Математичне моделювання та методи оптимізації. Лабораторна 1: Регресійні моделі

Михайло Голуб

27 вересня 2024 р.

Завдання лабораторної роботи:

1. Завантажити дані з джерела https://www.saveecobot.com/maps Бажано обрати такі пости моніторингу, щоб був достатньо великий часовий ряд даних (хоча б за попередній рік), а також щоб були різні типи забруднюючих речовин.

2. Проаналізувати завантажений датасет, провести підготовку для роботу із ним (форматування, видалення пустих значень тощо).

3. Знайти можливі залежності між забруднювачами повітря (чи залежить забруднювач PM2.5 від чадного газу і тд). Зробити це з використанням регресійного аналізу. Отримати показники залежностей (або показати, що такі залежності відсутні).

4. Отримати регресійну модель залежності, використавши частину набору даних на навчання, іншу частину – на тестування моделі: залежність забрудника від часу дня (зранку повітря брудніше ніж вночі – це припущення. Обгрунтувати або спростувати його); залежність одного забрудника від іншого.

5. Отримати чисельні оцінки (RMSE, R^2) отриманої моделі.

6. Описати отримані результати та виокремити отримані висновки та припущення.

Хід роботи:

Підготовка даних:

Обрано датасет з пристрою з ідентифікатором 16181, що розташований на вулиці Солом'янській. Датасет містить наступні колонки: *device_id, phenomenon, value, logged_at, value_text.*

Даний датасет завантажується командою *pandas.read_csv* модуля роботи з даними Pandas. Оскількі деякі рядки мають кількість значень відмінну від більшості, їх потрібно пропустити: встановлено значення аргумента *on_bad_lines* в *warn.*

Щоб пересвідчитись що дані завантажено коректно в консоль виводяться перші 5 рядків таблиці та висота таблиці.

Вхідна таблиця містить колокнки *device_id ma value_text* які не містять корисних значень, тож їх можна видалити.

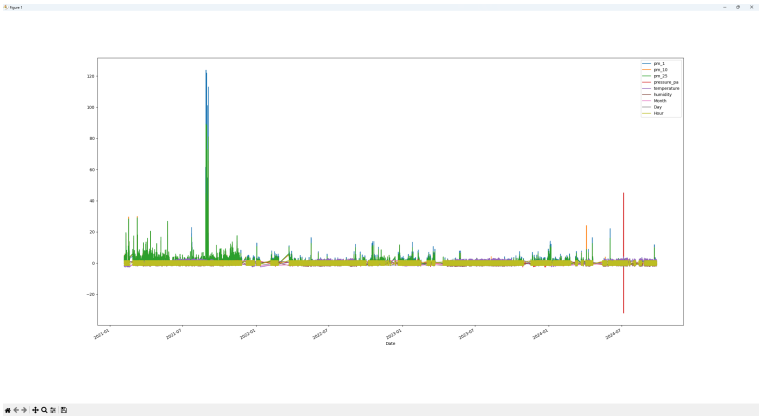
В отриманій таблиці містяться трійки значень *logged_at, phenomenon, value.* Таке представлення є не зручним, набагато легшим у використанні було б представлення *logged_at, pm_1, pm_10, ...* Для цього таблиця розбивається на декілька таблиць в яких міститься лише один *phenomenon*, в яких дані представлені парами *logged_at, value.* Отримані таблиці з'єднуються в одну таблицю з колонками *logged_at, pm_1, pm_10, ...*

До цього моменту *logged_at* було представлене текстовим рядком. Для коректної роботи функцій та методів цей рядок парситься в формат *DateTime* з яким вміє працювати більшість модулів. Також, для подальшої побудови регресійних моделей додано колонки *Month, Date* та *Hour.*

Останнім кроком підготовки даних до роботи з ними – видалення рядків з пустими клітинками.

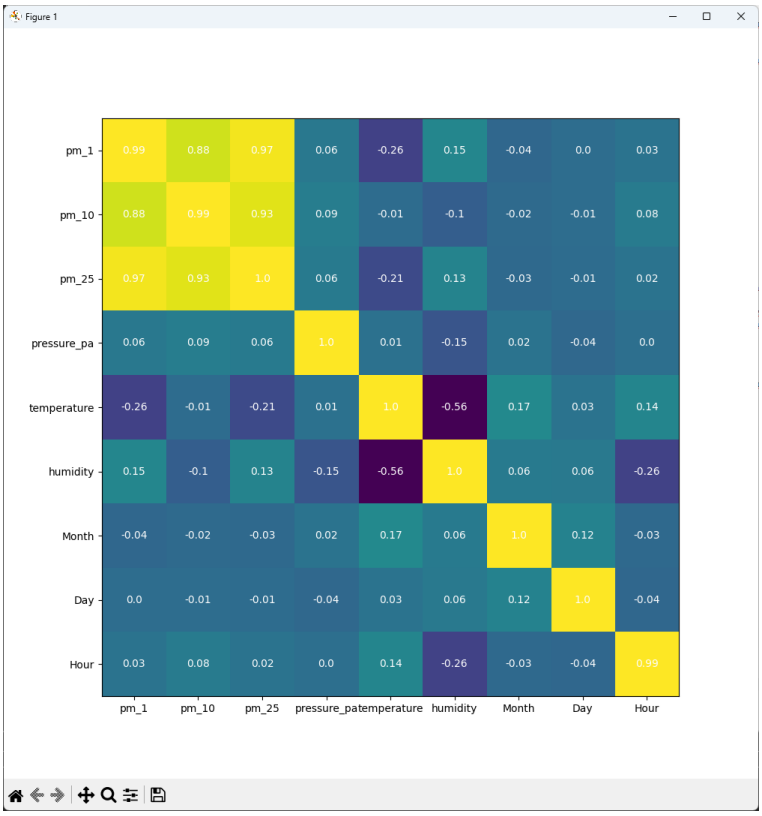
Аналіз даних:

Для розуміння чистоти даних, побудовано графік усіх колонок з нормалізованими данними.



Як видно з графіків, існують декілька піків та прогалин в значеннях. Але вони знехтовно малі порівняно з загальною кількістю рядків таблиці.

Для аналізу залежності різних показників побудовано теплову карту коефіцієнту Пірсона.

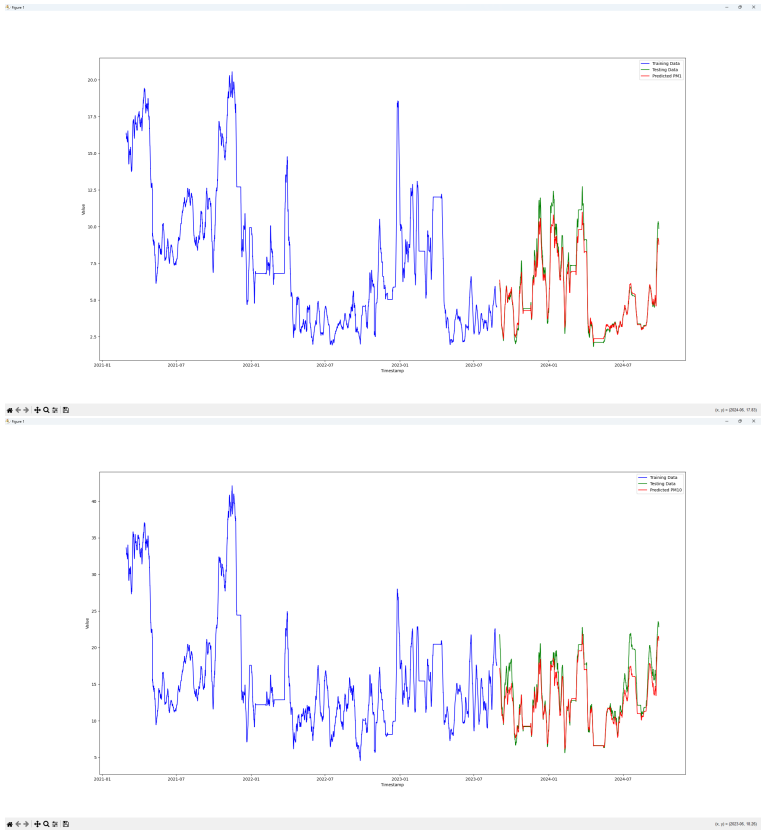


З теплової карти видно, що усі три показники забруднювачів PM значно корелюють. Також, вологість корелює з температурою та годинами; температура корелює з PM1 та PM2.5. Усі інші пари колонок слабо корелюють.

Таким чином, залежність забрудників PM від часу доби спростовано.

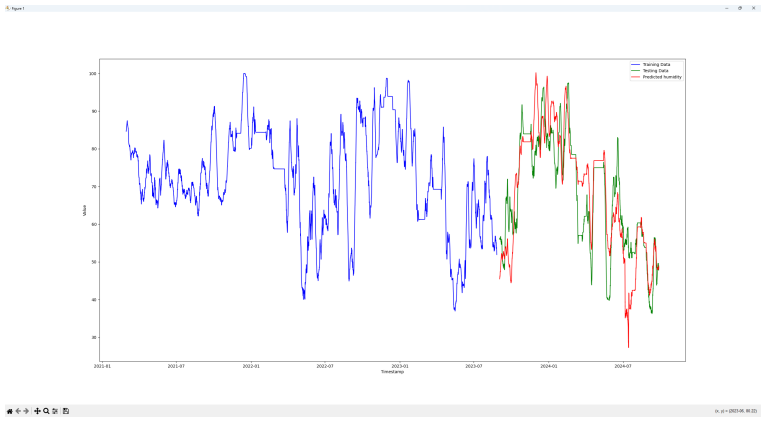
Побудова моделей:

Кожен датчик PM має свою ціну, використання лише одного датчика PM2.5 та моделювання інших значень зменшило б вартість пристроїв. Модель яка обраховує PM1 та PM10 залежно від значень PM2.5, температури та вологи:



Модель передбачила PM1 досить точно, а от PM10 має значні відхилення на деяких піках.

Датчики вологості споживають значну кількість енергії. Під час відключень передбачення вологості за рахунок регресійної моделі могло б продовжити час автономної роботи пристрою. Модель яка обраховує вологість залежно від значень температури, тиску, часу доби та забрудників:



Дана модель має низьку точність в найгіршому випадку та точність нижче допустимої в середньому.

Висновки:

- Показники PM значно корелюють між собою: мають коеф. Пірсона 88-97%;
- Температура та волігість корелюють між собою: мають коеф. Пірсона 56%;
- Більшість показників мають коеф. Пірсона менше 30
- Забрудненість PM маже не корелює з часом доби;
- Лінійна регресійна модель може досить точно передбачувати PM1 за значеннями PM2.5, температури та вологи;
- Вологість можна неточно передбачити за усіма іншими показниками.