**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

# Звіт по лабораторній роботі

**Тема:** Лінійна регресія (OLS, Ridge, Lasso) для прогнозування вмісту алкоголю у вині

## 1. Постановка задачі

Мета роботи — навчитися застосовувати методи лінійної регресії для прогнозування неперервної величини на основі багатовимірних вхідних даних.  
В якості прикладу використовується датасет **Wine Quality (WineQT)**, що містить фізико-хімічні характеристики зразків червоного вина.

Завдання:

1. Завантажити та підготувати дані.
2. Провести дослідження кореляцій між ознаками.
3. Побудувати базову модель **лінійної регресії (OLS)**.
4. Додати дві модифікації: **Ridge (L2-регуляризація)** та **Lasso (L1-регуляризація)**.
5. Виконати розбиття даних на **train / validation / test**, щоб перевірити перенавчання (overfitting).
6. Провести аналіз якості моделей за допомогою метрик (MSE, MAE, R²).
7. Побудувати графіки (parity plot, residuals, криві навчання та валідації).
8. Зробити висновки щодо ефективності моделей та впливу регуляризації.

## 2. Опис датасету

Дані містять результати хімічного аналізу вина та дві цільові характеристики: quality (оцінка дегустаторів, цілочисельна від 0 до 10) та alcohol (вміст алкоголю у відсотках).

Вхідні ознаки (predictors):

* fixed acidity — фіксована кислотність
* volatile acidity — летюча кислотність
* citric acid — лимонна кислота
* residual sugar — залишковий цукор
* chlorides — хлориди
* free sulfur dioxide — вільний діоксид сірки
* total sulfur dioxide — загальний діоксид сірки
* density — густина
* pH — кислотність
* sulphates — сульфати
* quality — сенсорна оцінка (використовується як одна з ознак, а не як ціль)

Ціль (target):

* alcohol — неперервна величина, яку потрібно прогнозувати.

Ми обираємо саме alcohol, адже це неперервна змінна, і вона краще відповідає задачі **лінійної регресії**.

## 3. Теоретичні відомості

### 3.1. Лінійна регресія (OLS)

Лінійна регресія будує залежність виду:  
**ŷ = w1·x1 + w2·x2 + ... + wn·xn + b**  
де ŷ — прогнозоване значення, xi — значення ознак, wi — вагові коефіцієнти моделі, b — вільний член.

Модель підбирає коефіцієнти так, щоб мінімізувати середньоквадратичну похибку (MSE).

### 3.2. Метрики оцінки

* **MSE (Mean Squared Error)** — середньоквадратична похибка, чутлива до великих відхилень.
* **MAE (Mean Absolute Error)** — середня абсолютна похибка, показує на скільки в середньому модель помиляється у вимірюваній одиниці (тут — відсотки алкоголю).
* **R² (коефіцієнт детермінації)** — частка дисперсії цільової змінної, яку пояснює модель. Значення ближче до 1 означає високу якість прогнозування.

### 3.3. Регуляризація

Щоб уникнути перенавчання, до функції втрат додається штраф на великі коефіцієнти:

* **Ridge (L2):** штрафується сума квадратів коефіцієнтів. Це зменшує їх величини, але не занулює.
* **Lasso (L1):** штрафується сума модулів коефіцієнтів. Може занулювати частину коефіцієнтів, тим самим виконуючи відбір ознак.

### 3.4. Стандартизація ознак

Ознаки мають різні масштаби (наприклад, alcohol ≈ 10, sulphates ≈ 0.5).  
Для коректної роботи моделі їх потрібно стандартизувати:  
**z = (x - μ) / σ**,  
де μ — середнє значення, σ — стандартне відхилення (рахуються тільки по train-набору).

## 4. Методика експерименту

1. Завантаження даних з CSV.
2. Видалення технічної колонки Id.
3. Побудова кореляційної матриці для дослідження зв’язків.
4. Розбиття на train/validation/test у пропорції 60/20/20.
5. Масштабування ознак.
6. Навчання трьох моделей:
   * OLS (без регуляризації)
   * Ridge (L2) з підбором параметра α через крос-валідацію
   * Lasso (L1) з підбором параметра α через крос-валідацію
7. Оцінка метрик на train/validation/test.
8. Перевірка на перенавчання через розриви у метриках, криву навчання та криву валідації.
9. Візуалізації: parity plots (прогноз vs факт), residuals (помилки), вплив окремих ознак.

## 5. Результати

### 5.1. Метрики (приклад)

* OLS: MSE=0.25, MAE=0.35, R²=0.68
* Ridge: MSE=0.23, MAE=0.33, R²=0.70
* Lasso: MSE=0.24, MAE=0.34, R²=0.69

(значення можуть змінюватися залежно від випадкового розбиття).

### 5.2. Аналіз

* OLS має найменші обмеження, тому коефіцієнти більші, але модель трохи схильна до перенавчання.
* Ridge згладжує коефіцієнти → кращий баланс між train і validation, менший ризик перенавчання.
* Lasso занулює частину коефіцієнтів → виконує відбір ознак, що допомагає інтерпретації.

### 5.3. Графіки

* **Parity plots:** більшість точок лежить близько до лінії y=x, але є відхилення.
* **Residuals:** розподілені навколо нуля, але з невеликими систематичними відхиленнями.
* **Learning curve:** R² на тренувальних даних вищий за R² на валідації → легке перенавчання, але зростання train-set зменшує розрив.
* **Validation curve (Ridge/Lasso):** оптимальне α знаходиться між 0.1 і 10, де R² найбільший.

Наведемо основні графічні результати:

A screenshot of a graph

Description automatically generated

A graph of alcohol and alcohol

Description automatically generated

A graph of a bar chart

Description automatically generated

A graph of alcohol and alcohol

Description automatically generated

A graph of a line graph

Description automatically generated

## 6. Висновки

1. Лінійна регресія добре підходить для прогнозування вмісту алкоголю на основі хімічних характеристик вина.
2. Звичайна модель OLS має ризик перенавчання.
3. Використання регуляризації (Ridge/Lasso) покращує узагальнюючу здатність.
4. Ridge дає найкращий компроміс між якістю та стабільністю.
5. Lasso дозволяє проводити відбір ознак, що може бути корисним у подальшому аналізі.
6. Ключові фактори, що впливають на вміст алкоголю, — щільність (density), летюча кислотність та вміст сульфатів.
7. Код програми: <https://github.com/maksymkhodakov/MLLab1>