УДК 004.62 + 634.8

І. В. Демков
С. О. Жуков

РОЗВІДУВАЛЬНИЙ АНАЛІЗ ДАНИХ ДЛЯ ПЕРЕДБАЧЕННЯ ЯКОСТІ ВИНА ЗА ЙОГО ВЛАСТИВОСТЯМИ

Вінницький національний технічний університет

Анотація

Робота присвячена підготовці та розвідувальному аналізу даних португальського вина "Vinho Verde" для подальшого використання в інформаційній технології передбачення якості вина методами машинного навчання. Проведено статистичний аналіз датасету, досліджено розподіли змінних, виявлено кореляційні зв'язки та визначено основні фактори, що впливають на якісні характеристики вина.

Ключові слова: розвідувальний аналіз даних, передбачення якості, вино, кореляційний аналіз, статистичні показники.

Abstract

The work is devoted to the preparation and exploratory data analysis of Portuguese "Vinho Verde" wine for further use in information technology for predicting wine quality using machine learning methods. Statistical analysis of the dataset was conducted, variable distributions were studied, correlation relationships were identified, and the main factors affecting wine quality characteristics were determined.

Keywords: exploratory data analysis, quality prediction, wine, correlation analysis, statistical indicators.

Вступ

Сучасний розвиток інформаційних технологій дозволяє ефективно застосовувати методи машинного навчання для аналізу та передбачення різноманітних характеристик продуктів. Особливо це актуально для виноробної галузі, де визначення якості вина традиційно проводиться експертами-дегустаторами на основі суб'єктивних оцінок. Використання методів інтелектуального аналізу даних дозволяє виявити приховані закономірності між хімічними властивостями вина та його фінальною якістю, що дозволяє покращити процес його виготовлення.

Метою даного дослідження є проведення розвідувального аналізу даних для виявлення факторів, що визначають якість вина, та підготовка даних для подальшого машинного навчання. Основою дослідження виступає відкритий датасет з хімічними характеристиками та експертними оцінками португальського вина "Vinho Verde" [1].

Розвідувальний аналіз

Для проведення аналізу використано відкритий набір даних, опублікований на платформі Kaggle, що містить інформацію про хімічні властивості червоного португальського вина та його експертні оцінки. Датасет складається з 1599 записів та містить 12 числових полів, що представляють різні хімічні показники та фінальну оцінку якості за 10-бальною шкалою (рис. 1).

Основні атрибути датасету включають:

- Fixed acidity - фіксована кислотність
- Volatile acidity - летка кислотність
- Citric acid - вміст лимонної кислоти
- Residual sugar - залишковий цукор
- Chlorides - вміст хлоридів
- Free sulfur dioxide - вільний діоксид сірки
- Total sulfur dioxide - загальний діоксид сірки
- Density - густина
- pH - показник кислотності
- Sulphates - вміст сульфатів
- Alcohol - вміст алкоголю

– Quality - експертна оцінка якості

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1599 entries, 0 to 1598
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   fixed acidity         1599 non-null   float64
 1   volatile acidity      1599 non-null   float64
 2   citric acid           1599 non-null   float64
 3   residual sugar        1599 non-null   float64
 4   chlorides             1599 non-null   float64
 5   free sulfur dioxide   1599 non-null   float64
 6   total sulfur dioxide  1599 non-null   float64
 7   density               1599 non-null   float64
 8   pH                    1599 non-null   float64
 9   sulphates             1599 non-null   float64
 10  alcohol               1599 non-null   float64
 11  quality               1599 non-null   int64
dtypes: float64(11), int64(1)
memory usage: 150.0 KB
```

Рисунок 1 – Вид вхідних значень

Першим етапом аналізу стало дослідження основних статистичних показників кожної змінної. Виявлено, що датасет не містить пропущених значень, що спрощує подальший аналіз. Визначено середні значення, медіани, стандартні відхилення та інші показники для всіх змінних (рис. 2). Ці показники будуть використані у подальшому формуванні суджень.

	count	mean	std	min	25%	50%	75%	max
fixed acidity	1599.0	8.320	1.741	4.600	7.100	7.900	9.200	15.900
volatile acidity	1599.0	0.528	0.179	0.120	0.390	0.520	0.640	1.580
citric acid	1599.0	0.271	0.195	0.000	0.090	0.260	0.420	1.000
residual sugar	1599.0	2.539	1.410	0.900	1.900	2.200	2.600	15.500
chlorides	1599.0	0.087	0.047	0.012	0.070	0.079	0.090	0.611
free sulfur dioxide	1599.0	15.875	10.460	1.000	7.000	14.000	21.000	72.000
total sulfur dioxide	1599.0	46.468	32.895	6.000	22.000	38.000	62.000	289.000
density	1599.0	0.997	0.002	0.990	0.996	0.997	0.998	1.004
pH	1599.0	3.311	0.154	2.740	3.210	3.310	3.400	4.010
sulphates	1599.0	0.658	0.170	0.330	0.550	0.620	0.730	2.000
alcohol	1599.0	10.423	1.066	8.400	9.500	10.200	11.100	14.900
quality	1599.0	5.636	0.808	3.000	5.000	6.000	6.000	8.000

Рисунок 2 – Обраховані статистичні показники змінних

Наступним кроком було визначення характеру розподілу змінних через обчислення асиметрії та ексцесу [2, 3]. Більшість змінних демонструють правосторонню асиметрію, що свідчить про схильність до аномально високих значень (рис. 3). Особливо виділяються змінні chlorides, residual sugar та sulphates, які мають надзвичайно велику кількість даних суттєво далі нормального розподілу. Це може вказувати на наявність прихованих підгруп у даних.

	skewness	skewness_z_score	skewness_p_value	kurtosis	kurtosis_z_score	kurtosis_p_value
fixed acidity	0.981830	13.649470	0.000000	1.124860	6.182460	0.000000
volatile acidity	0.670960	10.048170	0.000000	1.217960	6.515650	0.000000
citric acid	0.318040	5.093990	0.000000	-0.790280	-11.229000	0.000000
residual sugar	4.536390	32.487210	0.000000	28.524440	21.561660	0.000000
chlorides	5.675020	35.467870	0.000000	41.581710	22.915850	0.000000
free sulfur dioxide	1.249390	16.256770	0.000000	2.013490	8.830430	0.000000
total sulfur dioxide	1.514110	18.479940	0.000000	3.794170	12.079700	0.000000
density	0.071220	1.166600	0.243370	0.927410	5.417270	0.000000
pH	0.193500	3.144950	0.001660	0.800670	4.877910	0.000000
sulphates	2.426390	24.311280	0.000000	11.679880	17.772340	0.000000
alcohol	0.860020	12.318610	0.000000	0.195650	1.558780	0.119050
quality	0.217600	3.528320	0.000420	0.292030	2.193940	0.028240

Рисунок 3 – Визначення асиметрії та ексцесу

Для візуального аналізу розподілів змінних побудовано гістограми та графіки щільності розподілу (рис. 4). Вони підтверджують виявлені раніше особливості - схильність до вершинності та наявність викидів. Для змінної quality, проведено додатковий аналіз з використанням Q-Q графіка (рис. 5), який показав, що розподіл оцінок близький до нормального, але має дещо важкі хвости та незначне зміщення в області оцінки 5 [4].

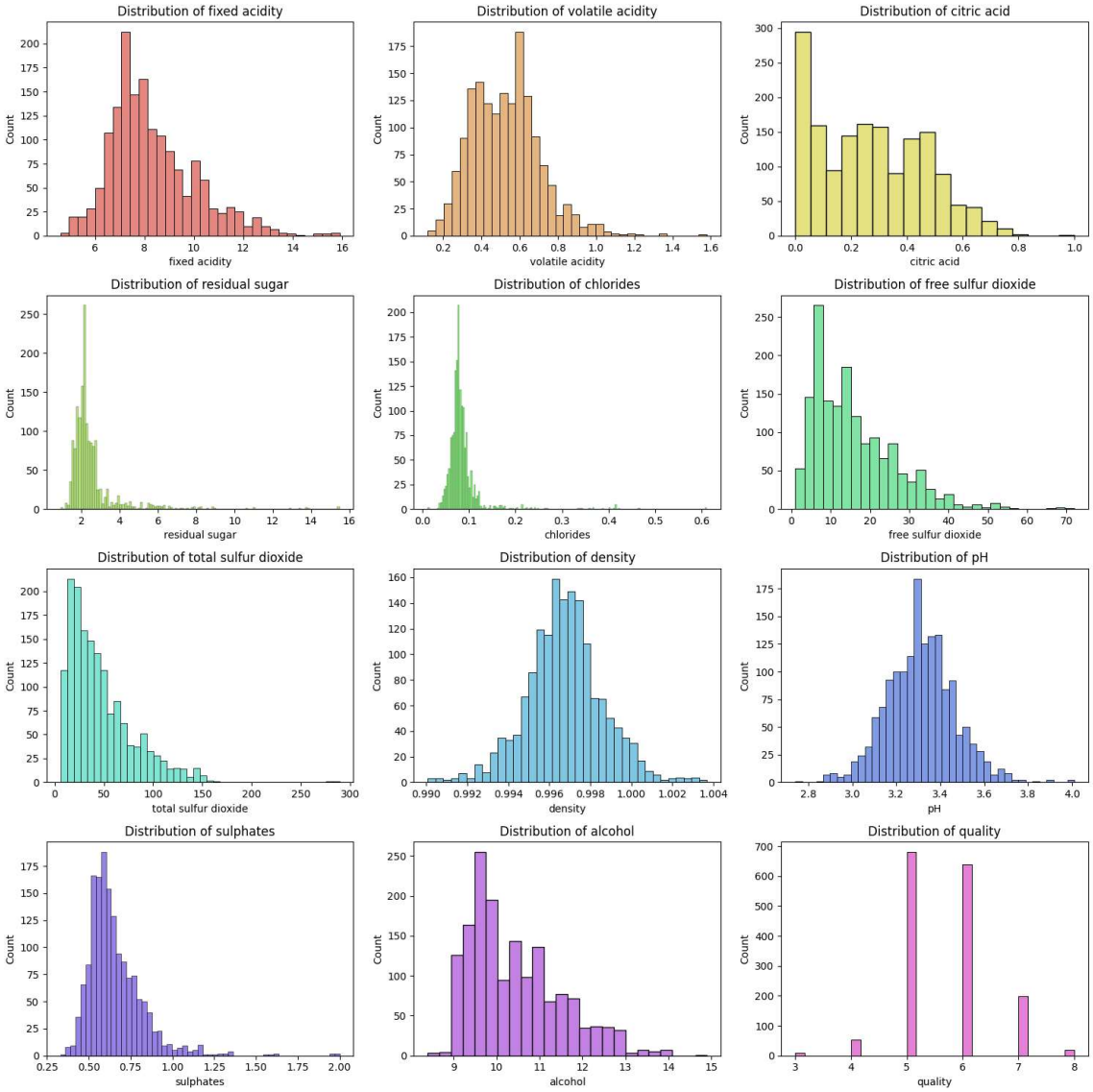

Рисунок 4 – Розподіли чисельних змінних

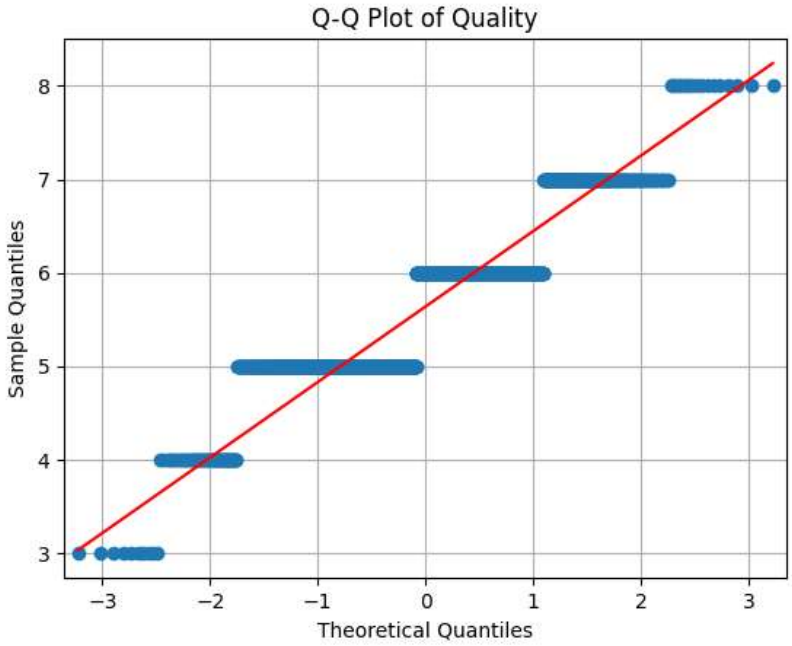

Рисунок 5 – Q-Q plot цільової змінної quality

Для глибшого аналізу впливу різних факторів на якість вина створено категоріальну змінну "alcohol hardness" на основі квартилів вмісту алкоголю (рис. 6). Тест хі-квадрат між створеною змінною та цільовою змінною quality показав значення 439.86, тобто наявність статистично значущого зв'язку (p-value близько до нуля), що підтверджує гіпотезу про залежність якості вина від вмісту алкоголю.

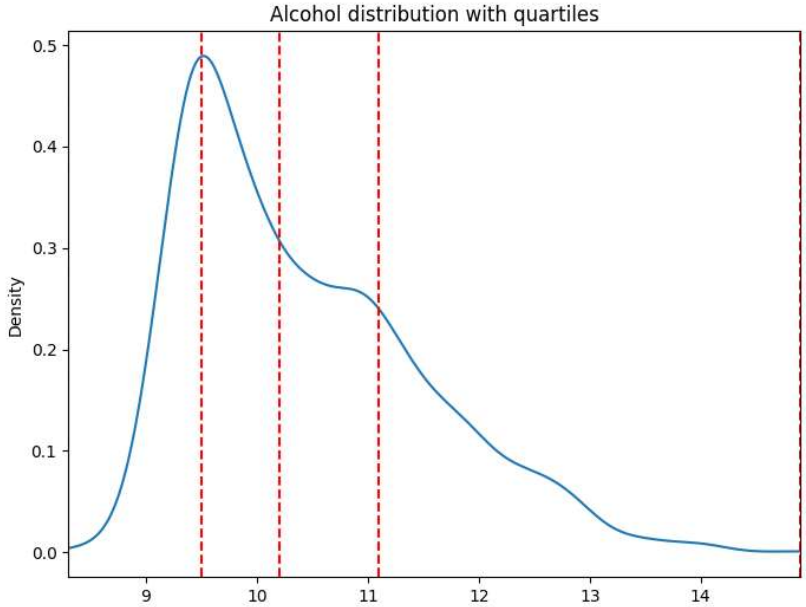

Рисунок 6 – Ілюстрація поділу цільової змінної на категорії

Побудовано коробкові графіки для всіх змінних, які дозволили візуально оцінити розподіл даних та наявність викидів (рис. 7). Особливу увагу приділено зміні розподілів змінних в залежності від експертної оцінки. Виявлено, що зниження volatile acidity до рівня 0.4, підтримка citric acid між 0.3 та 0.5, зниження density, підтримка sulphates між 0.7 та 0.8 та збільшення вмісту алкоголю позитивно впливають на якість вина.

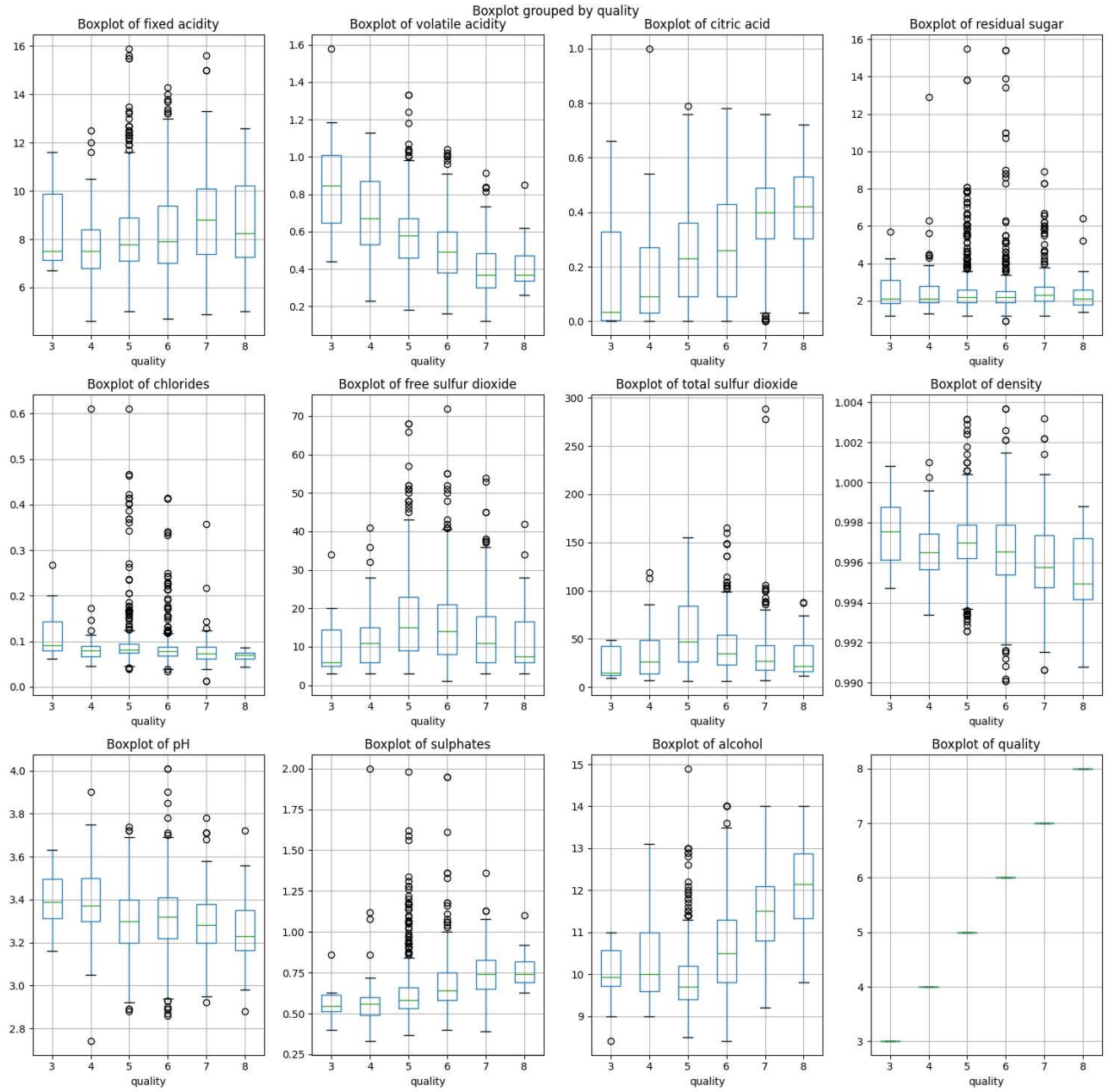

Рисунок 7 – Коробкові графи чисельних змінних з поділом на категорії цільової змінної

Проведено також статистичне тестування гіпотези про відмінність змінної residual sugar (як змінної з найбільшою кількістю викидів) між оцінками якості 5 та 6, які є найчастішими у датасеті. Т-тест Стьюдента показав відсутність статистично значущої різниці (p-value = 0.504), що може свідчити про суб'єктивність експертних оцінок у цьому діапазоні та використання оцінок 5 та 6 як "безпечного" варіанту.

Для дослідження зв'язків між змінними обчислено коваріацію та кореляцію. Побудовано теплокарту кореляцій (рис. 8), яка дозволила виявити основні залежності між змінними. Найсильніший позитивний зв'язок з якістю вина демонструють змінні alcohol (r = 0.48), sulphates (r = 0.25) та citric acid (r = 0.23). Негативний вплив має volatile acidity (r = -0.39).

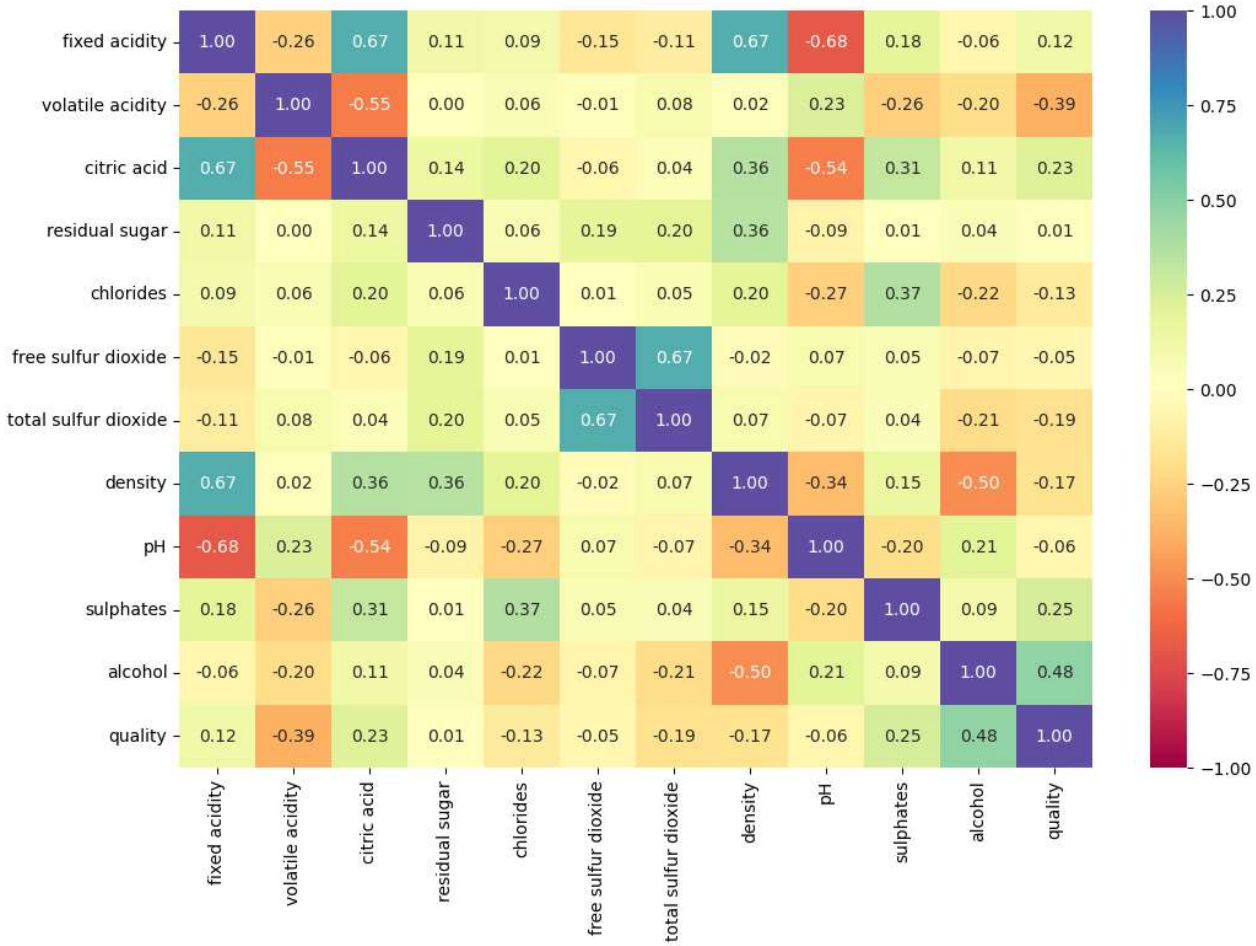

Рисунок 8 – Теплокарта кореляцій між чисельними змінними

Додатково проведено аналіз типу залежності між вмістом алкоголю та якістю вина з використанням як параметричних, так і непараметричних методів обчислення кореляції (рис. 9). Результати показали, що цей зв'язок має лінійний та монотонний характер.

```
Pearson r 0.476 with p-value 0.000
Spearman rho 0.479 with p-value 0.000

Pearson's r:
Transformation: x          0.476
Transformation: 1/x        nan
Transformation: x**2       0.452
Transformation: x**3       0.402
Transformation: log(x)     nan
Transformation: sqrt(x)    0.473
Transformation: exp(x)     0.469
Transformation: log(1/x)   nan
```

Рисунок 9 – Визначення кореляції Спірмена між цільовою змінною та змінною з найбільшою позитивною кореляцією, перевірка форми даної залежності.

За допомогою кластерграми [5] встановлено, що дані природньо розділяються на приблизно 5 груп (рис. 10). Графік паралельних координат (рис. 11) додатково використано для виявлення тенденцій у даних, однак він не показав чіткого групування за експертними оцінками.

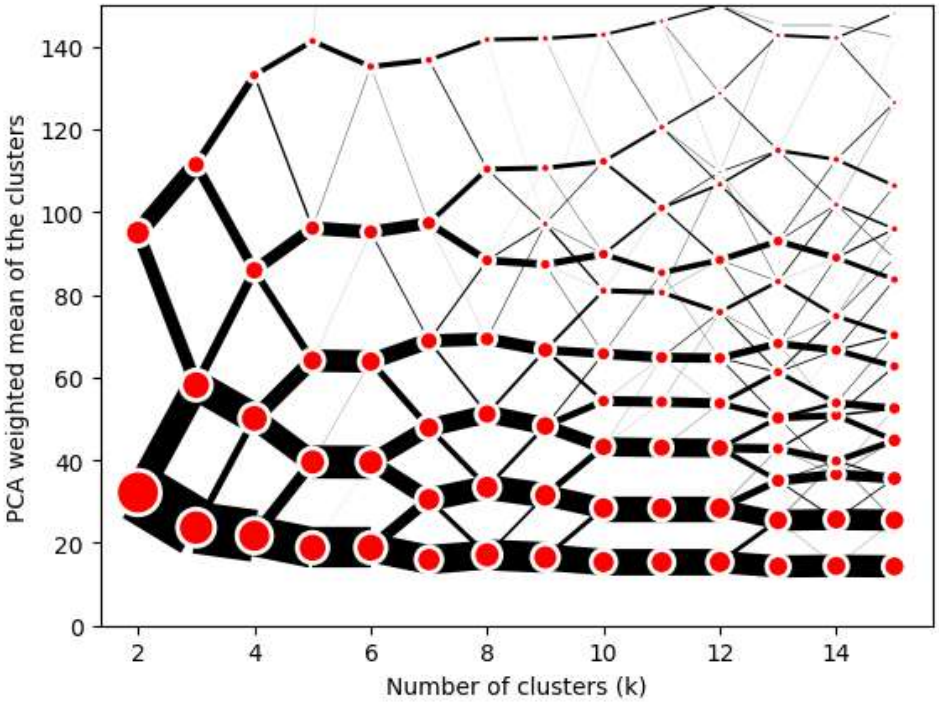

Рисунок 10 – Кластерграма даних

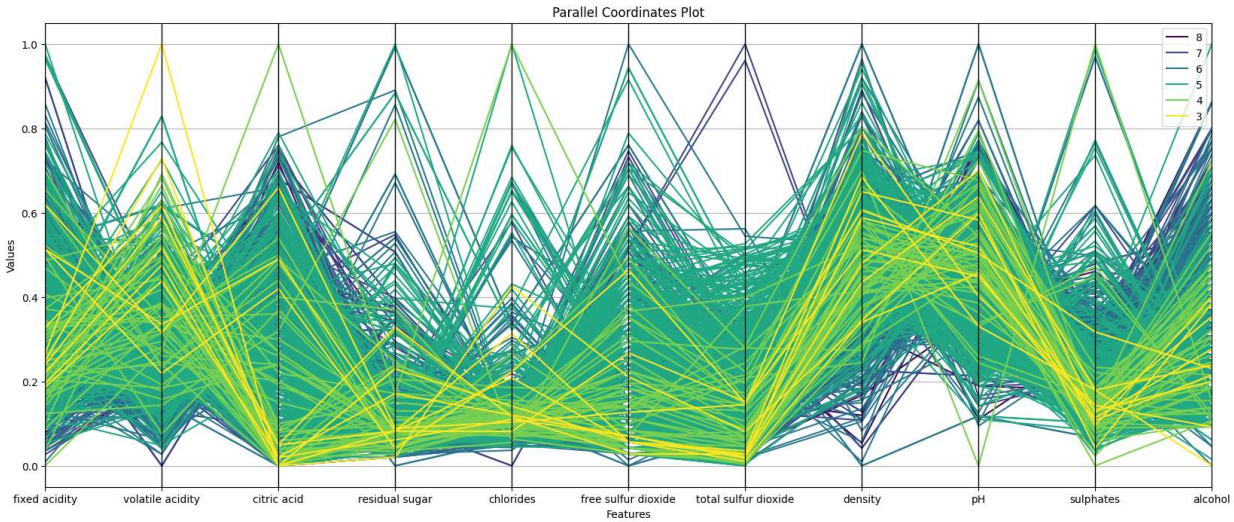

Рисунок 11 – Графік паралельних координат з кольоровим поділом на якість вина

З отриманих проміжних результатів можна сформувати загальні висновки про вплив хімічних та органолептичних властивостей вина на його експертну оцінку.

Висновки

У ході розвідувального аналізу даних властивостей вина "Vinho Verde" виявлено ряд закономірностей та особливостей, які мають значення для подальшого моделювання та передбачення якості. Встановлено, що найбільший вплив на якість вина мають вміст алкоголю, сульфатів, лимонної кислоти (позитивний вплив) та летка кислотність (негативний вплив).

Аналіз розподілів змінних показав наявність значної кількості викидів у змінних chlorides, residual sugar та sulphates, що вимагає додаткової уваги при підготовці даних для моделювання. Також виявлено, що експертні оцінки мають певну упередженість з тяжінням до значень 5 та 6, між якими часто немає статистично значущої різниці за хімічними показниками.

На основі кластерного аналізу встановлено, що дані природньо групуються у близько 5 кластерів, що може бути корисним для подальшого аналізу та моделювання. При цьому найкращим варіантом кластеризації може бути метод K-Means через лінійний характер розділення даних.

Для подальшого використання датасету в інформаційних технологіях передбачення якості вина рекомендується:

– Видалити викиди зі змінних chlorides та residual sugar, які спотворюють загальну картину
– Провести стандартизацію змінних для коректного обчислення відстаней у метричному просторі
– Мінімізувати вплив упередженості експертних оцінок через перегрупування або зважування
– Зосередити увагу на змінних з найвищою кореляцією з цільовою змінною: alcohol, sulphates, citric acid та volatile acidity

Також цікавим спостереженням є те, що якість вина позитивно корелює з показниками, які свідчать про завершення процесу визрівання, але при збереженні певних складних сполук, що виникають під час бродіння. Це підтверджує емпіричні знання виноробів про важливість належного терміну дозрівання для досягнення найкращої якості продукту.

Результати проведеного розвідувального аналізу даних створюють надійну основу для подальшої розробки інформаційної технології передбачення якості вина методами машинного навчання, що може стати важливим інструментом для виноробної промисловості та контролю якості продукції. Робочий ноутбук було опубліковано на сайті Kaggle [6].

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Red Wine Quality [Електронний ресурс]. URL: https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009 (звернення: 23.04.2025).
2. Skewness [Електронний ресурс]. URL: https://en.wikipedia.org/wiki/Skewness (звернення: 23.04.2025).
3. Kurtosis [Електронний ресурс]. URL: https://en.wikipedia.org/wiki/Kurtosis (звернення: 23.04.2025).
4. Q-Q plot [Електронний ресурс]. URL: https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot (звернення: 23.04.2025).
5. Clustergram [Електронний ресурс]. URL: https://clustergram.readthedocs.io/en/stable/ (звернення: 23.04.2025).
6. EDA for wine quality [Електронний ресурс]. URL: https://www.kaggle.com/code/demkovivan/eda-for-wine-quality-draft (звернення: 23.04.2025).

Демков Іван Владиславович – студент групи СА-21б факультету інтелектуальних інформаційних технологій та автоматизації Вінницького національного технічного університету, м. Вінниця, email: demkov_ivan@ukr.net

Жуков Сергій Олександрович – к.т.н., доцент кафедри системного аналізу та інформаційних технологій Вінницького національного технічного університету, м. Вінниця, e-mail: sazhukov@gmail.com

Demkov Ivan V. – student of group CA-21b, Faculty of Intelligent Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, Ukraine, email: demkov_ivan@ukr.net

***Zhukov Serhii O.*.** – Ph.D., Assistant Professor of the Department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: sazhukov@gmail.com