КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ТАРАСА ШЕВЧЕНКА

Факультет комп'ютерних наук та кібернетики

Кафедра Обчислювальної Математики

Дипломна робота

За спеціальністю 113 "Прикладна математика"

на тему:

Моделювання розповсюдження шкідливих викидів у атмосфері

Виконав студент 4-го курсу

Коломієць Микола

Науковий керівник:

Затула Дмитро

Київ, 2024

АНОТАЦІЯ

Коломієць М.О. Моделювання розповсюдження шкідливих викидів у атмосфері. Бакалаврська кваліфікаційна робота зі спеціальності 113 – прикладна математика, освітньо-професійна програма – прикладна математика. Київ: КНУ ім. Тараса Шевченка, 2024. 30 с.

У бакалаврській кваліфікаційній роботі досліджено методи моделювання розповсюдження шкідливих викидів у атмосфері з використанням лінійної регресії та нейронних мереж. Проведено аналіз даних про викиди за період 2021-2023 років, а також кліматичних даних, зібраних з офіційних джерел. Розроблено та оцінено моделі прогнозування якості повітря для перших трьох місяців 2024 року, з проведенням порівняння точності короткострокових та довгострокових прогнозів.

Ключові слова: моделювання, шкідливі викиди, лінійна регресія, нейронна мережа, нейрон, прогнозування, якість повітря, Python, Jupyter Notebook.

ЗМІСТ

ВСТУП

Актуальність теми. Забруднення повітря є дуже важливим питанням для здоров'я людей та навколишнього середовища. Воно може бути викликане різними факторами, такими як викиди транспорту, промисловрості та сільського господарства. На стан повітря певної місцевості також може впливати вітер, температура та географія самого міста. Велика кількість забруднюючих речовин може викликати різні захворювання та проблеми зі здоров'ям, тому це є важливим фактором для вибору місця проживання і відповідно важливою проблемою для влади міста, яка піклується про доброжиток населення міста.

Моніторинг і можливість прогнозування забруднення повітря є необхідним у сучасному світі. Для забезпечення здоров'я та комфорту населення міста важливо мати точні дані, щодо забруднення та його динаміки. Також важливо мати можливість прогнозувати забруднення задля вивчення аномалій та можливості вчасно реагувати на різкі зміни у стані повітря. Відмінності між реальним станом повітря і прогнозованим може бути наслідком появи нового чинника забрудення або впливу зміни клімату і за допомогою аналізу подібних відмінностей ці чинники можуть бути визначені. Крім того моделі можуть продемонструвати, як на стан повітря вплинули покращувальні заходи, що дозволить з меншими похибками визначати, які з методів є кращими.

Мета і завдання роботи. Метою даної дипломної роботи є розробка та застосування моделей для прогнозування розповсюдження шкідливих викидів у атмосфері, зокрема за допомогою лінійної регресії та нейронних мереж. Це дозволить покращити точність прогнозування рівнів забруднення повітря та забезпечити своєчасне реагування на можливі аномалії в забрудненні.

Завданнями роботи є:

- Збір та обробка данних: зібрати дані стосовно забруднення та параметрів за допомогою, яких буде виконуватись прогнозування. Провести очистку та підготовку даних для аналізу.

- Аналіз даних: Вивчити розподіл даних, провести логарифмічне перетворення та нормалізацію, за потреби, для підвищення точності моделей. Виявити патерни та взаємозв'язки між параметрами.

- Побудова моделей: Розробити та навчити моделі лінійної регресії та нейронної мережі для прогнозування викидів забруднювачів у повітря. Оптимізувати моделі для досягнення найкращих результатів.

- Прогнозування: Здійснити прогнозування рівнів забруднення та порівняти точність короткострокових та довгострокових прогнозів.

- Оцінка та валідація: Оцінити точність моделей.

Об'єкт дослідження. Об'єктом дослідження данної кваліфікаційної роботи є процеси розповсюдження шкідливих викидів у атмосфері та їх взаємодія з кліматичними факторами, такими як вітер, температура та географія місцевості.

Об'єкт дослідження. У данній роботі були використані такі методи:

- Статистичний аналіз: Для вивчення розподілу даних та виявлення основних патернів і аномалій.

- Візуалізація даних: Створення графіків і гістограм для наочного представлення розподілу даних та результатів моделювання.

- Моделювання: Розробка та застосування моделей лінійної регресії та нейронних мереж для прогнозування рівнів забруднення повітря.

РОЗДІЛ 1 ЗАБРУДНЕННЯ

1.1 Основні викиди

До основних викидів, що забруднюють атмосферу, належать:

- Тверді частинки (PM)
- Оксиди азоту
- Сірчастий газ
- Озон
- Вуглекислий газ
- Чадний газ

1.2 Тверді частинки (PM)

PM (particulate matter) це особлива категорія викидів в атмосферу, що включає в себе всі не газоподібні забруднювачі з малим розміром частинок. Частинки мають різноманітний хімічний склад і деякі можуть бути токсичними.

Ці викиди класифікують за розміром частинок. Найпоширенішими групами є PM10 та PM2.5, що відповідають частинкам з діаметром менше 10 та 2.5 мікронів відповідно. Подібні частинки можуть долати великі відстані в атмосфері за допомогою вітру. Тож деяка частина цих викидів прилітає зза кордону.

Невелика частина забруднення має природній характер (пил, морські частинки, вулканічний попіл) і більшість забруднення мають антропогенні джерела. Для PM це може бути викиди від спалення палива в транспорті, промисловості або пожеж в лісах. В великих містах значна частина PM з'являється від зношених шин та дисків автотранспорту.

Концентрація PM в повітрі моніториться організаціями з охорони здоров'я і граничні допустимі норми регулюються законодавством, задля цього був створений індекс якості повітря відносно PM.[2]

Індекс якості повітря	$PM_{2.5}$	PM_{10}
Добрий	0	0
Задовільний	12	54
Шкідливий для групи ризику	35	154
Шкідливий	55	254
Дуже шкідливий	150	354
Небезпечний	250	424

Таблиця 1.1: Індекс якості повітря відносно PM у $\frac{\text{мкг}}{\text{м}^3}$

Підвищена PM зазвичай спостерігається біля доріг з інтенсивним рухом транспорту та біля зони зони підприємств у великих містах. Як вже зазначалося, ці частинки можуть переноситися вітром тож великі концентрації заюруднення спостерігаються лише в дні з слабким вітром або у місцевості де вітер не виносить забруднення за межі міста.

Частинки через дихання потраплють у кровообіх і можуть осідати у внутрішніх органах, таким чином частинки менші 10 мікронів можуть спричинити проблеми з диханням та впливати на роботу серця.[1]

1.3 Оксиди азоту (NOx)

Більша частина оксидів азоту утворюється в результаті з'єднання кисню з азотом у полум'ї. Менша частина є результатом горіння сполук азоту в паливі. Природньо NOx утворюється в наслідок блискавки і незначною мірою мікробних процесів в ґрунті.

Антропогенні викиди оксидів домінують серед інших викидів за масою. Лише в британції ці викиди становлять близько 2.2 мільйонів тон кожного року. З них половина припадає на транспорт, чверть на електростанції і решта на інші промислові та побутові процеси спалювання.

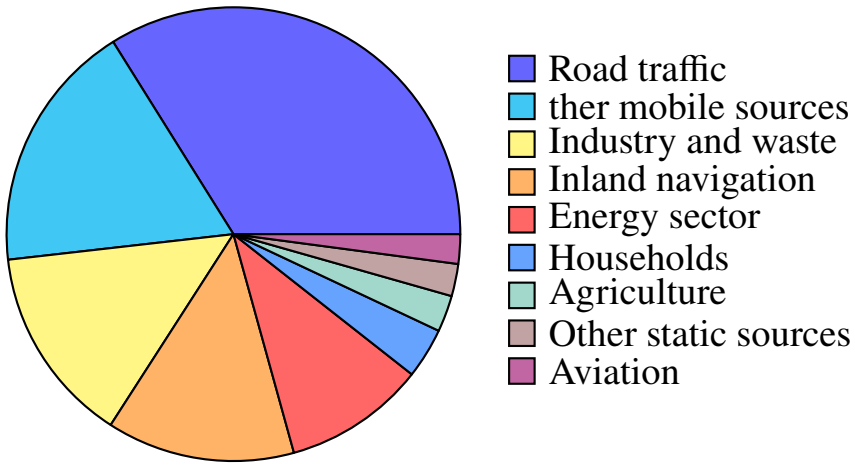

Рис 1.1 Діаграма чинників викидів оксидів азоту за 2022 рік по світу

Основними забруднювачами данного типу є оксид азоту NO та меншою мірою діоксид азоту NO_2. Протягом дня ці оксиди в атмосфері перетворюються один в одного. Оксид азоту окислюється в атмосфері до NO_2 за участю озону протягом десятків хвилин, а діоксид розщеплюється під дією ультрафіолетового випромінювання на NO та атом оксигену, що утворює озон з киснем. Таким чином ці гази існують у квазірівноважному стані за участі світла. Згодом діоксид азоту окислюється до азотної кислоти, яка швидко поглинається під час контакту з поверхнями, назважаючи на це самі оксиди зникають повільно і можуть подорожувати на великі відстані до розкладу на кислоту або нітрати. Тож забруднення однієї країни спричиняють забруднення і в сусідніх. Найбільша концентрація спостігається в районах великих міст через які проходять автомагістралі з інтенсивним рухом.

Високий рівень діоксиду азоту може спричинити пошкодження дихальних шляхів та підвищити вразливість людини до респіраторних інфекцій і астми. Тривалий вплив може спричинити хронічні захворювання легенів.

1.4 Сірчастий газ (SO2)

Діоксид сірки виділяється при спалюванні палива, що містить сірку, тож основними джерелами данних викидів є виробництво електроенергії, промислове та побутове спалення палива. Міжнародні організації зменшують кількість викидів сірчистого газу за допомогою законів і врегулювань. Також було розроблене обладнення для очищення димових газів від сірки і завдяки йому викиди продовжують зменшуватись, не звачаючи на збільшення

використання вугілля з 2000-них років. Агенством з охорони навколишнього середовища був створений індекс якості повітря відносно сірчистого газу, що дозволяють визначити наскільки небезпечна концентрація викидів для здоров'я людини:

Індекс якості повітря	Частка в повітрі (ppm)
Добрий	0
Задовільний	0,1
Шкідливий для групи ризику	0,2
Шкідливий	1,0
Дуже шкідливий	3.0
Небезпечний	5.0

Таблиця 1.2: Індекс якості повітря відносно SO2 у ppm (мільйонна частка)

Короткостроковий контакт з викидами може спричинити проблеми з дихальною системою людини. Особливо небезпечними подібні контакти для групи ризику - люди з астмою або діти. Також викиди діоксиду спричиняють формуванню інших оксидів сірки, що в наслідок реакціїї з іншими компоннтами атмосфери можуть перетворитись на тверді частки (PM), що в свою чергу вже мають свої наслідки для здоров'я людини.

Коли сірковий газ реагує з повітрям і водою, утворюється корозійна рідина - сіркова кислота, це одна з головних компонент кислотних дощів, що спричиняють велику шкоду навколишньому середовищу. На додачу сам сірковий газ сповільнює ріст рослин і пошкоджує листя.

При взаємодії сіркового газу з карбоном утворюються сульфатні айрозолі, що збільшують життя хмар сприяючі глобальному потеплінню.

1.5 Озон(O3)

Озон корисний високо, поганий поблизу. Озоновий шар, який знаходиться високо у верхніх шарах атмосфери, захищає нас від значної частини ультрафіолетового випромінювання Сонця. Проте забруднене озоном повітря на рівні землі, де ми можемо ним дихати, спричиняє серйозні проблеми зі здоров'ям. Озон агресивно атакує легеневу тканину, вступаючи з нею в хімічну реакцію.

Приземний озон утворюється в атмосфері з газів, які викадються із вихлопних труб, димових труб, фабрик та багатьох інших джерел забруднення. Коли ці гази контактують із сонячним світлом, вони реагують і утворюють озоновий дим. Для утворення озону необхідні оксиди азоту, легкі органічні сполуки та сонячне світло. Утворення оксидів азоту було розглянено в одному з попередніх підрозділів. Легкі органічні сполуки викидаються в повітря з деяких звичайних споживчих товарів, таких як фарба, і коли випаровуються побутові хімікати, такі як розріджувачі фарб і розчинники. вони також викидаються від автомобілів, хімічних заводів, нафтопереробних заводів, фабрик і автозаправних станцій. За присутності цих газів в правильних умовах утворюється озон і згодом вітер може рознести утворені викиди на великі відстані. Високий рівень озону частіше спостерігається влітку через високі температури і оскільки підвищення рівня данного виду викидів сприяє потеплінню це утворює циклічну залежність.

У Галвестоні, Техас, було проведене дослідження, яке показало, що навіть короткочасний вплив озону може погіршити здоров'я дорослих людей. Дослідження показало, що рятувальники мали більшу обструкцію легенів наприкінці дня, коли рівень озону був високий.

Групи ризику для впливу озону:

- вагітні жінки;
- діти;
- люди старші за 65 років;
- люди з астмою або іншими захворюваннями дихальних шляхів;
- Люди з нижчим соціально-економічним статусом;
- люди, які працюють або займаються спортом надворі.

Деякі контраверсійні дослідження показують більший вплив саме на жінок, але на данний момент відсутній остаточній консенсус стосовно цього питання.

Вплив озону у поєднанні з іншими факторами ризику скорочує середню тривалість життя. Існують переконливі докази смертоносності довго-

строкового впливу озону завдяки маштабним дослідженням. Було встановлено, що ризик передчасної смерті зростає зі збільшенням рівня озону.

В багатьох країнах в літній період утворюється достатньо озону для того щоб викликати проблеми зі здоров'ям. Інші проблеми крім укорочення середньої тривалості життя включають в себе:

- задишка хрипи і кашель
- респіраторні інфекції
- сприятливість до запалення легенів
- потреба в госпіталізації групи ризику

По мірі збільшення тривалості впливу озону можуть також з'являтися і інші проблеми. Це може бути метаболічні розлади, проблеми нервової системи, репродуктивні проблеми, рак, а також збільшення смертності від серцевих захвонювань.

Вдихання інших викидів може зробити аргонізм більш вразливи до озону і навпаки - вдихання озону підвищить реакцію на інші забруднювачі. Вплив озону також може підсилити реакцію у людей з з алергією.

Нові дослідження натякають на те, що можливо варто переосмислити стандарти щодо оцінки данного забруднення. Наприклад, дослідження 2017 року продемонструвало, що люди похолого віку мають симптоми навіть коли рівень озону залишається нижчим за поточний національний стандарт.

1.6 Вуглекислий газ (CO2)

Вуглекислий газ це важливий для землі газ, що зберігає тепло в атмосфері, що вивільняється за спалювання викопного палива. Також велика кількість вуглекислого газу викидається в атмосферу в результаті природніх процесів таких, як дихання та виверження вулканів, тож позбутися нанівець від цього газу неможливо. Хоча вуглекислий газ не вважається забрудником, через те що він є натуральним компонентом атмосфери та всі живі організми викидають його, але його кількість в атмосфері сильно зросла через спалювання викопаного палив у виробницві. Невеликі концентрації його у повітрі

безпечні для дихання, проте вони створюють парниковий ефект і сприяють глобальному потеплінню.

Найпоширеніший антропогенний чинник данного виду викидів полягає у спалюванні викопаного палива - вугілля, нафти та газу, для отримання тепла, електроенергії, руху транспорту. Крім прикладу електростанцій, що перетворюють тепло на електроенергію, можна навести будівничі процеси розповсюджені з приходом урбанізації - під час виробництва цементу використовується велика кількість викопаного палива для спалювання матеріалів цементу. Під час цього спалювання відбуваються хімічні реакції, що вивільняють вуглекислий газ. Загалом будівничі процеси багаті на хімічні реакції, що вивільняють вуглекислий газ, також вони залучають використання великої кількості транспорту, що також викидає вуглекислий газ. Крім того, вирубка лісів також вивільняє накопичений вуглець з лісових ландшафтів в атмосферу. Обернений до цього процес - секвестрація поліпшує стан справ і є одним із найперспективніших рішень для данного виду викидів.

Загалом вплив цих чинників виглядає так:

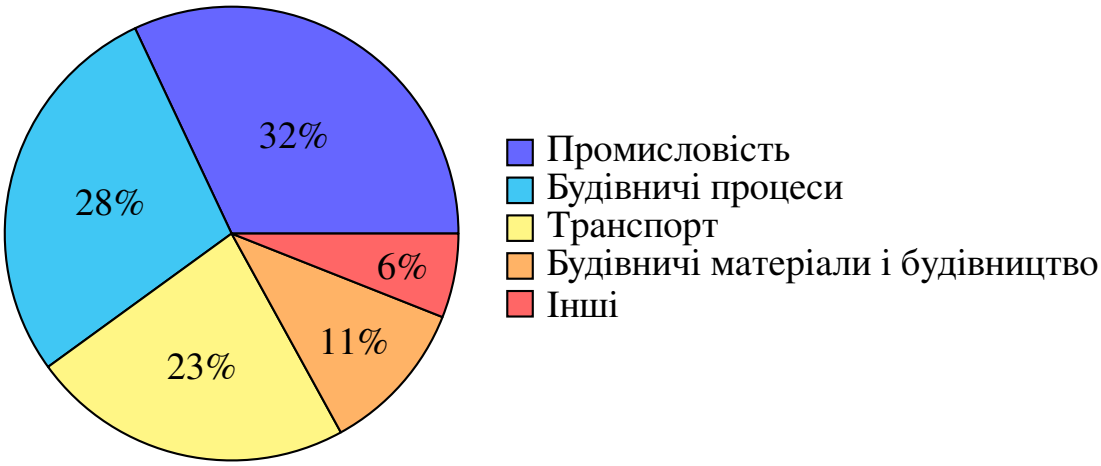

Рис 1.2 Діаграма антропогенних чинників викидів вуглекислого газу за 2019 рік по світу

Як було зазначено вище, невелика концентрація вуглекислого газу безпечна для дихання, проте великі концентрації можуть спричинити такі проблеми зі здоров'ям як:

- Головні болі
- Запаморочення

- Неспокійність
- Відчуття «поколювання».
- Утруднене дихання
- пітливість
- втома
- Почастішання пульсу
- Підвищений артеріальний тиск
- Кома
- Асфіксія
- Судоми

1.7 Чадний газ (CO)

Чадний газ, також відомий як оксид вуглецю, утворюється в результаті неповного згоряння палив, що містить у своєму складі вуглець, наприклад бензину, природного газу або деревини. Тож його джерелами є автомобілі, електростанції, лісові пожежі та застарілі сміттєспалювальні заводи. За кількістю викидів переважають саме автомобілі, що викидають близько 60% всіх викидів чадного газу. Оксид вуглецю також може утворюватися в результаті фотохімічних реакцій в атмосфері з метанових і неметанових вуглеводнів, інших летких органічних вуглеводнів в атмосфері та органічних молекул у поверхневих водах і ґрунтах.

Ранні ознаки легкого та помірного отруєння CO схожі на грип або харчове отруєння (за винятком відсутності температури), і деякі загальні симптоми включають:

- Головний біль
- Запаморочення
- Нудота
- Задишка
- Втома

Більш високі рівні отруєння призводять до гірших симптомів, включаючи нудоту, втруту свідомості, кому та навіть іноді призводять до летальних випадків.

1.8 Методи вимірювання забруднення

РОЗДІЛ 2 ДАНІ, ЇХ ПЕРЕТВОРЕННЯ І АНАЗІЛ

У даному розділі буде детально описано вигляд даних, їх джерело, формат, форма та виміри. Крім того, буде проведений аналіз зв'язків і патернів у даних, що допоможе виявити ключові взаємозв'язки та тенденції.

2.1 Дані

Дані щодо викидів були отримані з офіційного сайту Європейського Союзу за період з 2020 по 2023 роки. Кліматичні дані за той же період були зібрані з сайту NASA. Нижче наведено ключі та відповідні назви параметрів, які використовувалися:

SFC SW DWN	короткохвильове випромінювання неба
WD10M	напрямок вітру на висоті 10 метрів
WD50M	напрямок вітру на висоті 50 метрів
WS10M	швидкість вітру на висоті 10 метрів
WS50M	швидкість вітру на висоті 50 метрів
QV2M	питома вологість на 2 метри
PS	тиск на поверхні
T2M	температура на висоті 2 метри
co conc	чадний газ, виміри $\frac{10^{-9}kg}{m^3}$
no2 conc	діоксид азоту, виміри $\frac{10^{-9}kg}{m^3}$
no conc	моноксид азоту, виміри $\frac{10^{-9}kg}{m^3}$
o3 conc	озон, виміри $\frac{10^{-9}kg}{m^3}$
pm10 conc	PM10, виміри $\frac{10^{-9}kg}{m^3}$
pm2p5 conc	PM2.5, виміри $\frac{10^{-9}kg}{m^3}$
so2 co	сірчастий газ, виміри $\frac{10^{-9}kg}{m^3}$
time	дата замірів у форматі dd-mm-202y

Таблиця 2.1: Розшифровка ключів з NETCDF4

Формат, у якому були завантажені дані, є NETCDF. Цей формат зберігає інформацію у вигляді багатовимірних масивів, які мають прив'язку до географічних координат та рівнів, на яких були зняті відповідні виміри. Географічні дані були отримані з прямокутника, координати якого складають широту від 44.2 до 52.3 та довготу від 30.4 до 40.3. Цей прямокутник представляє собою східну частину України. (див. рисунок 2.1).

Рис. 2.1: Мапа з прямокутником, що показує область дослідження

Основні маніпуляції над даними, включаючи їх аналіз та обробку, виконувались у форматі Dataframe з бібліотеки pandas. Він надає зручний інтерфейс для роботи з даними та дозволяє проводити різноманітні операції, такі як фільтрація, групування та візуалізація. Netcdf це формат в, якому дані початково зберігались і були надані з джерел.

2.2 Інструменти

Для обробки даних використовувалась мова програмування Python на веб-інтерактивній обчислювальній платформі Jupyter Notebook. Ця платформа була обрана через можливість поетапного виконання коду з отриманням проміжних результатів без необхідності чекати завершення роботи всієї програми. Це забезпечує зручність та ефективність в аналізі даних, дозволяючи швидко вносити зміни та отримувати зворотній зв'язок.

Для аналізу, візуалізації та обробки даних були використані такі бібліотеки:

- NETCDF4 - Ця бібліотека була використана для роботи з NETCDF файлами, що дозволило зручно завантажувати, зберігати та обробляти великі об'єми багатовимірних даних.

- numpy - Використовувалася для роботи з масивами даних, забезпечуючи ефективне виконання операцій над ними, включаючи обчислення та трансформацію даних.

- matplotlib - Використовувалася для візуалізації даних, дозволяючи створювати різноманітні графіки та діаграми, які допомагають візуально представляти результати аналізу.

- pandas - Ця бібліотека була застосована для збереження даних у форматі Dataframe, що забезпечує зручний доступ до даних, їх фільтрацію, агрегацію та інші операції під час аналізу.

- seaborn - Використовувалась для побудови деяких графіків та дозволила виявити деякі взаємозв'язки та патерни в даних.

2.3 Вигляд, паттерни та зв'язок даних

2.3.1 Дані забруднення

Для того, щоб краще зрозуміти структуру та розподіл даних, перш ніж застосовувати методи моделювання, доцільно провести пропередній аналіз. Одним із важливих інструментів для цього є візуалізація даних у формі гістограми. Гістограма дозволить оцінити розподіл значень параметрів та прогнозованих величин. Гістограма - це тип діаграми, яка показує частоту розподілу даних. Додаткові причини застосування гістограми - виявити аномалії, що мають малу частоту і мають значення сильно відмінні від основного масиву, вони часто сильно впливають на результати застосуваня лінійної регресії і інших методів моделювання, розуміння структури даних, які впливають на вибір методу моделювання та його ефективність.

На гістограмі, зображеній на рис. 2.1, можна чітко побачити, що лише дані, які стосуються викидів озону, мають розподіл, подібний до нормального. Це означає, що для більшості рівнів озону спостерігається концентрація значень навколо середнього, зі зменшенням частоти як для значень, що пе-

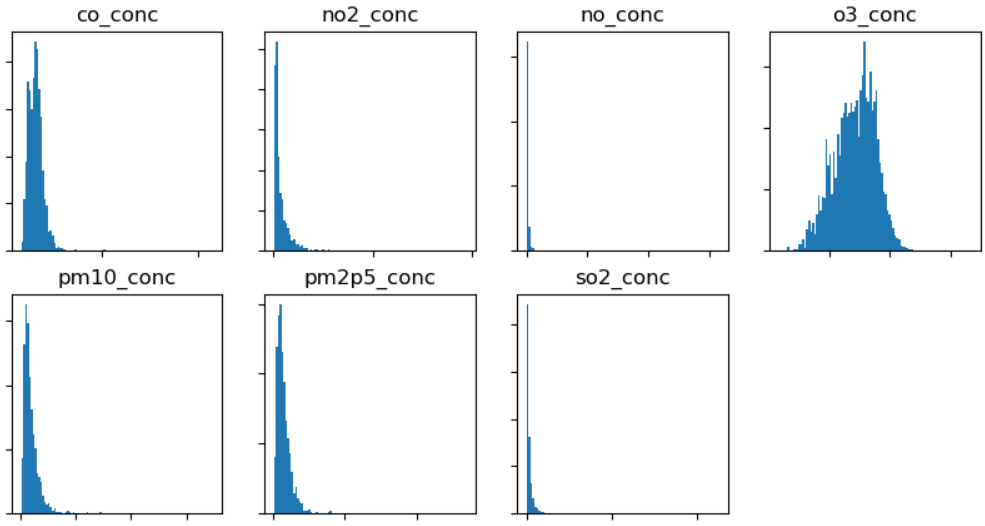

Рис. 2.2: Гістограми для даних про викиди

ревищують середнє, так і для тих, що знаходяться нижче середнього - гарні умови до застосування лінійної регресії.

Інші гістограми демонструють наявність аномалій, що ускладнює оцінку типів розподілів. Для вибору кращого діапазону застосування гістограми, що має меншу кількість аномалій, але не відкидає багато даних, доцільно розглянути діапазон значень цих параметрів.

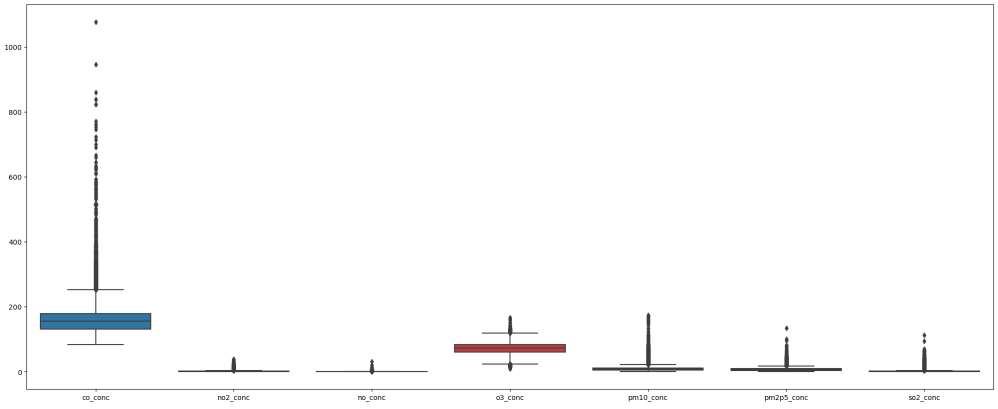

Рис. 2.3: Діапазон даних про викиди

Для застосування гістограми візьмемо такі проміжки: [100; 450], [0; 8], [0; 1], [0; 130], [0; 50], [0; 25], [0; 10]. Розглянемо на прикладі моноксиду азоту і сірчастого газу, оскільки їхні гістограми були найменш ілюстративними. Максимуми по викидам цих типів становлять 31 і 112 відповідно. При цьому лише 1.5% записів відносно моноксиду азоту більші за 1, і лише 1.2% записів

про сірчастий газ більші за 10.

Ці низькі частоти вказують на наявність аномалій, які значно впливають на загальне сприйняття розподілу даних. Тому для більш точного аналізу і візуалізації доцільно зосередитися на меншому діапазоні значень.

Вибір зазначених проміжків дозволяє краще побачити основні тенденції в розподілі даних, уникнувши впливу крайніх значень. Такий підхід допоможе виявити більш загальні патерни та забезпечить коректніше використання методів моделювання. Зокрема, гістограми з обраними діапазонами покажуть, як основна маса даних розподілена навколо середніх значень, що сприятиме точнішій оцінці можливостей застосування лінійної регресії та інших методів аналізу.

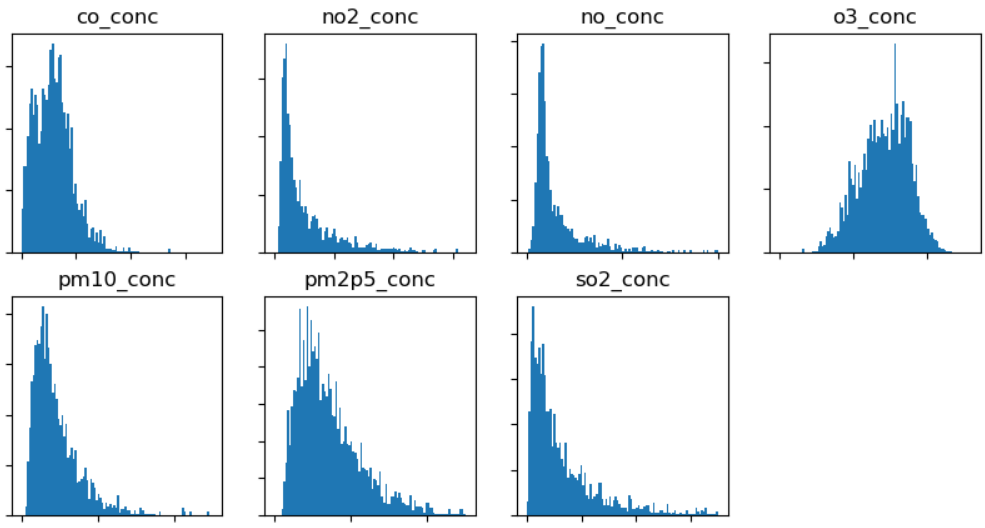

Рис. 2.4: Гістограми на зменшених проміжках

Як видно з гістограм, дані відносно всіх викидів, окрім озону та вуглекислого газу, утворюють правосторонні розподіли. Це означає, що більшість значень зосереджені у нижньому діапазоні, з поступовим зменшенням частоти у напрямку до вищих значень. Такий тип розподілу свідчить про наявність великої кількості малих значень та кількох значно більших, що є характерним для багатьох видів забруднюючих речовин.

Варто зазначити, що вуглекислий газ не демонструє чіткого розподілу. Хоч його дані розміщені не симетрично вони розподілені достатньо рівномірно відносно середнього значення для можливості застосування логістичної регресії.

2.3.2 Параметри прогнозування

Для застосування моделей з розділу 3 важливо проаналізувати зв'язки між параметрами прогнозування. Аналіз кореляційних коефіцієнтів дозволяє виявити, які параметри мають сильний або слабкий взаємозв'язок, що може суттєво вплинути на точність та ефективність моделі. Високий кореляційний коефіцієнт (наближений до 1 або -1) свідчить про сильний зв'язок між параметрами, тоді як низький коефіцієнт (близький до 0) вказує на слабкий або відсутній зв'язок. Важливо зазначити, що даний коефіцієнт показує лише кореляцію, тобто лінійну залежність між змінними. Це означає, що кореляційний коефіцієнт не відображає нелінійні зв'язки, які можуть бути присутніми в даних. Попарні кореляційні коефіцієнти наведені в таблиці:

	SFC SW DWN	WD10M	WS50M	QV2M	WD50M	PS	WS10M	T2M
SFC SW DWN	1.000000	-0.185005	-0.367893	0.648201	-0.186386	0.120202	-0.338364	0.641930
WD10M	-0.185005	1.000000	0.064822	-0.205908	0.999572	-0.179114	0.033519	-0.196132
WS50M	-0.367893	0.064822	1.000000	-0.211465	0.063771	0.000431	0.961444	-0.208328
QV2M	0.648201	-0.205908	-0.211465	1.000000	-0.206943	0.166831	-0.130927	0.922823
WD50M	-0.186386	0.999572	0.063771	-0.206943	1.000000	-0.179502	0.032133	-0.196800
PS	0.120202	-0.179114	0.000431	0.166831	-0.179502	1.000000	0.217696	0.280368
WS10M	-0.338364	0.033519	0.961444	-0.130927	0.032133	0.217696	1.000000	-0.106707
T2M	0.641930	-0.196132	-0.208328	0.922823	-0.196800	0.280368	-0.106707	1.000000

Таблиця 2.2: Коефіцієнти кореляції між параметрами прогнозування

Як видно з таблиці 2.2, проблемою є лише велика кореляція між параметрами вітру на висоті 10 та 50 метрів. З причин, пояснених у розділі 3, ми відкидаємо одну з цих висот. Надалі будуть розглядатися дані лише відносно висоти 10 метрів.

Розглянемо гістограми для параметрів прогнозування, щоб детально проаналізувати розподіл значень і виявити можливі аномалії або тенденції в даних.

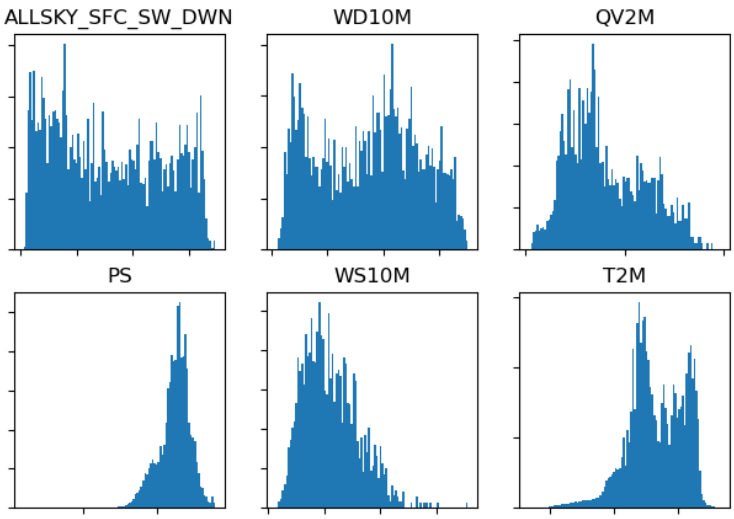

Рис. 2.5: Гістограми кліматичних параметрів

Швидкість вітру на висоті 10 метрів та тиск мають розподіли, наближені до нормального. Це сприяє більш точному аналізу та застосуванню статистичних методів, що передбачають нормальність розподілу.

Інші параметри не належать до жодного з типів розподілу.

2.4 Перетворення даних

Як було описано в попередньому підрозділі, більшість прогнозованих величин не мають нормального розподілу. Це може ускладнити подальший аналіз та моделювання. Однак, за допомогою логарифмування, можна виправити цю ситуацію та наблизити розподіл даних до нормального. Для цього ми застосуємо логарифмічне перетворення у вигляді $ln(1+x)$, що дозволяє уникнути великих негативних величин і забезпечити коректне оброблення даних з нульовими або малими значеннями. Це перетворення допоможе зменшити вплив аномалій та зробити дані більш придатними для аналізу.

Таким чином, після застосування логарифмічного перетворення, розподіли викидів вуглекислого газу, твердих часток PM10 та PM2.5 стали більш наближеними до нормального. Це сприяє більш точному та надійному аналізу даних.

Проте, розподіл озону став менш подібним до нормального після застосування логарифма.

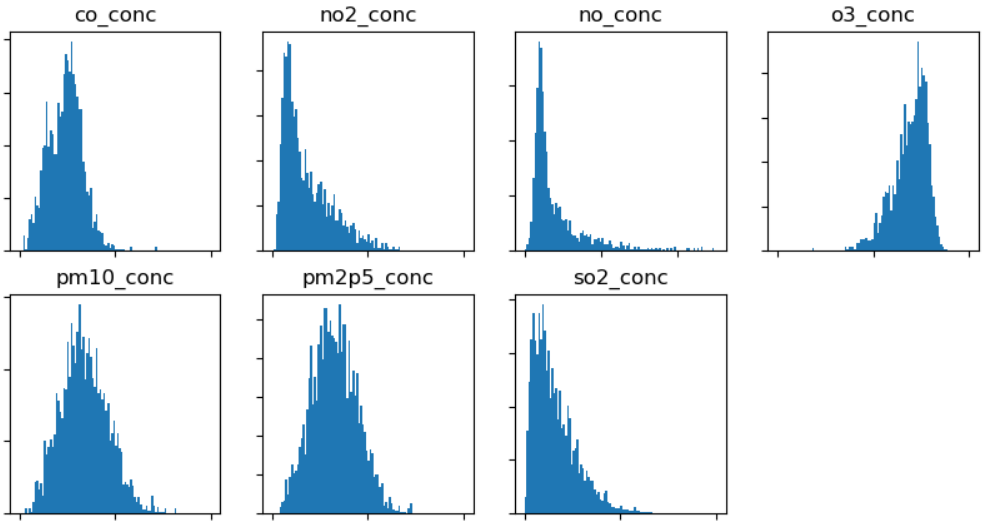

Рис. 2.6: Гістограми данних, до яких застосувалося логарифмічне перетворення

Отже, використання логарифмічного перетворення для даних озону не має сенсу, оскільки це не покращує його розподіл і лише ускладнює подальший аналіз.

Далі ми застосовуємо нормалізацію даних, що приводить всі параметри та прогнозовані величини до розподілу на проміжку [0, 1]. Це дозволяє усунути масштабні відмінності між параметрами, забезпечуючи більш стабільну роботу алгоритмів моделювання, але трохи ускладнює інтерпретацію результатів. Нормалізація відбувається за формулою $\frac{X-X_{min}}{X_{max}-X_{min}}$.

РОЗДІЛ 3 МОДЕЛЮВАННЯ

Для моделювання були розглянуті застосовані лінійна регресія та нейронна мережа. Кожна модель має свої переваги і свої недоліки. Умови, основи математичного обгрунтування, код, переваги та недоліки кожної моделі будуть розглянуті у відповідних підрозділах данного розділу і будуть підбиті підсумки у останньому підрозділі.

Дані, використані для отримання коефіцієнтів або навчання моделі, охоплюють період з 2021 по 2023 роки. Період прогнозування складає перші три місяці 2024 року. Для більш детального аналізу буде проведено порівняння точності короткострокових прогнозів з довгостроковими. Це дозволить оцінити ефективність моделі на різних часових проміжках, порівнюючи результати прогнозування для перших тижнів 2024 року з результатами для останніх тижнів цього ж періоду.

3.1 Лінійна регресія

3.1.1 Опис моделі

Лінійна регресія це модель, де прогнозована величина наближається за допомогою лінійної комбінації змінних:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

де y_i прогнозована величина за i-ий проміжок часу, x_j змінні, за якими відбувається прогнозування, β_j коефіцієнти, які визначають лінійну залежність, ε_i помилка моделі. Для лінійної регресії з багатьма параметрами частіше використовують матричний запис:

$$Y = X\beta + \varepsilon$$

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Коефіцієнти вектора β знаходяться за допомогою методу найменших квадратів тобто обераються такі значення, за яких набуває найменшого значення вираз:

$$S(\beta) = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 = |\|y - X\beta\|^2$$

З припущенням, що всі стовпці матриці X незалежні (не має залежних параметрів моделі) метод має єдиний розв'язок:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Для застосування подібних моделей були створені припущення, яких варто притримуватись задля коректного застосування тої чи іншої моделі:

- Слабка екзогенність. Незважаючи на випадкову природу більшості життєвих процесів до, яких застосовується регресія, вхідні дані розглядаються як фіксовані значення тобто припускається, що дані не містять помилок у вимірюванні
- Лінійність. З визначення моделі слідує, що середнє значення відповіді є лінійною комбінацією параметрів. На перший погляд це досить сильно обмежує можливості моделі, проте це припущення не обмежує перетворення оригінальних даних до параметрів, тобто пред застосуванням регресії не рідко дані логарифмують і нормують, але коефіцієнти регресії β залишаються лінійними
- Постійна дисперсія. Для великих і малих величин похибки має лишатись однаковою.

Порушення цих припущень призводить до упереджених оцінок коефіцієнтів, ненадійних довірчих інтервалів і тестів на значимість.

Перші два припущення ми можемо лише перевірити постфактум, на відміну від останнього. Як вже було описано в минулому розділі лише розподіл викидів озону схожий на нормальний, проте після логарифмічного перетворення дані стосовно твердих часток і вуглекислого газу можна використати для застосування лінійної регресії.

3.1.2 Результати

3.2 Нейронна мережа

3.2.1 Опис моделі

Нейронна мережа це модель, що складається з штучних нейронів, які з'єднані між собою. Концепція була створена на основі біологічних нейронних мереж, які складаються з нейронів, поведінка і структура яких не до кінця вивчена. Штучні нейрони мають більш просту структуру - кожен приймає певну кількість параметрів, однакову для всіх нейронів з одного шару, і видає вихідне значення. Вихідне значення обчислюється, як скалярний добуток параметрів певного нейрона на вхідні значення, після чого використовується активаційна функція. Параметри нейронів змінюються під час навчання моделі.

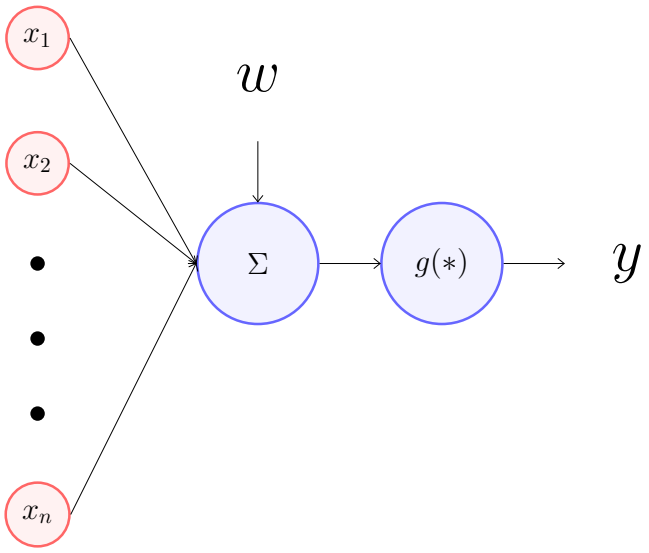

Рис. 3.1: Схема нейрона

На рис. 3.1 $x_1, x_2, \ldots x_n$ - вхідні значення, Σ - скалярний добуток (X, w) де w параметри нейрона, g - активаційна функція, y - вихідне значення.

Найпоширеніші функції акивації:

- Сигмоїда: $g(x) = \frac{1}{1+e^{-x}}$
- Relu: $g(x) = \max(0, x)$
- Гіперболічний тангенс: $g(x) = \tanh(x)$
- Softmax: $g(x) = \frac{e^{x_i}}{\sum_j e^{x_j}}$

Нейронні мережі зазвичай складаються з багатьох шарів. Перший шар приймає як вхідні дані параметри моделі, а останній видає прогнозоване значення. Нейрони на проміжних шарах приймають, як вхідні значення вихідні значення нейронів з попереднього шару. Для останнього нейрона часто використовують relu або softmax як функцію активації для задач класифікації одного класа або багатьох відповідно.

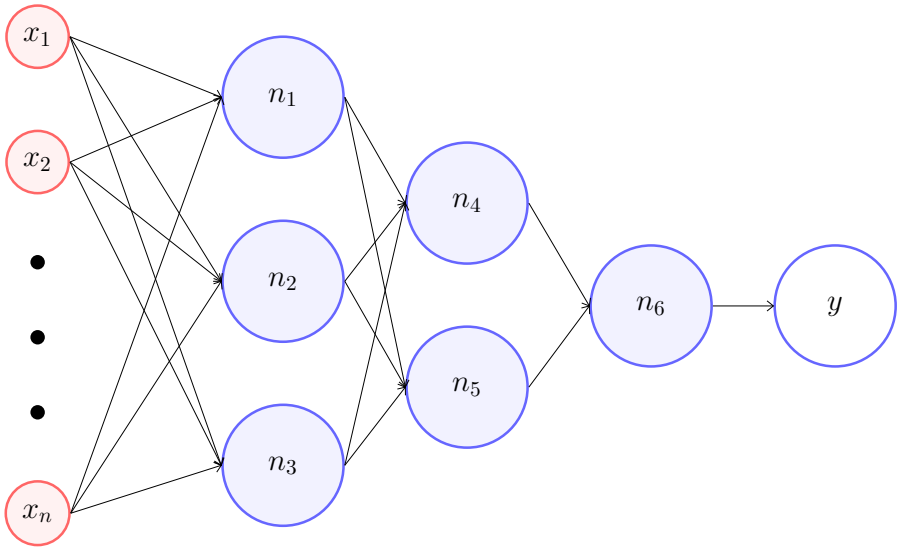

Рис. 3.2: Схема нейронної мережі з трьома шарами

Перед початком навчання ваги та зсуви ініціалізуються випадковими значеннями. Це допомагає уникнути симетрії між нейронами і забезпечує різні шляхи навчання для кожного з них. Далі йде розрахунок кінцевого значення моделі, яке порівнюється з реальним значенням за допомогою функції втрат. Функції витрат визначають, наскільки відрізняються прогнозовані значення від реальних. Після цього відбувається корекції ваг та зсувів. Якщо розгладати функцію витрат, як багатовимірну функцію від ваг, то корекція полягає у застосування градієнтного методу оптимізації. Градієнт розраховується в оберненому порідку від обчислення результату моделі, тобто

спочатку оптимізуються ваги останнього шару, потім передостаннього і так далі. Функції активації підібрані так, що градієнт можна обчислити аналітично. Цей процес повторюється для кожного набору даних, поки результат не задовільнить визначені критерії визначені для моделі.

Для нашого випадку велика нейронна мережа з багатьма шарами не підійде, адже під час навчання мадель запам'ятає дані і не зважаючі на те, що результат на навчальних даних буде дуже точним, на тестових даних результат буде поганим. У ході експериментів з різними розмірами моделі було виявлено, що найкращі результати показує модель з трьому шарами.

3.2.2 Результати

РОЗДІЛ 4 ВИСНОВКИ

БІБЛІОҐРАФІЯ

[1] Dusts | air pollution information system.

[2] Метеопост - Що таке PM2.5 та PM10. https://meteopost.com/info/PM/.