 (1%)	92	1 (1%)	126	1 (1%)	162	1 (1%)	214	1 (1%)	251	1 (1%)		
11	1 (1%)	57	1 (1%)	93	1 (1%)	127	2 (1%)	165	1 (1%)	215	1 (1%)	252	1 (1%)		
12	1 (1%)	58	1 (1%)	94	1 (1%)	129	1 (1%)	166	1 (1%)	217	1 (1%)	253	1 (1%)		
13	1 (1%)	59	1 (1%)	96	1 (1%)	130	1 (1%)	173	1 (1%)	218	1 (1%)	254	1 (1%)		
15	1 (1%)	61	1 (1%)	97	1 (1%)	131	1 (1%)	174	1 (1%)	220	1 (1%)	255	1 (1%)		
16	1 (1%)	62	1 (1%)	98	1 (1%)	133	1 (1%)	175	1 (1%)	221	1 (1%)	256	1 (1%)		
17	1 (1%)	63	1 (1%)	100	1 (1%)	134	1 (1%)	177	1 (1%)	223	1 (1%)	257	1 (1%)		
20	1 (1%)	64	1 (1%)	102	1 (1%)	135	1 (1%)	178	1 (1%)	224	1 (1%)	258	1 (1%)		
21	1 (1%)	65	2 (1%)	103	1 (1%)	138	1 (1%)	180	1 (1%)	225	1 (1%)	259	2 (1%)		
22	1 (1%)	68	1 (1%)	104	1 (1%)	140	1 (1%)	181	1 (1%)	226	2 (1%)	260	1 (1%)		
24	1 (1%)	69	1 (1%)	105	1 (1%)	141	1 (1%)	183	1 (1%)	227	1 (1%)	261	1 (1%)		
25	1 (1%)	71	1 (1%)	106	1 (1%)	142	1 (1%)	184	1 (1%)	229	1 (1%)	262	3 (2%)		
26	1 (1%)	73	1 (1%)	107	1 (1%)	143	1 (1%)	185	1 (1%)	230	1 (1%)	263	1 (1%)		
28	1 (1%)	74	1 (1%)	108	1 (1%)	145	1 (1%)	186	1 (1%)	231	1 (1%)	264	1 (1%)		
29	1 (1%)	75	1 (1%)	109	1 (1%)	146	1 (1%)	188	2 (1%)	233	1 (1%)	265	1 (1%)		
31	1 (1%)	76	1 (1%)	110	1 (1%)	147	1 (1%)	189	1 (1%)	234	1 (1%)	266	1 (1%)		
32	1 (1%)	78	1 (1%)	111	1 (1%)	148	1 (1%)	196	1 (1%)	236	1 (1%)	267	1 (1%)		
33	1 (1%)	79	1 (1%)	112	1 (1%)	150	1 (1%)	197	1 (1%)	237	1 (1%)	268	1 (1%)		
35	1 (1%)	80	1 (1%)	115	1 (1%)	151	1 (1%)	198	1 (1%)	238	1 (1%)	269	1 (1%)		
36	1 (1%)	81	1 (1%)	117	1 (1%)	152	1 (1%)	199	1 (1%)	241	1 (1%)				
37	1 (1%)	85	1 (1%)	118	1 (1%)	154	1 (1%)	200	1 (1%)	242	1 (1%)				
39	1 (1%)	86	1 (1%)	119	1 (1%)	155	1 (1%)	202	1 (1%)	243	1 (1%)				

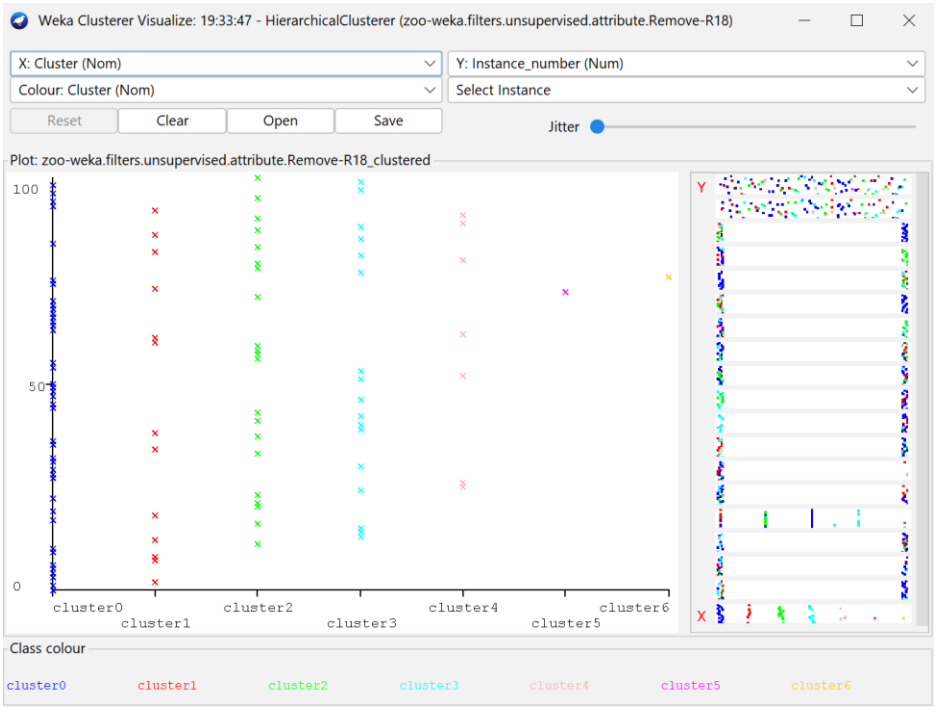


2)



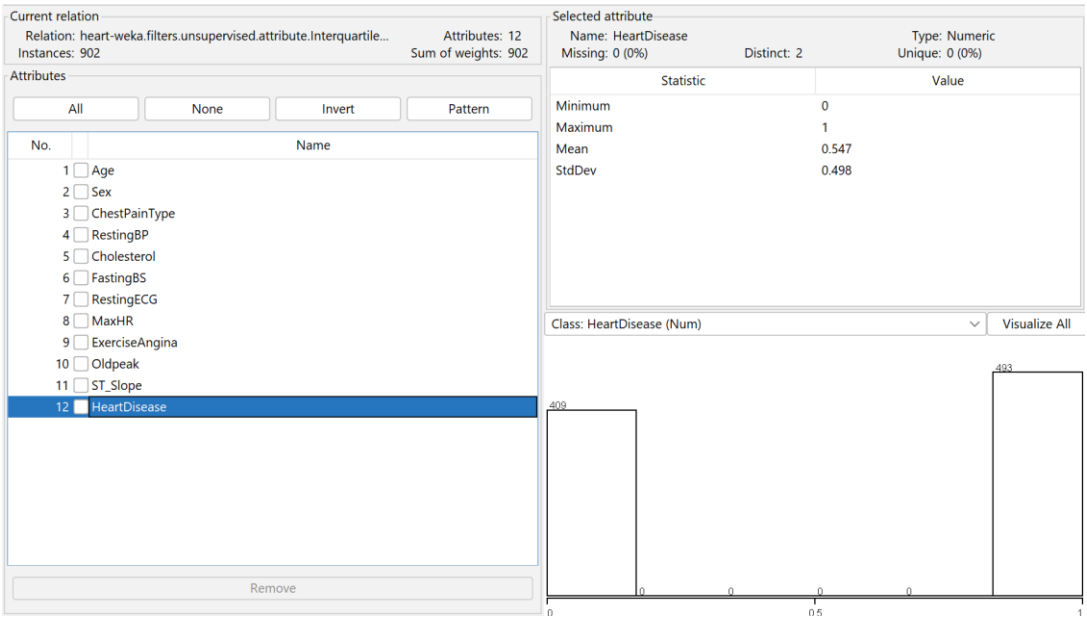


Оскільки є 7 типів тварин то ми поділили на 7 кластерів.



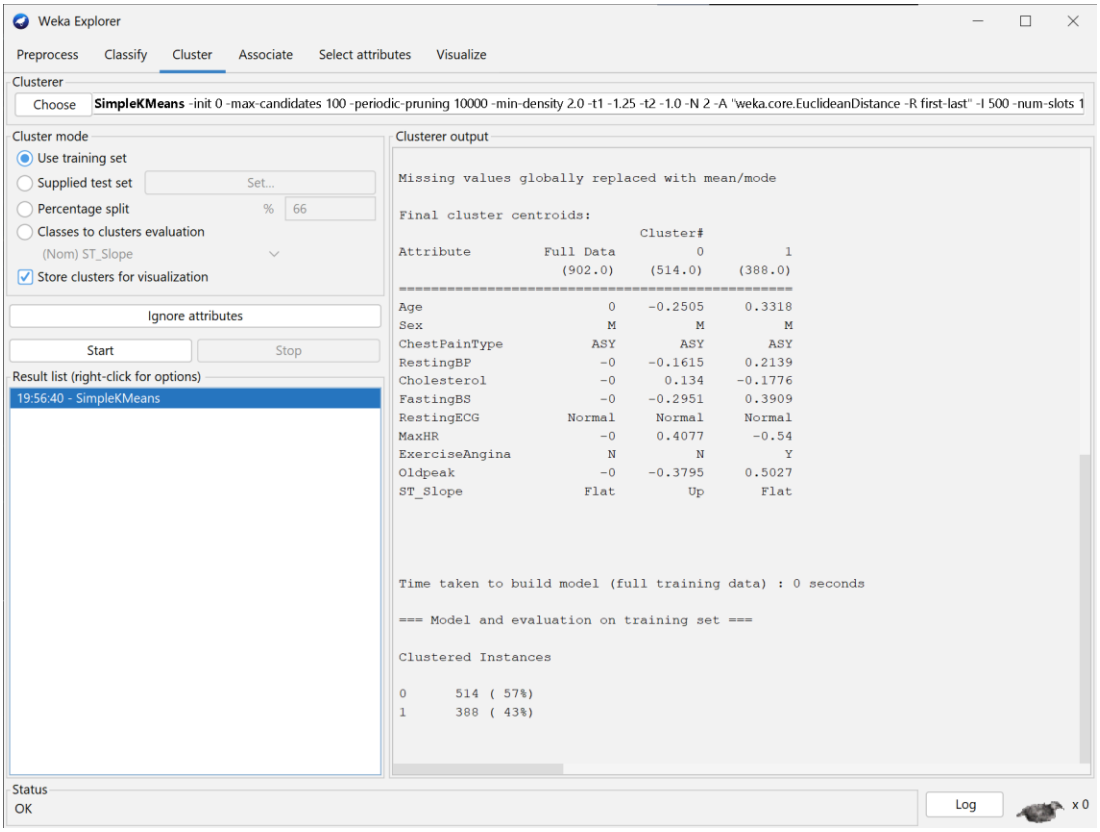
Бачимо, що пропорціно кластери трохи збігаються із типами тварин, проте деякі тварини були неправильно рознесі. Звідси можемо зробити висновок, що такий метод є не дуже ефективний.

Індивідуальне завдання



Якість кластерів будемо порівнювати із значенням атрибуду HeartDisease, який приймає значення (0, 1). Тому кількість кластерів K буде дорівнювати 2.

1)K-Means



- Кількість кластерів:

```
Cluster 0: -1.53053,F,NAP,-1.210289,-0.145797,-0.552283,ST,1.690362,N,-0.86283,Up
Cluster 1: 1.860414,F,NAP,-1.210289,0.611364,1.80866,LVH,-0.269062,N,-0.86283,Up
```

```
0       514 ( 57%)
1       388 ( 43%)
```

- Початкові центроїди:

```
Final cluster centroids:
                                Cluster#
Attribute          Full Data          0          1
                     (902.0)    (514.0)    (388.0)
=================================================
Age                        0    -0.2505     0.3318
Sex                        M          M          M
ChestPainType            ASY        ASY        ASY
RestingBP                 -0    -0.1615     0.2139
Cholesterol               -0     0.134     -0.1776
FastingBS                 -0    -0.2951     0.3909
RestingECG            Normal     Normal     Normal
MaxHR                     -0     0.4077      -0.54
ExerciseAngina             N          N          Y
Oldpeak                   -0    -0.3795     0.5027
ST_Slope                Flat         Up       Flat
```

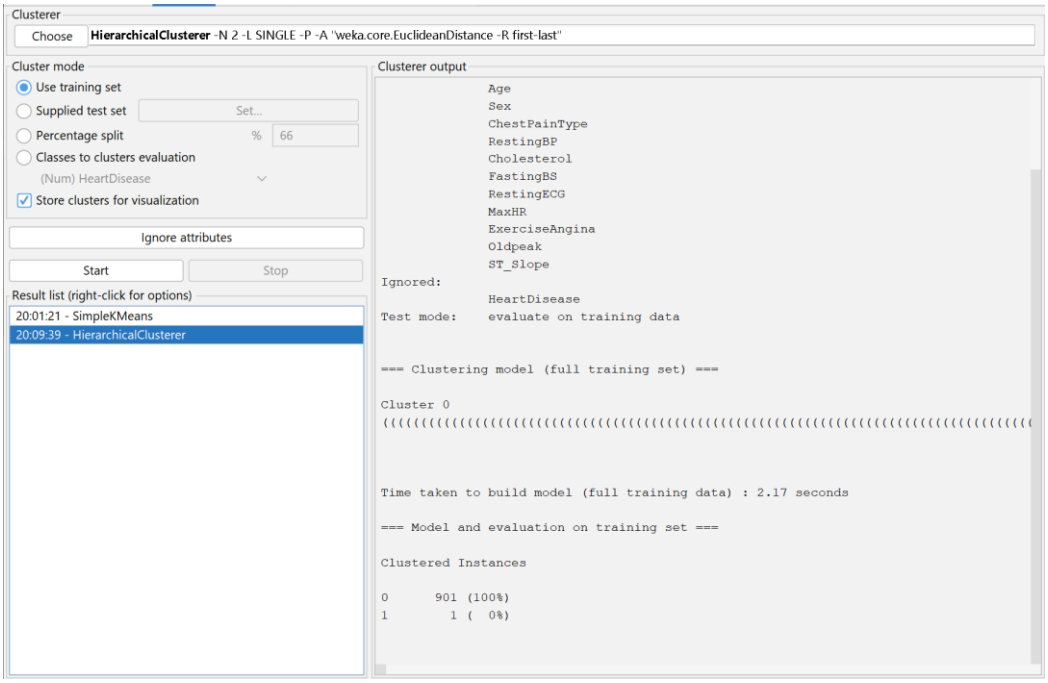
- Максимальна кількість ітерацій:

```
Number of iterations: 8
```

Результати:

Більшість значень кластеру 0 належать до здорових людей, а кластеру 1 – до хворих, проте все одно значна кількість людей неправильно розподілена.

2) Agglomerative Clustering (ієрархічна кластеризація):



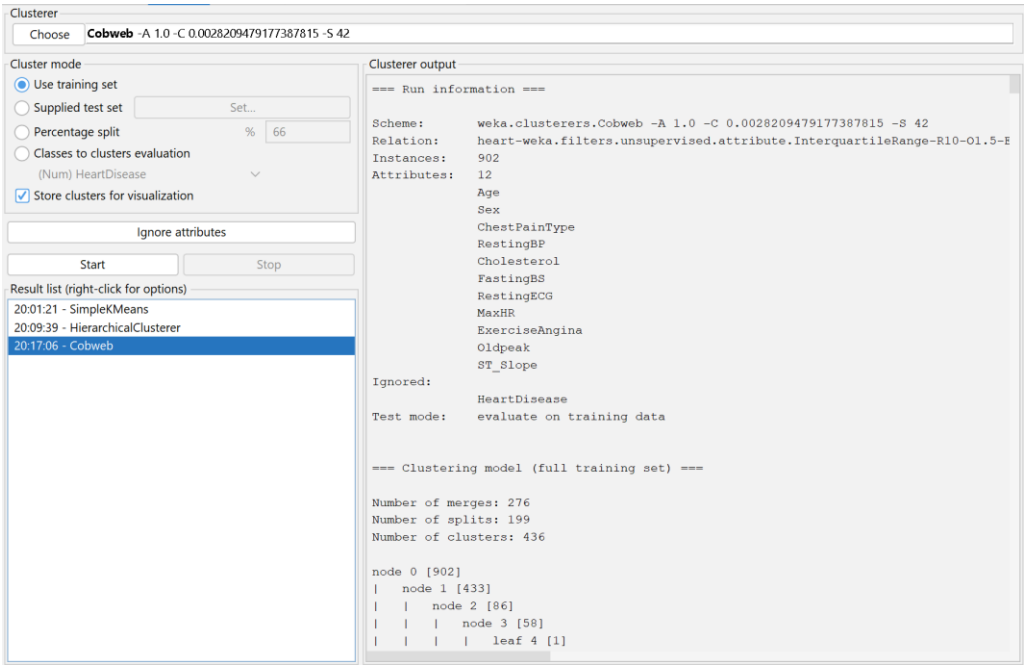
- Кількість кластерів:

```
0      901 (100%)
1        1 (  0%)
```

Результати:

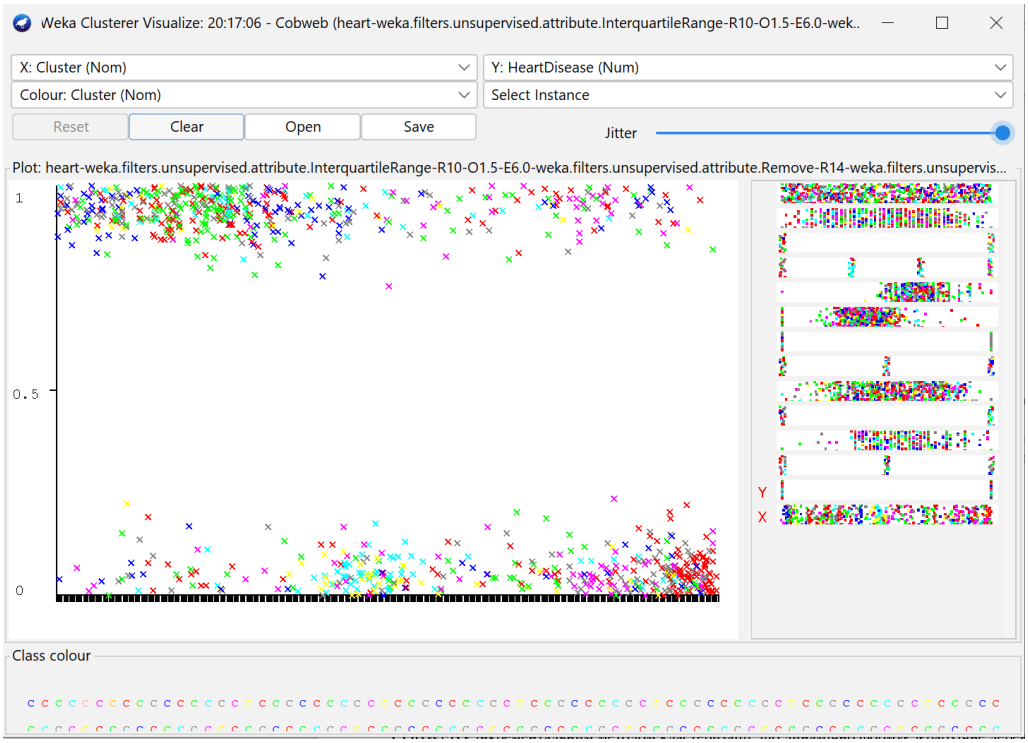
Як бачимо, ієрархічна кластеризація не підходить нашому алгоритму.

3) CobWeb




```
Number of merges: 276
Number of splits: 199
Number of clusters: 436
```

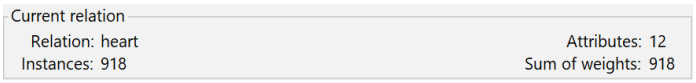
Результати:



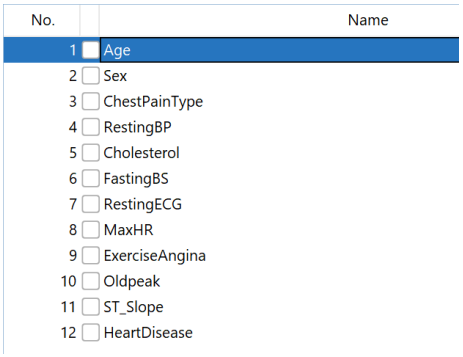
Наша модель складається із 436 кластерів, що свідчить про велику кількість чинників. Серед тих кластерів модна виокремити деякі, які більш правильно розроділені.

Короткий опис датасету

Практичним завданням для вирішення є знайти взаємозв'язок між медичними показниками пацієнта, щоб передбачити, чи можлива у нього серцева хвороба.



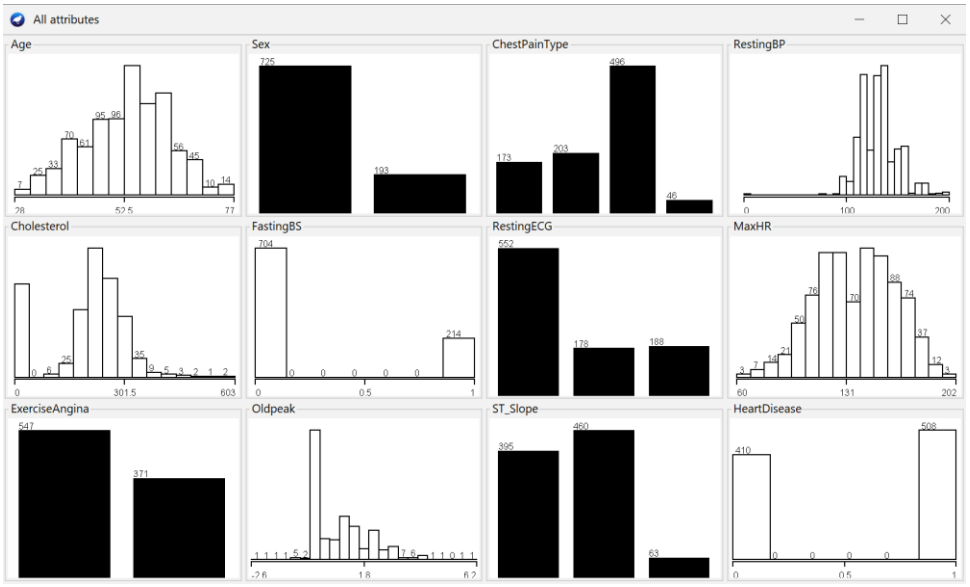
У вибірці 918 примірників.



Перелік атрибутів:

Age – вік
Sex – стать
ChestPainType – тип болю у грудях
RestingBP – артеріальний тиск у спокої
Cholesteron – сироватковий холестерин
FastingBs – рівень цукру в крові натще
RestingECG – результати електрокардіограми в спокої
MaxHR – досягнута максимальна частота серцевих скорочень
ExerciseAngina – стенокардія фізичного навантаження
Oldpeak – зниження сегмента ST на електрокардіограмі під час фізичного навантаження
ST_Slope – нахил піку навантаження на сегмент ST
HeartDisease – виявлена хвороба серця

Екземплярів із відсутніми значеннями немає.



Цільовим атрибутом є HeartDisease, який приймає значення 0 і 1 (нехворий і хворий на серце відповідно).

Клас хворих містить 410 екземплярів, а здорових – 506 екземплярів.

Було усунено викиди і стандартизовано всі атрибути.

Також для кластеризації був виключений цільовий атрибут HeartDisease.

У моєму випадку найкраще резульати показав K-means, потім CobWeb і найгірше справилась модель ієрархічної кластеризації.

Який метод підходить для яких даних?

Тип даних	Рекомендований метод
Великі набори з чіткими групами	K-Means
Невеликі або середні набори, важлива інтерпретація	Ієрархічна кластеризація
Дані, що змінюються у часі (онлайн-потоки)	COBWEB
Дані без явно виражених кластерів	Ієрархічна або density-based (DBSCAN)

Висновок: У ході лабораторної роботи було розглянуто та реалізовано три методи кластеризації: K-середніх (K-Means), ієрархічна кластеризація (Agglomerative Clustering) та COBWEB у середовищі Weka.

1. **Метод K-Means** дозволив ефективно розділити дані на кластери, при цьому добре спрацював метод «лікоть» (Elbow Method) для визначення оптимальної кількості кластерів. Виявлено, що при збільшенні K сумарна квадратична помилка (SSE) зменшується, проте після певного значення зниження стає менш значним.
2. **Ієрархічна кластеризація** допомогла візуалізувати взаємозв’язки між об'єктами за допомогою дендрограми, що дало змогу чіткіше зрозуміти природну структуру даних. Метод добре виділив групи на основі спільних характеристик, таких як вік, стать та фінансовий стан.
3. **Метод COBWEB** продемонстрував свою ефективність у випадках, коли необхідно автоматично визначити структуру кластерів, проте його результати менш інтерпретовані у порівнянні з іншими підходами.

Метод **K-Means** є ефективним для великих наборів даних, проте вимагає попереднього визначення кількості кластерів. **Ієрархічна кластеризація** краще підходить для аналізу малих та середніх вибірок, дозволяючи отримати глибше розуміння структури даних. **COBWEB** добре підходить для потокових даних, але іноді створює надто дрібні кластери, що ускладнює інтерпретацію. Важливим фактором для якості кластеризації є вибір параметрів та попередня обробка даних (нормалізація, видалення шуму тощо). Таким чином, використання різних методів кластеризації дозволяє отримати більш повне уявлення про структуру даних та знайти найкращий алгоритм для конкретної задачі.