Міністерство освіти і науки України

Західноукраїнський національний університет

Факультет комп’ютерних інформаційних технологій

Кафедра інформаційно-обчислювальних систем і управління

Звіт

Про виконання лабораторної роботи №1

з дисципліни « Методи та системи штучного інтелекту »

Виконав:

Студент групи КНШІ-31

Семенюк Руслан

Тернопіль 2025

Хід роботи

Варіант: 18

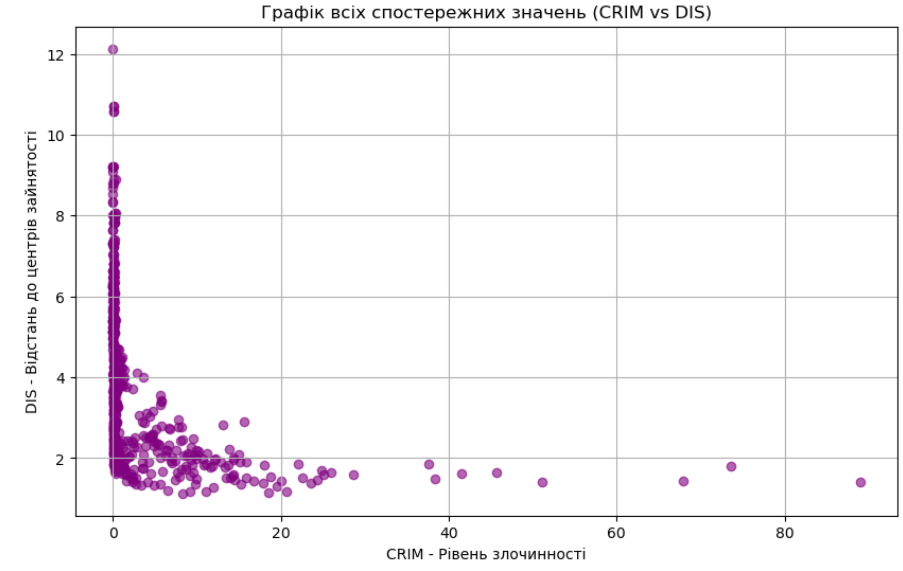
Тема: лінійна регресія: апроксимація, інтерполяція та екстраполяція даних

Мета: практично засвоїти побудову регресійної моделі шляхом повного перебору коефіцієнтів, знайти оптимальні b0​ та b1​, візуалізувати поверхню MSE та оцінити точність моделі на тестовій вибірці.

Варіант: СRIM – рівень злочинності (як незалежна змінна X), DIS – відстань до центрів зайнятості (як незалежну змінну Y)

Завдання 1. Завантажити дані із датасету і навести графік, множини спостережних значень

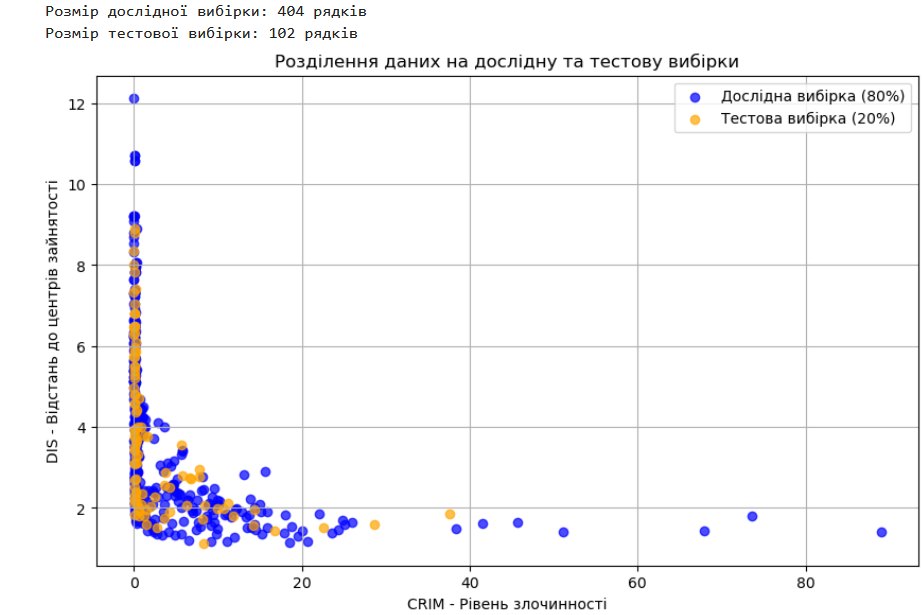




Коли Злочинність (CRIM) зростає , Відстань до роботи (DIS) зазвичай зменшується Іншими словами, що ближче до ділових центрів, то вищий може бути рівень злочинності.

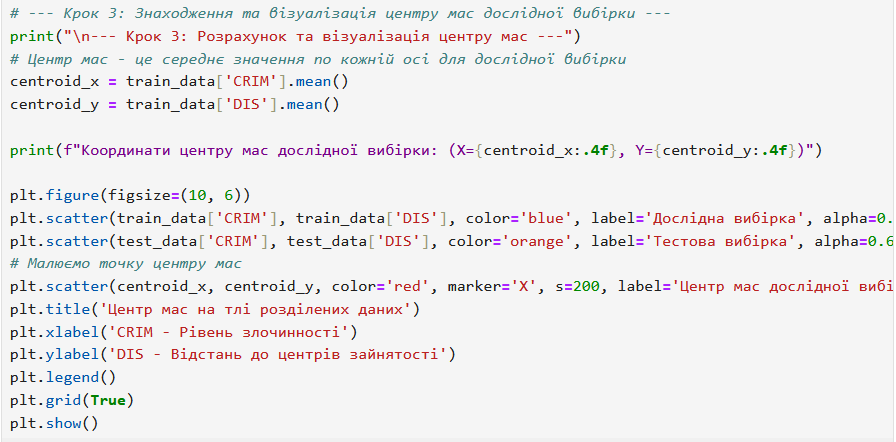
Завдання 2. Розділити вибірку на дві частини (дослідну і тестову) – представити графічно.

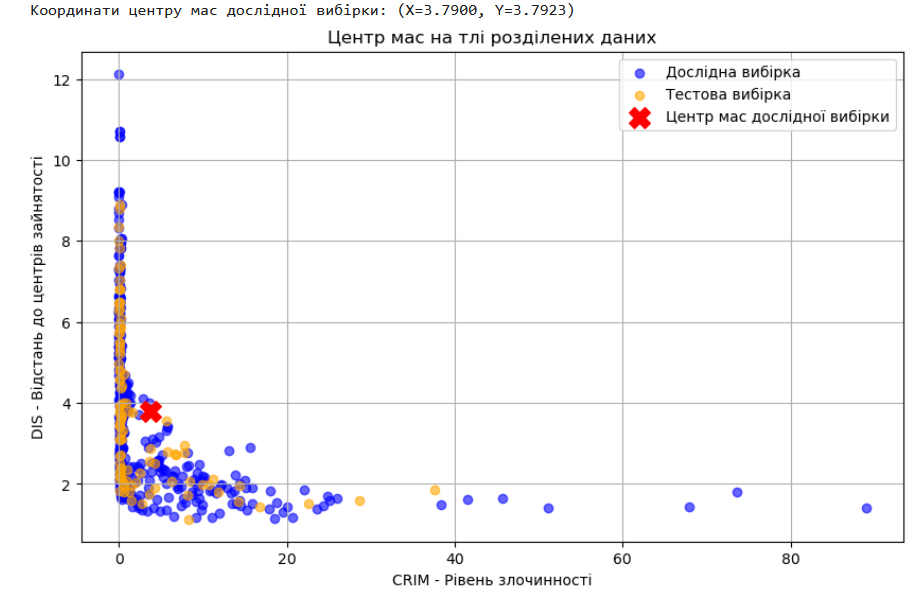




Сині (Дослідна вибірка): Це дані, які ми використовуємо для навчання нашої моделі (щоб знайти найкращу пряму). 80% даних. Помаранчеві (Тестова вибірка): Це дані, які модель ніколи не бачила. Ми використовуємо їх для перевірки моделі. 20% даних.

Завдання 3. Знайти точку центру мас дослідної вибірки даних (відобразити точку на графіку)





Розраховується центр мас (середнє арифметичне) дослідної вибірки окремо для змінної CRIM (xцентр​) та DIS (yцентр​).

Візуалізується положення розрахованого центру мас на графіку, що містить розділені дослідну та тестову вибірки.

Завдання 4. Дослідити лінійні регресії, що відображають множину даних дослідної вибірки:

a. навести аналітичну формулу, лінійної регресії та спосіб розрахунку коефіцієнтів;

Лінійна регресія описується прямою: Y= b1X + b0 де:

X – незалежна змінна (CRIM)

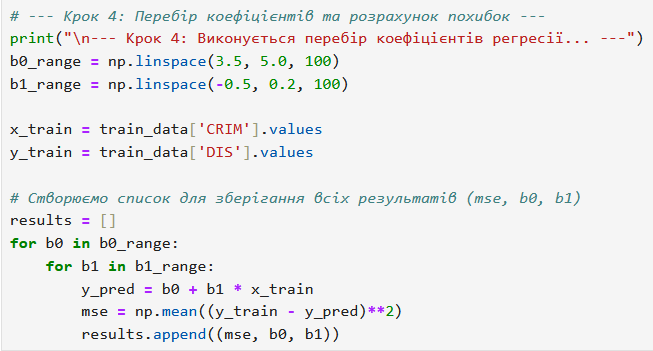
Y – залежна змінна (DIS)

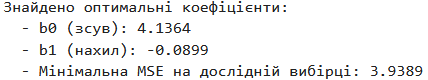
b1– нахил (коефіцієнт регресії)

b0– перетин з віссю Y

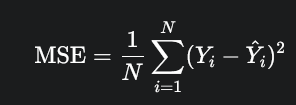
Виконується повний перебір усіх можливих комбінацій b1 та b0​ у заданих діапазонах.:

b. створити програмну реалізацію, що забезпечує перебір множини коефіцієнтів лінійної регресії (представити програмний код)





c. передбачити розрахунок сумарної похибки для кожної досліджуваної прямої;

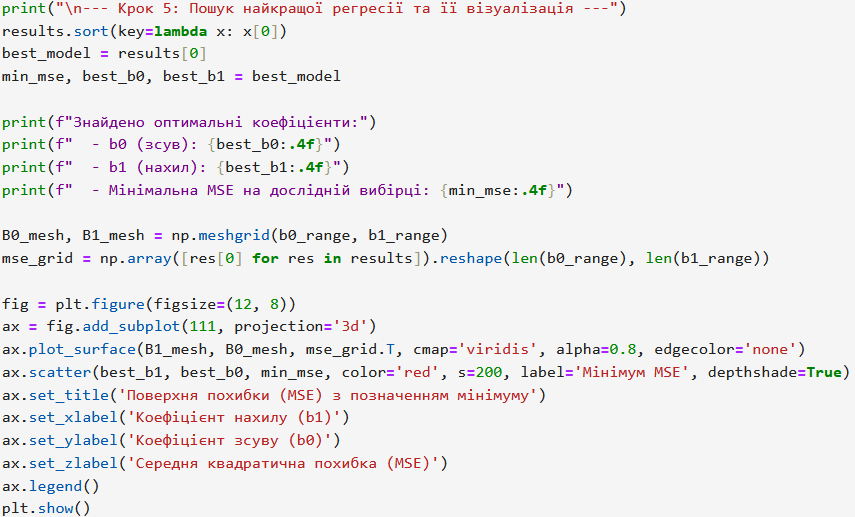


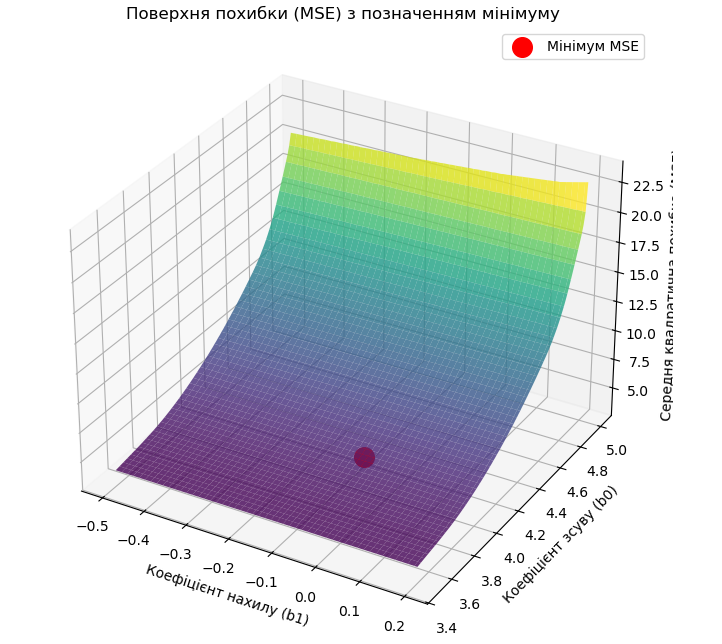
У коді вище ми для кожної прямої робимо:

1. Знаходиться похибка: Для кожної точки віднімаємо прогнозоване значення (Y^) від фактичного (Y).
2. Квадратування: Кожна похибка підноситься до квадрату (Y−Y^)2. Це робиться для того, щоб усунути від'ємні значення і сильніше штрафувати великі помилки.
3. Усереднення: Усі ці квадратні похибки додаються і діляться на загальну кількість точок N.

d. представити у звіті графік поверхні, що відображає розраховані похибки (MSE) від значень коефіцієнтів прямої.

Завдання 5. Забезпечити пошук та вибір лінійної регресії із найменшою сумарною похибкою для даних навчальної вибірки із датасету (позицію відобразити графічно)





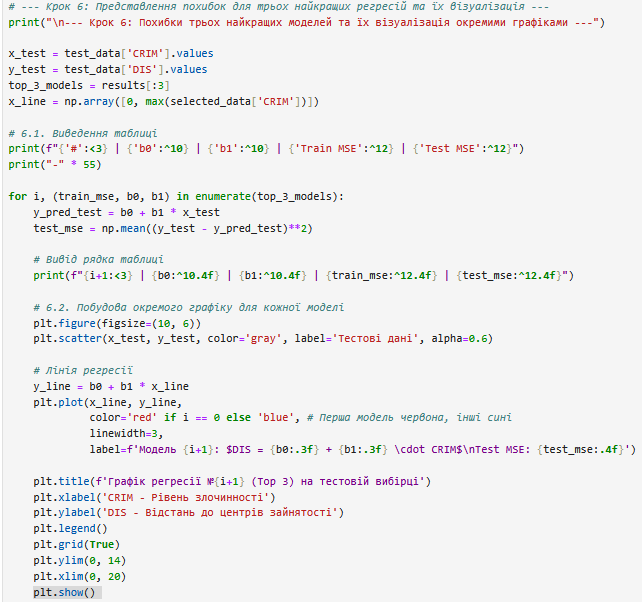
На графіку бачимо

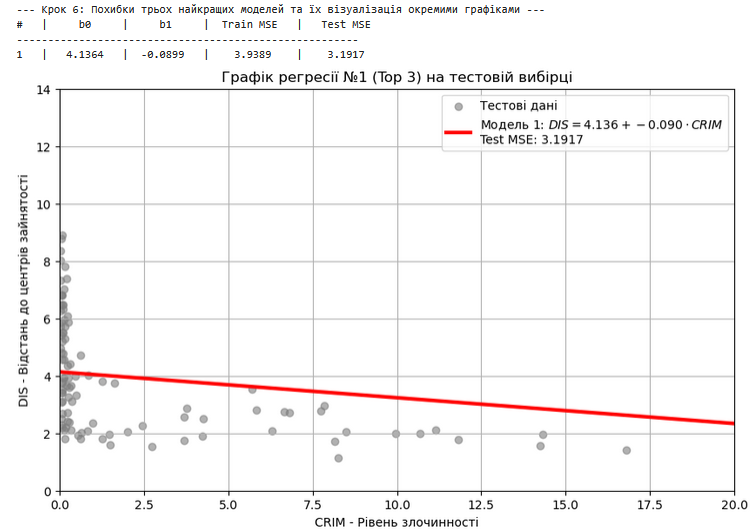
1. Дві осі внизу (b0​ і b1​) — це всі можливі коефіцієнти нашої прямої.
2. Вісь Z (висота) — це Похибка (MSE).

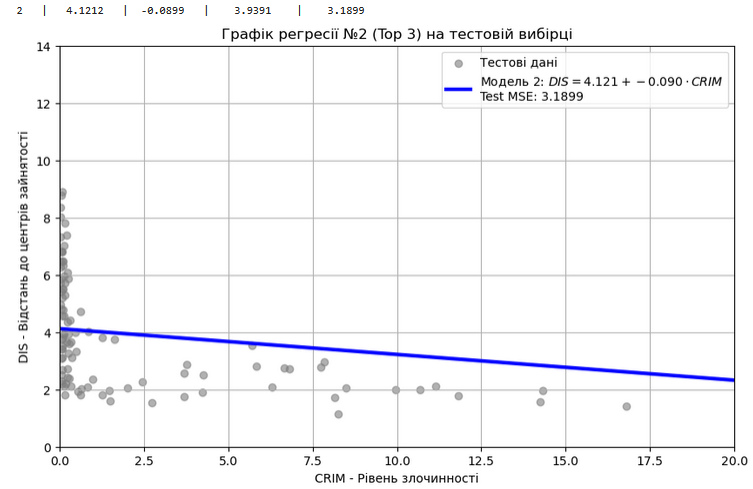
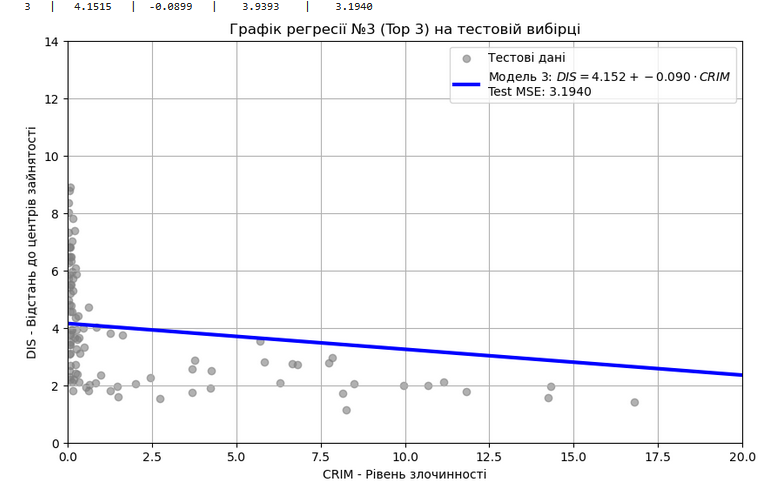
Що це означає:

1. Високі ділянки — це погані лінії регресії з великою похибкою.
2. Найнижча точка (позначена червоним) — це ідеальна пара коефіцієнтів (b0​ і b1​), яка дає найменшу похибку на дослідній вибірці.

Завдання 6. Представити похибки трьох лінійних регресій найближчих до оптимальної для даних тестової та дослідної вибірки.





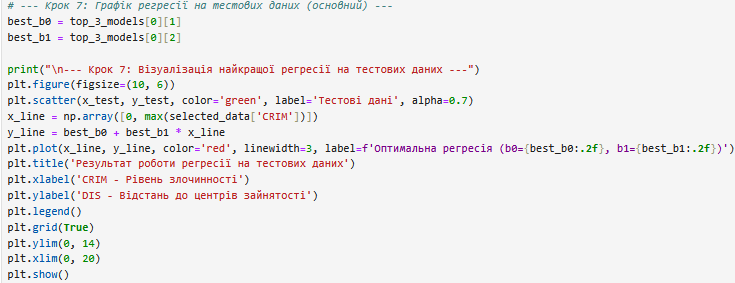
 

Три послідовні графіки. На кожному графіку є тестові дані (сірі точки) та одна пряма лінія (Модель 1, 2 або 3).

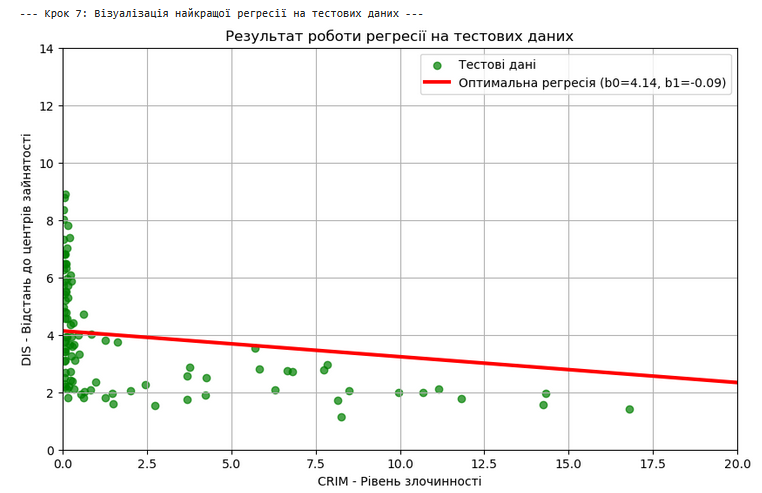
Це лінії регресії, які мають найменшу похибку на навчальній вибірці (три найкращі моделі).

Ці графіки дозволяють візуально порівняти, наскільки добре три майже ідентичні (за коефіцієнтами) моделі справляються з прогнозуванням невідомих тестових даних. Ці три лінії практично зливаються.

Завдання 7. Побудувати графік, що відображає набір тестових даних, лінійну регресію



Малюємо тільки тестові точки, накладаємо оптимальну лінію регресії



Лінія показує загальний спадний тренд, підтверджуючи обернену залежність, але відхилення точок вказують на помірну точність моделі.

Висновок: У ході виконання роботи здійснено аналіз залежності відстані до центрів зайнятості (DIS) від рівня злочинності (CRIM) на наборі даних Boston Housing. Проведено розділення вибірки та виконано повний перебір коефіцієнтів лінійної регресії, що дозволило візуалізувати поверхню похибки (MSE). Знайдено оптимальні коефіцієнти, які мінімізують MSE на навчальній вибірці. Встановлено, що між змінними існує обернена залежність (негативний коефіцієнт нахилу). Оцінено точність моделі на тестовій вибірці, при цьому зафіксовано помірну точність (MSE ≈3.19). Помірна точність пояснюється тим, що обрана однофакторна модель не враховує нелінійну природу даних та вплив інших значущих змінних.