**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ**

**ІМЕНІ ТАРАСА ШЕВЧЕНКА**

Факультет комп’ютерних наук та кібернетики

Кафедра математичної інформатики

**Звіт**

до лабораторної роботи № 1

Виконав студент 1-го курсу

ОНП “Штучний інтелект”

Ходаков Максим Олегович

Київ – 2025

***Тема: Лінійна регресія на прикладі прогнозування якості вина (WineQT Dataset)***

# 1. Постановка задачі

Метою цієї роботи є дослідження можливостей лінійної регресії для прогнозування якісних показників вина на основі його фізико-хімічних характеристик. В якості прикладу використовується датасет WineQT, що містить результати лабораторних аналізів червоного вина та його оцінку за шкалою від 0 до 10.

Задача:- побудувати модель, яка передбачає оцінку якості вина (quality) на основі вхідних змінних; - порівняти класичну лінійну регресію (OLS) зі штрафними функціями L2-регуляризації (Ridge) та L1-регуляризації (Lasso).

# 2. Використані технології

Для виконання роботи застосовувалися наступні програмні засоби та бібліотеки:  
- Python 3.x – мова програмування;  
- pandas – робота з таблицями (DataFrame);  
- numpy – математичні обчислення;  
- matplotlib, seaborn – візуалізація даних;  
- scikit-learn (sklearn) – побудова моделей машинного навчання.

# 3. Теоретичні відомості

## 3.1. Лінійна регресія (OLS)

Лінійна регресія намагається знайти залежність між набором незалежних змінних (X) та цільовою змінною (y), використовуючи рівняння виду:  
y = w0 + w1\*x1 + w2\*x2 + ... + wn\*xn + ε

Модель навчається шляхом мінімізації суми квадратів похибок (MSE):  
MSE = (1/m) \* Σ (yi - ŷi)^2

## 3.2. Проблеми класичної регресії

- Мультиколінеарність: коли вхідні ознаки сильно корелюють між собою → коефіцієнти стають нестабільними.  
- Аутлайєри: поодинокі "аномальні" значення сильно впливають на результат.  
- Overfitting: модель занадто точно підганяє дані тренувальної вибірки і гірше працює на нових.

## 3.3. Регуляризація

Для уникнення проблем вводяться штрафні функції, які обмежують "розмах" коефіцієнтів:  
  
1. Ridge-регресія (L2): J(w) = MSE + α Σ wj^2  
 – зменшує всі коефіцієнти плавно.  
  
2. Lasso-регресія (L1): J(w) = MSE + α Σ |wj|  
 – занулює частину коефіцієнтів і виконує відбір ознак.

# 4. Опис даних

Датасет WineQT.csv складається з 12 змінних:  
Вхідні ознаки (X):  
1. fixed acidity  
2. volatile acidity  
3. citric acid  
4. residual sugar  
5. chlorides  
6. free sulfur dioxide  
7. total sulfur dioxide  
8. density  
9. pH  
10. sulphates  
11. alcohol  
  
Вихідна змінна (y): quality – якість вина (0–10).  
Додатково: колонка Id (видалена як службова).

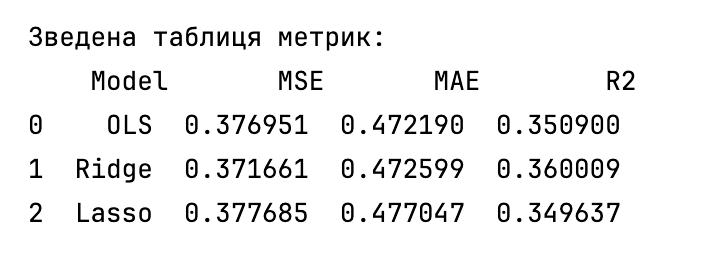
# 5. Хід роботи

1. Завантаження даних із CSV.  
2. Очищення даних (видалення колонки Id).  
3. Аналіз кореляційної матриці.  
4. Формування X і y.  
5. Розбиття даних на train/test (80/20).  
6. Масштабування ознак.  
7. Навчання моделей: OLS, RidgeCV, LassoCV.  
8. Оцінка результатів (MSE, MAE, R²).  
9. Аналіз коефіцієнтів та порівняння.

# 6. Результати експериментів

## 6.1. Метрики

Приблизні результати (залежатимуть від конкретного запуску):



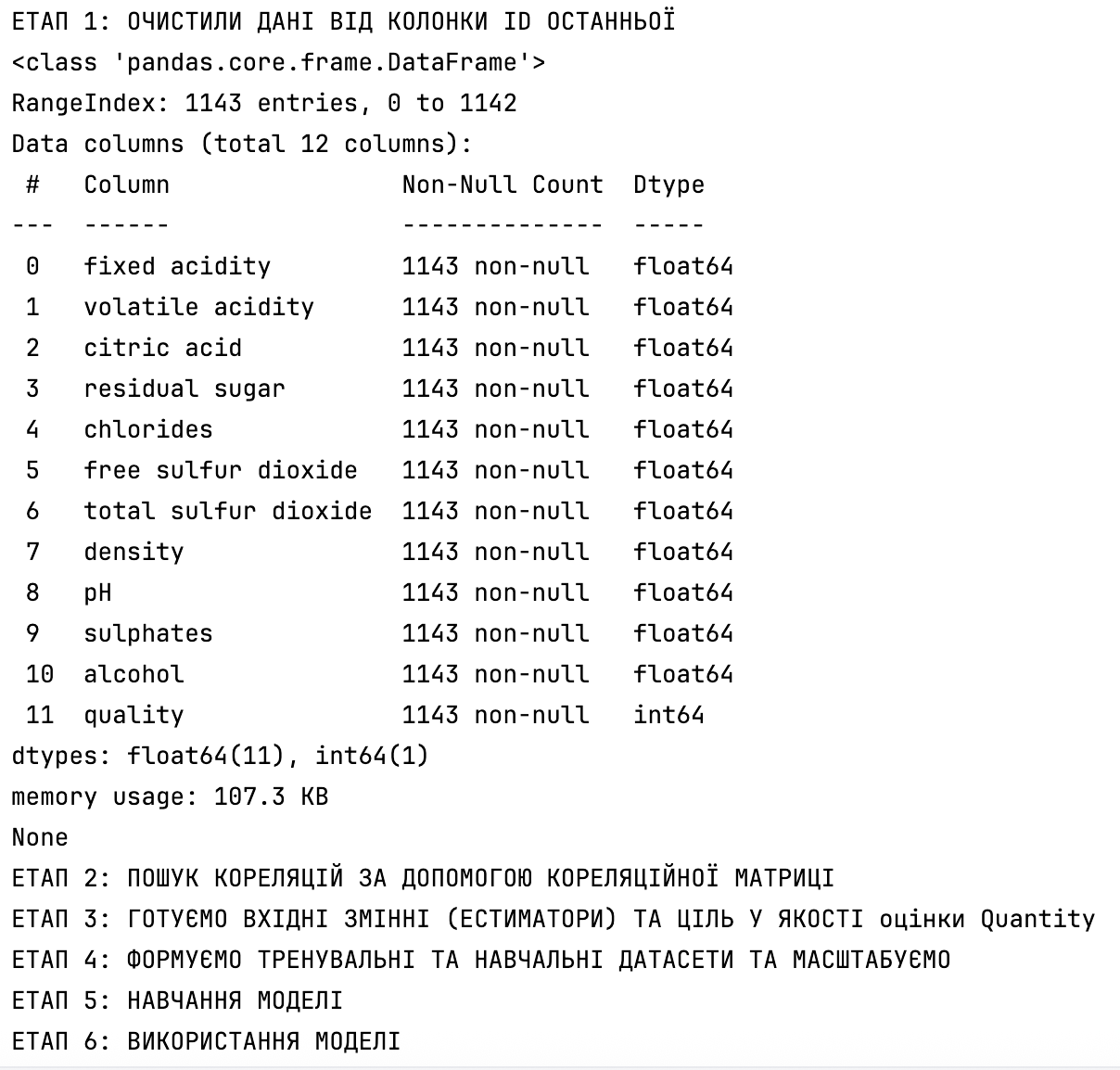
## 6.2. Коефіцієнти

– OLS: alcohol і sulphates позитивно впливають на якість, volatile acidity – негативно.  
– Ridge: зменшив усі коефіцієнти, але зберіг тренди.  
– Lasso: занулював слабкі предиктори, залишив найбільш важливі.

6.3. Лістинг програми

A screenshot of a computer

Description automatically generated



A screenshot of a computer

Description automatically generated

## A screenshot of a computer Description automatically generated

A colorful squares with white text

Description automatically generated with medium confidence

A graph with blue dots

Description automatically generatedA graph with blue dots

Description automatically generated

A graph with blue dots

Description automatically generated

A graph with a bar graph

Description automatically generated with medium confidence

A graph with blue dots

Description automatically generated

A graph with a bar graph

Description automatically generated with medium confidence

A graph with a bar graph

Description automatically generated with medium confidence

A graph with blue and orange dots

Description automatically generated

A graph of different types of sulphur

Description automatically generated

A graph with numbers and lines

Description automatically generated

## 6.4. Код програми

Доступний за посиланням: <https://github.com/maksymkhodakov/MLLab1>

# 7. Висновки

1. Лінійна регресія проста й інтерпретована, але має обмеження при мультиколінеарності.  
2. Ridge-регресія стабілізує модель і покращує узгодженість.  
3. Lasso-регресія виконує відбір ознак і спрощує модель.  
4. У нашому випадку найкраще показав себе Ridge (L2).  
5. Подальші дослідження можна проводити з поліноміальними ознаками або більш складними моделями.