ISSN 2306-5478

УДК 004.85 https://doi.org/10.31713/vt1202612

Злотенко Б. М. [1; ORCID ID: 0000-0002-0870-8535],
д.т.н, професор,
Скідан В. В. [1; ORCID ID: 0000-0002-8358-9759],
к.т.н, доцент
Мительська О. В. [2; ORCID ID: 0009-0004-4147-0866],
к.т.н, доцент,
Афтанділянц В.Є. [2; ORCID ID: 0000-0003-0660-1395],
к.е.н, доцент,

[1]Київський національний університет технологій та дизайну

ВПЛИВ РОЗПОДІЛУ ДАНИХ НА КОРЕКТНІСТЬ ПОЯСНЕНЬ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ В ОСВІТНІЙ АНАЛІТИЦІ

У статті досліджується вплив статистичних властивостей розподілу даних на коректність пояснень моделей машинного навчання в задачах освітньої аналітики. Розглянуто основні методи пояснюваного штучного інтелекту, зокрема локальні підходи (LIME, SHAP), градієнтні атрибуції та контрфактичні пояснення, з точки зору їх залежності від дисбалансу класів, коваріатного зміщення та доменного зсуву даних. Показано, що якість пояснень визначається не лише архітектурою моделі та вибором функції атрибуції ϕ(), але й статистичною структурою даних. Зміна розподілу між навчальною та тестовою вибірками призводить до нестабільності ранжування ознак, збільшення локальної невідповідності (infidelity) та зниження інтерпретованості пояснень, навіть за умови збереження високої відповідності моделі f. Запропоновано узагальнену багатокритеріальну схему оцінювання пояснень, яка поєднує метрики відповідності моделі (fidelity, deletion/insertion AUC), локальної узгодженості (infidelity), стабільності відносно малих збурень вхідних даних та контрфактичної коректності (близькість, розрідженість, валідність).

Проведено формалізацію основних показників та їх систематизацію за чутливістю до змін розподілу даних. Експериментальну перевірку проведено на задачі прогнозування академічного ризику з використанням табличних даних (520/129 спостережень, 12 ознак). Результати демонструють суттєвий вплив дисбалансу класів і коваріатного зміщення на стабільність

пояснень та поведінку різних моделей, що підтверджується аналізом метрик deletion та порівнянням архітектур.

***Ключові слова*: штучний інтелект, освітні дані, інтерпретація моделей, дисбаланс класів, машинне навчання, класифікація, дисбаланс класів, explainable AI.**

Вступ. Сучасні системи освітньої аналітики (learning analytics) активно інтегрують методи машинного навчання з метою підтримки ухвалення рішень у закладах освіти, зокрема для прогнозування академічної успішності, ідентифікації студентів групи ризику та оцінювання рівня залученості здобувачів освіти в цифрових навчальних середовищах (LMS) [1]. Зі зростанням складності моделей підвищується їхня прогностична точність, однак одночасно загострюється проблема інтерпретованості, прозорості та підзвітності результатів перед педагогічними працівниками, адміністрацією та регуляторними органами. У цьому контексті методи пояснюваного штучного інтелекту Explainable AI (XAI) розглядаються як ключовий інструмент забезпечення інтерпретованості моделей шляхом оцінювання внеску вхідних ознак у результат функції прогнозування f(x) [2-3]. Проте сучасні підходи XAI, такі як LIME, SHAP та контрфактичні методи не є повністю інваріантними до статистичних властивостей даних, на яких здійснюється навчання моделі та побудова пояснень ϕ(x) [4]. Це зумовлює принципову проблему того що пояснення можуть залишатися формально узгодженими з поведінкою моделі f, однак бути нестабільними або змістовно некоректними при зміні розподілу даних. У свою чергу дисбаланс класів, коваріатне зміщення між P_train(x) та P_test(x), а також доменний зсув P(x, y) призводять до зміни простору локальних апроксимацій і порушення стабільності ранжування ознак.

Аналіз останніх досліджень і публікацій. Сучасні методи пояснюваного штучного інтелекту Explainable AI (XAI) умовно поділяються на кілька основних класів залежно від механізму побудови інтерпретацій. До першої групи належать локальні модельно-незалежні методи, серед яких одним із найбільш поширених є LIME, що здійснює апроксимацію поведінки складної моделі за допомогою локального лінійного сурогатного наближення в околі об'єкта x [5].

Друга група представлена методами, що базуються на теорії кооперативних ігор, зокрема SHAP, який оцінює маргінальний

ISSN 2306-5478

внесок ознак через усереднення за всіма можливими перестановками або їх наближеннями [6]. Третій клас включає градієнтні та інтегровані градієнтні методи, орієнтовані на диференційовні моделі глибокого навчання, де важливість ознак визначається через похідні виходу моделі за входами. Окремий напрям становлять контрфактичні пояснення, які формалізують інтерпретацію через пошук мінімальних змін вхідного об'єкта $x \rightarrow x'$, що призводять до зміни прогнозу моделі.

Поряд із розвитком методів інтерпретації значна увага приділяється оцінюванню їх якості. Базовими підходами є метрики відповідності моделі (fidelity), які вимірюють ступінь узгодженості пояснення з локальною поведінкою функції f, а також метрики локальної невідповідності (infidelity), що оцінюють розбіжність між прогнозованими та фактичними змінами виходу моделі при збуреннях [7]. Додатково використовуються показники стабільності, які характеризують стійкість атрибуцій до малих змін вхідних даних, а також процедури deletion/insertion, що аналізують зміну виходу моделі при послідовному виключенні або додаванні ознак у порядку їх важливості $|\phi_i(x)|$.

У сфері освітньої аналітики (learning analytics) XAI застосовується для прогнозування академічної успішності, виявлення студентів групи ризику та аналізу поведінкових патернів у LMS-системах [8]. Водночас більшість наявних досліджень зосереджується переважно на точності прогнозування моделей, тоді як проблеми стабільності та коректності пояснень за умов зміни розподілу даних залишаються недостатньо формалізованими, особливо в педагогічних застосуваннях порівняно з класичними дослідженнями у сфері машинного навчання.

Мета статті – дослідити вплив статистичних властивостей розподілу даних на коректність та стабільність XAI-пояснень у задачах освітньої аналітики. Сформувати узгоджену багатокритеріальну систему метрик для їх оцінювання та проаналізувати поведінку пояснень у різних умовах розподільних зсувів.

Виклад основного матеріалу. Сучасні системи освітньої аналітики широко застосовують методи машинного навчання для прогнозування академічної успішності, виявлення студентів групи ризику та підтримки прийняття рішень у цифрових освітніх середовищах. Формально задача зводиться до побудови відображення:

$$f: X \rightarrow [0,1]$$

де $X \subseteq R$ – простір ознак, що описують навчальну діяльність студента

значення f(x) – інтерпретується як ймовірність належності до класу ризику академічної неуспішності.

Однак у реальних системах недостатньо лише отримати точний прогноз. Критично важливо пояснити, чому модель прийняла те чи інше рішення. Для цього використовуються методи пояснюваного штучного інтелекту Explainable AI (XAI), які будують локальні інтерпретації у вигляді вектора атрибуцій [9]:

$$phi(x)=(phi\_1(x),...,phi\_d(x))^T \text{ in } R^d$$

де кожна компонента x – відображає внесок відповідної ознаки у прогноз моделі.

У класичному підході такі пояснення розглядаються як локальна лінійна апроксимація поведінки складної моделі в околі точки xxx, що реалізується, зокрема, у методі LIME та SHAP. Проте ключовим обмеженням є те, що якість пояснення залежить не лише від моделі f, але й від статистичного розподілу даних, на яких модель була навчена та на яких проводиться інтерпретація. У таких умовах модель може залишатися формально коректною, проте її локальні пояснення втрачають семантичну стабільність. Це означає, що пояснення ϕ(x)\phi(x)ϕ(x) можуть бути узгодженими з функцією f, але не відображати реальні причинно-наслідкові залежності в новому домені.

Для вирішення пропонується узагальнений протокол оцінювання пояснень, який поєднує обчислення всіх трьох метрик у єдиному процесі. Його можна формалізувати як ітеративну процедуру, що представлена в програмному лістингу 1:

Програмний лістинг 1: Алгоритмічний підхід оцінювання метрик

```
    for x in dataset:
phi = explain(model, x)

# infidelity
I = compute_infidelity(phi, model, x)

# stability
x_perturbed = x + noise
S = distance(phi, explain(model, x_perturbed))

# sparsity
C = count_significant(phi)
```

 ISSN 2306-5478

Цей підхід дозволяє одночасно оцінити як локальну точність пояснення, так і його структурні властивості.

Для практичної перевірки підходу використано задачу прогнозування академічної неуспішності на синтетично-реалістичних освітніх даних. Розмір вибірки становив 520 навчальних і 129 тестових спостережень при розмірності простору ознак d=12. У межах експерименту було реалізовано програмний модуль мовою Python із використанням бібліотек NumPy, scikit-learn та Matplotlib [10]. Модель класифікації будувалася на основі логістичної регресії, а пояснення ϕ(x) інтерпретувалися як вектор коефіцієнтів моделі, фрагмент виконавчого коду представлено у програмному лістингу 2.

Програмний лістинг 2: Фрагмент виконавчого коду

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.datasets import make_classification
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(
   X, y, test_size=129, random_state=42)

model = LogisticRegression(max_iter=1000)
model.fit(X_train, y_train)

def f(x):
   return model.predict_proba(x.reshape(1, -1))[0, 1]

def explain(model, x):
   return model.coef_[0]

# 5. Infidelity
def infidelity(phi, f, x, eps=0.1, M=100):
   errors = []
   for _ in range(M):
      xi = np.random.normal(0, eps, len(x))
      errors.append(
         (np.dot(xi, phi) - (f(x + xi) - f(x)))**2
      )
   return np.mean(errors)

eps_values = np.linspace(0.01, 0.5, 20)
inf_values = []
```

```
x_sample = X_test[0]
phi = explain(model, x_sample)

for eps in eps_values:
    val = infidelity(phi, f, x_sample, eps=eps)
    inf_values.append(val)

plt.figure(figsize=(8,5))
plt.plot(eps_values, inf_values, marker='o')
plt.title("Залежність infidelity від рівня збурення")
plt.xlabel("Рівень шуму (ε)")
plt.ylabel("Infidelity")
plt.grid(True)
plt.savefig("infidelity_plot.png")
plt.show()

for e, v in zip(eps_values[:5], inf_values[:5]):
    print(f"eps={e:.2f} -> infidelity={v:.5f}")
```

Для оцінювання якості пояснень було реалізовано обчислення метрики infidelity, що представлена на рис. 1. Вона визначає узгодженість пояснення з поведінкою моделі при випадкових збуреннях вхідних даних.

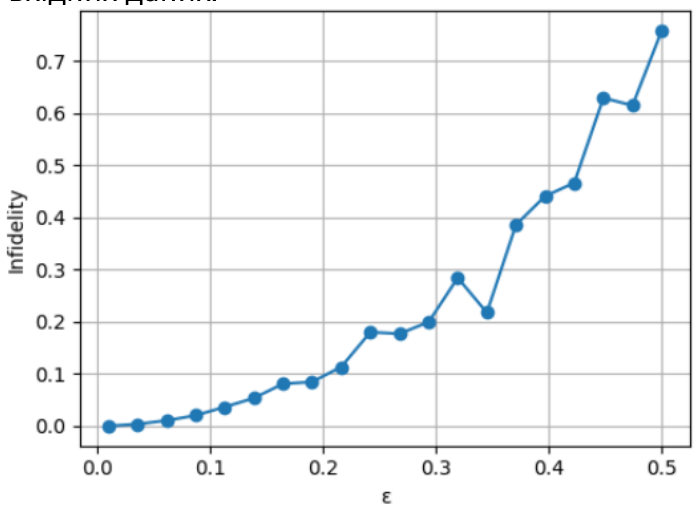

Рис.1. Залежність infidelity від рівня збурення

На даному графіку зображено залежність показника infidelity від величини збурення ε, що використовується для оцінювання локальної узгодженості пояснення з поведінкою моделі. Зі

ISSN 2306-5478

збільшенням параметра ε спостерігається чітка тенденція до зростання значення infidelity. У початковій області (приблизно ε∈[0.01;0.1]) значення метрики є близькими до нуля, що свідчить про високу узгодженість пояснення з моделлю при малих збуреннях вхідних даних. У зазначеному режимі лінійна апроксимація, що використовується для обчислення атрибуцій ϕ(x)\phi(x)ϕ(x), забезпечує точне відтворення локальної поведінки функції f(x).

У середньому діапазоні значень ε∈[0.1;0.3] спостерігається поступове, майже монотонне зростання infidelity. Це вказує на те, що зі збільшенням масштабу збурень модель дедалі менше відповідає лінійному наближенню, а отже, пояснення втрачає точність. У цій області вже проявляється нелінійний характер моделі. Для більших значень ε>0.3 зростання стає більш різким і нерівномірним, із локальними коливаннями. Це пояснюється тим, що збурення вхідного вектора виходять за межі локального околу точки xxx, де припущення про лінійність пояснення є валідним. Внаслідок цього різниця між реальним приростом f(x+ξ)–f(x)f(x+\xi) - f(x)f(x+ξ)–f(x) та його лінійною оцінкою через ϕ(x)\phi(x)ϕ(x) значно збільшується.

Отримана залежність підтверджує теоретичне положення про те, що показник infidelity є чутливим до масштабу збурень і може використовуватися як індикатор межі локальної адекватності пояснення. Низькі значення метрики при малих ε та її швидке зростання при їх збільшенні свідчать про обмеженість застосування локальних пояснювальних моделей у глобальному контексті.

Отриманий графік демонструє залежність метрики стабільності пояснень від рівня збурення вхідних даних ε.

У початковій області малих збурень (ε∈[0.01;0.1]) значення S1 є незначними та близькими до нуля. Це свідчить про те, що пояснення ϕ(x)\phi(x)ϕ(x) є стійкими до малих змін вхідних даних, тобто модель демонструє локальну стабільність: близькі об'єкти мають подібні інтерпретації.

У діапазоні ε∈[0.1;0.25] спостерігається поступове зростання метрики стабільності. Це означає, що навіть помірні збурення починають впливати на структуру пояснення, змінюючи вагомість окремих ознак. Така поведінка є типовою для моделей із нелінійними межами прийняття рішень.

Для більших значень ε>0.25 графік набуває нерівномірного, флуктуаційного характеру з різкими піками та спадами. Зокрема, видно локальні максимуми в області ε≈0.3–0.35 та подальше суттєве

зростання значень. Це свідчить про високу чутливість пояснень до збурень і втрату їхньої стабільності.

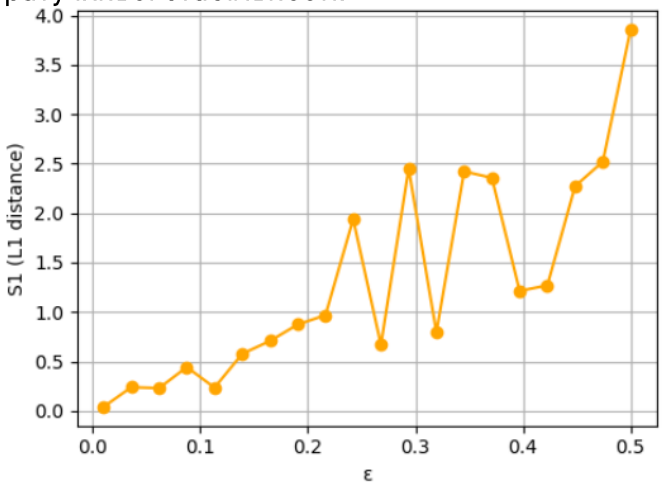

Рис. 2. Залежність стабільності пояснень від рівня збурення

Такі коливання зумовлені тим, що за значних збурень вхідний вектор x переходить до інших областей простору ознак, де змінюється локальна геометрія моделі. Унаслідок цього навіть незначні додаткові варіації можуть спричиняти суттєві зміни у ранжуванні важливості ознак.

Висновки. У роботі досліджено вплив статистичних властивостей даних на коректність пояснень моделей машинного навчання в задачах освітньої аналітики. Основну увагу зосереджено на аналізі методів пояснюваного штучного інтелекту (XAI) у контексті прогнозування академічної неуспішності, де інтерпретація результатів відіграє критичну роль у підтримці прийняття управлінських рішень у закладах освіти.

У межах дослідження формалізовано задачу оцінювання локальних пояснень у вигляді векторів атрибуцій ϕ(x)\phi(x)ϕ(x), що відображають внесок окремих ознак у прогноз моделі. Запропоновано багатокритеріальний підхід до оцінювання якості пояснень, який охоплює показники локальної узгодженості (fidelity), невірності (infidelity), стабільності до збурень та структурної складності пояснення. Такий підхід забезпечує комплексну оцінку пояснювальних методів як з точки зору узгодженості з моделлю, так і з позиції їхньої надійності та інтерпретованості.

Експериментальна частина дослідження базується на синтетично-реалістичних освітніх даних, що моделюють типові

ISSN 2306-5478

характеристики навчального процесу (академічна успішність, відвідуваність, активність у LMS тощо). Результати експериментів засвідчили, що якість пояснень істотно залежить від статистичного розподілу даних. Зокрема, встановлено, що зі збільшенням рівня шуму у вхідних даних спостерігається монотонне зростання показника infidelity, що свідчить про погіршення локальної узгодженості пояснень із поведінкою моделі. Одночасно з цим підвищується чутливість пояснень до малих змін вхідного вектора.

Окремо досліджено вплив дисбалансу класів та коваріатного зміщення. Виявлено, що за умов значного дисбалансу класів пояснення демонструють знижену стабільність, особливо для об'єктів рідкісного класу, що проявляється у варіативності атрибуцій та зміні ранжування ознак. У випадку коваріатного зміщення, коли розподіл ознак тестових даних відрізняється від навчального, модель функціонує в режимі екстраполяції, що негативно впливає як на точність прогнозування, так і на коректність пояснень.

Отримані результати підтверджують, що пояснення моделей машинного навчання не є інваріантними до розподілу даних і повинні розглядатися як функція не лише моделі, але й емпіричного розподілу P(x,y)P(x,y)P(x,y). Це має суттєві практичні наслідки для систем освітньої аналітики, оскільки некоректна інтерпретація результатів може призводити до помилкових управлінських рішень або дискримінаційних ефектів щодо окремих груп студентів.

Практична значущість роботи полягає у розробці відтворюваного підходу до оцінювання XAI-пояснень із реалізацією мовою Python. Запропонований інструментарій дозволяє аналізувати залежність метрик якості пояснень від параметрів шуму та досліджувати їх поведінку в різних статистичних умовах, що може бути використано при розробці прикладних аналітичних систем.

Водночас дослідження має певні обмеження. Використання синтетичних даних, попри їх наближеність до реальних, не повністю відображає складність освітніх процесів. Крім того, розглянуті методи пояснення обмежуються переважно лінійними атрибуційними підходами та не охоплюють повний спектр сучасних XAI-методів, зокрема для глибоких нейронних мереж та складних ансамблевих моделей.

Подальші дослідження доцільно спрямувати на:

- використання реальних освітніх датасетів;
- розширення набору методів XAI (зокрема SHAP та контрфактичних пояснень);
- розробку підходів до адаптації пояснень при зміні розподілу даних;

- інтеграцію етичних та нормативних вимог у процес оцінювання інтерпретованості моделей.

Таким чином, результати роботи підтверджують необхідність комплексного підходу до оцінювання пояснюваності моделей машинного навчання та підкреслюють важливість урахування статистичних властивостей даних при інтерпретації результатів у задачах освітньої аналітики.

1. Lang, C., Siemens, G., Wise, A., & Gasevic, D. (2017). *Handbook of learning analytics*. Society for Learning Analytics Research. 2. Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., ... & Ranjan, R. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. ACM computing surveys, 55(9), 1-33. 3. Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019, September). Explainable AI: A brief survey on history, research areas, approaches and challenges. In *CCF international conference on natural language processing and Chinese computing* (pp. 563-574). Cham: Springer International Publishing. 4. Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2025). A perspective on explainable artificial intelligence methods: SHAP and LIME. *Advanced Intelligent Systems*, *7*(1), 2400304. 5. Nguyen, H. T. T., Cao, H. Q., Nguyen, K. V. T., & Pham, N. D. K. (2021, May). Evaluation of explainable artificial intelligence: Shap, lime, and cam. In *Proceedings of the FPT AI Conference* (pp. 1-6). 6. Zhang, K., Xu, P., & Zhang, J. (2020, October). Explainable AI in deep reinforcement learning models: A SHAP method applied in power system emergency control. In *2020 IEEE 4th conference on energy internet and energy system integration (EI2)* (pp. 711-716). IEEE. 7. Patel, A. (2025, May). Revisiting Fidelity in Explainable AI: Unpacking Cognitive Biases and Deceptive Transparency in Model Interpretations. In *2025 5th Intelligent Cybersecurity Conference (ICSC)* (pp. 298-302). IEEE. 8. Pylypenko, V. (2025, September). Прогнозування високого рівня академічної успішності студентів з використанням машинного навчання. *Наука і техніка сьогодні*, 8(45), 1634–1649. https://doi.org/10.52058/2786-6025-2025-8(49)-1634-1649 9. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82-115. 10. HR, M. S., NR, M. N., & MU, M. D. (2026). *NUMERICAL PYTHON: SCIENTIFIC COMPUTING AND DATA SCIENCE APPLICATIONS WITH NUMPY, SCIPY AND MATPLOTLIB*. Vaagai international publishing house.

REFERENCES

1. Lang, K., Simens, G., Uayz, A., i Gasevich, D. (2017). Spravochnik po analitike obucheniya. Obshchestvo issledovaniy v oblasti analitiki obucheniya. 2. Dvivedi, R., Deyv, D., Nayk, KH., Singkhal, S., Omer, R., Patel', P., ... i Randzhan, R. (2023). Ob"yasnimyy II (XAI): osnovnyye idei, metody i resheniya. ACM computing surveys, 55(9), 1-33. 3. Syuy, F., Ushkoreyt, KH., Du, YU., Fan', V., Chzhao, D., i

 ISSN 2306-5478

Chzhu, Dzh. (2019, sentyabr'). Ob"yasnimyy II: kratkiy obzor istorii, oblastey issledovaniy, podkhodov i problem. V sbornike trudov mezhdunarodnoy konferentsii CCF po obrabotke yestestvennogo yazyka i kitayskim vychisleniyam (str. 563-574). Cham: Springer International Publishing. 4. Salikh, A. M., Raisi-Estabrag, Z., Galatstso, I. B., Radeva, P., Petersen, S. Ye., Lekadir, K., i Menegaz, G. (2025). Perspektiva metodov ob"yasnimogo iskusstvennogo intellekta: SHAP i LIME. Advanced Intelligent Systems, 7(1), 2400304. 5. Nguyen, KH. T. T., Tsao, KH. K., Nguyen, K. V. T., i Fam, N. D. K. (2021, may). Otsenka ob"yasnimogo iskusstvennogo intellekta: SHAP, LIME i CAM. V sbornike trudov konferentsii FPT AI (str. 1-6). 6. Chzhan, K., Syuy, P., i Chzhan, Dzh. (2020, oktyabr'). Ob"yasnimyy II v modelyakh glubokogo obucheniya s podkrepleniyem: metod SHAP, primenyayemyy v sistemakh avariynogo upravleniya energosistemoy. V 2020 godu sostoitsya 4-ya konferentsiya IEEE po energeticheskomu Internetu i integratsii energeticheskikh sistem (EI2) (str. 711-716). IEEE. 7. Patel' A. (may 2025 g.). Peresmotr vernosti v ob"yasnimom iskusstvennom intellekte: raspakovka kognitivnykh iskazheniy i obmanchivoy prozrachnosti v interpretatsiyakh modeley. V 2025 g. sostoitsya 5-ya konferentsiya po intellektual'noy kiberbezopasnosti (ICSC) (str. 298-302). IEEE. 8. Pilipenko V. (2025, sentyabr'). Prognozirovaniye vysokogo urovnya akademicheskoy uspevayemosti studentov s razlichnymi mashinnymi obucheniyami. Nauka i tekhnika segodnya, 8(45), 1634–1649. https://doi.org/10.52058/2786-6025-2025-8(49)-1634-1649 9. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion, 58, 82-115. 10. HR, M. S., NR, M. N., & MU, M. D. (2026). NUMERICAL PYTHON: SCIENTIFIC COMPUTING AND DATA SCIENCE APPLICATIONS WITH NUMPY, SCIPY AND MATPLOTLIB. Vaagai international publishing house.

Zlotenko B. M. [1; ORCID ID: 0000-0002-0870-8535],
Doctor of Technical Sciences, Professor,
Skidan V. V. [1; ORCID ID: 0000-0002-8358-9759],
Candidate of Engineering, Associate Professor
Mytelska O. V. [2; ORCID ID: 0009-0004-4147-0866],
Candidate of Engineering, Associate Professor,
Aftandilyants V.E. [2; ORCID ID: 0000-0003-0660-1395],
Candidate of Economic Sciences, Associate Professor,

[1] Kyiv National University of Technologies and Design

THE IMPACT OF DATA DISTRIBUTION ON THE ACCURACY OF MACHINE LEARNING MODEL EXPLANATIONS IN EDUCATIONAL ANALYTICS

The article investigates the impact of statistical data distribution on the correctness of machine learning model explanations in the context of educational analytics. The methods of explanatory artificial intelligence are considered local (LIME, SHAP), gradient attributions and counterfactual approaches from the point of view of their dependence on the properties of the empirical distribution of the training and test samples. It is shown that the quality of explanations is determined not only by the model architecture and the choice of the algorithm ϕ(), but also by class imbalance, covariate bias (the difference between P_train(x) and P_test(x)) and full domain bias when the joint distribution P(x, y) changes. Such effects lead to instability of attribution rankings, increased infidelity and to a situation where the explanations are formally consistent with the trained function f, but are subjectively misleading for the teacher or administrator of the educational institution. A generalized multicriteria approach to evaluating explanations is proposed, combining model fit metrics (fidelity, deletion/insertion AUC), local infidelity, stability of explanations relative to small perturbations of the input vector, and counterfactual consistency (closeness, sparsity, validity). A formalization of key quantities and a summary table of metric groups with an indication of sensitivity to data distribution are presented. For illustration, an experimental formulation for the problem of predicting the risk of academic failure on tabular features is described (illustrative sample size 520/129 observations, d = 12 features) and four figures are presented: the effect of class imbalance and covariate shift on explanation metrics, a comparison of five model architectures, and deletion procedure curves.

Keywords: artificial intelligence, educational data, model interpretation, class imbalance, machine learning, classification, class imbalance, explainable AI.

Отримано: 21 січня 2026 року
Прорецензовано: 28 лютого 2026 року
Прийнято до друку: 27 березня 2026 року