https://doi.org/10.31891/2307-5732-2025-355-6
УДК 637.5.02

БРУСЕНЦОВ ГЕОРГІЙ
Національний університет «Львівська Політехніка»
https://orcid.org/0000-0002-3346-0164
e-mail: heorhii.y.brusentsov@lpnu.ua

ПРЕПРОЦЕСИНГ ІЄРАРХІЧНИХ ДАНИХ З МЕДІА ПЛАТФОРМ НА БАЗІ СИСТЕМ НА ОСНОВІ ПРАВИЛ

Стаття присвячена препроцесингу ієрархічних даних з рекламних медіа-платформ з використанням систем на основі правил. Розглядаються особливості структури даних на прикладі рекламних кампаній платформи Meta. Досліджено основні етапи препроцесингу: очищення, трансформація, фільтрація та інтеграція даних, які необхідні для забезпечення високої якості аналітики та прийняття маркетингових рішень. Запропоновано методологію та практичні підходи використання правил для автоматизованої та стандартизованої обробки великих масивів ієрархічних даних.

Ключові слова: препроцесинг даних, ієрархічні дані, системи на основі правил, рекламна аналітика.

BRUSENTSOV HEORHII
Lviv Polytechnik National University

PREPROCESSING OF HIERARCHICAL DATA FROM MEDIA PLATFORMS USING RULE-BASED SYSTEMS

The article explores preprocessing of hierarchical data from digital advertising platforms using rule-based systems. With rapid growth in digital marketing, companies increasingly rely on structured data analysis from platforms like Meta for decision-making. Advertising data often have complex hierarchical structures organized into campaigns, ad sets, ads, and creatives, each level providing different detail granularity. Efficient preprocessing, including cleaning, transformation, filtering, and integration, is essential to ensure analytical accuracy and optimized budget allocation. Manual preprocessing or general-purpose scripts commonly used are labor-intensive and error-prone, leading to unreliable analytics. To address these challenges, the study proposes a rule-based preprocessing methodology, enabling automated, transparent, and scalable data handling aligned with marketing domain expertise. Rule-based systems use explicit logic ("if-then" rules) to process data systematically, effectively incorporating domain knowledge and reducing errors inherent in manual approaches. A dataset from the Meta platform with nearly 20,000 rows and 232 columns, including many missing values, is examined. Practical preprocessing approaches employed involve aggregating relevant features into logical flags or counts and encoding categorical data through one-hot or label encoding. Missing categorical values were labeled explicitly to distinguish genuine absence clearly from unintentional gaps. These methods targeted data related to geolocation, audience segmentation, creative automation, asset management, tracking pixels, and targeting relaxation. The developed approach significantly reduced dataset dimensionality and eliminated missing values, ensuring data quality suitable for advanced analytics such as clustering or performance evaluation. The rule-based framework proved effective, facilitating standardized, reliable preprocessing critical for robust advertising analysis.

Keywords: data preprocessing, hierarchical data, rule-based systems, advertising analytics.

Стаття надійшла до редакції / Received 01.05.2025
Прийнята до друку / Accepted 18.05.2025

Постановка проблеми та її актуальність

Ринок цифрової реклами стрімко зростає, і компанії все більше покладаються на дані з медіа платформ для ухвалення рішень у маркетингу. Сучасна рекламна індустрія фактично стала data-driven, адже понад 80% маркетологів базують свої рішення на даних, використовуючи аналітичні інструменти для оцінки рекламних кампаній на кожному етапі [1]. Дані з рекламних платформ (наприклад, соціальних мереж чи пошукових систем) мають складну ієрархічну структуру, яка відображає організацію рекламних кампаній. Ефективна обробка таких ієрархічних даних є актуальною задачею, оскільки від якості їхнього препроцесингу залежить точність аналітики та оптимізація рекламного бюджету.

Якісний препроцесинг даних передує будь-якому аналізу чи моделюванню і є критично важливим кроком. Відомо, що попереднє очищення та підготовка даних забирають значну частину часу у проєктах аналізу даних, але ці зусилля є необхідними. Як зазначають дослідники, препроцесинг – це «важливий етап у процесі отримання знань», що включає очищення, трансформацію, фільтрацію та інтеграцію даних. Якісно підготовлені вхідні дані безпосередньо впливають на результати подальшого аналізу: застосування належних методів препроцесингу може суттєво підвищити ефективність моделей та точність прогнозів. Натомість недоліки на цьому етапі призводять до принципу “сміття на вході – сміття на виході”, тобто помилки чи «шум» у сирих даних неминуче зумовлять хибні висновки моделі. Через це фахівці з аналізу даних наголошують на необхідності приділяти достатньо уваги очищенню та підготовці даних перед запуском аналітики. В рекламній індустрії ці аспекти особливо актуальні, адже неточні дані можуть призвести до фінансових втрат при неправильному налаштуванні кампаній чи невірній оцінці їх результатів. Проте на практиці препроцесинг рекламних даних часто здійснюється вручну або за допомогою скриптів загального призначення. Такий підхід є трудомістким і схильним до помилок [2], що мотивує пошук більш формалізованих і надійних методів попередньої обробки.

Ієрархічна структура даних медіа платформ ускладнює їх обробку. Рекламні платформи (на кшталт Meta Ads, Google Ads тощо) організовують кампанії багаторівнево. Наприклад, у системі Meta (Facebook), яка розглядається у даній роботі, розрізняють такі рівні ієрархії рекламних даних [3][4]:

- Рекламна кампанія – це набір рекламних оголошень, об'єднаних спільною маркетинговою метою (ціллю кампанії).
- Набір оголошень (ad set) – це група оголошень, що показуються визначеній цільовій аудиторії; усі рекламні креативи в межах одного набору орієнтовані на одну аудиторію.
- Оголошення (ad) – це конкретне рекламне оголошення, яке містить візуальні та текстові матеріали (креативи).
- Креатив – власне медіафайл або контент оголошення (зображення, відео, текст тощо), який зберігається у бібліотеці і може повторно використовуватися в різних оголошеннях.

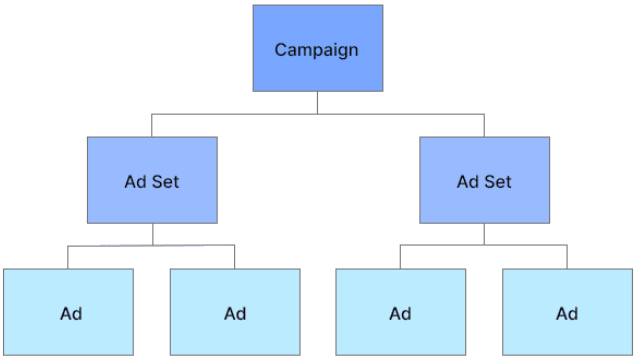

Рис. 1. Вигляд ієрархії рекламних даних на платформі Meta [4]

Такий деревоподібний формат даних створює ряд технічних і аналітичних викликів. По-перше, між рівнями існують залежності: зміна на верхньому рівні може автоматично впливати на нижчі. Наприклад, якщо зупинити або видалити рекламну кампанію, всі набори оголошень і оголошення нижче неї також будуть автоматично зупинені або видалені. Це вимагає від систем препроцесингу враховувати каскадне успадкування станів і налаштувань. По-друге, обсяги даних можуть бути дуже великими – один рекламний акаунт здатен містити до 100 000 кампаній, 100 000 наборів оголошень і 100 000 оголошень. Обробка настільки великих і вкладених структур потребує оптимізованих алгоритмів, здатних масштабуватися. По-третє, дані на різних рівнях потрібно з'єднувати та узгоджувати. Для цілісного аналізу часто необхідно об'єднувати інформацію з різних рівнів ієрархії (наприклад, зіставляти показники оголошення з параметрами кампанії, або додавати до оголошення дані про відповідний креатив). Це означає, що вкладені структури даних мають бути перетворені до плоского вигляду (таблиць або фреймів), з яким зручніше працювати аналітичним інструментам. Процес такої трансформації вкладених структур (flattening) є нетривіальним і вимагає значних зусиль та навичок обробки даних. Зокрема, фахівці відзначають, що ієрархічні (вкладені) дані створюють додаткову складність, потребують спеціальних навичок маніпулювання (наприклад, володіння мовами запитів чи спеціалізованими бібліотеками) і можуть впливати на точність моделей, якщо неправильно оброблені [5]. Вирішення цих складнощів на етапі препроцесингу є обов'язковим для отримання коректних і надійних результатів аналізу рекламних кампаній.

У зв'язку з наведеними викликами, дослідники і практики приділяють увагу правильній методології препроцесингу ієрархічних даних. Один із перспективних підходів – використання систем на основі правил (rule-based systems) для автоматизації і формалізації очищення даних. Системи, що керуються правилами, дозволяють явно задати логіку обробки у вигляді набору правил "якщо–то", що враховують структуру і семантику даних. Такий підхід добре узгоджується з потребами рекламної індустрії, де бізнес-логіка часто може бути формалізована як набір правил (наприклад, правила очищення або фільтрації: «відфільтрувати оголошення без показів», «об'єднати витрати всіх оголошень з однаковим креативом» тощо). Перевага систем на основі правил – у можливості інкорпорувати експертні знання про предметну область безпосередньо в процес препроцесингу. При цьому очищення даних значною мірою "зумовлене доменними знаннями, а не лише властивостями самих даних" [6]. Це означає, що фахівці з реклами можуть задати правила на основі свого розуміння кампаній (наприклад, виключити тестові кампанії, об'єднати дублікатні оголошення тощо), і система виконуватиме ці правила автоматично для всього масиву даних. Інтерпретованість та прозорість є додатковими плюсом: правила легко читати і перевіряти, що важливо при спілкуванні між аналітиками й маркетологами. Відповідно до досліджень, експерти галузі легше сприймають і описують кроки препроцесингу у вигляді правил, оскільки такий формат зрозумілий і піддається контролю [2]. На відміну від деяких машинних методів, система на основі правил забезпечує передбачуваність: однакові входи обробляються однаковими правилами, що гарантує узгодженість трансформацій. Це особливо важливо для багаторівневих структур, де послідовне застосування правил на кожному рівні допомагає зберегти зв'язність даних. Наприклад, система правил може забезпечити, що для кожної кампанії виконаються перевірки цілі, для кожного набору оголошень – перевірки коректності аудиторії, а для кожного оголошення – стандартизація форматів креативів.

Таким чином, якісний препроцесинг ієрархічних даних є необхідною умовою успішної аналітики в

рекламній сфері. Врахування багаторівневих залежностей, очищення та агрегування даних кампаній потребують продуманого підходу. Системи на основі правил зарекомендували себе як ефективний інструмент, що дозволяє формалізувати знання про дані та забезпечити їх стандартизовану обробку.

Метою роботи є: побудова системи на основі правил для препроцесингу ієрархічних даних з медіа платформ.

Виклад основного матеріалу

У даній роботі розглянуто датасет, який містить розгорнуті ієрархічні дані конфігурацій реклам, отриманих з рекламної платформи Meta. Як було зазначено вище, дані були організовані за чотирма основними рівнями ієрархії:

- Campaigns (кампанії)
- Ad Sets (групи оголошень)
- Ads (оголошення)
- Creatives (креативи)

На рівні кампаній представлені ключові характеристики, такі як ціль кампанії (objective), стратегія ставок (bid_strategy), бюджет (daily_budget, lifetime_budget), статус кампанії (status, effective_status), часові межі проведення (start_time, stop_time) та інші атрибути для високорівневого керування рекламою.

На рівні груп оголошень (adsets) дані включають деталізовані параметри таргетингу, серед яких географічні локації (країни, міста, райони, поштові індекси), демографічні характеристики (стать, вік, статус стосунків, освіта), платформи показу (Facebook, Instagram, Messenger, Audience Network), гнучкі специфікації таргетингу (інтереси, поведінка, місця роботи), бюджети та налаштування оптимізації та автоматизації реклами.

Оголошення (ads) пов'язують групи оголошень із конкретними креативами, надаючи інформацію про статус оголошення, пов’язані ідентифікатори, часові рамки, пікселі відстеження (fb_pixel) і взаємозв'язок із кампаніями.

Креативи (creatives) містять інформацію про безпосередні рекламні повідомлення, включаючи заголовки, тексти (body), URL-адреси, типи закликів до дії (call_to_action), мультимедійні файли (зображення, відео), а також параметри автоматизації та адаптації контенту під різні формати і розміщення. Особливу увагу приділено використанню автоматичних налаштувань (degrees_of_freedom_spec), що дозволяють адаптувати контент оголошення до різних місць розміщення, оптимізувати медіафайли (наприклад, анімації, відео-автокроп, автоматичні налаштування яскравості та контрасту) та покращувати тексти й описи з використанням алгоритмічних рішень.

Загальна кількість колонок склала 232 та розмірність датасету була майже 20 тисяч рядків. Частка відсутніх значень у 110 колонках складає більше 80%, що відповідає майже половині усіх колонок датасету. Кількість не пустих колонок становить лише 60. Зачасту у прикладах датасетів з інших домейних областей, такі колонки видаляються, але у цьому випадку, дані колонки можуть містити важливу інформацію для тих рядків, у яких дане значення не відсутнє. Ця інформація потенційно може впливати на якість та ефективність такої реклами.

Препроцесинг датасету, а саме транформація та агрегація, відбувалася використовуючи наступні підходи:

- **Створення колонки прапорця на основі вибраних колонок.** Даний підхід дозволяє агрегувати дані колонок у одну, яка буде позначати присутність інформації про ту чи іншу властивість.
- **Створення колонки лічильника на основі вибраних колонок.** Даний підхід дозволяє агрегувати дані колонок у одну, яка буде позначати кількість сконфігурованих параметрів поєднаних спільною характеристикою.

Дані підходи зачасту будуть застосовані до груп колонок, які містять текстову інформацію, що може не нести ніякої користі без застосування методів семантичного аналізу або засобів обробки природної мови.

Для даного датасету агрегація колонок застосовувалася для наступних груп:

- **Таргетування геолокації.** Дані про геолокацію таргетної аудиторії, такі як країни, міста, регіони і так далі.
- **Таргетування виключення.** Дані про аудиторію яку слід виключити з таргетування, щоб їй не показувалася дана реклама.
- **Таргетування аудиторії.** Дані про аудиторію, які включають стать, вік, вид діяльності, сімейний стан, тощо.
- **Автоматичні налаштування (degrees of freedom).** Дані конфігурацій безпосередньо для платформи Meta, які визначають що і наскільки може бути автоматично змінено у креативі. Наприклад, обрізання зображення для кращого його представлення на тому чи іншому пристрої, додавання анімацій, покращення текстів і описів реклами, тощо.
- **Стрічка ресурсів (asset feed).** Дані креативів які будуть поєднані між собою платформою Meta автоматично для створення комбінацій оголошень. Туди входять зображення, описи, заголовки, заклики до дій та інші частини креативу, які бачить користувач.
- **Піксель відстеження (fb pixel).** Інструмент, який допомагає платформі дізнатися, що користувачі роблять на сайті, який рекламується, після того як вони натиснули або переглянули рекламу. Це можуть бути

такі дані, як покупки товарів, додавання товарів у кошик, час перебування на сторінці, тощо.

• **Послаблення таргетування (targeting relaxation).** Механізм платформи, який трохи розширює аудиторію, якщо це може потенційно покращити ефективність реклами.

• **Конфігурації креативів.** До цієї групи відносяться колонки які бачить користувач у тому чи іншому місці розміщення реклами. Це такі значення, як заголовки, зображення, відео, описи, основні тексти реклами, заклики до дій, тощо.

Такі дані можуть бути корисними для кластерного аналізу визначення аудиторії за конкретним рекламним продуктом. Але для задач оцінювання ефективності реклами може бути важливим лиш факт конфігурації даних видів таргетування.

Наступним кроком є трансформування категоріальних даних за допомогою унітарного кодування (one-hot-encoding) або категоріальним кодуванням (label encoding). Вибір типу кодування було здійснено на основі значень певної колонки. Якщо обрана колонка має обмежений набір значень і кожний рядок містить лише одне значення, то у такому випадку категоріальне кодування повинне бути застосоване. Якщо обрана колонка має обмежений набір значень, але кожний рядок містить декілька значень з цього набору, то у такому випадку унітарне кодування повинне бути застосоване.

Наприклад, для колонки, яка відповідає за рівні фільтрації контенту безпеки бренду (brand_safety_content_filter_levels) на рівні групи оголошень, було застосовано унітарне кодування, так як одна реклама може мати більше одного рівня фільтрації. Можливі значення у такому випадку можуть бути FACEBOOK_STANDARD та AN_STANDARD, що означають фільтрацію чутливого вмісту для елементів застосунків facebook (стрічка новин, історії в instragram, тощо) та для мережі аудиторії (вебсайти та застосунки, які використовують Meta Ads).

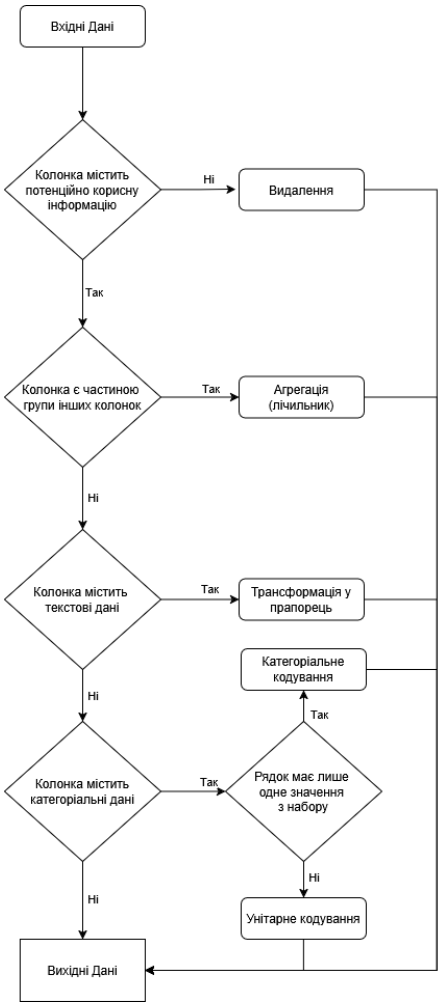

Рис. 2. Діаграма системи на основі правил для препроцесингу даних з медіа платформ

Також для категоріальних колонок пусті значення були заповнені значеннями “disabled”, “not set” “none”, у випадках коли такі від’ємні значення були присутні у наборі. У таких випадках відсутність значення співпадає з від’ємним. Для прикладу візьмемо колонку “тип темпу” (pacing type) кампанії, яка визначає наскільки швидко чи повільно витрачається бюджет протягом усього терміну дії кампанії. Одне з можливих значень є “disabled” (виключене), яке означає що платформа Meta не контролює витрати бюджету автоматично.

Як результат розмірність датасету було зменшено до 59 колонок, а кількість відсутніх значень становить 0% серед всіх колонок. Детальна діаграма системи на основі правил та кореляційна матриця вихідного датасету зображені на рисунках 2 та 3 відповідно.

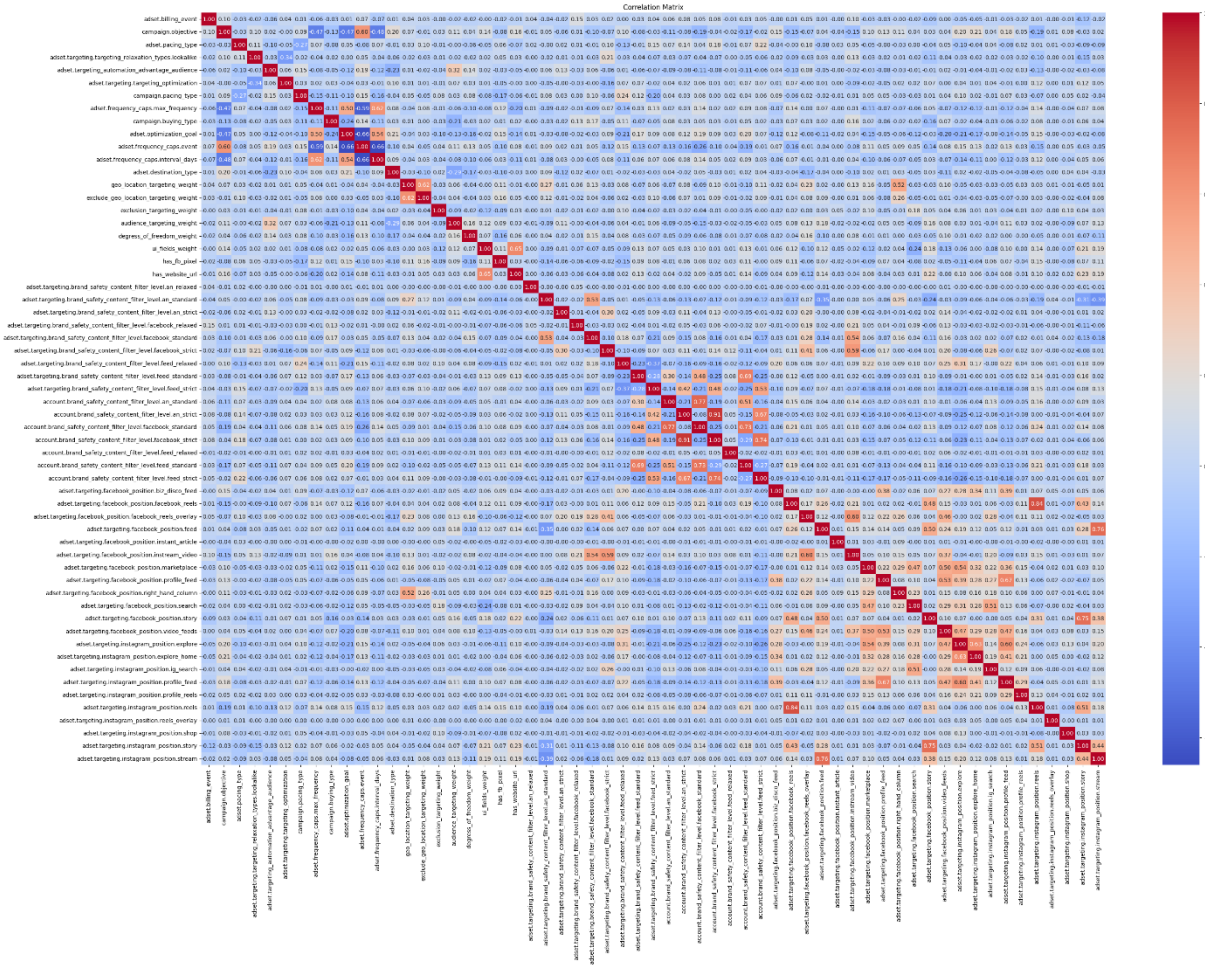

Рис. 3. Кореляційна матриця вихідного датасету

Висновки

На основі даної роботи можна зробити висновки, що використання систем на основі правил є доцільним у препроцесингу даних з медіа платформ. У результаті розмірність розглянутого датасету було зменшено на майже 75% та зменшена частка відсутніх значень до 0%, що спрощує подальшу роботу з даним датасетом для використання будь-яких моделей машинного навчання.

Також варто зазначити, що такий підхід легко застосовуватиметься у вузькопрофільних галузях, так як до побудови системи на основі правил можна залучати людей близьких до обраної галузі.

Література

1. The Importance of Data Analytics in Digital Marketing [Електронний ресурс] // William & Mary Online Business Blog. – 07 березня 2021. – Режим доступу: https://online.mason.wm.edu/blog/data-analytics-in-digital-marketing (дата звернення: 22.04.2025). – Назва з екрана.

2. Ramírez A., Moreno N., Vallecillo A. Rule-based preprocessing for data stream mining using complex event processing // Expert Systems. — 2021. — Vol. 38, e12762.

3. Хейна М. Як створити і запустити рекламу на Facebook у 2022 році? [Електронний ресурс] // Блог BannerBoo. — 27 вересня 2022. — Режим доступу: https://bannerboo.com/ua/blog/yak-stvoriti-i-zapustiti-reklamu-na-facebook-u-2022-rotsi/ (дата звернення: 25.04.2025).

4. Meta Platforms, Inc. Marketing API: Overview [Електронний ресурс] // Meta for Developers. — 2021. — Режим доступу: https://developers.facebook.com/docs/marketing-api/overview (дата звернення: 25.04.2025). — Назва з екрана.

5. Pecan Team. Data Flattening 101: Preparing Your Data for Predictive Analytics Success [Електронний ресурс] // Pecan AI Blog. — 5 червня 2024. — Режим доступу: https://www.pecan.ai/blog/data-flatten-analytics-101/ (дата звернення: 28.04.2025).

6. Bradji L., Boufaida M. A Rule Management System for Knowledge Based Data Cleaning // Intelligent Information Management. — 2011. — Vol. 3, No. 6. — P. 230–239.

References

1. The Importance of Data Analytics in Digital Marketing [Elektronnyi resurs] // William & Mary Online Business Blog. – 07 bereznia 2021. – Rezhym dostupu: https://online.mason.wm.edu/blog/data-analytics-in-digital-marketing (data zvernennia: 22.04.2025). – Nazva z ekrana.

2. Ramírez A., Moreno N., Vallecillo A. Rule-based preprocessing for data stream mining using complex event processing // Expert Systems. — 2021. — Vol. 38, e12762.

3. Kheina M. Yak stvoryty i zapustyty reklamu na Facebook u 2022 rotsi? [Elektronnyi resurs] // Bloh BannerBoo. — 27 veresnia 2022. — Rezhym dostupu: https://bannerboo.com/ua/blog/yak-stvoriti-i-zapustiti-reklamu-na-facebook-u-2022-rotsi/ (data zvernennia: 25.04.2025).

4. Meta Platforms, Inc. Marketing API: Overview [Elektronnyi resurs] // Meta for Developers. — 2021. — Rezhym dostupu: https://developers.facebook.com/docs/marketing-api/overview (data zvernennia: 25.04.2025). — Nazva z ekrana.

5. Pecan Team. Data Flattening 101: Preparing Your Data for Predictive Analytics Success [Elektronnyi resurs] // Pecan AI Blog. — 5 chervnia 2024. — Rezhym dostupu: https://www.pecan.ai/blog/data-flatten-analytics-101/ (data zvernennia: 28.04.2025).

6. Bradji L., Boufaida M. A Rule Management System for Knowledge Based Data Cleaning // Intelligent Information Management. — 2011. — Vol. 3, No. 6. — P. 230–239.