НАЦІОНАЛЬНИЙ УНІВЕРСТИТЕТ БІОРЕСУРСІВ

І ПРИРОДОКОРИСТУВАННЯ УКРАЇНИ

Факультет інформаційних технологій

УДК 004.9:613.14/.15-047.36

«ПОГОДЖЕНО»	**«ДОПУСКАЄТЬСЯ ДО ЗАХИСТУ»**
Декан факультету інформаційних технологій	Завідувач кафедри комп'ютерних наук
Болбот І. М., д.п.н., професор	Голуб Б.Л., к.т.н., доцент
_____	_____
____ _____ 2024 р.	**____ _____ 2024 р.**

МАГІСТЕРСЬКА КВАЛІФІКАЦІЙНА РОБОТА

на тему Інтелектуальна система моніторингу параметрів атмосферного повітря

Спеціальність _____121 «Інжереія програмного забезпечення»_____
(код і назва)

Освітня програма _____Програмне забезпечення інформаційних систем___
(назва)

Орієнтація освітньої програми ___освітня-професійна_____
(освітньо-професійна або освітньо-наукова)

Гарант освітньої програми

_____професор, д.т.н. _____ _Семко В. В.____
(науковий ступінь та вчене звання) (підпис) (ПІБ)

Керівник магістерської кваліфікаційної роботи

_____доцент к.т.н._____ _____ ___Голуб Б. Л._____
(науковий ступінь та вчене звання) (підпис) (ПІБ)

Виконав _____ _Москаленко Д. Ю._____
(підпис) (ПІБ студента)

КИЇВ-2024

Зміст

ВСТУП

Актуальність. Забруднення атмосферного повітря є однією з найбільш актуальних екологічних проблем сучасності, яка суттєво впливає на здоров'я населення та стан навколишнього середовища. Інтенсивна індустріалізація, розвиток транспорту, збільшення обсягів викидів шкідливих речовин в атмосферу, а також глобальні кліматичні зміни призводять до погіршення якості повітря, особливо у великих містах та промислових регіонах. Забрудненість повітря стає причиною різноманітних захворювань, зокрема дихальних та серцево-судинних, а також знижує тривалість життя населення. Вплив забрудненого зовнішнього повітря дрібними частинками є другим за величиною фактором ризику передчасної смерті в усьому світі після високого кров'яного тиску та найбільшим фактором екологічного ризику для передчасної смерті. Наприклад, сьогодні люди, які живуть у країнах з низьким і середнім рівнем доходу, піддаються впливу забруднення повітря в один-чотири рази частіше, ніж люди, які живуть у країнах з високим рівнем доходу. Діти особливо вразливі до шкідливого впливу забруднення повітря на здоров'я через їхню унікальну чутливість і вплив. 26 відсотків смертей новонароджених у всьому світі пояснюються впливом забрудненого повітря [1].

Найбільш смертоносними захворюваннями, пов'язаними із забрудненням повітря PM2.5, є інсульт, хвороби серця, легенів, захворювання нижніх дихальних шляхів (такі як пневмонія) і рак. Високий рівень дрібних часток також сприяє розвитку інших захворювань, наприклад діабету, може перешкоджати когнітивному розвитку дітей, а також викликати проблеми з психічним здоров'ям.

В умовах зростаючого масштабу цієї проблеми виникає необхідність у впровадженні ефективних інструментів для моніторингу, аналізу та прогнозування якості повітря. Це дозволяє оперативно виявляти критичні зміни та приймати відповідні рішення. Сучасні інформаційні технології надають можливості для створення систем, що здатні обробляти великі обсяги даних у

реальному часі, виконувати багатовимірний аналіз і забезпечувати точність прогнозування на основі виявлених закономірностей.

Зокрема, технології OLAP (Online Analytical Processing) дозволяють виконувати багатовимірний аналіз даних, забезпечуючи гнучке управління інформацією у розрізі часу, географії та інших параметрів. Методи машинного навчання, такі як кластеризація та прогнозування, відкривають можливості для виявлення складних патернів у даних і підвищення точності моделей прогнозування.

Об’єкт і предмет дослідження. Об’єкт дослідження атмосферне повітря та його показники в Україні. Предметом дослідження інтелектуальна система моніторингу параметрів атмосферного повітря

Мета дослідження. Метою дослідження є надання точних результатів моніторингу параметрів якості атмосферного повітря на основі технологій OLAP, Data Mining та Machine Learning.

Завдання дослідження. Для досягнення мети дослідження необхідно вирішити наступні завдання:

1. Провести системний аналіз проблеми моніторингу якості атмосферного повітря.
2. Вивчити існуючі технології OLAP та їхнє застосування для багатовимірного аналізу даних.
3. Дослідити методи Data Mining, зокрема асоціативні правила та кластеризацію, для виявлення закономірностей у даних.
4. Розробити та протестувати моделі Machine Learning для прогнозування екологічних показників, враховуючи вплив кластеризації.
5. Побудувати звіти та візуалізації даних для аналізу трендів, просторово-часових розрізів та категорій забруднення.

Методи дослідження. У роботі використовуються наступні методи:

- Технологія OLAP для багатовимірного аналізу даних.
- Методи Data Mining, такі як кластеризація та асоціативні правила.

- Алгоритми машинного навчання (рандомний ліс, градієнтний бустинг).
- Інструменти візуалізації даних, такі як Power BI та SSRS.
- Розрахунок KPI на основі багатовимірного кубу OLAP.

Наукова цінність роботи полягає в комплексному підході до аналізу та прогнозування якості повітря, який поєднує OLAP-технології для багатовимірного аналізу даних з передовими методами машинного навчання. Зокрема, дослідження впливу кластеризації на точність моделей Random Forest та Gradient Boosting у контексті прогнозування якості повітря є актуальним і недостатньо вивченим у сучасній науковій літературі.

1 СИСТЕМНИЙ АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

1.1 Опис предметної області

Забруднення повітря є складною сумішшю шкідливих речовин, що надходять як з антропогенних, так і з природних джерел. Антропогенні джерела включають викиди від транспорту, використання палива для опалення будинків, промислові процеси та виробництво електроенергії. Зокрема, електростанції, що працюють на вугіллі, та викиди від хімічного виробництва є значними джерелами забруднення [2]. Природні джерела також вносять свій вклад у забруднення повітря. До них належать лісові пожежі, які часто спричинені діяльністю людини, виверження вулканів та виділення газів, таких як метан, під час розкладання органічних речовин у ґрунтах [3].

Основні забруднюючі речовини та їх вплив.

1. Транспортне забруднення повітря є сумішшю газів і часток, що містять приземний озон, різні форми вуглецю, оксиди азоту (NO2), оксиди сірки (SO2), леткі органічні сполуки, поліциклічні ароматичні вуглеводні та дрібнодисперсні частки (PM) [4]. Ці речовини утворюються внаслідок викидів від автомобілів, промисловості та спалювання викопного палива.

2. Приземний озон (O_3) утворюється в атмосфері під дією сонячного світла в результаті хімічних реакцій між оксидами азоту та ЛОС, які викидаються автомобілями, електростанціями та промисловими підприємствами [5]. Високі концентрації озону можуть викликати подразнення дихальних шляхів та знижувати функцію легенів.

3. Шкідливі гази, такі як вуглекислий газ (CO_2), монооксид вуглецю (CO), оксиди азоту та сірки, є компонентами викидів від транспорту та промислових процесів. Вони можуть впливати на серцево-судинну систему та загальний стан здоров'я.

4. Дрібнодисперсні частки ($PM_{2.5}$ та PM_{10}) складаються з сульфатів, нітратів, вуглецю та мінерального пилу. Основними джерелами є викиди від транспорту, промисловості, сигаретного диму та спалювання органічних

матеріалів, наприклад, під час лісових пожеж. $PM_{2.5}$, які у 30 разів тонші за людську волосину, можуть проникати глибоко в легені та спричиняти серйозні проблеми зі здоров'ям, включаючи захворювання серця та легенів. Зображення розмірів забруднюючих частинок представлено на рис. 1.

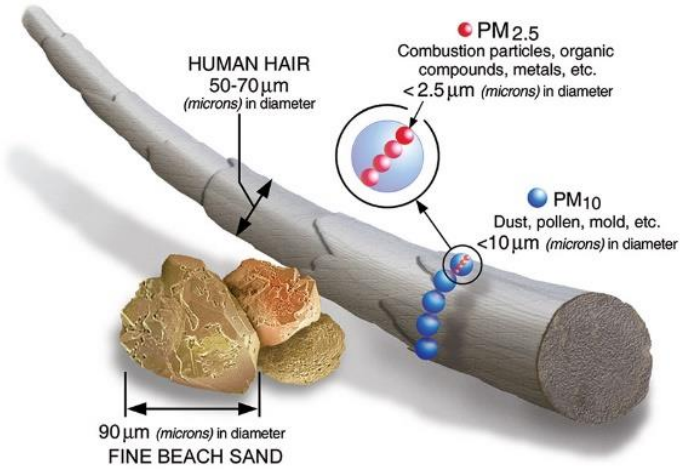

Рис 1. Дрібнодисперсні частинки

5. Поліциклічні ароматичні вуглеводні є органічними сполуками, що складаються з вуглецю та водню. Вони утворюються під час неповного згоряння органічних матеріалів і є відомими канцерогенами. ПАВ присутні у викидах від промислових процесів, виробництва електроенергії та містяться у дрібнодисперсних частках.

Спочатку забруднення повітря розглядалося переважно як загроза для дихальної системи. Проте дослідження показали, що воно пов'язане з окислювальним стресом та запаленням у клітинах, що може призводити до розвитку хронічних захворювань та раку [6]. У 2013 році Міжнародне агентство з дослідження раку (IARC) класифікувало забруднення повітря як канцероген для людини [7]. Короткочасний вплив підвищених рівнів забруднення повітря пов'язаний зі зниженням функції легенів, астмою, серцевими проблемами та збільшенням госпіталізацій [8]. Дрібнодисперсні частки $PM_{2.5}$ асоціюються з підвищеним ризиком смертності [9].

Дослідження показали, що зменшення забруднення повітря сприяє зниженню смертності. Наприклад, після введення норм щодо забруднення

повітря та закриття вугільних електростанцій спостерігалося зменшення рівня смертності, пов'язаної з $PM_{2.5}$ [10].

Основні проблеми громадського здоров'я, пов'язані з забрудненням повітря:

1. Рак: Підвищений ризик розвитку раку молочної залози, лейкемії та раку легенів пов'язаний з впливом забруднюючих речовин [11].

2. Серцево-судинні захворювання: $PM_{2.5}$ можуть порушувати функцію кровоносних судин та прискорювати кальцифікацію в артеріях, що призводить до підвищеного ризику інсульту та інших серцево-судинних захворювань [12].

3. Респіраторні захворювання: Забруднення повітря впливає на розвиток легенів, сприяє виникненню емфіземи, астми та хронічного обструктивного захворювання легень (ХОЗЛ). Діти, які живуть у містах з високим рівнем забруднення, частіше страждають на астму [13].

Спостереження за вмістом забруднюючих речовин у повітрі міст України є повноваженням Українського гідрометцентру, який є підрозділом Державної служби з надзвичайних ситуацій у складі Міністерства внутрішніх справ. Моніторинг здійснюється відповідно до Постанови Кабінету міністрів № 343 від 9 березня 1999 року та Керівництва з контролю забруднення атмосфери РД 52.04.186-89, затвердженого майже три десятиліття тому.

Основним методом визначення концентрацій забрудників є відбір проб повітря на стаціонарних постах спостереження. Кількість постів визначається розміром міста і особливостями структури промисловості. Вона може коливатись від одного поста для міст з населенням, меншим за 50 тисяч мешканців, до двадцяти постів для міст-мільйонників [14].

Гранично допустима концентрація (ГДК) — це кількість забруднюючої речовини в середовищі, яка при постійному або тимчасовому впливі не шкодить здоров'ю людини і не викликає негативних змін у майбутніх поколінь [15]. Для оцінки якості повітря фактичні концентрації забрудників порівнюються з ГДК. Порівняння таких показників дозволяє визначити кратність перевищення ГДК, що характеризує рівень впливу речовини на здоров’я. Однак, через різну

шкідливість речовин пряме порівняння є некоректним. Для подолання цього обмеження використовується індекс забруднення атмосфери (ІЗА), який враховує як перевищення концентрацій, так і клас небезпечності речовин, дозволяючи оцінити комплексне забруднення [14].

Вимірювання забруднення здійснюється з урахуванням концентрації речовини та визначених порогових значень. Високі концентрації потребують точного вимірювання, тоді як для низьких рівнів допускаються приблизні оцінки. Показники забруднення аналізуються для визначення рівня (допустимий чи недопустимий) та ступеню небезпечності (від безпечного до дуже небезпечного) [16]. У ЄС основними забруднюючими речовинами є діоксид сірки, оксиди азоту, частинки PM10 та PM2.5, свинець, бензол, монооксид вуглецю та озон, для яких встановлені окремі регуляції [14].

Відповідно до Постанови КМУ №827 від 14.08.2019, система моніторингу якості повітря в Україні охоплює вимірювання ключових забрудників, таких як дрібні тверді частинки (PM10 та PM2.5), чадний газ (CO), оксиди азоту (NO, NO2), сірчисті сполуки (SO2, H2S), озон (O3), бензапірен, фенол і формальдегід. Ці речовини визначають стан повітря та його вплив на здоров’я, зокрема ризики для дихальної, серцево-судинної систем і онкологічні загрози [14].

1.2 Огляд сучасних досліджень у сфері моніторингу якості повітря

Моніторинг якості повітря є однією з ключових сфер екологічних досліджень, яка стрімко розвивається завдяки впровадженню сучасних технологій збору, аналізу та прогнозування даних. В умовах зростання екологічних загроз, спричинених антропогенним впливом, існує гостра потреба у вдосконаленні методів аналізу та прогнозування параметрів атмосферного повітря.

В дослідженні «*Hybridization of Air Quality Forecasting Models Using Machine Learning and Clustering*» Вані В. Тамаш, Жиль Ноттон, Крістоф Паолі, Марі-Лор Ніве, Кирило Воянт було проведене комбінування штучних нейронних

мереж (MLP) із методами кластеризації для прогнозування концентрацій озону (O3), діоксиду азоту (NO2) і зважених частинок (PM10) на 24 години наперед. В якості вхідних даних використовувалися погодні показники та концентрації забруднюючих речовин. Важливо, що кластеризація дозволила створювати окремі моделі для кожного типу погодних умов, що суттєво підвищувало точність прогнозування пікових значень концентрацій.

Для реалізації цього завдання дослідники використовували штучні нейронні мережі (MLP), які показали здатність моделювати нелінійні залежності між метеорологічними змінними та концентраціями забруднюючих речовин. Архітектура нейронної мережі включала один прихований шар із десятьма нейронами, що забезпечувало баланс між складністю моделі та уникненням перенавчання. Крім того, для підвищення точності прогнозування пікових значень було інтегровано кластеризацію, яка дозволила групувати дані за подібними метеорологічними умовами. Кожна група отримала окрему спеціалізовану модель.

Кластеризація проводилася двома методами: ієрархічною кластеризацією та комбінацією самоорганізованих карт (SOM) і методу k-means. Цей етап дозволив створити кілька підгруп, кожна з яких мала спільні характеристики метеорологічних даних і забруднюючих речовин. Вибір цих методів зумовлений їхньою здатністю ефективно виявляти приховані структури у великих масивах даних. У кожній кластерній групі тренувалася окрема нейронна мережа MLP, яка оптимально адаптувалася до умов відповідного кластера.

Результати дослідження демонструють значну перевагу використання кластеризації в поєднанні з нейронними мережами. Для глобальної моделі без кластеризації (fMLP) індекс згоди (IA) склав 0.87 для озону, 0.74 для PM10 і 0.80 для NO2. Однак моделі з кластеризацією (hMLP і kMLP) значно перевершували глобальну модель у прогнозуванні пікових значень концентрацій забруднення, що підтверджується кривими ROC. Це особливо важливо для раннього попередження про перевищення порогових значень забруднення, що може мати критичне значення для здоров’я населення [17].

Дослідження Саміри Мухаммад «*Development and Implementation of Air Quality Data Mart for Ontario, Canada*» присвячене створенню сховища даних для аналізу якості повітря в Онтаріо, Канада, із застосуванням OLAP-технологій. Метою роботи було зберігання історичних даних про забруднення повітря та їх подальший аналіз за допомогою багатовимірної моделі. Сховище базувалося на зірковій схемі, що дозволило аналізувати дані за ключовими вимірами, такими як час, місцезнаходження та забруднювачі (наприклад, NO_2, PM2.5, O_3). Ефективність сховища перевірялася через порівняння міських і сільських районів, пікових і позапікових періодів. Результати показали значне перевищення норм забруднення, особливо у великих містах, із сезонною варіативністю рівнів озону та PM2.5 [18].

Розроблена система виявила важливість OLAP-технологій для моніторингу та аналізу забруднення повітря, надаючи можливість ідентифікувати проблемні місця та тренди забруднення. Використання багатовимірного аналізу для таких показників, як PM2.5 та O_3, може бути цінним для мого дослідження, де також передбачено створення сховища даних та прогнозування на основі історичних даних. Застосування подібного підходу в комбінації з моделями прогнозування дозволить більш точно оцінювати якість повітря та вплив різних факторів.

Дослідження "*Forecasting air quality time series using deep learning*" було присвячене застосуванню методів глибинного навчання для прогнозування концентрацій озону (O3) на основі часових рядів. У роботі використовувалася рекурентна нейронна мережа (RNN) з довготривалою короткостроковою пам'яттю (LSTM). Вона дозволила прогнозувати 8-годинні середні концентрації озону з високою точністю для періодів до 72 годин, при цьому середня абсолютна помилка (MAE) становила менше 2 одиниць. Методика передбачала очищення та підготовку даних, включно з обробкою пропущених значень та зниженням розмірності вхідних характеристик за допомогою дерев рішень.

Головною перевагою підходу є спроможність LSTM обробляти складні залежності в часових рядах, включаючи змінні погодні умови та транспортні

механізми забруднювачів. Автори продемонстрували, що зменшення кількості вхідних параметрів з 25 до 5 за допомогою дерев рішень не тільки покращило точність прогнозу, але й скоротило час навчання моделі. Важливими факторами виявилися озон, сонячна радіація, напрямок вітру та концентрація метану [19].

Це дослідження має безпосередню цінність для мого проекту, оскільки демонструє ефективність глибинного навчання у прогнозуванні забруднення повітря, враховуючи часові закономірності. Застосування RNN з LSTM може бути інтегроване в систему моніторингу якості повітря, що дозволить підвищити точність прогнозування та своєчасність прийняття рішень щодо управління екологічною ситуацією.

У дослідженні "*Machine learning algorithms to forecast air quality: a survey*" проведено огляд методів машинного навчання (ML) для прогнозування якості повітря за період 2011–2021 років. Автори проаналізували 155 статей, вивчаючи географічний розподіл, типи передбачуваних показників, предикторні змінні та оцінювальні метрики. Головні результати демонструють, що найбільш прогнозованими показниками є індекс якості повітря (AQI) та концентрації забруднюючих речовин, таких як PM2.5 і PM10. Для цього найбільш ефективними виявилися алгоритми глибокого навчання (DL), особливо LSTM, завдяки їх здатності обробляти часові ряди з високою точністю [20].

Методологія дослідження включала використання різних джерел для відбору статей, після чого відбиралися ті, що фокусувалися на прогнозуванні майбутніх значень AQI або концентрації забруднюючих речовин. Результати вказують на те, що більшість моделей використовують поєднання змінних, включаючи метеорологічні фактори (температура, швидкість вітру) і концентрації забруднювачів. Гібридні моделі, які поєднують алгоритми DL з регресійними моделями, також демонструють високу ефективність, що є цінним для вашого дослідження, орієнтованого на моніторинг якості повітря та використання кластеризації і прогнозування.

У дослідженні "*Air Quality Analysis Based On MapReduce and K-Means: A Decision Making System*" основна увага приділяється впровадженню інтегрованої

системи для аналізу якості повітря з використанням сучасних технологій обробки великих даних та алгоритму кластеризації K-means. Основними цілями були підвищення ефективності обробки даних, їх візуалізація, а також створення аналітичних інструментів для підтримки прийняття рішень.

Для збору та зберігання даних було використано платформу Hadoop HBase, яка забезпечує високу масштабованість та швидкість обробки великих обсягів інформації. Дані з трьох моніторингових станцій в Марракеші (Мхамід, Даудіят, Джамаа Ель Фна) оброблялися із застосуванням моделі багатовимірного аналізу (OLAP) та Hadoop MapReduce для паралельної обробки. Кластеризація даних проводилася за допомогою алгоритму K-means, що дозволило виділити три основні групи, засновані на концентрації озону та рівні сонячної радіації [21].

Було виділено три кластери:

1. Кластер 1 (вечірній час) характеризувався змінною сонячною радіацією залежно від сезону.
2. Кластер 2 (денний час) мав постійно високий рівень радіації, що сприяло підвищенню концентрації озону.
3. Кластер 3 (нічний час) характеризувався відсутністю сонячної радіації і, відповідно, низькою концентрацією озону.

Система продемонструвала залежність концентрації озону від добових та сезонних змін сонячної радіації. Використання OLAP та кластеризації дозволило візуалізувати просторово-часовий розподіл забруднювачів і їх динаміку.

Проведені дослідження надають цінний досвід у сфері моніторингу якості повітря, пропонуючи різноманітні підходи до аналізу, прогнозування та обробки даних. Використання OLAP-технологій для багатовимірного аналізу, алгоритмів кластеризації для виявлення патернів, а також методів глибокого навчання та машинного навчання для точного прогнозування забруднення повітря дозволить створювати комплексні рішення, які можуть бути інтегровані у мою систему моніторингу. Вивчені підходи та результати можуть бути використані для покращення точності прогнозів, ідентифікації критичних факторів забруднення та підвищення ефективності прийняття рішень, що відповідає меті дослідження.

1.3 Постановка завдання для аналізу

Для досягнення мети дослідження визначено наступні завдання:

1. *Аналіз предметної області:*
    - Опис процесів моніторингу якості атмосферного повітря та характеристика параметрів забруднення (PM2.5, CO, O3 та інші).
    - Аналіз існуючих підходів до моніторингу, зокрема використання технологій OLAP та методів Data Mining.
    - Вивчення існуючих рішень у вигляді статей, патентів та практичних впроваджень.
2. *Розробка структури сховища даних:*
    - Проектування сховища даних для зберігання інформації про параметри якості повітря.
    - Впровадження механізмів ETL для вилучення, трансформації та завантаження даних із джерел, таких як API SaveEcoBot та станції моніторингу КМДА.
3. *Реалізація багатовимірного аналізу даних:*
    - Побудова OLAP-кубу для аналізу даних за такими вимірами: час, локація, показники забруднення та категорії.
    - Розрахунок ключових показників ефективності (KPI) для оцінки стану повітря.
4. *Розробка аналітичних звітів:*
    - Середні значення показників у розрізі міст за вибраний період.
    - Забруднення в розрізі місяця по місту для вибраного показника.
    - Забруднення в розрізі категорії та часу для вибраного показника по місту.
    - Аналіз трендів за показниками забруднення.
5. *Дослідження методів Data Mining:*

- Вивчення методів асоціативних правил для виявлення взаємозв’язків між екологічними показниками.
- Дослідження алгоритмів кластеризації, таких як k-середніх, для групування даних.

6. *Розробка моделей прогнозування:*
    - Реалізація моделей машинного навчання, зокрема Random Forest та Gradient Boosting, для прогнозування показників забруднення повітря.
    - Аналіз впливу кластеризації на точність прогнозування моделей.
7. *Порівняння підходів до прогнозування:*
    - Оцінка ефективності прогнозування з використанням кластеризації та без неї.
    - Порівняння точності моделей Random Forest та Gradient Boosting у різних сценаріях.
8. *Інтерпретація результатів та розробка рекомендацій:*
    - Аналіз отриманих результатів багатовимірного аналізу, звітів та прогнозування.
    - Формулювання рекомендацій щодо застосування кластеризації та інших підходів у системах моніторингу якості повітря.

2 МОДЕЛЮВАННЯ СИСТЕМИ

2.1 Опис та аналіз методології системного аналізу

Системний аналіз — це структурована методологія, яка використовується для розуміння та оцінки складних систем у різних областях, включаючи інформаційні технології, інженерні та бізнес-процеси. Цей підхід передбачає поділ системи на її фундаментальні компоненти, щоб зрозуміти їхню взаємодію та функціональність, що сприяє прийняттю обґрунтованих рішень і покращенню системи.

Ключовим аспектом системного аналізу є розробка моделі системи. Ця модель служить абстракцією реальної системи, що дозволяє аналітикам симулювати та оцінювати різні сценарії, не впливаючи на реальну систему. Процес часто включає два етапи: роз’єднання моделі, де компоненти системи спрощуються та аналізуються як підсистеми, та інтеграція моделі, де ці підсистеми об’єднуються, щоб представити всю систему. Встановлення балансу між деталізацією моделі та аналітичною ефективністю має вирішальне значення, оскільки надто складні моделі можуть перешкоджати ефективному аналізу [22].

Системний аналіз є фундаментальним підходом, який використовується для забезпечення структурованого розуміння складних систем, їхніх компонентів і взаємодії між ними. Його основна мета полягає у визначенні та аналізі вимог, пошуку оптимальних рішень і їх інтеграції для досягнення поставлених цілей. У контексті інтелектуальних систем моніторингу, системний аналіз дозволяє зрозуміти ключові процеси, пов'язані з обробкою великих обсягів даних, виявленням аномалій і прогнозуванням параметрів навколишнього середовища.

Методологія системного аналізу включає кілька етапів: визначення цілей системи, розробку моделі процесів, аналіз взаємозв'язків між компонентами системи, вибір відповідних інструментів та технологій для її реалізації. Це дозволяє зосередитися на вирішенні конкретних проблем і досягненні максимальної ефективності системи [23].

Перший крок у системному аналізі – це визначення вимог, яке включає збір та аналіз даних від зацікавлених сторін. Цей процес допомагає уточнити основні завдання, які система має виконувати. Далі здійснюється розробка моделі, яка представляє функціональні та нефункціональні вимоги до системи, включаючи структуру даних, їхній потік і взаємодію між компонентами [24].

Окрім того, аналіз альтернатив є важливою складовою системного аналізу. Він дозволяє порівняти кілька підходів до вирішення задачі, визначити їх переваги, ризики та вплив на загальну продуктивність системи. Цей процес передбачає використання кількісних і якісних методів оцінки, таких як SWOT-аналіз і функціональний аналіз [25].

Одним із ключових інструментів системного аналізу є моделювання. Це можуть бути як математичні моделі, так і моделі бізнес-процесів, які допомагають формалізувати взаємодію компонентів системи [26]. Наприклад, для моніторингу повітря використовуються моделі, які враховують динаміку зміни забруднень і прогнозують їхній вплив на здоров’я населення.

Застосування системного аналізу також включає інтеграцію інструментів для збору, обробки та аналізу даних. Для цього можуть використовуватися платформи типу SQL Server, OLAP-куби для агрегування даних і Power BI для візуалізації. У випадку Python використовуються бібліотеки для аналізу даних, такі як Pandas і Scikit-learn, що дозволяють реалізувати моделі машинного навчання.

Результати системного аналізу лягають в основу проектування системи, її тестування та впровадження. У контексті вашої теми аналіз допомагає інтегрувати кластеризацію даних і прогнозування забруднення повітря, що дозволяє підвищити точність оцінок і розробити ефективні стратегії реагування.

Під методологією системного дослідження розуміють сукупність методів і засобів, спрямованих на вирішення складних проблем з урахуванням їхньої цілісності. Системний метод — це впорядкований підхід до досягнення мети, заснований на принципах аналізу і синтезу, а системні засоби включають принципи та поняття, які визначають структуру й методи роботи [27].

Системне дослідження проблеми зазвичай складається з таких етапів:

1. **Формулювання проблеми**: проблема визначається як невідповідність між необхідним і фактичним станом справ. У реальних умовах будь-яку проблему слід розглядати в контексті пов’язаних із нею задач, тобто як частину "системи проблем", що взаємозалежні та потребують комплексного підходу до вирішення [27].

2. **Визначення цілей**: аналіз проблеми включає вибір напрямку для вирішення, який найбільше відповідає досягненню цілі. Складність полягає в тому, що можливих шляхів може бути багато, але слід обрати оптимальний [27].

3. **Формулювання критеріїв і обмежень**: критерії визначають якісні ознаки альтернатив, виражені у кількісних показниках, а обмеження встановлюють межі, яких необхідно дотримуватися під час пошуку рішень. Оптимізація за критерієм забезпечує максимальне наближення до поставленої цілі [27].

4. **Генерація альтернатив і сценаріїв**: на цьому етапі висувається максимальна кількість ідей про можливі шляхи досягнення цілей. Критерії допомагають знайти нові альтернативи, а обмеження скорочують їх кількість, виключаючи неприйнятні варіанти [27].

5. **Оцінка ресурсів**: для кожної альтернативи визначаються ресурси, необхідні для її реалізації. У разі нестачі ресурсів або занадто жорстких обмежень, цілі можуть бути переглянуті або адаптовані [27].

Методологія системного аналізу дозволяє структуровано та послідовно вирішувати навіть найскладніші завдання, забезпечуючи їхню адаптивність до змін у зовнішньому середовищі.

Основні етапи методики системного аналізу (розроблені Стенлі Янгом), що найчастіше застосовується на практиці:

1. визначення мети організації;
2. виявлення проблем організації;
3. дослідження проблем і постановка діагнозу;
4. пошук розв'язку проблеми;

5. оцінка всіх альтернатив і вибір найкращої з них;
6. узгодження рішення в організації;
7. затвердження рішення;
8. підготовка до введення в дію;
9. управління застосуванням рішення;
10. перевірка ефективності рішення [27].

2.2 Діаграма прецедентів

Діаграма варіантів використання в уніфікованій мові моделювання (UML) - це візуальне представлення, яке ілюструє взаємодію між користувачами (акторами) і системою. Вона фіксує функціональні вимоги системи, показуючи, як різні користувачі взаємодіють із різними варіантами використання або певними функціями в системі. Діаграми варіантів використання забезпечують огляд поведінки системи на високому рівні, що робить їх корисними для зацікавлених сторін, розробників і аналітиків, щоб зрозуміти, як система призначена для роботи з точки зору користувача, і як різні процеси пов'язані один з одним. Вони мають вирішальне значення для визначення обсягу системи та вимог [28].

Основна мета діаграми прецедентів полягає в ідентифікації та організації вимог до системи з точки зору зовнішніх користувачів. Це дозволяє розробникам та зацікавленим сторонам зрозуміти, які функції повинна виконувати система, забезпечуючи ефективну комунікацію між технічними та нетехнічними учасниками проекту.

У контексті розробки інтелектуальної системи моніторингу параметрів атмосферного повітря, діаграма прецедентів служить для відображення взаємодії між користувачами системи (наприклад, екологами, аналітиками даних, адміністраторами) та функціональними можливостями системи, які включають аналіз забрудненості повітря за допомогою OLAP-технологій та прогнозування з використанням кластеризації. Діаграму прецедентів до моєї системи зображено на рис. 2.

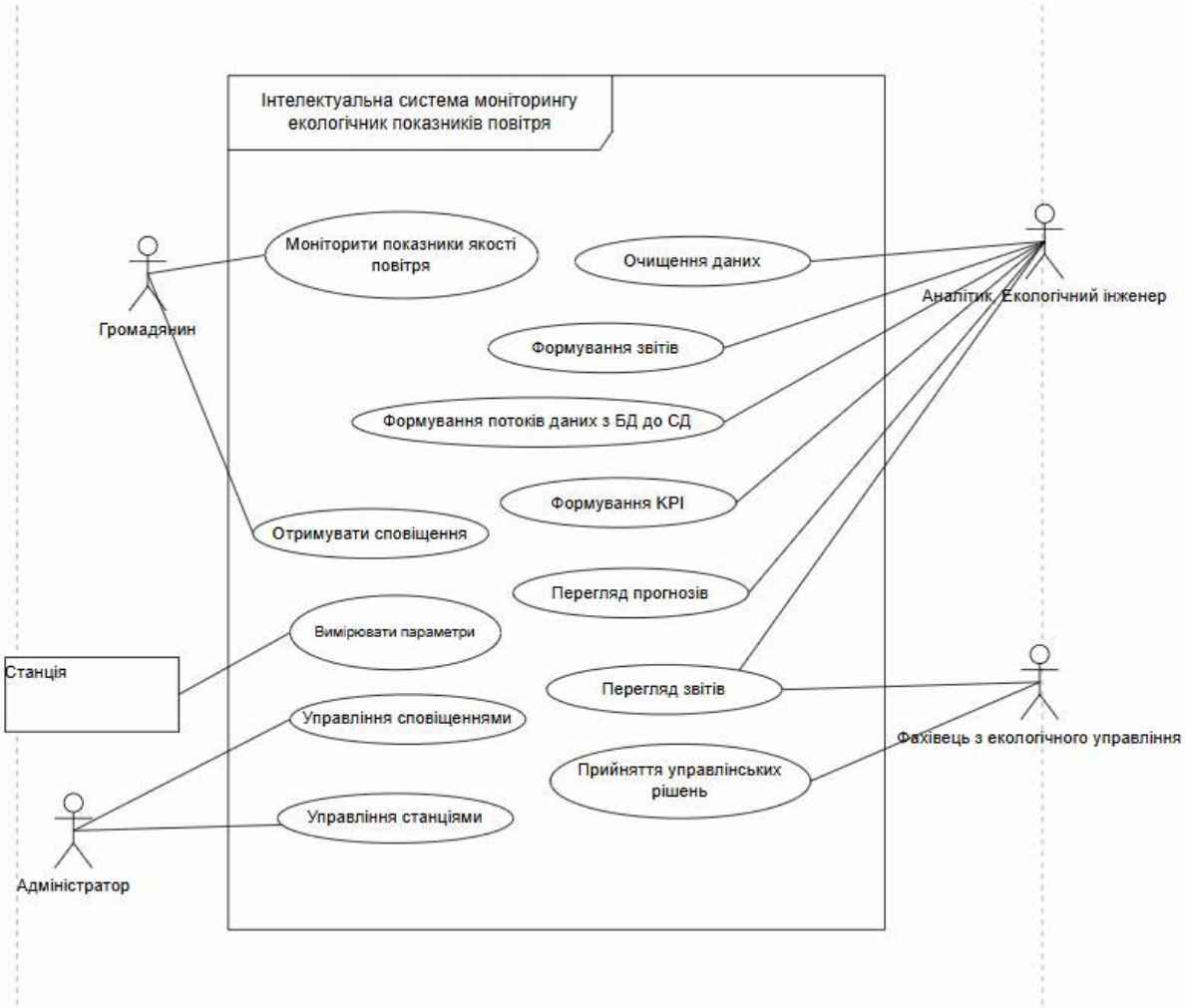

Рис. 2 Діаграма прецедентів

Виділено наступні прецеденти:

1. *Моніторинг показників якості повітря.* Запускається громадянином. Дозволяє користувачеві переглядати актуальні показники якості повітря, такі як рівень забруднення, вміст шкідливих речовин (PM2.5, CO, O3 тощо) у вибраних регіонах та містах.

2. *Отримання сповіщень.* Запускається громадянином. Дозволяє отримувати сповіщення про стан якості повітря в реальному часі, особливо у випадках підвищення рівнів забруднення до небезпечних для здоров'я значень.

3. *Вимірювання параметрів.* Запускається станцією. Включає збір даних про різні екологічні показники, такі як рівень забруднення, температура, вологість тощо. Дані передаються на сервер для подальшого аналізу та моніторингу.

4. *Управління сповіщеннями.* Запускається адміністратором. Дозволяє налаштовувати систему сповіщень для громадян, визначати порогові значення показників забруднення, при досягненні яких будуть надсилатися сповіщення, а також керувати процесом надсилання повідомлень.

5. *Управління станціями.* Запускається адміністратором. Дозволяє адміністратору додавати нові станції моніторингу, оновлювати їхню інформацію, а також керувати активністю станцій, забезпечуючи своєчасне надходження даних.

6. *Очищення даних.* Запускається аналітиком. Включає процес обробки та очищення отриманих даних від можливих шумів і аномалій для забезпечення точності аналізу та прогнозів.

7. *Формування звітів.* Запускається аналітиком. Дозволяє створювати різні звіти на основі зібраних даних, наприклад, аналіз середніх та максимальних показників забруднення по містах, а також тенденцій змін показників якості повітря.

8. *Формування потоків даних з БД до СД.* Запускається аналітиком. Включає процес передачі очищених даних з оперативної бази даних (БД) до сховища даних (СД) для подальшого багатовимірного аналізу та формування звітів.

9. *Формування KPI.* Запускається аналітиком. Дозволяє визначати ключові показники ефективності (KPI) для оцінки якості повітря та моніторингу змін у параметрах забруднення в різних регіонах.

10. *Перегляд прогнозів.* Запускається аналітиком. Дозволяє аналітику переглядати результати прогнозів щодо рівня забруднення повітря з використанням алгоритмів машинного навчання та кластеризації для оцінки майбутніх тенденцій.

11. *Перегляд звітів.* Запускається аналітиком і працівником МОЗ. Дозволяє переглядати підготовлені звіти для аналізу поточного стану якості повітря та виявлення можливих тенденцій або аномалій.

12. *Прийняття управлінських рішень.* Запускається фахівцем з екологічного управління. На основі переглянутих звітів фахівець може приймати рішення щодо заходів для покращення якості повітря, зокрема у випадках підвищеного рівня забруднення, планувати відповідні дії та впроваджувати стратегії зменшення екологічних ризиків.

2.3 Діаграма розгортання системи моніторингу параметрів повітря

Діаграма розгортання — це тип діаграми, який визначає фізичне обладнання, на якому працюватиме програмна система. Він також визначає, як програмне забезпечення розгортається на базовому обладнанні. Він прив’язує частини програмного забезпечення системи до пристрою, який збирається її виконувати [29].

Діаграма розгортання відображає архітектуру програмного забезпечення, створену при проектуванні, на архітектуру фізичної системи, яка її виконує. У розподілених системах він моделює розподіл програмного забезпечення між фізичними вузлами [29].

Програмні системи проявляються за допомогою різноманітних артефакти, а потім вони відображаються в середовищі виконання, яке збирається виконувати програмне забезпечення, наприклад вузли. У діаграмі розгортання задіяно багато вузлів; отже, зв'язок між ними представлений за допомогою шляхів зв'язку [29].

Діаграму розгортання для моєї системи зображено на рис. 3.

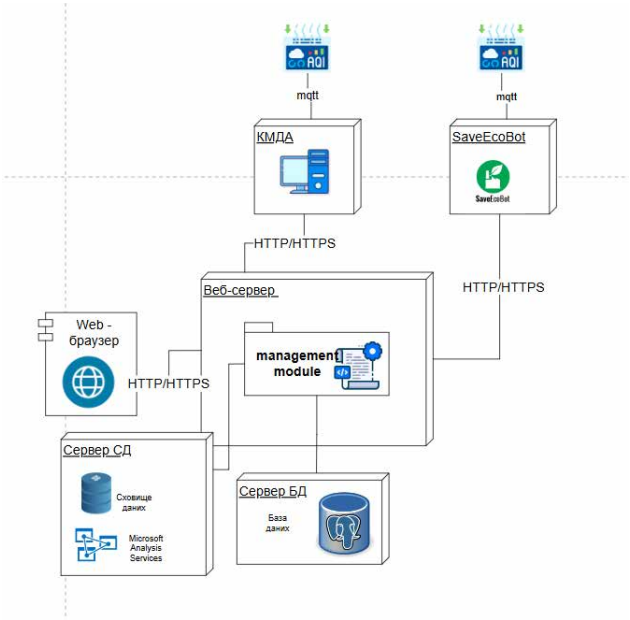

Рис. 3 Діаграма розгортання

На діаграмі розгортання представлено архітектуру системи моніторингу параметрів якості повітря, яка складається з кількох основних компонентів, що взаємодіють між собою. Опис основних елементів подано нижче.

1. **Вузол «Датчики»**. Датчики, які збирають дані про якість повітря, передають інформацію через протокол MQTT на сервери КМДА та SaveEcoBot.

2. **Вузол станцій КМДА та SaveEcoBot**: ці компоненти отримують дані від датчиків і передають їх на веб-сервер системи через HTTP або HTTPS. КМДА є державним сервером, а SaveEcoBot — це платформа для моніторингу якості повітря, яка отримує інформацію з різних джерел.

3. **Вузол Веб-сервер** містить **Management Module** - основний модуль керування, що відповідає за обробку отриманих даних від КМДА і SaveEcoBot та управління передачею даних на інші компоненти системи. Також веб-сервер забезпечує доступ користувачів до системи через веб-браузер, використовуючи HTTP/HTTPS.

4. **Вузол Сервер СД (Сховище даних)**: використовується для зберігання агрегованих та оброблених даних, які надходять з основної бази даних. Тут знаходяться Microsoft Analysis Services, які підтримують OLAP-куб

для аналізу даних. Цей компонент дозволяє зберігати великі обсяги історичних даних для подальшого аналізу, формування звітів та прогнозів.

5. **Вузол Сервер БД (База даних)**: основна база даних, де зберігаються первинні дані, отримані з датчиків, а також інші оперативні дані системи. Це джерело інформації для СД (Сховища даних).

6. **Користувач (веб-браузер)**: користувачі отримують доступ до системи через веб-браузер, де можуть переглядати аналітичні звіти, графіки, результати прогнозування та іншу інформацію, пов’язану з якістю повітря. Взаємодія з веб-сервером також відбувається через протокол HTTP/HTTPS.

3 РОЗРОБКА СИСТЕМИ

3.1 Структура джерел даних та їх підготовка для аналізу

3.1.1 Структура оперативної бази даних

У цьому розділі описано структуру джерел даних, які використовуються для збору, зберігання та аналізу параметрів якості повітря в системі моніторингу. В основі розробленої системи лежить оперативна база даних, яка містить різні таблиці, що зберігають інформацію про користувачів, станції моніторингу, показники якості повітря, вимірювання та налаштування серверів. Ці дані є основою для подальшого аналізу, кластеризації та прогнозування показників, а також для генерації звітів. Схему оперативного джерела БД представлено на рис. 4.

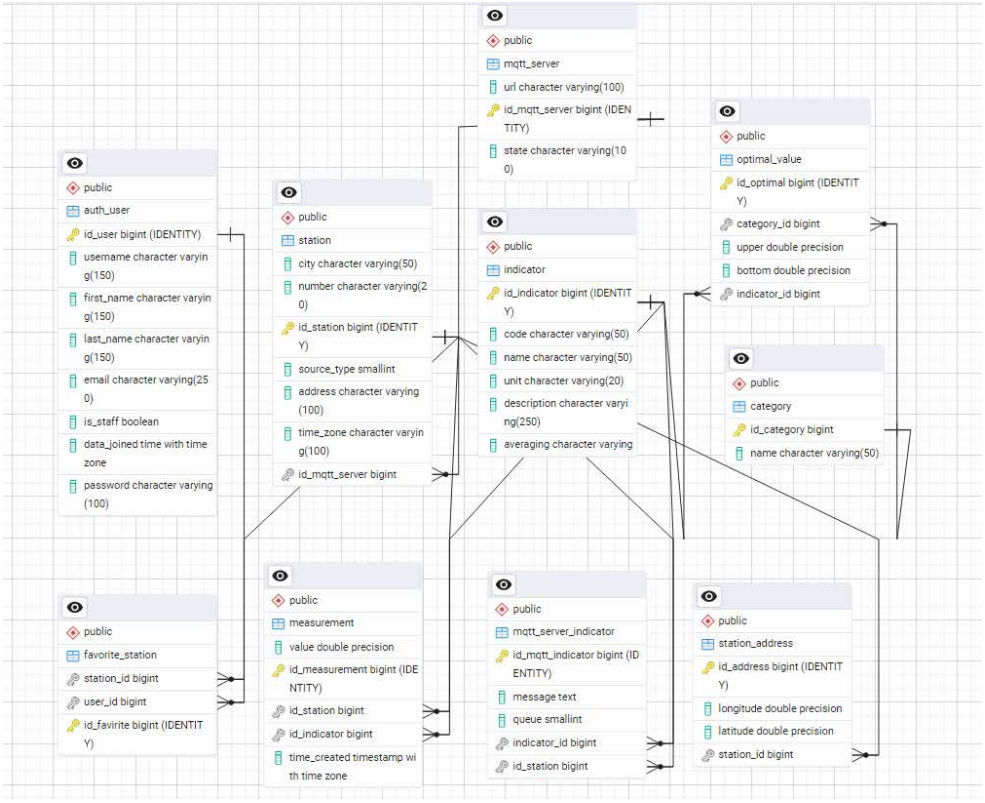

Рис. 4 Схема оперативної БД

В оперативній базі даних, яка є основою системи моніторингу якості повітря, зберігаються всі необхідні дані для забезпечення аналітичного процесу. Вона організована у вигляді кількох основних таблиць, кожна з яких виконує

конкретну функцію. Основну інформацію зберігають таблиці, що містять дані про користувачів, станції моніторингу, показники забруднення, вимірювання якості повітря та конфігурацію серверів. Ця структура забезпечує ефективне зберігання, доступ та обробку даних у реальному часі.

Таблиці для зберігання даних про станції та користувачів дозволяють системі ідентифікувати кожного користувача та станцію моніторингу, що спрощує управління доступом, а також дозволяє налаштовувати індивідуальні запити та уподобання. Інформація про станції включає їхні географічні координати, що важливо для геопросторового аналізу, а також часові пояси, які дозволяють враховувати відмінності у часових даних.

Таблиця з показниками забруднення повітря містить детальну інформацію про кожен параметр, такий як PM2.5, CO та інші показники, що вимірюються на станціях. Вона включає опис, одиниці вимірювання та категорії допустимих рівнів, які використовуються для аналізу і оцінки якості повітря. Дані про вимірювання, які надходять із станцій у реальному часі, зберігаються з прив’язкою до конкретної станції та часу вимірювання, що дозволяє відслідковувати динаміку змін показників на окремих територіях.

Така структура забезпечує взаємозв’язок між даними, дозволяє ефективно зберігати великі обсяги інформації та використовувати ці дані для аналітичної обробки, кластеризації та прогнозування. Завдяки цьому система моніторингу здатна надавати актуальну інформацію щодо якості повітря та прогнозувати можливі зміни рівнів забруднення.

3.1.2 Основні поняття OLAP-технологій

OLAP (Online Analytical Processing) — це технологія, що дозволяє швидко виконувати складні запити та багатовимірний аналіз великих обсягів даних у сховищах даних, озерах даних або інших репозиторіях. OLAP використовується в бізнес-аналітиці (BI), системах підтримки прийняття рішень та різноманітних бізнес-прогнозах і звітності. Більшість бізнес-даних мають кілька вимірів — категорій, за якими дані розбиваються для представлення, відстеження або аналізу [30].

OLAP-куб — це структура даних, що дозволяє швидко аналізувати дані за кількома вимірами, які визначають бізнес-проблему. Наприклад, багатовимірний куб для звітності про продажі може складатися з таких вимірів: продавець, сума продажів, регіон, продукт, місяць, рік.

OLAP-інструменти дозволяють користувачам інтерактивно аналізувати багатовимірні дані з різних перспектив. Основні аналітичні операції в OLAP включають:

- **Консолідація (roll-up)**: агрегація даних за одним або кількома вимірами. Наприклад, об'єднання всіх офісів продажів у відділ продажів для прогнозування тенденцій.
- **Деталізація (drill-down)**: перехід до детальніших даних. Наприклад, перегляд продажів за окремими продуктами в межах регіону.
- **Слайсинг і дайсинг (slicing and dicing)**: вибір певного набору даних з OLAP-куба та перегляд його з різних точок зору [31].

Існують різні типи OLAP-систем:

- **MOLAP (Multidimensional OLAP)**: використовує багатовимірні бази даних для зберігання даних.
- **ROLAP (Relational OLAP)**: працює поверх реляційних баз даних, використовуючи спеціальні запити для імітації багатовимірних даних.
- **HOLAP (Hybrid OLAP)**: поєднує підходи MOLAP та ROLAP для досягнення кращої продуктивності та гнучкості [32].

Продовжуючи дослідження можливостей OLAP-технологій, можна відзначити їхнє значення для аналізу даних у різних сферах, зокрема у сфері моніторингу та прогнозування якості повітря.

Оскільки якість повітря залежить від багатьох факторів, таких як місце, час, погодні умови та джерела викидів, для ефективного моніторингу потрібно враховувати багатовимірність цих даних. У даному контексті OLAP-куб надають можливість здійснювати багатовимірний аналіз, структурувати дані за різними вимірами та виконувати складні аналітичні запити для виявлення трендів і закономірностей. Саме багатовимірний аналіз, який дозволяє структурувати

інформацію за часом, місцем, типом забруднювача, є надзвичайно цінним для моніторингу екологічних параметрів повітря.

У межах даного дослідження було побудовано OLAP-куб та сформовані звіти для моніторингу якості повітря за показниками:

- «Максимальні показники забруднення за вказаним показником у розрізі міст за певний рік»
- «Середнє забруднення по окремих параметрах у розрізі локацій»
- Середні значення показників у розрізі часу по вибраних містах»
- Забруднення у розрізі категорії (типу забруднювача) та часу
- Тренди змін забруднення у часі.

Крім того, щоб покращити точність прогнозування якості повітря, було застосовано поєднання методів кластеризації та машинного навчання, зокрема алгоритми Random Forest та Gradient Boosting. Таке поєднання дозволяє групувати дані за схожими характеристиками та будувати окремі моделі для кожного кластера, що сприяє підвищенню точності прогнозу. Це підходить для передбачення рівнів забруднення, враховуючи специфіку забруднення повітря в конкретні часові інтервали та регіони.

Таким чином, використання OLAP-технологій у комбінації з методами кластеризації та прогнозування є актуальним підходом для побудови інтелектуальних систем моніторингу. Він дозволяє не тільки спостерігати за рівнями забруднення, але й здійснювати ефективне прогнозування, що сприяє своєчасному прийняттю заходів задля зменшення негативного впливу на здоров’я людей.

Продовжуючи дослідження, важливо розглянути методи оцінки ефективності впроваджених технологій моніторингу та прогнозування якості повітря. Для цього доцільно застосувати ключові показники ефективності (KPI), які дозволяють кількісно оцінити досягнення поставлених цілей та ефективність процесів.

KPI (Key Performance Indicators) — це метрики, що використовуються для вимірювання успіху організації або окремих співробітників у досягненні

стратегічних та оперативних цілей. Вони надають кількісні дані, що дозволяють оцінити ефективність роботи, контролювати прогрес та виявляти відхилення від плану. Крім того, KPI можуть слугувати основою для мотиваційних систем, заохочуючи співробітників за досягнення цілей, та допомагають керівникам приймати обґрунтовані рішення, базуючись на об'єктивних даних [33].

Для ефективного впровадження KPI у систему моніторингу якості повітря необхідно визначити конкретні, вимірювані, досяжні, релевантні та обмежені в часі показники, які відображатимуть успішність реалізації проєкту. Це можуть бути, наприклад, точність прогнозування рівнів забруднення, швидкість обробки даних, кількість виявлених аномалій або своєчасність реагування на перевищення допустимих норм.

Використання KPI дозволить не лише оцінити поточний стан системи моніторингу, але й виявити слабкі місця та напрями для вдосконалення. Регулярний аналіз цих показників сприятиме підвищенню ефективності роботи системи та забезпеченню високої якості повітря, що є критично важливим для здоров'я населення та довкілля.

3.1.3 Архітектура сховища даних

Архітектура сховища даних є ключовим компонентом у процесі управління та аналізу великих обсягів інформації, отриманої з різних джерел. Сховище даних (Data Warehouse) — це централізована платформа, призначена для зберігання, обробки та аналізу даних з метою підтримки прийняття рішень на основі даних [34].

Основними елементами архітектури сховища даних є джерела даних, процеси ETL (Extract, Transform, Load), саме сховище даних, марти даних, OLAP-сервіси та презентаційний рівень. Джерела даних можуть включати різноманітні внутрішні та зовнішні системи, такі як бази даних, файли чи веб-сервіси. Процес ETL відповідає за вилучення даних з цих джерел, їх трансформацію до потрібного формату та завантаження у сховище даних [35]. Саме сховище даних є централізованою базою даних, оптимізованою для швидкого виконання запитів та аналізу великих обсягів інформації. Марти даних

представляють собою підмножини сховища даних, сфокусовані на конкретних бізнес-напрямках або відділах. OLAP-сервіси (Online Analytical Processing) забезпечують багатовимірний аналіз даних, дозволяючи користувачам виконувати складні запити та отримувати аналітичні звіти [36]. Презентаційний рівень включає інтерфейси користувача та додатки для візуалізації даних, створення звітів та дашбордів.

Існують різні типи архітектур сховищ даних, зокрема однорівнева, дворівнева та трирівнева архітектури. Однорівнева архітектура характеризується розміщенням усіх компонентів на одному сервері і підходить для невеликих організацій з обмеженим обсягом даних [34]. Дворівнева архітектура розділяє сервер бази даних та клієнтські додатки, що покращує продуктивність та безпеку системи [35]. Трирівнева архітектура включає додатковий середній шар — аплікаційний сервер, який обробляє бізнес-логіку та забезпечує масштабованість і гнучкість [36].

Ефективна архітектура сховища даних надає низку переваг, таких як масштабованість, що дозволяє обробляти зростаючі обсяги даних без втрати продуктивності. Вона забезпечує високу продуктивність завдяки швидкому доступу до даних та оперативному виконанню аналітичних запитів. Інтеграція даних з різних джерел в єдину узгоджену структуру сприяє покращенню якості інформації та прийняттю більш обґрунтованих рішень. Крім того, архітектура забезпечує безпеку та контроль доступу, захищаючи конфіденційні дані та керуючи правами доступу користувачів.

Сучасні сховища даних часто інтегруються з хмарними технологіями та підтримують обробку як структурованих, так і неструктурованих даних. Це надає більшу гнучкість, економію ресурсів та можливість використовувати передові аналітичні інструменти та алгоритми машинного навчання [35].

У контексті розробки інтелектуальної системи моніторингу параметрів атмосферного повітря архітектура сховища даних дозволяє ефективно зберігати та аналізувати великі обсяги екологічних даних. Використання OLAP-технологій забезпечує багатовимірний аналіз забрудненості повітря, а застосування методів

прогнозування, таких як кластеризація, рандомний ліс та градієнтний бустінг, підвищує точність прогнозів та сприяє прийняттю обґрунтованих екологічних рішень.

Архітектура сховища даних відіграє ключову роль у забезпеченні узгодженого та систематизованого аналізу даних. У межах інтелектуальної системи моніторингу параметрів атмосферного повітря її реалізація дозволяє не лише зберігати великий обсяг даних із різних джерел, але й забезпечує інтеграцію екологічних, метеорологічних та часових параметрів для отримання комплексної аналітичної картини. Це дає можливість проводити як оперативний моніторинг, так і прогнозування трендів забруднення повітря, враховуючи сезонні, просторові та погодні закономірності.

Особливістю запропонованої архітектури є поєднання багатовимірного аналізу через OLAP-куби із передовими методами прогнозування на основі машинного навчання. Інтеграція цих підходів дозволяє не лише ідентифікувати ключові проблеми, такі як пікові рівні забруднення, але й прогнозувати можливі сценарії розвитку екологічної ситуації в майбутньому. Таким чином, побудована архітектура забезпечує можливість детального аналізу історичних даних, а також ефективного прийняття рішень для попередження екологічних ризиків.
Зображення сховища даних представлено на рис. 5 нижче.

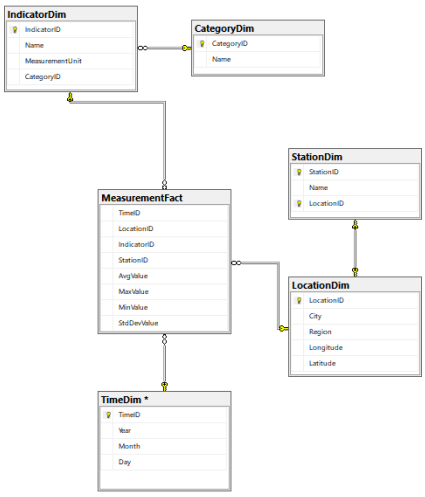

Рис. 5 Архітектура сховища даних

Представлена архітектура сховища даних розроблена для підтримки аналізу та моніторингу параметрів атмосферного повітря. Вона реалізує зіркову схему, де центральна фактова таблиця *MeasurementFact* пов'язана з декількома вимірними таблицями. Вимірні таблиці включають *LocationDim*, яка зберігає географічну інформацію (місто, регіон, довгота, широта); *TimeDim*, що фіксує часові аспекти (рік, місяць, день); *CategoryDim*, яка визначає категорії показників; *IndicatorDim*, що містить деталі про конкретні показники забруднення разом із одиницями вимірювання та пов'язаними категоріями; та *StationDim*, яка представляє моніторингові станції з їхніми назвами та асоційованими локаціями.

Таблиця *MeasurementFact* записує виміряні значення різних екологічних індикаторів. Вона посилається на вимірні таблиці через зовнішні ключі, що дозволяє проводити багатовимірний аналіз за часом, місцем, індикаторами та станціями.

3.1.4 Процеси вилучення, трансформації та завантаження даних

Процеси вилучення, трансформації та завантаження даних (ETL) є фундаментальними компонентами при створенні та підтримці сховищ даних. ETL-процеси відповідають за переміщення даних з різних джерел, їх очищення, перетворення до необхідного формату та завантаження в цільову систему для подальшого аналізу [37]. Цей підхід забезпечує консолідацію даних з різноманітних систем, що дозволяє отримати цілісне та узгоджене уявлення про інформацію.

Для реалізації ETL-процесів часто використовуються спеціалізовані інструменти, такі як SQL Server Integration Services (SSIS). SSIS є компонентом Microsoft SQL Server і надає потужні можливості для інтеграції, перетворення та управління даними [38]. Використання SSIS полегшує створення складних ETL-пакетів, автоматизує процеси імпорту та експорту даних, а також забезпечує високу продуктивність при обробці великих обсягів інформації.

Процес наповнення сховища даних (СД) базується на трьох основних етапах, що забезпечують передачу даних з оперативної бази даних (БД) до

сховища даних для подальшого аналізу. Ці етапи охоплюють вилучення даних, їх трансформацію та завантаження в СД, з урахуванням структурної архітектури СД, що складається з декількох рівнів завдань. Потоки даних представлені на рис. 6.

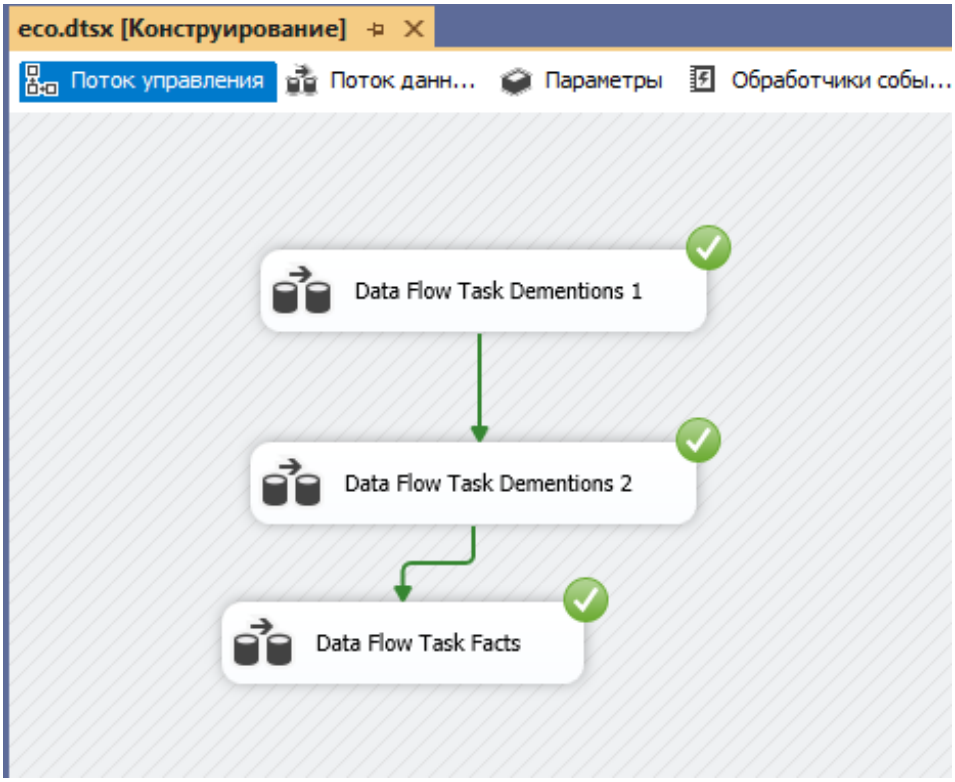

Рис. 6 Потоки даних для наповнення СД

Схема сховища даних містить 2 рівня завдань. На першому етапі здійснюється наповнення таблиць вимірів першого рівня. Наприклад, у цьому етапі обробляються таблиці з основними параметрами, такими як "Місця" (Locations) та "Станції" (Stations). Для наповнення цих таблиць в SSIS (SQL Server Integration Services) використовуються компоненти Sort та Merge Join. Компонент Sort використовується для сортування даних за обраним ключовим полем. Як видно на рис. 7, сортування застосовується окремо для кожного джерела даних перед злиттям.

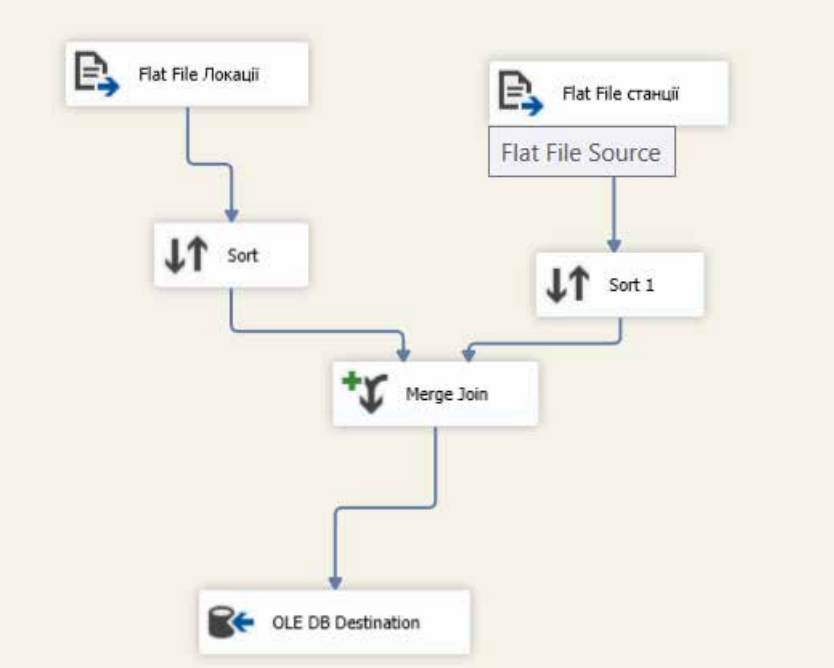

Рис. 7 Заповнення таблиці локацій

Таблиця "Станції" сортується за ідентифікатором станції (id), що дозволяє забезпечити точність об'єднання даних. На рис. 8 зображено сортування за ідентифікатором станції.

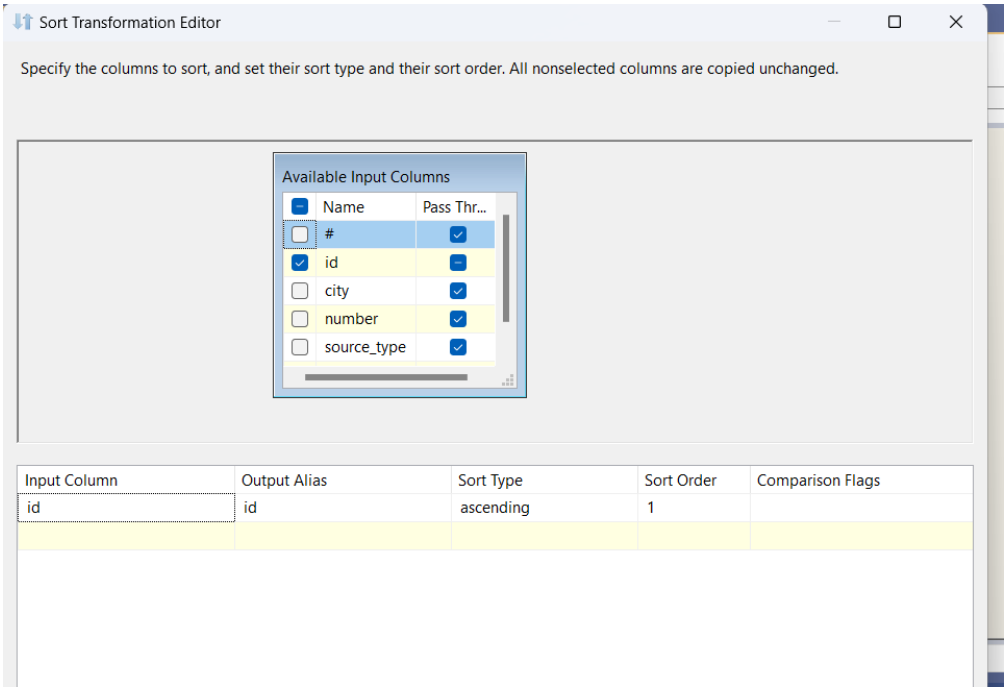

Рис. 8 Сортування за ідентифікатором перед об'єднанням даних

Компонент Merge Join: використовується для злиття двох джерел даних на основі ключового поля. На Рис. показано, як станції об'єднуються з інформацією

про їхні координати (довгота та широта). Використовується внутрішнє злиття (inner join), що забезпечує завантаження лише тих записів, які є в обох джерелах.

На рис. 9 зображено внутрішнє злиття таблиці станції і таблиці локацій.

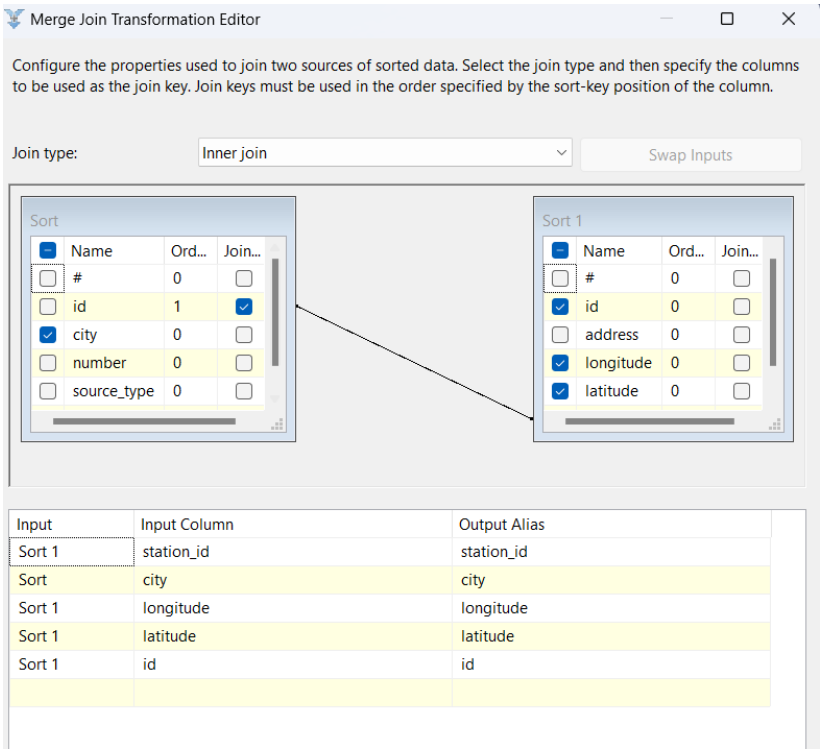

Рис. 9 Об'єднання таблиць станції і локацій

Потік завдань для заповнення сховища даних з оперативного джерела для першого рівня представлено на рис. 10.

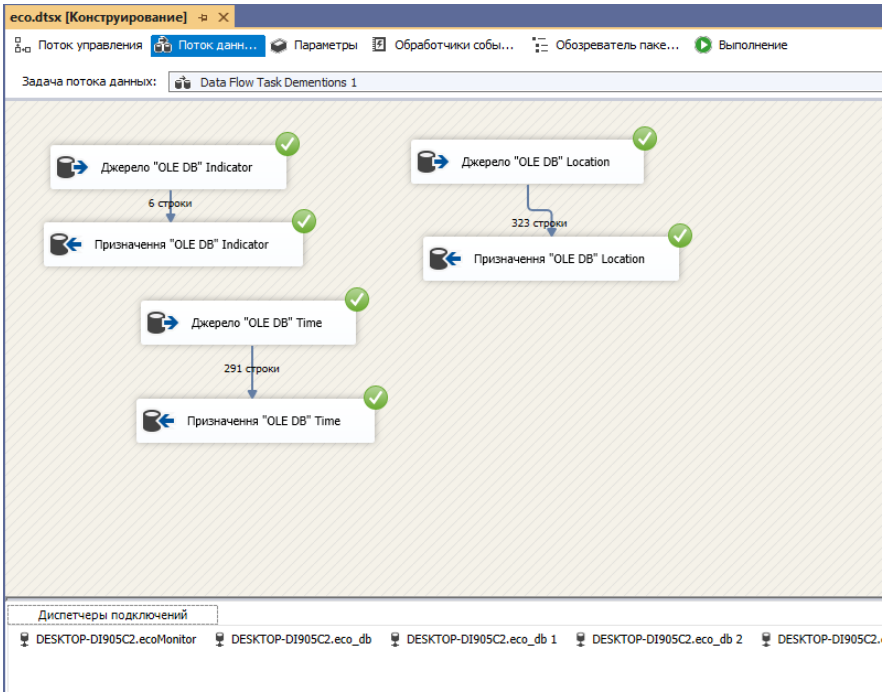

Рис. 10 Потік завдань першого рівня

На другому етапі обробляються більш складні дані, які включають додаткові категорії, такі як показники забруднення (наприклад, PM2.5, CO2 тощо). На цьому етапі дані агрегуються з урахуванням часового контексту (дата, місяць, рік) та інших ключових вимірів. На рис. 11 Зображено заповнення вимірів 2-го рівня.

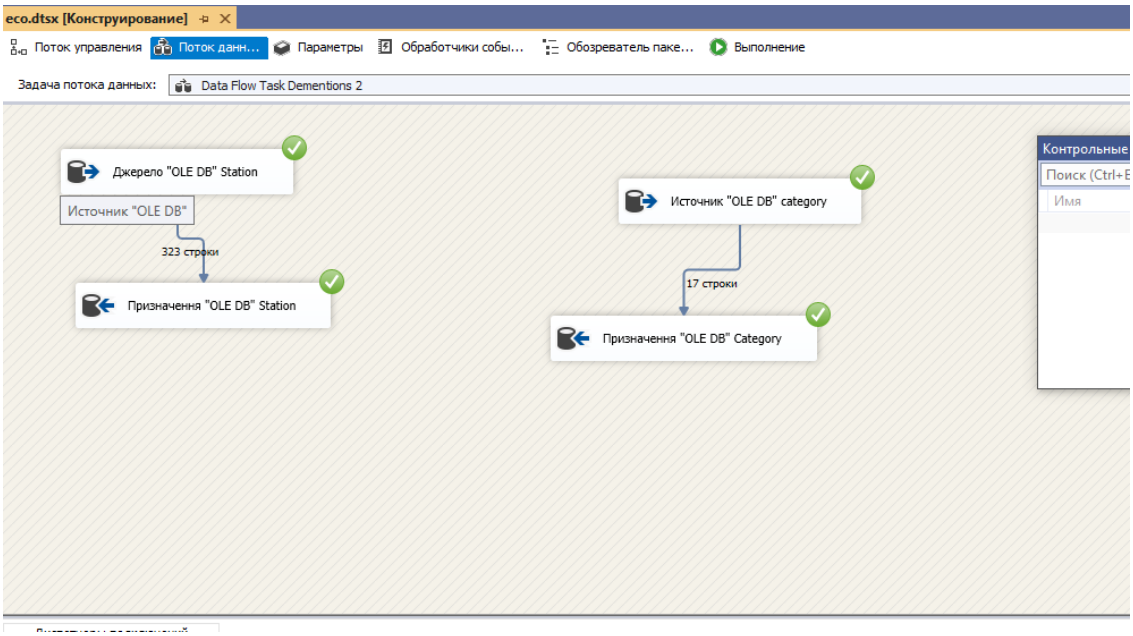

Рис. 11 Заповнення вимірів другого рівня

На останньому етапі проводиться наповнення таблиці фактів, яка містить агреговані дані про забруднення повітря за всіма станціями. Для цього використовуються результати попередніх двох етапів. Запити до оперативної БД дозволяють зібрати показники забруднення за певний період часу, які об'єднуються з відповідними даними таблиць вимірів. На рис. 12 представлено запит для агрегування даних і заповнення таблиці фактів.

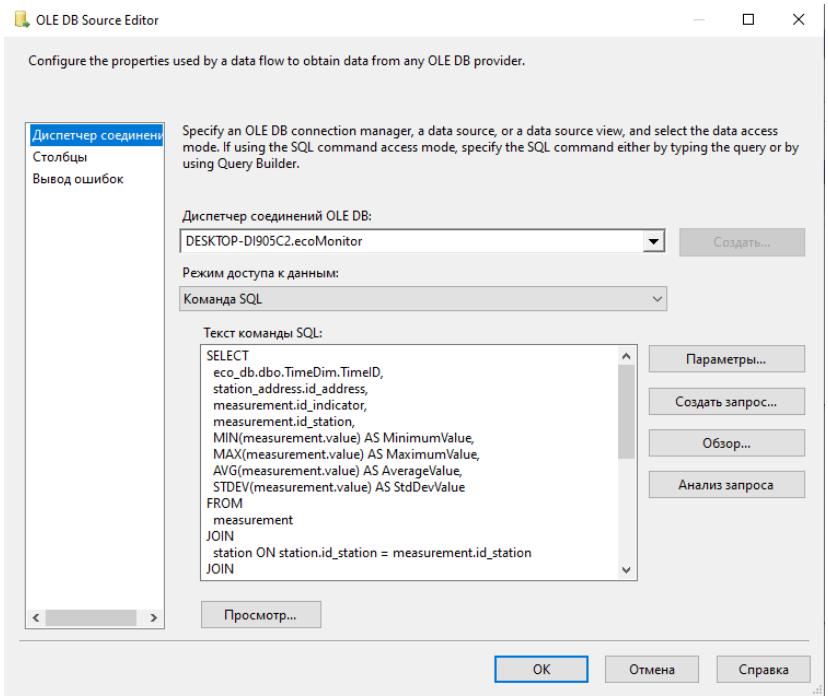

Рис. 12 Формування запиту на основі джерела оперативної БД

Запит виконує вибірку даних з кількох таблиць для обчислення ключових статистичних показників (мінімального, максимального, середнього значення та стандартного відхилення) для кожного параметра якості повітря, пов'язаного зі станціями, адресами та часовими інтервалами. Результати групуються за станціями, індикаторами, адресами та часовими ідентифікаторами (TimeID), забезпечуючи агреговану інформацію для подальшого аналізу в системі моніторингу.

На рис. 13 представлено процес передачі агрегованих даних до сховища даних за допомогою компонента OLE DB Destination Editor в середовищі SSIS. Дані, отримані в результаті виконання SQL-запиту, передаються у відповідні стовпці цільової таблиці. Вхідні поля, такі як TimeID, id_address, id_indicator, id_station, AverageValue, MaximumValue, MinimumValue та StdDevValue, зіставляються з відповідними стовпцями у цільовій таблиці сховища: TimeID, LocationID, IndicatorID, StationID, AvgValue, MaxValue, MinValue та StdDevValue.

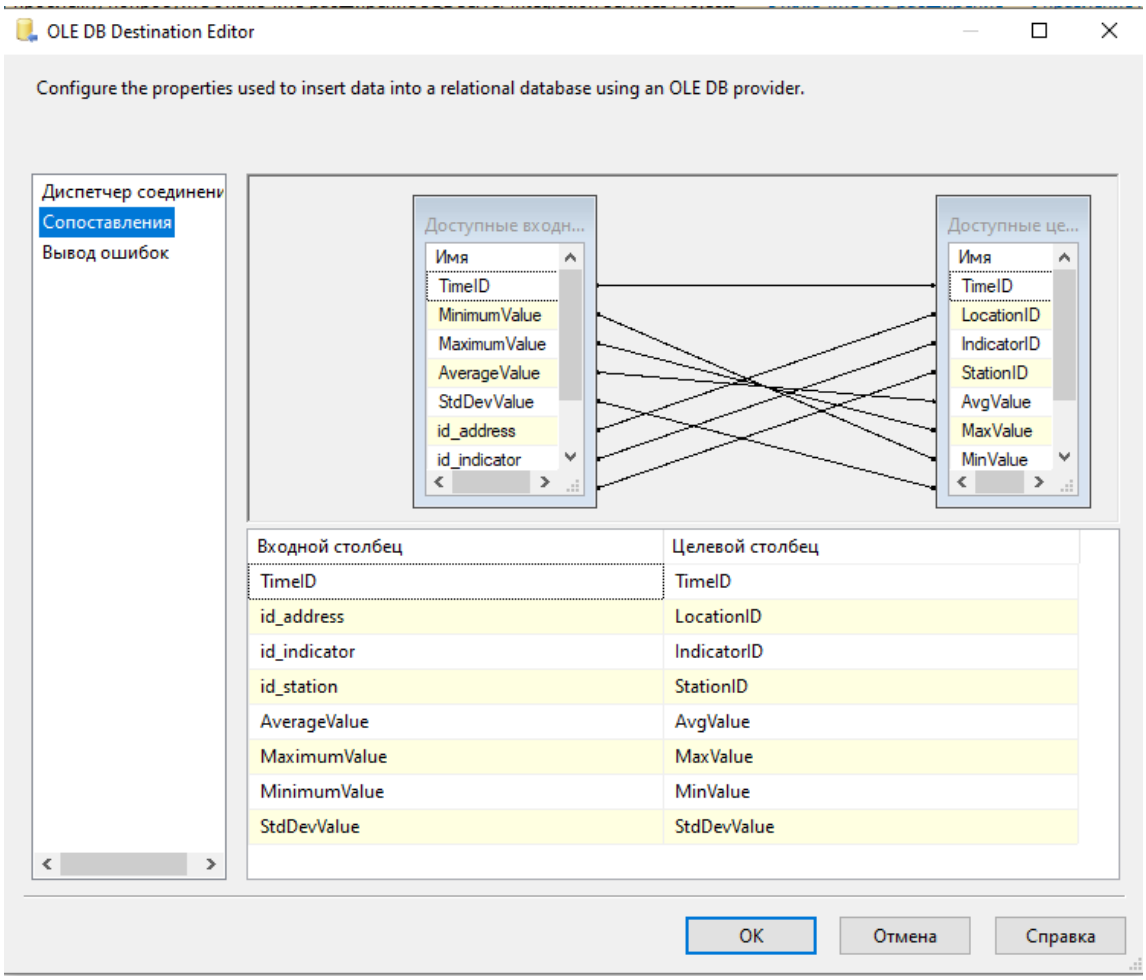

Рис. 13 Процес передачі даних на основі сформованої вибірки

3.2 Огляд методів Data Mining

У сучасних інтелектуальних системах моніторингу параметрів атмосферного повітря обробка великих обсягів даних є критично важливою для виявлення закономірностей та прогнозування екологічних тенденцій. Data Mining (видобування даних) надає інструменти та методи, які дозволяють ефективно аналізувати ці дані та підтримувати прийняття обґрунтованих рішень у сфері охорони довкілля.

Кластеризація – це процес організації об'єктів даних у набір розрізнених класів, які називаються кластерами. Кластерний аналіз є одним з основних інструментів аналізу даних в інтелектуальному аналізі даних. Алгоритм кластеризації K-середніх — це метод кластеризації з розбивкою, який розділяє дані на K груп. Алгоритм кластеризації K-середніх є більш помітним завдяки своєму інтелекту для швидкого та ефективного кластеризації великих обсягів

даних. Якість кінцевих результатів кластеризації алгоритму K-середніх залежить від випадкового вибору початкових центроїдів [39].

В моєму аналізі даних забруднення кластеризація методом K-середніх використовується для групування даних про якість повітря за схожими характеристиками. Що дозволяє виявити регіони або часові періоди з подібними рівнями забрудненості і спрощує аналіз та моніторинг екологічної ситуації. За допомогою кластеризації можна ідентифікувати зони ризику та спрямувати зусилля на покращення стану повітря в певних умовах, регіонах або періодах.

Метод асоціативних правил — це метод, запропонований Р. Агравалом та Р. Срікантом для виявлення асоціацій між даними. Алгоритм використовує пошаровий ітеративний підхід. Спочатку генерується набір кандидатів C_k , потім обчислюється підтримка всіх k-елементних наборів. Якщо підтримка перевищує мінімальний поріг підтримки, цей набір стає частим набором елементів L_k [40, 41].

На основі частого набору L_k генерується наступний набір кандидатів C_{k+1} . Далі обчислюється підтримка кандидатських наборів, і якщо вона перевищує мінімальний поріг, вони стають частими наборами L_{k+1}. Цей процес повторюється, доки не будуть знайдені всі часті набори елементів або доки не залишиться жодного нового частого набору. З елементів кожного частого набору створюються різні асоціативні правила. Ступінь довіри (confidence) кожного правила обчислюється окремо, і якщо він перевищує мінімальний поріг ступеня довіри, таке асоціативне правило вважається значущим та генерується [42].

Пошук асоціативних правил сприяє виявленню взаємозв'язків між різними параметрами забрудненості повітря. Наприклад, можна визначити, які забруднювачі часто зустрічаються разом або як певні метеорологічні умови впливають на рівень забруднення. Це допомагає зрозуміти комплексний характер екологічних проблем та розробити більш ефективні стратегії їх вирішення.

3.3 Інструментарій для аналізу даних

У процесі розробки інтелектуальної системи моніторингу параметрів атмосферного повітря використовуються різноманітні інструменти та технології, що забезпечують ефективний збір, зберігання, аналіз та візуалізацію даних. Ключовими компонентами цієї системи є Microsoft SQL Server, SQL Server Analysis Services (SSAS), SQL Server Reporting Services (SSRS), Power BI та мова програмування Python з відповідними бібліотеками для машинного навчання.

SQL Server служить основою для зберігання даних у вигляді реляційної бази даних. Він забезпечує надійне та продуктивне середовище для управління великими обсягами інформації. За допомогою SQL Server було створено сховище даних, яке дозволяє ефективно зберігати історичні дані про параметри атмосферного повітря та забезпечує швидкий доступ до них для подальшого аналізу.

SQL Server Analysis Services (SSAS) використовується для побудови багатовимірних моделей даних та реалізації OLAP (Online Analytical Processing) технологій. SSAS дозволяє створювати багатовимірні проекти, що включають куби, виміри та міри, які забезпечують швидкий та гнучкий доступ до агрегованих даних. Використання багатовимірних моделей спрощує аналіз великих обсягів інформації та підтримує складні запити користувачів [43].

У рамках SSAS також створюються структури видобування даних (Mining Structures), які дозволяють застосовувати алгоритми Data Mining для виявлення прихованих закономірностей у даних. Це включає використання методів кластеризації, класифікації та прогнозування для аналізу екологічних показників та прогнозування рівнів забрудненості повітря.

Мова програмування Python відіграє важливу роль у побудові прогнозних моделей та реалізації алгоритмів машинного навчання. З використанням бібліотек, таких як scikit-learn, pandas та numpy, можна ефективно обробляти великі набори даних, виконувати статистичний аналіз та будувати моделі прогнозування. Зокрема, алгоритми рандомного лісу (Random Forest) та

градієнтного бустінгу (Gradient Boosting) використовуються для прогнозування концентрацій забруднювачів та оцінки екологічних ризиків.

Для розрахунку ключових показників ефективності (KPI) використовуються можливості як SSAS. KPI дозволяють кількісно оцінити рівень забрудненості, динаміку змін екологічних параметрів та ефективність впроваджених заходів щодо поліпшення якості повітря. Це сприяє прийняттю обґрунтованих управлінських рішень та плануванню екологічних програм.

SSAS надає наступні основні методи аналізу даних, що зображено нижче.

1. Дерева рішень (Decision Trees): для класифікації та прогнозування.
2. Кластеризація (Clustering): для групування подібних даних.
3. Наївний Баєсівський класифікатор (Naive Bayes): для швидкої класифікації.
4. Асоціативні правила (Association Rules): для виявлення взаємозв'язків між змінними.
5. Нейронні мережі (Neural Networks): для моделювання складних патернів.
6. Аналіз часових рядів (Time Series Analysis): для прогнозування на основі історичних даних.
7. Регресія (Regression): для аналізу залежностей між змінними.

SQL Server Reporting Services (SSRS) та Power BI застосовуються для створення звітів та візуалізації даних. SSRS дозволяє генерувати різноманітні звіти, включаючи табличні, матричні та графічні представлення даних, які можуть бути опубліковані та доступні через веб-портал або інтегровані в інші додатки [44]. Power BI надає інтерактивні можливості для візуалізації даних, створення дашбордів та спільної роботи над звітами. Використання Power BI спрощує аналіз даних та робить його доступним для широкого кола користувачів без глибоких технічних знань.

OLAP-технології забезпечують багатовимірний аналіз даних, дозволяючи користувачам швидко отримувати відповіді на складні бізнес-запити. Вони підтримують операції згортання та розгортання даних за різними вимірами,

такими як час, місце, тип забруднювача тощо. Це особливо важливо для аналізу екологічних даних, де необхідно враховувати багато факторів та їх взаємодію.

У процесі побудови сховища даних та аналітичних моделей важливим є забезпечення якості даних. Для цього використовуються методи очищення, нормалізації та перетворення даних як у SQL Server, так і за допомогою Python. Це включає виявлення та обробку пропущених значень, видалення аномалій та перевірку консистентності даних.

Використання інтегрованого підходу, що поєднує можливості SQL Server, SSAS, SSRS, Power BI та Python, дозволило створити інтелектуальну систему моніторингу. Така система забезпечує повний цикл роботи з даними: від їх збору та зберігання до аналізу та візуалізації результатів. Це сприяє підвищенню ефективності екологічного моніторингу та підтримує прийняття стратегічних рішень у сфері охорони довкілля.

Загалом, використання сучасних інструментів для аналізу даних дозволяє отримати глибоке розуміння екологічних процесів, виявити приховані закономірності та тенденції, а також розробити ефективні заходи для покращення якості атмосферного повітря.

3.4 Дані для аналізу

У рамках розробки системи використовувалися дані, отримані з двох основних джерел. Першим джерелом є офіційний сайт автоматизованої системи моніторингу повітря Києва (https://asm.kyivcity.gov.ua/), де представлені дані зі станцій Київської міської державної адміністрації (КМДА). Другим джерелом виступає відкритий API SaveEcoBot, який агрегує дані зі станцій по всій Україні. Дані з сайту КМДА охоплюють широкий спектр показників якості повітря. Основні з них представлено нижче.

- Вологість (*Humidity*, %);
- Тиск (*Pressure*, hPa);
- Температура (*Temperature*, °C);
- Тверді частинки різних фракцій:

    - PM1 (мкг/м³);
    - PM2.5 (мкг/м³);
    - PM4 (мкг/м³);
    - PM10 (мкг/м³);
- Вуглекислий газ (*Carbon dioxide*, CO2, мкг/м³);
- Леткі органічні сполуки (*Volatile Organic Compounds*, VOC, ppb);
- Індекс якості повітря (*Air Quality Index*);
- Діоксид азоту (*Nitrogen dioxide*, NO2, мкг/м³);
- Монооксид вуглецю (*Carbon monoxide*, CO, мг/м³);
- Діоксид сірки (*Sulfur dioxide*, SO2, мкг/м³);
- Радіаційний фон (*Radiation*, мкЗв/год);
- Бензол (*Benzene*, C6H6, мкг/м³);
- Озон (*Ozone*, O3, мкг/м³);
- Напрямок вітру (*Wind Direction*, градуси);
- Швидкість вітру (*Wind Speed*, м/с);
- Сірководень (*Hydrogen sulfide*, H2S, мкг/м³);
- Формальдегід (*Formaldehyde*, CH2O, мкг/м³);
- Загальний індекс якості повітря (*Common Air Quality Index*);

Проте хоча ці дані досить детальні, кількість станцій КМДА є обмеженою, що може впливати на просторову репрезентативність аналізу.

Для розширення географічного покриття було використано дані з відкритого API SaveEcoBot, який збирає інформацію зі станцій по всій Україні. Цей ресурс надає дані про показники, що зазначені нижче.

- Тверді частинки (PM2.5, PM10, мкг/м³);
- Вологість (%);
- Температура (°C);
- Індекс якості повітря (*Air Quality Index*);

Хоча набір показників SaveEcoBot є менш розширеним, він забезпечує широку географічну представленість, що важливо для аналізу загальноукраїнських тенденцій якості повітря.

Для ефективного зберігання та обробки даних було створено ряд ключових таблиць:

- Станції: містить інформацію про станції моніторингу, включаючи їх унікальні ідентифікатори, назви, географічні координати та інші характеристики.
- Показники: зберігає дані про різні показники якості повітря, їх одиниці виміру та категорії забруднення.
- Виміри: фактичні вимірювання показників повітря, пов'язані зі станціями та часом вимірювання.
- Категорії забруднення: класифікація рівнів забруднення згідно з нормативними або міжнародними стандартами.
- Оптимальні значення: містить нормативні значення або граничні допустимі концентрації для кожного показника, що дозволяє оцінювати відповідність фактичних даних стандартам.

Зібрані дані проходили попередню обробку, яка включала:

- Очищення даних: видалення або корекція помилкових та неповних записів.
- Перевірка консистентності: забезпечення узгодженості даних між різними таблицями та джерелами.
- Нормалізація даних: приведення даних до єдиного формату та одиниць виміру.
- Заповнення пропущених значень: використання статистичних методів або алгоритмів машинного навчання для прогнозування відсутніх даних таких як заповнення середніми значеннями. Що дозволило зберігати стабільність даних.

4 РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

4.1 Дослідження використання KPI

У рамках розробленої системи було визначено ключові показники ефективності (KPI), які відображають стан повітря за такими параметрами, як концентрація шкідливих речовин, рівень забруднюючих частинок і погодні умови. Ці показники дозволяють оцінити загальну динаміку змін та виявляти потенційні загрози, забезпечуючи основу для прийняття управлінських рішень і планування стратегій зменшення рівня забруднення.

Кожен із показників відображає середнє або максимальне значення забруднення в розрізі показника та часу. Для аналізу фактів, занесених у куб, було визначено такі KPI:

1. KPI_co – показник середньої концентрації оксиду вуглецю (CO), який відображає рівень цього газу в атмосфері. Його підвищений рівень може вказувати на шкідливий вплив автомобільного транспорту або промислових викидів.

2. KPI_no2 – середнє значення оксиду азоту (NO2), який є важливим параметром для оцінки якості повітря. Підвищений рівень NO2 може негативно впливати на здоров'я людини та довкілля.

3. KPI_o3 – показник середньої концентрації озону (O3), який є критичним компонентом для оцінки фотохімічного смогу. Небезпечний рівень озону може свідчити про надмірне забруднення атмосфери.

4. KPI_pm10 – середня концентрація твердих частинок розміром до 10 мкм (PM10), які можуть впливати на дихальну систему та загальний стан здоров’я.

5. KPI_pm2_5 – показник середньої концентрації дрібнодисперсних частинок розміром до 2,5 мкм (PM2.5), які мають значний вплив на здоров'я, проникаючи глибоко у легені.

6. KPI_so2 – середня концентрація діоксиду сірки (SO2), що може вказувати на вплив промислових викидів та спалювання палива.

Кожен з цих показників відображає певні аспекти стану повітря і разом вони допомагають сформувати повну картину якості атмосферного середовища. Значення показників порівнюються із встановленими цілями, щоб визначити, чи параметри знаходяться у задовільному стані, або ж їх потрібно покращувати. Тренди (зміни показників з часом) та загальні оцінки (у формі дашбордів чи індикаторів) дають змогу оцінити динаміку змін та розробити відповідні стратегії реагування чи вдосконалення.

Для побудови ключових показників ефективності (KPI) у середовищі SSAS було проведено низку дій для визначення фактичного, цільового значення, а також статусу й тренду кожного показника.

На рис. 14 представлено загальний огляд створення KPI. Тут налаштовано назву KPI (наприклад, KPI_pm2_5), а також визначено, що цей показник пов’язаний із відповідною групою вимірів. Визначення значення KPI здійснюється через формулу, яка розраховує середнє значення показника PM2.5 за допомогою агрегування даних із фактичного кубу. Формула для обчислення базується на доступних вимірах часу, локації та значення параметра.

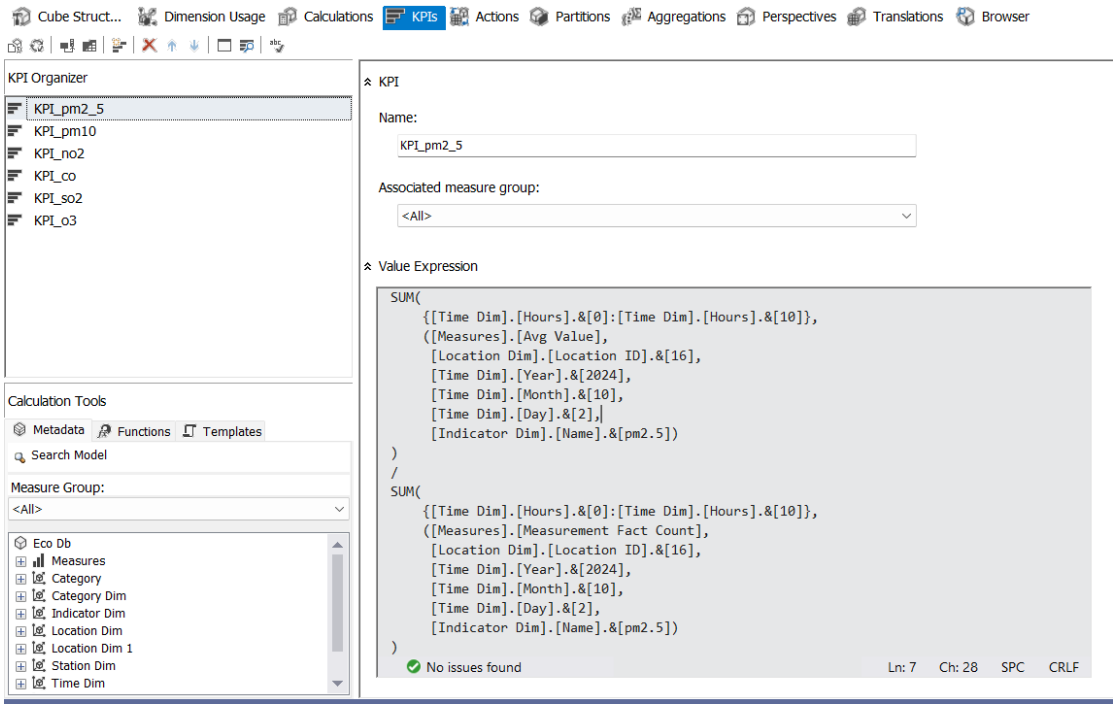

Рис. 14 Створення KPI

На рис. 15 показано, як налаштовується статус і тренд KPI. Для статусу використовуються умови, які порівнюють фактичне значення з цільовим. Наприклад:

- Якщо значення KPI дорівнює або перевищує цільове значення, статус вважається позитивним (показується стрілка вгору).
- Якщо значення менше цільового, статус вважається негативним (стрілка вниз).

Тренд визначається шляхом обчислення змін фактичного значення у часовій перспективі. Якщо середнє значення PM2.5 за поточний період зростає порівняно з попереднім, тренд позначається як негативний.

Також на рисунку наведено формули, які реалізують ці розрахунки. Формули використовують MDX-запити для доступу до даних куба, щоб визначити середнє значення (Avg Value) для відповідного параметра, цільове значення (Target Value) і кількість записів для вимірювання точності даних.

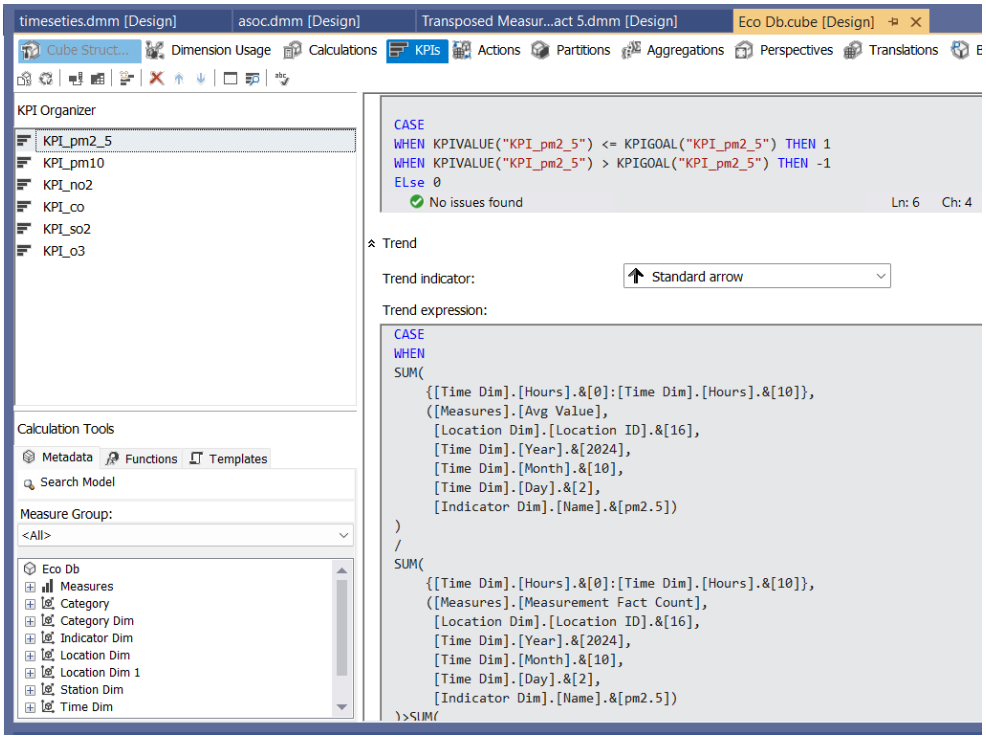

Рис. 15 Налаштування статусу і тренду KPI

У результаті кожен KPI був налаштований таким чином:

- Фактичне значення: середнє або максимальне значення параметра за період.
- Цільове значення: допустимі рівні для кожного показника, встановлені відповідно до екологічних стандартів.
- Статус: відповідність фактичного значення цільовому.
- Тренд: аналіз змін показника у часовій перспективі.

Зображення KPI представлено на рис. 16.

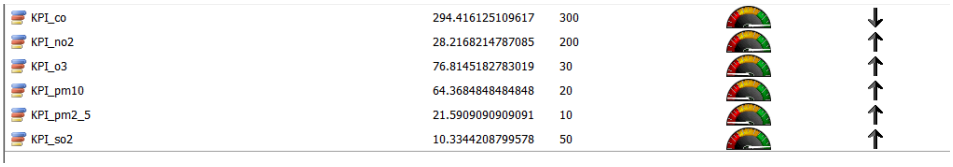

Рис. 16 Обраховані KPI

На основі представлених ключових показників ефективності (KPI) можна зробити висновок, що деякі з параметрів, зокрема концентрації CO, NO_2, PM10 і PM2.5, перевищують допустимі межі, що вказує на високий рівень забруднення повітря. Наприклад, рівень CO значно наближається до граничного значення, а PM10 і PM2.5 перевищують допустимі норми, вказуючи на потенційну небезпеку для здоров'я. Водночас, рівні озону (O_3) та діоксиду сірки (SO_2) залишаються в межах безпечних показників. Ці результати показують що є необхідність вжиття заходів для зниження концентрацій шкідливих речовин у повітрі.

4.2 Аналіз і звітність за показниками забруднення

В ході дослідження використовувались такі засоби візуалізації як SQL Server Reporting Services та Power Bi. Нижче наведено звіти візуалізовані за допомогою цих сервісів.

Середні значення показників в розрізі по вказаному місту

Після створення сховища даних у рамках магістерської роботи було використано інструмент Power BI для візуалізації та аналізу даних. Основна мета — оцінити середні рівні забруднення повітря (PM2.5) в різних містах, що

дозволяє виявити найбільш проблемні регіони та планувати заходи для покращення якості повітря.

Аналіз середніх значень PM2.5 показав значні розбіжності між різними містами. Найвищий середній рівень зафіксовано в місті Вараш (204.44), що може бути наслідком інтенсивної промислової діяльності або інших джерел забруднення. Також високі показники PM2.5 були зареєстровані в містах Новопокровка (141.1) та Постомиль (138.9), які потребують подальшого вивчення джерел забруднення.

Міста, такі як Сонячне (67.36) та Ворохта (50.63), демонструють відносно низькі середні рівні PM2.5, що може свідчити про меншу присутність джерел забруднення або сприятливі природні умови для очищення повітря. Варто зазначити, що середній рівень PM2.5 навіть у цих містах може перевищувати рекомендовані норми ВООЗ, що вимагає постійного моніторингу.

Графік забруднення дрібнодисперсними частками в розрізі міст зображено на рис. 17

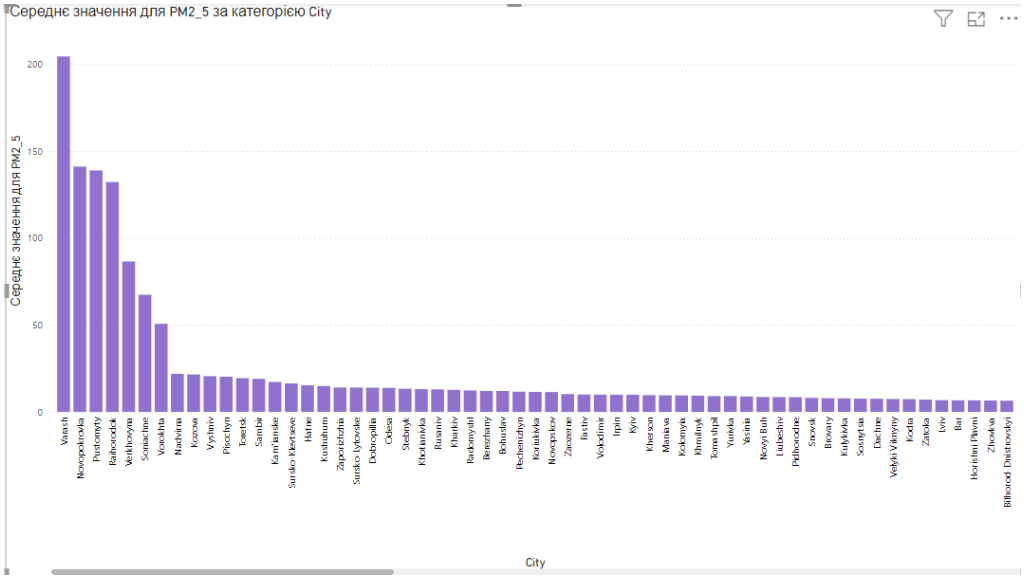

Рис 17. Звіт забруднення в розрізі географічного розташування

Забруднення в розрізі місяця по місту за показником

Продовжуючи аналіз, за допомогою Power BI було побудовано звіт, що відображає середньомісячні значення забруднення повітря для чотирьох основних показників: CAQI, CO, PM2.5 та O3. Цей аналіз дозволив

ідентифікувати сезонні зміни у рівнях забруднення для кожного параметра. На рис. 18 наведено графіки забруднення різних показників в розрізі місяців.

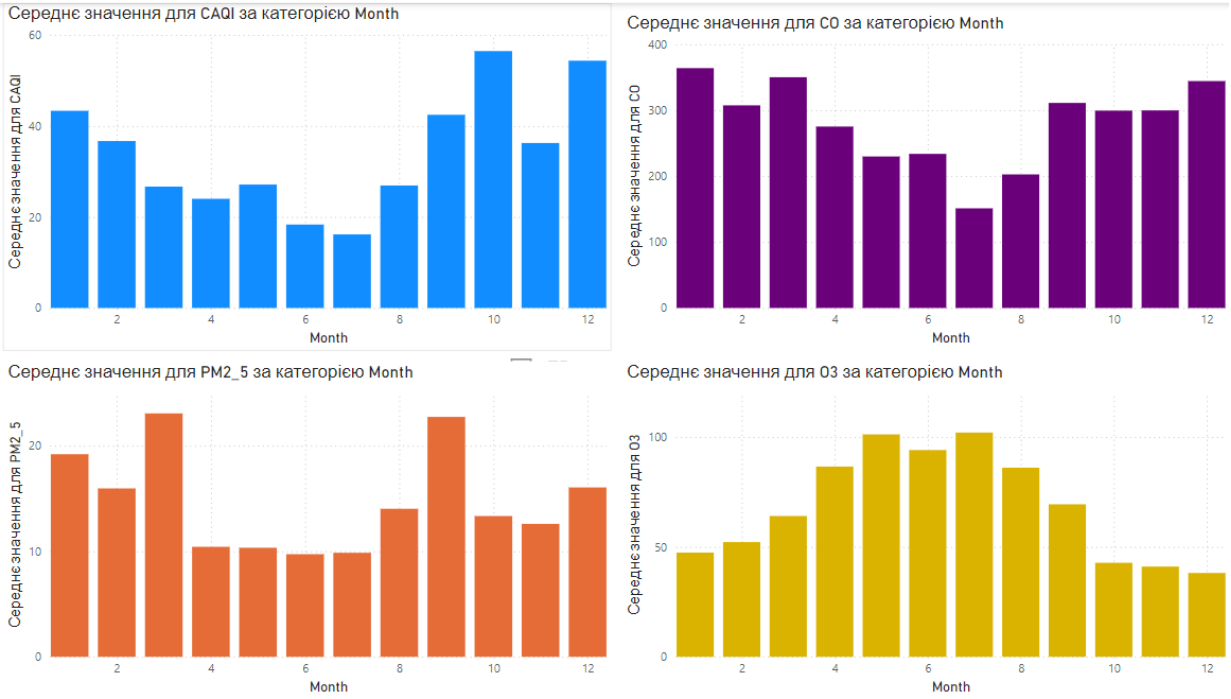

Рис 18. Звіт по показникам в розрізі місяців

CAQI (Загальний індекс якості повітря): динаміка CAQI свідчить про значну сезонну варіативність. Найнижчі середні значення зафіксовані влітку, особливо у червні (18,33) та липні (16,17). У холодні місяці, такі як жовтень (56,45) та грудень (54,34), показники значно вищі, що може свідчити про вплив опалювального сезону та збільшення концентрації забруднюючих речовин.

CO (Оксид вуглецю): рівні CO демонструють найбільше значення у січні (363,74) та поступове зниження до липня (151,11). У літні місяці спостерігається найменше забруднення, ймовірно, через покращення розсіювання газів у теплу погоду. У грудні рівень CO знову зростає до 344,12, підтверджуючи вплив опалення та зимових умов.

PM2.5 (Дрібнодисперсні частинки): PM2.5 характеризується найвищими значеннями у січні (19,19) та березні (23,07), що також можна пов’язати з зимовими умовами та підвищеним використанням палива. У літні місяці значення падають до мінімуму в червні (9,72) та липні (9,85), що підтверджує сезонне зниження концентрації твердих частинок.

O3 (Озон): Концентрація озону демонструє протилежну тенденцію порівняно з іншими параметрами. Найвищі рівні спостерігаються влітку — у травні (101,16) та липні (101,99), що пов’язано зі збільшенням сонячного випромінювання та фотохімічних реакцій. Взимку ж концентрація озону знижується до найнижчих значень у грудні (38,13).

Отримані результати підтверджують необхідність врахування сезонних змін у процесі моделювання та прогнозування забруднення повітря.

Забруднення в розрізі категорії та часу по місту за показником

Для аналізу рівня забруднення повітря в Києві за 2024 рік було побудовано кругову діаграму, яка відображає розподіл суми концентрацій дрібнодисперсних частинок PM2.5 між двома категоріями: "good" (якісне повітря) та "polluted" (забруднене повітря).

Результати аналізу показують, що більшість вимірювань належать до категорії "polluted", яка охоплює 85,07% загального обсяг вимірів. Це свідчить про домінування умов із забрудненим повітрям протягом року. Натомість категорія "good", яка відповідає більш сприятливим умовам, становить лише 14,93% загальної кількості вимірів.

Виявлена диспропорція між двома категоріями також свідчить про актуальність подальшого аналізу причин високих рівнів забруднення та можливих шляхів їх зменшення. На рис. 19 зображено кругову діаграму забруднення дрібнодисперсними частками в Києві в розрізі категорії забруднення і часу забруднення на протязі року.

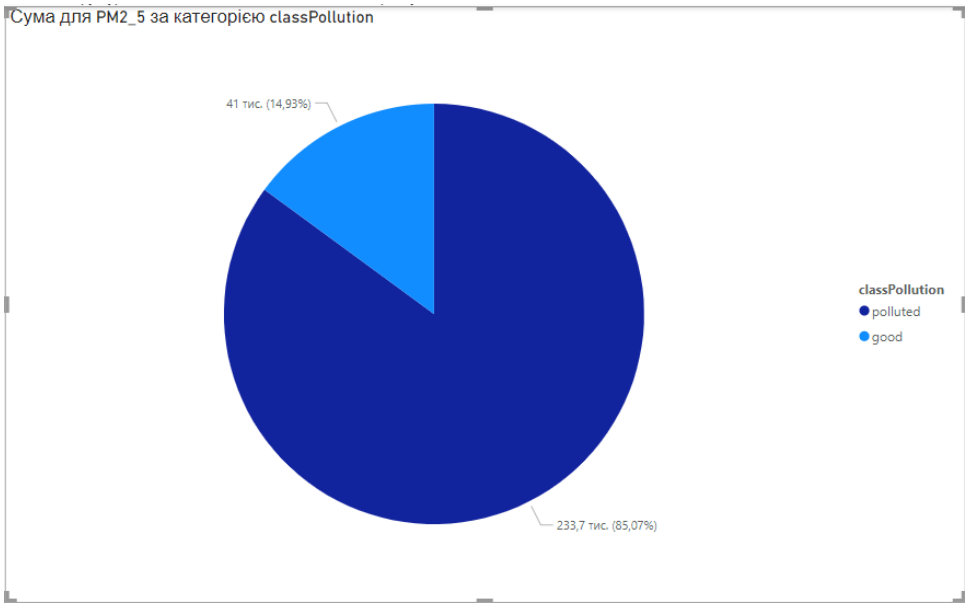

Рис. 19 Діаграма забруднення дрібнодисперсними частинками в розрізі часу

Тренд за показником

Візуалізація трендових змін середніх значень концентрацій забруднюючих речовин (PM2.5, CO, SO2 та O3) за період 2023–2024 років надає важливу інформацію про динаміку якості повітря. Усі графіки відображають середні значення в розрізі дня, місяця, кварталу та року, що дозволяє детально оцінити сезонні та довгострокові тенденції. На рис. 20 зображені тренди забруднення по різним показникам протягом 2023-2024 років.

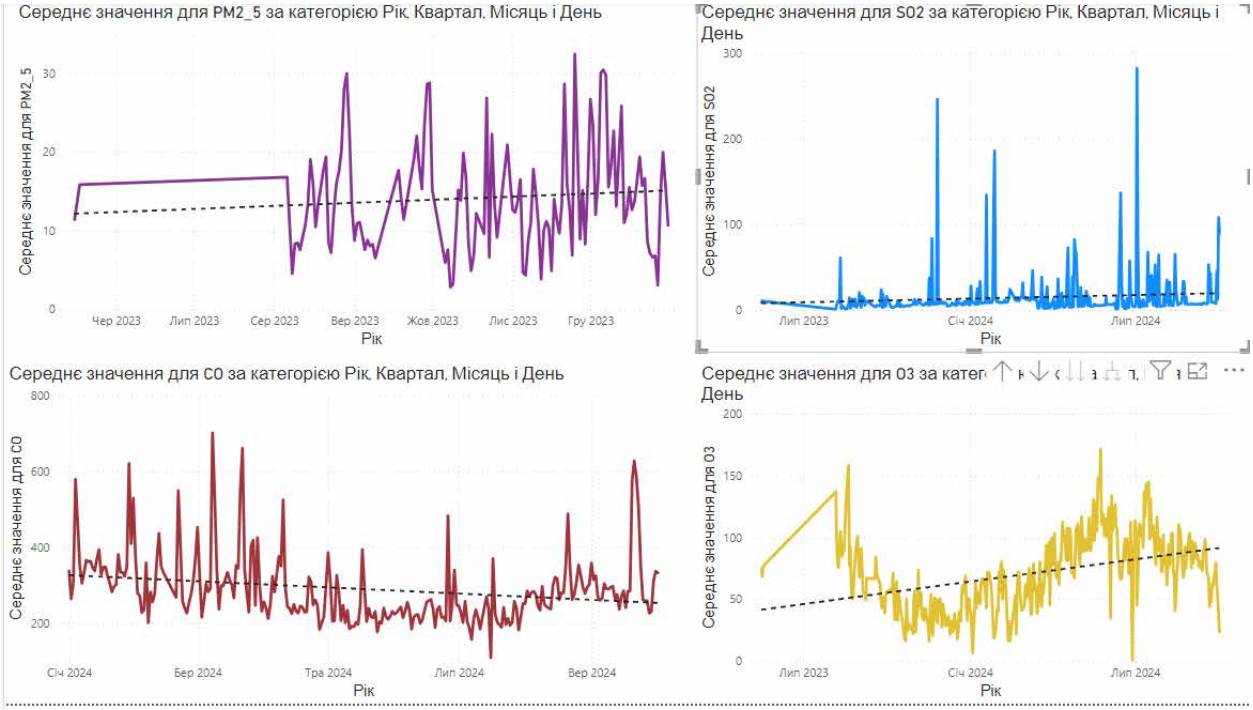

Рис 30. Тренди по показникам

1. **PM2.5.** Графік показує нестабільну динаміку концентрацій PM2.5 із помітними піками в осінньо-зимовий період. Це може бути пов'язано із сезонним збільшенням викидів, наприклад, через опалення. У той же час, наявність висхідної трендової лінії свідчить про можливе погіршення загальної ситуації із забрудненням дрібнодисперсними частками.

2. **SO2.** Графік SO2 демонструє значні піки, які виникають в окремі дні, що може бути наслідком специфічних джерел забруднення, таких як промислові викиди або погодні умови. Тренд залишається стабільним, але часті сплески потребують глибшого аналізу для виявлення причин.

3. **CO.** Концентрації CO також демонструють коливання, із періодами значного зростання. Загальний тренд, однак, має тенденцію до поступового зниження. Це може свідчити про певні успіхи у зменшенні викидів, наприклад, через модернізацію транспортних засобів або зміну джерел енергії.

4. **O3.** Концентрації озону демонструють поступове збільшення протягом періоду спостереження, із сезонними коливаннями. Це може бути результатом фотохімічних реакцій в атмосфері, які посилюються в теплий період року.

Висновки.

1. Графіки ілюструють суттєві сезонні коливання у всіх показниках, що є типовим для процесів забруднення повітря.

2. Підвищення концентрацій PM2.5 та O3 потребує подальшого дослідження, зокрема з урахуванням можливих антропогенних і природних факторів.

3. CO демонструє позитивну тенденцію до зниження, що є ознакою ефективності заходів із поліпшення якості повітря.

4.3 Дослідження застосування методів кластеризації

У процесі дослідження для аналізу даних було застосовано методи кластеризації, які дозволяють виявляти приховані закономірності у великих

наборах даних. Зокрема, використовувався алгоритм K-Means, який є одним із найбільш популярних і простих методів кластеризації. Цей метод дозволяє розділити вибірку даних на кластери, кожен з яких характеризується схожими ознаками, ідентифікованими на основі мінімізації відстаней до центроїдів.

Дані для кластеризації були отримані з сховища даних, яке агрегувало інформацію про параметри якості повітря, зокрема температура, вологість, CO, PM 2,5 та O3. Підготовка даних включала такі етапи:

- заповнення пропущених значень шляхом обчислення середнього значення для кожного параметра;
- масштабування даних із використанням алгоритму StandardScaler, щоб привести всі ознаки до однакового діапазону і запобігти домінуванню параметрів із великими значеннями.

Для визначення оптимальної кількості кластерів було використано метод ліктя. Графік (рис. 1) показує зміну інерції (сумарної відстані від точок до їх центрів кластерів) у залежності від кількості кластерів k. Після аналізу графіка було встановлено, що оптимальною є кількість кластерів k=3, оскільки подальше збільшення кількості кластерів не призводить до значного зменшення інерції.

На рис. 31 Зображено візуалізацію графіка до метода ліктя при відборі оптимальнох кількості кластерів.

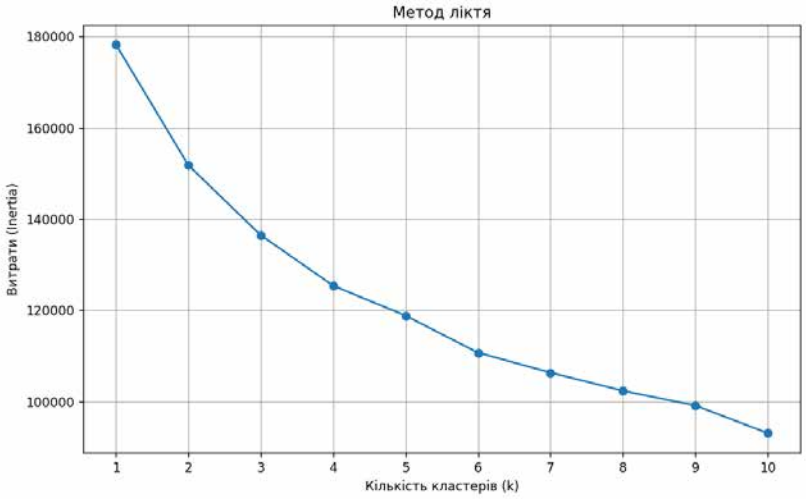

Рис 31. Метод ліктя

Після застосування алгоритму K-Means для кластеризації було виділено три основні кластери:

1. Cluster 0 – цей кластер характеризується найвищою вологістю (79.55%) та помірним рівнем озону (55.37 мкг/м³) і CO (259.66 мкг/м³).

2. Cluster 1 – для цього кластеру характерні найнижчі значення озону (36.68 мкг/м³) і температура (12.39°C), проте виявлено високий рівень вуглецю (CO) – 498.79 мкг/м³.

3. Cluster 2 – характеризується високими середніми значеннями температури (22.39°C) та концентрації озону (98.91 мкг/м³), при низькій вологості (48.39%) і PM2_5 (11.62 мкг/м³).

Результати кластеризації дозволили визначити закономірності між параметрами, такими як озон, температура, вологість і концентрація забруднювальних речовин.

Для оцінки результатів кластеризації було побудовано кілька графіків, які демонструють особливості та характеристики кожного кластеру.

Перший графік (рис. 1) є тривимірною візуалізацією, де використано такі параметри, як температура (Temperature), вологість (Humidity) та озон (O3).

На цьому графіку можемо бачити три окремих кластери, що були виділені алгоритмом кластеризації. Кластери позначені різними кольорами, а їхні назви відображають основні характеристики, такі як рівень вологості чи температури в поєднанні зі значеннями озону. Наприклад, один із кластерів характеризується високою температурою та підвищеним рівнем озону, інший – високою вологістю при низькому рівні озону. На рис. 32 Зображено тривимірне представлення кластерів.

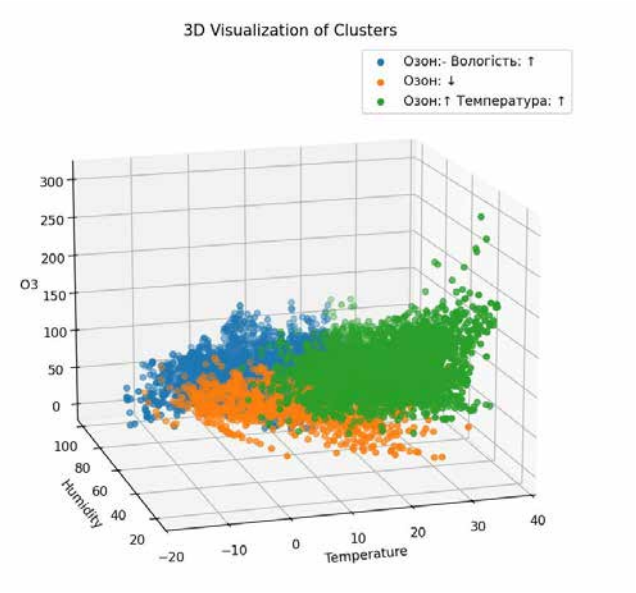

Рис. 32 Тривимірне представлення кластерів

Другий графік (рис. 33) демонструє розподіл кластерів у двовимірному просторі на основі параметрів вуглекислий газ (CO) та озон (O3). Цей графік дозволяє візуально оцінити залежність між викидами CO та концентрацією O3 для кожного кластеру. На графіку видно, що деякі кластери мають чіткі межі, зокрема, один із кластерів характеризується високими значеннями CO та низьким рівнем озону.

Третій графік (рис. 34) відображає взаємозв'язок між PM2_5 та O3 у межах кластерів. Цей графік надає можливість аналізувати розподіл дрібнодисперсних часток (PM2_5) залежно від рівня озону, що є важливим для розуміння впливу забруднення повітря на різні території.

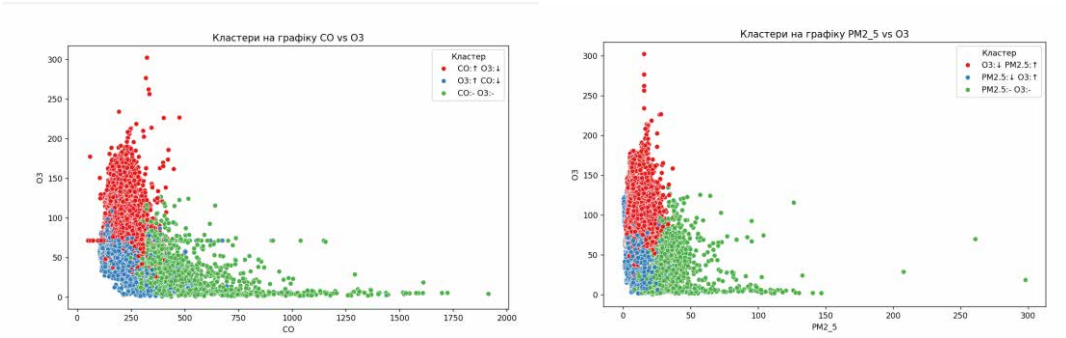

Рис. 33-34 Представлення кластерів в двохмірному розрізі декількох параметрів

4.4 Дослідження використання методу асоціативних правил

Метод асоціативних правил є інструментом аналізу даних, який дозволяє виявляти приховані залежності між змінними у великих наборах даних. У межах даного дослідження асоціативні правила було застосовано для аналізу залежностей між ключовими параметрами якості повітря, такими як концентрація PM2.5, CO, O3, тиск, SO2 та іншими показниками, що визначають стан атмосферного повітря. Метою цього дослідження було виявлення взаємозв'язків між показниками та їх вплив на рівень PM2.5 — одного з найважливіших параметрів, що впливають на здоров’я населення.

Метод асоціативних правил було реалізовано за допомогою інструментів Microsoft SQL Server Analysis Services (SSAS), що забезпечує ефективний процес обробки даних та зручні інтерфейси для візуалізації результатів. Дані, що використовувались, були попередньо підготовлені у сховищі даних та агреговані у формат, зручний для аналізу.

Ключовим завданням було виявити правила, які допомагають пояснити залежність концентрації PM2.5 від інших параметрів, наприклад, рівня CO або озону (O3). Правила мають форму: "Якщо параметр X знаходиться в діапазоні A-B, то параметр Y буде мати значення Z". Це дозволяє використовувати знайдені залежності для прогнозування рівня забруднення на основі поточних або минулих показників.

Застосування методу асоціативних правил дозволило отримати наступні значущі результати:

Основні правила. Наприклад:

- Якщо CO знаходиться в діапазоні від 715 до 1338, то PM2.5 ≥ 37.21 (ймовірність 56.2%, важливість 1.74).
- Якщо CO = 715-1338 і озон < 10.65, то PM2.5 ≥ 37.21 (ймовірність 55.9%, важливість 1.73).
- Якщо CO = 715-1338 і тиск ≥ 1001.32, то PM2.5 ≥ 37.21 (ймовірність 56.2%, важливість 1.72).

Залежності з нижчою ймовірністю. Виявлено менш сильні залежності, наприклад:

- Якщо CO знаходиться в діапазоні від 289 до 715, то PM2.5 = 22.77-37.21 (ймовірність 30.6%, важливість 1.21).
- Якщо CO = 289-715 та SO2 < 7.49, то PM2.5 ≥ 37.21 (ймовірність 16.3%, важливість 1.20).

Взаємозв’язок між параметрами. Наприклад, тиск і концентрація SO2 мають помітний вплив на концентрацію PM2.5 у повітрі. Залежності, які включають місяць або час доби, демонструють вплив сезонності та часу на концентрацію PM2.5.

Отримані правила дозволяють виявити ключові фактори, що впливають на якість повітря. Наприклад, високий рівень CO та низький рівень озону є сильними індикаторами підвищеної концентрації PM2.5. Такі знання можуть бути корисними для раннього попередження населення про небезпечні рівні забруднення, а також для розробки заходів, спрямованих на зменшення викидів.

На рис. 35 представлено побудову асоціативних правил в SSAS.

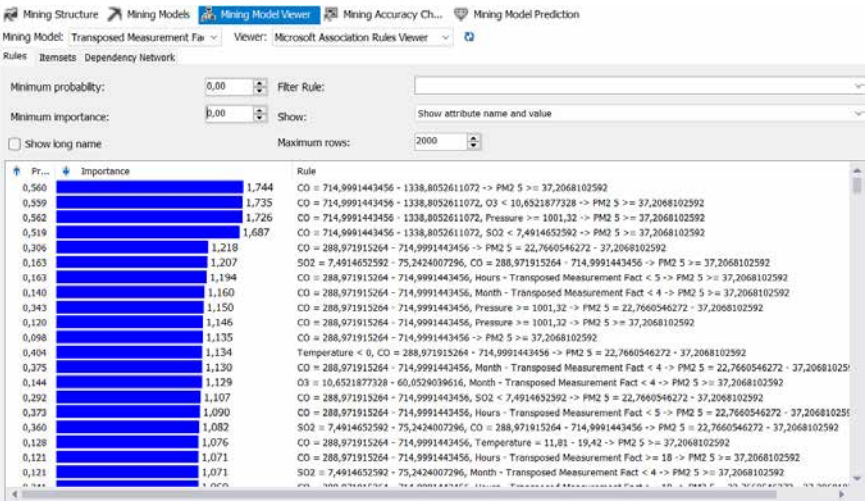

Рис. 35 Побудовані асоціативні правила

Графік взаємозв’язків демонструє ключові залежності між змінними та правилами. Зокрема, найбільш сильні зв’язки свідчать про тісний взаємозв’язок між концентраціями вуглекислого газу (CO), озону (O3), температурою та показниками забруднення PM2.5. Наприклад, зв'язки між CO і PM2.5 вказують

на те, що підвищення концентрації CO найчастіше супроводжується збільшенням PM2.5. Це підтверджує висновки з таблиці асоціативних правил, які показують, що CO у діапазоні від 714 до 1338 однозначно впливає на PM2.5 (ймовірність 56%).

Також можна відзначити, що взаємозв'язки між змінними, такими як тиск (Pressure) і PM2.5, вказують на опосередковані фактори, що впливають на забруднення. Це відкриває можливість для більш детального аналізу зовнішніх чинників. На рис. 36 представлено діаграму залежностей для асоціативних правил.

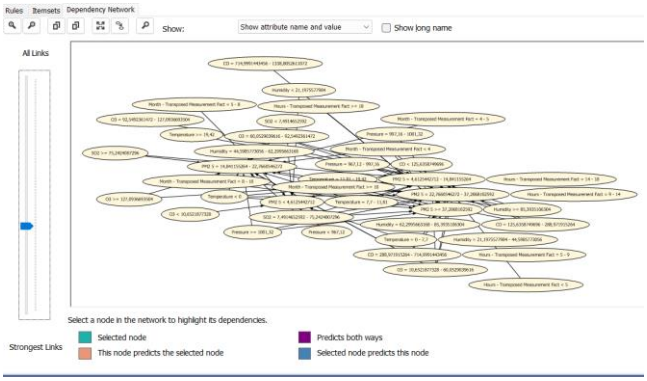

Рис. 36 Діаграма залежностей для асоціативних правил

На графіку Lift Chart ми спостерігаємо відхилення кривої моделі від ідеальної. Відповідно до графіка, модель, побудована за допомогою асоціативних правил, досягає рівня точності 74% для верхньої половини популяції. Це значно перевищує ідеальний випадковий розподіл (50%) і підтверджує ефективність побудованих правил для прогнозування високих значень PM2.5.

Показник точності підтверджує, що модель здатна ефективно передбачати залежності між змінними, такими як CO, O3, і PM2.5. Це також вказує на необхідність подальшого вдосконалення моделі, зокрема уточнення вибору змінних для аналізу. На рис. 37 представлено графік оцінки точності побудованих асоціативних правил.

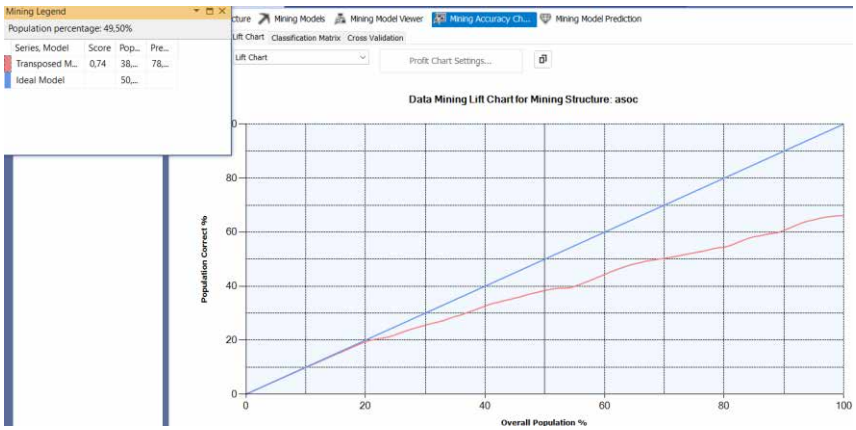

Рис. 37 Графік оцінки точності побудованих асоціативних правил

Результати свідчать про те, що метод асоціативних правил є ефективним для визначення основних закономірностей у даних, а також для створення прогнозних моделей. Однак, як показують графіки, залежності між змінними є багатофакторними, тому точність прогнозів можна покращити за рахунок включення більшого числа зовнішніх факторів.

Це дослідження є важливим етапом у системі моніторингу повітря, оскільки дозволяє не лише ідентифікувати закономірності, а й виділити критичні фактори, що впливають на забруднення.

4.5 Прогнозування показників забруднення за допомогою методів машинного навчання

Інтелектуальна система моніторингу атмосферного повітря, побудована в рамках дослідження, поєднує в собі аналітичні інструменти, такі як OLAP-куби для зберігання та аналізу великих обсягів даних, і методи машинного навчання для прогнозування змін якості повітря. Використання OLAP дозволило структурувати дані, визначити ключові показники ефективності (KPI) та виявити залежності між параметрами. Наступним логічним кроком стало застосування методів машинного навчання, таких як Random Forest і Gradient Boosting, для прогнозування рівнів забруднення. Це забезпечує не лише розуміння поточного

стану повітря, але й дозволяє оцінити ймовірність майбутніх змін, що є критичним для прийняття оперативних рішень екологічними службами.

Random Forest (випадковий ліс) є популярним методом машинного навчання, який використовує ансамблевий підхід для розв'язання задач класифікації та регресії. Метод базується на побудові множини рішень (дерев рішень) і об'єднанні їх результатів для отримання фінального прогнозу. Його ключова ідея полягає в створенні декількох дерев рішень на основі різних підвибірок даних та ознак, що забезпечує підвищену стійкість до перенавчання і шуму [45].

Однією з головних переваг Random Forest є його здатність оцінювати важливість ознак, що робить його ефективним інструментом для інтерпретації моделей. Крім того, метод добре працює з великими обсягами даних, оскільки "кожне дерево вирішує частину проблеми, що дозволяє швидко отримувати результати навіть на складних наборах даних" [46].

Gradient Boosting (градієнтний бустинг) є одним із потужних методів машинного навчання, який послідовно будує ансамбль моделей. На кожному етапі алгоритм додає нову модель, яка коригує помилки попередніх. Головна ідея Gradient Boosting полягає в мінімізації похибки за допомогою оптимізації градієнтного спуску, що дозволяє створювати високоточні прогнози навіть на складних і нерівномірно розподілених даних [47].

Метод також дозволяє ефективно виявляти нелінійні залежності між ознаками, завдяки чому є популярним у задачах прогнозування, зокрема для часових рядів. Як зазначено в документації Scikit-learn, "Gradient Boosting дозволяє налаштовувати широкий спектр гіперпараметрів, забезпечуючи високу гнучкість для адаптації до конкретних наборів даних" [47].

Обидва методи — Random Forest і Gradient Boosting — часто застосовуються в задачах прогнозування, зокрема в екологічному моніторингу, завдяки їхній здатності обробляти великі обсяги даних, знаходити ключові залежності між ознаками та забезпечувати високий рівень точності.

У цьому дослідженні методи машинного навчання були використані для прогнозування концентрації дрібнодисперсних частинок PM2.5. Основний акцент зроблено на аналізі моделей прогнозування, які поєднують кластеризацію даних із моделями Gradient Boosting і Random Forest. Попередньо проведена кластеризація стала основою для подальшого покращення точності моделей.

Перший етап полягав у підготовці даних. З бази даних були отримані дані з середніми значеннями параметрів, таких як температура, вологість, тиск, концентрації забруднювачів (NO2, SO2, CO, O3) та інші. Для покращення якості прогнозів були додані часові характеристики (місяць, день тижня, година та сезон). Це дозволило моделям врахувати сезонні та добові коливання параметрів якості повітря. Пропущені значення заповнювалися середніми по стовпчиках.

Для кожної моделі були визначені незалежні змінні (ознаки), такі як погодні умови та концентрації забруднювачів, а також цільова змінна — концентрація PM2.5. Дані були масштабовані для забезпечення рівномірного впливу кожної змінної на результати моделі.

Перший етап прогнозування передбачав використання моделей Gradient Boosting і Random Forest без попередньої кластеризації. Було оцінено їх точність за допомогою метрик середньоквадратичної похибки (MSE) і коефіцієнта детермінації (R^2). Результати показали, що ці моделі дають задовільні прогнози, але їх можна вдосконалити.

Далі було здійснено кластеризацію даних за допомогою алгоритму K-Means. Вибрані дані розділилися на три групи (кластери) залежно від схожості характеристик. Це дозволило розробити окремі моделі Gradient Boosting і Random Forest для кожного кластеру. Такий підхід дав змогу врахувати специфічні закономірності кожного кластеру, що, в свою чергу, підвищило точність прогнозів.

Під час прогнозування тестові дані були розподілені по кластерах, і для кожного з них використовувалася відповідна модель. Порівняння моделей з кластеризацією та без неї показало, що використання кластеризації суттєво покращило результати. Наприклад, для Gradient Boosting MSE з кластеризацією

було знижено до 2.0996 порівняно з 3.3229 без кластеризації, а R^2 підвищилося до 0.9270 порівняно з 0.8844. Аналогічно, моделі Random Forest з кластеризацією також показали кращі результати.

На рис. 38 зображено графіки прогнозування з порівнням двох методів з кластеризацією і без.

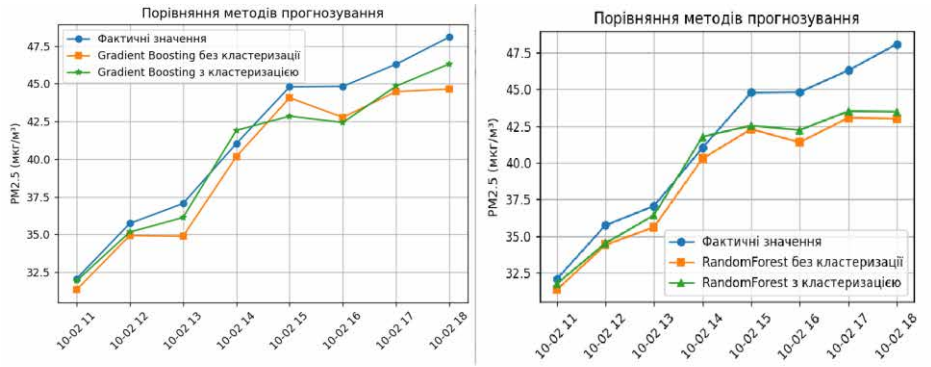

Рис. 38 Порівняння методів прогнозування з кластеризацією і без

Для прогнозування рівня $PM_{2.5}$ в кожному кластері використовувалися окремі моделі Random Forest та Gradient Boosting з адаптованими параметрами, що дозволило точніше налаштувати моделі під особливості кожного кластеру. Попередньо дані кластеризували за допомогою KMeans (3 кластери), після чого для кожного кластера було побудовано окрему модель. Для моделей Random Forest використовувалися параметри з 200 деревами рішень, глибиною до 15, а для малих кластерів — до 10. Gradient Boosting було налаштовано на 200 етапів з коефіцієнтом навчання 0.1 та максимальною глибиною 5. На тестових даних, визначених до конкретних кластерів, застосовувалися моделі, що відповідали кластеру, до якого вони належать, що дозволило забезпечити гнучкий та точний прогноз забруднення повітря.

У підсумку, поєднання кластеризації з методами машинного навчання дозволило не лише підвищити якість прогнозів, але й краще зрозуміти залежності між параметрами якості повітря. Такий підхід може бути корисним для розробки ефективних стратегій управління якістю повітря.

ВИСНОВКИ

У ході виконання магістерської роботи було реалізовано інтелектуальну систему моніторингу параметрів атмосферного повітря, яка відповідає сучасним вимогам до аналізу великих обсягів даних та прогнозування екологічних показників. Система базується на багатовимірному аналізі даних з використанням технологій OLAP, методів Data Mining та алгоритмів машинного навчання. Завдяки цьому вдалося вирішити поставлені завдання щодо консолідації даних, аналізу динаміки змін забруднення та прогнозування можливих критичних ситуацій.

Розроблене сховище даних забезпечує централізоване зберігання інформації, що включає параметри якості повітря, отримані з різних джерел, таких як КМДА та SaveEcoBot. Використання ETL-процесів дозволило автоматизувати збирання, трансформацію та завантаження даних, що підвищило точність і надійність аналізу. Побудовані OLAP-куби забезпечили багатовимірний підхід до дослідження показників забруднення, що дозволяє вивчати дані в різних розрізах: часових, географічних і категорійних.

Особливу увагу було приділено розробці звітів, які включають середні значення показників у розрізі міст, аналіз забруднення за місяцями, категоріями та часом, а також трендовий аналіз для виявлення довгострокових змін. Ці звіти дозволяють ефективно візуалізувати результати, роблячи їх доступними для кінцевих користувачів. Важливою складовою системи стало впровадження розрахунку ключових показників ефективності (KPI), які дали змогу кількісно оцінити рівень забрудненості та вплив на екологічну ситуацію.

Дослідження методів кластеризації та їх інтеграції з алгоритмами машинного навчання, такими як Random Forest та Gradient Boosting, показало значне підвищення точності прогнозування. Сегментація даних за допомогою кластеризації дозволила краще враховувати особливості різних груп спостережень, що позитивно вплинуло на результати моделювання. Аналіз

впливу кластеризації також довів її ефективність у задачах екологічного прогнозування, особливо для складних даних з високою варіативністю.

Отримані результати підтвердили доцільність використання OLAP-технологій для гнучкого аналізу великих обсягів екологічних даних. Виявлення закономірностей за допомогою методів Data Mining, таких як асоціативні правила, стало важливим етапом у побудові системи, що підтримує прийняття рішень на основі отриманих аналітичних даних. Прогнозування забруднення повітря з використанням моделей машинного навчання забезпечило можливість оперативного реагування на екологічні ризики.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Pollution Action Note - Data you need to know [Електронний ресурс] – Режим доступу до ресурсу: https://www.unep.org/interactives/air-pollution-note/
2. NIEHS. Air Pollution [Електронний ресурс] – Режим доступу до ресурсу: https://www.niehs.nih.gov/health/topics/agents/air-pollution/index.cfm
3. WHO. Ambient (outdoor) air pollution [Електронний ресурс] – Режим доступу до ресурсу: https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health
4. EPA. Transportation Air Pollution [Електронний ресурс] – Режим доступу до ресурсу: https://www.epa.gov/transportation-air-pollution-and-climate-change
5. EPA. Ground-level Ozone Basics [Електронний ресурс] – Режим доступу до ресурсу: https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics
6. NIEHS. Air Pollution and Health Effects [Електронний ресурс] – Режим доступу до ресурсу: https://www.niehs.nih.gov/health/topics/agents/air-pollution/index.cfm
7. IARC. Outdoor air pollution a leading environmental cause of cancer deaths [Електронний ресурс] – Режим доступу до ресурсу: https://www.iarc.who.int/wp-content/uploads/2018/07/pr221_E.pdf
8. WHO. Air Quality and Health [Електронний ресурс] – Режим доступу до ресурсу: https://www.who.int/health-topics/air-pollution#tab=tab_1
9. Pope III, C. A., & Dockery, D. W. (2006). Health effects of fine particulate air pollution: lines that connect. Journal of the Air & Waste Management Association, 56(6), 709-742.
10. He, J., et al. (2016). Association between PM2.5 and mortality in the urban population: A meta-analysis. Environment and Pollution, 219, 1021-1027.

11. Nie, J., et al. (2007). Exposure to traffic emissions throughout life and risk of breast cancer: the Western New York Exposures and Breast Cancer (WEB) study. Cancer Causes & Control, 18(9), 947-955.

12. Brook, R. D., et al. (2010). Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the American Heart Association. Circulation, 121(21), 2331-2378.

13. NIEHS. Children's Environmental Health [Електронний ресурс] – Режим доступу до ресурсу: https://www.niehs.nih.gov/research/programs/geh/geh_newsletter/2019/11/childrens-environmental-health/index.cfm

14. Як Україна вимірює забруднення повітря? [Електронний ресурс] – Режим доступу до ресурсу: https://ua-energy.org/uk/posts/yak-ukraina-vymiriuie-zabrudnennia-povitria

15. Моніторинг довкілля [Електронний ресурс] – Режим доступу до ресурсу: https://de.khnu.km.ua/labrun.aspx?a=257&b=2&c=65

16. АВТОМАТИЗОВАНИЙ МОНІТОРИНГ ТА ОЦІНКА ЯКОСТІ АТМОСФЕРНОГО ПОВІТРЯ [Електронний ресурс] – Режим доступу до ресурсу: http://eprints.library.odeku.edu.ua/id/eprint/6535/1/HrybOM_ChugaiAV_MV_Vaisala_AQT420_2019_ECOIMPACT_58.pdf

17. Hal Science. Air Quality Analysis Based on Clustering Methods. [Електронний ресурс] – Режим доступу до ресурсу: https://hal.science/hal-01243171/

18. Lund University. Integration of GIS and Machine Learning for Air Quality Prediction [Електронний ресурс] – Режим доступу до ресурсу: https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=3559141&fileOId=3559170

19. Taylor & Francis. Machine Learning Applications in Air Quality Monitoring [Електронний ресурс] – Режим доступу до ресурсу: https://www.tandfonline.com/doi/full/10.1080/10962247.2018.1459956

20. Springer. Applications of Random Forest and Gradient Boosting in Air Quality Forecasting [Електронний ресурс] – Режим доступу до ресурсу: https://link.springer.com/article/10.1007/s10462-023-10424-4

21. i-CSRS. Air Quality Analysis Using MapReduce and OLAP Technologies [Електронний ресурс] – Режим доступу до ресурсу: https://www.i-csrs.org/Volumes/ijasca/ID-45_Pg140-153_Air-Quality-Analysis-Based-On-MapReduce.pdf

22. [Електронний ресурс] – Режим доступу до ресурсу: https://engineering.purdue.edu/~engelb/abe565/sysanal.htm

23. What is System Analysis: Steps, Importance & Implementation [Електронний ресурс] – Режим доступу до ресурсу: https://www.grorapidlabs.com/blog/what-is-system-analysis-steps-importance-implementation

24. Career Advice: What is System Analysis [Електронний ресурс] – Режим доступу до ресурсу: https://www.indeed.com/career-advice/career-development/what-is-system-analysis

25. Systems Analysis and Design Methodology and Supporting Processes [Електронний ресурс] – Режим доступу до ресурсу: https://www.researchgate.net/publication/364311340_Systems_Analysis_and_Design_Methodology_and_Supporting_Processes

26. Steps and Best Practices in System Analysis [Електронний ресурс] – Режим доступу до ресурсу: https://implementationsciencecomms.biomedcentral.com/articles/10.1186/s43058-023-00504-5

27. Методологія системного аналізу [Електронний ресурс] – Режим доступу до ресурсу: https://studfile.net/preview/16455660/

28. Use Case Diagram – Unified Modeling Language (UML) [Електронний ресурс] – Режим доступу до ресурсу: https://www.geeksforgeeks.org/use-case-diagram/

29. Діаграма розгортання: Підручник з UML із ПРИКЛАДОМ [Електронний ресурс] – Режим доступу до ресурсу: https://www.guru99.com/uk/deployment-diagram-uml-example.html

30. What is OLAP (online analytical processing)? [Електронний ресурс] – Режим доступу до ресурсу: https://www.ibm.com/topics/olap

31. OLAP Operations in the Multidimensional Data Model [Електронний ресурс] – Режим доступу до ресурсу: https://www.javatpoint.com/olap-operations

32. What is OLAP? Cube, Analytical Operations in Data Warehouse [Електронний ресурс] – Режим доступу до ресурсу: https://www.guru99.com/online-analytical-processing.html

33. Що таке KPI (ключові показники ефективності)? [Електронний ресурс] – Режим доступу до ресурсу: https://hurma.work/blog/shho-take-kpi-klyuchovi-pokazniki-efektivnosti/

34. What Is a Data Warehouse? [Електронний ресурс] – Режим доступу до ресурсу: https://www.oracle.com/database/what-is-a-data-warehouse/

35. Data warehouse solutions [Електронний ресурс] – Режим доступу до ресурсу: https://www.ibm.com/data-warehouse

36. Databases architecture design [Електронний ресурс] – Режим доступу до ресурсу: https://learn.microsoft.com/en-us/azure/architecture/databases/

37. What is ETL (extract, transform, load)? [Електронний ресурс] – Режим доступу до ресурсу: https://www.ibm.com/topics/etl

38. SQL Server Integration Services [Електронний ресурс] – Режим доступу до ресурсу: https://learn.microsoft.com/en-us/sql/integration-services/sql-server-integration-services?view=sql-server-ver16

39. Ramzi A. Haraty, Mohamad Dimishkieh, Mehedi Masud. An Enhanced k-Means Clustering Algorithm for Pattern Discovery in Healthcare Data, International Journal of Distributed Sensor Networks Volume, Hindawi Publishing Corporation, pp.1-11, 2015.

40. Wang Zengping, Zhao Bing, Li Zekun, Sun Yi, Li Bin. Analysis of factors influencing the quality of energy meter verification based on Apriori

association rules [Електронний ресурс] – Режим доступу до ресурсу: https://www.sciencedirect.com/science/article/pii/S1003980917302307

41. Yang Fang. Apriori association rule-based evaluation method of smart energy meter failure rate level in typical environment [D]. Hunan University, 2020. DOI: 10.27135

42. Wu Wenbo. Data mining technology applying Apriori association rule algorithm for mining e-commerce potential customers [D]. Hangzhou: Zhejiang University of Technology.

43. SQL Server Analysis Services (SSAS) Overview [Електронний ресурс] – Режим доступу до ресурсу: https://learn.microsoft.com/en-us/analysis-services/analysis-services-overview?view=asallproducts-allversions

44. SQL Server Reporting Services (SSRS) [Електронний ресурс] – Режим доступу до ресурсу: https://learn.microsoft.com/en-us/sql/reporting-services/create-deploy-and-manage-mobile-and-paginated-reports?view=sql-server-ver15

45. IBM, "Random Forest is a commonly used machine learning algorithm for both classification and regression problems" [Електронний ресурс] – Режим доступу до ресурсу: https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems.

46. Scikit-learn Documentation, "RandomForestRegressor" [Електронний ресурс] – Режим доступу до ресурсу: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

47. Scikit-learn Documentation, "GradientBoostingRegressor" [Електронний ресурс] – Режим доступу до ресурсу: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html