Міністерство освіти і науки України

Львівський національний університет імені Івана Франка

Факультет електроніки та комп’ютерних технологій

Кафедра системного проектування

Звіт

Про виконання лабораторної роботи №7

З курсу «Системи машинного навчання»

### Кластеризація

**Виконала:**

Студентка групи ФЕС-32

Філь Дарина

**Перевірив:**

Доцент Колич І.І.

Львів 2024

**Мета:** ознайомитися з методами кластеризації та їх застосуванням.

**Теоретичні відомості**

**Основні Алгоритми Кластеризації**

1. **K-середніх (K-means):** Алгоритм, який розділяє дані на (k) кластерів, мінімізуючи суму квадратів відстаней від точок до центроїдів кластерів.
2. **Моделі суміші Гауса (Gaussian Mixture Models, GMM):** Алгоритм, який використовує комбінацію кількох гаусових розподілів для моделювання даних і знаходження кластерів.

**K-середніх (K-means)**

**Опис**: K-means є одним із найпопулярніших методів кластеризації. Він працює, розбиваючи набір даних на k кластерів та мінімізуючи суму квадратів відстаней між точками та центроїдами кластерів.

**Алгоритм:**

1. Вибрати k початкових центроїдів випадковим чином.
2. Призначити кожну точку до найближчого центроїда, утворюючи кластерів.
3. Обчислити нові центроїди для кожного кластера.
4. Повторювати кроки 2-3, доки центроїди не перестануть змінюватися.

**Переваги:**

* Простий у реалізації та швидкий.
* Ефективний для великих наборів даних.

**Недоліки:**

* Потрібно знати кількість кластерів заздалегідь.
* Чутливий до початкових умов.
* Працює краще на даних сферичної форми.

**Моделі суміші Гауса (GMM)**

**Опис:** GMM використовують комбінацію кількох гаусових розподілів для моделювання даних. Вони можуть моделювати кластери еліптичної форми, що робить їх більш гнучкими порівняно з K-means.

**Алгоритм:**

1. Ініціалізація параметрів (середні, ковариаційні матриці, ваги компонентів).
2. Крок E (Expectation): Обчислення ймовірностей приналежності кожної точки до кожного кластеру на основі поточних параметрів.
3. Крок M (Maximization): Оновлення параметрів, максимізуючи правдоподібність даних на основі ймовірностей, отриманих на кроці E.
4. Повторювати кроки E та M, доки зміни параметрів не стануть незначними.

**Переваги:**

* Може моделювати кластери еліптичної форми.
* Менш чутливий до початкових умов порівняно з K-means.

**Недоліки:**

* Потрібно знати кількість кластерів заздалегідь.
* Може бути складнішим у реалізації та навчанні.

**Оцінка Кластеризації**

**Силуетний коефіцієнт (Silhouette Coefficient)**

**Опис:** Силуетний коефіцієнт оцінює якість кластеризації, порівнюючи середню відстань від точки до інших точок у її кластері із середньою відстанню до точок у найближчому сусідньому кластері.

**Формула:**

де:

* середня відстань від точки і до інших точок у її кластері;
* середня відстань від точки і до точок у найближчому сусідньому кластері;

Інтерпретація:

* Точка добре кластеризована;
* Точка знаходиться на межі між двома кластерами;
* Точка, можливо, була неправильно кластеризована;

Індекс Девіса-Боулдіна (Davies-Bouldin Index)

**Опис:** Індекс Девіса-Боулдіна оцінює середню схожість між кожним кластером і кластером, який найбільш схожий на нього.

**Формула:**

де:

* середнє відхилення точок у кластері і до цетроїда ;
* відстань між центроїдами кластерів і та j;

Інтерпретація:

* нижчі значення (DB) вказують на кращу кластеризацію;

**Хід роботи**

**Завдання**

1. **Підготовка даних** 
   1. Використайте набір даних Wine.
   2. Розділіть дані на ознаки (features) та мітки (labels).
   3. Нормалізуйте дані для покращення продуктивності моделей кластеризації.
2. **K-середніх** 
   1. Створіть та навчіть модель K-середніх на даних.
   2. Виконайте прогнозування кластерів для даних.
   3. Визначте оптимальне значення k за допомогою методу "ліктя" " (Elbow Method, https://en.wikipedia.org/wiki/Elbow\_method\_(clustering))
3. **Моделі суміші Гауса (GMM)** 
   1. Створіть та навчіть модель GMM на даних.
   2. Виконайте прогнозування кластерів для даних.
4. **Оцінка Моделей** 
   1. Оцініть моделі за допомогою силуетного коефіцієнта та індексу ДевісаБоулдіна.
   2. Візуалізуйте результати кластеризації.
5. **Оформлення звіту** 
   1. Оформіть звіт з результатами лабораторної роботи, включаючи графіки, та аналіз результатів.

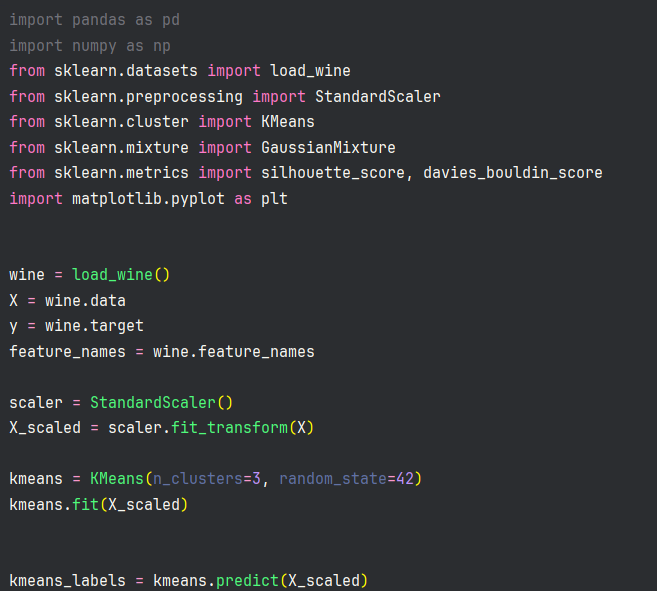


Рис.1 Ініціалізація бібліотек, набору даних та алгоритму k-means

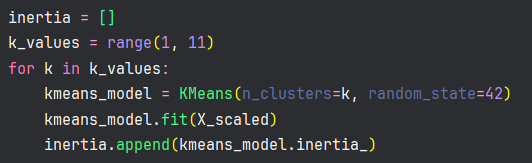


Рис.2 Метод ліктя

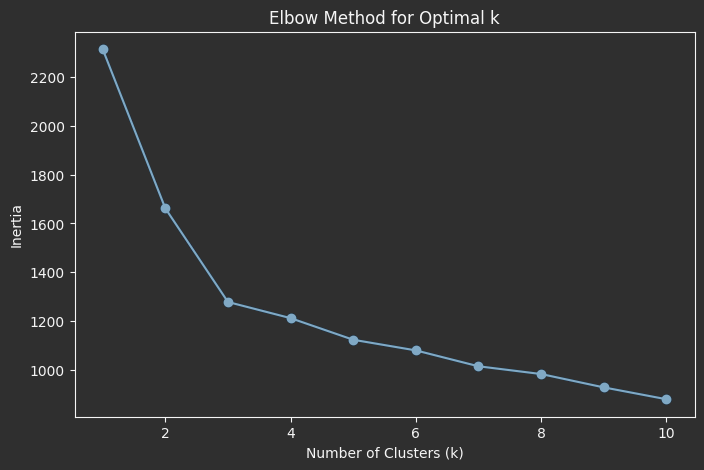


Рис.3 Оптимальне значення k, визначене за допомогою методу ліктя, який аналізує інерцію відносно кількості кластерів

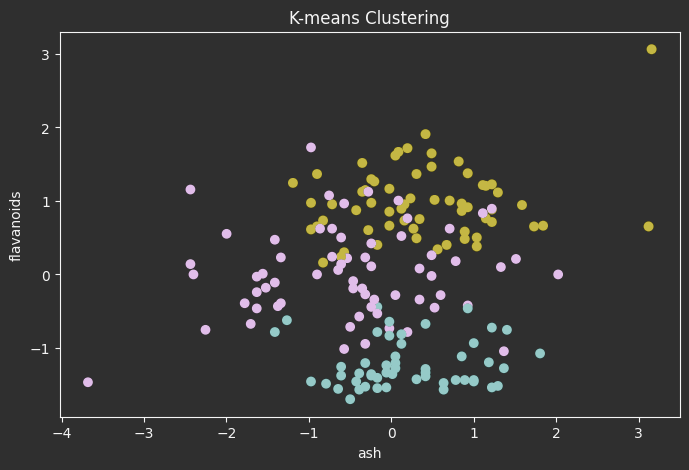


Рис.4 Результат кластеризації для алгоритму k-means

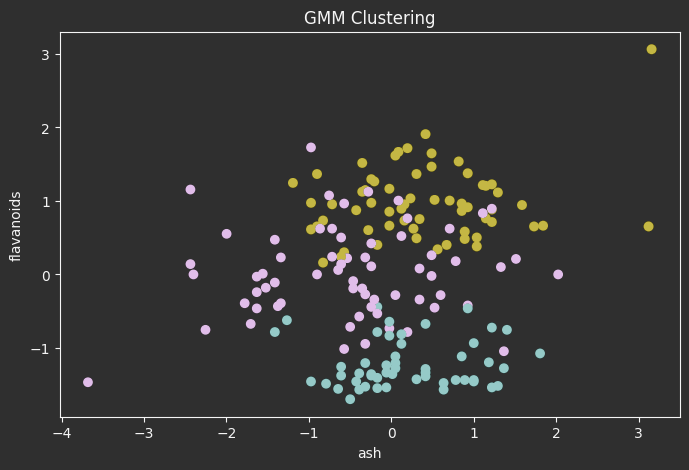


Рис.5 Результат кластеризації для алгоритму GMM

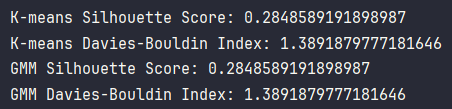


Рис.6 Метрики оцінювання показують, що значення силуету становить 0.28, що є невдалим результатом, оскільки метрика силуету може варіюватися від -1 до 1, де 1 є найкращим показником. Індекс Девіса коливається від 0 до нескінченності; у нашому випадку значення 1 вважається прийнятним, але не ідеальним

Я спробувала покращити результат за допомогою методу головних компонент і отримала наступні результати.

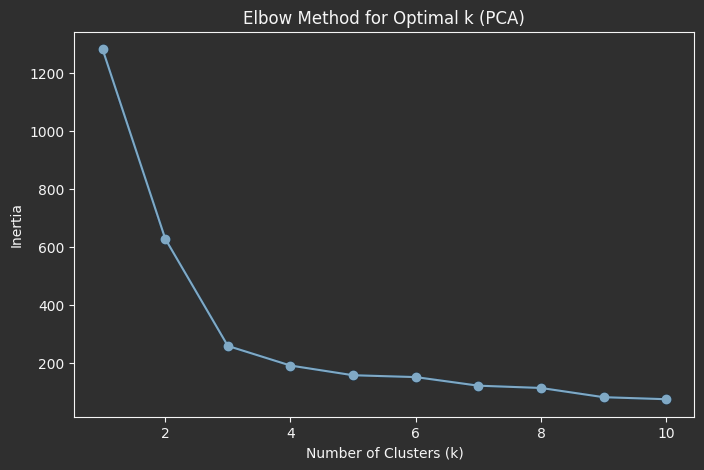


Рис.7 Оптимальне значення k, використовуючи метод ліктя

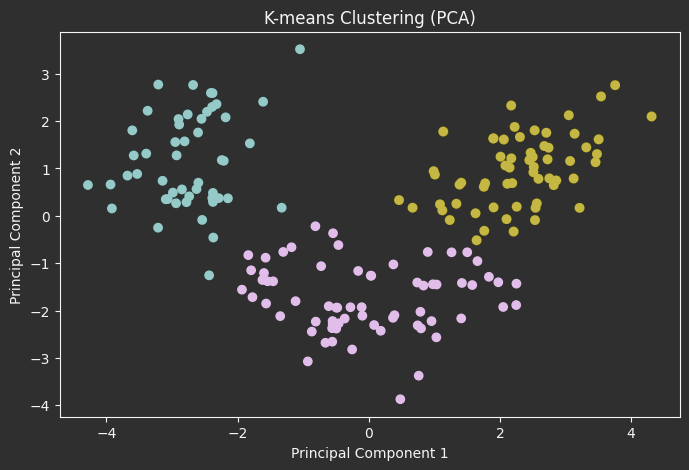


Рис.8 Результат кластеризації алгоритмом k-means після застосування методу головних компонент показує, що кластери не перетинаються і чітко виділяються

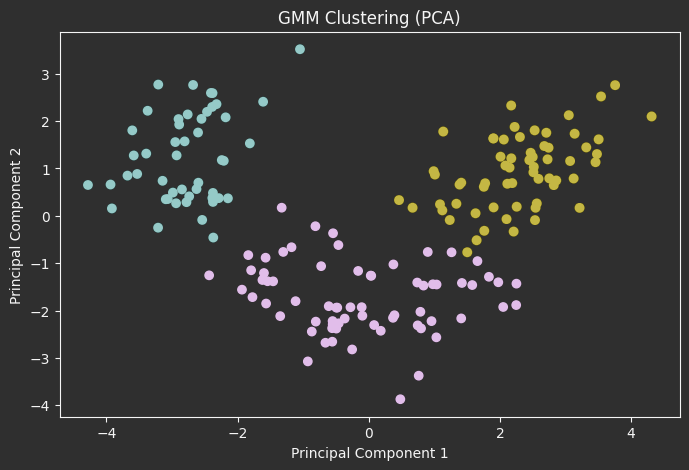


Рис.9 Результат кластеризації алгоритмом GMM після застосування методу головних компонент демонструє, що кластери не перетинаються і чітко видимі

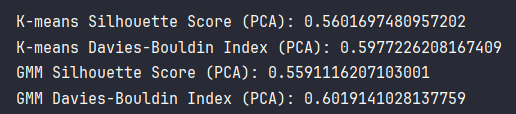


Рис.10 Ми спостерігаємо, що отримані метрики все ще не ідеальні, але значно покращилися після застосування методу PCA; результат кластеризації став майже вдвічі кращим

**Висновок:** У цій лабораторній роботі я навчилася використовувати на практиці два методи кластеризації: k-means та GMM. Для визначення кількості кластерів для алгоритму k-means був застосований метод ліктя. Також я використала метод PCA для покращення результатів кластеризації.