Корж Інна В’ячеславівна, бакалавр

Національний технічний університет України “Київський політехнічний інститут імені Ігоря Сікорського”, Київ

https://orcid.org/0009-0004-8159-2200

Сарибога Ганна Володимирівна

Національний технічний університет України “Київський політехнічний інститут імені Ігоря Сікорського”, Київ

http://orcid.org//0000-0003-0805-7899

**ОЦІНКА СТАНУ ТА ПРОГНОЗУВАННЯ ЯКОСТІ ВОДНИХ РЕСУРСІВ ЗА ДОПОМОГОЮ ШІ**

Постановка проблеми. Забруднення водних ресурсів є серйозною проблемою у багатьох регіонах світу через індустріалізацію, сільське господарство та інші людські діяльності. Важливість знаходження інструментів та методів оцінки стає нагальною у часи невизначеності, нестабільності.

Впродовж останніх 10—15 років відбувається неконтрольоване накопичення промислових і побутових відходів в Україні. Найбільшими забруднювачами довкілля (у тому числі водних ресурсів) є енергетична, хімічна, металургійна, вуглевидобувна промисловості, комунальні підприємства міст і селищ міського типу, сільське господарство. Отримання детальної об’єктивної інформації про стан басейнів річок (зокрема малих) можливе лише на основі здійснення їх систематичного моніторингу з використанням електронної бази даних на основі Інтерактивної карти динаміки екологічного стану малих річок України та водних об’єктів у річкових басейнах [1, с. 2].

Актуальність. Термін «якість води» означає оцінку стану води, яка враховує фізичні, хімічні та біологічні властивості води. При дослідженні будь-якої водної системи моделювання параметрів якості води має величезне значення. Оцінка та прогнозування якості вод необхідні для утримання рівня забруднення в допустимих межах. Прогнозування явищ та дій, які впливають на якість води, забезпечує оптимальне управління водними ресурсами. Машинне навчання є ефективним засобом прогнозування оскільки надає можливість обробки великих даних, виявити складні зв’язки між різними параметрами води програмно та уникнути можливих ризиків забруднення, а також уникнути людського фактору та зекономити час на прийняття рішення.

Аналіз основних досліджень і публікацій. Питанням прогнозування якості водних ресурсів присвячено значну кількість публікацій зарубіжних учених. D. Irwan, M. Ali та інші відібрали 83 статті із 801, 44 статті було відхилено через дублювання, а 674 статті не враховувалися в цьому дослідженні, оскільки вони були не про прогнозування якості води чи їхні основні висновки не мали відношення до прогнозування якості води. Ці статті були вибрані, щоб охопити численні алгоритми моделювання, такі як традиційна машина навчання (ML), ансамблеве навчання (EL), глибоке навчання (DL), а також гібридні моделі [2]. Крім різного підходу до побудови алгоритмів також велика різниця присутня у кількості вхідних даних і параметрів. У кожній роботі використовувалися структуровані дані, які можна впорядкувати та звести в таблиці. Багато публікацій, які були розглянуті, використовували хімічні речовини для передбачення. Для прогнозу використовувалися метеорологічні дані. Крім того, інші дослідницькі роботи використовували поєднання як хімічних, так і метеорологічних вхідних даних. Статті, які було розглянуто належать проміжку з 2009 по 2023 рік, значне збільшення кількості статей по даному питанню припадає саме на 2016 рік (5), порівнюючи з 2015 роком (1). Протягом 2019 - 2022 року їх кількість поступово зростала, середня кількість статей за цей проміжок часу становить - 15, що вкотре доводить актуальність даного питання.

Метою проекту стала розробка програмного продукту для оцінка стану та прогнозування якості водних ресурсів за допомогою алгоритмів машинного навчання. Розроблене програмне забезпечення повинно включати графічний інтерфейс, який дає користувачеві можливість введення інформації про хімічні сполуки та мікроорганізми (алюміній, аміак, миш’як, барій, кадій, мідь, свинець, бактерії, віруси та ін.), що присутні у воді, та їх концентрацію та прогнозувати на основі отриманих даних подальші перспективи щодо якості води.

Алгоритми машинного навчання, створені для повторення структури та функціональності людського мозку. Ці моделі складаються з взаємопов’язаних штучних нейронів (вузлів), розташованих шарами, причому кожен елемент шару повністю з’єднаний з наступним. Кожен вузол у вхідному шарі отримує окрему вхідну змінну, а вузли в прихованих шарах перетворюють ці вхідні дані за допомогою ряду нелінійних функцій перед тим, як створити вихідні дані в кінцевому рівні

Вибір відповідної моделі штучного інтелекту залежить не лише від вибраних змінних і доступності даних, але й від типу моделі та рівня продуктивності. Ефективність має вирішальне значення для програм реального часу, оскільки вона пов’язана зі здатністю моделі швидко обробляти дані. Точність визначає, наскільки точно прогнози моделі відповідають фактичним даним. Можливість інтерпретації важлива, оскільки вона дозволяє користувачам зрозуміти, як модель робить прогнози. Адаптивність означає здатність моделі пристосовуватися до нових даних і обставин, що змінюються. Вимоги до даних пов’язані з кількістю та типом даних, необхідних для навчання моделі. Моделі, які вимагають великих обсягів даних, можуть бути складними для навчання та не підходять для всіх програм. Ідеальна модель є ефективною, точною, інтерпретованою, адаптованою та вимагає мінімальних даних. Оскільки такої моделі ще не розроблено, тому найкращим підходом буде вибір моделі, яка максимально відповідає конкретній ситуації. Гібридні моделі можуть бути відносно хорошою альтернативою для оцінки та прогнозування Було проведено огляд та порівняльний аналіз гібридних алгоритмів машинного навчання, а саме: Random Forest, Extra Trees, XGBoost і CatBoost - для виявлення найефективнішого для роботи з встановленим датасетом з Kaggle.

Random Forest є ансамблем рішаючих дерев, це техніка машинного навчання, яка поєднує прогнози з кількох моделей для створення більш точного та стабільного прогнозу [3]. Важливою особливістю є випадковість у виборі ознак для побудови кожного дерева, що робить модель більш стійкою до перенавчання та здатною враховувати різноманітність даних.

Extra Trees (або Extremely Randomized Trees) також є ансамблем рішаючих дерев. При побудові кожного дерева в Extra Trees випадковим чином обираються не тільки ознаки, але й пороги для розділення вузлів. Це робить модель ще більш випадковою та менш схильною до перенавчання, що іноді може призводити до ефективного використання меншої кількості даних [4].

XGBoost є одним з найпопулярніших алгоритмів градієнтного бустінгу. Він використовує дерева рішень як базові моделі та акцентується на оптимізації шляхом градієнтного підйому. XGBoost включає регуляризацію, що допомагає у контролі перенавчання та підвищенні стійкості моделі.

CatBoost, або Categorical Boosting, є алгоритмом градієнтного бустінгу, який спеціалізується на ефективній обробці категоріальних ознак. Основна особливість CatBoost полягає в тому, що він автоматично обробляє категоріальні дані без необхідності додаткового передпроцесу [5]. Це робить його особливо зручним для роботи з даними, де присутні категоріальні ознаки. Для розуміння алгоритму роботи програми розроблено діаграму(рис. 1).

Для порівняння точності використовувались Сonfusion Matrix - це матриця, яка підсумовує продуктивність моделі машинного навчання на наборі тестових даних. Це засіб відображення кількості точних і неточних випадків на основі прогнозів моделі. На основі матриць найбільшу точність продемонстрував алгоритм CatBoost - 0,9575. Це може бути зумовлено тим, що CatBoost відомий своєю стійкістю до шуму та вміє добре працювати з неоднорідними даними.

Крім того, для ефективної обробки даних в програмі використовуються концепція Pipelines. Pipelines є важливим інструментом у світі обробки даних, оскільки вони дозволяють автоматизувати та структурувати процес обробки даних в логічні етапи.

Pipelines уможливлюють автоматичне виконання послідовних завдань, включаючи нормалізацію даних, заповнення пропущених значень, видалення викидів, вибір ознак та застосування моделей машинного навчання. Вони дозволяють створювати зручну та зрозумілу конвеєрну структуру, яка сприяє чистоті та легкості обробки даних.

Такий підхід до обробки даних з використанням pipelines сприяє підтримці чистого та структурованого коду, робить програму більш гнучкою у виборі обробних етапів, та сприяє швидкому впровадженню змін в аналітичні процеси без значних трудовитрат.

Датасет для навчання і тренування було завантажено зі сайту Kaggle, з 20-тьма параметрами, серед яких показники: алюмінію, аміаку, миш’яка, барію, кадію, міді, свинця, бактерій, вірусів і тп. на літр води. Кількість екземплярів — 8000. На основі всіх цих даних відбувалось навчання і тренування моделі.

Використання бібліотек, таких як NumPy, Pandas, Seaborn, Scikit-learn та Matplotlib, дозволило ефективно обробляти та візуалізувати дані, а також використовувати різні алгоритми машинного навчання для прогнозування безпеки води.

Процес обробки даних та використання пайплайнів допомагає забезпечити структурованість та ефективність аналізу. Враховуючи різноманітність параметрів та їх взаємодію, програма забезпечує комплексний підхід до оцінки безпеки води для споживачів.

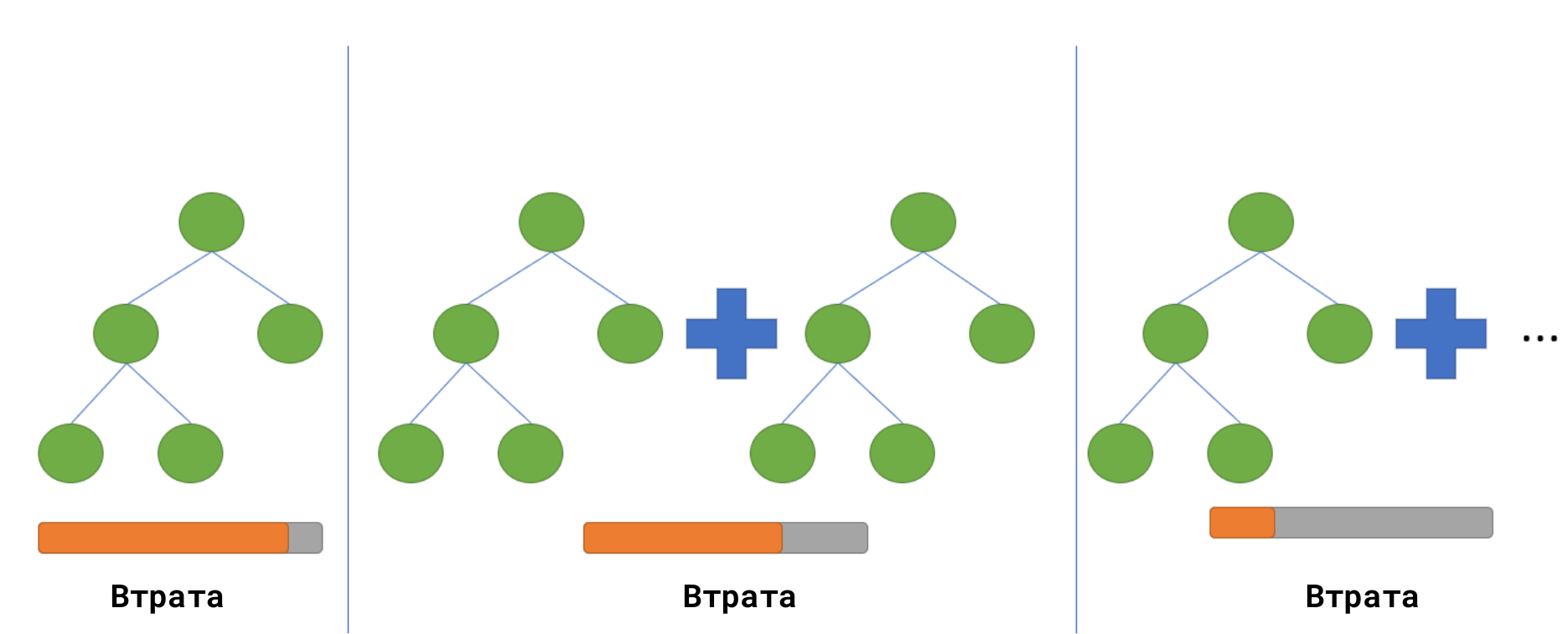


Рисунок 1 – алгоритм роботи CatBoost

Розроблено інтерфейс, інтуїтивно зрозумілий користувачу(рис. 2). Коли користувач вводить дані в усі відповідні поля і натискає на кнопку "Qualify", введені інформації передаються в систему обробки даних. В цьому етапі програма використовує алгоритм CatBoost. Після чого з’являється спливаюче вікно з результатом на основі тих даних, які користувач увів раніше. Результатом виступає інформація чи безпечна дана вода для вживання.

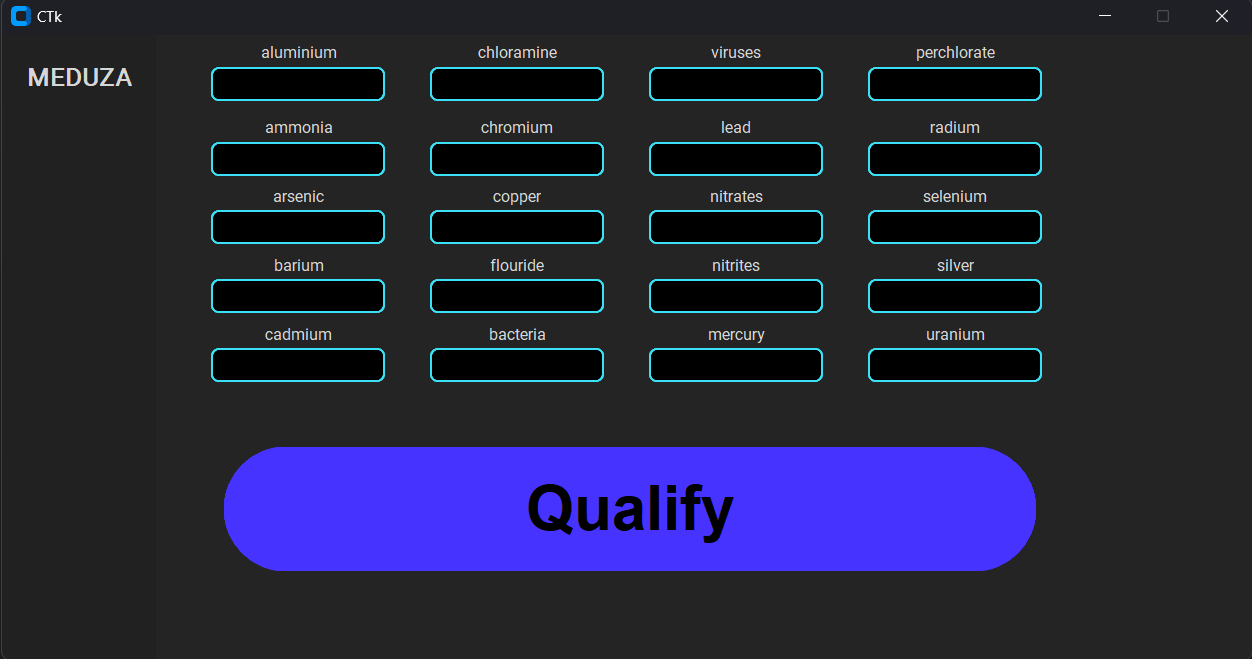


Рисунок 2 – головне вікно програми

Висновки. Досліджено алгоритми машинного навчання для застосування у сфері водних ресурсів таких як Random Forest, Extra Trees, XGBoost і CatBoost та запропоновано CatBoost для використання, як важливий засіб виявлення рівня забрудненості води та прогнозування подальшого стану для визначення чинників, що впливають на нього. Протестовано алгоритм та натреновано нейромережу з достовірністю 0,9575, що є підставою вважати алгоритм ефективним, а прогноз достовірним. На основі алгоритму розроблено програмний продукт з простим та зрозумілим інтерфейсом для користувача. Програмний продукт запропоновано, як ефективний інструмент моніторингу та оцінки якості води з метою забезпечення відповідного допустимого рівня і, як наслідок, збереження здоров'я людини та навколишнього середовища.

**Література**

1. Predicting Water Quality with Artificial Intelligence: A Review of Methods and Applications / D. Irwan et al. Archives of Computational Methods in Engineering. 2023. URL: <https://doi.org/10.1007/s11831-023-09947-4> (дата звернення: 10.09.2024).

2. ДП «Редакція газ. «Голос України». Проблеми використання водних ресурсів потребують негайного розв’язання. *Голос України - газета Верховної Ради України*. URL: <http://www.golos.com.ua/article/306442> (дата звернення: 08.09.2024).

3. Random Forest Regression in Python - GeeksforGeeks. GeeksforGeeks. URL: https://www.geeksforgeeks.org/random-forest-regression-in-python/ (дата звернення: 08.09.2024).

4. Extra-trees. Help NextBrain. URL: https://help.nextbrain.ai/en/article/extra-trees-1ka8ie5/ (дата звернення: 08.09.2024).

5. GeeksforGeeks. How CatBoost algorithm works - GeeksforGeeks. GeeksforGeeks. URL: https://www.geeksforgeeks.org/catboost-algorithms/ (дата звернення: 08.09.2024).