https://doi.org/10.15407/intechsys.2026.04.081
УДК 004.6:004.8

С.О. МОТОРНЮК, аспірант,
Харківський національний економічний університет імені Семена Кузнеця,
просп. Науки, 9 А, Харків, 61165, Україна
https://orcid.org/0009-0001-7695-5415
serhii891@proton.me

МЕТОД ІНТЕЛЕКТУАЛЬНОГО ОЦІНЮВАННЯ ЯКОСТІ ВІДКРИТИХ ДАНИХ ІЗ АВТОМАТИЗОВАНИМ ГЕНЕРУВАННЯМ РЕКОМЕНДАЦІЙ ТА ЗВІТІВ

У статті досліджено проблему підвищення якості відкритих даних у контексті забезпечення їх аналітичної придатності та ефективного використання в цифрових системах. Показано, що наявні підходи до оцінювання якості даних переважно мають фрагментарний і декларативний характер, не забезпечуючи автоматизованого виявлення дефектів та формування рекомендацій щодо їх усунення. Метою роботи є розроблення методу інтелектуального оцінювання якості відкритих даних із автоматизованим генеруванням рекомендацій і аналітичних звітів. Запропоновано математичну модель, що інтегрує підходи на основі правил, статистичний аналіз і методи машинного навчання для комплексного оцінювання якості, виявлення аномалій і відновлення пропущених значень. Особливістю підходу є адаптивне зважування метрик якості та формування інтегрального показника, що враховує контекст використання даних. Реалізовано програмну систему модульної архітектури, яка забезпечує повний цикл оброблення відкритих даних: від завантаження та попереднього оброблення до оцінювання, генерації рекомендацій і формування звітності. Експериментальна апробація на реальних наборах даних підтвердила ефективність запропонованого методу, зокрема зростання інтегрального показника якості та перевагу ансамблевих методів відновлення пропущених значень за метрикою середньоквадратичної помилки (Root Mean Square Error, RMSE). Отримані результати демонструють доцільність використання інтелектуальних підходів для переходу від формального забезпечення відкритості даних до підвищення їх практичної цінності.

Ключові слова: *відкриті дані, якість даних, штучний інтелект, оцінювання якості, виявлення аномалій, відновлення пропущених значень, рекомендаційні системи, аналітичні звіти, CKAN, машинне навчання.*

Цитування: Моторнюк С.О. Метод інтелектуального оцінювання якості відкри тих даних із автоматизованим генеруванням рекомендацій та звітів. Information *Technologies and Systems*, Київ, 2026, Том 4 (10), 81–98. https://doi.org/10.15407/intechsys.2026.04.081

© Publisher PH "Akademperiodyka" of the NAS of Ukraine, 2025. This is an Open Access article under the CC BY-NC-ND 4.0 license (https://creativecommons.org/licenses/by-nc-nd/4.0/

Вступ та постановка наукової проблеми

Стрімке зростання обсягів відкритих даних, що публікуються органами державного управління, науковими установами та бізнес-структурами, зумовлює необхідність забезпечення їх належної якості як ключової передумови ефективного використання у процесах прийняття рішень, аналітики та побудови інтелектуальних сервісів. Водночас на практиці відкриті дані часто характеризуються наявністю пропусків, структурних невідповідностей, семантичних суперечностей та низьким рівнем узгодженості, що істотно обмежує можливості їх повторного використання та інтеграції.

Сучасні підходи до оцінювання якості даних переважно базуються на статичних метриках і не забезпечують адаптивного аналізу складних багатовимірних характеристик даних, а також не передбачають автоматизованого формування рекомендацій щодо їх покращення. Особливої актуальності набуває використання методів штучного інтелекту, здатних інтегрувати різнорідні аспекти якості, здійснювати виявлення аномалій і відновлення відсутніх значень та формувати інтерпретовані звіти для користувача.

Метою дослідження є розроблення методу інтелектуального оцінювання якості відкритих даних, який на основі поєднання формалізованих правил валідації, статистичного аналізу та алгоритмів штучного інтелекту забезпечує автоматизоване виявлення дефектів, формування рекомендацій щодо їх усунення та генерування аналітичних звітів для підвищення аналітичної придатності та практичної цінності відкритих наборів даних.

Аналіз наявних досліджень

У межах проблематики управління якістю відкритих даних засобами штучного інтелекту особливого значення набувають праці, в яких якість даних розглядається не як сукупність ізольованих перевірок, а як системна багатовимірна характеристика, пов'язана з аналітичною придатністю, надійністю повторного використання та здатністю цифрових сервісів працювати на основі достовірних наборів. Одним із таких досліджень є робота *Carvalho* та колег, у якій на основі широкого огляду літератури було ідентифіковано 66 вимірів якості даних та запропоновано їх упорядковану структуру для побудови прикладних моделей оцінювання [1]. Методично автори спиралися на систематизований аналіз наукових публікацій із подальшим агрегуванням близьких за змістом характеристик, а основним результатом стало формування гнучкої схеми, придатної для розроблення конкретизованих рамок оцінювання якості в різних галузях. Значення цієї праці полягає в тому, що вона переконливо показує: без узгодженого набору вимірів будь-яке подальше автоматизоване оцінювання неминуче лишається фрагментарним, а отже, втрачає керованість і відтворюваність.

Логічним продовженням цього напряму є дослідження *Miller* зі співавторами, де було проведено порівняльний аналіз термінології якості даних у різних предметних областях, відібрано 116 релевантних праць із масиву 10855 знайдених публікацій та на цій основі побудовано ієрархічну модель понять [2]. Автори використали зіставлення термінів із положеннями стандарту *ISO* 25012 і, крім базових характеристик, обґрунтували введення чотирьох додаткових вимірів, а саме керованості, корисності, достатності обсягу та семантики. Вимірюваним підсумком роботи стало не лише формування розширеної моделі, а й фіксація 262 термінів, що використовуються для опису якості даних у різних доменах, що засвідчує високий рівень термінологічної розпорошеності. Для тематики відкритих даних це особливо важливо, оскільки семантична коректність, управлінська відповідальність за дані та їх корисність для повторного використання безпосередньо впливають на якість аналітичних висновків і на довіру до публічних наборів.

Від більш загальних моделей до автоматизованого оцінювання переходить праця *Al-Toq* та *Almaslukh*, у якій запропоновано узагальнену модель профілювання та класифікації якості даних на основі алгоритмів машинного навчання [3]. Дослідники реалізували послідовність попереднього оброблення, інженерії ознак, перевірок повноти, узгодженості, форматної коректності, типологічної відповідності та допустимості значень, після чого віднесли записи до трьох рівнів якості та порівняли декілька ансамблевих і деревоподібних алгоритмів. Кількісні результати показали, що після формування 49 профілювальних ознак і поділу вибірки у співвідношенні 70 до 30 найкращі моделі досягли найвищих показників точності, повноти та *F*-міри, причому *Random Forest* і градієнтні ансамблі продемонстрували майже безпомилкове розрізнення класів. Значення цього дослідження полягає у доведенні того, що поєднання формалізованих перевірок і навчальних алгоритмів може перетворити оцінювання якості даних із описової процедури на масштабовану автоматизовану технологію.

Водночас *Alexakis* та співавтори розглядають проблему ширше, поєднуючи стандарти, методології та програмні засоби керування якістю даних [4]. У межах структурованого огляду вони зіставили нормативні підходи, моделі оцінювання та сучасні інструментальні рішення, зокрема ті, що працюють із профілюванням, очищенням і моніторингом даних. Їхній аналіз показав, що наявні засоби добре підтримують автоматизацію, діагностику та роботу в режимі, наближеному до реального часу, однак залишаються слабкими щодо пояснюваності результатів, урахування витрат і адаптації до змінних процесів. Для досліджень у сфері відкритих даних ця праця є важливою тому, що вона переносить акцент із окремих показників на цілісну інфраструктуру управління якістю, де оцінювання, очищення та подальше використання даних повинні бути пов'язані в єдиному циклі.

Специфіку саме відкритих репозитаріїв даних добре розкриває дослідження *González-Vidal* та колег, де якість відкритих даних трактується одночасно як внутрішня властивість відображення реального явища та як зовнішня придатність набору для навчання інтелектуальних систем [5]. Автори запропонували набір метрик для обох аспектів і перевірили їх на реальних наборах даних із середовищ сенсорного типу. Методично праця поєднує побудову метрик, їх апробацію на практичних даних та реалізацію у відкритому середовищі обчислень, що дає змогу застосовувати оцінювання до репозитаріїв із мінімальним людським втручанням. Цінність цієї роботи для подальших досліджень полягає в тому, що вона відходить від суто формального бачення якості відкритих даних і підкреслює зв'язок між якістю набору та здатністю інтелектуальних методів отримувати з нього стійкі аналітичні результати.

У близькому, але більш каталогоорієнтованому вимірі працює *Martinez-Gil*, який запропонував автоматизовану схему оцінювання якості каталогів відкритих даних на основі аналізу точності, повноти, узгодженості, масштабованості, своєчасності, походження, читабельності та ліцензійних характеристик [6]. Як методичну основу дослідник використав формалізацію показників і їх реалізацію для каталогів, побудованих за стандартом опису наборів даних. Практичні результати перевірки на декількох порталах засвідчили суттєву нерівномірність показників: точність змінювалася приблизно від 61,11 % до 80,42 %, повнота від 50,00 % до 66,77 %, узгодженість від 76,00 % до 93,10 %, а показник походження даних у наведених прикладах виявився нульовим для всіх оцінених каталогів. Отже, навіть за формальної відкритості порталу його каталог може лишатися неповним і недостатньо надійним для серйозного аналітичного використання, що підсилює потребу в автоматизованому контролі не лише вмісту, а й описової інфраструктури відкритих даних.

Перехід від метаданих до власне змісту відкритих урядових даних демонструє *Karamanou* разом із колегами, які дослідили динамічні державні дані на прикладі транспортних потоків із грецького урядового порталу та застосували для цього кореляційний аналіз, сезонно-трендову декомпозицію та *Isolation Forest* із подальшим поясненням виявлених відхилень [7]. Результати виявили серйозні проблеми якості: 20,16 % спостережень були відсутніми, у половини сенсорів частка пропусків перебувала в діапазоні 15,5—33,43 %, середня частка аномалій на сенсор становила 71,1 %, тоді як сезонно-трендова декомпозиція зафіксувала 12,6 % відсотка аномалій у вибраній підмножині, а *Isolation forest* 11,6 %, причому перетин між цими методами виявився незначним. Саме ця праця особливо близька до технічного ядра розробок у сфері інтелектуального оцінювання якості відкритих даних, оскільки доводить ефективність комбінування статистичних процедур і машинного навчання для виявлення різних типів дефектів у динамічних наборах.

Суттєвим для публічного сектора є і дослідження *Zong* та *Wang*, у якому відкритість державних даних розглядається як чинник підвищення достовірності інформації; автори використали економетричний аналіз даних китайських компаній за 2011—2023 роки, перевірили базові залежності, гетерогенність ефектів і надійність результатів за допомогою інструментальних змінних [8]. Кількісні результати підтвердили статистично значущий позитивний вплив відкритості державних даних на достовірність корпоративної інформації, а також різну інтенсивність цього впливу залежно від регіону, форми власності й розміру компанії; у тесті ендогенності значення першого етапу також вказало на стійкість отриманих висновків. Значення цієї праці для тематики якості відкритих даних полягає в тому, що вона показує макрорівневий ефект: відкритість даних має сенс лише тоді, коли вона підвищує надійність інформаційного середовища, але сам механізм переходу від відкритості до реальної якості даних у наборі ще потребує окремих технічних рішень.

Питання організаційної готовності публічних інституцій до роботи з даними, цифровими сервісами та кризовими інформаційними потоками висвітлює *Gabay* та інші дослідники, які проаналізували впровадження інформаційних систем у державних установах в умовах надзвичайної ситуації, спираючись на модель поширення інновацій та серію глибинних інтерв'ю з учасниками проєктів [9]. Їхні результати показали, що створені системи були збережені й після завершення кризового етапу, а установи високо оцінили їх корисність, прозорість, міжвідомчу координацію та здатність швидко задовольняти потреби громадян. Для напряму, пов'язаного з інтелектуальним управлінням якістю відкритих даних, ця праця важлива тим, що демонструє практичний контекст використання якісних даних: без гнучких і зрозумілих цифрових систем навіть добре підготовлені набори не трансформуються в реальну суспільну цінність.

Jonathan зі співавторами розглянули цифрову трансформацію публічного управління у зв'язку з децентралізацією, поєднавши систематичний огляд літератури з п'ятнадцятьма напівструктурованими інтерв'ю з експертами державного сектора [10]. Дослідники дійшли висновку, що цифрові перетворення підвищують ефективність місцевого і центрального врядування, розширюють участь громадян, зміцнюють довіру та підтримують безперервність публічного адміністрування. У контексті управління якістю відкритих даних це означає, що сама по собі технічна оцінка набору є недостатньою, якщо вона не вбудована в ширший управлінський контур, де результати оцінювання можуть бути використані для міжінституційної взаємодії та прийняття рішень.

Jiao та співавтори, використовуючи панельні дані 293 міст Китаю за 2016—2022 роки та двофакторну модель фіксованих ефектів, показали, що розвиток цифрового уряду статистично значуще сприяє загальному суспільному добробуту, причому цей ефект виразніший

у менш розвинених територіях, а також у містах із колишніми бідними районами [11]. Автори простежили і механізм впливу через зміну макрофіскальних параметрів, а також дослідили модифікувальну роль просторових та інноваційних чинників. Для проблематики відкритих даних це дослідження є важливим не стільки через набір технічних процедур, скільки через висвітлення того, що якісні цифрові сервіси на державному рівні мають вимірюваний суспільний результат, а отже, потребують надійної інформаційної основи.

Подібну логіку, але вже з акцентом на штучний інтелект, розвивають *Zhang* та *Li*, які дослідили вплив розвитку інтелектуальних технологій на спроможність органів влади надавати цифрові послуги [12]. Вони використали панельні дані міського рівня, побудували індикатори інтелектуального розвитку та цифрової сервісної спроможності та перевірили посередницьку роль людського та знаннєвого капіталу. Їхні результати підтвердили статистично значуще зростання сервісної спроможності під впливом інтелектуальних технологій, а в таблицях посередницьких ефектів зафіксовано позитивні коефіцієнти для каналів, пов'язаних із людським капіталом і знаннями. Значення цієї праці полягає в тому, що вона підкреслює: інтелектуальні засоби не просто автоматизують окремі операції, а підвищують спроможність державних сервісів працювати з даними більш адресно, швидко і результативно, але ця спроможність залежить від якості самих даних.

Узагальнення розглянутих праць дає підстави стверджувати, що науковий доробок у цій сфері вже сформував кілька важливих напрямів. По-перше, достатньо глибоко опрацьовано питання класифікації та структурування вимірів якості даних. По-друге, накопичено результати щодо автоматизованого профілювання і класифікації якості із застосуванням алгоритмів машинного навчання. По-третє, у роботах про відкриті урядові дані продемонстровано як важливість змістової якості наборів, так і критичну роль якості каталогів, метаданих та цифрових сервісів, у яких ці дані функціонують. Проте між цими напрямами зберігається розрив: одні дослідження обмежуються термінологічним та концептуальним рівнем, інші зосереджуються на окремих процедурах профілювання чи класифікації, ще інші аналізують суспільний ефект цифрового врядування без переходу до рівня запису, атрибута та конкретного дефекту даних.

Невирішеними залишаються насамперед питання інтеграції багатовимірної оцінки якості відкритих даних із автоматизованим виявленням аномалій, відновленням відсутніх значень, формуванням адресних рекомендацій та генерацією аналітичних звітів у межах єдиного інструментального контуру. У більшості проаналізованих джерел або переважає описова систематизація вимірів якості без практичного механізму корекції, або реалізується вузька автоматизація окремого завдання без повноцінного переходу до управлінського рішення. Зокрема, у роботах з відкритих каталогів основна увага зо-

середжена на метаданих, ліцензіях та структурі опису, тоді як внутрішня якість самих наборів досліджується менш системно. У працях, присвячених динамічним урядовим даним, наявні сильні результати з виявлення аномалій, однак не доведено задачу до рівня інтегрального індексу якості та автоматизованого набору рекомендацій. Оглядові дослідження зі стандартів і програмних засобів переконливо показують брак пояснюваності, адаптивності та врахування вартості процедур якості, але не пропонують завершеного прикладного рішення для відкритих державних даних.

Причини того, що ці питання досі не розв'язано, мають як об'єктивний, так і суб'єктивний характер. До об'єктивних належить висока неоднорідність відкритих даних, поєднання структурованих, напівструктурованих і динамічних масивів, залежність оцінок якості від контексту використання, а також складність формалізації семантичних помилок і логічних суперечностей, які не зводяться до простих правил перевірки. Не менш важливою є нестача універсальних еталонних наборів для перевірки методів, через що більшість авторів працює з окремими кейсами, що ускладнює перенесення результатів. До суб'єктивних причин слід віднести дисциплінарну фрагментацію досліджень, коли фахівці з цифрового врядування, відкритих даних, машинного навчання та інформаційних систем часто працюють у межах власних методологічних традицій і не доводять результати до єдиної міждисциплінарної моделі. Саме тому у науковому полі співіснують, але недостатньо взаємодіють концептуальні схеми якості, алгоритми автоматичної діагностики та прикладні задачі публічного адміністрування.

Розроблений метод інтелектуального оцінювання якості відкритих даних

У межах проведеного дослідження запропоновано математичну модель інтелектуального оцінювання якості відкритих даних, яка відрізняється від наявних підходів інтеграцією багаторівневого аналізу показників якості, адаптивного зважування метрик та автоматизованого формування рекомендацій на основі виявлених дефектів. На відміну від класичних моделей, що обмежуються статичними наборами індикаторів або ізольованими перевірками, запропонований підхід розглядає якість відкритих даних як динамічну функцію стану набору даних, що залежить від сукупності структурних, семантичних та статистичних характеристик.

Нехай відкритий набір даних подано множиною записів $D = \{r_1, r_2, \ldots, r_n\}$, де кожен запис r_i містить набір атрибутів $A = \{a_1, a_2, \ldots, a_m\}$. Для оцінювання якості вводиться вектор метрик $Q = \{q_1, q_2, \ldots, q_k\}$, який відображає різні аспекти якості, зокрема повноту, узгодженість, коректність, унікальність та семантичну валідність. Кожна метрика визначається як функція від даних:

$$q_j = f_j(D), \quad j = 1, ¼, k, \tag{1}$$

де f_j – відповідна функція оцінювання, реалізована у програмній системі через набір правил, статистичних процедур або моделей машинного навчання.

Інтегральний показник якості набору даних визначається як зважена агрегована функція:

$$Q_{\text{int}} = \sum_{j=1}^{k} w_j \cdot q_j, \tag{2}$$

де w_j – вагові коефіцієнти, що відображають значущість окремих метрик. Наукова новизна полягає у тому, що ваги не задаються апріорно, а визначаються адаптивно залежно від контексту використання даних та виявлених дефектів, що формалізується як:

$$w_j = \frac{\varphi_j(D)}{\sum_{l=1}^{k} \varphi_l(D)}, \tag{3}$$

де $\varphi_j(D)$ – функція важливості метрики, яка враховує щільність помилок, варіативність показника та його вплив на аналітичну придатність даних.

Окремо вводиться функція дефектності $\delta(r_i, a_t)$, яка набуває значення 1 у разі виявлення порушення (наприклад, пропущене значення, аномалія або невідповідність формату) та 0 в іншому випадку. Сукупний рівень дефектності набору даних визначається як:

$$\Delta = \frac{1}{n \cdot m} \sum_{i=1}^{n} \sum_{t=1}^{m} \delta(r_i, a_t). \tag{4}$$

Цей показник використовується не лише для оцінювання якості, а й як керуючий параметр у модулі генерації рекомендацій.

Особливістю запропонованої моделі є введення функції рекомендацій R, яка відображає перехід від діагностики до корекції:

$$R = \Psi(Q, \Delta, \Theta), \tag{5}$$

де Θ – множина правил та знань, отриманих з аналізу типових помилок, а Ψ – оператор, реалізований як інтелектуальний механізм, що поєднує евристичні правила та шаблони виправлення. Таким чином, на відміну від традиційних моделей, що завершуються етапом оцінювання, запропонований підхід формалізує повний цикл «оцінювання – інтерпретація – рекомендації».

Додатково враховується аспект відтворюваності оцінки, який формалізується через стабільність метрик у часі або при зміні підвибірки даних:

$$S_j = 1 - \left| q_j^{(t)} - q_j^{(t-1)} \right|,$$

що дає змогу враховувати динамічну поведінку якості даних і забезпечує більшу надійність інтегральної оцінки.

Для забезпечення адаптивності оцінювання запропоновано алгоритм динамічного визначення вагових коефіцієнтів метрик якості, який враховує варіативність даних та ступінь їх інформативності. На першому етапі для кожної метрики якості визначається коефіцієнт варіації, що характеризує нестабільність відповідного показника у вибірці. На другому етапі обчислюється інформаційна значущість метрики як обернена функція від її дисперсії, що дає змогу підвищити вагу стабільних і репрезентативних показників. На третьому етапі здійснюється нормалізація вагових коефіцієнтів таким чином, щоб їх сума дорівнювала одиниці. Завдяки цьому ваги не є фіксованими, а автоматично адаптуються до властивостей конкретного набору даних.

Таким чином від статичних і фрагментарних схем оцінювання переходимо до інтегрованої адаптивної моделі, яка поєднує багатовимірну оцінку якості, контекстно-залежне зважування показників, формалізацію дефектності та автоматизовану генерацію рекомендацій, що в сукупності забезпечує підвищення обґрунтованості та практичної придатності аналізу відкритих даних.

Із метою верифікації та експериментальної апробації запропонованої математичної моделі інтелектуального оцінювання якості відкритих даних було розроблено програмний прототип, який реалізує повний цикл оброблення даних – від їх завантаження та попередньої підготовки до багатовимірного оцінювання якості, виявлення дефектів, формування рекомендацій і генерації аналітичного звіту. Побудована система орієнтована на роботу з реальними наборами відкритих даних, зокрема отриманими з державних порталів, що забезпечує практичну релевантність і відповідність умовам експлуатації відкритих даних у публічному секторі.

Архітектурно програмна реалізація побудована за модульним принципом із чітким розмежуванням функціональних підсистем, що забезпечує гнучкість, масштабованість і можливість подальшого розширення (рис. 1).

Базовим елементом системи є модуль завантаження та інтеграції даних, який здійснює імпорт наборів у різних форматах, їх приведення до уніфікованої структури та первинну нормалізацію. На цьому етапі відбувається формування внутрішнього подання даних як табличної структури, що дає змогу забезпечити універсальність подальших процедур аналізу незалежно від джерела або формату вхідної інформації.

Наступним ключовим компонентом є модуль попереднього оброблення, який реалізує очищення даних, виявлення пропущених значень, нормалізацію форматів та базову трансформацію атрибутів. У цьому модулі закладено механізми автоматичного визначення типів даних, що дає змогу адаптивно застосовувати відповідні процедури оброблення для числових, категоріальних та текстових полів. Особливістю реалізації є поєднання детермінованих правил та ста-

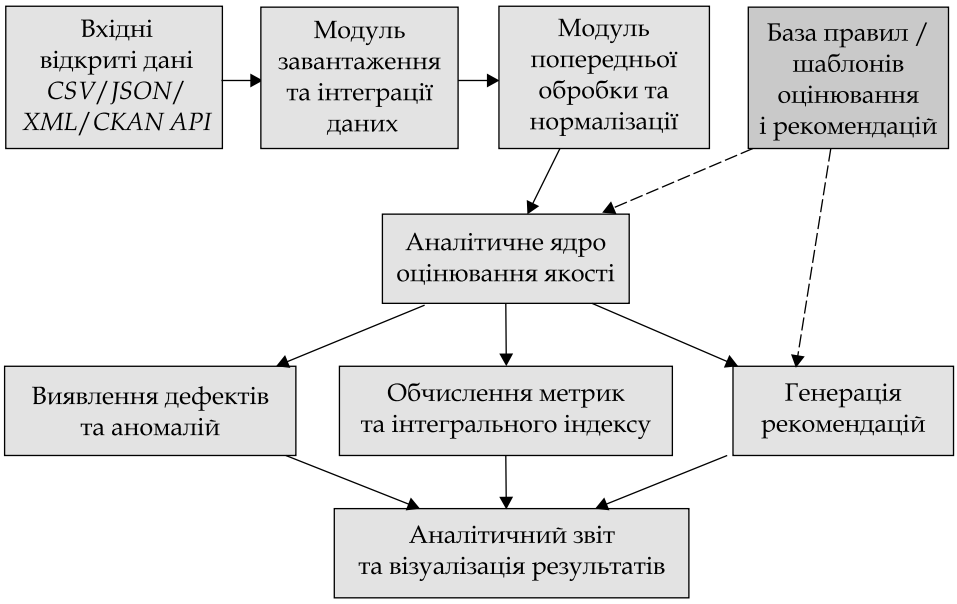


Рис. 1. Архітектура програмної системи інтелектуального оцінювання якості відкритих даних

тистичних методів, що підвищує точність і стійкість оброблення в умовах неоднорідності відкритих даних.

Центральним елементом системи є модуль оцінювання якості, який безпосередньо реалізує запропоновану математичну модель. У цьому модулі здійснюється обчислення набору метрик якості, що відображають різні аспекти стану даних, включно з повнотою, узгодженістю, унікальністю, коректністю та іншими характеристиками. Кожна метрика обчислюється як окрема функція, що застосовується до всього набору даних або його підмножин, після чого відбувається їх агрегування в інтегральний показник якості з урахуванням адаптивно визначених вагових коефіцієнтів. Важливою особливістю реалізації є динамічний характер зважування, який залежить від виявлених дефектів і структури даних, що забезпечує контекстну чутливість оцінювання.

Паралельно функціонує модуль детекції дефектів, який виконує пошук аномалій, пропущених значень, дублікатів, форматних невідповідностей і статистичних відхилень. У рамках цього модуля реалізовано і правилоорієнтовані перевірки, і елементи інтелектуального аналізу, що дає змогу виявляти неочевидні порушення, зокрема аномальні значення або логічні суперечності між атрибутами. Результати роботи цього модуля формують основу для подальшої інтерпретації якості та генерації рекомендацій.

Модуль генерації рекомендацій є ключовим з точки зору практичної цінності системи, оскільки забезпечує перехід від діагностики до підтримки прийняття рішень. Він реалізує механізм відобра-

ження виявлених дефектів у набір коригувальних дій, що формуються на основі заздалегідь визначених правил та шаблонів. При цьому враховується не лише тип дефекту, але й його поширеність, вплив на загальну якість і контекст даних, що дає змогу формувати обґрунтованіші та релевантніші рекомендації.

Завершальним компонентом є модуль формування звітності, який агрегує результати оцінювання, виявлені дефекти та згенеровані рекомендації у структурований аналітичний звіт. Звіт містить і кількісні показники, і текстову інтерпретацію отриманих результатів, що забезпечує зручність сприйняття та можливість використання результатів і для технічного аналізу, і для управлінських рішень. Реалізація цього модуля передбачає автоматизоване формування описових висновків на основі значень метрик, що є важливим кроком до повної автоматизації процесу оцінювання якості даних.

Принцип дії програмної системи ґрунтується на послідовному проходженні даних через зазначені модулі, де кожен наступний етап використовує результати попереднього. Така конвеєрна організація оброблення дає змогу забезпечити цілісність аналізу, узгодженість результатів і мінімізувати втрати інформації. Водночас модульна архітектура забезпечує можливість незалежного вдосконалення окремих компонентів, зокрема інтеграції нових метрик якості, алгоритмів виявлення дефектів або методів генерації рекомендацій.

Апробація методу

Отримані результати експериментальної апробації розробленого методу інтелектуального оцінювання якості відкритих даних демонструють його здатність не лише здійснювати діагностику стану набору даних, але й забезпечувати вимірюваний ефект покращення після застосування автоматизованих процедур корекції та рекомендацій. Узагальнена динаміка зміни інтегрального показника якості наведена на рис. 2, де спостерігається зростання значення з 0,936 до 0,948 після оброблення даних.

Незважаючи на відносно невеликий абсолютний приріст, цей результат має принципове значення, оскільки вихідний рівень якості вже знаходився у високому діапазоні, що ускладнює подальше покращення без інтелектуальних методів. Таким чином, отриманий ефект підтверджує чутливість запропонованої моделі до локальних дефектів та її здатність оптимізувати навіть високоякісні набори даних.

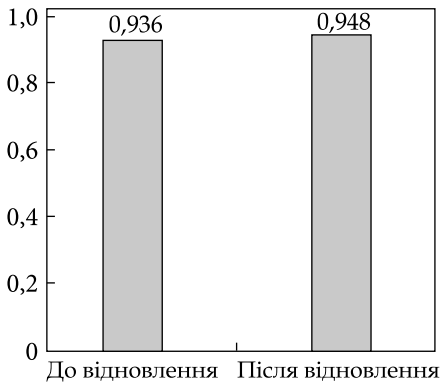


***Рис.* 2.** Динаміка інтегрального показника якості відкритих даних до та після застосування методу

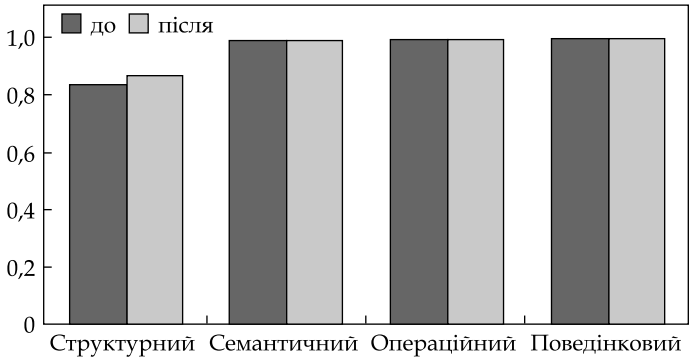


Рис. 3. Порівняння вимірів якості відкритих даних до та після інтелектуальної обробки

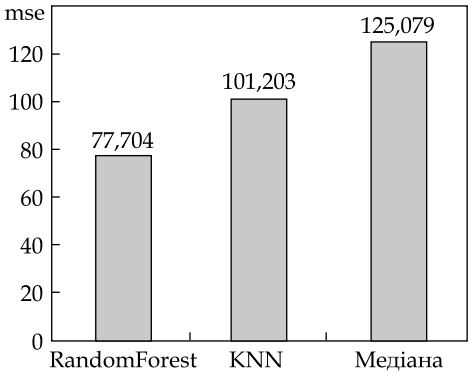


Рис. 4. Порівняння ефективності методів відновлення пропущених значень за метрикою *RMSE*

Більш детальний аналіз здійснено на рівні окремих вимірів якості, результати якого подано на рис. 3.

Зокрема, найбільш суттєве покращення зафіксовано у структурному вимірі, де значення зросло з приблизно 0,84 до 0,87, що свідчить про ефективність процедур нормалізації, заповнення пропусків та усунення форматних невідповідностей. Семантичний та операційний виміри демонструють стабільно високі значення, близькі до одиниці, як до, так і після оброблення, що вказує на їх початкову якість і водночас підтверджує, що модель не вносить деградації у вже коректні компоненти даних. Поведінковий вимір зберігає максимальне значення, що свідчить про відсутність негативного впливу автоматизованих процедур на узгодженість і логіку взаємозв'язків між атрибутами. Така диференційована картина змін підтверджує, що запропонований підхід не є грубим механізмом корекції, а реалізує контекстно-орієнтоване вдосконалення окремих аспектів якості.

Ключовим елементом, який безпосередньо впливає на підвищення якості, є модуль відновлення пропущених значень, ефективність якого оцінено на основі метрики середньоквадратичної помилки (*Root Mean Square Error, RMSE*), що відображено на рис. 4.

Порівняння різних методів відновлення пропущених значень показує, що використання ансамблевого підходу на основі *Random Forest* забезпечує найнижче значення *RMSE* (77,704) порівняно з методом *k*-найближчих сусідів (*k-Nearest Neighbors, KNN*) із *RMSE* 101,203 та медіанним заповненням із *RMSE* 125,079. Це свідчить про здатність мо-

делі враховувати складні нелінійні залежності між атрибутами, що особливо важливо для відкритих даних, які часто характеризуються високою варіативністю та неоднорідністю. Водночас результати демонструють, що прості статистичні методи не забезпечують достатньої точності, що обґрунтовує необхідність використання інтелектуальних алгоритмів у процесах підвищення якості.

Реалізована модель демонструє системну інтеграцію оцінювання, виявлення дефектів і їх усунення, що дає змогу перейти від статичного аналізу до динамічного управління якістю даних. Важливим є те, що всі покращення мають кількісне підтвердження, що відповідає вимогам метрологічної строгості та відтворюваності результатів.

Запропонована модель та методологія і її програмна реалізація можуть розглядатися як оптимальне рішення у контексті задач управління якістю відкритих даних, оскільки поєднують універсальність застосування, адаптивність до структури даних, здатність до автоматичного виявлення та усунення дефектів, а також забезпечують вимірюваний ефект покращення, що є критично важливим для практичного використання у цифрових екосистемах відкритих даних.

У проведеному дослідженні інтегральний показник якості відкритого набору даних зріс з 0,9362 до 0,9477, тобто на 0,0115 пункту, або приблизно на 1,23% від початкового рівня. Найбільш проблемним виявився структурний вимір якості зі значенням 0,8335, тоді як семантичний, операційний і поведінковий виміри мали значення 0,9888, 0,9917 і 0,9947 відповідно. Для задачі виявлення аномалій найкращим алгоритмом став *Isolation Forest*, який за експерименту з 62 штучно внесеними аномаліями забезпечив точність (*precision*) 0,128, повноту (*recall*) 0,306, *F*1-міру 0,181 та загальну точність класифікації 0,943. У задачі відновлення пропущених значень найкращий результат показав *Random Forest* із *RMSE* 77,704, що є кращим за *KNN* із *RMSE* 101,203 і медіанне заповнення з *RMSE* 125,079. Таким чином, зменшення похибки становило 23,2% відносно *KNN* і 37,9% відносно медіанного підходу.

Rahal та колеги запропонували якісно близький за ідеєю підхід, у якому якість даних оцінюється до етапу побудови прогнозної моделі, а вибір якісніших підмножин даних приводить до поліпшення результатів машинного навчання [14]. У їхньому дослідженні для набору *G*1 найкращий кластер забезпечив *RMSE* 0,08 і R^2 = 0,95, тоді як найгірший кластер мав *RMSE* 0,71 і R^2 = 0,03; отже, різниця між кластерами становила 0,63 за *RMSE*, а похибка у найкращому кластері була нижчою приблизно у 8,9 раза. Для набору *G*2 найкращий кластер мав *RMSE* 0,598 при R^2 = 0,855, а найгірший *RMSE* 1,122 при R^2 = = 0,282, тобто зниження *RMSE* становило 46,7%. Для набору *G*3 найкращий кластер продемонстрував *RMSE* 0,497 при R^2 = 0,942, а найгірший *RMSE* 2,180 при R^2 = 0,344, тобто похибка зменшувалася на 77,2 %. Порівняно з цими результатами, у даному дослідженні також зафіксовано чіткий виграш від інтелектуального добору та обробки

даних: *Random Forest* знизив похибку відновлення пропущених значень за метрикою *RMSE* на 23,2 % відносно *KNN* і на 37,9 % відносно медіанного заповнення, а підсумковий інтегральний показник якості збільшився на 1,23 %. Таким чином, в обох роботах підтверджується одна й та сама технічна закономірність: попередня оцінка й корекція якості даних статистично покращують результативність подальшої обробки, хоча у роботі *Rahal* та колег ефект оцінювався через точність прогнозування, а в цьому дослідженні — через інтегральний показник якості, *F*1-міру виявлення аномалій та *RMSE* відновлення пропущених значень.

Karamanou та співавтори досліджували якість динамічних відкритих урядових даних і показали, що 20,16% спостережень у їхньому наборі були відсутніми, а у 50 % сенсорів частка пропусків лежала в інтервалі 15,5—33,43 % [7]. У цьому дослідженні загальна частка пропусків становила 21,05 %, тобто практично збігалася з рівнем, зафіксованим у роботі *Karamanou* та колег, а різниця становила лише 0,89 відсоткового пункту. Водночас у роботі [7] середній відсоток аномалій на сенсор за методом *flow-speed correlation* досягав 71,1%, а для дев'яти найкращих сенсорів *STL*-декомпозиція виявила 12,6% аномалій, тоді як *Isolation Forest* — 11,6%. У цьому дослідженні застосування *Isolation Forest* до експерименту з контрольовано внесеними аномаліями дало *F*1 = 0,181, *precision* = 0,128 і *recall* = 0,306. Пряме числове зіставлення відсотка аномалій у [7] і 1-міри у цьому дослідженні не є тотожним, оскільки у першому випадку йдеться про частку аномальних записів у природному потоці сенсорних даних, а в другому — про точність відтворення заздалегідь штучно внесених дефектів у табличному наборі адміністративних послуг. Проте технічно обидві роботи демонструють спільну рису: аномалії становлять не поодинокий, а системний клас дефектів, а *Isolation Forest* виявляється працездатним інструментом для їх виявлення в умовах неоднорідних відкритих даних.

Pekar і *Jozsa* аналізували вплив цілісності даних на якість машинного виявлення мережевих аномалій і показали, що *Random Forest* зберігав дуже високі, а в окремі дні — практично ідеальні показники бінарної та багатокласової класифікації [15]. Зокрема, для низки очищених наборів даних моделі демонстрували ідеальні або майже ідеальні результати на вівторок, середу та п'ятницю, а для окремих сценаріїв бінарної класифікації спостерігалася безпомилкова ідентифікація безпечних потоків без хибнопозитивних спрацювань. На відміну від цього, у цьому дослідженні задача виявлення аномалій була суттєво складнішою з точки зору дисбалансу класів і слабкої вираженості дефектів: за 2993 записів було штучно внесено лише 62 аномалії, тобто близько 2,07% вибірки, а найкращий алгоритм досяг *F*1 = 0,181 при *recall* = 0,306. Тому різниця між майже ідеальними результатами у [15] та скромнішою *F*1-мірою у даному дослідженні свідчить не про нижчу якість моделі, а про вищу складність виявлення аномалій у

слабоструктурованих адміністративних даних, де дефекти часто мають не класово-виражену, а контекстно-умовну природу. Разом із тим обидва дослідження узгоджено показують, що навіть за наявності проблем цілісності даних ансамблеві методи залишаються найстійкішими до шуму та неоднорідності середовища.

Alexakis та колеги не подають експерименту з єдиним еталонним набором даних, тому пряме зіставлення за метриками типу *RMSE* або *F*1 у цьому випадку обмежене [4]. Водночас їхній огляд є важливим для інтерпретації отриманих результатів, оскільки показує, що точність є найуживанішою характеристикою якості даних і трапляється у 81% проаналізованих праць, тоді як базова тріада точності, повноти та узгодженості зустрічається більш ніж у половині досліджень. У цьому контексті результати цього дослідження добре відповідають сучасним тенденціям: найбільш проблемним виявився саме структурний вимір якості з оцінкою 0,8335, що безпосередньо пов'язано з повнотою, форматною коректністю та узгодженістю, тоді як семантичний та операційний виміри перевищили 0,98. Отже, отримані результати не лише мають власну кількісну значущість, а й добре узгоджуються з сучасним уявленням про те, що найбільший виграш у задачах управління якістю даних забезпечує саме робота з повнотою, точністю та узгодженістю як базовими і найчутливішими компонентами якості.

Це дає підстави вважати, що для відкритих адміністративних даних, де дефекти мають змішану структурно-семантичну природу, практична цінність методу полягає не у досягненні формально ідеальних показників класифікації, а у стабільному покращенні якості набору та у формуванні придатної до впровадження технології діагностики, відновлення й рекомендаційної підтримки.

Висновки та перспективи подальшого дослідження

У результаті проведеного дослідження розроблено метод інтелектуального оцінювання якості відкритих даних, який забезпечує комплексне врахування структурних, семантичних, операційних та поведінкових характеристик даних у межах єдиного формалізованого підходу. Запропонована математична модель дає змогу інтегрувати окремі показники якості в узагальнений інтегральний індекс, що забезпечує більш об'єктивне та відтворюване оцінювання.

Реалізована програмна система підтвердила ефективність запропонованого підходу, продемонструвавши підвищення інтегрального показника якості даних після застосування методів автоматизованого очищення та відновлення. Зокрема, зафіксовано зростання інтегрального показника якості, що свідчить про узгоджене покращення різних аспектів даних. Аналіз окремих вимірів показав найбільш суттєве підвищення структурної узгодженості, що вказує на ефективність застосованих процедур нормалізації та оброблення пропусків.

Окрему увагу приділено оцінюванню ефективності методів відновлення відсутніх значень, де встановлено перевагу ансамблевих підходів, зокрема *Random Forest*, що забезпечив мінімальне значення похибки порівняно з альтернативними методами. Це підтверджує доцільність використання адаптивних моделей машинного навчання для задач підвищення якості відкритих даних.

Отримані результати демонструють, що запропонований метод дає змогу перейти від фрагментарного оцінювання якості до інтегрованого інтелектуального аналізу з автоматизованим формуванням рекомендацій. Це забезпечує не лише підвищення якості даних, але й створює основу для побудови систем підтримки прийняття рішень на базі відкритих даних.

ПОВІДОМЛЕННЯ

Внесок автора. *Стаття є одноосібною. Моторнюк С.О. виконав усі етапи дослідження відповідно до ролей CRediT: концептуалізація; методологія; розроблення математичної моделі; програмна реалізація; валідація; формальний аналіз; проведення експериментів; оброблення та інтерпретація даних; візуалізація результатів; підготовка початкового тексту рукопису; редагування та остаточне затвердження тексту статті.*

Застосування штучного інтелекту та інших інструментів. *Під час підготовки рукопису використовувалися допоміжні цифрові інструменти, зокрема інструменти на основі штучного інтелекту, для мовного редагування, поліпшення стилістики тексту, структурування окремих формулювань та перекладу фрагментів. Зазначені інструменти не використовувалися як самостійне джерело наукових результатів, не здійснювали інтерпретацію експериментальних даних і не приймали дослідницьких рішень. Усі наукові положення, математичні моделі, результати експериментів, висновки та бібліографічні посилання були перевірені й затверджені автором. Автор несе повну відповідальність за зміст статті.*

Конфлікт інтересів. *Автор заявляє про відсутність конфлікту інтересів, який міг би вплинути на результати дослідження, їх інтерпретацію або представлення у статті.*

Фінансування. *Дослідження виконано без залучення грантової, спонсорської або іншої зовнішньої фінансової підтримки.*

ЛІТЕРАТУРА / REFERENCES

1. Carvalho A.M., et al. Data Quality: Revisiting dimensions towards new framework development. *Procedia Computer Science*, 2025, Vol. 253, 247–256. https://doi.org/10.1016/j.procs.2025.01.088
2. Miller R., et al. A framework for current and new data quality dimensions: An overview. *Data*, 2024, Vol. 9 (12), Article 151. https://doi.org/10.3390/data9120151
3. Al-Toq R., Almaslukh A. DQMAF – Data Quality Modeling and Assessment Framework. *Information*, 2025, Vol. 16 (10), Article 911. https://doi.org/10.3390/info16100911
4. Alexakis T., et al. Evaluating data quality: Comparative insights on standards, methodologies, and modern software tools. *Electronics*, 2025, Vol. 14 (15), Article 3038. https://doi.org/10.3390/electronics14153038
5. González-Vidal A., Ramallo-González A.P., Skarmeta A.F. Intrinsic and extrinsic quality of data for open data repositories. *ICT Express,* 2022, Vol. 8 (3), 328–333. https://doi.org/10.1016/j.icte.2022.06.001
6. Martinez-Gil J. Framework to automatically determine the quality of open data catalogs. *Expert Systems with Applications*, 2025, Vol. 289, Article 128379. https://doi.org/10.1016/j.eswa.2025.128379

7. Karamanou A., et al. Exploring the quality of dynamic open government data using statistical and machine learning methods. *Sensors*, 2022, Vol 22 (24), Article 9684. https://doi.org/10.3390/s22249684
8. Zong H., Wang Y. Impact mechanism of government public data openness and digital financial development on corporate information reliability. *International Review of Economics & Finance*, 2025, Vol. 103, Article 104495. https://doi.org/10.1016/j.iref.2025.104495
9. Gabay H., Reychav I., Jonathan G.M. Transformative shift: Digital evolution in Israeli public sector agencies during an emergency period. *Procedia Computer Science*, 2025, Vol. 256, 496–503. https://doi.org/10.1016/j.procs.2025.02.145
10. Jonathan G.M., Perjons E., Rusu L. Digital transformation-driven decentralisation of public governance. *Procedia Computer Science,* 2024, Vol. 239, 1220–1229. https://doi.org/10.1016/j.procs.2024.06.290
11. Jiao W., Zhang C., Wang Y. The impact of digital government construction on common prosperity – using data collected from 293 prefecture-level cities in China over the 2016–2022 period. *International Review of Economics & Finance*, 2025, Vol. 102, Article 104353. https://doi.org/10.1016/j.iref.2025.104353
12. Zhang Y., Li Y. The impact of artificial intelligence on government digital service capacity. *International Review of Economics & Finance,* 2025, Vol. 102, Article 104374. https://doi.org/10.1016/j.iref.2025.104374
13. Dani pro nadani administratyvni posluhy. (n.d.). Portal vidkrytykh danykh. URL: https://data.gov.ua/dataset/e1fd3119-3df6-43ce-8e5b-41bcc79dcb02 [Accessed 10 May. 2026]
14. Rahal M., et al. Enhancing machine learning performance through intelligent data quality assessment: An unsupervised data-centric framework. *Heliyon*, 2025, Vol. 11 (4), Article e42777. https://doi.org/10.1016/j.heliyon.2025.e42777
15. Pekar A., Jozsa R. Evaluating ML-based anomaly detection across datasets of varied integrity: A case study. *Computer Networks*, 2024, Vol. 251, Article 110617. https://doi.org/10.1016/j.comnet.2024.110617

Отримано/Received 15.07.2026
Прийнято/Accepted 12.09.2026
Опубліковано/Published 23.10.2026

S.O. MOTORNIUK, PhD Student,
Simon Kuznets Kharkiv National University of Economics,
9A, Nauky ave., Kharkiv, 61165, Ukraine
https://orcid.org/0009-0001-7695-5415
serhii891@proton.me

METHOD OF INTELLIGENT ASSESSMENT
OF THE QUALITY OF OPEN DATA WITH AUTOMATED
GENERATION OF RECOMMENDATIONS AND REPORTS

Introduction. The rapid growth of open data published by public authorities, research institutions and business organizations increases the need for reliable mechanisms of data quality assessment. Although open data are formally available for reuse, their practical analytical value is often limited by missing values, structural inconsistencies, format errors, semantic contradictions, duplicates and anomalous observations. These defects reduce the reliability of decision-making, complicate integration into digital platforms and weaken the effectiveness of intelligent services based on open datasets. Existing approaches to data quality assessment are often fragmented, static and mainly descriptive. They usually focus on isolated metrics and do not provide a complete automated cycle that includes defect detection, quality improvement recommendations and analytical reporting. Therefore, the development of intelligent methods capable of combining formal validation rules, statistical procedures and machine learning algorithms is an urgent scientific and practical task.

The purpose of the paper is to develop a method for intelligent assessment of open data quality with automated generation of recommendations and analytical reports. The proposed method is aimed at increasing the analytical usability and practical value of open datasets by integrating multidimensional quality metrics, adaptive weighting, anomaly detection, missing value imputation and interpretable reporting within a single methodological and software framework.

Methods. The study proposes a mathematical model for intelligent open data quality assessment. The model represents a dataset as a set of records and attributes and evaluates it through a vector of quality metrics, including completeness, consistency, correctness, uniqueness and semantic validity. Each metric is calculated using a combination of rule-based validation, statistical analysis and machine learning techniques. A key feature of the approach is the construction of an integral quality indicator based on adaptive metric weights. Unlike traditional models with fixed weights, the proposed method adjusts the importance of each metric according to the detected defects, variability of indicators and their influence on the analytical suitability of the dataset. The model also includes a defect function for identifying missing values, anomalies and format inconsistencies, as well as a recommendation function that transforms diagnostic results into corrective actions. To verify the proposed method, a modular software prototype was developed. It supports data loading, preprocessing, quality assessment, defect detection, recommendation generation and automated analytical reporting.

Results. Experimental validation confirmed the effectiveness of the proposed approach. After applying the developed method, the integral quality indicator of the open dataset increased from 0.9362 to 0.9477, which corresponds to an improvement of approximately 1.23% compared with the initial level. This result is significant because the original dataset already had a relatively high quality score, making further improvement difficult without intelligent processing. The most noticeable improvement was observed in the structural quality dimension, which increased from approximately 0.84 to 0.87 due to normalization, missing value processing and correction of format inconsistencies. Semantic, operational and behavioral dimensions remained close to one, indicating that the proposed method improved problematic components without degrading already correct data aspects. In anomaly detection, Isolation Forest showed the best result under the experimental conditions with injected anomalies, achieving an overall classification accuracy of 0.943. In the missing value imputation task, the Random Forest method achieved the lowest Root Mean Square Error (RMSE) value of 77.704, outperforming the k-Nearest Neighbors (KNN) method with RMSE 101.203 and median imputation with RMSE 125.079. Thus, Random Forest reduced the imputation error by 23.2% compared with KNN and by 37.9% compared with the median approach.

Conclusions. The proposed method enables a transition from fragmented and declarative data quality assessment to an integrated intelligent analysis of open datasets. Its scientific novelty lies in combining multidimensional quality assessment, adaptive weighting of metrics, formalized defect detection and automated recommendation generation. The developed software implementation confirms the practical applicability of the method and demonstrates that intelligent processing can provide measurable improvements even for datasets with initially high quality. The results prove the feasibility of using machine learning algorithms, especially ensemble models, for improving open data quality. The proposed approach can be used as a basis for decision support systems, open data portals and digital governance platforms where the reliability, completeness and analytical usability of data are critical.

Keywords: *open data, data quality, artificial intelligence, quality assessment, anomaly detection, data imputation, recommender systems, analytical reports, CKAN, machine learning.*