УДК 004.9

Мокін В.Б.
Сокур Д.С.

Інформаційна технологія класифікації типу вин за їх хімічним складом

Вінницький національний технічний університет

Анотація

У статті досліджено процес розробки інформаційної технології для класифікації типу вин на основі їх хімічного складу. Робота зосереджена на використанні сучасних алгоритмів машинного навчання для аналізу ключових параметрів, таких як кислотність, вміст алкоголю, рівень цукру та інші. Розглянуто етапи збору, очищення даних, побудови моделей, а також оцінки їх ефективності. Результати демонструють високий потенціал застосування таких технологій у виноробній галузі для автоматизації контролю якості продукції.

Ключові слова:

Класифікація вин, машинне навчання, хімічний склад, автоматизація, виноробство.

Abstract

The article investigates the process of developing information technology for classifying wine types based on their chemical composition. The work focuses on the use of modern machine learning algorithms to analyze key parameters such as acidity, alcohol content, sugar level, and others. The stages of data collection, data cleaning, model building, and evaluation of their effectiveness are considered. The results demonstrate the high potential of using such technologies in the wine industry to automate product quality control.

Keywords:

Wine classification, machine learning, chemical composition, automation, winemaking.

Вступ

Виноробство є однією з найстаріших і водночас інноваційних галузей харчової промисловості, яка завжди потребує високих стандартів якості продукції. Вино, як складний продукт, має різноманітний хімічний склад, що визначає його смакові, ароматичні та фізичні характеристики. Саме точна класифікація вина дозволяє виноробам забезпечити стабільність якості, відповідність споживчим очікуванням і оптимізувати виробничі процеси.

Сьогодні класифікація типів вина (червоне чи біле) за їх хімічним складом є важливою задачею, яка зазвичай виконується вручну або із залученням хімічних експертів. Проте такі методи є тривалими, потребують високої кваліфікації персоналу і піддаються впливу суб'єктивних факторів. Ці виклики створюють потребу у впровадженні автоматизованих підходів, які можуть значно підвищити точність і швидкість класифікації.

Сучасні інформаційні технології, зокрема методи машинного навчання, відкривають нові можливості для автоматизації цього процесу. Використовуючи алгоритми, які аналізують великий обсяг даних і враховують складні взаємозв’язки між показниками, можна створити систему, здатну ефективно класифікувати вина на основі їх хімічного складу.

Метою роботи є розробка інформаційної технології класифікації типу вин за їх хімічним складом із використанням сучасних алгоритмів машинного навчання та аналіз даних із відкритого набору Kaggle "Wine Quality Data Set". Ця технологія дозволяє підвищити якість класифікації, скоротити час обробки даних і мінімізувати людський фактор у процесі прийняття рішень.

Актуальність дослідження та постановка проблеми

Вино є невід'ємною складовою багатьох культур і традицій у всьому світі. Його якість, смакові властивості та відповідність споживчим очікуванням залежать від точного контролю хімічного складу. Тип вина (червоне чи біле) визначається такими характеристиками, як кислотність, рівень алкоголю, залишковий цукор та інші хімічні показники, які впливають на аромат, текстуру і тривалість зберігання.

Хімічний склад вина – це складний набір даних, де навіть невеликі зміни у співвідношенні компонентів можуть суттєво вплинути на кінцевий результат. Ручні методи аналізу часто не враховують взаємозв’язки між параметрами, такі як кореляція між кислотністю, залишковим цукром і рівнем алкоголю, що може призводити до похибок.

Використання сучасних інформаційних технологій, зокрема методів машинного навчання, дозволяє ефективно вирішувати ці проблеми. Завдяки автоматизованому аналізу великих обсягів даних з урахуванням складних залежностей між параметрами, ці технології значно підвищують точність класифікації.

Розробка інформаційної технології класифікації вин на основі хімічного складу є не лише актуальною, а й критично важливою для:

- Забезпечення відповідності сучасним стандартам якості.
- Автоматизації виробничих процесів у виноробстві.
- Підвищення точності прогнозування та мінімізації втрат через дефекти продукції.

Ця робота є важливим внеском у розвиток автоматизації виноробної промисловості, що дозволяє ефективно інтегрувати інноваційні технології у процеси контролю якості.

Характеристика проблемної області

Хімічний склад вина є багатофакторним, і навіть невеликі зміни у співвідношенні компонентів можуть суттєво впливати на його тип і якість. Основні параметри, що визначають тип вина, включають:

- Алкоголь (Alcohol): Визначає міцність і впливає на текстуру та смак.
- Летюча кислотність (Volatile Acidity): Відповідає за ароматичний профіль вина. Підвищені значення можуть свідчити про недоліки у виробництві.
- Залишковий цукор (Residual Sugar): Показник солодкості вина, критично важливий для визначення його типу.
- pH: Визначає кислотно-лужний баланс, що впливає на смак і стабільність продукту.
- Діоксид сірки (Free та Total Sulfur Dioxide): Забезпечує збереження вина, запобігаючи його окисленню.

У процесі класифікації виникають такі виклики:

- Складність багатофакторного аналізу: Множинні параметри взаємодіють один з одним, і традиційні методи часто не можуть врахувати ці зв’язки.
- Потреба в ефективних рішеннях: Необхідність швидкого аналізу даних при великих обсягах виробництва.
- Залежність від суб’єктивних оцінок: Людський фактор часто є причиною похибок у класифікації.

Вирішення цих проблем можливе завдяки застосуванню алгоритмів машинного навчання. Вони дозволяють автоматизувати класифікацію, забезпечують точність аналізу та швидкість обробки даних. Розробка інформаційної технології для автоматичної класифікації типів вин стане важливим кроком у цифровізації виноробної галузі та підвищенні її конкурентоспроможності.

Методи і засоби вирішення поставлених задач

Для реалізації інформаційної технології класифікації вин за їх хімічним складом було розроблено багатоступеневий процес, що включає підготовку даних, навчання моделей машинного навчання та оцінку їхньої ефективності. Основні етапи розробки наведено нижче.

1. Підготовка даних

1.1. Джерело даних

Набір даних, використаний для розробки, узятий із відкритого джерела Kaggle – «Wine Quality Data Set», який містить 6497 зразків червоного та білого вина. Кожен зразок представлений 11 хімічними показниками (наприклад, рівень кислотності, залишкового цукру, алкоголю) і типом вина (червоне або біле).

1.2. Обробка даних

Видалення дублікатів: У датасеті було знайдено 15,5% дубльованих записів, які було вилучено для уникнення зміщення результатів.

Масштабування ознак: Всі числові параметри нормалізовано за допомогою методу StandardScaler, щоб привести їх до одного масштабу та покращити роботу моделей.

Формування наборів: Датасет було розбито на три частини: тренувальний (70%), валідаційний (20%) і тестовий (10%).

2. Оцінка моделей

Оцінювання точності здійснювалося за метриками accuracy, precision, recall і F1-score. Найкращі результати продемонстрував алгоритм XGBoost із точністю 97.5% на тестових даних.

Модель	Точність на тренувальних даних	Точність на тестових даних
Decision Tree	98.4%	96.3%
Random Forest	98.4%	96.8%
XGBoost	99.1%	97.5%

Табл. 1 - Результати порівняння моделей

Технічні аспекти реалізації інформаційної технології

Розробка інформаційної технології класифікації вин за хімічним складом включала кілька ключових етапів, кожен із яких було реалізовано за допомогою сучасних інструментів і технологій.

1. Середовище розробки та інструменти

Для реалізації проєкту використовувалися такі програмні засоби:

- Мова програмування: Python – завдяки його широкому набору бібліотек для обробки даних і машинного навчання.
- Бібліотеки для аналізу даних: Pandas, NumPy – для роботи з даними та їх попередньої обробки.
- Модулі для візуалізації: Matplotlib, Seaborn – для створення графіків і теплових карт кореляцій.
- Бібліотеки машинного навчання: Scikit-learn, XGBoost – для побудови та налаштування моделей класифікації.
- Інструменти крос-валідації: GridSearchCV – для підбору оптимальних гіперпараметрів.
- IDE: Jupyter Notebook – для інтерактивного кодування та візуалізації результатів.

Системні вимоги:

- Операційна система: Windows 10 або Linux.
- Процесор: Intel Core i5 або вище.
- Оперативна пам'ять: 8 GB RAM.
- Пакети Python: sklearn, xgboost, pandas, numpy, matplotlib.

2. Архітектура рішення

Система складається з трьох основних компонентів:

- Модуль підготовки даних:

Включає завантаження набору даних, попередню обробку (очищення, нормалізація, видалення дублікатів) і розділення даних на тренувальний, валідаційний та тестовий набори.

- Модуль машинного навчання:

Реалізовано за допомогою трьох алгоритмів: Decision Tree, Random Forest, XGBoost. Основна задача – побудова моделей класифікації, їх навчання та оцінка продуктивності.

- Модуль візуалізації:

Забезпечує інтерпретацію результатів через графіки важливості ознак, теплові карти кореляцій і розподіли результатів класифікації.

3. Реалізація етапів обробки даних

3.1. Завантаження та очищення даних

Використаний датасет Kaggle містив понад 6000 зразків вин із характеристиками. Для підготовки даних виконувалися наступні дії:

- Видалення дублікатів і пропусків.
- Нормалізація значень за допомогою StandardScaler для приведення всіх параметрів до одного масштабу.

3.2. Розподіл даних

Дані розділялися на три вибірки:

- Тренувальна (70%) – для навчання моделей.
- Валідаційна (20%) – для налаштування гіперпараметрів.
- Тестова (10%) – для оцінки продуктивності моделей.

4. Побудова моделей машинного навчання

Для кожного алгоритму було виконано навчання на тренувальному наборі з використанням підібраних гіперпараметрів. Найкращі результати продемонструвала модель XGBoost.

Параметри моделі XGBoost:

- n_estimators: 200
- max_depth: 5
- learning_rate: 0.1

5. Оцінка та візуалізація результатів

Для оцінки моделей використовувалися такі метрики:

- Accuracy: Загальна точність класифікації.
- Precision та Recall: Для оцінки точності передбачення класів.
- F1-score: Комплексна метрика для балансування точності та повноти.

Також було створено графіки важливості ознак для інтерпретації роботи моделей. Основними параметрами, які впливають на класифікацію, виявилися:

- Вміст алкоголю.
- Летюча кислотність.
- Рівень залишкового цукру.

Переваги розробленої інформаційної технології

Розроблена інформаційна технологія класифікації типу вин на основі їх хімічного складу має низку значущих переваг, які роблять її перспективним інструментом для впровадження у виноробну галузь. Використання сучасних методів машинного навчання не лише підвищує точність прогнозування, але й дозволяє автоматизувати

процеси, які раніше вимагали значних людських ресурсів. Завдяки гнучкості та адаптивності система може бути інтегрована у різні етапи виробництва, знижуючи витрати часу і матеріалів.

Основні переваги включають:

- Висока точність класифікації

Використання сучасних алгоритмів машинного навчання, таких як XGBoost, дозволило досягти точності 97.5% на тестових даних. Це забезпечує мінімізацію помилок класифікації, що особливо важливо для стабільної якості продукції. Такий рівень точності дозволяє виробникам вин забезпечувати відповідність продукції міжнародним стандартам.

- Автоматизація процесів

Технологія дозволяє автоматизувати процес класифікації типу вин, усуваючи необхідність ручного аналізу. Це знижує витрати часу, зменшує вплив людського фактора та підвищує ефективність. Автоматизована система працює швидко й стабільно, що важливо в умовах великих виробничих обсягів.

- Гнучкість та масштабованість

Система може бути адаптована для класифікації інших типів напоїв або харчових продуктів із використанням відповідних хімічних параметрів. Крім того, її легко інтегрувати у виробничі процеси, використовуючи сенсори IoT та інші автоматизовані інструменти. Це робить розробку універсальним рішенням, здатним відповідати потребам не лише виноробства, а й суміжних галузей.

- Інтерпретованість результатів

Завдяки візуалізації важливості ознак та графікам залежностей між параметрами, система забезпечує зрозумілу інтерпретацію роботи моделі. Це допомагає виноробам ухвалювати обґрунтовані рішення, зосереджуючи увагу на параметрах, що найбільше впливають на якість продукції.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Bishop C. M. Pattern Recognition and Machine Learning. Springer, 2006.
2. Géron A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, 2019.
3. Breiman L. Random Forests. Machine Learning, 2001, 45(1), 5–32.
4. Friedman J. H. Greedy function approximation: A gradient boosting machine. Annals of Statistics, 2001, 29(5), 1189–1232.
5. Wine Quality Data Set. URL: https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009 (дата звернення: 17.11.2024).
6. Pedregosa F., Varoquaux G., Gramfort A., et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 2011, 12, 2825–2830.
7. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, 785–794.
8. Python Software Foundation. Python 3 Documentation. URL: https://docs.python.org/3/ (дата звернення: 17.11.2024).
9. Scikit-learn Documentation. An Overview of the Library's Capabilities. URL: https://scikit-learn.org/ (дата звернення: 17.11.2024).
10. XGBoost Documentation. Introduction to XGBoost. URL: https://xgboost.readthedocs.io/ (дата звернення: 17.11.2024).

Науковий керівник – Мокін Віталій Борисович– д.т.н., завідувач кафедри системного аналізу та інформаційних технологій, професор, Факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: vbmokin@vntu.edu.ua

Сокур Дмитро Сергійович – студент групи 1ІСТ-23м, Факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: dmytrosokur11@gmail.com

Supervisor –***Mokin Vitalii*** – Doctor of Technical Sciences, Professor of the Department of System Analysis and Information Technologies, Faculty of Intellectual Informational Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail: vbmokin@vntu.edu.ua

Sokur Dmytro S. – student of 1IST-23m, Faculty of Intellectual Informational Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail: dmytrosokur11@gmail.com