Міністерство освіти і науки України
Національний університет водного господарства та природокористування
Навчально-науковий інститут кібернетики, інформаційних технологій та інженерії
Кафедра комп'ютерних технологій та економічної кібернетики

Допущено до захисту:
Завідувач кафедри
комп'ютерних технологій та
економічної кібернетики
д. е. н., проф. П. М. Грицюк

_____
«__»_____2025р

КВАЛІФІКАЦІЙНА РОБОТА
на здобуття ступеня «магістр»
за освітньо-професійною програмою
«Інформаційні технології в бізнесі»
спеціальності 126 «Інформаційні системи та технології»

на тему**: «Моделювання нелінійного впливу кліматичних факторів на врожайність пшениці методами машинного навчання»**

Виконав:
здобувач освіти 2 курсу,
групи ІТБ-61м
Більчук Юлія Петрівна
(прізвище, ім'я, по – батькові)

Керівник:
д.е.н., професор Грицюк П. М.
(науковий ступінь, вчене звання прізвище та ініціали)

Рецензент:
к.т.н., доцент Джоші О. І.
(науковий ступінь, вчене звання прізвище та ініціали)

Рівне – 2025

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Національний університет водного господарства та природокористування
Навчально-науковий інститут кібернетики, інформаційних технологій та інженерії
Кафедра комп'ютерних технологій та економічної кібернетики
Освітньо-кваліфікаційний рівень – магістр
Освітньо-професійна програма
«Інформаційні технології в бізнесі»
Спеціальність 126 «Інформаційні системи та технології»

ЗАТВЕРДЖУЮ
Завідувач кафедри
комп'ютерних технологій та
економічної кібернетики
д. е. н., проф. П. М. Грицюк
_____
«__»_____20__р

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

Більчук Юлії Петрівни
(прізвище, ім'я, по батькові)

1. Тема роботи: *Моделювання нелінійного впливу кліматичних факторів на врожайність пшениці методами машинного навчання*

керівник роботи: д. е. н., проф. Петро Миколайович Грицюк
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

затверджена наказом по університету від _____

2.Термін здачі студентом закінченої роботи: 17 грудня 2025 р.

3. Вихідні дані до роботи: *Статистичні кліматичні показники областей степового регіону України та значення врожайності пшениці*

4. Зміст розрахунково-пояснювальної записки (перелік питань, що їх належить розробити) *мета роботи, пошук та аналіз даних, розроблення відповідних моделей, опис отриманих результатів*

5. Перелік графічного матеріалу:

1. *Кліматичні карти України*
2. *Карти кліматичного районування, родючості грунтів України*
3. *Графіки врожайності зернових культур в Україні*
4. *Таблиці досліджуваних даних*
5. *Графіки дослідження врожайності пшениці*
6. *Графік моделювання залишків врожайності пшениці*

6. Консультація розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Завдання видав (підпис, дата)	Завдання прийняв (підпис, дата)
Розділ I	*Грицюк П.М.*		
Розділ II	*Грицюк П.М.*		
Розділ III	*Грицюк П.М.*		

7. Дата видачі завдання: _____

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів дипломної роботи	Термін виконання етапів роботи	Примітка
1.	*Постановка задачі, загальна характеристика пріоритетних питань*	*11.09.25-20.09.25*	
2.	*Аналіз та пошук даних*	*21.09.25-26.09.25*	
3.	*Завантаження даних*	*27.09.25-29.09.25*	
4.	*Виведення даних*	*30.09.25-09.10.25*	
5.	*Статистичний аналіз даних*	*11.10.25-20.10.25*	
6.	*Створення моделі лінійної регресії*	*23.10.25-05.11.25*	
8.	*Створення моделі лінійної регресії з добутками та квадратичними членами*	*08.11.20-19.11.25*	
9.	*Створення повної нелінійної моделі*	*22.11.25-03.12.25*	
10.	*Створення моделі фактичної динаміки врожайності пшениці*	*05.12.25-17.12.25*	
11.	*Представлення роботи та інших документів до захисту*	*18.12.2025*	

Студент _____
(підпис) (прізвище, та ініціали)

Керівник роботи _____
(підпис) (прізвище, та ініціали)

РЕФЕРАТ

Пояснювальна записка до кваліфікаційної роботи на тему «Моделювання нелінійного впливу кліматичних факторів на врожайність пшениці методами машинного навчання» для здобуття освітнього ступеня «Магістр», за освітньо-професійною програмою «Інформаційні технології в бізнесі» за спеціальністю 126 «Інформаційні системи та технології» написана на 65 сторінок та містить 35 ілюстрації, 2 таблиці, 11 джерел та додатки.

Метою роботи є створення математичної моделі для оцінки впливу кліматичних факторів на врожайність зернових культур з використанням методів машинного навчання.

Об'єктом дослідження є врожайність пшениці в степовій зоні України.

Предметом дослідження є математичні моделі регресійного типу, які дозволяють оцінити вплив кліматичних факторів на врожайність пшениці.

У процесі дослідження нами була використана мова програмування Python, середовища Google Colab, бібліотек sklearn, matplotlib, numpy та pandas.

Практичне значення одержаних результатів полягає у покращенні планування та управління ризиками у сільському господарстві шляхом оцінки впливу температурних чинників на врожайність зернових культур.

Ключові слова: машинне навчання, лінійна регресія, нелінійна регресія, статистична значущість.

ПЕРЕЛІК СКОРОЧЕНЬ ТА УМОВНИХ ПОЗНАЧЕНЬ

ML	–	Машинне навчання
AIC	–	Інформаційний критерій Акаіке
BIC	–	Баєсівський інформаційний критерій
eps	–	Трендове відхилення
RMSE	–	Середні помилка прогнозування
K-fold	–	Крос валідація
MSE	–	Середньоквадратична помилка
std	–	Стандартне відхилення
MAE	–	Середня абсолютна помилка

ЗМІСТ

ВСТУП

Сільське господарство є основою продовольчої безпеки країни та ключовим сектором економіки. В умовах глобальних кліматичних змін, зростання населення та обмежених природних ресурсів виникає нагальна потреба в ефективних підходах до планування і прогнозування врожайності сільськогосподарських культур. Одним із важливих завдань сучасного агропромислового комплексу є точне визначення впливу кліматичних факторів на урожай, особливо зернових культур, які становлять основу аграрного виробництва в Україні.

Враховуючи складність взаємодії між погодними умовами, типами ґрунтів, режимами зрошення та іншими природними параметрами, традиційні методи оцінювання часто не забезпечують належної точності. У цьому контексті застосування математичних моделей, зокрема методів лінійної регресії, дозволяє об’єктивно оцінити залежності між врожайністю та кліматичними умовами на основі наявних статистичних даних.

Лінійна регресія як один із базових інструментів машинного навчання дозволяє ефективно моделювати взаємозв’язки між змінними, виявляти тенденції та будувати прогнози. Її застосування у сільськогосподарському секторі відкриває нові можливості для підвищення ефективності аграрного менеджменту, дозволяючи заздалегідь передбачити можливі ризики, оптимізувати використання ресурсів та ухвалювати обґрунтовані агротехнічні рішення.

Актуальність роботи полягає у створенні ефективної математичної моделі для прогнозування врожайності зернових культур на основі кліматичних факторів із використанням методів лінійної регресії.

Об’єкт дослідження —врожайність зернових культур в степовій зоні України.

Предмет дослідження — математичні моделі регресійного типу, які пояснюють вплив кліматичних факторів на врожайність.

Мета дослідження полягає у створенні математичних моделей залежності врожайності зернових культур від кліматичних факторів на основі методу лінійної регресії, з використанням сучасних інструментів машинного навчання.

Досягнення зазначеної мети передбачає виконання таких завдань:

- Пошук та попередня обробка даних, що включає кліматичні показники та фактичні значення врожайності зернових культур.
- Проведення статистичного аналізу вхідних змінних.
- Побудова лінійної регресійної моделі, яка дозволяє оцінити вплив кліматичних факторів на врожайність.
- Побудова моделі регресії з лінійними факторами та їх квадратами.
- Побудова моделі регресії з лінійними факторами, їх квадратами та добутками.
- Застосування методу перехресної крос-валідації для оцінки точності, та надійності побудованої моделі.
- Розроблення висновків на основі отриманих результатів та надання рекомендацій щодо їх практичного використання.

Програмні продукти та бібліотеки, що використовувались під час розробки та навчання моделі: Google Colab, мова програмування Python, бібліотеки sklearn, matplotlib, numpy та pandas.

РОЗДІЛ 1. Особливості клімату степової зони України

Територія України лежить у зонах мішаних лісів, лісостепу та степу (Рис.1.1). Степова зона — це безлісна територія, що простягається від лісостепу на південь до Чорного та Азовського морів. Поверхня степів рівнинна, подекуди з горбами, ярами та балками.

Рис. 1.1. Кліматичні зони України

У степовій зоні сонце знаходиться вище над горизонтом, ніж у зоні мішаних лісів. Через це сонячні промені падають прямiше й сильніше нагрівають земну поверхню.

Літо в степу довге, сонячне та спекотне. Опадів випадає мало. Осінь зазвичай тепла, але в другій її половині починаються дощі. Зима коротка, холодна й малосніжна. Весна настає рано. Унаслідок швидкого підвищення температури повітря волога з ґрунту швидко випаровується.

У весняно-літній період часто спостерігаються гарячі суховії, які спричиняють посухи. Взимку холодні вітри призводять до хуртовин і чорних бур, що руйнують родючий шар ґрунту.

Через степову зону до морів течуть великі річки України, зокрема Дніпро. У дельті Дунаю багато прісних озер, а на узбережжі Чорного моря — солоні озера-лимани. На Дніпрі створено каскад водосховищ.

У степу переважають трав'янисті рослини. Дерева й кущі трапляються здебільшого вздовж водойм і в балках, де є волога.

Ранньою весною, коли в ґрунті ще достатньо вологи, степ квітне. З'являються півники, гіацинти, крокуси, горицвіт, тюльпани, півонії, маки. До настання спеки ці рослини відцвітають, дають насіння, а їх наземна частина відмирає. У ґрунті залишаються бульби, цибулини й кореневища, де накопичуються поживні речовини для наступного року.

Пізніше їх змінюють рослини, пристосовані до спеки й посухи: полин, типчак, ковила. Одні мають довге коріння, що дістає вологу з глибини, інші — вузькі, опушені або жорсткі листки, які мінімізують випаровування.

У середині літа степ висихає. Рослини в'януть, і вітер котить їх клубками по рівнині, розсіюючи насіння. Степ стає сірим і непривітним.

Для збереження унікальної природи степу створено низку заповідників: Асканія-Нова, Луганський, Український степовий [1].

Степ для українців і всіх народів, які жили тут — від Дону до Дунаю — був не лише природною зоною, а важливою частиною культури, господарства, світогляду. Племена і народи залишили тут слід, формуючи архетипи, господарську модель і культурне середовище. У цих умовах сформувалась ксеротермічна (сухо-спекотна) рослинність, чисельна травоїдна фауна та родючі чорноземи [2].

Степова зона включає в себе Кримську, Дніпропетровську, Донецьку, Херсонську, Кіровоградську, Луганську, Миколаївську, Одеську та Запорізьку області (Південна зона). Площа становить близько 240 200 кв. км., рельєф здебільшого рівнинний.

Клімат зони помірно-континентальний із спекотним літом і холодною зимою (Рис.1.2.) Середня температура січня становить від -5 до -7°С, в липні

– +21- +23°С. Тривалість періоду вегетації, як правило, становить 210-245 днів, а періоду із середньою температурою понад +15°С – 120-140 днів. Річна сума температур, які перевищують 10°С, становить від 2 800 до 3 600.

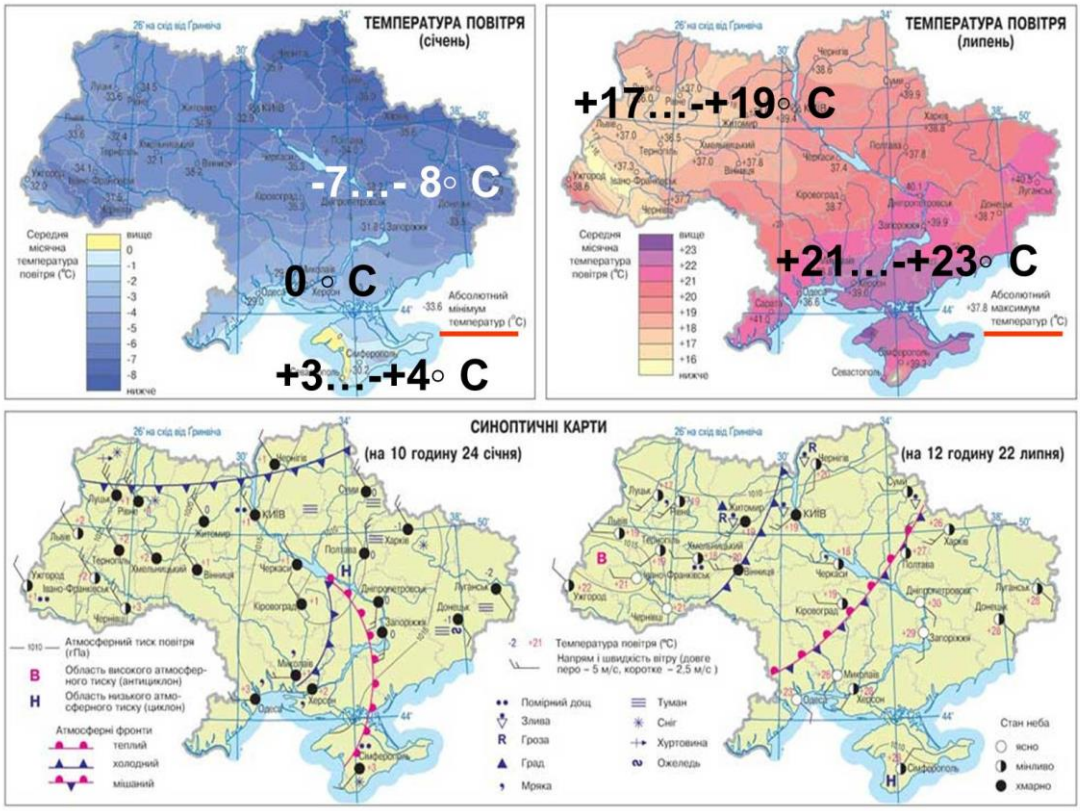

Рис.1.2. Температурні та синоптичні особливості території України.

Середньорічна кількість опадів становить від 350 мм на півдні до 500 мм на півночі. Більшість опадів випадає в літні місяці, часто бувають зливи. У південній частині часто бувають пилові бурі та суховії. Чорноземи (займають приблизно 90 % площі зони) переважають у верхньому шарі ґрунтів. Темні каштанові чорноземи є типовими для південної частини. Зона Степу є найбільш розораною в Україні, і приблизно 48 % ріллі країни знаходиться тут.

А тепер звернімо увагу на особливості кліматичних умов, представлені у зручному табличному форматі (Таблиця 1.1).

Таблиця1.1

Oblast	Odesa		Mykolaiv		Kherson		Zaporizhzhia		Dnipropetrovsk		Donetsk	
Metric	AvgTemp_°C	Precip_mm	AvgTemp_°C	Precip_mm	AvgTemp_°C	Precip_mm	AvgTemp_°C	Precip_mm	AvgTemp_°C	Precip_mm	AvgTemp_°C	Precip_mm
Jan	-0.9	38	-1.7	37	-0.9	36	-2.6	44	-3.4	45	-3.7	49
Feb	0.5	31	-0.3	31	0.4	31	-1.6	35	-2.4	35	-2.8	39
Mar	4.8	34	4.2	36	4.7	36	3.4	45	2.7	47	2.2	48
Apr	10.8	34	10.9	38	11.2	41	10.5	43	10.3	47	9.6	50
May	17.4	38	17.4	46	17.7	49	17.1	45	16.9	53	16.2	60
Jun	21.9	46	21.7	56	22.1	54	21.3	51	21	59	20.5	68
Jul	24.5	35	24.3	37	24.9	34	23.8	37	23.3	48	23	50
Aug	24.2	38	24.1	40	24.7	37	23.7	32	23	40	22.8	36
Sep	18.3	40	18	44	18.5	43	17.4	46	16.8	50	16.4	45
Oct	11.7	38	11.1	32	11.5	30	10.3	37	9.7	39	9.2	43
Nov	6.5	35	5.5	36	6	36	4.3	43	3.7	41	2.9	42
Dec	1.2	38	1	37	1.7	37	0.1	47	-0.6	43	-1.3	49

Аналіз наведених даних виявляє певні закономірності та відмінності між областями. Так, температурний режим протягом року є очікуваним: найхолоднішим місяцем є січень, з середніми температурами, що коливаються від -0.9°C в Одеській та Херсонській областях до -3.7°C у Донецькій. Найтеплішим місяцем є липень, де середні температури сягають максимуму в Херсонській області (24.9°C) та дещо нижчі в Донецькій (23.0°C). Щодо опадів, їх місячна кількість демонструє відносну стабільність без різких піків чи спадів для більшості областей. Найбільша кількість опадів у січні зафіксована в Донецькій області (49 мм), а найменша - у Херсонській (36 мм). У липні діапазон опадів становить від 32 мм у Запорізькій області до 50 мм у Донецькій. Порівнюючи області, можна відзначити, що Донецька область характеризується дещо нижчими зимовими та літніми температурами, а також вищою кількістю опадів у січні. В той же час, Херсонська область є найтеплішою в липні та має одну з найнижчих кількостей опадів у січні. Одеська та Миколаївська області демонструють схожі температурні та опадові характеристики.

Для більш детального розгляду кліматичних особливостей однієї з областей (рис.1.3), а саме Дніпропетровської, пропонуємо ознайомитися з кліматограмою.

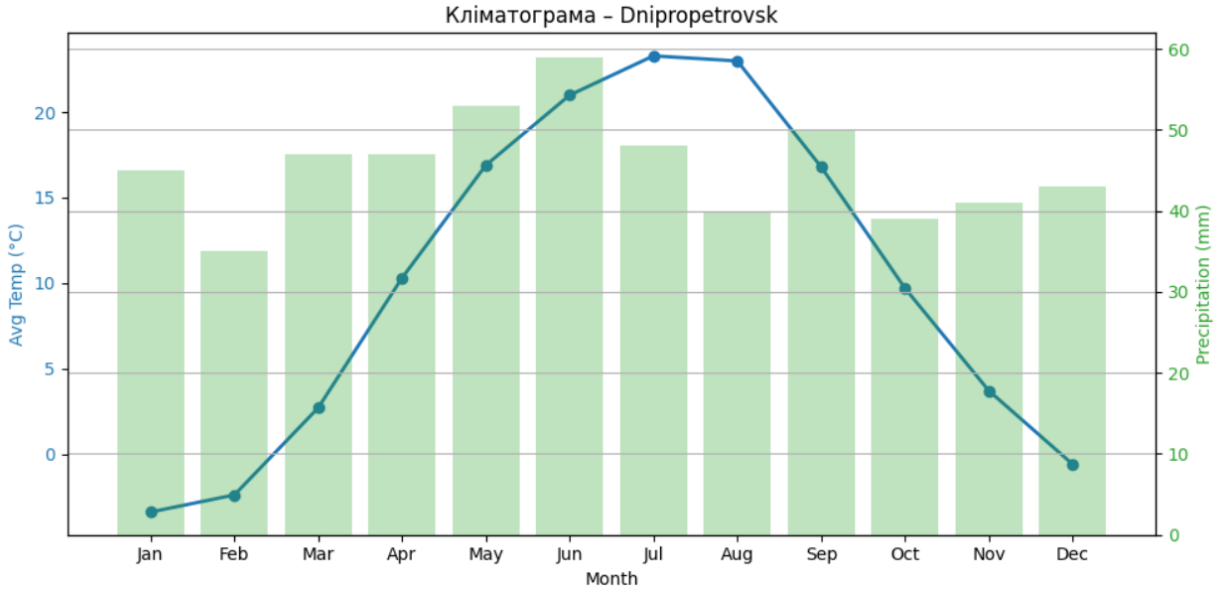

Рис.1.3. Кліматограма Дніпропетровської області

Графік наочно демонструє річний хід середньої температури та кількості опадів у Дніпропетровській області. Як бачимо, температурна крива має типову амплітуду для помірного клімату: найнижчі середні температури спостерігаються в січні (-3.4°C), а пік тепла припадає на липень та серпень (23.3°C та 23.0°C відповідно). Кількість опадів протягом року є відносно рівномірною, проте спостерігається деяке збільшення в літні місяці, особливо в червні (59 мм) та липні (48 мм). Найменша кількість опадів фіксується в зимові місяці, зокрема в лютому (35 мм) та грудні (39 мм). Загалом, клімат Дніпропетровської області характеризується помірно холодною зимою та теплим літом з достатньою кількістю опадів для сільськогосподарської діяльності.

Сільське господарство спеціалізується на вирощуванні зернових, фруктів і овочів, а також на виноградарстві. Основними зерновими є озима пшениця, кукурудза, ячмінь і технічні культури, соняшник. Широко поширене

овочівництво, особливо в приміських зонах великих міст, наприклад, Донецька.

Як відомо, врожайність будь-яких сільськогосподарських культур суттєво залежить від того на яких грунтах вони вирощуються.

Грунт є найдорожчим багатством людства. Чорноземні ґрунти, найродючіші в світі, вкривають дві третини території України. Сільськогосподарські угіддя тут становлять 71%, а орні 56% усієї площі країни. Земельні ресурси не належать до категорії невичерпних. Через великий розвиток промисловості, зростання міст площа сільськогосподарських угідь зменшується за рахунок відведення земельних ділянок для промислового і житлового будівництва, гірничих розробок тощо.

Як відомо, Україна володіє третиною світового запасу найпродуктивніших земель – чорноземів (рис.1.4.-1.5.). Завдяки цьому, а також зручному, здебільшого рівнинному рельєфу, більше 60% площі нашої країни зайнято сільськогосподарськими угіддями. Однак, крім чорноземів, зустрічаються і деякі інші типи грунтів.

Рис.1.4. Карта агрокліматичного районування України

Рис.1.5. Карта родючості грунтів України

Сільське господарство традиційно відіграє ключову роль в економіці України, зокрема виробництво зернових культур. Щороку аграрний сектор забезпечує значну частину як внутрішнього споживання, так і експорту. Попри складні погодні умови та економічні виклики, українські аграрії продовжили активно працювати на полях.

Згідно з останніми даними Мінагрополітики, Україна з 2018 -2022 рік зібрала 50,9 млн. т. зернових культур з площі 10,9 млн га (94%). У розрізі культур статистика виглядає наступним чином:

- врожай пшениці становить 20,2 млн т, з площі 5 млн га (100%), середня врожайність — 4,05 т/га;
- ячменю — 5,8 млн т, з площі 1,7 млн га (100%), середня врожайність — 3,47 т/га;
- гороху — 269 тис. т, з площі 118 тис. га, середня врожайність — 2,28 т/га;

- гречки — 158,5 тис. т, з площі 116 тис. га (98%), середня врожайність — 1,3 т/га;
- проса — 101,8 тис. т, з площі 44,7 тис. га (99%), середня врожайність — 2,2 т/га;
- кукурудзи на зерно — 23,5 млн т, з площі — 3,6 млн га (85%), середня врожайність — 6,5 т/га.

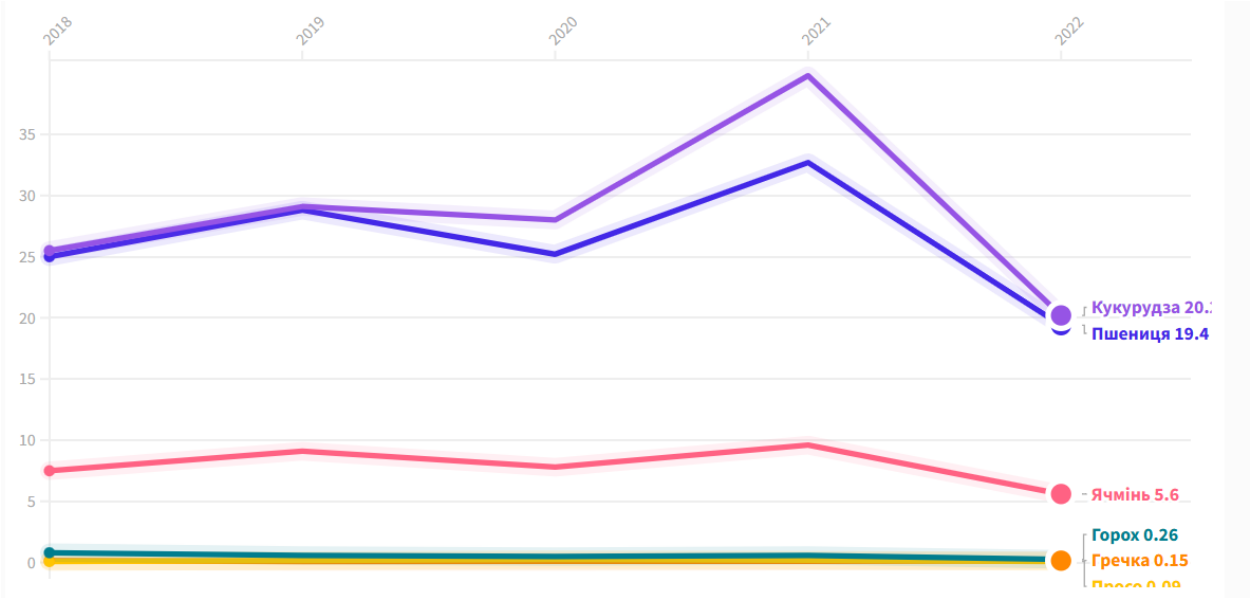

Рис.1.6. Врожай зернових культур в України в 2018-2022 рр.

Глянувши на статистику бачимо, що за 5 років Україна зібрала найменшу кількість зернових. І на це є кілька причин:

- широкомасштабна війна, яка вплинула на посівні та збиральні площі;
- погодний фактор: дощове літо та осінь;
- деякі фермери не мали змоги провести повноцінне внесення добрив;
- нестача елеваторів, що змусило частину фермерів залишити кукурудзу зимувати в полі.

Якщо брати до уваги решту сезонів, то помітне просідання врожайності кукурудзи, а також знизився і середній показник врожайності[3].

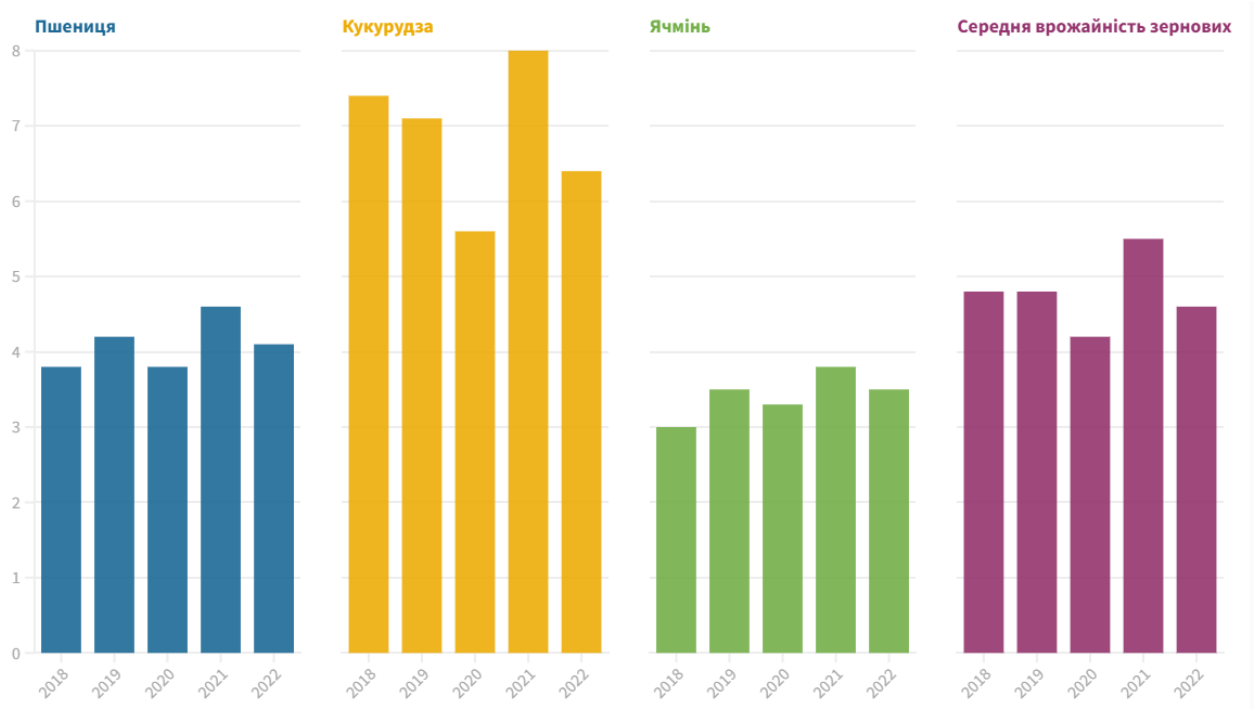

Рис.1.7. Середня врожайність зернових 2018-2022 р.

Тому приходимо висновку, що фактори, які ми перерахували вище вплинули на зниження врожайності культур й відповідно на загальну кількість зібраного зерна.

Для розуміння тенденцій в аграрному секторі важливим є аналіз динаміки врожайності ключових культур. Наведений графік (Рис.1.8) ілюструє зміни в урожайності пшениці протягом періоду з 2000 по 2022 рік.

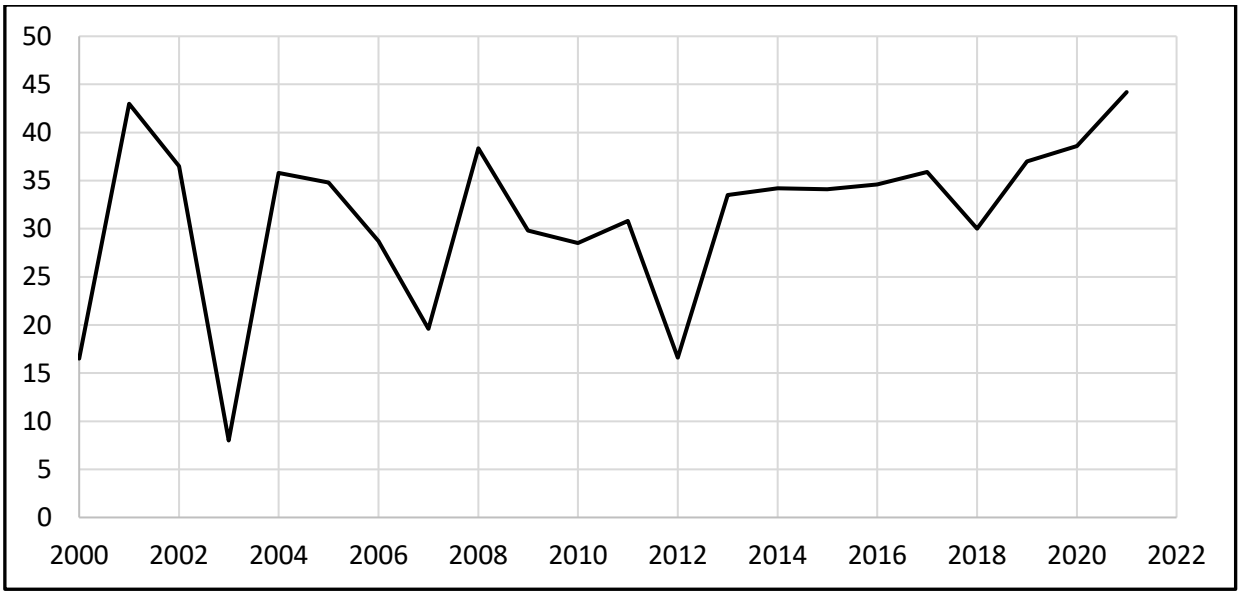

Рис.1.8. Динаміка врожайності пшениці у Дніпропетровській області

Після піку в 2001 році спостерігається різке падіння у 2003 році. Подальші роки відзначаються нестабільністю, зі значним зростанням у 2008 році та наступним спадом до 2012 року. Однак, після 2012 року врожайність стабілізується з поступовим зростанням, досягаючи максимуму у 2022 році. Така мінливість, ймовірно, зумовлена поєднанням кліматичних умов та агротехнічних факторів. Тому приходимо висновку, що фактори, які ми перерахували вище, впливають на велику мінливість врожайності культур, зокрема і пшениці в Дніпропетровській області.

РОЗДІЛ 2. Методи та інструменти машинного навчання для аналізу даних

Машинне навчання. Машинне навчання (Machine Learning, ML) – розділ штучного інтелекту, присвячений розумінню та розробленню методів, що дозволяють машинам «навчатися», тобто методів, у яких використовують дані для покращення продуктивності комп'ютера в певному наборі завдань. Алгоритми машинного навчання створюють модель на основі вибіркових даних, відомих як навчальні дані, щоб робити прогнози чи ухвалювати рішення без явного програмування для цього. Алгоритми машинного навчання використовують у широкому спектрі завдань, таких як медицина, фільтрація електронної пошти, розпізнавання мовлення, сільське господарство, виявлення шахрайства, виявлення загроз зловмисного програмного забезпечення, автоматизація бізнес-процесів, комп'ютерний зір, проєктування безпілотних апаратів тощо, де важко або неможливо розробити звичайні алгоритми для виконання необхідних завдань. Машинне навчання, яке застосовують у бізнес-проблемах, також називають прогнозною аналітикою.

Машинне навчання дає компаніям уявлення про тенденції поведінки клієнтів і бізнес-операційні моделі, а також підтримує розроблення нових продуктів. Багато провідних сучасних компаній, таких як Facebook, Google і Viber, роблять машинне навчання центральною частиною своєї діяльності. Машинне навчання стало значним конкурентоспроможним фактором для багатьох компаній. Воно тісно пов'язане з обчислювальною статистикою, зосереджуваною на прогнозуванні за допомогою комп'ютерів, але не все машинне навчання є статистичним навчанням. Вивчення математичної оптимізації надає методи, теорію та сфери застосування в галузі машинного навчання. Інтелектуальний аналіз даних є спорідненою галуззю дослідження, що базується на дослідницькому аналізуванні даних за допомогою навчання[4].

Основною мовою програмування для більшості фахівців у галузі машинного навчання є Python, що вирізняється своєю простотою та наявністю великої кількості потужних бібліотек для обробки та аналізу даних. Для ефективного застосування методів машинного навчання на практиці необхідні спеціалізовані інструменти, які спрощують роботу з даними, реалізацію алгоритмів та візуалізацію результатів. Саме тому знання та вміння використовувати відповідні бібліотеки Python є ключовим для фахівця в цій галузі. Сьогодні існує велика кількість бібліотек для машинного навчання. Для полегшення вибору ми розглянемо лише найпопулярніші та найнеобхідніші бібліотеки, які охоплюють базові потреби для початку роботи з Machine Learning та Deep Learning.

NumPy:

Основний функціонал NumPy полягає в підтриманні багатовимірних масивів даних та швидких алгоритмів лінійної алгебри. Саме тому NumPy – ключовий компонент Scikit-learn, SciPy та Pandas. Зазвичай NumPy використовують як допоміжну бібліотеку для виконання різних математичних операцій із структурами даних Pandas, тому варто вивчити її базові можливості.

Pandas:

Аналізування та підготовка даних найчастіше займають більшу частину часу при виконанні завдань машинного навчання. Дані можуть бути одержані в CSV, JSON, Excel або іншому структурованому (або не дуже) форматі, і виникає необхідність обробити їх, для того щоб застосовувати в моделях машинного навчання. Для цього використовують бібліотеку Pandas. Це потужний інструмент, що дозволяє швидко аналізувати, модифікувати та готувати дані для подальшого використання в інших бібліотеках, таких як Scikitlearn, TensorFlow або PyTorch. У Pandas можна завантажувати дані з різних джерел: SQL-баз, CSV-, Excel-, JSON-файлів та інших менш популярних форматів. Коли дані завантажені в пам’ять, з ними можна

виконувати безліч різних операцій для аналізування, трансформації, заповнення відсутніх значень та очищення набору 20 даних. Pandas дозволяє виконувати безліч SQL-подібних операцій над наборами даних: об’єднання, групування, агрегування тощо. Також вона надає вбудований набір популярних статистичних функцій для базового аналізування.

Matplotlib і Seaborn:

Побудова графіків та візуалізація Matplotlib – це стандартний інструмент у наборі дата інженера. Він дозволяє створювати різноманітні графіки та діаграми для візуалізації одержаних результатів. Графіки, створені Matplotlib, легко інтегруються в Google Colab. Це дозволяє візуалізувати дані та результати, одержані під час оброблення моделей. Для цієї бібліотеки створено багато додаткових пакетів. Один із найбільш популярних – Seaborn. Його основна перевага – готовий набір найчастіше використовуваних статистичних діаграм і графіків [4].

Модель лінійної регресії. Окрім візуалізації даних, критично важливим є розуміння взаємозв'язків між різними змінними. Саме тут на допомогу приходить регресійний аналіз (Рис. 2.1). Регресія є цінним інструментом для аналітиків та фінансистів, оскільки дозволяє виявляти залежності між різними факторами процесу. Наприклад, за допомогою регресійного аналізу агроном може дослідити, як кількість опадів та середня температура протягом вегетаційного періоду впливають на врожайність пшениці. Зібрані дані за кілька років можуть показати, що певна кількість опадів сприяє кращому врожаю, а надмірна спека може мати негативний вплив.

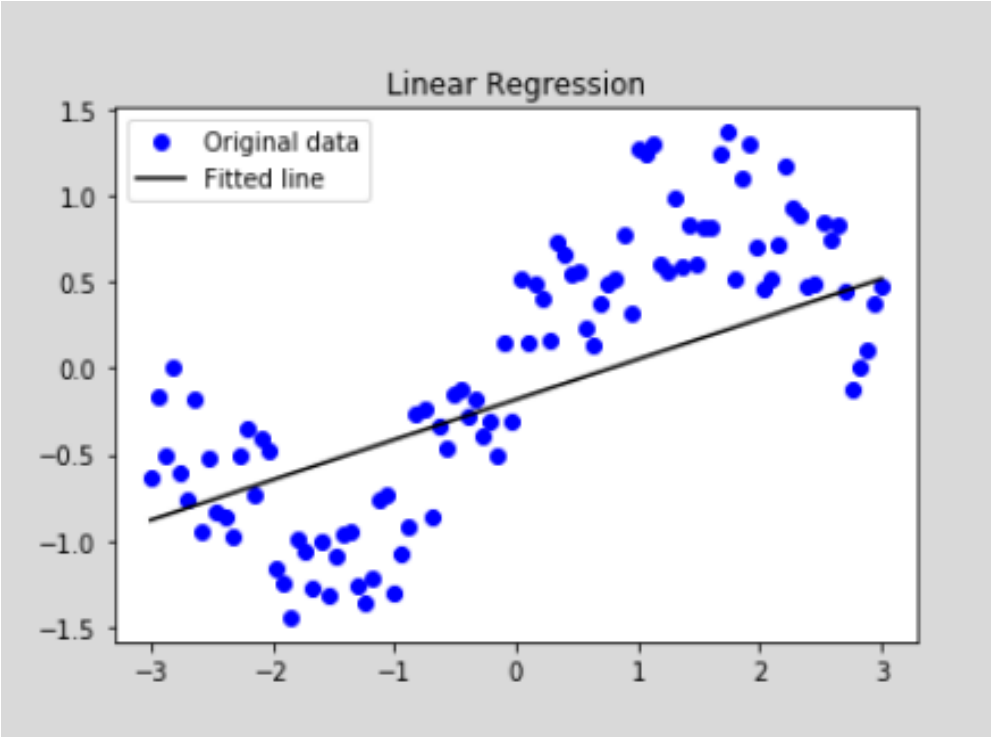

Рис.2.1. Лінійна регресія

Регресія показує або передбачає взаємозв'язок між процесом та його чинником. Машина намагається побудувати криву на графіку, яка відображає залежність. Проте, на відміну від людини з крейдою та дошкою, вона це робить із застосуванням математики. Моделі регресійного аналізу зазвичай використовують, щоб показати або передбачити взаємозв'язок між процесом і тим, що цей процес може спровокувати. Тут варто пам'ятати, що така кореляція – не завжди причинність, тобто навіть пряма в простій лінійній регресії, яка добре відображає залежності між даними, може не дати конкретної відповіді про причинно-наслідковий зв'язок. Саме тому регресійний аналіз не використовують для інтерпретації причинно-наслідкових зв'язків між змінними. Однак такий аналіз може засвідчити проте, як і наскільки змінні пов'язані одна з одною, а визначення причин та наслідків – це предмет глибших досліджень за допомогою інших алгоритмів і методів. Графік регресії може показати позитивний зв'язок, негативний зв'язок або відсутність зв'язку процесу з тими чи іншими факторами. Якщо зі зростанням x зростає y (нижня частина графіка перетинає вісь, а верхня прагне в поле

графіка) – залежність позитивна, тобто значення буде зростати, якщо навпаки – негативна, то значення будуть зменшуватись [4].

OLS Regression Results

Dep. Variable:	eps	R-squared:	0.488
Model:	OLS	Adj. R-squared:	0.455
Method:	Least Squares	F-statistic:	14.65
Date:	Wed, 26 Mar 2025	Prob (F-statistic):	6.97e-15
Time:	12:53:39	Log-Likelihood:	-403.58
No. Observations:	132	AIC:	825.2
Df Residuals:	123	BIC:	851.1
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	15.1064	6.535	2.312	0.022	2.171	28.042
t2	-0.7124	0.310	-2.298	0.023	-1.326	-0.099
t3	0.9927	0.300	3.304	0.001	0.398	1.587
t5	-1.5266	0.286	-5.335	0.000	-2.093	-0.960
t6	0.7813	0.275	2.842	0.005	0.237	1.326
t7	-1.1954	0.278	-4.292	0.000	-1.747	-0.644
t9	0.4352	0.186	2.343	0.021	0.068	0.803
R10	0.0845	0.019	4.377	0.000	0.046	0.123
R20	0.0454	0.016	2.886	0.005	0.014	0.077

Omnibus:	5.704	Durbin-Watson:	2.080
Prob(Omnibus):	0.058	Jarque-Bera (JB):	5.586
Skew:	-0.504	Prob(JB):	0.0612
Kurtosis:	3.020	Cond. No.	1.07e+03

Рис.2.2. Модель лінійної регресії

Після того, як модель побудована, необхідно ретельно оцінити її якість та статистичну значущість отриманих результатів. Результат регресії, такий як OLS Regression Results, надає детальну інформацію про модель та її параметри, дозволяючи зробити висновки щодо її адекватності та здатності до прогнозування (Рис.2.2).

Розглянемо детальніше значення кожної з наведених характеристик, щоб краще зрозуміти представлену регресійну модель:

Dep. Variable: eps - це залежна змінна (цільова змінна), яку ми намагаємося передбачити за допомогою нашої регресійної моделі. У цьому випадку це змінна "eps".

Model: OLS - вказує на те, що для побудови моделі використовувався метод звичайних найменших квадратів (Ordinary Least Squares). Це

стандартний метод для лінійної регресії, який мінімізує суму квадратів різниць між спостережуваними та прогнозованими значеннями.

Method: Least Squares - повторює інформацію про використаний метод оцінки параметрів моделі.

Date: Wed, 26 Mar 2025 та Time: 12:53:39 - фіксують дату та час проведення регресійного аналізу.

No. Observations: 132 - загальна кількість спостережень (точок даних), використаних для навчання моделі.

Df Residuals: 123 - кількість ступенів вільності залишків. Розраховується як різниця між кількістю спостережень та кількістю оцінюваних параметрів (включаючи константу).

Df Model: 8 - кількість ступенів вільності моделі, що відповідає кількості незалежних змінних (предикторів) у моделі (t2, t3, t5, t6, t7, t9, R10, R20).

Covariance Type: nonrobust - вказує на спосіб обчислення коваріаційної матриці параметрів. "Nonrobust" означає, що використовуються стандартні припущення про гомоскедастичність (сталість дисперсії залишків).

R-squared: 0.488 - коефіцієнт детермінації (R2), який показує частку дисперсії залежної змінної, що пояснюється регресійною моделлю. Значення 0.488 означає, що приблизно 48.8% варіації "eps" може бути пояснено включеними до моделі предикторами. Чим ближче R2 до 1, тим краще модель відповідає даним.

Adj. R-squared: 0.455 - скоригований коефіцієнт детермінації, який враховує кількість предикторів у моделі. Він є більш консервативною оцінкою порівняно з R2, особливо при великій кількості незалежних змінних, оскільки коригує R2 на кількість предикторів.

F-statistic: 14.65 - F-статистика, яка використовується для перевірки гіпотези про значущість моделі в цілому. Вона оцінює, чи є хоча б одна з незалежних змінних статистично значущою для прогнозування залежної змінної.

Prob (F-statistic): 6.97e-15 - p-значення, пов'язане з F-статистикою. Дуже мале значення (менше за стандартний рівень значущості 0.05) свідчить про те, що модель в цілому є статистично значущою, тобто хоча б одна з незалежних змінних має значущий вплив на залежну змінну.

Log-Likelihood: -403.58 - значення логарифмічної функції правдоподібності. Це міра того, наскільки добре модель відповідає даним. Вищі значення вказують на кращу відповідність.

AIC: 825.2 - інформаційний критерій Акаіке (Akaike Information Criterion). Це критерій вибору моделі, який враховує як якість підгонки моделі до даних, так і кількість параметрів у моделі. Моделі з меншим значенням AIC вважаються кращими.

BIC: 851.1 - Баєсівський інформаційний критерій (Bayesian Information Criterion) або критерій Шварца. Подібно до AIC, BIC використовується для вибору моделі, але він сильніше штрафує за складність моделі (більшу кількість параметрів). Моделі з меншим значенням BIC вважаються кращими.

coef - коефіцієнти регресії для кожної незалежної змінної та константи (intercept). Ці значення показують зміну залежної змінної при збільшенні відповідної незалежної змінної на одиницю, за умови, що інші змінні залишаються незмінними.

std err - стандартна похибка коефіцієнта, яка є мірою точності оцінки коефіцієнта. Менші стандартні похибки вказують на більш точні оцінки.

t - t-статистика, яка використовується для перевірки гіпотези про те, що коефіцієнт дорівнює нулю. Розраховується як відношення коефіцієнта до його стандартної похибки (t=stderrcoef).

P>|t| - p-значення, пов'язане з t-статистикою. Воно показує ймовірність отримання такого ж або більш екстремального значення t-статистики, якщо нульова гіпотеза (коефіцієнт дорівнює нулю) є істинною. Мале p-значення (зазвичай менше 0.05) свідчить про те, що відповідний коефіцієнт є статистично значущим.

Omnibus: 5.704 - статистика Omnibus, яка використовується для перевірки нормальності розподілу залишків.

Prob(Omnibus): 0.058 - p-значення, пов'язане зі статистикою Omnibus. Значення близьке до 0.05 може викликати занепокоєння щодо нормальності залишків.

Durbin-Watson: 2.080 - статистика Дурбіна-Уотсона, яка використовується для перевірки наявності автокореляції першого порядку в залишках. Значення близько 2 свідчить про відсутність значної автокореляції.

Jarque-Bera (JB): 5.586 - статистика Жарка-Бера, ще один тест на нормальність розподілу залишків.

Prob(JB): 0.0612 - p-значення, пов'язане зі статистикою Жарка-Бера. Подібно до тесту Omnibus, значення близьке до 0.05 може вказувати на відхилення від нормальності.

Skew: -0.504 - коефіцієнт асиметрії, який показує ступінь асиметричності розподілу залишків.

Kurtosis: 3.020 - коефіцієнт ексцесу, який показує "гостроту" піку та "товщину" хвостів розподілу залишків порівняно з нормальним розподілом (для якого ексцес дорівнює 3).

Cond. No.: 1.07e+03 - число обумовленості (Condition Number), яке вказує на мультиколінеарність (сильну кореляцію між незалежними змінними). Високе число обумовленості може свідчити про проблеми з мультиколінеарністю, що може впливати на стабільність оцінок коефіцієнтів.

Для того, щоб переконатися в надійності та узагальнюючій здатності побудованої регресійної моделі, часто застосовують крос-валідацію. Цей метод дозволяє оцінити, наскільки добре модель буде працювати на нових, невідомих даних. Загальна ідея крос-валідації полягає у розділенні наявного набору даних на кілька підмножин, які по черзі використовуються для навчання та оцінки моделі. Одним із найпоширеніших підходів є K-блокова крос-валідація, де дані діляться на K рівних частин. Потім модель тренується

на K-1 частинах і оцінюється на залишеній частині, і цей процес повторюється K разів. Отримані оцінки продуктивності усереднюються, надаючи більш стабільну та об'єктивну картину якості моделі. Крос-валідація допомагає виявити потенційне перенавчання, коли модель занадто добре підлаштовується під тренувальні дані, але погано працює на нових. Застосування крос-валідації є важливим кроком у побудові надійних та практичних моделей машинного навчання, включаючи й регресійні моделі.

Отже, застосування крос-валідації є фундаментальним кроком для підтвердження здатності регресійної моделі до узагальнення на невідомих даних, мінімізуючи ризик перенавчання. Детальний аналіз результатів OLS-регресії, відображених вище, розкриває комплексну якість моделі в поясненні варіації залежної змінної. Інтерпретуючи ці результати, можна оцінити, які незалежні змінні мають значущий вплив на залежну змінну, наскільки добре модель апроксимує дані (через R2 та скоригований R2), та загальну статистичну значущість моделі (через F-статистику). Ці показники є ключовими для визначення практичної цінності та прогностичної сили побудованої регресійної моделі. Ретельна інтерпретація цих показників є вирішальною для оцінки практичної цінності та прогностичної сили побудованої регресійної моделі.

РОЗДІЛ 3. Регресійна модель прогнозування врожайності пшениці

3.1. Побудова моделі лінійної регресії

У цьому розділі представлено процес аналізу даних та розробки математичної моделі для прогнозування врожайності зернових культур в степовій зоні України. Метою є оцінка впливу кліматичних факторів на врожайність пшениці. Для досягнення цієї мети будуть застосовані методи лінійної регресії та крос-валідації.

Методика досліджень. Процес збору даних є критично важливим етапом будь-якого дослідження, оскільки якість та повнота даних безпосередньо впливають на достовірність отриманих результатів і надійність побудованих моделей. У даному дослідженні використовуються дані з файлу «stepa.csv».

Дані представлені у вигляді таблиці, що складається з записів (кількість рядків) та змінних (кількість стовпців). Файл містить інформацію щодо кліматичних характеристик та трендових залишків врожайності пшениці в шести областях степової зони України (Херсонська, Миколаївська, Одеська, Запорізька, Дніпропетровська, Кіровоградська). за період 2000 – 2021 роки. Кліматичні характеристики стосуються трьох місяців, на протязі яких вегетаційний процес пшениці є найбільш активним: квітень, травень, червень. Використовуються наступні позначення:

t1 – середня температура першої декади квітня (t°C);

t2 – середня температура другої декади квітня;

t3 – середня температура третьої декади квітня;

t4 – середня температура першої декади травня;

t5 – середня температура другої декади травня;

t6 – середня температура третьої декади травня;

t7 – середня температура першої декади червня;

t8 – середня температура другої декади червня;

t9 – середня температура третьої декади червня;

R10 – сума опадів у квітні (мм);

R20 – сума опадів у травні;

R30 - сума опадів у червні;

eps – трендове відхилення врожайності пшениці (ц/га).

Важливою змінною є 'eps', яка представляє собою трендове відхилення врожайності пшениці, виміряне в центнерах з гектара (ц/га). Трендове відхилення відображає, наскільки врожайність конкретного року була вищою або нижчою за очікувану, враховуючи загальний напрямок зміни врожайності протягом багатьох років. Ця тенденція може бути зумовлена різними факторами, такими як покращення технологій або зміни клімату.

Таким чином, аналізуючи 'eps' разом із кліматичними даними, можна оцінити, як саме погодні умови у квітні, травні та червні впливають на відхилення врожайності пшениці від її загальної тенденції [8].

Лінійна регресійна модель.

Наші дослідження мають на меті дослідити вплив кліматичних факторів на коливання врожайності пшениці. Для проведення досліджень ми використовуємо реальні статистичні дані за період 2000 – 2021 роки для областей степової зони України. Для порівняння різних моделей ми проводили числові експерименти з використанням середовища програмування Python та бібліотек numpy, pandas, scipy, statsmodels. Для побудови моделі регресії ми використовуємо інструмент sm.OLS(y, X).

Спочатку будується модель лінійної регресії на всіх факторах. Вона має вигляд :

$$eps = \beta_0 + \beta_1 t_1 + \cdots + \beta_9 t_9 + \beta_{10} R_{10} + \beta_{20} R_{20} + \beta_{30} R_{30} + \varepsilon \,. \quad (3.1)$$

Тут eps – це відхилення врожайності від тренду. Це відхилення визначається впливом кліматичних факторів. ε – це неусувна похибка моделі регресії, яка обумовлена випадковим характером впливаючих факторів.

Не всі фактори моделі є однаково значущими. Для оцінки значущості факторів, які входять у модель лінійної регресії потрібно переглянути

статистичну значущість кожного коефіцієнта. Одним із найбільш поширених способів оцінки значущості факторів є перегляд стандартних помилок коефіцієнтів лінійної регресії (стандартних помилок оцінок). Чим менше стандартна помилка, тим більш значущий вважається коефіцієнт. Саме такий метод реалізований у даному дослідженні. Для реалізації цього методу використовують бібліотеку statsmodels:

```
import statsmodels.api as sm
```

Для отримання інформації про якість отриманої моделі використовують команду print(model.summary()). Ця команда надасть детальну інформацію про кожен фактор, включаючи його стандартну помилку та p-value. Фактори з найменшими p-value вважаються найбільш значущими. Ці відомості можна використати для вибору найбільш значущих факторів моделі. Частина факторів є незначущими і знижують якість моделі. Їх потрібно видалити з моделі. Це дещо знижує коефіцієнт детермінації , але спрощує модель і збільшує її статистичну значущість. Відбір факторів слід продовжувати до тих пір, поки для всіх факторів моделі не виконається умова . Таким чином, ми порівнюємо два варіанти кожної моделі : модель з усіма факторами і модель з лише значушими факторами.

Квадратична регресійна модель.

На другому етапі наших досліджень ми включаємо у розгляд квадрати кліматичних факторів. Це дозволяє виявити нелінійний вплив кліматичних факторів на врожайність пшениці. Цей вплив буде особливо помітним при наближенні температури повітря до границі екологічної зони комфорту (10^{O}C $\leq$ t $\leq 30^{O}$C). Модель квадратичної регресії має наступний вигляд :

$$eps = \beta_0 + \beta_1 t_1 + \cdots + \beta_9 t_9 + \beta_{10} R_{10} + \beta_{20} R_{20} + \beta_{30} R_{30} + \beta_{11} t_1^2 + \cdots + \beta_{99} t_9^2 + \beta_{100} R_{10}^2 + \beta_{200} R_{20}^2 + \beta_{300} R_{30}^2 + \varepsilon \, . \quad (3.2)$$

Ми будуємо і досліджуємо два варіанти квадратичної моделі : модель з усіма факторами і модель лише із значущими факторами.

Загальна нелінійна регресійна модель.

На третьому етапі наших досліджень ми додаємо до моделі добутки сусідніх за часом кліматичних факторів. Це дозволяє врахувати пролонговану дію деяких факторів. Наприклад, коли засуха триває два місяці підряд, це дуже сильно впливає на зменшення врожайності. Загальна модель нелінійної регресії включає лінійні кліматичні фактори, квадрати кліматичних факторів і добутки сусідніх кліматичних факторів. Вона має наступний вигляд :

$$eps = \beta_0 + \beta_1 t_1 + \cdots + \beta_9 t_9 + \beta_{10} R_{10} + \beta_{20} R_{20} + \beta_{30} R_{30} + \beta_{11} t_1^2 + \cdots + \beta_{99} t_9^2 + \beta_{100} R_{10}^2 + \beta_{200} R_{20}^2 + \beta_{300} R_{30}^2 + \beta_{12} t_1 t_2 + \beta_{23} t_2 t_3 + \cdots + \beta_{89} t_8 t_9 + \beta_{1020} R_{10} R_{20} + \beta_{2030} R_{20} R_{30} + \varepsilon. \quad (3.3)$$

Ми будуємо і досліджуємо два варіанти загальної нелінійної моделі: модель з усіма факторами і модель лише із значущими факторами.

Побудова і навчання моделей. Переходимо до програмної реалізації описаних вище моделей. Для цього використовуємо програмне середовище `Google Colab Python` та бібліотеки `numpy, pandas, matplotlib, scikit learn, statsmodels`. Спочатку завантажуємо дані (Рис.3.1) із датасет-файла в хмарне середовище Google Colab. Для контролю виводимо перші 5 рядків даних за допомогою команди print(XY.head()).

```
Вибрати файли  Файл не вибрано    Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.
Saving stepa.csv to stepa (1).csv
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 132 entries, 0 to 131
Data columns (total 13 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   t1      132 non-null    float64
 1   t2      132 non-null    float64
 2   t3      132 non-null    float64
 3   t4      132 non-null    float64
 4   t5      132 non-null    float64
 5   t6      132 non-null    float64
 6   t7      132 non-null    float64
 7   t8      132 non-null    float64
 8   t9      132 non-null    float64
 9   R10     132 non-null    float64
 10  R20     132 non-null    float64
 11  R30     132 non-null    float64
 12  eps     132 non-null    float64
dtypes: float64(13)
memory usage: 13.5 KB
```

Рис.3.1. Відображення структури датасету

Виконуємо стандартний статистичний аналіз даних (рис. 3.2)

	t1	t2	t3	t4	t5	t6	t7	t8	t9	R10	R20	R30	eps
count	132.00	132.00	132.00	132.00	132.00	132.00	132.00	132.00	132.00	132.00	132.00	132.00	132.00
mean	8.59	11.03	12.94	15.27	16.91	19.24	19.90	21.52	22.20	30.81	48.73	63.37	-0.00
std	2.21	2.17	2.38	3.03	2.62	2.90	2.49	2.42	2.69	25.19	32.92	44.18	7.22
min	2.54	6.65	8.51	10.49	11.50	12.82	14.80	16.86	16.79	0.00	0.30	8.00	-21.02
25%	7.15	9.35	11.41	13.17	15.15	17.55	17.82	19.68	19.88	9.75	25.27	34.83	-3.03
50%	8.82	10.79	12.34	14.31	16.40	19.05	20.26	21.33	22.73	24.50	40.90	53.00	0.96
75%	9.88	12.40	14.50	16.83	18.76	20.61	21.81	23.33	24.29	46.00	68.97	78.95	5.05
max	15.20	16.95	21.30	24.90	24.50	28.68	25.95	28.20	29.05	102.00	156.00	329.00	16.64

Рис.3.2. Статистичний аналіз даних

Для програмної реалізації моделі лінійної ренгресії ми використовуємо бібліотеку statsmodels: import statsmodels.api as sm. Спочатку ми готуємо дані для регресійного аналізу, розділяючи змінні на впливаючі фактори (незалежні змінні) та змінну відгуку (залежна змінна) (Рис.3.3.).

	const	t1	t2	t3	t4	t5	t6	t7	t8	t9	R10	R20	R30
0	1.00	9.23	13.26	16.47	13.34	14.74	19.49	21.20	19.15	18.51	39.20	33.00	104.20
1	1.00	9.74	10.60	13.78	15.90	12.83	14.68	15.98	20.54	19.42	44.90	38.30	79.10
2	1.00	5.63	11.76	12.54	15.36	17.24	19.06	16.55	21.18	23.99	8.20	6.80	86.50
3	1.00	4.43	10.32	10.49	16.35	20.93	20.99	20.18	21.35	19.46	14.90	54.90	51.70
4	1.00	6.55	11.62	12.38	15.17	13.69	15.59	16.48	18.94	20.50	14.30	97.20	53.70

Рис.3.3. Розділення змінних на впливаючі фактори та змінну-відгук

Матриця впливаючих факторів X0 (включаючи константу) готова до використання як предиктори, а вектор залежної змінної y (який не відображається на цьому зображенні, але був визначений раніше) є цільовою змінною, яку ми намагатимемося пояснити або спрогнозувати за допомогою факторів у X0.

Наступним кроком буде побудова лінійної регресії з усіма її факторами (Рис.3.4.).

OLS Regression Results

Dep. Variable:	eps	R-squared:	0.513
Model:	OLS	Adj. R-squared:	0.464
Method:	Least Squares	F-statistic:	10.46
Date:	Wed, 26 Mar 2025	Prob (F-statistic):	7.44e-14
Time:	12:50:40	Log-Likelihood:	-400.20
No. Observations:	132	AIC:	826.4
Df Residuals:	119	BIC:	863.9
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	17.1230	7.664	2.234	0.027	1.948	32.298
t1	0.2156	0.326	0.662	0.509	-0.429	0.860
t2	-0.6916	0.326	-2.123	0.036	-1.337	-0.047
t3	1.1495	0.336	3.421	0.001	0.484	1.815
t4	-0.4280	0.229	-1.869	0.064	-0.881	0.025
t5	-1.2432	0.388	-3.204	0.002	-2.011	-0.475
t6	0.5951	0.302	1.971	0.051	-0.003	1.193
t7	-1.3222	0.315	-4.198	0.000	-1.946	-0.699
t8	0.1193	0.334	0.358	0.721	-0.541	0.780
t9	0.3831	0.230	1.667	0.098	-0.072	0.838
R10	0.0782	0.019	4.020	0.000	0.040	0.117
R20	0.0396	0.016	2.455	0.016	0.008	0.072
R30	0.0125	0.011	1.119	0.265	-0.010	0.035

Omnibus:	8.526	Durbin-Watson:	2.033
Prob(Omnibus):	0.014	Jarque-Bera (JB):	8.363
Skew:	-0.597	Prob(JB):	0.0153
Kurtosis:	3.312	Cond. No.	1.76e+03

Рис.3.4. Побудова моделі лінійної регресії з усіма факторами

Побудована модель лінійної регресії є статистично значущою і пояснює близько 51.3% дисперсії змінної 'eps'. Однак слід звернути увагу на можливу мультиколінеарність серед предикторів, про що свідчить високе число обумовленості. Це може впливати на стабільність та інтерпретацію окремих коефіцієнтів [9].

Прокоментуємо характнеристики моделі, наведені на рис. 3.4.

- **count**: у цьому рядку показано кількість непустих (не-NaN) значень у кожному стовпці. У даному випадку для всіх стовпців (t1 до eps) є 132 непустих значення. Це означає, що кожен стовпець має 132 спостереження.

- **mean**: це середнє значення (середнє арифметичне) для кожного стовпця. Наприклад, середнє значення для стовпця t1 становить 8.59, а для стовпця R30 - 78.95.
- **std**: це стандартне відхилення для кожного стовпця. Стандартне відхилення вимірює розсіювання значень навколо середнього. Більше значення вказує на більшу мінливість даних. Наприклад, стандартне відхилення для t1 становить 2.21, а для R30 - 44.18, що свідчить про значно більшу мінливість значень у стовпці R30.
- **min**: це мінімальне значення, знайдене в кожному стовпці. Наприклад, найменше значення в стовпці t1 - 2.54, а в стовпці eps - -21.02.
- **25%**: це перший квантиль. Він показує значення, нижче якого знаходиться 25% даних. Наприклад, 25% значень у стовпці t1 менші або рівні 7.15.
- **50%**: це медіана. Вона розділяє набір даних на дві рівні половини. Половина значень є меншою або рівною медіані, а інша половина - більшою або рівною. Для стовпця t1 медіана становить 8.82.
- **75%**: це третій квантиль. Він показує значення, нижче якого знаходиться 75% даних. Наприклад, 75% значень у стовпці t1 менші або рівні 9.88.
- **max**: це максимальне значення, знайдене в кожному стовпці. Наприклад, найбільше значення в стовпці t1 - 15.20, а в стовпці R20 - 156.00.

Отже, цей вивід надає нам швидкий огляд центральної тенденції (середнє, медіана), розсіювання (стандартне відхилення, квантилі, мінімум, максимум) та кількості спостережень для кожного числового стовпця у цьому наборі даних XY.

Для подальшого аналізу одним із найбільш поширених способів оцінки значущості факторів є перегляд стандартних помилок коефіцієнтів лінійної регресії (стандартних помилок оцінок). Чим менше стандартна помилка, тим більш значущий вважається коефіцієнт.

Наведене на рис. 3.4 описання моделі лінійної регресії дозволяє здійснити оцінку значущості факторів моделі. Деякі фактори (t1, t2, t3, t4, t5, t7, R10, R20) виявилися статистично значущими предикторами 'eps', тоді як інші (t6, t8, t9, R30) - ні (на рівні значущості 0.05), тому їх краще видалити (Рис.3.5.).

OLS Regression Results

Dep. Variable:	eps	R-squared:	0.484
Model:	OLS	Adj. R-squared:	0.450
Method:	Least Squares	F-statistic:	14.39
Date:	Sun, 14 Dec 2025	Prob (F-statistic):	1.14e-14
Time:	17:19:16	Log-Likelihood:	-404.14
No. Observations:	132	AIC:	826.3
Df Residuals:	123	BIC:	852.2
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	29.8296	5.533	5.392	0.000	18.878	40.781
t1	0.5551	0.263	2.113	0.037	0.035	1.075
t2	-0.5535	0.295	-1.874	0.063	-1.138	0.031
t3	0.8792	0.311	2.823	0.006	0.263	1.496
t4	-0.5374	0.210	-2.563	0.012	-0.952	-0.122
t5	-0.6352	0.267	-2.377	0.019	-1.164	-0.106
t7	-1.2376	0.286	-4.328	0.000	-1.804	-0.672
R10	0.0732	0.019	3.783	0.000	0.035	0.112
R20	0.0300	0.015	1.997	0.048	0.000	0.060

Omnibus:	15.581	Durbin-Watson:	2.191
Prob(Omnibus):	0.000	Jarque-Bera (JB):	17.608
Skew:	-0.775	Prob(JB):	0.000150
Kurtosis:	3.894	Cond. No.	867.

Рис.3.5. Видалення незначущих факторів

Видалення факторів призвело до зменшення потенційної мультиколінеарності, про що свідчить зниження числа обумовленості.

- **R-squared: 0.484**: Коефіцієнт детермінації трохи зменшився з 0.513 до 0.484. Це означає, що нова модель пояснює трохи меншу частку дисперсії залежної змінної 'eps' (48.4% проти 51.3%).

- **Adj. R-squared: 0.450**: Скоригований R^2 зменшився з 0.464 до 0.450. Зменшення скоригованого R^2 свідчить про те, що після зменшення кількості предикторів, модель пояснює трохи меншу частку варіації.
- **F-statistic: 14.39** та **Prob (F-statistic): 1.14e-14**: Загальна статистична значущість моделі залишається дуже високою (p-значення значно менше 0.05). Це означає, що навіть зі зменшеною кількістю факторів, модель в цілому є значущою.

Хоча це призвело до невеликого зменшення пояснювальної здатності моделі (зниження R^2 та скоригованого R^2), модель залишається статистично значущою. Крім того, модель стала більш економною з меншою кількістю предикаторів.

Далі створюємо новий DataFrame XY2, який містить відібрані фактори з X1 (після видалення 't1', 't4', 't8', 'R30') та оригінальну залежну змінну 'eps' з XY (рис.3.6.).

	const	t2	t3	t5	t6	t7	t9	R10	R20	eps
0	1.00	13.26	16.47	14.74	19.49	21.20	18.51	39.20	33.00	-0.72
1	1.00	10.60	13.78	12.83	14.68	15.98	19.42	44.90	38.30	9.72
2	1.00	11.76	12.54	17.24	19.06	16.55	23.99	8.20	6.80	3.06
3	1.00	10.32	10.49	20.93	20.99	20.18	19.46	14.90	54.90	-15.60
4	1.00	11.62	12.38	13.69	15.59	16.48	20.50	14.30	97.20	7.24

Рис.3.6. Побудова нового фрейму даних

Також розділяємо набір даних XY2 на навчальну вибірку (train, x_train, y_train) та контрольну вибірку (test, x_test, y_test). Навчальна вибірка, що становить 75% від усіх даних, буде використана для навчання моделі лінійної регресії. Контрольна вибірка, що залишилася (25%), буде використана для оцінки того, наскільки добре навчена модель узагальнює нові, невидимі дані. Та наступним етапом проводимо навчання та оцінки моделі лінійної регресії на розділених даних (рис.3.7).

```
train R^2: 0.51
train RMSE: 5.14
test R^2: 0.31
test RMSE: 5.49
```

Рис.3.7. Побудова моделі та оцінка її точності

Отже після навчання моделі лінійної регресії на навчальній вибірці та її оцінки на контрольній вибірці, ми бачимо, що модель має помірну пояснювальну здатність на навчальних даних (R^2=0.51), але її здатність до узагальнення на нових даних є нижчою (R^2=0.31). Середня помилка прогнозування (RMSE) також трохи вища на контрольній вибірці.

Для глибшого розуміння ефективності моделі, результати її прогнозування на навчальній вибірці були візуалізовані. Графік, що відображає залежність між фактичними (y_train) та прогнозованими (train_pred) значеннями, наочно демонструє, що хоча модель і фіксує загальну тенденцію в даних, її прогнози не є абсолютно точними (Рис.3.8).

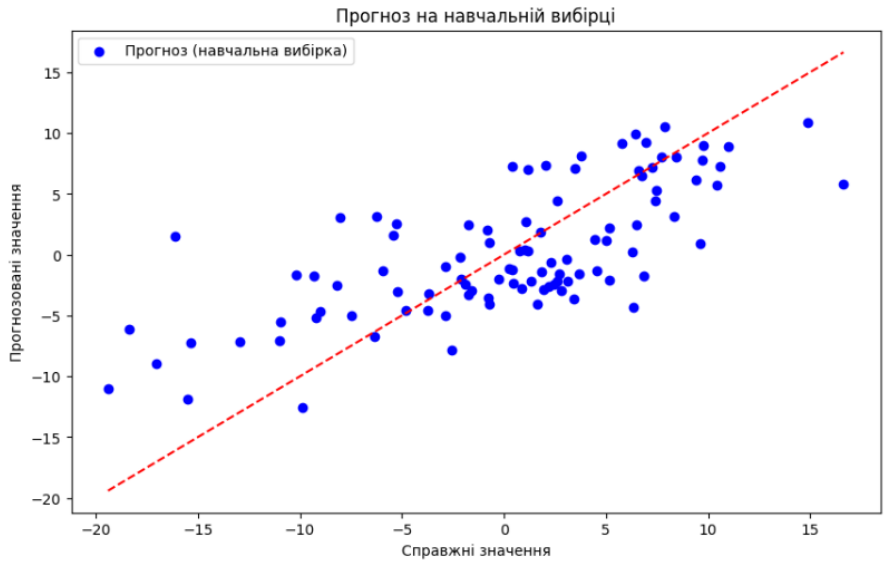

Рис.3.8. Графік відповідності для навчальної вибірки

Значний розкид синіх точок навколо лінії ідеального прогнозу підкреслює наявність розбіжностей між фактичними та передбаченими значеннями. Це візуальне спостереження повністю узгоджується з раніше отриманими кількісними метриками, такими як R^2 та RMSE, що характеризують якість прогнозування моделі на навчальній вибірці.

Для всебічного аналізу ефективності моделі, окрім графіка залежності між фактичними та прогнозованими значеннями, було побудовано поточковий

графік (Рис.3.9.), що відображає часову послідовність цих значень для навчальної вибірки.

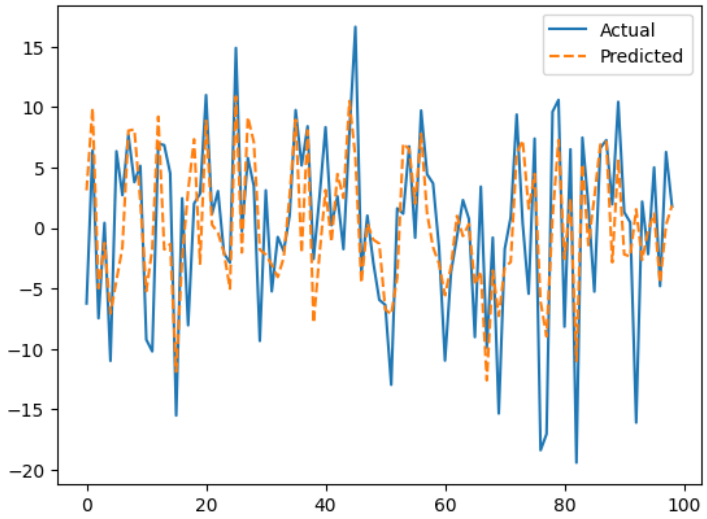

Рис.3.9. Поточковий графік факт - прогноз

Цей детальний графік показує, як модель справляється з кожним окремим спостереженням у часі. Візуально помітно, що прогнозована лінія загалом повторює деякі ключові коливання фактичних даних, проте існують суттєві розбіжності в амплітуді та фазі. Це ще раз підтверджує, що хоча модель і вловлює певні закономірності в навчальній вибірці, її прогнози далекі від ідеалу. Графік наочно демонструє ділянки, де модель показує відносну точність та ті моменти, де її похибки є найбільшими.

З метою оцінки здатності моделі до узагальнення, ми візуалізували результати її прогнозування на контрольній вибірці, яка складалася з раніше невидимих для моделі даних. Аналогічно до аналізу навчальної вибірки, графік відображає співвідношення між фактичними (y_test) та прогнозованими (test_pred) значеннями (Рис.3.10).

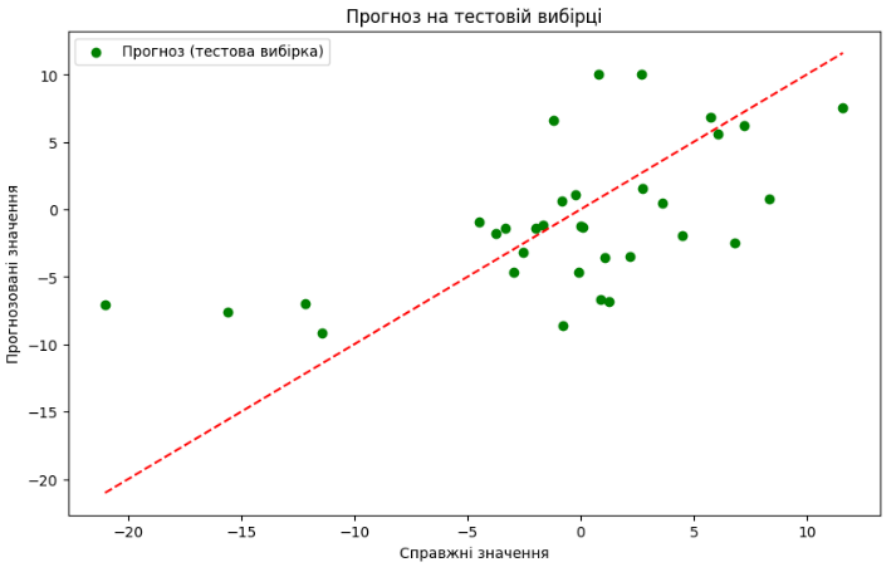

Рис.3.10. Графік відповідності для тестової вибірки

Візуальний аналіз графіка контрольної вибірки чітко демонструє гіршу прогностичну здатність моделі на нових даних порівняно з навчальними. Більший розкид зелених точок навколо лінії ідеального прогнозу свідчить про зниження точності узагальнення. Це підкреслює важливість подальшої роботи над удосконаленням моделі, що може включати оптимізацію набору факторів, застосування складніших алгоритмів або розширення обсягу навчальних даних. Порівняння з графіком для навчальної вибірки підтверджує очікуване краще прилягання точок до лінії ідеального прогнозу на навчальних даних, оскільки саме на них модель проходила навчання. Отже, помітно гірша якість прогнозування на контрольній вибірці є ключовим індикатором поточної здатності моделі до узагальнення.

На додаток до попереднього аналізу контрольної вибірки також було побудувано поточковий графік, що відображає часову послідовність фактичних та прогнозованих значень залежної змінної ('eps') для цих нових, невидимих даних. Цей графік наочно ілюструє (Рис.3.11.) обмежену здатність моделі точно прогнозувати значення залежної змінної в часовій послідовності на контрольній вибірці.

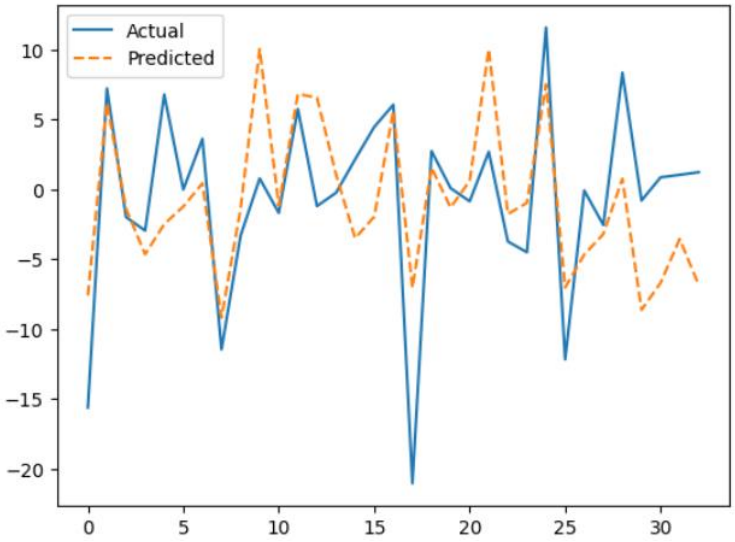

Рис.3.11. Поточковий графік факт - прогноз

Досить помітні розбіжності між фактичними та прогнозованими значеннями узгоджуються з раніше отриманими кількісними метриками для контрольної вибірки (R2=0.31 та RMSE = 5.49). Графік підкреслює, що, незважаючи на вловлювання деяких загальних закономірностей, модель не забезпечує достатньої точності при прогнозуванні окремих значень на нових даних. Це може бути зумовлено різними чинниками, серед яких недостатність інформації у вибраних предикторах, наявність нелінійних залежностей або ж непередбачуваність певної частки варіативності залежної змінної.

Після побудови моделі лінійної регресії постає питання оцінки її якості. Як відомо, модель навчається на навчальній вибірці і перевіряється на контрольній. Проте, часто виникає проблема перенавчання, коли модель добре працює на навчальних даних, але погано на нових. Щоб отримати більш об'єктивну оцінку ефективності моделі та уникнути перенавчання, застосовується метод крос-валідації K-fold. Перехресна крос-валідація (cross-validation) є потужним інструментом у машинному навчанні для оцінки ефективності моделі та уникнення перенавчання (overfitting) або недонавчання (underfitting), який передбачає розбиття даних на декілька піднаборів для навчання та тестування.

На початковому етапі ми відокремили незалежні змінні (впливаючі фактори) від залежної змінної (змінної відгуку), використовуючи підхід, відмінний від попереднього поділу на навчальну та контрольну вибірки. На цьому кроці ми підготували дані для потенційного нового поділу на навчальну та тестову вибірки за допомогою функції train_test_split. Також було визначено матрицю незалежних змінних X (з доданим стовпцем констант для врахування вільного члена в моделі) та вектор залежної змінної *y*.

Далі ми ініціалізуємо модель лінійної регресії, використовуючи бібліотеку scikit-learn, і підготувалися до застосування K-Fold крос-валідації для оцінки її продуктивності. Зокрема, було успішно створено об'єкт моделі лінійної регресії та налаштовано процес K-Fold крос-валідації з 5 складками та обов'язковим перемішуванням даних. На наступних етапах об'єкт kf буде використано для ітерації по різних навчальних та валідаційних наборах даних. На кожній ітерації модель буде навчатися на навчальній частині поточної складки, а її продуктивність оцінюватиметься на валідаційній частині зі збереженням метрик MSE (середньоквадратична помилка) та R^2 (коефіцієнт детермінації). Такий підхід дозволить отримати більш надійну оцінку здатності моделі до узагальнення порівняно з одноразовим поділом даних на навчальну та контрольну вибірки.

Результати K-Fold крос-валідації виявили помітну варіативність у продуктивності моделі лінійної регресії залежно від використовуваних підмножин даних. Діапазон значень середньоквадратичної помилки (MSE) склав від 18.52 до 48.91, що вказує на значні коливання в точності прогнозів на різних тестових наборах. Особливо високі значення MSE на Складках 3 та 5 свідчать про більші розбіжності між прогнозованими та фактичними значеннями на цих конкретних підмножинах даних (Рис.3.12.).

```
Складка 1: кількість елементів контрольної вибірки: 27
Ckв помилка (MSE): 18.68
R-квадрат: 0.63

Складка 2: кількість елементів контрольної вибірки: 27
Ckв помилка (MSE): 24.07
R-квадрат: 0.38

Складка 3: кількість елементів контрольної вибірки: 26
Ckв помилка (MSE): 48.91
R-квадрат: 0.27

Складка 4: кількість елементів контрольної вибірки: 26
Ckв помилка (MSE): 18.52
R-квадрат: 0.57

Складка 5: кількість елементів контрольної вибірки: 26
Ckв помилка (MSE): 39.02
R-квадрат: 0.35
```

Рис.3.12. Виконанання K-fold Cross-Validation

Аналогічно, коефіцієнт детермінації (R-квадрат) також демонструє значну мінливість, коливаючись від 0.27 до 0.63. Це означає, що частка дисперсії цільової змінної, яку пояснює модель, суттєво змінюється залежно від тестового набору. Низькі значення R-квадрат, особливо на Складках 3 та 5, говорять про слабку відповідність моделі даним на цих складках та її обмежену здатність пояснювати їхню варіацію.

Загалом, продуктивність моделі лінійної регресії є нестабільною на різних підмножинах даних. У той час як на деяких складках (1 та 4) спостерігаються відносно кращі результати з нижчим MSE та вищим R-квадрат, на інших (3 та 5) її продуктивність значно погіршується. Низькі значення R-квадрат на окремих складках можуть також свідчити про те, що лінійна модель не є оптимальним вибором для представлених даних, або ж існують невраховані фактори, що впливають на залежну змінну.

Обчислено середні значення метрик якості моделі, отримані в результаті K-Fold крос-валідації. Середня середньоквадратична помилка (MSE) становить 29.84, що є більш стабільною оцінкою типової величини помилки прогнозу на нових даних (Рис.3.13.).

```
Середня середньоквадратична помилка: 29.84
Середнє R-квадрат: 0.44
```

Рис.3.13. Усереднення результатів оцінювання моделей

Середній коефіцієнт детермінації (R-квадрат) дорівнює 0.44, що вказує на те, що в середньому модель пояснює близько 44% дисперсії залежної змінної. Ці усереднені значення, що знаходяться між найкращими та найгіршими результатами окремих складок, підкреслюють важливість крос-валідації для отримання об'єктивної оцінки узагальнюючої здатності моделі. Значення середнього R-квадрата 0.44 свідчить про те, що модель пояснює менше половини варіації цільової змінної, що може сигналізувати про необхідність подальшого вдосконалення моделі. Середня MSE надає кількісну міру типової помилки прогнозу, яку слід враховувати при практичному застосуванні моделі. Отримані результати підкреслюють важливість крос-валідації для надійної оцінки моделей машинного навчання.

3.2. Нелінійні моделі впливу кліматичних факторів на врожайність

Наша гіпотеза полягає у тому, що основною причиною коливань врожайності є зміна кліматичних факторів (температура повітря та кількість опадів). Перейдемо до аналізу впливу кліматичних факторів на коливання врожайності пшениці та створення математичної моделі, яка передбачає пояснення трендового відхилення врожайності пшениці (eps) на основі температурних показників. Нашою метою є оцінка впливу окремих кліматичних факторів на врожайність пшениці у степовій зоні України. Для реалізації даної мети застосовується метод лінійної регресії. Спочатку ми аналізуємо лише вплив лінійних кліматичних факторів, а у подальшому додаємо квадратичні кліматичні фактори для врахування нелінійних залежностей врожайності від клімату.

Для підвищення прогностичної здатності моделі необхідно додати нелінійні залежностей. Спробуємо додати до моделі значення квадратів температур (наприклад, t1^2, t2^2 тощо). В результаті отримуємо модель регресії з 24 факторами (12 лінійних та 12 нелінійних факторів - рис.3.2.1). Спочатку проведемо статистичний аналіз нових факторів

```
Перші 5 рядків нового датафрейму:
    t1    t2    t3    t4    t5    t6    t7    t8    t9   R10  ...   t3^2  \
0 9.23 13.26 16.47 13.34 14.74 19.49 21.20 19.15 18.51 39.20  ... 271.26
1 9.74 10.60 13.78 15.90 12.83 14.68 15.98 20.54 19.42 44.90  ... 189.89
2 5.63 11.76 12.54 15.36 17.24 19.06 16.55 21.18 23.99  8.20  ... 157.25
3 4.43 10.32 10.49 16.35 20.93 20.99 20.18 21.35 19.46 14.90  ... 110.04
4 6.55 11.62 12.38 15.17 13.69 15.59 16.48 18.94 20.50 14.30  ... 153.26

    t4^2   t5^2   t6^2   t7^2   t8^2   t9^2    R10^2    R20^2     R30^2
0 177.96 217.27 379.86 449.44 366.72 342.62 1,536.64 1,089.00 10,857.64
1 252.81 164.61 215.50 255.36 421.89 377.14 2,016.01 1,466.89  6,256.81
2 235.93 297.22 363.28 273.90 448.59 575.52    67.24    46.24  7,482.25
3 267.32 438.06 440.58 407.23 455.82 378.69   222.01 3,014.01  2,672.89
4 230.13 187.42 243.05 271.59 358.72 420.25   204.49 9,447.84  2,883.69

[5 rows x 24 columns]
```

Рис.3.2.1. Модель з квадратичними членами

Модель з квадратами. Після цього нами було побудовано лінійну регресію, яка включає квадратичні фактори (модель 2 - рис.3.2.2.)

```
--- Результати регресії (Всі 24 фактори) ---
                            OLS Regression Results
==============================================================================
Dep. Variable:                    eps   R-squared:                       0.671
Model:                            OLS   Adj. R-squared:                  0.597
Method:                 Least Squares   F-statistic:                     9.101
Date:                Sun, 07 Dec 2025   Prob (F-statistic):           1.61e-16
Time:                        18:43:08   Log-Likelihood:                -374.34
No. Observations:                 132   AIC:                             798.7
Df Residuals:                     107   BIC:                             870.7
Df Model:                          24
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         34.3557     55.085      0.624      0.534     -74.843     143.554
t1             4.4784      1.495      2.996      0.003       1.515       7.442
t2             0.6673      2.248      0.297      0.767      -3.789       5.124
t3             5.6574      2.301      2.459      0.016       1.096      10.218
t4             0.7005      1.665      0.421      0.675      -2.600       4.001
t5             4.2779      2.809      1.523      0.131      -1.290       9.846
t6             0.1390      1.953      0.071      0.943      -3.732       4.010
t7           -12.9472      3.162     -4.095      0.000     -19.215      -6.680
t8            -3.8762      3.436     -1.128      0.262     -10.687       2.935
t9             3.1323      3.207      0.977      0.331      -3.225       9.489
R10            0.1132      0.062      1.835      0.069      -0.009       0.236
R20            0.0440      0.045      0.988      0.326      -0.044       0.132
R30            0.0420      0.027      1.558      0.122      -0.011       0.095
t1^2          -0.2480      0.088     -2.827      0.006      -0.422      -0.074
t2^2          -0.0270      0.103     -0.263      0.793      -0.230       0.176
t3^2          -0.2019      0.089     -2.256      0.026      -0.379      -0.025
t4^2          -0.0115      0.052     -0.221      0.826      -0.115       0.092
t5^2          -0.1413      0.078     -1.803      0.074      -0.297       0.014
t6^2           0.0014      0.049      0.029      0.977      -0.096       0.099
t7^2           0.3021      0.081      3.743      0.000       0.142       0.462
t8^2           0.0812      0.079      1.034      0.304      -0.075       0.237
t9^2          -0.0637      0.071     -0.899      0.370      -0.204       0.077
R10^2         -0.0003      0.001     -0.436      0.664      -0.002       0.001
R20^2         -0.0002      0.000     -0.737      0.462      -0.001       0.000
R30^2         -0.0001   9.88e-05     -1.469      0.145      -0.000    5.07e-05
==============================================================================
Omnibus:                       12.096   Durbin-Watson:                   2.112
Prob(Omnibus):                  0.002   Jarque-Bera (JB):               17.327
Skew:                          -0.486   Prob(JB):                     0.000173
Kurtosis:                       4.485   Cond. No.                     1.83e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.83e+06. This might indicate that there are
```

Рис.3.2.2. Лінійна регресія разом з квадратичними змінами

Для нової моделі R-квадрат (R-squared): 0.671, а скорегований коефіцієнт детермінації (Adj.R^2): 0.597. Після додавання квадратичних членів пояснювальна здатність моделі значно зросла, пояснюючи 67.1% дисперсії.

Оскільки повна модель включала багато незначущих факторів, було проведено їхній відбір, залишивши лише ті, що були статистично значущими

(Рис.3.2.3.). До фінальної моделі увійшли 10 факторів: t1, t3, t6, t7, R10, t1^2, t3^2, t5^2, t6^2, t7^2).

OLS Regression Results

Dep. Variable:	eps	R-squared:	0.636
Model:	OLS	Adj. R-squared:	0.606
Method:	Least Squares	F-statistic:	21.18
Date:	Sun, 14 Dec 2025	Prob (F-statistic):	2.93e-22
Time:	17:19:31	Log-Likelihood:	-380.98
No. Observations:	132	AIC:	784.0
Df Residuals:	121	BIC:	815.7
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	63.9211	27.157	2.354	0.020	10.156	117.686
t1	4.8419	1.051	4.606	0.000	2.761	6.923
t3	6.5304	1.601	4.079	0.000	3.361	9.700
t6	2.5460	1.360	1.872	0.064	-0.146	5.238
t7	-14.3859	2.388	-6.024	0.000	-19.114	-9.658
R10	0.0719	0.017	4.229	0.000	0.038	0.106
t1^2	-0.2575	0.060	-4.321	0.000	-0.375	-0.139
t3^2	-0.2370	0.058	-4.058	0.000	-0.353	-0.121
t5^2	-0.0161	0.007	-2.163	0.033	-0.031	-0.001
t6^2	-0.0621	0.034	-1.816	0.072	-0.130	0.006
t7^2	0.3346	0.060	5.554	0.000	0.215	0.454

Omnibus:	12.518	Durbin-Watson:	2.074
Prob(Omnibus):	0.002	Jarque-Bera (JB):	18.735
Skew:	-0.484	Prob(JB):	8.55e-05
Kurtosis:	4.571	Cond. No.	4.64e+04

Рис.3.2.3. Видалення незначущих факторів

R-квадрат (R-squared): 0.636. Ця фінальна модель, незважаючи на меншу кількість предикторів, має високу пояснювальну здатність (63.6%) і є більш економною та статистично значущою. Усі фактори у фінальній моделі є статистично значущими.

Висновок. Аналіз вищенаведених моделей продемонстрував, що додавання квадратичних членів до регресійної моделі є ефективним підходом для прогнозування трендового відхилення врожайності, оскільки залежність від температурних факторів не є суто лінійною. Коефіцієнт детермінації

моделі квадратичної регресії $R^2 = 0{,}636$ і це набагато більше, ніж для лінійної моделі $R^2 = 0{,}484$.

Повна нелінійна модель.

Наше дослідження сфокусоване на багатофакторному регресійному аналізі, що включає ускладнені нелінійні залежності. Для підвищення точності моделювання до початкового набору чинників було включено не лише лінійні фактори та фактори-добутки, але й квадратичні члени. Це дозволяє моделі ефективно відображати нелінійну криву реакції врожайності на ключові агрокліматичні чинники (температуру та опади).

Дослідження пройшло три ключові етапи: підготовка даних із розширеним набором факторів, оцінка повної моделі та фінальна оптимізація.

Створення комплексного факторного простору стало можливим завдяки важливому розширенню набору змінних, яке включило три типи, що складають загалом 34 фактори (Рис.3.2.4.).

```
Перші 5 рядків нового датафрейму:
    t1    t2    t3    t4    t5    t6    t7    t8    t9   R10  ...  t1*t2  \
0 9.23 13.26 16.47 13.34 14.74 19.49 21.20 19.15 18.51 39.20  ... 122.39
1 9.74 10.60 13.78 15.90 12.83 14.68 15.98 20.54 19.42 44.90  ... 103.24
2 5.63 11.76 12.54 15.36 17.24 19.06 16.55 21.18 23.99  8.20  ...  66.21
3 4.43 10.32 10.49 16.35 20.93 20.99 20.18 21.35 19.46 14.90  ...  45.72
4 6.55 11.62 12.38 15.17 13.69 15.59 16.48 18.94 20.50 14.30  ...  76.11

   t2*t3  t3*t4  t4*t5  t5*t6  t6*t7  t7*t8  t8*t9  R10*R20  R20*R30
0 218.39 219.71 196.63 287.28 413.19 405.98 354.47 1,293.60 3,438.60
1 146.07 219.10 204.00 188.34 234.59 328.23 398.89 1,719.67 3,029.53
2 147.47 192.61 264.81 328.59 315.44 350.53 508.11    55.76   588.20
3 108.26 171.51 342.21 439.32 423.58 430.84 415.47   818.01 2,838.33
4 143.86 187.80 207.68 213.43 256.92 312.13 388.27 1,389.96 5,219.64

[5 rows x 34 columns]
```

Рис.3.2.4. Опис усіх даних

- Лінійні фактори (t1 до t9, R10): Прямий вплив температури та опадів.
- Квадратичні фактори (t1^2 до t9^2, R10^2, R20^2, R30^2): Враховують нелінійну криву реакції, наявність температурного оптимуму або ефекту насичення.

- Фактори-добутки (t1*t2 до t8*t9, R10*R20, R20*R30): Відображають пролонговану нелінійну дію та взаємодію сусідніх у часі температурних чинників та опадів.

Оцінка повної регресійної моделі, що містить 34 фактори, продемонструвала високу потенційну якість (Рис.3.2.5.)

Dep. Variable:	eps	R-squared:	0.764
Model:	OLS	Adj. R-squared:	0.681
Method:	Least Squares	F-statistic:	9.215
Date:	Sun, 14 Dec 2025	Prob (F-statistic):	3.50e-18
Time:	17:19:54	Log-Likelihood:	-352.56
No. Observations:	132	AIC:	775.1
Df Residuals:	97	BIC:	876.0
Df Model:	34		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	42.4148	56.065	0.757	0.451	-68.858	153.688
t1	4.6167	1.646	2.805	0.006	1.350	7.883
t2	0.5267	2.262	0.233	0.816	-3.962	5.016
t3	-1.7852	2.637	-0.677	0.500	-7.019	3.449
t4	1.0753	1.753	0.613	0.541	-2.404	4.555
t5	1.1997	2.954	0.406	0.685	-4.662	7.062
t6	5.5610	2.541	2.188	0.031	0.518	10.604
t7	-11.4642	3.365	-3.406	0.001	-18.144	-4.785
t8	0.2970	3.487	0.085	0.932	-6.623	7.217
t9	-0.7001	3.318	-0.211	0.833	-7.285	5.884
R10	0.1348	0.069	1.957	0.053	-0.002	0.272
R20	0.0471	0.052	0.913	0.363	-0.055	0.150
R30	0.0348	0.035	0.983	0.328	-0.035	0.105
t1^2	-0.1421	0.091	-1.555	0.123	-0.323	0.039
t2^2	-0.0604	0.133	-0.454	0.651	-0.325	0.204
t3^2	-0.3763	0.136	-2.758	0.007	-0.647	-0.105
t4^2	-0.2707	0.092	-2.930	0.004	-0.454	-0.087
t5^2	-0.3322	0.119	-2.803	0.006	-0.567	-0.097
t6^2	-0.1820	0.072	-2.545	0.013	-0.324	-0.040
t7^2	0.3568	0.131	2.734	0.007	0.098	0.616
t8^2	-0.2896	0.173	-1.671	0.098	-0.634	0.054
t9^2	-0.1178	0.085	-1.387	0.169	-0.286	0.051
R10^2	-0.0004	0.001	-0.660	0.511	-0.002	0.001
R20^2	-0.0002	0.000	-0.752	0.454	-0.001	0.000
R30^2	-9.296e-05	9.46e-05	-0.983	0.328	-0.000	9.47e-05
t1*t2	-0.1381	0.148	-0.932	0.354	-0.432	0.156
t2*t3	0.1545	0.146	1.061	0.291	-0.134	0.443
t3*t4	0.6994	0.137	5.097	0.000	0.427	0.972
t4*t5	-0.1058	0.109	-0.971	0.334	-0.322	0.111
t5*t6	0.5886	0.167	3.519	0.001	0.257	0.921
t6*t7	-0.4269	0.164	-2.605	0.011	-0.752	-0.102
t7*t8	0.2364	0.210	1.126	0.263	-0.180	0.653
t8*t9	0.2825	0.136	2.074	0.041	0.012	0.553
R10*R20	2.441e-05	0.001	0.036	0.971	-0.001	0.001
R20*R30	-0.0002	0.000	-0.474	0.637	-0.001	0.001

Omnibus:	7.929	Durbin-Watson:	2.013
Prob(Omnibus):	0.019	Jarque-Bera (JB):	9.990
Skew:	-0.354	Prob(JB):	0.00677
Kurtosis:	4.147	Cond. No.	2.16e+06

Рис.3.2.5. Модель регресії

- Висока пояснююча сила: R^2 = 0.764 (76.4%). Це підтвердило, що розширення факторного простору значно підвищило здатність моделі пояснювати варіацію залежної змінної (eps).

- Загальна значущість: Prob(F-statistic)=3.50e-18 підтверджує, що модель є високо значущою.
- Проблема мультиколінеарності: Включення такої великої кількості взаємопов'язаних членів призвело до екстремально високого Condition Number (2.16e+06). Це зробило окремі коефіцієнти в повній моделі нестабільними та важкими для надійної інтерпретації, що обґрунтувало необхідність ручної оптимізації.

Шляхом ручного вилучення статистично незначущих факторів (P > |t| > 0.05) була проведена оцінка фінальної ощадливої моделі (Рис.3.2.6.).

```
--- Результати регресії після ручного видалення факторів (18 факторів видалено)
                            OLS Regression Results
==============================================================================
Dep. Variable:                    eps   R-squared:                       0.745
Model:                            OLS   Adj. R-squared:                  0.710
Method:                 Least Squares   F-statistic:                     21.03
Date:                Sun, 14 Dec 2025   Prob (F-statistic):           6.16e-27
Time:                        17:20:01   Log-Likelihood:                -357.49
No. Observations:                 132   AIC:                             749.0
Df Residuals:                     115   BIC:                             798.0
Df Model:                          16
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         66.4287     22.700      2.926      0.004      21.465     111.393
t1             4.1957      1.017      4.125      0.000       2.181       6.210
t6             5.1077      1.513      3.377      0.001       2.112       8.104
t7           -12.2839      2.237     -5.492      0.000     -16.714      -7.854
R10            0.0910      0.016      5.668      0.000       0.059       0.123
t1^2          -0.2098      0.057     -3.713      0.000      -0.322      -0.098
t3^2          -0.3470      0.057     -6.042      0.000      -0.461      -0.233
t4^2          -0.2607      0.041     -6.340      0.000      -0.342      -0.179
t5^2          -0.3495      0.060     -5.793      0.000      -0.469      -0.230
t6^2          -0.1699      0.049     -3.461      0.001      -0.267      -0.073
t7^2           0.5038      0.078      6.462      0.000       0.349       0.658
t8^2          -0.1255      0.063     -1.987      0.049      -0.251      -0.000
t9^2          -0.1014      0.054     -1.866      0.065      -0.209       0.006
t3*t4          0.6251      0.098      6.384      0.000       0.431       0.819
t5*t6          0.5881      0.107      5.491      0.000       0.376       0.800
t6*t7         -0.4261      0.111     -3.845      0.000      -0.646      -0.207
t8*t9          0.2050      0.117      1.759      0.081      -0.026       0.436
==============================================================================
Omnibus:                       10.188   Durbin-Watson:                   2.074
Prob(Omnibus):                  0.006   Jarque-Bera (JB):               14.363
Skew:                          -0.411   Prob(JB):                     0.000761
Kurtosis:                       4.392   Cond. No.                     8.33e+04
==============================================================================
```

Рис.3.2.6. Фінальна модель

Ручне вилучення 18 незначущих факторів дозволило створити ощадливу модель із 16 значущими чинниками.

Критерії якості отриманої моделі:

- R^2 = 0.745 (74.5%): Модель зберегла майже всю пояснюючу здатність повної моделі, але з набагато меншою кількістю факторів.
- Скоригований R^2 = 0.710: Цей високий показник свідчить про те, що 16-факторна модель є кращою та ощадливішою за 34-факторну.
- Prob(F-statistic) = 6.16e-27: Гарантує, що модель є високо статистично значущою.
- Durbin-Watson = 2.074: Значення близьке до 2.0, що підтверджує відсутність автокореляції залишків.

Аналіз значущості та впливу факторів:

- Лінійний вплив: Фактори t1, t6, R10 є високо значущими ($P \leq 0.004$). Особливо виділяється негативний вплив t7 (коеф. -12.2839, P=0.000), що означає, що підвищення температури у період t7 (наприклад, критичний період наливу зерна) різко знижує врожайність.
- Квадратичний вплив: Значущість всіх квадратичних членів (t1^2, t3^2, t4^2, t5^2, t6^2, t7^2, t8^2) з $P \leq 0.049$ підтверджує, що реакція врожайності на температуру не є прямо пропорційною, а має нелінійну криву. Наприклад, негативні коефіцієнти для t1^2, t3^2, t4^2, t5^2, t6^2 вказують на ефект оптимуму (надмірне відхилення від середнього негативно впливає на врожай).
- Пролонгована дія через добутки: Значущість членів-добутків (t5*t6, t6*t7) підкреслює, що вплив температури є пролонгованим і відображає синергію/антагонізм між сусідніми періодами.

Покращення мультиколінеарності:

Condition No. (8.33e+04) суттєво знизилося порівняно з повною моделлю (2.16e+06), що підтверджує, що ручне вилучення факторів різко підвищило надійність та стійкість оцінок коефіцієнтів.

Висновок: Побудована фінальна модель є потужною та високо значущою, яка пояснює 74.5% коливань врожайності. Включення квадратичних та добуткових членів є найбільшим досягненням, оскільки це дозволяє моделі кількісно відобразити складну нелінійну криву реакції врожайності на агрокліматичні чинники. Ручна оптимізація успішно зменшила мультиколінеарність у 26 разів (2.16e+06 → 8.33e+04), підвищивши надійність моделі, що робить її значним кроком уперед для прогнозування врожайності.

3.3. Аналіз результатів прогнозування для областей

Цей аналіз присвячений оцінці ефективності прогностичної моделі шляхом порівняння її результатів із фактично спостережуваними даними. Ми розглядаємо динаміку залишкових значень (відхилення від тренду) для шести регіонів України (Одеська, Кіровоградська, Миколаївська, Дніпропетровська, Херсонська, Запорізька області).

Метою дослідження є порівняння двох ключових показників:

- Фактичні залишкові значення (eps, синя лінія): справжні відхилення від тренду врожайності.
- Прогнозовані залишкові значення (eps*, помаранчева лінія): відхилення, передбачені нашою моделлю.

Ідеальна модель мала б лінії, що точно збігаються. Аналізуючи графіки, ми шукаємо відповіді на два головні запитання: чи вловлює модель загальний напрямок змін залишків і чи правильно вона оцінює масштаб (величину) цих змін, особливо під час кризових явищ.

Результати показують, що модель успішно відображає динаміку процесів у часі, але має схильність до недооцінки амплітуди найсильніших коливань.

Далі, ми детально розглянемо результати для кожного регіону окремо, щоб визначити, де модель працює найкраще, а де вимагає вдосконалення.

Одеська область

Графік Одеської області охоплює період 2000–2021 років. Він демонструє високу початкову волатильність та є одним із найпроблемніших при оцінці якості прогнозування моделі (Рис.3.3.1).

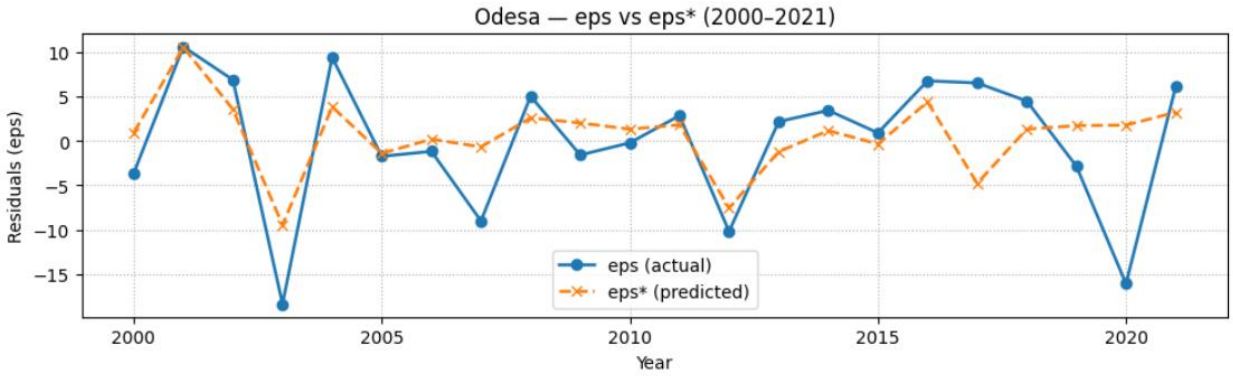

Рис.3.3.1. Моделювання залишків врожайності для Одеської області. Синя лінія – фактичні значення; штрихова помаранчева лінія – модель.

- Недооцінка кризового падіння (2003): У 2003 році фактична помилка різко впала до -15, що є одним із найглибших провалів на графіку. Модель передбачила падіння, але значно недооцінила його глибину (прогноз був лише близько -10).
- Систематичне заниження (2005–2010): У цей період лінії йдуть синхронно, але прогнозована лінія часто розташовується нижче фактичної. Це може вказувати на постійну схильність до заниження фактичних залишків моделлю.
- Критична розбіжність (2020): У 2020 році спостерігається критична розбіжність - найбільша помилка кінця періоду. Фактичний залишок різко впав до -15, але прогноз був позитивним (близько +5), що свідчить про кардинальну помилку у відображенні напрямку зміни.

Отже, модель для Одеси демонструє значну нестабільність і має серйозну проблему з відображенням як глибоких провалів, так і останніх значних коливань, що узгоджується з найнижчим кількісним показником R^2 для цього регіону.

Кіровоградська область

Графік Кіровоградської області (2000–2021) демонструє значну волатильність, але при цьому модель досить точно відстежує більшість екстремальних змін (Рис.3.3.2).

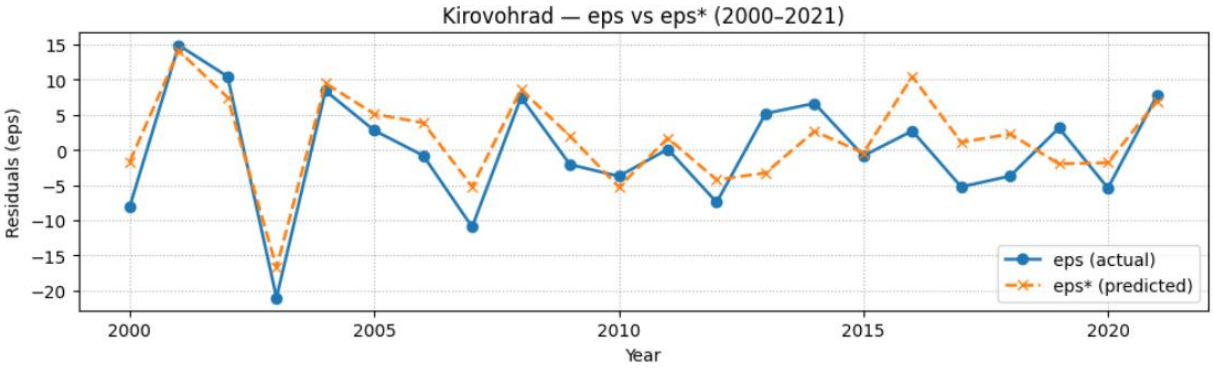

Рис.3.3.2. Моделювання залишків врожайності для Кіровоградської області. Синя лінія – фактичні значення; штрихова помаранчева лінія – модель.

- Ефективність у 2003: Модель дуже добре передбачила різке падіння залишків до -20 у 2003 році, майже ідеально відтворивши амплітуду цього глибокого провалу.
- Середній період (2005–2012): Лінії йдуть досить близько, точно вловлюючи переходи. Це період високої точності прогнозу.
- Кардинальна помилка (2016): Спостерігається найбільша розбіжність за весь період у 2016 році. Фактичне значення залишалося близько 0, тоді як прогноз різко зріс до +10, що вказує на помилку у визначенні напрямку та амплітуди змін.
- Відновлення точності: Після 2017 року і до кінця періоду, модель знову демонструє гарне узгодження з фактичними даними, особливо точно відтворюючи зростання у 2021 році.

Отже, модель для Кіровоградщини є однією з найкращих серед усіх регіонів у відображенні найглибшого негативного провалу 2003 року, але вона має очевидну слабкість у прогнозуванні сильних позитивних викидів (як у 2016 році).

Миколаївська область

Графік Миколаївської області (2000–2021) демонструє один із найвищих рівнів узгодженості між прогнозом і фактом, підтверджуючи високий показник R^2 для цього регіону (Рис.3.3.3).

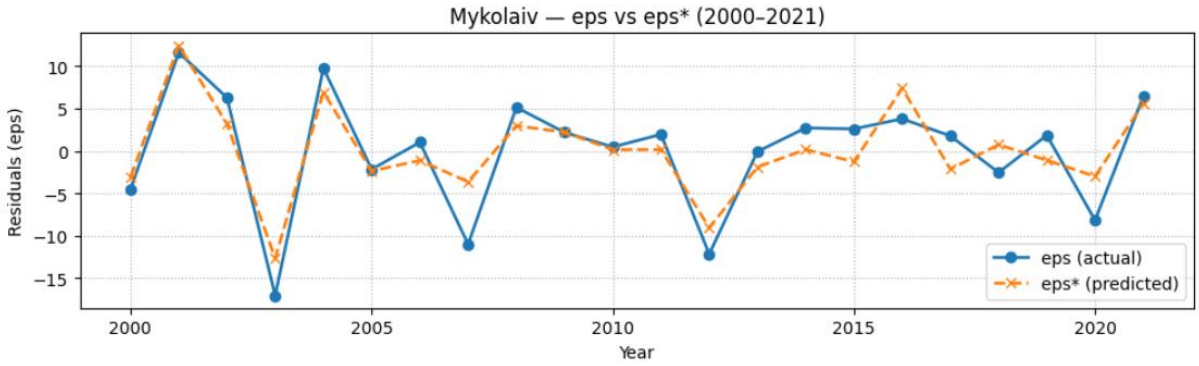

Рис.3.3.3. Моделювання залишків врожайності для Миколаївської області. Синя лінія – фактичні значення; штрихова помаранчева лінія – модель.

- Висока точність у кризі (2003): Модель дуже точно передбачила падіння до -15 у 2003 році, демонструючи високу чутливість до сильного шоку.
- Стабільна синхронність: Протягом майже всього періоду 2005–2021 лінії йдуть надзвичайно близько, часто перетинаючись. Модель майже ідеально відображає як невеликі, так і помірні коливання.
- Відхилення (2016): Найбільше відхилення спостерігається у 2016 році, коли фактичний залишок зріс вище, ніж прогноз (до +7.5). Однак, це відхилення є відносно невеликим порівняно з іншими регіонами.
- Надійна робота: Навіть у періоди волатильності (2019–2021) модель відмінно відтворила падіння та подальше зростання, підтверджуючи її високу надійність.

Отже, Миколаївська область служить еталоном якості прогнозування в цій зоні, демонструючи високу точність та мінімальну систематичну помилку як у періоди екстремальних, так і помірних коливань.

Дніпропетровська область

Графік Дніпропетровської області (2000–2021) демонструє глибокі початкові коливання, які модель переважно точно відстежує, але має проблеми з позитивними сплесками (Рис.3.3.4).

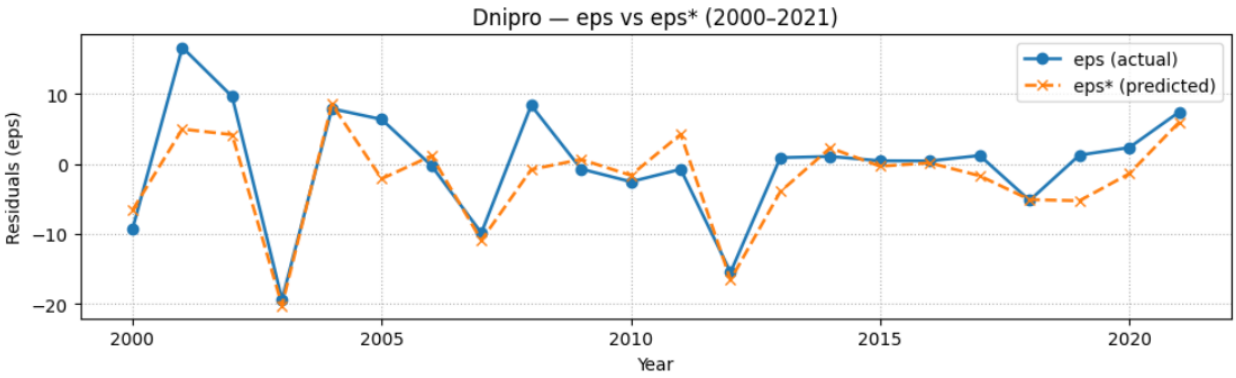

Рис.3.3.4. Моделювання залишків врожайності для Дніпропетровської області. Синя лінія – фактичні значення; штрихова помаранчева лінія – модель.

- Ідеальне відображення провалу (2003): Фактичний залишок впав до -20, і модель дуже точно передбачила цю амплітуду. Це є одним із найкращих відображень кризових провалів на всіх графіках.
- Систематичне заниження позитивних піків (2007–2010): У цей період модель показала слабкість: прогноз залишається близько нуля або нижче, тоді як фактичне значення піднімається до +10. Модель не змогла вловити значний позитивний сплеск.
- Висока стабільність (2012–2018): Модель демонструє дуже високу точність у цьому стабільному періоді, де обидві лінії коливаються в зоні ±5 і часто збігаються.

Отже, модель має високу здатність прогнозувати негативні "шоки", але демонструє слабкість та недооцінку амплітуди сильних позитивних сплесків, що вимагає уваги до факторів, які їх спричиняють.

Херсонська область

Графік Херсонської області (2000–2021) показує, що прогноз має високу кореляцію з фактичними даними, але часто вдається до згладжування коливань (Рис.3.3.5).

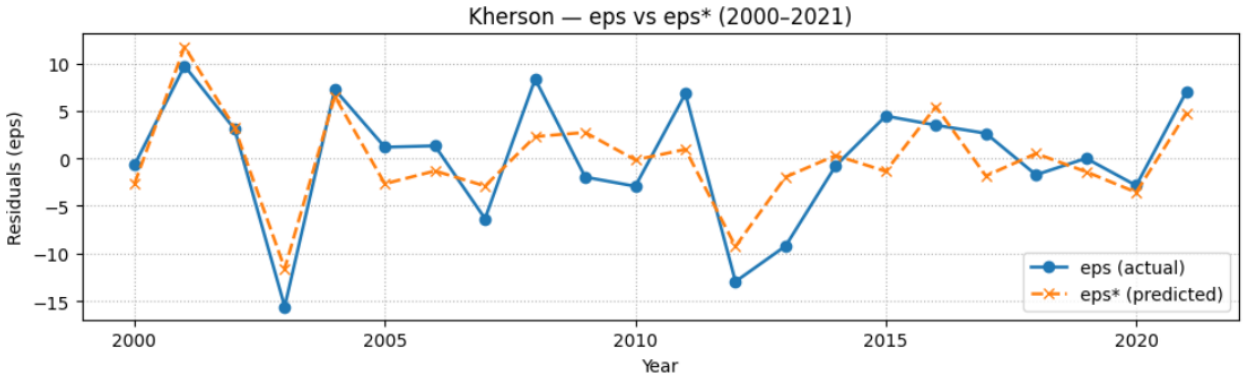

Рис.3.3.5. Моделювання залишків врожайності для Херсонська області. Синя лінія – фактичні значення; штрихова помаранчева лінія – модель.

- Точне відображення кризи (2003): Модель добре відтворила падіння 2003 року до -15, хоча і з невеликою недооцінкою.
- Систематичне згладжування: У період 2006–2012 прогнозована лінія (eps*) часто розташовується ближче до нуля, ніж фактична (eps), демонструючи схильність до згладжування амплітуди коливань. Наприклад, у 2012 році фактичне падіння до -13 було пом'якшене прогнозом до -8.
- Стабільність та надійність: У 2015–2019 роках модель демонструє високу точність, з мінімальними розбіжностями.
- Недооцінка фінального зростання: Хоча модель вловила динаміку кінця періоду (2020–2021), вона недооцінила амплітуду фінального зростання.

Отже, модель для Херсонщини є надійною у відстеженні напрямку змін, але її якість знижується через систематичне згладжування амплітуди коливань, особливо у періоди найбільшої волатильності.

Запорізька область

Графік Запорізької області (2000–2021) характеризується добре передбаченим початковим падінням, але має значні проблеми у відображенні коливань після 2015 року (Рис.3.3.6).

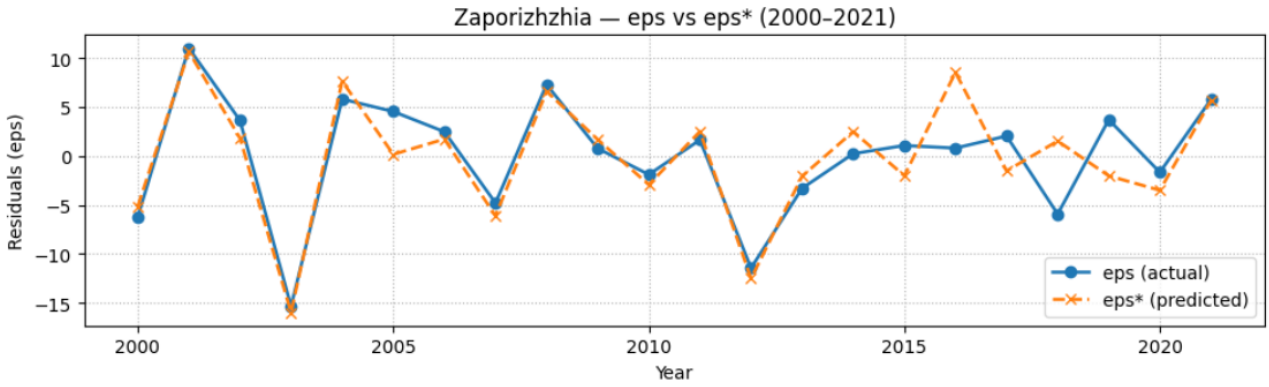

Рис.3.3.6. Моделювання залишків врожайності для Запорізька області. Синя лінія – фактичні значення; штрихова помаранчева лінія – модель.

- Висока точність у падінні (2003): Падіння до -15 у 2003 році було дуже точно передбачене моделлю, що вказує на гарне врахування кризових факторів.
- Сильна синхронність (2005–2014): Протягом цього десятиліття прогноз і фактичні дані йдуть надзвичайно близько, демонструючи високу якість моделі.
- Кардинальна помилка (2016): Спостерігається найбільша розбіжність у другій половині періоду. Фактичне значення залишається близько 0, тоді як прогноз різко злітає до +10. Це помилка напрямку та амплітуди, яка є серйозним недоліком моделі для Запоріжжя.
- Недооцінка зростання (2020): У 2020 році модель знову недооцінила зростання фактичних залишків (прогноз близько +2 при факті +6).

Отже, модель для Запоріжжя була надзвичайно сильною до 2015 року, але після цього її якість суттєво знизилася, демонструючи великі помилки у прогнозуванні сильних позитивних викидів, що співпадає з проблемами, виявленими у Кіровоградській області.

Загальний аналіз графіків для шести областей Степової зони (Одеса, Кіровоград, Миколаїв, Дніпро, Херсон, Запоріжжя) підтверджує, що модель

загалом має високу здатність відображати напрямок зміни залишкових значень. Для більшості регіонів, особливо до 2015 року, простежується висока синхронність між фактичними та прогнозованими даними.

Регіон Миколаївської області є еталоном, демонструючи високу точність та мінімальні розбіжності як у періоди екстремальних (2003), так і помірних коливань.

Модель виявилася винятково ефективною у відображенні найглибшого негативного шоку 2003 року для Дніпропетровської та Кіровоградської областей, де амплітуда падіння була передбачена майже ідеально.

На відміну від негативних шоків 2003 року, модель має системну проблему з прогнозуванням сильних позитивних викидів. Це особливо помітно у Кіровоградській та Запорізькій областях у 2016 році, де прогноз кардинально помилявся з напрямком або амплітудою.

Одеської області демонструє найбільшу нестабільність, маючи критичну розбіжність у 2020 році (фактичне падіння -15 при прогнозі +5). Ця візуальна оцінка підтверджує найнижчий кількісний показник R^2 для Одеси, що вказує на низьку пояснювальну силу моделі.

Для Херсонської області модель демонструє систематичну схильність до згладжування, роблячи прогнози амплітуди меншими, ніж фактичні коливання.

Таким чином, модель є ефективною для визначення тренду, але її точність щодо масштабу змін залежить від регіону та є обмеженою в періоди екстремальної волатильності.

Попередній аналіз графіків для степової зони (Херсон, Запоріжжя, Одеса, Кіровоград, Миколаїв, Дніпро) показав, що модель ефективно відстежує напрямок змін, але має проблеми з точністю амплітуди в періоди екстремальної волатильності. Щоб отримати повну картину якості, ми тепер оцінимо модель, використовуючи кількісні метрики точності для Степової зони, які представлені у таблиці 3.2.

Таблиця 3.3.2. Кількісні метрики точності прогнозних моделей для областей степової зони

	Регіон	MAE	RMSE	R2
0	Херсон	3.14	3.67	0.68
1	Запоріжжя	2.26	3.12	0.73
2	Одеса	4.22	5.85	0.40
3	Кіровоград	3.73	4.40	0.69
4	Миколаїв	2.62	3.13	0.79
5	Дніпро	3.26	4.55	0.68

Було використано три ключові метрики для оцінки моделі:

- R 2 (Коефіцієнт детермінації): Показує, яку частку змін у даних пояснює модель (ідеально R 2≈1.0).
- MAE (Середня абсолютна помилка): Середня величина помилки прогнозу (чим менше, тим краще).
- RMSE (Квадратична середня помилка): Середня помилка, яка більше штрафує великі відхилення (чим менше, тим краще).

Оцінка ефективності (R^2)

Ця метрика найкраще показує загальну прогностичну здатність моделі для кожного регіону:

- **Найкраща ефективність:** Модель демонструє найвищу прогностичну силу для Миколаївської області ($R^2 = 0.79$). Це означає, що 79% варіацій даних у цьому регіоні пояснюється факторами, включеними до моделі. Запорізька область ($R^2 = 0.73$) також має дуже високий та надійний показник.
- **Найнижча ефективність:** Для Одеської області модель працює найгірше ($R^2 = 0.40$). Це критично низький показник, який свідчить про те, що лише 40% варіації даних в Одесі пояснюється моделлю, а більшість змін

викликана неврахованими факторами. Це вимагає першочергового перегляду моделі для цього регіону.

- **Прийнятна ефективність:** Херсонська ($R^2 = 0.68$), Кіровоградська ($R^2 = 0.69$) та Дніпропетровська ($R^2 = 0.68$) області мають схожі, помірно високі та прийнятні показники.

Оцінка величини помилок (MAE та RMSE)

Ці метрики підтверджують, що регіони з високим R^2 мають менші помилки:

- **Найточніший прогноз:** Запорізька область має найменшу середню абсолютну помилку (MAE = 2.26), а Миколаївська (MAE = 2.62) також демонструє високу точність.
- **Найбільші помилки:** Одеська область має найбільшу MAE (4.22) та найбільшу RMSE (5.85), що підтверджує її найнижчу якість прогнозу.
- **Значні помилки:** Кіровоградська (MAE} = 3.73) та Дніпропетровська (MAE = 3.26) області мають помітно вищі помилки порівняно з лідерами.

Аналіз впливу викидів (співвідношення RMSE / MAE)

Порівнюючи MAE та RMSE, можемо оцінити, наскільки сильно на загальну помилку впливають поодинокі великі відхилення (викиди):

- **Одеса** (RMSE = 5.85 / MAE = 4.22): Відношення (5.85 / 4.22 ≈ 1.39) є найвищим. Це підтверджує, що низький R^2 для Одеси пов'язаний не лише із загальною невідповідністю, а й із частими або особливо великими, непередбачуваними помилками (викидами).
- **Миколаївська** (RMSE = 3.13 / MAE = 2.62): Низьке відношення (3.13 / 2.62 ≈ 1.20) підтверджує, що помилки тут не тільки маленькі, але й рівномірно розподілені, без критичних викидів.

- **Запорізька** (RMSE = 3.12 / MAE = 2.26): Відношення (3.12 / 2.26 ≈ 1.38) є майже таким же високим, як і в Одесі. Це вказує на те, що, хоча середня помилка (MAE) тут найменша, регіон все ж таки схильний до сильних поодиноких викидів, які істотно збільшують RMSE. Це нетипово для регіону з високим R^2 і вимагає додаткового вивчення.

Висновок.

Аналіз кількісних метрик демонструє значну регіональну нерівність у прогностичній силі моделі:

- **Лідер якості:** Модель працює найкраще та є найнадійнішою для Миколаївської області ($R^2 = 0.79$) та Запорізької області ($R^2 = 0.73$). Миколаївська область є еталоном для порівняння.
- **Критична зона:** Одеська область вимагає негайного втручання та перегляду факторів моделі, оскільки її якість ($R^2 = 0.40$) є критично низькою і пов'язана зі значними, неконтрольованими викидами.
- **Скритий ризик:** Хоча Запорізька область є лідером за точністю (MAE), високе співвідношення RMSE / MAE (1.38) вказує на проблему з екстремальними викидами, яка може погіршити прогноз у майбутньому.

Ці результати підкреслюють, що для підвищення загальної прогностичної сили необхідно провести регіональну адаптацію моделі, особливо зосередившись на включенні унікальних місцевих факторів для найпроблемніших областей (Одеса) та аналізі причин поодиноких, але сильних помилок у регіонах-лідерах (Запоріжжя).

ВИСНОВКИ

Врожайність пшениці в Україні в останні досліджувані роки демонструє зростаючий тренд. Однак, цей тренд поєднується із значними міжрічними коливаннями, які переважно пояснюються впливом кліматичних факторів. З точки зору вирощування пшениці, Україна поділяється на три основні агрокліматичні зони: Степова, Лісостепова та Полісся. При цьому вплив кліматичних факторів на врожайність є різним у кожній із цих зон.

У процесі нашого дослідження були зібрані та детально оброблені кліматичні показники шести областей, що належать до Степової зони України.

Основною метою нашої роботи було змоделювати вплив кліматичних факторів на врожайність пшениці саме в Степовому регіоні. Для досягнення цієї мети ми послідовно застосували три різні моделі: лінійну регресію, квадратичну та загальну нелінійну.

Як незалежні змінні лінійна регресійна модель включає 12 лінійних кліматичних факторів: показники температури (t1-t9) та опадів (R10, R20, R30). Залежною змінною (eps) виступає трендове відхилення врожайності пшениці. Нами була побудована лінійна регресійна модель, яка оцінює саме лінійний вплив кліматичних факторів на залишок врожайності (eps). Ця модель виявилася статистично достовірною, а її коефіцієнт детермінації (R^2) склав 0,51.

Всі розрахунки були виконані у середовищі Python з використанням таких бібліотек, як pandas, scikit-learn, statsmodels. Важливою перевагою моделювання у середовищі Python є можливість отримати повний опис моделі. Для оцінки значущості кожного погодного фактора, ми ретельно аналізували статистичні показники коефіцієнтів регресії. Щоб перевірити здатність моделі робити точні прогнози на нових (невидимих) даних, ми розділили наявні дані на навчальну та контрольну вибірки. Крім того, для більш об'єктивної оцінки якості моделі та запобігання перенавчанню, ми використали метод K-Fold

крос-валідації, який передбачає багаторазове навчання та тестування моделі на різних підмножинах даних. Результати цієї перевірки підтвердили адекватність побудованої регресійної моделі.

Проте, вплив кліматичних факторів на врожайність часто є суттєво нелінійним, що науково підтверджується існуванням зони екологічного оптимуму. Зважаючи на це, ми додали до нашої моделі квадратичні фактори. Модель, що містить як лінійні, так і квадратичні фактори, виявилася статистично значущою і продемонструвала покращений коефіцієнт детермінації — 0,67. У певних агрокліматичних ситуаціях критично важливо врахувати пролонговану (кумулятивну) дію кліматичних факторів. Наприклад, тривала посуха протягом двох місяців може призвести до загибелі врожаю. Для врахування такої пролонгованої дії ми включили в модель добутки сусідніх факторів (10 добутків: 8 температурних і 2 опадових).

Таким чином, загальна нелінійна модель фінально включала: 12 лінійних факторів, 12 квадратичних факторів та 10 добутків факторів. Ця комплексна модель є статистично значущою і досягла найвищого коефіцієнта детермінації — 0,76.

Виконана нами робота має важливе практичне значення, оскільки дозволяє виконувати прогнозування врожайності пшениці з періодом упередження три місяці. Дане дослідження представляє практичний інтерес для фермерів, зернових асоціацій та зернотрейдерів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Природні зони України: веб-сайт. URL: https://geomap.com.ua/uk-g4/20.html
2. Оптимізація сільського господарства Степу України: веб-сайт. URL: https://goodway.club/ekology/217-optimization-agriculture/
3. Врожай 2022: веб-сайт. URL: https://kurkul.com/spetsproekty/1406-yak-viyna-vplinula-na-vrojay-zernovih-ta-oliynih--pidsumki-sezonu-2022
4. Основи машинного навчання : навч. посіб. / В. О. Харченко. – Суми : Сумський державний університет, 2023. – 264 с.
5. Архів метеорологічних даних. Режим доступу: веб-сайт. URL: https://meteopost.com/weather/archive/
6. Державна служба статистики України. Режим доступу: веб-сайт. URL: http://www.ukrstat.gov.ua.
7. Клімат України / За ред. В. М. Ліпінського, В. А. Дячука, В. М. Бабіченко. — К.: Вид-во Раєвського, 2003. — 343 с.
8. Hrytsiuk, P., Havryliuk, M.: Modeling of the nonlinear impact of climatic factors on wheat yield using machine learning techniques. In book: Information and Communication Technologies in Education, Research, and Industrial Applications. Pp 20-35 (2025)
9. Hrytsiuk, P., Babych, T., Baranovsky, S., Havryliuk, M.: Assessing of Climate Impact on Wheat Yield using Machine Learning Techniques. CEUR Workshop Proceedings, 3513, 314–329 (2023)
10. Draper, N. R., Smith, H.: Applied Regression Analysis. Wiley, New York, NY (1998)
11. Petro Hrytsiuk1[0000-0002-3683-4766] and Maksym Havryliuk1[0000-0003-1149-6251]: The National University of Water and Environmental Engineering, Soborna str., 11, Rivne, 33028, Ukraine

ДОДАТКИ

```
# Завантаження даних з диска
from google.colab import files
import pandas as pd
import numpy as np
import statsmodels.api as sm
from scipy import stats


from google.colab import files

uploaded = files.upload() # Після зчитування файлу рядок задокументувати
file = "stepa.csv"
XY = pd.read_csv(file)
#  XY = XY.sort_values(by="eps", ascending=False).reset_index(drop=True)
# print(XY.head(), "\n")
XY0 = XY.copy()

pd.options.display.float_format = '{:,.2f}'.format

print("\n--- Інформація про дані ---")
XY.info()
print("\n--- Статистичний опис даних ---")
pd.options.display.float_format = '{:,.2f}'.format
print(XY.describe())


# Розділення змінних на впливаючі фактори та змінну-відгук
# Визначення залежної змінної (y)
y = XY['eps']
# Визначення факторів (X) - всі колонки, крім 'eps'
X_original = XY.drop(columns=['eps'])
factors = X_original.columns.tolist()


print("\n\n--- ЗАВДАННЯ 1: Рівняння лінійної регресії з 9 факторами ---")

# Додати константу до матриці X
X = sm.add_constant(X_original)

# Побудова загальної моделі з усіма факторами
model = sm.OLS(y, X).fit()

print("\n--- Результати регресії (Всі 9 факторів) ---")
print(model.summary())


# ручне видалення факторів
X1 = X.drop(columns=['t6','t8','t9','R30']) #
```

```
model1 = sm.OLS(y, X1).fit()
print(model1.summary())


print("\n\n--- МОДЕЛЬ 2: Додавання квадратичних змінних ---")

# Створення копії датафрейма для квадратичних членів
X_quadratic_factors = X_original.copy()

# Додаємо квадратичні члени для кожного з 12 факторів
for factor in factors:
    X_quadratic_factors[f'{factor}^2'] = X_original[factor] ** 2

print("\nПерші 5 рядків нового датафрейму:")
print(X_quadratic_factors.head())


print("\n\n--- Модель 2 з 24 факторами (лінійні + квадратичні) ---")

# Додаємо константу до матриці X з квадратичними членами
X_total = sm.add_constant(X_quadratic_factors)

# Побудова загальної моделі з усіма 24 факторами
model_total = sm.OLS(y, X_total).fit()

print("\n--- Результати регресії (Всі 24 фактори) ---")
print(model_total.summary())



# ручне видалення факторів
X2 =
X_total.drop(columns=['t2','t4','t5','t8','t9','t2^2','t4^2','t8^2','t9^2'
,'R20','R30','R10^2','R20^2','R30^2']) #
model2 = sm.OLS(y, X2).fit()
print(model2.summary())


import pandas as pd
import statsmodels.api as sm

# Визначення факторів, які використовуються у попередній моделі
factors_t1_t9 = [f't{i}' for i in range(1, 10)]
X_t_factors = X_original[factors_t1_t9].copy()

# --- Додавання добутків сусідніх лінійних факторів ---
print("\n\n--- ЗАВДАННЯ 1: Додавання добутків сусідніх лінійних факторів -
--")

# Створення копії датафрейма для нової моделі
```

```
X_product_factors = X_t_factors.copy()

# Додаємо добутки сусідніх лінійних факторів: t1*t2, t2*t3, ..., t8*t9
for i in range(1, 9): # Індекси для t1*t2 до t8*t9
    factor_i = f't{i}'
    factor_j = f't{i+1}'
    new_factor_name = f'{factor_i}*{factor_j}'
    # Обчислення добутку і додавання як нової колонки
    X_product_factors[new_factor_name] = X_t_factors[factor_i] *
X_t_factors[factor_j]

print("\nПерші 5 рядків нового датафрейму:")
print(X_product_factors.head())


# --- Модель 3 з 18 факторами (лінійні t1-t9 + добутки) ---
print("\n\n--- Модель 3 з 18 факторами (лінійні + добутки) ---")

# Додаємо константу до матриці X з факторами-добутками
X_total_products = sm.add_constant(X_product_factors)

# Побудова загальної моделі з усіма 18 факторами
model_total_products = sm.OLS(y, X_total_products).fit()

print("\n--- Результати регресії (Всі 18 факторів) ---")
print(model_total_products.summary())


import statsmodels.api as sm
print("\n\n--- МОДЕЛЬ 3: Ручне видалення обраних факторів з 18-факторної
моделі ---")

factors_to_drop = ['t3', 't7*t8']
#
==========================================================================

# Створюємо нову матрицю X, видаляючи обрані фактори
X_manual_drop = X_total_products.drop(columns=factors_to_drop)

# Побудова нової моделі регресії з меншою кількістю факторів
model_manual_drop = sm.OLS(y, X_manual_drop).fit()

print(f"\n--- Результати регресії після ручного видалення факторів:
{factors_to_drop} ---")
print(model_manual_drop.summary())


#  Повна модель

# Об’єднання всіх блоків факторів
```

```
X_full = X_original.copy()

# Квадрати
for factor in factors:
    X_full[f"{factor}^2"] = X_original[factor] ** 2

# Добутки температур
t_factors = [f"t{i}" for i in range(1, 10)]   # оголосити

for i in range(len(t_factors) - 1):
    f1, f2 = t_factors[i], t_factors[i + 1]
    X_full[f"{f1}*{f2}"] = X_original[f1] * X_original[f2]

# Добутки опадів
r_factors = ["R10", "R20", "R30"]

for i in range(len(r_factors) - 1):
    f1, f2 = r_factors[i], r_factors[i + 1]
    X_full[f"{f1}*{f2}"] = X_original[f1] * X_original[f2]

print("\nПерші 5 рядків нового датафрейму:")
print(X_full.head())


# --- Побудова загальної моделі з усіма 26 факторами ---
print("\n\n--- Модель 4 з 26 факторами (Лінійні + Квадратичні + Добутки) -
--")

# Додаємо константу до матриці X
X_total_full = sm.add_constant(X_full)

# Побудова загальної моделі
model_total_full = sm.OLS(y, X_total_full).fit()

print("\n--- Результати регресії (Всі 26 факторів) ---")
print(model_total_full.summary())


import statsmodels.api as sm
# --- Ручне видалення незначущих факторів з повної моделі ---
print("\n\n--- ЗАВДАННЯ: Ручне видалення обраних факторів з 26-факторної
моделі ---")

factors_to_drop =
['t2','t3','t4','t5','t8','t9','R20','R30','t2^2','R10^2','R20^2','R30^2',
't1*t2','t2*t3','t4*t5','t7*t8','R10*R20','R20*R30'] # <--- ВИКОРИСТАННЯ
ВАШОГО ЗРАЗКА
#
==========================================================================
```

```
# Створюємо нову матрицю X, видаляючи обрані фактори
# Видалення відбувається з матриці X_total_full, яка включає константу
('const')
X_manual_full_drop = X_total_full.drop(columns=factors_to_drop)

# Побудова нової моделі регресії
model_manual_full_drop = sm.OLS(y, X_manual_full_drop).fit()

print(f"\n--- Результати регресії після ручного видалення факторів
({len(factors_to_drop)} факторів видалено) ---")
print(model_manual_full_drop.summary())


# Графічна ілюстрація роботи моделі (період 2000-2021)
import os
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm

CSV_PATH = "stepa.csv"

XY = pd.read_csv(CSV_PATH)
XY.columns = XY.columns.str.strip()

# 1) ВІДНОВЛЮЄМО РОКИ ТА ОБЛАСТІ ЛИШЕ ЗА ІНДЕКСОМ РЯДКА
regions = ["Kherson", "Zaporizhzhia", "Odesa", "Kirovohrad", "Mykolaiv",
"Dnipro"]

years = np.arange(2000, 2022)   # 2000..2021 (22 роки)
ny = len(years)                 # 22

n = len(XY)
assert n == ny * len(regions), f"Очікував {ny*len(regions)} рядків
({ny}*{len(regions)}), маємо {n}."

row_id = np.arange(n)           # 0..n-1
block = row_id // ny            # 0..5
pos_in_block = row_id % ny      # 0..21

XY["region"] = [regions[b] for b in block]
XY["year"] = years[pos_in_block]

# 2) ДОДАЄМО eps* (якщо його нема) — ваша скорочена модель
if "eps_star" not in XY.columns:
    y = XY["eps"]
    X_original = XY.drop(columns=["eps"])
    fac = [f"t{i}" for i in range(1, 10)]
    Xb = X_original[fac].copy()
```

```
    X = Xb.copy()
    for f in fac:
        X[f"{f}^2"] = Xb[f]**2
    for i in range(1, 9):
        fi, fj = f"t{i}", f"t{i+1}"
        X[f"{fi}*{fj}"] = Xb[fi] * Xb[fj]

    Xc = sm.add_constant(X)

    drop =
['t2','t3','t4','t5','t8','t2^2','t8^2','t9^2','t1*t2','t2*t3','t4*t5','t6
*t7']
    # захист, якщо якихось колонок може не бути (на всяк випадок)
    drop = [c for c in drop if c in Xc.columns]

    Xr = Xc.drop(columns=drop)
    model = sm.OLS(y, Xr).fit()
    XY["eps_star"] = model.predict(Xr)

# 3) САНІТІ-КОНТРОЛЬ: всередині кожного блоку рівно 22 роки 2000..2021
for r in regions:
    yrs = XY.loc[XY.region == r, "year"].tolist()
    assert yrs == list(years), f"Роки у {r} не 2000..2021 по порядку.
Отримано: {yrs[:5]}..."

# 4) ПЛОТИНГ + ЗБЕРЕЖЕННЯ У папку ./plots
out_dir = "./plots"
os.makedirs(out_dir, exist_ok=True)

def plot_region(df_region, region_name):
    dfp = df_region.sort_values("year").reset_index(drop=True)
    plt.figure(figsize=(10, 3.2))
    plt.plot(dfp["year"], dfp["eps"], marker="o", linewidth=1.8,
label="eps (actual)")
    plt.plot(dfp["year"], dfp["eps_star"], marker="x", linestyle="--",
linewidth=1.8, label="eps* (predicted)")
    plt.title(f"{region_name} — eps vs eps* (2000–2021)")
    plt.xlabel("Year"); plt.ylabel("Residuals (eps)")
    plt.grid(True, linestyle=":"); plt.legend()
    out_png = os.path.join(out_dir,
f"plot_eps_vs_eps_star_{region_name}.png")
    plt.tight_layout(); plt.savefig(out_png, dpi=150); plt.show()
    return out_png

saved = {r: plot_region(XY.loc[XY["region"]==r,
["year","eps","eps_star"]], r) for r in regions}
print(saved)
```

```
print("\nKherson, перші 5 рядків:")
print(XY.loc[XY.region=="Kherson", ["year","eps","eps_star"]].head())


import os
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from openpyxl import Workbook
from openpyxl.drawing.image import Image as XLImage
from openpyxl.utils.dataframe import dataframe_to_rows
from sklearn.metrics import mean_absolute_error, mean_squared_error,
r2_score

# --- Вхідні дані ---
# (припускаємо, що XY вже створений і містить колонки: region, year, eps,
eps_star)
regions = ["Kherson", "Zaporizhzhia", "Odesa", "Kirovohrad", "Mykolaiv",
"Dnipro"]

# --- КРОК 1. Обчислення MAE, RMSE, R2 ---
results = []
out_dir = "./report_plots"
os.makedirs(out_dir, exist_ok=True)

for r in regions:
    df = XY[XY["region"] == r].sort_values("year").reset_index(drop=True)
    y_true, y_pred = df["eps"], df["eps_star"]

    mae = mean_absolute_error(y_true, y_pred)
    rmse = np.sqrt(mean_squared_error(y_true, y_pred))
    r2 = r2_score(y_true, y_pred)

    results.append({"Region": r, "MAE": mae, "RMSE": rmse, "R2": r2})

    # Діапазон років беремо з даних )
    y0 = int(df["year"].min())
    y1 = int(df["year"].max())

    # побудова графіка для Excel
    plt.figure(figsize=(7, 3))
    plt.plot(df["year"], y_true, "o-", label="eps (actual)")
    plt.plot(df["year"], y_pred, "x--", label="eps* (predicted)")
    plt.title(f"{r} — eps vs eps* ({y0}-{y1})")
    plt.xlabel("Year")
    plt.ylabel("Residuals (eps)")
    plt.legend()
    plt.grid(True, linestyle=":")
```

```
    fig_path = os.path.join(out_dir, f"{r}_plot.png")
    plt.tight_layout()
    plt.savefig(fig_path, dpi=150)
    plt.close()


results_df = pd.DataFrame(results)
results.append({"Region": r, "StartYear": y0, "EndYear": y1, "MAE": mae,
"RMSE": rmse, "R2": r2})
print("Зведені показники точності:")
display(results_df)

# --- КРОК 2. Формування Excel-звіту ---
report_path = "./Steppe_Yield_Residuals_Report.xlsx"
wb = Workbook()
ws_summary = wb.active
ws_summary.title = "Summary"

# Вставляємо загальну таблицю
for r in dataframe_to_rows(results_df, index=False, header=True):
    ws_summary.append(r)

# --- КРОК 3. Окремі листи по кожній області ---
for r in regions:
    ws = wb.create_sheet(title=r[:30])  # до 31 символа дозволено у Excel
    df = XY[XY["region"] == r][["year", "eps", "eps_star"]].copy()
    # Запис таблиці
    for row in dataframe_to_rows(df, index=False, header=True):
        ws.append(row)

    # Вставлення метрик у верхній частині листа
    met = results_df[results_df["Region"] == r].iloc[0]
    ws.append([])
    ws.append(["MAE", "RMSE", "R2"])
    ws.append([met["MAE"], met["RMSE"], met["R2"]])

    # Додаємо зображення (графік)
    fig_path = os.path.join(out_dir, f"{r}_plot.png")
    if os.path.exists(fig_path):
        img = XLImage(fig_path)
        img.anchor = "E2"  # позиція вставки
        ws.add_image(img)

wb.save(report_path)
print(f"\n✅ Excel-звіт збережено: {report_path}")
```