

ПОВЕДІНКОВІ МОДЕЛІ ОЦІНЮВАННЯ УСПІШНОСТІ СТУДЕНТІВ У СМАРТ-КЛАСАХ

Ткач В.С.¹, Кузнєцова Н.В.²

Національний технічний університет України
«Київський політехнічний інститут ім. Ігоря Сікорського», Київ, Україна

¹ tkach.viktoria@lil.kpi.ua, ² n.kuznietsova@kpi.ua

У роботі досліджено поведінкові дані студентів у смарт-класах та представлено підхід до прогнозування їхньої залученості й ризику відтоку (відвідуваності) з класу. На основі набору даних OULAD побудовано моделі машинного навчання та моделі виживання, які дають змогу оцінювати ключові фактори успішності, відстежувати часову динаміку активності та своєчасно виявляти студентів із підвищеним ризиком припинення навчання.

Ключові слова: смарт-клас, поведінкові дані, машинне навчання, моделі виживання, прогнозування відтоку.

1. ВСТУП

Смарт-клас – це сучасне навчальне середовище, у якому поєднано цифрові технології та педагогічні підходи. Завдяки збору даних щодо активності студентів, їх результати та взаємодію з матеріалами та навчальними ресурсами стає можливим створення індивідуальних, орієнтованих на конкретних студентів, планів навчання та підвищення якості засвоєння знань. У період активної цифровізації такі класи особливо важливі, адже вони забезпечують гнучкість і допомагають швидко реагувати на потреби кожного студента.

Важливою частиною дослідження у даній сфері є аналіз поведінкових даних. Показники активності, участі, взаємодії та оцінювання можуть вказувати на рівень залученості, який в свою чергу впливає на ризик відтоку. Уміння прогнозувати ці процеси робить смарт-клас корисним інструментом для прийняття своєчасних рішень. Актуальність теми зумовлена тим, що сьогодні навчальні заклади мають потребу забезпечувати вищу якість освітнього процесу в умовах зростання вимог до результатів навчання і залучати для цього сучасні цифрові технології, які роблять навчальний процес більш інтерактивним. Для цього необхідно своєчасно виявляти студентів, які можуть втратити зацікавленість або зіткнутися з труднощами у навчанні. Саме тому поведінкових аналіз і прогнозування відтоку стають важливими інструментами підтримки студентів та прийняття своєчасних педагогічних рішень.

2. ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

Метою даного дослідження є аналіз поведінкових даних студентів та прогнозування їхньої залученості й ризику відтоку у смарт-класі з використанням методів машинного навчання та моделей виживання. У роботі застосовуються різні підходи класифікації й аналізу часу до настання певної події-триггеру, проводиться їх порівняння та оцінювання точності для визначення найбільш ефективних методів прогнозування навчальної активності студентів. Такий підхід дозволяє визначити чинники, що впливають на успішність, та своєчасно виявляти студентів із підвищеним ризиком припинення навчання. Об'єкт дослідження – процес поведінкової та навчальної активності студентів у смарт-класі.

Предмет дослідження – моделі машинного та моделі виживання, що використовуються для прогнозування залученості та ризику відтоку студентів.

3. МЕТОДИ МАШИННОГО НАВЧАННЯ ТА МОДЕЛІ ВИЖИВАННЯ

3.1. Методи машинного навчання

У дослідженнях, пов'язаних із класифікацією та прогнозуванням, широко застосовуються різноманітні методи машинного навчання, які дозволяють моделювати взаємозв'язки між ознаками та визначити належність об'єктів до певних категорій. У даній роботі реалізовано п'ять підходів: логістичну регресію, дерево рішень, градієнтний бустинг, наївний байєсівський класифікатор та метод опорних векторів із RBF-ядром.

Логістична регресія – це статистичний метод бінарної класифікації, який моделює ймовірність належності об'єкта до певного класу за допомогою логістичної функції. Вона дозволяє оцінювати вплив кожної ознаки на результат [1].

Дерево рішень – метод, який формує ієрархічну структуру з послідовності умов «якщо – то». Модель на основі дерева рішень розбиває простір ознак на підмножини, будуючи логічні правила, які забезпечують зрозумілу та наочну класифікацію [2].

Градієнтний бустинг – ансамблевий метод, у якому набір простіших моделей, наприклад, дерев рішень, додається послідовно, кожна з яких коригує помилки попередньої. Завдяки поетапному покращенню метод здатен відтворювати складні закономірності [3].

Наївний байєсівський класифікатор ґрунтується на теоремі Байєса та припускає незалежність ознак між собою. Попри простоту і деяку наївність даний метод часто демонструє стабільну роботу та використовується для швидкої і надійної класифікації [4].

Метод опорних векторів з RBF-ядром – це алгоритм, який шукає оптимальну межу між класами, максимізуючи відстань між ними. Використовуючи радіально-базисної функції дозволяє моделювати нелінійні залежності та ефективно розділяти дані на класи у багатовимірному просторі [5].

3.2. Моделі виживання

Моделі виживання застосовуються для аналізу часу до настання події та дозволяють працювати з цензурованими спостереженнями. У дослідженні використано три підходи: модель Каплана-Майєра, модель пропорційних ризиків Кокса та модель прискореного часу до події Вейбула.

Модель Каплана-Майєра є непараметричною моделлю, яка оцінює функцію виживання й показує частку об'єктів, що залишаються активними у різні моменти часу. Вона будує ступінчасту криву, що дозволяє порівнювати різні групи за динамікою виживання [6].

Модель пропорційних ризиків Кокса описує, як змінні впливають на ризик настання події. Основна формула має вигляд:

$$h(t|x) = h_0(t)\exp(\beta x),$$

де $h(t|x)$ – ризик настання події, $h_0(t)$ – базовий ризик, βx – лінійна комбінація коефіцієнтів і коваріант. Експонента коефіцієнтів $\exp(\beta)$ інтерпретується як hazard ratio, тобто у скільки разів змінюється ризик при зміні ознаки [7].

Модель прискореного часу до події Вейбулла напряду моделює тривалість до події. Вона ґрунтується на рівнянні:

$$\ln(T) = \beta x + \sigma W,$$

де T – час до події, β – коефіцієнти, σ – параметр масштабу, а W – випадкова складова залежно від вибраного розподілу. У цій моделі $\exp(\beta)$ інтерпретується як hazard ratio, який показує, як змінюється тривалість часу до події за зміни певних характеристик [8].

Представлені підходи охоплюють як лінійні так і не лінійні способи моделювання даних, що забезпечує комплексне дослідження різних типів залежностей. Сукупне використання цих методів створює теоретичне підґрунтя для подальшого аналізу та інтерпретації поведінкових процесів у навчальному середовищі.

4. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Для аналізу поведінкових даних студентів було використано відкритий навчальний набір даних OULAD [9], який містить інформацію щодо активності здобувачів освіти, їхні демографічні характеристики, результати оцінювання. Попередня обробка даних включала очищення даних, узгодження таблиць та перетворення записів у формат який придатний для подальшого моделювання. Окремо було сформовано підмножини даних для класифікаційних методів і моделей виживання.

На наступному етапі побудовано та протестовано моделі машинного навчання й моделі виживання відповідно до описаних раніше підходів. Для кожного методу розраховано ключові метрики точності класифікації, які відображають здатність моделі коректно розмежовувати студентів на групи «відрахований» та «не відрахований». Узагальнене порівняння результатів наведено у таблиці 1, де подано значення метрик macro-F1, accuracy та ROC-AUC для всіх моделей.

Таблиця 1. Порівняння результатів моделей класифікації

Модель	Macro-F1	Accuracy	ROC-AUC
Логістична регресія	0.9362	0.9361	0.9804
Гradientний бустинг	0.9393	0.9392	0.9852
Дерево рішень	0.9187	0.9185	0.9185
Метод опорних векторів (RBF)	0.9408	0.9407	0.9811
Наївний байєсівський класифікатор	0.8673	0.8664	0.9443

Аналіз результатів показав, що найкращі значення основних показників точності продемонструвала модель методу опорних векторів із RBF-ядром. Gradientний бустинг і логістична регресія також показали високу якість класифікації, тоді як дерево рішень і наївний байєсівський класифікатор показали нижчі результати.

На основі результатів прогнозування, отриманих методами машинного навчання, було визначено ключові характеристики студентів, найбільш схильних до відрахування. Аналіз показав, що проблемні студенти мають суттєво нижчу активність у віртуальному навчальному середовищі: рідше заходять у навчальну систему, виконують менше завдань і раніше припиняють участь у курсі. Також виявлено відмінність у соціально-демографічних ознаках, такі студенти частіше мають нижчий рівень освіти та проживають у менш забезпечених регіонах. Сукупність цих факторів формує узагальнений профіль студента з підвищеним ризиком відтоку.

Аналіз часу до відрахування студентів був проведений на основі трьох моделей виживання. Загальна крива Каплана-Майєра (рис. 1) показує, що найбільше відрахувань студентів відбувається саме на початку навчання, а саме у перші тижні функція виживання помітно зменшується, після цього зниження стає більш плавним, але до кінця періоду спостерігається, що ймовірність для студента залишитися на курсі поступово зменшується

приблизно до 0.4. Це свідчить про те, що критичний період ризику припадає на першу чверть навчання.

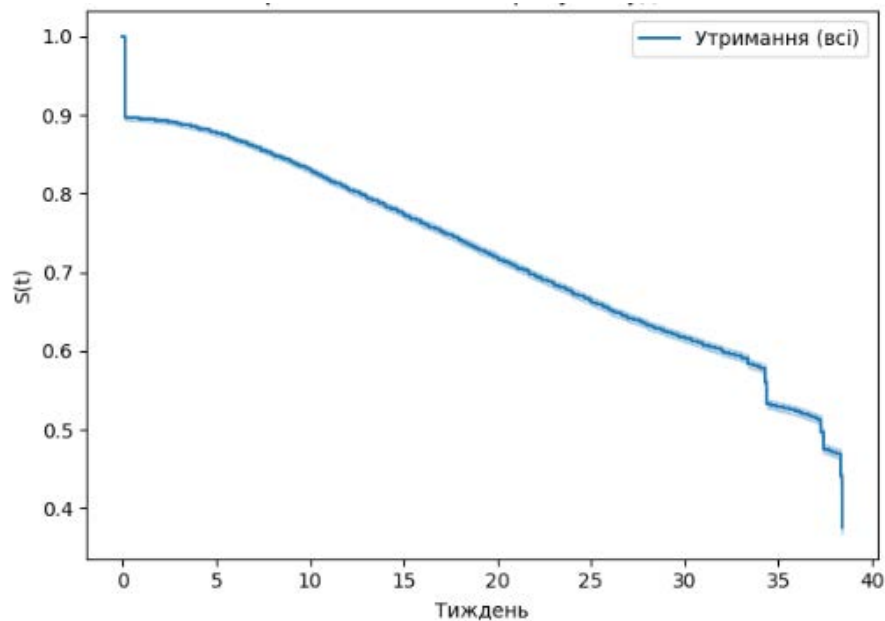


Рисунок 1. Крива Каплана-Майєра

Стратифіковані криві Каплана-Майєра за рівнем активності (рис. 2) демонструють чітку різницю у часовій динаміці між групами. Студенти з низькою кількістю кліків вибувають значно раніше і ця різниця формується вже з перших тижнів навчання. Натомість студенти з високою активністю залишаються на курсі довше. Це підкреслює те, що інтенсивність взаємодії з платформою прямо пов'язана з тривалістю навчання.

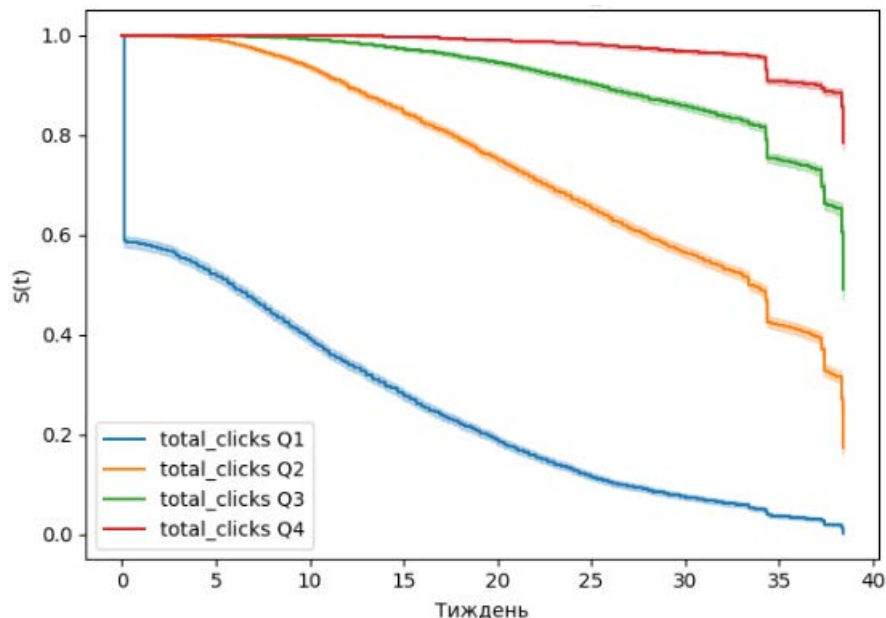


Рисунок 2. Стратифіковані криві Каплана-Майєра

На рисунку 3 наведено результати аналізу з використанням моделі пропорційних ризиків Кокса, яка дозволяє оцінити вплив окремих ознак на швидкість настання події. З графіка видно, що поведінкові показники: активні дні, кількість виконаних завдань і середній

бал зменшують ризик відрахування, тоді як окремі демографічні та регіональні характеристики пов'язані з більш раннім формуванням ризику. Це свідчить про те, що регулярна активність уповільнює настання події, тоді як соціальна вразливість її прискорює.

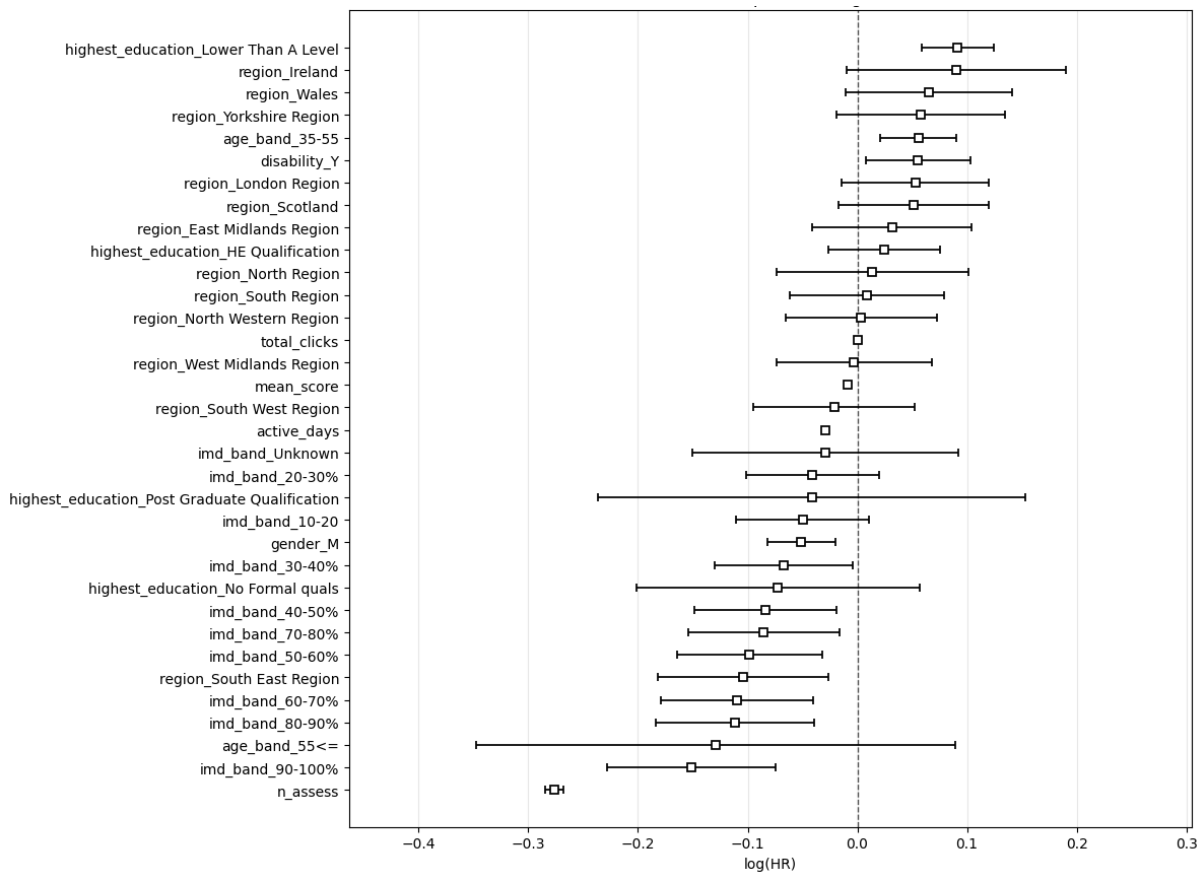


Рисунок 3. Коефіцієнти моделі Кокса

Прогнозні криві виживання отримані на основі моделі Вейбулла (рис. 4) показують часові відмінності між студентами різної активності. Студенти з низькою активністю мають суттєво короткий час до відрахування і вибувають приблизно до 25 тижня, тоді як активні студенти демонструють високу ймовірність залишитися на курсі.

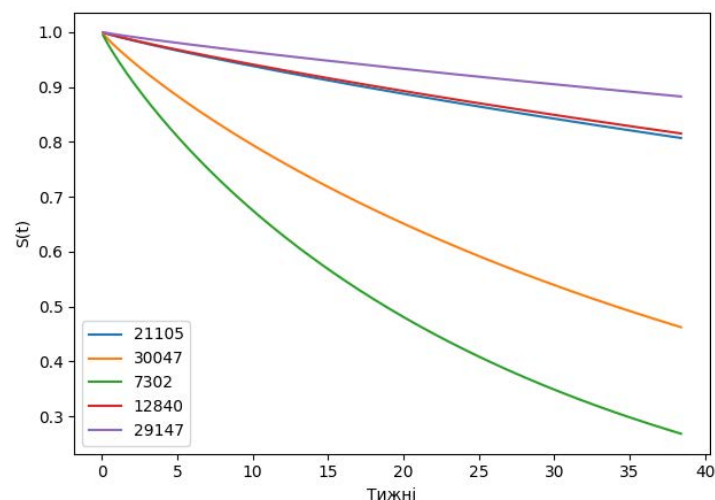


Рисунок 4. Прогнозні криві виживання моделі Вейбулла

5. ВИСНОВКИ

У ході дослідження було проаналізовано поведінкові дані студентів у смарт-класі та побудовано моделі машинного навчання і моделі виживання для прогнозування залученості студентів та ризику їх відтоку. Методи класифікації показали високу точність у визначенні студентів із підвищеною ймовірністю відрахування, а найкращі результати продемонстрував метод опорних векторів із RBF-ядром. Отримані моделі дали змогу виявити ключові чинники ризику, серед яких визначальне значення мають показники активності, а саме частота входу в навчальну систему, кількість виконаних завдань, активність та рівень успішності.

Аналіз часу до події підтвердив, що критичний період ризику припадає на початок навчання. Крива Каплана-Майєра показала швидке зниження виживання у перші тижні, а стратифікація за активністю продемонструвала суттєві часові відмінності між групами студентів за кількістю кліків. Модель Кокса дала змогу оцінити вплив окремих на швидкість настання відрахування, тоді як модель Вейбулла показала, що студенти з низькою активністю вибувають приблизно до 25 тижня.

Отримані результати свідчать, що поведінкові характеристики студентів є ключовими показниками їхньої стабільності у навчанні. Застосування моделей машинного навчання та аналізу виживання у смарт-класах дозволяє своєчасно виявляти студентів із підвищеним ризиком відтоку, визначити критичні часові точки та підтримувати здобувачів освіти на ранніх етапах в критичні моменти.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. LaValley M. P. Statistical Primer for Cardiovascular Research. Logistic Regression. *Circulation*. 2008. Vol. 117. P. 2395-2399.
2. Song Y., Lu Y. Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*. 2015. Vol. 27, No. 2. P. 130-135.
3. Friedman J.H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*. 2021. Vol.29, No. 5. P. 1189-1232.
4. Peretz O., Koren M., Koren O. Naïve Bayes classifier – An ensemble procedure for recall and precision enrichment. *Engineering Applications of Artificial Intelligence*. 2024. Vol. 136, 108972. 12 p.
5. Ding X., Liu J., Yang F., Cao J., Random radial basis function kernel-based support vector machine. *Journal of the Franklin Institute*. 2021. Vol. 358, No. 18. P. 10121-10140.
6. D'Arrigo G., Leonardis D., Abd Elhafeez S., Fusaro M., Tripepi G., Roumeliotis S. Method to Analyse Time-to-Event Data: The Kaplan-Maier Survival Curve. *Oxidative Medicine and Cellular Longevity*. 2021. Vol. 2021, 2290120, 7 p.
7. Fox J., Weisberg S., Cox Proportional-Hazards Regression for Survival Data in R. *An R Companion to Applied Regression*. 2023. Vol. 3, 20 p.
8. Liu E., Liu R., Lim K., Using the Weibull Accelerated Failure Time Regression Model to Predict Time to Health Events. *Applied Biostatistics for the Health Science and Epidemiology*. 2023. Vol. 13, No. 14. 13041
9. OULAD: Open University Learning Analytics Dataset. URL: <https://www.kaggle.com/code/vjcalling/oulad-open-university-learning-analytics-dataset>