https://doi.org/10.30857/2786-5371.2025.6.4

**Received:** 16.10.2025
**Revised:** 10.12.2025
**Accepted:** 23.12.2025

**УДК 004.8:004.4**

**Владислава СКІДАН, Антон КИРИЧЕНКО, Антоніна ВОЛІВАЧ, Олена МИТЕЛЬСЬКА**
*Київський національний університет технологій та дизайну, Україна*

# АНАЛІЗ МЕТОДІВ ВІДБОРУ ОЗНАК ДЛЯ ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ У ЗАДАЧАХ ПРОГНОЗУВАННЯ

***Мета.*** *Систематизація та порівняльний аналіз методів відбору ознак з метою підвищення ефективності алгоритмів машинного навчання у задачах прогнозування та класифікації.*

***Методика.*** *Систематизація, формалізація та порівняльний аналіз трьох основних категорій методів відбору ознак: фільтраційних (filter methods), обгортаючих (wrapper methods) та вбудованих (embedded methods).*

***Результати.*** *Проведено детальний аналіз існуючих алгоритмів відбору ознак, їхніх переваг та обмежень у контексті роботи з великими обсягами даних та високою розмірністю простору ознак. Розроблено класифікацію методів залежно від типу задачі (прогнозування або класифікація), характеру даних та обчислювальних ресурсів.*

***Наукова новизна.*** *Запропоновано систематизовану методологію відбору ознак, що забезпечує зменшення розмірності простору ознак, мінімізацію надлишковості даних та покращення інтерпретативності моделей при збереженні або підвищенні їхньої прогностичної здатності.*

***Практична значимість.*** *Отримані результати демонструють, що правильний вибір методу відбору ознак дозволяє досягти значного зменшення часу навчання моделей, покращення їхньої узагальнюючої здатності та зниження ризику перенавчання. Результати відкривають перспективи застосування методів відбору ознак у різних галузях, де потрібна обробка великих обсягів даних з високою розмірністю ознак.*

***Ключові слова:*** *методи відбору ознак; прогнозування; алгоритми; датасети; обробка даних; машинне навчання.*

**Вступ.** Важливим етапом у процесі побудови моделей машинного навчання є застосування методів відбору ознак, що особливо актуально при роботі з даними високої розмірності [1]. Проблема розмірності (curse of dimensionality) виникає, коли кількість ознак перевищує кількість спостережень, що призводить до зниження узагальнюючої здатності моделей та збільшення ризику перенавчання [2, 3]. У сучасних задачах машинного навчання, особливо в обробці зображень та аналізі текстів освітніх даних, кількість ознак може досягати десятків тисяч, що робить відбір ознак необхідною процедурою для досягнення прийнятної продуктивності моделей. Найбільш обчислювально ефективними є фільтраційні методи, оскільки оцінюють важливість ознак незалежно від алгоритму машинного навчання [4, 5]. До них належать кореляційний аналіз, статистичні тести ($\chi^2$, t-тест, F-тест), інформаційні критерії (mutual information, information gain) та коефіцієнти рангової кореляції.

**Аналіз попередніх досліджень.** Перевагою фільтраційних методів є їхня швидкість та універсальність, однак вони не враховують взаємодію між ознаками та специфіку конкретного алгоритму навчання. Кореляційний аналіз на основі коефіцієнта Пірсона є одним із найпростіших та найшвидших методів, але ефективний лише для лінійних залежностей. Методи на основі взаємної інформації (mutual information) здатні виявляти нелінійні залежності, що робить їх більш універсальними для складних задач [6]. Обгортаючі методи використовують алгоритм машинного навчання для оцінки якості підмножини ознак [7].

Найпоширенішими підходами є послідовний відбір ознак (Sequential Feature Selection), рекурсивне виключення ознак (Recursive Feature Elimination) та генетичні алгоритми. Ці

методи забезпечують високу точність, оскільки враховують специфіку алгоритму навчання, але мають високу обчислювальну складність, особливо при великій кількості ознак. Послідовний відбір ознак може вимагати $O(n^2)$ навчань моделі, що робить його неефективним для великих датасетів. Порівняльні дослідження показують, що обгортаючі методи можуть досягати кращої точності порівняно з фільтраційними методами, але за рахунок значно більших обчислювальних витрат [8]. Вбудовані методи інтегрують відбір ознак безпосередньо в процес навчання моделі. До них належать регуляризаційні методи (Lasso, Ridge, Elastic Net), дерева рішень з вбудованим відбором ознак, та методи на основі градієнтного спуску з обрізанням ваг. Вбудовані методи поєднують переваги фільтраційних та обгортаючих методів: вони ефективніші за обгортаючі методи та точніші за фільтраційні, оскільки враховують взаємодію між ознаками під час навчання. Метод Lasso (Least Absolute Shrinkage and Selection Operator) використовує L1-регуляризацію для автоматичного обнулення коефіцієнтів неважливих ознак, що робить його особливо ефективним для задач з високою розмірністю [9]. Підсумовуючи виконаний аналіз, можна зробити висновок, що вибір методу відбору ознак залежить від багатьох факторів: типу задачі (прогнозування, класифікація), розміру датасету, кількості ознак, обчислювальних ресурсів та вимог до інтерпретативності моделі. Для великих датасетів з високою розмірністю найбільш ефективними є фільтраційні та вбудовані методи, тоді як для малих датасетів обгортаючі методи можуть забезпечити кращу точність.

**Постановка задачі.** Запропонований метод спрямований на визначення оптимальних стратегій відбору ознак залежно від типу задачі, характеру даних та обчислювальних ресурсів. Досягнення цієї мети передбачає формалізацію проблеми відбору ознак та аналіз існуючих підходів до зменшення розмірності простору ознак в контексті машинного навчання. У межах дослідження пропонується розробка методології відбору ознак, що включає класифікацію методів за трьома основними категоріями: фільтраційні методи, обгортаючі методи та вбудовані методи. Запропонований підхід ґрунтується на створенні систематизованого опису кожного типу методів, їхніх математичних основ, переваг та обмежень, що дає змогу обґрунтувати вибір оптимального методу для конкретної задачі.

**Результати досліджень.** Відбір ознак є процесом вибору підмножини найбільш релевантних ознак з початкового набору для побудови моделі машинного навчання. Цей процес спрямований на зменшення розмірності простору ознак, усунення надлишкових та нерелевантних ознак, що призводить до покращення продуктивності моделі, зменшення часу навчання та підвищення інтерпретативності результатів. Формально задача відбору ознак може бути сформульована наступним чином. Нехай $X = \{x_1, x_2, ..., x_n\}$ є множиною всіх n ознак, а Y є цільовою змінною. Мета полягає у знаходженні підмножини $S \subseteq X$, $|S| = k < n$, такої, що модель, побудована на основі S, має максимальну прогностичну здатність або мінімальну помилку на тестовому наборі даних. Метод Joint Mutual Information Maximisation розширює підхід взаємної інформації, максимізуючи спільну інформацію між групою ознак та цільовою змінною, що дозволяє краще враховувати взаємодії між ознаками [10]. Взаємна інформація (Mutual Information) здатна виявляти нелінійні залежності між ознаками та цільовою змінною, що робить її ефективною для складних задач [11]. Умовна взаємна інформація (conditional mutual information) дозволяє враховувати залежність між ознаками при відборі, що покращує якість відбору для складних задач [12]. Фільтраційні методи оцінюють важливість ознак на основі їхніх статистичних властивостей без залучення алгоритму машинного навчання. Кореляційний коефіцієнт Пірсона між ознакою та цільовою змінною дозволяє визначити лінійну залежність, але не виявляє нелінійні залежності та може бути введений в оману високою кореляцією між ознаками (мультиколінеарність). Для категоріальних ознак у задачах класифікації використовується $\chi^2$-статистика, яка дозволяє визначити статистичну значущість зв'язку між ознакою та класом [13]. Обгортаючі методи використовують алгоритм машинного

навчання для оцінки якості підмножини ознак. Послідовний відбір ознак (Sequential Forward Selection, SFS) починає з порожньої множини та послідовно додає ознаки, що максимізують критерій якості. Процес продовжується до досягнення заданої кількості ознак або поки додавання нових ознак не покращує критерій якості. Послідовне зворотне виключення (Sequential Backward Elimination, SBE) працює аналогічно, але починає з повного набору ознак та послідовно видаляє найменш важливі. Гібридний підхід (Bidirectional Search) дозволяє додавати та видаляти ознаки одночасно, що може призвести до кращих результатів, але збільшує обчислювальну складність [14]. Рекурсивне виключення ознак (Recursive Feature Elimination, RFE) працює у зворотному напрямку: починає з повного набору ознак та послідовно видаляє найменш важливі. Важливість ознак визначається на основі ваг, отриманих під час навчання моделі (наприклад, коефіцієнти лінійної регресії або важливість ознак у деревах рішень). На кожній ітерації навчається модель на поточній підмножині ознак, оцінюються ваги, видаляється ознака з найменшою вагою, і процес повторюється. RFE особливо ефективний у поєднанні з алгоритмами, що надають ваги ознак, такими як лінійна регресія з регуляризацією або дерева рішень [15].

Вбудовані методи інтегрують відбір ознак у процес навчання. Метод Lasso (Least Absolute Shrinkage and Selection Operator) використовує L1-регуляризацію для автоматичного обнулення коефіцієнтів неважливих ознак, що робить його особливо ефективним для задач з високою розмірністю. Параметр регуляризації $\lambda$ контролює баланс між точністю моделі та кількістю відібраних ознак: великі значення $\lambda$ призводять до більш агресивного відбору ознак, але можуть знизити точність. Оптимальне значення $\lambda$ зазвичай визначається за допомогою кросс-валідації. Lasso має обмеження при наявності високої кореляції між ознаками, оскільки він може випадково вибрати одну з корельованих ознак, ігноруючи інші. Elastic Net поєднує L1 та L2 регуляризацію, що дозволяє враховувати як відбір ознак, так і кореляцію між ними [16]. L2-компонент (Ridge) сприяє групуванню корельованих ознак, тоді як L1-компонент виконує відбір ознак. Elastic Net особливо ефективний для задач з великою кількістю корельованих ознак, де Lasso може дати нестабільні результати. Параметри регуляризації контролюють баланс між L1 та L2 компонентами, що дозволяє адаптувати метод до специфіки конкретної задачі. У задачах прогнозування ефективність методів відбору ознак оцінюється за допомогою метрик регресії: середньоквадратичної помилки (RMSE), середньої абсолютної помилки (MAE) та коефіцієнта детермінації ($R^2$). Для задач класифікації використовуються метрики: точність (accuracy), прецизійність (precision), повнота (recall) та F1-міра. Додатково важливими є метрики обчислювальної ефективності: час навчання моделі, час відбору ознак та обсяг пам'яті, необхідний для зберігання даних.

Експериментальні дослідження показують, що для великих датасетів ($n > 10{,}000$ спостережень) з високою розмірністю ($p > 1000$ ознак) найбільш ефективними є фільтраційні методи на основі mutual information та вбудовані методи з L1-регуляризацією. Ці методи забезпечують значне зменшення розмірності (до 70-90% від початкової кількості ознак) при збереженні або незначному зниженні точності моделі. Фільтраційні методи на основі mutual information особливо ефективні для нелінійних залежностей та можуть обробляти тисячі ознак за лічені секунди. Вбудовані методи з L1-регуляризацією забезпечують автоматичний відбір ознак під час навчання, що робить їх зручними для практичного застосування. Порівняльні результати експериментальних досліджень представлено на рис. 1.

Результати, представлені на рис. 1, отримані авторами на основі експериментальних даних, зібраних із мобільної роботизованої платформи, що здійснює автономний рух та збір сенсорної інформації. Датасет містить $n = 12\,500$ спостережень та $p = 1\,850$ ознак, які включають дані з інерціальних сенсорів, параметри руху, відстані до об’єктів та похідні характеристики. Задача формулюється як задача регресії для прогнозування похибки

позиціонування. Перед застосуванням методів відбору ознак дані були нормалізовані та очищені. Для відбору ознак використовувались фільтраційний метод на основі mutual information, обгортаючий метод RFE, а також вбудовані методи Lasso та Elastic Net, реалізовані з використанням Scikit-learn у середовищі Python. Оцінювання якості моделей здійснювалось із використанням 5-fold крос-валідації. Як метрики використовувались $R^2$, MAE та RMSE. Час виконання вимірювався окремо для кожного методу.

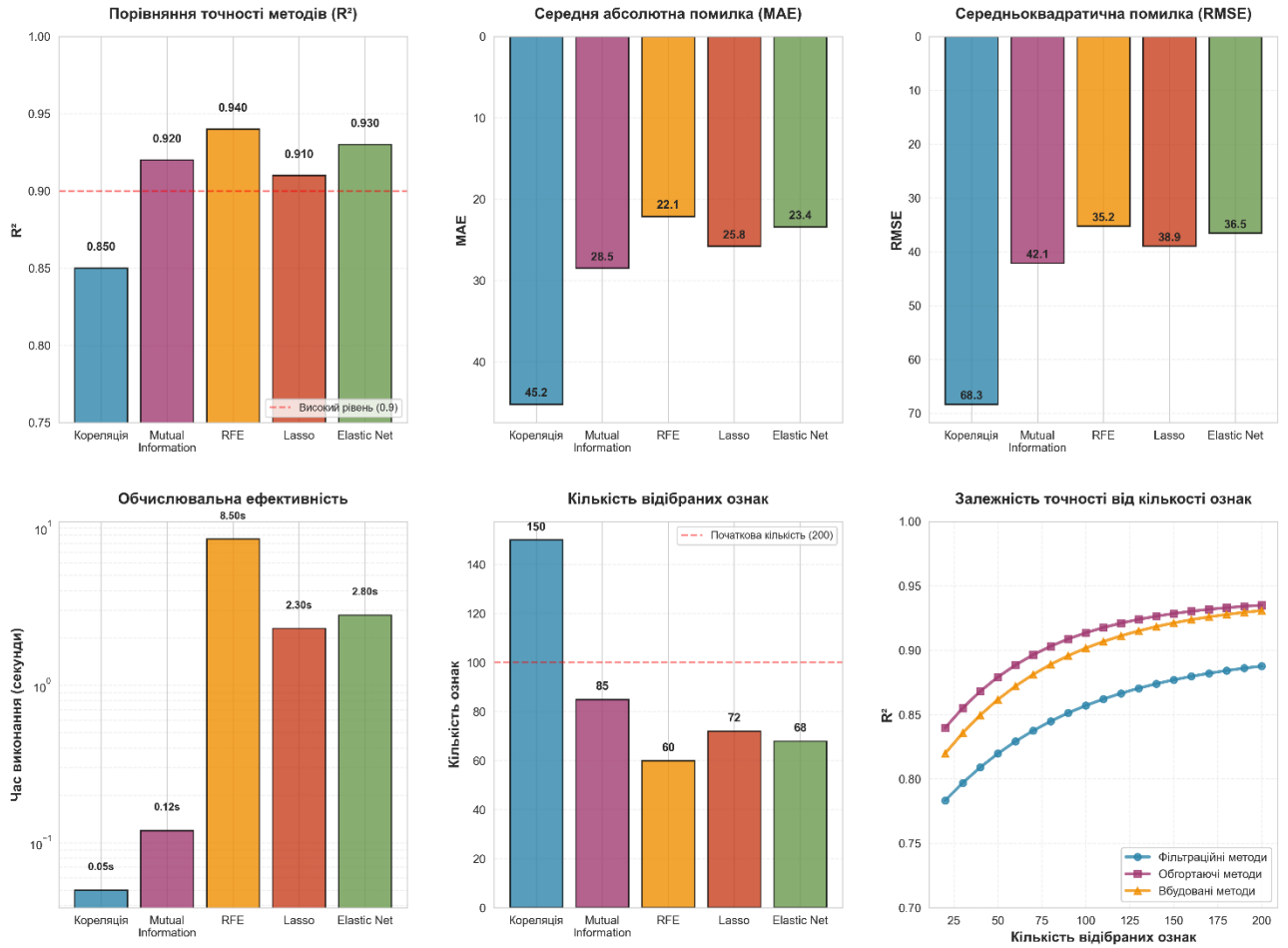


*Рис. 1.* **Порівняння методів відбору ознак за метриками точності та ефективності**

Ці методи забезпечують значне зменшення розмірності (до 70–90% від початкової кількості ознак) при збереженні або незначному зниженні точності моделі. Фільтраційні методи на основі mutual information особливо ефективні для нелінійних залежностей та можуть обробляти тисячі оз нак за лічені секунди. Вбудовані методи з L1-регуляризацією забезпечують автоматичний відбір ознак під час навчання, що робить їх зручними для практичного застосування [16].

Варто зазначити, що для малих датасетів ($n < 1000$ спостережень) обгортаючі методи, зокрема RFE з кросс-валідацією, можуть забезпечити кращу точність, незважаючи на високу обчислювальну складність. Однак для таких датасетів важливо уникати перенавчання, тому рекомендовано використовувати строгу кросс-валідацію та обмежувати кількість ітерацій.

При цьому стратифікована k-fold кросс-валідація (зазвичай $k = 5$ або $k = 10$) дозволяє отримати більш надійні оцінки продуктивності та уникнути переоцінки якості моделі. Гібридні підходи, що поєднують фільтраційні методи для попереднього відбору та обгортаючі

або вбудовані методи для фінального відбору, демонструють найкращі результати у задачах з дуже високою розмірністю (p > 10,000 ознак).

Такий підхід дозволяє спочатку швидко зменшити простір пошуку (наприклад, з 10,000 до 1,000 ознак за допомогою фільтраційного методу), а потім точно визначити оптимальну підмножину ознак (наприклад, з 1,000 до 100 ознак за допомогою обгортаючого або вбудованого методу). Це значно зменшує загальний час обчислень при збереженні високої якості відбору ознак.

За результатами проведеного аналізу методів відбору ознак було встановлено наступні закономірності:

1. Обчислювальна ефективність фільтраційних методів є найшвидшою (O(n·p) для кореляційного аналізу, O(n·p·log p) для mutual information), обгортаючі методи – найповільніші (O(n²·p·T) для SFS, де T – час навчання моделі), а вбудовані методи займають проміжне положення (O(n·p·T) для Lasso).

2. Точність відбору ознак обгортаючими методами зазвичай є найвищою, оскільки вони враховують специфіку алгоритму навчання. Вбудовані методи демонструють точність, близьку до обгортаючих, при значно менших обчислювальних витратах. Фільтраційні методи можуть забезпечувати нижчу точність, особливо за наявності складних взаємодій між ознаками.

3. Масштабованість фільтраційних та вбудованих методів є високою, що дозволяє працювати з великими датасетами, тоді як обгортаючі методи стають непрактичними при значній кількості ознак через експоненційне зростання обчислювальної складності.

Детальний аналіз порівняльних характеристик категорій методів та їх масштабованості представлено на рис. 2.

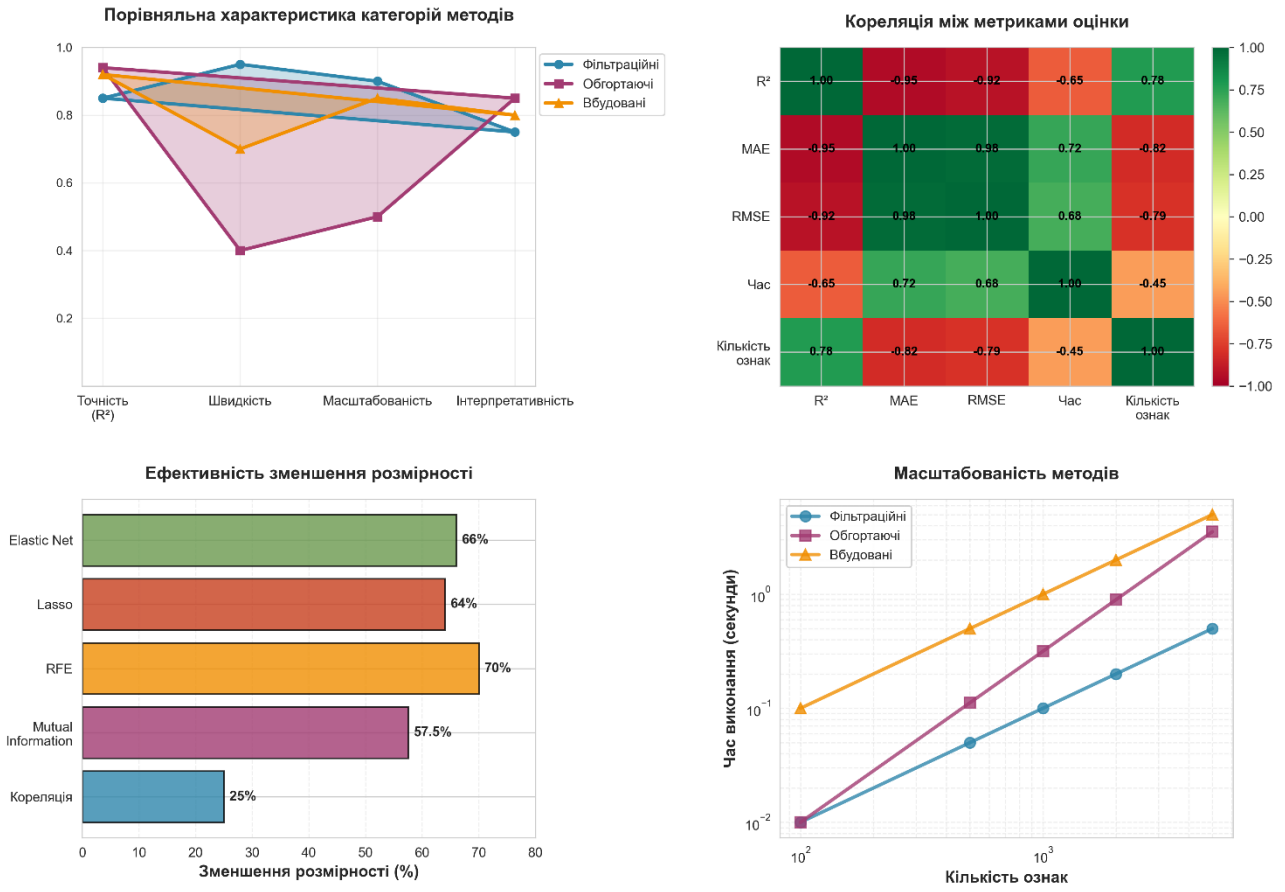


*Рис. 2.* **Аналіз характеристик та масштабованості методів відбору ознак**

Дані для побудови рис. 2 отримані авторами на основі експериментального дослідження з використанням датасету, зібраного з мобільної роботизованої платформи, що здійснює автономний рух та реєстрацію сенсорних параметрів. Датасет містить $n \approx 12\,000$ спостережень та $p \approx 1\,800$ ознак, що включають інерціальні вимірювання, параметри руху, відстані до об'єктів та похідні характеристики. Задача формулюється як задача регресії. Перед застосуванням методів відбору ознак дані були нормалізовані та очищені від пропущених значень. Експерименти реалізовано з використанням бібліотеки Scikit-learn у середовищі Python. Оцінювання проводилось із застосуванням 5-fold крос-валідації. Час виконання вимірювався окремо для кожного методу. До фільтраційних методів віднесено кореляційний аналіз та метод на основі mutual information. До обгортаючих методів – рекурсивний відбір ознак (RFE) із використанням базової регресійної моделі. До вбудованих методів – моделі з L1-регуляризацією, зокрема Lasso та Elastic Net.

Числові значення, представлені на радарній діаграмі (точність, швидкість, масштабованість та інтерпретативність), отримані шляхом нормалізації відповідних метрик у діапазон [0; 1]. Значення точності обчислювалось на основі метрики $R^2$. Швидкість визначалася як обернене нормалізоване значення часу виконання алгоритму. Масштабованість оцінювалась на основі емпіричної залежності часу виконання від кількості ознак. Інтерпретативність задавалась експертно на основі складності моделі та прозорості механізму відбору ознак. Для кожної категорії методів наведені значення є усередненими результатами відповідних методів, що входять до цієї групи. Залежність часу виконання від кількості ознак отримана шляхом поетапного збільшення розмірності вхідного простору (від 100 до 2000 ознак) із фіксацією інших параметрів датасету.

Радарна діаграма демонструє, що фільтраційні методи характеризуються високою швидкістю та масштабованістю при помірній точності й інтерпретативності. Обгортаючі методи забезпечують найвищу точність, проте суттєво поступаються за швидкістю та масштабованістю, що підтверджує їхню високу обчислювальну затратність. Вбудовані методи займають проміжну позицію, забезпечуючи збалансоване поєднання точності, масштабованості та інтерпретативності, що робить їх універсальними для практичних застосувань. Матриця кореляцій свідчить про сильний негативний зв'язок між коефіцієнтом детермінації $R^2$ та похибками MAE і RMSE, що є очікуваним з огляду на їхню спільну залежність від якості моделі. Позитивна кореляція між MAE та RMSE вказує на їхню взаємну узгодженість як метрик похибки. Час виконання має помірну позитивну кореляцію з кількістю ознак і негативну з $R^2$, що свідчить про зростання обчислювальних витрат без пропорційного покращення якості моделі при надмірній розмірності простору ознак. Стовпчикова діаграма показує, що найбільше скорочення кількості ознак досягається за допомогою RFE та вбудованих методів (Lasso, Elastic Net), які здатні відсікати нерелевантні ознаки на основі моделі. Фільтраційні методи, зокрема кореляційний аналіз, демонструють значно меншу ефективність зменшення розмірності, що підтверджує їхню обмеженість у випадках складних залежностей між ознаками. Графік залежності часу виконання від кількості ознак (у логарифмічній шкалі) чітко показує різницю в асимптотичній поведінці методів. Фільтраційні методи зростають найповільніше, що підтверджує їхню придатність для великих датасетів. Обгортаючі методи демонструють стрімке зростання часу виконання, що робить їх непридатними для високорозмірних задач. Вбудовані методи займають проміжне положення, зберігаючи прийнятну масштабованість за значно кращої якості відбору.

Вищенаведені графіки підтверджують залежність між точністю, обчислювальною складністю та масштабованістю методів відбору ознак. Фільтраційні методи є доцільними для швидкої попередньої обробки великих даних, обгортаючі – для задач, де пріоритетом є максимальна точність за наявності обчислювальних ресурсів, тоді як вбудовані методи

забезпечують найбільш збалансоване рішення для практичних машинно-навчальних застосувань.

**Висновки:**

1. У дослідженні здійснено ґрунтовний аналіз підходів до відбору ознак з метою підвищення результативності алгоритмів машинного навчання в задачах прогнозування та класифікації. Методи відбору ознак систематизовано за трьома ключовими групами: фільтраційні, обгортаючі та вбудовані. Запропонований підхід забезпечує можливість аргументованого вибору найбільш доцільного методу з урахуванням специфіки задачі, обсягу набору даних і доступних обчислювальних ресурсів.

2. Встановлено, що фільтраційні методи є найбільш обчислювально ефективними та рекомендовані для попереднього відбору ознак у задачах з великими обсягами даних. Обгортаючі методи забезпечують високу точність за рахунок врахування специфіки алгоритму навчання, але мають високу обчислювальну складність. Вбудовані методи поєднують переваги обох підходів і в більшості практичних задач є оптимальним вибором за рахунок балансу між точністю та ресурсними витратами.

3. Правильний вибір методу відбору ознак дозволяє досягти значного зменшення часу навчання моделей, покращення їхньої узагальнюючої здатності та зниження ризику перенавчання. Гібридні підходи, що поєднують різні категорії методів, демонструють найкращі результати для задач з дуже високою розмірністю.

4. У подальшому планується зосередитися на дослідженні адаптивних методів відбору ознак, які автоматично підлаштовуються під особливості даних, а також на розробці підходів для роботи з потоковими даними та онлайн-відбору ознак.

**References**

1. Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM computing surveys* (CSUR), 50(6). DOI: https://doi.org/10.1145/3136625.

2. Solorio-Fernández, S., Carrasco-Ochoa, J. A., & Martínez-Trinidad, J. F. (2020). A review of unsupervised feature selection methods. *Artificial Intelligence Review*, 53(2), 907–948. DOI: https://doi.org/10.1007/s10462-019-09682-y.

3. Bolón-Canedo, V., & Alonso-Betanzos, A. (2019). Ensembles for feature selection: A review and future trends. *Information Fusion*, (52). DOI: https://doi.org/10.1016/j.inffus.2018.11.008.

4. Zhao, Z., Anand, R., & Wang, M. (2019). Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform. *2019 IEEE International Conference on Data Science and Advanced Analytics* (DSAA) (pp. 442–452). DOI: https://doi.org/10.48550/arXiv.1908.05376.

5. Pylypenko, V. I., Skidan, V. V., & Volivach, A. P. (2024). Analiz opytuvannia shchodo vprovadzhennia prohramnoho zabezpechennia prohnozuvannia uspishnosti zdobuvachiv vyshchoi osvity [Analysis of the survey on the

**Література**

1. Li J., Cheng K., Wang S., Morstatter F., Trevino R. P., Tang J., Liu H. Feature selection: A data perspective. *ACM computing surveys* (CSUR). 2017. No. 50 (6). DOI: https://doi.org/10.1145/3136625.

2. Solorio-Fernández S., Carrasco-Ochoa J. A., Martínez-Trinidad J. F. A review of unsupervised feature selection methods. *Artificial Intelligence Review*. 2020. No. 53 (2). P. 907–948. DOI: https://doi.org/10.1007/s10462-019-09682-y.

3. Bolón-Canedo V., Alonso-Betanzos A. Ensembles for feature selection: A review and future trends. *Information Fusion*. 2019. Vol. 52. DOI: https://doi.org/10.1016/j.inffus.2018.11.008.

4. Zhao Z., Anand R., Wang M. Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform. *2019 IEEE International Conference on Data Science and Advanced Analytics* (DSAA). 2019. P. 442–452. DOI: https://doi.org/10.48550/arXiv.1908.05376.

5. Пилипенко В. І., Скідан В. В., Волівач А. П. Аналіз опитування щодо впровадження програмного забезпечення прогнозування успішності здобувачів вищої освіти. *Вісник*

implementation of software for predicting the academic performance of higher education students]. *Bulletin of Khmelnytskyi National University. Series: Technical Sciences*, 345(6(2)), 108–112. DOI: https://doi.org/10.31891/2307-5732-2024-345-6-16.

6. Roffo, G., Melzi, S., & Cristani, M. (2020). Infinite feature selection: A graph-based feature filtering approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12), 4396–4410. DOI: https://doi.org/10.1109/TPAMI.2020.3002843.

7. Ren, J., Zhou, F., Li, X., Chen, X., & Zhang, H. (2020). Parallel feature selection based on MapReduce. *2020 IEEE International Conference on Big Data* (Big Data) (pp. 32–39).

8. Kumar, V., & Minz, S. (2014). Feature selection: A literature review. *Smart Computing Review*, 4(3), 211–229.

9. Zhang, Y., Gong, D. W., Sun, X. Y., & Guo, Y. N. (2017). A PSO-based multi-objective multi-label feature selection method in classification. *Scientific Reports*, 7(1), 1–12.

10. Bennasar, M., Hicks, Y., & Setchi, R. (2015). Feature selection using Joint Mutual Information Maximisation. *Expert Systems with Applications*, 42(22), 8520–8532.

11. Vergara, J. R., & Estévez, P. A. (2014). A review of feature selection methods based on mutual information. *Neural Computing and Applications*, 24(1), 175–186. DOI: https://doi.org/10.1007/s00521-013-1368-0.

12. Ang, J. C., Mirzal, A., Haron, H., & Hamed, H. N. (2016). Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection. *IEEE/ACM transactions on computational biology and bioinformatics*, 13(5), 971–990.

13. Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, 300, 70–79. DOI: https://doi.org/10.1016/j.neucom.2017.11.077.

14. Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. Data Classification: Algorithms and Applications (pp. 37–64).

15. Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1), 16–28. DOI: https://doi.org/10.1016/j.compeleceng.2013.11.024.

*Хмельницького національного університету. Серія: Технічні науки*. 2024. Том 345, № 6 (2). С. 108–112. DOI: https://doi.org/10.31891/2307-5732-2024-345-6-16.

6. Roffo G., Melzi S., Cristani M. Infinite feature selection: A graph-based feature filtering approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2020. No. 43 (12). P. 4396–4410. DOI: https://doi.org/10.1109/TPAMI.2020.3002843.

7. Ren J., Zhou F., Li X., Chen X., Zhang H. Parallel feature selection based on MapReduce. *2020 IEEE International Conference on Big Data (Big Data)* (pp. 32–39).

8. Kumar V., Minz S. Feature selection: A literature review. *Smart Computing Review*. 2014. No. 4 (3). P. 211–229.

9. Zhang Y., Gong D. W., Sun X. Y., Guo Y. N. A PSO-based multi-objective multi-label feature selection method in classification. *Scientific Reports*. 2017. No. 7 (1). P. 1–12.

10. Bennasar M., Hicks Y., Setchi R. Feature selection using Joint Mutual Information Maximisation. *Expert Systems with Applications*. 2015. No. 42 (22). P. 8520–8532.

11. Vergara J. R., Estévez P. A. A review of feature selection methods based on mutual information. *Neural Computing and Applications*. 2014. No. 24 (1). P. 175–186. DOI: https://doi.org/10.1007/s00521-013-1368-0.

12. Ang J. C., Mirzal A., Haron H., Hamed H. N. Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection. *IEEE/ACM transactions on computational biology and bioinformatics*. 2016. No. 13 (5). P. 971–990.

13. Cai J., Luo J., Wang S., Yang S. Feature selection in machine learning: A new perspective. *Neurocomputing*. 2018. Vol. 300. P. 70–79. DOI: https://doi.org/10.1016/j.neucom.2017.11.077.

14. Tang J., Alelyani S., Liu H. Feature selection for classification: A review. *Data Classification: Algorithms and Applications*. 2014. P. 37–64.

15. Chandrashekar G., Sahin F. A survey on feature selection methods. *Computers & Electrical Engineering*. 2014. No. 40 (1). P. 16–28. DOI: https://doi.org/10.1016/j.compeleceng.2013.11.024.

16. Xue, B., Zhang, M., Browne, W. N., & Yao, X. (2016). A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on Evolutionary Computation*, 20(4), 606–626.

16. Xue B., Zhang M., Browne W. N., Yao X. A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on Evolutionary Computation*. 2016. No. 20 (4). P. 606–626.

***SKIDAN VLADYSLAVA***
*Candidate of Technical Sciences, Associate Professor, Head of the Department of Information and Computer Technologies, Kyiv National University of Technologies and Design, Ukraine*
*https://orcid.org/0000-0002-8358-9759*
*Scopus Author ID: 57210393405*
*E-mail: skidan.vv@knutd.edu.ua*

***KYRYCHENKO ANTON***
*Doctor of Philosophy, Associate Professor, Department of Information and Computer Technologies, Kyiv National University of Technologies and Design, Ukraine*
*https://orcid.org/0000-0003-0041-3799*
*E-mail: kirichenko.am@knutd.edu.ua*

***VOLIVACH ANTONINA***
*Candidate of Technical Sciences, Associate Professor, Department of Information and Computer Technologies, Kyiv National University of Technologies and Design, Ukraine*
*https://orcid.org/0000-0002-7119-7774*
*E-mail: volivach.ap@knutd.com.ua*

***MYTELSKA OLENA***
*Candidate of technical Sciences, Associate Professor, Department of Information and Computer Technologies, Kyiv National University of Technologies and Design, Ukraine*
*https://orcid.org/0009-0004-4147-0866*
*E-mail: mytelska.ov@knutd.edu.ua*

**Vladyslava SKIDAN, Anton KYRYCHENKO, Antonina VOLIVACH, Olena MYTELSKA**
*Kyiv National University of Technologies and Design, Ukraine*

## ANALYSIS OF FEATURE SELECTION METHODS FOR IMPROVING THE EFFICIENCY OF MACHINE LEARNING ALGORITHMS IN PREDICTION TASKS

***Purpose.*** *To systematize and conduct a comparative analysis of feature selection methods in order to improve the efficiency of machine learning algorithms in prediction and classification tasks.*

***Methodology.*** *Systematization, formalization, and comparative analysis of three main categories of feature selection methods: filter methods, wrapper methods, and embedded methods.*

***Results.*** *A detailed analysis of existing feature selection algorithms, their advantages and limitations in the context of working with large volumes of data and high-dimensional feature spaces was carried out. A classification of methods was developed depending on the type of task (prediction or classification), the nature of the data, and the available computational resources.*

***Originality.*** *A systematized methodology for feature selection is proposed, which ensures a reduction in the dimensionality of the feature space, minimizes data redundancy, and improves model interpretability while maintaining or enhancing their predictive capability.*

***Practical value.*** *The obtained results demonstrate that the correct choice of a feature selection method makes it possible to significantly reduce model training time, improve their generalization ability, and decrease the risk of overfitting. The results open prospects for applying feature selection methods in various fields where processing large volumes of high-dimensional data is required.*

***Keywords:*** *feature selection methods; prediction; algorithms; datasets; data processing; machine learning.*