Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет інформаційно-аналітичних технологій та менеджменту
(повна назва)

Кафедра прикладної математики
(повна назва)

АТЕСТАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Застосування методів машинного навчання
для дослідження результатів зовнішнього незалежного оцінювання
(тема)

Виконав:
студент _2_ курсу, групи _САУм-19-1_
Коновалова Т.О.
(прізвище, ініціали)
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. Гибкіна Н.В.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри ПМ _____ ____Тевяшев А.Д._____
(підпис) (прізвище, ініціали)

2020 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:

Зав. кафедри ПМ _____
(підпис)
"____" _____ 2020 р.

ЗАВДАННЯ
НА АТЕСТАЦІЙНУ РОБОТУ

студентові _Коноваловій Тетяні Олегівні
(прізвище, ім'я, по батькові)

1. Тема роботи _Застосування методів машинного навчання для дослідження результатів зовнішнього незалежного оцінювання

затверджена наказом по університету від 23 жовтня 2020 р. № 1420 Ст

2. Термін подання студентом роботи до екзаменаційної комісії 10 грудня 2020 р.

3. Вихідні дані до роботи _результати зовнішнього незалежного оцінювання за 2019-2020 н.р. м. Харкова

4. Перелік питань, що потрібно опрацювати в роботі
1. Системний аналіз проблеми дослідження результатів зовнішнього незалежного оцінювання
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 проблеми
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	вересень 2020 р.	виконано
2	Вибір та обґрунтування методу	жовтень – листопад 2020 р.	виконано
3	Розробка алгоритму і програми	листопад – грудень 2020 р.	виконано
4	Проведення аналітичних досліджень та розрахунків	листопад – грудень 2020 р.	виконано
5	Робота над текстом пояснювальної записки	грудень 2020 р.	виконано
6	Представлення роботи на рецензію в ЕК	грудень 2020 р.	виконано

Дата видачі завдання _1_ вересня 2020 р.

Студент _____
(підпис)

Керівник роботи _____ доц. Гибкіна Н.В.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка: 81 с., 26 рис., 10 табл., 4 додатки, 10 джерел.

МЕТОДИ МАШИННОГО НАВЧАННЯ, МЕТОД ГОЛОВНИХ КОМПОНЕНТ, МЕТОДИ КЛАСТЕРІЗАЦІЇ, МЕТОД k-СЕРЕДНІХ, ВЛАСНИЙ ВЕКТОР, ВЛАСНЕ ЗНАЧЕННЯ, СИСТЕМНИЙ АНАЛІЗ, КОРЕЛЯЦІЙНА МАТРИЦЯ, ЗОВНІШНЄ НЕЗАЛЕЖНЕ ОЦІНЮВАННЯ.

Об’єкт дослідження – результати зовнішнього незалежного оцінювання випускників закладів середньої освіти м. Харкова за 2020 рік.

Методи дослідження – методи машинного навчання, зокрема, методи зниження розмірності та методи кластерного аналізу: метод головних компонент та метод k-середніх відповідно.

В атестаційній роботі розв'язується задача дослідження результатів зовнішнього незалежного оцінювання (ЗНО) з окремих предметів за 2020 рік із використанням методів машинного навчання, а саме, методу кластерізації k-середніх та допоміжного методу зниження розмірності – методу головних компонент.

Було розроблено програмний продукт для обробки багатовимірного масиву результатів зовнішнього незалежного оцінювання випускників міста Харкова за 2020 рік та кластерізації закладів середньої освіти міста Харкова щодо якості навчання за окремими предметами.

Отримані результати дослідження були подані у графічному вигляді і на їх основі були зроблені висновки щодо якості освіти.

ABSTRACT

Introductory note: 81 pages, 10 tables, 26 figures, 4 appendixes, 10 sources.

METHODS OF MACHINE LEARNING, PRINCIPAL COMPONENT ANALYSIS, EIGENVALUES, EIGENVALUES, SYSTEM ANALYTICS, CONTEXT DIAGRAM, CORRELATION MATRIX, EXTERNAL INDEPENDENT EVALUATION.

The object of the study is the results of an external independent evaluation of graduates of secondary education institutions in 2020.

Research methods – methods of machine learning, in particular methods of reducing the dimension and methods of cluster analysis – the method of principal components and the method k-means.

The attestation work solves the problem of studying the results of external independent assessment in individual subjects for 2020 using machine learning methods, namely the methods of reducing the dimension – the method of principal components, and methods of cluster analysis – the method k-means, and drawing conclusions based on the research, in particular.

A software product was be developed in order to obtain the result of processing and analysis of the multidimensional array, namely the results of external independent evaluation of graduates of the city of Kharkiv in 2020.

The results of the study were be presented in the form of a graphical interpretation and conclusions were drawn on the classification of educational objects, and the quality of education in certain subjects and profiles of study for each object.

ЗМІСТ

ВСТУП

Освіта є важливою галуззю розвитку будь-якої країни та суспільства у цілому, бо саме вона відповідає за якість знань народу, підготовку кваліфікованих кадрів, а, значить, і розвиток технологій, інновацій тощо. Очевидно, що освіта перш за все має бути якісною. Саме тому дослідження у сфері освіти представляють особливу цінність. Через те, що галузь є дуже поширеною, вона має багато проблем та розгалужень. Дана атестаційна робота представляє дослідження у сфері оцінки подачі знань та якості освіти у цілому у закладах середньої освіти міста Харкова на основі результатів аналізу ЗНО у 2020 році, а також якості підготовки молодого покоління до отримання вищої освіти.

Кожного року заклади середньої освіти (ЗСО) випускають тисячі абітурієнтів, що потенційно можуть стати студентами закладів вищої освіти (ЗВО). Для кожного з випускників постає проблема обрання закладу вищої освіти для вступу. І на цьому етапі важливу роль відіграє профорієнтаційна робота ЗВО. Але тут виникають певні проблеми – кількість ЗСО в України вимірюються тисячами. У таких містах-мільйонниках як Київ, Харків, Львів нараховується сотні шкіл. Тому ЗВО не можуть чисто фізично охопити цю сукупність ЗСО для якісної і влучної профорієнтаційної роботи навіть у межах свого міста, а що казати про свої та інші області країни. Тому є сенс зосереджувати свою увагу на школах з якісною підготовкою з тих предметів, які необхідні для вступу у конкретний ЗВО. Очевидно зробити припущення, що університетам технічної направленості варто зосередити свою увагу на ЗСО з поглибленим вивченням предметів точного профілю, а університетам з підготовкою гуманітарних спеціальностей – на ЗСО з поглибленим вивченням предметів гуманітарного профілю. Але чи все так однозначно та очевидно? Є багато загальноосвітніх шкіл, які мають якість підготовки з конкретних профілів не гірше і навіть краще, чим вищезгадані. І школи з високим рівнем підготовки за предметами точного профілю можуть мати також високий рівень у предметах природнього та гуманітарного профілів.

Тому в атестаційній роботі розглядається задача дослідження результатів зовнішнього незалежного оцінювання (ЗНО) з окремих предметів за 2020 рік з метою оцінки якості вивчення обраних предметів учнями закладів середньої освіти, виявлення можливого взаємозв’язку та схожості освітніх об’єктів та формування цих об’єктів у групи. Дослідження планується проводити за результатами складання тесту учнями ЗСО у межах міста Харкова у 2020 році та за профілями навчання.

Оскільки для кожного освітнього об’єкту розглядаються результати тестування за декількома предметами, то ми маємо справу з багатовимірним масивом даних. Більш того, кожен з предметів має декілька числових значень, що будуть використовуватися у якості ознак, і кожна з цих характеристик може нескінченно подрібнюватися. Очевидно, що масив статистичних даних у такому вигляді є складним та майже неможливим для аналізу. Також через велику кількість ЗСО може виникнути складність обробки результатів і виявлення можливого взаємозв’язку та схожості між ними. Тому у даній роботі запропоновано вирішувати вищезгадані проблеми за допомогою використання методів машинного навчання.

Дана робота є продовження дослідження, що були розпочаті у роботі [2].

1 СИСТЕМНИЙ АНАЛІЗ ПРОБЛЕМИ ДОСЛІДЖЕННЯ РЕЗУЛЬТАТІВ ЗОВНІШЬОГО НЕЗАЛЕЖНОГО ОЦІНЮВАННЯ ТА ПОСТАНОВКА ЗАДАЧ ДОСЛІДЖЕННЯ

1.1 Системний аналіз проблеми дослідження результатів зовнішнього незалежного оцінювання

1.1.1 Вербальна модель системи

Об’єкт аналізу – результати зовнішнього незалежного оцінювання випускників закладів середньої освіти за 2020 рік м. Харкова.

Предмет аналізу – методи машинного навчання, зокрема, методи кластерізації та зниження розмірності, для подальшого аналізу результатів зовнішнього незалежного оцінювання.

Точка зору: дослідник.

Ціль: обрати найкращий за певними критеріями, що будуть розглянуті далі, метод для подальшого дослідження.

Через те, що досягти мети можна тільки враховуючи кожний елемент системи, то розглянута система «Дослідження результатів зовнішнього незалежного оцінювання» є цілісною.

Також система має властивість синергії, бо тільки сума роботи кожного елементу дає результативність.

Великий вплив має на систему навколишнє середовище, а, отже, система є відкритою.

Входами системи є результати зовнішнього незалежного оцінювання випускників 2020 року міста Харкова. Головним виходом системи є результати кластерізації досліджуваних ЗСО.

1.1.2 Морфологічний опис системи

Досліджувана система – результати зовнішнього незалежного оцінювання учнів ЗСО м. Харкова за 2020 рік.

До зовнішнього середовища системи належать елементи, які знаходяться поза системою, але які мають вплив на ефективність, стабільність та в цілому на функціювання системи.

У Додатку А на рисунку А.1 наданий опис зовнішнього середовища системи.

До елементів зовнішнього середовища відносяться:

а) обчислювальна техніка – впливає на швидкість обчислювання, на точність та на вигляд результатів дослідження;

б) український центр якості освіти – збирає та надає статистичні дані для подальшого аналізу, а також встановлює правила проведення і має великий вплив на вигляд даних;

в) розробник – відтворює обраний метод та його алгоритм у вигляді програмного коду, оброблює і аналізує результати;

г) заклади середньої освіти – підготовляють абітурієнтів до складання тесту;

д) методи машинного навчання, зокрема, методи зниження розмірності даних – впливають на відображення даних у відповідному вигляді та використанні у подальшому аналізі.

Модель типу «чорний ящик» є влучним представленням взаємодії системи з навколишнім середовищем. Цей підхід дозволяє на першому етапі побудови моделі сконцентруватися тільки на межах «чорного ящика» та абстрагуватися від детальної реалізації внутрішнього змісту.

Вплив зовнішнього середовища на систему та навпаки називаються відповідно входом і виходами системи. У Додатку А на рисунку А.2 наведена модель «чорний ящик» для обраної системи.

Модель типу «білий ящик», навпаки, описує всі елементи системи, внут-

рішній зв’язок між елементами, параметри внутрішнього середовища. З цим типом моделі варто бути обережним, адже найважливішими є повнота та простота опису, що інколи є складною задачею. Для цього потрібно вміло декомпозувати складні елементи системи та пам’ятати про рівні абстрагування.

Система «Дослідження результатів зовнішнього незалежного оцінювання» складається з наступних підсистем:

а) попередня обробка вхідних даних;

б) вибір найкращого методу розв’язання;

в) розробка програмного продукту;

г) проведення обчислень за допомогою розробленого програмного продукту;

д) аналіз результатів.

У Додатку А на рисунку А.3 надана модель типу «білий ящик» для вищезгаданої системи.

1.1.3 Функціональна модель системи

На рисунку А.4 (Додаток А) наведена контекстна діаграма IDEF0. Ця методологія передписує подальшу побудову ієрархічної системи діаграм. Діаграма показує, що є результатом роботи, не розкриваючи деталізацію складових елементів. Діаграма складається з робот, які зображуються прямокутниками, та зв’язків – стрілок.

Після цього проводиться функціональна декомпозиція системи. Декомпозиція – це процес розбиття складної системи на підсистеми. Кожна з підсистем, у свою чергу, також розбивається на підсистеми до певного ступеня. Це допомагає аналітику більш докладніше розглянути функціональні частини, структурувати послідовність робіт та розглядати кожний елемент з необхідного рівня абстракції.

Розгляд функціональної моделі проводимо з точки зору керівника систе-

ми, тобто аналітика.

На рисунках А.5 та А.6 Додатка А зображені декомпозиції системи «Дослідження результатів зовнішнього незалежного оцінювання».

1.1.4 Інформаційна модель

На рисунку А.7 Додатка А представлена діаграма потоку даних. Data Flow Diagramming (DFD) як і IDEF0 представляє систему у вигляді робіт зі зв'язками. Найчастіше діаграми такого типу використовують разом з моделлю IDEF0 для додаткового представлення поточних операцій.

У DFD діаграмах розглядається, звідки беруться данні, їх обробка і який результат очікується. Вони допомагають зрозуміти, з чого повинна складатися система. Головна мета DFD – показати, як кожна робота перетворює свої вхідні дані у вихідні, а також виявити відносини між цими роботами.

Узагалі синтаксично розділяють два варіанти DFD діаграм – Гейна-Сарсона та Йордана. Вибір типу синтаксису залежить від середовища опису та особистих вподобань аналітика. Ми обрали діаграму Гейна-Сарсона.

1.2 Аналіз сценаріїв вирішення проблеми дослідження результатів зовнішнього незалежного оцінювання

1.2.1 Модель аналізу проблеми

Для якісної обробки даних та отримання висновків дуже важлива візуалізація результатів. Але результати зовнішнього незалежного оцінювання мають велику розмірність і це є проблемою дослідження. Тому є необхідність обрати оптимальний метод зменшення розмірності серед декількох альтернатив.

Отже, для дослідження метод повинен відповідати наступним вимогам:

– К1: гнучкість;

– К2: простота розрахунків;

– К3: точність;

– К4: потужності ЕОМ.

У подальшому будемо називати їх критеріями.

Обирати будемо з множини альтернатив:

– А1: факторний аналіз;

– А2: метод головних компонент (МГК);

– А3: метод екстремального групування ознак.

Побудуємо ієрархічну структуру, використовуючи метод парних порівнянь. Ієрархічна структура приведена на рис. 1.1.

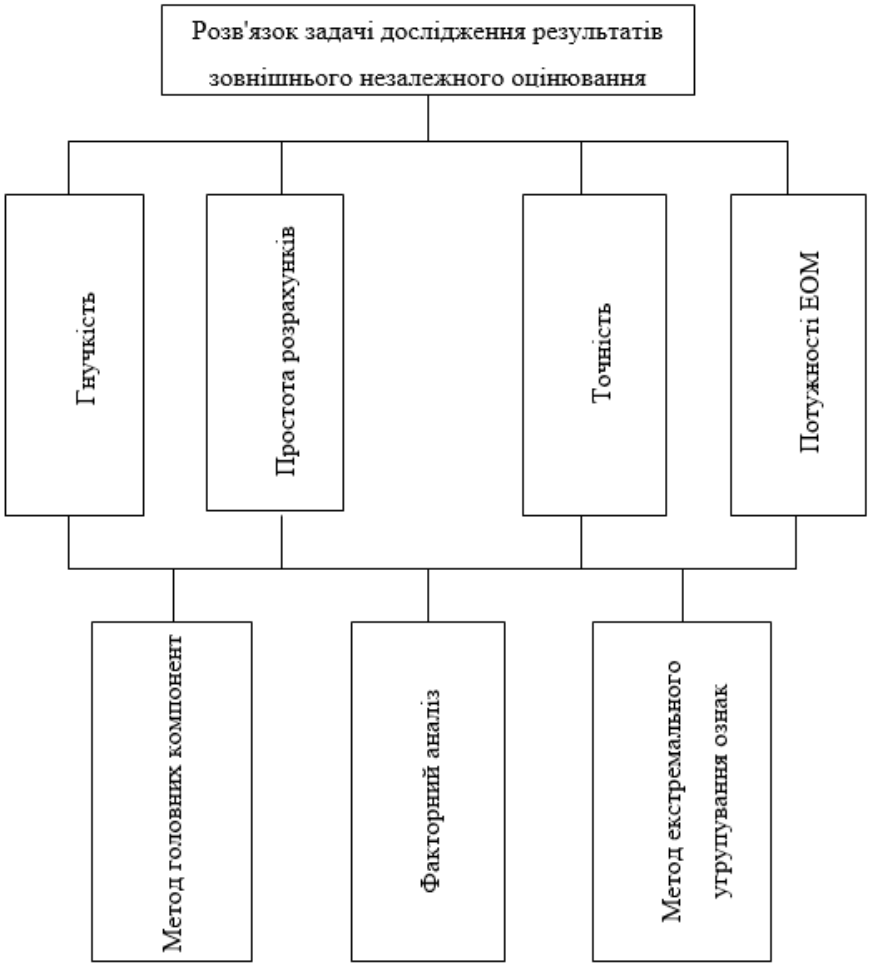

Рисунок 1.1 – Ієрархічна модель процесу аналізу розв'язку задачі

1.2.2 Оцінювання вектора пріоритетів незадоволеностей методом аналізу ієрархій

Для початку потрібно розрахувати вектор локальних пріоритетів критеріїв. Для цього сформуємо матрицю попарних порівнянь важливості критеріїв за шкалою Т. Сааті (таблиця 1.1).

Аналіз вектора пріоритетів першого рівня показав, що найбільш значущою виявилась гнучкість методу $(W_1 = 0,456)$, наступними – простота розрахунків, точність та потужності ЕОМ. На рисунку 1.2 наведена діаграма пріоритетності критеріїв.

Максимальне власне значення матриці суджень дорівнює $\lambda_{\max} = 4,09142$. Індекс узгодженості дорівнює $\text{ІУ} = 0,030472$ і відношення узгодженості $\text{ВУ} = 0,033858$.

Таблиця 1.1 – Розрахунок вектора локальних пріоритетів критеріїв

	К1	К2	К3	К4	Середнє геометричне по рядках	Вектор пріоритетів
К1	1	3	4	6	$x_1 = 2,91295$	$W_1 = 0,560949$
К2	$\frac{1}{3}$	1	4	5	$x_2 = 1,18921$	$W_2 = 0,22901$
К3	$\frac{1}{4}$	$\frac{1}{3}$	1	2	$x_3 = 0,638943$	$W_3 = 0,123042$
К4	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{2}$	1	$x_4 = 0,4518$	$W_4 = 0,087004$
					$\sum_{i=1}^{4} x_i = 5,1929$	

Далі наведено порівняння альтернатив за кожним з критеріїв. У таблицях 1.2 – 1.5 відображена інформація стосовно порівняння кожної альтернативи за кожним з критеріїв відповідно.

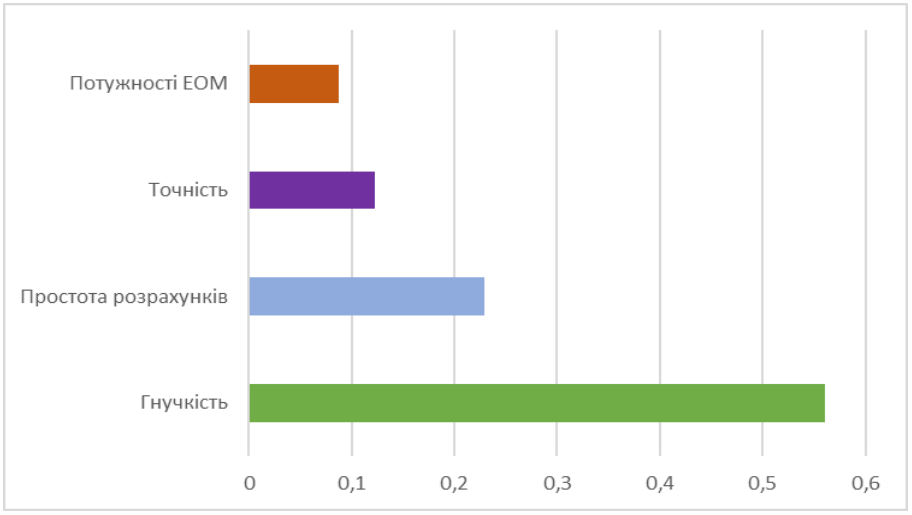

Рисунок 1.2 – Діаграма пріоритетності за критерієм порівняння

Таблиця 1.2 – Порівняння альтернатив за першим критерієм

К1	А1	А2	А3
А1	1	$\frac{1}{2}$	$\frac{1}{5}$
А2	2	1	6
А3	5	$\frac{1}{6}$	1

Таблиця 1.3 – Порівняння альтернатив за другим критерієм

К1	А1	А2	А3
А1	1	$\frac{1}{2}$	$\frac{1}{3}$
А2	2	1	$\frac{1}{2}$
А3	3	2	1

Таблиця 1.4 – Порівняння альтернатив за третім критерієм

К3	А1	А2	А3
А1	1	$\frac{1}{4}$	$\frac{1}{3}$
А2	4	1	2
А3	3	$\frac{1}{2}$	1

Таблиця 1.5 – Порівняння альтернатив за четвертим критерієм

К4	А1	А2	А3
А1	1	$\frac{1}{5}$	$\frac{1}{2}$
А2	5	1	2
А3	2	$\frac{1}{2}$	1

Для кожної з альтернатив порахуємо вектори локальних пріоритетів:

$$\vec{p}_1^{\,A} = \begin{pmatrix} 0{,}125631 \\ 0{,}254704 \\ 0{,}254704 \end{pmatrix},\ \vec{p}_2^{\,A} = \begin{pmatrix} 0{,}16342 \\ 0{,}29696 \\ 0{,}53962 \end{pmatrix},\ \vec{p}_3^{\,A} = \begin{pmatrix} 0{,}121957 \\ 0{,}558425 \\ 0{,}319618 \end{pmatrix},\ \vec{p}_4^{\,A} = \begin{pmatrix} 0{,}128271 \\ 0{,}595379 \\ 0{,}276350 \end{pmatrix}.$$

Далі наведено індекси узгодженості й відношень узгодженості для матриць попарних порівнянь за кожним критерієм:

$$CI_{K1}^{A} = 0{,}435846,\ CI_{K2}^{A} = 0{,}004601,\ CI_{K3}^{A} = 0{,}009147,\ CI_{K4}^{A} = 0{,}002768$$

$$CR_{K1}^{A} = 0{,}751459,\ CR_{K2}^{A} = 0{,}007933,\ CR_{K3}^{A} = 0{,}015771,\ CR_{K4}^{A} = 0{,}004772.$$

1.2.3 Модель вирішення проблеми

Складемо матрицю для розрахунку вектора глобальних пріоритетів альтернатив:

$$P = \begin{pmatrix} 0{,}125631 & 0{,}163424 & 0{,}121957 & 0{,}1282710 \\ 0{,}619665 & 0{,}266961 & 0{,}558425 & 0{,}595379 \\ 0{,}254704 & 0{,}539615 & 0{,}319618 & 0{,}276350 \end{pmatrix}.$$

Розрахуємо вектор глобальних пріоритетів:

$$\vec{p} = \begin{pmatrix} 0{,}125631 & 0{,}163424 & 0{,}121957 & 0{,}1282710 \\ 0{,}619665 & 0{,}266961 & 0{,}558425 & 0{,}595379 \\ 0{,}254704 & 0{,}539615 & 0{,}319618 & 0{,}276350 \end{pmatrix} \cdot \begin{pmatrix} 0{,}560949 \\ 0{,}229006 \\ 0{,}123042 \\ 0{,}087004 \end{pmatrix} = \begin{pmatrix} 0{,}134063 \\ 0{,}536116 \\ 0{,}329821 \end{pmatrix}.$$

Тоді індекс узгодженості й відношення узгодженості для всієї ієрархії дорівнюватимуть:

$$CI = CI^K + \vec{p}^K,\ \overrightarrow{CI^A} = 0{,}27738,$$

$$RI = RI^K + RI^A = 1{,}48,$$

$$CR = \frac{CI}{RI} = 0{,}187419.$$

Найбільша компонента вектора глобальних пріоритетів відповідає другій альтернативі. Отже, за допомогою методу аналізу ієрархій з множини альтернатив для дослідження результатів зовнішнього незалежного оцінювання була обрана альтернатива А2, тобто метод головних компонент.

1.3 Змістовна та формальна постановка задачі

1.3.1 Змістовна постановка задачі

В атестаційній роботі розв'язується задача дослідження результатів зовнішнього незалежного оцінювання з окремих предметів за 2020 рік у м. Харкові з використанням методів машинного навчання, а саме методів зниження розмірності та методів кластерного аналізу. На основі отриманих результатів з оцінювання якості вивчення обраних предметів учнями закладів середньої освіти необхідно виділити групи схожих освітніх об’єктів. Дослідження планується проводити за результатами складання тесту учнями ЗСО.

Оскільки для кожного освітнього об’єкту розглядаються результати тестування за декількома предметами, то ми маємо справу з багатовимірним масивом даних. Більш того, кожен з предметів має декілька числових значень, що будуть використовуватися у якості ознак, і кожна з цих характеристик може подрібнюватися. Очевидно, що масив статистичних даних у такому вигляді є складним та майже неможливим для аналізу. Тому потрібно зменшити розмірність даних за допомогою методу головних компонент. Це стисне обсяг наявної інформації без втрат інформативності та дозволить графічно подати дані на площині.

Групування ЗСО за схожістю у якості підготовки з окремих предметів проводитимемо за допомогою одного з найпоширеніших методів кластерного аналізу – методу k-середніх.

1.3.2 Формальна постановка задачі

Досліджується задача оцінки якості викладання та засвоєння навчального матеріалу з предметів різного профілю (гуманітарних та технічних) учнями закладів середньої освіти у 2020 році міста Харкова, виявлення можливого взає-

мозв'язку та схожості освітніх об'єктів та класифікація цих закладів за схожими ознаками. Також досліджується задача впливу профіля навчання на якість освіти за певними предметами та їх можливий взаємозв'язок при підході до рівня підготовки випускників ЗСО.

Загальна кількість досліджуваних освітніх об'єктів дорівнює n. Кожен з об'єктів описується вектором ознак $\vec{x}^{(1)}$, $\vec{x}^{(2)}$, …, $\vec{x}^{(n)}$ розмірності p. Ознаками виступають наступні показники:

– загальна кількість учнів, що складали ЗНО;

– кількість учнів, які складали ЗНО з певного предмета;

– кількість учнів, які не склали ЗНО з певного предмета;

– кількість учнів, що складали ЗНО з певного предмета й отримали бали у межах [100-120);

– кількість учнів, що складали ЗНО з певного предмета й отримали бали у межах [120-140);

– кількість учнів, що складали ЗНО з певного предмета й отримали бали у межах [140-160);

– кількість учнів, що складали ЗНО з певного предмета й отримали бали у межах [160-180);

– кількість учнів, що складали ЗНО з певного предмета й отримали бали у межах [180-200].

У результаті у якості даних для аналізу отримаємо багатовимірний масив розмірності $n \times p$.

Задля зменшення розмірності багатовимірного статистичного масиву даних потрібно представити кожну характеристику об'єкта у вигляді вектора Y деяких інтегральних показників $y^{(1)}, y^{(2)}, ..., y^{(p')}$ з суттєво меншим за p числом компонент p'.

Спираючись на метод аналізу ієрархій, що був наведений у розділі 1, було вирішено розв'язувати задачу зменшення розмірності методом головних компонент.

За допомогою МГК кожен ЗСО буде характеризуватися двома узагальненими характеристиками. Наступним етапом буде кластерізація цих характеристик з метою утворення груп схожих об'єктів. Спираючись на особливості формату даних, для кластеризації у роботі буде використовуватися ймовірнісний метод кластерного аналізу, а саме метод k-середніх.

Метод k –середніх може бути записаний як задача ітеративної мінімізації внутрішньокластерної суми квадратичних помилок SSE:

$$SSE = \sum_{i=1}^{n}\sum_{j=1}^{\mathrm{k}} \omega_{ij} \left\| \vec{x}^{(i)} - \vec{\mu}^{(j)} \right\|_2^2 .$$

Тут $\vec{\mu}^{(j)}$ – центроїд кластера j, а $\vec{x}^{(i)}$ – елементи тренувального набору даних, тобто ЗСО міста Харкова. Якщо зразок $\vec{x}^{(i)}$ знаходиться у кластері j, то $\omega_{ij} = 1$, інакше $\omega_{ij} = 0$.

1.4 Постановка задач дослідження

Спираючись на результати системного аналізу системи «Дослідження результатів зовнішнього незалежного оцінювання», а також на результати аналізу проблеми вибору ефективного методу розв'язання, сформулюємо задачі дослідження даної атестаційної роботи:

– сформулювати задачу класифікації освітніх об'єктів міста Харкова, а також профілів навчання за результатами складання ЗНО випускниками ЗСО у 2020 році з метою виявлення взаємозв'язку та схожості аналізованих об'єктів;

– застосовуючи метод головних компонент перетворити вихідні дані у масив зменшеної розмірності;

– застосовуючи методи кластерного аналізу отримати групи схожих за якістю навчання з окремих предметів ЗСО міста Харкова;

– розробити програмний продукт, що здійснює основні розрахунки та візуалізує отримані результати;

– за допомогою графічних результатів проаналізувати результати ЗНО за освітніми об'єктами міста Харкова.

2 ВИБІР ТА ОБГРУНТУВАННЯ МЕТОДУ РОЗВ'ЯЗАННЯ

2.1 Сутність проблеми зниження розмірності у багатовимірних задачах

Зазвичай кожен об’єкт у нашому житті характеризується декількома ознаками. Тому при спробі проаналізувати групу таких об’єктів, так чи інакше ми будемо мати справу з багатовимірним масивом даних, де кожен об’єкт характеризується вектором ознак довільної розмірності.

Очевидно, що чим більше об’єктів і ознак, тим складнішим стає їх аналіз. Було б чудово знизити розмірність даних, але ж не втратити інформативності. Але не тільки складність аналізу спонукає перейти до вибірки меншої розмірності. У першу чергу це:

– необхідність лаконічності зв’язаних з обмеженістю місця та об’єму пам’яті сховищ даних, наприклад, бази даних;

– необхідність подання даних у дво-, тривимірному просторі у графічному вигляді;

– економія та оптимізація ресурсів ЕОМ за допомогою спрощення розрахунків.

Описану задачу можна представити наступним чином: загальне число ознак $x^{(1)}, x^{(2)}, ..., x^{(p)}$, зареєстрованих на кожному з досліджуваних об'єктів, дорівнює p. Багатовимірний вектор спостережень

$$X_i = \begin{pmatrix} x_i^{(1)} \\ ... \\ x_i^{(p)} \end{pmatrix}, \; i = 1, 2, ..., n, \tag{2.1}$$

потрібно піддати статистичній обробці. У такому вигляді необхідно представити кожне з спостережень (2.1) у вигляді вектора Z деяких допоміжних показників $z^{(1)}, z^{(2)}, ..., z^{(p')}$ з меншим числом p' компонент ніж p.

При цьому нові (допоміжні) ознаки $z^{(1)}, z^{(2)}, ..., z^{(p')}$ можуть вибиратися з числа початкових або визначатися по якому-небудь правилу по сукупності початкових ознак, наприклад, їх лінійній комбінації, для формування нової системи ознак. До останніх пред'являються різного роду вимоги: найбільша інформативність, взаємна некорельованість, найменша деформація геометричної структури множини початкових даних тощо. У залежності від варіанту формальної конкретизації цих вимог обирається певний алгоритм зниження розмірності [3].

Для побудови математичної моделі методу зменшення розмірності необхідно проаналізувати наступну інформацію:

а) форма завдання вихідних даних :

1) який вигляд мають дані, що описують об'єкт – вигляд так званих одномоментальних спостережень (2.2), чи матриці попарних порівнянь (2.3), чи щось інше:

$$n.c.д. = \begin{pmatrix} x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(p)} \\ x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(p)} \\ \dots & \dots & \dots & \dots \\ x_n^{(1)} & x_n^{(p)} & \dots & x_n^{(p)} \end{pmatrix}, \qquad (2.2)$$

$$(n.c.д.)_2 = \begin{pmatrix} x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(p)} \\ x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(p)} \\ \dots & \dots & \dots & \dots \\ x_n^{(1)} & x_n^{(p)} & \dots & x_n^{(p)} \end{pmatrix} \begin{pmatrix} m = n & чи & p; \\ t = t_1, & \dots, & t_N \end{pmatrix}; \qquad (2.3)$$

2) чи є серед початкових статистичних даних навчаюча інформація, тобто будь-які відомості про аналізовану результуючу властивість;

3) якщо навчаюча інформація присутня в початкових статистичних даних, то в якому саме вигляді представлені дані;

б) тип оптимізуючого критерію $L_{p'}(Z)$ інформативності шуканого набору ознак $Z=\left(z^{(1)},\ldots,z^{(p')}\right)^T$;

в) клас $F(x)$ допустимих перетворень початкових ознак X. Допоміжні ознаки $Z=\left(z^{(1)},\ldots,z^{(p')}\right)^T$ у випадку представлення інформації у вигляді матриці (2.2) конструюються у вигляді функції від X, тобто $Z=Z(X)$ [4].

2.2 Огляд основних методів зниження розмірності

Специфічність реальних задач приводить до проблеми вибору конкретного методу для зниження розмірності: до методу головних компонент, факторного аналізу, екстремального групування параметрів тощо. Далі буде розглянуто основні методи зниження розмірності.

2.2.1 Метод головних компонент

Метод головних компонент (PCA – principal component analysis) відноситься до методів машинного навчання без вчителя. Один із найпоширеніших і найбільш використовуваних методів зниження розмірності. Був винайдений Карлосом Пірсоном у 1901 році. Має широку популярність у аналізі даних, стисненні зображень тощо.

Для зниження розмірності даних необхідно знайти нормовану лінійну комбінацію p вихідних ознак $x^{(1)},x^{(2)},\ldots,x^{(p)}$

$$y^{(1)}=l_{11}x^{(1)}+l_{12}x^{(2)}+\ldots+l_{1p}x^{(p)}=l_1^{'}X.$$

Власний вектор коваріаційної матриці $\sum$, що позначається $L^{(i)}$, визначає

перехід від $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ до $y^{(i)}$. Компоненти цього вектора $l_{i1}, l_{i2}, \ldots, l_{ip}$ є нормованим розв'язком системи рівнянь

$$\left(\sum - \lambda_i I\right) L^{(i)} = 0,$$

де λ_1 – i-й за величиною корінь рівняння

$$\left|\sum - \lambda_i I\right| = 0.$$

Детальніше цей метод буде розглянуто у розділі 2.3.

2.2.2 Факторний аналіз

Як відомо, модель факторного аналізу пояснює структуру зв'язків між початковими показниками $x^{(1)}, \ldots, x^{(p)}$ тим, що поведінка кожного з них статистично залежить від одного й того ж самого набору так званих спільних факторів $y^{(1)}, \ldots, y^{(p')}$, тобто

$$x^j - \mu^{(j)} = \sum_{v=1}^{p'} q_{jv} y^{(v)} + u^{(j)}, \quad j = 1, 2, \ldots, p,$$

де q_{jv} – навантаження спільного фактору $y^{(v)}$ на початковий показник $x^{(j)}$;

$u^{(j)}$ – остаточна специфічна випадкова компонента, причому $My^{(v)} = 0$, $Mu^{(j)} = 0$, $Dy^{(v)} = 1$ і $y^{(1)}, \ldots, y^{(p')}$, $u^{(1)}, \ldots, u^{(p)}$ – попарно некорельовані.

Виявляється, якщо F визначити як клас всіляких лінійних комбінацій $x^{(1)}, \ldots, x^{(p)}$ з урахуванням вищезгаданих обмежень на $y^{(v)}$, а у якості міри інформативності p-мірної системи показників вибрати величину

$l_{p'}(Z(X)) = 1 - \left\| R_X - R_{\hat{X}} \right\|^2$, то розв'язок оптимізаційної задачі співпадає з вектором спільних факторів $y^{(1)}, y^{(2)}, \ldots, y^{(p')}$ у моделі факторного аналізу [3].

2.2.3 Метод екстремального групування ознак

У даному методі йдеться про таке розбиття сукупності початкових показників $x^{(1)}, \ldots, x^{(p)}$ на задане число p' груп $S_1, \ldots, S_{p'}$, що ознаки, які належать до однієї групи, були б взаємнокорельовані більш-менш сильно, у той час як ознаки, що належать до різних груп, мали б слабку кореляцію. Одночасно розв'язується задача заміни кожної групи сильно взаємокорельованих початкових показників одним допоміжним рівнодіючим показником $z^{(i)}$, який, що цілком очевидно, повинен бути у тісному кореляційному зв'язку з ознаками цієї групи. Визначивши у якості класу допустимих перетворень F початкових показників усі нормовані $(Dz^{(i)} = 1)$ лінійні комбінації $x^{(1)}, \ldots, x^{(p)}$, будемо шукати розв'язок $(S_1^*, \ldots, S_p^*; \tilde{z}^{(1)}, \ldots, \tilde{z}^{(p')})$, максимізуючи функціонал

$$l_{p'}(Z(X); S) = \sum_{x^{(k)} \in S_1} r^2(x^{(k)}, z^{(1)}) + \ldots + \sum_{x^{(k)} \in S_{p'}} r^2(x^{(k)}, z^{(p')}),$$

де $r(x, z)$ – коефіцієнт кореляції між змінними x та z [3].

2.3 Метод головних компонент

Метод головних компонент (PCA — principal component analysis) відноситься до методів машинного навчання без вчителя. Один із найпоширеніших і найбільш використовуваних методів зниження розмірності. Основна його ідея –

це послідовне виявлення напрямків, в яких дані мають найбільший розкид.

Головні компоненти є множиною ознак $y^{(1)}, y^{(2)}, \ldots, y^{(p)}$, кожна з яких отримана у результаті деякої лінійної комбінації вихідних ознак $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$, що, у свою чергу, є результатами вимірів та описів об'єктів. Отримані в результаті такого перетворення нові ознаки $y^{(1)}, y^{(2)}, \ldots, y^{(p)}$ мають ряд зручних властивостей. Зокрема, вони впорядковані за ступенем розсіювання у досліджуваній сукупності об'єктів; перша ознака має найбільшу ступінь розсіювання, тобто дисперсію [3].

Розглянемо геометричну інтерпретацію сутності лінійного перетворення вихідної системи ознак на прикладі двовимірної системи спостережень $\left(x_i^{(1)}, x_i^{(2)}\right)$, $i = 1, 2, \ldots, n$, отриманих з нормальної генеральної сукупності із середнім значенням $a = \left(a^{(1)}, a^{(2)}\right)$ та коваріаційною матрицею

$$\sum = \begin{pmatrix} \sigma_1^2 & r\sigma_1\sigma_2 \\ r\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}, \ |r| \le 1, \ \sigma_1 > 0, \ \sigma_2 > 0.$$

Тут σ_1^2 і σ_2^2 – компонент дисперсії, відповідно $x^{(1)}$ і $x^{(2)}$, а r – коефіцієнт кореляції між ними. Геометрично це означає, що точки $\left(x_i^{(1)}, x_i^{(2)}\right)$ будуть розташовуватися приблизно в обрисах еліпсоїдів розсіювання вигляду:

$$\frac{1}{1+r^2}\left[\left(\frac{x^{(1)}-a^{(1)}}{\sigma_1}\right)^2 - 2r\left(\frac{x^{(1)}-a^{(1)}}{\sigma_1}\right)\times\left(\frac{x^{(2)}-a^{(2)}}{\sigma_2}\right) + \left(\frac{x^{(2)}-a^{(2)}}{\sigma_2}\right)^2\right] = c^2.$$

У даному випадку для вивчення $\left(x^{(1)}, x^{(2)}\right)$ зручно перейти до нових координат $\left(y^{(1)}, y^{(2)}\right)$ за допомогою перетворень:

$$y^{(1)} = (x^{(1)} - a^{(1)})\cos\alpha + (x^{(2)} - a^{(2)})\sin\alpha,$$
$$y^{(2)} = -(x^{(1)} - a^{(1)})\sin\alpha + (x^{(2)} - a^{(2)})\cos\alpha,$$

де

$$tg2\alpha = \frac{2r\sigma_1\sigma_2}{\sigma_1^1 - \sigma_2^2}.$$

Будемо вважати, що досліджувані спостереження $X_1, X_2, \ldots, X_n$ були взяті з деякої p-мірної генеральної сукупності, що визначається відповідною ймовірнісною мірою. Однак зі всіх характеристик генеральної сукупності суттєве значення має коваріаційна матриця $\sum = \left(\sigma_{ij}\right)$, де

$$\sigma_{ij} = M(x^{(i)} - a^{(i)})(x^{(j)} - a^{(j)}),\ i, j = 1, 2, \ldots, p.$$

Першою головною компонентою досліджуваної генеральної сукупності спостережень будемо називати таку нормовану лінійну комбінацію p вихідних ознак $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$

$$y^{(1)} = l_{11}x^{(1)} + l_{12}x^{(2)} + \ldots + l_{1p}x^{(p)} = l_1^{'}X,$$

де $l_1^{'} = (l_{11}, l_{12}, \ldots, l_{1p})$, причому $l_{11}^2 + l_{12}^2 + \ldots + l_{1p}^2 = 1$, яка серед усіх інших нормованих лінійних комбінацій $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ має найбільшу дисперсію [3].

І взагалі, i-ю головною компонентою досліджуваної генеральної сукупності будемо називати нормовану лінійну комбінацію вихідних ознак $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$

$$y^{(i)} = l_{i1}x^{(1)} + l_{i2}x^{(2)} + \ldots + l_{ip}x^{(p)} = l_1^{'}X, \qquad (2.4)$$

яка серед усіх інших нормованих $l_{i1}^2 + l_{i2}^2 + \ldots + l_{ip}^2 = 1$ комбінацій, некорельованих з усіма попередніми головними компонентами $y^{(1)},\ldots,y^{(i-1)}$, тобто $\mathrm{cov}(y^{(i)}, y^{(j)}) = M(y^{(i)}y^{(j)}) = 0$, $j<i$, має найбільшу дисперсію [3].

З означення випливає, що, по-перше, головні компоненти $y^{(1)},\ldots,y^{(p)}$ занумеровані у порядку спадання їх дисперсій, тобто $Dy^{(1)} \geq Dy^{(2)} \geq \ldots \geq Dy^{(p)}$, причому

$$Dy^{(i)} = M(l_i^{'}X)^2 = M(l_i^{'}XX^{'}l_i) = l_i^{'}\sum l_i \qquad (2.5)$$

і, по-друге, вектор $L^{(i)}$, що визначає перехід від $x^{(1)}, x^{(2)},\ldots,x^{(p)}$ до $y^{(i)}$, є так званим власним вектором коваріаційної матриці $\sum$, тобто його компоненти $l_{i1}, l_{i2},\ldots,l_{ip}$ визначаються як нормований розв'язок системи рівнянь

$$\left(\sum - \lambda_i I\right)L^{(i)} = 0, \qquad (2.6)$$

де λ_1 – i-й за величиною корінь рівняння

$$\left|\sum - \lambda_i I\right| = 0, \qquad (2.7)$$

де $|M|$ – визначник матриці M;

I – одинична матриця;

λ – невідоме число.

З (2.5), (2.6), (2.7) випливає, що

$$Dy^{(i)} = \lambda_i .$$

Отже, коваріаційна матриця $\sum_\gamma$ головних компонент $y^{(1)}, y^{(2)},\ldots,y^{(p)}$ буде

мати вигляд

$$\sum_{\gamma} = \begin{pmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \lambda_p \end{pmatrix}.$$

Спираючись на те, що перетворення

$$L = \begin{pmatrix} l_{11} & l_{12} & \dots & l_{1p} \\ l_{21} & l_{22} & \dots & l_{2p} \\ \dots & \dots & \dots & \dots \\ l_{p1} & l_{p2} & \dots & l_{pp} \end{pmatrix} = \begin{pmatrix} l_1^{'} \\ l_2^{'} \\ \vdots \\ l_p^{'} \end{pmatrix},$$

за допомогою якого здійснюється перехід від вихідних компонент X до головних компонент Y $(Y = LX)$, – є ортогональним, виразимо початкові змінні $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ через головні компоненти

$$x^{(i)} = l_{1i} y^{(1)} + l_{2i} y^{(2)} + \dots + l_{pi} y^{(p)}.$$

Це дає можливість для визначення необхідної кількості компонент, яких можна прибрати з розгляду і завдяки цьому зменшити розмірність простору, і що важливо, не втратити інформативності у значній мірі.

Внесок перших p' головних компонент $(1 \le p' \le p)$ у загальну дисперсію можна розрахувати наступним чином:

$$q(p') = \frac{Dy^{(1)} + Dy^{(2)} + \dots + Dy^{(p')}}{Dx^{(1)} + Dx^{(2)} + \dots + Dx^{(p)}} = \frac{\lambda_1 + \lambda_2 + \dots + \lambda_{p'}}{\lambda_1 + \lambda_2 + \dots + \lambda_p}. \qquad (2.8)$$

Спираючись на цю формулу, можна сказати, що при виборі кількості p' головних компонент з їх загальної кількості p, вимірність простору задачі зменшиться і стане зручнішим для аналізу у подальшому.

2.4 Метод k-середніх як метод кластерізації

Кластерний аналіз – це сімейство алгоритмів, метою яких є формування груп таким чином, щоб елементи групи були найбільш схожі один до одного і не схожими на елементи, що не входять до угрупування.

Одним з найпопулярніших методів машинного навчання, зокрема кластерного аналізу, є метод k-середніх.

Метод k-середніх – це метод кластерного аналізу, головною метою якого є розділення m об'єктів на k груп $k \le m$, що називаються кластерами, при цьому кожний об'єкт відноситься до кластеру з найближчим центром (центроїд) кластеру. Метод відноситься до класу методів, що навчаються без учителя.

Для обчислювання відстаней між об'єктами найчастіше використовується евклідова метрика:

$$\rho(\vec{x},\vec{y}) = \|\vec{x}-\vec{y}\| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}, \quad \vec{x},\vec{y} \in \mathbb{R}^n.$$

Розглянемо детальніше алгоритм методу k-середніх. Маємо ряд $X = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_m\}$ об'єктів. Метою кластерізації є розподіл m об'єктів на k кластерів $S = \{S_1, S_2, \ldots, S_k\}$ так, щоб мінімізувати середнє квадратичне відхилення від центру кластерів:

$$SSE = \arg\min_{s} \sum_{i=1}^{k} \sum_{j \in S_i} \left\|\vec{x}_j - \vec{\mu}_i\right\|^2,$$

де $\vec{\mu}_i$ – центроїд для кластеру S_i.

Спочатку обираються випадкові координати центроїдів. Метод є ітераційним і виконується у два основних кроки.

Крок призначення: об’єкти з вектору m призначаються до кластерів, де за евклідовою метрикою відстань від об’єкта до центроїда мінімальна:

$$S_i^{(t+1)} = \left\{ \vec{x}_j : \left\| \vec{x}_j - \vec{\mu}_i^{(t)} \right\| \le \left\| \vec{x}_j - \vec{\mu}_l^{(t)} \right\|, l = 1, 2, \ldots, m \right\}, i = 1, 2, \ldots, k.$$

Крок оновлення: розрахунок нових центрів для кластерів:

$$\vec{\mu}_i^{(t+1)} = \frac{1}{\left| S_i^{(t+1)} \right|} \sum_{\vec{x}_j \in S_i^{(t+1)}} \vec{x}_j, i = 1, 2, \ldots k.$$

Цикл закінчується, коли розподіл об’єктів за кластерами перестає змінюватися.

Якщо об’єкти мають різну значущість, то це можна враховувати при розрахунку відстані. Цей метод називають зваженим k-середнім.

Основною перевагою методу є простота реалізації, швидкість виконання і при цьому високий показник ефективності. До недоліків можна віднести високу чутливість до так званих «викидів» у даних, а також залежність від початкового вибору центроїдів. Ще одним недоліком зазвичай виділяють необхідність знати заздалегіть кількість кластерів. Але існує багато варіантів вирішення цього недоліку. Наприклад використання так знаного «метод ліктя».

2.5 Алгоритм розв'язання задачі

Для якісного і комплексного дослідження результатів ЗНО необхідно знизити розмірність даних, кластеризувати об’єкти та нанести отримані результати

на двомірну площину.

Система складається з n досліджуваних об'єктів. Кожен з об'єктів має p ознак.

Для стиснення розмірності даних складної системи використаємо алгоритм методу головних компонент. Для кластерізації даних будемо використовувати метод k-середніх. Для обчислювання відстаней між об'єктами використовуємо евклідову метрику.

Отже, послідовність розрахунків, має наступний вигляд:

а) перейти від абсолютних показників до відносних, бо вони є більш наочними для аналізу;

б) віднормувати нову вибірку даних;

в) обчислити кореляційну матрицю $\sum = \left(\sigma_{ij}\right)$ ознак за віднормованими у п.2 спостереженнями;

г) отримати власні значення $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$ та відповідні до них нормовані власні вектори $L^{(i)} = (l_{i1}, l_{i2}, ..., l_{ip})^T$, $i = 1, 2, ..., p$ для кореляційної матриці Σ;

д) проаналізувавши величину (2.8) вибрати дві головні компоненти $p' = 2$;

е) побудувати для кожного з досліджуваних об'єктів головні компоненти $y^{(1)}$, $y^{(2)}$, …, $y^{(p')}$ – узагальнені інтегральні показники;

є) обрати кількість k кластерів для класифікації;

ж) побудувати об'єкти за стиснутими до розмірності $p' = 2$ та розбитими на кластери S_k координатами у просторі $\mathbb{R}^2$.

2.6 Розв'язання задачі дослідження результатів зовнішнього незалежного оцінювання

Проведемо ряд обчислювальних експериментів для випускників ЗСО міста Харкова за 2020 рік за технічними та гуманітарними предметами. Було обра-

но наступні предмети: українська мова та література, математика, хімія, біологія, фізика та англійська мова. Такий вибір зумовлений бажанням дослідити найбільш різноманітні направлення до вступу. Наприклад, математика, англійська мова та фізика потрібні для вступу у заклади вищої освіти на технічні і ІТ-спеціальності, біологія та хімія – для вступу на медичні спеціальності, а також на спеціальності біоінженерії, виробництва тощо.

Позначимо $n = 235$ – кількість досліджуваних об'єктів, у нашому випадку це ЗСО міста Харкова, $p = 7$ – кількість ознак, за якими буде проводиться аналіз. У якості ознак виступають:

$x^{(1)}$ – відсоток учнів, які складали ЗНО з певного предмета;

$x^{(2)}$ – відсоток учнів, які не склали ЗНО з певного предмету;

$x^{(3)}$ – відсоток учнів, що складали ЗНО з певного предмета й отримали бали у діапазоні [100-120);

$x^{(4)}$ – відсоток учнів, що складали ЗНО з певного предмета й отримали бали у діапазоні [120-140);

$x^{(5)}$ – відсоток учнів, що складали ЗНО з певного предмета й отримали бали у діапазоні [140-160);

$x^{(6)}$ – відсоток учнів, що складали ЗНО з певного предмета й отримали бали у діапазоні [160-180);

$x^{(7)}$ – відсоток учнів, що складали ЗНО з певного предмета й отримали бали у діапазоні [180-200].

У результаті у якості даних для аналізу отримаємо багатовимірний масив розмірності $n \times p$.

Фрагмент вихідної інформації, що була взята на офіційному сайті українського центру якості освіти [7] та оброблена за допомогою SQL запиту, наведеного у додатку Б, можна переглянути у додатку Г. Інформація за іншими вибірками також представлена у додатку Г.

Застосуємо методи машинного навчання та проаналізуємо результати за кожним окремим предметом. Графічні результати та результати кластерізації

будуть переглянуті та проаналізовані у розділі 4.

Переглянемо результати застосування МГК до даних складання ЗНО з української мови та літератури. Оскільки даний предмет є обов'язковим для складання випускниками ЗСО, то розгляд ознаки $x^{(1)}$ не має практичного сенсу і не буде розглядатися у цьому експерименті.

Побудуємо кореляційну матрицю Σ на основі нормованих даних:

$$\begin{pmatrix} 1. & 0.55 & 0.04 & -0.44 & -0.57 & -0.47 \\ 0.55 & 1. & 0.24 & -0.46 & -0.71 & -0.64 \\ 0.04 & 0.24 & 1. & -0.05 & -0.49 & -0.56 \\ -0.44 & -0.46 & -0.05 & 1. & 0.17 & -0.04 \\ -0.57 & -0.71 & -0.49 & 0.17 & 1. & 0.48 \\ -0.47 & -0.64 & -0.56 & -0.04 & 0.48 & 1. \end{pmatrix}$$

Наведемо у порядку спадання власні значення кореляційної матриці:

$$\lambda_1 = 3,05802,\ \lambda_2 = 1,3118,\ \lambda_3 = 0,72877,\ \lambda_4 = 0,485592,\ \lambda_5 = 0,41582,$$
$$\lambda_6 = 0,4 \cdot 10^{-6}.$$

Було обрано використовувати візуалізацію у просторі $\mathbb{R}^2$. Отже будемо використовувати дві головні компоненти. Оцінимо, який внесок у загальну дисперсію дають ці дві компоненти. Цей показник можна інтерпретувати як показник збереження інформативності від початкових даних:

$$q(2) = \frac{\lambda_1 + \lambda_2}{\lambda_1 + \ldots + \lambda_6} = \frac{3,05802 + 1,3118}{3,05802 + \ldots + 0,4 \cdot 10^{-6}} = 0,728304 = 72,83\%.$$

Відповідні до максимальних головних компонент $\lambda_1 = 3,05802$, $\lambda_2 = 1,3118$ власні вектори мають вигляд:

$$L^{(1)} = (0,42;\ 0,5;\ 0,3;\ -0,24;\ -0,48;\ -0,44)^T;$$

$$L^{(2)} = (-0,36;\ -0,15;\ 0,55;\ 0,63;\ -0,09;\ -0,37)^T.$$

Отже, значення перших двох головних компонент дорівнюють:

$$y^{(1)} = 0,42x_2 + 0,5x_3 + 0,3x_4 - 0,24x_5 - 0,48x_6 - 0,44x_7; \quad (2.9)$$

$$y^{(2)} = -0,36x_2 - 0,15x_3 + 0,55x_4 + 0,63x_5 - 0,09x_6 - 0,37x_7. \quad (2.10)$$

Отримані дві головні компоненти $\left(y_i^{(1)}, y_i^{(2)}\right)$, $i = 1,\ldots,n$, є координатами кожного об'єкту на двовимірній площині. За допомогою методу кластерізації кожна з координат буде розподілена до певного кластеру схожих об'єктів. Детальніше графічні результати розглянуті у розділі 4.

Переглянемо результати застосування МГК до даних складання ЗНО з математики.

Кореляційна матриця Σ має вигляд:

$$\begin{pmatrix} 1. & -0.13 & -0.15 & -0.44 & -0.42 & -0.51 & -0.47 \\ -0.13 & 1. & 0.33 & -0.02 & -0.19 & -0.4 & -0.27 \\ -0.15 & 0.33 & 1. & 0.05 & -0.28 & -0.34 & -0.34 \\ -0.44 & -0.02 & 0.05 & 1. & 0.21 & 0. & -0.18 \\ -0.42 & -0.19 & -0.28 & 0.21 & 1. & 0.16 & 0.08 \\ -0.51 & -0.4 & -0.34 & 0. & 0.16 & 1. & 0.46 \\ -0.47 & -0.27 & -0.34 & -0.18 & 0.08 & 0.46 & 1. \end{pmatrix}$$

Наведемо у порядку спадання власні значення кореляційної матриці:

$$\lambda_1 = 2,39223,\ \lambda_2 = 1,60509,\ \lambda_3 = 1,21143,\ \lambda_4 = 0,78258,\ \lambda_5 = 0,606173,$$

$$\lambda_6 = 0,492495,\ \lambda_7 = 0,1 \cdot 10^{-5}.$$

Оцінимо, який внесок у загальну дисперсію дають дві головні компоненти:

$$q(2)=\frac{\lambda_1+\lambda_2}{\lambda_1+\ldots+\lambda_7}=\frac{2,39223+1,60509}{2,39223+\ldots+0,1\cdot10^{-5}}=0,58105=58,11\%.$$

Відповідні до максимальних головних компонент $\lambda_1=2,39233$, $\lambda_2=1,60509$ власні вектори мають вигляд:

$$L^{(1)}=(0,42;\,0,33;\,0,34;\,-0,12;\,-0,34;\,-0,55;\,-0,46)^T;$$
$$L^{(2)}=(0,55;\,-0,37;\,-0,41;\,-0,55;\,-0,23;\,0,08;\,0,16)^T.$$

Отже, значення перших двох головних компонент дорівнюють:

$$y^{(1)}=0,42x_1+0,33x_2+0,34x_3-0,12x_4-0,34x_5-0,52x_6-0,46x_7; \qquad (2.11)$$
$$y^{(2)}=0,55x_1-0,37x_2-0,41x_3-0,55x_4-0,23x_5+0,08x_6+0,16x_7. \qquad (2.12)$$

Переглянемо результати застосування МГК до результатів складання ЗНО з фізики.

Кореляційна матриця Σ має вигляд:

$$\begin{pmatrix} 1. & -0.35 & -0.59 & -0.55 & -0.5 & -0.48 & -0.43 \\ -0.35 & 1. & 0.26 & 0.09 & -0.06 & -0.08 & -0.09 \\ -0.59 & 0.26 & 1. & 0.37 & 0.23 & -0.11 & -0.1 \\ -0.55 & 0.09 & 0.37 & 1. & 0.23 & -0.03 & -0.07 \\ -0.5 & -0.06 & 0.23 & 0.23 & 1. & 0.02 & -0.01 \\ -0.48 & -0.08 & -0.11 & -0.03 & 0.02 & 1. & 0.53 \\ -0.43 & -0.09 & -0.1 & -0.07 & -0.01 & 0.53 & 1. \end{pmatrix}$$

Наведемо у порядку спадання власні значення кореляційної матриці:

$$\lambda_1=2,44765,\ \lambda_2=1,70862,\ \lambda_3=1,06331,\ \lambda_4=0,732191,\ \lambda_5=0,583308,$$
$$\lambda_6=0,464921,\ \lambda_7=0,3\cdot10^{-5}.$$

Оцінимо, який внесок у загальну дисперсію дають дві головні компоненти:

$$q(2)=\frac{\lambda_1+\lambda_2}{\lambda_1+\ldots+\lambda_7}=\frac{2,44765+1,70862}{2,44765+\ldots+0,3\cdot10^{-5}}=0,59375=59,38\%.$$

Відповідні до максимальних головних компонент $\lambda_1=2,44765$, $\lambda_2=1,70862$ власні вектори мають вигляд:

$$L^{(1)}=(0,63;-0,21;-0,42;-0,4;-0,34;-0,24;-0,21)^T;$$
$$L^{(2)}=(-0,09;-0,25;-0,35;-0,25;-0,09;0,6;0,61)^T.$$

Отже, значення перших двох головних компонент дорівнюють:

$$y^{(1)}=0,63x_1-0,21x_2-0,42x_3-0,4x_4-0,34x_5-0,24x_6-0,21x_7; \qquad (2.13)$$
$$y^{(2)}=-0,09x_1-0,25x_2-0,35x_3-0,25x_4-0,09x_5+0,6x_6+0,61x_7. \qquad (2.14)$$

Переглянемо результати застосування МГК до результатів складання ЗНО з біології.

Кореляційна матриця Σ має вигляд:

$$\begin{pmatrix} 1. & -0.41 & -0.69 & -0.68 & -0.62 & -0.51 & -0.02 \\ -0.41 & 1. & 0.49 & 0.27 & 0.02 & -0.08 & -0.2 \\ -0.69 & 0.49 & 1. & 0.5 & 0.15 & 0. & -0.21 \\ -0.68 & 0.27 & 0.5 & 1. & 0.21 & 0.12 & -0.18 \\ -0.62 & 0.02 & 0.15 & 0.21 & 1. & 0.28 & -0.08 \\ -0.51 & -0.08 & 0. & 0.12 & 0.28 & 1. & 0.16 \\ -0.02 & -0.2 & -0.21 & -0.18 & -0.08 & 0.16 & 1. \end{pmatrix}$$

Наведемо у порядку спадання власні значення кореляційної матриці:

$\lambda_1=2,83692$, $\lambda_2=1,50741$, $\lambda_3=0,91789$, $\lambda_4=0,670085$, $\lambda_5=0,6445$,

$$\lambda_6 = 0,42312,\ \lambda_7 = 0,1\cdot 10^{-7}.$$

Оцінимо, який внесок у загальну дисперсію дають дві головні компоненти:

$$q(2) = \frac{\lambda_1 + \lambda_2}{\lambda_1 + \ldots + \lambda_7} = \frac{2,83692 + 1,50741}{2,83692 + \ldots + 0,1\cdot 10^{-7}} = 0,620617 = 62,06\%.$$

Відповідні до максимальних головних компонент $\lambda_1 = 2,83692$, $\lambda_2 = 1,50741$ власні вектори мають вигляд:

$$L^{(1)} = (-0,57;\ 0,32;\ 0,46;\ 0,44;\ 0,32;\ 0,21;\ -0,12)^T;$$
$$L^{(2)} = (0,21;\ 0,42;\ 0,29;\ 0,09;\ -0,37;\ -0,59;\ -0,45)^T.$$

Отже, значення перших двох головних компонент дорівнюють:

$$y^{(1)} = 0,63x_1 - 0,21x_2 - 0,42x_3 - 0,4x_4 - 0,34x_5 - 0,24x_6 - 0,21x_7; \qquad (2.15)$$
$$y^{(2)} = -0,09x_1 - 0,25x_2 - 0,35x_3 - 0,25x_4 - 0,09x_5 + 0,6x_6 + 0,61x_7. \qquad (2.16)$$

Переглянемо результати застосування МГК до результатів складання ЗНО з англійської мови.

Кореляційна матриця Σ має вигляд:

$$\begin{pmatrix} 1. & -0.05 & -0.18 & -0.49 & -0.63 & -0.71 & -0.64 \\ -0.05 & 1. & 0.24 & 0.05 & -0.16 & -0.21 & -0.21 \\ -0.18 & 0.24 & 1. & 0.18 & -0.04 & -0.24 & -0.26 \\ -0.49 & 0.05 & 0.18 & 1. & 0.33 & 0.12 & -0.02 \\ -0.63 & -0.16 & -0.04 & 0.33 & 1. & 0.32 & 0.22 \\ -0.71 & -0.21 & -0.24 & 0.12 & 0.32 & 1. & 0.59 \\ -0.64 & -0.21 & -0.26 & -0.02 & 0.22 & 0.59 & 1. \end{pmatrix}$$

Наведемо у порядку спадання власні значення кореляційної матриці:

$$\lambda_1 = 2,75206,\ \lambda_2 = 1,59035,\ \lambda_3 = 0,95014,\ \lambda_4 = 0,71545,\ \lambda_5 = 0,59302,$$

$$\lambda_6 = 0,39898,\ \lambda_7 = 0,1 \cdot 10^{-7}.$$

Оцінимо, який внесок у загальну дисперсію дають дві головні компоненти:

$$q(2) = \frac{\lambda_1 + \lambda_2}{\lambda_1 + ... + \lambda_7} = \frac{2,75206 + 1,59035}{2,75206 + ... + 0,1 \cdot 10^{-7}} = 0,620344 = 62,03\%.$$

Відповідні до максимальних головних компонент $\lambda_1 = 2,75206$, $\lambda_2 = 1,59035$ власні вектори мають вигляд:

$$L^{(1)} = (0,56; 0,13; 0,08; -0,25; -0,41; -0,49; -0,44)^T;$$

$$L^{(2)} = (0,25; -0,43; -0,6; -0,49; -0,16; 0,19; 0,29)^T.$$

Отже, значення перших двох головних компонент дорівнюють:

$$y^{(1)} = 0,56x_1 + 0,13x_2 + 0,08x_3 - 0,25x_4 - 0,41x_5 - 0,49x_6 - 0,44x_7; \qquad (2.17)$$

$$y^{(2)} = 0,25x_1 - 0,43x_2 - 0,6x_3 - 0,49x_4 - 0,16x_5 + 0,19x_6 + 0,29x_7. \qquad (2.18)$$

Переглянемо результати застосування МГК до результатів складання ЗНО з хімії.

Кореляційна матриця Σ має вигляд:

$$\begin{pmatrix} 1. & -0.23 & -0.49 & -0.72 & -0.7 & -0.42 & -0.52 \\ -0.23 & 1. & 0.03 & 0.04 & -0.02 & 0.1 & 0.03 \\ -0.49 & 0.03 & 1. & 0.15 & 0.23 & 0.03 & -0.02 \\ -0.72 & 0.04 & 0.15 & 1. & 0.39 & 0.11 & 0.24 \\ -0.7 & -0.02 & 0.23 & 0.39 & 1. & 0.09 & 0.36 \\ -0.42 & 0.1 & 0.03 & 0.11 & 0.09 & 1. & 0.1 \\ -0.52 & 0.03 & -0.02 & 0.24 & 0.36 & 0.1 & 1. \end{pmatrix}$$

Наведемо у порядку спадання власні значення кореляційної матриці:

$$\lambda_1 = 2,74224,\ \lambda_2 = 1,08804,\ \lambda_3 = 1,03578,\ \lambda_4 = 0,88146,\ \lambda_5 = 0,71252,$$
$$\lambda_6 = 0,533282,\ \lambda_7 = 0,3 \cdot 10^{-8}.$$

Оцінимо, який внесок у загальну дисперсію дають дві головні компоненти:

$$q(2) = \frac{\lambda_1 + \lambda_2}{\lambda_1 + ... + \lambda_7} = \frac{2,74224 + 1,08804}{2,74224 + ... + 0,3 \cdot 10^{-8}} = 0,54718 = 54,72\%.$$

Відповідні до максимальних головних компонент $\lambda_1 = 2,74224$, $\lambda_2 = 1,08804$ власні вектори мають вигляд:

$$L^{(1)} = (0,6; -0,11; -0,26; -0,43; -0,45; -0,22; -0,34)^T;$$
$$L^{(2)} = (-0,08; 0,74; -0,08; -0,11; -0,3; 0,56; -0,14)^T.$$

Отже, значення перших двох головних компонент дорівнюють:

$$y^{(1)} = 0,6x_1 - 0,11x_2 - 0,26x_3 - 0,43x_4 - 0,45x_5 - 0,22x_6 - 0,34x_7; \qquad (2.19)$$
$$y^{(2)} = -0,08x_1 + 0,74x_2 - 0,08x_3 - 0,11x_4 - 0,3x_5 + 0,56x_6 - 0,14x_7. \qquad (2.20)$$

Проведемо інший експеримент. Він дозволить оцінити якість освіти у розрізі профілів навчання та дослідити їх взаємозв'язок між собою або перевірити, чи є він узагалі. У якості досліджуваних об'єктів виступають профілі навчання ЗСО міста Харкова. Тобто кожен випускник навчався у класі за певним профілем і мав змогу на більш поглиблене вивчання окремого предмету, ніж на вивчення інших загальноосвітніх предметів. У експерименті беруть участь тільки профілі навчання міста Харкова за 2020 рік за певним предметом. Кількість досліджуваних об'єктів $n = 19$, ознаки залишаються незмінними. У цьому ряді

експериментів розглянемо чи будуть мати взаємозв'язок профілі навчання та оцінимо якість підготовки з таких предметів як українська мова та література, математика.

Переглянемо результати застосування МГК до результатів складання ЗНО з української мови та літератури у розрізі профілів навчання.

Кореляційна матриця Σ має вигляд:

$$\begin{pmatrix} 1. & 0.34 & 0.22 & -0.18 & -0.43 & -0.37 \\ 0.34 & 1. & 0.74 & -0.2 & -0.83 & -0.7 \\ 0.22 & 0.74 & 1. & -0.25 & -0.88 & -0.56 \\ -0.18 & -0.2 & -0.25 & 1. & 0.33 & -0.44 \\ -0.43 & -0.83 & -0.88 & 0.33 & 1. & 0.48 \\ -0.37 & -0.7 & -0.56 & -0.44 & 0.48 & 1. \end{pmatrix}$$

Наведемо у порядку спадання власні значення кореляційної матриці:

$$\lambda_1 = 3,34285,\ \lambda_2 = 1,40802,\ \lambda_3 = 0,847793,\ \lambda_4 = 0,261728,$$

$$\lambda_5 = 0,139608,\ \lambda_6 = 0,2 \cdot 10^{-7}.$$

Оцінимо, який внесок у загальну дисперсію дають дві головні компоненти:

$$q(2) = \frac{\lambda_1 + \lambda_2}{\lambda_1 + ... + \lambda_6} = \frac{3,34285 + 1,40802}{3,34285 + ... + 0,2 \cdot 10^{-7}} = 0,791812 = 79,18\%.$$

Відповідні до максимальних головних компонент $\lambda_1 = 3,34285$, $\lambda_2 = 1,40802$ власні вектори мають вигляд:

$$L^{(1)} = (0,28;\ 0,5;\ 0,49;\ 0,11;\ -0,51;\ 0,4)^T;$$

$$L^{(2)} = (0,06;\ -0,03;\ 0,08;\ -0,81;\ -0,17;\ 0,56)^T.$$

Отже, значення перших двох головних компонент дорівнюють:

$$y^{(1)} = 0,28x_2 + 0,5x_3 + 0,49x_4 - 0,11x_5 - 0,51x_6 - 0,4x_7; \quad (2.21)$$

$$y^{(2)} = 0,06x_2 - 0,03x_3 + 0,08x_4 - 0,81x_5 - 0,17x_6 + 0,56x_7. \quad (2.22)$$

Переглянемо результати застосування МГК до результатів складання ЗНО з математики у розрізі профілів навчання.

Кореляційна матриця Σ має вигляд:

$$\begin{pmatrix} 1. & 0.54 & 0.45 & -0.05 & -0.49 & -0.9 & -0.83 \\ 0.54 & 1. & 0.35 & 0.17 & -0.74 & -0.57 & -0.5 \\ 0.45 & 0.35 & 1. & 0.6 & -0.29 & -0.6 & -0.74 \\ -0.05 & 0.17 & 0.6 & 1. & -0.06 & -0.16 & -0.42 \\ -0.49 & -0.74 & -0.29 & -0.06 & 1. & 0.56 & 0.28 \\ -0.9 & -0.57 & -0.6 & -0.16 & 0.56 & 1. & 0.76 \\ -0.83 & -0.5 & -0.74 & -0.42 & 0.28 & 0.76 & 1. \end{pmatrix}$$

Наведемо у порядку спадання власні значення кореляційної матриці:

$$\lambda_1 = 4,00899,\ \lambda_2 = 1,39863,\ \lambda_3 = 0,914152,\ \lambda_4 = 0,331077,\ \lambda_5 = 0,220083,$$

$$\lambda_6 = 0,127068,\ \lambda_7 = 0,6 \cdot 10^{-7}.$$

Оцінимо, який внесок у загальну дисперсію дають дві головні компоненти:

$$q(2) = \frac{\lambda_1 + \lambda_2}{\lambda_1 + ... + \lambda_7} = \frac{4,00899 + 1,39863}{4,00899 + ... + 0,6 \cdot 10^{-7}} = 0,772517 = 77,25\%.$$

Відповідні до максимальних головних компонент $\lambda_1 = 4,00899$, $\lambda_2 = 1,39863$ власні вектори мають вигляд:

$$L^{(1)} = (0,43;\ 0,37;\ 0,37;\ 0,18;\ -0,33;\ -0,45;\ -0,44)^T;$$

$$L^{(2)} = (0,24;\ 0,26;\ -0,44;\ -0,69;\ -0,38;\ -0,12;\ 0,21)^T.$$

Отже, значення перших двох головних компонент дорівнюють:

$$y^{(1)} = 0,43x_1 + 0,37x_2 + 0,37x_3 + 0,18x_4 - 0,33x_5 - 0,45x_6 - 0,44x_7; \quad (2.23)$$

$$y^{(2)} = 0,24x_1 + 0,26x_2 - 0,44x_3 - 0,69x_4 - 0,38x_5 - 0,12x_6 + 0,21x_7. \quad (2.24)$$

Графічна візуалізація результатів кластерізації отриманих методом головних компонент стиснених даних та їх аналіз наведені у розділі 4.

3 ПРОГРАМНА РЕАЛІЗАЦІЯ

3.1 MySQL як засіб формування вибірки даних

База даних – це сукупність даних, що організована згідно з правилами опису характеристик цих даних та їх взаємозв'язку, а також зазвичай зберігається та має електронний доступ з ЕОМ. Розрізняють реляційну та нереляційну бази даних. Розглянемо реляційну базу даних.

Реляційна база даних – база даних, що побудована за принципами реляційної моделі даних. Була запропонована до використання у 1970 році Едвардом Коддом.

MySQL – це система управління реляційними базами даних (СУРБД) з відкритим вихідним кодом. Є однією з найпопулярніших СУБД на ринку. По-перше, це пояснюється безкоштовністю та відкритим кодом, тобто можливістю вільно використовувати та підлаштувати систему під свої вимоги у разі необхідності. Також система є швидкою та безпечною, що також безперечний плюс.

У MySQL використовується мова структурованих запитів (Structured query language – SQL). SQL – це мова програмування для взаємодії з базами даних, здійснення запиту, оновлення та видалення даних, керування даними. Це проста та гнучка мова з низьким порогом входу з легкістю дозволяє починати писати на ній у перший же період ознайомлення. SQL використовується як інструкція роботи та обробки даних. Одною із найбільших переваг є незалежність від СУБД та стандартизованість мови, що дозволяє виконувати однакові запити з однієї СУБД у іншу.

3.2 Mathematica 11 як система символьної математики

Mathematica – це потужна система комп'ютерної алгебри, розроблена Стівеном Вольфрамом. Виходячи з означення, система дозволяє виконувати

найрізноманітніші математичні розрахунки, операції, перетворення та спрощення. Програма включає в себе набір потужних графічних, звукових інструментів, що дозволяє побудову геометричних фігур, обробку зображень, має реалізовані методи статистичного аналізу, машинного навчання тощо. Також пакет Mathematica 11.0 підтримує імпорт та експорт даних з файлів різних форматів, у тому числі зображень.

Mathematica є інтерпретуючою системою, це означає послідовне зчитування та аналіз кожного виразу (інтерпретація) та його миттєве виконання. Тому сам код розбитий на комірки для зручного виконування потрібного фрагменту за необхідністю. Через те, що система використовує одне ядро, яке завантажується під час роботи, є можливою робота у різних документах, що є дуже зручним. Документи, у яких виконуються розрахунки і написаний код, називаються блокнотами.

Основою Mathematica є гнучка символьна мова, що підтримує безліч парадигм програмування, сучасні інструменти налагодження, автоматичне проектування інтерфейсу і багато іншого. Вона спрощує весь процес розробки від дизайну до впровадження. Mathematica всі дані, програми, формули, графіки, документи представляє у вигляді символьних виразів [9].

Також часто можна побачити використання інших обчислювальних систем, таких як MathCAD або MatLab. Варто зауважити, що ці пакети конкурують з Mathematica. Але у цих пакетах символьні розрахунки дуже слабко розвиненні. Тому серед різних середовищ програмування Mathematica є найоптимальнішим вибором для написання програми розв’язання задачі дослідження результатів ЗНО.

3.3 Опис програми

Оскільки дані, що були взяті з офіційного сайту українського центру якості освіти [7], представлені у вигляді показників за кожним випускником ЗСО

всієї України у 2020 році, то потрібно було попередньо підготувати необхідні дані для подальшої обробки. Для цього дані були імпортовані у реляційну базу даних у підготовлену таблицю. Потім за допомогою запиту, що наведений у Додатку Б, були отримані необхідні статистичні дані за кожним ЗСО за певним предметом міста Харкова для подальшого аналізу та експортовані у файл формату csv.

Основна програма, де реалізовані методи машинного навчання, а саме метод головних компонент та метод кластерізації k-середніх для дослідження результатів складання ЗНО випускниками ЗСО, було реалізовано у пакеті Mathematica 11.

Спочатку були імпортовані описані вище дані, після цього йде реалізація алгоритмів, що описані у розділі 2.4 і 2.5. Отримані після зменшення розмірності координати, що характеризують кожен з об’єктів, були відображені на двовимірній площині.

Для подальшого аналізу потрібно розподілити ЗСО за кластерами. З цією метою був застосований метод k-середніх. Для його реалізації використаний метод Mathematica 11 FindClusters з встановленням необхідних атрибутів, таких як Method і DistanceFunction→EuclideanDistance. Після отримання кластерів формуємо висновки та будуємо кругові діаграми якості навчання за отриманими кластерами.

Лістинг програми надається у додатку В.

4 РЕЗУЛЬТАТИ ОБЧИСЛЮВАЛЬНОГО ЕКСПЕРИМЕНТУ

У розділі 2 було наведено обчислювальні розрахунки для запропонованих експериментів з дослідження результатів ЗНО за 2020 рік випускників ЗСО м. Харкова. Проаналізуємо графічні результати кластерізації досліджуваних даних.

Експеримент 1. Проаналізуємо якість підготовки абітурієнтів ЗСО міста Харкова за наступними предметами: українська мова та література, математика, фізика, англійська мова, біологія, хімія. Метою є виявлення груп ЗСО за схожими ознаками, якщо це можливо, і висновки на отриманих результатах. Нижче розглянемо результати за кожним предметом окремо.

Українська мова та література. Отримаємо координати кожного досліджуваного об’єкту за формулами (2.9) – (2.10). Тепер піддамо їх процесу кластерізації і результат відобразимо на двовимірній площині. У цьому випадку у подальших дослідженнях такого типу було обрано розбиття на 5 кластерів, як найоптимальніша для аналізу кількість кластерів. На рисунку 4.1 наведено графічне відображення результатів кластерізації за даними ЗНО з української мови та літератури. У таблиці 4.1 представлені числові характеристики за кожним отриманим кластером. На рис 4.2 наведена діаграма розподілу результатів ЗНО відповідно до кожного кластера. Варто зауважити, що у таблиці 4.1 під поняттями «нижче середнього» та «вище середнього» маються на увазі бали ЗНО у діапазоні [100-140) і [160-200] відповідно. Варто зауважити, що номер кластера не несе ніякої характеристики щодо наповнення кластеру, він використовується тільки як позначка сформованої групи об’єктів. Як бачимо, у кластері 5 найбільший відсоток випускників, що мають результати вище середнього. До цього кластеру увійшли, окрім добре відомих шкіл, такі школи як спеціалізована школа (СШ) №16 ім. Сергеєва, ліцей №4, загальноосвітня школа (ЗОШ) №4 та СШ №119 тощо. У цих школах від 35 відсотків учнів склали ЗНО з української мови у межах [160-200]. Кластери 4 і 3 трохи гірші за показниками, але є все ще достатньо хорошими. Тобто під час вступної агітації слід звернути свою увагу на такі школи як, наприклад, ЗОШ №70, гімназії № 172 та №65 тощо, які належать до цих кластерів.

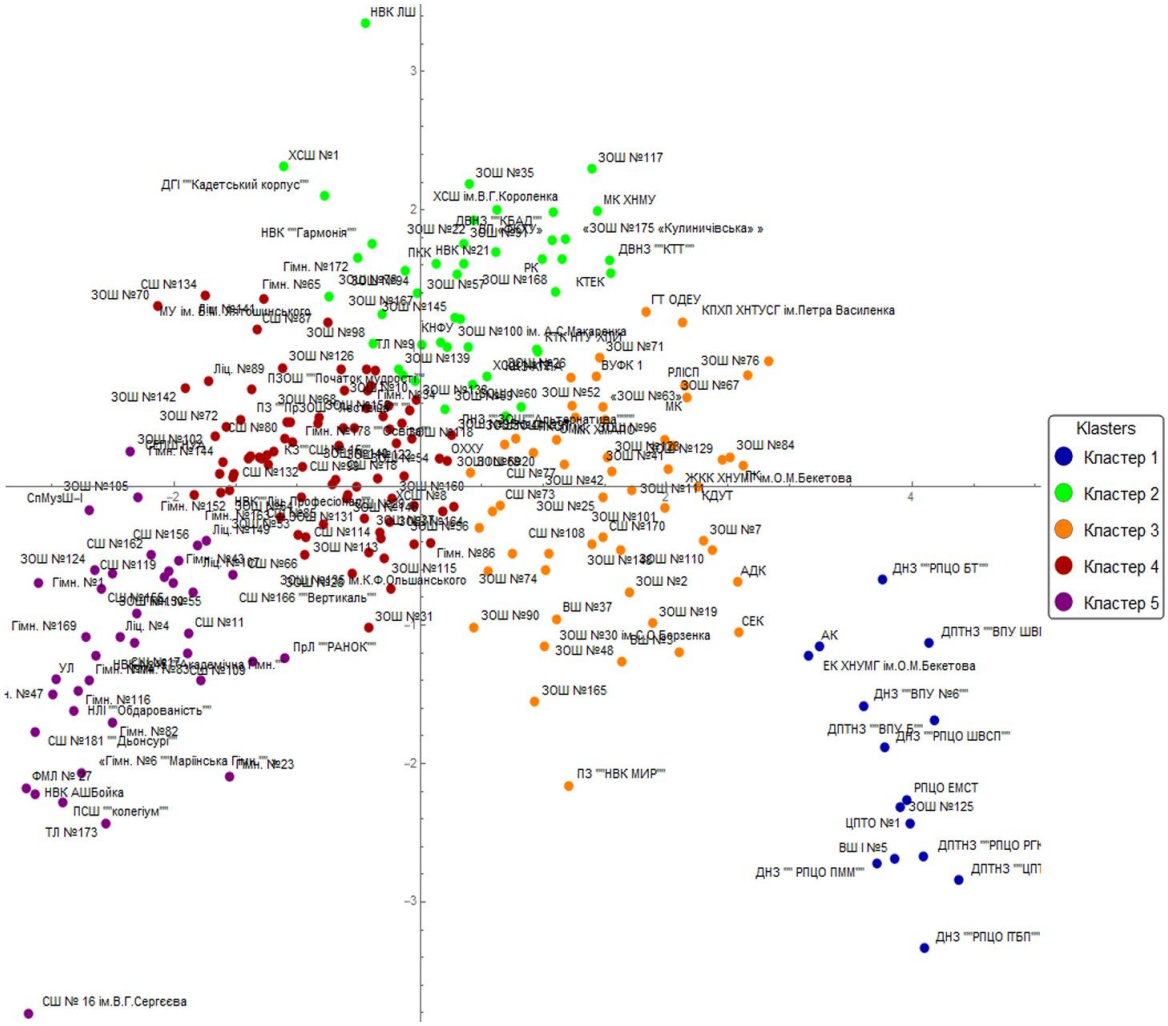

Рисунок 4.1 – Класифікація ЗСО м. Харкова за результатами ЗНО з української мови та літератури за 2020 рік

Таблиця 4.1 – Зведені числові характеристики за кожним кластером з української мови та літератури

№ кластера	1	2	3	4	5
загальна кількість об’єктів	15	55	37	88	40
загальна кількість учнів	705	1611	1400	2768	1610
кількість об’єктів з ненульовим % абітурієнтів, що склали ЗНО					
нижче середнього	15	55	37	88	35
вище середнього	10	54	37	88	40
межі % абітурієнтів, що склали ЗНО					
нижче середнього	8,88-61,9	10,34-60	0-25,15	0-31	0-25
вище середнього	0-4,76	0-37,04	0-16,67	7,41-22,2	12,5-71,88
середнє значення % абітурієнтів, що склали ЗНО					
нижче середнього	30,99	21,25	16,51	11,09	3,56
вище середнього	1,35	10,04	10,70	21,47	35

Математика. Відображення результатів для математики на двовимірній площині та діаграма розподілу результатів ЗНО відповідно до кожного кластера надано на рисунку 4.3 і 4.4 відповідно. У таблиці 4.2 представлені числові характеристики за кожним отриманим кластером. Також для математики проведемо класифікацію зваженим методом k-середніх. У якості ваг використаємо показник кількості учнів у школі, які складали ЗНО з математики у кожній школі. Таким чином, школи з великою кількістю учнів будуть мати більше впливу на формування кластеру. Зроблено це для того, щоб уникнути ситуації, коли дуже маленька школа потрапляє до дуже гарного чи навпаки слабкого кластеру, хоча в ній усього 2-3 випускники мають високий результат. Результати цього експерименту наведені на рисунку 4.5. Як бачимо, наше припущення спрацювало і кластери значно змінилися.

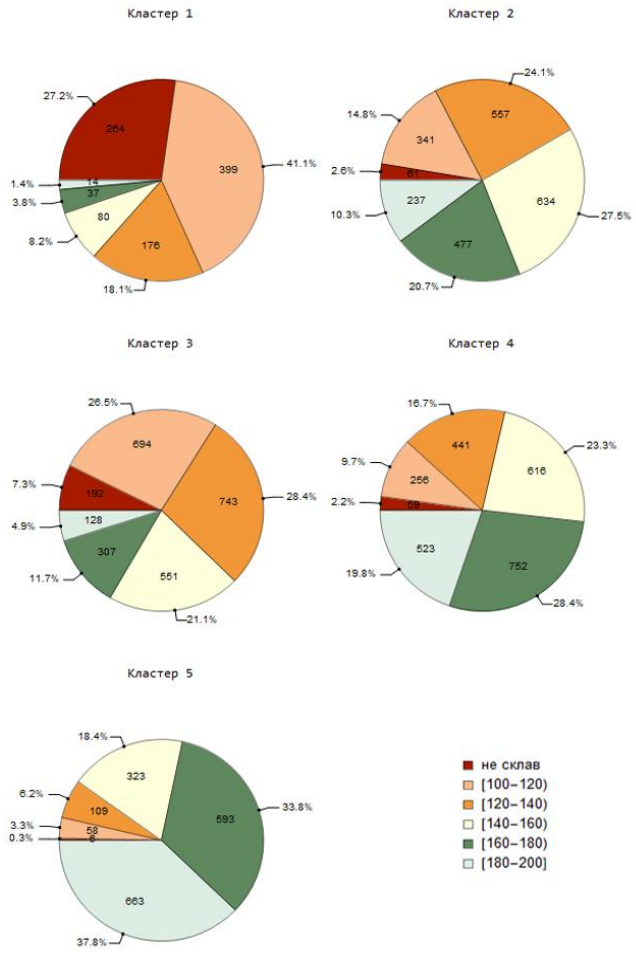

Рисунок 4.2 – Діаграма розподілу результатів

ЗНО з української мови та літератури відповідно до кожного кластера

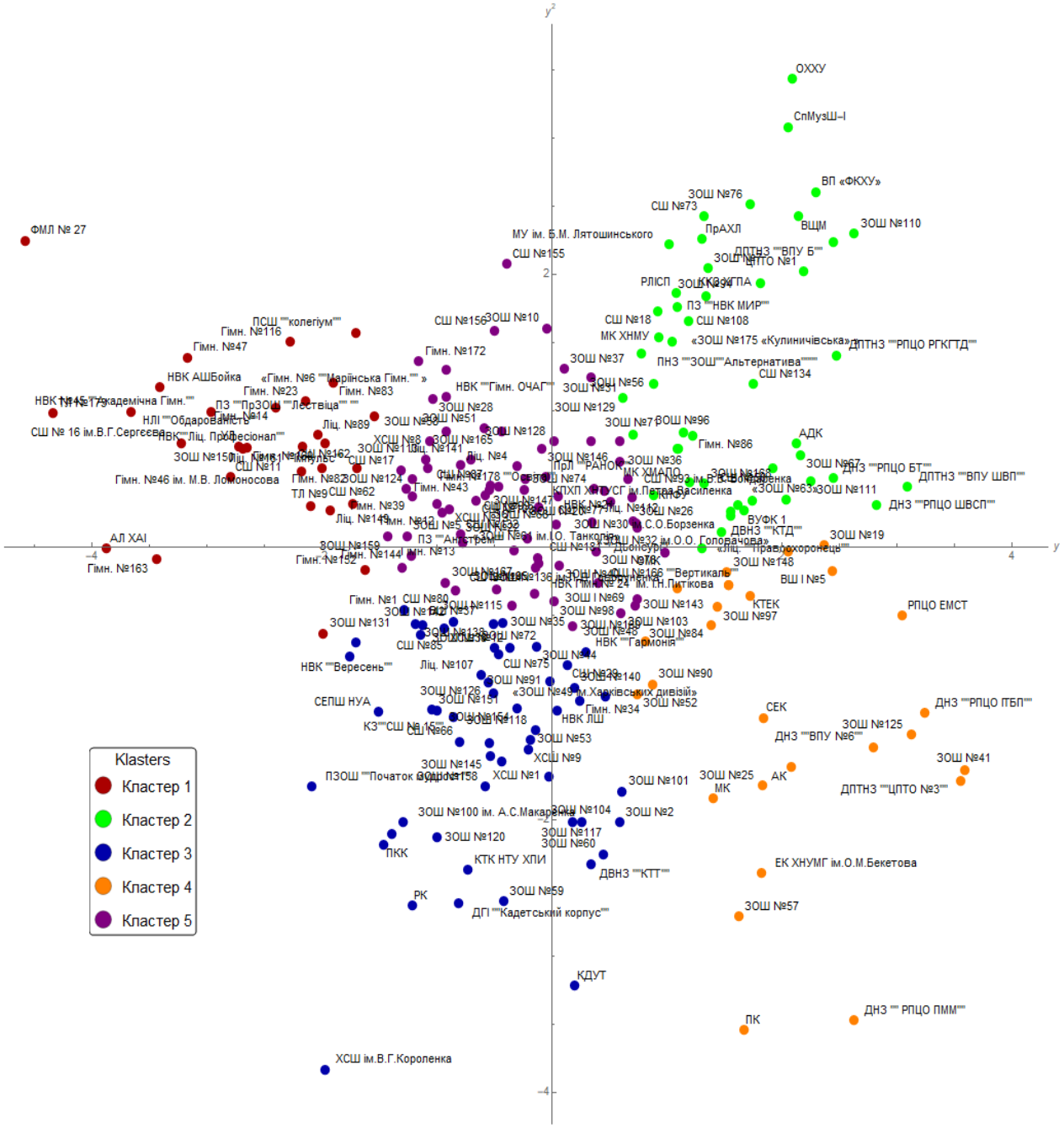

Рисунок 4.3 – Класифікація ЗСО м. Харкова
за результатами ЗНО з математики за 2020 рік без ваг

На рисунку 4.5 невеликі ЗСО як ліцей №149 та ліцей №89 перейшли до кластеру під номером 5, який трохи програє у показниках кластеру 1. Така ж ситуація спостерігається у кластерах 2 та 4. Узагалі можна сказати, що кластери стали більш чіткими завдяки цьому методу кластерізації.

Таблиця 4.2 – Зведені числові характеристики за кожним кластером з математики

№ кластера	1	2	3	4	5
загальна кількість об’єктів	25	72	50	14	74
загальна кількість учнів	1327	3063	2095	712	3094
кількість об’єктів з ненульовим % абітурієнтів, що склали ЗНО					
нижче середнього	17	69	50	14	71
вище середнього	25	62	50	11	74
межі % абітурієнтів, що склали ЗНО					
нижче середнього	0-19,23	0-45,45	0-35,71	11-50	0-24,56
вище середнього	8,77-87,93	0-9,09	3,45-23	0-6,67	6,45-42,8
середнє значення % абітурієнтів, що склали ЗНО					
нижче середнього	5,89	11,78	19,81	22,57	10,07
вище середнього	26,35	2,82	7,27	1,23	15,24

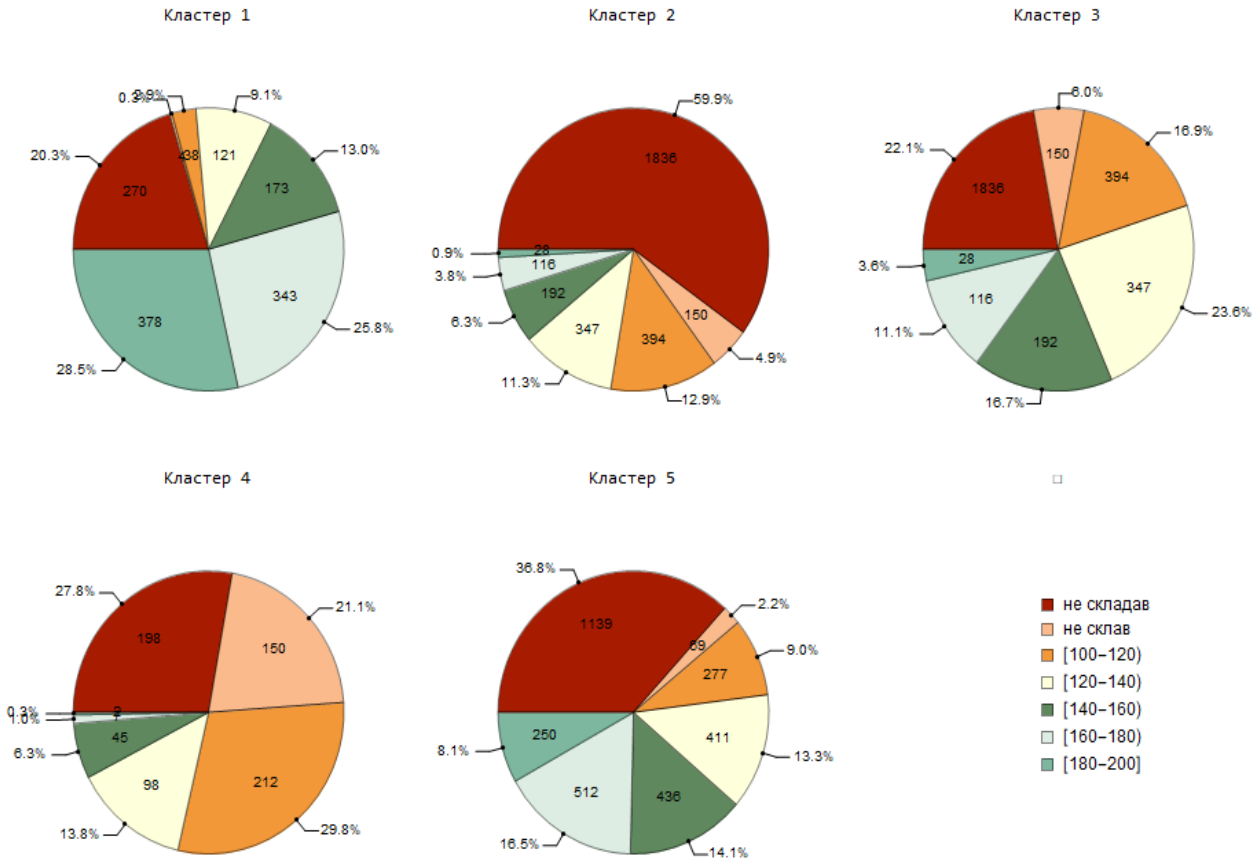

Рисунок 4.4 – Діаграма розподілу результатів ЗНО з математики відповідно до кожного кластера

Рисунок 4.5 – Класифікація ЗСО м. Харкова
за результатами ЗНО з математики за 2020 рік з вагами

Фізика та хімія. Розглянемо не найпопулярніші для здачі ЗНО предмети – фізику та хімію. Розраховані за формулами (2.13) – (2.14) та (2.19) – (2.20) координати з фізики та хімії зображені на рисунках 4.6 та 4.9 відповідно. Також були побудовані діаграми розподілу випускників за балам ЗНО для кожного кластера (рисунки 4.7 та 4.8).

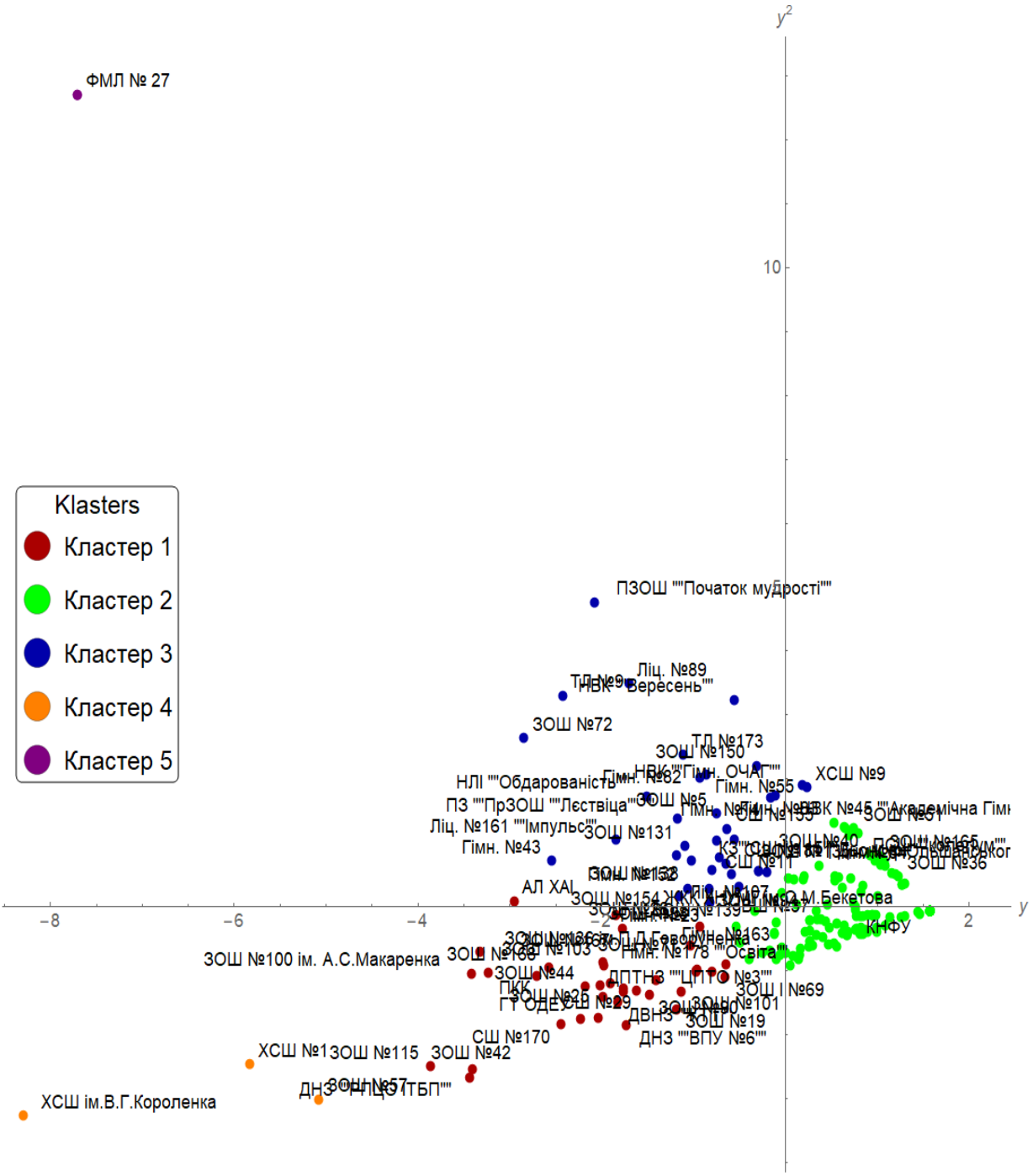

Рисунок 4.6 – Класифікація ЗСО м. Харкова
за результатами ЗНО з фізики за 2020 рік

Як видно з графіків та діаграм, дані по цим предметам є не такими наочними як попередні. На жаль, у цьому випадку результат не є інформативним, бо більшість випускників не складає ці предмети. Але що цікаво, у фізики фізико-математичний ліцей №27 утворює власний кластер. Це не дивно, бо велика кількість учнів цієї школи складали фізику і мають високі результати.

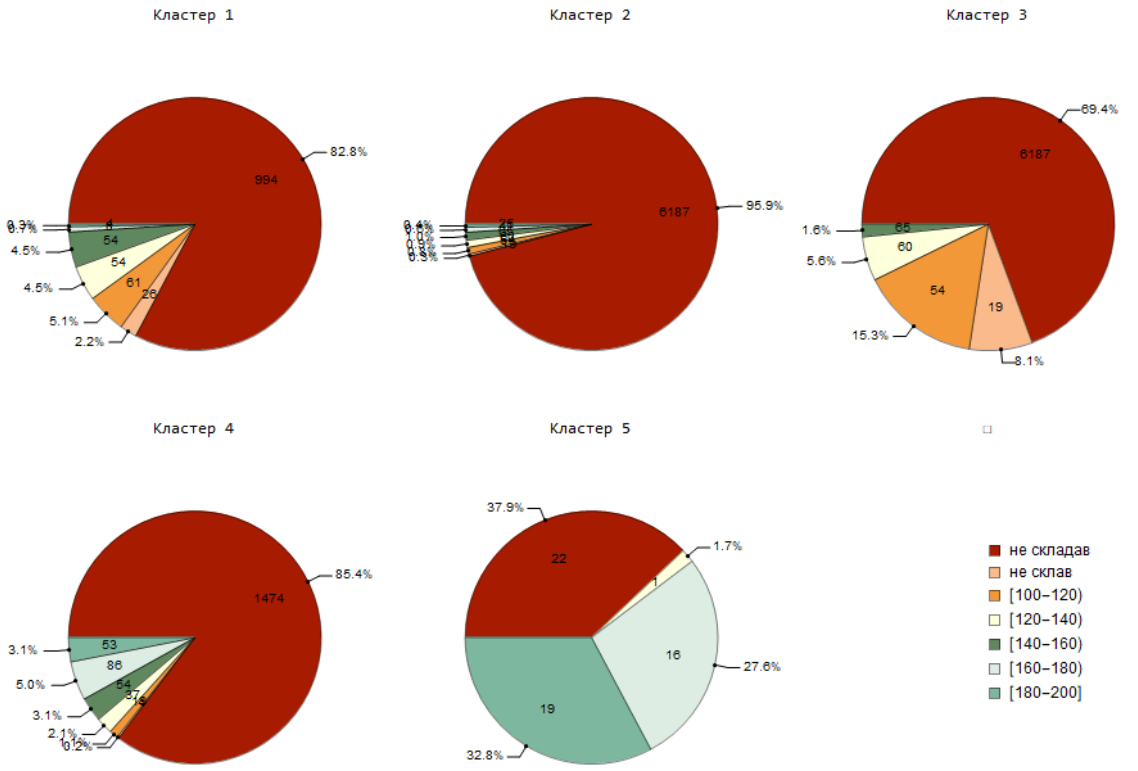

Рисунок 4.7 – Діаграма розподілу результатів ЗНО з фізики відповідно до кожного кластера

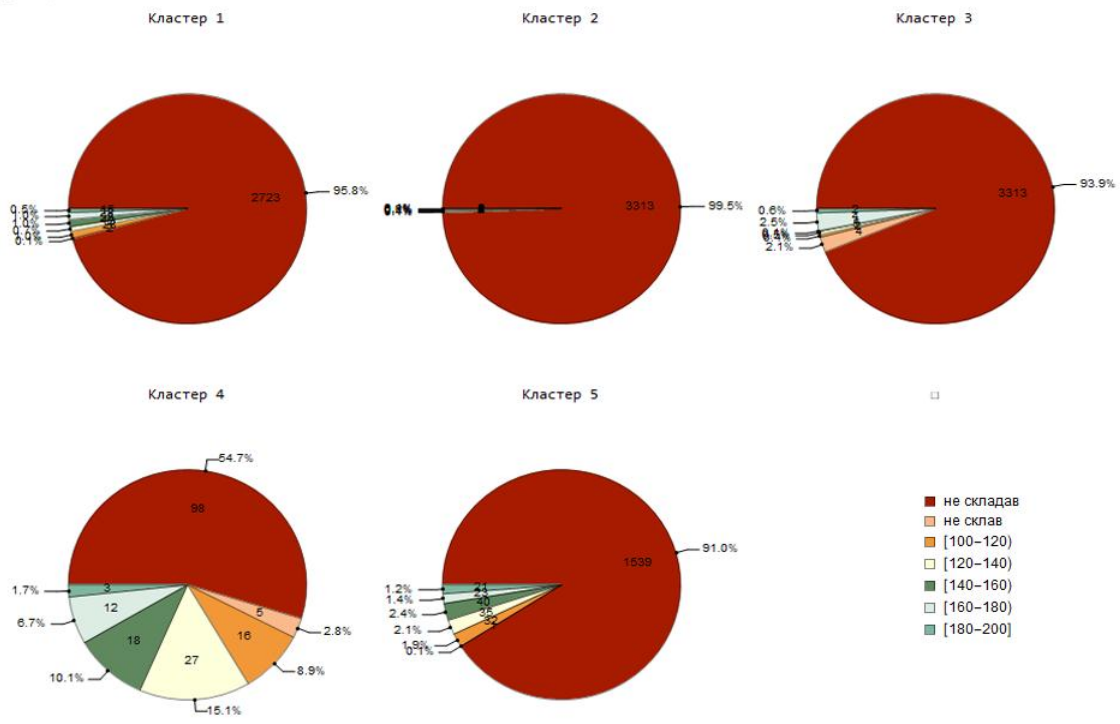

Рисунок 4.8 – Діаграма розподілу результатів ЗНО з хімії відповідно до кожного кластера

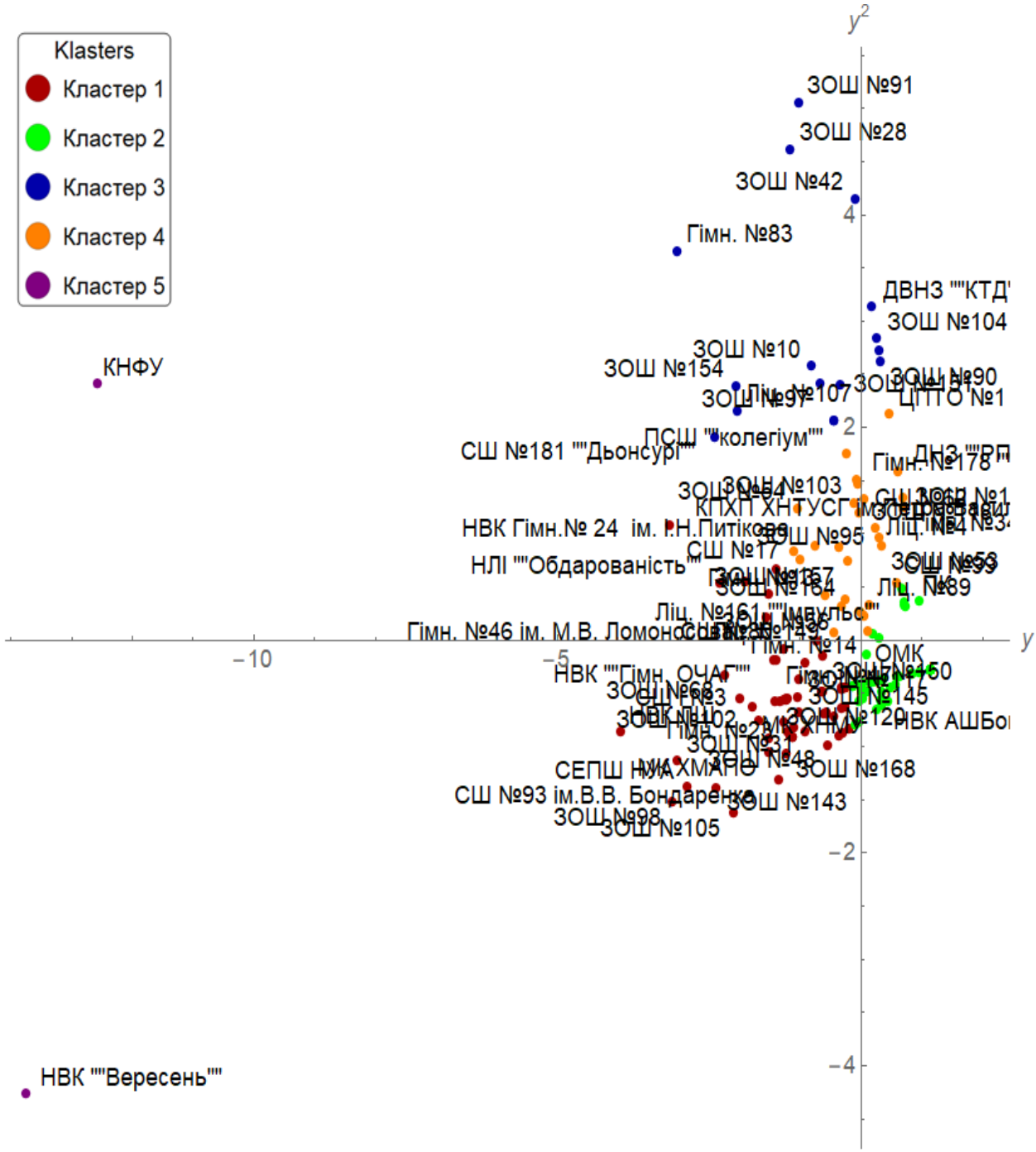

Рисунок 4.9 – Класифікація ЗСО м. Харкова
за результатами ЗНО з хімії за 2020 рік

Біологія. Біологія є дуже цікавим предметом з точки зору аналізу. Вона не є такою популярною для здачі як математика, історія та подібні предмети, але вона доволі часто потребується навіть не для медичних спеціальностей. Отримані координати, що були розраховані за формулами (2.15) – (2.16), наведені на рисунку 4.10. Також у таблиці 4.3 представлені числові характеристики за кож-

ним отриманим кластером. На рис 4.11 наведена діаграма розподілу результатів ЗНО з біології відповідно до кожного кластера. Як і очікувалося, ЗСО з медичнім профілем угрупували окремий кластер. Це ще раз доводить, що методи машинного навчання гарно показують себе під час розв'язання поставленої задачі. Варто зазначити, що школи, які входять до кластеру 4, також мають непогані результати і вигідно відрізняються від інших кластерів.

Таблиця 4.3 – Зведені числові характеристика за кожним кластером з біології

№ кластера	1	2	3	4	5
загальна кількість об'єктів	125	11	52	43	4
загальна кількість учнів	6024	428	2193	1499	185
кількість об'єктів з ненульовим % абітурієнтів, що склали ЗНО					
нижче середнього	46	11	52	32	3
вище середнього	98	8	39	43	4
межі % абітурієнтів, що склали ЗНО					
нижче середнього	0-11,11	4,54-35,7	0-18,75	0-18,51	0-13,85
вище середнього	0-12,9	0-13,33	0-11,11	0-17,6	0-45,83
середнє значення % абітурієнтів, що склали ЗНО					
нижче середнього	1,69	21,42	8,23	6,29	8,44
вище середнього	2,94	1,87	2,19	5,41	20,64

Англійська мова. Цей предмет доволі часто здають за вибором і він потребується для вступу у ЗВО різних профілей. Результати кластерізації наведені на рисунках 4.12 та 4.13. Як бачимо, гарні результати дають кластери 1 та 5. ЗСО, що входять до цих кластерів, мають у середньому 13-28 відсотків учнів, що отримали бали у діапазоні [180-200].

Експеримент 2. Суть цього експерименту у тому, щоб оцінити якість освіти у розрізі профілів навчання та класифікувати профілі навчання, якщо це можливо. У якості досліджуваних об'єктів виступають профілі навчання ЗСО міста Харкова у 2020 році. У цьому ряді експериментів розглянемо, чи будуть мати взаємозв'язок профілі навчання та оцінка якості підготовки з таких предметів як українська мова та література, математика.

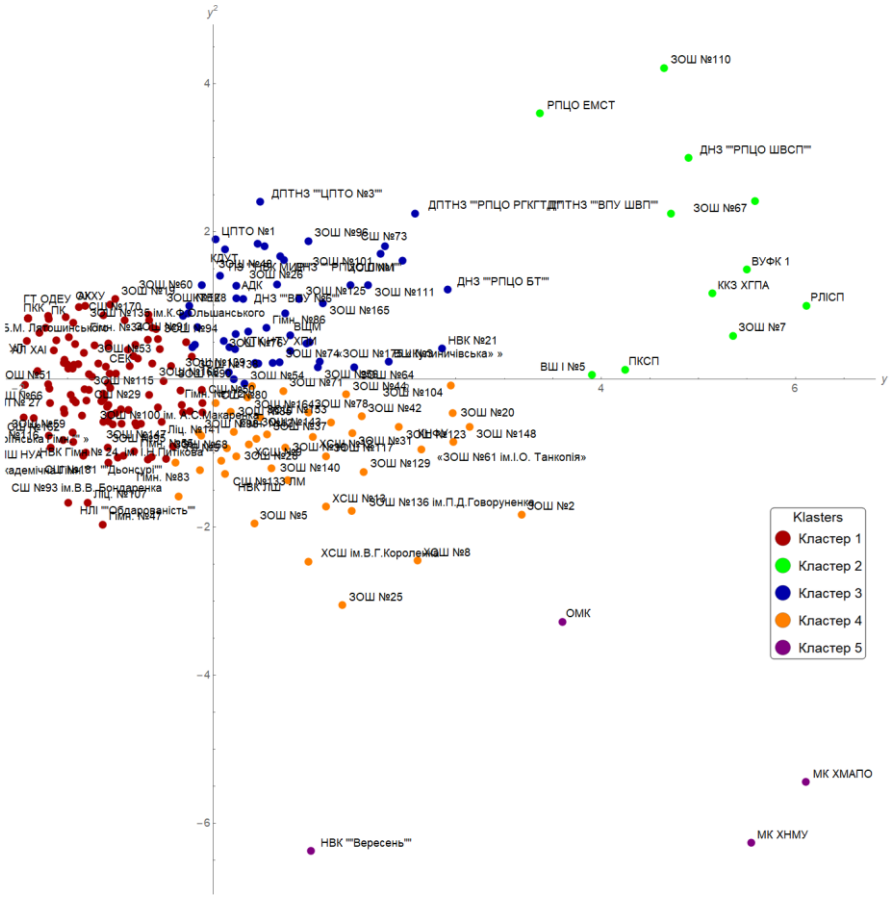

Рисунок 4.10 – Класифікація ЗСО м. Харкова
за результатами ЗНО з біології за 2020 рік

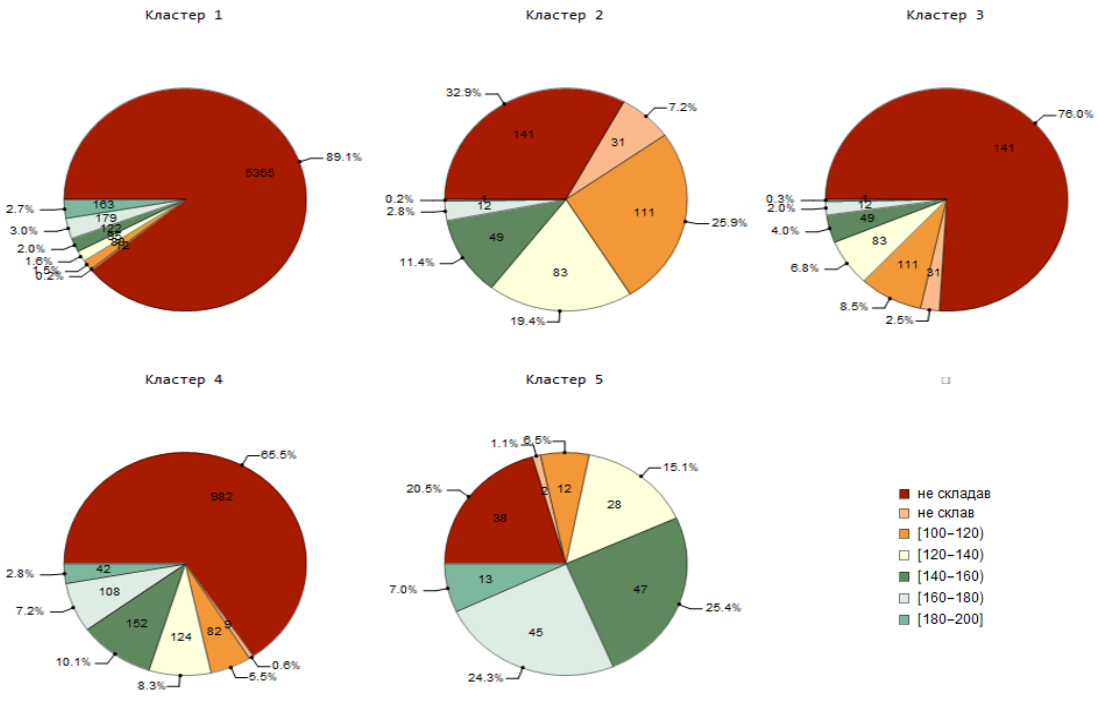

Рисунок 4.11 – Діаграма розподілу результатів ЗНО з біології

Рисунок 4.12 – Класифікація ЗСО м. Харкова
за результатами ЗНО англійської мови за 2020 рік

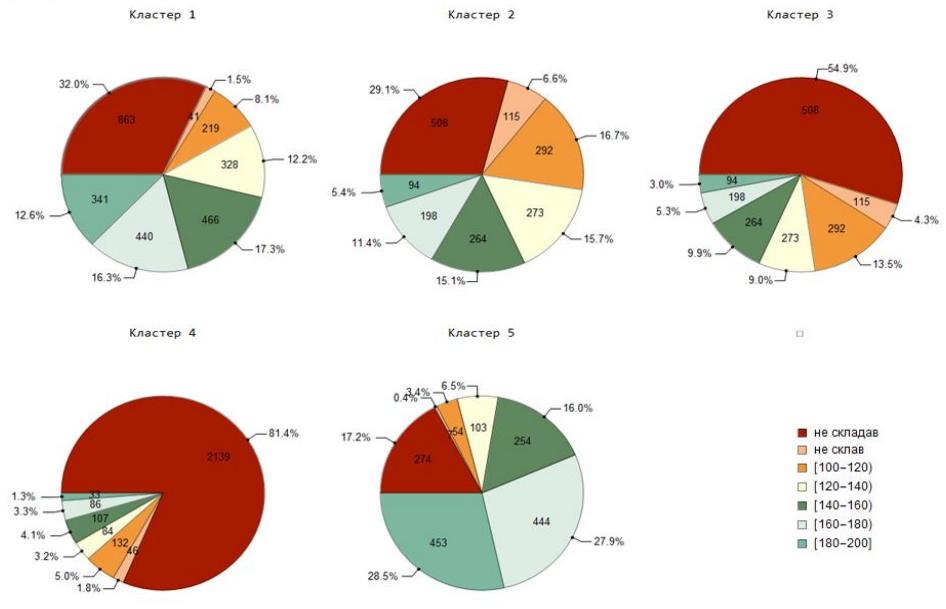

Рисунок 4.13 – Діаграма розподілу результатів ЗНО з англійської мови

Українська мова та література. Як і у попередніх дослідженнях, отримуємо за формулами (2.21) – (2.22) координати, що відображають усі 7 ознак для кожного з об’єктів. Отриманий графік та діаграму розподілу за кластерами наведено у рисунках 4.14 та 4.15. Так як даних не так багато, як у попередніх випадках, то кількість кластерів розбиття оберемо рівною 4. Результати за профілями для української мови та літератури виявились достатньо неочікуваними.

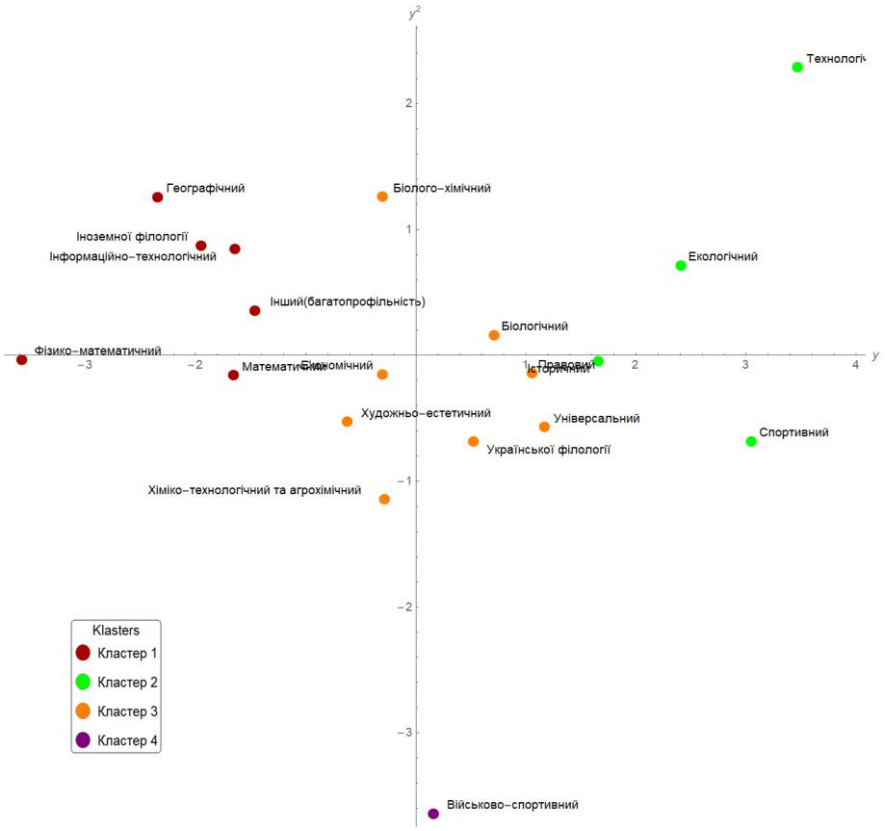

Рисунок 4.14 – Класифікація профілів навчання за результатами ЗНО з української мови та літератури

Як видно, учні з технічних профілів навчання дуже добре складають українську мову та літературу. Цікаво ще те, що профіль з поглибленим вивченням української мови показує непогані результати, але не найкращі, і цей профіль не ввійшов до сильного кластеру. Ще цікавою особливістю є наявність у сильному кластері географічного профілю навчання. А ось учні екологічного, спортивного та технологічного профілю отримують невисокі бали з ЗНО з української мови та літератури.

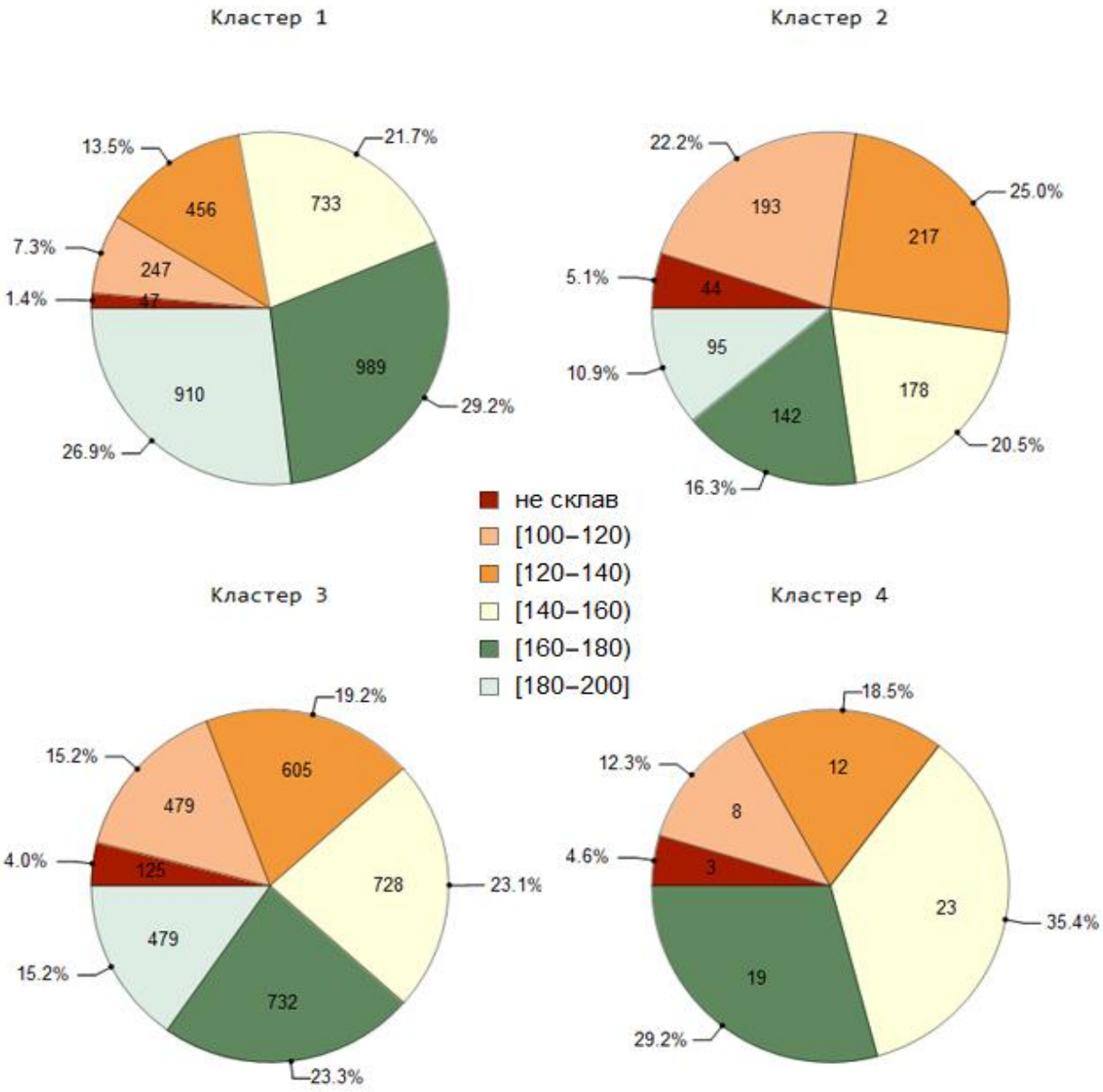

Рисунок 4.15 – Діаграма розподілу результатів ЗНО
з української мови та літератури у розрізі профілів навчання

Математика. Аналогічно попередньому експерименту, розглянемо результати стиснення 7 ознак до двох та відображення кластерізованих координат на площині. Як видно з рисунків 4.16 та 4.17, результати вийшли майже очікуваними. Фізико-математичний профіль навчання займає лідируючу позицію і сильно відрізняється від інших профілів навчання. Саме тому цей профіль самостійно утворює кластер. Цікаво, що такий гуманітарний профіль як іноземна філологія, також має дуже непогані результати з математики.

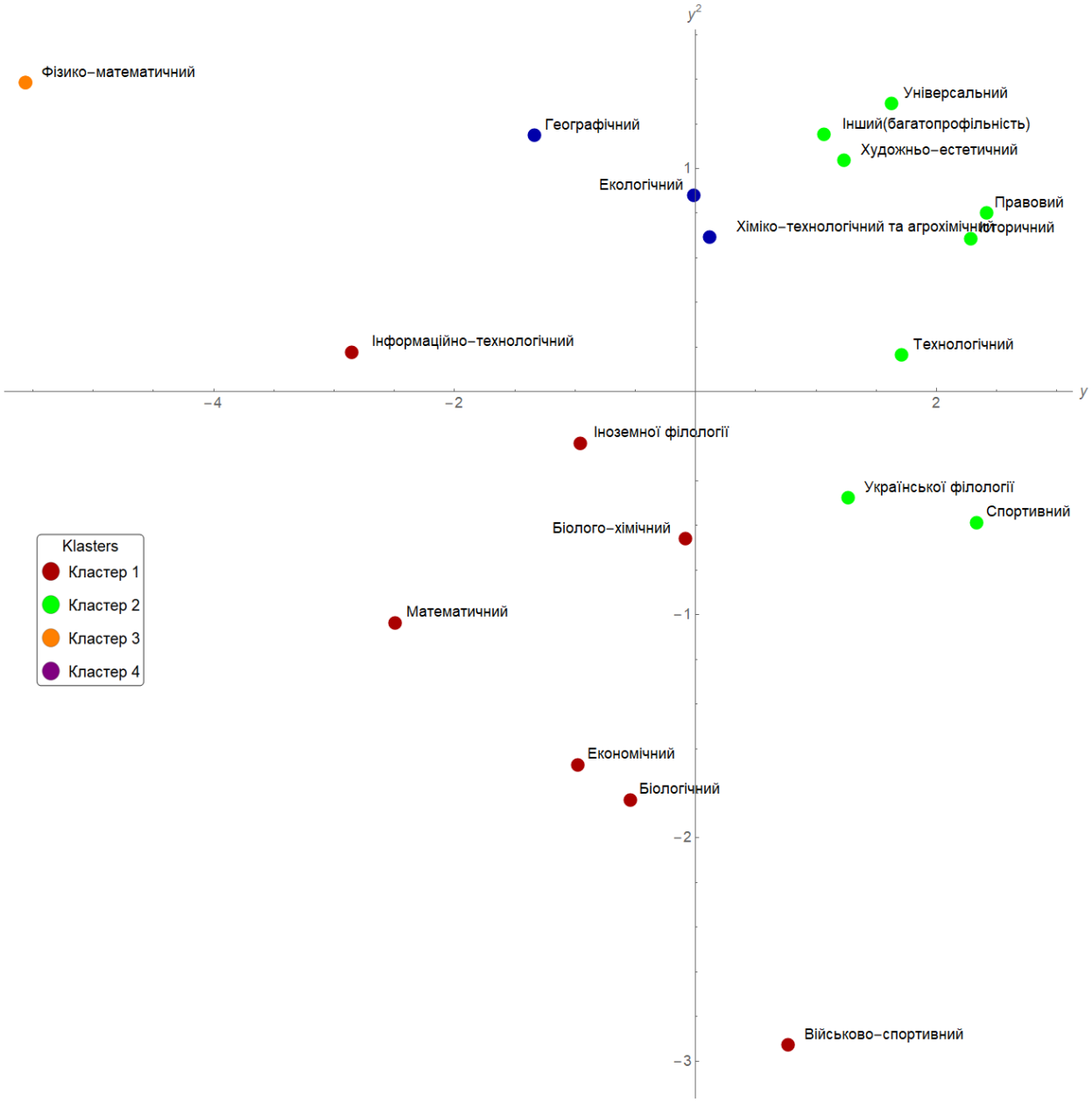

Рисунок 4.16 – Класифікація профілів навчання за результатами ЗНО з математики

Отже, спираючись на результати досліджень, можна с певністю сказати, що методи машинного навчання дійсно підходять для розв’язання задачі дослідження результатів ЗНО та виділення схожих за якістю підготовки ЗСО. За рахунок застосування методів зниження вимірності результати вийшли наочними, що робить їх корисними для практичного застосування.

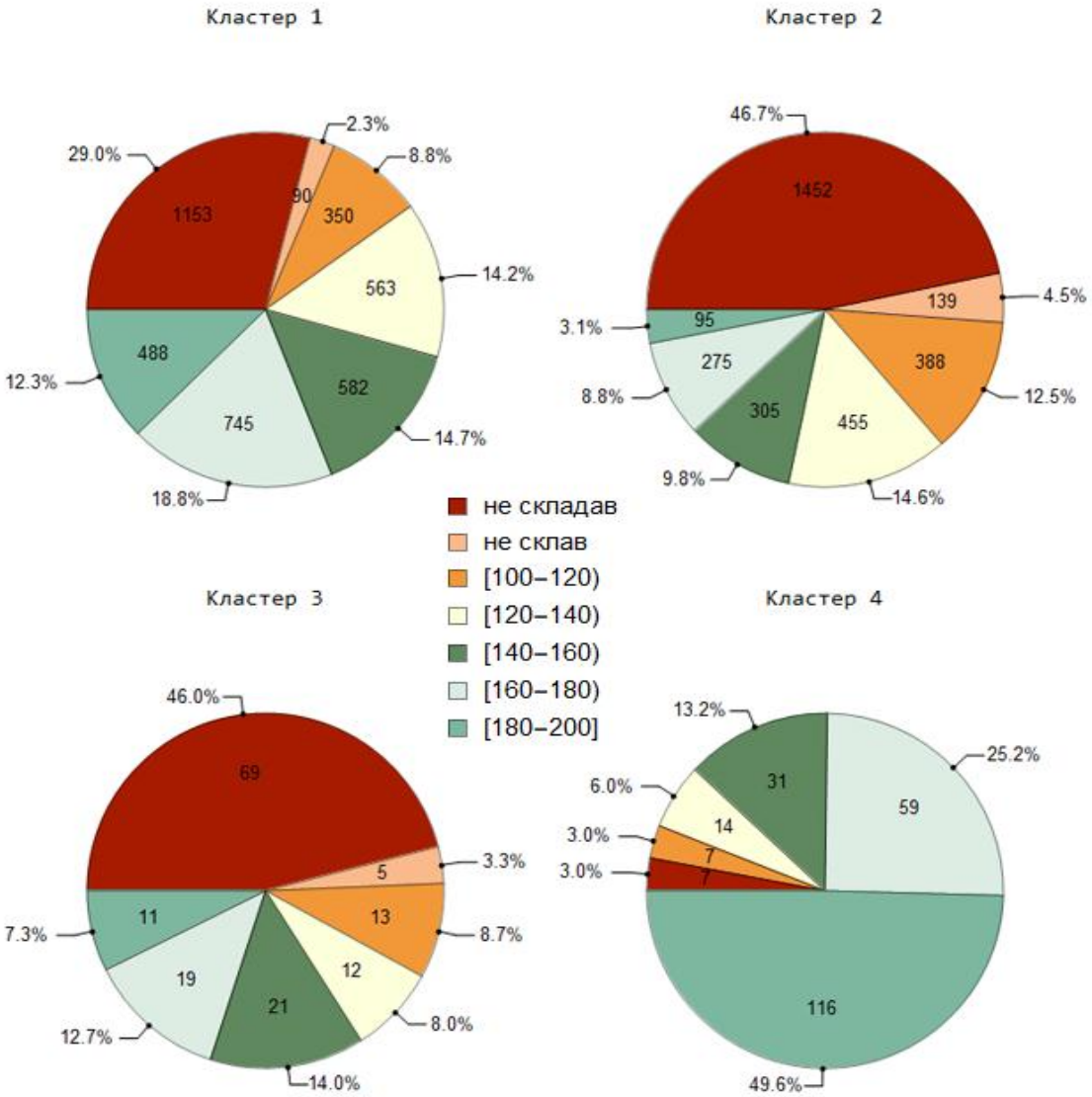

Рисунок 4.17 – Діаграма розподілу результатів ЗНО
з математики у розрізі профілів навчання

5 АНАЛІЗ МОЖЛИВИХ ЗАСТОСУВАНЬ

Актуальність розглянутої у атестаційній роботі задачі полягає у наступному: по-перше, методи машинного навчання з кожним роком набираюсь популярність у дослідженнях статистичного профілю, і у даній роботі наведено приклад використання сумісної роботи двох з цих методів, а саме, методу головних компонент та методу k-середніх, для дослідження результатів ЗНО за багатьма критеріями; по-друге, отримані у роботі результати можуть бути розширені та використані для більш поглибленого аналізу, а також на інших масивах даних.

Також варто зауважити, що результати даної роботи можуть бути використані для організації профорієнтаційної роботи закладами середньої освіти на наступний рік.

Методи, що описувалися у даній атестаційній роботі, можуть бути застосовані до дослідження інших сфер, таких як економіка, політологія, екологія, у багатьох прикладних задачах.

Розроблена програма повністю готова до використання, гнучка до змін формату даних та може з легкістю використовуватися для інших задач. Для модифікування програми не потребується спеціальних навичок, усе було описано на інтуїтивно-зрозумілому вигляді.

ВИСНОВКИ

Результатом виконання атестаційної роботи стало розв’язання задачі дослідження результатів зовнішнього незалежного оцінювання випускниками закладів середньої освіти м. Харкова у 2020 році методами машинного навчання, зокрема, методом головних компонент та методом k-середніх.

Використовуючи СУРБД MySQL та пакет Mathematica, було сформовано вибірку для аналізу та розроблено програмний продукт для обробки і аналізу багатовимірного масиву результатів зовнішнього незалежного оцінювання випускників міста Харкова за 2020 рік. Отримані результати дослідження у вигляді графіків були проаналізовані у розділі 4.

Спираючись на аналіз результатів з розділу 4, можна виявити декілька переваг та недоліків дослідження у цілому.

До переваг можна віднести швидке отримання результатів та простоту реалізації процедури розв’язання. Також результати повторюють тенденції з даного питання, наявні у відкритих джерелах, та є наочними, що дозволяє інтерпретувати їх для своїх цілей.

Недоліками є те, що чим більше об’єктів для аналізу буде обрано, тим менш наочною буде графічна реалізація. Також правила складання ЗНО змінюються майже кожен рік, а також ЗВО можуть змінювати свої правила до вступу, тому досліджувані результати повинні бути свіжими і актуальними, тобто використання застарілих торішніх даних може знизити точність висновків. І ще одним недоліком є потреба у повноті та однорідності даних. Тобто у нашому випадку більш-менш достовірно оцінити якість освіти у ЗСО з певного предмету можна, тільки якщо більшість учнів складали цей тест, а це неможливо. Такий випадок ми бачили на прикладі результатів з фізики та хімії.

Результати, отримані у роботі, можуть бути використані ЗВО для організації профорієнтаційної роботи з закладами середньої освіти. Використовуючи результати, можна скорегувати профорієнтаційну роботу та сконцентруватися на певному кластері ЗСО, що є найбільш профільним с точки зори конкретного ЗВО.

Методи, що використовувалися у даній атестаційній роботі, можуть бути застосовані у дослідженнях у інших галузях, таких як економіка, політологія, екологія у багатьох прикладних задачах.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Антонов А. В. Системный анализ. Москва : Высш. шк., 2004. 454 с.

2. Малищак Т. О. Дослідження результатів зовнішнього незалежного оцінювання методом компонентного аналізу // 23-й Міжнародний молодіжний форум «Радіоелектроніка та молодь у XXI столітті» : зб. матеріалів форуму (м. Харків, 16-18 квітня 2019 р.). Т. 7. Харків : ХНУРЕ, 2019. С. 136-137.

3. Прикладная статистика: Классификация и снижение размерности / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин. Москва : Финансы и статистика, 1989. 607 с.

4. Айвазян С. А., Бежаева З. И., Староверов О. В. Классификация многомерных наблюдений. Москва : Статистика, 1974. 240 с.

5. Мандель И. Д. Кластреный анализ. Москва : Финансы и статистика, 1988. 176 с.

6. Ким Дж. О., Мюллер Ч. У., Клекка У. Р. Факторный, дискриминантный и кластерный анализ. Москва : Финансы и статистика, 1989. 215 с.

7. Український центр оцінювання якості освіти. URL : http://testportal.gov.ua/ofzvit/ (дата звернення: 09.09.2020).

8. Статистика ЗНО 2020. URL : https://zno-2020.monitoring.in.ua (дата звернення: 15.09.2020).

9. Четвериков М. А. Применение средств Wolfram Mathematica для создания интерактивных иллюстраций // Молодой ученый. 2013. №8. С. 62-66.

10. Correlation between volatile composition and sensory properties in Spanish Albariño wines // Microchemical Journal. 2010. Vol. 95. Iss. 2. P. 240-246.