УДК 622.17: 004.383.8
DOI https://doi.org/10.32782/IT/2024-4-11

Тімур ЖЕЛДАК
кандидат технічних наук, доцент, завідувач кафедри системного аналізу та управління, Національний технічний університет «Дніпровська політехніка», просп. Дмитра Яворницького, 19, Дніпро, Україна, 49005
***ORCID:** 0000-0002-4728-5889*
***Scopus Author ID:** 55602208300*

Олександр ВЛАДИКО
кандидат технічних наук, доцент кафедри системного аналізу та управління, Національний технічний університет «Дніпровська політехніка», просп. Дмитра Яворницького, 19, Дніпро, Україна, 49005
***ORCID:** 0000-0001-9779-9565*
***Scopus Author ID:** 55522741200*

Бібліографічний опис статті: Желдак, Т., Владико, О. (2024). Використання машинного навчання для кластеризації з використанням індексу небезпеки хвостосховищ та їх впливу на навколишнє середовище. *Information Technology: Computer Science, Software Engineering and Cyber Security*, 4, 81–91, doi: https://doi.org/10.32782/IT/2024-4-11

ВИКОРИСТАННЯ МАШИННОГО НАВЧАННЯ ДЛЯ КЛАСТЕРИЗАЦІЇ З ВИКОРИСТАННЯМ ІНДЕКСУ НЕБЕЗПЕКИ ХВОСТОСХОВИЩ ТА ЇХ ВПЛИВУ НА НАВКОЛИШНЄ СЕРЕДОВИЩЕ

В статті виконано аналіз небезпечного впливу хвостосховищ утворених від промислового виробництва на навколишнє середовище методами машинного навчання з метою їх кластеризації. При різних розмірах хвостосховищ, їх токсичності та параметрів була визначена їх положення в кластерах, що дозволило класифікувати стан рівня їх небезпеки. Зроблено порівняльний аналіз підходів машинного навчання та інших методів пошуку аналогів для визначення потенційної небезпеки хвостосховищ. За допомогою методу DBSCAN було визначено хвостосховища, які за своїми параметрами потрапляли до кореневого, граничного та шумового кластерів. Всі хвостосховища, що потрапили до кореневого та граничного кластерів, було класифіковано наступним методом кластеризації – k-середніх, і вони були розподілені до одного з чотирьох кластерів. Кластеризація методом k-середніх будувалася на вибірках даних хвостосховищ шляхом порівняння метрики відстані між витягнутими в одновимірний вектор матрицями суміжності. Наведено результати оцінки роботи алгоритму, що підтверджують, що моделювання та пошук центрів кластеризації за допомогою k-середніх є більш комплексним рішенням задачі в порівнянні з методом агломеративної кластеризації, який було приведено до порівняння k-середніх. Таким чином точність використаного підходу кластеризації виявився вищою, ніж у існуючих класичних алгоритмах кластеризації.

***Метою роботи** є кластеризація хвостосховищ, що утворені від діяльності промислового виробництва, за рівнем їх небезпечного впливу на навколишнє середовище з використанням методів машинного навчання.*

***Методологія.** Для досягнення мети використані методи машинного навчання та аналізу даних, що реалізуються в рамках фреймворку Sklearn мови програмування Python, використаний пакет роботи з даними pandas, що дозволяє обробляти великі масиви даних з використанням гнучкої системи форматування та запитів до них. Групування даних здійснювалося з використанням групи методів навчання без вчителя (unsupervised learning): алгоритмів кластеризації DBSCAN, кластерного аналізу методом ієрархічної деревоподібної кластеризації та методом k-середніх, методу головних компонентів.*

***Наукова новизна.** Наукова новизна полягає у виборі способів кластеризації хвостосховищ за допомогою методів кластерного аналізу.*

***Висновки.** За допомогою методів машинного навчання виконана кластеризація хвостосховищ, в залежності від ступеня їх небезпеки для навколишнього середовища.*

***Ключові слова:** машинне навчання, кластеризація, метод k-середніх, хвостосховища, метод DBSCAN, індекс небезпеки, метод агломеративної кластеризації, навколишнє середовище.*

Timur ZHELDAK
Candidate of Technical Sciences, Associate Professor, Head of the System Analysis and Control Department, Dnipro University of Technology, 19, Dmytra Yavornytskoho Ave., Dnipro, Ukraine, 49005, zheldak.t.a@nmu.one
***ORCID:** 0000-0002-4728-5889*
***Scopus Author ID:** 55602208300*

Alexander VLADYKO
Candidate of Technical Sciences, Associate Professor at the Department of System Analysis and Management, Dnipro University of Technology, 19, Dmytra Yavornytskoho Ave., Dnipro, Ukraine, 49005, vladyko.o.b@nmu.one
ORCID: *0000-0001-9779-9565*
Scopus Author ID: *55522741200*

To cite this article: Zheldak, T., Vladyko A. (2024). Vykorystannia mashynnoho navchannia dlia klasteryzatsii z vykorystanniam indeksu nebezpeky khvostoskhovyshch ta yikh vplyvu na navkolyshnie seredovyshche [Using machine learning for clustering using the hazard index of tailings and their environmental impact]. *Information Technology: Computer Science, Software Engineering and Cyber Security*, 4, 81–91, doi: https://doi.org/10.32782/IT/2024-4-11

USING MACHINE LEARNING FOR CLUSTERINGUSING THE HAZARD INDEX OF TAILINGSAND THEIR ENVIRONMENTAL IMPACT

The article analyzes the hazardous impact of industrial tailings on the environment using machine learning methods for their clustering. With different sizes of tails, their toxicity and parameters, their location in clusters was determined, which made it possible to classify the condition according to the degree of their danger. A comparative analysis of machine learning approaches and other methods of finding analogues for determining the potential danger of tailings storage facilities was carried out. Using the DBSCAN method, the tails were identified, which according to their parameters fell into the root, marginal, and noise clusters. All tails that fell into the root and border clusters were classified by the following clustering method – k-means, and they were allocated to one of four clusters. K-means clustering was constructed on tail data samples by comparing the distance metric between adjacency matrices compiled into a one-dimensional vector. The results of the evaluation of the algorithm are given, which confirm that the modeling and search for clustering centers using k-means is a more difficult solution to the problem compared to the method of agglomerative clustering, which was reduced to the comparison of k-means. Thus, the accuracy of the used clustering approach turned out to be higher than that of existing classical clustering algorithms.

The purpose *of the work is the clustering of tailings, formed by the activity of industrial production, according to the level of their dangerous impact on the environment using machine learning methods.*

The methodology. *To achieve the goal, the methods of machine learning and data analysis implemented within the Sklearn framework of the Python programming language were used, and the pandas data package was used, which allows processing large data sets using a flexible system of formatting and queries to them. Data grouping was carried out using a group of unsupervised learning methods: DBSCAN clustering algorithms, cluster analysis using the hierarchical tree clustering method and k-means method, principal component method.*

The scientific novelty. *The scientific novelty consists in the choice of methods of clustering of tailings storage facilities using methods of cluster analysis.*

Conclusions. *With the help of machine learning methods, clustering of tailings storage facilities was performed, depending on the degree of their danger to the environment.*

Key words: *machine learning, clustering, k-means method, tailings, DBSCAN method, hazard index, agglomerative clustering method, environment.*

Актуальність проблеми. Аварії на хвостосховищах у світі за останні три десятиліття показали, що вони призводять до великих катастроф, негативно впливають на здоров'я людей і спричиняють різноманітну шкоду навколишньому середовищу (Nikolaieva, 2015). Незважаючи на те, що умови безпеки значно покращилися, завдяки суворим вимогам і технічним заходам, безпека певної кількості хвостосховищ все ще залишається на незадовільному рівні через нестачу грошей для їх постійного захисту.

Крім того, очікується різке збільшення кількості хвостосховищ внаслідок збільшення видобутку кольорових металів, оскільки розумні та передові технології спричиняють різке зростання попиту на метали (кобальт, мідь, літій, нікель та ін.).

Як відомо видобуток корисних копалин супроводжується значним потоком відходів, що утворюються в результаті такої діяльності. Одним із багатьох типів відходів гірничодобувної промисловості є рідкі відходи, шлами, хвости, дрібнозернисті відходи, отримані з гірничо-збагачувальних фабрик, які потім зберігаються певним чином. В ідеальних умовах хвостосховища повинні забезпечувати безпечне тривале зберігання отриманих відходів. Однак на хвостосховищах можуть виникати аварії через несприятливі природні умови, недоліки проектування та будівництва, невідповідну практику експлуатації та управління.

Таким чином, суспільство може згодом зіткнутися зі зростаючим ризиками аварій на хвостосховищах із потенційними жертвами

та екологічними збитками. Якщо безпекою хвостосховища не почати керувати належним чином, відповідно до нових технологічних стандартів з урахуванням впливу водного балансу супутніх річок, тоді можуть виникати додаткові непередбачувані наслідки на хвостосховищах. У зв'язку з цим у керівництва країнами зростає занепокоєння щодо погіршення стану навколишнього середовища, спричиненого неконтрольованими великомасштабним переміщеннями небезпечних матеріалів, пов'язаних із аваріями на хвостосховищах. Такі аварії можуть призвести до непередбачуваних розливів і викидів небезпечних матеріалів, що зберігаються на хвостосховищах у навколишнє середовище. Це забруднення може мати серйозні наслідки та спричиняти збитки для здоров'я людини, побудованої інфраструктури, економічної діяльності, і викликати негативний вплив на довкілля та природні ресурси. Викиди з хвостосховищ дуже часто мають негативний широкомасштабний або транскордонний вплив на ресурси довкілля. Крім того, аварії на хвостосховищах можуть призвести до довгострокового забруднення води та ґрунту, мати негативні хронічні та накопичувальні наслідки для здоров'я людини та навколишнього середовища.

Тому особливої актуальності набувають питання створення для хвостосховищ адекватних умов та заходів безпеки разом з прогнозування вірогідних аварій. Аварії в на хвостосховищах як в Європейському союзі так і в Україні яскраво продемонстрували, наскільки серйозним може бути вплив неправильної їх експлуатації на людей, навколишнє середовище та водні ресурси. Ці події вимагають розробки та впровадження нових методів аналізу хвостосховищ і нових управлінських рішень по недопущенню випадків аварії.

Одним із сучасних методів прогнозування небезпеки хвостосховищ є методи штучного інтелекту до яких відноситься машинне навчання, яке дозволяє розділити хвостосховища на кластери по їх небезпечності та прогнозувати їх стан на певний проміжок часу.

Аналіз останніх досліджень і публікацій. Для кластеризації хвостосховищ активно використовуються підходи, що ґрунтується на використанні технологій машинного навчання. Зокрема, можливе застосування різних алгоритмів кластеризації з метою отримання груп зі схожими параметрами хвостосховищ. Для визначення їх впливу на навколишнє середовище використовують моделі, що описують різні процеси в яких найчастіше застосовують ієрархічну агломеративну кластеризацію та кластеризацію методом *k*-середніх. Далі виконується навчання моделі, в результаті якого вихідна база даних поділяється на групи родовищ зі схожими властивостями (Astakhova, 2015).

За допомогою метода *k*-середніх можна виділити кластер з подібними параметрами хвостосховищ, де пропонується використовувати даний вид кластеризації, попередньо знизивши розмірність вихідної бази даних. Потім для оцінки якості отриманих кластерів з аналогами можна використати регресійні моделі. Проте слід зазначити, що ці підходи мають певні недоліки. Обидва види кластеризації виявляються досить чутливими до вхідних даних великої розмірності. Зниження розмірності вхідних даних може призвести до втрат вихідної інформації. Інший недолік – складність застосування таких підходів у завданнях, де потрібно відновити один або кілька параметрів досліджуваних хвостосховищ. Хвостосховище, частина параметрів якого пропущена, неспроможна брати участь у процесі навчання моделі, і потребує додаткової обробки щодо його кластера.

Таким чином, запропонований алгоритм повинен ефективно працювати як з безперервними, так і з дискретними параметрами необхідної розмірності. Виходячи з пред'явлених вимог, був зроблений вибір на користь реалізації із застосуванням технологій машинного навчання, таких як метод *k*-середніх, які дозволяють при середній обчислювальній складності та незначному застосуванні вхідних даних отримати хорошу інтерпретованість навченої моделі. За допомогою формування умовних розподілів ймовірностей у вузлах мережі з'являється можливість не лише оцінювати невідомі змінні, а й виконувати пошук аномальних значень та проводити аналіз достовірності та інформативності отриманих даних.

Мета дослідження: кластеризація хвостосховищ, що утворені від діяльності промислового виробництва, за рівнем їх небезпечного впливу на навколишнє середовище з використанням методів машинного навчання. Запропоновані методи машинного навчання дозволяють більш точно визначити хвостосховища з великим рівнем небезпеки для навколишнього середовища.

Виклад основного матеріалу. Машинне навчання є великим розділом в галузі вивчення штучного інтелекту, яке включає методи побудови різних алгоритмів, здатних до самонавчатися. Виділено три групи методів такого навчання, що часто використовуються: навчання з учителем (регресія, класифікація); навчання

без вчителя (пошук правил, зменшення розмірності, кластеризація) та навчання з підкріпленням (генетичний алгоритм, *Q*-learning та ін.) (IMES, 2019). З цих методів найкраще вирішує задачу визначення небезпеки впливу хвостосховищ на навколишнє середовище є кластеризація. В свою чергу в кластеризації застосовують наступні алгоритми і методи машинного навчання (Kravchenko, 2020):

– евристичні графові алгоритми (алгоритм виділення зв'язкових компонентів, алгоритм найкоротшого незамкнутого шляху, FOREL алгоритм);

– статистичні алгоритми, засновані на розбиття (метод *k*-середніх, алгоритм DBSCAN, що ґрунтується на щільності розподілів характеристик, які вивчаються);

– ієрархічні методи (агломеративні та дивізіонні: алгоритми CURE, ROCK, Chameleon, метод Варда);

– алгоритми нечіткої кластеризації (FCM, FCS та MM алгоритми).

Кожна група з цих методів кластеризації має свої переваги і недоліки. Зокрема, статистичні алгоритми, такі як метод *k*-середніх та алгоритм DBSCAN, засновані на розбитті, ефективно працюють із великими обсягами даних, що завжди можна застосувати для графових методів кластеризації.

Обробка статистичних даних для вирішення задач кластеризації з використанням індексу небезпеки хвостосховищ можливо за рахунок баз даних аварій на хвостосховищах, яку сформуємо з використанням бібліографічних та відкритих джерел. Для повноти аналізу аварій та проривів у навколишнє середовище на хвостосховищах необхідно мати статистику цих подій за 20 – 70 років, тому знайдено відповідну інформацію у відповідних джерелах. І це необхідна статистика за останні 60 років в Європі, що повністю покриває нашу потребу в статистиці на 100%. Отже, за цей період зареєстровано 323 аварії. Кількість відмов залишалася незмінною протягом перших трьох десятиліть, а потім зменшилася після 1990 року протягом двох десятиліть, імовірно, відображаючи скорочення видобувної діяльності в країнах колишнього Радянського Союзу. Однак за останнє десятиліття кількість відмов повернулася до рівня початку накопичення статистики. Тенденція до аварій зростає, тому необхідно вжити заходів, щоб уникнути велику їх кількість (Winkelmann-Oei G., 2015).

Проведемо оцінку індексу потенціалу небезпеки хвостосховища та індексу ризику відходів. В межах попередніх пілотних проектів Німецького агентства з навколишнього середовища була розроблена методологія хвостосховища для підтримки регіональної та місцевої оцінки безпеки. Яка включає оцінку потенціалу небезпеки на основі індексу для великої кількості хвостосховищ, так званий індекс небезпеки хвостосховищ *THI*, і детальний контрольний список для аналізу безпеки окремих хвостосховищ. Спираючись на сильні сторони методики, а також покращуючи та адаптуючи її на основі сучасних технічних знань і найкращих доступних технологій отримуємо набір практичних інструментів для покращення умов безпеки хвостосховищ. *THI* вже довів свою корисність у спрямуванні обмежених фінансових і кадрових ресурсів країни на хвостосховища, що представляють найвищий потенціал небезпеки. Критерії, що лежать в основі підходу *THI*, були покращені завдяки використанню результатів історичного аналізу відмов хвостосховища. Оскільки *THI* враховує лише потенціал небезпеки, потенційні наслідки окремих аварій хвостосховищ, що становлять різні загрози для навколишнього середовища та населення, не враховуються. Цю проблему було вирішено шляхом визначення зони потенційного ризику поблизу хвостосховища на основі розмірів попередніх аварій для оцінки навколишнього середовища (водної екосистеми) та населення, що перебувають у зоні ризику. Результат цього підходу перетворює *THI* на індекс ризику хвостосховища – *TRI*, який навіть краще відображає найнебезпечніші хвостосховища в одній країні з точки зору потенційно постраждалого населення та навколишнього середовища (https://www.sendaiplatform.org).

За допомогою цього методу велика кількість хвостосховищ може бути відсортована та розставлена за пріоритетністю відповідно до розрахованого потенціалу небезпеки.

Метод *THI* враховує такі параметри, які були визначені як найважливіші:

– загальна місткість хвостосховища;

– токсичність речовин хвостів, що зберігаються;

– статус управління хвостосховищем;

– природні умови, характерні для ділянки хвостосховища;

– параметри безпеки дамби.

Для підготовки даних для машинного навчання розкриємо більш детально ці параметри.

Загальна місткість хвостосховища THI_{Cap}. Припускається, що цей параметр зростає зі збільшенням об'єму за логарифмічним співвідношенням з основою 10. Таким чином, збільшення об'єму хвостосховищ у 10 разів буде

означати збільшення індексу на одиницю, що розраховується за формулою $THI_{Cap} = Log_{10}[V_t]$, де V_t – загальний об'єм хвостосховищ у хвостосховища, м^3.

Токсичність речовин хвостів, що зберігаються THI_{tox}. Для інтегрованої характеристики токсичності вкрай важливо мати параметр, який представляє всі потенційні загрози для водної екосистеми в короткостроковій і довгостроковій перспективі. Методи об'єднують всі потенційні загрози для водних екосистем, включаючи гостру та хронічну токсичність, а також біологічні накопичення та накопичує небезпеки для різних організмів (риб, ракоподібних, бактерій та ін.). Ці дані доступні в мережі Інтернет для, приблизно, 7000 речовин (https://www.sendaiplatform.org).

Статус управління хвостосховища **–** це статус хвостосховища, який слід ідентифікувати з чотирьох варіантів. Статистика аварій на хвостосховищах (Rico et al., 2008a, 2008b) показує, що закриті та відновлені хвостосховища безпечніші з точки зору частоти аварій. На цих хвостосховищах аварій не зафіксовано. З цієї причини передбачається, що параметр, пов'язаний з управлінням хвостосховищами, є нижчим для закритих або реконструйованих споруд порівняно з активними хвостосховищами.

Природні умови, характерні для ділянки хвостосховища. Параметр «Природні умови» THI_{Nat} пов'язаний з екологічними ризиками, які дуже часто пов'язані з аварією хвостосховища. Особливо землетруси, сильні зливи та повені багато разів класифікувалися як причини аварій на хвостосховищах. Тому, відповідний потенціал небезпеки розраховується за таким рівнянням: $THI_{Nat} = THI_{Seism} + THI_{Flood}$, де THI_{Seism} – індекс небезпеки для сейсмічної активності, а THI_{Flood} – індекс небезпеки для затоплення на основі геологічних і гідрологічних умов майданчика хвостосховища. Вплив повеней THI_{Flood} пов'язаний із зонами, схильними до повеней, зі статистичним параметром *HQ*-500, який кількісно визначає частоту повеней із повторюваністю один раз у п'ятсот років.

Параметри безпеки дамби. Стійкість дамби є, мабуть, найбільш критичним параметром у оцінці небезпеки. Вважається, що параметр «Умови греблі» THI_{Dam} пов'язаний із проектним параметром греблі «Коефіцієнт безпеки» *FoS*, який має бути розрахований вже на етапі проектування хвостосховища та відноситься до стійкості схилу греблі (Coduto, 1998; Круз та ін., 2008; Фредлунд та ін., 2012). Термін *FoS* зазвичай використовується для вираження запасу міцності схилів на насипних дамбах.

Метод індексу ризику відходів *TRI*. Оцінка *TRI* враховує загальний потенціал небезпеки, а також населення та водні тіла нижче за течією як потенційні об'єкти ризику впливу в разі аварії.

Оскільки соціально-економічні цінності ризику та вразливості потенційних рецепторів можна оцінити лише шляхом детальної оцінки, підхід *TRI* не включає ці аспекти. Будь-яка подальша детальна оцінка ризику для окремих хвостосховища для підтримки планування на випадок надзвичайних ситуацій або специфічної оцінки безпеки повинна включати більш конкретні аспекти та інформацію безпосередньо на ділянці та навколо неї.

TRI розраховується на основі значень *THI* і *TEI* за є їх додатком, тому розрахунок виглядаю так: *TRI* = *THI* + *TEI*. Подібно до *THI*, *TEI* та *TRI* також слід оцінювати за логарифмічною шкалою і визначається за такою формулою: $TEI = TEI_{Pop} + TEI_{Env}$. Параметр TEI_{Pop} є фактором, що враховує населення нижче за течією, розташоване до 10 км від хвостосховища *PAR*. Коефіцієнт TEI_{Pop} визначається простою класифікацією.

Вплив на навколишнє середовище TEI_{Env} – це коефіцієнт, який враховує розмір найближчої до хвостосховища водойми, яка розташована нижче за течією в межах 10 км від хвостосховища і може бути забруднена аварією на хвостосховища. Коефіцієнт TEI_{Env} визначається на основі середнього значення річкового стоку або площі поверхні озера.

Отже вхідний набір даних за 2022 рік містить 346 записів за 13 показниками хвостосховищ Європи. Вибрані перших п'ять даних для аналізу та кластеризації регіонів показники наведено у таблиці 1.

На цьому формування вхідних даних для проведення кластеризації з використанням індексу небезпеки хвостосховищ є достатнім. Тому перейдемо до самих методів, що забезпечують таке визначення.

Наступним виконаємо кластеризацію з використанням індексу небезпеки хвостосховищ методами машинного навчання. Незважаючи на різні підходи до вирішення завдань кластеризації, всі методи ґрунтуються на поданні об'єктів, які пов'язані між собою, що дозволяє оцінити їх близькі властивості. Тому при простій кластеризації об'єктів за декількома числовими параметрами найбільше поширеними у спеціалізованому програмному забезпеченні залишаються ієрархічна кластеризація та метод *k*-середніх, що відрізняється відносною простотою, високою якістю одержуваних результатів, їх інтерпретованістю та широкими можливостями

Таблиця 1

Вхідні дані хвостосховищ на основі індексу потенціалу небезпеки та оцінки ризиків

код	Склад відходів	Індекс небезпеки хвостосховищ							Індекс ризику відходів			
		THI_{Cap}	THI_{Tox}	THI_{Man}	THI_{Nat}	THI_{Dam}		THI	TEI_{Pop}	TEI_{Env}	TEI	TRI
					THI_{Seism}	THI_{Flood}	THI_{Dam}					
AT1	шлам	6,26	1,00	3,00	0,00	0,00	1,00	11,26	4,00	2,00	6,00	17,26
AT2	шлам	5,70	1,00	0,00	1,00	1,00	1,00	9,70	4,00	3,00	7,00	16,70
AT3	шлам	5,48	3,00	0,00	1,00	1,00	1,00	11,48	4,00	2,00	6,00	17,48
AT4	шлам	5,70	2,00	0,00	1,00	0,00	1,00	9,70	4,00	2,00	6,00	15,70
AT5	шлам	5,45	1,00	0,00	1,00	0,00	1,00	8,45	4,00	2,00	6,00	14,45

з налаштування правил групування. Відмінність даних методів полягає лише в тому, що перші починають алгоритм *n* елементів (класів) і далі об'єднують близькі на відстані групи об'єктів, поки не залишиться всього один клас, а другі навпаки починають алгоритм з одного класу та поділяють далекі групи, доки буде досягнуто групування *n* об'єктів на *n* класів. З чого стає зрозуміло, що до беззаперечних переваг цього методу належить можливість побудови дендограм, тобто дерев, на яких чітко видно етапи класифікації та відстань між класами.

Розглянемо алгоритм DBSCAN для кластеризації хвостосховищ. Як відомо, цей алгоритм заснований на щільності для заданої множини точок у деякому просторі, який відносить в одну групу точки, що розташовані найбільш щільно (точки з багатьма сусідами) та розмічає точки, які лежать в областях з малою щільністю (чиї сусіди розташовані занадто далеко) як викиди (шум), тому визначимо епсилон-околицю об'єкта, де для будь-якого вектору *x* в метричному його просторі ознак визначається як безліч точок, що віддаляються від *x* не більше ніж на відстань *ε*, тому отримаємо вираз: $U_z(x)=\{u\in U:\rho(x,u)\le\varepsilon\}$, де $\rho(x, u)$ – евклідова відстань, як метрика простору ознаки. З виразу зрозуміло, що параметр $\varepsilon > 0$, тому задаємо його як параметр роботи алгоритму DBSCAN. Далі, на основі отриманої епсилон-околиці об'єктів згрупуємо їх на три типи: кореневий (що містить не менше *m* об'єктів у своїй епсилон-околиці, $|U_z(x)|\ge m$); граничний (не кореневий, але в околиці кореневого); шумовий (не кореневий і не граничний). З чого зрозуміло, що даний алгоритм заснований на евристиках, а не на математиці, де розрахунки більш точні та чіткі. При виконанні алгоритму отримаємо результати кластеризації, що зазначені на рис. 1, де зеленим кольором позначено основний кластер, синім та голубим – граничний та червоним – шумовий.

Аналізуючи отримані результати, можна зробити висновок, що цей алгоритм дозволяє нам отримати необхідне рішення в межах розглянутого завдання та виділити 5% шумових точок. Однак алгоритм стає нестійким і час від часу видаляє частину вибірки, вважаючи визначені точки шумом. Тому цей алгоритм може застосовуватися для рішення розглянутої задачі за умов, що у ньому зазначено, що точка не повинна мати сусідів для того, щоб потрапити в кластер, який допомагає розділити основні кластери та шумові точки.

Наступним розглянемо агломеративну ієрархічну кластеризацію, яка є сукупністю алгоритмів упорядкування даних, вкладених у створення ієрархії, що складається з груп точок, що спостерігаються. Вихідними для проведення кластерного аналізу служить матриця відстаней між об'єктами, сформована з використанням тієї чи іншої метрики. Агломеративна кластеризація починається з *n* кластерів, де *n* – число спостережень і передбачається, що кожне з них є окремим кластером. Потім алгоритм намагається знайти і згрупувати найбільш схожі між собою точки даних та з цього починається формування груп.

Отже, маємо набір даних і кожен об'єкт – це незалежний кластер. Потім, відповідно до заданої метрики між об'єктами: $\rho(x_i, x_j)$, $i, j = 1, \ldots, l$, вибираємо два найближчих таксони U, V і об'єднуємо їх в один W: $W = U \cup V$. Тепер у нас з'явилася група з двох об'єктів. Тут є безліч евристики: відстань ближнього сусіда, відстань далекого сусіда, групова середня відстань, відстань між центрами, відстань Уорда. Так от, виявляється, всі ці варіанти можна описати одним виразом: $R_{WS} = \alpha_U R_{US} + \alpha_V R_{VS} + \beta R_{UV} + \gamma| R_{US} - R_{VS}|$, де α_U, α_V, β, γ – деякі параметри, що визначають вид метрики між кластерами (Wozniak M., 2021).

Цей вираз називається формулою Ланса-Уільямса і для зазначених способів обчислень відстаней між таксонами, коефіцієнти набувають вигляду:

$$\alpha_U = \frac{|S|+|U|}{|S|+|W|},\ \alpha_V = \frac{|S|+|V|}{|S|+|W|},\ \beta = \frac{-|S|}{|S|+|W|},\ \gamma=0. \quad (1)$$

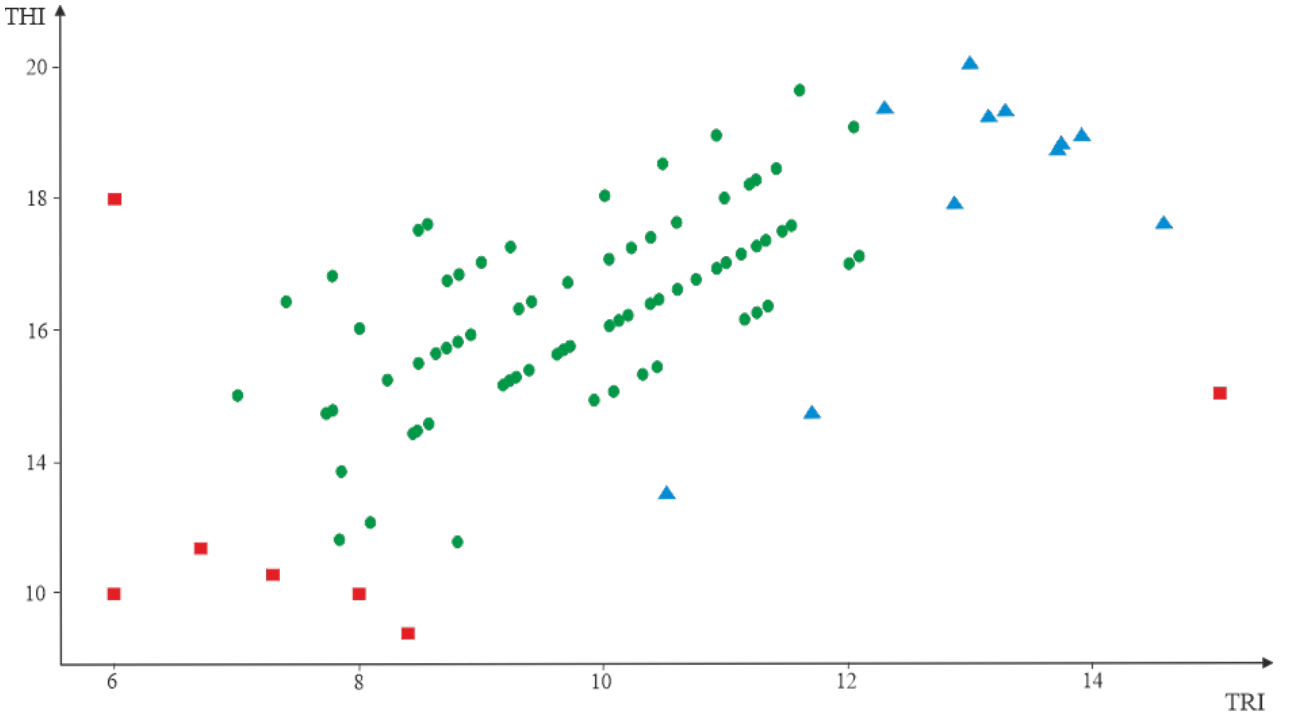

Рис. 1. Візуалізація кластеризації даних з допомогою алгоритму DBSCAN

Для проведення агломеративної ієрархічної кластеризації необхідно виконати наступний порядок дій. По перше, визначимо кількість кластерів, де $C_1=\{\{x_i\},\ldots,\{x_j\}\}$, – безліч одноелементних кластерів, $R\{x_i\}\cdot\{x_j\}= \rho(x_i, x_j)$ – метрика (відстань) між окремими об'єктами для всіх t = 2, …, l, де t – номер ітерації. Знайти в C_{t-1} пару кластерів U, V з мінімальною відстанню R_{UV}. Далі поєднуємо їх в один кластер: $W = U\cup V$. $C_t = C_{t-1}\cup\{W\}/\{U, V\}$. Тому для всіх $S\in C_t$. Наступним обчислимо R_{WS} за формулою Ланса-Уільямса: $R_{WS}=\alpha\cup R_{US} + \alpha_V R_{VS} + \beta R_{UV} + \gamma|R_{US}–R_{VS}|$.

Для візуалізації процесу ієрархічної кластеризації побудуємо дендрограмму. Де по вертикалі відкладаємо мінімальну відстань R_t між кластерами, а по горизонталі – об'єкти (хвостосховища). Такий графік дозволяє побачити в якому порядку відбувалося об'єднання даних у групи і наскільки кластери відокремлені один від одного по мінімальній відстані (рис. 2а). Інший вигляд цієї ж інформації наведено на рис. 2б, де показані результати розбиття хвостосховищ в плані.

На рис. 2а можна в обох випадках спостерігати тупикові гілки дендрограми (хвостосховища 43, 42, 40; 34, 35), які закінчуються на рівнях побудови дерева більше або рівних 5 і 7, що є аномаліями класифікації.

Виходячи з рис. 2а ми бачимо, що найефективніше можемо поділити кластери на три групи. На основі оцінки дендрограми та відстані між кластерами на співвідношенні додавання кластера до кластера та граничної відстані на якому відбувається їх об'єднання можна поділити вибірку на три основі кластери. Виходячи з такого ділення можемо сказати, що ці хвостосховища достатньо чітко об'єднуються в один таксон і це проходить до верху крізь всі рівні, крім аномальних значень, які були розглянуті вище.

На рис. 2б представлені хвостосховища, що поділені на три групи та в яких показаний чіткий поділ кольорами. До синього кластеру (на рисунку позначено трикутниками) класифікуємо хвостосховища, що показують найбільшу ємність разом з достатньо низьким рівнем токсичності. Такі хвостосховища мають низький рівень небезпеки у випадку аварії. До зеленого (на рисунку позначено колами) та червоного (на рисунку позначено квадратами) кластерів відносяться група хвостосховищ, що характеризуються ємністю менше середньої та більш високим рівнем токсичності. Такі хвостосховища мають достатньо високу токсичність речовин, і відповідно високий рівень небезпеки.

Наступним методом кластеризації застосуємо метод *k*-середніх. Вхідні дані з табл. 1 за результатами аналізу кореляції факторів були скориговані та спрямовані на проведення кластерного аналізу з метою визначення наявності у структурі даних кластерів, де виконаємо перебір кількості кластерів (Astakhova, 2015).

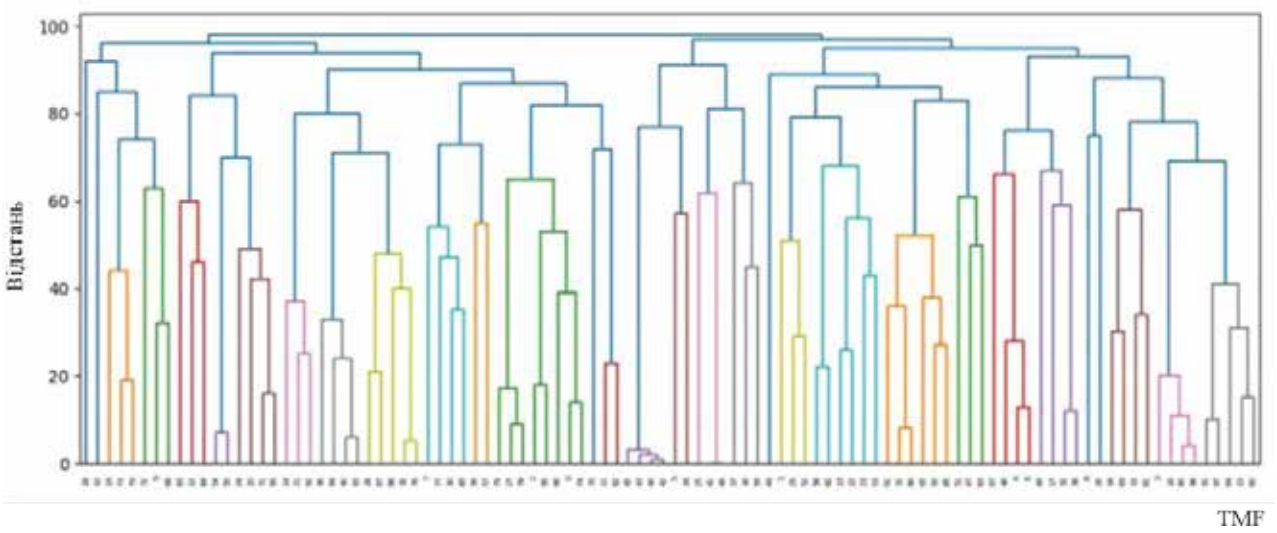

Рис. 2а. Агломеративна ієрархічна кластеризація (дендрограма)

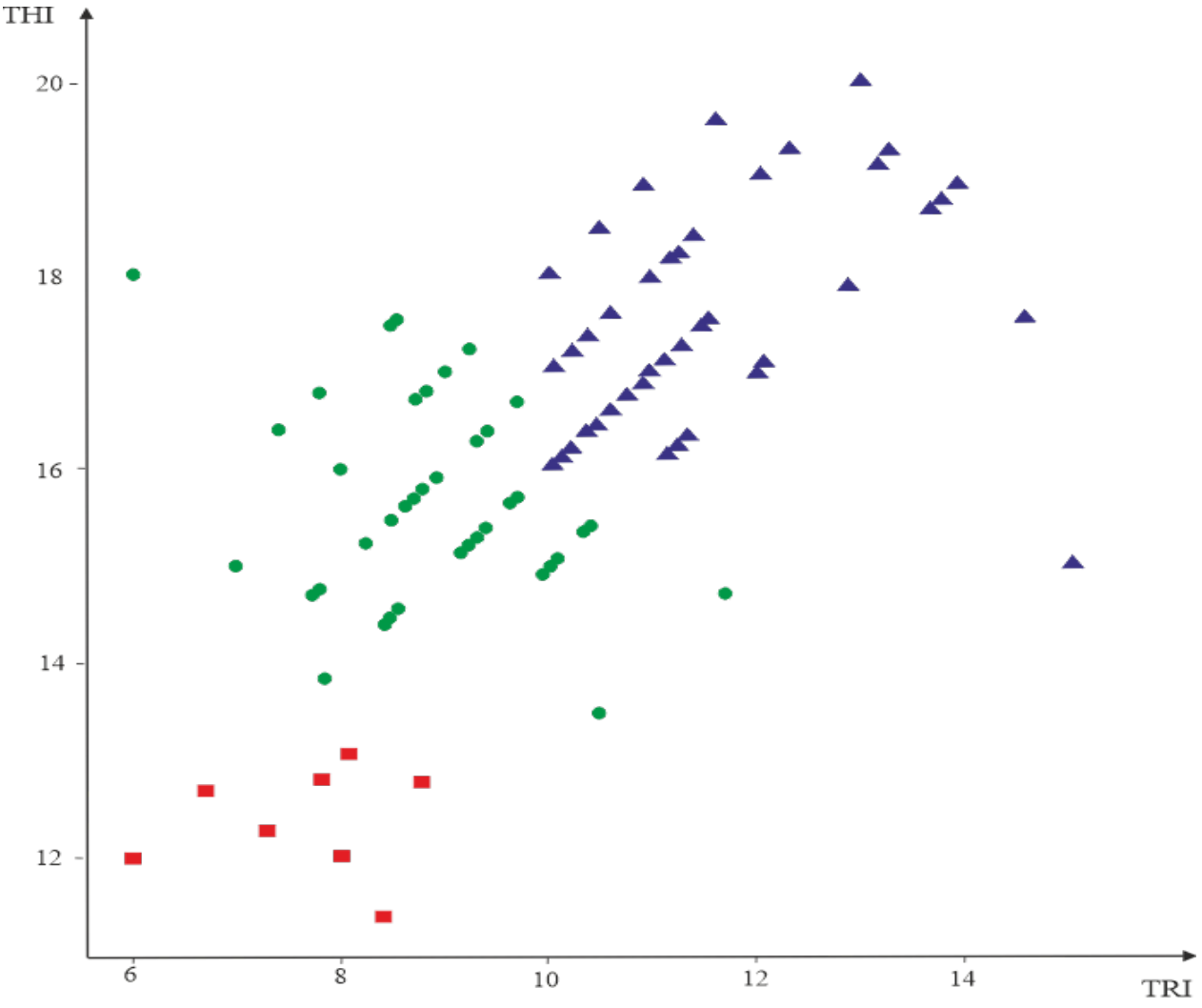

Рис. 2б. Агломеративна ієрархічна кластеризація

Як відомо, метод *k*-середніх полягає у відшуканні такого розбиття набору з *n* точок $\{x_1, x_2,\ldots, x_n\}$ на *k* непустих кластерів, що не перетинаються, а також щоб була мінімальною сума квадратів відстаней від кожної точки до центра відповідного кластера. Класичний метод *k*-середніх шукає найкращу множину кластерів $\{S_1, S_2, \ldots, S_k\}$ шляхом мінімізації цільової функції: $\sum_{h=1}^{k}\sum_{x_i \in S_h} x_i - c_h\ ^2 \rightarrow \min_{S_1, S_2, \ldots, S_k}$, де c_h – центр *h*-го кластера, застосуємо це для пошуку кластерів у хвостосховищах. Такий підхід дозволить графічним способом визначити кількість кластерів, яке буде відповідати точці, де спад своїх значень уповільнюється найбільш сильно. Така точка знаходиться на рівні чотирьох кластерів. Отже, було визначено, що в заданій структурі даних можна виділити чотири кластери, що відрізняються між собою за якимись, поки що не встановленими, факторами.

Отриманий розподіл на кластери говорить про наявність факторів, що впливають на структуру даних, і підтверджують гіпотезу про можливість кластеризації хвостосховищ, проте не розкривають сутність цих факторів. Для визначення факторів, які впливають структуру даних чинників, використаємо метод алгоритму кластеризації Ллойда. Цей метод дозволяє знизити розмірність багатовимірних даних та отримати їх візуальну структуру без істотної втрати якості інформації.

Для покращення інтерпретації структури даних вони були зображені на рис. 3.

Як осі координат використані значення головних компонент (1 і 2 відповідно). У цих координатах були представлені хвостосховища (кожен кластер від 0 до 3-го показаний різним кольором). На рис. 3 спостерігається досить чіткий поділ на чотири кластери – темно-червоні квадрати праворуч (перший за зростанням кластер), чорні зірки вгорі по центру (другий за чисельністю кластер), зелені кола зліва (третій кластер), сині трикутники в нижній частині графіка (найменший за чисельністю кластер). Кількість кластерів, що візуально спостерігається на рис. 3, збігається з кількістю отриманих іншими способами.

Таким чином, за допомогою методу кластерного аналізу вихідний набір даних може бути згрупований на чотири кластери. Візуалізація вихідних даних шляхом основних компонентів підтверджує наявність чотирьох груп точок даних. З урахуванням отриманих основних компонентів було визначено межі технічних показників кластерів.

На дисперсію даних впливають дві основні компоненти, перша – ємність хвостосховищ, та друга – оцінка токсичності. Друга компонента пов'язана з оцінкою ризику впливу на населення та навколишнє середовище. Розглянемо детальніше отримані кластери.

До першого кластера відноситься група хвостосховищ, яку можна охарактеризувати як

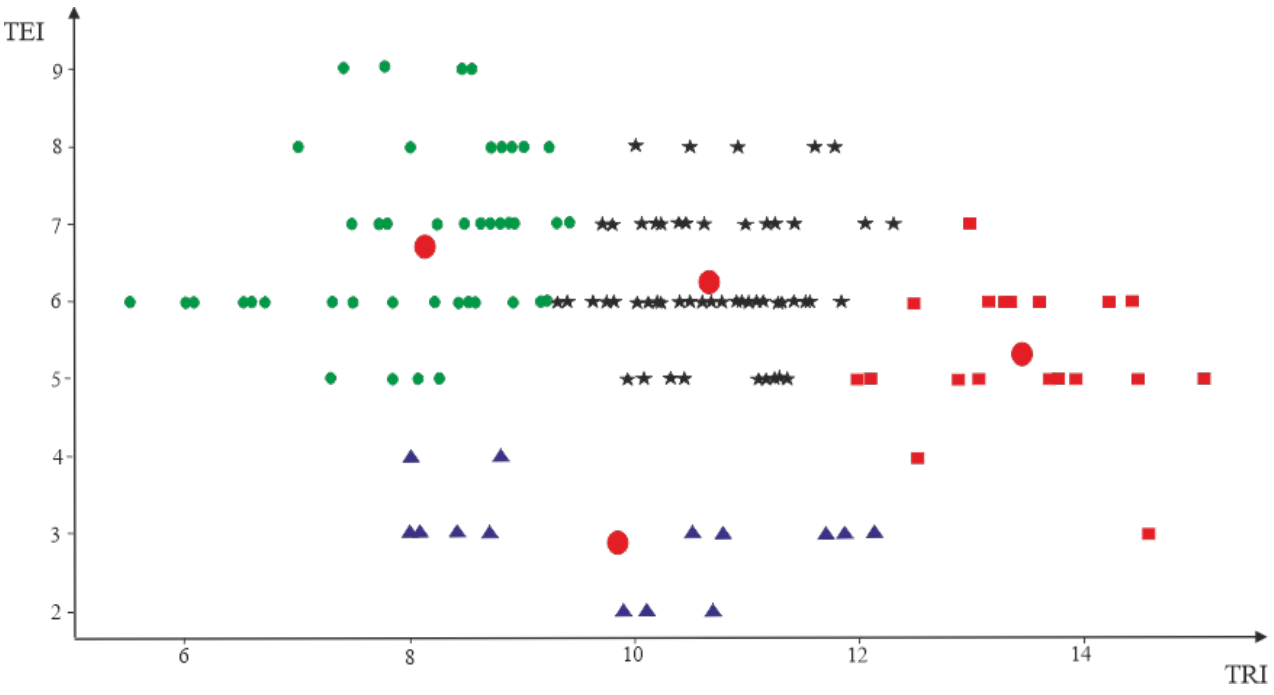

Рис. 3. Візуалізація даних з допомогою методу кластеризації Ллойда

середні за ємністю та низькою токсичністю. Ці хвостосховища можуть складатися з шламу різної крупності із загальним об'ємом близько 1650000 м^3. Такі хвостосховища мають низьку токсичність речовин хвостів, а відповідно і низький рівень небезпеки на навколишнє середовище та на населення. Як правило такі хвостосховища з низьким рівнем аварії з нестійкою дамбою, що відноситься до сприятливих природних умов.

До другого кластеру відносяться група хвостосховищ, що характеризуються найбільшою ємністю разом з низьким рівнем токсичності. Ці хвостосховища можуть складатися з шламу із загальною місткістю більше 10580000 м^3. Такі хвостосховища мають низьку токсичність речовин відходів, а відповідно і низьким рівнем небезпеки для навколишнього середовища і населення. Як правило такі хвостосховища мають низький рівень аварій з нестійкою дамбою, що відноситься до сприятливих природних умов.

До третього кластеру відносяться група хвостосховищ, що характеризуються найменшою ємністю разом та високим рівнем токсичності. Ці хвостосховища можуть складатися з червоного мулу або глини із відносно невеликою загальною місткістю близько 780000 м^3. Такі хвостосховища мають високу токсичність речовин відходів, а відповідно високим рівнем небезпеки на навколишнє середовище та відповідно – на населення з нестійкою дамбою. У випадку аварії на такому хвостосховищі наслідки для навколишнього середовища та населення несприятливі.

До четвертого кластеру відносяться група хвостосховищ, що характеризуються середньою ємністю разом та високим рівнем токсичності. Ці хвостосховища можуть складатися з шламів із загальним об'ємом близько 3880000 м^3. Такі хвостосховища мають високу токсичність речовин хвостів, а відповідно і високий рівень небезпеки на навколишнє середовище та на населення. У випадку аварії на такому хвостосховищі наслідки для навколишнього середовища та населення має неприємні наслідки, але не несприятливі.

Таким чином, встановлені за допомогою методу *k*-середніх індексу потенціалу небезпеки хвостосховищ в Європейському союзі дозволять їх розподілити на групи за їх небезпечністю для навколишнього середовища так наслідками у випадку їх аварії.

Порівнюючи метод агломеративної кластеризації та метод кластеризації Ллойда можна дійти до висновку, що другий метод є більш точний, що дозволяє більш якісно здійснити аналіз хвостосховищ.

Висновки і перспективи подальших досліджень. В статті виконано аналіз небезпечного впливу хвостосховищ утворених від промислового виробництва на навколишнє середовище методами машинного навчання. При різних розмірах хвостосховищ, їх токсичності та технологічних параметрів визначено їх положення в кластерах, що дозволило класифікувати стан рівня їх небезпеки внаслідок аварії. Зроблено порівняльний аналіз підходів методів машинного навчання для визначення потенційної небезпеки хвостосховищ. За допомогою методу DBSCAN було визначено хвостосховища, які за своїми параметрами потрапляють до кореневого, граничного та шумового кластерів. Далі, всі хвостосховища, що потрапили до кореневого та граничного кластерів, було класифіковано наступними методами кластеризації – *k*-середніх та агломеративним методом. Кластеризація методом *k*-середніх будовалася на вибірках даних хвостосховищ шляхом порівняння метрики відстані і було розподілено на чотири кластери. Агломеративний метод дозволив визначити три групи кластерів. За цими результатами виконано порівняння цих методів та було зроблено висновок, що моделювання та пошук центрів кластеризації за допомогою *k*-середніх є більш комплексним рішенням задачі в порівнянні з методом агломеративної кластеризації. Точність використаного підходу кластеризації методом *k*-середніх виявився вищою, ніж у агломеративному методі кластеризації.

Таким чином, вдалося визначити спосіб кластерного аналізу хвостосховищ для при їх небезпеці для навколишнього середовища внаслідок аварії та використати за рахунок машинного навчання кращий спосіб їх класифікації.

ЛІТЕРАТУРА:

1. Nikolaieva I. O., Rudakov D. V. Development of a checklist for improvement of tailings safety. *Naukovyi Visnyk Natsionalnoho Hirnychoho Universytetu*. 2015. 2, 97–103

2. Proceedings of 7th International Conference Innovation Management, Entrepreneurship and Sustainability (IMES), 2019. Pp. 519–533

3. Kravchenko S., Hryshkun Ye. and Vlasenko O., Metody klasyfikatsii mashynnoho navchannia z vykorystanniam biblioteky scikit-learn, URL: http://tech.vernadskyjournals.in.ua/journals/2020/3_2020/part_1/21.pdf

4. Winkelmann-Oei G., Rudakov D., Shmatkov G., Nikolaieva I. A method for evaluation of tailings hazard. New Developments in Mining Engineering: Theoretical and Practical Solutions of Mineral Resources Mining / editors Bondarenko V., Kovalevska I., Pivnyak G. London: Taylor & Francis Group, 2015. pp. 33–38.

5. URL: https://www.sendaiplatform.org

6. Rico, M., Benito, G., Díez-Herrero, A. Floods from tailings dam failures. J. Hazard. Mater. 2008a. 154, 79–87.

7. Rico, M., Benito, G., Salgueiro, A.R., Díez -Herrero, A., Pereira, H.G. Reported tailings dam failures: a review of the European incidents in the worldwide context. J. Hazard. Mater. 2008b.152, 846–852.

8. Coduto D. P. Geotechnical Engineering: Principles and Practices. 1998.

9. Fredlund D. G., Rahardjo H., Fredlund M. D. Unsaturated Soil Mechanics in Engineering Practice. 2012.

10. Wozniak M. Editorial: Applying Machine Learning for Combating Fake News and Internet/Media Content Manipulation. 2021 URL: https://www.researchgate.net/publication/354462936_Editorial_Applying_Machine_Learning_for_Combating_Fake_News_and_InternetMedia_Content_Manipulation.

11. Astakhova N. N., Demidova L. A., Nikulchev E. V. Forecasting Method for Grouped Time Series with the Use of K-Means Algorithm. *Applied Mathematical Sciences*. 2015. Vol. 9. No. 97. P. 4813–4830.

REFERENCES:

1. Nikolaieva, I. O., Rudakov, D. V. (2015). Development of a checklist for improvement of tailings safety. *Naukovyi Visnyk Natsionalnoho Hirnychoho Universytetu.* 2, 97–103

2. Proceedings of 7th International Conference Innovation Management, Entrepreneurship and Sustainability (IMES), 2019. Pp. 519–533

3. Kravchenko, S., Hryshkun, Ye. and Vlasenko, O. Metody klasyfikatsii mashynnoho navchannia z vykorystanniam biblioteky scikit-learn, Retrieved from: http://tech.vernadskyjournals.in.ua/journals/2020/3_2020/part_1/21.pdf

4. Winkelmann-Oei, G., Rudakov, D., Shmatkov, G., Nikolaieva, I. (2015). A method for evaluation of tailings hazard. New Developments in Mining Engineering: Theoretical and Practical Solutions of Mineral Resources Mining / editors Bondarenko V., Kovalevska I., Pivnyak G. London: Taylor & Francis Group, pp. 33–38.

5. Retrieved from: https://www.sendaiplatform.org

6. Rico, M., Benito, G., Díez-Herrero, A., 2008a. Floods from tailings dam failures. J. Hazard. Mater. 154, 79–87.

7. Rico, M., Benito, G., Salgueiro, A.R., Díez -Herrero, A., Pereira, H.G., 2008b. Reported tailings dam failures: a review of the European incidents in the worldwide context. J. Hazard. Mater. 152, 846–852.

8. Coduto, D. P. (1998). Geotechnical Engineering: Principles and Practices.

9. Fredlund, D. G., Rahardjo, H., Fredlund, M. D. (2012). Unsaturated Soil Mechanics in Engineering Practice.

10. Wozniak M. (2021). Editorial: Applying Machine Learning for Combating Fake News and Internet/Media Content Manipulation. Retrieved from: https://www.researchgate.net/publication/354462936_Editorial_Applying_Machine_Learning_for_Combating_Fake_News_and_InternetMedia_Content_Manipulation.

11. Astakhova, N. N., Demidova, L. A., Nikulchev, E. V. (2015). Forecasting Method for Grouped Time Series with the Use of K-Means Algorithm. *Applied Mathematical Sciences*. Vol. 9. No. 97. P. 4813–4830.