# import libraries

import pandas as pd

from statsmodels.tsa.api import VAR

import matplotlib.pyplot as plt

from sklearn.metrics import mean\_squared\_error

import numpy as np

# clean numeric columns

columns\_to\_clean = ['C6H6(GT)', 'T', 'RH', 'AH']

# replace commas with dots and convert them to numeric

for col in columns\_to\_clean:

data[col] = data[col].str.replace(',', '.')

data[col] = pd.to\_numeric(data[col], errors='coerce')

# drop the unnamed columns and any rows with missing values

data\_cleaned = data.drop(columns=['Unnamed: 15', 'Unnamed: 16']).dropna()

# convert 'Date' and 'Time' to a single DateTime index

data\_cleaned['DateTime'] = pd.to\_datetime(data['Date'] + ' ' + data['Time'], format='%d/%m/%Y %H.%M.%S')

data\_cleaned = data\_cleaned.set\_index('DateTime')

# drop the 'Date' and 'Time' columns

data\_cleaned = data\_cleaned.drop(columns=['Date', 'Time'])

# ensure all columns are numeric

data\_cleaned = data\_cleaned.apply(pd.to\_numeric, errors='coerce')

# drop rows with NaN

data\_cleaned = data\_cleaned.dropna()

# split the data (80% train, 20% test)

train\_size = int(len(data\_cleaned) \* 0.8)

train\_data = data\_cleaned[:train\_size]

test\_data = data\_cleaned[train\_size:]