**Міністерство освіти і науки України**

**Національний технічний університет України «КПІ» імені Ігоря Сікорського Кафедра обчислювальної техніки ФІОТ**

**ЗВІТ**

**з лабораторної роботи №5**

**з навчальної дисципліни «Технології Data Science»**

**Тема:**

**РЕАЛІЗАЦІЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ (MACHINE LEARNING (ML)**

**Виконав:**

Студент 4 курсу кафедри ФІОТ,

Навчальної групи ІП-11

Олександр Головня

**Перевірив:**

Професор кафедри ОТ ФІОТ Олексій Писарчук

**Київ 2024**

**І. Мета:**

Виявити дослідити та узагальнити особливості аналізу даних з використанням методів та технологій машинного навчання (Machine Learning (ML)

**ІІ. Завдання:**

Розробити програмний скрипт мовою Python що реалізує обчислювальний алгоритм машинного навчання (Machine Learning (ML)) відповідно до технічних умов:

**Група технічних вимог 1:** Реалізувати кластеризацію вхідних даних, отриманих Вами у ході виконання Дз 1, модельних та (або) реальних – на власний вибір. Методи Machine Learning з переліку: kmeans (k-середніх); Support Vector Machine (машина опорних векторів); k-nearest neighbors (найближчих сусідів); ієрархічна кластеризація – для кластеризації обраних даних обрати самостійно. Провести аналіз та пояснення отриманих результатів, сформувати висновки.

**Завдання І рівня складності 7 балів:** реалізувати на вибір **ОДНУ** з п’яти сформованих груп технічних вимог.

**ІІІ. Результати виконання лабораторної роботи.**

Блок схема алгоритму:

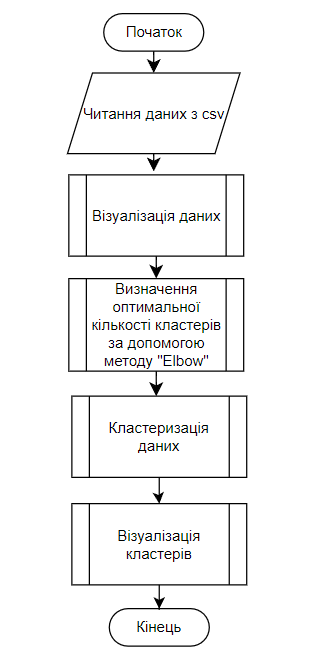


Рис.1 – Блок-схема алгоритму програми

* 1. **Опис структури проекту програми.**

Для реалізації розробленого алгоритму мовою програмування Python з використанням можливостей інтегрованого середовища сформовано проект.

Проект базується на лінійній бізнес-логіці функціонального програмування та має таку структуру.

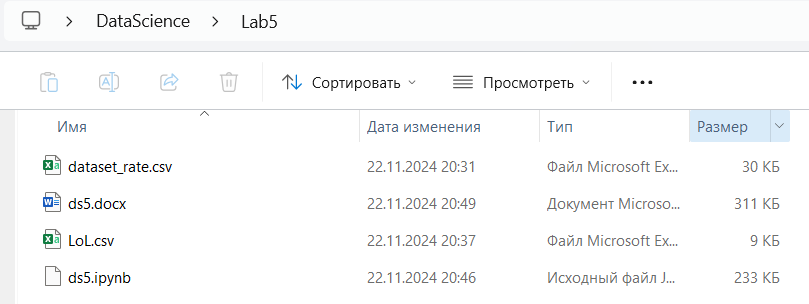


Рис.2 – Структура проєкту

ds5.ipynb – файл програмного коду лабораторної роботи;

ds5.docx – файл звіту лабораторної роботи

dataset\_rate.xlsx – dataset

* 1. **Результати роботи програми відповідно до завдання.**

Спочатку я знайшов на Kaggle датасет з реальними даними і створив відповідний csv файл

Длаі зчитаємо ці дані в датафрейм, що представляють собою інформацію по обмінному курсу Швейцарського франка до долара США:

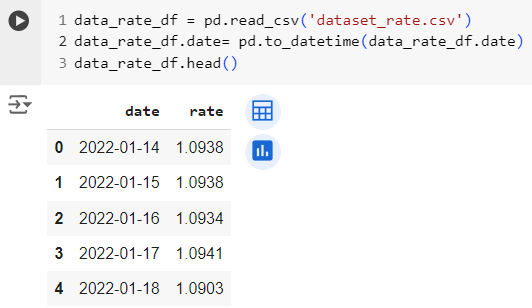


Рис 3.1 – Зчитування даних у датафрейм

Наступним кроком буде побудова діаграми розсіювання для аналізу даних:

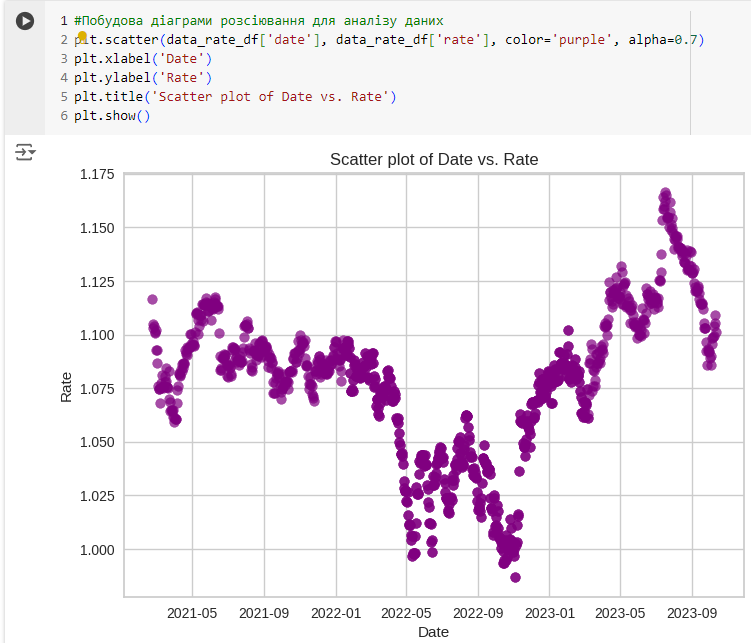


Рис. 3.2 – Візуалізація даних

Виконаємо задачу кластеризації за допомогою відомого методу k-means.

Загалом, дані не містять викидів або складних структур, тому, враховуючи це, для підбору оптимальної кількості кластерів можна використати Метод "Локтя" (Elbow Method)

Ідея: Побудувати графік залежності "внутрішньокластерної суми квадратів" (Inertia або Distortion) від кількості кластерів.

Як працює: Розрахувати Inertia для різних значень K.

Знайти точку, де графік змінює нахил (формує "лікоть"). Ця точка — оптимальна кількість кластерів.

Мінуси: Метод суб'єктивний, вимагає візуальної оцінки "ліктя".

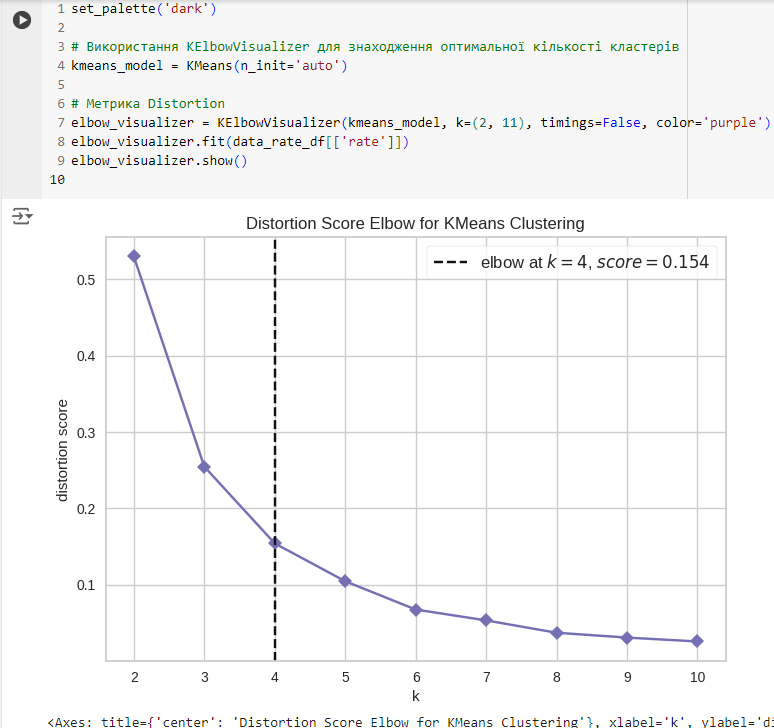


Рис. 3.3 – Метод ліктя

Оптимальне значення для кластера дорівнює 4,і оскільки значення показника зменшується не так інтенсивно, то використовувати більше кластерів немає сенсу. Варто зазначити, що цей метод не є універсальним і єдино правильним, для кожної задачі може бути використаний окремий підхід до визначення кластерів, прикладом таких можуть бути:

- Метод "Силуета" (Silhouette Method)

Ідея в оцінці, наскільки кожен об'єкт добре належить своєму кластеру та наскільки він віддалений від інших кластерів. Значення Silhouette Score коливається між -1 і 1. Значення ближче до 1 вказує на добре сформовані кластери.

* Критерій Девіса-Болдіна (Davies-Bouldin Index)

Оцінюється відношення розкиду в межах кластеру до відстані між кластерами. Чим менше значення Davies-Bouldin Index, тим краще сформовані кластери.

Для данної задачі достатньо і методу Ліктя, виконаємо кластеризацію на основі 4 кластерів

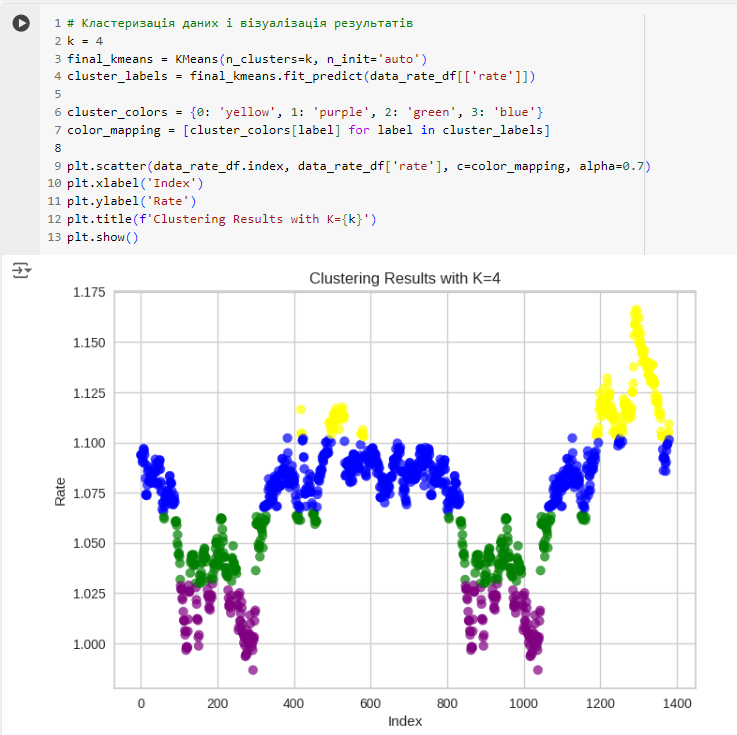


Рис. 3.7 – Візуалізація кластерів

Тепер,коли у нас є оптимальна кількість кластерів, ми навчили алгоритм k-means, вказавши саме цю кількість кластерів(4). На наступному кроці ми використали навчену модель, щоб визначити, до якого кластеру належить кожна точка даних, і візуалізували результати кластеризації, використовуючи різні кольори для кожного кластера.

* 1. **Програмний код.**

Програмний код послідовно реалізує алгоритм рис.1 та спрямовано на отримання результатів, поданих вище.

При цьому використано можливості Python бібліотек: pip; pandas; numpy; sklearn; matplotlib.

Контексні коментарі пояснюють сутність окремих скриптів наведеного коду програми.

* 1. **Аналіз результатів відлагодження та верифікації результатів роботи програми.**

Результати відлагодження та тестування довели працездатність розробленого коду. Це підтверджується результатами розрахунків, які не суперечать теоретичним положенням.

Верифікація функціоналу програмного коду, порівняння отриманих результатів з технічними умовами завдання на лабораторну роботу доводять, що усі завдання виконані у повному обсязі.

**IV. Висновки.**

Отже, в ході даної лабораторної я застосував на практиці навчики кластеризації даних, зокрема використав алгоритм K-means на реальних даних, підібравши оптимальну кількість кластерів, використовуючи метод Ліктя. Після отримання 4 кластерів я візуалізував ці кластери, використовуючи діаграму розсіювання та різні кольори для різних кластерів.