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 «Вступ до технології Data Science»

Тема:

РЕАЛІЗАЦІЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ (MACHINE LEARNING (ML))

Виконав:
Студент 3 курсу кафедри ІПІ ФІОТ,
Навчальної групи ІП-11
Сідак К.І.

Перевірив:
Професор кафедри ОТ ФІОТ
Писарчук О.О.

Київ 2023

I. Мета роботи:

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

II. 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: k-means (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 I рівня складності 7 балів: реалізувати на вибір **ОДНУ** з п’яти сформованих груп технічних вимог.

III. Виконання лабораторної роботи.

3.1. Зчитування даних та їх візуалізація

Для виконання даної лабораторної роботи я використав реальні дані по щоденному курсу франку до долара. Оскільки ці дані вже були завантажені у відповідний csv файл, то для отримання даних я зчитав цей файл у датафрейм.

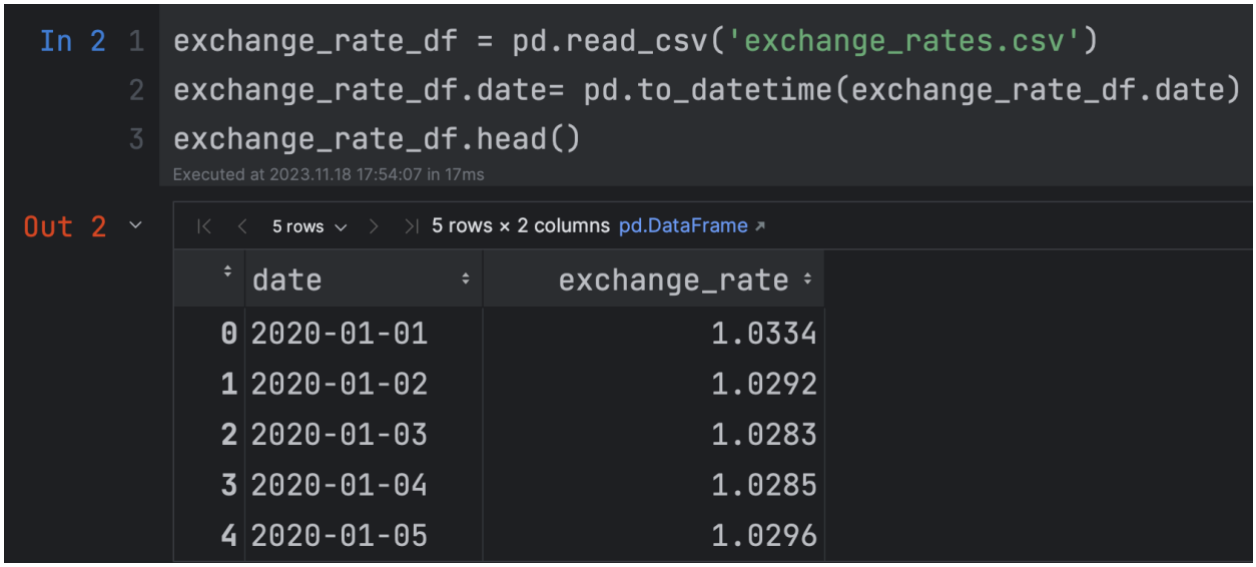

Рис 3.1 – Зчитування даних у датафрейм

Наступним кроком я візуалізував отримані дані у вигляді діаграми розсіювання.

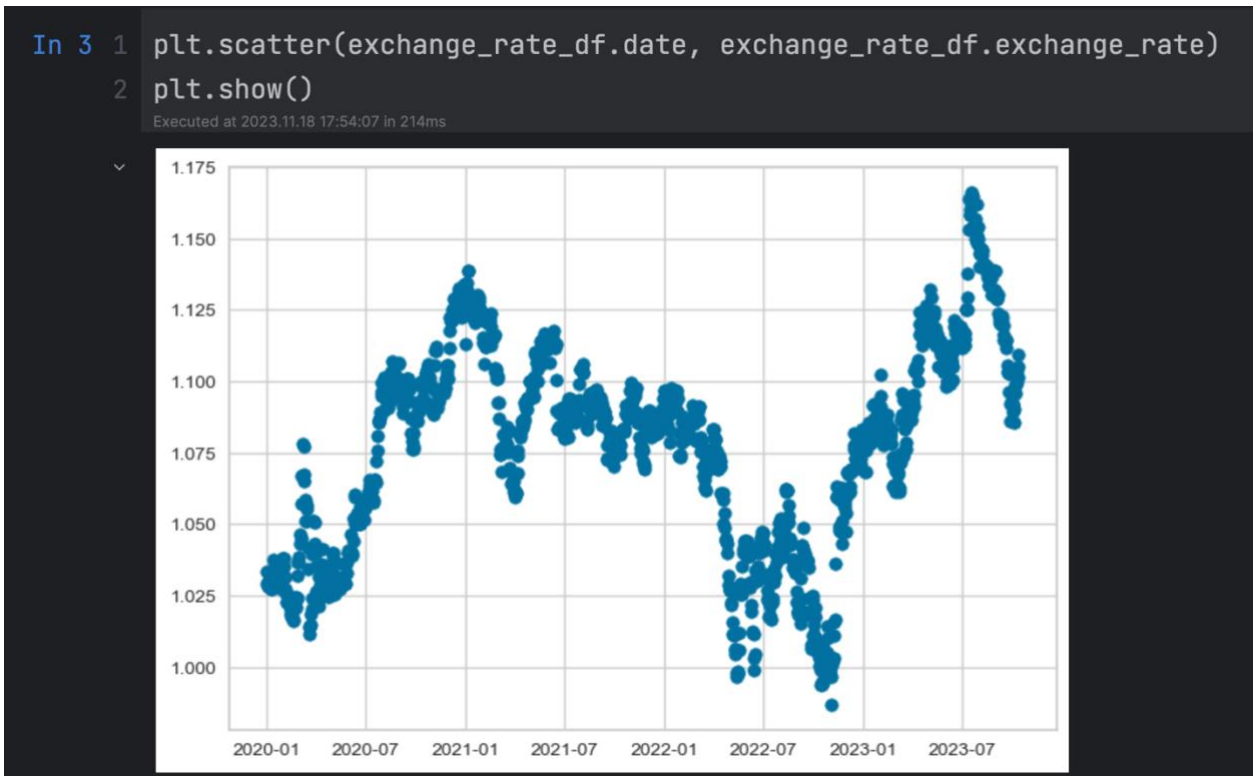

Рис. 3.2 – Візуалізація даних

Я вирішив використати метод k-means для кластеризації даних через його простоту та швидкість роботи, враховуючи, що він підходить для наших даних, бо загалом дані не мають як таких викидів або якоїсь складно структури. Варто зазначити, що цей алгоритм потребує задання кількості кластерів перед навчанням наперед. Можна побачити, що явно по графіку складно виділити певну кількість кластерів, тому потрібно буде використати інші підходи для визначення оптимальної кількості кластерів. Я підбірав оптимальне значення серед чисел від 2 до 10 включно.

3.2. Підбір оптимальної кількості кластерів

Для підбору оптимальної кількості кластерів я спершу використав метод ліктя на основі метрики WCSS (withing-cluster sum of squares), також відомою як distortion score.

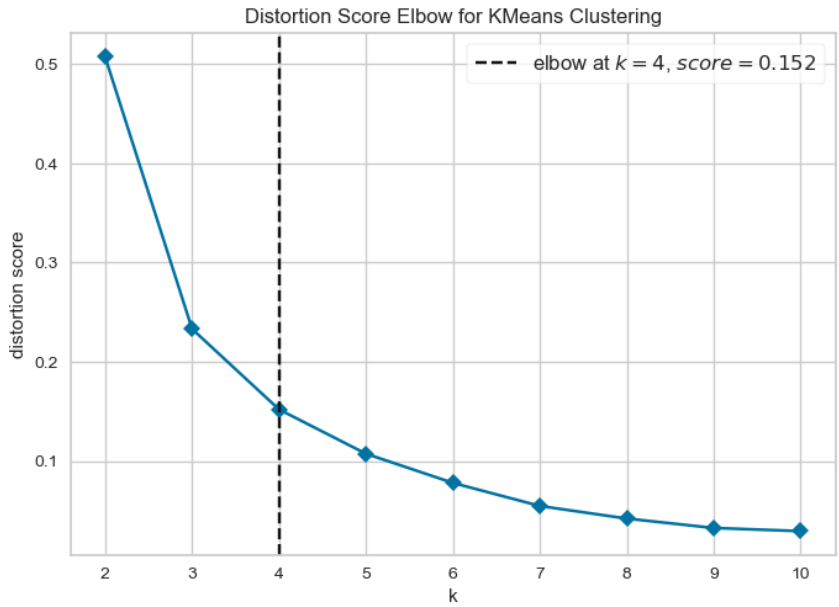

Рис. 3.3 – Метод ліктя на основі WCSS

Можна побачити, що за цим методом оптимальним значенням K є 4, тобто значення метрики вже не так суттєво зменшується, тому брати більшу кількість кластерів немає сенсу.

Варто зазначити, що даний метод не є універсальним та єдино правильним, тому я використав також інші підходи, зокрема підбір оптимальної кількості кластерів на основі значення силуетного коефіцієнту.

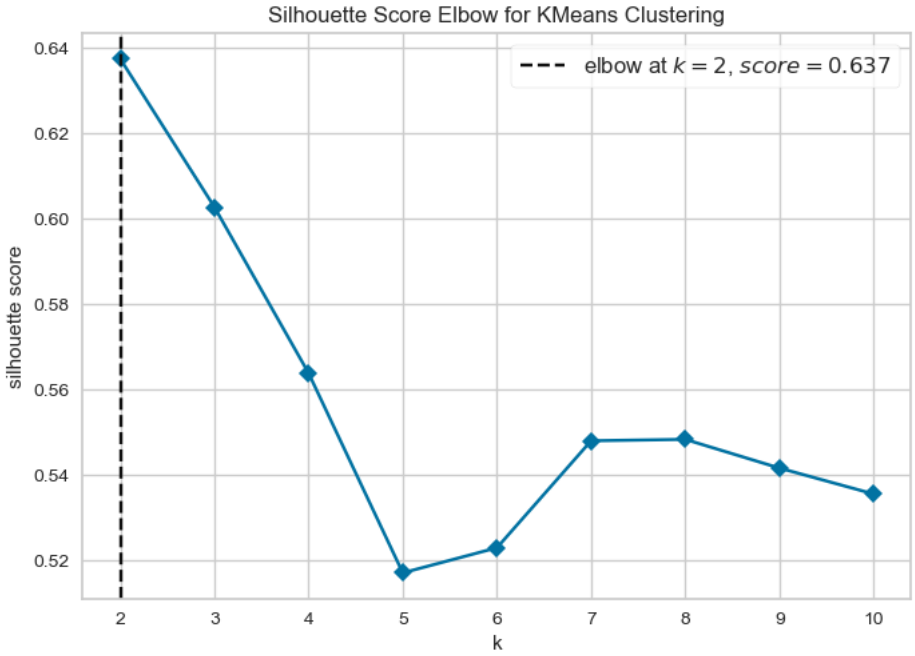

Рис. 3.4 – Значення силуетного коефіцієнту в залежності від кількості кластерів

З графіку видно, що найбільше значення даного коефіцієнту досягається при кількості кластерів рівним 2.

Ще один підхід, який я використав, це підбір кількості кластерів на основі метрики індексу Девіса Боулдіна. Чим менше значення цієї метрики, тим кращою є кластеризація.

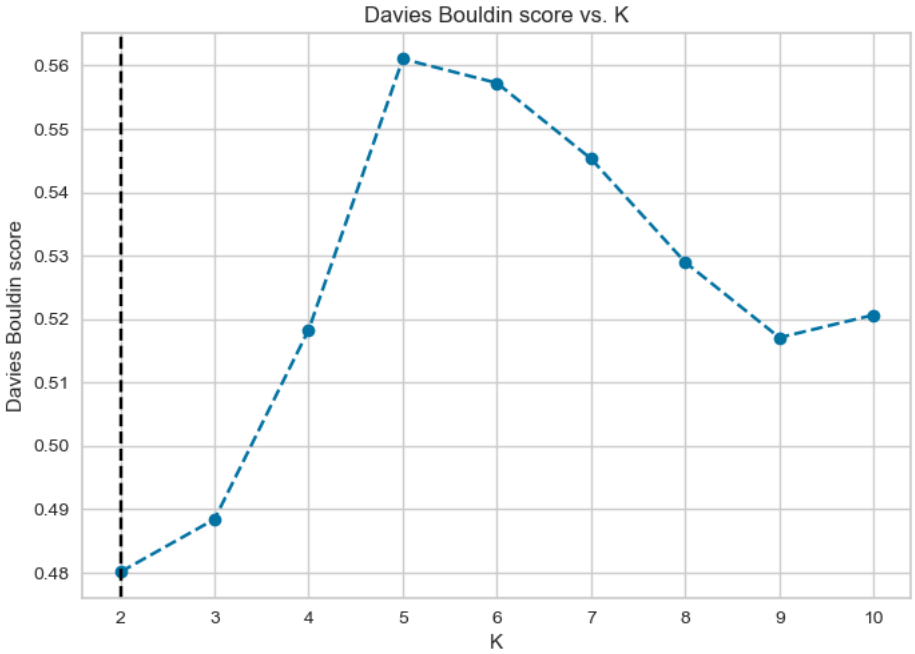

Рис. 3.5 – Значення індексу Девіса Боулдіна в залежності від кількості кластерів

За даного підходу оптимальним значенням K є 2.

Останній підхід, який я використав, це підбір кількості кластерів на основі метрики Calinski-Harabasz index. Чим більше її значення, тим кращою є кластеризація, проте я використав саме метод ліктя, адже для наших даних немає сенсу брати занадто велику кількість кластерів.

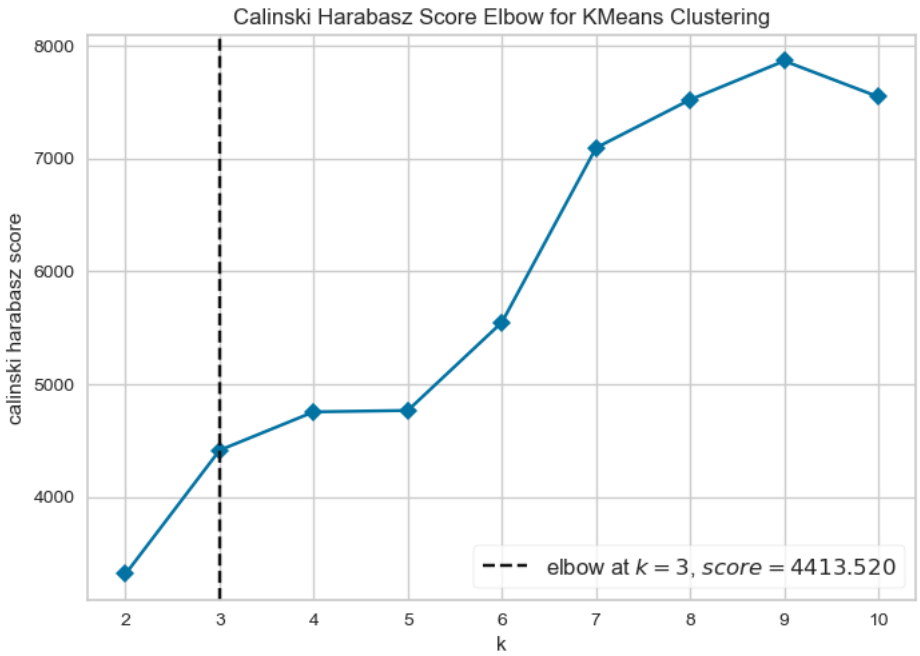

Рис. 3.6 – Значення індексу Девіса Боулдіна в залежності від кількості кластерів

Можна побачити, що збільшення даної метрики не є таким суттєвим після трьох кластерів (до 5), а 6 кластерів та більше вже занадто багато для наших даних.

Різні методи підбору кількості кластерів дали різні значення, тому я вирішив брати середнє значення по чотирьом підходам, що становить 2,75. Таким чином, округливши це значення, я отримав оптимальну кількість кластерів рівну 3.

3.3. Візуалізація результатів кластеризації даних

Отримавши оптимальну кількість кластерів рівну трьом, я навчив алгоритм K-means, задавши саме цю кількість кластерів. Наступним кроком я за допомогою навченої моделі я поставив номер кластера у відповідність кожному датапоінту та візуалізував кластери, де кожний колір позначав окремий кластер.

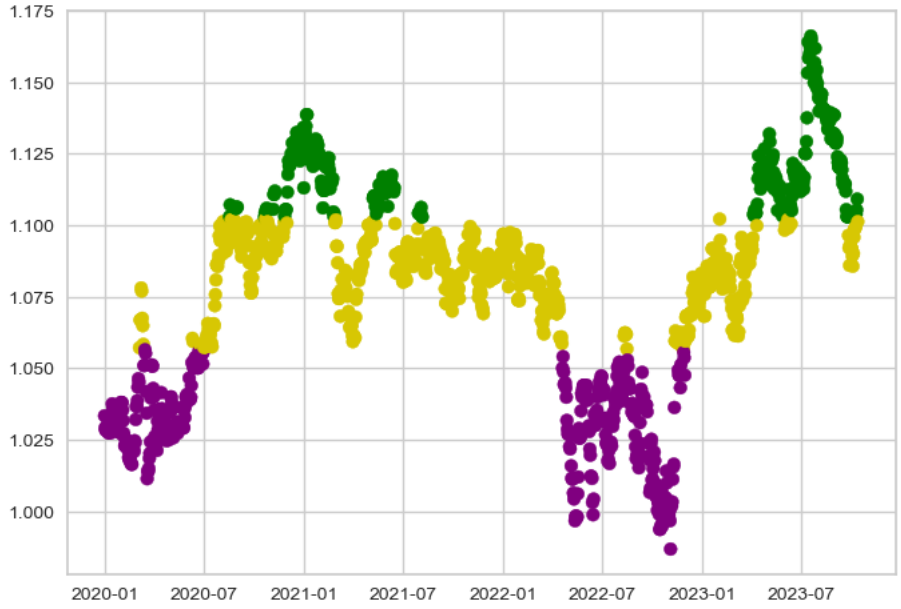

Рис. 3.7 – Візуалізація кластерів

Отже, дані було кластеризовано у три чітко відокремлені кластери.

IV. Висновок

Отже, в ході даної лабораторної я застосував на практиці навчики кластеризації даних, зокрема використав алгоритм K-means на реальних даних, отриманих у ході першої лабораторної роботи та підібрав оптимальну кількість кластерів, використовуючи 4 різних методи. Після отримання трьох кластерів я візуалізував ці кластери, використовуючи діаграму розсіювання та різні кольори для різних кластерів.