УДК 004.02+004.04+004.89
DOI https://doi.org/10.32782/2663-5941/2023.5/18

Вакалюк Т.А.
Державний університет «Житомирська політехніка»

Янчук В.М.
Державний університет «Житомирська політехніка»

Морозов Д.С.
Державний університет «Житомирська політехніка»

Зубрицький В.В.
Державний університет «Житомирська політехніка»

Новіцька І.В.
Житомирський державний університет імені Івана Франка

ПРОГНОЗНЕ МОДЕЛЮВАННЯ АНАЛІТИКИ УСПІШНОСТІ СТУДЕНТІВ З ВИКОРИСТАННЯМ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ

Прогнозна аналітика стосується використання алгоритмів машинного навчання та статистики для прогнозування майбутніх результатів і продуктивності. Такі методи, як інтелектуальний аналіз даних і прогнозне моделювання, оцінюють ймовірність майбутніх результатів і сповіщають про майбутні події, щоб допомогти у прийнятті рішення. Використання машинного навчання (ML) для прогнозування успішності навчання студентів має потенціал для вдосконалення освітніх систем. Застосування ML у цій галузі відкриває нові можливості для аналізу великих обсягів даних, виявлення складних залежностей і розробки персоналізованих підходів до навчання. Сучасні системи управління навчанням та електронні платформи зберігають велику кількість інформації про навчальні досягнення студентів, такі як оцінки, відвідуваність, здані іспити тощо. Використання цих даних у поєднанні з ML дозволяє ідентифікувати корисні закономірності та прогнозувати майбутні успіхи студентів. В даній роботи було визначено проблему прогнозного моделювання та аналітики успішності студентів з використанням алгоритмів машинного навчання. Дослідження і аналіз алгоритмів машинного навчання для прогнозного моделювання та аналітики успішності студентів є важливим кроком у покращенні освітнього процесу. Відштовхуючись від цього, було проведено дослідження та проаналізовано декілька алгоритмів машинного навчання. Було проведено аналіз метрик та результатів натренованої моделі. З матриці помилок можна побачити, що отримано досить непогані результати після підбору параметрів для моделей. У результаті було отримано реалізацію алгоритму прогнозного аналізу з використанням моделей RandomForest та XGBoost, який підходить для подальшої модернізації та практичному застосуванні. З освітньої точки зору, це дослідження може допомогти викладачам визначити студентів, які перебувають у групі ризику щодо поганої успішності, і може допомогти педагогам вжити своєчасних коригувальних заходів.

Ключові слова: *прогнозне моделювання, аналітика, успішність студентів, машинне навчання, алгоритми.*

Постановка проблеми. Прогнозна аналітика стосується використання алгоритмів машинного навчання та статистики для прогнозування майбутніх результатів і продуктивності. Такі методи, як інтелектуальний аналіз даних і прогнозне моделювання, оцінюють ймовірність майбутніх результатів і сповіщають про майбутні події, щоб допомогти у прийнятті рішення.

Прогнозна аналітика допомагає передбачити майбутні тенденції та закономірності, використовуючи історичні дані. Вона використовує різні шаблони даних і визначає кореляції між змінними. Це допомагає зменшити бізнес-ризики та витрати, прогнозуючи майбутні значення певних змінних. Використання машинного навчання (ML) для прогнозування успішності навчання студентів має потенціал для вдосконалення освітніх систем. Застосування ML у цій галузі відкриває нові можливості для аналізу великих обсягів даних, вияв-

лення складних залежностей і розробки персоналізованих підходів до навчання. Сучасні системи управління навчанням та електронні платформи зберігають велику кількість інформації про навчальні досягнення студентів, такі як оцінки, відвідуваність, здані іспити тощо. Використання цих даних у поєднанні з ML дозволяє нам ідентифікувати корисні закономірності та прогнозувати майбутні успіхи студентів.

Аналіз останніх досліджень і публікацій. Сьогодні аналіз даних перебуває в центрі уваги всіх галузей [6], [7], [8], [9], [10], [11], [12]. Якщо підприємство не використовує аналіз накопичених даних, тоді можна говорити про обраний хибний та некомпетентний підхід у веденні бізнесу на відміну від конкурентів, які займаються даним процесом. Аналіз даних дає змогу підприємству вживати активних заходів і створювати конкурентні переваги у своїй галузі.

Систематичне застосування статистичних і логічних методів для опису обсягу даних, модульності структури даних, стиснення представлення даних, ілюстрації за допомогою зображень, таблиць і графіків, а також оцінки статистичних нахилів, даних про ймовірність і отримання значущих висновків, відомі як аналіз даних. Найбільш зацікавлені у використанні методів та інструментів аналізу даних є комерційні підприємства, що розгортають свої проекти на основі інформаційних сховищ даних [3]. Забезпечення цілісності даних є одним із найважливіших компонентів аналізу даних. Розглянемо деякі методи інтелектуального аналізу даних.

Регресійний аналіз – метод, який працює шляхом моделювання зв'язку між залежною змінною та однією або кількома незалежними змінними. Модель регресії може бути лінійною, множинною, логістичною, нелінійною, життєвими даними тощо [10]. Часто особливий інтерес приділяється оцінці впливу кожної незалежної змінної на залежну змінну, і такий ефект розглядається як середнє значення ефекту для всіх суб'єктів у вибірці. Наприклад, якщо бали з математики 500 студентів регресують за показником їхньої мотивації, значення нахилу або коефіцієнт регресії кількісно визначає середню зміну досягнень з математики для всіх 500 студентів на одну одиницю зміни мотивації. Це означає, що ці 500 студентів розглядаються як одна однорідна група щодо впливу мотивації на досягнення з математики, а неявне припущення полягає в тому, що ці студенти є з однієї групи населення зі схожими характеристиками.

Інтерес до вивчення аналітики навчання, пов'язаної із залученням студентів, останнім часом значно зростає. Це ще більше розширило дослідницьке поле для освіти. Заклади вищої освіти виявили свою зацікавленість у використанні аналітики для підтримки їхньої участі. Це може діяти як інструмент, який допоможе в посередництві обміну інформацією між студентами та викладачами, що призведе до ефективного навчання, підвищення обізнаності та способу вирішення поточних складних ситуацій.

Експериментальне дослідження – це, по суті, дослідження однієї або кількох змінних (залежних змінних), якими маніпулюють для оцінки впливу на одну або більше змінних, відомих як незалежні змінні. Воно базується на причинно-наслідкових зв'язках у вибраному предметі, щоб зробити висновок про різні зв'язки, які може створити продукт, теорія чи ідея [9]. Відношення серед змінних встановлюється за допомогою точної та систематичної маніпуляції. Ця методика підходить, коли в основі дослідження лежить теорія тестування та оцінка методів. Крім того, ту саму настройку та протокол можна відтворити за допомогою тих самих змінних. Це може підтвердити достовірність продуктів, ідей і теорій [6]. Крім того, цей тип наукового підходу може надати набір інструкцій для оцінювання та звітування інформації для дослідження.

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

Виклад основного матеріалу. Машинне навчання має власний життєвий цикл, тобто процес, який проходять дані для розробки та розгортання системи прогнозування. Порівняно з життєвим циклом розробки програмного забезпечення, розробка моделей машинного навчання передбачає експериментування з наборами даних для досягнення цілей і завдань, визначених під час застосування свіжих даних після навчання. Розглянемо кожен з етапів більш детально.

1. **Збір даних.** Як відомо, машини спочатку навчаються на даних, які їм надаються. Надзвичайно важливо збирати надійні дані, щоб модель машинного навчання могла знайти правильні шаблони. Якість даних, які подаються на машину, визначатиме, наскільки точна запропонована модель. Якщо є неправильні або застарілі дані, в результаті отримаємо неправильні результати або прогнози, які не є актуальними. Тому потрібно переконатись, що дані використовуються з надійного джерела, оскільки це безпосередньо вплине на результат моделі. Хороші дані є релевантними,

містять дуже мало пропущених і повторюваних значень і добре представляють різні наявні підкатегорії/класи.

2. **Попередня обробка даних.** Отримавши дані, їх потрібно підготувати. Це можна зробити за допомогою:

a. Об'єднання всіх наявних даних і їх рандомізація. Це допомагає переконатися, що дані розподіляються рівномірно, а порядок не впливає на процес навчання.

b. Очищення даних для видалення непотрібних даних, відсутніх значень, рядків і стовпців, повторюваних значень, перетворення типів даних тощо. Можливо, навіть доведеться реструктурувати набір даних і змінити рядки та стовпці або індекси рядків і стовпців.

c. Візуалізуйте дані, щоб зрозуміти, як вони структуровані, і зрозуміти взаємозв'язок між різними змінними та присутніми класами.

d. Розбиття очищених даних на два набори – набір для навчання та набір для тестування. Навчальний набір – це набір, з якого вчиться модель. Тестовий набір використовується для перевірки точності моделі після навчання.

3. **Вибір моделі машинного навчання.** Модель машинного навчання визначає результати, які отримуються після запуску алгоритму машинного навчання на зібраних даних. Важливо підібрати модель, яка відповідає поставленим завданням. Протягом багатьох років вчені та інженери розробляли різні моделі, які підходять для різних завдань, таких як розпізнавання мовлення, розпізнавання зображень, прогнозування тощо. Окрім цього, також потрібно перевірити, чи підходить модель для числових чи категоріальних даних, і вибрати відповідно.

4. **Навчання моделі** є найважливішим кроком у машинному навчанні. Під час навчання передаються підготовлені дані моделі машинного навчання, щоб знаходити закономірності та робити прогнози. Це призводить до того, що модель навчається на даних, щоб вона могла виконати поставлене завдання. З часом, під час навчання, модель стає краще прогнозувати.

5. **Покращення параметрів моделі.** Після того, як створено та оцінено модель, важливо подивитись, чи можна якимось чином підвищити її точність. Це робиться шляхом налаштування параметрів, наявних у моделі. Параметри – це змінні в моделі, які зазвичай вибирає програміст. При конкретному значенні параметра точність буде максимальною. Налаштування параметрів стосується пошуку цих значень.

6. **Використання попередньо натренованої моделі на абсолютно нових даних.**

Оцінка прогнозної моделі є важливою частиною для визначення точності успішності студента. Для цього важливо кількісно оцінити якість прогнозів системи [3]. Розглянемо деякі важливі показники ефективності для оцінки методів машинного навчання:

1. Точність – це кількість правильних позитивних результатів, поділена на всі зразки, позначені алгоритмом як позитивні.

2. Правильність – визначається як відношення правильних прогнозів до загальної кількості введених зразків. Це часто використовувана метрика для оцінки якості рішень класифікатора. Це найбільш використовувана метрика оцінки як для бінарної, так і для багатокласової класифікації. Це визначальне значення для оцінки можливостей алгоритму.

3. Чутливість – це кількість правильних позитивних результатів, поділена на всі зразки, які алгоритм повинен був позначити як позитивні.

На відміну від будь-яких інших показників, прогнозна аналітика навчання є більш ефективною, оскільки зосереджена на окремому студенті, а не на системі керування навчанням в цілому. Це робить прогнозну аналітику надзвичайно впливовою у вирішенні проблем, які спричиняють неефективне навчання. Це дозволяє керівникам визначати прогрес, досягнутий співробітниками на своїх курсах. Ця технологія допомагає зрозуміти, наскільки добре студенти отримали вигоду від курсів і чи збираються вони застосовувати отримані знання на практиці.

Ефективність рішення машинного навчання залежить від природи набору даних і продуктивності алгоритмів. Вибір правильного алгоритму навчання, який підходить для застосування в конкретній області, є не простою справою. Причина цього полягає в тому, що призначення алгоритмів ML є різними. Навіть результати різних алгоритмів навчання в подібній категорії можуть відрізнятися залежно від характеристик даних [7]. Багато алгоритмів машинного навчання було реалізовано в дослідницькому співтоваристві. Розглянемо найважливіші і відомі методи, які фігурують у науковій літературі.

Логістична регресія [1] (або логіт-регресія) – це оцінка параметрів логістичної моделі (коефіцієнтів у лінійній комбінації). Формально в бінарній логістичній регресії існує одна двійкова залежна змінна, кодована змінною-індикатором, де два значення позначені «0» і «1», тоді як кожна з незалежних змінних може бути двійковою змінною (два

класи, кодовані як індикаторна змінна) або безперервна змінна (будь-яке реальне значення). Відповідна ймовірність значення, позначеного як «1», може коливатися від 0 (звичайно, значення «0») до 1 (безумовно, значення «1»), отже, позначення; [2] функція, яка перетворює логарифмічні шанси на ймовірність, є логістична функція, звідси і назва. Одиниця вимірювання для логарифмічної шкали шансів називається логіт, від logistic unit, звідси альтернативні назви. Формула логістичної регресії, як незалежні змінні впливають на залежну змінну [2]:

$$F1 = 2 * (precision * recall)/(precision + recal),$$

де P – у логістичній моделі p(x) інтерпретується як ймовірність того, що залежна змінна Y дорівнює успіху або підпадає під класифікацію випадку, або не підпадає під класифікацію випадку. Це важливо, оскільки воно показує, що значення виразу лінійної регресії може змінюватися від негативної до позитивної нескінченності, і все ж після перетворення результуючий вираз для ймовірності P(X) коливається між 0 і 1;

B1,β2,…… βn є невідомими параметрами (коефіцієнтами);

β0 є β0 є константою для створення лінії найкращого підходу. Метою логістичної регресії є зіставлення функції з характеристик набору даних на цілі для обчислення ймовірності того, що новий запис належить до одного з цільових класів.

Дерево рішень має деревоподібну структуру, де кожен вузол показує атрибут, кожне посилання показує рішення (правило), а кожен лист показує результат. Його можна використовувати як для безперервних, так і для дискретних наборів даних [9]. Дерево рішень починається з кореневого вузла. З цього вузла користувачі рекурсивно розділяють кожен вузол відповідно до алгоритму навчання дерева рішень на основі запитань «якщо». Результатом є дерево рішень, у якому кожна гілка представляє можливий сценарій рішення та його результат.

Випадковий ліс (RandomForest) – вважається експертним рішенням для більшості проблем і підпадає під класифікатори ансамблевого навчання, за допомогою яких слабкі моделі поєднуються для створення потужної. Ансамблеві методи є одними з найбільш перспективних напрямів дослідження. Він визначається як набір класифікаторів, прогнози яких об'єднуються для прогнозування нових випадків. Ансамблеві алгоритми навчання показали себе як ефективний метод для підвищення точності прогнозування та послаблення складності проблем навчання у підпроблеми. Численні дерева рішень створюються у випадкових лісах. Щоб класифікувати об'єкт, що має атрибути, кожне з дерев дає класифікацію, яка також вважається голосом. Тоді лісу надається можливість вибрати класифікацію з максимальною кількістю голосів. Це показано на рис. 1.

Збільшення градієнта – це техніка машинного навчання, яка використовується, зокрема, у завданнях регресії та класифікації. Вона дає модель прогнозування у формі ансамблю слабких моделей прогнозування, які зазвичай є деревами рішень. Коли дерево рішень є слабким навчальним елементом, отриманий алгоритм називається деревом із посиленням градієнта; зазвичай він перевершує випадковий ліс.

XGBoost є скороченням від «Extreme Gradient Boosting». «Екстремальний» стосується підвищення швидкості, наприклад паралельних обчислень і усвідомлення кешу, що робить XGBoost приблизно в 10 разів швидшим, ніж традиційний Gradient Boosting. Крім того, XGBoost включає унікальний алгоритм пошуку поділу для оптимізації дерев разом із вбудованою регуляризацією, яка зменшує

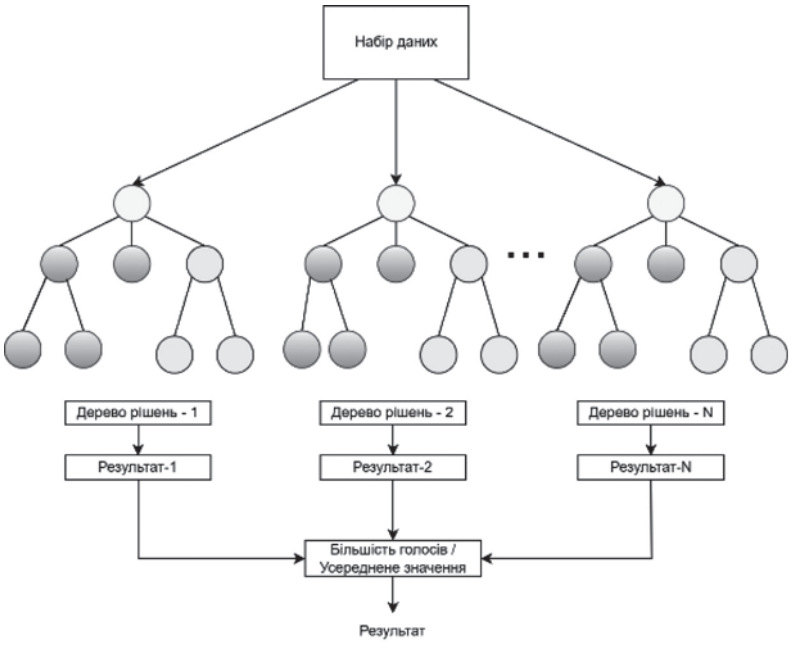


Рис. 1. Приклад випадкового лісу з урахуванням кількох результатів дерева рішень

переобладнання. Загалом, XGBoost – це швидша та точніша версія Gradient Boosting. XGBoost тренується на залишках. Замість агрегування дерев, підсилені градієнтом дерева навчаються на помилках під час кожного раунду посилення.

Для прогнозної аналітики додатку слід використати прогнозне моделювання за допомогою мови програмування Python [5], [12].

Написання коду відбуватиметься в хмарному середовищі розробки Google Colaboratory.

Важливо знати, як правильно орієнтуватися в даних, з якими працюєте, щоб знати, як побудувати прогнозну модель. З цієї причини Python має кілька функцій, які допоможуть у дослідженнях:

- info() – Функція, що показує нам тип даних кожного стовпця, кількість стовпців, використання пам'яті та кількість записів у наборі даних;
- shape – властивість, що відображає кількість записів і стовпців;
- describe() – функція, що узагальнює статистичні властивості набору даних, такі як кількість, середнє, мінімальне та максимальне;
- corr() – функція, що відображає кореляцію між різними змінними в нашому наборі даних Чим ближче до 1, тим сильніша кореляція між цими змінними. Знак мінус означає, що ці 2 змінні негативно корельовані, тобто одна зменшується зі збільшенням іншої і навпаки.

Побудова алгоритму машинного навчання

Почнемо з попередньої обробки даних. На цьому етапі потрібно:

1. Переглянути кількість значень для кожного атрибуту в певному наборі даних (рис. 2).

2. Перетворення категоріальних змінних на фіктивні/індикаторні змінні, тому що описані вище моделі працюють на основі математичних моделей, яким потрібні числові значення.

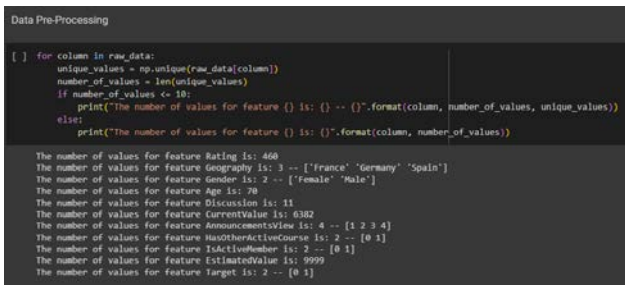


Рис. 2. Огляд властивостей дата сету

3. Масштабування числових властивостей за допомогою MinMaxScaler для підвищення швидкості роботи моделей та їх точності (рис. 3). Дві найпопулярніші техніки для масштабування числових даних перед моделюванням – нормалізація та стандартизація. Нормалізація масштабує кожну вхідну змінну окремо до діапазону 0–1, який є діапазоном для значень з плаваючою комою, де ми маємо найбільшу точність. Стандартизація масштабує кожну вхідну змінну окремо шляхом віднімання середнього (це називається центруванням) і ділення на стандартне відхилення, щоб зсунути розподіл, щоб отримати середнє значення нуль і стандартне відхилення одиниці.

Наступний етап це розбиття даних за допомогою методики «утримування» – це коли ви розділяєте свій набір даних на набір «навчальний» і «тестовий». Навчальний набір – це те, на чому модель навчається, а тестовий набір використовується, щоб побачити, наскільки добре ця модель працює на нових даних. Звичайним поділом під час використання методу очікування є використання 90% даних для навчання, а решта 10% даних для тестування (рис. 4).

Після цього отриманні дані можна використовувати в моделях, на наступному рисунку зображено побудову дерева рішень (рис. 5), яке, в свою чергу, можна використати для знаходження найбільш релевантних властивостей набору даних. Наприклад, якщо стоїть мета покращити результативність лінійної регресії чи отримати показники властивостей, які найбільше впливають на кінцевий результат, DecisionTreeClassifier має повністю справитися з цією задачею.

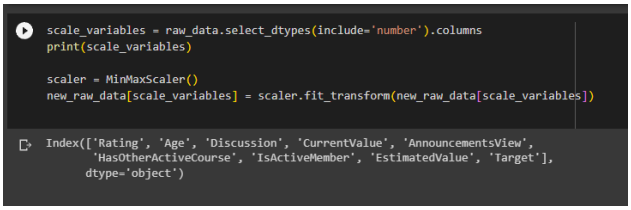


Рис. 3. Масштабування числових даних

Рис. 4. Розбиття даних

Дерево рішень отримує знання у формі дерева, яке також можна переписати як набір окремих правил, щоб полегшити його розуміння. Основною перевагою класифікатора дерева рішень є його здатність використовувати різні підмножини ознак і правила прийняття рішень на різних етапах класифікації. Загальне дерево рішень складається з одного кореневого вузла, ряду внутрішніх і кінцевих вузлів і гілок. Листові вузли вказують на клас, який буде присвоєно зразку. Кожен внутрішній вузол дерева відповідає ознаці, а гілки представляють об'єднання ознак, які призводять до цих класифікацій. Кожна властивість має зна-

чення важливості, яке відмінне від нуля, тому нічого не потрібно видаляти з набору даних.

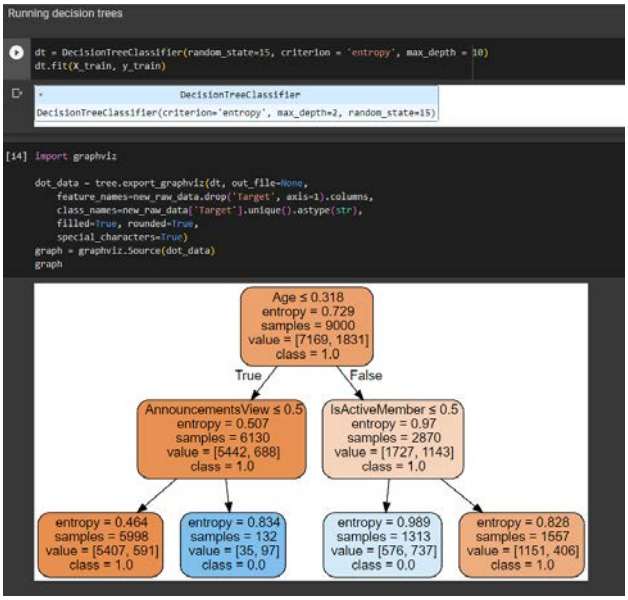


Рис. 5. Побудова дерева рішень

Розглянемо використання моделі RandomForest та покращення її резуль-татів шляхом підбору параметрів для роботи з набором даних (рис. 6), також побудуємо матрицю помилок [7]. З отриманої матриці (1 – успішне завершення курсу, 0 – студент не зміг завершити курс успішно) можна встановити, що ми отримали досить непогані результати класифікації, адже коефіцієнт правильного встановлення того що студент не завершить курс 0.96, а коефіцієнт правильності встановлення успішного завершення становить 0.47.

Наступний етап полягає в виборі найкращих параметрів моделі при використанні з навчальним набором даних, лише після цього можемо добитися більш кращого результату виконання прогнозного аналізу на абсолютно новому наборі даних (рис. 7, рис. 8). Наступний крок, це використання моделі на абсолютно новому наборі даних, для визначення того чи завершить студент курс та з якою імовірністю.

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

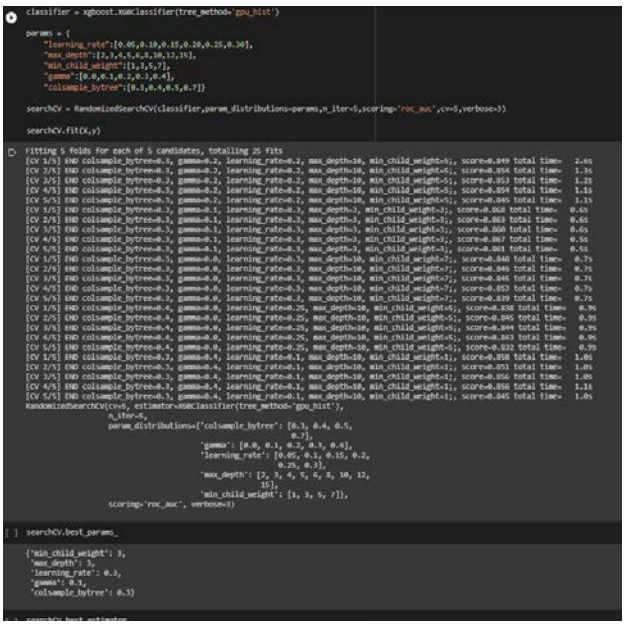


Рис. 7. Підбір параметрів моделі XGBoost

Аналіз результатів та визначення метрик для оцінки моделей

Точність моделі є основним індикативним фактором для оцінки моделі. Результати, отримані за алгоритмом, були представлені відповідно до найвищої спостережуваної точності.

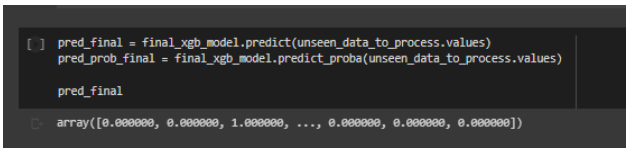


Рис. 8. Попередня обробка нових даних та використання фінальної моделі з обраними параметрами для прогнозування

Для оцінки моделі були використані такі характеристики:

- Точність (Accuracy) є найосновнішим показником продуктивності, який відображає загальну правильність передбачень моделі. Він обчислюється шляхом ділення кількості правильно передбачених випадків на загальну кількість випадків. Однак точність може ввести в оману, якщо набір даних незбалансований, тобто один клас має значно більше екземплярів, ніж інші.
- Презиція (Precision) – вимірює частку правильно передбачених позитивних випадків серед усіх випадків, передбачених як позитивні. Він акцентує увагу на правильності позитивних прогнозів. Точність обчислюється шляхом ділення кількості істинних позитивних результатів (правильно передбачених позитивних випадків) на суму істиних позитивних і хибних позитивних

результатів (неправильно передбачених позитивних випадків).

• Запам'ятовування (Recall) – також відоме як чутливість (Sensitivity) або справжній позитивний показник, вимірює частку правильно передбачених позитивних випадків серед усіх фактичних позитивних випадків. Він зосереджується на здатності моделі визначати позитивні випадки. Відкликання обчислюється діленням кількості істинно позитивних результатів на суму істинно позитивних результатів і хибно-негативних результатів (позитивні випадки, неправильно передбачені як негативні).

• Помилка класифікації – також відома як помилка неправильної класифікації або частота помилок, є показником, який використовується для оцінки ефективності моделі класифікації. Він кількісно визначає ступінь, до якого модель неправильно класифікує спостереження або робить неправильні прогнози. У контексті машинного навчання помилка класифікації є важливим показником для оцінки точності та надійності класифікатора.

• F-measure (F1) – це збалансована метрика, яка поєднує точність і запам'ятовування в один єдиний показник. Він забезпечує гармонійне середнє значення точності та запам'ятовування, надаючи однакову важливість обом показникам. F1-оцінка широко використовується, коли існує нерівномірний розподіл класів у наборі даних. Він розраховується як середньозважене значення точності та запам'ятовування за формулою:

Наступим кроком для більш кращого розуміння роботи нашої моделі є побудова матриці помилок, що дасть нам більш детальні характеристик того наскільки точним було наше прогнозування. Матриця помилок зазвичай використовується для оцінки ефективності методів, використаних після класифікації. Переглянемо матриці помилок після покращення параметрів моделей (1 – успішне завершення курсу, 0 – студент не зміг завершити курс успішно). Матриця плутанини, також відома як матриця помилок, – це таблиця, яка підсумовує продуктивність моделі класифікації. Це особливо корисно під час оцінювання моделей, які класифікують дані на два класи. Матриця плутанини дає уявлення про кількість істинно позитивних, істинно негативних, хибно-позитивних і хибно-негативних результатів, створених моделлю.

Ось розподіл компонентів у матриці плутанини для задачі бінарної класифікації з двома класами, які часто позначаються як «позитивний» і «негативний»:

• Істинно позитивні результати (TP – True Positives): це кількість випадків, які модель правильно класифікувала як позитивні.

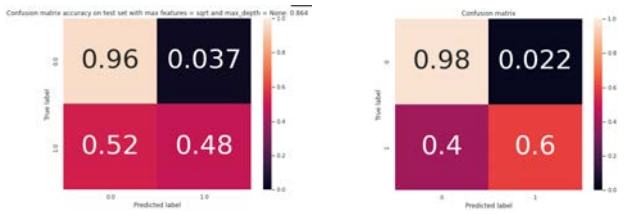
• Істинно негативні результати (TN – True Negatives): Це вказує на кількість екземплярів, які модель правильно класифікувала як негативні.

• Хибно-позитивні спрацьовування (False FP): також відомі як помилки типу I, це випадки, які модель неправильно класифікувала як позитивні, хоча насправді вони належать до негативного класу.

• Хибно-негативні (FN): також звані помилками типу II, це випадки, які були неправильно класифіковані як негативні моделлю, хоча вони насправді належать до позитивного класу.

Матриця помилок пропонує стислий підсумок результатів класифікації моделі, що дає нам змогу оцінити її ефективність і визначити потенційні області для вдосконалення.

Аналізуючи отримані метрики, можна зробити висновок, що XGBoost набагато краще справляється з проблемою прогнозної класифікації, але також не можна відкидати модель RandomForest для даної задачі, тому що вона також показує гарні результати.



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. В даній роботи було визначено проблему прогнозного моделювання та аналітики

успішності студентів з використанням алгоритмів машинного навчання. Дослідження і аналіз алгоритмів машинного навчання для прогнозного моделювання та аналітики успішності студентів є важливим кроком у покращенні освітнього процесу. Відштовхуючись від цього, було проведено дослідження та проаналізовано декілька алгоритмів машинного навчання.

Було проведено аналіз метрик та результатів натренованої моделі. З матриці помилок можна побачити, що отримано досить непогані результати після підбору параметрів для моделей. В основному моделі показали досить гарні результати. Також проведено приклад використання отриманих даних.

У результаті було отримано реалізацію алгоритму прогнозного аналізу з використанням моделей RandomForest та XGBoost, який підходить для подальшої модернізації та практичному застосуванні. З освітньої точки зору, це дослідження може допомогти викладачам визначити студентів, які перебувають у групі ризику щодо поганої успішності, і може допомогти педагогам вжити своєчасних коригувальних заходів.

Список літератури:

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. 28–33.

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/. (дата звернення: 10.08.2023)

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. 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. 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.

Vakaliuk T.A., Yanchuk V.M., Morozov D.S., Zubrytskii V.V., Novitska I.V.

PREDICTIVE MODELING OF STUDENT PERFORMANCE ANALYTICS USING MACHINE LEARNING ALGORITHMS

The term "predictive analytics" describes the application of statistics and machine learning techniques to forecast future results and performance. In order to assist you in making decisions, techniques like data mining and predictive modeling calculate the likelihood of future outcomes and notify you of impending events. The ability to anticipate student performance using machine learning has the potential to enhance educational systems. The use of machine learning in this area opens up new opportunities for analyzing large amounts of data, identifying complex dependencies, and developing personalized learning approaches. Modern learning management systems and e-platforms store a large amount of information about students'

academic achievements, such as grades, attendance, exams, etc. Using this data in conjunction with machine learning allows us to identify useful patterns and predict future student performance. In this paper, we have identified the problem of predictive modeling and analytics of student performance using machine learning algorithms. Research and analysis of machine learning algorithms for predictive modeling and analytics of student performance is an important step in improving the educational process. Based on this, several machine learning algorithms have been researched and analyzed. The metrics and results of the trained model were analyzed. From the error bars, we can see that quite good results were obtained after selecting the parameters for the models. As a result, an implementation of the predictive analysis algorithm using the RandomForest and XGBoost models was obtained, which is suitable for further modernization and practical application. From an educational point of view, this study can help teachers identify students who are at risk of poor performance and can help educators take timely corrective measures.

Key words: *predictive modeling, analytics, student performance, machine learning, algorithms.*