ISSN 2707-1898 (print)

Український журнал інформаційних технологій
Ukrainian Journal of Information Technology
http://science.lpnu.ua/uk/ujit
https://doi.org/10.23939/ujit2024.01.026
Article received 26.04.2024 p.
Article accepted 30.04.2024 p.
UDK 004.8

Correspondence author
D. V. Fedasyuk
dmytro.v.fedasyuk@lpnu.ua

УДК 004.8

Д. В. Федасюк, М. О. Костюк
Національний університет "Львівська політехніка", Львів, Україна

ПРОГНОЗУВАННЯ ВОЛОГОСТІ ҐРУНТУ З ВИКОРИСТАННЯМ МАШИННОГО НАВЧАННЯ У СИСТЕМАХ РОЗУМНОГО ЗЕМЛЕРОБСТВА

Вирощування сільськогосподарських культур у сучасних умовах є комплексним завданням і практично поєднує у собі практики досвіду та новітні методи, зокрема інформаційні технології, що охоплює поняття "розумне землеробство". Важливим чинником стабільної прогнозованої врожайності є рівень вологості ґрунтів, який є результатом змін таких кліматичних чинників, як температура повітря, кількість опадів, вітряність тощо. Запропоновано методику опрацювання реальних історичних показників змін клімату певної географічної ділянки з подальшим тренуванням та застосуванням моделей машинного навчання для прогнозування вологості ґрунтів. Для побудови моделі машинного навчання вибрано і досліджено алгоритми: алгоритм регресійних дерев, випадкового лісу, лінійної регресії, алгоритми M5P та алгоритм K*. Розроблено структуру кліматичних даних для навчання моделі із метою подальшого прогнозування вологості ґрунтів з урахуванням температури і вологості повітря, температури і вологості ґрунту, кількості опадів, кількості прямої та розсіяної сонячної радіації, швидкості вітру. Джерела інформації вибрано із відкритих розподілених світових ресурсів. Розроблено архітектуру та створено програмну систему прогнозування вологості ґрунтів на основі алгоритмів машинного навчання із застосуванням фреймворку Spring Framework, бібліотеки WEKA та Java FX з можливістю вибирати та досліджувати вибрані алгоритми. Виконано експерименти та наведено результати тривалості навчання моделей. Найменше часу навчання потребують алгоритми регресійних дерев та лінійної регресії. Здійснено порівняння алгоритмів за критеріями: швидкість навчання, швидкість перехресного тестування, швидкість прогнозування, показники ефективності тестування для реальних історичних даних. Отримані результати дадуть змогу оцінити та вибрати найкращі моделі машинного навчання для проєктування інформаційно-аналітичної системи "розумне землеробство" для прогнозування вологості ґрунтів.

Ключові слова: алгоритм M5P, лінійна регресія, K*, алгоритми регресійних дерев, випадкових лісів, передбачення.

Вступ / Introduction

Поточний етап розвитку землеробства передбачає упровадження інформаційних технологій, таких як штучний інтелект, машинне навчання, хмарні обчислення, інтернет речей, системи передавання даних на великі відстані, робототехніку тощо. Можна стверджувати, що нині відбувається цифрова трансформація у сільському господарстві, яке використовує передові технології для підвищення продуктивності та ефективності, оскільки саме розвиток сільського господарства, і землеробства, зокрема, може забезпечити продовольчу стабільність планети.

Одним із найважливіших факторів, що впливає на врожайність, є стан вологості ґрунтів. Останній відіграє важливу роль у вирощуванні рослин, водяному циклі та екосистемах. Можна вважати, що напрям досліджень із прогнозування вологості ґрунту, разом зі збиранням та аналізом даних з інших систем та сенсорів, активно розвивається та є невід'ємною частиною сучасних систем розумного землеробства – рушійною силою сучасного етапу розвитку сільського господарства.

Система прогнозування вологості ґрунту за допомогою машинного навчання дає змогу оперативно реагувати на прогнозовану зміну кліматичних умов та, у випадку інтеграції із системами розумного землеробства, сприятиме оптимізації використання водних ресурсів, електроенергії, палива та добрив разом зі створенням сприятливих умов для зростання рослин.

Тому актуальним завданням є дослідження, спрямовані на прогнозування вологості ґрунтів як один із головних факторів забезпечення контрольованої врожайності, за допомогою алгоритмів машинного навчання.

Об'єкт дослідження – процеси моніторингу, збирання даних та використання машинного навчання для прогнозування вологості ґрунту.

Предмет дослідження – моделі та алгоритми машинного навчання для моніторингу та прогнозування показників вологості ґрунту.

Мета роботи – дослідити ефективність та придатність алгоритмів машинного навчання для моніторингу та прогнозування вологості ґрунтів, особливостей і структури кліматичних даних для їх навчання та рекомендацій для практичного використання.

Для досягнення зазначеної мети визначено такі основні *завдання дослідження*:

- розробити принципи системи моніторингу, яка ґрунтується на використанні відкритих великих кліматичних даних;

- збирання вхідних попередніх кліматичних даних для навчання системи та визначення алгоритму їх оброблення;
- вибір, навчання та тестування алгоритмів машинного навчання;
- розроблення користувацького інтерфейсу програмного застосунку, що дасть змогу гнучко порівнювати результати роботи алгоритмів та передбачати значення вологості ґрунту;
- опрацювання результатів та формування висновків щодо ефективності застосовуваних алгоритмів машинного навчання.

Матеріали і методи дослідження. Під час дослідження застосовано такі методи: методи машинного навчання – алгоритм регресійних дерев, алгоритм випадкового лісу, лінійної регресії, алгоритм M5P та алгоритм. Метод об'єктно-орієнтованого програмування використано для розроблення програмного застосунку. Програмний застосунок створено з використанням фреймворку Spring Framework, бібліотеки WEKA на мові програмування Java. Для навчання системи вибрано відкриті інформаційні ресурси, які містять кліматичні дані із метеорологічних станцій Орадея (Румунія), та сервіс OpenMeteo, який надає відкрите API для некомерційного використання з доступом до архівних даних кліматичних спостережень. Графіки побудовано із використанням програмного середовища Microsoft Excel та Google Sheets.

Аналіз останніх досліджень і публікацій. Сучасне землеробство 4.0 – це концепція цифрової трансформації в сільському господарстві, яка використовує передові технології, такі як штучний інтелект, хмарні обчислення, інтернет речей, робототехніку та давачі для підвищення продуктивності та ефективності. Ця концепція передбачає використання даних та аналітики для прийняття рішень, автоматизації процесів та покращення якості продукції [1].

Огляд можливостей використання штучного інтелекту, машинного навчання та інтернету речей в сільському господарстві подано у дослідженні [2]. Він містить загальну інформацію про використання цих технологій, порівняння різних технологій штучного інтелекту, використовуваних для контролю стану ґрунтів, на прикладі IBM's AgroPad, Management-Oriented Modeling (MOM), FL: SoilRisk Characterization-Decision Support System (SRC-DSS), моделі ANN, Fuzzy Inference System (FIS), Support Vector Machine (SVM), Plantix та техніки аналізу придатності угідь.

У роботі [3] запропоновано нову архітектуру на основі повністю зв'язаної прямопотокової моделі штучної нейронної мережі (ANN) для оцінювання вологості ґрунту. Для побудови архітектури використано супутникові зображення Sentinel-1, Sentinel-2 та Shuttle Radar Topographic Mission території річки Косі, що в Гімалаях. Для вимірювання вологості ґрунту використано калібровані зонди TDR, розміщені в 224 місцях. Здійснено порівняння ефективності моделі ANN із десятьма іншими алгоритмами машинного навчання (GRNN, RBN, Exact RBN, GPR, SVR, RF, Boosting EL, RNN, BDT та AutoML).

Стаття [4] спрямована на покращення передбачення вологості ґрунту за допомогою нової моделі кодування-декодування (encoder-decoder) з використанням навчання зі зміщенням. Дослідження пропонує перспективний інструмент для управління екосистемами та точного землеробства. У статті автори описують методику побудови моделі передбачення вологості ґрунту, що передбачає використання залишкового навчання (residual learning). Модель ґрунтується на архітектурі з використанням навчання зі зміщенням, що дає змогу зменшити кількість параметрів моделі та підвищити її ефективність. Залишкове навчання використовують для зменшення впливу шуму та забезпечення стабільності моделі. Автори протестували EDT-LSTM для прогнозування вологості ґрунту на різних часових інтервалах, використовуючи дані зі станцій FLUXNET. Результати експериментів показали, що запропонована модель забезпечує високу точність передбачення вологості ґрунту порівняно з іншими методами.

У статті [5] досліджено прогнозування вологості ґрунту в провінції Цзянсу в Китаї. Для цього було зібрано дані з 70 метеорологічних та автоматичних спостережних станцій вологості ґрунту з 2014 до 2022 рр. на глибині 0–10 см (RH_{s10cm}) та оброблено за допомогою алгоритму екстремального градієнтного підсилення (XGBoost). Показано, що модель XGBoost виявилась придатною для прогнозування вологості ґрунту на локальному рівні певної території, оскільки вона відтворювала просторові характеристики розподілу різних рівнів посухи та ефективно передбачала динамічний процес зміни "виникнення – розвиток – завершення" конкретної події посухи.

У статті [6] досліджено взаємозв'язок між вологістю ґрунту, вологістю повітря, рівнем сонячної радіації та визначено, що зміни вологості ґрунту та вологості атмосфери відіграють важливу роль у зміні клімату. В дослідженні проаналізовано взаємозв'язки між їхніми річними середніми значеннями із використанням даних для перенавчання ERA5-Land.

Залежність між температурою ґрунту, вологістю повітря та температурою повітря наведено у дослідженні [7]. Автори роблять висновок, що температура повітря найбільше впливає на температуру ґрунту порівняно з іншими параметрами, однак температура повітря, вологість повітря та інтенсивність сонця також впливають на температуру ґрунту.

Використання технологій великої кількості даних досліджено в низці робіт. Зокрема, у статті [8] розглянуто поєднання використання інтернету речей, хмарних обчислень та BigData для отримання кращого контролю над вирощуванням сільськогосподарських культур.

Результати дослідження та їх обговорення / Research results and their discussion

Структура та агрегування кліматичних даних. Для тренування моделей необхідно мати якісний та достатній набір початкових даних, що впливає на прогнозовану величину. Для навчання системи та подальшого прогнозування необхідно вибрати дані, що безпосередньо впливають на прогнозоване значення вологості ґрунту. Для дослідження алгоритмів вибрано такі параметри: вологість ґрунту, $м^3/м^3$; температура повітря, ^oC; відносна вологість повітря, %; температура ґрунту, ^oC; кількість опадів, мм; інтенсивність прямої та розсіяної сонячної радіації, $Вт/м^2$; сила вітру, м/с.

Як джерело даних вибрано ресурс International SoilMoistureNetwork (ISMN) з вебсторінки ismn.earth/en

[9]. ISMN збирає виміряні дані вологості ґрунту на місцях (поверхневі та підповерхневі), узгоджені щодо одиниць вимірювання та частоти вибірки, для яких застосовують контроль якості, згодом дані розміщують у вільному доступі в мережі Інтернет.

Оскільки цей ресурс є об'єднанням різних метеорологічних станцій у різних країнах, є можливість вибрати найподібнішу за кліматичними умовами до України та з релевантним набором показників протягом тривалого періоду. Додатковим критерієм була стабільність вимірювань, оскільки багато станцій не виконують вимірювання регулярно або ж припинили їх, тому показники неможливо використати для прогнозування, оскільки вони не є релевантними та не відображають останні кліматичні зміни регіону. Тому з урахуванням цього вибрано станцію Орадея (Oradea) (21.89612, 47.0358), розміщену в Румунії на відстані 130 км від України. Використано щогодинні дані вимірювань із сенсорів вологості ґрунту ($м^3/м^3$), температури повітря (^oC) та кількості опадів (мм/год) за період з 2018-03-01 до 2023-08-31.

Окремо дані за 2020 р. використовували лише для тестування системи, а саме порівняння прогнозованих значень з актуальними за той час. Завдяки цьому система може перевірити точність передбачень та ймовірні похибки показників вибраних алгоритмів.

На основі показників з цього джерела сформовано такі дані: вологість ґрунту, $м^3/м^3$; температура ґрунту, ^oC; температура повітря, ^oC; кількість опадів, мм.

Додатково відбирали дані щодо решти показників. Як джерело таких даних вибрали сервіс OpenMeteo [10], який пропонує безкоштовне API, що дає змогу отримати історичні показники погоди для певних координат. Його використовуватимемо, щоб отримати дані, необхідні для навчання та прогнозування вологості ґрунту на основі історичних даних.

На основі показників цього джерела зібрано такі дані:

1. Вологість повітря, %.
2. Інтенсивність прямої та розсіяної сонячної радіації, $Вт/м^2$.
3. Сила вітру, м/с.

На виході отримуємо чотири різних файли у форматі csv для кожної з вимірюваних величин. Структуру заголовків наведено у табл. 1.

Для розрахунків брали два перших стовпці, які відповідають даті з кроком 1 год та безпосередньо вимірюваній величині. Стовпці з назвами *_flag та *_orig_flag мають технічний зміст та характеризують ймовірну якість вимірюваних величин. Дані, отримані з ISMN, зберігають локально для подальшого опрацювання та доповнення даними з OpenMeteo.

Ресурс OpenMeteo надає відкрите API для некомерційного використання, що дає можливість завантажувати необхідні дані для навчання та прогнозування погоди. Для того, щоб отримати решту даних, необхідно для кожної date_time з результатів вимірювань ISMN одержати історичну довідку з решти показників для певних координат. Для цього можна використати OpenMeteo WEB UI та завантажити дані для вибраного періоду часу в форматі CSV. Структуру вихідного файла наведено у табл. 2.

Дані, одержані від наведених сервісів, необхідно поєднати для кожного інтервалу часу та підготувати до подальшого використання – навчання алгоритмів та тестування історичних даних. Для цього опрацьовують рядки даних документа вологості ґрунту, в якому для кожного наявного значення дати у форматі *YYYY_MM_DD HH_mm_SS* вибирають записи з решти CSV документів кліматичних даних та об'єднують у кінцевий об'єкт, що містить всі необхідні характеристики на певну добу та годину часу в минулому. Схему агрегації кліматичних даних подано на рис. 1.

Навчання алгоритмів відбуватиметься на однакових наборах вхідних даних, що дасть змогу порівняти їх ефективність та точність прогнозування об'єктивніше. Для прогнозування вологості ґрунтів недостатньо знати поточні кліматичні показники, зібрані з метеорологічних джерел, оскільки вологість є результатом попередніх змін погоди. Тому в застосованому алгоритмі навчання і прогнозування враховуватимемо це, а саме кумулятивні показники клімату за минулу добу (рис. 2).

Всі експерименти дослідження виконано з використанням бібліотеки WEKA [11], яка є платформою з відкритим вихідним кодом для машинного навчання та інтелектуального аналізу даних. Бібліотека WEKA дає змогу використовувати CSV документи для навчання алгоритмів. Тому в нашій системі він слугуватиме кінцевим форматом підготовки даних.

На вологість ґрунту впливають погодні явища, які відбулися до замірів, наприклад, вчорашній дощ, тому в застосованому алгоритмі для навчання і прогнозування будуть враховані кумулятивні показники клімату за минулу добу (рис. 3).

Табл. 1. *Заголовки CSV даних, отриманих з ISMN / CSV headers of data received from ISMN*

Показник	Заголовки			
Вологість ґрунту	data_time	soil_moisture	soil_moisture_flag	soil_moisture_orig_flag
Температура повітря	data_time	air_temperature	air_temperature_flag	air_temperature_orig_flag
Температура ґрунту	data_time	soil_temperature	soil_temperature_flag	soil_temperature_orig_flag
Кількість опадів	data_time	precipitation	precipitation_flag	precipitation_orig_flag

Табл. 2. *Заголовки CSV даних, отриманих з OpenMeteo / CSV headers of data received from OpenMeteo*

Заголовки	Опис
time	Час, год
temperature_2 m, °C	Температура повітря
relative_humidity_2 m, %	Відносна вологість повітря
wind_speed_10 m, m/s	Швидкість вітру
diffuse_radiation, W/m²	Розсіяна сонячна радіація
direct_radiation, W/m²	Пряма сонячна радіація

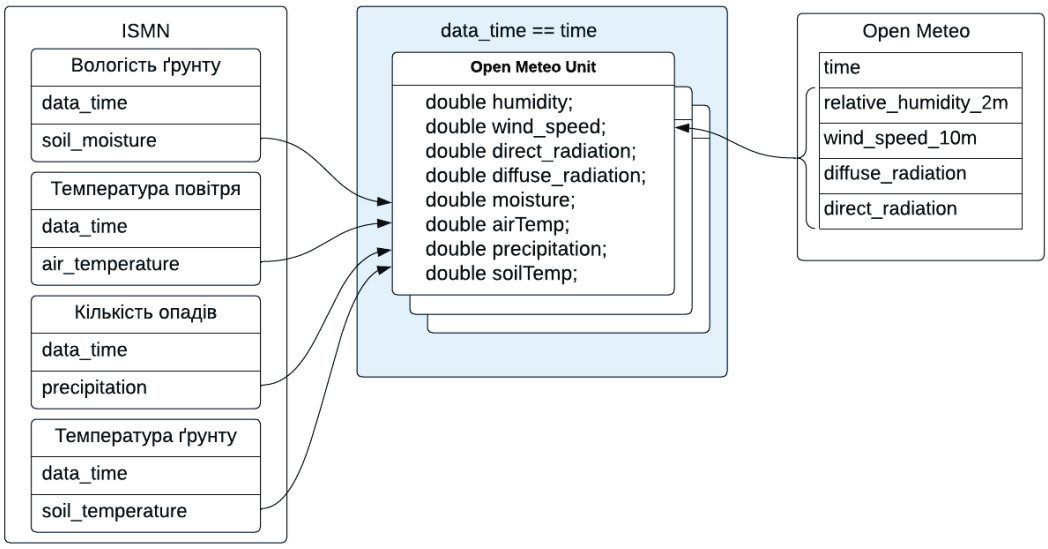

Рис. 1. Схема агрегації кліматичних даних для прогнозування вологості ґрунту / Scheme of aggregation of climatic data for forecasting soil moisture

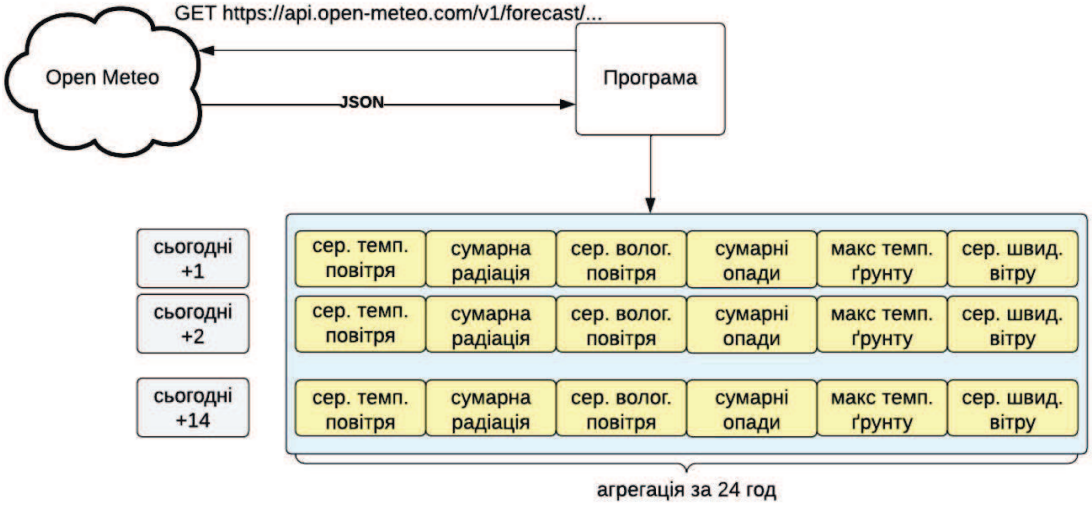

Рис. 2. Приклад отримання та агрегації кліматичних даних для щоденного прогнозу / An example of obtaining and aggregating climate data for a daily forecast

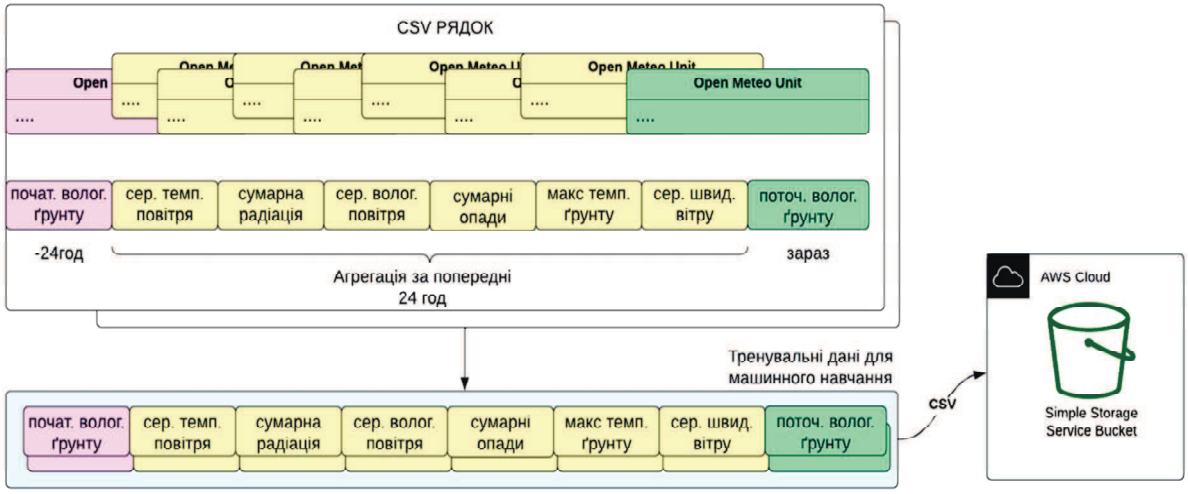

Рис. 3. Формування та збереження набору даних для навчання алгоритмів / Forming and saving a set of data for training algorithms

Кінцевий набір даних зберігається до AWS S3 сховища у вигляді CSV файла. Хмарне сховище надалі використовується як джерело вхідних даних для машинного навчання, що дає змогу не додавати CSV або ж JSON документи до кінцевої програми та мати можливість використовувати різні типи навчальних даних, збережених у сховищі.

Алгоритми машинного навчання та програмна реалізація прототипу для вирішення завдання прогнозування вологості ґрунтів. Для побудови моделі вибрано та проаналізовано п'ять алгоритмів машинного навчання:

- Алгоритм регресійних дерев.
- Алгоритм випадкового лісу.
- Алгоритм лінійної регресії.
- Алгоритм M5P.
- Алгоритм K*.

Машинне навчання дає змогу системі формувати прогнози, досліджуючи історичні дані. Загалом процес навчання розділяють на такі групи: навчання кероване (з вчителем), без контролю, напівконтрольоване і навчання із підкріпленням [12, 13].

Кероване навчання (анг. Super leaning) – це метод, в якому модель тренують об'єкти входу, наприклад, параметри змінних передбачень, а також бажане значення виходу. Алгоритми керованого навчання потребують, щоб користувач задав певні керівні параметри. У результаті отримують функцію між входами і виходами за вибраними навчальними даними.

Алгоритм лінійної регресії – класичне навчання "з учителем", або кероване навчання. Лінійна регресія – це тип прогнозного аналізу, який намагається передбачити значення однієї залежної змінної за допомогою іншої незалежної змінної. Він оцінює коефіцієнти лінійного рівняння, що містить одну або більше незалежних змінних, які найкраще передбачають залежну змінну та відповідають прямій лінії або поверхні, що зменшує варіацію між прогнозованими та фактичними вихідними значеннями [14].

Алгоритм випадкового лісу – це приклад моделей машинного навчання, які належать до ансамблевих методів. Точність прогнозування з використанням алгоритму випадкового лісу порівняно з іншими методами досліджено в роботі [15]. Перевагою методу випадкового лісу є можливість досліджувати нелінійні та ієрархічні зв'язки між предикторами очікуваним прогнозом за допомогою ансамлевого навчання. Такі методи передбачають багаторівневі спроби прогнозувати змінну на основі різних даних або моделей. Використання таких спроб для прогнозування реакції може підвищити стійкість і точність прогнозу порівняно з використанням будь-якого іншого набору або моделі [16].

Алгоритм M5P – це керований алгоритм машинного навчання, який ґрунтується на структурі дерева для класифікації даних [17]. Алгоритм M5P є різновидом дерева рішень. Алгоритм M5P працює швидше, ніж регресійні алгоритми, може ефективно обробляти велику кількість наборів даних із багатьма атрибутами і вимірами.

*Алгоритм K** широко досліджений у багатьох працях. Доведено, що це ефективний метод навчання із учителем, який використовують для класифікації [18]. Метод оснований на ентропії для обчислення відстані між навчальними вибірками під час класифікації, забезпечує високу здатність класифікації збалансованих даних. Його недолік проявляється під час роботи з незбалансованими даними.

Навчання цих алгоритмів на єдиній моделі вхідних даних дає змогу порівняти їх точність прогнозування та вплив вхідних параметрів на кінцеве значення порівняно із реальним у випадку історичної довідки.

Отже, програмний продукт повинен уможливити порівняння ефективності використання вибраних алгоритмів. Їх колекцію для реалізації машинного навчання надає бібліотека WEKA, що вільно інтегрується в проєкти на основі мови Java та підтримує велику кількість алгоритмів класифікації та регресії, такі як дерева рішень, класифікатори на основі правил, опорні векторні машини та нейронні мережі.

Згідно із вимогами до програмного продукту, система має прогнозувати вологість ґрунту на деякий час наперед, максимум на 14 днів. Крім того, під час перевірки історичних даних програма так само повинна передбачати дані більш ніж на 1 добу. Однак вибраний алгоритм навчання містить дані лише за попередню добу і може оперувати тільки цими показниками для навчання та передбачення.

Застосовано набір даних, оснований на прогнозі на добу наперед. Однак, якщо потрібен прогноз на більшу кількість діб, можна робити декілька прогнозів на кожну наступну ітерацію, використовуючи значення початкової вологості ґрунту із результату передбачення попередньої ітерації (рис. 4).

Цей підхід задовольняє потребу в гнучких прогнозах на будь-якому проміжку часу за наявності кліматичних даних для розрахунку.

Обговорення результатів дослідження. Програмний продукт розроблено так, що користувач може вибирати алгоритм для тестування, а також бачити інформацію стосовно кількості тренувальних даних, результатів перехресного тестування алгоритмів тощо. На рис. 5 наведено фрагмент користувацького інтерфейсу, відповідальний за цю функціональність.

Важливою характеристикою є швидкість тренування алгоритмів машинного навчання, яка визначає, наскільки швидко алгоритм здатний опрацювати інформацію з навчального набору даних та згенерувати модель для задач прогнозування. Це значення важливе для ефективного виконання експериментів з моделями, налаштування параметрів та використання машинного навчання для реальних задач. Значення залежить від складності алгоритму, хоча не обов'язково означає пропорційну точність прогнозування. Вибрано п'ять кроків з пропорційним зменшенням кількості тренувальних даних (табл. 3).

За результатами вимірювання можна спостерігати, що найменше часу для навчання потребують алгоритми регресійних дерев, лінійної регресії та K*. Якщо для двох перших швидкодію можна пояснити порівняно низькою складністю алгоритмів, то для K* малий показник тривалості навчання є наслідком того, що цей алгоритм використовує підхід "ледачого навчання". Це означає, що він не створює модель під час тренування, а зберігає увесь тренінговий набір та використовує його для класифікації нових екземплярів у режимі реального часу. M5P забезпечив середню тривалість навчання, у 5–10 разів меншу за алгоритм випадкових лісів. Хоча ці два алгоритми використовують дерево рішень, алгоритм випадкових лісів потребує більше часу на навчання зокрема через технологію bagging, що покликана зменшити дисперсію даних.

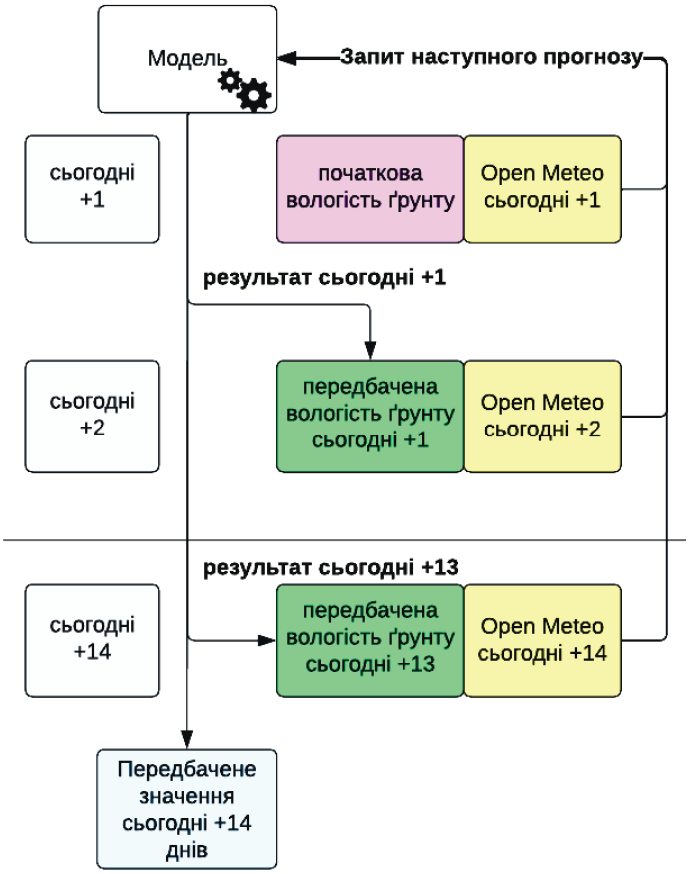

Рис. 4. Алгоритм визначення початкового значення вологості ґрунту для щоденних передбачень протягом тривалого періоду / Algorithm for determining the initial value of soil moisture for daily predictions over a long period of time

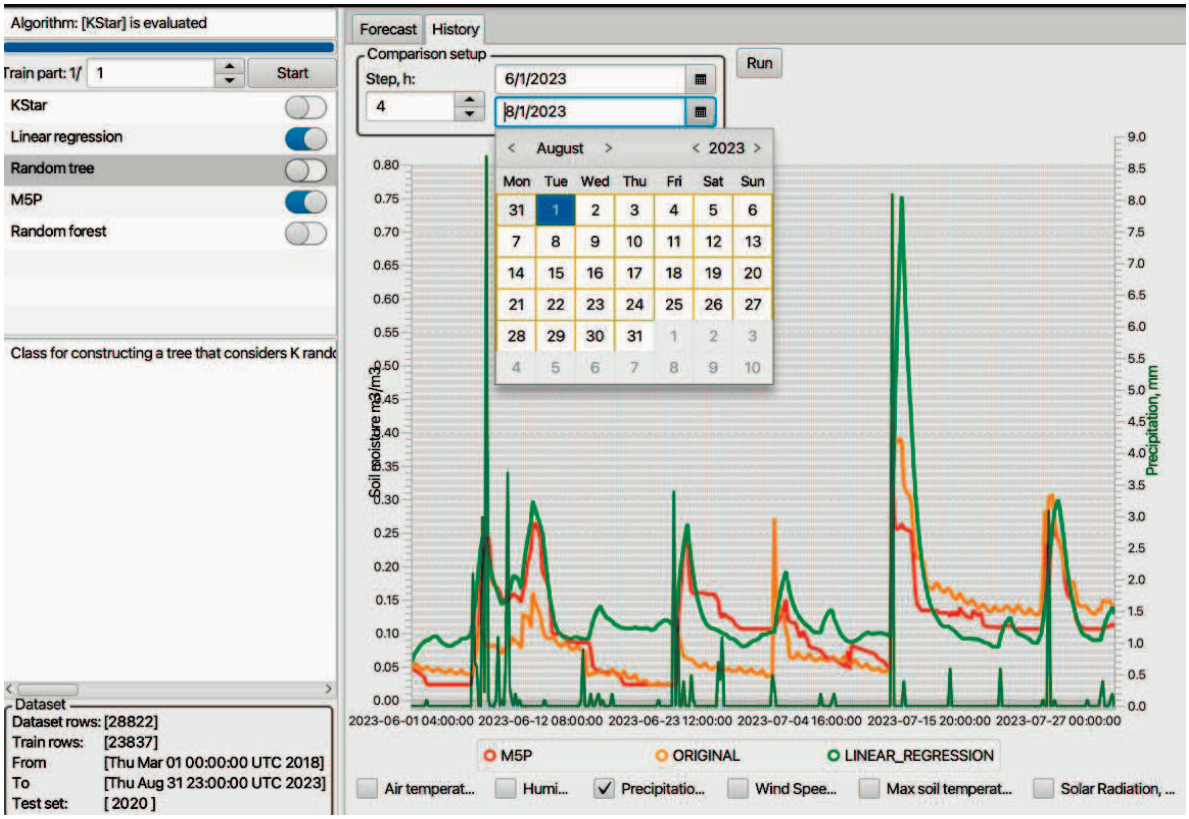

Рис. 5. Інтерфейс тестування алгоритмів за допомогою історичних даних / An interface for testing algorithms using historical data

Табл. 3. Зміна кількості прикладів даних для кожного із кроків тестування/Changing the number of data samples for each test step

Крок	Кількість прикладів
1/1	23837
1/2	11919
1/3	5960
1/4	2980
1/5	1490

Табл. 4. Тривалість навчання моделей, мс / Duration of model training, ms

Крок	Регресійні дерева	Випадковий ліс	Лінійна регресія		M5P	K*
1/1	144	5223	46	920		11
1/2	42	2263	5	242		1
1/4	19	1091	2	124		0
1/8	11	582	3	102		0
1/16	4	233	1	30		0

Табл. 5. Тривалість перехресної валідації, мс / Duration of cross-validation, ms

Крок	Регресійні дерева	Випадковий ліс	Лінійна регресія	M5P	K*
1/1	1194	14389	596	4127	441761
1/2	395	7822	59	2039	115941
1/4	200	4831	29	1060	32464
1/8	126	2272	24	601	8505
1/16	48	1150	8	263	2224

У табл. 4 наведено результати тривалості навчання моделей алгоритмів.

Тестування алгоритмів здійснено на тренувальних даних за допомогою перехресної валідації та на тестових даних, із порівнянням реальних показників вологості ґрунту з прогнозованими алгоритмами. Мета тестування – довести ефективність побудованих моделей та зробити висновок щодо доцільності використання алгоритмів для передбачень майбутніх періодів. Показники ефективності, отримані за її результатами, показують, наскільки точно модель може опрацьовувати нові для неї дані. У табл. 5 наведено час виконання перехресної валідації моделей алгоритмів залежно від кількості початкових даних.

За результатами тривалості валідації можна зробити висновок, що час тестування значно більший за час навчання. Це пов'язано із алгоритмом виконання перехресної валідації, яка відбувається за десять ітерацій. Загалом, можна спостерігати пропорційні залежності для усіх видів алгоритмів порівняно із тривалістю навчання, за винятком алгоритму K*, максимальний час навчання у якому – майже 442 секунди. Це пояснюється тим, що алгоритм K* оснований на прикладах, тому зберігає увесь набір навчальних даних для класифікації. Це забирає багато часу під час перехресного тестування, оскільки для кожної ітерації алгоритм порівнює новий приклад зі збереженими навчальними прикладами.

Розглянемо показники ефективності за результатами перехресної валідації.

Графік змін значень коефіцієнта кореляції залежно від кількості навчальних даних за результатами перехресної валідації наведено на рис. 6.

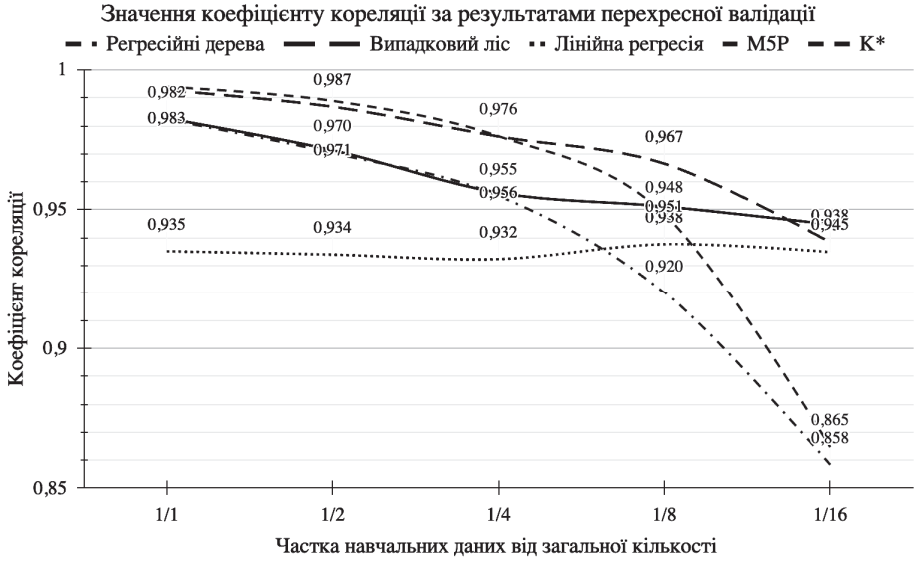

Рис. 6. Графік змін показників коефіцієнта кореляції / Graph of changes in the correlation coefficient

На графіку можна спостерігати поступове зменшення коефіцієнта кореляції для всіх типів алгоритмів, окрім алгоритму лінійної регресії. Істотне зменшення значення свідчить, що для моделей, побудованих на основі алгоритмів, таких як регресійні дерева та K*, характерна сильна залежність від кількості вхідних даних. Це може свідчити, що вони схильні до перенавчання та слід очікувати, що прогнозні значення на малій кількості вхідних даних будуть неточними. Показник для лінійної регресії є стабільним, а це вказує на те, що для надання максимально точного прогнозу із застосуванням цієї моделі кількість навчальних даних не є критичною. Решта моделей, такі як M5P та випадковий ліс, більш-менш стабільні, помітне зменшення коефіцієнта кореляції спостерігається лише за кількості даних 1/16 для алгоритму "випадкового лісу".

Дослідження моделей із використанням тестового набору даних. Користувацький інтерфейс програмного продукту дає змогу протестувати алгоритми на реальних історичних кліматичних даних. Для цього потрібно вибрати часовий інтервал протягом 2020 р., дані якого не використовували для тренування моделей. Для тестування було вибрано період з 01.07 до 31.07, тобто липень, що є активним з погляду меліоративних робіт. За базове реальне значення вологості ґрунту взято значення за 30 червня, всі решта відповідно отримано за допомогою машинного навчання та показників зміни клімату. Крок вимірювання становив 2 год, що зменшило витрати часу на опрацювання алгоритмів та забезпечило порівняно стабільні показники зміни кліматичних умов. На рис. 7 зображено зміни реальних показників вологості ґрунту за вибраний період із додатковим накладанням графіка зміни кількості опадів, де можна простежити кореляцію між цими величинами.

З графіка (рис. 7) бачимо, що майже завжди опади сприяли підвищенню рівня вологості, хоча в деяких випадках змін не відбулося. Це може бути пов'язано із впливом інших факторів, таких як температура повітря, що сприяла швидкому висиханню ґрунту тощо. Також існує ефект накопичення вологи, коли унаслідок опадів протягом певного періоду ґрунт збирає вологу. Результати тестування алгоритмів за цей період, своєю чергою, повинні показати, наскільки модель вхідних даних та алгоритм їх опрацювання співпрацюють із вибраними алгоритмами машинного навчання для прогнозування вологості ґрунту на основі впливів кліматичних показників. Вибрані алгоритми протестовано на різній кількості вхідних даних для визначення залежності їх ефективності від об'єму попереднього навчання. Швидкодія усіх алгоритмів, окрім K*, на високому рівні, становить менш ніж 15 мс. Однак для K* цей показник більший, та, очікувано, знижується зі зменшенням кількості навчальних даних (табл. 6).

За результатами вимірювання часу розрахунку прогнозованих величин можна зробити висновок, що алгоритм K*, попри найшвидше навчання, потребує найбільшого часу для тестування та визначення прогнозованого значення. На рис. 8 подано результати прогнозування вологості ґрунту за допомогою створеного програмного застосунку та алгоритмів машинного навчання.

Проаналізовано вплив кліматичних ознак на остаточне значення вологості ґрунту в тренувальному наборі даних. За результатами з'ясовано, що найістотніше прямо впливають, окрім, попередніх показників, показники вологості повітря та кількості опадів.

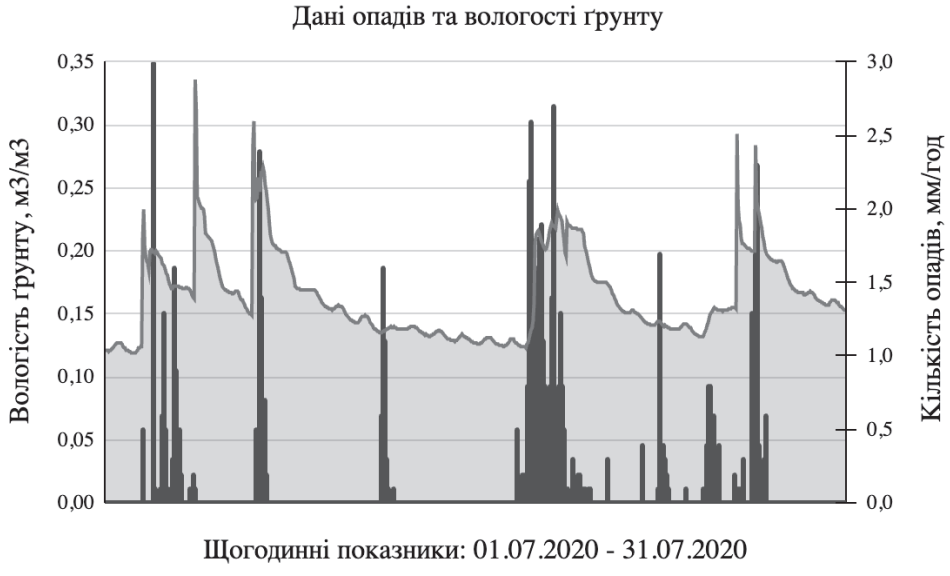

Рис. 7. Реальні показники зміни вологості ґрунту та кількість опадів /
Real indicators of changes in soil moisture and amount of precipitation

Табл. 6. Час прогнозування для моделі алгоритму K*, мс / Prediction time for the K* algorithm model, ms

1/1	1/2	1/4	1/8	1/16
7528	3600	1130	1120	1000

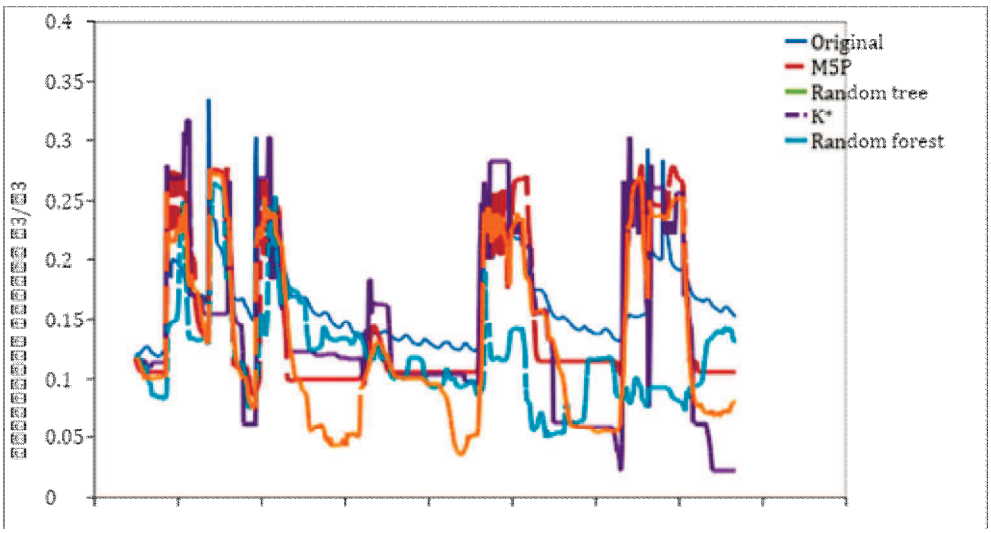

Рис. 8. Результати тестування алгоритмів прогнозування вологості ґрунту / Results of testing algorithms for predicting soil moisture

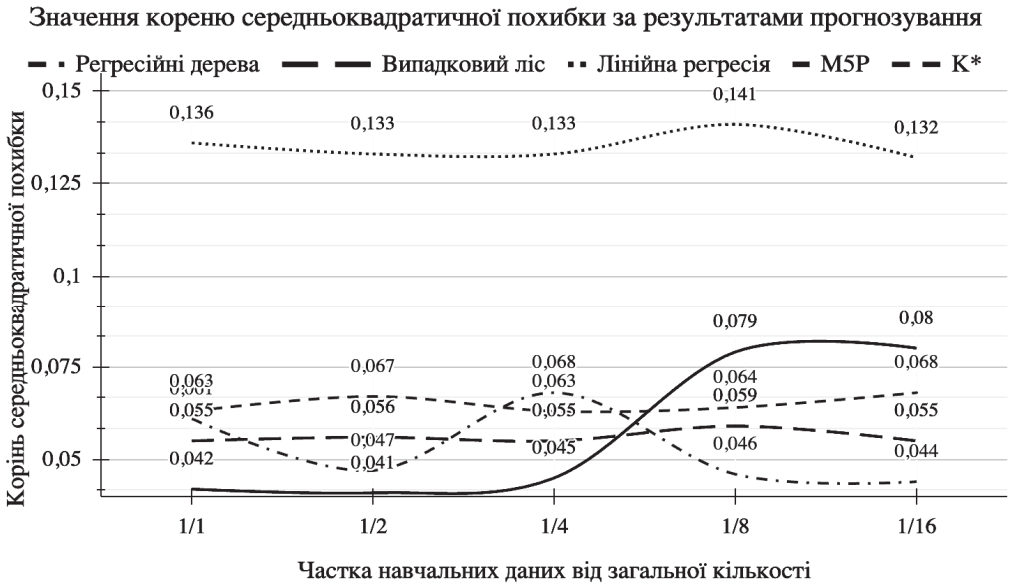

Рис. 9 . Зміна значення кореня середньоквадратичної похибки / Change of the value of the root mean squared error

За результатами досліджень зроблено висновки щодо кожного окремого алгоритму і доцільності його використання для прогнозування на основі кліматичних показників за допомогою реалізованого алгоритму. Виявлено, що найефективнішим алгоритмом для виконання поставленого завдання є алгоритм M5P, навчений на максимальній кількості тренувальних даних та застосований для передбачень на терміни до 14 днів. Також спостерігаємо, що решті алгоритмів не вдалось показати більш-менш стабільні та точні результати протягом тривалого періоду. Ефективність їх застосування, залежно від кількості тренувальних даних, можна спостерігати на графіку зміни значення кореня середньоквадратичної похибки (рис. 9), де M5P має найменше значення серед інших за повного обсягу тренувальних даних. Найбільше значення похибки в алгоритму лінійної регресії, що свідчить про найнижчу точність прогнозування.

Результати дослідження дають змогу сформулювати наукову новизну та практичну значущість.

Наукова новизна отриманих результатів дослідження – метод інтеграції даних та рекомендації щодо ефективності алгоритмів машинного навчання для прогнозування вологості ґрунту на основі кліматичних показників.

Практична значущість результатів дослідження – виконані дослідження та їх результати демонструють особливості використання алгоритмів машинного навчання та можуть бути корисними рекомендаціями для створення програмного застосунку для прогнозування вологості ґрунтів, що дасть змогу оперативно реагувати на прогнозовану зміну кліматичних умов та, за умови інтеграції із системами розумного землеробства, сприятиме раціональному використанню водних ресурсів, електроенергії, палива, добрив, інших ресурсів.

Висновки / Conclusions

1. Прогнозування вологості ґрунтів, важливе у сфері розумного землеробства, досліджено з використанням алгоритмів машинного навчання та реалізованого на їх основі програмного забезпечення.

2. Для тренування моделей використано два відкриті міжнародні інформаційні ресурси, які збирають кліматичні дані: температуру повітря, відносну вологість повітря, температуру ґрунту, кількість опадів, кількість сонячної радіації, силу вітру за період з 2018-03-01 до 2023-08-31.

3. Для побудови моделі прогнозування вологості ґрунтів вибрано і досліджено алгоритми машинного навчання: алгоритм лінійної регресії, алгоритм регресійних дерев, алгоритм "випадкового лісу", алгоритм M5P, алгоритм K*.

4. З використанням бібліотеки WEKA на мові програмування Jawa розроблено програмну систему прогнозування вологості ґрунтів, яка дає змогу вибирати алгоритм, кількість тренувальних даних та використовувати реалізовану модель як для тестування, так і для прогнозування значень вологості ґрунту.

5. За допомогою розробленого алгоритму агрегування даних здійснено навчання алгоритмів на однакових наборах вхідних даних, що дало змогу порівнювати їхню ефективність та точність прогнозування.

6. За результатами тестування на основі прогнозу протягом 31 дня на історичному періоді часу та порівняння з реальними показниками ґрунту за той самий період найточніші прогнози надав алгоритм M5P, але за умови використання якнайбільшої кількості тренувальних даних за період, що не перевищує 14 днів. Його результати корелюють із реальними кліматичними змінами, тому його доцільно надалі досліджувати на предмет покращення показників прогнозування.

7. За результатами досліджень зроблено висновки щодо доцільності використання кожного окремого алгоритму для прогнозування на підставі кліматичних показників за допомогою реалізованого алгоритму. З'ясовано, що найефективнішим для розв'язання поставленої задачі є алгоритм M5P.

References

[1] Araújo, S. O., Peres, R. S., Barata, J., Lidon, F., & Ramalho, J. C. (2021). Characterising the Agriculture 4.0 Landscape – Emerging Trends, Challenges and Opportunities. Agronomy, 11(4), 2-37. https://doi.org/10.3390/agronomy11040667

[2] Dawn, N., Ghosh, T., Ghosh, S., Saha, A., Mukherjee, P., Sarkar, S., Guha, S., & Sanyal, T. (2023). Implementation of Artificial Intelligence, Machine Learning, and Internet of Things (IoT) in revolutionizing Agriculture: A review on recent trends and challenges. *International Journal of Experimental Research and Review*, 30, 190–218. https://doi.org/10.52756/ijerr.2023.v30.018

[3] Singh, A., Gaurav, K. (2023). Deep learning and data fusion to estimate surface soil moisture from multi-sensor satellite images. *Sci. Repб.*, 13, 2251. https://doi.org/10.1038/s41598-023-28939-9

[4] Li, Q., Li, Z., Shangguan, W., Wang, X., Li, L. & Yu, F. (April 2022). Improving soil moisture prediction using a novel encoder-decoder model with residual learning. *Computers and Electronics in Agriculture, 195*, April 2022. 106816. https://doi.org/10.1007/978-3-030-26630-1_7

[5] Ren, Y., Ling, F., & Wang, Y. (2023). Research on Provincial-Level Soil Moisture Prediction Based on Extreme Gradient Boosting Model. *Agriculture*, *13*(5), 927. https://doi.org/10.3390/agriculture13050927

[6] Jiang, K., Pan, Z., Pan, F., Teuling, A. J., Han, G., An, P., Chen, X., ..., & Dong, Z. (16 June 2023). Combined influence of soil moisture and atmospheric humidity on land surface temperature under different climatic background. *iScience, 26*(6), 106837. https://doi.org/10.1016/j.isci.2023.106837

[7] Ariyanto, D. P., Qudsi, Z. A., Sumani, Dewi, W. S., Rahayu, & Komariah. (2021). The dynamic effect of air temperature and air humidity toward soil temperature in various lands coverat KHDTK Gunung Bromo, Karanganyar – Indonesia, IOP Conf. Ser. Earth Environ. Sci., 724, 1, 012003. https://DOI10.1088/1755-1315/724/1/012003

[8] Gikunda, P. K. & Jouandeau, N. (2019). Modern CNNs for IoT Based Farms. *Communications in Computer and Information Science, 1026*, 68-79. https://doi.org/10.48550/arXiv.1907.0777

[9] International Soil Moisture Network [Електронний ресурс]. Retrieved from: https://ismn.earth/en/

[10] Free Weather API [Електронний ресурс]. Retrieved from: https://open-meteo.com/

[11] Frank, E., Hall, M. A., Witten, I. H. (2016). The WEKA Workbench. In Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 4th ed.; Elsevier: San Francisco, CA, USA, 1-128.

[12] Torgo, L. (2011). Regression Trees. In: Sammut, C., Webb, G.I. (eds) Encyclopaedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0387-30164-8711

[13] Ziqiu, Kang , Cagatay, Catal, Bedir, Tekinerdogan. (2020). Machine learning applications in production lines: A systematic literature review. *Computers & Industrial Engineering, 149*, 106773. https://doi.org/10.1016/j.cie.2020.106773

[14] KDAG IIT KGP Linear Regression h- [Електронний ресурс]. Retrieved from: https://kdagiit.medium.com/linear-regression-ba3fe4ba38c0

[15] Jaiswal, Jitendra, Samikannu, Rita (2017). Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression, Conference: 2017 World Congress on Computing and Communication Technologies (WCCCT), 65–68. https://doi.org/10.1109/WCCCT.2016.25

[16] Everingham, Y., Sexton, J., & Skocaj, D. (2016). Accurate prediction of sugarcane yield using a random forest algorithm Agron. Sustain. *Dev*., 36: 27. https://doi.org/10.1007/s13593-016-0364-z

[17] Hasup Song, Hasup Song, Injong Gi, Jihyuk Ryu, Yonghwan Kwon, Jongpil Jeong. (2023). Production Planning Forecasting System Based on M5P Algorithms and Master Data in Manufacturing Processes. *Appl. Sci., 13*(13), 7829. https://doi.org/10.3390/app13137829

[18] Goksu Tuysuzoglu, Kokten Ulas, Birant & Derya Birant Rainfall. (2023). Prediction Using an Ensemble Machine Learning Model Based on K-Stars, Sustainability, 15(7), 5889. https://doi.org/10.3390/su15075889

D. V. Fedasyuk, M. O. Kostiuk

Lviv Polytechnic National University, Lviv, Ukraine

FORECASTING OF SOIL MOISTURE USING MACHINE LEARNING IN SMART AGRICULTURE SYSTEMS

Growing crops in modern conditions is a complex task and practically combines the practices of experience and the latest methods, including information technology, which has become part of the concept of "smart farming". An important factor in the stable predicted yield is the level of soil moisture, which is the result of changes in climatic factors such as air

temperature, soil temperature, intensity of solar radiation, rainfall, wind speed, etc. A methodology for processing real historical indicators of climate change in a certain geographical area with subsequent training and application of machine learning models to predict soil moisture is proposed. To build a machine learning model, the following algorithms were selected and studied: the algorithm of regression trees, random forest, linear regression, M5P algorithms and the K* algorithm. The data source for training the models is the open information resource International Soil Moisture Network (ISMN) from ismn.earth/en. , which provides data on soil moisture and temperature, air temperature, and rainfall. Other data was used from the Open Meteo information service, which provides a free API and allows you to get historical data and weather forecast in specified coordinates during specified days. A data structure was developed to train the model for further prediction of soil moisture. An architecture has been developed and a software system for predicting soil moisture based on machine learning algorithms has been created using the Spring Framework, the WEKA library and Java FX with the ability to select and study the appropriate algorithms. Experiments have been carried out and the results of the duration of model training have been presented, while the algorithms of regression trees and linear regression require the least training time. A comparison of algorithms is made according to the following criteria: learning speed, cross-testing speed, prediction speed, testing performance indicators for real historical data. Based on the results of the study, conclusions are drawn about individual algorithms, the feasibility of using them to predict soil moisture based on climatic indicators. The obtained results will make it possible to evaluate and select the best models of machine learning in the design of the information and analytical system "smart agriculture" for forecasting soil moisture.

Keywords: M5P algorithm, linear regression, K*, regression tree algorithms, random forests, foresight.

Інформація про авторів:

Федасюк Дмитро Васильович, д-р техн. наук, професор, завідувач кафедри програмного забезпечення. Email: dmytro.v.fedasyuk@lpnu.ua; https://orcid.org/0000-0003-3552-7454

Костюк Микита Олександрович, магістр кафедри програмного забезпечення. Email: mykyta.kostiuk.mpzip.2022@lpnu.ua; https://orcid.org/0009-0006-0165-2870

Цитування за ДСТУ: Федасюк Д. В., Костюк М. О. Прогнозування вологості ґрунту з використанням машинного навчання у системах розумного землеробства. *Український журнал інформаційних технологій*. 2024, т. 6, № 1. С. 26–36.

Citation APA: Fedasyuk, D. V., & Kostiuk, M. O. (2024). Forecasting of soil moisture using machine learning in smart agriculture systems. *Ukrainian Journal of Information Technology, 6*(1), 26–36. https://doi.org/10.23939/ujit2024.01.026