Міністерство освіти і науки України

Національний технічний університет України «Київський політехнічний

інститут імені Ігоря Сікорського"

Факультет інформатики та обчислювальної техніки

Катедра інформатики та програмної інженерії

Звіт

з комп’ютерного практикуму № 5 з дисципліни

«Аналіз даних в інформаційних системах»

«Регресійні моделі»

Виконав студент ІП-11 Лесів Владислав Ігорович

Перевірила Ліхоузова Тетяна Анатоліївна

Київ 2023

**Комп’ютерний практикум 5**

**Регресійні моделі**

**Мета роботи:** ознайомитись з різновидами регресійних моделей.

**Основне завдання**.

Завантажити дані про якість червоного вина

1. Дослідити дані, підготувати їх для побудови регресійної моделі;
2. Розділити дані на навчальну та тестову вибірки;
3. Побудувати декілька регресійних моделей для прогнозу якості вина (12 - quality). Використати лінійну регресію та поліноміальну регресію обраного вами виду;
4. Використовуючи тестову вибірку, з'ясувати, яка з моделей краща;

**Додаткове завдання.**

Завантажити дані файлу Data4.csv.

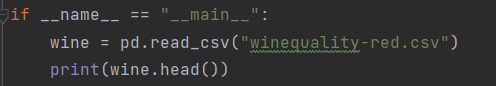
1. Дослідити дані, сказати, чи є мультиколінеарність, побудувати діаграми розсіювання;
2. Побудувати декілька регресійних моделей (використати лінійну регресію та поліноміальну регресію обраного вами виду);
3. Використовуючи тестову вибірку з файлу Data4t.csv, з'ясувати, яка з моделей краща;

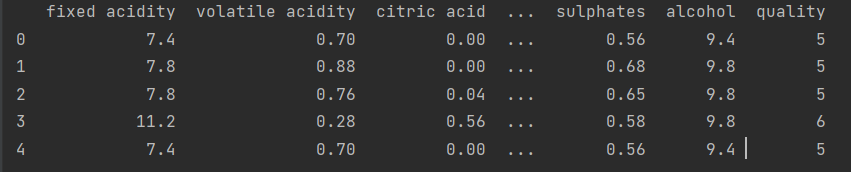
**Хід роботи.**

**Основне завдання.**

1. Дослідити дані, підготувати їх для побудови регресійної моделі;

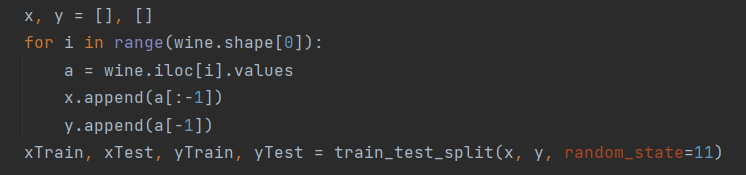
Завантажую дані, бачу, що усі числові значення задані коректно, тож дані готові для подальшого опрацювання. Маємо різні характеристики вина, а останньою колонкою – оцінка якости вина.





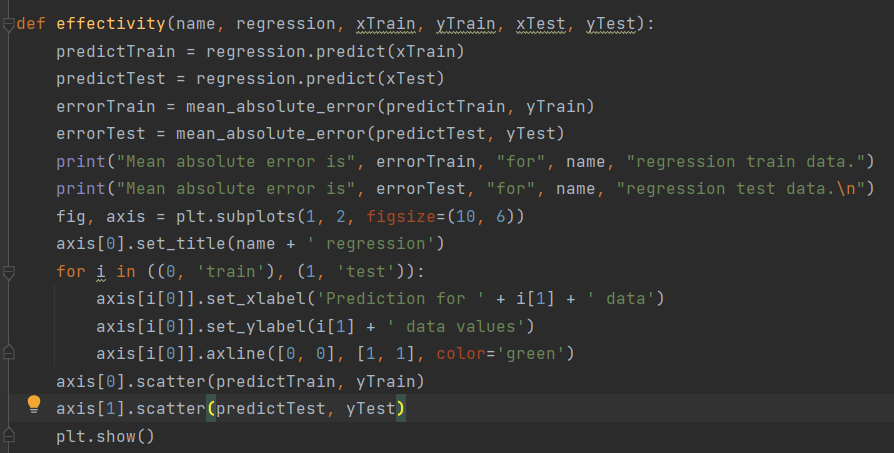
1. Розділити дані на навчальну та тестову вибірки;

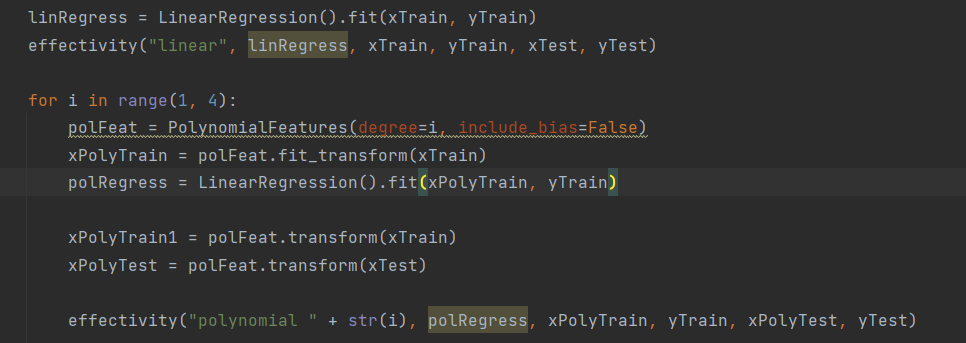
Формую масиви даних x та y, відповідно розподіляю характеристики до x, а якість – до y. Далі за допомогою sklearn.model\_selection\_train\_test\_split() розділяю створені масиви на тестову й навчальну вибірки.



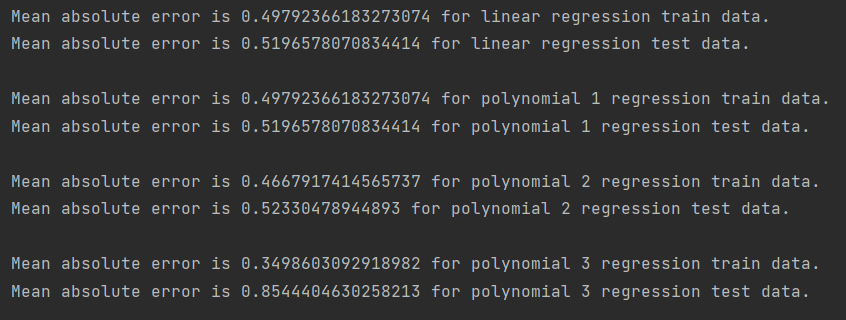
1. Побудувати декілька регресійних моделей для прогнозу якості вина (12 - quality). Використати лінійну регресію та поліноміальну регресію обраного вами виду;

Я обрав лінійну регресію та поліноміальну регресію зі степенями від 1 до 3. Треную відповідні моделі навчальною вибіркою, а потім роблю передбачення на навчальній та тестовій вибірках. Як показник ефективности буду використовувати середню абсолютну похибку. Знаходжу відповідні похибки, а також будую графіки з регресіями.





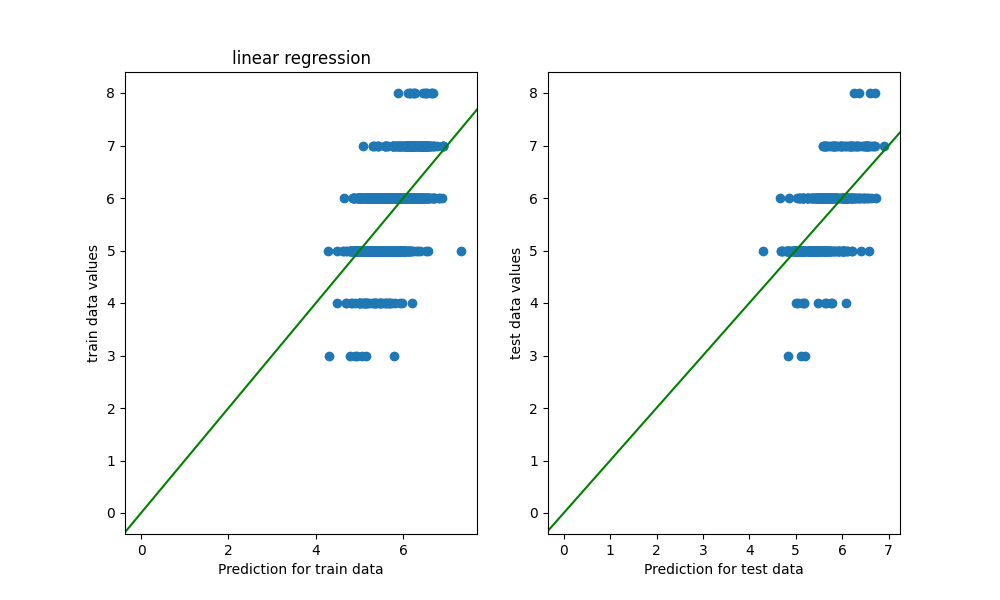
Отримуємо такі результати:

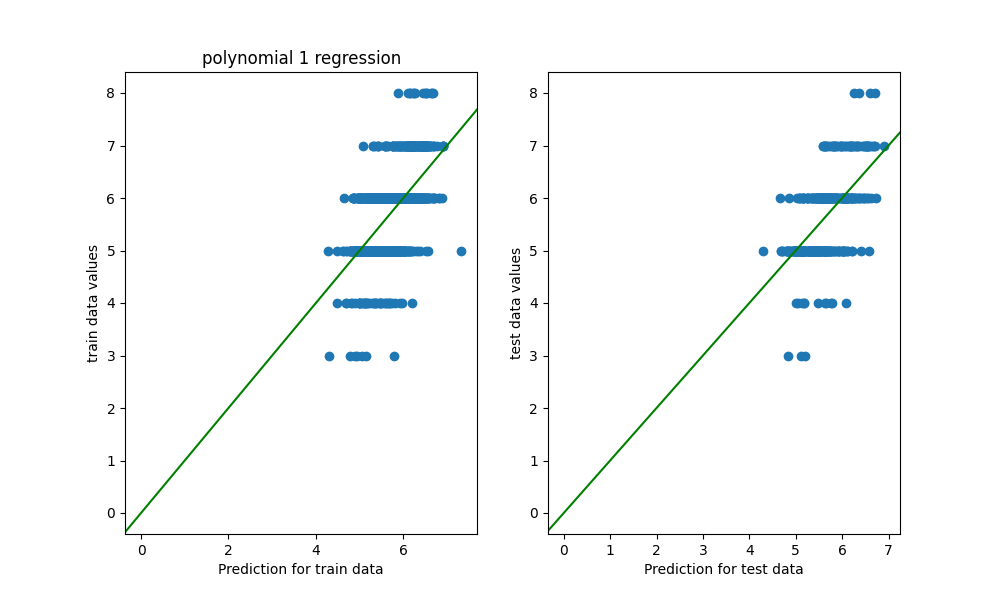


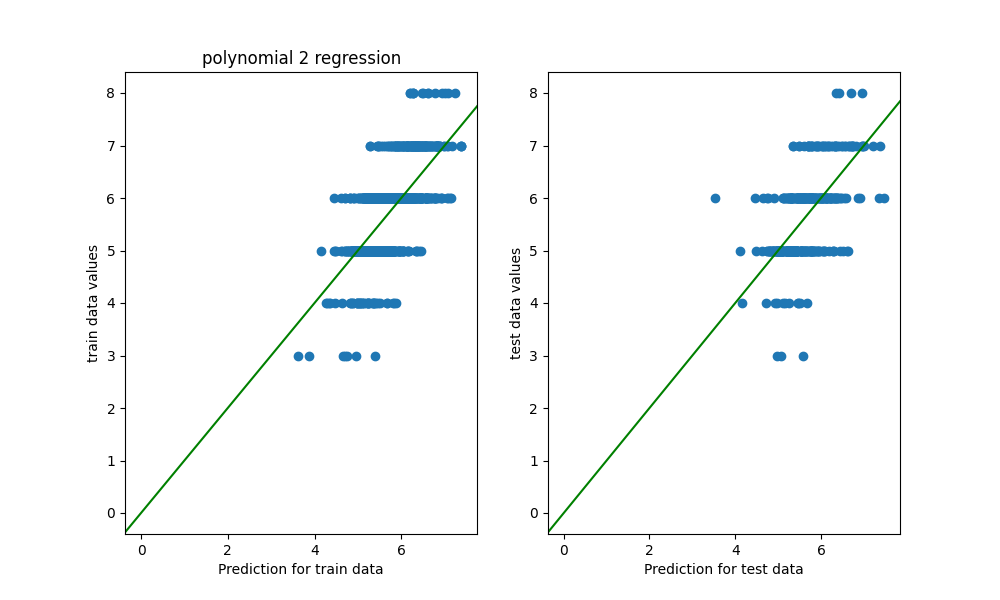
Бачимо, що чим більший степінь поліноміальної регресії, тим менша похибка після тестування навчальної вибірки, що насправді не допомагає, адже навчальна вибірка і так нам уже доступна на момент передбачення. Натомість похибка для тестової вибірки навпаки збільшується, тобто зі збільшенням степені регресія стає все менш ефективною.

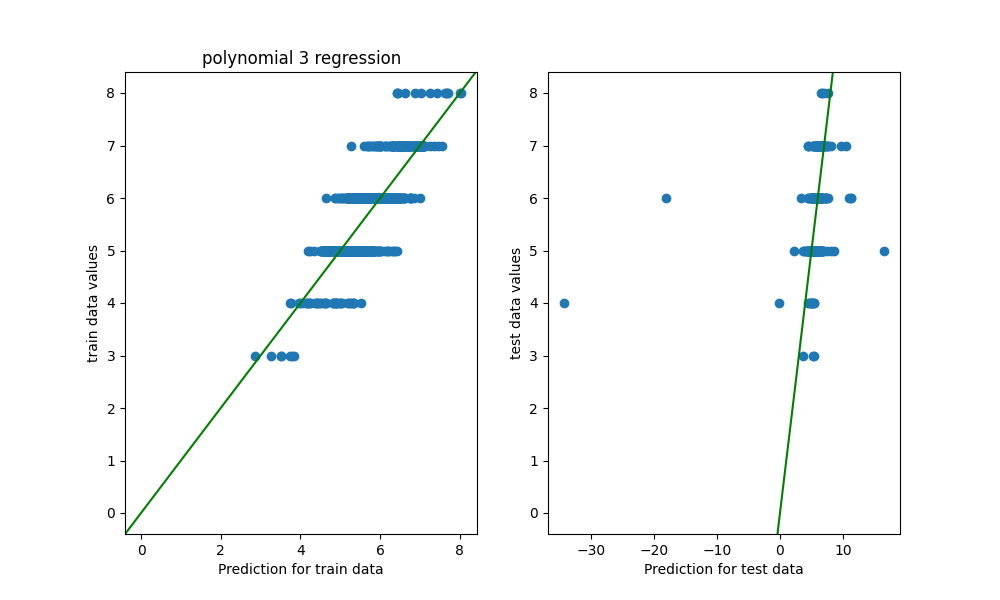
1. Використовуючи тестову вибірку, з'ясувати, яка з моделей краща;

Отримуємо також такі графіки:







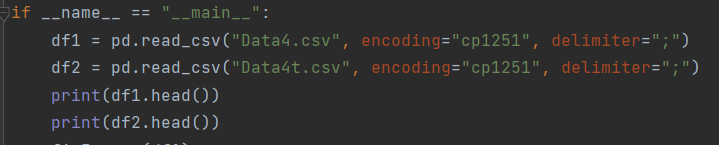


Можна спостерігати те, що було обчислено. Для тренувальної вибірки результати покращуються, але для тестової – похибка збільшується. Лінійна регресія сама по собі працює так само, як і поліноміальна зі степенем 1.

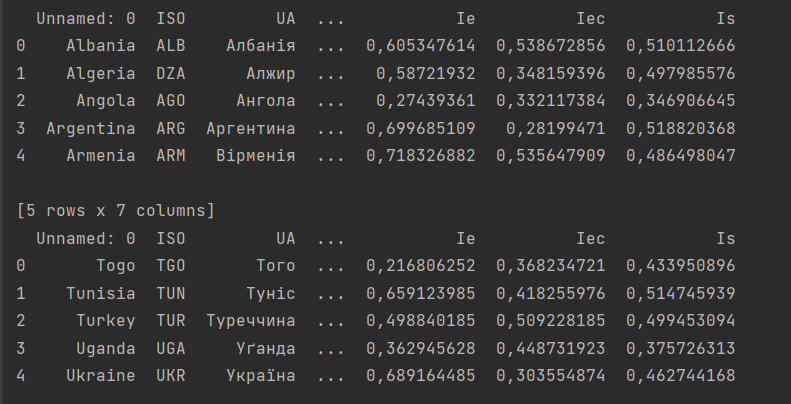
Отже, з’ясовуємо, що найкраще впоралася лінійна регресія або ж поліноміальна регресія зі степенем 1. За застосування такої регресії маємо найменшу похибку серед усіх – 0.5197.

**Додаткове завдання.**

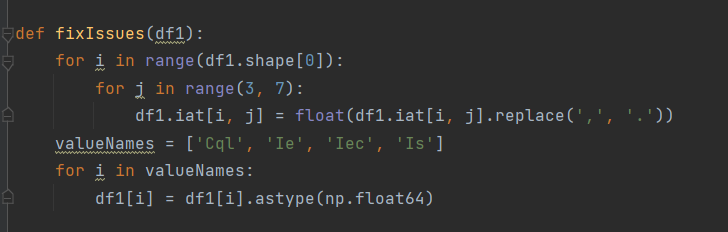
1. Дослідити дані, сказати, чи є мультиколінеарність, побудувати діаграми розсіювання;

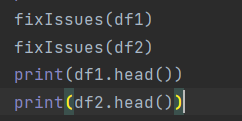


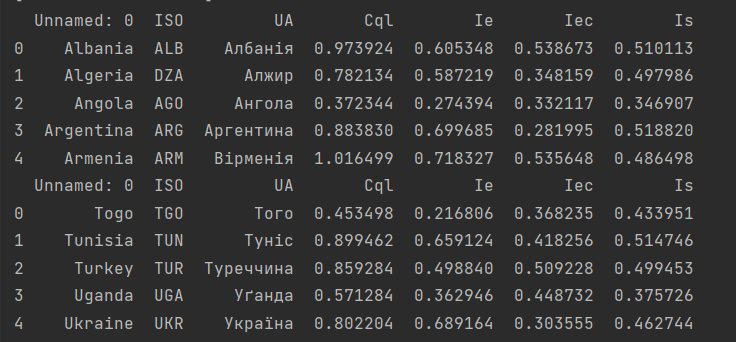
Зчитуємо дані, отримую такі результати:



Бачимо, що числові дані задані з комою, яку треба замінити на крапку для переведення у дійсний тип.

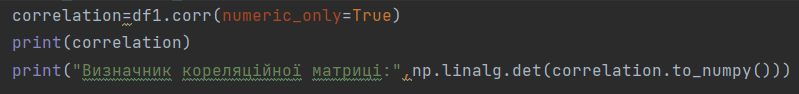


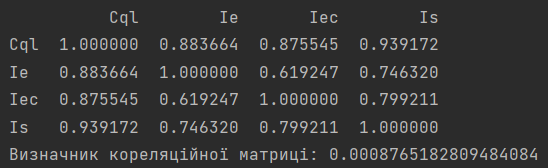




Після відповідних перетворень отримуємо дані, які можна далі досліджувати.

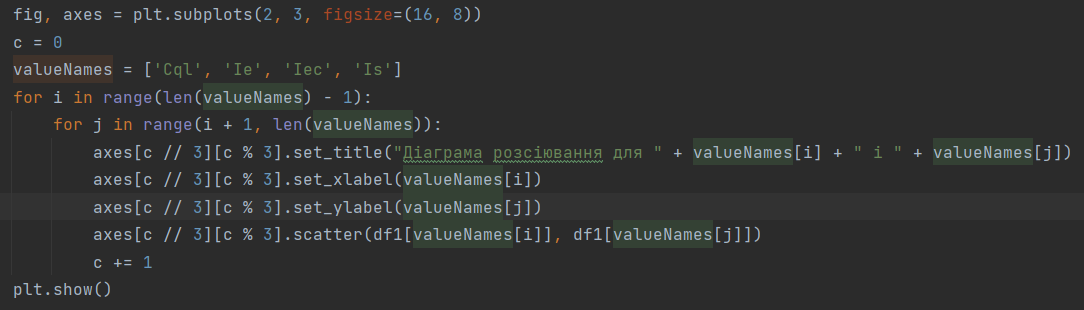
Для дослідження мультиколінеарности використаю метод corr() і для знаходження визначника кореляційної матриці – функцію numpy.linalg.det().

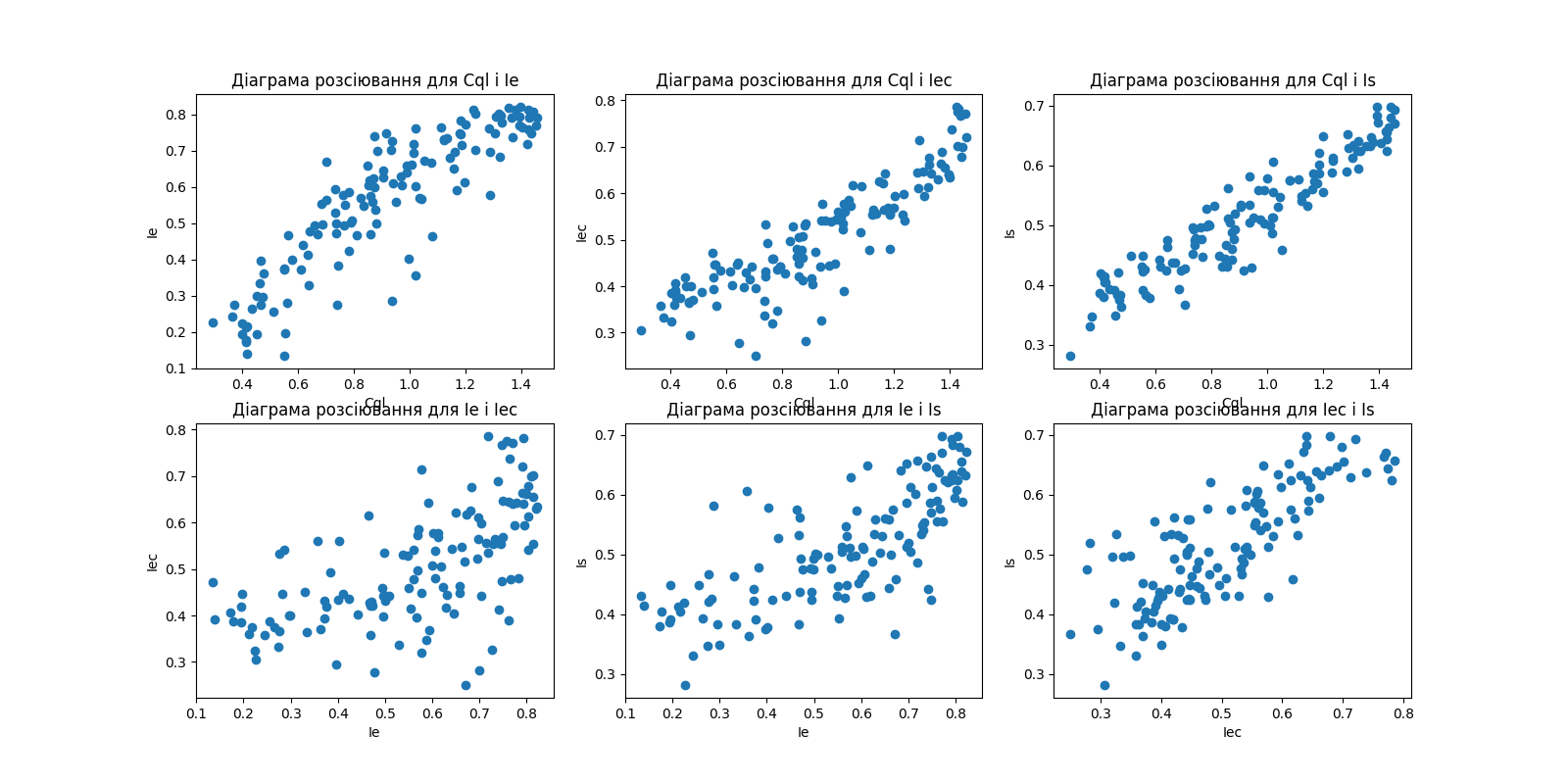




Бачимо, що визначник , а також для пари (“Cql”,”Is”) значення більше за 0,9. Тож мультиколінеарність присутня.

Побудуємо діаграми розсіювання. Будую усі на одному графіку.



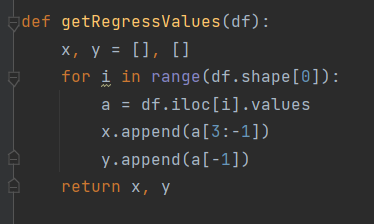


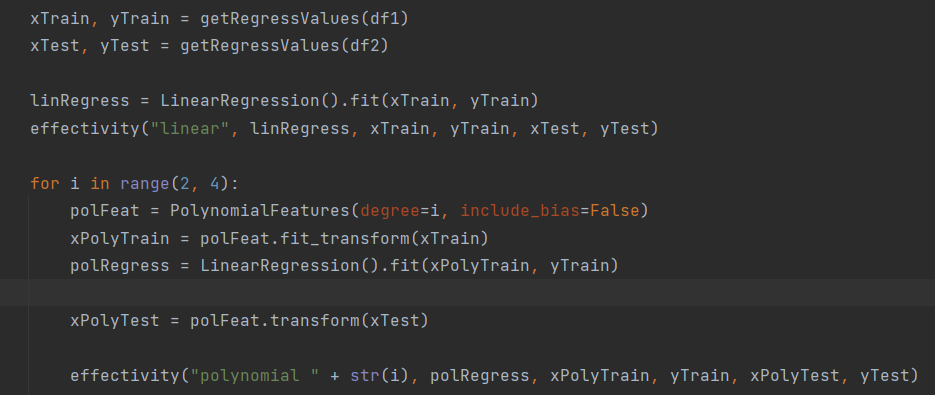
Отримуємо такі діаграми. Візуально також можна переконатися, що суто лінійної залежности ніде нема. На кожній діаграмі є, хоч і невеличкі, відхилення.

1. Побудувати декілька регресійних моделей (використати лінійну регресію та поліноміальну регресію обраного вами виду);

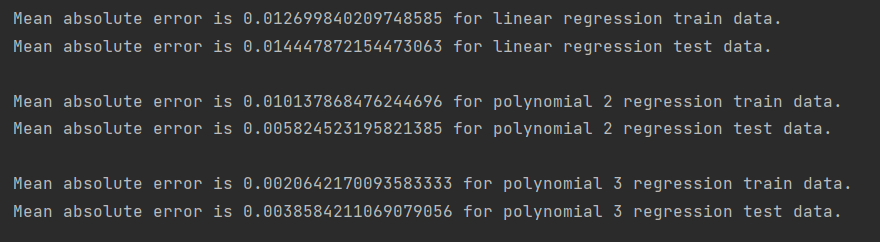
Як і в основному завданні, отримую з датафреймів значення для тестових й навчальних вибірок. Досліджувати буду останній параметр датафрейму “Is”. Далі так само, як і в основному завданні, створюю моделі, навчаю, тестую, аналізую ефективність за допомогою тієї ж функції.

Я обрав лінійну регресію та поліноміальні степенів 2 й 3.



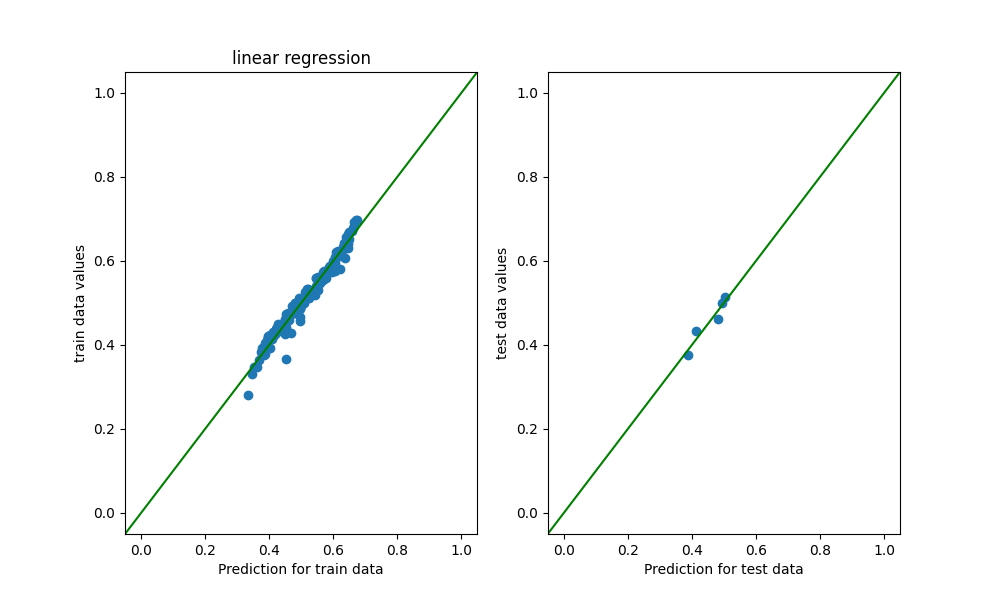


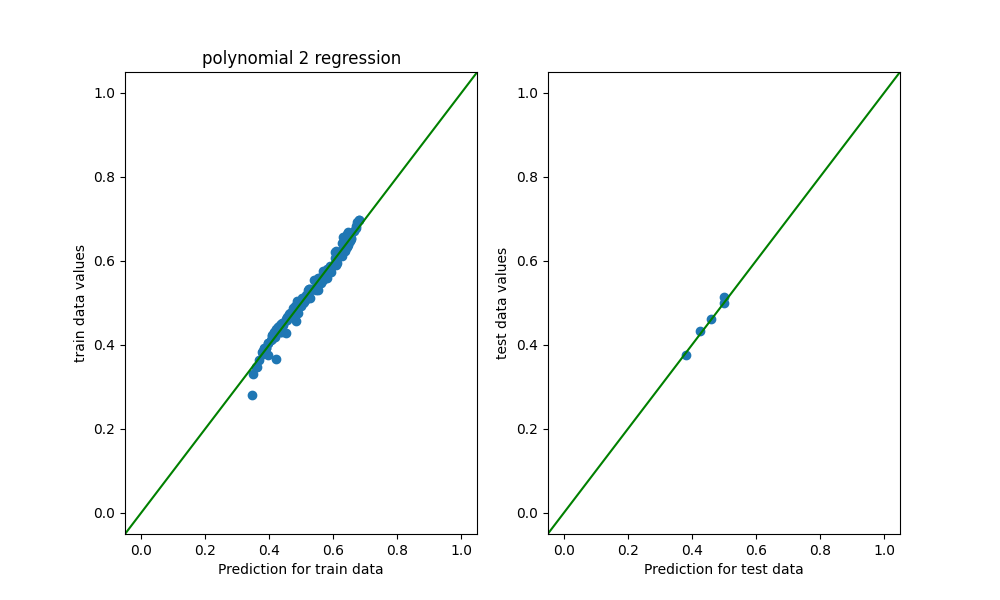
1. Використовуючи тестову вибірку з файлу Data4t.csv, з'ясувати, яка з моделей краща; Отримуємо наступні результати:

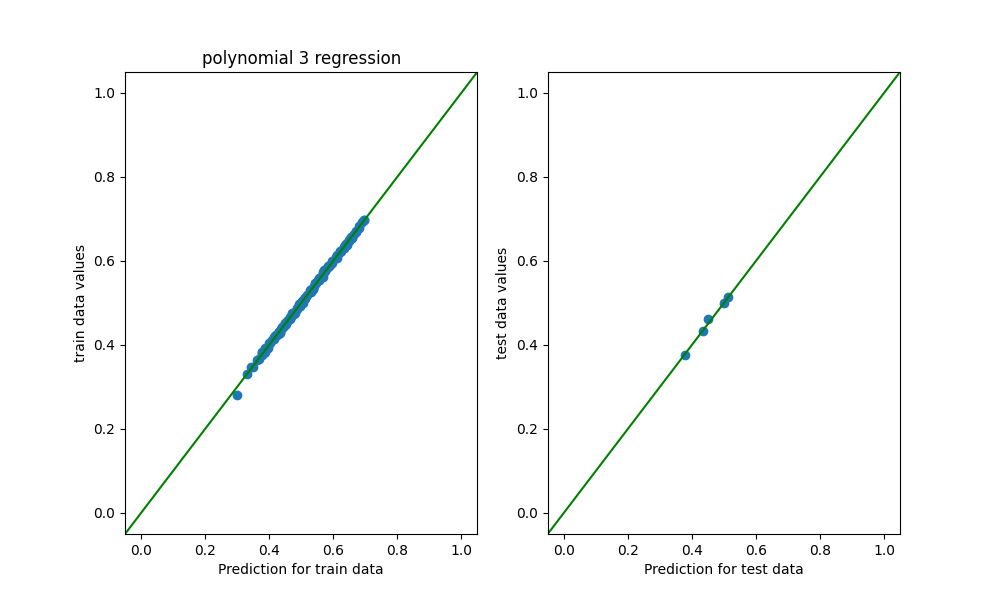


Бачимо, що для такого передбачення найкраще спрацювала поліноміальна регресія степеню 3. Похибка зменшується зі збільшенням степеню, і на степінь 3 для тестових даних складає 0.00386. Для тренувальних даних похибка так само зменшується, тобто така регресія працює в усіх випадках ефективніше, ніж решта.

Також отримуємо такі графіки, які візуально доводять сказане вище й отримане в результаті обчислень.







**Висновок.**

Отже, у цій роботі я ознайомився з різновидами регресійних моделей. У результаті лабораторної роботи було використано різні DataFrames, проведено їх дослідження, створено моделі лінійної та поліноміальних регресій, які були навчені та протестовані за допомогою даних DataFrames. Було передбачено якість вина залежно від його характеристик, ми побачили, що найкраще з цим завданням впоралася лінійна регресія. Тоді як у додатковому завданні з прогнозом найкраще впоралася поліноміальна регресія. Використовуючи PyCharm, отримуємо коректний результат.