3. Song J., Meng C., Ermon S. Denoising diffusion implicit models. International Conference on Learning Representations (ICLR 2021). URL: https://arxiv.org/abs/2010.02502 (дата звернення: 12.10.2026).

4. Rombach R., Blattmann A., Lorenz D., Esser P., Ommer B. High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022. P. 10684–10695. URL: https://arxiv.org/abs/2112.10752 (дата звернення: 12.10.2026).

5. Lin S., Liu B., Li J., Yang X. Common diffusion noise schedules and sample steps are flawed. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2024). URL: https://arxiv.org/abs/2305.08891 (дата звернення: 12.10.2026).

УДК 004.85

ПОРІВНЯЛЬНИЙ АНАЛІЗ ПРОСТИХ МЕТОДІВ КЛАСИФІКАЦІЇ ЗА ФІЗИКО-ХІМІЧНИМИ ПОКАЗНИКАМИ ЗРАЗКІВ ВИНА

Ренат Бондар

Миколаївський національний аграрний університет, м. Миколаїв, Україна

Науковий керівник: О. Є. Богатєнкова,
асистент Миколаївський національний аграрний університет

Оцінювання та класифікація виноробної продукції за лабораторними показниками є типовим завданням підтримки рішень у виноробстві: фізико-хімічні тести доступні на етапі сертифікації, а моделі інтелектуального аналізу даних можуть доповнювати роботу експертів [1]. Для невеликих вибірок практично важливо обрати простий, інтерпретований класифікатор і з’ясувати, наскільки його результат залежить від попередньої обробки ознак [2].

Мета дослідження – порівняти за точністю та макро-F1 чотири прості методи класифікації (метод k найближчих сусідів, наївний байєсівський класифікатор, логістичну регресію, дерево рішень) на відкритих даних хімічного аналізу вина та оцінити вплив стандартизації ознак на метод k найближчих сусідів.

Використано відкритий набір Wine з репозиторію UCI [3]: 178 зразків вина з однієї області Італії, 13 неперервних показників хімічного аналізу (вміст спирту, яблучної кислоти, золи, фенолів, проліну тощо) та три класи за сортом винограду (59, 71 і 48 зразків). Слід зазначити, що цільовою змінною є сорт, а не органолептична оцінка якості, яку моделюють у [1] на значно більших вибірках вина Vinho Verde; тому отримані результати стосуються порівняння методів на лабораторних показниках вина, а не прогнозування експертної оцінки якості. Моделі реалізовано в бібліотеці scikit-learn [4]. Для методу k найближчих сусідів (k = 5) ознаки стандартизовано за формулою (1), відстань між зразками обчислювали за формулою (2):

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \qquad (1)$$

$$d(x, x') = \sqrt{\sum_{j=1}^{13} (x_j - x'_j)^2} \qquad (2)$$

де x_{ij} – значення j-ї ознаки i-го зразка; μ_j і σ_j – середнє та стандартне відхилення j-ї ознаки на навчальній вибірці; $d(x, x')$ – евклідова відстань між зразками. Також застосовано гаусів наївний Байєс, логістичну регресію з L2-регуляризацією (C = 1) та дерево рішень глибиною 3; гіперпараметри не підбирали. Якість оцінювали п’ятикратною стратифікованою перехресною перевіркою, повтореною 10 разів (50 розбиттів, початкове значення генератора 42); стандартизацію виконували лише на навчальних блоках, щоб уникнути витоку інформації. Показники – частка правильних відповідей (accuracy) і макро-F1.

Усереднені за 50 розбиттями показники наведено в табл. 1 та на рис. 1. Найвищу точність отримано для логістичної регресії (0,983), далі йдуть наївний Байєс (0,972) і метод k найближчих сусідів зі стандартизацією (0,964); дерево рішень глибиною 3 дало 0,905. Різниця між першими трьома методами (до 1,9 в. п.) не перевищує розкиду оцінок між розбиттями (стандартне відхилення 0,016-0,031), тому на цих даних не можна стверджувати переваги одного з них.

Таблиця 1 - Результати перехресної перевірки (середнє ± стандартне відхилення, 50 розбиттів)

Метод	Accuracy	Macro-F1
kNN (k=5), стандартизація	0,964 ± 0,031	0,964 ± 0,031
kNN (k=5), без стандартизації	0,711 ± 0,071	0,690 ± 0,074
Наївний Байєс (Гауса)	0,972 ± 0,032	0,973 ± 0,031
Логістична регресія	0,983 ± 0,016	0,983 ± 0,016
Дерево рішень (глибина 3)	0,905 ± 0,051	0,906 ± 0,051

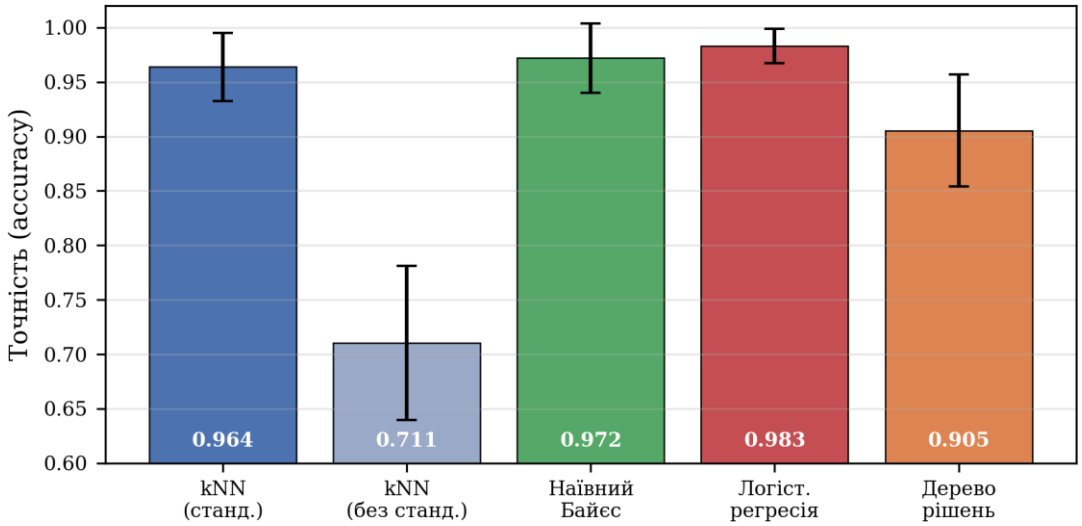


Рис. 1. Точність (accuracy) методів класифікації; позначено середнє та стандартне відхилення

Найбільш показовим є вплив масштабування: без стандартизації точність методу k найближчих сусідів становить 0,711, що на 25,3 в. п. нижче, ніж зі стандартизацією (макро-F1 – 0,690 проти 0,964). Причина полягає в тому, що в наборі ознаки мають різні порядки величин (зокрема, пролін вимірюється сотнями одиниць, тоді як інші показники – одиницями), і відстань (2) визначається переважно однією ознакою. Дерево рішень глибиною 3 поступається лінійним та ймовірнісним моделям, імовірно через обмежену глибину; цю гіпотезу в роботі не перевіряли.

Щоб перевірити, чи не залежить висновок про метод k найближчих сусідів від вибору k = 5, оцінку повторено за того самого протоколу для k від 1 до 21 (табл. 2). Точність зростає від 0,954 (k = 1) до 0,970 (k = 21), тобто в межах 1,6 в. п., що менше за стандартне відхилення оцінок за окремими розбиттями. Отже, на цих даних метод стійкий до вибору k, а головним чинником якості є стандартизація ознак.

Таблиця 2 – Залежність точності kNN (зі стандартизацією) від числа сусідів k

k	1	3	5	7	11	15	21
Accuracy	0,954	0,955	0,964	0,967	0,964	0,969	0,970
Станд. відхил.	0,035	0,036	0,031	0,028	0,027	0,029	0,025