

<https://doi.org/10.15407/sofs2025.02.027>  
УДК 001.8:316.77

**М.В. ПОЛЯКОВ**<sup>1</sup>, доктор економічних наук, доцент  
e-mail: [info@noosphere.com](mailto:info@noosphere.com)  
<https://orcid.org/0000-0001-7896-2486>

**І.Г. ХАНІН**<sup>2</sup>, доктор економічних наук, професор  
e-mail: [i.h.khanin@nuwm.edu.ua](mailto:i.h.khanin@nuwm.edu.ua)  
<https://orcid.org/0000-0002-4221-2314>

**Г.Я. ШЕВЧЕНКО**<sup>1</sup>, кандидат технічних наук, керівник відділу  
e-mail: [nikk.gena@gmail.com](mailto:nikk.gena@gmail.com)  
<https://orcid.org/0000-0003-3984-9266>

**В.С. БІЛОЗУБЕНКО**<sup>3</sup>, доктор економічних наук, професор  
e-mail: [bvs910@gmail.com](mailto:bvs910@gmail.com)  
<https://orcid.org/0000-0003-1269-7207>

**О.А. МАРЧЕНКО**<sup>1</sup>, магістр, науковий співробітник  
e-mail: [arche.oleg1@gmail.com](mailto:arche.oleg1@gmail.com)  
<https://orcid.org/0000-0001-7665-7832>

<sup>1</sup> Асоціація Ноосфера

пр. Гагаріна 103-А, Дніпро, 49000, Україна

<sup>2</sup> Національний університет водного господарства  
і природокористування

вул. Соборна 11, Рівне, 33028, Україна

<sup>3</sup> Університет митної справи та фінансів

вул. Володимира Вернадського 2/4, Дніпро, 49000, Україна

---

## DATA MINING ДЛЯ ДОСЛІДНИЦЬКИХ ЦІЛЕЙ

---

*Розглянуто можливість використання методів Data Mining (DM) у науковому дослідженні, спрямованому на пізнання, яке потребує відповіді на питання «чому?», а також зв'язку методів DM із загальними питаннями пізнання. Методологія дослідження знаходиться на стику математики, статистики, інформатики,*

---

Цитування: Поляков В.М., Ханін І.Г., Шевченко Г.Я., Білозубенко В.С., Марченко О.А. *Data Mining* для дослідницьких цілей. *Наука та наукознавство*. 2025. № 2 (128). С. 27—47. <https://doi.org/10.15407/sofs2025.02.027>

© Видавець ВД «Академперіодика» НАН України, 2025. Стаття опублікована на умовах відкритого доступу за ліцензією CC BY-NC-ND license (<https://creativecommons.org/licenses/by-nc-nd/4.0/>)

теорії управління. Поряд із використанням загальнонаукових методів пізнання (узагальнення, систематизація, абстрагування, індукція і дедукція, аналіз і синтез, аналогія, порівняння, формалізація, моделювання, класифікація, категоризація) застосовано спеціальні методи аналізу (логічний, структурний, функціональний), описовий метод дослідження, інтерпретаційну методологію. Уточнено зв'язок методів DM із загальною теорією пізнання. Обґрунтовано, що методи DM — це нова, людино-машинна методологія емпіричного пізнання. Вказано на обмеженість цих методів — результатами їх використання є емпіричні закономірності (ЕМЗ), тобто ймовірні (а не точні) знання, на основі яких формують інструкції або правила ухвалення рішення, але залишаються на першому, найнижчому з можливих, емпіричному рівні пізнання, що є характерним для бізнесу. Але головна мета наукових досліджень — отримання дійсно нових знань, які надалі можуть бути втілені в інновації та забезпечити конкурентну перевагу вищого порядку. ЕМЗ можуть слугувати «заготовками» для формулювання, перевірки та відбору гіпотез з метою подальшого глибшого пізнання предметної області та отримання надійніших знань, що наближають дослідника до розуміння. А якщо з'являється розуміння, це означатиме, що знайдено дещо таке, що лежить в основі отриманого знання і може стати новим методом розв'язання поставлених завдань. Знання — це відомості, а розуміння — здатність робити висновки. На сьогоднішній день це межа застосовності всіх інструментів DM: вони дають знання у вигляді емпіричних і робочих гіпотез, але не дають розуміння, без якого неможливий перехід до наступного рівня пізнання — наукового методу.

**Ключові слова:** Data Mining, емпірична закономірність, навчальна вибірка, ознаковий простір, проектування даних.

**Вступ.** У 2024 р. Нобелівську премію з фізики отримали Джон Гопфільд і Джоффрі Гінтон за розробки в машинному навчанні нейромереж. Дж. Гопфільд створив асоціативну пам'ять, здатну відновлювати зображення в даних. Дж. Гінтон розробив метод для автономної ідентифікації властивостей у даних.

Основною метою будь-яких наукових досліджень є отримання знань і досягнення розуміння досліджуваної області дійсності: «Виявлення закономірностей у потоці даних — основний шлях наукового пізнання навколишньої дійсності. І треба сказати, надзвичайно трудомісткий, навіть якщо заздалегідь передбачити, які саме величини пов'язуються шуканою закономірністю»<sup>1</sup> [1, с. 18—19]. Однак для подальшого викладу важливим видається зауваження, що стосується різниці між законом і закономірністю. Під законом розуміють загальнообов'язкове правило, а під закономірністю — менш суворе правило ніж закон, наприклад зв'язок і взаємозалежність певних явищ об'єктивної дійсності.

---

<sup>1</sup> Переклад із російської зроблено О.А. Марченком.

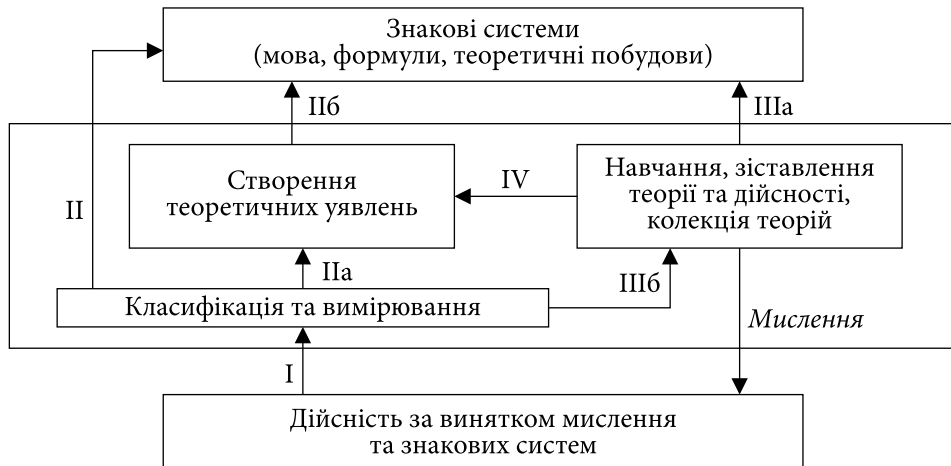
Фактичним результатом будь-якого наукового дослідження є відповіді на запитання «що?», «як?» і «чому». Нині актуальними є дослідження складних об'єктів, часто на стику наук або предметних галузей, які характеризуються багатовимірністю і різнотипністю ознак, що їх описують. Це потребує залучення сучасних методів їх аналізу, здебільшого пов'язаних із т. зв. інтелектуальним аналізом даних або *Data Mining (DM)* у західній термінології. Водночас *DM* активно використовують для потреб бізнесу. Така практика дала змогу зрозуміти важливий факт — необхідно розрізняти т. зв. академічні дослідження, які найчастіше проводять на зафіксованих історичних даних насамперед для покращення архітектури моделей *DM*, і виробничі дослідження, метою яких є не академічний інтерес, а успішність бізнесу як результат використання моделі *DM*. Тобто бізнес насамперед цікавлять відповіді на перші два запитання — «що?» і «як?»

Однак трапляється ще й третій випадок — дослідження, метою яких є не академічний інтерес, як-то поліпшення якості архітектури моделей *DM*, а успішність наукових досліджень як результат використання моделі *DM*. Це випадок безпосереднього використання *DM* у науковому дослідженні, спрямованому на пізнання з отриманням відповіді на запитання «чому?» Саме ця обставина пов'язує *DM* із загальними питаннями пізнання.

**Аналіз останніх досліджень і публікацій.** Універсальна схема пізнання включає три етапи: перший — досвід, спостереження, експеримент, вивчення феномена, тобто нагромадження фактів для подальшого аналізу; другий — узагальнення фактів, виокремлення в них істотного, формування на їхній основі гіпотез і висновків, тобто деяка абстракція від першого; третій — перевірка на практиці отриманих раніше гіпотез або висновків. Досить чітко і стисло її сформулював академік Б. Гнеденко [2]: перший етап — спостереження або живе споглядання, другий — перехід до абстракції, третій — перевірка абстракції на практиці.

Потрібно також нагадати, що на найбільш ранніх етапах досліджень, пов'язаних із *DM* (у той період побутувала назва «розпізнавання образів» (РО)), виникали питання про роль РО в пізнанні. Зокрема, Л. Малиновським [3] запропоновано схему пізнання, побудовану на принципах розпізнавання та класифікації образів (рис. 1).

*DM* (або РО) від початку активно застосовували з науковою метою для розв'язання прикладних задач. Достатньо згадати [4] — одну з перших монографій з автоматизації оброблення наукових досліджень у хімії, а також [5—6], де наведено алгоритм розв'язання задачі автоматизації досліджень у геології. Потім на порядку денному постали завдання бізнесу, де *DM* успішно використовували для розв'язання численних завдань маркетингу, комерції, фінансів, виробництва на основі багатови-



**Рис. 1.** Співвідношення між мисленням, дійсністю та знаковими системами  
Джерело: [3].

мірних і різнотипних даних, зокрема т. зв. великих даних [7—9]. Останнім часом питання застосування *DM* у наукових дослідженнях знову набуло актуальності.

У [10] доведено, що десятиліття інвестицій у поєднанні із зусиллями дослідників, технологів, бібліотекарів, архівістів призвели до створення цифрової інфраструктури в гуманітарних науках, здатної підтримувати робочі процеси; показано ключові досягнення в галузях, як-то епіграфіка та палеографія, а також у пов'язаних із ними історичних дослідженнях стародавньої та середньовічної епох. Однак автор [10] вказує на нерівномірний розподіл можливостей інфраструктури, на необхідність покращувати її зручність для користувачів.

Огляд [11] показав, що існують нові можливості для соціологічних досліджень, які в певному сенсі є побічними продуктами інформатики. В роботі представлено методи *DM*, які можуть бути застосовані до конкретних соціологічних проблем поза межами бізнес-додатків, зокрема таких, які ще не досліджено соціологами, і показано нові погляди на класичні соціологічні питання.

У [12] розкрито вплив цифрової революції на соціальні науки, зокрема на емпіричну соціологію, з епістемологічної, інформаційної та аналітичної точок зору. Відповідно до тематичної спрямованості дослідницької теми, вона зосереджена на великих даних і машинному навчанні (*DM*), які є двома основними елементами нової міждисциплінарної галузі комп'ютерних соціальних наук. Залежно від тематичної спрямованості дослідження потребує використання великих даних і машинного навчання (*DM*), які є двома основними елементами новітньої міждисциплі-

лінарної галузі комп'ютерних соціальних наук. Автори дають поради для покращення інституційної системи. Одним із пріоритетів є необхідність реформування освіти в галузі соціальних наук і створення великої централізованої інфраструктури даних.

У [13] запропоновано використання *DM* для аналізу даних моніторингу і продемонстровано застосування *DM* для оброблення даних про обмотки вимикачів у енергетиці. На основі отриманих результатів оцінено стан цих компонентів обладнання.

Приклад застосування *DM* у сфері медицини, а саме у кардіології, наведено у [14]. Показано, що діагностика має фундаментальне значення для кардіологічної допомоги. Обчислювальні алгоритми дають змогу краще аналізувати наявні медичні дані великої кількості пацієнтів і вдосконалити механізм діагностики. Проаналізовано дані пацієнтів кардіологічного інституту в Мексиці із застосуванням кластеризації. Хоча робота в цьому напрямі перебуває на ранній стадії, часткові результати дають впевненість у тому, що група дослідників зможе скористатися перевагами обчислювальних методів для вдосконалення кардіологічної діагностики. Про використання інтелектуального аналізу даних та його значення в медичних дослідженнях також йдеться у [15].

Звертаючи увагу на освітню діяльність, можна дійти висновку, що сьогодні університети видають не лише випускників, а й величезні обсяги даних. Побудова інформаційної системи, яка може навчатися на основі даних, є складним завданням, але його успішно вирішують за допомогою різних підходів *DM*, як-то кластеризація, класифікація, алгоритми прогнозування тощо. Однак застосування цих алгоритмів до освітніх даних є досить обмеженим [16].

Яскравий приклад застосування *DM* у сільськогосподарській сфері наведено у [17]. Автори здійснили огляд досліджень, присвячених застосуванню методів *DM*, як-то алгоритми *ID3*, *k*-середні, *k* найближчих сусідів, нейронні мережі. Підкреслено, що такий підхід у сільському господарстві є відносно новим для прогнозування та управління у сільськогосподарському секторі.

Однак, незважаючи на успішність досліджень із використанням інструментів *DM*, актуальним залишається питання: що ми пізнаємо завдяки *DM*? Мається на увазі не конкретний результат, отриманий емпіричним шляхом, а результат із погляду теорії пізнання, оскільки без чіткої відповіді на це запитання подальше використання *DM* може призвести до хибного розуміння мети дослідження і пов'язаних із цим помилок як у прогнозах, так і в ухваленні рішень.

**Методологія дослідження** знаходиться на стику математики, статистики, інформатики, теорії управління. Поряд із використанням загальнонаукових методів пізнання (узагальнення, систематизація, абстрагу-

вання, індукція і дедукція, аналіз і синтез, аналогія, порівняння, формалізація, моделювання, класифікація, категоризація) застосовано спеціальні методи аналізу (логічний, структурний, функціональний), описовий метод дослідження, інтерпретаційну методологію.

**Наукова новизна** дослідження:

- Доведено, що методи *DM* являють собою сучасну людино-машинну методологію емпіричного пізнання. Встановлено, що ці методи мають обмеження, пов'язані з їхньою здатністю виявляти емпіричні закономірності (ЕМЗ), що виражають імовірнісні знання. Ці закономірності щонайкраще стають основою для розроблення інструкцій, методик або правил ухвалення рішень. Однак вони залишаються на початковому, емпіричному рівні пізнання, що є характерним якраз для бізнес-додатків.

- Показано, що ЕМЗ можна використовувати як відправну точку для формулювання, перевірки і відбору гіпотез, що відкриває можливість глибшого дослідження предметної області та отримання надійних знань. Такий процес сприяє переходу від простого накопичення знань до їх осмислення, що наближає дослідника до розуміння сутності досліджуваного явища.

- Підкреслено, що розуміння є ключовим етапом, який дає змогу виявити фундаментальні принципи, покладені в основу отриманих знань. Можливо, саме на цьому етапі можуть бути розроблені нові методи розв'язання поставлених задач.

- Обґрунтовано, що знання і розуміння — це різні категорії: знання є сукупністю відомостей, а розуміння передбачає здатність робити висновки. Методи *DM* на сучасному етапі обмежуються генерацією знань у формі емпіричних і робочих гіпотез, але не дають змоги досягти рівня наукового методу без розвитку розуміння. Отже, виявлено межу застосовності сучасних інструментів *DM*.

**Мета статті** — розкрити потенційну корисність інструментів *Data Mining* для дослідника та характеристики отриманих за їхньою допомогою результатів із позицій теорії пізнання.

Відповіді на ці питання є важливими не лише з теоретико-методологічного, а й з практичного погляду, оскільки дають змогу чітко окреслити обмеження, притаманні всім інструментам *DM*.

**Результати дослідження.** Застосування інструментів *DM* починається тільки тоді, коли є вже підготовлені дані у вигляді вибірок, де об'єкти представлені наборами багатовимірних даних (ознаковий опис). Такі вибірки мають різні назви: «таблиця “об'єкт-властивість”», «навчальна вибірка», «таблиця експериментальних даних». Надалі вважатимемо їх еквівалентними. Важливо зазначити, що питання вибору ознакового опису (ознакового простору) і попереднього опрацювання даних перебувають поза компетенцією *DM*, хоча вони присутні у розв'язанні будь-

якої задачі *DM*: вважають, що такі питання належать до компетенції фахівця конкретної предметної області, з яким взаємодіє аналітик *DM*.

У дослідників, які бажають застосувати методи *DM*, іноді виникає хибне відчуття, що ці методи є універсальними і тому придатні для будь-яких потреб.

Роботи з питань *DM*, зокрема включені до огляду літератури, традиційно містять опис побудови тієї чи іншої моделі та її випробування в академічному або виробничому середовищі, але не торкаються іншого класу проблем, що виникають під час спроб використовувати методи *DM* для потреб наукових досліджень. Але в роботах, пов'язаних із практичним використанням методів *DM*, потрібен інший погляд на розв'язання дослідницьких завдань, на труднощі, що трапляються, і на весь спектр практичних завдань на цьому шляху. Особливо це стосується завдань, пов'язаних з осмисленням самої задачі з найширшого погляду: вибору ознакового простору та відповідних метрик, формування навчальної вибірки, вибору методу *DM* і визначення зв'язку цього методу з теорією пізнання як засадничою методологією всього процесу розв'язання задачі за допомогою *DM*, встановленню місця *DM* у цій теорії, зв'язку між *DM* і напрямом дослідження, який є основою для розв'язання задачі за допомогою *DM*. Це надзвичайно важливо, оскільки точне формулювання цих завдань не лише прокладає «дорожню» карту використання методів *DM* у наукових дослідженнях, а й визначає можливості та обмеження методів *DM*, їхнє місце в загальній культурі науково-дослідної діяльності.

Методи *DM* відрізняються від інструментів статистичного опрацювання даних тим, що замість перевірки заздалегідь передбачуваних користувачами залежностей вони на підставі наявних даних здатні знаходити такі залежності, а, точніше, закономірності, на основі яких дослідник зможе самостійно будувати науково обґрунтовані гіпотези про їхній характер. Найбільш використовуваними є такі методи *DM*:

- кластеризація — групування об'єктів з виявленням внутрішньогрупової подібності та міжгрупової відмінності;
- класифікація на основі дерева рішень — побудова причинно-наслідкової ієрархії умов, що призводить до певних рішень;
- пошук асоціацій — пошук стійких комбінацій елементів у подіях або об'єктах;
- нейронні мережі — використовуються для розпізнавання зорових образів, побудови регресії, класифікації.

Перевагами методів *DM* є особливості, що враховують характер сучасних підприємств:

- виявлення раніше невідомих, нетривіальних і практично корисних знань у даних;



**Рис. 2.** Універсальна схема пізнання з використанням інструментів *DM*  
Джерело: розроблено авторами.

- розв’язання різних типів задач, як-то прогнозування, визначення ймовірності ризиків, пошук послідовностей і групування;
- візуалізація результатів аналізу даних, що дає змогу використовувати його особам без спеціальної математичної підготовки.

Основне завдання при застосуванні інструментів *DM* — виявлення причинно-наслідкових зв’язків для зібраних даних із метою отримання нового знання.

У загальному випадку ми стикаємося з тим, що для розв’язання поставленої задачі засобами *DM* потрібно проводити і збирання, і аналіз даних, і, скоріш за все, програмну побудову моделей (наприклад, за допомогою мови програмування *Python* або обираючи готові рішення).

Фактично отриманню нового знання передують автоматичне знаходження т. зв. емпіричних закономірностей (ЕМЗ), які, згідно з [18], є джерелами для формулювання гіпотез, що є найважливішою складовою наукового пізнання, формою розвитку природознавства [19].

Описаній вище універсальній схемі пізнання, в якій передбачено використання інструментів *DM*, для більшої наочності можна надати графічного вигляду (рис. 2).

Схема на рис. 2 відображає загальний підхід до пізнання, але її необхідно уточнити відповідно до мети нашого дослідження — визначити місце інструментів *DM* і характер їх використання. Бажано також вказати можливості та обмеження, властиві всім методам *DM* у процесі пізнання.

Нижче розкрито загальний перебіг робіт із розв'язання задач аналізу даних за допомогою методів *DM*, ув'язки методів *DM* із загальною моделлю пізнання.

Першим кроком є визначення цільових, або результативних, змінних і чинників (параметрів, ознак, атрибутів), які впливають на них. У нашому випадку цільовою є змінна, що описує результат (мету) процесу. Вона може набувати таких значень: 0 — немає дефектів; 1 — є дефект типу 1; 2 — є дефект типу 2 тощо.

Наступний етап пов'язаний із визначенням набору необхідних ознак або параметрів, т. зв. ознакового простору, і збиранням даних. Під даними можна вважати все, з чим взаємодіє людина.

Завдання визначення набору ознак або параметрів із максимальним впливом на цільову змінну є дуже важливим, оскільки не існує жодних правил їх визначення, і для того, щоб обрати ознаки з найщільнішими зв'язками з цільовою змінною, потрібно долучити науковців, інженерів та інших спеціалістів. На початковому етапі розв'язання цієї задачі рекомендується записати якомога більше ознак, що характеризують досліджувані об'єкти і можуть вплинути на результативну змінну. Потрібно також розуміти, що ключовим моментом є взаємодія деяких груп ознак із цільовою змінною та дотримання певних вимог до таких ознак.

Крім того, потрібно звернути увагу на ту обставину, що добір ознак (проектування даних) фактично пов'язаний з першою частиною загальної моделі пізнання: від живого споглядання до абстракції і далі до практики [2]. Звертаємо увагу, що використання методів *DM* якнайточніше вписується в цю модель. Відомий учений у галузі *DM* А.Д. Закревський [1] вважав, що практично будь-яке завдання можна описати за допомогою ознак. Маючи дані про ознаки за великий період, можна за допомогою методів *DM* передбачати тенденції та будувати прогнози. «Живе» споглядання відповідає вибору ознакового простору, його оцінюванню на адекватність і збиранню необхідного експериментального або експертного матеріалу, придатного до подальшого опрацювання. У наведеному вище твердженні можна побачити перетин загальної моделі пізнання, першої її частини з початковим етапом проектування даних, тобто вибором ознакового простору для будь-якого методу *DM*.

У теорії пізнання цей етап докладно не розкрито, хоча й наголошено на його значній ролі та на недостатності досліджень у цьому напрямі. Спроби заповнити цю прогалину зроблено в [21—23], а також у [24], де запропоновано процедуру для початкового опису ландшафту всіх чин-

ників, що стосуються завдання. У термінах *DM* такий підхід означає вибір ознакового простору.

Запропонована процедура (Процедура 1) переводить перший крок загальної схеми пізнання в конструктивну схему, що дає змогу отримати інформаційну картину завдання:

- проведення «споглядального» аналізу задачі й тих чинників, ознак або понять, що в ній присутні;
- опис ландшафту завдання виходячи з аналізу завдання і цілей дослідження, тобто опис усіх «навколишніх» чинників і ознак, які можуть вплинути на завдання;
- вибір у цьому ландшафті детермінант *Ді*, тобто чинників, які, на думку дослідника, мають вирішальний вплив, їх обґрунтування;
- запис відібраних детермінант *Ді* у формальному вигляді  $\langle D_1, D_2, \dots, D_3 \rangle$ , який фактично є описом ознакового простору.

Ця процедура є інструментом дослідження, що дає змогу зацікавленим сторонам отримати інформаційну картину завдання з метою побудови моделі, яка нас цікавить.

Ми хочемо наголосити, з одного боку, на безсумнівній важливості цього етапу, з іншого — на недооцінці глибокого зв'язку вибору ознакового простору з теорією пізнання. *DM* загалом можна вважати емпіричним, або навіть індуктивним, методом пізнання, що одразу вказує на цілком певні обмеження подібних методів. Детальніше про це висвітлено в [19].

Інших формальних методів, що докладно розкривають сутність першого кроку загальної моделі пізнання, зокрема в застосуванні до завдань *DM*, ми не знайшли.

*Обмеження DM під час проектування даних.* *DM* не є всемогутньою технологією, це інструмент, що має притаманні йому обмеження, особливості та сферу застосування, незнання яких знижує його ефективність. *DM* має чотири загальні обмеження:

- пріоритетними є ознаки, в т. ч. цільові, які мінімізують витрати ресурсів; можна також включати до розгляду непрямі ознаки, які легко піддаються визначенню і формалізації;
- ознаки мають бути чітко визначені, з описом методики їх вимірювання або оцінювання.
- для аналізу необхідна достатня кількість ознак, щоб виявити ефективні комбінації, здатні приводити до бажаного результату, що є ключовим завданням методів *DM*;
- у деяких випадках краще використовувати дискретні ознаки замість безперервних; для цього важливо коректно перетворити дані, наприклад, вводячи якісні категорії «високе значення», «середнє значення» і «низьке значення» замість точних числових величин.

Важливо ретельно проводити збирання даних, оскільки інакше можна замість інформації отримати «сміття», і жодними методами *DM* виявити ЕМЗ буде неможливо. Крім того, можна спробувати збагатити дані, тобто включити до аналізу додаткові дані виходячи зі специфіки розв'язуваної задачі, наприклад, використати відкриті дані, як-то результати наукових досліджень.

Звертаємо увагу на те, що «малозначущі» на перший погляд ознаки, або, найімовірніше, їхні поєднання, можуть виявитися визначальними. *DM* якраз і є засобом відкриття несподіваних і невідомих закономірностей у досліджуваних даних. Крім того, якість навчальної вибірки (НВ) безпосередньо пов'язана з вибором системи ознак (атрибутів) для опису об'єктів. Загальновідомо також, що питання про вибір ознак (атрибутів) під час побудови розв'язувального або класифікаційного правила є одним із найважливіших. Складність цієї проблеми полягає в тому, що не існує формальних правил, які дають змогу заздалегідь вказати такий набір ознак (атрибутів), за допомогою якого можна проводити класифікацію із заданою точністю. Ось чому цей етап називається «проектування даних», і він є зоною відповідальності постановника завдання — керівника проекту.

*Збирання та оцінювання даних.* Після формулювання завдання, «живого» споглядання і добору ознак починається наступний етап — збирання даних. Це теж дуже важливий крок, адже від якості зібраного матеріалу значною мірою залежатиме успіх розв'язання всієї задачі *DM*. Що ми вимірюємо, як вимірюємо, є необхідні інструменти для такого вимірювання чи їх потрібно створювати — подібні питання складають сутність цього етапу, який потребує значних витрат часу, фінансових і людських ресурсів. Але отримання НВ ще не гарантує успішності подальших кроків. Необхідна ретельна перевірка, «вичитування» НВ з метою усунення всіляких помилок і недоречностей. Це дуже відповідальна і трудомістка робота, якої не уникнути. До того ж, найчастіше її виконують самі розробники, оскільки крім них ніхто не може оцінити придатність отриманого матеріалу до подальшого опрацювання. Більшу частину роботи інженера з аналітики даних на будь-якому підприємстві складає вивчення і сортування «сирих» даних для подальшого аналізу.

На жаль, цей етап часто або неправильно оцінюють, або не звертають на нього достатньої уваги. На додаток до цього етапу можна запропонувати кількісний метод оцінювання якості сформованої та перевіреної НВ. Такий метод запропоновано і детально розглянуто нижче.

Нині практично єдиним способом оцінити якість НВ у загальному випадку є результат, отриманий на основі точності знайденого під час навчання розв'язувального правила (РП). Однак отримання такої оцінки потребує проведення процедури навчання, а отже значних витрат часу.

Водночас, особливо в процесі виконання наукових досліджень, бажано знати якість НВ ще на етапі її створення без проведення навчання й отримання РП. Бажано на основі отриманої оцінки дати інтерпретацію НВ, зокрема вказати, які граничні можливості щодо розрізнення об'єктів має НВ, а також провести аналіз НВ, вказати інформативні ознаки та видати конструктивні рекомендації щодо поліпшення основної якості НВ — розрізняльної здатності. Тобто це найважливіший етап підготовки даних, який також практично не піддається формалізації і тому подвійно важливий. У [25] зроблено спробу оцінити якість НВ для побудови прогнозованої нейронної мережі. Однак висунуті для такого оцінювання вимоги до НВ знижують ефект від її застосування і не дають змоги оцінити якість НВ для більш загальних випадків, які значно частіше трапляються на практиці.

Ми пропонуємо заповнити цю прогалину і підійти до вибору ознакового простору, тобто детермінант ландшафту, більш усвідомлено і з формалізованих позицій, в основу яких покладено запропоновану вище Процедура 1 і алгоритмічну процедуру обчислення критерію, за допомогою якого оцінюють якість НВ. Зауважимо, що цей критерій обчислюють безпосередньо за даними НВ і він характеризує розрізняльну здатність НВ.

Розрахунок критерію ґрунтується на застосуванні формули для обчислення важливості (інформативності) довільної групи дискретних ознак [26], яка має вигляд:

$$V = (x_{i1}, \dots, x_{ij}) = \frac{1}{k} \sum_{\Delta \in \Gamma} \max_Y \left( \frac{m_{\Delta Y}}{m_Y} \right),$$

де  $k$  — число класів;  $m_Y$  — число об'єктів, що належать класу  $Y$ ;  $\Delta t_{i1}, t_{i2}, \dots, t_{ij}$  ( $0 \leq t_{ij} \leq k_{ij} - 1$ ),  $j = 1, \gamma$  — довільний набір значень ознак (атрибутів); ( $1 \leq \gamma \leq n$ ) — кількість вибірок з  $m$ -го класу, для яких виконується співвідношення  $x_{ij} = t_{ij}$  ( $j = 1, \dots, \gamma$ );  $t_{ij}$  — значення ознак (атрибутів)  $x_{ij}$  у наборі  $\Delta$ ;  $\Gamma$  — множина всіх наборів значень ознак (атрибутів)  $x_{i1}, \dots, x_{ij}$ .

Можна показати, що  $1/k \leq V(x_{i1}, \dots, x_{ij}) \leq 1$ . Граничне значення, що дорівнює 1, має місце за повної помітності класів, що доведено в [26]. Тобто формула за умови  $\gamma = n$  може слугувати оцінкою якості НВ, що складається з дискретних наборів.

Якщо  $V(x_{i1}, \dots, x_{ij}) = 1$ , тоді має місце повна помітність класів, принаймні за використання всього набору ознак. Такий підхід до оцінювання якості НВ дає змогу реалізувати зазначену процедуру, фактично оптимізувати початкові етапи підготовки даних під час створення вибірки, свідомо підійти до її створення і ще на перших етапах обрати оптимальний набір ознак, що може значно скоротити часові й інші витрати під час дослідження предметної області за допомогою методів ДМ або великих даних.

Аналогічна ситуація складається і для таблиць «Об'єкт — властивість» (ТОВ), коли на етапі початкового дослідження ми лише збираємо дані для подальшої їх кластеризації / структуризації без верифікації. У цьому разі роль оцінки якості відіграє візуалізація ТОВ, її візуальне дво- або тримірне представлення. Величезну роль візуалізації в аналізі даних розкрито у [27—30].

Дійсно, одним з основних етапів будь-якого дослідження є структуризація (кластеризація) отриманих даних. Візуальне представлення такої структуризації відіграє важливу, а іноді й головну роль, особливо коли потрібно аналізувати багатовимірні дані, які становлять певну складність для багатьох дослідників-предметників, мало знайомих з основами науки про дані (*Data Science*). Це пов'язано з чуттєвим походженням усього нашого знання, і візуалізація фактично виконує функцію засобу зв'язку між об'єктом дослідження і дослідником, даючи змогу побачити експериментальні дані загалом і наочно. Тоді дослідник сприймає найбільшу кількість інформації, зокрема багатовимірні дані, що характеризують знамениту задачу про іриси Фішера, представлену ним ще у [31], яка слугує своєрідним оселком для тестування багатьох алгоритмів *DM*. У цій задачі присутні три класи і чотири ознаки. Візуалізація дає змогу побачити наочно взаємне розташування об'єктів різних класів, поданих різними кольорами, і дати попередню оцінку зібраним даним із погляду роздільності кластерів.

Усі завдання, згадані в цьому підрозділі, належать до компетенції фахівця зі збирання даних.

*Початкова аналітика.* Після вибору ознак, конструювання ознакового простору і формування вибірки бажано видобути корисні дані з представленої вибірки і осмислити їх. Таку роботу можна почати з опису даних, визначення трендів і взаємозв'язків, перевірки помилок. Отриману картину необхідно надати зацікавленим особам і обговорити з ними всі її особливості. Рекомендується почати з простої візуалізації даних, зокрема багатовимірної. Це завдання аналітика.

*Вибір методів DM.* Далі потрібно зупинитися на виборі методів *DM* і вирішити, що робити: використати вже готові алгоритми і написати для них програму з урахуванням особливостей створеної НВ або розробити свій власний алгоритм, тобто знайти новий спосіб розв'язання поставленої задачі. У першому випадку робота займе кілька місяців, у другому — набагато більше часу. Але в разі успіху результат може бути значним.

Ми вважаємо, що методи *DM* потрібно обирати з погляду не тільки завдання класифікації, а й здатності до інтерпретації та пояснення отриманих результатів. На жаль, широко використовувані зараз нейронні мережі такої здатності не мають, що різко знижує їхню корисність для наукових досліджень. На це хотілося б звернути особливу увагу, оскільки

ки нерозуміння цієї обставини надалі може призвести до дискредитації методів *DM*, адже після певних успіхів на шляху розпізнавання і класифікації, мабуть, виникнуть питання, цілком природні для будь-якої розумової діяльності: що робити далі, як знайти причинно-наслідковий зв'язок, як інтерпретувати результати. Тому в статті обговорено не тільки і не стільки сьогодишню ситуацію в цій галузі знань, а й зроблено спробу вказати на важливі питання, які можуть з'явитися надалі. Нам потрібне не тільки знання, а й розуміння, і лише тоді можна буде говорити про можливе управління процесами і новий рівень пізнання.

Але потрібно пам'ятати, що *DM* має кілька недоліків:

- передбачає значні витрати часу і праці, оскільки видобування даних — це тривалий і складний процес, який потребує продуктивних і кваліфікованих фахівців;
- може видавати помилкові або нерелевантні закономірності, якщо дані не підготовлено в належний спосіб або якщо використано неадекватні алгоритми або метрики;
- можливість спотворення даних унаслідок їх усічення;
- неможливість виявлення даних, що є похідними від оброблених даних.

Вибір оптимального методу *DM* для виконання завдання залежить від кількох чинників:

- тип і обсяг наявних даних: деякі методи краще працюють із числовими даними, інші — з категоріальними або текстовими; деякі методи потребують великої кількості даних для навчання, інші — меншої;
- мета і складність завдання: деякі методи підходять для простих завдань класифікації або регресії, інші — для складніших завдань кластеризації, аналізу послідовностей або асоціативних правил; деякі методи дають змогу інтерпретувати результати і зрозуміти причинно-наслідкові зв'язки, інші — не дають;
- вимоги до швидкості та точності моделі: деякі методи є більш швидкими і простими для навчання і застосування, інші — більш повільними і складними; деякі методи дають точніші та надійніші прогнози, інші — менш точні та надійні.

За вибір методу *DM* відповідає інженер з аналітики даних — це його завдання і зона відповідальності.

У багатьох випадках розв'язання конкретних практичних завдань із точки зору пізнання обмежується рівнем гіпотези або навіть емпіричної закономірності (або попередньої гіпотези), на основі якої надалі щонайкраще формулюють інструкції або правила прийняття рішення і залишаються на першому, найнижчому з можливих, емпіричному рівні пізнання. Це характерно для завдань бізнесу, оскільки в короткостроковій перспективі бізнес як сферу практичної діяльності це влаштовує,

але в довгостроковій перспективі втрачається головне — знаходження дійсно нових глибоких знань, які можуть бути втілені в інновації, що забезпечить конкурентну перевагу вищого порядку. Для наукових досліджень із залученням *DM* саме останнє є основним.

**Комунікації із замовником.** Необхідно вказати на постійний зв'язок із замовником, тобто з тими особами або фахівцями, в чий інтерес розв'язують завдання. Цей зв'язок має тривати протягом усіх етапів розв'язання задачі, оскільки лише в такий спосіб можна отримати якісне рішення. У разі успішної взаємодії рішення дійсно виходить вдалим, а інакше всі витрати виявляються марними. Можна додати, що на етапі взаємодії, окрім суто професійних, виникають проблеми узгодженої думки, врахування особистісних характеристик і багато іншого з того, що ми називаємо арсеналом гуманітарних наук. Іноді такий підхід є вирішальним, принаймні його в жодному разі не можна скидати з рахунків, хоча він і не належить до суто професійних обов'язків фахівця з *DM*.

**Висновки і перспективи подальших досліджень.** На підставі результатів аналізу можна стверджувати, що методи *DM* здатні забезпечити тільки рівень емпіричного пізнання в досліджуваній предметній області, тобто рівень методик та інструкцій. Тому зрозуміло, чому немає «проривних» відкриттів та інновацій, зроблених за допомогою *DM*: їх можна отримати тільки в конкретній предметній області за умови тісної співпраці та взаємодії, повноцінної наукової комунікації, включно з обговоренням гіпотез, ідей та інших обов'язкових елементів творчої діяльності з представниками цієї ж предметної області, що є найбільшою перешкодою для такого роду досягнень.

Звідси слідує низка висновків:

- Методи *DM*, як і великих даних, — це нова людино-машинна методологія емпіричного пізнання.
- У цих методів є своя межа у вигляді емпіричних закономірностей.
- Емпіричні закономірності можуть слугувати «заготовками» для формулювання, перевірки і добору гіпотез з метою подальшого глибшого пізнання предметної області.
- *DM* притаманне значне протиріччя в обробленні даних: будь-яке завдання, вирішуване за допомогою цієї технології, залишається зовнішнім щодо доступних засобів і самого досліджуваного об'єкта (ситуації).
- Моделі *DM* здебільшого мають яскраво виражений класифікаційний характер. Класифікація — це початковий етап будь-якого пізнання, включно з генетичним рівнем, що робить ці моделі затребуваними, хоча вони не мають концептуальної або смислової наповненості.
- *DM* працює із зовнішніми проявами досліджуваного об'єкта, не зачіпаючи його сутності, що позбавляє цей метод широти і системності.

- Немає чіткого регламенту застосування методів *DM*, рішення про які приймає фахівець. Натомість визнається, що фахівці не можуть адекватно орієнтуватись у великій кількості методів і в специфіці різних предметних галузей.

- Використання інструментів *DM* вимагає тісної співпраці з фахівцями в конкретній предметній галузі, що викликає низку питань, пов'язаних з ініціюванням такого співробітництва, готовністю фахівців у предметній галузі, постановкою задачі у відповідному розрізі, створенням тимчасового колективу для розв'язання задачі.

- «Відхід» фахівців із *DM* і великих даних у сферу створення стандартизованого програмного забезпечення не вирішує завдання поглиблення пізнання; межею залишається емпіричне пізнання — отримання емпіричної гіпотези, тобто фактично попередньої гіпотези для конкретної предметної області. У цьому разі вирішення конкретного завдання з поглиблення пізнання, уточнення такої гіпотези повністю перекладається на фахівців у предметній галузі. Крім того, у них виникають проблеми з опануванням відповідного програмного забезпечення, що заважає їм зрозуміти, чи придатні обрані програми для розв'язання практичних задач у їхній галузі. Водночас залишається «за бортом» низка важливих для розв'язання задачі питань: вибір і побудова ознакового опису, розуміння предметної галузі, верифікація даних (якість, корисність), вибір відповідного алгоритму оброблення, оцінювання надійності методу, інтерпретація результатів, перехід від даних до рішень і дій. Найчастіше такі питання вирішують інтуїтивно, на основі досвіду або аналогії, методом «проб і помилок».

Повноцінна взаємодія фахівців у предметних галузях та інженера з аналітики даних є значно більш трудомісткою з погляду організаційних і комунікаційних витрат, але, на наш погляд, вона здатна забезпечити глибокі прориви в предметній галузі.

Тож основний висновок такий: для успішного використання *DM* і отримання гранично можливих результатів у пізнанні з їх допомогою, а саме гіпотез, головна увага має бути зосереджена на попередній аналітиці, формулюванні задач і виборі ознак для опису об'єктів. Це стає аксіомою використання *DM* у наукових дослідженнях. Задачу потрібно вивчати детально, обґрунтовано обирати ознаки, складаючи ознаковий опис об'єктів, збирати та верифікувати дані, і тільки потім можна сподіватися на отримання прийнятного результату з використанням різних спеціальних застосунків.

У багатьох випадках розв'язання конкретних практичних задач з погляду пізнання обмежується рівнем гіпотези. Але головною метою наукових досліджень є отримання дійсно нових знань, які надалі можуть бути втілені в інновації та конкурентну перевагу вищого порядку.

Якщо з'являється розуміння, це означатиме, що знайдено дещо таке, що лежить в основі отриманого знання і може стати новим методом розв'язання зазначених задач. Знання — це відомості, а розуміння — здатність зробити висновки. Знання — це межа застосовності всіх інструментів *DM*: вони дають відомості у вигляді емпіричних і робочих гіпотез, але не дають розуміння, без якого неможливий перехід до наступного рівня пізнання — наукового методу.

І наостанок зазначимо, що поява систем на кшталт *GPT* аж ніяк не применшує і не замінює *DM*, оскільки *GPT* не може знайти потрібні знання через увесь перелік розглянутих вище питань — добір ознак для конкретної задачі, формування навчальної вибірки, перевірку й очищення даних, пошук емпіричної залежності та побудову моделі. Цей набір процедур і розумових операцій поки що повністю не алгоритмізований, а отже не може бути реалізований за допомогою систем штучного інтелекту.

*DM* — це дуже широка і різноманітна сфера, і її застосування всередині організації залежатиме від специфіки діяльності, цілей і завдань, доступних даних і ресурсів.

## СПИСОК ЛІТЕРАТУРИ

1. Закревский А.Д. Логика распознавания. Москва: Наука и техника, 1988. 118 с.
2. Гнеденко Б.В. Математика и научное познание. Москва: Знание, 1983. 64 с.
3. Малиновский Л. Процессы классификации — основа построения наук о действительности. *Алгоритмы обработки экспериментальных данных*. Москва: Наука, 1986. С. 155—182.
4. Jurs P., Eisenauer T. Pattern recognition in chemistry. New Jersey: Wiley, 1977. 184 p.
5. Кренделев Ф.П., Кренделев С.Ф. Эвристические методы в геологии. Москва: Наука, 1977. 152 с.
6. Алабин Б.К., Воронин Ю.А., Гольдин С.В., Гольдина Н.А., Еганов Э.А., Иванова М.Н. и др. Геология и математика. Новосибирск: Наука, 1970. 224 с.
7. Rastegari H., Sap M.N. Data mining and e-commerce: methods, applications, and challenges. *Jurnal Teknologi Maklumat*. 2008. No. 20 (2). P. 116—128.
8. Gupta A., Dengre V., Kheruwala H.A., Shah M. Comprehensive review of text-mining applications in finance. *Financial Innovation*. 2020. No. 6 (39). P. 1—25. <https://doi.org/10.1186/s40854-020-00205-1>
9. Daher J.B., Brun A., Boyer A. A Review on Heterogeneous, Multi-source and Multi-dimensional data mining. LORIA — Université de Lorraine, 2018. 13 p.
10. Waters D.J. The emerging digital infrastructure for research in the humanities. *International Journal on Digital Libraries*. 2023. Vol. 24. P. 87—102. <https://doi.org/10.1007/s00799-022-00332-3>
11. Németh R., Koltai J. The Potential of Automated Text Analytics in Social Knowledge Building. *Pathways Between Social Science and Computational Social Science*. Springer, 2021. P. 49—70. [https://doi.org/10.1007/978-3-030-54936-7\\_3](https://doi.org/10.1007/978-3-030-54936-7_3)

12. Leitgöb H., Prandner D., Wolbring T. Editorial: Big data and machine learning in sociology. *Front. Sociol.* 2023. Vol. 8. <https://doi.org/10.3389/fsoc.2023.1173155>
13. Strachan S., Stephen B., McArthur S. Practical Applications of Data Mining in Plant Monitoring and Diagnostics. *IEEE Power Engineering Society General Meeting*. 2007. P. 1—7.
14. Toledo M.R., Vázquez E.R., García-Salcedo R., Gómez-Vargas I., Uruchurtu E.S., et al. Data Mining applied to interventional cardiology procedures. *Journal of Physics: Conference Series*. 2021. Vol. 1723. No. 1. P. 012034. <https://doi.org/10.1088/1742-6596/1723/1/012034>
15. Khajehi M., Etemady F. Data Mining and Medical Research Studies. 2010 Second International Conference on Computational Intelligence, Modelling and Simulation. Bali, Indonesia, 2010. P. 119—122.
16. Algarni A. Data Mining in Education. *International Journal of Advanced Computer Science and Applications*. 2016. Vol. 7. No. 6. <https://doi.org/10.14569/ijacsa.2016.070659>
17. Yethiraj N. G. Applying Data Mining Techniques in the Field of Agriculture and Allied Sciences. *International Journal of Business Intelligence Research*. 2012. Vol. 1. No. 2. P. 40—42. <https://doi.org/10.20894/ijbi.105.001.002.004>
18. Polyakov M., Khanin I., Shevchenko G., Bilozubenko V. Data mining as a cognitive tool: capabilities and limits. *Управление знаниями и производительностью*. 2021. № 5 (1). С. 1—13.
19. Маркс К. Доход и его источники. Вульгарная политическая экономика. Москва: Изд-во политической литературы, 1955. 526 с.
20. Ханин І.Г. Ноосферний путь познання и хозяйствования. Дніпр: Нова ідеологія, 2018. 319 с.
21. Смирнов В.А. Проблемы логики научного познания. Москва: Институт философии АН СССР, 1964.
22. Чораян О.Г. Концепция вероятности и размытости в работе мозга. Ростов-на-Дону: Изд-во РГУ, 1987. 156 с.
23. Шапиро Д.И. К человеко-машинным методам решения одного класса задач. *Вопросы кибернетики. Теория кибернетики. Теория и практика ситуационного управления*. 1977. № 18. С. 82—88.
24. Shevchenko G.Ya., Bilozubenko V.S., Marchenko O.A. The formation of the corporate scientific environment. *Наука та наукознавство*. 2022. № 2 (116). С. 12—24. <https://doi.org/10.15407/sofs2022.02.012>
25. Олешко Д.Н., Крисилов В.А., Блажко А.А. Построение качественной обучающей выборки для прогнозирующих нейросетевых моделей. *Искусственный интеллект*. 2004. № 3. С. 567—573.
26. Василенко Ю.А., Шевченко Г.Я. Аналитический метод нахождения тестов. *Автоматика*. 1979. № 4. С. 3—8.
27. Tukey J.W. Exploratory Data Analysis. London: Addison-Wesley, 1977. 688 p.
28. Tukey J.W. The Future of Data Analysis. *Annals of Mathematical Statistics*. 1962. Vol. 33. No. 1. P. 1—67.
29. Ishikawa K. What is Total Quality Control? The Japanese Way. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1985. 215 p.
30. Taguchi G., Chowdhury S., Wu Y. Taguchi's Quality Engineering Handbook. Hoboken, New Jersey: John Wiley & Sons, Inc., 2005. 1662 p.

31. Fisher R.A. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*. 1936. Vol. 7. P. 179—188.

## REFERENCES

1. Zakrevsky, A.D. (1988). *The logic of recognition*. Moscow: Nauka i tekhnika [in Russian].
2. Gnedenko, B.V. (1983). *Mathematics and scientific cognition*. Moscow: Znaniye [in Russian].
3. Malinovsky, L. (1986). *Classification processes: the basis for the construction of sciences about reality. Algorithms of experimental data processing*. Moscow: Nauka [in Russian].
4. Jurs, P., & Eisenauer, T. (1977). *Pattern recognition in chemistry*. New Jersey: Wiley.
5. Krendelev, F.P., & Krendelev, S.F. (1977). *Heuristic methods in geology*. Moscow: Nauka [in Russian].
6. Alabin, B.K., Voronin, Y.A., Goldin, S.V., Goldina, N.A., Yeganov, E.A., Ivanova, M.N., & et al. (1970). *Geology and mathematics*. Novosibirsk: Nauka [in Russian].
7. Rastegari, H., & Sap, M.N. (2008). Data mining and e-commerce: methods, applications, and challenges. *Jurnal Teknologi Maklumat*, 20 (2), 116—128.
8. Gupta, A., Dengre, V., Kheruwala, H.A., & Shah, M. (2020). Comprehensive review of text-mining applications in finance. *Financial Innovation*, 6 (39), 1—25. <https://doi.org/10.1186/s40854-020-00205-1>
9. Daher, J.B., Brun, A., & Boyer, A.A. (2018). *Review on Heterogeneous, Multi-source and Multi-dimensional data mining*. LORIA — Université de Lorraine.
10. Waters, D.J. (2023). The emerging digital infrastructure for research in the humanities. *International Journal on Digital Libraries*, 24, 87—102. <https://doi.org/10.1007/s00799-022-00332-3>
11. Németh, R., & Koltai, J. (2021). The Potential of Automated Text Analytics in Social Knowledge Building. *Pathways Between Social Science and Computational Social Science*. Springer, 49—70. [https://doi.org/10.1007/978-3-030-54936-7\\_3](https://doi.org/10.1007/978-3-030-54936-7_3)
12. Leitgöb, H., Prandner, D., & Wolbring, T. (2023). Big data and machine learning in sociology. *Front. Sociol.*, 8. <https://doi.org/10.3389/fsoc.2023.1173155>
13. Strachan, S., Stephen, B., & McArthur, S. (2007). Practical Applications of Data Mining in Plant Monitoring and Diagnostics. *IEEE Power Engineering Society General Meeting*, 1—7.
14. Toledo M.R., Vázquez E.R., García-Salcedo R., Gómez-Vargas I., & Uruchurtu E.S., et al. (2021). Data Mining applied to interventional cardiology procedures. *Journal of Physics: Conference Series*, 1723 (1), 12—34, 012034. <https://doi.org/10.1088/1742-6596/1723/1/012034>
15. Khajehei, M., & Etemady, F. (2010). Data Mining and Medical Research Studies. *Second International Conference on Computational Intelligence, Modelling and Simulation*. Bali, Indonesia, 119—122.
16. Algarni, A. (2016). Data Mining in Education. *International Journal of Advanced Computer Science and Applications*, 7 (6). <https://doi.org/10.14569/ijacsa.2016.070659>
17. Yethiraj, N.G. (2012). Applying Data Mining Techniques in the Field of Agriculture and Allied Sciences. *International Journal of Business Intelligence Research*, 1 (2), 40—42. <https://doi.org/10.20894/ijbi.105.001.002.004>

18. Polyakov, M., Khanin, I., Shevchenko, G., & Bilozubenko, V. (2021). Data mining as a cognitive tool: capabilities and limits. *Knowledge and Performance Management*, 5 (1), 1—13.
19. Marx, K. (1955). *Revenue and its Sources. Vulgar Political Economy*. Moscow: Publishing House of Political Literature [in Russian].
20. Khanin, I.G. (2018). *Issues of noospheric development of economy and cognition*. Dnipro: Nova ideolohia [in Ukrainian].
21. Smirnov, V.A. (1964). *Problems of the logic of scientific cognition*. Moscow: Institute of Philosophy, Academy of Sciences of the USSR [in Russian].
22. Chorayan, O.G. (1987). *The concept of probability and fuzziness in the work of the brain*. Rostov-on-Don: Publishing house of Rostov State University [in Russian].
23. Shapiro, D.I. (1977). An introduction to human-machine methods for solving one class of problems. *Issues of Cybernetics. Theory and Practice of Situational Management*, 18, 82—88 [in Russian].
24. Shevchenko, G.Ya., Bilozubenko, V.S., & Marchenko, O.A. (2022). The formation of the corporate scientific environment. *Science and Science of Science*, 2 (116), 12—24. <https://doi.org/10.15407/sofs2022.02.012>
25. Oleshko, D.N., Krisilov, V.A., & Blazhko, A.A. (2004). Constructing a qualitative training sample for predictive neural network models. *Artificial Intelligence*, 3, 567—573 [in Russian].
26. Vasilenko, Yu.A., & Shevchenko, G.Ya. (1979). An analytical method of finding tests. *Automatics*, 4, 3—8 [in Russian].
27. Tukey, J.W. (1977). *Exploratory Data Analysis*. London: Addison-Wesley.
28. Tukey, J.W. (1962). The Future of Data Analysis. *Annals of Mathematical Statistics*, 33 (1), 1—67.
29. Ishikawa, K. (1985). *What is Total Quality Control? The Japanese Way*. Englewood Cliffs, NJ: Prentice-Hall, Inc.
30. Taguchi, G., Chowdhury, S., & Wu, Y. (2005). *Taguchi's Quality Engineering Handbook*. Hoboken, New Jersey: John Wiley & Sons.
31. Fisher, R.A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7, 179—188.

Одержано / Received 11.02.2025

Прорецензовано / Revised 26.03.2025

Підписано до друку / Accepted 01.04.2025

*M.V. Polyakov*<sup>1</sup>, Dsc (Economics), associate professor

e-mail: [info@noosphere.com](mailto:info@noosphere.com)

<https://orcid.org/0000-0001-7896-2486>

*I.G. Khanin*<sup>2</sup>, Dsc (Economics), professor

e-mail: [i.h.khanin@nuwm.edu.ua](mailto:i.h.khanin@nuwm.edu.ua)

<https://orcid.org/0000-0002-4221-2314>

*G.Ya. Shevchenko*<sup>1</sup>, PhD (Engineering), head of department

e-mail: [nikk.gena@gmail.com](mailto:nikk.gena@gmail.com)

<https://orcid.org/0000-0003-3984-9266>

*V.S. Bilozubenko*<sup>3</sup>, Dsc (Economics), professor

e-mail: [bvs910@gmail.com](mailto:bvs910@gmail.com)

<https://orcid.org/0000-0003-1269-7207>

*O.A. Marchenko*<sup>1</sup>, master, researcher

e-mail: [arche.oleg1@gmail.com](mailto:arche.oleg1@gmail.com)

<https://orcid.org/0000-0001-7665-7832>

<sup>1</sup> Association Noosphere

103-A, Gagarin avenue, Dnipro, 49000, Ukraine

<sup>2</sup> National University of Water and Environmental Engineering

11, Soborna str., Rivne, 33028, Ukraine

<sup>3</sup> University of Customs and Finance

2/4, Volodymyr Vernadskyi str., Dnipro, 49000, Ukraine

## DATA MINING FOR RESEARCH PURPOSES

The article considers the possibility of using Data Mining (DM) methods in a scientific study aimed at cognition, which requires an answer to the question ‘why?’, as well as the connection of DM methods with general questions of cognition. The research methodology is at the intersection of mathematics, statistics, computer science, and management theory. Along with the use of general scientific methods of cognition (generalization, systematization, abstraction, induction and deduction, analysis and synthesis, analogy, comparison, formalization, modelling, classification, categorization), special methods of analysis (logical, structural, functional), descriptive research method, and interpretive methodology are applied. The relationship of DM methods with the general theory of cognition is clarified. It is proved that DM methods are a new, human-machine methodology of empirical cognition. The authors point out the limitations of these methods: the results of their use are empirical regularities (ERs), i.e. probable (rather than exact) knowledge, on the basis of which instructions or decision-making rules are formulated, but remain at the first, lowest possible, empirical level of knowledge, which is typical for business. However, the main goal of scientific research is to obtain truly new knowledge that can be translated into innovations and provide a higher-order competitive advantage. ESMs can serve as ‘blanks’ for formulating, testing and selecting hypotheses in order to further deepen the knowledge of the subject area and obtain more reliable knowledge that brings the researcher closer to understanding. And if understanding is gained, it means that something has been found that underlies the knowledge gained, and this can become a new method of solving the tasks at hand. Knowledge is information, and understanding is the ability to draw conclusions. Today, this is the limit of applicability of all DM tools: they provide knowledge in the form of empirical and working hypotheses, but do not provide understanding, without which it is impossible to move to the next level of knowledge, the scientific method.

**Keywords:** *data mining, empirical regularity, training sample, feature space, data design.*