УДК 519.237.8 : 004.85 DOI: https://doi.org/10.30837/IYF.CVSAMM.2024.307

МЕТОДИ СТАТИСТИЧНОГО АНАЛІЗУ РЕЗУЛЬТАТІВ ЗОВНІШНЬОГО НЕЗАЛЕЖНОГО ОЦІНЮВАННЯ

Шарай К.В.
Науковий керівник – канд. техн. наук, доц. Гибкіна Н.В.
Харківський національний університет радіоелектроніки, каф. ПМ,
м. Харків, Україна
e-mail: kateryna.sharai@nure.ua

This work considers the application of cluster analysis methods, in particular, the k-means method and its modifications to consolidate the Kharkiv secondary education institutes into groups and compare them by the results of External Independent Evaluation. The results of clustering will be presented in graphic form so the principal components method will be used for this purpose. The obtained findings will be meaningful for solving application tasks related to the field of education, for example, for the estimation how quality the educational preparation in different disciplines is and revealing what changes are needed to improve the level of training in every distinct school.

Одним з найефективніших підходів до роботи з великими даними є використання методів багатовимірного статистичного аналізу та автоматизація цих методів за допомогою технологій машинного навчання.

Розглянемо застосування таких методів статистичного аналізу як метод головних компонент та ітеративні методи кластеризації для дослідження результатів складання зовнішнього тестування (зовнішнє незалежне оцінювання – ЗНО) учнями закладів середньої освіти (ЗСО) [1]. Актуальність дослідження випливає з потреби у контролі рівня освітньої підготовки учнів та його покращення. Моніторинг результатів ЗНО із застосуванням методів аналізу даних та машинного навчання допоможе розділити аналізовані ЗСО на групи, у кожній з яких ЗСО характеризуються подібним рівнем якості навчання за окремими предметами та профільними предметами, популярністю того чи іншого предмета серед учнів для складання ЗНО. Проаналізувавши структуру кластерів та виявивши загальні тенденції, за якими їх було сформовано, можна пропонувати заходи щодо підвищення якості навчання, а також підтримки наявних високих результатів. Наприклад, доцільним буде створити план для більш ефективної підготовки учнів з предметів, за якими переважають низькі бали, або створити профільні класи з посиленим вивченням окремих дисциплін чи перепрофілювати навчальний заклад повністю. Для закладів вищої освіти використання результатів аналізу дозволить визначати ЗСО, більша частина випускників яких отримує високі результати ЗНО за профільним для вступу предметами.

У дослідженні розглядається набір даних, що містить відомості про результати ЗНО випускників ЗСО Харкова за 2020–2021 навчальний рік. Їх подано у вигляді таблиці, що складається з 234 записів, кожен з яких пред-

ставляє окремий ЗСО. Для кожного об'єкта (ЗСО) наводяться значення наступних ознак: назва та тип ЗСО; число осіб, що складали іспит з дисципліни; відсоткові показники числа випускників, які отримали: менше 100 балів, 100-120, 120-140, 140-160, 160-180, 180-200 балів; а також тих, хто не складав ЗНО з даного предмета. Задача полягає в обробці наявних даних та їх статистичному аналізі з метою виявлення закономірностей в оцінках випускників різних ЗСО за окремими предметами.

Попередньо здійснюється підготовча обробка вихідних даних. Вона складається з наступних етапів: масштабування ознак (нормалізація чи стандартизація для їх зведення до єдиної шкали вимірювання), поділ масиву на тренувальну та тестову вибірки, відбір ознак з метою зменшення розмірності датасету. Для візуального подання результатів кластеризації в подальшому використовуються 2 агреговані ознаки об'єктів, отримані за допомогою методу головних компонент. Після завершення попередньої обробки отримуємо дані, готові до кластерного аналізу.

Сутність кластеризації полягає в тому, щоб розділити набір даних на групи, в яких міститимуться подібні об'єкти, а несхожі об'єкти належатимуть різним кластерам [2]. Схожість об'єктів визначається відстанню між ними в сенсі обраної метрики. В якості міри відстані d між окремими об'єктами використовуватимемо евклідову метрику $d(\vec{x},\vec{x}') = \sqrt{\sum_{i=1}^{n}(x^{(i)} - x'^{(i)})^2}$, де $x^{(i)}$, $x'^{(i)}$ – значення i-ї ознаки (атрибута) об'єктів $\vec{x}$ та $\vec{x}'$, $i = \overline{1,n}$.

Метою кластеризації є побудова відображення $f : X \to Y$, яке кожному об'єкту з X ставить у відповідність мітку одного з кластерів $y^{(j)}$, $y^{(j)} \in Y$, де Y – множина всіх кластерів. Кінцевим результатом буде розділений на m груп вихідний масив даних, тобто матимемо розбиття аналізованих ЗСО на m кластерів, що сформувалися на основі об'єднання закладів з певною спільною закономірністю. У дослідженні пропонується використовувати ітеративні методи кластеризації, зокрема, k-середніх, k-середніх++, k медіан. Кількість кластерів m можна визначити за допомогою методу ліктя.

Проаналізувавши склад кластерів, а також тенденції, що простежуються в кожному з них, можна отримати важливі для практичної діяльності висновки щодо якості напрямів освітньої підготовки у ЗСО.

Список використаних джерел:

1. Гибкіна Н.В., Сидоров М.В. Статистичний аналіз результатів зовнішнього незалежного оцінювання у м. Харкові за 2019 рік. *Радіоелектроніка та інформатика*. 2020. № 2 (89). С. 11–22.

2. Raschka S., Mirjalili V. Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow. Packt Publishing, 2017. 622 p.