

ПОРІВНЯННЯ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ ДЛЯ ПРОГНОЗУВАННЯ ВРОЖАЙНОСТІ

А. С. Різниченко^{1,а}, Г. О. Яйлимова¹

¹ Навчально-науковий Фізико-технічний інститут

Анотація

У дослідженні порівнюються класичні алгоритми машинного навчання для задачі прогнозування врожайності зернових культур. Використано статистичні дані Світового банку та супутникові дані (опаді, покриття лісами). Оцінено ефективність Linear Regression, Decision Tree, Random Forest та Gradient Boosting.

Ключові слова: машинне навчання, алгоритм, регресія, сценарії, прогнозування, геопросторові дані, врожайність.

Вступ

Зміна клімату та інтенсивне використання природних ресурсів створюють серйозні виклики для сільського господарства та продовольчої безпеки. Надмірний тиск на ґрунти, зменшення площ орних земель через урбанізацію, а також нестабільність опадів і частіша поява екстремальних погодних явищ впливають на стабільність урожаїв. Особливо це актуально для України, яка є одним із провідних зернових експортерів і залежить від своєчасного планування виробництва.

У той же час розвиток супутникових технологій та зростання обсягів відкритих кліматичних і геопросторових даних дають можливості для моніторингу стану посівів та прогнозування врожайності.

Нещодавня стаття «Forecast of Yield of Major Crops in Ukraine in War Conditions» [1] показала високу практичну цінність дистанційного зондування землі для оцінки врожайності в екстремальних умовах: автори використали 16-денні композити NDVI з сенсора MODIS та Sentinel-2 і навчили для кожного регіону України Random Forest-пересори на офіційних даних врожайності за 2016–2021 роки. Моделі на основі MODIS давали відносну помилку прогнозу в межах 8–18 %, тоді як дані Sentinel-2 краще відтворювали змінність врожайності.

Досягнення в галузі машинного навчання відкрили нові можливості для точного прогнозування врожайності шляхом інтеграції різномірних джерел даних. Систематичний огляд [2] проаналізував понад п'ятдесят наукових праць і показав, що ключовими факторами для прогнозу є кількість опадів та характеристики ґрунту. Автори підкреслюють надзвичайну важливість поєднання різних типів індикаторів для досягнення надійних результатів.

Тому проаналізувавши статті вирішено використо-

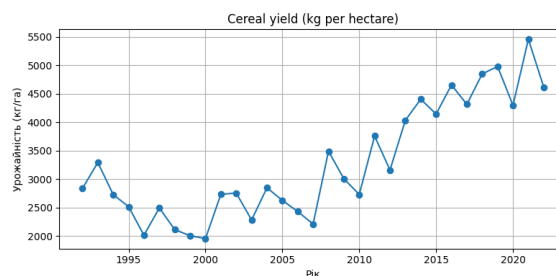


Рис. 1. Зміна урожаю за роками

увати мультифакторний підхід: у дослідженні порівнюється ефективність Linear Regression, Decision Tree, Random Forest та Gradient Boosting на наборі кліматичних, земельних та соціально-економічних показників для прогнозування врожайності зернових культур в Україні.

Застосування методів машинного навчання дозволяє виявляти залежності між багатьма факторами та виробничими результатами. Прогнозування врожайності на основі ML моделювання [3] забезпечує:

- Підвищення точності оцінок майбутнього обсягу виробництва;
- Оптимізацію ресурсів;
- Зменшення економічних втрат;
- Покращення прийняття рішень на рівні агропідприємств і державної політики.

1. Постановка задачі

Треба побудувати модель $f(X) \rightarrow y$, де X – набір індикаторів (урбанізація, водозабір, опади, площі землі тощо), y – врожайність. Необхідно порівняти продуктивність регресійних алгоритмів за метриками MSE, MAE та R^2 .

Математична модель: $\min_{\theta} \sum_{i=1}^n (y_i - f_{\theta}(X_i))^2$, де

^аandriz-ipt25@lil.kpi.ua

f_θ – функція прогнозу.

2. Опис моделей

► Лінійна регресія

Це статистичний метод, який шукає пряму (або багатовимірну гіперплощину) в просторі ознак, що найкраще підходить під дані. Він припускає, що залежність між вхідними змінними і вихідним значенням можна апроксимувати лінійною комбінацією. Було виконано модифікацію «positive=True», щоб уникнути фізично неможливих від'ємних прогнозів.

► Decision Tree (Дерево рішень)

Модель ділить простір ознак на послідовність простих умов, формуючи дерево розгалужень, а в кожному листку дає константний прогноз. Загалом такий алгоритм автоматично захоплює як лінійні, так і складні нелінійні залежності, легко інтерпретується через візуалізацію гілок [4]. Але, без обмежень глибини може перенавчатися. Тому з модифікацій маємо «max depth» – обмеження глибини дерева та «min samples leaf» – мінімальна кількість спостережень у листку.

► Random Forest

Це поєднання багатьох «дерев рішень», кожне з яких навчається на випадково вибраній підмножині даних і випадковій підмножині ознак. Прогноз фінальної моделі обчислюється як середнє передбачення всіх окремих дерев [5]. Такий підхід має зменшувати вплив шуму й підвищувати стабільність.

► Gradient Boosting

Це послідовне додавання невеликих «дерев рішень»: кожне нове дерево навчається виправляти залишки помилок попередньої моделі. В кінці прогноз формується як сума внесків кожного дерева [6], помножених на невеликий крок навчання. Комбінуючи в такий спосіб, отримується досить гнучка та точна апроксимація складних залежностей.

Математично кожний з алгоритмів надано в таблиці 1.

3. Обраний датасет

Обрано набір даних Світового банку [7], що містить показники по Україні за період із січня 1960 до грудня 2023 року. Вся інформація охоплює різні соціально-економічні та кліматичні індикатори, відсоткові та абсолютні величини. В даній роботі для задачі прогнозування врожайності, відібрано ті змінні, які мають безпосереднє або опосередковане відношення до клімату та сільського господарства, і одночасно мають достатню кількість спостережень за роками. Зокрема, до моделі увійшли щорічні дані про демографічний ріст (Population growth – приріст населення, Urban population growth – приріст міського населення), площу та тип земного покриття (Agricultural land (sq. km) – площа сільськогосподарських угідь, Arable land – рілля як % від загальної площі, Forest area – лісові масиви як % від загальної

площі), ресурсова ємність енергетики (Renewable energy consumption), гідрологічні показники (Annual freshwater withdrawals – річний забір прісної води, Average precipitation in depth – середньорічна кількість опадів), а також ключову-цільову змінну – врожайність зернових культур (Cereal yield).

Спочатку дані були приведені до широкого формату, де кожен рядок відповідає конкретному року, а колонки – числовим значенням індикаторів. Після перетворення всіх стовпців у числовий тип, створено набір відібраних даних для періоду 2001–2021 роки та 11 змінних (10 індикаторів та цільова врожайність).

Саме на таких якісно підготовлених даних можна розпочинати процедуру навчання моделей. Завантаживши таблицю в середовище машинного навчання, ми легко розділяємо її на вхідні матриці X та вектор цільових y , а потім по черзі навчаємо й порівнюємо ефективність чотирьох обраних алгоритмів (Linear Regression, Decision Tree, Random Forest та Gradient Boosting), застосовуючи єдині метрики якості. Такий підхід гарантує чесне зіставлення моделей і робить висновок щодо найкращого методу обґрунтованим.

4. Результати порівняння алгоритмів

Після навчання чотирьох алгоритмів (Linear Regression, Decision Tree, Random Forest, Gradient Boosting) на однаковому наборі кліматично-геопросторових та землекористувальних індикаторів отримано такі ключові показники якості (Рис.2, 3, 4).

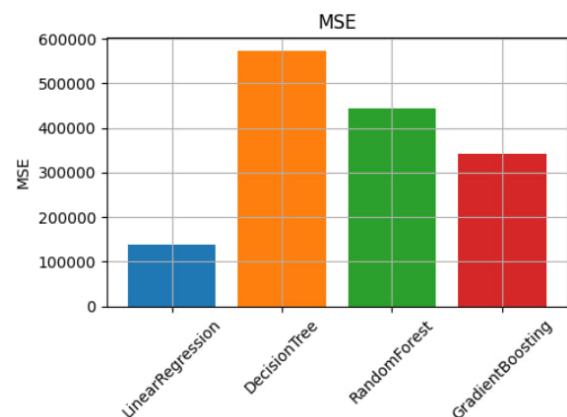


Рис. 2. MSE

Лінійна регресія: дана модель показала найвищий $R^2 \approx 0.84$. Завдяки обмеженню коефіцієнтів на невід'ємність, лінійна регресія адекватно описала тенденцію зміни врожайності й видала мінімальні помилки серед усіх.

Decision Tree: дерево перенавчилося на окремі коливання даних і дало слабкий прогноз $R^2 \approx 0.33$.

Random Forest: середнє зі 100 дерев знизило дисперсію прогнозів порівняно з одним деревом та показало середню точність $R^2 \approx 0.48$.

Gradient Boosting: базові налаштування давали помітно кращі результати, але все ще відставали від

Таблиця 1. Математичний опис алгоритмів

Алгоритм	Математична модель
Linear Regression	$f(X) = \beta_0 + \sum_{j=1}^p \beta_j x^{(j)}$
Decision Tree	$f(X) = \sum_{m=1}^M c_m 1\{X \in R_m\}$
Random Forest	$f(X) = \frac{1}{K} \sum_{k=1}^K T_k(X)$
Gradient Boosting	$f(X) = \sum_{m=1}^M \gamma_m h_m(X)$

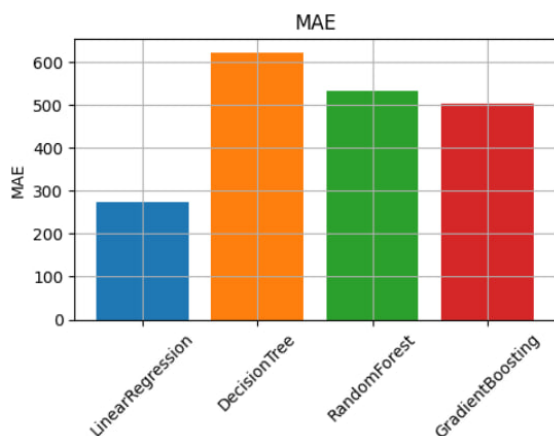


Рис. 3. MAE

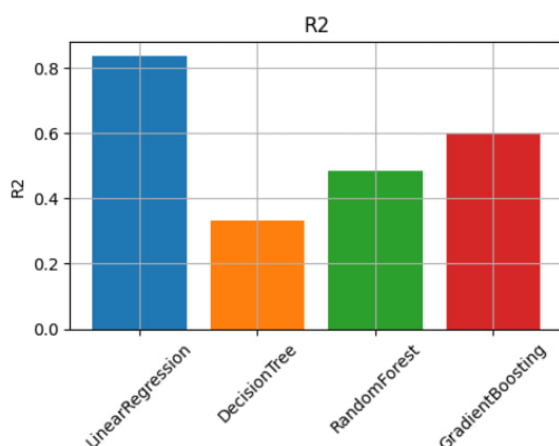


Рис. 4. R^2

простої лінійної моделі $R^2 \approx 0.60$.

Висновки

У цьому дослідженні було порівняно чотири алгоритми машинного навчання – лінійну регресію, дерево рішень, випадковий ліс та градієнтний бустинг, для задачі прогнозування врожайності зернових культур на основі кліматично-ґрунтових та геопросторових даних. Лінійна регресія виявилася найефективнішою завдяки тому, що залежність врожайності від наших показників у досліджуваному періоді була приближена до лінійної. Обмеження коефіцієнтів на невід’ємність забезпечило фізично осмислені прогнози без від’ємних значень. Простота та інтерпретованість лінійної регресії роблять її коректним рішенням до використання.

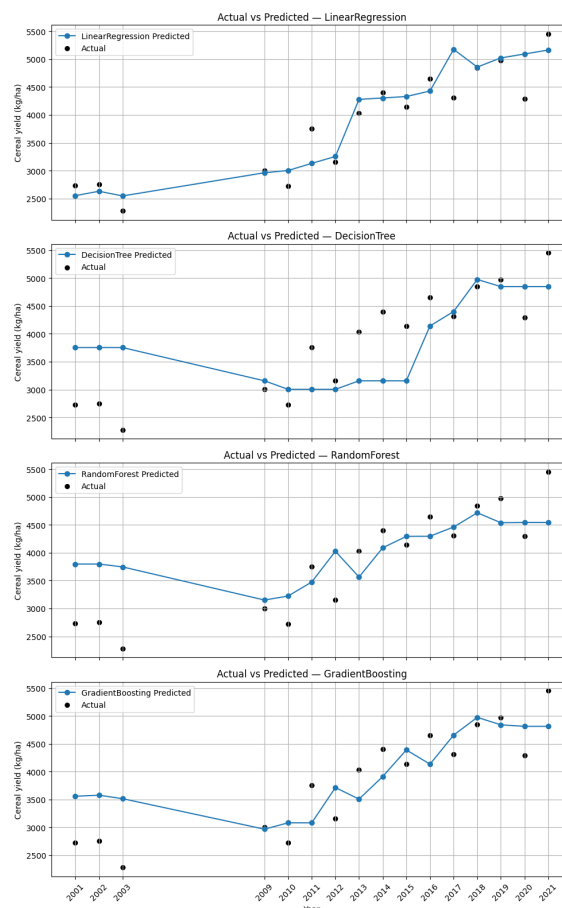


Рис. 5. Фактичне від передбачуваного

Такі алгоритми можуть стати основою для систем підтримки прийняття рішень у аграрному секторі, як в оперативному моніторингу стану посівів, так і в довгостроковому плануванні посівних площ, оптимальному використанні ресурсів. Отже, на вибірці історичних даних машинне навчання дозволяє будувати досить точні прогнози врожайності.

Перелік використаних джерел

1. Kussul N., Drozd S., Yailymova H. Forecast of Yield of Major Crops in Ukraine in War Conditions 2022 Based on MODIS and Sentinel-2 Satellite Data // Proceedings of the 11th International Conference on Applied Innovations in IT (ICAIIIT). — Kyiv, Ukraine, 03.2023. — С. 89—95. — URL: https://icaiiit.org/paper.php?paper=11th_ICAIIIT_1/2_5.

2. *Klompenburg T. van, Kassahun A., Catal C.* Crop yield prediction using machine learning: A systematic literature review // *Computers and Electronics in Agriculture*. — 2020. — Т. 177. — URL: <https://doi.org/10.1016/j.compag.2020.105709>.
3. Machine learning in agriculture: A review / K. Liakos, P. Busato, D. Moshou, S. Pearson, D. Bochtis // *Sensors*. — 2018. — Т. 18, № 8. — URL: <https://www.mdpi.com/1424-8220/18/8/2674>.
4. Classification and Regression Trees / L. Breiman, J. Friedman, R. Olshen, C. Stone. — Belmont, CA : Wadsworth, 1984.
5. *Breiman L.* Random Forests // *Machine Learning*. — 2001. — Т. 45, № 1. — С. 5—32. — URL: <https://link.springer.com/article/10.1023/A:1010933404324>.
6. *Friedman J.* Greedy Function Approximation: A Gradient Boosting Machine // *The Annals of Statistics*. — 2001. — Т. 29, № 5. — С. 1189—1232. — URL: <https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-5/Greedy-function-approximation-A-gradient-boosting-machine/10.1214/aos/1013203451.full>.
7. *Exchange H. D.* World Bank climate change indicators for Ukraine. — 2025. — URL: <https://data.humdata.org/dataset/world-bank-climate-change-indicators-for-ukraine> ; Dataset.