Міністерство освіти і науки України

Національний технічний університет України

«Київський політехнічний інститут»

Фізико-технічний інститут

**ЛАБОРАТОРНА РОБОТА №1**

З предмету

«Математичне моделювання»

Виконав:

Студент ФІ-21 групи

Булавінцев Юрій

Київ 2024

**Мета роботи**: Отримати досвід використання основних методів та засобів аналізу даних у середовищі розробки IPython Notebook на реальних наборах даних.

***Завдання для лабораторної роботи №1***

1. Завантажити дані з одного або декількох постів з джерела [https://www.saveecobot.com/](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
   1. Залежність забрудника від часу дня (зранку повітря брудніше ніж вночі – це припущення. Обгрунтувати або спростувати його).
   2. Залежність одного забрудника від іншого
5. Отримати чисельні оцінки (RMSE, R^2) отриманої моделі.
6. Описати отримані результати та виокремити отримані висновки та припущення.

**Розвʼязок**

1. На початку роботи з датасетом я зробив аналіз на наявність порожніх комірок та побачив їх відсутність, проте в подальших діях я все ж скористався командою .dropna() для остаточного їх знищення, якщо такі є.
2. Код

import numpy as np

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import mean\_squared\_error, r2\_score

data = pd.read\_csv("LAB1 (MM)/saveecobot\_22643.csv")

#print(data.columns)

data\_wide = data.pivot\_table(index="logged\_at", columns="phenomenon", values="value").reset\_index()

data\_wide["logged\_at"]=pd.to\_datetime(data\_wide["logged\_at"])

data\_wide["hour"]=data\_wide["logged\_at"].dt.hour

#print(data\_wide.head(10))

data\_clean = data\_wide.dropna()

print("Отримання значень одного забрудника (рм2.5) від інших (no2\_ug, so2\_ug, co\_mg, no2\_ppb, so2\_ppb, co\_ppm)")

X = data\_clean[["no2\_ug", "so2\_ug", "co\_mg", "no2\_ppb", "so2\_ppb", "co\_ppm"]]

Y = data\_clean["pm25"]

X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X, Y, test\_size=0.2, random\_state=21)

model = LinearRegression()

model.fit(X\_train, Y\_train)

Y\_pred = model.predict(X\_test)

mse = mean\_squared\_error(Y\_test, Y\_pred)

r2 = r2\_score(Y\_test, Y\_pred)

print("MSE: ", mse)

print("R2: ", r2)

print("Залежність забрудника від часу дня (pm2.5)")

X\_time = data\_clean[["hour"]]

Y\_time = data\_clean["pm25"]

#print(X\_time)

X\_time\_train, X\_time\_test, Y\_time\_train, Y\_time\_test = train\_test\_split(X\_time, Y\_time, test\_size=0.2, random\_state=21)

model\_time = LinearRegression()

model\_time.fit(X\_time\_train, Y\_time\_train)

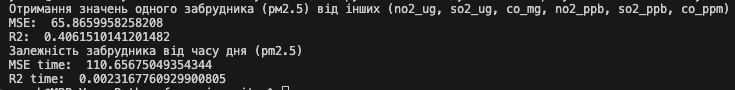
Y\_time\_pred = model\_time.predict((X\_time\_test))

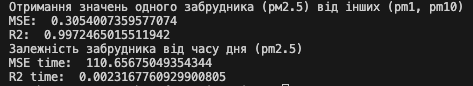
mse\_time = mean\_squared\_error(Y\_time\_test, Y\_time\_pred)

r2\_time = r2\_score(Y\_time\_test, Y\_time\_pred)

print("MSE time: ", mse\_time)

print("R2 time: ", r2\_time)

**Якщо намагатись знайти закономірність за наявністю небезпечних речовин в атмосфері таких як (no2\_ug, so2\_ug, co\_mg, no2\_ppb, so2\_ppb, co\_ppm)**

**Якщо намагатись знайти закономірність за наявністю небезпечних речовин в атмосфері таких як (pm1, pm10)**

MSE - показник того наскільки розрахункові дані відрізняються від фактичних.

R^2 - показник того, наскільки модель може пояснити те чи інше значення.

**Висновок:**

Як можна дійти висновку, то залежність «пилу» рм2.5 набагато міцніше повʼязаний з рм1 та рм10. Через що модель працює набагато краще. МСЕ прямує до свого найкращого значення 0.0, а R^2 до 1.0.

Крім того, варто зазначити, що залежності від часу дня до рм2.5 є дуже незначною, через що виникла велика різниця між тим, що було отримано приборами та тим, що змогла розрахувати модель, а пояснити може взагалі лише 0,23%.