УДК [004.9:378.014.6](043.2)

Є. В. Боримський
С. О. Жуков

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АНАЛІЗУ ТА ПРОГНОЗУВАННЯ ПРОДУКТИВНОСТІ СТУДЕНТІВ З ВИКОРИСТАННЯМ МОДЕЛІ ДЕРЕВА РІШЕНЬ

Вінницький національний технічний університет

Анотація

Дослідження присвячене процесам навчання, оптимізації та тестування моделей машинного навчання для подальшого використання в інформаційній технології аналізу та прогнозування продуктивності студентів закладу вищої освіти. Проведено порівняльний аналіз базової та оптимізованої моделей, оцінено їхню точність за ключовими метриками (R^2, MAE, RMSE, MAPE) та підтверджено покращення прогнозної здатності після налаштування гіперпараметрів.

Ключові слова: машинне навчання, прогнозування, продуктивність студентів, оптимізація гіперпараметрів, тестування моделей, заклад вищої освіти.

Abstract

The research is devoted to the training, optimization, and testing of machine learning models for further use in information technology to analyze and predict the productivity of higher education students. A comparative analysis of the basic and optimized models was carried out, their accuracy was evaluated according to key metrics (R^2, MAE, RMSE, MAPE), and the improvement of predictive ability after hyperparameter tuning was confirmed.

Keywords: machine learning, forecasting, student productivity, hyperparameter optimization, model testing, institution of higher education.

Вступ

В умовах інтенсивної цифровізації освітнього середовища проблема аналізу та своєчасного прогнозування продуктивності студентів залишається надзвичайно актуальною. Після збору даних та проведення розвідувального аналізу (EDA) критичним етапом побудови надійної інформаційної технології є конструювання математичного апарату — моделей машинного навчання [1].

Створення точного прогнозу неможливе лише завдяки базовим алгоритмам, оскільки вони часто страждають від проблеми перенавчання (overfitting), коли модель ідеально запам’ятовує тренувальний набір, але втрачає здатність до узагальнення нових даних. Розв'язання цієї проблеми полягає в тестуванні моделей на відкладених вибірках та кропіткій оптимізації їхніх гіперпараметрів [2].

Актуальність дослідження зумовлена тим, що висока точність машинного прогнозування є запорукою коректної роботи систем раннього попередження. Це дає змогу закладам освіти своєчасно виявляти ризики зниження академічної успішності та персоналізувати підтримку студентів у цифровому середовищі.

Отже, метою роботи є проведення навчання, оптимізації та тестування моделей машинного навчання для інформаційної технології аналізу та прогнозування продуктивності студентів.

Результати дослідження

Для побудови моделей було використано набір даних «Student Productivity and Digital Distraction», що містить 20 000 записів. Цей датасет містить 18 змінних, які комплексно описують академічні результати (відвідуваність, виконані завдання), фізіологічний стан (години сну, рівень стресу, споживання кофеїну, фізична активність) та ступінь цифрових відволікань (кількість годин у соціальних мережах, іграх, на телефоні та YouTube) студентів. Короткий опис структури цього набору даних та результати його первинного розвідувального аналізу детально представлені у попередній роботі [3], а сам першоджерело розміщено у відкритому доступі на платформі Kaggle [4].

Під час моделювання спершу було побудовано базову модель із параметрами за замовчуванням. Процес тестування продемонстрував класичну проблему перенавчання: коефіцієнт детермінації (R^2) на тренувальній вибірці досягнув ідеального значення 1.0000, тоді як на тестовій він складав 0.8990.

Середня абсолютна похибка (MAE) базової моделі на тестових даних дорівнювала 3.6749, а середньоквадратична похибка (RMSE) — 4.7727 [5].

Щоб підвищити точність і стабільність прогнозування, було здійснено оптимізацію гіперпараметрів. Це дозволило зменшити надмірну складність моделі і збалансувати її роботу [6]. У таблиці 1 наведено результати порівняння метрик базової та оптимізованої моделей.

Таблиця 1. Порівняння результатів тестування моделей

Метрика	Базова модель	Оптимізована модель
R^2 (Тренування)	1.0000	0.9690
R^2 (Тест)	0.8990	0.9131
MAE (Тест)	3.6749	3.4086
RMSE (Тест)	4.7727	4.4283
MAPE% (Тест)	8.02	7.42

Оптимізована модель продемонструвала значно кращу здатність до узагальнення. Метрика R^2 на тренуванні знизилась до 0.9690, проте на тестовій вибірці зросла до 0.9131. Застосування оптимізації дозволило покращити R^2 Score на 1.56%. Відбулося відчутне зменшення відхилень: MAE покращилась на 7.25%, знизившись до 3.4086, показник RMSE впав до 4.4283, а середня абсолютна помилка у відсотках (MAPE) зменшилась з 8.02% до 7.42%.

На рисунку 1 наведено криві навчання (Learning Curve). Графік демонструє залежність метрики R^2 від розміру тренувального набору: синя крива відображає точність на тренувальній вибірці, а червона — на валідаційній. Зі збільшенням обсягу даних спостерігається чітка тенденція до зближення цих кривих, що вказує на поступове зростання здатності моделі до узагальнення. Це переконливо доводить, що наявного обсягу даних достатньо для уникнення перенавчання та забезпечення стабільного прогнозування.

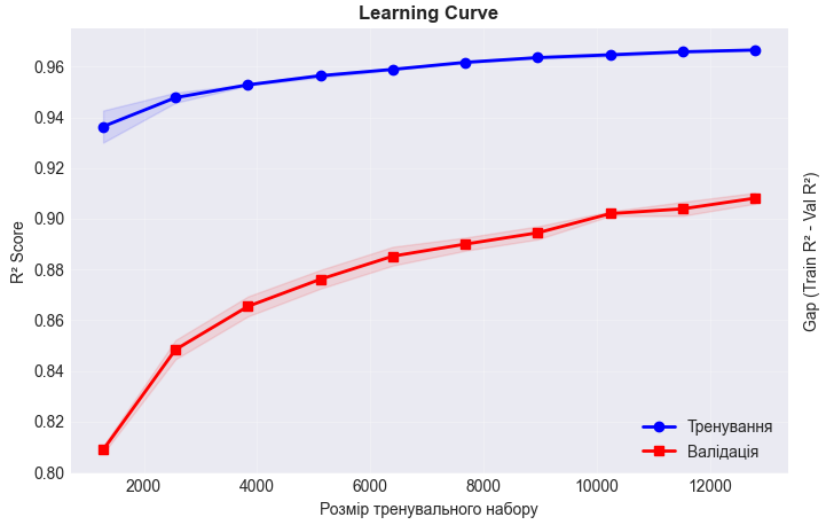

Рис. 1. Криві навчання для тренувальної та тестової вибірок

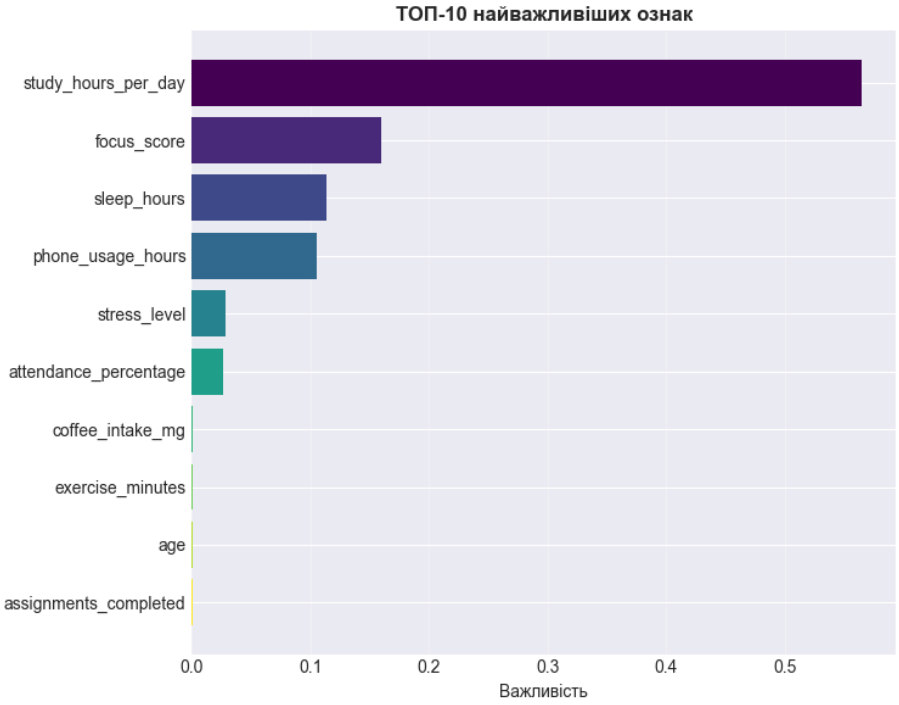

Рис. 2. Важливості ознак відсортовані за спаданням

Також було проведено аналіз важливості ознак (Рис. 2). Оптимізована модель підтвердила, що в топ-5 найважливіших чинників входять кількість годин навчання на день, показник концентрації та години сну, що узгоджується з результатами розвідувального аналізу даних [7].

Висновки

У ході дослідження було успішно реалізовано етапи побудови, оптимізації гіперпараметрів та тестування моделей машинного навчання для оцінки й прогнозування продуктивності студентів. Здійснений порівняльний аналіз підтвердив складність математичного моделювання поведінкових чинників та показав, що базова модель дерева рішень із параметрами за замовчуванням є схильною до ефекту перенавчання (overfitting), демонструючи ідеальний коефіцієнт детермінації на тренувальній вибірці ($R^2 = 1.0000$) при суттєвому зниженні точності на тестових даних ($R^2= 0.8990$).

Цілеспрямоване налаштування гіперпараметрів дозволило знизити надмірну складність структури дерева рішень та забезпечило збалансовану й стабільну роботу алгоритму з новими даними. Впровадження оптимізованої моделі дозволило підвищити узагальнювальну здатність системи:

Аналіз побудованих кривих навчання (Learning Curves) показав чітку тенденцію до зближення тренувальної та валідаційної кривих у міру збільшення обсягу даних, що підтверджує репрезентативність використаного датасету (20 000 записів) та достатність вибірки для запобігання перенавчанню. Дослідження відсортованої важливості ознак (Feature Importance) дозволило кількісно підтвердити, що фундаментальними домінуючими компонентами впливу на результати студентів є кількість годин навчання на день, показник концентрації та години сну, що повністю узгоджується з теоретичними засадами та результатами первинного аналізу даних.

Таким чином, протестована й оптимізована модель дерева рішень забезпечує формування надійного алгоритмічного та аналітичного ядра для створення повноцінної інформаційної технології прогнозування продуктивності студентів. Отримані результати та налаштована математична архітектура готові до практичної інтеграції в освітнє середовище закладів вищої освіти, оскільки дозволяють не лише здійснювати моніторинг поточного стану, а й забезпечують можливість раннього виявлення ризиків зниження академічної ефективності та обґрунтованої підтримки прийняття управлінських рішень.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Моторіна В. Г., Дем'яненко О. О., Марущак О. В. Аналіз впливу цифрових технологій на якість вищої освіти в Україні в умовах глобальних викликів /В. Г. Моторіна, О. О. Дем'яненко, О. В. Марущак // Педагогічна Академія: наукові записки. – 2024. – вип. 10, https://pedagogical-academy.com/index.php/journal/article/view/343
2. Мокін В. Б., Дратований М. В. Наука про дані: машинне навчання та інтелектуальний аналіз даних. Навчальний посібник. Вінниця: ВНТУ, 2024. 263 с. [Електронний ресурс] – Режим доступу: https://iq.vntu.edu.ua/repository/getfile.php/8163.pdf
3. Боримський Є. В., Войцеховська О. О. Розвідувальний аналіз даних для інформаційної технології аналізу та прогнозування продуктивності студентів закладів вищої освіти // Матеріали LV Всеукраїнської науково-технічної конференції факультету інтелектуальних інформаційних технологій та автоматизації, Вінниця, 2026. [Електронний ресурс]. – Режим доступу: https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2026/paper/view/29137
4. Student Productivity & Digital Distraction Dataset [Електронний ресурс]. – Режим доступу: https://www.kaggle.com/datasets/sehaj1104/student-productivity-and-digital-distraction-dataset
5. Chugh A. MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better? / A. Chugh // Analytics Vidhya. – 2020. [Електронний ресурс]. – Режим доступу: https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e
6. Гіперпараметри (Машинне навчання). [Електронний ресурс] – Режим доступу: https://developers.google.com/machine-learning/crash-course/linear-regression/hyperparameters?hl=uk
7. Khan, Mohammad A., and Hamdan Al-Jahdali. "The consequences of sleep deprivation on cognitive performance." (2023): [Електронний ресурс]. – Режим доступу: https://pubmed.ncbi.nlm.nih.gov/37045455/

Боримський Євгеній Володимирович – студент групи СА-24б, Факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, м. Вінниця, e-mail: zenia.zt2006@gmail.com

Жуков Сергій Олександрович **—** кандидат технічних наук, доцент кафедри системного аналізу та інформаційних технологій, Вінниця, e-mail: sazhukov@gmail.com

Borymskyi Yevhenii V. – student of the SA-24b group, Faculty of Intelligent Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, email: zenia.zt2006@gmail.com

Zhukov Serhii O. **—** Cand. Sc. (Eng.), Assistant Professor of the Department of Systems Analysis and Information Technologies, Vinnytsia, e-mail: sazhukov@gmail.com.