Міністерство освіти і науки України
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського"

Факультет інформатики та обчислювальної

техніки Кафедра інформатики та програмної

інженерії

Звіт

з лабораторної роботи №5 з дисципліни
«Аналіз даних в інформаційних системах»

«Регресійні моделі»

Варіант 7

Виконав студент ІП-12 Васильєв Єгор Костянтинович
(шифр, прізвище, ім'я, по батькові)

Перевірив Ліхоузова Тетяна Анатоліївна
(прізвище, ім'я, по батькові)

Київ 2023

Лабораторна робота № 5

Тема: Регресійні моделі

Для виконання лабораторної роботи було обрано мову програмування Python, та бібліотеки Numpy, Pandas і Matplotlib. Спочатку дані було завантажено та імпортовано до датафрейму.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.40000	0.70000	0.00000	1.90000	0.07600	11.00000	34.00000	0.99780	3.51000	0.56000	9.40000	5
1	7.80000	0.88000	0.00000	2.60000	0.09800	25.00000	67.00000	0.99680	3.20000	0.68000	9.80000	5
2	7.80000	0.76000	0.04000	2.30000	0.09200	15.00000	54.00000	0.99700	3.26000	0.65000	9.80000	5
3	11.20000	0.28000	0.56000	1.90000	0.07500	17.00000	60.00000	0.99800	3.16000	0.58000	9.80000	6
4	7.40000	0.70000	0.00000	1.90000	0.07600	11.00000	34.00000	0.99780	3.51000	0.56000	9.40000	5
5	7.40000	0.66000	0.00000	1.80000	0.07500	13.00000	40.00000	0.99780	3.51000	0.56000	9.40000	5
6	7.90000	0.60000	0.06000	1.60000	0.06900	15.00000	59.00000	0.99640	3.30000	0.46000	9.40000	5
7	7.30000	0.65000	0.00000	1.20000	0.06500	15.00000	21.00000	0.99460	3.39000	0.47000	10.00000	7
8	7.80000	0.58000	0.02000	2.00000	0.07300	9.00000	18.00000	0.99680	3.36000	0.57000	9.50000	7
9	7.50000	0.50000	0.36000	6.10000	0.07100	17.00000	102.00000	0.99780	3.35000	0.80000	10.50000	5
10	6.70000	0.58000	0.08000	1.80000	0.09700	15.00000	65.00000	0.99590	3.28000	0.54000	9.20000	5
11	7.50000	0.50000	0.36000	6.10000	0.07100	17.00000	102.00000	0.99780	3.35000	0.80000	10.50000	5
12	5.60000	0.61500	0.00000	1.60000	0.08900	16.00000	59.00000	0.99430	3.58000	0.52000	9.90000	5
13	7.80000	0.61000	0.29000	1.60000	0.11400	9.00000	29.00000	0.99740	3.26000	1.56000	9.10000	5
14	8.90000	0.62000	0.18000	3.80000	0.17600	52.00000	145.00000	0.99860	3.16000	0.88000	9.20000	5
15	8.90000	0.62000	0.19000	3.90000	0.17000	51.00000	148.00000	0.99860	3.17000	0.93000	9.20000	5
16	8.50000	0.28000	0.56000	1.80000	0.09200	35.00000	103.00000	0.99690	3.30000	0.75000	10.50000	7
17	8.10000	0.56000	0.28000	1.70000	0.36800	16.00000	56.00000	0.99680	3.11000	1.28000	9.30000	5
18	7.40000	0.59000	0.08000	4.40000	0.08600	6.00000	29.00000	0.99740	3.38000	0.50000	9.00000	4
19	7.90000	0.32000	0.51000	1.80000	0.34100	17.00000	56.00000	0.99690	3.04000	1.08000	9.20000	6
20	8.90000	0.22000	0.48000	1.80000	0.07700	29.00000	60.00000	0.99680	3.39000	0.53000	9.40000	6

Рисунок 3.1 – імпортовані дані у data frame

Було розділено дані на навчальну та тестову вибірки у співвідношенні 75 до 25 за допомогою методу train_test_split бібліотеки sklearn. Було створено три регресійні моделі для прогнозування параметру quality: лінійну одновимірну, яка використовує параметр з найбільшим коефіцієнтом кореляції з залежною змінною, багатовимірну лінійну та поліноміальну регресії, які використовують усі наявні параметри.

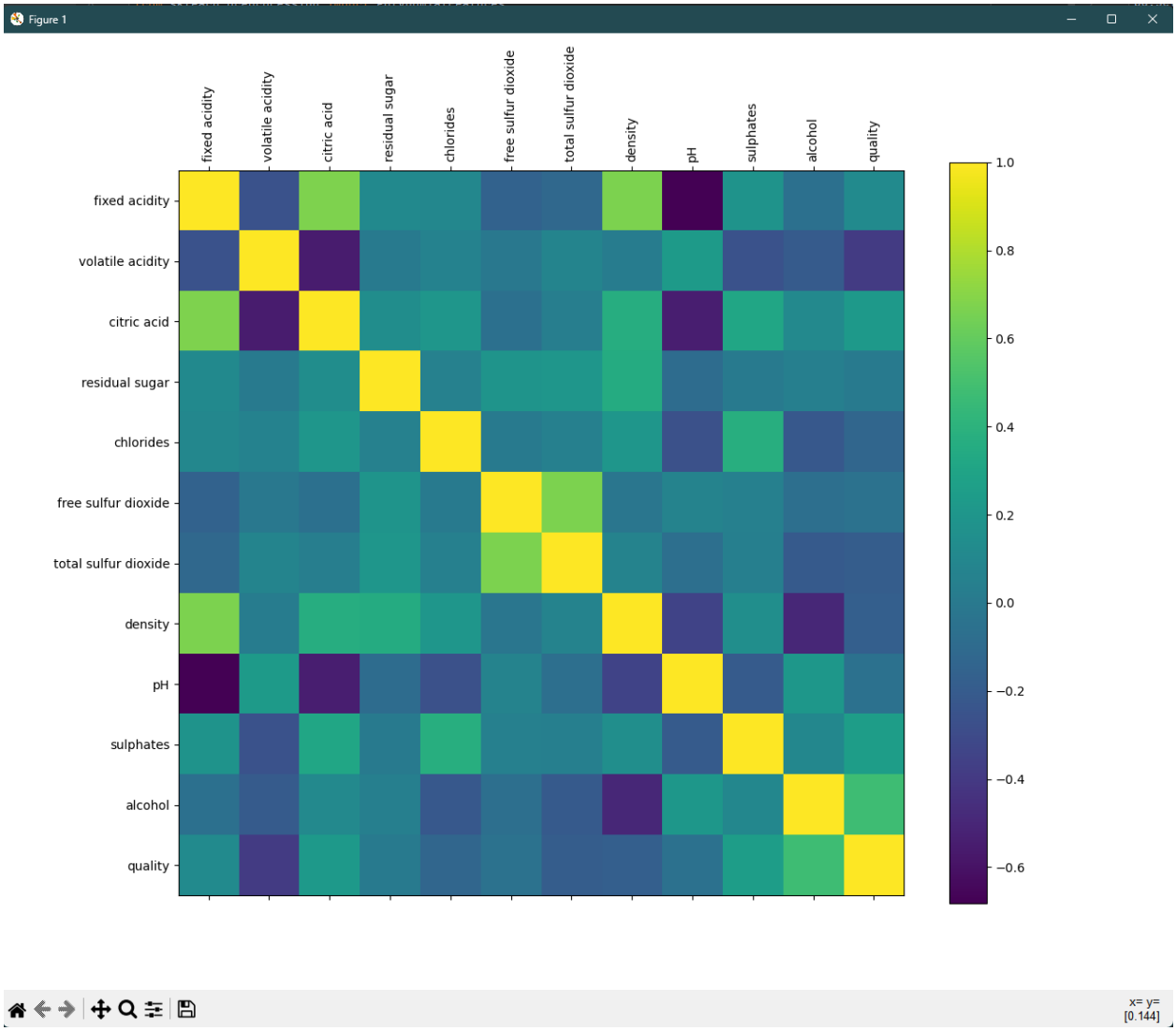

Рисунок 3.2 – кореляційна матриця

```
Parameter that correlates the most with the quality: alcohol, with coefficient 0.48
```

Рисунок 3.3 – параметр, який корелює найбільше з якістю вина

Після тренування трьох різних моделей було порівняно їх між собою, використовуючи тестову вибірку за допомогою вбудованих в sklearn метрик: середньоквадратична похибка, коефіцієнт детермінації (R-squared) та середня абсолютна похибка.

```
Mean squared error for linear regression with alcohol parameter: 0.483, R-squared metric: 0.274, mean absolute error: 0.553
Mean squared error for multivariate regression: 0.438, R-squared metric: 0.341, mean absolute error: 0.520
Mean squared error for polynomial regression: 0.464, R-squared metric: 0.303, mean absolute error: 0.523
```

Рисунок 3.4 – метрики досліджуваних регресій

Для другого завдання було завантажено нові дані, додано пропущену назву для одного стовпця, виправлено тип числових параметрів та отримано наступний датафрейм:

	Country	ISO	UA	Cql	Ie	Iec	Is
0	Albania	ALB	Албанія	0.97392	0.60535	0.53867	0.51011
1	Algeria	DZA	Алжир	0.78213	0.58722	0.34816	0.49799
2	Angola	AGO	Ангола	0.37234	0.27439	0.33212	0.34691
3	Argentina	ARG	Аргентина	0.88383	0.69969	0.28199	0.51882
4	Armenia	ARM	Вірменія	1.01650	0.71833	0.53565	0.48650
5	Australia	AUS	Австралія	1.45761	0.79152	0.72115	0.69241
6	Austria	AUT	Австрія	1.39356	0.77116	0.64008	0.69825
7	Azerbaijan	AZE	Азербайджан	0.91725	0.74825	0.47343	0.42516
8	Bangladesh	BGD	Банґладеш	0.40104	0.19428	0.38488	0.38603
9	Barbados	BRB	Барбадос	1.02251	0.35702	0.55919	0.60599
10	Belarus	BLR	Білорусія	0.93838	0.72663	0.32592	0.53434
11	Belgium	BEL	Бельґія	1.28754	0.69651	0.61062	0.65202
12	Belize	BLZ	Беліз	0.90560	0.62804	0.41642	0.53375
13	Benin	BEN	Бенін	0.55524	0.19595	0.44674	0.44830
14	Bhutan	BTN	Бутан	0.81033	0.46766	0.42775	0.53297
15	Bolivia, Plurinational State of	BOL	Болівія, Багатонаціональна Держава	0.76279	0.57791	0.31982	0.49653
16	Bosnia and Herzegovina	BIH	Боснія і Герцеґовина	0.78247	0.42396	0.43653	0.52686
17	Botswana	BWA	Ботсвана	1.03682	0.57039	0.58559	0.53027
18	Brazil	BRA	Бразілія	0.93537	0.70376	0.44217	0.50423
19	Bulgaria	BGR	Болгарія	1.13563	0.73429	0.55590	0.55450
20	Burkina faso	BFA	Буркіна-Фасо	0.45192	0.19526	0.41985	0.39242

Рисунок 3.5 – завантажені дані по країнам

Для дослідження на мультиколінеарність було виведено кореляційну матрицю для числових параметрів, з якої видно, що майже всі вони є мультиколінеарними (дуже високі відповідні коефіцієнти) .

	Cql	Ie	Iec	Is
Cql	1.000000	0.883664	0.875545	0.939172
Ie	0.883664	1.000000	0.619247	0.746320
Iec	0.875545	0.619247	1.000000	0.799211
Is	0.939172	0.746320	0.799211	1.000000

Рисунок 3.6 – кореляційна матриця для числових параметрів

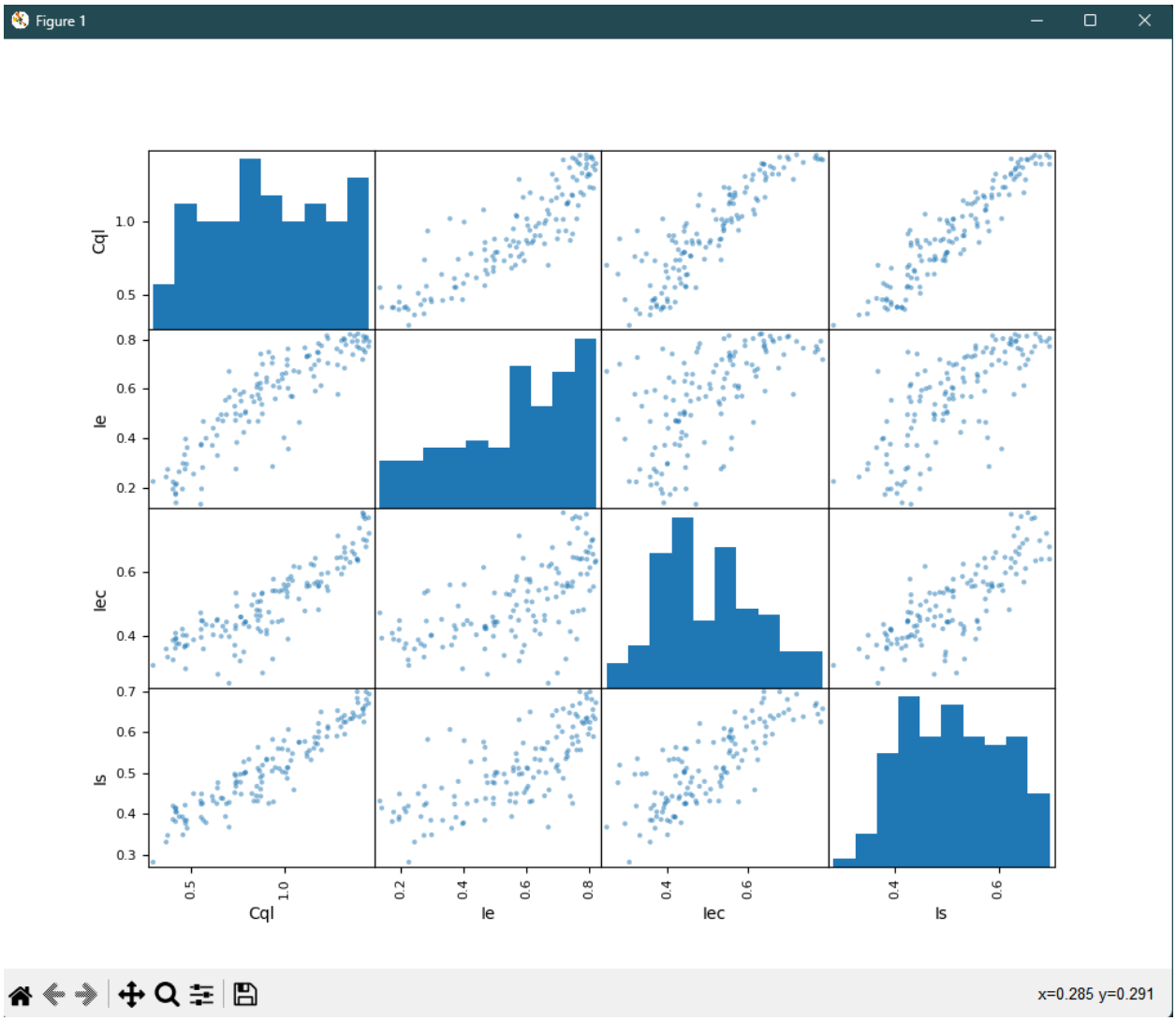

Рисунок 3.7 – діаграми розсіювання для кожної пари числових параметрів

Було побудовано три регресійні моделі для передбачення залежної змінної Cql: лінійну одновимірну, що використовує лише один параметр – Is, лінійну багатовимірну регресію та поліноміальну, що використовують усі параметри. Аналогічно до першого завдання, після тренування трьох різних моделей було порівняно їх між собою, використовуючи тестову вибірку за допомогою вбудованих в sklearn метрик:

```
Mean squared error for linear regression with all parameters: 0.00038, R-squared metric: 0.988, mean absolute error: 0.017
Mean squared error for linear regression with only "Is" parameter: 0.01141, R-squared metric: 0.623, mean absolute error: 0.078
Mean squared error for polynomial regression: 0.00005, R-squared metric: 0.998, mean absolute error: 0.006
```

Рисунок 3.8 – метрики досліджуваних регресій

Висновок: досліджуючи данні про різні види червоного вина було встановлено, що багатовимірна лінійна регресія є найоптимальнішою для прогнозування якості вина маючи усі інші його параметри. Для даних з країнами, найоптимальнішою для прогнозування параметру Cql виявилась поліноміальна регресія.