# Лабораторна робота 2.3: Оцінка та вдосконалення моделей машинного навчання

## Мета лабораторної роботи

У цій лабораторній роботі ви:

* Розділите дані на навчальну, валідаційну та тестову вибірки
* Оціните ефективність моделей за допомогою відповідних метрик
* Діагностуєте проблеми зміщення (bias) та дисперсії (variance)
* Застосуєте регуляризацію для вдосконалення моделей
* Побудуєте криві навчання для оцінки моделі
* Застосуєте методи покращення продуктивності моделей на практиці

### **Варіант 2: Прогнозування заробітної плати**

Опис: Розробка моделі для прогнозування заробітної плати на основі років досвіду.

Задачі:

* Згенеруйте дані, де залежність зарплати від досвіду має вигляд полінома 3-го ступеня з доданим шумом
* Дослідіть моделі з різними ступенями поліному (від 1 до 8)
* Визначте оптимальний ступінь поліному та застосуйте регуляризацію
* Проаналізуйте криві навчання для різних моделей

## Питання для самоперевірки

1. Яка мета розділення даних на навчальну, валідаційну та тестову вибірки? 2. Які ознаки моделі з високим зміщенням (high bias)? Як їх виправити?

3. Які ознаки моделі з високою дисперсією (high variance)? Як їх виправити? 4. Як визначити оптимальний ступінь поліному для регресійної моделі?

5. Як визначити оптимальне значення параметра регуляризації?

6. Що показують криві навчання? Як їх інтерпретувати?

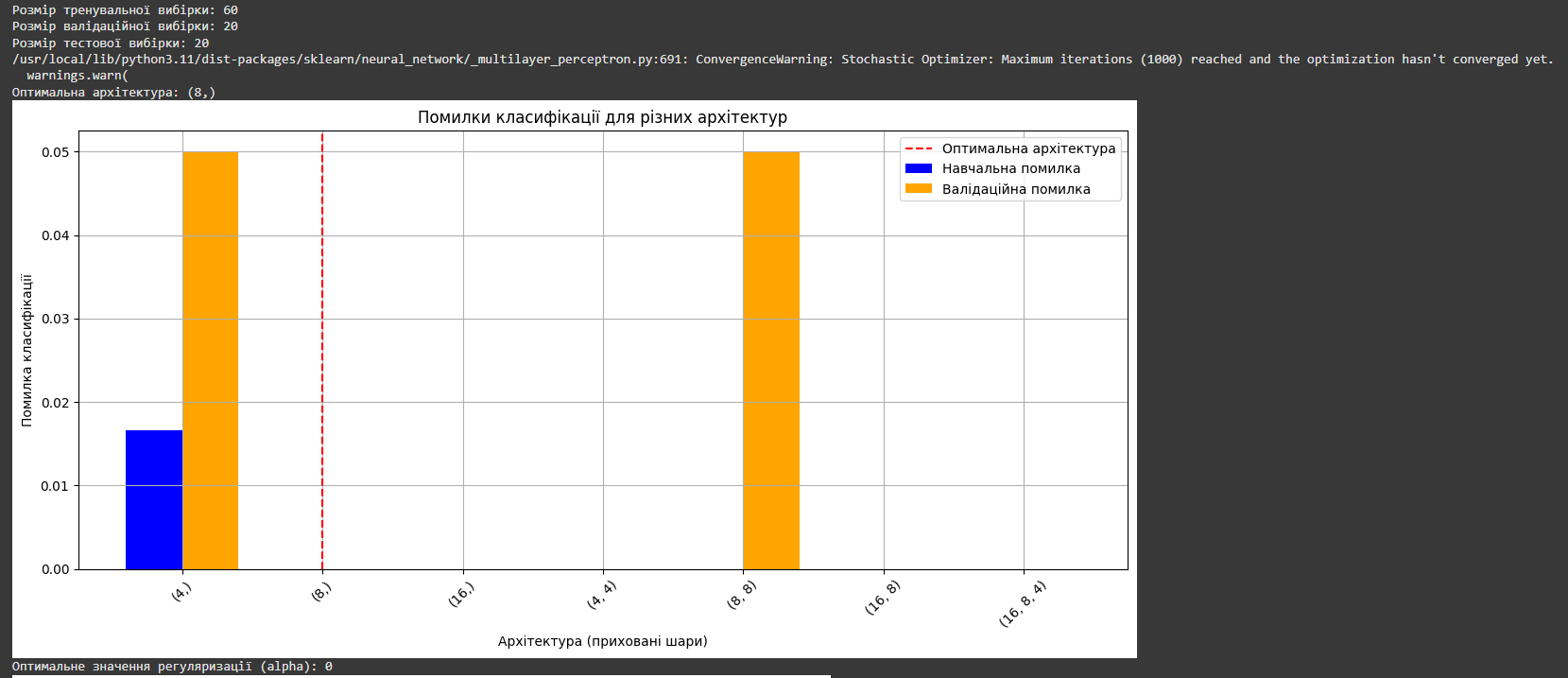
7. Які методи можна застосувати для боротьби з перенавчанням (overfitting)? 8. У чому різниця між недонавчанням (underfitting) та перенавчанням (overfitting)?

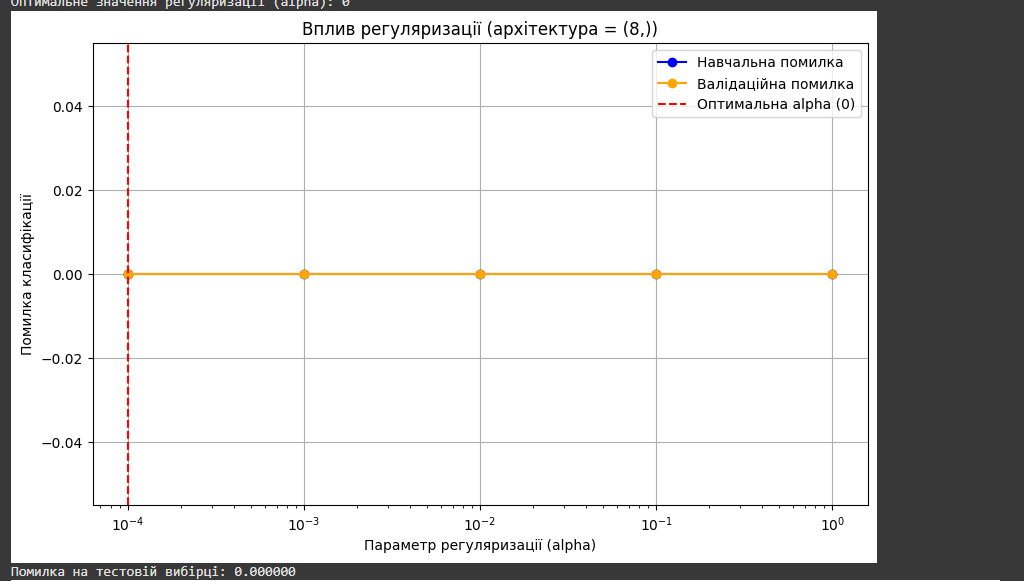
9. Як впливає розмір навчальної вибірки на якість моделі?

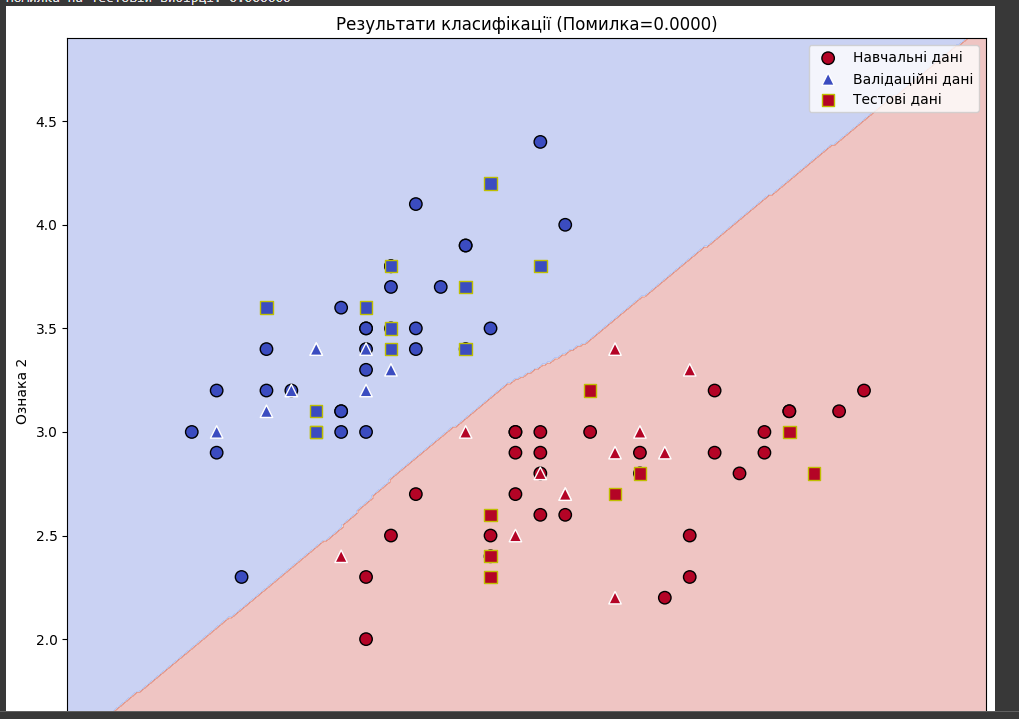
10. Як оцінити, чи потрібно збирати більше даних для покращення моделі?

1. Розділення даних на навчальну, валідаційну та тестову вибірки потрібне для того, щоб навчити модель на одній частині даних, підібрати найкращі гіперпараметри на іншій (валідаційній), а потім перевірити справжню якість моделі на тестовій, не використовуючи її у процесі навчання чи налаштування.
2. Модель із високим зміщенням зазвичай має низьку точність як на навчальних, так і на тестових даних. Вона не здатна вловити закономірності в даних, бо є надто простою. Щоб виправити це, слід збільшити складність моделі, наприклад, використати поліном вищого ступеня або додати нові релевантні ознаки.
3. Модель із високою дисперсією демонструє дуже хорошу точність на навчальній вибірці, але значно гірші результати на тестовій. Це свідчить про те, що модель занадто сильно підлаштувалась під навчальні дані й не узагальнює. Для зменшення дисперсії зазвичай зменшують складність моделі, вводять регуляризацію або додають більше навчальних даних.
4. Оптимальний ступінь полінома визначають за результатами валідації: підбирають такий, при якому модель має найменшу валідаційну помилку. Це дозволяє уникнути як недонавчання, так і перенавчання.
5. Щоб знайти найкраще значення параметра регуляризації, найчастіше використовують кросвалідацію. Порівнюючи якість моделі для різних значень параметра, обирають те, що дає найменшу помилку на валідаційній вибірці.
6. Криві навчання показують, як змінюється помилка на тренувальних і валідаційних даних при збільшенні кількості навчальних прикладів. Вони допомагають зрозуміти, чи модель страждає від перенавчання чи недонавчання. Наприклад, коли помилки на тренуванні й валідації високі й близькі — це ознака недонавчання, а коли помилка на тренуванні мала, але на валідації висока — це ознака перенавчання.
7. Щоб уникнути перенавчання, можна використовувати регуляризацію, зменшувати складність моделі, збільшувати обсяг навчальних даних або застосовувати техніки на кшталт dropout у нейронних мережах і ранньої зупинки навчання.
8. Недонавчання виникає тоді, коли модель не здатна навчитися навіть із тренувальних даних, тобто вона занадто спрощена. Перенавчання ж навпаки означає, що модель вивчає навіть шум і випадкові особливості навчальної вибірки, тому не працює добре на нових даних. Основна різниця між цими двома проблемами — у співвідношенні точності на тренувальних і тестових даних.
9. Збільшення розміру навчальної вибірки зазвичай покращує якість моделі, особливо якщо модель має високу дисперсію. При цьому вона починає краще узагальнювати й менш чутливо реагувати на шум у даних.
10. Щоб зрозуміти, чи потрібні додаткові дані, варто подивитися на криві навчання. Якщо модель має високу валідаційну помилку, а тренувальна — значно менша, і при цьому валідаційна крива ще не стабілізувалась, тоді збільшення кількості прикладів може справді покращити результат.

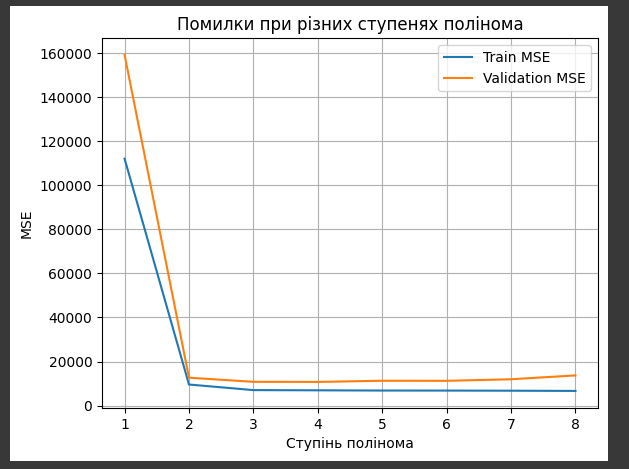
**Виконання коду:**







Дослідження моделей з різним ступенем (1–8)



Визначення оптимального ступеня та застосування регуляризації



Криві навчання для різних моделей

