НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

Факультет прикладної математики

Кафедра прикладної математики

До захисту допущено:

Завідувач кафедри

_____ Олег ЧЕРТОВ

«___»_____2020 р.

Дипломна робота

на здобуття ступеня бакалавра

за освітньо-професійною програмою «Наука про дані та математичне моделювання»
спеціальності 113 «Прикладна математика»
на тему: «Програмна система моделювання впливу природних та антропогенних факторів на зміни кліматичних показників»

Виконала:
студентка IV курсу, групи КМ-62
Шумель Софія Олександрівна _____

Керівник:
асистент
Ковальчук-Хімюк Людмила Олександрівна _____

Консультант з нормоконтролю:
старший викладач
Мальчиков Володимир Вікторович _____

Рецензент:
старший викладач кафедри програмного забезпечення комп'ютерних систем, канд. техн. наук
Рибачок Наталія Антонівна _____

Засвідчую, що у цій дипломній роботі немає запозичень з праць інших авторів без відповідних посилань.
Студент (-ка) _____

Київ – 2020 року

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Факультет прикладної математики
Кафедра прикладної математики

Рівень вищої освіти – перший (бакалаврський)

Спеціальність – 113 «Прикладна математика»

Освітньо-професійна програма «Наука про дані та математичне моделювання»

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ Олег ЧЕРТОВ
«___»_____2020 р.

ЗАВДАННЯ

на дипломну роботу студенту

Шумель Софії Олександрівні

1. Тема роботи «Програмна система моделювання впливу природних та антропогенних факторів на зміни кліматичних показників», керівник роботи Ковальчук-Хімюк Людмила Олександрівна, асистент, затверджені наказом по університету від «25» травня 2020 р. № 1160-С

2. Термін подання студентом роботи «10» червня 2020 р.

3. Вихідні дані до роботи: розроблювана система повинна працювати з завантаженими даними з файлу

4. Зміст роботи: :виконати аналіз існуючих методів розв’язання задачі, вибрати метод для аналізу факторів впливу на клімат, спроектувати систему для аналізу даних, здійснити програмну реалізацію розробленої системи, провести тестування розробленої системи.

5. Перелік ілюстративного матеріалу.

6. Дата видачі завдання «3» лютого 2020 р.

Календарний план

№ з/п	Назва етапів виконання дипломної роботи	Термін виконання етапів роботи	Примітка
1	Огляд літератури за тематикою та збір даних	20.11.2015	
2	Огляд існуючих математичних методів та підходів для розв’язання поставленої задачі.	15.12.2015	
3	Проведення порівняльного аналізу математичних методів аналізу кліматичних даних.	24.12.2015	
4	Підготовка матеріалів першого розділу роботи.	01.02.2016	
5	Розроблення математичного забезпечення для прогнозування змін клімату, залежно від факторів впливу.	01.03.2016	
6	Підготовка матеріалів другого розділу роботи	15.03.2016	
7	Розробка архітектури системи для прогнозування змін клімату. Підготовка матеріалів третього розділу роботи	05.04.2016	
8	Розроблення програмного забезпечення для аналізу кліматичних даних.	15.04.2016	
9	Підготовка матеріалів четвертого розділу роботи.	03.05.2016	
10	Оформлення пояснювальної записки.	01.06.2016	

Студент Софія ШУМЕЛЬ

Керівник Людмила КОВАЛЬЧУК-ХІМЮК

АНОТАЦІЯ

Дипломну роботу виконано на 66 аркушах, вона містить 2 додатки та перелік посилань на використані джерела з 20 найменувань. У роботі наведено 25 рисунків та 7 таблиць.

Метою даної дипломної роботи є дослідження впливу природних та антропогенних факторів на глобальне потепління, яке представлено змінною середньої температури поверхні землі.

У роботі проведено аналіз існуючих рішень указаної задачі — регресійній аналіз, факторний аналіз, моделі на основі ланцюгів Маркова та нейромережеві моделі. Виконано їх порівняння з погляду ефективності алгоритмів та пристосованості методів до використання кліматичних даних. Для розв’язання задачі в роботі обрано регресійну модель.

На основі обраної моделі реалізовано три алгоритми машинного навчання: лінійна регресія, Random Forest та SVR для вибору найадекватнішої моделі, яка описує розглянуті дані, та виявлення факторів, що сприяють глобальному потеплінню.

Ключові слова: глобальне потепління, машинне навчання, регресія, прогнозування.

ABSTRACT

The thesis is presented on 66 pages. It contains 2 appendixes and and bibliography of 20 references. 25 figures and 7 tables are given in the thesis

The goal of this thesis is to research the influence of natural and anthropogenic factors on global warming, which is represented by the variable of average temperature.

The paper analyzes the existing solutions of this problem - regression analysis, factor analysis, models based on Markov chains and neural network models. Their comparison in terms of efficiency of algorithms and adaptability of methods to the use of climatic data is performed. A regression model was chosen to solve the problem.

Based on the selected model, three machine learning algorithms are implemented - linear regression, Random Forest and SVR to select the most appropriate model that describes the data considered, and to identify factors that contribute to global warming.

Key words: global warming, machine learning, regression, forecasting.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СКОРОЧЕНЬ ТА ТЕРМІНІВ

ANN (Artificial Neural Networks) — штучні нейронні мережі.

CH4 — органічна сполука метан.

CO2 — хімічна сполука діоксид вуглецю.

CSC (Climate Service Center) — центр кліматичних послуг Німеччини.

Gini Impurity — забруднення Джині.

IPCC (Intergovernmental Panel on Climate Change) — міжурядова група експертів з питань змін клімату.

MSE (Mean Squared Error) — середня квадратична помилка.

NOAA (National Oceanic and Atmospheric Administration) — національне управління океанічних і атмосферних досліджень.

помилка.

N2O — неорганічна сполука оксид азоту.

Random Forest — випадковий ліс.

RMSE (Root Mean Square Error) — корінь середньої квадратичної

SVR (Support Vector Regression) — підтримка векторної регресії.

VEI (Volcanic Explosivity Index) — індекс вулканічної експлозивності.

ЕОМ — електронно-обчислювальна машина.

МНК — метод найменших квадратів.

ПММ — прихована марковська модель.

ВСТУП

Протягом останніх десятиліть дослідження глобальної кліматичної системи все частіше стає об'єктом масштабних міжнародних програм, пов'язаних з оцінкою майбутніх змін в погодних і кліматичних подіях. Зміна клімату є одним з найсерйозніших викликів перед суспільством.

У 2015 році для пом'якшення зміни клімату в Паризькій угоді ООН встановили мету — обмежити збільшення середньої температури поверхні землі до 2 градусів Цельсія вище індустріальної температури. Оскільки концентрація вуглекислого газу в атмосфері знаходиться на найвищому рівні за 800 000 років, саме збільшенням викидів парникових газів пояснюють зростання середньої глобальної температури.

Очевидно, що на глобальну зміну температури впливають також зовнішні фактори такі, як зміна розмірів та розташування материків та океанів, зміна кількості сонячної радіації, вулканізм, зміна параметрів орбіти та вісі Землі. В більш загальному аспекті мінливості кліматичної системи, неможливо стверджувати, що кліматичні явища не є періодичними.

Мотивацією дослідження, аналізу та прогнозування кліматичних даних є знаходження відповіді на актуальне питання «Чи дійсно потепління з 1850 року може бути спричинені антропогенними факторами та, відповідно, суспільство може впливати на зміну клімату, дотримуючись певних домовленостей та умов?».

1 ПОСТАНОВКА ЗАДАЧІ

Метою даної дипломної роботи є проведення аналізу кліматичних даних та дослідження впливу різних факторів (природних та антропогенних) на зміну клімату. Для реалізації поставленої задачі необхідно розробити відповідне математичне та програмне забезпечення.

Для досягнення поставленої мети потрібно виконати наступні завдання:

а) проаналізувати існуючі математичні методи аналізу кліматичних даних;

б) обрати математичне забезпечення для моделювання кліматичної системи;

в) розробити програмне забезпечення, яке реалізує обраний метод розв'язання задачі.

Програмне забезпечення має задовольняти наступні вимоги:

а) приймати вхідні дані у вигляді таблиць з глобальними температурами та кількісними характеристиками факторів;

б) представляти результати дослідження у графічному вигляді.

2 ОГЛЯД МАТЕМАТИЧНИХ МЕТОДІВ РОЗВ'ЯЗАННЯ ЗАДАЧІ

Для аналізу кліматичних змін початковими даними слугують часові ряди, які містять значення кліматичних показників, наприклад вологості, кількості опадів, температури за певних проміжок часу. Очевидно, що чим довше часовий ряд, тим більше інформації можна отримати з даних. Через те, що спостереження за сонячної радіацією та промисловою активністю людини ведуться лише приблизно 30 років, тобто питання про наявність більш довготривалих часових інтервалів є актуальним, проблема впливу зміни сонячної активності та антропогенних парникових газів на клімат та атмосферу не має простого рішення [2].

Існує багато різних статистичних методів та процедур для аналізу кліматичних показників [3]. На досвіді робочої групи CSC було встановлено наступні категорії методів аналізу даних, які відіграють важливу роль у проектах пов'язаних з аналізом клімату:

а) загальні статистичні методи;

б) частотні розподіли;

в) аналіз часових рядів;

г) методи аналізу екстремальних значень;

д) просторово-часові методи;

е) інтерполяція зменшення масштабів.

Кожна методологія має свої вимоги для застосування та оцінки. Перед використанням будь-якої моделі необхідно з'ясувати виконання набору усіх передумов. Винятково можливе порушення деяких вимог, проте тільки в тому випадку, коли метод дає результати, які незначно відхиляються від результатів у разі дотримання всіх вимог.

2.1 Особливості кліматичних даних

Клімат — парадигма складної системи. Він має багато нелінійних змінних у широкому діапазоні просторово-часових масштабів [1].

Для того, щоб виявити особливості клімату, метеорологічну інформацію узагальнюють на більш довготривалі проміжки спостережень. Це є завданням кліматологічної обробки даних. Методи кліматологічної обробки даних визначають необхідну та достатню сукупність кліматичних показників, які допомагають знайти способи отримання достовірних кліматичних даних за масовими даними метеорологічних показників.

Кліматологічна обробка базується на представленні метеорологічних величин у вигляді випадкових величин. Отже, щоб описати закономірності кліматичних показників можна використовувати розділ математики — математичну статистику.

Метеорологічні величини не завжди є дійсно випадковими. Проте вони формуються ймовірнісними законами [3]. В якості основних кліматичних показників приймаються характеристики розподілу та структури часових метеорологічних рядів.

Оскільки записи даних для моделювання сучасного та майбутнього клімату дуже великі, можна сказати, що статистичні методи та процедури оцінки відіграють ключову роль при роботі з великою кількістю кліматичних даних.

Проте для вдалого застосування математичних методів аналізу кліматичних даних необхідно враховувати їх специфіку.

По-перше, метеорологічні елементи не є ймовірнісними величинами, які зазвичай досліджує статистика.

По-друге, значення різних кліматичних елементів зазвичай пов'язані між собою, а також у часі та просторі.

По-третє, метеорологічні ряди не є однорідними в силу відмінностей умов та методів спостережень. Переважна більшість статистичних методів направлена на однорідні дані.

Зазвичай за результатами метеорологічних спостережень може бути створена деяка статистична сукупність числових показників, яку прийнято називати рядом метеорологічних спостережень, або метеорологічним рядом.

Проте кліматичні дослідження не завжди базуються безпосередньо на сукупності первинних результатів спостережень. При досліджені зміни та коливань клімату достатньо використовувати сукупність середніх величин, які групуються за проміжками часу або простору. Таблиці таких усереднених даних є менш громіздкими та більш доступні.

Кліматичні характеристики розраховуються за обмеженими рядами спостережень. Тому потрібно пам'ятати, що, якщо початковий масив даних малий, то аномально високі чи низькі значення елементів, будуть мати великий вплив на статистичні характеристики. Відповідно, в великому ряду спостережень відхилення не будуть позначатися на результатах. Отже, рахуючи будь-яку статистичну характеристику необхідно завчасно визначити, якою має бути довжина початкового ряду для досягнення необхідної точності.

2.2 Загальні статистичні методи

Розглянемо основні методи статистичного аналізу та задачі, в яких вони використовуються при дослідженні кліматичних даних.

До таких методів можна віднести наступні:

а) дисперсійний аналіз;

б) кореляційний аналіз;

в) факторний аналіз;

г) кластерний аналіз;

д) дискримінантний аналіз.

Дисперсійний аналіз є статистичним методом аналізу результатів спостережень, одночасно залежних від різних факторів. Метою даного аналізу є вибір найбільш значущих факторів та оцінки їх впливу на досліджуваний процес. За допомогою методів дисперсійного аналізу визначається наявність фактору впливу на процес, який відображений сукупністю даних [6].

Для застосування класичних методів дисперсійного аналізу випадкові дані мають бути розподілені нормально та дисперсія експериментальних даних має бути однакова незалежно від умов.

Кореляційний аналіз застосовується для дослідження залежностей між випадковими величинами, а стохастична складова залежності між цими вибірками називається коефіцієнтом кореляції, який показує наскільки зв'язок між величинами близький до лінійної залежності. Слід зауважити, що кореляційний аналіз є частиною будь-якого статистичного дослідження.

Розглянуті вище методи дисперсійного та кореляційного аналізу дозволяють визначити наявність зв'язку між випадковими величинами та оцінити цей зв'язок. При існуванні кореляційного зв'язку між величинами, цю залежність можна задати функцією. Залежність середніх значень залежної величини та оцінки її статистичних властивостей є змістом регресійного аналізу [6]. Тобто, регресійний аналіз — статистичний метод аналізу даних, призначений для дослідження залежності однієї змінної від однієї або декількох незалежних величин.

Факторний аналіз — метод аналізу даних, призначений для виявлення незліченних величин, які визначають складні системи, — фактори. Методами факторного аналізу вирішують три основні групи проблем [4]:

а) пошук передбачуваних неявних закономірностей, що визначаються впливом зовнішніх або внутрішніх чинників на досліджуваний процес;

б) виявлення та вивчення статистичного зв'язку ознак з факторами або головними компонентами;

в) стиснення інформації за допомогою узагальнених факторів або головних компонент, кількість яких є меншою за кількість обраних параметрів.

У загальному випадку класифікація — це поділ сукупності об’єктів на однорідні в певному розумінні групи (класи) об’єктів до заздалегідь відомих класів. Розрізняють три групи завдань: дискримінацію, кластеризацію й групування. В свою чергу, кластерний аналіз — сукупність алгоритмів обробки даних, призначених для розподілення досліджуваних об’єктів на відносно однорідні групи (кластери), при чому кількість кластерів заздалегідь невідома. Дискримінантний аналіз — статистичний метод аналізу даних, призначений для розподілення об'єктів за заданими групами.

У таблиці 2.1 наведено, які основні задачі аналізу кліматичних даних можна розв’язати за допомогою зазначених методів.

Таблиця 2.1 — Основні задачі та методи статистичного аналізу

Основні задачі	Методи
Підготовка та відбір даних	Статистичні критерії, візуалізація
Визначення моделі та розподілу даних	Статистичні критерії, візуалізація
Дослідження зв'язків між різними факторами	Кореляційний аналіз, дисперсійний аналіз
Пошук прихований величин, що описують систему	Факторний аналіз
Розподілення об'єктів на групи	Кластерний аналіз, дискримінантний аналіз

2.2.1 Регресійний аналіз

Моделювання за допомогою регресійного аналізу полягає в створенні певної функціональної залежності, яка описує поведінку, досліджуваної випадкової змінної [5]. На практиці, це може бути ціна пшениці на світовому ринку або кількість смертей від раку легенів. Для питання, досліджуваного в цій роботі, такою змінною буде середня температура поверхні землі, або всі показники, які визначають клімат. У всіх випадках ця зміна називається залежною, або регресією, та позначається як y. Моделювання спрямоване на опис зміни середнього значення залежної змінної при зміні умов, при цьому вважається, що зміна залежної змінної не впливає на умови, що змінюються.

Змінні, які впливають на поведінку залежної змінної, називаються предикторами, або незалежними змінними, та позначаються як x.

Регресійні моделі також містять параметри — невідомі константи, які мають бути знайдені під час аналізу.

Математична складність моделі залежить від мети моделювання та повноти досліджуваного об'єкта. У випадках, коли прогнозування є основною метою, моделі належать до класу моделей, лінійних за параметрами. Тобто, параметри моделі є простими коефіцієнтами незалежних змінних або функціями незалежних змінних. Такі моделі називають лінійними регресійними моделями [4].

Моделі, які описують більш складні та реалістичні системи, зазвичай нелінійні в параметрах. Нелінійні моделі поділяють на дві категорії:

а) внутрішньо лінійні моделі, які можна лінеаризувати за допомогою відповідного перетворення на залежну змінну;

б) моделі, які неможливо трансформувати.

Класичний регресійний аналіз включає методи побудови математичних моделей досліджуваних систем, методи визначення параметрів цих моделей і перевірки їх адекватності [5].

Задачу побудови регресійної моделі можна сформулювати наступним чином [4]:

$$F(\alpha) = \sum_{i=1}^{n} (z_i(\alpha, X) - y_i)^2 \to min, \quad (2.1)$$

де $z_i(\alpha, X)$ — значення функції, що апроксимує залежність в $i-$ точці, y_i— значення залежної змінної, α — вектор шуканих параметрів. Функцію $z(\alpha, X)$ називають середньоквадратичною регресійною моделлю.

Апроксимуюча функція у випадку однієї незалежної змінної, в свою чергу, може бути представлена у вигляді елементарних функцій: поліном, обернений поліном, експоненціальна або показникова функція, степенева, лінійно-логарифмічна функція, тригонометричний ряд Фур'є.

За наявністю декількох незалежних змінних (моделі множинної регресії) використовують функції, лінійні як за параметрами, так і за незалежними змінними, а також поліноміальні моделі, що є лінійними за параметрами, але нелінійними за незалежними змінними [4].

2.2.1.1 Лінійні багатофакторні моделі

Загальний вигляд багатофакторної лінійної моделі можна записати в наступному вигляді:

$$Y = \alpha_0 + \sum_{i=1}^{p} \alpha x_j + \varepsilon = X\alpha + \varepsilon, \quad (2.2)$$

де Y— вектор-стовпчик, X — матриця значень p незалежних змінних, α— вектор-стовпчик невідомих параметрів моделі, ε — вектор-стовпчик похибок моделі.

Побудова поліномів, які утворюють систему, відбувається на основі метода ймовірнісної апроксимації. Метод полягає в тому, що визначення коефіцієнтів апроксимуючих поліномів відбувається за умови мінімума дисперсії помилки апроксимації $D_\varepsilon = \underline{\varepsilon^2} - (\varepsilon)^2$ функції $y_i(x)$, де $\varepsilon_i(x) = y_i(x) - \widehat{y_i}(x)$. Враховуючи особливості величин y_i і компонентів вектора x, а також незалежність факторів в початковій виборці від вигляду поліному, можна побудувати систему функцій. Ця система дозволяє отримати оцінки шуканих коефіцієнтів системи рівнянь.

Найбільш розповсюдженим та теоретично обґрунтованим методом для побудови лінійних та нелінійних по факторами рівнянь регресії є метод найменших квадратів [7].

Якщо в моделі кожен вхідний фактор заданий одновимірним часовим рядом, апроксимуюча функція може бути набувати наступний вигляд:

$$Y\left(\underline{X}(t), t\right) = b_0 + b = b_1 f_1\left(X_1(t)\right) + \ldots + b_j f_j\left(X_j(t)\right) + \cdots + b_m t + \theta_1, \quad (2.3)$$

де $\{b_j\}$ — шукані коефіцієнти регресії, Y — значення показника, що моделюється, $f_i(X_i)$ — відомі базисні однофакторні неперервні елементарні функції (лінійна, поліноміальна, степенева, логарифмічна, експоненційна та інші),θ_1— випадкова помилка апроксимації.

Перевагою цього методу є те, що даний метод не накладає обмежень на нелінійність зв'язків між залежною та незалежними змінними. Єдиним обмеженням є неперервність базисних функцій $f_i(X_i)$.

Зазвичай шляхом заміни змінних вигляду:

$$z_i = f_i(X_i) \quad (2.4)$$

можна звести нелінійне рівняння регресії до лінійного:

$$\underline{Y}^* = \sum_{j=0}^{n} b_j z_j + \theta_2, \quad (2.5)$$

де θ_2— випадковий помилка, при чому $\theta_1 \neq \theta_2$.

Метод найменший квадратів для лінійного багатофакторного рівняння регресії дозволяє знайти ефективні оцінки шуканих коефіцієнтів $\{b_j\}, j = \underline{0, \ldots n}$.

2.2.1.2 Нечітка регресійна багатофакторна модель

В цьому методі замість мінімізації квадратичного функціонала розв’язується задача мінімізації ширини нечіткого інтервалу прогнозованих значень Y. Метод нечіткої регресії зводиться до розв'язання задачі лінійного програмування [7].

Теорія нечітких множин є частиною математики, яка орієнтована на обробку невизначених об'єктів, які не деформуються під точні поняття. За допомогою цієї теорії можна розробити довготривалі прогнози явища в умовах невизначеності та швидкої зміни зовнішнього середовища. Нечітка логіка дозволяє розширити традиційні методи прогнозування. Зокрема, на основі теорії нечітких множин можна побудувати нечіткі регресійні моделі, використовувані для прогнозування досліджуваних показників.

Задача аналізу — знайти таке рівняння зв'язку, яке найбільш точно описує зв'язок між значеннями факторів і результативними показниками з умовою того, що дані нечіткі, розмиті, а коефіцієнтами рівняння регресії є нечіткі числа.

Розглянемо суть алгоритму побудови нечіткої регресії.

Нехай, маємо ряд факторів $X_1, X_2, \ldots, X_n$, які визначають показник Y, а також вибірку даних з m спостережень значень Y, які можуть бути чіткими або нечіткими. Необхідно визначити функцію:

$$Y = \alpha_0 + \alpha_1 x_1 + \ldots + \alpha_n x_n; \underline{X} = (x_1, \ldots, x_n), \qquad (2.6)$$

яка найбільш точно описує значення результативного показника Y. Параметри моделі $\alpha_0, \alpha_1, .., \alpha_n$— нечіткі симетричні довірчі трійки чисел, які будуть мати вигляд:

$$\alpha_i = (\alpha_i - b_i, \alpha_i, \alpha_i + b_i), \qquad (2.7)$$

де α_i — найбільш ймовірне значення коефіцієнта, а величина b_i — ширина розмитості цього коефіцієнта.

Регресійна функція $Y(\underline{X})$ буде описана у вигляді трикутного симетричного нечіткого числа, яке містить реальне значення результата. Схематичне представлення цього числа зображено на рисунку 2.1.

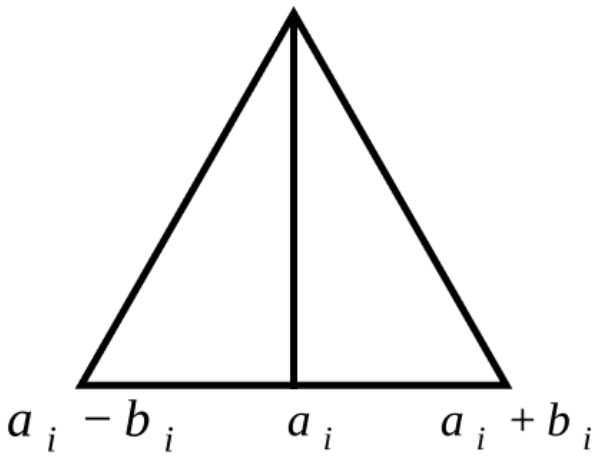


Рисунок 2.1 — Представлення нечіткого числа a_i

Формалізувати модель можна наступним чином:

$$\sum^m \sum^n ((\alpha_i - b_i)x_{ij}) + (\alpha_0 - b_0) \le Y_j, \forall j = \underline{1,..m} \qquad (2.8)$$

$$Y_j \le \sum^m \sum^n ((\alpha_i + b_i)x_i) + (\alpha_0 + b_0) \ge Y_j, \forall j = \underline{1,..m}, b_j \ge 0 \qquad (2.9)$$

Необхідно знайти такі значення α_iта b_i, щоб ширина отриманого “коридора”, який описує реальні значення досліджуваного показника, були мінімальною за сумою всіх змін:

$$F = \sum^m |\sum^n ((\alpha_i + b_i)x_{ij}) + (\alpha_0 + b_0) - \sum^n ((\alpha_i - b_i)x_{ij})| \to min \qquad (2.10)$$

Задане рівняння містить 3 складові:

а) Функцію Y_1, яка містить мінімальні коефіцієнти, тобто значення функції, які знаходяться вище будь-якого із значень апроксимуючого показника Y;

б) Функцію Y_2, яка визначає середину можливих значень досліджуваного показника Y;

в) Функцію Y_3, яка містить максимальні коефіцієнти, тобто значення функції, які знаходяться нижче будь-якого значення апроксимуючого показника Y.

Нечітку регресію часто називають “нечітким коридором” через те, що всі значення досліджуваного параметру знаходяться між функціями Y_1та Y_3.

Саме ці складові відрізняють нечітку регресію від класичних методів регресійного аналізу.

Основним критерієм якості побудови функції регресії є абсолютна та відносна оцінка рівня його невизначеності. У випадку трикутної нечіткої регресії, в якості абсолютної оцінки може виступати ширина “коридору”, а в якості відносної оцінки — відношення інтервала зміни нечіткого числа до його середнього значення.

2.2.2. Факторний аналіз

Факторний аналіз відіграє важливу роль у більшості досліджень. Він робить можливим узагальнення великої кількості матеріалу до декількох незалежних та простих факторів. Методами факторного аналізу можна підтвердити деяку наукову гіпотезу або сформулювати деяку нову гіпотезу на основі великого об'єму спостережень за певними впливовими компонентами [8].

Факторний аналіз є методом, який ґрунтується на питанні наскільки сильно досліджуванні змінні корелюють між собою. Це означає, що вони або взаємно визнають одна одну, або зв'язок між ними обумовлено іншою величиною, яку

безпосередньо не можна виміряти. Отже, задачею факторного аналізу є визначення величини, яка передбачає спостережувані зв'язки. Ця величина називається фактором.

Передумовами факторного аналізу є наявність взаємозв'язку між декількома змінними. В якості кількісного показника цього зв'язку між вибірками використовується коефіцієнт кореляції та кореляційна матриця.

При проведенні факторного аналізу усі розрахунки мають послідовний характер. На рисунку 2.2 показана процедура виконання цих обчислювальних дій. Чотири вертикальні стрілки відповідають чотирьом основним проблемам факторного аналізу.

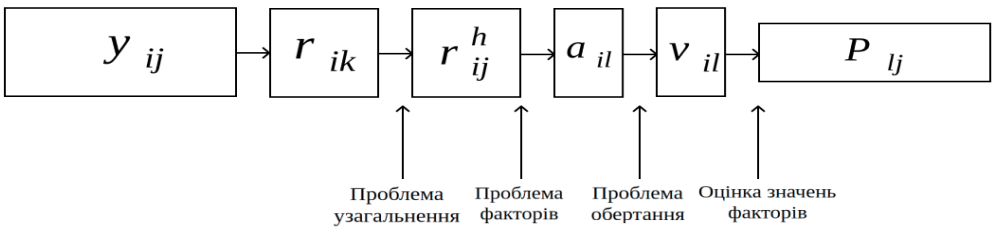


Рисунок 2.2 — Процедура виконання факторного аналізу

Маємо матрицю початкових даних $Y = \{y_{ij}\}$, за допомогою якої обчислюється кореляційна матриця $R = \{r_{ik}\}$. Елементи головної діагоналі цієї кореляційної матриці є оцінками узагальнення $\{r_{ik}^h\}$, які встановлюють задачу узагальнення, головним завданням якої є визначення оцінок $\hat{h}_i^2$. Наступною проблемою є проблема факторів: необхідно з редуцированої матриці $R_h = \{r_{ij}^h\}$ за допомогою певних способів отримати фактори, в результаті чого можна обчислити матрицю $A = \{\alpha_{il}\}$. Оскільки розв'язок є множиною таких матриць A, постає проблема обертання. Після вирішення цього етапу отримується матриця $V = \{v_{il}\}$. Четверта, остання проблема факторного аналізу — оцінка значень факторів. В більшості випадків окремі етапи аналізу не застосовуються. На практиці через велику кількість обчислень деякі проблеми розв’язуються не повністю [8].

Одним з основним методів факторного аналізу є метод головних компонент, основна модель якого може бути записана наступним чином:

$$Z = AP\ , \qquad (2.11)$$

де Z— матриця стандартизованих початкових даних, A— факторне відображення, P— матриця значень факторів.

Стандартизовані початкові дані отримуються з вхідних даних за формулою:

$$z_{ij} = \frac{y_{ij} - \underline{y_{ij}}}{S_i},\ i = 1,2,\dots,m,\ j = 1,2,\dots,n, \qquad (2.12)$$

де y_{ij} — елемент матриці початкових даних, $\underline{y_{ij}}$ — середнє значення, S_i — стандартне відхилення.

Щоб обчислити кореляційну матрицю R використовують наступне співвідношення:

$$R\ = \frac{1}{n-1} ZZ'. \qquad (2.13)$$

На головній діагоналі матриці R стоять узагальнені значення, які дорівнюють одиниці.

Для знаходження матриць А та Р використовують основну теорему факторного аналізу:

$$R\ =\ ACA', \qquad (2.14)$$

де C- кореляційна матриця.

Модель класичного факторного аналізу містить ряд загальних факторів та по одному характерному фактору на кожну змінну.

2.3 Моделі на основі ланцюгів Маркова

В загальному випадку марковські процеси використовують для навчання та розпізнавання послідовних даних, до яких можна віднести коливання температури, біологічні зміни, економічні показники та інші [9]. В моделях Маркова кожне спостереження послідовності даних залежить від попередніх елементів послідовності.

Розглянемо систему, де існує безліч станів $S = \{1, 2, \ldots, N\}$. У кожний дискретний момент t система здійснює перехід до одного зі станів набору, відповідно до матриці переходів ймовірностей P. У багатьох випадках прогнозування наступного стану залежить лише від поточного, тобто ймовірність переходу не залежить від усієї історії процесу. Така модель називається марковським процесом першого порядку.

Такі моделі мають обмежену потужність в багатьох сферах, тому дослідники вдосконалили дану методологію та розширили її на модель з більшою потужністю представлення — приховану марковську модель. У даній моделі кількість станів та ймовірності переходу є невідомими, замість цього кожна модель має імовірнісну функцію, пов'язану з моментом t:

$$b_j(\theta_t) = P(\theta_t | X_t = j). \qquad (2.15)$$

Загальна структура моделювання ПММ — це метод навчання без нагляду. Він може обробляти різні довжини послідовностей вводу, тобто не потрібно вказувати параметри для навчання. Недоліком даного способу є необхідність збалансувати вимоги чутливості та точності.

У випадку аналізу та прогнозування часових рядів, коли можна вважати, що дані породжуються деяким основоположним стохастичних процесом. Прихована модель Маркова має більшу силу ніж звичайний ланцюг Маркова: ймовірності

переходу, а також функція щільності ймовірності генерації спостережень регулюються [9].

2.4 Нейромережеві моделі

В задачах моделювання за допомогою нейронних мереж можна відокремити наступні підзадачі: визначення ознак, навчання мережі, побудова моделі, що реалізує розв'язок задачі.

Штучні нейронні мережі — це сукупність моделей біологічних нейронних мереж, які представляють собою мережу елементів — штучних нейронів, пов'язаних між собою синаптичними зв'язками. Мережа обробляє вхідну інформацію та формує сукупність вихідних сигналів, шляхом зміни свого внутрішнього стану [10].

Для розв'язку задачі необхідно знайти таку нейрону мережу, яка найкращим способом побудує відображення $F: X \to Y$, яке узагальнює вхідні дані. Пошук такої мережі реалізується за допомогою одного або декількох алгоритмів навчання.

На рисунку 2.3 показана модель нейрона, яка лежить в основі ANN.

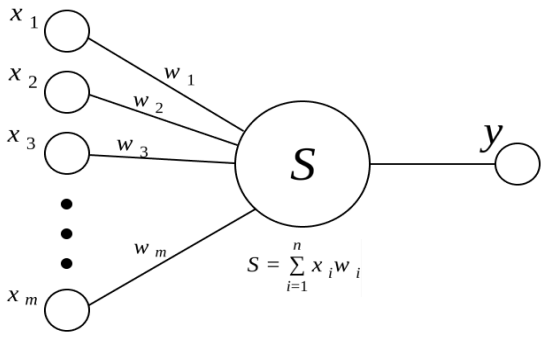


Рисунок номер 2.3 — Схема штучного нейрону

В моделі виокремлюють три основні елементи: набір синапсів або зв'язків, суматор та функцію активації. Синапси характеризуються своєю вагою. Сигнал x_j на

вході синапсу jдо k-го нейрону, помножується на вагу w_{kj}. Суматор додає вхідні сигнали, цю операцію можна описати як лінійну комбінацію. Функція активація обмежує амплітуду вихідного сигналу з нейрону. Зазвичай нормалізований діапазон амплітуд виходу нейрону лежить в інтервалі [0,1] або [-1, 1].

Математично роботу нейрону можна сформулювати наступними рівняннями:

$$u_k = \sum_{j=1}^{m} \quad w_{kj} x_j, \tag{2.16}$$

$$y_k = \varphi(u_k + b_k), \tag{2.17}$$

де $x_1, \ldots, x_m$ — вхідні сигнали,$w_{k1}, \ldots, w_{km}$ — синаптичні ваги k-го нейрону, u_k — лінійна комбінація вхідних сигналів, b_k— поріг, $\varphi(\cdot)$— функція активації, y_k — вихідний сигнал.

ANN — це набір зв'язаних між собою нейронів, елементом якої є одношаровий перцептор.

Побудова нейромережі реалізується в два етапи:

а) вибір типу архітектури мережі;

б) навчання мережі.

Проектуючи модель ANN, можна скористатися вже дослідженими архітектурами. До таких можна віднести багатошаровий перцептрон, нейромережу з загальною регресією, мережу Кохонена та інші [11] .

На другому етапі необхідно навчити мережу. На практиці кількість вагів в ANN становить кілька десятків тисяч, тому навчання є складним процесом. Для багатьох архітектур розроблені спеціальні алгоритми навчання, які дозволяють знайти ваги певним чином [11].

2.5 Порівняльна характеристика методів

Базуючись на огляді існуючих методів розв'язку, було створено узагальнюючу таблицю 2.2, яка містить короткі систематизовані записи про переваги та недоліки зазначених методів.

Таблиця 2.2 — Порівняння методів

Метод	Переваги	Недоліки
Регресійний аналіз	Простота та гнучкість, можливість розширення функціоналу поза області відомих значень, доступність до проміжних обчислень.	Велика кількість вимог до вхідних даних, проектування лише лінійних залежностей.
Факторний аналіз	Узагальнення вхідної системи з можливістю дослідження спрощеної моделі.	Відсутність однозначного математичного розв’язку, висока складність обчислень.
Марковські процеси	Легко моделюється та проектується.	Використання лише дискретних даних та неможливість моделювання вибірок з довгою пам'яттю.
Нейронні мережі	Відсутність обмеження на лінійність системи, здатність швидкої адаптації до навколишнього середовища.	Відсутність доступу до проміжних обчислень, складність проектування.

Слід зауважити, що для жодного методу не представлені показники точності методу, оскільки вони залежать не тільки від моделі, а також від досвіду дослідника, обраних даних та інших факторів. Точність моделювання буде оцінюватися при розв'язання поставленої задачі, в рамках даної роботи.

2.5.1 Переваги та недоліки регресійного аналізу

Однією з найбільш важливих переваг регресійної моделі є розширення функціоналу поза межі області відомих значень вхідних параметрів. Також дана методологія є простою та гнучкою, що забезпечує, одноманітність аналізу та проектування. Регресійні моделі можуть надати результати значно швидше, ніж інші моделі, проте визначення параметрів нелінійних регресійних моделей є досить ресурсномістким.

На відміну від методу ANN моделі регресійного аналізу надають доступ для аналізу всіх проміжних обчислень, що є важливою перевагою.

До недоліків регресійних моделей можна віднести необхідність точних дискретних даних. При регресійному аналізі приймається, що дані, які використовує дослідник, є абсолютно точними. На практиці це припущення дуже часто не виконується. Не детермінованість незалежних змінних призводить до застосування надлишкових моделей кореляційного аналізу.

Також проблемою може бути наявність декількох локальних екстремумів функціоналів. Переважна більшість методів оптимізації функцій дозволяє знаходити лише локальні екстремуми, тоді результат мінімізації залежить від вибору початкових умов. Це зумовлює необхідність встановлення інших критеріїв вибору моделі.

МНК, який широко застосовується при регресійному аналізі, також має ряд свої недоліків у вигляді накладання обмежень на дані:

а) вхідні фактори мають бути детермінованими величинами, а спостереження вихідного значення — випадковими взаємонезалежними величинами;

б) дані не можуть бути мультиколінеарними, тобто не можлива лінійна залежність між вхідними факторами; порушення цього обмеження призводить до неадекватності отриманої регресійної моделі;

в) має виконуватися наближення до нуля математичного сподівання помилки;

г) дисперсія випадкової спостережуваної величини має бути постійною в будь-якій точці спостережень.

2.5.2 Переваги та недоліки факторного аналізу

Застосування факторного аналізу є особливо ефективним в областях, де неможливо маніпулювати спостережуваними даними. Для більшості явищ природи та клімату характерна варіація ознак. Коваріація різних ознак в оточуючому середовищі дає можливість після точного аналізу робити висновки, аналогічні результатам класичних експериментів.

Теоретичні положення факторного аналізу є досить важкими для сприйняття. Практичне застосування факторного аналізу передбачає використання ЕОМ для обробки даних. Метод не має однозначного математичного розв’язку, тому на деяких етапах використовують допоміжні засоби [8]. Через цю невизначеність вчені стверджують, що факторний аналіз, як метод дослідження, не є ефективним та корисним.

В процесі факторного аналізу не завжди можна отримати розв'язок, оскільки складність обчислень власних значень кореляційної матриці досить висока [12]. Наприклад, кореляційна матриця може бути вираженою, що є наслідком лінійної кореляції параметрів. Для матриць великої розмірності при обчисленнях може відбутися втрата значущості.

2.5.3 Переваги та недоліки моделей на основі ланцюгів Маркова

Головною перевагою використання ланцюгів Маркова для прогнозування змінних є можливість використання процесів без інформації про минулі спостереження, проте у випадках, коли процес має довгу пам’ять, цю властивість можна віднести до недоліків даного методу.

Ще одна особливість використання даної моделі є необхідність в дискретних даних за станом та за часом.

2.5.4 Переваги та недоліки нейромережевих моделей

При використанні ANN система має різні корисні властивості. По-перше, система може бути нелінійною. Не існує жодного обмеження на лінійність штучних нейронів: вони можуть бути лінійними та нелінійними. Це є значною перевагою, оскільки більшість даних, які досліджуються на практиці, є нелінійними, наприклад, такі як мова людини. Підхід відображення вхідного сигналу в вихідний схожий на непараметричне статистичне навчання.

Ще однією перевагою нейромереж є здатність до адаптації своїх синаптичних вагів до змін навколишнього середовища, тобто для роботи в нестаціонарному середовищі, де статистика змінюється з часом, моделі ANN є дуже ефективними.

Наступна перевага полягає в гнучкій моделі для нелінійної апроксимації багатовимірних функцій, тобто ANN є зручною методологією прогнозування процесів, які залежать від великої кількості змінних. Завдяки паралельній обробці даних та сильному зв'язку нейронів, на них не впливають обмеження потужностей комп’ютерів.

Дослідження в області моделювання часових рядів за допомогою нейромереж продовжуються і в теперішній час, проте жодних стандартних методів ще не виявлено. Це є основним недоліком нейронних мереж. В ANN багато факторів взаємодіють досить важким чином, тож найкращим є евристичний метод.

2.6 Обґрунтування вибору методу

Процеси, які відбуваються всередині кліматичної системи, мають складні зв'язки та їх не можна однозначно називати лінійними. Під діями цих процесів в кліматичній системі відбувається коливання різних часових масштабів, тому для опису станів кліматичної системи необхідні складні фізико-математичні моделі. Зважаючи на те, що ще немає ідеальної моделі, яка добре описує та прогнозує кліматичні величини, для кількісного опису зв'язків між впливаючими факторами та залежними значеннями для дослідження властивостей кліматичної системи доцільно використовувати апарат математичних моделей з групи ймовірнісно-статистичних.

На основі попереднього розділу, де представлено переваги та недоліки методів та моделей, які можна використати для розв'язку поставленої задачі та особливостей кліматичних даних, яким притаманний дискретний характер, прийнято рішення використання регресійних моделей з застосуванням алгоритмів машинного навчання.

Дані моделі є прозорими, простими при моделюванні, а також мають широке застосування у сфері моделювання кліматичних даних. Ще однією важливою перевагою є те, що вони дозволяють використовувати при розрахунках параметрів вибірки різних об'ємів. Слід зауважити, що на основі обраної моделі можливо побудувати модель нейронної мережі та розширити дослідження кліматичної системи.

2.7 Висновки до розділу

У даному розділі було розглянуто особливості кліматичних даних та математичні методи розв’язання поставленої задачі, а саме: статистичні методи, регресійний та факторний аналізи, моделі на основі ланцюгів Маркова та нейромережеві моделі.

На основі порівняльного аналізу (таблиця 2.2) та розділу 2.6, щодо обґрунтування вибору моделі, зроблено висновок, що для досягнення поставленої мети найкраще підходить застосування регресійних моделей. Перевагами згаданої вище моделі є прозорість та простота моделювання, широке застосування у сфері прогнозування кліматичних даних, гнучкість, можливість до розширення вибірки.

3 МАТЕМАТИЧНЕ ЗАБЕЗПЕЧЕННЯ

3.1 Задача регресії

В задачі регресії необхідно визначити значення залежної змінної об'єкту (середньої температури поверхні землі) та основі значень інших змінних (впливаючих факторів), які характеризують даний об'єкт.

Формально задачу регресії для заданої кліматичної моделі можна записати наступним чином.

Нехай існує множина об'єктів:

$$I = \{i_1, i_2, \ldots, i_j, \ldots, i_n\}, \tag{3.1}$$

кожен елемент якої є значенням середньої температури поверхні землі та характеризується набором змінних:

$$I_j = \{x_1, x_2, \ldots, x_h, \ldots, x_m, y\}, \tag{3.2}$$

де x_h — значення h-фактору, який впливає на середню температуру поверхні землі.

Кожна змінна x_i може приймати значення з деякої множини:

$$C_h = \{c_{hi}\}. \tag{3.3}$$

Побудову функції регресії можна формально описати як задачу вибору функції з мінімальною помилкою:

$$min\ R(f) = \frac{1}{m}\sum_{i=1}^{m} \quad c(y_i, f(x_i)), \tag{3.4}$$

де F — множина всіх можливих функцій, які описують дані, $c(y_i, f(x_i))$ — функція втрат.

Задача регресії розв'язується у два етапи. На першому виділяється навчальна вибірка даних, в яку включаються як і залежні зміні, так і незалежні. За допомогою цієї вибірки будується модель визначення значення залежної змінної. Кількість об'єктів, які входять в вибірку, має бути достатньо великою: чим більше об'єктів, тим точніше буде модель.

Для виконання цієї умови в роботі 80% даних належать до навчальної вибірки. Розподілення об'єктів між навчальними та тестовими вибірками відбувається випадковим чином.

На другому етапі побудована модель застосовується до об'єктів тестувальної вибірки та оцінюється адекватність прогнозів.

Основна проблема, з якою зустрічаються при розв'язанні задач регресії — недостатня кількість вхідних даних та дані, в яких зустрічаються пропущені та помилкові значення.

Через характер вимірювань кліматичних даних, які використовуються в даній роботі, вони не узгоджуються між собою. Наприклад, у 1900 році може бути відоме значення вуглекислого газу та відповідна температура, але можуть не існувати дані про концентрації метану та оксид азоту. Такий дискретний характер даних зумовлює необхідність підготовки цих даних до аналізу.

3.2 Методи машинного навчання

Машинне навчання — це сукупність статистичних методів для аналізу тенденцій, пошуку зв'язків та розробки моделей прогнозування змінних на основі наборів даних. В даній роботі на основі аналізу алгоритмів лінійних та нелінійних

моделей машинного навчання для дослідження глобального потепління розглянуто наступні:

а) лінійна регресія;

б) Random Forest;

в) SVR.

3.3.1 Модель лінійної регресії

В даній моделі шукана функція F має наступний вигляд:

$$y = \omega_0 + \omega_1 x_1 + \ldots + \omega_n x_n = \omega_0 + \sum \omega_j x_j, \quad (3.5)$$

де $\omega_0, \omega_1, \ldots, \omega_n$ — коефіцієнти при незалежних змінних.

Задача полягає в пошуку таких коефіцієнтів ω, щоб задовольнити умову загальної задачі регресії.

Будемо використовувати квадратичну функцію втрат:

$$F := \{ f \mid f(x) = \sum_{i=1}^{n} \omega_i f_i(x), \omega_i \in R \}, \quad (3.6)$$

де $f_i: X \to R$.

Необхідно знайти розв'язок наступної задачі:

$$min\, R(f) = min \frac{1}{m} \sum_{i=1}^{m} (y_i - \sum_{i=1}^{n} \omega_i f_i(x))^2 . \quad (3.7)$$

Знайдемо похідну $R(f)$ по ω. Якщо ввести позначення $Y_{ij} := f_i(x_j)$, отримаємо мінімум при умові:

$$Y^T y = Y^T Y \omega. \quad (3.8)$$

Розв'язком цього виразу буде:

$$\omega = (Y^T Y)^{-1} Y^T y, \quad (3.9)$$

за допомогою якого знаходимо шукані коефіцієнти ω, а отже вигляд функції регресії.

3.3.2 Random Forest

В основі алгоритму Random Forest лежить поняття дерева рішення, воно є базовою одиницею випадкового лісу.

Дерева рішень — це спосіб представлення правил в ієрархічній, послідовній структурі.

Кожен вузол дерева має перевірку певної залежної змінної. У випадках, коли змінна приймає чисельне значення, то відбуваються перевірка більше/менше відносно деякої константи. Листя дерева відповідають значенням залежної змінної.

При використанні алгоритмів лісів регресії умови поділу вузлів визначаються таким чином, щоб Gini Impurity зменшувалось. Gini Impurity — це ймовірність невірного маркування у вузлі випадково обраного об'єкта. Її можна обчислити за наступною формулою:

$$I_G(n) = 1 - \sum_{i=1}^{J} (p_i)^2. \quad (3.10)$$

В кожному вузлі дерево рішень шукає таке значення певного параметру, яке призведе до максимального зменшення Gini Impurity. Таким чином, процес рекурсивно повторюється, поки дерево не досягне максимальної глибини.

Random forest — алгоритм машинного навчання, який полягає в використанні ансамбля дерев прийняття рішень: на кожній ітерації алгоритм робить випадкову вибірку елементів, після чого на новій виборці запускається побудова дерев прийняття рішень.

В процесі тренування кожне дерево вчиться на випадковій вибірці з набору даних. Вибірка відбувається з бутстреппінгом, що дає можливість повторно використовувати об'єкти одним й тим самим деревом.

При тестуванні результат виводиться шляхом усереднення прогнозів, отриманих від кожного дерева.

Важливою проблемою, яка зустрічається при реалізації алгоритму Random Forest, є перенавчання системи. Таку систему називають високоваріативною та надто гнучкою, що призводить до того, що модель визначає не тільки закономірності, а ще стає надто чутливою до шумів. Для балансу використовується обмеження на кількість дерев.

В моделі Random Forest, розглянутій на кліматичних даних, методом підбору було вирішено, що найоптимальніша кількість дерев має дорівнювати 20.

3.3. Support Vector Regression

Ідея методу ґрунтується на припущенні про те, що найкращим способом поділу точок в m-вимірному просторі є $(m - 1)$ площина, рівновіддалена від точок, які належать різним класам. Для двовимірного простору цю ідею можна представити у вигляді, зображеному на рисунку 3.1.

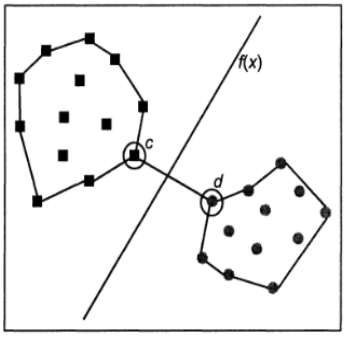


Рисунок 3.1 — Графічна інтерпретація ідеї методу SVR[6]

Незалежні змінні, які характеризують об'єкт, є координатами векторів.

Формально таку задачу можна описати як пошук функції, яка відповідає наступним умовам:

$$< \omega, x > + b - y_i \leq \varepsilon\,, \tag{3.11}$$

$$y_i - < \omega, x > - b \leq \varepsilon. \tag{3.12}$$

Введемо скалярний добуток перетворених векторів:

$$k(x, x') := < \Phi(x), \Phi(x') >. \tag{3.13}$$

Функція $k(x, y)$називається ядром.

Розв'язком даної задачі буде функція вигляду:

$$f(x) \;=\; \sum_{i=1}^{l} \quad (\alpha_i - \alpha_i^*) k(x_i, x) \;+ b. \tag{3.14}$$

Вид перетворення $k(x_i, x)$ може бути різного типу та обирається в залежності від структури даних.

В даній роботі було обрано базову радіальну функцію Гауса, яка має вигляд:

$$k(x, y) = exp(-\gamma ||x - y||). \tag{3.15}$$

3.2 Інтерполяційний поліном Лагранжа

Алгоритми машинного навчання не можуть ефективно обробляти пропущені точки даних, тому в даній роботі використана лінійна інтерполяція для вирівнювання цих даних та створення неперервних величин.

В основі апроксимації лежить заміна частково відомої функції $f(x)$ іншою $\varphi(x)$, наближеною до $f(x)$, яка має властивості зручні для застосування інших методів, безпосередньо, методів машинного навчання. Така функція $\varphi(x)$ називається апроксимацією або наближенням функції $f(x)$.

Тож задача апроксимації функції $f(x)$ функцією $\varphi(x)$ полягає в побудові такої функції $\varphi(x)$, що:

$$f(x) \simeq \varphi(x). \qquad (3.16)$$

Якщо в якості апроксимуючої функції $\varphi(x)$ використовуються многочлени або функції утворені з многочленів, така апроксимація називається поліноміальною апроксимацією або кусково-поліноміальною апроксимацією.

Точки, в яких відома інформація про $f(x)$ називаються вузлами.

Розглянемо один з методів побудови апроксимуючої функції — побудова інтерполяційного полінома Лагранжа.

Нехай відомі значення функції $f(x)$ в точках $x_0, x_1, \ldots, x_n$, які визначенні на відрізку $[a, b]$, тобто функція задана таблично.

Функція $\varphi(x)$ називається інтерполяційною для $f(x)$ на $[a, b]$, якщо її значення $\varphi(x_0), \varphi(x_1), \ldots, \varphi(x_n)$ в вузлах інтерполяції $x_0, x_1, \ldots, x_n$ співпадає зі значеннями заданої функції $f(x)$.

З геометричної точки зору, інтерполяція означає, що графік функції $\varphi(x)$ перетинає, як мінімум $(n + 1)$ заданих точок (рисунок 3.2).

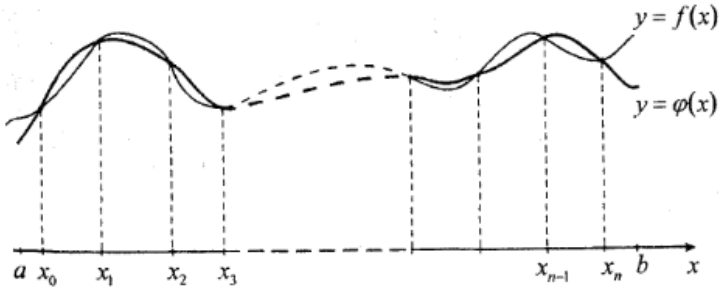


Рисунок 3.2 — Геометрична інтерпретація задачі інтерполювання [7]

Задачу інтерполяції можна сформулювати наступним чином. Для функції $f(x)$, заданої таблично, знайти многочлен $P_n(x)$ такий, що виконується сукупність умов інтерполяції:

$$P_n(x_i) = y_i \ \ \forall i \in \{0, 1, \ldots, n\}. \tag{3.17}$$

Необхідно знайти коефіцієнти $a_0, \alpha_1, \ldots, a_n$ полінома:

$$P_n(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n, \tag{3.18}$$

таким чином, щоб вони задовольняли систему рівнянь:

$$\begin{gathered} \alpha_0 + \alpha_1 x_0 + \alpha_2 x_0^2 + \cdots + \alpha_n x_0^n = y_0, \\ \alpha_0 + a_1 x_1 + a_2 x_1^2 + \ldots + a_n x_1^n = y_1, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \alpha_0 + a_1 x_n + a_2 x_n^2 + \ldots + a_n x_n^n = y_n. \end{gathered} \tag{3.20}$$

Будемо будувати многочлен $L_n(x)$ у вигляді лінійної комбінації $\sum_{i=0}^{n} \quad c_i l_i(x)$ многочленів $l_i(x) \ \ (i = 0, 1, \ldots, n)$. Для того, щоб цей многочлен був інтерполяційним для функції $f(x)$ достатньо зафіксувати в якості коефіцієнтів c_i цієї лінійної комбінації відомі значення функції $y_i = f(x_i)$.

Таким чином для многочлена

$$L_n(x) = \sum_{i=0}^{n} \quad y_i l_i(x) \tag{3.21}$$

в кожному вузлі x_j $(j \in \{0, 1, \ldots, n\}$ виконується:

$$L_n(x_j) = l_0(x_j)y_0 + \ldots + l_{j-1}(x_j)y_{j-1} + l_j(x_j)y_j + l_{j+1}(x_j)y_{j+1} + \ldots + l_n(x_j)y_n \text{ , } L_n(x_j) = 0 + \ldots + 0 + y_j + 0 + \ldots + 0 = y_j \text{ .} \tag{3.22}$$

Перепишемо базисні многочлени $l_i(x)$ в наступному вигляді:

$$l_i(x) \; = \; A_i(x - x_0)\ldots(x - x_{i-1})(x - x_{i+1})\ldots(x - x_n). \tag{3.23}$$

Зберігаючи виконання умову інтерполяції, отримаємо, що:

$$A_i = \frac{1}{(x-x_0)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}. \tag{3.24}$$

Таким чином, базисні многочлени Лагранжа мають вигляд:

$$l_i(x) = \frac{(x-x_0)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)}, \tag{3.25}$$

а шуканий інтерполяційний поліном Лагранжа:

$$L_i(x) = \sum_{i=0}^{n} \quad \left(\frac{(x-x_0)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)}\right) y_i. \tag{3.26}$$

При $n = 1$ поліном Лагранжа приймає наступну форму:

$$L_1(x) = \frac{x-x_1}{x_0-x_1} y_0 + \frac{x-x_0}{x_1-x_0} y_1 \tag{3.27}$$

та називається формулою лінійної інтерполяції.

3.4 Оцінка моделі

В якості головного критерію вибору моделі, яка найкраще описує дані, та для оцінки адекватності моделей, в роботі використано RMSE або квадрат середньої квадратичної помилки.

Суть даного методу полягає в мінімізації суми квадратів відхилень фактичних значень від передбачуваних. Якщо отриману суму SSE розділити на число спостережень, то отримаємо MSE. Формула для обчислення помилки наступна:

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \widehat{y_i})^2, \quad (3.28)$$

де N — загальна кількість спостережень, y_i — фактичні дані, $\widehat{y_i}$ — передбачувані дані.

Для знаходження RMSE використовується формула:

$$RMSE = \sqrt{MSE} \quad (3.29)$$

Для оцінки отриманих моделей використовуються дані з тестових вибірок.

3.5 Висновки до розділу

У даному розділі було проаналізоване математичне забезпечення системи моделювання впливу різних факторів на середню температуру поверхні землі. Було детально розглянуто методи машинного навчання, а саме лінійну регресію, Random Forest та SVR, які досліджуються в даній роботі.

Для усунення проблеми неузгодженості даних з різних джерел було обрано інтерполяційний поліном Лагранжа.

Також визначенні гіперпараметри для моделей: для алгоритму SVR обрана базова радіальна функція Гауса, як функція ядра, в моделі Random Forest кількість дерев дорівнює 20.

Для оцінки адекватності досліджуваних моделей використовується RMSE.

4 ПРОЕКТУВАННЯ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

4.1 Структура програми

Програмне забезпечення у даній роботі реалізовано у вигляді вебдодатку, який за допомогою інформаційної панелі виводить результати аналізу кліматичних даних. Вебдодаток дозволяє користувачу зручно отримувати результати дослідження у вигляді графіків.

Для реалізації поставленої мети було використано наступні технології:

а) Python 3.6.9;

б) Jupyter Notebook;

в) Flask;

г) Bootstrap3;

д) Jinja2.

За допомогою Flask було створено вебдодаток, який реалізує усю роботу з користувачем через розроблений сайт. Користуючись засобами Bootstrap3 та Jinja2, було створено дизайн сайту та зверстано зручний користувацький інтерфейс.

Для аналізу даних було використано бібліотеки Python, такі як numpy, pandas, scipy та sklearn.

Програмне забезпечення розділено на наступні модулі:

а) завантаження та попередня обробка даних;

б) модуль прогнозування;

в) модуль обчислення похибок;

г) модуль візуалізації даних;

д) інтерфейс програми та робота вебдодатку.

Схема взаємодії всі цих модулів зображена на рисунку 4.1.

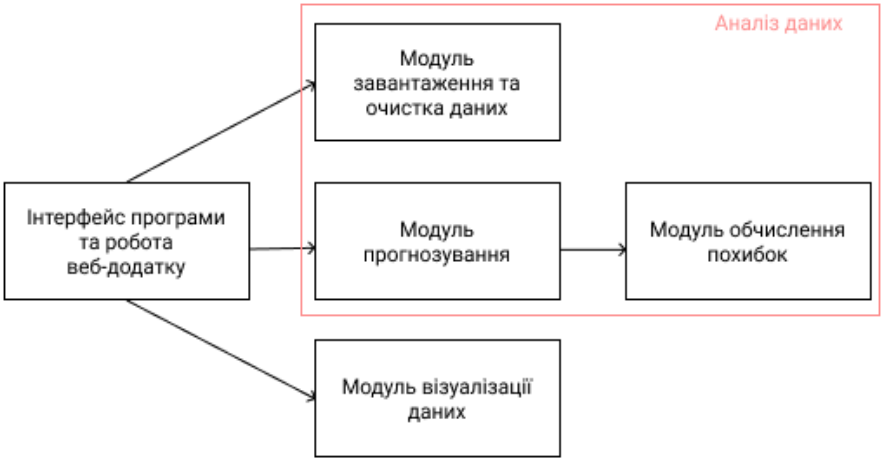


Рисунок 4.1 — Структура програми

Сайт містить різні сторінки, на яких знаходяться дані про дослідження. При відкритті вебдодатку користувач потрапляє на першу сторінку, де розташована основна інформація про дослідження, така як мета дослідження, задачі дослідження та інше (рисунок 4.2).

Рисунок 4.2 — Головна сторінка вебдодатку

На інших вкладках, перейти за якими можна скориставшись верхнім меню, знаходяться графіки прогнозування середньої температури поверхні землі, залежно від факторів, які розглянуті в даній роботі (рисунок 4.3).

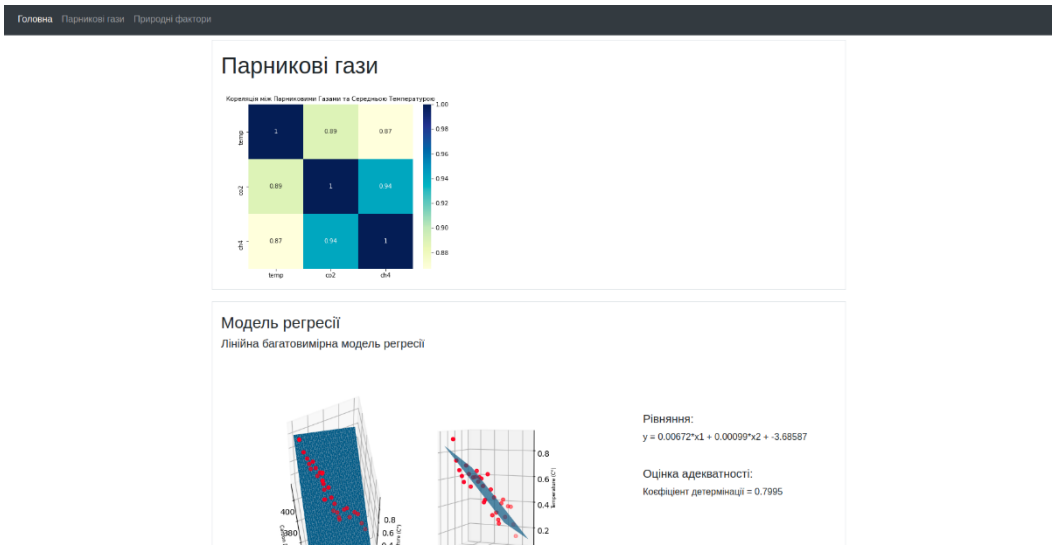


Рисунок 4.3 — Сторінка вебдодатку «Парникові гази»

4.2 Вхідні дані

Для виконання дослідження та досягнення поставленої мети було проведено аналіз джерел [14] та виявлено природні та антропогенні фактори, які впливають на зміни клімату. На рисунку 4.4 зображено схему цих факторів впливу.

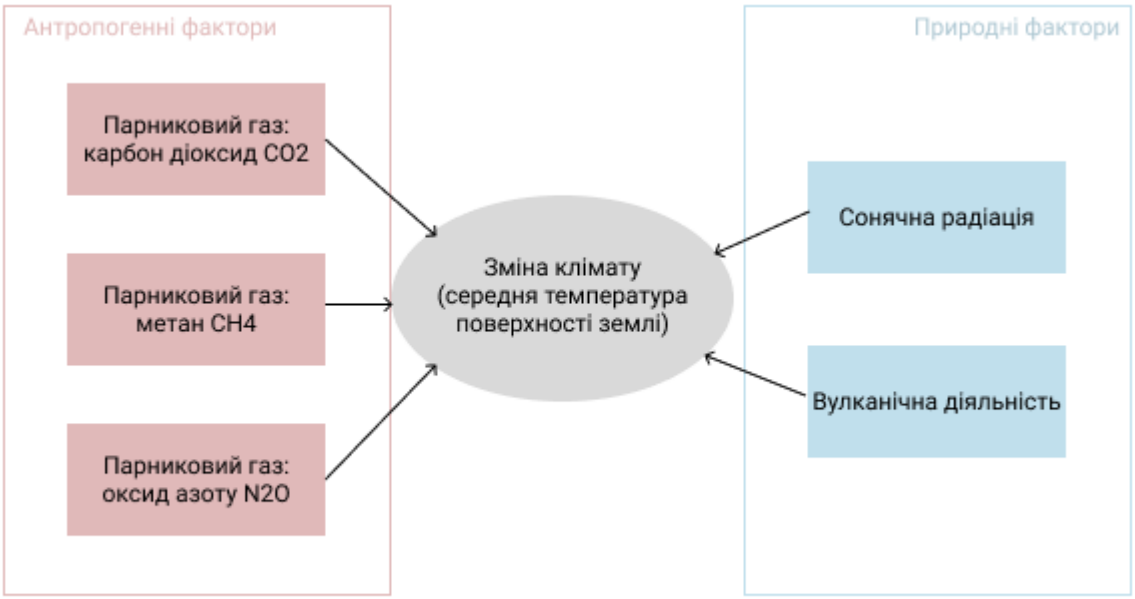


Рисунок 4.4 — Схема факторів впливу на середню температури поверхні землі

До антропогенних факторів, які згідно публікацій [15] найбільше впливають на зміни клімату, належать дані про концентрації парникових газів в атмосфері Землі, а саме карбон діоксиду [16], метану [17] та оксид азоту [18].

До природних факторів було віднесено дані про сонячну активність [19] та вулканічні викиди [20]. Вулканічну активність було визначено величиною, яка є сумою рівнів VEI всіх вулканів, виверження яких відбувалося протягом року.

Дані було взято з публічної бібліотеки NOAA.

В роботі для характеристики зміни клімату використана середня глобальна температура поверхні землі, представлена Lawrence Berkeley National DataBase.

В програму всі дані поступають в форматі .csv.

Для забезпечення рівномірності та гладкості даних необхідно побудувати інтерполяційний поліном Лагранжа.

Оскільки всі дані мають різні одиниці вимірювання, для коректної роботи алгоритмів машинного навчання відбувається нормування всіх вхідних даних таким чином, що кожна величина лежала в діапазоні від 0 до 1. Це досягається шляхом

стандартизації, коли кожен об'єкт в наборі має нульове середнє значення та одиничну дисперсію.

4.3 Висновки до розділу

У даному розділі було проведено загальний опис архітектури програмного забезпечення, зазначено технології для розв’язку поставленої задачі — Python, Flask, Bootstrap3 та інші.

Була надана інформація про джерела вхідних даних в систему та зазначено формат даних, необхідний для роботи програмного забезпечення. Дані, що використанні в даному дослідженні, потребують попередньої обробки за допомогою інтерполяційного поліному Лагранжа та нормування величин шляхом стандартизації нульового середнього значення та одиничної дисперсії.

5 ТЕСТУВАННЯ ТА ВИПРОБУВАННЯ СИСТЕМИ

5.1 Аналіз вхідних даних

Дані про концентрацію парникових газів в атмосфері з використаних наборів даних протягом 800 000 років зображено на рисунку 5.1.

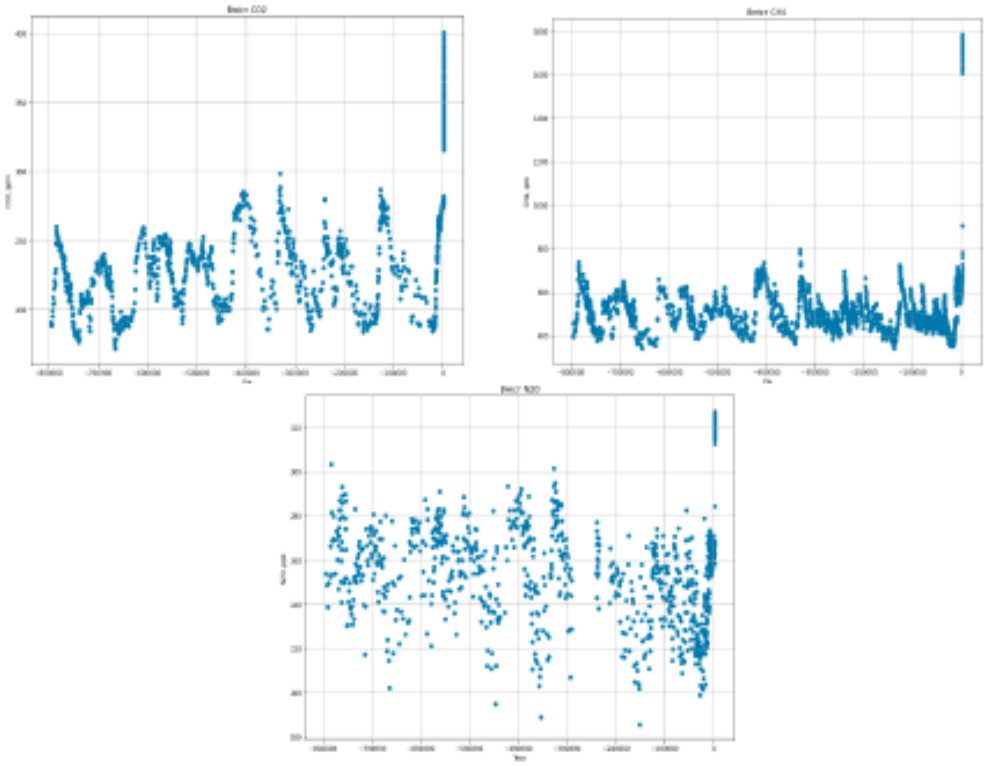


Рисунок 5.1 — Графіки концентрації парникових газів в атмосфері

Зважаючи на особливості вхідних даних, необхідно було інтерполювати значення наборів даних та утворити неперервні функції. На рисунку 5.2 зображено результати інтерполяції даних про парникові гази.

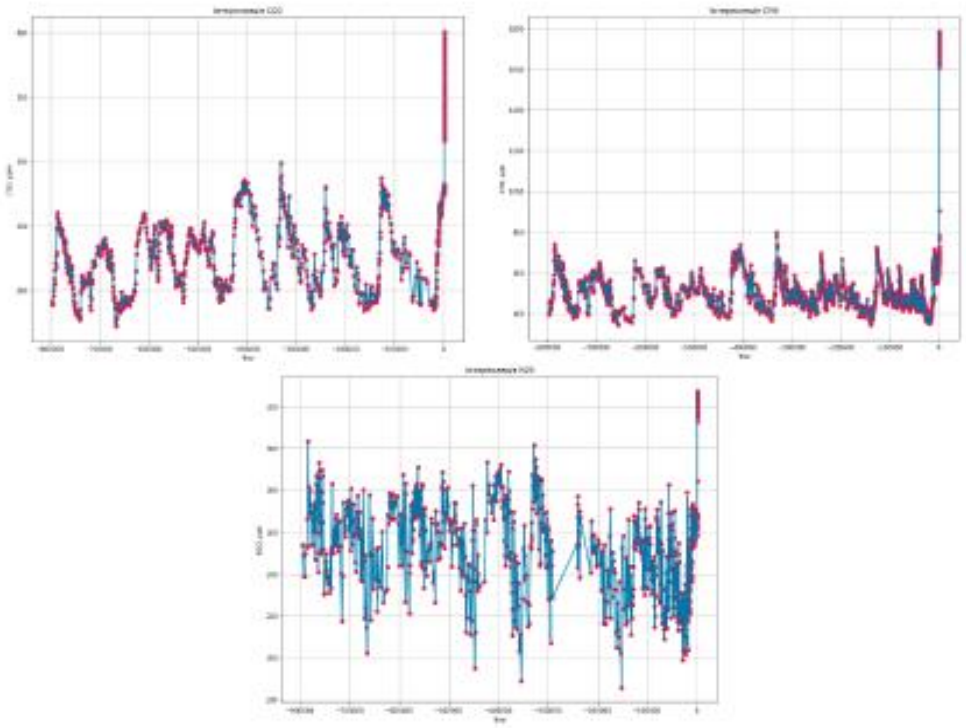


Рисунок 5.2 — Графіки інтерполяційних функцій парникових газів

Дані про сонячну радіацію з набору даних з 1882 року по 2015 рік представлені на рисунку 5.3 Оскільки в даних немає пропусків в інтерполяції немає необхідності.

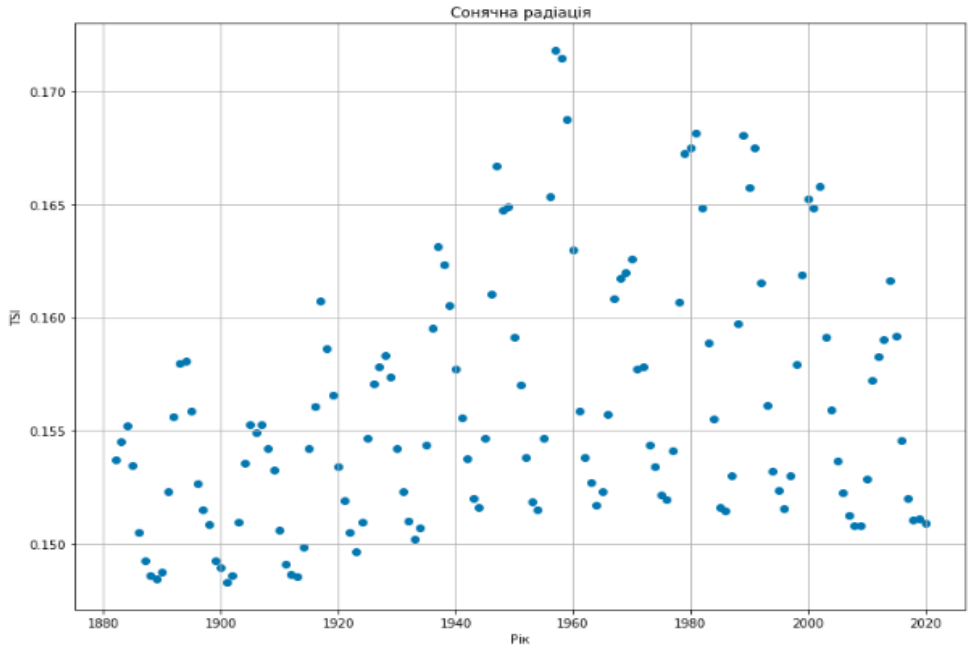


Рисунок 5.3 — Графік даних про сонячну радіацію

Дані про вулканічну активність з 1800 року зображено на рисунку 5.4.

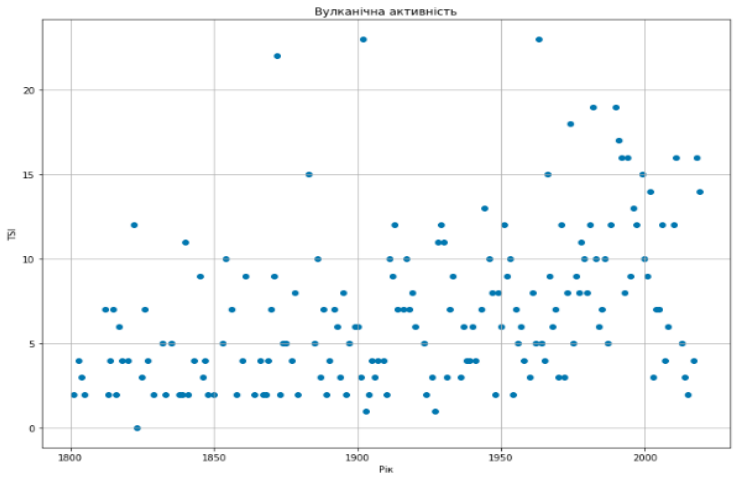


Рисунок 5.4 — Графік даних про вулканічну активність

Оскільки дані мали пропуски, необхідно було побудувати інтерполяційні поліноми (рисунок 5.5).

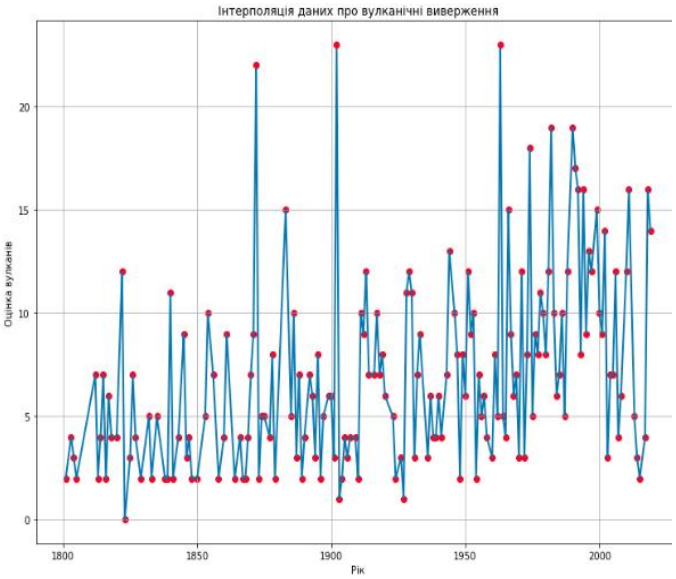


Рисунок 5.5 — Графіки інтерполяційної функцій вулканічної активності

Дані про залежність середньої температури поверхні землі від року зображені на рисунку 5.6. За даним графіком можна спостерігати тенденцію збільшення середньої температури поверхні землі протягом останніх 50 років.

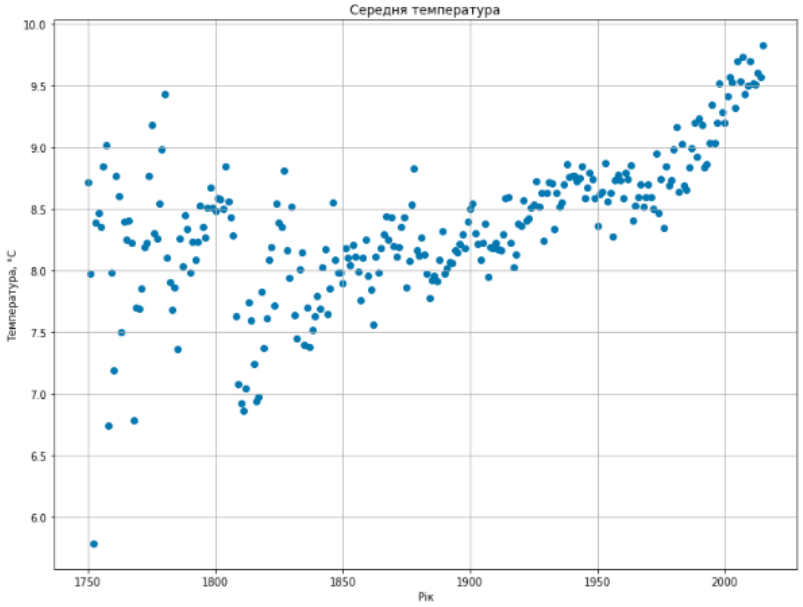


Рисунок 5.6 — Графік залежності середньої температури від року

Після нормалізації дані випадковим чином розподіляються на дві вибірки: тестову та навчальну у відношенні 4:1.

5.2 Результати прогнозування загальної моделі

Було побудовано загальну модель, в якій враховано антропогенні та природні фактори впливу. У ролі антропогенних факторів було взято концентрації парникових газів, а саме карбон діоксиду, метану та оксид азоту. Природні фактори розглядаються у вигляді сонячної радіації та показника вулканізму.

За допомогою вхідних даних представлених у попередньому розділі та алгоритмів машинного навчання було побудовано три моделі регресії:

а) лінійна регресія;

б) Random Forest;

в) SVR.

Скориставшись однаковими вибірками даних для навчання, для всіх алгоритмів було навчено розглянуті моделі та отримані регресійні функції для прогнозування залежної змінної. В силу особливостей алгоритмів не завжди ці функції мають аналітичне представлення.

Щоб оцінити модель, за допомогою тестової вибірки незалежних змінних, тобто даних усіх факторів, які розглянуто в цій роботі, було знайдено передбачувані середні температури поверхні землі. На основі цих передбачених залежних змінних та тестової вибірки відомих значень було оцінено адекватності моделей за допомогою RMSE.

Результат адекватності лінійної регресії для тестових даних наступний:

$$RMSE = 0.341. \quad (5.1)$$

При оцінці похибки навчальних даних отримано $RMSE = 0.404$.

Це вказує на те, що модель є досить адекватною та можлива для використання.

Функція лінійної регресії буде мати наступний вигляд:

$$Y = 0.369X_1 + 1.203X_2 + 1.699X_3 + 0.031X_4 + 0.024X_5 - 0.005, \quad (5.2)$$

де Y — вектор прогнозованих середніх температур, X_1 — вектор концентрації CO_2, X_2— вектор концентрації CH_4, X_3— вектор концентрації N_2O, X_4— вектор показників VEI вулканічної активності, X_5 — вектор значення сонячної радіації.

Для побудови моделі Random Forest було визначено кількість дерев 20. Похибка цієї моделі на тестових даних складає $RMSE = 0.277$. При оцінці похибки на навчальних даних $RMSE = 0.154$.

Вектор оцінки важливості незалежних змінних при алгоритмі Random Forest дорівнює $(0.286, 0.368, 0.285, 0.025, 0.036)$.

Побудувавши модель алгоритму SVR, було отримано наступні помилки для тестових та навчальних даних відповідно:

$$RMSE = 0.338, \quad (5.3)$$

$$RMSE = 0.153. \quad (5.4)$$

Результати навчання та тестування моделей у вигляді їх оцінки RMSE для порівняння наведено в таблиці 5.1. Видно, що модель Random Forest дозволяє знайти найадекватнішу модель. Далі буде показано, що з візуального огляду також можна стверджувати, що ця модель найбільш ефективно працює на розглянутому наборі даних і створює найбільш точну модель прогнозування температури за визначеними факторами.

Таблиця 5.1 — RMSE помилка кожної моделі на основі навчальних та тестових даних

Алгоритм	Навчальне RMSE	Тестове RMSE
Лінійна регресія	0.404	0.341
Random Forest	0.154	0.227
SVR	0.153	0.338

Таким чином, щоб побудувати більш точну модель прогнозування температури з більшим набором властивостей і зв'язків, Random Forest є найкращим варіантом із трьох розглянутих алгоритмів. Точність даного алгоритму дозволяє оцінити ступінь важливості впливу різних факторів на температуру, які представлені у таблиці 5.2.

Таблиця 5.2 — Важливість кожного фактору визначена Random Forest

Фактор	Важливість
CO_2	0.286
CH_4	0.368

N_2O	0.285
VEI	0.025
STI	0.036

На основі цієї таблиці було побудовано діаграму, яка зображена на рисунку 5.7.

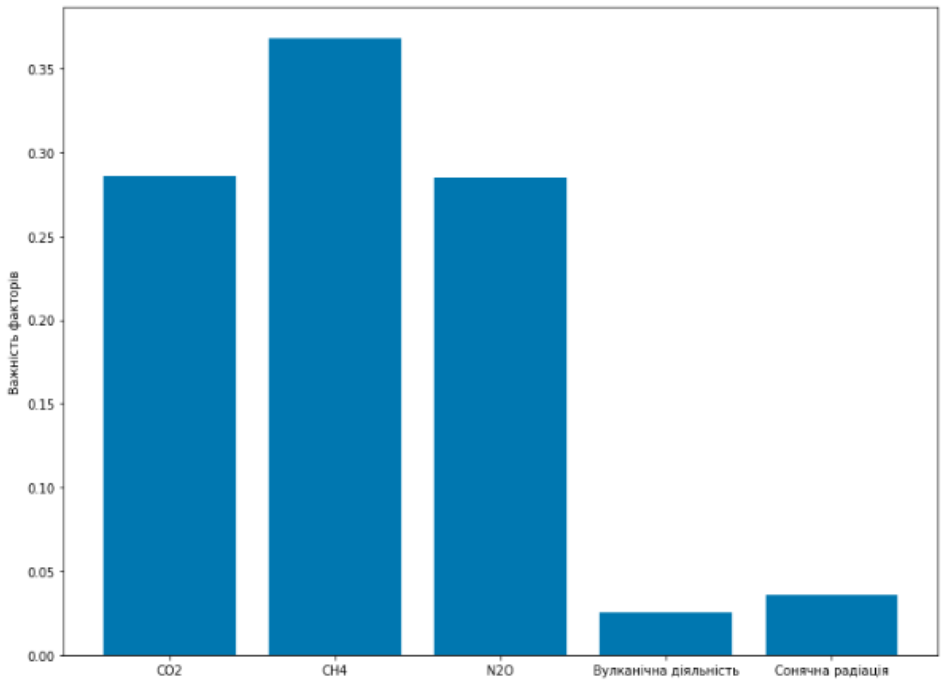


Рисунок 5.7 — Важливість кожного фактору, визначена Random Forest.

Як видно з діаграми значущості метан має найбільш значущий коефіцієнт 0.368, за ним йде вуглекислий газ з 0.286 та оксид азоту з коефіцієнтом 0.285. Значно менший вплив на середню температуру поверхні землі мають природні фактори: вулканічна діяльність має коефіцієнт 0.025 , а сонячна радіація — 0.036.

Таким чином за допомогою машинного навчання була підтверджена заява міжнародної групи з питань клімату IPCC про причину антропогенного впливу глобального потепління.

Оскільки парникові гази мають значно більший вплив, розглянемо детальніше результати побудованих моделей на основі цих даних.

5.3 Результати прогнозування моделей для парникових газів

За допомогою розглянутих вище алгоритмів було побудовано одновимірні моделі прогнозування залежності кожного парникового газу та температури.

Як найбільш впливовий фактор з парникових газів, розглянемо моделі прогнозування залежності між метаном та температурою. На рисунках 5.8 — 5.10 зображено візуалізація моделей на навчальних та тестових даних.

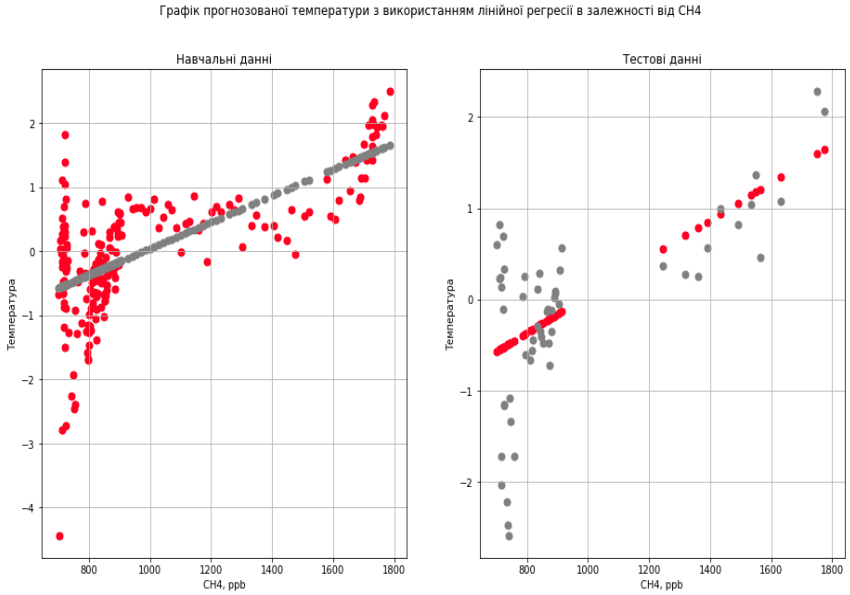


Рисунок 5.8 — Графік прогнозування залежності температури та CH_4 за допомогою лінійної регресії

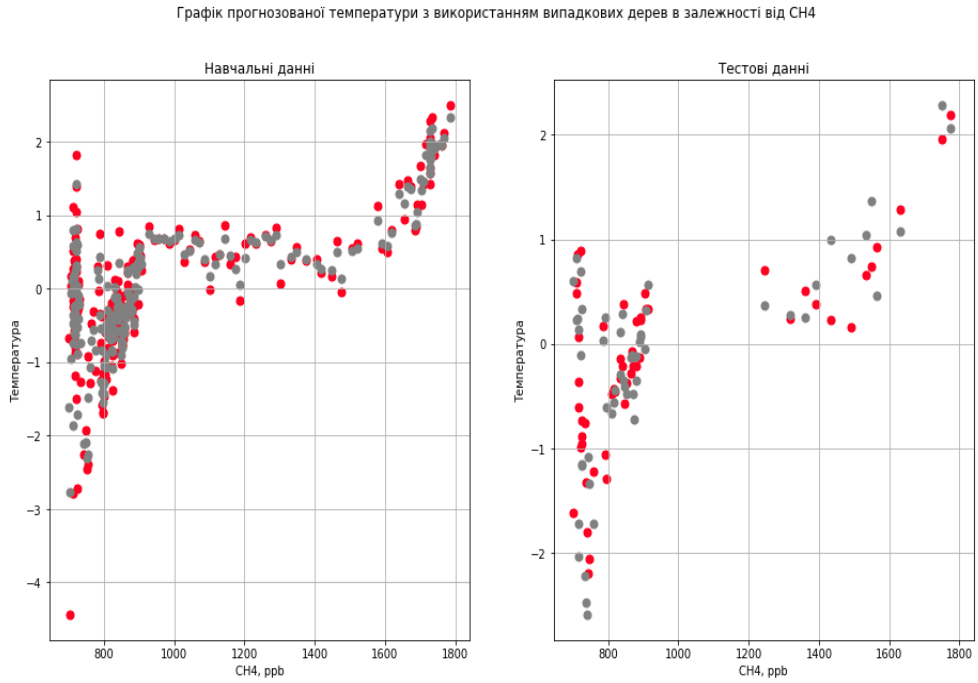


Рисунок 5.9 — Графік прогнозування залежності температури та CH_4 за допомогою Random Forest

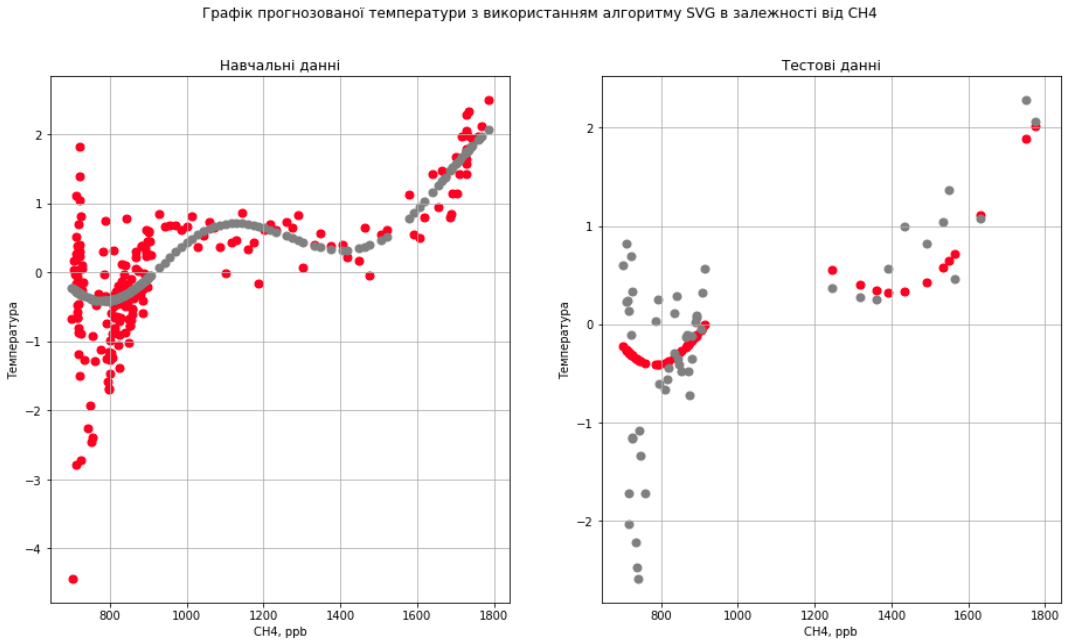


Рисунок 5.10 — Графік прогнозування залежності температури та CH_4 за допомогою SVR

Оцінки RMSE заданих моделей для метану представлено у таблиці 5.3.

Таблиця 5.3 — Оцінки адекватності моделей залежності CH_4 та температури

Алгоритм	RMSE
Лінійна регресія	0.749
Random Forest	0.673
SVR	0.277

Візуалізація моделей залежності CO_2 та середньої температури поверхні землі зображено на рисунках 5.11 — 5.13.

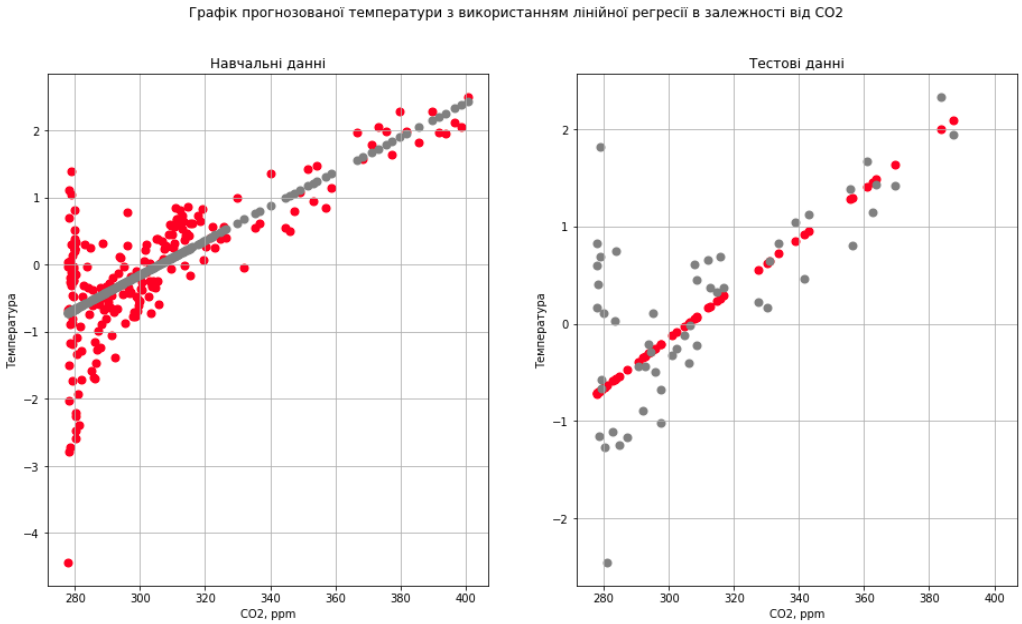


Рисунок 5.11 — Графік прогнозування залежності температури та CO_2за допомогою лінійної регресії

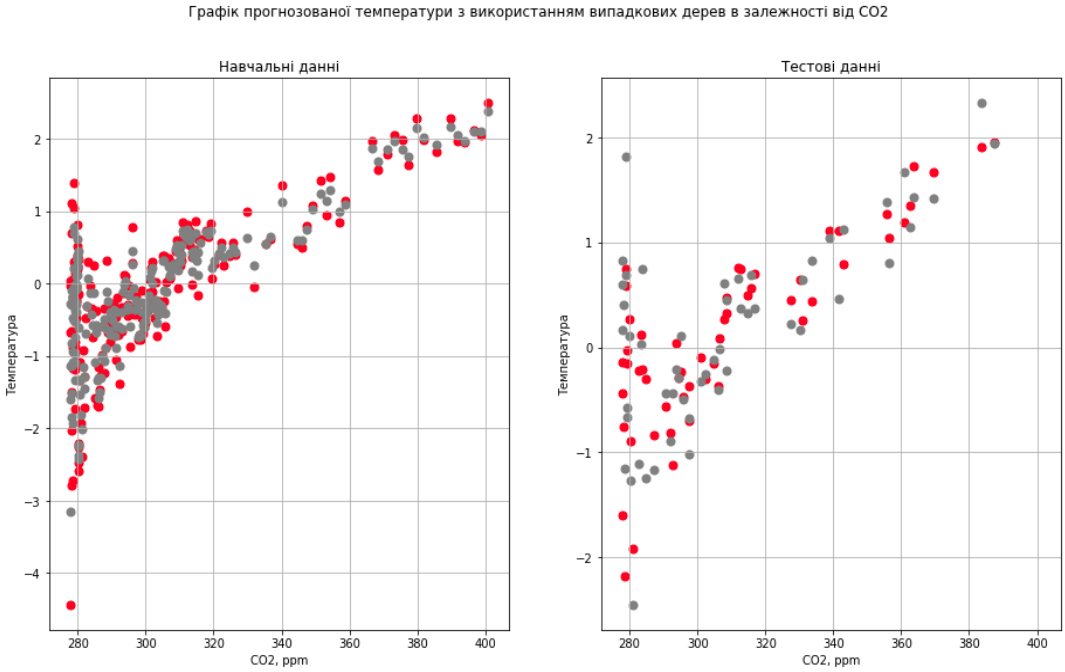


Рисунок 5.12 — Графік прогнозування залежності температури та CO_2за допомогою Random Forest

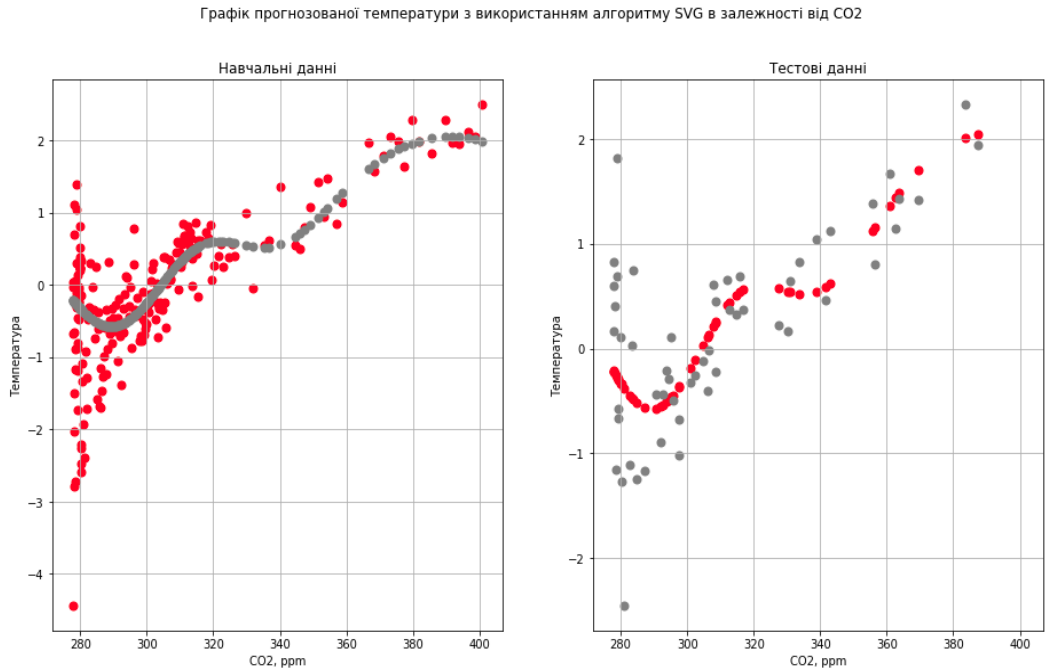


Рисунок 5.13— Графік прогнозування залежності температури та CO_2за допомогою SVR

Результати оцінки розглянутих моделей наведені у таблиці 5.4.

Таблиця 5.4 — Оцінки адекватності моделей залежності CO_2 та температури

Алгоритм	RMSE
Лінійна регресія	0.695
Random Forest	0.569
SVR	0.619

При побудові моделей машинного навчання на основі даних про концентрацію оксид азоту та середньої температури було отримано графіки зображені на рисунках 5.14 — 5.16.

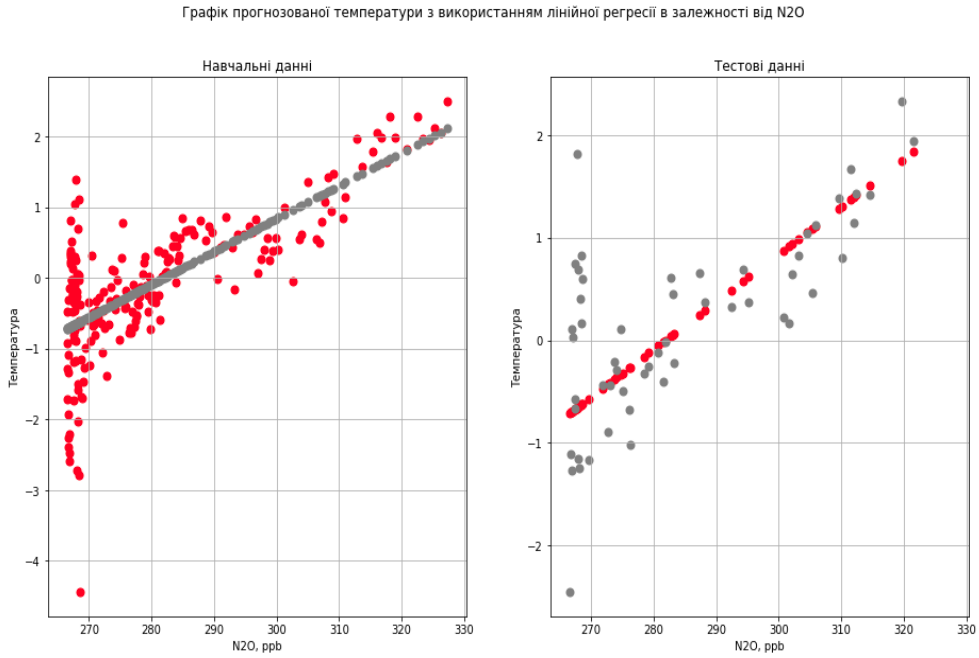


Рисунок 5.14 — Графік прогнозування залежності температури та N_2O за допомогою лінійної регресії

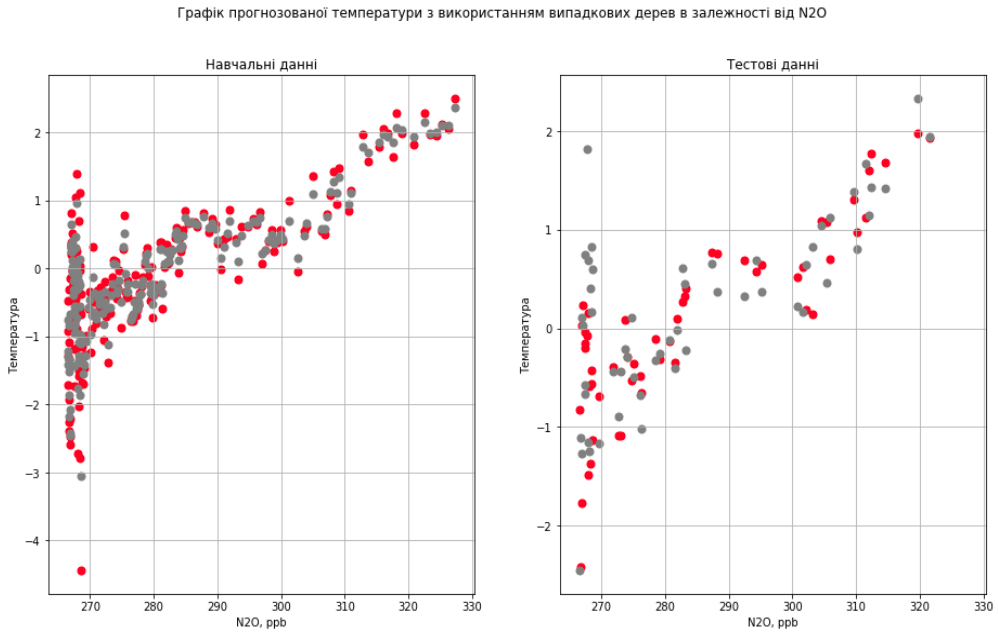


Рисунок 5.15 — Графік прогнозування залежності температури та N_2O за допомогою Random Forest

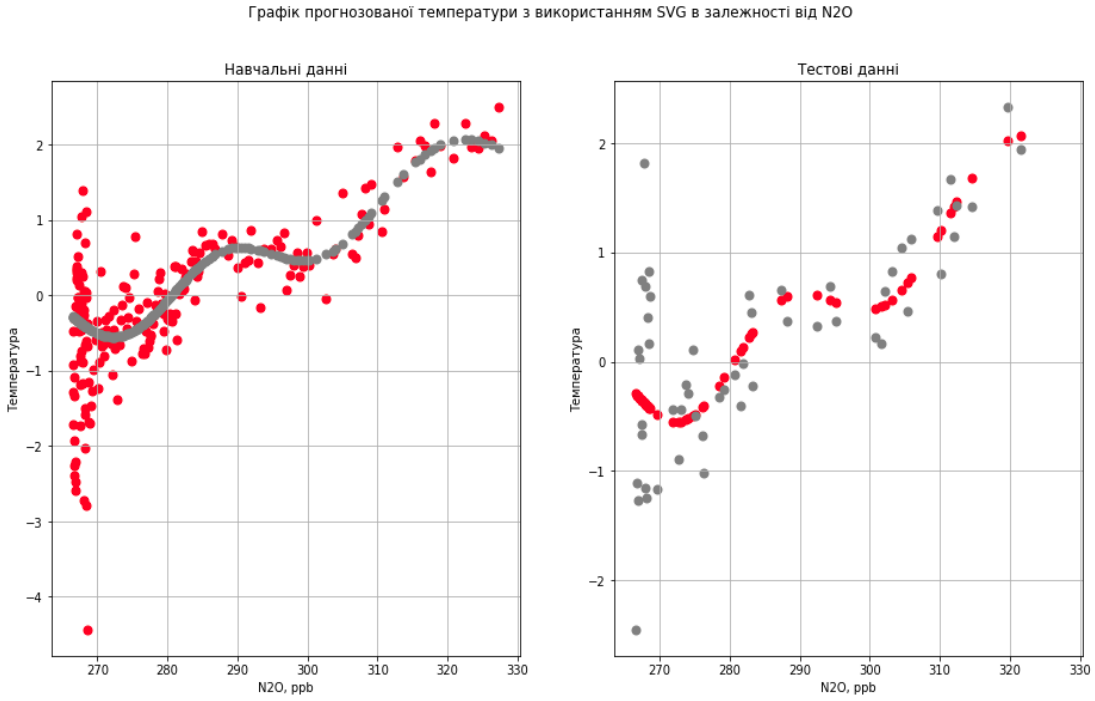


Рисунок 5.16 — Графік прогнозування залежності температури та N_2O за допомогою SVR

Оцінки адекватності моделей представлені у таблиці 5.5.

Таблиця 5.5 — Оцінки адекватності моделей залежності CO_2 та температури

Алгоритм	RMSE
Лінійна регресія	0.688
Random Forest	0.657
SVR	0.645

Наведені результати демонструють результати прогнозування за різними факторами на основі розглянутих алгоритмів машинного навчання в даній роботі. Візуально можна побачити, що алгоритм Random Forest, який в попередньому розділі було визначено найкращим за ознакою похибки RMSE, добре прогнозує середню температуру поверхності землі.

5.4 Висновки до розділу

В даному розділі була представлено візуалізацію розглянутих наборів даних, показано графіки інтерполяції цих даних та представлено результати побудованих моделей. В результаті обчислення адекватностей моделей, шляхом знаходження квадрата середньоквадратичної помилки, було визначено, що найкращим алгоритмом машинного навчання для кліматичних даних представлених у даній роботі є метод Random Forest, RMSE якого складає 0.227.

На основі отриманої моделі було знайдено оцінки важливості незалежних змінних. Визначено, що найбільший вплив на зміну середньої температури поверхні землі має метан, коефіцієнт якого у моделі дорівнює 0.368. Не набагато менший вплив мають два інших парникових газа: вуглекислий газ та оксид азоту, коефіцієнти яких відповідно дорівнюють 0.286 та 0.285 відповідно.

В свою чергу природні фактори в розглянутій моделі значно менше впливають на середню температуру поверхні землі. Показник сонячної активності має коефіцієнт 0.036, а вулканічний показник VEI — 0.025.

ВИСНОВКИ

У даній дипломній роботі було досліджено методи аналізу кліматичних даних та визначено силу впливу різних факторів на зміну середньої температури поверхні землі.

В рамках дипломного проектування:

а) розглянуто основні підходи до аналізу кліматичних даних в рамках поставленої задачі: регресійні моделі, факторний аналіз, моделі на основі ланцюгів Маркова та нейромережеві моделі. У результаті проведеного порівняльного аналізу було обрано регресійну модель, побудова якої відбувається на основі алгоритмів машинного навчання;

б) порівняно три алгоритми машинного навчання: лінійна регресія, Random Forest та SVR. При побудові моделі за допомогою лінійної регресії було знайдено квадрат середньоквадратичної помилки, який дорівнює 0.341. При використанні алгоритму SVR помилка склала 0.338. Було визначено, що найкращим алгоритмом машинного навчання для представлених у даній роботі кліматичних даних є метод Random Forest. Квадрат середньоквадратичної помилки даного алгоритму на тестових даних складає 0.227;

в) за допомогою моделі, побудованої на основі Random Forest, було оцінено показники важливостей факторів впливу. У досліджені було показано, що метан домінує серед інших парникових газів у впливі на показник середньої температури поверхні землі. Проте два інших парникових гази також мають високий вплив, на відміну від природних факторів впливу. Можна сказати, що при проведенні дослідження з іншими значеннями факторів впливу, з великою ймовірністю Random Forest буде алгоритмом, який з найменшою помилкою описує систему прогнозування зміни середньої температури поверхні землі;

г) отримані результати аналізу кліматичних даних продемонстровано на розробленому сайті у вигляді інформаційної панелі.

ПЕРЕЛІК ПОСИЛАНЬ

1. Reanalysis of Historical Climate Data for Key Atmospheric Features: Implications for Attribution of Causes of Observed Change / Randall Dole, Martin Hoerling, and Siegfried Schubert // A Report by the U.S. CCSP and Subcommittee on Global Change Research — 2008 — P. 156.
2. Плазменная гелиогеофизика. / Под ред. Л.М. Зеленого, И. С, Веселовского. // М. Физматлит, — 2008. — С. 672.
3. Statistical methods for the analysis of simulated and observed climate data // The working group statistics at the Climate Service Center. — Hamburg, 2013. — P. 135.
4. Методи аналізу даних: навчальний посібник для студентів / В. Є. Бахрушин. — Запоріжжя: КПУ, 2011 — 268 с.
5. Статистические методы анализа гидрометеорологической информации / Малинин В // СПБ.: РГГМУ. — 2008. — 408 с.
6. Applied regression analysis: a research tool. — 2nd ed. / John O. Rawlings, Sastry G. Pentula, David A. Dickey — 1932.
7. Прикладная математическая статистика. Для инженеров и научных работников. — М.: ФИЗМАТЛИТ, 2006. — 816 с.
8. Бирюков А.Н. Нечеткая регрессионная прогнозная многофакторная модель для решения экономической прикладной задачи // Управление экономическими системами: электронный научный журнал — 2010.
9. Иберла К. Факторный анализ \ Пер. с нем. В. М. Ивановой; Предисл. А. М. Дуброва. — М.: Статистика, 1980. — 398 с.
10. Yingjian Z. Prediction of Financial Time Series with Hidden Markov Models, Simon Fraiser University, Burnaby, 2004
11. И. В. Заенцев, Нейронные сети: основные модели, 1999. — 76 с.
12. Хайкин, Саймон. Нейронные сети: полный курс, 2-е издание, пер. с англ. / С. Хайкин. — М. Издательский дом "Вильяме".— 2006. — 1104 с.

13. Гайдышев И. Анализ и обработка данных: специальный справочник / И. Гайдышев. — СПб: Питер, 2001. — 752 с.
14. Wang, W.-C. Assessment of River Water Quality Based on Theory of Variable Fuzzy Sets and Fuzzy Binary Comparison Method. Water Resources Management, [Electronic Resourse]. — Mode of Access: https://doi.org/10.1007/s11269-014-0738-4
15. United States Environmental Protection Agency (2016). Global Atmospheric Concentrations of Carbon Dioxide over Time [Electronic Resourse]. — Mode of Access: https://www.epa.gov/sites/production/files/2016-08/ghg-concentrations_fig-1.csv
16. United States Environmental Protection Agency (2016) Global Atmospheric Concentrations of Methane over Time [Electronic Resourse]. — Mode of Access: https://www.epa.gov/sites/production/files/2016-08/ghg-concentrations_fig-2.csv
17. United States Environmental Protection Agency (2016) Global Atmospheric Concentrations of Nitrous Oxide over Time Time [Electronic Resourse]. — Mode of Access: https://www.epa.gov/sites/production/files/2016-08/ghg-concentrations_fig-3.csv
18. National Centeres For Environmental Information/ Dataset. Solar Radiation [Electronic Resourse]. — Mode of Access: https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/solar-radiation
19. NCEI/WDS Global Significant Volcanic Eruptions Database, 4360 BC to Present [Electronic Resourse]. — Mode of Access https://catalog.data.gov/dataset/global-significant-volcanic-eruptions-database-4360-bc-to-present

Додаток А
Лістинги програм

Лістинг файлу app.py — Реалізація веб-застосунку

```
from flask import Flask, render_template, url_for, send_file
from matplotlib.backends.backend_agg import FigureCanvasAgg as FigureCanvas
from io import BytesIO
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt


from visual import *
from dataperfom import *
from analizedata import *


app = Flask(__name__)


temp_df = load_temp_data()
co2_df = load_carbon_data()


temp_co2_df = join_dataset(temp_df, co2_df)
co2_lin_pred, carbon_linear_err, l_coef, l_r2 = carbon_linear_regg(temp_co2_df)
co2_poly_pred, carbon_poly_err, p_corf, p_r2 = carbon_poly_regr(temp_co2_df)


ch4_df = load_methan_data()
temp_methan_data = join_dataset(temp_df, ch4_df)


general_gases_df = join_dataset(temp_co2_df, ch4_df)


ml_regr, ml_coef, ml_r2 = multu_linear_regr(general_gases_df)


period_regr, period_coef = period_regr(temp_co2_df)



@app.route('/visual')
def temp_plot():
  img = do_temp_plot(temp_df.index, temp_df['temp'])
  return send_file(img, mimetype='temp.png')



# @app.route('/')
# def index():
#     return render_template("general.html")



@app.route('/')
def main():
  return render_template('general.html')


@app.route('/gases_corr_')
def corr_gases_img():
  img4 = corr_matrix(general_gases_df)
  return send_file(img4, mimetype='correlation.png')


@app.route('/ml_reg_img')
def linear_gases_img():
  img7 = do_3d_plot(general_gases_df ,ml_regr)
  return send_file(img7, mimetype='multiplulinearregr.png')


@app.route('/greenhouse_gases')
def greenhouse_gases():
  return render_template('atropogenic.html', ml_coef=ml_coef, ml_r2=ml_r2)
```

```
@app.route('/nature')
def nature():
  return render_template('nature.html')


@app.route('/carbon_plot')
def co2_plot():
  img = carbon_plot(co2_df.index, co2_df['co2'])
  return send_file(img, mimetype='carbon.png')


@app.route('/carbon_plot_1')
def co2_plot1():
  img1 = temp_carbon_plot(temp_co2_df['co2'], temp_co2_df['temp'])
  return send_file(img1, mimetype='carbon1.png')


@app.route('/l_carbon_plot')
def linear_carbon_plot():
  img2 = linear_plot(temp_co2_df['co2'], temp_co2_df['temp'], co2_lin_pred)
  return send_file(img2, mimetype='carbon2.png')


@app.route('/poly_carbon_plot')
def poly_carbon_plot():
  img3 = linear_plot(temp_co2_df['co2'], temp_co2_df['temp'], co2_poly_pred)
  return send_file(img3, mimetype='carbon2.png')


@app.route('/period_plot')
def period_carbon_plot():
  img = period_plot(temp_co2_df['co2'], temp_co2_df['temp'], period_regr)
  return send_file(img, mimetype='carbon3.png')


@app.route('/greenhouse_gases/cardondioxide')
def carbon():
  corr = coef_correlation(temp_co2_df['co2'], temp_co2_df['temp'])
  return render_template('co2.html', corr = corr, error_l=carbon_linear_err, error_p=carbon_poly_err, l_coef=l_coef, l_r2=l_r2, p_coef=p_corf, p_r2=p_r2)


if __name__ == '__main__':
  app.run(debug=True)
```

Лістинг файлу dataanalize.py — Побудова регресійних моделей та обчислення помилок

```
from io import BytesIO

from scipy.stats import pearsonr, spearmanr
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures
from sklearn import metrics
from sklearn.metrics import mean_squared_error, r2_score
import numpy as np
from scipy import optimize
from mpl_toolkits.mplot3d import Axes3D
import matplotlib.pyplot as plt


def coef_correlation(x, y):
  pear, _ = pearsonr(x, y)
  pear = round(pear, 3)


  sper, _ = spearmanr(x, y)
  sper = round(sper, 3)
  corr = {'pearson': pear, 'spearman': sper}
  return corr


def count_error(test, pred):
  mean_absolute = round(metrics.mean_absolute_error (test, pred)*100, 3)
  mean_squared = round(metrics.mean_squared_error (test, pred)*100 , 3)
  root_mean_squared = round(np.sqrt (metrics.mean_squared_error (test, pred)),3)
  error = {'absolute': mean_absolute,
       'squared': mean_squared,
       'root_squared': root_mean_squared}
  return error
```

```
def carbon_linear_regg(df):
    X = df['co2'].values.reshape(-1, 1)
    y = df['temp'].values.reshape(-1, 1)
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)

    regr = LinearRegression()
    regr.fit(X_train, y_train)

    y_pred = regr.predict(X_test)
    co2_pred = regr.predict(X)

    err = count_error(y_test, y_pred)

    slope = np.round(regr.coef_, 5)
    intercept = np.round(regr.intercept_, 5)

    coef = {'slope': slope,
            'intercept': intercept}

    r2 = np.round(regr.score(X, y), 4)



    return co2_pred, err, coef, r2



def carbon_poly_regr(df):
    X = df['co2'].values.reshape(-1, 1)
    y = df['temp'].values.reshape(-1, 1)

    polynomial_features = PolynomialFeatures(degree=4)
    X_poly = polynomial_features.fit_transform(X)
    poly_reg = LinearRegression()
    poly_reg.fit(X_poly, y)

    co2_poly_pred = poly_reg.predict(X_poly)

    err = count_error(y, co2_poly_pred)
    r2 = r2_score(y, co2_poly_pred)

    slope = np.round(poly_reg.coef_, 4)
    intercept = np.round(poly_reg.intercept_, 4)

    coef = {'slope': slope,
            'intercept': intercept}

    return co2_poly_pred, err, coef, r2

def period_regr(df):
    X = df['co2'].values
    y = df['temp'].values

    def test_func(x, a, b):
        return a * np.sin(b * x)

    params, params_covariance = optimize.curve_fit(test_func, X, y, p0=[2, 2])

    y_pred = np.array(test_func(X, params[0], params[1]))

    return y_pred, params
```

```
def multu_linear_regr(general0):

    X = general0[['co2', 'ch4']].values.reshape(-1, 2)
    Y = general0['temp']



    x_pred = np.linspace(330, 410, 30)
    y_pred = np.linspace(1600, 1800, 30)
    xx_pred, yy_pred = np.meshgrid(x_pred, y_pred)
    model_viz = np.array([xx_pred.flatten(), yy_pred.flatten()]).T

    # Train
    ols = LinearRegression()
    model = ols.fit(X, Y)
    predicted = model.predict(model_viz)

    r2 = np.round(model.score(X, Y), 4)

    slope = np.round(model.coef_, 5)
    intercept = np.round(model.intercept_, 5)

    coef = {'slope': slope,
            'intercept': intercept}

    return predicted, coef, r2
```

Лістинг файлу visual.py — Візуалізація даних

```
from io import BytesIO

import numpy as np
import matplotlib.pyplot as plt
from matplotlib.backends.backend_agg import FigureCanvasAgg as FigureCanvas
import seaborn as sns
from mpl_toolkits.mplot3d import Axes3D

from dataperfom import *



def do_temp_plot(x, y):

    fig, ax = plt.subplots()
    plt.plot(x, y, color='blue')
    plt.xlabel('Рік')
    plt.ylabel('Середня температура')
    plt.grid(True)

    canvas = FigureCanvas(fig)

    img = BytesIO()
    fig.savefig(img)
    img.seek(0)

    return img

def carbon_plot(x, y):

    fig, ax = plt.subplots()

    plt.plot(x, y)
    plt.xlabel('Рік')
    plt.ylabel('Діоксид вуглецю, ppm')
    plt.grid(True)

    img = BytesIO()
    fig.savefig(img)
    img.seek(0)
```

```
    return img

def temp_carbon_plot(x, y):
    fig, ax = plt.subplots()



    plt.scatter(x, y)
    plt.title('Temp and CO2')
    plt.xlabel('CO2, ppm')
    plt.ylabel('Середня Температура')
    plt.grid(True)

    img = BytesIO()
    fig.savefig(img)
    img.seek(0)

    return img

def linear_plot(X, y, pred):
    fig = plt.figure()

    plt.scatter(X, y, color='gray')
    plt.plot(X, pred, color='red', linewidth=2)
    plt.grid(True)

    img = BytesIO()
    fig.savefig(img)
    img.seek(0)

    return img

def period_plot(X, y, pred):
    fig = plt.figure()

    plt.scatter(X, y, color='gray')
    plt.plot(X, pred, color='red', linewidth=2)

    plt.legend(loc='best')
    plt.grid(True)

    img = BytesIO()
    fig.savefig(img)
    img.seek(0)

    return img

def corr_matrix(df):

    fig = plt.figure(dpi=75)

    ax = fig.add_axes([0.1, 0.1, 0.75, 0.8])
    sns.heatmap(df.corr(), annot=True, cmap="YlGnBu")
    ax.set_title("Кореляція між Парниковими Газами та Середньою Температурою", fontsize=10)

    img = BytesIO()
    fig.savefig(img)
    img.seek(0)

    return img




def do_3d_plot(df, predicted):
    x_pred = np.linspace(330, 410, 30)
```

```
    y_pred = np.linspace(1600, 1800, 30)
    xx_pred, yy_pred = np.meshgrid(x_pred, y_pred)

    X = df[['co2', 'ch4']].values.reshape(-1, 2)
    Y = df['temp']

    x = X[:, 0]
    y = X[:, 1]
    z = Y

    fig = plt.figure()

    ax1 = fig.add_subplot(121, projection='3d')
    ax2 = fig.add_subplot(122, projection='3d')

    axes = [ax1, ax2]

    for ax in axes:
        ax.scatter(x, y, z, color='r')
        ax.plot_trisurf(xx_pred.flatten(), yy_pred.flatten(), predicted, linewidth=0.2, antialiased=True)
        ax.set_xlabel('Carbon Dioxid (ppm)', fontsize=6)
        ax.set_ylabel('Methan', fontsize=6)
        ax.set_zlabel('Temperature (C°)', fontsize=6)

    ax1.view_init(elev=60, azim=165)
    ax2.view_init(elev=4, azim=114)

    img = BytesIO()
    fig.savefig(img)
    img.seek(0)

    return img
```

Лістинг файлу cleaning.py — Обробка даних

```
import pandas as pd



def load_temp_data():
    temp_df = pd.read_csv('data/annual_temp.csv')

    temp_df = temp_df[temp_df.Source != 'GISTEMP']
    temp_df = temp_df.drop(columns=['Source'], axis=1)

    temp_df.index = temp_df['Year']
    del temp_df['Year']

    temp_df = temp_df.dropna()
    temp_df = temp_df.reindex(index=temp_df.index[::-1])

    temp_df = temp_df.rename({"Mean": "temp"}, axis='columns')
    temp_df = temp_df.rename_axis(index={'Year': 'year'})

    return temp_df

def load_carbon_data():
    co2 = pd.read_csv('data/co2-concentrations.csv', header=7)

    co2 = co2.drop(
        {'Unnamed: 1', 'Unnamed: 2', 'Unnamed: 3', 'Unnamed: 5', 'Unnamed: 6', 'Unnamed: 7', 'Unnamed: 8', 'Unnamed: 9',
         'Unnamed: 10'}, 1)
    co2 = co2.dropna()

    co2[['Ice Core Measurements']] = co2[['Ice Core Measurements']].apply(pd.to_numeric)
```

```
    co2.index = co2['Ice Core Measurements']
    del co2['Ice Core Measurements']

    co2 = co2.rename({"Unnamed: 4": "co2"}, axis='columns')
    co2 = co2.rename_axis(index={'Ice Core Measurements': 'year'})

    return co2

def load_methan_data():
    ch4 = pd.read_csv('data/ch4-concentrations.csv', header=7)

    ch4 = ch4.drop({'Unnamed: 1', 'Unnamed: 2', 'Unnamed: 4', 'Unnamed: 5'}, 1)
    ch4 = ch4.dropna()

    ch4[['Ice Core Measurements']] = ch4[['Ice Core Measurements']].apply(pd.to_numeric)

    ch4.index = ch4['Ice Core Measurements']
    del ch4['Ice Core Measurements']

    ch4 = ch4.rename({"Unnamed: 3": "ch4"}, axis='columns')
    ch4 = ch4.rename_axis(index={'Ice Core Measurements': 'year'})

    return ch4


def join_dataset(x, y):
    return x.join(y, how='inner')
```

Додаток Б
Ілюстративний матеріал

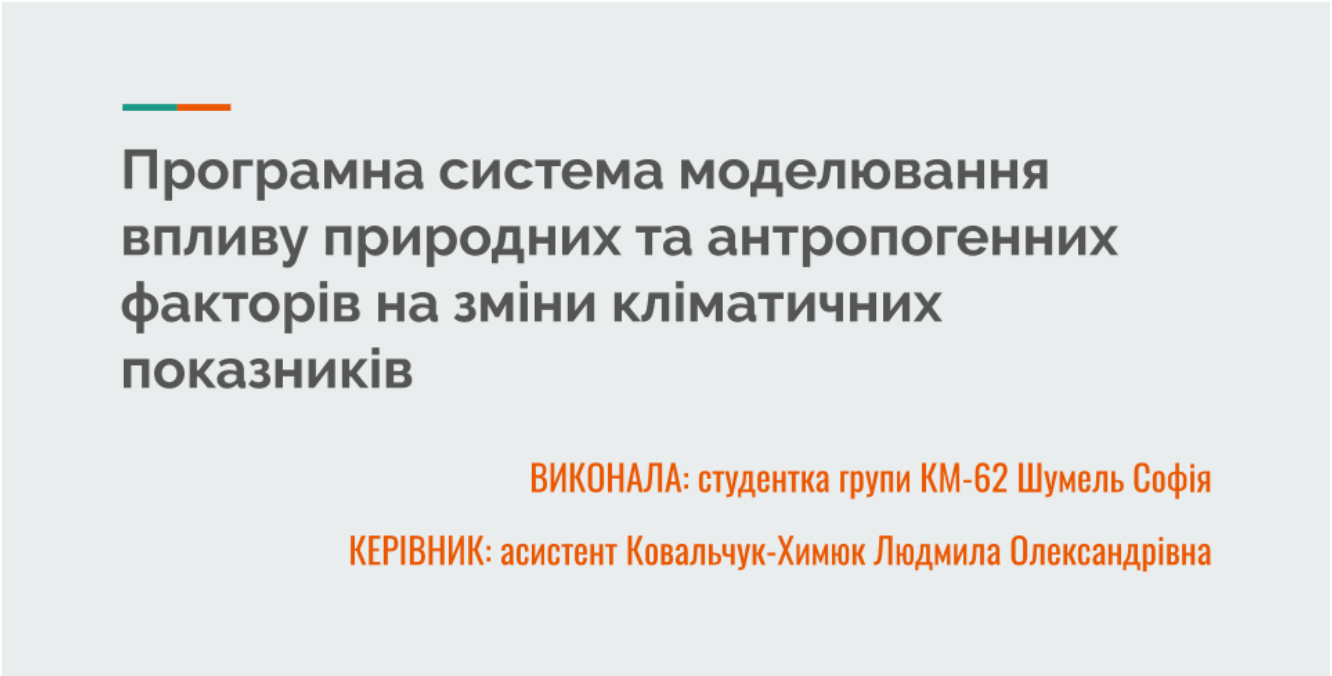


Рисунок Б.1 – Слайд 1

Рисунок Б.2 – Слайд 2

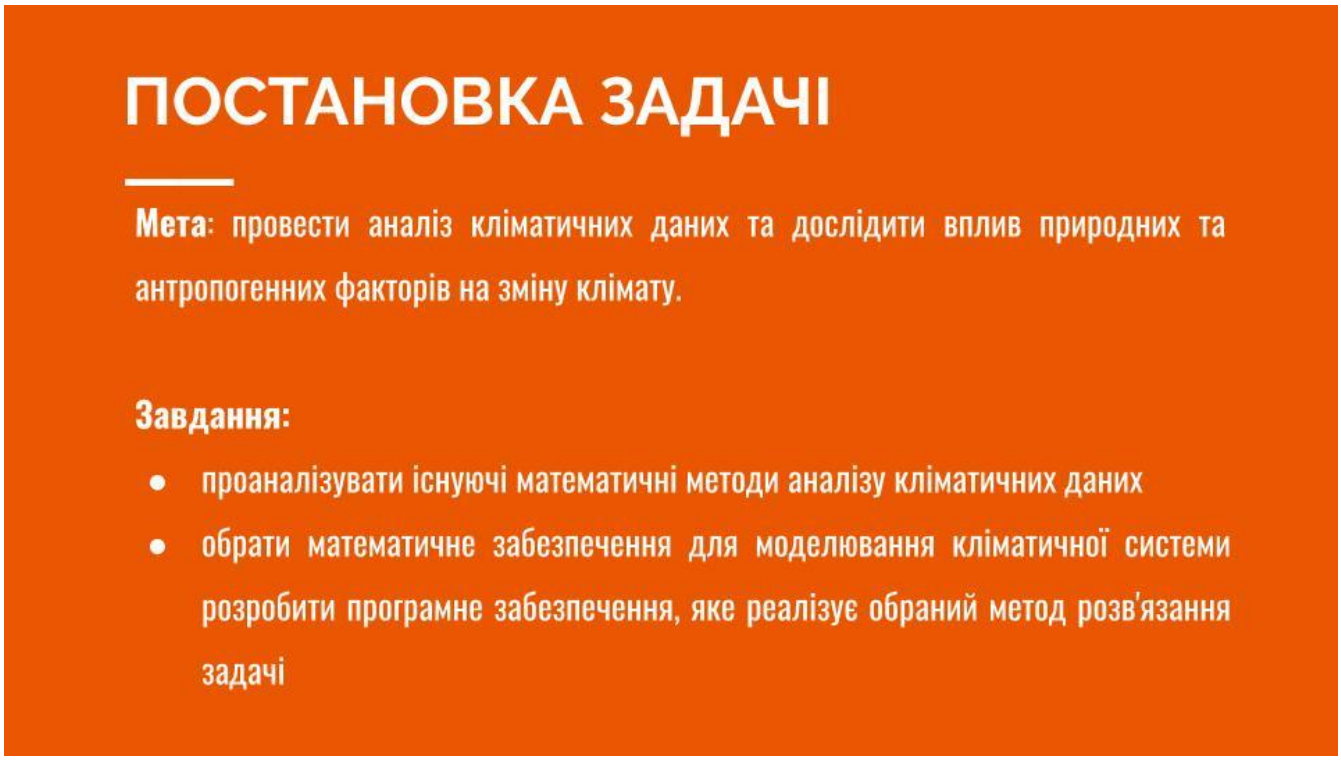


Рисунок Б.3 – Слайд 3

МЕТОДИ

Регресійний аналіз

Факторний аналіз

Моделі на основі ланцюгів Маркова

Нейромережеві моделі

Рисунок Б.4 – Слайд 4

ПОРІВНЯННЯ МЕТОДІВ

МЕТОД	ПЕРЕВАГИ	НЕДОЛІКИ
Регресійний аналіз	Простота та гнучкість, можливість розширення функціоналу поза області відомих значень, доступність до проміжних обчислень.	Велика кількість вимог до вхідних даних, проектування лише лінійних залежностей.
Факторний аналіз	Узагальнення вхідної системи з можливістю дослідження спрощеної моделі.	Відсутність однозначного математичного розв'язку, висока складність обчислень.
Марковські процеси	Легко моделюється та проектується.	Використання лише дискретних даних та неможливість моделювання вибірок з довгою пам'яттю.
Нейронні мережі	Відсутність обмеження на лінійність системи, здатність швидкої адаптації до навколишнього середовища.	Відсутність доступу до проміжних обчислень, складність проектування.

Рисунок Б.5 – Слайд 5

ЗАДАЧА РЕГРЕСІЇ

Нехай існує множина об'єктів:

$$I=\{i1,i2,...,ij,...,in\},$$

кожен елемент якої є значенням середньої температури поверхні землі та характеризується набором змінних:

$$Ij=\{x1,x2,...,xh,...,xm,y\},$$

де xh — значення h-фактору, який впливає на середню температуру поверхні землі.

Необхідно визначити:

$$min\ R(f) = \frac{1}{m}\sum_{i=1}^{m} c(y_i, f(x_i))$$

де F — множина всіх можливих функцій, які описують дані, c(yi,f(xi)) — функція втрат.

Рисунок Б.6 – Слайд 6

МЕТОДИ МАШИННОГО НАВЧАННЯ

Лінійна регресія

Random Forest

SVR

Рисунок Б.7 – Слайд 7

ОЦІНКА АДЕКВАТНОСТІ МОДЕЛЕЙ

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y_i})^2$$

де N — загальна кількість спостережень, уі — фактичні дані, уі— передбачувані дані.

$$RMSE = \sqrt{MSE}$$

Рисунок Б.8 – Слайд 8

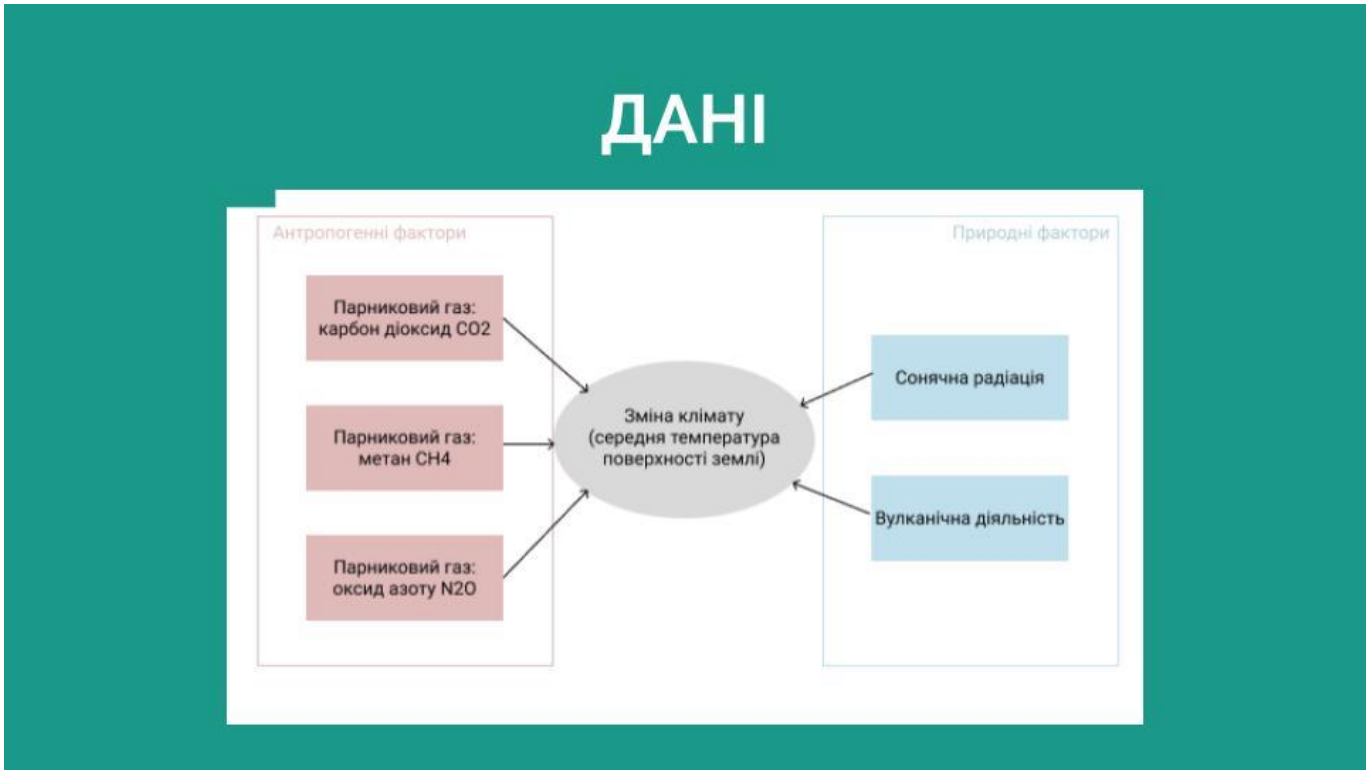


Рисунок Б.9 – Слайд 9

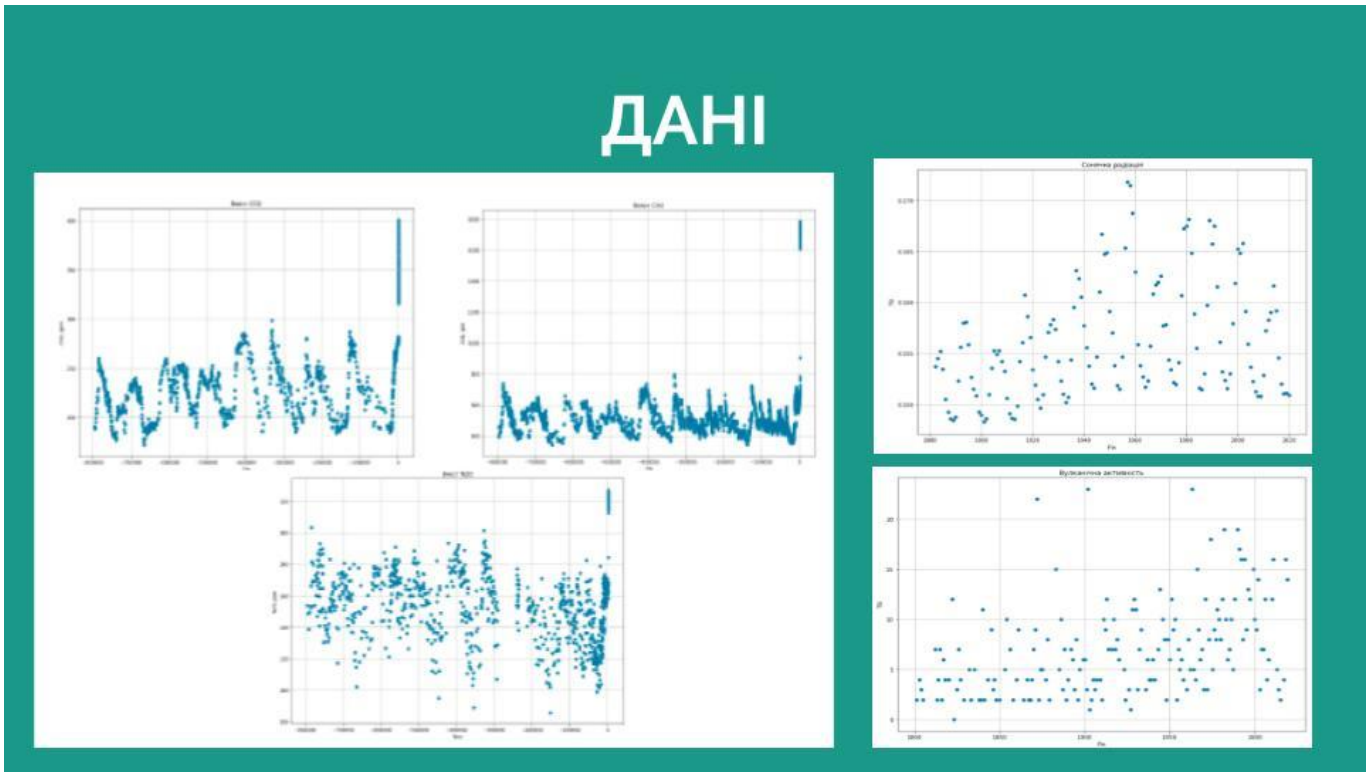


Рисунок Б.10 – Слайд 10

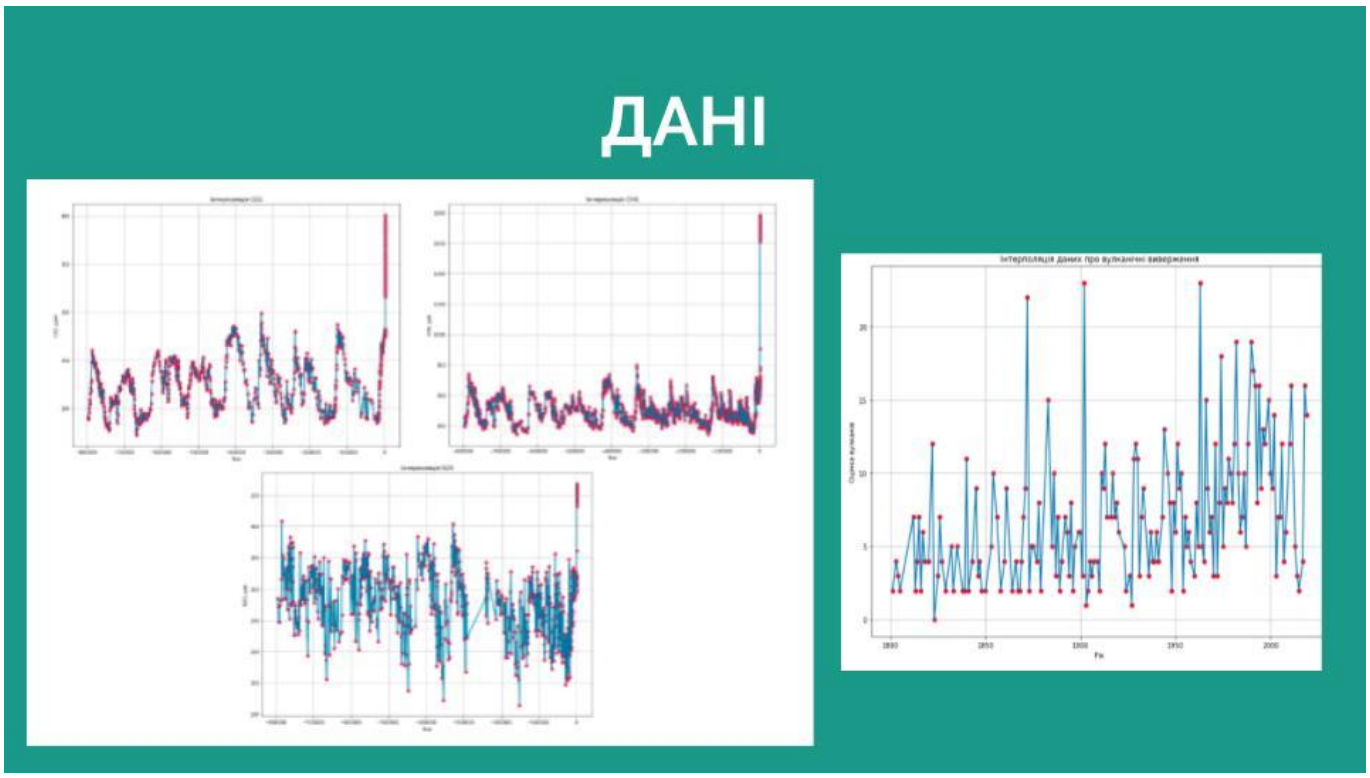


Рисунок Б.11 – Слайд 11

Лінійна регресія

$$y = \omega_0 + \omega_1 x_1 + ... + \omega_n x_n = \omega_0 + \sum \omega_j x_j$$

де ω_0, $\omega_1, ..., \omega_n$ — коефіцієнти при незалежних змінних.

$$min\ R(f) = min \frac{1}{m} \sum_{i=1}^{m} (y_i - \sum_{i=1}^{n} \omega_i f_i(x))^2$$

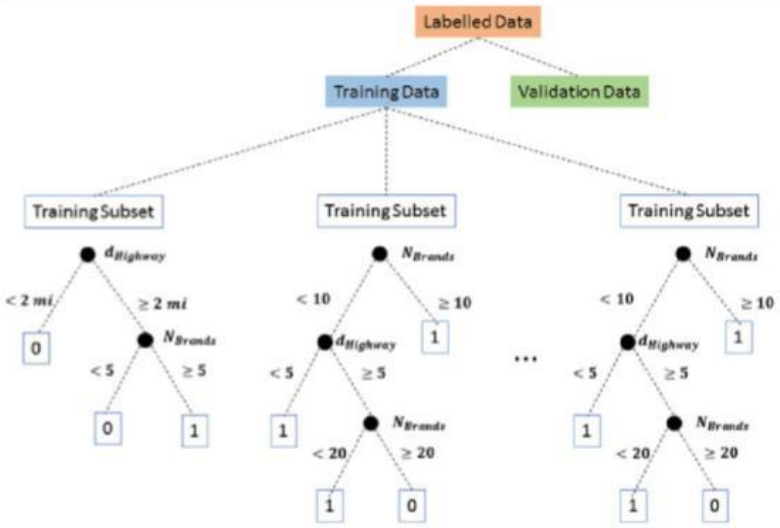
$$Y = 0.369X_1 + 1.203X_2 + 1.699X_3 + 0.031X_4 + 0.024X_5 - 0.005$$

де Y — вектор прогнозованих середніх температур, X_1 — вектор концентрації CO_2, X_2 — вектор концентрації CH_4, X_3 — вектор концентрації N_2O, X_4 — вектор показників VEI вулканічної активності, X_5 — вектор значення сонячної радіації.

$$RMSE = 0.341$$

Рисунок Б.12 – Слайд 12

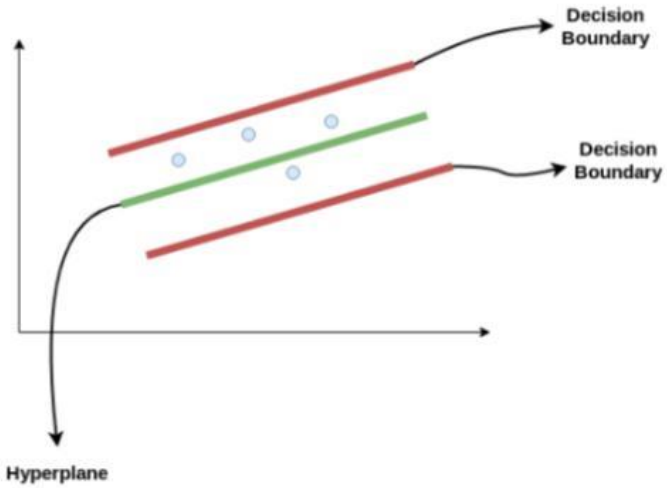
Random Forest



$$RMSE = 0.277$$

Рисунок Б.13 – Слайд 13

SVR



$$RMSE = 0.338$$

Рисунок Б.14 – Слайд 14

ПОРІВНЯННЯ АЛГОРИТМІВ

АЛГОРИТМ	НАВЧАЛЬНЕ RMSE	ТЕСТОВЕ RMSE
Лінійна регресія	0.404	0.341
Random Forest	0.154	0.227
SVR	0.153	0.338

Рисунок Б.15 – Слайд 15

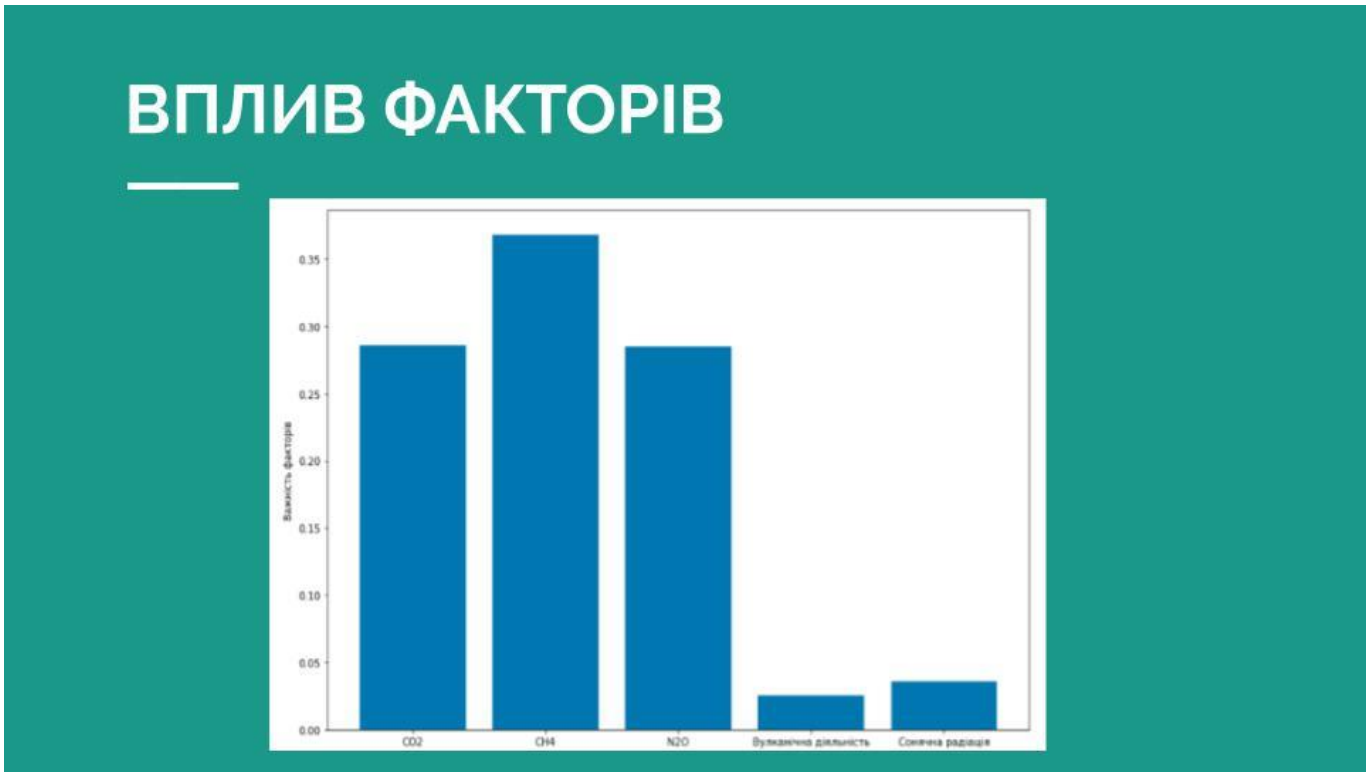


Рисунок Б.16 – Слайд 16

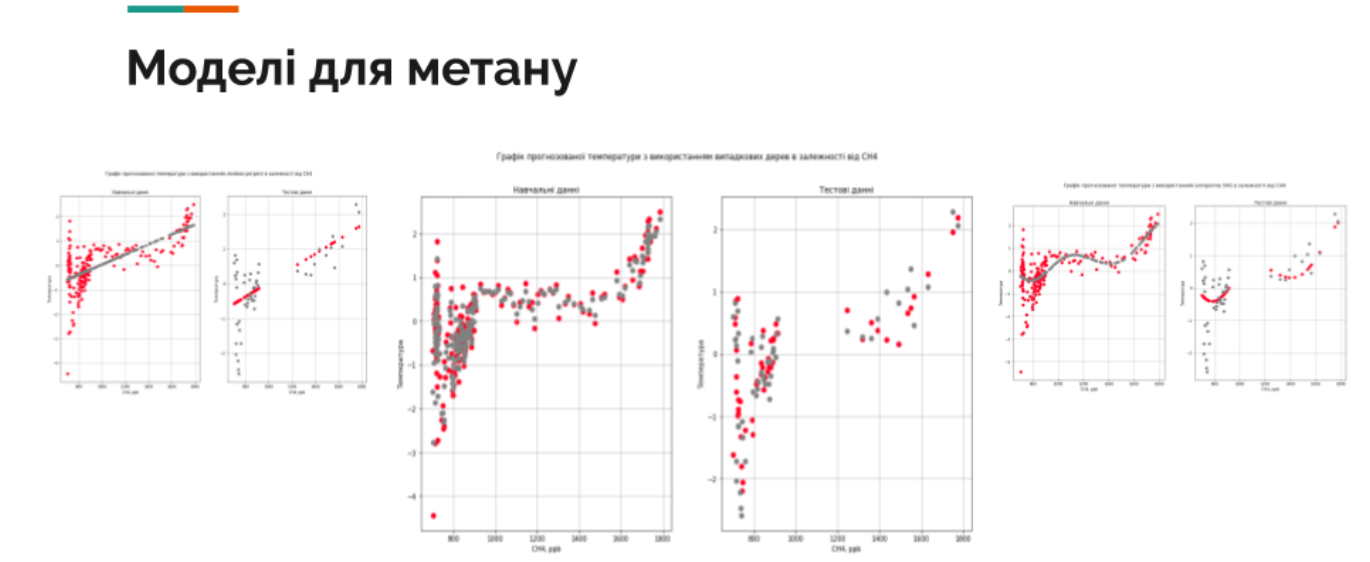


Рисунок Б.17 – Слайд 17

Моделі для карбон діоксиду

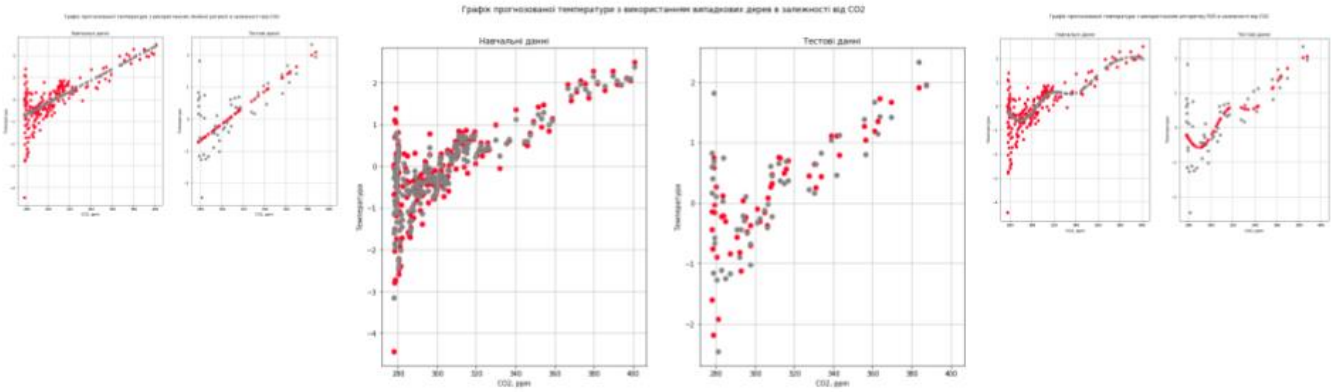


Рисунок Б.18 – Слайд 18

Моделі для оксид азоту

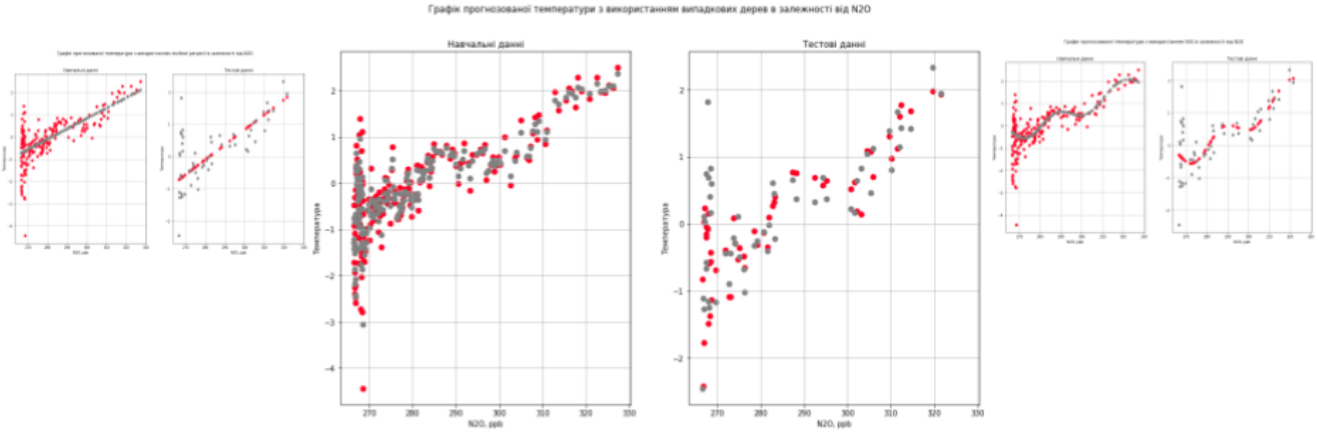


Рисунок Б.19 – Слайд 19

ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ

- Python
- Jupyter Notebook
- Flask
- Bootstrap3
- Jinja2

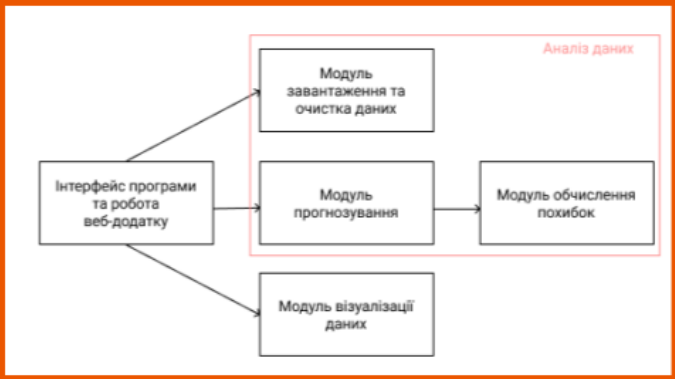


Рисунок Б.20 – Слайд 20

ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ

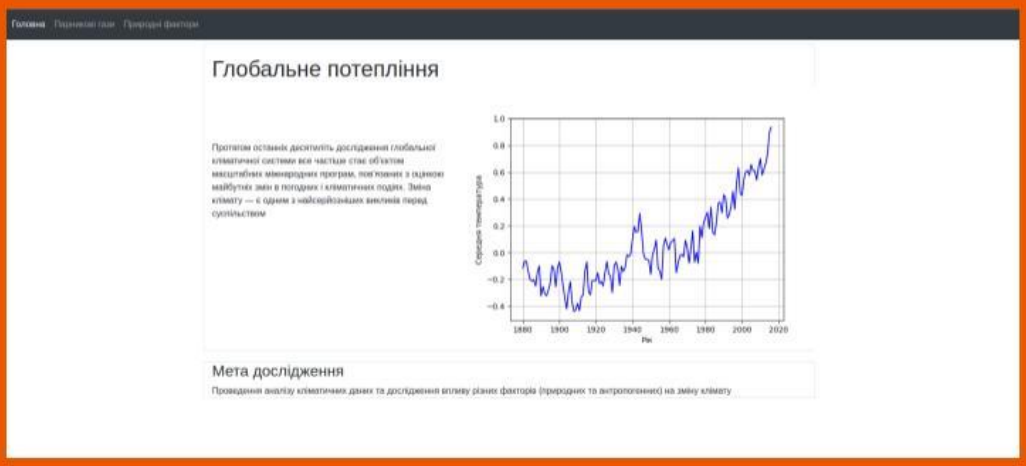


Рисунок Б.21 – Слайд 21

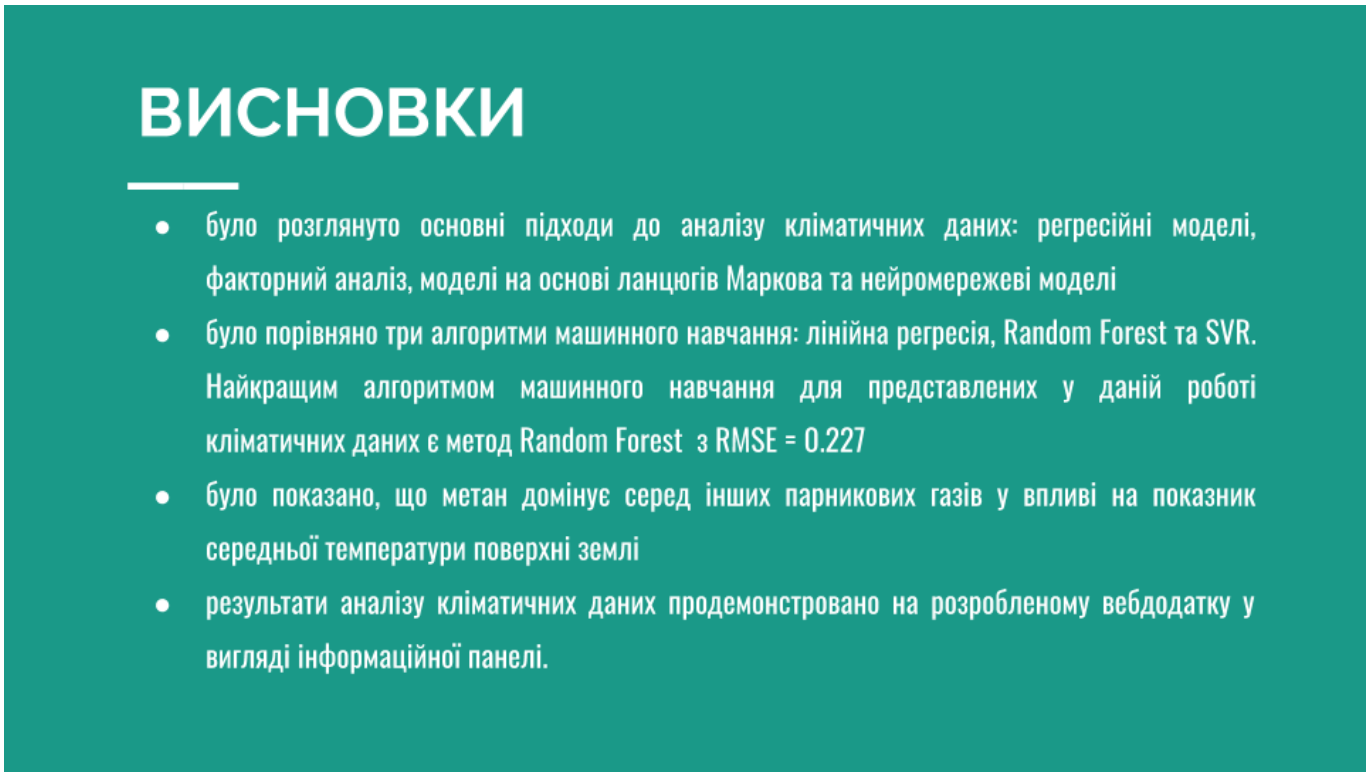


Рисунок Б.22 – Слайд 22