DOI 10.34132/mspc2025.01.14.07

Олександр Гожий
Ірина Калініна

МЕТОДОЛОГІЯ АНАЛІЗУ ТА ПОПЕРЕДНЬОЇ ОБРОБКИ ДАНИХ ДЛЯ ВИРІШЕННЯ ЗАДАЧ МАШИННОГО НАВЧАННЯ

В статті описується та досліджується методологія аналізу та попередньої обробки даних для вирішення задач машинного навчання. Методологія об'єднує на основі системного підходу наступні групи методів: методи обробки пропусків в даних, методи обробки аномальних значень, методи генерації ознак, методи ідентифікації нелінійностей та нестаціонарностей та методи нормалізації. Досліджено метод генерації ознак. Методи відбору і генерації ознак поділяють на три основні групи: методи фільтрації, методи обгортки і вбудовані методи. Метод генерації ознак складається з п'яти кроків для ефективного вибору найбільш релевантних ознак набору даних. Він пропонує кращий підхід до відбору ознак. Це призводить до покращення продуктивності моделі з меншою кількістю ознак і скорочення обчислювальних витрат. Для експериментальної перевірки методології аналізу та попередньої обробки даних була розглянуто створення системи класифікації якості червоних вин. Для вирішення завдання класифікації якості вина було проведено моделювання з використанням різних алгоритмів. Було доведено ефективність методу для вирішення задач аналізу та попередньої обробки даних для вирішення задач класифікації.

***Ключові слова**, методологія аналізу та попередньої обробки даних, методи генерації ознак, методи обробки пропусків в даних, методи обробки аномальних значень, методи ідентифікації нелінійностей та нестаціонарностей та методи нормалізації.*

Аналіз та попередня обробка даних, це початковий етап вирішення завдань машинного навчання. Від якісно проведеного аналізу та попередньої обробки даних залежить кінцевий результат вирішення завдання. На цьому етапі досліджуються реальні данні різної природи, їх властивості та поведінка, які впливають на процес вирішення завдання в-цілому. Ігнорування цього етапу призводить до неможливості побудови коректних ймовірнісне-статистичних моделей для конкретного процесу та їх придатності для розв'язання задачі машинного навчання.

На етапі аналізу та попередньої обробки даних вирішуються наступні задачі: визначення набору даних для аналізу та попередньої обробки даних; ідентифікація та заповнення пропусків в даних; ідентифікація та підбір методів обробки аномальних значень в даних; визначення зовнішніх впливів і збурень та їх типу, та встановлення можливості їх ймовірнісне-статистичного опису за допомогою конкретних типів розподілів випадкових величин; визначення методів та алгоритмів фільтрації з метою виділення корисної складової при обробці даних; генерація ознак в наборі даних; аналіз та встановлення основних типів нелінійностей та нестаціонарностей даних; розробка підходу до застосування методів нормалізації та стандартизації даних.

Для ефективного вирішення задач аналізу та попередньої обробки даних в задачах машинного навчання, необхідно визначити головні властивості тих наборів даних, що досліджуються:

1. *Складність* – кількість і різноманітність факторів, які складають та будуть впливати на набір даних.
2. *Множина зв'язків* між факторами, які характеризують данні, тобто сила, з якою зміна одного параметра (фактора) впливає на зміну інших параметрів набору даних
3. *Динамічність* – швидкість, з якою відбуваються зміни у наборі даних.
4. *Невизначеність* Поява невизначеності зумовлена недостатністю або спотворенням інформації на будь-якому етапі згаданого процесу. Це може бути коротка вибірка даних, якої недостатньо для побудови адекватної моделі; спотворення вимірів шумовими складовими; некоректність встановлення типу розподілу даних і, як наслідок, невірний вибір методу оцінювання параметрів моделі; помилки дослідника при обчисленні оцінок прогнозів або альтернативних рішень на їх основі.

Виділення та обробка такого роду інформаційних характеристик при аналізі інформації для опису набору даних свідчить про те, що необхідно застосовувати *системний підхід* і розглядати набір даних як систему або сукупність систем, які впливають на кінцевий результат аналізу. Саме в рамках цього підходу прийнято представляти будь-які об'єкти у вигляді структурованої системи, виділяти елементи системи, взаємозв'язок між ними і динаміку розвитку елементів і всієї системи в цілому. Тому аналіз і попередня обробка даних, яка використовується для вирішення різних завдань машинного навчання, а також для накопичення необхідної інформації для подальшого використання на різних етапах її обробки, розглядається

як необхідний компонент методології аналізу та попередньої обробки даних в завданнях машинного навчання.

Метою даної роботи є дослідження системного підходу до аналізу та попередньої обробки даних при розв'язанні задач машинного навчання. Для цього необхідно розробити та дослідити методологію аналізу та попередньої обробки даних.

Структура методології аналізу та попередньої обробки даних

Методологія аналізу та попередньої обробки даних побудована на основі системного підходу і об'єднує групи методів та методологічних підходів, які згруповані по функціональному призначенню в процесі аналізу та підготовки даних до моделювання. Структурна модель методології аналізу та попередньої обробки даних представлена на рисунку 1. Методологія передбачає системне використання наступних груп методів: методів обробки пропусків в даних, методів обробки аномальних значень, методів генерації ознак, методів ідентифікації та подолання нелінійностей та нестаціонарностей, методів нормалізації даних.

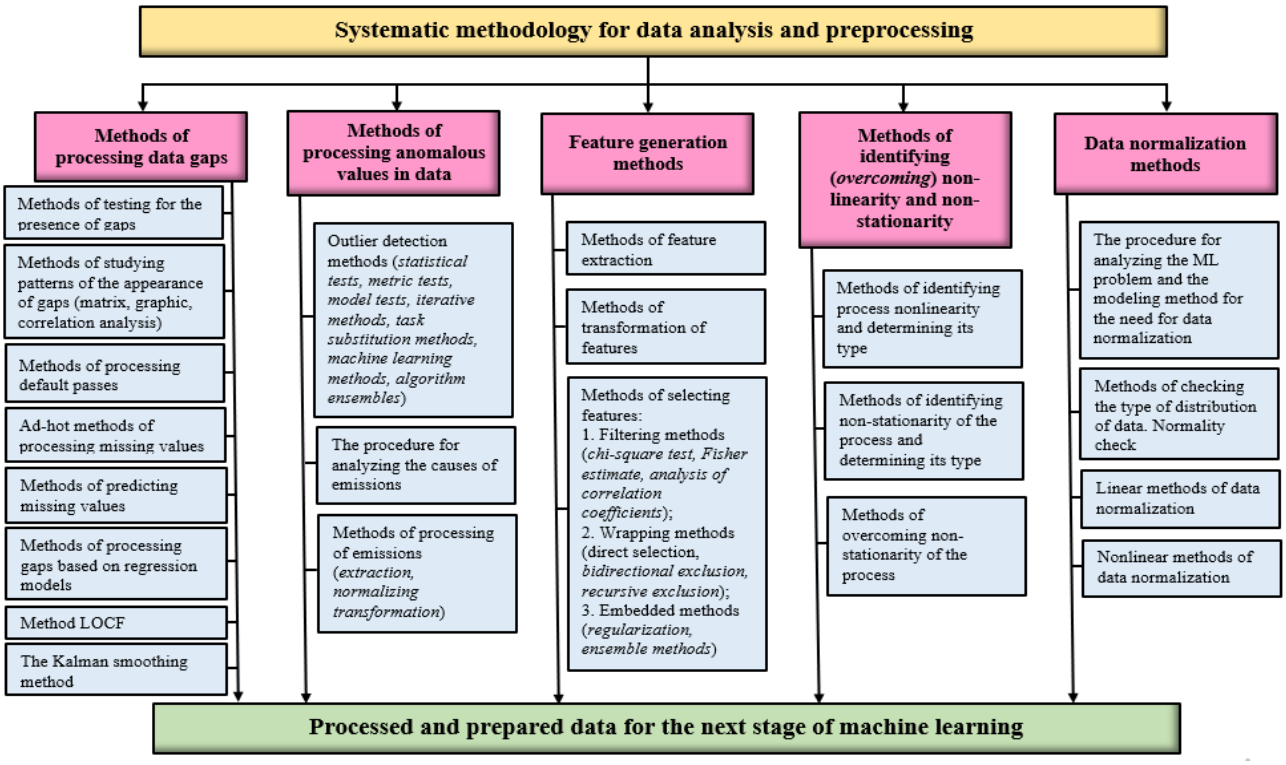


Рис. 1. Методологія аналізу та попередньої обробки даних
Джерело: сформовано автором

Методи обробки пропусків в даних. Методи обробки пропусків в даних об'єднують метод и використання тестів для перевірки на наявність пропусків, різновиди методів дослідження закономірностей появи пропусків, методи обробки пропусків за замовченням, ad-hot методи, методи прогнозування відсутніх значень, методи обробки пропусків на основі регресійних моделей, метод LOCF, методи згладжування на основі фільтра Калмана [1].

Методи обробки аномальних значень. Методи обробки аномальних значень об'єднують методи виявлення аномальних значень в наборі даних, процедуру аналізу причин їх появи та методи обробки викидів [2].

Методи генерації ознак. Методи генерації ознак об'єднують методи вилучення постійних, квазіпостійних та дублюючих ознак, методи перетворення ознак та три групи методів по відбору ознак.

Методи ідентифікації (подолання) нелінійностей та нестаціонарностей. Методи ідентифікації (подолання) нелінійностей та нестаціонарностей об'єднують методи ідентифікації нелінійностей процесу та визначення його типу, методи ідентифікації нестаціонарностей процесу та визначення його типу, методи подолання нестаціонарностей процесу [3].

Методи нормалізації даних. Методи нормалізації даних об'єднують процедуру аналізу задачі машинного навчання та метода моделювання на необхідність нормалізації даних, методи перевірки типу розподілу даних (перевірка на нормальність), лінійні та нелінійні методи нормування даних [4].

Результатом використання методів і методологічних підходів методології аналізу та попередньої обробки даних є оброблений та підготовлений до наступного етапу машинного навчання набір даних.

Методи генерації ознак

Відбір ознак – це процес вибору найбільш релевантних ознак моделі машинного навчання, адаптований до конкретної проблеми, яку намагаються вирішити. Це передбачає вибіркове включення чи виключення важливих ознак із збереженням їх незмінними. Таким чином, він ефективно усуває нерелевантний шум з

даних та зменшує розмір та обсяг вхідного набору даних. Мета відбору ознак у машинному навчанні полягає у наступному: зменшити розмірність простору ознак; прискорити алгоритм навчання; покращити точність прогнозування алгоритму класифікації; покращити зрозумілість результатів навчання.

Комбінований метод відбору ознак. Оскільки відбір ознак грає вирішальну роль машинному навчанні, підвищуючи продуктивність моделі і знижуючи обчислювальні витрати, то роботі запропоновано комбінований метод відбору ознак, який включає поетапне використання і методів фільтрації і методів обгортки. На рисунку 2 представлено блок-схему запропонованого комбінованого методу.

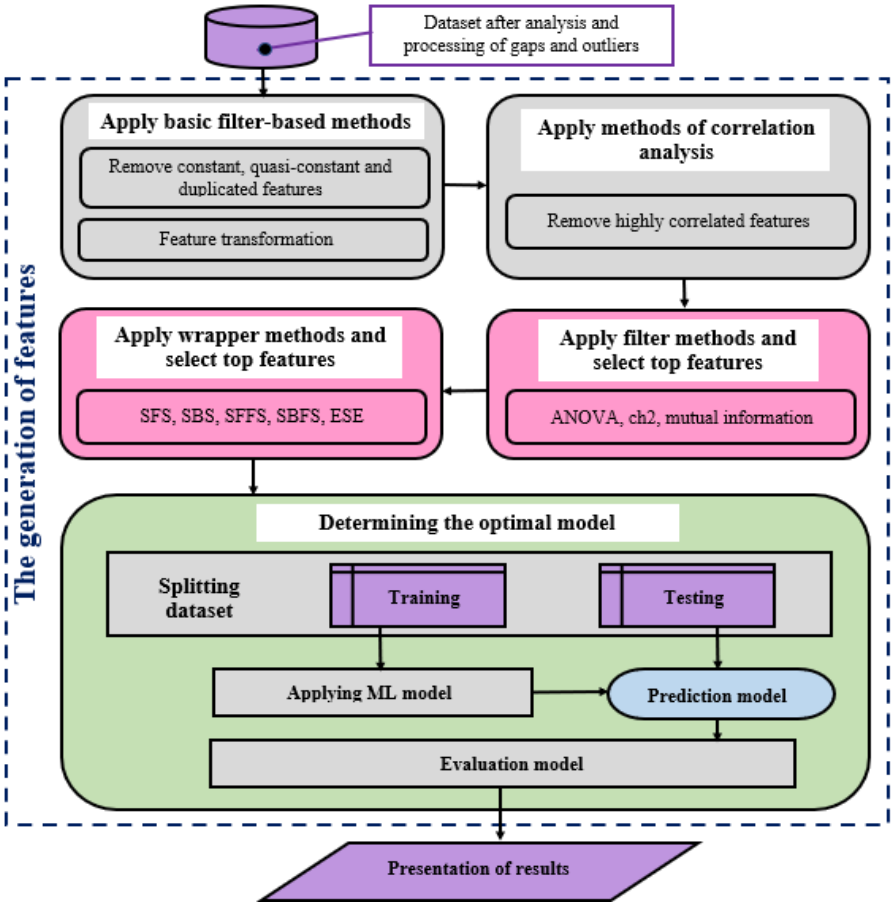


Рис. 2. Блок-схема комбінованого методу відбору ознак
Джерело: сформовано автором

Метод складається з кількох кроків для ефективного вибору найбільш релевантних ознак набору даних.

Крок 1. Застосовування базових методів вибору ознак на основі фільтрів для виявлення та видалення постійних, квазіпостійних та дублюючих ознак.

Крок 2. Використання кореляційного аналізу для виключення ознак з високою кореляцією вище за певний поріг.

Крок 3. Використання методів фільтрації для відбору ознак, такі як ANOVA, хі-квадрат та взаємна інформація для вибору найбільш важливих ознак.

Крок 4. Використання методів обгортки, включаючи послідовний прямий вибір ознак (SFS), послідовний зворотний вибір ознак (SBS), послідовний плаваючий прямий вибір ознак (SFFS), послідовний плаваючий вибір ознак (SBFS) і вичерпний вибір ознак (EFS) для подальшого уточнення підмножини ознак.

Крок 5. Для відбору оптимальної моделі з кількох альтернативних використовуються методи *перевірочної вибірки* та *перехресної перевірки*. Для цього на кроці визначення оптимальної моделі набор даних поділяється на два, навчальний та тестовий набори з використанням вибраних ознак. Дали використовується алгоритми машинного навчання для навчання моделі з меншою кількістю ознак, виконується налаштування моделі та оцінка результатів моделювання.

Комбінований метод пропонує кращий підхід до відбору ознак. Це призводить до покращення продуктивності моделі з меншою кількістю ознак і скорочення обчислювальних витрат. Завдяки систематичному відбору найбільш інформативних ознак та використанню методів фільтрації та обгортки комбінований метод забезпечує більш ефективне та дієве навчання та розгортання моделі машинного навчання.

Експериментальна частина

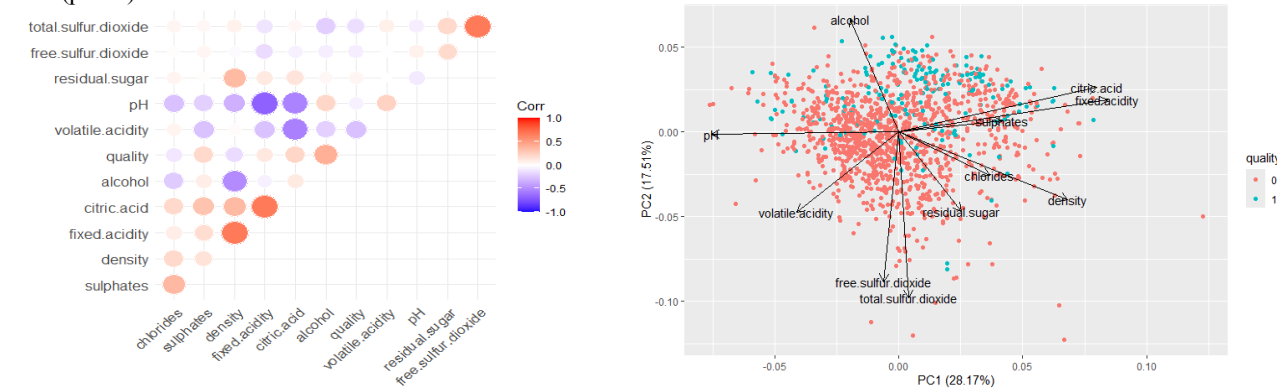
Для вирішення задачі машинного навчання (класифікації) на основі методології аналізу та попередньої обробки даних було використано набір даних *Red Wine Quality*. Набір стосується червоних сортів

португальського вина «Vinho Verde» [1]. Мета класифікації – виявити чинники, пов'язані з ризиком невчасної доставки попередніх замовлень клієнтам та інформацію про одержувачів товарів. Набір даних складається з 1599 записів і включає 11 атрибутів та мітку. Це набір числових ознак, що визначають такі характеристики вин, як: значення фіксованої та летючої кислотності, вміст лимонної кислоти, залишкового цукру, хлоридів, вільного діоксиду сірки, загального діоксиду сірки, а також значення густини, pH, сульфатів та алкоголю. Результуюча змінна, якість вина, категоріальна, що приймає значення від 0 (дуже погана якість) до 10 (відмінна якість).

Аналіз та попередня обробка даних. Після завантаження дані були проаналізовані за допомогою звітів по зведеної статистики та побудована гістограма розподілу якості червоного вина. Графік показує що якісні вина значно переважають погані вина з великим відривом. Більшість вин є непоганими (оцінка 5 або 6 балів), але також є і вина поганої якості (3 або 4 бали). Переважна більшість якісних вин має рейтинг якості 7.

За результатами аналізу, визначено відсоток викидів в кожному атрибуті та їх кількість. Найбільша кількість викидів у атрибута «residual.sugar», 9,7% зі всіх значень. Враховано, що деякі з методів машинного навчання працюють погано за наявності викидів. Тому аномальні значення в наборі оброблені за допомогою метода **затискного перетворення**

Оскільки метою класифікації є прогноз якості вина, важливо дізнатися, яка зі змінних має найсильніший зв'язок з цільовою ознакою, якістю вина. З теплової карти, алкоголь має найсильнішу кореляцію з якістю вина (рис.3).



а) Теплова карта б) Результати методу головних компонент

Рис. 3. Аналіз та попередня обробка даних

Джерело: сформовано автором

Для подолання мультиколеніарності використано метод головних компонент (рис.3). Таким чином, головними властивостями, що розрізняють якісне та посереднє вино, є вміст сульфатів та рівень алкоголю. Алкоголь має найсильніший зв'язок з якістю вина.

Для мінімізації «кореляційних плеяд» використано фактор інфляції дисперсії VIF. Для відбору оптимальної підмножини змінних в моделях машинного навчання використані методи обгортки (методи *покрокового включення* та *покрокового виключення*).

Перевірка розподілів в ознак в наборі даних та аналіз описової статистики виявили необхідність нормалізації. Нормалізацію числових змінних виконано з допомогою перетворення Йео-Джонсона, тому що набір містить нульові значення.

Результати роботи класифікаторів

Для вибору оптимальної моделі для вирішення завдання класифікації якості вина було проведено моделювання з використанням різних алгоритмів. Оцінка результатів класифікації за допомогою метрик якості зібрано у таблицю 1.

Таблиця 1.

Оцінка результатів класифікації

Тип моделі	Коефіцієнт успішності (*accuracy*)	Коефіцієнт помилок (*error rate*)	Каппа-статистика (*Kappa*)	Чутливість (*sensitivity*)	Специфічність (*specificity*)	Точність (*precision*)	Повнота (*recall*)	F-міра (*F-measure*)
LDA	0.8616	0.1384	0.4166	0.9922	0.3226	0.8581	0.9922	0.9203

QDA	0.8868	0.1132	0.5774	0.9766	0.5161	0.8929	0.9766	0.9329
Bagging	0.8742	0.1258	0.5166	0.9766	0.4516	0.8803	0.9766	0.9260

Оцінка кращої моделі була здійснена на основі аналізу ефективності кожної моделі, проте рішення про вибір було прийнято з урахуванням значень *F*-міри. Отже, найкраща модель – QDA зі значенням F-міри=0.9329, або 93%.

Висновки

В роботі досліджується методологія аналізу та попередньої обробки даних для вирішення задач машинного навчання. Методологія об'єднує наступні групи методів: методи обробки пропусків в даних, методи обробки аномальних значень, методи генерації ознак, методи ідентифікації нелінійностей та нестаціонарностей та методи нормалізації. Методи поєднано на основі системного підходу. Детально досліджено методи генерації ознак. Методи відбору і генерації ознак поділяють на три основні групи: методи фільтрації, методи обгортки і вбудовані методи. Запропоновано комбінований метод відбору ознак, який включає поетапне використання і методів фільтрації і методів обгортки. Метод складається з кількох кроків для ефективного вибору найбільш релевантних ознак набору даних. Він пропонує кращий підхід до відбору ознак. Це призводить до покращення продуктивності моделі з меншою кількістю ознак і скорочення обчислювальних витрат.

References

1. Roonizi, K. (2023) Kalman filter/smoother-based design and implementation of digital IIR filters, *Signal Processing*, vol. 208, 108958. URL: https://doi.org/10.1016/j.sigpro.2023.108958
2. Bidyuk, P., Kalinina, I. & Gozhyj, A. (2022) An Approach to Identifying and Filling Data Gaps in Machine Learning Procedures. *Lecture Notes on Data Engineering and Communications Technologies*, vol. 77, (pp. 164–176). [in Switzeland]
3. Kiptoon, D. (2023) Understanding Feature Engineering in Machine Learning. [Online]. URL: https://medium.com/@jdkiptoon/understanding-feature-engineering-in-machine-learning-59fc343a29c9 (application date: 15.08.2024).
4. Kalinina, I. & Gozhyj, A. (2022) Modeling and forecasting of nonlinear nonstationary processes based on the Bayesian structural time series. *Applied Aspects of Information Technology*, vol. 5, no. 3, (pp. 240-255). URL: https://doi.org/10.15276/aait.05.2022.17.

DOI 10.34132/mspc2025.01.14.07

Oleksandr Gozhyj,
Iryna Kalinina

METHODOLOGY OF DATA ANALYSIS AND PRE-PROCESSING FOR SOLVING MACHINE LEARNING PROBLEMS

The paper describes and investigates the methodology of data analysis and pre-processing for solving machine learning problems. The methodology combines the following groups of methods based on a systematic approach: methods for processing data gaps, methods for processing abnormal values, methods for generating features, methods for identifying nonlinearities and non-stationarities, and methods for normalization. The method of generating features is investigated. Methods for selecting and generating features are divided into three main groups: filtering methods, wrapper methods, and embedded methods. The feature generation method consists of five steps for effectively selecting the most relevant features of the data set. It offers a better approach to feature selection. This leads to improved model performance with fewer features and reduced computational costs. To experimentally verify the methodology of data analysis and pre-processing, the creation of a red wine quality classification system was considered. To solve the problem of wine quality classification, modeling was carried out using various algorithms. The effectiveness of the method for solving data analysis and preprocessing problems for solving classification problems was proven.

Key words*, data analysis and preprocessing methodology, feature generation methods, data gap processing methods, outlier processing methods, nonlinearity and non-stationarity identification methods, and normalization methods.*

Відомості про авторів/ Information about the Authors

Олександр Гожий, доктор технічних наук, професор кафедри інтелектуальних інформаційних систем, Чорноморського національного університету імені Петра Могили, м. Миколаїв, Україна. E-mail: alex.gozhyj@gmail.com, orcid: https://orcid.org/0000-0002-3517-580X.

Oleksandr Gozhyj, Doctor of Technical Sciences, Professor of the Department of Intellectual Information Systems,of the Petro Mohyla Black Sea National University, Mykolaiv, Ukraine. E-mail: alex.gozhyj@gmail.com, orcid: https://orcid.org/0000-0002-3517-580X

Ірина Калініна, доктор технічних наук, професор кафедри інтелектуальних інформаційних систем, Чорноморського національного університету імені Петра Могили, м. Миколаїв, Україна. E-mail: irina.kalinina1612@gmail.com, orcid: https://orcid.org/0000-0001-8359-2045

Iryna Kalinina, Doctor of Technical Sciences, Professor of the Department of Intellectual Information Systems,of the Petro Mohyla Black Sea National University, Mykolaiv, Ukraine. E-mail: irina.kalinina1612@gmail.com, orcid: https://orcid.org/0000-0001-8359-2045

© О. Гожий, І. Калініна *Стаття отримана редакцією 19.05.2025.*