УДК: 004.8:519.2:663.2

А. С. Кравчук
С. О. Жуков

РОЗВІДУВАЛЬНИЙ АНАЛІЗ ДАНИХ ДЛЯ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ОЦІНКИ ЯКОСТІ ЧЕРВОНОГО ВИНА МЕТОДАМИ МАШИННОГО НАВЧАННЯ

Вінницький національний технічний університет

Анотація

Дослідження присвячене підготовці та розвідувальному аналізу даних для подальшого використання в інформаційній технології оцінки якості червоного вина методами машинного навчання. Проведено аналіз структури датасету та його основних характеристик, а також досліджено можливі взаємозв'язки між якістю вина та фізико-хімічними показниками.

Ключові слова: інформаційні технології, якість вина, машинне навчання, аналіз даних, прогнозування, фізико-хімічні властивості, розвідувальний аналіз.

Abstract

The research is devoted to the preparation and exploratory data analysis for further use in the information technology of evaluating red wine quality using machine learning methods. The dataset structure and its key features were analyzed, and possible relationships between wine quality and physicochemical properties were investigated.

Keywords: information technology, wine quality, machine learning, data analysis, prediction, physicochemical properties, exploratory analysis.

Вступ

Інформаційні технології відіграють значну роль у харчовій промисловості та виноробстві, особливо у дослідженні факторів, що впливають на якість продукту. Якість червоного вина визначається багатьма фізико-хімічними параметрами, такими як кислотність, рівень цукру, вміст спирту, а також органолептичними характеристиками, що формують його смаковий профіль.

Методи машинного навчання відкривають нові можливості для аналізу цих параметрів і побудови моделей прогнозування якості вина. Одним із ключових етапів створення таких моделей є розвідувальний аналіз даних (EDA), що дозволяє визначити закономірності, оцінити кореляції між змінними та підготувати дані до подальшого моделювання.

Розвідувальний аналіз

Для проведення аналізу було обрано набір даних «Wine Quality Dataset», який містить інформацію про хімічні властивості червоного вина та його оцінку якості [1]. Даний датасет є загальнодоступним і широко використовується для досліджень у сфері аналізу даних та машинного навчання. Він включає 1599 записів і 12 атрибутів, що описують різні фізико-хімічні характеристики вина, які можуть впливати на його якість.

Цільовою змінною у цьому дослідженні є якість вина (Quality), яка оцінюється за шкалою від 0 до 10. Ця оцінка базується на сенсорному аналізі, що відображає загальне сприйняття смаку, аромату та балансу напою. Вона може залежати від таких факторів, як кислотність (Fixed Acidity, Volatile Acidity, Citric Acid), залишковий цукор (Residual Sugar), рівень сульфатів (Sulphates), вміст алкоголю (Alcohol) та інші хімічні параметри.Серед інших атрибутів, що можуть впливати на якість вина, варто виділити pH-рівень (pH), щільність (Density), вміст діоксиду сірки (Free Sulfur Dioxide, Total Sulfur Dioxide) та рівень хлоридів (Chlorides).Даний набір даних дозволяє провести розвідувальний аналіз (EDA), дослідити залежності між хімічним складом вина та його якістю, а також виявити ключові фактори, що визначають сенсорні властивості напою. На рисунку 1 представлено фрагмент датасету «Wine Quality Dataset», який містить основні атрибути.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

Рис. 1. Приклад атрибутів, що містить набір даних

Попередньо було проведено очищення даних та усунення можливих пропущених значень. Побудовано гістограму розподілу якості вина (рис. 1).

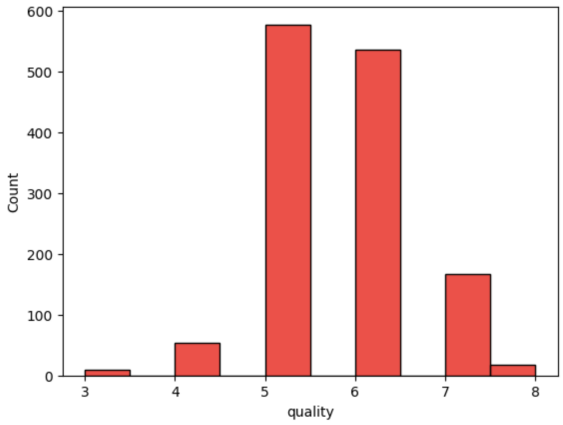


Рис. 2 Гістограма розподілу якості вина, що містить набір даних

З гістограми видно, що розподіл якості вина є нерівномірним, оскільки найбільше зразків мають оцінки 5 і 6, тоді як вина з нижчими (3-4) та вищими (7-8) оцінками зустрічаються значно рідше. Це може вплинути на точність моделей прогнозування, оскільки вони можуть бути схильні до переважного прогнозування середніх значень.

Після підготовки даних було побудовано матрицю кореляції (рис. 2), яка дозволяє виявити явні залежності між фізико-хімічними показниками вина та його якістю. Це допоможе визначити ключові фактори, що найбільше впливають на оцінку вина, та використовувати їх для подальшого аналізу і прогнозування.

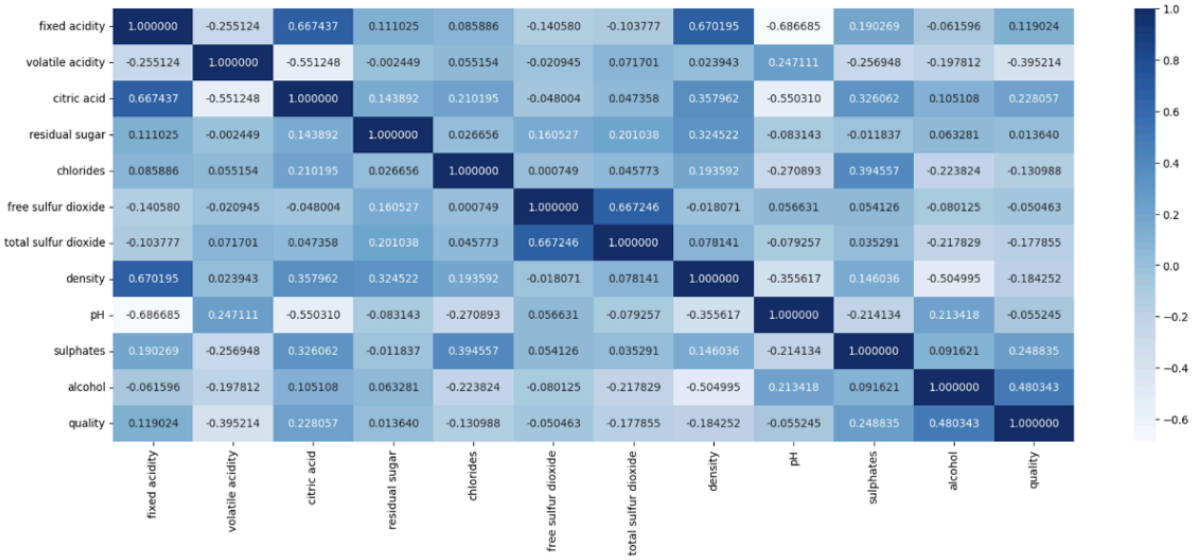


Рис. 3. Матриця кореляції

Леткі кислоти (Volatile Acidity) мають виражену негативну кореляцію з якістю, що свідчить про їхній несприятливий вплив на смакові характеристики вина. Водночас такі показники, як рівень pH та щільність (Density), також виявляють певну взаємозалежність із якістю, що може бути корисним для подальшого аналізу впливу хімічного складу на органолептичні властивості вина.

Побудовано графіки розподілу густини (Density) та леткої кислотності (Volatile acidity), для більш детального аналізу, що дозволяє оцінити природну варіативність хімічних параметрів та їхній можливий вплив на якість кінцевого продукту (рис. 4, 5).

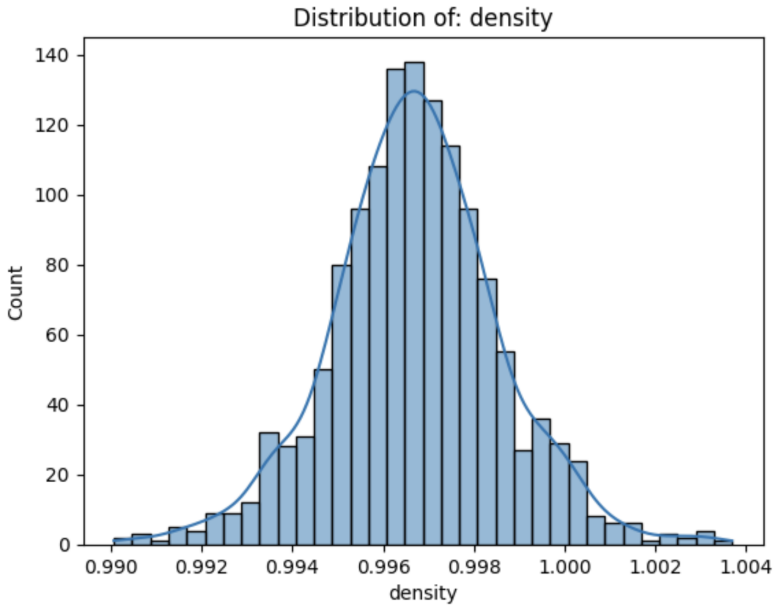


Рис. 4. Графік розподілу густини (Density)

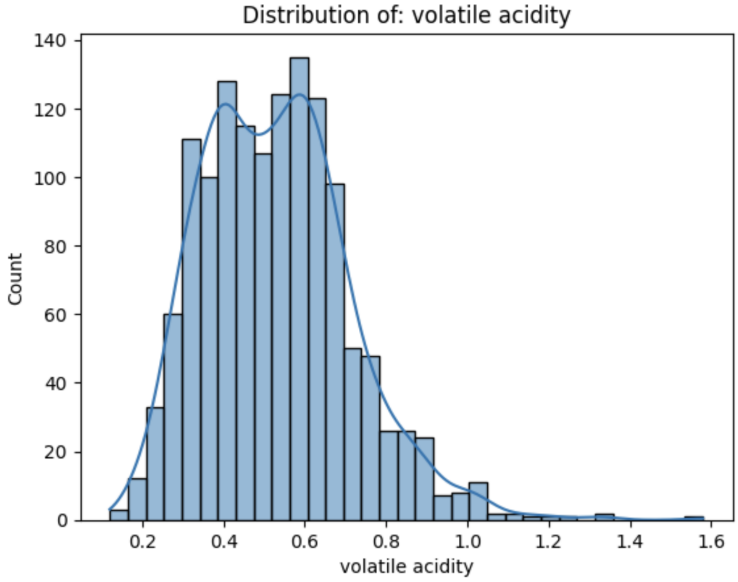


Рис. 5. Графік розподілу леткої кислотності (Volatile acidity)

З цих графіків видно, що фіксована кислотність і густина мають різні характерні особливості у своєму розподілі. Фіксована кислотність демонструє асиметричний правосторонній розподіл із найбільшою концентрацією значень у діапазоні 6–9, що свідчить про природну варіативність цього показника, тоді як густина має нормальний розподіл із середнім значенням близько 0.996, що вказує на її стабільність у вибірці. Такий аналіз дозволяє припустити, що кислотність може мати ширший вплив на якість продукту через свою варіативність, тоді як густина є більш передбачуваним параметром. Це може бути корисним

при подальшому дослідженні хімічного складу та його взаємозв'язку з якісними характеристиками кінцевого продукту.

Отримані результати є важливими для подальшого моделювання за допомогою методів машинного навчання, оскільки вони підтверджують вплив фізико-хімічних параметрів на якість вина та можуть бути використані для створення прогнозних моделей.

Висновки

У ході розвідувального аналізу даних було досліджено взаємозв'язки між хімічним складом та якістю червоного вина на основі відповідного набору даних. Аналіз показав, що вміст алкоголю та сульфатів має позитивний вплив на якість вина, тоді як підвищена концентрація летких кислот негативно корелює з оцінкою якості. Виявлено, що рівень pH та кислотності мають взаємозв'язок із якістю вина, що може свідчити про їхній вплив на його смакові властивості.

Попередня обробка даних включала очищення, усунення пропусків та нормалізацію значень, що дозволило підвищити точність подальшого аналізу. Аналіз розподілу цільової змінної показав, що більшість зразків вина отримали оцінку якості в діапазоні 5–6 балів, що вказує на переважно середній рівень якості продукції.

Кореляційний аналіз виявив статистично значущі взаємозв'язки між основними хімічними показниками та якістю вина. Зокрема, леткі кислоти мають виражений негативний вплив, що вказує на необхідність контролю їхнього рівня для підвищення якості кінцевого продукту.

Отримані результати є важливими для подальшого застосування методів машинного навчання з метою прогнозування якості червоного вина та оптимізації технологічних процесів його виробництва.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Red Wine Quality Dataset. Kaggle. 2024 [Електронний ресурс]. – Режим доступу: https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009/code
2. Pandas Tutorial 2024 [Електронний ресурс] – Режим доступу: https://www.w3schools.com/python/pandas/default.asp
3. Matplotlib Pyplot Documentation. 2024 [Електронний ресурс]. – Режим доступу: https://matplotlib.org/3.5.3/api/_as_gen/matplotlib.pyplot.html
4. A Complete Guide to Data Visualization in Python With Libraries & More. 2023 [Електронний ресурс]. – Режим доступу: https://www.simplilearn.com/tutorials/python-tutorial/data-visualization-in-python

Кравчук Анна Станіславівна – студентка групи СА-23б, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, м. Вінниця. e-mail: annakravchuk880@gmail.com

Жуков Сергій Олександрович– к.т.н., доцент кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, e-mail: sazhukov@gmail.com

Kravchuk Anna – student of Faculty of Intellectual Information Technologies and Automation, SA-23b, Vinnytsia National Technical University, Vinnytsia, e-mail annakravchuk880@gmail.com

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