КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ТАРАСА ШЕВЧЕНКА

Факультет інформаційних технологій
Кафедра технологій управління

Спеціальність: 122 «Комп'ютерні науки»
Освітня програма: «Інформаційна аналітика та впливи»

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

на тему:

"Інформаційна аналітика та прогнозування кліматичних змін"

Студентки 2-го курсу групи ІАВ-21

Дуля Дарина Василівна
(прізвище, ім'я, по батькові)

(підпис студента)

Науковий керівник:

доктор технічних наук, професор
(науковий ступінь, вчене звання)

Хлевна Юлія Леонідівна
(прізвище, ім'я, по батькові)

(дата) (підпис)

Попередній захист:

(Висновок: «До захисту в Екзаменаційній комісії»)

Завідувач кафедри
технологій управління

(підпис) (прізвище, ініціали) (дата)

Київ – 2023

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА
Факультет інформаційних технологій

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

ЗАТВЕРДЖУЮ
Завідувач кафедри
професор Морозов В.В.

_____
«08» грудня 2023 року

З А В Д А Н Н Я
НА ВИКОНАННЯ КВАЛІФІКАЦІЙНОЇ РОБОТИ

Студентка Дуля Дарина Василівна

Група___IAB-21_____

1. Тема кваліфікаційної роботи
Інформаційна аналітика та прогнозування кліматичних змін_____

Затверджена наказом по від «08» грудня 2023 р. № 5
2. Строк подання студентом готової роботи – "16" травня 2023 р.

3. Цільова установка та вихідні дані до роботи: дослідження теоретичної бази сфери кліматичних змін та застосування інформаційної аналітики у даній сфері; аналіз методів та методик аналізу та моделювання процесів, включно з програмними засобами реалізації; підбір та стандартизація даних дослідження, побудова прогностичної моделі та її оцінка; проведення інформаційної розвідки щодо потенційної імплементації отриманої моделі на підприємствах.

4. Зміст роботи: аналіз моніторингу кліматичних змін, інформаційна аналітика у сфері кліматичних змін, методології управління проєктами з інформаційної аналітики даних, методи інформаційного аналізу та прогнозування даних, програмні засоби для моделювання, вибір даних для аналізу кліматичних змін, обробка даних, побудова моделі методом штучних нейронних мереж, побудова моделі методом рекурентної нейромережі, порівняння моделей, застосування розробленої моделі на підприємствах екологічно-моніторингового напряму, застосування розробленої моделі на підприємствах комерційного спрямування, потенціал розвитку проєкту.

5. Перелік графічного матеріалу (слайдів) Інфографіка викидів вуглекислого газу, Середній показник об'єму льодовиків у найгарячішу пору року з 1979 по 2019 роки, Тягар захворювань за фактором ризику (2019 рік, Схема життєвого циклу процесу дослідження даних за методологією CRISP-DM, Схема будови штучної нейронної мережі, Загальна схема роботи LSTM, Показники місячного значення температурних аномалій, Глобальні середньомісячні показники вуглекислого газу, Річний приріст усередненого глобального вмісту метану в атмосфері, Глобальні середньомісячні показники оксиду азоту, Річний приріст усередненого глобального вмісту SF6 в атмосфері, Температурний графік кореляції,

Результуючий графік прогнозу рівню світового океану, Результуючий графік прогнозу рівню світового океану LSTM моделі.

6. Календарний план виконання роботи:

№ з/п	Назва частин роботи	%	Виконання роботи	
			За планом	Фактично
1	Вибір теми дипломної роботи	3	01.10.2022	01.10.2022
2	Протокол кафедри ТУ про затвердження тем дипломних робіт та призначення наукових керівників	2	08.12.2022	08.12.2022
3	Формування переліку нормативних матеріалів, літератури з проблематики дипломної роботи	10	08.01.2023	08.01.2023
4	Складання розгорнутого плану кваліфікаційної роботи	5	18.01.2023	18.01.2023
5	Ознайомлення наукового керівника з розгорнутим планом кваліфікаційної роботи. Внесення змін.	5	20.01.2023	20.01.2023
6	Підготовка розділу 1 “Аналіз теоретичних засад використання інформаційної аналітики у сфері кліматичних змін”	10	13.02.2023	13.02.2023
7	Підготовка розділу 2 “Аналіз методів та методологій інформаційного аналізу та моделювання”	14	06.03.2023	06.03.2023
8	Підготовка розділу 3 “Аналіз та прогнозування кліматичних змін”	14	03.04.2023	03.04.2023
9	Підготовка розділу 4 “Технологія застосування розроблених моделей на діючих підприємствах”	13	17.04.2023	17.04.2023
10	Оформлення кваліфікаційної роботи. Підготовка висновків і пропозицій	15	01.05.2023	01.05.2023
11	Передача кваліфікаційної роботи науковому керівникові	2	02.05.2023	02.05.2023
12	Передача кваліфікаційної роботи рецензенту для рецензування	2	10.05.2023	10.05.2023
13	Попередній захист кваліфікаційної роботи	2	17.05.2023	17.05.2023

Дата видачі завдання «08» грудня 2023 р.

Керівник роботи _д.т.н., професор Хлевна Юлія Леонідівна_____
(посада, прізвище, ім’я, по батькові)

_____
(підпис)

Завдання прийняв до виконання студент групи _ІАВ-21____

___Дуля Дарина Василівна_____
(прізвище, ім’я, по батькові)

_____
(підпис)

ЗМІСТ

АНОТАЦІЯ

Київський національний університет імені Тараса Шевченка
Факультет інформаційних технологій
Кафедра технологій управління
Спеціальність 122 – Комп'ютерні науки,
освітня програма «Інформаційна аналітика та впливи»

Дипломна робота магістрантки Дулі Дарини Василівни.

Тема роботи – «Інформаційна аналітика та прогнозування кліматичних змін».

Мета дипломної роботи магістра – проведення аналізу та пошук рішення щодо підвищення якості моніторингових процесів зміни рівня світового океану.

Об'єкт дослідження. Кліматичні процеси, що призводять до зміни рівня світового океану.

Предмет дослідження. Аналіз та прогнозування зміни рівня світового океану та розробка прогностичної моделі для моніторингу та прогнозування цих змін.

Наукова новизна та практичне значення одержаних результатів полягає у розроблені діючої, ефективної прогностичної моделі змін рівня світового океану, що може бути імплементована в більш складні інформаційні системи для збільшення ефективності моніторингу кліматичних змін.

У роботі досліджуються існуючі підходи до використання нейронних мереж у задачах аналізу та прогнозування кліматичних змін, а саме впливу цих змін на рівень світового океану. Розробляється нова прогностична модель для аналізу змін рівня світового океану, що може бути використано в широкому спектрі наукових та практичних задач. Наводяться можливі варіанти імплементації отриманої прогностичної моделі в компаніях різного роду організаційного підпорядкування.

Дипломна робота складається зі вступу, основної частини, що включає 4 розділи, висновків та списку використаних джерел. Всього налічує 88 сторінок та перелік з 62 джерела на 7 сторінках.

Ключові слова: Data Science, Big Data, інформаційна аналітика даних, кліматичні зміни, світовий океан, штучні нейронні мережі.

ВСТУП

Актуальність теми дослідження полягає у тому, що сучасний світ стикається зі складними викликами, пов'язаними зі зміною клімату. Ці виклики виникають внаслідок декількох факторів. По-перше, швидка індустріалізація країн, які до недавнього часу були переважно сільськогосподарськими, призводить до збільшення викидів шкідливих речовин у атмосферу. По-друге, геометричне зростання населення планети призводить до збільшення споживання ресурсів та виробництва товарів, що негативно впливає на довкілля. Крім того, культура надмірного споживання, яка просуває ідею безконтрольного накопичення матеріальних благ, призводить до надлишкового виробництва товарів, продуктів та послуг, багато з яких в результаті стають непотрібними і відбраковуються.

Однією з найважливіших, надзвичайно складних до аналізу та загрозливих змін є коливання рівня світового океану.

Метою та завданням дослідження є проведення аналізу та пошук рішення щодо підвищення якості моніторингових процесів зміни рівня світового океану. В контексті дослідження необхідно виконати наступні кроки:

- проаналізувати теоретичну базу проблемної області та застосування інформаційної аналітики в даній сфері;
- провести аналіз методів та методик інформаційного аналізу та моделювання процесів, включно з програмними засобами реалізації;
- відібрати та стандартизувати досліджувані дані, спроектувати та реалізувати прогностичну модель;
- провести та відобразити аналітичні розвідки щодо можливості імплементації отриманої моделі на підприємствах різного господарського типу.

Об’єкт дослідження. Кліматичні процеси, що призводять до зміни рівня світового океану.

Предмет дослідження. Аналіз та прогнозування зміни рівня світового океану та розробка прогностичної моделі для моніторингу та прогнозування цих змін.

Методи дослідження включають в себе огляд літературної та джерельної бази, що присвячена вибраній тематиці, статистичний аналіз, що спрямований на аналіз та пошук залежностей у зібраній інформації, візуалізація даних, моделювання та прогнозування, що відіграють важливу роль на етапі проектування та реалізації прогностичної моделі.

Наукова новизна та практичне значення одержаних результатів полягає у розроблені діючої, ефективної прогностичної моделі змін рівня світового океану, що може бути імплементована в більш складні інформаційні системи для збільшення ефективності моніторингу кліматичних змін.

Наукова публікація: протягом навчання мною була взята участь у IV Всеукраїнській науково-практичній інтерне-конференції студентів, аспірантів та молодих вчених за тематикою «Сучасні комп’ютерні системи та мережі в управлінні» за темою «Тенденції розвитку Front-end інструментів у сучасній веброзробці» [60].

РОЗДІЛ 1. АНАЛІЗ ТЕОРЕТИЧНИХ ЗАСАД ВИКОРИСТАННЯ ІНФОРМАЦІЙНОЇ АНАЛІТИКИ У СФЕРІ КЛІМАТИЧНИХ ЗМІН

1.1. Аналіз моніторингу кліматичних змін

Впродовж свого існування людство займалось вивченням будови і функціонування навколишнього світу з метою подальшого розвитку і покращення власного існування. Однак, на сьогоднішній день простежується чіткий вплив технологічного прогресу на екосистему Землі.

Кліматичні зміни – зміни статистичних показників кліматичної системи, що зберігаються протягом кількох (щонайменше трьох) десятиліть і є невід'ємною частиною природного процесу.

Кліматичні зміни на планеті Земля відбувались протягом усієї її історії, до початку сучасного глобального потепління, яке почалось у кінці XIX століття. Історичні дослідження та геологічні записи свідчать про те, що клімат Землі змінювався безперервно протягом мільйонів років [35]. Причинами таких змін були різні природні фактори, які включають коливання в сонячній активності, зміни у складі шахів атмосфери, вулканічна активність, ерозія ґрунтів та інші.

Наприклад, в період плейстоцену (від 2,6 мільйонів років тому до 11,7 тисяч років тому), на Землі були значні кліматичні зміни, пов'язані з циклічними змінами нахилу Землі, ексцентриситету орбіти та передачі. Ці зміни призводили до змін в розподілі сонячної енергії на поверхні Землі та, відповідно, до змін клімату на різних регіонах планети.

Крім того, у відносно недавній історії Землі було кілька періодів з охолодженням клімату, таких як "Малий льодовиковий період" (XVI-XIX століття), який був охарактеризований значним зниженням температур на планеті та змінами в кліматичних умовах.

Таким чином, можна стверджувати, що кліматичні зміни відбувались на Землі задовго до сучасного глобального потепління і є природним процесом для нашої планети. Однак, швидкість, з якою відбуваються кліматичні зміни на

сьогоднішній день, змусила науковців шукати причини кліматичних змін, що не мають природного характеру, такою причиною є людська діяльність.

Головною причиною сучасних кліматичних змін, які спостерігаються на планеті протягом останніх 30 років, є людська діяльність. Люди впливають на клімат Землі через надмірне використання ресурсів, зокрема використання горючих палив, емісії парникових газів та інших шкідливих викидів в атмосферу.

Одним з основних викликів, які відносяться до впливу людської діяльності на клімат планети, є використання горючих палив. Спалювання нафти, газу та вугілля для виробництва електроенергії та транспорту призводить до викиду в атмосферу великих кількостей парникових газів, зокрема вуглекислого газу та метану. Ці гази затримують тепло в атмосфері, утворюючи парниковий ефект, або так званий ефект теплиці, що призводить до збільшення температури повітря та світового океану.

Також, інші дії людини, такі як вирубка лісів, забруднення водойм та використання засобів масової комунікації, також можуть впливати на клімат планети.

Відомо, що від початку промислової революції в середині XIX століття рівень вуглекислого газу в атмосфері збільшився на більше 40% (рисунок 1.1). Це призвело до зростання середньорічної температури повітря на планеті, змін в розподілі опадів, посилення екстремальних погодних умов та інших наслідків.

Одним з наслідків глобального потепління є танення льодовиків. Саме в Антартиці кліматичні зміни є найпомітнішими, адже льодовики тануть з неймовірною швидкістю. 1979 рік став першим роком супутникового спостереження за льодовиками. Того ж року у найтепліший сезон об’єм льоду зменшився на 40% (рисунок 1.2). За такими тенденціями до середини XI століття Арктика може залишитись без льоду [8].

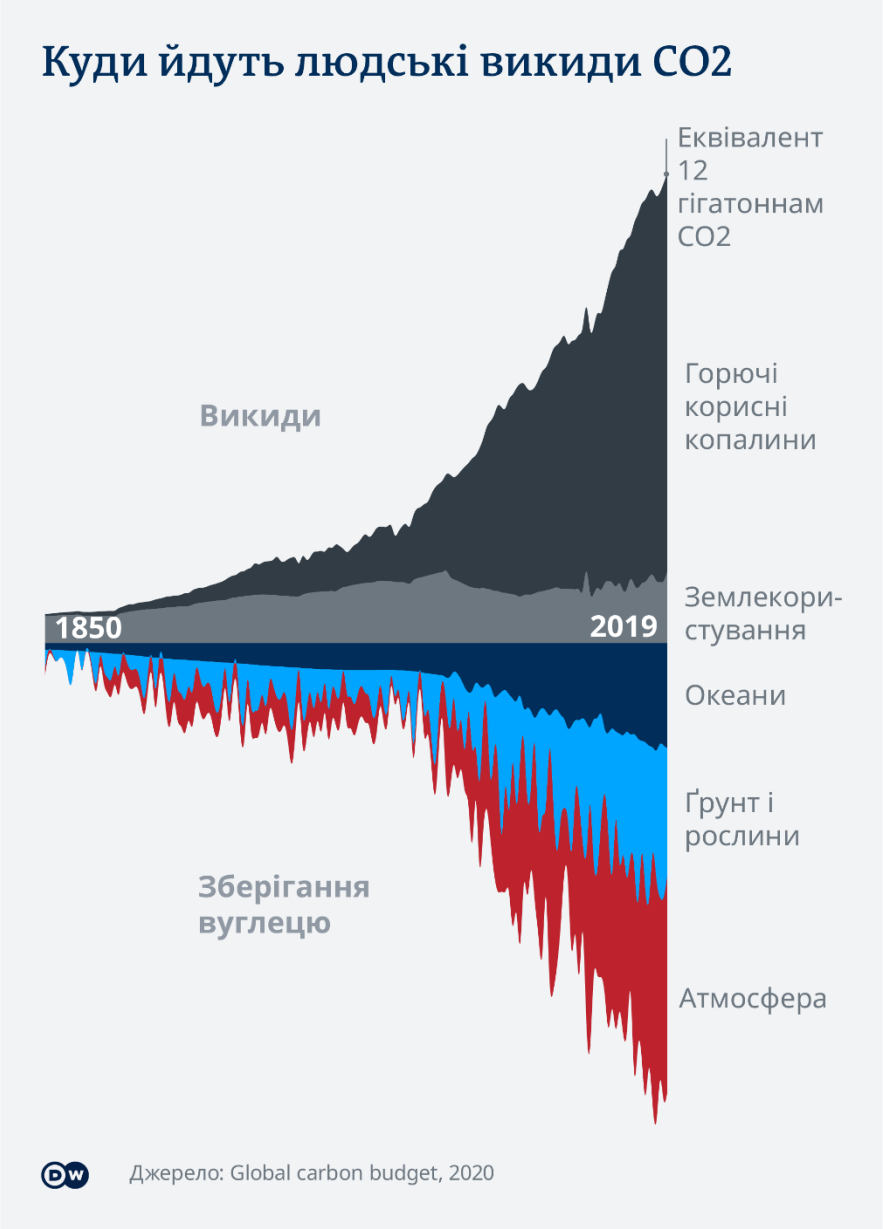

Рисунок 1.1 - Інфографіка викидів вуглекислого газу

Дані кліматичні зміни мають декілька серйозних наслідків. По-перше, це призводить до скорочення площі так званого білого покриву планети, який відбиває від 20% до 50% сонячної радіації [5]. У свою чергу, площа світового океану буде збільшуватись, а вона поглинає більше 95% сонячної радіації. Таким чином, зменшення льодового покриву призводить до збільшення світового океану і пришвидшення темпів нагрівання планети.

По-друге, за підрахунками науковців з National Snow and Ice Data Center [3], вічна мерзлота утримує 1400 гігатонн вуглекислого газу, а це є майже вдвічі більше, ніж на сьогоднішній день міститься у атмосфері. Таким чином, танення

льодовиків вивільняє додатковий вуглекислий газ, що збільшуватиме парниковий ефект.

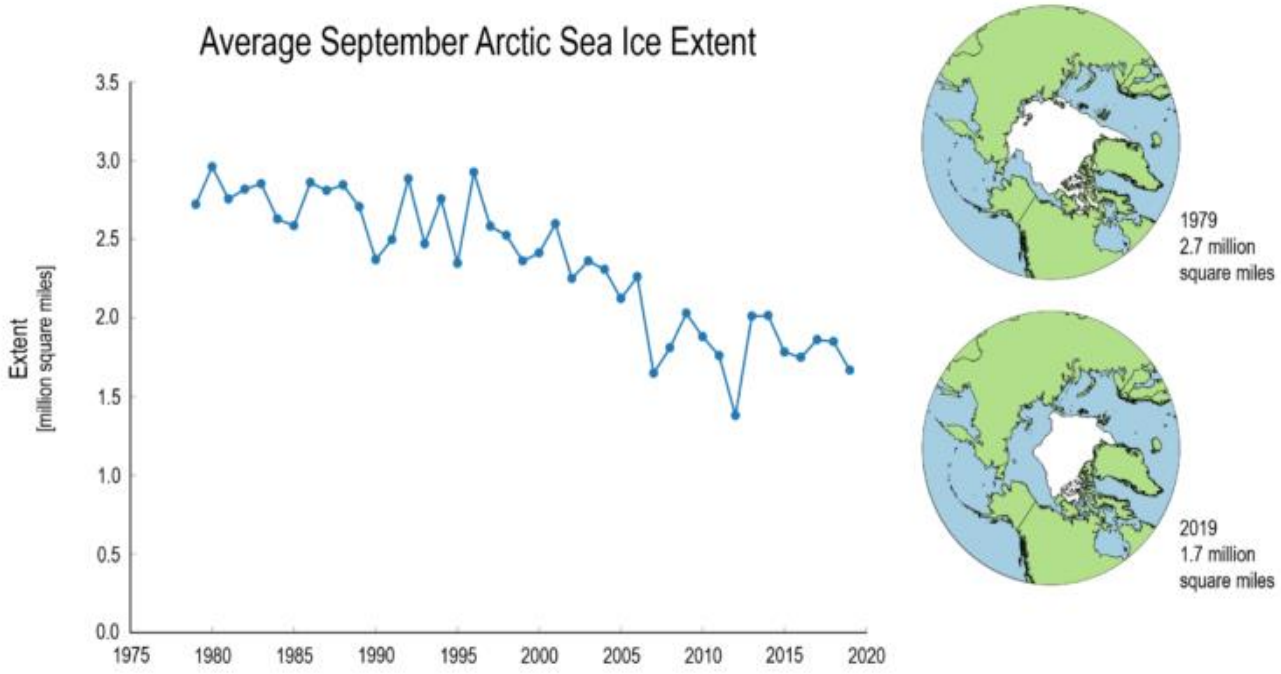

Рисунок 1.2 - Середній показник об’єму льодовиків у найгарячішу пору року з 1979 по 2019 роки [49]

По-третє, підвищення рівня світового океану призводить до затоплення суші. На сьогоднішній день поступово зникають під водою наступні острови: Мальдіви, Фіджі, Сейшельські Острови, Маршалові острови, Канарські острови, Федеративні штати Мікронезії, Французька Полінезія, Тувалу, Філіппіни, Соломонові острови.

Глобальне потепління призводить до збільшення випаровування води, що, у свою чергу, призводить до перерозподілу вологу. У результаті цього, в одних регіонах клімат стає надміру сухим, що призводить до посух, в інших регіонах надмірна волога конденсується і призводить до зливів і штормів, що призводить до затоплень. У результаті таких змін страждають не лише люди, а й місцева флора та фауна, яка не може вижити у таких умовах. Завдяки таким наслідкам глобального потепління з’явились кліматичні біженці - люди, які були змушені покинути домівки через раптові або ж довготривалі зміни у кліматі. За даними

ООН [18], з 2008 року щороку в середньому 21,5 млн людей були вимушені переселитись через погодні умови, такі як лісові пожежі, повені, шторми та екстремальні температури. Очікується, що в найближчі десятиліття дані показники будуть зростати.

Глобальне потепління може мати негативний вплив на здоров'я людей через зміну клімату та вплив на екосистеми, які забезпечують нас усім необхідним для існування.

Один з головних впливів глобального потепління на здоров'я є збільшення кількості випадків та поширення хвороб, що передаються через комах та інших носіїв, таких як комарі. До прикладу, збільшення кількості комарів, що передають хвороби, може спричинити епідемії хвороб, таких як малярія.

Глобальне потепління також впливає на якість повітря, особливо в міських районах, де рівень забруднення повітря вже є високим. Збільшення температури може спричинити збільшення кількості забруднюючих речовин, таких як озон та аерозольні частки, що можуть викликати проблеми з дихальною системою і призводити до ряду хвороб.

Так, за даними досліджень [4], забрудненість повітря являє собою один із основних факторів смертності, а також одним з головних чинників глобального тягаря захворювань.

Дослідження глобального тягаря хвороб - це глобальна програма, яка оцінює втрату працездатності та смертність від основних захворювань, травм та факторів ризику [59]. Воно охоплює 286 причин смерті, 369 захворювань та травм, 87 факторів ризику в 204 країнах та територіях. Дана оцінка враховує не лише роки життя, втрачені через ранню смертність, але й роки, прожиті з поганим станом здоров’я.

На рисунку 1.3 ми бачимо візуалізацію факторів ризику, які є розташованими у порядку DALY [29] (Disability-adjusted life years - роки життя з поправкою на інвалідність) - показника, що використовується для оцінки тягаря захворювання. З візуалізації видно, що забруднення повітря займає місце в топі

списку, що робить його одним з головних факторів ризику погіршення здоров’я у всьому світі.

Варто зазначити, що забруднення повітря не лише позбавляє людей життя, а й сильно впливає на його якість.

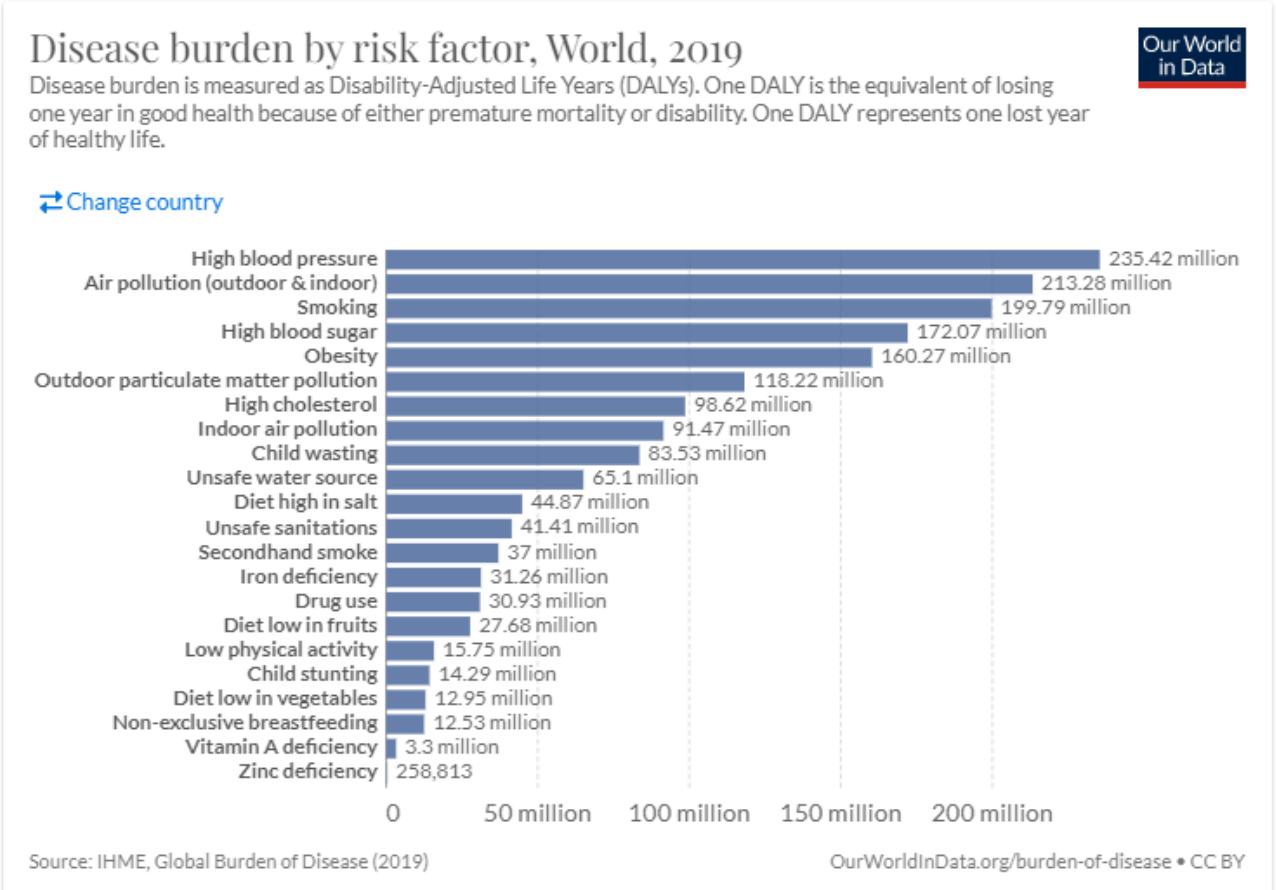

Рисунок 1.3 - Тягар захворювань за фактором ризику (2019 рік)

Глобальне потепління також може мати вплив на доступність питної води, особливо в регіонах з високою температурою та низькою вологістю. Зміна клімату може призвести до зменшення кількості доступної прісної води, збільшення солоністю води та інших проблем, пов'язаних з доступністю води для споживання та сільського господарства.

Таким чином, глобальне потепління може мати значний вплив на здоров'я людей, що підкреслює важливість зменшення викидів парникових газів та бережливого використання ресурсів, щоб зменшити його вплив на клімат.

Кліматичні зміни не лише серйозний вплив на екологію Землі, але й на економіку країн та глобальний ринок. Глобальне потепління впливає на аграрний

сектор, туризм, енергетику, будівництво, транспорт та багато інших галузей та сфер людського життя. Збільшення ризику катаклізмів та природних лих призводять до великих майнових втрат та людських жертв.

Екологічна криза призводить до змін умов для сільського господарства, що може впливати як на вирощування різних культур і їх існування, так і на врожайність.

Глобальне потепління може призвести до зміни умов туристичного бізнесу. Наприклад, зменшення снігового покриву в гірських регіонах або зникнення під водою ділянок суходолу може призвести до зменшення туристичного потоку і відповідних економічних наслідків.

Кліматичні зміни можуть призвести до змін умов для галузей енергетики, що може впливати на видобуток палива, виробництво електроенергії та інші аспекти енергетичної галузі. Дана екологічна криза може торкнутись будь-якої галузі людського життя. Наприклад, призвести до змін умов будівництва, адже можливі природні катаклізми вимагатимуть відповідної стійкості будов, що в свою чергу впливатиме на вартість будівельних та ремонтних робіт.

Зокрема в Україні внаслідок глобальних змін клімату за останні 30 років середньорічна температура зросла на 1°С [58]. Період від кінця двадцятого століття і до сьогодення вважається найтеплішим за всю історію погодних спостережень. Всі сезони в Україні стали теплішими. Найбільші зміщення шкали спостерігаються у січні (на 2,3°С) та у липні (1,4°С). Як наслідок, вже на сьогоднішній день можна спостерігати посилення посух, зміну водності річок та озер, зменшення, а місцями і повне зникнення води у криницях; швидку зміну природних ареалів існування флори та фауни, що призводить до скорочення або і повноцінне зникнення цілих видів.

На думку директора інституту водних проблем і меліорації НААН Михайла Яцюка [61], одним із яскравих наслідків глобального потепління в Україні є паводок, який спостерігається весною 2023 року. Так на півночі України підтоплено близько 500 домогосподарств. Спостерігається сезонне підняття рівня води в межах річок Дніпро, Десна, Сейм, Прип’ять, Горинь та

Західний Буг на територіях Волинської, Київської, Рівненської, Черкаської та Чернігівської областей.

Виходячи з усього вище переліченого, можемо зробити висновки про глобальне світове занепокоєння кліматичними змінами, важливість та значущість вивчення та аналізу цих змін з подальшою метою зменшення їх впливу та адаптації людства й екологічної системи Землі до наслідків глобального потепління.

1.2. Інформаційна аналітика у сфері кліматичних змін

Інформаційна аналітика даних та прогнозування грає дедалі більш важливу роль в сфері дослідження кліматичних змін та їх вплив на екологію планети Земля. Вона забезпечує збір, аналіз та інтерпретацію великих обсягів даних про клімат, зокрема зібраних за допомогою супутникових технологій та метеорологічних станцій.

Історія використання інформаційної аналітики даних та комп'ютерних технологій у сфері дослідження кліматичних змін розпочалась з статті "The Use of the Computer in Analysis of Climate Data" (1952) Льюїса Каплана [9]. Це дослідження є однією з перших спроб використання комп'ютерів у дослідженні кліматичних змін, в ній було розглянуто теоретичний підхід до використання електронних обчислювальних машин для аналізу кліматичних даних. Дана наукова робота показала, що використання комп’ютерних технологій дозволяє ефективніше аналізувати дані у сфері клімату, виявляти залежності між різними кліматичними факторами і розуміти їх взаємодію, здійснювати точніші прогнози.

Наступним етапом використання аналітики і комп’ютерних технологій у сфері дослідження кліматичних змін став запуск General Atmosphere Research Program (GARP) [22]. Це програма дослідження атмосфери, запущена в 1950-х роках, що фокусувалася на покращенні прогнозів погоди за допомогою математичних моделей. Була започаткована з метою покращення розуміння

процесів, які відбуваються в атмосфері, та розвитку інструментів для їх прогнозування.

GARP була організована за ініціативою Всесвітнього метеорологічного об'єднання (WMO) та його партнерів, таких як НАСА, Національний управління океанів та атмосфери (NOAA) та інші. У рамках програми було проведено численні дослідження з вивчення атмосфери, розроблено нові інструменти і моделі для прогнозування погоди.

Одним з найважливіших результатів GARP стала розробка моделі GARP Global Atmospheric Research Program (GARP) Model [23].

Global Atmospheric Research Program (GARP) Model була важливим кроком у розвитку моделювання клімату. Вона була використана для розуміння глобальних кліматичних змін та вирішення проблем, пов'язаних з погодою. GARP відіграла ключову роль у побудові збалансованих моделей атмосфери, що включали у себе як фізичні, так і хімічні процеси, і вона стала попередником багатьох інших кліматичних моделей, які були створені в подальшому.

Одним із головних результатів GARP було виявлення зв'язку між зміною концентрації вуглекислого газу в атмосфері та зміною клімату. Крім того, було встановлено, що глобальне потепління може привести до підвищення рівня моря, що може мати серйозні наслідки для узбережних міст та інших регіонів, що знаходяться на рівнині. Також було виявлено, що зміна клімату може призвести до зниження рівня снігового покриву в деяких регіонах, що може вплинути на водні ресурси та екологію цих регіонів.

У 1970-х роках, на базі моделі GARP (General Atmosphere Research Program), була створена модель Community Atmosphere Model (CAM) [13]. CAM стала однією з найбільш відомих і використовуваних комп'ютерних моделей клімату, що дозволяє проводити дослідження причин кліматичних змін і прогнозування їх на майбутнє. CAM була розроблена з метою відтворення різних аспектів клімату з високою точністю, зокрема розподілу температури, атмосферного тиску, вологості, хмарності та інших факторів.

CAM є однією з найбільш відомих і використовуваних моделей атмосфери, яка має велику кількість версій для різних застосувань і завдань. Вона базується на фундаментальних законах фізики і математики, включаючи рівняння збереження енергії, маси і кількості руху. Крім того, в моделі враховуються такі фактори, як взаємодія землі і атмосфери, а також процеси гідрології та кріосфери.

CAM є однією з найбільш відомих і використовуваних моделей атмосфери, яка має велику кількість версій для різних застосувань і завдань. Вона базується на фундаментальних законах фізики і математики, включаючи рівняння збереження енергії, маси і кількості руху. Крім того, в моделі враховуються такі фактори, як взаємодія землі і атмосфери, а також процеси гідрології та кріосфери.

Основними особливостями CAM є:

- Висока роздільна здатність: модель здатна розраховувати кліматичні зміни з точністю до кількох кілометрів на горизонтальній площині.
- Гнучкість і модульність: модель має багато версій, які можуть бути змінені або настроєні відповідно до потреб користувачів.
- Здатність до адаптації: модель може бути настроєна для різних кліматичних умов, що дозволяє її використання для вивчення кліматичних змін у різних частинах світу.

У загальному, CAM є потужним інструментом для вивчення кліматичних змін та їх впливу на земне середовище. Однак, варто зазначити, що кожна модель має свої обмеження та недоліки, і модель CAM не є винятком. Основними недоліками CAM є:

- Недостатня роздільна здатність: модель використовує грубу сітку для опису атмосферних процесів, що може призвести до неточностей в передбаченні деяких атмосферних явищ.
- Відсутність деяких процесів: модель не враховує деякі важливі атмосферні процеси, такі як евапотранспірація та взаємодія океану з атмосферою, що може призвести до неточностей в передбаченні кліматичних змін.
- Недостатня увага до регіональних відмінностей: модель більш схильна до передбачення глобальних кліматичних змін, але менш уважна до

регіональних відмінностей. Це може бути проблемою для вирішення конкретних проблем, таких як передбачення повеней чи посух.

- Обмеження в обсязі даних: модель залежить від доступності даних, і відсутність деяких даних може призвести до неточностей в передбаченні кліматичних змін.
- Відсутність врахування впливу людської діяльності: модель не враховує впливу людської діяльності на клімат, такий як викиди від промисловості та транспорту, що може призвести до неточностей в передбаченні кліматичних змін.

Виходячи з усього вищеописаного, CAM є важливою моделлю для розуміння та прогнозування кліматичних змін, але її використання повинно бути доповнене іншими методами та моделями для отримання повнішого та точнішого розуміння клімату.

На сьогоднішній день, для аналізу та прогнозування кліматичних змін використовуються наступні методи:

1. Методи побудови комп’ютерних моделей, які базуються на фізичних законах та математичних рівняннях для опису процесів, які відбуваються в атмосфері та океані. Ці методи використовуються для побудови комп'ютерних моделей, які можуть прогнозувати майбутні зміни клімату. Приклади: CAM, GCM, CCM, WRF.
2. Статистичні методи, які використовуються для аналізу статистичних даних про кліматичні зміни та побудови статистичних моделей для прогнозування майбутніх змін. Приклади: ARIMA, GARCH, регресійна аналітика.
3. Методи машинного навчання, які використовуються для аналізу великих обсягів даних та побудови моделей, які можуть прогнозувати майбутні зміни клімату. Приклади: нейронні мережі, дерева рішень, метод опорних векторів.

Ці методи використовуються в комбінації один з одним для покращення точності та надійності прогнозів клімату. Кожен метод має свої переваги та

недоліки, тому важливо враховувати їх при виборі підходу для конкретної задачі прогнозування кліматичних змін.

Надалі розглянемо деякі з сучасних моделей прогнозування кліматичних змін.

Однією з таких моделей є Community Earth System Model (CESM) [28] - це комп'ютерна програма, яка моделює зміни клімату на Землі та інших планетах за допомогою інтегрованого підходу. CESM була розроблена групою науковців з різних інститутів під керівництвом National Center for Atmospheric Research (NCAR) з метою вдосконалення прогнозування кліматичних змін та розуміння їх причин.

Основні компоненти CESM включають модулі, що моделюють атмосферу, океан, льодовики, біосферу та хімію атмосфери. Ці компоненти взаємодіють між собою, щоб забезпечити повний опис кліматичних процесів. Дані, що використовуються для калібрування та валідації моделі, забезпечуються за допомогою спостережень та експериментів.

Переваги CESM полягають у тому, що вона дозволяє інтегрувати різні складові кліматичної системи, що дозволяє більш точно прогнозувати кліматичні зміни. Крім того, CESM дозволяє проводити дослідження різних сценаріїв зміни клімату та їх наслідків для природних та соціально-економічних систем.

Серед недоліків CESM можна відзначити високу складність і обмежену точність моделювання деяких процесів, таких як хмарність, яка може суттєво вплинути на кліматичний прогноз. Крім того, виникає проблема взаємодії різних компонентів моделі, що може призвести до неточності в прогнозуванні.

Іншою популярною моделлю клімату є Goddard Institute for Space Studies ModelE (GISS ModelE) [21]. GISS ModelE - є однією з найвідоміших моделей клімату, розробленою Годдардським інститутом космічних досліджень NASA. Ця модель є однією з найбільш точних моделей клімату і використовується для прогнозування температури поверхні Землі, рівня моря, снігового покриву, хмарності та інших кліматичних параметрів.

GISS ModelE є комплексною моделлю, яка поєднує в собі різні компоненти, що відображають фізичні процеси в атмосфері, гідрологічні і

біохімічні процеси на земній поверхні, процеси в океані та хімічні процеси. Ці компоненти об'єднуються, щоб створити інтегровану модель клімату, яка може бути використана для прогнозування кліматичних змін.

Основні переваги моделі GISS ModelE полягають у високій точності та гнучкості. Модель може використовуватися для різних досліджень в області клімату, включаючи дослідження міжвікових кліматичних змін, аналіз впливу глобального потепління на різні регіони світу, вивчення ефектів викидів парникових газів та багато іншого.

Одним з недоліків моделі є високі вимоги до обчислювальних ресурсів, що може зробити її недосяжною для деяких дослідницьких груп та організацій. Крім того, модель може містити неточності в результаті неповної або недостатньо точної інформації про певні процеси в кліматі.

Ще однією прогностичною моделлю такого рівня є Weather Research and Forecasting Model (WRF) [50] - гнучка та ефективна модель, призначена для дослідження погодних умов на різних масштабах - від локального до глобального. Модель була розроблена в Університеті штату Пенсильванія, США, у 1990-х роках та постійно модифікується для покращення її точності та ефективності.

WRF використовується для прогнозування погоди, дослідження кліматичних змін, моделювання повітряних мас, хмар та інших процесів у трьох просторових вимірах. Основні переваги WRF полягають у гнучкості та можливості налаштування для вирішення різноманітних завдань, а також у високій роздільній здатності, що дозволяє досліджувати явища на детальному рівні.

Недоліки WRF пов'язані зі складністю її використання та підготовкою вхідних даних, а також з необхідністю великої кількості обчислювальних ресурсів для проведення розрахунків з високою роздільністю.

Підсумовуючи все вище описане, можна зробити наступні висновки:

- Методи інформаційного аналізу даних та прогнозування є невід’ємною складовою процесу дослідження кліматичних змін на планеті. Сучасні

методи дозволяють більш точно прогнозувати кліматичні зміни і вивчати їх вплив на екосистеми.

- Одним з основних факторів, який ускладнює роботу аналітиків у сфері кліматичних змін, є процес збору кліматичних даних. Численні показники вимагають великої кількості ресурсів та високої точності вимірювань.
- Побудова ідеальної моделі клімату Землі є неможливою у зв'язку з рядом факторів, таких як неможливість врахування всіх чинників, які мають вплив на екологію, відсутність достатньої кількості даних та складнощі відображення просторових та часових змін. Проте, розробка моделей на основі доступних даних та аналіз їх прогнозів, дає можливість більш точно передбачати кліматичні зміни та їх наслідки.

Висновки до першого розділу

У першому розділі кваліфікаційної роботи магістра було проведено інформаційно-літературний огляд проблематики кліматичних змін та сучасний стан дослідження даної проблеми.

На основі розглянутої інформації ми можемо стверджувати про беззаперечність процесу змін клімату, шкідливий вплив індустріалізації та іншої людської діяльності. Однією з основних причин цього процесу є збільшення викидів парникових газів та їх накопичення в атмосфері планети, що впливає на глобальні екологічні процеси, такі як: підвищення середньої температури, танення льодовиків та збільшення рівня світового океану, зміна берегової лінії та зменшення біологічного різноманіття.

Також у даному розділі:

- був проведений огляд використання інформаційної аналітики у даній сфері;
- були розглянуті основні найвідоміші екологічні моделі, їх переваги, недоліки, принципи та основні аналітичні методи;
- Було визначено основні фактори, які ускладнюють процес збору та прогнозування кліматичних змін.

РОЗДІЛ 2. АНАЛІЗ МЕТОДІВ ТА МЕТОДОЛОГІЙ ІНФОРМАЦІЙНОГО АНАЛІЗУ ТА МОДЕЛЮВАННЯ

2.1. Методології управління проєктами з інформаційної аналітики даних

Методології управління [2] проєктами являють собою системні підходи, що допомагають ефективно керувати проєктами з метою досягнення поставлених цілей в рамках встановлених обмежень, таких як час, бюджет, ресурси та якість. Методології управління проєктами використовуються в різних сферах, включаючи науку про дані (Data Science), де збір та аналіз даних є ключовими компонентами проєкту.

Методології управління проєктами в галузі Data Science дозволяють ефективно вирішувати завдання з обробки даних, зокрема, вони допомагають керувати процесами збору, обробки та аналізу даних, а також забезпечувати високу якість виконання проєкту відповідно до вимог замовника. Методології управління проектами в галузі Data Science також дозволяють ефективно використовувати наявні ресурси, включаючи інструменти та персонал, та забезпечувати комунікацію між учасниками проекту.

Однією з відмінностей методологій управління проектами в галузі Data Science від звичайних методів управління проектами є те, що вони більш орієнтовані на ризики та змінність проекту. В галузі науки про дані, де проєкти зазвичай пов'язані з нечіткими вимогами та невизначеністю, важливо мати гнучкі методи управління проектами, які дозволяють швидко реагувати на зміни та адаптуватися до нових умов.

Однією з важливих особливостей методологій управління проектами в галузі Data Science є їх гнучкість та адаптивність. Дані проекти зазвичай є складними та динамічними, а тому необхідно змінювати підходи до їх управління на ходу.

На сьогоднішній день, говорячи про методології управління проєктами, не можна не згадати Agile-методології, такі як Scrum, Kanban та Lean. Дані

методології користуються нечуваною популярністю, вони дозволяють організовувати процес роботи в такий спосіб, щоб бути готовим до швидких змін та адаптуватися до нових умов. Крім того, такі методології підтримують взаємодію між учасниками команди та високому ступені взаєморозуміння, що важливо для успішного виконання проекту.

Розглядаючи методології управління проєктами, варто зазначити, що проєкти у сфері Data Science мають високий рівень специфічності навіть для галузі інформаційних технологій. Це стало поштовхом до розвитку методологій управління проєктами у сфері науки про дані [45], які мають ряд переваг:

1. Структурованість: методології управління проектами з аналітики даних надають структурований підхід до управління проектами, який допомагає забезпечити планування, виконання та контроль процесу аналізу даних.
2. Результат-орієнтованість: методології управління проектами відмінно підходять для аналізу даних, оскільки вони фокусуються на досягненні конкретних результатів та метою, а не на процесі самого аналізу.
3. Інструментарій: методології управління проектами мають набір стандартів, інструментів та практик, що допомагає забезпечити стандартизацію процесу аналізу даних та його управління.
4. Оптимізація процесів: використання методологій управління проектами допомагає зменшити ризики та підвищити ефективність процесу аналізу даних. Наприклад, забезпечується ефективніше використання ресурсів, управління ризиками та уникнення помилок.
5. Якість результатів: методології управління проектами з аналітики даних допомагають забезпечити якість результатів аналізу даних, оскільки вони надають чіткі та структуровані процедури для контролю та забезпечення якості даних.

У порівнянні з гнучкими методологіями, методології управління проектами з аналітики даних дозволяють більш ефективно керувати проектами в умовах великої кількості даних та складного аналітичного процесу. Вони забезпечують структурований підхід до керування проектом, що дозволяє більш

точно визначити етапи роботи, розподілити завдання та визначити відповідальних за їх виконання. Це сприяє підвищенню ефективності роботи команди та зменшенню ризиків. Крім того, використання методологій управління проектами у галузі науки про дані дозволяє підвищити якість та достовірність отриманих результатів, що є важливим для прийняття обґрунтованих рішень на основі аналізу даних. Таким чином, використання методологій управління проектами у сфері науки про дані є важливим кроком для досягнення успіху в аналітичних проектах та підвищення якості прийнятих рішень.

Існує декілька популярних методологій управління проєктами у галузі науки про дані. Надалі розглянемо деякі з них.

Team Data Science Process (TDSP) [56] - це методологія управління проектами з аналітики даних, розроблена компанією Microsoft для забезпечення ефективної та структурованої роботи з даними в команді. Ця методологія має за мету поєднати кращі практики з області науки про дані, інженерії програмного забезпечення та управління проектами.

TDSP складається з шести етапів, які пов'язані між собою і виконуються в логічному порядку. Етапи включають:

1. Business Understanding: На цьому етапі відбувається фокусування на розумінні бізнес-проблеми та визначенні потреб для її вирішення. Доцільно провести зустріч зі стейкхолдерами, щоб визначити ключові метрики та зберігати їх у вигляді Business Understanding Document.
2. Modeling: На цьому етапі проводиться розробка та оцінка моделей машинного навчання. Вона включає побудову моделей, їх оцінку, налаштування та вибір оптимальної моделі.
3. Deployment: Після вибору оптимальної моделі, процес переходить до етапу розгортання. В цьому етапі розробляється проект, що розгортає модель, і забезпечується її доступність для використання користувачами.

4. Customer Acceptance: Цей етап забезпечує забезпечення відповідності побудованої моделі очікуванням клієнта та стейкхолдерів. Також він включає в себе проведення тестів та оцінку результатів.
5. Operationalization: Останній етап TDSP - це забезпечення ефективної експлуатації моделі, зокрема, розробка та забезпечення виконання процесів підтримки, моніторингу та управління ними

До недоліків використання TDSP можна віднести необхідність виконання кожного з етапів, що може затримати час виконання проекту та призвести до більшої складності в управлінні проектом, особливо якщо проект не має чіткої постановки задачі та вимог. Крім того, TDSP не завжди підходить для проектів з невеликими обсягами даних та невеликими командами, оскільки може бути важко знайти людей, які володіють достатньою кількістю різних навичок, щоб працювати на кожному етапі проекту.

Також, TDSP не враховує можливості постійних змін вимог, що можуть виникати протягом проекту, тому може бути не ефективним для проектів з високим рівнем невизначеності. Крім того, TDSP може бути дещо важким для розуміння та використання для людей, які не мають досвіду в роботі з даними та аналітикою.

Однак, якщо використовувати TDSP для проектів з великим обсягом даних та складними аналітичними завданнями, можна забезпечити більш ефективне та структуроване управління проектом та підвищити шанси на успіх проекту.

SEMMA (Sample, Explore, Modify, Model, Assess) [55] - це методологія, розроблена SAS Institute для проведення аналізу даних. SEMMA є процесом етапного аналізу даних, що містить п'ять етапів, кожен з яких розроблений для певної мети.

1. Sample (вибірка): Перший етап SEMMA - це вибірка даних. На цьому етапі вибирається відповідна вибірка даних з загального набору даних, з якою буде працювати аналітик. Правильний вибір вибірки дозволяє збільшити ефективність і точність наступних етапів.

2. Explore (дослідження): На другому етапі аналітик розглядає відношення між різними змінними та встановлює ті змінні, що є найбільш важливими для подальшого аналізу. На цьому етапі також проводяться візуалізації даних та аналіз розподілу змінних.
3. Modify (модифікація): На третьому етапі аналітик займається очищенням даних та підготовкою їх до подальшого аналізу. На цьому етапі також можуть бути виконані додаткові дослідження для встановлення більш точного відношення між змінними.
4. Model (моделювання): На четвертому етапі проводиться моделювання, тобто використання алгоритмів машинного навчання та інших методів для прогнозування майбутніх значень змінних на основі наявних даних. На цьому етапі можуть використовуватися різні моделі, такі як регресійна модель, дерево рішень, нейронна мережа тощо.
5. Assess (оцінка): Останній етап SEMMA - це оцінка результатів аналізу та моделювання. На цьому етапі проводяться перевірка, валідація та інтерпретація результатів, щоб переконатися, що модель є достатньо точною та може бути використана для прийняття рішень. У цьому етапі можуть бути використані різні методи оцінки, такі як матриці помилок, ROC-криві, криві підтвердження та інші.

Методологія SEMMA є структурованим та логічним підходом до аналізу даних, який дозволяє ефективно вирішувати проблеми великих обсягів даних. Вона має чітко визначені етапи, які допомагають уникнути втрати часу та зусиль на невірних напрямках дослідження. Методологія SEMMA може бути використана для розв'язання різних типів задач, від розуміння даних до прогнозування. Однак, є недоліки, такі як відсутність інтерактивності між етапами, що може призвести до втрати потенційної інформації або можливостей. Методологія SEMMA може бути недостатньо гнучкою для різних проектів, оскільки вона досить стандартизована та передбачувана. Крім того, на етапі Assessment може бути важко визначити, яка модель є найбільш ефективною для

задачі, оскільки процес включає багато статистичних методів та інтерпретацію результатів.

CRISP-DM (CRoss Industry Standard Process for Data Mining) [54] являє собою міжгалузевий стандартизований процес дослідження даних, який широко використовується у різних прикладних областях вже понад двадцять років. CRISP-DM описує життєвий цикл дослідження даних (рисунок 2.1), що складається з 6 фаз, від постановки завдання з точки зору бізнесу до впровадження технічного рішення. Незважаючи на чітку структурну визначеність, дана методологія належить до групи гнучких методологій моделювання та розробки, що в свою чергу дозволяє нам за необхідності повертатись до попередніх етапів та вносити необхідні корективи задля вирішення поставленої задачі.

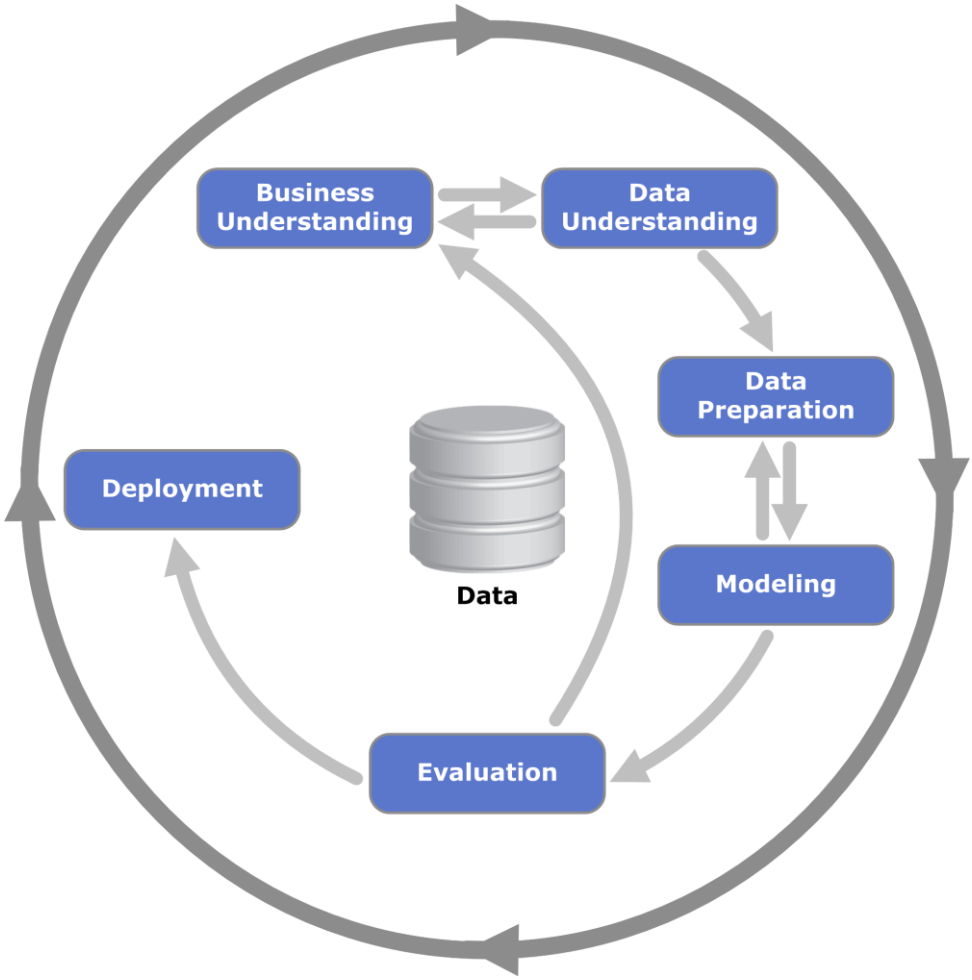

Рисунок 2.1 - Схема життєвого циклу процесу дослідження даних за методологією CRISP-DM

Першою фазою методології CRISP-DM є Business Understanding або ж розуміння бізнесу. Основним завданням даної фази є визначення цілей проекту і вимог з боку бізнесу. Надалі на основі отриманих бізнес завдань відбувається постановка задачі інтелектуального аналізу даних і формулювання плану досягнення цілей проекту.

Другою фазою методології є Data Understanding або розуміння даних. Дана фаза є логічним продовженням розуміння бізнесу і спрямовує увагу на виявлення, збір та поверхневий аналіз наборів даних, які у результаті допоможуть досягти цілей проекту. Розуміння даних включає у себе 4 завдання:

1. Збір вихідних даних – отримання початкових даних (за необхідності вивантаження їх до інструменту аналізу).
2. Опис даних – поверхневий огляд даних, визначення їх основних характеристик як то наявних полів і форматів даних.
3. Дослідження даних – аналіз даних та виявлення зв’язків та закономірностей.
4. Перевірка якості даних – перевірка наявності помилок як то порожніх полів, невірних форматів тощо.

Третьою фазою методології CRISP-DM є Data Preparation або підготовка даних. Дана фаза вважається найдовшою і займає до 80% часу роботи над проектом. На цьому етапі відбувається підготовка остаточного набору даних для подальшого моделювання. Підготовка даних складається з 5 кроків:

1. Обрання даних – визначення наборів даних, що стануть основою результуючого датасету.
2. Очистка даних – так звана ліквідація сміття або ж прибрання помилкових і непотрібних даних.
3. Створення похідних даних – за необхідності додавання нових додаткових атрибутів.
4. Інтегрування даних – об’єднання даних з декількох початкових датасетів.

5. Форматування даних – за необхідності зміна формату даних (наприклад багато числових даних початково зберігаються строками, однак для проведення математичних операцій необхідний числовий тип даних).

Четвертою фазою є Modeling або моделювання. Дана фаза включає у себе наступні кроки:

1. Обрання методу моделювання (у деяких випадках рекомендується обирати декілька методів і порівнювати отримані моделі).
2. Створення тестового набору даних – деякі методи моделювання потребують розділення датасетів на навальні, тестувальні і перевірочні.
3. Побудова моделі – побудова моделі відповідно до обраного методу або методів.
4. Оцінка моделі – у випадку створення декількох моделей, спеціаліст повинен визначити критерії успіху і порівняти за ними отримані моделі.

П’ятою фазою є Evaluation або оцінка. На відміну від останнього кроку попередньої фази дана фаза оцінює отриману модель на відповідність поставлених бізнес завдань. Оцінка включає в себе такі етапи:

1. Оцінка результатів – оцінка відповідності отриманої моделі критеріям успіху бізнесу.
2. Перевірка (Review) – процес оцінки пройдених фаз проекту, визначення успішності виконання кожної фази, формулювання відповідних висновків.
3. Визначення наступних кроків – на основі оцінки визначити необхідність переходу до фази впровадження або ж повернення до попередніх фаз.

Шостою заключною фазою методології CRISP-DM є Deployment або ж впровадження. Цей завершальний етап містить 4 завдання:

1. План впровадження – розробка і документацію плану впровадження фінальної моделі.
2. План підтримки та моніторингу – розробка детального плану моніторингу та технічного обслуговування з метою уникнення можливих проблем на етапі експлуатації.

3. Розробка фінального звіту – команда проекту має розробити фінальне резюме і презентацію результатів.
4. Фінальний огляд проекту – завершальна ретроспектива проекту, на якій команда має визначити позитивні та негативні аспекти виконання проекту, за необхідності підсвітити моменти, що потребують більшої уваги в майбутньому.

Враховуючи все вище описане, можемо зробити висновки про ряд переваг використання CRISP-DM як методології управління проєктом у сфері аналітики даних. До таких переваг віднесемо:

- Гнучкість: CRISP-DM є гнучкою методологією, яка може бути використана для різних типів проектів та задач. Вона також може бути легко адаптованою для конкретних потреб проекту.
- Результативність: CRISP-DM забезпечує підхід до аналізу даних, що дозволяє виявляти ключові відкриття та створювати ефективні рішення для покращення бізнес-процесів.
- Структурованість: CRISP-DM надає чіткий та структурований підхід до аналізу даних, що допомагає забезпечити систематичний підхід до проблем та запобігає втраті часу на невірні напрямки.
- Ефективність: CRISP-DM дозволяє використовувати кращі практики в аналізі даних, що дозволяє ефективніше вирішувати проблеми та швидше досягати результатів.
- Набуті знання: CRISP-DM допомагає збирати, описувати та інтерпретувати дані, що дозволяє отримувати цінні знання про бізнес-процеси та клієнтські потреби.

Таким чином, на подальших етапах дипломної роботи методом управління проєктом слугуватиме CRISP-DM.

2.2. Методи інформаційного аналізу та прогнозування даних

На сьогоднішній день, інформаційна аналітика та прогнозування є дуже важливими напрямками розвитку сфери, досліджень та бізнесу. Завдяки швидкому розвитку інформаційних технологій, величезній кількості даних, що генеруються кожної секунди, і новітнім методам аналізу цих даних, стали можливими більш точні прогнози, більш точна аналітика та, відповідно, більш точне прийняття рішень.

Зараз є доступними різні методи аналізу та прогнозування. До прикладу, розглянемо найпопулярніші з них:

1. Регресія [44] - метод аналізу даних, який дозволяє прогнозувати значення залежної змінної на основі значень однієї або кількох незалежних змінних. Цей метод широко використовується в багатьох галузях, включаючи економіку, фінанси, медицину та багато інших. До переваг даного методу можна віднести легкість розуміння та використання, можливість побудови прогнозів та прийняття на їх основі рішень, оцінка ступеню впливу різних факторів на результуючу змінну тощо. До недоліків регресійного методу можна віднести той факт, що регресійні моделі зазвичай припускають лінійну залежність між змінними, тому можуть бути не завжди точними для складних нелінійних залежностей; регресійні моделі можуть бути дуже чутливими до викидів і аномальних даних, що може мати значний вплив на результат; при побудові моделей регресійним методом необхідно враховувати взаємозв’язок між змінними, що може бути складним у великих наборах даних. У загальному, регресія є корисним методом для прогнозування залежної змінної на основі незалежних змінних, проте, для отримання точних прогнозів, необхідно враховувати обмеження методу і проводити аналіз результатів з урахуванням можливих недоліків.
2. Дискримінантний аналіз [17] - це метод статистичного аналізу, який використовується для визначення різниць між двома або більше групами об'єктів, визначених за певними характеристиками. Основною метою

дискримінантного аналізу є знаходження лінійної комбінації змінних, яка найкращим чином відрізняє групи об'єктів. До переваг даного методу належать: даний метод дозволяє визначити, які змінні є важливими для відокремлення груп; дозволяє класифікувати нові об'єкти відповідно до їх характеристик; використовується для вирішення задач кластеризації і визначення взаємозв'язків між змінними. До недоліків дискримінантного аналізу віднесемо: даний метод є неефективним у випадку, якщо змінна не може бути лінійно розділеною; не ефективний у випадку, якщо кількість змінних більша за кількість об'єктів вибірки; підходить лише для задач з класифікації з двома або більше категоріями. У підсумку, дискримінантний аналіз є потужним інструментом аналізу даних, який дозволяє визначити групи об'єктів, відрізняючи їх за допомогою певних змінних.

3. Кластерний аналіз [12]- це метод аналізу даних, який використовується для групування об'єктів у класи (або кластери) на основі схожості між ними. Кластерний аналіз може бути застосований в різних галузях, включаючи біологію, медицину, бізнес, соціологію та інші. Цей метод базується на вимірюванні відстані між об'єктами та на підтримці схожості між ними. До переваг кластерного аналізу належать: надає можливість групувати об’єкти на основі їх схожості без заздалегідь визначених класів; дозволяє знайти нові залежності між об’єктами, що може бути використано для подальшого дослідження. До недоліків даного методу належать: у залежності від обраного методу кластерного аналізу, можуть виникнути проблеми з масштабуванням даних даних або відсутністю універсальних методів; не існує однозначного методу, що дозволить визначити оптимальну кількість кластері, що може призвести до неоднозначної інтерпритації результатів. Підсумовуючи, кластерний аналіз є корисним методом аналізу даних, основною метою якого є зменшення внутрішньої варіації між об’єктами в одному кластері і збільшення відстані між кластерами.

4. Аналіз часових рядів [48] - це метод аналізу даних, в якому досліджуються зміни значень якоїсь величини відносно часу. Даний метод надає можливість виявляти тенденції, цикли, сезонність та інші показники регулярних закономірностей у залежності від часу. На основі виявлених закономірностей, даний метод дозволяє побудувати прогнозування майбутніх значень. Однію з переваг даного методу є високий рівень інформативності отриманих результатів, завдяки чому метод аналізу часових рядів широко використовують у різних галузях, включаючи економіку, фінанси, біологію, медицину, соціологію та інші. До недоліків даного методу можна віднести: важкість виявлення інформативних закономірностей в невеликих обсягах даних; потенційний вплив нерегулярних факторів, таких як погода, політична ситуація, на результати аналізу; потреба в знаннях і досвіді в статистиці, математиці та програмуванні для проведення аналізу. У підсумку, аналіз часових рядів є важливим інструментом для виявлення регулярних закономірностей в залежності від часу, що дозволяє зробити прогноз майбутніх значень та прийняти рішення в різних галузях. Однак, для успішного застосування аналізу часових рядів необхідно мати достатній обсяг даних та знання інструментів статистики та програмування.
5. Метод головних компонент [41]- це метод зменшення розмірності даних, який використовується для виявлення головних змінних, які пояснюють більшість дисперсії в даних. Це досягається шляхом проектування оригінальних даних на нові ортогональні змінні, які називаються головними компонентами. Головні компоненти ранжуються за величиною відповідних власних значень, і вони можуть бути використані для зменшення розмірності даних шляхом відбору найважливіших компонентів. До переваг даного методу можна віднести: він дозволяє зменшити розмірність даних і зберегти важливі змінні; метод може бути застосований до даних з будь-якої кількості змінних; за допомогою даного методу, може бути значно поліпшена візуалізація даних; зменшення

розмірності даних може допомогти покращити ефективність багатьох алгоритмів машинного навчання. Серед недоліків даного методу важливо відзначити: зазвичай головні компоненти важко зіставити з фізичними змінними в досліджуваній системі; метод головних компонент може бути досить чутливим до аномалій в даних. Метод головних компонент є потужним інструментом для зменшення розмірності даних і візуалізації даних. Він широко використовується в багатьох галузях, таких як фінанси, біологія, фізика та машинне навчання. Однак, перед використанням методу головних компонент варто ретельно проаналізувати дані і визначити, чи є він застосовним для конкретного завдання.

6. Метод опорних векторів [19] є одним з найбільш популярних методів машинного навчання, який використовується для класифікації і регресії. Основна ідея полягає в тому, щоб знайти оптимальну гіперплощину, яка розділяє дані на дві категорії з максимальною можливою шириною між ними. До переваг даного методу можна віднести: хорошу точність (у випадку правильного налаштування, метод має високу точність); ефективність у високорозмірному просторі, завдяки чому метод опорних векторів часто використовується у завданнях з багатьма змінними; робастність - не втрачає точність при наявності помилок, шумів або аномалій; гнучкість - може використовувати різні функції ядра, що дозволяє моделі працювати з різними типами даних. До недоліків даного методу варто віднести: потенційну громіздкість, оскільки метод може потребувати багато часу на тренування великої кількості даних; складність обрання правильної функції ядра; складність визначення оптимальних параметрів. Підводячи підсумок, варто зазначити, що метод опорних векторів є потужним методом машинного навчання, який зазвичай працює добре, коли правильно вибрані параметри та функції ядра. Однак, треба мати на увазі, що метод опорних векторів може мати обмеження, коли вхідні дані мають велику кількість ознак або коли дані не є лінійно роздільними.

7. Метод штучних нейронних мереж [20] - це один з методів машинного навчання, який моделює роботу людського мозку, щоб навчити комп'ютер розпізнавати складні залежності між вхідними та вихідними даними. Нейронна мережа складається з нейронів, які співпрацюють між собою та обмінюються сигналами. Кожен нейрон отримує вхідні дані, обчислює зважену суму вхідних сигналів та передає результат через активаційну функцію до наступного нейрона або вихідного шару. Для навчання нейронної мережі використовуються набори даних, на яких вона покращує свої здібності. Після навчання мережа може бути використана для прогнозування вихідних значень для нових вхідних даних. Перевагами методу нейронних мереж є здатність до роботи з великою кількістю вхідних даних, розпізнавання складних залежностей та здатність до навчання без потреби в експертних знаннях про даній проблематиці. Однак, недоліками можуть бути складність підбору оптимальної архітектури та підгонки параметрів, часом потрібний для навчання, а також можлива низька інтерпретованість результатів. Таким чином, метод нейронних мереж - це надзвичайно потужний метод машинного навчання, який здатний розпізнавати складні залежності та прогнозувати вихідні значення для нових вхідних даних, але може потребувати підбору оптимальної архітектури та підгонки параметрів.
8. Дерева рішень [16] - це аналітичний метод, який використовується для класифікації та прогнозування шляхом створення дерева рішень з урахуванням характеристик вхідних даних. У процесі побудови дерева рішень, з вхідних даних вибираються найбільш важливі ознаки, за якими можна здійснити класифікацію. Дерево рішень будується шляхом розділення набору даних на дві групи відповідно до значення вибраної ознаки. Цей процес повторюється для кожної нової групи, поки не буде досягнута певна умова зупинки, наприклад, достатньої чистоти класифікації в кожній групі або достатньої кількості елементів у кожній групі. Перевагами даного методу є: простота та зрозумілість результуючої

моделі, можливість використання методу як для задачі класифікації, так і для задачі прогнозування, а також швидкість побудови моделі. Однак, дерева рішень можуть мати і наступні недоліки: низький рівень ефективності для даних з великою кількістю ознак. що може не завжди гарантувати оптимальний результат. Підсумовуючи, метод побудови дерева рішень є досить простим та швидким, що дає зрозумілу та інтерпретовану модель. Однак, для складних завдань з великою кількістю ознак, даний метод може мати низьку ефективність.

9. Байєсівський аналіз [53] - це метод статистичного аналізу даних, що базується на теоремі Байєса, основна ідея якої полягає в тому, що ми можемо визначити ймовірності різних гіпотез або параметрів, враховуючи як даний набір спостережень, так і апріорні знання про ці параметри. Для виконання байєсівського аналізу спочатку визначається апріорна ймовірність гіпотези або параметра. Далі, на основі зібраних даних, використовуючи формулу Байєса, розраховується апостеріорна ймовірність гіпотези або параметра. Ця апостеріорна ймовірність слугує оновленою інформацією про те, яка з гіпотез або значень параметрів є найбільш вірогідною. Для виконання байєсівського аналізу спочатку визначається апріорна ймовірність гіпотези або параметра. Далі, на основі зібраних даних, використовуючи формулу Байєса, розраховується апостеріорна ймовірність гіпотези або параметра. Ця апостеріорна ймовірність слугує оновленою інформацією про те, яка з гіпотез або значень параметрів є найбільш вірогідною. Однією з головних переваг байєсівського аналізу є можливість врахувати апріорні знання про параметри. Це дає можливість отримати більш точні оцінки параметрів при обмеженій кількості спостережень. Крім того, байєсівський аналіз зазвичай має гнучкий підхід до моделювання, що дозволяє враховувати складні залежності між параметрами. Однак, недоліком байєсівського аналізу є висока обчислювальна складність, особливо коли кількість параметрів або гіпотез дуже велика. Крім того, визначення апріорних

ймовірностей може бути неточним, що може призвести до неточності оцінок апостеріорної ймовірності. В цілому, байєсівський аналіз є корисним методом аналізу даних, особливо в тих випадках, коли доступні апріорні знання про досліджувану проблему. Використання цього методу дозволяє зменшити вплив шуму та забезпечити більш точні результати прогнозування.

10. Машинне навчання [32] - це галузь штучного інтелекту, яка досліджує алгоритми та статистичні моделі, щоб дозволити комп'ютеру "навчатись" з даних. Мета машинного навчання полягає у виявленні закономірностей та шаблонів у великих наборах даних і використанні цієї інформації для прийняття рішень та прогнозування. У машинному навчанні використовуються різні типи алгоритмів, включаючи навчання з вчителем, навчання без вчителя та підсилення. У навчанні з вчителем, модель навчається на основі попередньо підготовлених даних, де кожен зразок має відповідну мітку. У навчанні без вчителя, модель намагається виявити шаблони та закономірності без заздалегідь відомих міток. У підсиленному навчанні, модель взаємодіє з оточенням та навчається на основі відповідей на свої дії. Переваги машинного навчання полягають у тому, що воно може обробляти великі обсяги даних та виявляти складні зв'язки, які складно визначити за допомогою інших методів. Крім того, машинне навчання може підвищити ефективність прийняття рішень, що може призвести до зниження витрат та покращення продуктивності. Недоліки машинного навчання включають необхідність великих обсягів даних для навчання моделей, можливість перенавчання на тренувальних даних, що може призвести до поганої загальної результативності, та складність пояснення, яка може бути проблемою у випадку, якщо модель використовує для прийняття критичних рішень важливих для людей, наприклад, у медицині або правосудді. Також, машинне навчання може виявитися недостатньо точним або ненадійним у випадках, коли дані містять помилки або спотворення, а також в умовах, коли змінні у навчальних даних можуть

змінюватися з часом. Крім того, використання машинного навчання може підвищити ризик порушення приватності та безпеки даних, якщо моделі навчаються на чутливих персональних даних без відповідних заходів захисту.

Підсумовуючи все вище описане, для прогнозування кліматичних змін оберемо методи штучних нейронних мереж. До штучних нейронних мереж відносяться:

1. Прямі мережі (Artificial Neural Network - ANN) [6] - мережі, в яких інформація переміщується тільки в одному напрямку (від вхідних нейронів до вихідних).
2. Зворотні мережі (Recurrent Neural Networks - RNN) [52] - мережі, в яких інформація може переміщуватися у зворотному напрямку, тобто з вихідних нейронів до вхідних. Це дозволяє використовувати контекст для аналізу послідовностей.
3. Згорткові мережі (Convolutional Neural Network - CNN) [57] - мережі, що використовуються для обробки зображень та інших типів даних, які мають структуру сітки (наприклад, звукові сигнали). Згорткові мережі використовують згортки та пулінг для ефективної обробки даних.
4. Мережі довготривалої пам'яті (LSTM) [62] - спеціальний тип зворотних мереж, який зберігає певну кількість інформації з минулих ітерацій для більш ефективного аналізу послідовностей.
5. Autoencoder [7] - мережі, які використовуються для зменшення розмірності даних або для генерації нових даних зі зменшеною розмірністю.

У випадку прогнозування кліматичних змін оберемо ANN та LSTM як підвид RNN.

Пряма штучна мережа (Artificial Neural Networks - ANN) [30] - це комп’ютерні системи, що імітують структури та функції людського мозку. Вони складаються з взаємопов'язаних нейронів, які працюють разом, щоб обробити вхідні дані та зробити передбачення на основі зібраних даних.

Архітектура ANN зазвичай складається з трьох типів штучних нейронів (рисунок 2.2): вхідні, приховані та вихідні. Вхідні штучні нейрони приймають вхідні дані, які потім передаються до прихованих штучних нейронів. Приховані штучні нейрони обробляють вхідні дані та генерують нові сигнали, які потім передаються до вихідних штучних нейронів. Вихідні штучні нейрони збирають сигнали з прихованих штучних нейронів та генерують вихідні результати.

Процес навчання ANN полягає в тому, щоб знайти ваги та зсуви штучних нейронів, які дають найкращі результати для задачі. Для цього використовуються різні алгоритми навчання, такі як зворотнє поширення помилки та генетичні алгоритми.

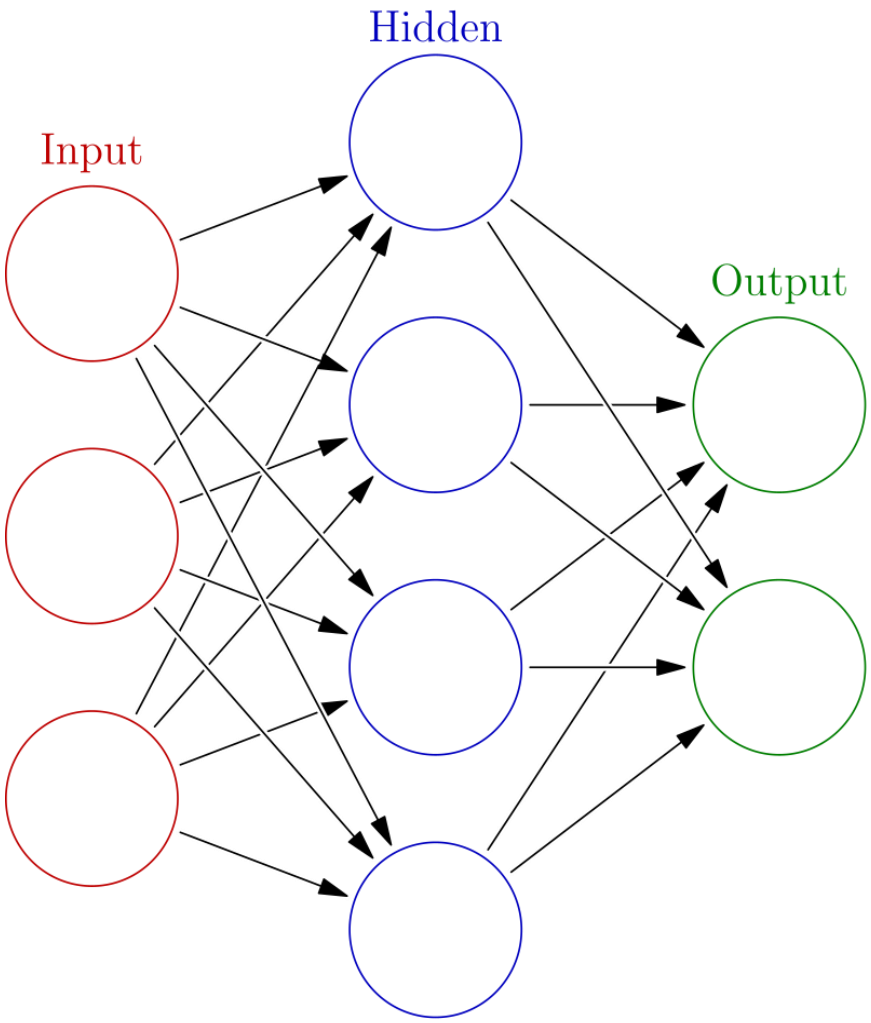

Рисунок 2.2 - Схема будови штучної нейронної мережі

ANN можуть бути використані для багатьох різних завдань, таких як класифікація, прогнозування та розпізнавання образів. Застосування ANN стали особливо популярними в області глибинного навчання, де вони використовуються для розпізнавання образів та обробки природних мов.

Зворотні мережі (Recurrent Neural Network - RNN) [14] - це клас нейронних мереж, які призначені для роботи з послідовними даними. Вони використовують попередні результати своєї роботи як додатковий вхід на кожному кроці часу. Це дозволяє зберігати інформацію про попередні кроки часу і використовувати її для прийняття рішень на поточному кроці.

У RNN є вхідний шар, прихований шар та вихідний шар. Кожен шар містить нейрони, які мають ваги і зсуви. Вхідний шар приймає вхідні дані, прихований шар виконує обчислення на кожному кроці часу і зберігає інформацію про попередні кроки часу, а вихідний шар видає результат на поточному кроці часу.

RNN можуть бути використані для багатьох завдань, таких як машинний переклад, генерація тексту, розпізнавання мови та багато інших. Вони є особливо корисними для роботи з послідовнісними даними, де інформація про попередні кроки часу є важливою для прийняття рішень.

Однак, у RNN є певні недоліки, такі як проблема зниклого градієнту, яка може виникнути при тренуванні мережі на довгих послідовностях. Це може призвести до того, що інформація про ранні кроки часу не буде врахована під час тренування. Для цього були розроблені покращені версії RNN, такі як LSTM і GRU, які дозволяють уникнути проблеми зниклого градієнту та покращити точність прогнозування.

LSTM (Long Short-Term Memory) [15] є підтипом рекурентних нейронних мереж (RNN), який був запропонований в 1997 році Хохрейтером та Шмідхубером. LSTM був створений, щоб розв'язати проблему зниклих градієнтів у RNN, яка може виникнути при тренуванні на довгих послідовностях даних.

LSTM має здатність зберігати попередній стан мережі і використовувати його для вирішення поточної задачі. Основні компоненти LSTM - це "ворота" (gates), які виконують функцію визначення, яка інформація повинна бути збережена, а яка - забута. Ці ворота реалізовані за допомогою спеціальних функцій активації, які дозволяють контролювати потік інформації в мережі.

Загальна архітектура LSTM містить чотири основні елементи (рисунок 2.3):

1. Шар пам'яті (memory cell) - це "пам'ятлива" частина мережі, яка зберігає інформацію про минулий стан. Вона може зберігати і видаляти інформацію за допомогою воріт.
2. Ворота забуття (forget gate) - дозволяють мережі вибирати, яку інформацію з попереднього стану необхідно забути. Це допомагає запобігти проблемі затухання градієнту, яка виникає у звичайних рекурентних нейронних мережах. Якщо нейронна мережа занадто довго працює над одним завданням, градієнт може ставати настільки малим, що мережа перестає навчатися.
3. Ворота входу (input gate) - дозволяють мережі вибирати, яка нова інформація повинна бути додана до memory cell. Це допомагає мережі зберігати інформацію про зміну контексту вхідних даних.
4. Вихідний шар (output layer) - відповідає за вибір, яку інформацію з memory cell слід використовувати на виході. Зазвичай це є комбінація інформації з поточного стану та інформації, яка була збережена в memory cell.

Всі ці елементи взаємодіють між собою, щоб забезпечити мережі здатність запам'ятовувати довготривалу залежність вхідних даних і здійснювати передбачення на їх основі.

Оскільки LSTM мережі можуть зберігати інформацію про контекст вхідних даних протягом довгого періоду часу, вони здатні досягати високої точності при розв'язанні завдань машинного навчання, що пов'язані з послідовністю даних, наприклад, в розпізнаванні мови, машинному перекладі та прогнозуванні часових рядів.

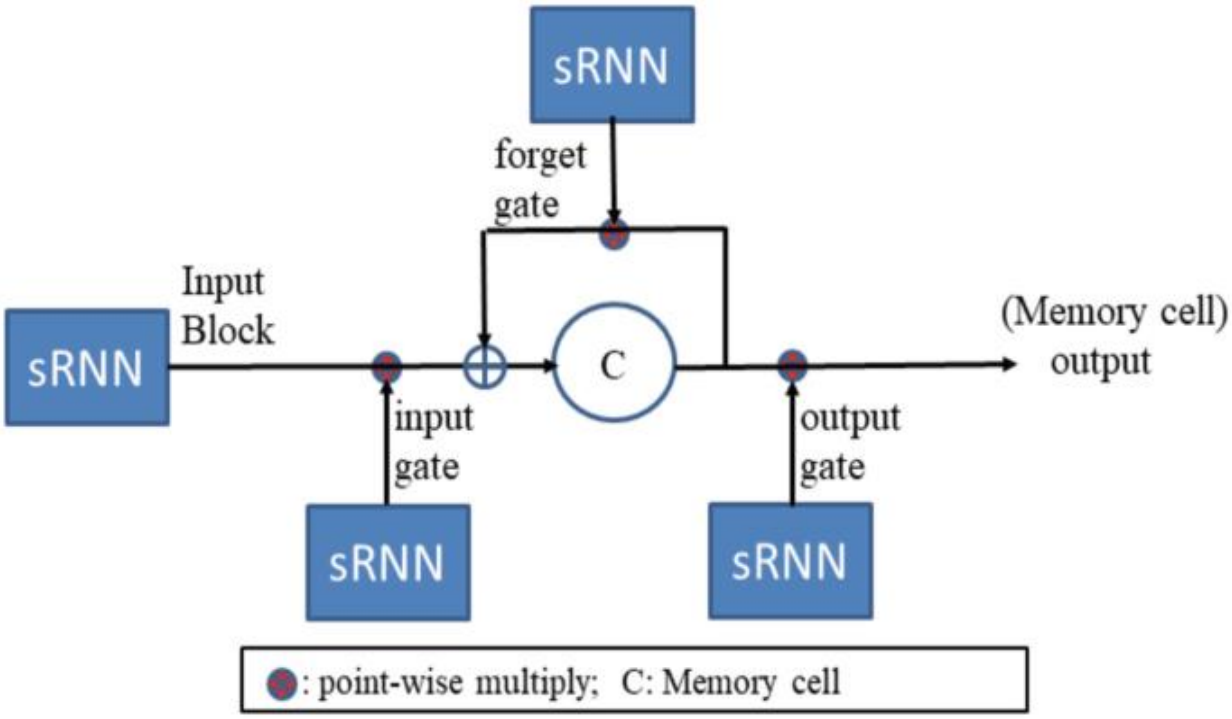

Рисунок 2.3 - Загальна схема роботи LSTM

Таким чином, у подальшому для побудови прогностичних моделей кліматичних змін будут використовувати два методи штучних нейронних мереж, а саме: ANN та LSTM.

2.3. Програмні засоби для моделювання

На сьогоднішній день існує широкий вибір інструментів, мов програмування, окремих бібліотек і пакетів для Data Science. Основні мови програмування, які використовуються для розробки нейронних мереж, включають наступні:

1. Python [42] - це одна з найпопулярніших мов програмування, яка використовується для розробки штучних нейронних мереж. Існує багато бібліотек, таких як Tensorflow, PyTorch, Keras та інші, які надають зручний інтерфейс для розробки нейронних мереж.
2. Java [31] - це мова програмування, яка також може бути використана для розробки нейронних мереж. Існує декілька бібліотек, таких як

Deeplearning4j, Encog та Neuroph, які надають зручний інтерфейс для розробки нейронних мереж.

3. C++ [10] - це мова програмування, яка використовується для розробки ефективних та швидких нейронних мереж. Існує декілька бібліотек, таких як TensorFlow C++, Caffe та Torch, які надають зручний інтерфейс для розробки нейронних мереж.
4. MATLAB [33] - це програмна платформа, яка надає засоби для чисельних розрахунків, аналізу даних та візуалізації. MATLAB також має вбудовані функції для розробки та імплементації нейронних мереж.
5. R [43] - це мова програмування, яка широко використовується в статистичному аналізі даних. Існує декілька бібліотек, таких як TensorFlow для R, Keras для R та інші, які надають зручний інтерфейс для розробки нейронних мереж.

Серед оглянутих мов для подальшої роботи з даними та моделювання оберемо Python.

Python [51] є високорівневою мовою програмування, що, заради універсальності в підході до програмування, використовує різні парадигми програмування, зокрема:

- об’єктно-орієнтованість – парадигма програмування на основі множини об’єктів, що взаємодіють на основі концепцій інкапсуляції, спадкування, поліморфізму та абстракції;
- процедурність – парадигма реалізації послідовних кроків заснованих на підпрограмах, методах або функціях, що викликаються з будь-якого місця програми, включно із самовикликом такого коду через рекурсію;
- функціональність – парадигма, що заснована виключно на обробці функцій та уникає стану зміни даних, а способом розбиття програми є створення нової функції при правилі композиції – оператора суперпозиції функції;
- аспектно-орієнтованість – парадигма доповнення або альтернатива ООП, що вводить ділення функціональності на класи або модулі та впроваджує

додаткову логіку функціональності, що викликається у місцях з’єднання таких модулів.

Це інтерпретована мова високого рівня, що використовує принципи строгої динамічної типізації, що дозволяє використовувати гнучкість багатьох скриптових мов програмування – роботу із змінними, що не мають типів. Такий підхід дозволяє використовувати змінні переписуючи в них дані не одного типу в залежності від необхідності логіки роботи програми.

Перевагами використання мови Python є швидкість написання програмного забезпечення, чому сприяє широка підтримка модулів, бібліотек та готових програмних пакетів над якими працюють розробники, що входять до спільноти підтримки цієї мови програмування.

Для роботи з наборами даних оберемо бібліотеку Pandas.

Pandas [40] - це бібліотека для маніпулювання та аналізу даних у мові програмування Python. Вона надає широкий спектр інструментів для роботи з табличними даними, які можна подати у форматі CSV, Excel, SQL або зчитати з веб-серверів та інших джерел даних.

Основні переваги бібліотеки pandas:

- Легка робота з даними. Pandas надає інтуїтивно зрозумілий інтерфейс для роботи з даними, що дозволяє зчитувати, зберігати, обробляти та аналізувати їх швидко та ефективно.
- Швидкість обробки. Pandas є дуже швидкою бібліотекою, що дозволяє оброблювати великі об'єми даних з високою швидкістю та ефективністю.
- Можливість маніпулювання даними. Pandas надає багато інструментів для маніпулювання даними, таких як фільтрація, групування, сортування, об'єднання, розбиття, обрізка та інші.
- Робота з пропущеними значеннями. Pandas має вбудований механізм для роботи з пропущеними значеннями, який дозволяє легко обробляти дані з пропущеними значеннями.

- Візуалізація даних. Pandas має інтегровану підтримку візуалізації даних, яка дозволяє легко побудувати графіки, діаграми та інші візуальні елементи для аналізу даних.

Для проведення окремих операцій з даними, оберемо бібліотеку NumPy.

NumPy (Numerical Python) [37] - це бібліотека мови програмування Python, призначена для роботи з числовими даними. Вона надає багато функцій для роботи з багатовимірними масивами та матрицями, що дозволяє виконувати різні математичні операції.

NumPy є основою для інших бібліотек Python, які використовуються для наукових обчислень, таких як Pandas, SciPy, Scikit-learn та TensorFlow.

Основні можливості NumPy:

- Масиви (Arrays): NumPy надає потужний об'єкт масиву, який може містити багатовимірні масиви. Масиви NumPy більш ефективні у використанні пам'яті та операцій з масивами, ніж звичайні списки Python.
- Універсальні функції (Universal functions): NumPy містить вбудовані універсальні функції, які забезпечують елементарні математичні операції з елементами масивів.
- Індексування та зрізи (Indexing and Slicing): NumPy масиви можуть бути індексовані та зрізані, як звичайні списки Python, але з більшою функціональністю.
- Радіальні функції (Radial basis functions): NumPy містить функції, які використовуються в статистичних дослідженнях та машинному навчанні для побудови моделей.
- Лінійна алгебра (Linear algebra): NumPy містить бібліотеку для роботи з лінійною алгеброю, такою як матриці та вектори, та надає методи для розв'язування систем лінійних рівнянь, обчислення власних значень та векторів.
- Статистичні функції (Statistical functions): NumPy містить функції, які допомагають виконувати статистичний аналіз даних, такі як середнє значення, медіана, дисперсія, кореляція тощо.

- Трансформації Фур'є (Fourier transforms): NumPy містить функції для обчислення трансформації Фур'є, які використовуються в сигнальній обробці та обробці зображень.
- Інтерполяція (Interpolation): NumPy містить функції для інтерполяції між точками даних, що допомагає аналізувати даний на більшій кількість точок, ніж ті, що є в початкових даних. Інтерполяція може бути корисною при візуалізації даних або при отриманні більш детальної інформації про досліджуваний процес. Функції для інтерполяції в NumPy включають interp1d, interp2d, griddata та багато інших.

Для візуалізації даних оберемо бібліотеку Mathplotlib.

Matplotlib [34] - це бібліотека для створення візуалізацій у мові програмування Python. Ця бібліотека містить набір інструментів для створення графіків, діаграм, гістограм, кругових діаграм, теплових карт та інших типів графіків.

Основна ідея Matplotlib полягає в тому, щоб забезпечити користувачам можливість створювати візуалізації за допомогою коду на Python. Вона дає користувачам повний контроль над тим, як візуалізації виглядають і як вони поводяться.

Matplotlib має кілька важливих переваг, таких як:

- Широкий набір функцій і можливостей для створення різних типів графіків та візуалізацій;
- Простий та логічний інтерфейс, що дозволяє швидко розібратись у роботі з бібліотекою;
- Відкритий код та велика спільнота користувачів, що дозволяє знайти рішення для будь-яких проблем.

Ще одна перевага Matplotlib полягає в тому, що вона може бути легко інтегрована з іншими популярними бібліотеками для аналізу даних, такими як NumPy, Pandas та SciPy. Також Matplotlib має підтримку для різних форматів файлів, таких як PNG, PDF, SVG, EPS та інших, що дозволяє зберігати візуалізації в зручному форматі для подальшого використання.

Для побудови та тренування моделей оберемо бібліотеку TensorFlow [46]. Бібліотека була розроблена компанією Google та є однією з найбільш популярних інструментів для створення штучних нейронних мереж.

TensorFlow має широкі можливості для роботи з числовими даними, в тому числі з многовимірними матрицями, зображеннями та звуком. Бібліотека містить багато інструментів для оптимізації моделей, в тому числі для розпаралелювання обчислень на графічних процесорах.

TensorFlow також має багато інструментів для візуалізації та налагодження моделей, що дозволяє досліджувати та покращувати їх ефективність. Бібліотека підтримує створення як секвенційних моделей, так і графових моделей, що дає можливість створювати складні моделі з багатьма вхідними та вихідними шарами.

Один з головних модулів TensorFlow - це Keras [47], який надає високорівневий інтерфейс для створення та тренування штучних нейронних мереж.

Keras забезпечує інтерфейс для побудови моделей нейронних мереж, що дозволяє використовувати різні типи шарів, такі як Dense (повнозв'язний шар), Convolutional (згортковий шар) та Recurrent (рекурентний шар), для побудови різноманітних архітектур нейронних мереж.

Крім того, Keras містить багато вбудованих функцій для тренування мереж, включаючи оптимізатори (наприклад, Adam, RMSprop), функції втрат (наприклад, Mean Squared Error, Binary Crossentropy) та метрики (наприклад, Accuracy, Precision, Recall).

Особливість Keras полягає в його простоті використання, що дозволяє новачкам в штучному інтелекті швидко вивчити основні поняття та почати розробляти власні моделі. Крім того, Keras підтримує гнучку конфігурацію та легку міграцію між різними фреймворками машинного навчання, такими як TensorFlow, Microsoft Cognitive Toolkit та Theano.

Для додаткових обчислювальних операцій оберемо бібліотеку Scikit-learn.

Scikit-learn (sklearn) [1] - це бібліотека машинного навчання для мови програмування Python. Вона надає широкий набір інструментів для роботи з даними, навчання моделей та оцінювання їх продуктивності. Однією з важливих складових sklearn є модуль "sklearn.preprocessing", який надає інструменти для попередньої обробки та підготовки даних перед застосуванням алгоритмів машинного навчання.

Модуль "sklearn.preprocessing" містить багато функцій для стандартизації, нормалізації та кодування даних. Наприклад, функція "StandardScaler" виконує стандартизацію даних, перетворюючи їх так, щоб кожен ознака мала середнє значення 0 і стандартне відхилення 1. Функція "MinMaxScaler" нормалізує дані, перетворюючи їх так, щоб кожен ознака мала значення в діапазоні від 0 до 1.

Модуль також містить функції для роботи з категоріальними даними. Наприклад, функція "LabelEncoder" дозволяє кодувати категоріальні ознаки у числові значення. Функція "OneHotEncoder" створює "гарячі" коди для категоріальних ознак, перетворюючи їх на вектори бінарних значень.

Крім того, модуль "sklearn.preprocessing" надає інструменти для роботи зі збільшенням розмірності даних, такі як функція "PolynomialFeatures", яка генерує всі можливі комбінації ознак заданого ступеня.

Усі ці функції допомагають зробити дані придатними для застосування алгоритмів машинного навчання. Вони є частинами зручного та потужного інструментарію sklearn для попередньої обробки та підготовки даних.

Таким чином, у даному підрозділі було проведено огляд та обґрунтовано вибір програмних засобів та інструментів для подальшої роботи з даними, побудови моделі та її оцінки.

Висновки до другого розділу

У другому розділі дипломної роботи було проведено огляд методологій управління проєктом у сфері інформаційної аналітики даних. У результаті чого

за методологію управління проєктом з інформаційної аналітики та прогнозування кліматичних змін було обрано CRISP-DM.

Надалі було проаналізовано переваги та недоліку ряду методів аналізу даних, на основі чого було обрано методи штучних нейронних мереж (ANN) та метод короткотривалої пам’яті (LSTM) для подальшого інформаційного аналізу та прогнозування даних у сфері кліматичних змін.

В останній частині другого розділу дипломної роботи було проведено огляд найпопулярніших мов програмування у сфері Data Science, на основі трендів та переваг оглянутих мов було обрано мову програмування Python як основний програмний засіб подальшого аналізу та моделювання. Також було обрано ряд програмних засобів - бібліотек та пакетів вище обраної мови, які в майбутньому будуть використані для всіх етапів аналізу та прогнозування у сфері кліматичних змін.

РОЗДІЛ 3. АНАЛІЗ ТА ПРОГНОЗУВАННЯ КЛІМАТИЧНИХ ЗМІН

3.1. Відбір даних для аналізу кліматичних змін

Першим етапом побудування моделі є підбір та підготовка тестового набору даних. Працюючи над дослідженням Big Data у сфері кліматичних змін варто брати до уваги той факт, що велика кількість показників та факторів, які їх спричиняють, є даними, що досліджуються великою кількістю науковців з різних частин Землі. Це спричиняє велику кількість підходів до вимірювань, використання різноманітних величин та методів числення. У такому разі, варто розуміти, що існує досить невелика кількість готових до подальшої роботи наборів даних, а більшість наявних у вільному доступі даних - є сирими, через що вимагають більшої уваги на етапі підготовки.

Для пошуку набору даних показників кліматичних змін скористаємось відкритим ресурсом даних Our World in Data [39]. Даний веб-застосунок надає доступ до широкого спектру великих даних у різних сферах, має базову аналітику даних у окремих напрямках, візуалізації, посилання на джерела тощо. Скориставшись пошуком на сайті, отримуємо доступ до даних щодо кліматичних змін, а саме до набору даних, зібраних Національним управлінням з аеронавтики і дослідження космічного простору (NASA) та Інститутом космічних досліджень Годдарда [36]. На сайті ми маємо можливість скористатись випадаючим списком і обрати показник кліматичних змін. Для прикладу оберемо місячну температурну аномалію (рисунок 3.1). На отриманому графіку ми можемо бачити, що відповідні спостереження ведуться з середини 1880 року та проводяться і на сьогоднішній день. Щодо самих показників температурних аномалій ми можемо спостерігати схожість з графіком тригонометричних функцій синуса або косинуса, що може бути пов’язано з сезонною складовою температурних змін. Однак, ми також можемо спостерігати досить різкі зміни протягом останніх двадцяти-тридцяти років, що свідчать про пришвидшення кліматичних змін і про наявність результатів впливу на екологію сторонніх факторів, таких як наслідки людської життєдіяльності.

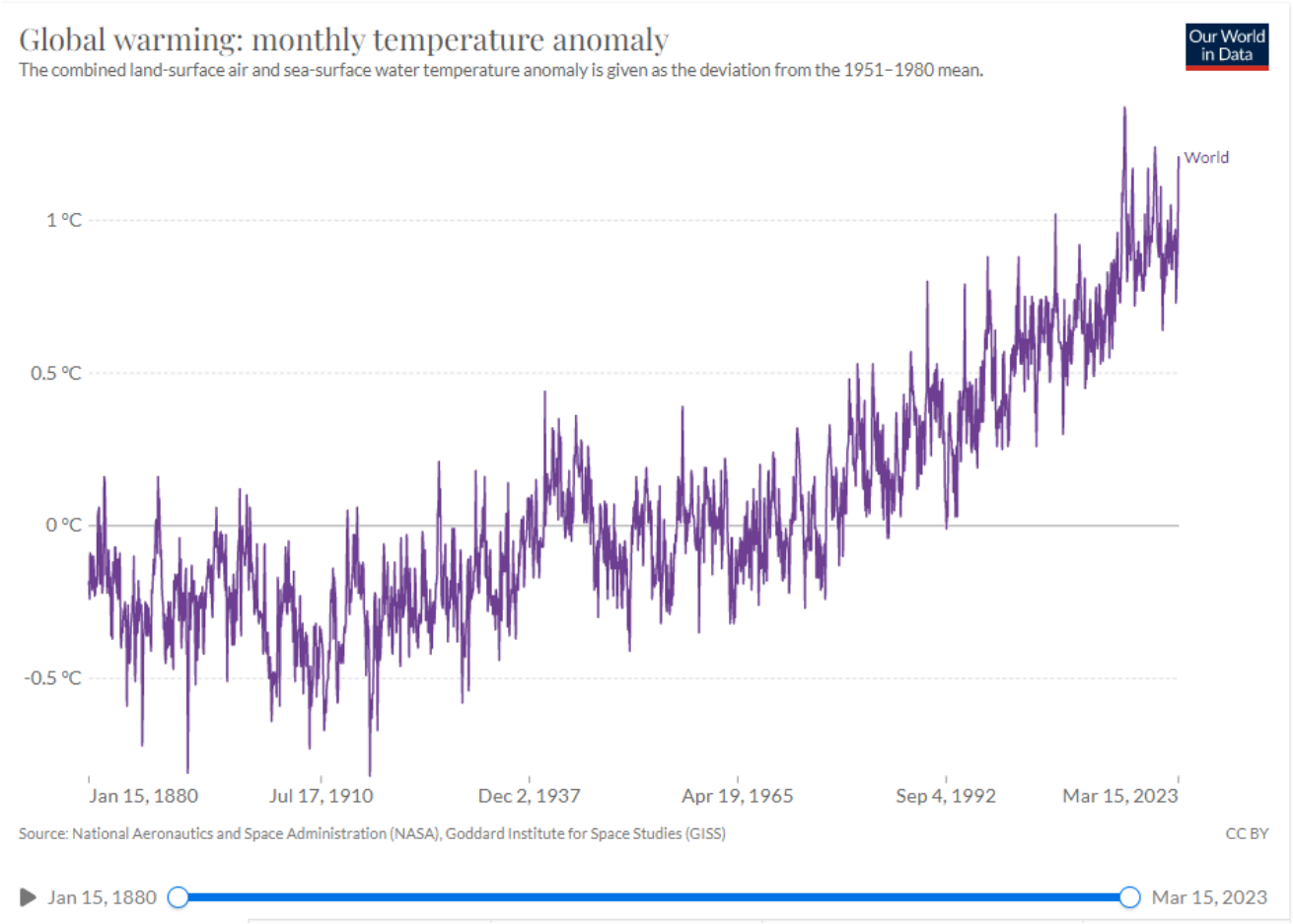

Рисунок 3.1 - Показники місячного значення температурних аномалій

Завантажимо набір даних [11] і оглянемо їх. Даний набір містить 20 колонок (таблиця 3.1), які мають відповідні значення дати, локації і кліматичних показників. Ряд кліматичних показників є зібраними з різних джерел і, відповідно, вирахувані різними методами.

Таблиця 3.1 - Склад набору даних про кліматичних змін

№	Назва колонки	Тип даних
1	Entity	object
2	Date	object
3	Seasonal variation	float64

4	Combined measurements	float64
5	Monthly averaged	float64
6	Annual averaged	float64
7	monthly_sea_surface_temperature_anomaly	float64
8	Sea surface temp (lower-bound)	float64
9	Sea surface temp (upper-bound)	float64
10	Monthly pH measurement	float64
11	Annual average	float64
12	Temperature anomaly	float64
13	Church & White	float64
14	University of Hawaii	float64
15	Average	float64
16	arctic_sea_ice_osisaf	float64
17	Monthly averaged.1	float64
18	Annual averaged.1	float64
19	Monthly averaged.2	float64
20	Annual averaged.2	float64

Розглянемо детальніше отримані нами дані. Даний набір даних містить широкий вибір метрик кліматичних змін. Наприклад, колонки Seasonal variation та Combined measurements відповідають за різні виміри крижаного покриву планети. Як ми знаємо, крижані покриви Землі відповідають за відбивання

сонячного проміння, тому їх зменшення призводить до збільшення рівню поглинання ультрафіолету і пришвидшенню процесу нагрівання. Також, зменшення кількості льодовиків призводить до збільшення рівню світового океану.

Наступні колонки monthly_sea_surface_temperature_anomaly, Sea surface temp (lower-bound), Sea surface temp (upper-bound) є показниками, що відповідають за температурні показники поверхні морів. Ці показники є надзвичайно важливими, оскільки температура світового океану, а вірніше її зростання, призводить до збільшення кількості тепла, яке світовий океан поглинає. У результаті цього відбувається теплове розширення води та підвищення рівня води. Також дані кліматичні зміни мають наслідки для морських екосистем, оскільки від температури залежать і внутрішно біологічні процеси у воді.

Іншим важливим показником кліматичних змін є океанічний pH (колонки Monthly pH measurement та Annual average) - міра кислотності (або лужності) океанічних вод. З останніх досліджень відомо[38], що океанський pH почав зменшуватися через поглинання вуглекислого газу (CO2) з атмосфери, в результаті чого змінюється хімічний склад води. Цей процес називається океанічним збуренням (Ocean acidification). Океанічне збурення може мати серйозний вплив на морські екосистеми та людське життя, оскільки змінює хімічну структуру води і погіршує умови для розвитку морських організмів.

Наступною йде колонка Temperature anomaly, яка відповідає за значення температурних аномалій. Даний показник є дуже важливим, оскільки дозволяє виявити зміни клімату та визначити наскільки вони можуть бути пов’язаними з антропогенним впливом. Завдяки даному показнику можна відслідковувати тенденції зміни температури на планеті, що є одним з показників глобального потепління.

Наступними йде група колонок, що відповідають за показники рівня води у світовому океані (колонки Church & White, University of Hawaii та Average). Ці колонки містять дослідження рівня морів за різними методиками і, відповідно, їх

середнє значення. Даний показник є надзвичайно важливим, оскільки наочно демонструє результат кліматичних змін, а саме затоплення територій суходолу.

Завершальною групою йдуть колонки, що відповідають за показники площ льодовиків. Даний показник демонструє зменшення кількості льодовиків у світовому океані. Він є важливим елементом у розумінні кліматичних змін і розробки стратегій зміцнення резистентності до змін клімату.

Основною причиною стрімкого нагрівання планети Земля є парниковий ефект. У такому випадку факторами впливу на збільшення температурних аномалій оберемо парникові гази, а саме об’єм наступних газів у атмосфері:

- діоксид вуглецю (CO2);
- метан (CH4);
- окид азоту (N2O);
- сульфур гексафлуорид (SF6).

Для отримання даних, необхідних для подальшого дослідження, скористаємось сайтом Глобальної лабораторії моніторингу (Global Monitoring Laboratory) - підрозділу Інституту Годдарда з дослідження космосу (GISS), що спеціалізуються на моніторингу та дослідженні газів, які мають вплив на клімат Землі та якість повітря. Дана лабораторія використовує різноманітні методи дослідження, включаючи вимірювання газів у повітрі, збір зразків атмосферного повітря та його аналіз в лабораторії.

Діоксид вуглецю або CO2 - це основний парниковий газ, що виробляється при згорянні вуглеводнів, таких як нафта, газ і вугілля. CO2 також викидається при процесі дихання та розкладання органічних речовин. Основним джерелом викидів CO2 є енергетика та транспорт, промисловість та сільське господарство. Скористаємось набором даних, зібраний Глобальною лабораторією моніторингу [24]. На рисунку 3.2 можемо бачити графік відповідних досліджень, що ведуться з 1980 року, на якому видно чітку лінію тренду постійного зростання кількості даного парникового газу в атмосфері Землі.

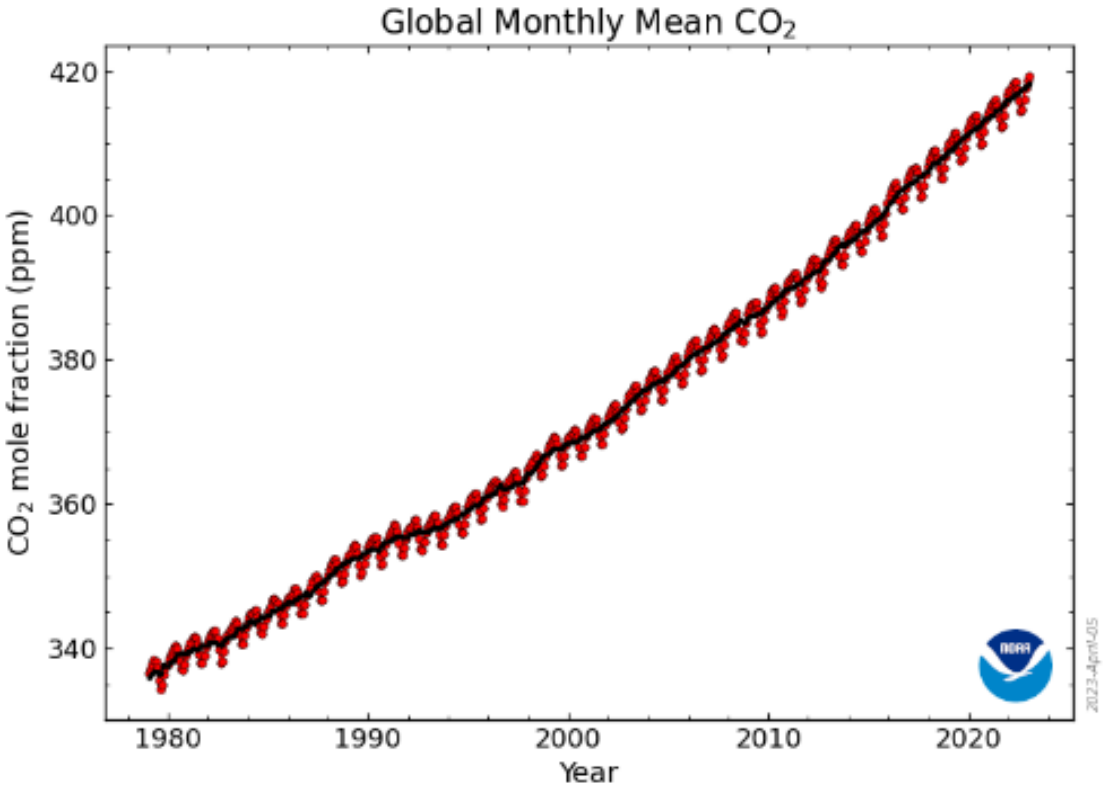

Рисунок 3.2 - Глобальні середньомісячні показники вуглекислого газу

Після завантаження, оглянемо отриманий набір даних. Даний набір даних містить 8 колонок (таблиця 3.2), яка відповідають за різні представлення дати і показників об’єму діоксиду вуглецю в атмосфері. Надалі головними даними, які будуть використовуватись з даного набору, будуть дати і середньомісячне значення кількості вуглекислого газу (колонка “average”).

Таблиця 3.2 - Загальний вид набору даних про вуглекислий газ

№	Назва колонки	Тип даних
1	year	int64
2	month	int64
3	decimal date	float64
4	average	float64

5	deseasonalized	float64
6	ndays	int64
7	sdev	float64
8	unc	float64

Метан або CH4 - це парниковий газ, який виробляється в результаті різних біологічних процесів, таких як розклад органічних речовин в забруднених водоймах, та при виділенні газу з нафтових та газових свердловин. Головними джерелами викидів CH4 є скотарство, сільське господарство, газова промисловість та звалища. За даними Глобальної лабораторії моніторингу [25], зниження швидкості зростання кількості метану в атмосфері спостерігалось лише раз за останні 35 років - у 2000-2005 роках. На жаль, подальші роки показали лише стрімке збільшення викидів даного газу в атмосферу.

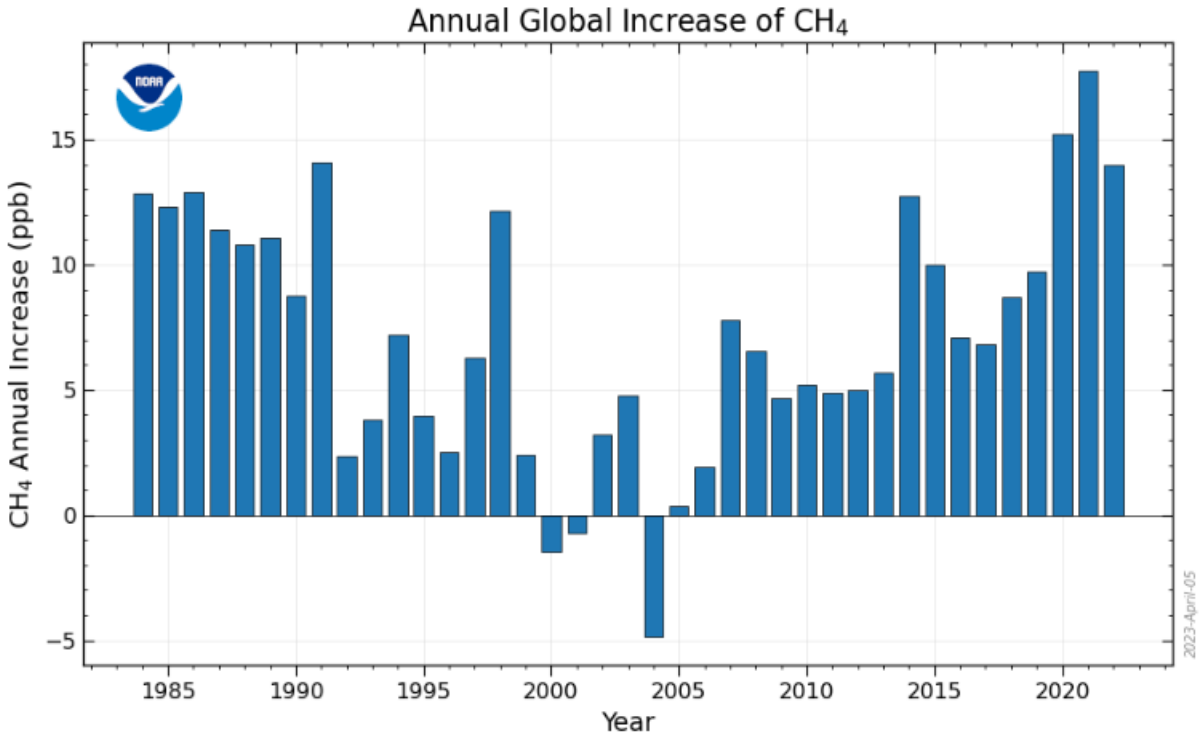

Рисунок 3.3 - Річний приріст усередненого глобального вмісту метану в атмосфері

Завантажемо та оглянемо отриманий набір даних. Набір даних складається (таблиця 3.3) з семи колонок, що містять інформацію про дату відповідно зроблених та обчислених вимірів, безпосереднє значення середньомісячної кількості парникового газу в атмосфері та його тренд. Для подальших досліджень скористаємось даними щодо дати зроблених вимір і середньомісячного значення відповідно.

Таблиця 3.3 - Загальний вид набору даних про метан

№	Назва колонки	Тип даних
1	year	int64
2	month	int64
3	decimal	float64
4	average	float64
5	average_unc	float64
6	trend	float64
7	trend_unc	float64

Наступним парниковим газом, який ми розглянемо є оксид азоту. N2O є газом, який виробляється в результаті діяльності людини, такої як використання азотних добрив у сільському господарстві та процесі виробництва азотних кислот. Іншими джерелами викидів оксиду азоту є звалища, промисловість та автомобілі. Використаємо набір даних, який був зібраний Глобальною лабораторією моніторингу [26]. На графіку, наведеному на рисунку 3.4, можемо спостерігати стабільну тенденцію до збільшення кількості даного парникового газу в атмосфері планети.

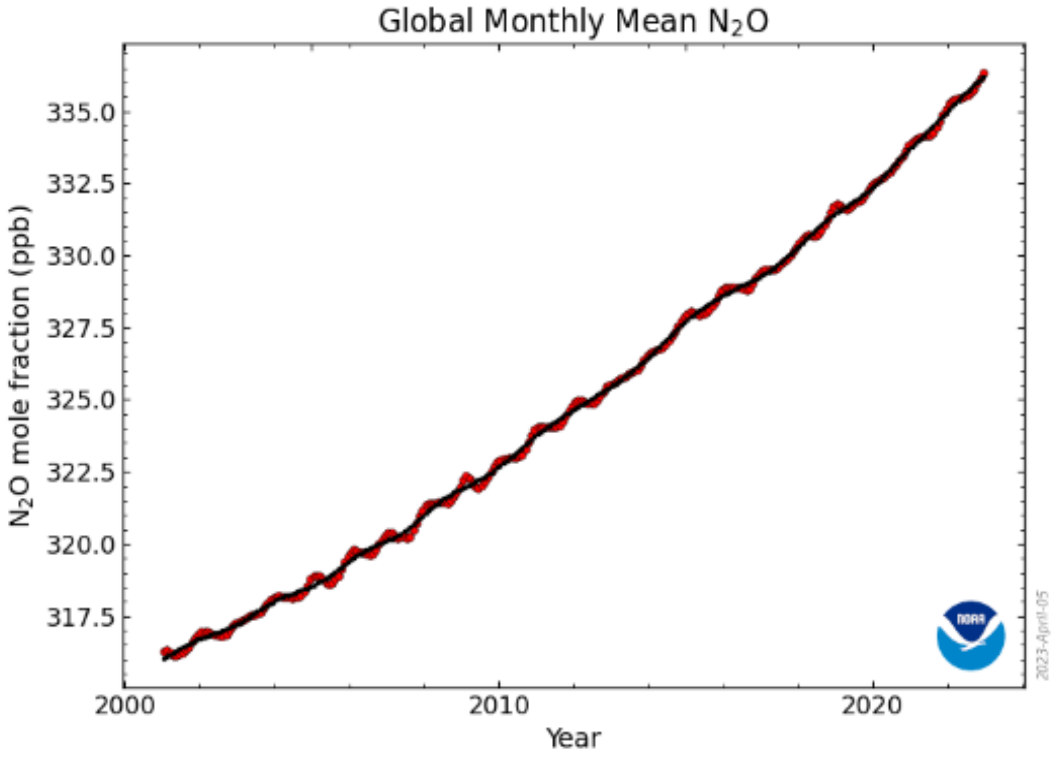

Рисунок 3.4 - Глобальні середньомісячні показники оксиду азоту

Завантажимо набір даних і проведемо попередній огляд. Як ми бачимо в таблиці 3.4, даний набір даних містить 7 колонок, частина з яких є різними видами представлення дати, інша частина являє собою набір числових показників, які відповідають за різні представлення середнього місячного об’єму оксиду азоту в атмосфері, його тренд. Для подальшого дослідження будуть використовуватись інформація щодо дати виміру і середньомісячного значення.

Таблиця 3.4 - Загальний вид набору даних про оксид азоту

№	Назва колонки	Тип даних
1	year	int64
2	month	int64
3	decimal	float64
4	average	float64

5	average_unc	float64
6	trend	float64
7	trend_unc	float64

Сульфур гексафлуорид або SF6 - це штучний газ, що використовується у різних промислових процесах, зокрема, у електроенергетичній промисловості. Він має надзвичайно високий потенціал парникового ефекту та може залишатися в атмосфері до 3200 років. Головними джерелами викидів SF6 є електростанції, трансформатори, а також процес виробництва електроніки. Скористаємось даними, наданими Глобальною лабораторією моніторингу [27]. На графіку, наведеному на рисунку 3.5, можна спостерігати невтішну тенденцію збільшення швидкості зростання кількості даного парникового газу в атмосфері Землі, що, в свою чергу, буде призводити до збільшення парникового ефекту і його негативного впливу на клімат та екологію планети.

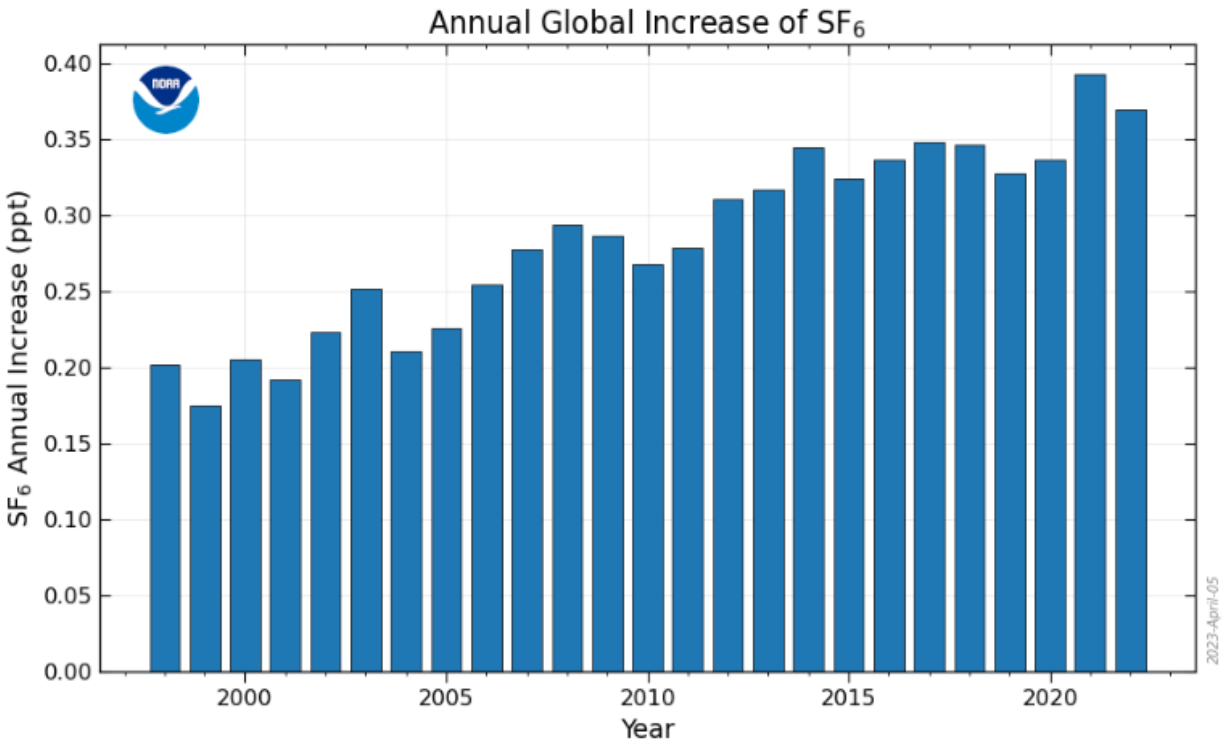

Рисунок 3.5 - Річний приріст усередненого глобального вмісту SF6 в атмосфері

Завантажимо та проведемо попередній огляд набору даних. У таблиці 3.5 наведена відповідна інформація щодо структури розглянутого нами набору. Аналогічно до попередніх, даний набір даних містить дані про дату проведення відповідних вимірювань та обчислень, середні значення та тренди. У подальших дослідженнях буде використано інформацію щодо дати та середньомісячного значення кількості сульфур гексафлуориду в атмосфері.

Таблиця 3.5 - Загальний вид набору даних про сульфур гексафлуорид

№	Назва колонки	Тип даних
1	year	int64
2	month	int64
3	decimal	float64
4	average	float64
5	average_unc	float64
6	trend	float64
7	trend_unc	float64

3.2. Обробка даних

Для побудови прогностичної моделі необхідно об’єднати попередньо обрані дані в один набір даних. Повний лістинг коду даного підрозділу знаходиться у Додатку А.

Для цього спершу імпортуємо необхідні на даному етапі бібліотеки, а саме:

import pandas as pd

Надалі зчитаємо набір даних кліматичних змін з локального файлу з розширенням .csv за допомогою методу read_csv бібліотеки pandas:

climate_change = pd.read_csv('./data/climate-change.csv')

На етапі попереднього огляду даних було виявлено, що набори даних вуглекислих газів є глобальними, у той час як набір даних з показниками кліматичних змін містять інформацію для різних регіонів, у тому числі і загальносвітові. У такому випадку є потреба у фільтруванні даного набору даних. Для цього залишимо лише ті значення, які відповідають умові:

data_temp_world = climate_change[climate_change['Entity'] == 'World']

Наступним кроком форматуємо дату для зручного подальшого використання:

result_data['Date'] = pd.to_datetime(result_data['Date'])

При огляді отриманих даних було виявлено, що наш набір даних містить порожні значення. Для уникнення можливих проблем з даними полями у майбутньому, заповнимо їх на основі інших існуючих даних. Скористаємось методом fillna() з параметром методу заповнення “ffill” (forward fill), який заповнить порожні значення з ближнього непорожнього значення, що йде перед незаповненим значенням:

data_temp_world.fillna(method="ffill", inplace=True)

Попередній код з обробки даних кліматичних змін буде винесено в окрему функцію getClimateChange() і подальший виклик цього набору даних виглядатиме наступним чином:

data_temp_world = getClimateChange()

Набори даних з інформацію щодо парникових газів мають схожу структуру, де дата є розбитою на дві окремі колонки: рік та місяць. Оскільки у попередньому наборі даних дата є єдиною колонкою з відповідним типом, приведемо поточні набори даних до відповідного виду. Для цього створемо функцію add_date_column(df), яка прийматиме параметр df - набір даних, о потребує змін і повертатиме змінений набір даних. Спершу на основі двох колонок створимо колонку з назвою Date, яка поєднає в собі колонки рік та місяць і 15 число, як середину кожного місяця (аналогічно до попереднього набору даних) і відформатує її відповідно:

df['Date'] = pd.to_datetime(df['year'].astype(str) + '-' + df['month'].astype(str) + '-15')

Надалі видалимо зайві колонки року та місяцю:

df = df.drop(columns=['year', 'month'])

Для завершення функції, повернемо відредагований набір даних:

return df

Тепер ми маємо окрему функцію, для редагування дати у наборі даних. Залишається лише зчитати набори даних і відредагувати їх відповідним чином. Для цього створимо функцію get_factor(path, delimiter, averageName), яка має параметри:

- path - шлях до файлу з розширенням .csv, у якому зберігаються дані;
- delimeter - розділювач у .csv файлах;
- averageName - ім’я, яке у подальшому буде присвоєне колонці з середньомісячним значенням кількості відповідного парникового газу в атмосфері.

Спершу зчитаємо набір даних з файлу:

factor_data = pd.read_csv(path, delimiter=delimiter)

Далі для відповідного набору даних застосуємо функцію add_date_column для проведення необхідних маніпуляцій з датою:

factor_prepared = add_date_column(factor_data)

Для подальшої роботи з даними нам потрібні лише дата та середнє значення. Оскільки надалі ці набори даних будуть зливатись в один, є необхідність перейменувати колонки average на відповідно переданий параметр averageName. Отже, отримуємо наступний код:

factor = factor_prepared.loc[:, ['Date', 'average']].rename(columns={"average": averageName})

Завершальним етапом створення цієї функції є повернення результуючого набору даних:

return factor

Тепер ми можемо обробити набори даних про парникові гази викликом однієї функції:

co2 = get_factor('./data/co2.csv', ',', 'AverageCO2')

ch4 = get_factor('./data/ch4.csv', ';', 'AverageCH4')

n2o = get_factor('./data/n2o.csv', ';', 'AverageN2O')

sf6 = get_factor('./data/n2o.csv', ';', 'AverageSF6')

Тепер наші набори даних готові для злиття. Для цього скористаємось методом merge(), передаючи параметром on поле, за яким буде відбуватись злиття:

data = data_temp_world.merge(co2, on="Date").merge(ch4, on="Date").merge(n2o, on="Date").merge(sf6, on="Date")

У результаті ми отримуємо набір даних з відповідними полями, представленими у таблиці 3.6.

Таблиця 3.6 - Типізація отриманого набору даних

№	Назва колонки	Тип даних
1	Entity	object
2	Date	object
3	Seasonal variation	float64
4	Combined measurements	float64
5	Monthly averaged	float64
6	Annual averaged	float64
7	monthly_sea_surface_temperature_anomaly	float64
8	Sea surface temp (lower-bound)	float64
9	Sea surface temp (upper-bound)	float64

10	Monthly pH measurement	float64
11	Annual average	float64
12	Temperature anomaly	float64
13	Church & White	float64
14	University of Hawaii	float64
15	Average	float64
16	arctic_sea_ice_osisaf	float64
17	Monthly averaged.1	float64
18	Annual averaged.1	float64
19	Monthly averaged.2	float64
20	Annual averaged.2	float64
21	AverageCO2	float64
22	AverageCH4	float64
23	AverageN2O	float64
24	AverageSF6	float64

Наступним кроком використаємо кореляцію для дослідження наявності зв’язку та його сили між факторами впливу та показниками кліматичних змін. Оберемо в нашому наборі да них залежні та незалежні змінні, кореляцію між якими і будемо визначати. Незалежними змінними будуть виступати показники середньомісячної кількості парникових газів в атмосфері, залежними змінними виступатимуть показники кліматичних змін. Для цього скористаємось методом corr() для побудови матриці кореляції:

corr = data.corr()

Використовуючи бібліотеки NumPy та Seaborn, побудуємо графік кореляції:

mask = np.triu(np.ones_like(corr, dtype=bool))

sns.heatmap(corr, cmap='coolwarm', annot=True, mask=mask, square=True, vmin=-1, vmax=1, center=0)

plt.show()

У результаті роботи даного коду нами був отриманий температурний графік кореляції наших змінних, зображений на рисунку 3.6. З графіку кореляції ми можемо бачити, що більшість залежних змінних мають кореляцію у межах від 0,7 до 1 за модулем, що є показником сильної кореляції між змінними.

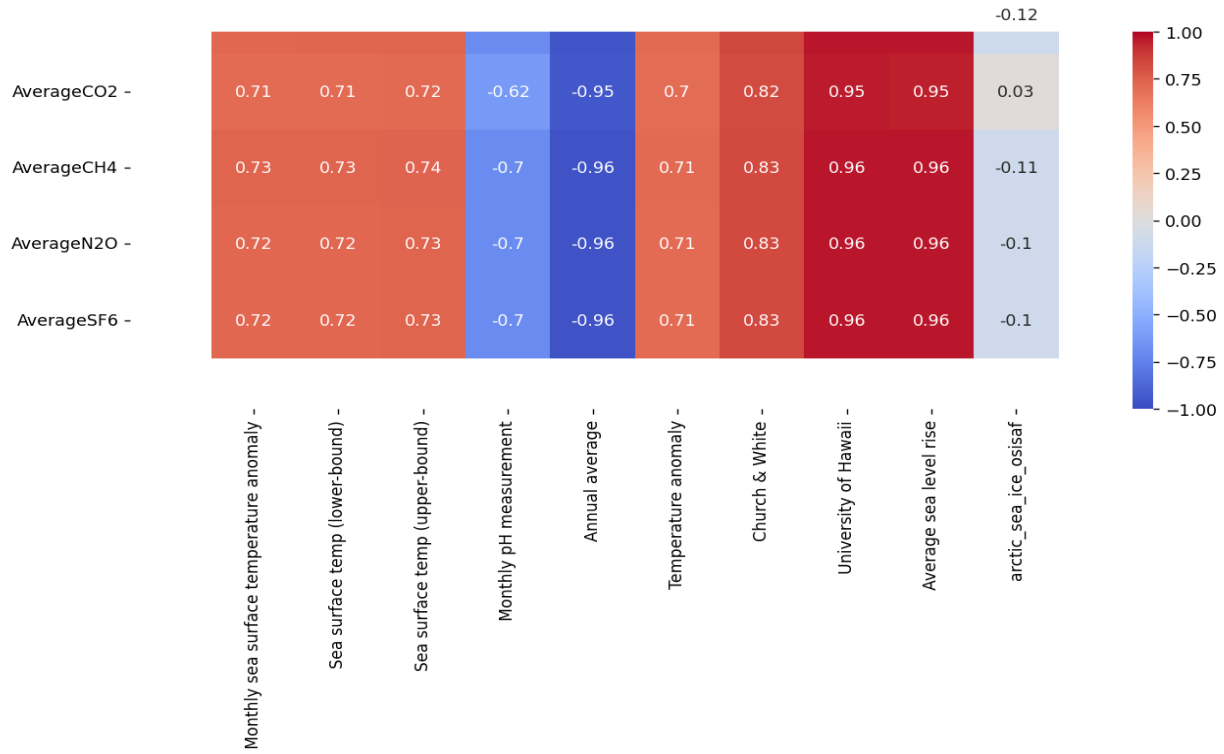

Рисунок 3.6 - Температурний графік кореляції

Розглянемо графік кореляції детальніше. Ми можемо побачити, що найсильніші коефіцієнти кореляції є між нашими чотирма незалежними змінними (даними щодо об’ємів парникових газів в атмосфері) і змінною Annual average, яка відповідає за середньорічний показник кислотності світового океану.

Оскільки даний показник є середньорічним значенням, то можемо припустити, що для подальшого моделювання дана змінна не підходить, оскільки ми маємо недостатню кількість відповідних вимірів. Для перевірки переглянемо отриманий нами набір даних і запевнимось, що середньорічних значень за двадцять років недостатньо для подальшої побудови прогностичної моделі.

Також виділимо умовну третю групу залежних змінних, яка включає показники температурних аномалій поверхні світового океану та глобальних температурних аномалій. Змінні даної групи мають коефіцієнт кореляції наближений (але не менший) 0,7, що також є сильним коефіцієнтом кореляції, що, в свою чергу, також говорить про сильну статистичну залежність.

Таким чином, у даному підрозділі була проведена обробка наборів даних, у результаті чого було отримано фінальний набір даних, готовий до подальшої побудови моделі. Також було проведено оцінку залежності незалежних та залежних змінних, у результаті якої було визначено наступний пріоритет залежних змінних:

- рівень світового океану;
- глобальні температурні аномалії;
- температурні аномалії світового океану.

3.3. Побудова та оцінка прогностичних моделей

3.3.1. Побудова моделі методом штучних нейронних мереж

Повний лістинг коду даного підрозділу знаходиться у Додатку Б.

Початковим етапом побудови прогностичної моделі є імпортування необхідних бібліотек та пакетів:

```
from tensorflow import keras
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
```

Далі ми завантажимо наш набір даних, який ми попередньо зберегли у окремому файлі з розширенням .csv:

df = pd.read_csv(data.csv')

Визначимо змінні X та y, де X - незалежні змінні, y - залежна змінна. За залежну змінну візьмемо показники рівню світового океану, оскільки дана змінна показала найсильнішу кореляцію з залежними змінними.

X = df[['AverageCO2', 'AverageCH4', 'AverageN2O', 'AverageSF6']]

y = df['Average sea level rise']

Для подальшої зручності використання винесемо в додаткову змінну дати досліджень:

dates = df['Date']

Розділимо наші дані на тренувальні та тестові набори за допомогою функції train_test_split():

X_train, X_test, y_train, y_test, dates_train, dates_test = train_test_split(X, y, dates, test_size=0.2, shuffle=False, random_state=0)

Для можливості виконання обрахунків, виконаємо операцію зміни структури даних:

X_train = X_train.reshape(-1, 4)

X_test = X_test.reshape(-1, 4)

Визначимо нашу модель:

```
model = keras.Sequential([
    keras.layers.Dense(64, activation='relu', input_shape=[4]),
    keras.layers.Dense(64, activation='relu'),
    keras.layers.Dense(1)
])
```

У наші моделі ми використовуємо два приховані шари з 64 вузлами кожен і функцією активації "relu". У нашому випадку вхідний шар має 4 вузли, тому що ми маємо 4 незалежні змінні. Вихідний шар має один вузол, що представляє нашу залежну змінну - "Average sea level rise".

Операція компілювання є обов'язковою для нейронної мережі перед навчанням. Для цього використаємо функцію compile(), щоб вказати оптимізатор, функцію втрат і метрику, які будуть використовуватись при навчанні моделі. Оптимізатор відповідає за налаштування ваг мережі, щоб зменшити втрати. Ми будемо використовувати оптимізатор Adam, який покращить швидкості та точність навчання нашої моделі. Функція втрат відображає рівень помилок між прогнозованими та фактичними значеннями. У нашому випадку, ми будемо використовувати середньоквадратичну помилку (MSE). Дана метрика вимірює ефективність моделі. Отже, маємо наступний код для компілювання моделі:

model.compile(loss='mean_squared_error', optimizer='adam', metrics=['mean_squared_error'])

Надалі ми можемо навчити модель за допомогою методу fit і передати тренувальні дані (X_train, y_train), кількість епох навчання (epochs) та розмір пакету даних для обробки (batch_size). Кожна епоха відповідає одному проходу через весь навчальний набір даних. Отримуємо відповідний код для навчання моделі:

model.fit(X_train, y_train, epochs=100, batch_size=10)

Після навчання моделі ми можемо зробити прогнози на тестових даних за допомогою методу predict:

y_pred = model.predict(X_test)

Для оцінки якості отриманої моделі, обчислимо середньоквадратичну помилку (mean squared error, MSE) моделі на тестовому наборі значень. Для цього скористаємось методом evaluate(), який обчислює значення втрат та відповідних метрик, які були попередньо передані у вигляді аргументів на етапі компіляції моделі.

test_loss, test_mse = model.evaluate(X_test, y_test, verbose=2)

print('Mean Squared Error on test set:', test_mse)

Після запуску й навчання моделі отримуємо вивід середньоквадратичної помилки:

Mean Squared Error on test set: 10.679295539855957

Також розрахуємо точність даної моделі:

*accuracy = 100 * (1 - np.abs((y_pred - y_test) / y_test)).mean()*

print('Accuracy:', accuracy)

Після відпрацювання коду, отримаємо відповідний вивід:

Accuracy: 82.3593880656721

Отже, середньоквадратична помилка побудованої моделі становить 10,7, а відсоток точності становить 82,3%. Такі показники свідчить про те, що дана модель здатна досить точно узагальнювати та передбачати значення на нових даних.

Далі для більшої наочності роботи моделі побудуємо графік, на якому відобразимо реальні та прогнозовані дані. Для побудови графіка скористаємось бібліотекою matplotlib.pyplot. Задамо дві лінії: червону для прогнозованих значень та синю для реальних:

plt.plot(dates_test, y_pred, color='red', label='Predicted')

plt.plot(dates, y, color='blue', label='Actual')

Для кращого розуміння графіку, задамо заголовок графіку та відповідні мітки для осей:

plt.title('Average sea level rise over time')

plt.xlabel('Date')

plt.ylabel('Average sea level rise')

Встановимо легенду та відображаємо графік:

plt.legend()

plt.show()

У результаті роботи даного коду, отримуємо графік (рисунок 3.7) зростання рівня світового океану, де синім кольором позначені актуальні поточні дані, а червоним - дані, прогнозовані отриманою моделлю. З даного графіку бачимо, що результати, спрогнозовані моделлю на основі штучної нейронної мережі, є доволі точними.

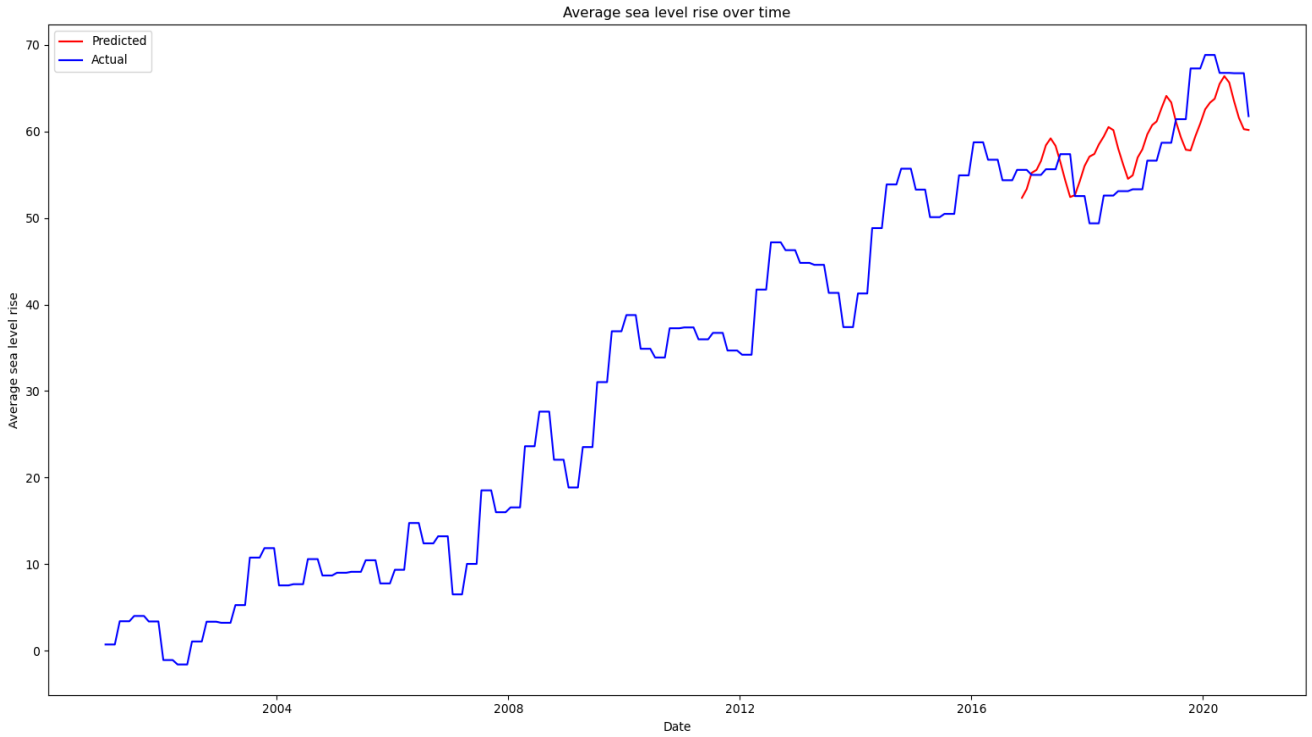

Рисунок 3.7 - Результуючий графік прогнозу рівню світового океану

Для кращої оцінки точності побудованої моделі, будуємо ще одну прогностичну модель для порівняння.

3.3.2. Побудова моделі методом рекурентної нейромережі

Побудуємо прогностичну модель методом Long Short-Term Memory. LSTM - один з методів для прогнозування часових рядів, який дозволяє враховувати залежності між значеннями в часі, інтегруючи інформацію з попередніх моментів часу.

Повний лістинг коду даного підрозділу знаходиться у Додатку В.

Розпочнімо побудову прогностичної моделі з імпорту всіх необхідних нам у подальшому бібліотек:

from tensorflow import keras

import pandas as pd

import numpy as np

from sklearn.preprocessing import MinMaxScaler

```
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score
import matplotlib.pyplot as plt
```

Завантажимо набір даних та заповнимо відсутні значення:

```
df = pd.read_csv(data.csv')
df.fillna(method="ffill", inplace=True)
```

Наступним кроком розіб’ємо дані на тренувальні та тестові набори у співвідношенні 80:20 відповідно:

```
train_size = int(len(df) * 0.8)
train_df, test_df = df[:train_size], df[train_size:]
```

За аналогією розіб’єм дати на тренувальний та тестовий набори:

```
train_dates = pd.to_datetime(train_df['Date'])
test_dates = pd.to_datetime(test_df['Date'])
```

Нормалізуємо дані за допомогою функції MinMaxScaler():

```
scaler = MinMaxScaler()
train_scaled = scaler.fit_transform(train_df[['Average sea level rise']])
test_scaled = scaler.transform(test_df[['Average sea level rise']])
```

Далі ми створюємо функцію create_dataset(), яка буде створює набори даних з використанням попередніх значень (lookback). Для кожного елемента набору даних функція створює відповідний набір даних, що містить look_back попередніх значень та відповідний цільовий показник:

```
def create_dataset(X, y, look_back=1):
    X_data, y_data = [], []
    for i in range(len(X)-look_back):
        X_data.append(X[i:(i+look_back)])
        y_data.append(y[i+look_back])
    return np.array(X_data), np.array(y_data)
```

Задаємо змінну, що визначає кількість попередніх кроків часового ряду, які будуть використовуватися для прогнозування наступного значення:

```
look_back = 12
```

Далі створимо тренувальні та тестові набори даних:

X_train, y_train = create_dataset(train_scaled, train_scaled, look_back)

X_test, y_test = create_dataset(test_scaled, test_scaled, look_back)

Створюємо LSTM модель, яка містить один LSTM шар та один Dense шар. Параметр input_shape передає форму вхідних даних, яку складається з трьох аргументів: кількість попередніх кроків (look_back), кількість ознак (1) та кількість зразків в наборі даних:

```
model = keras.Sequential([
   keras.layers.LSTM(64, input_shape=(X_train.shape[1], X_train.shape[2])),
   keras.layers.Dense(1)
])
```

Надалі компілюємо модель зі встановленням функції втрат та оптимізатора для використання під час навчання:

model.compile(loss='mean_squared_error', optimizer='adam')

Скористаємось методом fit() для навчання моделі:

history = model.fit(X_train.reshape(X_train.shape[0], X_train.shape[1], 1), y_train, epochs=100, batch_size=10, validation_data=(X_test.reshape(X_test.shape[0], X_test.shape[1], 1), y_test), verbose=2, shuffle=False)

Використовуємо метод predict() для прогнозування значень:

train_predict = model.predict(X_train)

test_predict = model.predict(X_test)

Далі виконаємо інверсію масштабування, тобто приведемо наші дані у початковий формат:

train_predict = scaler.inverse_transform(train_predict)

y_train = scaler.inverse_transform(y_train.reshape(-1, 1))

test_predict = scaler.inverse_transform(test_predict)

y_test = scaler.inverse_transform(y_test.reshape(-1, 1))

Тепер ми можемо виконувати обрахунки для оцінки отриманої моделі. Розрахуємо середньоквадратичну помилку:

train_mse = mean_squared_error(y_train, train_predict)

test_mse = mean_squared_error(y_test, test_predict)

Далі виведемо отримані величини:

print('Train MSE: %.3f' % train_mse)

print('Test MSE: %.3f' % test_mse)

І отримуємо обчислені значення:

Train MSE: 4.540

Test MSE: 3.058

Таким чином, середньоквадратична похибка моделі на тестовому наборі даних становить 3,058, що є дуже маленьким показником.

Для кращого розуміння точності моделі розрахуємо коефіцієнт детермінації отриманої моделі:

train_r2 = r2_score(y_train, train_predict)

test_r2 = r2_score(y_test, test_predict)

print('Train R2: %.3f' % train_r2)

print('Test R2: %.3f' % test_r2)

У результаті отримаємо наступний вивід:

Train R2: 0.984

Test R2: 0.900

У такому випадку, ми можемо говорити, що точність даної моделі становить 90%.

Для більшої наочності побудуємо графік:

plt.plot(train_dates[look_back:], y_train, label='Actual')

plt.plot(train_dates[look_back:], train_predict, label='Predicted')

plt.plot(test_dates[look_back:], y_test, label='Actual')

plt.plot(test_dates[look_back:], test_predict, label='Predicted')

plt.legend()

plt.show()

Рисунок 3.8 - Результуючий графік прогнозу рівню світового океану LSTM моделі

Таким чином, було побудовано графік (рисунок 3.8), на якому можна прослідкувати синім та зеленим кольором позначені тренувальні та тестові набори реальних даних рівня світового океану, у свою чергу, помаранчевим та червоним кольором позначені відповідні прогнозовані дані. На графіку наочно можна спостерігати точність прогнозів отриманої моделі.

3.3.3 Порівняння моделей

Нами було побудовано дві прогностичні моделі (надалі - модель 1 та модель 2). Для їх порівняння обчислимо так звану Root Mean Square Error (RMSE), що являє собою квадратний корінь з середньоквадратичної помилки. Даний показник часто використовують з метою прибирання зайвих квадратів у величинах, що спрощує розуміння метрик.

Таблиця 3.7 - Порівняльна таблиця моделей

Модель	MSE	RMSE
Модель 1	10,7	3,27
Модель 2	3,058	1,74

Для наочності винесемо дані, отримані з розрахунків у таблицю 3.7. З таблиці видно, що модель 2 вийшла у понад 2,5 рази точнішою. Таким чином, ми можемо зробити висновок, що побудова моделі глибокого навчання на основі LSTM рекурентної мережі виявилась точнішою.

Висновки до третього розділу

У третьому розділі дипломної роботи було проведено збір та обґрунтування обраних даних показників кліматичних змін та факторів, що на них впливають. Було проведено обробку та аналітику та візуалізацію даних, на основі чого було обрано залежні та незалежні змінні для подальшої побудови прогностичної моделі.

На основі отриманого набору даних було побудовано та оцінено прогностичну модель нейронної мережі зі штучними нейронами та прогностичну модель LSTM рекурентної нейромережі.

У результаті порівняння було визначено, що модель 2, побудована методом LSTM виявилась у рази точнішою і мала показник кореню середньоквадратичної помилки рівний 1,74, коефіцієнт детермінації 0,9. На основі цього можемо зробити висновок, що дана модель є моделлю високої точності.

РОЗДІЛ 4. ТЕХНОЛОГІЯ ЗАСТОСУВАННЯ РОЗРОБЛЕНИХ МОДЕЛЕЙ НА ДІЮЧИХ ПІДПРИЄМСТВАХ

Завдяки пророблені роботі та отримані моделі з високим показником точності можливо запропонувати кілька варіантів її застосування в межах діючих підприємств.

Умовно поділимо такі установи на дві категорії:

- установи, що будуть використовувати отриману модель як додаткове прогностичне джерело для аналізу та прогнозування наслідків кліматичних змін;
- установи, що мають у своїй компетенції роботу з економічними аспектами наслідків змін клімату.

Далі ми розглянемо детально потенційні можливості використання отриманої моделі у підприємствах зазначених категорій.

4.1. Застосування розробленої моделі на підприємствах екологічно-моніторингового напряму

Підприємства та організації, що займаються моніторинговою та аналітичною роботою із даними, що стосуються рівня світового океану можуть використовувати отриману мною модель в контексті прогнозування змін рівня світового океану та його вплив на інші аспекти змін клімату.

В сучасному світі проблеми, пов'язані зі зміною клімату, стають все більш актуальними і нагальними. Один з найбільш серйозних викликів, пов'язаних зі зміною клімату, полягає у підвищенні рівня світового океану. Це проблема, яка повинна бути вирішена якомога швидше, оскільки зростання рівня світового океану може мати дуже серйозні наслідки для економіки, національної безпеки, екології та інших аспектів життя людей.

Одним із способів вирішення цієї проблеми є моніторинг та аналітична робота із даними, що стосуються рівня світового океану. Багато підприємств та

організацій вже займаються цією роботою, але важливо мати надійну модель для прогнозування змін рівня світового океану та його впливу на інші аспекти змін клімату.

Для досягнення мети щодо прогнозування змін рівня світового океану та його впливу на інші аспекти змін клімату, можна використовувати різноманітні методики та інструменти. Наприклад, для аналізу даних про рівень світового океану можна використовувати математичні моделі, статистичний аналіз даних, інтерполяцію та екстраполяцію даних, машинне навчання та інші підходи.

Щодо визначення впливу рівня світового океану на інші аспекти змін клімату, то можна використовувати глобальні кліматичні моделі, які дозволяють враховувати взаємодію між різними компонентами клімату та прогнозувати їх зміни в майбутньому. Такі моделі можуть бути використані для прогнозування змін температури, атмосферного тиску, вітру та інших параметрів клімату, які можуть бути пов'язані зі зміною рівня світового океану.

Додатково, для аналізу впливу змін рівня світового океану на біосферу та соціально-економічні процеси, можна використовувати моделі взаємодії між людиною та природним середовищем, які дозволяють враховувати вплив різних факторів на екосистеми та людську діяльність.

Важливо відзначити, що для отримання точних результатів прогнозування необхідно мати доступ до великої кількості якісних даних. Оскільки дані про рівень світового океану є динамічними та можуть змінюватися з часом, необхідно забезпечувати постійне оновлення та перевірку даних

Модель прогнозування рівня світового океану є надзвичайно важливою для компаній та організацій, що займаються моніторингом та аналітичною роботою з даними змін клімату. Рівень світового океану є одним з ключових показників кліматичних змін, і його зростання може мати серйозні наслідки для природних екосистем, а також для людей, які проживають на берегах океанів та морів.

Компанії та організації, які використовують модель прогнозування рівня світового океану, можуть бути в кращому становищі для оцінки потенційних

наслідків зміни рівня морів, таких як повені, затоплення, ерозія берегів та інші проблеми. За допомогою цієї моделі, компанії можуть допомогти укладенню угод та ухваленню рішень на міжнародному рівні, які спрямовані на зниження викидів парникових газів та обмеження зростання рівня світового океану. Крім того, ця модель може бути корисною для визначення областей, де потрібні додаткові заходи адаптації до зміни рівня морів, щоб зменшити наслідки для людей та екосистем.

Загалом, модель прогнозування рівня світового океану є незамінним інструментом для компаній та організацій, що займаються моніторингом та аналітичною роботою з даними змін клімату. Вона допомагає оцінити наслідки змін рівня морів та визначити необхідні заходи для зменшення їх впливу на природу та людей.

4.2. Застосування розробленої моделі на підприємствах комерційного спрямування

Моніторинг рівня світового океану є надзвичайно важливим для компаній, які займаються страхуванням в прибережній зоні. Зміна рівня моря може спричинити повені та затоплення, що можуть викликати значні збитки для страховиків та їх клієнтів. Оскільки кліматичні зміни дедалі більше впливають на рівень світового океану, страхові компанії мають бути уважні із відстеженням цього процесу та використовувати ці дані для прогнозування ризиків та встановлення прийнятних тарифів для страхових продуктів.

Крім того, моніторинг рівня світового океану може мати значний вплив на розвиток нових торгових маршрутів. Зміна рівня моря може призвести до затоплення певних територій та утворення нових водних шляхів для торгівлі. Наприклад, зміна рівня моря може вплинути на доступність транспорту для відвантаження та отримання товарів, що може вплинути на вартість товарів та логістику. Також, затоплення деяких територій може призвести до необхідності збудувати нові інфраструктури та об'єкти, що може відкрити нові можливості

для підприємств та компаній, що займаються будівництвом та розвитком інфраструктури.

Для компаній, які займаються морським страхуванням та мають інтереси в прибережній зоні, моніторинг рівня світового океану є ключовим компонентом стратегії ризик-менеджменту. Затоплення портів та інфраструктури прибережних територій можуть завдати значної шкоди як людському життю, так і майновій сфері, тому забезпечення відповідного страхового покриття є важливим завданням для компаній, які займаються морським страхуванням. Моніторинг рівня світового океану дозволяє їм змінювати політику страхування та розробляти нові страхові продукти, які враховують потенційні ризики залежно від рівня моря.

Зростання рівня світового океану може мати також позитивні наслідки для міжнародної торгівлі. Високі рівні моря можуть відкривати нові торгові маршрути через затоплені території, що може бути корисним для компаній, які займаються логістикою та транспортуванням. Зокрема, різке зниження рівня льодовиків на Арктиці в результаті глобального потепління відкрило нові можливості для корабельної транспорта в цьому регіоні. Проте, разом з новими можливостями, з'являються нові ризики та виклики, пов'язані з безпекою мореплавання та екологічними наслідками, що можуть виникнути внаслідок затоплення нових територій. Тому, для компаній, які займаються міжнародною торгівлею, моніторинг рівня світового океану є важливим елементом ризик-менеджменту та планування бізнес-стратегії.

4.3. Потенціал розвитку проєкту

Подальший розвиток моделі прогнозування рівня світового океану потребує постійного збільшення кількості показників, які впливають на прогнозування змін. Зростання кількості факторів, які враховуються при створенні моделей прогнозування, забезпечує більш точні прогнози змін рівня світового океану. Окрім того, включення до моделі нових показників дозволяє

виявити залежності та взаємозв'язки між різними факторами, що впливають на зміну рівня світового океану.

Крім того, важливим аспектом розвитку моделі прогнозування рівня світового океану є вдосконалення алгоритмів обробки отриманих даних. Нові алгоритми мають бути більш точними та ефективними, щоб забезпечити якісний та оперативний прогноз змін рівня світового океану.

Важливим аспектом розвитку моделі прогнозування є постійний моніторинг та аналіз отриманих даних. Це дозволяє не тільки оновлювати модель та удосконалювати її алгоритми, але і надавати оперативну інформацію страховим компаніям та іншим зацікавленим сторонам. Також, регулярний моніторинг рівня світового океану дозволяє оперативно вживати заходів у випадку погіршення ситуації, що може запобігти значним збиткам та втратам.

Отже, можна зробити висновок, що розвиток моделі прогнозування рівня світового океану є важливим завданням для компаній та організацій, що займаються моніторингом та аналізом змін клімату. Вдосконалення моделі дозволяє забезпечити більш точний та оперативний прогноз змін рівня світового океану, що може бути корисним для страхових компаній, торгових компаній та інших зацікавлених сторін. Крім того, розвиток моделі може допомогти в захисті людей та зменшенні наслідків негативних змін клімату.

Розвиток моделі прогнозування рівня світового океану має значний потенціал для подальшого розвитку. Одним з найбільш важливих напрямків розвитку моделі є оновлення методів отримання актуальних даних, що дозволить покращити оновлення моделі в динамічно змінюваному середовищі. Наприклад, розширення мережі датчиків температури та солоності океану відкриє нові можливості для збору даних та внесення їх до моделі. Додавання нових даних також дозволить покращити точність прогнозів, що зробить модель привабливішою для компаній та організацій, що займаються моніторингом змін клімату.

Ще одним важливим напрямком розвитку моделі є розширення через збільшення кількості факторів, що впливають на зміни в моделі. Наприклад, до

досліджень можна додати нові показники кліматичних змін, що впливають на прогнозування змін світового океану. Такі фактори можуть включати в себе деякі віддалені силові впливи, такі як зміни вітру або течій, а також більш очевидні впливи, такі як зміни температури та атмосферного тиску. Додавання нових факторів дозволить моделі бути більш гнучкою та точною, що значно полегшить прогнозування змін рівня світового океану.

Пошукова технологія може використовуватись для збору та аналізу даних з інтернет-джерел. Це може бути корисним при підтримці моделі прогнозування рівня світового океану, оскільки Інтернет містить значну кількість даних про кліматичні зміни та рівень моря. Застосування пошукових технологій для збору цих даних може бути корисним, оскільки це дозволить отримувати більше інформації про рівень моря та інші фактори, що впливають на нього.

Крім того, розвиток моделі прогнозування може бути корисним для інших галузей, включаючи компанії, що займаються океанською торгівлею та перевезеннями. Одним з можливих застосувань моделі може бути прогнозування можливих затоплених територій та змін у морських шляхах, що може впливати на швидкість перевезення товарів та збільшення витрат на паливо.

Крім того, компанії страхування можуть використовувати модель для оцінки ризику затоплення в прибережних зонах та встановлення відповідної страхової премії. Це може бути корисно як для компаній, що надають страхові послуги, так і для мешканців прибережних зон, які зможуть звернутися за страховим захистом на випадок затоплення.

Важливо також зазначити, що окрім подальшого розширення і вдосконалення безпосередньої моделі, вона має і інший потенціал технічного розвитку, а саме - використання моделі як частини інших проєктів у сфері інформаційних технологій. До прикладу такого проєкту може слугувати застосунок з 3Д моделлю планети. У поєднанні такого проєкту з отриманою моделлю можна наочно прогнозувати і візуально відстежувати на зручній обємній графіці ділянки суходолу, які потенційно будуть покриті водою. Це

спростить розуміння результатів прогнозування і дозволить їх використовувати більш широкому колу людей.

Отже, розвиток моделі прогнозування рівня світового океану є важливим кроком у розвитку наукового та технічного прогресу, а також вирішенні питань, пов'язаних з кліматичними змінами та природно-екологічними проблемами.

Висновки до четвертого розділу

У четвертому розділі кваліфікаційної роботи магістра було проведено огляд практичного застосування отриманих прогностичних моделей кліматичних змін на підприємствах. Серед потенційних установ-користувачів моделі було виділено два основні напрямки, а саме: установи, що займаються безпосереднім моніторингом кліматичних змін та наслідків їх впливу на планету та людство; комерційні підприємства.

Серед комерційних підприємств було виділено галузь страхування нерухомості, що може використовувати отримані прогнози щодо підвищення рівня світового океану з метою планування страхових політик у відповідних регіонах.

Говорячи про некомерційні установи, основною метою яких є моніторинг екологічних змін, то такі організацію широко використовують прогностичні моделі у даній сфері з метою прогнозування загроз господарствам, флорі та фауні. Дані організації тісно співпрацюють з міжнародними органами та урядами країн з метою підготовки до можливих катаклізм і боротьби з ними.

Також у даному розділі було розглянуто технічних потенціал до розвитку поточних моделей. Серед них було виділено постійне оновлення даних і пошук нових джерел даних; розширення моделі, а саме збільшення кількості факторів, що впливають на кліматичні зміни, а також потенційно можливе розширення показників кліматичних змін; застосування нових методів моделювання з метою вдосконалення моделі. Ще одним важливим потенційним шляхом розвитку

поточної моделі було визначено її поєднання з сучасними технологіями, такими як 3Д моделювання поверхні Землі.

ВИСНОВКИ

У кінці дослідження підіб’ємо підсумки. У першому розділі роботи було проведено інформаційно-літературний огляд проблематики кліматичних змін та сучасного стану дослідження цієї проблеми. За результатами отриманої та проаналізованої інформації можна зробити висновок про беззаперечність процесу змін клімату, шкідливий вплив індустріалізації та іншої людської діяльності. Однією з основних причин цього процесу є збільшення викидів парникових газів та їх накопичення в атмосфері планети, що впливає на глобальні екологічні процеси, такі як підвищення середньої температури, танення льодовиків, збільшення рівня світового океану, зміна берегової лінії та зменшення біологічного різноманіття.

У другому розділі було проведено огляд методологій управління проєктом в галузі інформаційної аналітики даних. На основі цього огляду була обрана CRISP-DM як методологія управління проєктом з інформаційної аналітики та прогнозування кліматичних змін. Також було проаналізовано переваги та недоліки різних методів аналізу даних, і на цій основі були вибрані методи штучних нейронних мереж (ANN) та метод короткотривалої пам'яті (LSTM) для подальшого інформаційного аналізу та прогнозування даних в галузі кліматичних змін.

У третьому розділі було проведено збір та обґрунтування вибраних даних про показники кліматичних змін та факторів, що на них впливають. Була проведена обробка даних, аналітична вибірка та візуалізація, що дало змогу вибрати залежні та незалежні змінні для подальшої побудови прогностичної моделі.

На основі набору даних було побудовано та оцінено прогностичну модель нейронної мережі зі штучними нейронами та прогностичну модель LSTM рекурентної нейромережі. Шляхом порівняння було визначено, що модель 2, побудована за допомогою методу LSTM, виявилась значно точнішою і мала

корінь середньоквадратичної помилки рівний 1,74 та коефіцієнт детермінації 0,9. Це свідчить про високу точність даної моделі.

У четвертому розділі було проведено огляд практичного застосування отриманих прогностичних моделей кліматичних змін у підприємствах. Серед потенційних користувачів моделі були виділені два основні напрямки: установи, що займаються безпосереднім моніторингом кліматичних змін та їх наслідків на планету та людство, і комерційні підприємства.

Серед комерційних підприємств була виділена галузь страхування нерухомості, яка може використовувати прогнози щодо підвищення рівня світового океану для планування страхових політик у відповідних регіонах. Щодо некомерційних установ, основною метою яких є моніторинг екологічних змін, прогностичні моделі широко використовуються для прогнозування загроз господарствам, флорі та фауні. Зазначені організації тісно співпрацюють з міжнародними органами та урядами країн з метою підготовки до можливих катастроф і боротьби з ними.

Додатково, у цьому розділі було розглянуто потенційні напрямки розвитку поточних моделей. Зокрема, було виокремлено такі шляхи розширення технічного потенціалу: постійне оновлення даних та пошук нових джерел інформації; розширення моделі шляхом включення додаткових факторів, що впливають на кліматичні зміни; застосування нових методів моделювання для поліпшення точності моделей. Крім того, важливим напрямом розвитку є поєднання поточної моделі з сучасними технологіями, зокрема 3D моделюванням поверхні Землі.

Відповідно до поставленого завдання було проаналізовано теоретичну базу проблемної області та застосування інформаційної аналітики в сфері дослідження змін клімату, проведено аналіз методів та методик інформаційного аналізу та моделювання процесів та програмні засоби реалізації, відібрано та стандартизовано досліджувані дані, спроектовано та реалізовано прогностичну модель, відображено можливі імплементації отриманої моделі на підприємствах різного господарського типу.

На підставі вищевикладеного можна зробити висновок, що магістерська робота є актуальною і відповідає науковим і практичним проблемам. Розроблена прогностична модель може бути застосована в широкому спектрі підприємств. Завершуючи свою роботу, можу заявити, що всі поставлені цілі були досягнуті повністю.

ПЕРЕЛІК ВИКОРИСТАНИХ ІНФОРМАЦІЙНИХ ДЖЕРЕЛ:

1. 6.3. Preprocessing data. scikit-learn. URL: https://scikit-learn.org/stable/modules/preprocessing.html (дата звернення: 14.05.2023).
2. 9 Of The Most Popular Project Management Methodologies Made Simple. The Digital Project Manager. URL: https://thedigitalprojectmanager.com/projects/pm-methodology/project-management-methodologies-made-simple/ (дата звернення: 14.05.2023).
3. About Our Analyses. National Snow and Ice Data Center. URL: https://nsidc.org/news-analyses/scientific-analyses/about-our-analyses (дата звернення: 08.05.2023).
4. Air Pollution. Our World in Data. URL: https://ourworldindata.org/air-pollution#how-are-death-rates-from-air-pollution-changing (дата звернення: 08.05.2023).
5. Albedo. URL: https://en.wikipedia.org/wiki/Albedo (дата звернення: 08.05.2023).
6. Artificial Neural Network Tutorial - Javatpoint. www.javatpoint.com. URL: https://www.javatpoint.com/artificial-neural-network (дата звернення: 14.05.2023).
7. Autoencoders in Deep Learning: Tutorial & Use Cases [2023]. V7 - AI Data Platform for Computer Vision. URL: https://www.v7labs.com/blog/autoencoders-guide (дата звернення: 14.05.2023).
8. Borunda A. Arctic summer sea ice could be gone by as early as 2035. Science. URL: https://www.nationalgeographic.com/science/article/arctic-summer-sea-ice-could-be-gone-by-2035 (дата звернення: 08.05.2023).
9. Bryson, R. A., & Hoare, J. E. (1952). The use of the computer in analysis of climate data. Journal of Meteorology, 9(2), 75-81.
10. C++ Overview. Online Courses and eBooks Library. URL: https://www.tutorialspoint.com/cplusplus/cpp_overview.htm (дата звернення: 14.05.2023).

11. Climate Change Impacts Data Explorer. Our World in Data. URL: https://ourworldindata.org/explorers/climate-change?facet=none&pickerSort=asc&pickerMetric=entity&Metric=Arctic+sea+ice+extent&Long-run+series?=false&country=OWID_WRL~Gulkana+Glacier~Lemon+Creek+Glacier~North+America~South+Cascade+Glacier~Wolverine+Glacier (дата звернення: 19.04.2023).
12. Cluster Analysis. URL: https://byjus.com/maths/cluster-analysis/ (дата звернення: 19.04.2023).
13. Community Atmosphere Model (CAM) | Community Earth System Model. Home | Community Earth System Model. URL: https://www.cesm.ucar.edu/models/cam (дата звернення: 08.05.2023).
14. CS 230 - Recurrent Neural Networks Cheatsheet. Stanford University. URL: https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks (дата звернення: 14.05.2023).
15. CS 230 - Recurrent Neural Networks Cheatsheet. Stanford University. URL: https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks#architecture (дата звернення: 14.05.2023).
16. Decision Tree Algorithm in Machine Learning - Javatpoint. www.javatpoint.com. URL: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm (дата звернення: 14.05.2023).
17. Discriminant Analysis: A Complete Guide. Digital Vidya. URL: https://www.digitalvidya.com/blog/discriminant-analysis/ (дата звернення: 14.05.2023).
18. Frequently asked questions on climate change and disaster displacement | UNHCR UK. UNHCR UK. URL: https://www.unhcr.org/uk/news/stories/frequently-asked-questions-climate-change-and-disaster-displacement (дата звернення: 08.05.2023).
19. Gandhi R. Support Vector Machine – Introduction to Machine Learning Algorithms. Medium. URL: https://towardsdatascience.com/support-vector-

machine-introduction-to-machine-learning-algorithms-934a444fca47 (дата звернення: 14.05.2023).
20. Gill N. S. Artificial Neural Networks Applications and Algorithms. Continuous Intelligence with Real Time AI. URL: https://www.xenonstack.com/blog/artificial-neural-network-applications (дата звернення: 14.05.2023).
21. GISS GCM ModelE. URL: https://www.giss.nasa.gov/tools/modelE/ (дата звернення: 14.05.2023).
22. Global Atmospheric Research Program (GARP) Records. ArchivesSpace Public Interface |. URL: https://aspace.archives.ucar.edu/repositories/2/resources/23 (дата звернення: 08.05.2023).
23. Global Atmospheric Research Program. URL: https://www.sciencedirect.com/topics/earth-and-planetary-sciences/global-atmospheric-research-program (дата звернення: 08.05.2023).
24. Global Monitoring Laboratory - Carbon Cycle Greenhouse Gases. NOAA ESRL Global Monitoring Laboratory. URL: https://gml.noaa.gov/ccgg/trends/data.html (дата звернення: 19.04.2023).
25. Global Monitoring Laboratory - Carbon Cycle Greenhouse Gases. NOAA ESRL Global Monitoring Laboratory. URL: https://gml.noaa.gov/ccgg/trends_ch4/ (дата звернення: 19.04.2023).
26. Global Monitoring Laboratory - Carbon Cycle Greenhouse Gases. NOAA ESRL Global Monitoring Laboratory. URL: https://gml.noaa.gov/ccgg/trends_n2o/ (дата звернення: 19.04.2023).
27. Global Monitoring Laboratory - Carbon Cycle Greenhouse Gases. NOAA ESRL Global Monitoring Laboratory. URL: https://gml.noaa.gov/ccgg/trends_sf6/ (дата звернення: 19.04.2023).
28. Home | Community Earth System Model. Home | Community Earth System Model. URL: https://www.cesm.ucar.edu/ (дата звернення: 08.05.2023).

29. Indicator Metadata Registry Details. World Health Organization (WHO). URL: https://www.who.int/data/gho/indicator-metadata-registry/imr-details/158 (дата звернення: 08.05.2023).
30. Introduction to Artificial Neural Networks. Analytics Vidhya. URL: https://www.analyticsvidhya.com/blog/2021/09/introduction-to-artificial-neural-networks/ (дата звернення: 14.05.2023).
31. Java - Overview. Online Courses and eBooks Library. URL: https://www.tutorialspoint.com/java/java_overview.htm (дата звернення: 14.05.2023).
32. Machine learning, explained | MIT Sloan. MIT Sloan. URL: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained (дата звернення: 14.05.2023).
33. MATLAB - Overview. Online Courses and eBooks Library. URL: https://www.tutorialspoint.com/matlab/matlab_overview.htm (дата звернення: 14.05.2023).
34. Matplotlib – Visualization with Python. Matplotlib – Visualization with Python. URL: https://matplotlib.org/ (дата звернення: 14.05.2023).
35. NASA News & Feature Releases.URL: https://www.giss.nasa.gov/research/news/20170118/ (дата звернення: 14.05.2023).
36. National Aeronautics and Space Administration Goddard Institute for Space Studie. URL: https://data.giss.nasa.gov/gistemp/ (дата звернення: 14.05.2023).
37. NumPy. NumPy. URL: https://numpy.org/ (дата звернення: 14.05.2023).
38. Ocean Acidification: The Other CO2 Problem / Scott C. Doney та ін. Annual Review of Marine Science. 2009. Т. 1, № 1. С. 169–192. URL: http://oceans.mit.edu/wp-content/uploads/doney_ann_rev_proof.pdf.
39. Our World in Data. Our World in Data. URL: https://ourworldindata.org/ (дата звернення: 18.04.2023).
40. pandas - Python Data Analysis Library. pandas - Python Data Analysis Library. URL: https://pandas.pydata.org/ (дата звернення: 14.05.2023).

41. Principal component analysis: a review and recent developments. URL: https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202 (дата звернення: 14.05.2023).
42. Python - Overview. Online Courses and eBooks Library. URL: https://www.tutorialspoint.com/python/python_overview.htm (дата звернення: 14.05.2023).
43. R - Overview. Online Courses and eBooks Library. URL: https://www.tutorialspoint.com/r/r_overview.htm (дата звернення: 14.05.2023).
44. Regression Analysis. Corporate Finance Institute. URL: https://corporatefinanceinstitute.com/resources/data-science/regression-analysis/ (дата звернення: 14.05.2023).
45. Singh P. Adapting Project Management Methodologies to Data Science. Medium. URL: https://towardsdatascience.com/adapting-project-management-methodologies-to-data-science-a710ac9872ea (дата звернення: 14.05.2023).
46. TensorFlow. TensorFlow. URL: https://www.tensorflow.org/ (дата звернення: 14.05.2023).
47. The Sequential model | TensorFlow Core. TensorFlow. URL: https://www.tensorflow.org/guide/keras/sequential_model (дата звернення: 14.05.2023).
48. Time Series Analysis and Forecasting | Data-Driven Insights (Updated 2023). Analytics Vidhya. URL: https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-to-time-series-analysis/ (дата звернення: 14.05.2023).
49. USGCRP Indicator Details. GlobalChange.gov. URL: https://www.globalchange.gov/browse/indicators/arctic-sea-ice-extent (дата звернення: 08.05.2023).
50. Weather Research & Forecasting Model (WRF) | Mesoscale & Microscale Meteorology Laboratory. Home | Mesoscale & Microscale Meteorology Laboratory. URL: https://www.mmm.ucar.edu/models/wrf (дата звернення: 08.05.2023).

51. Welcome to Python.org. Python.org. URL: https://www.python.org/ (дата звернення: 14.05.2023).
52. What are Recurrent Neural Networks? | IBM. IBM - Deutschland | IBM. URL: https://www.ibm.com/topics/recurrent-neural-networks (дата звернення: 14.05.2023).
53. What is Bayesian Analysis? | International Society for Bayesian Analysis. International Society for Bayesian Analysis | The International Society for Bayesian Analysis (ISBA) was founded in 1992 to promote the development and application of Bayesian analysis. URL: https://bayesian.org/what-is-bayesian-analysis/ (дата звернення: 14.05.2023).
54. What is CRISP DM? - Data Science Process Alliance. Data Science Process Alliance. URL: https://www.datascience-pm.com/crisp-dm-2/ (дата звернення: 14.05.2023).
55. What is SEMMA? - Data Science Process Alliance. Data Science Process Alliance. URL: https://www.datascience-pm.com/semma/ (дата звернення: 14.05.2023).
56. What is the Team Data Science Process? - Azure Architecture Center. Microsoft Learn: Build skills that open doors in your career. URL: https://learn.microsoft.com/en-us/azure/architecture/data-science-process/overview (дата звернення: 14.05.2023).
57. Згорткові нейронні мережі (CNN): Вступ - techukraine.net. techukraine.net. URL: https://techukraine.net/згорткові-нейронні-мережі-cnn-вступ/ (дата звернення: 14.05.2023).
58. Зміна клімату в Україні та світі: причини, наслідки та рішення для протидії. Екодія. URL: https://ecoaction.org.ua/zmina-klimatu-ua-ta-svit.html (дата звернення: 08.05.2023).
59. Результати дослідження глобального тягаря хвороб в Україні | Центр громадського здоров'я. Центр громадського здоров'я України | МОЗ. URL: https://phc.org.ua/news/rezultati-doslidzhennya-globalnogo-tyagarya-khvorob-v-ukraini#:~:text=Дослідження%20глобального%20тягаря%20хвороб%20–

%20це,в%20204%20країнах%20та%20територіях (дата звернення: 08.05.2023).
60. Тенденції розвитку Front-end інструментів у сучасній веброзробці / О. С. Борсук, Д. В. Дуля, М. В. Пирог. // Матеріали IV Всеукраїнської науково-практичної інтернет-конференції студентів, аспірантів та молодих вчених за тематикою «Сучасні комп'ютерні системи та мережі в управлінні»: збірка наукових праць. – 2021. – С. 16–18.
61. Українські науковці вважають цьогорічні підтоплення проявами змін клімату. GrowHow.in.ua. URL: https://www.growhow.in.ua/ukrainski-naukovtsi-vvazhaiut-tsohorichni-pidtoplennia-proiavamy-zmin-klimatu/ (дата звернення: 08.05.2023).
62. Учасники проектів Вікімедіа. Довга короткочасна пам'ять – Вікіпедія. Вікіпедія. URL: https://uk.wikipedia.org/wiki/Довга_короткочасна_пам'ять (дата звернення: 14.05.2023).

ДОДАТКИ

Додаток А.

Лістинг коду обробки та візуалізації даних

```
import pandas as pd
climate_change = pd.read_csv('./data/climate-change.csv')
data_temp_world = climate_change[climate_change['Entity'] == 'World']
result_data['Date'] = pd.to_datetime(result_data['Date'])
data_temp_world.fillna(method="ffill", inplace=True)
data_temp_world = getClimateChange()
df['Date'] = pd.to_datetime(df['year'].astype(str) + '-' + df['month'].astype(str) + '-15')
df = df.drop(columns=['year', 'month'])
return df
factor_data = pd.read_csv(path, delimiter=delimiter)
factor_prepared = add_date_column(factor_data)
factor = factor_prepared.loc[:, ['Date', 'average']].rename(columns={"average": averageName})
return factor
co2 = get_factor('./data/co2.csv', ',', 'AverageCO2')
ch4 = get_factor('./data/ch4.csv', ';', 'AverageCH4')
n2o = get_factor('./data/n2o.csv', ';', 'AverageN2O')
sf6 = get_factor('./data/n2o.csv', ';', 'AverageSF6')
data = data_temp_world.merge(co2, on="Date").merge(ch4, on="Date").merge(n2o, on="Date").merge(sf6, on="Date")
corr = data.corr()
mask = np.triu(np.ones_like(corr, dtype=bool))
sns.heatmap(corr, cmap='coolwarm', annot=True, mask=mask, square=True, vmin=-1, vmax=1, center=0)
plt.show()
```

Додаток Б.

Лістинг коду моделі методом штучної нейронної мережі

```
from tensorflow import keras
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
df = pd.read_csv(data.csv')
X = df[['AverageCO2', 'AverageCH4', 'AverageN2O', 'AverageSF6']]
y = df['Average sea level rise']
dates = df['Date']
X_train, X_test, y_train, y_test, dates_train, dates_test = train_test_split(X, y, dates, test_size=0.2, shuffle=False, random_state=0)
X_train = X_train.reshape(-1, 4)
X_test = X_test.reshape(-1, 4)
model = keras.Sequential([
   keras.layers.Dense(64, activation='relu', input_shape=[4]),
   keras.layers.Dense(64, activation='relu'),
   keras.layers.Dense(1)
])
model.compile(loss='mean_squared_error', optimizer='adam', metrics=['mean_squared_error'])
model.fit(X_train, y_train, epochs=100, batch_size=10)
y_pred = model.predict(X_test)
test_loss, test_mse = model.evaluate(X_test, y_test, verbose=2)
print('Mean Squared Error on test set:', test_mse)
accuracy = 100 * (1 - np.abs((y_pred - y_test) / y_test)).mean()
print('Accuracy:', accuracy)
plt.plot(dates_test, y_pred, color='red', label='Predicted')
plt.plot(dates, y, color='blue', label='Actual')
```

```
plt.title('Average sea level rise over time')
plt.xlabel('Date')
plt.ylabel('Average sea level rise')
plt.legend()
plt.show()
```

Додаток В.

Лістинг коду моделі методом рекурентної нейромережі

```
from tensorflow import keras
import pandas as pd
import numpy as np
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score
import matplotlib.pyplot as plt
df = pd.read_csv(data.csv')
df.fillna(method="ffill", inplace=True)
train_size = int(len(df) * 0.8)
train_df, test_df = df[:train_size], df[train_size:]
train_dates = pd.to_datetime(train_df['Date'])
test_dates = pd.to_datetime(test_df['Date'])
scaler = MinMaxScaler()
train_scaled = scaler.fit_transform(train_df[['Average sea level rise']])
test_scaled = scaler.transform(test_df[['Average sea level rise']])
def create_dataset(X, y, look_back=1):
    X_data, y_data = [], []
    for i in range(len(X)-look_back):
        X_data.append(X[i:(i+look_back)])
        y_data.append(y[i+look_back])
    return np.array(X_data), np.array(y_data)
look_back = 12
X_train, y_train = create_dataset(train_scaled, train_scaled, look_back)
X_test, y_test = create_dataset(test_scaled, test_scaled, look_back)
model = keras.Sequential([
    keras.layers.LSTM(64, input_shape=(X_train.shape[1], X_train.shape[2])),
    keras.layers.Dense(1)
```

```
])
model.compile(loss='mean_squared_error', optimizer='adam')
history = model.fit(X_train.reshape(X_train.shape[0], X_train.shape[1], 1), y_train, epochs=100, batch_size=10, validation_data=(X_test.reshape(X_test.shape[0], X_test.shape[1], 1), y_test), verbose=2, shuffle=False)
train_predict = model.predict(X_train)
test_predict = model.predict(X_test)
train_predict = scaler.inverse_transform(train_predict)
y_train = scaler.inverse_transform(y_train.reshape(-1, 1))
test_predict = scaler.inverse_transform(test_predict)
y_test = scaler.inverse_transform(y_test.reshape(-1, 1))
train_mse = mean_squared_error(y_train, train_predict)
test_mse = mean_squared_error(y_test, test_predict)
print('Train MSE: %.3f' % train_mse)
print('Test MSE: %.3f' % test_mse)
train_r2 = r2_score(y_train, train_predict)
test_r2 = r2_score(y_test, test_predict)
print('Train R2: %.3f' % train_r2)
print('Test R2: %.3f' % test_r2)
plt.plot(train_dates[look_back:], y_train, label='Actual')
plt.plot(train_dates[look_back:], train_predict, label='Predicted')
plt.plot(test_dates[look_back:], y_test, label='Actual')
plt.plot(test_dates[look_back:], test_predict, label='Predicted')
plt.legend()
plt.show()
```