МАГІСТЕРСЬКА РОБОТА

МР. ІПм - 07.00.00.000 ПЗ

Група ІПм-22-1

Васько Христина

2024

Івано-Франківський національний технічний університет нафти і газу

Інститут інформаційних технологій

Кафедра інженерії програмного забезпечення

Васько Христина Андріївна

(прізвище, ім'я, по батькові)

УДК 004.4

(індекс)

МАГІСТЕРСЬКА РОБОТА

Засоби моделей Data Analytics для обробки кліматичних даних

(назва роботи)

Інженерія програмного забезпечення

(назва освітньої програми)

121 - Інженерія програмного забезпечення

(шифр і назва спеціальності)

Васько Х. А.

(підпис, ініціали та прізвище здобувача освітнього ступеня)

Науковий керівник Лютак Ігор Зіновійович, к.т.н., професор

(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

Допущено до захисту

В.о. завідувача кафедри

доц. Бандура В.В.

(посада) (підпис) (дата) (ініціали та прізвище)

Рецензент

доц.

(посада) (підпис) (дата) (ініціали та прізвище)

Робота містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

Івано-Франківськ – 2024

Івано-Франківський національний технічний університет нафти і газу

Інститут інформаційних технологій

Кафедра інженерії програмного забезпечення

Освітній рівень магістр

Спеціальність 121 – Інженерія програмного забезпечення

ЗАТВЕРДЖУЮ:
В.о. зав. кафедрою **ІПЗ**
доц. В.В. Бандура
“ 04 ” вересня 2023 р.

ЗАВДАННЯ

НА МАГІСТЕРСЬКУ РОБОТУ СТУДЕНТУ

Васько Христині Андріївній

(прізвище, ім’я, по-батькові)

1. Тема магістерської роботи “ Засоби Data Analytics для обробки кліматичних даних”

керівник проекту (роботи) Лютак Ігор Зіновійович, к.т.н., професор

затверджені наказом закладу вищої освіти від “ 18 ” грудня 2023 р. № 738/7

2. Строк подання студентом проекту (роботи) 22 січня 2024 р.

3. Вихідні дані до проекту (роботи) Теоретичні концепції та моделі аналізу кліматичних умов, набори даних, які включають інформацію про зміни клімату в конкретних регіонах протягом визначеного періоду

4. Зміст розрахунково - пояснювальної записки(перелік питань, які потрібно розробити)

1. Теоретичні відомості про основи моделей аналізу даних для обробки кліматичної інформації
2. Дослідження сучасних технологій аналізу кліматичних умов
3. Розробка алгоритму та системи аналізу кліматичних умов

5. Перелік графічного матеріалу (з точним зазначенням обов’язкових креслень)

1. Суть кластеризації даних (рис. 2.1, ст. 24)
2. Принцип роботи Bagging та Boosting методів (рис. 2.2, ст.28)
3. Гістограма температурних показників кожного місяця (рис. 3.6, ст.54)
4. Зміна показників швидкості вітру з часом (рис. 3.8, ст.56)
5. Перевірка збалансованості даних (рис. 3.15, ст.64)

6. Консультанти розділів проекту (роботи)

Розділ	Консультант	Підпис, дата
Нормоконтроль	доц., к.т.н. Вовк Р.Б.	
Перевірка на плагіат	доц., к.т.н. Вовк Р.Б.	

7. Дата видачі завдання 04 вересня 2023 р.

Керівник _____
(підпис)

Завдання прийняв до виконання _____
(підпис)

КАЛЕНДАРНИЙ ПЛАН

№ п/п	Назви етапів магістерської роботи	Строк виконання етапів роботи	Примітка
1	Підбір і вивчення літератури по темі магістерської роботи	01.10.2023	виконано
2	Аналіз основних сучасних методів аналізу даних для обробки кліматичної інформації	06.10.2023	виконано
3	Визначення методів обробки інформації, які описуються у даній роботі	22.10.2023	виконано
4	Дослідження різноманіття сучасних технологій аналізу кліматичних умов	19.11.2023	виконано
5	Формулювання умов та порівняння методів для рішення даних завдань	05.12.2023	виконано
6	Реалізація рішення	01.01.2024	виконано
7	Затвердження пояснювальної записки роботи завідувачем кафедри	15.01.2024	виконано

Студент – магістр _____
(підпис)

Керівник роботи _____
(підпис)

АНОТАЦІЯ

Магістерська робота: 82 с., 24 рис., 1 табл., 40 джерел.

Тема: Застосування моделей data analytics для обробки кліматичних даних.

Об’єкт дослідження: система інструментів та моделей data analytics, які використовуються для обробки та аналізу кліматичних даних.

Мета роботи: вивчення потенціалу моделей data analytics для вирішення завдань, пов'язаних зі зміною клімату.

Предмет дослідження: вивчення і застосування методів та інструментів аналізу даних для розуміння кліматичних процесів, прогнозування змін у кліматі та використання цих аналітичних засобів для вирішення конкретних завдань, пов'язаних з кліматичними даними.

Результати дослідження:

У даній магістерській роботі було проведено глибокий огляд існуючих методів аналізу кліматичних даних з метою розкриття їх ефективності та можливостей застосування в розв'язанні конкретних завдань. Під час дослідження було використано два підходи: візуалізацію даних та прогнозування даних алгоритмом Random Forest.

Висновок:

В результаті досліджень дана робота внесла важливий вклад у сучасну кліматологію та показує, що застосування методів аналізу даних має великий потенціал для розуміння та прогнозування кліматичних змін. Візуалізація виявилась необхідним інструментом для зрозуміння та передачі складних взаємозв'язків між кліматичними параметрами. Модель Random Forest здатна ефективно враховувати нелінійні взаємозв'язки та забезпечує високий рівень точності в порівнянні з іншими методами.

ВІЗУАЛІЗАЦІЯ КЛІМАТИЧНИХ ДАНИХ, КЛІМАТОЛОГІЯ, ЕКОЛОГІЧНІ ЗМІНИ, СИСТЕМИ ВІЗУАЛІЗАЦІЇ ДАНИХ, ЕФЕКТИВНІСТЬ ПРОГНОЗУВАННЯ, RANDOMFOREST, АНАЛІЗ ТРЕНДІВ.

ANNOTATION

Master's work: 82 p., 24 fig., 1 tab., 40 sources.

Topic: Applying data analytics models to climate data processing

Object of research: a system of data analytics tools and models used to process and analyse climate data.

Purpose: Exploring the potential of data analytics models to address climate change challenges.

Subject of research: learning and applying data analysis methods and tools to understand climate processes, predict climate change, and use these analytical tools to solve specific problems related to climate data.

Research results:

In this master's thesis, an in-depth review of existing methods of climate data analysis was conducted to reveal their effectiveness and possibilities of application in solving specific problems. Two approaches were used in the study: data visualisation and data prediction using the Random Forest algorithm.

Conclusion:

As a result of the research, this paper has made an important contribution to modern climatology and shows that the use of data mining techniques has great potential for understanding and predicting climate change. Visualisation proved to be an essential tool for understanding and communicating complex relationships between climate parameters. The Random Forest model is able to effectively account for nonlinear relationships and provides a high level of accuracy compared to other methods.

CLIMATE DATA VISUALISATION, CLIMATOLOGY, ENVIRONMENTAL CHANGE, DATA VISUALISATION SYSTEMS, FORECASTING EFFICIENCY, RANDOM FOREST, TREND ANALYSIS.

ЗМІСТ

Стр

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

ECMWF – (The European Centrefor Medium-Range Weather Forecasts) один з провідних світових центрів чисельного прогнозування погоди, моніторингу клімату та досліджень.

CART –The Classification And Regression Tree. Метод машинного навчання, який використовується як для задач класифікації, так і для регресії. Вперше введений в роботі Лео Бреймана, Джерома Фрідмана, Річарда Ольшена і Чарльза Стоуна у 1984 році, CART є одним з популярних методів дерев рішень.

Геоінформаційні системи – це інтегровані системи, призначені для збору, зберігання, обробки, аналізу та візуалізації географічно визначених даних.

CET – Центральноєвропейський час

EDA –це процес аналізу та вивчення основних характеристик набору даних.

ML – це галузь штучного інтелекту, яка вивчає розробку та застосування алгоритмів та моделей, які дають комп'ютерам здатність навчатися і покращувати свою продуктивність на основі досвіду без явного програмування.

Кластеризація - це метод машинного навчання, який використовується для групування подібних об'єктів чи спостережень в класи або кластери.

Аномалії клімату - відхилення від типових кліматичних умов.

Random Forest (Випадковий ліс) - це ансамбльний метод машинного навчання, який використовується для класифікації та регресії. Він базується на ідеї створення багатої кількості різних рішень та об'єднання їх для зниження ризику перенавчання та поліпшення загальної ефективності моделі.

ВСТУП

Актуальність роботи

Зміна клімату є однією з найгостріших проблем, з якими стикається світ сьогодні. Вона має значний вплив на навколишнє середовище, економіку та суспільство.

Кліматичні дані є важливим інструментом для розуміння зміни клімату. Вони можуть використовуватися для прогнозування майбутніх погодних явищ, виявлення тенденцій у кліматі, оцінки впливу зміни клімату та розробки заходів з пом`якшення та адаптації до зміни клімату.

Зміни в кліматі стають все більш екстремальними та непередбачуваними, що створює серйозні виклики для суспільства.

Сучасні технології та методи аналізу даних можуть виступити ключовими інструментами у розумінні цих змін та розробці стратегій пристосування.

Цей розділ роботи спрямований на розвиток таких інструментів, які здатні працювати з великим обсягом реальних часових та просторових кліматичних даних.

Моделі data analytics є потужним інструментом для обробки кліматичних даних. Вони можуть використовуватися для вирішення широкого спектру завдань, пов'язаних зі зміною клімату.

Актуальність роботи "Засоби моделей data analytics для обробки кліматичних даних" полягає в тому, що вона вивчає потенціал моделей data analytics для вирішення завдань, пов'язаних зі зміною клімату. Використання засобів data analytics дозволяє ефективно обробляти, аналізувати та використовувати ці дані для прогнозування та розробки стратегій адаптації до змін клімату.

Робота досліджує різні типи моделей data analytics, які можуть бути використані для обробки кліматичних даних, а також їхні переваги та недоліки.

Дана робота має важливе значення для розуміння того, як моделі data analytics можуть бути використані для боротьби зі зміною клімату. Вона може допомогти розробникам, науковцям та іншим зацікавленим сторонам у прийнятті рішень, заснованих на даних, для вирішення цієї глобальної проблеми.

Порівняння роботи з відомими розв’язаннями проблеми

Історія дата аналізу кліматичних даних сягає корінням у 18 столітті, коли метеорологи почали використовувати статистичні методи для аналізу даних про погоду. Вони використовували такі методи, як середнє, дисперсія та кореляція для виявлення тенденцій і закономірностей у даних.

У 20 столітті розвиток комп’ютерів дозволив метеорологам обробляти більші набори даних. Це призвело до розробки більш складних методів дата аналізу, таких як регресія та машинне навчання. Ці методи дозволяють метеорологам краще розуміти клімат і прогнозувати майбутні погодні явища.

У 21 столітті дата аналізу кліматичних даних продовжує розвиватися. Завдяки зростанню доступності даних і потужності комп’ютерів, метеорологи можуть використовувати нові методи дата аналізу для вирішення більш складних завдань. Наприклад, вони можуть використовувати дата аналіз для виявлення впливу зміни клімату на погодні явища та для розробки заходів з пом’якшення та адаптації до зміни клімату. Можна визначити широкий перелік завдань, які можна вирішити засобами дата аналізу:

- Прогнозування погоди (прогнозування штормів, посух тощо).
- Виявлення тенденцій у кліматі, таких як глобальне потепління. Це може допомогти в розробці заходів з пом’якшення зміни клімату.
- Виявлення незвичайних або неочікуваних значень у кліматичних даних.
- Кластеризація (групування кліматичних даних за їхніми схожими характеристиками. Це може допомогти в розумінні взаємозв’язків між різними кліматичними змінними).
- Аналіз кореляції (визначення зв’язку між двома або більше змінними. Це може допомогти в розумінні причин зміни клімату).

Розвиток технологій, зокрема в області комп'ютерів, статистики та інформаційних технологій, сприяє постійному вдосконаленню методів та засобів аналізу кліматичних даних. Ми можемо очікувати, що майбутнє принесе нові технології та інновації у цій області.

Мета і задачі дослідження

Метою даної магістерської роботи є вивчення потенціалу моделей data analytics для вирішення завдань, пов'язаних зі зміною клімату. Робота досліджує різні типи моделей data analytics, які можуть бути використані для обробки кліматичних даних, а також їхні переваги та недоліки. Планується провести вивчення потенціалу моделей data analytics для вирішення завдань, пов'язаних зі зміною клімату. Робота досліджує різні типи моделей data analytics, які можуть бути використані для обробки кліматичних даних, а також їхні переваги та недоліки.

Робота має ряд конкретних цілей, включаючи:

- Огляд різних типів моделей data analytics, які можуть бути використані для обробки кліматичних даних.
- Аналіз переваг і недоліків різних типів моделей data analytics.
- Оцінка потенціалу моделей data analytics для вирішення завдань, пов'язаних зі зміною клімату.

Об’єктом дослідження є система інструментів та моделей data analytics, які використовуються для обробки та аналізу кліматичних даних.

Предметом дослідження є вивчення і застосування методів та інструментів аналізу даних для розуміння кліматичних процесів, прогнозування змін у кліматі та використання цих аналітичних засобів для вирішення конкретних завдань, пов'язаних з кліматичними даними.

Методи дослідження

У даному дослідженні було скомбіновано та адаптовано декілька методів, які підпадали під вимоги. Основними з них є збір та обробка даних, вибір методів моделювання та розробка моделей аналізу даних з урахуванням специфічних кліматичних параметрів. Було також проведено аналіз результатів та врахування етичних аспектів використання та обробки кліматичних даних.

Наукова новизна одержаних результатів

Здійснено всебічний огляд різних типів моделей data analytics, які можуть бути використані для обробки кліматичних даних. Це включає статистичні моделі, машинне навчання та інтелектуальний аналіз тексту. Виявлено переваги та недоліки різних типів моделей data analytics для вирішення конкретних завдань, оцінено їх потенціал у вирішенні завдань, пов’язаних зі зміною клімату.

Практичне значення одержаних результатів.

На основі проведених досліджень був проведений аналіз кліматичних даних за допомогою моделей аналізу даних може допомагати в попередженні природних катастроф, що дозволяє ефективно моніторити вплив кліматичних змін на екосистеми

Особистий внесок

- Було проведено аналіз кліматичних даних в період 2000 – 2010 року на території певних міст Європи та створено інтерактивну візуалізацію для висвітлення результатів
- Розглянуто та використано існуючі моделі дата аналізу кліматичних показників та висвітлено результати кожного з них. Показано, що моделі дата аналізу є ефективними у використанні для виявлення взаємозв’язків між різними кліматичними змінними.

Структура магістерської роботи.

Магістерська робота викладена на 87 сторінках друкованого тексту, який складається з вступу, трьох розділів, висновків, списку використаних джерел (16 найменувань). Робота містить 15 рисунків.

РОЗДІЛ 1
ТЕОРЕТИЧНІ ОСНОВИ МОДЕЛЕЙ АНАЛІЗУ ДАНИХ ДЛЯ ОБРОБКИ КЛІМАТИЧНОЇ ІНФОРМАЦІЇ

1.1 Поняття кліматичних даних

Кліматологія — наука, що вивчає питання кліматоутворення, опис і класифікацію клімату земної кулі, антропогенні впливи на клімат. Належить до географічних наук, оскільки клімат є географічною характеристикою. Так само належить до географічної частини метеореології, оскільки кліматотворні процеси мають геофізичну природу.

Клімат - це довгостроковий стан погоди в певній місцевості. Він характеризується такими параметрами, як температура, кількість опадів, швидкість вітру, вологість повітря та ін.

Кліматичні дані - це інформація, яка характеризує клімат певного регіону або місцевості протягом тривалого періоду часу. Ці дані включають різноманітні показники та параметри, що описують атмосферні умови, температуру, вологість, опади, вітер, атмосферний тиск та інші аспекти клімату.

Основними компонентами кліматичних даних можна виділити наступні:

- Температура (середня, максимальна та мінімальна температура протягом року, температурні аномалії та коливання).
- Опади(кількість опадів, включаючи дощ, сніг, град, розподіл опадів за місяцями чи сезонами).
- Вологість (відносна вологість повітря, специфічна вологість атмосфери).
- Атмосферний тиск (середні значення атмосферного тиску).
- Вітряний режим (середня швидкість вітру, його напрямок).
- Інші параметри, такі як сонячна радіація, кількість сонячних годин тощо.

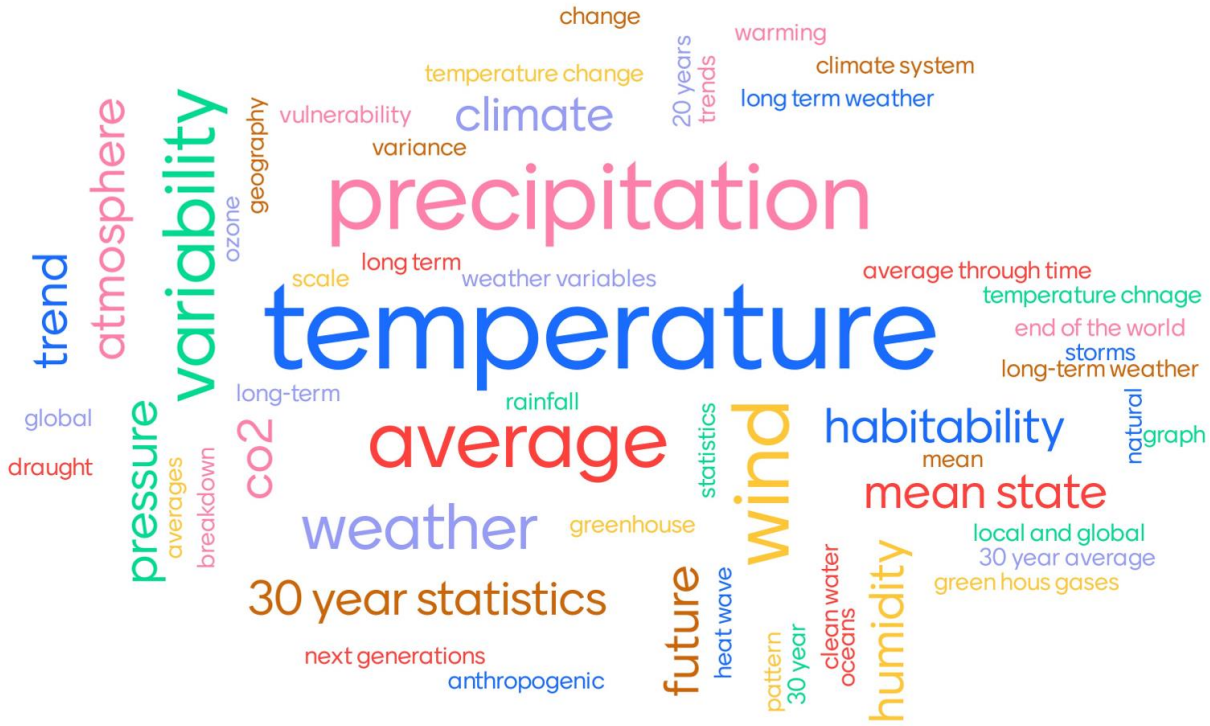


Рис 1.1. Основні кліматичні показники

Збір кліматичних даних використовує різноманітні джерела для отримання різноманітної інформації про умови. Набори кліматичних даних створюють метеорологи та кліматологи за допомогою датчиків з усього світу. Різні датчики збирають різні типи даних. Потужні термометри, наприклад, можуть обчислити середню температуру в географічній зоні, тоді як монітори опадів визначають кількість опадів.

Кліматичні дані можуть бути отримані за допомогою різних методів, таких як:

- Історичні записи: це джерело кліматичних даних, яке включає в себе записи про погоду, погодні явища та інші метеорологічні параметри, які були зроблені в минулому.
- Метеорологічні спостереження: це найпоширеніший метод отримання кліматичних даних. Він полягає в вимірюванні метеорологічних параметрів на регулярній основі, наприклад, кожні 6 годин або добу.
- Моделі клімату: це комп'ютерні моделі, які використовуються для прогнозування клімату в майбутньому. Вони можуть бути використані для

отримання кліматичних даних для періодів часу, для яких немає даних спостережень або історичних записів.

Кліматичні дані є важливим джерелом інформації для розуміння клімату і його змін. Вони можуть бути використані для прийняття рішень щодо управління навколишнім середовищем, а також для розробки заходів з пом'якшення та адаптації до зміни клімату.

Можуть бути використані для вирішення широкого спектру завдань, таких як:

- Прогнозування погоди: кліматичні дані можуть бути використані для прогнозування погодних явищ, таких як шторми, посухи та повені.
- Виявлення тенденцій у кліматі: кліматичні дані можуть бути використані для виявлення тенденцій у кліматі, таких як глобальне потепління.
- Виявлення аномалій: кліматичні дані можуть бути використані для виявлення аномалій у кліматі, таких як екстремальні погодні явища.
- Аналіз кореляції між кліматичними змінними: кліматичні дані можуть бути використані для аналізу кореляції між кліматичними змінними, такими як температура і кількість опадів.

Існують декілька основних джерел отримання кліматичних даних. Найпопулярнішою з них є Обсерваторія Землі NASA - інтернет-платформа для публікацій NASA, заснована в 1999 році. Це основне джерело супутникових зображень та іншої наукової інформації про клімат і навколишнє середовище, яку NASA надає широкому загалу. Також не менш важливими можна виділити Ініціативу Європейського космічного агентства зі зміни клімату (CCI), геомережу ФАО (FAO Geo Network) та інші.

1.2 Поняття та види моделей data analytics

Data Analytics - це міждисциплінарна галузь, яка використовує широкий спектр методів аналізу, включаючи математику, статистику та комп'ютерні науки, для отримання інсайтів з наборів даних. Аналітика даних - це широкий термін, який

включає в себе все: від простого аналізу даних до теоретичного осмислення способів збору даних і створення фреймворків, необхідних для їх зберігання.

Існує чотири ключові типи аналізу даних: описовий, діагностичний, прогнозний і прескриптивний. Разом ці чотири типи аналізу даних можуть допомогти організації в прийнятті рішень на основі даних. З першого погляду кожен із них говорить нам наступне:

- Descriptive analytics (Описова аналітика) - зосереджується на узагальненні та описі наявних даних, щоб дати вам чітке розуміння того, про що вони говорять.
- Diagnostic Analytics (Діагностична аналітика) - вивчає дані, щоб зрозуміти першопричини подій, поведінки та результатів. Аналітики даних використовують різноманітні методи та інструменти для виявлення закономірностей, тенденцій та зв'язків, щоб пояснити, чому відбулися певні події.
- Predictive Analysis (прогнозний аналіз) - процес використовує аналіз даних, машинне навчання, штучний інтелект і статистичні моделі для пошуку закономірностей, які можуть передбачити майбутню поведінку.
- Prescriptive analytics (прескриптивна) - це використання передових процесів та інструментів для аналізу даних і контенту, щоб рекомендувати оптимальний курс дій або стратегію руху вперед. Простіше кажучи, вона намагається відповісти на питання: "Що нам робити?".

Люди, які працюють з аналітикою даних, зазвичай досліджують кожну з цих чотирьох сфер, використовуючи процес аналізу даних, який включає визначення питання, збір необроблених даних, очищення даних, аналіз даних та інтерпретацію результатів.

Модель аналізу даних — це математична модель, яка використовується для аналізу даних. Це дає змогу бачити закономірності та зв’язки у ваших даних, щоб приймати рішення та прогнозувати майбутні події.

Існує безліч моделей та методів data analytics, які можна використовувати для обробки та аналізу кліматичних даних. Нижче наведено кілька прикладів, включаючи різні підходи та інструменти:

- Лінійна регресія — це алгоритм, який використовується для прогнозування або візуалізації зв'язку між двома різними характеристиками/змінними. У задачах лінійної регресії розглядаються два типи змінних: залежні змінні та незалежні змінні. Незалежна змінна — це змінна, яка існує незалежно від себе і на яку не впливають інші змінні. Коригування незалежної змінної змінює значення залежної змінної. Залежна змінна — це змінна, що досліджується, і змінна, яку регресійна модель намагається знайти або передбачити. У задачі лінійної регресії кожне спостереження/випадок складається як зі значення залежної змінної, так і зі значення незалежної змінної.
- Методи машинного навчання - великий підрозділ штучного інтелекту, що вивчає методи побудови алгоритмів, здатних навчатися. У найзагальнішому випадку розрізняють два типу машинного навчання: навчання по прецедентах, або індуктивне навчання, і дедуктивне навчання. Оскільки останнє прийнято відносити до області експертних систем, то терміни «машинне навчання» і «навчання по прецедентах» можна вважати синонімами. Цей метод навчання зараз, як прийнято говорити, в тренді, а ось експертні системи переживають кризу. Бази знань, що лежать в їх основі, важко узгоджувати з реляційною моделлю даних, тому промислові СУБД неможливо ефективно використовувати для наповнення баз знань експертних систем.
- Кластерний аналіз передбачає поділ даної вибірки об'єктів на підмножини, які називаються кластерами. Це дозволяє кожному кластеру складатися з подібних об'єктів, а об'єкти в різних кластерах значно відрізнятися. Завдання кластеризації належать до широкого класу завдань статистичної обробки та неконтрольованого навчання. Задачі кластерного аналізу можна об'єднати в такі групи: розробка типології або класифікації, Дослідження корисних концептуальних схем групування об'єктів, представлення гіпотез на основі

дослідження даних, перевірка гіпотез або досліджень для визначення, чи дійсно типи (групи), виділені тим або іншим способом, присутні в наявних даних.

- Геоінформаційні системи (ГІС) - це новітня комп'ютерна технологія для картографування та аналізу реальних об'єктів світу та подій, що відбуваються на Землі.Ця технологія поєднує в собі переваги традиційних операцій бази даних, таких як запити та статистичний аналіз, із повною візуалізацією і географічним (просторовим) аналізом за допомогою карт. Ці можливості відрізняють ГІС від інших інформаційних систем і забезпечують унікальні можливості для її застосування в широкому спектрі задач, пов'язаних із аналізом та прогнозом явищ і подій довкілля, з осмисленням і виділенням головних факторів і причин, а також їх можливих наслідків.
- Аналіз часових рядів – це набір методів математико-статистичного аналізу, спрямованих на виявлення та прогнозування структури часових рядів, що включає, зокрема, методи регресійного аналізу.
- Байєсівські методи можуть використовуватися для оцінки ймовірностей та ризиків кліматичних подій, що допомагає управляти невизначеністю в кліматичних моделях.
- Великі дані та обробка потокових даних: застосування техніки обробки великих обсягів даних та аналізу потокових даних для реального часу аналізу кліматичної інформації.

1.3 Переваги та недоліки моделей data analytics

Використання аналітики даних у бізнес-операціях має кілька переваг. Перш за все, аналітика даних може допомогти бізнесу приймати обґрунтовані рішення на основі інсайтів, заснованих на даних. Це може допомогти бізнесу уникнути дорогих помилок і краще використовувати свої ресурси. Аналітика даних також може допомогти бізнесу визначити нові можливості для зростання та оптимізувати свою діяльність для підвищення ефективності та зменшення витрат.

Однак використання аналітики даних також має певні недоліки. Наприклад, аналітика даних вимагає значних інвестицій у вигляді ресурсів, часу та експертизи. Компанії, які не мають необхідних ресурсів і досвіду, можуть зіткнутися з труднощами при ефективному впровадженні аналітики даних. Крім того, аналітика даних може бути схильною до упередженості та неточностей, якщо її не впроваджувати належним чином.

Можна виділити певні переваги моделей data analytics. Нижче наведені основні з них:

- Ефективність при обробці великих обсягів даних (моделі можуть ефективно обробляти великі обсяги даних, дозволяючи виявляти тенденції та закономірності, які можуть бути непомітними при ручному аналізі).
- Швидкість та автоматизація (алгоритми дозволяють автоматизувати аналіз та прийняття рішень, що призводить до швидшого реагування на зміни в даних).
- Можливість виявлення складних зв'язків (машинне навчання та інші просунуті методи дозволяють виявляти складні та неочевидні зв'язки між різними змінними в даних).
- Прогностична здатність (деякі моделі можуть використовуватися для прогнозування майбутніх подій або трендів на основі історичних даних).
- Покращення прийняття рішень (аналітика даних може слугувати важливим інструментом для прийняття обгрунтованих та заснованих на фактах рішень).

Незважаючи на велику кількість переваг, існує також перелік недоліків аналітики даних:

- Залежність від якості вихідних даних (точність та надійність результатів data analytics залежить від якості та правильності вихідних даних. Неправильні або викривлені дані можуть призвести до неточностей).

- Неспроможність враховувати контекст (моделі можуть бути обмеженими у здатності враховувати контекст та специфіку конкретної області, що може призвести до недостатньо точних або застарілих результатів).
- Спрощення складних явищ (деякі моделі можуть надмірно спрощувати складні явища або не враховувати всі можливі фактори, що призводить до обмеженої точності).
- Вплив вибірки даних (результати data analytics можуть бути чутливими до вибірки даних, використаної для навчання моделі, що може впливати на їхню загальну адаптивність).
- Відсутність пояснювальної здатності (деякі моделі, особливо ті, що базуються на глибинному навчанні, можуть бути важко зрозумілі та пояснювальні, що робить їх менш зручними для використання у випадках, коли потрібне пояснення прийняття рішень).

1.4 Застосування моделей Data Analytics для кліматичних даних

Клімат Землі постійно змінюється, і ці зміни мають значний вплив на нашу планету. Зміна клімату може призвести до таких наслідків, як підвищення рівня моря, зростання частоти і інтенсивності погодних явищ, і зміна рослинності та тваринного світу.

Моделі data analytics можуть бути використані для аналізу кліматичних даних і отримання цінних знань про зміни клімату. Ці моделі можуть допомогти нам краще зрозуміти, як клімат змінюється, і виявити тенденції, які можуть бути важливими для прогнозування майбутніх змін.

Описові моделі можна використовувати для пошуку закономірностей і тенденцій у кліматичних даних. Наприклад, такі моделі можуть використовуватися для пошуку тенденцій до зміни кількості опадів, підвищення рівня моря або підвищення температури.

Діагностичні моделі можна використовувати для виявлення причинно-наслідкових зв’язків між різними факторами, включаючи зміни клімату. Такі моделі,

наприклад, можуть бути використані для пошуку зв’язку між зміною клімату та підвищенням частоти та інтенсивності погодних явищ.

Прогнозні моделі можна використовувати для прогнозування змін клімату в майбутньому. Такі моделі, наприклад, можуть використовуватися для прогнозування майбутнього рівня моря або для прогнозування частоти та інтенсивності погодних явищ.

У дослідженні «Big Data Analytics in Weather Forecasting: A Systematic Review» [1] Marzieh Fathi, Mostafa Haghi Kashani систематично переглянути та проаналізувати дослідження, присвячені використанню аналітики великих даних у прогнозуванні погоди, виявити недоліки попередніх оглядів і запропонувати докладніший та структурований підхід до аналізу існуючих робіт у даній галузі. В результаті було встановлено, що час, точність, масштабованість, MSE, RMSE, точність та надійність є ключовими факторами якості обслуговування, що застосовуються у вибраних дослідженнях, тоді як температура, вітер, опади, вологість, дані про дощ та тиск є найчастіше використовуваними параметрами погоди. Apache Hadoop є найпопулярнішим інструментом серед аналізованих досліджень, зі значною часткою використання, за ним слідують MATLAB і Python з окремими перевагами та недоліками кожної технології.

Основне завдання статті "Visual Analytics of Large-Scale Climate Model Data" [2] - продемонструвати розробку та застосування інструменту візуальної аналітики для дослідження та аналізу великомасштабних даних кліматичних моделей. Інструмент покликаний допомогти науковцям зрозуміти складні набори даних, надаючи інтерактивні, багатогранні можливості візуалізації та аналізу. У статті представлено можливості візуалізації, обчислювальну продуктивність та зручність використання інструменту в контексті вивчення кліматичних явищ, таких як коливання Меддена-Джуліана. У ній також повідомляється про оцінку застосовності інструменту науковцями в цій галузі, а також про спостереження та уроки, отримані під час розробки інструменту. Дослідження виявило, що так інструменти як Matrix of Scatterplots, Parallel Coordinates та 3D Volume Time Series Visualization [2] успішно задовольнили потребу науковців в інтерактивній візуалізації даних

великомасштабної кліматичної моделі, полегшивши їх вивчення та аналіз. Науковці не відсіяли більше 10% даних, що свідчить про їхню зосередженість на загальному огляді та тенденціях, а не на локальних дрібних деталях. Незважаючи на великий обсяг даних (~2,9 млрд. числових значень), звичайні методи візуалізації інформації, подібні до розглянутих у статті, можуть обробляти такі набори даних на звичайному робочому столі з відповідними апаратними специфікаціями.

Інструмент буде продовжувати розвиватися з метою розширення його дослідницьких можливостей та аналізу додаткових кліматичних явищ, окрім коливань Меддена-Джуліана.

В даній роботі постає питання вибору найефективнішого способу аналізу кліматичних даних згідно із поставленими завданнями та запитами. Розглянуто методи прогнозування даних за допомогою машинного навчання ML та візуального аналізу даних.

1.5 Висновки до розділу

В даному розділі ми розглянули фундаментальні концепції та методології, які лежать в основі вивчення кліматичних змін та їх впливу на навколишнє середовище. В ході дослідження ми проаналізували ключові аспекти обробки кліматичної інформації, використовуючи сучасні методи та моделі аналізу даних.

Важливим елементом нашого вивчення стало встановлення взаємозв'язків між різними кліматичними параметрами та розробка моделей прогнозування змін клімату на основі отриманих даних. Ми детально розглянули математичні та статистичні підходи до аналізу часових рядів, просторової варіабельності та інших аспектів кліматичних даних.

РОЗДІЛ 2
ДОСЛІДЖЕННЯ СУЧАСНИХ ТЕХНОЛОГІЙ АНАЛІЗУ КЛІМАТИЧНИХ УМОВ

2.1 Машинне навчання в аналізі кліматичних змін

2.1.1 Машинне навчання в моделюванні зміни клімату: вирішення екологічних проблем

Прогнозування кліматичних змін за допомогою алгоритмів машинного навчання є активним напрямком досліджень. Різні алгоритми використовуються для вирішення різних кліматологічних задач. Розглянемо детальніше деякі з них:

- Clustering

Процес упорядкування групи об’єктів таким чином, щоб ці об’єкти були більш схожі один на одного, ніж об’єкти в іншій групі, яка називається кластером. На етапі дослідницького аналізу даних фахівці з обробки даних часто використовують кластеризацію, щоб знайти нові шаблони та інформацію в даних. Кластеризація не потребує набору даних із маркуванням, оскільки це неконтрольоване машинне навчання.

Сама по собі кластеризація - це завдання, яке потрібно вирішити, а не один алгоритм. Цієї мети можна досягти за допомогою різноманітних алгоритмів, які значно відрізняються від вашого розуміння кластера та його ефективного пошуку.

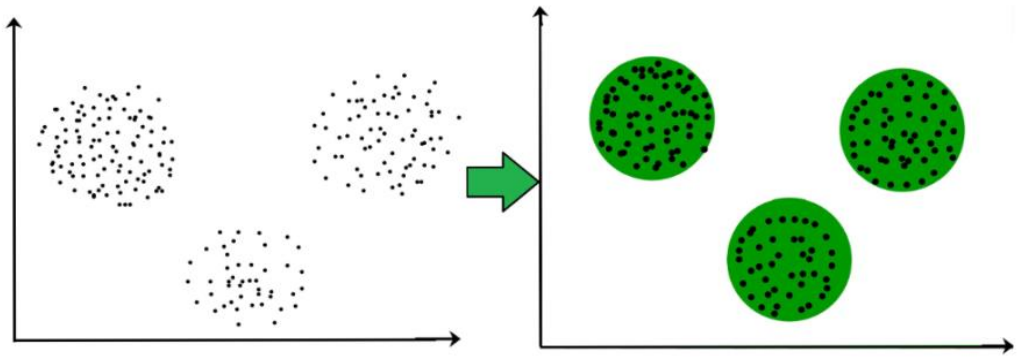


Рис. 2.1. Суть кластеризації

- Decision trees

Навчання на основі дерев рішень - це метод, який часто використовується в інтелектуальному аналізі даних, видобутку даних. Його мета полягає у створенні моделі, яка прогнозує значення цільового параметра на основі декількох вхідних параметрів. Дерево може й можна навчити, розбивши набір вихідних даних на підмножини на основі на основі тесту значень атрибутів. Цей процес повторюється для кожної підмножини рекурсивно, що називається рекурсивним розбиттям. Рекурсія завершується, коли підмножина у вузлі має всі однакові значення цільової змінної або коли розбиття більше не додає значення до прогнозів.

Цей процес низхідної індукції дерев рішень є прикладом жадібного алгоритму, і це, безумовно, найпоширеніша стратегія для навчання дерев рішень на основі даних в інтелектуальному аналізі даних. В системах TDIDT машинне навчання можна класифікувати на основі наступних ознак: використана стратегія навчання, представлення знань, отриманих системою, область застосування системи.

Дерева рішень можна описати як комбінацією математичних та обчислювальних методів, які допомагають опису, категоризації та узагальнення певного набору даних для полегшення машинного навчання. У дереві рішень залежна змінна прогнозується на основі незалежної змінної. Дерева рішень, що використовуються в інтелектуальному аналізі даних, зазвичай бувають наступних типів: аналіз дерев класифікації використовується для прогнозування даних у класі, аналіз дерева регресії необхідний для прогнозування незалежної змінної як одиниці числа.

Аналіз дерева класифікації та регресії (CART) – це термін, який зазвичай використовується для позначення обох вищезгаданих процедур. Дерева, що використовуються для регресії та класифікації схожі за процедурою, але відрізняються процедурами розбиття вузла. Навчання на дереві рішень - це побудова дерева рішень на основі навчальних даних, позначених класами. Дерево рішень - це блок-схема,де кожен внутрішній вузол позначає тест на атрибуті, а кожна гілка представляє результат тесту, а кожен листовий вузол містить мітку класу або прогноз. Самий верхній вузол у дереві – це кореневий вузол, як і в дереві.

Дерево рішень розділяє атрибути за допомогою жадібного пошуку, який оптимізує за певним критерієм. Умови тесту задаються в залежності від типу атрибутів: номінальні, порядкові або неперервні. Визначення найкращого розбиття залишається проблемою. Жадібний метод передбачає, що перевага надається вузлам з однорідним розподілом класів, тому існує потреба у вимірюванні методу домішок у вузлах домішок. Домішка вузлів вимірюється наступним чином: Giniiindex, ентропія, помилка класифікації.

Gini index

Використовується в CART як абревіатура алгоритму класифікації та регресії. Gini index - це міра того, як часто випадково вибраний елемент з набору буде неправильний якби він був випадково позначений відповідно до розподілу міток у підмножині. Gini index може бути обчислена шляхом підсумовуючи ймовірність вибору кожного елемента, помножену на ймовірність помилки при віднесенні цього елемента до тієї чи іншої категорії. Вона досягає свого мінімального значення нуля, коли всі випадки у вузлі потрапляють в одну цільову категорію.

$$Gini(t) = 1 - \sum [P(j \mid t)]^2 \tag{2.1}$$

Наприклад, одне рішення класифікує 1 екземпляр до класу c1, а решта 5 екземплярів, решта 5 екземплярів класифікуються в c2, тоді ймовірність p(c1)= 1/6і p(c2)=5/6. Тоді:

$$Gini = 1 - [(1/6)^2 - (5/6)^2] \tag{2.2}$$

Знайшовши вищевказане значення, відсортовується значення атрибутів та знаходиться індекс Gini для кожного значення атрибутів і виберіть позицію розбиття за найменшим індексом Gini.

Отримання інформації: ентропія у заданому вузлі t позначається формулою:

$$Entropy(t) = -\sum p(j|t) log\, p(j|t) \tag{2.3}$$

де p(j | t) - відносна частота класу j у вузлі t. Вимірюється однорідність вершини. Максимальне значення (log n) має місце, коли записи рівномірно розподілені між усіма класами, що означає найменшу кількість інформації. і мінімальне значення 0.0, коли всі записи належать до одного класу, що означає найбільше інформації. Наприклад, потрібно прийняти рішення, яке розділяє один екземпляр у клас c1 і 5 екземплярів у клас c2. тоді ентропія обчислюється наступним чином: P(C1) = 1/6 P(C2) = 5/6. Ентропія = - (1/6) log2 -(1/6) - (5/6) log2; (1/6) = 0.65 біт. Тоді інформаційний виграш у вигляді розщеплення виграшу обчислюється наступним чином наступним чином:

$$Entropy(p) - (\sum (ni/n) Entropy(i)) \tag{2.4}$$

- Random forests

Даний алгоритм широко відомий через його зручність та адаптивність,що дозволяє йому ефективно вирішувати як класифікаційні, так і регресійні задачі. Сила алгоритму полягає в його здатності обробляти складні набори даних і зменшувати надмірне припасування, що робить його цінним інструментом для різних завдань прогнозування в машинному навчанні.

Однією з найважливіших особливостей алгоритму випадкового лісу є те, що він може працювати з набором даних, що містить безперервні змінні, як у випадку регресії, і категоріальні змінні, як у випадку класифікації. Він краще підходить для задач класифікації та регресії. У цьому уроці ми розглянемо принцип роботи випадкового лісу і застосуємо його до задачі класифікації.

Перш ніж зрозуміти роботу алгоритму випадкового лісу в машинному навчанні, ми повинні розглянути техніку ансамблевого навчання. Ансамбль просто означає об'єднання декількох моделей. Таким чином, для прогнозування використовується

набір моделей, а не окрема модель. Ансамбль використовує два типи методів: Bagging I Boosting (підсилення).

Bagging створює іншу навчальну підмножину із зразкових навчальних даних із заміною, а кінцевий результат базується на голосуванні більшості. Наприклад, Random Forest.

Boosting об'єднує слабких учнів у сильних, створюючи послідовні моделі, щоб кінцева модель мала найвищу точність. Наприклад, ADA BOOST, XG BOOST.

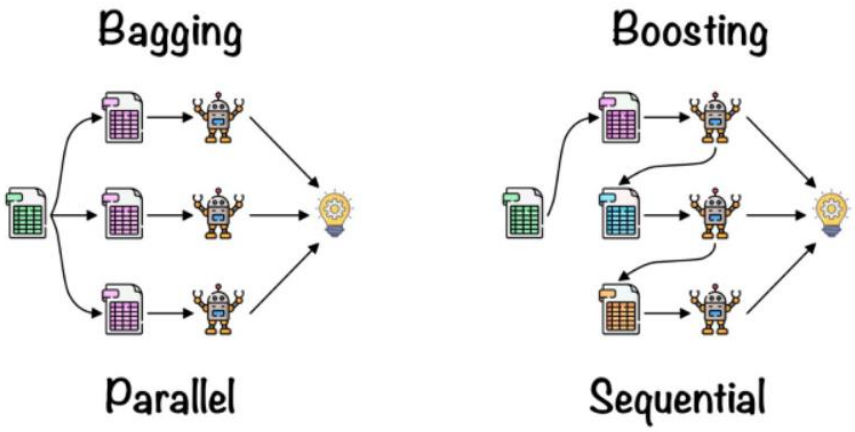


Рис. 2.2. Принцип роботи Bagging I Boosting методів

Метод ансамблевого навчання складається з набору класифікаторів, таких як дерева рішень, і прогнози цих класифікаторів збираються для отримання найпоширенішого результату. Пакування, також відоме як завантажувальна агрегація, і посилення є двома найвідомішими методами ансамблю.

У 1996 році Лео Брейман представив метод упаковки, який дозволяє вибирати випадкову вибірку даних у навчальному наборі із заміною, що дозволяє вибирати окремі точки даних декілька разів. Ці моделі навчаються незалежно після створення кількох вибірок даних. Залежно від типу завдання (наприклад, регресії чи класифікації), середнє або більшість цих прогнозів дають більш точну оцінку. Це найпоширеніший метод для зменшення дисперсії всередині шумного набору даних.

Алгоритм випадкового лісу використовує як пакетування, так і випадковість ознак для створення некорельованого лісу дерев рішень. Це розширення методу

пакетування. Випадковість функцій, також відома як пакетування функцій або «метод випадкового підпростору», створює випадкову підмножину функцій, яка забезпечує низьку кореляцію між деревами рішень. Посилання на цю функцію не на ibm.com. Це основна відмінність між випадковими лісами та деревами рішень. Дерева рішень враховують усі можливі розподіли функцій, але випадкові ліси вибирають лише одну підмножину цих функцій.

Перед навчанням алгоритми випадкового лісу повинні встановити три основні гіперпараметри. Розмір вузла, кількість дерев і кількість відібраних функцій належать до них. Звідси можна вирішити проблеми класифікації або регресії за допомогою класифікатора випадкового лісу.

Алгоритм випадкового лісу складається з набору дерев рішень, а кожне дерево в ансамблі складається з вибірки даних, отриманої з навчального набору із заміною, яка називається початковою вибіркою. Одна третина цієї навчальної вибірки була відкладена як вихідна (oob) вибірка. Пакування функцій додає ще один випадок випадковості, зменшуючи кореляцію між деревами рішень і доповнюючи набір даних різноманітністю. Визначення прогнозу буде різним залежно від типу проблеми. Для регресійного методу окремі дерева рішень будуть усереднені, а для класифікаційного методу найпоширеніша категоріальна змінна дасть прогнозований клас. Нарешті, це передбачення завершується використанням зразка oob для перехресної перевірки.

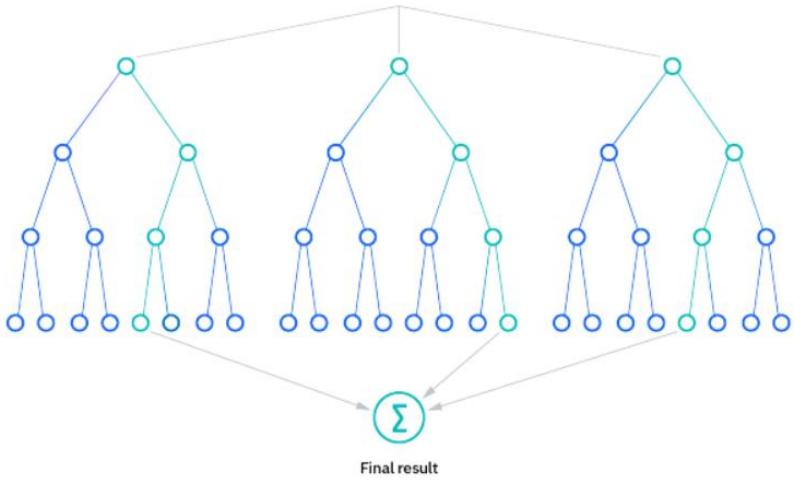


Рис. 2.3. Набір дерев рішень, з яких складається алгоритм random forest

Хоча Random forest - це сукупність дерев рішень, але в їхній поведінці є багато відмінностей. Таким чином, випадкові ліси набагато успішніші за дерева рішень, тільки якщо дерева різноманітні та прийнятні.

Таблиця 2.1

Відмінності між деревом рішень та Random forest

Дерева рішень	**Random forest**
Єдине дерево рішень швидше в обчисленнях.	створюються з підмножин даних, а кінцевий результат базується на середньому або мажоритарному ранжуванні; таким чином, проблема надмірної підгонки вирішена.
Навчається на всьому наборі даних і може виявити перенавчання, особливо якщо дані досить складні або шумні.	використовує метод багатократного бутстрепу (bagging), вибираючи випадковим чином підмножину даних для навчання кожного дерева.

Алгоритм random forest має ряд ключових переваг і труднощів, коли він використовується для задач класифікації або регресії. Деякі з них включають:

Основні переваги:

- *Зменшення ризику переобладнання:* дерева рішень мають ризик переобладнання, оскільки вони, як правило, щільно підходять для всіх зразків у навчальних даних. Тим не менш, у випадку, коли у випадковому лісі існує велика кількість дерев рішень, класифікатор не буде перебільшувати модель, оскільки загальна дисперсія та помилка прогнозу зменшуються за допомогою усереднення некорельованих дерев.
- *Забезпечує гнучкість*: випадковий ліс добре підходить для регресійних і класифікаційних завдань, тому він популярний серед дослідників даних. Класифікатор випадкового лісу також є корисним інструментом для оцінки відсутніх значень, оскільки він підтримує точність у випадках, коли частина даних відсутня, завдяки розташуванню функцій.

- *Легко визначити значення функції*: випадковий ліс полегшує оцінку значення змінної або її внеску в модель. Важливість функції можна визначити різними способами. Зазвичай використовуються важливість Джіні та середнє зменшення домішок (MDI) для визначення того, наскільки зменшується точність моделі, коли певна змінна виключається. Тим не менш, ще одним критерієм важливості є важливість перестановки, також відома як точність зниження середнього значення (MDA). MDA оцінює середнє зниження точності шляхом випадкової перестановки значень ознак у зразках oob.

Основні виклики:

- Процес, що потребує багато часу: випадкові алгоритми лісу можуть надавати точніші прогнози, оскільки вони можуть обробляти великі набори даних, але вони також можуть працювати повільно, оскільки вони обчислюють дані для кожного окремого дерева рішень.
- Вимагає більше ресурсів: випадкові ліси повинні мати більше ресурсів для обробки більшої кількості наборів даних.
- Більш складний: передбачення окремого дерева рішень легше інтерпретувати порівняно з лісом із них.

Зміна клімату є глобальною проблемою, яка має значні наслідки для суспільства та навколишнього середовища. Машинне навчання є потужним інструментом для моделювання зміни клімату, оскільки глобальне потепління та викиди парникових газів зростають.

Застосовуючи потенціал ідей, керованих даними, машинне навчання дає промінь надії на вирішення екологічних проблем і спрямовує нас до більш сталого майбутнього.

- Необхідність передового моделювання зміни клімату. Щоб прийняти рішення щодо змін клімату, необхідно мати точне та складне кліматичне моделювання. Новий метод підвищення точності кліматичних моделей і можливостей прогнозування доступний завдяки здатності машинного навчання обробляти великі та складні набори даних.

- Методи машинного навчання в науці про клімат**.** Алгоритми машинного навчання, такі як нейронні мережі та моделі глибокого навчання, надають вченим-кліматологам потужні інструменти для аналізу історичних кліматичних даних і визначення закономірностей, які спричиняють мінливість клімату. Розкриваючи ці заплутані зв’язки, дослідники отримують цінну інформацію про фактори, що впливають на зміну клімату.
- Прогнозування екстремальних погодних явищ. Точне прогнозування надзвичайних погодних явищ є одним із найважливіших завдань кліматології. Підхід машинного навчання, керований даними, дозволяє точніше прогнозувати, допомагаючи громадам краще підготуватися та реагувати на екстремальні події, такі як урагани, повені та посухи.
- Пом'якшення наслідків зміни клімату та адаптація. Машинне навчання відіграє важливу роль у розробці ефективних стратегій пом’якшення наслідків зміни клімату та адаптації. Аналізуючи різні кліматичні сценарії та визначаючи райони, вразливі до екологічних ризиків, політики можуть впроваджувати цільові заходи для зменшення викидів і захисту громад.

2.1.2 Виявлення патернів та трендів за допомогою методів машинного навчання

Отже, як машинне навчання виявляє закономірності та тенденції? Пропонуємо зазирнути за завісу:

Все починається з того, що алгоритми машинного навчання отримують величезну кількість даних. Уявіть собі гори даних про продажі, прогнози погоди, відгуки клієнтів або будь-яку іншу інформацію, що стосується ваших конкретних потреб. Чим більше даних алгоритми можуть поглинути, тим краще вони виявляють приховані в них коштовності.

Наступним кроком відбувається підготовка даних, виокремлення ключових характеристик і структурування їх таким чином, щоб алгоритми могли легко зрозуміти і проаналізувати. Це схоже на організацію вашої комори, що полегшує виявлення закономірностей у ваших кулінарних звичках.

Різні алгоритми машинного навчання, кожен з яких має свої сильні та слабкі сторони, запускаються на підготовлених даних. Деякі, як-от регресійні моделі, чудово знаходять числові зв'язки та прогнозують майбутні значення. Інші, як-от алгоритми кластеризації, групують схожі точки даних разом, виявляючи приховані сегменти та угруповання у ваших даних.

Навчання на моделях. Їх потрібно тренувати на даних, як студентів, що старанно готуються до іспиту. Таке навчання передбачає надання алгоритмам мічених прикладів, навчання їх розпізнавати закономірності та асоціювати їх з конкретними результатами.

Розпізнавання образів: нарешті, навчені алгоритми сканують дані за допомогою нових знань. Вони виявляють повторювані закономірності, кореляції та тенденції, які інакше залишилися б прихованими у звичайному тексті або цифрах.

2.1.3 Інноваційні підходи до використання штучного інтелекту в кліматології.

Рада Євросоюзу повідомляє, що досягла попередньої згоди щодо проекту узгоджених правил штучного інтелекту (ШІ), також відомого як закон про штучний інтелект. Суть проекту регламенту полягає в тому, щоб переконатися, що системи штучного інтелекту, які працюють на європейських ринках і використовуються в ЄС, є безпечними, а також що вони дотримуються основних прав і принципів ЄС. Крім того, ця важлива пропозиція спрямована на стимулювання інвестицій і інновацій у сфері штучного інтелекту в Європі.

Основна законодавча ініціатива, Закон про штучний інтелект, має на меті сприяти приватним і державним компаніям, які розробляють і використовують безпечний і надійний штучний інтелект на єдиному ринку ЄС.

Основна ідея полягає в тому, щоб контролювати штучний інтелект відповідно до його здатності завдавати шкоди суспільству. Це робиться за допомогою підходу, що базується на ризиках, тобто суворіші правила застосовують до більш високих ризиків. Будучи першим законодавчим проектом такого типу в світі, він має здатність встановити глобальний стандарт для регулювання штучного інтелекту в інших

юрисдикціях, як це зробив GDPR, просуваючи європейський підхід до глобального регулювання технологій.

Порівняно з першою пропозицією Комісії, основні нові елементи попередньої угоди можна підсумувати таким чином:

- правила щодо систем штучного інтелекту високого ризику, які можуть спричинити системний ризик у майбутньому, а також щодо систем штучного інтелекту високого ризику;
- переглянуту систему управління з деякими правозастосовними повноваженнями на рівні ЄС;
- розширення переліку заборон, але з можливістю використання дистанційної біометричної ідентифікації правоохоронними органами в громадських місцях за умови дотримання гарантій;
- кращий захист прав, зобов'язавши розробників систем штучного інтелекту з високим рівнем ризику провести оцінку впливу на фундаментальні права перед запуском системи штучного інтелекту в експлуатацію.

Тимчасова угода охоплює такі аспекти:

- Визначення та сфера застосування

Щоб гарантувати, що визначення системи штучного інтелекту містить достатньо чіткі критерії для відмежування штучного інтелекту від більш простих систем програмного забезпечення, компромісна угода узгоджує визначення з підходом, запропонованим ОЕСР. Попередня угода також уточнює, що регламент не поширюється на сфери, що виходять за межі законодавства ЄС, і ні в якому разі не повинен впливати на компетенцію держав-членів щодо національної безпеки або будь-якої організації, на яку покладено завдання в цій сфері. Крім того, закон про штучний інтелект не застосовуватиметься до систем, які використовуються виключно у військових чи оборонних цілях.

Щоб гарантувати, що визначення системи штучного інтелекту містить достатньо чіткі критерії для відмежування штучного інтелекту від більш простих систем програмного забезпечення, компромісна угода узгоджує визначення з підходом, запропонованим ОЕСР. Попередня угода також уточнює, що регламент не

поширюється на сфери, що виходять за межі законодавства ЄС, і ні в якому разі не повинен впливати на компетенцію держав-членів щодо національної безпеки або будь-якої організації, на яку покладено завдання в цій сфері. Крім того, закон про штучний інтелект не застосовуватиметься до систем, які використовуються виключно у військових чи оборонних цілях.

- Класифікація систем ШІ як високо ризикових і заборонених практик ШІ

Компромісна угода передбачає горизонтальний рівень захисту, включаючи класифікацію високого ризику, щоб гарантувати, що системи штучного інтелекту, які не можуть спричинити серйозні порушення фундаментальних прав або інші значні ризики, не будуть зафіксовані. Для деяких видів використання ШІ ризик вважається неприйнятним, тому ці системи будуть заборонені в ЄС. Попередня угода забороняє, наприклад, когнітивні поведінкові маніпуляції , нецілеспрямоване вирізання зображень обличчя з Інтернету чи записів камер відеоспостереження, розпізнавання емоцій на робочому місці та в навчальних закладах, соціальне оцінювання, біометричну категоризацію для отримання конфіденційних даних, таких як сексуальна орієнтація чи релігійні переконання. переконання та деякі випадки інтелектуальної поліцейської діяльності для окремих осіб.

- Винятки правоохоронних органів

Враховуючи специфіку правоохоронних органів і необхідність зберегти їх здатність використовувати штучний інтелект у своїй життєво важливій роботі, було погоджено кілька змін до пропозиції Комісії щодо використання систем штучного інтелекту в правоохоронних цілях. Крім того, що стосується використання систем віддаленої біометричної ідентифікації в режимі реального часу в загальнодоступних місцях, у тимчасовій угоді роз’яснюються цілі, у яких таке використання є суворо необхідним для правоохоронних цілей і для яких правоохоронним органам має бути дозволено використовувати такі системи у виняткових випадках. Компромісна угода передбачає додаткові гарантії та обмежує ці винятки випадками жертв певних злочинів, запобіганням справжнім, поточним або передбачуваним загрозам, таким як терористичні атаки, і розшуку людей, підозрюваних у найтяжчих злочинах.

- Системи ШІ загального призначення та основні моделі

Було додано нові положення для врахування ситуацій, коли системи штучного інтелекту можна використовувати для багатьох різних цілей (штучний інтелект загального призначення), і де технологія штучного інтелекту загального призначення згодом інтегрується в іншу систему високого ризику. Попередня угода також стосується конкретних випадків систем ШІ загального призначення (GPAI). Попередня угода передбачає, що базові моделі повинні відповідати певним зобов'язанням щодо прозорості, перш ніж вони будуть розміщені на ринку. Було введено більш суворий режим для моделей фундаменту «високого удару». Це базові моделі, навчені великою кількістю даних і з розширеною складністю, можливостями та продуктивністю, значно вищими за середні, які можуть поширювати системні ризики вздовж ланцюжка створення вартості.

- Штрафи

Штрафи за порушення закону про штучний інтелект встановлювалися як відсоток від глобального річного обороту компанії-порушника за попередній фінансовий рік або заздалегідь визначеної суми, в залежності від того, яка з цих сум більша. Це становитиме 35 мільйонів євро або 7% за порушення заборонених додатків штучного інтелекту, 15 мільйонів євро або 3% за порушення зобов'язань щодо штучного інтелекту та 7,5 мільйонів євро або 1,5% за надання невірної інформації. Однак тимчасова угода передбачає більш пропорційні обмеження адміністративних штрафів для малих і середніх підприємств і стартапів у разі порушення положень закону про AI.

- Прозорість і захист основних прав

Попередня угода передбачає оцінку впливу на фундаментальні права перед тим, як високо ризикова система штучного інтелекту буде випущена на ринок розробниками. Користувачі системи штучного інтелекту високого ризику, які є публічними організаціями, також будуть зобов'язані зареєструватися в базі даних ЄС для систем штучного інтелекту високого ризику.

- Заходи підтримки інноваційної діяльності

З метою створення законодавчої бази, яка є більш сприятливою для інновацій, і сприяння нормативному навчанню на основі фактичних даних. Уточнено, що

регуляторні пісочниці штучного інтелекту, які мають створити контрольоване середовище для розробки, тестування та перевірки інноваційних систем штучного інтелекту, також повинні дозволяти тестування інноваційних систем штучного інтелекту в реальних умовах. Нові положення дозволяють тестувати системи штучного інтелекту в реальних умовах, за певних умов і застережних заходів.

- Набрання чинності

Попередня угода передбачає, що закон про AI має застосовуватися через два роки після набрання ним чинності, за деякими винятками для окремих положень.

Після попередньої домовленості буде продовжено роботу на технічному рівні, щоб завершити деталі нового регламенту. Текст передають представникам держав-членів ЄС на схвалення, та має бути підтверджений обома інституціями та пройти юридично-мовний перегляд перед офіційним ухваленням співзаконодавцями.

2.2 Інтерактивні візуалізації кліматичних даних

2.2.1 Сучасні підходи до візуалізації глобальних кліматичних змін

Сучасні підходи до візуалізації глобальних кліматичних змін використовують передові інструменти та техніки для ефективного подання складних кліматичних даних. Ось деякі з ключових підходів та технік:

- Інтерактивні веб-візуалізації: розвиток технологій веб-візуалізації дозволяє створювати інтерактивні графіки та карти, які користувачі можуть вивчати та аналізувати онлайн. Такі візуалізації дозволяють взаємодіяти з різними шарами даних та використовувати різні фільтри для деталізації інформації.
- Глобальні картографічні проекції: використання спеціальних картографічних проекцій дозволяє коректно та ефективно відображати глобальні зміни клімату на картах. Наприклад, широтно-довготні проекції часто використовуються для зображення глобальних шаблонів кліматичних змін.
- Тривимірні візуалізації: використання тривимірних графіків і моделей дозволяє краще розуміти просторові та географічні аспекти кліматичних змін. Це може

бути особливо корисним для візуалізації динаміки атмосферних та океанічних процесів.

- Теплові карти та градієнтні візуалізації: використання теплових карт дозволяє показати просторовий розподіл температур, опадів та інших кліматичних параметрів. Градієнтні візуалізації ефективно відображають зміни концентрації чи температурні аномалії на географічній карті.
- Анімації та часові ряди**:** створення анімацій та візуалізація часових рядів дозволяє демонструвати динаміку кліматичних змін протягом тривалого періоду. Це допомагає визначити тренди та сезонні варіації.
- Інтерактивні графіки та діаграми**:** використання інтерактивних графіків та діаграм дозволяє деталізувати конкретні аспекти кліматичних змін. Користувачі можуть вибирати різні параметри для відображення та взаємодіяти з графіками.
- Динамічні показники**:** створення динамічних показників та індексів, таких як індекси сухості чи температурні аномалії, дозволяє швидко визначати зони інтенсивних кліматичних змін.
- Використання сучасних програм для візуалізації**:** використання програм для візуалізації даних, таких як Tableau, D3.js, Plotly та інших, надає багато можливостей для створення вражаючих та ефективних візуалізацій кліматичних даних.

2.2.2 Використання візуалізації даних для Big Data

Зростання популярності великих даних і проектів аналізу даних зробили візуалізацію більш важливою, ніж будь-коли. Компанії все частіше використовують машинне навчання для збору величезних обсягів даних, які важко і повільно сортувати, розуміти та пояснювати. Візуалізація пропонує засіб пришвидшити це та представити інформацію власникам бізнесу та зацікавленим сторонам у зрозумілий для них спосіб.

Візуалізація великих даних часто виходить за рамки типових методів, які використовуються у звичайній візуалізації, наприклад кругових діаграм, гістограм і

корпоративних графіків. Натомість він використовує більш складні представлення, такі як карти тепла та діаграми температури. Для візуалізації великих даних потрібні потужні комп'ютерні системи для збору необроблених даних, їх обробки та перетворення на графічні представлення, які люди можуть використовувати для швидкого аналізу.

Хоча візуалізація великих даних може бути корисною, вона може створити кілька недоліків для організацій. Вони такі:

- Щоб максимально використовувати інструменти візуалізації великих даних, вам потрібно найняти спеціаліста з візуалізації. Щоб переконатися, що компанії оптимізують використання своїх даних, цей спеціаліст має вміти визначити найкращі набори даних і стилі візуалізації.
- Оскільки візуалізація великих даних вимагає потужного комп'ютерного обладнання, ефективних систем зберігання даних і навіть переходу в хмару, такі проекти часто вимагають участі IT-спеціалістів і керівництва.
- Статистика, отримана за допомогою візуалізації великих даних, буде настільки точною, наскільки точна візуалізована інформація. Таким чином, наявність людей і процедур, які відповідають за управління та контроль якості корпоративних даних, метаданих і джерел даних, є надзвичайно важливою.

Вирізняють такі найпопулярніші техніки візуалізації даних:

Лінійні діаграми. Це одна з найпростіших і поширених технік. Лінійні діаграми показують, як змінні можуть змінюватися з часом.

Діаграми площ. Цей метод візуалізації є різновидом лінійної діаграми; він відображає кілька значень у часовому ряді - або послідовність даних, зібраних у послідовні, рівновіддалені моменти часу.

Діаграми розсіювання. Ця техніка відображає зв'язок між двома змінними. Точкова діаграма має форму осей x і y з крапками для представлення точок даних.

Treemaps. Цей метод дозволяє переглядати ієрархічні дані у вкладеному форматі. Кожна категорія має розмір прямокутника, який пропорційний її відсотковому вмісту від загального розміру. Деревоподібні карти ідеально підходять для ситуацій, коли є кілька категорій і мета порівняння різних частин цілого.

Піраміди населення**.** У цій техніці використовується стовпчаста діаграма для відображення складного соціального наративу населення. Його найкраще використовувати, коли намагаєтеся відобразити розподіл сукупності.

2.2.3 Роль інтерактивних візуалізацій у комунікації наукових результатів

Вивчення складних предметів вимагає візуалізації та графіки. Тим не менш, науковці та науково-політичні програми рідко думають про те, як візуалізації можуть допомогти відкриттям, створенню цікавих і надійних звітів або підтримці онлайн-ресурсів, незважаючи на те, що вони визнають цю цінність. Досвід і знання в галузі науки, політики, комп'ютерних технологій і дизайну необхідні для створення доступних і неупереджених візуалізацій на основі складних і невизначених даних. Тим не менш, візуалізація рідко присутня в нашій науковій підготовці, організації чи співпраці. З розвитком нових політичних програм візуалізація інформації має дедалі більше місця як у науковій політиці, так і в роботі науковців. З іншого боку, існує більша ймовірність пропущених відкриттів, непорозумінь і, в найгіршому випадку, створення упередженості щодо досліджень, які легко продемонструвати.

Інтерактивні візуалізації відіграють важливу роль у комунікації наукових результатів, роблячи інформацію доступною та зрозумілою для різних аудиторій.

Інтерактивні візуалізації дозволяють наочно представити складні наукові дані, зробити їх більш зрозумілими та доступними для різних користувачів, навіть тих, хто не має спеціалізованих знань у відповідній області.

Інтерактивність дозволяє користувачам експериментувати, взаємодіяти з даними та вивчати різні аспекти інформації. Це полегшує вивчення деталей і допомагає зрозуміти складні взаємозв'язки.

Сприйняття тимчасових або просторових змін може покращитися за допомогою анімацій і динамічних змін. Це дозволяє краще відображати тенденції та динаміку даних.

Інтерактивні візуалізації можна адаптувати до потреб користувачів, надаючи їм можливість вибрати певні параметри, підлаштувати графіки та фільтрувати дані за уподобаннями.

Інтерактивні візуалізації можуть залучити увагу громадськості до наукових досліджень і зробити їх більш захоплюючими для спостерігачів. Розповсюдження наукових знань і залучення громадськості до обговорення важливих питань є важливими завданнями.

Під час презентацій або доповідей інтерактивні візуалізації можуть допомагати в реальному часі відповідати на запитання аудиторії та глибше досліджувати конкретні аспекти дослідження.

Активні візуалізації допоможуть у прийнятті рішень, дозволяючи людям самостійно досліджувати та аналізувати дані, що робить їх більш розумними.

Загалом, інтерактивні візуалізації стають потужним інструментом для комунікації наукових результатів, допомагаючи зробити науку більш доступною, зрозумілою та цікавою для різних аудиторій.

2.3 Використання Big Data як основи для моделювання кліматичних процесів

Аналіз великих даних допомагає визначити закономірності, тенденції та кореляції, які сприяють прийняттю обґрунтованих рішень і ефективним стратегіям. Успішний аналіз вимагає використання алгоритмів високоякісних даних із різних сховищ.

Концепція роботи з великими наборами даних існує десятиліттями. Поява «великих даних» була викликана прогресом у різних технологіях і комп’ютерних науках, у тому числі поширенням Інтернету та дедалі більшою оцифровкою інформації та систем. Ці розробки призвели до експоненціального збільшення обсягу даних, які генеруються, збираються та зберігаються різними системами та інструментами. Оскільки урядові чи міжнародні організації та промисловість усвідомили потенційну цінність цих даних для аналізу та прийняття рішень, термін «великі дані» став широко використовуватися для опису проблем і можливостей, пов’язаних з керуванням і вилученням інформації з великих і складних наборів даних.

Великі дані дозволяють збирати, аналізувати та інтерпретувати дані, що робить їх життєво важливими для кліматичного моделювання. Прилади на супутниках і датчики з усієї земної кулі збирають дані, щоб оцінити умови на Землі та передбачити майбутні події. Після цього аналітика використовується для обробки даних про клімат, таких як опади, температура, атмосферні умови та океанічні моделі. Вчені можуть створювати складні кліматичні моделі для моделювання майбутніх подій, використовуючи як історичні дані, так і дані в реальному часі. Це допомагає виявити моделі, тенденції та кореляції, які сприяють ефективному плануванню та прийняттю розумних рішень.

Одним із прикладів є прогноз нещодавнього зсуву в швейцарському селі Брінц на початку червня 2023 року. Використовуючи дані, зібрані з датчиків на землі, а також супутникові зображення, протягом багатьох років вчені активно відстежували рух гори, що оточує ідилічне село. Завдяки достатньому попередженню, заснованому на постійному аналізі даних, влада села змогла евакуювати місто ще до початку зсуву. За кілька тижнів гора справді сповзла, ледве оминувши село. Аналіз даних і дії на основі цих даних мали вирішальне значення для забезпечення безпеки людей.

Big Data мають невід’ємну проблему, пов’язану з неможливістю інтеграції чи змішування між неоднорідними наборами даних із диверсифікованих доменних сховищ. Дорожня карта IEEE IC Big Data Governance and Meta data: Дорожня карта стандартів спрямована на вирішення проблеми, яка дозволить зробити дані відкритими, доступними та повторно використаними за допомогою машинозчитуваної та дієвої стандартної інфраструктури даних.

Дані про якість також мають вирішальне значення для забезпечення якісного аналізу. У статті Big Data Challenges in Climate Science: Improving the Next-Generation Cyber infrastructure автори представляють потребу в покращеній кіберінфраструктурі для обробки великої кількості критичних наукових даних.

Розуміння наслідків зміни клімату та пом’якшення її наслідків вимагає співпраці міждисциплінарної групи дослідників, вчених та інженерів. Ці спеціалісти розробляють системи для вимірювання та інтерпретують результати, щоб зробити

висновки. Щоб визначити напрямок дій і прийняття рішень, спочатку потрібно провести багато досліджень і аналізу.

Зі швидким розвитком систем спостереження Землі буде запускатися все більше і більше супутників для різних видів спостереження Землі місії. Петабайти даних спостереження Землі були зібрані та накопичені в глобальному масштабі з безпрецедентною швидкістю, тому поява великих даних спостереження Землі надає людям нові можливості для кращого розуміння кліматичних систем.

2.4 Роль блокчейну в управлінні кліматичними даними

Технологія блокчейн, яку часто асоціюють з криптовалютами, знайшла унікальне та важливе застосування в управлінні кліматичними даними. У світі, який бореться з наслідками зміни клімату, точність кліматичних даних стає найважливішою для прийняття обґрунтованих рішень. У цій статті досліджується багатогранна роль блокчейну в революції в управлінні кліматичними даними, підкреслюючи його потенціал для підвищення точності та прозорості.

Однією із значних перешкод у традиційному управлінні кліматичними даними є сприйнятливість до маніпуляцій. Централізовані системи схильні до підробки даних, що викликає занепокоєння щодо надійності наданої інформації.

Хоча прозорість є життєво важливою для управління кліматичними даними, традиційні методи часто не надають необхідного рівня відкритості. У зв’язку з його децентралізованою природою блокчейн пропонує рішення, які роблять записи даних прозорими та захищеними від підробки.

Blockchain, за своєю суттю, є децентралізованою та розподіленою технологією бухгалтерської книги, яка записує транзакції в мережі комп’ютерів. Ключові особливості блокчейну включають незмінність, прозорість і безпеку за допомогою криптографічних алгоритмів.

Несанкціонованим змінам кліматичних даних запобігає незмінність записів у блокчейні. Коли дані записуються в блокчейн, вони стають незмінною та постійною частиною книги, що підвищує точність кліматичної інформації в історії.

Технологія блокчейн виявляється кардинальною у сфері екологічного фінансування, пропонуючи безпечну та прозору платформу для сприяння сталим інвестиціям. Завдяки використанню децентралізованої системи бухгалтерського обліку блокчейну ініціативи екологічного фінансування отримують додатковий рівень довіри та автентичності. Інвестори можуть перевірити достовірність показників сталого розвитку та переконатися, що їхні кошти спрямовані на законні кліматичні та екологічні проекти. Завдяки токенізації зелені активи можуть бути представлені в блокчейні, що полегшує відстеження їх використання та впливу протягом життєвого циклу. Така прозорість не тільки зміцнює довіру серед інвесторів, але й залучає ширше коло зацікавлених сторін, готових підтримати екологічні ініціативи, що в кінцевому підсумку сприяє позитивним екологічним результатам.

Крім того, стійкість блокчейну до втручання гарантує, що всі транзакції та дані, пов’язані з екологічним фінансуванням, залишаються незмінними. Така стійкість до маніпуляцій і шахрайства забезпечує міцну основу для створення довіри до сектору екологічних фінансів. Оскільки технологія блокчейн продовжує розвиватися та набуває все більшого поширення, вона може революціонізувати спосіб фінансування проектів, пов’язаних із кліматом. Застосовуючи рішення екологічного фінансування на основі блокчейну, ми можемо сприяти більш екологічному та стійкому майбутньому, де кліматичні та екологічні цілі ефективно підтримуються з гарантією автентичності та підзвітності.

Розумні контракти Blockchain пропонують інноваційний спосіб заохочення глобальних кліматичних зобов’язань. Країни можуть взяти зобов’язання щодо захисту клімату за допомогою депозиту, посилюючи виконання зобов’язань і зменшуючи ризик помилкових обіцянок. Недотримання лімітів компенсації вуглецю призведе до того, що депозит буде утримано або розподілено між країнами, які виконують свої зобов’язання.

Оскільки світ стикається зі зростаючими проблемами в управлінні зусиллями щодо сталого розвитку та величезними обсягами даних, блокчейн стає ключовим союзником. Удосконалення технологій призведе до нових варіантів використання,

посилюючи переваги та вплив блокчейну. Майбутнє, де безпечно обмінюватимуться даними, запобігатиме шахрайству, а сталість буде пріоритетом, цілком доступне.

Технологія блокчейн має величезний потенціал для революції в боротьбі зі зміною клімату та просуванні сталого майбутнього. Від забезпечення прозорості в ланцюгах постачання до стимулювання екологічно чистих практик і перевірки автентичності екологічного фінансування, універсальність блокчейну надає потужні інструменти для боротьби зі зміною клімату. Оскільки споживачі вимагають більшої прозорості, а організації переходять у бік сталого розвитку, впровадження рішень на блокчейні відіграватиме життєво важливу роль у створенні більш стійкого до клімату та справедливого світу.

2.5 Висновки до розділу

В даному розділі були розглянуті сучасні технології аналізу кліматичних умов такі як моделювання клімату, візуалізація даних та методи машинного навчання для аналізу даних. Ми дійшли висновку, що сучасні технології аналізу кліматичних умов дозволяють нам краще розуміти клімат Землі, виявляти тенденції і закономірності в кліматичних змінах, а також прогнозувати майбутні зміни клімату.

Візуалізація даних є важливим інструментом для представлення кліматичних даних у зручному для розуміння форматі. Візуалізація даних може допомогти нам краще зрозуміти кліматичні зміни та прийняти обґрунтовані рішення щодо їх пом'якшення.

Машинне навчання є надзвичайно потужним способом вирішення певного типу завдань у сфері кліматичних даних таких як прогнозування на основі наявних даних тощо.

Аналіз даних є важливим інструментом для виявлення тенденцій і закономірностей в кліматичних даних. Аналіз даних може використовуватися для виявлення змін клімату, а також для визначення причин цих змін.

РОЗДІЛ 3
РОЗРОБКА АЛГОРИТМУ ТА СИСТЕМИ АНАЛІЗУ КЛІМАТИЧНИХ ДАНИХ

3.1 Опис проблеми та новизна роботи

Кліматичні зміни в сучасному світі стають предметом глибокого дослідження, оскільки вони мають великий вплив на екологічні, економічні та соціальні аспекти життя на планеті. Зростання кількості та складності кліматичних даних вимагає ефективних та інноваційних підходів до їхнього аналізу. В даному розділі ми представимо результати нашої роботи з розробки алгоритму та системи аналізу кліматичних даних, які не лише відповідають на це завдання, але й вносять новаторський внесок у цю галузь.

3.1.1 Значення дослідження

Розробка алгоритму та системи аналізу кліматичних даних важлива не лише з точки зору розуміння змін клімату, але й у контексті вирішення практичних завдань, таких як прогнозування погоди, розробка стратегій адаптації та прийняття рішень в галузі екології та сталого розвитку.

Основною метою є створення надійного та ефективного інструменту, який дозволить науковцям та владі приймати обґрунтовані рішення на основі аналізу великих обсягів кліматичних даних.

3.1.2 Актуальність проблеми

Зміни клімату стають все більш непередбачуваними, що поставляє суспільство перед серйозними викликами. Сучасні технології та методи аналізу даних можуть бути важливими інструментами для розуміння цих змін і розробки стратегій їх адаптації. Цей розділ роботи присвячений створенню таких програм, які можуть працювати з великою кількістю реальних часових і просторових кліматичних даних.

3.1.3 Структура розділу

Розділ буде розглядати крок за кроком весь процес розробки алгоритму та системи аналізу. Починаючи від визначення вимог та обрання підходів, ми детально розглянемо різні етапи розробки, включаючи вибір моделей, обробку даних, реалізацію алгоритмів та валідацію результатів. Надалі ми звернемо увагу на можливості вдосконалення та розширення розробленої системи для подальшого використання та досліджень. У даній роботі розглядаються 2 типи моделей аналітики даних, які призначені для специфічних завдань: MLалгоритм Random Forest для прогнозування даних на основі наявного набору даних та Data Visualisation проєкт, в якому ми проаналізуємо набір даних для знаходження інсайдів, які допоможуть у розв’язку певних конкретних завдань.

Цей розділ є ключовим елементом всієї роботи, оскільки відображає наш підхід до вирішення завдань аналізу кліматичних даних та подальших досліджень у цьому напрямку. Ми сподіваємося, що розгляд цього розділу призведе до глибшого розуміння та оцінки наших розробок, а також стане стимулом для подальших досліджень у цій важливій галузі.

3.2 Вимоги до даного рішення

Розробка інноваційного рішення для аналізу кліматичних даних потребує чітко визначених та обґрунтованих вимог, які враховують якісні та кількісні аспекти використання розробленої системи. Цей розділ розглядає ключові вимоги до даного рішення, які визначать його ефективність, надійність та пристосованість до потреб користувачів.

3.2.1 Функціональні вимоги:

- Аналіз великого обсягу даних: система повинна забезпечувати можливість аналізу великої кількості кліматичних даних, враховуючи просторові та часові параметри.

- Моделювання та прогнозування: вимагається можливість використання розробленого рішення для моделювання кліматичних змін та прогнозування їхнього розвитку на основі наявних даних.
- Підтримка різних форматів даних: система повинна бути гнучкою та підтримувати різні формати кліматичних даних для забезпечення сумісності з різними джерелами даних.
- Масштабованість: система повинна бути масштабованою для ефективної роботи з ростом обсягу даних та забезпечувати високу продуктивність.

3.2.2 Нетехнічні вимоги

- Безпека та конфіденційність:забезпечення високого рівня безпеки та конфіденційності для збереження чутливих кліматичних даних.
- Підтримка користувачів: надати систему зручну для використання та наділену документацією, яка допомагатиме користувачам розуміти та ефективно використовувати рішення.
- Легкість впровадження: забезпечити простоту впровадження системи в різні середовища та здатність працювати на різних платформах.

3.3.3 Технічні вимоги

- Кросплатформеність: розробка повинна бути кросплатформеною, щоб забезпечити можливість використання на різних операційних системах.
- Використання сучасних технологій:Використання передових технологій програмування та обробки даних для підтримки високої ефективності та швидкодії.
- Розширюваність. Можливість розширення функціоналу та додавання нових алгоритмів аналізу без необхідності внесення глибоких змін.
- Відкритість та інтеграція: Підтримка відкритих стандартів та можливість інтеграції з іншими системами для обміну даними.

Ці вимоги є основою для створення системи, яка відповідає викликам аналізу кліматичних даних, які існують у наш час. Ми сподіваємося створити рішення, яке буде відповідати потребам науковців, екологів і природокористувачів і буде ефективним, гнучким і простим у використанні під час виконання цих вимог.

3.3 Програмна реалізація аналізу датасету (візуалізація даних)

Візуалізація даних стає ключовою складовою сучасного аналізу, оскільки вона дозволяє нам перетворювати сухі цифри в конкретні історії та висновки. В контексті кліматичного аналізу, де величезні обсяги даних потребують систематизації та інтерпретації, візуалізація стає керівним принципом. Вона допомагає виявляти патерни, визначати тренди та робити інформовані висновки, спираючись на візуальний аналіз даних.

Однією з головних завдань цього підрозділу це продемонструвати, як результати візуалізації можуть слугувати фундаментом для наукових висновків та реальних практичних рішень. Визначимо, як побудовані графіки та діаграми відображають важливі аспекти кліматичних змін та як їх можна використовувати для прийняття обґрунтованих рішень в різних галузях.

В цьому розділі до реалізації аналізу датасету через призму візуалізації надасть читачеві повний огляд наших методів, допоможе зрозуміти важливість візуалізації в аналізі кліматичних даних та прокладе шлях для подальших досліджень.

3.3.1 – Загальний огляд набору даних

Для використання даного методу ми застосували набір даних, який містить інформацію про погодні умови у Мадриді у 1997-2005 роках, включаючи максимальну/мінімальну/середню температуру, точку роси, вологість, видимість і швидкість вітру, а також опади, хмарність і напрямок вітру. Файл складається із наступних даних:

- Дата
- Температура

- Мінімальна температура
- Максимальна температура
- Точка роси
- Середня точка роси
- Мінімальна точка роси
- Максимальна вологість
- Мінімальна вологість
- Середня вологість
- Максимальний рівень моря
- Середній рівень моря
- Мінімальний рівень моря
- Тиск
- Напрямок вітру
- Максимальна видимість
- Мінімальна видимість
- Середня видимість
- Максимальна швидкість вітру
- Мінімальна швидкість вітру
- Середня швидкість вітру
- Хмарність

```
In [7]: df.columns
Out[7]: Index(['CET', 'Max TemperatureC', 'Mean TemperatureC', 'Min TemperatureC',
               'Dew PointC', 'MeanDew PointC', 'Min DewpointC', 'Max Humidity',
               ' Mean Humidity', ' Min Humidity', ' Max Sea Level PressurehPa',
               ' Mean Sea Level PressurehPa', ' Min Sea Level PressurehPa',
               ' Max VisibilityKm', ' Mean VisibilityKm', ' Min VisibilitykM',
               ' Max Wind SpeedKm/h', ' Mean Wind SpeedKm/h', ' Max Gust SpeedKm/h',
               'Precipitationmm', ' CloudCover', ' Events', 'WindDirDegrees'],
              dtype='object')
```

Рис. 3.1. Перелік стовпців даного датафрейму

Даний набір даних містить переважно дані типів int64та float64, що вказує на те, що дані є числовими. Дані типу "int64" є цілими числами, в той час як тип "float64" зазвичай використовується для зберігання числових значень з плаваючою точкою.

```
df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6812 entries, 0 to 6811
Data columns (total 23 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   CET                        6812 non-null   object
 1   Max TemperatureC           6810 non-null   float64
 2   Mean TemperatureC          6809 non-null   float64
 3   Min TemperatureC           6810 non-null   float64
 4   Dew PointC                 6810 non-null   float64
 5   MeanDew PointC             6810 non-null   float64
 6   Min DewpointC              6810 non-null   float64
 7   Max Humidity               6810 non-null   float64
 8    Mean Humidity             6810 non-null   float64
 9    Min Humidity              6810 non-null   float64
 10   Max Sea Level PressurehPa 6812 non-null   int64
 11   Mean Sea Level PressurehPa 6812 non-null  int64
 12   Min Sea Level PressurehPa 6812 non-null   int64
 13   Max VisibilityKm          5872 non-null   float64
 14   Mean VisibilityKm         5872 non-null   float64
 15   Min VisibilitykM          5872 non-null   float64
 16   Max Wind SpeedKm/h        6812 non-null   int64
 17   Mean Wind SpeedKm/h       6812 non-null   int64
 18   Max Gust SpeedKm/h        3506 non-null   float64
 19  Precipitationmm            6812 non-null   float64
 20   CloudCover                5440 non-null   float64
 21   Events                    1798 non-null   object
 22  WindDirDegrees             6812 non-null   int64
dtypes: float64(15), int64(6), object(2)
memory usage: 1.2+ MB
```

Рис. 3.2. Типи даних у датафреймі

3.3.2 Вирішення завдань, пов'язаних із вказаним датасетом.

1) Протягом вказаного періоду часу визначити відсоток днів, при яких спостерігалися опади. Який відсоток днів спостерігалась ясна погода без опадів.

Спершу вирахуємо кількість випадків випадання опадів:

```
precipitation_count = df[' Events'].notna().sum()

precipitation_count
```

```
In [8]: precipitation_count = df[' Events'].notna().sum()
        precipitation_count

Out[8]: 1798
```

Рис. 3.3. Кількість випадків опадів

Відповідно розрахуємо відсоток днів з опадами та виведемо на екран результати:

```
precipitation_percentage = (precipitation_count / len(df)) * 100
print(f"Відсоток днів з опадами: {precipitation_percentage:.2f}%")
# Розраховуємо відповідно відсоток днів без опдівтпавиводимл результати на
екран:
clear_percentage = 100 - precipitation_percentage
print(f"Відсоток ясних днів: {clear_percentage:.2f}%")
```

Дані результати були висвітлені у текстовому форматі. Для кращого візуального сприйняття інформації людиною використовуємо у даному випадку кругову діаграму (pie charts):

```
labels = ['Дні з опадами', 'Ясні дні']
sizes = [precipitation_percentage, clear_percentage]
colors = ['lightcoral', 'lightskyblue']
plt.pie(sizes, labels=labels, colors=colors, autopct='%1.1f%%',
startangle=90)
plt.title('Відсоток ясних днів та днів з опадами:')
plt.axis('equal')  # Equalaspectratioensuresthatthepieisdrawnas a circle.
plt.show()
```

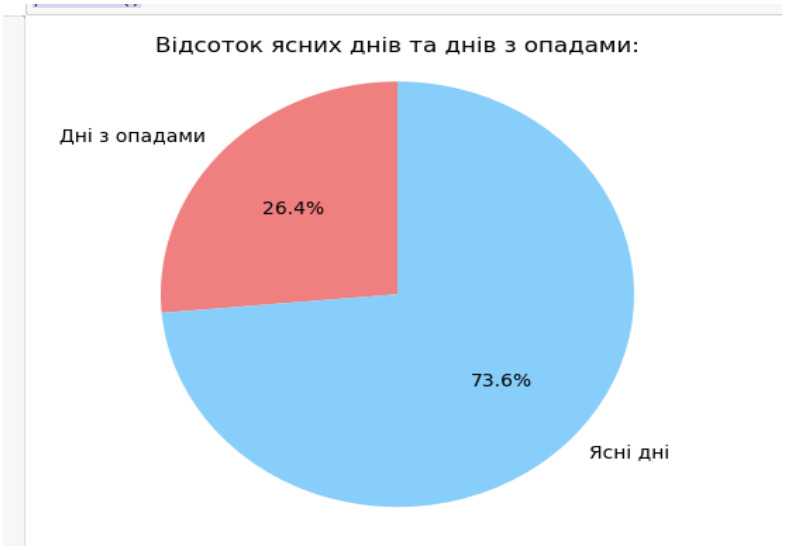


Рис. 3.4. Piechart, який показує відсоткове відношення ясних днів та днів з опадами

2) Визначити, в якому місяці даного періоду температура найвища. Припустимо, ви плануєте відпустку в Мадриді і сподіваєтеся на найтеплішу температуру. В якому місяці ви могли б запланувати поїздку?

Для вирішення даного завдання ми використовуємо перший стовпець «CET» який, як описано вище, має тип object. Для подальшої обробки інформації потрібно перетворити тип даного стовпця у data time object.

Після даного перетворення, є можливість витягти лиш місяць замість цілої дати, щоб вирахувати середню температуру за кожен місяць та знайти місяць з найвищою температурою:

```
df['CET'] = pd.to_datetime(df['CET'])
df['Month'] = df['CET'].dt.month
monthly_mean_temperature = df.groupby('Month')['MeanTemperatureC'].mean()
warmest_month = monthly_mean_temperature.idxmax()
print(f"Найтепліший місяць у Мадриді - {warmest_month}")
```

```
Найтепліший місяць у Мадриді - 7
```

Рис. 3.5. Результат пошуку найтеплішого місяця у Мадриді

Потрібно пам’ятати, що результат даного аналізу повинен бути легким для сприйняття. Так як даний результат – число, даний вибір висвітлення даних не є коректним. Для цього створимо додатково гістограму даного результату (barplot).

```
plt.figure(figsize=(10, 6))
monthly_mean_temperature.plot(kind='bar', color='green')
plt.title('Середня температура у Мадриді кожного місяця')
plt.xlabel('Місяць')
plt.ylabel('Середня температура(°C)')
plt.xticks(rotation=0)
plt.show()
```

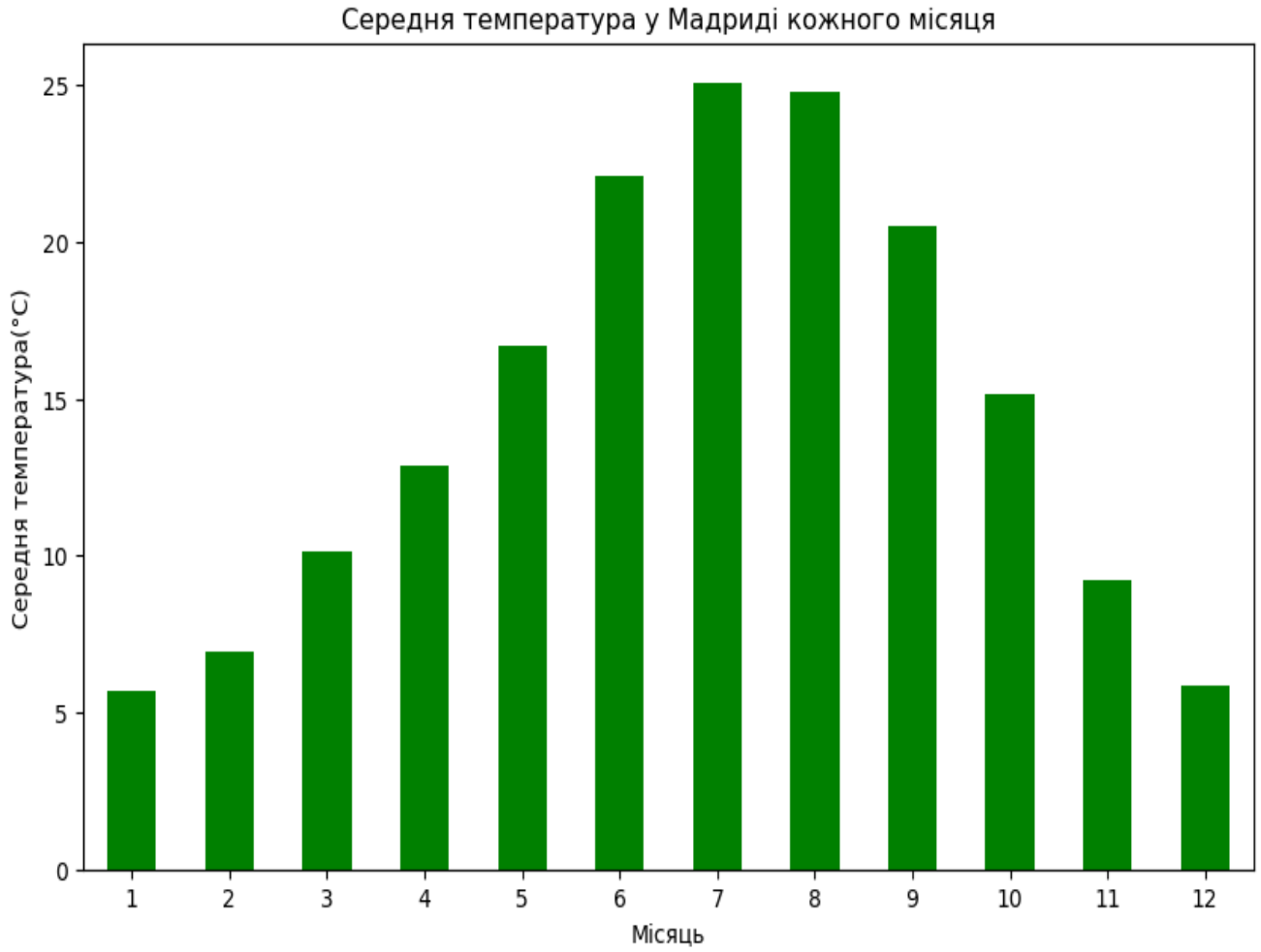


Рис. 3.6. Гістограма температур кожного місяця у Мадриді

3) Визначити в який день у вибірці в Мадриді був найшвидший порив вітру? Якою була погода в цей день?

Проаналізуємо дані про швидкість вітру у Мадриді, визначає час та швидкість найвищого пориву вітру та візуалізує ці дані на графіку:

```
# Знайдемо дату з найшвидшим поривом вітру
max_gust_date = df.loc[df[' MaxGustSpeedKm/h'].idxmax(), 'CET']
# Знайдемо відповідну інформацію про погоду на цю дату
max_gust_weather = df.loc[df['CET'] == max_gust_date]
# Виводимо результати
print(f" {max_gust_date},У Мадриді відбувся найшвидший порив вітру.")
print("Погода того дня:")
print(max_gust_weather)
```

```
 1997-11-06 00:00:00,У Мадриді відбувся найшвидший порив вітру.
Погода того дня:
           CET  Max TemperatureC  Mean TemperatureC  Min TemperatureC  \
309 1997-11-06              16.0               11.0               6.0

     Dew PointC  MeanDew PointC  Min DewpointC  Max Humidity   Mean Humidity  \
309        12.0             5.0            1.0          88.0            70.0

      Min Humidity  ...   Mean VisibilityKm   Min VisibilitykM  \
309           54.0  ...                 9.0                5.0

      Max Wind SpeedKm/h   Mean Wind SpeedKm/h   Max Gust SpeedKm/h  \
309                   58                    27                103.0

     Precipitationmm   CloudCover   Events  WindDirDegrees  Month
309              0.0          5.0     Rain             224     11

[1 rows x 24 columns]
```

Рис. 3.7. Дані найшвидшого пориву вітру у Мадриді та погода того ж дня

Побудуємо графік зміни швидкості поривів вітру з часом. відображає зміну швидкості поривів вітру в Мадриді протягом часу (Рис. 3.8):

```
plt.figure(figsize=(12, 6))
plt.plot(df['CET'], df[' MaxGustSpeedKm/h'], marker='o', linestyle='-', color='b', label='WindGustSpeed')
plt.scatter(max_gust_date, df.loc[df['CET'] == max_gust_date, ' MaxGustSpeedKm/h'], color='red', label='MaxGustDate')
plt.title('Зміна швидкості поривів вітру в Мадриді з часом')
plt.xlabel('Дата')
plt.ylabel(' MaxGustSpeed (Km/h)')
plt.xticks(rotation=45)
plt.legend()
plt.grid(True)
```

1) `plt.show()`

Як відрізняється середня видимість (км) у ясні дні від туманних?

Заміняємо значення NaN у стовпчику "Events" на заповнювач (наприклад, "NoEvent")та згрупуємо дані на основі стовпця "Events":

```
df[' Events'].fillna('NoEvent', inplace=True)
grouped_data = df.groupby(' Events')[' Mean VisibilityKm'].mean()
```

```
grouped_data
```

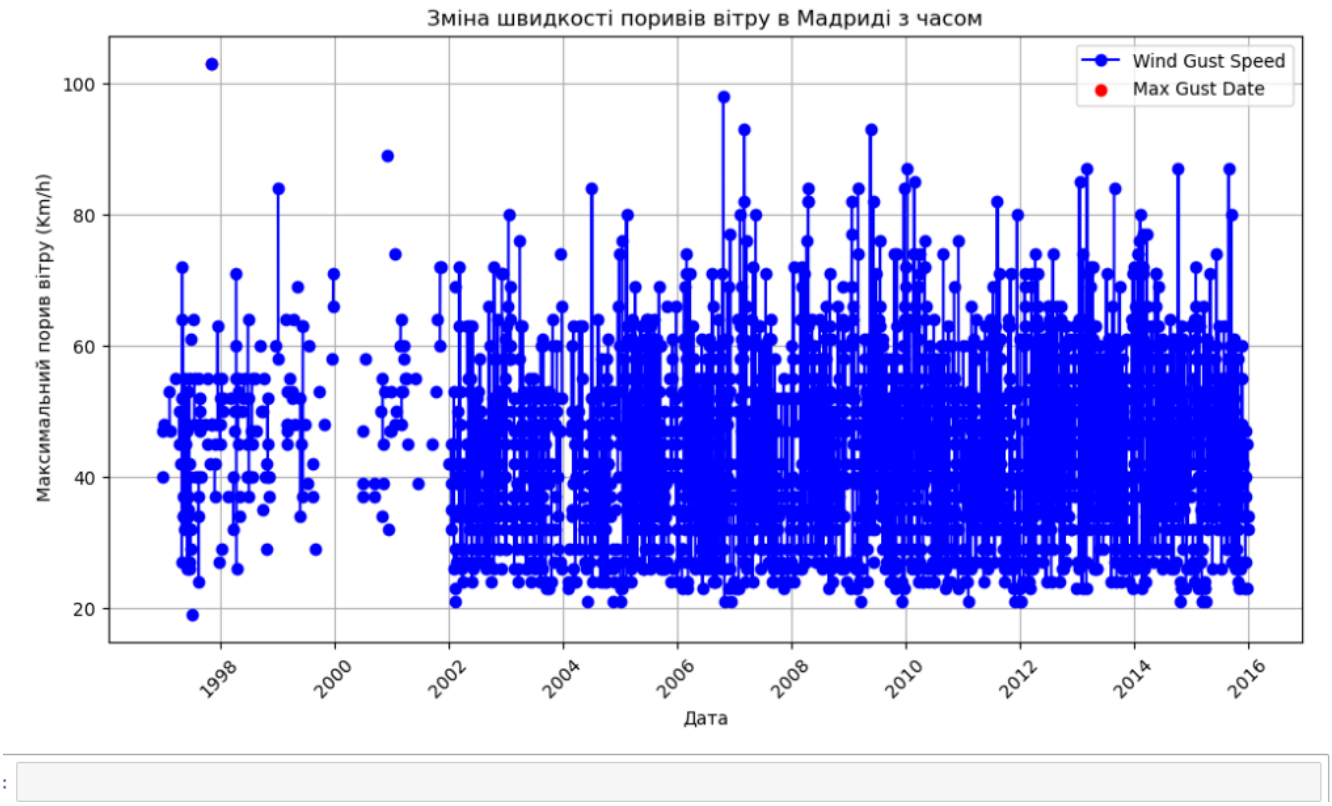


Рис. 3.8. Зміна швидкості поривів вітру в Мадриді з часом

```
Out[28]:  Events
Fog                          6.536481
Fog-Rain                     7.275362
Fog-Rain-Snow                5.000000
Fog-Rain-Thunderstorm        8.000000
Fog-Snow                     4.500000
Fog-Thunderstorm             7.000000
NoEvent                     12.635002
Rain                        10.209649
Rain-Hail                   14.000000
Rain-Hail-Thunderstorm      10.142857
Rain-Snow                    9.727273
Rain-Snow-Thunderstorm       7.000000
Rain-Thunderstorm           10.534413
Snow                         9.285714
Thunderstorm                10.822222
Tornado                     10.000000
Name:  Mean VisibilityKm, dtype: float64
```

Рис. 3.9. Групування даних у Data Frame за значенням в стовпці ' Events' та обчислення середнього значення для стовпця ' Mean Visibility Km' в кожній групі.

Витягніть дані видимості для ясних і туманних днів та отримуємо дані видимості для ясних і туманних днів:

```
   clear_visibility = grouped_data.get('Clear', 0)  # Use 0 as default if 'Clear' is not present
   clear_visibility #0
   fog_visibility = grouped_data.get('Fog', 0)  # Use 0 as default if 'Fog' is not present
   fog_visibility #6.536480686695279
```

Наведемо графік порівняння:

```
plt.bar(['Clear', 'Fog'], [clear_visibility, fog_visibility], color=['skyblue', 'gray'])
plt.title('Порівняння середньої видимості в ясні та туманні дні в Мадриді')
plt.xlabel('Погодні події')
plt.ylabel('Середній рівень витдимості (Km)')
plt.show()
```

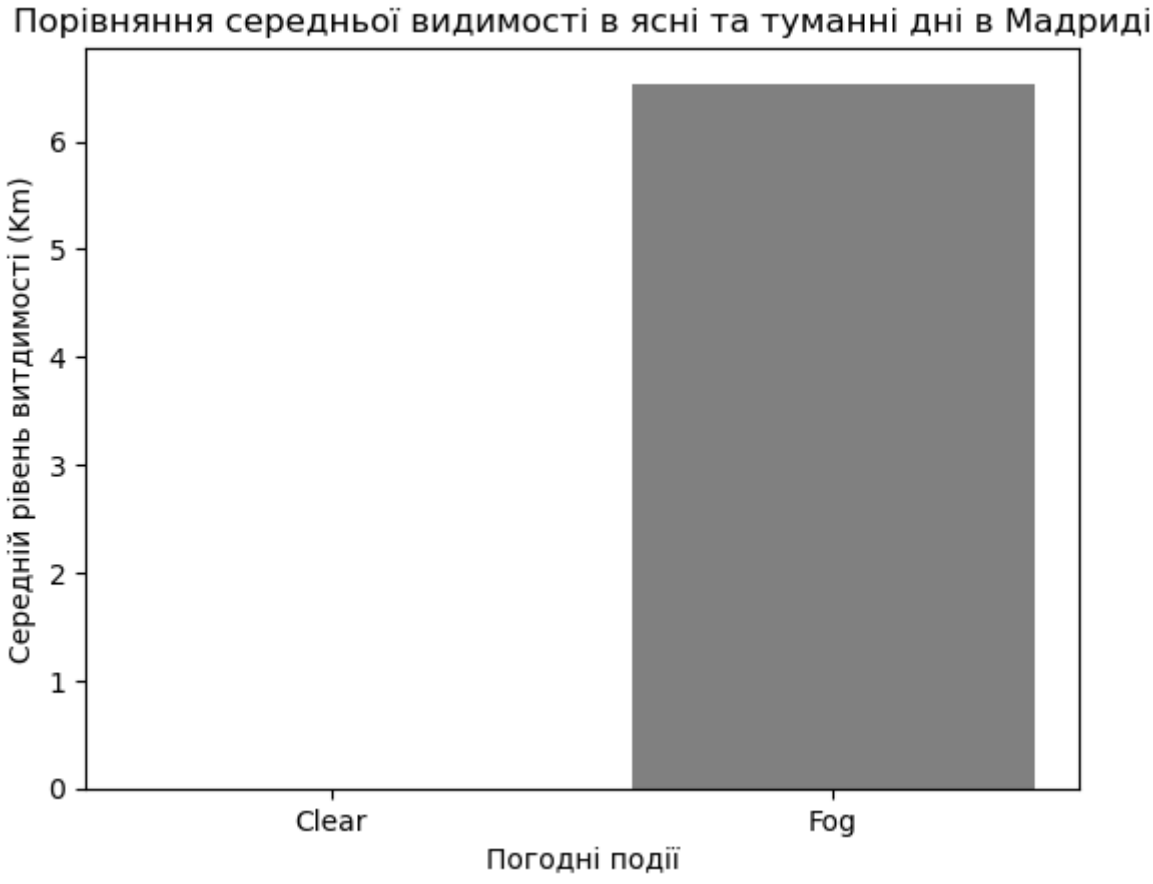


Рис. 3.10. Порівняння середньої видимості в ясні та туманні дні в Мадриді

Отже, для числових стовпців буде показано такі статистичні дані, як кількість, середнє значення, стандартне відхилення, мінімум, 25-й процентиль, медіана (50-й процентиль), 75-й процентиль і максимум. Для нечислових стовпців буде показано кількість, унікальність, верхній рівень і частоту.

3.4 Програмна реалізація прогнозування клімату за допомогою алгоритму Random Forest

Як було вище зазначено, існує ряд завдань, які неможливо виконати використовуючи тільки методи аналізу та візуалізації даних. Загалом у таких випадках комбінуються методи для вирішення більш складних завдань.

Методи машинного навчання в аналізі даних відіграють надзвичайно важливу роль, адже охоплюють певну сукупність функціоналу, яка вирішує проблеми, що виходять за межі здатностей традиційного аналітичного підходу та візуалізації.

Машинне навчання дозволяє автоматизувати процес виявлення залежностей, визначення складних патернів та прогнозування в поданні великих обсягів даних. Його здатність вирішувати задачі класифікації, регресії, кластеризації та інші, враховуючи велику кількість змінних та їхні взаємозв'язки, дозволяє отримувати більш точні та об'єктивні результати даних.

3.4.1 Загальний огляд набору даних

У даній роботі виконується прогнозування погоди на основі датасету кліматичних даних Австралії в період з 2008 по 2017 роки. Файл складається з даних, перелічених нижче:

- Дата
- Розташування (Location)
- Мінімальна та максимальна температура ()
- Швидкість вітру о 9:00 ранку та о 15:00
- Напрямок вітру о 9:00 ранку та о 15:00
- Вологість о 9:00 ранку та о 15:00

- Тиск о 9:00 ранку та о 15:00
- Хмарність о 9:00 ранку та о 15:00
- Температура о 9:00 ранку та о 15:00
- Чи були опади сьогодні (Boolean type)
- Чи будуть опади завтра (Boolean type)

Даний набір даних містить переважно дані типів objectта float64, що вказує на різноманітність форматів інформації. Дані типу "object" можуть включати рядки, текстові описи, або інші нечислові дані, в той час як тип "float64" зазвичай використовується для зберігання числових значень з плаваючою точкою.

Загальна кількість рядків становить 145460, загальна кількість стовпців - 23.

Короткий опис набору даних:

Цей набір даних містить близько 10 років щоденних спостережень за погодою з багатьох місць по всій Австралії.

```
In [11]: dataset.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 145460 entries, 0 to 145459
Data columns (total 23 columns):
 #   Column         Non-Null Count   Dtype  
---  ------         --------------   -----  
 0   Date           145460 non-null  object 
 1   Location       145460 non-null  object 
 2   MinTemp        143975 non-null  float64
 3   MaxTemp        144199 non-null  float64
 4   Rainfall       142199 non-null  float64
 5   Evaporation    82670 non-null   float64
 6   Sunshine       75625 non-null   float64
 7   WindGustDir    135134 non-null  object 
 8   WindGustSpeed  135197 non-null  float64
 9   WindDir9am     134894 non-null  object 
 10  WindDir3pm     141232 non-null  object 
 11  WindSpeed9am   143693 non-null  float64
 12  WindSpeed3pm   142398 non-null  float64
 13  Humidity9am    142806 non-null  float64
 14  Humidity3pm    140953 non-null  float64
 15  Pressure9am    130395 non-null  float64
 16  Pressure3pm    130432 non-null  float64
 17  Cloud9am       89572 non-null   float64
 18  Cloud3pm       86102 non-null   float64
 19  Temp9am        143693 non-null  float64
 20  Temp3pm        141851 non-null  float64
 21  RainToday      142199 non-null  object 
 22  RainTomorrow   142193 non-null  object 
dtypes: float64(16), object(7)
memory usage: 25.5+ MB
```

Рис. 3.11. Типи даних кожного стовпця датасету та кількість їх не порожніх значень

Rain Tomorrow є цільовою змінною для прогнозування. Вона означає - чи буде дощ наступного дня, так чи ні? У цьому стовпчику стоїть "Так", якщо дощ того дня був 1 мм або більше. У термінах машинного навчання ми визначаємо, що завтрашній дощ – це є наша залежна змінна, яка залежить від усіх функцій (стовпців cvs файлу), отже питання полягає в тому, яку функцію потрібно взяти до уваги і яку відкинути. При врахуванні усіх функцій, система потребуватиме велику кількість обчислень та іноді наша модель не передбачить хорошого результату.

Проаналізувавши даний датасет, можна припустити, що такі функції як розташування, температура, напрямок та швидкість вітру, тиск та хмарність є важливими факторами для проектування моделі. Дата, вологість та сонячне світло ми не включатимемо в розрахунки, так як вони не надають цінної інформації(дата не впливає на прогнозування, вологість та сонячне мають значення «NA»).

Визначимо основні проблеми даного датасету:

- Dirty Data - відсутні дані для Evaporation, Sunshine, Wind Gust Speed, Wind Speed 9am, Cloud9am, Rain Today, Rain Tomorrow.completion. Для Date Time, Rain Today, Rain Tomorrow.validity призначено неправильний тип даних
- Messy Data– date: може бути розділена на число, день і рік

Статистичний опис числових ознак у вашому наборі даних та транспонує цю таблицю:

Out[4]:

	count	mean	std	min	25%	50%	75%	max
MinTemp	143975.0	12.194034	6.398495	-8.5	7.6	12.0	16.9	33.9
MaxTemp	144199.0	23.221348	7.119049	-4.8	17.9	22.6	28.2	48.1
Rainfall	142199.0	2.360918	8.478060	0.0	0.0	0.0	0.8	371.0
Evaporation	82670.0	5.468232	4.193704	0.0	2.6	4.8	7.4	145.0
Sunshine	75625.0	7.611178	3.785483	0.0	4.8	8.4	10.6	14.5
WindGustSpeed	135197.0	40.035230	13.607062	6.0	31.0	39.0	48.0	135.0
WindSpeed9am	143693.0	14.043426	8.915375	0.0	7.0	13.0	19.0	130.0
WindSpeed3pm	142398.0	18.662657	8.809800	0.0	13.0	19.0	24.0	87.0
Humidity9am	142806.0	68.880831	19.029164	0.0	57.0	70.0	83.0	100.0
Humidity3pm	140953.0	51.539116	20.795902	0.0	37.0	52.0	66.0	100.0
Pressure9am	130395.0	1017.649940	7.106530	980.5	1012.9	1017.6	1022.4	1041.0
Pressure3pm	130432.0	1015.255889	7.037414	977.1	1010.4	1015.2	1020.0	1039.6
Cloud9am	89572.0	4.447461	2.887159	0.0	1.0	5.0	7.0	9.0
Cloud3pm	86102.0	4.509930	2.720357	0.0	2.0	5.0	7.0	9.0
Temp9am	143693.0	16.990631	6.488753	-7.2	12.3	16.7	21.6	40.2
Temp3pm	141851.0	21.683390	6.936650	-5.4	16.6	21.1	26.4	46.7

Рис. 3.12. Статистичний опис числових ознак

Виведемо кількість унікальних значень в кожному стовпці:

```
In [5]: df.nunique()

Out[5]: Date             3436
        Location           49
        MinTemp           389
        MaxTemp           505
        Rainfall          681
        Evaporation       358
        Sunshine          145
        WindGustDir        16
        WindGustSpeed      67
        WindDir9am         16
        WindDir3pm         16
        WindSpeed9am       43
        WindSpeed3pm       44
        Humidity9am       101
        Humidity3pm       101
        Pressure9am       546
        Pressure3pm       549
        Cloud9am           10
        Cloud3pm           10
        Temp9am           441
        Temp3pm           502
        RainToday           2
        RainTomorrow        2
        dtype: int64
```

Рис. 3.13. Кількість унікальних значень в кожному стовпці датасету

Проаналізуємо у датафреймі кількість пропущених значень і згенеруємо датафрейм, який містить кількість пропущених значень для кожного стовпця та відсоток від загальної кількості записів у кожному стовпці:

```
no = df.isnull().sum()
per= df.isnull().sum()/len(df)*100
missing_values = pd.DataFrame({"Загальний no.":no,
"відсоток":per}).sort_values(ascending=False, by='Загальний no.')
missing_values
```

3.3.2 Попередня обробка даних (Data Pre-Preprocessing)

У даному кроці використано один з декількох видів пре-процесингу даних: імпутація стовпців з використанням медіани, які містять пропущені значення:

```
   cols_1= [varforvarindf.columnsifdf[var].isnull().mean()*100 >5
anddf[var].isnull().mean()*100 <10]
   cols_1
```

```
num_cols = ["WindGustSpeed", "Pressure9am"    ,"Pressure3pm"]
cat_cols = ["WindDir9am"]
```

Наведений код генерує набір бокс-діаграм для числових стовпчиків (num_cols) у DataFramedf.

Для створення діаграм використовується бібліотека seaborn. Блок-схема для прийняття рішення про використання середнього, моди або медіани для імплікації:

```
plt.figure(figsize=(9,4))
for i, colsinenumerate(num_cols):
plt.subplot(1, 3, i+1)
sns.boxplot(data=df, x=cols)
plt.title(f"{cols}")
plt.tight_layout()
plt.show()
```

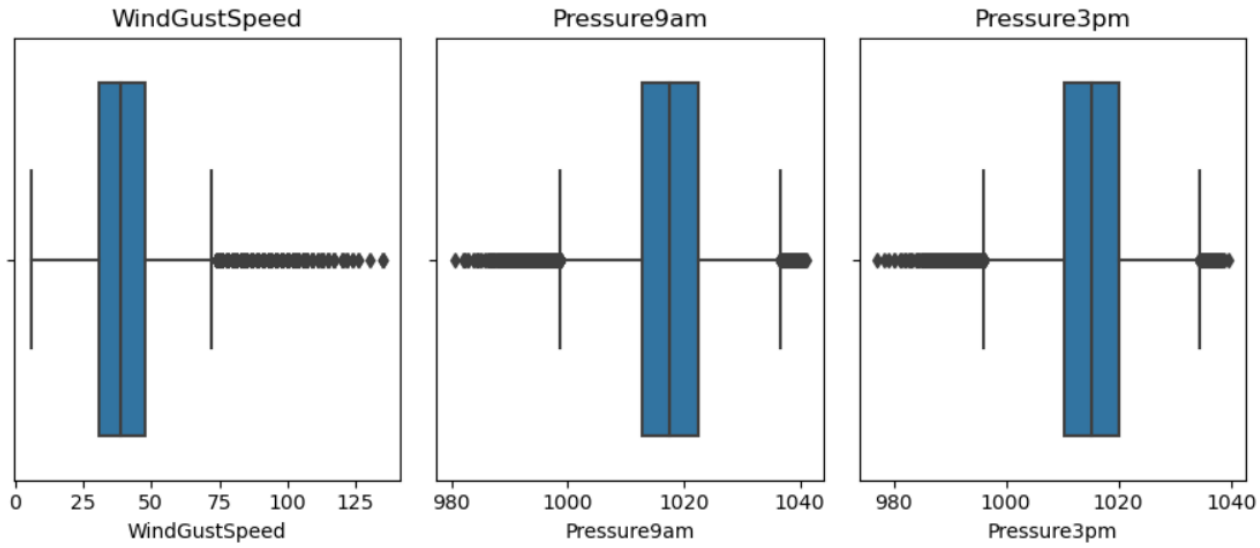


Рис. 3.14. Блок-схеми даних(стовпці Wind Gust Speed, Pressure9am, Pressure3pm)

З отриманих результатів можна зробити наступні висновки:

- дані викривлені. Є кілька або велика кількість точок даних, які виступають в ролі викидів (outliers).
- Outliers даних матимуть значний вплив на середнє значення, а отже, в таких випадках не рекомендується використовувати середнє значення для заміни пропущених значень.

- Використання середніх значень для заміни пропущених значень може не привести до створення якісної моделі.
- Тому ми використовуємо медіану для імплікації, оскільки медіана менш чутлива до пропусків, ніж середнє значення:

```
df["WindGustSpeed"]=df["WindGustSpeed"].fillna(df["WindGustSpeed"].median())
df["Pressure9am"]=df["Pressure9am"].fillna(df["Pressure9am"].median())
df["Pressure3pm"]=df["Pressure3pm"].fillna(df["Pressure3pm"].median())
df["WindDir9am"]=df["WindDir9am"].fillna(df["WindDir9am"].mode()[0])
```

3.4.3 Expolatory Data Analysis (Одновимірний аналіз)

Давайте перевіримо, збалансований наш набір даних чи ні?

Даний код створює два графіки для аналізу розподілу цільової змінної "Rain Tomorrow" в Data Framedf. Перший графік - countplot, відображає кількість записів для кожного унікального значення "Rain Tomorrow" та додає мітки зі значеннями кількості.

Другий графік - кругова діаграма (piechart), відображає відсотковий розподіл різних значень "Rain Tomorrow".

Обидва графіки вміщені у велику фігуру з розмірами 10x4 дюйма для порівняння розподілу цільової змінної за двома різними методами візуалізації:

```
plt.figure(figsize=(10,4))
plt.subplot(1, 2, 1)
ax= sns.countplot(data=df, x="RainTomorrow", palette="rocket")
forbarsinax.containers:
ax.bar_label(bars)
plt.title("Порушенняцільової ознаки:",fontweight="black", size=13)
plt.subplot(1, 2, 2)
#create piechart
df['RainTomorrow'].value_counts().plot(kind='pie',autopct='%0.1f%%')
plt.tight_layout()
plt.show()
```

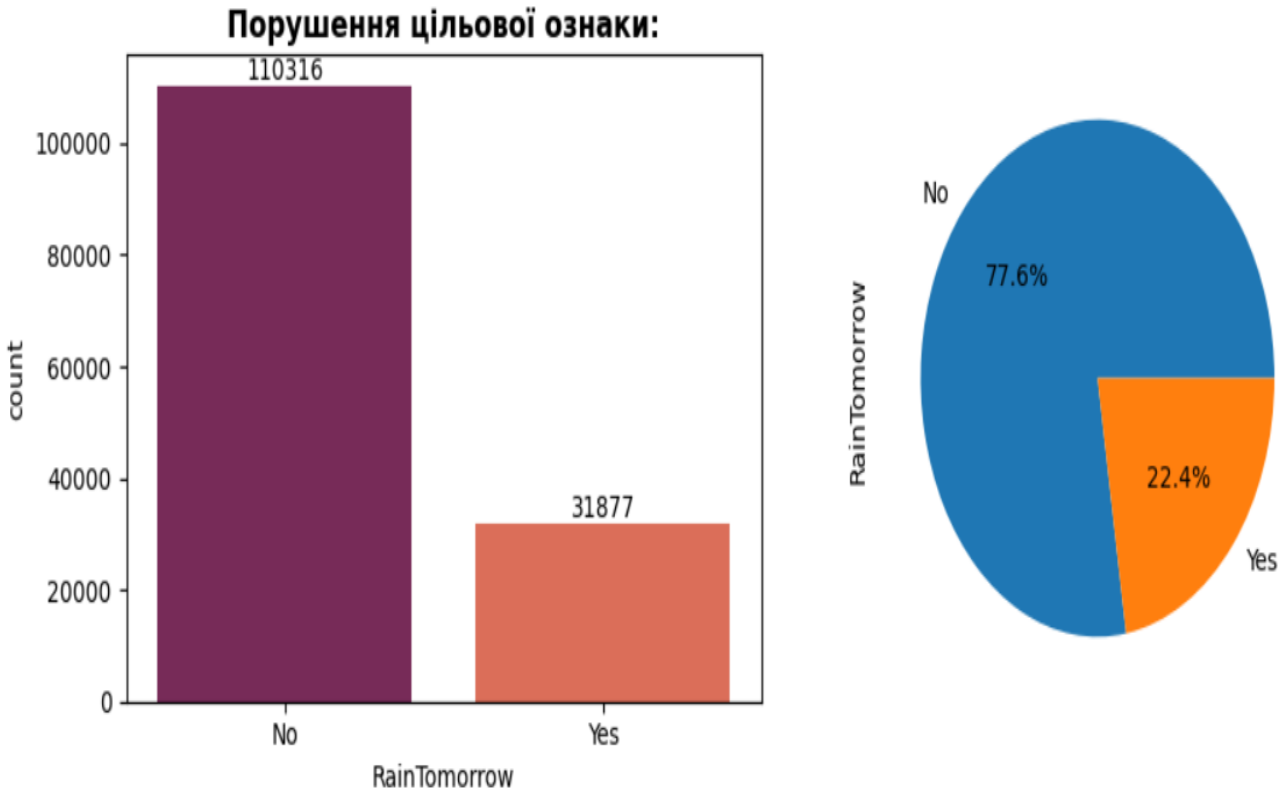


Рис. 3.15. Перевірка збалансованості набору даних

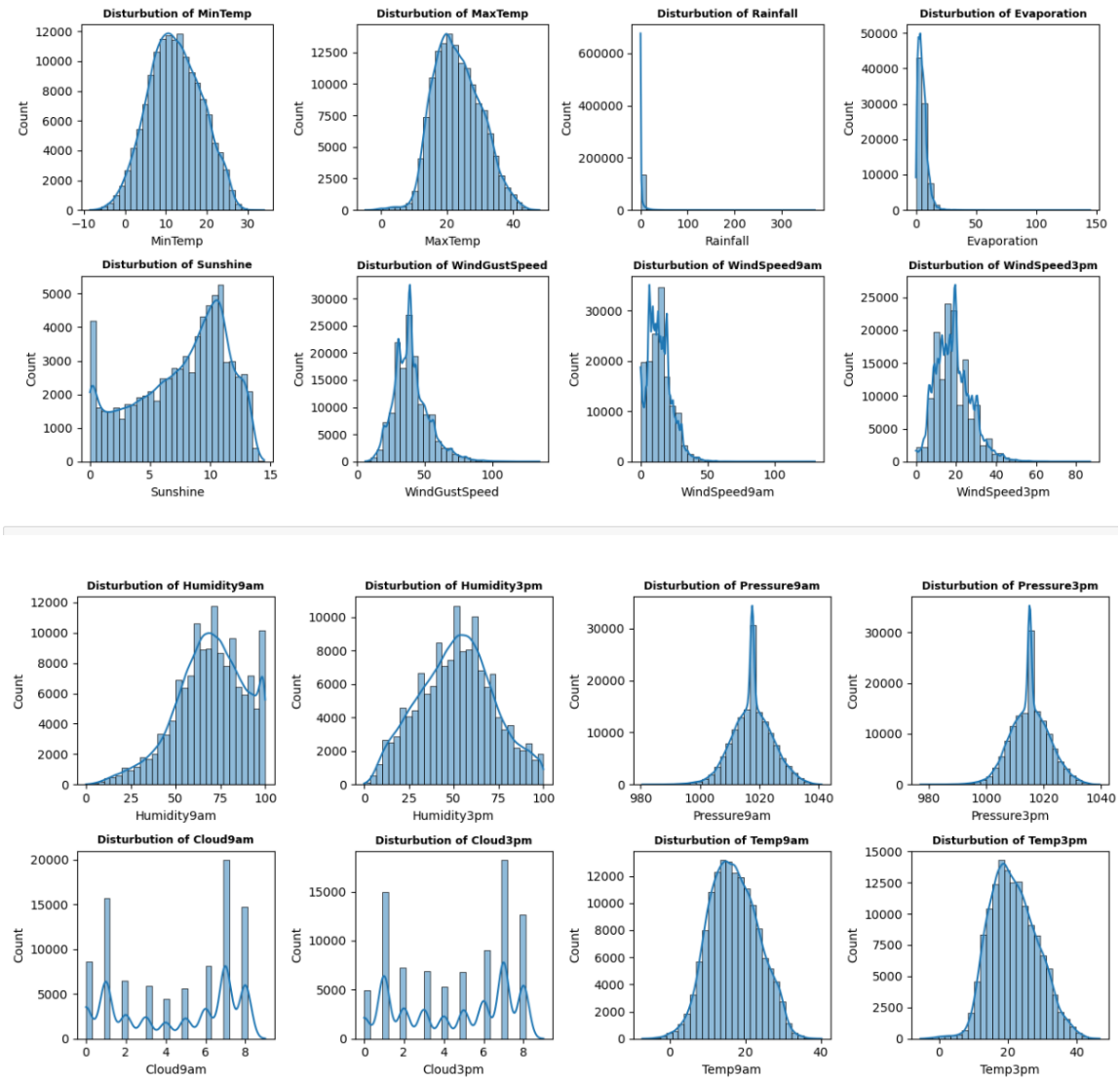
Оскільки ми можемо зробити висновок, що цільові класи ознак дуже незбалансовані, більшість даних належить до класу "Ні".

Тому в наступному розділі ми застосуємо певну техніку вибірки.

Розглянемо розподіл даних для числових стовпчиків. Розподіл кожної ознаки досліджується для визначення його форми, центральної тенденції та дисперсії:

```
cont_cols= df.select_dtypes(include="number").columns.to_list()
defcont_dist(data, feat):
plt.figure(figsize=(13,6))
for i, colsinenumerate(feat):
plt.subplot(2, 4, i+1)
sns.histplot(data=data, x=cols ,kde=True, palette="rocket",bins=30)
plt.title(f"Disturbutionof {cols}", fontweight="black", size=9)
plt.tight_layout()
plt.show()
cont_dist(df, cont_cols[0:8])
```

```
]: cont_dist(df, cont_cols[0:8])
```



Рисю 3.16. Візуалізація розподілу перших 8 числових ознак за допомогою гістограм та оцінки ядерної густини.

З наведених вище графіків збурень можна зробити висновок, що Min та Max Temp мають нормальне збурення. Тоді як інші характеристики не мають нормального розподілу. Збурення Pressure9am та Pressure3pm мають пік в центрі, збурення Temp9am та Temp3pm слідують за нормальним збуренням. Випаровування та Сонце не відповідають нормальному збуренням.

Розглянемо розподіл даних для категорійних стовпчиків. Розподіл кожної ознаки досліджується для визначення його форми, центральної тенденції та дисперсії:

Цей код використовує бібліотеку seaborn для створення набору countplot графіків, які візуалізують кількість випадків для кожної категорійної ознаки у датафреймі. Для кожної ознаки вибирається окремий підграфік, в якому відображається стовпчикова діаграма, а назва графіку відповідає назві відповідної категорії.

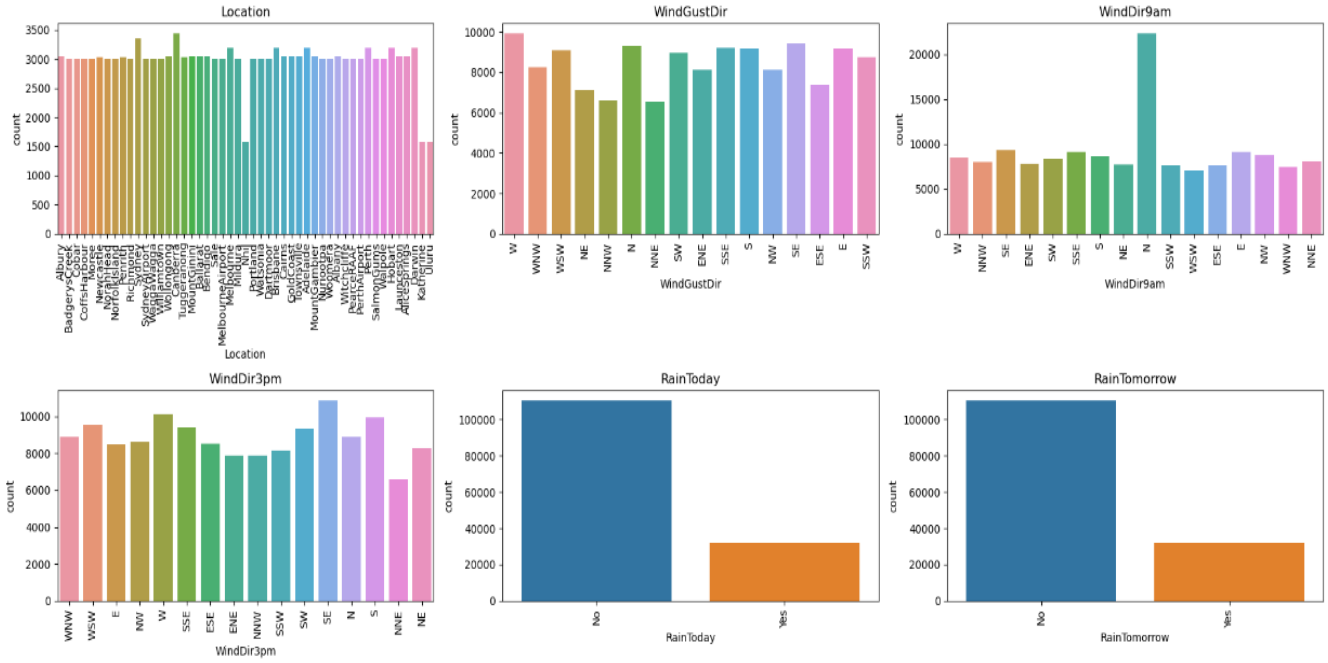


Рис. 3.17. Кількість випадків для кожної категорійної ознаки

3.4.4 Багатовимірний аналіз

Даний аналіз передбачає оцінку декількох змінних (більше двох) для виявлення будь-якого можливого зв'язку між ними.

Даний код створює теплову карту кореляції для числових ознак у Data Framedf. Спочатку визначається матриця кореляції для вибраних числових ознак. Потім використовуючи бібліотеку Seaborn, код генерує теплову карту, де кожен квадрат представляє кореляційний коефіцієнт між двома ознаками. Значення кореляції позначаються кольорами: від темного фіолетового (найменша кореляція) до яскравого жовтого (найбільша кореляція). Також на карті виводяться самі значення кореляції за допомогою анотацій. Все це дозволяє вам швидко оцінити ступінь взаємозв'язку між числовими ознаками в наборі даних.

```
plt.figure(figsize=(10,8))
```

```
corr= df[cont_cols].corr()
sns.heatmap(corr, annot=True, cmap='viridis', linewidths=0.1)
plt.show()
```

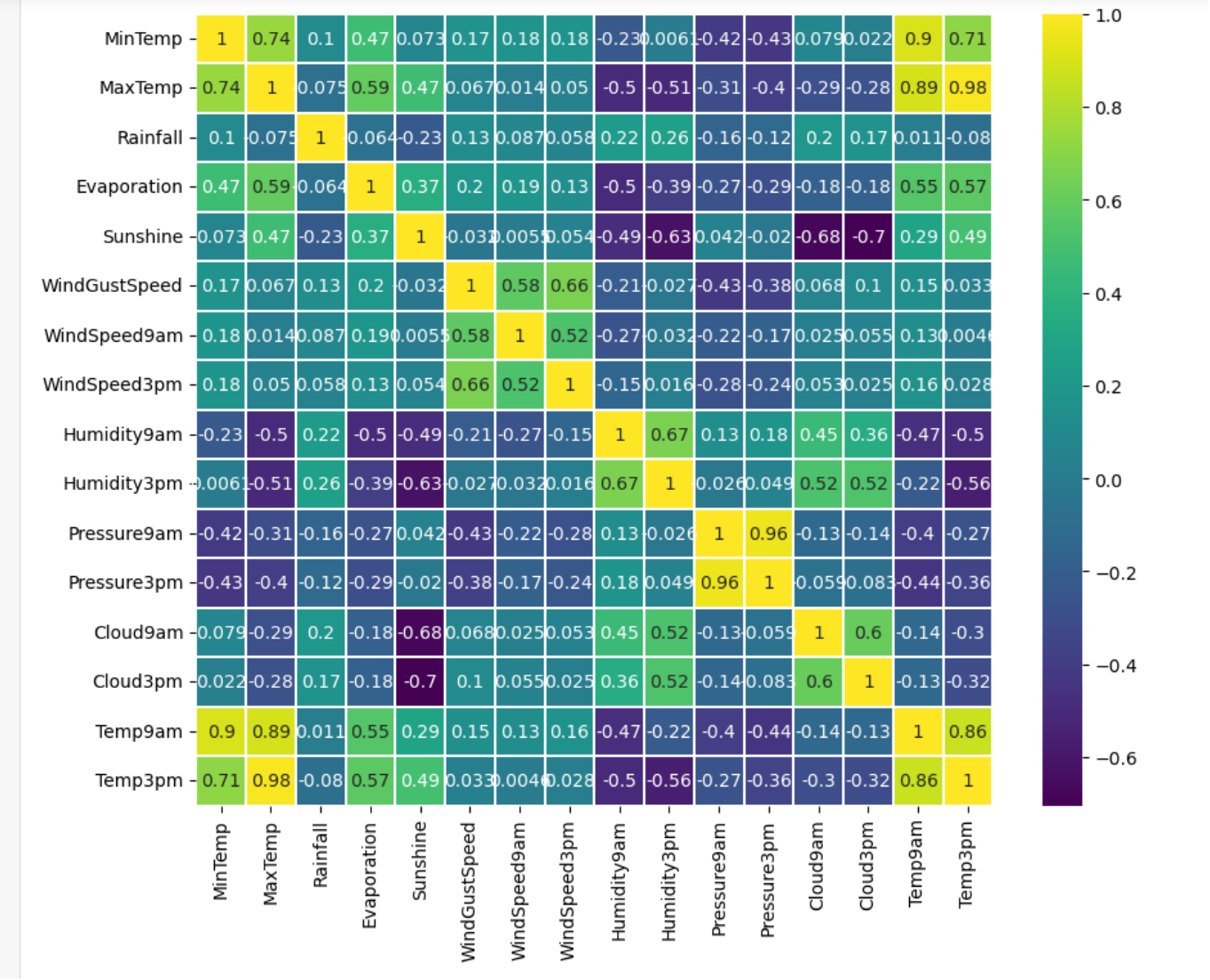


Рис. 3.18. Теплова карта кореляції для числових ознак

3.4.5 Label Encoding

Даний метод перетворює категоріальні стовпці в числові, щоб їх можна було використовувати в моделях машинного навчання, які приймають лише числові дані. Це важливий етап попередньої обробки в проекті машинного навчання. Зробимо одне гаряче кодування категорійної змінної та отримуємо k-1 фіктивних змінних після одного гарячого кодування:

```
cat_cols
```

```
cat_cols.remove("RainToday")
df = pd.get_dummies(df, columns=cat_cols, drop_first=True)
# previewthedatasetwithhead() method
df.head()
(145460, 112)
```

3.4.6 Відокремлення лейблів та ознак для модельних тренувань (включаючи масштабування функції)

```
X = df.drop(["RainTomorrow"], axis=1)
y = df["RainTomorrow"]
fromsklearn.preprocessingimportRobustScaler
scalar = RobustScaler()
X_scaled = scalar.fit_transform(X)
```

Розділюємо дані на окремі навчальні та тестові набори. Тепер у нас є набір даних X_train, готовий до завантаження в класифікатор логістичної регресії. Я зроблю це наступним чином.

```
fromsklearn.model_selectionimporttrain_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)
```

3.4.7 Модель навчання та оцінювання

Побудуємо модель логістичної регресії на навчальній множині. Створюємо екземпляр моделі:

```
fromsklearn.linear_modelimportLogisticRegression
logreg = LogisticRegression(solver='liblinear', random_state=0)
logreg.fit(X_train, y_train)
```

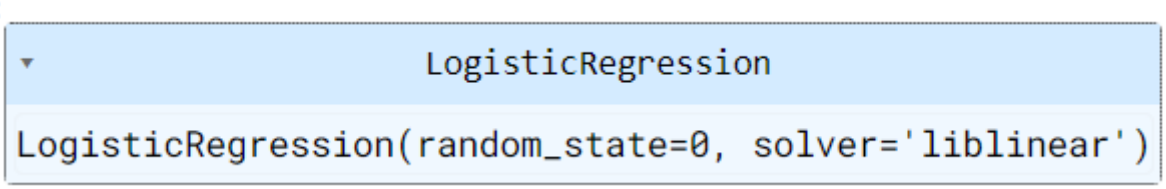


Рис. 3.19. Модель логістичної регресії

```
y_pred_test = logreg.predict(X_test)
y_pred_test
fromsklearn.metricsimportaccuracy_score
print('Modelaccuracyscore: {0:0.4f}'. format(accuracy_score(y_test,
y_pred_test)))
```

Model accuracyscore: 0.8513

Перевіряємо, чи немає надмірного або недостатнього прилягання:

```
print('Trainingsetscore: {:.4f}'.format(logreg.score(X_train, y_train)))

print('Testsetscore: {:.4f}'.format(logreg.score(X_test, y_test)))
```

```
Training set score: 0.8553
Test set score: 0.8513
```

Рис. 3.20. Результати тренування нашої моделі

Ми бачимо, що оцінка точності тренувального набору становить 0.8532, тоді як оцінка точності тестового набору дорівнює 0.8519. Ці два значення є цілком порівнянними. Отже, немає випадків надмірного або недостатнього припасування.

Висновки до розділу

Розділ "Розробка алгоритму та системи аналізу кліматичних даних" включає дві ключові частини: візуалізація даних та прогнозування кліматичних умов за допомогою алгоритмів Random Forest.

У першій частині, візуалізація даних, ми використовуємо бібліотеку seaborn для побудови лінійного графіка, що відображає зміни максимальної швидкості вітру в залежності від часу. Цей графік дозволяє нам легко спостерігати та аналізувати динаміку вітрових умов у Мадриді. Крім того, ми відзначаємо найвищий порив вітру червоною точкою на графіку, щоб виділити його іншим кольором та підкреслити його значення у контексті інших даних.

У другій частині, ми використовуємо алгоритми Random Forest для прогнозування кліматичних умов. Ці алгоритми дозволяють нам врахувати різноманітні фактори та їх взаємодію, щоб отримати точний та надійний прогноз. Їх висока ефективність полягає у здатності обробляти велику кількість даних та автоматично враховувати їхню важливість для прогнозу.

У підсумку, розділ надає інтегрований підхід до аналізу кліматичних даних, починаючи з їх візуалізації для кращого розуміння динаміки, а завершуючи прогнозуванням за допомогою потужних алгоритмів Random Forest для покращення точності та надійності прогнозів кліматичних умов у майбутньому.

ВИСНОВКИ

У рамках даної магістерської роботи було вивчено та розглянуто кілька важливих елементів обробки кліматичних даних за допомогою методів аналізу даних.

Використання візуалізацій даних виявилося ефективним методом для представлення та розуміння складних кліматичних параметрів. Графічне відображення даних сприяло швидкому виявленню паттернів, тенденцій та важливих аномалій у кліматичних змінах.

У результаті проведеного аналізу виявлено, що візуалізація даних є важливим інструментом для подання та аналізу кліматичних даних. Вона сприяє не лише легкому сприйняттю інформації, але й допомагає виявити різноманітні взаємозв'язки та залежності між різними кліматичними параметрами. Візуалізація значно поліпшує розуміння основних закономірностей та допомагає визначити ключові аспекти кліматичних процесів.

Дослідження підтвердили ефективність використання моделі машинного навчання случайного лісу для прогнозування майбутніх змін клімату. Модель продемонструвала високу точність у визначенні тенденцій і прогнозуванні змін клімату. Зокрема, висока швидкість обробки та точність отриманих результатів ідентифікації та верифікації створюють надійні основи для майбутнього аналізу кліматичних умов.

Результати, отримані на даний момент, є важливим кроком уперед у розвитку галузі обробки кліматичних даних. Представлені методи виявилися не лише ефективними з точки зору дослідження, але вони також мають великий потенціал для практичного застосування. Використання засобів аналізу даних і машинного навчання стає життєво важливим для розуміння та прогнозування кліматичних явищ. Це пояснюється зростанням кількості кліматичних даних, їхньою складністю та актуальністю в контексті змін клімату.

Дані результати відкривають нові двері для подальшого дослідження та створює можливості для реалізації розроблених методів і моделей у реальному світі. Розширення сфери використання даних аналітики в кліматології сприятиме не лише

науковим відкриттям, але й розробці практичних стратегій для вирішення глобальних кліматичних викликів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1.Orlan Galiy, Aditzhan Omarkhan. "Climate Data Analysis and Climate modeling with Python" (PDF). Retrieved 2023-10-01

2. P Yiou, E Baert, MF Loutre (1996). Surveysin Geophysics. Springer.[in English]

3.Greg J. McInerny, MinChen, Robin Freeman, David Gavaghan, Miriah Meyer, Francis Rowland. Information visualization forscien ceand policy. Retrieved 2014-02-21.

4."Стимулювання інновацій та впровадження штучного інтелекту в Європі" [Електронний ресурс] – Режим доступу до ресурсу: https://finap.com.ua/stymulyuvannya-innovatsij-ta-vprovadzhennya-shtuchnogo-intelektu-v-yevropi/

5.Liuyi Chen, Bocheng Han, Xuesong Wang, Jiazhen Zhao, Wenke Yang, Zhengyi Yang. Machine Learning Methodsin Weatherand Climate Applications: A Survey. Retrieved 2023-09-09.

7.P Yiou, E Baert, MF Loutre. Spectral analysis of climate data. Surveysin Geophysics, - Springer, 1996.

8. B Hennemuth, S Bender, K Bülow, N Dreier. Statistical method sforthe analysis of simulate dandob served climate data: applied in projectsand institutions dealing with climate change impact and adaptation- 2013.

9.T Nocke, T Sterzel, M Böttinger, M Wrobel. Visualization of climate and climate change data: Anoverview. Digital earth summit 2008 - Citeseer

10. VM Nik. Climate Simulation of an Attic Using Future Weather Data Sets-Statistical Methods for Data Processing and Analysis (PDF). 2010

11.D Štaffenová, R Ponechal, P Ďurica. Climate data processing for needs of energy analysis, 2014.

12. A Smola. Introduction to machine learning [Електронний ресурс] – Режим доступу до ресурсу: https://www.scribd.com/document/450136647/smola-pdf

13. Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., & Denzler, J. (2019). Deep learning and process understanding for data-driven Earth system science. Nature, 566(7743), 195-204.

14. Elshorbagy, A., & Carey, S. K. (2017). Climate data and hydrological modeling: A review. Journal of Hydrology: Regional Studies, 12, 1-17.

15. Knutti, R., & Sedláček, J. (2013). Robustness and uncertainties in the new CMIP5 climate model projections. Nature Climate Change, 3(4), 369-373.

16. Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, communication & society, 15(5), 662-679.

17. Hansen, J., Ruedy, R., Sato, M., & Lo, K. (2010). Global surface temperature change. Reviews of Geophysics, 48(4).

18. IPCC Special Report on Global Warming of 1.5°C (2018). Intergovernmental Panel on Climate Change (IPCC).

19. Knutti, R., & Sedláček, J. (2013). Robustness and uncertainties in the new CMIP5 climate model projections. Nature Climate Change, 3(4), 369-373.

20. Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. Bulletin of the American Meteorological Society, 93(4), 485-498.

21. Wilks, D. S. (2011). Statistical methods in the atmospheric sciences (Vol. 100). Academic Press.

22. WMO (World Meteorological Organization) (2017). Guide to Climatological Practices (WMO-No. 100).

23. Thakur, J., & Kumar, A. (2016). Big data in environmental sciences: A review. Journal of King Saud University-Science, 30(3), 293-305.

24. Wahid, A., & Azam, F. (2019). Big Data Analytics in Climate Change: A Survey. Journal of King Saud University-Computer and Information Sciences.

25. Pebesma, E., & Bivand, R. (2005). Classes and methods for spatial data in R. R News, 5(2).

26. Ziehn, T., & Umlauf, N. (2017). Spatial dependence in meteorological variables: a comparison between two Gaussian random field models. Spatial Statistics, 21, 235-251.

27. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R News, 2(3), 18-22.

28. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

29. Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, D. R., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. Ecology, 88(11), 2783-2792.

30. Li, X., Zhou, W., & Li, S. (2017). A review of remote sensing of climate-related variables and their applications to forest health and vulnerability assessment. Science of the Total Environment, 598, 194-207.

31. Liao, W., & Noble, W. S. (2003). Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and functional relationships. Journal of Computational Biology, 10(6), 857-868.

32. Lunetta, R. S., & Lyon, J. G. (2004). Stratified random sampling of land cover in the United States for accuracy assessment of the National Land Cover Data Set (NLCD) class: results, analysis, and recommendations. Remote Sensing of Environment, 92(2), 345-358.

33. Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Springer.

34. Kassambara, A., & Mundt, F. (2020). Factoextra: Extract and visualize the results of multivariate data analyses. R package version, 1(3).

35. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). Springer.

36. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

37. Mitchell, T. M. (1997). Machine learning. McGraw-Hill.

38. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

39. Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1). MIT press Cambridge.

40. Ghil, M., & Vautard, R. (1991). Interdecadal oscillations and the warming trend in global temperature time series. Nature, 350(6316), 324-327.

ДОДАТКИ

Додаток А

Програмне рішення проєкту «Візуальний аналіз даних»

```
Import pandas as pd
Import numpy as np
Import seaborn as sns
Import matplotlib.pyplot as plt
df = pd.read_csv("Madrid-weather.csv")
df
df.info()
df.describe()
#Кількість випадків випадання опадів:
df.columns
precipitation_count = df[' Events'].notna().sum()
precipitation_count
#Відсотковий результат:
precipitation_percentage = (precipitation_count / len(df)) * 100
print(f"Відсоток днів з опадами: {precipitation_percentage:.2f}%")
# Розраховуємо відповідно відсоток днів без опдівтпавиводимл результати на
екран:
clear_percentage = 100 - precipitation_percentage
print(f"Відсоток ясних днів: {clear_percentage:.2f}%")
# Create a piechart
labels = ['Дні з опадами', 'Ясні дні']
sizes = [precipitation_percentage, clear_percentage]
colors = ['lightcoral', 'lightskyblue']

plt.pie(sizes, labels=labels, colors=colors, autopct='%1.1f%%', startangle=90)
plt.title('Відсоток ясних днів та днів з опадами:')
plt.axis('equal')  # Equal a spectratioensures that the pie is drawn as a
circle.
```

```
# Show the plot
plt.show()

# the most warm month in Madrid
df['CET'] = pd.to_datetime(df['CET'])
df['Month'] = df['CET'].dt.month

monthly_mean_temperature = df.groupby('Month')['MeanTemperatureC'].mean()
warmest_month = monthly_mean_temperature.idxmax()
print(f"Найтепліший місяць у Мадриді - {warmest_month}")

#  barplot
plt.figure(figsize=(10, 6))
monthly_mean_temperature.plot(kind='bar', color='green')
plt.title('Середня температура у Мадриді кожного місяця')
plt.xlabel('Місяць')
plt.ylabel('Середня температура(°C)')
plt.xticks(rotation=0)
plt.show()
# Знайдемо дату з найшвидшим поривом вітру
max_gust_date = df.loc[df[' MaxGustSpeedKm/h'].idxmax(), 'CET']

# Знайдемо відповідну інформацію про погоду на цю дату
max_gust_weather = df.loc[df['CET'] == max_gust_date]

# Виводимо результати
print(f" {max_gust_date},У Мадриді відбувся найшвидший порив вітру.")
print("Погода того дня:")
print(max_gust_weather)
#Зміна швидкості поривів вітру в Мадриді з часом
plt.figure(figsize=(12, 6))
plt.plot(df['CET'],  df['  MaxGustSpeedKm/h'],  marker='o',  linestyle='-',
color='b', label='WindGustSpeed')
```

```
plt.scatter(max_gust_date,    df.loc[df['CET']    ==    max_gust_date,    '
MaxGustSpeedKm/h'], color='red', label='MaxGustDate')
plt.title('Зміна швидкості поривів вітру в Мадриді з часом')
plt.xlabel('Дата')
plt.ylabel(' Максимальний порив вітру (Km/h)')
plt.xticks(rotation=45)
plt.legend()
plt.grid(True)
plt.show()
df[' Events'].fillna('NoEvent', inplace=True)
df[' Events'].fillna('NoEvent', inplace=True)
grouped_data = df.groupby(' Events')[' MeanVisibilityKm'].mean()
grouped_data
clear_visibility = grouped_data.get('Clear', 0)  # Use 0 asdefaultif 'Clear'
isnotpresent
clear_visibility
fog_visibility  =  grouped_data.get('Fog',  0)   #  Use  0  asdefaultif  'Fog'
isnotpresent
fog_visibility
plt.bar(['Clear',      'Fog'],      [clear_visibility,      fog_visibility],
color=['skyblue', 'gray'])
plt.title('Порівняння середньої видимості в ясні та туманні дні в Мадриді')
plt.xlabel('Погодні події')
plt.ylabel('Середній рівень видимості (Km)')
plt.show()
```

Додаток Б

Програмне рішення прогнозування клімату за допомогою алгоритму Random Forest

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import warnings

warnings.filterwarnings("ignore")


df= pd.read_csv("weatherAUS.csv")

df.info()

print("Total no. of entries: ", df.shape[0])

print("Total no. of features: ", df.shape[1])

# This will display all columns

pd.set_option('display.max_columns', None)


#This will display 5 rows

df.head()

df.describe().T

#Checking the cardinality of features

df.nunique()

no = df.isnull().sum()

per= df.isnull().sum()/len(df)*100

missing_values = pd.DataFrame({"total no.":no,
"percentage":per}).sort_values(ascending=False, by='total no.')
```

```
missing_values

df.duplicated().sum()


#  Data Pre-Preprocessing

cols_1= [var for var in df.columns if df[var].isnull().mean()*100 >5 and
df[var].isnull().mean()*100 <10]

cols_1

num_cols = ["WindGustSpeed", "Pressure9am" ,"Pressure3pm"]

cat_cols = ["WindDir9am"]

plt.figure(figsize=(9,4))

for i, cols in enumerate(num_cols):

plt.subplot(1, 3, i+1)

sns.boxplot(data=df, x=cols)

plt.title(f"{cols}")

plt.tight_layout()

plt.show()

df["WindGustSpeed"]=df["WindGustSpeed"].fillna(df["WindGustSpeed"].median())

df["Pressure9am"]=df["Pressure9am"].fillna(df["Pressure9am"].median())

df["Pressure3pm"]=df["Pressure3pm"].fillna(df["Pressure3pm"].median())

df["WindDir9am"]=df["WindDir9am"].fillna(df["WindDir9am"].mode()[0])


#Expolatory Data Analysis

plt.figure(figsize=(10,4))

plt.subplot(1, 2, 1)

ax= sns.countplot(data=df, x="RainTomorrow", palette="rocket")

for bars in ax.containers:
```

```
ax.bar_label(bars)
plt.title("Target feature disturbution",fontweight="black", size=13)

plt.subplot(1, 2, 2)
#create pie chart
df['RainTomorrow'].value_counts().plot(kind='pie',autopct='%0.1f%%')

plt.tight_layout()
plt.show()
cont_cols= df.select_dtypes(include="number").columns.to_list()
cont_cols.remove("Cloud3pm_imputed")
cont_cols.remove("Cloud9am_imputed")
def cont_dist(data, feat):
plt.figure(figsize=(13,6))
    for i, cols in enumerate(feat):
plt.subplot(2, 4, i+1)
sns.histplot(data=data, x=cols ,kde=True, palette="rocket",bins=30)
plt.title(f"Disturbution of {cols}", fontweight="black", size=9)
plt.tight_layout()
plt.show()
cont_dist(df, cont_cols[0:8])
plt.figure(figsize=(19,8))
for i, cols in enumerate(cat_cols):
plt.subplot(2, 3, i+1)
sns.countplot(data=df, x=cols)
plt.title(f"{cols}")
```

```
   # Set the x-axis ticks at a 90-degree angle
plt.xticks(rotation=90)


plt.tight_layout()
plt.show()
# Multivariate Analysis
plt.figure(figsize=(10,8))
corr= df[cont_cols].corr()
sns.heatmap(corr, annot=True, cmap='viridis', linewidths=0.1)
plt.show()
```