Historical Data Mining for Predicting Air Quality in Victoria, Canada

Alvin Guo - V00987315
Daming Wang - V00960801
Christopher Xu - V01007912

BackGround

1. Air quality significantly impacts both the environment and public health worldwide

2. Urban environments, such as Victoria, Canada, face intricate air quality challenges.

3. These challenges encompass factors such as climate change, industry emissions, vehicular pollutants, and evolving urban development patterns.

4. These diverse factors interact in multifaceted ways, affecting the composition of the air.

Objective

The primary objective of this research is to develop a data-driven model that accurately predicts future air quality in Victoria, Canada, based on historical data. To meet this objective, the study is guided by the hypothesis that discernible patterns and trends in past air quality data, combined with relevant location and date factors, can provide reliable predictive capabilities.

This ultimate goal is to contribute to the scientific understanding of air quality dynamics and provide actionable insights to guide policy development, inform public health strategies, support environmental conservation efforts in Victoria, and potentially extendable to other regions.

Related Work

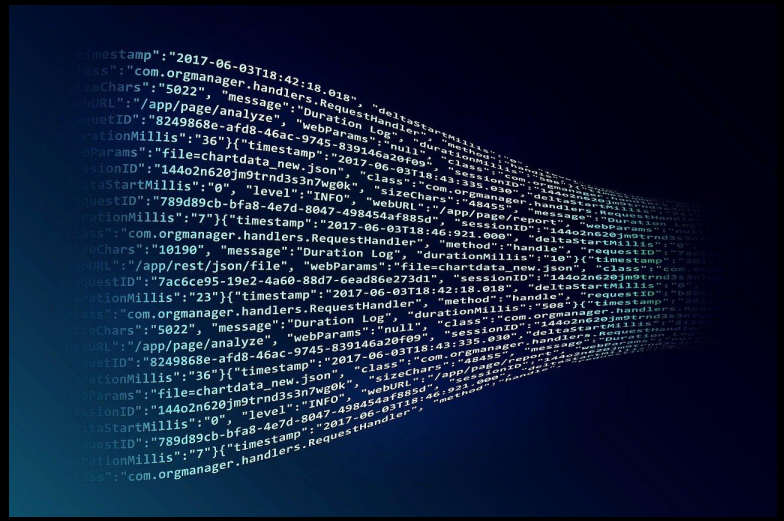
“Prediction of PM2.5 Concentrations using Random Forest Models”

A significant body of research has examined data mining models, such as the random forest model, for predicting PM2.5 concentrations. Huang(2018) conducted a comprehensive analysis of the prediction of high-resolution PM2.5 concentrations using the random forest model in the North China Plain. Huang underscored the random forest model’s capabilities in handling large-scale datasets, nonlinear relationships, and interactions among predictors.

“Extreme Gradient Boosting (XGBoost) Models in Estimating NO2 Concentrations”

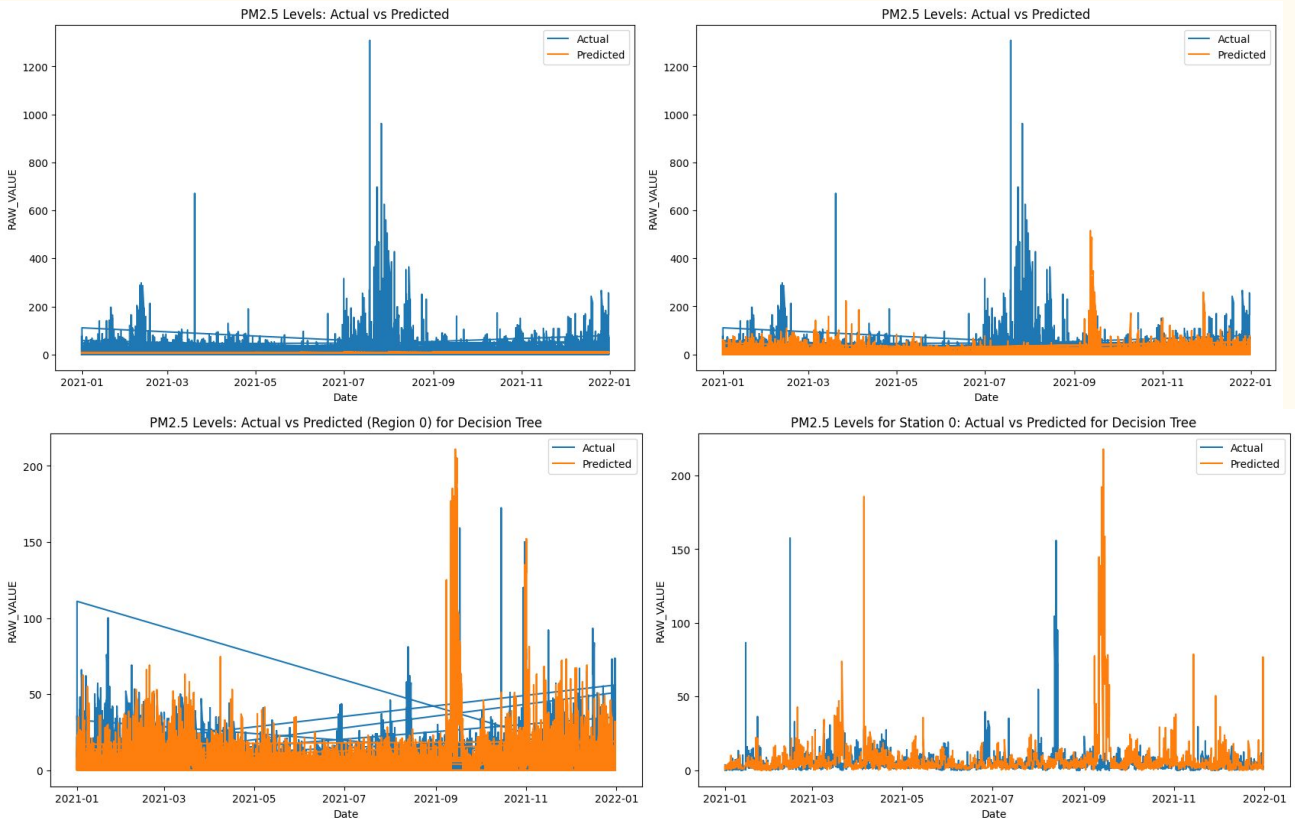
Recent advancements in remote sensing have facilitated the use of satellite data in environmental research. Liu(2022) employed Extreme Gradient Boosting, a popular machine learning and data mining algorithm. In conjunction with MODIS satellite retrievals, generate 250 m-resolution regional NO2 concentration products. The XGBoost model’s ability to capture complex relationships and handle large datasets was emphasized, contributing to improved estimation accuracy. Liu (2022) concludes the research by delineating the model's future applications in air quality monitoring, urban planning, and policy formulation.

Data Mining

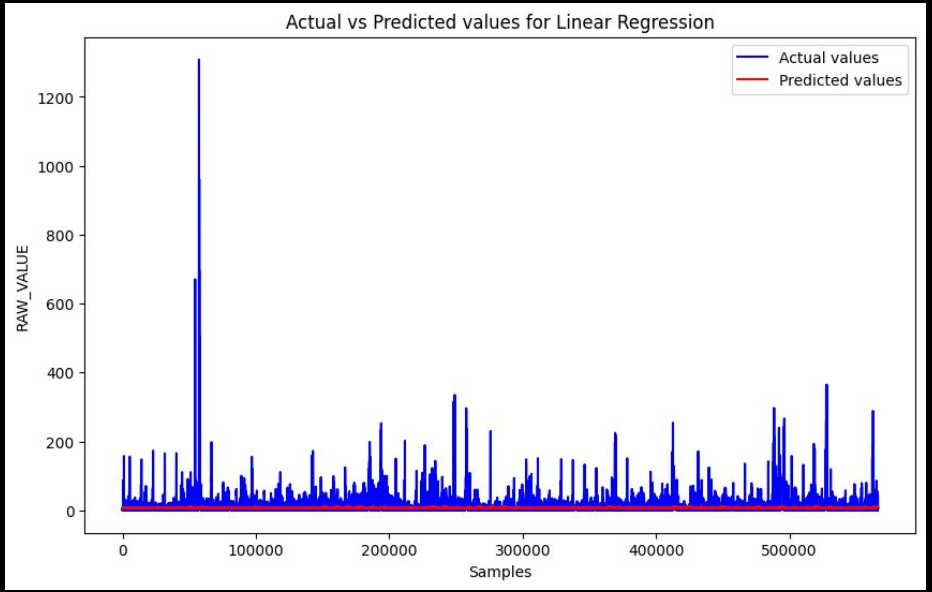


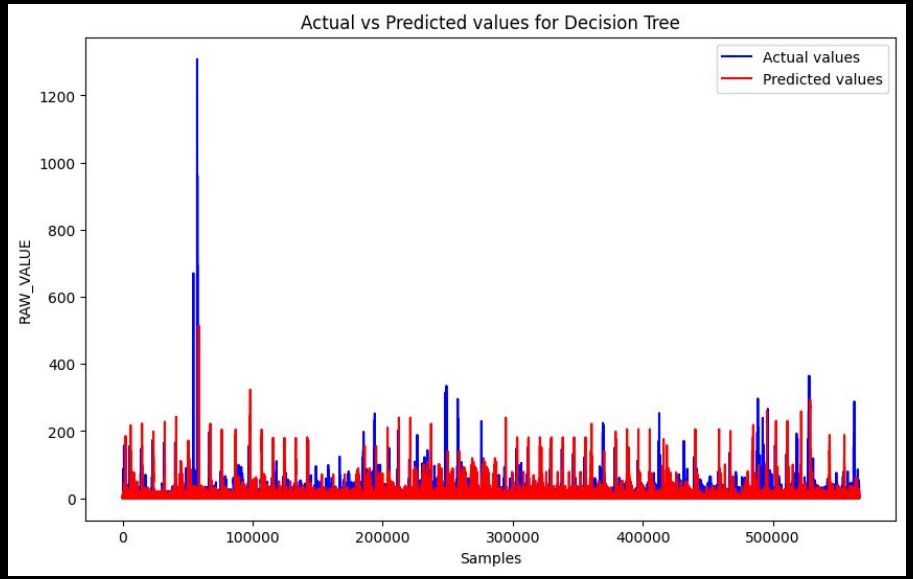
1. Observation
   a. PM 2.5
   b. NO2
   c. O3
2. Linear Regression
   a. Data
   b. Graph
3. Decision Tree
   a. Graph
4. Other Methods
   a. Logistic Regression
   b. Random Forest
   c. Gradient Boosting
   d. Ensemble Methods

Observation PM2.5

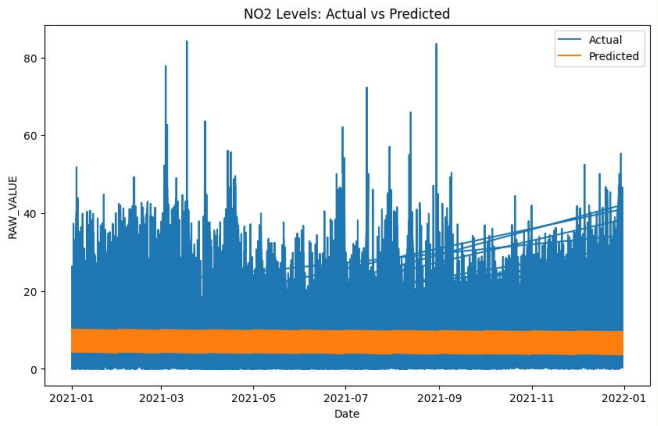


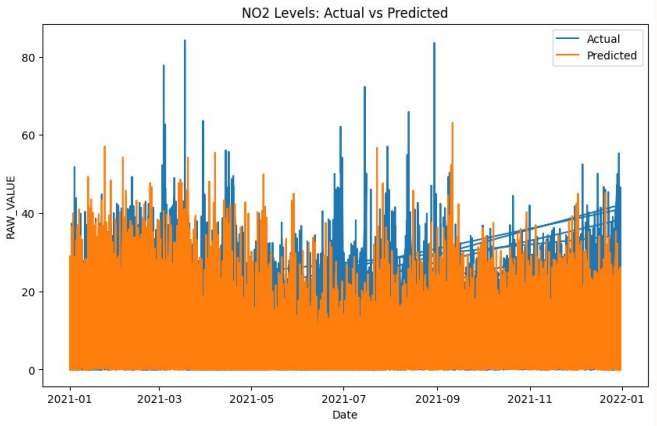
Observation
PM 2.5

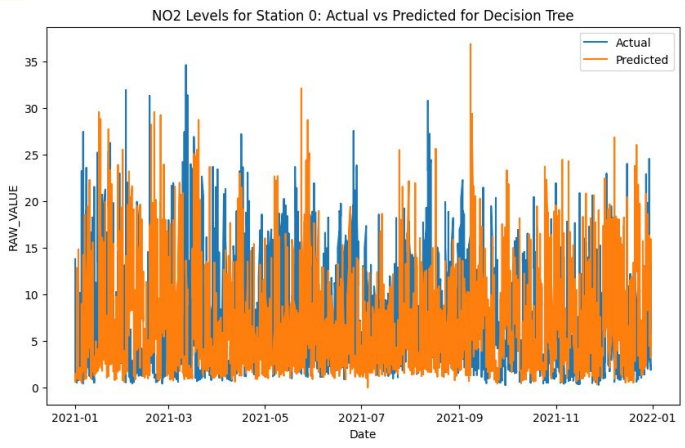


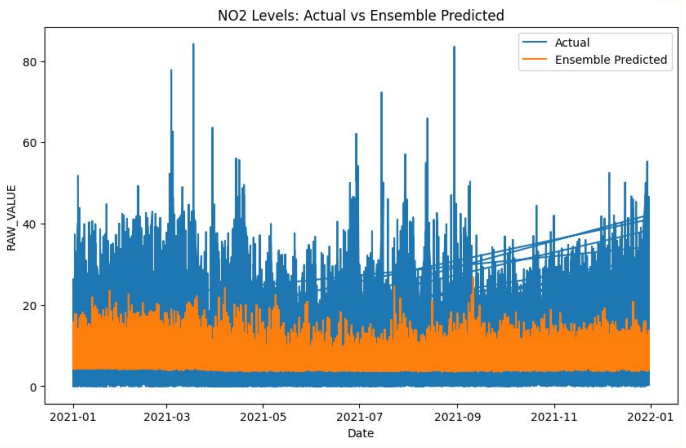


Observation NO2

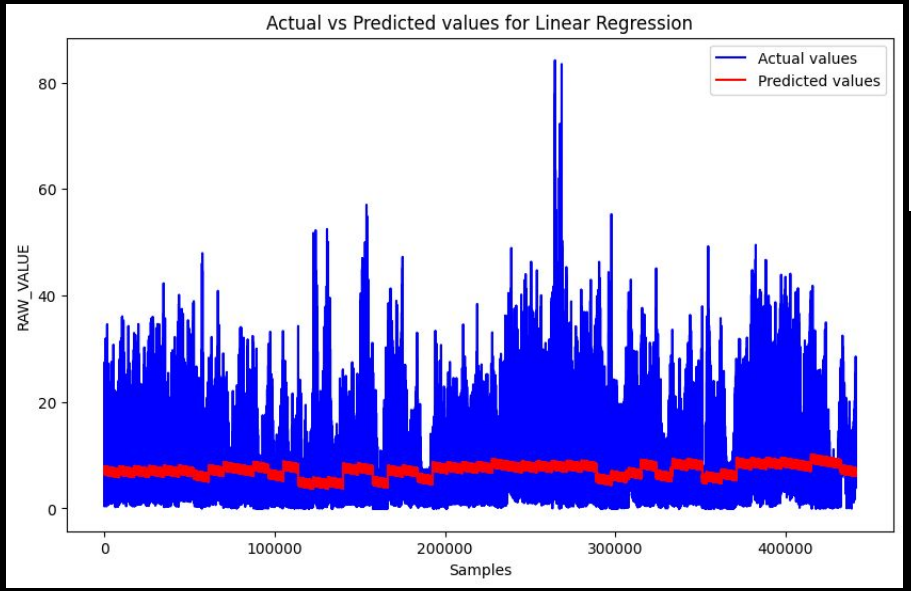


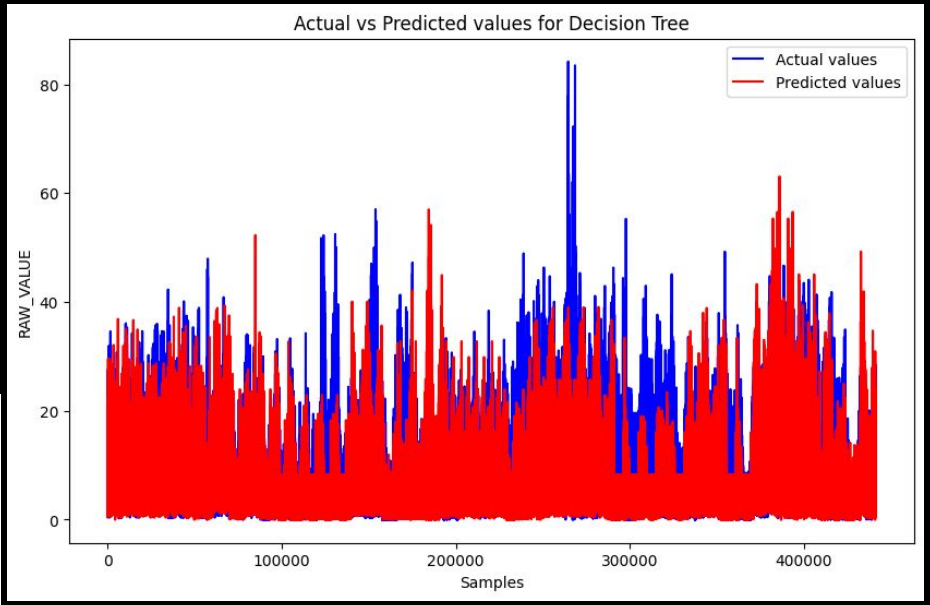




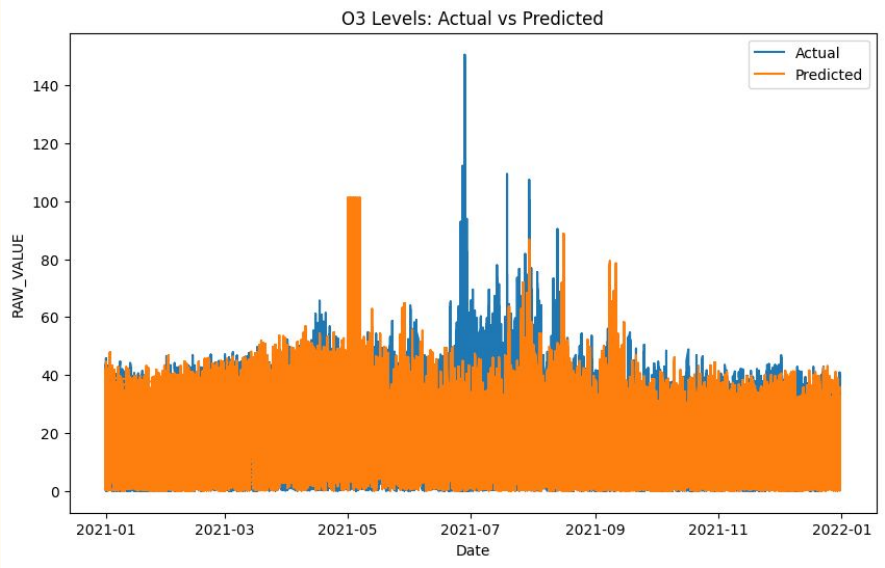


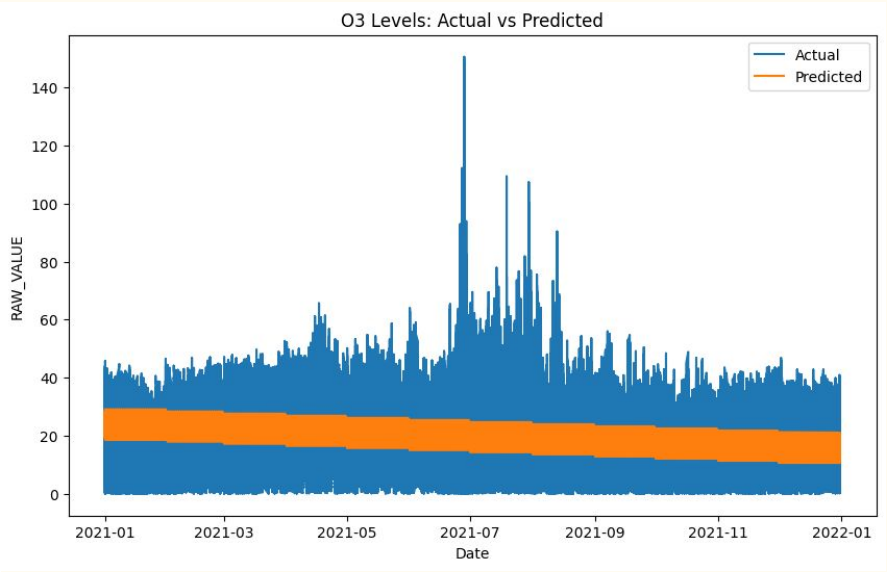
Observation
NO2



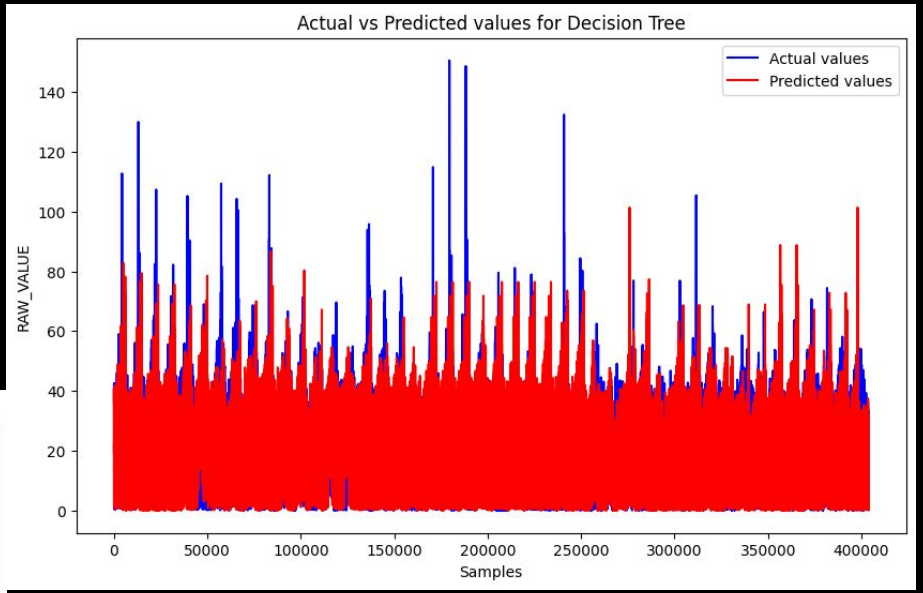


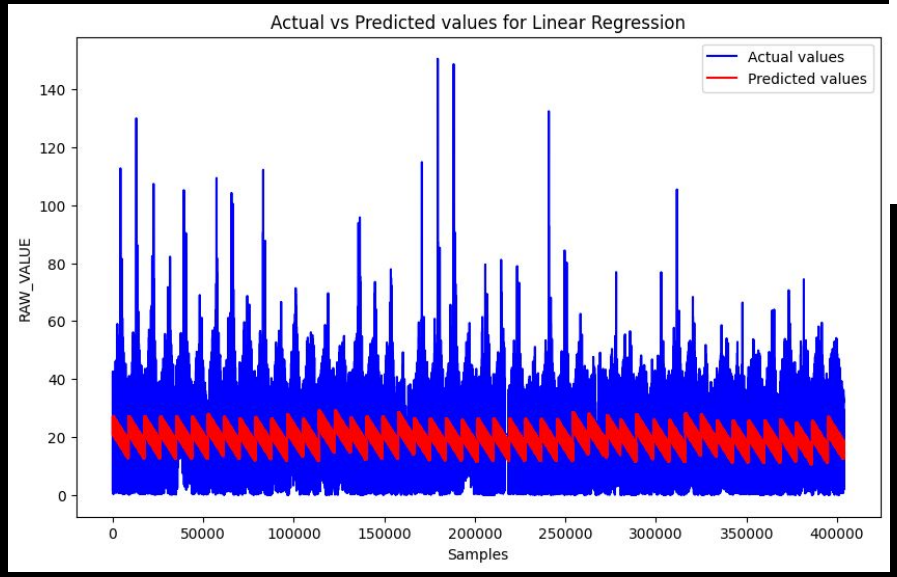
Observation
O3





Observation 03





Evaluation

Air pollution often happens around summer the most and winter secondly.

	MSE	Model	Most effectiveness
PM 2.5	188.7	Gradient Boosting	Year
NO2	38.6	Ensemble	Location
O3	119	Ensemble	Month

Implications and Future Work

Implications:

1. Environmental Conservataion
    a. Biodiversity
    b. Ecosystems
2. Urban Planning
    a. Location and Design

Future Work:

1. Broader Set of Predictive Variables
2. Advanced Data Mining Techniques

Conclusion

1. The PM 2.5 model predicts well using the Gradient Boosting method
2. PM 2.5 varies with the year attribute the most.
3. NO2 and O3 predicst well using Ensemble Model.
4. NO2 varies with the location the most.
5. O3 varies with the month the most.
6. As time goes on, PM 2.5 increases, NO2 and O3 decreases gradually.
7. All three pollution shows higher value during the summer time the most, and winter time secondly.

Bibliography

Data Source:
https://catalogue.data.gov.bc.ca/dataset/air-quality-monitoring-verified-hourly-data

Our Open-Source Project Repository:
https://github.com/EdNovas/seng474-project

Huang, K. (2018). “Predicting monthly high-resolution PM2.5 concentrations with a random forest model in the North China Plain”. Science Direct. https://doi.org/10.1016/j.envpol.2018.07.016

Liu, J. (2022). “Generating 250 m-resolution regional NO2 concentration products first from MODIS retrievals using extreme gradient boosting”. Springer Link. https://doi.org/10.1007/s11869-022-01285-x

Thanks