|  |  |  |  |
| --- | --- | --- | --- |
| МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ  КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ  імені Тараса Шевченка  ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ  **Кафедра програмних систем і технологій**  Дисципліна  **«Компʼютерне моделювання процесів»**  **Лабораторна робота № 5** | | | |
| **Виконав:** | Шкітак Нікіта | **Перевірила**: | Ніколаєнко Анастасія Юріївна |
| Група | ІПЗ-31-1 | Дата перевірки |  |
| Форма навчання | денна | Оцінка |  |
| Спеціальність | 121 |
| 2023 | | | |

# Завдання

# 1. Підготуйте набір даних для кластеризації (за варіантом).

# 2. Проведіть кластеризацію даних за алгоритмом k-means із застосуванням бібліотеки Scikit-Learn. Кількість кластерів виберіть таку ж, як і кількість класів у попередній лабораторній. (2 бали)

# 3. Побудуйте графіки результатів кластеризації для декількох параметрів попарно (не менше 4-х), додайте на графіки центроїди. (3 бали)

# 4. Порахуйте кількість екземплярів у кожному кластері. Проаналізуйте, скільки об’єктів кожного з класів потрапило у кожний кластер. (3 бал)

# 5. Визначте оптимальну кількість кластерів за декількома метриками. Виведіть таблицю та побудуйте графіки залежності показника якості кластеризації (метрики) від кількості кластерів. Проаналізуйте отримані результати. (4 бали)

# 6. Дослідіть вплив масштабування/нормалізації ознак на результат кластеризації та на оптимальну кількість кластерів. (4 бали)

# 7. Оформіть звіт і захистіть роботу на практичному занятті.

# Виконання завдання

## Вибірка

Обрана набір даних "Wine" має 14 колонок. Перша колонка містить номер сорту вина від 1 до 3, за яким проводиться класифікація. Інші 13 колонок містять різні характеристики вина, такі як вміст алкоголю, яблучної кислоти, феноли і т.д. Загалом в наборі даних міститься 178 записів, з яких 59 відносяться до вина типу 1, 71 - до типу 2 та 48 записів відносяться до типу 3. Цю інформацію можна переглянути у консолі під час обробки цього набору даних:

Також варто зауважити що в таблиці відсутні як дублікати так і порожніх значень немає. Відповідно можна зробити висновок що таблиця є повністю заповнена.

## Кластеризація

Відповідно для порівняння впливу масштабування, ми будемо проводити кластеризацію при масштабуванні та без. Як можемо побачити з скріншотів, тільки розподіл об’єктів на кластери трохи відрізняється на даному етапі:

До масштабування:  
A screenshot of a computer screen

Description automatically generated

Після масштабування:  
A screenshot of a computer screen

Description automatically generated

## Графіки кластеризації

Перш за все потрібно знайти центроїди кластерів. Після масштабування ми можемо побачити, що значення центроїдів лежить в однмоу діапазоні : від 0 до 1:

До масштабування:  
A black screen with white numbers

Description automatically generated

Після масштабування:  
A black background with white numbers

Description automatically generated

Для прикладу утворених проекцій даних було обрано такі пари колонок: **Malic.acid** and **Acl, Acl** and **Phenols, Alcohol** and **Ash, Mg** and **Flavanoids:**

## Оцінка класифікатора

A screenshot of a computer screen

Description automatically generated

За даними звіту про оцінки класифікатора можна зробити висновок, що клас вина №1 визначається найбільш точно, в той час як клас 3, який до того ж містить найменше записів – визначається менш точно.

До масштабування:

A diagram of a task without scaling

Description automatically generated

Після масштабування

A diagram of a task

Description automatically generated

До масштабування:A diagram of a task

Description automatically generated

Після масштабування:

A diagram of a task

Description automatically generated

До масштабування:

A diagram of a task without scaling

Description automatically generated

Після масштабування:  
A diagram of a task

Description automatically generated

До масштабування:

A diagram of a task

Description automatically generated

Після масштабування:

A diagram of a task

Description automatically generated

Як ми можемо побачити, масштабовані дані є більш згруповані та зменшують кількість аномальних ситуацій елементів кластеру. Відповідно в результаті збільшується коректність розподілу даних по кластерах, так як зменшується кількість шуму.

## Кількість екземплярів.

Розподіл елементів по кластерах до масштабування:

A black and white screen with numbers and a black background

Description automatically generated

Розподіл елементів по кластерах після масштабування:  
A screenshot of a black and white screen

Description automatically generated

Аналіз результатів перед та після масштабування показує, що при використанні масштабованих даних ймовірність того, що об'єкти одного класу вин потраплять до одного кластера, збільшується. Розподіл записів стає більш рівномірним, і спостерігається менш виразна нерівність в заповненості кластерів, що відрізняється від ситуації до масштабування.

## Оптимальна оцінка кластерів

Варто зазначити: що аналіз буде відбуватись за даними **після масштабування**

Аналіз оптимальної кількості до масштабування:

A screenshot of a computer

Description automatically generated

Аналіз оптимальної кількості після масштабування:

A screenshot of a computer screen

Description automatically generated

Метод ліктя до масштабування:

A graph with blue dots

Description automatically generated

Метод ліктя після масштабування:  
A graph with blue dots

Description automatically generated

З малюнку видно, що різкий спад суми квадратів відстаней від екземплярів до найближчого центроїда (WCSS) відбувається вже між точками, які відповідають двом і трьом кластерам. Але не зовсім зрозуміло, яка з точок на цьому графіку є «згином ліктя». Однієї оцінки недостатньо, щоб визначити оптимальну кількість кластерів.

Метод силуету до масштабування

A graph with blue dots

Description automatically generated

Метод силуету після масштабування:

A graph with blue dots

Description automatically generated

Чим ближче значення силуетної оцінки до 1, тим краще. Як видно з малюнку найбільше значення у точці 2.

Метод Девіса-Булдіна до масштабування:

A graph with blue dots

Description automatically generated

Метод Девіса-Булдіна після масштабування :A graph with blue dots

Description automatically generated

Чим нижче значення цього індексу, тим краще. Як видно з малюнку найнижче значення у точці 3. Відповідно враховуючи всі вище враховані оцінки, можна зробити висновок: що найоптимальніша кількість кластерів є 3, що дорівнює кількості класів в заданому датасеті.

Даний результат є досить логічним та правильним, оскільки кожен клас має свої унікальні властивості, за якими потрібно створювати відповідно окремі кластери.

## Висновок

Під час виконання даної лабораторної роботи було проведено кластеризацію великої вибірки даних, за допомогою алгоритму k-means, використовуючи бібліотеку Scikit-Learn. Також побудова графік розподілення кластерів допомогла візуалізувати розміщення даних у просторі.

Було побудовано алгоритм, для знаходження оптимальної кількіоості кластерів. В результаті, було отримано оптимальну кількість кластерів – 3, що є рівною кількості класів в заданому датасеті. Аналіз було проведено за методами : ліктя, силуету та Девіса-Булдіна.

Також було досліджено вплив масштабування ознак на результат кластеризації та оптимальну кількість кластерів. Виявлено, що правильне масштабування може покращити якість кластеризації, та забезпечити більш точне визначення груп обʼєктів. Однієюю з причин є те, що масштабування зменшує кількість шумів, що негативно впливають на результат кластеризації.

Відповідно, результатом виконаної роботи є працююча модель кластеризації виборок даних, що може бути легко видозмінена під інший датасет.