Державний торговельно-економічний університет  
Факультет інформаційних технологій  
Кафедра інженерії програмного забезпечення та кібербезпеки

**ЛАБОРАТОРНА РОБОТА №2**

**З ДИСЦИПЛІНИ «Моделі і структури даних»**

**НА ТЕМУ: «Модель лінійної регресії»**

**Виконавець**:

Студентка групи 4-4

Авєріна Наталія Ігорівна

Київ 2024

**Лабораторна робота №2**

**Тема**: «**Модель лінійної регресії**»

**Цілі**: навчитися застосовувати різні методи регресійного аналізу для прогнозування цільової змінної, виконати попереднє дослідження даних, підготовку, моделювання та оцінку моделей. Сформувати висновки на основі порівняння оцінки продуктивності. Визначити модель регресії, яка найкраще підходить для прогнозування на обраній вибірці даних.

**Хід виконання:**

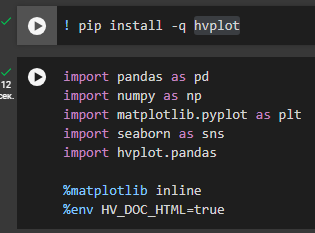
Завдання. Варіант 2

Оберіть власний датасет, де було б не менше сотні рядків, а також як числові, так і категоріальні стовпчики. Зробіть EDA, дослідіть дані. Натренуйте декілька моделей (звичайна лінійна регресія, з регуляризацією, поліноміальна). Знайдіть модель, яка якнайкраще розв’яже задачу з точки зору таких метрик якості як RMSE та R^2.

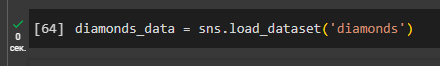
**Виконана робота:**

<https://colab.research.google.com/drive/1re9smwWP5J5qEVOuJ1irsrBZsy2FQdSL?usp=sharing>

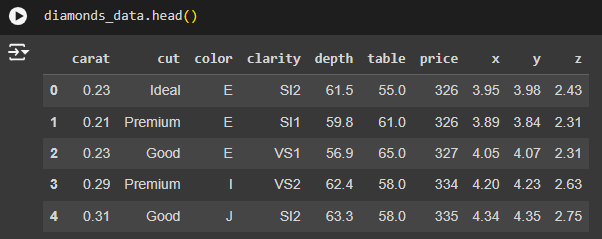
1. Завантажуємо та імпортуємо необхідні бібліотеки

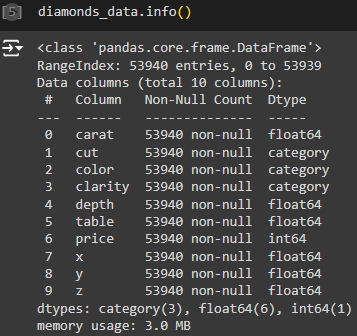


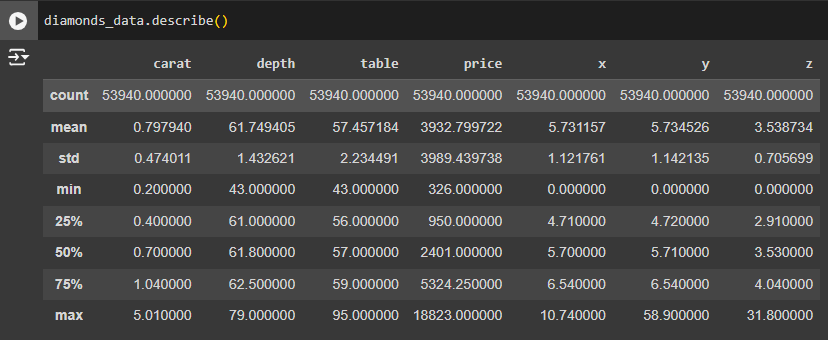
1. Завантажуємо seaborn датасет – diamonds.



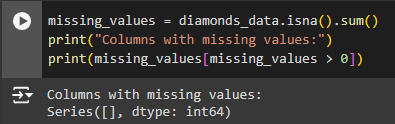
1. Переглядаємо основну інформацію про датасет.



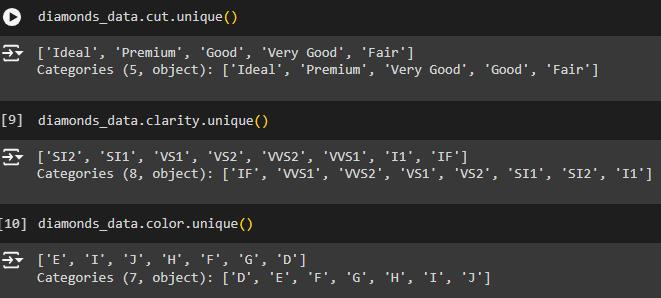




1. Перевіряємо, чи датасет не має пропущених значень.

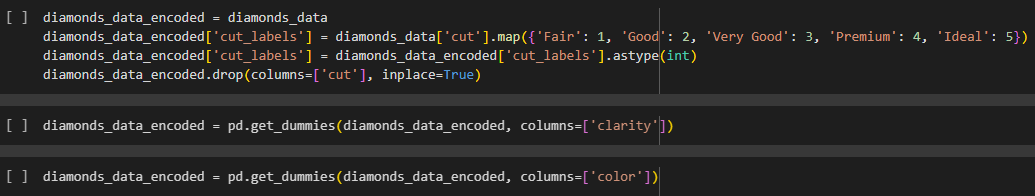


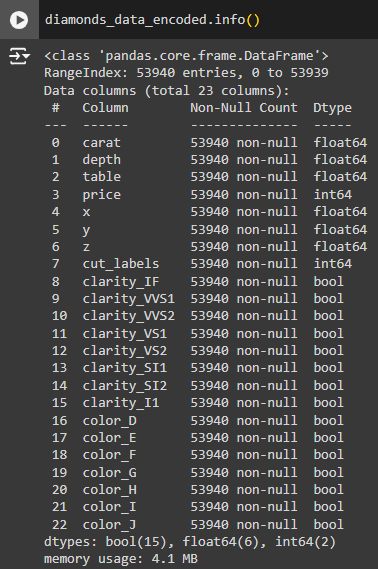
1. Так як ми маємо 3 категоріальних стовпчики, ми маємо їх перетворити на числовий формат. Спочатку переглянемо унікальні значення кожного.



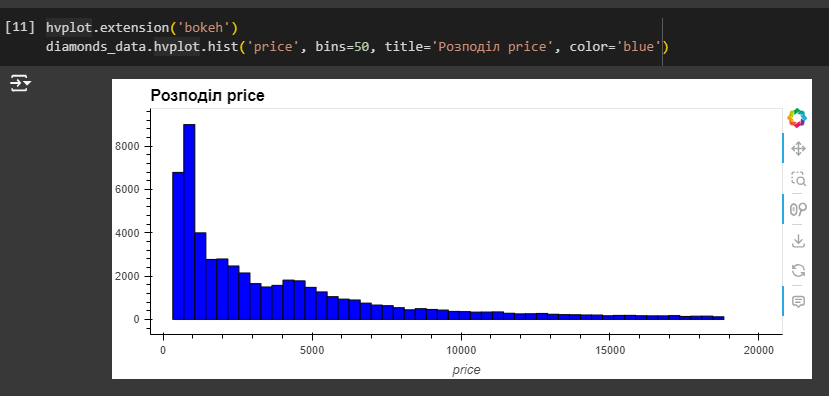
1. Є два варіанти перетворення даних: One-Hot Encoding і Label Encoding. Label Encoding більше підходить для категорій, які мають природний порядок. В нашому датасеті для Label Encoding ми можемо виділити стовпчик *cut*, який має значення *Fair* як найнижчу якість діаманту (1), а *Ideal* – найвищу (5).

Для інших категоріальних колонок використаємо One-Hot Encoding.

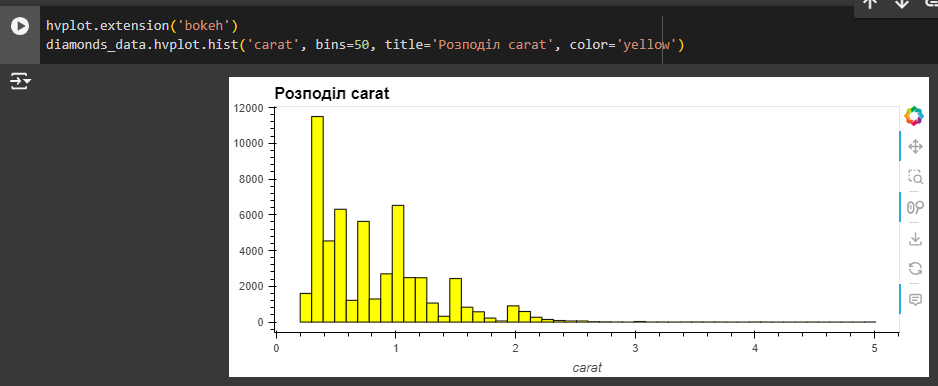




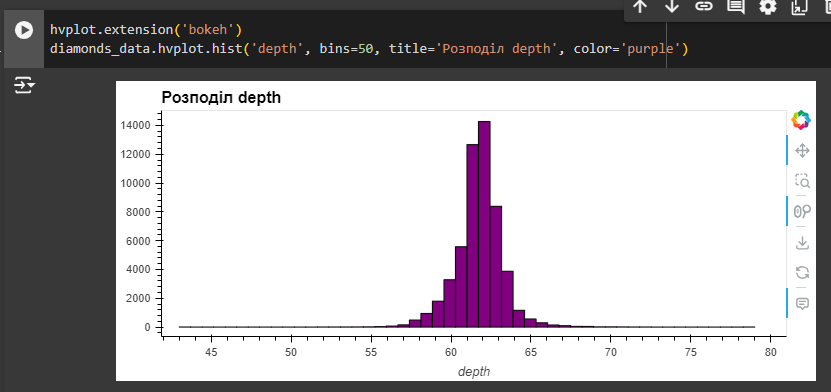
1. Виконаємо розвідувальний аналіз даних



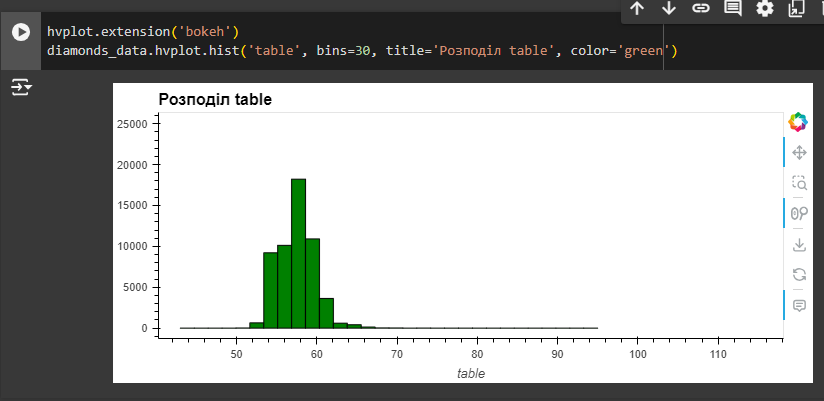
За розподілом Price можемо зазначити, що найчастіше діаманти коштують близько 880 у.о. Чим дорожче діаманти, тим менша їх кількість.



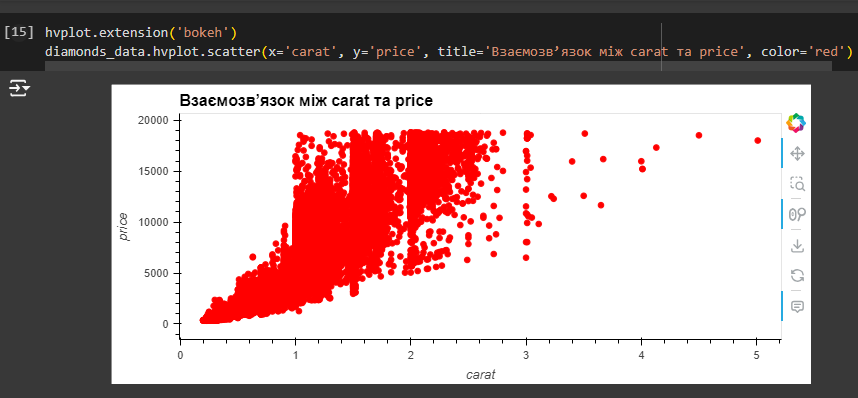
Схожа ситуація зі стовпчиком carat. Найбільше всього діамантів в датасеті розміром приблизно 0.3 карата, далі 1 карат, 0.5 і 0.7. Відповідно можна припустити, що це найпопулярніші ваги діамантів.



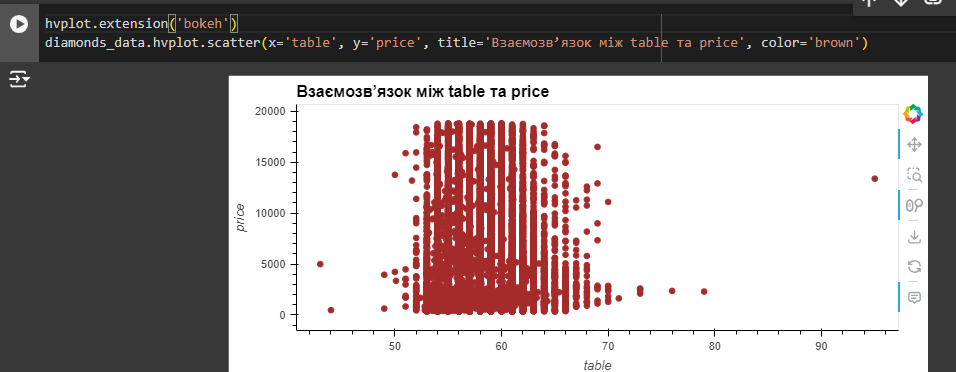
Найбільше значень depth в датасеті – 62. Глибина діаманту – це відсоткове значення висоти діаманта, яке впливає на оцінку діаманта (*cut*). Ідеальними значеннями є 59-61%. Судячи з графіку можна зазначити, що скоріше за все більшість діамантів в нашому датасеті є ideal, premium чи very good.



Ідеальне значення table коливається між 53 і 60 %. За графіком можемо побачити, що в нашому датасеті не ідеальних значень відносно дуже мало.

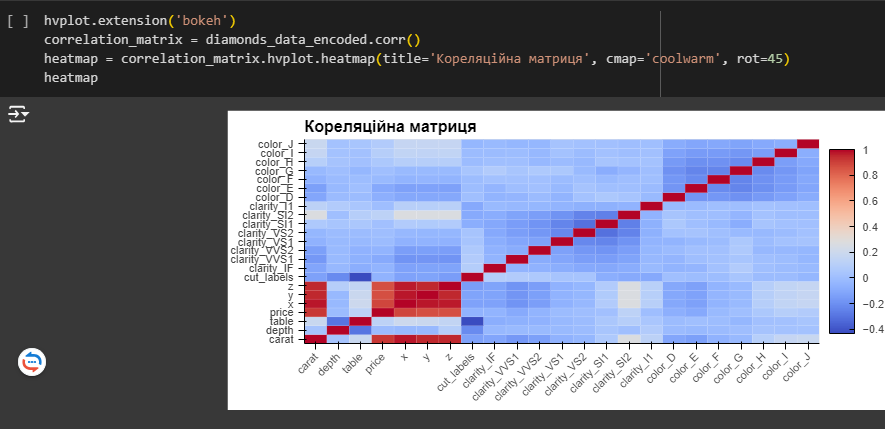
На цьому графіку бачимо, що ціна залежить від карат. Чим більша вага, тим більша ціна.





А значення depth та table майже не впливають на ціну.

1. Кореляційний аналіз



Матриця показує, що найбільша кореляція відбувається між значеннями карат, ціною та x, y, z, що є логічним. Чим більша вага діаманту, тим більший розмір його сторін і відповідно ціна.

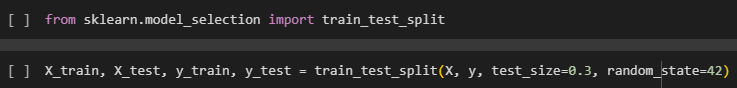
Також матриця показує, що чим більший table, тобто діаметр верхньої площини діаманту, тим менша оцінка діаманту.

Категоріальні ознаки як cut, color, clarity також мають помітний вплив.

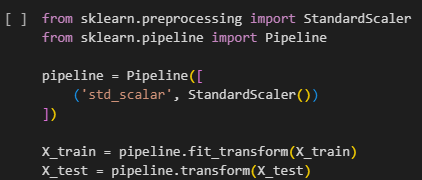
1. Вибираємо цільову змінну (price) та ознаки.



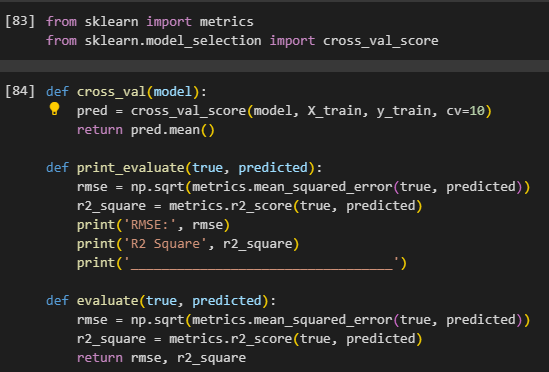
1. Розділяємо дані на тренувальну і валідаційну вибірку. Для цього є функція train\_test\_split в sklearn.



1. Стандартизуємо дані.

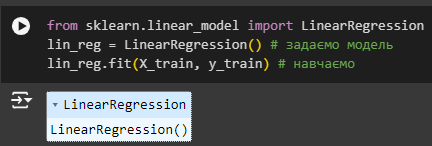


1. Створимо функції для майбутньої оцінки моделей з точки зору таких метрик якості як RMSE та R^2.

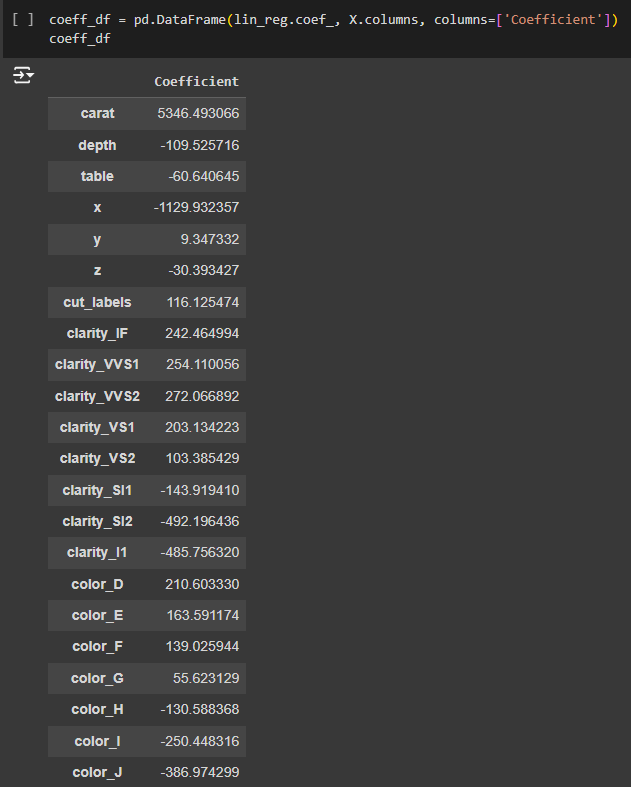


Лінійна регресія

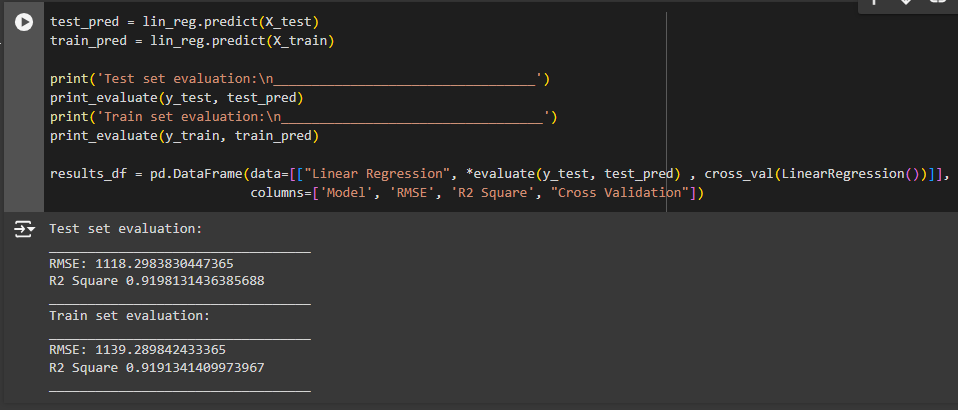
1. Задаємо модель та навчаємо її на тренувальних даних.



1. Можемо перевірити вагу (вплив) кожної незалежної змінної (функції) на залежну змінну (цільову змінну) за допомогою коефіцієнтів. Якщо коефіцієнт додатний, це означає, що збільшення значення відповідної незалежної змінної збільшує значення залежної змінної. Якщо коефіцієнт від'ємний, це означає, що збільшення значення відповідної незалежної змінної зменшує значення залежної змінної. Велике значення коефіцієнта означає великий вплив.

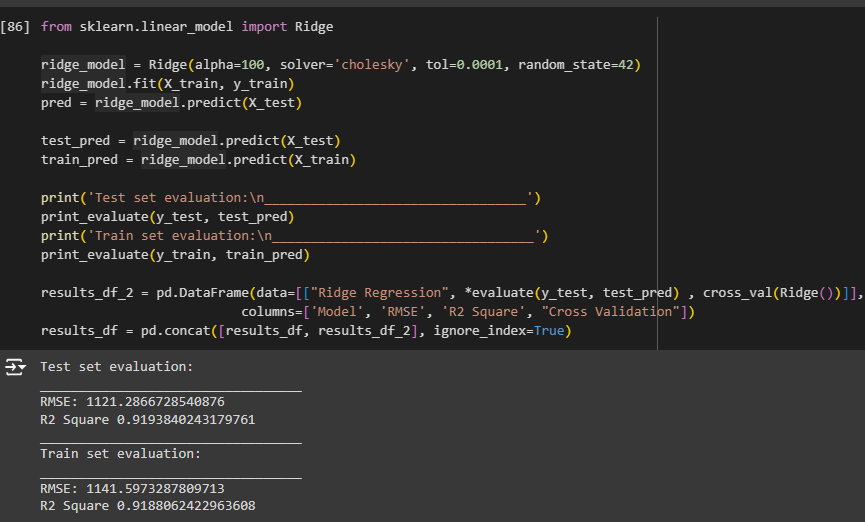


1. Оцінка лінійної моделі



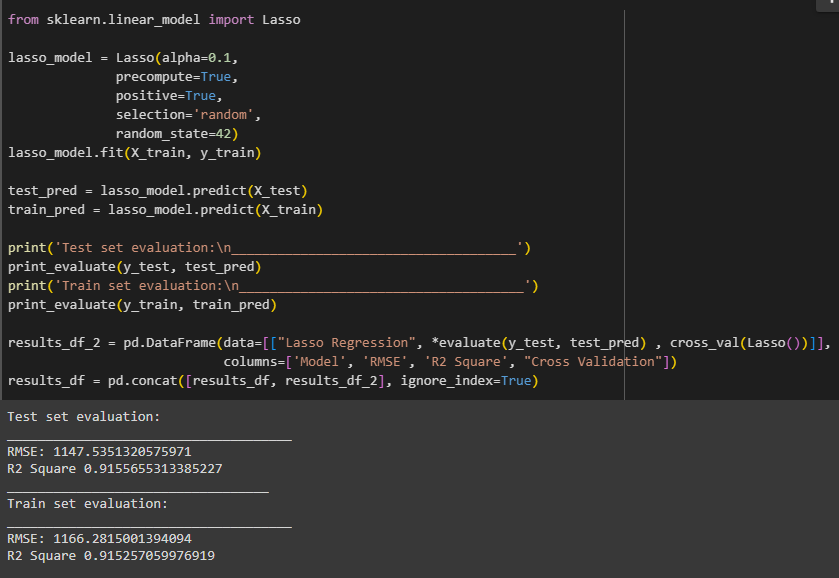
Ridge — це лінійна регресія з регуляризацією, яка додає штраф за великі коефіцієнти в моделі, щоб запобігти перенавчанню.

1. Тренуємо нову модель та виводимо оцінку якості на тренувальній і тестовій вибірках. Проводимо додатково крос-валідацію і додаємо результати до таблиці для порівняння з іншими моделями.

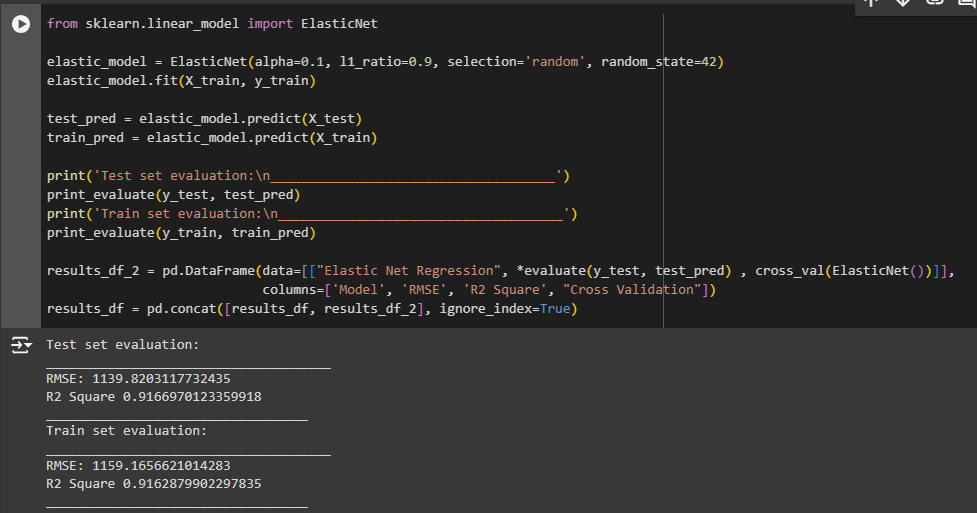


LASSO регресія

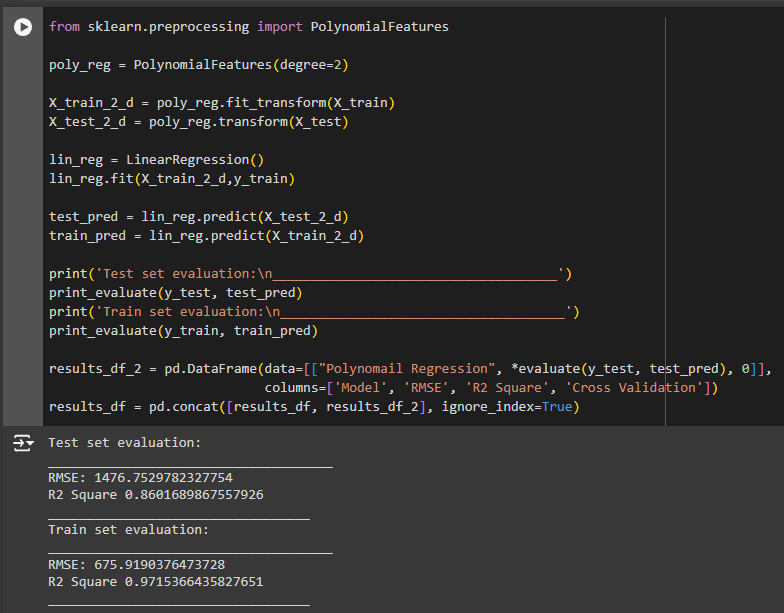
1. Тренуємо модель LASSO регресії та виводимо оцінку якості.



1. Тренуємо модель Elastic Net.



1. Тренуємо модель поліноміальної регресії.



**Висновок**:

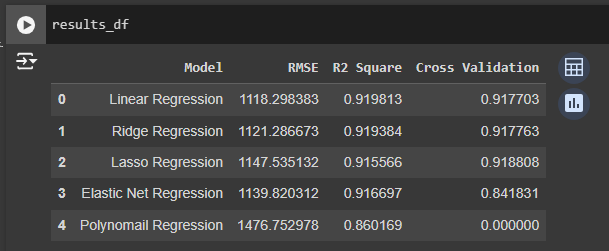
Лінійна регресія має найнижчий RMSE (1118.30) і високий R² (0.9198), що свідчить про те, що вона найбільш точно прогнозує дані серед усіх моделей. Крос-валідація також показала стабільний результат (0.9177).

Ridge регресія має трохи вищий RMSE (1121.29) порівняно з лінійною, але різниця незначна. Це показує, що додавання регуляризації дещо зменшило варіацію в моделі, але не погіршило її значно. Крос-валідація також практично не змінилась (0.9178).

Lasso регресія має ще вищий RMSE (1147.54) і трохи нижчий R² (0.9156). Це означає, що вона гірше справляється з прогнозуванням, ніж лінійна або Ridge регресія, хоча її результат крос-валідації (0.9188) дещо кращий.

Elastic Net також демонструє вищий RMSE (1139.82) і нижчий R² (0.9167), але найгірший результат крос-валідації (0.8418). Це може свідчити про те, що модель перенавчилася або не адаптувалася належним чином до даних.

А поліноміальна регресія демонструє найгірші результати з усіх моделей. Високий RMSE (1476.75) і низький R² (0.8602) свідчать про те, що ця модель надмірно ускладнює залежності в даних, що призвело до погіршення якості прогнозів. Крос-валідація з результатом 0 також показує, що модель не дає адекватних результатів на інших підвибірках.



Отже, лінійна регресія та Ridge регресія є найкращими моделями для цих даних, оскільки вони мають найнижчий RMSE та найвищий R².