DOI 10.34185/1562-9945-4-159-2025-08
УДК 004.89:004

Л.А. Люшенко, О.С. Монько

ПОБУДОВА ЗАДАЧЕОРІЄНТОВАНОЇ БАЗИ ДАНИХ ДЛЯ ПРОГНОЗУВАННЯ ДИНАМІКИ СПОЖИВАННЯ ЕЛЕКТРИЧНОЇ ЕНЕРГІЇ

Анотація. У зв'язку зі стрімким розвитком технологій штучного інтелекту з'явився широкий перелік нових підходів, які значно розширили можливості сучасних програмних експертних систем, що здатні відігравати критичну роль у науково-технічному розвитку різних галузей — від медицини до промисловості. Основою функціонування програмних експертних систем є наявність релевантних вхідних даних. Програмні експертні системи повинні спиратися на якісно опрацьовану інформацію, отриману шляхом перетворення вимірювань та необроблених даних у структуровані бази даних. Дана стаття присвячена процесу створення бази даних для програмних експертних систем. Досліджуються сучасні програмні рішення, спрямовані на перетворення необроблених даних у структурований вигляд, необхідний для високої точності роботи прогнозних обчислювальних модулів. У статті представлено алгоритмічний підхід до побудови бази даних для програмних експертних систем, спрямованих на прогнозування динаміки споживання електроенергії.
Ключові слова: задачеорієнтовані бази даних, датасети, програмні експертні системи, попередня обробка даних, трансформація даних, прогнозування електроспоживання, кореляційний аналіз, періодичні змінні.

Постановка проблеми. Протягом останнього десятиліття спостерігається різкий приріст цифрової інформації, який спричинений впровадженням різноманітних систем автоматичного збору даних та моніторингу. Окрім того, швидкі темпи розвитку технологій штучного інтелекту дозволяють здійснювати розробку сучасних програмних експертних систем. Результати прогнозів надають можливість приймати обґрунтовані управлінські рішення з мінімізацію майбутніх ризиків. Однією з проблем створення та функціонування таких програмних експертних систем є наявність якісних вхідних даних. Тому актуальним є дослідження підходів до створення автоматичної програмної обробки даних та побудови баз даних експертних систем. Точність обчислень та прогнозів забезпечується: розробкою програмних модулів із використанням спеціалізованих алгоритмів для усунення шуму, помилкових даних, аналітичним доповненням даних, вирішення проблем їх неконсистентності. Отримані у результаті обробки дані використовується для формування баз даних, що стануть основою для побудови баз знань експертних систем.

Мета дослідження полягає в розробці підходу до побудови задачеорієнтованої бази даних програмної експертної системи прогнозування динаміки споживання електричної енергії. Дослідження та експеримент проведено на прикладі реального датасету споживання електроенергії.

Аналіз останніх досліджень і публікацій. Якість даних має дуже вагомий вплив на точність прогнозів у інтелектуальних системах, які використовують методи машинного навчання. Ефективним є застосування технік: стандартизації (нормалізації); пов'язування записів, які стосуються одного й того ж об'єкта; локалізації та виправлення помилок якості даних, які не відповідають заданому набору правил якості [1].

У роботі А. Чжена та А. Казарі "Інженерія ознак для машинного навчання" наведено розлогий огляд підходів до створення похідних ознак, серед яких перетворення та масштабування ознак, інженерія взаємодій [2]. Перспективним напрямом є розробка методів автоматичної генерації та відбору релевантних ознак (наразі відбір ознак часто має ітеративний характер), адже остаточний перелік ознак часто визначається лише після множинних експериментів та оцінок впливу конкретних ознак на цільову метрику [3].

Класичні методології з побудови сховищ даних (data warehouses) описано у роботах Кімбелла та Росса [4], які запровадили концепцію "data marts" для спеціалізованого використання у межах бізнес-аналітики. Формування спеціалізованих сховищ даних дають змогу адаптувати вихідні схеми зберігання та набори ознак під конкретні завдання чи галузі застосування. Ідея створення задачеорієнтованих баз даних у ширшому контексті викладена в доменно-орієнтованих підходах, які підкреслюють важливість врахування конкретних вимог до змінних та особливостей предметної області [5]. Підкреслюється важливість інтеграції правил і знань із різних джерел та необхідність періодичного оновлення схеми зберігання при зміні умов середовища чи появі нових завдань.

Проте питання автоматизації вибору структури зберігання все ще залишається відкритим. Класичні праці з організації баз даних [6] і сучасні дослідження в галузі NoSQL чи графових систем [7] розглядають принципи оптимізації моделювання сутностей, проте переважно потребують ручного конструювання схеми або попереднього аналізу експертом. Деякі автори пропонують інтегрувати інструменти моніторингу та оцінки структури даних у ETL/ELT, що дає змогу напівавтоматично підбирати формати зберігання, але зазначається, що такі системи все ще залишаються недостатньо гнучкими для повної адаптації під змінні вимоги реального використання [8].

Існуючі роботи демонструють, що ідеальна "універсальна" база даних у реальних застосуваннях часто призводить до надмірності й накопичення шуму [5,9], що підтверджує необхідність ретельного відбору релевантних даних. Залишається відкритим питання пошуку універсальних принципів і формальних критеріїв, які б гарантували достатній рівень автоматизації й цілісності даних без втрати найважливіших патернів. Недосконалості наявних підходів, зокрема обмежена інтеграція з неструктурованими

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

джерелами, залежність від багаторазового ручного втручання, підтверджують актуальність проведення подальших досліджень у цьому напрямку.

Виклад основного матеріалу. Розробка програмних експертних систем вимагає перетворення вихідних даних, отриманих із різних джерел, у структуровану форму, що підходить для подальшого аналізу, використання методами машинного навчання та забезпечення швидкого доступу до необхідної інформації. Необроблені набори даних зазвичай мають численні недоліки, а саме: розбіжності за форматом і масштабом, дублювання записів, відсутні чи суперечливі значення. Якість використовуваних даних має безпосередній впливав на точність результатів прогнозування за допомогою машинного навчання [8].

Розглянемо запропонований процес побудови задачеорієнтованих баз даних, здійснений на прикладі реального датасету споживання електроенергії (рис. 1):

Рисунок 1 - Етапи процесу побудови бази даних

На першому етапі проводиться аналіз наявних неструктурованих первинних даних, зібраних з пристроїв моніторингу, автоматизованих систем збору чи інших джерел. Його мета — отримати загальні уявлення про розподіл даних, визначення їх формату та виявлення ймовірних аномалій. У рамках цього етапу слід провести статистичну оцінку наявних змінних, яка допомагає встановити числові межі для типових значень і виявити серйозні відхилення. З метою виявлення нетипових, хибних значень слід провести кластеризаційний аналіз за допомогою алгоритму k-середніх. Це дає можливість високорівнево оцінити масштаб та характер проблем, які планується вирішити в подальшому.

На другому етапі виконується початкова обробка даних, що охоплює очищення від явно некоректних записів, раніше виявлених аномалій, пропусків та дублікатів. На

цьому етапі необхідно обрати метод обробки пропусків. Їх можна або видалити, або заповнити середнім значенням чи медіаною, сусідніми значеннями, або скористатися іншими більш комплексними підходами. Для виконання даних операцій добре підходять бібліотеки Pandas або NumPy, які надають потужні інструменти для статистичного аналізу й виправлення виявлених недоліків у даних. В результаті цього етапу формується проміжний датасет, який відповідає мінімальним стандартам якості. Під час видалення нерелевантних даних важливо зберігати проміжні версії даних у файловій системі, що дозволить за необхідності повернутися до попередніх кроків обробки.

На третьому етапі здійснюється трансформація даних, а саме приведення їх до єдиного масштабу та формату, необхідного для застосування методів машинного навчання. Більшість алгоритмів машинного навчання не можуть коректно працювати з категоріальними даними, тому для їх трансформації використовуються методи кодування, які дозволяють перетворювати категоріальні змінні на числові. Метод простого присвоєння числових міток кожній категорії може призвести до виникнення штучної ієрархії змінної, тому пропонується застосування техніки фіктивного кодування [10]. Наступним кроком йде забезпечення рівномірного масштабування числових ознак, що необхідно для зведення різнорідних даних до єдиного виду, який створюватиме умови для їх порівняння. Цей процес необхідний для коректної роботи алгоритмів класифікації та кластеризації, а також для деяких методів машинного навчання. Стандартизація як процес масштабування даних не обмежує значення показників суворо заданими інтервалами, а змінює розподіл на основі статистичних ознак. Коли розподіл числових ознак має нерівномірний характер, застосування стандартного відхилення як міри масштабування зменшує вплив екстремальних значень на подальші обчислення [9].

На четвертому етапі доповнюємо датасет новими ознаками за допомогою генерування похідних змінних, які надають додаткову інформацію для розв’язання поставленої задачі й сприяють підвищенню точності прогнозів. Під час роботи з часовими рядами зазвичай результативним є опрацювання часової компоненти: виокремлення дня тижня, місяця, розпізнавання періодичності. На основі інформації про час сходу та заходу сонця можна виділити окрему змінну тривалості світлового дня, яка в свою чергу може мати зв’язок із рівнем електроспоживання, або активністю користувачів енергосистеми. Окрім того, виявляються періодичні змінні у випадках наявності відповідних періодичних закономірностей. Введення таких змінних, отриманих за допомогою функцій перетворень синуса та косинуса, надає можливість математично представити повторювані періоди. Звичайне числове перетворення не враховує природну неперервність між кінцем і початком періодів. Використання таких перетворень дає змогу інтегрувати сезонність без необхідності розділяти дані на окремі періоди [11]. Слід взяти до уваги, що для кількох періодичних змінних може виникнути мультиколінеарність, тому введення таких компонент слід здійснювати з обережністю.

На п’ятому етапі проводиться відбір найбільш релевантних змінних та метрик. Цей етап передбачає аналіз взаємозв’язків між змінними. Оцінюється сила відносин

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

між ними, включаючи і цільову змінну. На відміну від коефіцієнтів кореляції Пірсона, за допомогою коефіцієнтів рангової кореляції Спірмена можна виявляти лінійні та нелінійні зв'язки, за умови, що вони мають монотонний характер. При інтерпретації коефіцієнтів слід звертати увагу на ситуації, коли вони наближаються до крайніх значень (+1 або –1), адже такі результати вказують про наявність сильного зв'язку між незалежними змінними – мультиколінеарності. Аналіз матриці кореляції Спірмена забезпечує можливість здійснити попередній відбір змінних та знизити надмірну залежність між ними шляхом вилучення однієї з змінних кожної мультиколінеарної пари [12]. Слід підкреслити, що остаточний перелік релевантних змінних встановлюється після реалізації численних експериментів із застосуванням конкретних методів машинного навчання.

На шостому етапі потрібно обрати схему зберігання інформації. Через те, що структура даних може сильно змінюватися у процесі попередньої обробки, доцільно проводити цей етап після завершення всіх попередніх перетворень. Необхідно брати до уваги використовувані інструменти й технології розроблюваної системи, вимоги до продуктивності, обсяг інформації, передбачуваний темп зростання даних, специфіка доступу (транзакційний або аналітичний) та характер взаємозв'язків між сутностями. Реляційні моделі застосовують у системах, де важливою є надійність і швидкість транзакцій, де необхідно зберігати дані з високим рівнем структурованості та з чітко визначеними зв'язками між сутностями. Проте через динамічне оновлення даних чи необхідність працювати з великими обсягами напівструктурованих даних, або складними ієрархічними зв'язками, доцільним може бути використання NoSQL-рішень (документоорієнтованих, графових, колонкових). В окремих випадках виправданим може бути застосування змішаних рішень (реляційна схема для транзакційних операцій і NoSQL-сховище для швидкого пошуку напівструктурованих даних).

Сьомий етап відповідає за процес ітеративного формування задачеорієнтованих баз даних. Експертна система відповідно до поставлених задач може потребувати специфічних змінних, інакшої глибини історичних даних. Універсальна база даних, яка включає всі можливі дані є надто громіздкою і неефективною для використання під конкретну практичну задачу. Тому після виконання всіх попередніх етапів обробки й перетворень, під кожну нову задачу може бути сформовано власну логічно відокремлену базу даних, що реалізує концепцію задачеорієнтованих баз даних, згідно з якою кожна задача отримує власний необхідний "зріз" даних.

На основі описаних вище етапів та відкритих джерел було створено базу даних про споживання електроенергії у Лондоні в період з листопада 2011 року по лютий 2014 року. Початковий датасет був сформований у рамках програми "UK Power Networks led Low Carbon London project" і містив інформацію про 5 567 домогосподарств – загалом у розрізі добового електроспоживання 3 510 433 записів даних. За допомогою OpenWeather API датасет було доповнено інформацією про погодні умови, а також даними про державні свята.

У результаті аналізу даних було виявлено, що датасет містить нерелевантні значення електроспоживання за останній збережений день кожного домогосподарства

відповідно. Більше того, було відмічено, що з травня 2012 року спостерігалося активне залучення домогосподарств до проекту, причому більшість з них брала участь у ньому понад півроку. Для того, щоб охопити більшу кількість домогосподарств, у даному випадку доречно використовувати середнє споживання електроенергії на добу, а не загальний добовий обсяг.

У цьому датасеті містилися зайві дані, що стосуються погодних умов, зокрема текстовий опис погоди, години досягнення максимальних та мінімальних температур і подібні відомості, тож їх слід вилучити. Пропущені дані було заповнено даними наступного дня за допомогою бібліотеки Pandas, що є доречним через роботу з часовим рядом.

На етапі трансформації даних було застосовано техніку фіктивного кодування для категоріальної змінної “precipType”, яку було перетворено на набір двійкових змінних: “rain” та “snow”. Також до датасету було застосовано техніку стандартизації для можливості застосування алгоритму градієнтного спуску в рамках алгоритму логістичної регресії.

Датасет було доповнено окремою змінною тривалості світлового дня на основі інформації про час сходу та заходу сонця. Аналогічно дані містили інформацію про температурні екстремуми протягом доби, на основі чого було обчислено температурний діапазон, що відображає інтенсивність змін температурних умов.

На рис. 2 за допомогою функції автокореляції продемонстровано наявність тижневої періодичності в середньодобовому споживанні електроенергії, тому впровадження періодичних змінних є обгрунтованим.

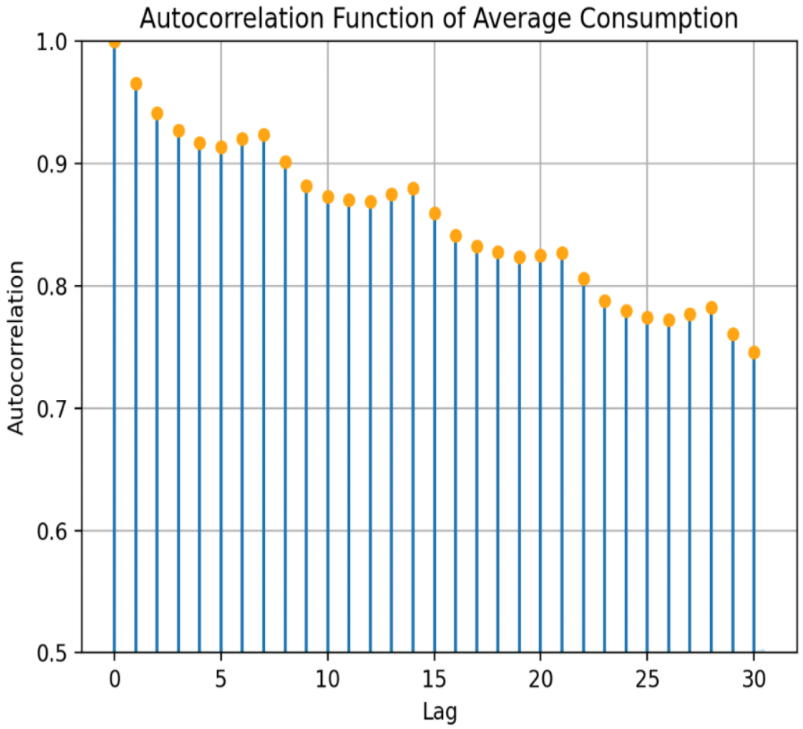

Рисунок 2 - Функція автокореляції середньодобового електроспоживання

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

У результаті проведених експериментів було виявлено, що введені періодичні змінні “monthSin” та “monthCos” суттєво підвищували точність прогнозування рівня споживання електроенергії.

У результаті проведеного кореляційного аналізу за допомогою матриці кореляції рангу Спірмена (рис. 3) було вилучено 12 найменш релевантних погодних метрик.

За основу бази даних рівня споживання електричної енергії доцільно було взяти реляційну модель через невелику кількість сутностей і зв’язків після всіх попередніх перетворень та виділення найбільш релевантних змінних.

За допомогою вищеописаного алгоритму було побудовано базу даних рівня для прогнозування рівня споживання електроенергії за допомогою моделей множинної лінійної регресії та випадкового лісу. На рис. 4 зображено порівняльний графік прогнозів двох моделей [13].

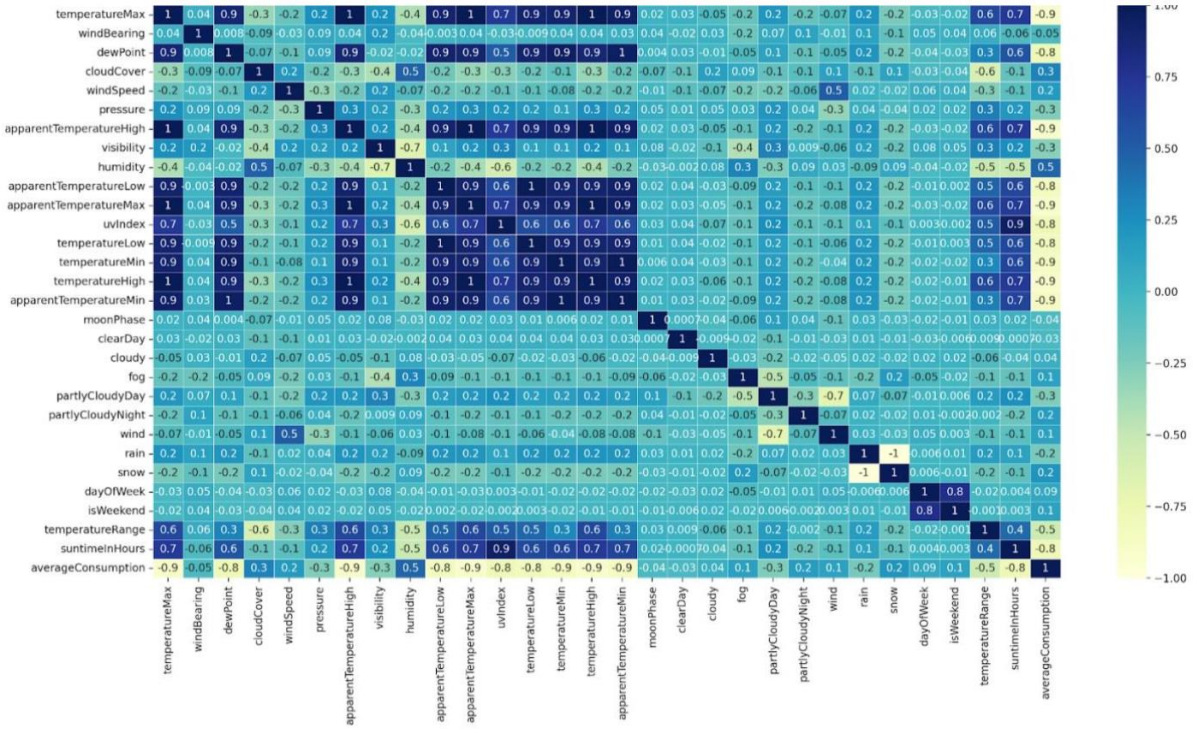

Рисунок 3 - Матриця кореляції рангу Спірмена

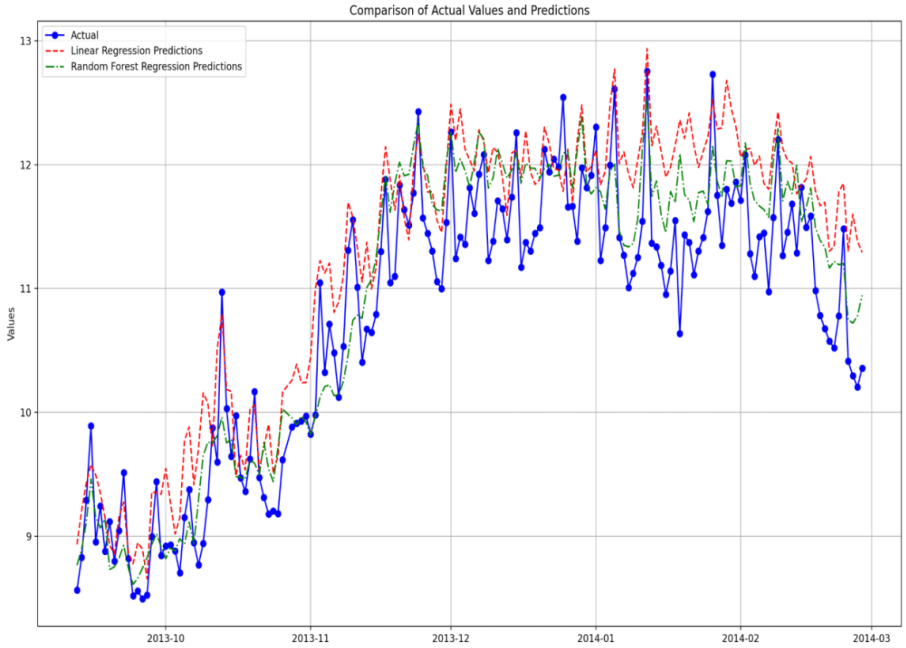

Рисунок 4 - Порівняльний графік прогнозів множинної лінійної регресії та випадкового лісу

Отже, формування задачеорієнтованих баз даних має на меті врахування індивідуальних вимог до точності, швидкості обробки та гнучкості. На початку було виконано масштабний відбір змінних і сформовано велику універсальну структуру, а вже під конкретну задачу було проведено повторний відбір змінних. У випадку погіршення результатів чи появи нових вимог до прогнозування цикл може початися спочатку. Завдяки ітеративному підходу система має потенціал до постійного вдосконалення, адаптуючи набір змінних до специфічних потреб конкретних модулів машинного навчання чи аналітики. Така динамічність дає змогу експертній системі швидко реагувати на структурні зрушення у предметній галузі, зміни у способах збору або валідації даних, варіативні сценарії використання. У підсумку, описаний підхід сприяє впорядкуванню процесу побудови баз даних для інформаційних систем, що працюють з задачами прогнозування.

Висновки та перспективи подальших досліджень. У рамках дослідження було запропоновано алгоритм побудови бази даних для експертних систем прогнозування споживання електроенергії, який ґрунтується на послідовному застосуванні методів попереднього аналізу, очищення та обробки, трансформації та відбору даних, подальшим вибором схеми їх зберігання. За допомогою такого підходу початковий різнорідний набір даних було перетворено в структурований формат, необхідний для отримання необхідних результатів точності прогнозування. Зокрема дані було використано для прогнозування за допомогою моделей множинної лінійної регресії, логістичної регресії та випадкового лісу. Ітеративне доопрацювання задачеєорієнтованих баз даних дає можливість перебудовувати структуру та набір змінних у відповідь на нові умови чи завдання.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Дослідження показало, що датасет енергоспоживання містив ознаки тижневої періодичності і подальше впровадження періодичних змінних, отриманих за допомогою функцій перетворень синуса та косинуса, дало позитивний ефекти на точність прогнозів. Аналіз кореляційних залежностей із застосуванням коефіцієнтів рангової кореляції Спірмена допоміг виявити та усунути нерелевантні метрики, зокрема 12 метрик погодних умов.

Проведене дослідження відкриває перспективи для подальшого вдосконалення алгоритмічних рішень у галузі розробки програмних експертних систем, які використовують методи машинного навчання та знаходять практичне застосування в процесах прийняття управлінських рішень у галузі електроенергетики. Беручи до уваги потребу в адаптації до зростаючих обсягів даних і їх різноманітності, запропонований алгоритм слугуватиме фундаментом для подальших досліджень у напрямку автоматизованого проектування задачеорієнтованих баз знань.

ЛІТЕРАТУРА/REFERENCES

1. Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. ACM Computing Surveys, 41(3), Article 16. DOI: 10.1145/1541880.1541883.

2. Zheng, A., & Casari, A. (2018). Feature engineering for machine learning: Principles and techniques for data scientists. O'Reilly Media.

3. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). An introduction to statistical learning: With applications in R (2nd ed.). Springer.

4. Kimball, R., & Ross, M. (2013). The data warehouse toolkit: The definitive guide to dimensional modeling (3rd ed.). Wiley.

5. Cao, L. (2010). Domain-driven data mining: Challenges and prospects. IEEE Transactions on Knowledge and Data Engineering, 22(6), 755–769.

6. Elmasri, R., & Navathe, S. B. (2016). Fundamentals of database systems (7th ed.). Pearson.

7. Crowe, M. K., & Laux, F. (2023). Graph data models and relational database technology. arXiv Preprint. https://arxiv.org/abs/2303.12376

8. Wang, R., Weng, J. та Huang, Z. (2018). A Comprehensive Survey on Data Preprocessing in Big Data Analytics. Journal of Big Data, 5(1), 22. DOI:10.1186/s40537-018-0139-3.

9. García, S., Fernández, A., Luengo, J., & Herrera, F. (2015). Data preprocessing in data mining. Springer.

10. Johannemann, J., Hadad, V., Athey, S., & Wager, S. (2019). Sufficient representations for categorical variables. arXiv Preprint. https://arxiv.org/abs/1908.09874

11. Kitagawa, G. (2024). A Triginometric Seasonal Component Model and its Application to Time Series with Two Types of Seasonality.

12. С.Ю. Гавриленко, В.О. Полторацький. Метод підвищення оперативності класифікації даних за рахунок зменшення кореляції ознак. Системи управління, навігації та зв'язку. Збірник наукових праць, 2023.

13. Л.А. Люшенко, О.С. Монько, В.І. Сущук-Слюсаренко. Програмне рішення для прогнозування динаміки споживання електричної енергії. Наука і техніка сьогодні, №5(33), 2024. DOI:10.52058/2786-6025-2024-5(33)-1258-1268.

Received 05.05.2025.
Accepted 07.05.2025.

Construction of a task-oriented database for electricity consumption dynamics forecasting

Modern software expert systems have the potential to optimize decision-making processes in a wide array of industries, including energy, healthcare, and manufacturing. Nevertheless, many machine learning–oriented approaches underestimate the significance of a robust database architecture specifically tailored to each domain. This study aims to develop and validate an algorithmic approach to building a task-oriented (specialized) database structure for an expert system dedicated to forecasting electric power consumption. Real consumption measurements are paired with contextual data such as weather patterns, then transformed through systematic validation, cleaning, and encoding. Outlier detection and imputation steps further reduce distortions stemming from incomplete or anomalous inputs, while an iterative feature selection process refines the pool of variables most relevant to predictive models.

Unlike generic pipelines, the proposed approach incorporates a schema evolution stage, ensuring that newly identified indicators can be integrated without undermining existing data integrity. Experimental application on a practical dataset combining technical power usage and external factors demonstrates improved resilience of forecasting results. By outlining each stage of data preparation and emphasizing schema adaptability, the proposed method supports high-precision predictions even under dynamically changing conditions. By continually revisiting and adjusting the database, the system can adapt to evolving operational conditions, thus enhancing the accuracy of computational modules responsible for projecting consumption patterns. The findings underscore the importance of a specialized, domain-specific data foundation in realizing the full potential of expert systems.

Keywords: task-oriented databases, datasets, software expert systems, data preprocessing, data transformation, electricity consumption forecasting, correlation analysis, periodic variables.

Люшенко Леся Анатоліївна – к.т.н., доцент, доцент кафедри програмного забезпечення комп'ютерних систем Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського», ORCID: 0000-0003-4319-5955
Монько Олександр Сергійович – аспірант кафедри програмного забезпечення комп'ютерних систем Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського», ORCID: 0009-0005-7110-1540

Liushenko Lesia – candidate of technical sciences, associate professor, associate professor of department of computer systems software at National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", ORCID: 0000-0003-4319-5955
Monko Oleksandr – graduate student at National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", ORCID: 0009-0005-7110-1540

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)