Big Data Analysis and Project – Part D

# Air Aware: AQI Prediction and Analysis

## Implementing data-driven techniques to predict air quality index and inform public health decisions

### Restatement of the Questions

Two primary questions drive the project:

1. Can the air quality trends from one year be used to predict the air quality of the subsequent year in Bangalore?
2. Which machine learning model is the best in predicting AQI based on historical data?

These questions are essential for understanding the persistence of air pollution and determining the most accurate approach to forecasting future air quality, crucial for public health decisions and regulatory actions.

### Broad Conclusions from Analysis and Modeling

Analyzing datasets from Tutiempo (Tutiempo 2024) and Weather Map (Weather Map 2024), the study focused on the relationship between various environmental factors and PM2.5 levels. During preprocessing, meteorological data from HTML files were merged with AQI data from CSV files (2013-2018) to create a unified dataset. Hourly PM2.5 values (2013-2018) were averaged (Air Now n.d.) daily and monthly, and unnecessary variables were excluded. Exploratory Data Analysis identified significant correlations, particularly between temperature-related features and PM2.5 levels, suggesting that temperature is a key factor in air quality.

Several machine learning models were tested, including Linear Regression, Ridge Regression, Lasso Regression, Decision Tree Regressor, and XGBoost Regressor. Among these, the XGBoost Regressor emerged as the most effective, with an $R^2$ score (Chugh 2020) of 0.7211 on the test set. This result indicates that air quality trends from one year can indeed influence the next, and that the XGBoost model is particularly well-suited for predicting AQI, validating the potential of historical data for accurate forecasts.

### Data Visualization and Interpretation

To support the analysis and model outcomes, the following visualizations were created:

1. **Correlation Heatmap**: This heatmap visualized the correlations between environmental variables and PM2.5 levels, highlighting the strong relationship with temperature-related features and guiding feature selection for modeling.

2. **Time Series Plots**: These plots depicted trends in PM2.5 levels over the years, revealing seasonal patterns and the persistence of high pollution levels.

3. **Model Performance Comparison**: A bar chart compared the $R^2$ scores of different models, emphasizing the superior performance of the XGBoost Regressor (Kharwal 2023).

4. **Feature Importance Plot (XGBoost)**: This plot illustrated the features most influential in predicting AQI, with temperature-related variables ranking high.

5. **Residual Plots**: Residual plots for the XGBoost model confirmed a good fit, with residuals uniformly distributed around zero.

These visualizations were crucial for interpreting data and validating the models, demonstrating the effectiveness of XGBoost Regression in predicting AQI.

## Specific Actions to Improve Air Quality

Based on the analysis and modeling, several specific actions are recommended:

1. Implementation of Real-time AQI Monitoring and Prediction Systems:

- Develop and deploy an advanced AQI monitoring system that integrates real-time data with machine learning models like XGBoost to provide accurate forecasts.
- Use the predictive model to issue timely warnings and advisories, especially during anticipated periods of poor air quality.

2. Temperature Control Measures:

- Implement urban planning strategies to mitigate temperature fluctuations, such as increasing vegetation cover (Clean Air Fund 2023) and reflective surfaces, which can reduce PM2.5 concentrations.
- Regulate industrial emissions more strictly during high-temperature periods, as these correlate with increased PM2.5 levels.

3. Policy Formulation Based on Predictive Insights:

- Formulate regulations aimed at reducing emissions (Clean Air Fund 2023) during predicted high-risk periods based on model insights.
- Enforce stricter emission controls on industries and vehicles during these periods to prevent air quality deterioration.

4. Public Awareness and Education:

- Conduct public awareness campaigns about the health impacts of poor air quality (Clean Air Fund 2023), particularly during predicted high-risk periods.
- Promote the use of masks, indoor air purifiers, and reduced outdoor activities when poor air quality is forecasted.

5. Continuous Model Refinement and Data Collection:

- Continue collecting and incorporating more recent data to refine predictions and improve accuracy.
- Explore integrating additional environmental variables, such as humidity and wind speed, to further enhance the model's predictive capabilities.

## Conclusion and Future Work

This project demonstrates the potential of advanced machine learning models, particularly XGBoost Regressor, to predict AQI based on historical data. The findings suggest that air quality trends are influenced by prior years, and these trends can be effectively forecasted to aid in public health decision-making.

To improve the system under study, it is recommended that the actions mentioned above be implemented, focusing on real-time monitoring, temperature control, and policy-driven interventions. Future work should involve exploring more sophisticated models and incorporating additional environmental variables to enhance predictive accuracy further. Expanding the dataset to include more recent and diverse data will also ensure that the predictive models remain relevant and effective in changing environmental conditions.

## References


Tutiempo 2024, World Weather, Tutiempo 2024, viewed 7 June 2024, <https://en.tutiempo.net/>

Air Now n.d., Air Quality Index (AQI) basics, viewed 7 June 2024, <https://www.airnow.gov/aqi/aqi-basics/#:~:text=Think%20of%20the%20AQI%20as,300%20represents%20hazardous%20air%20quality.>

Akshita Chugh 2020, *MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better?*, Medium, viewed 25 July 2024, <https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e>

Aman Kharwal 2023, *Air Quality Index Analysis Using Python*, The Clever Programmer, viewed 7 June 2024, <https://thecleverprogrammer.com/2023/09/18/air-quality-index-analysis-using-python/>

Clean Air Fund 2023, *5 ways cities are cleaning the air we breathe,* Clean Air Fund, viewed 13 August 2024, <https://www.cleanairfund.org/news-item/5-ways-cities-are-cleaning-the-air-we-breathe/>