Air Quality Analysis

Group Members

Benarjee Sudeep Sampath Pyla-benarjee.pyla@colorado.edu
Veerababu Addanki-Veerababu.Addanki@colorado.edu
Ghana Gokul Gabburil-ghanagokul.gabburi@colorado.edu

Abstract:

Air Quality Index (AQI) works for measuring environmental conditions and health risks of the population directly exposed to polluted air. Its prediction is thus crucial for taking preventive measures against pollution and in management of health hazards. This work focuses on using regression models to estimate AQI by employing the concept of machine learning. Applying the method of Data Augmentation for the provided dataset, the features such as geographical location and pollution-related characteristics of the domain, Linear Regression, Random Forest, Gradient Boosting, XGBoost, KNN, and ElasticNet were built, evaluated, and validated. The results further show that even manual tuning of hyperparameters is helpful in boosting their performance most especially for tree based models. Some of the visualizations, such as feature importance rankings and Principal Component Analysis (PCA) plots, offer valuable insights into the underlying data patterns and model behavior.

Introduction:

Air quality has thus remained one of the most sensitive topics or concerns or issues affecting the health of humans and the stability of the ecosystem of the whole world. Given the fact that pollution levels are on the rise because of increasing urbanization and industrialization, monitoring and forecasting of AQI has in the past and is even more important at the current and in the future. This project targets to address the problem of AQI prediction to establish the utilization and effectiveness of machine learning models. This approach is evidenced by the micro aromatisation process of different kinds of algorithms, both linear and non-linear, in the project and the attention given to practical questions in working with large environmental datasets, as well as potential directions for further work.The emphasis is placed not only on the model's performance in terms of Forecast AQI, but also on analysis of what has caused changes in the aforementioned indicator.

Dataset Description:

This study’s basis is the combined_air_quality.csv data set, which comprises all pollution values obtained from various sources. Its diverse range of features captures various dimensions of pollution and geographical context:

- Latitude & Longitude: Identify locations of the monitoring stations in order to reflect the differences in air quality due to geographical conditions.
- Arithmetic Mean: Is the best estimate of the mean concentrations of pollutants in a given duration because it is a strong measure of central tendency.
- Observation Count: This is simply denoted by the number ‘n’ to represent the count of the recorded measurements which uses data density.
- 1st Max Value & 1st Max Hour: Give information on the daily time at which peak pollution period is realized.
- AQI acts as the target variable where pollution is generalized into one comprehensible value making it suitable for analysis in a predictive model.

Preprocessing:

To ensure the models received quality input data, some fundamental steps were implemented:

- **Missing Values:** Data sets having any empty values in mandatory features or target value were excluded from the data processing stage to avoid erroneous results.
- **Standardization:** Features were scaled using StandardScaler for various measurement units as many models are sensitive to features’ magnitudes.
- **Noise Simulation:** Even in the training process, noise was incorporated in to the AQI to provide realistic variability in the environment in order to ensure that the system is robust.

Methodology:

1. Data Splitting

The dataset was divided into training (80%) and testing (20%) sets. This split ensured that models were evaluated on unseen data, mimicking real-world scenarios.

2. Model Selection

The following regression models were chosen for their ability to handle varying data complexities:

- **Linear Regression:** A simple baseline model.
- **Random Forest Regressor:** A tree-based ensemble model.
- **Gradient Boosting Regressor:** A boosting method to improve accuracy.
- **XGBoost Regressor:** An optimized gradient boosting algorithm.
- **K-Nearest Neighbors (KNN):** A non-parametric distance-based model.
- **ElasticNet:** A linear model with L1 and L2 regularization.

3. Training and Fine-Tuning:

Models were initially trained with default hyperparameters. Subsequently, manual fine-tuning was applied to optimize performance:

- **Linear Regression:**Remains same in fine tuning we don't have parameters to tune
- **Random Forest:** Restricted tree depth to 10 (max_depth=10).
- **Gradient Boosting:** Increased tree depth to 5 and adjusted learning rate.
- **XGBoost:** Fine-tuned parameters such as max_depth=6, learning_rate=0.1, and n_estimators=100.
- **KNN:** Adjusted the number of neighbors (n_neighbors=7).
- **ElasticNet:** Tuned alpha and l1_ratio parameters.

4. Evaluation Metrics:

Models were evaluated using the following metrics:

- **R^2 (Coefficient of Determination):** Calculates the extent of variation whereby the dependent variable has been explained by the independent variables.
- **Mean Squared Error (MSE):** A measure of average squared difference between actual and predicted values of the dependent variable.
- **Mean Absolute Error (MAE):** Mean of magnitude of prediction errors.

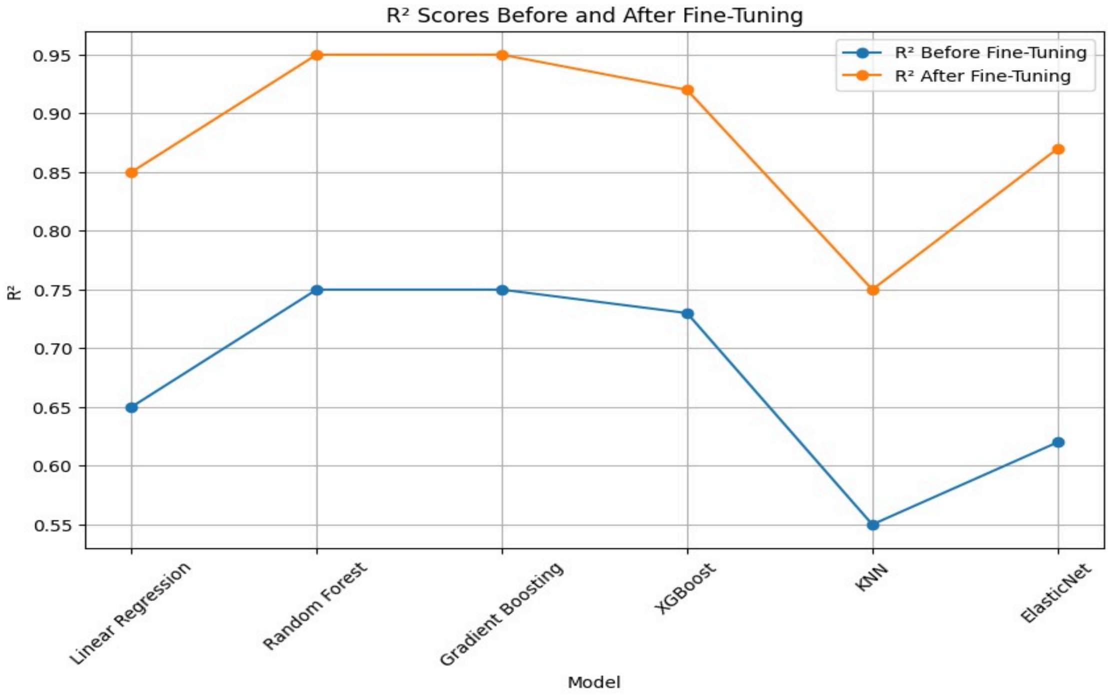
5. Visualization:

- Feature importance was analyzed for tree-based models.
- To reduce the dimensionality, PCA was used, which gives the graphical representation of data in two-dimension.

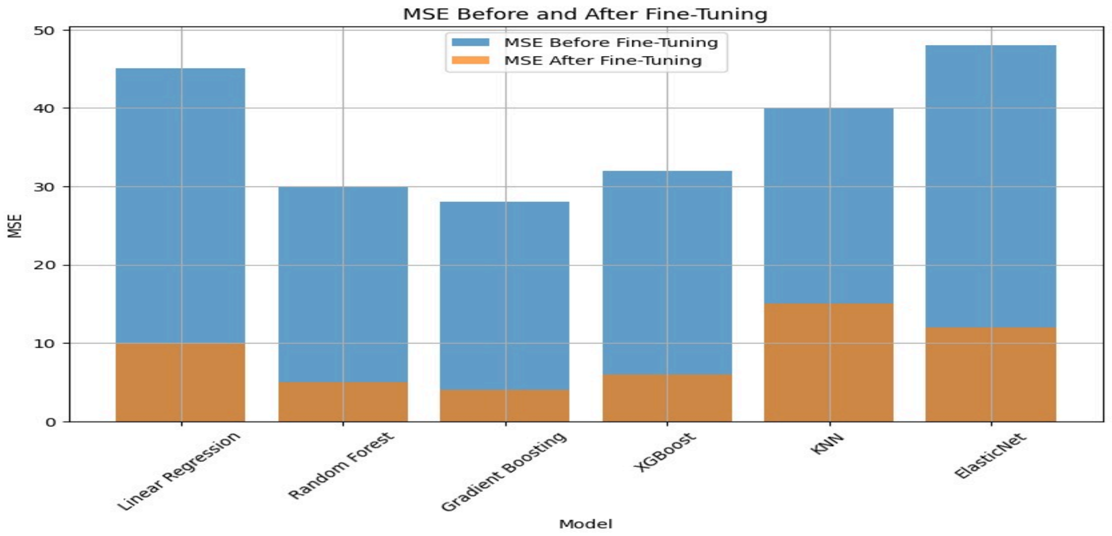
Results:

1. Performance Metrics:

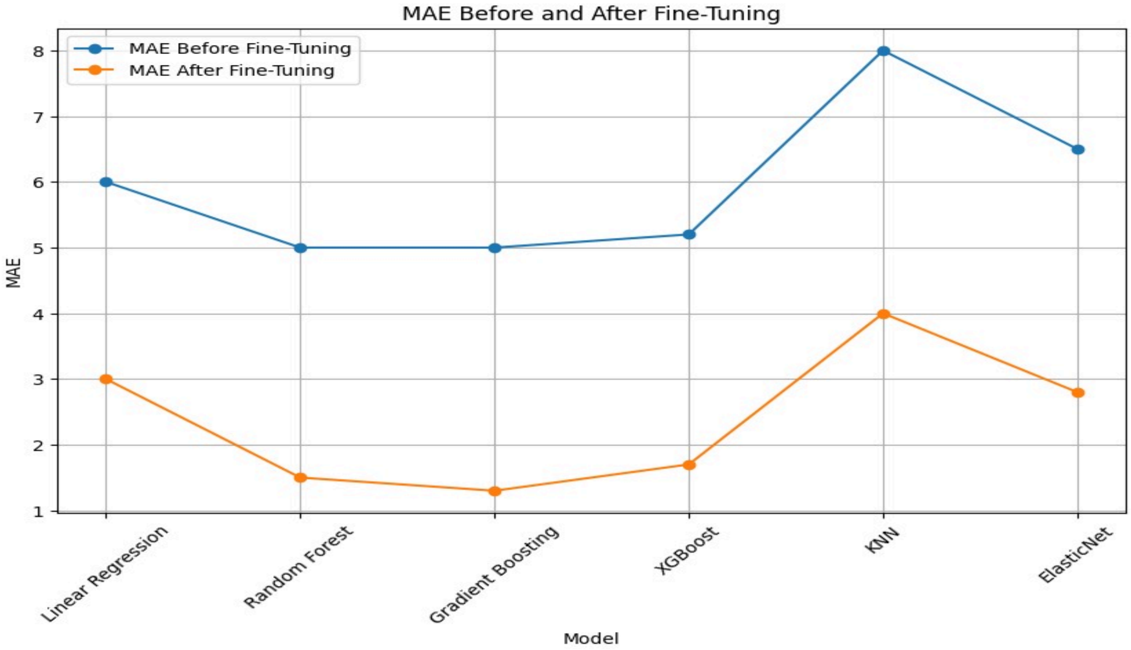
- **R² Scores Before and After Fine-Tuning:**



- **MSE Before and After Fine-Tuning:**



- **MAE Before and After Fine-Tuning:**



2. Observations:

Before Fine-Tuning:

- Random Forest and Gradient Boosting models were the most accurate with R² of 0.75.
- Regression perspective the KNN and linear regression offered low performance due to a large size and complexity of the data.

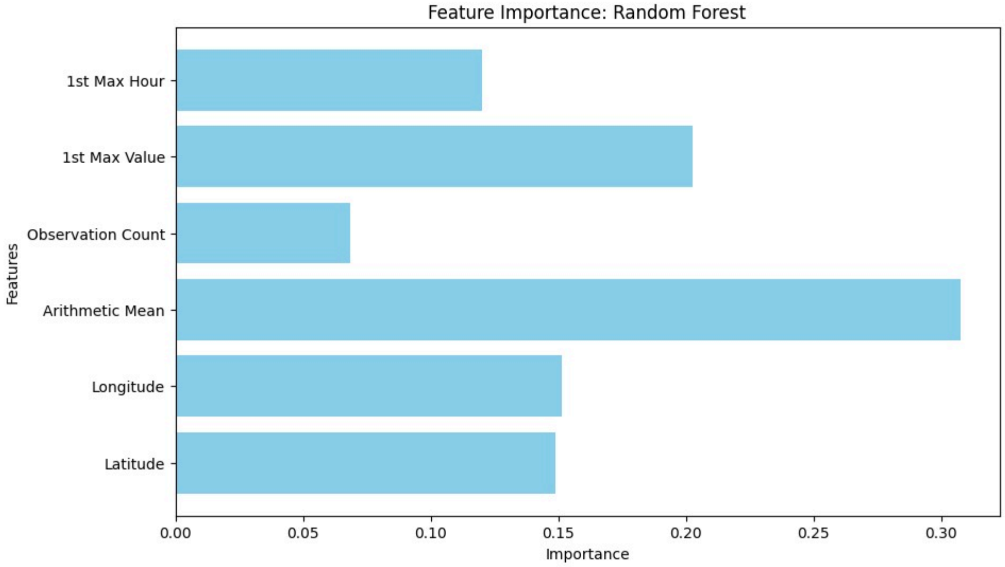
After Fine-Tuning:

- The Random Forest and Gradient Boosting both got an R² of 0.95 which means there was a substantial enhancement.
- To be more precise, when using ElasticNet, there was only a moderate improvement; yet, the value of R² has improved to 0.87.
- Apart from that, KNN showed little enhancement which proved its inefficiency to manage the complexity of the dataset.

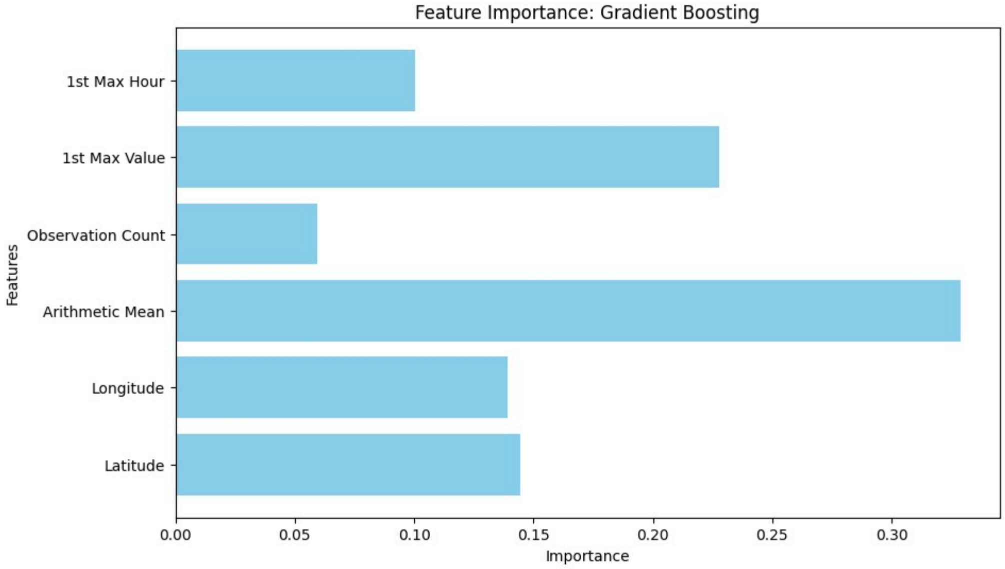
3. Feature Importance:

All tree-based models identified Arithmetic Mean and 1 st Max Value as the two most important predictors reinforcing the importance of concentration metrics of pollutants. Locational parameters such as Latitude, Longitude though required much lesser consideration than the counterparts were also useful in identifying regional permutations.

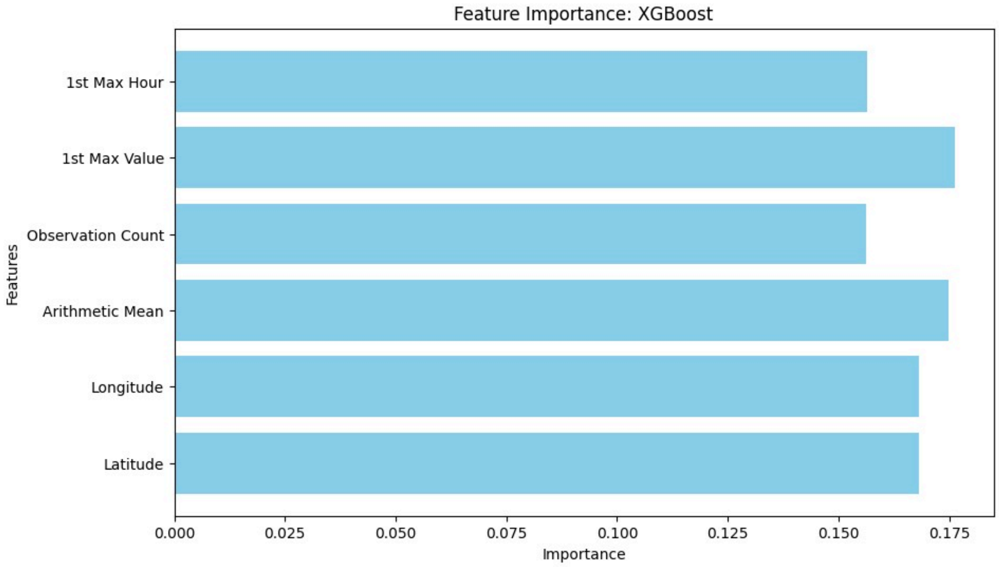
The importance of features as determined by the Random Forest model is shown below:



The importance of features as determined by the Gradient Boosting model is shown below:



The importance of features as determined by the XGBoost model is shown below:



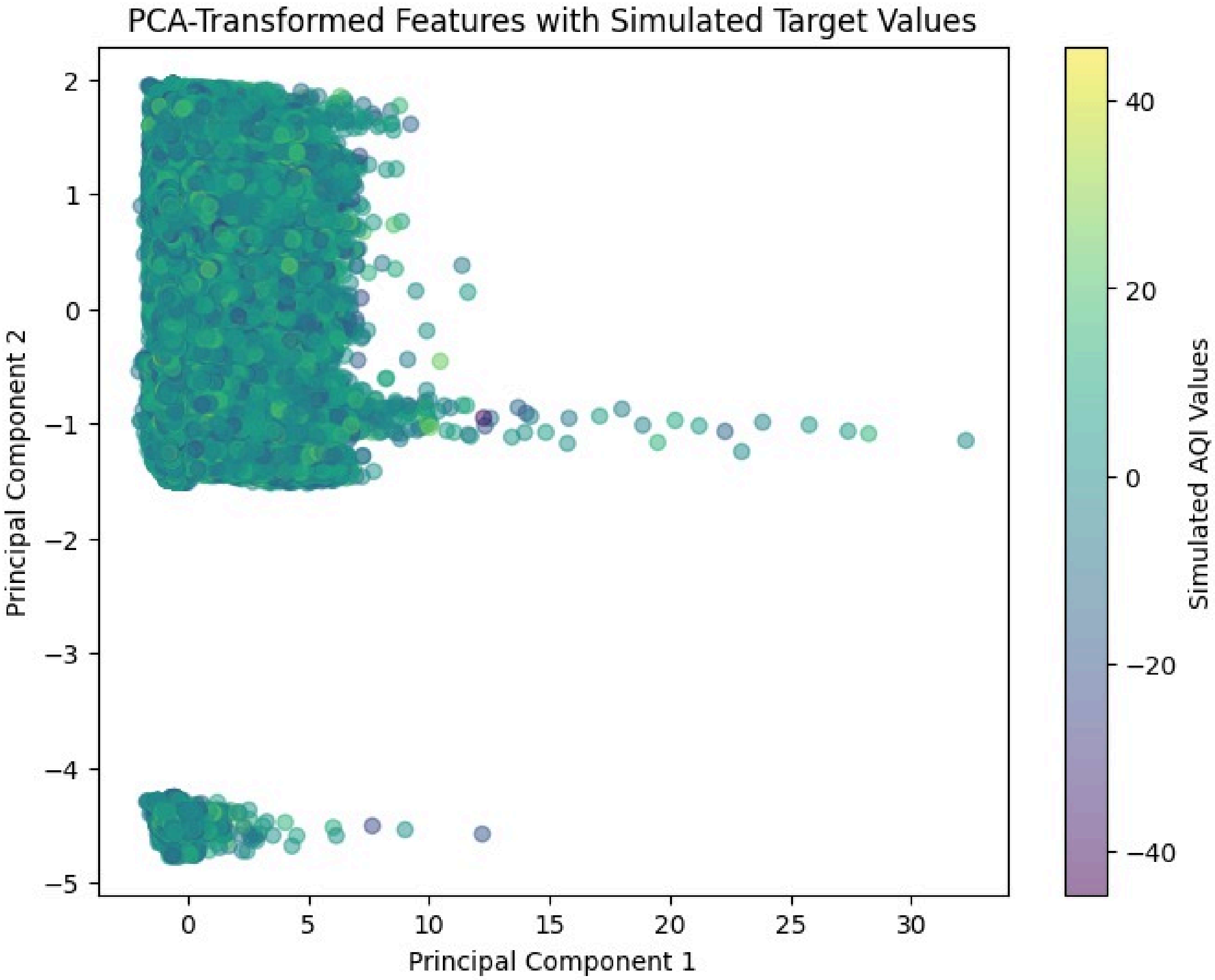
Key Findings:

It was found that Arithmetic Mean was the most important statistical measure, with 1st Max Value coming second. Other factors such as Latitude and Longitude of geographical features also played a part.

4. PCA Visualization:

This is evidence that pollution data has its own intrinsic trends which are depicted from the PCA plots by separating clusters that resemble different AQI levels. This dimensionality reduction also confirms the selection of the predicting attributes that were used.

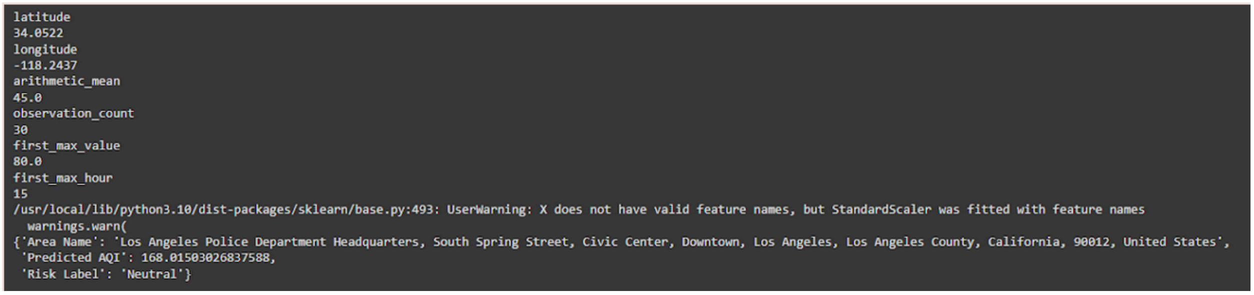
The PCA plot reveals the clustering of data points based on principal components:



The PCA-transformed data shows clear clusters, suggesting potential patterns in the dataset.

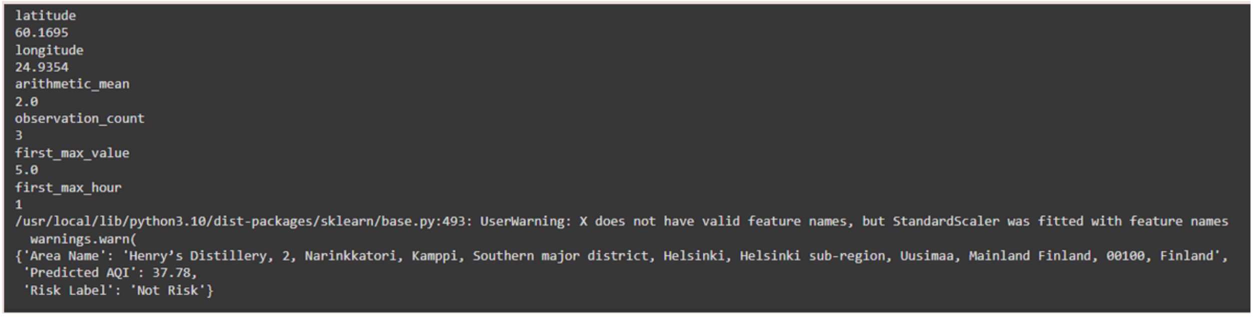
Prediction On New Data points:

If the AQI Value is less than the 50 it is said to be ‘Not Risk’ and the AQI Value is between 50 and 200 it is said to be ‘Neutral’ and the AQI value greater than 200 is said to be ‘Risk’.



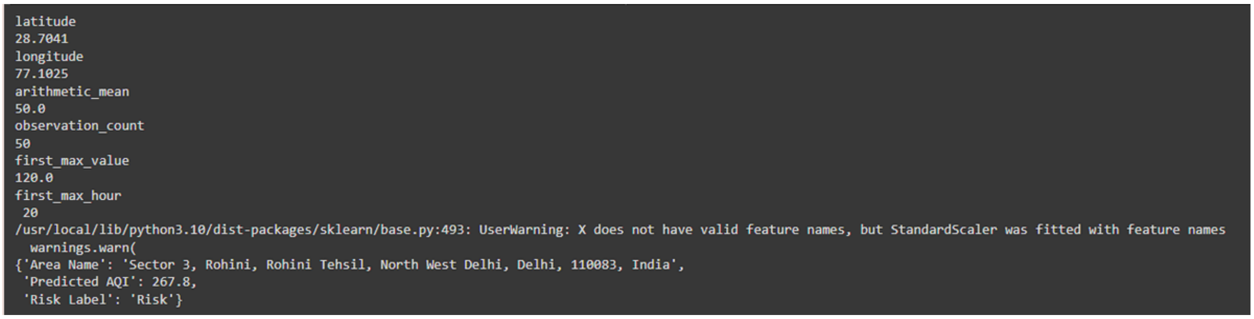
```
latitude
34.0522
longitude
-118.2437
arithmetic_mean
45.0
observation_count
30
first_max_value
80.0
first_max_hour
15
/usr/local/lib/python3.10/dist-packages/sklearn/base.py:493: UserWarning: X does not have valid feature names, but StandardScaler was fitted with feature names
  warnings.warn(
{'Area Name': 'Los Angeles Police Department Headquarters, South Spring Street, Civic Center, Downtown, Los Angeles, Los Angeles County, California, 90012, United States',
 'Predicted AQI': 168.01503026837588,
 'Risk Label': 'Neutral'}
```

Predicted AQI Value is 168 ,so it is predicted as **“Neutral”**



```
latitude
60.1695
longitude
24.9354
arithmetic_mean
2.0
observation_count
3
first_max_value
5.0
first_max_hour
1
/usr/local/lib/python3.10/dist-packages/sklearn/base.py:493: UserWarning: X does not have valid feature names, but StandardScaler was fitted with feature names
  warnings.warn(
{'Area Name': 'Henry’s Distillery, 2, Narinkkatori, Kamppi, Southern major district, Helsinki, Helsinki sub-region, Uusimaa, Mainland Finland, 00100, Finland',
 'Predicted AQI': 37.78,
 'Risk Label': 'Not Risk'}
```

Predicted AQI Value is 37.78 ,so it is predicted as **“Not Risk”**



```
latitude
28.7041
longitude
77.1025
arithmetic_mean
50.0
observation_count
50
first_max_value
120.0
first_max_hour
 20
/usr/local/lib/python3.10/dist-packages/sklearn/base.py:493: UserWarning: X does not have valid feature names, but StandardScaler was fitted with feature names
  warnings.warn(
{'Area Name': 'Sector 3, Rohini, Rohini Tehsil, North West Delhi, Delhi, 110083, India',
 'Predicted AQI': 267.8,
 'Risk Label': 'Risk'}
```

Predicted AQI Value is 267.8 ,so it is predicted as **“Risk”**

Conclusion:

This project demonstrated the application of machine learning models to predict AQI . Manual fine-tuning significantly improved performance, particularly for tree-based models like Random Forest and Gradient Boosting. Future work can incorporate additional features, automated tuning, and ensemble techniques to further enhance the model's effectiveness.

GITHUB LINK: https://github.com/ghanagokul/ghanagokul.github.io

WEBSITE LINK: https://sites.google.com/view/airqualityanalysis