Digital technologies have made knowledge of engineering graphics more open. Leading companies and educational institutions develop and support open access to educational materials, interactive textbooks, webinars, online courses, and organize competitions among specialists. Dialogue between participants in the process of creating and using a product contributes to the professional growth and improvement of the qualifications of engineers and students, on the one hand, and to the improvement of software products, on the other.

Keywords: *graphic training, computer graphics, engineering graphics, free software, 2D graphics, selection criteria, Auto Desk.*

DOI: https://doi.org/10.31392/NZ-udu-162.2025.03
УДК 378.091.26:519.67+004.77

Нікіфоров Р. О., Ткаченко Л. А.

МОДЕЛЮВАННЯ НАВЧАЛЬНИХ РЕЗУЛЬТАТІВ СТУДЕНТІВ ІЗ ВИКОРИСТАННЯМ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ

Прогнозна аналітика передбачає застосування методів статистики та алгоритмів машинного навчання з метою передбачення майбутніх результатів і показників ефективності. До її інструментарію належать, зокрема, інтелектуальний аналіз даних і моделі прогнозування, які надають можливість оцінити ймовірність майбутніх подій і сприяють прийняттю обґрунтованих рішень.

У статті розглянуто проблему прогнозування академічної успішності студентів із застосуванням алгоритмів машинного навчання. Аналіз та апробація відповідних методів стали важливим етапом у напрямі вдосконалення освітньої аналітики та підвищення якості навчального процесу. З'ясовано основні етапи, які проходять дані під час створення, навчання та впровадження моделі прогнозування, зокрема збір даних, попередня обробка даних, вибір моделі машинного навчання, навчання моделі, оптимізація параметрів, застосування попередньо навченої моделі до нових даних. В межах дослідження було розглянуто і протестовано декілька моделей машинного навчання з метою визначення їх ефективності у задачах передбачення результатів навчання.

Було здійснено оцінку основних метрик продуктивності моделей, що дало змогу провести якісний аналіз їхньої точності. Результати матриці помилок свідчать про задовільну роботу моделей після здійснення оптимізації їхніх гіперпараметрів. Здебільшого моделі продемонстрували високий рівень точності класифікації. Окрім цього, було продемонстровано практичне використання побудованих моделей для аналізу нових даних.

Загалом реалізовано ефективне рішення задачі прогнозного моделювання з використанням алгоритмів ***RandomForest*** *та* ***XGBoost****, яке може бути адаптоване для подальшого вдосконалення та реального впровадження у практику. З освітнього погляду, застосування таких моделей дозволяє завчасно ідентифікувати студентів із потенційно низькою успішністю.*

Ключові слова: *прогнозна аналітика, моделювання, продуктивність, машинне навчання, навчальний процес.*

Прогнозна аналітика передбачає застосування методів статистики та алгоритмів машинного навчання з метою передбачення майбутніх результатів і показників ефективності. До її інструментарію належать, зокрема,

інтелектуальний аналіз даних і моделі прогнозування, які дозволяють оцінити ймовірність майбутніх подій і сприяють прийняттю обґрунтованих рішень.

Завдяки використанню історичних даних прогнозна аналітика дає змогу виявляти майбутні тренди й закономірності. Вона аналізує структуру даних, виявляє патерни та встановлює кореляції між різними змінними. Такий підхід дозволяє мінімізувати ризики у бізнесі та знижувати витрати за рахунок прогнозування майбутніх значень ключових показників. Одним із перспективних напрямів застосування прогнозної аналітики є використання машинного навчання (ML) для передбачення навчальних досягнень студентів. Впровадження ML у сферу освіти відкриває нові горизонти для обробки великих масивів освітніх даних, розкриття складних взаємозв'язків і створення персоналізованих моделей навчання.

Сучасні електронні освітні платформи й системи управління навчальним процесом акумулюють значні обсяги даних щодо навчальної діяльності студентів: їхніх оцінок, відвідуваності, результатів іспитів тощо. Поєднання таких даних з алгоритмами машинного навчання дозволяє виявляти значущі закономірності та здійснювати достовірне прогнозування академічних результатів студентів у майбутньому.

У сучасних умовах аналіз даних набуває ключового значення в діяльності практично всіх галузей економіки та науки [6; 7; 8; 9; 10; 11; 12]. Відсутність аналізу накопичених даних у роботі підприємства свідчить про неефективну та недостатньо професійну стратегію ведення бізнесу, особливо у порівнянні з конкурентами, які активно впроваджують аналітичні підходи. Застосування аналітики надає підприємству змогу своєчасно реагувати на зміни, впроваджувати проактивні рішення та формувати конкурентні переваги у своїй сфері.

Аналіз даних охоплює послідовне використання статистичних і логічних методів для опису їхнього обсягу, структурної організації, стислого подання, а також представлення у вигляді графіків, таблиць і зображень. Окрім цього, аналіз охоплює оцінку статистичних трендів, ймовірнісних характеристик і формулювання обґрунтованих висновків. Найбільш активно ці інструменти застосовують комерційні організації, які реалізують свої проєкти на базі сховищ даних [3]. Водночас забезпечення достовірності та цілісності даних виступає однією з ключових умов ефективного аналізу.

Одним із поширених методів інтелектуального аналізу даних є регресійний аналіз. Цей підхід ґрунтується на моделюванні взаємозв'язку між залежною змінною та однією чи кількома незалежними змінними. Існують різні типи регресійних моделей: лінійні, множинні, логістичні, нелінійні, а також моделі, що працюють із життєвими даними [10]. Особливу увагу в таких дослідженнях зазвичай приділяють тому, як зміна незалежних змінних впливає на залежну – тобто визначенню середнього ефекту кожної змінної на результат.

Наприклад, якщо навчальні досягнення з математики у вибірці з 500 студентів аналізуються з урахуванням рівня їх мотивації, то коефіцієнт

регресії показує, наскільки в середньому змінюються бали з математики при зміні мотивації на одну одиницю. У такій моделі всі студенти вважаються частиною однорідної групи щодо впливу мотиваційного чинника на їхні академічні результати, а отже, передбачається подібність характеристик вибірки.

Інтерес до дослідження аналітики навчального процесу, що пов'язаний із залученням студентів, помітно зріс останнім часом. Це суттєво розширило спектр наукових досліджень у галузі освіти. Заклади вищої освіти дедалі більше демонструють зацікавленість у впровадженні аналітичних інструментів для підтримки рівня залученості студентів до навчального процесу. Аналітика в цьому контексті може виконувати функцію інструмента, який сприяє посередництву в інформаційному обміні між студентами та викладачами, що, у свою чергу, веде до підвищення ефективності освітнього процесу, формування усвідомленого ставлення до навчання та знаходження шляхів вирішення складних ситуацій, які виникають у процесі навчання.

Експериментальне дослідження по суті являє собою вивчення однієї або декількох змінних (залежних змінних), на які здійснюється вплив з метою оцінки ефекту дії однієї або кількох інших змінних, відомих як незалежні змінні. Такий тип дослідження базується на встановленні причинно-наслідкових зв'язків у межах обраного об'єкта, що дозволяє зробити висновки про можливі взаємозв'язки, які може створювати конкретний продукт, наукова теорія або концепція [9]. Взаємозв'язки між змінними в межах експерименту встановлюються шляхом ретельної та систематичної маніпуляції змінними. Цей метод особливо доречний у випадках, коли дослідницька робота спрямована на перевірку певної теорії або оцінку ефективності методів.

Крім того, експериментальні установки та протоколи досліджень можуть бути відтворені в інших дослідницьких умовах за допомогою аналогічних змінних. Це дозволяє підтвердити достовірність одержаних результатів щодо продуктів, концепцій і гіпотез [6]. До того ж експериментальний підхід здатен забезпечити чітко структурований набір процедур, необхідних для здійснення оцінки та подальшого звітування результатів дослідження.

Метою роботи є аналіз моделей, методів прогнозного аналізу для аналітики успішності студентів.

Машинне навчання має власний життєвий цикл – послідовність етапів, які проходять дані під час створення, навчання та впровадження моделі прогнозування. На відміну від класичного циклу розробки програмного забезпечення, побудова моделей машинного навчання передбачає активну фазу експериментування з різними наборами даних, особливо при використанні нових (актуальних) даних після початкового навчання моделі. Нижче розглянемо основні етапи цього процесу.

1. Збір даних. Перший і надзвичайно важливий етап – збирання якісних, надійних і репрезентативних даних. Модель може виявити закономірності лише за наявності достатньої кількості достовірної інформації. Від якості вхідних даних безпосередньо залежить точність і надійність прогнозів. Неточні,

застарілі або нерелевантні дані можуть спричинити хибні результати. Тому важливо використовувати перевірені джерела, слідкувати за актуальністю, мінімізувати пропуски, дублікати, а також забезпечити повне представлення всіх підкатегорій чи класів.

2. Попередня обробка даних. Після збирання даних необхідно підготувати їх до аналізу та навчання моделі. Цей етап включає кілька процедур:

Об'єднання та випадкове упорядкування. Всі наявні дані об'єднуються в єдину структуру та перемішуються, щоб зменшити вплив їх початкового порядку.

Очищення даних. Проводиться видалення зайвих або недоречних елементів, заповнення або усунення пропущених значень, усунення дублікатів, а також перетворення форматів відповідно до потреб моделі. Іноді потрібна реструктуризація таблиці (зміна розташування рядків, стовпців, індексів).

Візуалізація. Для кращого розуміння структури даних доцільно побудувати графіки, діаграми, які допоможуть виявити залежності між змінними та оцінити збалансованість класів.

Розділення вибірки. Після очищення дані поділяють на навчальну вибірку (training set) – для побудови моделі, і тестову (test set) – для перевірки її ефективності.

3. Вибір моделі машинного навчання. На цьому етапі обирається модель, здатна забезпечити необхідні результати під час запуску алгоритму на підготовлених даних. Важливо, щоб модель відповідала характеру завдання. Існує багато типів моделей, кожна з яких краще підходить для певних задач – наприклад, розпізнавання мовлення, аналіз зображень або прогнозування числових значень. Також слід враховувати тип даних: числові чи категоріальні, оскільки це впливає на вибір моделі.

4. Навчання моделі. Це центральний етап, під час якого модель «вчиться» на основі підготовлених даних. Вона виявляє закономірності та зв'язки між змінними, що дозволяє їй робити прогнози. У процесі навчання модель поступово вдосконалюється і набуває здатності узагальнювати нову інформацію, реагуючи на раніше невідомі вхідні дані.

5. Оптимізація параметрів. Після навчання оцінюється якість моделі та визначається, чи можна покращити її роботу. Це відбувається шляхом налаштування параметрів – змінних, що впливають на продуктивність. Для кожної моделі існують оптимальні комбінації параметрів, які забезпечують найкращі результати. Мета – знайти ті значення, що забезпечать максимальну точність і ефективність у межах поставленого завдання.

6. Застосування попередньо навченої моделі до нових даних. Одним із завершальних етапів життєвого циклу машинного навчання є використання попередньо навченої моделі для обробки нових вхідних даних. Це дозволяє оцінити, наскільки ефективно модель узагальнює набуті знання та застосовує їх до ситуацій, з якими вона не стикалась під час початкового навчання.

Особливо важливою є оцінка якості прогнозної моделі, коли йдеться про передбачення успішності студентів. Для цього проводиться кількісний аналіз результатів, отриманих у процесі прогнозування. Найбільш поширеними метриками оцінювання ефективності алгоритмів машинного навчання є:

Точність (precision) – показник, що визначає частку істинно позитивних результатів серед усіх передбачень, які модель класифікувала як позитивні. Інакше кажучи, це співвідношення між правильно класифікованими позитивними випадками та загальною кількістю позитивних передбачень.

Правильність (accuracy) – метрика, яка демонструє частку всіх правильних прогнозів (як позитивних, так і негативних) серед загальної кількості передбачень. Це один з найпоширеніших показників, особливо для задач класифікації.

Чутливість (recall або sensitivity) – відображає здатність моделі виявляти всі позитивні випадки серед тих, які насправді належать до позитивного класу. Ця метрика особливо важлива, коли необхідно мінімізувати кількість пропущених істинно позитивних результатів.

На відміну від загальних системних метрик, індивідуалізована аналітика навчальних досягнень орієнтована не лише на ефективність освітньої системи в цілому, а й на результати конкретного здобувача освіти. Такий підхід дозволяє виявити персоналізовані причини недостатньої успішності та вчасно вжити необхідних заходів щодо її підвищення.

Завдяки прогнозній аналітиці освітні менеджери можуть оперативно відстежувати динаміку прогресу студентів у процесі проходження курсів. Такий інструмент дає змогу своєчасно виявляти слабкі місця в засвоєнні навчального матеріалу, оцінювати ефективність освітнього процесу, а також прогнозувати, наскільки здобувач освіти буде здатен застосовувати набуті знання у практичній діяльності. Це відкриває широкі можливості для вдосконалення навчання через обґрунтоване й вчасне прийняття управлінських рішень.

Ефективність використання методів машинного навчання значною мірою залежить від особливостей вхідного набору даних і характеристик обраного алгоритму. Вибір оптимального алгоритму для конкретної сфери – завдання непросте, оскільки кожен із них має власне призначення та функціональні особливості. Навіть моделі одного типу можуть демонструвати різні результати залежно від структури й обсягу вхідних даних [7].

У науковій спільноті розроблено велику кількість методів машинного навчання. Розглянемо найбільш відомі й широко описані у фаховій літературі підходи.

Одним із таких методів є **логістична регресія** [1], або логіт-регресія. Вона застосовується для побудови моделей, що дозволяють оцінювати ймовірність настання певної події. У класичному випадку – це бінарна логістична регресія, де цільова (залежна) змінна набуває лише двох значень: «0» або «1». Незалежні змінні можуть бути як дискретними, так і неперервними. Завдяки використанню логістичної функції, модель відображає ймовірність позитивного результату у межах від 0 до 1 [2].

У логістичній регресії логарифм відношення ймовірності позитивного результату до ймовірності негативного представлено як лінійна функція незалежних змінних. Величина, що відображає це логарифмічне співвідношення, називається **логітом** (logit) – звідси й назва моделі.

Математичне представлення логістичної регресії виглядає так:

P(x) – ймовірність того, що цільова змінна *Y* набуде значення «1» при векторі ознак *x*;

β_0 – константа (вільний член), що визначає початкове зміщення;

$\beta_1, \beta_2, ..., \beta_n$ – коефіцієнти, що відповідають незалежним змінним.

Основною метою логістичної регресії є встановлення зв'язку між вхідними характеристиками об'єкта та ймовірністю його належності до певного класу, що дозволяє ефективно класифікувати нові спостереження.

Іншим популярним підходом є **дерево рішень** – модель із ієрархічною структурою, де вузли представляють ознаки (атрибути), гілки – варіанти рішень або правила, а листові вузли – результати класифікації чи прогнозу. Дерева рішень застосовуються як до дискретних, так і до неперервних даних [9].

Побудова дерева починається з **кореневого вузла**, після чого відбувається послідовне розгалуження: кожен вузол поділяється на підвузли за логікою умов типу «якщо-то». У результаті формується структура, в якій кожен шлях від кореня до листа відповідає певному сценарію класифікації або передбачення. Така наочна й інтерпретована модель є зручною для практичного застосування в освітній аналітиці.

Один із найефективніших методів ансамблевого навчання – **випадковий ліс (Random Forest)**. Цей підхід часто вважається оптимальним для широкого кола задач, оскільки поєднує декілька слабких моделей (дерев рішень) у єдину потужну систему класифікації або регресії. Ансамблеві методи, зокрема випадкові ліси, демонструють високу результативність завдяки здатності зменшувати ризики перенавчання та підвищувати точність передбачень. Вони працюють шляхом об'єднання результатів кількох незалежно побудованих моделей: кожне дерево у складі лісу формує власне рішення (голос), після чого остаточний результат визначається більшістю голосів. Таким чином досягається колективне рішення, що забезпечує підвищену стійкість до шуму в даних і узагальненість моделі. Зазвичай цей процес візуалізується у вигляді схематичного зображення, де продемонстровано процес голосування дерев (див. рис. 1).

Підсилення градієнта (Gradient Boosting) – потужний метод машинного навчання для регресії та класифікації, який створює ансамбль слабких моделей (зазвичай дерев рішень), навчаючи їх послідовно з урахуванням помилок попередніх. Якщо базовими моделями є дерева рішень, алгоритм називають Gradient Boosted Decision Trees (GBDT). Цей підхід часто дає кращу точність, ніж випадкові ліси.

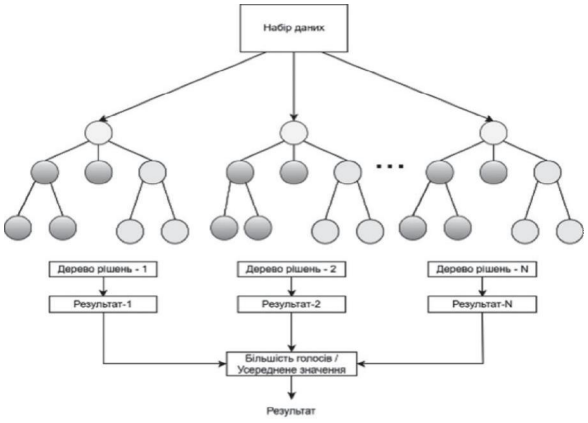

Рис. 1. Приклад випадкового лісу з урахуванням кількох результатів дерева рішень

Особливу увагу заслуговує алгоритм **XGBoost (Extreme Gradient Boosting)**, який завдяки оптимізаціям у паралельній обробці та використанні кешу значно прискорює навчання (приблизно у 10 разів порівняно зі звичайним градієнтним підсиленням). XGBoost також включає регуляризацію для уникнення перенавчання і ефективний алгоритм пошуку розбиття в деревах. На відміну від простого агрегування, він навчає кожне дерево на залишках помилок попередніх.

Для реалізації прогнозної аналітики часто використовують **Python** через широкий спектр бібліотек для обробки даних, моделювання та візуалізації [5], [12]. Рекомендується працювати в середовищі **Google Colaboratory**, яке не вимагає локального встановлення та надає обчислювальні ресурси.

Першим кроком у побудові моделі є аналіз і підготовка даних. Python має корисні функції:

- info() – загальна інформація про типи даних та пам’ять;
- shape – розмірність датафрейму;
- describe() – базова статистика числових змінних;
- corr() – кореляція між змінними.

Підготовка даних включає:

1. Аналіз частоти появи значень атрибутів (рис. 2).
2. Кодування категоріальних змінних у числову форму (індикаторні змінні), оскільки більшість алгоритмів працюють з числовими даними.

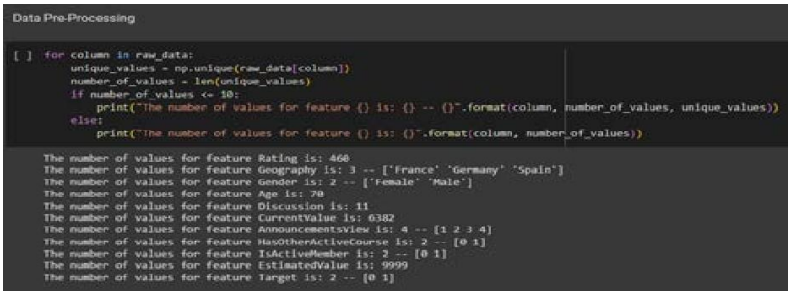

Рис. 2. Огляд властивостей data сету

На етапі попередньої обробки даних важливим є масштабування числових ознак, що покращує ефективність моделей машинного навчання, прискорюючи обчислення та підвищуючи точність прогнозів. Один із поширених методів – MinMaxScaler, який нормалізує значення до інтервалу [1] (рис. 3), забезпечуючи однаковий масштаб ознак. Це важливо для моделей, чутливих до різних діапазонів значень.

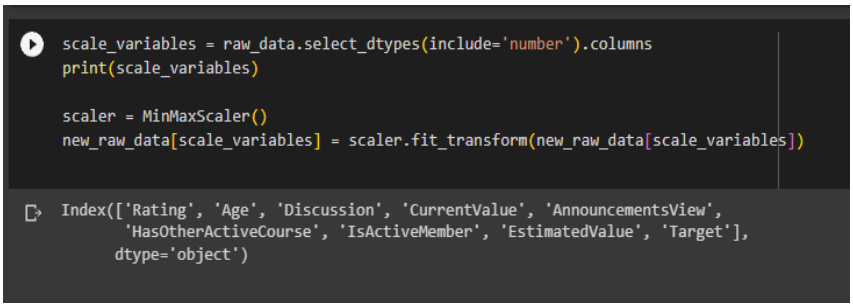

Рис. 3. Масштабування числових даних

Існують дві основні стратегії масштабування:

• Нормалізація перетворює змінні так, щоб їхні значення були в межах від 0 до 1. Підходить для моделей, що базуються на відстанях (наприклад, k-ближчих сусідів).

• Стандартизація змінює розподіл змінних так, що середнє стає 0, а стандартне відхилення – 1. Використовується для змінних із приблизно нормальним розподілом.

Наступний крок – розподіл даних на навчальну та тестову вибірки за методикою hold-out. Зазвичай 90 % даних ідуть на навчання, а 10 % – для перевірки моделі (рис. 4). Це дозволяє об'єктивно оцінити здатність моделі працювати з новими даними.

Рис. 4. Розбиття даних

Після масштабування і розподілу модель готова до навчання. На прикладі дерева рішень (рис. 5) видно, що така модель не лише прогнозує, а й допомагає визначити найінформативніші змінні. Це особливо корисно для аналітики, адже дає змогу зрозуміти, які фактори найбільше впливають на цільову змінну, що сприяє покращенню моделі.

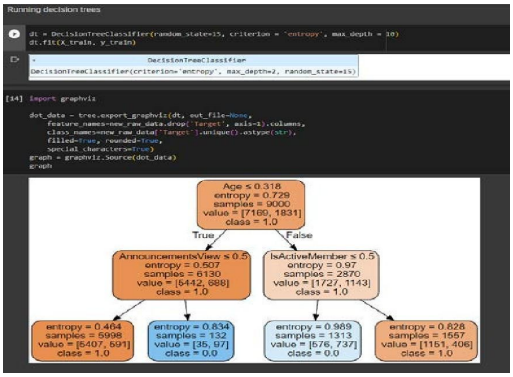

Рис. 5. Побудова дерева рішень

Дерево рішень – одна з найпростіших для інтерпретації моделей машинного навчання, оскільки результати подаються у вигляді деревовидної структури, яку легко перетворити на набір логічних правил. Це спрощує аналіз і розуміння механізмів прийняття рішень.

Головна перевага DecisionTreeClassifier – здатність використовувати різні підмножини ознак та адаптивні правила на кожному рівні дерева. Загальна структура включає:

• кореневий вузол – початок розгалуження;

• внутрішні вузли – конкретні ознаки;

• гілки – умови значень ознак;

• листові вузли – класи, які присвоюються зразкам.

Кожна ознака має ненульове значення важливості і відіграє роль у класифікації, тому видаляти ознаки з набору не потрібно, що зберігає повноту інформації (рис. 5).

Для підвищення надійності та точності класифікації була застосована модель RandomForest, що реалізує ансамблевий підхід, об'єднуючи декілька дерев рішень. Модель тренувалася на розподілених за методикою «утримування» даних (рис. 6).

Рис. 6. Використання моделі RandomForest на розбитих даних за допомогою методики утримування

Для оцінки ефективності моделі використано матрицю помилок, яка дозволяє аналізувати точність класифікації за двома класами:

0 – студент не зміг завершити курс успішно;

1 – студент успішно завершив курс. Отримані результати класифікації демонструють високу чутливість моделі до негативного класу (коефіцієнт коректної ідентифікації студентів, які не завершать курс, становить 0.96), при цьому точність для позитивного класу (успішне завершення) є помірною (0.47). Цей дисбаланс свідчить про потребу в подальшій оптимізації моделі, зокрема через налаштування гіперпараметрів

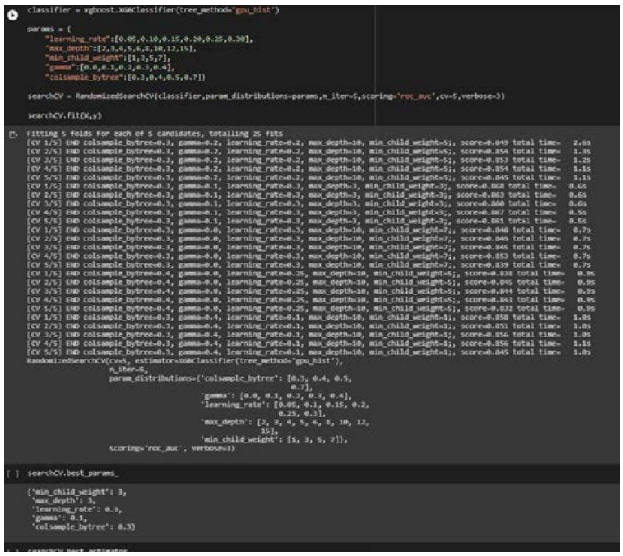

Рис. 7. Підбір параметрів моделі XGBoost

Для досягнення більш високої точності моделі застосовано **алгоритм XGBoost** (Extreme Gradient Boosting), який поєднує швидкість та регуляризацію. Було здійснено підбір гіперпараметрів на основі навчального набору даних, що дозволило суттєво покращити точність класифікації (рис. 8). Такий підхід є ефективним для забезпечення високої якості узагальнення моделей при роботі з новими, раніше не баченими даними.

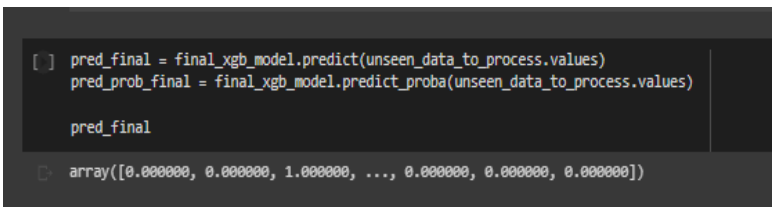

Рис. 8. Попередня обробка нових даних та використання фінальної моделі з обраними параметрами для прогнозування

На завершальному етапі модель XGBoost була протестована на **абсолютно новому наборі даних**, що дозволило не лише класифікувати

результат (чи завершить студент курс), а й оцінити **ймовірність цього завершення** для кожного індивіда.

Ключовим показником ефективності побудованих моделей виступає **точність класифікації (accuracy)**, яка дозволяє оцінити загальну кількість правильних передбачень у співвідношенні до загальної кількості спостережень. Порівняльний аналіз результатів моделей показав, що найвищу точність було досягнуто саме при використанні XGBoost з оптимальними параметрами. Такий результат свідчить про доцільність застосування сучасних бустингових алгоритмів у прогнозній аналітиці в освіті.

Для кількісної оцінки якості побудованих моделей використовувалися такі **класичні метрики машинного навчання**:

Точність (Accuracy) – загальний показник правильності класифікації, який відображає частку правильно передбачених результатів серед усіх випадків. Обчислюється за формулою:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

Однак слід зазначити, що у випадках **незбалансованих даних**, де кількість представників одного класу суттєво переважає інший, ця метрика може бути оманливо високою.

Прецизія (Precision) – показник, який характеризує точність позитивних передбачень, тобто частку правильно класифікованих позитивних випадків серед усіх передбачених як позитивні:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Повнота (Recall або Sensitivity) – чутливість моделі до позитивного класу, тобто частка істинно позитивних випадків, що були правильно передбачені як позитивні:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Помилка класифікації (Classification Error) – частка неправильно класифікованих об'єктів. Цей показник визначає, наскільки часто модель робить помилки у передбаченні:

$$\text{Error} = 1 - \text{Accuracy}$$

F1-міра (F1-score) – гармонічне середнє між precision та recall, що

дозволяє врахувати як точність, так і повноту одночасно. Вона є особливо важливою при **нерівномірному розподілі класів**:

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

Остаточне застосування моделі на **новому незалежному наборі даних** дозволило перевірити її узагальнювальну здатність і визначити ймовірність успішного завершення курсу кожним окремим студентом. Такий підхід є перспективним для впровадження в адаптивні освітні системи, де необхідно своєчасно виявляти ризики академічної неуспішності.

Для глибшого розуміння ефективності роботи обраних моделей доцільно проаналізувати **матрицю помилок**, яка забезпечує детальнішу характеристику точності класифікаційного прогнозу. Такий аналіз дозволяє оцінити продуктивність алгоритмів після проведення класифікації. У межах даного дослідження було сформовано матриці помилок після оптимізації параметрів моделей. Важливо зазначити, що значення **1** позначає успішне завершення студентом курсу, а **0** – його незавершення.

Матриця помилок (confusion matrix), також відома як **матриця плутанини**, є табличною формою представлення результатів класифікації, яка узагальнює кількість коректних і помилкових передбачень, здійснених моделлю. Цей інструмент є особливо корисним у випадках, коли алгоритм класифікує спостереження на два класи.

У структурі бінарної класифікації, елементи матриці помилок поділяються на такі складові:

True Positive (TP) – випадки, коли модель правильно передбачила позитивний результат (студент успішно завершив курс).

True Negative (TN) – випадки, коли модель вірно визначила негативний результат (студент не завершив курс).

False Positive (FP) – ситуації, коли модель помилково передбачила позитивний результат, хоча він був негативним (помилка I типу).

False Negative (FN) – ситуації, коли модель неправильно класифікувала позитивний результат як негативний (помилка II типу).

Матриця помилок дозволяє не лише оцінити загальну якість класифікації, а й виявити аспекти, які потребують доопрацювання.

Аналіз отриманих результатів свідчить, що після налаштування параметрів модель **XGBoost** продемонструвала вищі показники класифікації, зокрема у передбаченні студентів, які не зможуть завершити курс. Водночас, **RandomForest** також показала досить високі результати, особливо у частині розпізнавання негативного класу.

Таким чином, хоча XGBoost забезпечує кращу загальну точність прогнозування, **використання RandomForest не слід виключати**. Вона може

бути ефективною у контекстах, де важливо мінімізувати хибно-позитивні або хибно-негативні прогнози. Матриця помилок у цьому випадку виступає важливим інструментом для прийняття рішень щодо доцільності застосування тієї чи іншої моделі в конкретному завданні освітньої аналітики.

а) б)

Рис. 9. Матриця помилок для а) RandomForest; б) XGBoost після покращення параметрів

Отримані дані, в свою чергу, викладачі можуть використовувати для роботи зі студентами опираючись на результати роботи моделі, наприклад після класифікаційної обробки було отримано відповідь на те, чи завершить студент курс та з якою ймовірністю це відбудеться (рис. 10).

Predictions - Succeed or Not	Predictions - Probability to Succeed education	Predictions - Succeed or Fail Desc
1.0	0.56	Succeed
0.0	0.27	Fail
1.0	0.96	Succeed
0.0	0.02	Fail
0.0	0.10	Fail

Рис. 10. Результати прогнозного аналізу використовуючи модель з обраними попередньо параметрами

Висновки. У межах даної роботи було окреслено проблему прогнозування академічної успішності студентів із застосуванням алгоритмів машинного навчання. Аналіз та апробація відповідних методів стали важливим етапом у напрямі вдосконалення освітньої аналітики та підвищення якості навчального процесу. В рамках дослідження було розглянуто і протестовано декілька моделей машинного навчання з метою визначення їх ефективності у задачах передбачення результатів навчання.

У процесі дослідження було здійснено оцінку основних метрик продуктивності моделей, що дало змогу провести якісний аналіз їхньої точності. Результати матриці помилок свідчать про задовільну роботу моделей після здійснення оптимізації їхніх гіперпараметрів. Здебільшого, моделі продемонстрували високий рівень точності класифікації. Окрім цього, було продемонстровано практичне використання побудованих моделей для аналізу нових даних.

Загалом реалізовано ефективне рішення задачі прогнозного моделювання з використанням алгоритмів **RandomForest** та **XGBoost**, яке може бути адаптоване для подальшого вдосконалення та реального впровадження у практику. З освітнього погляду, застосування таких моделей дозволяє завчасно ідентифікувати студентів із потенційно низькою успішністю, що, у свою чергу, сприяє своєчасному вжиттю заходів педагогічної підтримки та покращенню результатів навчання.

Використана література:

1. Зубрицький В. В. Огляд методів та технологій штучного інтелекту в електронному навчанні. Сучасні виклики і актуальні проблеми науки, освіти та виробництва: міжгалузеві диспути [зб. наук. пр.] : матеріали XXIV міжнародної науково-практичної інтернет-конференції (м. Київ, 28 січня 2022 р.). Київ, 2022. С. 43-45.
2. Zubrytskyi Vasyl, Vakaliuk Tetiana. Overview of methods of intellectual data analysis. *Тези доповідей II Міжнародної студентської наукової конференції* (Т. 2), м. Одеса, 17 грудня 2021.
3. Chilukuri K. C. A novel framework for active learning in engineering education mapped to course outcomes. *Procedia Computer Science*, 172, 2020. P. 280-33.
4. Dewan M. A. A., Murshed M., Lin F. Engagement detection in online learning: a review. *Smart Learning Environments*, 6 (1), 2019.
5. Python Documentation. URL : https://www.python.org/doc/.
6. Ko C. Y., Leu F.-Y. Examining successful attributes for undergraduate students by applying machine learning techniques. *IEEE Transactions on Education*, 64 (1), 2021. P. 50-57.
7. Krawczyk B., Minku L. L., Gama J., Stefanowski J., Woźniak M. Ensemble learning for data stream analysis: A survey. *Information Fusion*, 37, 2017. P. 132-156.
8. Liu Z., Yang C., Rüdian S., Liu S., Zhao L., Wang T. Temporal emotion-aspect modeling for discovering what students are concerned about in online course forums. *Interactive Learning Environments,* 27(5-6), 2019. P. 598-627. URL : https://www.tandfonline.com/doi/full/10.1080/10494820.2019.1610449.
9. Liu Z., Zhang N., Peng X., Liu S., Yang Z., Peng J., Su Z., Chen J. Exploring the relationship between social interaction, cognitive processing and learning achievements in a MOOC discussion forum. *Journal of Educational Computing Research,* 60 (1), 2022. P. 132-169. URL : https://journals.sagepub.com/doi/10.1177/07356331211027300.
10. Moscoso-Zea O., Paredes-Gualtor J., Lujan-Mora S. A holistic view of data warehousing in education. IEEE Access, 6, 2018. P. 64659-64673.
11. Moscoso-Zea O., Lujan-Mora S. Knowledge management in higher education institutions for the generation of organizational knowledge. In 2017 12th Iberian Conference on Information Systems and Technologies (CISTI). 2017.
12. Russell R. Machine learning step-by-step guide to implement machine learning algorithms with Python. Editorial : Columbia, Sc. 2018.
13. Shahiri A. M., Husain W., Rashid N. A. A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, 72, 2015. P. 414-422.

Reference:

1. Zubrytskyi V. V. (2022). Ohliad metodiv ta tekhnolohii shtuchnoho intelektu v elektronnomu navchanni. [Overview of artificial intelligence methods and technologies in e-learning]. *Suchasni vyklyky i aktualni problemy nauky, osvity ta vyrobnytstva: mizhhaluzevi dysputy* [zb. nauk. pr.] : materialy XXIV mizhnarodnoi naukovo-praktychnoi internet-konferentsii (m. Kyiv, 28 sichnia 2022 r.). Kyiv. S. 43-45 [in Ukranian].
2. Zubrytskyi Vasyl, Vakaliuk Tetiana. (2021). Overview of methods of intellectual data analysis. *Tezy dopovidei II Mizhnarodnoi studentskoi naukovoi konferentsii* (T. 2), m. Odesa, 17 hrudnia [in English].
3. Chilukuri K. C. (2020). A novel framework for active learning in engineering education mapped to course outcomes. *Procedia Computer Science*, 172. P. 28-33 [in English].
4. Dewan M. A. A., Murshed M., Lin F. (2019). Engagement detection in online learning: a review. *Smart Learning Environments,* 6 (1) [in English].

5. Python Documentation. URL : https://www.python.org/doc/ [in English].
6. Ko C. Y., Leu F.-Y. (2021). Examining successful attributes for undergraduate students by applying machine learning techniques. *IEEE Transactions on Education*, 64 (1). P. 50-57 [in English].
7. Krawczyk B., Minku L. L., Gama J., Stefanowski J., Woźniak M. (2017). Ensemble learning for data stream analysis: A survey. *Information Fusion*, 37. P. 132-156 [in English].
8. Liu Z., Yang C., Rüdian S., Liu S., Zhao L., Wang T. (2019). Temporal emotion-aspect modeling for discovering what students are concerned about in online course forums. Interactive Learning Environments, 27 (5-6). P. 598-627. URL : https://www.tandfonline.com/doi/full/10.1080/10494820.2019.1610449 [in English].
9. Liu Z., Zhang N., Peng X., Liu S., Yang Z., Peng J., Su Z., Chen J. (2022). Exploring the relationship between social interaction, cognitive processing and learning achievements in a MOOC discussion forum. *Journal of Educational Computing Research*, 60 (1). P. 132-169. URL : https://journals.sagepub.com /doi/10.1177/07356331211027300 [in English].
10. Moscoso-Zea O., Paredes-Gualtor J., Lujan-Mora S. (2018). A holistic view of data warehousing in education. IEEE Access, 6. P. 64659-64673 [in English].
11. Moscoso-Zea O., Lujan-Mora S. (2017). Knowledge management in higher education institutions for the generation of organizational knowledge. In 2017 12th Iberian Conference on Information Systems and Technologies (CISTI). [in English].
12. Russell R. (2018). Machine learning step-by-step guide to implement machine learning algorithms with Python. Editorial : Columbia, Sc. [in English].
13. Shahiri A. M., Husain W., Rashid N. A. (2015). A review on predicting student's performance using data mining techniques. Procedia Computer Science, 72. P. 414-422 [in English].

R. Nikiforov, L. Tkachenko. Modeling students' learning results using machine learning algorithms.

Predictive analytics involves the use of statistical methods and machine learning algorithms to predict future outcomes and performance indicators. Its tools include, in particular, data mining and forecasting models, which provide an opportunity to assess the probability of future events and contribute to making informed decisions.

The article considers the problem of predicting students' academic performance using machine learning algorithms. Analysis and testing of relevant methods have become an important stage in improving educational analytics and improving the quality of the educational process. The main stages that data go through during the creation, training and implementation of a forecasting model are clarified, in particular, data collection, data preprocessing, selection of a machine learning model, model training, parameter optimization, application of a pre-trained model to new data. As part of the study, several machine learning models were considered and tested in order to determine their effectiveness in predicting learning outcomes.

The main performance metrics of the models were assessed, which made it possible to conduct a qualitative analysis of their accuracy. The results of the error matrix indicate satisfactory performance of the models after optimization of their hyperparameters. For the most part, the models demonstrated a high level of classification accuracy. In addition, the practical use of the constructed models for analyzing new data was demonstrated.

In general, an effective solution to the problem of predictive modeling using the RandomForest and XGBoost algorithms was implemented, which can be adapted for further improvement and real implementation in practice. From an educational point of view, the use of such models allows for early identification of students with potentially low performance.

Keywords: *predictive analytics, modeling, productivity, machine learning, educational process.*