НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ БІОРЕСУРСІВ
І ПРИРОДОКОРИСТУВАННЯ УКРАЇНИ

Факультет інформаційних технологій

УДК

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

МАГІСТЕРСЬКА КВАЛІФІКАЦІЙНА РОБОТА

на тему _Система моніторингу показників складу річкової води_____

_____

Спеціальність ___122 «Комп'ютерні науки»_____

_____
(код і назва)

Освітня програма _Комп'ютерний еколого - економічний моніторинг__

_____
(назва)

Орієнтація освітньої програми _____
(освітньо-професійна або освітньо-наукова)

Гарант освітньої програми

_____ _____ _____
(науковий ступінь та вчене звання) (підпис) (ПІБ)

Керівник магістерської кваліфікаційної роботи

_____к.т.н., доцент_____ _____ __Сватко Віталій Володимирович__
(науковий ступінь та вчене звання) (підпис) (ПІБ)

Виконав _____ ___Скорик Максим Віталійович___
(підпис) (ПІБ студента)

КИЇВ-2024

ЗМІСТ

ВСТУП

Річкова вода є одним з найважливіших природних ресурсів, який використовується для пиття, поливу, промисловості та інших потреб. Тому контроль за її станом є надзвичайно важливим. Для контролю стану води і використовується такий метод як моніторинг.

Моніторинг поверхневих вод – система послідовних спостережень, збору, обробки даних про стан водних об’єктів, прогнозування їх змін та розробки науково обґрунтованих рекомендацій для прийняття управлінських рішень, які можуть позначитися на стані вод.

Саме тому основною метою дослідження стало налагодження системи ефективного аналізу якості води на основі зібраних показників води.

Об’єктом дослідження стали річкові води України, якісні показники води та вплив різних факторів на її склад і стан (природні умови, антропогенні джерела забруднення, сезонні зміни).

Предметом дослідження є методи та набори показників, що використовуються для оцінки якості річкової води, такі як хімічний склад, концентрація забруднюючих речовин.

В рамках дослідження буде проведено комплекс заходів із аналізу хімічно фізичних показників води. Взявши за основу зчитані та збережені показники буде розраховано загальний показник рівня забруднення води. Кожне із набору спостережень збору води буде класифіковано згідно офіційної індексації та присвоєно відповідний рівень забруднення від «чистого» до «катастрофічно забрудненого» рівня.

Для проведення аналізу будуть використані дані із дослідного ресурсу Державного агентства водних ресурсів України. Насамперед даними дослідження виступають хіміко-фізичні такі показники води як: Хлориди, Нітрати, Азот, Амоній, Нітрити, Кисень, рівень біохімічного споживання кисню та інші. Також для забезпечення повноцінної роботи було зібрано дані станцій, водних ресурсів – річок та їх приток.

Результатом дослідження стануть загальні характеристики, графіки а також знайдені залежності та ступені важливості показників, у розрізі сили впливу цих показників на загальний рівень забрудненості вод.

Отже, завданням роботи є: проведення початкового системного та предметного аналізу області дослідження, накопичення даних спостережень зі станцій збору вод для забезпечення вимог повноти, для подальшої побудови моделей для аналізу, аналіз побудованих моделей для насамперед класифікації спостережень за рівнем забруднення, побудова звітної інформації в вигляді графіків, таблиць та матриць, що показують рівень забруднення, можливі критичні ділянки, показники що найбільше впливають на перевищення нормального рівню якості води.

Актуальність дослідження продиктована необхідністю постійного спостереження та контролю за якістю поверхневих вод для забезпечення безпечного подальшого використання населенням.

1 СИСТЕМНИЙ АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

1.1 Предметна область

Якість води – це поєднання хімічних показників та фізичних властивостей вод, що відображають ступінь їх відповідності певним потребам та вимогам (рибогосподарським, господарсько-питним, санітарно-гігієнічним).

Якість річкових вод визначається на підставі аналізу та оцінки комплексу показників з подальшим розрахунком інтегрального параметру: коефіцієнт забруднення, гідрохімічний індекс забрудненості вод ІЗВ, інтегральний екологічний індекс, екологічні стани за міжнародними методиками згідно вимог Водної Рамкової Директиви ЄС та Міжнародної комісії із захисту річки Дунай та ін. [1].

Саме для контролю стану і якості води використовується такий метод як моніторинг.

Моніторинг поверхневих вод – система послідовних спостережень, збору, обробки даних про стан водних об’єктів, прогнозування їх змін та розробки науково обґрунтованих рекомендацій для прийняття управлінських рішень, які можуть позначитися на стані вод.

Основна мета налагодження системи спостережень і контролю за забрудненням водних об’єктів – це отримання інформації про природну якість води та оцінка змін якості води внаслідок дії антропогенних факторів.

Служба спостережень та контролю (моніторингу) виконує такі завдання:

1. спостереження та контроль рівня забруднення водного середовища за хімічними, фізичними та гідробіологічними показниками;
2. вивчення динаміки вмісту забруднюючих речовин і виявлення умов, за яких мають місце коливання рівня забруднення;

3. дослідження закономірностей процесів самоочищення та накопичення забруднюючих речовин у донних відкладах [2].

Основним елементом моніторингу вод є стації забору води, або пункти спостереження за якістю поверхневих вод. Розташування таких пунктів та станцій залежно від географії може різнитися, місцями доречного аналізу можуть виступати: початки витоку рік або кінцеві створи річок, що впадають в подальшому в моря, місця скидань стічних і дощових вод підприємств чи населення, місця із можливим впливом результатів сільськогосподарської діяльності людей, кордони країн чи економічних зон, тощо.

Сучасні методи моніторингу річкової води дозволяють визначати широкий спектр показників її складу, включаючи фізичні, хімічні та біологічні. Однак, ці методи часто є дорогими та трудомісткими, що ускладнює їх регулярне застосування.

Оскільки не існує єдиного показника, який визначав би весь комплекс характеристик води, оцінювання якості води проводиться на основі системи показників. Запис та відповідне збереження показників хіміко фізичного вмісту води за графіком на станціях відбувається приблизно раз на місяць.

Хімічними показниками які будуть виступати основними у ході подальшого дослідження є: вміст розчиненого кисню, хімічне та біохімічне споживання кисню; водневий показник; уміст азоту, фосфору, хлору, діоксиду натрію.

1.2 Постановка завдання

Після проведення початкового аналізу предметної області було визначено: основним завданням роботи буде проведення дослідження роботи системи моніторингу показників складу річкової води. Насамперед це передбачає збір, аналіз та створення звітної інформації за зібраними показниками. Із метою

визначення відповідності річкової води стандартам якості, знаходження можливих джерел забрудників, а також аналізу впливу води із відповідними якостями на навколишнє середовище.

Дуже важливою складовою дослідження буде визначення показника який і буде представляти рівень забрудненості.

На теперішній час в Україні та в інших країнах світу розроблена досить велика кількість критеріїв комплексної оцінки якості поверхневих прісних вод. Одні класифікації належать до бактеріологічних та фізико-хімічних, в основу інших покладена гідробіологічна оцінка забрудненості вод. Кожен із них дає змогу отримувати важливу інформацію, а при їх застосуванні разом - оцінювати водне середовище з екологічних позицій.

Оцінка якості води за хімічними показниками вважається досить трудомістким завданням, оскільки воно базується на порівнянні середніх концентрацій, які спостерігаються в пунктах контролю якості вод, з встановленими нормами (ГДК) для кожного інгредієнта. Більшість із запропонованих сьогодні комплексних показників отримано шляхом об’єднання та узагальнення численних часткових показників у один інтегруючий, який дає змогу характеризувати різні становища водних об’єктів [3].

Основні методики комплексної оцінки якості вод засновані на використанні наступних комплексних показників: індексу забруднення води (ІЗВ), модифікованого (ІЗВ), комплексного індексу забруднення (КІЗ), коефіцієнта забрудненості χ, комплексного показника екологічного стану (КПЕС) і узагальненого екологічного індексу.

Для реалізації завдань роботи по аналізу було прийняте рішення використовувати модифікований індекс забруднення води для визначення ступеня чистоти річкової води.

У роботі модифікований ІЗВ розраховується по шості показниках: БСК5, O2, CL, NO2, SO4, PO4. Розрахунок відбувається за формулою:

$$ІЗВ = (1/6)\ \Sigma\ (C_i / ГДК_i)$$

де C_i – середнє арифметичне значення показника якості води;

$ГДК_i$ – гранично допустима концентрація.

Після обчислення числа ІЗВ за формулою, оцінка якості води виконується за наступними класами: I – дуже чиста (ІЗВ≤0,3); II – чиста (0,3<ІЗВ≤1); III – помірно забруднена (1<ІЗВ≤2,5); IV – забруднена (2,5<ІЗВ≤4); V – брудна (4<ІЗВ≤6); VI – дуже брудна (6<ІЗВ≤10); VII – надзвичайно брудна (ІЗВ>10).

Основним результатом який має бути представлений будуть таблиці звітної інформації, що мають демонструвати як зведені загальні статистичні дані показників так і звіти динаміки змін показників, які дозволять аналізувати варіювання значень показників.

Використання таких звітів дозволить набагато ясніше та детальніше надавати доступ до таких комплексних даних дослідження, як інформація про рівні забруднення, найкритичніші ділянки, тощо. Це дозволить значно спростити та пришвидшити проведення аналізу на зразок знаходження можливих місць забруднення чи викидів збудників у річкові води.

Для реалізації визначених звітів було вирішено використовувати як лінійні моделі для аналізу даних, так і нетривіальні моделі машинного навчання. Методи машинного навчання дозволять швидко знаходити, відображати та аналізувати можливі залежності між показниками води та індексом забруднення.

1.3 Проектування системи

Для вирішення задач дослідження було вирішено використовувати дані, доступні у відкритому доступі, що знаходяться на офіційному електронному ресурсі Державного агентства водних ресурсів України. Архів цих даних включає в себе фактично всі офіційні станції забору річкових вод України.

Ресурс містить всі дані про розташування пунктів збору проб води,

включаючи географічні та промислові особливості – причини чому дані збираються саме в цих місцях.

Основою системи стануть записи показників спостережень саме із цього джерела - цей ресурс містить статистичні дані спостережень хімічних рівнів показників.

Провівши початковий аналіз предметної області системи, для демонстрації опису загальних функцій та принципів роботи системи, за допомогою уніфікованої мови моделювання (UML) була побудована діаграма прецедентів (Рис.1).

Діаграми варіантів використання (use case diagrams) використовуються для відображення сценаріїв використання системи (use cases) та користувачів системи (actors), які використовують її функції. Актори на діаграмі варіантів використання позначаються символом людини, а варіанти використання – еліпсом [4]. Загалом, варіанти використання та актори можуть бути пов’язані між собою

Трьома видами зв’язків:

-узагальненням (generalization);

- розширенням (extend relationship);

- включенням (include relationship).

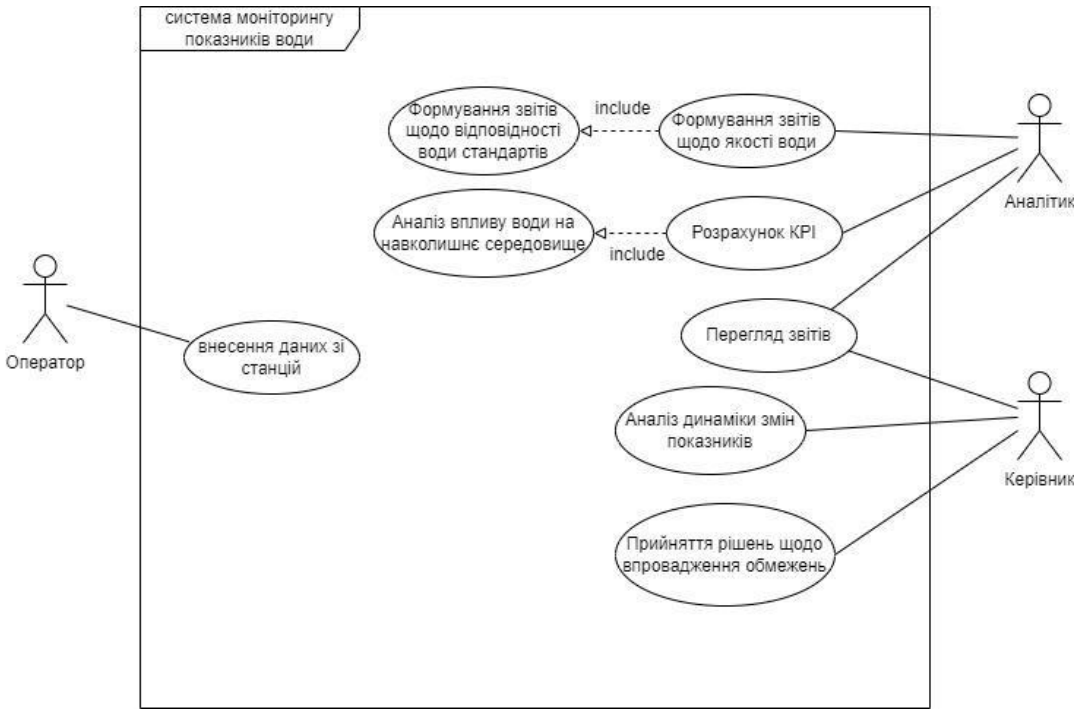

Рис. 1 Діаграма прецедентів (use case)

Як можна побачити, після проведеного аналізу було визначено, що для повноцінної роботи системою будуть користуватися 3 типи користувачів:

- Оператори – користувачі з найменшою кількість прав, основною задачею яких буде внесення початкових даних із станцій у оперативне джерело системи.
- Аналітики – користувачі, які будуть складати основу системи, матимуть доступ до таких функцій як, формування звітної інформації, розрахунку KPI, та проведення аналізу отриманих показників
- Керівник – виступатиме керуючим користувачем, який відповідатиме за контроль створеної звітної інформації, прийматиме управлінські рішення на основі отриманих результатів.

Основними активностями, найважливішими для подальшого дослідження є «Формування звітів щодо якості води» та «Аналіз змін показників».

2 МЕТОДИ ТА ТЕХНОЛОГІЇ АНАЛІЗУ

2.1 Архітектура системи

Рівень розвитку сучасних технологій настільки високий, що дозволяє побудувати інформаційну систему будь-якого масштабу, складності й функціональності. Однак, з огляду на вимоги, засновані на показниках різних бізнес-оцінок, виникають додаткові складнощі, вирішення яких зводиться до забезпечення раціонального підходу до процесу проектування, реалізації й подальшій експлуатації інформаційних систем. Виходячи із цього, можна однозначно вважати обрану архітектуру одним з основних показників ефективності створюваної інформаційної системи, а, отже, і успішності.

Визначити поняття "архітектура інформаційної системи" можна безліччю способів. Це зв’язано:

- З відсутністю загальноприйнятого визначення самої інформаційної системи. З огляду на складність структури, достатнім способом описати її можливо тільки при консолідації декількох точок зору, що в кожному конкретному випадку може приводити до різних результатів.
- З різноманіттям трактувань самого терміну "архітектура".

У результаті, архітектуру інформаційної системи можна описати як концепцію, що визначає модель, структуру, виконувані функції й взаємозв’язок компонентів інформаційної системи [5].

Зважаючи на особливість дослідження та провівши подальший аналіз області, була побудована діаграма топології (Рис. 2) досліджуваної системи, що представляє собою базову архітектуру системи. Насамперед в основу архітектури були вкладені поняття технологій OLAP.

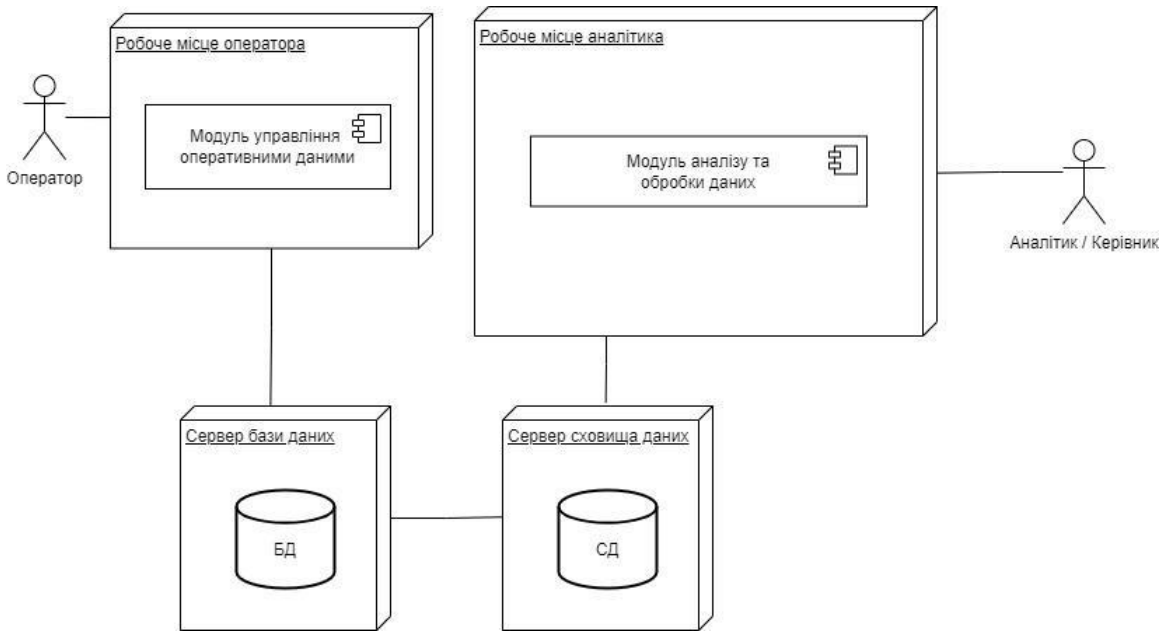

Рис. 2 Архітектура системи

Акторами системи, так і залишилися користувачі оператори, аналітики та керівник. Користувач оператор відповідно працюючи зі свого робочого місця матиме доступ до модуля управління оперативними даними, що включає в себе роботу із базою даних. База даних розташована на своєму окремому сервері.

Розділення доступу користувача аналітика та керівника буде відносно мінімальним у розрізі діаграми архітектури, тому вони тут зображені одним актором, та взаємодіючи із одним робочим місцем використовують умовно один модуль аналізу та обробки. За допомогою цього модуля вони матимуть доступ до основного елемента системи із використанням методології OLAP - сховища даних, що розташоване у своєму окремому сервері. Цей модуль дозволить проводити весь аналіз показників отриманих із оперативного джерела – бази даних.

2.2 Загальні поняття з напрямку OLAP - технології

Під час проведення дослідження роботи було зібрано неймовірну кількість необроблених, непідготовлених даних абсолютно різного масштабу. Для

правильної та точної обробки, збереження та аналізу даних і було вирішено використання такої методології як OLAP (англ. online analytical processing, аналітична обробка у реальному часі).

В основі концепції OLAP лежить багатомірне концептуальне подання даних (Multidimensional conceptual view). Головна ідея OLAP-системи полягає в побудові багатовимірних таблиць, які можуть бути доступні для запитів користувачів. Ці багатовимірні таблиці або так звані багатовимірні куби, або сховища даних, будуються на основі вихідних і агрегованих даних. І вихідні, і агреговані дані для багатовимірних таблиць можуть зберігатися як у реляційних, так і в багатовимірних базах даних.

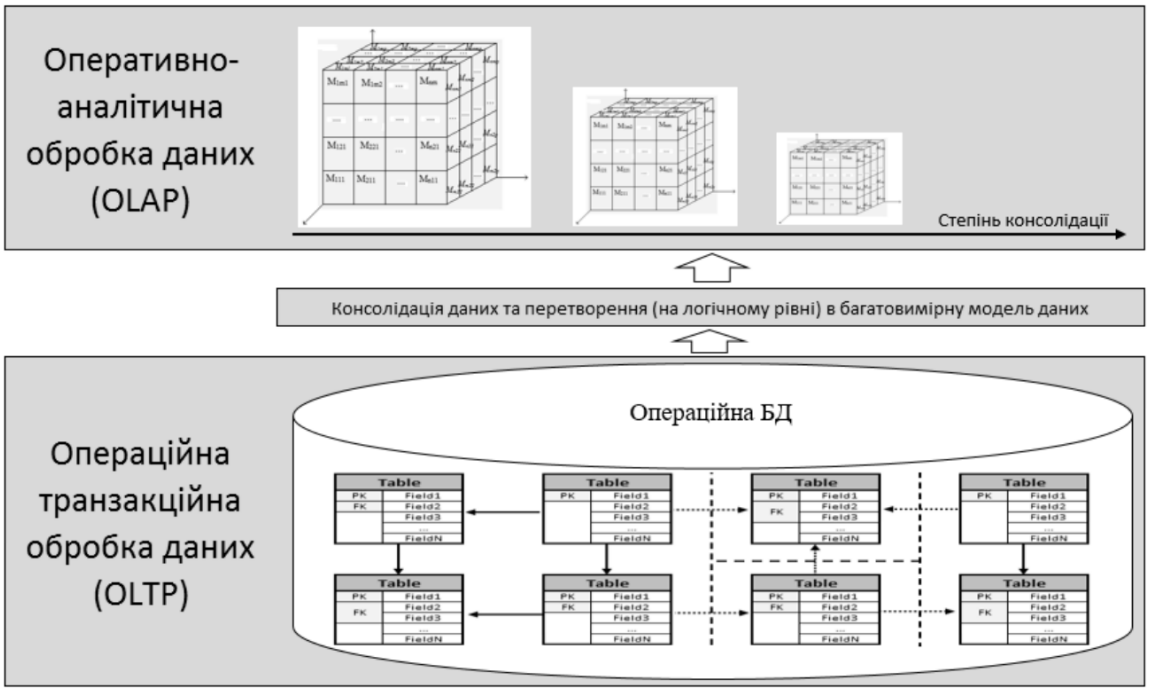

Рис. 3 Узагальнена схема взаємозв’язку між OLTP та OLAP технологіями

Технологія OLAP орієнтована, головним чином, на обробку нерегламентованих запитів до сховищ даних. Таким чином, технологія OLAP є подальшим розвитком технологій баз даних або OLTP, і тому потребує нових методів обробки і відповідних програмних засобів або автономних, або в складі існуючих СКБД [6].

Ідея, покладена в основу технології інформаційних сховищ даних, полягає в тому, що проводити оперативний аналіз безпосередньо на базі інформаційних систем неефективно. Натомість, всі необхідні для аналізу дані витягуються з декількох традиційних баз даних (в основному, реляційних), перетворюються і потім поміщаються в одне джерело даних – сховище даних.

В процесі перенесення дані:

- Очищаються – усунення непотрібної інформації;
- Агрегуються – обчислення сум, середніх;
- Трансформуються – перетворення типів даних, реорганізація структур зберігання;
- Об'єднуються із зовнішніх і внутрішніх джерел – приведення до єдиних форматів;
- Синхронізуються – відповідність одному моменту часу.

Сьогодні, технології побудови сховищ даних є основою для створення повноцінних інтелектуальних систем аналізу даних, орієнтованих на рішення слабко структурованих задач прийняття рішень [7].

Таблиці в сховищах даних додатково поділяються на 2 різновиди.

Таблиця фактів містить фактичні числові дані (метрики або характеристики), які підлягають аналізу. Вона зберігає інформацію про конкретні події чи транзакції, які можна виміряти або підрахувати. Зазвичай має посилання на таблиці вимірів через зовнішні ключі, які дозволяють деталізувати дані і не має власного ключа – унікального ідентифікатора. Містить числові метрики, які можуть агрегуватися

Таблиці вимірів - зберігають відносно додаткову інформацію про події, описані в таблиці фактів. Вона визначає контекст даних, які містяться в таблиці фактів, і використовується для фільтрації та групування даних під час аналізу. Містить атрибути або характеристики подій, описаних у таблиці фактів.

Архітектури сховищ даних:

- Схема «Зірка»: Це найпростіша схема, де кожна таблиця вимірів безпосередньо пов'язана з таблицею фактів. Ця схема забезпечує швидкий та інтуїтивно зрозумілий процес запитів і є особливо актуальною для Data Marts, підмножин сховищ даних, призначених для конкретної галузі бізнесу.

- Схема «Сніжинка» : Розширення схеми «зірка», схема «сніжинка» передбачає нормалізацію таблиць вимірів у декілька пов'язаних таблиць. Це може зменшити простір для зберігання і зберегти цілісність даних за рахунок усунення надмірності. Однак це може призвести до збільшення складності SQL-запитів.

- Схема «Галактика»: Відома як багатофакторна схема, вона дозволяє інтегрувати кілька таблиць фактів, які мають спільні таблиці вимірів. Ця схема підходить для складних моделей даних з різноманітними бізнес-процесами. Вона особливо корисна, коли ваша модель даних відображає кілька тем або предметів в одному сховищі даних [9].

Отже, під час проведення дослідження, використання технологій OLAP дозволило значно спростити, узагальнити та стандартизувати значні масиви даних показників води.

2.3 Моделювання джерел даних

Як було описано раніше, у постановці завдання, основним джерелом початкових, необроблених даних станцій виступає офіційний ресурс моніторингу Державного агентства водних ресурсів України.

Цей ресурс є фактичною, офіційною онлайн-платформою, призначеною для збору, зберігання, аналізу та надання відкритого доступу до інформації про стан водних ресурсів країни. Платформа надає інформацію в реальному часі про останні показники якості поверхневих вод України. Це включає ключові хімічні, фізичні та біологічні показники. Дані в реальному чаті представляються графічно на інтерактивній карті, яка показує місце розташування фактично всіх станцій води, класифікуючи їх на категорії забруднення за кольорами [22].

Та головним джерелом даних, які і складуть основу аналізу системи, становить один підрозділів цього ресурсу, а саме підрозділ «Звітність». Цей розділ становить собою по суті архів всіх історичних даних хімічних показників, які були зібрані на станціях. Він містить категорізацію за типами звітів загального користування:

- Дані моніторингу (за адміністративно-територіальним принципом)
- Дані моніторингу (за водогосподарською організацією)
- Дані моніторингу (за ознакою транскордонного створу)
- Дані моніторингу (за районом річкового басейну або суббасейну)

Для досягнення цілей дослідження, було вирішено збирати дані класифіковані за районом річкового басейну та суббасейну, для забезпечення повноцінної оцінки стану якості окремих річок.

Для отримання таблиць даних із ресурсу, було обрано відповідно актуальні проміжки часу, від 2020 до 2023 років на визначених річках. Річки, що були обрані для дослідження включають:

- Дністер
- Південний Буг
- Дніпро (Суббасейни Прип’ять та Середній Дніпро)

Дані, що зберігаються у архіві представлені у вигляді необроблених таблиць значень показників.

Архівні таблиці включають переліки відповідних станцій збору води із відповідною для кожної станції таблицею таблиць, що описують, які хімічні показники відслідковуються на станції, та у яких кількостях, якої дати. Ця платформа також дозволяє зберігати ці таблиці показників у форматі .csv, цей інструментарій і був використаний в подальшому для спрощення збереження та перенесення даних.

Таким чином, для кожної із визначених річок – басейнів та їх суббасейнів були сформовані часові таблиці даних, за 2020 – 2022 роки.

Як вже було визначено, для проведення повноцінного аналізу зібраних даних, потрібні дві складові – оперативне джерело (база даних) та відповідно сховище даних. Оперативне джерело міститиме в собі початкові, необроблені зібрані дані. Ці дані представляють собою дані про річки, на яких розташовуються станції, відповідно детальна інформація про станції – їх розташування відносно витоку річки, номер станції, важливий фактор, у зв’язку із яким збір показників відбувається саме там та інші.

Найважливішою частиною на основі якої і буде відбуватися аналіз це власне дані спостережень. Кожна станція приблизно раз на місяць проводить збір показників – такий збір показників і становить одне спостереження. Дані про вміст цих показників, наприклад Хлору, Кисню, Нітратів, у спостереженні і будуть становити основу аналізу та дослідження системи. Саме на основі цих спостережень і можлива оцінка якості річкової води.

Провівши подальший аналіз, для підтримання високої ефективності та

надійності, у якості оперативного джерела для системи моніторингу річкової води була використана окрема база даних, її схема представлена на Рис. 4.

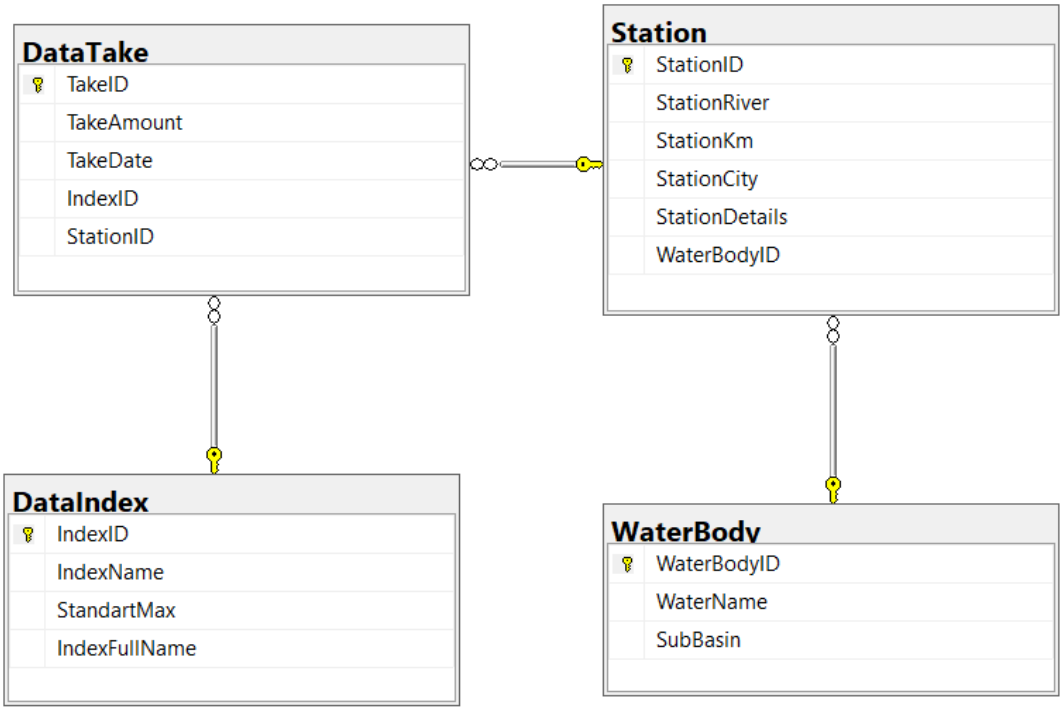

Рис. 4 Структура оперативного джерела – бази даних

Проста структура бази даних дозволить значно спростити подальше наповнення та контроль за даними. Саме даними із цієї бази даних і буде в подальшому наповнене сховище даних – основа системи, ціль аналізу та дослідження.

База даних була створена використовуючи SQL Server та СУБД Microsoft SQL Server Manager.

База даних складається із 4 сутностей:

- Index – сутність показника, зберігатиме у собі дані про всі показники складу води, які можуть зчитуватися станціями. Містить інформацію про скорочену наукову назву показника, максимальну допустиму межу для

стандарту якості, а також повну назву яка буде відображатися українською мовою.

- TakeData – сутність збору даних визначеного показника визначеної станції, є основою бази даних, позначає забір даних показника, а саме його кількість, та дату коли цей показник був записаний.
- Station – сутність станції збору показників, що розташована на визначеному водному ресурсі. Містить у собі дані про назву станції, контактний номер телефону керівника станції, ім’я цього керівника, адресу станції.
- WaterBody – сутність водного ресурсу, буде позначати на якій саме річці чи озері розташована визначена станція. Містить дані про назву цього ресурсу, його типу, довжину чи розміри, короткий опис водойми.

Ця реляційна база даних забезпечить швидкий, надійний та безпечний спосіб зберігання для подальшого перенесення даних до сховища.

Наступним кроком підготовки даних стало власне проектування основної структури OLAP - багатовимірного кубу, сховища даних.

Сховище даних - це система, яка використовується для звітності та аналізу даних і вважається основним компонентом бізнес-аналітики. Воно зберігає історичні дані, основна задача якої - аналізувати тенденції в часі. Мета сховища даних - консолідувати і централізувати великі обсяги даних з різних операційних джерел, роблячи їх доступними для ефективної аналітики та звітності [9].

Продовжуючи аналіз особливостей створення та структуризації сховищ даних та керуючись принципами OLAP, було побудовано наступну структуру сховища даних типу «Зірка» (Рис. 5). Для створення та налаштування сховища даних був використаний Microsoft SQL Server та СУБД Microsoft SQL Server Manager.

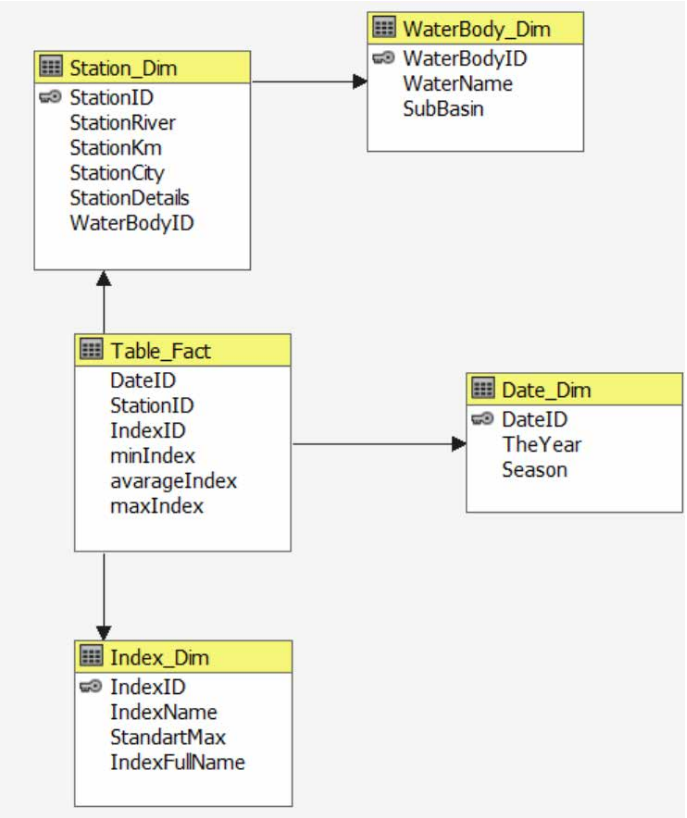

Рис. 5 Структура сховища даних

Як можна побачити система включатиме наступні таблиці вимірів:

- Date dim – вимір часу, вказує на дату коли було здійснено запис вимірювання значень показників

- Station dim – вимір станції, описує всі необхідні дані станцій, де було проведено спостереження вимірювання

- Water body dim – вимір водного ресурсу, більшого водного басейну річки, такої як Дніпро, Південний Буг, до яких відносять територіальним розташуванням станції

- Index dim – вимір показнику, що описує власні показники води, на основі яких буде проводитися аналіз (наприклад показник хлору, кисню)

Згідно архітектурної моделі «Зірка», сховище даних також включає в себе таблицю фактів - Table Fact, яка містить у собі показники обчислених фактів.

Вона включає в себе мінімальне, середнє та максимальне значення конкретного показника який був записаний на станції, у визначену дату. Ці факти будуть розраховані та агреговані за порами року під час перенесення даних із оперативного джерела до власне сховища даних. Як можна побачити ця таблиця містить лише три зовнішні ключі із таблиць вимірів, і не містить унікального ключа. Це задовольнить вимоги методології OLAP, дозволить проводити комплексний аналіз числових фактів – обчислень за всіма вказаними вимірами.

2.4 Основи теорії машинного навчання

Машинне навчання - це розділ штучного інтелекту, спрямований на створення систем, здатних навчатися на історичних даних, виявляти закономірності та приймати логічні рішення з мінімальним втручанням людини або без нього. Це метод аналізу даних, який автоматизує побудову аналітичних моделей, які можуть використовувати дані, що охоплюють різні форми цифрової інформації, включаючи числові значення, слова, та зображення [21].

Програми машинного навчання навчаються на основі вхідних даних і постійно підвищують точність результатів за допомогою автоматизованих методів оптимізації.

Якість моделі машинного навчання залежить від двох основних аспектів:

- Якість вхідних даних. При розробці алгоритмів машинного навчання часто зустрічається фраза «сміття на вході, сміття на виході». Цей вислів означає,

що якщо ви вводите низькоякісні або безладні дані, то результат роботи вашої моделі буде значною мірою неточним.

- Сам вибір моделі. У машинному навчанні існує безліч алгоритмів, які може вибрати фахівець з даних, і кожен з них має своє специфічне застосування. Дуже важливим є вибір правильних алгоритмів для кожного випадку використання. Для прикладу, нейронні мережі - це тип алгоритму зі значним показником точності та універсальності, які він може забезпечити. Однак для невеликих обсягів даних вибір простішої моделі часто дає кращі результати.

Чим краща модель машинного навчання, тим точніше вона може знаходити особливості та закономірності в даних. Це, в свою чергу, означає, що рішення і прогнози будуть точнішими [10].

Алгоритми машинного навчання поділяються на п'ять широких категорій: навчання з учителем, навчання без нагляду, напівнавчання під наглядом, самонавчання та навчання з підкріпленням.

1. Контрольоване навчання, або навчання з учителем (Supervised machine learning)

Навчання з учителем - це тип машинного навчання, в якому модель навчається на визначеному наборі даних (тобто цільова або результуюча змінна відома). Наприклад, для системи, що досліджується даними для навчання будуть спостереження хімічних показників що були зафіксовані та зберігаються у джерелі даних.

Контрольоване навчання зазвичай використовується для оцінки ризиків, розпізнавання зображень, аналітики передбачень та виявлення шахрайства і включає в себе кілька типів алгоритмів.

- Регресійні алгоритми - прогнозують вихідні значення шляхом визначення лінійних зв'язків між дійсними або безперервними величинами.
- Алгоритми класифікації - прогнозують категоричні вихідні змінні шляхом маркування фрагментів вхідних даних.
- Наївні класифікатори Байєса - дозволяють вирішувати завдання класифікації для великих наборів даних. Вони також є частиною сімейства генеративних алгоритмів навчання, які моделюють розподіл вхідних даних для певного класу або категорії.
- Нейронні мережі - імітують роботу людського мозку з величезною кількістю пов'язаних між собою вузлів обробки, які можуть полегшити такі процеси, як переклад природної мови, розпізнавання зображень, розпізнавання мови та створення зображень.
- Алгоритми випадкових лісів - прогнозують значення або категорію, комбінуючи результати з декількох дерев рішень.

2. Неконтрольоване машинне навчання або навчання без учителя (unsupervised learning)

Алгоритми неконтрольованого навчання - такі як Apriori, Gaussian Mixture Models (GMM) та аналіз головних компонент (PCA) - роблять висновки з неописаних наборів даних, полегшуючи дослідницький аналіз даних та дозволяють розпізнавання шаблонів і моделювання передбачень.

Найпоширенішим методом неконтрольованого навчання є кластерний аналіз, який використовує алгоритми кластеризації для класифікації точок даних за схожістю значень (як у сегментації клієнтів або виявленні аномалій).

Алгоритми асоціацій дозволяють аналітикам даних виявляти зв'язки між об'єктами даних у великих базах даних, полегшуючи візуалізацію даних і зменшуючи їхню розмірність.

3. Самокероване машинне навчання

Самокероване навчання (Self-supervised learning, SSL) - парадигма машинного навчання для обробки даних нижчої якості, а не для покращення кінцевих результатів. Самокероване навчання точніше імітує те, як вчаться класифікувати об'єкти люди.

Дозволяє моделям навчатися на неописаних даних, замість того, щоб потребувати величезних наборів анотованих і/або маркованих даних. Алгоритми SSL, які також називають алгоритмами передбачень або претекстового навчання, вивчають одну частину вхідних даних за іншою частиною, автоматично генеруючи мітки і перетворюючи неконтрольовані задачі на контрольовані.

Ці алгоритми особливо корисні для таких завдань, як комп'ютерний зір і NLP (Natural language processing), де обсяг маркованих навчальних даних, необхідних для навчання моделей, може бути надзвичайно великим (іноді надмірно великим).

4. Навчання з підкріпленням (Reinforcement learning)

Навчання з підкріпленням, яке також називають навчанням з використанням зворотного зв'язку з людиною (RLHF), - це тип динамічного програмування, який навчає алгоритми за допомогою системи заохочень і покарань.

Щоб застосувати навчання з підкріпленням, агент (машина що навчається) виконує дії в певному середовищі для досягнення заздалегідь визначеної мети. Агент отримує винагороду або покарання за свої дії на основі встановленої

метрики (зазвичай у балах), заохочуючи агента продовжувати хороші практики і відмовлятися від поганих. З повторенням агент навчається найкращим стратегіям.

Алгоритми навчання з підкріпленням поширені в розробці відеоігор і часто використовуються для навчання роботів повторювати людські завдання.

5. Напівконтрольоване навчання (Semi-supervised learning)

П'ятий тип техніки машинного навчання пропонує комбінацію між контрольованим і неконтрольованим навчанням.

Алгоритми напівконтрольованого навчання навчаються на невеликому наборі класифікованих даних і великому наборі неописаних даних, причому описані дані керують процесом навчання для більшого масиву неописаних даних. Напівконтрольована модель навчання може використовувати неконтрольоване навчання для виявлення кластерів даних, а потім використовувати контрольоване навчання для маркування кластерів.

Незалежно від типу, моделі машинного навчання можуть отримувати інформацію з корпоративних даних, але їхня вразливість до людського фактору та упередженості даних робить відповідальну практику використання штучного інтелекту робить використання цих методів комплексним завданням [11].

Машинне навчання було своєрідною революцією в обробці, аналізі та вилученні інформації з величезних масивів даних. Незважаючи на свої численні переваги, моделі машинного навчання можуть припускатися помилок, що іноді призводить до справді катастрофічних результатів. Ось чому точність є вирішальним фактором у будь-якому проекті машинного навчання.

Точна модель є надійною, а коли мова йде про рішення, що змінюють життя, ніщо не має більшого значення, ніж надійність.

У машинному навчанні точність - це оціночна метрика, яка вимірює загальну правильність прогнозів моделі. Вона являє собою відношення

правильно передбачених випадків (як істинних позитивних, так і істинних негативних) до загальної кількості випадків у наборі даних.

Точність розраховується за наступною формулою:

Точність = (Кількість правильних прогнозів) / (Загальна кількість прогнозів)

Вона виражається у відсотках в діапазоні від 0% до 100%. Вища точність вказує на те, що передбачення моделі краще узгоджуються з фактичними мітками або істинними значеннями.

Точність зазвичай використовується в машинному навчанні для задач бінарної класифікації, де метою є класифікація екземплярів в один з двох класів або категорій. Однак точність також може застосовуватися до задач багатокласової класифікації, враховуючи кількість правильно передбачених екземплярів у всіх класах.

Точність є фундаментальним аспектом машинного навчання і має вирішальне значення для успіху будь-якого проекту. Точна модель є надійною і може використовуватися з упевненістю, тоді як неточна може призвести до катастрофічних наслідків. Вимірювання точності, розуміння факторів, які на неї впливають, і вжиття заходів для її забезпечення є критично важливими для успіху проекту машинного навчання [23].

3 ПРОГРАМНА РЕАЛІЗАЦІЯ ЗАДАЧ

3.1 Механізми вилучення, обробки і передачі даних

Процеси, що описують процедури збереження, обробки та управління даними є основою для ефективного досягнення цілей дослідження.

Основними інструментами для здійснення цього стали BI та служба SSAS.

BI (англ. Business Intelligence, інтелектуальний аналіз даних, бізнес-аналітика) – це набір методик, технологій та практик основною метою використання яких є проведення поглибленого, комплексного аналізу даних. Кінцевим результатом застосування методів BI є корисна інформація, яка може бути використана для прийняття управлінських рішень.

Основні компоненти системи BI:

- Збір даних - це процес отримання даних з різних джерел, таких як оперативні системи, бази даних, веб-сайти.
- Очищення даних - це процес підготовки даних для подальшого аналізу, включаючи видалення помилок, виправлення невідповідностей і стандартизацію даних.
- Аналіз даних - це процес виявлення закономірностей і тенденцій у даних, а також створення прогнозів.
- Візуалізація даних - це процес представлення даних у зрозумілому для людини вигляді, наприклад, у вигляді діаграм, графіків, карт.

Для вже програмної реалізації методів BI була використана служба аналізу SSAS.

SQL Server Analysis Services (SSAS) - це аналітичний рушій даних (VertiPaq), який використовується для підтримки прийняття рішень та бізнес-

аналітики. Він дозволяє створювати семантичні моделі даних корпоративного рівня для бізнес-звітів і клієнтських додатків, таких як Power BI, Excel, звіти Reporting Services та для інших інструментів візуалізації даних.

Встановлений як локальний або віртуальний екземпляр сервера, SQL Server Analysis Services підтримує табличні моделі на всіх рівнях сумісності (залежно від версії), багатовимірні моделі, інтелектуальний аналіз даних і Power Pivot для SharePoint.

Загальний робочий принцип застосування включає в себе встановлення екземпляра SQL Server Analysis Services, створення табличної або багатовимірної моделі даних, розгортання моделі як бази даних на сервері, обробку бази даних для завантаження її даними, а потім призначення дозволів для доступу до даних. Коли модель даних готова до роботи, до неї може отримати доступ будь-яка клієнтська програма, що підтримує Analysis Services як джерело даних. [8].

Для вже програмної реалізації методології BI для системи моніторингу річкової води, було використано середовище програмування Visual Studio 2022. Для створення та налаштування бази та сховища був використаний SQL Server та СУБД SQL Server Manager.

Для повної програмної реалізації для середовища розробки було додатково встановлені служби керування та обробки даних та застосунки для обробки та аналізу даних. Після чого був відповідно створений необхідний багатовимірний проект сервісів аналізу. Вже в ньому були налаштовані підключення до сховища даних, створені уявлення кубу, а також його виміри, згідно визначеної структури збереження історичних даних.

Наступним кроком було відповідне внесення даних, на основі яких і буде проведено дослідження. Програмне середовище Visual Studio дозволяє автоматизувати процес передачі даних із бази даних до сховища. Це було здійснено за допомогою окремого проекту – проекту сервісів інтеграції (Integration Services).

Створивши такий проект було визначено три кроки – рівні Control Flow на які ділитиметься весь алгоритм переведення даних до сховища (Рис. 6).

Рис. 6 Загальна послідовність Control Flow

Послідовність Control Flow, в даному випадку, описує процес перенесення та відповідно обробки даних із оперативного джерела – бази даних до багатовимірного кубу – сховища даних.

Спочатку до сховища будуть перенесенні незмінені таблиці вимірів 1 рівня (часовий вимір, вимір показників, водних ресурсів), після чого, також незмінені, таблиці вимірів 2 рівня (вимір станції), та на останньому етапі заповнена таблиця фактів вже агрегованими, обчисленими кінцевими значеннями.

Також, для прикладу, розглянемо детальніше відповідну деталізацію Flow заповнення таблиць вимірів 1 рівня – таблиць вимірів показників, часових періодів та водних ресурсів (Рис. 7).

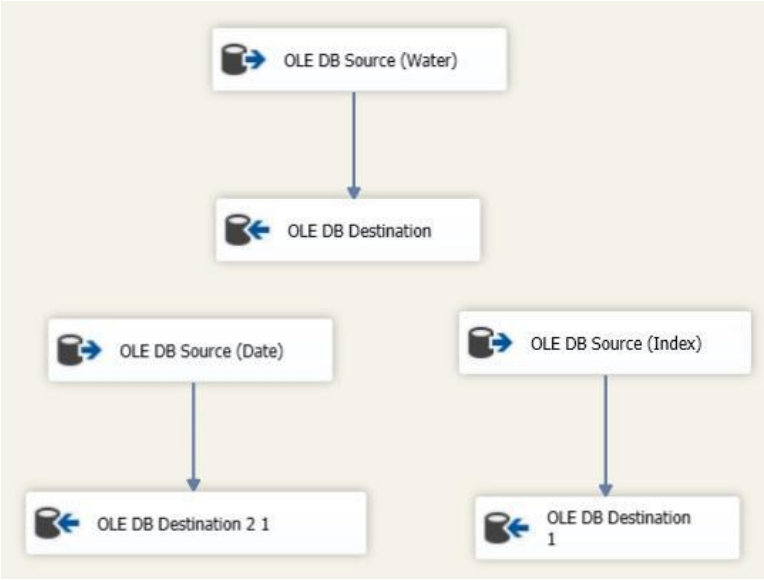

Рис.7 Заповнення вимірів 1 рівня

Інструменти служби SSAS дозволили точно визначити для кожного із окремих вимірів визначити порядок та властивості перенесення даних [24].

Як можна побачити для реалізації передачі даних між базою даних та сховищем у наведених Data Flow будуть використані 2 типи об'єктів: Data Source – виступатиме джерелом даних, та Data Destination що визначатиме куди дані із джерела будуть перенаправлені до сховища.

Відповідно, можна побачити, для кожної із таблиць джерела даних що відповідає за збереження вимірів – Водних ресурсів (Water), показників (Index), часового виміру (Date), було налаштовано зв’язок перенесення до необхідної таблиці у сховищі даних.

Кожен із зв’язків буде здійснювати роботу, використовуючи строки підключення які були програмно описані для відповідного підключення та зв’язку із джерелами даних. Остаточне місце, таблиці сховища, куди будуть перенесені дані, описані у своїх об’єктах місце призначень (Data Destination). В загальному, таблиці наведені на рисунку є дуже схожими по структурі у оперативному джерелі даних та сховищі, тому комплексних зв’язки тут не були використані.

Налаштувавши всі значення параметрів та зв’язки із базою та сховищем, алгоритм був запущений, та результатом роботи стали дані які автоматично були

перенесені із оперативного джерела до сховища, де вони в подальшому були використані для проведення комплексного аналізу параметрів та відповідно досягнення визначених цілей дослідження.

В цілому сховище даних включило в себе агреговану інформацію про:

- 16 сезонів – часовий період від осені 2020 року до зими 2023
- 9 хімічних показників, що характеризуватимуть якість води
- 18 станцій збору води
- 4 водні басейни – річки на яких знаходяться станції
- 2151 строку записів спостережень зафіксованих значень показників

3.2 Проведення початкового аналізу даних

Підготувавши дані та налаштувавши ефективне їх застосування та керування ними було проведено низку загальних експериментів дослідження. Метою такого початкового дослідження було виявлення можливих аномалій, закономірностей у даних дослідження. Також використання відносно тривіальних методів аналізу дозволить представити загальний опис спостережень, частоти та інші параметри.

3.2.1 Побудова матриці кореляції

Оцінка залежностей між хімічними показниками води може бути критично важливою для розуміння загального стану екосистеми річки, прогнозування можливих змін та запобігання забрудненню. Аналіз взаємозв'язків між такими показниками, як рівень сульфатів (SO4), хлоридів (CL), нітратів (NO3), рівень розчиненого кисню та іншими хімічними елементами, дозволяє виявляти джерела забруднення, їхні шляхи розповсюдження і можливий вплив на водні

організми.

Одним із інструментів загальної оцінки та візуалізації залежностей показників, що була використана є матриця кореляції.

Кореляційна матриця - це таблиця з коефіцієнтами кореляції (залежності) для різних змінних. Матриця показує, як усі можливі пари значень у таблиці пов'язані між собою. Це потужний інструмент для узагальнення великого набору даних, пошуку та відображення закономірностей у даних.

Її часто показують у вигляді таблиці, де кожна змінна вказана як у рядках, так і в стовпчиках, а коефіцієнт кореляції між кожною парою змінних записаний у кожній клітинці. Коефіцієнт кореляції коливається від -1 до +1, де -1 означає ідеальну негативну кореляцію (залежність), +1 означає ідеальну позитивну кореляцію, а 0 означає відсутність зв’язку між змінними.

Використовуючи матрицю кореляції ми можемо легко оцінити залежності та зв’язки між всіма показниками. Зважаючи на особливість дослідження, до змінних матриці кореляції також було додано показник загального забруднення – Індекс забруднення води (ІЗВ). Це додатково дозволить оцінити силу впливу кожного окремого хімічного показника на відповідно якість води в цілому.

Визначення сили впливу показників на індекс забруднення води є важливим аспектом для розуміння екологічного стану річкових систем, а також для формування ефективних стратегій збереження та управління водними ресурсами.

Кожен хімічний елемент або сполука у воді, впливає на якість води по-різному. Знання про те, наскільки суттєво кожен із цих показників впливає на загальний стан води, є ключем до комплексного і раціонального управління та моніторингу якості води.

Програмна реалізація створення матриці кореляції було здійснено використовуючи бібліотеку matplotlib.pyplot мови програмування Python у хмарному середовищі програмування Google Colab. Результат реалізації створення матриці можна побачити на Рис.8.

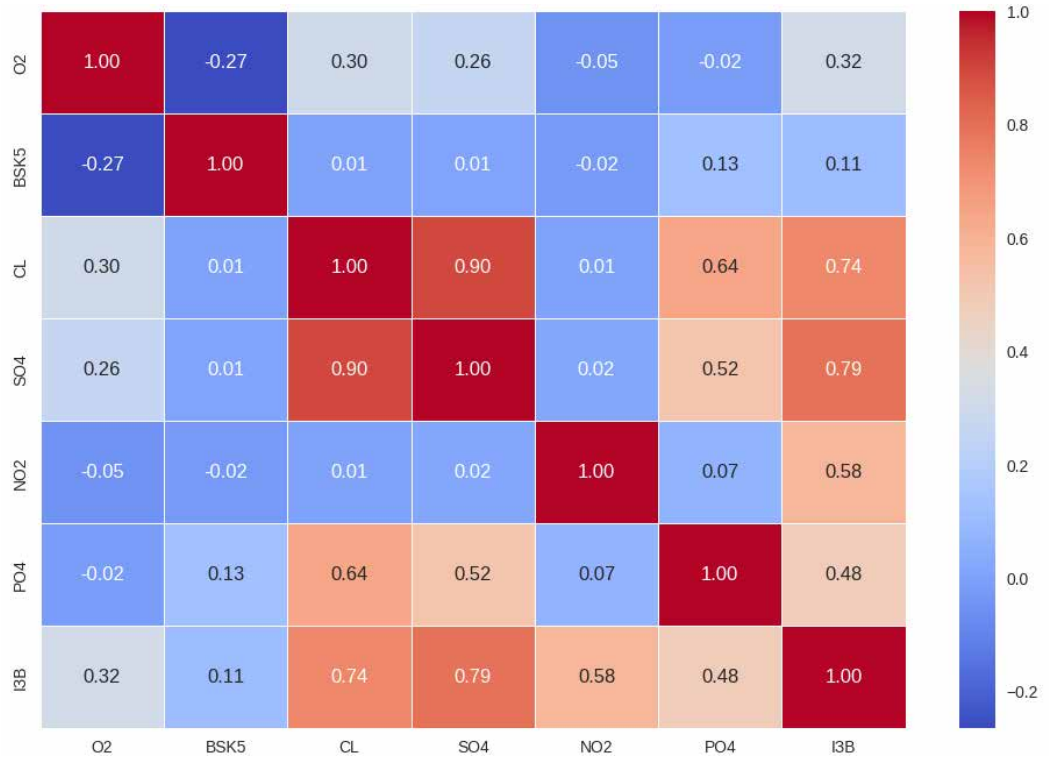

Рис.8 Матриця кореляції хімічних показників

Після проведення аналізу матриці, фокусуючись на силі впливу показників на індекс забруднення, можна зробити такі висновки:

- Вмісти кисню (O2) та Біохімічного споживання кисню (BSK5) дуже слабко впливають на загальне забруднення води
- Показники Хлоридів (CL) та Сульфатів (SO4) мають найбільший вплив на індекс забруднення, що робить ці елементи пріоритетними можливими джерелами під час моніторингу
- Показники Нітратів (NO2) та Фосфатів (PO4) також мають значний вплив на рівень якості води і показують, що вплив цих елементів не має бути ігнорований

Оцінивши основний вплив хімічних елементів води на показник забруднення, було проведено аналіз взаємозв'язків та залежності між окремими показниками. Отже, на основі аналізу матриці, додатково можна стверджувати:

- Негативна, несильна кореляція із коефіцієнтом -0.27 спостерігається між показниками Кисню та Біохімічного споживання кисню, що напряму вказує на те, що із підвищенням рівню споживання кисню, його рівень падає. Така залежність цих показників насамперед може показувати, що дані, які були використані для дослідження, є абсолютно реальними [25].

- Дуже сильна кореляція із коефіцієнтом 0.9 була виявлена між показниками Хлоридів та Сульфатів. Насамперед це може стверджувати про те, що джерело обох цих показників є одним і тим самим. Оцінивши джерела показників, можна стверджувати, що присутність обох іонів може бути наслідком скидання стічних вод промислових підприємств, особливо хімічних, металургійних. Вода може протікати через породи, багаті на солі, такі як гіпс і галіт, що сприяє одночасному розчиненню цих сполук і насиченню води сульфатами та хлоридами.

- Також було зафіксована кореляція середньої сили між показниками Хлору, Сульфатів та Фосфатів. Це, аналогічно, може свідчити про співпадіння джерел збудників цих речовин. Одним із можливих джерел можна відзначити використання мінеральних добрив, які містять сульфати, фосфати. Потрапляння цих речовин у водойми через змивання з полів, особливо після дощів або зрошення може пояснювати одночасне збільшення рівню цих показників.

Сильна кореляція між показниками вказує на можливу наявність одного або кількох із перелічених факторів забруднення. Щоб точно визначити джерело цього зв’язку, доцільно провести додатковий аналіз, враховуючи походження води, місцеві геологічні особливості, кліматичні умови та потенційні антропогенні джерела у розрізі розташування станцій збору води.

3.2.2 Алгоритм One Rule

Класифікація (Classification). Найпростіша і найпоширеніша задача в Data Mining. В результаті розв'язання задачі класифікації виявляються ознаки, що характеризують групи об'єктів досліджуваного набору даних - класи; на основі цих ознак новий об'єкт може бути віднесений до того чи іншого класу.

"One Rule" - простий, але точний алгоритм класифікації, який генерує по одному правилу для кожного предиктора в даних, а потім вибирає правило з найменшою сумарною помилкою в якості "єдиного правила". Щоб створити правило для предиктора, ми будуємо таблицю частот для кожного предиктора в порівнянні з мішенню. Було показано, що One Rule створює правила лише трохи менш точні, ніж найсучасніші алгоритми класифікації, при цьому створюючи правила, які легко інтерпретуються людиною [13].

Зважаючи на завдання та особливості дослідження показників води, було вирішено про проведення класифікації, використовуючи алгоритм One Rule, по трьом основним правилам:

- Для кожної із пори року – вода чиста (Рис. 9)

Пори року	В нормі	Перевищує норму	Загальна	Помилка
AUT	50	9	59	15,2542372881356
SPR	50	6	56	10,7142857142857
SUM	54	5	59	8,47457627118644
WIN	54	11	65	16,9230769230769

Рис. 9 Правила класифікації за порами року

Основна мета реалізації цієї класифікації – визначення сезонності даних та відповідно показань забруднення. Як можна побачити, ймовірність того що зчитане спостереження для кожної із пір року є відносно схожою і складає 84% – 92% ймовірності. На основі таких показників можна, казати про майже повну відсутність сезонності даних. Також можна чітко побачити, що загалом було зафіксовано відносно мало спостережень які характеризуються забрудненістю вод.

- Для кожної станції – вода чиста (Рис. 10)

Номер станції	В нормі	Перевищує норму	Загальна	Помилка
2	0	9	9	100
4	3	6	9	66,6666666666667
1	11	1	12	8,33333333333333
18	15	1	16	6,25
8	7	1	8	12,5
6	15	1	16	6,25
13	14	1	15	6,66666666666667
16	15	1	16	6,25
11	8	1	9	11,1111111111111
14	16	0	16	0
5	16	0	16	0

Рис. 10 Правила класифікації за станціями

Правило, на основі якого проводилася класифікація: на відповідній станції – вода буде зафіксована чиста. Як можна побачити із результатів, відсортованими за найменшою кількістю показників чистих спостережень, на 2 станції були зафіксовані виключно показники помірно забрудненої води, тому можна сказати, що при подальших зчитувань даних із станції є 100% вірогідність, що вода буде забруднена. Аналогічно, для станції 4, можна сказати, що вірогідність фіксування показників чистої води буде 34%.

Таким чином, цей метод дозволяє легко знаходити станції із найбільшими

показниками забрудненості, допоможе оцінити першочергові пріоритетні місця, що можуть містити джерела забруднень.

- Для кожного показника – вода чиста (Рис. 11)

Показник	В нормі	Перевищує норму	Загальна	Ймовірність забруднення
O2	208	31	239	12,9707112970711
BSK5	225	14	239	5,85774058577406
CL	239	0	239	0
SO4	213	26	239	10,8786610878661
NO2	231	8	239	3,34728033472803
PO4	239	0	239	0

Рис. 11 Правила класифікації за показниками

Для цієї класифікації було сформовано алгоритм, що фіксує який показник перевищував норму, та одночасно було зафіксовано спостереження води, що не входить до I та II класів забруднення, тобто чистих вод.

Як можна побачити, наприклад, у випадку якщо було зафіксовано перевищення рівню SO4 то із вірогідністю 10.8 % можна очікувати що вода буде забруднена. Також можна побачити, що рівні CL та PO4 не перевищували значень забруднення за весь період спостережень, що може свідчити насамперед про відсутність великих забрудників цих елементів.

Застосування цього методу в такий спосіб також дозволяє оцінювати відносну важливість показників, хоч і з дуже сумнівною точністю. Також можна чітко побачити, показники яких елементів перевищували допустимі норми найбільше, а які не були в критичному стані ніколи.

Використання методу One Rule для оцінки якості води може бути корисним для створення простих і зрозумілих правил класифікації, які можуть дозволити визначити, чи є вода якісною або забрудненою, на основі окремих хімічних показників, пір року або станцій.

3.3 Використання методів машинного навчання

3.3.1 Застосування моделі XGBoost

Extreme Gradient Boost (XGBoost) - це масштабований інструментарій для навчання на основі узагальненого дерева рішень з градієнтним прискоренням (GBDT). Це найкраща бібліотека машинного навчання для задач регресії, класифікації та ранжування, яка включає в себе паралельний бустинг дерев.

Gradient Boosting Decision Trees (GBDT) - це метод ансамблевого навчання на основі дерев рішень для класифікації та регресії, який можна порівняти з випадковим лісом. Щоб створити кращу модель, методи ансамблевого навчання поєднують різні методи машинного навчання.

Випадковий ліс і GBDT створюють модель з багатьма деревами рішень. Різниця полягає у способі побудови та з'єднання дерев. Випадковий ліс будує всі дерева рішень паралельно, використовуючи випадкові бутстрап-вибірки з набору даних за допомогою методу, який називається пакуванням. Середнє значення всіх прогнозів дерева рішень використовується для отримання остаточного прогнозу.

Градієнтний бустінг - це процес "підсилення" або покращення однієї слабкої моделі шляхом об'єднання її з кількома додатковими слабкими моделями для створення спільної сильної моделі. У градієнтному бустінгу, який є розширенням бустінгу, підхід адитивної побудови слабких моделей визначається як алгоритм градієнтного спуску. Щоб зменшити кількість помилок, градієнтний бустінг визначає очікувані результати для наступної моделі. Градієнт помилки (звідси і назва градієнтного бустінгу) щодо прогнозу визначає цільові результати для кожного випадку [14].

Великою перевагою використання моделі з градієнтним підсиленням є те, що після побудови дерев з підсиленням відносно просто отримати оцінку важливості для кожної ознаки. Важливість ознаки обчислюється для одного

дерева рішень за величиною, на яку точка розщеплення кожного атрибута покращує показник ефективності, зваженою на кількість спостережень, за які відповідає вузол. Іншими словами, ви просто підсумовуєте, наскільки розбиття кожної ознаки дозволило вам зменшити домішки в усіх точках розбиття в дереві. Потім важливість ознаки усереднюється по всіх деревах рішень в моделі. Як вбудовану функцію, досить просто створити код, який відображає важливість функції в моделі [15].

Для досягнення цілей дослідження, за допомогою мови Python було реалізовано навчання моделі машинного навчання цим класифікатором у хмарному середовищі програмування Google Colab (Рис. 12).

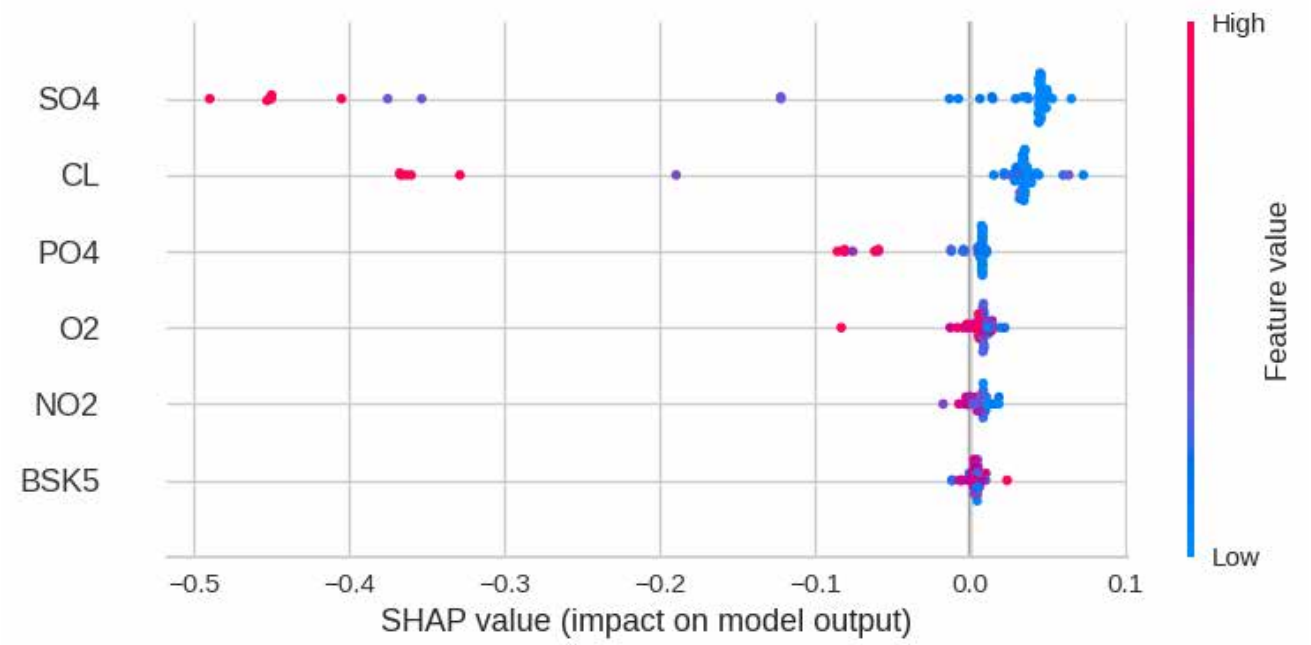

Рис. 12 Важливість ознак хімічних показників (XGBoost)

Із графіку важливості можливо зробити такі висновки:

- Найвпливовішим, а отже і найважливішим, показником є рівень Сульфатів - SO4, він має найбільше відношення впливу на властивість чистоти води.

- Другим за важливістю є показник Хлору - CL, його найнижчі зафіксовані спостереження мають найбільший вплив саме на висновок, що вода буде чиста (показник SHAP найбільший для низького рівня показника).
- Показник Фосфатів – PO4 характеризується середньою силою впливу, великий вміст фосфатів характеризується відносно посереднім коефіцієнтом сили впливу -0.8, в той час як малий вміст зовсім незначним чином впливає на відношення спостережень до чистоти.
- Рівні вмісту Кисню, Нітратів та Біохімічного споживання кисню мають найменший вплив на індекс забруднення. За визначенням моделі, навіть великі показники цих елементів мають майже нульовий коефіцієнт загального впливу.

Реалізоване рішення побудови та візуалізації моделі також дозволяє проводити деталізацію загального графіку, обравши необхідні показники для аналізу. Один із таких дочірніх графіків деталізації важливості можна побачити на Рис. 13. Він розширює графік важливості ознак, у якому були виділені показники іонів Сульфатів та Хлоридів.

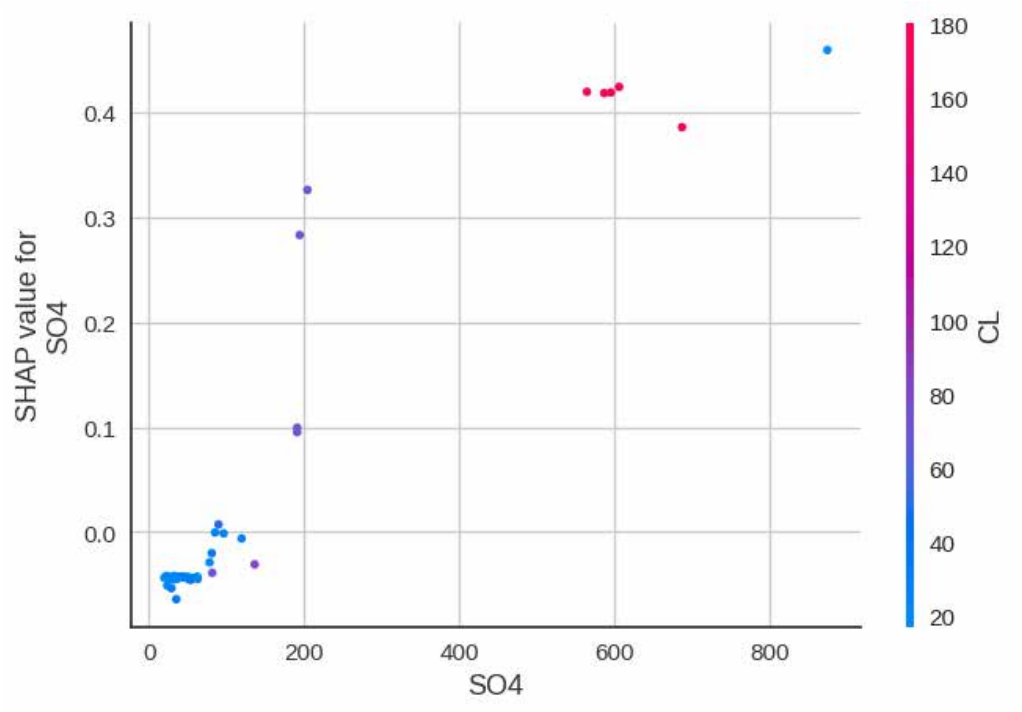

Рис. 13 Дочірній графік важливості (залежності Хлоридів та Сульфатів)

Як можна побачити, одночасно великі значення сульфатів та хлоридів, характеризуються найбільшим коефіцієнтом впливу на якість води. Цей результат в цілому співпадає із отриманими раніше результатами, із матриці кореляції. З іншого боку низькі показники обох іонів негативно впливають на індекс забруднення, тобто низький їх рівень, логічно, може описувати спостереження чистих вод.

Також додатково можна побачити загальну картину розподілу рівнів показників. Як можна побачити, показник Хлоридів не перевищував вмісту 180, а Сульфатів 850. Так як допустима концентрація цих елементів, згідно офіційної класифікації, відповідно визначена як 300 та 150 можна стверджувати, що вміст хлору є в повній нормі протягом всіх спостережень, та його перевищення не було зафіксовано. З іншого боку, показник Сульфатів перевищував норму відносно багато разів, тому можна казати про можливе забруднення. Але для визначення точних висновків, необхідно звісно проведення подальшого аналізу та досліджень.

Модель XGBoost надає точний, комплексний та інформативний підхід до оцінки сили впливу хімічних показників на загальну якість води. Вона допомагає пріоритезувати ключові забруднювачі, забезпечує обґрунтовані висновки для управлінських рішень і може впевнено використовуватися у системі моніторингу якості води.

3.3.2 Побудова дерева рішень

Наступним методом машинного навчання, що був використаний та реалізований під час дослідження є алгоритм Дерева Рішень (Decision Tree).

Дерево рішень - це непараметричний алгоритм керованого навчання для задач класифікації та регресії. Воно має ієрархічну деревовидну структуру, що складається з кореневого вузла, гілок, внутрішніх вузлів і листових вузлів. Дерева рішень використовуються для задач класифікації та регресії, створюючи легкі для розуміння моделі.

Дерево рішень - це ієрархічна модель, що використовується в підтримці прийняття рішень, яка відображає рішення та їх потенційні результати, включаючи випадкові події, витрати ресурсів та корисність. Ця алгоритмічна модель використовує умовні керуючі оператори і є непараметричною, керованою, корисною для задач класифікації та регресії. Деревоподібна структура складається з кореневого вузла, гілок, внутрішніх вузлів і листових вузлів, утворюючи ієрархічну деревоподібну структуру.

Деякі переваги дерев рішень:

- Прості для розуміння та інтерпретації. Дерева можна візуалізувати.
- Потребують невеликої підготовки даних. Інші методи часто вимагають нормалізації даних, створення фіктивних змінних і

видалення порожніх значень. Деякі комбінації дерев і алгоритмів підтримують пропущені значення.

- Вартість використання дерева (тобто прогнозування даних) логарифмічно залежить від кількості точок даних, використаних для навчання дерева.
- Здатний обробляти як числові, так і категоріальні дані. Інші методи зазвичай спеціалізуються на аналізі наборів даних, які містять лише один тип змінних. Дивіться алгоритми для отримання додаткової інформації.
- Використовує модель білої скриньки. Якщо певна ситуація спостерігається в моделі, пояснення умови легко пояснюється за допомогою булевої логіки. В той час як навпаки, в моделі чорної скриньки (наприклад, у штучній нейронній мережі) результати може бути складніше інтерпретувати.
- Можна перевірити модель за допомогою статистичних тестів. Це дає можливість врахувати надійність моделі.
- Добре працює, навіть якщо його припущення дещо порушуються справжньою моделлю, на основі якої були отримані дані [16].

Керуючись принципами побудови та аналізу дерев рішень, також використавши мову програмування Python було реалізовано використання та візуалізація побудови моделі дерева рішень (Рис. 14)

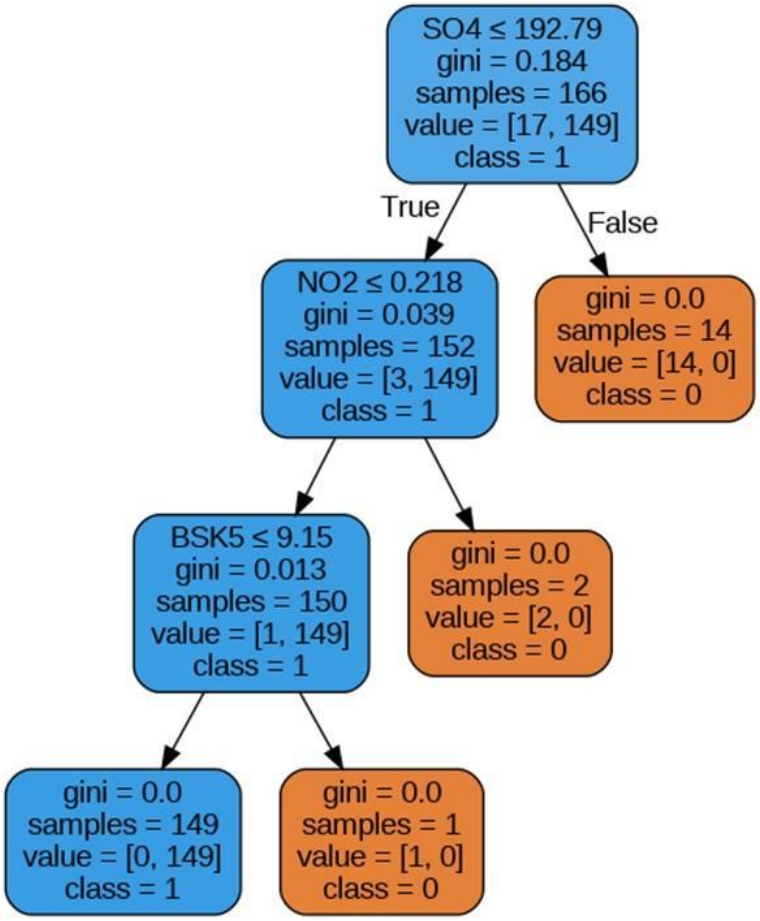

Рис. 14 Дерево рішень

Як можна побачити, після виконання та побудови, моделлю було виділено два класи: клас 1 – вода чиста, та клас 0 – вода помірно забруднена.

Переглянувши результат візуалізації, можна відразу відзначити велику перевагу значень класу 1 – чиста вода. Із побудованого рішення можна зробити наступні висновки:

- Показник SO4 також був визнаний дуже впливовим показником.
- У разі перевищення SO4 значення 192.79 є приблизно 90% ймовірність вважати що рівень забруднення води, із таким значенням, буде вищим.
- У випадку якщо SO4 менше 194 але NO2 більше 0.218 то можлива 90% ймовірність, що вода буде відноситися до класу помірно забрудненої

- Якщо SO4 менше 194 водночас NO2 менше 0.218 та BSK5 менше 9.15 то ймовірність, що спостереження буде відноситися до класу чистої води складає 99%.

Використання методу побудови дерева рішень забезпечує зрозумілий та ефективний підхід для аналізу та прогнозування якості води на основі її хімічних параметрів. Завдяки застосуванню цього методу можливе проведення прогнозування на основі невеликої кількості ключових показників.

Він дозволив збудувати чіткі правила для класифікації, визначити найбільш впливові показники, а також надати прості й інтуїтивно зрозумілі результати для дослідження.

3.3.3 Кластеризація

Кластеризація – це технологія машинного навчання, яка використовується для групування точок даних. Подаючи на вхід набір точок даних, ми можемо використовувати алгоритм кластеризації для класифікації кожної точки даних у конкретну групу. Точки даних, що знаходяться в одній групі, повинні мати схожі властивості, тоді як властивості точок даних у різних групах мають істотно відрізнятись. Кластеризація відноситься до методів навчання без вчителя і є методикою статистичної обробки даних, що використовується в багатьох галузях.

Загалом, кластеризацію можна розділити на дві підгрупи:

1. Жорстка кластеризація. При жорсткій кластеризації кожна точка даних або належить кластеру повністю, або не належить зовсім.

2. М'яка кластеризація (нечітка кластеризація). При м'якій кластеризації замість того, щоб відносити деяку точку даних в окремий кластер, знаходиться ймовірність того, що ця точка даних буде в кожному з цих кластерів.

Для кластеризації даних застосовуються різні методи, але ціль цих методів однакова – об’єднати схожі за характеристиками об’єкти у групи. Для кожного методу кластеризації можна визначити модель кластеру та алгоритм кластеризації.

Для досягнення цілей дослідження було вирішено використовувати метод побудови моделі k-середніх. Метод k-середніх – найбільш відомий алгоритм кластеризації. Цей алгоритм будує задане число кластерів, розташованих якнайдалі один від одного. Робота алгоритму ділиться на кілька етапів:

1. Спершу випадковим чином задаємо масив з k центральних точок. Щоб оцінити значення k, яке необхідно використовувати, достатньо переглянути дані та спробувати визначити кількість окремих угрупувань.

2. Для кожної точки даних обчислюємо відстань між цією точкою та кожною з центральних точок, а потім відносимо точку до групи з найближчим до неї центром.

3. Виходячи з класифікованих точок, ми перераховуємо центр групи, беручи середнє значення всіх векторів групи.

4. Повторюємо кроки 2-3 задану кількість разів або до тих пір, поки зміна центрів груп на кроці 3 не буде менше наперед заданого значення [17].

Реалізація навчання моделі використовуючи метод кластеризації також була реалізована із використанням мови програмування Python із застосуванням бібліотеки pycaret (Рис. 15).

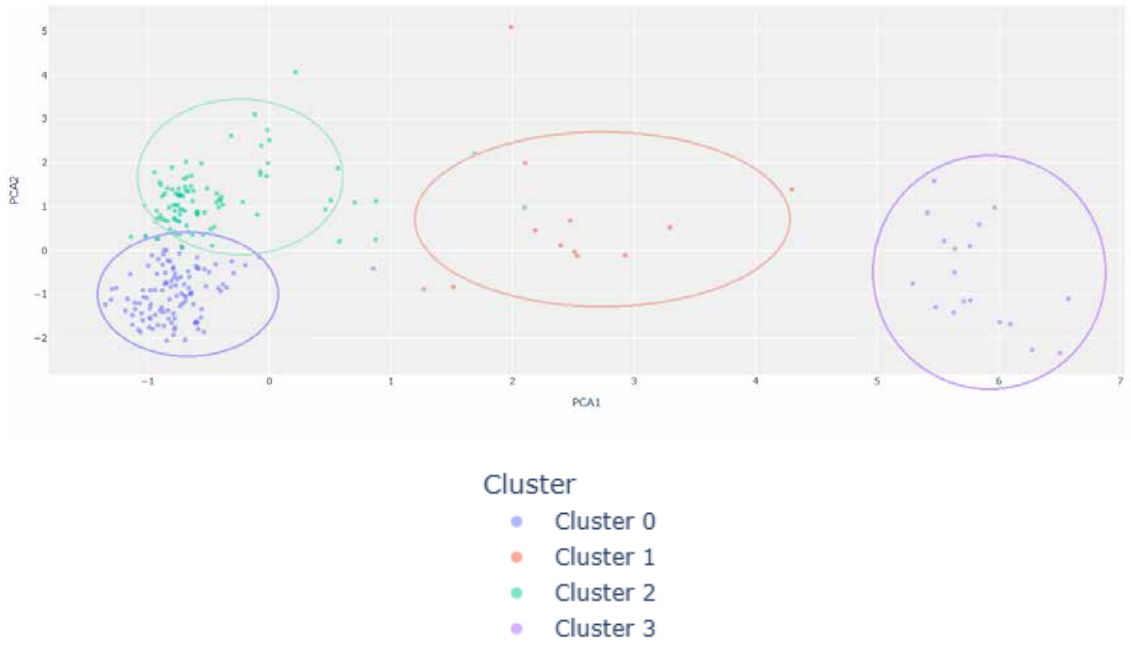

Рис. 15 Результат використання методу кластеризації

Інструменти використаної бібліотеки дозволяють детально оглядати вміст кластерів, що й було використано для опису кластерів та спостережень які до них входять. Провівши подальший аналіз моделі, можна зробити такі висновки щодо створеної моделі:

- Кластери 2 і 0 включають в себе переважну більшість спостережень, вони характеризуються виключно показниками класу чистої води, із нормальними та низькими рівнями вмісту всіх хімічних елементів.

- Додатково, невелика кількість даних спостережень кластеру 2 містять записи із підвищеним рівнем показника Сульфатів на 4 станції.

- Кластер 1 характеризується показниками із підвищеним рівнем збудників, що вже відносять до рівню «вода забруднена». Спостереження індексу забруднення є відносно середнього рівня. Хімічні показники іонів Хлору та Сульфатів мають підвищений рівень. До кластеру входять спостереження 4, 8 та 6 станцій.

- Кластер 3 містить в собі спостереження даних із найбільшим зафіксованим індексом забруднення, по суті повністю відноситься до класу забрудненої води. Характеризується особливо критичними рівнями іонів Хлоридів, Фосфатів та Сульфатів. Станції 2 та 3 єдині, що входять до даного кластеру.

Провівши аналіз кожного із кластерів були виявлені схожі патерни та аномалії в даних, які дали змогу стверджувати: Переважна більшість показників (близько 80%) відносять до класу чистої води, що вже каже про відсутність критичних проблем забруднення на станціях, що відносять до кластерів 0 та 2; На станціях 4, 6 та 8 було виявлено підвищені рівні Сульфатів, із перевищенням норми майже вдвічі, що може вказувати на наявність збудників, та говорить про відносно низьку загальну якість води; Найбільш забрудненими були визначені спостереження із станцій 2 та 3, із високими показниками Хлоридів та Фосфатів, ці станції можна визначити як першочерговими, пріоритетним буде виконання заходів із стримування саме на цих станціях.

Використання моделі кластеризації дозволило провести поглиблений аналіз якості води, виявити групи з подібними характеристиками та знайти аномалії. Це ефективний інструмент для моніторингу, прийняття управлінських рішень і розробки стратегій очищення та збереження водних ресурсів.

3.4 Побудова звітів моніторингу в середовищі Power BI

Звітна інформація є важливою частиною будь-якої системи моніторингу. Вона дозволяє відстежувати стан системи, виявляти відхилення від норми та приймати необхідні заходи. Звітна інформація може бути представлена у різних форматах, таких як таблиці, графіки, діаграми, карти.

Важливо, щоб звітна інформація була точною, актуальною та зрозумілою для користувачів. Вона повинна бути доступна вчасно, щоб компанії могли приймати рішення на основі свіжих даних. Звітна інформація є цінним ресурсом для будь-якої системи. Вона може допомогти компаніям підвищити свою ефективність, прийняти правильні рішення та досягти своїх цілей.

Power BI - це хмарна служба бізнес-аналітики від Microsoft, яка дозволяє будь-кому створювати візуалізації та аналізувати дані із швидкістю та ефективністю. Це потужний і водночас гнучкий інструмент бізнес-аналітики для роботи з різноманітними даними та їхнього аналізу.

Багато компаній навіть вважають його незамінним для роботи, пов'язаної з Data Mining. Простота використання Power BI пояснюється тим, що вона має інтерфейс перетягування. Ця функція допомагає виконувати такі завдання, як сортування, порівняння та аналіз, дуже легко і швидко. Power BI також сумісний з різними джерелами, включаючи Excel, Microsoft SQL Server і хмарними сховищами даних, що робить його відмінним вибором для роботи із великою кількістю комплексних даних.

Power BI дає можливість аналізувати та досліджувати дані як локально, так і в хмарі. Power BI надає можливість легко і безпечно співпрацювати і ділитися налаштованими інформаційними панелями та інтерактивними звітами з колегами і організаціями [18].

Використання широких та ефективних інструментів Power BI і стане основою створеної системи звітності дослідження моніторингу води. Завдяки зручному процесу інтеграції компонентів Microsoft, не буде ніяких проблем інтеграції вже створеного сховища даних, на основі Microsoft SQL Server, із

середовищем Power BI. Для досягнення цього була сформована строка підключення, яка була використана для перенесення даних із сховища до функціоналу Power BI.

Використання методів Power BI дозволити створити багатофункціональні, інтерактивні листи звітів, частини яких будуть розглянуті нижче. Одним із сформованих звітів є візуалізація історичних змін показнику забруднення ІЗВ протягом всього часу дослідження із можливістю вибору необхідного водного ресурсу, в цьому випадку річки Дністер (Рис. 16).

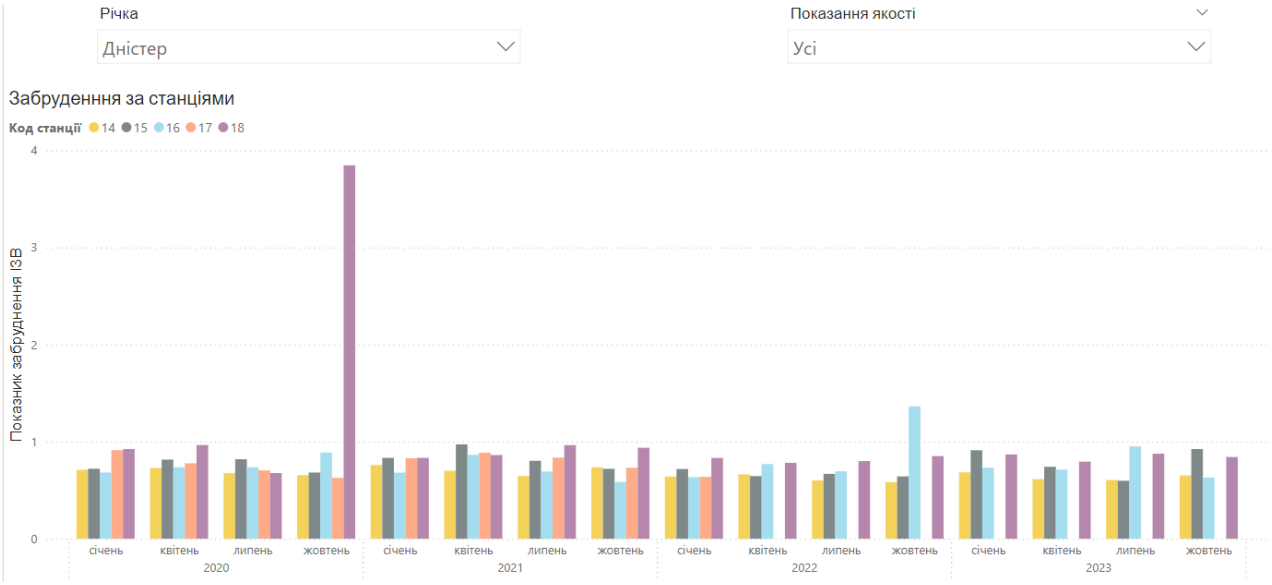

Рис. 16 Графік зміни індексу забруднення

Як можна побачити, як і було визначено також раніше, більшість спостережень ніколи не перевищували рівень забруднення 1. Згідно офіційної класифікації, індекс забруднення, що не перевищує значення 1, може відноситися до класу «Вода чиста». Але доволі очевидні 2 спостереження, що перевищують 1, один восени 2020 року на 18 станції, інший, на станції 16, восени 2022 року. Розглянемо детальніше одну із цих аномалій.

Інструментарій Power BI дозволяє проводити деталізації, отримуючи повні данні про обрані спостереження. Цю функцію і було використано для представлення деталізації першої критичної аномалії восени 2020 року (Рис. 17).

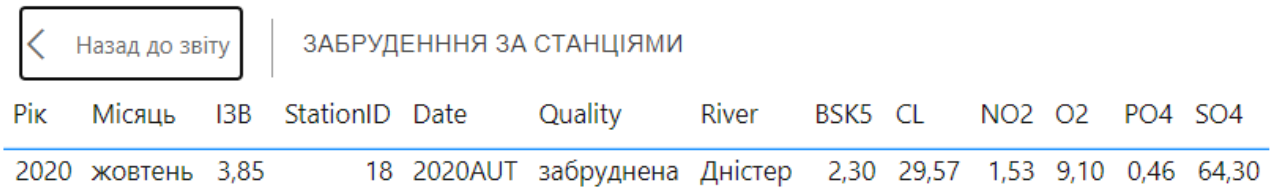

Рис. 17 Спостереження осені 2020 року 18 станції

Отже, проаналізувавши показники, можна сказати про те, що такі показники, як Хлор, Сульфати, Фосфати та Біохімічне споживання кисню є в повній нормі, та відповідає рівням стандартів. З іншого боку, рівень кисню підвищений вдвічі, але не цей показник став критичним.

Увагу привертає показник Нітритів (NO2), його значення було зафіксоване 1.53, із офіційною граничною допустимою концентрацією цього елемента в 0.08, що становить підвищення норми в 19.1 разів. Таке підвищення є неприпустимо критичним забрудненням, неприпустимим для визначення цієї води як чистої.

Але, є декілька інших фактів, що можуть вказувати на штучність цього показника. Якщо звернути увагу на наступний сезон спостережень, то дуже легко помітити, що абсолютно ніяких слідів такого критичного забруднення на цій самій станції, а також на сусідніх, виявлено не було. Індекс забруднення наступного сезону (зимою 2021) залишився на приблизно тому самому рівні як і відносно попередньому (літом 2020), що може свідчити про просту помилку в записі або зчитуванні даних зі станції. Також ніяких аномалій забруднення не помітно і на інших станціях цієї річки, вони залишаються на десь тих самих рівнях, що також може підтвердити теорію про помилку в записі даних.

В загальному можна сказати, навіть якщо це є помилковим спостереженням, можна побачити що такий спосіб візуалізації дозволяє легко виявляти аномалії, співпадіння. Використання такої звітності дозволяє створювати основу для ухвалення рішень, підвищення прозорості процесів і підготовки рекомендацій, спрямованих на покращення екологічного стану.

Продовжуючи аналіз, з метою візуалізації статистичних даних показників,

із використанням інструментів візуалізації Power BI, було реалізовано візуалізацію графіків box plot (Рис. 18).

В описовій статистиці box plot або коробковий графік (також відома як секторна діаграма або ящик з вусами) - це тип діаграми, що часто використовується в аналізі даних. Секторні діаграми візуально показують розподіл числових даних і асиметрію шляхом відображення квартилів (або процентилів) даних та інших статистичних показників.

Секторні діаграми показують п'ятицифрове зведення набору даних: мінімальний показник, перший (нижній) квартиль, медіана, третій (верхній) квартиль і максимальний бал [19].

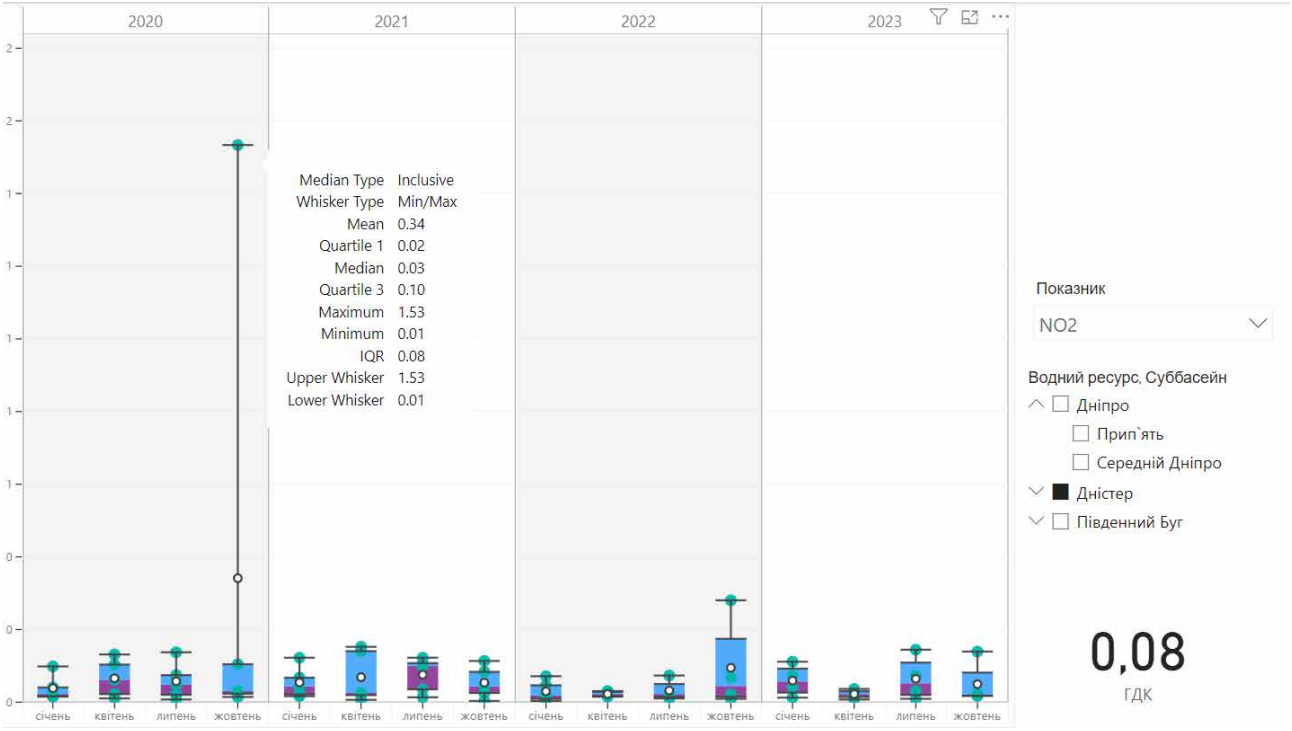

Рис. 18 Box plot

Для поглиблення дослідження вже окремих показників та їх рівнів, було відповідно реалізовано візуалізацію, яка дозволяє оцінювати значення обраних показників, на обраному водному ресурсі – річці.

Продовжуючи аналіз того самого аномального спостереження зафіксованого на річці Дністер, на Рис. 18 можна побачити статистичний

загальний графік зміни показнику Нітритів на річці Дністер. Аналогічно, бачимо значну аномалію восени 2020 року, також із максимальним показником 1.53. Відразу на цьому графіку можемо побачити, що у майбутньому після цього спостереження рівень цього показнику зовсім не змінився, і залишився в межах норми, що також може підтверджувати теорію про «штучність» цієї аномалії.

Так само використання коробкового графіку дозволить оцінювати рівні всіх потрібних для цього показників, на визначених станціях та річках.

Іще одним типом створеної звітності та часткового прогнозування є звіт статистики всіх показників також за обраним водним ресурсом та окремою річкою (Рис. 19). Також, функції Power BI дозволили включити прогнозування зміни показників на наступні 2 роки.

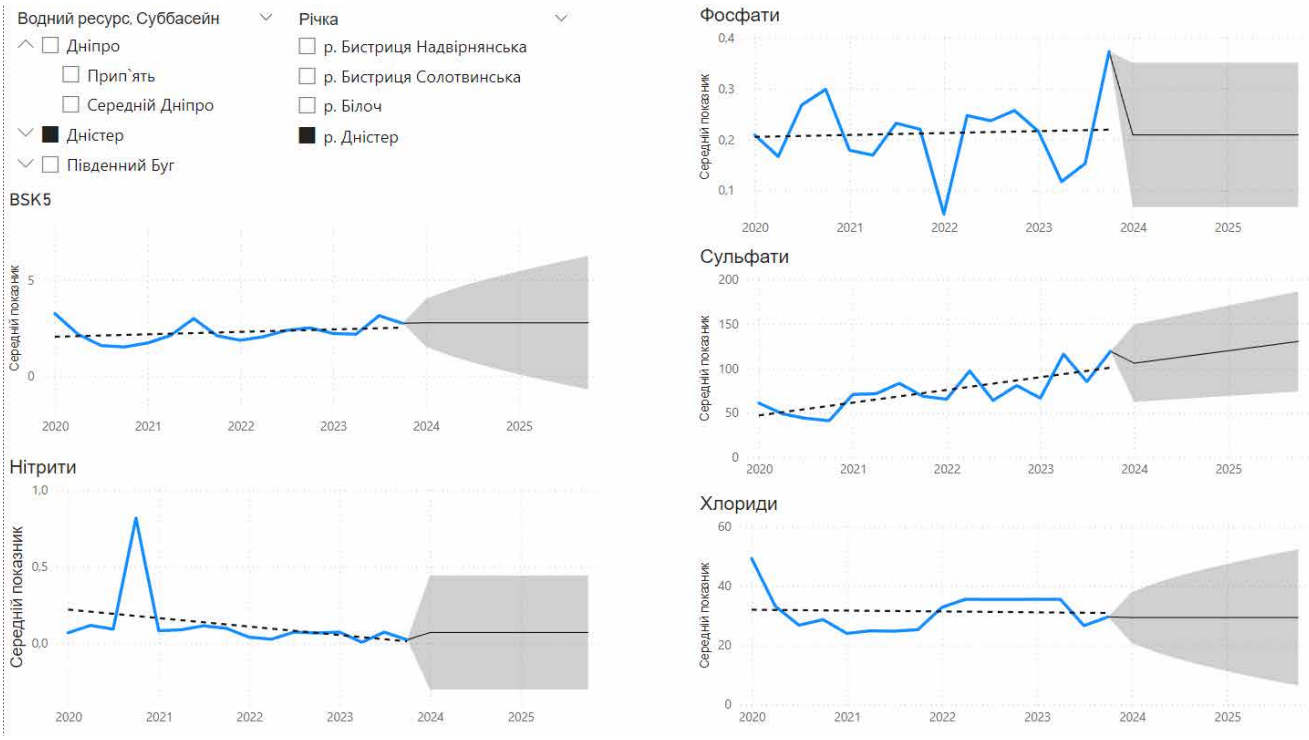

Рис. 19 Графіки рівнів показників та прогнозування

Використання такого звіту дозволяє отримати загальну картину вмісту води, по всім показникам. Як можна побачити вміст таких речовин, як хлоридів, біохімічного споживання кисню, нітритів та фосфатів протягом часу досліджень в цілому залишилися на тому самому рівні, що може свідчити про відсутність

зафіксованих забруднень. З іншого боку, рівень фосфат іонів під час цього часового періоду збільшував свою концентрацію, так як демонструє графік, можна передбачати продовження збільшення його вмісту.

Це може свідчити про наявність стороннього забруднення, та може бути основою для продовження моніторингу цього показника та можливого прийняття управлінських рішень щодо вжиття заходів із контролю за станом води.

Також, для продовження аналізу передбачень можливих рівнів забруднення використовуючи інструменти візуалізації, було також побудовано загальний графік передбачення зміни індексу забруднення (Рис. 20).

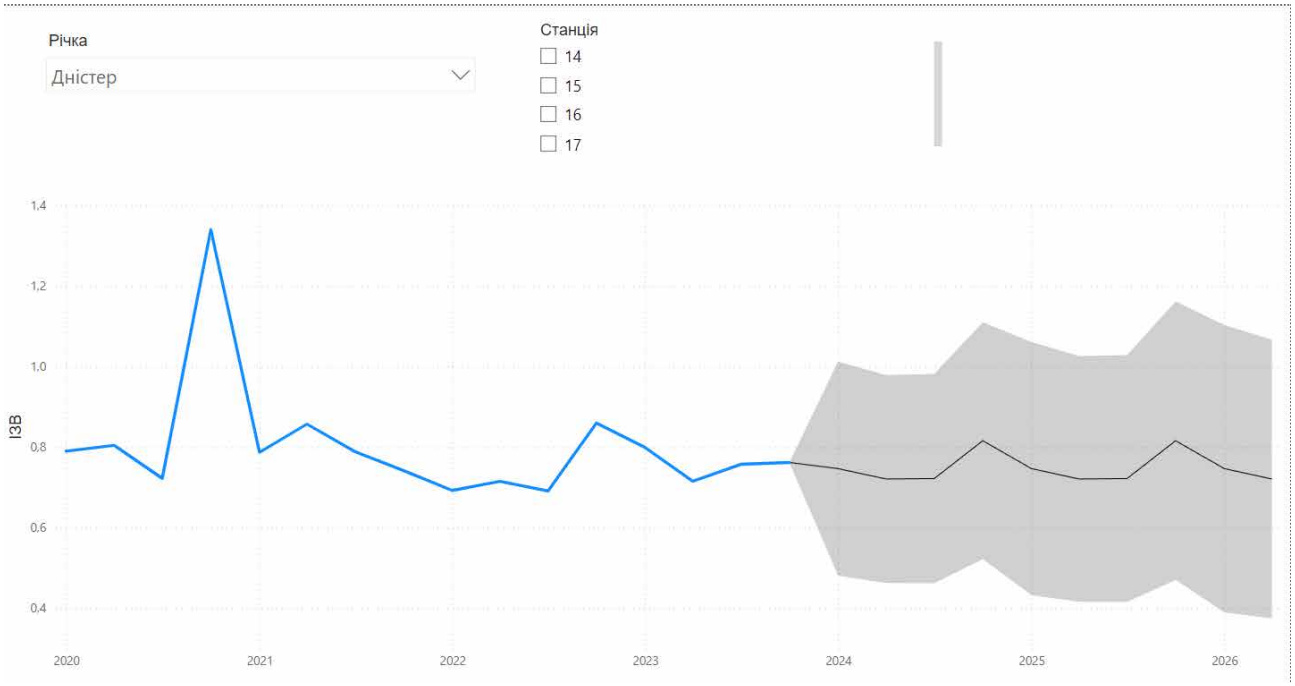

Рис. 20 Прогнозування показнику забруднення

Прогнозування індексу забруднення є важливим елементом моніторингу води, оскільки дозволяє передбачати зміни якості води та приймати превентивні заходи для запобігання критичним екологічним ситуаціям. Графік прогнозування відіграє ключову роль у візуалізації результатів і прийнятті обґрунтованих рішень.

Як можна побачити, навіть на загальному графіку забруднення, визначена аномалія восени 2020 року має дуже значний вплив. Але в цілому, передбачення говорить про коливання індексу забруднення в межах від 0.5 до 1.3 протягом наступних 3 років, із збереженням середнього показника забруднення 0.8.

Формування звітної інформації є одним із ключових елементів системи моніторингу річкової води, оскільки вона забезпечує зручний і зрозумілий спосіб представлення результатів аналізу якості води для ухвалення рішень або проведення подальших досліджень.

3.5 Розрахунок KPI

Ключові показники ефективності (KPI) - це управлінський інструмент або засіб, за допомогою якого можна відстежувати діяльність або процес, контролювати його (якщо він відхиляється, то можна розпізнати причину і виправити її) і забезпечити досягнення бажаних результатів.

Одним із способів досягнення хороших показників в оцінці роботи співробітників є використання методу ключових показників ефективності (KPI). KPI порівнюють те, що було створено, з тим, що було визначено. Успішне впровадження буде залежати від реалізації хорошої стратегії обслуговування відповідно до того, що було визначено.

Ключові показники ефективності (KPI) - це ряд ключових показників, які можна виміряти і які надають інформацію про те, якою мірою наприклад організація досягла успіху в досягненні стратегічних цілей. Елементи, що містяться в KPI, складаються зі стратегічних цілей, ключових показників, що мають відношення до цих стратегічних цілей, цілей, які є орієнтиром, і часових рамок або періоду виконання KPI.

KPI є одним з декількох основних інструментів управління, KPI формуються з наступними цілями:

1. Поєднання цінностей бачення та місії, організації стратегії для досягнення очікуваних цілей діяльності

2. Вимірювання ефективності, чи є значне збільшення або зменшення визначених параметрів
3. Порівняння поточної ефективності діяльності з минулими показниками діяльності, або порівняння з показниками інших організацій, щоб поточна організація отримала уявлення про сильні та слабкі сторони організації порівняно з конкурентами
4. Організаційні KPI використовуються як основа для визначення показників або робочих цілей підрозділів або окремих осіб [20].

Використання KPI дозволяє систематично вимірювати, аналізувати та вдосконалювати моніторинг, забезпечуючи стабільну якість водних ресурсів та оперативне управління ними. Для обчислення показників KPI також було застосовано інструменти Power BI. Першим із показників, що був сформований це загальна оцінка

Для коректного обчислення показнику ефективності у середовищі було виставлено такі параметри:

- Налаштовано режим оберненої візуалізації – якщо показник менше мети – це є позитивним результатом
- Метою було виставлено значення 1, тобто вважаємо всі значення які перевищують цю мету – негативним результатом, тобто показник вказує на забруднення води
- Обчислення розраховується на основі середнього значення всіх показників забруднення зафіксованим за визначений період

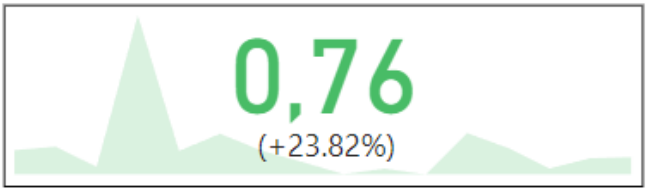

Рис. 21 Загальний показник ефективності забруднення для річки Дністер

Як видно із обчисленого показника, в цілому протягом всього проміжку спостережень зберігається позитивна динаміка забруднення, що характеризується 24% зменшенням забруднення. Це дозволяє стверджувати про відносну чистоту вод ріки Дністер на визначеному часовому проміжку.

Додатково також було імплементований розрахунок KPI для кожного із показників (Рис. 22), для забезпечення правильної реалізації, за мету було встановлені показники офіційних гранично допустимих концентрацій, які також були використані під час попередніх досліджень.

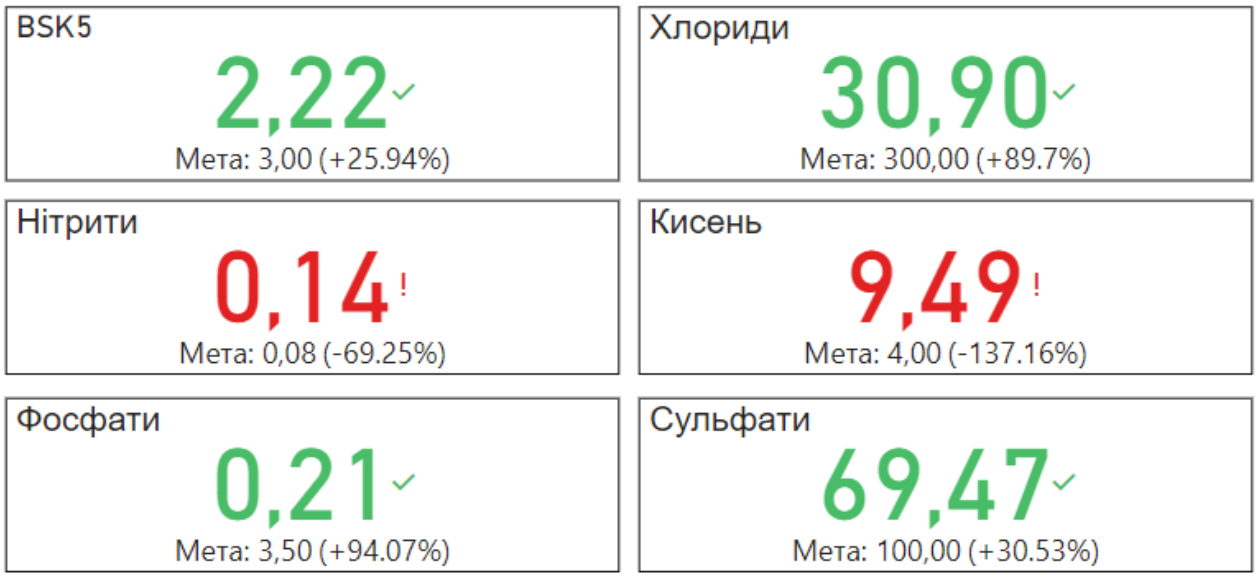

Рис. 22 Показники KPI для значень показників ріки Дністер

Продовжуючи дослідження аномальної ділянки, було вибрано той самий відрізок ріки Дністер. Згідно сформованих показників KPI, чітко видно не тільки зафіксоване перевищення середнього вмісту нітритів на 70% а й перевищення рівню вмісту кисню на 137%.

В цілому, використання показників KPI дозволяє забезпечити ефективність та прозорість системи моніторингу річкової води, а також може допомогти своєчасно приймати рішення для збереження екологічного балансу.

Це критично важливо для сталого управління водними ресурсами та захисту екосистем.

ВИСНОВОК

В ході виконання магістерської кваліфікаційної роботи було проведено аналіз предметної області, проектування та управління оперативним джерелом даних та сховищем даним, згідно методології OLAP. Використовуючи реалізований багатовимірний куб, на основі СУБД Microsoft SQL Server, були застосовані різні методи та підходи до аналізу, що включать:

- Побудову матриці кореляції, для оцінки залежностей між хімічними показниками води
- Алгоритм One Rule для проведення початкової задачі класифікації спостережень
- Методи машинного навчання, XGBoost, дерево рішень та використання кластеризації, використовуючи мову програмування Python

Для оцінки загальних статистичних показників та представлення візуалізацій, використовуючи середовище Power BI, були побудовані різноманітні звіти. Звітна інформація надає змогу оцінювати як історичні зміни показників та рівня забруднення, так і дозволяє будувати передбачення моделей графіків.

Обчислені показники KPI показали наскільки загальні рівні забруднення обраних річок відповідають стандартам якості.

У процесі дослідження системи моніторингу показників річкової води було проаналізовано основні аспекти організації, обробки та інтерпретації даних, що стосуються якості води. Особливу увагу приділено важливості застосування сучасних методів аналітики, таких як машинне навчання, прогнозування, виявлення залежностей між хімічними показниками та формування звітної інформації.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Аналіз показників, що визначають якість вод річок на території Рахівського району [Електронний ресурс] – Режим доступу до ресурсу – https://www.researchgate.net/publication/344207298_Analiz_pokaznikiv_so_viznacaut_akist_vod_ricok_na_teritorii_Rahivskogo_rajonu
2. Моніторинг поверхневих вод [Електронний ресурс] – Режим доступу до ресурсу – https://kc.pnu.edu.ua/wp-content/uploads/sites/11/2020/09/%D0%9B%D0%B5%D0%BA%D1%86%D1%96%D1%8F-1-%D0%9D2%D0%9E-n.pdf
3. Оцінка якості поверхневих вод правобережних приток басейну Прип'яті у Волинській області [Електронний ресурс] – Режим доступу до ресурсу – https://evnuir.vnu.edu.ua/bitstream/123456789/9263/3/Netrobchuk.pdf
4. Побудова діаграм варіантів використання [Електронний ресурс] – Режим доступу до ресурсу – http://www.tsatu.edu.ua/kn/wp-content/uploads/sites/16/laboratorna-robota-5-diahramy-variantiv-vykorystannja.pdf
5. Поняття архітектури інформаційних систем [Електронний ресурс] – Режим доступу до ресурсу – https://elearning.sumdu.edu.ua/free_content/lectured:de1c9452f2a161439391120eef364dd8ce4d8e5e/20160217112601/170352/index.html
6. Порівняльний аналіз методів побудови OLAP-систем із використанням засобів MS SQL Server та Oracle [Електронний ресурс] – Режим доступу до ресурсу – https://journals.indexcopernicus.com/api/file/viewByFileId/156430.pdf
7. Сховища даних та OLAP – технології [Електронний ресурс] – Режим доступу до ресурсу – https://moodle.znu.edu.ua/pluginfile.php/767343/mod_resource/content/1/%D0%A2%D0%B5%D0%BC%D0%B0%202.pdf

8. SQL Server Analysis Services overview [Електронний ресурс] – Режим доступу до ресурсу – https://learn.microsoft.com/en-us/analysis-services/ssas-overview?view=asallproducts-allversions
9. Data Warehouse Architecture: Foundations and Best Practices [Електронний ресурс] – Режим доступу до ресурсу – https://www.aampe.com/blog/data-warehouse-architecture
10. Machine Learning [Електронний ресурс] – Режим доступу до ресурсу – https://www.opentext.com/what-is/machine-learning
11. Machine learning types [Електронний ресурс] – Режим доступу до ресурсу – https://www.ibm.com/think/topics/machine-learning-types
12. Correlation Matrix [Електронний ресурс] – Режим доступу до ресурсу – https://www.questionpro.com/blog/correlation-matrix/#What_is_a_correlation_matrix
13. Задачі Data Mining та їх класифікація - [Електронний ресурс] – Режим доступу до ресурсу: https://moodle.znu.edu.ua/pluginfile.php/486125/mod_resource/content/1/%D0%9B%D0%B5%D0%BA%D1%86%D1%96%D1%8F%205.pdf
14. XgBoost [Електронний ресурс] – Режим доступу до ресурсу – https://itwiki.dev/data-science/ml-reference/ml-glossary/xgboost
15. Calculating XGBoost Feature Importance [Електронний ресурс] – Режим доступу до ресурсу – https://medium.com/@emilykmarsh/xgboost-feature-importance-233ee27c33a4
16. Decision Trees - [Електронний ресурс] – Режим доступу до ресурсу: https://scikit-learn.org/stable/modules/tree.html
17. Стандартні методи кластеризації даних [Електронний ресурс] – Режим доступу до ресурсу – https://csc.knu.ua/media/study/asp/mod_probl_inf_tech_sys_analysis_ivohin/lecture/lec2.pdf

18. Data Visualization with Power BI [Електронний ресурс] – Режим доступу до ресурсу – https://www.datacamp.com/tutorial/data-visualisation-powerbi
19. Box Plot Explained [Електронний ресурс] – Режим доступу до ресурсу – https://www.simplypsychology.org/boxplots.html
20. A Systematic Literature Review of Key Performance Indicators (KPIs) Implementation [Електронний ресурс] – Режим доступу до ресурсу – https://www.researchgate.net/publication/345941517_A_Systematic_Literature_Review_of_Key_Performance_Indicators_KPIs_Implementation
21. Machine Learning, ML [Електронний ресурс] – Режим доступу до ресурсу – https://www.it.ua/knowledge-base/technology-innovation/machine-learning
22. Ресурс Державного агентства водних ресурсів України [Електронний ресурс] – Режим доступу до ресурсу http://monitoring.davr.gov.ua/EcoWaterMon/GDKMap/Index
23. A Comprehensive Guide to Accuracy in Machine Learning [Електронний ресурс] – Режим доступу до ресурсу https://www.artsyltech.com/blog/Accuracy-In-Machine-Learning
24. Analysis Services documentation [Електронний ресурс] – Режим доступу до ресурсу https://learn.microsoft.com/en-us/analysis-services/?view=asallproducts-allversions
25. Коефіцієнт переведення БСК5 [Електронний ресурс] – Режим доступу до ресурсу https://ecolog-ua.com/consultation/shcho-take-koeficiyent-perevedennya-bsk5-v-bskpov-yakyy-dorivnyuye-133-chy-ye-u-nas-ye