Міністерство освіти і науки України
Національний технічний університет України
“Київський політехнічний інститут імені Ігоря Сікорського”
Факультет інформатики та обчислювальної техніки

Курсова робота

з дисципліни
“Аналіз даних в інформаційних системах”

Варіант №28

Виконав студент Флорчук Назарій Петрович

Перевірив викладач Ліхоузова Тетяна Анатоліївна

Завдання:

1. За вказаним набором вхідних змінних побудувати регресійні моделі для прогнозування значення вихідної змінної (якість).
2. Використати лінійну та три варіанти поліноміальної регресії (на свій вибір).
3. За результатами тестування обрати найкращу модель.

1. Постановка задачі.

Варіант	Вхідні змінні
28	5 6 7 9 10 11

Вихідні дані:

Набір даних про червоні варіанти португальського вина «Vinho Verde» [P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009]. З міркувань конфіденційності та логістики доступні лише фізико-хімічні (вхідні) та органолептичні (вихідні) змінні (наприклад, немає даних про сорти винограду, марку вина, продажну ціну вина тощо). Класи впорядковані та незбалансовані (наприклад, нормальних вин набагато більше, ніж відмінних чи поганих).

Файл із даними знаходиться за посиланням.

Опис даних файлу:

Назва	Опис
Вхідні змінні (на основі фізико – хімічних тестів)	
1 – фіксована кислотність	більшість кислот, пов'язаних з винами, або фіксовані, або нелеткі (легко не випаровуються)
2 – летюча кислотність	кількість оцтової кислоти у вині, яка при надто високому рівні може призвести до неприємного смаку оцту
3 – лимонна кислота	лимонна кислота, що міститься в невеликих кількостях, може надати вин «свіжість» та аромат
4 – залишковий цукор	кількість цукру, що залишилася після припинення бродіння, рідко можна знайти вина із вмістом цукру менше 1 грама на літр, а вина із вмістом цукру більше 45 грамів на літр вважаються солодкими
5 – хлориди	кількість солі у вині
6 – вільний діоксид сірки	вільна форма SO2 існує у рівновазі між молекулярним SO2 (у вигляді розчиненого газу) та бісульфіт-іоном; запобігає зростанню мікробів та окисленню вина

7 – загальний діоксид сірки	кількість вільних та пов'язаних форм S02; при низьких концентраціях SO2 практично не виявляється у вині, але при концентраціях вільного SO2 більше 50 частин на мільйон SO2 стає помітним в ароматі та смаку вина
8 – щільність	щільність води близька до щільності води залежно від відсоткового вмісту спирту та цукру
9 – pH	описує, наскільки кислим чи лужним є вино за шкалою від 0 (дуже кисле) до 14 (дуже лужне); більшість вин мають показник pH від 3 до 4 за шкалою pH
10 – сульфати	добавка до вина, яка може сприяти підвищенню рівня сірчистого газу (SO2), що діє як протимікробний та антиоксидантний засіб
11 – спирт	відсоток утримання алкоголю у вині
Вихідна змінна (на основі сенсорних даних)	
12 – якість (оцінка від 0 до 10)	оцінка (від 0 до 10)

2. Вибір моделей.
   2.1. Вибір ознак, які будуть використані для аналізу.
   Згідно завдання та варіанту, ознаками для аналізу будуть:

Назва	Опис
Вхідні змінні (на основі фізико – хімічних тестів)	
5 – хлориди	кількість солі у вині
6 – вільний діоксид сірки	вільна форма SO2 існує у рівновазі між молекулярним SO2 (у вигляді розчиненого газу) та бісульфіт-іоном; запобігає зростанню мікробів та окисленню вина
7 – загальний діоксид сірки	кількість вільних та пов'язаних форм S02; при низьких концентраціях SO2 практично не виявляється у вині, але при концентраціях вільного SO2 більше 50 частин на мільйон SO2 стає помітним в ароматі та смаку вина
9 – pH	описує, наскільки кислим чи лужним є вино за шкалою від 0 (дуже кисле) до 14 (дуже лужне); більшість вин мають показник pH від 3 до 4 за шкалою pH
10 – сульфати	добавка до вина, яка може сприяти підвищенню рівня сірчистого газу (SO2), що діє як протимікробний та антиоксидантний засіб
11 – спирт	відсоток утримання алкоголю у вині
Вихідна змінна (на основі сенсорних даних)	

12 – якість (оцінка від 0 до 10)	оцінка (від 0 до 10)

2.2. Визначення моделей, які можуть бути використані.
Згідно завдання, буде використано лінійну регресійну модель (з використанням усіх 6-ти вхідних змінних, тобто множину лінійну регресійну модель), та кілька поліноміальних регресійних моделей (з використанням усіх 6-ти вхідних змінних та із різними ступенями ознак полінома).

Висновки щодо побудови моделей (яка з обраних моделей підходить краще), будуть зроблені на основі ***MSE*** (середньої квадратичної помилки), $\boldsymbol{R^2}$ (коефіцієнта детермінації) та частково на основі побудованих графіків порівнянь оригінальних (правдивих) та прогнозованих значень.

Якщо порівнювати моделі, то для даних краще підходить модель з більшим значенням коефіцієнта детермінації ($\boldsymbol{R^2}$).

При порівнянні моделей, модель з найменшим значенням середньої квадратичної помилки (***MSE***) краще підходить для даних.

2.3. Підготовка даних для навчання та верифікації моделей.
Згідно аналізу даних із файлу за посиланням, перевірено, що усі поля мають числовий формат та пусті значення полів відсутні.

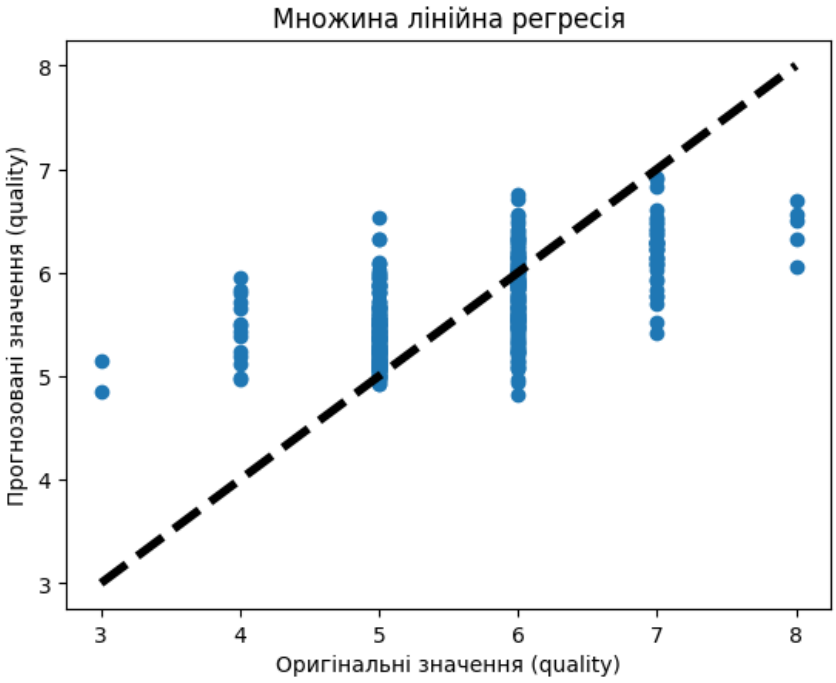
Такі властивості даних цілком підходять для подальшої роботи з даними, без її додаткової обробки. Це значно спрощує подальшу роботу з даними, та покращує точність отриманих результатів.

Дані нормалізації не потребують.

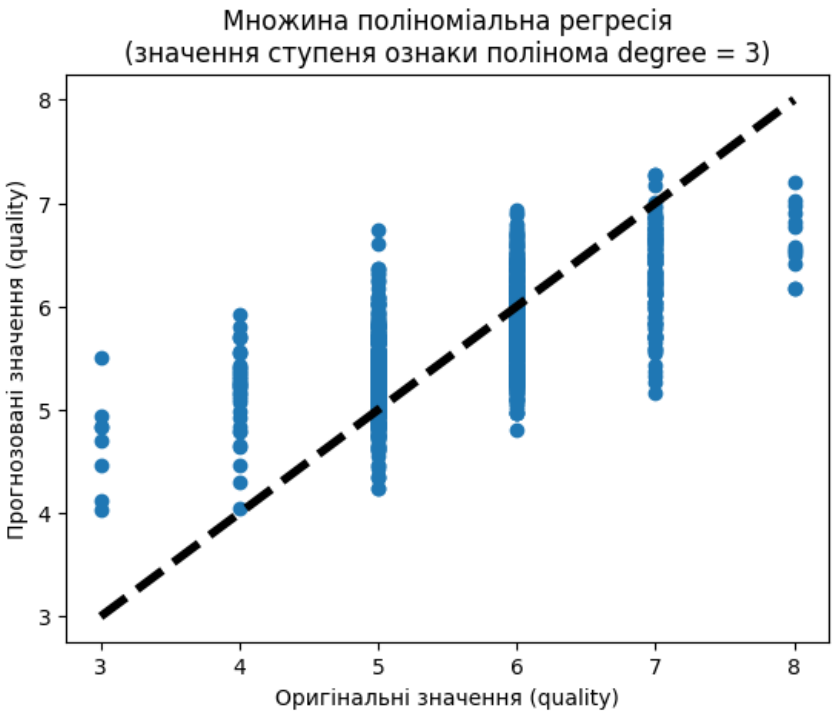
2.4. Формування моделей. Вибір оптимального класу складності моделей.
Складність – є однією із характеристик моделей. Чим більше властивостей використовується для опису моделі, тим складнішою модель є. І не завжди чим складніша модель, тим точнішою вона є.

Лінійні моделі відносяться до класу регресійних моделей з одним рівнянням. До нього також належать і нелінійні моделі, які представленні нелінійними функціями (гіпербола, парабола, степенева тощо).

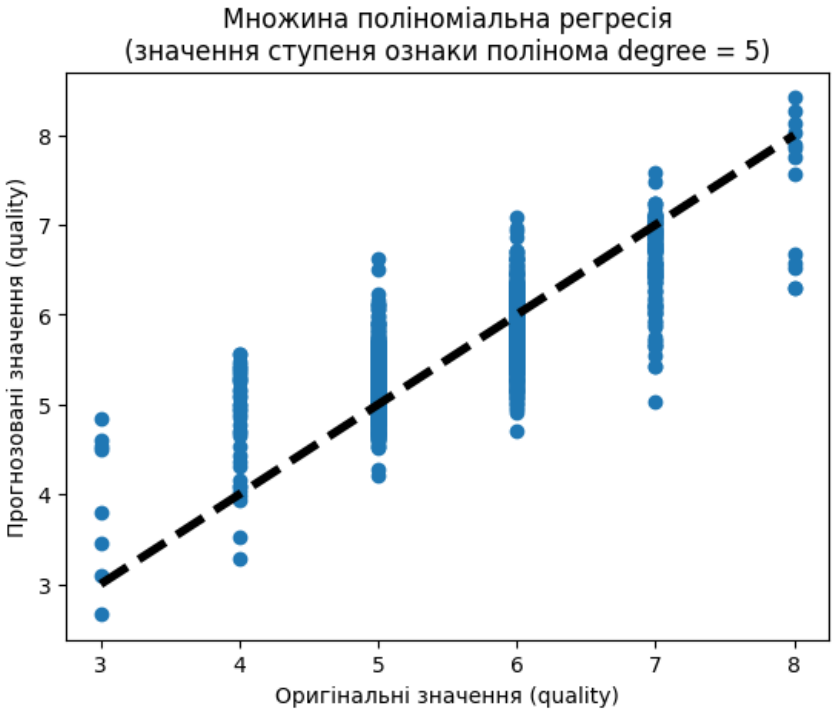
2.5. Верифікація моделей.
Розрахунки та детальне їх пояснення, можна переглянути у файлі програми.



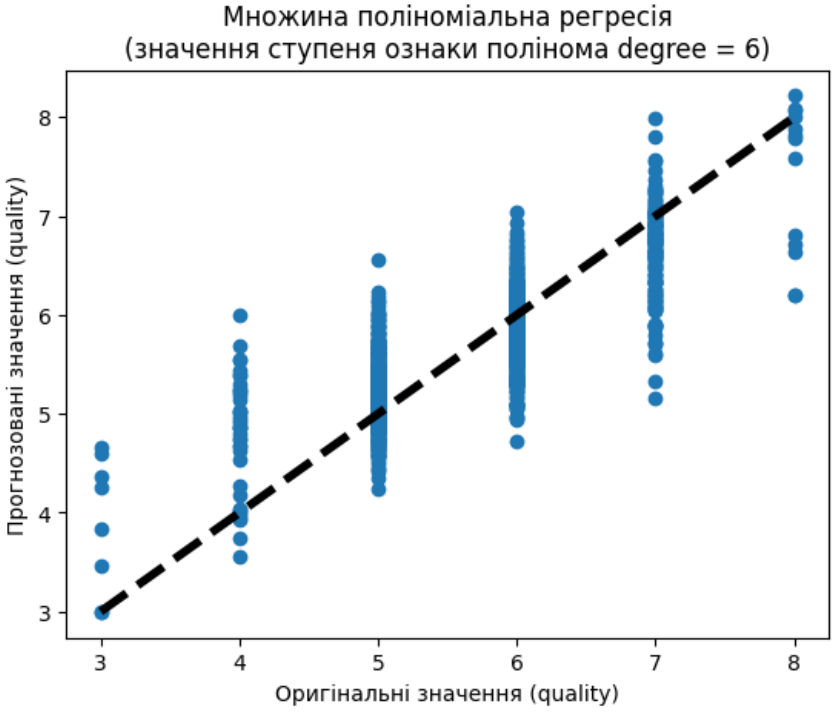
MSE = 0.39597234123968106
R^2 = 0.3078104782338481



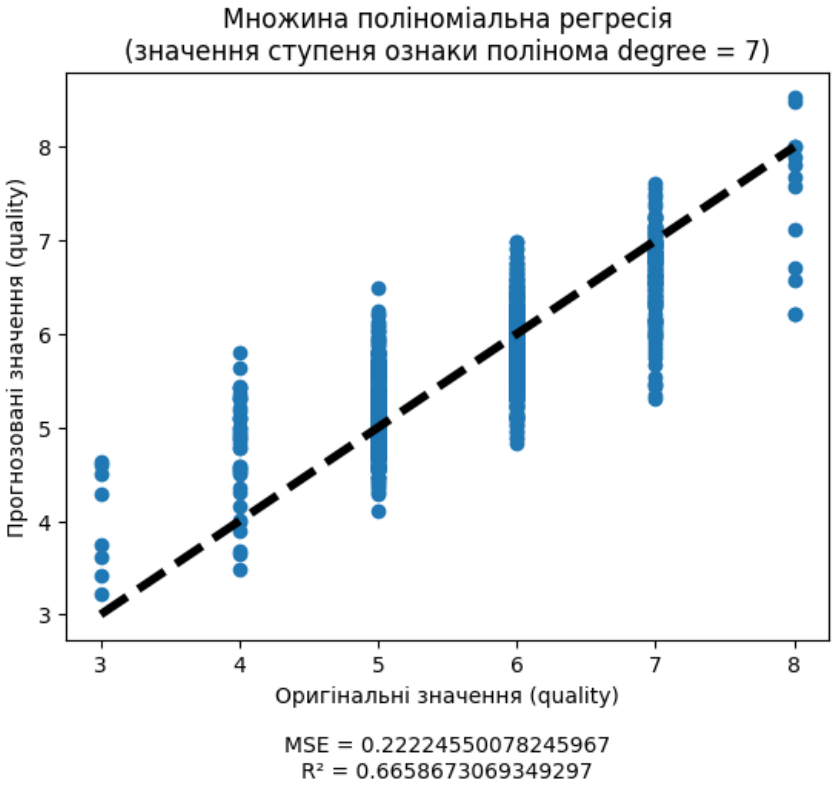
MSE = 0.35739782498920436
R^2 = 0.4626739468794392



MSE = 0.23637531799257247
R^2 = 0.6446239798020512



MSE = 0.21076818197823458
R^2 = 0.6831227628506642



3. Висновки щодо якості побудованих моделей.
   Згідно проведеного аналізу даних, чітко видно з графіків, що у регресійних моделях із багатьма змінними (множинних регресійних моделях) (якість вина залежить від багатьох його характеристик), лінійна регресійна модель, показує значно гірший результат у плані прогнозування, на відміну від поліноміальних моделей.

   Коефіцієнт детермінації лінійної регресійної моделі (на основі даних аналізу) $\boldsymbol{R^2}$ = 0.3078104782338481, що є не задовільним значенням, і навряд зможе використовуватись у реальних умовах.

   Щодо поліноміальних регресійних моделей, то із зміною ступеня ознаки полінома, при його значені **6**, коефіцієнт детермінації (на основі даних аналізу) найвищий, і рівний $\boldsymbol{R^2}$ = 0.6831227628506642. При більшому і меншому значенні ступеня, значення коефіцієнту детермінації падає.

   Щодо значення середньої квадратичної помилки (***MSE***), то у поліноміальної регресійної моделі, із значенням ступеня ознаки полінома **6**, значення середньої квадратичної помилки найнижче у порівнянні із іншими ступенями ознак полінома, та у порівнянні із значенням середньої квадратичної помилки у лінійної регресійної моделі – ***MSE*** = 0.21076818197823458 у поліноміальної регресійної моделі, проти ***MSE*** = 0.39597234123968106 у лінійної регресійної моделі.

Як результат, у даному випадку, із даними що надані, найкращою моделлю для прогнозування є поліноміальна регресійна модель із ступенем ознаки полінома **6**.