Міністерство освіти і науки України

Національний технічний університет України «Київський політехнічний

інститут імені Ігоря Сікорського"

Факультет інформатики та обчислювальної техніки

Кафедра інформатики та програмної інженерії

Звіт

з лабораторної роботи № 4 з дисципліни

«Основи науки про дані»

«Розв’язування задачі лінійної регресії»

Виконав студент \_\_\_\_ІП-11 Трикош І. В\_\_\_\_

Перевірив \_\_\_Новотарський М. А.\_\_\_

Київ 2024

**Завдання**

1. Детально розглянути наведені теоретичні відомості до даної лабораторної роботи.

2. Ознайомитися з кодом, що наведений у Jupyter notebook DD\_lab4.ipynb.

3. Завантажити набір даних відповідно до номеру у списку з таблиці 4.

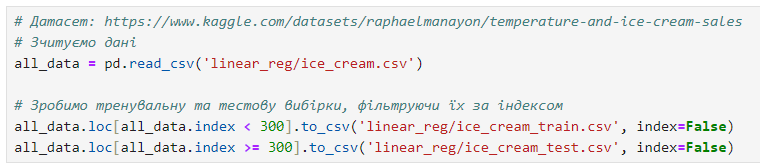
4. Вирішити задачу лінійної регресії, спираючись на приклад, який описаний у теоретичних відомостях та наведений у ноутбуці DD\_lab4.ipynb.

**Виконання**

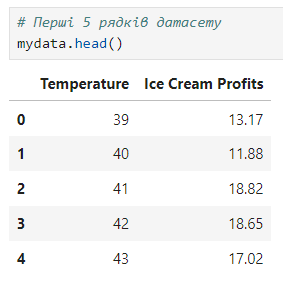
Для виконання лабораторної роботи візьмемо датасет про температуру та дохід від морозива: <https://www.kaggle.com/datasets/raphaelmanayon/temperature-and-ice-cream-sales/data>

Перший стовпець датасету – значення температури (у Фаренгейтах), при якій продавалося морозиво та відповідні значення прибутку (у доларах США) від продажів морозива.

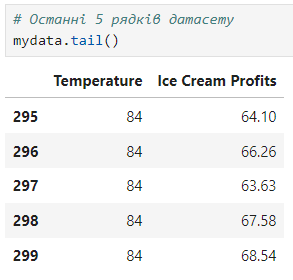
Датасет містить лише один файл, тому нам треба розділити його на два (тренувальні та тестові дані):



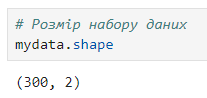
Виведемо перші 5 рядків тренувальних даних з допомогою методу head:



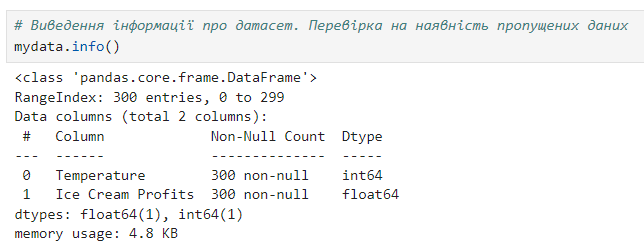
Виведемо останні 5 рядків тренувальних даних з допомогою методу tail:



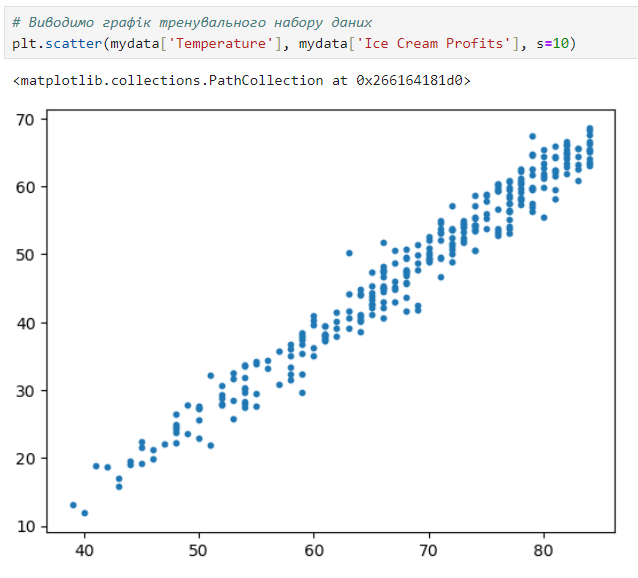
Переглянемо розмір набору даних, застосувуючи атрибут датафрейму shape:



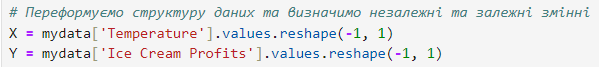
Переглянемо інформацію про датасет з допомогою методу info. Бачимо, що датасет не містить пропущений значень, тому можна продовжувати роботу:



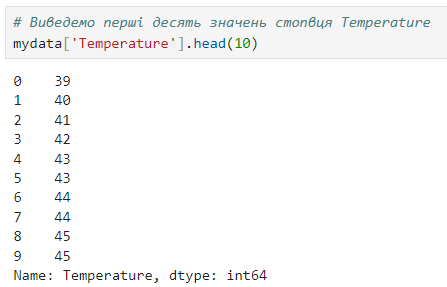
Побудуємо графік тренувального набору даних, застосовуючи метод scatter (створює діаграму розсіювання, або ж точкову діаграму) та його параметр s (розмір маркера/точки):



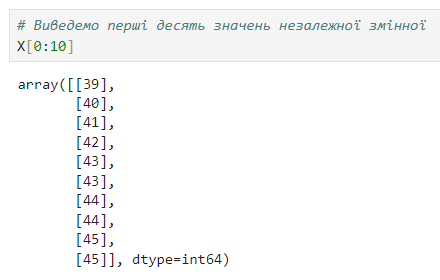
З тренувального датасету сформуємо набір незалежних змінних (факторів) та залежних змінних (відгуків). Використаємо атрибут values для отримання значень стовпця і метод reshape з параметрами -1, 1 для отримання набору змінних у вигляді двовимірного масиву з кількістю стовпців 1 та кількістю рядків, що дорівнює кількості рядків у тренувальному датасеті:



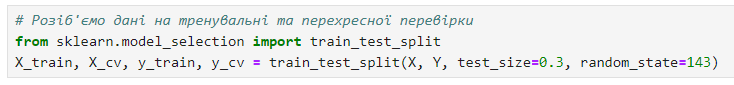
Для наочності виведемо перші десять значень стовпця Temperature:



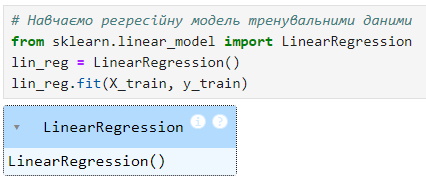
А тепер виведемо перші десять значень Х. Бачимо, що значення такі самі, але мають дещо іншу форму:



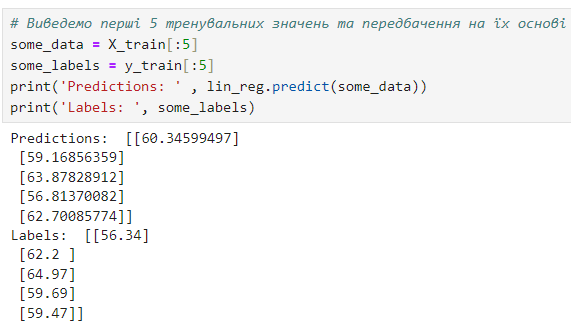
Розіб’ємо тренувальний датасет на тренувальні дані та дані для перехресної перевірки. Для цього використаємо метод train\_test\_split, що має параметри: незалежна змінна, залежна змінна, розмір тестового набору та значення зерна для генератора випадкових чисел.



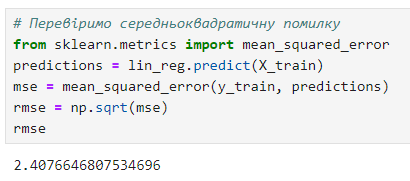
Навчимо модель лінійної регресії на тренувальних даних, використовуючи метод fit, що приймає незалежну та залежну змінні як параметри:



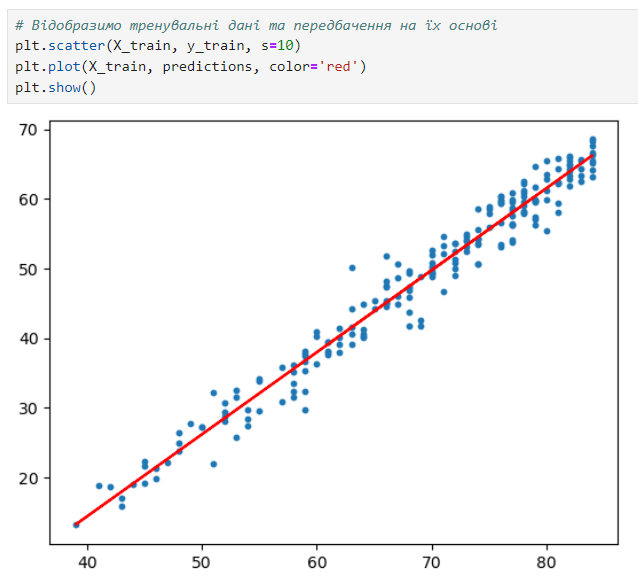
Для перевірки успішності навчання, перевіримо прогноз для перших 5 тренувальних значень, використавши метод predict, який приймає незалежну змінну як параметр. Бачимо, що передбачення доволі близькі до реальних значень:



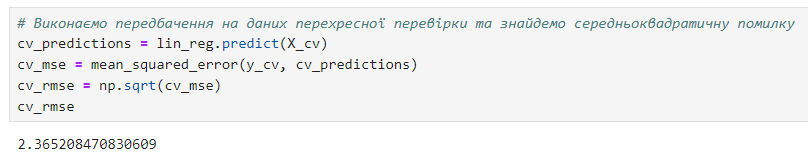
Перевіримо середньоквадратичну помилку, використавши метод mean\_squared\_error і потім взявши квадратний корінь з результату. Бачимо, що помилка не дуже велика:



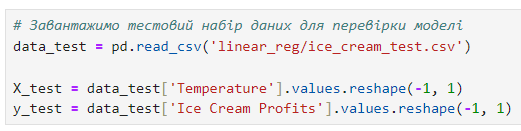
Відобразимо результат навчання моделі на тренувальних даних (червона лінія) та самі тренувальні дані (позначено синім). Бачимо, що навчена модель добре відповідає тренувальним даним:



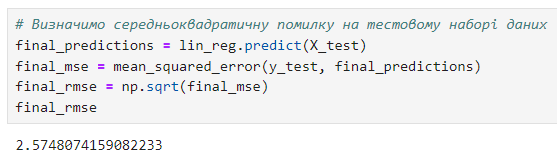
Тепер перевіримо середньоквадратичну помилку на даних для крос-валідації. Бачимо, що вона також незначна:



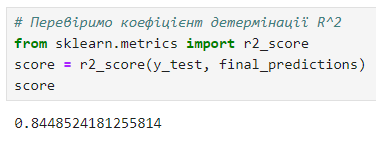
Тепер перевіримо модель на тестовому наборі даних. Спочатку імпортуємо його та визначимо незалежні та залежні змінні:



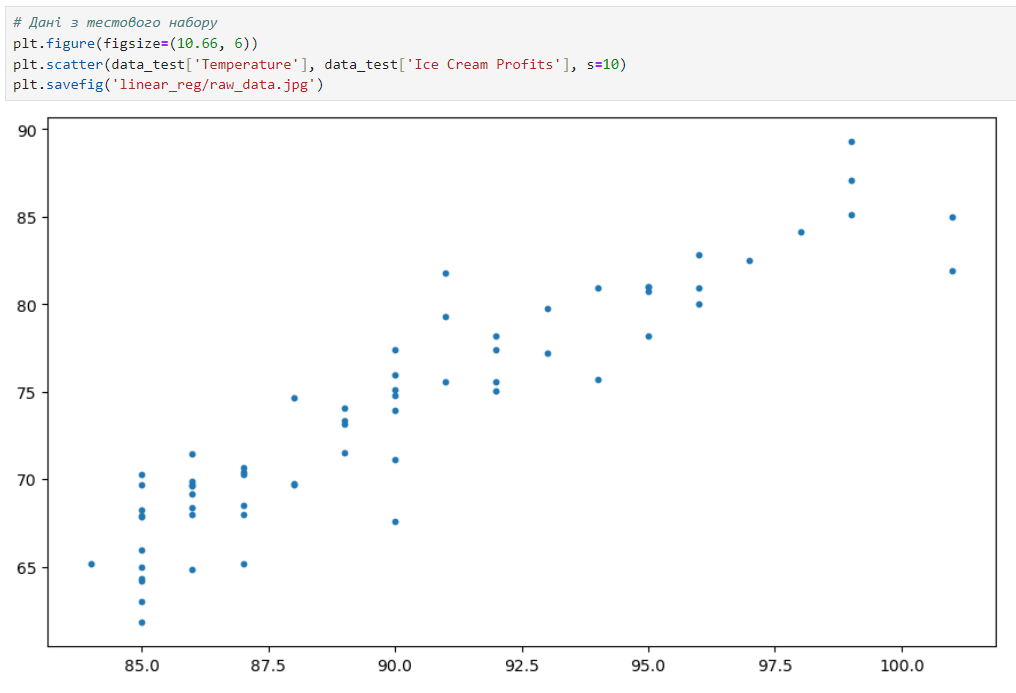
Обчислимо середньоквадратичну помилку на тестовому наборі даних. Тут вона теж незначна, хоча трохи більша, ніж на тренувальному наборі:



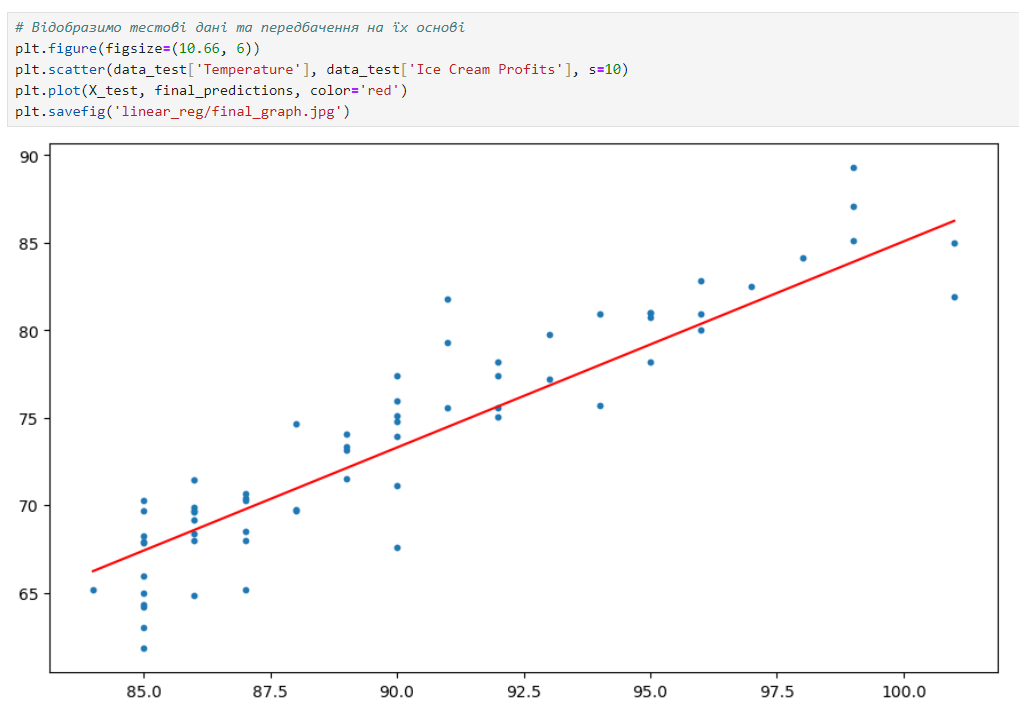
Обчислимо ще один показник якості моделі – коефіцієнт детермінації R2, використавши метод r2\_score. Бачимо, що він доволі високий, тому модель навчена добре:



Виведемо дані з тестового набору:



Виведемо тестові дані та передбачення на їх основі. Бачимо, що навчена модель гарно відповідає тестовим даним:



**Висновки**

Під час виконання лабораторної роботи я навчився працювати з простою моделлю лінійної регресії на прикладі даних про температуру і прибутками від продажів морозива при цій температурі і встановив, що є кореляція між цими показниками.