**КЛАСТЕРИЗАЦІЯ ДАНИХ СТУДЕНТІВ ЗА РЕЗУЛЬТАТАМИ АКТИВНОСТЕЙ НА КУРСІ ДИСТАНЦІЙНОГО НАВЧАННЯ**

**DATA CLUSTERING OF STUDENTS BASED ON ACTIVITY RESULTS IN A DISTANCE LEARNING COURSE**

Науковий керівник: Денисенко Андрій Володимирович

Магістр, Шпинковський Олександр Олександрович, Shpynkovskyi Oleksandr

Студентка, Дришлюк Вероніка Андріївна, Dryshliuk Veronika

Національний університет «Одеська політехніка», УКРАЇНА

***Анотація.*** *Окреслено важливість аналізу активності користувачів ресурсу дистанційного навчання університету.Проведено попередню обробку та анонімізацію даних користувачів навчального онлайн курсу, розроблено та реалізовано алгоритм кластеризаціїданих за допомогою методу K-середніх, а також здійснено інтерпретацію отриманих результатів. Запропоновано подальші заходи для покращення результатів аналізу.*

***Ключові слова:*** *дистанційне навчання, кластеризація, аналіз даних.*

***Abstract.*** *The importance of analyzing user activity on the university's distance learning resource has been outlined. A preliminary processing and anonymization of the data of users of the online educational course was carried out, a clustering algorithm using the K-means method was developed and implemented, and the obtained results were interpreted. Further measures to improve the analysis results have been proposed.*

***Keywords:*** *distancelearning, clustering, dataanalysis****.***

**Вступ.** У сучасних умовах поширення дистанційного навчання особливої актуальності набуває аналіз активності студентів у цифровому освітньому середовищі. Такий аналіз є важливим інструментом для вдосконалення якості навчального процесу, оскільки дозволяє виявляти закономірності у поведінці студентів, оцінювати рівень їх залученості та вчасно реагувати на потенційні проблеми. Сегментація (кластеризація) навчальних даних дає змогу об’єднувати студентів у групи зі схожими моделями поведінки, що відкриває можливості для персоналізації навчання, оптимізації оцінювання та надання цільової підтримки тим, хто її потребує найбільше.

**Мета роботи.** Провести кластеризацію студентів за їхньою активністю на онлайн курсі на основі даних з файлу журналу[1]. Визначити групи студентів за рівнем участі: низька, середня та висока активність. Отримати розуміння патернів взаємодії студентів із освітнім контентом на дистанційному навчальному порталі. Для досягнення мети потрібно виконати наступні завдання:

* анонімізація користувачів;
* обчислення кількості дій (активностей) для кожного користувача;
* використання алгоритму K-means для розбиття користувачів на 3 групи (кластери) за кількістю дій;
* опис та інтерпретація результатів;
* програмний код оформити у відповідності до стилю оформлення програми (ступінчаста структура запису програмного коду, вибір імен ідентифікаторів, використання коментарів тощо);
* продумати оформлення виводу результату.

**Основна частина роботи.** Перед виконанням кластеризації було проведено ретельну підготовку даних, адже якість вхідної інформації безпосередньо впливає на точність та значущість результатів аналізу[2]. Для цього було використано дані з платформи дистанційного навчання, яка фіксує дії студентів під час роботи з освітнім контентом (тестами, форумами, лабораторними, практичними завданнями, лекційними матеріалами тощо).

1. Підготовка даних.

Анонімізація: особисті дані студентів (імена та прізвища) були замінені на унікальні ідентифікатори (наприклад, "User1", "User2") з метою забезпечення конфіденційності.

Вибір характеристик:для подальшого аналізу було обираються найбільш релевантні параметри, що відображають активність студентів — зокрема, кількість дій на платформі, оцінки, кількість спроб тощо.

Попередній аналіз: здійснюється аналіз розподілу даних для виявлення загальних тенденцій та підготовки до кластеризації, що дає змогу краще зрозуміти навчальну поведінку студентів.

2. Вибір методу кластеризації.

Простота та ефективність: алгоритм K-means добре підходить для числових даних і дозволяє швидко сегментувати велику кількість записів [3].

Гнучкість: можливість попередньо визначити кількість груп (у цьому випадку — три: низька, середня та висока активність) дає змогу краще контролювати результат.

Зрозумілі результати: вихідні кластери легко інтерпретувати, що полегшує подальший аналіз і прийняття рішень.

Для реалізації кластеризації було використано мову програмування Python та бібліотеки pandas, NumPy і scikit-learn [4]. Дані було імпортовано з Excel-файлу, що містив інформацію про дії користувачів наплатформі дистанційного навчання.Імена студентів замінено на унікальні ID задля забезпечення конфіденційності.Для кожного користувача підраховано кількість виконаних дій (взаємодій із платформою) та проведено кластеризацію:

У результаті було виділено три кластери що характеризували активність на курсі: низьку, середню та високу. Це дозволило класифікувати студентів за залученістю до навчального процесу та спростити подальший аналіз (таб.1).

Таблиця 1 — Порівняння кількості користувачів та їх середньої

кількості дій в різних кластерах

|  |  |  |
| --- | --- | --- |
| Кластер | Кількість користувачів | Середня кількість дій |
| 0 | 8 | 308 |
| 1 | 7 | 1385 |
| 2 | 2 | 2525 |

На рис. 1 наведено відсортований перелік найбільш активних користувачів дистанційного курсу.

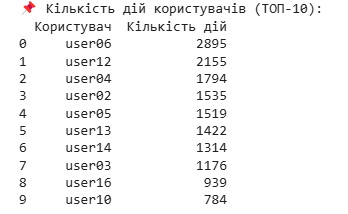


Рисунок 1 – Сортування користувачів за їх активністю на курсі

**Висновки.** У цьому завданні виконано кластеризацію студентів за активністю на платформі дистанційного навчання для аналізу їхньої взаємодії з навчальним матеріалом. Дані були попередньо оброблені: імена анонімізовано, пропущені значення заповнено. Застосовано алгоритм K-середніх, який дозволив виокремити три кластери:

Кластер 0 (низька активність): 8 користувачів, середня кількість дій — 308

Кластер 1 (середня активність): 7 користувачів, 1385 дій

Кластер 2 (висока активність): 2 користувача, 2525 дій

Більшість студентів мали середню активність, найменше — високу, але з найбільшим обсягом дій. Кластеризація допомагає виявляти студентів, які потребують додаткової підтримки чи стимулювання. Подальше розширення параметрів аналізу дозволить точніше адаптувати освітній процес.

Завдяки використанню кластерного аналізу в рамках дистанційних онлайн курсів стало можливим визначити типові моделі взаємодії студентів із системою дистанційного навчання. Це, у свою чергу, може бути використано для індивідуалізації навчального підходу, виявлення студентів, яким потрібна підтримка, та покращення загальної якості освітнього процесу.

**СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ**

1. Теорія ймовірності та математична статистика [Електронний ресурс] : дистанційний навчальний курс / Національний університет «Одеська політехніка». – Режим доступу: https://edu.op.edu.ua/course/view.php?id=12. – Назва з екрана. – Дата звернення: 11.05.2025.
2. Розкриття можливостей кластерного аналізу. URL: https://mindthegraph.com/blog/uk/%D0%BA%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%BD%D0%B8%D0%B9-%D0%B0%D0%BD%D0%B0%D0%BB%D1%96%D0%B7/ (дата звернення: 20.02.2025).
3. Що таке кластеризація K-Means. URL: https://www.unite.ai/uk/what-is-k-means-clustering/ (дата звернення: 20.02.2025).
4. Прокопович, І. В. Використання інтелектуальних технологій у визначенні діагнозу хвороби / І. В. Прокопович, О. А. Шпинковський // Інформ. системи та технології в медицині (ISM-2018) : І Міжнарод. наук.-практ. конф., 28-30 листоп. 2018 р. : зб. наук. пр. / ХНУРЕ. – Харків, 2018. – С. 127–129.