

# Математичне моделювання та методи оптимізації. Лабораторна 1: Регресійні моделі

Михайло Голуб

28 вересня 2024 р.

### **Завдання лабораторної роботи:**

1. Завантажити дані з джерела <https://www.saveecobot.com/maps> Бажаючи 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 роботи із ним (форматування, видалення пустих значень тощо).
3. Знайти можливі залежності між забруднювачами повітря (чи залежить забруднювач  $PM_{2.5}$  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. Обґрунтувати 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. Оскільки 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 колонки *device\_id* та *value\_text* які не містять корисних значень, тож їх можна видалити.

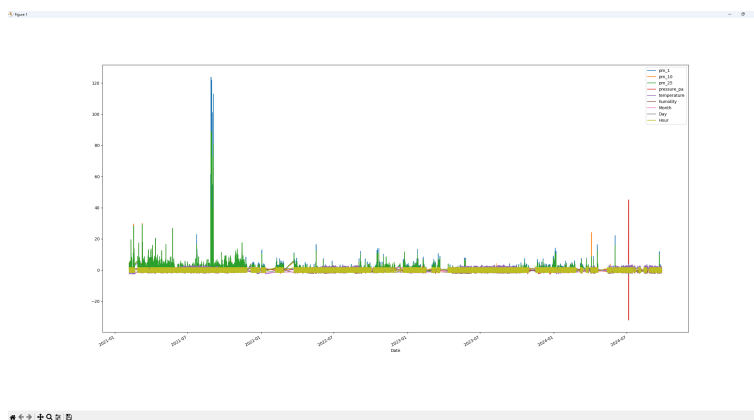
В отриманій таблиці містяться трійки значень *logged\_at*, *phenomenon*, *value*. Таке представлення є не зручним, набагато легшим у використанні було б представлення *logged\_at*, *pm\_1*, *pm\_10*, ... Для цього таблиця розбивається на декілька таблиць в яких міститься лише один *phenomenon*, та в 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 представлено 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 даними.



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

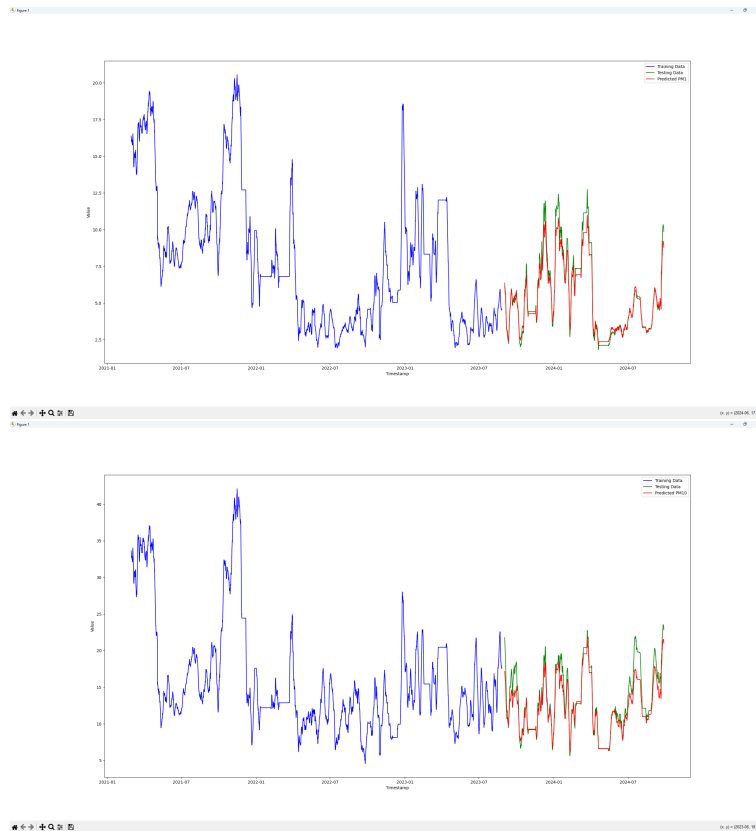
Для аналізу залежності різних показників побудовано теплову карту коефіцієнту Пірсона.



З теплової карти видно, що усі три показники забруднювачів РМ значно корелюють. Також, вологість корелює з температурою та годинами; температура корелює з РМ1 та РМ2.5. Усі інші пари колонок слабо корелюють. Таким чином, залежність забрудників РМ від часу доби спростовано.

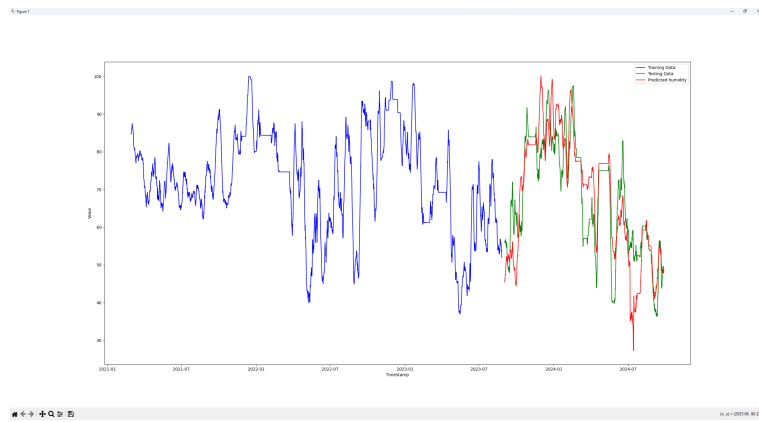
### Побудова моделей:

Кожен датчик РМ має свою ціну, використання лише одного датчика РМ2.5 та моделювання інших значень зменшило б вартість пристроїв. Модель яка обраховує РМ1 та РМ10 залежно від значень РМ2.5, температури та вологості:



Модель передбачила РМ1 досить точно: коефіцієнт детермінації 0.94 та RMSE 1.5. Модель що передбачує РМ10 має нижчу точність: коеф. детермінації 0.88 та RMSE 8.6.

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 коефіцієнт детермінації 0.23 та RMSE 446. Це погані показники, така модель є неточною.

### **Висновки:**

- Показники РМ значно корелюють між собою: мають коеф. Пірсона 88-97%;
- Температура та вологість корелюють між собою: мають коеф. Пірсона 56%;
- Більшість показників мають коеф. Пірсона менше 30
- Забрудненість РМ маже не корелює з часом доби;
- Лінійна регресійна модель може точно передбачувати РМ1 та РМ10 за значеннями РМ2.5, температури та вологи;
- Вологість можна неточно передбачити за усіма іншими показниками.