ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АНАЛІЗУ ТА ПЕРЕДБАЧЕННЯ ОПАДІВ

Керівник МКР:
к.т.н., доцент Козачко О. М.
Розробила:
студентка гр. 2ІСТ-19м Мельник О. Л.

Вступ

Сьогодні важко уявити світ, в якому ми не вміємо передбачати погодні умови. Кожен другий, а то і перший, щодня перевіряє погоду перед виходом з дому чи поїздкою у відпустку за межі міста, країни, материка. Можливість дізнатися погоду на декілька днів вперед стала повсякденною справою, проте так було не завжди. За часів Арістотеля люди вже почали підкреслювати певні закономірності у змінах погодних умов та спостерігати за ними систематично, щоправда тоді вони грунтувались лише на рівні прикмет.

Основним недоліком моніторингу в синопточних методах є непостійність у часі та просторі, оскільки моніторинг повітря здійснюється під час передачі даних між різними станціями. Ще однією важливою особливістю синоптичної карти є станції, з яких можна використовувати дані погодних умов, адже вони знаходяться на відстані не менше 100-150 км одна від одної. Саме тому можна підставити під сумнів якість збору та обробки даних, оскільки карти погоди збираються напротягу 6 годих після заміру показників та карти топографії протягом доби, що призводить до ігнорування можливих змін отриманих значень параметрів та впливу даних змін на передбачення погоди вцілому.

Отже, розробка інформаційної технології аналізу та передбачення опадів, яка буде забезпечувати швидкий комплексний аналіз даних, буде доцільною з точки зору пошуку оптимального методу прогнозування.

Мета роботи – підвищення точності прогнозування наявності опадів за рахунок використання інформаційних технологій, машинного навчання та аналітичної обробки даних.

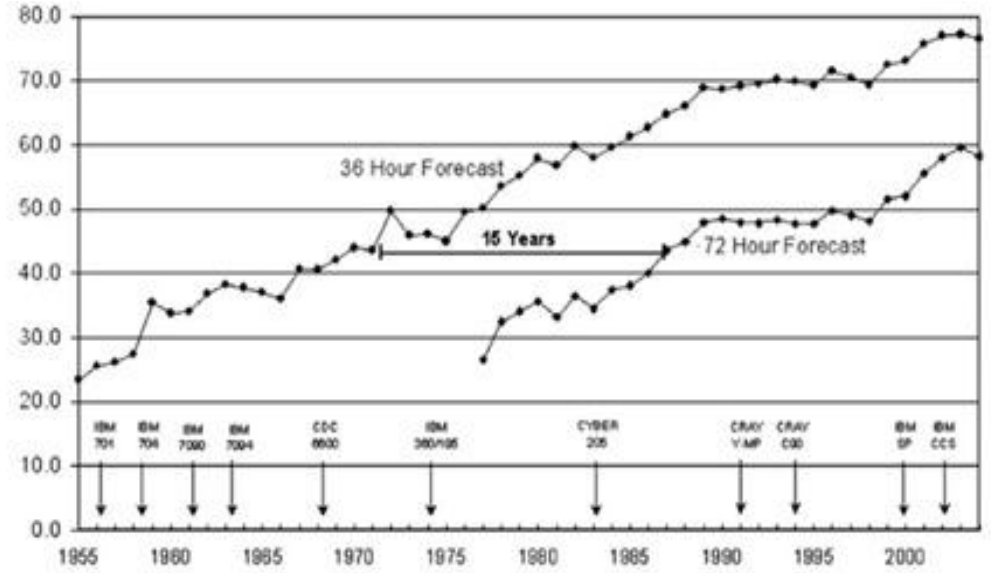
Об’єкт досліджень – процес передбачення опадів на основі аналізу попередніх метеоданих.

Предмет дослідження – інформаційна технологія аналізу і передбачення опадів на реальних даних.

Основні задачі дослідження:

- ❑ Проаналізувати предметну область дослідження;
- ❑ Обгрунтувати доцільность створення інформаційної технології аналізу та передбачення опадів;
- ❑ Використати інструменти машинного навчання для проектування і розробки інформаційної технології;
- ❑ Протестувати розроблений програмний продукт і проаналізувати отримані дані.

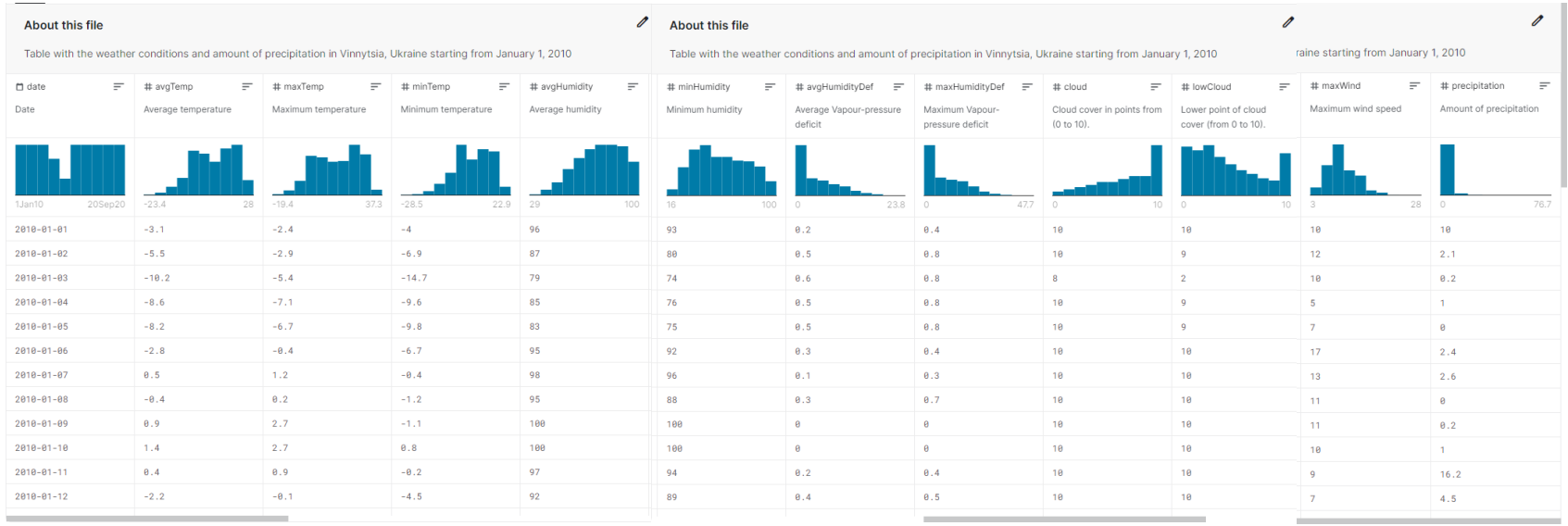
Аналіз предметної області



Skill of the 36 hour (1955–2004) and 72 hour (1977–2004) 500 hPa forecasts produced at NCEP. Forecast skill is expressed as a percentage of an essentially perfect forecast score. Thanks to Bruce Webster of NCEP for the graphic of S1 scores.

Графік відображення прогресу точності прогнозування погодних умов

Огляд даних



About this file

Table with the weather conditions and amount of precipitation in Vinnytsia, Ukraine starting from January 1, 2010

date	avgTemp	maxTemp	minTemp	avgHumidity	minHumidity	avgHumidityDef	maxHumidityDef	cloud	lowCloud	maxWind	precipitation
Date	Average temperature	Maximum temperature	Minimum temperature	Average humidity	Minimum humidity	Average Vapour-pressure deficit	Maximum Vapour-pressure deficit	Cloud cover in points from (0 to 10).	Lower point of cloud cover (from 0 to 10).	Maximum wind speed	Amount of precipitation
1Jan10 – 20Sep20	-23.4 – 28	-19.4 – 37.3	-28.5 – 22.9	29 – 100	16 – 100	0 – 23.8	0 – 47.7	0 – 10	0 – 10	3 – 28	0 – 76.7
2010-01-01	-3.1	-2.4	-4	96	93	0.2	0.4	10	10	10	10
2010-01-02	-5.5	-2.9	-6.9	87	80	0.5	0.8	10	9	12	2.1
2010-01-03	-10.2	-5.4	-14.7	79	74	0.6	0.8	8	2	10	0.2
2010-01-04	-8.6	-7.1	-9.6	85	76	0.5	0.8	10	9	5	1
2010-01-05	-8.2	-6.7	-9.8	83	75	0.5	0.8	10	9	7	0
2010-01-06	-2.8	-0.4	-6.7	95	92	0.3	0.4	10	10	17	2.4
2010-01-07	0.5	1.2	-0.4	98	96	0.1	0.3	10	10	13	2.6
2010-01-08	-0.4	0.2	-1.2	95	88	0.3	0.7	10	10	11	0
2010-01-09	0.9	2.7	-1.1	100	100	0	0	10	10	11	0.2
2010-01-10	1.4	2.7	0.8	100	100	0	0	10	10	10	1
2010-01-11	0.4	0.9	-0.2	97	94	0.2	0.4	10	10	9	16.2
2010-01-12	-2.2	-0.1	-4.5	92	89	0.4	0.5	10	10	7	4.5

Приклад вхідних даних з датасету, сформованого на основі даних Вінницького обласного центру з гідрометеорології, завантаженого на платформу для машинного навчання Kaggle від компанії Google

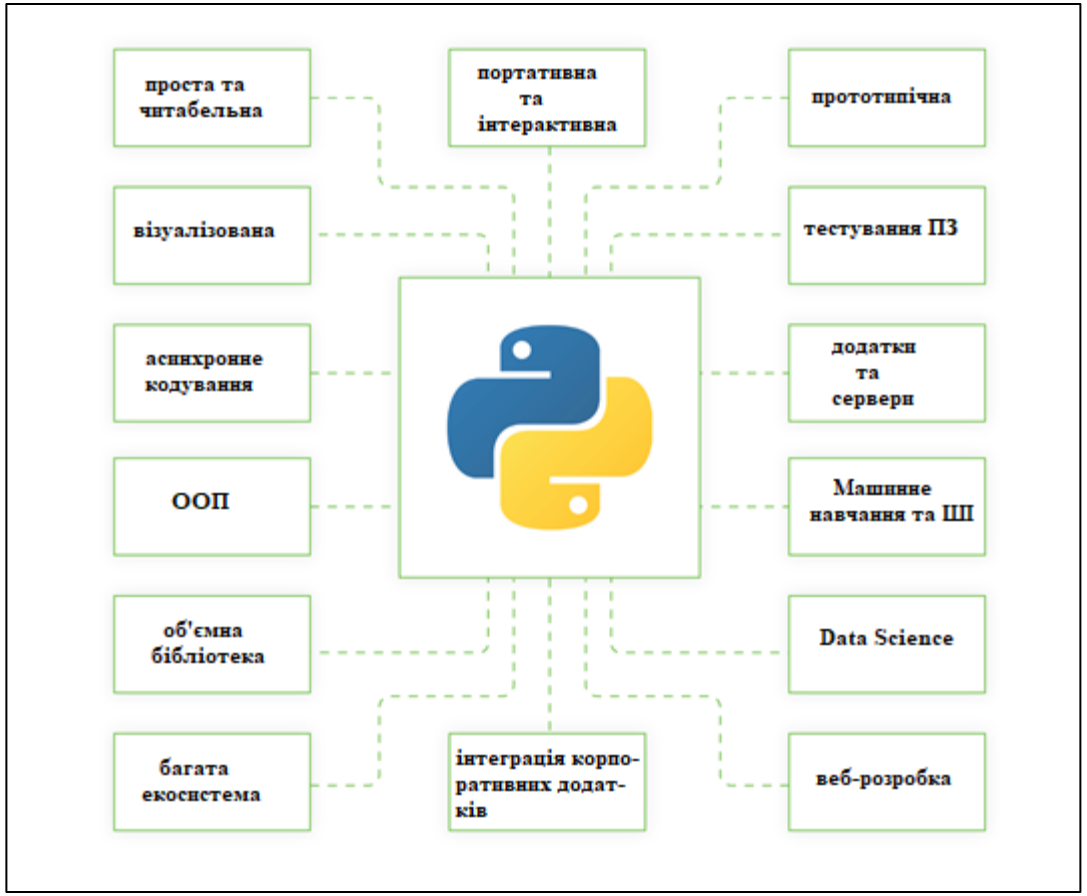
Огляд даних

	avgTemp	maxTemp	minTemp	avgHumidity	minHumidity	avgHumidityDef	maxHumidityDef	cloud	lowCloud	maxWir
0	-3.1	-2.4	-4.0	96.0	93	0.2	0.4	10.0	10.0	10
1	-5.5	-2.9	-6.9	87.0	80	0.5	0.8	10.0	9.0	12
2	-10.2	-5.4	-14.7	79.0	74	0.6	0.8	8.0	2.0	10
3	-8.6	-7.1	-9.6	85.0	76	0.5	0.8	10.0	9.0	5
4	-8.2	-6.7	-9.8	83.0	75	0.5	0.8	10.0	9.0	7

Приклад тренувальних даних моделі

Python

Найбільш вживаною мовою програмування серед розробників Data science залишається Python, популярність якій переважно додають великий набір інструментів та методів реалізації ML, а також масштабована база бібліотек, заточених під даний напрям. Однією з кращих бібліотек даної мови програмування стала Sk.Learn, в якій зосереджено чималу кількість алгоритмів.



NumPy, Sklearn, Pandas

NumPy – одна з найбільш застосовуваних бібліотек Python для роботи з великим обсягом даних (матриці або масиви), що працює на основі великого набору математичних функцій.

Sklearn – бібліотека створена на основі двох інших: NumPy та SciPy. Зручна у використанні та зрозуміла бібліотека, що допоможе перетворити дані або вибрати функцію всього за допомогою кількох рядків коду.

Pandas є однією з кращих високорівневих бібліотек для роботи зі структурою даних, дозволяє згрупувати, відфільтрувати, скомбінувати дані таким чином, щоб вони були максимально зрозумілими для сприйняття, а відповідно використання та аналізу.

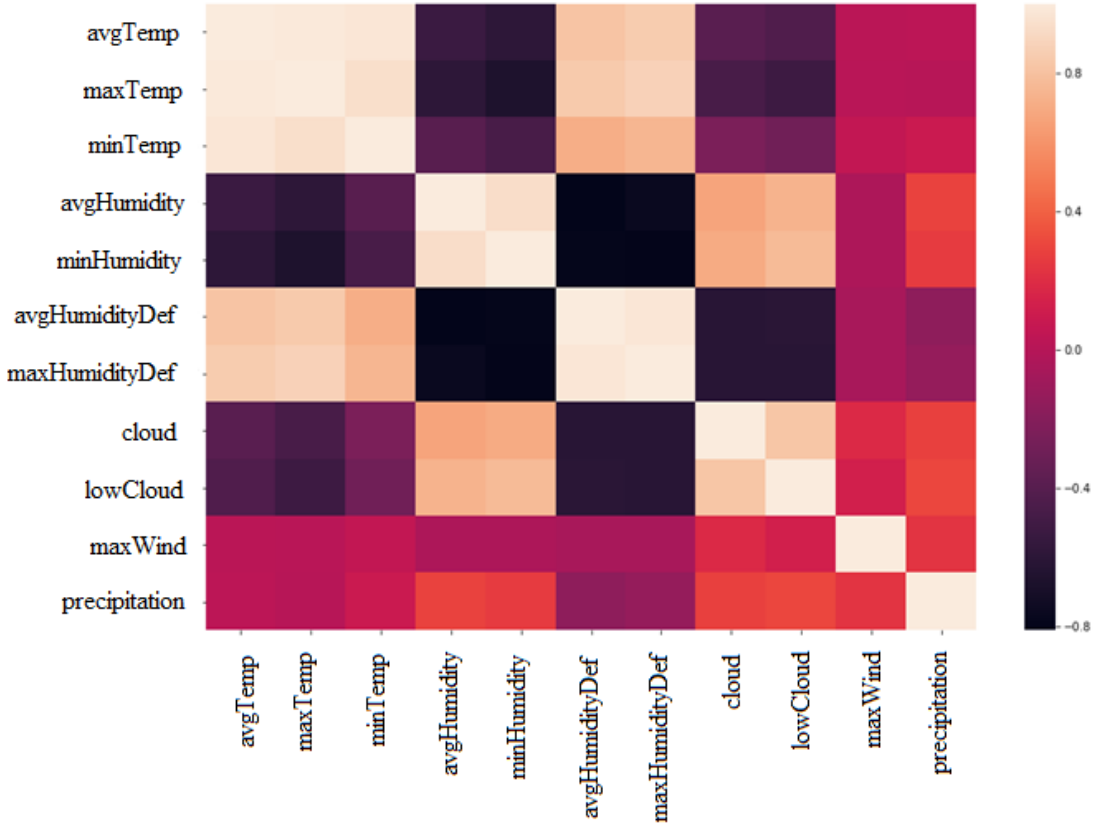
LightGBM, Matplotlib та Seaborn

LightGBM – фреймворк для реалізації градієнтного бустінга, в основі якого лежить алгоритм дерева рішень. Головним принципом роботи є розподіл листя дерева по найкращій відповідності.

Matplotlib дає широкий вибір у типі графіків, за її допомогою можна побудувати гістограму, різноманітні діаграми та реалізувати графік в площині недекартових координат. Підтримує графічний інтерфейс користувача всіх операційних систем.

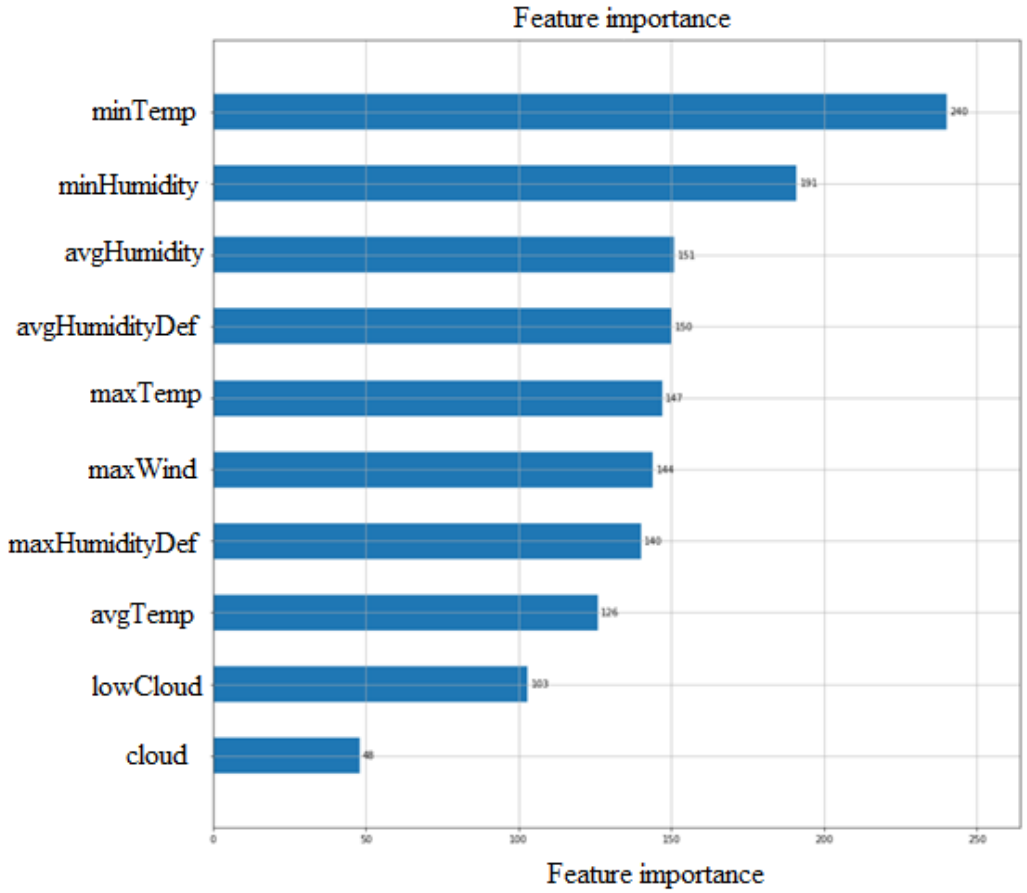
Seaborn – це високорівнева бібліотека візуалізації графіків на базі іншої бібліотеки Matplotlib. Seaborn дозволяє скоротити кількість коду для реалізації більш складних типів візуалізації.

Візуалізація обрахованої попарної кореляції стовпців



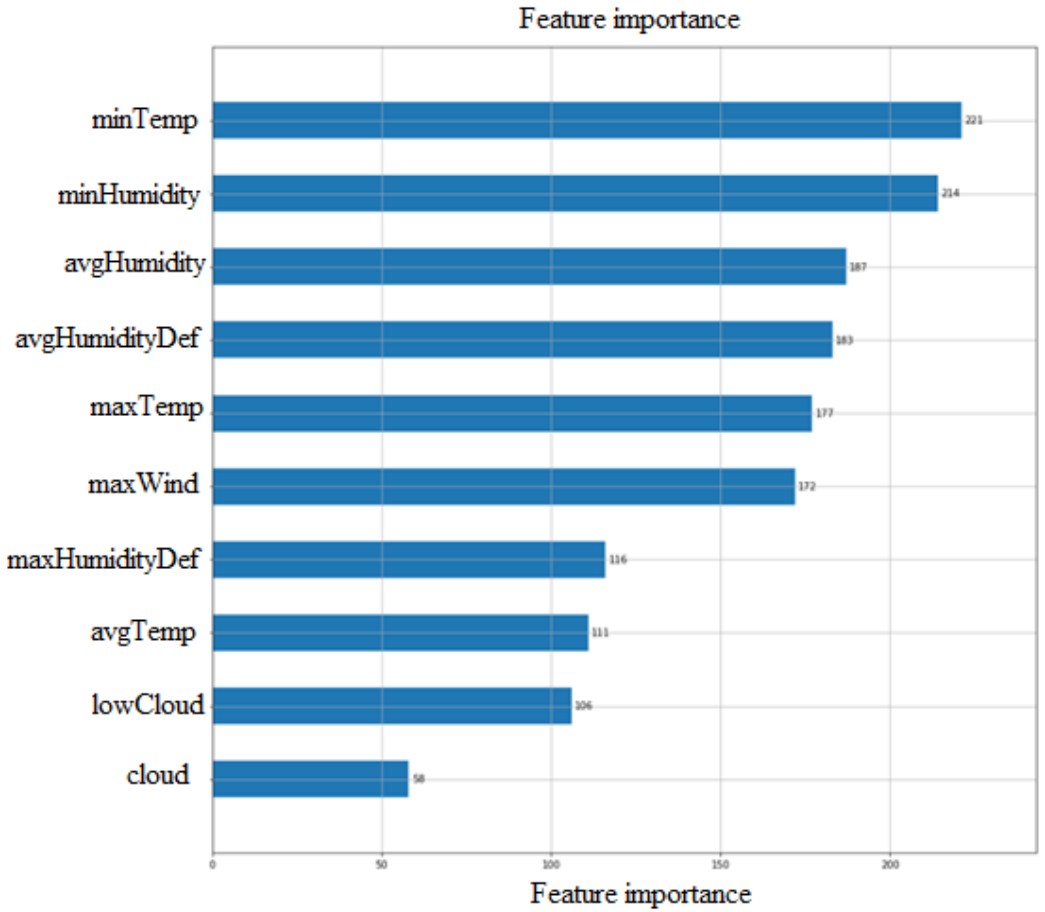
Параметр	Значення
avgTemp	середня температура повітря (°C)
maxTemp	максимальна температура повітря(°C)
minTemp	мінімальна температура повітря (°C)
avgHumidity	середня вологість повітря
minHumidity	мінімальна вологість повітря
avgHumidityDef	середній дефіцит вологості повітря
maxHumidityDef	максимальний дефіцит вологості повітря
cloud	оцінка загальної хмарності (0 – 10)
lowCloud	оцінка хмарності нижнього ярусу (0 – 10)
maxWind	максимальна швидкість вітру (м/с)
precipitation	кількість опадів за добу (мм)

Діаграма важливості атрибутів згідно результатів моделі lgbm



Параметр	Значення
avgTemp	середня температура повітря (°C)
maxTemp	максимальна температура повітря(°C)
minTemp	мінімальна температура повітря (°C)
avgHumidity	середня вологість повітря
minHumidity	мінімальна вологість повітря
avgHumidityDef	середній дефіцит вологості повітря
maxHumidityDef	максимальний дефіцит вологості повітря
cloud	оцінка загальної хмарності (0 – 10)
lowCloud	оцінка хмарності нижнього ярусу (0 – 10)
maxWind	максимальна швидкість вітру (м/с)

Діаграма важливості атрибутів згідно результатів моделі xgb



Параметр	Значення
avgTemp	середня температура повітря (°C)
maxTemp	максимальна температура повітря(°C)
minTemp	мінімальна температура повітря (°C)
avgHumidity	середня вологість повітря
minHumidity	мінімальна вологість повітря
avgHumidityDef	середній дефіцит вологості повітря
maxHumidityDef	максимальний дефіцит вологості повітря
cloud	оцінка загальної хмарності (0 – 10)
lowCloud	оцінка хмарності нижнього ярусу (0 – 10)
maxWind	максимальна швидкість вітру (м/с)

Значення коефіцієнтів важливості за допомогою LinearRegression

	feature	score_linreg
1	maxTemp	0.393227
8	lowCloud	0.268171
9	maxWind	0.244437
2	minTemp	0.074363
3	avgHumidity	0.041354
4	minHumidity	0.035451
6	maxHumidityDef	0.027409
7	cloud	0.005716
5	avgHumidityDef	-0.147902
0	avgTemp	-0.355524

Параметр	Значення
avgTemp	середня температура повітря (°C)
maxTemp	максимальна температура повітря(°C)
minTemp	мінімальна температура повітря (°C)
avgHumidity	середня вологість повітря
minHumidity	мінімальна вологість повітря
avgHumidityDef	середній дефіцит вологості повітря
maxHumidityDef	максимальний дефіцит вологості повітря
cloud	оцінка загальної хмарності (0 – 10)
lowCloud	оцінка хмарності нижнього ярусу (0 – 10)
maxWind	максимальна швидкість вітру (м/с)

Значення коефіцієнтів важливості за допомогою LogisticRegression

	feature	score_logreg
0	avgTemp	0.105288
1	maxTemp	0.939418
2	minTemp	1.203480
3	avgHumidity	3.807568
4	minHumidity	0.653257
5	avgHumidityDef	2.174296
6	maxHumidityDef	1.381805
7	cloud	2.618418
8	lowCloud	1.314107
9	maxWind	3.915504

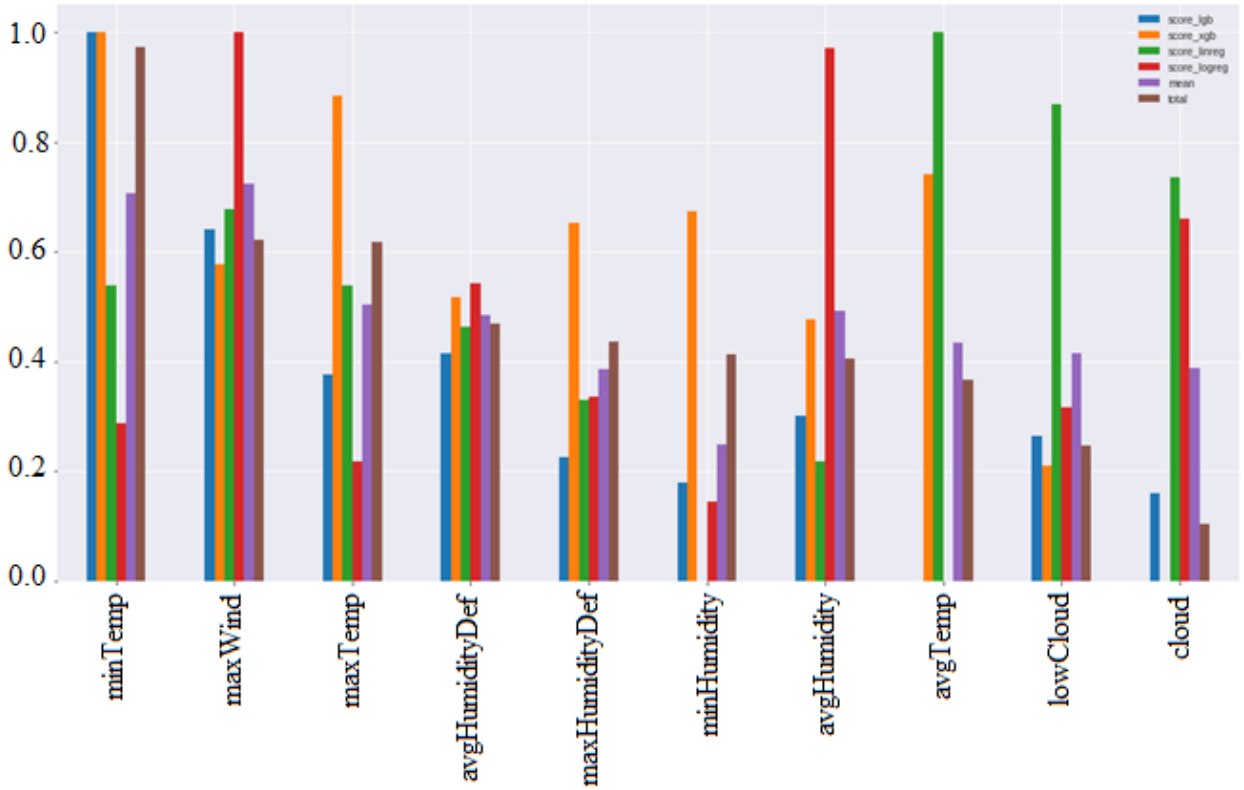
Параметр	Значення
avgTemp	середня температура повітря (°C)
maxTemp	максимальна температура повітря(°C)
minTemp	мінімальна температура повітря (°C)
avgHumidity	середня вологість повітря
minHumidity	мінімальна вологість повітря
avgHumidityDef	середній дефіцит вологості повітря
maxHumidityDef	максимальний дефіцит вологості повітря
cloud	оцінка загальної хмарності (0 – 10)
lowCloud	оцінка хмарності нижнього ярусу (0 – 10)
maxWind	максимальна швидкість вітру (м/с)

Результуючі значення коефіцієнтів важливості атрибутів

feature	score_lgb	score_xgb	score_linreg	score_logreg	mean	total
minTemp	1.000000	1.000000	0.538865	0.288223	0.706772	0.974035
maxWind	0.641509	0.578231	0.676980	1.000000	0.724180	0.622245
maxTemp	0.377358	0.884354	0.538450	0.218919	0.504770	0.617574
avgHumidityDef	0.415094	0.517007	0.463164	0.543016	0.484570	0.468331
maxHumidityDef	0.226415	0.653061	0.330047	0.335025	0.386137	0.435500
minHumidity	0.179245	0.673469	0.000000	0.143816	0.249133	0.413618
avgHumidity	0.301887	0.476190	0.219066	0.971672	0.492204	0.404818
avgTemp	0.000000	0.741497	1.000000	0.000000	0.435374	0.365918
lowCloud	0.264151	0.210884	0.868651	0.317257	0.415236	0.246221
cloud	0.160377	0.000000	0.735391	0.659577	0.388836	0.104122

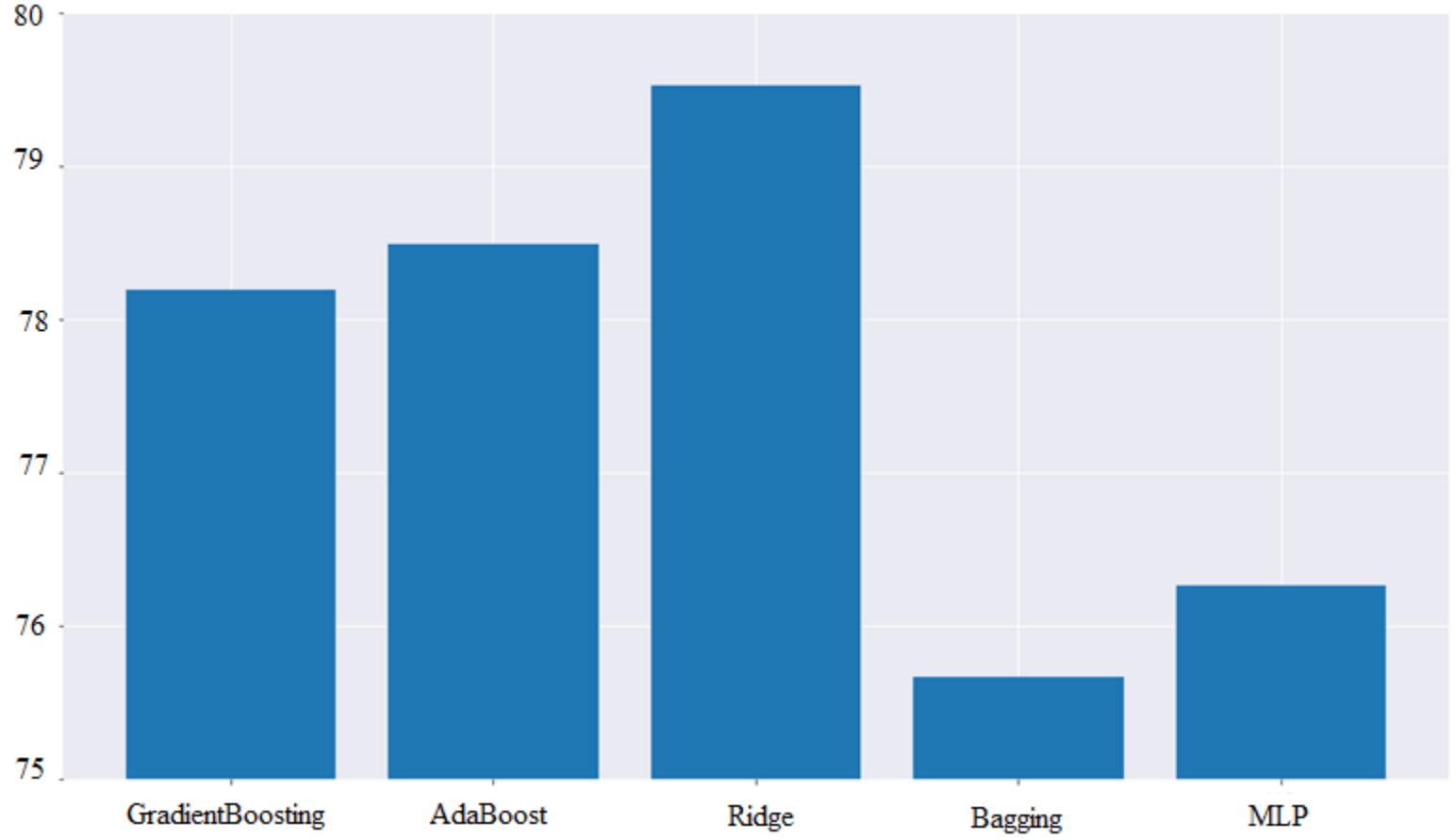
Параметр	Значення
avgTemp	середня температура повітря (°C)
maxTemp	максимальна температура повітря(°C)
minTemp	мінімальна температура повітря (°C)
avgHumidity	середня вологість повітря
minHumidity	мінімальна вологість повітря
avgHumidityDef	середній дефіцит вологості повітря
maxHumidityDef	максимальний дефіцит вологості повітря
cloud	оцінка загальної хмарності (0 – 10)
lowCloud	оцінка хмарності нижнього ярусу (0 – 10)
maxWind	максимальна швидкість вітру (м/с)

Діаграма значень коефіцієнтів важливості атрибутів

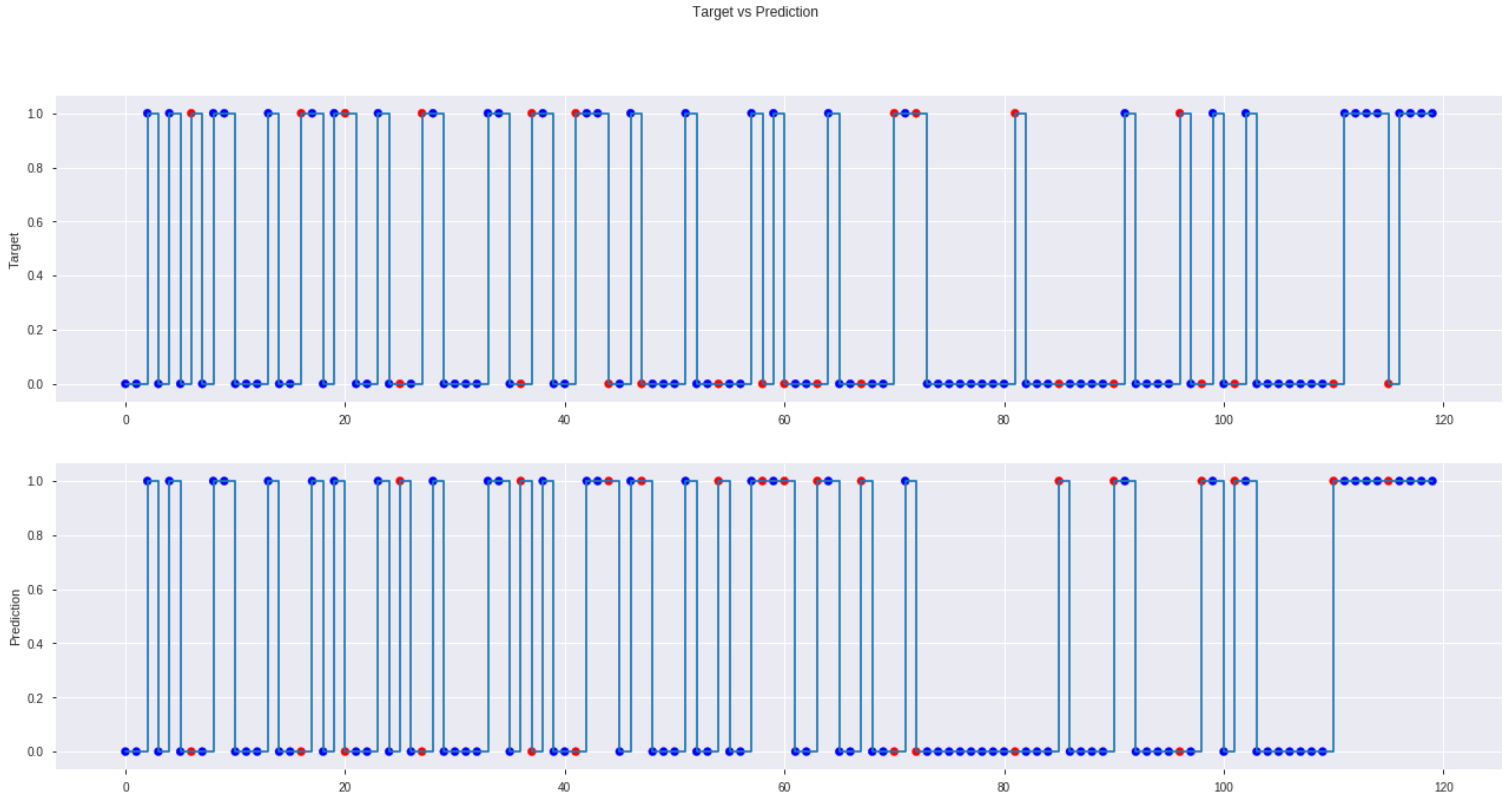


Параметр	Значення
avgTemp	середня температура повітря (°C)
maxTemp	максимальна температура повітря(°C)
minTemp	мінімальна температура повітря (°C)
avgHumidity	середня вологість повітря
minHumidity	мінімальна вологість повітря
avgHumidityDef	середній дефіцит вологості повітря
maxHumidityDef	максимальний дефіцит вологості повітря
cloud	оцінка загальної хмарності (0 – 10)
lowCloud	оцінка хмарності нижнього ярусу (0 – 10)
maxWind	максимальна швидкість вітру (м/с)

Порівняння результатів перевірки точності застосованих класифікаторів



Результат роботи програми



Графік попарного порівняння результатів прогнозування наявності опадів з реальними даними

Висновки

Було проведено аналіз проблеми точності та ресурсозатратності відділів Українського гідрометцентру, зокрема підкреслено, що далеко не кожен відділ забезпечений сучасним та потужним устаткуванням для підвищення точності передбачень та швидкості обрахунку даного передбачення. Сформовано датасет для навчання моделі на основі реальних даних, отриманих від Вінницького обласного центру з гідрометеорології. Відштовхуючись від даних, було проведено аналіз щодо вибору технологій та інструментів для оптимальної реалізації технології передбачення. Розроблено модель прогнозування наявності опадів, спираючись на визначені гіперпараметри, тобто атрибути з найбільшою вагою на вихідні дані.

В ході тестування розробленої технології аналізу та передбачення було продемонстровано точність роботи моделі.

Наукова новизна

Подальшого розвитку набув метод передбачення опадів Вінницького регіону, який на відміну від існуючих, визначає інформативні ознаки впливу, на основі яких здійснюється прогнозування наявності опадів за рахунок використання алгоритмів машинного навчання.

Практичне значення

Практичне значення одержаних результатів можна охарактеризувати наступними пунктами:

- розроблено просту для розуміння та зручну у використанні модель передбачення;
- модель не потребує вибірки даних зі складною структурою, використовується датасет простого формату.

За результатами магістерської кваліфікаційної роботи опубліковано: тези на XV Міжнародній конференції "Контроль і управління в складних системах" (КУСС-2020).

Дякую за увагу!