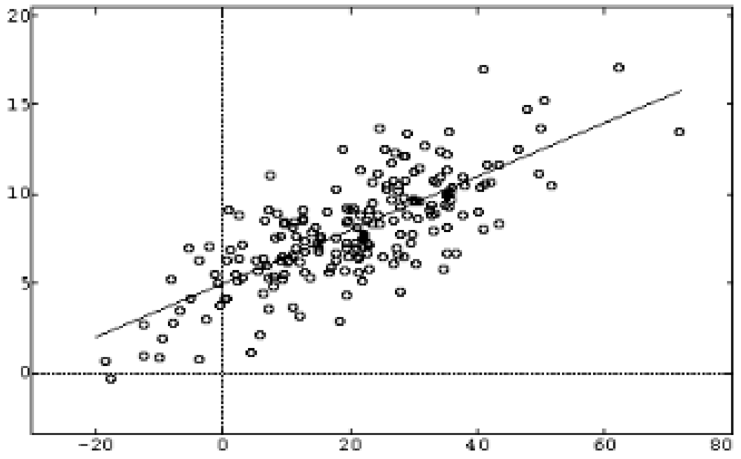
Побудова лінійної та нелінійної регресії за допомогою методу найменших квадратів

Підготував: студент ОІ-32
Криворучко Микола

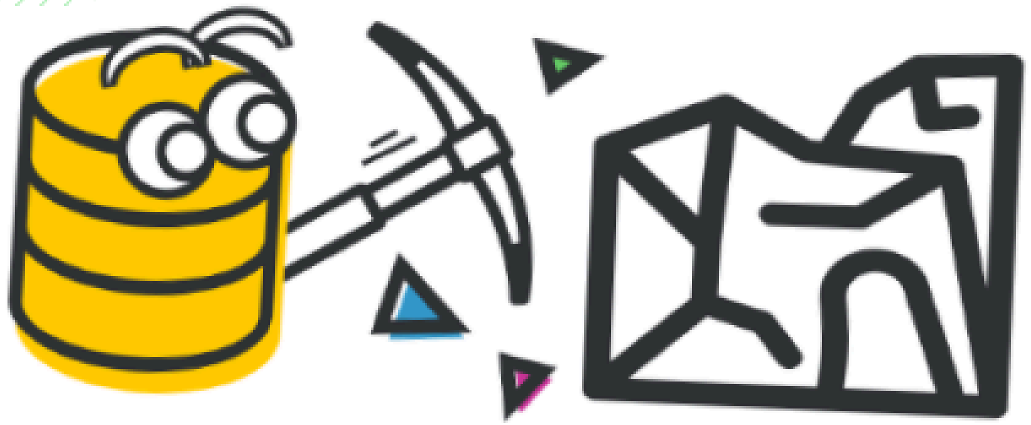
ЗМІСТ

1. Що таке Data Mining?
2. Які існують задачі Data Mining?
3. Задача регресії
4. Лінійна vs Нелінійна регресія
5. Метод найменших квадратів
6. Висновок



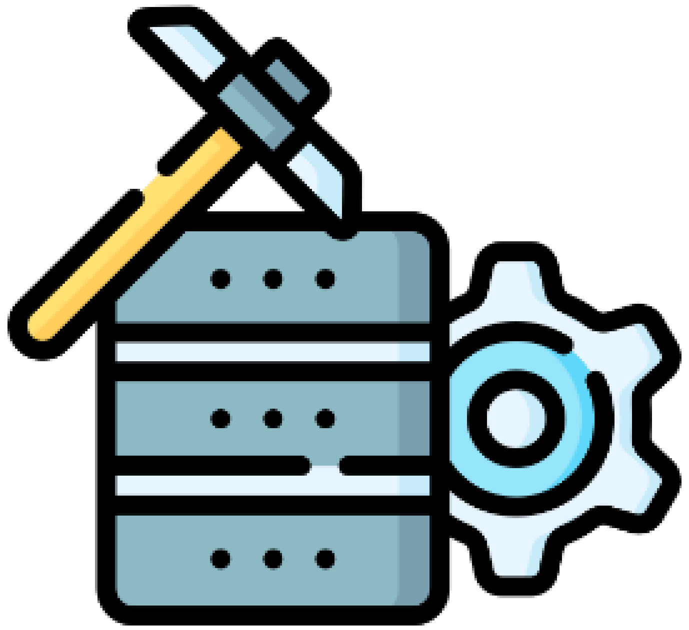
Що таке Data Mining?

Набір методів, алгоритмів та засобів опрацювання "сирих даних" із метою видобування з них необхідної інформації(знань)



Задачі Data Mining

1. Задача класифікації
2. Задача регресії
3. Задача кластеризації
4. Побудова асоціативних правил



Задача регресії

Задача регресії — це задача передбачення безперервного числового значення на основі однієї або кількох ознак. Мета — знайти функцію f(x), яка найкраще описує залежність між ознаками та відповіддю, зазвичай мінімізуючи помилку прогнозу (наприклад, середньоквадратичну).

Наприклад: передбачити ріст рослини за кількістю сонячних годин.

Лінійна vs Нелінійна регресія

Лінійна

передбачає залежність у вигляді прямої: $y=ax+b$.

VS

Нелінійна

передбачає криву або складну функцію: $y=b0+b1x + b2x^2 + \ldots.$

Метод найменших квадратів

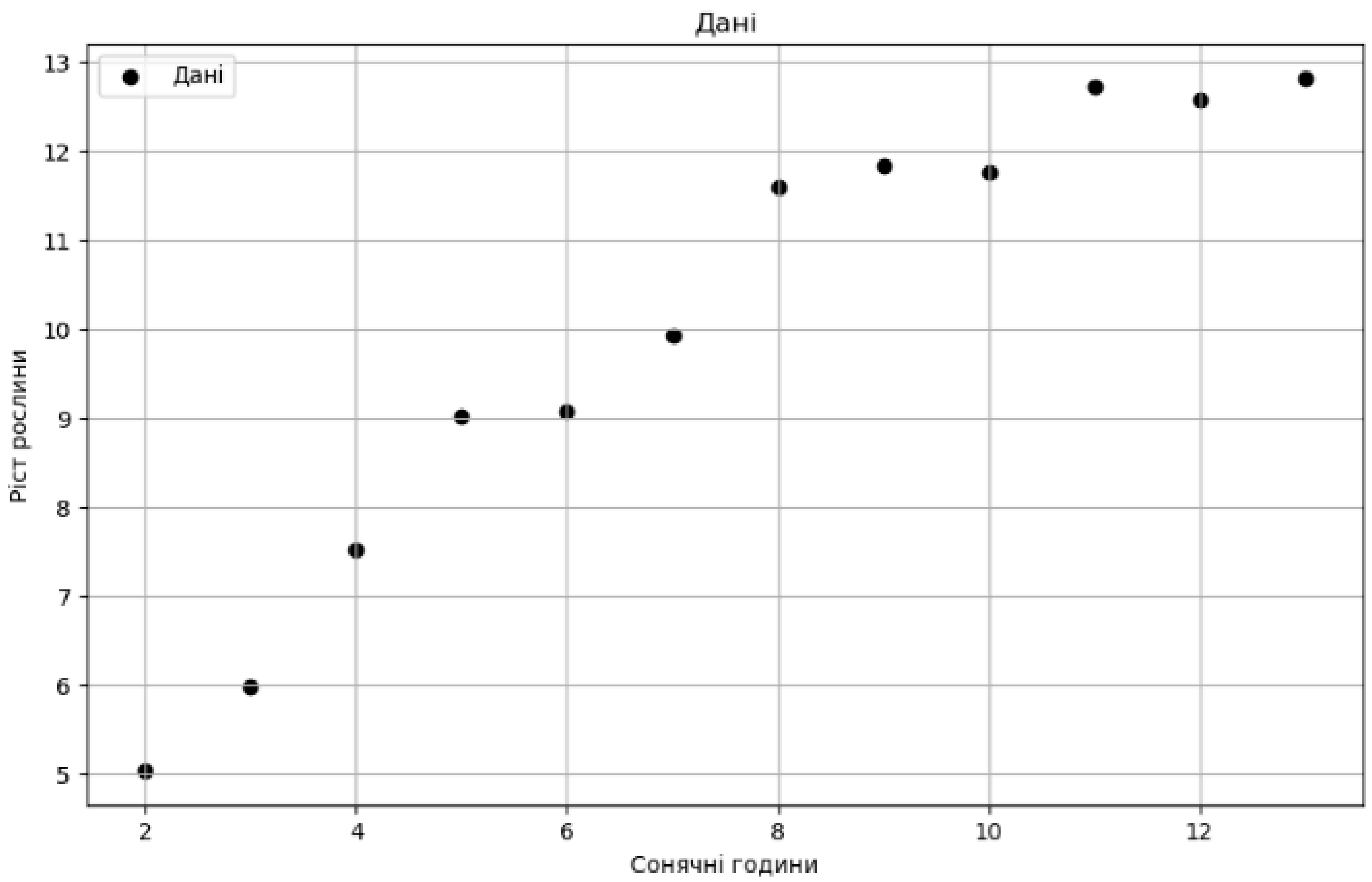
Метод найменших квадратів — це метод регресійного аналізу для оцінки невідомих величин за результатами вимірів, що містять випадкові помилки. Метод найменших квадратів також застосовується для наближеного представлення заданої функції іншими більш простими функціями і часто є корисним для обробки спостережень.

Задача

На скільки сантиметрів виросте рослина?

	Кількість сонячних годин на день	Ріст рослини за певний період(см)
0	2	5.048357
1	3	5.980868
2	4	7.523844
3	5	9.011515
4	6	9.082923
5	7	9.932932
6	8	11.589606
7	9	11.833717
8	10	11.765263
9	11	12.721280
10	12	12.568291
11	13	12.817135

Графік даних



Побудова Лінійної Регресії

Першим кроком є знаходження лінійної апроксимуючої прямої виду $y = ax + b$, де y – ріст рослини, а x – сонячні години. Для розрахунку коефіцієнтів a та b за методом найменших квадратів необхідно обчислити проміжні суми.

Побудова Лінійної Регресії

Розрахункова таблиця

i	x_i	y_i	$x_i y_i$	x_i^2
1	2	5.048357	10.096714	4
2	3	5.980868	17.942604	9
3	4	7.523844	30.095376	16
4	5	9.011515	45.057575	25
5	6	9.082923	54.497538	36
6	7	9.932932	69.530524	49
7	8	11.589606	92.716848	64
8	9	11.833717	106.503453	81
9	10	11.765263	117.652630	100
10	11	12.721280	139.934080	121
11	12	12.568291	150.819492	144
12	13	12.817135	166.622755	169

Побудова лінійної регресії

Сума (Σ)	90	119.875731	1001.469589	818

Кількість спостережень (n) = 12

Розрахунок коефіцієнтів

Використовуємо раніше наведені формули для a і b. Підставляємо обчислені суми:

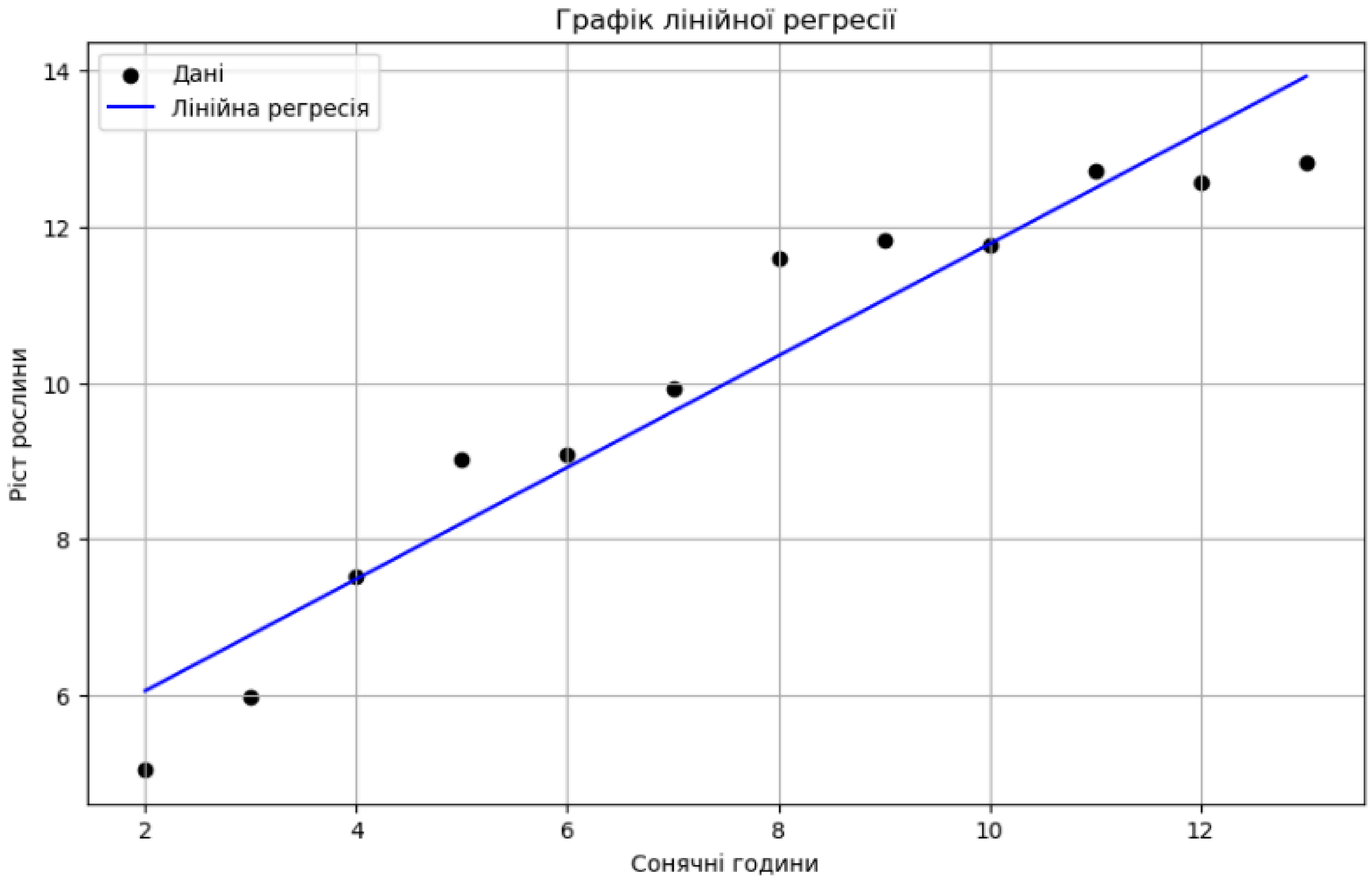
$a = (12 * 1001.4696 - 90 * 119.8757) / (12 * 818 - 90^2) \approx 0.716$

$b = (119.8757 - 0.716 * 90) / 12 \approx 4.619$

Рівняння лінійної регресії

Таким чином, рівняння лінійної регресії має вигляд: $y \approx 0.716x + 4.619$

Графік лінійної регесії



Побудова нелінійної регресії

Наступним кроком є знаходження квадратичної апроксимації виду $y = b_0 + b_1x + b_2x^2$, яка потенційно може краще описати нелінійний характер даних.
Для знаходження коефіцієнтів b_0, b_1 та b_2 необхідно розв'язати систему нормальних рівнянь.

Побудова нелінійної регресії

$$y = b_0 + b_1 x + b_2 x^2$$

$$\sum_i y_i = nb_0 + b_1 \sum_i x_i + b_2 \sum_i x_i^2$$
$$\sum x_i y_i = b_0 \sum_i x_i + b_1 \sum_i x_i^2 + b_2 \sum_i x_i^3$$
$$\sum_i x_i^2 y_i = b_0 \sum_i x_i^2 + b_1 \sum_i x_i^3 + b_2 \sum_i x_i^4$$

Результати розрахунку коефіцієнтів

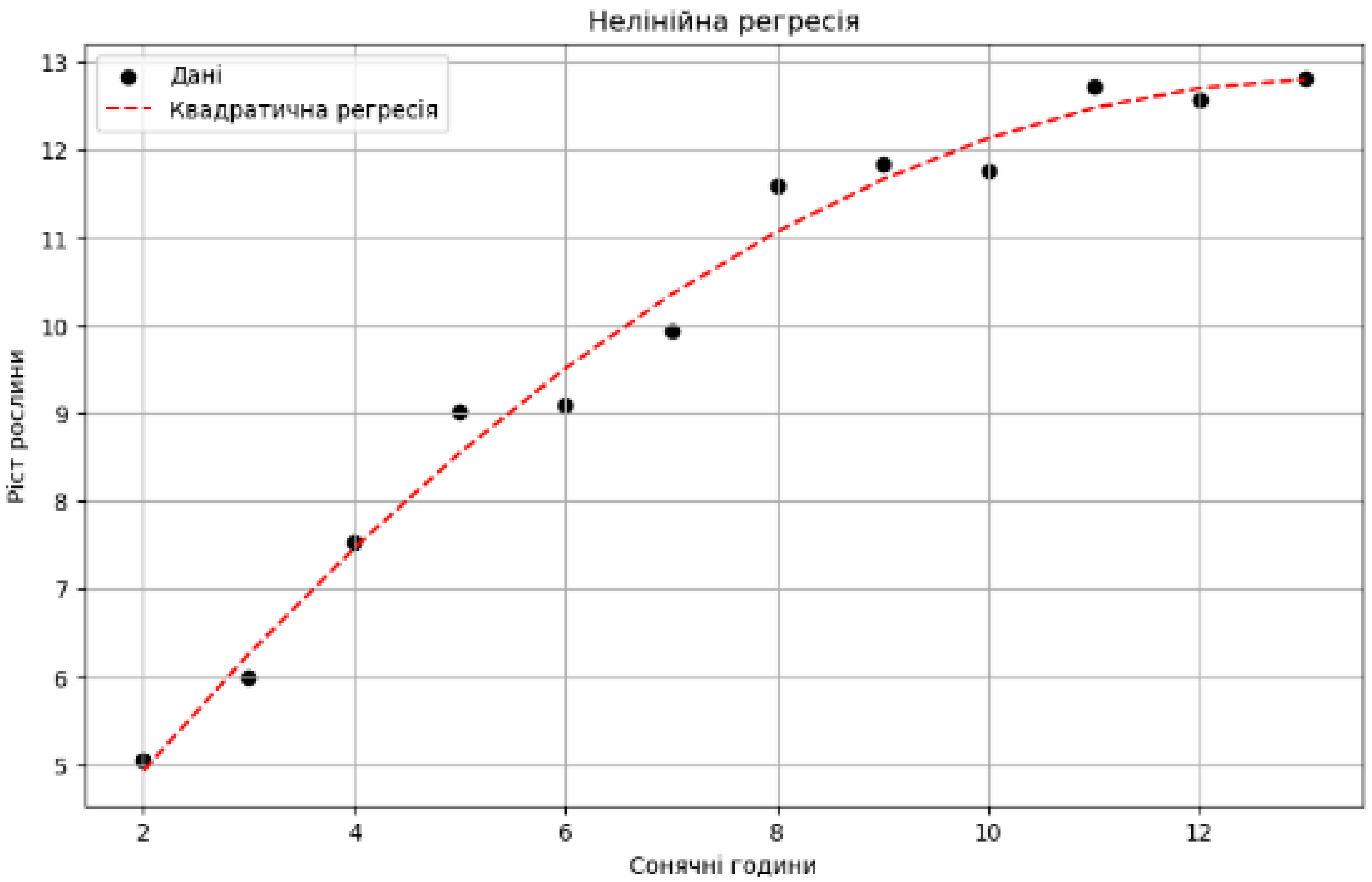
Розв'язання системи рівнянь дало наступні значення коефіцієнтів:

- $b_0 \approx 1.890$
- $b_1 \approx 1.640$
- $b_2 \approx -0.062$

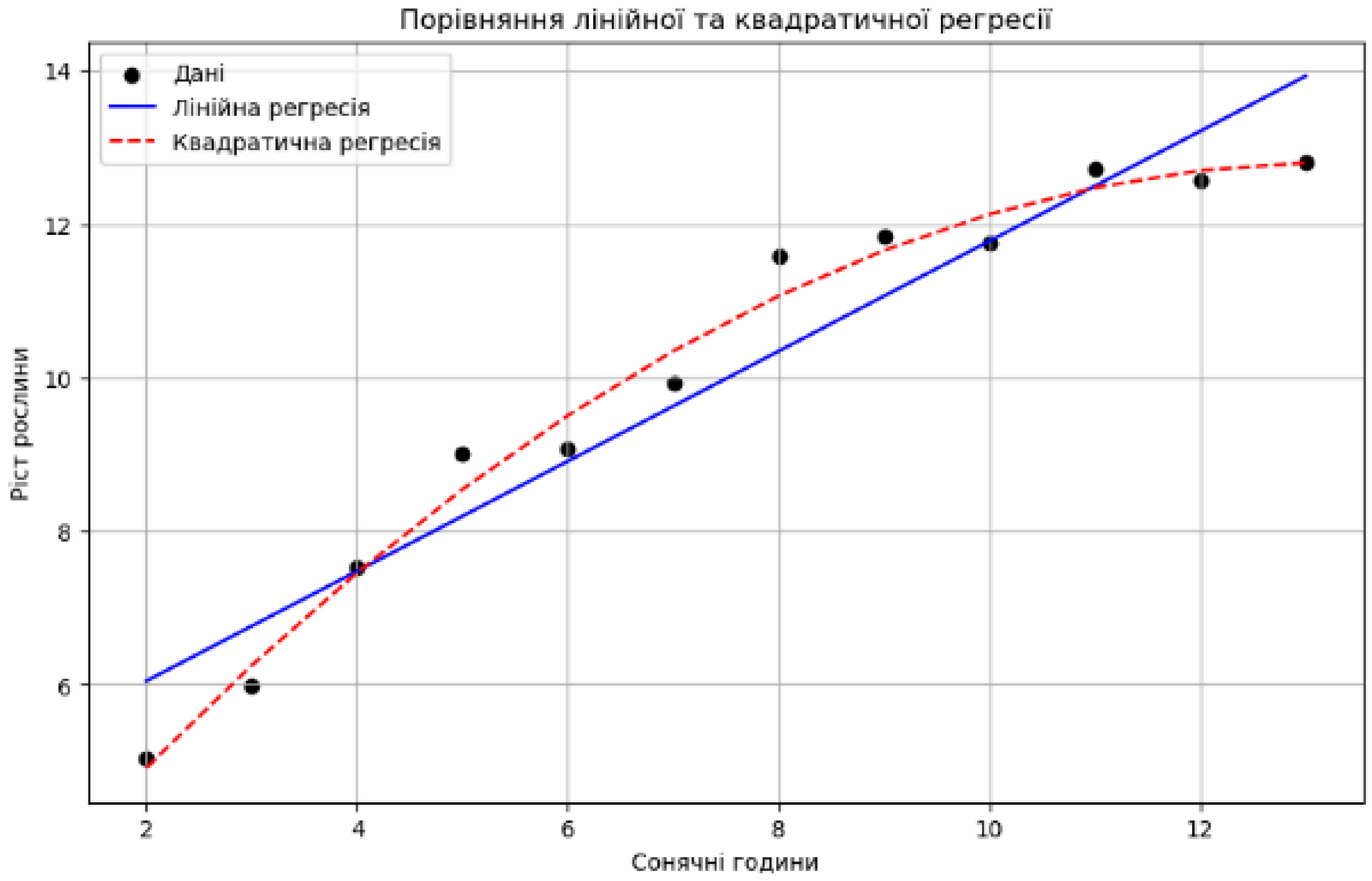
Рівняння квадратичної регресії

Остаточне рівняння квадратичної регресії, впорядковане для відповідності теоретичній формі: $y \approx 1.890 + 1.640x - 0.062x^2$

Графік нелінійної регесії



Порівняння



Порівняння лінійної та квадратичної регресії
Дані
Лінійна регресія
Квадратична регресія
Ріст рослини
Сонячні години
14
12
10
8
6
2
4
6
8
10
12

Порівняння

Для **лінійної моделі** $y = ax + b$ середньоквадратична помилка (MSE) визначається як:

$$\text{MSE}_{\text{лінійна}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (ax_i + b)\right)^2$$

Підставляючи дані, отримали:

$$\text{MSE}_{\text{лінійна}} = 0.520$$

Для **квадратичної (нелінійної) моделі** $y = b_0 + b_1x + b_2x^2$:

$$\text{MSE}_{\text{квадратична}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (b_0 + b_1x_i + b_2x_i^2)\right)^2$$

Підставляючи дані, отримали:

$$\text{MSE}_{\text{квадратична}} = 0.098$$

Порівняння:

$$\text{MSE}_{\text{квадратична}} < \text{MSE}_{\text{лінійна}}$$

Тобто **квадратична модель описує дані точніше**, бо середня квадратична помилка менша.

Висновок

У цій роботі я реалізував алгоритми лінійної та квадратичної (нелінійної) регресії для моделювання залежності росту рослин від кількості сонячних годин. У процесі реалізації я використав метод найменших квадратів для обчислення параметрів моделей, а також розрахував проміжні суми та середні значення ознаки та відповіді. Це дозволило побудувати лінійну модель, яка описує дані прямою, та квадратичну модель, яка враховує нелінійність у вигляді параболи. Порівняння середньоквадратичних похибок показало, що квадратична модель точніше відтворює залежність, оскільки її помилка менша, ніж у лінійної моделі.