КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА

Факультет інформаційних технологій

Кафедра технологій управління

Спеціальність 122 – Комп'ютерні науки,
освітня програма «Інформаційна аналітика та впливи»

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

на тему:

Розробка моделей та інформаційної системи прогнозування змін клімату за допомогою методів Data Science

Студента 2-го курсу групи ІАВ-21
Українця А. О.

Науковий керівник
к.е.н., доцент
Мірошниченко І. В.

_____ (підпис студента) _____ (дата) _____ (підпис)

Попередній захист:

_____
(Висновок: «До захисту в Екзаменаційній комісії»)

Завідувач кафедри
технологій управління _____ (підпис) _____ (прізвище, ініціали) _____ (дата)

Київ – 2025

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА
Факультет інформаційних технологій

Кафедра технологій управління
Освітньо-кваліфікаційний рівень Магістр
Спеціальність 122 – Комп'ютерні науки
Освітня програма Інформаційна аналітика та впливи

ЗАТВЕРДЖУЮ
Завідувач кафедри
Професор МОРОЗОВ В. В.
_____
«___» _____ 2025 р.

ЗАВДАННЯ
НА ВИКОНАННЯ КВАЛІФІКАЦІЙНОЇ РОБОТИ

Студент Українець А. О.

Група ІАВ-21

1. **Тема кваліфікаційної роботи**

Розробка моделей та інформаційної системи прогнозування змін клімату за допомогою методів Data Science

Затверджена наказом по від «_____»____2025р. № __.

2. **Строк подання студентом готової роботи –** "__" _____2025р.

3. **Цільова установка та вихідні дані до роботи**

Проаналізувати вплив різних факторів на зміну клімату та підвищення середньої температури на планеті та в окремих країнах. Застосувати сучасні методи аналізу даних та прогнозування для створення моделей, що дозволяють здійснювати прогнозування змін середньої температури. Вихідні дані включають публічні кліматичні набори даних, зокрема: джерела викидів CO_2, зміни середньої температури, дані про чисельність населення тощо.

4. **Зміст роботи**

У змісті роботи передбачено аналіз наукових джерел щодо застосування методів Data Science у кліматичних дослідженнях, підбір та обробку відповідних публічних наборів даних, побудову моделей прогнозування середньої

температури, аналіз впливу різних факторів на зміну клімату, реалізацію інформаційної системи для прогнозування, візуалізацію результатів дослідження, а також аналіз точності побудованих моделей і формулювання висновків.

5. Перелік графічного матеріалу (слайдів)

Актуальність (1 слайд), постановка задачі (1 слайд), опис вибраних методів (1 слайд), дані для навчання (1 слайд), візуалізація даних (1 слайд), оцінка якості та результати (5 слайдів), інтерфейс (1 слайд), висновок (1 слайд).

6. Календарний план виконання роботи:

№ п/п	Назва частин роботи	%	Виконання роботи	
			За планом	Фактично
1.	Вибір теми кваліфікаційної магістерської роботи, дослідження актуальності обраної теми, наявності наукових матеріалів з теми	3	01.10.24	01.10.24
2.	Протокол кафедри ТУ про затвердження тем дипломних робіт та призначення наукових керівників	2	27.12.24	27.12.24
3.	Формування переліку нормативних матеріалів, літератури з проблематики дипломної роботи	10	18.02.25	18.02.25
4.	Складання розгорнутого плану виконання та представлення кваліфікаційної роботи	5	25.02.25	25.02.25
5.	Ознайомлення наукового керівника з розгорнутим планом кваліфікаційної роботи. Внесення змін.	5	27.02.25	27.02.25
6.	Підготовка розділу 1 «Аналіз предметної галузі та постановка задачі.	10	15.03.25	15.03.25
7.	Підготовка розділу 2 «Інструменти та методи реалізації моделей прогнозування».	14	24.03.25	24.03.25
8.	Підготовка розділу 3 «Розробка методів Data Science».	14	11.04.25	11.04.25

9.	Підготовка розділу 4 «Розгортання інформаційної системи для аналізу кліматичних змін».	13	18.04.25	18.04.25
10.	Оформлення кваліфікаційної роботи. Підготовка аналізу результатів роботи, висновків. Перевірка відповідності початковій меті та задачам роботи	15	02.05.25	02.05.25
11.	Передача кваліфікаційної роботи науковому керівникові	2	03.05.25	03.05.25
12.	Передача кваліфікаційної роботи рецензенту для рецензування	2	09.05.25	09.05.25
13.	Попередній захист кваліфікаційної роботи	5	12.05.25	12.05.25

Дата видачі завдання «____» _____ 2025 р.

Керівник роботи: к.е.н., доцент Мірошниченко І. В.

_____
(підпис)

Завдання прийняв до виконання студент групи ІАВ-21 Українець А. О.

_____
(підпис)

ЗМІСТ

АНОТАЦІЯ

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА
Факультет інформаційних технологій
Кафедра технологій управління
Спеціальність 122 – Комп'ютерні науки,
освітня програма "Інформаційна аналітика та впливи"

Дипломна робота магістра Українця А. О..

Тема роботи – «Розробка моделей та інформаційної системи прогнозування змін клімату за допомогою методів Data Science».

Мета дипломної роботи магістра – розробка ефективних підходів до аналізу та прогнозування кліматичних змін з використанням сучасних інструментів Data Science, а також створення інформаційної системи для інтерактивного представлення результатів дослідження.

Об'єкт дослідження – процеси моделювання, аналізу та прогнозування змін клімату на основі відкритих даних, зокрема про викиди парникових газів.

Предмет дослідження – методи обробки, аналізу та візуалізації кліматичних даних, а також підходи до побудови інформаційної системи для представлення результатів.

Наукова новизна роботи – полягає в поєднанні факторного аналізу впливу різних джерел викидів на зміну середньої температури з прогнозуванням температурних трендів за допомогою глибоких рекурентних нейронних мереж типу LSTM. Побудовано регресійні та нейромережеві моделі, що враховують вплив різних категорій викидів CO_2 від сільського господарства. Крім того, реалізовано вебзастосунок, який надає зручний інтерфейс для візуалізації результатів аналізу й прогнозів у інтерактивній формі.

У роботі проведено огляд наукових джерел щодо проблеми змін клімату та сучасних методів прогнозування. Здійснено попередню обробку та візуалізацію великих обсягів кліматичних і екологічних даних, виконано побудову моделей прогнозування з використанням методів машинного навчання та нейронних

мереж, оцінено якість прогнозів за допомогою метрик RMSE, MAE та R^2, реалізовано вебінтерфейс для візуалізації прогнозів.

Дипломна робота складається зі вступу, чотирьох розділів, висновків і списку використаних джерел. Загальний обсяг – 104 сторінки, кількість джерел – 25.

Ключові слова: зміни клімату, викиди CO_2, машинне навчання, нейронна мережа, прогнозування, LSTM, Data Science, інформаційна система , екологічні дані.

ПЕРЕЛІК ВИКОРИСТАНИХ СКОРОЧЕНЬ

CO_2 – вуглекислий газ

LSTM – Long Short-Term Memory (довготривала короткочасна пам’ять, тип рекурентної нейронної мережі)

RNN – Recurrent Neural Network (рекурентна нейронна мережа)

ML – Machine Learning (машинне навчання)

RF – Random Forest (випадковий ліс)

GB – Gradient Boosting (градієнтний бустинг)

FAOSTAT – Food and Agriculture Organization Statistical Database (Статистична база даних Продовольчої та сільськогосподарської організації ООН)

CSV – Comma-Separated Values (формат збереження табличних даних)

MAE – Mean Absolute Error (середня абсолютна похибка)

RMSE – Root Mean Squared Error (корінь середньоквадратичної похибки)

R^2 – Coefficient of Determination (коефіцієнт детермінації)

HTML – HyperText Markup Language (мова розмітки гіпертексту)

CSS – Cascading Style Sheets (каскадні таблиці стилів)

JS – JavaScript (мова програмування для веброзробки)

ВСТУП

Зміни клімату є однією з найактуальніших глобальних проблем сьогодення. Вони чинять суттєвий вплив на довкілля, соціально-економічні процеси та якість життя населення у всьому світі. Зростання середньої глобальної температури, танення льодовиків, підвищення рівня світового океану, збільшення частоти екстремальних погодних явищ – усе це є наслідками антропогенного впливу на кліматичну систему. З огляду на масштабність і складність цієї проблеми, виникає необхідність у використанні новітніх інформаційних технологій для її дослідження та прогнозування.

Методи Data Science – як-от машинне навчання, статистичний аналіз, обробка великих обсягів даних – відкривають нові можливості для аналізу змін клімату, виявлення ключових факторів впливу, а також створення моделей прогнозування кліматичних показників. Актуальність використання таких методів зумовлена як зростаючим обсягом доступних даних про довкілля, так і потребою в точних та своєчасних прогнозах для прийняття рішень на різних рівнях: від локального до глобального.

Зв’язок теми з науковими програмами полягає в інтеграції сучасних підходів аналізу даних у дослідження змін клімату, що відповідає пріоритетним напрямкам розвитку науки та техніки в галузі охорони довкілля, інформатизації та цифрової трансформації суспільства.

Метою дослідження є розробка методів аналізу та прогнозування змін клімату з використанням сучасних технологій Data Science, а також створити інформаційну систему для візуалізації прогнозів температури та ключових факторів, що впливають на клімат у різних країнах світу.

Завдання дослідження:

1. Провести огляд наукових джерел щодо змін клімату та способів їхнього прогнозування;
2. Зібрати, обробити та проаналізувати кліматичні дані та дані про джерела викидів парникових газів;

3. Виконати кореляційний та регресійний аналіз для виявлення основних чинників, що впливають на зміну середньої температури;
4. Розробити прогнозні моделі із застосуванням методів машинного навчання, зокрема LSTM-мереж;
5. Розробити інформаційну систему (вебсайт) для представлення результатів аналізу та прогнозів у доступному вигляді для кожної країни.

Об’єктом дослідження є процеси моделювання, аналізу та прогнозування кліматичних змін на основі даних про викиди парникових газів від сільськогосподарської промисловості в різних країнах світу.

Предметом дослідження є методи обробки, аналізу та візуалізації кліматичних даних за допомогою інструментів Data Science, а також підходи до створення інформаційної системи для доступу до результатів дослідження.

Методи дослідження:

1. Методи статистичного аналізу – для оцінювання впливу різних факторів на зміну температури;
2. Машинне навчання (нейронні мережі, LSTM) – для побудови моделей прогнозування середньої температури;
3. Методи візуалізації даних – за допомогою бібліотек Matplotlib, Seaborn, Plotly для представлення результатів аналізу;
4. Інструменти розробки вебзастосунків – для реалізації вебсайту, що дозволяє переглядати прогнози температури та найвпливовіші чинники в різних країнах.

Наукова новизна одержаних результатів: Запропоновано підхід до поєднання аналізу факторів, що впливають на зміну температури, з прогнозними моделями на основі нейронних мереж типу LSTM; Проведено аналіз залежності між викидами CO_2 від сільськогосподарської діяльності в різних країнах та зміною середньої температури; Розроблено інтерактивну інформаційну систему, що надає можливість досліджувати прогноз зміни температури по країнах та виявляти ключові фактори впливу в кожному окремому випадку.

Практичне значення одержаних результатів: Результати дослідження можуть бути використані науковими установами, екологічними організаціями, аналітичними центрами та органами державної влади для оцінки ризиків, пов’язаних зі змінами клімату, та для формування ефективної кліматичної політики. Розроблений вебсайт дозволяє зручно переглядати результати моделювання, що підвищує доступність дослідження для широкого кола користувачів.

Апробація результатів роботи: Публікація тез “Artificial Neural Network for Classification of Images of Plant Flowers” у “X International Conference Information Technology and Implementation (Satellite) 2023” та “Development of Data Science Technologies in the Field of Climate Change Research” у “XI International Conference Information Technology and Implementation (Satellite) 2024”.

Публікації: S.M. Bilan, A.O. Ukrainets. Artificial Neural Network for Classification of Images of Plant Flowers. Information Technology and Implementation (Satellite): Conference Proceedings, November 21, 2023, Kyiv, Ukraine / Ministry of Education and Science of Ukraine, Taras Shevchenko National University of Kyiv and [etc]; Vitaliy Snytyuk (Editor). – Kyiv: Publishing House «Caravela», 2023 [1]. Ihor Miroshnychenko, Andrii Ukrainets Development of Data Science Technologies in the Field of Climate Change Research. Information Technology and Implementation (Satellite): Conference Proceedings, November 21, 2024, Kyiv, Ukraine / Ministry of Education and Science of Ukraine, Taras Shevchenko National University of Kyiv and [etc]; Vitaliy Snytyuk (Editor). – Kyiv: Publishing House «Caravela», 2024 [2].

РОЗДІЛ 1
АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ ТА ПОСТАНОВКА ЗАДАЧІ

1.1 Опис предметної області

Зміна клімату є однією з найгостріших глобальних проблем сучасності, яка впливає на навколишнє середовище, економіку, суспільство та здоров'я населення. Підвищення середньорічних температур, танення льодовиків, підвищення рівня світового океану, а також частіші екстремальні погодні явища (засухи, повені, бурі) – усе це є проявами кліматичних змін, які дедалі більше турбують наукову спільноту, уряди країн і міжнародні організації. Усвідомлення загроз, пов'язаних зі зміною клімату, зумовлює необхідність системного вивчення цього явища та пошуку ефективних рішень щодо його стримування та адаптації до наслідків [3].

У цьому контексті особливої ваги набуває якісний аналіз кліматичних даних, що дає змогу глибше зрозуміти причини та динаміку змін, а також розробляти науково обґрунтовані сценарії майбутнього. Для цього необхідно обробляти величезні обсяги різноманітної інформації, зокрема метеорологічних спостережень, супутникових вимірювань, даних про рівень викидів парникових газів, землекористування, антропогенну діяльність тощо. Оскільки ці дані можуть значно відрізнятися за джерелами, форматом та частотою оновлення, їх аналіз традиційними методами є складним. У зв'язку з цим, ключову роль у дослідженнях змін клімату починають відігравати сучасні цифрові технології, серед яких особливо важливими є технології аналізу даних – Data Science.

Data Science – це міждисциплінарна галузь, що поєднує статистику, інформатику, штучний інтелект і знання про предметну область для отримання цінної інформації з даних. У контексті дослідження кліматичних змін технології Data Science дозволяють не лише виявляти закономірності в історичних кліматичних даних, але й будувати моделі для прогнозування майбутніх змін, що є надзвичайно важливим для прийняття обґрунтованих рішень у сфері екологічної політики, містобудування, сільського господарства та інших галузей.

Серед основних методів, що використовуються в Data Science для аналізу кліматичних даних, вирізняються:

- Статистичні методи, які дозволяють вивчати розподіли даних, виявляти тренди, аномалії та взаємозв’язки між різними змінними;
- ML методи, які забезпечують побудову прогнозних моделей, здатних автоматично навчатися на нових даних і виявляти складні, нелінійні залежності.

1.1.1 Статистичний аналіз

Статистичний аналіз – це метод вивчення та аналізу даних, спрямований на виявлення їхньої структури, властивостей та взаємозв'язків. Цей метод дозволяє робити висновки про популяцію або великий набір даних на основі вивчення лише частини цих даних – вибірки. Статистичний аналіз є основою для глибокого розуміння природних процесів і явищ, таких як зміни клімату, де кожен набір даних може містити величезну кількість інформації, що потребує належного оброблення та інтерпретації. Основні цілі статистичного аналізу включають виявлення закономірностей у даних, перевірку гіпотез, прогнозування майбутніх значень та підтримку прийняття рішень на основі даних [4].

Типи статистичного аналізу:

1. Описова статистика: Використовується для узагальнення та опису основних характеристик набору даних, таких як середнє значення, медіана, дисперсія, стандартне відхилення та інші параметри. Вона допомагає зрозуміти загальні тенденції в даних, що є особливо важливим для виявлення сезонних коливань температури чи змін в рівнях викидів парникових газів.

2. Інференційний статистичний аналіз: Застосовується для здійснення висновків про більшу популяцію на основі даних, зібраних з вибіркової групи. Це дозволяє дослідникам робити прогнози про кліматичні тенденції на основі вибіркових даних, таких як температурні показники з різних регіонів чи країнах. Техніки, як t-тести або аналіз дисперсії, можуть використовуватися для

виявлення відмінностей між різними групами (наприклад, різними роками чи сезонами).

3. Статистика асоціацій: Використовується для виявлення взаємозв'язків між категоріальними змінними в наборі даних. У контексті кліматичних досліджень це може включати виявлення зв’язків між змінами температури та антропогенними факторами, такими як рівень викидів парникових газів або зміни в землекористуванні.

Процедура статистичного аналізу включає наступні кроки: збір даних, підготовка даних, вибір статистичних методів, аналіз даних та висновки й інтерпретація даних.

Завдяки таким методам статистичного аналізу, дослідники мають змогу не лише зрозуміти поточний стан кліматичних змін, а й прогнозувати майбутні сценарії, що є необхідним для ефективного управління природними ресурсами та адаптації до змін клімату.

1.1.2 Машинне навчання

Машинне навчання – це галузь штучного інтелекту, що досліджує розробку алгоритмів, здатних навчатися на основі даних і робити прогнози або приймати рішення без потреби в явному програмуванні. Це один із ключових інструментів реалізації інтелектуальної поведінки в інформаційних системах, що дозволяє створювати моделі, здатні виявляти закономірності у великих масивах даних і використовувати ці знання для подальшого самонавчання та вдосконалення [5].

Сучасні методи машинного навчання широко застосовуються в різних галузях, включаючи медицину, фінанси, транспорт, виробництво, а також дослідження кліматичних змін. Завдяки здатності працювати з великими обсягами даних, інструменти машинного навчання дозволяють вирішувати завдання класифікації, кластеризації, виявлення аномалій, створення рекомендаційних систем, обробки природної мови, комп’ютерного зору, прогнозування часових рядів тощо.

Типи машинного навчання:

1. Контрольоване навчання: Один з найпоширеніших типів, за якого модель навчається на основі навчального набору даних, що містить вхідні значення (ознаки) та відповідні їм правильні виходи (мітки). Це дозволяє будувати моделі для класифікації (наприклад, визначення типу кліматичного явища) або регресії (наприклад, прогнозування температури чи рівня викидів).

2. Неконтрольоване навчання: Застосовується у випадках, коли дані не мають міток, і мета полягає у виявленні внутрішніх структур, кластерів або прихованих залежностей у даних. У сфері аналізу клімату це може бути використано для групування схожих кліматичних регіонів або виявлення аномальних погодних патернів.

3. Навчання з підкріпленням: Це метод, при якому агент взаємодіє з середовищем, виконує певні дії та отримує зворотний зв’язок у вигляді винагород або штрафів. Цей підхід особливо актуальний у задачах управління динамічними системами, зокрема – у моделюванні кліматичних сценаріїв або оптимізації використання ресурсів.

Процедура машинного навчання включає наступні кроки: збір даних, підготовка даних, вибір моделі, тренування моделі, тестування моделі та висновки.

Застосування ML у сфері кліматичних досліджень дає змогу покращити точність прогнозування кліматичних змін, виявляти критичні фактори, що впливають на глобальне потепління, а також моделювати майбутні сценарії розвитку подій. Це робить машинне навчання важливим інструментом у формуванні стратегії адаптації та мінімізації впливу змін клімату на навколишнє середовище.

1.2 Сучасні методи прогнозування

Однією з основних задач Data Science є прогнозування – процес передбачення майбутніх значень або подій на основі аналізу історичних та поточних даних. Це завдання є критично важливим у багатьох сферах, таких як

фінанси, охорона здоров'я, енергетика, транспорт, маркетинг і, зокрема, – у дослідженні змін клімату.

Прогнозування дає змогу виявляти тренди, моделювати сценарії майбутнього, своєчасно реагувати на потенційні ризики та приймати обґрунтовані управлінські рішення. Наприклад, у контексті кліматичних змін прогнозування дозволяє оцінити можливі коливання температур, рівня опадів, концентрації парникових газів, частоти екстремальних погодних явищ тощо.

Залежно від типу даних, структури проблеми та цілей аналізу, застосовуються різні методи прогнозування, які можна умовно поділити на такі основні групи:

- Методи регресії, які використовуються для передбачення числових (неперервних) значень, наприклад, середньої температури або рівня CO_2 в атмосфері.
- Методи класифікаційного моделювання, які застосовуються у випадках, коли потрібно передбачити категоріальні (дискретні) події або стани, наприклад, класифікувати кліматичні умови як "нормальні" чи "аномальні".
- Нейронні мережі, зокрема глибокі нейронні архітектури, що демонструють високу точність у складних прогнозних задачах і можуть обробляти великі обсяги даних різної природи.
- Аналіз часових рядів, орієнтований на виявлення закономірностей у зміні даних з часом, зокрема сезонності, трендів і циклів.
- Видобуток даних (data mining) – широка сукупність методів для автоматизованого виявлення знань у великих наборах даних, що поєднує статистику, ML та бази даних.

Кожен з методів має свої переваги та обмеження і може бути адаптований під специфіку задачі. Правильний вибір методу часто визначає успішність прогнозної моделі та точність її результатів. У наступних підпунктах розглянемо кожну з цих груп методів більш детально [6].

1.2.1 Методи регресії

Методи регресії є статистичними підходами, що використовуються для побудови моделей, які описують залежність між однією залежною змінною (цільовою змінною) та однією або декількома незалежними змінними (факторами або предикторами). Ці методи дозволяють не лише виявляти зв'язки між змінними, а й робити кількісні прогнози майбутніх значень залежної змінної на основі нових даних. Регресійне моделювання широко застосовується для прогнозування, оцінки ризиків та підтримки прийняття рішень у таких галузях, як економіка, медицина, соціологія, екологія та інші.

До найпоширеніших видів регресії належать: лінійна (рис. 1.1), логістична, поліноміальна.

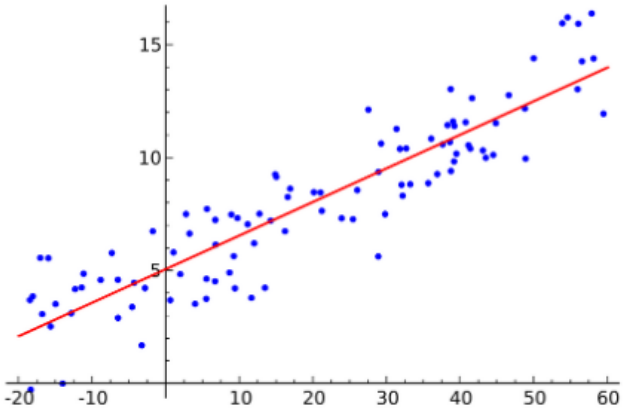

Рисунок 1.1 – приклад лінійної регресії

Переваги методів: простота реалізації та інтерпретації результатів; ефективність при наявності лінійних або близько лінійних зв'язків між змінними; можуть використовуватися з невеликими наборами даних.

Недоліки методів: обмеженість у випадках складних або нелінійних залежностей; чутливість до викидів та мультиколінеарності; потреба у ретельному виборі форми моделі та її параметрів; мають широке застосування в економіці, екології, медицині, кліматичних дослідженнях.

У контексті прогнозування кліматичних змін регресійні методи можуть застосовуватись для моделювання впливу таких факторів, як місяць року, географічне положення, висота над рівнем моря, близькість до водойм, рівень

урбанізації тощо, на середню температуру в регіоні. Побудована на тренувальних даних модель може з високою точністю прогнозувати температуру за новими вхідними характеристиками.

Таким чином, методи регресії є потужним інструментом для побудови пояснюваних моделей, які дозволяють здійснювати точні кількісні прогнози та проводити аналіз впливу окремих факторів на цільову змінну.

1.2.2 Методи класифікації

Методи класифікаційного моделювання – це статистичні та машинні підходи, які використовуються для визначення категорії або класу, до якого належить новий об’єкт даних, на основі його характеристик або ознак. На відміну від регресії, де результат є числовим, у класифікації прогнозується дискретне значення – клас або мітка. Такі методи дозволяють автоматизувати процес прийняття рішень, аналізуючи історичні дані та виявляючи приховані закономірності в них.

Приклади таких методів: дерева рішень (рис 1.2), випадкові ліси, наївний байєсівський класифікатор, метод опорних векторів, k-найближчих сусідів.

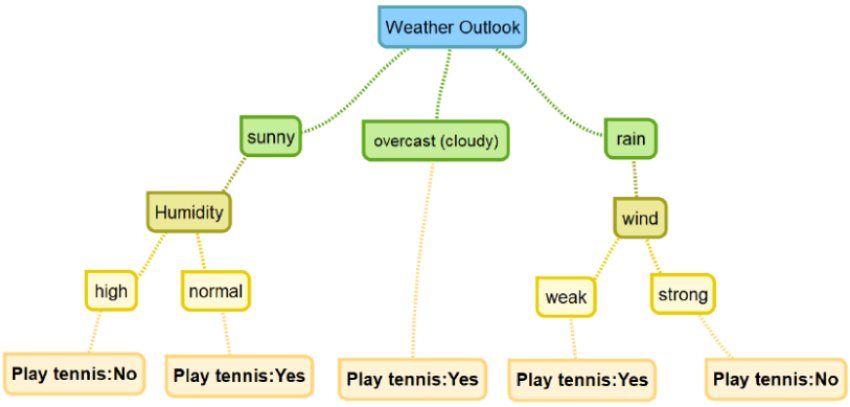

Рисунок 1.2 – приклад дерева рішень

Переваги методів: добре підходять для задач з чітко визначеними класами; можуть працювати навіть з обмеженим обсягом навчальних даних; часто забезпечують високу точність; застосовуються для задач виявлення шахрайства, медичної діагностики, класифікації зображень, спаму тощо.

Недоліки методів: можуть бути менш точними, ніж інші методи для складних задач; можуть бути чутливими до шуму в даних; можуть бути складними для інтерпретації.

У контексті досліджень змін клімату, методи класифікаційного моделювання широко застосовуються для: виявлення аномальних температурних змін та класифікації погодних умов; прогнозування ймовірності виникнення екстремальних кліматичних явищ (наприклад, спеки, повеней, засух); класифікації типів кліматичних зон на основі метеорологічних параметрів; виявлення змін у поведінці природних систем під впливом зростання викидів парникових газів.

Наприклад, маючи дані про температуру, рівень опадів, концентрацію CO_2 та інші екологічні змінні, можна побудувати класифікаційну модель, яка передбачатиме ймовірність того, що в певному регіоні протягом найближчих років спостерігатимуться аномальні кліматичні події. Це дозволяє формувати попереджувальні системи, а також розробляти ефективні стратегії адаптації до змін клімату.

Завдяки своїй гнучкості та здатності працювати з різними типами даних, методи класифікації є важливим інструментом у сучасному аналізі кліматичних процесів і допомагають приймати обґрунтовані рішення в екологічній політиці та управлінні ризиками.

1.2.3 Нейронні мережі

Нейронна мережа – це система, яка імітує людський мозок, ґрунтуючись на сукупності з’єднаних штучних нейронів (рис. 1.3). Ці нейрони взаємодіють між собою, обробляючи інформацію та передаючи сигнал для подальшої обробки іншим нейронам. Кожен штучний нейрон розташований у своєму шарі, де він виконує специфічну обробку даних. Нейрони одного шару можуть передавати свої виходи як вхідні сигнали нейронам наступного шару. Це дозволяє нейронним мережам здійснювати складні перетворення вхідних даних через багатошарові структури.

Приклади таких мереж: штучна нейронна мережа, згорткова, рекурентна.

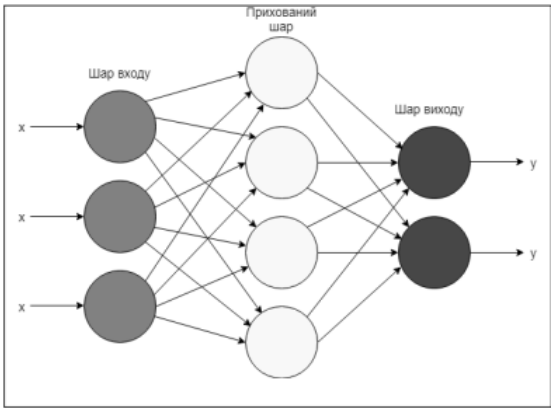

Рисунок 1.3 – проста штучна нейронна мережа

Переваги нейронних мереж: висока точність навіть для складних і високовимірних задач; здатність моделювати нелінійні залежності та складні взаємозв'язки між змінними; можуть працювати з різними типами даних: числовими, текстовими, графічними; добре підходять для роботи з великими обсягами даних.

Недоліки нейронних мереж: складність побудови, навчання та інтерпретації; необхідність великої кількості навчальних даних; потреба у значних обчислювальних ресурсах; можуть бути схильні до перенавчання або недонавчання.

У сфері дослідження кліматичних змін нейронні мережі відіграють ключову роль завдяки своїй здатності моделювати складні процеси та залежності. Наприклад: прогнозування температури на основі історичних кліматичних рядів; аналіз супутникових зображень для виявлення змін у земному покриві; оцінка впливу різних факторів (викиди CO_2, урбанізація, вирубка лісів) на зміну клімату; моделювання ризиків виникнення екстремальних погодних явищ (засухи, бурі, повені).

Завдяки здатності до самонавчання та гнучкості у роботі з великими наборами даних, нейронні мережі стають потужним інструментом у розробці систем підтримки прийняття рішень, стратегій адаптації та боротьби з наслідками глобального потепління.

1.2.4 Аналіз часових рядів

Аналіз часових рядів – це статистичний підхід до обробки, моделювання та інтерпретації даних, які фіксуються у часовій послідовності з певним інтервалом (день, місяць, рік тощо). Такий тип даних широко застосовується в різних сферах – від економіки до метеорології, оскільки дозволяє простежити динаміку змін досліджуваного явища у часі. Аналіз часових рядів дає змогу не лише виявляти загальні тенденції (тренди), періодичні коливання (сезонність) чи випадкові відхилення, але й будувати прогностичні моделі, які допомагають передбачити майбутню поведінку системи на основі минулих спостережень. Завдяки цьому аналіз часових рядів є потужним інструментом у задачах довгострокового та короткострокового прогнозування, а також для прийняття рішень на основі даних. На рисунку 1.4 наведено приклад прогнозу часового ряду.

Приклад таких методів: авторегресійні моделі, моделі з ковзним вікном, експоненційне згладжування.

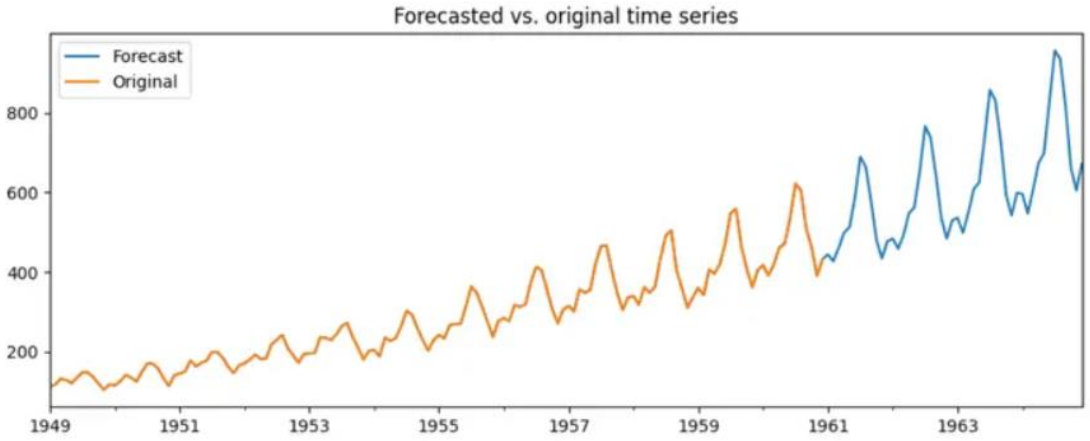

Рисунок 1.4 – приклад прогнозу часового ряду

Переваги методів: виявлення трендів, сезонності та циклічних змін; здатність прогнозувати майбутні значення з урахуванням історичних залежностей; можливість моделювання навіть складних і нерівномірних часових структур.

Недоліки методів: потребує стаціонарності даних або складних перетворень для її досягнення; висока чутливість до викидів і шуму; складність вибору оптимальної моделі та параметрів.

Аналізуючи дані за декілька років, можна виявити загальний тренд зміни температури. Наприклад, якщо середня температура поступово зростає з року в рік, це може свідчити про глобальне потепління. Цей тренд є важливим індикатором для прогнозування майбутніх кліматичних змін, оцінки екологічних ризиків і прийняття ефективних рішень у сфері збереження навколишнього середовища. Крім того, аналіз часових рядів дозволяє досліджувати вплив антропогенних чинників, оцінювати сезонні коливання температури, рівня опадів або концентрацій парникових газів, а також виявляти аномальні явища, які можуть сигналізувати про глобальні кліматичні зрушення. Таким чином, цей підхід є незамінним інструментом у дослідженнях клімату, розробці адаптаційних стратегій та підтримці сталого розвитку.

1.2.5 Видобуток даних

Видобуток даних (data mining) – це процес виявлення прихованих закономірностей, тенденцій і корисної інформації з великих обсягів даних за допомогою автоматизованих методів аналізу. Основною метою є отримання нових знань, які раніше не були очевидними, та формування на їх основі практичних висновків.

Важливою складовою видобутку даних є попередня підготовка: очищення даних від помилок, обробка пропущених значень, нормалізація та перетворення даних у зручний для аналізу формат. Після цього застосовуються різні алгоритми для класифікації, кластеризації, пошуку аномалій, виявлення асоціативних правил, а також побудови прогностичних моделей.

Результати видобутку даних активно використовуються у бізнесі, медицині, фінансах, маркетингу, а також у наукових дослідженнях, зокрема для моделювання кліматичних процесів.

Приклади методів: асоціативні правила, кластеризація, аналіз тексту.

Переваги: можливість виявлення прихованих закономірностей у великих масивах даних; підвищення ефективності прийняття рішень; прогнозування майбутніх подій чи змін; виявлення аномальних або критичних ситуацій.

Недоліки: висока вимогливість до якості даних; складність інтерпретації результатів; потреба у значних обчислювальних ресурсах; ризики, пов’язані з конфіденційністю та етикою обробки даних.

Одним із прикладів застосування видобутку даних у кліматичних дослідженнях є аналіз великих обсягів даних про температуру поверхні океану, концентрацію парникових газів, атмосферний тиск, опади та інші параметри, зібрані за допомогою супутників, метеостанцій і сенсорних мереж. Такі дані дозволяють виявляти довготривалі кліматичні тренди, оцінювати наслідки людської діяльності, виявляти аномальні явища (наприклад, посухи, хвилі тепла або підвищену частоту штормів), а також розробляти адаптаційні та превентивні заходи для зменшення впливу змін клімату. Видобуток даних відіграє ключову роль у трансформації «сирих» даних у цінну інформацію, що сприяє глибшому розумінню глобальних процесів та формуванню сталих екологічних рішень.

1.3 Аналіз наукових джерел щодо застосування методів прогнозування у вивченні змін клімату

Методи прогнозування відіграють ключову роль у різних сферах діяльності, дозволяючи аналізувати наявні дані та передбачати майбутні події. Вони використовуються для зменшення ризиків, оптимізації процесів та ухвалення обґрунтованих рішень у фінансах, маркетингу, метеорології, медицині, виробництві та багатьох інших галузях. Важливою темою є також застосування методів прогнозування для аналізу та прогнозування змін клімату. У цьому розділі розглянемо наукові публікації, що досліджують використання різних методів прогнозування для вирішення проблем, пов'язаних із глобальними кліматичними змінами.

У статті “Air Temperature Forecasting Using Machine Learning Techniques: A Review” представлено огляд методів машинного навчання для прогнозування

температури повітря. Розглянуто моделі MLPNN, RBFNN, CNN, LSTM, RNN та SVM. SVM показав високу точність на глобальному рівні (MSE = 0.00452 °K), перевершивши MLPNN (MSE = 0.08912 °K). На регіональному рівні найкращі результати продемонстрували глибокі нейронні мережі (LSTM, CNN), з MSE = 0.0017 °K для одноетапного прогнозу. Це підтверджує ефективність SVM та LSTM для задач прогнозування температури [7].

У статті “Transport and Climate Change: A Review” розглядається вплив транспорту на зміну клімату та можливі способи зменшення викидів парникових газів. Автори аналізують три основні проблеми: використання автомобілів, вантажні перевезення та авіацію. Хоча технологічні інновації, такі як нові види палива та покращення енергоефективності, важливі, для стабілізації викидів CO_2 необхідні також політичні заходи і зміни в поведінці споживачів. Висновки підкреслюють важливість короткотермінових заходів, таких як заохочення екологічних автомобілів, велосипедних маршрутів і пішохідних зон, а також міжнародних угод для обмеження зростання авіації [8].

У статті “Greenhouse gases emissions and global climate change: Examining the influence of CO_2, CH4, and N2O” розглядається вплив основних парникових газів (CO_2, CH_4 та N_2O) на глобальне потепління та зміни клімату. Автори підкреслюють важливість зменшення викидів цих газів для стримування зміни клімату, оскільки їх концентрація в атмосфері значно впливає на глобальну температуру. Викиди CO_2, зокрема, є основним фактором, що викликає підвищення температури. Дослідження також показують важливість ефективних політичних заходів та технологічних інновацій для зниження рівня викидів. Висновки вказують на необхідність глобальної співпраці для досягнення сталого зниження викидів парникових газів, що дозволить уповільнити темпи змін клімату та їхні негативні наслідки [9].

У статті “Predicting future global temperature and greenhouse gas emissions via LSTM model” автори досліджують вплив викидів парникових газів на глобальну температуру, використовуючи модель LSTM для прогнозування майбутніх змін температури та концентрацій CO_2 на основі історичних даних. Результати

показали, що до 2100 року глобальна температура може підвищитися на 4.8°C, а концентрація CO_2 досягне 850 ppm, якщо не вживатимуться значні заходи для зменшення викидів. Висновки наголошують на важливості застосування ML для точних прогнозів змін клімату та необхідності активних дій для обмеження викидів парникових газів, щоб уникнути катастрофічних наслідків для планети [10].

Стаття “Short and mid-term sea surface temperature prediction using time-series satellite data and LSTM-AdaBoost combination approach” розробляє метод прогнозування температури поверхні моря (SST) на короткострокову та середньострокову перспективу, поєднуючи LSTM та AdaBoost. Автори використовують часові ряди супутникових даних і застосовують AdaBoost для покращення точності прогнозів. Результати показали високу точність цього підходу, що може бути корисним для морських досліджень та управління морськими ресурсами [11].

Стаття “A spatiotemporal CNN-LSTM deep learning model for predicting soil temperature in diverse large-scale regional climates” розробляє модель, яка поєднує згорткову нейронну мережу (CNN) та LSTM для прогнозування температури ґрунту на глибину 0–7 см у різних кліматичних зонах Канади та США. Модель була натренована на щогодинних даних температури ґрунту та оцінена в п'яти кліматичних зонах. Модель перевершила методи випадкового лісу та опорних векторів регресії за точністю прогнозів, досягнувши низької середньоквадратичної помилки (NRMSE = 2.3%) та високих коефіцієнтів детермінації (R^2 = 95%). Це робить модель перспективною для застосування у сільському господарстві, гідрології та адаптації до змін клімату [12].

Стаття “Weather forecast using LSTM networks” описує створення моделі прогнозування погоди за допомогою LSTM на основі даних з міста Єна, Німеччина. Модель використовує шість ознак: температура, тиск, вологість, парціальний тиск водяної пари, швидкість вітру та щільність повітря. Після нормалізації даних створюються навчальний і валідаційний набори. Модель з 32 одиницями пам'яті тренується на 15 епохах з середньоквадратичною похибкою

як функцією втрат. Після навчання досягнуто training loss 0.11, що свідчить про високу точність прогнозування та ефективність використання LSTM для прогнозування погодних умов [13].

Стаття “Monthly climate prediction using deep convolutional neural network and long short-term memory” описує застосування гібридної моделі CNN-LSTM для місячного прогнозування кліматичних факторів у місті Цзінань, Китай, на основі даних за 72 роки (1951–2022). Модель прогнозує середню температуру, екстремальні температури, опади, вологість і тривалість сонячного сяйва. Результати показали, що CNN-LSTM перевершує окремі моделі LSTM, RNN, ANN та CNN за точністю, зокрема для середньої температури з RMSE 0.629°C. Цей підхід демонструє ефективність глибокого навчання для поліпшення кліматичних прогнозів, що може бути корисним для запобігання стихійним лихам і управління водними ресурсами [14].

1.4 Постановка задачі

На основі аналізу сучасних підходів до прогнозування та огляду наукових публікацій і досліджень у сфері зміни клімату було обрано найбільш доцільні методи для виконання поставлених завдань. Зокрема, для оцінки впливу різних джерел викидів CO_2 від сільськогосподарської промисловості на зміну температури було вирішено застосувати методи машинного навчання RF і G, які дозволяють виявити найважливіші фактори завдяки високій точності та здатності працювати з великими обсягами даних. Для прогнозування температури у вигляді часових рядів обрано нейронну мережу LSTM, що ефективно працює з послідовними даними та дозволяє моделювати складні нелінійні залежності.

Метою дослідження є розробка методів аналізу та прогнозування змін клімату з використанням сучасних технологій Data Science, а також створити інформаційну систему для візуалізації прогнозів температури та ключових факторів, що впливають на клімат у різних країнах світу.

Для досягнення поставленої мети необхідно виконати такі завдання:

- Дослідити різні методи прогнозування.
- Проаналізувати різні методи моделювання, застосовувані в інших дослідженнях.
- Зібрати дані про викиди CO_2 від сільськогосподарської діяльності та температуру по кожній країні;
- Здійснити попередню обробку даних: очистити від пропусків, нормалізувати;
- Провести візуалізацію та первинний аналіз даних;
- Здійснити кореляційний аналіз для виявлення взаємозв'язків між джерелами викидів і температурою;
- Побудувати регресійні моделі для виявлення найбільш впливових джерел викидів;
- Побудувати модель LSTM для прогнозування змін температури на основі часових рядів;
- Перевірити точність та ефективність побудованих моделей;
- Здійснити комплексний аналіз отриманих результатів.
- Розробити інформаційну систему (вебсайт) для представлення результатів аналізу та прогнозів.

1.5 Висновок до розділу

У цьому розділі було здійснено детальний огляд сучасних підходів до прогнозування змін клімату, зокрема аналіз основних методів, що використовуються для моделювання кліматичних процесів і прогнозування температурних змін. Цей огляд допоміг обрати найбільш доцільні методи для подальшого глибокого аналізу та практичного застосування. Вивчення наукових публікацій і результатів досліджень виявило високу ефективність методів машинного навчання, зокрема Random Forest та Gradient Boosting, для виявлення ключових факторів, що суттєво впливають на зміну температури, завдяки їх здатності працювати з великими обсягами даних та виявляти складні взаємозв'язки між різними змінними. Додатково, переваги використання

нейронних мереж типу LSTM були підтверджені їхньою здатністю працювати з часовими рядами та моделювати складні нелінійні залежності, що дозволяє досягати високої точності при прогнозуванні температурних змін на основі історичних даних.

З метою реалізації головної мети дослідження було сформульовано конкретні завдання, серед яких: збір та обробка кліматичних даних, проведення аналізу взаємозв'язків між різними джерелами викидів CO_2 та зміною температури, побудова регресійних моделей для виявлення найбільш впливових чинників, а також розробка інформаційної системи для візуалізації отриманих результатів та прогнозів. Окрему увагу було приділено важливості попередньої обробки даних, включаючи очищення від пропусків та нормалізацію, що є основою для проведення точного аналізу. Кореляційний та регресійний аналіз допоможе ідентифікувати ключові фактори, що впливають на температуру, а побудова прогнозних моделей дозволить забезпечити надійні прогнози для майбутніх змін клімату.

Таким чином, на основі проведеного аналізу та чітко визначених завдань, наступним кроком стане детальне дослідження кожного методу окремо, вибір найбільш підходящих інструментів й мови програмування для їх реалізації, а також перевірка доступних технічних засобів для побудови ефективних моделей прогнозування. Це дозволить досягти головної мети дослідження – створення надійних моделей для прогнозування зміни клімату, які можуть стати важливим інструментом для прийняття обґрунтованих рішень на державному та міжнародному рівнях, сприяючи більш точному розумінню та адаптації до зміни клімату в різних країнах світу.

РОЗДІЛ 2
ІНСТРУМЕНТИ ТА МЕТОДИ РЕАЛІЗАЦІЇ МОДЕЛЕЙ ПРОГНОЗУВАННЯ

2.1 Аналіз обраних методів

Для досягнення поставленої мети дослідження, що полягає у побудові ефективних моделей прогнозування змін клімату на основі аналізу впливових факторів, було обрано сучасні ML методи, здатні працювати з великими обсягами даних, враховувати складні залежності між змінними та забезпечувати високу точність прогнозування. Серед великої кількості доступних підходів було виділено три методи, які зарекомендували себе як найбільш результативні для задач, пов'язаних із аналізом кліматичних змін.

До обраних методів належать: випадковий ліс (RF), який є ансамблевим методом на основі дерева рішень, градієнтний бустинг (GB), що дозволяє послідовно покращувати якість моделі, а також нейронна мережа типу LSTM (Long Short-Term Memory), що є різновидом рекурентних нейронних мереж, здатних ефективно обробляти часові ряди. Кожен із зазначених методів має свої переваги, недоліки та сфери застосування, які будуть розглянуті в наступних підпунктах.

2.1.1 Метод випадкового лісу

Випадковий ліс (RF) – це потужний ансамблевий метод машинного навчання, який поєднує велику кількість дерев рішень для вирішення задач класифікації та регресії. Основна ідея методу полягає в тому, що комбінація багатьох слабких моделей (окремих дерев) може забезпечити кращу узагальнюючу здатність, ніж одна сильна модель [15].

RF належить до класу методів ансамблевого навчання, зокрема – до підкласу методів бутстреп-агрегування (bagging). У цьому підході кожне дерево тренується на випадковій підмножині навчального набору даних, відібраній з поверненням (методом bootstrap), і при цьому на кожному вузлі дерева вибирається випадкова підмножина ознак для поділу. Така стратегія дозволяє

створити колекцію моделей, які є слабо скорельованими між собою, що в результаті підвищує загальну точність і стійкість моделі.

У 2001 році Лео Брейман представив більш формалізовану версію методу, яку сьогодні прийнято називати RF. У своїй роботі він поєднав метод bagging (bootstrap aggregating) з випадковим вибором підмножини ознак при кожному розгалуженні дерева, що значно знижує кореляцію між окремими деревами в ансамблі й підвищує загальну точність моделі.

Основні характеристики методу:

- Стохастичність: як підмножини зразків, так і підмножини ознак обираються випадковим чином, що робить модель менш чутливою до шуму.
- Ансамблювання: результат прогнозування отримується шляхом об’єднання результатів (усереднення або голосування) окремих дерев.
- Висока узагальнююча здатність: завдяки зменшенню варіації та використанню незалежних моделей.

Основні кроки роботи випадкового лісу:

1. Вибір випадкової підмножини даних: Для кожного дерева випадкового лісу випадковим чином вибирається підмножина з навчального набору з поверненням (тобто деякі зразки можуть повторюватися, а деякі – не потрапити до підмножини зовсім). Це забезпечує різноманітність серед дерев, що покращує загальну узагальнюючу здатність моделі.

2. Побудова дерева рішень: Під час побудови кожного дерева, на кожному вузлі, випадковим чином вибирається підмножина ознак (feature subset). Серед цих ознак визначається найкраща – та, яка найкраще розділяє дані відповідно до певного критерію (наприклад, мінімізація середньоквадратичної помилки або індексу Джині). Такий підхід знижує кореляцію між деревами.

3. Завершення дерев: Кожне дерево росте до максимальної можливої глибини без обрізки, або до тих пір, поки в листках не залишиться задана мінімальна кількість зразків. Це дозволяє деревам максимально пристосуватись

до своїх бутстреп-вибірок, хоча в сукупності з іншими деревами це не призводить до перенавчанню.

4. Прогнозування: Для регресії – усереднюють прогнози всіх дерев, для класифікації – обирається клас, який отримав найбільше голосів серед усіх дерев.

Random Forest можна математично описати як ансамбль дерев рішень. Нехай у нас є набір тренувальних даних D=$\{(x_1, y_1),(x_2, y_2),...,(x_n, y_n)\}$, де x_i – вектор ознак, а y_i – відповідне значення цільової змінної.

Bootstrap sampling: Для кожного дерева T_k з множини дерев, вибирається випадкова підмножина даних D_k з вибіркою з поверненням з початкового набору D. Це означає, що деякі зразки можуть бути вибрані кілька разів, а інші – ні. Такий підхід дозволяє кожному дереву навчатися на різних підмножинах даних, зменшуючи кореляцію між деревами.

Побудова дерева: Для кожного дерева T_k будується дерево рішень на основі підмножини D_k. Кожне дерево розвивається до максимальної глибини без обрізки або до досягнення мінімальної кількості зразків у листках.

Розділення вузлів: Для кожного вузла дерева обирається випадкова підмножина ознак m, і серед них обирається ознака з найкращим розділенням (наприклад, мінімізація MSE або критерію Джині). Це дозволяє знизити кореляцію між деревами і покращити їх ефективність.

$$MSE = \frac{1}{|D_1|}\sum_{i\in D_1}(y_i - \bar{y}_1)^2 + \frac{1}{|D_2|}\sum_{i\in D_2}(y_i - \bar{y}_2)^2 \quad (2.1)$$

де D_1 та D_2 – підмножини даних після поділу;

y_i – значення цільової змінної у вузлі;

$\bar{y}_1$ та $\bar{y}_2$ – середні значення цільових змінних у для кожної підмножини.

Розбиття вузла: Вибраний поріг розділяє зразки на дві частини – одну для лівого піддерева, іншу для правого піддерева. Цей процес повторюється рекурсивно до тих пір, поки не буде досягнуто певного критерію зупинки (наприклад, максимальна глибина дерева або мінімальна кількість зразків у листку).

Прогноз для нового зразка: Коли дерево побудоване, для нового зразка x прогнозується його цільова змінна $T_k(x)$. Для цього зразок проходить через всі вузли дерева, де на кожному етапі вибирається шлях в залежності від значення ознак і порогів.

Прогноз у листку: Коли зразок досягає листка, дерево видає прогноз для цього зразка. Для задачі регресії цей прогноз може бути середнім значенням $T_k(x)$ для всіх зразків, що потрапили в цей листок.

Таким чином, прогноз для дерева можна записати як:

$$T_k(x) = \frac{1}{|D_k|}\sum_{i\in D_k} y_i \tag{2.2}$$

де D_{k} – набір даних, що потрапили в листок дерева k;

y_i – значення цільової змінної у листку дерева k.

Прогноз усередині випадкового лісу: Для кожного дерева T_k в лісі робиться прогноз $T_k(x)$ для нового зразка x.

Формально, прогноз моделі Random Forest для нового зразка x можна виразити як:

$$\hat{y} = \frac{1}{M}\sum_{k=1}^{M} T_k(x) \tag{2.3}$$

де M – кількість дерев у лісі;

$T_k(x)$ – прогноз k-го дерева для зразка x.

Переваги методу випадкового лісу:

- Стійкість до перенавчання: Один з основних плюсів RF полягає в тому, що завдяки використанню великої кількості дерев, кожне з яких навчається на різних підмножинах даних, ризик перенавчання значно зменшується. Це забезпечує високу здатність до узагальнення моделі, навіть якщо дані мають значний рівень шуму чи мають складну структуру.
- Висока точність та стабільність: Завдяки ансамблевому підходу, коли результати кожного дерева комбінуються для отримання кінцевого прогнозу, Random Forest демонструє високу точність та

стабільність у порівнянні з іншими методами. Усереднення прогнозів допомагає зменшити вплив аномальних або шумових значень на кінцевий результат.

- Інтерпретованість та оцінка важливості ознак: Хоча окремі дерева можуть бути складними для інтерпретації через свою глибину та розгалуження, Random Forest надає потужні інструменти для оцінки важливості кожної ознаки, що дозволяє виявити найбільш впливові параметри для прогнозування. Це робить модель зрозумілою для аналітиків, особливо в задачах, де важливо знати, які ознаки мають найбільший вплив на результат.

Недоліки методу випадкового лісу:

- Велика обчислювальна складність: Оскільки Random Forest використовує велику кількість дерев, створення та прогнозування моделі потребує значних обчислювальних ресурсів, зокрема пам'яті та часу процесора. Це може бути проблемою при роботі з великими наборами даних або в реальному часі, коли важлива швидкість прогнозу.
- Складність інтерпретації: Хоча окремі дерева можуть бути інтерпретовані досить легко, ансамбль дерев, як у випадку RF, може бути складним для повного розуміння, особливо коли йдеться про те, як окремі дерева взаємодіють між собою. Навіть з оцінкою важливості ознак, модель може залишатися непрозорою для кінцевого користувача, якщо необхідно отримати конкретні інтерпретації поведінки моделі для окремих випадків.

2.1.2 Метод градієнтного бустингу

XGBoost (Extreme Gradient Boosting) – це потужна і вдосконалена реалізація методу градієнтного бустингу, яка відзначається високою точністю, швидкістю обчислень та гнучкістю налаштувань. Цей алгоритм спеціально розроблений для масштабованої роботи з великими обсягами даних і

забезпечення ефективної обробки як структурованих, так і частково неструктурованих наборів даних [16][17].

Однією з ключових переваг XGBoost є наявність регуляризації (L1 та L2), що дозволяє зменшити ризик перенавчання моделі. Крім того, алгоритм використовує оптимізований розклад Тейлора другого порядку для функції втрат, що забезпечує врахування як першої, так і другої похідної (градієнта і гесіана). Це дозволяє точніше описувати криву функції втрат і, відповідно, приймати більш ефективні рішення при побудові дерева на кожній ітерації.

Як і класичний градієнтний бустинг, XGBoost формує ансамбль слабких моделей – зазвичай дерев рішень – де кожне наступне дерево навчається на помилках попередніх. Проте завдяки вдосконаленій реалізації, алгоритм досягає вищої продуктивності, стабільності та можливості тонкого налаштування, що зробило його одним з найпопулярніших інструментів машинного навчання на практиці.

Основні кроки роботи XGBoost:

1. Ініціалізація: Алгоритм стартує з базового прогнозу. Для задачі регресії це зазвичай середнє значення цільової змінної, для класифікації – логарифмічні шанси класу. Цей прогноз буде послідовно уточнюватися.

2. Обчислення похідних: На кожній ітерації для кожного прикладу в навчальній вибірці обчислюються перша похідна (градієнт) та друга похідна (гесіан) функції втрат. Вони описують напрямок та "кривизну" функції втрат, що дає змогу більш точно формувати модель.

3. Побудова нового дерева: Використовуючи градієнти та гесіани, будується нове дерево рішень. Його завдання – апроксимувати негативний градієнт функції втрат, тобто передбачити помилки попереднього ансамблю моделей.

4. Оновлення моделі: Поточне передбачення оновлюється шляхом додавання результату нового дерева, помноженого на коефіцієнт навчання (learning rate), що контролює величину внеску дерева в остаточну модель.

5. Повторення: Кроки 2–4 повторюються ітеративно. Алгоритм додає нові дерева до ансамблю доти, поки не буде досягнута задана кількість ітерацій або приріст якості не стане незначним (може використовуватись рання зупинка).

Математичний опис:

Ініціалізація моделі: На початку процесу будується базова модель, яка зазвичай є простим середнім значенням цільової змінної для всіх тренувальних даних. Це початковий прогноз, від якого відштовхуються подальші ітерації.

$$\widehat{y_0} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{2.4}$$

де n – кількість зразків у тренувальному наборі даних;

y_i – значення цільової змінної для і-го зразка у тренувальному наборі даних.

Обчислення градієнтів: На кожній ітерації t обчислюється градієнт функції втрат за попередній прогноз. Градієнти вказують напрямок, у якому потрібно коригувати модель для зменшення помилки. У найпростішому випадку (наприклад, для квадратичної втрати) це просто залишки:

$$g_i^{(t)} = \frac{\partial L(y_i,\hat{y}_i)}{\partial \hat{y}_i} \tag{2.5}$$

де $L(y_i,\hat{y}_i)$ – функція втрат для і-го зразка.

А також обчислюється друга похідна (гесіан), яка дозволяє точніше апроксимувати функцію втрат:

$$h_i^{(t)} = \frac{\partial^2 L(y_i,\hat{y}_i)}{\partial \widehat{y^2}_i} \tag{2.6}$$

Побудова дерева: XGBoost будує дерево $f_t(x)$, яке апроксимує функцію втрат другого порядку. Для цього використовується розклад Тейлора, і оптимізується наступне вираження:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t(x_i)^2\right] + \Omega(f_t) \tag{2.7}$$

де $f_t(x_i)$ – це прогноз, який дає дерево t для вхідного зразка x_i.

$\Omega(f_t)$ – регуляризація для дерева, щоб запобігти перенавчанню:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{2.8}$$

де γ – штраф за кількість листів у дереві;

T – кількість листів у дереві;

λ – параметр регуляризації;

ω_j – параметри для j-го листа дерева:

$$\omega_j = -\frac{\Sigma_{i\epsilon I_j} g_i}{\Sigma_{i\epsilon I_j} h_i + \lambda} \tag{2.9}$$

де I_j – множина індексів зразків, що потрапляють в j-й лист дерева.

Оцінка розбиття вузла: Під час побудови дерева в XGBoost на кожній ітерації розглядаються всі можливі ознаки та пороги розділення для кожного вузла. Для кожного можливого варіанту обчислюється приріст (Gain) – показник того, наскільки ефективно таке розбиття зменшує функцію втрат. Обирається те розбиття, яке дає максимальне значення Gain, що дозволяє забезпечити оптимальність розбиття вузла. Формула обчислення Gain має вигляд:

$$Gain = \frac{1}{2}\left[\frac{\left(\Sigma_{i\epsilon I_L} g_i\right)^2}{\Sigma_{i\epsilon I_L} h_i + \lambda} + \frac{\left(\Sigma_{i\epsilon I_R} g_i\right)^2}{\Sigma_{i\epsilon I_R} h_i + \lambda} - \frac{(\Sigma_{i\epsilon I} g_i)^2}{\Sigma_{i\epsilon I} h_i + \lambda}\right] - \gamma \tag{2.10}$$

де I_L та I_R – індекси зразків у лівій та правій частинах дерева;

λ – параметр регуляризації;

γ – штраф за кількість листів у дереві.

Оновлення моделі: Прогноз моделі оновлюється шляхом додавання внеску нового дерева, зваженого на коефіцієнт навчання η. Це дозволяє контролювати внесок кожного нового дерева і запобігати перенавчанню. Оновлений прогноз виглядає наступним чином:

$$\widehat{y_i^{(t)}} = \widehat{y_i^{(t-1)}} + \eta * f_t(x_i) \tag{2.11}$$

де η - коефіцієнт навчання, що контролює вагу нового дерева,

$f_t(x_i)$ - прогноз t-го дерева для зразка x_i.

Повторення процесу: Цей процес повторюється для заданої кількості ітерацій або до досягнення необхідного рівня точності. Кожне нове дерево додається до ансамблю, поступово покращуючи точність прогнозів.

Після завершення навчання моделі, для кожного нового зразка x остаточний прогноз моделі XGBoost є сумою прогнозів всіх дерев у ансамблі, зважених на коефіцієнт навчання:

$$\hat{y} = \sum_{t=1}^{T} \eta * f_t(x) \quad (2.12)$$

Переваги:

- XGBoost є одним із найшвидших алгоритмів градієнтного бустингу завдяки оптимізованому коду. Крім того, він підтримує паралельну обробку даних, що дозволяє значно зменшити час навчання моделі;
- Завдяки поетапному вдосконаленню моделі та використанню градієнтного підходу другого порядку (з використанням гесіанів), XGBoost здатний досягати дуже високої якості прогнозів як у задачах класифікації, так і регресії;
- Алгоритм використовує регуляризацію як для структури дерев, так і для ваг у листах, що дозволяє зменшити ризик перенавчання, особливо при роботі з невеликими або шумовими наборами даних;
- XGBoost ефективно працює з великими наборами даних і великою кількістю ознак, оскільки оптимізований як за часом виконання, так і за споживанням пам’яті.

Недоліки:

- XGBoost має велику кількість параметрів, таких як кількість дерев, глибина дерев, коефіцієнт навчання, параметри регуляризації тощо. Для досягнення оптимальної продуктивності необхідне ретельне налаштування цих гіперпараметрів, що може бути ресурсоємним та вимагати експериментування або використання методів

автоматичного підбору (наприклад, grid search або Bayesian optimization);

- Незбалансоване або некоректне налаштування гіперпараметрів може призвести до перенавчання моделі або, навпаки, до недостатнього навчання. Наприклад, надмірна глибина дерев або занадто малий коефіцієнт регуляризації можуть зробити модель занадто складною, що погіршить її узагальнюючу здатність;
- Хоча XGBoost надає інструменти для інтерпретації (наприклад, важливість ознак), результуюча модель часто є ансамблем багатьох дерев, що ускладнює повне розуміння логіки прийняття рішень, особливо для користувачів без досвіду в машинному навчанні.

2.1.3 Нейронна мережа типу LSTM

Нейронна мережа з довгостроковою короткочасною пам’яттю (Long Short-Term Memory, LSTM) є різновидом рекурентної нейронної мережі (RNN), яка здатна зберігати та використовувати інформацію протягом тривалих проміжків часу. Це дозволяє їй ефективно працювати з послідовностями даних, такими як часові ряди, мовний текст або фінансові показники. Основною перевагою LSTM є здатність долати проблему зниклого градієнту, яка часто виникає в класичних RNN [18][19].

На відміну від звичайної RNN, LSTM має складну архітектуру з вбудованими воротами (gates), які контролюють потік інформації та допомагають вирішити, яку інформацію слід зберегти, оновити або забути (рисунок 2.1).

Ключова ідея LSTM полягає в окремому керуванні довгостроковою пам’яттю (cell state) та короткостроковою пам’яттю (hidden state), що дозволяє моделі адаптивно фільтрувати та передавати релевантну інформацію через час. Завдяки цьому LSTM демонструє високу ефективність у завданнях прогнозування, класифікації послідовностей, автоматичного перекладу, розпізнавання мови та інших задач, де є важливими часові залежності.

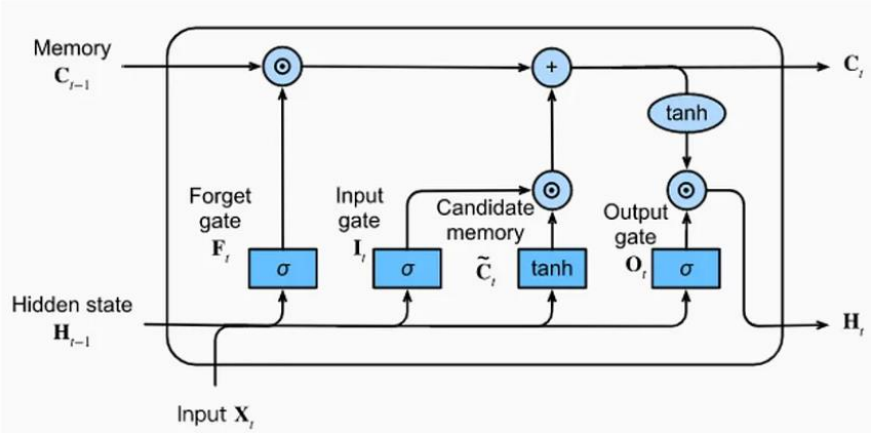

Рисунок 2.1 – приклад LSTM моделі

Основні компоненти LSTM:

Клітинний стан (Cell State): Це головний елемент, що забезпечує можливість зберігати довготривалу інформацію протягом усього часового ряду. Клітинний стан передається через усі часові кроки з мінімальними змінами, забезпечуючи стабільне зберігання даних у мережі. LSTM модифікує клітинний стан лише за допомогою спеціальних воріт, що дозволяє вибірково додавати нову інформацію або забувати застарілу. Завдяки цьому компоненту мережа здатна ефективно зберігати контекст навіть на тривалих інтервалах часу, що робить її придатною для роботи з довгими залежностями у послідовностях. Це також допомагає уникати проблеми зниклого градієнта, яка є типовою для класичних RNN.

Прихований стан (Hidden State): Це тимчасова, короткострокова пам’ять LSTM, яка містить інформацію про поточний момент часу. Він формується на основі клітинного стану і вихідних воріт та використовується як вхід на наступному часовому кроці. Прихований стан також передається як частина вхідного сигналу до наступного блоку мережі або використовується як проміжний або фінальний вихід при обробці послідовності. Таким чином, прихований стан відображає поточний контекст і є ключовим для прийняття рішень у кожен момент часу.

Ворота (Gates): Ворота – це спеціальні логічні модулі, які керують потоком інформації в мережі. Їх завдання – визначити, які дані необхідно зберегти, які –

забути, а які – передати далі. Всі ворота побудовані на основі нейронної мережі з сигмоїдною активацією, яка обмежує вихід у межах [0, 1], що дозволяє точно контролювати, скільки інформації буде пропущено.

Ворота забування (Forget Gate): Ці ворота відповідають за "очищення" клітинного стану від неактуальної або зайвої інформації. На кожному часовому кроці вони аналізують поточний вхід та прихований стан попереднього кроку, після чого визначають, яку частину інформації з попереднього клітинного стану необхідно зберегти, а яку – забути. Це дозволяє LSTM адаптуватися до змін у даних і залишати лише релевантну інформацію.

Ворота входу (Input Gate): Визначають, яка нова інформація буде додана до клітинного стану. Складаються з двох етапів: Перший крок – сигмоїдна функція вирішує, які компоненти нового вхідного сигналу будуть розглядатися; Другий крок – гіперболічний тангенс створює вектор нових кандидатних значень, які можуть бути додані до клітинного стану. Разом вони дозволяють моделі динамічно оновлювати свою довготривалу пам’ять на основі нових даних.

Ворота виходу (Output Gate): Визначають, яка частина клітинного стану буде використана для формування прихованого стану, тобто виходу з поточного елемента LSTM. Ці ворота комбінують значення клітинного стану (через tanh) та сигмоїдну функцію, яка визначає, які частини інформації будуть виведені. Таким чином, ворота виходу регулюють, яка інформація буде доступна зовнішньому середовищу та наступному часовому кроці.

Функції активації (Activation Functions): У LSTM використовуються дві основні функції активації: Сигмоїдна функція – застосовується у всіх воротах для прийняття рішень, наскільки сильно пропускати певну інформацію (0 – не пропускати зовсім, 1 – пропускати повністю); Гіперболічний тангенс – використовується для масштабування значень клітинного стану та формування кандидатного вектору нової інформації. Забезпечує діапазон значень від -1 до 1, що корисно для стабільного навчання мереж.

Математичний опис:

Вхід**:** На кожному кроці часу t, LSTM отримує вхідний вектор x_t та попередній вихід h_{t-1}.

Вагові коефіцієнти для воріт:

Ворота забування:

$$(f_t): f_t = \sigma_g(W_f * [h_{t-1}, x_t] + b_f) \quad (2.13)$$

Ворота входу:

$$(i_t): i_t = \sigma_g(W_i * [h_{t-1}, x_t] + b_i) \quad (2.14)$$

Ворота оновлення клітинного стану:

$$(g_t): g_t = tanh(W_c * [h_{t-1}, x_t] + b_c) \quad (2.15)$$

Ворота виходу:

$$(o_t): o_t = \sigma_g(W_o * [h_{t-1}, x_t] + b_o) \quad (2.16)$$

де W_f, W_i, W_c, W_o – вагові матриці;

h_{t-1} – прихований стан на попередньому кроці часу;

x_t – вхідний вектор на поточному кроці часу;

b_f, b_i, b_c, b_o – вектори зсуву;

σ_g – сигмоїдна функція активації;

tanh – гіперболічний тангенс.

Сигмоїдна функція:

$$\sigma_g(z) = \frac{1}{1+e^{-z}} \quad (2.17)$$

Гіперболічний тангенс:

$$tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \quad (2.18)$$

Оновлення клітинного стану за допомогою по елементного множення:

$$C_t = f_t \odot C_{t-1} + i_t \odot g_t \quad (2.19)$$

Оновлення вихідного стану стану за допомогою по елементного множення:

$$h_t = o_t \odot tanh\,(C_t) \quad (2.20)$$

Переваги LSTM:

- Довготривала пам’ять: Ефективно запам’ятовує залежності на довгих часових відстанях, завдяки механізму збереження клітинного стану;
- Контроль інформаційного потоку: Завдяки використанню трьох типів воріт (забування, входу, виходу), модель може гнучко вирішувати, яку інформацію зберігати або забувати;
- Менше проблем із затухаючими градієнтами: На відміну від звичайних RNN, LSTM менше схильна до проблеми затухання градієнтів при навчанні на довгих послідовностях;
- Гнучкість застосування: Використовується в багатьох задачах: прогнозування часових рядів, розпізнавання мовлення, машинний переклад, генерація тексту тощо.

Недоліки LSTM:

- Висока обчислювальна складність: Кількість параметрів значно більша, ніж у звичайних RNN, що призводить до довшого часу навчання;
- Складність інтерпретації: Через складну внутрішню структуру важко інтерпретувати, як саме приймаються рішення;
- Потребує багато даних для навчання: Для отримання гарних результатів LSTM потребує великої кількості даних;
- Схильність до перенавчання: Якщо не використовувати регуляризацію (наприклад, Dropout), модель може перенавчитися на навчальній вибірці.

2.2 Метрики оцінки точності моделі

Для оцінки ефективності моделей прогнозування використовуються різноманітні метрики точності, які дозволяють кількісно виміряти рівень збігу між передбаченими та фактичними значеннями. Правильний вибір метрик має важливе значення, оскільки він впливає на інтерпретацію результатів та

подальше вдосконалення моделі. У задачах прогнозування часових рядів зазвичай використовуються такі метрики [20].

Середньоквадратична помилка (Mean Squared Error, MSE): Ця метрика обчислює середнє значення квадратів різниці між фактичними та передбаченими значеннями. Вона є чутливою до великих відхилень, тому добре підходить для виявлення великих помилок.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (2.21)$$

де y_i – фактичне значення;

$\hat{y}_i$ – передбачене значення;

n – кількість спостережень

Корінь середньоквадратичної помилки (Root Mean Squared Error, RMSE): Це квадратний корінь із MSE, що повертає помилку в тих самих одиницях, що й початкові дані. RMSE легше інтерпретувати порівняно з MSE.

$$RMSE = \sqrt{MSE} \quad (2.22)$$

Середня абсолютна помилка (Mean Absolute Error, MAE): Оцінює середню абсолютну різницю між фактичними та передбаченими значеннями. На відміну від MSE, MAE не акцентує увагу на великі відхилення, тому є більш стійкою до викидів.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2.23)$$

Середня абсолютна відсоткова помилка (Mean Absolute Percentage Error, MAPE): Ця метрика вимірює середнє відсоткове відхилення передбачених значень від фактичних. Перевагою є зручність інтерпретації у відсотках, але вона чутлива до малих значень фактичних даних.

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (2.24)$$

Коефіцієнт детермінації (R^2): Показує, яка частка дисперсії залежної змінної пояснюється моделлю.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (2.25)$$

де $\bar{y}$ – середнє значення фактичних значень.

2.3 Узагальнення етапів побудови моделей

Процес побудови моделей прогнозування змін клімату за допомогою методів Data Science включає кілька ключових етапів. Кожен з них є важливою складовою загального підходу до аналізу даних та створення ефективної прогностичної моделі:

1. Збір та попередня обробка даних.

На цьому етапі відбувається завантаження необхідних датасетів, видалення пропущених або аномальних значень, приведення даних до відповідного формату. Для часових рядів особливо важливо забезпечити правильну послідовність спостережень.

2. Вибір ознак та побудова вхідних даних.

Виконується вибір найважливіших параметрів, що впливають на зміну температури, наприклад: рівень викидів парникових газів, рік, країна тощо. У випадку використання LSTM-моделі формується послідовність спостережень певної довжини (наприклад, 10 попередніх значень).

3. Масштабування даних.

Застосовується масштабування (наприклад, за допомогою MinMaxScaler) для приведення всіх вхідних параметрів до одного діапазону, що є важливою умовою коректної роботи більшості моделей машинного навчання та нейронних мереж.

4. Розбиття даних на тренувальний та тестовий набори.

Дані діляться на дві частини: для навчання моделі (train) та для перевірки її якості (test). Це дозволяє уникнути перенавчання та адекватно оцінити точність прогнозу.

5. Побудова та тренування моделі.

Залежно від обраного підходу (випадковий ліс, градієнтний бустинг, LSTM-мережа), створюється відповідна модель, налаштовуються її параметри та виконується навчання на тренувальних даних.

6. Прогнозування та візуалізація результатів.

Після навчання, модель використовується для прогнозу температури на основі тестових або нових вхідних даних. Результати прогнозу порівнюються з реальними значеннями, створюються графіки для оцінки якості моделі.

7. Оцінка точності моделі.

За допомогою таких метрик, як середньоквадратична помилка (MSE) та коефіцієнт детермінації (R^2), здійснюється кількісна оцінка якості побудованої моделі. Це дозволяє визначити, наскільки точно модель відтворює тренди та коливання температури.

Узагальнена структура побудови моделей дозволяє не лише отримувати точні прогнози, але й забезпечує прозорість, відтворюваність та масштабованість процесу в рамках більшого проекту з аналізу кліматичних змін.

2.4 Вибір інструментів реалізації

У рамках даного дослідження було обрано мову програмування та набір бібліотек, які найкраще відповідають завданням з аналізу даних, побудови моделей прогнозування та візуалізації результатів. У цьому пункті описано обґрунтування вибору мови програмування та функціональні можливості використаних бібліотек.

2.4.1 Вибір мови програмування

Для виконання роботи було обрано мову програмування Python. Ця мова останнім часом все частіше використовується для створення моделей машинного навчання, аналізу даних та побудови графіків. Використання Python є доцільним завдяки його простоті та зручності у використанні, що дозволяє будувати моделі з мінімальними зусиллями. Мова Python надає можливість виконувати складні математичні операції та аналіз даних за допомогою невеликих фрагментів коду, що значно спрощує процес розробки моделей [21].

Крім того, Python має велику кількість різноманітних відкритих бібліотек, таких як scikit-learn, xgboost, TensorFlow, Keras, pandas, numpy, matplotlib та багато інших, які забезпечують широкий спектр інструментів для машинного навчання, аналізу даних та побудови різноманітних графіків.

Велика та активна спільнота Python забезпечує підтримку та доступ до численних ресурсів для навчання та вирішення проблем, що виникають під час роботи. Це дозволяє швидко знаходити рішення для будь-яких технічних питань, що можуть виникнути під час виконання проекту. Python також є дуже гнучким і масштабованим, що дозволяє інтегрувати різні модулі та бібліотеки для досягнення оптимальних результатів.

Однак, Python має деякі недоліки, такі як відносно повільна швидкодія у порівнянні з компільованими мовами, що може бути критичним для деяких задач. Також Python може вимагати більше пам'яті у порівнянні з іншими мовами, що може стати обмеженням при обробці великих обсягів даних.

Загалом, вибір мови програмування Python для виконання роботи є обґрунтованим і раціональним, оскільки ця мова надає всі необхідні інструменти для успішного виконання завдань з аналізу даних та розробки моделей прогнозування, а її переваги значно переважують недоліки.

2.4.2 Опис бібліотек, функцій та веб-технологій

Для аналізу даних:

- pandas: Бібліотека для обробки та аналізу даних, яка надає структури даних та інструменти для роботи з таблицями та часовими рядами;
- numpy: Бібліотека для роботи з багатовимірними масивами та великим набором математичних функцій для виконання операцій над цими масивами.

Для візуалізації:

- matplotlib.pyplot: Основна бібліотека для створення статичних, анімованих та інтерактивних візуалізацій у Python;

- plotly.express: Бібліотека для створення інтерактивних графіків та візуалізацій;
- seaborn: Бібліотека для візуалізації даних на основі matplotlib, яка надає високоінформативні статистичні графіки.

Для машинного навчання:

- sklearn.model_selection.train_test_split: Функція для розподілу даних на тренувальний та тестовий набори;
- sklearn.metrics.mean_squared_error: Функція для обчислення середньоквадратичної помилки;
- sklearn.metrics.r2_score: Функція для обчислення коефіцієнта детермінації (R^2);
- sklearn.ensemble.RandomForestRegressor: Клас для реалізації регресії з використанням випадкового лісу;
- xgboost.XGBRegressor: Клас для реалізації регресії з використанням градієнтного бустингу;
- tensorflow.keras.models.Sequential: Клас для створення послідовних моделей нейронних мереж;
- tensorflow.keras.layers.LSTM: Клас для додавання шарів LSTM (Long Short-Term Memory) у нейронну мережу [22];
- tensorflow.keras.layers.Dense: Клас для додавання щільних (fully connected) шарів у нейронну мережу;
- sklearn.preprocessing.MinMaxScaler: Клас для масштабування даних до діапазону [0, 1].

Для веб-розробки:

- HTML: Використовується для створення структури веб-сторінок і інтеграції результатів аналізу;
- CSS: Застосовується для стилізації веб-сторінок, зокрема для оформлення графіків та таблиць результатів;

- JavaScript: Використовується для забезпечення інтерактивності на сайті, зокрема для динамічного відображення графіків та змін даних.

Ці бібліотеки та веб-технології дозволяють здійснювати повний цикл аналізу даних: від попередньої обробки та візуалізації результатів до побудови моделей машинного навчання та інтеграції отриманих даних у веб-інтерфейси.

2.5 Висновок до розділу

У цьому розділі було здійснено ґрунтовний аналіз методів машинного навчання, які є найбільш релевантними для задачі прогнозування кліматичних змін, зокрема температури. Розглянуто ключові підходи до моделювання, включаючи алгоритми випадкового лісу (RF), градієнтного бустингу (XGBoost) та довготривалої короткочасної пам’яті (LSTM). Кожен з методів має свої особливості та переваги: алгоритми на основі дерев рішень забезпечують інтерпретованість та високу точність для табличних даних, тоді як рекурентні нейронні мережі демонструють високу ефективність при роботі з часовими рядами завдяки здатності враховувати залежності в динаміці даних.

Окрему увагу приділено метрикам оцінки точності моделей, що є критично важливими для об’єктивного порівняння ефективності підходів та вибору найкращої моделі. Такі метрики, як MAE, RMSE, R^2, дозволяють кількісно оцінити рівень відхилення прогнозу від фактичних значень, що є важливим кроком у процесі валідації моделей.

Також було систематизовано основні етапи побудови моделей — від попередньої обробки даних до валідації та впровадження моделей у вигляді інформаційної системи. Це дало змогу сформувати чіткий алгоритм дій для побудови прогнозної системи на практиці.

Для реалізації обраних моделей було визначено оптимальний набір інструментів. Основною мовою програмування обрано Python, завдяки її гнучкості, багатству функціональних бібліотек та активній спільноті розробників. Зокрема, було охарактеризовано бібліотеки, які забезпечують ефективну роботу з даними (pandas, numpy), побудову та навчання моделей

(scikit-learn, xgboost, tensorflow), а також створення інтерактивних візуалізацій (matplotlib, seaborn, plotly), які допомагають краще зрозуміти структуру даних та результати моделювання.

Крім того, було враховано можливості веб-технологій (HTML, CSS, JavaScript) для інтеграції побудованих моделей у зручний інтерфейс користувача. Це є важливим кроком для створення повноцінної інформаційної системи, яка дозволяє не лише проводити прогнозування, а й забезпечувати доступ до результатів у наочній формі.

Таким чином, у цьому розділі було закладено теоретико-практичну основу для реалізації ефективної системи прогнозування кліматичних змін. Комплексний підхід до вибору методів і технологій дозволяє не лише досягти високої точності прогнозів, а й забезпечує можливість масштабування, адаптації та інтеграції моделі в реальні умови, що є важливим чинником для подальшого розвитку дослідження.

РОЗДІЛ 3
РОЗРОБКА МЕТОДІВ DATA SCIENCE

3.1 Аналіз даних

3.1.1 Опис даних та джерела даних

У цьому дослідженні використано дані, що охоплюють період з 1990 по 2020 роки та містять інформацію про різні джерела викидів CO_2 у сільському господарстві, середньорічну температуру повітря, а також чисельність населення по кожній країні світу (рис. 3.1) [23]. Таблиця 3.1 містить перелік змінних, представлених у наборі даних.

	Area	Year	Savanna fires	Forest fires	Crop Residues	Rice Cultivation	Drained organic soils (CO2)	Pesticides Manufacturing	Food Transport	Forestland	...	Manure Management	Fires in organic soils	Fires in humid tropical forests	On-farm energy use
0	Afghanistan	1990	14.7237	0.0557	205.6077	686.0000	0.0	11.807483	63.1152	-2388.8030	...	319.1763	0.0	0.0	NaN
1	Afghanistan	1991	14.7237	0.0557	209.4971	678.1600	0.0	11.712073	61.2125	-2388.8030	...	342.3079	0.0	0.0	NaN
2	Afghanistan	1992	14.7237	0.0557	196.5341	686.0000	0.0	11.712073	53.3170	-2388.8030	...	349.1224	0.0	0.0	NaN
3	Afghanistan	1993	14.7237	0.0557	230.8175	686.0000	0.0	11.712073	54.3617	-2388.8030	...	352.2947	0.0	0.0	NaN
4	Afghanistan	1994	14.7237	0.0557	242.0494	705.6000	0.0	11.712073	53.9874	-2388.8030	...	367.6784	0.0	0.0	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
6929	Zimbabwe	2016	1190.0089	232.5068	70.9451	7.4088	0.0	75.000000	251.1465	76500.2982	...	282.5994	0.0	0.0	417.3150
6930	Zimbabwe	2017	1431.1407	131.1324	108.6262	7.9458	0.0	67.000000	255.7975	76500.2982	...	255.5900	0.0	0.0	398.1644
6931	Zimbabwe	2018	1557.5830	221.6222	109.9835	8.1399	0.0	66.000000	327.0897	76500.2982	...	257.2735	0.0	0.0	465.7735
6932	Zimbabwe	2019	1591.6049	171.0262	45.4574	7.8322	0.0	73.000000	290.1893	76500.2982	...	267.5224	0.0	0.0	444.2335
6933	Zimbabwe	2020	481.9027	48.4197	108.3022	7.9733	0.0	73.000000	238.7639	76500.2982	...	266.7316	0.0	0.0	444.2335

Рисунок 3.1 – дані з викидами

Таблиця 3.1 – колонки таблиці з даними

Назва колонки	Значення
Area	назва країни
Year	рік
Savanna fires	викиди від пожеж в саванах
Forest fires	викиди від пожежі в лісах
Crop Residues	викиди від спалювання або розкладання залишків рослинного матеріалу після збору врожаю
Rice Cultivation	викиди від вирощування рису
Drained organic soils (CO_2)	викиди, що виділяються під час осушення органічних ґрунтів
Pesticides Manufacturing	викиди від виробництва пестицидів
Food Transport	викиди від транспортування харчових продуктів
Forestland	землі, вкриті лісами
Net Forest conversion:	зміна лісової площі внаслідок вирубки

Food Household Consumption	викиди від споживання продуктів харчування на рівні домогосподарств
Food Retail	викиди від роботи підприємств роздрібної торгівлі продуктами харчування
On-farm Electricity Use	споживання електроенергії на фермах
Food Packaging	викиди від виробництва та утилізації пакувальних матеріалів для харчових продуктів
Agrifood Systems Waste Disposal	викиди від утилізації відходів у агропромисловій системі
Food Processing	викиди від обробки харчових продуктів
Fertilizers Manufacturing	викиди від виробництва добрив
IPPU	викиди від промислових процесів і використання продукції
Manure applied to Soils	викиди від внесення тваринного гною в сільськогосподарські ґрунти
Manure left on Pasture	викиди від гною тварин на пасовищах
Manure Management	викиди від управління та обробки гною тварин
Fires in organic soils	викиди від пожеж в органічних ґрунтах
Fires in humid tropical forests	викиди від пожеж у вологих тропічних лісах
total_emission	загальні викиди парникових газів з різних джерел
total_population	загальна чисельність населення
Average Temperature °C	середнє підвищення температури (за рік) у градусах Цельсія

Всі викиди записуються у кілотоннах (кт). Середнє підвищення температури за рік було пораховано відносно нормального значення (середнє значення температури з 1951 по 1980 рік) [24].

Основним джерелом даних є міжнародна база FAOSTAT (Food and Agriculture Organization of the United Nations). Ця база даних є авторитетним джерелом глобальної статистики у сфері сільського господарства, харчової промисловості, використання природних ресурсів та впливу на довкілля. FAOSTAT надає відкритий доступ до великого обсягу структурованих даних, які регулярно оновлюються та мають високу наукову цінність. Вибір саме цього джерела обумовлений його надійністю, деталізацією аграрних викидів та широким визнанням у наукових дослідженнях. Ці дані забезпечують основу для аналізу впливу різних джерел викидів на зміну середньої температури та дозволяють створювати прогностичні моделі для оцінки майбутніх змін .

3.1.2 Попередня обробка даних

Попередня обробка даних є критично важливим етапом в будь-якому проекті Data Science. Цей процес включає в себе кілька ключових кроків, які забезпечують якість та готовність даних до подальшого аналізу та моделювання. Попередня обробка даних складалася з декількох етапів.

Спочатку було завантажено дані до Jupyter notebook за допомогою бібліотеки pandas. Після завантаження було додано дві колонки. В першій пораховано загальна кількість населення країни, а в другій до якого континенту відноситься країна. Також проведено огляд структури та змісту даних. За допомогою вбудованих функцій оглянуто інформацію про стовпці (рис. 3.2) та відсутні значення (рис. 3.3).

#	Column	Non-Null Count	Dtype
0	Area	6934 non-null	object
1	Continent	6934 non-null	object
2	Year	6934 non-null	int64
3	Savanna fires	6903 non-null	float64
4	Forest fires	6841 non-null	float64
5	Crop Residues	5545 non-null	float64
6	Rice Cultivation	6934 non-null	float64

Рисунок 3.2 – інформація про стовпці

Area	0
Continent	0
Year	0
Savanna fires	31
Forest fires	93
Crop Residues	1389
Rice Cultivation	0
Drained organic soils (CO2)	0
Pesticides Manufacturing	0
Food Transport	0
Forestland	493
Net Forest conversion	493

Рисунок 3.3 – інформація про відсутні значення

Виявлено, що інформація була відсутня в 11 стовпчиках. Для вирішення цієї проблеми було замінено пусті значення на середні. Спочатку замінено значення середніми по країні, а якщо по країні повністю була відсутня інформація, то було замінено середніми значеннями по континенту, на якому розташована країна.

Ці кроки забезпечили якість та готовність даних до подальшого аналізу та моделювання. Тепер дані готові до використання в наступних етапах дослідження.

3.1.3 Візуалізація даних

Візуалізація даних є важливим етапом аналізу, оскільки вона дозволяє краще зрозуміти структуру, тенденції та взаємозв'язки в даних. Вона допомагає швидше виявити патерни та аномалії, що можуть залишитися непоміченими при традиційному аналізі. У цьому дослідженні використовувались кілька інструментів для візуалізації даних, таких як Matplotlib, Seaborn і Plotly. Кожен з цих інструментів надає різні можливості для створення графіків і діаграм, що дозволяють візуалізувати як загальні тенденції, так і деталі окремих аспектів досліджуваних даних.

Перша візуалізація була спрямована на порівняння викидів CO2 від сільськогосподарської діяльності з загальними викидами CO2 . Для цього були використані дані про загальні викиди CO2 за той самий період, переведені у гігатонни (Гт) для зручності порівняння [25]. Після цього було проведено групування даних по роках, що дозволило отримати більш чітке уявлення про зміни викидів за часом. Отриманий результат представлений на рисунку 3.4.

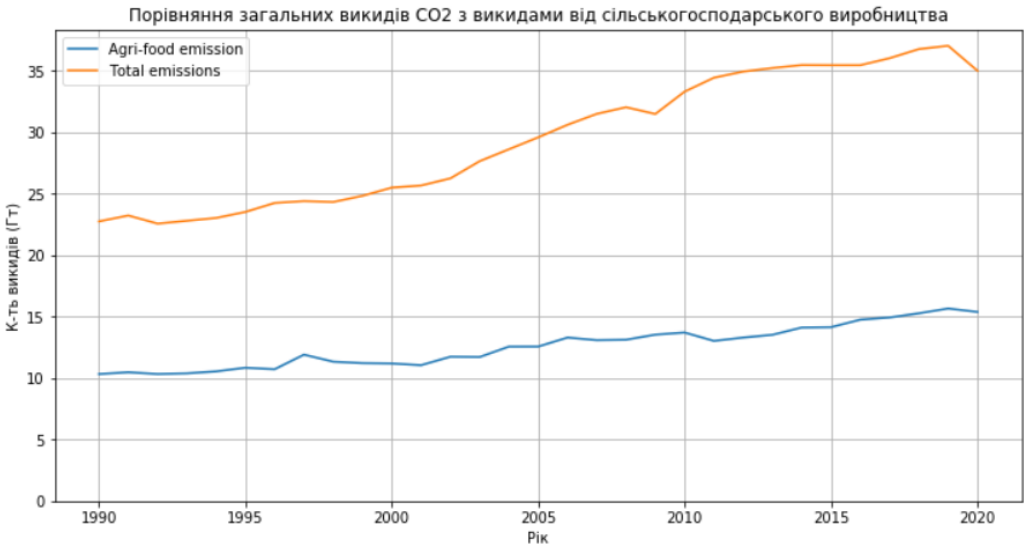

Рисунок 3.4 – порівняння загальних викидів CO_2 з викидами від сільськогосподарського виробництва

Для більш детального аналізу було вирішено побудувати другий графік, на якому показано відсоткове співвідношення викидів від сільськогосподарського сектору до загальних викидів CO2 за кожен рік (рис. 3.5). Така візуалізація дозволяє краще зрозуміти, як зміни викидів CO2 від сільськогосподарської діяльності впливають на загальну картину та дозволяє виявити періоди, коли цей сектор стає більш або менш важливим джерелом викидів.

Рисунок 3.5 – динаміка відсотка викидів CO_2 від сільськогосподарського виробництва відносно загальних викидів

На рисунку 3.4 видно, що загальні викиди CO2 з роками значно зростали, тоді як викиди від сільськогосподарської діяльності теж збільшувались, але не так стрімко. У відсотковому співвідношенні, як показано на рисунку 3.5, після 1997 року викиди від сільськогосподарської діяльності починають спадати, але після 2011 року знову спостерігається їх зростання. В 2020 році викиди від сільського господарства складають майже 44% від загальних викидів CO2. Це може свідчити про необхідність впровадження нових заходів для зменшення викидів у сільському господарстві або їх оптимізації для досягнення більш сталого розвитку.

Наступним етапом візуалізації було створення карти світу, яка ілюструє зміну середньої температури в різних країнах у динаміці за роками. Такий тип візуалізації дозволяє краще зрозуміти просторовий розподіл температурних змін

і оцінити, які регіони найбільше постраждали від глобального потепління. На рисунку 3.6 зображено зміну середньої температури в 1992 році, а на рисунку 3.7 – у 2020 році.

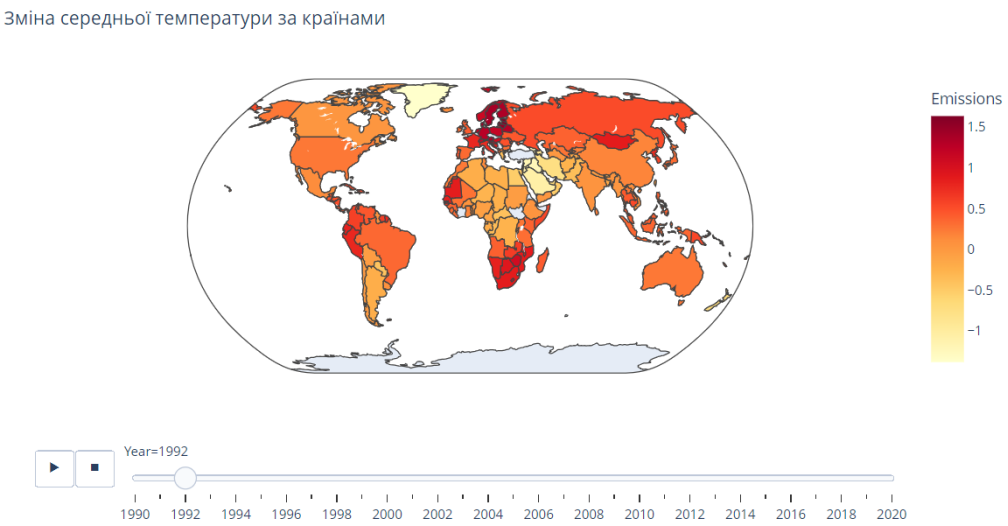

Рисунок 3.6 – зміна середньої температури в 1992 році

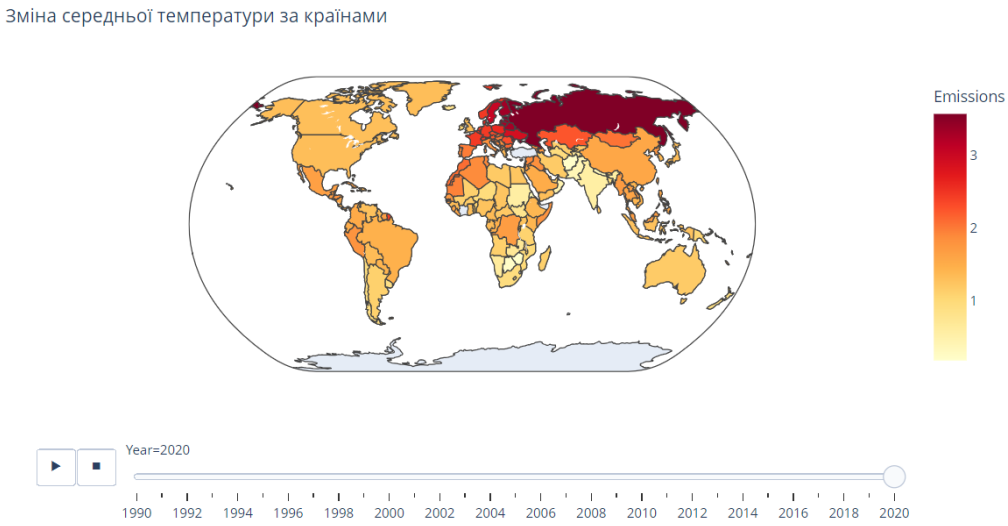

Рисунок 3.7 – зміна середньої температури в 2020 році

Порівнюючи ці два зображення, можна помітити суттєве зростання температурних змін за цей період. У 1992 році максимальне підвищення середньої температури в деяких країнах становило близько 1,6 градусів Цельсія. Натомість у 2020 році це значення зросло до 3,5 градусів, що свідчить про серйозне посилення кліматичних змін.

Крім того, якщо раніше підвищення температури мало більш розрізнений характер і відзначалося в різних частинах світу, то в останні роки воно стає більш зосередженим у певних регіонах. Зокрема, особливо помітне потепління спостерігається на території Європи, а також у частині Азії. Це може бути наслідком як глобальних кліматичних тенденцій, так і регіональних особливостей, пов'язаних із інтенсивністю антропогенного впливу та природними умовами.

Наступним кроком дослідження стало виявлення основних джерел викидів CO_2. Для цього дані було згруповано за роками, після чого розраховано сумарні обсяги викидів для кожної категорії джерел. На основі цих підрахунків побудовано графік (рис. 3.8), який дозволяє візуально оцінити внесок кожного джерела у загальну кількість викидів.

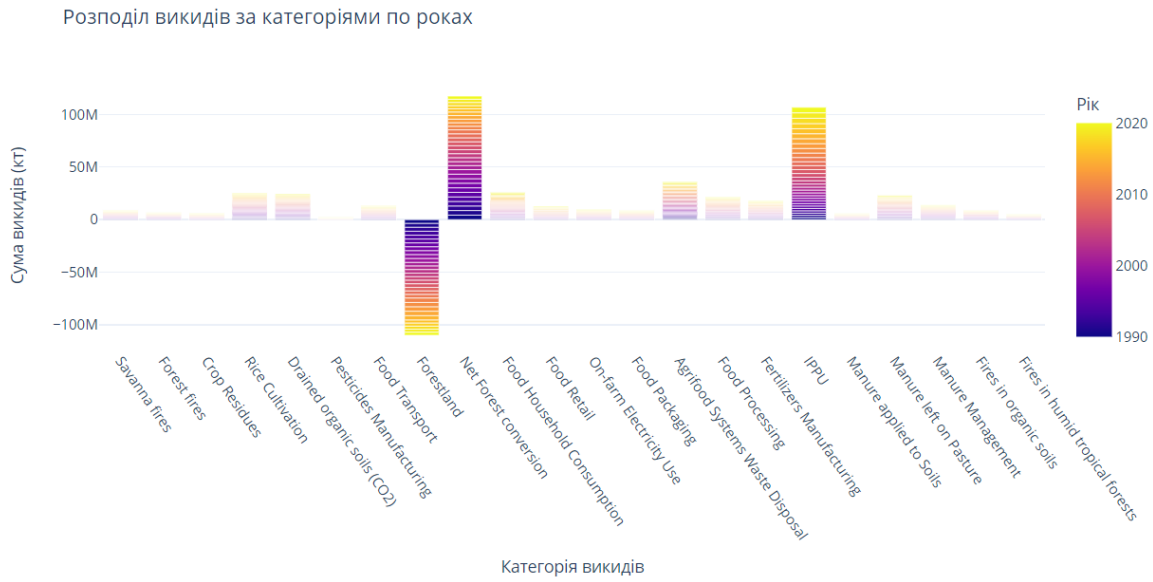

Рисунок 3.8 – Розподіл викидів за категоріями по роках

Як видно з рисунка, найбільшу частку викидів складають категорії, пов'язані зі зміною лісового покриву внаслідок вирубки, а також із промисловими процесами та використанням продукції. Це свідчить про значний вплив промисловості та змін у землекористуванні на загальну екологічну ситуацію. Варто звернути увагу на негативне значення викидів, пов'язане з лісами – воно вказує на їхню здатність поглинати CO_2 з атмосфери, що є вкрай важливою природною функцією в умовах зростаючих викидів.

Окрім цього, графік демонструє тенденцію до поступового зростання викидів у категорії промислових процесів, що є тривожним сигналом і потребує врахування при плануванні кліматичних стратегій. Візуалізація також дає змогу визначити, на які джерела варто звернути першочергову увагу в рамках заходів із зменшення викидів, і водночас підкреслює важливість збереження та відновлення лісів як одного з ключових способів боротьби з наслідками глобального потепління.

Наступна візуалізація, яку було побудовано, – це порівняння зміни середньої температури із загальними викидами CO_2 за роками (рис. 3.9). Для побудови графіка використовувалися середньорічні значення температури та обсяги загальних викидів, що дозволило наочно простежити динаміку змін обох показників у часовому розрізі.

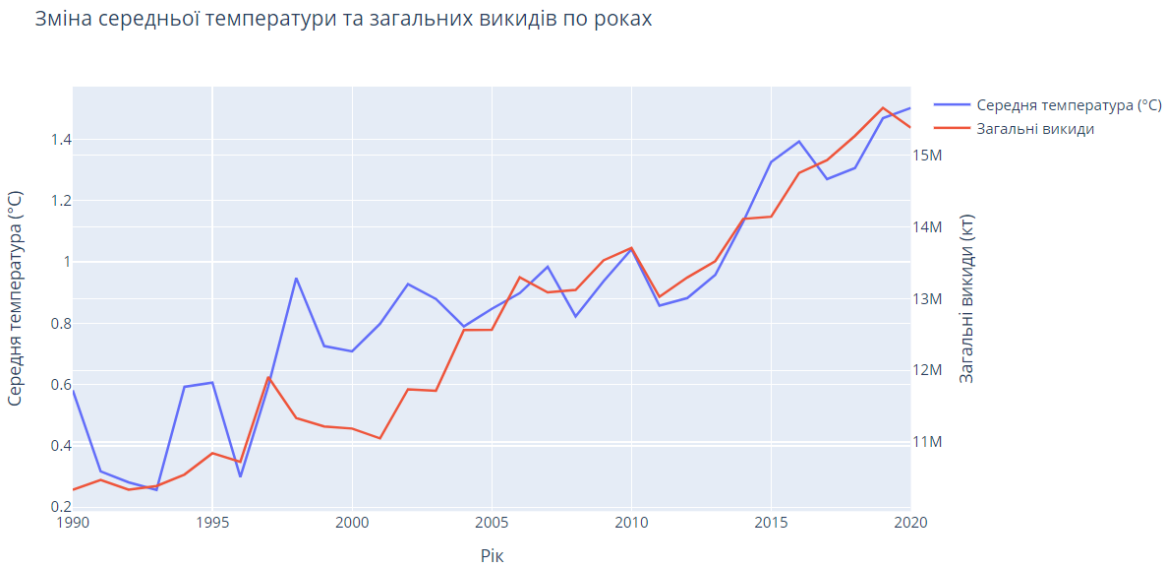

Рисунок 3.9 – Зміна середньої температури та загальних викидів по роках

Як видно з графіка, спостерігається чітка позитивна кореляція: із зростанням загального обсягу викидів CO_2 середня температура також демонструє стабільне підвищення. Це підтверджує наявність тісного зв’язку між діяльністю людини, зокрема спалюванням викопного палива, і глобальними кліматичними змінами. Така візуалізація слугує ще одним доказом важливості скорочення викидів парникових газів, щоб уповільнити темпи потепління та пом’якшити його наслідки для екосистем і людства.

Наступним графіком було побудовано порівняння зміни середньої температури та кількості викидів CO_2 за континентами та роками (рис. 3.10). Для побудови використовувалися агреговані дані по континентах, що дозволило побачити як загальносвітові тенденції, так і регіональні відмінності.

Рисунок 3.10 – Порівняння зміни температури та кількості викидів за континентами та роками

З графіка видно, що найбільший рівень викидів спостерігається в Азії, де з кожним роком кількість викидів продовжує зростати. Це, ймовірно, пов’язано з активною індустріалізацією, високою щільністю населення та швидким економічним розвитком у цьому регіоні. Водночас, у більшості континентів помітне також поступове підвищення середньої температури. Це свідчить про зв'язок між зростанням викидів і змінами клімату, а також підкреслює глобальний характер проблеми.

Отримані результати підтверджують необхідність вжиття подальших заходів щодо скорочення викидів, зокрема в регіонах із найвищим їх рівнем, для уповільнення глобального потепління та зменшення його негативних наслідків.

Після цього було вирішено виокремити топ-20 країн за сумарною кількістю викидів CO_2 за весь досліджуваний період, а також окремо проаналізувати ситуацію лише за 2020 рік. Це дозволяє побачити як

довгострокові тенденції, так і актуальну ситуацію на останній рік у вибірці (рис. 3.11 – 3.12).

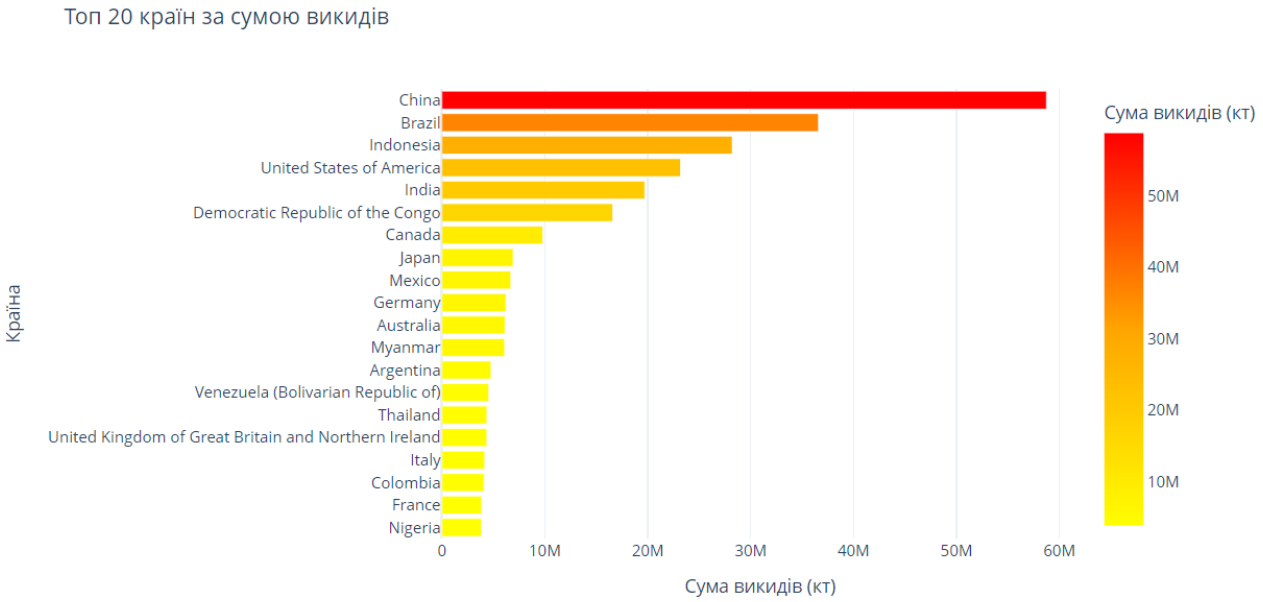

Рисунок 3.11 – топ-20 країн за сумарною кількістю викидів CO_2 за весь період

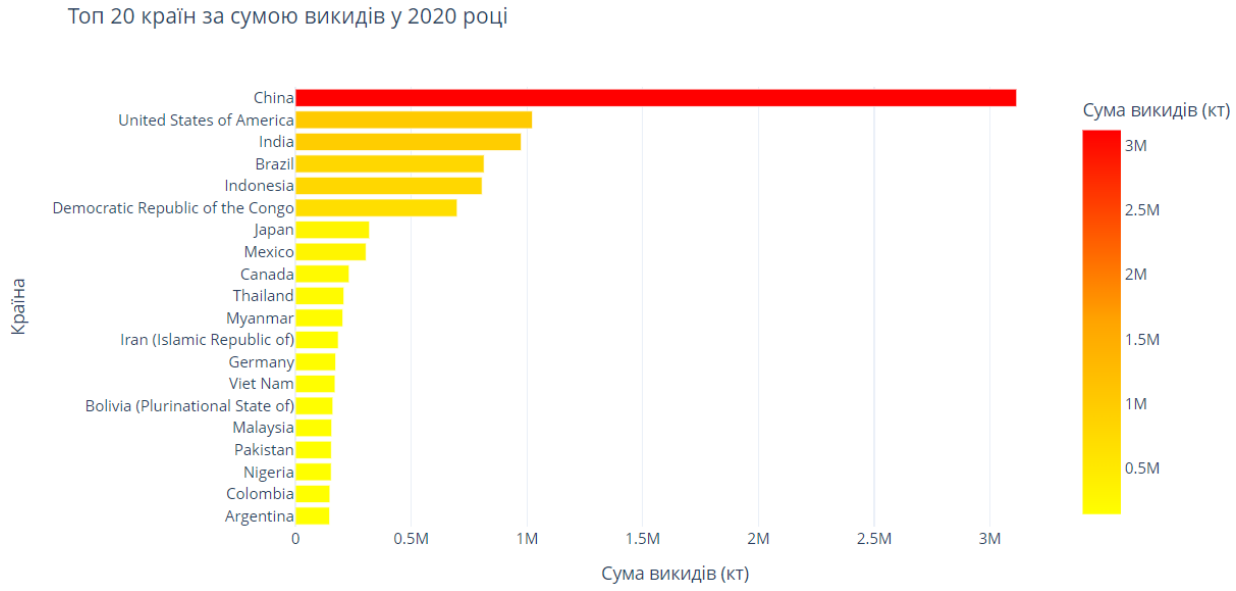

Рисунок 3.12 – топ-20 країн за кількістю викидів CO_2 у 2020 році

З рисунків видно, що перше місце за обома критеріями посідає Китай. За весь період ця країна стабільно генерує найбільшу кількість викидів CO_2, а у 2020 році вона випустила близько 3 мільйонів кт CO_2, що знову забезпечило їй лідерську позицію. Також цікавим є приклад Бразилії: за сумарними викидами вона знаходиться на другому місці, однак у 2020 році опустилася на четверту

позицію з показником близько 800 тисяч кт CO_2. Це може свідчити про певне зменшення темпів забруднення в останні роки або зміни в джерелах та структурі викидів.

Наступним кроком було сформовано рейтинг із 20 країн, які мають населення понад 5 мільйонів осіб, за показником викидів CO_2 на одну людину за 2020 рік. Такий підхід дозволяє оцінити рівень екологічного навантаження не лише в абсолютних цифрах, а й у розрахунку на душу населення, що є більш інформативним для порівняння між країнами з різною чисельністю населення (рис. 3.13).

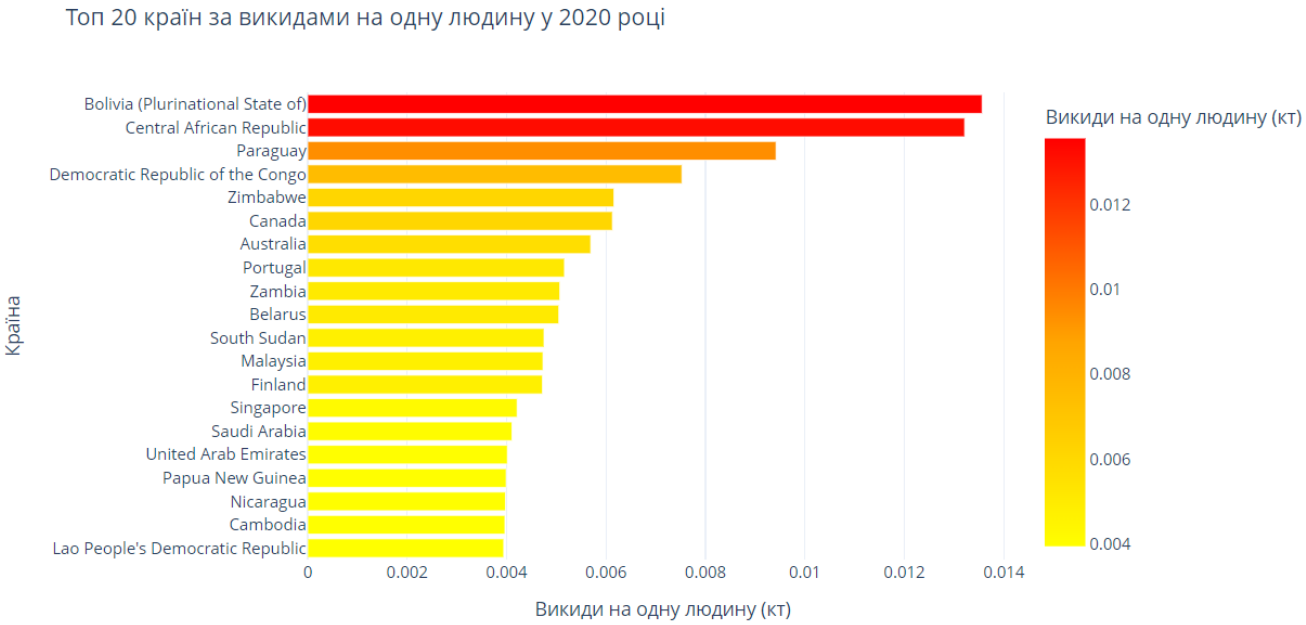

Рисунок 3.13 – топ-20 країн за викидами CO_2 на одну людину у 2020 році

На графіку видно, що перше місце посідає Болівія з показником 13,5 тонн CO_2 на одну людину. Це досить високий результат, який може бути зумовлений як структурою економіки, так і способом ведення господарської діяльності. На протилежному боці рейтингу знаходиться Лаос із 3,9 тонни на людину. Такий розрив між країнами демонструє нерівномірність навантаження на клімат з боку окремих держав і підкреслює важливість індивідуального підходу до формування екологічної політики та кліматичних ініціатив.

Було побудовано розподіл середньої температури по роках для кращого розуміння динаміки її змін та варіацій у різних країнах (рис. 3.14).

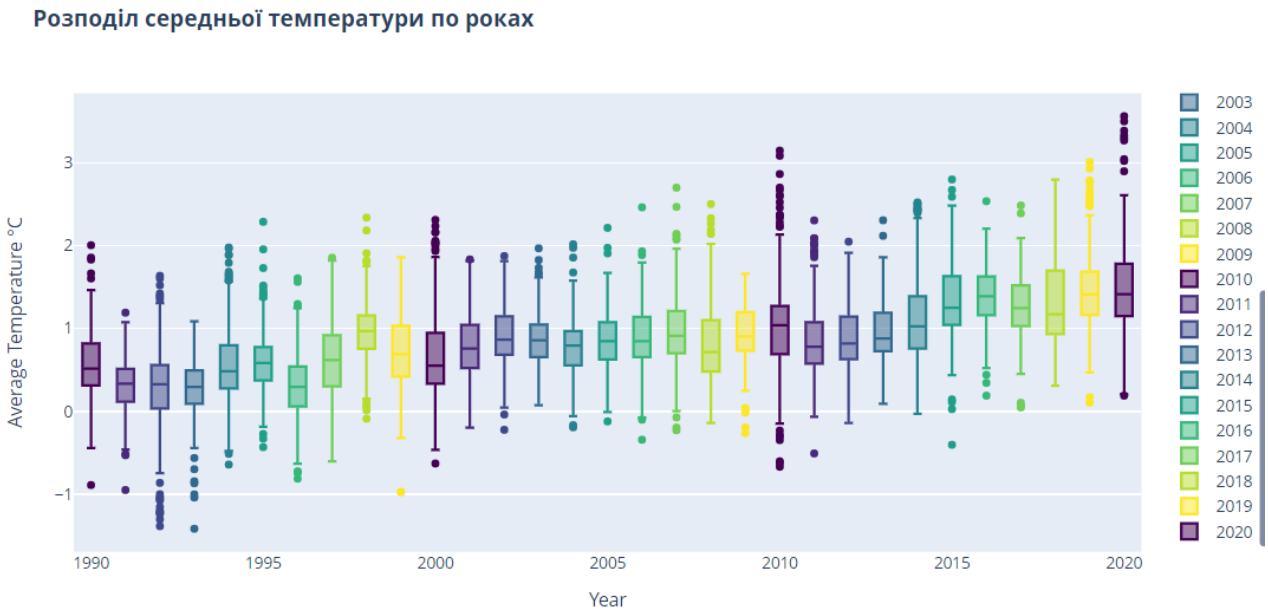

Рисунок 3.14 - розподіл середньої температури по роках

На цьому графіку можна побачити мінімальні, максимальні та медіанні значення зміни середньої температури для кожного року. Такий тип візуалізації дає змогу оцінити не лише загальну тенденцію змін температури, а й ступінь її коливань у межах одного року. Крім того, графік демонструє, наскільки значення змін відхиляються від медіани — це дозволяє побачити, чи є певні країни з аномально високими або низькими змінами температури. Це також може вказувати на локальні кліматичні особливості або специфічні екологічні фактори, які впливають на динаміку температури в певних регіонах.

Останнім видом графіків, що були побудовані, є теплові мапи, які використовуються для виявлення кореляцій між різними змінними. Теплові мапи були створені для даних, розподілених по континентах, а також для згрупованих даних по всьому світу та по роках. У звіті представлена лише одна з цих теплових мап – для згрупованих даних по всьому світу, яка дає змогу вивчити взаємозв’язки між різними факторами на глобальному рівні протягом усіх років (рис. 3.15).

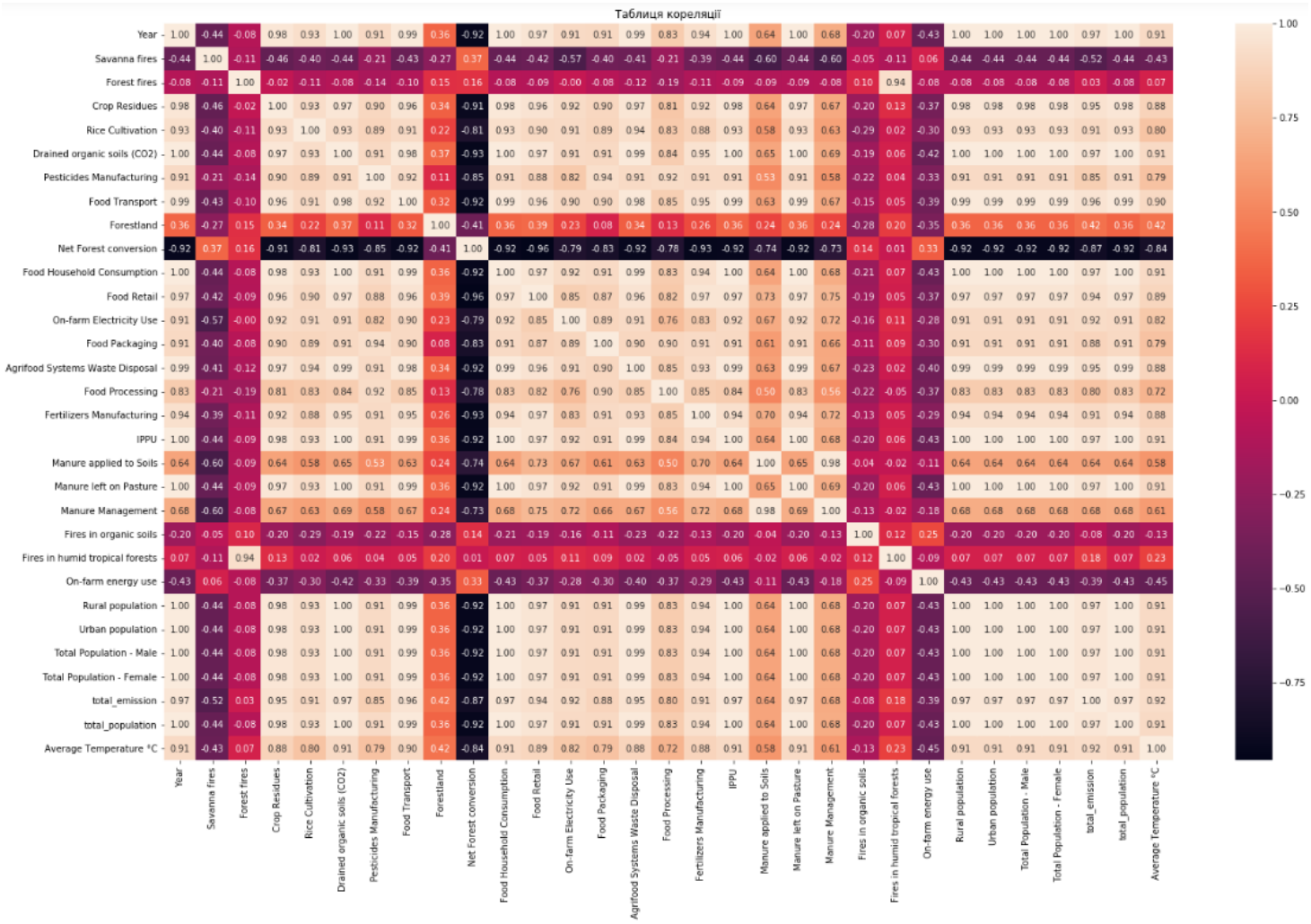

Рисунок 3.15 – теплова мапа для коефіцієнтів кореляції

З рисунка 3.15 видно, що особливу увагу варто звернути на останній рядок теплової мапи, який демонструє, як різні джерела викидів корелюють зі зміною середньої температури. Коефіцієнти кореляції, наближаючись до значення 1, вказують на те, що при збільшенні викидів з певного джерела середня температура також зростатиме. Це підкреслює тісний взаємозв'язок між окремими джерелами викидів і зміною клімату.

Особливо сильні кореляції спостерігаються для більшості факторів викидів, окрім таких, як викиди від лісових пожеж, викиди від пожеж в органічних ґрунтах та викиди від пожеж у вологих тропічних лісах. Це може свідчити про те, що ці фактори мають менший прямий вплив на загальну температуру в порівнянні з іншими джерелами викидів.

Проте є важливе зауваження щодо цієї теплової мапи: вона показує, що наші дані є мультиколінеарними, що означає, що багато змінних сильно корелюють одна з одною. Це може призвести до певних труднощів при застосуванні методів лінійної регресії, оскільки вони не здатні точно виділити окремі фактори, які найбільше впливають на зміну температури. Для вирішення цієї проблеми буде доцільно застосувати більш складні методи, зокрема, побудовані на основі дерев рішень, які дозволяють більш ефективно обробляти такі корельовані дані.

3.2 Побудова моделей

Перед побудовою моделей було проведено підготовку даних: з усього доступного глобального набору даних було відібрано 25 релевантних ознак, що охоплюють різні джерела викидів CO_2, які потенційно впливають на зміну середньої температури. Аналіз та моделювання на цьому етапі здійснювалися для узагальненої картини по всьому світу, а цільовою змінною виступала середня температура повітря (°C) (рис. 3.16).

```
features = [
    'Savanna fires', 'Forest fires', 'Crop Residues', 'Rice Cultivation', 'Drained organic soils (CO2)',
    'Pesticides Manufacturing', 'Food Transport', 'Forestland', 'Net Forest conversion', 'Food Household Consumption',
    'Food Retail', 'On-farm Electricity Use', 'Food Packaging', 'Agrifood Systems Waste Disposal', 'Food Processing',
    'Fertilizers Manufacturing', 'IPPU', 'Manure applied to Soils', 'Manure left on Pasture', 'Manure Management',
    'Fires in organic soils', 'Fires in humid tropical forests', 'On-farm energy use', 'total_emission',
    'total_population']

target = 'Average Temperature °C'
X = df_grouped[features]
y = df_grouped[target]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=41)
```

Рисунок 3.16 – відібрані ознаки

Перша побудована модель була основана на випадковому лісі. Для побудови спочатку було поділено дані на навчальну (80%) та тестову (20%) вибірки за допомогою функції train_test_split() (рис. 3.16). Наступним кроком було використання класу RandomForestRegressor() з 2000 деревами та мінімальною кількістю зразків для розщеплення – 4. Також було вирішено не змінювати параметри за замовчуванням, бо при їх зміні модель видавала гірші результати: критерій для оцінки якості розщеплення – середньоквадратична помилка, максимальна глибина дерева – без обмежень, бо датасет має всього 30

рядків після групування, мінімальна кількість зразків необхідних у листовому вузлі – 1, використання бутстрапу. Після цього модель була навчена на тренувальних даних за допомогою функції fit(), а прогнозування було виконано на тестовій вибірці за допомогою функції predict() (рис. 3.17).

```
model = RandomForestRegressor(n_estimators=2000, min_samples_split = 4, random_state=42)
model.fit(X_train, y_train)

y_pred = model.predict(X_test)
```

Рисунок 3.17 – модель Random Forest Regressor

Наступною моделлю для порівняння була основана на градієнтному бустингу. Розподілені дані використовувались ті ж самі, що й у попередній моделі (рис. 3.16). Для цієї моделі було використано клас XGBRegressor() з параметрами навчальної швидкості , який використовується для оновлення, щоб уникнути перенавчання – 0.5. Також в цьому методі було вирішено залишити інші параметри за замовчуванням, бо при їх зміні модель видавала гірші результати: максимальна глибина дерева – 6, мінімальне зменшення втрат, необхідне для створення подальшого розділення листового вузла дерева – 0, мінімальна сума ваги екземпляра, необхідна для листа = 1 та L2 член регуляризації ваг – 1, L1 член регуляризації ваг – 0. Після цього модель була навчена на тренувальних даних за допомогою функції fit(), а прогнозування було виконано на тестовій вибірці за допомогою функції predict() (рис. 3.18).

```
model = XGBRegressor(learning_rate = 0.5)
model.fit(X_train, y_train)

y_pred = model.predict(X_test)
```

Рисунок 3.18 – модель XGBRegressor

Останньою побудованою моделлю була модель на основі рекурентної нейронної мережі (RNN), зокрема, довготривалої короткочасної пам’яті (LSTM) (рис. 3.19). Для підвищення якості прогнозування були додатково зібрані та об’єднані дані про зміну середньої температури за період з 1960 по 2024 рік, що дозволило розширити часовий горизонт аналізу. Спочатку ці дані було

нормалізовано за допомогою функції MinMaxScaler(). Далі було сформовано датасет, який містив вхідні дані на основі попередніх значень температури з врахуванням кроку – 10. Далі перетворено вхідні дані до необхідного формату для LSTM.

Наступний крок складався зі створення самої моделі. Спочатку ми використали клас Sequential() для створення моделі та додавання наступних шарів до моделі. Перший шар – LSTM, який складався з 50 нейронів і приймав на вхід матрицю з розмірністю (10, 1). Останній шар цієї моделі був Dense(), який містив один нейрон, що видавав нам результат (прогноз). Після створення моделі, її було скомпільовано з використанням функції втрат mean_squared_error та оптимізатором adam.

Останній крок – навчання, воно відбувалося з 20 епохами та розміром пакету, тобто кількість прикладів, які модель обробляє одночасно перед оновленням ваг моделі – 2 (рис. 3.20).

```
scaler = MinMaxScaler(feature_range=(0, 1))
dataset = scaler.fit_transform(np.array(df_aggregated['Average Temperature °C']).reshape(-1, 1))

def create_dataset(dataset, look_back=1):
    X, Y = [], []
    for i in range(len(dataset)-look_back):
        X.append(dataset[i:(i+look_back), 0])
        Y.append(dataset[i + look_back, 0])
    return np.array(X), np.array(Y)

look_back = 10
X, Y = create_dataset(dataset, look_back)
X = np.reshape(X, (X.shape[0], X.shape[1], 1))

model = Sequential()
model.add(LSTM(50, input_shape=(look_back, 1)))
model.add(Dense(1))
model.compile(loss='mean_squared_error', optimizer='adam')

model.fit(X, Y, epochs=20, batch_size=2, verbose=2)

trainPredict = model.predict(X)
trainPredict = scaler.inverse_transform(trainPredict)
```

Рисунок 3.19 – модель LSTM

```
Epoch 15/20
27/27 - 0s - loss: 0.0060 - 113ms/epoch - 4ms/step
Epoch 16/20
27/27 - 0s - loss: 0.0058 - 116ms/epoch - 4ms/step
Epoch 17/20
27/27 - 0s - loss: 0.0066 - 107ms/epoch - 4ms/step
Epoch 18/20
27/27 - 0s - loss: 0.0065 - 106ms/epoch - 4ms/step
Epoch 19/20
27/27 - 0s - loss: 0.0058 - 113ms/epoch - 4ms/step
Epoch 20/20
27/27 - 0s - loss: 0.0060 - 105ms/epoch - 4ms/step
```

Рисунок 3.20 – процес навчання моделі

Наступним етапом дослідження стало побудува моделей для прогнозування зміни середньої температури в кожній окремій країні. Такий підхід дозволив урахувати специфічні особливості та чинники, які впливають на

температурні зміни в різних географічних регіонах. Було реалізовано два варіанти моделей: модель на основі алгоритму Random Forest із підбором гіперпараметрів та комбіновану модель LSTM + ARIMA, що дозволяє враховувати часову структуру даних.

У першому підході застосовувався алгоритм Random Forest Regressor для побудови моделей, які прогнозують середню температуру на основі набору із 24 ознак, що відображають різні джерела парникових викидів (рис. 3.16). Для кожної країни було відібрано відповідні записи з датасету, після чого дані було розділено на тренувальну та тестову вибірки у співвідношенні 90% до 10%.

Підбір гіперпараметрів моделі здійснювався за допомогою методу повного перебору (GridSearchCV), у межах якого тестувались різні комбінації параметрів, такі як кількість дерев (n_estimators), максимальна глибина дерев (max_depth), мінімальна кількість зразків для поділу (min_samples_split), кількість ознак, що враховуються при кожному розбитті (max_features), та інші (рис. 3.21). Пошук оптимальних значень здійснювався з використанням крос-валідації (2 фолди) з метою мінімізації середньоквадратичної помилки (рис. 3.22).

```
param_grid = {
    'n_estimators': [150, 250, 350, 450],
    'min_samples_split': [3, 4, 5],
    'max_depth': [3, 4, 5],
    'min_samples_leaf': [2, 3, 4],
    'max_features': ['sqrt', 'log2', 0.3, 0.5],
    'bootstrap': [True],
    'max_samples': [0.8, 0.9, 1],
    'random_state': [42]
}
```

Рисунок 3.21 – гіперпараметри для підбору

```
X_train, X_test, y_train, y_test = train_test_split(X_country, y_country, test_size=0.1, random_state=42)

model = RandomForestRegressor(random_state=42)

grid_search = GridSearchCV(estimator=model, param_grid=param_grid, cv=2, scoring='neg_mean_squared_error', n_jobs=-1, verbos

grid_search.fit(X_train, y_train)
```

Рисунок 3.22 – підбір гіперпараметрів

Після визначення найкращих гіперпараметрів, модель для кожної країни вчилася на повному тренувальному наборі. Крім того, для кожної побудованої моделі визначалась важливість вхідних ознак, що дало змогу проаналізувати, які типи викидів мають найбільший вплив на зміну температури в кожній конкретній країні (рис. 3.23).

```
X_train, X_test, y_train, y_test = train_test_split(X_country, y_country, test_size=0.1, random_state=42)

model = RandomForestRegressor(**best_params_by_country[country])
model.fit(X_train, y_train)
```

Рисунок 3.23 – модель для знаходження найвпливовіших параметрів для кожної країни

У другому підході для кожної країни було сформовано часовий ряд зміни середньої температури, який попередньо згладжувався за допомогою ковзного середнього з вікном у 4 роки. Дані було нормалізовано до інтервалу [0, 1] із використанням масштабування MinMaxScaler. Для моделі LSTM було обрано стратегію побудови набору даних зі зсувом (look_back) у 10 років — це означає, що модель навчалась прогнозувати наступне значення на основі попередніх 10 спостережень.

```
country_data = temp[temp['Area'] == country]

country_data['Smoothed_Temp'] = country_data['Average Temperature °C'].rolling(window=4).mean()

country_data = country_data.dropna(subset=['Smoothed_Temp'])

scaler = MinMaxScaler(feature_range=(0, 1))
dataset = scaler.fit_transform(np.array(country_data['Smoothed_Temp']).reshape(-1, 1))

look_back = 10
```

Рисунок 3.24 – підготовка часових рядів

Архітектура мережі включала два послідовні шари LSTM (по 64 та 32 нейрони відповідно) з шарами Dropout між ними, що забезпечувало зменшення перенавчання. На виході використовувався щільний шар (Dense) з одним нейроном, який генерував прогнозоване значення температури. Модель навчалась протягом 20 епох з малим розміром пакету (batch_size = 2), що забезпечувало гнучке оновлення ваг у процесі тренування (рис. 3.25).

```
model = Sequential()
model.add(LSTM(64, return_sequences=True))
model.add(Dropout(0.1))
model.add(LSTM(32, return_sequences=False))
model.add(Dropout(0.1))
model.add(Dense(1))
model.compile(loss='mean_squared_error', optimizer='adam')

model.fit(X, Y, epochs=20, batch_size=2, verbose=2)
```

Рисунок 3.25 – створена модель LSTM

Після формування початкового прогнозу за допомогою моделі LSTM було розраховано залишки – різницю між реальними та прогнозованими значеннями, яка продемонструвала наявність автокореляційної залежності. Для врахування цієї структури було використано модель ARIMA з параметрами (3, 0, 2). У даній конфігурації параметр p = 3 свідчить про те, що модель враховує три попередні значення залишків при формуванні поточного прогнозу, що відображає присутність автокореляції до третього лагу. Значення d = 0 вказує на те, що ряд залишків є стаціонарним і не потребує додаткового диференціювання. Параметр q = 2 означає врахування двох попередніх значень випадкових похибок, що сприяє більш точному відображенню короткострокових коливань. Прогнозовані значення залишків було додано до результатів моделі LSTM, що дозволило підвищити загальну точність прогнозу та компенсувати окремі недоліки нелінійного моделювання (рис. 3.26).

```
residuals = country_data['Average Temperature °C'].iloc[look_back:] - trainPredict[:, 0]
arima_model = ARIMA(residuals, order=(3, 0, 2))  # Простий ARIMA (1,0,0)
arima_fit = arima_model.fit()
arima_preds = arima_fit.predict(start=0, end=len(residuals)-1, typ='levels')

arima_corrected_preds = trainPredict[:, 0] + arima_preds
```

Рисунок 3.26 – створена модель ARIMA

3.3 Оцінка якості та результатів

Для оцінки якості моделей використовувались метрики середньоквадратичної помилки (MSE), кореня з середньоквадратичної помилки (RMSE) та коефіцієнта детермінації (R^2).

Оцінювання моделей проводилось як для загальних даних, так і окремо для кожної країни. Це дозволило проаналізувати, наскільки якісно модель здатна узагальнювати залежності у глобальних даних і водночас адаптуватися до особливостей даних окремих країн.

Перша модель на основі випадкового лісу дала, такі результати (рис. 3.27):

MSE = 0.01547

RMSE = 0.1244

R^2 = 0.8465

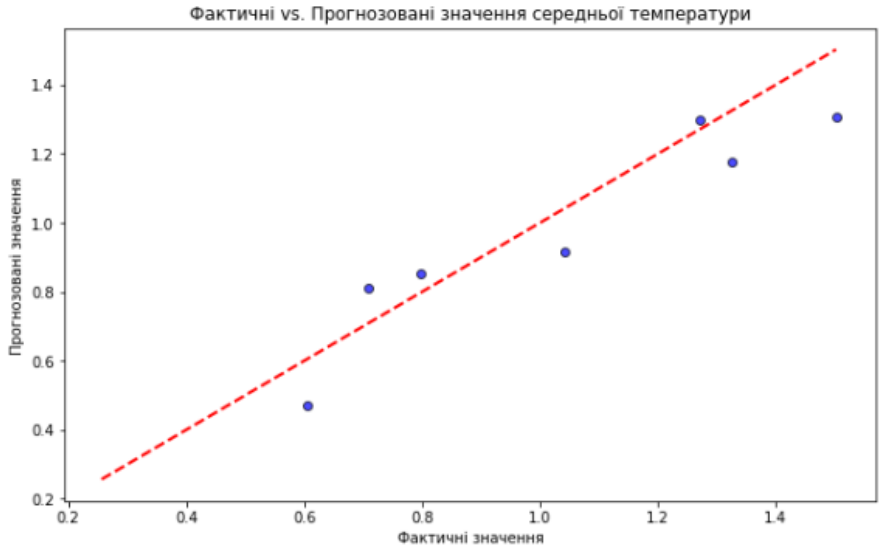

Рисунок 3.27 – результати першої моделі

Друга модель на основі градієнтного бустингу дала, такі результати (рис. 3.28):

MSE = 0.0182

RMSE = 0.135

R^2 = 0.8189

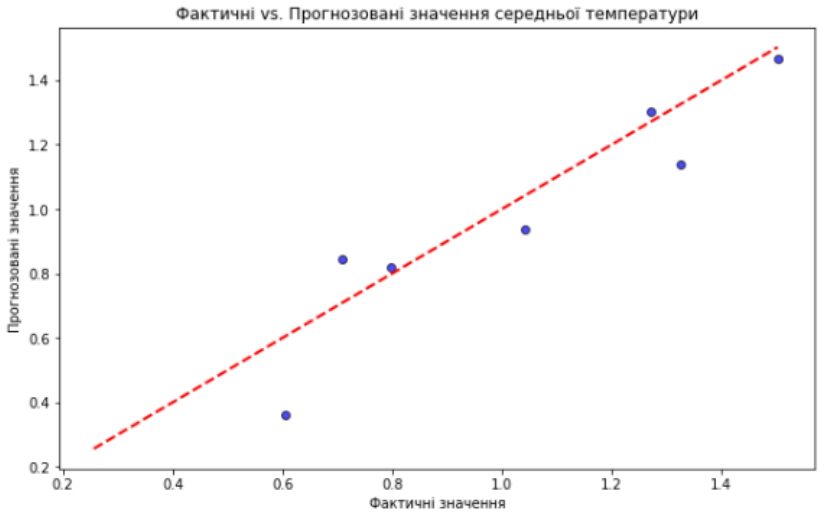

Рисунок 3.28 – результати другої моделі

З наведених результатів видно, що обидві моделі мають подібну продуктивність. Вони дали значення R^2 близько 0.82–0.84, що вказує на те, що моделі добре пояснюють більшу частину варіації у даних. Крім того, середньоквадратична помилка і корінь з неї є відносно невеликими, що вказує на

те, що прогнози можуть мати похибку приблизно 0.12–0.14 градусів. Це свідчить про достатню точність обох моделей для завдання прогнозування зміни середньої температури.

Також було проведено дослідження важливості параметрів, які впливають на зміну середньої температури. Обидві моделі дозволили оцінити, які саме чинники є найбільш значущими, однак результати мали певні відмінності (рис. 3.29–3.30):

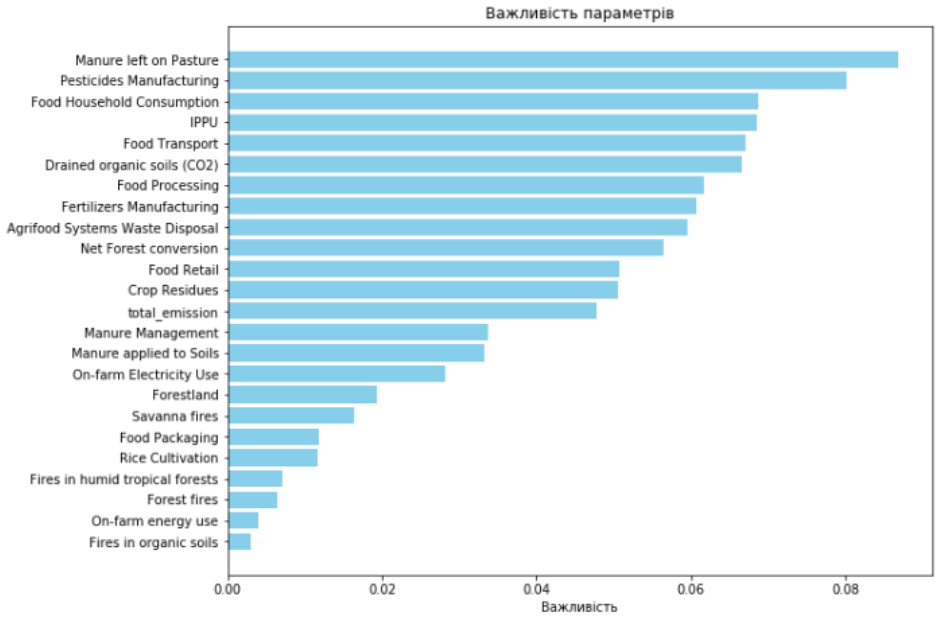

Рисунок 3.29 – важливість параметрів для першої моделі

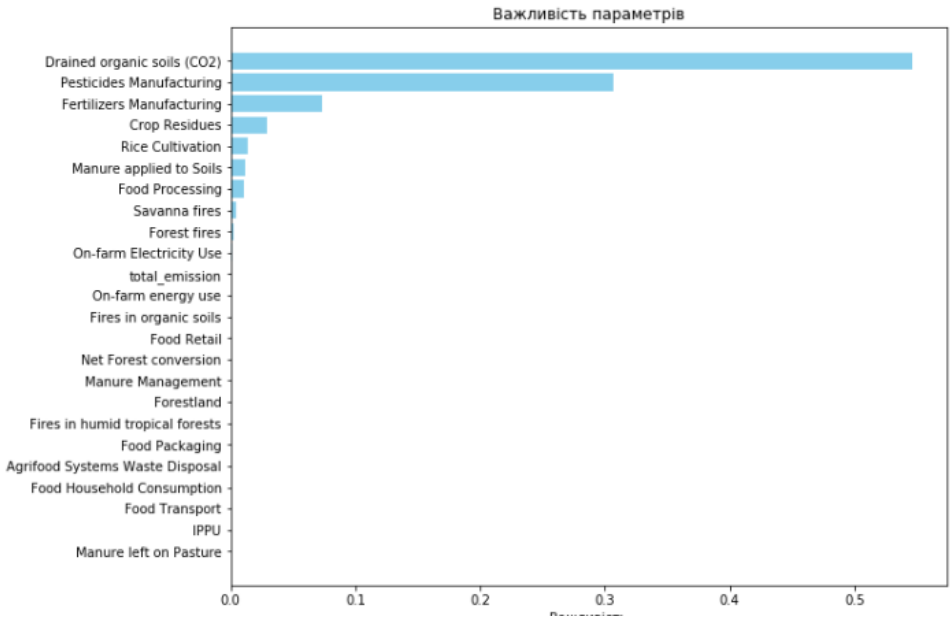

Рисунок 3.30 – важливість параметрів для другої моделі

З результатів видно, що модель на основі випадкового лісу виявила ширше коло важливих ознак, розподіливши значущість між більшою кількістю параметрів. Це може свідчити про здатність моделі виявляти комплексні залежності між різними джерелами викидів і температурою. Натомість модель на основі градієнтного бустингу зосередилася переважно на двох ключових параметрах, вважаючи інші менш впливовими або майже незначущими у контексті прогнозування. Такий підхід може бути корисним для виявлення домінантних факторів, але менш ефективним у врахуванні менш виражених взаємозв'язків. Таким чином, використання різних моделей дозволяє отримати більш комплексне та глибоке розуміння того, як різні типи викидів впливають на зміну температури. Такий підхід забезпечує надійну основу для формування рекомендацій щодо пріоритетних напрямів скорочення викидів парникових газів (табл. 3.2).

Таблиця 3.2 – порівняння важливостей параметрів моделей

Перша модель		**Друга модель**	
Параметр	**Важливість**	**Параметр**	**Важливість**
Викиди від гною тварин на пасовищах	0.087	Викиди, що виділяються під час осушення органічних ґрунтів	0.546
Викиди від виробництво пестицидів	0.08	Викиди від виробництво пестицидів	0.307
Викиди від споживання продуктів харчування на рівні домогосподарств	0.0686	Викиди від виробництва добрив	0.073
Викиди від промислових процесів і використання продукції	0.0685	Викиди від спалювання або розкладання залишків рослинного матеріалу після збору врожаю	0.029
Викиди, від транспортування харчових продуктів	0.067	Викиди від вирощування рису	0.014

Результати порівняння важливості параметрів двох моделей показують, що обидві моделі виділяють викиди від виробництва пестицидів як значущий фактор, що впливає на зміну клімату. У першій моделі на основі випадкового

лісу найбільш важливими є також викиди від гною тварин на пасовищах, тоді як у другій моделі на основі градієнтного бустингу найбільший вплив мають викиди, що виділяються під час осушення органічних ґрунтів. Це свідчить про те, що різні моделі по-різному оцінюють вплив факторів, виявляючи певні параметри як більш значущі, що може бути результатом різниці в алгоритмах або в самих даних.

Наступною моделлю була рекурентна нейронна мережа LSTM, вона показала такі результати (рис. 3.31):

MSE = 0.03

RMSE = 0.17

R^2 = 0.9

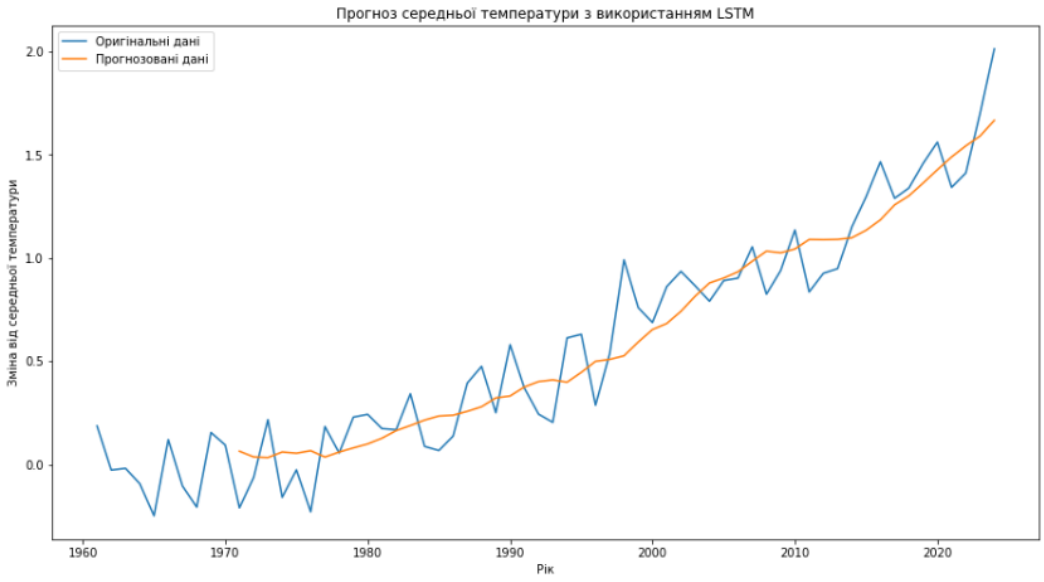

Рисунок 3.31 – порівняння результатів моделі LSTM

На рисунку 3.31 видно, що є відмінності від оригінальних даних, але тенденцію росту середньої температури все одно видно.

Також за допомогою цієї нейронної мережі було зроблено прогноз на 10 років вперед, щоб дізнатися, як зміниться середня температура (рис. 3.32).

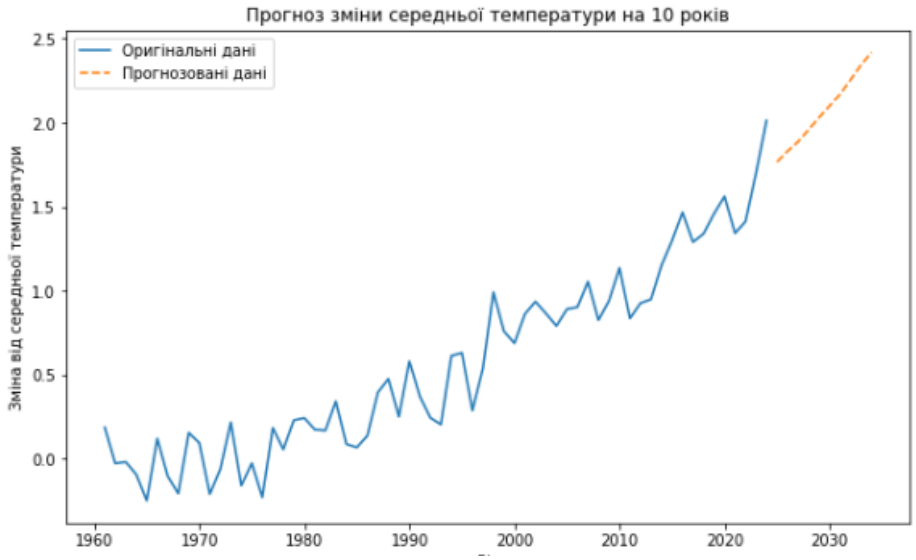

Рисунок 3.32 – прогноз на 10 років вперед

На основі рисунку 3.32 можна зробити висновок, що температура підвищиться через 10 років приблизно на 0.8 градуси. Це демонструє, що неперервне збільшення викидів парникових газів може вплинути на зростання середньорічної температури планети, відоме як глобальне потепління. Такий розвиток подій може мати серйозні наслідки для клімату, довкілля та життя на Землі, включаючи зміни в розподілі опадів, підвищення рівня морів, зміни в екосистемах та загрози для біорізноманіття. Однак цей прогноз може бути змінений за умови впровадження ефективних заходів для зменшення викидів та адаптації до змін клімату.

Після аналізу результатів глобальних моделей, наступним кроком стало оцінювання якості прогнозування для окремих країн. Це дозволяє виявити, наскільки стабільно моделі працюють у різних регіонах світу та які країни мають найточніші або найменш точні прогнози. Для цього було побудовано окремі моделі на основі випадкового лісу для кожної країни, і проведено оцінювання їхньої ефективності за допомогою метрик R^2 та RMSE.

```
R2 < 0: 105 країн
0 ≤ R2 < 0.2: 32 країн
0.2 ≤ R2 < 0.5: 37 країн
R2 ≥ 0.5: 53 країн
```

Рисунок 3.33 – порівняння коефіцієнта R^2

На рисунку 3.33 видно, що в 105 країнах значення коефіцієнта детермінації R^2 було менше нуля, що вказує на те, що модель гірше прогнозує температуру, ніж просто середнє значення. Ще у 32 країнах R^2 було між 0 і 0.2, у 37 країнах – між 0.2 і 0.5, що свідчить про слабке або помірне пояснення варіацій температури. Лише в 53 країнах R^2 перевищило 0.5, що вважається прийнятним рівнем якості моделі.

```
RMSE < 0.1: 17 країн
0.1 ≤ RMSE < 0.2: 73 країн
0.2 ≤ RMSE < 0.3: 60 країн
0.3 ≤ RMSE < 0.4: 30 країн
0.4 ≤ RMSE < 0.5: 29 країн
RMSE ≥ 0.5: 21 країн
```

Рисунок 3.34 – порівняння RMSE

На рисунку 3.34 наведено результати оцінки середньої квадратичної помилки (RMSE). У 17 країнах RMSE була меншою за 0.1, що свідчить про високу точність прогнозу. У 73 країнах вона коливалася між 0.1 та 0.2, у 60 країнах – між 0.2 та 0.3. Ще у 30 країнах RMSE становила від 0.3 до 0.4, у 29 – від 0.4 до 0.5, а у 21 країні перевищувала 0.5, що вказує на значні похибки прогнозування.

Для ілюстрації розглянемо результати для України та Бразилії, аби визначити найбільш впливові викиди в кожній країні.

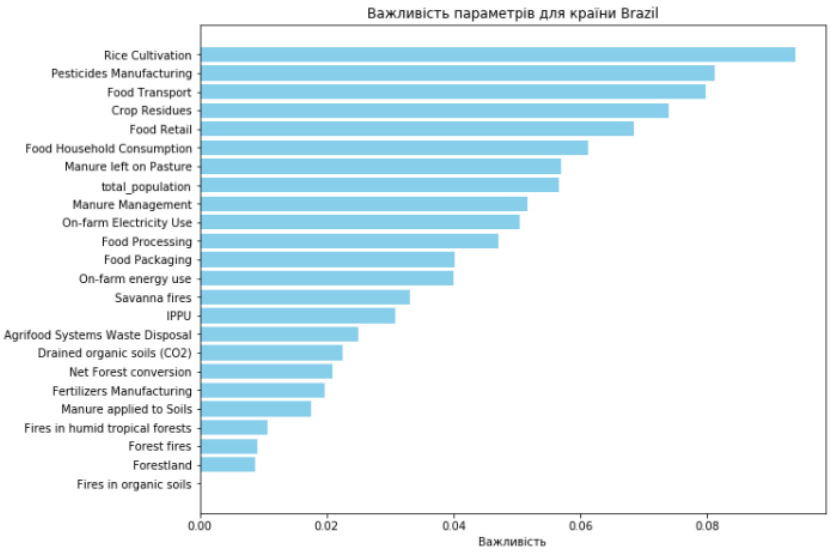

Рисунок 3.35 – найвпливовіші викиди в Бразилії

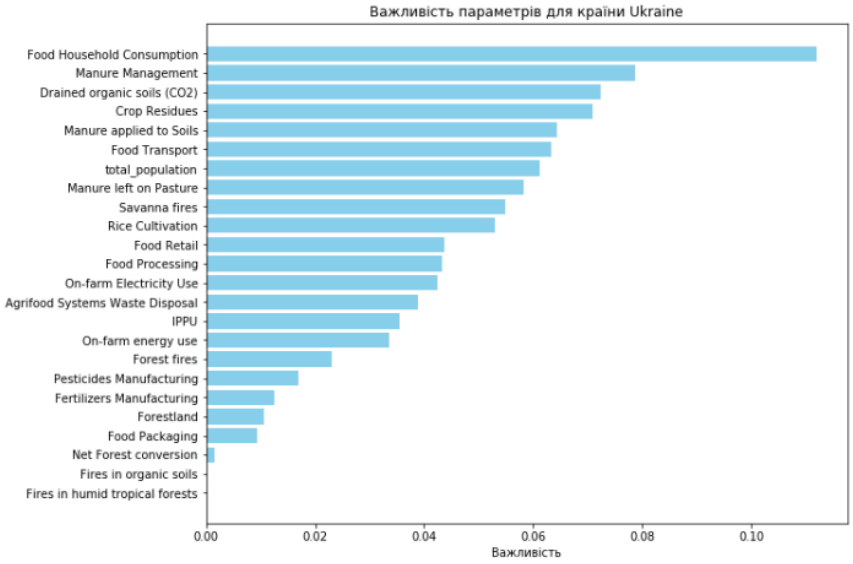

Рисунок 3.36 – найвпливовіші викиди в Україні

У випадку Бразилії найбільший вплив на зміну температури мають викиди, пов'язані з вирощуванням рису, тоді як в Україні домінуючим фактором є викиди від споживання продуктів харчування на рівні домогосподарств.

Загалом, модель випадкового лісу показала добру точність лише в частині країн, тоді як в інших – значні відхилення, що свідчить про потребу адаптації моделі або використання додаткових локальних факторів для покращення прогнозування на рівні окремих держав.

Після оцінки результатів моделі випадкового лісу для окремих країн, було проведено також тестування альтернативного підходу на основі нейронних мереж. Зокрема, розглядалася модель LSTM, до якої було додано етап корекції залишків за допомогою моделі ARIMA.

```
R2 < 0.5: 31 країн
0.5 ≤ R2 < 0.6: 43 країн
0.6 ≤ R2 < 0.7: 65 країн
0.7 ≤ R2 < 0.8: 53 країн
R2 ≥ 0.8: 17 країн
```

Рисунок 3.37 – порівняння коефіцієнта R^2

Результати тестування гібридної моделі LSTM з корекцією залишків за допомогою ARIMA (рис. 3.37) демонструють покращення якості прогнозування у порівнянні з моделлю випадкового лісу для багатьох країн. Зокрема, у 135

країнах значення R^2 перевищило 0.5, що свідчить про задовільну або високу якість моделі. З них у 43 країнах R^2 перебувало в межах від 0.5 до 0.6, у 65 – від 0.6 до 0.7, у 53 – від 0.7 до 0.8, а у 17 країнах коефіцієнт детермінації перевищив 0.8, що є ознакою дуже хорошої відповідності моделі до реальних даних. Водночас у 31 країні R^2 залишалося нижчим за 0.5, що вказує на потребу подальшого удосконалення прогнозної моделі для цих регіонів.

```
RMSE < 0.1: 0 країн
0.1 ≤ RMSE < 0.2: 11 країн
0.2 ≤ RMSE < 0.3: 86 країн
0.3 ≤ RMSE < 0.4: 39 країн
0.4 ≤ RMSE < 0.5: 46 країн
RMSE ≥ 0.5: 27 країн
```

Рисунок 3.38 – порівняння RMSE

На рисунку 3.38 представлено аналіз середньої квадратичної помилки (RMSE). Результати показують, що хоча жодна країна не досягла рівня RMSE < 0.1, у більшості випадків точність залишалася прийнятною. Зокрема, у 86 країнах RMSE знаходилася в межах від 0.2 до 0.3, що є добрим показником для температурного прогнозування. У 39 країнах RMSE коливалася від 0.3 до 0.4, у 46 – від 0.4 до 0.5, а у 27 країнах значення помилки перевищувало 0.5, що свідчить про складність прогнозування температури в окремих регіонах або про наявність нерозкритих факторів, які впливають на зміну клімату.

Для ілюстрації розглянемо результати прогнозування зміни середньої температури на найближчі 10 років для двох країн — України та Бразилії.

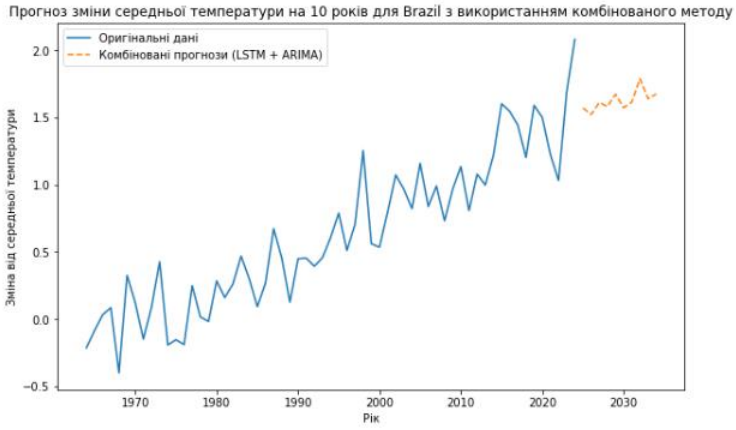

Рисунок 3.39 – прогноз в Бразилії на 10 років

Рисунок 3.40 – прогноз в Україні на 10 років

У Бразилії модель прогнозує початкове зниження температури приблизно на 0.5°C, після чого очікується поступове зростання на 0.1°C (рис. 3.39). В Україні ж прогноз показує спочатку спад на 0.4°C, який повторюється ще раз, але надалі спостерігається суттєве зростання температури на 1.8°C у кінцевому періоді прогнозу (рис. 3.40).

Такий контраст між країнами підкреслює різну динаміку кліматичних змін у залежності від регіональних особливостей і підкреслює необхідність врахування локальних трендів при формуванні кліматичних стратегій.

Загалом результати гібридної моделі свідчать про її вищу ефективність у порівнянні з класичними методами в більшості випадків. Поєднання здатності LSTM виявляти складні нелінійні залежності з можливістю ARIMA моделювати залишкові лінійні тренди забезпечує точніше узгодження з фактичними даними. Це робить гібридний підхід перспективним напрямом для подальших досліджень і практичного застосування у задачах кліматичного прогнозування.

3.4 Висновок до розділу

У цьому розділі було здійснено всебічне дослідження взаємозв’язку між викидами парникових газів та змінами середньорічної температури з використанням методів Data Science. Проведено повний цикл аналізу даних: від збору та попередньої обробки до візуалізації ключових закономірностей, що

дозволило отримати уявлення про динаміку кліматичних змін у глобальному та регіональному масштабах.

Для прогнозування майбутніх змін температури були реалізовані та протестовані кілька моделей машинного навчання, серед яких випадковий ліс, градієнтний бустинг та рекурентні нейронні мережі LSTM. Усі моделі показали високу якість прогнозування та здатність точно описувати складні залежності у даних, особливо у довгостроковій перспективі.

Особливої уваги заслуговує побудова окремих моделей для кожної країни, що дало змогу дослідити специфіку впливу викидів на клімат у межах окремих держав. Це дозволило виявити, що вплив різних типів викидів на зміну температури суттєво варіюється залежно від країни, її рівня індустріалізації, структури економіки та політики щодо охорони довкілля. Такий підхід відкриває можливості для більш адресного та ефективного прийняття рішень на національному рівні щодо зменшення викидів і адаптації до змін клімату.

Отримані результати підтверджують ефективність застосування сучасних аналітичних методів і моделей для вивчення кліматичних процесів. Вони демонструють потенціал використання технологій машинного навчання як інструменту для прогнозування кліматичних змін і формування науково обґрунтованих рішень. У подальших дослідженнях доцільно зосередитися на розширенні обсягу даних, удосконаленні моделей за рахунок глибшої просторової деталізації.

РОЗДІЛ 4
РОЗГОРТАННЯ ІНФОРМАЦІННОЇ СИСТЕМИ ДЛЯ АНАЛІЗУ КЛІМАТИЧНИХ ЗМІН

4.1 Архітектура інформаційної системи та загальна концепція

У цьому пункті описується структура, логіка та взаємодія компонентів інформаційної системи, що реалізує веб-застосунок для візуалізації кліматичних даних, прогнозування температури та аналізу впливових факторів. Основна мета системи – забезпечити зручний інтерфейс користувача для перегляду аналітичних графіків та прогнозів на основі моделей машинного навчання.

Для наочного представлення архітектури та логіки роботи інформаційної системи було побудовано кілька UML-діаграм: діаграма компонентів; діаграма послідовності; діаграма прецедентів; діаграма діяльності.

Діаграма компонентів ілюструє основні вузли та взаємодію між ними у рамках архітектури інформаційної системи. Система поділена на кілька логічних частин: клієнтський вузол, серверний вузол, вузол даних і вузол машинного навчання (рис. 4.1).

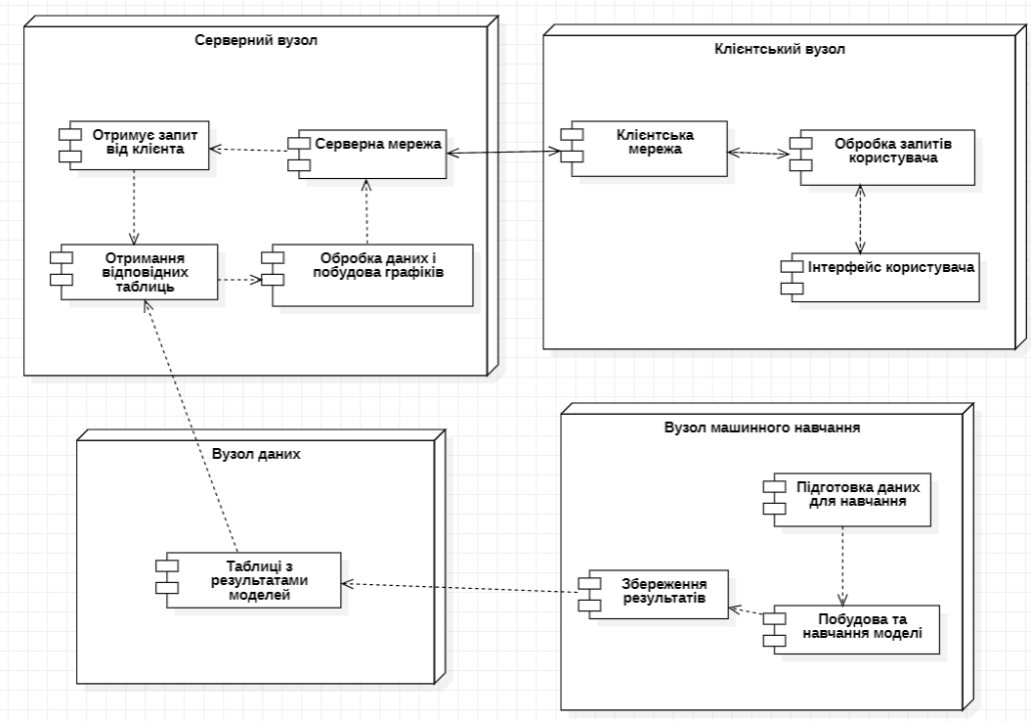

Рисунок 4.1 – діаграма компонентів

Клієнтський вузол: Цей вузол забезпечує взаємодію користувача з системою через веб-інтерфейс. Він містить такі компоненти:

- Клієнтська мережа: Забезпечує з'єднання між браузером користувача та сервером застосунку через мережу Інтернет;
- Інтерфейс користувача: Відображає графічні елементи веб-сайту, з якими взаємодіє користувач для перегляду статистичних даних та навігації;
- Обробка запитів користувача: Компонент на клієнтській стороні, який відповідає за обробку дій користувача (наприклад, кліків, вибору опцій) та формування відповідних запитів до сервера.

Клієнтський вузол не виконує обчислень — його головна роль полягає у відображенні інформації та формуванні запитів для серверної частини.

Серверний вузол: Цей вузол є центральним елементом обробки запитів та підготовки даних для відображення. Він включає:

- Серверна мережа: Отримує запити від клієнтської частини, обробляє їх, звертається до відповідних модулів і надсилає відповідь у вигляді побудованих графіків;
- Отримання запиту від клієнта: Компонент, який приймає запити, що надходять від клієнтських застосунків через серверну мережу;
- Отримання відповідних таблиць: Компонент, відповідальний за звернення до вузла даних та отримання необхідних таблиць з результатами моделей машинного навчання на основі запиту клієнта;
- Обробка даних і побудова графіків: Ключовий компонент, який обробляє отримані табличні дані, виконує необхідні агрегації та розрахунки, а також будує графічні представлення статистичної інформації.

Уся обробка даних і побудова результатів відбувається на сервері. Саме серверна мережа координує логіку обробки та повернення результатів користувачу.

Вузол даних: Цей вузол зберігає структуровану інформацію, яка використовується системою. Його функції зосереджені на збереженні та наданні доступу до даних, отриманих як із зовнішніх джерел, так і в результаті прогнозі. Він містить:

- Таблиці з результатами моделей: Сховище даних, де зберігаються структуровані дані, що є вихідними результатами роботи алгоритмів машинного навчання.

Вузол машинного навчання: Цей вузол відповідає за процеси, пов'язані з розробкою та навчанням моделей машинного навчання, результати яких використовуються для формування статистичних даних. Він включає:

- Підготовка даних для навчання: Компонент, який здійснює попередню обробку та підготовку набору даних, необхідного для навчання моделей.
- Побудова та навчання моделі: Процес розробки архітектури моделі машинного навчання та її навчання на підготовлених даних.
- Збереження результатів: Компонент, відповідальний за збереження результатів роботи моделей (наприклад, у вигляді CSV-файлів або записів у базі даних), які потім використовуються вузлом даних.

Діаграма також відображає взаємозв'язки між цими вузлами та їхніми компонентами за допомогою пунктирних стрілок, що символізують залежності та потік даних. Клієнтський вузол взаємодіє з серверним вузлом через мережу. Серверний вузол отримує дані з вузла даних та повертає оброблені дані у вигляді графіків до клієнтського вузла. Вузол машинного навчання готує дані та навчає моделі, а результати зберігаються та використовуються вузлом даних.

Ця архітектура забезпечує чітке розділення відповідальності між різними частинами системи, що сприяє її масштабованості, підтримуваності та гнучкості.

Наступною представлена діаграма послідовності (рис. 4.2), яка демонструє часову послідовність взаємодії між основними учасниками інформаційної системи під час реалізації сценарію побудови статистичних графіків на основі вибору користувачем певної країни. Учасниками цього процесу є:

- Користувач, який ініціює запит через інтерфейс;
- Графічний інтерфейс, що приймає дії користувача та формує запити;
- Сервер, який обробляє запит, отримує необхідні дані та генерує відповідь;
- Сховище даних, з якого витягуються необхідні аналітичні таблиці.

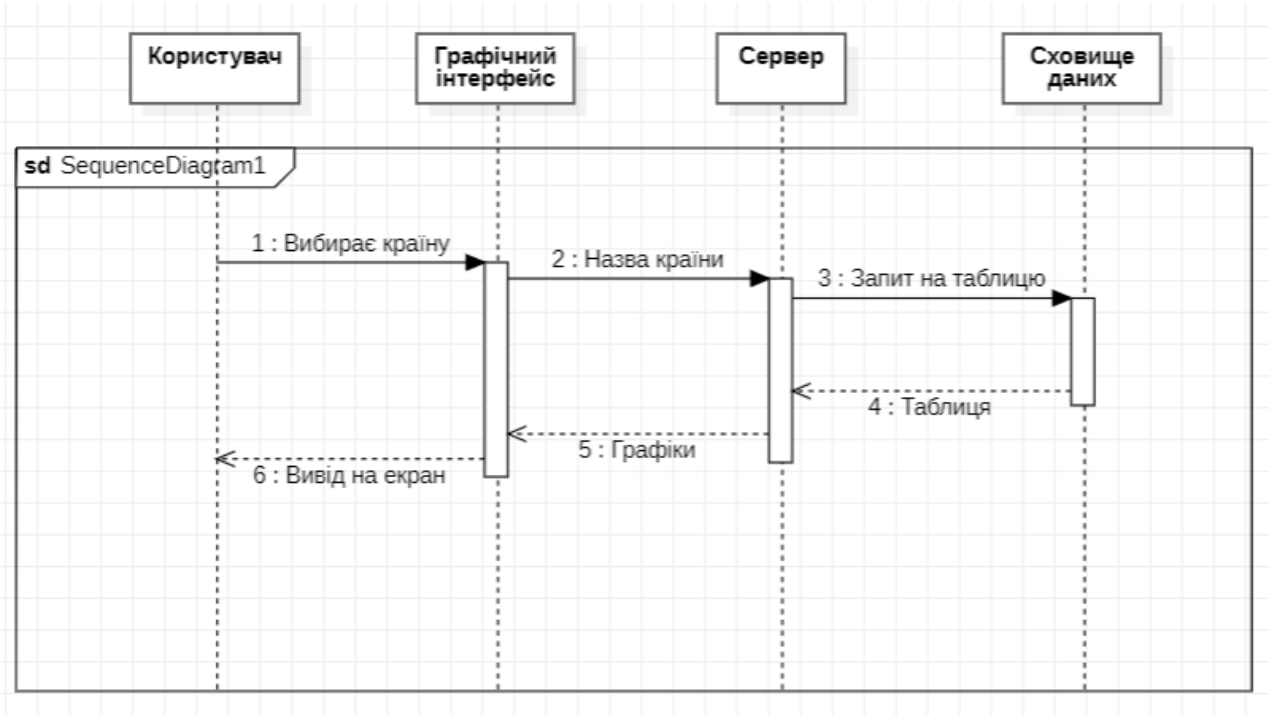

Рисунок 4.2 – діаграма послідовності

Діаграма демонструє наступну послідовність кроків:

1. Користувач ініціює процес, здійснюючи дію в графічному інтерфейсі (наприклад, клікає на країну на карті);
2. Графічний інтерфейс отримує інформацію про обрану країну та передає її на сервер;
3. Сервер, отримавши назву країни, формує запит до сховища даних для отримання таблиці, що містить статистичні дані, пов'язані з цією країною;
4. Сховище даних обробляє запит сервера, знаходить відповідну таблицю з даними та повертає її назад на сервер;

5. Сервер, отримавши табличні дані, виконує їхню обробку (наприклад, агрегацію, фільтрацію) та будує на їхній основі графічне представлення статистичної інформації;
6. Графічний інтерфейс отримує графіки від сервера та відображає їх користувачеві на екрані веб-браузера.

Ця діаграма дозволяє наочно простежити, як дані переміщуються між компонентами системи, а також як відбувається взаємодія між користувачем та сервером у рамках конкретного сценарію використання. Вона сприяє розумінню логіки роботи застосунку та може бути корисною при масштабуванні або модифікації системи.

Наступна діаграма, яка була побудована в рамках цього проекту, – це діаграма прецедентів, що ілюструє ключові сценарії взаємодії користувача з веб-сайтом для візуалізації статистичних даних (рис. 4.3). Ця діаграма дозволяє описати функціональні можливості системи з точки зору кінцевого користувача, акцентуючи увагу на основних діях, які може виконувати користувач.

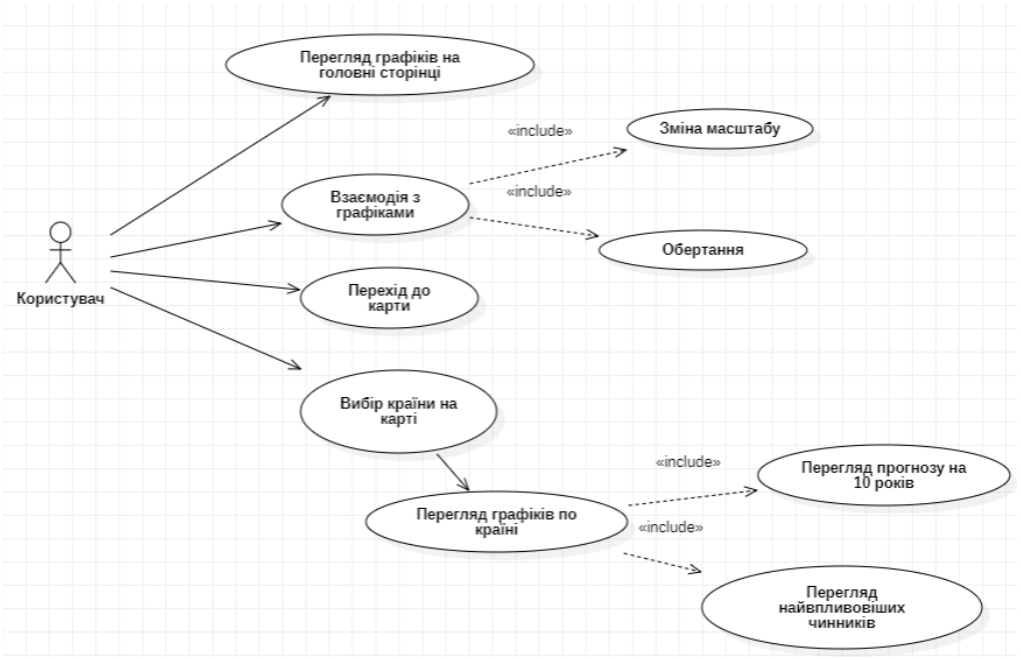

Рисунок 4.3 – діаграма прецедентів

Актор:

- Користувач: Представляє будь-яку особу, яка використовує веб-сайт для перегляду статистичних даних.

Прецеденти:

- Перегляд графіків на головній сторінці: Користувач може переглядати статистичні графіки, що відображаються на головній сторінці веб-сайту після його відкриття;
- Взаємодія з графіками: Цей прецедент узагальнює можливі дії користувача з графіками на головній сторінці. Він включає наступні прецеденти:
    - Зміна масштабу: Користувач може змінювати рівень збільшення або зменшення відображення графіків для детальнішого аналізу або загального огляду.
    - Обертання: Користувач може обертати графіки (якщо вони є тривимірними або мають таку функціональність) для кращого розуміння представлених даних.
- Перехід до карти: Користувач має можливість перейти на окрему сторінку веб-сайту, де відображається інтерактивна карта.
- Вибір країни на карті: На сторінці з картою користувач може вибрати конкретну країну, клікнувши на неї. Ця дія є необхідною для перегляду статистичних даних, специфічних для обраної країни.
- Перегляд графіків по країні: Після вибору країни на карті користувач може переглядати статистичні графіки, що відображають дані, пов'язані саме з цією країною. Цей прецедент включає наступні прецеденти:
    - Перегляд прогнозу на 10 років: Користувачеві відображається графік прогнозованих даних для обраної країни на період у 10 років.
    - Перегляд найвпливовіших чинників: Користувачеві відображається графік чинників, які мають найбільший вплив на статистичні показники та прогноз для обраної країни.

Діаграма показує, що актор “Користувач” може ініціювати всі основні прецеденти, що відображають функціональність веб-сайту. Відношення

<<include>> використовуються для декомпозиції складніших дій (таких як “Взаємодія з графіками” та “Перегляд графіків по країні”) на більш конкретні та складові частини, які завжди виконуються в контексті базового прецеденту.

Ця діаграма прецедентів надає чітке уявлення про те, як користувач може взаємодіяти з веб-сайтом для перегляду статистичних даних на головній сторінці та отримання детальної інформації, включаючи прогнози та впливові чинники, для обраних країн через інтерактивну карту.

Перейдемо до останньої діаграми – діаграми діяльності користувача на веб-сайті (рис. 4.4). Ця діаграма моделює логіку поведінки користувача при взаємодії з інтерфейсом системи, демонструючи послідовність можливих дій і варіанти навігації по розділах веб-сайту.

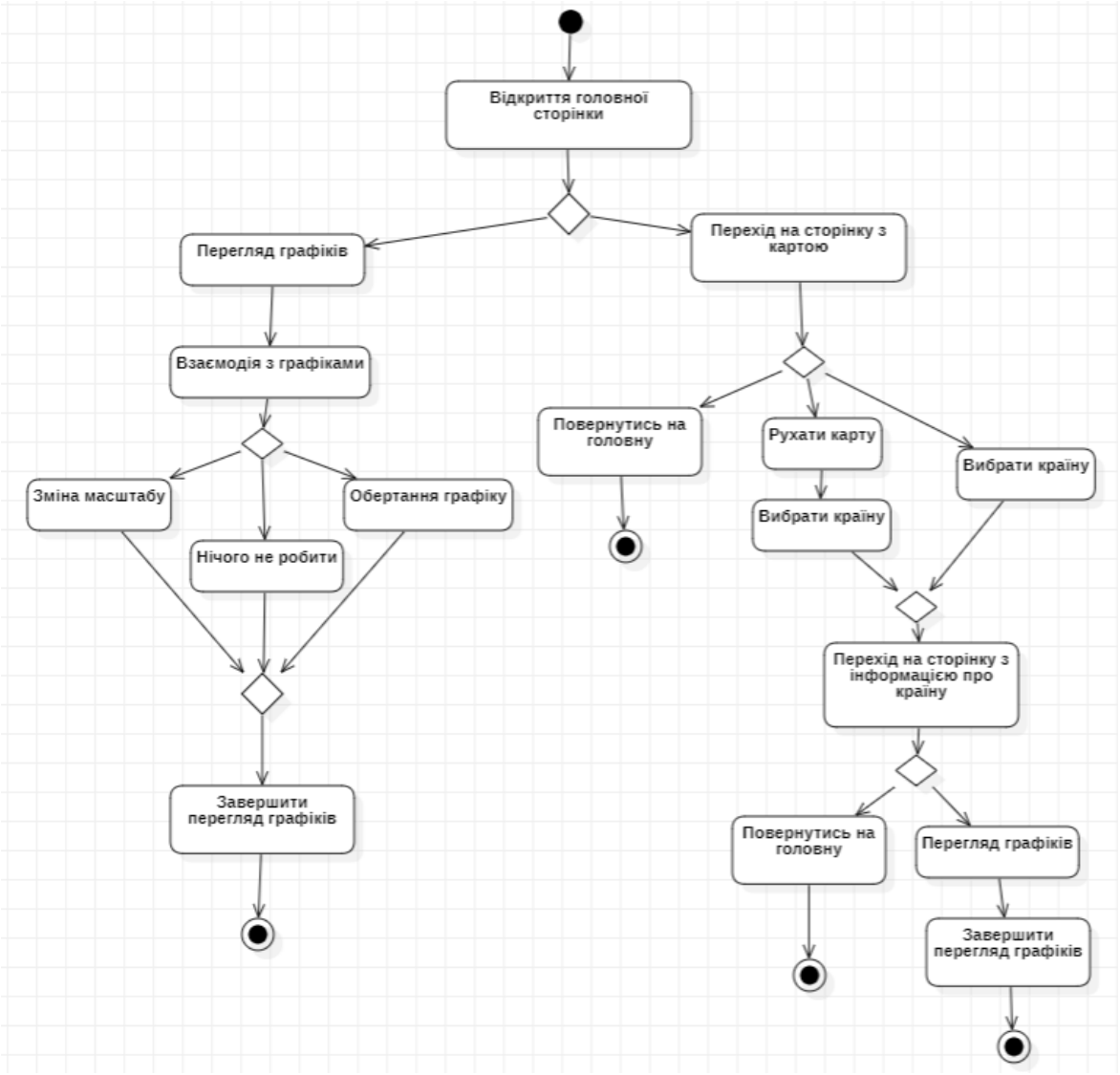

Рисунок 4.4 – діаграма діяльності

Процес взаємодії користувача з веб-сайтом починається з початкового вузла, що символізує вхід на сайт. Користувач відкриває головну сторінку, де йому надається доступ до загальних статистичних графіків. Це перший етап після завантаження веб-сайту.

Після відкриття головної сторінки користувач має кілька варіантів для подальших дій. Перший варіант – це перегляд графіків, які відображаються безпосередньо на головній сторінці. Користувач може взаємодіяти з графіками, наприклад, змінювати масштаб або обертати їх (якщо це можливо), що реалізовано через вкладене рішення в діаграмі діяльності. Після завершення взаємодії з графіками або відсутності подальших дій, цей сценарій завершується кінцевим вузлом діяльності, що позначає кінець цього шляху.

Інший варіант – перехід на сторінку з інтерактивною картою. Користувач може переміщати карту для перегляду різних регіонів або вибрати конкретну країну для перегляду детальнішої інформації. З цієї сторінки також є можливість повернутися на головну, що завершує цей сценарій.

Далі, після вибору країни на карті, користувач переходить на окрему сторінку, де відображаються статистичні графіки та прогнози для обраної країни. Після перегляду інформації користувач може повернутися на головну сторінку, що також завершує цей шлях.

Кожен із цих варіантів завершується кінцевим вузлом діяльності, що є символом завершення певного шляху взаємодії користувача. Це може статися після перегляду графіків на головній сторінці, повернення на головну сторінку з інтерактивної карти або після перегляду інформації про країну та повернення на головну сторінку.

Діаграма діяльності показує всі основні сценарії використання веб-сайту, відображаючи логіку переходів між розділами сайту: головною сторінкою, інтерактивною картою та сторінкою з даними про країни. Вона допомагає зрозуміти, як користувач взаємодіє з сайтом, що є важливим для подальшої розробки інтерфейсу та функціональної логіки сайту.

4.2 Структура та компоненти інформаційної системи

Інформаційна система побудована за модульним принципом, що забезпечує розділення функціональності на окремі компоненти.

Проект має наступну структуру директорій та файлів (рис. 4.5):

- static/: містить статичні ресурси:
  - css/main.css: файл стилів веб-інтерфейсу;
  - js/map.js: JavaScript-код для інтерактивної карти;
  - Набір .csv файлів:
    - actual_and_forecast_global_temperature.csv – містить актуальні та прогнозовані дані про зміну середньої температури у світі;
    - CO2.csv – містить дані про джерела викидів парникових газів по кожній країні;
    - feature_importances_by_country.csv – містить інформацію про найвпливовіші види викидів для кожної країни;
    - global_feature_importance.csv – містить інформацію про найвпливовіші види викидів на глобальному рівні;
    - Predicts.csv – зберігає актуальні та прогнозовані дані про зміну середньої температури по країнах;
    - total_CO2.csv – містить дані про загальні викиди CO_2 у світі.
- templates/: зберігає HTML-шаблони для генерації веб-сторінок:
  - base.html – базовий шаблон;
  - index.html – головна сторінка;
  - country_page.html – сторінка з інформацією по країнах;
  - map.html – сторінка з інтерактивною картою.
- venv/: віртуальне середовище з встановленими бібліотеками Python.
- main.py: основний файл серверної частини (backend) на Python.

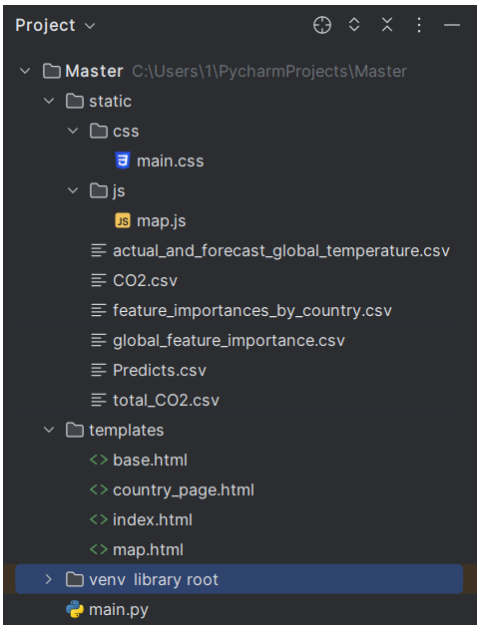

Рисунок 4.5 – структура застосунку

Система складається з трьох ключових компонентів, які забезпечують її функціонування. Клієнтська частина включає HTML-шаблони, CSS та JavaScript, що відповідають за формування інтерфейсу користувача, його візуальне оформлення та інтерактивність. Зокрема, JavaScript використовується для реалізації інтерактивної карти, яка реагує на дії користувача. Серверна частина реалізована у файлі main.py на мові Python. Вона обробляє HTTP-запити, що надходять від клієнта, зчитує та опрацьовує дані з CSV-файлів, формує відповідні структури для візуалізації, після чого генерує веб-сторінки на основі HTML-шаблонів. Сховище даних представлено у вигляді набору CSV-файлів, що зберігаються в директорії static/. Ці файли містять структуровані дані, які використовуються для побудови графіків, аналізу тенденцій, а також демонстрації актуальної статистики та прогнозів.

Взаємодія між компонентами системи відбувається наступним чином: веб-браузер користувача надсилає HTTP-запити до серверної частини (main.py). Серверна частина приймає ці запити, звертається до відповідних CSV-файлів для отримання необхідних даних, після чого обробляє їх і використовує HTML-шаблони для формування веб-сторінок, які містять графіки та іншу інформацію. Ці сторінки потім передаються в браузер користувача для відображення. Така

архітектура чітко розділяє логіку представлення та обробки даних, де файли .csv слугують основним сховищем даних для візуалізації та аналізу.

4.3 Інтерфейс користувача та навігація по системі

Інтерфейс користувача розроблений з урахуванням принципів простоти, наочності та зручності взаємодії з даними. Всі елементи системи згруповані у логічні розділи, що дозволяє користувачеві легко орієнтуватися в структурі платформи та швидко знаходити необхідну інформацію.

Система має три основні сторінки: головну, інтерактивну карту та інформаційну сторінку країни. Така структура забезпечує інтуїтивну навігацію між глобальним та локальним рівнями аналізу кліматичних даних.

Головна сторінка є центральним хабом системи, з якого починається взаємодія користувача з платформою (рис. 4.6). У верхній частині розташовано навігаційне меню, яке дозволяє швидко перемикатися між основними розділами системи.

Основний вміст сторінки складається з набору з 9 інтерактивних графіків, які демонструють різні аспекти екологічної ситуації, пов’язаної зі зміною клімату. Кожен графік надає окрему перспективу для аналізу:

1. Порівняння загальних викидів CO_2 з викидами від сільськогосподарського виробництва;
2. Зміна середньої температури по країнах;
3. Розподіл викидів за категоріями;
4. Зміна середньої температури та загальних викидів по роках;
5. Викиди CO_2 за континентами;
6. Топ країн за загальними викидами CO_2;
7. Топ країн за викидами CO_2 на особу;
8. Прогноз зміни середньої температури;
9. Найважливіші чинники для прогнозу.

Усі графіки є інтерактивними: користувач має можливість наводити курсор для перегляду значень, масштабувати окремі ділянки графіка,

фільтрувати дані та взаємодіяти з візуалізацією в режимі реального часу. Це значно підвищує зручність роботи з даними, забезпечуючи глибше розуміння екологічних процесів.

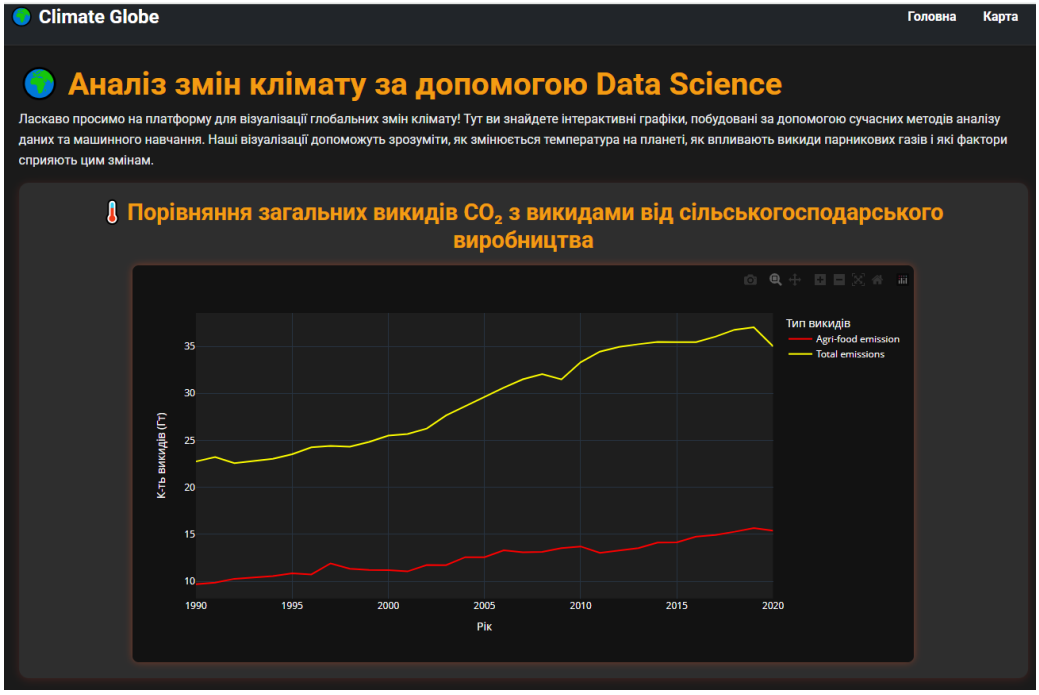

Рисунок 4.6 – головна сторінка

Сторінка з інтерактивною картою призначена для отримання даних на рівні країни для подальшого аналізу (рис. 4.7). Вона реалізована у вигляді повноекранної карти світу, з якою користувач може взаємодіяти безпосередньо.

На цій сторінці єдиним елементом інтерфейсу є сама карта, що дозволяє зосередитися на географічному контексті даних. Користувач може масштабувати карту, переміщати її, а також наводити курсор на країни, щоб побачити їхні назви. Така взаємодія забезпечує простий та зручний спосіб орієнтації у глобальному просторі.

При кліку на країну користувач автоматично переходить на інформаційну сторінку цієї країни, де вже подано детальні аналітичні графіки та прогнози, що стосуються саме обраної території. Таким чином, карта виконує роль інтуїтивного інструмента навігації між глобальними та локальними даними.

Дизайн реалізовано відповідно до сучасних стандартів: карта адаптивна, підтримує плавне масштабування, інтерактивність при наведенні, а також

миттєву реакцію на дії користувача, що забезпечує комфортну та ефективну роботу з нею на будь-яких пристроях.

Рисунок 4.7 – сторінка з картою

Після вибору країни на інтерактивній карті відкривається сторінка, присвячена детальному аналізу кліматичної ситуації саме в цій державі (рис. 4.8). Ця сторінка є ключовим елементом платформи для глибокого вивчення локальних екологічних змін і прогнозів.

На сторінці представлено:

- Назву країни, що чітко ідентифікує об’єкт аналізу та підтверджує вибір користувача;
- Графік прогнозу зміни середньої температури на 10 років, побудований на основі моделі машинного навчання. Цей графік демонструє очікувану динаміку температурних змін і допомагає оцінити потенційні наслідки кліматичних процесів на території країни;
- Графік найвпливовіших видів викидів CO_2 у сільському господарстві, що показує, які саме джерела є найбільш значущими з

точки зору впливу на зміну клімату. Це дозволяє визначити пріоритетні напрямки для зменшення шкідливих викидів.

Сторінка побудована з урахуванням принципів інтерактивності та аналітичної глибини. Користувач може взаємодіяти з графіками: наводити курсор для перегляду значень, масштабувати часову шкалу, фільтрувати дані тощо.

Рисунок 4.8 – сторінка з інформацією по країні

Користувач починає з головної сторінки, де відображаються різні графіки. Для переходу до інтерактивної карти світу користувач натискає відповідну кнопку навігації у верхньому правому куті. Після вибору певної країни на карті, наприклад України, користувач перенаправляється на сторінку з даними для обраної країни, де відображаються графік прогнозу температури на 10 років та стовпчаста діаграма топ-10 найвпливовіших факторів, що впливають на зміну клімату в цій країні. Користувач також може повернутись назад до карти, натиснувши кнопку “Назад”. Для повернення на головну сторінку користувач може скористатися відповідною кнопкою навігації у верхньому правому куті. Ці переходи між сторінками забезпечують зручну та логічну навігацію, що дозволяє користувачеві ефективно досліджувати кліматичні дані на різних рівнях (рис. 4.9).

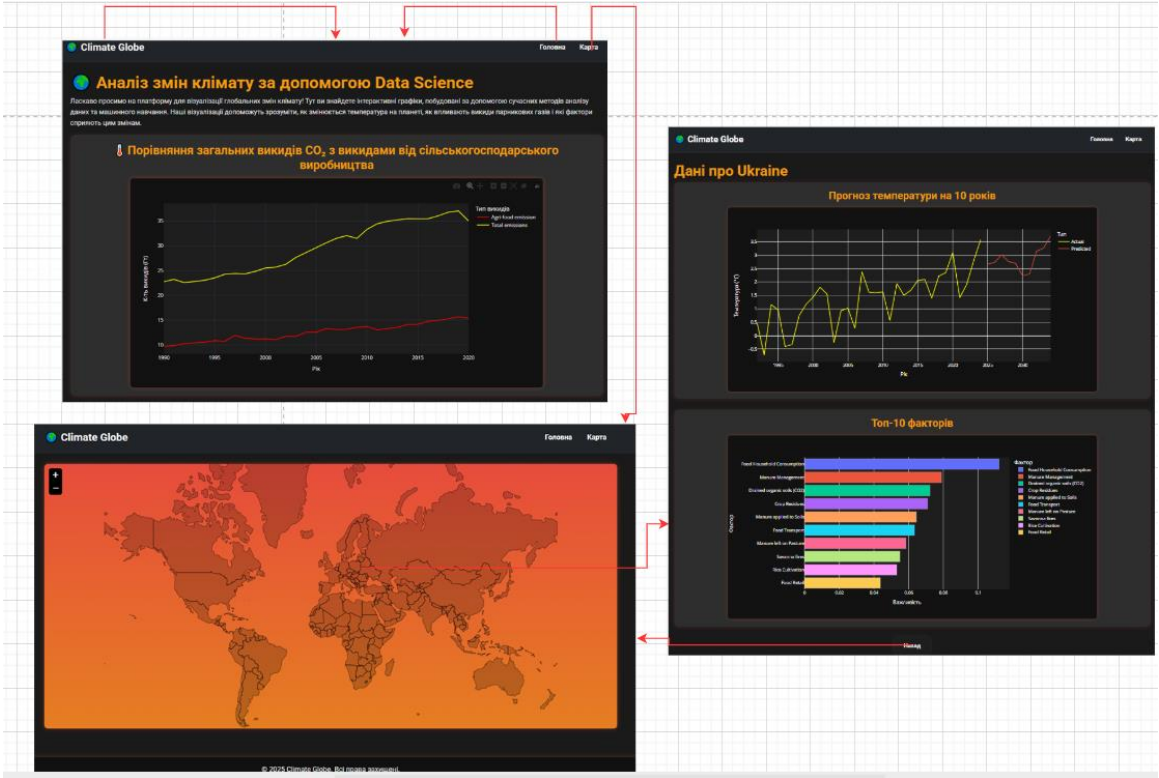

Рисунок 4.9 – переходи між сторінками

Загалом, структура інтерфейсу користувача та навігація по системі сприяють зручному та ефективному доступу до кліматичних даних. Завдяки чітко організованим переходам між головною сторінкою, інтерактивною картою та інформаційною сторінкою країни, користувач може швидко переміщатися між глобальними та локальними даними. Інтерактивні елементи, такі як графіки та карта, дозволяють здійснювати гнучку взаємодію з даними, забезпечуючи глибше розуміння екологічних процесів. Навігаційне меню та кнопки, розташовані в зручних місцях, дозволяють безперешкодно повертатися на попередні сторінки, що забезпечує зручність та ефективність роботи з платформою. Така логічна структура навігації створює позитивний користувацький досвід та дозволяє користувачеві ефективно аналізувати кліматичні дані на різних рівнях.

4.4 Перевірка функціональності

Для перевірки функціональності розробленої інформаційної системи було здійснено ручне тестування основних компонентів.

Першим етапом тестування була перевірка коректного відображення графіків на головній сторінці системи. Особливу увагу було приділено інтерактивним елементам – зокрема, відображенню інформації при наведенні курсору на точки графіка, а також можливості зміни масштабу відображення даних (рис. 4.10).

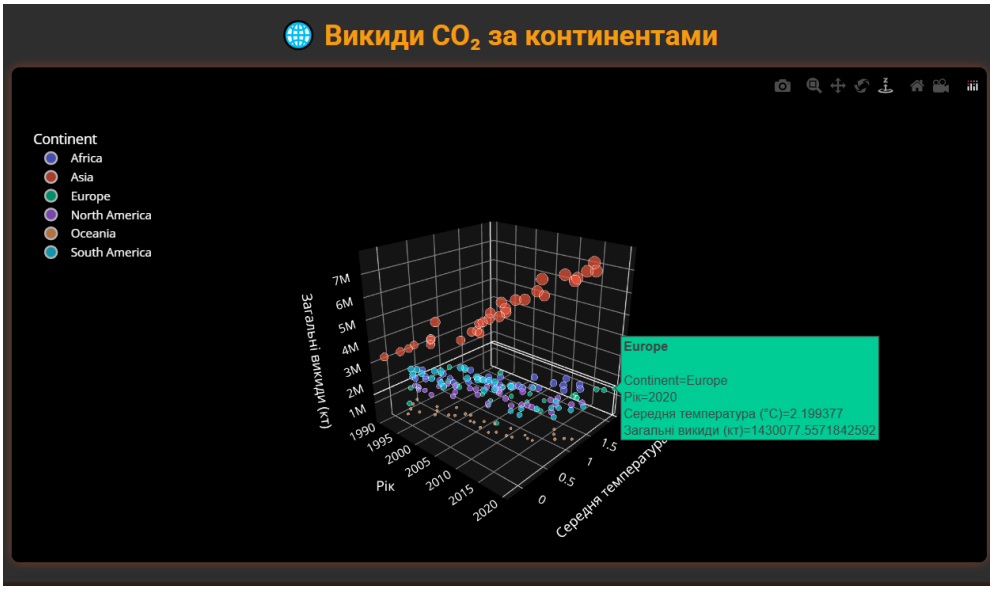

Рисунок 4.10 – перевірка інтерактивності графіків

Наступним кроком була перевірка коректного відображення інтерактивної карти світу. Було протестовано можливість переміщення карти та її взаємодії з користувачем. Зокрема, перевірено, чи при наведенні курсору на країну відображається її назва, що є важливим для ідентифікації об'єкта дослідження (рис. 4.11).

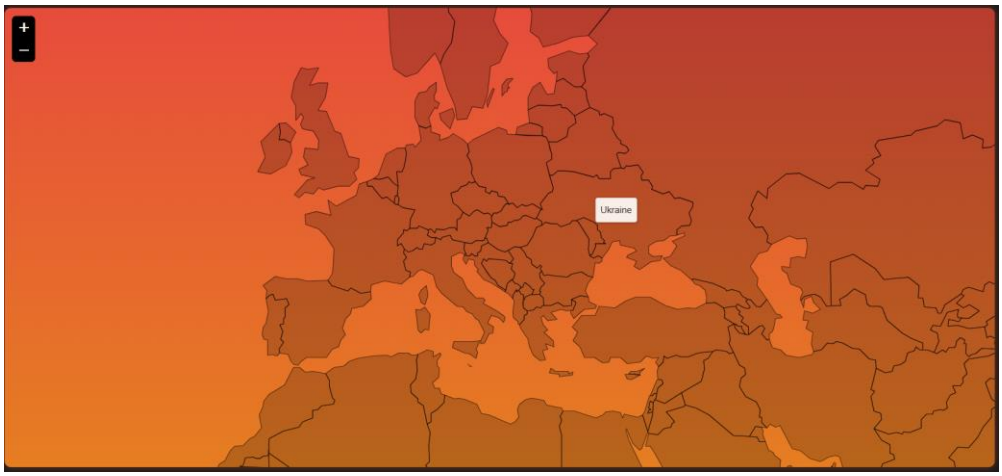

Рисунок 4.11 – перевірка роботи карти

Останнім етапом перевірки функціональності було протестовано коректність переходу на сторінку відповідної країни при натисканні на неї на карті. Було впевнено, що відображаються саме дані, які стосуються обраної країни, та що графіки будуються на основі коректно витягнутих даних. Також перевірено, чи зберігається інтерактивність графіків: можливість змінювати масштаб, переглядати значення при наведенні курсору тощо (рис. 4.12).

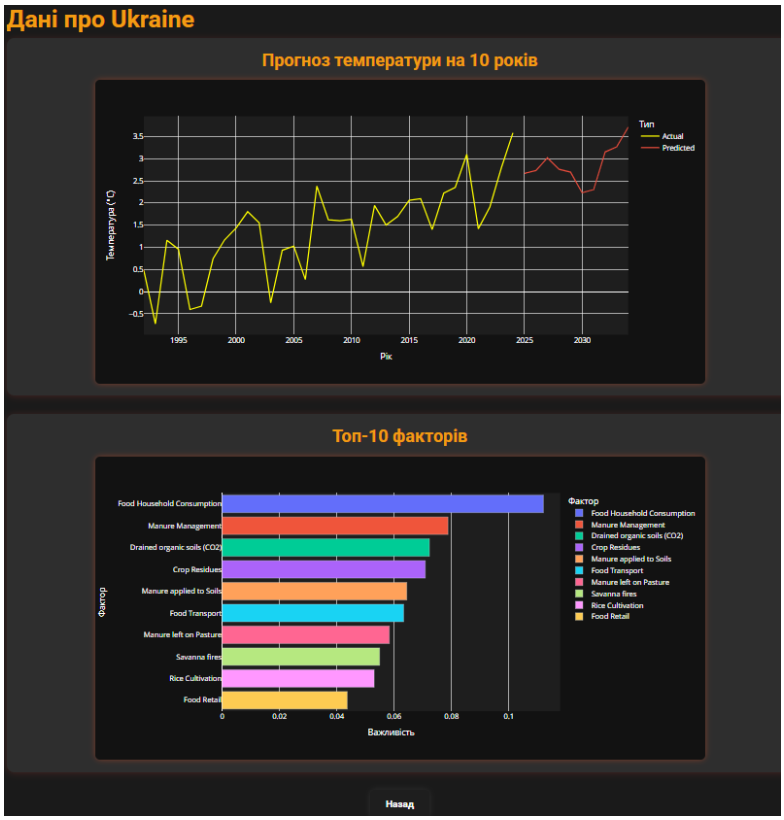

Рисунок 4.12 – перевірка правильності витягнутих даних

4.5 Перспективи застосування та розвиток системи

Розроблена інформаційна система демонструє значний потенціал для практичного використання у сфері дослідження змін клімату, аналізу екологічної ситуації в різних країнах світу та підтримки прийняття рішень, заснованих на даних. Її можна застосовувати як у науковій, так і в прикладній діяльності, адже вона забезпечує наочне подання великих обсягів даних, дозволяє здійснювати гнучкий аналіз інформації, а також створює зручні можливості для взаємодії з користувачем. Система може бути використана дослідниками, які займаються вивченням причин і наслідків глобального потепління, фахівцями з екології,

аналітиками, які працюють у сфері сталого розвитку, а також представниками державних установ, які формують політику у сфері охорони довкілля.

Окрім того, запропонована система має великий потенціал для використання в освітньому процесі, оскільки надає зручний інструмент для демонстрації впливу антропогенних факторів на кліматичні зміни. Використання реальних даних, інтерактивних візуалізацій та можливість аналізу змін за різні періоди часу робить систему ефективною як у викладанні дисциплін, пов’язаних з аналізом даних, так і в екологічних науках.

З точки зору подальшого розвитку, система може бути значно розширена як за функціональністю, так і за глибиною аналізу. Одним із перспективних напрямів є інтеграція додаткових джерел даних, таких як супутникові спостереження, дані про рівень забруднення повітря, зміну біорізноманіття, індекси якості води, землекористування та інші параметри, що мають вплив на кліматичні процеси. Це дозволить отримати більш комплексне уявлення про екологічну ситуацію та фактори, що сприяють глобальному потеплінню.

Також перспективним є удосконалення аналітичного блоку системи, зокрема впровадження більш складних моделей прогнозування на основі методів машинного навчання та глибинного навчання. У майбутньому можлива побудова персоналізованих аналітичних панелей для окремих користувачів, де кожен зможе вибрати параметри, які його цікавлять, зберігати налаштування, створювати власні звіти та порівнювати показники між країнами або регіонами.

Важливою перспективою є також підвищення зручності використання системи, зокрема шляхом локалізації інтерфейсу на різні мови, реалізації адаптивного дизайну для мобільних пристроїв та впровадження інструментів фільтрації та пошуку за країнами, роками, джерелами викидів тощо. З огляду на зростаючий обсяг відкритих даних, можливим є також розширення системи через підключення до зовнішніх API для автоматичного оновлення інформації.

Отже, розроблена система має великий потенціал для подальшого вдосконалення та практичного використання в різних галузях. Її відкритість, модульність і гнучкість дозволяють адаптувати її під нові задачі та забезпечити

довготривалу цінність як для дослідницьких, так і для прикладних цілей, спрямованих на розв’язання глобальних проблем зміни клімату.

4.6 Висновок до розділу

У даному розділі було представлено повний процес реалізації інформаційної системи для аналізу кліматичних змін із використанням сучасних технологій обробки, візуалізації та представлення даних. Детально описано вибір технологічного стеку, який забезпечує ефективну обробку великих обсягів даних, а також реалізацію основних функціональних компонентів системи, включно з інтерфейсом користувача та інтерактивними елементами. Значна увага була приділена створенню зручного, інтуїтивно зрозумілого інтерфейсу, що забезпечує швидкий доступ до інформації та аналітичних функцій.

Проведене ручне тестування функціональності підтвердило працездатність основних модулів системи. Зокрема, було перевірено коректне відображення інтерактивних графіків, що дозволяють аналізувати часові ряди та показники, роботу інтерактивної карти світу, а також коректність переходу на сторінки окремих країн із відображенням відповідних даних. Система демонструє стабільну взаємодію з користувачем, надаючи можливість масштабування візуалізацій, перегляду додаткової інформації при наведенні курсору та інші інтерактивні можливості.

Розроблене рішення надає потужний інструмент для дослідження впливу викидів парникових газів на зміну середньої температури в різних регіонах світу. Воно дозволяє виявляти закономірності, тренди та потенційні кореляції у просторово-часовому вимірі, що є важливим для глибшого розуміння кліматичних процесів. Запропонована система має значний потенціал для подальшого розвитку – зокрема, у напрямах підключення додаткових джерел екологічної інформації, реалізації автоматизованих засобів збору та оновлення даних, а також застосування методів штучного інтелекту для побудови моделей прогнозування майбутніх кліматичних змін.

Таким чином, розроблена система є ефективним і сучасним засобом для інтерактивного аналізу екологічних даних, що може бути з успіхом використаний у наукових дослідженнях, освітніх цілях, а також у прикладних розробках, пов’язаних із моніторингом навколишнього середовища. Її використання сприятиме підвищенню обізнаності про масштаби та динаміку кліматичних змін, а також формуванню обґрунтованих стратегій щодо їх пом’якшення на національному та глобальному рівнях.

ВИСНОВОК

У межах магістерської роботи було всебічно досліджено можливості застосування сучасних технологій Data Science для аналізу та прогнозування кліматичних змін, зокрема підвищення середньорічної температури під впливом викидів парникових газів. Актуальність обраної теми зумовлена глобальними викликами, пов’язаними зі зміною клімату, яка спричиняє суттєві екологічні, соціальні та економічні наслідки у всьому світі. У такому контексті особливої важливості набуває розробка ефективних моделей для аналізу та прогнозування кліматичних процесів, що можуть слугувати основою для прийняття рішень на державному та міжнародному рівнях.

У першому розділі було визначено мету, завдання, об’єкт та предмет дослідження, обґрунтовано актуальність теми й окреслено методологічну основу роботи. Огляд літератури дозволив сформувати уявлення про сучасний стан досліджень у галузі кліматичних змін, про найпоширеніші джерела даних, а також про методи статистичного аналізу та машинного навчання, які застосовуються у подібних дослідженнях. У результаті аналізу наукових джерел було виділено ключові підходи до моделювання кліматичних процесів, а також визначено доцільність використання часових рядів та нейронних мереж, зокрема LSTM, для побудови моделей прогнозування температури.

У другому розділі було здійснено ґрунтовний аналіз методів машинного навчання, які є найбільш релевантними для задачі прогнозування кліматичних змін, зокрема температури. Розглянуто ключові підходи до моделювання, включаючи алгоритми випадкового лісу (RF), градієнтного бустингу (XGBoost) та довготривалої короткочасної пам’яті (LSTM). Кожен з методів має свої особливості та переваги: алгоритми на основі дерев рішень забезпечують інтерпретованість та високу точність для табличних даних, тоді як рекурентні нейронні мережі демонструють високу ефективність при роботі з часовими рядами завдяки здатності враховувати залежності в динаміці даних. Було також систематизовано етапи побудови моделей, що дозволило сформувати чіткий алгоритм для реалізації прогнозної системи на практиці.

У третьому розділі було здійснено всебічне дослідження взаємозв’язку між викидами парникових газів та змінами середньорічної температури з використанням методів Data Science. Проведено повний цикл аналізу даних: від збору та попередньої обробки до візуалізації ключових закономірностей, що дозволило отримати уявлення про динаміку кліматичних змін у глобальному та регіональному масштабах. Зокрема, для прогнозування майбутніх змін температури були реалізовані та протестовані кілька моделей машинного навчання, серед яких випадковий ліс, градієнтний бустинг та рекурентні нейронні мережі LSTM. Отримані результати підтверджують ефективність застосування цих моделей для прогнозування змін температури в залежності від викидів парникових газів, що має важливе значення для формування політики зменшення викидів.

У четвертому розділі була розроблена інформаційна система, яка інтегрує технології обробки великих обсягів даних та візуалізації для аналізу кліматичних змін. Система дозволяє користувачам здійснювати інтерактивний моніторинг змін температури на основі викидів парникових газів, а також робити прогнози на майбутнє. Була реалізована ефективна інтерфейсна частина, що дає змогу користувачам отримувати не лише статистичні дані, а й візуалізувати їх у вигляді графіків та карт. Розробка такої системи забезпечує підтримку процесу прийняття рішень на основі аналізу кліматичних змін і має потенціал для впровадження в державні установи та міжнародні організації, які займаються питаннями охорони довкілля та клімату.

У заключенні підсумовано основні результати дослідження та зроблено висновки про ефективність застосування технологій Data Science для прогнозування кліматичних змін. Розроблені моделі та інформаційна система можуть бути основою для подальших досліджень та розробки стратегії зменшення викидів парникових газів. Отримані результати показують важливість інтеграції сучасних методів аналізу даних у практику моніторингу кліматичних змін, що має прямий вплив на формування екологічної політики та адаптацію до змін клімату.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. S.M. Bilan, A.O. Ukrainets “Artificial Neural Network for Classification of Images of Plant Flowers”. Information Technology and Implementation (IT&Is 2023): http://iti.fit.univ.kiev.ua/wp-content/uploads/%D0%97%D0%B1%D1%96%D1%80%D0%BA%D0%B0_ITI_2023.pdf
2. Ihor Miroshnychenko, Andrii Ukrainets “Development of Data Science Technologies in the Field of Climate Change Research”. Information Technology and Implementation (IT&Is 2024): http://iti.fit.univ.kiev.ua/wp-content/uploads/%D0%97%D0%B1%D1%96%D1%80%D0%BA%D0%B0-19_12_2024_ITI_2024-%D0%B5.pdf
3. Camilla Schramek, Sven Harmeling “G20 and climate change”: https://careclimatechange.org/wp-content/uploads/2017/06/G20-REPORT-.pdf
4. Indeed Editorial Team “7 Types of Statistical Analysis Techniques (And Process Steps)”: https://www.indeed.com/career-advice/career-development/types-of-statistical-analysis
5. Amit Yadav “Machine Learning for Climate Change Prediction”. Medium: https://medium.com/biased-algorithms/machine-learning-for-climate-change-prediction-bc4f89686a53
6. Liz Ticong “5 Top Predictive Analytics Techniques and Real-World Applications”. Datamation: https://www.datamation.com/big-data/predictive-analytics-techniques/
7. Jenny Cifuentes, Geovanny Marulanda, Antonio Bello, Javier Reneses “Air Temperature Forecasting Using Machine Learning Techniques”. Journal Energies volume 13 issue 16: https://www.mdpi.com/1996-1073/13/16/4215

8. Lee Chapman “Transport and climate change”. Journal of Transport Geography volume 15 issue 5 september 2007 pages 354-367: https://www.sciencedirect.com/science/article/abs/pii/S0966692306001207
9. Mikalai Filonchyk, Michael Peterson, Lifeng Zhang “Greenhouse gases emissions and global climate change: Examining the influence of CO2, CH4, and N2O”. Science of The Total Environment, Volume 935, 20 July 2024: https://www.researchgate.net/publication/380700139_Greenhouse_gases_emissions_and_global_climate_change_Examining_the_influence_of_CO2_CH4_and_N2O
10. Ahmad Hamdan, Ahmed Al-Salaymeh, Issah M. AlHamad, Samuel Ikemba, Daniel Raphael Ejike Ewim “Predicting future global temperature and greenhouse gas emissions via LSTM model”. Sustainable Energy Research Volume 10, article number 21, (2023): https://link.springer.com/article/10.1186/s40807-023-00092-x
11. Changjiang Xiao, Nengcheng Chen, Chuli Hu, Ke Wang, Jianya Gong, Zeqiang Chen “Short and mid-term sea surface temperature prediction using time-series satellite data and LSTM-AdaBoost combination approach”. Remote Sensing of Environment Volume 233, November 2019: https://www.sciencedirect.com/science/article/abs/pii/S0034425719303773?utm_source=chatgpt.com
12. Vahid Farhangmehr, Hanifeh Imanian, Abdolmajid Mohammadian, Juan Hiedra Cobo, Hamidreza Shirkhani, Pierre Payeur “A spatiotemporal CNN-LSTM deep learning model for predicting soil temperature in diverse large-scale regional climates”. Science of The Total Environment Volume 968, 10 March 2025: https://www.sciencedirect.com/science/article/abs/pii/S0048969725005364
13. Adil Lheureux “Weather forecast using LSTM networks” DigitalOcean: https://www.digitalocean.com/community/tutorials/weather-forecast-using-ltsm-networks

14. Qingchun Guo, Zhenfang He, Zhaosheng Wang “Monthly climate prediction using deep convolutional neural network and long short-term memory”. Scientific Reports volume 14, Article number: 17748 (2024): https://www.nature.com/articles/s41598-024-68906-6
15. Leo Breiman “Random Forests”. Machine Learning, Volume 45, pages 5–32, (2001): https://link.springer.com/article/10.1023/A:1010933404324
16. Tianqi Chen, Carlos Guestrin “XGBoost: A Scalable Tree Boosting System”: https://arxiv.org/pdf/1603.02754
17. Aayush Tyagi “Introduction to XGBoost Algorithm in Machine Learning”: https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/
18. Ottavio Calzone “An Intuitive Explanation of LSTM”. Medium: https://medium.com/@ottaviocalzone/an-intuitive-explanation-of-lstm-a035eb6ab42c
19. Christopher Olah “Understanding LSTM Networks”. Colah`s blog: https://colah.github.io/posts/2015-08-Understanding-LSTMs/
20. “Evaluation Metrics in Machine Learning”. Geeksforgeeks https://www.geeksforgeeks.org/metrics-for-machine-learning-model/
21. Oleksiy Voloshyn “Переваги і недоліки мови Python”. Hillel blog: https://blog.ithillel.ua/articles/perevagi-i-nedoliki-movi-python
22. “TensorFlow layers” TensorFlow: https://www.tensorflow.org/api_docs/python/tf/keras/layers
23. FAOSTAT “Emissions totals”: https://www.fao.org/faostat/en/#data/GT/metadata
24. FAOSTAT “Temperature change statistics”: https://openknowledge.fao.org/server/api/core/bitstreams/510225fe-18ca-40b8-9b00-1ca0bcb382d8/content
25. Hannah Ritchie, Max Roser “CO2emissions”. Our world in data: https://ourworldindata.org/co2-emissions