ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНІ ТЕХНОЛОГІЇ В ОСВІТІ

УДК 37.091:004.85
DOI https://doi.org/10.32782/2412-9208-2026-2-187-208

COMPARATIVE ANALYSIS OF POST-HOC EXPLANATION METHODS FOR SMALL-SCALE MACHINE LEARNING MODELS IN LEARNING ANALYTICS SYSTEMS

ПОРІВНЯЛЬНИЙ АНАЛІЗ МЕТОДІВ ПОСТ-ХОК ПОЯСНЕННЯ ДЛЯ МАЛИХ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ У СИСТЕМАХ ОСВІТНЬОЇ АНАЛІТИКИ

Vladyslav DEHTIAROV,
Postgraduate Student at the Department of Computer Science, Sumy State University
116, Kharkivska Str., Sumy, 40007, Ukraine

Владислав ДЕГТЯРЬОВ,
аспірант кафедри комп'ютерних наук,
Сумський державний університет,
вул. Харківська, 116, м. Суми, 40007, Україна

vvdehtiarov@gmail.com
https://orcid.org/0000-0002-1578-8588

Valentyna BOROVYK,
Candidate of Technical Sciences, Associate Professor at the Department of Computer Science, Sumy State University,
116, Kharkivska Str., Sumy, 40007, Ukraine

Валентина БОРОВИК,
кандидат технічних наук,
доцент кафедри комп'ютерних наук,
Сумський державний університет,
вул. Харківська, 116, м. Суми, 40007, Україна

v.borovyk@cs.sumdu.edu.ua
https://orcid.org/0000-0002-3668-6302

ABSTRACT

This paper addresses a practical problem of educational informatics – how to produce a transparent and reproducible explanation of a machine-learning prediction in a learning-analytics system, in a form readable by a teacher, a student, and an internal auditor. Grounded in a reference pedagogical scenario (an academic supervisor receives a 'high-risk' flag for a sixth-week student), the paper compares post-hoc explanation methods – LIME, LinearSHAP, TreeSHAP, KernelSHAP, Anchors, and Integrated Gradients – against the regulatory backdrop of the EU AI Act and GDPR. Evaluation criteria: computational

efficiency, stability (Jaccard similarity of the top-5 features across 10 reruns), fidelity (prediction shift when top-3 features are zeroed), and plausibility. Experiments on Iris (4 features), Wine (13), Breast Cancer (30) in Python 3.12; significance – Welch's t-test. Results: LinearSHAP runs under 0.1 ms at perfect stability and beats LIME at p = 0.045; LIME stability drops from 1.00 to 0.54 – 0.76 as dimensionality grows; KernelSHAP in high dimensions falls to 0.47 – not fit for audit; Anchors reaches the highest fidelity 0.75 at moderate stability; Integrated Gradients is deterministic but scales poorly (17 – 133 ms). None of the surveyed methods combines sub-millisecond time, perfect stability, and an IF-THEN format simultaneously – this gap is closed by the authors' method Greedy-Prune-Explain (GPE): a three-phase algorithm with $O(d^2 \cdot n)$ complexity, a precision $\geq \tau$ guarantee, and deterministic output. Expected pedagogical impact – shorter time-to-intervention, support for self-regulated learning, and documentary-grade reproducibility of pedagogical decisions. The paper concludes with practical guidelines for designers of educational information systems.

Key words: *explainable artificial intelligence, learning analytics, educational informatics, local model explanations, model interpretability, decision trees, teacher' digital competence.*

Вступ. Цифровізація освіти досягла стадії, коли більшість закладів – від окремих шкіл до регіональних ЗВО – щодня генерують дані, достатні для побудови прогнозних моделей щодо окремого студента. Learning Management Systems фіксують відкриття матеріалів, час на завдання, поведінку в тестах, відвідуваність; до цього додаються дані від систем електронного журналу, навчальних платформ, опитувань. На такому тлі розвинулися дві паралельні галузі: освітня аналітика (learning analytics) та педагогічна інформатика, які спираються здебільшого на моделі машинного навчання скромного розміру – логістичну регресію, дерева рішень, градієнтний бустинг на кількох сотнях ознак, короткі правилові системи.

Конкретний педагогічний інструментарій, у який вбудовуються ці моделі, складається з кількох регулярних задач. Прогнозування академічного ризику обслуговує тьюторську підтримку: куратор групи отримує сигнал про студента з підвищеною ймовірністю незадовільного результату й має ухвалити рішення про зміст, форму та час педагогічного втручання. Рекомендаційні модулі індивідуальної освітньої траєкторії працюють інакше – вони підбирають наступний навчальний об'єкт (тему, завдання, форму контролю) і вимагають, щоб студент розумів причину рекомендації, інакше особиста відповідальність за вибір губиться. Формувальне оцінювання у ЗВО спирається на проміжні моделі прогресу: викладач отримує не лише поточний бал, а й аналітичний коментар про сильні й слабкі сторони навчальної діяльності студента, який далі трансформується у зворотний зв'язок на наступному занятті. Окреме місце посідають модулі академічної доброчесності, де модель класифікує текст як такий, що ймовірно згенеровано автоматично; без пояснення викладач не може відрізнити необґрунтоване звинувачення від змістовної педагогічної бесіди зі студентом. У кожній із цих задач про-

ICV 2024: 62.64
ISSN 2412-9208

гноз моделі – інформація для педагогічного судження, а не його заміна. Якість судження прямо залежить від того, наскільки прозоро система формулює свою відповідь.

Питання пояснюваності таких моделей у педагогічному контексті не є лише академічним. Європейський акт про штучний інтелект (EU AI Act, 2024) [8] відносить системи, що приймають рішення щодо оцінювання, допуску до програм і прогнозування відсіву, до категорії «високого ризику» з відповідними вимогами щодо прозорості, документування та аудиту. GDPR у статті 22 окремо гарантує суб'єкту даних право на зрозуміле пояснення будь-якого значущого автоматизованого рішення. В українських реаліях це означає, що система, яка рекомендує викладачеві, кому з групи призначати додаткові консультації, має не лише коректно прогнозувати, але й коректно пояснювати свої рекомендації – і робити це відтворювано.

Регуляторний контур змикається з педагогічним. Європейська рамка цифрових компетентностей педагога DigCompEdu [17] виокремлює групу «Аналіз і доказова педагогіка», у якій інтерпретація результатів автоматизованого аналізу освітніх даних винесена в окрему професійну компетентність. Професійний стандарт «Вчитель закладу загальної середньої освіти» (Україна, 2020) [23] у блоці інформаційно-цифрової компетентності формулює близьку вимогу: уміння педагога використовувати цифрові інструменти для збирання, обробки й аналізу даних про навчальні досягнення учнів та приймати на цій основі педагогічно обґрунтовані рішення. В обох документах ключове слово – «обґрунтовані». Воно вимагає, щоб поведінка алгоритму була зрозумілою педагогові не лише за результатом, а й за аргументацією. XAI-метод, який повертає вектор ваг ознак без правила або словесного формулювання, такий стандарт задовольняє формально, але не насправді: формат пояснення є таким же важливим аспектом педагогічної придатності, як і його коректність.

Проблема в тому, що більшість наявних методів пост-хок пояснення ML-моделей формувалися як дослідницькі інструменти для великих обчислювальних середовищ, а не для локально розгорнутих педагогічних систем з обмеженими ресурсами. Систематичний огляд Arrieta та співавторів [5] описує десятки таких методів; окремо Rudin [21] аргументує тезу, що для рішень із високими ставками доцільніше одразу обирати інтерпретовану модель, а не накладати пояснення поверх чорної скриньки. Але навіть у такій консервативній парадигмі проектувальник педагогічної інформаційної системи змушений на етапі архітектурних рішень обирати конкретний метод пояснення – і бракує систематичного порівняння, яке дало б відповідь не в абстрактному обчислювальному середовищі, а з огляду на обмеження саме малих моделей, які дійсно використовуються в learning analytics.

Аналіз останніх досліджень і публікацій. Методи LIME [18], SHAP [13] та Anchors [19] утворюють основний інструментарій пост-хок пояснення з 2016 – 2018 років; Integrated Gradients [24] доповнює його напрямом градієнтних методів для диференційованих моделей. Більшість опублікованих порівнянь цих методів зроблено на великих моделях комп'ютерного зору й обробки природної мови, де обговорення зосереджується переважно на вірності пояснень, а не на часі відгуку чи відтворюваності. Огляди Romero & Ventura [20] та Samek і співавт. [22] фіксують зростання інтересу до педагогічно релевантної XAI-проблематики, але систематичне порівняння методів саме в режимі малих моделей – того режиму, у якому живе освітня аналітика – у доступній нам літературі відсутнє. Саме цей методологічний розрив і обумовлює актуальність дослідження.

Мета статті – провести систематичне порівняння п'яти методів пост-хок пояснення (LIME, LinearSHAP, TreeSHAP, KernelSHAP, Anchors, Integrated Gradients) на малих моделях машинного навчання за трьома кількісними критеріями (обчислювальна ефективність, стабільність, вірність) та одним якісним (плаузибельність), а також сформулювати рекомендації щодо вибору методу для проектувальників систем освітньої аналітики.

Важливе методологічне уточнення. Ця стаття є частиною ширшого дослідження локальної пояснюваності дерев рішень у задачах бінарної класифікації, у межах якого авторами розроблено власний метод Greedy-Prune-Explain (GPE) і відповідний відкритий програмний фреймворк із scikit-learn-сумісним API. GPE є модельно-специфічним: на відміну від LIME, SHAP, Anchors та Integrated Gradients, які порівнюються в цій роботі, він безпосередньо використовує структуру дерева рішень для побудови мінімального правила «ЯКЩО-ТО». Представлений тут порівняльний аналіз встановлює обґрунтований методологічний контекст для GPE – фіксує, які з наявних методів і за якими характеристиками є найсильнішими бейзлайнами, і де саме у цих методах залишається простір, який закриває GPE. Усі експерименти цієї статті виконувалися на завданнях бінарної класифікації, що відповідає предметній рамці дисертаційного дослідження.

Педагогічний контекст і референтний кейс-сценарій

Щоб подальша технічна частина не сприймалася відірваною від практики, опишемо референтний педагогічний сценарій, до якого далі повертатимемося при обговоренні методів. Сценарій узагальнює регулярну ситуацію в роботі куратора академічної групи бакалаврату; конкретні цифри взято з типового діапазону параметрів LMS-аналітики українських ЗВО за матеріалами публікацій Romero & Ventura [20] та практики провідних університетів.

ICV 2024: 62.64
ISSN 2412-9208

Учасники сценарію. Куратор академічної групи другого курсу бакалаврату. Студент М., 19 років, перший семестр на спеціальності, шостий тиждень навчання з шістнадцяти. Викладач профільної дисципліни, який отримує дашборд для своєї групи.

Інформаційна система. LMS закладу інтегрована з електронним журналом і модулем прогнозування академічного ризику. Базова модель – неглибоке дерево рішень глибини 5, навчене на даних трьох попередніх потоків спеціальності. Вхідні ознаки: середня оцінка з контрольних заходів двох курсів-передумов, частка відкриття матеріалів у LMS за поточний семестр, середній час між отриманням завдання й першим відкриттям, частка пропусків аудиторних занять, серединний бал за стартові опитування курсу. Розмірність вхідного вектора – близько п'яти-семи ознак; такий обсяг типовий для дашбордів learning analytics [20].

Подія. На шостому тижні модель повертає для студента М. прогноз «високий ризик незадовільного підсумкового результату» з ймовірністю 0,78. Прапорець з'являється у дашборді куратора.

Що має відбутися далі. Куратор не може використати прогноз як підставу для адміністративної дії – це суперечило б педагогічній етиці й одночасно вимогам статті 22 GDPR щодо значущих автоматизованих рішень. Його дії пролягають через інший контур. По-перше, перевірити, на яких ознаках модель базує прогноз. По-друге, ухвалити рішення про педагогічне втручання – індивідуальна консультація, корекція темпу проходження матеріалу, рекомендація додаткових ресурсів, перерозподіл часу на профільну дисципліну. По-третє, поінформувати студента про ситуацію зрозумілою для нього мовою – не як «вас позначила система», а як «модель помітила те-то і те-то, давайте поговоримо». По-четверте, задокументувати причину втручання таким чином, щоб через місяць пояснити її методичній комісії або при внутрішньому аудиті закладу.

Чотири дії з попереднього абзацу – не алгоритмічна послідовність, а каркас педагогічного судження. Кожна з них висуває окрему вимогу до XAI-методу. Перша – щоб метод сам по собі повертав інтерпретовану структуру, а не лише прогноз. Друга – щоб ця структура була достатньо змістовною для зв'язування з конкретним педагогічним заходом. Третя – щоб формат пояснення залишався зрозумілим студенту без спеціальної підготовки. Четверта – щоб пояснення відтворилося через місяць у тих же умовах та повернуло ідентичний результат, інакше документ не виконує своєї аудиторської функції.

Розглянемо два формати виходу XAI-модуля на тій самій моделі. У першому варіанті система повертає вектор ваг «середня оцінка передумов: 0,42; час до першого відкриття: 0,29; відсутність на лекціях: 0,15». Цього достатньо для першої вимоги, частково для другої, але недостатньо для третьої та четвертої. У другому варіанті система повер-

тає правило «ЯКЩО середня оцінка передумов ≤ 3,2 І час до першого відкриття > 5 днів, ТО високий ризик». Тут задовольняються всі чотири вимоги одразу: куратор бачить дві конкретні підстави, обирає форму втручання (тренінг навичок самоорганізації плюс перевірка прогалин із курсів-передумов), пояснює студенту у форматі без спеціальної підготовки й документує рішення з тим самим правилом, яке через місяць поверне той самий результат на тих самих даних. Саме у цьому полягає педагогічна різниця між атрибутивним та правил-овим типами XAI, до якої далі повертаємось у порівняльному аналізі.

Методи та методики дослідження. Таксономія методів пояснення.

Методи пояснення ML-моделей зазвичай розділяють за трьома осями, кожна з яких має безпосереднє значення для педагогічного застосування [2; 9].

Інтринсичні vs. пост-хок. Декілька сімейств моделей – лінійна й логістична регресія, неглибокі дерева рішень, GAM – самі по собі читаються людиною: коефіцієнти, шляхи, форм-функції не потребують додаткової оболонки [12]. Пост-хок метод, натомість, обгортає вже навчену модель і синтезує пояснення без зміни її архітектури; саме ця гнучкість робить пост-хок підходи привабливими навіть тоді, коли базова модель формально інтерпретована.

Модельно-агностичні vs. модельно-специфічні. Агностичні методи (LIME, KernelSHAP, Anchors) трактують модель f(x) як чорну скриньку й працюють незалежно від її внутрішньої структури. Специфічні методи використовують структуру моделі для більшої ефективності або точності: TreeSHAP використовує топологію дерева для точного обчислення значень Шеплі у поліноміальному часі [14]; Integrated Gradients потребує диференційованості f, що одразу виключає дерева рішень і правильні системи.

Локальні vs. глобальні. Локальні пояснення описують один прогноз; глобальні узагальнюють поведінку моделі на всій вибірці. Більшість пост-хок робіт (і наша також) – локальні за задумом [15].

Критерії оцінювання

У педагогічному контексті критерії, за якими оцінюється якість пояснення, не зводяться до абстрактних академічних метрик. Вибираємо чотири, з яких три вимірюємо кількісно.

Обчислювальна ефективність. Час, за який система генерує одне пояснення, визначає, чи може воно взагалі бути представлене користувачеві в інтерактивному темпі. Банбері та співавтори [6] дають типові часові бюджети для малих моделей і крайових розгортань; саме в такий бюджет (порядок десятків мілісекунд) мають укладатися педагогічні дашборди та LMS-модулі. Вимірюємо середній час і стандартне відхилення на 30 пояснення на пару «метод – датасет».

ICV 2024: 62.64
ISSN 2412-9208

Стабільність. Якщо однаковий вхід породжує різні пояснення в різні моменти часу, довіра користувача зникає негайно. Alvarez-Melis і Jaakkola [4] показали, що методи на основі випадкового семплування здатні розходитися навіть для ідентичного входу. Кількісно оцінюємо стабільність як Jaccard-подібність топ-5 ознак між 10 повторними прогонами з різними сидами; значення 1,0 відповідає повній відтворюваності.

Вірність. Пов'язане, але окреме питання: чи справді виділені пояснювачем ознаки ті, на які реально спирається модель? Слідуючи традиції робіт з «перевірки здорового глузду» над saliency-мапами [3], ми обнуляємо три найбільш зважених ознаки й фіксуємо, наскільки змінюється прогнозна ймовірність моделі. Велика зміна – свідчення того, що пояснювач обрав справді вирішальні ознаки.

Плаузибельність – наскільки формат пояснення відповідає очікуванням користувача-не-спеціаліста [7]. Цю властивість розглядаємо якісно як властивість самого формату виведення (ваги ознак, правила, шляхи), бо повноцінне її вимірювання вимагає окремого користувацького дослідження.

Описи методів

LIME [18] перебуває на крайньому агностичному боці спектра. Для точки x він генерує збурені семпли в околі x, зважує їх за близькістю й підганяє інтерпретовану сурогатну модель – зазвичай розріджену лінійну регресію – на прогнозах моделі для цих семплів. Формально це оптимізація

$$\xi(x) = \operatorname{argmin}_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g),$$

де $\mathcal{L}$ – зважена ядром π_x втрата, Ω – регуляризатор складності сурогатної моделі g. Популярність LIME в прикладних проєктах пояснюється його гнучкістю: йому байдуже, що за модель f – нейромережа, калібрована логістична регресія, чи віддалений сервіс. Зворотний бік: два виклики на тому самому вхідному векторі можуть повернути різний набір топ-ознак, бо збурення – стохастичне [4]. Ширина ядра – ще один нетривіальний параметр; лінійний сурогат в околі суттєво нелінійного рішення є оптимістичним наближенням дійсної поведінки моделі.

SHAP [13] обґрунтовує атрибуцію ознак кооперативною теорією ігор. Значення Шеплі ознаки i – це її гранична важливість, усереднена за всіма коаліціями, до яких вона не входить:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \left[f(S \cup \{i\}) - f(S) \right]$$

Формула елегантна, але її пряме обчислення має складність $O(2^d)$, через що SHAP існує у вигляді родини варіантів, а не одного алго-

ритму. LinearSHAP згортає суму до $\phi_i = \beta_i(x_i) - \mathbb{E}[x_i]$ для лінійних моделей – точно, за $O(d)$, детерміновано. TreeSHAP використовує структуру дерева для точного обчислення за $O(TLD^2)$ на ансамблі з T дерев [14]. KernelSHAP – агностичний варіант на основі семплування коаліцій, який успадковує ту саму стохастичну проблему, що й LIME. Aas та співавт. [1] окремо вказують на нетривіальність вибору базового розподілу для скорельованих ознак – суб'єкт, який легко проглядається, якщо сприймати SHAP просто як «бібліотеку».

Anchors [19] вирішує іншу задачу: замість того, щоб розподіляти важливість по всіх ознаках, метод повертає одне кон'юнктивне правило A таке, що на більшості входів, які задовольняють A, модель зберігає свій прогноз:

$$\mathbb{E}_{D(z|A)}[\mathbf{1}_{f(x) = f(z)}] \geq \tau$$

Поріг τ – гарантія точності (0,95 у типовій реалізації). Привабливість очевидна: правило виду «ЯКЩО попередні провали > 0 І studytime ≤ 2 ТО fail» можна обговорити предметно з викладачем, що складніше для вектора числових ваг SHAP. Компроміси теж конкретні: на неперервних ознаках метод обирає вузькі інтервали з малим покриттям, а пошук кон'юнкцій через beam search – не швидкий.

Integrated Gradients [24] потребує диференційованої f. Метод атрибутує ознакам інтеграл градієнта моделі вздовж шляху від обраного базового входу x' до x:

$$IG_i(x) = (x_i - x_i') \times \int_0^1 \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} d\alpha.$$

Метод детермінований за побудовою і задовольняє дві аксіоми – чутливість та інваріантність реалізації [16]. Sturmfels та співавт. [23] показали, що різний вибір базового входу (нульовий, «розмитий», усереднений) може призводити до якісно різних атрибуцій на тому самому вході; крім того, обчислювальна ціна лінійно зростає з кількістю ознак і кроків інтегрування, що є прийнятним для прикладу з 4 ознаками, але стає обмеженням на 30 і більше.

Запропонований метод Greedy-Prune-Explain (GPE)

Проведений огляд чотирьох чинних методів (LIME, SHAP, Anchors, Integrated Gradients) дозволяє точно зафіксувати обмеження, на подолання яких спрямовано авторський метод локальної пояснюваності дерев рішень – Greedy-Prune-Explain (GPE). Метод і відповідний відкритий Python-фреймворк із scikit-learn-сумісним API розроблено у межах дисертаційного дослідження одного з авторів. Розгорнутий формальний опис, доведення властивостей та емпірична оцінка на фінансових і освітніх наборах даних становлять предмет окремих публікацій серії;

ICV 2024: 62.64
ISSN 2412-9208

у цій статті наводимо стислу версію, достатню для того, щоб позиціонувати GPE відносно тих методів, які порівнюються вище.

Ключова ідея – працювати з деревом рішень саме як із деревом, а не як із чорною скринькою. Нехай T – навчене бінарне дерево класифікації, X – навчальні дані (n екземплярів, m ознак), x – екземпляр, для якого потрібне пояснення, τ – поріг точності правила (за замовчуванням 0,95). Завдання: знайти мінімальне правило R у форматі «ЯКЩО-ТО» таке, що воно коректно класифікує x як T(x), його точність на X не нижча за τ, а кількість умов у R – мінімальна. Метод складається з трьох послідовних фаз.

Фаза GREEDY. Зчитуємо повний шлях від кореня до листа, якого досягає екземпляр x; отримуємо впорядкований список умов P. Вартість – O(d), де d – глибина дерева.

Фаза PRUNE. Ініціалізуємо R ← P. У циклі для кожної з умов обчислюємо точність залишкового правила на X, видаляємо ту умову, зняття якої найменше понижує точність (за умови, що ця точність не падає нижче τ). Цикл припиняється, коли жодна умова не може бути видалена без порушення обмеження на точність.

Фаза EXPLAIN. Формуємо фінальне правило R разом із метриками: precision (точність на фоновій вибірці), coverage (частка екземплярів, що задовольняють R) та complexity (|R|).

Теоретичні властивості методу – три, і всі три безпосередньо відповідають тим обмеженням наявних методів, які виявляє проведене нижче порівняння. По-перше, повна детермінованість: на тих самих вхідних даних GPE завжди повертає те саме правило (на відміну від LIME і KernelSHAP). По-друге, часова складність у найгіршому випадку $O(d^2 \cdot n)$ – для типового педагогічного дерева глибини 5 – 7 і фонової вибірки в тисячі екземплярів це укладається у субмілісекундний бюджет. По-третє, інваріант precision(R) ≥ τ зберігається протягом усього циклу відсікання за самою конструкцією алгоритму. Формат виведення – IF-THEN правило з однією-двома умовами – читається без спеціальної підготовки і тому природно придатний для комунікації з викладачем, студентом або внутрішнім аудитором освітнього закладу. Саме ця комбінація властивостей, недоступна жодному з чотирьох методів, описаних вище, і обґрунтовує доцільність введення окремого модельно-специфічного методу для дерев рішень у педагогічних інформаційних системах.

Експериментальний дизайн

Обрано ланцюжок датасетів зростаючої розмірності, щоб простежити масштабування методів. Iris (150 екземплярів, 4 ознаки), Wine (178, 13) і Breast Cancer Wisconsin (569, 30) – три стандартні навчальні набори scikit-learn, які у педагогічній інформатиці використовуються як базові; їх розмірність 4 – 30 ознак відповідає типовому обсягу ознакових векторів

у learning analytics (приміром, 30 ознак – середня кількість показників у learning dashboards [20]). Усі три задачі переформульовано як бінарні для однорідності оцінювання.

На кожному з датасетів тренувалися три моделі: логістична регресія (парна до LinearSHAP), дерево рішень max_depth=5 (парне до TreeSHAP та Anchors) та мілка нейромережа з одним прихованим шаром на 32 нейрони (для KernelSHAP та Integrated Gradients). Застосовність методу визначається моделлю – LIME універсальний і тестувався на всіх трьох. Розбиття train/test – 80/20 зі стратифікацією по класу.

Час генерації усереднювався за 30 прогонами на датасет. Стабільність вимірювалася Jaccard-подібністю топ-5 ознак за 10 повторними прогонами з різними сидами. Вірність – через обнулення трьох найбільш зважених ознак і реєстрацію зміни прогнозної ймовірності. Обчислення виконувалося в Python 3.12 з scikit-learn 1.8, SHAP 0.50, LIME 0.2, SciPy 1.11 на ноутбуці Intel Core i7 з 16 ГБ RAM. Статистична значущість – двовибірковий t-тест Велча, $\alpha = 0{,}05$.

Результати та дискусія

Обчислювальна ефективність

Таблиця 1 зводить результати вимірювання часу генерації пояснень.

Таблиця 1

Середній час генерації одного пояснення (мс) за різних методів і датасетів. Жирним виділено найшвидший метод у межах типу моделі

Модель	Метод	Iris (4)	Wine (13)	Breast C. (30)
Лог. регресія	LinearSHAP	**< 0,1**	**< 0,1**	**< 0,1**
Лог. регресія	Integ. Gradients	17,7	56,9	130,5
Лог. регресія	LIME	16,0	22,3	49,2
Дерево рішень	TreeSHAP	**0,1**	**0,1**	**0,1**
Дерево рішень	Anchors	0,9	8,7	20,0
Дерево рішень	LIME	15,9	22,4	49,5
Мілка НМ	KernelSHAP	**2,7**	**3,8**	**4,5**
Мілка НМ	Integ. Gradients	17,8	57,6	133,0
Мілка НМ	LIME	16,1	22,6	61,4

Декілька закономірностей виходять з таблиці на поверхню. LinearSHAP фактично ігнорує розмірність – від 4 до 30 ознак час тримається субмілісекундним, і цей метод переважає LIME на тих самих рядках лінійної моделі з $p = 0{,}045$. Integrated Gradients – протилежний випадок: від 17,7 мс на Iris до 130,5 мс на Breast Cancer, прискорення приблизно 7,4×, що відповідає теоретичній складності $O(d \cdot s)$ набли-

ICV 2024: 62.64
ISSN 2412-9208

ження інтеграла. Anchors займає середнє становище: 0,9 мс на 4-ознаковому дереві, 20 мс на 30-ознаковому – швидше за LIME на тому ж дереві, але повільніше за модельно-специфічні варіанти SHAP. Власне масштабування LIME приблизно у 3 рази на діапазоні 4 – 30 ознак, із виходом на 49 – 61 мс – уже на межі інтерактивних бюджетів.

Стабільність пояснень

Таблиця 2 подає головний з точки зору педагогічної реалізації результат: наскільки той самий вхід породжує той самий вихід при повторному виклику.

Таблиця 2

Стабільність пояснень (Jaccard-подібність топ-5 ознак між 10 прогонами). Жирним – детерміновані методи з гарантованою відтворюваністю

Модель	**Метод**	**Iris**	Wine	Breast C.
Лог. регресія	LinearSHAP	**1,00**	**1,00**	**1,00**
Лог. регресія	Integ. Gradients	**1,00**	**1,00**	**1,00**
Лог. регресія	LIME	1,00	1,00	0,76
Дерево рішень	TreeSHAP	**1,00**	**1,00**	**1,00**
Дерево рішень	Anchors	0,85	0,85	0,85
Дерево рішень	LIME	1,00	0,54	0,76
Мілка НМ	KernelSHAP	1,00	0,81	0,47
Мілка НМ	Integ. Gradients	**1,00**	**1,00**	**1,00**
Мілка НМ	LIME	1,00	0,93	0,93

Детерміновані методи поводяться так, як і передбачає їхня конструкція: LinearSHAP, TreeSHAP та Integrated Gradients стабільно тримають 1,00 на всіх датасетах, бо у них немає стохастичного компонента, з яким можна не погодитися при повторі. LIME на дереві демонструє падіння з 1,00 на Iris до 0,54 на Wine – менше половини ознак зберігається між прогонами. Anchors має стабільні 0,85 – очікувано для beam search із недетермінованим розрізненням. Окремо виділяємо KernelSHAP на мілкій нейромережі: падіння до 0,47 на Breast Cancer означає, що понад половину топ-ознак між прогонами не збігаються. Таку стабільність не можна довіряти будь-якому освітньому сценарію, у якому пояснення стає частиною офіційного запису (оцінювання, аудит, обґрунтування педагогічного втручання).

Вірність пояснень

Таблиця 3

Вірність (зміна прогнозної ймовірності при обнуленні топ-3 ознак). Вище = пояснювач правильніше ідентифікував важливі ознаки

Модель	Метод	Iris	Wine	Breast C.
Лог. регресія	LinearSHAP	0,35	0,22	0,09
Лог. регресія	Integ. Gradients	**0,40**	0,21	0,07
Лог. регресія	LIME	0,36	0,13	0,10
Дерево рішень	TreeSHAP	0,40	0,25	0,20
Дерево рішень	Anchors	**0,75**	**0,75**	**0,75**
Дерево рішень	LIME	0,40	0,25	0,19
Мілка НМ	KernelSHAP	0,31	0,28	0,08
Мілка НМ	Integ. Gradients	0,32	**0,28**	0,06
Мілка НМ	LIME	0,31	0,16	0,08

Anchors виокремлюється. Значення 0,75 на всіх трьох датасетах не випадкове: метод явно оптимізує правило на збереження прогнозу в локальному околі, а перевірка саме такою властивістю винагороджує. Атрибутивні методи (LIME, SHAP, IG) у 0,06 – 0,40 – суттєво нижче; при цьому значення падають із зростанням розмірності. Фізичний зміст цього падіння зрозумілий: коли модель читає 30 ознак, ймовірнісна вага розмазана по багатьох із них, і обнулення трьох рідко переважає рішення. Важливо: усередині конкретної моделі різниці між LIME, TreeSHAP та IG нерідко замалі, щоб на їх підставі обрати один метод як «більш вірний» за інші.

Застосування до педагогічних інформаційних систем

Поєднання результатів таблиць 1 – 3 дає проектувальнику педагогічної інформаційної системи чіткішу оптику для вибору методу, ніж будь-який окремо взятий показник. Поділимо обговорення за переліком регулярних педагогічних задач, окреслених у вступній частині та референтному кейс-сценарії, а не за переліком методів – оскільки у реальній системі задача диктує вибір методу, а не навпаки.

Прогнозування академічного ризику й раннє попередження. Цей сценарій має найвищу педагогічну ставку: куратор приймає рішення про втручання, а прозорість моделі є передумовою того, що це втручання буде доказово обґрунтоване, а не реакцією на «червоний прапорець без причини». Для тих закладів, де базовою моделлю є логістична регресія на 5 – 15 ознаках (мінімальний дашборд курсу), LinearSHAP лишається оптимальним: субмілісекундний час дозволяє оновлювати пояснення в реальному часі при кожній зміні даних, а нульова стохастика гарантує, що куратор, який повертається до екрана через годину, побачить ту саму атрибуцію. Для систем на основі дерев рішень – і саме така архітектура переважає у LMS українських ЗВО за оглядом Romero &

ICV 2024: 62.64
ISSN 2412-9208

Ventura [20] – TreeSHAP забезпечує ту саму субмілісекундну швидкість при детермінованості. Для фінального кроку – комунікації пояснення студенту або документування педагогічного рішення – викладач має додатково «перекласти» вектор ваг у словесну формулу; саме на цьому етапі атрибутивні методи поступаються правил-овим.

Диференційовані педагогічні втручання. Якщо пояснення моделі повертає правило «ЯКЩО частка пропусків > 0,3 І час до першого відкриття > 5 днів ТО ризик», викладач безпосередньо бачить два важелі для педагогічної дії. Перший – навичка планування часу (тренінг самоорганізації, технологія Pomodoro, корекція темпу проходження модуля, рекомендація консультації поза формальним розкладом). Другий – мотиваційний контур (індивідуальна бесіда, наставництво однокурсника-старшокурсника, перегляд персональних навчальних цілей, ведення навчального щоденника). Атрибутивне пояснення на тій самій моделі дало б «частка пропусків: 0,38; час до відкриття: 0,29», що залишає викладачу нести когнітивний тягар перекладу ваг у конкретну форму втручання. Експериментально виміряна вірність Anchors на рівні 0,75 показує: правил-овий формат педагогічно ефективніший на цьому етапі – попри гіршу стабільність порівняно з TreeSHAP.

Формувальне оцінювання й зворотний зв'язок. Принципова відмінність формувального оцінювання від підсумкового – у тому, що його основним адресатом є студент, а не адміністрація. Це робить плаузибельність формату пояснення ключовою методичною характеристикою: студент має не лише отримати оцінку, а зрозуміти, які саме компоненти його навчальної діяльності перебувають у задовільному стані, які – ні, і що з цим робити. Класична модель ефективного зворотного зв'язку Hattie & Timperley [10] передбачає, що дієвий фідбек одночасно відповідає на чотири методичні питання: «куди прямую» (feed-up), «як справляюся зараз» (feed-back), «що далі» (feed-forward) і метакогнітивне «що з цього досвіду перейде у наступний крок». Правил-ове пояснення Anchors природно лягає на цей формат: «ЯКЩО опрацьовано модулі 1 – 3 І виконано не менше 4 із 5 контрольних завдань ТО успішне завершення розділу» одночасно задає мету, фіксує стан і вказує наступний крок. Вектор ваг SHAP такої структури не несе сам по собі – її доводиться відтворювати викладачу, що додає методичну роботу до кожного прогнозу.

Розвиток метакогнітивних навичок студента. Прозорий ШІ у навчальному середовищі виконує не лише комунікативну, а й формувальну функцію – він стає дзеркалом, у якому студент бачить свою навчальну діяльність очима моделі. Holstein & Aleven [11] обґрунтовують концепцію teacher-AI complementarity: коли пояснення алгоритму доступне студенту, той починає переносити логіку самооцінювання й планування з конкретного курсу на загальну стратегію навчання. У

термінах рамки саморегульованого навчання Zimmerman [25] це підвищує якість фази forethought – попереднього планування навчальної дії. Формат пояснення тут визначальний: правило «ЯКЩО час до першого відкриття завдання > 5 днів ТО ризик відставання» дає студенту керовану одиницю поведінки, тоді як «час до відкриття: 0,29» – лише число без точки прикладання зусиль.

Академічна доброчесність. Окремий клас педагогічних задач, де неправильне рішення без пояснення прямо порушує етику. Якщо модель класифікує текст роботи як «ймовірно автоматично згенерований» з імовірністю 0,84, цього недостатньо для дисциплінарного провадження. Викладач повинен показати студенту, на яких саме ознаках побудовано рішення (поверхневі статистики тексту, патерни вживання сполучників, рівномірність синтаксичної складності, розподіл частотності службових частин мови), і дати студенту можливість предметно відповісти. Правил-ове пояснення «ЯКЩО частка узгоджених варіацій довжини речень < 0,15 І середня лексична різноманітність > порогу ТО підозра» одночасно слугує основою педагогічної бесіди й документом для можливого апеляційного процесу.

Цифрова компетентність викладача. Поява XAI-модулів у LMS закладу трансформує сам зміст професійної компетентності педагога. Інтерпретація пояснень стає окремим педагогічним умінням, яке необхідно формувати – як свого часу формували уміння оцінити тестове завдання за критеріями валідності й надійності. У рамці DigCompEdu [17] це вкладається в групу C («Цифрова педагогіка»), де виокремлюється вміння «використовувати дані учнів для адаптації навчання». На рівні підготовки магістрів освіти й перепідготовки чинних викладачів це означає необхідність включити модуль роботи з пояснюваним ШІ у методичні курси – і на рівні «прочитати пояснення», і на рівні «обрати метод під свій педагогічний сценарій». Без такої підготовки навіть найточніша модель з найзручнішим поясненням залишається непрозорим артефактом, який педагог боїться інтерпретувати.

Інформована згода й академічна політика. Окремий аспект педагогічної підзвітності – інформована згода: студент має право знати, що його дані обробляються прогнозною моделлю, і отримати конкретне пояснення кожного значущого рішення. Це переводить XAI з технічного компонента у компонент академічної політики закладу: положення про використання AI-аналітики в навчальному процесі, протокол комунікації прогнозу студенту, форма документування педагогічного втручання, регламент апеляції. У відсутності таких документів навіть юридично коректне пояснення лишається непридатним для повсякденного використання – у викладача й куратора немає процедурної підстави звертатися до нього.

ICV 2024: 62.64
ISSN 2412-9208

Прагматичний підсумок. Для модулів, що працюють із лінійними моделями – LinearSHAP. Для дерев рішень у режимі реал-тайм дашборда викладача – TreeSHAP (0,1 мс, точно, детерміновано). Для кейс-студій конкретного студента й комунікації з педрадою або батьками – Anchors. Для офлайн-аналізу за необхідності аксіоматичних гарантій – Integrated Gradients. LIME залишається як інструмент прототипування і інспекції сурогатних моделей, але не як продакшн-компонент педагогічної системи. KernelSHAP для освітніх кейсів, де пояснення потрапляє до офіційного запису, не рекомендуємо. Як показує перебір вимог чотирьох педагогічних сценаріїв, повного збігу «субмілісекундний час + ідеальна стабільність + правил-овий формат + гарантія точності» не дає жоден з аналізованих методів – саме цей розрив і закриває розроблений авторами Greedy-Prune-Explain.

Педагогічна ефективність і вплив на навчальний процес

Окремо акумулюємо очікувані ефекти запропонованого вибору методів на три виміри навчального процесу, що є опорними для рубрики «Інформаційно-комунікаційні технології в освіті».

Ефективність навчання. Перехід від атрибутивного пояснення (вектор ваг ознак) до правил-ового (Anchors або GPE) скорочує когнітивне навантаження викладача й студента під час інтерпретації прогнозу. Замість обчислювальної задачі «оцінити сумарний внесок п'яти ознак» учасник отримує одну-дві умови, на яких ґрунтується рішення моделі. У термінах моделі ефективного зворотного зв'язку [10] таке пояснення одночасно відповідає на питання feed-up (куди прямую), feed-back (як справляюся) та feed-forward (що далі), тоді як вектор ваг покриває переважно лише feed-back. Очікуваний педагогічний ефект – стискання циклу від отримання прогнозу до конкретної педагогічної дії, що зменшує час реакції в системі раннього попередження й підвищує ймовірність вчасного втручання.

Формування компетентностей. Прозорий ШІ у навчальному середовищі впливає на дві групи компетентностей одночасно. У студента розвиваються метакогнітивні навички: побачивши власну навчальну діяльність у форматі правила «ЯКЩО час до першого відкриття завдання > 5 днів І частка пропусків > 0,3 ТО ризик відставання», він отримує операціоналізоване уявлення про власну саморегуляцію, що відповідає фазі forethought моделі саморегульованого навчання [25]. У викладача формується цифрова компетентність із групи С («Цифрова педагогіка») рамки DigCompEdu [17], зокрема навичка адаптивного навчання на основі учнівських даних. Передумова – зрозумілий формат пояснення, який ми отримуємо лише з правил-ових методів (Anchors, GPE) та частково з модельно-специфічних варіантів SHAP.

Прийняття педагогічних рішень. Стабільність пояснення – критична властивість для документування педагогічного втручання. Якщо при повторному виклику методу на тих самих даних повертається інший набір ознак (як це відбувається з LIME при стабільності 0,54 і KernelSHAP при 0,47 на високих вимірах), то задокументоване вчора пояснення вже не відтворюється сьогодні – і втрачає аудиторську силу. Детермінованість LinearSHAP, TreeSHAP та GPE дозволяє вбудувати пояснення безпосередньо у формуляр педагогічного рішення (документ методичної комісії, протокол консультації, звіт куратора). Очікуваний ефект – підвищення доказовості педагогічних рішень і зниження ризику оскарження неформального судження «викладач так вирішив» з боку студента чи його законних представників.

Концепція teacher-AI complementarity. Запропоноване зміщення з атрибутивних на правил-ові пояснення безпосередньо відповідає концепції доповнюваності між педагогом і алгоритмом, артикульованій у роботах Holstein і Aleven [11]: алгоритм бере на себе те, що добре формалізується (швидке зчитування паттернів у багатовимірних даних), а педагог – те, що погано формалізується (вибір форми втручання, етична оцінка ситуації, побудова комунікації зі студентом). Ефективність такого розподілу вирішально залежить від інтерфейсу – і саме інтерфейсом між алгоритмом і педагогом виступає метод пояснення.

Обмеження роботи

Сумлінно зазначимо межі узагальнення. Перше – об'єм даних: три табличні датасети в діапазоні 4 – 30 ознак добре відповідають характерним освітнім вимірам, але висновки можуть відрізнятися для текстових корпусів студентських робіт, часових послідовностей поведінки в LMS чи сильно корельованих ознак. Друге – апаратура: тестування проведено на типовому ноутбуці, а не на реальному шкільному або університетському сервері; у продакшн-середовищі сталі можуть змінитися. Третє – метрика вірності: перекосова перевірка через обнуленя топ-3 ознак має свої сліпі плями [3]; достатність (sufficiency) та повнота (comprehensiveness) дали б ширше бачення. Четверте – плаузибельність залишається якісною; повноцінне педагогічне дослідження сприйняття пояснень викладачами й студентами потребує окремого емпіричного формату.

Висновки. Систематичне порівняння п'яти методів пост-хок пояснення рішень малих моделей машинного навчання – LIME, LinearSHAP, TreeSHAP, KernelSHAP, Anchors та Integrated Gradients – у застосуванні до педагогічних інформаційних систем дозволяє сформулювати такі підсумки.

За швидкістю й стабільністю одночасно виграють модельно-специфічні варіанти SHAP. LinearSHAP повертає пояснення за менш ніж 0,1 мс, TreeSHAP – за 0,1 мс, обидва показують ідеальну стабільність 1,00 і переважають LIME за часом із $p = 0{,}045$ для лінійної моделі.

ICV 2024: 62.64
ISSN 2412-9208

Integrated Gradients дорівнює їм у стабільності, але програє у швидкості на високих вимірах (17 – 131 мс), через що переміщується у категорію офлайн-інструмента.

За стабільністю окремо виокремлюється KernelSHAP: падіння з 1,00 на Iris до 0,47 на Breast Cancer. Для освітнього сценарію, де пояснення потенційно стає частиною аудиторського запису, така стабільність є достатньою підставою для виключення методу з переліку кандидатів. LIME у схожій ситуації деградує менш різко (до 0,54 – 0,76), але також потребує або фіксації сиду, або переходу в офлайн-режим з кешуванням.

За вірністю пояснень лідирує Anchors (0,75 на всіх датасетах). Це робить метод оптимальним для педагогічних сценаріїв, у яких пояснення покликане бути предметом обговорення між викладачем, студентом і (за необхідності) батьками, – формат «ЯКЩО … ТО …» читається без спеціальної підготовки й дає конкретну зачіпку для педагогічного втручання. Атрибутивні методи (LIME, SHAP, IG) демонструють помірну вірність 0,06 – 0,40 із падінням зі зростанням розмірності.

Окремо сформулюємо педагогічний внесок проведеного порівняння. По-перше, дослідження конкретизує вплив XAI-інструментарію на ефективність навчання через канал точних і своєчасних педагогічних втручань: правил-овий формат пояснення скорочує цикл «прогноз → педагогічна дія» з кількох робочих сесій викладача до однієї консультації, переводячи прогноз моделі у поведінкову одиницю, з якою працювати методично легше, ніж із набором числових ваг. По-друге, дослідження доводить, що формування цифрової компетентності педагога потребує включення модуля інтерпретації пояснень до методичної підготовки – інакше навіть формально доступна XAI-інфраструктура залишається невикористаною, бо викладач не має операційного контуру звертатися до неї. По-третє, обґрунтовано формувальну роль XAI у розвитку метакогнітивних навичок студента: правил-ове пояснення дає студенту керовану одиницю поведінки й тим самим підтримує фазу попереднього планування у саморегульованому навчанні. По-четверте, зафіксовано вимоги до академічної політики закладу – положення про використання AI-аналітики, протокол комунікації, форма документування, регламент апеляції, – за відсутності яких ефект від технологічного компонента не реалізується у педагогічній практиці.

Універсального кращого методу не існує; вибір визначається педагогічним сценарієм, а не алгоритмічним смаком розробника. Для реал-тайм аналітичних дашбордів куратора – LinearSHAP і TreeSHAP. Для комунікації з педагогічним колективом про конкретного студента – Anchors. Для офлайн-аналізу з вимогою аксіоматичних гарантій – Integrated Gradients.

Одночасно результати проведеного порівняння виявляють цілком конкретні методологічні розриви: жоден із чинних методів не поєднує субмілісекундний час генерації, ідеальну стабільність і природний для викладача формат «ЯКЩО-ТО» одночасно. Саме цей розрив закриває розроблений авторами метод Greedy-Prune-Explain (GPE) – модельно-специфічний підхід, який використовує саму структуру дерева рішень, повертає компактне правило «ЯКЩО-ТО» за $O(d^2 \cdot n)$ із гарантією точності, і за рахунок детермінованості своєї побудови задовольняє вимогу відтворюваності, обов'язкову для педагогічного аудиту. Представлений у цій статті порівняльний аналіз є частиною ширшого дисертаційного дослідження локальної пояснюваності дерев рішень у задачах бінарної класифікації; він фіксує стартову лінію, від якої відштовхується GPE, і пояснює, чому саме такий контур властивостей було обрано цільовим. Детальний опис методу, його теоретичних гарантій та експериментальних результатів становить предмет окремих публікацій серії.

Перспективи подальших досліджень охоплюють кілька взаємопов'язаних напрямків. Технічний напрямок – валідація висновків на реальних наборах даних learning analytics української освітньої системи (моделі прогнозування успішності на рівні курсу й відсіву на рівні ЗВО), а також розробка легких реалізацій XAI-методів для систем з обмеженими апаратними ресурсами. Методико-педагогічний напрямок – розробка навчального модуля «Пояснюваний штучний інтелект в освітній аналітиці» для магістерських програм підготовки педагогів і програм підвищення кваліфікації чинних викладачів з апробацією на базі конкретного ЗВО. Емпірично-педагогічний напрямок – користувацьке дослідження сприйняття пояснень викладачами й студентами в умовах реального педагогічного процесу, з фіксацією зміни поведінки студентів у відповідь на пояснення моделі, а не лише на її прогноз. Інституційно-етичний напрямок – формування рамкових документів академічної політики закладу щодо використання AI-аналітики у навчальному процесі, включно з протоколом інформованої згоди студента та регламентом апеляції прогнозу.

Література

1. Aas K., Jullum M., Løland A. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence.* 2021. Vol. 298. Art. 103502. DOI: https://doi.org/10.1016/j.artint.2021.103502

2. Adadi A., Berrada M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access.* 2018. Vol. 6. P. 52138–52160. DOI: https://doi.org/10.1109/ACCESS.2018.2870052

3. Adebayo J., Gilmer J., Muelly M., Goodfellow I., Hardt M., Kim B. Sanity Checks for Saliency Maps. *Proceedings of the 32nd Conference on Neural Information Processing Systems.* Montréal, Canada, 2018.

ICV 2024: 62.64
ISSN 2412-9208

4. Alvarez-Melis D., Jaakkola T. S. On the Robustness of Interpretability Methods. *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning*. Stockholm, Sweden, 2018. P. 66–71.

5. Arrieta A. B., Díaz-Rodríguez N., Del Ser J. et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*. 2020. Vol. 58. P. 82–115. DOI: https://doi.org/10.1016/j.inffus.2019.12.012

6. Banbury C., Reddi V. J., Lam M. et al. Benchmarking TinyML Systems: Challenges and Direction. *Proceedings of the IEEE*. 2021. Vol. 109, No. 2. P. 211–233. DOI: https://doi.org/10.1109/JPROC.2020.3031354

7. Caruana R., Lou Y., Gehrke J., Koch P., Sturm M., Elhadad N. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Sydney, Australia, 2015. P. 1721–1730. DOI: https://doi.org/10.1145/2783258.2788613

8. European Parliament, Council of the European Union. Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). *Official Journal of the European Union*. 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj (дата звернення: 24.04.2026).

9. Guidotti R., Monreale A., Ruggieri S., Turini F., Giannotti F., Pedreschi D. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*. 2019. Vol. 51, No. 5. Art. 93. P. 1–42. DOI: https://doi.org/10.1145/3236009

10. Hattie J., Timperley H. The Power of Feedback. *Review of Educational Research*. 2007. Vol. 77, No. 1. P. 81–112. DOI: https://doi.org/10.3102/003465430298487

11.Holstein K., Aleven V. Designing for Human-AI Complementarity in K-12 Education. *AI Magazine*. 2022. Vol. 43, No. 2. P. 239–248. DOI: https://doi.org/10.1002/aaai.12058

12. Lipton Z. C. The Mythos of Model Interpretability. *Queue*. 2018. Vol. 16, No. 3. P. 31–57. DOI: https://doi.org/10.1145/3236386.3241340

13. Lundberg S. M., Lee S.-I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*. Long Beach, CA, USA, 2017. Vol. 30. P. 4765–4774.

14. Lundberg S. M., Erion G., Chen H. et al. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence*. 2020. Vol. 2. P. 56–67. DOI: https://doi.org/10.1038/s42256-019-0138-9

15. Miller T. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*. 2019. Vol. 267. P. 1–38. DOI: https://doi.org/10.1016/j.artint.2018.07.007

16. Murdoch W. J., Singh C., Kumbier K., Abbasi-Asl R., Yu B. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*. 2019. Vol. 116, No. 44. P. 22071–22080. DOI: https://doi.org/10.1073/pnas.1900654116

17. Redecker C. European Framework for the Digital Competence of Educators: DigCompEdu / ed. by Y. Punie. Luxembourg : Publications Office of the European Union, 2017. EUR 28775 EN. DOI: https://doi.org/10.2760/178382

18. Ribeiro M. T., Singh S., Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA, USA, 2016. P. 1135–1144. DOI: https://doi.org/10.1145/2939672.2939778

19. Ribeiro M. T., Singh S., Guestrin C. Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*. New Orleans, LA, USA, 2018. Vol. 32, No. 1. P. 1527–1535. DOI: https://doi.org/10.1609/aaai.v32i1.11491

20. Romero C., Ventura S. Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*. 2020. Vol. 10, No. 3. e1355. DOI: https://doi.org/10.1002/widm.1355

21. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*. 2019. Vol. 1, No. 5. P. 206 – 215. DOI: https://doi.org/10.1038/s42256-019-0048-x

22. Samek W., Montavon G., Vedaldi A., Hansen L. K., Müller K.-R. Explainable AI: Interpreting, *Explaining and Visualizing Deep Learning*. Berlin : Springer, 2019. 439 p.

23. Sturmfels P., Lundberg S., Lee S.-I. Visualizing the Impact of Feature Attribution Baselines. *Distill*. 2020. Vol. 5, No. 1. e22. DOI: https://doi.org/10.23915/distill.00022

24. Sundararajan M., Taly A., Yan Q. Axiomatic Attribution for Deep Networks. *Proceedings of the 34th International Conference on Machine Learning*. Sydney, Australia, 2017. P. 3319 – 3328.

25. Zimmerman B. J. Becoming a Self-Regulated Learner: An Overview. *Theory Into Practice*. 2002. Vol. 41, No. 2. P. 64 – 70. DOI: https://doi.org/10.1207/s15430421tip4102_2

References

1. Aas, K., Jullum, M., & Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, *298*, Article 103502. https://doi.org/10.1016/j.artint.2021.103502 [in English].

2. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052 [in English].

3. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. In *Proceedings of the 32nd Conference on Neural Information Processing Systems*. Montréal, Canada [in English].

4. Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. In *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning* (pp. 66–71). Stockholm, Sweden [in English].

5. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82 – 115. https://doi.org/10.1016/j.inffus.2019.12.012 [in English].

6. Banbury, C., Reddi, V. J., Lam, M., et al. (2021). Benchmarking TinyML systems: Challenges and direction. *Proceedings of the IEEE*, *109*(2), 211–233. https://doi.org/10.1109/JPROC.2020.3031354 [in English].

7. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1721–1730). https://doi.org/10.1145/2783258.2788613 [in English].

8. European Parliament, Council of the European Union. (2024). *Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*. Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj [in English].

9. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), Article 93, 1–42. https://doi.org/10.1145/3236009 [in English].

10. Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487 [in English].

ICV 2024: 62.64
ISSN 2412-9208

11. Holstein, K., & Aleven, V. (2022). Designing for human-AI complementarity in K-12 education. *AI Magazine*, *43*(2), 239–248. https://doi.org/10.1002/aaai.12058 [in English].

12. Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, *16*(3), 31–57. https://doi.org/10.1145/3236386.3241340 [in English].

13. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 4765–4774 [in English].

14. Lundberg, S. M., Erion, G., Chen, H., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*, 56–67. https://doi.org/10.1038/s42256-019-0138-9 [in English].

15. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007 [in English].

16. Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, *116*(44), 22071–22080. https://doi.org/10.1073/pnas.1900654116 [in English].

17. Redecker, C. (2017). *European framework for the digital competence of educators: DigCompEdu* (Y. Punie, Ed.; EUR 28775 EN). Publications Office of the European Union. https://doi.org/10.2760/178382 [in English].

18. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778 [in English].

19. Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1), 1527–1535. https://doi.org/10.1609/aaai.v32i1.11491 [in English].

20. Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, *10*(3), e1355. https://doi.org/10.1002/widm.1355 [in English].

21. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x [in English].

22. Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (2019). *Explainable AI: Interpreting, explaining and visualizing deep learning*. Springer [in English].

23. Sturmfels, P., Lundberg, S., & Lee, S.-I. (2020). Visualizing the impact of feature attribution baselines. *Distill*, *5*(1), e22. https://doi.org/10.23915/distill.00022 [in English].

24. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 3319 – 3328). Sydney, Australia [in English].

25. Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, *41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2 [in English].

АНОТАЦІЯ

Стаття розв'язує практичну задачу педагогічної інформатики – як забезпечити прозоре й відтворюване пояснення прогнозу, що повертає модель машинного навчання у системі освітньої аналітики (learning analytics), у формі, придатній одночасно для викладача, студента й внутрішнього аудиту закладу. На основі референтного педагогічного сценарію (куратор отримує сигнал «високий ризик» щодо студента) проведено порівняння методів пост-хок пояснення – LIME,

LinearSHAP, TreeSHAP, KernelSHAP, Anchors та Integrated Gradients. Регуляторний контур окреслено вимогами EU AI Act і GDPR щодо прозорості й відтворюваності значущих рішень. Методи оцінювання: обчислювальна ефективність, стабільність (Jaccard-подібність топ-5 ознак за 10 прогонами), вірність (зміна прогнозу при обнуленні топ-3 ознак), плаузибельність. Експерименти проведено на трьох публічних наборах зростаючої розмірності – Iris (4 ознаки), Wine (13), Breast Cancer Wisconsin (30) – у середовищі Python 3.12; статистичну значущість перевірено двовибірковим t-тестом Велча. Результати: LinearSHAP < 0,1 мс при абсолютній стабільності, переважає LIME за швидкістю з p = 0,045; стабільність LIME з розмірністю падає з 1,00 до 0,54 – 0,76; KernelSHAP у високих вимірах – 0,47 (непридатно для аудиту); Anchors – найвища вірність 0,75 за помірної стабільності; Integrated Gradients детермінований, але погано масштабується (17 – 133 мс). Жоден метод не поєднує субмілісекундний час, ідеальну стабільність і формат «ЯКЩО-ТО» одночасно – цей розрив закриває авторський метод Greedy-Prune-Explain (GPE): трифазний алгоритм зі складністю $O(d^2 \cdot n)$, гарантією precision ≥ τ та детермінованим виходом. Очікувані педагогічні ефекти – стискання циклу між прогнозом і втручанням, підтримка фази forethought саморегульованого навчання, документовно відтворюване обґрунтування педагогічного рішення. Сформовано практичні рекомендації для проектувальників педагогічних інформаційних систем закладів загальної та вищої освіти.

Ключові слова: *пояснюваний штучний інтелект, освітня аналітика, педагогічна інформатика, локальні пояснення моделей, інтерпретованість, дерева рішень, цифрова компетентність викладача.*

Дата першого надходження статті до видання: 24.04.2026
Дата прийняття статті до друку після рецензування: 08.05.2026
Дата публікації (оприлюднення) статті: 30.05.2026

ICV 2024: 62.64
ISSN 2412-9208