|  |  |  |  |
| --- | --- | --- | --- |
| 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. Проаналізуйте та підготуйте набір даних до класифікації (за варіантом).

Виведіть статистичну інформацію числових даних з DataFrame (метод

describe бібліотеки Pandas). Скільки об’єктів (записів) всього у вибірці?

Скільки ознак можна вибрати для класифікації і що це за ознаки? Чи були

пропущені значення? Чи були дублікати записів? Скільки екземплярів у

кожному класі? (1 бал)

2. Візуалізуйте матрицю графіків розсіювання. Проаналізуйте зв’язок між

ознаками. (1 бал)

3. Розділіть дані з файлу на тренувальний (навчальний) та тестовий набори.

Можна 80% віднести до тренувального (навчального) набору, а 20% до

тестового. Розробіть класифікатор за варіантом із застосуванням Scikit-

Learn. Навчіть класифікатор на тренувальному наборі даних. Застосуйте

класифікатор до тестового набору даних. (2 бали)

4. Оцініть класифікатор (побудуйте звіт щодо продуктивності

класифікатора). (1 бал)

5. Побудуйте матрицю неточностей. Проаналізуйте її. (2 бали)

6. Виконайте масштабування (нормалізацію) функції та проаналізуйте, чи

вплинуло це на продуктивність класифікатора. (2 бали)

7. Підберіть параметри класифікатора, за якого найбільша частка

правильних передбачень серед тестових даних. (3 бали)

8. Проаналізуйте, що і як саме впливає на точність класифікатора, який

навчено саме на вашому наборі даних і зробіть висновки. (1 бал)

9. Напишіть власну функцію, яка розраховує евклідову відстань між двома

об’єктами. Виберіть по три об’єкти з кожного класу тестового набору.

Для кожного з цих об’єктів визначте найближчого сусіда з навчального

набору використовуючи власну функцію розрахунку евклідової відстані.

Порівняйте, чи до одного й того ж класу відносяться вибрані об’єкти і їх

найближчі сусіди. (3 бали)

10. 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:

A screenshot of a computer

Description automatically generated

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

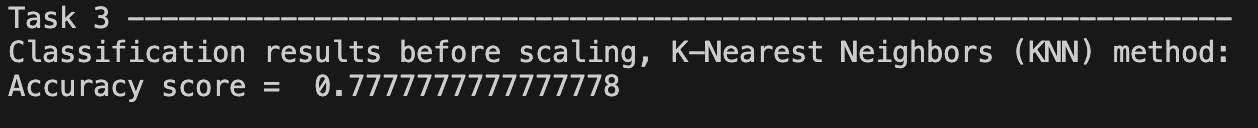
## A grid of colorful dots Description automatically generated with medium confidenceМатриця графіків розсіювання

Загалом, графіки розсіювання показують досить щільнe розташування даних, де значення знаходяться в близькості одне до одного. Однак у багатьох випадках можна помітити певні закономірності та чіткіше виокремлені групи даних, що належать до окремих класів.

## Поділ датасету та навчання класифікатора

Відповідно було поділено датасет у на 80% тренувальних даних та 20% тестових даних для тестування робити класифікатора.

Також варто зазначити що для даного завдання було обрано класифікацію за найближчими сусідами, метрика була обрана – евкліда.



Відповідно було отримано таку точність за використання метода найближчих сусідів.

## 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.

## Матриця неточностей

A screenshot of a computer screen

Description automatically generated

Вище показана матриця неточності показує відмінності між фактичними значеннями записів у вибірці та тими, які передбачив класифікатор. На малюнку видно таку ситуацію: з першого класу було правильно ідентифіковано 9 об'єктів, проте 3 записи було помилково віднесено до класу 3; у другому класі було точно визначено 9 об'єктів, але ще 5 записів було помилково віднесено до класу 3; третій клас має 5 правильно визначених об'єктів, але 3 (майже половина) помилково призначено до класу 2 і 1 до класу 1.

З цієї інформації можна зробити висновок, що класи 2 і 3 мають схожі властивості, що може створювати проблеми для класифікатора при виборі правильного класу для об'єкта.

## Масштабування

Для масштабування функції використовувалась StandardScaler з Sklearn. Масштабування виконується безпосередньо для приведення даних до єдиного однорідного діапазону значень всіх ознак.

Процес масштабування:

A black background with blue and yellow text

Description automatically generated

Після виконання масштабування, ми можемо побачити поліпшення показників класифікатора:

A screen shot of a computer

Description automatically generated

Показники стали максимально допустимим, тобто відсоток правильно визначених класів дорівнює 100%.

Відповідно також можемо побачити це на матриці неточностей:  
A screenshot of a graph

Description automatically generated

Як ми можемо побачити, всі обʼєкти були визначені відносно певного класу.

## Параметри класифікатора

Під час виконання завдань, було визначено, що найбільшу кількість правильних відповідей класифікатор першого набору даних визначає при обрані параметрів кількості сусідів 7. Також для прикладу можна взяти і порівняти вибірку з кількість сусідів що дорівнює 2 та порівняти з 7:

A colorful squares with numbers and labels

Description automatically generatedA colorful squares with numbers and labels

Description automatically generated

Рисунок 1 Кількість сусідів - 2

Рисунок 2 Кількість сусідів - 7

A screenshot of a computer screen

Description automatically generated

Рисунок 3 Кількість сусідів - 7

A screenshot of a computer screen

Description automatically generated

Рисунок 4 Кількість сусідів - 2

Як ми можемо побачити, хоча не дуже суттєво, але точність є більшою у випадку коли, кількість сусідів дорівнює 7.

## Власні функції

Ефективність даного методу значно залежить від кількох факторів, одним з яких є вплив масштабування даних. Для забезпечення однакового діапазону даних використовувалося масштабування, що дозволило підвищити точність класифікації. Однак існує додаткова проблема: метод найближчих сусідів досить чутливий до шуму. Одне викидне (аномальне) значення у вибірці може значно вплинути на класифікацію об'єкта.

## Аналіз що впливає на точність

Результат визначення найближчого сусіда з навчального набору з функцією розрахунку евклідової відстані та порівняння, чи до одного й того ж класу відносяться вибрані об’єкти і їх найближчі сусіди.

A screenshot of a computer

Description automatically generated

## Висновок

Під час лабораторної роботи було проведено дослідження методу класифікації наборів даних з використанням бібліотеки SkLearn та бібліотеки Pandas. Основним методом дослідження обрали метод найближчих сусідів, за яким передбачається клас нового об'єкту на основі оцінки найближчих сусідів з тренувального набору.

Результатом роботи стали добре налаштований класифікатор для записів характеристик різних класів вин та створений класифікатор, який використовує евклідову метрику та метод найближчого сусіда для визначення класу нового запису. Проте обидва результати можуть бути покращені у майбутньому: налаштований класифікатор може бути покращений шляхом збільшення навчальної інформації та спроби використання інших методів класифікації; самостійно створений класифікатор може бути удосконалений за допомогою розширення кількості найближчих значень для уникнення впливу аномальних даних.