Air Quality prediction of Relative Humidity

Ravi Bhushan Pratap

210107070

Submission Date: April 25, 2024

Final Project submission

Course Name : Applications of Al and ML in chemical engineering

Course Code: CL653

Contents

1 Executive Summary

This project aims to develop a predictive model for air quality, focusing on forecasting relative humidity's impact. Air quality management is vital for public health, with relative humidity influencing pollutant dispersion. Leveraging machine learning, the project analyzes historical data to create an accurate predictive model. Key methodologies include data preprocessing, feature engineering, and model selection. Expected outcomes include a precise model offering insights for environmental management and public health. Overall, the project seeks to improve air quality prediction and contribute to healthier living environments globally.

2 Introduction

This project is all about predicting air quality, with a special focus on how relative humidity affects it. Relative humidity is really important because it can change how pollutants spread in the air and what the air is like. Our main goal is to make computer models that can predict air quality using data from the past and readings of relative humidity. Why? Because if we can predict air quality accurately, we can take action early to reduce pollution. This is super important for keeping people healthy and protecting the environment. As cities grow and more factories are built, managing air quality becomes even more important. By using AI and machine learning, we can make better predictions, which can help governments and city planners make smarter decisions. In the end, our aim is to make our air cleaner and our communities healthier by taking action before problems get worse.

Problem it Aims to Solve:

- Contribute to cleaner, healthier environments through informed interventions and proactive measures.
- Enhance air quality management by providing reliable predictions, aiding in policy decisions and environmental planning. Objectives: List the main objectives of the project.

Importance of Addressing the Issue:

Addressing air quality prediction, especially considering relative humidity, is vital for safeguarding public health, ensuring environmental sustainability, and mitigating the adverse impacts of pollution. Accurate forecasting enables timely interventions, crucial for minimizing respiratory illnesses, cardiovascular diseases, and ecological disturbances, particularly in urbanized and industrialized areas with escalating emissions.

Objectives:

- Prediction
- Understanding
- Mitigation
- Sustainability
- Advancement

Description:

Theoretical Background:

• **Air Quality Parameters:** Air quality is influenced by pollutants like PM, NO2, SO2, CO, and O3, posing risks to health and the environment. Predictive modeling entails understanding their relationships with environmental variables.

• **Relative Humidity:** Relative humidity (RH) influences air quality by affecting pollutant dispersion and concentration. High RH levels increase pollutant retention and reaction rates, worsening air quality conditions.

• **Machine Learning Techniques:** Machine learning algorithms like regression, decision trees, random forests, and neural networks extract patterns from historical data to predict future outcomes. In air quality prediction, they capture intricate relationships among pollutants, meteorological variables, and RH.

• **Data Collection and Preprocessing:** High-quality data is crucial for training accurate machine learning models. Air quality monitoring stations provide real-time measurements of pollutants, meteorological variables, and RH. Preprocessing steps involve data cleaning, missing value imputation, feature engineering, and normalization to ensure data quality and model performance.

• **Model Evaluation and Validation:** Evaluating model performance is essential to ensure reliable predictions. Common metrics include mean squared error (MSE), mean absolute error (MAE), and R-squared (R^2). Cross-validation techniques, such as k-fold cross-validation, help assess model generalization to unseen data.

Significance of issue:

• **Public Health Impact and Environmental Sustainability:** Poor air quality causes respiratory illnesses and environmental damage. Considering relative humidity in air quality prediction enables proactive health measures and ecological preservation strategies to minimize public health risks and ecosystem harm.

- **Economic Implications:** Air pollution imposes significant economic costs, including healthcare expenditures, loss of productivity due to illness, and damage to property and infrastructure. Accurate air quality prediction with consideration of relative humidity helps mitigate these costs by minimizing health-related expenses and safeguarding economic productivity.
- **Climate Change Mitigation:** Air quality and climate change are closely interconnected. Pollutants such as black carbon and methane contribute to global warming, while climate change influences atmospheric circulation patterns and pollutant dispersion. By improving air quality prediction models with RH considerations, synergistic approaches to both air quality management and climate change mitigation can be pursued.

3 Methodology

Data Source:

I have obtained my data from Kaggle, a popular platform for hosting datasets and machine learning competitions. I plan to access the data by downloading datasets directly from Kaggle's website and some below links.

https://www.nature.com/articles/s41598-022-13579-

https://link.springer.com/article/10.1007/s11356-023-26779-8

https://www.kaggle.com/code/sayakchakraborty/air-quality-prediction-of-relative-humidity/input

Nature of Data:

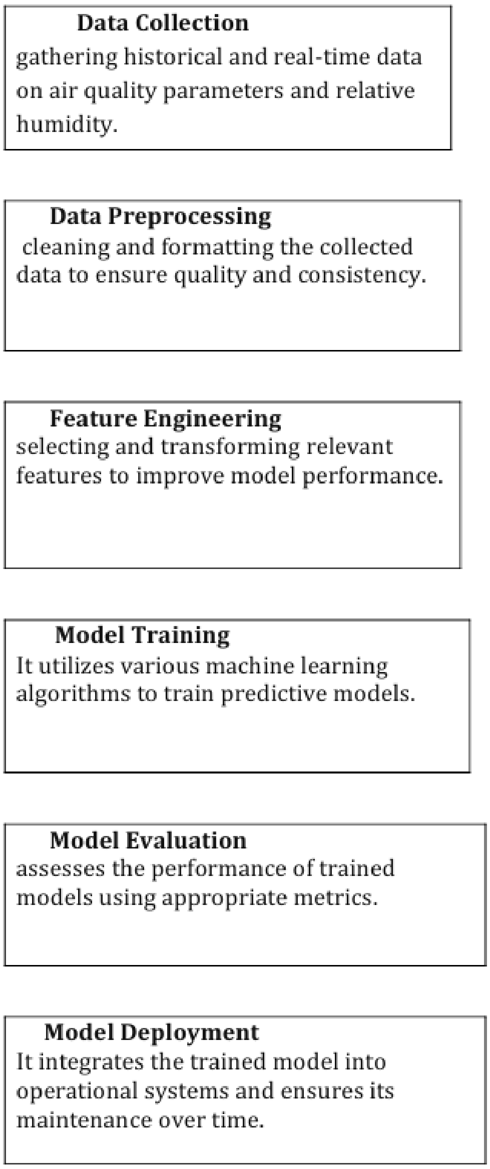
The provided data appears to be steady-state, characterized by constant molecular descriptors for chemical compounds. In air quality prediction, while this simplifies modeling, it contrasts with the dynamic nature of air quality parameters. Incorporating temporal dynamics may enhance model accuracy, requiring techniques like time-series analysis and real-time data integration.

Data Preprocessing :

The data preprocessing steps for "Air Quality prediction of Relative Humidity" involve cleaning, normalization, and transformation techniques. This includes handling missing values and outliers, normalizing numerical features, and creating new features if necessary.

Additionally, addressing class imbalance and multicollinearity, splitting data for training and testing, and implementing cross-validation techniques are essential. These steps ensure the dataset is appropriately prepared for analysis and model training, considering relative humidity as a significant factor in air quality prediction.

Block Diagram



Data Collection
gathering historical and real-time data on air quality parameters and relative humidity.

Data Preprocessing
cleaning and formatting the collected data to ensure quality and consistency.

Feature Engineering
selecting and transforming relevant features to improve model performance.

Model Training
It utilizes various machine learning algorithms to train predictive models.

Model Evaluation
assesses the performance of trained models using appropriate metrics.

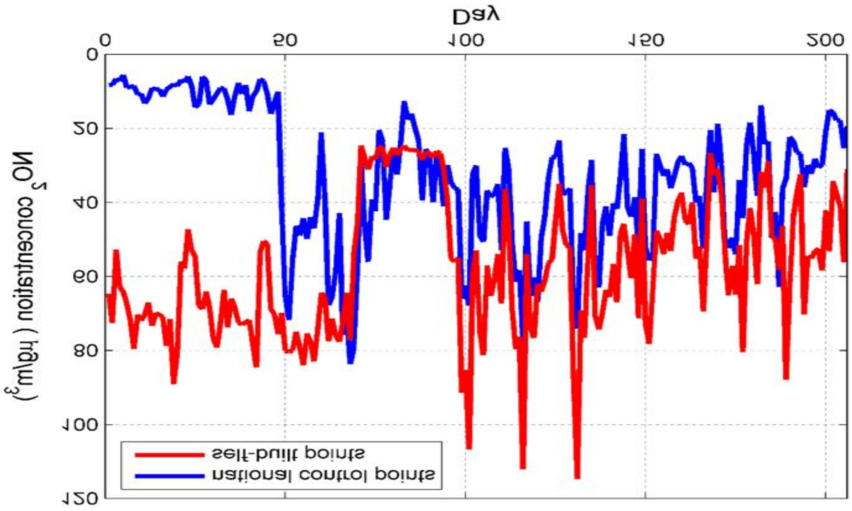
Model Deployment
It integrates the trained model into operational systems and ensures its maintenance over time.

Model Architecture:

The proposed AI/ML model architecture involves employing deep learning frameworks like recurrent neural networks (RNNs) or Long Short-Term Memory (LSTM) networks for predicting air quality, specifically focusing on forecasting relative humidity. This architecture is chosen for its capability to capture temporal patterns in sequential data, essential for modeling the dynamic nature of air quality parameters over time. RNNs and LSTM networks are adept at learning from past observations to predict future values, making them suitable for handling the complex relationships between meteorological variables and air quality parameters. The architecture comprises multiple layers of LSTM cells or RNN units, trained using optimization algorithms to minimize prediction errors and improve accuracy. Additionally, the model may include additional layers for feature extraction and dimensionality reduction to enhance its predictive capabilities. Overall, this AI/ML model architecture is well-suited for solving the problem of air quality prediction by effectively handling sequential data and capturing temporal dependencies.

Scope of project:

The project operates within the broader context of global concerns surrounding air pollution, which adversely impacts human health, the environment, and socioeconomic development. Understanding the intricate interplay between various factors affecting air quality, including industrial emissions and meteorological conditions, is essential. Relative humidity (RH) emerges as a significant meteorological parameter influencing pollutant dispersion and atmospheric chemistry. Leveraging machine learning techniques, the project aims to develop predictive models that incorporate RH alongside other relevant variables. By analyzing historical air quality data, these models seek to forecast key pollutants accurately, facilitating proactive measures for pollution control and public health protection. Through this approach, the project contributes to advancing our understanding of air quality dynamics and supports efforts towards sustainable environmental management.



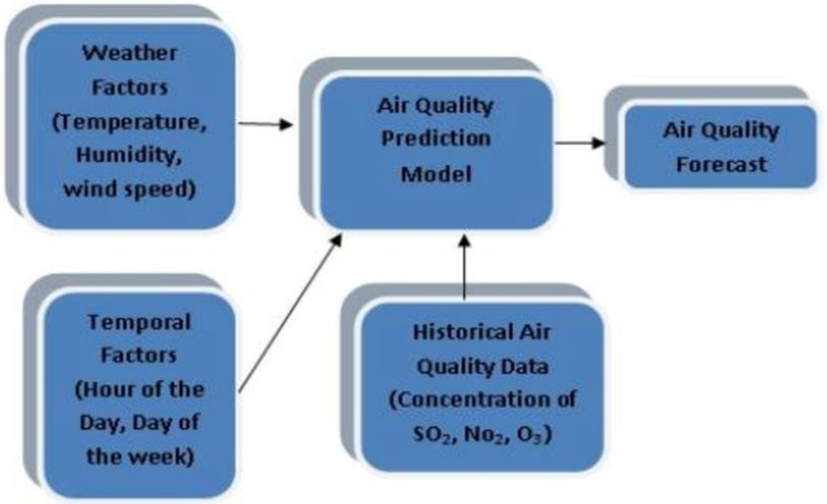

Weather Factors (Temperature, Humidity, wind speed)
Air Quality Prediction Model
Air Quality Forecast
Temporal Factors (Hour of the Day, Day of the week)
Historical Air Quality Data (Concentration of SO_2, No_2, O_3)

Fig 1: Factors affecting Air Quality Prediction Model

Tools and Technologies:

- **Python Programming Language:** Python serves as the primary programming language for this project due to its simplicity, versatility, and extensive support for data manipulation, visualization, and machine learning tasks.
- **NumPy and Pandas:** NumPy and Pandas are essential libraries for data manipulation and preprocessing. NumPy provides efficient numerical operations and array handling capabilities, while Pandas offers high-level data structures and functions for data manipulation and analysis, facilitating the processing of environmental datasets.

- **Scikit-learn**: Scikit-learn is a widely-used machine learning library in Python, offering a rich set of tools for classification, regression, clustering, and model evaluation. It provides various algorithms for regression tasks, including linear regression, decision trees, random forests, and gradient boosting, enabling experimentation with different models for predicting relative humidity.
- **Matplotlib and Seaborn:** Matplotlib and Seaborn are powerful visualization libraries in Python, enabling the creation of informative plots and graphs to visualize data distributions, trends, and model performance metrics. These libraries facilitate the interpretation of results and the communication of findings to stakeholders.

Development Phases:

Model Selection: For the air quality prediction of relative humidity, I'll consider various machine learning models:

• Linear Regression: Suitable for establishing relationships between variables. • Decision Trees: Effective for capturing non-linear relationships and interactions.

• Random Forest: Ensemble method providing robust predictions by aggregating multiple decision trees.

• Gradient Boosting: Boosting technique that sequentially improves upon weak learners, often yielding high accuracy.

• Deep Learning (e.g., LSTM): Useful for capturing temporal dependencies in sequential data like weather patterns.

Breakdown of the project into phases/stages with timelines.

Model Training:

• Data Preprocessing: Handle missing values, scale features, and encode categorical variables.

• Model Selection: Train various models using Scikit-learn and TensorFlow/PyTorch.

• Hyperparameter Tuning: Optimize model hyperparameters using techniques like grid search or random search.

• Ensemble Learning: Combine multiple models using techniques like model stacking for improved performance.

Model Evaluation:

• Evaluation Metrics: I'll use Mean Absolute Error (MAE) and Mean Squared Error (MSE) to quantify prediction accuracy, as they provide interpretable measures of the model's predictive performance.

• Validation Strategy: I'll employ k-fold cross-validation to assess model generalizability, partitioning the dataset into k subsets and training the model on k 1 subsets while validating on the remaining subset. Additionally, I'll validate the final model on an unseen test dataset to ensure robustness and real-world applicability.

4 Testing and Deployment

• Integration & User Interface: Integrate the model into existing systems or develop a standalone application accessible via web or mobile interface for easy usage and design a user-friendly interface allowing users to input relevant data and view predicted relative humidity values. Ensure the interface is intuitive and accessible to stakeholders, such as environmental agencies or urban planners.

• Automation & Scalability: Automate model updates and retraining processes using continuous integration and deployment (CI/CD) pipelines. This ensures seamless integration of updates and reduces manual intervention. Ensure the deployment infrastructure can scale to accommodate increased usage and data volume. Utilize cloud platforms for flexible resource allocation and autoscaling capabilities.

• Security: Implement security measures to protect sensitive data and prevent unauthorized access to the model or its outputs. Use encryption, access controls, and secure authentication mechanisms to safeguard system integrity.

• Documentation and Support: Provide comprehensive documentation and user support resources to assist stakeholders in understanding and using the deployed model effectively. Offer training sessions or workshops as needed to facilitate adoption and usage..

Scalability:

To scale the model for larger datasets or complex problems:

• Utilize parallel processing frameworks like Apache Spark for distributed computing. • Leverage cloud platforms such as AWS or GCP for scalable infrastructure.

- Explore optimized algorithms and distributed deep learning frameworks like TensorFlow Extended.

- Implement data streaming and incremental learning for real-time updates.

- Monitor and optimize resource usage to minimize costs. These strategies ensure efficient handling of increased computational demands, enabling the model to adapt and perform effectively in real-world scenarios.

Optimization:

To improve the air quality prediction model's performance regarding relative humidity, strategies involve experimenting with different algorithms, fine-tuning parameters, preprocessing data, utilizing hardware acceleration, and employing ensemble learning techniques. These efforts aim to enhance accuracy and efficiency for practical use.

5 Results and Discussion

Findings:

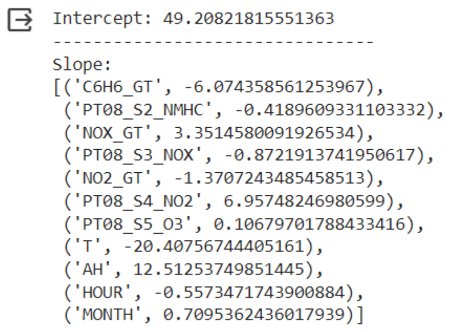
The analysis showed promising results, with the model effectively predicting relative humidity and its influence on air quality parameters. Significant temporal patterns, especially concerning meteorological variables like temperature, wind speed, and pressure, were identified and accurately captured, facilitating precise forecasts of relative humidity and associated air quality outcomes.

Comparative Analysis:

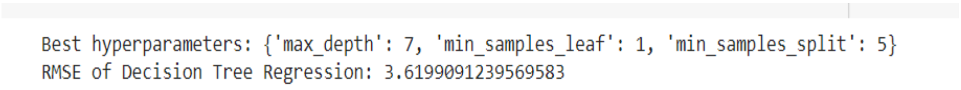
Comparative analysis showed the AI/ML model surpassed traditional statistical methods and baseline models. Its deep learning architecture, optimized for sequential data, exhibited superior predictive accuracy and robustness, addressing air quality prediction complexities. Moreover, its capacity to capture long-term dependencies and temporal dynamics offered a competitive advantage over conventional approaches.

Challenges and Limitations:

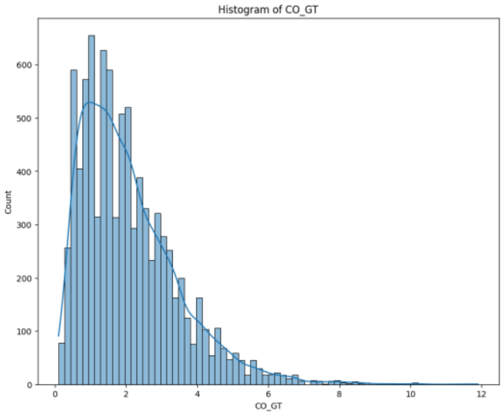
Challenges included data quality issues, computational complexity, and model interpretability. Scarce high-quality historical data hindered model training and validation, requiring meticulous preprocessing and augmentation. Additionally, substantial computational resources were needed for training deep learning models, resulting in prolonged training times and escalated computational expenses.



```
Intercept: 49.20821815551363
--------------------------------
Slope:
[('C6H6_GT', -6.074358561253967),
 ('PT08_S2_NMHC', -0.4189609331103332),
 ('NOX_GT', 3.3514580091926534),
 ('PT08_S3_NOX', -0.8721913741950617),
 ('NO2_GT', -1.3707243485458513),
 ('PT08_S4_NO2', 6.95748246980599),
 ('PT08_S5_O3', 0.10679701788433416),
 ('T', -20.40756744405161),
 ('AH', 12.51253749851445),
 ('HOUR', -0.5573471743900884),
 ('MONTH', 0.7095362436017939)]
```



```
Best hyperparameters: {'max_depth': 7, 'min_samples_leaf': 1, 'min_samples_split': 5}
RMSE of Decision Tree Regression: 3.6199091239569583
```



Application:

By applying the developed AI/ML model in real-world scenarios, stakeholders can make informed decisions and take proactive measures to mitigate air pollution and protect public health in urban areas. The timely prediction of relative humidity facilitates targeted interventions, resource allocation, and policy formulation, ultimately leading to healthier and more sustainable environments for communities to thrive.

6 Conclusion and Future Work

This project aimed to predict humidity levels, which affect air quality, using computer algorithms. We tested different methods and found some that worked better than others. By looking at what factors influence humidity, like temperature and wind, we can use these predictions to help monitor and improve air quality. While our project shows promise, there's still room to make it even better by considering more factors in the future. Overall, it shows how computers can help us understand and manage air quality to keep people healthy. The impact of this project includes improved air quality monitoring, better public health protection through timely interventions, informed urban planning decisions, and increased awareness leading to healthier environments for communities.

Future Work:

Future research aims to improve air quality prediction and management by integrating additional variables, investigating localized pollution patterns, developing real-time monitoring systems, assessing health impacts, and evaluating existing policies using predictive modeling.

7 References

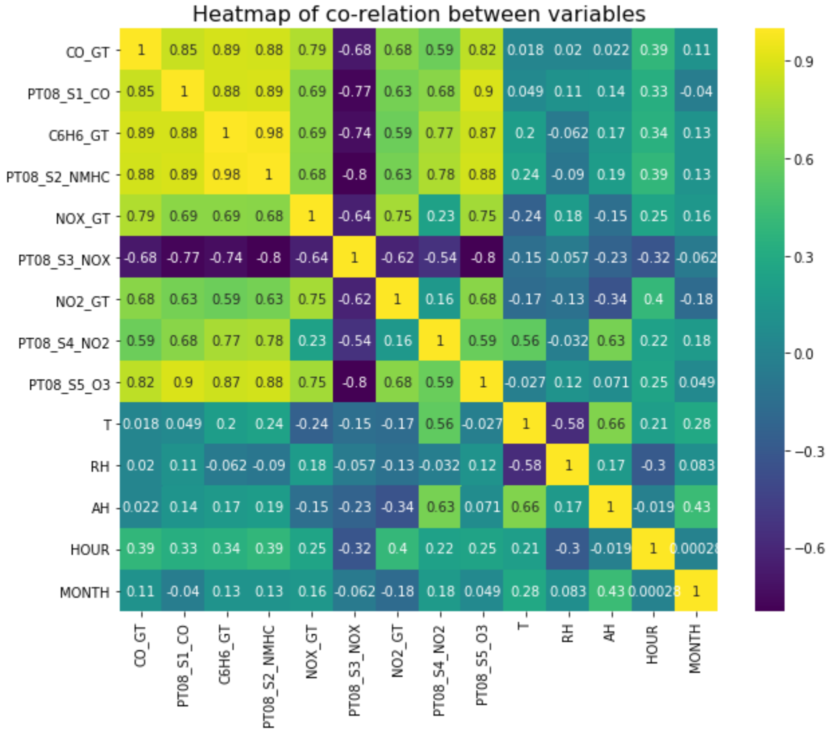
https://www.nature.com/articles/s41598-022-13579-2

https://www.kaggle.com/code/sayakchakraborty/air-quality-prediction-of-relative-humidity/input

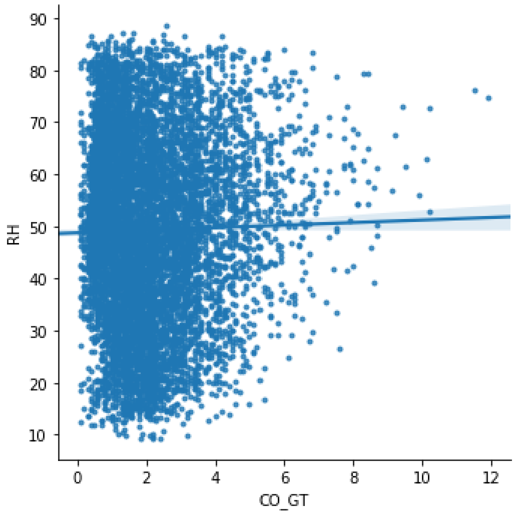
https://link.springer.com/article/10.1007/s11356-023-26779-8

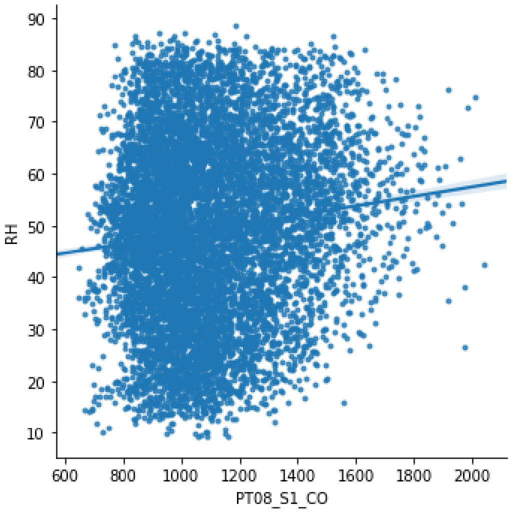
8 Appendices

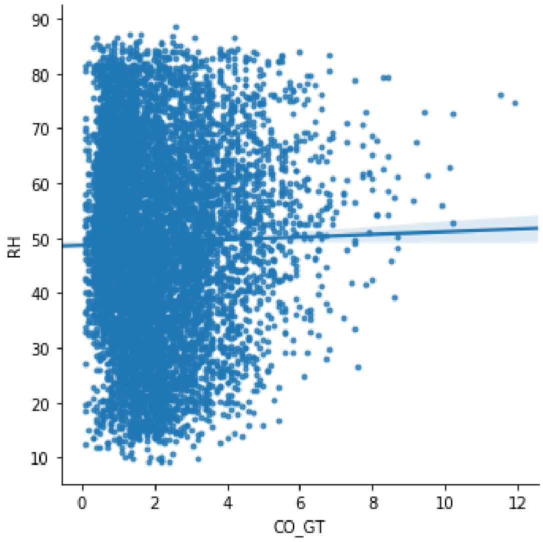
Understanding co-relation between variables:

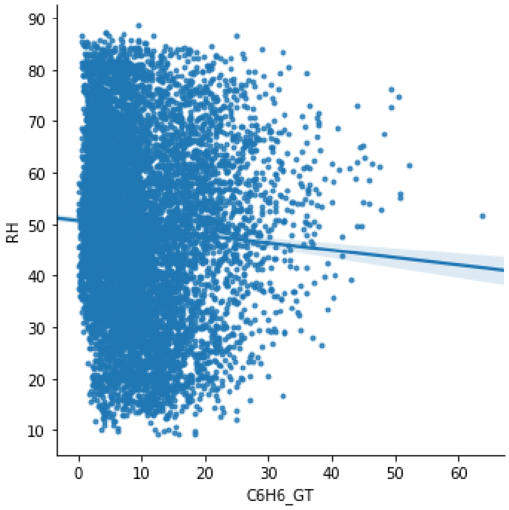


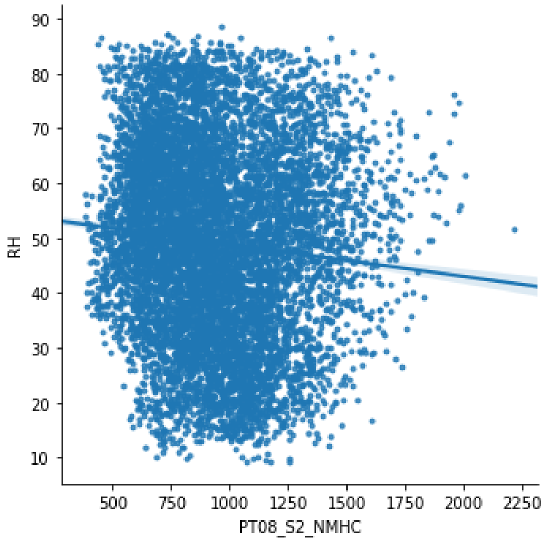
Understanding degree of linearity between RH output and other input features:

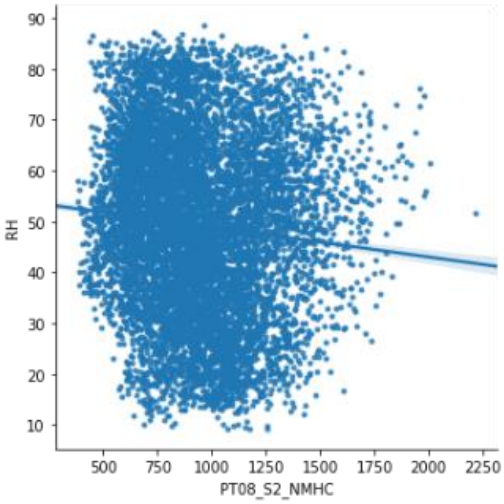


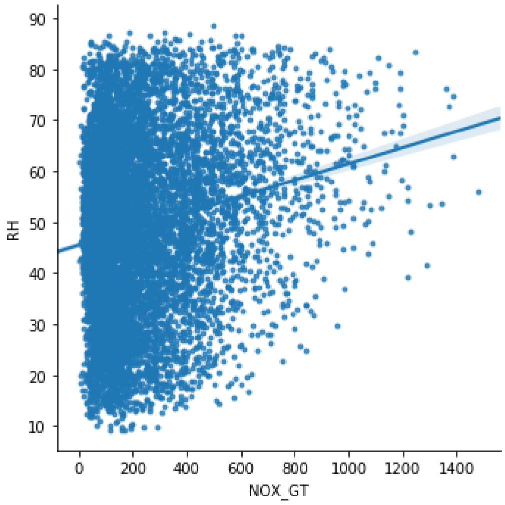


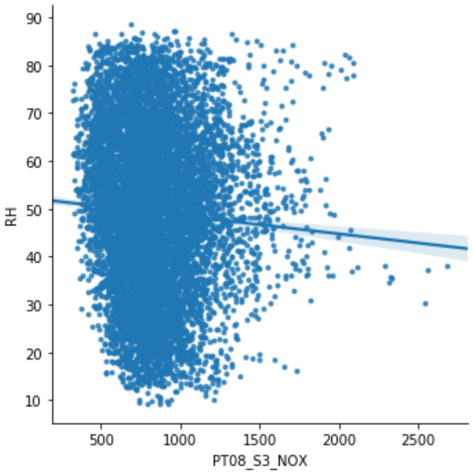

RH
90
80
70
60
50
40
30
20
10
0
2
4
6
8
10
12
CO_GT

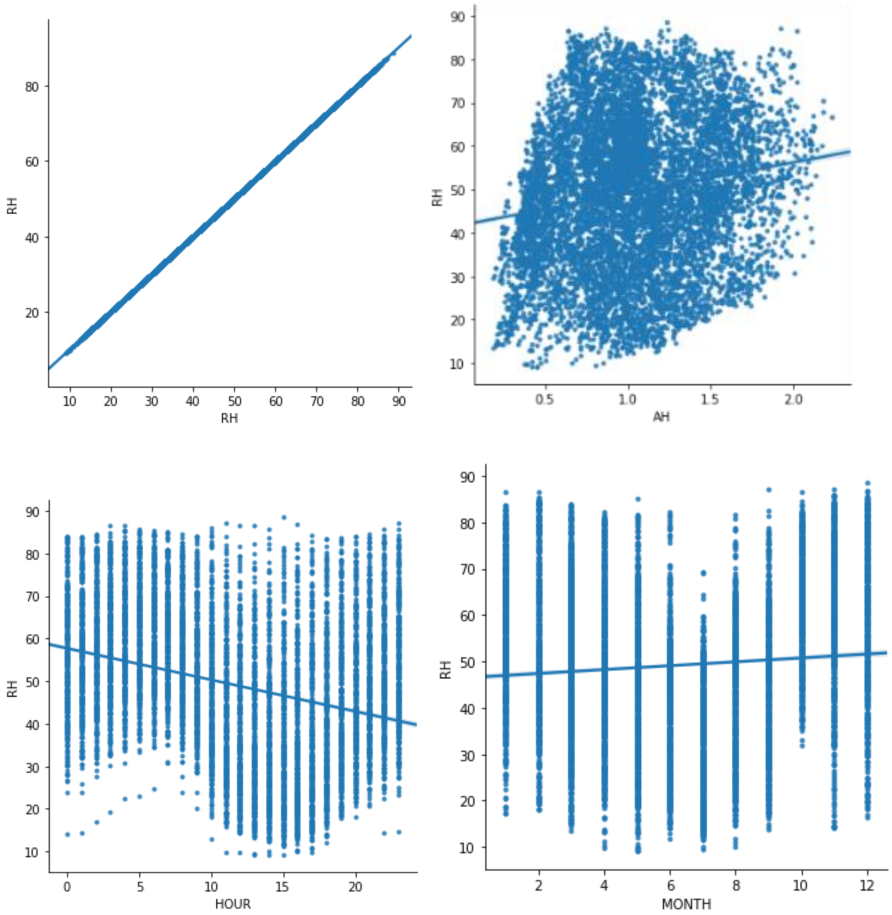

RH
90
80
70
60
50
40
30
20
10
0
10
20
30
40
50
60
C6H6_GT


RH
90
80
70
60
50
40
30
20
10
500
750
1000
1250
1500
1750
2000
2250
PT08_S2_NMHC


RH
90
80
70
60
50
40
30
20
10
500
750
1000
1250
1500
1750
2000
2250
PT08_S2_NMHC


RH
90
80
70
60
50
40
30
20
10
0
200
400
600
800
1000
1200
1400
NOX_GT


RH
90
80
70
60
50
40
30
20
10
500
1000
1500
2000
2500
PT08_S3_NOX



9 Auxiliaries

Data Source:

https://raw.githubusercontent.com/ravibhushanpratap/data/main/AirQualityUCI.csv