

РОЗВІДУВАЛЬНИЙ АНАЛІЗ ДАНИХ ДЛЯ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ АНАЛІЗУ ТА ПРОГНОЗУВАННЯ ПРОДУКТИВНОСТІ СТУДЕНТІВ ЗАКЛАДІВ ВИЩОЇ ОСВІТИ

Вінницький національний технічний університет

Анотація

Дослідження присвячене підготовці та розвідувальному аналізу даних для подальшого використання в інформаційній технології аналізу та прогнозування продуктивності студентів закладу вищої освіти методами машинного навчання. Проведено аналіз структури датасету та його основних характеристик, а також досліджено можливі взаємозв'язки між рівнем продуктивності та ключовими чинниками.

Ключові слова: розвідувальний аналіз даних, інформаційна технологія, машинне навчання, аналіз, продуктивність студентів, заклад вищої освіти.

Abstract

The research is devoted to the preparation and exploratory analysis of data for further use in information technology analysis and prediction of the productivity of students of a higher education institution using machine learning methods. The structure of the dataset and its main characteristics were analyzed, and possible relationships between the level of productivity and key factors were investigated.

Keywords: exploratory data analysis, information technology, machine learning, analysis, student productivity, institution of higher education.

Вступ

Проблема зниження продуктивності серед студентів закладів вищої освіти набуває особливої актуальності в умовах інтенсивної цифровізації освітнього середовища. Сучасний навчальний процес супроводжується формуванням значних обсягів даних про діяльність студентів, що створює необхідність у впровадженні інтелектуальних технологій для їхнього аналізу. Ефективне використання цієї інформації дозволяє не лише оптимізувати навантаження, а й підвищити якість освіти через персоналізований підхід до кожного студента [1].

Продуктивність студента є складним показником, на який впливають як академічні результати, так і фізіологічний стан, режим сну та рівень цифрових відволікань.

У процесі розробки систем прогнозування ключову роль відіграє розвідувальний аналіз даних (EDA). Він виступає сполучною ланкою між «сирими» даними та прогнозними моделями, дозволяючи виявляти приховані закономірності, аномалії та відбирати найбільш важливі ознаки. Нехтування цим етапом призводить до зниження точності роботи алгоритмів машинного навчання через вплив зайвого інформаційного шуму [2].

Актуальність дослідження зумовлена необхідністю розроблення систем раннього попередження, які дають змогу закладам освіти своєчасно виявляти негативні чинники, що впливають на якість навчання та добробут молоді, та оперативно реагувати на ці зміни. А використання методів машинного навчання відкриває нові можливості для моніторингу ефективності освіти та підтримки студентів у цифровому середовищі.

Отже, метою роботи є проведення розвідувального аналізу даних для подальшого використання в інформаційній технології аналізу та прогнозування продуктивності студентів закладу вищої освіти методами машинного навчання.

Результати дослідження

Для проведення аналізу було обрано набір даних, що має назву «Student Productivity and Digital Distraction», опублікований на платформі Kaggle у відкритому доступі [3]. Датасет містить 20 000 записів. На етапі первинного аналізу було проведено типізацію 18 змінних:

1. Числові:
 - унікальний ідентифікатор студента (student_id);
 - вік (age);
 - кількість годин навчання на день (study_hours_per_day);
 - години сну (sleep_hours);
 - години використання телефону (phone_usage_hours);
 - години в соціальних мережах (social_media_hours);
 - години перегляду YouTube (youtube_hours);
 - години за іграми (gaming_hours);
 - кількість перерв на день (breaks_per_day);
 - споживання кофеїну в міліграмах (coffee_intake_mg);
 - хвилини фізичних вправ (exercise_minutes);
 - кількість виконаних завдань (assignments_completed);
 - відсоток відвідуваності (attendance_percentage);
 - рівень стресу (stress_level);
 - показник концентрації (focus_score);
 - фінальна оцінка (final_grade);
 - показник продуктивності (productivity_score) – цільова ознака.
2. Категорійні: стать (gender).

На рисунку 1 наведено фрагмент датасету «Student Productivity and Digital Distraction», що включає 18 ознак.

	student_id	age	gender	study_hours_per_day	sleep_hours	phone_usage_hours	social_media_hours	youtube_hours	gaming_hou
0	1	23	Female	4.35	3.63	3.38	2.73	1.83	5.26
1	2	20	Male	6.14	6.58	5.48	1.51	3.13	1.73
2	3	29	Female	4.98	3.26	4.83	3.63	0.18	4.71
3	4	27	Female	3.19	4.58	10.06	3.95	5.75	2.52
4	5	24	Male	7.67	6.21	3.02	1.59	5.46	5.65

Рис. 1. Приклад статистичних показників числових ознак

Після огляду структури датасету було перевірено основні статистичні показники (рис. 2).

	student_id	age	study_hours_per_day	sleep_hours	phone_usage_hours	social_media_hours	youtube_hours	gaming_hours	breaks_per_day
count	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000
mean	10000.500000	23.007450	5.254562	6.517799	6.250479	4.003655	2.990177	2.988339	7.542400
std	5773.647028	3.754890	2.742876	2.029784	3.313082	2.305154	1.729815	1.732803	4.016231
min	1.000000	17.000000	0.500000	3.000000	0.500000	0.000000	0.000000	0.000000	1.000000
25%	5000.750000	20.000000	2.900000	4.770000	3.380000	2.000000	1.487500	1.490000	4.000000
50%	10000.500000	23.000000	5.250000	6.510000	6.240000	4.010000	2.980000	2.970000	8.000000
75%	15000.250000	26.000000	7.640000	8.310000	9.102500	5.970000	4.480000	4.490000	11.000000
max	20000.000000	29.000000	10.000000	10.000000	12.000000	8.000000	6.000000	6.000000	14.000000

Рис. 2. Приклад статистичних показників числових ознак

Значення статистичних показників, представлені на рисунку 2, вказують на центральні тенденції вибірки. Наприклад, середня кількість годин навчання на добу становить $\mu = 5,25$ при стандартному відхиленні $\sigma = 2,74$, тобто існує значна варіативність академічних звичок щодо виділення студентами часу на сну. Ці дані допомогли оцінити приблизний розподіл значень для кожної ознаки.

Далі було проведено аналіз розподілу середнього значення витраченого часу для кожної статі з метою виявлення потенційних поведінкових відмінностей між групами студентів (рис. 3) [4-6].

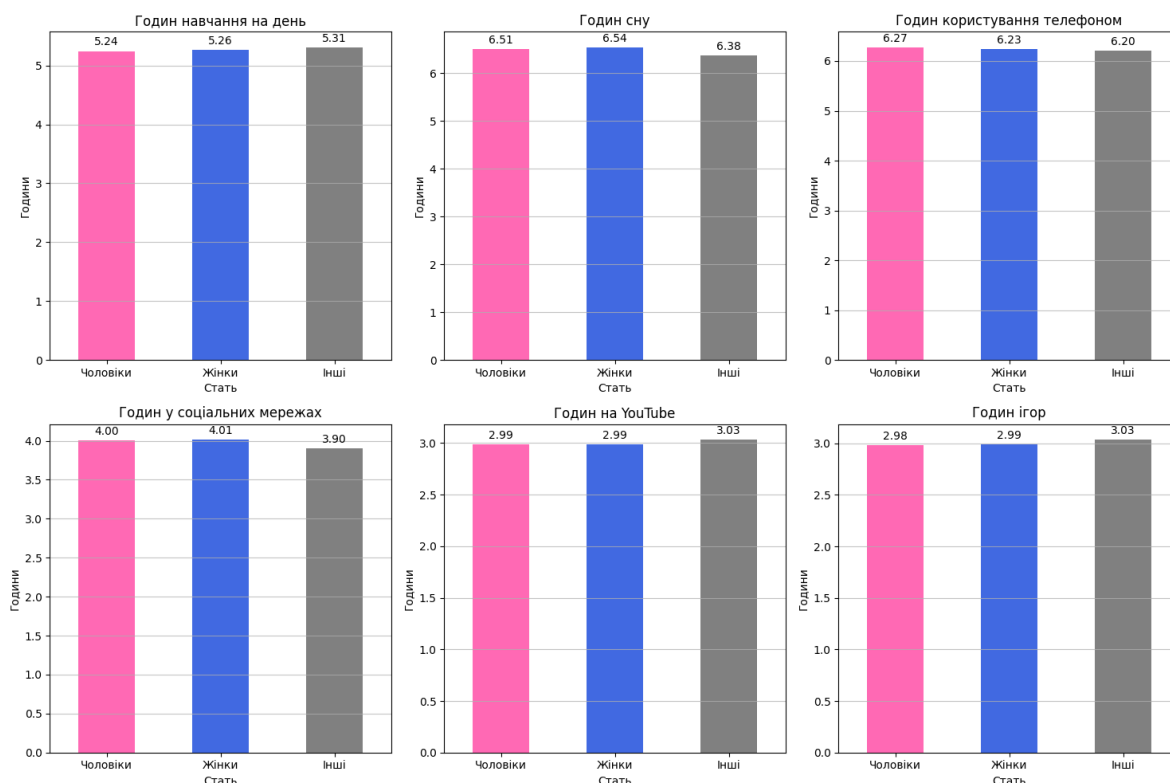


Рис. 3. Розподіл витраченого часу на активності за статтю

Аналіз показників годин навчання на день та годин сну показує майже ідентичні середні значення для чоловіків, жінок та категорії «інші», де коливання не перевищують десятих часток години. Це свідчить про те, що базовий режим дня та рівень академічного навантаження не залежать від статі студента.

Графіки інших показників, таких як години користування телефоном, соціальні мережі, YouTube та ігри, демонструють аналогічну стабільність. Така рівність у розподілі часу на розваги та навчання підтверджує, що в межах цієї вибірки стать не є диференціюючим фактором для продуктивності.

З графіків, приведених на рисунку 3 видно, що ознака статі має дуже низьку прогностичну здатність для цільової змінної. Оскільки розподіл активностей практично однаковий для всіх груп, цей параметр, ймовірно, не є ключовим фактором у математичній моделі передбачення продуктивності.

Далі було побудовано графік розподілу цільової ознаки (рис. 4) [4-6]. Це дозволило візуально оцінити, наскільки рівномірно розподілені значення продуктивності серед 20 000 студентів та зрозуміти переважаючий рівень ефективності в даній вибірці.

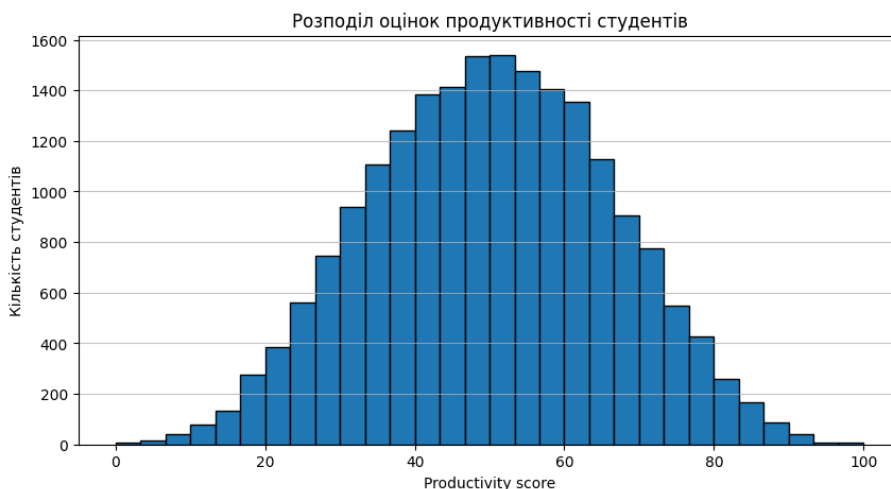


Рис. 4. Графік розподілу оцінок продуктивності студентів

Графік, представлений на рисунку 4 показує, що більшість значень продуктивності зосереджена в центральній частині діапазону, а саме в межах від 40 до 60 одиниць, де спостерігається найвища щільність даних. Модальний інтервал знаходиться навколо позначки 50, що свідчить про середньостатистичну ефективність більшості студентів у вибірці.

Також було побудовано матрицю кореляцій для пошуку лінійних зв'язків між усіма парами ознак у датасеті (рис. 5) [2].

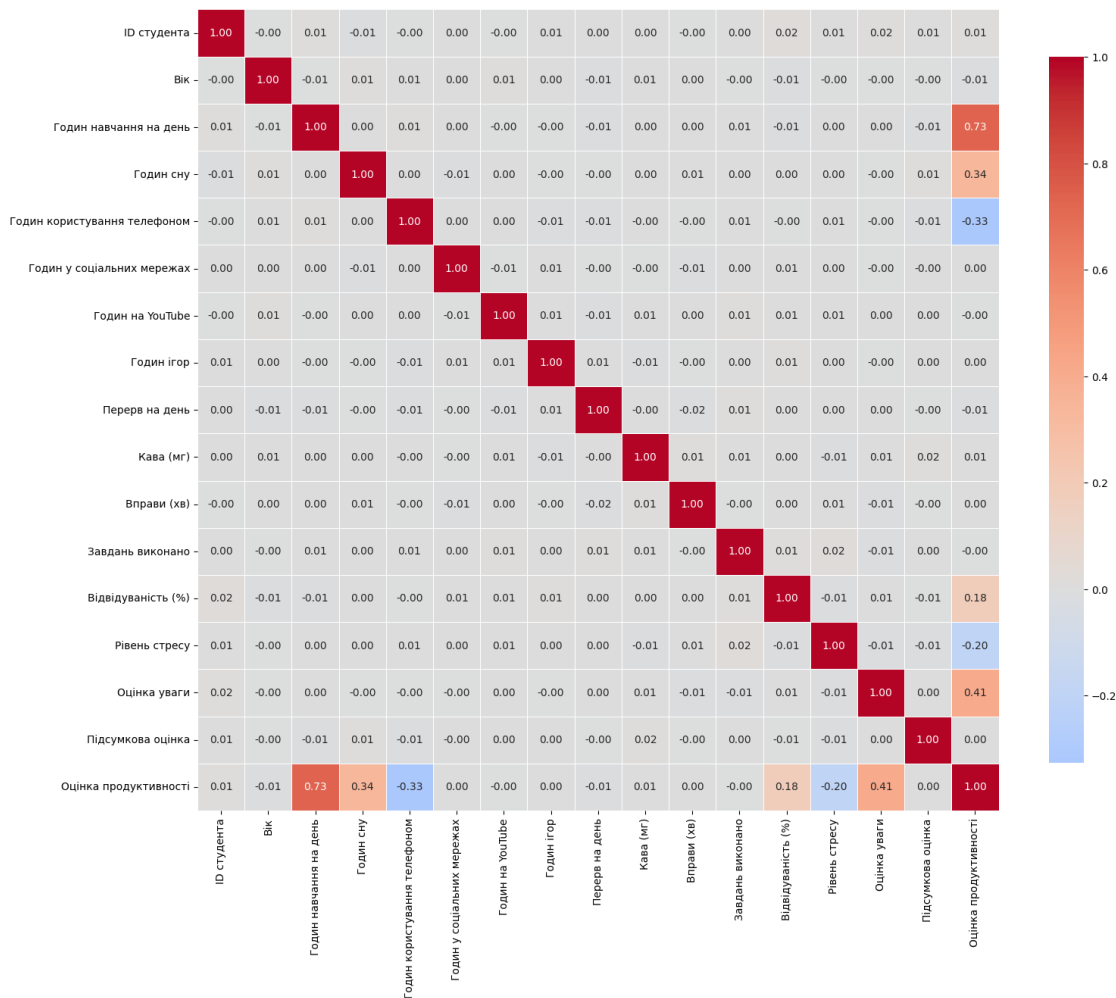


Рис. 5. Кореляційна матриця

Кореляційна матриця, представленої на рисунку 5, показує, що найсильніший позитивний зв'язок спостерігається між продуктивністю та годинами навчання на день (коефіцієнт 0.73), що підтверджує пряму залежність ефективності від часу, присвяченого навчанню. Також вагомий позитивний вплив мають оцінка уваги (0.41) та години сну (0.34). Отриманий у результаті аналізу позитивний коефіцієнт кореляції між годинами сну та продуктивністю узгоджується з висновками дослідників щодо ролі сну в когнітивній діяльності, де стабільний режим відпочинку виступає необхідною умовою для підтримки високої концентрації та засвоєння навчального матеріалу [7]. Це вказує на те, що когнітивний ресурс та фізіологічне відновлення є фундаментальними компонентами високої продуктивності студентів. Також проглядається чіткий негативний вплив факторів відволікання та психологічного тиску. Години користування телефоном мають від'ємну кореляцію (-0.33), що демонструє зниження продуктивності при надмірному використанні гаджетів. Схожа тенденція спостерігається і для рівня стресу (-0.20), що підкреслює деструктивну роль емоційного напруження. Також відсутність сильних кореляцій між вхідними ознаками свідчить про відсутність проблеми мультиколінеарності, що робить цей набір ознак придатним для побудови стабільних моделей машинного навчання.

Висновки

У ході розвідувального аналізу даних було досліджено чинники, що впливають на показник продуктивності (`productivity_score`) студентів, а саме академічні, фізіологічні чинники та рівень цифрових відволікань. Здійснений аналіз підтвердив багатофакторну природу академічної ефективності та показав, що найбільш впливовими факторами, які підвищують продуктивність, є: кількість годин навчання на день (`study_hours_per_day`), показник концентрації (`focus_score`) та години сну (`sleep_hours`). У той час як такі аспекти, як надмірне використання телефону (`phone_usage_hours`) та високий рівень стресу (`stress_level`), суттєво знижують ефективність навчання.

Проведений аналіз розподілів показав, що цільова змінна має близький до нормального розподіл, що свідчить про репрезентативність вибірки та придатність даних до побудови регресійних моделей.

Аналіз поведінкових характеристик студентів у розрізі статі показав відсутність суттєвих відмінностей у розподілі часу на навчання, відпочинок і цифрові активності. Це дозволяє зробити висновок про універсальний характер впливу цифрових факторів на продуктивність студентів незалежно від гендерної приналежності.

Побудовано матрицю кореляції, яка дала змогу кількісно оцінити взаємозв'язки між усіма параметрами та зробити висновки щодо відсутності мультиколінеарності між вхідними ознаками. Для візуалізації розподілу даних відповідно до окремих ознак, зокрема віку (`age`), годин використання соціальних мереж (`social_media_hours`), перегляду YouTube (`youtube_hours`), ігрової активності (`gaming_hours`) та споживання кофеїну (`coffee_intake_mg`), було використано стовпчасті діаграми та гістограми. Проведений аналіз підтвердив відсутність мультиколінеарності між ознаками, що забезпечує стабільність майбутніх моделей машинного навчання. Це створює надійну основу для побудови прогнозних систем, здатних не лише оцінювати поточний рівень продуктивності, а й передбачати його зміни залежно від поведінкових та фізіологічних факторів. Таким чином, дослідження закладає фундамент для подальшої інтеграції методів інтелектуального аналізу даних у сферу освіти та формування інноваційних підходів до управління навчальним процесом.

Таким чином, проведений розвідувальний аналіз даних забезпечив формування інформаційної та аналітичної основи для подальшої розробки інформаційної технології прогнозування продуктивності студентів. Отримані результати можуть бути використані для відбору найбільш інформативних ознак, оптимізації структури навчальних даних та побудови точніших моделей машинного навчання, призначених для раннього виявлення ризиків зниження академічної ефективності та підтримки прийняття управлінських рішень у закладах вищої освіти.

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. Student Productivity & Digital Distraction Dataset [Електронний ресурс] – Режим доступу: <https://www.kaggle.com/datasets/sehaj1104/student-productivity-and-digital-distraction-dataset>.
4. Pandas Getting started. [Електронний ресурс] – Режим доступу: https://pandas.pydata.org/docs/getting_started/index.html
5. Seaborn Tutorial. [Електронний ресурс] – Режим доступу: <https://seaborn.pydata.org/tutorial.html>
6. Matplotlib Pyplot Documentation. [Електронний ресурс]. – Режим доступу: https://matplotlib.org/3.5.3/api/as_gen/matplotlib.pyplot.html
7. Mohammad A Khan, Hamdan Al-Jahdali. The consequences of sleep deprivation on cognitive performance / Mohammad A Khan, Hamdan Al-Jahdali // Neurosciences. – 2023. – Vol. 28, Issue 2, pp. 91-99. doi: <https://doi.org/10.17712/nsj.2023.2.20220108>

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

Войцеховська Ольга Олександрівна – PhD, доцент кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, м. Вінниця, e-mail: olgav1085@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

Voitsekhovska Olha O. – PhD, Associate Professor of the Department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: olgav1085@gmail.com