

Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет інформаційно-аналітичних технологій та менеджменту

(повна назва)

Кафедра прикладної математики

(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Застосування методів інтелектуального аналізу даних
для оцінювання рівня знань випускників закладів загальної
середньої освіти

(тема)

Виконав:

здобувач 2 року навчання, групи САУм-24-1

Катерина ШАРАЙ

(Власне ім'я, ПРІЗВИЩЕ)

Спеціальність

124 Системний аналіз

(код і повна назва спеціальності)

Тип програми освітньо-професійна

(освітньо-професійна або освітньо-наукова)

Освітня програма

Системний аналіз і управління

(повна назва освітньої програми)

Керівник доц. Надія ГИБКІНА

(посада, Власне ім'я, ПРІЗВИЩЕ)

Допускається до захисту

Завідувач кафедри ПМ

(підпис)

Максим СИДОРОВ

(Власне ім'я, ПРІЗВИЩЕ)

2025 р.

Харківський національний університет радіоелектроніки

Факультет інформаційно-аналітичних технологій та менеджменту

Кафедра прикладної математики

Рівень вищої освіти другий (магістерський)

Спеціальність 124 Системний аналіз

(код і повна назва)

Тип програми освітньо-професійна

(освітньо-професійна або освітньо-наукова)

Освітня програма Системний аналіз і управління

(повна назва)

ЗАТВЕРДЖУЮ:

Завідувач кафедри _____

(підпис)

“ 10 ” листопада 2025 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Шарай Катерині Валентинівні
(прізвище, ім'я, по батькові)

1. Тема роботи Застосування методів інтелектуального аналізу даних для
оцінювання рівня знань випускників закладів загальної середньої освіти

затверджена наказом по університету від 10 листопада 2025 р. № 1027 Ст

2. Термін подання здобувачем роботи до екзаменаційної комісії 18 грудня 2025 р.

3. Вихідні дані до роботи результати складання ЗНО та НМТ з математики
у 2021 – 2024 роках у закладах середньої освіти у м. Харків

4. Перелік питань, що потрібно опрацювати в роботі _____

1. Системний аналіз предметної області

2. Вибір і обґрунтування методу розв'язання

3. Програмна реалізація

4. Результати обчислювального експерименту

5. Аналіз можливих застосувань

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій _____

1. Актуальність теми роботи _____

2. Постановка задачі _____

3. Системний аналіз предметної області _____

4. Метод чисельного аналізу _____

5. Результати обчислювального експерименту _____

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Підбір та вивчення технічної літератури за темою роботи	10 – 16 листопада 2025 р.	виконано
2	Вибір та обґрунтування методу	17 – 23 листопада 2025 р.	виконано
3	Розробка алгоритму і програми	24 – 30 листопада 2025 р.	виконано
4	Проведення аналітичних досліджень та розрахунків	01 – 07 грудня 2025 р.	виконано
5	Робота над текстом пояснювальної записки	08 – 17 грудня 2025 р.	виконано
6	Представлення роботи на рецензію в ЕК	18 грудня 2025 р.	виконано

Дата видачі завдання 10 листопада 2025 р.

Здобувач _____
(підпис)

Керівник роботи _____ доц. Надія ГИБКІНА
(підпис) (посада, Власне ім'я, ПРІЗВИЩЕ)

РЕФЕРАТ

Пояснювальна записка: 130 с., 20 табл., 30 рис., 2 дод., 30 джерел.

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ, ЗОВНІШНЄ НЕЗАЛЕЖНЕ ОЦІНЮВАННЯ, КЛАСТЕРИЗАЦІЯ, МЕТОД ГОЛОВНИХ КОМПОНЕНТ, МЕТОДИ НЕІЄРАРХІЧНОЇ КЛАСТЕРИЗАЦІЇ, НАЦІОНАЛЬНИЙ МУЛЬТИ-ПРЕДМЕТНИЙ ТЕСТ, ОЦІНЮВАННЯ НАВЧАЛЬНИХ ДОСЯГНЕНЬ, ПОПЕРЕДНЯ ОБРОБКА ДАНИХ, СТАНДАРТИЗАЦІЯ АТРИБУТІВ.

Об'єкт дослідження – навчальні досягнення випускників закладів середньої освіти м. Харків за 2021 – 2024 роки.

Мета роботи – аналіз показників успішності складання ЗНО та НМТ випускниками навчальних закладів м. Харків у 2021 – 2024 роках для дослідження якості освітньої підготовки з математики та виявлення тенденцій у її змінах під впливом зовнішніх чинників.

Методи дослідження – компонентний аналіз і неієрархічні ітеративні методи кластеризації.

Виконано інтелектуальний аналіз результатів ЗНО та НМТ за 2021 – 2024 роки у м. Харків та визначено ключові закономірності в академічній підготовці випускників до іспиту з математики в умовах пандемії та війни. Для вивчення навчальних досягнень учнів одинадцятого класу були застосовані неієрархічні методи кластерного аналізу, які за підсумками системного аналізу задачі визначені оптимальними за низкою критеріїв. Перед проведенням кластеризації було здійснено попередню обробку датасету, щоб покращити якість даних та забезпечити коректні результати.

Процес розв'язання поставленої задачі автоматизовано за допомогою програми, розробленої у математичному пакеті Wolfram Mathematica 10.4. Було виконано компонентний аналіз даних, у ході якого було вивчено особливості їх геометричної форми. Набори даних з балами ЗНО та НМТ з математики за 2021

– 2024 роки було розбито на п'ять кластерів за принципом подібності навчальних досягнень учнів, досліджено розподіл балів між закладами середньої освіти кожного кластера та зроблено висновки щодо ефективності підготовки випускників до іспиту з математики протягом чотирьох років. Також було проаналізовано щорічні зміни у складі та структурі кластерів та зафіксовано тенденції до погіршення результатів складання тестів та якості освітньої діяльності. У ході аналізу метрик якості кластерів було підтверджено локальні особливості розглянутих даних: витягнутість кластерів, висока щільність суттєвої частини даних тощо.

Результати виконання кваліфікаційної роботи можуть бути використані при створенні стратегій розвитку галузі середньої освіти в декількох напрямках, зокрема, для виявлення потенційно проблемних закладів з подальшими діями щодо підвищення рівня навчальної підготовки а також під час оптимізації умов складання тесту.

ABSTRACT

Introductory note: 130 pages, 20 tables, 30 figures, 2 appendixes, 30 sources.

ASSESSMENT OF GRADUATES' ACADEMIC ACHIEVEMENTS, CLUSTERING, DATA PREPROCESSING, EXTERNAL INDEPENDENT EVALUATION, INTELLECTUAL DATA ANALYSIS, NATIONAL MULTIPLE-SUBJECT TEST, NON-HIERARCHICAL CLUSTERING METHODS, PRINCIPAL COMPONENT METHOD, STANDARDISATION OF ATTRIBUTES.

Object of research – the academic achievements of secondary school graduates in Kharkiv in 2021 – 2024.

Purpose of work – the analysis of the performance indicators of graduates from secondary educational institutions in Kharkiv in 2021 – 2024 aims to study the quality of educational training in mathematics and identify trends in its changes under the influence of external factors.

Methods of research – principal components analysis and non-hierarchical clustering, internal measures of cluster validity: silhouette coefficient, Calinski–Harabasz index and Dunn index.

In the study an intellectual analysis of the results of the External Independent Evaluation (EIE) and National Multiple-subject Test (NMT) in 2021 – 2024 in Kharkiv is carried out, and key patterns in the academic preparation of graduates for the mathematics exam in the emergency conditions of the pandemic and war are identified. To study the academic achievements of the eleventh-grade students, non-hierarchical cluster analysis methods are applied, that, based on the results of a system analysis of the task, are determined to be optimal according to a number of criteria. Before clustering the dataset, it was pre-processed to improve its initial quality and ensure correct analysis results.

The process of solving the task, in particular, calculations and data operations, was fully automated using a program developed in the Wolfram Mathematica 10.4

mathematical package. The component analysis of the data is performed, and during it the peculiarities of their geometric shape are studied. The dataset with EIE and NMT scores obtained by graduates of the secondary educational institutions in Kharkiv in 2021–2024 in mathematics is processed and divided into five clusters based on the similarity of graduates' academic achievements, the distribution of scores among the secondary educational institutions in each cluster is studied separately, and conclusions are made regarding the effectiveness of graduates' preparing for the mathematics exam in Kharkiv over a four-year period. Annual changes in the composition and structure of clusters are also analysed, and trends towards a deterioration in test results and the quality of educational activities were identified. The results of the clustering of secondary educational institutions in Kharkiv are evaluated not only qualitatively, in terms of their practical content, but also quantitatively using internal validation measures. During the analysis of cluster quality metrics, the local characteristics of the data under consideration are additionally confirmed: the elongation of clusters, the high density of a significant part of the dataset, etc.

The results of the qualification work can be used to create strategies for the development of secondary education in several areas, in particular, to identify potentially problematic institutions with further actions to improve the level of educational training, as well as when optimizing the conditions for taking the test.

ЗМІСТ

	С.
Вступ	10
1 Системний аналіз предметної області та постановка задач дослідження	13
1.1 Системний аналіз задачі інтелектуального аналізу балів ЗНО	
та НМТ за 2021 – 2024 роки у м. Харків	13
1.1.1 Морфологічний опис системи	13
1.1.2 Функціональна модель системи	16
1.1.3 Інформаційна модель системи	19
1.2 Аналіз сценаріїв вирішення задачі інтелектуального аналізу балів ЗНО	
та НМТ за 2021 – 2024 роки у м. Харків.....	22
1.3 Змістовна та формальна постановка задачі	34
1.3.1 Змістовна постановка задачі	34
1.3.2 Формальна постановка задачі	35
1.4 Постановка задач дослідження	36
2 Вибір та обґрунтування методу розв’язання	38
2.1 Попередня обробка даних	38
2.2 Неієрархічні ітеративні методи кластерного аналізу даних	41
2.2.1 Метод k -means	42
2.2.2 Метод k -medoids.....	43
2.3 Внутрішні міри оцінки валідності кластерів.....	44
2.4 Застосування методів інтелектуального аналізу даних до розв’язання	
задачі оцінювання навчальних досягнень випускників ЗСО м. Харків за	
2021 – 2024 роки.....	48
2.4.1 Збір статистичних даних та їх передобробка	49
2.4.2 Кластеризація даних методами k -means та k -medoids	51
Висновки за розділом 2	52
3 Програмна реалізація	54
3.1 Mathematica 10 як система комп’ютерної алгебри	54

3.2 Алгоритм розв’язання задачі оцінювання навчальних досягнень випускників ЗСО м. Харків у 2021 – 2024 роках методами неієрархічної кластеризації	55
3.3 Опис програми	56
Висновки за розділом 3	58
4 Результати обчислювального експерименту та їх аналіз	59
4.1 Компонентний аналіз результатів ЗНО та НМТ за 2021 – 2024 роки	59
4.2 Неієрархічна кластеризація ЗСО м. Харків за результатами ЗНО та НМТ за 2021 – 2024 роки з математики.....	66
4.3 Порівняльний аналіз кластерів для дослідження змін рівня навчальної підготовки з математики у ЗСО м. Харків за 2021 – 2024 роки	87
4.4 Оцінювання якості кластеризації за внутрішніми мірами валідності кластерів	98
Висновки за розділом 4	102
Висновки	104
Перелік джерел посилання	106
Додаток А Лістинг програми	109
Додаток Б Назви та аббревіатури ЗСО м. Харків	121

ВСТУП

Актуальність роботи. Актуальність роботи зумовлена важливістю створення та підтримки системи середньої освіти, яка сприятиме отримуванню якісних знань та допоможе закласти міцний фундамент для майбутнього вступу до університету та досягнення професії. Бали складання зовнішнього незалежного оцінювання (ЗНО) та національного мультипредметного тесту (НМТ) у різні роки характеризують не тільки індивідуальний ступінь підготовки кожного окремого учня, а і загальну ефективність освітньої діяльності школи, з якої він випускається. Отже, дослідження результатів випускних іспитів допоможе сформулювати науково обґрунтовані висновки щодо якості навчання у школах та розробити відповідні стратегії її підвищення і збереження, а також відслідкувати панівні тенденції розвитку освітньої галузі у країні.

На початку 20-х років XXI століття освітня система України стикнулася з чисельними масштабними викликами: організація навчального процесу спочатку в умовах пандемії, а згодом – повномасштабної війни. Моніторинг навчальних досягнень випускників сприятиме не тільки розробці якісних освітніх програм або спеціалізованих заходів для вдосконалення компетенцій викладачів. Він також підсвітить глобальні проблеми сучасної освітньої інфраструктури, пов'язані з питаннями створення безпечного для життя учнів і вчителів, ефективного освітнього середовища. Аналіз балів ЗНО та НМТ може бути аргументом для змінення формату проведення занять, запровадження заходів щодо підвищення рівня якості підготовки учнів та вчителів, оптимізації умов написання іспитів таким чином, аби забезпечити учням безпечний фізично та психологічно простір для їх складання без перешкод, наприклад, у вигляді переривання тесту через довгу повітряну тривогу. Виходячи з цього, тема кваліфікаційної роботи є актуальною та має високу прикладну значущість для розвитку сфери середньої освіти.

Мета і завдання кваліфікаційної роботи. Метою кваліфікаційної роботи є аналіз показників успішності складання ЗНО та НМТ випускниками навчальних закладів м. Харків у 2021 – 2024 роках для дослідження якості освітньої

підготовки з математики та виявлення тенденцій у її змінах під впливом зовнішніх чинників.

Комплексний аналіз результатів складання випускних випробувань та динаміки їх змін надає змогу оцінити якість навчання у закладах середньої освіти та, спираючись на зроблені висновки, розробити практичні рекомендації для її посилення та підтримки на стабільно високому рівні. Враховуючи специфіку розглядуваного часового періоду, аналіз отриманих учнями балів ЗНО та НМТ має важливе значення для визначення рівня підготовленості освітньої інфраструктури до ефективної діяльності в кризових ситуаціях.

Для досягнення поставленої мети необхідно виконати наступні завдання:

- провести системний аналіз задачі «інтелектуальний аналіз балів зовнішнього незалежного оцінювання і національного мультипредметного тесту за 2021 – 2024 роки у м. Харків» та за його результатами обрати метод розв’язання поставленої задачі;
- ознайомитися з теоретичними концепціями методів інтелектуального аналізу, що будуть застосованими для дослідження вихідних даних;
- розробити програму для розв’язання поставленої задачі у системі комп’ютерної алгебри Wolfram Mathematica 10.4;
- виконати дослідження даних обраними методами та оцінити якість кластеризації;
- провести аналіз отриманих результатів та сформулювати прикладні висновки щодо якості освітньої підготовки у розглядуваний період часу.

Об’єктом дослідження є навчальні досягнення випускників закладів середньої освіти м. Харків за 2021 – 2024 роки.

Предметом дослідження є методи інтелектуального аналізу даних та їх застосування для розв’язання задачі оцінювання результатів ЗНО та НМТ, отриманих випускниками закладів середньої освіти м. Харків протягом 2021 – 2024 років.

Методи дослідження. У роботі використовуються метод головних компонент (РСА) для зменшення розмірності датасету, неієрархічні методи класте-

рного аналізу, внутрішні міри валідності кластерів: силуетний коефіцієнт, індекси Калінські – Харабаса та Данна.

Публікації. Результати, отримані у роботі, було представлено на 29-му Міжнародному молодіжному форумі «Радіoeлектроніка та молодь у ХХІ столітті» (м. Харків, 16 – 18 квітня 2025 р.) [1], на IV Міжнародній науково-практичній конференції англійською мовою «Європейські студії. Навчання і викладання у світі технологій» (м. Харків, м. Клуз-Напока, Румунія, 12 листопада 2025 р.) [2], на IX Міжнародній науково-технічній конференції «Комп’ютерне моделювання та оптимізація складних систем КМОСС–2025» (м. Дніпро, 5 – 7 листопада) [3], а також на Всеукраїнській науково-практичній конференції здобувачів вищої освіти і молодих учених «Комп’ютерно-інтегровані технології автоматизації технологічних процесів на транспорті та у виробництві» (м. Харків, 25 листопада 2025 р.) [4].

1 СИСТЕМНИЙ АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАДАЧ ДОСЛІДЖЕННЯ

1.1 Системний аналіз задачі інтелектуального аналізу балів ЗНО та НМТ за 2021 – 2024 роки у м. Харків

1.1.1 Морфологічний опис системи

Об’єкт аналізу – «задача інтелектуального аналізу балів ЗНО та НМТ за 2021 – 2024 роки у м. Харків».

Предмет аналізу – «використання методів інтелектуального аналізу даних для дослідження результатів ЗНО та НМТ за 2021 – 2024 роки у м. Харків з метою оцінки ефективності навчальної підготовки випускників».

Точка зору: дослідник.

Призначення системи: підготовка якісного датасету шляхом його очищення та трансформації, аналіз оброблених даних з метою знаходження у них неочевидних закономірностей, оцінка динаміки навчальних досягнень учнів від року до року, інтерпретація здобутих інсайтів та розробка практичних рекомендацій для покращення освітньої підготовки.

Метою системи є виконання порівняльного аналізу показників ЗНО та НМТ за 2021 – 2024 роки для закладів середньої освіти (ЗСО) м. Харків з метою оцінки якості освітньої підготовки, вдосконалення освітнього процесу та оптимізації умов складання випускних тестів з урахуванням безпекової ситуації.

Розглядувану систему можна класифікувати як:

- штучну, адже її було створено людиною;
- абстрактну, або символічну, так як вона складається з математичних формул, алгоритмів і гіпотез;
- механічну (неживу) за природою походження;
- децентралізовану, бо у розв’язуванні задачі методи інтелектуального аналізу застосовуються послідовно один за одним, кожен з них має однакову

важливість для обробки даних, а мінімальні зміни на кожному етапі можуть суттєво вплинути на кінцевий результат;

- гетерогенну (різномірну), оскільки використані методи інтелектуального аналізу не є взаємозамінними, відсутність навіть одного з них може спричинити загибель системи, зробивши задачу неможливою для розв’язання;

- нелінійну, адже відношення між входами та виходами системи описуються нелінійними залежностями;

- каузальну, бо її поведінка визначається заданою дослідником метою та зовнішніми умовами;

- велику: для розв’язання поставленої задачі використовується значна кількість різних підходів до аналізу даних;

- складну, оскільки її елементами виступають багатокрокові алгоритми аналізу даних з численними взаємозалежностями;

- відкриту, оскільки вона функціонує у взаємодії з зовнішнім середовищем, зокрема через використання інформаційних ресурсів та обчислювальної техніки;

- неієрархічну, так як застосовані методи інтелектуального аналізу даних є рівноправними між собою і не перебувають у відношеннях підпорядкування;

- керовану ззовні, бо аналіз даних проводиться дослідником, який не вважається частиною системи;

- динамічну, так як число методів інтелектуального аналізу даних постійно зростає разом із розвитком технологій машинного навчання.

Задача інтелектуального аналізу показників ЗНО та НМТ у м. Харків за 2021 – 2024 роки моделюється як відкрита система, що активно взаємодіє із зовнішнім середовищем. Її функціонування підтримується за допомогою технічних (наприклад, програмного забезпечення для обробки даних), соціальних (діяльність ЗСО), інформаційних (дані Українського центру оцінювання якості освіти) та регулятивних (освітні стандарти) ресурсів. Їх детальний опис представлено на рисунку 1.1. Із застосуванням зазначених засобів у системі генерується вихід у вигляді результатів аналізу навчальних досягнень випускників. Водночас з цим, отримані результати можуть вплинути на освітню політику та

інфраструктуру у Харкові, і відповідно, на показники ЗНО та НМТ у перспективі (рисунок 1.1). Таким чином, система утворює цикл зворотного зв'язку, що сприяє адаптації освіти до сучасних викликів.



Рисунок 1.1 – Зовнішнє середовище системи

У рамках системного аналізу система моделюється як «чорна скриня». Вона характеризується зосередженням фокусу уваги на елементах зовнішнього середовища системи, зокрема, її входах та виходах [5]. Внутрішня побудова та характеристики системи, взаємодія її складових між собою і процеси, що виконуються у ній, лишаються поза розглядом дослідника. На вхід розглядуваної системи подаються «брудні» дані з розподілом балів ЗНО та НМТ випусників ЗСО м. Харків за 2021 – 2024 роки. У результаті здійснення низки їх перетворень на виході отримуємо групи ЗСО, сформовані за критеріями подібності, та їх порівняльну оцінку (рисунок 1.2).

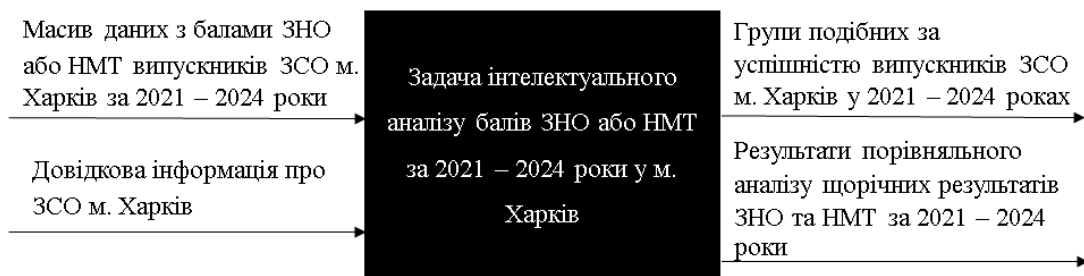


Рисунок 1.2 – Модель «Чорна скриня»

1.1.2 Функціональна модель системи

У ході свого життєвого циклу кожна система виконує певну низку функцій – дій, виходячи із свого призначення. Шляхом реалізації функцій система досягає мети свого існування. Для глибокого дослідження функціональних можливостей системи часто будуються графічні моделі, які легко сприймаються за рахунок своєї наочності. Однією з найбільш популярних методологій графічного опису функцій системи є IDEF0 (Integration Definition For Function Modeling) – нотації для побудови функціональної моделі системи. Відповідно до неї модель системи складається з набору діаграм, де функції зображуються у вигляді прямокутників (робіт), а взаємозв'язки між ними (входи та виходи) – у вигляді дуг.

Ієрархічні зв'язки між блоками моделі показуються за допомогою нумерації відповідно до їх рівня. Першою будується діаграма рівня А-0, яка називається контекстною. Вона відображає функцію цілі системи та ресурси, що забезпечують її реалізацію. Стрілки входу розташовуються зліва та показують об'єкти, що перетворюються системою у ході виконання її цільової функції. Результати діяльності системи утворюють її виходи, які зображуються з правої сторони роботи. Стрілки, направлені зверху до блоку, позначають засоби контролю, наприклад, стандарти, закони чи інструкції, що строго регламентують функціонування системи. До нижньої межі спрямовані стрілки механізмів – людських, технічних та інших ресурсів, використаних для виконання цільової функції [6].

На рисунку 1.3 представлено контекстну діаграму для задачі інтелектуального аналізу балів ЗНО та НМТ за 2021 – 2024 роки у м. Харків. У розглядваній системі вхід складають дані з розподілом балів ЗНО та НМТ випускників харківських ЗСО протягом 2021 – 2024 років, а також довідкова інформація щодо перейменування ЗСО чи зміни їх статусу. У результаті здійснення системою функції цілі початкові ресурси трансформуються на виходи – графічно представлені результати аналізу даних, зокрема, поділ ЗСО м. Харків на кластери за принципом схожості рівня навчальної підготовки у них, розподіл балів

всередині кожного з кластерів, графіки з підсумками дослідження змін у складі кластерів протягом чотирьох років та результати дисперсійного аналізу отриманих кластерів. Досліджувана система керується алгоритмами машинного навчання для попередньої обробки даних та їх кластеризації, критеріями перевірки статистичних гіпотез і документацією із синтаксисом мови символічної математики Wolfram. Окрім того, хід розв’язання задачі регламентується щорічними протоколами проведення ЗНО та НМТ. Усі дії у системі виконуються за допомогою дослідника та комп’ютерної техніки з встановленим на ній програмним забезпеченням для автоматизації обчислень.

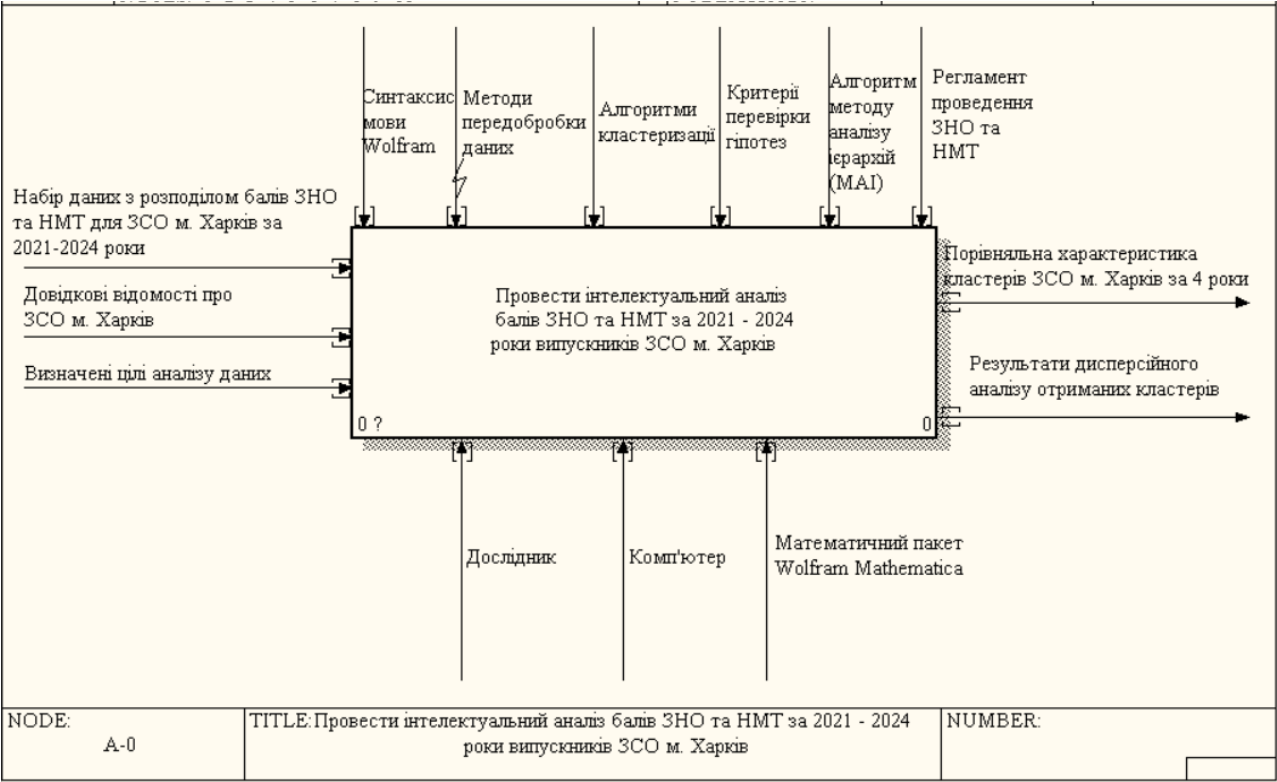


Рисунок 1.3 – Контекстна діаграма рівня А-0

Для вивчення основних функцій системи проводиться декомпозиція контекстної діаграми. Деталізація діаграми рівня А-0 надає змогу розглянути функції системи та краще зрозуміти її діяльність.

Декомпонуємо основний блок «Провести інтелектуальний аналіз балів ЗНО та НМТ за 2021 – 2024 роки у м. Харків» (рисунок 1.4). Бачимо, що досяг-

нення цільової функції забезпечується п'ятьма основними функціями. На початку підбирається найбільш підходящий в межах поставленої задачі метод для її розв'язання. Після цього розробляється програма для автоматизації методів обробки даних, а саме: їх попередньої обробки та застосування до них ітеративних методів кластеризації. Наступним кроком проводиться тестування створеного програмного продукту на початкових даних. Під час вивчення результатів запуску програми встановлюється їх статистична значущість шляхом перевірки гіпотез за відповідними критеріями. На останньому етапі отримані кластери критично аналізуються на предмет динаміки їх змін протягом розглядуваних чотирьох років.

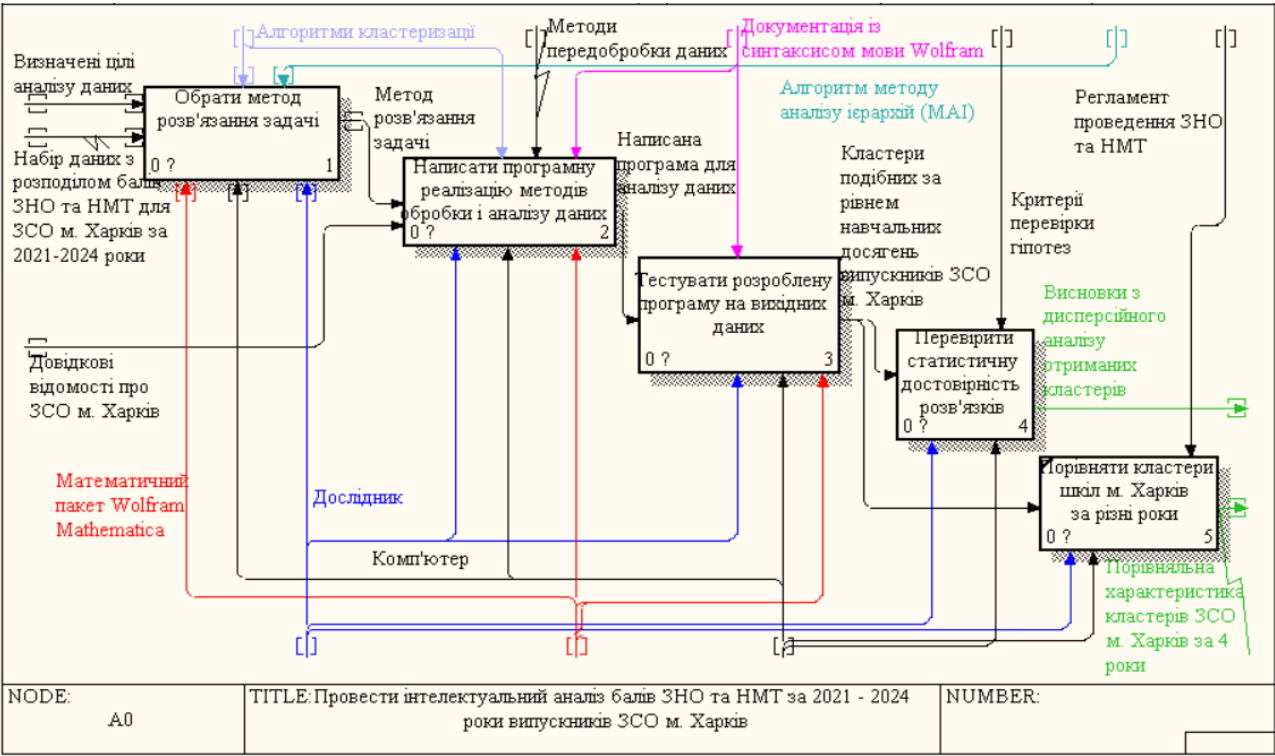


Рисунок 1.4 – Декомпозиція контекстної діаграми

Побудуємо діаграму рівня A1 для уточнення кроків процесу «Обрати метод розв'язання задачі» (рисунок 1.5).

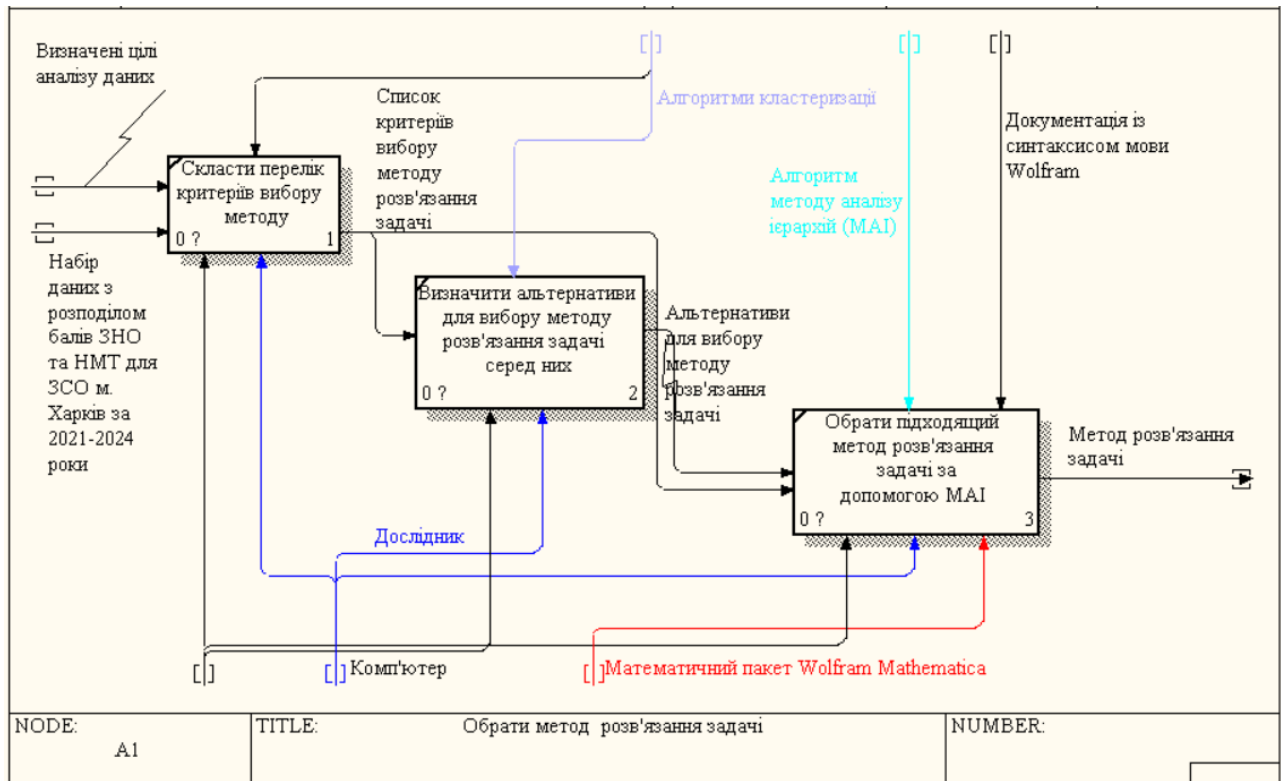


Рисунок 1.5 – Діаграма рівня А1 для функції «Обрати метод розв'язання задачі»

1.1.3 Інформаційна модель системи

Під час функціонування всередині системи відбувається обмін великою кількістю інформації. Процес обміну можна подати графічно за допомогою DFD-діаграм (DFD або Data Flow Diagram – діаграма потоків даних). Для зображення процесів, джерел, сховищ інформації використовують геометричні фігури: прямокутники або круги (залежно від застосованої нотації), а потоки даних між зазначеними об'єктами у системі зображуються у вигляді стрілок [7].

DFD-діаграми можуть складатися з компонент чотирьох типів. Головним елементом будь-якої інформаційної моделі є процес. Він позначає певну дію у системі, яка перетворює інформацію, формуючи нові дані на виході. Зображується процес за допомогою прямокутника з округленими кутами, поділеного на дві секції лінією. Джерела та приймачі інформації, що передається у системі,

називаються зовнішніми сутностями. Вони зображуються за допомогою звичайних прямокутників. Під час діяльності системи вхідні та вихідні дані накопичуються, тому часто вони потребують збереження на окремих носіях – сховищах даних. На діаграмі потоків даних ці елементи мають вигляд прямокутників, розділених на 2 частини вертикальною лінією. Дані, які передаються між процесами, зовнішніми сутностями та сховищами, формують інформаційні потоки системи. У будь-якій нотації DFD-діаграми їх подають стрілками.

Отже, використовуючи перелічені елементи, побудуємо DFD-діаграму системи за методом Гейна-Сарсона (рисунок 1.6). Особливість використаної нотації полягає у поданні процесів прямокутниками з круглими кутами.

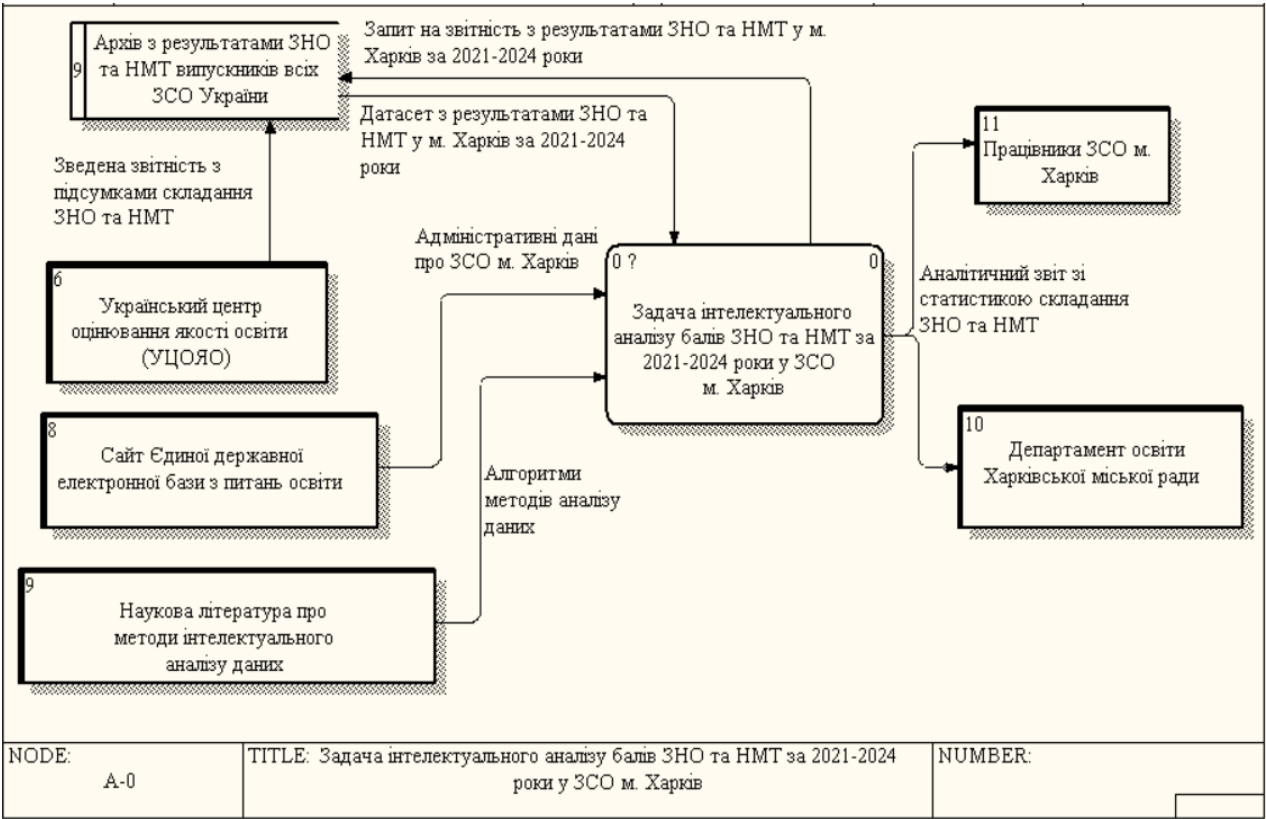


Рисунок 1.6 – DFD-діаграма системи «Задача інтелектуального аналізу балів ЗНО та НМТ за 2021 – 2024 роки у м. Харків»

Розглянемо детальніше створену діаграму потоків даних. Її центральним елементом є процес – проведення інтелектуального аналізу балів ЗНО та НМТ за 2021 – 2024 роки у м. Харків. Згідно з методологією DFD будь-який процес

потребує вхідних потоків інформації, які він трансформує та генерує виходи. Входом виступає масив даних з результатами випускних тестів, отриманих учнями ЗСО Харкова у 2021 – 2024 роках. Дослідник отримує його за запитом зі сховища даних – архіву УЦОЯО (Український центр оцінювання якості освіти). На сайті Єдиної державної електронної бази з питань освіти наводяться офіційні відомості про перейменування та зміну статусу ЗСО, які знадобляться для формування висновків з порівняльного аналізу кластерів ЗСО у різні роки. Також, важливою вхідною інформацією виступають методи інтелектуального аналізу даних, які можна знайти у науковій тематичній літературі. У результаті виконання процесу на вихід формуються аналітичні звіти зі статистикою навчальних досягнень випускників ЗСО м. Харків. Вони передаються працівникам Департаменту освіти Харківської міської ради або працівникам ЗСО.

На рисунку 1.7 представлено DFD-діаграму рівня A0 з потоками інформації, необхідної для реалізації головних функцій системи.

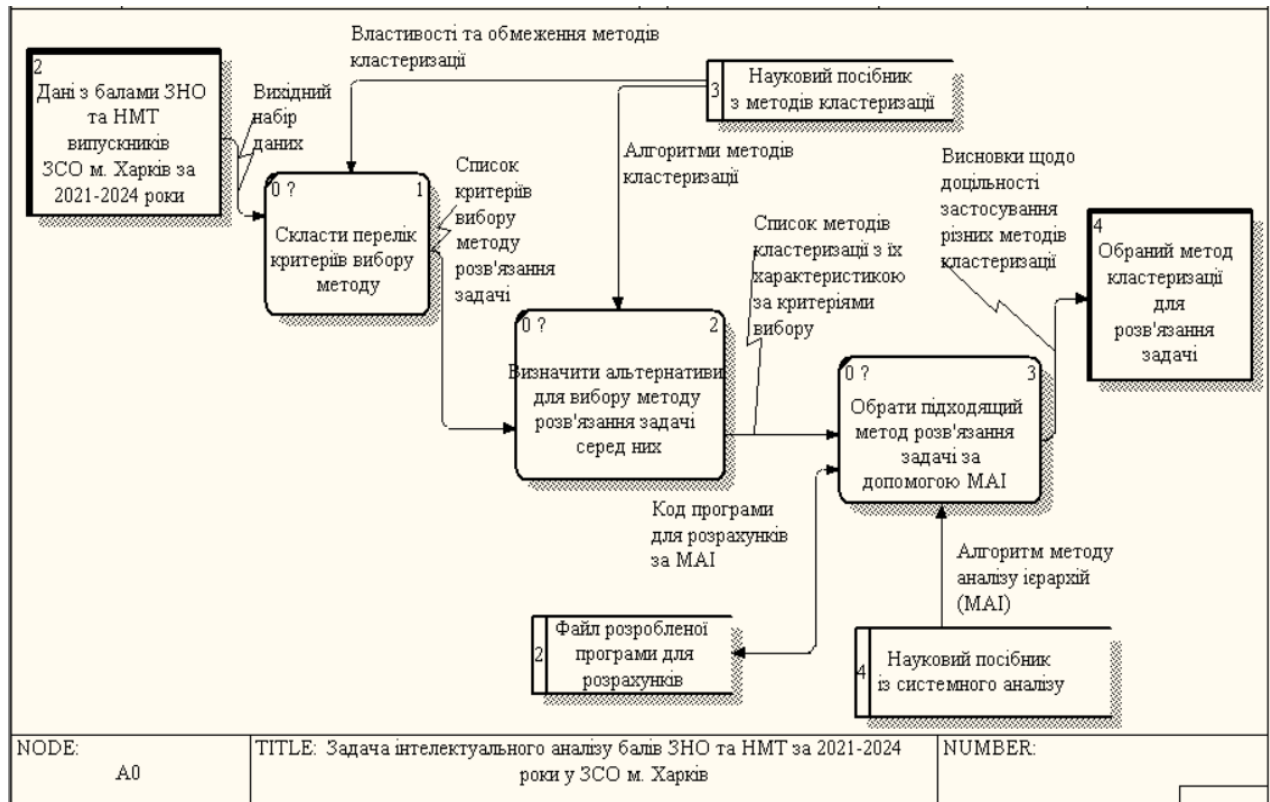


Рисунок 1.7 – DFD-діаграма рівня A0

1.2 Аналіз сценаріїв вирішення задачі інтелектуального аналізу балів ЗНО та НМТ за 2021 – 2024 роки у м. Харків

Робота присвячена дослідженню якості освітньої підготовки з урахуванням зовнішніх кризових обставин та вивченню динаміки її змін від року до року. Початковими даними розв’язуваної задачі є таблиці з деперсоналізованою інформацією про зареєстрованих на ЗНО та НМТ випускників ЗСО м. Харків, зокрема, назвою та типом навчального закладу, який вони завершили, а також набраними балами ЗНО та НМТ з різних навчальних дисциплін. Ідея розв’язання поставленої задачі полягає у розбитті ЗСО на множини, елементи яких характеризуються схожим рівнем навчальних досягнень випускників, та порівнянні їх складу, отриманого для різних років. Для її реалізації найбільш доречним рішенням буде застосувати методи кластеризації.

Методи кластерного аналізу дозволяють розв’язувати широкий спектр прикладних задач. Універсальність застосування цього підходу посприяла розробці більше сотні алгоритмів кластеризації. Вибір підходящого методу кластерного аналізу серед різноманіття існуючих часто стає проблематичним, адже він залежить від багатьох факторів. Розглянемо три класи методів, що характеризуються різними принципами формування кластерів: неієрархічні ітеративні, ієрархічні методи та методи на основі ймовірнісного розподілу.

Ідея методів неієрархічного кластерного аналізу полягає у поділі датасету на заздалегідь визначену кількість груп, що не перетинаються та складаються зі схожих за певними критеріями об’єктів. Кластери формуються за ітеративною процедурою, що полягає у мінімізації відстані між точками датасету та центроїдами – точками, які є усередненням ознак елементів кожного кластера.

Популярність використання неієрархічних методів кластерного аналізу пояснюється їх широкими перевагами, зокрема високою швидкістю здійснюваних обчислень, елементарністю математичних розрахунків, а також існуванням алгоритмів з відносною стійкістю до викидів. Водночас з тим, чутливість до структури датасету та потреба у попередньому встановленні майбутньої кілько-

сті кластерів створюють обмеженість використання ітеративної кластеризації для розв'язання складних задач. Значна ймовірність збіжності неієрархічних методів до локальних екстремумів є ще одним недоліком, який може призвести до неврахування важливих закономірностей у датасеті та його некоректного розподілу на категорії [8].

Ієрархічна кластеризація полягає у формуванні дерева кластерів. Отримані результати подаються графічно у вигляді дендрограми, рівні якої характеризують взаємозв'язки між об'єктами утворених кластерів. Методи ієрархічної кластеризації класифікуються за підходом до створення кластерів: агломеративні (об'єднуючі) та дивізивні (розділяючі). На першому етапі агломеративних алгоритмів кожен об'єкт датасету вважається окремим кластером. У ході об'єднання покроково зливаються по два найближчі за обраною відстанню кластери до моменту утворення єдиного кластера з усіх точок набору даних. Дивізивний підхід є протилежним до агломеративного: на початку кластеризації існує один кластер – вихідний масив даних – який надалі розщеплюється на менші кластери, доки вони не будуть містити в собі по одному об'єкту.

Методи ієрархічної кластеризації мають низку переваг відносно інших алгоритмів. Серед них гнучкість у виборі числа кластерів, необхідного для розв'язуваної задачі, можливість роботи з даними несферичної форми та відображення їх багаторівневої структури, а також наочне представлення кластерів у вигляді дендрограми для їх аналізу. Мінусами ієрархічних методів є їх чутливість до обраної відстані між кластерами, висока обчислювальна складність, залежність результатів аналізу від впливу аномалій. В алгоритмах ієрархічної кластеризації розрахунки проводяться за формулами, що є складнішими у порівнянні з ітеративними методами [9].

Для кластерного аналізу також застосовуються методи на основі ймовірного розподілу даних. Їх концепція базується на групуванні об'єктів, ознаки яких мають розподіли з подібними параметрами. Відмінною рисою алгоритмів на основі розподілу є принцип об'єднання зразків даних за апостеріорною ймовірністю їх приналежності до певного кластера, а не за встановленою мірою ві-

дстані. Основними ймовірнісними методами є гаусівська модель суміші (GMM) та модель максимізації. Гаусівська модель суміші ґрунтується на припущенні про те, що розглядувані дані згенеровані набором n багатовимірних нормальних розподілів, кожен з яких характеризується математичним сподіванням μ_k та коваріаційною матрицею Σ_k , $k = \overline{1, n}$. У загальному випадку можна розглянути модель суміші багатовимірних розподілів, відмінних від нормального. Модель максимізації складається з кроків вгадування ймовірності приналежності точок датасету до кластера з певним розподілом та покращення підібраних параметрів таким чином, щоб максимізувати ймовірність приналежності об'єкта коректному кластеру [10].

До переваг методів кластеризації на основі розподілу даних можна віднести їх здатність до якісної обробки даних будь-якої форми і структури та існування робасних до викидів модифікацій. Необхідність заздалегідь визначати число кластерів, ризик локальної збіжності методів та помилок через невідповідність даних встановленому розподілу виступають недоліками ймовірнісних методів кластеризації. Крім того, потреба у наявності специфічних знань накладає обмеження на використання методів кластеризації на основі розподілу.

Оберемо клас найбільш підходящих для розв'язання поставленої задачі алгоритмів кластерного аналізу за допомогою методу аналізу ієрархії (MAI) Саті. Визначаємо критерії порівняння методів кластеризації:

- критерій 1 (K1): чутливість до розподілу даних та форми кластерів;
- критерій 2 (K2): рівень обчислювальної складності;
- критерій 3 (K3): стійкість до викидів у датасеті;
- критерій 4 (K4): складність математичної концепції методу.

Обираємо метод розв'язання задачі серед наступних альтернатив:

а) альтернатива 1 (A1): неієрархічні ітеративні методи кластеризації;

1) K1: підходять для даних довільного розподілу та виділення кластерів сферичної форми;

2) K2: характеризуються низькою обчислювальною складністю;

- 3) K3: мають середню стійкість до викидів;
- 4) K4: потребують елементарних, інтуїтивно зрозумілих математичних розрахунків;
- б) альтернатива 2 (A2) – ієрархічні методи кластеризації;
- 1) K1: мають слабку чутливість до форми кластерів і розподілу даних;
 - 2) K2: характеризуються високою обчислювальною складністю;
 - 3) K3: є нестійкими до викидів у датасеті;
 - 4) K4: потребують розрахунків середньої складності;
- в) альтернатива 3 (A3) – методи кластеризації на основі розподілу даних;
- 1) K1: є чутливими до розподілу даних та ефективно формують кластери різної форми;
 - 2) K2: мають обчислювальну складність середнього ступеня;
 - 3) K3: характеризуються середньою стійкістю до викидів;
 - 4) K4: вимагають специфічних, математичних знань в області теорії ймовірностей та статистики.

На рисунку 1.8 наведемо ієрархічну структуру проблеми вибору.

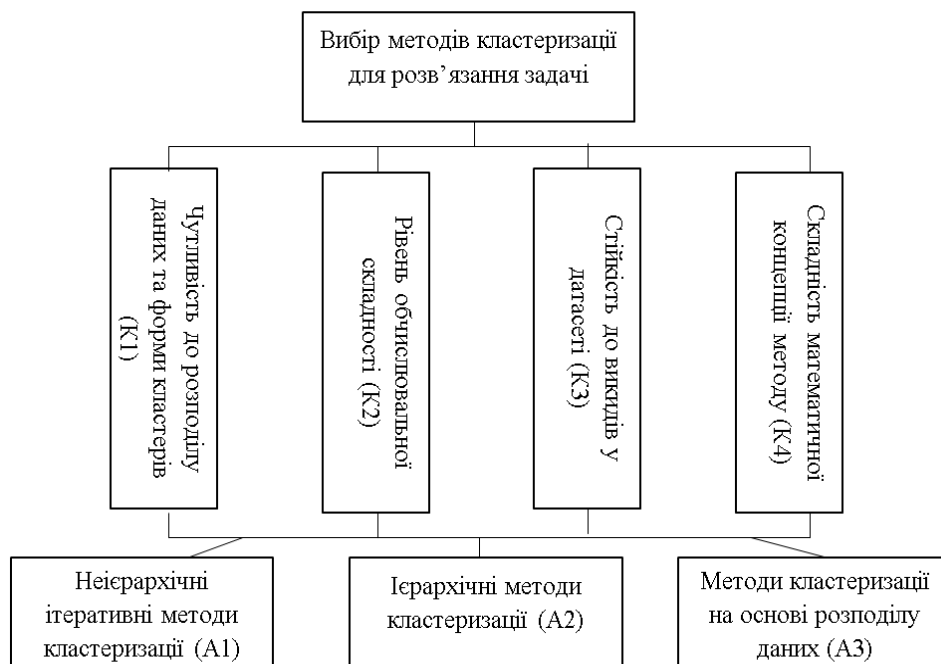


Рисунок 1.8 – Ієрархічна структура проблеми вибору методів кластеризації для розв'язання задачі

Спочатку складаємо матрицю попарних порівнянь важливості критеріїв. Знаємо, що:

- чутливість до розподілу даних та форми кластерів (K1) слабко чи істотно важливіша за рівень обчислювальної складності (K2), однаково чи слабко важливіша, ніж чутливість до шуму в даних (K3), та очевидно важливіша, ніж складність математичної концепції методу (K4);
- обчислювальна складність методів (K2) істотно важливіша, ніж складність математичної концепції методу (K4);
- стійкість до викидів у датасеті (K3) істотно важливіша, ніж обчислювальна складність методів (K2) та істотно чи очевидно важливіша, ніж складність математичної концепції методу (K4).

Перейдемо від якісного попарного порівняння критеріїв до еквівалентного йому чисельного за шкалою Сааті. Отриману матрицю попарних порівнянь наведено у таблиці 1.1.

Таблиця 1.1 – Матриця попарних порівнянь критеріїв

	K1	K2	K3	K4
K1	1	4	2	7
K2	$\frac{1}{4}$	1	$\frac{1}{5}$	6
K3	$\frac{1}{2}$	5	1	6
K4	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	1

Наступним кроком здійснюємо розрахунок компонент вектора локальних пріоритетів критеріїв. Отримані та наведені у таблиці 1.2 значення дозволяють кількісно оцінити вагомість кожного з критеріїв у загальній системі прийняття рішення.

Таблиця 1.2 – Розрахунок вектора локальних пріоритетів критеріїв

	K1	K2	K3	K4	Середнє геометричне за рядками	Вектор локальних Пріоритетів
K1	1	4	2	7	$x_1 = \sqrt[4]{1 \cdot 4 \cdot 2 \cdot 7} = 2,73556$	$p_1^K = \frac{x_1}{\sum_{i=1}^4 x_i} = 0,480$
K2	$\frac{1}{4}$	1	$\frac{1}{5}$	6	$x_2 = \sqrt[4]{\frac{1}{4} \cdot 1 \cdot \frac{1}{5} \cdot 6} = 0,74008$	$p_2^K = \frac{x_2}{\sum_{i=1}^4 x_i} = 0,130$
K3	$\frac{1}{2}$	5	1	6	$x_3 = \sqrt[4]{\frac{1}{2} \cdot 5 \cdot 1 \cdot 6} = 1,96799$	$p_3^K = \frac{x_3}{\sum_{i=1}^4 x_i} = 0,346$
K4	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	1	$x_4 = \sqrt[4]{\frac{1}{7} \cdot \frac{1}{6} \cdot \frac{1}{6} \cdot 1} = 0,25099$	$p_4^K = \frac{x_4}{\sum_{i=1}^4 x_i} = 0,044$
					$\sum_{i=1}^4 x_i = 5,69462$	$\sum_{i=1}^4 p_i^K = 1,000$

Для того, щоб знайти індекс узгодженості CI^K , проводимо низку обчислень. Спершу розраховуємо суми елементів матриці за стовпцями:

$$y_1 = 1 + \frac{1}{4} + \frac{1}{2} + \frac{1}{7} = 1,893, \quad y_2 = 4 + 1 + 5 + \frac{1}{6} = 10,167,$$

$$y_3 = 2 + \frac{1}{5} + 1 + \frac{1}{6} = 3,367, \quad y_4 = 7 + 6 + 6 + 1 = 20,000.$$

Далі знаходимо її найбільше власне значення $\lambda_{\max}^K$:

$$\lambda_{\max}^K \approx \sum_{i=1}^4 y_i p_i^k = 1,893 \cdot 0,480 + 10,167 \cdot 0,130 + 3,367 \cdot 0,346 + 20 \cdot 0,044 = 4,276.$$

Відповідно, індекс узгодженості CI^K та відношення узгодженості CR^K

дорівнюють:

$$CI^K = \frac{\lambda_{\max}^K - m}{m - 1} = \frac{4,276 - 4}{4 - 1} = 0,092,$$

$$CR^K = \frac{CI^K}{RI^K} = \frac{0,092}{0,90} = 0,102.$$

Величина $CR^K < 0,2$, тож вважаємо матрицю парних порівнянь критеріїв побудованою вірно.

Вектор локальних пріоритетів критеріїв відносно проблеми вибору має наступний вигляд:

$$\vec{p}^K = \begin{pmatrix} 0,480 \\ 0,130 \\ 0,346 \\ 0,044 \end{pmatrix}.$$

Далі складаємо матриці попарних порівнянь альтернатив – методів класифікації для розв’язання задачі – за кожним критерієм. Попарні порівняння виконуються для всіх можливих пар методів за шкалою відносної важливості Сааті, що дозволяє визначити ступінь переваги кожного методу за окремими характеристиками (таблиці 1.3–1.6).

Обчислюємо суми елементів матриці, наведеної у таблиці 1.3, за стовпцями:

$$y_1 = 1 + 4 + 2 = 7,000, \quad y_2 = \frac{1}{4} + 1 + \frac{1}{3} = 1,583 \quad y_3 = \frac{1}{2} + 3 + 1 = 4,500.$$

Відповідно, максимальне власне значення побудованої матриці $\lambda_{\max K1}^A$ дорівнює:

$$\lambda_{\max K1}^A \approx \sum_{i=1}^3 y_i p_i^A = 7,000 \cdot 0,138 + 1,538 \cdot 0,623 + 4,500 \cdot 0,239 = 3,018.$$

Таблиця 1.3 – Матриця попарних порівнянь альтернатив за критерієм K1
та розрахунок вектора локальних пріоритетів

	A1	A2	A3	Середнє геометричне за рядками	Вектор локальних пріоритетів
A1	1	$\frac{1}{4}$	$\frac{1}{2}$	$x_1 = \sqrt[3]{1 \cdot \frac{1}{4} \cdot \frac{1}{2}} = 0,50000$	$p_1^A = \frac{x_1}{\sum_{i=1}^3 x_i} = 0,138$
A2	4	1	3	$x_2 = \sqrt[3]{4 \cdot 1 \cdot 3} = 2,28943$	$p_2^A = \frac{x_2}{\sum_{i=1}^3 x_i} = 0,623$
A3	2	$\frac{1}{3}$	1	$x_3 = \sqrt[3]{2 \cdot \frac{1}{3} \cdot 1} = 0,87358$	$p_3^A = \frac{x_3}{\sum_{i=1}^3 x_i} = 0,239$
				$\sum_{i=1}^3 x_i = 3,66301$	$\sum_{i=1}^3 p_i^A = 1,000$

Матриця попарних порівнянь альтернатив за критерієм K1 має розмірність 3×3 , отже, випадковий індекс дорівнює $RI^A = 0,58$. З урахуванням цього отримаємо наступні значення індексу узгодженості CI_{K1}^A та відношення узгодженості CR_{K1}^A :

$$CI_{K1}^A = \frac{\lambda_{\max K1}^A - m}{m - 1} = \frac{3,018 - 3}{3 - 1} = 0,009,$$

$$CR_{K1}^A = \frac{CI_{K1}^A}{RI^A} = \frac{0,009}{0,58} = 0,016.$$

Отримана величина $CR_{K1}^A < 0,2$, отже матриця попарних порівнянь альтер-

рнатив за критерієм K1 побудована коректно, та думки дослідника гарно узгоджені.

Таблиця 1.4 – Матриця попарних порівнянь альтернатив за критерієм K2
та розрахунок вектора локальних пріоритетів

	A1	A2	A3	Середнє геометричне за рядками	Вектор локальних пріоритетів
A1	1	8	5	$x_1 = \sqrt[3]{1 \cdot 8 \cdot 5} = 3,41995$	$p_1^A = \frac{x_1}{\sum_{i=1}^3 x_i} = 0,733$
A2	$\frac{1}{8}$	1	$\frac{1}{4}$	$x_2 = \sqrt[3]{\frac{1}{8} \cdot 1 \cdot \frac{1}{4}} = 0,31498$	$p_2^A = \frac{x_2}{\sum_{i=1}^3 x_i} = 0,068$
A3	$\frac{1}{5}$	4	1	$x_3 = \sqrt[3]{\frac{1}{5} \cdot 4 \cdot 1} = 0,92832$	$p_3^A = \frac{x_3}{\sum_{i=1}^3 x_i} = 0,199$
				$\sum_{i=1}^3 x_i = 4,66325$	$\sum_{i=1}^3 p_i^A = 1,000$

Обчислюємо суми елементів матриці за стовпцями та її найбільше власне значення $\lambda_{\max K2}^A$:

$$y_1 = 1 + \frac{1}{8} + \frac{1}{5} = 1,325, \quad y_2 = 8 + 1 + 4 = 13,000, \quad y_3 = 5 + \frac{1}{4} + 1 = 6,250,$$

$$\lambda_{\max K2}^A \approx \sum_{i=1}^3 y_i p_i^A = 1,325 \cdot 0,733 + 13,000 \cdot 0,068 + 6,250 \cdot 0,199 = 3,099.$$

Матриця попарних порівнянь альтернатив за критерієм K2 має розмірність 3×3 , тому беремо випадковий індекс $RI^A = 0,58$ та отримуємо наступні величини індексу узгодженості CI_{K2}^A та відношення узгодженості CR_{K2}^A :

$$CI_{K2}^A = \frac{\lambda_{\max K2}^A - m}{m - 1} = \frac{3,099 - 3}{3 - 1} = 0,0495,$$

$$CR_{K2}^A = \frac{CI_{K2}^A}{RI^A} = \frac{0,0495}{0,58} = 0,085.$$

Отримана величина $CR_{K2}^A < 0,2$, отже думки дослідника гарно узгоджені.

Таблиця 1.5 – Матриця попарних порівнянь альтернатив за критерієм К3
та розрахунок вектора локальних пріоритетів

	A1	A2	A3	Середнє геометричне за рядками	Вектор локальних пріоритетів
A1	1	7	3	$x_1 = \sqrt[3]{1 \cdot 7 \cdot 3} = 2,75892$	$p_1^A = \frac{x_1}{\sum_{i=1}^3 x_i} = 0,659$
A2	$\frac{1}{7}$	1	$\frac{1}{4}$	$x_2 = \sqrt[3]{\frac{1}{7} \cdot 1 \cdot \frac{1}{4}} = 0,32932$	$p_2^A = \frac{x_2}{\sum_{i=1}^3 x_i} = 0,079$
A3	$\frac{1}{3}$	4	1	$x_3 = \sqrt[3]{\frac{1}{3} \cdot 4 \cdot 1} = 1,10064$	$p_3^A = \frac{x_3}{\sum_{i=1}^3 x_i} = 0,262$
				$\sum_{i=1}^3 x_i = 4,18888$	$\sum_{i=1}^3 p_i^A = 1,000$

Порахуємо суми елементів матриці за стовпцями та її найбільше власне значення $\lambda_{\max K3}^A$:

$$y_1 = 1 + \frac{1}{7} + \frac{1}{3} = 1,476, \quad y_2 = 7 + 1 + 4 = 12,000, \quad y_3 = 3 + \frac{1}{4} + 1 = 4,250,$$

$$\lambda_{\max K3}^A \approx \sum_{i=1}^3 y_i p_i^A = 1,476 \cdot 0,659 + 12,000 \cdot 0,079 + 4,250 \cdot 0,262 = 3,038.$$

Матриця попарних порівнянь альтернатив за критерієм К3 має розмірність 3×3 , тож для розрахунку обирається випадковий індекс $RI^A = 0,58$. Індекс узгодженості CI_{K3}^A та відношення узгодженості CR_{K3}^A дорівнюють

$$CI_{K3}^A = \frac{\lambda_{\max K3}^A - m}{m - 1} = \frac{3,038 - 3}{3 - 1} = 0,019,$$

$$CR_{K3}^A = \frac{CI_{K3}^A}{RI^A} = \frac{0,019}{0,58} = 0,033.$$

Отримана величина $CR_{K3}^A < 0,2$, отже думки дослідника є гарно узгодженими між собою.

Таблиця 1.6 – Матриця попарних порівнянь альтернатив за критерієм К4 та розрахунок вектора локальних пріоритетів

	A1	A2	A3	Середнє геометричне за рядками	Вектор локальних пріоритетів
A1	1	6	8	$x_1 = \sqrt[3]{1 \cdot 6 \cdot 8} = 3,63424$	$p_1^A = \frac{x_1}{\sum_{i=1}^3 x_i} = 0,747$
A2	$\frac{1}{6}$	1	5	$x_2 = \sqrt[3]{\frac{1}{6} \cdot 1 \cdot 5} = 0,94104$	$p_2^A = \frac{x_2}{\sum_{i=1}^3 x_i} = 0,193$
A3	$\frac{1}{8}$	$\frac{1}{5}$	1	$x_3 = \sqrt[3]{\frac{1}{8} \cdot \frac{1}{5} \cdot 1} = 0,29240$	$p_3^A = \frac{x_3}{\sum_{i=1}^3 x_i} = 0,060$
				$\sum_{i=1}^3 x_i = 4,86768$	$\sum_{i=1}^3 p_i^A = 1,000$

Порахуємо суми елементів матриці за стовпцями та її найбільше власне значення $\lambda_{\max K4}^A$:

$$y_1 = 1 + \frac{1}{6} + \frac{1}{8} = 1,292, \quad y_2 = 6 + 1 + \frac{1}{5} = 7,200, \quad y_3 = 8 + 5 + 1 = 14,000,$$

$$\lambda_{\max K4}^A \approx \sum_{i=1}^3 y_i p_i^A = 1,292 \cdot 0,747 + 7,200 \cdot 0,193 + 14,000 \cdot 0,060 = 3,195.$$

Оскільки матриця попарних порівнянь альтернатив за критерієм K4 має розмірність 3×3 , то випадковий індекс $RI^A = 0,58$. Отже, отримаємо наступні індекс узгодженості CI_{K4}^A і відношення узгодженості CR_{K4}^A :

$$CI_{K4}^A = \frac{\lambda_{\max K4}^A - m}{m - 1} = \frac{3,195 - 3}{3 - 1} = 0,0975,$$

$$CR_{K4}^A = \frac{CI_{K4}^A}{RI^A} = \frac{0,0975}{0,58} = 0,168.$$

Отримана величина $CR_{K4}^A < 0,2$, тож можна вважати думки дослідника гарно узгодженими.

Тепер будуюмо матрицю, стовпцями якої є обчислені вектори локальних пріоритетів альтернатив за кожним критерієм, а також визначаємо вектор глобальних пріоритетів:

$$P^A = \begin{pmatrix} 0,138 & 0,733 & 0,659 & 0,747 \\ 0,623 & 0,068 & 0,079 & 0,193 \\ 0,239 & 0,199 & 0,262 & 0,060 \end{pmatrix},$$

$$\vec{p} = P^A \cdot \vec{p}^K = \begin{pmatrix} 0,138 & 0,722 & 0,644 & 0,747 \\ 0,623 & 0,073 & 0,085 & 0,193 \\ 0,239 & 0,205 & 0,271 & 0,062 \end{pmatrix} \cdot \begin{pmatrix} 0,480 \\ 0,130 \\ 0,346 \\ 0,044 \end{pmatrix} = \begin{pmatrix} 0,422 \\ 0,344 \\ 0,234 \end{pmatrix}.$$

Індекс та відношення узгодженості для всієї ієрархії дорівнюють

$$CI = CI^K + (\bar{p}^K, CI^A) = 0,092 + 0,480 \cdot 0,009 + \\ + 0,130 \cdot 0,0495 + 0,346 \cdot 0,019 + 0,044 \cdot 0,0975 = 0,134,$$

$$RI = RI^K + RI^A = 0,90 + 0,58 = 1,48,$$

$$CR = \frac{CI}{RI} = \frac{0,134}{1,48} = 0,091.$$

Величина $CR < 0,2$, отже думки дослідника добре узгоджені.

Засновуючись на компонентах отриманого вектора глобальних пріоритетів, обираємо для розв'язання задачі інтелектуального аналізу балів ЗНО та НМТ за 2021 – 2024 роки у м. Харків першу альтернативу – неієрархічні ітеративні методи кластеризації.

1.3 Змістовна та формальна постановка задачі

1.3.1 Змістовна постановка задачі

У розв'язуванні задачі початковими даними є загальна інформація за 2021 – 2024 роки про випускників ЗСО м. Харків та їх навчальні досягнення. Вихідний датасет представлено у форматі таблиці, в якій наводяться відомості про тип та назву навчального закладу, місце та дату написання тесту, дисципліни для складання та бали ЗНО та НМТ, отримані за кожним з обраних предметів тощо. Завдання полягає в проведенні інтелектуального аналізу показників ЗНО та НМТ 2021 – 2024 з метою вивчення типових тенденцій у даних та дослідження динаміки їх змін з урахуванням зовнішніх кризових обставин (пандемія та війна). Для того, щоб розв'язати поставлену задачу, необхідно обчислити інтегральні показники на базі первинних атрибутів, виконати кластеризацію ЗСО за розрахованими агрегованими ознаками, порівняти за роками властивості сформованих кластерів та перевірити статистичну значущість зроблених висновків. Отримані результати можуть бути корисними під час розробки практи-

чно цінних рекомендацій стосовно покращення якості навчальних програм у закладах середньої освіти в кризових умовах.

1.3.2 Формальна постановка задачі

Групування ЗСО м. Харків за подібним рівнем навчальної підготовки моделюється як задача кластерного аналізу.

Досліджуваний набір даних подається як множина X з n об'єктів, кожен з яких має p ознак: $\vec{x}_i = (x_i^{(1)}, \dots, x_i^{(p)})$, $i = \overline{1, n}$, $\vec{x}_i \in X$. Розділяємо його на дві вибірки: навчальну $X_m = \{\vec{x}_i\}_{i=1}^m \subset X$ для тренування обчислювального алгоритму та тестову $X_k = \{\vec{x}_i\}_{i=1}^k \subset X$, $X_k \cap X_m = \emptyset$, для перевірки якості його роботи на нових даних.

Для виконання кластеризації слід обрати метрику $d(\vec{x}_i, \vec{x}_j)$, $i \neq j$, $i, j = \overline{1, n}$, – функцію для розрахунку відстані між об'єктами датасету $\vec{x}_i$ та $\vec{x}_j$, $\vec{x}_i, \vec{x}_j \in X$. Вибір метрики дуже сильно впливає на структуру та склад кластерів: елементи однієї множини за принципом відстані можуть поділитися на групи з абсолютно різними спільними властивостями. Далі будемо використовувати евклідову відстань між точками $\vec{x}_i$, $\vec{x}_j$, задану виразом:

$$d(\vec{x}_i, \vec{x}_j) = \sqrt{\sum_{k=1}^p (x_i^{(k)} - x_j^{(k)})^2}. \quad (1.1)$$

У процесі кластерного аналізу масив даних X поділяється на кластери – групи об'єктів, схожих між собою за обраною метрикою (1.1). Схожими вважаються елементи датасету $\vec{x}_i \in X$, $i = \overline{1, n}$, що знаходяться близько по відношенню один до одного та до центроїда кластера. Тобто, під час формування

кластерів об'єкти групуються якомога компактніше, прямуючи до мінімальної внутрішньокластерної відстані SSE :

$$SSE = \sum_{i=1}^n \sum_{j=1}^k \omega_{ij} d(\vec{x}_i, \vec{\mu}_j) \rightarrow \min, \quad (1.2)$$

де $\omega_{ij} = 1$, якщо $\vec{x}_i$ належить j -му кластеру з центроїдом $\vec{\mu}_j$, та $\omega_{ij} = 0$ – у протилежному випадку.

Таким чином, у ході кластеризації кожній точці $\vec{x}_i$ досліджуваного набору даних X призначається мітка кластера y_j з множини Y , утворюючи відображення $f : X \rightarrow Y$.

Кількість кластерів k , на які буде розбито набір даних, слід визначити попередньо за допомогою спеціальних методів, наприклад, методу ліктя.

Зазначимо, що кластеризація – це задача навчання без вчителя, тому мітки кластерів y_j для об'єктів $\vec{x}_i$ навчальної вибірки X_m не відомі заздалегідь. Отже, обчислювальний алгоритм визначає їх самостійно як для навчальних, так і для тестових даних [11].

1.4 Постановка задач дослідження

За результатами виконаного системного аналізу задачі інтелектуального аналізу балів ЗНО та НМТ за 2021 – 2024 роки у м. Харків для розв'язання поставленої задачі було обрано неієрархічні методи кластеризації, які виявилися оптимальними за низкою важливих для дослідника критеріїв.

Метою написання кваліфікаційної роботи є аналіз показників успішності складання ЗНО та НМТ випускниками навчальних закладів м. Харків у 2021 – 2024 роках для дослідження якості освітньої підготовки з математики та виявлення тенденцій у її змінах під впливом зовнішніх чинників.

Для досягнення поставленої мети необхідно виконати наступні завдання:

- провести системний аналіз задачі «інтелектуальний аналіз балів зовнішнього незалежного оцінювання і національного мультипредметного тесту за 2021 – 2024 роки у м. Харків» та за його результатами обрати метод розв’язання поставленої задачі;
- ознайомитися з теоретичними концепціями методів інтелектуального аналізу, що будуть застосованими для дослідження вихідних даних;
- розробити програму для розв’язання поставленої задачі у системі комп’ютерної алгебри Wolfram Mathematica 10.4;
- виконати дослідження даних обраними методами та оцінити якість кластеризації;
- провести аналіз отриманих результатів та сформулювати прикладні висновки щодо якості освітньої підготовки у розглядуваний період часу.

2 ВИБІР ТА ОБҐРУНТУВАННЯ МЕТОДУ РОЗВ'ЯЗАННЯ

2.1 Попередня обробка даних

Перед проведенням аналізу дані необхідно підготувати за допомогою спеціальних методів. Перелік обов'язкових процедур попередньої обробки даних може варіюватися в залежності від їх початкового стану й умов розв'язуваного завдання. У загальному випадку передобробка датасету включає в себе його перевірку й очистку, трансформацію та оптимізацію.

На першому кроці датасет досліджується на наявність помилок, дублікатів, пропусків та аномалій. У разі їх виявлення «брудні» дані очищаються: видаляються повторювані та хибні записи, рядки з порожніми значеннями вилучаються або заповнюються за певною стратегією. Викиди в залежності від їх походження прибираються або виділяються в окрему групу для вивчення. Чисті дані проходять процедури трансформації, зокрема, перейменування, зміни типу та групування ознак, а також масштабування для приведення до єдиної шкали вимірювання. Оптимізація набору даних включає в себе зниження його вимірності із застосуванням відбору ознак або їх проєктування у простір меншої вимірності.

Від ефективності попередньої обробки даних залежать адекватність результатів подальшого аналізу та здатність побудованих алгоритмів до роботи без збоїв. Важливість етапу первинної підготовки датасету зумовлює великі часові затрати на нього – до 80% від загального часу аналізу [12, 13].

Функція відстані $d(\vec{x}_i, \vec{x}_j)$ між об'єктами датасету, за якою визначається їх подібність у ході кластеризації, чутлива до одиниць вимірювання ознак. Отже, перед виконанням кластеризації необхідно за допомогою стандартизації привести атрибути даних до єдиного масштабу. Розглянемо детальніше процедуру стандартизації ознак об'єктів датасету.

У наборі даних маємо n об'єктів $\vec{x}_i$, $i = \overline{1, n}$, кожен з яких є вектором з p ознаками:

$$\vec{x}_i = (x_i^{(1)}, \dots, x_i^{(p)}). \quad (2.1)$$

Вираз для стандартизованого значення атрибута $\tilde{x}_i^{(j)}$, $j = \overline{1, p}$, яким описується спостереження $\vec{x}_i$ (2.1), дорівнює:

$$\tilde{x}_i^{(j)} = \frac{x_i^{(j)} - \bar{x}^{(j)}}{\hat{\sigma}_{x^{(j)}}}, \quad (2.2)$$

де $\bar{x}^{(j)} = \frac{1}{n} \sum_{i=1}^n x_i^{(j)}$ – вибіркове середнє значення для j -ї ознаки спостереження (2.1) у датасеті;

$\hat{\sigma}_{x^{(j)}} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i^{(j)} - \bar{x}^{(j)})^2}$ – вибіркове стандартне відхилення j -ї ознаки спостереження (2.1) у датасеті [14].

Після масштабування проводиться виділення ознак з метою зменшення вимірності даних. Перехід до простору ознак нижчої вимірності здійснюватимемо за допомогою методу головних компонент (РСА). Наведемо його кроки.

На першому етапі обчислюється вибіркова коваріаційна матриця $\Sigma = [\hat{\sigma}_{jl}]_{p \times p}$ за формулою:

$$\hat{\sigma}_{jl} = \frac{1}{n} \sum_{i=1}^n \tilde{x}_i^{(j)} \tilde{x}_i^{(l)}, \quad j, l = \overline{1, p}, \quad (2.2)$$

де $\tilde{x}_i^{(j)}$, $\tilde{x}_i^{(l)}$ – стандартизовані j -й та l -й атрибути об'єкта (2.1).

Коваріаційна матриця зі значеннями (2.2) містить повну характеристику структури датасету. Її діагональні елементи характеризують розкид кожної ознаки даних та її інформативність, позадіагональні елементи – наявність та силу зв'язку між парами ознак [15].

Наступним є етап розрахунку головних компонент, які представляють собою нормовані лінійні комбінації стандартизованих ознак об'єктів набору даних у вихідному просторі. Вони визначаються за формулою:

$$y^{(t)} = \sum_{j=1}^p \alpha_{tj} \tilde{x}_i^{(j)}, \quad t = \overline{1, p}. \quad (2.3)$$

Згідно з умовою нормування головних компонент $\sum_{j=1}^s \alpha_{hj}^2 = 1$. Головні компоненти $y^{(t)}$ та $y^{(s)}$, $t, s = \overline{1, p}$, є некорельованими між собою: $\text{cov}(y^{(t)}, y^{(s)}) = 0$ при $s < t$.

Вагові коефіцієнти α_{tj} головних компонент (2.3) являють собою елементи t -го нормованого власного вектора коваріаційної матриці зі значеннями (2.2):

$$\vec{A}_t = (\alpha_{t1}, \dots, \alpha_{tp})^T, \quad t = \overline{1, p}. \quad (2.4)$$

Власним векторам (2.4) відповідають власні значення λ_t , яким дорівнюють дисперсії побудованих головних компонент:

$$\lambda_t = Dy^{(t)}, \quad t = \overline{1, p}.$$

Кожна головна компонента є новою змінною, яка максимізує дисперсію даних у напрямку, ортогональному до попередніх компонент. Головні компоненти утворюють спадну послідовність за величиною дисперсії:

$$Dy^{(1)} \geq Dy^{(2)} \geq \dots \geq Dy^{(p)},$$

тобто перша компонента пояснює найбільшу частку дисперсії, а далі кожна на-

ступна зберігає все менше вихідної інформативності ознак даних. У залежності від умов задачі лишають різну кількість головних компонент (2.3) для подальшого аналізу. Часто для вибору кількості головних компонент використовують критерій «кам'янистого осипу»: точка різкого зламу на графіку власних значень вказує на доцільну кількість компонент. Також, для того, щоб оцінити, наскільки структура даних у новому базисі відповідає початковій, та вирішити, скільки з компонент залишити, використовують показник:

$$q(r) = \frac{\sum_{j=1}^r Dy^{(j)}}{\sum_{j=1}^p Dy^{(j)}}. \quad (2.5)$$

Величина (2.5) характеризує частку вихідної варіабельності датасету, яка лишається після переходу до нового простору ознак [16].

2.2 Неієрархічні ітеративні методи кластерного аналізу даних

Серед алгоритмів кластеризації неієрархічні методи вирізняються чисельністю та широким застосуванням в аналізі даних. Їхній підхід, як згадувалося вище, зводиться до поділу набору даних на групи об'єктів, подібних між собою за деякими характеристиками. Кількість кластерів k задається наперед як вхідний параметр. Схожість об'єктів визначається відстанню між ними в просторі ознак [17]. Для обчислень застосовуються різні метрики. Найпопулярнішими, крім евклідової, є мангеттенська метрика:

$$d(\vec{x}_i, \vec{x}_j) = \sum_{k=1}^p |x_i^{(k)} - x_j^{(k)}| \quad (2.6)$$

та метрика Чебишова:

$$d(\vec{x}_i, \vec{x}_j) = \max_{k=1, p} |x_i^{(k)} - x_j^{(k)}|.$$

Алгоритми неієрархічної кластеризації відрізняються за різними критеріями. Одним з них виступає підхід до обчислення центроїдів кластерів. Зазначена відмінність яскраво проявляється у двох методах: k -means та k -medoids. Саме їх буде використано для розв'язання поставленої задачі, тож переходимо до їх вивчення [18].

2.2.1 Метод k -means

Метод k -means є базовим неієрархічним методом кластеризації, адже його принципи покладені в основу всіх інших алгоритмів цього класу. Він відносить кожен зразок даних до кластера з найближчим до нього центроїдом. Іншими словами, кластери формуються за принципом мінімізації внутрішньокластерної дисперсії (1.2).

Центроїдом (центром ваги) кластера називають точку $\vec{\mu} = (\mu_1, \dots, \mu_p)$, у якої кожна координата μ_j , $j = \overline{1, p}$, дорівнює усередненому за всіма об'єктами кластера значенню j -го атрибута [19].

Алгоритм методу k -means складається з таких етапів.

Етап 1. Встановити кількість майбутніх кластерів k , застосовуючи метод ліктя чи силуетний аналіз.

Етап 2. Задати стартові значення центроїдів $\vec{\mu}^{(l)}$, $l = \overline{1, k}$, випадковим чином обравши k об'єктів $\vec{x}_i$ з розглядуваного набору даних.

Етап 3. Розбити датасет на кластери шляхом віднесення кожного спостереження $\vec{x}_i$, $\vec{x}_i \neq \vec{\mu}^{(l)}$, $l = \overline{1, k}$, до кластера з найближчим центроїдом.

Етап 4. Розрахувати координати центроїдів у створених кластерах за формулою:

$$\vec{\mu}^{(l)} = \frac{1}{|C_l|} \sum_{\vec{x}_i \in C_l} \vec{x}_i, \quad l = \overline{1, k},$$

де $|C_l|$ – кількість об'єктів в кластері під номером l .

Етап 5. Виконувати 3 та 4 етапи доти, доки не виконуватиметься критерій зупинки. Правилom завершення кластеризації може виступати встановлена кількість ітерацій процесу.

Алгоритм k -means виділяється з-поміж інших неієрархічних методів легкістю реалізації розрахунків та інтерпретації отриманих результатів. Також перевагою є обчислювальна швидкість. Водночас, у цього методу існує декілька значних недоліків: чутливість до початкової ініціалізації центроїдів та аномалій, необхідність завдання числа кластерів k заздалегідь та залежність якості кластеризації від геометричної форми даних у просторі ознак [20, 21].

2.2.2 Метод k -medoids

Згідно з алгоритмом k -medoids центроїдами кластерів виступають медоїди – об'єкти кластерів, які знаходяться на мінімальній відстані від усіх інших об'єктів свого кластера. Досліджуваний метод є робастним до аномальних значень, на відміну від метода k -means. Через це у ньому мірою внутрішньоклас-терної дисперсії WSS обрано мангеттенську метрику (2.6). У її формулі немає операції піднесення відстані між спостереженнями до другого ступеня, що підвищує стійкість до викидів [22].

Алгоритм методу k -medoids складається з таких етапів.

Етап 1. Випадковим чином обрати у масиві даних k об'єктів $\vec{x}_i$ як старто-

ві медоїди кластерів $\vec{m}_l, l = \overline{1, k}$.

Етап 2. Додати кожен зразок датасету $\vec{x}_i$ до кластера, медоїд якого розташовано на мінімальній відстані від нього.

Етап 3. У кожному кластері перебрати всі об'єкти та обрати новим медоїдом точку $\vec{x}_i$, яка знаходиться найближче до всіх інших точок кластера в сенсі обраної метрики:

$$\vec{x}_i : \sum_j d(\vec{x}_i, \vec{x}_j) \rightarrow \min, i = \overline{1, n}, i \neq j.$$

Медоїд обов'язково є елементом множини вихідних даних.

Етап 4. Повторювати етапи 2 та 3 до виконання заданого критерію зупинки.

Однією з переваг методу k -medoids є його висока ефективність обробки даних з викидами та, відповідно, асиметричними розподілами. До позитивних властивостей методу також відносяться низька чутливість до масштабу ознак, простота математичної концепції та здатність якісно виділяти кластери різної форми, не тільки сферичної. Разом з цим, метод зберігає мінуси алгоритму k -means, зокрема, необхідність визначати число кластерів k та ризик отримання неоптимального складу кластерів через потрапляння алгоритму до локального мінімуму. Також метод k -medoids має повільнішу за k -means збіжність [23].

2.3 Внутрішні міри оцінки валідності кластерів

Оскільки кластеризація відноситься до задач аналізу даних без учителя, то її результати потребують валідації виключно на основі структури даних. Від обґрунтованості поділу ЗСО м. Харків на кластери залежить достовірність висновків про рівень середньої освіти. З цієї причини важливим є оцінювання якості отриманого розбиття даних за допомогою спеціальних кількісних метрик. Перед тим, як розглянути внутрішні міри валідності кластерів, введемо ба-

зові поняття компактності та роздільності кластерів.

Компактність кластерів, або внутрішньокластерна дисперсія, визначає подібність об'єктів за критерієм відстані між ними: чим ближче один до одного розташовані точки даних у кластерах, тим якіснішими вони вважаються. Таким чином, у ході кластерного аналізу розв'язується задача мінімізації внутрішньокластерної дисперсії (1.2) [24].

Роздільність кластерів, або міжкластерна дисперсія, показує наскільки вони віддалені один від одного. За цим показником кластеризацію можна назвати ефективною, якщо сформовані групи об'єктів не перекриваються і мають чіткі межі. Міжкластерна роздільність має бути максимальною. Вона задається виразом:

$$BSS = \sum_{l=1}^k |C_l| d(\vec{\mu}_l, \vec{\mu}) \rightarrow \max, \quad (2.7)$$

де $d(\vec{\mu}_l, \vec{\mu})$ – обрана метрика відстані між центроїдом l -го кластера та середнім значенням за всіма точками аналізованого датасету $\vec{\mu}$;

$|C_l|$ – обсяг l -го кластера [25].

Переходимо безпосередньо до огляду деяких індексів, що використовують для оцінювання якості сформованих кластерів.

Силуетний коефіцієнт – одна з найпопулярніших внутрішніх мір валідності кластерів. Він показує, наскільки добре кожна точка даних відповідає своєму кластеру. Ідея розрахунку цієї метрики проста: чим ближче об'єкт до спостережень свого кластера і чим далі він від точок інших кластерів, тим якіснішою є кластеризація. Для знаходження силуету кожної точки розраховуються дві величини:

$$a(\tilde{x}_i) = \frac{1}{|C_l|} \sum_{\tilde{x}_i, \tilde{x}_j \in C_l} d(\tilde{x}_i, \tilde{x}_j), \quad i \neq j, \quad i, j = \overline{1, C_l}, \quad (2.8)$$

– середня відстань від об'єкта $\tilde{x}_i$ до всіх інших точок свого l -го кластера;

$$b(\tilde{x}_i) = \min_{\substack{h \in \overline{1, k}, \\ h \neq l}} \frac{1}{|C_h|} \sum_{\substack{\tilde{x}_j \in C_l, \\ \tilde{x}_j \in C_h}} d(\tilde{x}_i, \tilde{x}_j), \quad l = \overline{1, k}, \quad (2.9)$$

– середня відстань від об'єкта $\tilde{x}_i, \tilde{x}_i \in C_l$, до точок найближчого сусіднього кластера C_h .

Вираз (2.8) характеризує згуртованість кластера: його менші значення вказують на ефективну кластеризацію. Значення (2.9) є показником роздільності кластерів: більша його величина свідчить про чіткість та якість сформованих кластерів [26].

Отже, силуетний коефіцієнт s_i для точки $\tilde{x}_i$ обчислюється за формулою:

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}, \quad i = \overline{1, n}. \quad (2.10)$$

Значення показника (2.10) належить діапазону $[-1; 1]$. Близьке до одиниці значення s_i свідчить про гарну відокремленість кластерів та приналежність об'єкта набору даних до кластера, який дійсно найкраще відповідає його властивостям. Якщо силует s_i точки датасету близький до 0, то це означає, що об'єкт знаходиться на межі двох кластерів, і коректність його віднесення до свого кластера є спірною. Від'ємні значення s_i свідчать про перекриття кластерів та вказують або на неякісну кластеризацію, або на особливості структури даних, зокрема, на загальну подібність всіх об'єктів, через що їх важко розрізнити між собою [27].

Для загальної оцінки валідності окремого кластера для нього розраховується середній силуетний коефіцієнт – середнє арифметичне по всіх точках цього кластера.

Індекс Данна обчислюється за формулою:

$$D(C_l) = \frac{\min_{h=1,k;h \neq l} \delta(C_l, C_h)}{\max_{l=1,k} \Delta(C_l)}, \quad (2.11)$$

де $\delta(C_l, C_h) = \min_{\substack{\tilde{x}_i \in C_l, \tilde{x}_j \in C_h, \\ h \neq l}} d(\tilde{x}_i, \tilde{x}_j)$ – відстань між двома найближчими точками

$\tilde{x}_i, \tilde{x}_j$, що відносяться до різних кластерів з мітками h та l ;

$\Delta(C_l) = \max_{\tilde{x}_i, \tilde{x}_j \in C_l} d(\tilde{x}_i, \tilde{x}_j)$ – діаметр l -го кластера (максимальна відстань між

об'єктами даних $\tilde{x}_i$ та $\tilde{x}_j$, що належать l -му кластеру).

За величиною $\delta(C_l, C_h)$ визначається ступінь роздільності кластерів, а діаметр кластера $\Delta(C_l)$ служить мірою скупченості кластера. Аналізуючи співвідношення для індексу Данна, можемо дійти висновку про те, що він є чутливим до викидів у датасеті. Значний вплив аномалій на значення цієї метрики пояснюється тим, що для її розрахунку беруться екстремальні, а не усереднені відстані між об'єктами однієї чи декількох множин об'єктів. Якщо до кластера віднесено викид, то за його рахунок індекс Данна може значно зменшуватися через зростання діаметра кластера. Додатково, у разі близького розташування до об'єкта іншого кластера, аномалія може сильно зменшувати чисельник формули (2.11), незважаючи на положення основної маси об'єктів групи [28].

Індекс Данна набуває будь-яких додатних значень. Його високі значення вказують на чітку розмежованість кластерів між собою та компактну конфігурацію об'єктів середині кожного з них. Близькість до нуля сигналізує про неприйнятно низьку якість кластеризації.

Ще однією внутрішньою метрикою валідності кластерів є індекс Калінскі-Харабаса. Він дорівнює відношенню міжкластерної та внутрішньокластерної дисперсій:

$$CH(C_l) = \frac{BSS}{WSS} \cdot \frac{n-k}{k-1}, \quad (2.12)$$

де WSS – внутрішньокластерна дисперсія, розрахована за формулою (1.2);

BSS – міжкластерна дисперсія, розрахована за формулою (2.7);

k – кількість кластерів;

n – обсяг датасету.

Множник $\frac{n-k}{k-1}$ у виразі (2.12) використовується для нормування метрики

Калінскі – Харабаса на ступені свободи дисперсій, аби підвищити якість кластеризації при різних значеннях розміру масиву даних n та кількості кластерів k .

Як і індекс Данна, міра Калінскі – Харабаса може приймати довільні невід’ємні значення. Чим вищим є значення індексу Калінскі – Харабаса, тим краще кластеризовано дані. Велике значення показника $CH(C_l)$ (може сягати сотен та тисяч) вказує на те, що утворені кластери чітко виражені та віддалені один від одного, і при цьому вони досить щільні, складаються з об’єктів, дуже подібних у сенсі обраної метрики відстані. На відміну від міри Данна, він враховує розкиди в датасеті та в кожній групі окремо для всіх точок, що робить його більш показовим для оцінки якості кластеризації наборів даних різної структури [29].

2.4 Застосування методів інтелектуального аналізу даних до розв’язання задачі оцінювання навчальних досягнень випускників ЗСО м. Харків за 2021 – 2024 роки

Поставлена задача розв’язується з метою вивчення якості освітньої підготовки з математики у ЗСО м. Харків протягом 2021 – 2024 років на основі результатів випускних випробувань. Дослідження даних з балами ЗНО та НМТ, які було отримано випускниками, є комплексною процедурою, що складається

з п'яти основних етапів:

- отримання необхідних наборів з архівів і їх попередня обробка для підготовки до подальшого аналізу;
- кластеризація трансформованих датасетів за допомогою методів k -means та k -medoids;
- вивчення характеристик сформованих кластерів, трендів і закономірностей, що спостерігаються у них у кожному році окремо, і генерація висновків щодо якості академічної підготовки з математики у ЗСО м. Харків на базі отриманих результатів;
- аналіз змін у складі та структурі кластерів ЗСО м. Харків у розрізі часу та тенденцій розвитку освітньої інфраструктури міста протягом 2021 – 2024 років;
- оцінювання якості утворених кластерів з використанням внутрішніх мір їх валідності.

Нижче опишемо застосування методів інтелектуального аналізу даних для розв'язання поставленої задачі.

2.4.1 Збір статистичних даних та їх передобробка

Розв'язання задачі починається зі збору вихідних даних. Інформацію щодо результатів складання ЗНО та НМТ з математики випускниками ЗСО м. Харків за 2021 – 2024 роки можна отримати в онлайн-архівах УЦОЯО. Після експорту датасети в першу чергу піддаються передобробці.

Спочатку набори даних з індивідуальними результатами НМТ агрегуються за атрибутами назви та типу навчального закладу. Ця операція проводиться з метою приведення всіх чотирьох датасетів до однакового вигляду. Далі набори даних перевіряються на наявність пропусків, хибних значень чи дублікатів та очищаються від них за потреби. Окремо слід виділити етап вибору єдиної назви для кожного ЗСО, що буде використовуватися для його ідентифікації у даних за всі чотири роки.

Чисті датасети проходять ще декілька етапів підготовки: масштабування ознак та зменшення розмірності їх простору. Кожен об'єкт $\vec{x}_i$ характеризується вісьмома атрибутами:

- $x^{(1)}$ – кількість випускників ЗСО;
- $x^{(2)}$ – частка випускників ЗСО, які не складали ЗНО або НМТ з математики;
- $x^{(3)}$ – частка випускників ЗСО, які не набрали пороговий бал та не склали ЗНО або НМТ з математики;
- $x^{(4)}$ – частка випускників ЗСО, які склали ЗНО або НМТ з математики на 100 – 120 балів;
- $x^{(5)}$ – частка випускників ЗСО, які склали ЗНО або НМТ з математики на 120 – 140 балів;
- $x^{(6)}$ – частка випускників ЗСО, які склали ЗНО або НМТ з математики на 140 – 160 балів;
- $x^{(7)}$ – частка випускників ЗСО, які склали ЗНО або НМТ з математики на 160 – 180 балів;
- $x^{(8)}$ – частка випускників ЗСО, які склали ЗНО або НМТ з математики на 180 – 200 балів.

Проводимо стандартизацію всіх ознак, окрім $x^{(1)}$, перш ніж знизити їх вимірність методом головних компонент. Шляхом підстановки до формули (2.2) початкових значень ознак розраховуємо масштабовані величини $\tilde{\vec{x}}_i$.

На останньому кроці попередньої обробки даних проводимо компонентний аналіз стандартизованих атрибутів об'єктів. Ця процедура забезпечує можливість графічного подання утворених кластерів шляхом проєктування початкових даних на декартову площину з якомога меншими втратами інформації. Відповідно до алгоритму методу головних компонент обчислюємо на основі стандартизованих атрибутів інтегральні показники (2.3). Для зображення результатів кластеризації у двовимірному просторі лишаємо перші дві головні компоненти, які найкраще відповідають вихідному датасету, зберігаючи мак-

симальну частку його варіації.

Отже, після попередньої обробки даних за кожен рік отримаємо масив точок, кожна з яких задається координатами:

$$y_i^{(1)} = \sum_{j=1}^7 \alpha_{1j} \tilde{x}_i^{(j)}, \quad y_i^{(2)} = \sum_{j=1}^7 \alpha_{2j} \tilde{x}_i^{(j)}, \quad i = \overline{1, n_m}, \quad n_m = \overline{1, 4}.$$

2.4.2 Кластеризація даних методами k -means та k -medoids

Досліджувані дані з результатами ЗНО та НМТ випускників ЗСО м. Харків не містять інформації щодо загального рівня академічної підготовки до іспитів з математики у навчальних закладах міста. Тому для аналізу успішності учнів та якості викладання дисципліни у ЗСО, що значною мірою впливає на здобутки випускників, доцільним буде застосовувати методи аналізу даних без вчителя, зокрема неієрархічні ітеративні методи кластеризації.

Розбиваємо підготовлений набір даних на кластери із застосуванням методів k -means та k -medoids. Вони відрізняються між собою лише вибором та інтерпретацією центроїдів на кожній ітерації. Опишемо процедуру кластерного аналізу за цими двома алгоритмами. На першій ітерації серед зразків даних випадковим чином обираються k центрів ваг майбутніх кластерів $\vec{\mu}_l$, $l = \overline{1, k}$. Всі інші об'єкти $\vec{x}_i$, $\vec{x}_i \neq \vec{\mu}_l$, групуються навколо визначених центроїдів $\vec{\mu}_l$ за принципом близькості до одного з них за обраною метрикою відстані. Починаючи з другої ітерації, центри кластерів обчислюються по-іншому. У методі k -means вони отримуються шляхом усереднення ознак об'єктів кластерів. Отримані таким чином центроїди є абстрактними точками двовимірного простору, що у загальному випадку не належать датасету. Медоїдами $\vec{m}_l$, $l = \overline{1, k}$, виступають зразки аналізованого масиву даних, що вважаються типовими представниками своїх кластерів. На кожному кроці вони визначаються як точки, відносно яких

найкраще згруповані інші об'єкти. Таким чином, процес кластеризації набору даних повторюється до виконання умови зупинки. Вихідним результатом кластерного аналізу є розбиття датасету на k множин, кожна з яких складається із ЗСО м. Харків зі схожим рівнем досягнень ЗНО та НМТ з математики випускників обраного року (2021 – 2024 роки). Для подальшого аналізу отримані кластери візуалізуються на координатній площині.

У кожній групі зразки даних об'єднані загальними закономірностями та тенденціями у результатах ЗНО та НМТ з математики за той рік, дані за який аналізувались. Для оцінювання рівня академічної підготовки учнів до іспиту та якості освіти слід розглянути характеристики кожного кластера та сформулювати на основі виявлених патернів практично значущі висновки, які відображатимуть поточний стан галузі середньої освіти та сприятимуть його покращенню у перспективі. Аналізу кластерів в межах одного року, а також міжрічної динаміки змін у результатах складання іспиту з математики присвячено ще один етап розв'язання задачі. Він є підсумковим у проведеному дослідженні.

Після кластеризації даних постає питання кількісної оцінки валідності отриманих кластерів. Для того, щоб визначити, наскільки ефективною та коректною виявилася кластеризація, розраховуємо силуетний коефіцієнт, індекси Калінські – Харабаса та Данна. Значення цих внутрішніх мір валідності кластерів дозволяють оцінити якість кластеризації та за потреби скорегувати використаний підхід, а також вивчити вплив особливостей геометричної структури даних на кластеризацію, зокрема, наявності викидів або розтягнутої форми.

Висновки за розділом 2

З урахуванням цілей розв'язання поставленої задачі дослідження навчальних досягнень випускників ЗСО м. Харків у 2021 – 2024 роках було прийнято рішення розв'язувати її із застосуванням методів інтелектуального аналізу: неієрархічної кластеризації з попередньої передобробкою даних. Отже, другий

розділ роботи присвячено теоретичним концепціям перелічених підходів з подальшим описом їх прикладного використання для аналізу результатів ЗНО та НМТ з математики. За результатами його написання можемо дійти наступних висновків:

1. Перший крок розв’язання задачі – передобробка даних – визначає якість оброблюваних даних, а отже і коректність подальшого розв’язку. Тому критично важливим є виконання всіх необхідних процедур первинної підготовки датасетів. Набори даних з результатами ЗНО та НМТ спочатку було очищено від дублікатів, помилкових значень, шуму та пропусків. Після цього було здійснено стандартизацію атрибутів зразків даних. Наостанок до масштабованого датасету було застосовано метод головних компонент з метою зменшення вимірності початкового простору ознак.

2. Було досліджено специфіку застосування до підготовлених даних методів неієрархічної кластеризації k -means та k -medoids, наведено покрокові алгоритми цих методів з описом особливостей їх застосування для аналізу розглядуваних датасетів.

3. Для визначення якості проведеної кластеризації слід не тільки досліджувати виявлені закономірності, значущі для прикладного використання, а і оцінювати структуру кластерів за допомогою кількісних мір. Отже, було розглянуто внутрішні міри валідації кластерів та їх інтерпретацію для визначення ефективності кластерного аналізу.

3 ПРОГРАМНА РЕАЛІЗАЦІЯ

3.1 Mathematica 10 як система комп'ютерної алгебри

Wolfram Mathematica є комплексною системою комп'ютерної алгебри з потужним функціоналом для розв'язання задач різного профілю. Вона інтегрує у собі символічні і чисельні розрахунки, обробку великих даних та реалізацію методів їх інтелектуального аналізу, створення різноманітних візуалізацій, які наочно представляють хід та результати обчислень тощо. Система побудована на основі символічної мови програмування Wolfram Language. Її природність дозволяє описувати складні математичні концепції доступним способом та спрощує вивчення синтаксису.

Користувачський інтерфейс Wolfram Mathematica виділяється своєю універсальністю та простотою для новачків. Середовище розробки організовано як блокнот з інтерактивними комірками для написання програмного коду. Модульний підхід дозволяє створювати програми поетапно, запускати та тестувати їх фрагменти незалежно один від одного. Це оптимізує процес налагодження: помилки в одній з комірок не впливають на компіляцію інших, забезпечуючи функціональність програми. Наприклад, якщо в модулі обчислення виникає помилка, візуалізаційний блок може працювати автономно. Окрім того, у Wolfram Mathematica можна одразу документувати розроблений код, створюючи комірки з текстовими коментарями та примітками. У файлі програми зберігаються результати кожного запуску її локального ядра, і за потреби подальшого використання їх можна експортувати у вигляді окремого pdf-документу.

Wolfram Mathematica підтримує імпорт даних у багатьох форматах, зокрема, xls, csv, txt та інших. Для підготовки та аналізу завантаженого датасету у бібліотеці математичного пакету пропонуються сотні вбудованих функцій: для очищення та оптимізації даних, їх групування чи сортування або здійснення їх повноцінного статистичного дослідження, наприклад, розрахунку базових мір, таких як середнє вибіркове значення та медіана, або перевірки низки статисти-

чних гіпотез. Mathematica 10 також має розширені можливості для проведення data mining та побудови моделей машинного навчання, включаючи кластеризацію. Користувачі математичного пакету можуть застосувати готові методи з її бібліотеки для аналізу вхідних даних або розробити їх самостійно, використовуючи доступне різноманіття обчислювальних виразів та конструкцій для програмування.

Більшість результатів аналізу даних зручніше подавати у графічній формі. У Wolfram Mathematica існує багатий набір інструментів для візуалізації, зокрема, побудови гістограм, стовпчатих та кругових діаграм, точкових графіків тощо. Функції Graphics чи Plot дозволяють працювати з кольорами, мітками, анімацією. Графічні засоби Mathematica підтримують створення і статичних, і динамічних візуалізацій зі змінними параметрами, що додатково підкреслює потужність функціоналу цього математичного пакету [30].

3.2 Алгоритм розв'язання задачі оцінювання навчальних досягнень випускників ЗСО м. Харків у 2021 – 2024 роках методами неієрархічної кластеризації

Наведемо покроковий алгоритм розв'язання поставленої задачі.

Етап 1. Отримати з архівів УЦОЯО статистичні дані з результатами ЗНО та НМТ за 2021 – 2024 роки з математики для м. Харків.

Етап 2. Звести отримані датасети до єдиного вигляду шляхом їх агрегації за низкою ознак.

Етап 3. Дослідити домен середньої освіти, до якого відносяться вивантажені датасети. Іншими словами, вивчити адміністративні зміни, пов'язані із навчальними закладами м. Харків, та вплив цих змін на хід і результати подальшого аналізу даних.

Етап 4. За допомогою спеціально розроблених методів у програмі попередньо обробити набори даних: очистити їх від дублікатів, пропусків і помилок,

після чого здійснити стандартизацію атрибутів та подальший компонентний аналіз з метою переходу до двовимірного простору ознак.

Етап 5. Алгоритми неієрархічної кластеризації k -means та k -medoids імплементувати програмно символьною мовою Wolfram.

Етап 6. Обрати внутрішні міри валідності кластерів та написати блок функцій для їх автоматизованого розрахунку.

Етап 7. Встановити значення вхідних параметрів, зокрема, точність розрахунку відстані між центрами ваг кластерів на сусідніх ітераціях ε і кількість майбутніх кластерів k .

Етап 8. Здійснити попередню обробку даних з результатами ЗНО та НМТ та кластеризувати їх, використовуючи написану програму.

Етап 9. Зобразити графічно результати кластерного аналізу датасетів: кластери – на координатній площині, розподіл балів випускників ЗСО у кожному з кластерів – на кругових діаграмах.

Етап 10. Проаналізувати закономірності та тренди у кожному утвореному кластері та зробити на їх основі прикладні висновки щодо якості освітньої підготовки з математики у ЗСО м. Харків.

Етап 11. Провести порівняльний аналіз кластерів, сформованих за результатами ЗНО та НМТ, по роках та дослідити динаміку змін рівня академічної підготовки до іспитів з часом.

Етап 12. Оцінити якість сформованих кластерів, розрахувавши значення обраних для цієї мети метрик.

3.3 Опис програми

Інтелектуальний аналіз датасету з результатами ЗНО та НМТ з математики випускників навчальних закладів м. Харків за 2021 – 2024 роки було здійснено за допомогою розрахункової програми, розробленої у математичному пакеті Wolfram Mathematica 10.4.

Програмний код структуровано у вигляді модулів, кожен з яких відповідає окремому етапу розв’язання поставленої задачі. Його повний лістинг поданий у додатку А. Наведемо нижче опис створених функціональних модулів.

Спочатку виконується імпорт датасету та його попередня обробка. Агреговані результати ЗНО та НМТ з математики за кожен рік завантажуються у вигляді окремих масивів. Унаслідок адміністративних реформ частину ЗСО Харкова було перейменовано, що створює ризик некоректного співставлення об’єктів при міжрічному порівнянні. Для забезпечення правильної ідентифікації навчальних закладів у коді реалізовано алгоритм приведення їх назв до єдиного варіанту. Водночас у програмі проводиться обробка порожніх та помилкових значень. Надалі за допомогою самостійно розробленої функції `StdAndPCA()` проводяться стандартизація атрибутів очищених даних та їх компонентний аналіз за методом PCA.

Підготовлені дані слугують аргументами функцій наступного модуля для кластеризації ЗСО. У ньому реалізовані алгоритми *k*-means та *k*-medoids, в яких стартовий набір центроїдів визначається за допомогою окремого методу `Fstab()`. До блоку кластерного аналізу також входять наступні функції: `ClusteringGraphic()` для графічного подання його результатів на декартовій площині; `ClusteringAbrevs()`, який виводить аббревіатури ЗСО у складі кластерів; `PiechartDistribution()`, що повертає секторні діаграми розподілу балів ЗНО та НМТ з математики між ЗСО всередині кожної утвореної групи. Валідність сформованих кластерів оцінюється не тільки за практично значущими закономірностями у них, а і по розрахованих метриках якості. У програмі створений блок методів, зокрема, `Silhouette()` чи `CalinskiHarabasz()` для обчислення внутрішніх мір якості кластеризації.

Безпосереднє здійснення інтелектуального аналізу вихідних даних із використанням розроблених функцій складає останній, найбільший за обсягом, програмний модуль. Оскільки отримані результати порівнюються по роках, то для коректного співставлення кластерів на основі подібного рівня успішності учнів ЗСО у коді прописано їх ранжування за середніми балами. Наприкінці

програми додано функціонал, що передбачає більш глибокий аналіз рівня навчальної підготовки у ЗСО м. Харків, зокрема, метод, який повертає таблицю з переходами кожного ЗСО між кластерами по роках.

Висновки за розділом 3

Для розв’язання задачі оцінювання навчальних досягнень випускників ЗСО м. Харків було використано систему комп’ютерної алгебри Wolfram Mathematica 10.4. У третьому розділі кваліфікаційної роботи представлено огляд обчислювальних можливостей цього математичного пакету, а також описано програмний код, за допомогою якого здійснюється інтелектуальний аналіз балів ЗНО та НМТ з математики за 2021 – 2024 роки. Програмну реалізацію розв’язання задачі було здійснено за алгоритмом, який також подано у цьому розділі. У підсумку можемо зробити такі висновки:

1. Wolfram Mathematica 10.4 – це мультифункціональна система, що підтримує ефективне проведення інтелектуального аналізу даних за допомогою інтегрованих в неї засобів для передобробки даних, побудови алгоритмів кластеризації та візуалізації отриманих результатів.

2. Покроковий план розв’язання розглядуваної задачі, який було створено та наведено у розділі, є важливим підготовчим етапом для його програмної реалізації, що спрощує цей процес та робить його ефективнішим.

3. Різноманіття вбудованих функцій мови символьної математики Wolfram у поєднанні з обчислювальною потужністю та зручним інтерфейсом середовища розробки Mathematica 10.4 забезпечують ефективне автоматизоване виконання поставленої задачі.

4 РЕЗУЛЬТАТИ ОБЧИСЛЮВАЛЬНОГО ЕКСПЕРИМЕНТУ ТА ЇХ АНАЛІЗ

4.1 Компонентний аналіз результатів ЗНО та НМТ за 2021 – 2024 роки

У кваліфікаційній роботі досліджуються дані з балами ЗНО та НМТ з математики випускників ЗСО м. Харків за 2021 – 2024 роки. Вихідний датасет, завантажений із сайту УЦОЯО, складається з чотирьох таблиць, кожна з яких містить деперсоналізовану інформацію про випускників та їх навчальні досягнення за обраний рік [31]. Кожен запис містить загальні відомості про окремого учня, дані про ЗСО, де він навчався, а також про місце, дату та результати складання тестів. На рисунку 4.1 наведено приклад вихідних даних за 2022 рік.

Рік народження випускника	Стать випускника	Рік випуску	Назва ЗСО	Статус ЗСО	Населений пункт розташування ЗСО	Район населеного пункту розташування ЗСО	Дата іспиту	Статус складання іспиту	Бал складання іспиту	Регіон складання іспиту	Район складання іспиту
2006	жіноча	Випускник закладу загальної середньої освіти 2023 року	КОМУНАЛЬНИЙ ЗАКЛАД "ХАРКІВСЬКА СПЕЦІАЛІЗОВАНА ШКОЛА І-ІІІ СТУПЕНІВ З ПОГЛИБЛЕНИМ ВИВЧЕННЯМ ОКРЕМИХ ПРЕДМЕТІВ № 16 ХАРКІВСЬКОЇ МІСЬКОЇ РАДИ ХАРКІВСЬКОЇ ОБЛАСТІ ІМЕНІ В.Г. СЕРГЄВА"	спеціалізована школа	м. Харків	Київський район міста	19.06.2023	Зараховано	121	м. Харків, Шевченківський район міста	Шевченківський район міста
2006	жіноча	Випускник закладу загальної середньої освіти 2023 року	ХАРКІВСЬКА ЗАГАЛЬНООСВІТНЯ ШКОЛА І-ІІІ СТУПЕНІВ №157 ХАРКІВСЬКОЇ МІСЬКОЇ РАДИ ХАРКІВСЬКОЇ ОБЛАСТІ	середня загальноосвітня школа	м. Харків	Індустріальний район міста	16.06.2023	Зараховано	134	м. Гамбург	м. Гамбург
2006	чоловіча	Випускник закладу загальної середньої освіти 2023 року	ХАРКІВСЬКА СПЕЦІАЛІЗОВАНА ШКОЛА І-ІІІ СТУПЕНІВ №114 ХАРКІВСЬКОЇ МІСЬКОЇ РАДИ ХАРКІВСЬКОЇ ОБЛАСТІ	спеціалізована школа	м. Харків	Слобідський район міста	19.06.2023	Зараховано	151	м. Харків, Шевченківський район міста	Шевченківський район міста
2006	чоловіча	Випускник закладу загальної середньої освіти 2023 року	ХАРКІВСЬКА ЗАГАЛЬНООСВІТНЯ ШКОЛА І-ІІІ СТУПЕНІВ №68 ХАРКІВСЬКОЇ МІСЬКОЇ РАДИ ХАРКІВСЬКОЇ ОБЛАСТІ	середня загальноосвітня школа	м. Харків	Слобідський район міста	08.06.2023	Зараховано	157	м. Харків, Новобаварський район міста	Новобаварський район міста
2006	жіноча	Випускник закладу загальної середньої освіти 2023 року	КОМУНАЛЬНИЙ ЗАКЛАД "ХАРКІВСЬКИЙ НАВЧАЛЬНО-ВИХОВНИЙ КОМПЛЕКС - ГІМНАЗИЯ-ШКОЛА І СТУПЕНЯ" №24 ХАРКІВСЬКОЇ МІСЬКОЇ РАДИ ХАРКІВСЬКОЇ ОБЛАСТІ ІМЕНІ І.Н. ПІТКОВА-	навчально-виховний комплекс	м. Харків	Ненишлянський район міста	16.06.2023	Зараховано	107	м. Гамбург	м. Гамбург

Рисунок 4.1 – Вихідні дані з результатами НМТ з математики випускників ЗСО м. Харків за 2022 рік

Для аналізу необхідно подати вихідні дані у вигляді інтервального статистичного ряду з відсотковим розподілом балів ЗНО та НМТ для кожного ЗСО м. Харків. Перетворення початкового датасету розпочинаємо з виділення чотирьох необхідних атрибутів з дванадцяти, наведених на рисунку 4.1: назви та типу ЗСО, в якому навчався випускник, статусу та балу складання іспиту з математики (таблиця 4.1).

Таблиця 4.1 – Фрагмент підготовлених до групування даних з виділеними для аналізу атрибутами

Назва ЗСО	Статус ЗСО	Статус складання Іспиту	Бал складання Іспиту
Комунальний заклад «Харківська спеціалізована школа І – ІІІ ступенів з поглибленим вивченням окремих предметів № 16 Харківської міської ради Харківської області імені В.Г.Сергєєва»	Спеціалізована Школа	Зараховано	121
Харківська загальноосвітня школа І – ІІІ ступенів № 157 Харківської міської ради Харківської області	Середня загальноосвітня школа	Зараховано	134
Харківська спеціалізована школа І – ІІІ ступенів № 114 Харківської міської ради Харківської області	Спеціалізована Школа	Зараховано	151
Харківська загальноосвітня школа І – ІІІ ступенів № 68 Харківської міської ради Харківської області	Середня загальноосвітня школа	Зараховано	157
Комунальний заклад «Харківський навчально-виховний комплекс «Гімназія-школа І ступеня» № 24 Харківської міської ради Харківської області імені І.Н.Питікова»	Навчально-виховний комплекс	Зараховано	107

Далі групуємо дані за назвою навчального закладу. Для кожного ЗСО обчислюємо загальну кількість випускників, що складали іспит, та частку тих, хто зареєструвався, але не з'явився на тестування. Також формуємо інтервальний розподіл отриманих результатів за шкалою 100 – 200 балів з кроком інтервалу у 20 балів. Таблиця 4.2 ілюструє зразок підготовленого до передобробки датасету з результатами НМТ за 2022 рік.

Таким чином, отримуємо чотири зведені таблиці, аналогічні таблиці 4.2, із показниками успішності випускників харківських ЗСО за 2021 – 2024 роки. Обсяг наборів даних варіюється від року до року. У 2021 році ЗНО складали випускники 243 шкіл. У 2022 році у написанні НМТ брали участь учні з 245 шкіл, у 2023 – з 247, та у 2024 – лише з 223. Оскільки математика є обов'язковою для складання дисципліною, то кількість об'єктів у наборах даних відображає загальну чисельність ЗСО м. Харків, випускники яких проходили тестування.

Таблиця 4.2 – Згруповані дані з результатами НМТ з математики
у м. Харків за 2022 рік

Заклад освіти/інші категорії учасників		Кількість осіб, що складалі іспит	% учасників, які						
Назва	Тип		не з'явилися на іспит	не подолали поріг	Отримали відповідний результат за шкалою 100 – 200 балів				
					[100, 120)	[120, 140)	[140, 160)	[160, 180)	[180, 200)
Комунальний заклад «Харківська спеціалізована школа І – ІІІ ступенів з поглибленим вивченням окремих предметів № 16 Харківської міської ради Харківської області імені В.Г.Сергєєва»	Спеціалізована школа	25	3,85	0,00	0,00	0,00	53,85	26,92	15,38
Харківська загальноосвітня школа І – ІІІ ступенів № 157 Харківської міської ради Харківської області	Середня загальноосвітня школа	48	9,43	0,00	1,89	22,64	47,17	15,09	3,77
Харківська спеціалізована школа І – ІІІ ступенів № 114 Харківської міської ради Харківської області	Спеціалізована школа	32	11,11	0,00	0,00	13,89	47,22	11,11	16,67
Харківська загальноосвітня школа І – ІІІ ступенів № 68 Харківської міської ради Харківської області	Середня загальноосвітня школа	39	18,75	0,00	4,17	20,83	37,50	10,42	8,33
Комунальний заклад «Харківський навчально-виховний комплекс «Гімназія-школа І ступеня» № 24 Харківської міської ради Харківської області імені І.Н.Питікова»	Навчально-виховний комплекс	33	8,33	0,00	0,00	27,78	38,89	19,44	5,56

Перш ніж проводити кластерний аналіз даних, необхідно стандартизувати ознаки об'єктів та виділити їх за методом головних компонент. Раніше було визначено, що для кластеризації ЗСО за рівнем навчальних досягнень випускників враховуватимуться сім показників успішності. Масштабуємо ознаки об'єктів

датасету $x_i^{(j)}$, $j = \overline{1,7}$, згідно з формулою (2.1). Приклад стандартизованих значень $\tilde{x}_i^{(j)}$ для кількох ЗСО у 2022 році, представлено у таблиці 4.3. Стандартизація атрибутів є обов'язковим етапом передобробки даних для кластеризації, бо усуває вплив різного масштабу вимірювань і гарантує коректну роботу неієрархічних алгоритмів, які базуються на метриках відстані.

Таблиця 4.3 – Стандартизовані ознаки об'єктів даних $\tilde{x}_i^{(j)}$

Об'єкт	$\tilde{x}^{(1)}$	$\tilde{x}^{(2)}$	$\tilde{x}^{(3)}$	$\tilde{x}^{(4)}$	$\tilde{x}^{(5)}$	$\tilde{x}^{(6)}$	$\tilde{x}^{(7)}$
Комунальний заклад «Харківська спеціалізована школа I – III ступенів з поглибленим вивченням окремих предметів № 16 Харківської міської ради Харківської області імені В.Г.Сергєєва»	-0,849	-0,107	-0,484	-1,418	0,871	1,462	0,423
Харківська загальноосвітня школа I – III ступенів № 157 Харківської міської ради Харківської області	-0,401	-0,107	-0,484	0,622	0,755	-0,409	-0,856
Харківська спеціалізована школа I – III ступенів № 114 Харківської міської ради Харківської області	0,164	-0,107	1,815	0,320	0,222	-0,923	-0,548

До масштабованих даних застосовуємо метод головних компонент (РСА). Компонентний аналіз у задачі кластеризації здійснюється для переходу до нового двовимірного простору ознак, в якому можна графічно зобразити сформовані кластери. Це дозволить візуально оцінити якість кластеризації та полегшить змістовну інтерпретацію результатів. Перші дві головні компоненти $y^{(1)}$, $y^{(2)}$ пояснюють найбільшу частку варіації вихідного датасету, тому їх використання дає змогу зберегти основні закономірності та структуру даних з мінімальними втратами інформації. Обчислимо інтегральні показники $y^{(1)}$, $y^{(2)}$ для даних з результатами тестування за чотири роки.

Головні компоненти $y_{2021}^{(1)}$ та $y_{2021}^{(2)}$, розраховані на основі навчальних до-

сягнень випускників ЗСО м. Харків за 2021 рік, дорівнюють:

$$y_{2021}^{(1)} = 0,08610\tilde{x}^{(1)} + 0,49295\tilde{x}^{(2)} + 0,46462\tilde{x}^{(3)} - 0,03846\tilde{x}^{(4)} - \\ - 0,32711\tilde{x}^{(5)} - 0,48241\tilde{x}^{(6)} - 0,43878\tilde{x}^{(7)}, \quad (4.1)$$

$$y_{2021}^{(2)} = 0,14405\tilde{x}^{(1)} + 0,22398\tilde{x}^{(2)} - 0,12084\tilde{x}^{(3)} - 0,74938\tilde{x}^{(4)} - \\ - 0,42139\tilde{x}^{(5)} + 0,11836\tilde{x}^{(6)} + 0,40166\tilde{x}^{(7)}. \quad (4.2)$$

Величини (4.1), (4.2) зберігають 55,59% початкової дисперсії набору даних у новому просторі ознак. При підстановці значень $\tilde{x}_i^{(j)}$, $j = \overline{1,7}$, до формул (4.1), (4.2) отримаємо координати точок $\tilde{\tilde{x}}_{i(2021)} = (y_{2021}^{(1)}; y_{2021}^{(2)})$, $i = \overline{1,243}$, що відповідають досліджуваним ЗСО на площині. Наприклад, Комунальний заклад «Харківська спеціалізована школа I – III ступенів з поглибленим вивченням окремих предметів № 16 Харківської міської ради Харківської області імені В.Г.Сергєєва» задається точкою $\tilde{\tilde{x}}_{41(2021)} = (1,80748; -0,05515)$. Інші об'єкти визначаються подібним чином.

Розглядаємо датасет з показниками успішності за 2022 рік. Для нього формули для перших двох головних компонент $y_{2022}^{(1)}$ та $y_{2022}^{(2)}$ мають вигляд:

$$y_{2022}^{(1)} = 0,38311\tilde{x}^{(1)} + 0,08166\tilde{x}^{(2)} + 0,35965\tilde{x}^{(3)} - 0,20234\tilde{x}^{(4)} - \\ - 0,53467\tilde{x}^{(5)} - 0,46499\tilde{x}^{(6)} - 0,45143\tilde{x}^{(7)}, \quad (4.3)$$

$$y_{2022}^{(2)} = -0,32282\tilde{x}^{(1)} + 0,02501\tilde{x}^{(2)} - 0,27125\tilde{x}^{(3)} + 0,09304\tilde{x}^{(4)} + \\ + 0,74498\tilde{x}^{(5)} - 0,28846\tilde{x}^{(6)} - 0,41800\tilde{x}^{(7)}. \quad (4.4)$$

Вирази (4.3), (4.4) сумарно пояснюють 52,17% варіабельності вихідних даних. Користуючись ними, отримаємо вектори $\tilde{\tilde{x}}_{i(2022)} = (y_{2022}^{(1)}; y_{2022}^{(2)})$, $i = \overline{1,245}$, для позначення навчальних закладів у новому базисі, наприклад:

– Комунальний заклад «Харківська спеціалізована школа I – III ступенів з поглибленим вивченням окремих предметів № 16 Харківської міської ради Харківської області імені В.Г.Сергєєва» – $\tilde{x}_{41(2022)} = (2,27342; -0,32093)$;

– Харківська загальноосвітня школа I – III ступенів № 157 Харківської міської ради Харківської області – $\tilde{x}_{83(2022)} = (0,00210; -0,54753)$ тощо.

Аналогічно зменшуємо розмірність даних за 2023 рік. Формули для перших двох головних компонент $y_{2023}^{(1)}$ та $y_{2023}^{(2)}$ мають вигляд:

$$y_{2023}^{(1)} = 0,22086\tilde{x}^{(1)} + 0,22513\tilde{x}^{(2)} + 0,37512\tilde{x}^{(3)} + 0,42090\tilde{x}^{(4)} - 0,46437\tilde{x}^{(5)} - 0,44496\tilde{x}^{(6)} - 0,41114\tilde{x}^{(7)}, \quad (4.5)$$

$$y_{2023}^{(2)} = -0,42905\tilde{x}^{(1)} - 0,56555\tilde{x}^{(2)} - 0,02758\tilde{x}^{(3)} + 0,38634\tilde{x}^{(4)} + 0,40465\tilde{x}^{(5)} - 0,28700\tilde{x}^{(6)} - 0,31621\tilde{x}^{(7)}. \quad (4.6)$$

Після проектування вихідних даних на побудовані вектори (4.5), (4.6), зберігається 51,46% дисперсії. Результатом компонентного аналізу є масив точок $\tilde{x}_{i(2023)} = (y_{2023}^{(1)}; y_{2023}^{(2)})$, $i = \overline{1, 247}$, які представляють ЗСО у новому просторі ознак. Наприклад, Харківська спеціалізована школа I – III ступенів № 114 Харківської міської ради Харківської області має координати $\tilde{x}_{201(2023)} = (0,91100; 0,16194)$.

Трансформуємо набір даних, в якому наводяться результати НМТ 2024 року. Агреговані показники $y_{2024}^{(1)}$ та $y_{2024}^{(2)}$ задаються виразами:

$$y_{2024}^{(1)} = 0,20473\tilde{x}^{(1)} + 0,27396\tilde{x}^{(2)} + 0,38387\tilde{x}^{(3)} + 0,38674\tilde{x}^{(4)} - 0,66933\tilde{x}^{(5)} - 0,26819\tilde{x}^{(6)} - 0,25725\tilde{x}^{(7)}, \quad (4.7)$$

$$y_{2024}^{(2)} = -0,40310\tilde{x}^{(1)} - 0,61448\tilde{x}^{(2)} + 0,05221\tilde{x}^{(3)} + 0,59642\tilde{x}^{(4)} - 0,15260\tilde{x}^{(5)} + 0,16068\tilde{x}^{(6)} + 0,22887\tilde{x}^{(7)}. \quad (4.8)$$

Внесок головних компонент (4.7), (4.8) до загальної варіативності розглядуваного набору даних становить 42,14%. У новій координатній системі спостереження набувають вигляду $\tilde{x}_{i(2024)} = (y_{2024}^{(1)}; y_{2024}^{(2)})$, $i = \overline{1, 223}$. Зокрема:

– Комунальний заклад «Харківська спеціалізована школа I – III ступенів з поглибленим вивченням окремих предметів № 16 Харківської міської ради Харківської області імені В.Г.Сергєєва» – $\tilde{x}_{41(2024)} = (1,57914; 0,33189)$;

– Харківська спеціалізована школа I – III ступенів № 114 Харківської міської ради Харківської області – $\tilde{x}_{201(2024)} = (1,17584; 0,75582)$ тощо.

Аналізуючи отримані результати бачимо, що з кожним наступним роком спостерігається тенденція до зниження частки збереженої дисперсії у перших двох головних компонентах. Особливо виділяється різке падіння цієї величини на 9,32% у 2024 році. Розглянемо можливі причини виникнення виявленої закономірності.

Одним з найбільш ймовірних чинників зниження інформативності даних при переході до двовимірного простору є підвищення гетерогенності датасету. Іншими словами, кожного року все сильніше проявляється вплив індивідуальних особливостей кожного окремого учня на його навчальні досягнення. Представники кожного нового покоління випускників все більше розрізняються за рівнем освітньої підготовки, застосованими методиками навчання, ступенем мотивації тощо. Додатково структура даних з результатами іспитів змінюється під дією зовнішніх кризових обставин, спричинених війною. Наприклад, повітряні тривоги негативно впливають на психологічний стан учнів як під час навчального процесу, так і при написанні тестів, та знижують когнітивну ефективність абітурієнтів. Накопичений хронічний стрес також сприяє погіршенню загальної концентрації та мотивації учнів до навчання. Окрім того, повітряні тривоги заважали сфокусованому написанню іспиту. Ще одним серйозним організаційним недоліком екзаменаційних випробувань став формат його проведення. Складання іспитів з декількох дисциплін в один день виснажує більшість випускників, знижуючи їх концентрацію до мінімуму та збільшуючи відсоток поми-

лок. Усі люди мають різні пороги стресостійкості і психологічної вразливості, тож умови написання НМТ могли значно вплинути на його результати. Додатково різномірність даних могла підсилитися за рахунок випускників, що писали іспит закордоном у безпеці та щорічною зміною кількості дисциплін для складання та тривалості тесту. Сукупна дія перелічених факторів посприяла появі шуму у датасеті з балами НМТ та його гетерогенності, що призвело до падіння частки поясненої дисперсії.

4.2 Неієрархічна кластеризація ЗСО м. Харків за результатами ЗНО та НМТ за 2021 – 2024 роки з математики

Продовжуємо дослідження навчальних досягнень випускників ЗСО м. Харків і застосуванням неієрархічних ітеративних методів кластеризації k -means та k -medoids. Згідно з алгоритмами, поділяємо набори даних із переліком навчальних закладів для кожного року на $k = 5$ кластерів за успішністю складання випускних випробувань.

Число груп розбиття було підібране у ході аналізу закономірностей, що спостерігаються всередині кожної з них. Графічне подання результатів кластеризації методом k -means наведено на рисунках 4.2 – 4.5, методом k -medoids – на рисунках 4.6 – 4.9. Перелік повних назв ЗСО у складі різних кластерів винесено в додаток Б. На графіках є назви не всіх навчальних закладів, щоб уникнути загромодження простору та плутанини у підписах близько розташованих точок.

Зазначимо, що низький відсоток збереженої варіативності даних у початковому просторі ознак не впливає на якість кластерного аналізу, адже розбиття ЗСО на групи реалізоване з урахуванням усіх семи стандартизованих вихідних атрибутів.

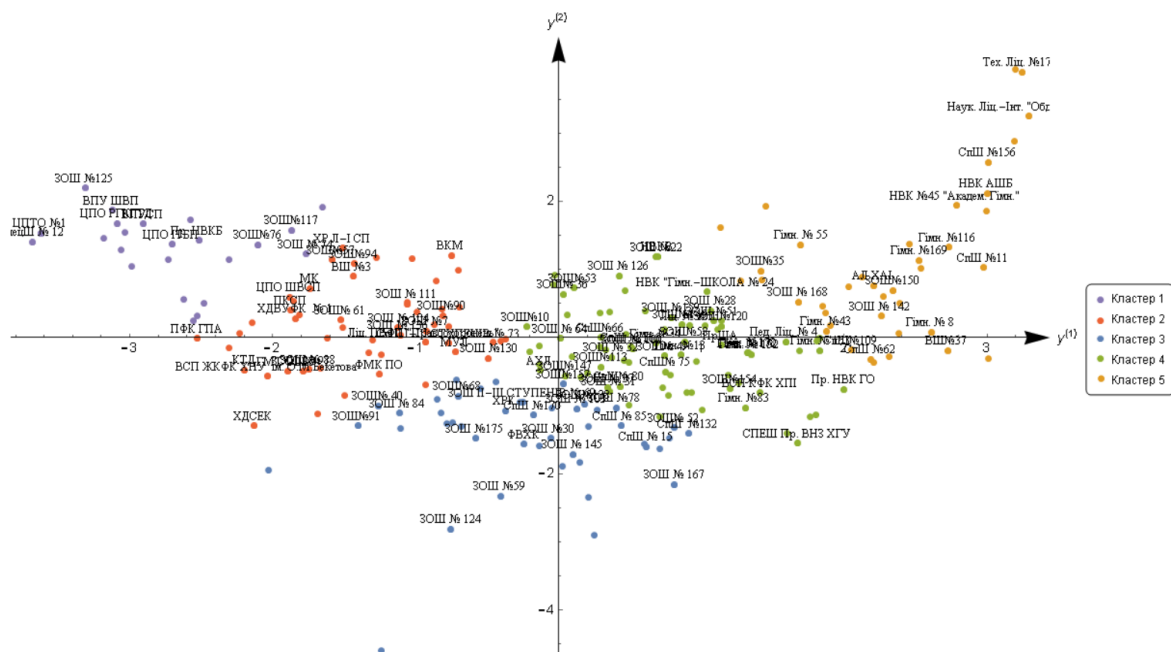


Рисунок 4.2 – Кластери ЗСО м. Харків, сформовані за результатами ЗНО 2021 року з математики методом k -means

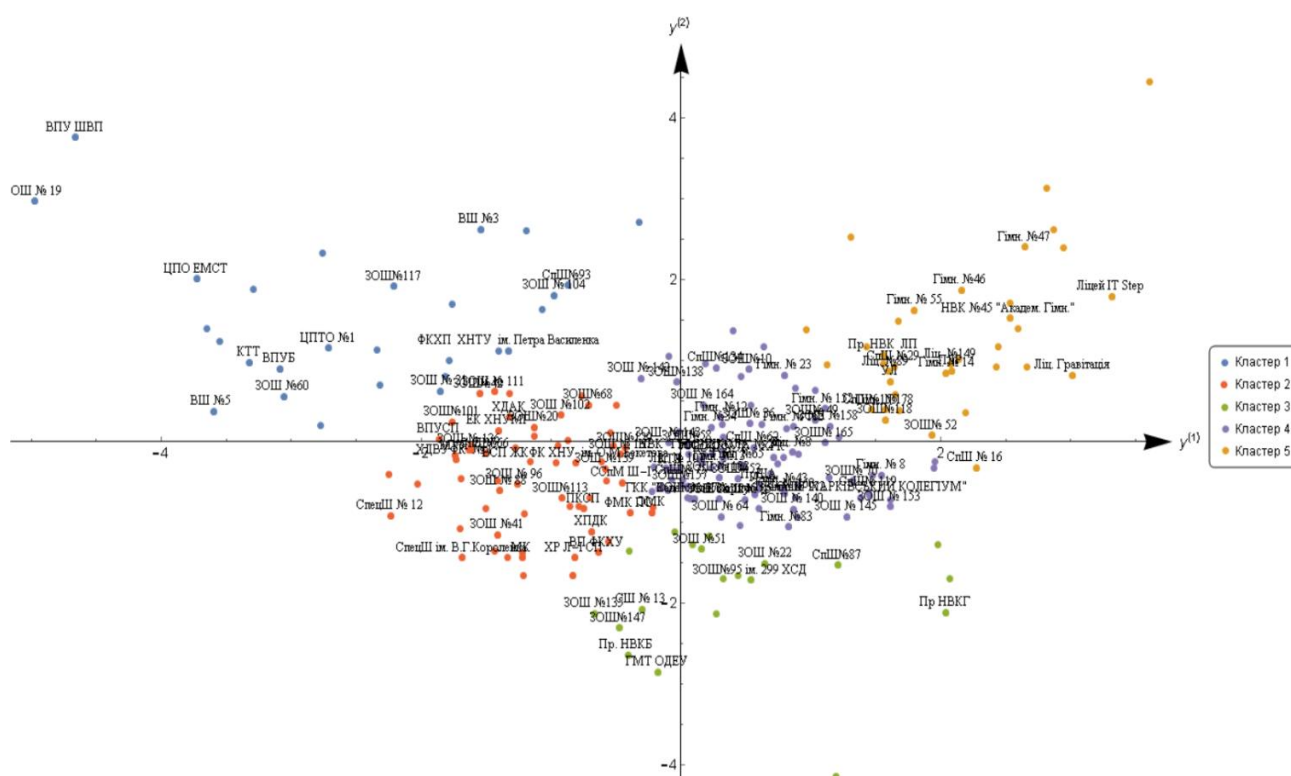


Рисунок 4.3 – Кластери ЗСО м. Харків, сформовані за результатами НМТ 2022 року з математики методом k -means

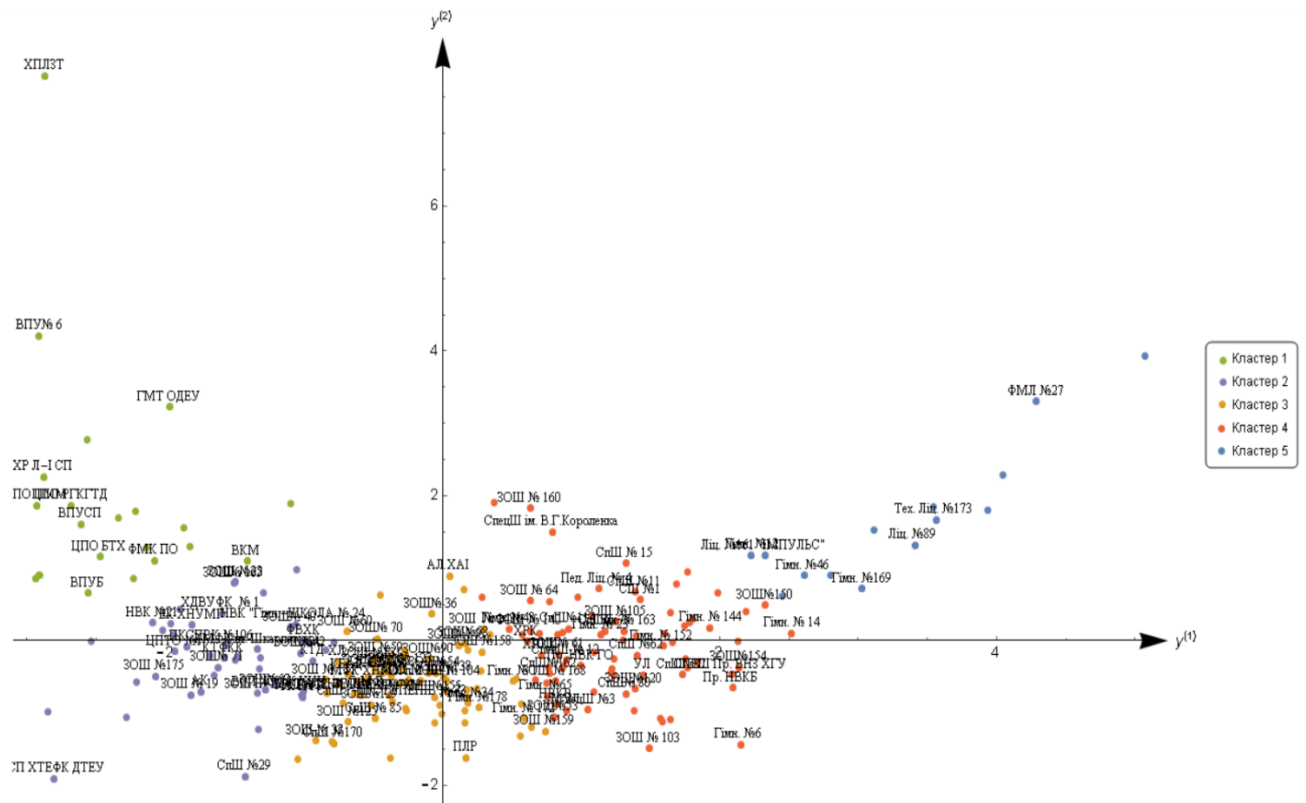


Рисунок 4.4 – Кластери ЗСО м. Харків, сформовані за результатами НМТ 2023 року з математики методом k -means

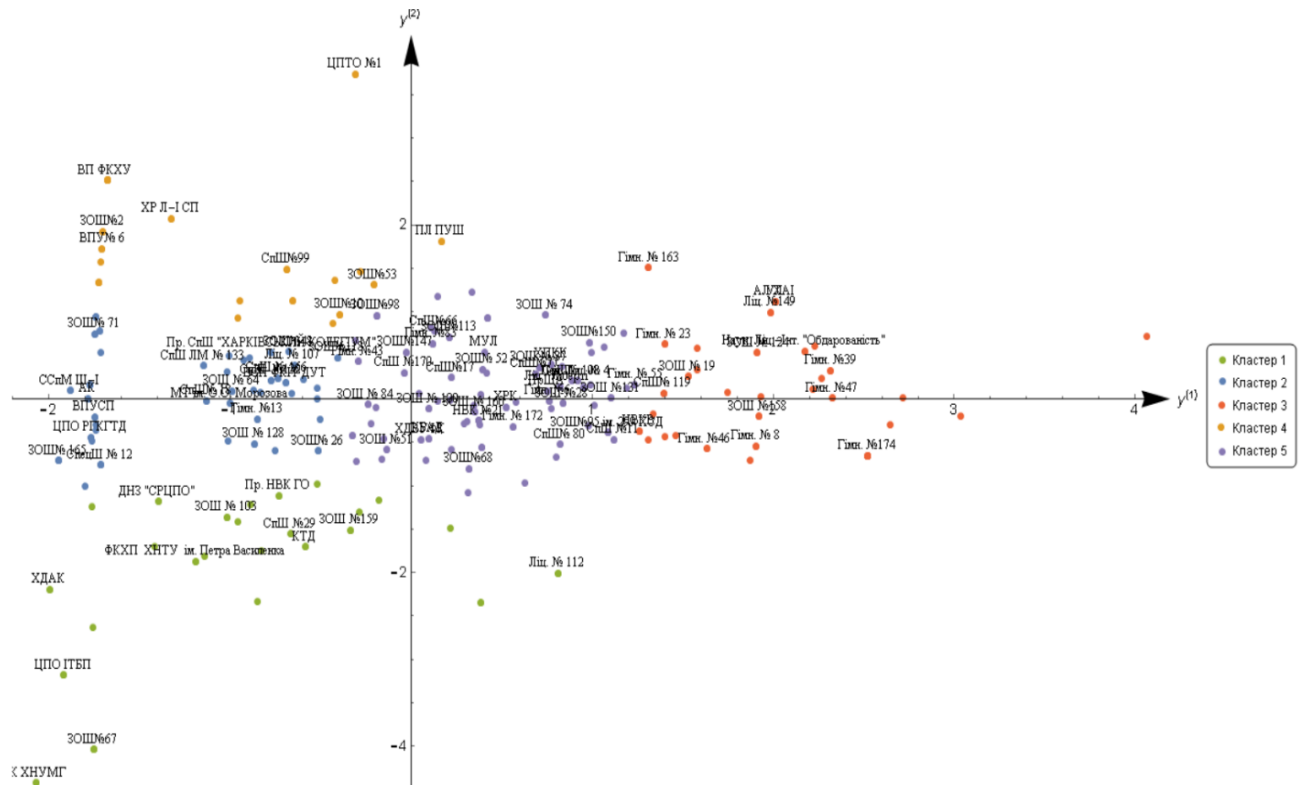


Рисунок 4.7 – Кластери ЗСО м. Харків, сформовані за результатами НМТ 2022 року з математики методом k -medoids

Рисунок 4.9 – Кластери ЗСО м. Харків, сформовані за результатами НМТ 2024 року з математики методом k -medoids

Проаналізуємо результати розбиття ЗСО м. Харків на кластери за кожен рік. Оскільки наступним етапом аналізу є вивчення змін структури та закономірностей груп ЗСО у часі, то у межах кожного року будемо їх ранжувати та нумерувати у порядку зростання середнього балу. Для порівняння у різні роки кластери будуть зіставлятися за однаковими мітками.

Розпочнемо вивчення отриманих груп ЗСО з результатів для 2021 року. У таблиці 4.3 наводимо низку властивостей кластерів, утворених методом *k*-means: кількість ЗСО та випускників у кожному з них, а також середній бал.

Таблиця 4.3 – Властивості кластерів за 2021 рік, метод *k*-means

№ кластера	1	2	3	4	5
Обсяг кластера	25	54	45	80	39
Кількість випускників у ЗСО кластера	916	1876	1756	2802	1639
Середній бал	119,976	131,801	137,022	148,854	164,907

Для демонстрації відсоткового розподілу результатів ЗНО з математики у 2021 році всередині кожного кластера будемо кругові діаграми (рисунок 4.10). Сектор графіка відображає відносну частку випускників ЗСО певного кластера, що склали тест на бали з певного проміжку шкали 100 – 200. За отриманими візуалізаціями вивчимо тенденції, що спостерігаються у кластерах, та у кожному з них оцінимо рівень успішності випускників ЗСО.

Проаналізуємо кожен отриманий кластер. Бачимо, що зі збільшенням номеру кластера зростає частка випускників з балами вище середнього та знижується відсоток тих, хто не склав іспит. Кластер 1 характеризується найгіршою якістю підготовки до ЗНО: у ньому близько половини учнів не подолали поріг у 100 балів, та ще третина – показали результати 100 – 120 балів. До кластера 2 теж потрапили ЗСО з досить слабким рівнем математичної підготовки, в яких 67,05% випускників написали тест на бали, нижчі середнього. При цьому, частка тих, хто не склав іспит, зменшилася майже вдвічі у порівнянні з першим кластером. У ЗСО кластера 3 фіксується середній рівень ефективності викладання

математики, адже у ньому відносно рівномірно розподілено відсотки середніх результатів (21,9%) та нижче середнього без урахування випускників, що не отримали пороговий бал (22,12 % – 100 – 120 балів та 25,58% – 120 – 140 балів). Частка випускників, які склали математику на високі бали, складає лише 12,76%. Кластер 4 відзначається достатньою ефективністю підготовки до ЗНО з математики. У ньому сумарна частка учнів з результатами 160 – 180 та 180 – 200 балів становить майже третину – 31,7%, а відсоток випускників, що отримали менше 100 балів, знизився у порівнянні з попередніми групами ЗСО до 11,5%. Кластер під номером 5 – позитивно виділяється з-поміж інших: до нього увійшли навчальні заклади із сильною математичною підготовкою, зокрема, профільні. У ЗСО цього кластера 59,35% випускників отримали 160 – 180 та 180 – 200 балів, та всього лиш 3,44% не змогли скласти ЗНО з математики. Відсоток нижчих за середні результати теж невисокий – 15,44%.

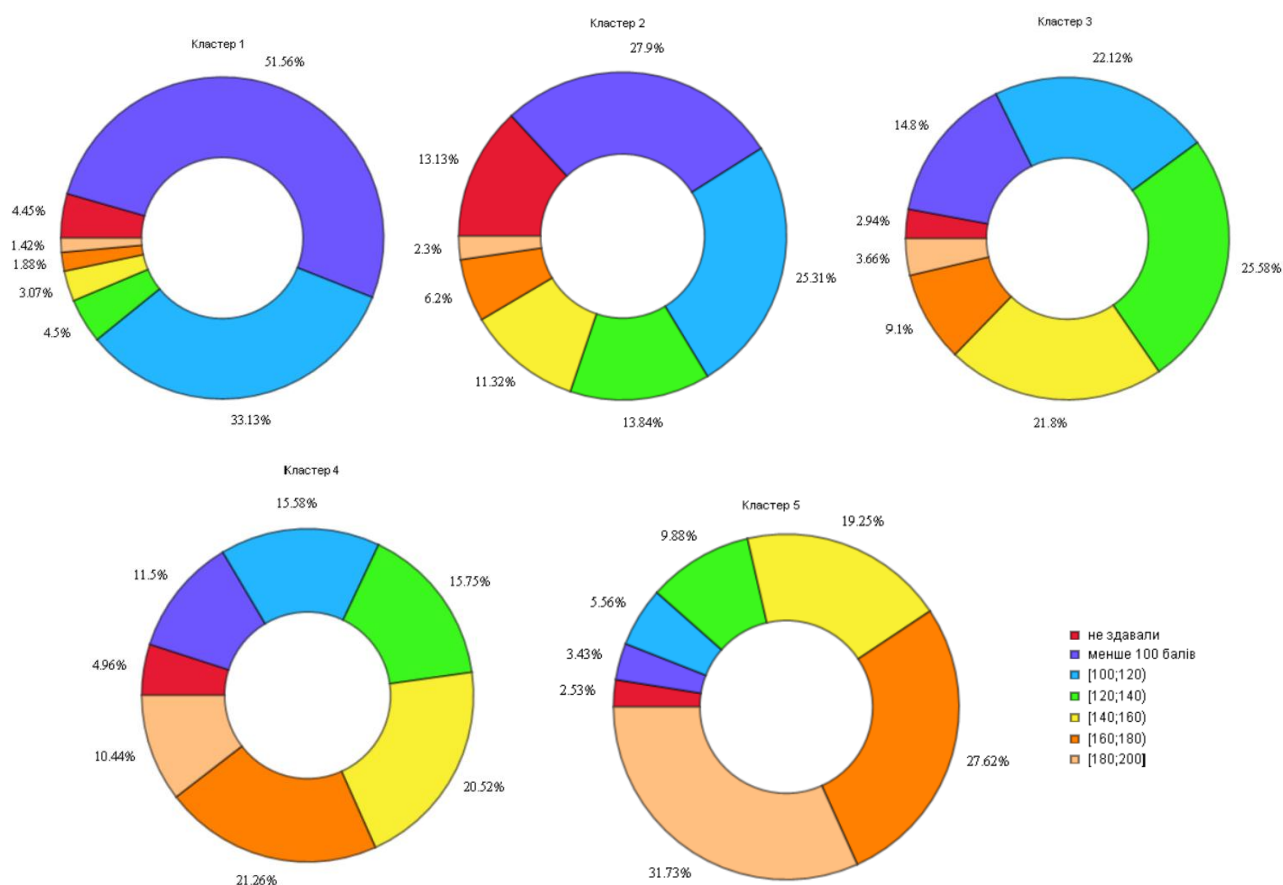


Рисунок 4.10 – Розподіл відносних часток балів ЗНО 2021 року з математики у кластерах, метод *k*-means

З 2022 року у зв'язку з початком воєнних дій у країні ЗНО було замінено на НМТ. Структура та складність завдань НМТ відрізнялись від ЗНО. Проаналізуємо результати проведення іспиту у новому форматі. Числові характеристики кластерів, сформованих методом k -means, представлено у таблиці 4.4.

Таблиця 4.4 – Властивості кластерів за 2022 рік, метод k -means

№ кластера	1	2	3	4	5
Обсяг кластера	28	64	21	95	37
Кількість випусників в ЗСО кластера	342	1343	888	3240	1374
Середній бал	140,605	144,401	149,869	154,209	166,357

На рисунку 4.11 наводимо секторні діаграми зі сегментацією результатів випускного тесту з математики за 2022 рік у кожному з кластерів.

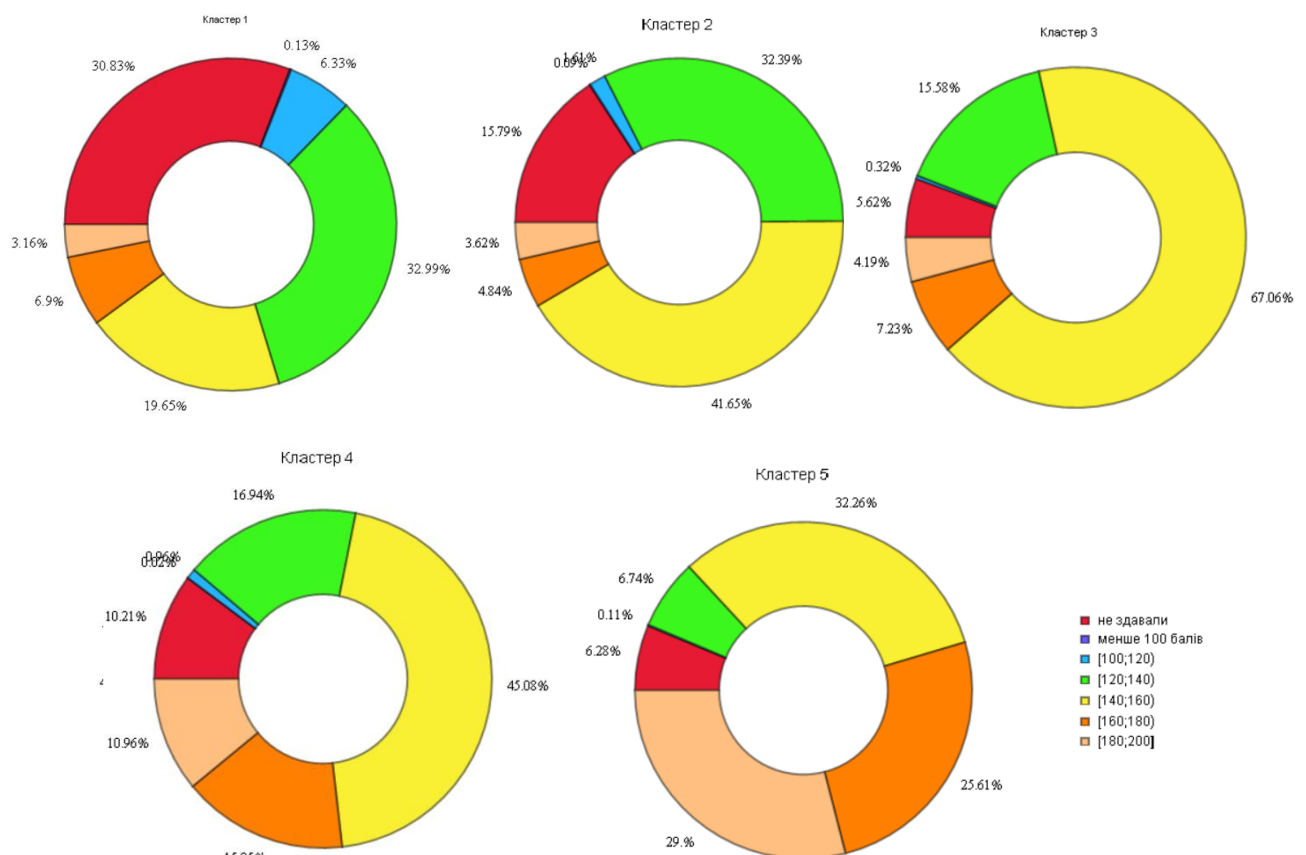


Рисунок 4.11 – Розподіл відносних часток балів НМТ 2022 року з математики у кластерах, метод k -means

За побудованими графіками оцінимо якість складання НМТ у 2022 році. Характерною рисою для всіх кластерів цього року є майже нульовий відсоток випускників, що не склали іспит з математики (менше 1% у кожній групі). Враховуючи обставини складання тесту, цю особливість можна пояснити полегшенням завдань у тесті порівняно з ЗНО, що дозволило майже всім учасникам подолати пороговий бал та скласти іспит хоча б на мінімальну оцінку. Розглянемо кластери окремо. До кластера 1 потрапили ЗСО, випускники яких продемонстрували найнижчий рівень знань з математики: 39,32% учнів отримали бали, нижче середнього. Особливістю кластера є значна частка випускників – 30,83%, що зареєструвалися для проходження НМТ, але не з'явилися на іспит в жодну із сесій. Навчальні заклади кластера 2 характеризуються кращими показниками успішності. У ньому домінують частки абітурієнтів, що отримали 120 – 140 та 140 – 160 балів – 32,39% та 41,65% відповідно. Також відзначимо, що хоча його середній бал вищий за перший кластер, але частка балів, вище середніх, у кластері 2 трохи менша за кластер 1 – 8,46% проти 10,04%. Кластер 3 – це група типових «середняків»: його було сформовано із ЗСО, де 67,05% випускників написали тест на 140 – 160 балів. Зауважимо, що хоча до нього входить на 7 ЗСО менше, ніж до кластера 1, кількість результатів у ньому більша в 2,6 разів. Це означає, що з навчальних закладах цієї групи випускалося більше учнів з потребою у складанні математики, відповідно, більшою мотивацією до вступу на технічні спеціальності у ЗВО (заклади вищої освіти). Кластер 4 має схожу до попереднього структуру з переважанням частки середніх балів – 45,06% та малим відсотком 120 – 140 балів (16,09%). Проте, він відзначається кращою якістю підготовки з математики, адже чверть випускників ЗСО цього кластера отримали бали вище середнього. Кластер 5 виділяється з-поміж інших досить високим рівнем навчальних досягнень з математики: 55% складають результати вище середнього, та третину – середні бали. При цьому у кластері зустрічаються лише поодинокі випадки складання НМТ на 100 – 120 балів, та абсолютно всі випускники склали іспит. Отже, можемо дійти висновку, що кластер 5 сформовано із ЗСО з високим рівнем викладання математики та мотивації учнів.

Продовжимо дослідження якості освіти у умовах війни, аналізуючи результати НМТ 2023 року. Зведена інформація про кожен кластер, сформований на їх основі, зокрема, обсяг і середні бали, подана у вигляді таблиці 4.5.

Таблиця 4.5 – Властивості кластерів за 2023 рік, метод *k*-means

№ кластера	1	2	3	4	5
Обсяг кластера	23	50	80	80	14
Кількість випускників в ЗСО кластера	299	1486	2997	2781	722
Середній бал	124,792	130,801	138,166	147,508	162,867

Зображуємо склад окремих кластерів графічно на кругових діаграмах (рисунк 4.12).

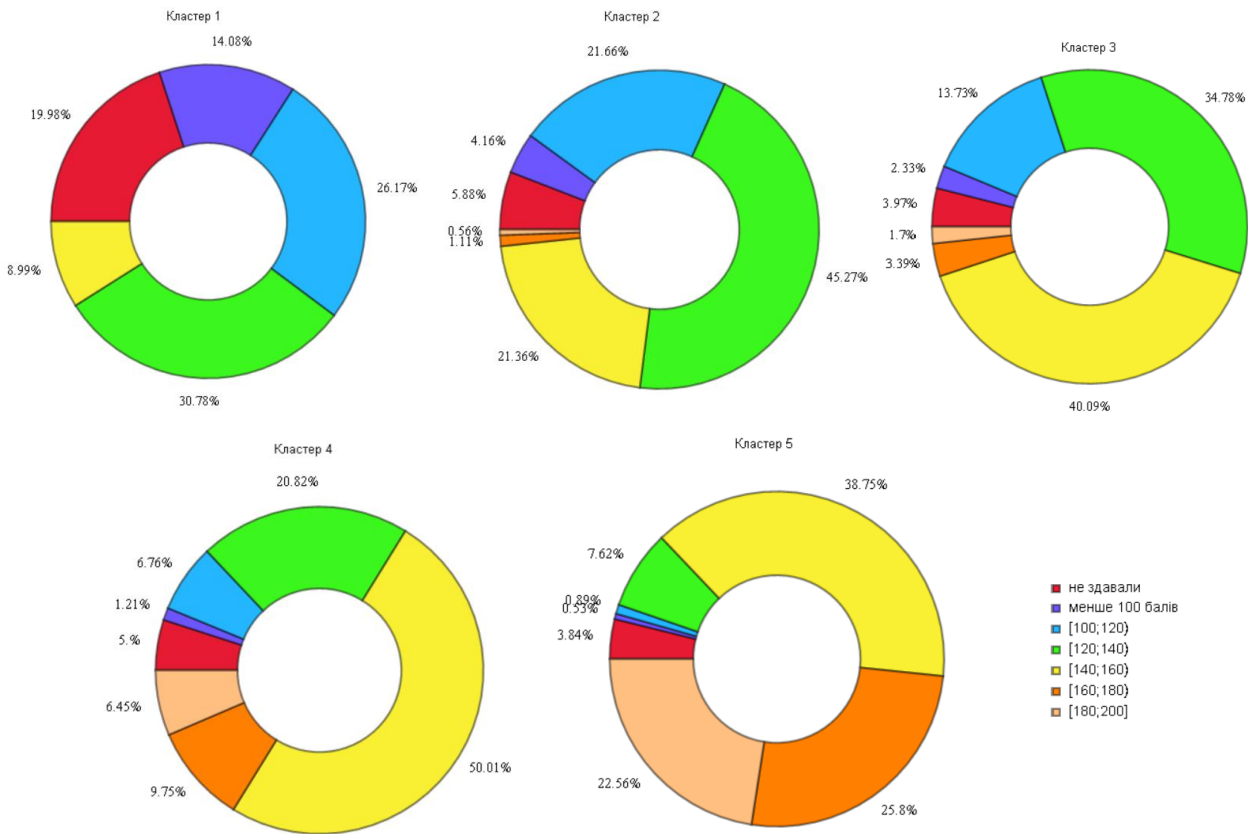


Рисунок 4.12 – Розподіл відносних часток балів НМТ 2023 року з математики у кластерах, метод *k*-means

Визначимо тенденції у кластерах ЗСО, утворених за результатами НМТ з математики за 2023 рік. Кластер 1 представляють ті ЗСО, в яких якість освітньої підготовки з математики дуже слабка. У цій групі нема жодного результату вище середнього, що може свідчити як про низький рівень підготовки на заняттях, так і про відсутність прагнення вивчати математику. Частка тих, хто не з'явився на тест, хоч і зареєструвався, в першому кластері є максимальною серед усіх – майже 20%, що також підсилює гіпотезу про низьку замотивованість випускників. Слід також розглянути припущення про те, чи не пов'язаний масовий пропуск іспиту з безпековими факторами. У кластері 2 ситуація краща: зменшилися відсотки ЗСО, випускники яких не склали НМТ (з 14,08% до 4,16%), та тих, хто отримав 100 – 120 балів (з 26,17% до 21,66%); домінує частка результатів 120 – 140 балів і майже втричі зросла частка випускників із середніми балами. Отже, у цей кластер об'єдналися ЗСО із доволі низьким рівнем викладання математики, проте не найгіршим. До того ж, у другій групі з'явилися ЗСО, де учні отримали бали вище середнього (1,73%), тобто у ній спостерігається вже вищий ступінь мотивації до навчання. До кластера 3 потрапили ЗСО з посередньою ефективністю математичної підготовки. У ньому переважає частка випускників, що написали тест на 140 – 160 балів і є лише 5% результатів вище середнього. Третину результатів становлять 120 – 140 балів, тож ЗСО у ньому відзначаються не надто сильною освітньою підготовкою з математики. До кластера 4 теж увійшли навчальні заклади, у яких дещо більше половини випускників склали іспит на 140 – 160 балів. Але загальний рівень навчання у ЗСО цієї групи помітно сильніший за попередню: відсоток учнів, що отримали 100 – 120 та 120 – 140 балів впав в 1,65 разів (з 45,49% до 27,58%), а частка тих, хто показав результати вище середнього, зросла більше ніж втричі (з 5,09% до 16,02%). Можна зробити припущення, що кластер 4 утворився із ЗСО з задовільною якістю викладання математики, де випускники старанно готувалися до тесту. За чисельністю учнів у ЗСО кластера він дещо поступається кластеру 3, в якому також знаходяться 80 навчальних закладів. Розглянемо останній кластер з найвищим середнім балом. Він сформувався із ЗСО, в яких математика може бути профільною дисципліною або, ймовір-

но, викладається на високому рівні. На це вказує переважання загальної частки результатів у 160 – 180 та 180 – 200 балів – 48,36% у поєднанні з малою кількістю (8,51%) тих, хто написав тест на бали нижче середнього. Порогові 100 балів для складання не змогли набрати менше 1% випускників кластера 5, і це лише підтверджує якість математичної освіти у ЗСО в його складі. Відмітимо, що попри найменший розмір кластера 5 за кількістю навчальних закладів, він містить результати 722 випускників, а це в 2,41 разів більше, ніж у наступному за обсягом кластері.

Розглянемо результати кластеризації за 2024 рік. Розподіл кількості ЗСО та випускників, що писали іспит, а також середній рівень їх навчальних досягнень наводяться у таблиці 4.6.

Таблиця 4.6 – Властивості кластерів за 2024 рік, метод *k*-means

№ кластера	1	2	3	4	5
Обсяг кластера	46	27	28	82	40
Кількість випускників в ЗСО кластера	377	129	95	757	274
Середній бал	123,679	129,404	135,713	137,168	156,012

На секторних графіках, поданих на рисунку 4.13, відображено структуру сформованих п'яти кластерів, отриману за показниками успішності випускників ЗСО, які увійшли до складу кожного з них. Встановимо закономірності, що спостерігаються у кожному окремому кластері.

У ЗСО кластера 1 переважає відсоток випускників (68,26%), що склали НМТ з математики на бали нижче середнього. Також жоден учень навчальних закладів кластера 1 не отримав 180 – 200 балів, і лише 2,26% отримали 160 - 180 балів. З цього можна припустити, що кластер 1 утворився із ЗСО зі слабкою академічною підготовкою з математики. Кластер 2 виділяється серед минулорічних груп ЗСО з тим же номером. Хоча він має середній бал, вищий за кластер 1, проте складається із ЗСО, в яких загальний рівень математичної підготовки навіть слабший за навчальні заклади попереднього кластера. Про це свідчить входжен-

ня до складу кластера 2 ЗСО, в яких трохи більше третини випускників не склали НМТ. Через значну частку учнів, що не подолали пороговий бал, можна припустити низьку мотивацію до навчання або відсутність належних освітніх умов. Додатковим підтвердженням цього служить найбільший з усіх п'яти кластерів відсоток тих, хто зареєструвався на НМТ, але не писав його. Цікавою особливістю кластера є домінування частки учнів, що отримали 140 – 160 балів, серед тих, хто склав іспит. Можна зробити припущення, що до кластера 2 увійшли ЗСО, в яких або підготовка з математики диференційована та залежить від окремих викладачів чи профілю закладу, або випускники, що здобули середні бали, додатково готувалися до іспиту з репетиторами чи самостійно. У кластері 3 85,56% випускників написали НМТ на бали нижче середнього, та всього лиш 0,71% – менше ніж на 100 балів, що демонструє базовий рівень освітньої підготовки з математики у навчальних закладах групи. Наявність у ньому близько восьми відсотків учнів, які отримали 160 – 180 та 180 – 200 балів, вказує на кращу у порівнянні з попередніми двома кластерами якість викладання математики. Ще однією особливістю кластера 3 є мала кількість випускників у кожному з його ЗСО, які склали НМТ, тому ймовірно, що у навчальних закладах цієї групи дуже мало випускників планують продовжувати навчання у ЗВО. У кластері 4 максимальну частку результатів (42,06%) становлять середні бали, а близько 40% – бали нижче середнього. Великий відсоток учнів, які отримали 100 – 120 та 120 – 140 балів, вказує на посередню математичну підготовку у ЗСО кластера. Водночас домінування кількості випускників, що склали тест на 140 – 160 балів, а також зростання частки тих, хто здобув 160 – 180 балів, до 6,7% (вищий показник за усі попередні кластери) робить рівень підготовки випускників ЗСО кластера 4 задовільним, надаючи йому відносну перевагу над іншими кластерами. Кластер 5 – найкращий у розглядуваному році. Майже чверть учнів ЗСО у його складі показали гідний рівень знань з математики, отримавши результати вище середнього, що є найбільшим відсотком високих балів з-поміж усіх інших кластерів, але у цілому рівень навчальних досягнень випускників у ЗСО кластера можна вважати середнім, що підтверджується суттєвою часткою результатів у 140 – 160 балів.

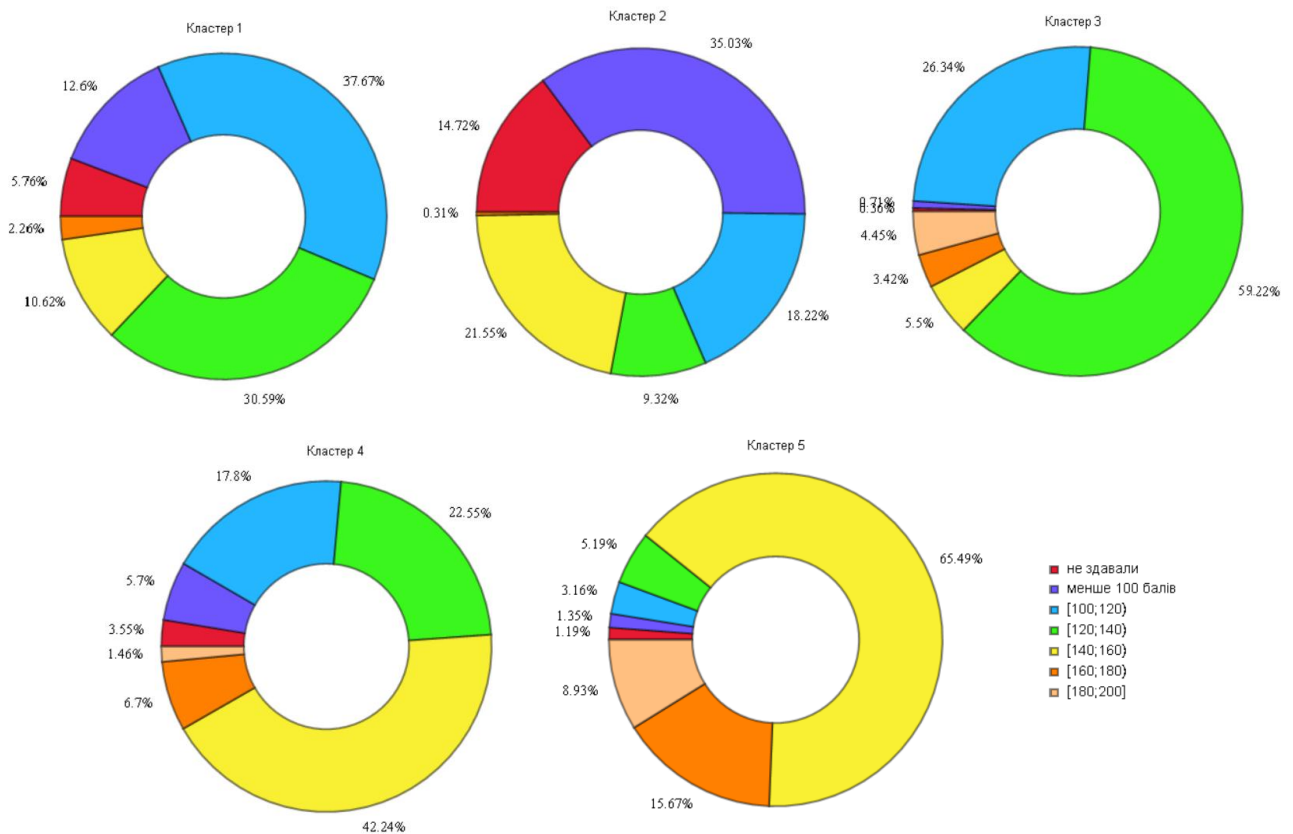


Рисунок 4.13 – Розподіл відносних часток балів НМТ 2024 року з математики у кластерах, метод k -means

Загалом можна дійти висновку про те, що у 2024 році загальний рівень складання НМТ з математики у навчальних закладах м. Харків є слабкішим за попередні роки. Результати складання іспиту сигналізують про наявність проблем, які потребують уваги та розробки новітніх підходів до їх вирішення.

Також відмітимо, що через падіння результатів НМТ у цілому деякі з утворених кластерів частково подібні один до одного за характеристиками. Таким чином, буде доречним перевірити, чи буде поділ датасету з відомостями за 2024 рік на меншу кількість кластерів, чіткішим. Для експерименту обираємо 3 кластери. Таблиця 4.7 містить інформацію про чисельність об'єктів у них та середні показники успішності випускників ЗСО, що увійшли до їх складу. Візуалізацію трьох кластерів на двовимірній площині представлено на рисунку 4.14. Для вивчення властивостей кластерів окремо розглядаємо розподіл результатів ЗНО та НМТ випускників ЗСО, з яких вони сформовані (рисунок 4.15).

Таблиця 4.7 – Властивості $k = 3$ кластерів за 2024 рік, метод k -means

№ кластера	1	2	3
Обсяг кластера	52	75	96
Кількість випускників в ЗСО кластера	307	514	811
Середній бал	128,47	131,344	145,605

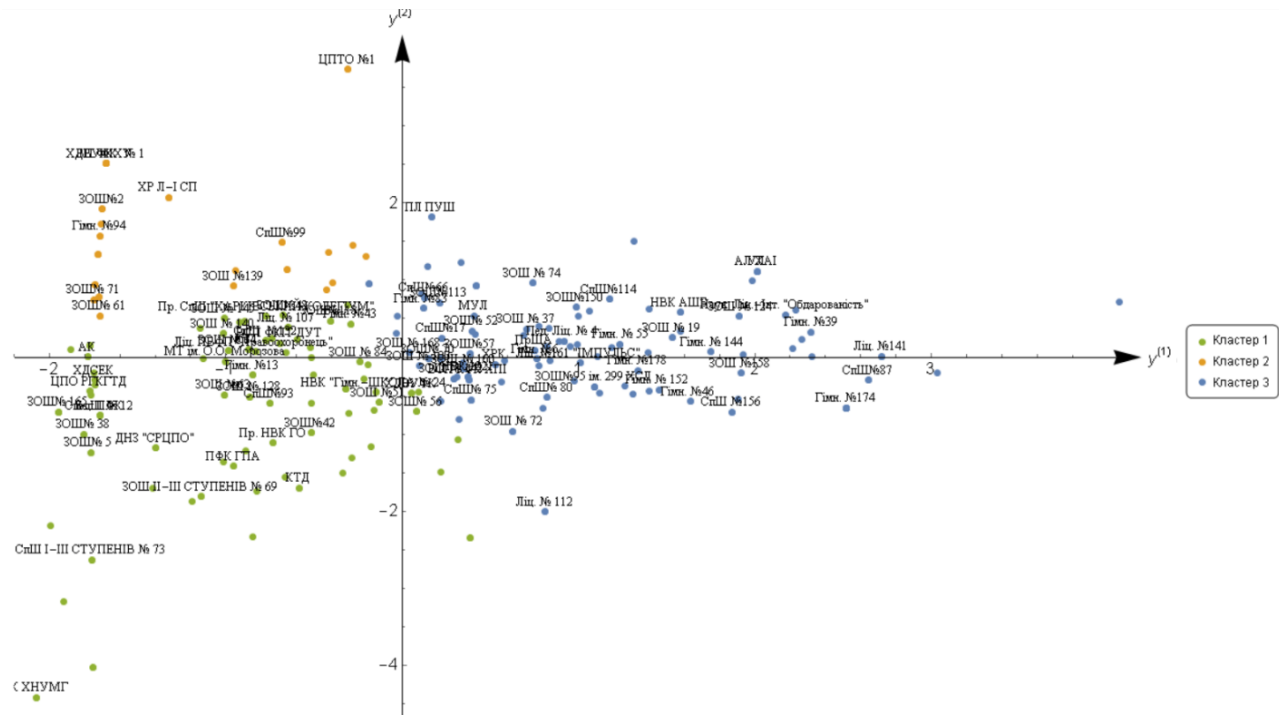


Рисунок 4.14 – Кластери ЗСО м. Харків, сформовані за результатами НМТ 2024 року з математики, метод k -means

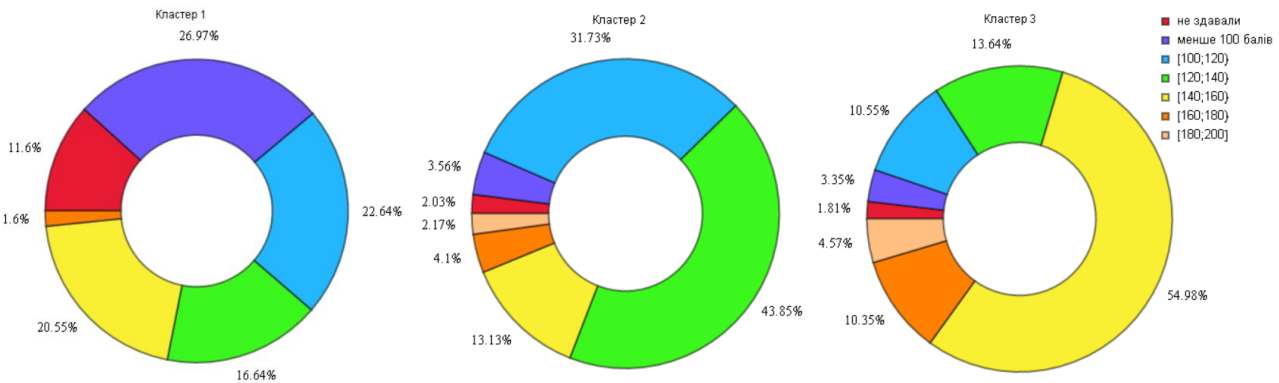


Рисунок 4.15 – Розподіл відносних часток балів НМТ 2024 року з математики у $k = 3$ кластерах, метод k -means

Опишемо властивості кластерів. До кластера 1 увійшли ЗСО, де майже всі, хто писав іспит, продемонстрували низькі чи середні результати. Домінуючою (26,97%) є частка тих, хто не зміг набрати пороговий бал, тобто учні ЗСО з кластера 1 не мають належного рівня знань. Більше 10% учнів, зареєструвавшись, не з'явилися на іспит, що може свідчити про відсутність у них мотивації або про наявність зовнішніх перешкоджаючих факторів. До кластера 2 було віднесено загалом навчальні заклади із слабким рівнем математичної підготовки, в яких більшість учнів (75,58%) отримали лише порогові бали та результати нижче середнього. Незначна частка випускників (6,27%), які написали НМТ на 160 – 180 та 180 – 200 балів, є мотивованими та зацікавленими математикою з високим рівнем підготовки. Третій кластер містить більшу частину ЗСО (43%). Відносно перших двох кластерів він є найкращим, адже близько 15% випускників ЗСО у ньому показали результати вище середнього. Така їх частка може свідчити про навчальні заклади з якісним вивченням математики. Проте загальний рівень навчальних досягнень в останній групі лишається посереднім: трохи більше половини учнів (54,98%) отримали 140 – 160 балів. Отже, спостережувані закономірності підтверджують загальне погіршення рівня підготовки з математики.

Проаналізуємо доцільність зменшення числа кластерів. Бачимо, що новий кластер 1 з рис. 4.15 зберіг майже ідентичні характеристики до кластера 2 з п'ятикластерної кластеризації (рис. 4.13), проте до нього вочевидь було приєднано кілька ЗСО з кластера 1. Новий кластер 2 об'єднав в собі значну кількість об'єктів з кластерів 1 і 3, а новий кластер 3 – з кластерів 4 та 5. Таким чином, для дослідження навчальних досягнень випускників 2024 року розбиття на 3 кластери є не менш змістовним, ніж на 5, але для порівняльного аналізу з попередніми роками, в яких поділ на 3 кластери призвів би до втрати важливих інсайтів щодо якості освіти, формування п'яти кластерів є кращим рішенням.

Далі розглянемо кластеризацію методом *k-medoids*. За рисунками 4.2 – 4.9 можемо зробити висновок про те, що геометрична структура даних з кожним роком стає все більш витягнутої форми, проте у жодному датасеті нема

групи аномалій, яка б була відокремленою в окремий кластер. У наборах даних присутні по 1 – 2 точки, що можна вважати аномаліями. Розглянемо, чи вплинуть вони на характеристики сформованих за різними методами кластерів. Наведемо результати кластерного аналізу методом k -medoids. У таблицях 4.8 – 4.11 подаємо числові характеристики сформованих угруповань об'єктів, а на рисунках 4.16 – 4.19 графічно зображуємо співвідношення результатів ЗНО та НМТ з математики, що належать різним діапазнам балів, в кожному кластері.

Таблиця 4.8 – Властивості кластерів за 2021 рік, метод k -medoids

№ кластера	1	2	3	4	5
Обсяг кластера	27	54	36	79	47
Кількість випусників в ЗСО кластера	1022	1706	1335	3043	1883
Середній бал	120,21	134,55	135,773	147,403	163,368

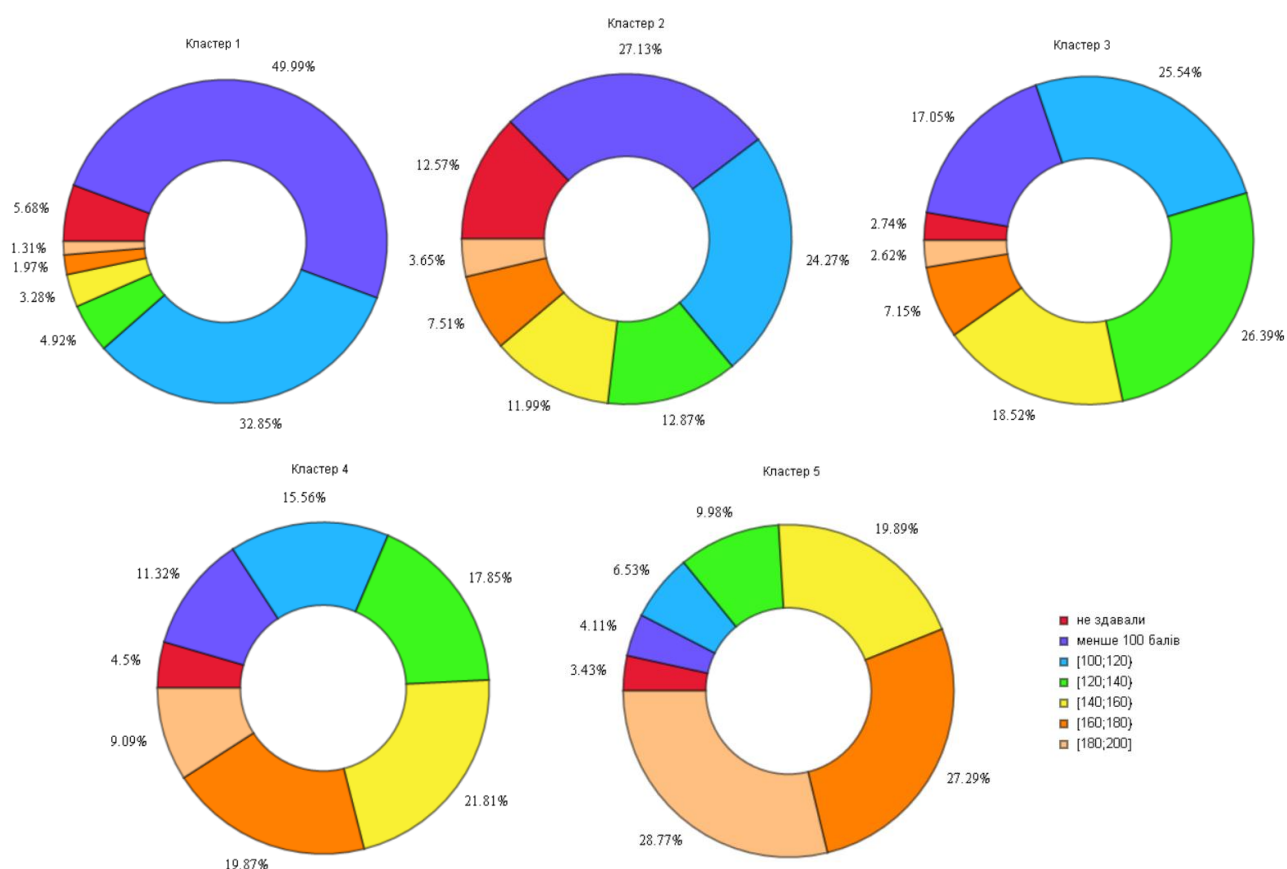
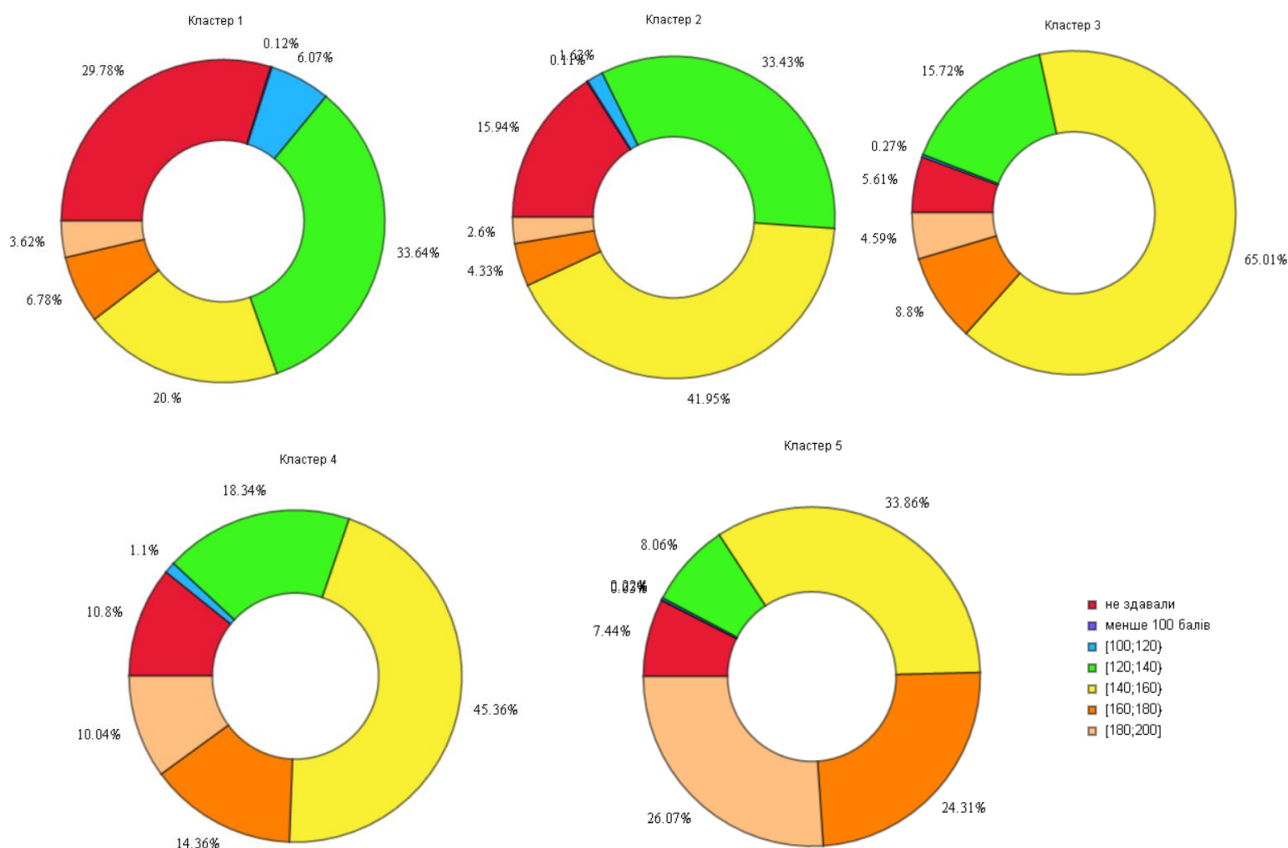


Рисунок 4.16 – Розподіл відносних часток балів ЗНО 2021 року з математики у кластерах, метод k -medoids

Таблиця 4.9 – Властивості кластерів за 2022 рік, метод *k-medoids*

№ кластера	1	2	3	4	5
Обсяг кластера	30	53	25	88	49
Кількість випускників в ЗСО кластера	376	1100	485	3033	1733
Середній бал	140,94	143,53	150,37	153,12	164,69

Рисунок 4.17 – Розподіл відносних часток балів НМТ 2022 року з математики у кластерах, метод *k-medoids*Таблиця 4.10 – Властивості кластерів за 2023 рік, метод *k-medoids*

№ кластера	1	2	3	4	5
Обсяг кластера	23	38	75	74	37
Кількість випускників в ЗСО кластера	299	1084	2880	2553	1433
Середній бал	119,25	129,44	135,89	137,49	156,15

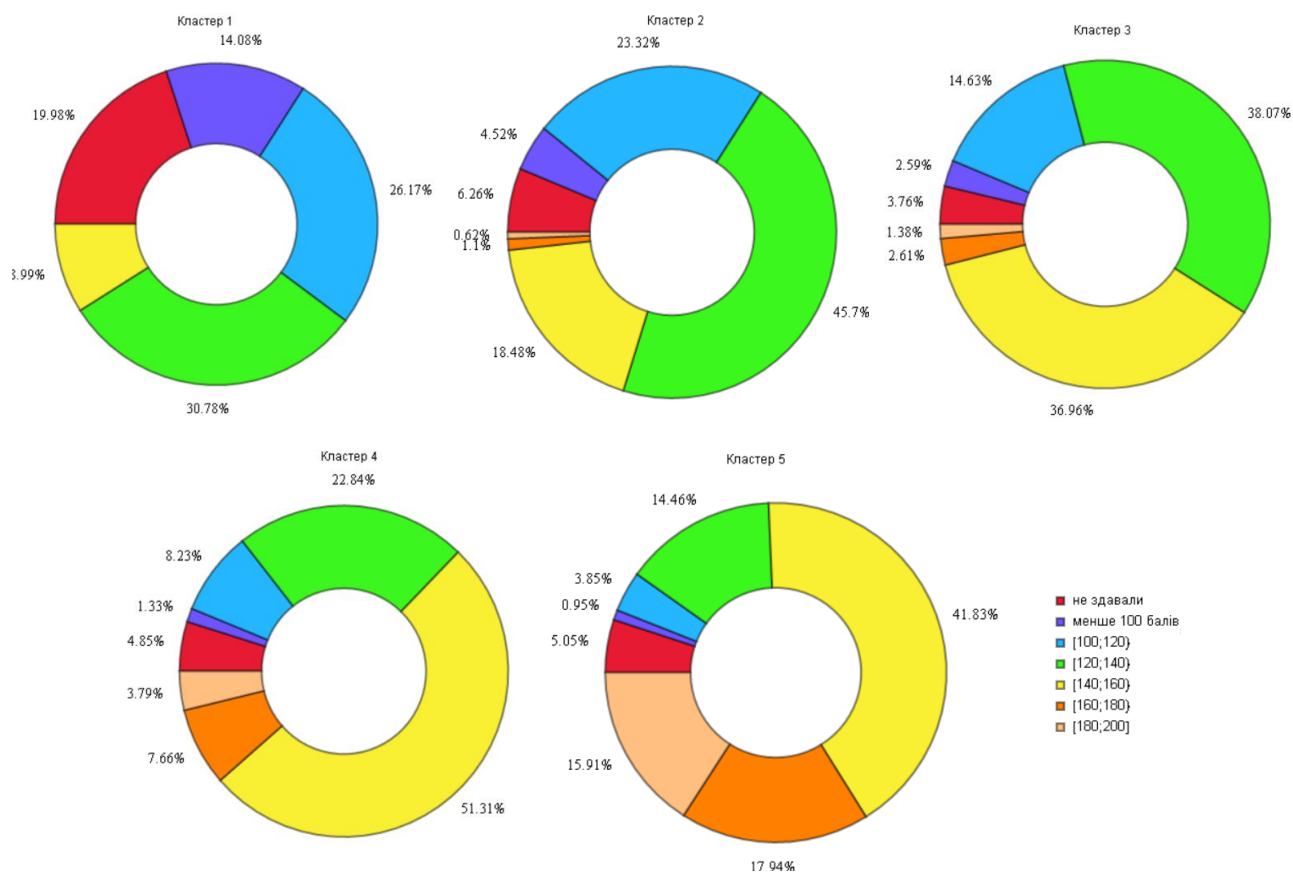


Рисунок 4.18 – Розподіл відносних часток балів НМТ 2023 року з математики у кластерах, метод *k-medoids*

Порівняємо результати кластеризації двома методами, починаючи з 2021 року. Проаналізувавши кругові діаграми на рисунках 4.9 та 4.15, можемо дійти висновку, що за допомогою обох алгоритмів сформовані кластери з аналогічними закономірностями. За таблицями 4.4 та 4.11 видно, що склад подібних за властивостями груп ЗСО відрізняється в залежності від використаного методу кластеризації, проте не критично. Найсильніше помітна різниця на прикладі кластерів з ЗСО з найвищими результатами НМТ, оскільки вони мають виражену витягнуту форму. Під впливом точок даних, що зміщені в правий верхній кут координатної площини, центроїд кластера 5 теж зміщується вправо вгору, через що менше об'єктів потрапляє до нього (35 проти 47 за методом *k-medoids*).

Для кластерів за 2022 рік спостерігається аналогічна ситуація: результати кластеризації методами *k-means* та *k-medoids*, мають подібні характеристики. Структура даних з результатами НМТ за 2022 рік схожа на 2021 рік. З цієї при-

чини кластер 5 методу k -means за 2022 рік включає в себе менше ЗСО, які є крайніми точками датасету. Збільшення середнього балу останнього кластера, отриманого за k -medoids, незважаючи на малу величину змін, підтверджує вплив його витягнутої форми на результати кластеризації.

Таблиця 4.11 – Властивості кластерів за 2024 рік, метод k -medoids

№ кластера	1	2	3	4	5
Обсяг кластера	24	26	53	78	42
Кількість випусників в ЗСО кластера	127	126	352	735	292
Середній бал	119,70	130,41	132,621	136,34	156,15

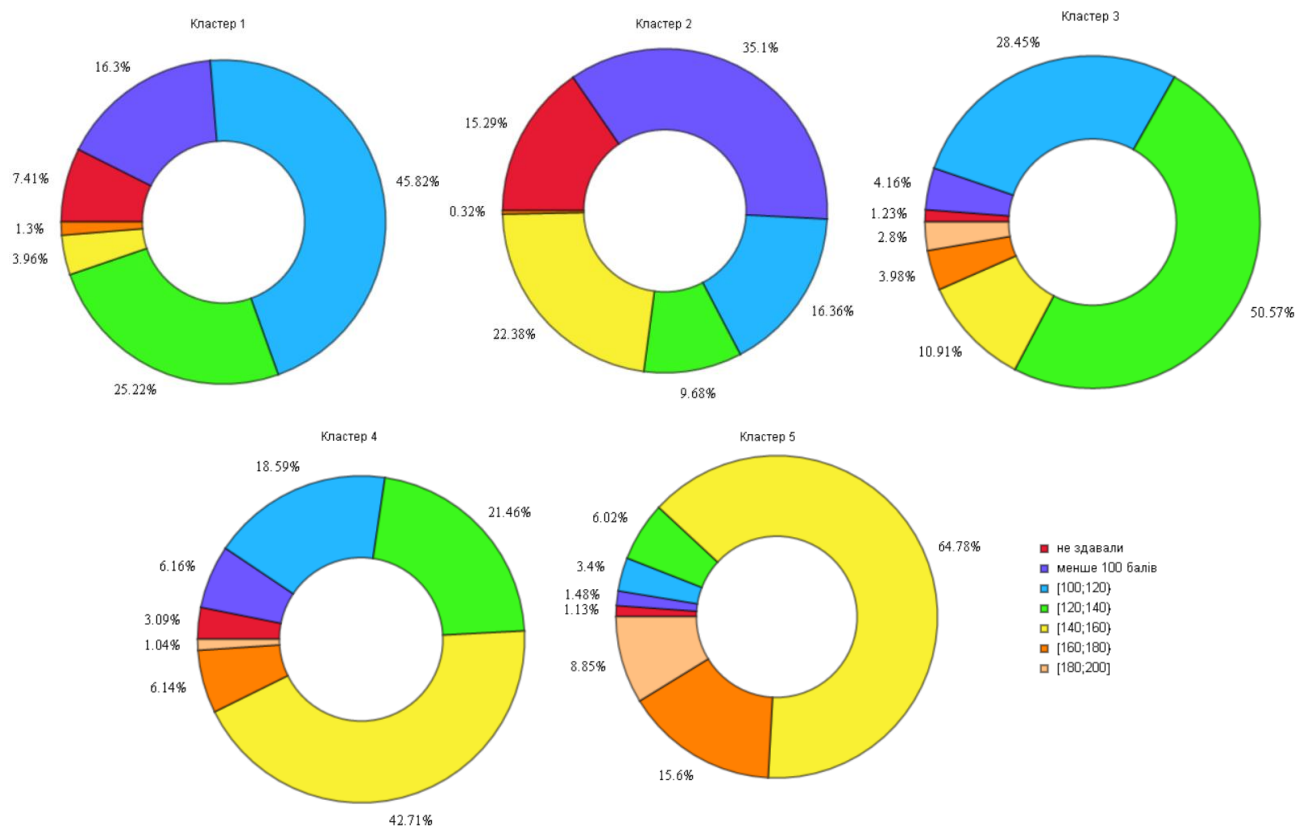


Рисунок 4.19 – Розподіл відносних часток балів НМТ 2024 року з математики у кластерах, сформованих за методом k -medoids

Датасет з результатами НМТ 2023 року має найбільш витягнуту форму з-поміж усіх інших. Різниця, що спостерігається у результатах його кластеризації за методами k -means та k -medoids, демонструє чутливість цих методів до ви-

кидів або несферичної конфігурації кластерів. У випадку використання *k-medoids* до кластера 5 увійшло майже втричі більше ЗСО, ніж за *k-means*. Алгоритм *k-means* відокремив навчальні заклади з максимальною часткою балів вище середнього, які розтягуються вздовж додатної осі координат. Водночас за *k-medoids* до кластера 5 були віднесені трохи слабші школи, з більшою часткою середнього рівня, але теж з прийнятним рівнем математичної підготовки. Через це значною мірою змінився склад кластера 4: він «послабшав» в результаті кластеризації із використанням *k-medoids*. Інші три кластери лишилися такими самими або дуже схожими через компактність точок даних.

Набір даних за 2024 рік має геометричну конфігурацію, відмінну від минулих років, що само по собі сигналізує про зміну ситуації з рівнем навчальних досягнень. Зміни, що відзначаються в залежності від застосованого методу, теж інші у порівнянні з 2021 – 2023 роками. Загалом панівні тенденції та патерни у кластерах з однаковими мітками лишилися незмінними. Разом з тим, у результаті кластеризації методом *k-medoids* ЗСО перерозподілилися між кластерами 1 і 3: у кластері 3 сильніше проявилася закономірність домінування результатів 120 – 140 балів, а до кластера 1 було додано слабші ЗСО. Витягнуте угруповання точок зліва на координатній площині (рисунки 4.6, 4.9), що сильно віддалене від інших об'єктів даних, сприяє зсуву центроїдів кластерів 1 і 3 вліво. З цієї причини до них потрапляє менше ЗСО з тих, які на декартовій площині розташовані правіше, і кластери помітно відрізняються від утворених за стійкішим до викидів методом *k-medoids*. Кластери найкращих за результатами НМТ 2024 року шкіл виявилися найстабільнішими через згрупованість точок у них.

Продовжимо аналіз кластерів, утворених методом *k-medoids*. Нагадаємо, що медоїди – це типові представники кластерів, які максимально відображають їх властивості, та визначимо ЗСО з найбільш характерним для кожного кластера рівнем підготовки з математики (таблиця 4.12).

Бачимо, що типовими представниками кластерів з найнижчими показниками НМТ виступають заклади професійно-технічної освіти. Кластери 2 та 3 представлені загальноосвітніми школами, де математика викладається на серед-

ньому рівні. Характерними представниками сильніших кластерів 4 і 5 є профільні заклади: навчально-виховні комплекси, гімназії та спеціалізовані школи, в яких вивченню математики приділяється посилена увага.

Таблиця 4.12 – Медоїди кластерів

Кластер/ Рік	2021	2022	2023	2024
1	ЦПО ІТБП	ЦПО РБАС	ЦПО РБАС	ХДСЕК
2	ЗОШ № 104	ЗОШ № 96	ВП ФКХУ	ВСП ЖКФК ХНУ
3	КБАД	АХЛ	ЗОШ № 25	ЗОШ № 139
4	СпШ № 75	Гімн. №13	ЗОШ № 28	СпШ № 62
5	НВК «Мир»	СпШ № 29	Ліц. № 141	НВК ЛП

4.3 Порівняльний аналіз кластерів для дослідження змін рівня навчальної підготовки з математики у ЗСО м. Харків за 2021 – 2024 роки

Наступним етапом дослідження є порівняльний аналіз кластерів за 2021 – 2024 роки для виявлення трендів у математичній підготовці у ЗСО м. Харків. Для порівняння кластери, сформовані методом *k*-means, ранжовано за величиною середнього балу НМТ. Дослідження динаміки змін у кластерах однакового рангу дозволяє відстежити, як трансформуються характеристики закладів з однаковим рівнем результативності: чи покращуються показники «середніх» ЗСО, чи залишаються стабільними результати лідерів, та чи поглиблюється розрив між сильними та слабкими навчальними закладами.

Порівняння результатів ЗНО та НМТ на основі утворених кластерів для кожного року проводитимемо по відношенню до показників минулого року. Це дозволить відслідковувати тенденції щорічних змін у підготовці з математики у ЗСО м. Харків, зокрема, з точки зору гнучкості до зовнішніх обставин та здатності до її покращення з урахуванням минулорічних проблем.

Почнемо з 2021 – 2022 років. За таблицями 4.4 та 4.5 бачимо загальне для

всіх кластерів зростання середнього балу, особливо помітне для груп ЗСО з нижчими показниками. Особливістю результатів НМТ за 2022 рік стала майже повна відсутність випускників, що не склали іспит (менше 1% у кожному кластері). Стрімке підвищення рівня навчальних досягнень випускників 2022 року можна пояснити сукупністю чинників. Так як 2022 рік став першим роком воєнних дій у країні, і освітня система не була адаптованою до подібних екстремних умов, то іспит проводився у спрощеному форматі. У порівнянні із ЗНО 2021 року було значно знижено кількість та складність завдань: замість 32 завдань як тестового, так і відкритого типу, учні мали відповісти лише на 20 тестових питань. Також було вдвічі знижено пороговий бал (з 10 до 5), що дозволило переважній більшості учнів скласти НМТ. До того ж до кінця лютого 2022 року, тобто більше половини одинадцятого класу, учні готувалися до повноцінного ЗНО, і це додатково посилювало їх можливості скласти НМТ на високі бали. Наведемо показники відносного приросту середніх балів і зміни часток результатів випускників у діапазонах шкали від 100 до 200 балів у таблиці 4.13.

Таблиця 4.13 – Показники для порівняльного аналізу кластерів ЗСО

м. Харків за результатами ЗНО 2021 та НМТ 2022 років

№ кластера		1	2	3	4	5
Приріст середнього балу (%)		+17,19%	+9,56%	+9,38%	+3,60%	+0,88%
Різниця часток випускників (%) з балами	менше 100	-51,43%	-27,81%	-14,8%	-10,54%	-3,43%
	100 – 120	-26,8%	-23,7%	-21,8%	-15,41%	-6,63%
	120 – 140	+28,49%	+18,55%	-10,00%	+1,19%	-3,14%
	140 – 160	+16,58%	+30,33%	+45,26%	+24,56%	+13,01%
	160 – 180	+5,02%	-1,36%	-1,87%	-5,41%	-2,01%
	180 – 200	+1,74%	-1,32%	+0,53%	+0,52%	-2,73%

Проведемо огляд структури кластерів з однаковим рангом та порівняємо їх. Найслабший кластер 1 у 2022 році характеризується набагато вищими балами, ніж у 2021 році. Зокрема, у 2021 році половина учнів не змогла скласти

ЗНО, та ще третина отримала 100 – 120 балів, але у 2022 році кількість тих, хто написав тест на 100 – 120 балів, впала на 26,8%. Домінуюча частка учнів у 2022 році показала результати нижче середнього, але у межах 120 – 140 балів. Разом з тим, відсоток отриманих балів вище середнього збільшився порівняно з 2021 роком з 3,3% до 10,06%. Причинами цього зростання можуть бути як спрощення тестових завдань так і можливість підвищення рівня викладання математики у ЗСО кластера 1. Слід зазначити, що у 2022 році 30% учнів не писали іспит взагалі, і це теж вплинуло на кінцевий розподіл результатів у кластері.

У кластері 2 за 2022 рік теж спостерігається краща ситуація, ніж у попередньому році. Якщо у 2021 році переважна частина випускників не склали ЗНО або написали його на 100 – 120 балів, то у наступному році більшість учнів отримали хоча б 120 – 140 або 140 – 160 балів. Кількість результатів вище середнього не зазнала кардинальних змін. Аналогічні тенденції помітні і у ЗСО кластера 3. У 2022 році у ньому більш, ніж на 40%, зросла частка учнів, які отримали середні бали і склали домінуючу підмножину кластера. Відсоток випускників, що написали НМТ на 100 – 120 балів або ж взагалі не склали його, теж помітно зменшилася. Проте чисельність учнів з високими навчальними досягненнями лишилася на попередньому рівні. На основі зазначених фактів можна дійти висновку, що значний вплив на покращення результатів іспиту мала саме зміна його формату.

Динаміка змін у кластері 4 за 2022 рік відрізняється від попередніх випадків. Хоч середній бал у ньому зріс на 3,6%, проте результати НМТ погіршилися. Якщо у 2021 році частка тих, хто отримав 140 – 160 та 160 – 180 балів була приблизно однаковою (20 – 21%) з переважанням вищих результатів, то у 2022 році вдвічі збільшилася кількість учнів із середніми балами, і на 5,41% знизився відсоток тих, хто отримав вищі бали. З позитивних трендів спостерігається лише падіння частки тих, хто не склав іспит або отримав 100 – 120 балів, що може бути пов'язане зі зниженням складності тесту. Таким чином, рівень підготовки у ЗСО кластера 4 у 2022 році міг навіть трохи погіршитися, або ж лишитися таким же, що і у минулому році. На зменшення відсотка учнів, що написали НМТ

на 160 – 180 балів, могли вплинути фактори стресу від умов складання та загальної ситуації у країні.

Кластер лідерів під номером 5 відзначається відносною стабільністю. У ньому за рахунок спрощення завдань абсолютно всі склали НМТ, лише 0,11% отримали 100 – 120 балів та дещо менша частка – 120 – 140 балів. Проте, розподіл балів вище середнього в цілому не змінився у порівнянні з 2021 роком. Як особливість можна виділити збільшення частки учнів, що не складали НМТ попри реєстрацію, майже втричі. Це може бути пов'язане з рішенням деяких випускників вступати в іноземні ЗВО без складання НМТ.

Дослідимо питання стабільності кластерів. На рисунку 4.20 побудовано стовпчасті діаграми, які для кожного з кластерів у 2022 році показують кількість переходів об'єктів до нього з кожного кластера 2021 року, у відсотках.

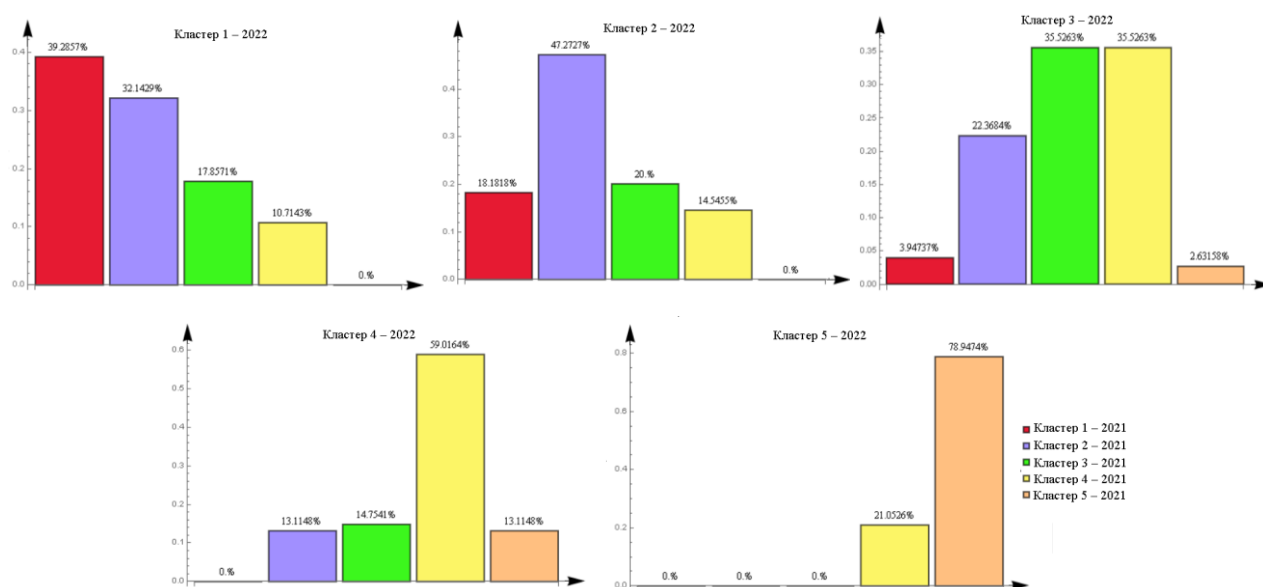


Рисунок 4.20 – Відносна кількість переходів ЗСО між кластерами за 2021 – 2022 роки

За графіками спостерігаються кілька цікавих закономірностей у «міграціях» ЗСО між кластерами. Кластер 5 з найсильніших за підготовкою до математики ЗСО виявився найстабільнішим. Це посилює гіпотезу про те, що загальне підвищення результатів НМТ пов'язано не з різким покращенням якості викла-

дання. Найбільший відсоток переходів зафіксовано до кластера 1, навіть з передостанньої за середнім балом групи ЗСО, яка вважалася однією з найкращих у 2021 році. З 2021 року у цій множині лишилися менше половини ЗСО, тож середній бал кластера виріс не за рахунок значних позитивних змін у навчальній програмі закладів з найгіршими результатами ЗНО. Кластери 2, 3, 4 мають відносно стабільний склад. Претендентами на суттєве посилення саме підходу до викладання математики виступають 3,94% ЗСО, які з кластера 1 «переїхали» до кластера 3 та деякі ЗСО з 13%, що перейшли до кластера 4 з кластера 2. Слід також визначити ЗСО, що увійшли до кластера 3 у 2022 (2,61%), а ще в минулому році були віднесені до групи лідерів, та вивчити причини погіршення результатів у них з метою їх усунення.

Надалі оцінімо, як змінилися результати іспиту у 2023 році у порівнянні з 2022 роком, та проаналізуємо рівень адаптації як учнів, так і освітян до воєнних реалій. Загалом рівень навчальних досягнень випускників ЗСО м. Харків став нижчим, про що свідчить зменшення середніх балів у всіх кластерах. Серед можливих причин погіршення результатів в першу чергу слід виділити негативний вплив воєнних дій на ефективність підготовки до НМТ протягом року. Хронічний стрес від повітряних тривог, осінньо-зимові блекауты сильно послабляли когнітивну продуктивність випускників. Також викладання в деяких ЗСО могло погіршитися через звільнення сильних вчителів з математики. Незважаючи на оптимізацію умов проведення тесту, багато минулорічних проблем лишилися невирішеними. У 2023 році збільшилася частота тривалих повітряних тривог, після яких написання іспиту зупинялося, через що випускники мали перескладати його в додаткову сесію. Хоча кількість учасників додаткової сесії була нижчою, ніж у 2022 році, проте багато учнів, спираючись на досвід минулого року, поспішали написати тест до першої повітряної тривоги, через що ймовірність помилок зросла. Також у 2023 році частіше фіксувалися випадки технічних збоїв у комп'ютерних центрах проведення НМТ, що теж спричинило певний відсоток зниження результатів. Структура та складність тесту, пороговий бал лишилися приблизно такими самими, як і 2022 році [32].

Динаміку змін у рівні успішності учнів з 2022 по 2023 рік по кластерах представлено у таблиці 4.14.

Таблиця 4.14 – Показники для порівняльного аналізу кластерів ЗСО
м. Харків за результатами НМТ 2022 та 2023 років

№ кластера		1	2	3	4	5
Приріст середнього балу (%)		-11,24%	-9,42%	-7,80%	-4,35%	-2,10%
Приріст відносної кількості випускників (%) з балами	менше 100	+13,95%	+4,07%	+2,33%	+1,19%	+0,53%
	100 – 120	+19,84%	+20,05%	+13,41%	+5,9%	+0,78%
	120 – 140	-2,21%	+12,88%	+19,2%	+3,88%	+0,88%
	140 – 160	-10,66%	-20,29%	-26,97%	+4,93%	+6,49%
	160 – 180	-6,9%	-3,73%	-3,84%	-6,1%	+0,19%
	180 – 200	-3,16%	-3,06%	-2,49%	-4,51%	-6,44%

Дослідимо розподіл результатів по кластерах окремо. Кластер 1 з найнижчим рівнем освітньої підготовки з математики, у 2023 році сформувався із ЗСО, де жоден випускник не продемонстрував результатів вище середнього. У порівнянні з 2022 роком дуже сильно виросли частки учнів, що не склали НМТ (14,08% проти 0,13%), або отримали 100 – 120 балів (26,17% проти 6,33%). Падіння балів у кластерах може бути спричиненим як проблемами з ефективністю математичної підготовки, так і з мотивацією та силами в учнів.

У кластері 2 за 2023 рік теж помітні тенденції до погіршення рівня навчальних досягнень випускників ЗСО м. Харків. На відміну від 2022 року, переважна частка учнів склала НМТ на 120 – 140 балів, а не 140 – 160, та значно зріс відсоток тих, хто набрав 100 – 120 балів: з 1,61% до 21,66%. Лише окремі учні (менше 2%) отримали результати вище середнього, можливо завдяки сукупності індивідуальних факторів та особливостей. Скоріш за все, якість навчання у ЗСО кластера 2 навіть погіршилася з 2021 та 2022 років, бо чисельність учнів з високими результатами відчутно знизилася, а з нижчими за середні – зросла.

ЗСО кластера 3 в обох роках лишилися типовими «середняками» з переважанням частки учнів, що набрали 140 – 160 балів. Загальний рівень навчаль-

них досягнень за кластером впав: результати нижче середнього у 2023 році показала майже половина випускників ЗСО цієї групи, натомість у 2022 році низьку успішність мали менше 16% учнів. Отже викладання математики у навчальних закладах кластера 3 не зазнало помітних змін у кращий бік у 2023 році.

У ЗСО кластера 4, який за показником середнього балу є одним з кращих, теж бачимо погіршення результатів. У 2023 році очевидними є наступні тренди кластера: зменшення чисельності випускників з результатами НМТ вище середнього; натомість, значне зростання кількості тих, хто отримав 100 – 120 балів, а також підвищення, хоч і не надто велике, частки учнів зі 120 – 140 та 140 – 160 балами. Виходячи із зазначених фактів, у порівнянні з 2022 роком якості освітньої підготовки до НМТ з математики у навчальних закладах цієї групи як мінімум лишилася такою самою або ж впала. Також вагому роль у зниженні успішності випускників ЗСО можуть відігравати індивідуальні особливості учнів, зокрема, ступінь стресостійкості.

«Зірковий» кластер 5 із ЗСО з найкращою математичною підготовкою у 2023 році теж послабив свої позиції. Хоч у ньому стабільно чверть випускників отримали 160 – 180 балів, як у минулому році, проте відсоток тих, хто склав НМТ на 180 – 200 балів, впав на 6,44%. Все більш виразним стає «середній» рівень навчання математиці навіть у сильних ЗСО, адже частка результатів у 140 – 160 балів складає половину кластера 5. Також з'явився малий відсоток учнів, які не склали іспит.

Додатково слід зазначити загальну для всіх кластерів тенденцію до зменшення частки випускників, що зареєструвалися на НМТ, але не складали його з певних причин, та збільшення чисельності випускників, що писали НМТ з математики у 2023 році (8632 проти 6727 у 2022). Спостережувана динаміка є позитивною та може свідчити про адаптацію випускників до кризових умов та прагнення отримувати освіту в Україні попри воєнні дії.

Перевіримо, наскільки кластери лишилися у 2023 році стабільними за своїм складом у порівнянні з 2022 роком. Перерозподіл ЗСО між кластерами представлено на рисунку 4.21.

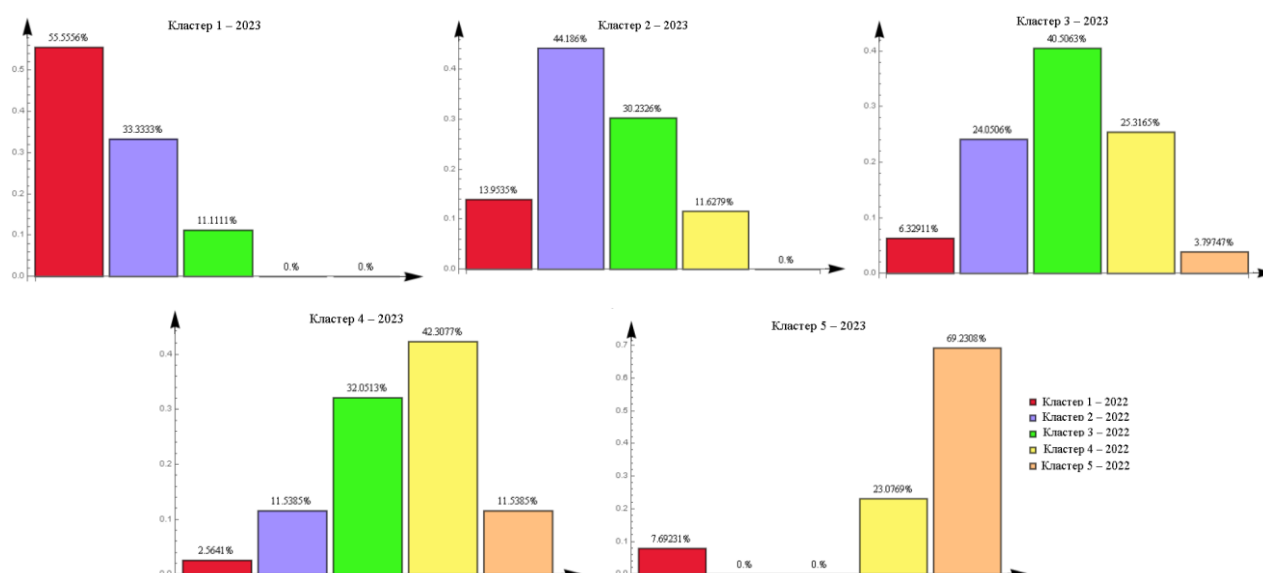


Рисунок 4.21 – Відносна кількість переходів ЗСО між кластерами у 2022 – 2023 роках

У 2023 році кластери характеризуються більшою стабільністю, ніж у 2022 році. У кожному з них домінує частка ЗСО з кластерів, що мали такий самий ранг за середнім балом у минулому році. Обмін об'єктами між сусідніми за номером кластерами не є цікавим з точки зору помітного покращення чи погіршення рівня математичної підготовки у них. Наявність у кожній групі ЗСО з кластера 1 за 2022 рік пояснюється, перш за все, штучним підняттям його рівня під впливом зовнішніх факторів. Окремо варто дослідити 7,69% навчальних закладів, що з аутсайдерів стали лідерами, мігрувавши з першого кластера до останнього, бо серед них можуть бути ті, в яких дійсно зріс рівень освітньої підготовки. Також треба розглянути 3,8% ЗСО, що потрапили до кластера посередніх з множини 5 найкращих навчальних закладів міста.

Перейдемо до аналізу змін за результатами НМТ 2023 – 2024 років. Якщо оцінювати середні бали для кластерів за два роки, то спостерігаємо їх падіння, причому з більшою стрімкістю у групах сильніших ЗСО. Враховуючи ситуацію в країні у 2024 році, можемо припустити, що чинниками, під дією яких загальний рівень результатів НМТ з математики знижується з кожним наступним роком, виступають: накопичений стрес в учнів та педагогів; відсутність стабільних занять через повітряні тривоги та періодичні проблеми з електропостачан-

ням і зв’язком. Також слід врахувати, що кожне наступне покоління випускників довше навчається в дистанційному форматі, відповідно, має нижчий рівень фундаментальних знань. Математика – це одна з найскладніших навчальних дисциплін, і для її досягнення очні заняття вочевидь є ефективнішими за онлайн-уроки. Додавання додаткового предмета до переліку обов’язкових для складання, скоріш за все, призвело до зростання психоемоційного та когнітивного навантаження на випускників, що, своєю чергою, могло негативно позначитися на показниках їхньої академічної успішності. Врахуємо ще наступний факт: НМТ у 2023 році складало 8249 людей, а у 2024 – лише 1632. Ймовірно, суттєва кількість випускників виїхала за кордон, і потенційно серед них була досить значна частка саме тих учнів, що могли б скласти НМТ на високі бали. Зведену інформацію з числовими показниками зміни успішності учнів у ЗСО різних кластерів наведено у таблиці 4.15.

Таблиця 4.15 – Показники для порівняльного аналізу кластерів ЗСО
м. Харків за результатами НМТ 2023 та 2024 років

№ кластера		1	2	3	4	5
Приріст середнього балу (%)		–0,9%	–1,07%	–1,78%	–7,01%	–4,39%
Приріст відносної кількості випускників (%) з балами	менше 100	–1,48%	+30,87%	–1,62%	+4,49%	+0,83%
	100 – 120	+11,5%	–3,44%	+12,61%	+11,04%	+2,27%
	120 – 140	–0,19%	–35,95%	+28,9%	+1,3%	–2,43%
	140 – 160	+1,63%	+0,19%	–34,59%	–7,77%	+26,74%
	160 – 180	+2,26%	–0,8%	+0,03%	–3,05%	–10,13%
	180 – 200	0,00%	–0,56%	+2,75%	–4,99%	–13,63%

Співставимо кластери з однаковими номерами. Як і у 2023 році, випускники ЗСО кластера 1 показали незадовільні результати з переважанням частки балів нижче середнього (близько 68%). Малий відсоток учнів у кластері 1, що здобули 160 – 180 балів, склали сильні учні чи такі, хто займався з репетиторами. ЗСО цього кластера у порівнянні з 2023 не відзначилися помітним покращенням.

щенням математичної підготовки. Зазначимо, що у 2024 році частка зареєстрованих абітурієнтів, які не складали тест, зменшилася майже в 4 рази у порівнянні з 2023 роком, що також пов'язано із впливом зовнішніх факторів.

Попри незначне зниження середнього балу кластера 2 у 2024 році (на 1,07%) розподіл балів у ньому вказує на суттєве погіршення рівня знань з математики випускників ЗСО кластера. У 2023 році у ньому домінувала частка учнів, що отримали 120 – 140 балів, а у 2024 році панівним є відсоток випускників, які не набрали менше п'яти порогових балів, не склавши іспит. Подібно до 2023 року, п'ята частина учнів написала НМТ на середні бали, та мізерний відсоток (0,31%) – на 160 – 180 балів. Як зазначалося вище, випадки, в яких учні отримали 0 балів, з математичної точки зору є аномаліями датасету, і можуть сильно скосити вліво розподіл балів всередині кожного кластера та його середній бал. Для збереження більшої стійкості до викидів середні бали, наведені у таблицях, розраховані лише на основі результатів випускників, що склали тест. Саме тому середній бал кластера 2 є кращим за кластер 1 у 2024 році, але вочевидь, до нього увійшли ЗСО м. Харків, випускники яких продемонстрували слабший рівень знань з математики. Зростання частки випускників, що не писали тест, хоча реєструвалися на нього, майже втричі (до 14,72%) може свідчити про падіння мотивації в учнів ЗСО кластера 2 до складання НМТ.

Кластер 3 у 2024 році сформувався із ЗСО, де рівень академічної підготовки з математики явно нижчий, ніж був у 2023 році. На це вказує зростання у ньому частки результатів нижче середнього у 1,86 разів. При цьому відсоток учнів, що здобули 160 – 180 та 180 – 200 балів, збільшився та майже всі склали іспит (окрім 0,31% учнів). Отже, якість математичної підготовки у кластері 3 у 2024 році, можливо, стала слабшою, оскільки кількість тих, хто отримав середні бали, зменшилася аж у 8 разів. Разом з тим вона стала базовою, тобто дозволила майже всім скласти іспит, хоч і на нижчі бали, ніж у 2023 році.

Кластер 4 у 2024 році за розподілом балів подібний до кластера 3 у 2023 році. До цього кластера увійшли не тільки навчальні заклади з прийнятною якістю математичної підготовки, а і більш посередні. Якщо порівнювати його з

кластером 4 у 2023 році, то спостерігається сталий тренд на збільшення частки результатів нижче середнього та зменшення відсотка учнів, що отримали вищі за середні бали. При цьому 140 – 160 балів, як і 2023 році, отримали переважна більшість випускників, проте на 7,77% менша, ніж минулого року.

Кластер 5 з найкращими ЗСО за ступенем математичної підготовки до НМТ у 2024 році теж вирізняється сильним послабленням загального рівня результатів. У 2023 році майже половина випускників ЗСО цієї групи отримала бали вище середнього, а у 2024 – тільки близько чверті з них показали подібні результати. При цьому переважати стали учні із 140 – 160 балами за тест. Більшість ЗСО кластера 5 утримують лідерські позиції від року до року. Хоча ефективність математичної підготовки у них під впливом зовнішніх кризових обставин могла погіршитися, проте ймовірніше, що результати за кластером впали через різке зменшення чисельності учнів, які писали тест: з 672 до 266. Серед більшої кількості випускників могло б бути більше потенційних відмінників, що могли б отримати 160 – 180 та 180 – 200 балів.

Проаналізуємо, які переміщення відбулися між кластерами за 2023 та 2024 роки (рисунок 4.22).

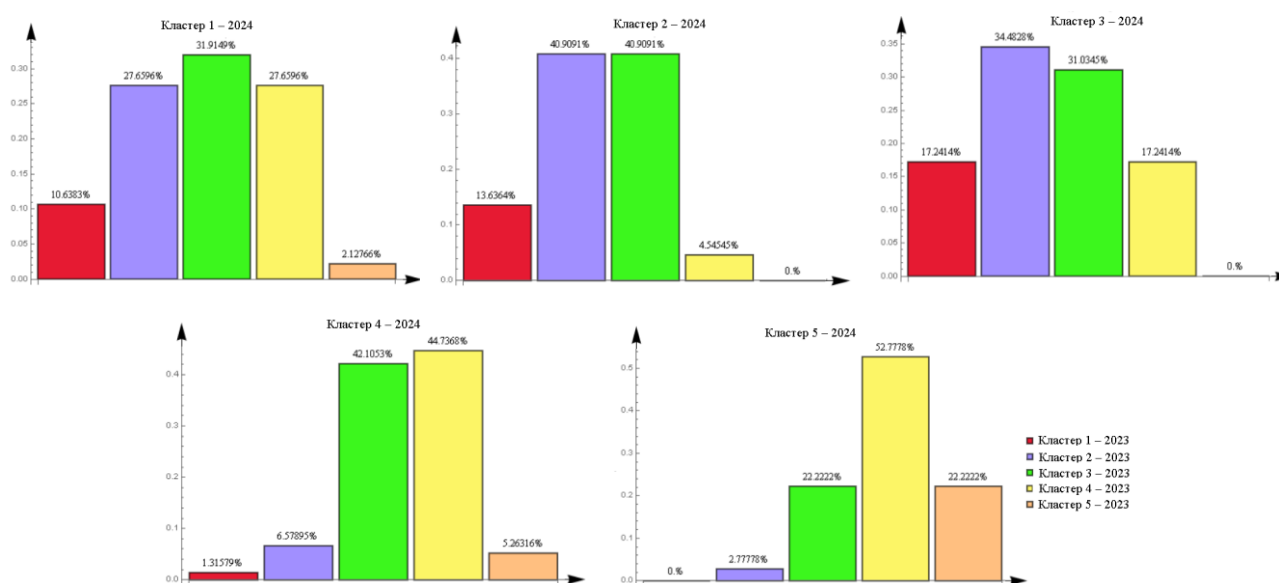


Рисунок 4.22 – Відносна кількість переходів ЗСО між кластерами у 2023 – 2024 роках

Для 2024 року очевидним є погіршення навчальних досягнень випускників ЗСО кластерів 1 – 3. До першого (найгіршого) кластера потрапила майже третина навчальних закладів, які у 2023 році вважалися дуже сильними або характеризувалися прийнятною якістю математичної підготовки. Навіть кластер 5 більш ніж на половину утворено із навчальних закладів, що раніше відносилися до кластера 4 з нижчими результатами. Це може бути сигналом значного погіршення якості математичної підготовки у ЗСО м. Харків.

4.4 Оцінювання якості кластеризації за внутрішніми мірами валідності кластерів

Наостанок проаналізуємо якість кластеризації з використанням спеціальних метрик. У ході дослідження було виявлено, що виділені кластери дійсно мають практично значущі відмінності та сформовані за змістовними закономірностями. Перевіримо їх валідність за допомогою обраних внутрішніх мір оцінки якості: силуетного коефіцієнту, індексів Калінскі – Харабаса та Данна. Значення цих показників, розраховані для результатів кластерного аналізу за кожен рік, наведені у таблиці 4.16.

Таблиця 4.16 – Метрики якості кластеризації даних, *k*-means

Метрика/Рік	2021	2022	2023	2024
Силуетний коефіцієнт	–0,07811	–0,06451	–0,08165	–0,10663
Індекс Калінскі – Харабаса	214,44795	177,58030	212,27709	179,65817
Індекс Данна	0,01134	0,01371	0,01989	0,02670

Розбиття ЗСО м. Харків за результатами ЗНО та НМТ 2021 – 2024 років найвищу оцінку якості отримало за метрикою Калінскі – Харабаса. Близькі до нуля значення коефіцієнту силуету та індексу Данна свідчать про локальні проблеми у структурі датасета, через які його складно розділити на однозначні чіт-

кі підмножини, які не перекривали б одна одну. Таким чином, виник конфлікт метрик якості кластеризації, проте він не є прямим доказом низької ефективності та некоректності проведеного аналізу. Отримана ситуація є поширеною під час обробки практичних даних та має логічне математичне обґрунтування. Усі наведені у таблиці 4.16 міри характеризують якість кластеризації, але роблять це за різними аспектами. Індекс Калінскі – Харабаса оцінює глобальну дисперсію даних. Його значення в приблизних межах від 175 до 215 є відносно низькими та свідчать про розмитість меж утворених кластерів. Це відповідає реальній структурі датасету із ЗСО м. Харків, адже виражені відмінності рівня навчальних досягнень спостерігаються для невеликої кількості профільних або найслабших ЗСО. Інші навчальні заклади теж відрізняються між собою за низькою характеристик, які відзначаються в трьох з п'яти кластерів, проте не настільки очевидно. За індексом Калінскі – Харабаса найбільш очевидні тренди виділяються у результатах ЗНО за 2021 рік та НМТ 2023 року.

У розрахунковій формулі індексу Данна акцент робиться на межах точках різних кластерів та відстані між ними. За рисунками 4.2 – 4.9 бачимо, що витягнута форма окремих частин даних та наявність викидів у деяких з них спричиняє збільшення діаметра кластерів. При існуванні у сформованих групах хоч одного об'єкта, більш віддаленого від інших, індекс Данна помітно знижується через його чутливість до відстані саме між їх крайніми точками. Аналізуючи представлені у таблиці 4.16 величини, бачимо, що абсолютно всі датасети через більшу чи меншу витягнутість окремих кластерів характеризуються низьким значенням індексу Данна (0,01 – 0,02), незважаючи на існуючі, значущі на практиці, закономірності у них.

Силуетний коефіцієнт обчислюється для кожного об'єкта кластера і підкреслює його індивідуальну приналежність до певного кластера. На рисунках 4.2 – 4.9 прослідковується тенденція до скупчення значної кількості точок з різних за характеристиками кластерів поблизу один одного. Відповідно, відстані між точками сусідніх кластерів є настільки малими, що в сенсі цієї метрики вони вважаються спірними, віднесеними з великою ймовірністю до «чужого» класте-

ра. Середні значення силуетних коефіцієнтів для них є від’ємними, але малими за модулем. Слід зазначити, що силует кластерів ЗСО, отриманих за навчальними досягненнями випускників 2024 року, є найменшим за модулем, що вказує на максимальну кількість у ньому близько розташованих точок та найслабшу відмінність всіх ЗСО м. Харків за рівнем підготовки з математики. Отриманий результат коректно співвідноситься з максимальним індексом Данна, бо викидів у цьому датасеті дійсно найменше. Одне з невисоких значень індексу Калінскі – Харабаса для даних 2024 року додатково підтверджує сильну скупченість даних. Для порівняння методів k -means та k -medoids обчислимо внутрішні міри для кластерів, сформованих за другим з них (таблиця 4.17).

Таблиця 4.17 – Метрики якості кластеризації даних, k -medoids

Метрика/Рік	2021	2022	2023	2024
Силуетний коефіцієнт	–0,10038	–0,09663	–0,10051	–0,09816
Індекс Калінскі – Харабаса	188,48561	144,46332	150,82834	137,67759
Індекс Данна	0,00807	0,00351	0,00580	0,01455

За таблицею 4.17 бачимо, що значення всіх розрахованих метрик вказують на погану роздільність та некомпактність кластерів, сформованих із застосуванням методу k -medoids. Обчислені показники свідчать про те, що якість кластеризації за алгоритмом k -means виявилася аналогічною або навіть трохи кращою, проте все одно низькою. Наприклад, індекс Калінскі – Харабаса вказує на те, що розбиття, утворене k -medoids, очевидно є менш чітким, ніж отримане другим методом. Воно характеризується як меншою міжкластерною, так і більшою внутрішньокластерною дисперсіями за рахунок з’єднання більш витягнутих кластерів з новими, більш скупченими, точками. Значення індексу Данна також зменшилися у порівнянні з таблицею 4.16, проте на малу величину. Менші значення індексу Данна для кластерів, утворених за результатами 2021, 2022 та 2023 років, можна пояснити впливом більш віддалених точок, що відповідають найслабшим чи найсильнішим ЗСО. Через їх приєднання до множи-

ни об'єктів, розташованих компактніше, діаметр кластерів зріс, тож індекс Данна став меншим. Водночас, максимальне значення для 2024 року підтверджує факт найбільшої компактності зразків даних, тобто, подібності ЗСО за результатами НМТ. Аналогічна ситуація спостерігається і для силуетних коефіцієнтів: через кращу стійкість k -medoids до викидів або віддалених точок датасету їх не було відокремлено в інші групи, подібно до методу k -means. Таким чином, кластери стали більш скупченими між собою, і це посприяло зниженню силуетних коефіцієнтів. Для 2024 року, так само, як і за індексом Данна, було отримане трохи вище значення через відмінну від інших, компактнішу вихідну структуру датасету.

Отже, значення метрик якості кластеризації, подані у таблицях 4.16 і 4.17, у першу чергу свідчать про особливості геометричної форми даних. Дійсно у більшості навчальних закладів м. Харків можна знайти певні відмінності за рівнем академічної підготовки їх випускників, проте не надто виразні, якщо оцінювати їх з математичної точки зору. Реальні дані мають високу щільність, і в них майже неможливо виділити «ідеальні» кластери сферичної форми з чіткими межами. Перевіримо, чи не пов'язані малі значення внутрішніх мір із завеликою кількістю кластерів (таблиці 4.18 – 4.20).

Таблиця 4.18 – Метрики якості кластеризації даних, k -means (2 кластери)

Метрика/Рік	2021	2022	2023	2024
Силуетний коефіцієнт	0,05146	0,05883	0,01054	0,01023
Індекс Калінскі – Харабаса	206,17730	152,05881	195,02834	145,54870
Індекс Данна	0,03184	0,02254	0,01551	0,02130

Таблиця 4.19 – Метрики якості кластеризації даних, k -means (3 кластери)

Метрика/Рік	2021	2022	2023	2024
Силуетний коефіцієнт	–0,01417	0,00464	0,03818	–0,02669
Індекс Калінскі – Харабаса	207,70013	173,56916	202,75056	179,65817
Індекс Данна	0,01807	0,00351	0,00580	0,01455

Таблиця 4.20 – Метрики якості кластеризації даних, k -means (4 кластери)

Метрика/Рік	2021	2022	2023	2024
Силуетний коефіцієнт	-0,03562	-0,01543	-0,05757	-0,02777
Індекс Калінскі – Харабаса	216,369	179,046	180,047	172,638
Індекс Данна	0,02081	0,01479	0,02138	0,01851

Аналізуючи показники, розраховані для розбиття даних за методом k -means на 2, 3 та 4 кластери, можемо дійти висновку, що зменшення їх кількості не призведе до значного покращення якості розбиття даних.

Висновки за розділом 4

У четвертому розділі кваліфікаційної роботи викладено результати обчислювального експерименту, імплементованого в системі комп'ютерної алгебри Wolfram Mathematica 10.4, а також їх детальний аналіз. За підсумками проведеного дослідження можемо сформулювати такі висновки:

1. Завдяки компонентному аналізу датасету було зменшено його вихідну розмірність з восьми до двох ознак, що дозволило подати його графічно та дослідити структуру даних за різні роки.

2. У ході кластеризації отримано розбиття ЗСО м. Харків на 5 категорій за навчальними досягненнями їх випускників та вивчено характеристики кожної з них. Виділено групу навчальних закладів, найслабших за рівнем академічної підготовки, що може стати у нагоді при запровадженні заходів, направлених на покращення освітнього процесу у них. Виділено спеціалізовані ЗСО з посиленням вивчення математики, досвід яких можна використати для підвищення рівня підготовки в інших навчальних закладах.

3. За результатами динамічного аналізу щорічних змін у кластерах з однотипним рівнем підготовки до іспиту зафіксовано різке покращення загальної успішності випускників ЗСО м. Харків у 2022 році, причиною якого, можливо,

стало суттєве спрощення тестів при спрямованості навчання на складання складнішого ЗНО. У наступних роках спостерігається прогресуюче зниження балів випускників ЗСО всіх кластерів та погіршення ефективності освітньої діяльності, пов'язане із зовнішніми важкими обставинами.

4. Розраховані значення внутрішніх мір демонструють, що реальні дані мають багато особливостей: наявність шуму, несферичну форму, погану роздільність точок – що значно впливають на якість кластерів. При цьому, кластери були сформованими за дійсно існуючими закономірностями, але ці патерни виявились не достатньо очевидними для отримання високої кількісної оцінки валідності кластерів.

ВИСНОВКИ

Кваліфікаційна робота присвячена інтелектуальному аналізу даних з результатами випускних випробувань з математики випускників ЗСО м. Харків 2021 – 2024 років з метою визначення рівня ефективності освітньої діяльності в кризових умовах. Дослідження було проведено із використанням методів компонентного та кластерного аналізу. У роботі розглянуті теоретичні концепції застосованих методів та їх практична реалізація у системі символічної математики Wolfram Mathematica 10.4. На підставі проведеного дослідження можемо зробити наступні висновки.

Одним з основних результатів системного аналізу задачі оцінювання навчальних досягнень випускників ЗСО м. Харків став вибір неієрархічної кластеризації як основного підходу для її розв’язання. Моделі, побудовані під час цього етапу, взаємно доповнили одна одну, висвітлили різні аспекти розглядуваної задачі та забезпечили її всебічне вивчення.

На основі вивчених теоретичних засад попередньої обробки даних та їх подальшої кластеризації за допомогою неієрархічних ітеративних методів, було створено стратегію розв’язання поставленої задачі.

З використанням засобів математичного пакету Wolfram Mathematica 10.4 було автоматизовано аналіз результатів ЗНО та НМТ з математики у м. Харків за 2021 – 2024 роки. Результатом здійсненого обчислювального експерименту став поділ датасету на кластери, до кожного з яких увійшли ЗСО зі схожими рівнями досягнень випускників на іспиті. Сформовані кластери були досліджені в двох напрямках: по-перше, було вивчено типові характеристики кожного з них окремо, а по-друге, було здійснено їх порівняльний аналіз в часовому розрізі з метою відстеження тенденцій змін у результатах написання випускних випробувань.

Окрім аналізу кластерів з точки зору наявних у них тенденцій, було проведено кількісне оцінювання їх якості. За отриманими значеннями внутрішніх мір валідації кластерів було вивчено вплив особливостей та проблем структури реальних даних на результати кластерного аналізу.

Зроблені висновки щодо ступеня підготовленості випускників різних ЗСО м. Харків до ЗНО та НМТ, а також загального рівня математичної підготовки за 2021 – 2024 роки можуть бути корисними при запровадженні заходів щодо покращення освітнього процесу, умов складання підсумкових випробувань та аналізу якості їх змісту.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Шарай К. В. Методи статистичного аналізу результатів зовнішнього незалежного оцінювання. *29-й Міжнародний молодіжний форум «Радіоелектроніка та молодь у XXI столітті»*: зб. матеріалів форуму (м. Харків, 16 – 18 квітня 2024 р.). Т. 7. Харків : ХНУРЕ, 2024. С. 307 – 308.

2. Шарай К. В. Intellectual data analysis of the results of final exams for assessment the quality of secondary education. *IV Міжнародній науково-практичній конференції англійською мовою «Європейські студії. Навчання і викладання у світі технологій»*: зб. матеріалів форуму. 2025. №3. С. 242.

3. Шарай К. В., Гибкіна Н. В. Неєірархічні методи кластерного аналізу для оцінювання навчальних досягнень випускників у м. Харків за 2021 – 2024 роки. *IX Міжнародна науково-технічна конференція «Комп'ютерне моделювання та оптимізація складних систем КМОСС–2025»*: зб. матеріалів форуму. 2025. №3. С. 227 – 228.

4. Гибкіна Н. В., Шарай К. В. Застосування методів аналізу даних для оцінювання результатів навчання за підсумками складання НМТ з математики. *Всеукраїнська науково-практична конференція здобувачів вищої освіти і молодих учених «Комп'ютерно-інтегровані технології автоматизації технологічних процесів на транспорті та у виробництві»*: зб. матеріалів форуму. Харків : ХНАДУ, 2025. №1. С. 268 – 271.

5. Бродський Ю. Б. Системний аналіз та теорія прийняття рішень: навч. посіб. в 3-х частинах. Частина 1: Системологія. Житомир: Державний університет «Житомирська політехніка», 2022. 92 с.

6. IDEF0. Знайомство з нотацією та приклад використання. URL: <https://www.trinion.org/blog/idef0-znakomstvo-s-notaciley-i-primer-ispolzovaniya> (дата звернення: 14.11.2025).

7. Що таке DFD-діаграма та як її створити? URL: <https://www.lucidchart.com/pages/ru/dfd-diagram> (дата звернення: 15.11.2025).

8. Data mining – cluster analysis. URL: <https://www.geeksforgeeks.org/data->

analysis/data-mining-cluster-analysis/ (дата звернення: 18.11.2025).

9. What is hierarchical clustering? URL: <https://www.ibm.com/think/topics/hierarchical-clustering> (дата звернення: 18.11.2025).

10. Expectation-Maximization algorithm in machine learning. URL: <https://www.geeksforgeeks.org/machine-learning/ml-expectation-maximization-algorithm/> (дата звернення: 19.11.2025).

11. Згуровський М.З., Зайченко Ю.П. Системи і методи штучного інтелекту. Київ: Академперіодика. 2025. 744 с.

12. Гороховатський В.О., Творошенко І.С. Методи інтелектуального аналізу та оброблення даних. Харків: ХНУРЕ, 2021. 92 с.

13. Литвин В. В., Пасічник В.В., Нікольський Ю.В. Аналіз даних та знань. Львів : Магнолія 2006, 2021. 276 с.

14. Data preprocessing in machine learning: steps and best practices. URL: <https://lakefs.io/blog/data-preprocessing-in-machine-learning/> (дата звернення: 22.11.2025).

15. Wickham H., Grolemund G. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media. 2017. 518 p.

16. Bruce P., Bruce A., Gedeck P. Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python. O'Reilly Media. 2020. 360 p.

17. Clustering in data analysis. URL: <https://www.geeksforgeeks.org/dbms/clustering-in-data-mining/> (дата звернення: 22.11.2025).

18. Understanding Distance Metrics Used in Machine Learning. URL: <https://www.analyticsvidhya.com/blog/2020/02/4-types-of-distance-metrics-in-machine-learning/> (дата звернення: 22.11.2025).

19. Dan A. Simovici. Clustering: Theoretical and practical aspects. World Scientific Publishing Co. Pte. Ltd. 2021. 855 p.

20. k-means Clustering: A Deep Dive into Unsupervised Learning. URL: <https://medium.com/k-means-clustering-a-deep-dive-into-unsupervised-learning-81213f56cfc9> (дата звернення: 23.11.2025).

21. Фісун М. Т., Кравець І. О., Казмірчук П. П., Ніколенко С. Г. Інтелек-

туальний аналіз даних: практикум. Львів: «Новий Світ – 2000». 2024. 162 с.

22. Метод кластеризації k -medoids у машинному навчанні. URL: <https://www.geeksforgeeks.org/machine-learning/k-medoids-clustering-in-machine-learning/> (дата звернення: 23.11.2025).

23. Мокін В. Б., Дратований М. В. Наука про дані: машинне навчання та інтелектуальний аналіз даних. Вінниця : ВНТУ, 2024. 258 с.

24. What is Within Cluster Sum of Squares (WCSS)? URL: <https://aiml.com/what-is-within-cluster-sum-of-squares-wcss/> (дата звернення: 25.11.2025).

25. Кононова К.Ю. Машинне навчання: методи та моделі. Харків : ХНУ імені В. Н. Каразіна, 2020. 280 с.

26. Silhouette Coefficient Explained with a Practical Example: Assessing Cluster Fit. URL: <https://medium.com/silhouette-coefficient-explained-with-a-practical-example-assessing-cluster-fit-c0bb3fdef719> (дата звернення: 26.11.2025).

27. Machine Learning for Developers. O'Reilly Media. 2017. 270 p.

28. Dunn index and DB index – Cluster Validity indices. URL: <https://www.geeksforgeeks.org/machine-learning/dunn-index-and-db-index-cluster-validity-indices-set-1/> (дата звернення: 26.11.2025).

29. Calinski-Harabasz Index for K-Means Clustering Evaluation. URL: <https://towardsdatascience.com/calinski-harabasz-index-for-k-means-clustering-evaluation-using-python-4fefe2988e/> (дата звернення: 15.11.2025).

30. Wolfram Mathematica. URL: <https://www.wolfram.com/mathematica/> (дата звернення: 17.11.2025).

31. Український центр оцінювання якості освіти. Статистичні дані. URL: <https://testportal.gov.ua/statystychni-dani/> (дата звернення: 19.11.2025).

32. НМТ-2023 vs НМТ-2022: що змінилося? URL: <https://nus.org.ua/2023/01/23/nmt-2023-vs-nmt-2022-shho-zminylosya/> (дата звернення: 10.11.2025).