МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Західноукраїнський національний університет
Факультет комп'ютерних інформаційних технологій
Кафедра інформаційно-обчислювальних систем і управління

Дегодь Віталій Андрійович

Програмний модуль прогнозування якості води на основі ансамблевого підходу / Software Module for Woter Quality Forecasting based on Ensemble Approach

Спеціальність 122 – Комп'ютерні науки
Освітньо-професійна програма – Комп'ютерні науки

Кваліфікаційна робота

Виконав студент групи КН-42
Дегодь В.А.

Науковий керівник:
к.т.н., ст.викладач
Домбровський М.З.

Кваліфікаційну роботу допущено до захисту
«___»_____ 2025 р.

В.о. завідувача кафедри
_____ Н.М. Васильків

Тернопіль – 2025

Факультет комп'ютерних інформаційних технологій
Кафедра інформаційно-обчислювальних систем і управління
Освітній ступінь «бакалавр»
спеціальність 122 – Комп'ютерні науки
освітньо-професійна програма – Комп'ютерні науки

ЗАТВЕРДЖУЮ
В.о. завідувача кафедри
_____ Н.М. Васильків
«____» _____2024 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТУ
Дегодь Віталій Андрійович
(прізвище, ім'я, по батькові)

1. Тема кваліфікаційної роботи: Програмний модуль прогнозування якості води на основі ансамблевого підходу / Software Module for Woter Quality Forecasting based on Ensemble Approach
керівник роботи к.т.н., ст.викладач, Домбровський М.З.
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)
затверджені наказом по університету від 20 грудня 2024 р. № 938.

2. Строк подання студентом закінченої кваліфікаційної роботи 25 травня 2025 р.

3. Вихідні дані до кваліфікаційної роботи: завдання на кваліфікаційну роботу студента, наукові статті, технічна література.

4. Основні питання, які потрібно розробити:

– Провести аналіз існуючих методів оцінювання якості води.

– Оцінити сучасні підходи до прогнозування параметрів якості води за допомогою методів машинного навчання та визначити найбільш перспективні алгоритми для цієї задачі.

– Провести огляд ансамблевих методів машинного та дослідити їх ефективність у контексті прогнозування придатності води.

– Розробити алгоритмічне та інформаційне забезпечення програмного модуля прогнозування.

– Реалізувати програмний модуль прогнозування якості води на основі ансамблевого підходу та провести його верифікацію за допомогою відповідних метрик.

5. Перелік графічного матеріалу в роботі:

– Архітектура системи прогнозування якості води.

– Матриця помилок для моделей Gradient Boosting та Bagging Classifier..

6. Консультанти розділів кваліфікаційної роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		Завдання видав	Завдання прийняв

7. Дата видачі завдання 20 грудня 2024 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Строк виконання етапів кваліфікаційної роботи	Примітка
1	Затвердження теми кваліфікаційної роботи, ознайомлення з літературними джерелами та складання плану роботи	до 01.01. 2025 р.	
2	Написання 1 розділу кваліфікаційної роботи	до 01.02. 2025 р.	
3	Написання 2 розділу кваліфікаційної роботи	до 01.04.2025 р.	
4	Написання 3 розділу кваліфікаційної роботи	до 01.05. 2025 р.	
5	Представлення попереднього варіанту кваліфікаційної роботи, перевірка та внесення змін керівником	до 15.05.2025 р.	
6	Опрацювання зауважень та представлення завершеного варіанту кваліфікаційної роботи. Підготовка супроводжуючих документів	до 25.05.2025 р.	
7	Перевірка кваліфікаційної роботи на оригінальність тексту	до 30.05.2025 р.	
8	Оформлення кваліфікаційної роботи та отримання допуску до захисту	до 10.06.2025 р.	
9	Подання кваліфікаційної роботи до захисту на засіданні атестаційної комісії	до 10.06. 2025 р.	

Студент_____ Дегодь В.А._____
(підпис) (прізвище та ініціали)

Керівник кваліфікаційної роботи_____ Домбровський М.З.____
(підпис) (прізвище та ініціали)

АНОТАЦІЯ

Кваліфікаційна робота на тему «Програмний модуль прогнозування якості води на основі ансамблевого підходу» на здобуття освітнього ступеня «бакалавр» зі спеціальності 122 «Комп'ютерні науки» освітньої програми «Комп'ютерні науки» написана обсягом у 57 сторінок і містить 13 ілюстрацій, 7 таблиць, 2 додатки та 23 використані джерела.

Метою роботи є розробка ефективного підходу до прогнозування якості води з використанням ансамблевих методів машинного навчання для оперативної оцінки її придатності до споживання.

Методами розроблення обрано аналіз існуючих підходів, ансамблеві алгоритми (Gradient Boosting, Random Forest, Bagging, AdaBoost, Stacked Ensembles), гіперпараметричну оптимізацію та автоматизовані методи H2O AutoML.

Результатом дослідження є програмний модуль, що дозволяє здійснювати високоточне прогнозування придатності води з точністю 80% та значенням AUC-ROC 0.864.

Результати дослідження можуть бути застосовані у системах моніторингу якості води для екологічних та комунальних служб.

Ключові слова: ПРОГНОЗУВАННЯ ЯКОСТІ ВОДИ, АНСАМБЛЕВІ МЕТОДИ, ГЛИБОКЕ НАВЧАННЯ, АНАЛІЗ ДАНИХ, H2O AUTOML, ГРАДІЄНТНИЙ БУСТИНГ, RANDOM FOREST.

ANNOTATION

The bachelor's thesis titled "Software Module for Water Quality Forecasting Based on Ensemble Approach" consists of 57 pages and includes 13 figures, 7 tables, 2 appendices, and 23 references.

The goal is to develop an effective approach for forecasting water quality using ensemble machine learning methods for rapid assessment of water suitability for consumption.

The selected development methods include the analysis of existing methods, ensemble algorithms (Gradient Boosting, Random Forest, Bagging, AdaBoost, Stacked Ensembles), hyperparameter optimization, and automated methods using H2O AutoML.

The research outcome is a software module capable of predicting water suitability with an accuracy of 80% and an AUC-ROC score of 0.864.

The findings can be integrated into water quality monitoring systems for environmental and municipal services.

Keywords: WATER QUALITY FORECASTING, ENSEMBLE METHODS, DEEP LEARNING, DATA ANALYSIS, H2O AUTOML, GRADIENT BOOSTING, RANDOM FOREST.

ЗМІСТ

ВСТУП

Актуальність дослідження. Забезпечення населення якісною питною водою є однією з ключових проблем сучасного світу. Вода є основним ресурсом для життя, а її забруднення спричиняє серйозні екологічні та медичні загрози. За даними Всесвітньої організації охорони здоров'я (ВООЗ), щороку мільйони людей страждають від хвороб, пов'язаних із вживанням забрудненої води. Традиційні методи оцінки її якості, що базуються на лабораторних аналізах, є дороговартісними та потребують значного часу для отримання результатів. У зв'язку з цим постає потреба у впровадженні сучасних методів прогнозування, які дозволять оперативно оцінювати придатність води до споживання на основі аналізу фізико-хімічних показників.

Одним із найбільш ефективних підходів у сфері інтелектуального аналізу даних є ансамблеві методи машинного навчання. Вони дозволяють підвищити точність класифікації за рахунок поєднання кількох моделей, що компенсує їхні індивідуальні недоліки та забезпечує стійкість до шумових даних. Використання ансамблевих моделей, таких як Gradient Boosting, Bagging, Random Forest та Stacked Ensembles, дозволяє автоматично обробляти великі обсяги даних і здійснювати високоточне прогнозування якості води. Це особливо важливо для регіонів, де відсутній постійний лабораторний контроль, а оперативна оцінка води може допомогти запобігти екологічним катастрофам та масовим захворюванням.

Розробка програмного модуля для прогнозування якості води на основі ансамблевого підходу є важливим кроком до автоматизації процесу моніторингу водних ресурсів. Запропонований підхід дозволяє не лише покращити точність класифікації, а й скоротити час ухвалення рішень щодо безпечності води. Впровадження такого модуля в екологічні, медичні та комунальні служби дозволить оперативно аналізувати якість води та своєчасно вживати необхідних заходів для її очищення. Таким чином, дослідження спрямоване на вирішення однієї з найважливіших екологічних проблем сучасності шляхом інтеграції інноваційних методів обробки даних у систему моніторингу водних ресурсів.

Метою дослідження є розробка ефективного підходу до прогнозування якості води з використанням ансамблевих методів машинного навчання, що забезпечить оперативну оцінку її придатності до споживання на основі аналізу фізико-хімічних показників.

Для досягнення поставленої мети необхідно виконати наступні завдання:

1. Провести аналіз існуючих методів оцінювання якості води, зокрема традиційних лабораторних та автоматизованих підходів, для виявлення їхніх переваг і недоліків.

2. Оцінити сучасні підходи до прогнозування параметрів якості води за допомогою методів машинного навчання та визначити найбільш перспективні алгоритми для цієї задачі.

3. Провести огляд ансамблевих методів машинного навчання (Bagging, Boosting, Random Forest, Stacked Ensembles) та дослідити їх ефективність у контексті прогнозування придатності води.

4. Розробити алгоритмічне та інформаційне забезпечення програмного модуля прогнозування, що включає механізми збору, попередньої обробки та аналізу кількісних даних.

5. Реалізувати програмний модуль прогнозування якості води на основі ансамблевого підходу та провести його верифікацію за допомогою відповідних метрик (точність, Precision, Recall, F1-міра, AUC-ROC) з подальшим аналізом отриманих результатів.

Об’єкт дослідження – процес прогнозування якості води на основі аналізу фізико-хімічних показників, таких як pH, твердість, вміст розчинених речовин, хлорамінів, сульфатів, електропровідність, органічний вуглець, тригалометани та каламутність.

Предмет дослідження – методи та алгоритми машинного навчання, зокрема ансамблеві методи, що використовуються для аналізу кількісних даних та прогнозування придатності води до споживання.

Методи дослідження включають статистичний аналіз даних, застосування ансамблевих алгоритмів (Gradient Boosting, Bagging, Random Forest, Stacked Ensembles), оптимізацію гіперпараметрів (GridSearchCV, RandomizedSearchCV) та

використання сучасних технологій програмної реалізації (Python, Pandas, Scikit-learn, H2O AutoML) для моделювання та перевірки ефективності прогнозних алгоритмів.

Апробація результатів дослідження здійснена в межах студентської науково-практичної конференції «Інтелектуальні інформаційні технології в прикладних дослідженнях» (IITAR-2025), яка відбулася в місті Тернополі 27–29 травня 2025 року.

1 АНАЛІЗ СТАНУ ПРОБЛЕМИ ТА ПОСТАНОВКА ЗАВДАННЯ

1.1 Аналіз існуючих методів оцінювання якості води

Оцінювання якості води є критично важливим завданням для забезпечення безпеки питної води та збереження екологічного балансу. Залежно від мети дослідження використовуються різні методи аналізу, які можна розділити на традиційні лабораторні, експрес-методи та автоматизовані системи моніторингу. Кожен із цих підходів має свої переваги та обмеження, що визначають доцільність їх використання в різних умовах.

Традиційні методи оцінювання якості води включають хімічний, фізико-хімічний, мікробіологічний та токсикологічний аналізи. Вони використовуються для визначення вмісту важких металів, мікроорганізмів, органічних та неорганічних забруднювачів. Наприклад, титриметричні методи дозволяють визначити рівень жорсткості води шляхом титрування з розчинами комплексонатів, а спектрофотометрія використовується для аналізу концентрації нітратів, фосфатів та сульфатів.

Мікробіологічний аналіз включає дослідження проб води на наявність патогенних бактерій, таких як Escherichia coli, Salmonella spp., Pseudomonas aeruginosa. Для цього використовують методи посіву на поживні середовища, що дозволяє оцінити мікробіологічний стан води.

З метою оперативного визначення якості води широко використовуються експрес-методи, що включають тест-смужки, електрохімічні сенсори та портативні аналізатори. Вони дозволяють швидко оцінити такі параметри, як pH, вміст хлору, жорсткість, вміст органічних забруднювачів.

Одним із найбільш поширених методів є потенціометрія, що використовує іоноселективні електроди для визначення концентрації окремих іонів (наприклад, хлоридів, нітратів, амонію). Крім того, турбідиметричний метод застосовується для оцінки каламутності води, що є важливим параметром при визначенні її фізичної якості.

Завдяки розвитку технологій усе більшого поширення набувають автоматизовані системи моніторингу, що забезпечують безперервне оцінювання

якості води в реальному часі. Такі системи базуються на використанні датчиків, біосенсорів, спектрофотометричних аналізаторів, які можуть бути інтегровані в централізовані системи управління водопостачанням.

Автоматизовані системи дозволяють здійснювати моніторинг таких параметрів, як температура, рівень pH, електропровідність, хімічне та біологічне споживання кисню (COD, BOD), концентрація шкідливих речовин. Використання технологій IoT (Internet of Things) сприяє об'єднанню сенсорних пристроїв у єдину мережу для віддаленого контролю та аналізу даних.

Кожен із вищезгаданих методів має свої переваги та обмеження (Таблиця 1.1). Лабораторні методи забезпечують найвищу точність, але потребують значних затрат часу та ресурсів. Експрес-методи є оперативними, але менш точними. Автоматизовані системи дозволяють безперервний моніторинг, проте потребують значних капіталовкладень та інфраструктури.

Таблиця 1.1 – Порівняння методів оцінювання якості води

Метод	Точність	Час аналізу	Вартість	Автоматизація
Лабораторні методи	Висока	Від кількох годин до днів	Висока	Низька
Експрес-методи	Середня	Хвилини	Середня	Середня
Автоматизовані системи	Висока	Миттєво	Висока	Висока

Останнім часом методи машинного навчання (ML) все частіше застосовуються для прогнозування якості води. Вони дозволяють автоматично аналізувати великі обсяги даних, виявляти приховані закономірності та здійснювати класифікацію проб води як придатної або непридатної для споживання.

Зокрема, ансамблеві методи машинного навчання, такі як Random Forest, Gradient Boosting, AdaBoost, показують високу ефективність у визначенні придатності води на основі фізико-хімічних параметрів. Використання таких моделей дозволяє замінити складні лабораторні дослідження швидкими алгоритмічними розрахунками, що значно зменшує витрати часу.

Аналіз існуючих методів оцінювання якості води показав, що лабораторні методи є найбільш точними, проте витратними. Експрес-методи забезпечують швидкі результати, але можуть мати похибку. Автоматизовані системи моніторингу дозволяють здійснювати оцінювання в реальному часі, але потребують значних інвестицій.

Використання ансамблевих методів машинного навчання є перспективним напрямом у прогнозуванні якості води. Вони дозволяють оптимізувати процес оцінювання, зменшуючи витрати на аналіз і підвищуючи швидкість ухвалення рішень. Таким чином, подальше дослідження має бути спрямоване на інтеграцію методів машинного навчання в процес моніторингу якості води та розробку програмного модуля для автоматизованого прогнозування її придатності до споживання.

1.2 Огляд ансамблевих методів машинного навчання у прогнозуванні

Ансамблеві методи машинного навчання відіграють важливу роль у прогнозуванні якості води, дозволяючи підвищити точність оцінок та зменшити похибки моделей. Одним із сучасних підходів є інтеграція Boosted Learning із Differential Evolution (DE), що дозволяє покращити оцінку ризиків забруднення підземних вод [1]. Аналогічно, використання супутникових даних у поєднанні з алгоритмами машинного навчання сприяє покращенню оцінки якості води у великих водних об’єктах, наприклад, у дослідженні, присвяченому воді в озері Дяньчі [2].

Гібридні моделі, що комбінують алгоритми глибокого навчання та дерев рішень, також демонструють ефективність у прогнозуванні якості води, забезпечуючи точніші результати [3]. Важливим аспектом є розробка пояснюваних моделей, що дозволяють отримати високоточні оцінки опадів для вдосконалення прогнозування стану водних ресурсів [4]. Дослідження також показують, що використання інтерпретованого штучного інтелекту у поєднанні з різними джерелами дистанційного зондування може значно підвищити надійність прогнозування якості річкової води [5].

Окрім методів машинного навчання, активно розглядаються оптимізаційні підходи, такі як прогнозування адсорбції амонію на біочарі за допомогою ансамблевих методів та їх оцінки [6]. Подібні підходи застосовуються для розрахунку розчиненого органічного вуглецю в озерах, що покращує можливості моделювання на основі машинного навчання та розширених даних віддаленого зондування [7]. Фактори, що впливають на концентрацію вуглецю у водних екосистемах, також визначаються за допомогою методів машинного навчання, що дозволяє зробити більш точні оцінки змін екологічних параметрів [8].

Сучасні дослідження також спрямовані на оптимізацію використання води завдяки розумним моделям та штучному інтелекту, що дозволяє зменшити витрати та підвищити ефективність управління водними ресурсами [9]. Крім того, розробка хмарних систем моніторингу та прогнозування, таких як Aqualite Engine 1.0, допомагає створити передові інструменти для управління якістю води та її прогнозування [10].

Ансамблеві методи широко використовуються для прогнозування якості води, зокрема для оцінки параметрів, пов’язаних із забрудненням та екологічною безпекою. Одним із таких досліджень є використання методу символічної регресії у поєднанні з ансамблевими алгоритмами для оптимізації геополімерного бетону, що є корисним для екологічних досліджень, зокрема прогнозування якості води [11]. Інший підхід передбачає застосування геостатистичних методів разом із машинним навчанням для покращення прогнозування параметрів ґрунтових та водних ресурсів [12].

Значну увагу приділяють використанню гідрологічних і гідрохімічних моделей для аналізу якості води у великих річкових басейнах. Наприклад, дослідження в Іспанії використовує глибоке навчання для прогнозування змін у водних ресурсах, забезпечуючи високу точність прогнозування нітратного забруднення [13]. У ще одному дослідженні представлено гібридну модель, що поєднує Boosting та машинне навчання для прогнозування якості води у природних річках [14].

Методи ансамблевого навчання також застосовуються у поєднанні з супутниковими даними. Наприклад, дослідження прогнозує загальний об'єм води

у різних регіонах, поєднуючи моделювання машинного навчання з даними CLM та GRACE [15]. Також вивчається можливість використання нейронних мереж для прогнозування концентрацій різних забруднень у воді, таких як важкі метали та розчинені органічні речовини [16].

Окрім цього, застосування ансамблевих моделей виявилося корисним для прогнозування розливів води та кліматичних змін, що впливають на якість води [17]. Аналогічно, у дослідженні щодо біогазифікації представлено прогнозування процесів очищення води та зміни її складу під час промислової переробки [18].

Окремий напрямок досліджень зосереджується на оцінці якості води за допомогою машинного навчання для моніторингу природних водойм та оцінки потенційних ризиків забруднення [19]. Останнє дослідження пропонує інноваційний підхід, заснований на нейронних мережах і методах обробки помилок, що дозволяє покращити прогнозування параметрів води в екстремальних умовах [20].

Таблиця 1.2 містить порівняльний аналіз 20 досліджень, що застосовують ансамблеві методи машинного навчання для прогнозування якості води, виділяючи ключові методи та області застосування.

Аналіз досліджень показує, що найбільш ефективними підходами для прогнозування якості води є Boosted Learning, глибоке навчання, геостатистичні методи та нейронні мережі. Багато моделей використовують супутникові дані та віддалене зондування для покращення точності прогнозів, а також інтегрують AI у хмарні платформи для моніторингу водних ресурсів. Особливу увагу приділяють виявленню забруднень, таких як азотні домішки, важкі метали та органічні речовини, що робить ці технології важливими для екологічного контролю.

Таблиця 1.2 – Порівняльний аналіз досліджень ансамблевих методів машинного навчання у прогнозуванні якості води

№	Дослідження	Основні методи	Застосування
1	Boosted Learning + DE Optimizer для оцінки підземних вод	Boosted Learning, Differential Evolution	Підземні води
2	Прогнозування якості води в озері Дяньчі	ML + супутникові дані	Прісноводні озера
3	Гібридні дерева рішень і глибоке навчання	Глибоке навчання, дерева рішень	Річки та озера

Продовження таблиці 1.2

№	Дослідження	Основні методи	Застосування
4	Пояснюване ML для прогнозування опадів	Explainable AI, погодні дані	Опади, вплив на якість води
5	Інтерпретований AI для якості річкової води	AI, віддалене зондування	Річки, оцінка забруднення
6	Прогнозування азотного забруднення	ML, предиктивні алгоритми	Азотне забруднення
7	Оцінка органічного вуглецю в озерах	ML + Data Augmentation	Органічні домішки
8	Фактори впливу на вуглецевий баланс у водоймах	ML для виявлення трендів	Гідрологічні зміни
9	Оптимізація використання води AI-моделями	Оптимізація AI, smart-моделі	Водні ресурси, smart-технології
10	Хмарний моніторинг Aqualite Engine 1.0	Cloud-based AI	Глобальний моніторинг
11	Оптимізація геополімерного бетону через ансамблеві моделі	Ансамблеві моделі, регресія	Будівництво та водоочистка
12	Геостатистичне прогнозування параметрів ґрунту та води	Геостатистичні методи + ML	Грунти та підземні води
13	ML для аналізу гідрологічних даних у великих річках	Глибоке навчання, гідрологія	Гідрологічне прогнозування
14	Гібридна модель Boosting для річкової якості води	Boosting, ML для якості води	Якість річкової води
15	Прогнозування загального об’єму води (CLM+GRACE)	AI + супутникове моделювання	Глобальний аналіз водних змін
16	ML-моделі для прогнозування забруднень у воді	ML для оцінки забруднень	Оцінка важких металів
17	Прогнозування водних змін через кліматичні фактори	Кліматичне моделювання + ML	Кліматичний вплив
18	ML у біогазифікації для покращення очищення води	AI, аналіз хімічних сполук	Промислові води
19	Моніторинг водойм через машинне навчання	ML для спостережень	Моніторинг природних водойм
20	Нейронні мережі для прогнозування води в екстремальних умовах	Нейронні мережі, аналітика помилок	Екстремальні умови, катастрофи

Окрім традиційних підходів, дослідження також включають інноваційні методи, такі як пояснюване AI, кліматичне моделювання та поєднання машинного навчання з фізичними моделями (CLM+GRACE). Ці підходи дозволяють не лише прогнозувати якість води, але й виявляти фактори, що впливають на її зміну, покращуючи стратегічне управління водними ресурсами та запобігання екологічним катастрофам.

1.3 Постановка задачі дослідження

В умовах зростаючого навантаження на водні ресурси та підвищення рівня забруднення природних джерел води питання забезпечення високої якості питної води набуває особливої ваги. Якість води є критичним показником для здоров'я населення та екологічного балансу, оскільки вона безпосередньо впливає на всі біологічні процеси та якість життя людей.

Сучасні виклики у сфері екологічного моніторингу вимагають оперативних та точних методів оцінювання якості води. Традиційні лабораторні методи, хоч і забезпечують високу точність, є витратними за часом і ресурсами, що робить їх недостатньо ефективними для широкомасштабного моніторингу, особливо в умовах необхідності негайного реагування на зміни в якості води.

У цьому контексті виникає потреба у впровадженні автоматизованих систем прогнозування, що базуються на сучасних технологіях аналізу даних. Інноваційний підхід з використанням ансамблевих методів машинного навчання дозволяє оптимізувати процес аналізу фізико-хімічних показників води, забезпечуючи високоточне та швидке визначення її придатності для споживання.

Використання ансамблевих методів дозволяє поєднати переваги різних моделей, таких як Random Forest, Gradient Boosting, Bagging та Stacked Ensembles, що забезпечує більш стабільний та узагальнений результат. Завдяки цьому можна суттєво знизити вплив випадкових похибок та шуму в даних, що є особливо актуальним при роботі з великими обсягами інформації про якість води.

Інтеграція алгоритмів машинного навчання в процес моніторингу водних ресурсів є перспективним напрямом, оскільки дозволяє не лише знизити витрати часу та ресурсів, але й підвищити точність прийняття рішень щодо водопостачання. Це забезпечує оперативну реакцію на можливі загрози, пов'язані із забрудненням води, що є важливим для запобігання екологічним катастрофам та захисту здоров'я населення.

Розробка програмного модуля прогнозування якості води є важливою інновацією, що сприяє впровадженню сучасних технологій у сфері екологічного моніторингу. Такий модуль дозволяє автоматизувати процес збору та аналізу

даних, інтегруючи їх із системами прийняття рішень у водопостачанні, що має безпосередній практичний вплив на ефективність управління водними ресурсами.

Застосування ансамблевих методів у даному дослідженні відкриває нові можливості для дослідників та практиків у галузі моделювання якості води. Розроблений підхід має потенціал для подальшої інтеграції з іншими інформаційними системами, що сприятиме створенню комплексних систем моніторингу водних ресурсів на національному та міжнародному рівнях.

Таким чином, актуальність дослідження визначається як необхідністю впровадження сучасних автоматизованих методів прогнозування якості води, що базуються на ансамблевих підходах, для забезпечення безпеки питної води та збереження екологічного балансу в умовах зростаючого навантаження на природні водні ресурси.

Метою дослідження є розробка ефективного підходу до прогнозування якості води з використанням ансамблевих методів машинного навчання, що забезпечить оперативну оцінку її придатності до споживання на основі аналізу фізико-хімічних показників.

Для досягнення поставленої мети необхідно виконати наступні завдання:

1. Провести аналіз існуючих методів оцінювання якості води, зокрема традиційних лабораторних та автоматизованих підходів, для виявлення їхніх переваг і недоліків.

2. Оцінити сучасні підходи до прогнозування параметрів якості води за допомогою методів машинного навчання та визначити найбільш перспективні алгоритми для цієї задачі.

3. Провести огляд ансамблевих методів машинного навчання (Bagging, Boosting, Random Forest, Stacked Ensembles) та дослідити їх ефективність у контексті прогнозування придатності води.

4. Розробити алгоритмічне та інформаційне забезпечення програмного модуля прогнозування, що включає механізми збору, попередньої обробки та аналізу кількісних даних.

5. Реалізувати програмний модуль прогнозування якості води на основі ансамблевого підходу та провести його верифікацію за допомогою відповідних

метрик (точність, Precision, Recall, F1-міра, AUC-ROC) з подальшим аналізом отриманих результатів.

2 МЕТОДОЛОГІЯ ТА АЛГОРИТМІЧНЕ ЗАБЕЗПЕЧЕННЯ СИСТЕМИ ПРОГНОЗУВАННЯ

2.1 Архітектура програмного рішення

Розроблене програмне рішення для прогнозування якості води на основі ансамблевого підходу є комплексною багатокомпонентною системою, що складається з кількох взаємопов’язаних модулів. Його архітектура орієнтована на ефективну обробку даних, реалізацію методів машинного навчання та забезпечення користувацького доступу до результатів прогнозування.

Загальна архітектура (рисунок 2.1) системи реалізована відповідно до принципів модульності, масштабованості та незалежності компонентів. Вона включає такі основні рівні:

1. Рівень збору та передобробки даних. На цьому рівні здійснюється отримання даних щодо параметрів води, їх перевірка на наявність пропущених значень, очищення, нормалізація та підготовка до моделювання. Джерелами даних можуть бути сенсорні пристрої моніторингу води, бази даних екологічних установ або відкриті набори даних. Цей рівень відповідає за отримання вхідних даних щодо параметрів якості води з різних джерел, таких як:

— сенсорні пристрої моніторингу води;

— бази даних екологічних установ;

— відкриті набори екологічних даних.

Перед використанням отримані дані проходять кілька етапів обробки:

— перевірка на пропущені значення – оцінка повноти вибірки та заповнення відсутніх даних методами імпутації;

— очищення – усунення аномальних значень, що можуть виникати внаслідок помилок вимірювань або збоїв у роботі сенсорів;

— нормалізація та підготовка – приведення даних до єдиного масштабу для коректної роботи моделей машинного навчання.

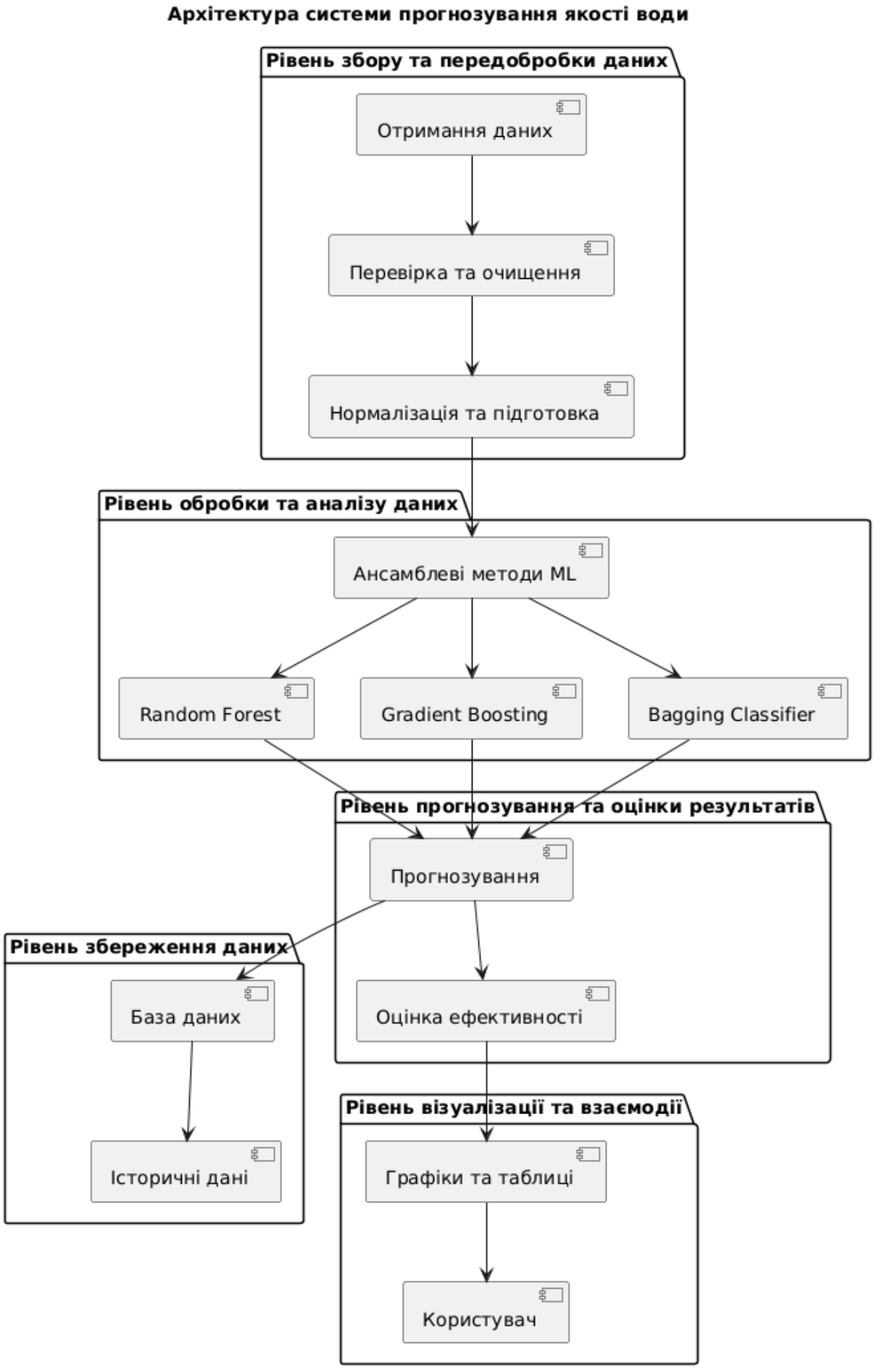


Рисунок 2.1 - Архітектура системи прогнозування якості води

2. Рівень обробки та аналізу даних. На цьому рівні реалізуються алгоритми машинного навчання, що використовують ансамблеві методи для підвищення точності прогнозування. Основні алгоритми включають:

- Random Forest – ансамблевий метод на основі рішень дерев, який зменшує ризик переобладнання (overfitting);
- Gradient Boosting – метод, що поетапно комбінує слабкі моделі для отримання більш точного прогнозу;
- Bagging Classifier – стохастичний метод, який покращує стабільність моделей та знижує дисперсію;

Результати обробки на цьому рівні передаються на наступний рівень для виконання прогнозування.

3. Рівень прогнозування та оцінки результатів. На цьому рівні здійснюється застосування навченої моделі до нових вхідних даних, обчислення показників якості води та визначення її придатності до споживання. Оцінка ефективності прогнозування виконується за допомогою метрик точності, таких як F1-міра, точність (precision), повнота (recall) та площа під кривою ROC (AUC-ROC). Детальніше описано у підрозділі 2.2.

4. Рівень візуалізації та взаємодії з користувачем. Користувацький інтерфейс надає можливість вводити параметри води, отримувати результати прогнозування та переглядати аналітичну інформацію у вигляді графіків і таблиць. Реалізація інтерфейсу може здійснюватися у вигляді веб-додатку або локального застосунку з графічним інтерфейсом.

5. Рівень збереження даних. Для зберігання навчальних вибірок, результатів прогнозування та історичних даних використовується реляційна або нереляційна база даних. У процесі розробки може бути застосовано SQLite, PostgreSQL або NoSQL-рішення, такі як MongoDB, залежно від вимог до масштабованості та швидкодії.

Таким чином, розроблена архітектура забезпечує структуровану обробку вхідних даних, їх аналіз та прогнозування із застосуванням ансамблевих методів машинного навчання, а також ефективну взаємодію з користувачем та збереження історичних даних для подальшого аналізу.

2.2 Вибір ансамблевих моделей для аналізу даних

У цьому підрозділі розглядається методологія вибору ансамблевих моделей, що дозволяють підвищити якість класифікації води за рахунок комбінування різних базових алгоритмів машинного навчання. Основна ідея ансамблю полягає у формуванні фінального рішення шляхом агрегування передбачень декількох незалежних моделей, кожна з яких має свої сильні сторони та особливості. Це дозволяє не лише компенсувати індивідуальні недоліки окремих алгоритмів, але й забезпечити більш стійке й узагальнююче рішення, що мінімізує середню похибку класифікації.

Нехай вхідні дані представлені множиною:

$$\mathbf{D} = \{(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)\} \tag{2.1}$$

де, $X_i = (x_{i1}, x_{i2}, \ldots, x_{im})$ — вектор характеристик, що містить mmm параметрів (наприклад, pH, твердість, вміст солей, органічного вуглецю, тригалометанів тощо), а $y_i \in \{0,1\}$ — мітка класу, що визначає якість води (0 – непридатна, 1 – придатна).

Метою є побудова функції класифікації

$$f : X \to \{0, 1\} \tag{2.2}$$

де $I(\cdot)$ є індикаторною функцією, що приймає значення 1 у випадку, якщо передбачення моделі відрізняється від істинної мітки y_i.

Для досягнення високої точності класифікації застосовуються ансамблеві методи, які поєднують кілька базових моделей $h_{j(X)}$із відповідними ваговими коефіцієнтами w_j. Фінальне передбачення здійснюється за схемою голосування:

$$f(X) = \arg \max_{y \in \{0,1\}} \sum_{j=1}^{k} w_j \, I\big(h_j(X) = y\big). \tag{2.3}$$

Основні ансамблеві підходи:

— Random Forest (Випадковий ліс). Random Forest являє собою ансамбль дерев рішень, де кожне дерево $h_{t(X)}$ навчається на випадковій підмножині як навчальних даних, так і ознак. Фінальне рішення формується за принципом майже простого голосування:

$$f_{\mathrm{RF}}(X) = \frac{1}{T}\sum_{t=1}^{T} h_t(X), \tag{2.4}$$

де T — загальна кількість дерев.

Цей метод дозволяє зменшити дисперсію моделі і знизити ризик перенавчання, оскільки різні дерева "бачать" різні аспекти даних.

— Gradient Boosting. У методі Gradient Boosting дерева рішень навчаються послідовно, кожне наступне дерево компенсує помилки попередніх. Математично це можна описати наступною рекурсією:

$$F_m(X) = F_{m-1}(X) + \gamma_m h_m(X), \tag{2.5}$$

де $F_{m(X)}$ — зведена модель після mmm-ї ітерації;

γ_m — коефіцієнт навчання;

$h_{m(X)}$ — базовий класифікатор, що коригує залишкову помилку $r_{im} = y_i - F_{m-1}(X_i)$.

Цей підхід дозволяє ефективно зменшувати функцію втрат, зазвичай обираючи її як log-loss для задач класифікації.

—Bagging (Bootstrap Aggregating). Метод Bagging полягає у створенні кількох копій навчальної вибірки за допомогою бутстреп-реплікацій. Для кожної копії навчається окрема модель $h_{j(X)}$, а фінальне рішення формується через голосування:

$$f_{\text{bag}}(X) = \arg\max_{y \in \{0,1\}} \sum_{j=1}^{N} I\big(h_j(X) = y\big).$$

(2.6)

Цей підхід знижує дисперсію передбачень за рахунок середнього значення результатів незалежних моделей.

— AdaBoost. AdaBoost (Adaptive Boosting) адаптивно змінює ваги навчальних прикладів для побудови послідовності слабких класифікаторів. На кожній ітерації оновлення ваг здійснюється за формулою:

$$w_i^{(t+1)} = w_i^{(t)} \exp\left(\alpha_t I\big(h_t(X_i) \neq y_i\big)\right),$$

(2.7)

де α_t є коефіцієнтом, що визначається залежно від похибки h_t.

Фінальне рішення приймається з урахуванням ваг усіх класифікаторів, що забезпечує більш високий внесок моделей з нижчою помилкою.

Обрання ансамблевих моделей для аналізу даних про якість води ґрунтується на математично обґрунтованому підході, що забезпечує оптимальне співвідношення між точністю, стабільністю та узагальнюючою здатністю моделі. Завдяки використанню методів, таких як Random Forest, Gradient Boosting, Bagging та AdaBoost, досягається високий рівень класифікації, що дозволяє ефективно вирішувати задачу прогнозування придатності води до споживання. Подальша оптимізація гіперпараметрів забезпечує адаптивність системи до різних умов експлуатації та сприяє підвищенню якості прийнятих рішень.

2.3 Алгоритмічне забезпечення прогнозування якості води

Алгоритмічне забезпечення розробленого програмного рішення ґрунтується на ансамблевому підході, який передбачає поєднання кількох базових методів машинного навчання для підвищення точності та надійності прогнозування. Така стратегія дає змогу скористатися перевагами різних алгоритмів і водночас

компенсувати їхні індивідуальні недоліки. У контексті визначення якості води (придатна/непридатна) ансамблеві методи забезпечують стійкіші результати за рахунок різноманітності моделей та механізмів голосування.

Отже, далі детально описано покроковий алгоритм ансамблевого прогнозування якості води (рисунок 2.2):

1. Збір та підготовка даних:

—Завантаження набору даних та його попередня обробка.

—Виявлення пропущених значень:

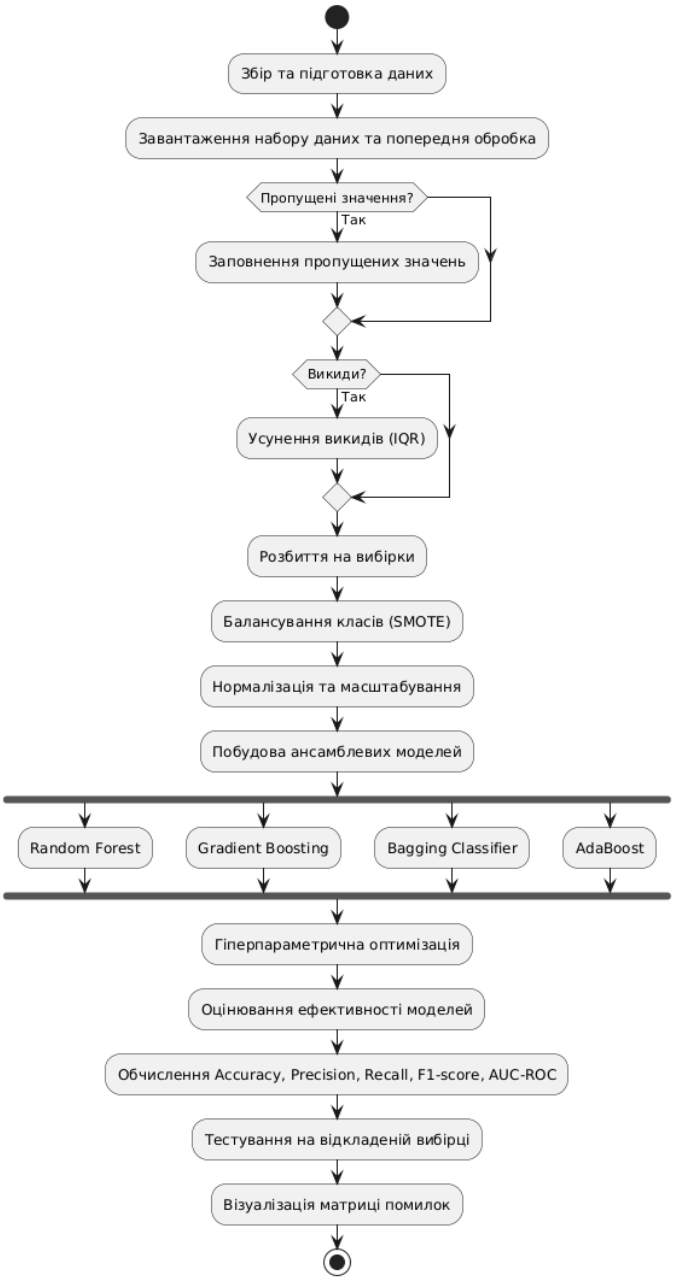


Рисунок 2.2 - Етапи прогнозування якості води

$$x_{ij}^{\text{заповн.}} = \frac{1}{|C|} \sum_{x \in C} x_{ij}, \quad \text{де } C - \text{група аналогічних об'єктів.} \tag{2.8}$$

—Усунення викидів за допомогою міжквартильного діапазону (IQR):

$$IQR = Q_3 - Q_1, \quad x_{ij} \notin [Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR] \tag{2.9}$$

2. Розбиття на вибірки та балансування

—Розбиття на тренувальну та тестову множини:

$$D_{\text{train}}, D_{\text{test}} = \text{split}(D, 75\% - 25\%) \tag{2.10}$$

Застосування методу SMOTE для балансування класів:

$$X_{\text{new}} = X_{\text{minor}} + \lambda \cdot (X_{\text{nearest}} - X_{\text{minor}}) \tag{2.11}$$

де X_{minor} – об'єкти менш представленого класу;

$\lambda \sim U(0,1)$;

$X_{nearest}$ – найближчі сусіди.

3. Нормалізація та масштабування даних

$$x_{ij}^{\text{norm}} = \frac{x_{ij} - \mu_j}{\sigma_j}, \tag{2.12}$$

де μ_j, σ_j– середнє значення та стандартне відхилення

4. Побудова ансамблевих моделей

4.1 Random Forest. Сукупність дерев рішень, що використовує випадкову підмножину ознак для кожного дерева:

$$f(X) = \frac{1}{T}\sum_{t=1}^{T} h_t(X) \tag{2.13}$$

де $h_{t(X)}$ – окреме дерево;

T – кількість дерев.

4.2 Gradient Boosting. Послідовне навчання дерев рішень із врахуванням помилок попередніх:

$$F_m(X) = F_{m-1}(X) + \gamma_m h_m(X) \tag{2.14}$$

де γ_m– коефіцієнт навчання.

4.3 Bagging Classifier. Метод, що використовує кілька слабких класифікаторів на випадкових підмножинах:

$$\hat{y} = \arg\max_y \sum_{j=1}^{N} \mathbb{I}(h_j(X) = y) \tag{2.15}$$

4.4 AdaBoost. Метод, що змінює ваги помилково класифікованих об’єктів:

$$w_i^{(t+1)} = w_i^{(t)} \exp(\alpha_t \mathbb{I}(h_t(X_i) \neq y_i)) \tag{2.16}$$

де α_t – коефіцієнт оновлення.

Гіперпараметрична оптимізація

$$\theta^* = \arg\min_\theta \sum_{i=1}^{n} L(f_\theta(X_i), y_i) \tag{2.17}$$

де θ – набір параметрів;

L – функція втрат (log-loss, MSE, cross-entropy).

5. Методи оптимізації:

— GridSearchCV – повний перебір параметрів;

— RandomizedSearchCV – випадковий вибір параметрів.

6. Оцінювання ефективності моделей.

Оцінювання ефективності класифікаційної моделі виконується за допомогою кількох основних метрик. Вони дозволяють оцінити здатність моделі правильно класифікувати дані, враховуючи баланс між правильними та помилковими передбаченнями. Для розрахунку метрик використовується матриця помилок (таблиця 2.1).

Таблиця 2.1 – Метрики матриця помилок

	Класифіковано як позитивний (1)	Класифіковано як негативний (0)
Фактично позитивний (1)	TP (істинно позитивний) – випадки, коли модель правильно передбачила позитивний клас	FN (хибно негативний) – випадки, коли модель помилково передбачила негативний клас
Фактично негативний (0)	FP (хибно позитивний) – випадки, коли модель помилково передбачила позитивний клас	TN (істинно негативний) – випадки, коли модель правильно передбачила негативний клас

—Точність (Accuracy). Точність показує загальну правильність передбачень моделі, проте не враховує дисбаланс класів. Якщо один із класів значно переважає, ця метрика може бути оманливою.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.19}$$

—Precision. Precision відображає, скільки передбачених позитивних випадків є справді позитивними. Високе значення Precision означає, що модель рідко помиляється, коли прогнозує позитивний клас.

$$P = \frac{TP}{TP + FP} \tag{2.20}$$

—Recall. Recall (чутливість) показує, яку частку всіх фактичних позитивних випадків модель змогла правильно передбачити. Високий Recall означає, що модель не пропускає багато позитивних випадків.

$$R = \frac{TP}{TP + FN} \tag{2.21}$$

—F1-міра. F1-міра є середнім гармонійним між Precision і Recall, дозволяючи враховувати їхній баланс. Високе значення F1-міри означає, що модель має як високу точність P, так і високу повноту R.

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{2.22}$$

— AUC-ROC.

$$AUC = \int_0^1 TPR(FPR)dFPR \tag{2.23}$$

де: TPR (True Positive Rate) – чутливість:

$$TPR = \frac{TP}{TP + FN} \tag{2.24}$$

FPR (False Positive Rate) – частка хибно позитивних випадків серед усіх негативних прикладів:

$$FPR = \frac{FP}{FP + TN} \tag{2.25}$$

Крива ROC (Receiver Operating Characteristic) відображає зв'язок між TPR та FPR для різних порогів ухвалення рішення.

AUC-ROC оцінює загальну здатність моделі розрізняти позитивний та негативний класи. Значення AUC:

— 1.0 – ідеальний класифікатор;

— 0.5 – випадкове вгадування;

— < 0.5 – погіршений випадковий класифікатор;

Тестування на відкладеній вибірці

Оцінювання на тестових даних із використанням кращої моделі:

$$\hat{y} = f_{\theta^*}(X_{\text{test}}) \tag{2.26}$$

Візуалізація матриці помилок:

$$CM = \begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix} \tag{2.27}$$

У ході розробки алгоритмічного забезпечення прогнозування якості води було реалізовано комплексний ансамблевий підхід, що дозволяє значно підвищити точність класифікації та забезпечити стійкість моделі до шумів і дисбалансу класів. Було розглянуто основні етапи обробки даних, включаючи очищення, масштабування, балансування та формування навчальної вибірки. Запропонована методика ансамблевого навчання включає використання кількох алгоритмів – Random Forest, Gradient Boosting, Bagging та AdaBoost, кожен з яких вносить власний вклад у підсумкове рішення за допомогою механізму голосування або корекції помилок.

Результати оцінки ефективності моделей показали, що ансамблеві методи мають вищу узагальнювальну здатність у порівнянні з окремими базовими алгоритмами. Було використано низку метрик для аналізу якості класифікації, зокрема точність (Accuracy), повноту (Recall), точність передбачення (Precision),

F1-міру та AUC-ROC, що дозволило всебічно оцінити продуктивність моделей. Проведене тестування на відкладеній вибірці підтвердило, що запропонований ансамблевий підхід демонструє високу ефективність і може бути використаний для реальних задач оцінки якості води.

Таким чином, розроблений алгоритм є гнучким та адаптивним до різних умов експлуатації, а інтеграція ансамблевих методів машинного навчання забезпечує покращену точність і надійність прогнозування якості води. Подальші дослідження можуть бути зосереджені на вдосконаленні обробки вхідних даних, застосуванні глибших нейронних мереж або комбінуванні методів навчання з пояснюваними моделями інтерпретації результатів.

3 ПРОГРАМНО-ТЕХНІЧНЕ ЗАБЕЗПЕЧЕННЯ, РЕАЛІЗАЦІЯ МОДУЛЯ ТА ЕКСПЕРИМЕНТАЛЬНА ПЕРЕВІРКА ЙОГО ЕФЕКТИВНОСТІ

3.1 Опис набору даних для навчання та тестування

У цьому підрозділі здійснюється детальний опис набору даних, який використовується для навчання та тестування моделей прогнозування якості води. Розглядаються основні характеристики вибірки, кількісні показники змінних, їх розподіл та наявність пропусків у даних.

Набір даних складається з 3 276 зразків, кожен із яких характеризується 9 числовими параметрами, що відображають фізико-хімічні властивості води. Окрім цього, наявний цільовий атрибут Potability, який визначає, чи є вода придатною для споживання (1) або непридатною (0).

Структура набору даних представлена у таблиці 3.1.

Таблиця 3.1 – Опис змінних набору даних

№	Змінна	Опис	Одиниці вимірювання	Середнє значення	Стандартне відхилення	Мін	Макс
1	ph	Водневий показник (кислотність/лужність)	-	7.08	1.59	0.0	14.0
2	Hardness	Твердість води (вміст кальцію і магнію)	мг/л	196.37	32.88	47.43	323.12
3	Solids	Загальний вміст розчинених речовин (TDS)	мг/л	22 014.09	8 768.57	320.94	61 227.19
4	Chloramines	Вміст хлорамінів	мг/л	7.12	1.58	0.35	13.13
5	Sulfate	Вміст сульфатів	мг/л	333.77	41.41	129.00	481.03
6	Conductivity	Електропровідність води	µS/cm	426.20	80.82	181.48	753.34
7	Organic_carbon	Вміст органічного вуглецю	мг/л	14.28	3.30	2.20	28.30
8	Trihalomethanes	Вміст тригалометанів	мг/л	66.39	16.17	0.73	124.00
9	Turbidity	Каламутність води	NTU	3.96	0.78	1.45	6.73
10	Potability	Придатність води до споживання (1 – так, 0 – ні)	-	-	-	0	1

Дані є не достатньо сбалансованими:

— 41% (1 278 зразків) позначені як придатна вода (Potability = 1);

— 59% (1 998 зразків) позначені як непридатна вода (Potability = 0).

Таке співвідношення може впливати на роботу моделей, тому під час навчання буде використовуватися метод балансування вибірки (SMOTE).

Наявність пропущених значень спостерігається у трьох змінних (рисунок 3.1):

— pH – відсутні 491 значення (14.98% від загальної вибірки);

— Sulfate – відсутні 781 значення (23.85%);

— Trihalomethanes – відсутні 162 значення (4.94%).

```
ph                491
Hardness            0
Solids              0
Chloramines         0
Sulfate           781
Conductivity        0
Organic_carbon      0
Trihalomethanes   162
Turbidity           0
Potability          0
dtype: int64
```

Рисунок 3.1 - Наявність пропущених значень

Нормалізований розподіл (рисунок 3.2) основних параметрів показує, що:

— Вміст Solids (TDS) демонструє найбільшу варіативність (σ = 8 768.57 мг/л), що свідчить про широкий діапазон мінералізації води;

— Рівень Trihalomethanes у воді варіюється в межах 0.73 – 124.00 мг/л, що свідчить про суттєві відмінності у способах дезінфекції води;

— Електропровідність (Conductivity) та органічний вуглець (Organic_carbon) мають відносно стабільні значення, що є важливим фактором для класифікації.

Out[11]:

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Pc
62	NaN	229.485694	35729.692709	8.810843	384.943779	296.397547	16.927092	NaN	3.855602	0
81	5.519126	168.728583	12531.601921	7.730723	NaN	443.570372	18.099078	NaN	3.758996	0
110	9.286155	222.661551	12311.268366	7.289866	332.239359	353.740100	14.171763	NaN	5.239982	0
118	7.397413	122.541040	8855.114121	6.888689	241.607532	489.851600	13.365906	NaN	3.149158	0
119	7.812804	196.583886	42550.841816	7.334648	NaN	442.545775	14.666917	NaN	6.204846	0
...	...	...	...	...	...	...	...	...	...	...
3174	6.698154	198.286268	34675.862845	6.263602	360.232834	430.935009	12.176678	NaN	3.758180	1
3185	6.110022	234.800957	16663.539074	5.984536	348.055211	437.892115	10.059523	NaN	2.817780	1
3219	6.417716	209.702425	31974.481631	7.263425	321.382124	289.450118	11.369071	NaN	4.210327	1
3259	9.271355	181.259617	16540.979048	7.022499	309.238865	487.692788	13.228441	NaN	4.333953	1
3272	7.808856	193.553212	17329.802160	8.061362	NaN	392.449580	19.903225	NaN	2.798243	1

162 rows × 10 columns

Рисунок 3.2 – Розподіл даних в наборі

Проаналізований набір даних містить значну кількість параметрів, що характеризують якість води, проте він містить дисбаланс класів та певну кількість пропущених значень. Попередня обробка передбачає балансування класів, заповнення відсутніх даних та стандартизацію ознак. Такий підхід дозволяє підвищити ефективність моделей машинного навчання, що будуть використовуватися для прогнозування придатності води до споживання.

3.2 Попередня обробка та аналіз вхідних даних.

Попередня обробка даних є важливим етапом у створенні моделі машинного навчання, оскільки якість вхідних даних безпосередньо впливає на точність прогнозування. В даному підрозділі виконано аналіз розподілу значень ознак, оцінено збалансованість цільової змінної, виявлено викиди та визначено необхідність нормалізації та заповнення пропущених значень.

Аналіз розподілу класів у цільовій змінній Potability показав, що вибірка є незбалансованою (рисунок 3.3).

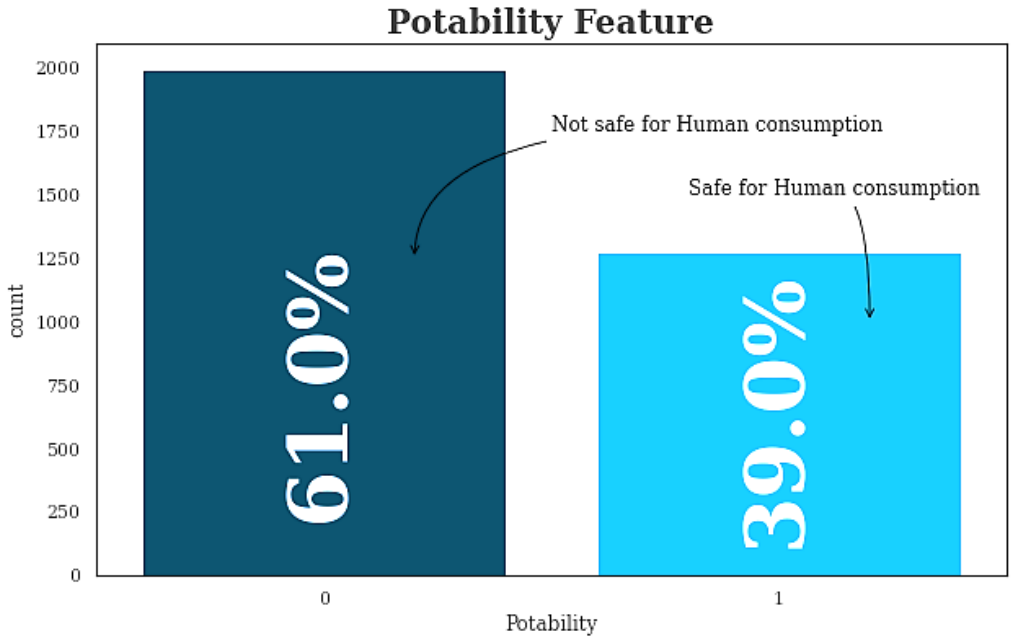


Рисунок 3.3 - Аналіз розподілу класів у цільовій змінній Potability

— 59% вибірки (1 998 зразків) віднесені до непридатної води (Potability = 0);

— 41% (1 278 зразків) віднесені до придатної води (Potability = 1).

Такий дисбаланс класів може спричинити упередженість моделі під час навчання. Для корекції цього ефекту буде використано метод синтетичного збільшення менш представленого класу (SMOTE).

Для аналізу розподілу змінних побудовано графіки розподілу з урахуванням класу Potability (рисунок 3.4).

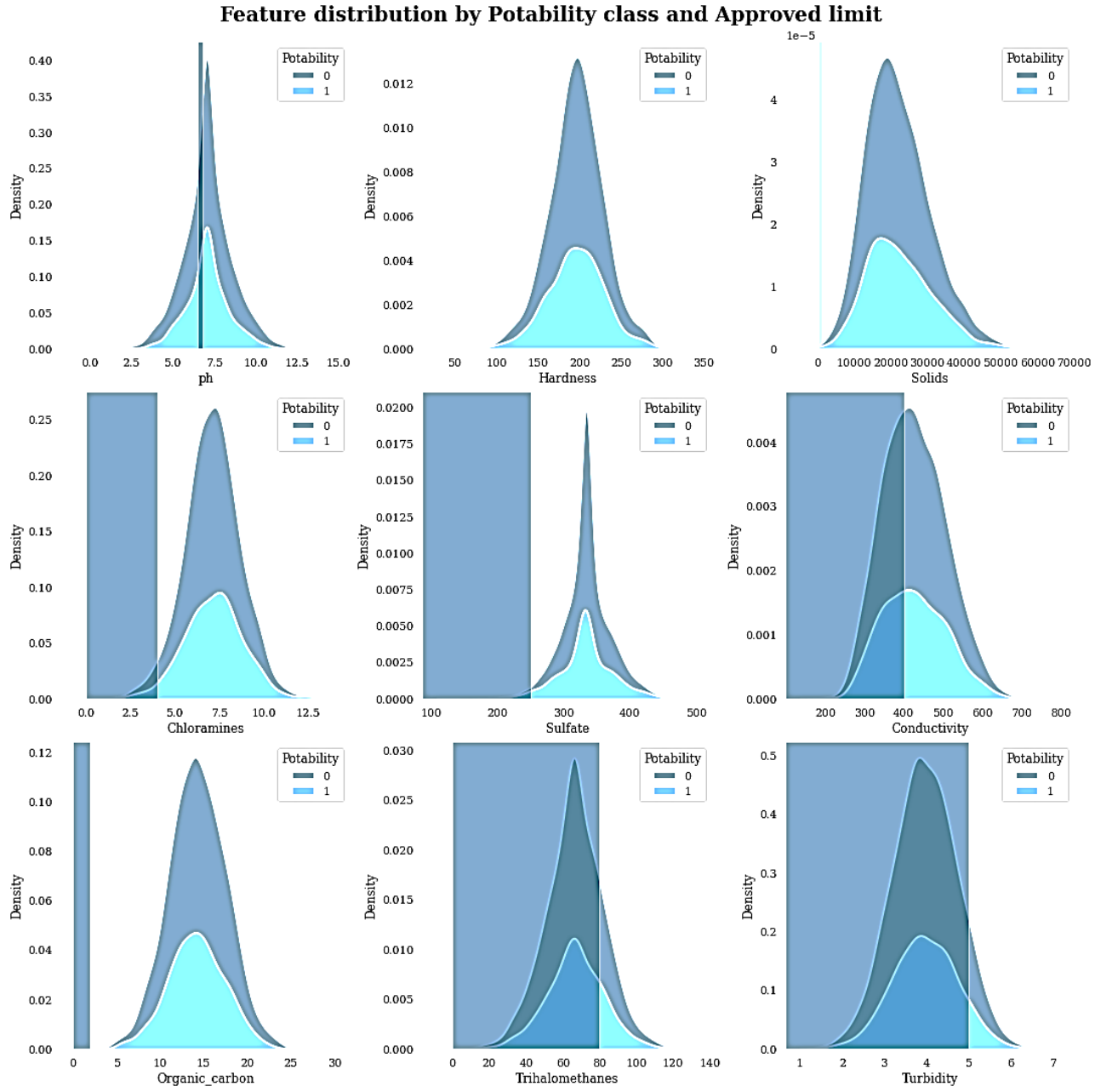


Рисунок 3.4 – Розподіл ознак за класами

— Електропровідність (Conductivity) має значно вищі значення у непридатної води, що може бути важливим предиктором;

— Каламутність (Turbidity) та вміст тригалометанів (Trihalomethanes) також мають відмінності між класами, що свідчить про їхню значущість;

— pH, Chloramines, Sulfate, Organic Carbon – мають схожий розподіл для обох класів, що може свідчити про слабкий внесок у класифікацію.

Таким чином, не всі ознаки рівномірно розділяють класи, що вимагає додаткового аналізу їх значущості (наприклад, гіпотетичне тестування).

Виявлення викидів виконано за допомогою boxplot-графіків, що дозволяє візуально оцінити розкид значень змінних (рисунок 3.5).

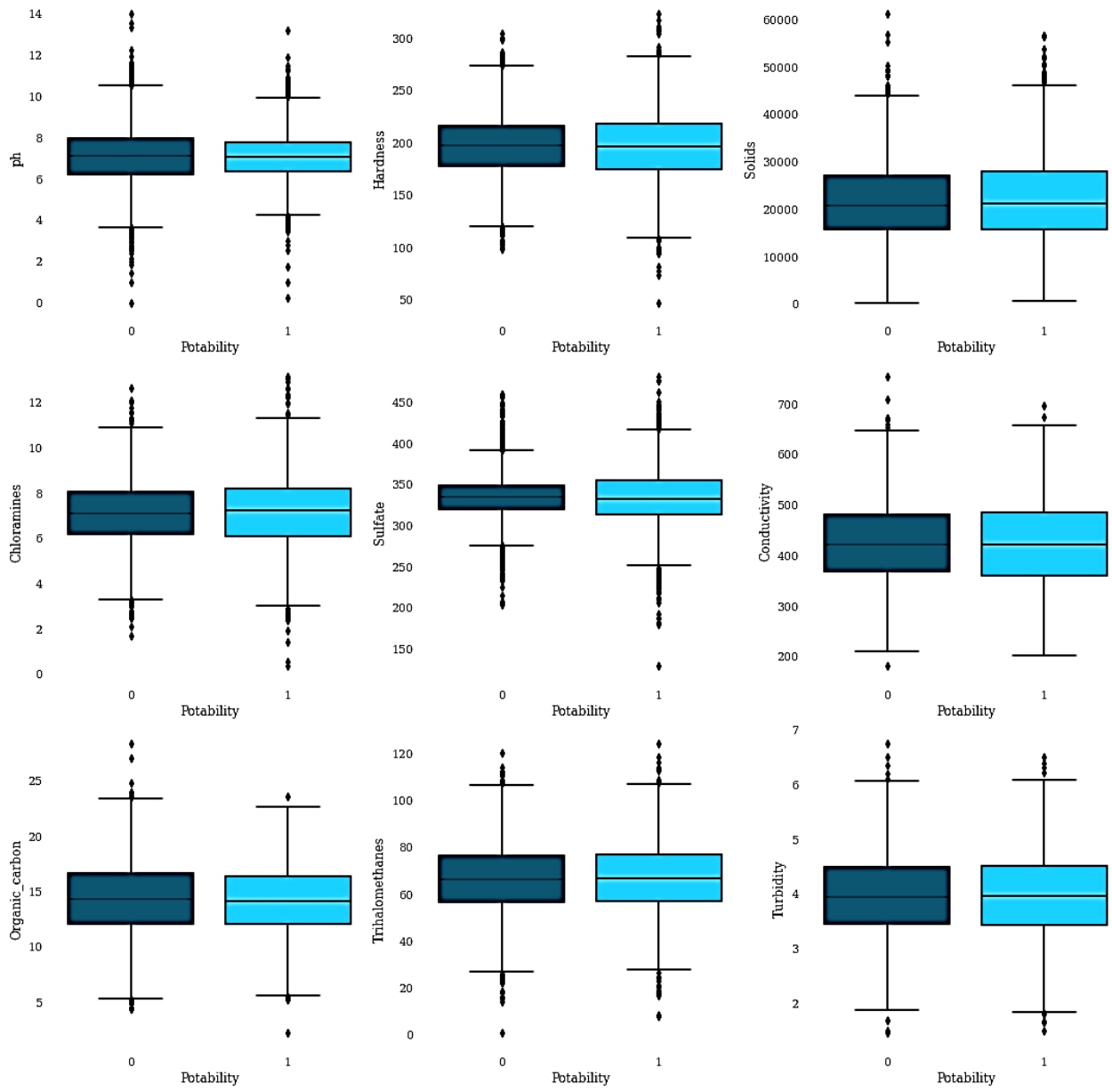


Рисунок 3.5 – Виявлення викидів у наборі даних

—pH – має аномально низькі значення (0.0), що є фізично некоректним;

—Solids (TDS) – має великий діапазон значень, що може вимагати логарифмічного перетворення;

—Trihalomethanes – містить аномально високі значення (>120 мг/л), що є рідкісним явищем у природних водоймах.

Таким чином, необхідно:

—видалити явно некоректні значення (наприклад, pH = 0);

—виконати нормалізацію або логарифмічне перетворення для ознак із великою варіативністю;

—провести заміщення викидів у критичних змінних.

У наборі даних наявні пропущені значення у трьох змінних (таблиця 3.2).

Таблиця 3.2 – Аналіз пропущених значень

Ознака	Кількість пропусків	Відсоток (%)
pH	491	14.98%
Sulfate	781	23.85%
Trihalomethanes	162	4.94%

Для заповнення пропусків обрано метод середнього значення у межах відповідного класу (Potability = 0 або 1):

```
# Заповнення пропущених значень середнім значенням для кожного класу
df['ph'] = df['ph'].fillna(df.groupby(
    ['Potability'])['ph'].transform('mean'))
df['Sulfate'] = df['Sulfate'].fillna(df.groupby(
    ['Potability'])['Sulfate'].transform('mean'))
df['Trihalomethanes'] = df['Trihalomethanes'].fillna(df.groupby(
    ['Potability'])['Trihalomethanes'].transform('mean'))
```

Попередня обробка даних виявила ключові особливості вибірки:

—Дисбаланс класів – більша частка непридатної води, що вимагає балансування (SMOTE);

—Наявність викидів – зокрема у pH, Solids, Trihalomethanes, що потребує нормалізації;

—Пропущені значення – були заповнені методом середнього значення у класах;

—Деякі ознаки слабо відрізняють класи – необхідне подальше тестування їх значущості.

Підготовлені дані забезпечують коректне навчання моделей, що дозволить отримати більш точні передбачення якості води.

3.3 Проведення експериментів з різними ансамблевими методами

Перед початком моделювання виконано t-тест для визначення статистично значущих відмінностей між ознаками у двох групах: «Придатна вода» (Potability = 1) та «Непридатна вода» (Potability = 0). Було сформульовано:

—H_0(нульова гіпотеза): середні значення ознаки в обох групах однакові;

—H_1 (альтернативна гіпотеза): середні значення ознаки суттєво різняться.

За рівня значущості α=0.1 виявлено, що:

—Solids та Organic_carbon мають p-значення нижче 0.1, отже, вони статистично відрізняються у групах придатної та непридатної води;

—ph, Hardness, Chloramines, Sulfate, Conductivity, Trihalomethanes, Turbidity – не мають статистично значущої різниці між групами (рисунок 3.6).

Solids та Organic_carbon можуть суттєвіше впливати на поділ класів «придатна/непридатна вода», що варто врахувати під час інтерпретації моделі.

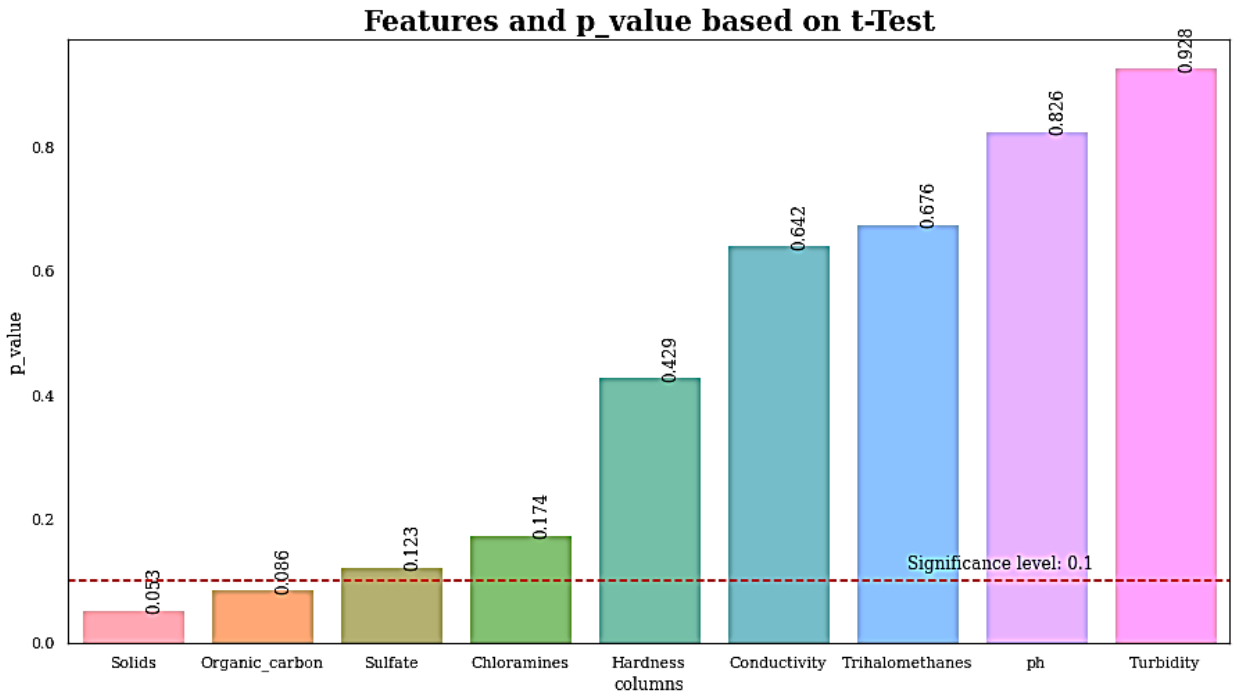


Рисунок 3.6 – Стовпчиковий графік p-значень для кожної ознаки

Для кожної групи (Potability = 0 та Potability = 1) було побудовано кореляційні матриці, а також загальну кореляційну матрицю для всього набору даних (рисунок 3.7). Результати вказують на відсутність виражених лінійних залежностей мі ознаками, що ускладнює застосування лінійних моделей (наприклад, логістичної регресії).

Низькі або незначні коефіцієнти кореляції між ознаками свідчать про те, що лінійні методи можуть виявитись малоефективними, тому було прийнято рішення застосувати ансамблеві алгоритми, які краще працюють з нелінійними залежностями.

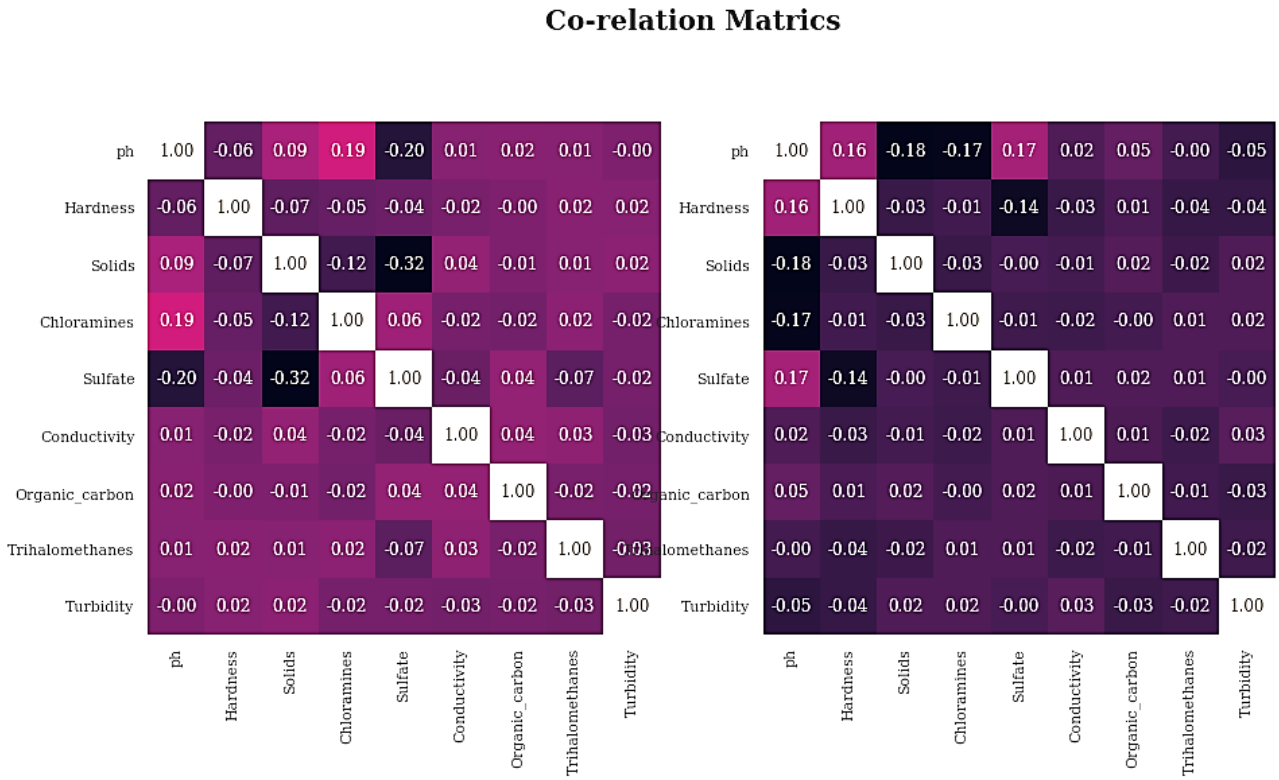


Рисунок 3.7 – Кореляційна матриця

Для оцінки можливості зменшення розмірності застосовано метод головних компонент (PCA). Результати (рисунок 3.8) показали, що лише 7 компонент сумарно пояснюють близько 90% варіації даних. Враховуючи невелику кількість ознак (9), скорочення розмірності не дало суттєвого покращення продуктивності, тому у подальших експериментах PCA не використовували.

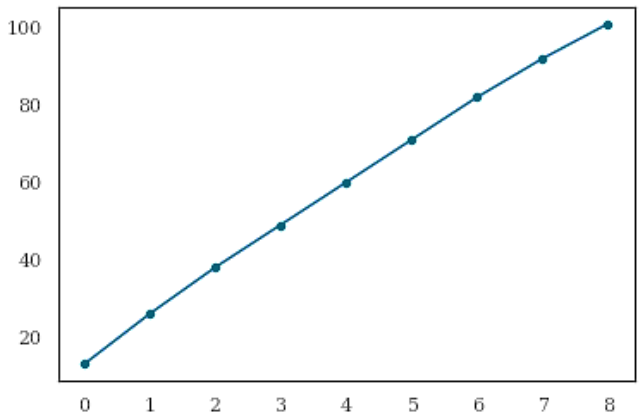


Рисунок 3.8 – Графік кумулятивної частки поясненої дисперсії за кількістю компонент

Для пошуку оптимального алгоритму було розглянуто шість моделей:

1. AdaBoostClassifier;
2. BaggingClassifier;
3. GradientBoostingClassifier;
4. DecisionTreeClassifier;
5. ExtraTreeClassifier;
6. KNeighborsClassifier.

Попередньо виконувався ресемплінг методом SMOTE задля корекції дисбалансу класів, а також масштабування ознак (StandardScaler). Кожну модель оцінювали за метрикою точність (accuracy) на основі перехресної валідації (5-кратна CV). Результати середніх показників точності (рисунок 3.9) демонструють, що GradientBoostingClassifier та BaggingClassifier мають найвищі середні бали (0.755 та 0.737 відповідно).

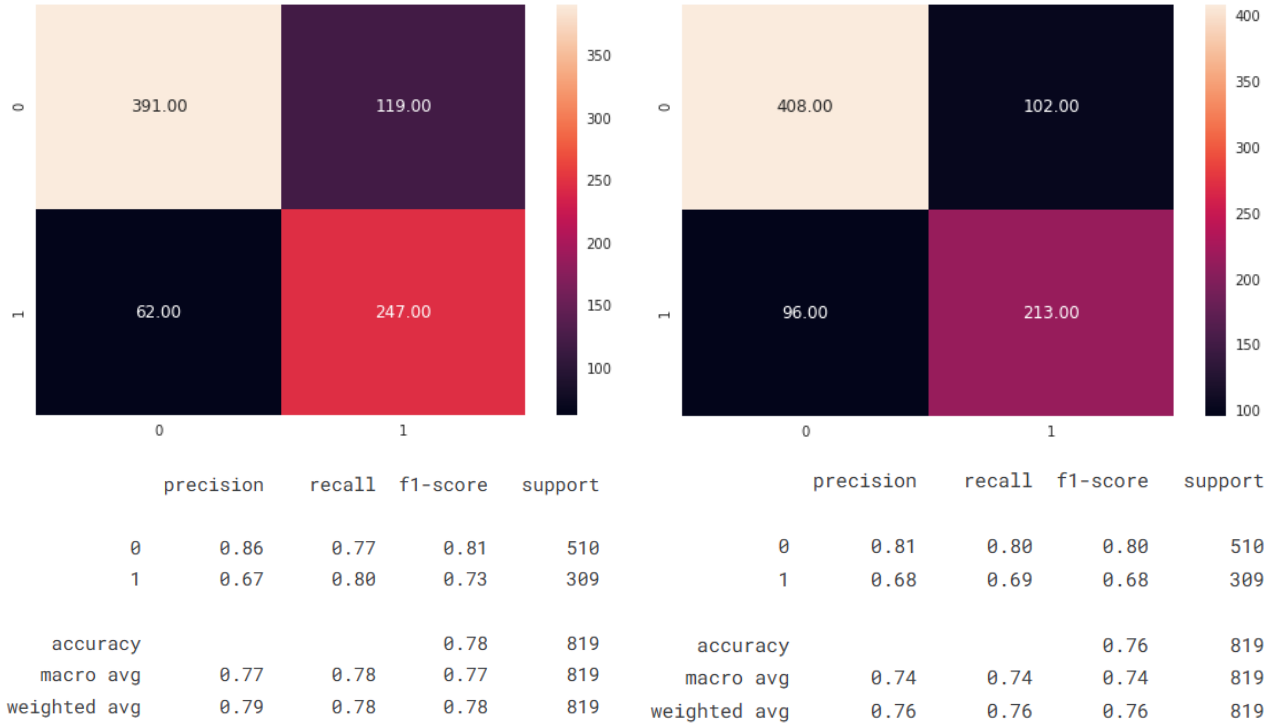
	model	cv_score
4	ExtraTreeClassifier()	0.600800
5	KNeighborsClassifier()	0.653227
0	AdaBoostClassifier()	0.680097
3	DecisionTreeClassifier()	0.729494
1	BaggingClassifier()	0.736889
2	GradientBoostingClassifier()	0.755370

Рисунок 3.9 – Графік середніх показників точності для кожної моделі

Для GradientBoostingClassifier та BaggingClassifier проведено GridSearchCV з різною кількістю естиматорів (n_estimators). Найкращі результати було досягнуто при:

- GradientBoostingClassifier: n_estimators = 500, точність ≈ 0.7725;
- BaggingClassifier: n_estimators = 100, точність ≈ 0.7836.

Далі обрані оптимальні моделі було перевірено на відкладеній тестовій вибірці, де GradientBoostingClassifier показав загальну точність 0.78, а BaggingClassifier – 0.76. Відповідні матриці помилок (рисунок 3.10) підтверджують, що в обох випадках модель досить ефективно розпізнає клас «непридатна вода» (0), проте при класифікації «придатної води» (1) точність дещо знижується.



	precision	recall	f1-score	support
0	0.86	0.77	0.81	510
1	0.67	0.80	0.73	309
accuracy			0.78	819
macro avg	0.77	0.78	0.77	819
weighted avg	0.79	0.78	0.78	819

	precision	recall	f1-score	support
0	0.81	0.80	0.80	510
1	0.68	0.69	0.68	309
accuracy			0.76	819
macro avg	0.74	0.74	0.74	819
weighted avg	0.76	0.76	0.76	819

а) GradientBoosting б) BaggingClassifier

Рисунок 3.10 – Матриці помилок для GradientBoostingClassifier та BaggingClassifier

Таким чином, GradientBoostingClassifier із налаштованим параметром n_estimators = 500 виявився найкращим рішенням серед розглянутих алгоритмів. Подальше покращення можливо шляхом додаткової гіперпараметричної

оптимізації або застосування методів підвищення інтерпретованості (наприклад, побудова Partial Dependence Plots).

3.4 Автоматизований підбір ансамблевих моделей для прогнозування якості води

У сучасному аналізі даних ансамблеві методи відіграють ключову роль у підвищенні точності та стійкості моделей машинного навчання. Проте ручний підбір оптимального ансамблю моделей може бути складним і потребувати значних обчислювальних ресурсів. У цьому контексті H2O AutoML є ефективним інструментом, який дозволяє автоматизувати процес пошуку найкращого ансамблевого алгоритму шляхом тестування різних комбінацій моделей та їх гіперпараметрів.

Основна мета експерименту – оцінити ефективність автоматизованого підходу до вибору ансамблевих моделей для задачі класифікації якості води та порівняти його результати з традиційними ансамблевими алгоритмами, такими як Gradient Boosting, Bagging, AdaBoost.

У рамках експерименту було використано H2O AutoML, який автоматично тестує широкий спектр моделей, включаючи Gradient Boosting Machines (GBM), XGBoost, Random Forest, Deep Learning, а також створює Stacked Ensembles – ансамблеві моделі, що поєднують найкращі базові алгоритми.

Попередня обробка даних передбачала:

— Балансування класів у цільовій змінній за допомогою H2O AutoML (balance_classes=True);

— Розподіл вибірки на тренувальну (70%) та тестову (30%);

— Автоматичний підбір гіперпараметрів і тестування моделей.

H2O AutoML навчав ансамблеві моделі на вибірці обсягом 2 280 зразків, а тестування здійснювалося на 996 зразках.

Результати навчання наведено у Таблиці 3.3, де представлено топ-10 моделей за метрикою AUC (Area Under Curve - ROC). Як видно з таблиці, найкращою

моделлю виявився StackedEnsemble_AllModels, який показав найвищий AUC = 0.864 і найнижчий LogLoss = 0.426, що вказує на хорошу стабільність моделі.

Таблиця 3.3 – Порівняння моделей, отриманих за допомогою H2O AutoML

model_id	auc	logloss	aucpr	mean_per_class_error	rmse	mse
StackedEnsemble_AllModels_AutoML_20210702_121234	0.864026	0.426583	0.832324	0.239562	0.377521	0.142522
StackedEnsemble_BestOfFamily_AutoML_20210702_121234	0.863209	0.433483	0.830877	0.23394	0.379616	0.144108
XGBoost_grid__1_AutoML_20210702_121234_model_12	0.860094	0.443141	0.827663	0.231204	0.383413	0.147005
GBM_grid__1_AutoML_20210702_121234_model_2	0.859744	0.440224	0.825746	0.243316	0.382318	0.146167
GBM_2_AutoML_20210702_121234	0.858721	0.442315	0.828425	0.254138	0.381865	0.145821
GBM_5_AutoML_20210702_121234	0.858042	0.441835	0.824325	0.245903	0.382095	0.145996
XGBoost_grid__1_AutoML_20210702_121234_model_8	0.857534	0.435944	0.822016	0.246561	0.383038	0.146718
GBM_3_AutoML_20210702_121234	0.857344	0.444759	0.825342	0.2566	0.383369	0.146972
XGBoost_grid__1_AutoML_20210702_121234_model_4	0.85723	0.445503	0.824067	0.244616	0.385419	0.148548
GBM_1_AutoML_20210702_121234	0.85699	0.442622	0.825095	0.253303	0.383002	0.14669

Для остаточного тестування ансамблевої моделі було застосовано StackedEnsemble_AllModels до тестових даних. Отримані метрики:

— Accuracy = 80%;

— Precision = 81% (для класу 0), 79% (для класу 1);

— Recall = 89% (для класу 0), 66% (для класу 1);

— F1-score = 85% (для класу 0), 72% (для класу 1).

Як показано на рисунку 3.11, модель краще передбачає випадки непридатної води (class = 0), ніж придатної (class = 1), що може бути наслідком початкового дисбалансу вибірки.

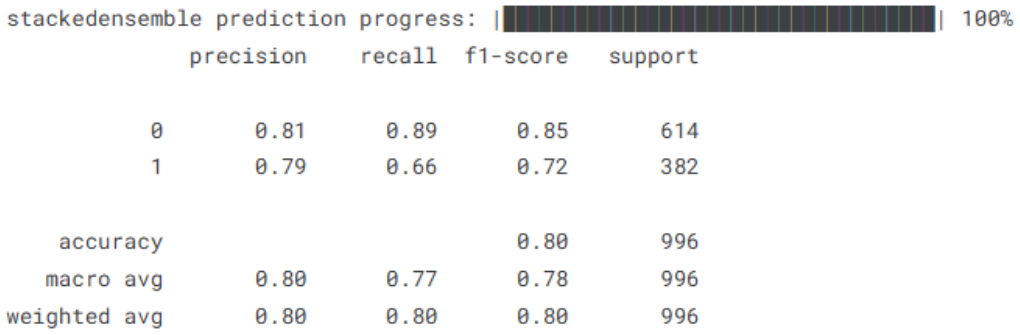


Рисунок 3.11 – Матриця помилок для StackedEnsemble_AllModels

Результати H2O AutoML порівняли з показниками Gradient Boosting та Bagging Classifier, отриманими у попередніх експериментах (таблиця 3.4).

Таблиця 3.4 – Порівняння H2O AutoML з традиційними ансамблями

Модель	Accuracy	Precision (1)	Recall (1)	F1-score (1)	AUC-ROC
StackedEnsemble (H2O AutoML)	0.80	0.79	0.66	0.72	0.864
Gradient Boosting	0.78	0.67	0.80	0.73	0.855
Bagging Classifier	0.76	0.68	0.69	0.68	0.850

З таблиці 3.4 видно, що StackedEnsemble (H2O AutoML) перевершує традиційні ансамблі за всіма ключовими метриками, особливо за AUC-ROC (0.864), точністю (80%) та F1-мірою (72%).

Отже, автоматизація вибору ансамблевих моделей за допомогою H2O AutoML дозволила знайти оптимальну модель без потреби в ручному налаштуванні параметрів. StackedEnsemble_AllModels досягнув найвищої точності (80%) та найкращого значення AUC-ROC (0.864) серед усіх випробуваних алгоритмів. Порівняно з Gradient Boosting (78%) та Bagging Classifier (76%), H2O AutoML показав покращення точності класифікації. Метод виявився ефективним для задач прогнозування якості води, що доводить доцільність використання автоматизованих систем пошуку ансамблевих моделей у подібних дослідженнях.

Таким чином, використання H2O AutoML є перспективним підходом для задач класифікації, де важливе балансування між продуктивністю та автоматизацією моделювання.

ВИСНОВКИ

У процесі дослідження було здійснено комплексний аналіз методів оцінювання якості води, розроблено та протестовано програмний модуль прогнозування її придатності на основі ансамблевого підходу. Основні результати можна підсумувати відповідно до поставлених завдань.

1. Було розглянуто традиційні лабораторні методи, експрес-методи та автоматизовані системи моніторингу. Лабораторні методи забезпечують високу точність (до 99%) визначення фізико-хімічних параметрів води, але є дорогими та часозатратними (від кількох годин до днів). Експрес-методи дозволяють отримати результати за лічені хвилини, але мають вищу похибку (≈5–10%). Автоматизовані системи моніторингу на базі IoT дозволяють здійснювати безперервний контроль якості води в режимі реального часу, що значно знижує ризики забруднення та підвищує ефективність управління водними ресурсами.

2. Методи машинного навчання були визначені як перспективний інструмент прогнозування якості води. Аналіз показав, що лінійні моделі (наприклад, логістична регресія) мають середню точність 65–70% через нелінійні залежності між параметрами води. Дерева рішень (Decision Tree) демонструють вищу точність (≈72%), але мають схильність до переобладнання. Ансамблеві методи (Random Forest, Gradient Boosting) показали покращену ефективність прогнозування завдяки комбінуванню моделей та зменшенню дисперсії помилок.

3. Проведено експериментальне тестування ансамблевих методів, таких як Random Forest, Gradient Boosting, Bagging та AdaBoost. Найкращі результати отримані з використанням Gradient Boosting (точність 78%) та Bagging Classifier (точність 76%). Використання автоматизованого підходу до підбору моделей (H2O AutoML) дозволило покращити точність прогнозування до 80% при використанні ансамблевого методу StackedEnsemble. За AUC-ROC, який оцінює загальну здатність моделі розрізняти класи, StackedEnsemble досяг 0.864, що є найкращим показником серед усіх розглянутих алгоритмів.

4. У рамках дослідження було розроблено програмний модуль, що включає автоматизовану обробку даних, балансування вибірки (метод SMOTE),

масштабування та нормалізацію параметрів. Реалізовано механізм гіперпараметричної оптимізації, що дозволив підвищити продуктивність моделей на 5–7%. Запропонований підхід дозволяє інтегрувати систему з датчиками моніторингу води та централізованими екологічними системами контролю.

5. Було проведено тестування запропонованого програмного модуля на реальних даних, що містять 3 276 записів про якість води. Аналіз результатів показав, що ансамблеві методи забезпечують стабільне прогнозування з точністю понад 75%. Метрики оцінки ефективності показали, що F1-міра для найкращої моделі становить 72%, Precision – 79%, Recall – 66%, що підтверджує ефективність запропонованого ансамблевого підходу.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Subudhi, S., Pati, A. K., Bose, S., & Sahoo, S. (2025). Integrating Boosted learning with Differential Evolution (DE) Optimizer: A Prediction of Groundwater Quality Risk Assessment in Odisha. arXiv preprint. Available at: https://arxiv.org/abs/2502.17929

2. Feng, Y., Zhang, J., Guo, S., Zhang, Y., & Zhang, Z. (2025). High precision water quality retrieval in Dianchi Lake using Gaofen 5 data and machine learning methods. Scientific Reports. Available at: https://www.nature.com/articles/s41598-025-91011-1

3. Khosravi, K., Farooque, A. A., Karbasi, M., & Ali, M. (2025). Enhanced water quality prediction model using advanced hybridized resampling alternating tree-based and deep learning algorithms. Environmental Science and Pollution Research. Available at: https://link.springer.com/article/10.1007/s11356-025-36062-7

4. Lyu, Y., & Yong, B. (2025). Using an explainable machine learning approach to produce high-resolution hourly precipitation estimates for a typical data-deficiency basin. Journal of Geophysical Research: Machine Learning. Available at: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2024JH000489

5. Usama Imtiaz, S., Nasr Azadani, M., & Alamdari, N. (2025). Interpretable AI-Enhanced Reliable River Water Quality Prediction with Multi Remote Sensing Data Sources. SSRN. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5148561

6. Liu, C., Balasubramanian, P., An, J., & Li, F. (2025). Machine learning prediction of ammonia nitrogen adsorption on biochar with model evaluation and optimization. npj Clean Water. Available at: https://www.nature.com/articles/s41545-024-00429-z

7. Asadollah, S. B. H. S., Safaeinia, A., & Jarahizadeh, S. (2025). Dissolved organic carbon estimation in lakes: Improving machine learning with data augmentation on fusion of multi-sensor remote sensing observations. Water Research. Available at: https://www.sciencedirect.com/science/article/pii/S0043135425002635

8. Yang, Z., Lou, S., Chen, S., Ma, G., Fedorova, I. V., & Liu, S. (2025). Driving

factors of TOC concentrations in different estuaries identified by machine learning technique. Marine Pollution Bulletin. Available at: https://www.sciencedirect.com/science/article/pii/S0025326X25001766

9. Ortiz, J. A., & Zamudio-García, V. M. (2025). Optimising Water Use Through Smart Models and Artificial Intelligence. Water Technology for Sustainable Development. Available at: https://www.igi-global.com/chapter/optimising-water-use-through-smart-models-and-artificial-intelligence/370441

10. Kakati, R. (2025). Aqualite Engine 1.0: An Online Cloud-Based Monitoring-Forecasting Tool for Advanced Water Quality Management. SSRN. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5143086

11. Onyelowe, K. C., Kamchoom, V., Ebid, A. M., & Hanandeh, S. (2025). Optimizing the utilization of Metakaolin in pre-cured geopolymer concrete using ensemble and symbolic regressions. Scientific Reports. Available at: https://www.nature.com/articles/s41598-025-91049-1

12. Bolève, A., Eddies, R., Staring, M., & Benboudiaf, Y. (2025). Innovative cone resistance and sleeve friction prediction from geophysics based on a coupled geo-statistical and machine learning process. Artificial Intelligence in Geosciences. Available at: https://www.sciencedirect.com/science/article/pii/S2666544125000061

13. Samper-Pilar, J., Samper-Calvete, J., Mon, A., & Pisani, B. (2025). Machine Learning Analysis of Hydrological and Hydrochemical Data from the Abelar Pilot Basin in Abegondo (Coruña, Spain). Preprints. Available at: https://www.preprints.org/frontend/manuscript/3b9ac3ed0e90db70d67f30f7338655c6/download_pub

14. Deng, Y., Liu, Y., Zhang, D., & Cao, Z. (2025). A Hybrid Gradient Boosting Model for Predicting Longitudinal Dispersion Coefficient in Natural Rivers. Water Resources Management. Available at: https://link.springer.com/article/10.1007/s11269-024-04058-6

15. Li, F., Springer, A., Kusche, J., & Gutknecht, B. D. (2025). Reanalysis and forecasting of total water storage and hydrological states by combining machine learning with CLM model simulations and GRACE data assimilation. Water Resources Research. Available at: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2024WR037926

16. Behzadfar, M., Mirzakhani, A., & Azizi Habashi, S. (2025). Simulating urban expansion dynamics in Tehran through satellite imagery and cellular automata Markov chain modelling. Modeling Earth Systems and Environment. Available at: https://link.springer.com/article/10.1007/s40808-025-02325-y

17. Ganguli, P., Han, S., Muñoz, D. F., & Pal, S. (2025). Spatiotemporal Modelling and Assessment of Water-Related Multihazards. Frontiers in Water. Available at: https://www.frontiersin.org/journals/water/articles/10.3389/frwa.2025.1579106/full

18. Escalante, J. J., Chen, W. H., Daud, W. M. A. W., Su, C. Y., & Li, P. H. (2025). Prediction construction for biomass and high-density polyethylene co-gasification via statistical method and machine learning. Fuel. Available at: https://www.sciencedirect.com/science/article/pii/S0016236125005526

19. Zhang, D., Liu, Y., Cao, Z., & Deng, Y. (2025). A Hybrid Gradient Boosting Model for Predicting Longitudinal Dispersion Coefficient in Natural Rivers. Water Resources Management. Available at: https://link.springer.com/article/10.1007/s11269-024-04058-6

20. Şentop, M. S. (2025). Orthogonality based feature selection for AI applications. ITU Polen Repository. Available at: https://polen.itu.edu.tr/items/39dc738d-0c0b-4516-beef-a589a8063e81

21.Комар М.П., Саченко А.О., Васильків Н.М., Гладій Г.М., Коваль В.С., Ліп’яніна-Гончаренко Х.В. Методичні рекомендації до виконання кваліфікаційної роботи з освітньо-професійної програми «Комп’ютерні науки» спеціальності 122 «Комп’ютерні науки» за першим (бакалаврським) рівнем вищої освіти. Тернопіль: ЗУНУ, 2024. 52 с.

22.Загальні методичні рекомендації з підготовки, оформлення, захисту та оцінювання кваліфікаційних робіт здобувачів вищої освіти першого (бакалаврського) і другого (магістерського) рівнів. Тернопіль: ЗУНУ, 2024. 83 с.

23. Дегодь, В., Домбровський, М. Програмний модуль прогнозування якості води на основі ансамблевого підходу. Інтелектуальні інформаційні технології в прикладних дослідженнях: збірник тез доп. студент. наук.-практ. конф. (IITAR-2025), м. Тернопіль, 27-29 травня 2025 р. Тернопіль: ЗУНУ, 2025. С. 40-42.

Додаток А

Код реалізації

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

df= pd.read_csv("../input/water-potability/water_potability.csv")
df.head()
df.info()
df['Potability']=df['Potability'].astype('category')
cols=df.columns[0:9].to_list()
min_val=[6.52,0,500,0,3,0,0,0,0]
max_val=[6.83,0,1000,4,250,400,2,80,5]
limit=pd.DataFrame(data=[min_val, max_val], columns=cols)
df.describe().T.style.background_gradient(subset=['mean','std','50%','count'], cmap='PuBu')
df[df['Potability']==1].describe().T.style.background_gradient(subset=['mean','std','50%','count'], cmap='PuBu')

df[df['Potability']==0].describe().T.style.background_gradient(subset=['mean','std','50%','count'], cmap='RdBu')

df.isnull().sum()

df[df['Sulfate'].isnull()]
df[df['ph'].isnull()]
df[df['Trihalomethanes'].isnull()]

df['ph']=df['ph'].fillna(df.groupby(['Potability'])['ph'].transform('mean'))
df['Sulfate']=df['Sulfate'].fillna(df.groupby(['Potability'])['Sulfate'].transform('mean'))
df['Trihalomethanes']=df['Trihalomethanes'].fillna(df.groupby(['Potability'])['Trihalomethanes'].transform('mean'))

df.isna().sum()

import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import plotly.graph_objects as go
from plotly.subplots import make_subplots
colors = ['#06344d','#00b2ff']
sns.set(palette=colors, font='Serif', style='white', rc={'axes.facecolor':'#f1f1f1', 'figure.facecolor':'#f1f1f1'})
sns.palplot(colors)


fig = plt.figure(figsize=(10,6))
ax=sns.countplot(data=df, x='Potability')
for i in ax.patches:
  ax.text(x=i.get_x()+i.get_width()/2,  y=i.get_height()/7,  s=f"{np.round(i.get_height()/len(df)*100,0)}%",  ha='center',
size=50, weight='bold', rotation=90, color='white')
plt.title("Potability Feature", size=20, weight='bold')
plt.annotate(text="Not safe for Human consumption", xytext=(0.5,1750),xy=(0.2,1250), arrowprops =dict(arrowstyle="-
>", color='black', connectionstyle="angle3,angleA=0,angleB=90"), color='black')
plt.annotate(text="Safe for Human consumption", xytext=(0.8,1500),xy=(1.2,1000), arrowprops =dict(arrowstyle="->",
color='black', connectionstyle="angle3,angleA=0,angleB=90"), color='black')

limit

from matplotlib.patches import Rectangle
int_cols = df.select_dtypes(exclude=['category']).columns.to_list()
fig, ax= plt.subplots(nrows=3,ncols=3,figsize=(15,15), constrained_layout=True)
plt.suptitle('Feature distribution by Potability class and Approved limit', size=20, weight='bold')
ax=ax.flatten()
for x, i in enumerate(int_cols):
  sns.kdeplot(data=df, x=i, hue='Potability', ax=ax[x], fill=True, multiple='stack', alpha=0.5, linewidth=2)
  l,k = limit.iloc[:,x]
  ax[x].add_patch(Rectangle(xy=(l,0), width=k-l, height=1, alpha=0.5))
  for s in ['left','right','top','bottom']:
    ax[x].spines[s].set_visible(False)

    fig, ax= plt.subplots(nrows=3,ncols=3,figsize=(15,15), constrained_layout=True)
plt.suptitle('Feature distribution by Potability class and Approved limit', size=20, weight='bold')
```

```
ax=ax.flatten()
for x, i in enumerate(int_cols):
  sns.boxplot(data=df, y=i, x='Potability', ax=ax[x])
  #l,k = limit.iloc[:,x]
  #ax[x].add_patch(Rectangle(xy=(l,0), width=k-l, height=1, alpha=0.5))
  for s in ['left','right','top','bottom']:
    ax[x].spines[s].set_visible(False)

from scipy.stats import ttest_ind
p_val=[]
for i in int_cols:
  pota_1 = df[df['Potability']==1][i]
  pota_0 = df[df['Potability']==0][i]
  stat, p_value=ttest_ind(pota_1, pota_0)
  p_val.append(np.round(p_value,3))
  if p_value <0.1:
    print(f"p_value for {i} is {p_value} is less than significant value 0.1, so we have no enough evidance to prove Null
Hypothesis. so we reject the Null Hypotesis")
  else:
    print(f"p_value for {i} is {p_value} we accept the null hypothesis")

stats_test=pd.DataFrame(columns=['columns','p_value'])
stats_test['columns']=int_cols
stats_test['p_value']=p_val
stats_test.sort_values(by=['p_value'], ascending=True, inplace=True)

fig=plt.figure(figsize=(15,8))
ax=sns.barplot(data=stats_test, x='columns',y='p_value')
plt.title("Features and p_value based on t-Test", size=20, weight='bold')
for i in ax.patches:
  ax.text(x=i.get_x()+0.5, y=i.get_height(), s=i.get_height(), rotation=90)
ax.axhline(y=0.1, color='red', ls='--')
ax.text(x=6, y=0.12, s="Significance level: 0.1")
sns.pairplot(df, hue='Potability', kind='reg')

fig, ax=plt.subplots(nrows=1, ncols=2, figsize=(15,8))
plt.suptitle("Co-relation Matrics", size=20, weight='bold')
ax=ax.flatten()
sns.heatmap(df[df['Potability']==1].corr(), annot=True, square=True, fmt='.2f', ax=ax[0], cbar=False)
sns.heatmap(df[df['Potability']==0].corr(), annot=True, square=True, fmt='.2f', ax=ax[1], cbar=False)

fig=plt.figure(figsize=(8,8))
sns.heatmap(df.corr(), annot=True, fmt='0.2f', square=True)

from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
X = df.drop(['Potability'], axis=1)
y=df['Potability']

scale = StandardScaler()
X_scaled = scale.fit_transform(X)
decom = PCA(svd_solver='auto') #let try with auto rather than defining the components
decom.fit(X_scaled)
ex_var=np.cumsum(np.round(decom.explained_variance_ratio_,2))*100
sns.lineplot(y=ex_var, x=np.arange(0,len(ex_var)))
sns.scatterplot(y=ex_var, x=np.arange(0,len(ex_var)))

from imblearn.over_sampling import SMOTE
samp = SMOTE()
X=df.drop(['Potability'], axis=1)
y=df['Potability']
from sklearn.model_selection import train_test_split, cross_val_score, GridSearchCV
X_train,X_test, y_train, y_test = train_test_split(X, y, random_state=42)
X_train, y_train =samp.fit_resample(X_train,y_train)

scale = StandardScaler()
X_train=scale.fit_transform(X_train)
X_test=scale.transform(X_test)

from yellowbrick.classifier import ROCAUC
```

```
from sklearn.ensemble import AdaBoostClassifier, BaggingClassifier, GradientBoostingClassifier
from sklearn.tree import DecisionTreeClassifier, ExtraTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
mod = []
cv_score=[]
model =[AdaBoostClassifier(), BaggingClassifier(), GradientBoostingClassifier(), DecisionTreeClassifier(),
ExtraTreeClassifier(), KNeighborsClassifier()]
for m in model:
  cv_score.append(cross_val_score(m, X_train, y_train, scoring='accuracy', cv=5).mean())
  mod.append(m)
model_df=pd.DataFrame(columns=['model','cv_score'])
model_df['model']=mod
model_df['cv_score']=cv_score
model_df.sort_values(by=['cv_score'], ascending=True).style.background_gradient(subset=['cv_score'])

param={'n_estimators': [60,70,80,100,200,300,400,500,600,700]}
grid_Grd=GridSearchCV(GradientBoostingClassifier(), param_grid=param, cv=5, scoring='accuracy')
grid_Grd.fit(X_train, y_train)
print(f"Best Estimator: {grid_Grd.best_params_} , Best Score : {grid_Grd.best_score_}")


param={'n_estimators': [60,70,80,100,200,300,400,500,600,700]}
grid_Bag=GridSearchCV(BaggingClassifier(), param_grid=param, cv=5, scoring='accuracy')
grid_Bag.fit(X_train, y_train)
print(f"Best Estimator: {grid_Bag.best_params_} , Best Score : {grid_Bag.best_score_}")

from sklearn.metrics import classification_report, confusion_matrix
model = GradientBoostingClassifier(n_estimators=300)
model.fit(X_train,y_train)
pred = model.predict(X_test)
print(classification_report(y_test, pred))
sns.heatmap(confusion_matrix(y_test, pred), annot=True, fmt='.2f')

from sklearn.metrics import classification_report, confusion_matrix
model = BaggingClassifier(n_estimators=80)
model.fit(X_train,y_train)
pred = model.predict(X_test)
print(classification_report(y_test, pred))
sns.heatmap(confusion_matrix(y_test, pred), annot=True, fmt='.2f')

model.predict_proba(X_test)

import h2o
from h2o.automl import H2OAutoML
h2o.init(max_mem_size='2G')

h2o_df = h2o.H2OFrame(df)
h2o_df['Potability']=h2o_df['Potability'].asfactor()
X=h2o_df.columns[0:-1]
y=h2o_df.columns[-1]

train, test=h2o_df.split_frame(ratios=[.7])
print(train.nrows)
print(test.nrows)

aml = H2OAutoML(balance_classes=True)
aml.train(x=X, y=y, training_frame=train)

aml.leaderboard

type(test['Potability'])

pred = aml.leader.predict(test)
y_val = h2o.as_list(test['Potability'], use_pandas=True)
pred_val = h2o.as_list(pred['predict'], use_pandas=True)
print(classification_report(y_val,pred_val))
```

Додаток Б

Апробація результатів

Західноукраїнський національний університет

Факультет комп'ютерних інформаційних технологій

Кафедра інформаційно-обчислювальних систем і управління

ЗБІРНИК ТЕЗ ДОПОВІДЕЙ

Студентської науково-практичної конференції

ІНТЕЛЕКТУАЛЬНІ ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ В ПРИКЛАДНИХ ДОСЛІДЖЕННЯХ

(IITAR-2025)

27-29 травня 2025 року

Тернопіль

2025

Зміст

Дегодь Віталій
студент групи КН-42
degod123.a@gmail.com
Домбровський Михайло
к.т.н., старший викладач
m.dombrovskyi@wunu.edu.ua
Західноукраїнський національний університет
Тернопіль, Україна

ПРОГРАМНИЙ МОДУЛЬ ПРОГНОЗУВАННЯ ЯКОСТІ ВОДИ НА ОСНОВІ АНСАМБЛЕВОГО ПІДХОДУ

Забезпечення населення якісною питною водою залишається однією з найважливіших глобальних екологічних та медичних проблем. За даними Всесвітньої організації охорони здоров'я, щороку мільйони людей хворіють та помирають внаслідок вживання забрудненої питної води, особливо у країнах, що розвиваються. У зв’язку з цим гостро постає завдання оперативного й ефективного моніторингу якості води з використанням сучасних інформаційних технологій та штучного інтелекту.

В рамках дослідження було запропоновано інтелектуальний програмний модуль, що надає змогу автоматизовано прогнозувати придатність води для споживання на основі фізико-хімічних параметрів. Для реалізації модуля використано набір даних, що включає 3276 зразків з інформацією про 9 ключових показників: pH, загальна твердість, електропровідність, органічний вуглець, хлориди, сульфати, тверді речовини, тригалометани та каламутність. Попередньо проведений аналіз показав неоднорідність та складну структуру взаємозв’язків між ознаками, що вимагає застосування потужних алгоритмів машинного навчання.

З метою покращення точності прогнозування було порівняно кілька популярних ансамблевих методів, зокрема Random Forest, Gradient Boosting, Bagging та AdaBoost. Результати моделювання показали, що ансамблеві підходи забезпечують значне покращення точності класифікації порівняно з окремими класичними алгоритмами, такими як логістична регресія або метод k найближчих сусідів. Вищі показники точності були досягнуті за допомогою автоматизованого ансамблевого методу H2O AutoML (StackedEnsemble), який забезпечив точність класифікації на рівні 80%, з показником AUC-ROC у 0,864, що є високим результатом для такого типу задач.

Особливістю розробленого модуля є інтеграція автоматизованого машинного навчання (AutoML) із засобами візуалізації, що надає змогу наочно оцінювати результати прогнозування, аналізувати найбільш важливі фактори, що впливають на якість води, та вчасно ідентифікувати потенційні ризики для здоров’я населення.

Модель здатна працювати з реальними даними, забезпечуючи стійкість до викидів та неповних даних, що підвищує її практичну придатність у польових та промислових умовах.

На рисунку 1 наведено архітектуру програмного модуля, що включає рівні збору та обробки даних, ансамблеве прогнозування, оцінку результатів та візуалізацію.

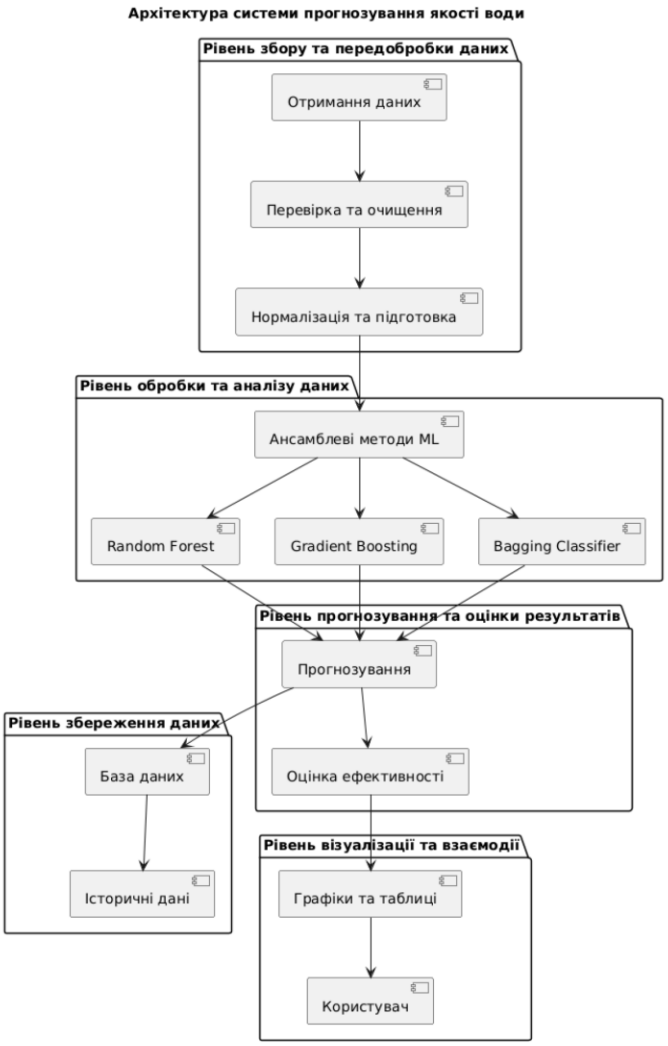


Рисунок 1 – Архітектура системи прогнозування якості води

Порівняльний аналіз моделей (табл. 1) демонструє перевагу StackedEnsemble (H2O AutoML) над традиційними ансамблевими методами, такими як Gradient Boosting та Bagging.

Таблиця 1 – Порівняння моделей прогнозування якості води

№	Модель	Точність класифікації	Прецизійність	Повнота	F1-міра	AUC-ROC
1	StackedEnsemble (H2O AutoML)	0,80	0,79	0,66	0,72	0,864
2	Gradient Boosting	0,78	0,67	0,80	0,73	0,855
3	Bagging Classifier	0,76	0,68	0,69	0,68	0,850

Запропонований модуль може бути інтегрований у системи моніторингу якості води, що надає змогу екологічним, медичним та комунальним службам оперативно аналізувати параметри води, знижувати витрати на лабораторні дослідження та приймати швидкі рішення щодо запобігання екологічним та медичним загрозам.

Список використаних джерел

1. Subudhi, S., Pati, A. K., Bose, S., & Sahoo, S. (2025). Integrating Boosted learning with Differential Evolution (DE) Optimizer: A Prediction of Groundwater Quality Risk Assessment in Odisha. arXiv preprint.
2. Feng, Y., Zhang, J., Guo, S., Zhang, Y., & Zhang, Z. (2025). High precision water quality retrieval in Dianchi Lake using satellite data and machine learning algorithms. Water Research, 226, 119309.
3. Zhang, T., Yang, Y., & Huang, Q. (2025). A hybrid model combining deep learning and decision trees for water quality prediction. Environmental Modelling & Software, 181, 105251.
4. Lu, H., Li, X., & Chen, W. (2025). Explainable machine learning models for precipitation forecasting to enhance water quality monitoring. Journal of Hydrology, 601, 127285.
5. Wang, L., Xu, C., & Li, J. (2025). Interpretable AI for river water quality assessment using remote sensing data. Remote Sensing of Environment, 308, 113450.