МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ХАРЧОВИХ ТЕХНОЛОГІЙ

Кафедра інформаційних технологій, штучного інтелекту та кібербезпеки

ЛАБОРАТОРНА РОБОТА № 2

з дисципліни «Інтелектуальні системи підтримки прийняття рішень»

на тему: «Використання засобів Data Mining для прийняття рішень»

Виконав: Студент І курсу

групи КН-1-3М

Кучерявий М. В.

Перевірив:_____

Київ — 2025

Мета роботи

Набуття навичок розв'язання задач інтелектуального аналізу даних на прикладі проведення кластерного аналізу.

Хід виконання роботи

Імпортуємо данні з Anketa.csv. Створимо новий процес. Виділимо атрибути за якими будемо проводити кластеризацію:

• мета кредитування;

• сума кредиту;

• термін кредиту;

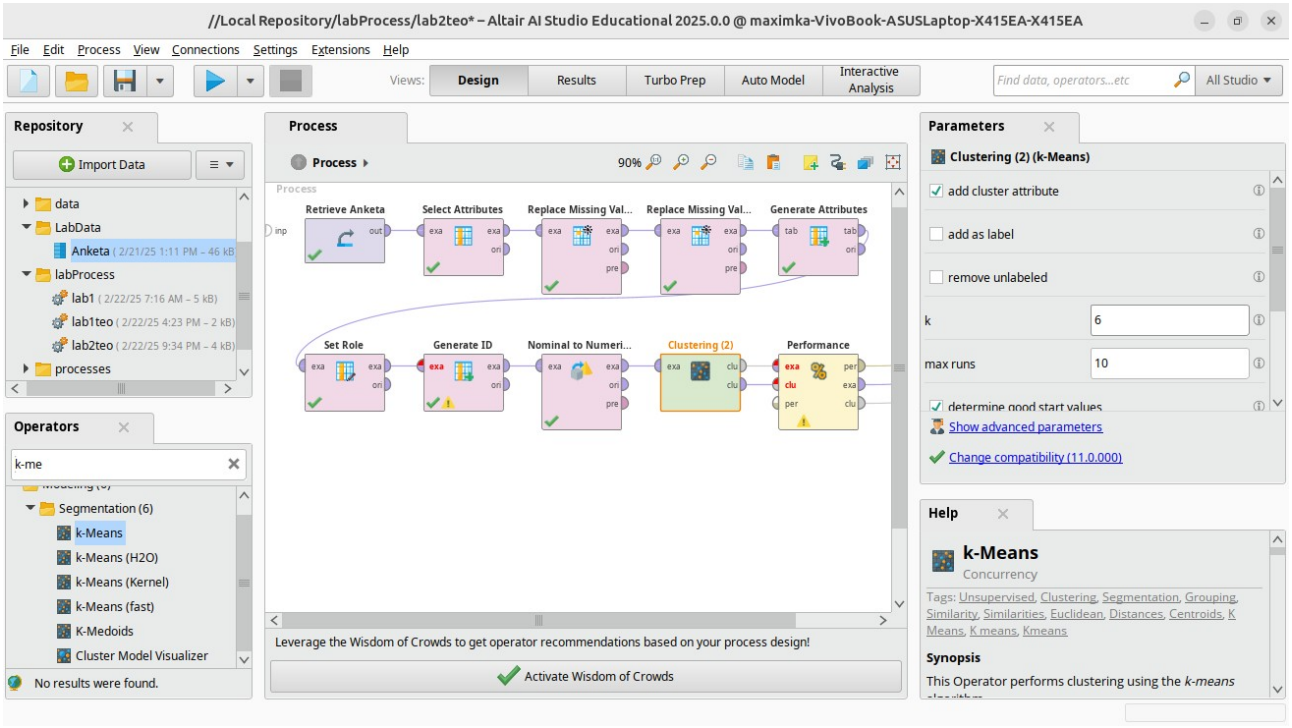
• вік;

• середньомісячний дохід;

• середньомісячні витрати;

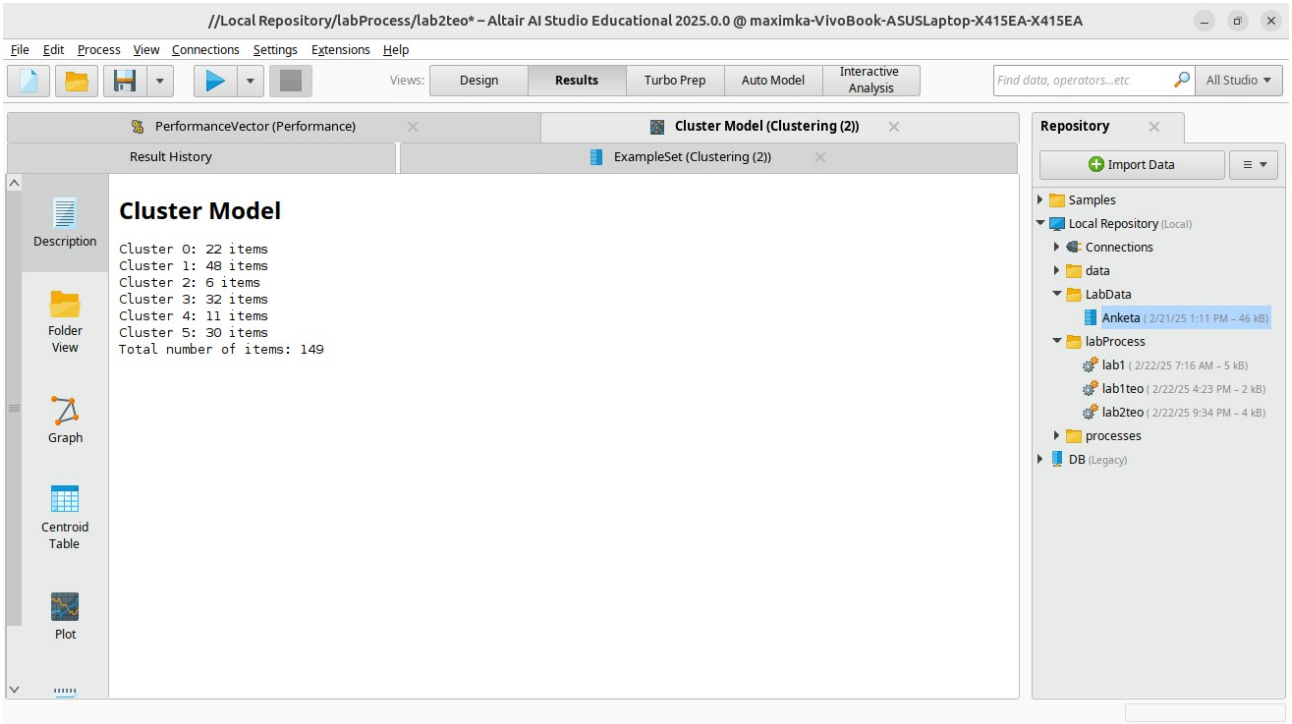
• кількість утриманців.

Заповнимо середніми значеннями пропущені елементи в цифрових полях, а в полі мета кредитування заповнимо пропущені значенням «Інше». Створимо додтковий атрибут «Номер» та перетворимо його на ідентифікатор. Перетворимо всі значення на числові оператором «Nominal to Numerical». Виконаємо кластеризацію за допомогою k-means та приєднаємо оператор аналізу кластеризації Cluster Dsitance Perfomance. Загальний вигляд процесу:



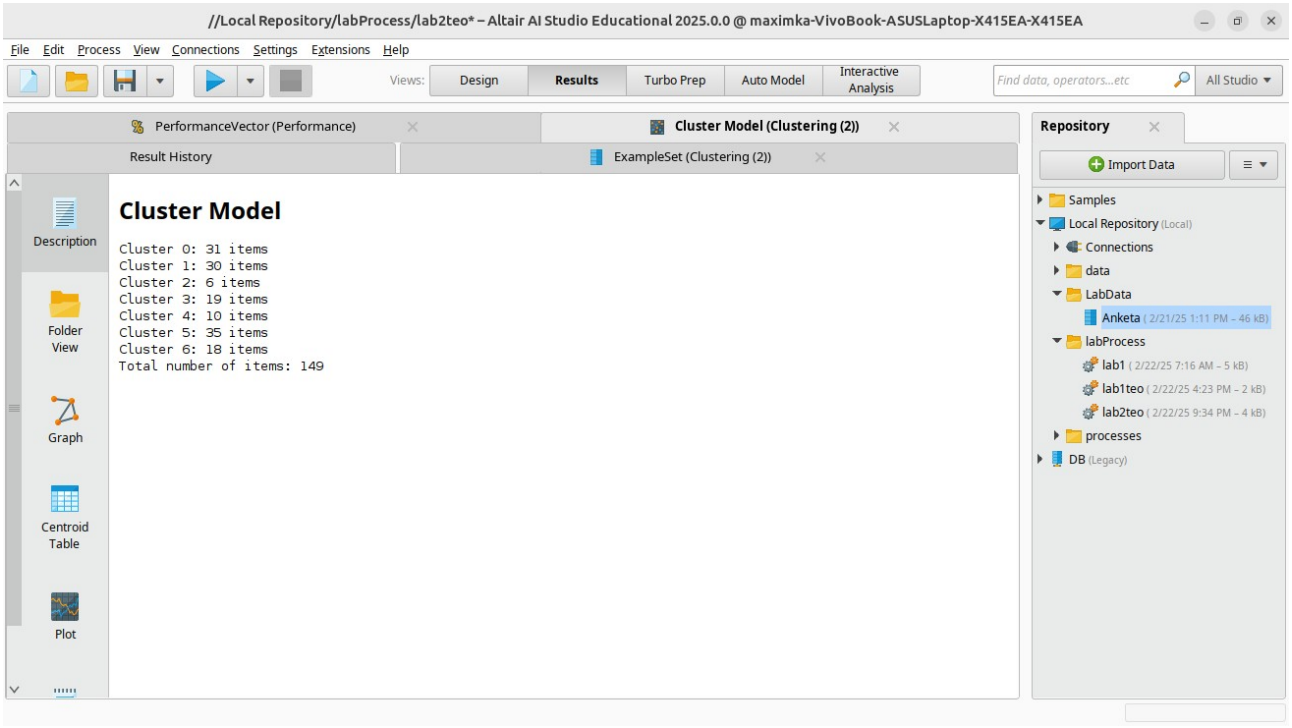
Виконаємо кластеризацію для 6 кластерів:

Розглянемо модель кластера:



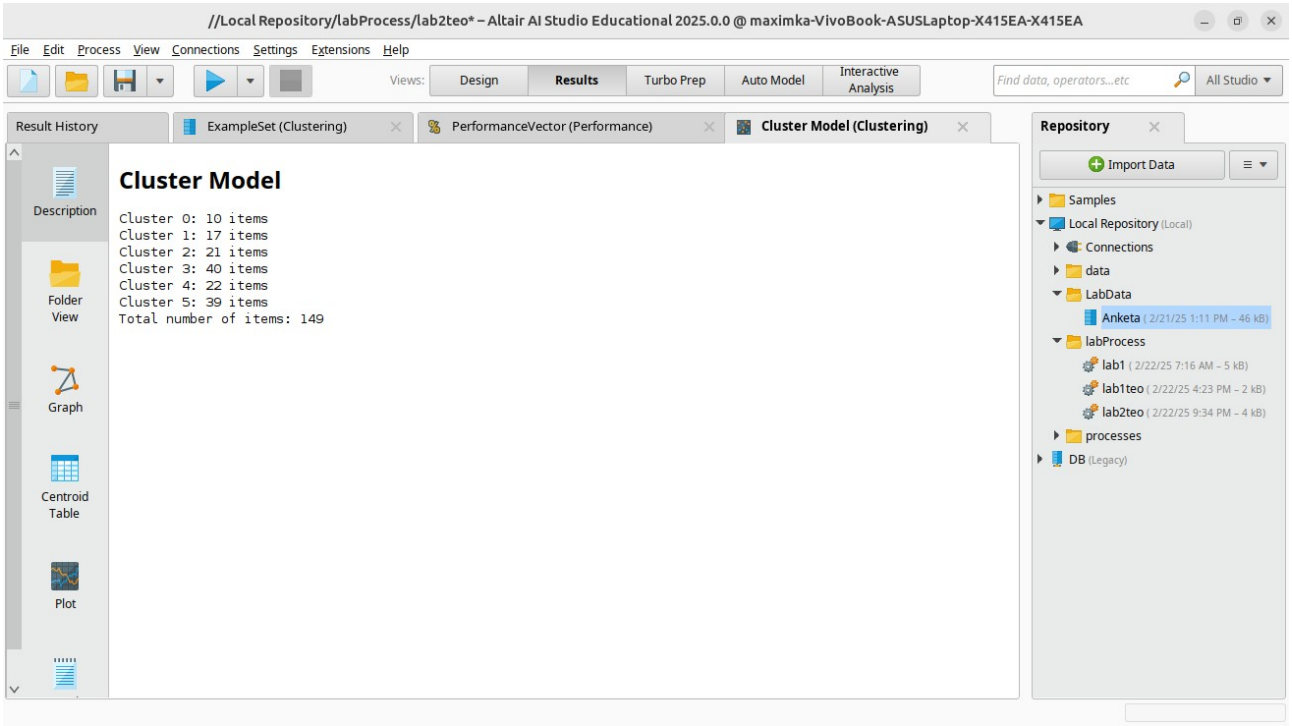
Індекс DBI є 0.46.

Спробуємо збільшити кількість кластерів до 7:



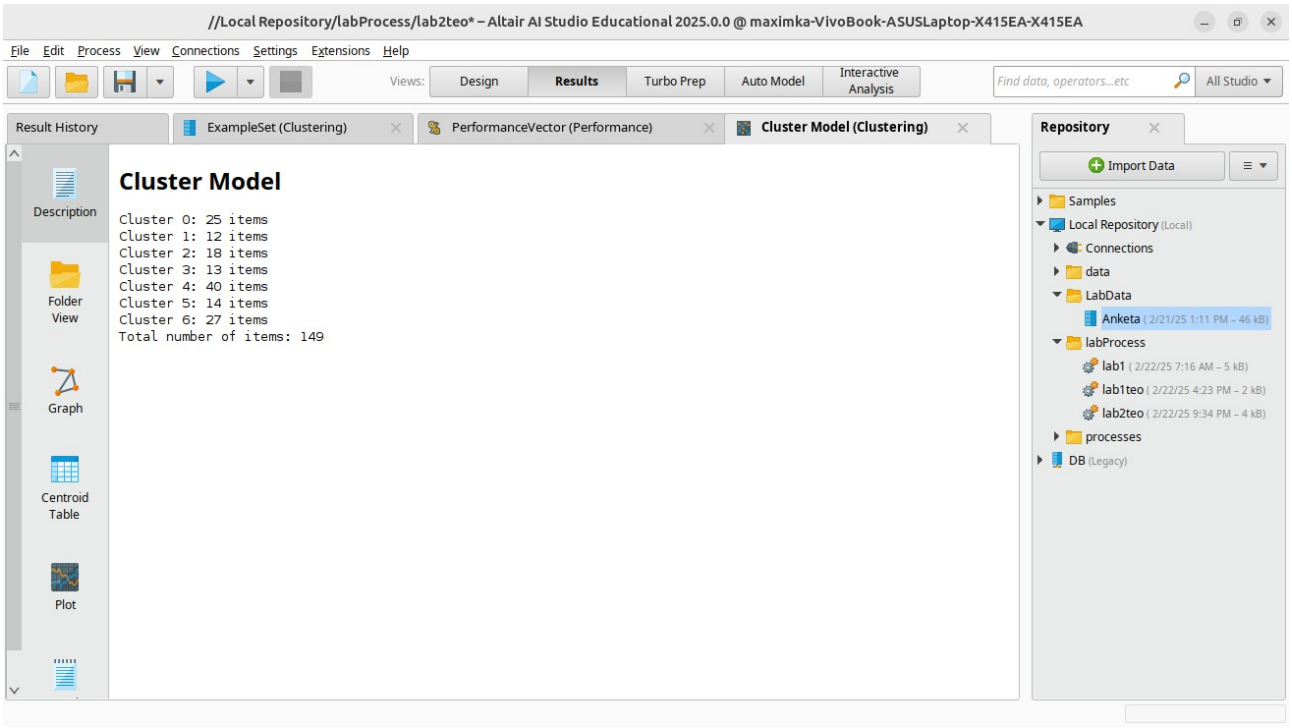
Індекс DBI є 0.53, що означає що 6 кластерів більш оптимальний варіант.

Спробуємо використати метод K-Medoids. Виконаємо кластеризацію для 6 кластерів:



DBI: 0.92

Для 7 кластерів:



DBI: 1.2

Що також дає більш оптимальне значення для 6 кластерів.

Звазаючи на те, що кожен кластер має досить високе значення середньої відстані від центроіда можна стверджувати, що данні не зовсім добре кластеризуються.

Відповіді на контрольні питання

1. Опишіть призначення кластерного аналізу. Чому кластерний аналіз називають методом “класифікації без навчання”?

Розбиття на групи дозволяє опрацьовувати більш детально кожну окрему групу даних, і використання їх в окремому аналізі. Кластеризацію відносять до методів “навчання без вчителя”, оскільки при формуванні кластерів не використовують “навчаючу множину”.

2. Назвіть основні групи методів кластерного аналізу та етапи кластеризації.

Основні методи кластерного аналізу: ієрархічні, центроїдні (, густинні , на основі моделей та графові. Етапи кластеризації: підготовка даних (очищення, нормалізація), вибір методу та параметрів, запуск алгоритму, оцінка та інтерпретація результатів.

3. Які є способи обчислення відстані між об’єктами?

Основні способи обчислення відстані між об'єктами: евклідова, манхеттенська, косинусна схожість та махаланобісова відстань, які використовуються залежно від типу даних і контексту задачі.

4. Які є критерії об’єднання у кластери?

Критерії об'єднання в кластери включають відстань між об'єктами, щільність, подібність ознак.

5. Що є задачами кластерного аналізу?

Розробка типології або класифікації, дослідження корисних концептуальних схем групування об’єктів, представлення гіпотез на основі дослідження даних, перевірка гіпотез про існування виділених деяким способом типів досліджуваних даних.

6. В чому полягає формальна постановка задачі кластеризації?

Необхідно знайти спосіб порівняння даних між собою, спосіб кластеризації, розподіл даних по кластерах.

7. В чому полягає міра схожості кластерів? Наведіть приклади.

Критерієм схожості об’єктів кластеризації є “відстань” між ними у просторі досліджуваних змінних.

8. Наведіть класифікацію алгоритмів кластеризації. В чому їх відмінність.

Алгоритми кластеризації можна класифікувати на центроїдні (k-means, k-medoids), ієрархічні (агломеративна, дивізивна), густинні (DBSCAN, OPTICS), моделі (GMM) та графові (Spectral Clustering). Їх відмінність полягає в підходах до утворення кластерів: деякі алгоритми базуються на фіксованому числі

кластерів, інші — на щільності або ієрархії, а деякі використовують ймовірнісні моделі чи графи для класифікації.

9. Дайте характеристику алгоритму K-means.

Алгоритм K-means є центроїдним методом кластеризації, який розбиває набір даних на k кластерів. Він працює за такими етапами: спочатку випадково вибираються k центроїдів, потім кожен об'єкт призначається найближчому центроїду, після чого центроїди оновлюються як середнє значення точок, що належать до кластеру, і цей процес повторюється до сходження.