

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ  
Національний університет «Запорізька політехніка»

Факультет комп'ютерних наук і технологій

(повне найменування факультету)

Кафедра програмних засобів

(повне найменування кафедри)

## Пояснювальна записка

до дипломного проєкту (роботи)

магістр

(ступінь вищої освіти)

на тему ДОСЛІДЖЕННЯ ТА ПРОГРАМНА РЕАЛІЗАЦІЯ  
МЕТОДІВ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ  
СТУДЕНТА В СИСТЕМІ ДИСТАНЦІЙНОГО НАВЧАННЯ  
RESEARCH AND SOFTWARE IMPLEMENTATION  
OF METHODS FOR INTELLIGENT ANALYSIS OF STUDENT  
DATA IN THE DISTANCE LEARNING SYSTEM

Виконав(ла): студент(ка) 2 курсу, групи КНТ-142м  
Спеціальності 121 Інженерія програмного  
забезпечення

(код і найменування спеціальності)

Освітня програма (спеціалізація)

Інженерія програмного забезпечення

ЧЕРНИК Д.С.

(ПРІЗВИЩЕ та ініціали)

Керівник ГЛАДКОВА О.М.

(ПРІЗВИЩЕ та ініціали)

Рецензент ЗЕЛЕНЬОВА І.Я

(ПРІЗВИЩЕ та ініціали)

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ  
Національний університет «Запорізька політехніка»

Факультет КНТ  
Кафедра програмних засобів  
Ступінь вищої освіти магістр  
Спеціальність 121 Інженерія програмного забезпечення  
(код і найменування)  
Освітня програма (спеціалізація) Інженерія програмного забезпечення  
(назва освітньої програми (спеціалізації))

**ЗАТВЕРДЖУЮ**

Завідувач кафедри ПЗ, д.т.н, проф.  
Сергій СУББОТІН  
“ ” 2023 року

**З А В Д А Н Н Я**

**НА ДИПЛОМНИЙ ПРОЄКТ (РОБОТУ) СТУДЕНТА(КИ)**

ЧЕРНІКА Дмитра Сергійовича

(ПРИЗВИЩЕ, ім'я, по батькові)

1. Тема проєкту (роботи) Дослідження та програмна реалізація методів інтелектуального аналізу даних студента в системі дистанційного навчання.  
Research and Software Implementation of Methods for Intelligent Analysis of Student Data in the Distance Learning System

керівник проєкту (роботи) к.т.н., доцент, ГЛАДКОВА Ольга Миколаївна,  
(науковий ступінь, вчене звання, ПРИЗВИЩЕ, ім'я, по батькові)

затверджені наказом закладу вищої освіти від “ 14 ” листопада 2023 року № 437

2. Строк подання студентом проєкту (роботи) 01 грудня 2023 року

3. Вихідні дані до проєкту (роботи) рекомендована література, технічне завдання

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити) 1. Аналіз засобів електронного навчання. 2. Big data в електронному навчанні. 3. Дослідження методів інтелектуального аналізу даних студента. 4. Експериментальне дослідження методів.

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень) Слайди презентації

## 6. Консультанти розділів проєкту (роботи)

| Розділ              | ПРІЗВИЩЕ, ініціали та посада консультанта | Підпис, дата   |                           |
|---------------------|-------------------------------------------|----------------|---------------------------|
|                     |                                           | завдання видав | прийняв виконане завдання |
| 1-4 Основна частина | ГЛАДКОВА О.М., доцент                     |                |                           |
| Нормоконтроль       | БСЛОВА А.В., асистент                     |                |                           |

7. Дата видачі завдання “ 08 ” вересня 2023 року.

## КАЛЕНДАРНИЙ ПЛАН

| № з/п | Назва етапів дипломного проєкту (роботи)       | Строк виконання етапів проєкту ( роботи ) | Примітки     |
|-------|------------------------------------------------|-------------------------------------------|--------------|
| 1     | Постановка завдання роботи.                    | 1 тиждень                                 | Завдання, ТЗ |
| 2     | Аналіз засобів електронного навчання           | 1-2 тижні                                 | Розділ 1     |
| 3     | Дослідження методів інтелектуального аналізу.  | 3-4 тижні                                 | Розділ 2     |
| 4     | Алгоритми обробки даних.                       | 5 тиждень                                 | Розділ 3     |
| 5     | Розробка бази даних, розробка програми.        | 6-7 тижні                                 | Розділ 4     |
| 6     | Тестування програмного забезпечення.           | 8 тиждень                                 |              |
| 7     | Оформлення пояснювальної записки та документів | 9-10 тижні                                | Додатки      |
|       | до неї.                                        |                                           |              |
| 8     | Нормоконтроль та рецензування.                 | 11 тиждень                                |              |
| 9     | Захист роботи.                                 | 12 тиждень                                |              |

Студент(ка)

\_\_\_\_\_ Дмитро ЧЕРНІК  
( підпис ) (Ім'я ПРІЗВИЩЕ)

Керівник проєкту (роботи)

\_\_\_\_\_ Ольга ГЛАДКОВА  
( підпис ) (Ім'я ПРІЗВИЩЕ)

## РЕФЕРАТ

Пояснювальна записка до дипломної кваліфікаційної роботи магістра:  
80 с., 7 табл., 21 рис., 2 дод., 50 джерел.

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ, ОНЛАЙН ОСВІТА, ПОВЕДІНКА  
СТУДЕНТА, МЕТОДИ, КЛАСТЕРІЗАЦІЯ, АЛГРИТМ, ПОЧАТКОВІ ДАНІ.

Об'єкт дослідження – аналіз поведінки студента під час проходження курсу електронного навчання.

Предмет дослідження – – методи та алгоритми інтелектуального аналізу даних у системі електронного навчання.

Мета роботи – застосування методів та алгоритмів інтелектуального аналізу даних для дослідження поведінки студента в системі електронної освіти.

Матеріали, методи та технічні засоби: математичні методи кластеризації даних, структурне та об'єктно-орієнтоване програмування, мови програмування Python, персональний комп'ютер з процесором Intel Core 2 Duo під управлінням операційної системи Microsoft Windows 11.

Результати. Створено програмна реалізація методу інтелектуального аналізу поведінки студента.

Висновки. Досліджено моделі аналізу процесу онлайн навчання, а також big data у системах електронної освіти. Описано форму подання даних про студентів, алгоритми вивантаження великих даних та подано логічну модель даних у системах електронної освіти. Проведено програмне моделювання алгоритму CFSFDP-HD, представлено експеримент та порівняння результатів аналізу роботи різних методів.

Галузь використання – онлайн навчання.

## ABSTRACT

Explanatory note to the diploma qualifying work of the master: 80 pages, 7 tables, 21 figures, 2 appendices, 50 sources.

INTELLECTUAL ANALYSIS, ONLINE EDUCATION, STUDENT BEHAVIOR, METHODS, CLUSTERIZATION, ALGORITHM, INITIAL DATA.

Object of study the analysis of the student's behavior during the course of e-learning.

Subject of study is methods and algorithms of intelligent data analysis in the electronic learning system.

The purpose of the work is the application of methods and algorithms of intellectual data analysis for the study of student behavior in the electronic education system.

Materials, methods and technical means: mathematical methods of data clustering, structural and object-oriented programming, Python programming languages, a personal computer with an Intel Core 2 Duo processor running the Microsoft Windows 11 operating system.

The results. A software implementation of the method of intellectual analysis of student behavior has been created.

Conclusions. Analysis models of the online learning process, as well as big data in electronic education systems, were studied. The form of submitting data about students, the algorithms for downloading big data, and the logical model of data in electronic education systems are described. A software simulation of the CFSFDP-HD algorithm was carried out, an experiment and a comparison of the results of the analysis of the work of different methods were presented.

The field of use is online education

## ЗМІСТ

|                                                                                      |    |
|--------------------------------------------------------------------------------------|----|
|                                                                                      | С. |
| Перелік скорочень та умовних позначок.....                                           | 8  |
| Вступ .....                                                                          | 9  |
| 1 Аналіз засобів електронного навчання.....                                          | 12 |
| 1.1 Аналіз систем електронного навчання .....                                        | 12 |
| 1.2 Методи аналізу онлайн-навчання студента.....                                     | 17 |
| 1.3 Алгоритми аналізу даних.....                                                     | 21 |
| 1.4 Висновки за розділом .....                                                       | 24 |
| 2 Big data в електронному навчанні .....                                             | 25 |
| 2.1 Дослідження dig data в сучасних системах електронного навчання .....             | 25 |
| 2.2 Форма подання даних про студента.....                                            | 27 |
| 2.3 Розподілені дані .....                                                           | 31 |
| 2.4 Висновки за розділом .....                                                       | 33 |
| 3 Дослідження методів інтелектуального аналізу даних студента .....                  | 34 |
| 3.1 Системи ETL-процесу .....                                                        | 34 |
| 3.2 База даних у системі електронної освіти .....                                    | 36 |
| 3.3 Інтелектуальний аналіз даних .....                                               | 37 |
| 3.4 Методи інтелектуального аналізу даних в системі електронного<br>навчання.....    | 42 |
| 3.5 Алгоритми інтелектуального аналізу даних в системі електронного<br>навчання..... | 44 |
| 3.6 Програмна реалізація алгоритму інтелектуального аналізу даних .....              | 53 |
| 3.7 Висновки за розділом .....                                                       | 55 |
| 4 Експериментальне дослідження методів.....                                          | 56 |
| 4.1 Проведення експерименту .....                                                    | 56 |
| 4.2 Дослідження результатів.....                                                     | 58 |
| 4.3 Висновки за розділом .....                                                       | 59 |
| Висновки.....                                                                        | 61 |

|                                   |    |
|-----------------------------------|----|
| Перелік джерел посилань .....     | 63 |
| Додаток А Лістинги модулів.....   | 70 |
| Додаток Б Слайди презентації..... | 72 |

## **ПЕРЕЛІК СКОРОЧЕНЬ ТА УМОВНИХ ПОЗНАК**

DWH – Data WareHouse;

ETL – Extract, Transform, Load;

OLAP – Online Analytical Processing;

OLTP – Online Transaction Processing;

АОД – аналіз освітніх даних;

ЕОС – електронні освітні середовища;

ІНС – інформаційне навчальне середовище;

ІТ – інформаційні технології.

## ВСТУП

В даний час електронне навчання – важливий елемент навчальної та викладацької діяльності у переважній кількості коледжів та університетів по всьому світу. У багатьох вищих навчальних закладах основні бібліотечні служби, ІТ та служби підтримки підтримують послуги з електронного навчання. Зі звіту UCISA за липень 2022 р. видно, що основні інформаційні служби надають послуги ІОС та організують його підтримку у 93% вищих навчальних закладів [1].

Електронна освіта містить у собі електронні підручники, освітні послуги та технології. Загалом, сучасні технології в освіті сприймаються студентами позитивно, - знання, вміння та навички, які вони здобувають, використовуються надалі не тільки в кар'єрному зростанні, а й у повсякденному житті [2].

В огляді світового ринку електронного навчання виділяється те, що застосування технологій у навчанні дає безмежні можливості для розвитку та отримання корисних знань, навичок. Здобувати знання у ХХІ столітті стало швидко, захоплююче та доступно. З погляду інтересів держави, електронна освіта надає можливість здобуття однакового рівня освіти людям, які живуть у віддалених точках країни, тобто електронна освіта дозволяє усунути освітню нерівність.

Актуальність дослідження обумовлена необхідністю застосувати методи інтелектуального аналізу даних для оцінки поведінки студента, що дозволить з більшою ймовірністю прогнозувати можливі помилки та підміну даних у системі електронного навчання.

В умовах формування системи безперервної освіти та підвищення кваліфікації, дистанційне навчання стало оптимальною формою випереджуючого навчання та професійної перепідготовки. Розвиток дистанційної освіти потребує спеціальних інформаційних можливостей для забезпечення безперебійної роботи мережі, а також для полегшення викладачів

та студентів.

Завдання, які вирішуються за допомогою методів інтелектуального аналізу даних, можна поділити на описові та передбачувальні [3].

У описових завданнях дається наочний опис наявних прихованих закономірностей, тоді як у передбачуваних вирішується питання про передбачення тих випадків, котрим даних ще немає. Описові завдання:

- пошук правил чи патернів;
- кластерний аналіз;
- знаходження регресійної моделі.

Передбачувальні завдання:

- класифікація об'єктів;
- регресійний аналіз.

В даний час, інтелектуальний аналіз даних є напрямком, що розвивається, за допомогою якого можна поліпшити якість електронної освіти.

Метою роботи є застосування методів та алгоритмів інтелектуального аналізу даних для дослідження поведінки студента в системі електронної освіти.

Об'єкт дослідження – аналіз поведінки студента під час проходження курсу електронного навчання.

Предмет дослідження – методи та алгоритми інтелектуального аналізу даних у системі електронного навчання.

Виходячи з мети дослідження необхідно вирішити такі завдання:

- провести порівняльний аналіз існуючих способів та засобів аналізу поведінки студента під час проходження курсу електронного навчання;
- дослідити можливі критерії порівняння поведінки студентів;
- дослідити методи та алгоритми інтелектуального аналізу даних;
- спроектувати та реалізувати модель інтелектуального аналізу поведінки студента при проходженні курсу електронного навчання;
- перевірити ефективність розробленої моделі та визначити результативність.

Наукова новизна дослідження полягає у використанні методів та алгоритмів інтелектуального аналізу даних для оцінки поведінки студента у системі електронної освіти.

Практична новизна полягає в аналізі поведінки студента при проходженні курсу електронного навчання з використанням методів та алгоритмів інтелектуального аналізу даних, який дозволить з більшою ймовірністю прогнозувати можливі помилки та підміну даних у системі дистанційного навчання.

# **1 АНАЛІЗ ЗАСОБІВ ЕЛЕКТРОННОГО НАВЧАННЯ**

## **1.1 Аналіз систем електронного навчання**

З розвитком інформаційних технологій, електронна освіта набула широкого застосування. Нині навчальна і викладацька діяльність включає у собі ваговий елемент – електронну освіту. Відповідно до звіту Асоціації інформаційних систем університетів та коледжів (UCISA) за липень 2022 р., основні ІТ- служби надають послуги інформаційного навчального середовища (далі ІНС) та організують його супровід у 93% вищих навчальних закладів [4].

Електронна освіта є таким видом освітнього процесу, де навчання відбувається за допомогою цифрових технологій, з використанням електронних підручників, освітніх послуг та технологій. Більшість студентів вітає сучасні технології в освіті – знання, вміння та навички, набуті під час такого навчання, використовуються надалі у самовдосконаленні та кар'єрному зростанні [5].

В огляді світового ринку електронного навчання наголошується, що використання технологій у навчанні відкриває безмежні можливості для розвитку та отримання нових знань та навичок, знімаючи різні бар'єри. Вчитися у XXI столітті стало швидко, захоплююче та доступне. З погляду інтересів держави, електронна освіта – це можливість здобуття однакового рівня освіти людям, які проживають у різних точках країни, тобто спосіб усунення освітньої нерівності [6].

В останні роки кількість досліджень, що вивчають ефективність електронного навчання зросла. Це насамперед пов'язано зі збільшеними можливостями для ІТ та навчання, а також підвищеною увагою з боку політиків та організацій до того, «що працює» у навчанні [7].

Активне впровадження електронної освіти потребує наявності як психологічної та професійної підготовленості всіх учасників даного процесу, так і його теоретичного освоєння та методичного забезпечення. Низький рівень затребуваності надійної професійної освіти на сучасному ринку праці,

зумовлений недостатніми темпами розвитку економіки, впливає на інтерес абітурієнтів, студентів, їх батьків, роботодавців, самих вищих навчальних закладів як освітні послуги та бажання частини учнів мати диплом за мінімальних розумових та фінансових витрат [8].

Наявність кількох способів розуміння ефективності електронного навчання дозволяє професіоналам та дослідникам гнучко вимірювати та визначати ефективність вирішення електронного навчання. Невиконання концепції електронного навчання може призвести до непорозумінь, а аспекти ефективності, які мають найбільшу цінність для учасників та зацікавлених сторін, не можуть розглядатися. Висвітлення багатьох визначень ефективності може призвести до роздумів та натхнення для належного використання концепції ефективності, що дозволить фахівцям з навчання краще узгоджувати свої очікування та спрямовувати свої вимірювальні зусилля на те, що важливо для них та для зацікавлених сторін [9].

Системи електронного навчання набули широкого поширення, що дало дослідникам більший обсяг інформації порівняно з традиційним освітнім процесом. Це з активним використанням у електронному навчанні різних технологій збору даних, і з великим масштабом аудиторії електронних освітніх середовищ (ЕОС). На початку 21 століття з'явився новий напрямок у галузі інтелектуального аналізу даних – аналізу освітніх даних, чому сприяло зростання обсягу даних у системах електронної освіти [10].

Різні методи кластеризації були використані в недавніх дослідженнях для прогнозування успішності студентів. Алана М. та співавт. в [6] було досліджено методи підтримки процесу прийняття рішень вчителем шляхом групування учнів та планування завдань відповідно, було досягнуто позитивного результату.

У [7] автор досліджував різні дослідження та набори даних, що обертаються навколо області ЕОС. Автор дійшов висновку, що ЕОС може використовуватися як частина широкого діапазону зон, включаючи розпізнавання студентів, що наражаються на ризик, виявлення потреб для

адаптації потреб різних груп студентів, збільшення показників випускників.

Ще одне наукове дослідження [11], проведено Амджадом Абу Саа, яке вивчає та прогнозує успішність учнів у різних сценаріях з використанням методів кластеризації. В аналогічному дослідженні [12] проаналізували різні аналітичні методи: інтелектуальний аналіз даних, навчання та академічну аналітику, і дійшли висновку, що застосування методів інтелектуального аналізу даних дає позитивні результати.

K-means - це сучасний алгоритм кластеризації на основі розділів, який застосовується у ЕОС. Інше дослідження, подане Ying, et al. [13] використовував K-means для розуміння поведінки студентів на основі опису 40 студентів. У дослідженні, проведеному Eranki & Moudgalya [14], K-means був застосований вивчення впливу людських характеристик на успішність учня під час прослуховування музики, даючи очевидні класифікації.

Аналізом освітніх даних (АОД) називають напрям досліджень, який спрямовано застосування методів інтелектуального аналізу даних, машинного навчання та статистики до інформації, виробленої освітніми установами.

АОД намагається виявити закономірності з даних, одержуваних у процесі навчання. Ці дані можуть бути дуже широкими і містити велику кількість подробиць. Наприклад, деякі системи управління навчанням (learning management system) відстежують інформацію про те, в який час студент отримав доступ до того чи іншого навчального об'єкта, як часто студенти зверталися до цього об'єкта і як багато хвилин об'єкт відображався на екрані комп'ютера студента, які тестові завдання виконувались першими тощо. Рівень подробиць цих даних такий, що навіть короткий сеанс роботи з ЕОС може зробити великий обсяг даних для аналізу [10].

Інші дані можуть містити набагато менше подробиць. Наприклад, електронна залікова книжка студента складається з упорядкованих у часі списків курсів, які пройшов цей студент, оцінки, отримані за той чи інший курс тощо. АОД використовує обидва вище описані типи даних, щоб виявити закономірності у навчанні студентів. Вивчаючи дані ЕОС, можна помітити

зв'язок між темами, яким навчався приділяв увагу під час проходження курсу, та підсумковою оцінкою цього студента.

АОД тісно пов'язані з аналітикою навчання. У цих дисциплін багато в чому збігаються цілі та завдання. Деякі фахівці бачать різницю між цими напрямками у цьому, що АОД націлений на автоматизації виявлення закономірностей в освітніх даних, тоді як аналітика навчання більшу увагу приділяє підготовці даних як, придатному для аналізу людиною [11].

Досліджувані дані є одним із особливостей АОД. Часто вони мають складну структуру, наприклад, ієрархічну, семантичну, представляючи труднощі для аналізу традиційними методами. Важливо й те, що це можуть бути дані, з якими працівниками сфери освіти зазвичай не стикаються. Аналіз журналів використання сайту або бази даних освітньої системи може надати таку інформацію [5]:

- сторінки, що часто відвідуються; використовувані браузерери;
- кількість відвідуваних сторінок;
- місце, звідки студенти входять до системи;
- кількість відвідувань, їх тривалість для кожного учня;
- популярні ключові слова;
- кількість кліків, переглядів, завантажень матеріалів;
- кількість сторінок, які були переглянуті учням за сеанс;
- список електронних ресурсів, які використовує учень;
- об'єм навчального матеріалу досліджуваного учням перед виконанням завдання.

Такі дані може надати система Moodle. У роботах [9], [10] було описано процес збору даних із Moodle, їх аналіз та деякі результати, отримані за допомогою інтелектуального аналізу даних.

Аналогом Moodle є система iSpring і в ній можна побачити, хто скільки разів і які матеріали переглядав. Відповідно, можна провести аналіз активності студентів (співробітників), популярності матеріалів тощо.

Слід зазначити, що сервіс iSpring Online LMS не є повноцінною

системою керування навчанням, і за його допомогою не можна автоматизувати такі процеси, як контроль знань або керування програмами навчання. Однак при використанні спільно з редактором iSpring Suite можливостей достатньо для створення та доставки навчального контенту в рамках нескладного навчального процесу.

Порівняємо три найпоширеніші системи онлайн навчання (табл. 1.1.).

Таблиця 1.1 – Порівняння систем онлайн навчання Moodle, Canvas і Blackboard

| Система    | Вартість                       | Інтеграція                                         | Зручність користування | Налаштування                         | Підтримка баз даних                       | Користувачський інтерфейс      | Можливості інтелектуальної роботи  |
|------------|--------------------------------|----------------------------------------------------|------------------------|--------------------------------------|-------------------------------------------|--------------------------------|------------------------------------|
| Moodle     | Безкоштовно (Open Source)      | Висока (підтримує багато плагінів та інструментів) | Помірна                | Висока (повністю налаштовувана)      | MySQL, PostgreSQL, MariaDB, MSSQL, Oracle | Налаштовуваний, веббазований   | Плагіни для аналітики та звітності |
| Canvas     | Змінна (залежить від установи) | Висока (широкий спектр інтеграцій)                 | Висока                 | Помірна (деякі опції налаштувань)    | PostgreSQL, MySQL, інші через LTI         | Інтуїтивний, веббазований      | Вбудовані інструменти аналітики    |
| Blackboard | Змінна (залежить від установи) | Висока (широкі можливості інтеграції)              | Висока                 | Помірна (обмежені опції налаштувань) | Oracle, MSSQL, MySQL, інші                | Стандартизований, веббазований | Розширена аналітика та звітність   |

## 1.2 Методи аналізу онлайн-навчання студента

В даний час вивчення поведінки в онлайн-навчання в середовищі великих даних в основному зосереджено на трьох аспектах, як показано на рисунку 1.1 [15].

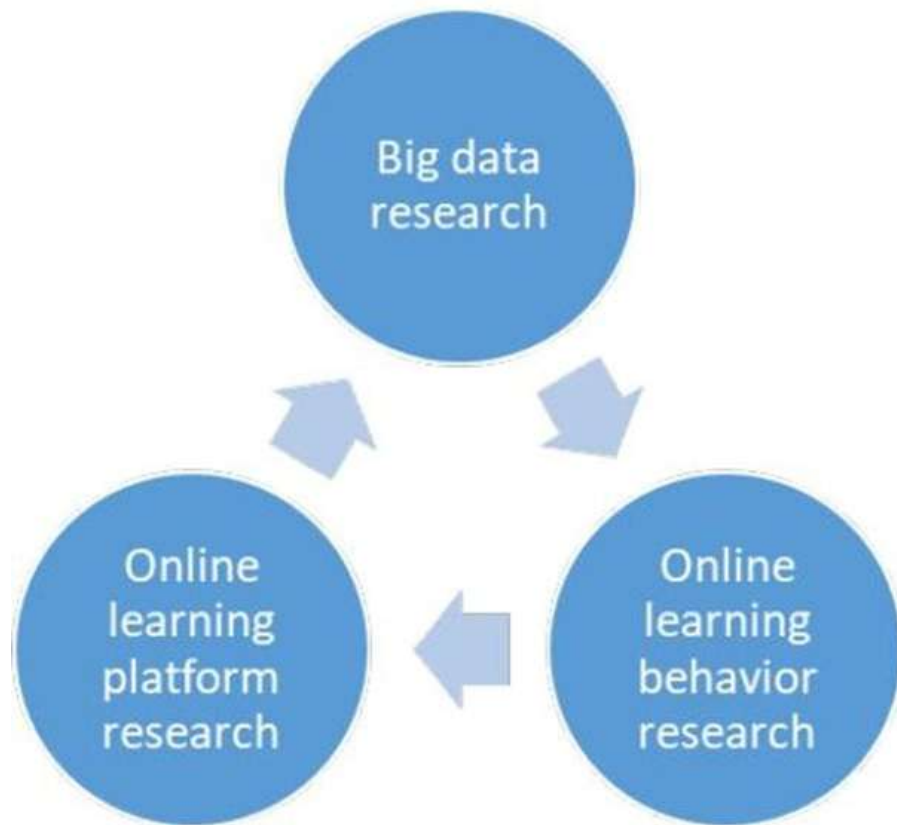


Рисунок 1.1 – Модель аналізу поведінки студента [15]

Big data research -вивчення великих даних. На даний момент основні технології аналізу великих даних включають візуальний аналіз, алгоритми інтелектуального аналізу даних, семантичний аналіз. Наприклад, модель Google Mapreduce орієнтована великі набори даних. Університет Пердью використовує технології великих даних для створення механізмів раннього прогнозування результатів, збираючи дані про студентів курсу. Основна увага сучасних дослідників до вивчення великих даних - вилучення цінності їх [16].

Online learning platform research – дослідження онлайн-платформи навчання. Платформи онлайн-навчання можна розділити на дві категорії: комерційні та некомерційні. Наприклад, Moodle є найбільш широко

використовуваною у світі безкоштовною платформою для онлайн-навчання, яка не тільки використовується в багатьох університетах світу, але пропонує різноманітні інструменти онлайн-спілкування, всілякі підручники для навчання. В останні роки типовим представником платформи онлайн-навчання є масовий відкритий онлайн-курс, також відомий як MOOC.

Online learning behavior research - вивчення поведінки в онлайн-навчання. Багато досліджень пов'язано з поведінкою, об'єкт дослідження та вибір зразка мають певні обмеження, в той час як деякі дослідження щодо поведінки в процесі онлайн-навчання проводилися за участю дослідників [17].

Кеннет вивчав вплив навчальної поведінки та рефлексивного навчання на онлайн-бізнес-курсах. Пріор розділив вплив онлайн-навчання на три аспекти: ставлення до навчання, інформаційна грамотність та самоефективність, Бучер вивчив різні рівні початкових знань та співвідніс результати навчання. Це показало, що високий рівень початкових знань може призвести до того, що учні набудуть глибшого рівня навчання [12].

Наслідуючи моделі аналізу та моделі процесу онлайн-навчання, модель аналізу ділиться на три частини: кластерний аналіз, аналіз рекомендацій та кореляційний аналіз. Відповідно до процесу вирішення проблем, процес моделі аналізу поведінки в режимі онлайн навчання поділяється на обробку даних, вибір методу та процес аналізу, виведення результатів і так далі [18].

На першому етапі збирається інформація про студентів, яка зберігається в системі електронного навчання, далі проводяться горизонтальний та логічний аналізи та здійснюється виведення результатів.

Також застосування інтелектуального аналізу даних у системах електронного навчання можна охарактеризувати як ітеративний цикл, в якому додатки інтелектуального аналізу даних роблять внесок у покращення навчання, а також використовують корисні знання для прийняття рішень. Процес інтелектуального аналізу даних представлений на рисунку 1.2.

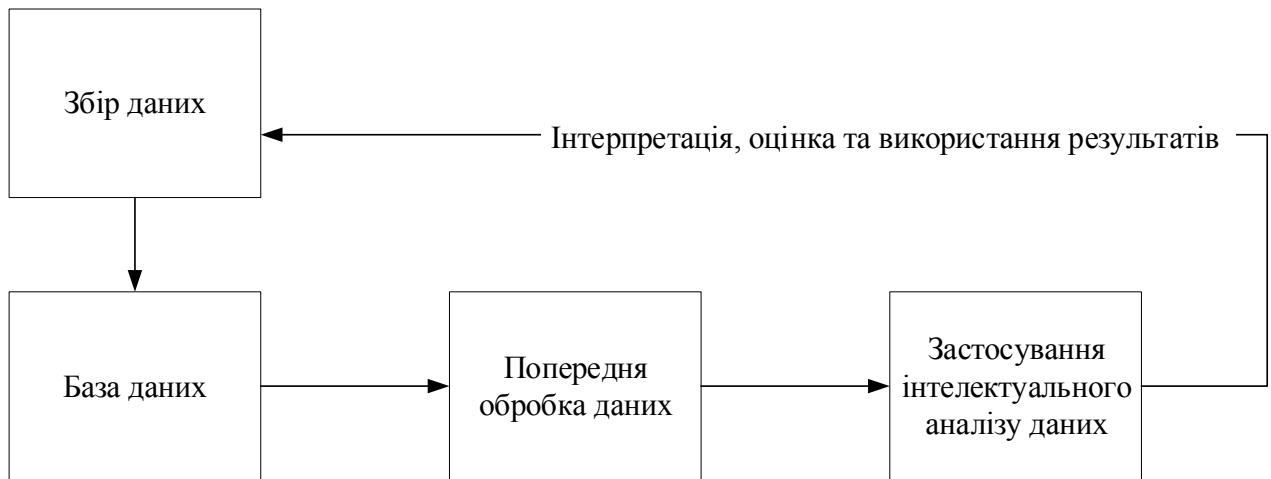


Рисунок 1.2 – Інтелектуальний аналіз даних

Розглянемо етапи докладніше.

- збір даних. Інформація про використання та взаємодію зберігається в базі даних. Інформація про взаємодію зберігається у базі даних;
- попередня обробка даних. Дані очищаються та перетворюються на відповідний формат для аналізу. Для попередньої обробки даних можна використовувати інструмент адміністратора бази даних або конкретний інструмент попередньої обробки. Дані перетворюються на відповідний формат;
- застосування інтелектуального аналізу даних. Алгоритми інтелектуального аналізу даних застосовуються для побудови та виконання моделі, яка виявляє та узагальнює знання, що становлять інтерес для користувача (вчителя, учня, адміністратора тощо). Досягнення цієї мети використовується інструмент інтелектуального аналізу даних;
- інтерпретація, оцінка та використання результатів. Отримані результати чи модель інтерпретуються та використовуються вчителем для подальших дій. Викладач може використовувати інформацію для прийняття рішень щодо студентів з метою покращення навчання студентів. Отримана модель інтерпретується та використовується викладачем для подальшого аналізу.

У 2020 році ЮНЕСКО опублікувала звіт, в якому детально розповідається, як навчальна аналітика обіцяє «перетворити освітні дослідження на науку, засновану на даних, а освітні заклади – на організації, які приймають рішення на основі фактичних даних». За останні кілька років багато з цих заходів вже почали застосовувати [17].

Поведінка в режимі онлайн навчання як багатовимірна складна система, що передбачає елементи, що впливають на детальний аналіз, для визначення характеристик навчання, мотивації та стилю навчання, щоб допомогти вчителям та менеджерам розробити розумну структуру навчання та стратегії навчання.

При аналізі чинників впливу також враховуються як внутрішні чинники учня, а й зовнішні чинники, включаючи середовище навчання, систему підтримки та режим навчання. Внутрішні фактори включають інформаційну грамотність учня, мотивацію навчання, початкові знання предмета учня, стиль навчання і самоефективність учня. До зовнішніх факторів належать моделі онлайн-навчання, ресурси для онлайн-навчання, системи підтримки онлайн-навчання та навички викладання.

Дослідження поведінки в системі онлайн навчання включає персональні вимоги, технічні вимоги і робота з даними. Персональні вимоги включають вимоги учнів, організаторів курсів, менеджерів платформ і дослідників в галузі освіти. Технічні вимоги включають аналіз даних про освіту, аналіз навчання і так далі. Робота з даними включає збір даних і зберігання даних [18].

Аналіз поведінки, пов'язаної з навчанням в режимі онлайн, в основному ділиться на явне і неявне. Серед них більшість речей, які можуть бути записані платформою – це явна поведінка при навчанні, така як перегляд, пошук, збереження і т.д. Неявне мотивація студента до навчання, старання, заучування, тренування, вивчення додаткової інформації. Поведінка в режимі онлайн навчання як складна система, ми можемо класифікувати її за декількома вимірами, а потім сформувати повну та всеосяжну систему

класифікації поведінки студентів у системах онлайн навчання.

Наприклад, є підхід, у якому створюється профіль студента з допомогою технології обробки великих даних. По-перше, аналізуються характеристики учнів та чинники, що впливають поведінка. Потім розраховується подібність поведінки студентів та використовуємо алгоритм коефіцієнта Жакара для класифікації студентів. Нарешті створюється профіль студента, а також візуальний аналіз [19].

Курс електронного навчання потребує певної мети, внутрішнього мотиву, синхронного зворотного зв'язку та незалежності учнів. Профіль студента допомагає студенту зрозуміти його ситуацію, знайти свої проблеми та покращити рівень проходження онлайн-курсів. З безперервним накопиченням даних про освіту та всебічний розвиток, профіль студента зобов'язаний сприяти здоровому розвитку електронного навчання [20].

Удосконалення програмних інструментів, призначених для аналізу даних, дозволило залучати до досліджень фахівців, які не мають досвіду в програмуванні, що стало головним фактором, який сприяв розвитку АОД.

### **1.3 Алгоритми аналізу даних**

За даними [20] значна частина статей, які були опубліковані в журналах Intelligent Tutoring Systems, та User Modeling and Adaptive Personalization, доповіді на конференціях з АОД та аналітики навчання застосовують такі пакети, що вільно розповсюджуються, як: RapidMiner, R, Weka, KEEL і SNAPP. Дані пакети містять алгоритми, які реалізують описані вище методи аналізу даних, також забезпечують імпорт і підтримку попередньої обробки даних для застосування в рамках цих методів. Забезпечується підтримка проведення статистичної перевірки адекватності моделі та візуалізація даних.

Небагато з цих пакетів мають відкритий вихідний код, але незважаючи на це, дослідники розробляють модулі, які розширюють базові можливості, додають нові і нові функції, яких потребують працівники освітніх установ.

Поширений інструмент для статистичного аналізу даних R зараз має близько 10000 пакетів-розширень. Серед них є інтерфейси для інших розвинених пакетів аналізу даних. Також Weka і RapidMiner мають розширення, які дозволяють використовувати можливості R. Відкритий програмний код дає можливість глибоко вивчити методи та алгоритми, які використовуються іншими дослідниками. Розглянемо деякі пакети докладніше.

RapidMiner - інструмент, який був створений для інтелектуального аналізу даних, маючи на меті, що аналітик не повинен програмувати. У RapidMiner включений великий набір операторів, за допомогою яких вирішуються широкі спектри завдань отримання та обробки інформації з різних джерел [21].

R - програмне середовище обчислень з відкритим вихідним кодом, яке було створено в рамках проєкту GNU. На початку R був реалізований співробітниками Оклендського університету Россом Айхекою та Робертом Джентлменом. На даний момент мова та середовище підтримуються та розвиваються організацією R Foundation [22].

Weka – програмне забезпечення, що вільно розповсюджується для аналізу даних та машинного навчання, написане мовою програмування Java в університеті Уайкато. Weka - це унікальний набір засобів для візуалізації та алгоритмів інтелектуального аналізу даних, а також вирішення завдань прогнозування, з включеною графічною оболонкою для доступу до них [23].

Orange - крос-платформний інструмент, призначений для машинного навчання та інтелектуального аналізу даних. Він об'єднує у потужні робочі процеси візуальне програмування, як інтуїтивний засіб об'єднання аналізу даних, та інтерактивні методи візуалізації [24].

Knime (Konstanz Information Miner) - Java-додаток з відкритим вихідним кодом, крос-платформний. Широко використовується для інтелектуального аналізу даних та їхньої оптимізації. Його можна завантажити як основна програма (Knime Desktop), так і весь SDK, заснований на Eclipse Helios [25].

Neural Designer – це програмний інструмент для розширеної аналітики. Він включає інструменти для описової, діагностичної, передбачувальної і приписуючої аналітики. Метою є отримання доступу до дій, що призводить до розумніших рішень та кращих результатів у бізнесі [14].

Аналіз наведених алгоритмів показано в таблиці 1.2.

Таблиця 1.2 – Порівняння алгоритмів аналізу даних

| Алгоритм   | Гнучкість в програмуванні                       | Розширення / Пакети                   | Підтримка алгоритмів | Візуалізація даних | Глибокий аналіз алгоритмів |
|------------|-------------------------------------------------|---------------------------------------|----------------------|--------------------|----------------------------|
| RapidMiner | Модульний підхід без необхідності програмування | Велика кількість операторів           | Широкий спектр       | Висока             | Обмежений                  |
| R          | Висока, з можливістю складного скриптування     | ~10,000 пакетів-розширень             | Екстенсивна          | Висока             | Висока                     |
| Weka       | Обмежена, зосереджена на Java                   | Розширення через Java, інтеграція з R | Широкий спектр       | Висока             | Обмежений                  |
| KEEL       | Обмежена, з фокусом на дослідженні              | Модулі та плагіни                     | Широкий спектр       | Висока             | Обмежений                  |
| SNAPP      | Не застосовується                               | Не застосовується                     | Обмежена             | Обмежена           | Не застосовується          |

Проте, існуючі програмні розробки можуть бути доопрацьовані для того, щоб відповідати вимогам для використання в АОД співробітниками, які не мають досвіду у програмуванні.

#### **1.4 Висновки за розділом**

Дослідження підкреслює важливість інтелектуального аналізу даних у контексті дистанційного навчання. Цей підхід дозволяє ефективно аналізувати поведінку студентів, їхні потреби та успішність, що сприяє підвищенню якості освітнього процесу.

Робота демонструє розвиток та впровадження передових методів інтелектуального аналізу даних, зокрема, використання алгоритмів кластеризації та машинного навчання. Це дає змогу більш точно визначати потреби студентів та прогнозувати їхню успішність.

Практична новизна дослідження полягає у розробці та впровадженні математичних моделей для оцінки поведінки студентів у системі дистанційного навчання. Це дозволяє з більшою точністю ідентифікувати потенційні проблеми у навчальному процесі та покращувати методики навчання.

Дослідження надає значний внесок у розвиток області інтелектуального аналізу даних в освіті, відкриваючи нові можливості для подальших досліджень. Перспективи включають поглиблене вивчення взаємодії між різними методами аналізу даних та їх застосування в різних освітніх контекстах.

## **2 BIG DATA В ЕЛЕКТРОННОМУ НАВЧАННІ**

### **2.1 Дослідження dig data в сучасних системах електронного навчання**

Типові дослідження великих даних навчання - це аналіз даних освіти і аналітика навчання студентів. Інтелектуальний аналіз даних визначається як «нова дисципліна, пов'язана з розробкою методів дослідження унікальних типів даних, що надходять з навчальних закладів, та використанням цих методів для кращого розуміння учнів та умов, у яких вони навчаються». Аналітика учнів має різні визначення через різні доступні інструменти аналітики з акцентом на використання можливостей для моделювання, де успіх учнів є однією з головних проблем.

Великі дані можуть приносити користь студентам, надаючи їм сучасну та динамічну систему освіти. У дослідженні [15] Атанасіос і Панагіотіс проаналізували цілі, завдання та переваги великих даних та відкритих даних в освіті. Автори дійшли висновку, що система освіти може бути покращена шляхом використання нових підходів до навчання, щоб зробити її більш ефективною та сфокусованою на ній. Понад те, в роботі автори [16], підтримують ту ж ідею і очікували, що великі дані можуть бути ефективно використані для прогнозування результатів учнів, а також покращення як викладання, так і навчання. Дослідження, проведене Туласи [17] і Беном Даніелем [18], було спрямоване на вищу освіту та вивчало рішення, пропоновані системами великих даних, для вирішення проблем, що стоять перед вищою освітою. Кріс Деде [19] продовжив просування теми, вивчивши «наступні кроки», які можуть бути зроблені з використанням великих даних в освіті, і дійшов висновку, що ця галузь має великий потенціал для покращення індивідуального досвіду навчання.

Численні дослідники стверджують, що персоналізація в академічному середовищі дозволяє використовувати більш ефективні та життєздатні форми навчання. Різні роботи намагаються покращити якість та життєздатність

електронного навчання, використовуючи стандарти інших дослідницьких зон. Ця модель просування персоналізації додатково проявляється у електронному навчанні [20] представили новий інструмент: Інтелектуальний вебучитель (IWT) для підтримки персоналізованого електронного навчання у їх дослідженні персоналізованого електронного навчання. Порівняння традиційних методів з IWT дозволяє зробити висновок, що персоналізація дозволяє використовувати більш ефективні та потужні форми електронного навчання, демонструючи зростаючий рівень виконання як викладачем, так і студентами. Система розроблена в роботі [21] і Moodle, виявилось здатним покращити навчальний процес і співпрацю між викладачем і студентом у вищій освіті.

В роботі [22] використовували індивідуальні можливості навчання студентів, щоб представити методика, засновану на онтології, для покращення адаптивної схеми електронного навчання. Запропонована система електронного навчання створює контент, що дозволяє учню дізнатися про персоналізовану систему електронного навчання, і стверджують, що система заохочує навчання завдяки реалізації індивідуальних потреб. Результати узгоджуються із дослідженням [23], яке представляє теоретичні основи адаптивного електронного навчання, самооцінки та теорії динамічних лісів. Система надає адаптовані навчальні матеріали для студентів на основі здібностей студентів. Модель сіткового агента була запропонована Жень та Юїном у їх дослідженні [24] для ефективної адаптації систем електронного навчання, що використовують штучну психологію, до окремих студентів, яким буде корисна ця персоналізація. Крім того, автори [25] запропонували ще одну персоналізовану систему електронного навчання, яка є екстрактором зворотного зв'язку з можливістю об'єднання налаштувань користувача.

Значення вищезгаданої літератури у тому, що персоналізовані схеми електронного навчання є ефективними інструментами індивідуального навчання. В електронному навчанні постійно генерується значний обсяг даних, які доступні в Інтернеті. Таким чином, потрібні складніші та прикордонні

методи кластеризації для порівняння даних EDM для отримання внутрішньої інформації. Щоб упоратися з вищезазначеними проблемами, рекомендується кластеризувати підхід до інтелектуального аналізу даних та інтегрувати його з персоналізованою системою електронного навчання. Інтеграція підходів інтелектуального аналізу даних робить систему навчання більш цікавою [26].

## **2.2 Форма подання даних про студента**

Всі дані про студентів займають значний обсяг, для їхньої обробки потрібен конкретний підхід. Під великими даними розуміється широке розмаїття масивів даних, які можуть бути належним чином оброблені традиційними додатками через свого величезного обсягу чи складного складу [27].

Спочатку опишемо, в якому вигляді зберігаються дані про студентів. Дані про студентів у системах електронної освіти найчастіше зберігаються у вигляді журналів сервера (log файли), що дозволяє скоротити обсяг пам'яті, що витрачається.

Студент взаємодіє та керує своїм профілем через інтерфейс, розгорнутий на ноутбучі чи смартфоні. Профіль користувача та інша інформація рідко змінюються через Інтернет. Відповідно до [26], [27], профіль студента відноситься до типової групи студентів. Функція профілю полягає в автоматичному визначенні потреб та переваг студента. Дані, пов'язані зі студентами, працюють як джерело для персоналізації студентських запитів та інтелектуальної відповіді на запити. Профілювання учнів - це безперервний процес, що містить як статичні, і динамічні дані.

Дані, зібрані статичним способом, включають особисті, особистісні, когнітивні, педагогічні дані та дані про переваги (табл. 2.1).

Таблиця 2.1 – Дані про студента

| Вид даних        | Опис                                    |
|------------------|-----------------------------------------|
| Особисті дані    | Біографічну інформацію про студентів    |
| Профіль студента | Загальний інтерес та поведінку студента |
| Когнітивні дані  | Про пізнання учнів                      |
| Педагогічні дані | Стилі та методи навчання                |

Особисті дані сприяють розвитку у студентів уваги, навичок спілкування. Профіль студента відображає загальний інтерес та поведінку студента. Когнітивні дані інформують про пізнання учнів, а педагогічні дані описують різні стилі та методи навчання. Якщо система підтримки профілю виявляє якусь незвичайну поведінку у діяльності учнів, вона відповідним чином оновлює профіль.

Типовий профіль студента може включати в себе відомості про його ім'я, вік, адресу, соціально-економічний статус, школу, про результати навчання, пройдений курс, вивчені модулі, результати іспитів, присвоєного ступеня.

У таблиці 2.2 наведені функції системи електронного навчання.

Результати аналізу зберігаються як набору шаблонів. Можна використовувати описи найбільш поширених послідовностей даних для прогнозу наступного найбільш ймовірного кроку в новій послідовності. Прогнозуючі запити можна налаштовувати для того, щоб вони повертали змінну кількість прогнозів або описові статистичні дані [28].

В базі зберігаються всі особисті дані студента, включаючи ім'я, вік, стать, адресу, поштовий індекс, а також дані, що стосуються освіти, такі як кваліфікація. Крім того, наявність такої інформації, як досвід роботи, кар'єрні цілі, діапазон доходу, попередні курси і курси, що цікавлять, були б дуже корисні для передбачення майбутньої поведінки різних класів людей, які влаштовують трійки. Крім того, інша інформація, така як особисті інтереси та хобі, була б дуже корисною для інструмента інтелектуального аналізу даних,

щоб виявляти приховані шаблони, створюючи інтелектуальні моделі, засновані на величезній кількості даних [29].

Таблиця 2.2 – Функції системи електронного навчання

| З боку системи електронної освіти<br>профіль студента                                                                                                                                                                                                                                                                                                                                                               | З боку студента електронної освіти<br>профіль студента                                                                                                                                                                                      |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <ul style="list-style-type: none"> <li>– обробляє інформацію про студента;</li> <li>– збирає дані для отримання відповідної інформації про рівень знань студента;</li> <li>– обробляє відгуки від студентів, щоб уточнити процес рекомендації навчальних матеріалів;</li> <li>– забезпечує динамічну оцінку;</li> <li>– рекомендує навчальний матеріал для кожного учня на основі його знань та профілю.</li> </ul> | <ul style="list-style-type: none"> <li>– відповідає за всі транзакції, які можуть відбуватися між мережею та комп'ютером;</li> <li>– відповідає за спостереження за поведінкою учня та зберігає його в модулі профілювання учня.</li> </ul> |

Розглянемо приклад на основі Moodle. Інформація з файлів дозволяє наставникам бачити, які ресурси використовуються і коли. Наприклад, викладач може перевірити, що окремий студент прочитав лекцію та подивитися, скільки часу він на це витратив [30].

Вбудоване меню використовується для встановлення фільтрів (рис. 2.1). Вчитель може розглянути звіти на рівні курсу, рівні користувача та/або рівні активності у певний день чи усі дні.

## ▼ Місця показу блоку

Початкове розміщення блоку



Курс: 1. Інструкція для факультетських менеджерів

Відобразити на типах сторінок

Регіон за замовчуванням



Головна сторінка курсу з будь-яким форматом ▾

Довільна сторінка

Довільна сторінка курсу

Головна сторінка курсу з будь-яким форматом

Рисунок 2.1 – Вбудоване меню для складання звітів

Фільтр можна налаштувати за кількома параметрами та переглянути результати, а потім вивантажити інформацію у вигляді текстового файлу. Існуючі фільтри:

- фільтр курсу. За промовчанням використовується поточний курс. Вчитель повинен мати права на курси, що запитуються;
- фільтр студентів. За замовчуванням усі користувачі. Список користувачів, що випадає, дозволяє вибрати конкретного користувача;
- фільтр активності. За промовчанням всі дії. Список, що випадає, може звузити фільтр до однієї дії;
- фільтр дій. За промовчанням всі дії. Ви можете обмежити дію переглядом, оновленням, видаленням або всіма змінами;
- фільтр відображення чи завантаження. За промовчанням відображаються результати фільтра. Журнал може бути завантажений як текст, ODS або Excel.

Звіти можуть допомогти визначити, які типи ресурсу найпопулярніші, а які ні. Це допоможе забронювати ресурси на майбутні курси. Перегляд окремих звітів про виконану роботу може допомогти контролювати студентську діяльність [31].

Сторінка «Звіт» дозволяє вчителю або адміністратору переглядати журнали курсу або сайту та звіти про активність користувачів. Звіт можна знайти у блоці адміністрування. Звіти сайту доступні користувачам, яким призначено адміністративну роль.

Журнали, що відображаються, показують активні посилання на інші частини курсу. До них належать профіль користувача, конкретна сторінка або посилання на ресурс.

Раніше ми позначили, що всі дії студентів реєструються у журналах, але як обробляти дані, які не мають структури. Першим кроком для аналітики електронного навчання є об'єднання всіх даних, які були зібрані, та їх «відчистка» від інформації, що не несе цінності.

Сховище даних є найкращим рішенням для зберігання та обробки даних. Збираючи всю інформацію у зручній формі та одному місці, стає можливим аналізувати та отримувати інформацію про показники студентів [32].

Для вирішення цього завдання є моделі розподіленої обробки даних, розглянемо найпопулярніші з них.

### **2.3 Розподілені дані**

У більшості випадків при завантаженні даних з джерел виникає необхідність зберігання цих даних в одній системі та їх передачі в іншу систему для подальшої обробки та аналізу. З метою цього процесу створюються сховища даних (ХД чи DWH — Data Warehouse). Сховище даних є базою даних для збору та обробки інформації, що надходить з різних джерел. Специфікою сховищ даних є орієнтованість на підготовку звітів для підтримки прийняття рішень в організаціях [33]. Оскільки в даний час існує велика кількість технологій і систем, різних за своєю структурою даних і архітектурою, дані при передачі з однієї системи в іншу, потребують перетворення. Грунтуючись на сказаному вище, можна виділити три послідовні етапи процесу роботи з даними (рис. 2.2):

- вилучення (Extract);
- перетворення (Transform);
- завантаження (Load).

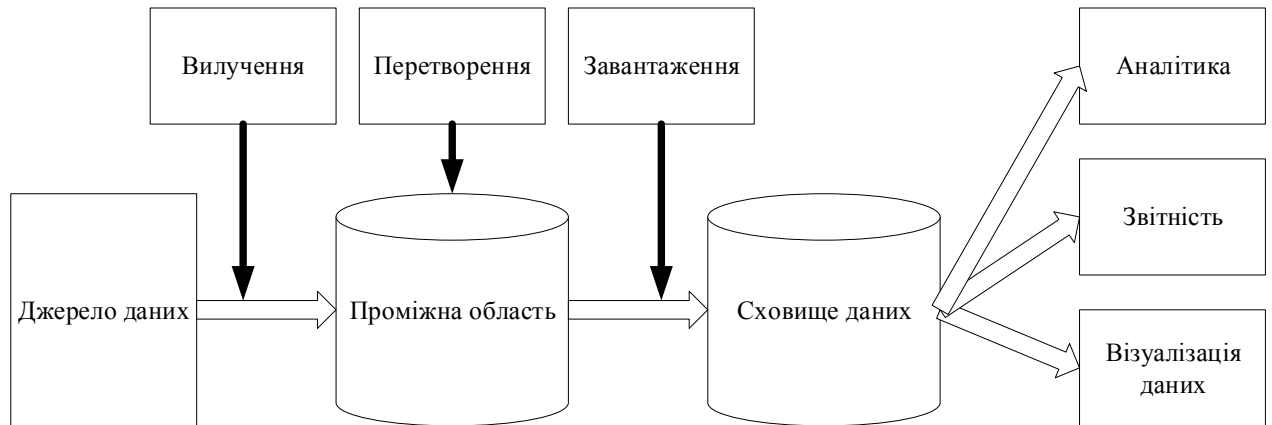


Рисунок 2.2 – Процес ETL

Ці етапи прийнято позначати аббревіатурою ETL, яка описує один з основних процесів управління даними при їх вилученні з джерел та подальшого завантаження у сховище даних з метою отримання достовірної інформації [34]. Зупинимося більш докладному розгляді вищевказаних етапів (табл. 2.3).

Таблиця 2.3 – Процес ETL

| Етап                | Функції                                                                                                                                                                                                                                               |
|---------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Вилучення даних.    | Етап полягає у вилученні даних із джерел, при цьому дані не змінюються незалежно від їх якості та завантажуються в проміжну область.                                                                                                                  |
| Перетворення даних. | У процесі перетворення дані пересмикуються у групування, а також перетворюються в потрібний формат, згідно з структурою сховища даних. Також виконується очищення даних, перевірка на повноту та формуються звіти для подальшого виправлення помилок. |
| Завантаження даних. | На цьому етапі здійснюється завантаження трансформованих даних із проміжної області у сховище даних.                                                                                                                                                  |

Варто зазначити, що завантажуються не всі дані, а тільки ті, які є новими або були змінені. Під час завантаження підтримується версійність. Це необхідною умовою отримання актуальної версії запису на довільну дату. Досить часто ETL стає проміжним шаром між системами класу OLTP та сховищем даних або OLAP-системою [35].

Online Transaction Processing (OLTP) – відносні невеликі транзакційні системи, що обробляють великі потоки даних у реальному часі.

Online analytical processing (OLAP) – системи динамічної побудови звітів та документів, складання складних запитів до бази даних для глибокого аналізу.

## **2.4 Висновки за розділом**

У другому розділі було досліджено моделі аналізу процесу онлайн-навчання, а також big data у системах електронної освіти. Було описано форму подання даних про студентів - у вигляді журналів сервера (log файли) (Автор, рік). Також подано логічну модель даних у системах електронної освіти (Автор, рік).

Таким чином, досліджено та систематизовано підходи до збору, зберігання та обробки великих даних у системах електронної освіти. Це дозволить ефективно аналізувати дані про навчальну діяльність студентів для подальшого застосування методів інтелектуального аналізу.

### **3 ДОСЛІДЖЕННЯ МЕТОДІВ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ СТУДЕНТА**

#### **3.1 Системи ETL-процесу**

Для здійснення ETL-процесу можна використовувати майже будь-яку сучасну мову програмування. Однак, якщо потрібна не разова конвертація, а постійно виконувати інтеграцію даних, доцільно розглянути спеціалізоване ПЗ. При цьому варто враховувати швидкість, розширюваність та масштабованість вибраного інструменту. Серед лідерів на ринку ETL-інструментів виділяються Oracle, Informatica та IBM. Зазвичай системи, створені вказаними компаніями, перекривають потреби більшості компаній у галузі ETL. Виходячи з цього, доцільно вибирати ETL-інструмент ґрунтуючись на поставлених завданнях, а також існуючій платформі компанії [36].

У разі переважання продуктів IBM варто звернути увагу на рішення Data Stage і Data Manager. Відмінною особливістю є наявність OLAP, що дозволяє будувати складні сховища даних.

При переважанні Unix операційних систем доцільними є рішення PowerCenter і PowerMart від Informatica. Варто відзначити можливість розробки мовою Java [30].

При переважанні продуктів Microsoft доцільно застосування SSIS від Microsoft. Зазначений продукт має розвинений користувальницький інтерфейс, проте не є кросплатформним [37].

У таблиці 3.1 наведено порівняння ключових ETL-інструментів від IBM, Informatica, та Microsoft.

Наразі системи, які належать до ETL, є не тільки вирішенням проблем, що виникають при введенні даних, їх перенесенні, що виникають при появі системних помилок або різниці між інформаційними системами, які є джерелами та споживачами даних. Варто зазначити, що ці завдання дуже важливі, оскільки у разі наявності великої кількості різних ІС ймовірно виникнення помилок при передачі даних між системами, що призводить до

зниження якості та достовірності даних. ETL-системи з часом стали включати ширший зміст, оскільки розв'язувані такими системами завдання отримали високий пріоритет у компаніях. Зрештою, бізнес дійшов висновку про важливість впровадження подібних інформаційних систем з метою підвищення швидкості прийняття рішень та ведення успішного конкурентного протистояння. Резюмуючи вищесказане, можна дійти невтішного висновку у тому, що розробники описаних систем повинні прагнути вдосконалення розроблюваних систем і підвищення поширеності таких систем над ринком [38].

Таблиця 3.1 – Порівняння ключових ETL-інструментів від IBM, Informatica, та Microsoft

| ETL Tool                  | Інтеграція з IBM | Інтеграція з Unix | Інтеграція з Microsoft | Java | OLAP | Кросплатформеність | Користувачський інтерфейс |
|---------------------------|------------------|-------------------|------------------------|------|------|--------------------|---------------------------|
| Data Stage (IBM)          | Так              | Ні                | Ні                     | Ні   | Так  | Ні                 | Продвинутий               |
| Data Manager (IBM)        | Так              | Ні                | Ні                     | Ні   | Так  | Ні                 | Продвинутий               |
| PowerCenter (Informatica) | Ні               | Так               | Ні                     | Так  | Ні   | Так                | Продвинутий               |
| PowerMart (Informatica)   | Ні               | Так               | Ні                     | Так  | Ні   | Так                | Продвинутий               |
| SSIS (Microsoft)          | Ні               | Ні                | Так                    | Ні   | Ні   | Ні                 | Продвинутий               |

### 3.2 База даних у системі електронної освіти

Після перетворення журналів сервера на базу даних, маємо справу зі структурованими даними. Розглянемо базу даних системи електронної освіти (рис. 3.1).

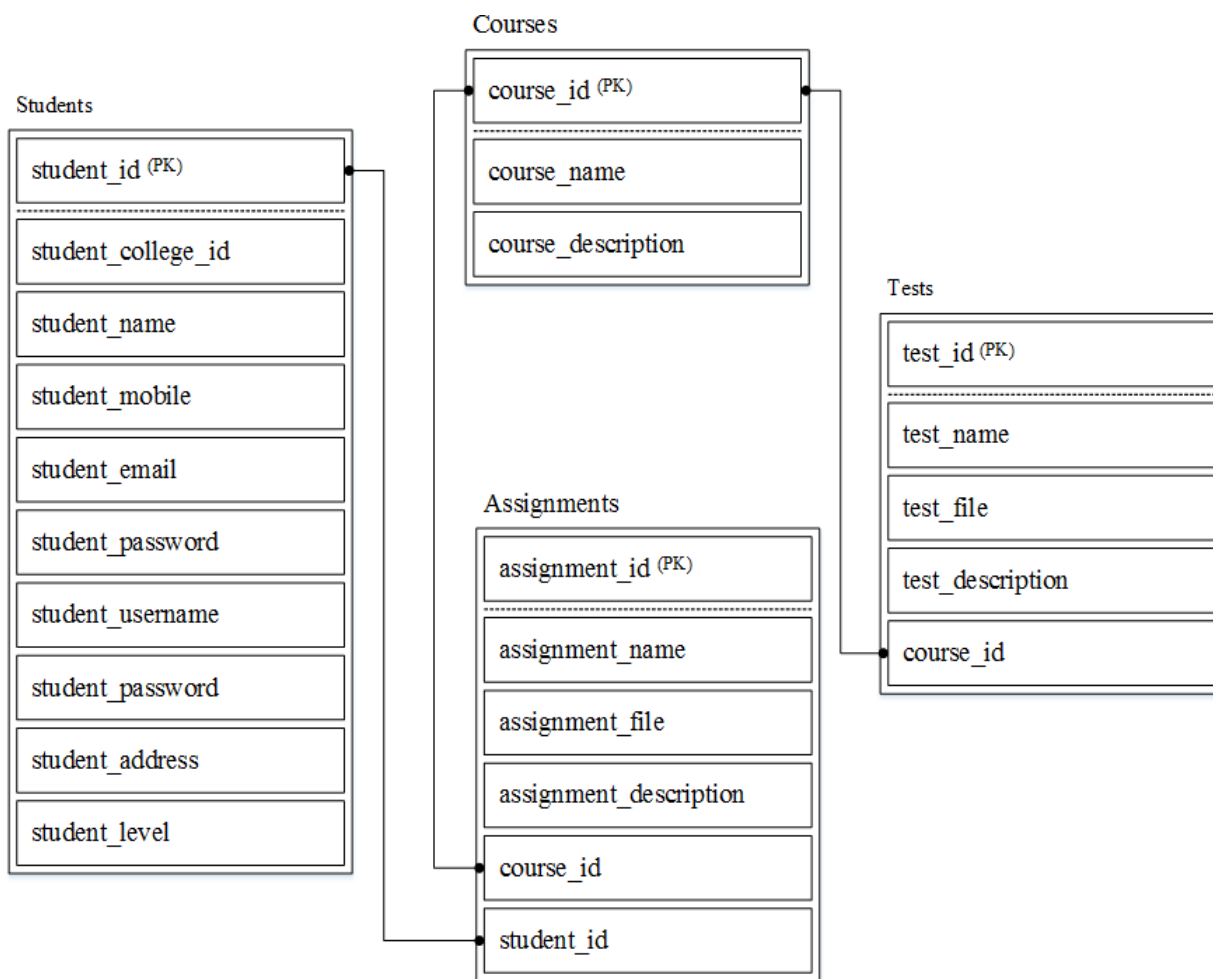


Рисунок 3.1 – Логічна модель даних системи електронної освіти

Опишемо базу даних платформи електронного навчання:

- деталі курсів зберігаються у таблицях курсів, відповідних всім таблицям;
- дані про студентів зберігаються у таблицях студентів, які відповідають усім таблицям;
- дані о спеціальностях зберігаються в таблицях спеціальностей,

відповідних усім таблицям;

- кожен об'єкт (курси, студенти, завдання, випробування, спеціальності) містить первинний ключ;
- сутність тестів, завдання має відношення до курсу, студентів із зовнішнього ключа;
- між курсами, студентами, завданнями, тестами та спеціальностями доступні стосунки один до одного та один до багатьох;
- усі сутності курсів, студентів, завдань, тестів, спеціальностей нормалізуються та не дублюються;
- усі таблиці платформи електронного навчання мають індексування для швидкого виконання запитів.

Розглянувши структуру даних у системі електронної освіти, перейдемо до вибору та порівняння найпопулярніших методів інтелектуального аналізу даних для оцінки поведінки студентів у системах електронної освіти.

### **3.3 Інтелектуальний аналіз даних**

Big data - це популярний термін, що описує набір даних, обсяг яких зростає з експоненційною швидкістю, ця інформація може бути структурованою, так і напівструктурованою або неструктурованою. В освіті це дає шанс отримати точніші рекомендації щодо курсів великого обсягу [39].

Нові методи аналізу необхідні обробки даних, оскільки обсяг даних зростає безпрецедентно швидко, і навіть їх складність. Моделі прогнозуючої аналітики чи прогнозування серед великих даних дозволяють установам приймати правильні інвестиційні рішення для вищого інституційного впливу. Алгоритми машинного навчання використовуються для побудови картографічної функції для прогнозного аналізу.

Big data все частіше застосовується в онлайн освіті, в основному з метою отримання освітніх даних та аналітики навчання. Дані щодо освіти значно розширилися, що дало безпрецедентний шанс для досліджень у

відповідних установах. Багато освітніх установ мають сховища даних і використовують аналітичні інструменти для складання персоналізованих планів навчання, а також для аналізу потенційних студентів [40].

Навчальна аналітика більше залежить від семантичних мереж, інтелектуальних навчальних програм та системних втручань. Інтелектуальний аналіз даних більшою мірою ґрунтується на результатах освітнього програмного забезпечення, моделювання учнів та прогностичних навчальних програм.

Розробка великих даних може реалізувати оптимізацію навчання. Викладачі можуть отримати реальну інформацію про кожного студента. Зміст, метод та процес навчання можуть бути налаштовані відповідно до отриманих характеристик про учнів. Сприйняття потреб учнів у навчанні, управління процесом навчання учнів, діагностики результатів навчання учнів тощо [41].

Аналіз великих даних допоможе визначити стилі навчання та дані про поведінку великої кількості студентів за допомогою відстеження та збору великих даних про них. Так що учні, ймовірно, отримають найбільш підходящі для їхнього рівня матеріали. Це не тільки підвищить ефективність роботи викладачів та ефективність навчання студентів, але також зможе посправжньому навчати студентів відповідно до їхніх здібностей та розвивати індивідуальні та інноваційні таланти, які відповідають потребам інформаційної доби.

Дослідники обговорюють можливості застосування великих даних в освіті у таких аспектах [42]:

- розширення можливостей для учнів - великий вибір програм та індивідуалізація навчальних пропозицій;
- своєчасні та належні коригування для задоволення різних потреб у навчанні учнів з різними стилями навчання та пов'язаними з цим проблемними областями;
- методи, дозволяють скоротити витрати у процесі, що вимагає багато часу, з допомогою надмірності інформації у системі глибоких знань.

Інтелектуальний аналіз даних у системах електронного навчання є критично важливим завданням для отримання, аналізу та управління великими обсягами даних. Прогнозування академічної поведінки студентів дозволяє визначити, наскільки добре окремі особи та команди виконуватимуть різні навчальні завдання. Обробка великих даних включає важливі ресурси даних і відповідні аналітичні інструменти для створення ефективних підходів, заснованих на даних. Ці підходи можуть забезпечити ранній зворотний зв'язок, інформування викладачів та менеджерів, що може допомогти покращити показники студентів, результати формальної оцінки, також можуть бути використані для виявлення схильних до ризику студентів, які можуть не скласти підсумкові іспити.

Процедуру прогнозування можна спростити так:

- визначення студентів, у тому числі їхній рівень знань, ерудиція, особистісне відповідність, здатність до навчання, початкові знання та інші аспекти;
- збір та вилучення даних про студентів на першому кроці в наборі даних та виконання попередньої обробки, приділяючи особливу увагу часу та зусиллям, що впливають на результати навчання;
- інтерпретація навичок навчання, щоб удосконалити криві прогнозування.

Прогнозування успішності може створити рівні шанси для освіти та однозначно підвищити інтерес учнів.

Студентам зазвичай пропонують широкий вибір курсів та модулів. Вони мають великий вибір як формального навчання, так неформального навчання на платформі електронного навчання [43]. Зробити правильний вибір може стати їм реальною дилемою. Система рекомендацій курсу, отримана з аналізу великих даних, може допомогти.

Керуючись глибокими дослідженнями та аналізом освітніх даних, викладачі можуть кількісно оцінити процес навчання та статус навчання студента. Вони можуть звернути увагу на кореляцію та визначити причинно-

наслідкові зв'язки. Можна виявляти актуальні проблеми в системі освіти, швидше та точніше знаходити фактори впливу та стратегії втручання, відкривати справді нові правила та значно розширювати можливості навчання.

Крім того, великі дані в освіті можуть також аналізувати дані, згенеровані учнями у процесі навчання, прогнозувати їх режими навчання та здатності до навчання, виявляти потенційні проблеми у навчанні та систематично покращувати модель навчання [44].

Різноманітні взаємодії та зворотний зв'язок залежно від курсів мають важливе значення для підвищення ефективності освіти, що призводить до важливості аналітики навчання. Початкові знання та звички до навчання, індивідуальні плани навчання є способом максимізації результатів навчання з урахуванням моделей мислення різних учнів.

Великі дані дозволяють дослідникам у сфері освіти судити про доцільність, а також про переваги та недоліки з абсолютно нової точки зору. Це демонструє, що кожен стан навчання можна помітити, використовуючи методи, більш досконалі, ніж традиційні, і кожен студент може зосередитися своєму шляху до нових знань [45].

Опис системи аналітики навчання може складатися із семи етапів:

- розуміння потреб у навчанні;
- розуміння освітніх даних;
- підготовка та попередня обробка даних;
- вибір та планування моделей;
- розробка функцій та побудова моделей;
- оцінка, рефреймінг та оптимізація;
- моніторинг та аналіз.

Кеннет вивчав вплив навчальної поведінки та рефлексивного навчання на онлайн-бізнес-курсах. Пріор проаналізував вплив онлайн-навчання за трьома аспектами: ставлення до навчання, інформаційна грамотність та самоефективність, Бачер вивчив різні рівні початкових знань і стосунки цих

результатів показують, що більш високий рівень початкових знань може призвести до того, що учні будуть успішнішими.

На середній бал або відсоткову частку успішних учнів можуть впливати різні фактори, такі як методика навчання та увага вчителів до деяких конкретних учнів. Спільним є те, що вчителі в основному зосереджуються на тих учнях, які беруть участь у заняттях та показують задовільні результати. Крім того, серед студентів існують деякі приховані патерни. Студенти можуть бути розділені на різні групи, залежно від їх успішності. Один і той самий метод навчання може бути неефективним для різних груп учнів.

Заходи подібності та кластеризації є важливими завданнями для пошуку аналогічних груп в освітніх великих даних. Подібні зразки даних у різних областях можуть бути корисними для дослідників та учнів, щоб легко отримати знання з різних областей [46].

Тонкощі аналізу великомасштабних даних обертаються навколо їх детального збору, керування, сортування, зберігання, передачі, візуального представлення та підтримки конфіденційності інформації. Аналіз великих даних зазвичай передбачає використання прогнозової аналітики або інших складних методів для вилучення корисної інформації з величезних наборів даних. Точний аналіз великих даних допомагає приймати більш обґрунтовані рішення, що, у свою чергу, підвищує ефективність виробництва, знижує витрати та знижує ризики [47].

Традиційні реляційні бази даних вміють обробляти швидкі й уніфіковані запити. Однак їм важко працювати зі складними, гнучко структурованими запитами, що призводить до надмірного навантаження та неефективності використання системи керування базами даних.

В аналізі великих даних використовуються різні методології та техніки, зокрема.

Інтелектуальний аналіз даних: це включає вивчення асоціативних правил, класифікацію (класифікацію нових даних на основі існуючих принципів даних), кластерний аналіз і регресійний аналіз.

Інтеграція та змішування даних: методи об'єднання різноманітних даних із багатьох джерел, що забезпечує поглиблений аналіз.

Машинне навчання: включає контрольоване та неконтрольоване навчання, використовуючи попередньо створені моделі для розширеного прогнозування.

Штучні нейронні мережі, мережевий аналіз, оптимізація: включає використання генетичних алгоритмів.

Просторовий аналіз: використовує топологічні, геометричні та географічні дані.

Статистичний аналіз: охоплює А/В тестування та аналіз часових рядів.

Візуалізація даних: представляє дані у формі зображень, графіків, діаграм, використовуючи інтерактивні елементи та анімацію як для кінцевих результатів, так і як необроблені дані для подальшого вивчення [48].

Зосереджуючись на даних студентів, інтелектуальний аналіз даних вибирається для поглибленого вивчення.

### **3.4 Методи інтелектуального аналізу даних в системі електронного навчання**

Інтелектуальний аналіз даних у АОД (аналізі даних) зазвичай включає такі методи, як класифікація, кластеризація, виявлення правил асоціації, розпізнавання шаблонів і інтелектуальний аналіз тексту. Ці методи поділяються на три основні групи: прогнозування, виявлення структури та ідентифікація зв'язків [49]. Зупинимось на них більш детально.

Прогнозний аналіз в АОД зосереджується на створенні моделі, яка прогнозує значення цільової змінної на основі значень незалежних змінних. Коли цільова змінна приймає неперервний діапазон, модель відома як регресія. Це передбачає пошук зв'язку між входом (незалежні змінні) і виходом (залежна змінна). У регресійному аналізі шукають функцію, яка описує зв'язок між цими змінними. Як правило, регресія виражається як комбінація

детермінованого компонента та стохастичного компонента. Метою багатофакторної регресії є пошук моделі, яка пов'язує кілька незалежних змінних із залежною змінною. Це передбачає:

Про завдання класифікації говорять, якщо вихідна змінна має кінцевий набір значень, спираючись на значення вхідних змінних, вихідна буде визначена до одного або іншого класу [46].

Використовуючи прогнозування можна припустити, частина даних розмічена, їм задані значення вхідних змінних і відповідні вихідні змінні. "Навчившись" на такому наборі даних за допомогою алгоритмів регресії або класифікації можна передбачати значення вихідної змінної для інших даних. Наприклад, спираючись на дані щодо звернення до навчальних матеріалів та підсумкової оцінки, яку студент отримав на іспиті у реального викладача, можна спрогнозувати отриману оцінку студенту зі схожою навчальною активністю.

Алгоритми виявлення структури застосовується для виявлення структури даних, без використання апріорних уявлень про неї. Найвідоміша група подібних алгоритмів – алгоритми кластеризації. Кластеризація – це логічне продовження ідеї класифікації. Унікальність кластеризації у тому, що класи об'єктів у початкових умовах не задані, тоді як із класифікації, як перейти до створення моделі, потрібно розмітити деяке підмножина даних. Результатом кластеризації є розбиття безлічі об'єктів на групи близьких у сенсі об'єктів [50].

Метою виявлення взаємовідносин є встановлення взаємозв'язку між змінними у великому обсязі вихідних даних із значною кількістю змінних. Наприклад, можна спробувати визначити, які змінні пов'язані сильніше з цікавою чи який із пар зв'язок сильніше, ніж у інших. Виявлення взаємовідносин в АОД часто використовують у вигляді пошуку сполучних правил чи пошуку послідовних шаблонів. У пошуку сполучних правил мета - це знаходження правила виду "if-then", яке показує, що якщо (if) одні змінні приймають якусь безліч значень, то (then) інша змінна матиме певне значення.

Наприклад, якщо у покупці є набір товарів А, то можна зробити висновок, що в цій же покупці повинен бути товар В. Пошук послідовних шаблонів є розвитком пошуку сполучних правил і призначений виявлення взаємовідносин між проявами послідовних у часі подій [48].

Дані методи застосовуються з метою оцінки поведінки студентів у системі електронного навчання, з допомогою можна зрозуміти, чесно чи студент проходить підсумкові тести, спрогнозувати, скільки студентів закінчить навчання і з якими оцінками. Розглянемо існуючі програмні розробки з цієї теми.

### **3.5 Алгоритми інтелектуального аналізу даних в системі електронного навчання**

Інтелектуальний аналіз даних, також відомий як виявлення даних або виявлення знань, є процесом аналізу даних з різних точок зору та узагальнення їх у корисну інформацію.

При аналізі даних, як правило, немає можливості розглянути всю цікаву для нас сукупність об'єктів. Вивчення великих обсягів даних є дорогим процесом, що вимагає великих тимчасових витрат, а також неминуче призводить до помилок, пов'язаних з людським фактором. Цілком достатньо розглянути деяку частину всієї сукупності, тобто вибірку, і отримати інформацію, що цікавить нас, на її підставі. Проте розмір вибірки має залежати від різноманітності об'єктів, які у генеральної сукупності. У вибірці мають бути представлені різні комбінації та елементи генеральної сукупності, тобто, всієї сукупності об'єктів, що вивчаються. Частина генеральної сукупності, відібрана з урахуванням властивостей і показників – це вибірка (sample) [38].

Інтелектуальний аналіз даних відповідає за пошук асоціацій, рекомендацій та інтелекту, щоб забезпечити індивідуальний та потужний механізм навчання для студентів. Наприклад, вибір відповідного контенту на основі інтересів студентів є великою проблемою. Цю проблему можна

вирішити, згрупувавши весь вміст, просто застосувавши підхід кластеризації для фільтрації вмісту відповідно до індивідуального профілю студента. Крім того, ключові компоненти виведення в таких системах електронного навчання ґрунтуються на методах інтелектуального аналізу даних, які аналізують профіль користувача та пропонують якісь дії із застосуванням штучного інтелекту.

В системі використовувалися різні методи кластеризації, такі як зсув середнього значення, К-середні, медоїд (K-medoids), заснована на щільності просторова кластеризація для шумових додатків та ієрархічна кластеризація. однак ці підходи також не є надійними для ідентифікації значних кластерів у неоднозначних та зашумлених наборах даних [22]. Розглянемо ці методи докладніше.

Зсув середнього значення — це непараметрична техніка аналізу простору ознак визначення місця розташування максимуму щільності ймовірності, так званий алгоритм пошуку моди. Це підхід на основі ковзних вікон; намагається виявити ущільнені області.

Метою цього підходу є виявлення центру кожного кластера з урахуванням методу центроїду. У методі зсуву середнього значення кандидати оновлюються для центральних точок (середнього значення точок) у ковзному вікні. На етапі постобробки вікна-кандидати фільтруються для видалення дублікатів і утворюють остаточний набір центральних точок та їх кластерів, що збігаються.

Зсув середнього значення, що має радіус « $r$ » (як ядро), починається з круглого вікна, що ковзає, центрованого у випадково обраній точці  $S$ . Метод зсуву середнього значення зрушує « $r$ » в область вищої щільності (на кожному кроці) до збіжності. Щільність кожного розсувного вікна пропорційна його розміру (точки всередині нього). При зміщенні щільність точок поступово переміщається до областей із вищою щільністю точок. Зсув ковзного вікна триває до того часу, поки зсув зможе вмістити більше точок всередині ядра (більше не збільшуючи щільність). У випадку, коли кілька розсувних вікон

перекриваються, точки даних групуються відповідно до ковзного вікна, в якому вони знаходяться, і ковзного вікна, що містить точки, що найбільш добре підтримуються.

У системах електронного навчання, працюючи з big data, обчислення складних функцій займає багато часу. Також, маючи невисоку щільність даних, результат роботи методу який завжди точний [40].

Алгоритм реалізований у таких програмних засобах:

- ELKI – Java-додаток для інтелектуального аналізу даних з багатьма алгоритмами кластеризації;
- ImageJ – створений у тому, щоб фільтрувати зображення з урахуванням фільтра зсуву середнього значення;
- OpenCV – включає в себе реалізацію зсуву середнього значення за допомогою способу cvMeanShift;
- Orfeo – є інструментальний набір, розроблений мовою програмування C++;
- Scikit-learn – розроблений мовою програмування Python та застосовує кульове дерево для результативного перегляду сусідніх точок.

Метод зрушення середнього значення має чотири недоліки:

- складний вибір вибору розміру вікна;
- неправильний розмір вікна може призвести до злиття мод або додаткових «тіньових» мод;
- складні обчислення;
- він достовірний на високою щільності даних (з ідеальним градієнтом, щоб знайти центр кластера).

K-means [20] – це сучасний алгоритм кластеризації на основі розділів. У K-means вхідні дані поділяються на k різних груп, де k - вхідний параметр, що використовується для зазначення кількості кластерів вихідних. K-means ітеративно покращує початкові розділи, доки оптимальні кластери не будуть знайдені.

K-means - найкращий вибір виявлення цікавого сигналу з освітніх наборів даних, якщо значну кількість кластерів вже визначено. Тим не менш, пошук відповідних груп з використанням K-Means може бути неможливий без попереднього аналізу та обчислення існуючого числа кластерів або за наявності шумів, або складних даних.

У освітніх даних, вибір кількості кластерів та початкову установку центроїдів K-means складно вгадати.

Недоліки алгоритму:

- не гарантується досягнення глобальних мінімумів (максимумів), лише локальних;
- результат залежить від вибору вихідних центрів кластерів, їх оптимальний вибір невідомий;
- кількість кластерів потрібно знати заздалегідь.

Таким чином, потрібні складніші та прикордонні методи кластеризації для порівняння освітніх даних та отримання внутрішньої інформації.

K-medoids [41] використовується для поділу набору даних на кластери, подібно до k-means. Метою обох методів (K-середніх та K-медоїдів) є мінімізація суми відстаней між точками даних кластера та центральною точкою даних того ж кластера. На відміну від k-середніх, k-medoids вибирає точки даних за допомогою центрів (медоїдів) та обробок за загальним правилом Манхеттена, щоб висловити відстань між точками даних. Він групує набір даних з  $n$  точок в  $k$  групи або кластери. K-medoids зменшує кількість попарних варіацій замість підсумовування квадратів евклідових відстаней. Ось чому він вважається більш стійким (до шуму та викидів), ніж k-середніх. Тим не менш, дані метод також потребує попередньої обробки даних та знаходження числа кластерів.

Метод k-medoids є ефективним для зворотного перетворення даних. Також підходить для кластеризації певних даних, де середнього немає.

DBSCAN [42] - це заснована на щільності просторова кластеризація додатків з шумами, яка починається через випадкову початкову точку. Сусідні

точки даних знаходяться на відстані - . Якщо в околиці існує достатня кількість точок даних, починається процедура кластеризації і поточна точка даних вважається першою. Інші дані будуть позначені як шум. Пізніше ці зашумлені точки даних можуть бути частиною кластера. В обох ситуаціях точки даних позначаються як відвідані. Для поточного кластера процедура створення всіх точок даних у Околиці повторюється, щоб додати нові точки. Цей процес повторюється до тих пір, поки всі точки даних у поточному кластері не будуть розпізнані та позначені як «відвідана». Той самий процес повторюється всім кластерів.

Алгоритм DBSCAN реалізований та застосовується в наступних програмних засобах:

- Apache Commons Math включає реалізацію на Java алгоритму, який працює за квадратичний час;
- ELKI – надає реалізацію DBSCAN, GDBSCAN та інших видів даного алгоритму. У цій реалізації використовуються довільні функції відстані та довільні типи даних, домогтися прискорення можна оптимізацією на низькому рівні, а також застосовуючи спеціальні методи на малих наборах даних;
- PostGIS містить ST\_ClusterDBSCAN — двовимірну реалізацію DBSCAN, яка використовує R-дерево як індекс, реалізовані такі геометричні типи, як «точка», "відрізок", "багатокутник" і т. д.;
- мова R містить пакет fpc DBSCAN, який підтримує довільні функції відстані через матриці відстаней;
- SPMF містить реалізацію алгоритму DBSCAN з підтримкою kd дерева тільки для евклідової відстані;
- Weka, як додатковий пакет, містить базову реалізацію DBSCAN, яка вимагає лінійну пам'ять і працює за квадратичний час.

DBSCAN має такі недоліки:

- DBSCAN неоднозначний - крайові точки, які можуть бути досягнуті з одного і більше кластерів, можуть належати до будь-якого з цих кластерів,

що залежить від послідовності перегляду точок;

- DBSCAN погано кластеризує набори даних із великою різницею у щільності;
- DBSCAN не повідомляє, коли кластери змінної щільності встановлені;
- при зміні щільності гранична відстань і точки для ідентифікації сусідніх точок даних відрізнятимуться від кластера до кластера, для дуже високої розмірності, порогова відстань стає складним для оцінки.

У цьому пункті ми обговоримо лише висхідну ієрархічну кластеризацію. Спочатку він обробляє кожну точку даних як один кластер, а потім безперервно поєднує пари кластерів, поки цілі кластери не будуть об'єднані в окремий кластер. Він також відомий як агломеративне ієрархічне кластерне дерево (АНСТ) [43]. АНСТ представляється як дерева. Коріння дерева розглядається як унікальний кластер.

Спочатку кожна точка даних обробляється як один кластер. Це означає, що точки даних „k” обробляються як кластери „k”. На наступному кроці вибирається метрика відстані для його виміру між двома кластерами. Крім того, два кластери об'єднуються в один ітеративно для кожної пари. Об'єднані кластери вибираються з найменшим середнім зв'язком; обидва кластери мають найменшу відстань та найбільш схожі точки даних. Цей крок буде продовжуватися доти, поки корінь дерева не буде явно заданий на початку. Кількість кластерів може бути обрана шляхом розпізнавання заданого кореневого номера, що допомагає припинити об'єднання кластерів.

Вихідний ієрархічний метод кластеризації не вимагає вказівки кількості кластерів і має можливість вибрати найкращий кластер через використання дерева.

У 2014 році CFSFDP був запропонований Алексом та Лайо [44]. Він має характеристики для виявлення значних кластерів більш інтуїтивним способом порівняно з K-means. Пропонується новий евристичний підхід, який розширює процедуру кластеризації, в якому області високої густини ідентифікуються як

потенційні кластери, автоматично визначаються винятки та організуються довільні форми кластерів.

У K-means для отримання значних кластерів користувачі повинні повторити процес кластеризації кілька разів з різними параметрами. Тим не менш, унікальний підхід, що використовується в CFSFDP для адаптивного виявлення кластерів та шуму, стане важливим інструментом кластеризації для аналізу у сфері електронної освіти. CFSFDP використовує таку методологію виявлення значних кластерів.

Дані про кожного студента є дані - data-point координатою якої є пара, де – рейтинг студента, обчислення якого відбувається на основі відгуків викладачів та середнього балів студента на предмет поточного семестру, а – середня кількість часу, проведене на платформі.

CFSFDP розраховує локальну щільність та мінімальна відстань між кожним студентом і точкою високої густини (передбачуваним центром кластера). Де дорівнює кількості студентів, які знаходяться на відстані меншій, ніж  $r$ , по відношенню до точки. У свою чергу, це важливий параметр, який використовується для оцінки кожної точки. Ефективність методу CFSFDP залежить від правильного вибору параметра.

Центри кластерів досягаються шляхом побудови розрахованих значень і що називається графом рішень. У кластерному аналізі ключове завдання знайти правильні кластерні центри в наборах даних [19]. Тим не менш, CFSFDP використовує граф рішень для вибору правильних центрів кластерів з найменшою кількістю взаємодії з людиною, що робить його гіднішим для аналізу великих даних чи потокових даних. CFSFDP має безліч застосувань в освіті, а також у багатьох інших областях, таких як біоінформатика, обробка зображень та аналіз білка [23].

Оскільки CFSFDP має характеристики, що дозволяють виявляти приховані дані з неоднозначних даних, він може застосовуватися в існуючих системах інтелектуального аналізу освітніх даних та системах електронного навчання для створення більш значущих кластерів, а надалі його можна

використовувати для кластеризації аналогічних документів, пошуку плагіату в документи, а також аналізувати профілі студентів та знаходити аналогічні ідеї у різних галузях досліджень. CFSFDP за допомогою дифузії (CFSFDP-HD) [21] був запропонований як варіант CFSFDP, де обмеження CFSFDP покращені, і користувачі можуть аналізувати дані без будь-якого попереднього знання предметної області. У CFSFDP-HD для оцінки густини базового набору даних використовувався адаптивний (табл. 3.2).

Цей алгоритм рекомендує оцінку для викладача та попереджає студента про можливі результати таким чином:

а) у вибірку для аналізу потрапляють усі студенти з бази даних, які проходили курс:

- 1) студенти, які завершили курс успішно;
- 2) студенти, які завершили курс неуспішно;
- 3) студенти, які проходять курс;

б) алгоритм виконує свою роботу, розподіляючи студентів на кластери на основі середніх балів за проміжні завдання курсу, балів, отриманих в результаті експертної оцінки;

в) студенту, який проходить курс, та викладачеві пропонується оцінка на основі оцінок студентів, які завершили курс.

Таблиця 3.2 – Порівняння методів кластеризації даних

| Метод      | Опис                                    | Чутливість до густини даних | Масштабованість | Вимоги до введення користувача | Використання в електронному навчанні          |
|------------|-----------------------------------------|-----------------------------|-----------------|--------------------------------|-----------------------------------------------|
| 1          | 2                                       | 3                           | 4               | 5                              | 6                                             |
| Mean Shift | Непараметричний, виявляє щільні області | Висока                      | Низька          | Розмір вікна                   | Не типово використовується                    |
| K-Means    | Поділяє дані на 'k' груп                | Середня                     | Висока          | Кількість кластерів            | Підходить для попередньо визначених кластерів |

Продовження таблиці 3.2

| 1                        | 2                                            | 3       | 4       | 5                                        | 6                                                                  |
|--------------------------|----------------------------------------------|---------|---------|------------------------------------------|--------------------------------------------------------------------|
| K-Medoids                | Схожий на K-means, але використовує медоїди  | Висока  | Середня | Кількість кластерів, початкова установка | Ефективний для даних із чіткими представниками                     |
| DBSCAN                   | Групує точки даних у щільних областях        | Висока  | Середня | Порогова відстань                        | Гарно підходить для даних з різною щільністю                       |
| Ієрархічна Кластеризація | Агломеративна, формує кластери злиттям точок | Низька  | Низька  | Немає вимог до кількості кластерів       | Підходить для експлоративного аналізу даних                        |
| CFSFDP                   | Групує області високої щільності             | Середня | Висока  | Параметри для розрахунку щільності       | Корисний для інтуїтивного виявлення кластерів у освітніх даних     |
| CFSFDP-HD                | Покращує CFSFDP, дозволяє аналіз без знань   | Середня | Висока  | Параметри дифузії                        | Покращує CFSFDP для надійніших результатів у електронному навчанні |

На відміну від існуючих систем електронного навчання, пропонується використовувати методи кластеризації за допомогою швидкого пошуку та знаходження піків щільності (CFSFDP) та CFSFDP за допомогою дифузії тепла (CFSFDP-HD) для досягнення надійних результатів.

### **3.6 Програмна реалізація алгоритму інтелектуального аналізу даних**

Основні етапи алгоритму CFSFDP-HD показані на рисунку 3.2. Поданий підхід використовує матрицю відстаней  $D$  набору даних як вхідні дані.

Крок 1: на першому етапі пропонований підхід оцінює щільність через алгоритм CFSFDP-HD.

Крок 2: пропонований підхід обчислює мінімальну відстань від вищих найближчих точок.

Крок 3: ідентифікація кластерних центрів досягається графом рішень. Результатом цього кроку є кластерні центри.

Крок 4: присвоєння точок, що залишилися, ідентифікованим кластерним центрам. Результатом цього кроку є організовані кластери з шумами і кластерами, що перекриваються.

Крок 5: на цьому етапі представлений підхід ідентифікує і виправляє помилково класифіковані точки, а також ідентифікує шум або викиди організованих кластерів (шуми та кластери, що перекриваються). Результатом пропонованого підходу є організовані кластери.

В результаті роботи алгоритму користувачеві подаються дані про студентів, розподілені за кластерами. Перевагою буде те, що навіть найвіддаленіші точки не будуть вважатися шумами, а долучаться до створених кластерів.

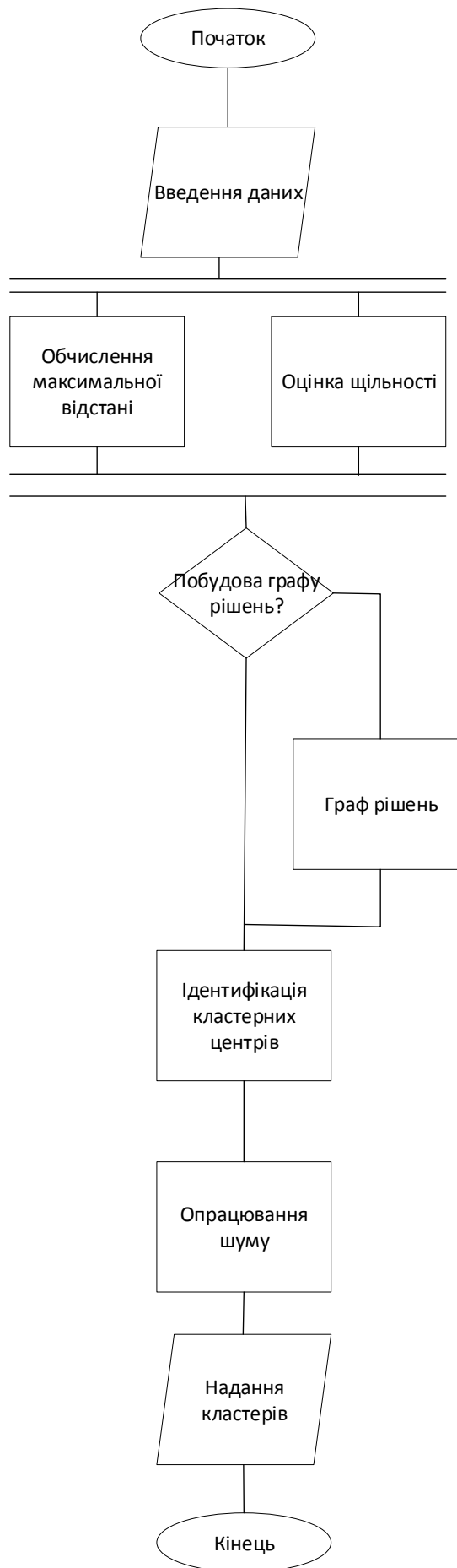


Рисунок 3.2 – Схема алгоритму

### 3.7 Висновки за розділом

У третьому розділі було досліджено методи інтелектуального аналізу даних у системі електронного навчання.

Розглянуто такі методи: прогнозування, виявлення структури та виявлення взаємозв'язків. Детально проаналізовано алгоритми: зсуву середнього значення, K-means, K-medoids, DBSCAN, ієрархічної кластеризації, CFSFDP та CFSFDP-HD. Визначено їхні переваги та недоліки.

Запропоновано використовувати методи кластеризації CFSFDP та CFSFDP-HD для досягнення надійних результатів оцінювання студентів. Ці алгоритми дозволяють інтуїтивно та швидко знаходити приховані закономірності у даних.

Розроблено програмну реалізацію алгоритму CFSFDP-HD на мові Python. Показано основні етапи роботи: оцінка щільності, розрахунок відстаней, побудова графу рішень, кластеризація.

Запропоновано використання ефективних алгоритмів інтелектуального аналізу даних для вирішення завдання оцінювання студентів у системах електронного навчання.

## 4 ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ МЕТОДІВ

### 4.1 Проведення експерименту

У цьому експерименті набір даних складається із 200 студентів різних років навчання. Експеримент моделюється за допомогою підходу CFSFDP-HD для поділу учнів на відповідні групи та ґрунтується на отриманих оцінках учнів та відвідуваності занять. Сегментація учнів з урахуванням прогресу необхідна для розробки відповідних методів навчання, направлених на усунення недоліків певної групи.

Графік прийняття рішень встановлюється після (1) оцінки щільності за допомогою термодифузії та (2) розрахунку мінімальної відстані від вищих найближчих щільних точок. Ідентифікація кластерних центрів досягається за допомогою графа рішень (рис. 4.1).

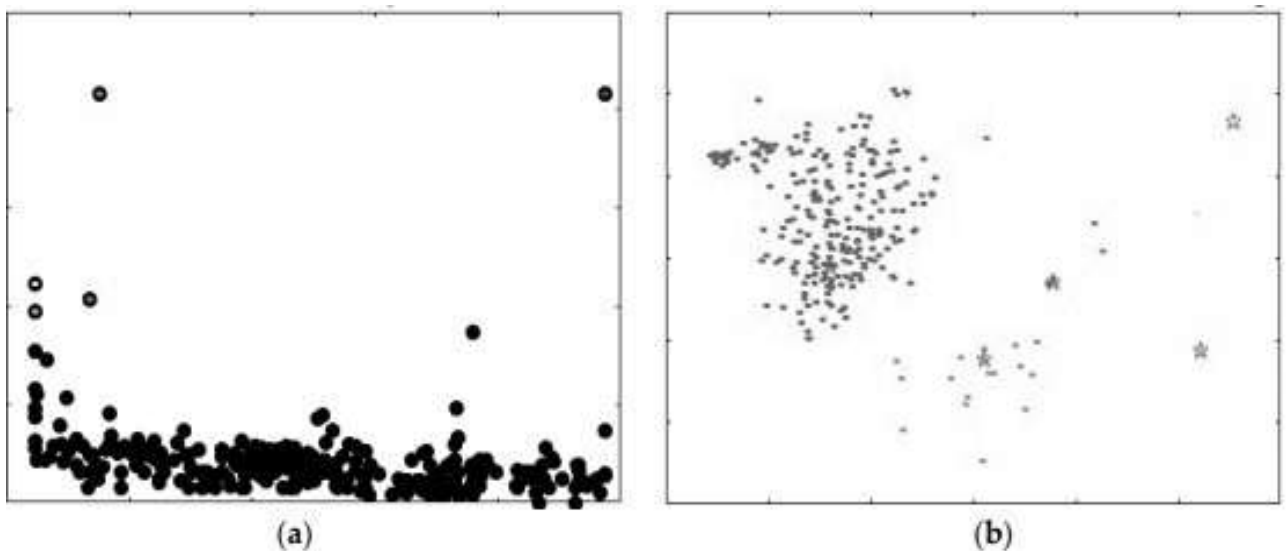


Рисунок 4.1 – Ідентифікація кластерних центрів за допомогою графа рішень

Вищезгаданий поділ студентів на чотири значущі групи може відіграти важливу роль у розвитку навичок навчання, приділяючи особливу увагу певній групі студентів. Самомотивовані та талановиті учні відокремлюються від учнів із низькими показниками. На основі отриманих різних категорій студентів, викладачі можуть адаптувати різні підходи до навчання для роботи з

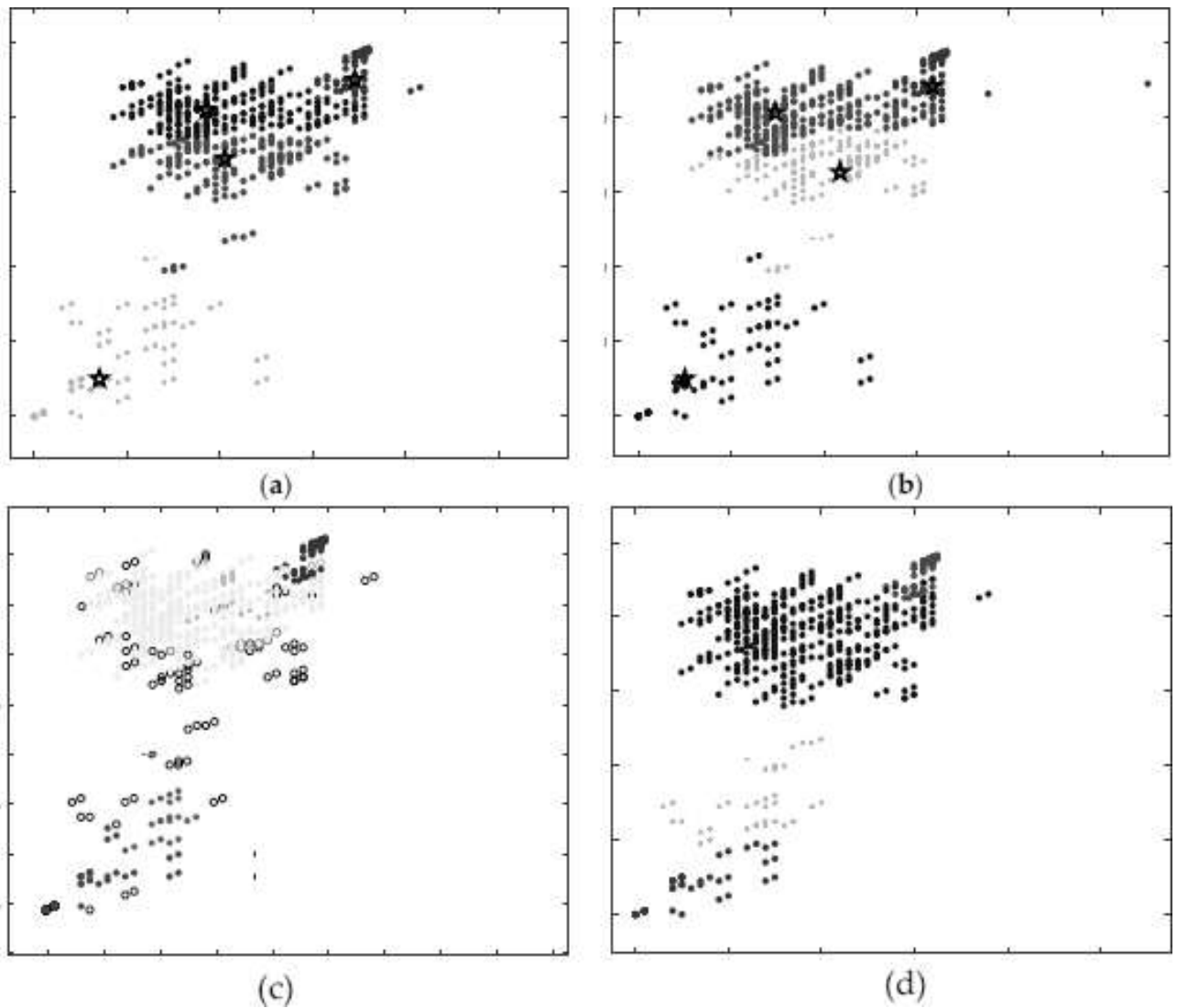
відповідною групою студентів. Отже, продуктивність студентів може бути покращена шляхом застосування різних методів для кожної групи студентів. З вищезазначеного тематичного дослідження, кластеризація може поділити дані про освіту на відповідні групи, і ці групи можуть бути використані для подальшого аналізу для покращення загальної системи освіти.

Проаналізовані методи кластеризації, тобто зсув середнього значення, K-means, K-medoids, DBSCAN та ієрархічна кластеризація, були протестовані з використанням одного набору даних з 200 учнів та з урахуванням обмежень на 4 кластери (рис. 4.2).

Шляхом порівняння CFSFDP-HD з K-means, K-medoids, DBSCAN та АНСТ підхід на основі графа рішень CFSFDP-HD забезпечує краще розуміння для вибору потенційних кластерів інтуїтивно. У звичайній практиці користувачі використовують K-means і K-medoids більше 10 000 разів з різними налаштуваннями (тобто числом кластерів та ітерацій), щоб отримати значущі кластери, однак підхід на основі графа рішень, що використовується в CFSFDP-HD, забезпечує евристику, щоб отримати точні рішення за кілька ітерацій.

Щоб отримати відповідні кластери з використанням підходів, що обговорюються, користувачі повинні мати попередні знання існуючих кластерів (кількість кластерів) і в цьому випадку, не в змозі виявити студентів з низькими і дуже низькими показниками. Це обмеження робить недоречним виявлення всіх внутрішніх прихованих закономірностей даних.

Для вирішення технічних проблем рекомендується використовувати метод CFSFDP-HD для виявлення існуючих шаблонів без технічних знань про базові дані.



(a) – K-means; (b) – K-medoids; (c) – DBSCAN; (d) – АНСТ

Рисунок 4.2 – Візуалізація результатів кластеризації різними методами

З експериментів видно, що CFSFDP-HD є адаптивнішим за своєю природою, і його результати більш значні порівняно з деякими з існуючих підходів.

## 4.2 Дослідження результатів

Метою роботи було застосування методів та алгоритмів інтелектуального аналізу даних для дослідження поведінки студента у системі електронної освіти. У роботі було представлено архітектуру електронного

навчання з використанням методів інтелектуального аналізу даних. Також було розглянуто потенційне застосування кластеризації в освітніх великих даних. Вище представлені результати цього експерименту

Мета дослідження полягала в тому, що можна застосовувати методи та алгоритми інтелектуального аналізу для прийняття рішення щодо коректності поведінки студента у системі електронного навчання. Підходи інтелектуального аналізу даних дають результати у існуючих системах електронного навчання, ефективно та результативно. З літератури було зазначено, що традиційні системи електронного навчання в основному ґрунтуються на запитах, і на запити працюють без будь-яких інтелектуальних чи евристичних даних. Так само K-means підходить для кластеризації освітніх даних, в яких відома кількість кластерів, і стикається з недоліками при застосуванні до кластерів невідомого розміру. Отже, більш надійні підходи до інтелектуального аналізу даних (CFSFDP та CFSFDP-HD) можна включити в систему електронного навчання для пошуку кластерів у освітніх даних. Крім того, було оцінено, що методи інтелектуального аналізу даних ефективні при аналізі великих даних, що робить системи освіти надійними та здатними вирішувати проблеми міждисциплінарних досліджень, емоційного навчання та електронного навчання.

#### **4.3 Висновки за розділом**

У четвертому розділі було проведено експериментальне дослідження запропонованого методу інтелектуального аналізу даних CFSFDP-HD.

Було сформовано вибірку даних про 200 студентів та застосовано алгоритм CFSFDP-HD для їх кластеризації. Результати порівнювалися з іншими методами: K-means, K-medoids, DBSCAN та ієрархічною кластеризацією.

Показано, що CFSFDP-HD є адаптивнішим за своєю природою і дає більш значущі результати порівняно з іншими методами. Він дозволяє

інтуїтивно та швидко знаходити приховані закономірності у даних.

Експериментально підтверджено ефективність застосування розробленого методу інтелектуального аналізу даних для вирішення завдання оцінювання студентів у системах електронного навчання.

## ВИСНОВКИ

У ході виконання дипломної кваліфікаційної роботи було сформульовано цілі та завдання, визначено об'єкт та предмет, позначено актуальність теми дослідження.

Виконано наступні завдання:

- проведено порівняльний аналіз існуючих способів та засобів аналізу поведінки студента при проходженні курсу електронного навчання;
- досліджено можливі критерії порівняння поведінки студентів;
- досліджено методи та алгоритми інтелектуального аналізу даних;
- спроєктовано та реалізовано програмний засіб інтелектуального аналізу поведінки студента при проходженні курсу електронного навчання на основі методів кластеризації;
- перевірено ефективність розробленої моделі.

Грунтуючись на аналізі сучасної літератури, було описано теоретико-методологічні засади інтелектуального аналізу даних поведінки студентів у системі електронного навчання. Визначено основні пункти аналізу поведінки студентів у системі електронного навчання. Розглянуто методи інтелектуального аналізу даних. Виділено головні переваги та недоліки існуючих програмних пакетів щодо оцінки поведінки студентів у системі електронного навчання. Описано існуючі програмні пакети.

Досліджено моделі аналізу процесу онлайн навчання, а також big data у системах електронної освіти. Описано форму подання даних про студентів, алгоритми вивантаження великих даних та подано логічну модель даних у системах електронної освіти.

Досліджено методи інтелектуального аналізу даних:

- прогнозування;
- виявлення структури;
- виявлення взаємозв'язків.

Досліджено алгоритми інтелектуального аналізу даних, позначено їх переваги та недоліки:

- зсув середнього значення;
- K-means;
- K-medoids;
- DBSCAN;
- ієрархічна кластеризація;
- CFSFDP та CFSFDP-HD.

Проведено програмне моделювання алгоритму CFSFDP-HD, представлено експеримент та порівняння результатів аналізу роботи різних методів.

Практична значущість роботи полягає у розробці програмної реалізації комбінованих моделей для оцінки поведінки студентів у системі дистанційного навчання. Це дозволяє з більшою точністю ідентифікувати потенційні проблеми у навчальному процесі та покращувати методики навчання.

Наукова новизна дослідження полягає у використанні методів та алгоритмів інтелектуального аналізу даних для оцінки поведінки студента у системі електронної освіти.

## ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАНЬ

1. Lee H.-Y. Big Challenge in Big Data Research: Continual Dispute on Big Data Analysis [Electronic resource]. / Hae-Young Lee // Korean Circulation Journal. – 2020. – Vol. 50, no. 1. – P. 69. – Access mode: <https://doi.org/10.4070/kcj.2019.0349>.
2. Rawat R. Big Data: Big Data Analysis, Issues and Challenges and Technologies [Electronic resource] / R. Rawat, R. Yadav // IOP Conference Series: Materials Science and Engineering. – 2021. – Vol. 1022. – P. 012014. – Access mode: <https://doi.org/10.1088/1757-899x/1022/1/012014>.
3. Bilge Y. Error analysis in big data analysis [Electronic resource]. / Yasar Bilge // Atlantic Journal of Medical Science and Research. – 2023. – Vol. 3, no. 2. – P. 81. – Access mode: <https://doi.org/10.5455/atjmed.2022.11.019>.
4. Song J. A Big Data Analysis and Mining Approach for IoT Big Data [Electronic resource]. / Jihyun Song, Kyeongjoo Kim // International Journal of Advances in Computer Science and Technology. – 2018. – Vol. 7, no. 1. – P. 1–3. – Access mode: <https://doi.org/10.30534/ijacst/2018/01712018>.
5. Bopche A. Big Data Analysis using Hadoop [Electronic resource]. / Archana Bopche // International Journal of Computer Applications. – 2017. – Vol. 179, no. 6. – P. 1–4. – Access mode: <https://doi.org/10.5120/ijca2017915960>.
6. Bakdash J. Risk Analysis in Big Data [Electronic resource]. / Jonathan Bakdash, Laura Marusich // SSRN Electronic Journal. – 2015. – Access mode: <https://doi.org/10.2139/ssrn.2641726>.
7. Singh C. Big Data Analysis: A Review [Electronic resource]. / Chaitanya Singh // International Journal of Engineering and Computer Science. – 2017. – Vol. 6, no. 11. – Access mode: <https://doi.org/10.18535/ijecs/v6i11.14>.
8. Ibrahim N. Big Data Analysis of Web Data Extraction [Electronic resource]. / Nadia Ibrahim, Alaa Hassan, Marwah Nihad // International Journal of Engineering & Technology. – 2018. – Vol. 7, no. 4.37. – P. 168. – Access mode: <https://doi.org/10.14419/ijet.v7i4.37.24095>.

9. Maitrey S. MapReduce: Simplified Data Analysis of Big Data [Electronic resource]. / Seema Maitrey, C. K. Jha // Procedia Computer Science. – 2015. – Vol. 57. – P. 563–571. – Access mode: <https://doi.org/10.1016/j.procs.2015.07.392>.
10. Lee S.-J. Big Data Analysis Using Principal Component Analysis [Electronic resource]. / Seung-Joo Lee // Journal of Korean Institute of Intelligent Systems. – 2015. – Vol. 25, no. 6. – P. 592–599. – Access mode: <https://doi.org/10.5391/jkiis.2015.25.6.592>.
11. Singhal P. An analysis of data mining with big data [Electronic resource]. / Priyank Singhal // Asian Journal of Multidimensional Research. – 2021. – Vol. 10, no. 11. – P. 434–440. – Access mode: <https://doi.org/10.5958/2278-4853.2021.01022.3>.
12. Xu Q. Volume-Based Data Representation of Big Data Analysis [Electronic resource]. / Qian Xu, Zheng Xu Zhao, Wei Wang // Advanced Materials Research. – 2013. – Vol. 798-799. – P. 680–684. – Access mode: <https://doi.org/10.4028/www.scientific.net/amr.798-799.680>.
13. Avunduk H. Big data and data analysis in continuous auditing [Electronic resource]. / Hüseyin Avunduk, Merve Kizgin // Journal of business in the digital age. – 2020. – P. 76–83. – Access mode: <https://doi.org/10.46238/jobda.685120>.
14. Sulaiman S. M. Smart Meter Data Analysis Using Big Data Tools [Electronic resource]. / S. M. Sulaiman, P. Aruna Jeyanthi, D. Devaraj // Journal of Computational and Theoretical Nanoscience. – 2019. – Vol. 16, no. 8. – P. 3629–3636. – Access mode: <https://doi.org/10.1166/jctn.2019.8338>.
15. Srivastava P. Moodle [Electronic resource]. / Praveen Srivastava, Shelly Srivastava // Transforming Higher Education Through Digitalization. – Boca Raton, 2021. – P. 133–144. – Access mode: <https://doi.org/10.1201/9781003132097-8>.
16. Moodle [Electronic resource]. / O. A. Guirette-Barbosa [et al.] // Cases on Economics Education and Tools for Educators. – [S. l.], 2023. – P. 205–228. – Access mode: <https://doi.org/10.4018/978-1-6684-7583-6.ch009>.

17. Галата С. Змішана освіта: одружити "онлайн" і "офлайн" / С. Галата // Освіта України. – 2017. – С. 8–9.
18. Picciano A. G. Introduction to Foundations, Planning, and Pedagogy [Electronic resource]. / Anthony G. Picciano // Online Education. – [S. l.], 2018. – P. 1–8. – Access mode: <https://doi.org/10.4324/9781315226750-1>.
19. Picciano A. G. Planning for Online Education [Electronic resource]. / Anthony G. Picciano // Online Education. – [S. l.], 2018. – P. 65–83. – Access mode: <https://doi.org/10.4324/9781315226750-5>.
20. Picciano A. G. Online Education [Electronic resource]. / Anthony G. Picciano. – [S. l.] : Routledge, 2018. – Access mode: <https://doi.org/10.4324/9781315226750>.
21. Basic Association Rule Mining in RapidMiner [Electronic resource]. // RapidMiner. – [S. l.], 2016. – P. 151–172. – Access mode: <https://doi.org/10.1201/b16023-11>.
22. Recommender System for Selection of the Right Study Program for Higher Education Students [Electronic resource]. // RapidMiner. – [S. l.], 2016. – P. 199–210. – Access mode: <https://doi.org/10.1201/b16023-13>.
23. Visualising Clustering Validity Measures [Electronic resource]. // RapidMiner. – [S. l.], 2016. – P. 213–238. – Access mode: <https://doi.org/10.1201/b16023-15>.
24. Kovács L. Efficiency comparison of Python and RapidMiner [Electronic resource]. / László Kovács, Hina Ghous // Multidiszciplináris Tudományok. – 2020. – Vol. 10, no. 3. – P. 212–220. – Access mode: <https://doi.org/10.35925/j.multi.2020.3.26>.
25. Kralj S. Comparative Analyses of Medicinal Chemistry and Cheminformatics Filters with Accessible Implementation in Konstanz Information Miner (KNIME) [Electronic resource]. / Sebastjan Kralj, Marko Jukič, Urban Bren // International Journal of Molecular Sciences. – 2022. – Vol. 23, no. 10. – P. 5727. – Access mode: <https://doi.org/10.3390/ijms23105727>.
26. Baydogan C. Sentiment analysis using Konstanz Information Miner in

social networks [Electronic resource]. / Cem Baydogan, Bilal Alatas // 2018 6th International Symposium on Digital Forensic and Security (ISDFS), Antalya, 22–25 March 2018. – [S. l.], 2018. – Access mode: <https://doi.org/10.1109/isdfs.2018.8355395>.

27. Zwick M. Automated curation of gene name normalization results using the Konstanz information miner [Electronic resource]. / Matthias Zwick // Journal of Biomedical Informatics. – 2015. – Vol. 53. – P. 58–64. – Access mode: <https://doi.org/10.1016/j.jbi.2014.08.016>.

28. Design der DWH-Schichten [Electronic resource]. / Dani Schnider [et al.] // Data Warehouse Blueprints. – München, 2016. – P. 109–213. – Access mode: <https://doi.org/10.3139/9783446451117.005>.

29. Kaur H. Visualizing Clouds on Different Stages of DWH - An Introduction to Data Warehouse as a Service [Electronic resource]. / Harjeet Kaur, Prateek Agrawal, Amita Dhiman // 2012 International Conference on Computing Sciences (ICCS), Phagwara, India, 14–15 September 2012. – [S. l.], 2012. – Access mode: <https://doi.org/10.1109/iccs.2012.78>.

30. Claybrook B. G. OLTP, online transaction processing systems / Billy G. Claybrook. – New York : J. Wiley, 1992. – 358 p.

31. Toward a General Distributed Messaging Framework for Online Transaction Processing Applications [Electronic resource]. / Jianwei Liao [et al.] // IEEE Access. – 2017. – Vol. 5. – P. 18166–18178. – Access mode: <https://doi.org/10.1109/access.2017.2717930>.

32. Redbooks I. Data Modeling Techniques for Data Warehousing / IBM Redbooks. – [S. l.] : Ibm, 1998. – 214 p.

33. Redbooks I. From Multiplatform Operational Data to Data Warehousing and Business Intelligence / IBM Redbooks. – [S. l.] : Ibm, 1998. – 358 p.

34. usyaffa N. Market basket analisis data mining terhadap data penjualan menggunakan algoritma frequent pattern growth (fp growth) [Electronic resource]. / Numan Musyaffa, Arfhan Prasetyo, Ricki Sastra // Jurnal khatulistiwa informatika. – 2021. – Vol. 9, no. 2. – Access mode: <https://doi.org/10.31294/jki.v9i2.11133>.

35. Skvortsov A. V. Algorithms for triangulation construction and analysis [Electronic resource]. / Alexey V. Skvortsov, Natalia S. Mirza. – Tomsk : Tomsk state university, 2006. – Access mode: <https://doi.org/10.17273/book.2006.1>.

36. Pushkareva A. Architecture of application for analysis and visualization of data of multi-parameter telemedical monitoring [Electronic resource]. / Anastasiya Pushkareva, Viktor Baranov, Andrey Kuzmin // Models, systems, networks in economics, technology, nature and society. – 2022. – No. 3. – Access mode: <https://doi.org/10.21685/2227-8486-2022-3-11>.

37. Litvinova Y. S. Quantitative risk analysis of IT-startups [Electronic resource]. / Yuliya Sergievna Litvinova, Anatolii Oleksandrovich Sumskiy // Bulletin of national technical university "Khpi". series: system analysis, control and information technologies. – 2020. – No. 1 (3). – P. 54–57. – Access mode: <https://doi.org/10.20998/2079-0023.2020.01.10>.

38. De La Cruz-Contreras B. A. A Comparative Analysis of Poverty Theories [Electronic resource]. / Brandon Alan De La Cruz-Contreras, Juan José Mendoza-Alvarado // Journal Public Economy. – 2019. – P. 27–37. – Access mode: <https://doi.org/10.35429/jpe.2019.5.3.27.37>.

39. Gaidarski I. An Analysis of the Domain of Information Security Systems [Electronic resource]. / Ivan Gaidarski, Pavlin Kutinchev // 2021 Big Data, Knowledge and Control Systems Engineering (BdKCSE), Sofia, Bulgaria, 28–29 October 2021. – [S. l.], 2021. – Access mode: <https://doi.org/10.1109/bdkcse53180.2021.9627314>.

40. Chernigovskiy A. V. Analysis of prognosticating models of network traffic [Electronic resource]. / A. V. Chernigovskiy, M. V. Krivov // Modern technologies and scientific and technological progress. – 2019. – Vol. 1, no. 1. – P. 172–173. – Access mode: <https://doi.org/10.36629/2686-9896/2019-1-1-172-173>.

41. Kamenskikh S. N. Analysis of information flows of industrial enterprise [Electronic resource]. / S. N. Kamenskikh, A. A. Shirokov // Scientific and technical volga region bulletin. – 2017. – Vol. 7, no. 4. – P. 125–127. – Access mode: <https://doi.org/10.24153/2079-5920-2017-7-4-125-127> (date of access: 01.12.2023).

42. Big data & sentiment analysis using python [Electronic resource]. / Akshansh Malik [et al.] // International research journal of computer science. – 2020. – Vol. 07, no. 06. – P. 159–166. – Access mode: <https://doi.org/10.26562/irjcs.2020.v0706.003>.

43. Irada Seyidova I. S. Big data processing with python in social networks [Electronic resource]. / Irada Seyidova Irada Seyidova, Sahib Shakhayev Sahib Shakhayev // ETM - equipment, technologies, materials. – 2023. – Vol. 16, no. 04. – P. 76–82. – Access mode: <https://doi.org/10.36962/etm16042023-76>.

44. Weiss C. J. Visualizing protein big data using Python and Jupyter notebooks [Electronic resource]. / Charles J. Weiss // Biochemistry and Molecular Biology Education. – 2022. – Access mode: <https://doi.org/10.1002/bmb.21621>.

45. Nokeri T. C. Big Data, Machine Learning, and Deep Learning Frameworks [Electronic resource]. / Tshepo Chris Nokeri // Data Science Solutions with Python. – Berkeley, CA, 2021. – P. 7–14. – Access mode: [https://doi.org/10.1007/978-1-4842-7762-1\\_2](https://doi.org/10.1007/978-1-4842-7762-1_2).

46. Kim S. h. Big data Analysis using Python in Agriculture Forestry and Fisheries [Electronic resource]. / So hee Kim, Min Soo Kang, Yong Gyu Jung // International journal of advanced smart convergence. – 2016. – Vol. 5, no. 1. – P. 47–50. – Access mode: <https://doi.org/10.7236/ijasc.2016.5.1.47>.

47. Liu Q. Teaching Practice of Python Programming Course in Big Data Era [Electronic resource]. / Qing Liu // ITM Web of Conferences. – 2019. – Vol. 26. – P. 01006. – Access mode: <https://doi.org/10.1051/itmconf/20192601006>.

48. Teimourzadeh A. Application of python in marketing education: a big data analytics perspective [Electronic resource]. / Aria Teimourzadeh, Samaneh Kakavand, Benjamin Kakavand // Marketing education review. – 2022. – P. 1–16. – Access mode: <https://doi.org/10.1080/10528008.2021.2021374>.

49. Python M. SQL for Beginners: The Simplified Beginner's Guide, to Learn and Understand SQL Language Computer Programming, Data Analytics, Database Design and Server. Including Basic Project and Exercise / Matthew Python. – [S. l.] : Independently Published, 2020.

50. Guevin J. Big Book of Python for Data Analysis : Learn Data Science and Coding with Python: Advanced Python Books / Johnie Guevin. – [S. l.] : Independently Published, 2021.

**ДОДАТОК А**  
**Лістинги модулів**

```

import numpy as np
from sklearn.cluster import KMeans, AgglomerativeClustering,
DBSCAN
from sklearn_extra.cluster import KMedoids
from hdbscan import HDBSCAN
import matplotlib.pyplot as plt
from scipy.spatial.distance import pdist, squareform

# Function to connect to the database and extract data
def get_data_from_db():
    # Pseudo-code for database connection and data extraction
    # This would be replaced with actual database connection code
    # For example using `sqlite3` or `sqlalchemy` for SQL
    databases
    data = np.random.rand(200, 2) # Placeholder for actual data
    return data

# Function to perform clustering
def perform_clustering(data, num_clusters):
    clusters = {
        'mean_shift': MeanShift().fit(data), # This is just a
        placeholder
        'k_means': KMeans(n_clusters=num_clusters).fit(data),
        'k_medoids': KMedoids(n_clusters=num_clusters).fit(data),
        'dbscan': DBSCAN().fit(data),
        'hierarchical':
        AgglomerativeClustering(n_clusters=num_clusters).fit(data),
        'cfsfdp_hd': HDBSCAN().fit(data) # HDBSCAN can be used as
        an example for CFSFDP-HD
    }
    return clusters

# Function to build a decision graph
def build_decision_graph(clusters):
    # Pseudo-code for building a decision graph
    # This would involve creating a graph that helps decide the

```

```

number of clusters
    pass

# Function to save results to a file
def save_results_to_file(clusters):
    with open('clustering_results.txt', 'w') as file:
        for method, cluster in clusters.items():
            file.write(f"{method}: {cluster.labels_}\n")

# Function to visualize clustering
def visualize_clustering(data, clusters):
    plt.figure(figsize=(15, 10))
    for i, (method, cluster) in enumerate(clusters.items(), 1):
        plt.subplot(2, 3, i)
        plt.scatter(data[:, 0], data[:, 1], c=cluster.labels_,
                    cmap='viridis', marker='.')
        plt.title(method)
    plt.show()

# Main program logic
def main():
    num_clusters = 4 # As per the requirement
    data = get_data_from_db()
    clusters = perform_clustering(data, num_clusters)
    build_decision_graph(clusters)
    save_results_to_file(clusters)
    visualize_clustering(data, clusters)

if __name__ == '__main__':
    main()

```

**ДОДАТОК Б**  
**Слайди презентації**

Національний університет «Запорізька політехніка»

Кафедра програмних засобів

## ДОСЛІДЖЕННЯ ТА ПРОГРАМНА РЕАЛІЗАЦІЯ МЕТОДІВ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ СТУДЕНТА В СИСТЕМІ ДИСТАНЦІЙНОГО НАВЧАННЯ

Виконав: студент гр. КНТ-142м Дмитро ЧЕРНІК

Керівник: доцент Ольга ГЛАДКОВА

Рисунок Б.1 – Слайд 1

### Мета та задачі проєкту

Метою роботи є застосування методів та алгоритмів інтелектуального аналізу даних для дослідження поведінки студента в системі електронної освіти.

Об'єкт дослідження – аналіз поведінки студента під час проходження курсу електронного навчання.

Предмет дослідження – методи та алгоритми інтелектуального аналізу даних у системі електронного навчання.

Виходячи з мети дослідження необхідно вирішити такі завдання:

- провести порівняльний аналіз існуючих способів та засобів аналізу поведінки студента під час проходження курсу електронного навчання;
- дослідити можливі критерії порівняння поведінки студентів;
- дослідити методи та алгоритми інтелектуального аналізу даних;
- спроектувати та реалізувати модель інтелектуального аналізу поведінки студента при проходженні курсу електронного навчання;
- перевірити ефективність розробленої моделі та визначити її результативність.

Рисунок Б.2 – Слайд 2

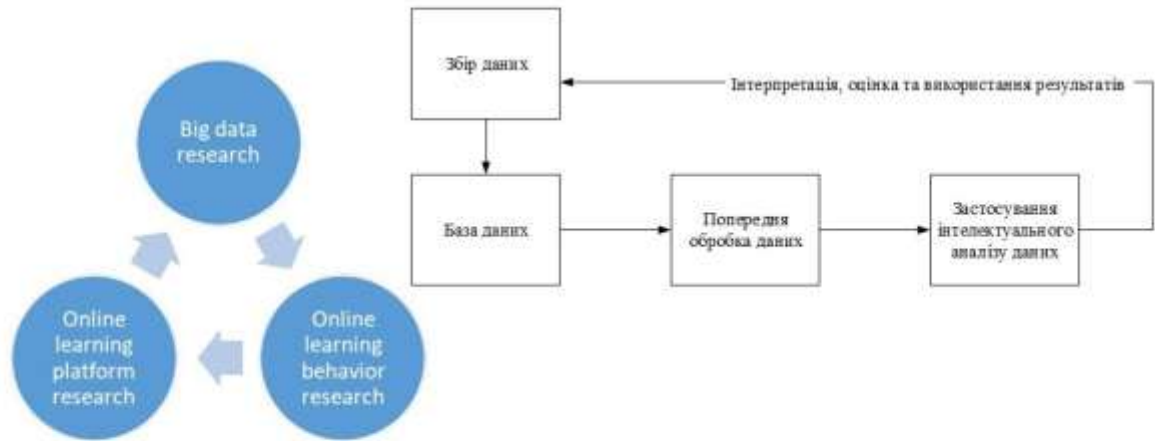
# Порівняння систем онлайн навчання Moodle, Canvas і Blackboard

| Система    | Вартість                       | Інтеграція                                         | Зручність користування | Налаштування                         | Підтримка баз даних                       | Користувачський інтерфейс       | Можливості інтелектуальної роботи  |
|------------|--------------------------------|----------------------------------------------------|------------------------|--------------------------------------|-------------------------------------------|---------------------------------|------------------------------------|
| Moodle     | Безкоштовно (Open Source)      | Висока (підтримує багато плагінів та інструментів) | Помірна                | Висока (повністю налаштовувана)      | MySQL, PostgreSQL, MariaDB, MSSQL, Oracle | Налаштовуваний, веб-базований   | Плагіни для аналітики та звітності |
| Canvas     | Змінна (залежить від установи) | Висока (широкий спектр інтеграцій)                 | Висока                 | Помірна (деякі опції налаштувань)    | PostgreSQL, MySQL, інші через LTІ         | Інтуїтивний, веб-базований      | Вбудовані інструменти аналітики    |
| Blackboard | Змінна (залежить від установи) | Висока (широкі можливості інтеграції)              | Висока                 | Помірна (обмежені опції налаштувань) | Oracle, MSSQL, MySQL, інші                | Стандартизований, веб-базований | Розширена аналітика та звітність   |

3

Рисунок Б.3 – Слайд 3

# Модель аналізу поведінки студента



4

Рисунок Б.4 – Слайд 4

## Дослідження алгоритмів аналізу даних

| Алгоритм   | Гнучкість в програмуванні                       | Розширення /<br>Пакети                | Підтримка алгоритмів | Візуалізація даних | Глибокий аналіз алгоритмів |
|------------|-------------------------------------------------|---------------------------------------|----------------------|--------------------|----------------------------|
| RapidMiner | Модульний підхід без необхідності програмування | Велика кількість операторів           | Широкий спектр       | Висока             | Обмежений                  |
| R          | Висока, можливість складного скриптування       | з ~10,000 пакетів-розширень           | Екстенсивна          | Висока             | Висока                     |
| Weka       | Обмежена, зосереджена на Java                   | Розширення через Java, інтеграція з R | Широкий спектр       | Висока             | Обмежений                  |
| KEEL       | Обмежена, фокусом на дослідженні                | з Модулів та плагінів                 | Широкий спектр       | Висока             | Обмежений                  |
| SNAPP      | Не застосовується                               | Не застосовується                     | Обмежена             | Обмежена           | Не застосовується          |

5

Рисунок Б.5 – Слайд 5

## Дані про студента

| Вид даних        | Опис                                               |
|------------------|----------------------------------------------------|
| Особисті дані    | біографічну інформацію про студентів               |
| Профіль студента | загальний прогрес в навчанні та поведінку студента |
| Когнітивні дані  | про пізнання учнів                                 |
| Педагогічні дані | стилі та методи навчання                           |

6

Рисунок Б.6 – Слайд 6

# Процес ETL

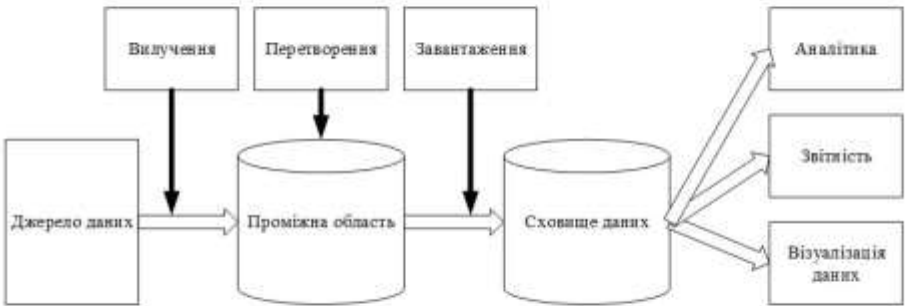


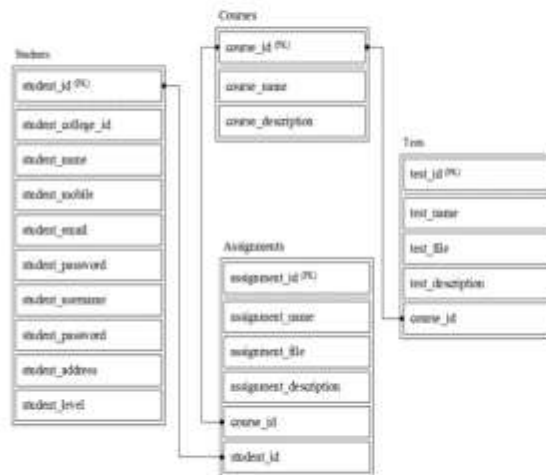
Рисунок Б.7 – Слайд 7

## Порівняння ключових ETL-інструментів від IBM, Informatica, та Microsoft

| ETL Tool                  | Інтеграція з IBM | Інтеграція з Unix | Інтеграція з Microsoft | Java | OLAP | Кросплат-форменість | Користу-вацький інтерфейс |
|---------------------------|------------------|-------------------|------------------------|------|------|---------------------|---------------------------|
| Data Stage (IBM)          | Так              | Ні                | Ні                     | Ні   | Так  | Ні                  | Продвинутий               |
| Data Manager (IBM)        | Так              | Ні                | Ні                     | Ні   | Так  | Ні                  | Продвинутий               |
| PowerCenter (Informatica) | Ні               | Так               | Ні                     | Так  | Ні   | Так                 | Продвинутий               |
| PowerMart (Informatica)   | Ні               | Так               | Ні                     | Так  | Ні   | Так                 | Продвинутий               |
| SSIS (Microsoft)          | Ні               | Ні                | Так                    | Ні   | Ні   | Ні                  | Продвинутий               |

Рисунок Б.8 – Слайд 8

## База даних



9

Рисунок Б.9 – Слайд 9

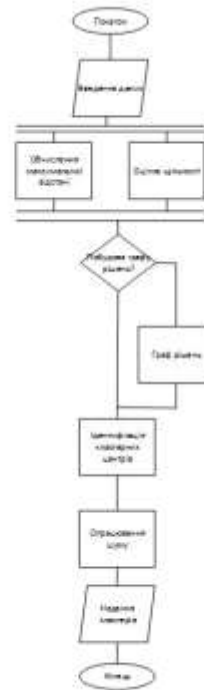
## Порівняння методів кластеризації даних

| Метод                    | Опис                                         | Чутливість до густини даних | Масштабованість | Вимоги до введення користувача     | Використання в електронному навчанні                               |
|--------------------------|----------------------------------------------|-----------------------------|-----------------|------------------------------------|--------------------------------------------------------------------|
| Mean Shift               | Непараметричний, виявляє щільні області      | Висока                      | Низька          | Розмір вікна                       | Не типово використовується                                         |
| K-Means                  | Поділяє дані на 'k' груп                     | Середня                     | Висока          | Кількість кластерів                | Підходить для попередньо визначених кластерів                      |
| K-Medoids                | Схожий на K-means, але використовує медоїди  | Висока                      | Середня         | Кількість кластерів                | Ефективний для даних із початковою установкою                      |
| DBSCAN                   | Групує точки даних у щільних областях        | Висока                      | Середня         | Порогова відстань                  | Гарно підходить для даних з різною щільністю                       |
| Ієрархічна кластеризація | Агломеративна, формує кластери злиттям точок | Низька                      | Низька          | Немає вимог до кількості кластерів | Підходить для експлоративного аналізу даних                        |
| CFSFDP                   | Групує області високої щільності             | Середня                     | Висока          | Параметри для розрахунку щільності | Корисний для інтуїтивного виявлення кластерів у освітніх даних     |
| CFSFDP-ND                | Покращує CFSFDP, дозволяє аналіз без знань   | Середня                     | Висока          | Параметри дифузії                  | Покращує CFSFDP для надійніших результатів у електронному навчанні |

10

Рисунок Б.10 – Слайд 10

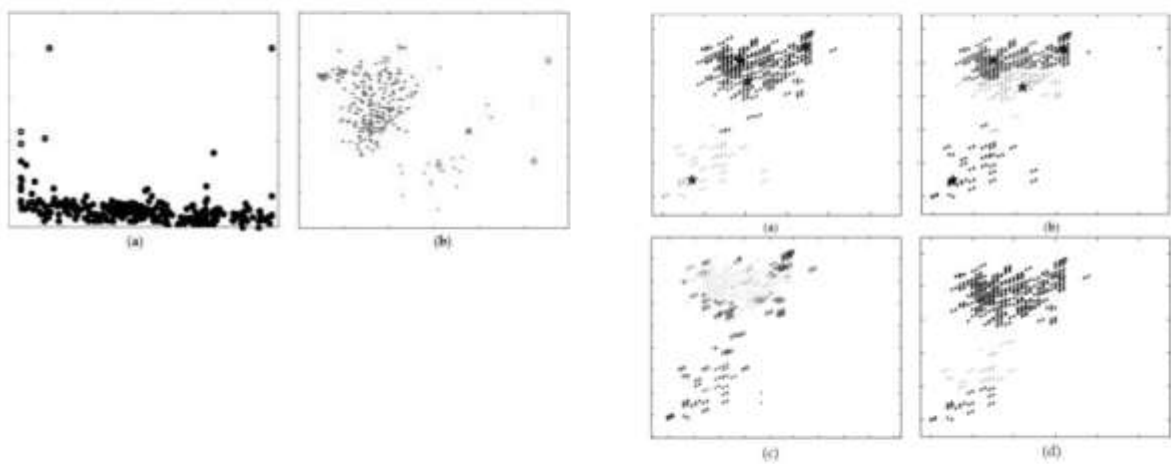
# Алгоритм



11

Рисунок Б.11 – Слайд 11

## Експериментальне дослідження методів



12

Рисунок Б.12 – Слайд 12

## Висновки

Були вирішені такі завдання:

- Проведено порівняльний аналіз існуючих способів та засобів аналізу поведінки студента під час проходження курсу електронного навчання;
- Досліджено можливі критерії порівняння поведінки студентів;
- Досліджено методи та алгоритми інтелектуального аналізу даних;
- Спроектвана та реалізована модель інтелектуального аналізу поведінки студента при проходженні курсу електронного навчання;
- Перевірена ефективність розробленої моделі та визначена її результативність.

Практична значущість роботи полягає у розробці програмної реалізації комбінованих моделей для оцінки поведінки студентів у системі дистанційного навчання. Це дозволяє з більшою точністю ідентифікувати потенційні проблеми у навчальному процесі та покращувати методики навчання.

Наукова новизна дослідження полягає у використанні методів та алгоритмів інтелектуального аналізу даних для оцінки поведінки студента у системі електронної освіти.