Міністерство освіти та науки України

Національний аерокосмічний університет

«Харківський авіаційний інститут»

Лабораторна робота

з дисципліни «Обробка та аналіз результатів НД з використанням ІТ»

на тему: «Машинне навчання, обробка великих масивів даних»

|  |  |
| --- | --- |
|  | Виконав  студент групи G5-504-1 Пліско О.О |
|  |  |
|  | Перевірив  доцент Крицький Д.М. |

м. Харків 2025

Лабораторна робота №1

Налаштування робочого місця і знайомство з Python

Виконання завдань

1. Обчисліть площу трикутника з основою 120 одиниць (B) і висотою 33 (H). Призначте результат змінній з назвою області та надрукуйте його.

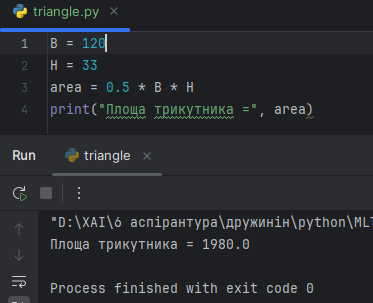


Рисунок 1.1 – Площа трикутника

2. Обчисліть площу квадрата зі стороною (S), що дорівнює 145 одиницям. Призначте результат змінній з назвою області та надрукуйте його.

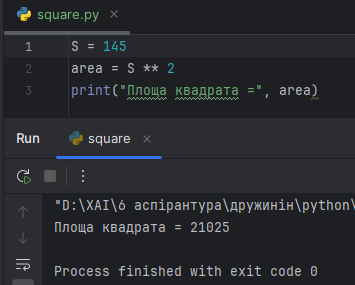


Рисунок 1.2 – Площа квадрата

3. Змініть програму в пункті 2, щоб отримати сторону S від користувача під час виконання.



Рисунок 1.3 – Площа квадрата

4. Якщо ви ще цього не зробили, помістіть два попередні сценарії в два окремі файли.



Рисунок 1.4 – Окремі файли завдань

5. Напишіть невеликий сценарій (trapezoid.py), який обчислює площу трапеції, яка має велику основу (MB), що дорівнює 30 одиницям, меншу основу (mb), дорівнює 12, і висоту (H) дорівнює 17. Надрукуйте отриману площу. Спробуйте виконати сценарій із коду Visual Studio та з терміналу.

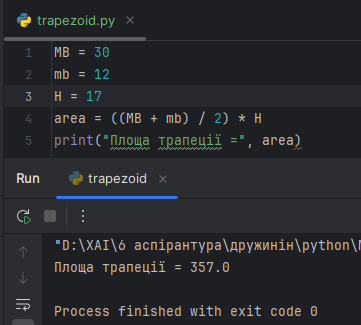


Рисунок 1.5 – Виконання в Pycharm

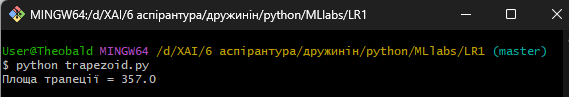


Рисунок 1.6 – Виконання з терміналу

6. Перепишіть приклад суми перших 1200 цілих чисел, використовуючи таке рівняння:

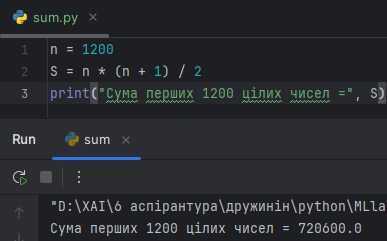


Рисунок 1.7 – Сума перших 1200 цілих чисел

7. Змініть програму в пункті 6, щоб вона отримувала кількість цілих чисел для суми N від користувача під час виконання.

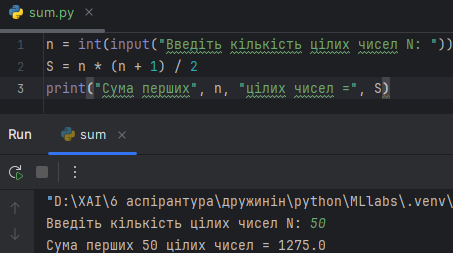


Рисунок 1.8 – Виконання від користувача

8. Напишіть невеликий сценарій для обчислення довжини гіпотенузи (c) прямокутного трикутника зі сторонами a=133 і b=72 одиниці.

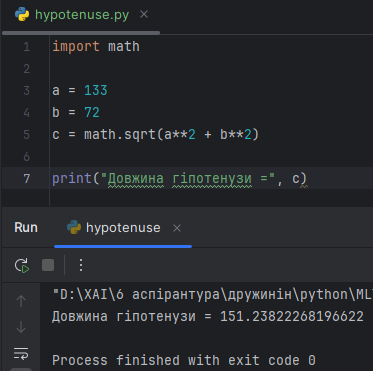


Рисунок 1.9 – Обчислення довжини гіпотенузи

9. Перепишіть сценарій трапеції, щоб він обчислив площу трапеції, починаючи з великої основи (MB), малої основи (mb) і висоти (H), взятих у вхідних даних.

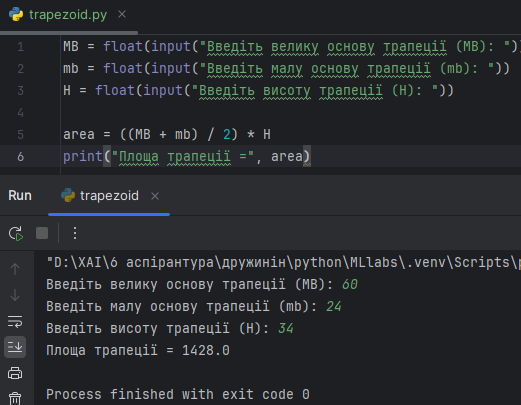


Рисунок 1.10 – Обчислення площі трапеції

10. Напишіть сценарій, який зчитує сторону геакута у вхідних даних і обчислює його периметр і площу, друкуючи їх на екрані.

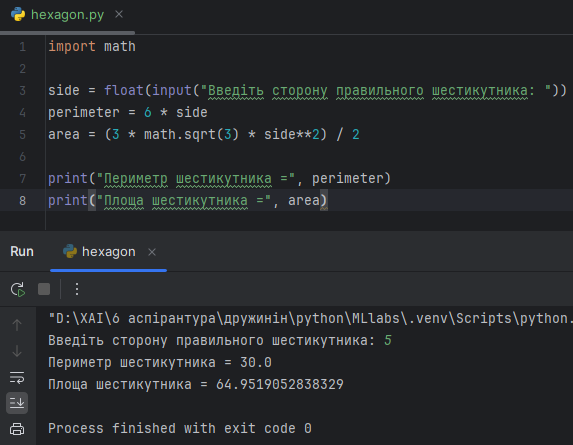


Рисунок 1.11 – Обчислення периметра і площі шестикутника

Лабораторна робота №2

Модулі Python. Об’єкт. Змінні.

Виконання завдань

1. Дано два рядка:

text = """Nobody said it was easy

No one ever said it would be this hard"""

Напишіть деякий код Python, який а) друкує весь рядок; б) друкує перший і останній символ; в) друкує перші 10 символів; г) відбитки з 19-го знака до 31-го; e) друкує весь рядок великими літерами.

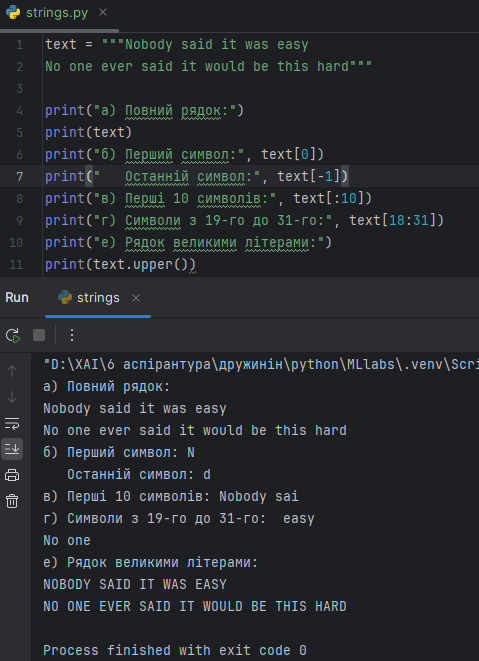


Рисунок 2.1 – Виконання першого завдання

2. Екзон гена починається з позиції 12030 на геномі і закінчується в позиції 12174. Чи впливає на цей екзон A/T SNP, присутній у позиції 12111 (тобто знаходиться всередині екзона)? А як щодо SNP, присутній на позиції 12188? Підказка: перевірте відповідний логічний вираз, щоб відповісти на це запитання.

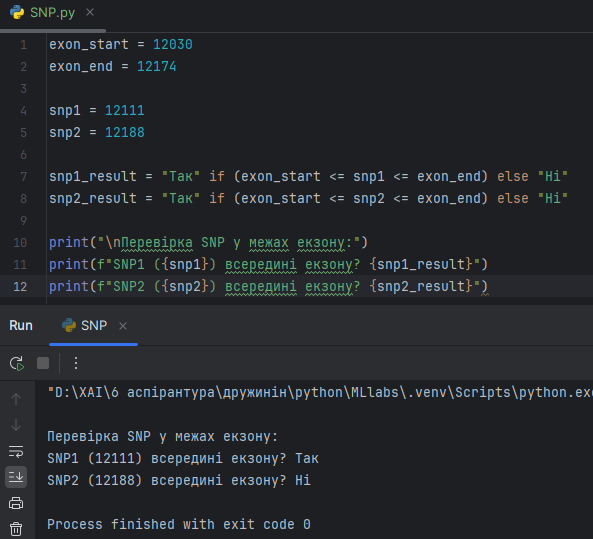


Рисунок 2.2 – Перевірка позицій в межах екзону

3. SNP FB\_AFFY\_0000024 чіпа Apple 480K SNP має 5’ фланкуючу область (тобто прямий зонд) CATTATTTTCACTTGGGTCGAGGCCAGATTC  
CATC і 3’ фланкуючу область (тобто зворотний зонд) GGATTGCCC  
GAAATCAGAGAAAAGTCG. SNP - це трансверсія G/A. Дати відповідь на наступні запитання:

Яка довжина 5’ флангової області? Та 3’ фланговий регіон?

Код IUPAC трансверсії G/A – R. Яка послідовність усієї області з використанням нотації “[G/A]” для SNP (підказка: об’єднайте її в новий рядок, який називається регіоном) і нотацію iupac R (region\_iupac)?

Отримати та надрукувати лише SNP з регіону та iupac\_region

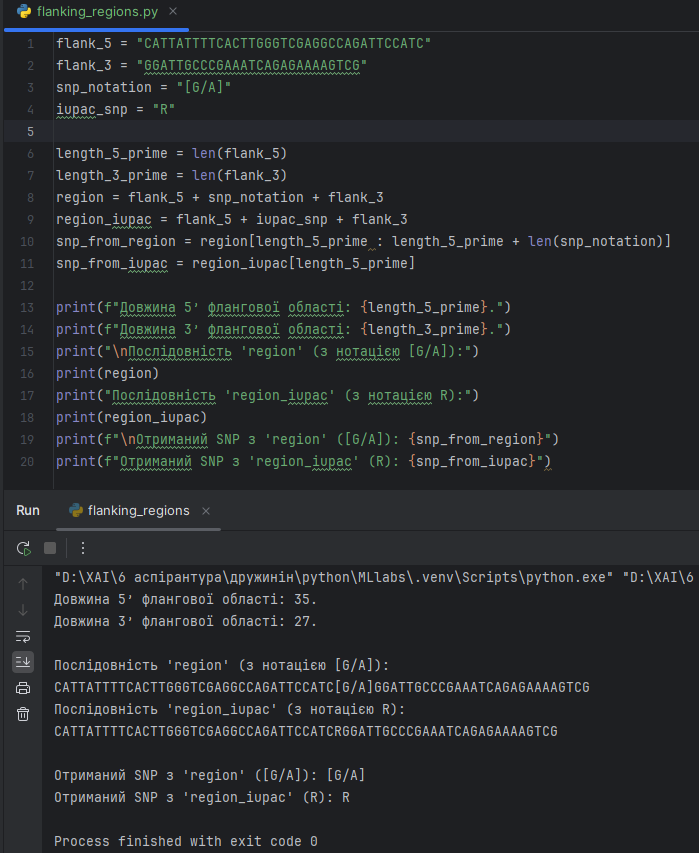


Рисунок 2.3 – Виконання 3-го завдання

4. Розрахуйте температуру плавлення Tm праймера з послідовністю

«TTAGCACACGTGAGCCAATGGAGCAAACGGGTAATT». Температуру плавлення Tm (у градусах Цельсія) можна обчислити як: Tm=64,9+41(GC−16,4)/N, де GC — загальна кількість G і C у праймері, а N — його довжина.

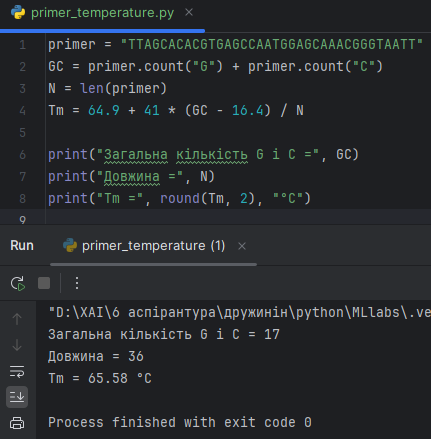


Рисунок 2.4 – Розрахунок температури плавлення

5. Перетворіть наступний екстракт гена PalB2 в мРНК (тобто замініть тимін на урацил):

CTGTCTCCCTCACTGTATGTAAATTGCATCTAGAATAGCA

TCTGGAGCACTAATTGACACATAGTGGGTATCAATTATTA

TTCCAGGTACTAGAGATACCTGGACCATTAACGGATAAAT

AGAAGATTCATTTGTTGAGTGACTGAGGATGGCAGTTCCT

GCTACCTTCAAGGATCTGGATGATGGGGAGAAACAGAGAA

CATAGTGTGAGAATACTGTGGTAAGGAAAGTACAGAGGAC

TGGTAGAGTGTCTAACCTAGATTTGGAGAAGGACCTAGAA

GTCTATCCCAGGGAAATAAAAATCTAAGCTAAGGTTTGAG

GAATCAGTAGGAATTGGCAAAGGAAGGACATGTTCCAGAT

GATAGGAACAGGTTATGCAAAGATCCTGAAATGGTCAGAG

CTTGGTGCTTTTTGAGAACCAAAAGTAGATTGTTATGGAC

CAGTGCTACTCCCTGCCTCTTGCCAAGGGACCCCGCCAAG

CACTGCATCCCTTCCCTCTGACTCCACCTTTCCACTTGCC CAGTATTGTTGGTGT і надрукуйте кількість присутніх урацилів і загальну довжину послідовності (не забудьте видалити новий рядок).

Розглядаючи генетичний код і всі можливі відкриті рамки зчитування, дайте відповідь на наступні запитання:

Скільки стоп-кодонів присутні в послідовності?

Скільки гліцинів (Gly)?

Чи присутній триптофан (Trp)?

Яка позиція крайнього лівого Trp? Роздрукуйте кодон, щоб перевірити правильність (підказка: нарізання).

Яка позиція крайнього правого Trp? Роздрукуйте кодон, щоб перевірити правильність (підказка: нарізання).

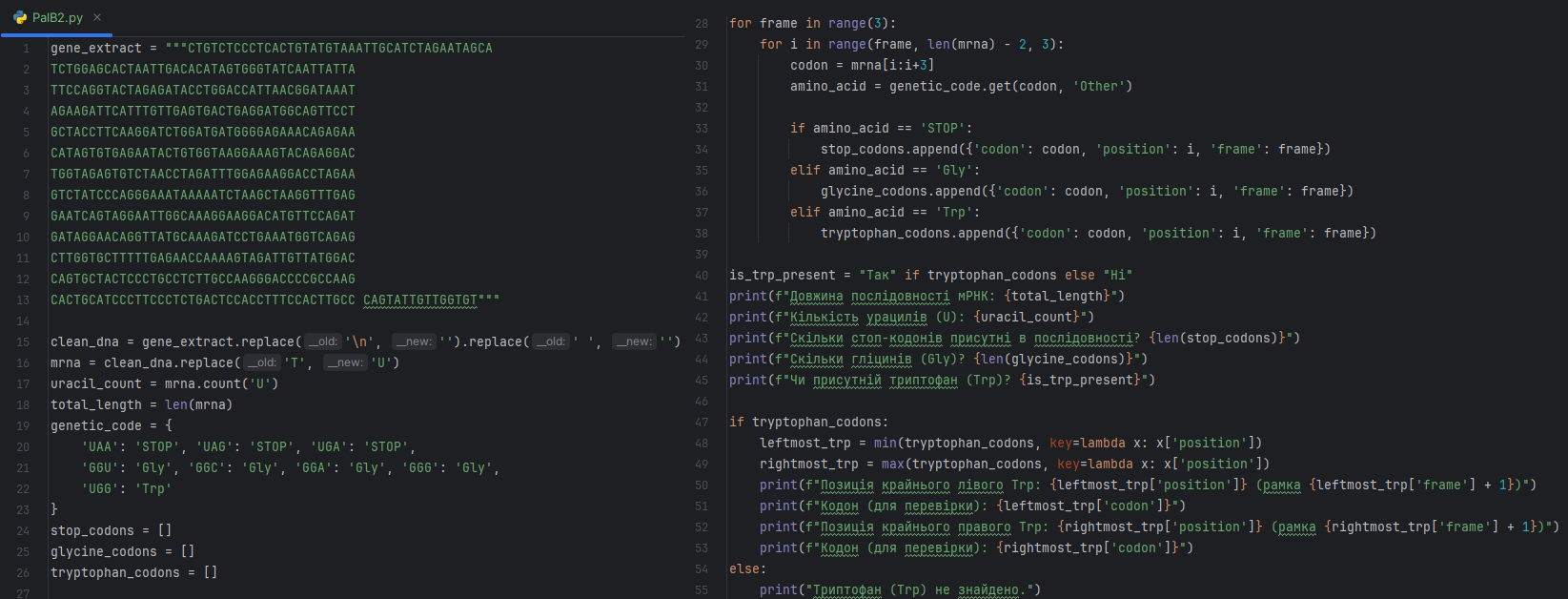


Рисунок 2.5 – Код для завдання



Рисунок 2.6 – Протокол роботи програми

Подумайте, як прочитати Illumina HiSeq 4000:

“”“AATGATACGGCGACCACCGAGATCTACACGCCTCCCTCGCGC

CATCAGAGAGTCTGGGTCTCAGGTACCGCAGTTGTATCTTGCGCGACTATA

ATCCACGGCTCTTATTCTAGCGTGCGCGTACGGCGGTGGGCGTCGTTACGCTATATT”

“”.

Дати відповідь на наступні запитання:

Скільки триває читання (остерігайтеся нових рядків)?

Який GC вміст прочитаного ($gc = \frac{G+C}{A+T+C+G}$)?

Адаптер Nextera:

"AATGATACGGCGACCACCGAGATCTACACGCCTCCCTCGCGCCATCAG" Чи присутній він у прочитаному? Як довго це?

Вийміть адаптер Nextera зі зчитування та перерахуйте вміст GC. Чи збільшився вміст GC після обрізання адаптера?

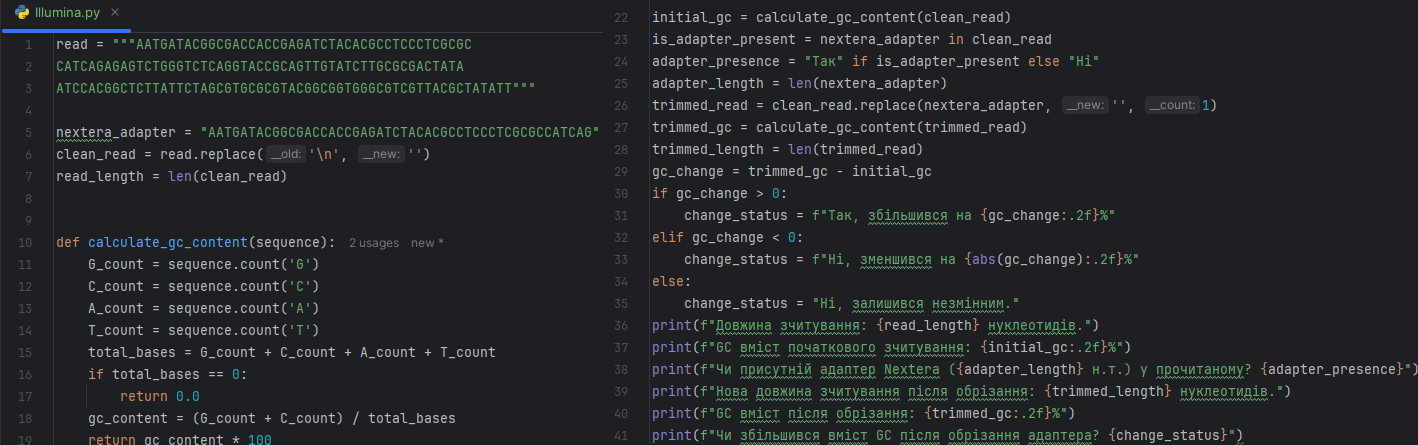


Рисунок 2.7 – Код для завдання

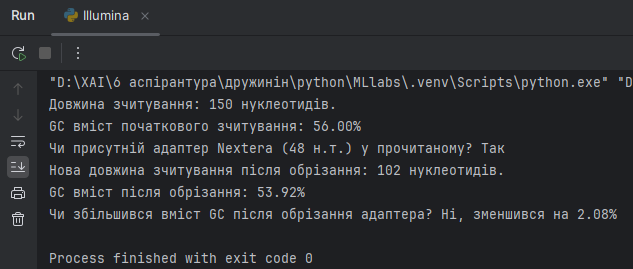


Рисунок 2.8 – Протокол роботи

7. Дано ген A, починаючи з позиції 1000 і закінчуючи позиції 3400, і ген B, починаючи з позиції 3700 і закінчуючи позиції 6000. Довільно виберіть позицію від 1 до 5202 і перевірте наступне:

a. є позиція в гені А?

б. є позиція в гені B?

c. позиція знаходиться між двома генами?

d. позиція знаходиться в одному з двох генів?

e. позиція знаходиться поза обома генами?

f. знаходиться в межах 100 до початку гена А? Щоб вибрати випадкове число, ви можете імпортувати модуль random і використовувати функцію random.randint(start,end):

import random

pos = random.randint(1,6000)

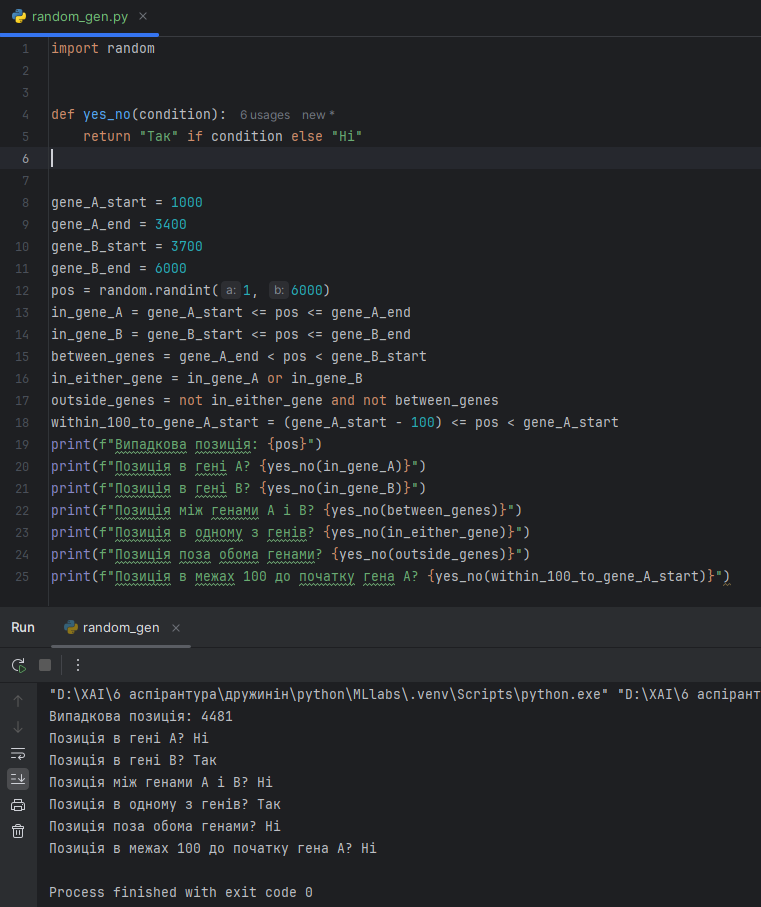


Рисунок 2.9 – Код та робота програми

8. ДНК-зв’язуючий домен білка-супресора пухлини TP53 може бути представлений рядком:

chain\_a = “”“SSSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTYSPALNKM

FCQLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVV

RRCPHHERCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFR

HSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILT

IITLEDSSGNLLGRNSFEVRVCACPGRDRRTEEENLRKKG

EPHHELPPGSTKRALPNNT”“”

Дати відповідь на наступні запитання:

- На скількох рядках записана послідовність?

- Якої довжини є послідовність (вилучити новий рядок)?

- Створіть нову послідовність з видаленими всіма новими рядками

- Скільки цистеїнів «С» і гістидинів «Н» у послідовності?

- Чи містить ланцюжок підпослідовність «NLRVEYLDDRN»? Де?

- Витягніть перший рядок послідовності (Підказка: використовуйте функцію find).

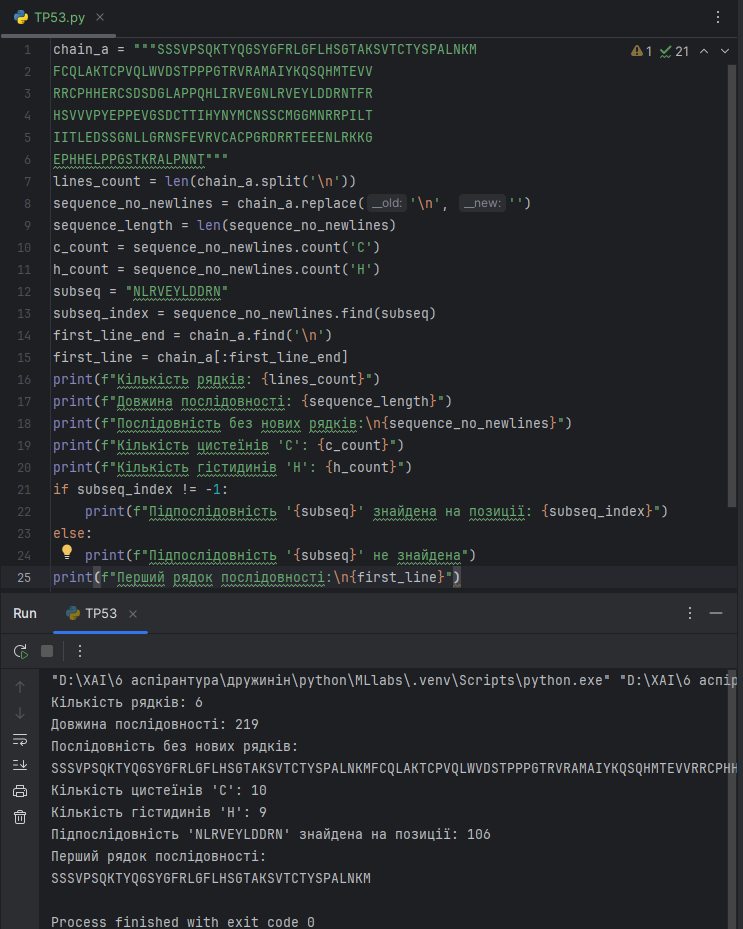


Рисунок 2.10 – Код та робота програми

Обчисліть нулі рівняння ax2-b=0, де a = 10 і b = 1. Підказка: використовуйте math.sqrt або \*\* 0,5. Перевірте, що підстановка отриманого значення x та у рівняння дає нуль.

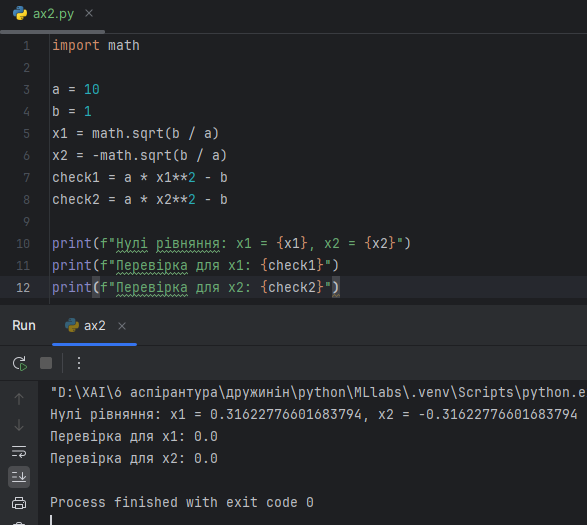


Рисунок 2.11 – Код та робота програми

Лабораторна робота №3

Функції, зовнішні модулі та файли у Python.

Виконання завдань

Рівень 1

1. Реалізувати підрахунок кількості унікальних символів у рядку за допомогою використання функції та словника.

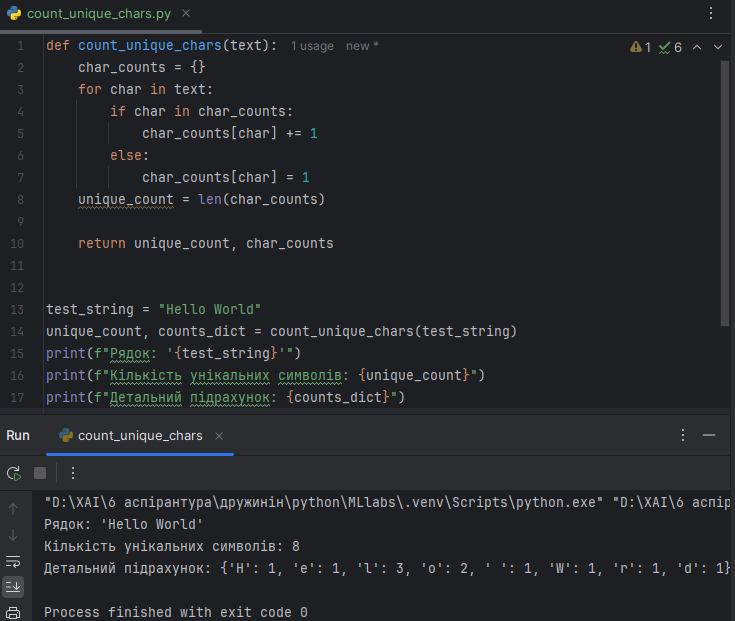


Рисунок 3.1 – Реалізація підрахунку унікальних символів

2. Запишіть функцію, яка обчислює суму довільної кількості числових аргументів. При цьому обов’язковою є перевірка на коректність числових даних.

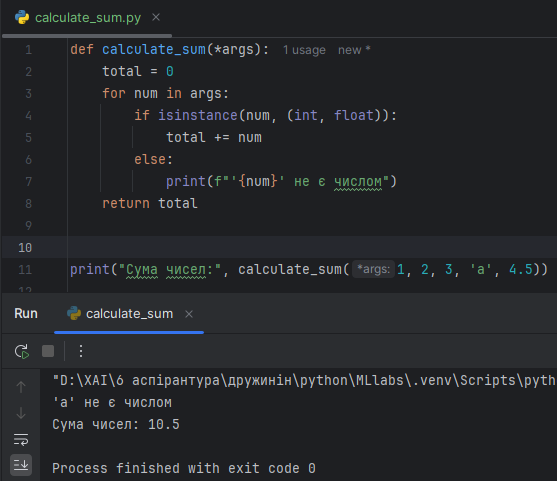


Рисунок 3.2 – Реалізація обчислення суми

3. Реалізуйте функцію, яка дозволяє будувати таблицю (словник) значень для будь-якого числа значень і довільної функції.

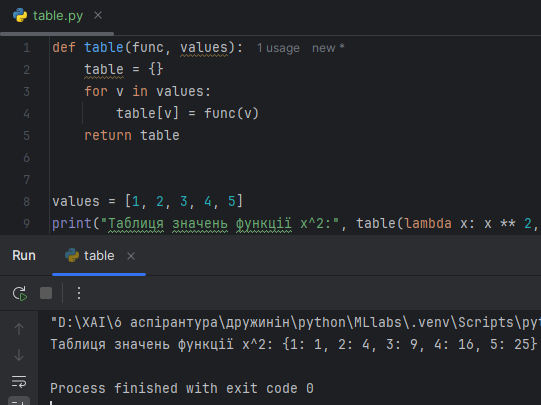


Рисунок 3.3 – Реалізація функції простої таблиці значень

4. Написати функцію, яка виконує циклічне зміщення елементів списку цілих чисел на вказану кількість кроків n раз. Результат виводиться після кожного перетворення. Зсув має бути кільцевим, тобто якщо елемент вийшов за межі списку він має з’явитися із іншого кінця.

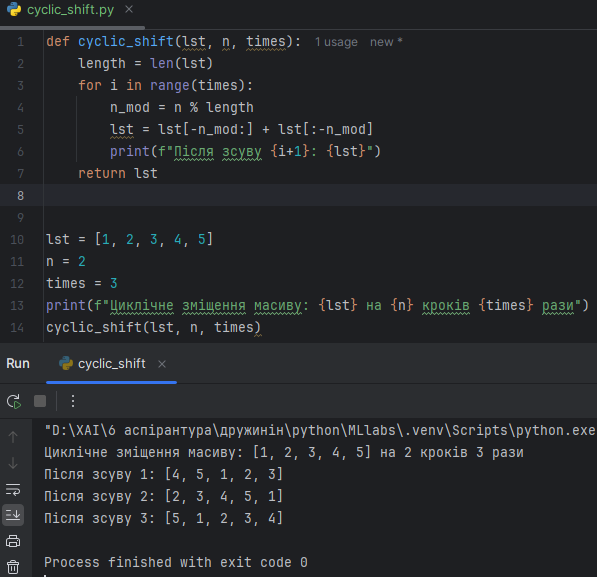


Рисунок 3.4 – Циклічне зміщення

5. Обчислити значення виразу 1+x+x2+....+xn та результат подати у вигляді 1+2+22+....+2n = сума.

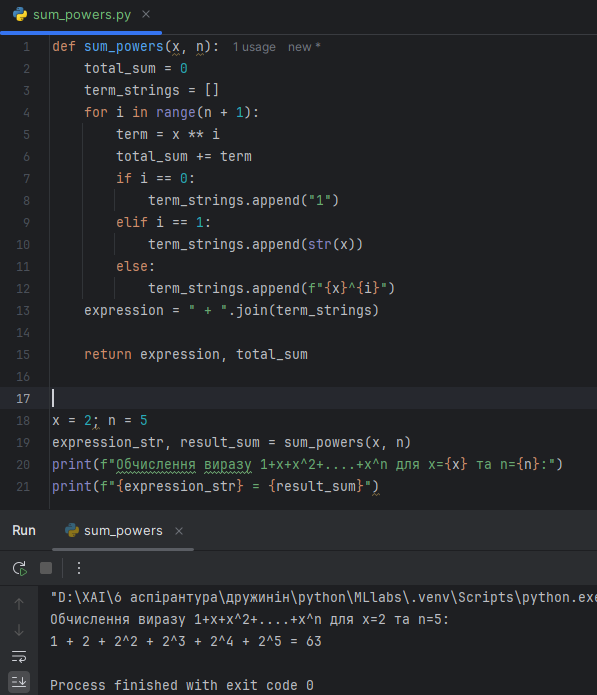


Рисунок 3.5 – Обчислення виразу

Рівень 2

1. Рекурсивно обчислити факторіал заданого числа.

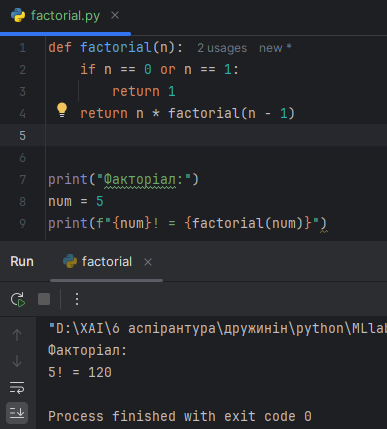


Рисунок 3.6 – Рекурсивне обчислення факторіала

2. Записати рекурсію для знаходження n числа Фібоначчі.

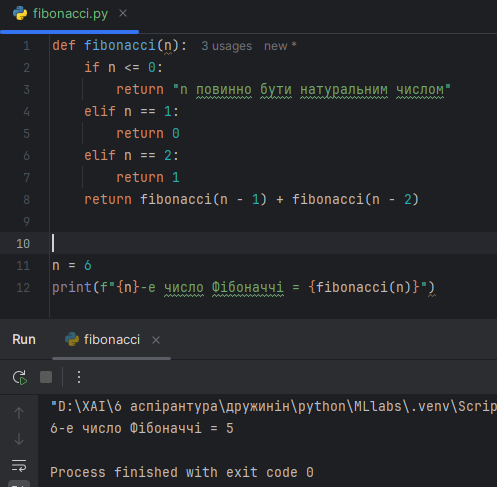


Рисунок 3.7 – Знаходження числа Фібоначчі

3. Рекурсивно перевірити чи є задане число натуральним.

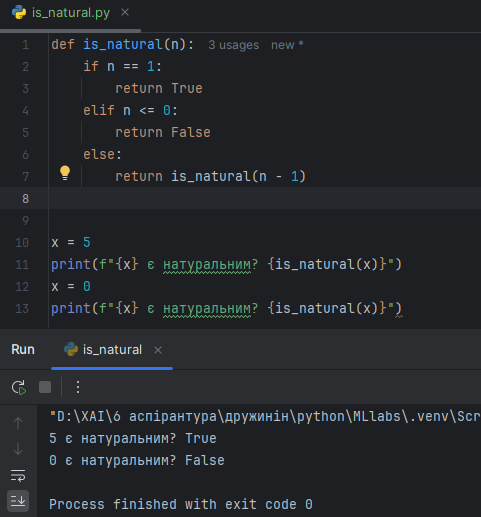


Рисунок 3.8 – Перевірка натурального числа

4. Вводиться послідовність, яка завершується 0. Виведіть цю послідовність у зворотному порядку.

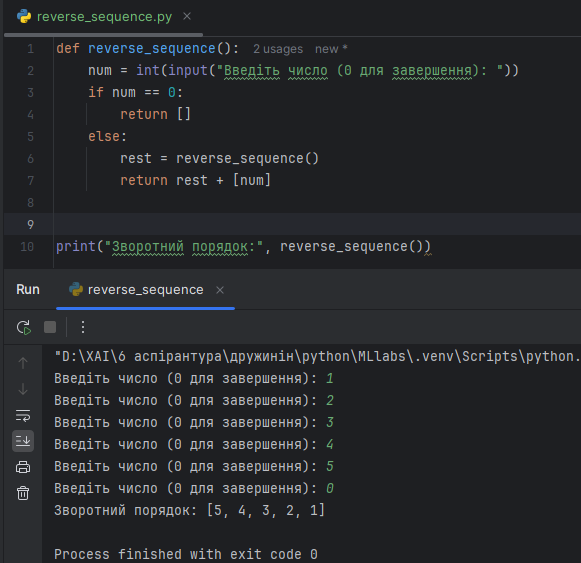


Рисунок 3.9 – Послідовність у зворотному порядку

5. Реалізувати піднесення до цілого степеня, незалежно від того від’ємний чи додатній степінь

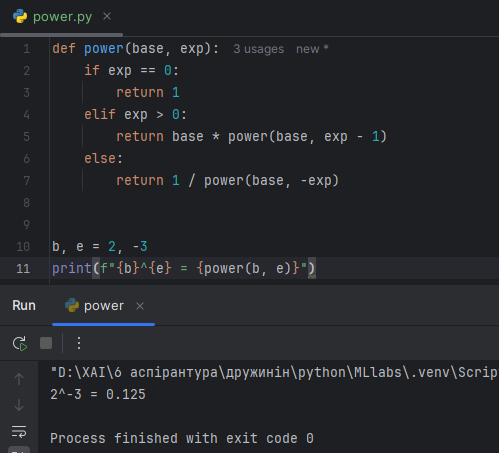


Рисунок 3.10 – Реалізація піднесення до цілого степеня

Рівень 3

1. Порахувати, скільки слів у текстовому файлі рядків, слів і символів.

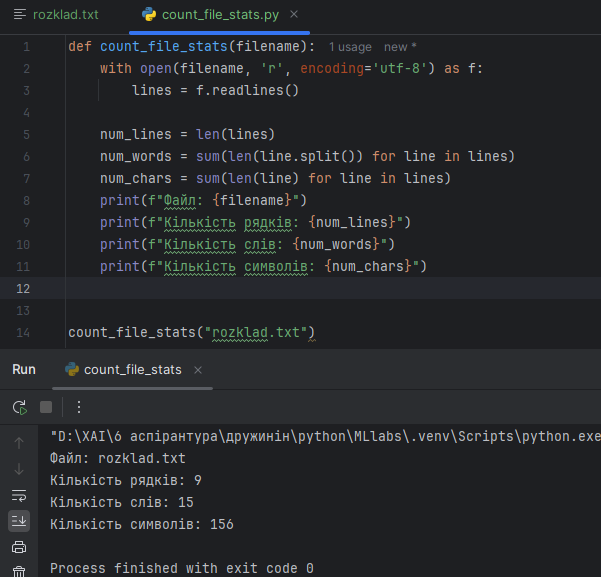


Рисунок 3.11 – Реалізація підрахунку слів і символів

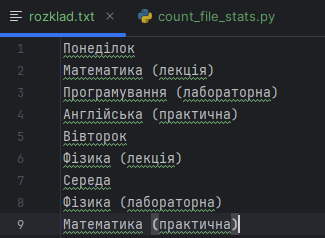


Рисунок 3.12 – Текстовий файл для тестування

2. Задано файл із розкладом на тиждень. Окрім дня тижня, у рядку подаються назви предметів. Біля назви предмету у дужках тип заняття (лекція, лабораторна, практична). У рядку може бути запис або дня, або предмету. Порахувати кількість різних предметів, кількість лекцій, практичних, лабораторних.

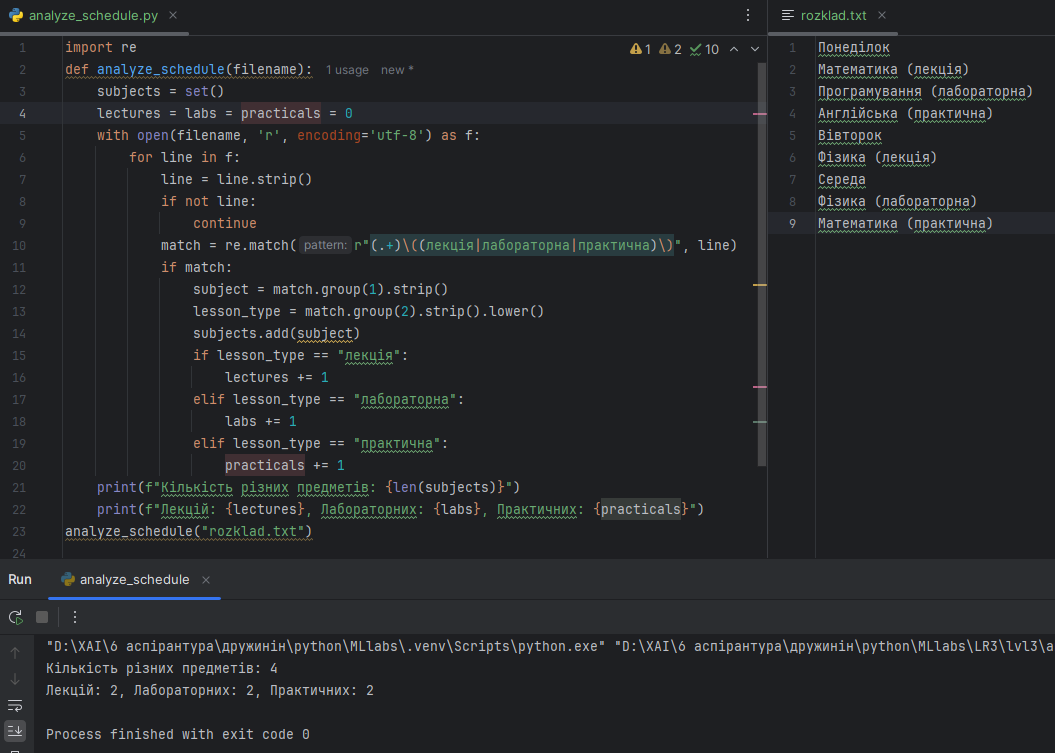


Рисунок 3.13 – Підрахунок предметів, лекцій, лабораторних, практичних

3. Написати програму, яка створює текстовий файл із 9 рядків, у якому в 1-рядку записано одна літера «a», 2 – дві ....

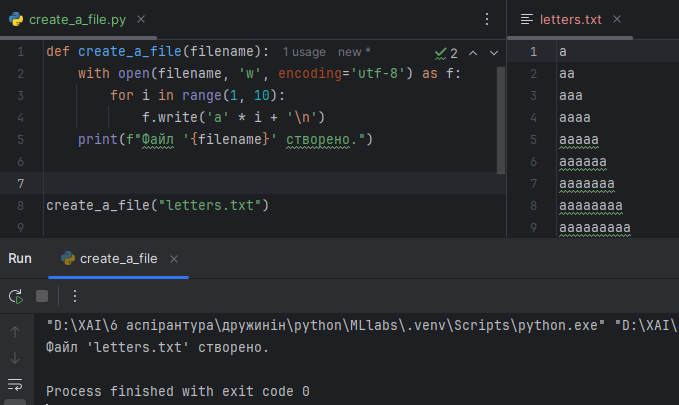


Рисунок 3.14 – Протокол роботи програми

4. Задано текстовий файл, який містить принаймні один рядок. Порахувати кількість рядків, які завершуються заданим символом.

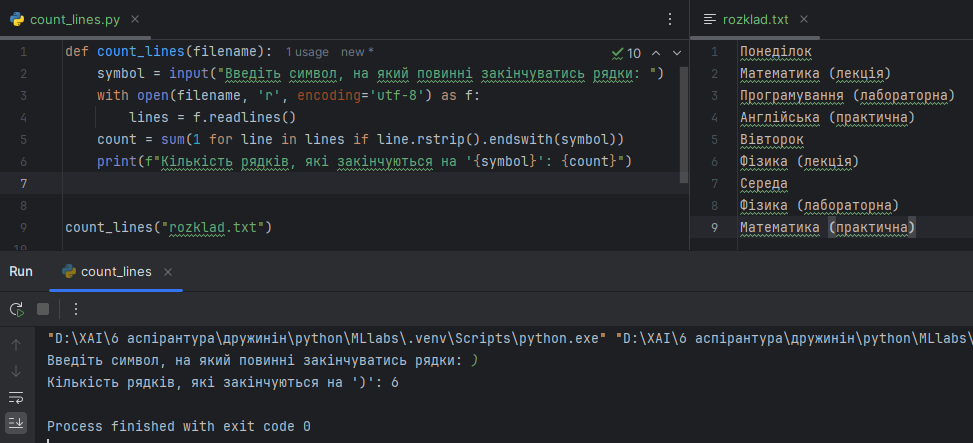


Рисунок 3.15 – Підрахунок рядків, які завершуються заданим символом

5. Задано текстовий файл, який містить принаймні один рядок. Вивести рядки, які завершують на вказаний символ та інвертувати їх.

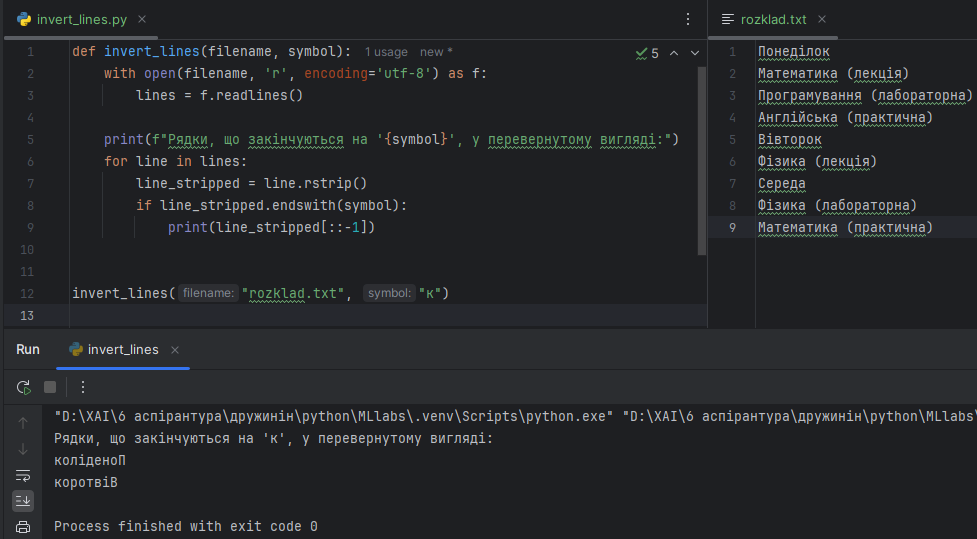


Рисунок 3.16 – Інвертування рядків за вказаним символом

Лабораторна робота № 4

Автоматизація підготовки даних для алгоритмів машинного навчання.

Виконання завдань

1. Відкрити текстовий файл та проаналізувати дані.

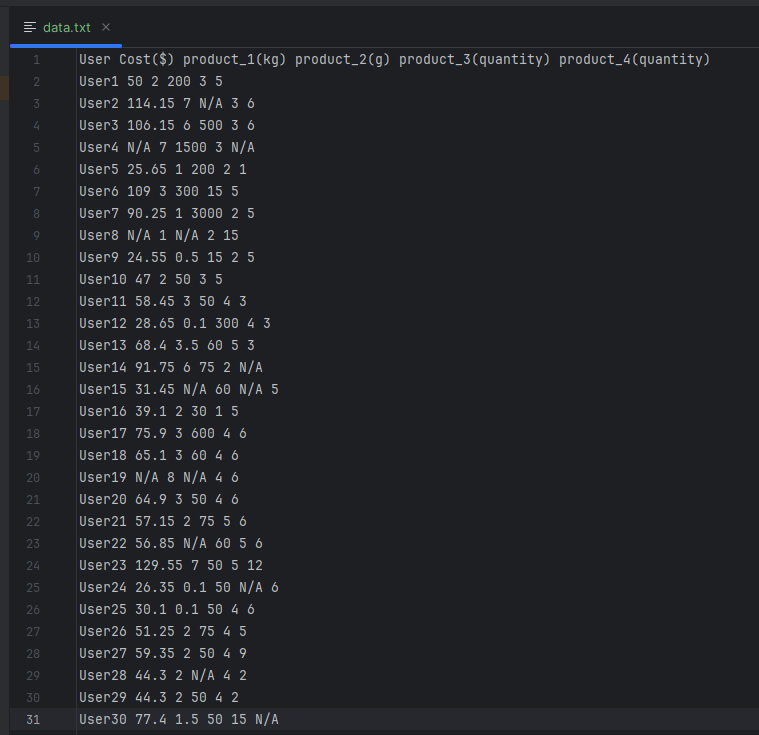


Рисунок 4.1 – Текстовий файл з вихідними даними

2. Замінити всі помилки на середнє арифметичне значення властивості та записати у файл TabSA.txt.

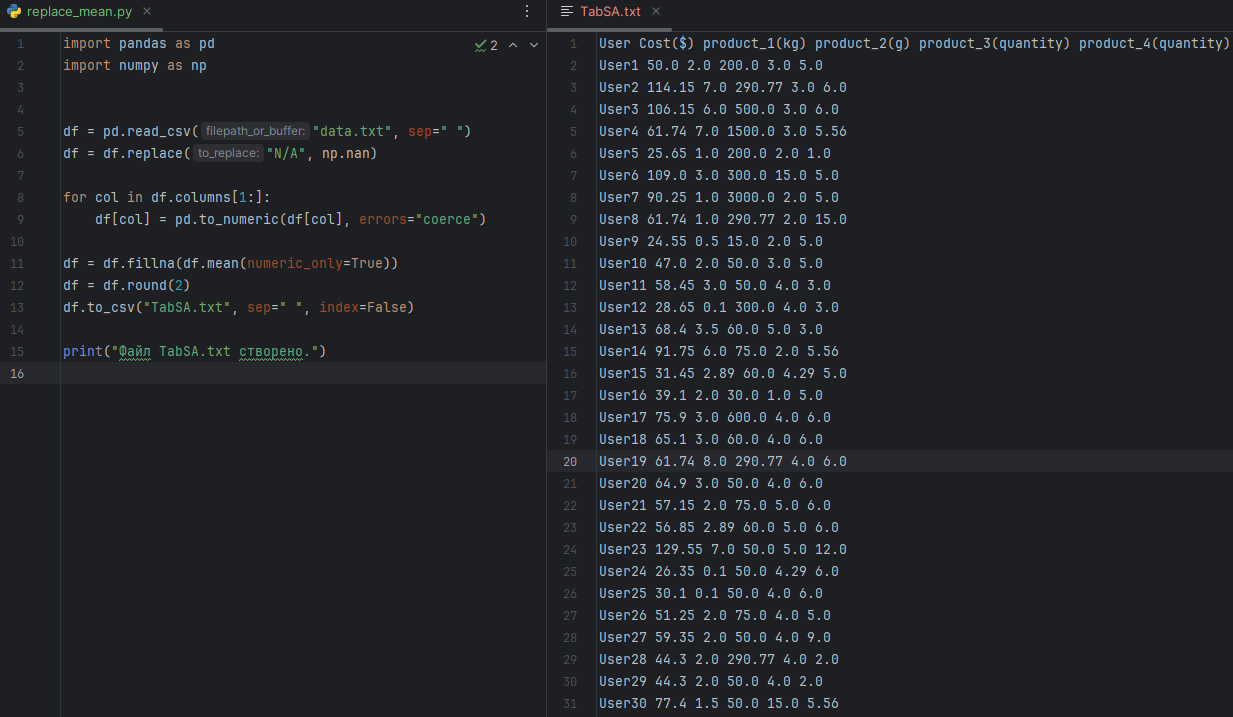


Рисунок 4.2 – Протокол роботи програми та створення файлу TabSA.txt

3. Замінити всі помилки на моду та записати у файл TabMODA.txt.

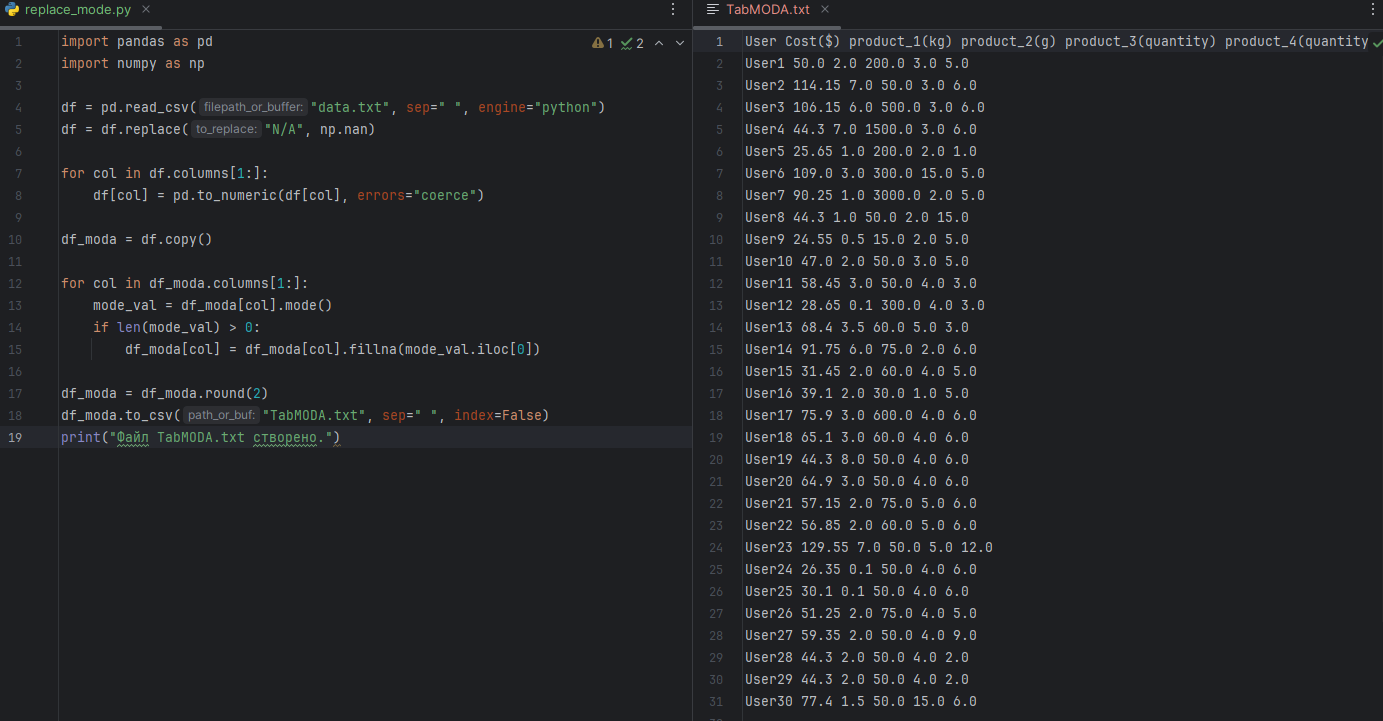


Рисунок 4.3 – Протокол роботи програми та створення файлу TabMODA.txt

4. Замінити всі помилки на медіану та записати у файл TabMD.txt.

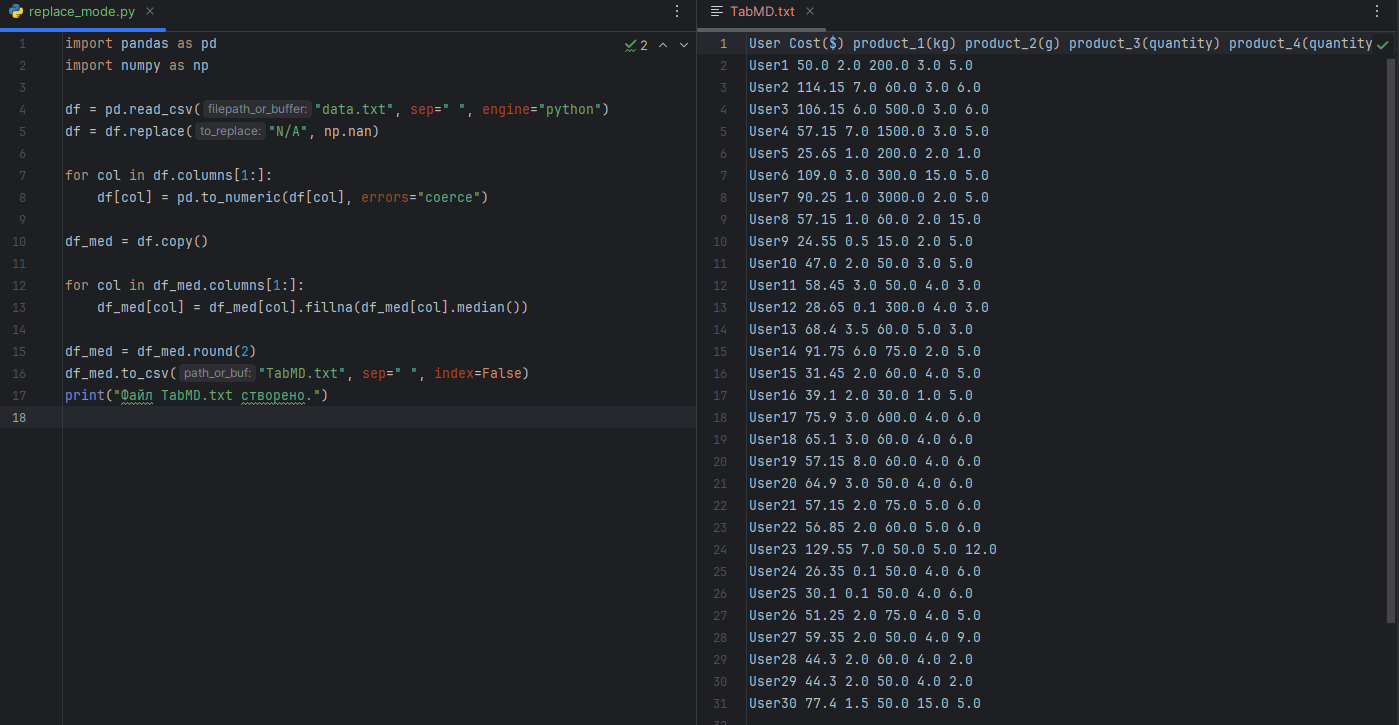


Рисунок 4.4 – Протокол роботи програми та створення файлу TabMD.txt

5. Виконати нормалізацію даних та записати у файл TabNorm.txt

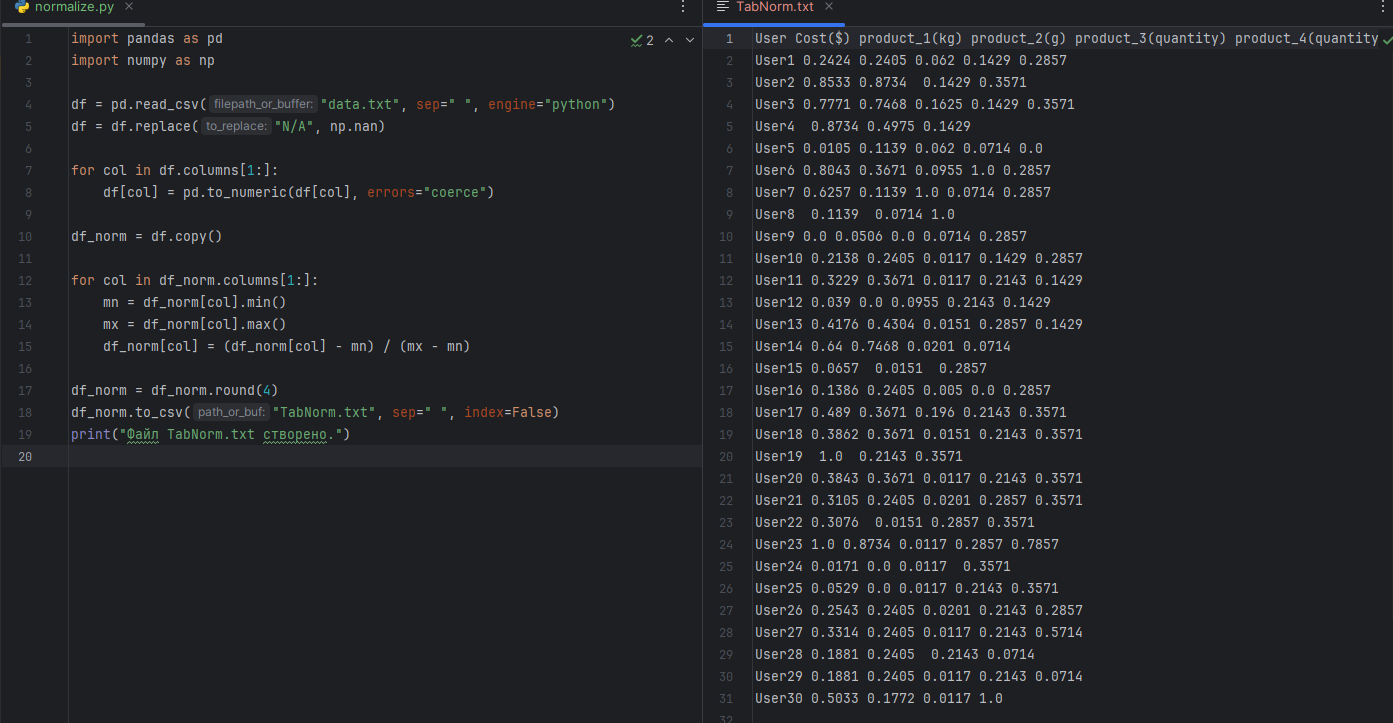


Рисунок 4.5 – Протокол роботи програми та створення файлу TabNorm.txt

6. Відновити дані за допомогою метрики Манхеттен.

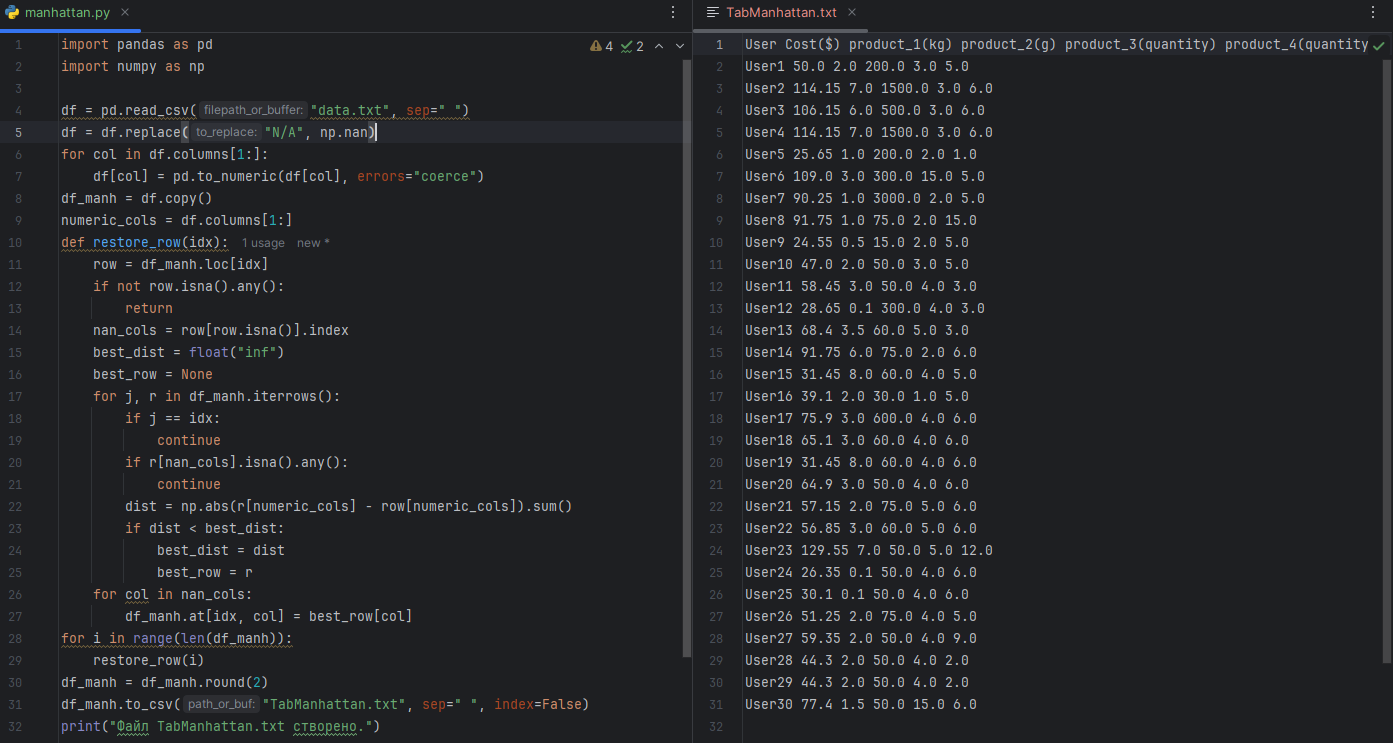


Рисунок 4.6 – Протокол роботи програми та створення файлу TabManhattan.txt

Лабораторна робота № 5

Побудова рекомендаційної системи відеопереглядів

Виконання завдань

1. Відкрити текстову таблицю даних.

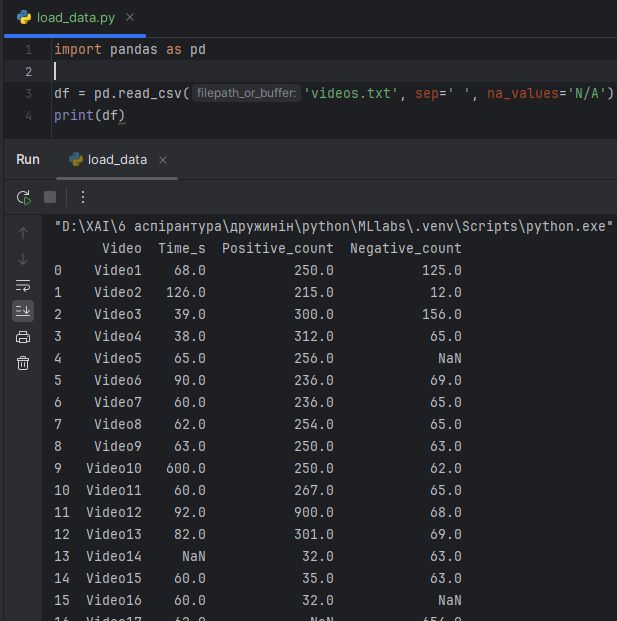


Рисунок 5.1 – Вивід текстової таблиці даних

2. Перевірити дані на похибки і якщо вони там є та виконати виправлення за допомогою метрики.

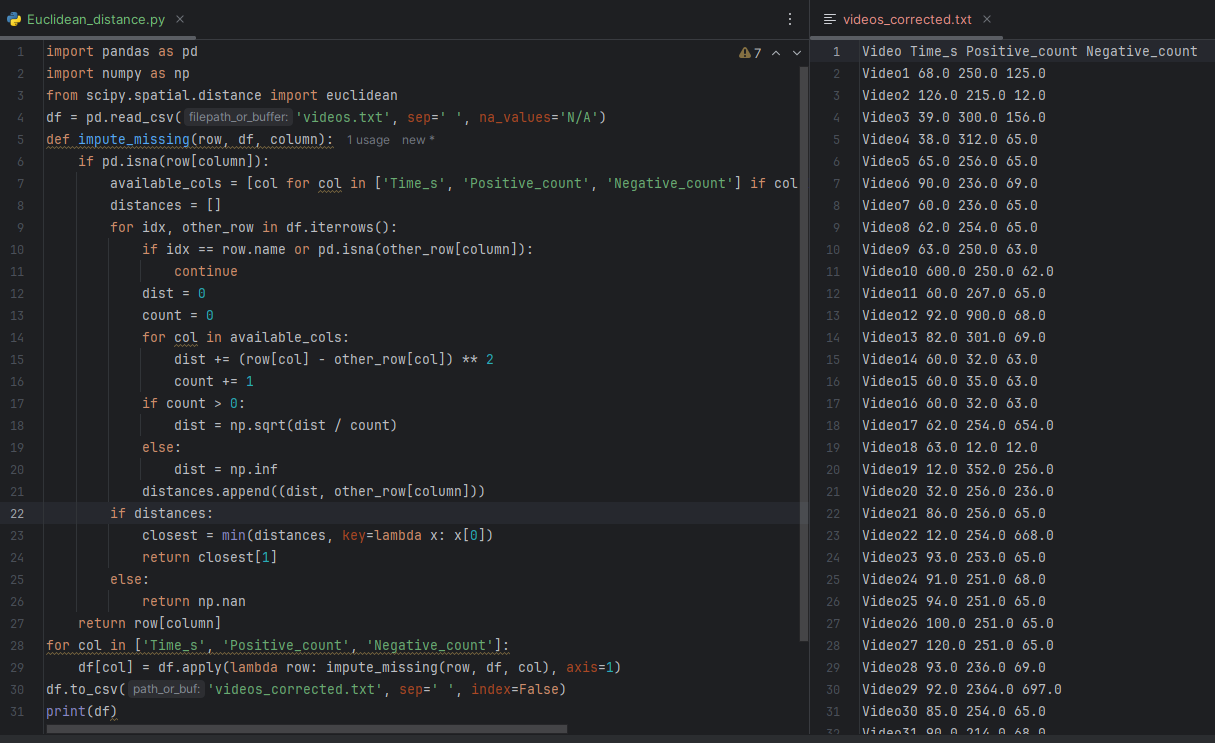


Рисунок 5.2 – Перевірка даних і імпутація за допомогою найближчого сусіда

3. Нормалізувати дані таблиці.



Рисунок 5.3 – Нормалізація даних таблиці за допомогою min-max

4. Створити рекомендаційну систему для відеохостингу, яка пропонує до перегляду 5 близьких за параметрами відеороликів. Система працює наступним чином — користувач вводить такі параметри як час перегляду, кількість позитивних та негативних оцінок, а система повертає список відео із вказаними статистичними даними.

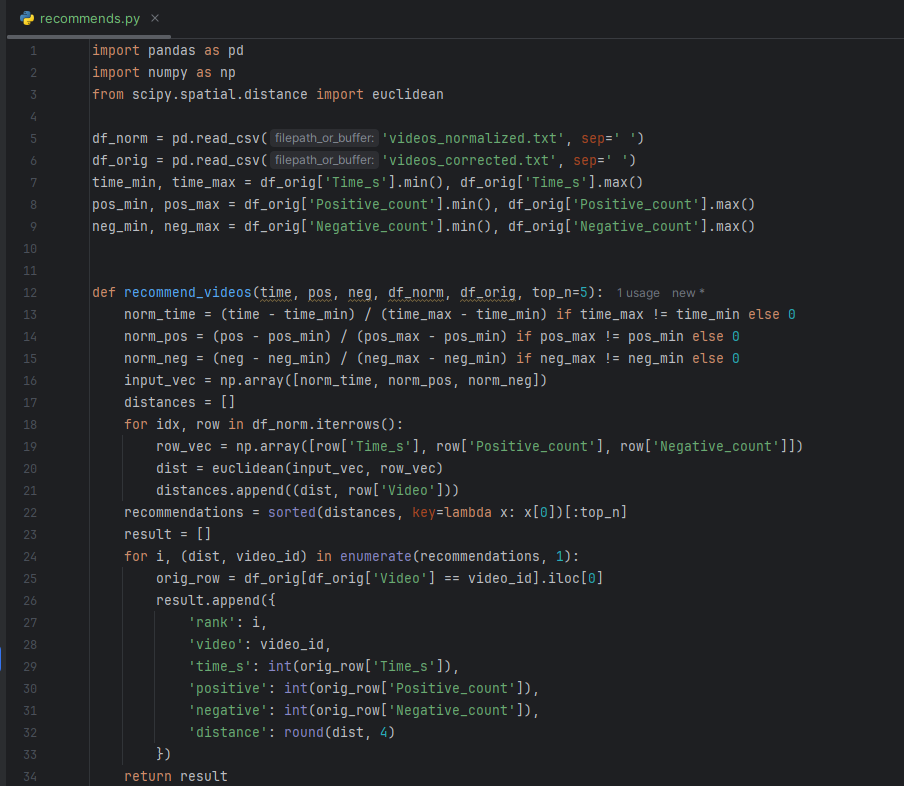


Рисунок 5.4 – Функція recommend\_videos

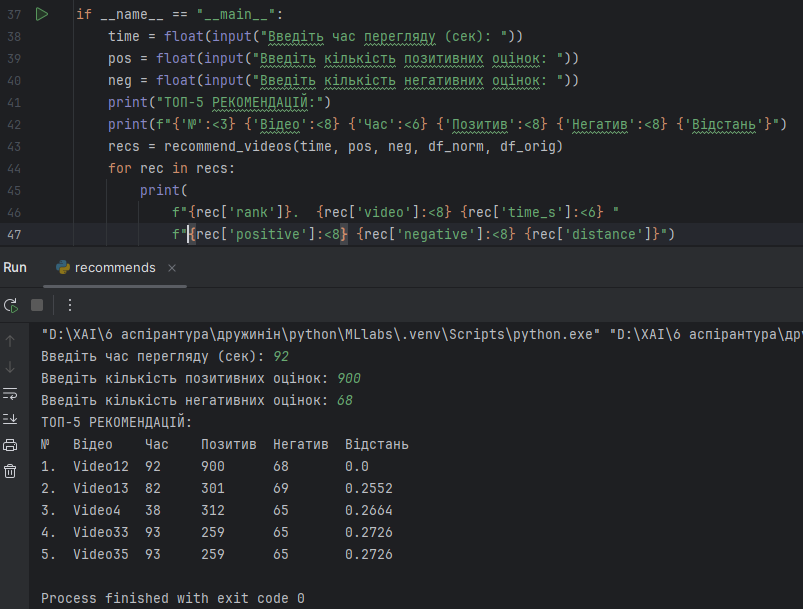


Рисунок 5.5 – Реалізація головної функції та перевірка рекомендацій

5. Удосконалити програму із урахуванням використання ключових слів для опису відео.

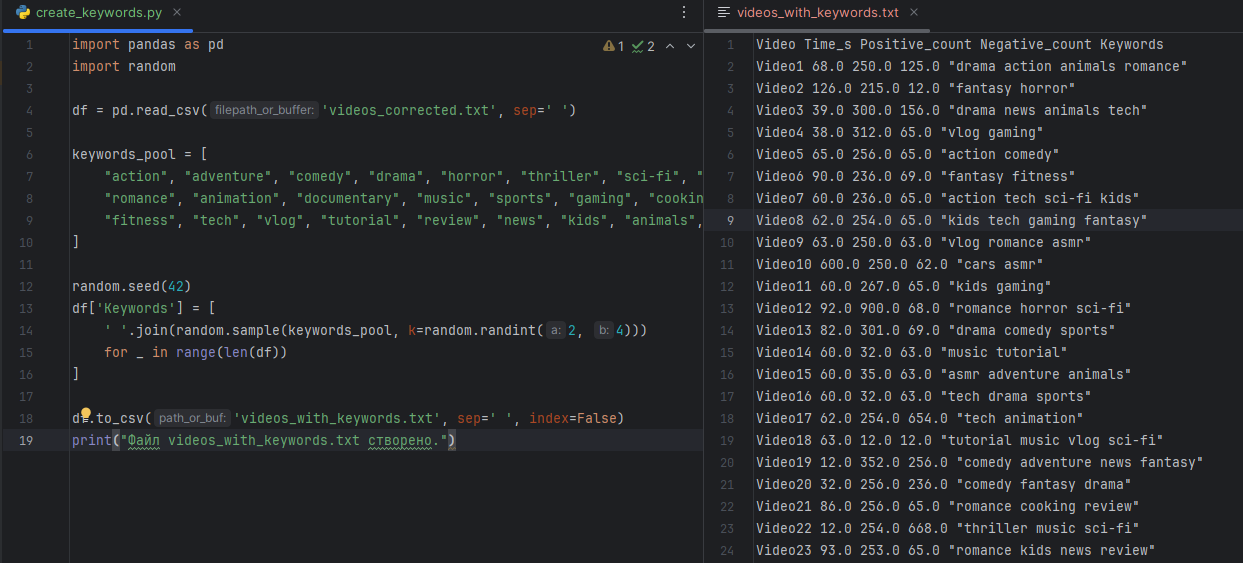


Рисунок 5.6 – Додавання до модифікованого файлу ключових слів

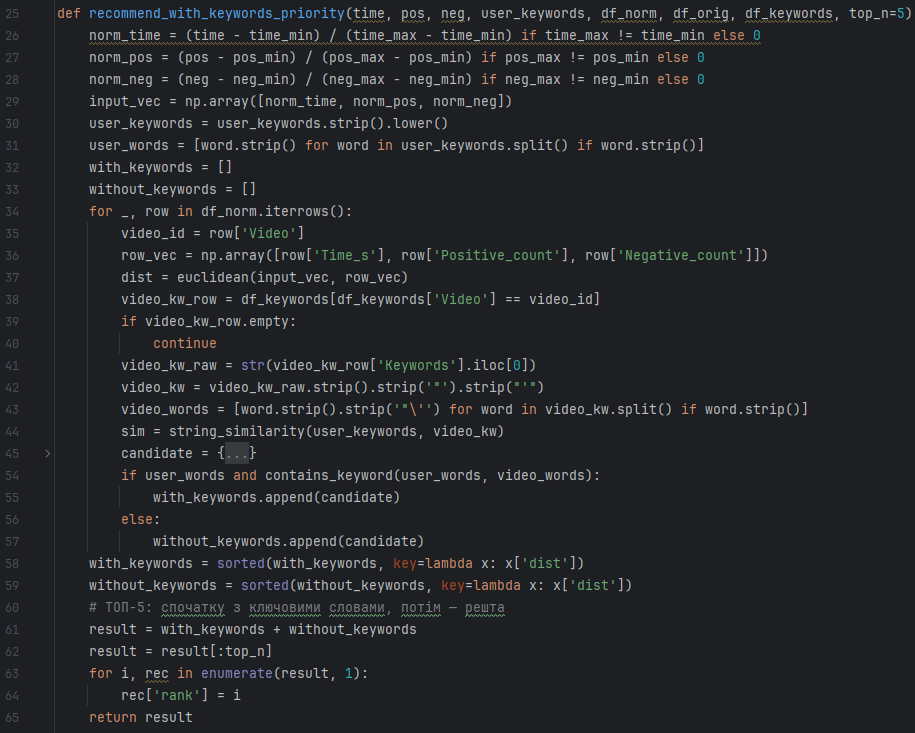


Рисунок 5.7 – Реалізація функції рекомендацій з додаванням ключових слів

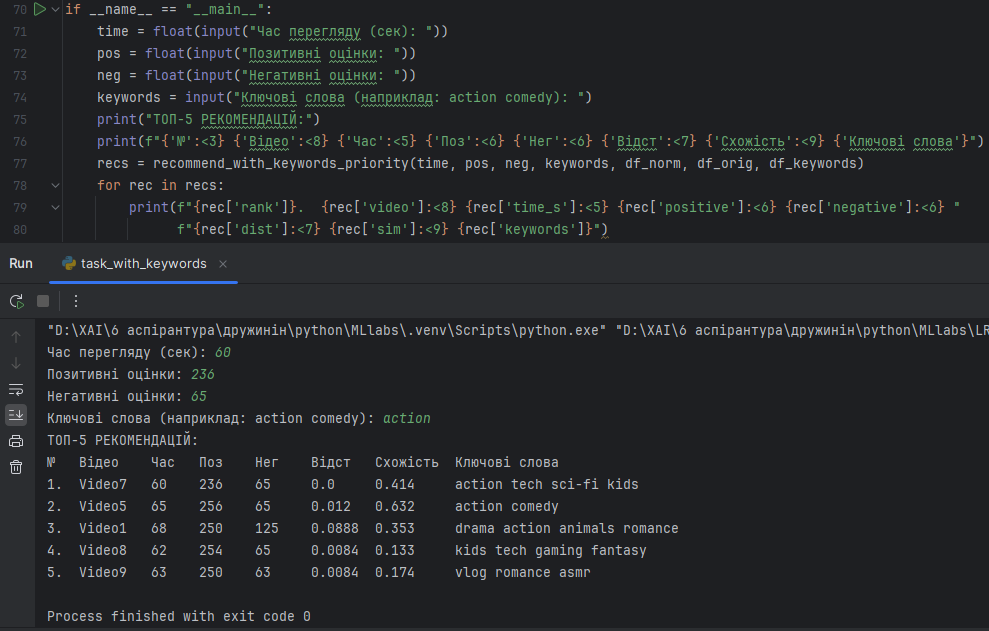


Рисунок 5.8 – Реалізація головної функції та перевірка рекомендацій

Лабораторна робота № 6

Задачі кластеризації

Виконання завдань

1. Завантажити базу даних Iris, вбудовану в модуль scikit-learn.

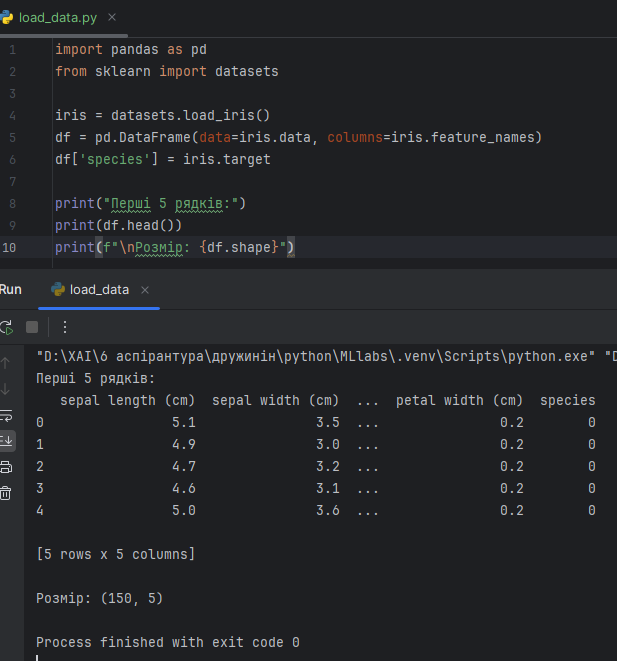


Рисунок 6.1 – Завантаження бази даних Iris та виведення перших 5 рядків

2. Виконати нормалізацію даних.

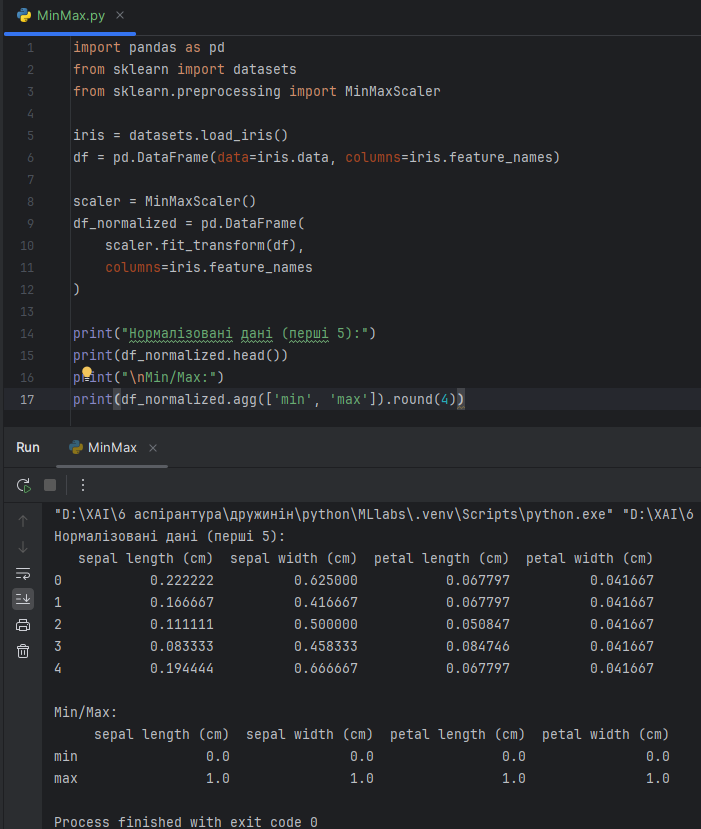


Рисунок 6.2 – Виконання нормалізації через мін-макс

3. Підключити бібліотеку scikit-learn.

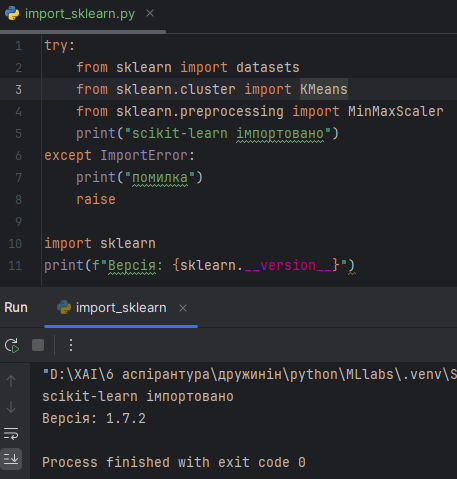


Рисунок 6.3 – Перевірка підключення бібліотеки scikit-learn

4. Виконати кластеризацію заданої множини (зокрема виділити центри кластерів).

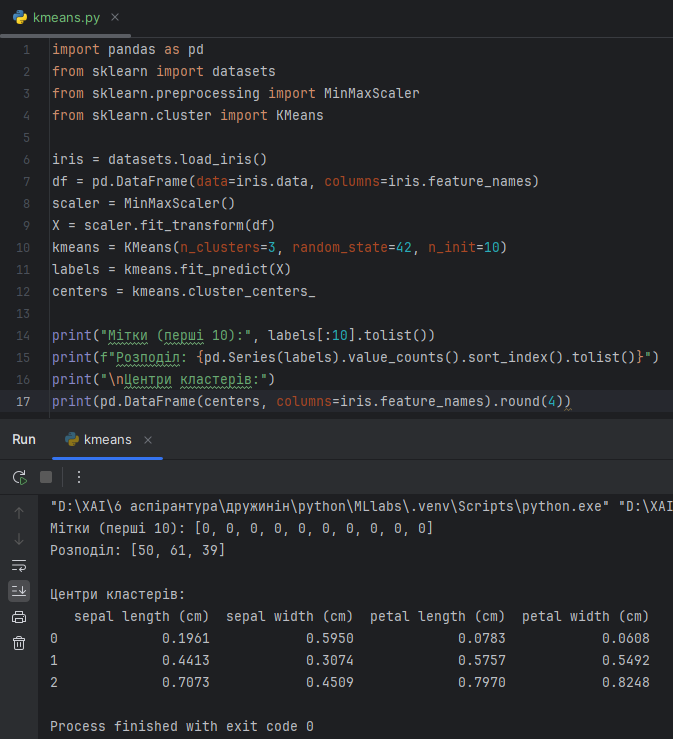


Рисунок 6.4 – Виконання кластеризації й виведення центрів кластерів

5. Продемонструвати графічно результати роботи алгоритму.

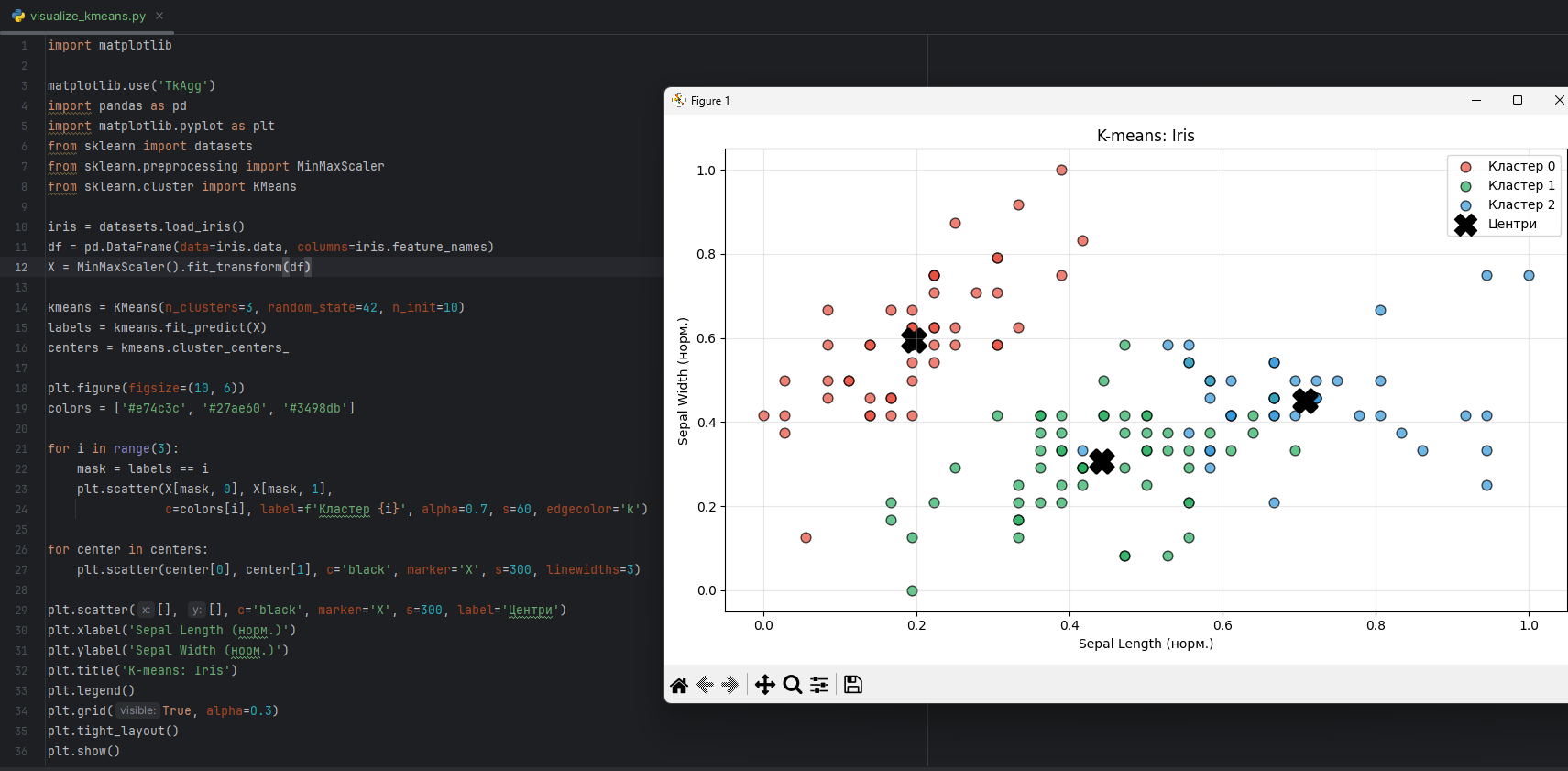


Рисунок 6.5 – Графічний результат роботи алгоритму

6. На прикладі бази даних Iris спробуйте реалізувати відшукання центрів кластерів за допомогою алгоритму FOREL, користуючись стандартними можливостями мови програмування Python.

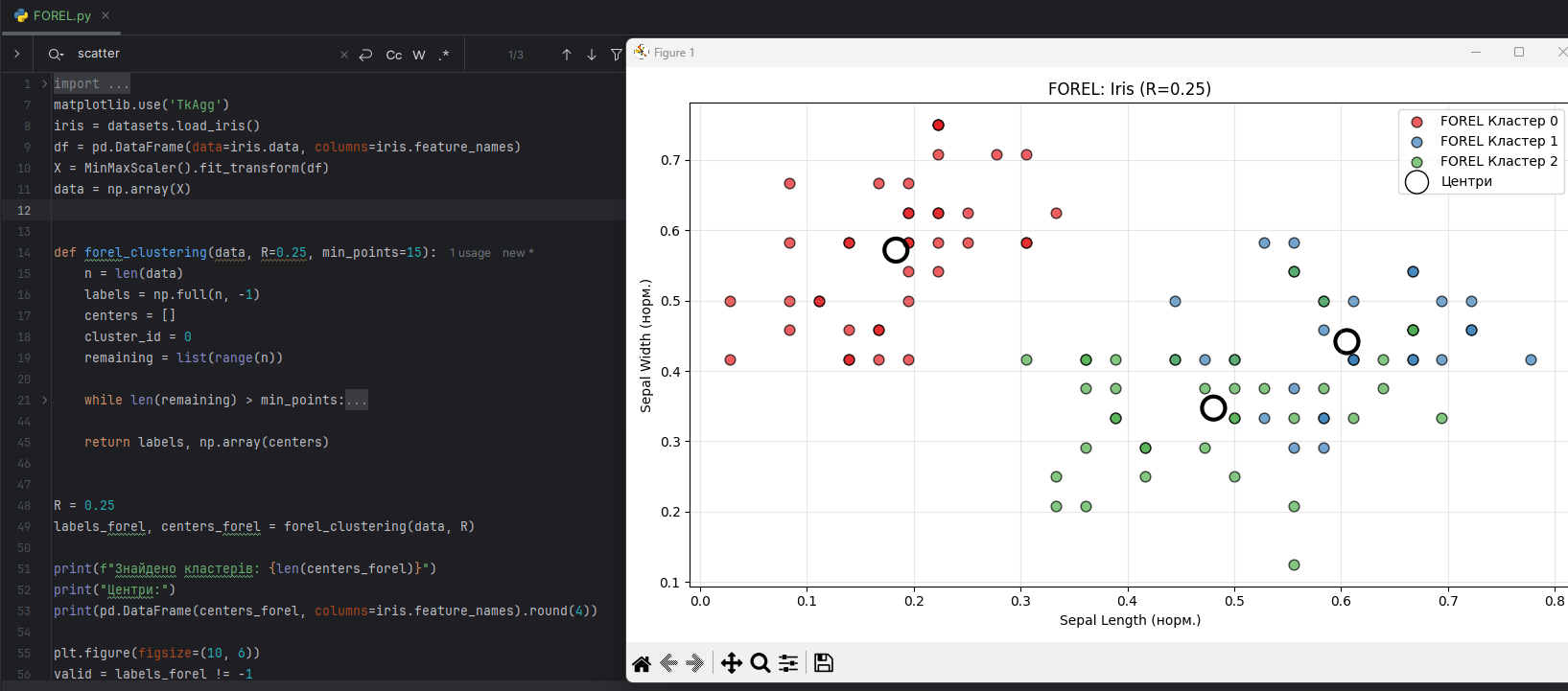


Рисунок 6.6 – Графічний результат роботи алгоритму FOREL

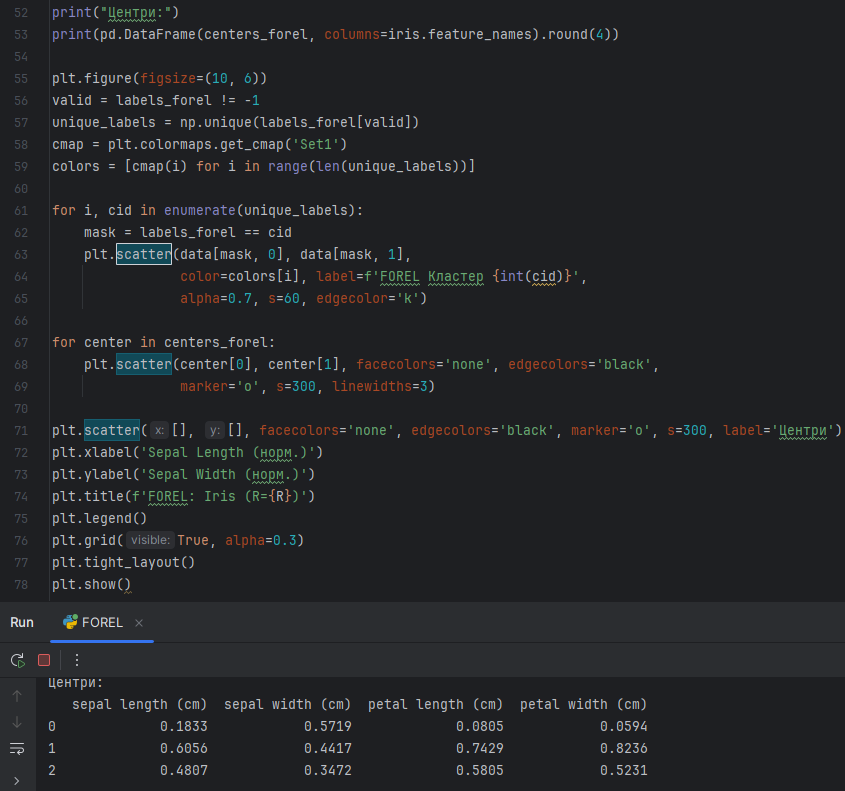


Рисунок 6.7 – Виведення центрів кластерів

Лабораторна робота № 7

Задачі класифікації

Виконання завдань

1. Завантажити базу даних Iris, вбудовану в модуль scikit-learn.



Рисунок 7.1 – Завантаження бази даних Iris та виведення перших 5 рядків

2. Виконати нормалізацію даних.

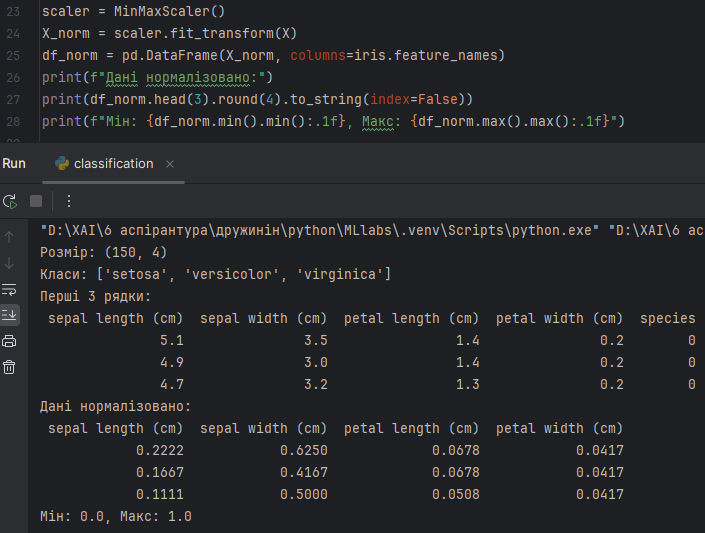


Рисунок 7.2 – Виконання нормалізації через мін-макс

3. Розбити дані на тренувальну та тестову вибірку

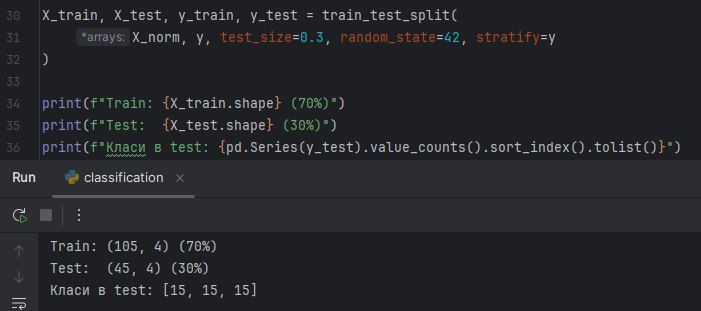


Рисунок 7.3 – Розбиття на тренувальну та тестову вибірку (70 на 30)

4. Побудувати моделі машинного навчання для класифікації об’єктів за допомогою метода kNN, дерева рішень та лінійного класифікатора.

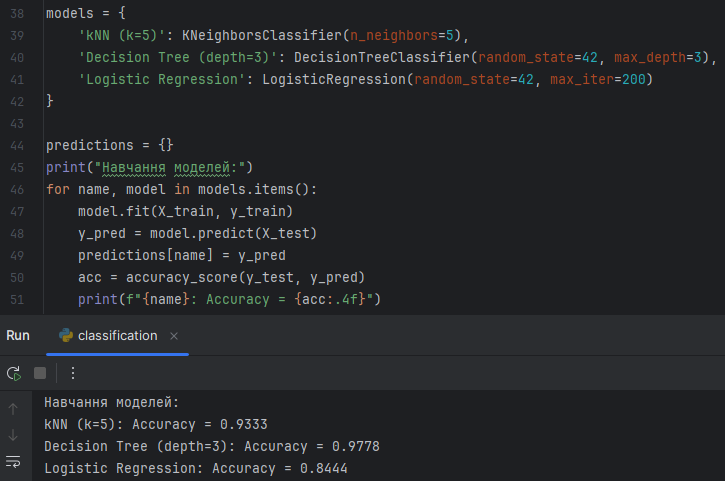


Рисунок 7.4 – Побудова моделей машинного навчання

5. Вивести у консоль показники ефективності кожного представлених методів.

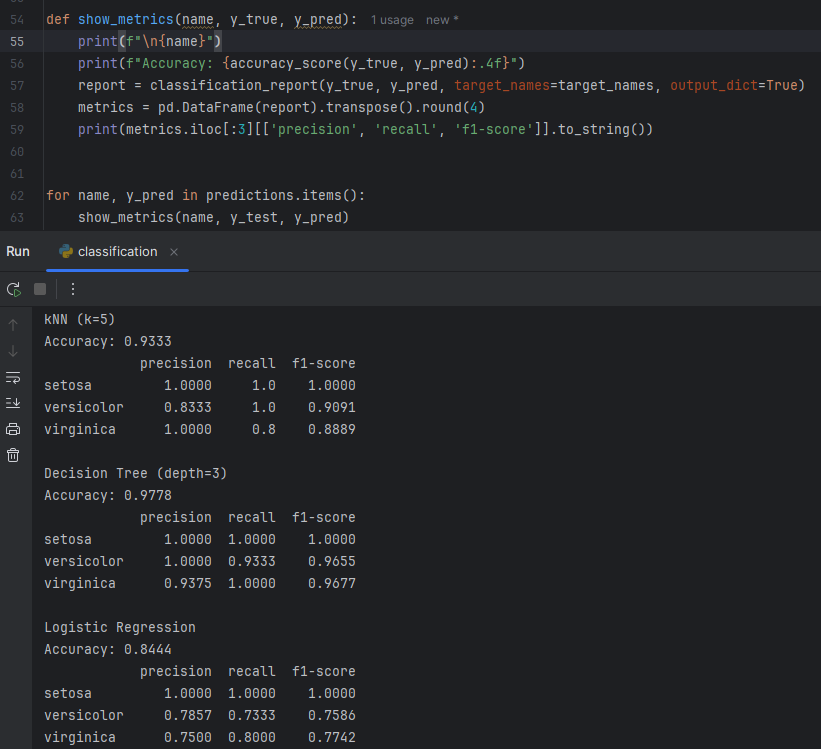


Рисунок 7.5 – Виведення показників ефективності кожного з методів

6. Порівняти їх, вказати переваги та недоліки для кожного з них.

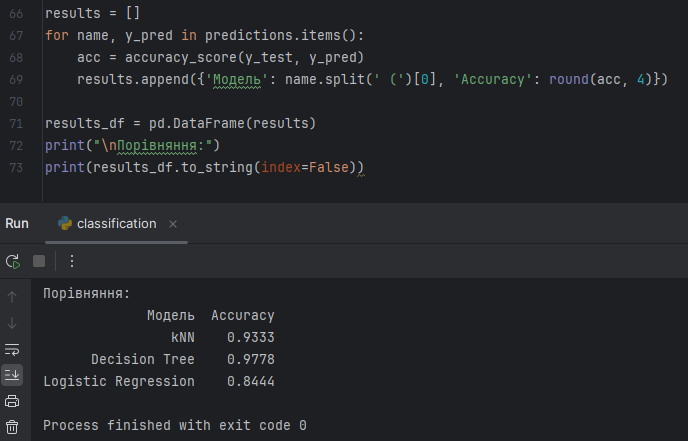


Рисунок 7.6 – Порівняння моделей

Із проведеного експерименту, можна зробити висновок, що Decision Tree має найвищу точність. Ще одна перевага – обробляє нелінійні залежності без проблем. Недоліком є те що може перенавчитися (тому в коді є обмеження max\_depth=3) та при глибокому дереві – складно читати. KNN отримав теж гарний результат однак поступився Decision Tree за точністю. Перевагою є простота реалізації і немає потреби в навчанні. Він також добре працює на малих, чітко розділених даних. Однак помиляється на перетинах класів і чутливий до вибірки. Також повільний при прогнозі (перевіряє всіх сусідів). Найгірший результат серед трьох – Logistic Regression. Він швидкий, добре узагальнює й працює з ймовірностями. Однак погано справляється з нелінійністю та потребує нормалізації.