УДК 004.9+574

М. М. Кадирова
Є. М. Крижановський

СИСТЕМНИЙ АНАЛІЗ ТА ПРОГНОЗУВАННЯ ВИТРАТ ВОДИ ЗА ІНФОРМАЦІЄЮ З БАЗИ ДАНИХ WISE

Вінницький національний технічний університет

Анотація

У роботі розглянуто задачу системного аналізу та прогнозування витрат поверхневих водних ресурсів Європи на основі даних бази моніторингу WISE (Water Information System for Europe). Виконано розвідувальний аналіз масиву з понад 5,7 мільйона спостережень, розроблено конвеєр підготовки даних, побудовано та порівняно дві прогностичні моделі — LightGBM та глибоку нейронну мережу. За результатами порівняння модель LightGBM продемонструвала вищу точність (R^2=0,7469, MAE=2,5464) та рекомендована як основна для задачі прогнозування гідрологічних показників.

Ключові слова: машинне навчання, LightGBM, прогнозування гідрологічних показників, WISE, градієнтний бустинг, нейронна мережа, розвідувальний аналіз даних.

Abstract

This work addresses the problem of systematic analysis and forecasting of surface water flow in Europe using data from the WISE (Water Information System for Europe) monitoring database. Exploratory data analysis was performed on a dataset of over 5.7 million observations, a data preprocessing pipeline was developed, and two predictive models were built and compared — LightGBM and a deep neural network. The LightGBM model demonstrated superior accuracy (R^2=0.7469, MAE=2.5464) and is recommended as the primary model for hydrological indicator forecasting.

Keywords: machine learning, LightGBM, hydrological forecasting, WISE, gradient boosting, neural network, exploratory data analysis.

Вступ

Управління водними ресурсами набуває критичного значення в умовах глобальних кліматичних змін та зростання антропогенного навантаження. Забезпечення надійного моніторингу та прогнозування гідрологічних показників поверхневих водних об'єктів є актуальним завданням як для державних екологічних агенцій, так і для аграрного та промислового секторів. Водночас традиційні методи просторової інтерполяції та лінійної регресії є недостатньо ефективними для обробки великих гетерогенних масивів екологічних даних через нездатність фіксувати складні нелінійні просторово-часові залежності.

Метою даної роботи є розробка та порівняльна оцінка інтелектуальних моделей прогнозування витрат поверхневих вод на основі загальноєвропейської бази моніторингу WISE без використання авторегресійних ознак, що забезпечує можливість прогнозування на станціях без тривалої історії спостережень.

Аналіз та підготовка даних

Дослідження виконано на основі набору даних Water Quantity бази WISE3, що налічує понад 5,7 мільйона записів гідрологічних спостережень за 2005–2023 роки по 12 атрибутах: географічні координати, часові мітки, коди країн, ідентифікатори станцій та виміряні значення цільового показника витрати води (SF). Реалізацію виконано мовою Python на платформі Kaggle з використанням бібліотек DuckDB, Pandas, LightGBM, TensorFlow/Keras, Optuna та Scikit-learn.

Розвідувальний аналіз даних виявив значну географічну нерівномірність вибірки: Словенія, Франція та Італія разом формують близько половини всього обсягу даних (рис. 1). Встановлено чітку сезонну циклічність показника SF з максимумами у зимово-весняний період (~7,2 м³/с) та мінімумом у серпні (~2,65 м³/с), що корелює з фазами сніготанення (рис. 2).

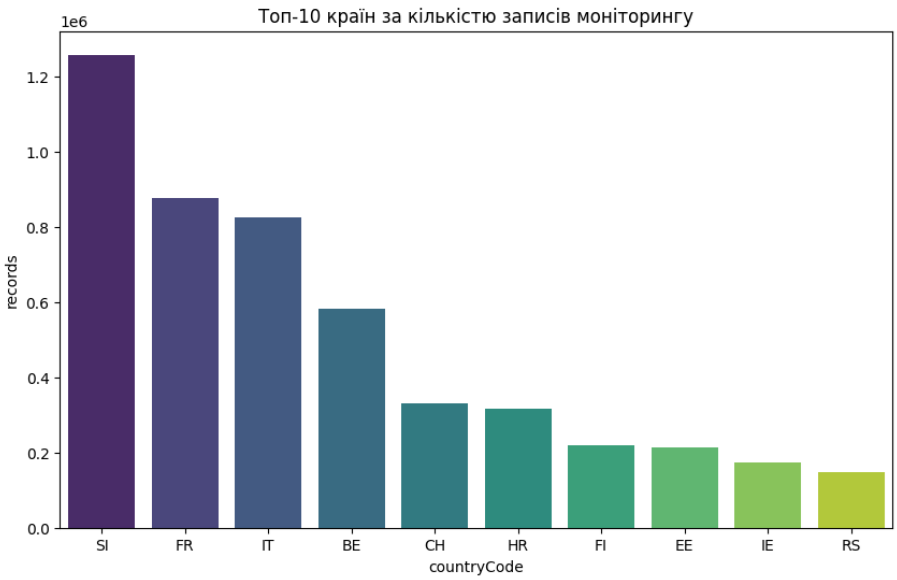

Рис. 1. Топ-10 країн за кількістю записів моніторингу WISE3

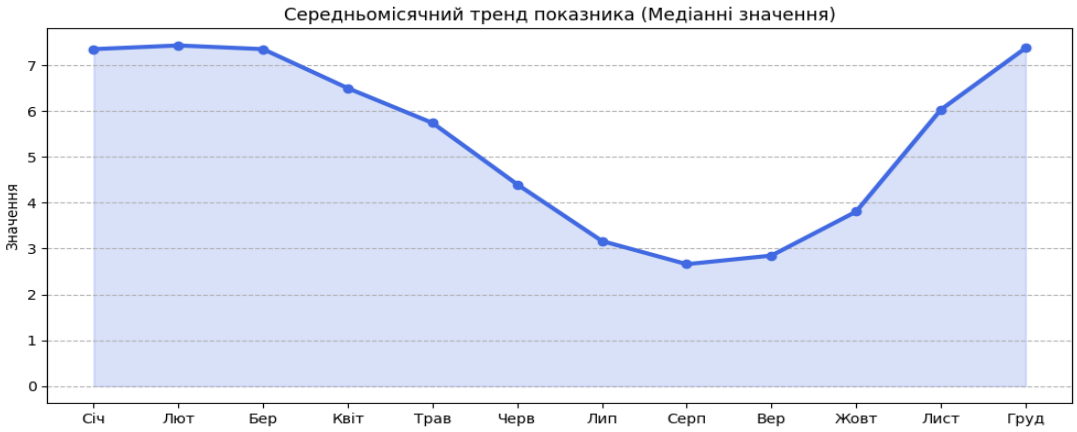

Рис. 2. Середньомісячний тренд показника SF

Розроблений конвеєр підготовки даних включає: видалення викидів методом міжквартильного розмаху (усунуто ~13% записів); тригонометричне кодування місяців (sin_month, cos_month) для коректної параметризації річної циклічності; введення лагової ознаки попереднього спостереження; міткове кодування категоріальних змінних; глобальну стандартизацію числових ознак. Логарифмічне перетворення цільової змінної стабілізувало дисперсію та підвищило стійкість моделей до екстремальних значень.

Результати дослідження

Для прогнозування витрат води розроблено та порівняно два підходи: градієнтний бустинг на основі LightGBM та глибоку нейронну мережу (TensorFlow/Keras). Модель LightGBM навчалась із байєсівською оптимізацією гіперпараметрів (Optuna) та ранньою зупинкою за MAE. Нейронна мережа з автоматичним підбором архітектури включала 2–4 приховані шари з активацією Swish, Batch Normalization та функцією втрат Huber Loss.

Аналіз важливості ознак методом Gain/Split підтвердив домінуючий вплив географічних координат (lat, lon) на прогноз (рис. 3), що математично обґрунтовує просторову детермінованість гідрологічного

режиму. Рік спостереження (year) посідає третє місце, відображаючи довготермінові тренди, зумовлені змінами клімату та антропогенним навантаженням.

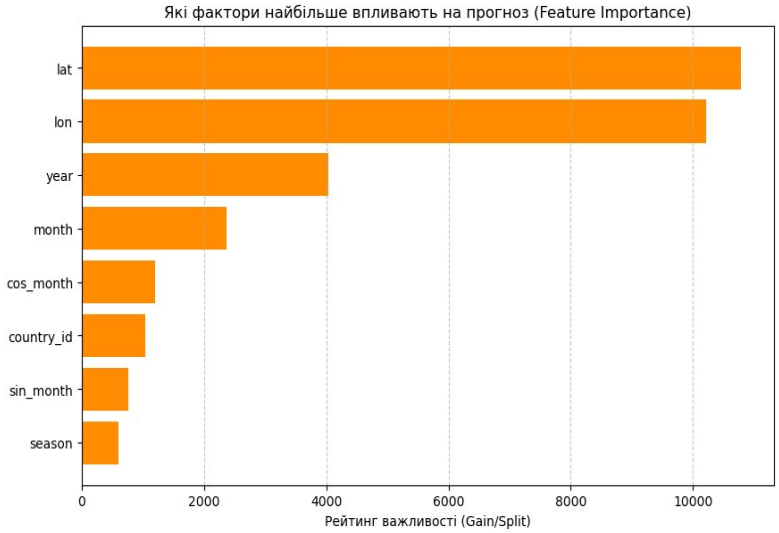

Рис. 3. Важливість ознак моделі LightGBM

За результатами порівняльної оцінки на незалежній тестовій вибірці (15% від загального обсягу) модель LightGBM значно перевершила нейронну мережу за обома метриками: R^2=0,7469 проти 0,5477 та MAE=2,5464 проти 3,3720 (рис. 4). Перевага бустингу пояснюється відносно невеликою кількістю структурованих табличних ознак, для яких ансамблеві алгоритми традиційно перевершують нейронні мережі.

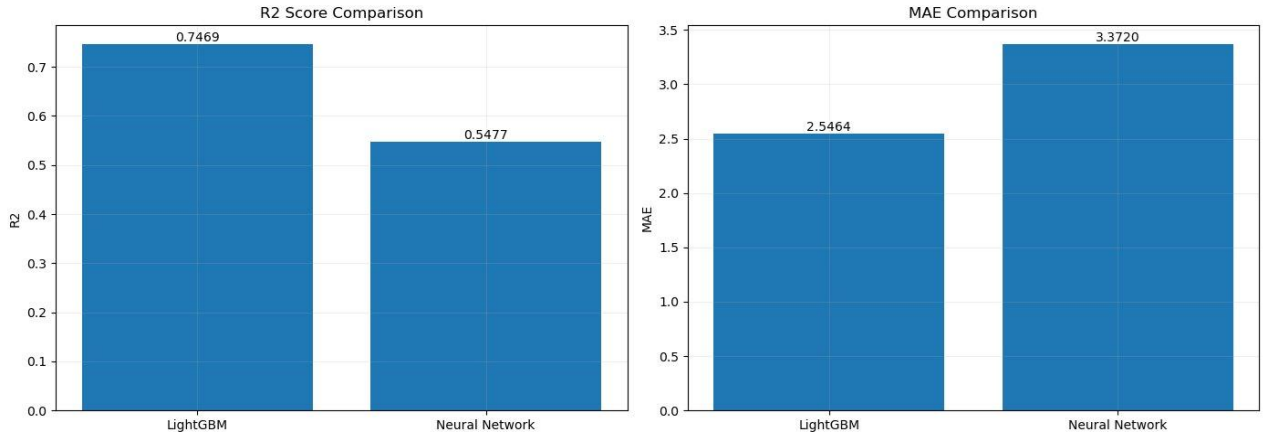

Рис. 4. Порівняння моделей LightGBM та нейронної мережі за метриками R^2 та MAE

Висновки

За результатами виконаного дослідження розроблено інтелектуальну систему прогнозування витрат поверхневих вод Європи на основі даних WISE. Встановлено, що модель градієнтного бустингу LightGBM є більш ефективною за нейронну мережу для табличних просторово-часових екологічних даних із відносно невеликою кількістю структурованих ознак. Виявлено домінуючий вплив географічних координат на прогноз витрат води, що підтверджує просторову природу гідрологічних процесів. Розроблена система може бути використана для підтримки управлінських рішень у сфері раціонального розподілу водних ресурсів та раннього виявлення екологічних загроз.

Перспективними напрямами подальшого дослідження є: залучення додаткових предикторів (дані про землекористування, індекси посушливості, супутникові показники вологості ґрунту); побудова

авторегресійної системи прогнозування; дослідження графових нейронних мереж (GNN) для явного моделювання топології річкових басейнів; впровадження методів інтерпретації SHAP; розробка веб-орієнтованого інтерфейсу для інтерактивної картографічної візуалізації прогнозів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Юань Т., Чжан Ю., Лю Ф. Machine learning-driven advances in water treatment and ecological environmental monitoring. Environmental Research. 2025. Вип. 245. С. 117–134.

2. Гарсія М., Лопес Дж., Фернандес К. Bibliometric-Systematic Literature Review (B-SLR) of Machine Learning-Based Water Quality Prediction: Trends, Gaps, and Future Directions. MDPI Water. 2025. Вип. 17. 24 с.

3. Мокін В. Б., Дратований М. В. Наука про дані: машинне навчання та інтелектуальний аналіз даних. Навч. посібник. Вінниця : ВНТУ, 2024. С. 119–128.

4. Сібінді Р., Мвангі Р. В., Вайтіту А. Г. A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices. Engineering Reports. 2023. Т. 5, № 4. e12599.

5. Water Quantity, 2025, The Water Information System for Europe. [Електронний ресурс]. – Режим доступу: https://sdi.eea.europa.eu/catalogue/datahub/api/records/3f6e7d37-eb3b-407a-b2ed-3eb390668e36

Крижановський Євгеній Миколайович – к.т.н, доцент кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, Вінниця, e-mail: kruzhan@gmail.com.

Кадирова Марія Михайлівна – студентка групи СА-22б, Факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: kadmar1508@gmail.com

Kryzhanovsky Evgeniy M. – Cand. Sc. (Eng.), Associate Professor of the Department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: kruzhan@gmail.com

Kadyrova Mariia M. – student of SA-22b group, Faculty of Intelligent Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia,
e-mail: kadmar1508@gmail.com