Predicting PM2.5 In Bangkok Thailand With Machine Learning

Jirapon Kluaymaingarm†
Information Management
Asian Institute of Technology
Bangkok, Thailand
st124845@ait.ac.th

Napassorn Sripum
Information Management
Asian Institute of Technology
Bangkok, Thailand
st124949@ait.ac.th

Patsakorn Tangkachaiyanunt
Information Management
Asian Institute of Technology
Bangkok, Thailand
st124876@ait.ac.th

Aye Nyein Thu
Information Management
Asian Institute of Technology
Bangkok, Thailand
st124957@ait.ac.th

ABSTRACT

This study investigates the effectiveness of various machine learning models in predicting PM2.5 levels across Bangkok's districts over a seven-day period. The models incorporate environmental and meteorological factors, including pollutant concentrations (PM2.5, NO2, CO) and meteorological variables (temperature, humidity, and sea level), which were retrieved from OpenWeather API. Additionally, it incorporates demographic factors such as traffic levels, the number of factories, and the green space ration, population, household. To determine the most accurate approach, we compare multiple regression-based machine learning models, including **Linear Regression, Ridge Regression, Lasso Regression, Decision Trees, Random Forest and XGBoost.** Results indicate that integrating temporal and environmental features significantly enhances predictive performance, with the **Random Forest** achieving the highest R² score, lowest Mean Absolute Error (MAE), lowest Mean Square Error (MSE)and lowest Root Mean Square Error (RMSE) score in forecasting PM2.5 levels. These findings highlight the potential of machine learning for air quality prediction and suggest that a grid-based mapping system could improve public awareness and support data-driven policymaking.

Keywords

PM2.5, Air Pollution, AQI, Machine Learning, GSI, Deep Learning, Regression

1. INTRODUCTION

Air pollution remains a pressing global concern, particularly in urban environments where high levels of fine particulate matter (PM2.5) pose severe risks to public health, climate stability, and overall quality of life. Chronic exposure to PM2.5 has been linked to respiratory diseases, cardiovascular conditions, and premature mortality, with vulnerable populations such as children and the elderly being disproportionately affected. In addition to health implications, poor air quality contributes to environmental degradation, reduced visibility, and economic losses, impacting urban planning and sustainability efforts. Therefore, accurate PM2.5 prediction models are essential for mitigating exposure risks, enabling timely interventions, and informing policy decisions aimed at improving air quality management.

However, despite advancements in machine learning (ML)-based air quality forecasting, several key limitations persist. Existing research has largely focused on deep learning approaches, often prioritizing predictive accuracy over interpretability. Additionally, many models fail to incorporate spatial and environmental factors, such as land use patterns, elevation, and pollution sources, which are essential for accurately modeling PM2.5 dispersion. Furthermore, while extensive studies have explored air pollution forecasting in cities like Beijing, Delhi, and Shanghai, research on Bangkok's air quality remains scarce, despite the city experiencing severe seasonal pollution due to traffic congestion, industrial emissions, and meteorological conditions.

illustrate 1: Bad Air quality in Bangkok, Thailand

Despite significant progress in air quality monitoring, real-time and localized PM2.5 forecasting remains a challenge, particularly in rapidly developing cities such as Bangkok. The city experiences seasonal air pollution surges due to a combination of vehicular emissions, industrial activity, and meteorological conditions. However, existing predictive models often rely solely on historical pollution data or limited meteorological factors, neglecting crucial Geographic and Spatial Information (GSI) such as land use patterns, elevation, and proximity to pollution sources. The lack of comprehensive models that incorporate spatial, temporal, and environmental factors reduces the reliability of AQI forecasting, limiting its effectiveness in real-world applications.

illustrate 2: People wearing mask due to bad air quality

Public awareness of PM2.5 pollution has increased significantly in recent years, particularly in highly affected cities like Bangkok. Citizens frequently express concerns over health risks, such as respiratory diseases and long-term cardiovascular effects, especially during peak pollution seasons. Many rely on mobile air quality apps, government alerts, and news reports to determine whether it is safe to engage in outdoor activities. However, due to inconsistent and sometimes delayed PM2.5 updates, people often feel uncertain about how to protect themselves from harmful exposure. A study by The Nation in 2023[12] found that a significant portion of urban residents feel that current PM2.5 predictions lack accuracy and fail to reflect real-time pollution conditions. This underscores the growing demand for more reliable and localized PM2.5 forecasting models that provide actionable insights for both policymakers and the general public.

The contributions of this work are:

I. **Comprehensive Comparative Analysis** – We systematically evaluate multiple machine learning models, including tree-based, regression, and deep learning approaches, to identify the most effective method for AQI prediction, addressing the current overreliance on deep learning.

II. **Integration of Geographic and Environmental Factors** – Unlike many existing studies that focus solely on historical air quality and meteorological data, our model incorporates geospatial (GSI), environmental, and demographic factors, enhancing prediction accuracy and real-world applicability.

III. **Bangkok-Specific AQI Forecasting** – While extensive research exists for cities like Beijing and Delhi, Bangkok remains underrepresented in AQI prediction studies. This research fills that gap by providing a localized, data-driven approach tailored to Bangkok's unique pollution dynamics.

IV. **Identification of Key Predictors** – We analyze and rank the most influential variables affecting PM2.5 concentrations, providing actionable insights for policymakers and urban planners to design targeted interventions for air quality improvement.

1.1 Related Work

Air pollution prediction and mitigation have been extensively studied using machine learning, chemical modeling, and policy analysis. Previous research has demonstrated that PM2.5 and ozone pollution significantly impact human health and economic stability, particularly in urban environments such as Japan, where ozone concentrations pose a long-term challenge (Yin Long et al., 2023). Studies have shown that end-of-pipe (EoP) technologies and electrification can reduce morbidity and prevent thousands of premature deaths, while economic losses due to air pollution have been quantified, with mitigation measures reducing economic losses by up to 141 billion USD. While studies on Japan provide valuable insights into the effectiveness of mitigation strategies, air quality forecasting in Southeast Asian cities like Bangkok remains underexplored. Given Bangkok's severe seasonal pollution, this research builds upon existing work by incorporating Geographic and Spatial Information (GSI) alongside pollutant and meteorological data to enhance AQI prediction accuracy. Additionally, by comparing multiple machine learning models, this study addresses limitations in prior research that focused primarily on deep learning approaches without interpretability.

Previous research has extensively explored machine learning models for PM2.5 prediction, particularly in South Asia, where air pollution is a significant environmental concern. Studies conducted in India and Pakistan have compared statistical models (ARIMA, SARIMA) and deep learning approaches (LSTM, GRU, Random Forest Regression), demonstrating that deep learning models often outperform traditional methods in capturing temporal dependencies in air pollution data (Ganguli et al., 2025; Mohammadi et al., 2024). Similarly, research in Nepal and Bangladesh has developed spatiotemporal PM2.5 prediction models, addressing the challenges posed by sparse air quality monitoring networks (Shuvo et al., 2021). Despite these advancements, most studies remain geographically limited, with models trained on specific urban areas, reducing their applicability to regions with different environmental and demographic characteristics.

The Role of Geographic and Spatial Information (GSI) in Air Quality Prediction

Geographic and Spatial Information (GSI) plays a crucial role in PM2.5 prediction by providing geospatial context to air quality data. Since air pollution is highly dependent on location-based factors, incorporating GSI allows for a more accurate and comprehensive analysis of PM2.5 distribution. Key geospatial variables, such as elevation, land use, traffic density, and meteorological conditions (temperature, humidity, and wind patterns), directly influence pollution levels and dispersion.

Additionally, satellite-derived data, including Aerosol Optical Depth (AOD) from MODIS and Sentinel-5P, enhance the predictive capability of machine learning models by providing near real-time atmospheric pollution insights. Prior research has shown that integrating spatial interpolation, GIS-based feature engineering, and deep learning with remote sensing data can improve PM2.5 prediction accuracy, benefiting urban planning, environmental policy, and public health management.

Several studies have attempted to integrate GSI into air quality prediction models to enhance accuracy and applicability. Di et al. (2016) used satellite-derived AOD data alongside meteorological and land use variables to estimate ground-level PM2.5 concentrations in the Southeastern United States, demonstrating the role of GSI in improving spatial resolution and predictive reliability. Similarly, Lin & Li (2021) developed an LSTM-based air quality prediction model that incorporated land-use patterns, traffic density, and meteorological conditions, leading to improved forecasting accuracy. Alam & Qazi (2023) extended this approach by applying hyperparameter optimization techniques to improve machine learning-based AQI forecasting, reinforcing the importance of geospatial data in air quality modeling. More recently, Kermani et al. (2024) utilized AOD, meteorological data, and land-use information to refine PM2.5 predictions, showing that geospatial factors remain critical for air quality forecasting.

Despite these advancements, existing studies often lack integration with real-time spatial visualization tools, limiting their usability for policymakers and public health officials. Sookchaiya & Phimoltares (2022) compared various machine learning models for predicting PM2.5 and PM10 levels in Chiang Mai Province but did not integrate their findings with spatial visualization tools for decision-making. Similarly, Zhang & Wang (2022) explored air quality prediction using Random Forest models in China, prioritizing model accuracy over practical implementation through geospatial mapping. Kumar & Singh (2023) expanded on this by conducting a comparative study of machine learning models for air quality prediction in smart cities, emphasizing the need for interactive, grid-based AQI mapping systems to aid decision-making for urban planners and policymakers.

Findings and best practices from Japan's Air Quality Research

Japan has conducted extensive and in-depth research on PM2.5, making it one of the few industrialized nations that has successfully reduced air pollution through strict regulatory measures and advanced monitoring techniques(Kunugi et al., 2018).

Extensive research has analyzed PM2.5 trends, chemical composition, and the effectiveness of air pollution control measures in Japan. Yamagami et al. (2021) studied long-term PM2.5 trends from 2003 to 2018 in Nagoya City, revealing a 53% reduction in PM2.5 levels due to stringent diesel emission regulations and changes in secondary aerosol formation. However, their findings also emphasized the impact of transboundary pollution from China, particularly in sulfate (SO_4^{2-}) and nitrate (NO_3^-) concentrations. Similarly, Ito et al. (2021) provided a comprehensive 30-year air quality analysis in Japan, highlighting a 26–30% decline in PM2.5 and 55–65% reduction in suspended particulate matter (SPM) mainly due to emission regulation, yet noting that ozone (O_3) levels continue to rise due to secondary particle formation and transboundary pollution.

While these studies provide valuable insights into air pollution mitigation strategies, they primarily focus on long-term trend analyses rather than real-time air quality forecasting. In contrast, Bangkok experiences severe seasonal pollution due to vehicle emissions, industrial activity, and meteorological factors, yet localized AQI prediction models remain limited. Most existing studies on Bangkok have prioritized deep learning for AQI forecasting, often sacrificing interpretability and the integration of spatial and demographic factors.

Real API Data use in Previous Research

In addition to the extensive research conducted in Japan, other regions have also developed innovative approaches to air quality monitoring. For example, a study by Geng et al. (2021) introduced the Tracking Air Pollution in China (TAP) database, which provides near real-time, full-coverage $PM_{2.5}$ data at a 10 km spatial resolution across China since 2000. This database integrates multiple data sources, including ground measurements, satellite retrievals, emission inventories, and chemical transport model simulations, to offer timely updates and historical records of air pollution levels. The TAP $PM_{2.5}$ data is estimated using a two-stage machine learning model (random forests and gradient boosting) coupled with gap-filling methods, achieving an average cross-validation R^2 of 0.83. Such comprehensive datasets are essential for supporting research and environmental management efforts.

Similarly, in the United States, Mendez et al. (2022) utilized low-cost sensors to assess $PM_{2.5}$ concentrations in four South Texan cities along the U.S.–Mexico border. This year-long study demonstrated the feasibility of using affordable sensing technology for localized air quality monitoring, providing valuable data for public health assessments and policy-making. These studies highlight the growing trend of employing advanced technologies and data integration methods to enhance air quality monitoring and prediction, offering valuable insights for other regions, where such comprehensive approaches are yet to be fully implemented.

Contributions of This Study

This study aims to bridge these gaps by developing a machine learning-based AQI prediction model for Bangkok, integrating GSI, environmental variables, and demographic factors. By comparing multiple machine learning approaches, including Linear Regression, Random Forest, Decision Trees, and XGBoost, this study seeks to identify the most effective predictive method for urban air quality management. Additionally, while prior research has focused on long-term pollution trends and regulatory impacts, this study emphasizes real-time forecasting and public awareness applications. Given the growing demand for accurate and timely air quality predictions, this research integrates both temporal and spatial features to enhance predictive reliability, ensuring that data-driven policymaking and public health interventions are more effective in rapidly developing urban environments.

2. Research Question

This study aims to investigate the effectiveness of machine learning models in predicting AQI levels across Bangkok's districts. Specifically, it addresses the following research questions:

1) Which model predicts AQI the most accurately and precisely?
2) What are the features that influence the results of AQI prediction the most?
3) How accurately does the general public interpret and use data visualizations to improve their awareness of health issues?

3. Data

For this study, data were obtained from OpenWeather[5], an open-source API that provides real-time and historical environmental data. The dataset includes air quality indicators from January 2024, such as $PM_{2.5}$, CO, and NO_2, along with meteorological variables like latitude, longitude, temperature, and relative humidity. Additionally, it incorporates demographic factors such as traffic levels to assess the impact of vehicle emissions on $PM_{2.5}$ concentrations, with higher traffic density potentially contributing to increased pollution levels.

Sea level data were obtained from topographic maps [20] and Research Gap [21]. Variations in sea level, such as areas at high elevations or below sea level, may potentially influence $PM_{2.5}$ concentrations by affecting atmospheric circulation, pollutant dispersion, and local meteorological conditions.

Factory data were sourced from The Department of City Planning (BMA CPUD) [23], using records from 2020 due to the unavailability of more recent data. Variations in the number of factories, such as high-density industrial areas versus low-density regions, may potentially influence $PM_{2.5}$ concentrations through emissions and localized air pollution levels.

Population statistics, including total population, population density, and green space ratio, were retrieved from the strategy and evaluation department [22], based on 2024 data. Population density and green space availability may influence $PM_{2.5}$ concentrations, as higher population density is often associated with increased emissions, while greater green space can help mitigate air pollution.

illustrate 3: OpenWeather API

สำนักการวางผังและพัฒนาเมือง
CITY PLANNING AND URBAN DEVELOPMENT DEPARTMENT

illustrate 4: The Department of City Planning (BMA CPUD)

กทม
สำนักยุทธศาสตร์และประเมินผล
STRATEGY AND EVALUATION DEPARTMENT

illustrate 5: Strategy and evaluation department

4. Methodology

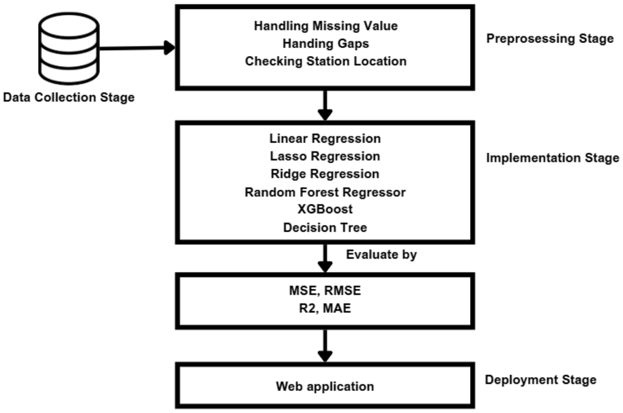


Figure 1 Diagram explain the whole methodology

The methodology adopted in this research comprises three principal components: data collection, preprocessing, implementation and deployment. The abstract representation of this methodology is illustrated in Figure 1.

4.1 Data collection Stage

To construct an accurate predictive model for PM2.5 prediction, data titled "Time Series Air Quality Data of Bangkok per Hour" covering the period from January 1, 2025, to present data was acquired from OpenWeather, which provides real-time and historical environmental data. This dataset comprises hourly measurements recorded across **7** monitoring stations distributed throughout Bangkok, Thailand as seen in table 2. The dataset encompasses a total of **14,841** rows, each containing **19** distinct features. This extensive dataset provides a comprehensive representation of the PM2.5 and associated influencing fact

Variable	Explanation	Unit of measure
PM2.5	Particulate matter with an aerodynamic diameter less than or equal to a nominal 2.5 micrometers	μg/m3
Temperature	The measure of air temperature in the atmosphere	Celcius (°C)
RH	Relative Humidity – the percentage of moisture in the air	Percent (%)
NO2	Nitrogen Dioxide – a major air pollutant primarily from vehicle emissions	Parts per billion (ppb)
CO	Carbon Monoxide – a toxic gas primarily from incomplete combustion	Parts per million (ppm)
Latitude	Angle between the straight line in the	Degrees

Variable	Explanation	Unit of measure
	certain point and the equatorial plane	
Longitude	Angle pointing west or east from the Greenwich Meridian	Degrees
Sea level	An average surface level of one or more among Earth's coastal bodies of water	meter (m)
Traffic Level	Traffic scale level with in that area	1-10
Number of Factory	Number of factory with in that area	Factory (units)
Factory Area	Area of the given factory space	Square metre (m^2)
Green Space	Density of greenspace per population	Square metre per person
Month	12 Different months in a year	Months (mo.)
Day	Seven Different day in a week	Days (d)
Hour	A unit of time in a day	Hour (h)
Population	Total number of people living in a given area.	People (person)
population density	Number of people per unit area, typically measured to understand urbanization levels.	square meter per people (m^2/people)
household	Total number of households in the given area, where a household refers to individuals living together in one dwelling unit.	Households (units)
greenspace	The total area covered by parks, forests, or other vegetated land within a given area.	square meter per people (m^2/population)
season	The weather patterns and temperature variations throughout the year.	Hot/ Rainy/ Cool season.

Table 1 Notation with the explantation and relative unit of measures

Data Dictionary

Below are some of the data dictionary values incorporated into our code.

I. Traffic levels are provided on a scale of either 1-10 or 1-4, representing the congestion level within the area surrounding the weather station #base on which sources

II. Days of the week were converted into integer values from 0 to 6, where 0 represents Monday and 6 represents Sunday.

III. We Group month into three different seasons, Rainy Season (May to October), Cool Season (November to February) and Summer (March to April)

4.2 Data Preprocessing Stage

During the preprocessing stage, various operations were systematically performed to prepare the data for predictive modeling. These steps include:

- **Acquire Data**: Since the data were not uploaded via a CSV file, we retrieved them through an API and converted them into a dataframe for further processing.
- **Merging Data:** We integrated datasets originating from disparate sources, creating a comprehensive and cohesive dataset. This integration is critical to maintaining uniformity in data handling and analysis, thereby enhancing the reliability of our analytical outcomes.
- **Data Mapping:** To further enhance the dataset's utility, we mapped each season to its corresponding months. This mapping process aids in clarifying seasonal trends and their impacts on the dataset, providing a structured framework for seasonal analysis.
- **Data Encoding:** To facilitate statistical analysis, we converted all categorical string data related to seasons, which were defined in the mapping section, into numerical formats.
- **Data Extraction:** We utilized the dt accessor to extract critical time-based elements such as the day of the week, month, and hour from the 'datetime' column. This extraction process is essential for conducting a deeper temporal analysis and yields insights that are pivotal for robust time series evaluations. These elements enable a granular examination of temporal patterns and their effects on the dataset.
- **Handling Missing Values**: To ensure data completeness, missing PM2.5 values were first interpolated using a linear interpolation method within each monitoring station, preserving local trends. Any remaining missing values were then filled using a backward fill (bfill) approach, ensuring that no gaps remained in the dataset. Additionally, the data were sorted by latitude, longitude, and timestamp to maintain temporal consistency before interpolation.
- **Handling Gaps**: To ensure a complete and continuous dataset, a full range of timestamps was generated using hourly intervals from the minimum to the maximum recorded datetime in the dataset. Next, of all possible station locations (latitude, longitude) and timestamps was created to form a comprehensive reference grid. To align the existing dataset with this grid, each station's latitude and longitude were combined into tuples for efficient merging. The dataset was then merged with the full datetime-station grid, filling in missing timestamps for each station. Finally, the latitude and longitude values were extracted from the merged tuples, and redundant columns were dropped to maintain dataset integrity.
- **Checking Station Location**: Since we acquired data from API sources, we need to figure out which station the data were pulled from to match with the exact GSI (longitude and latitude).

Following these preprocessing steps, the refined data are then advanced to the prediction module, where they are used to train machine learning models.

Station	Location
3T	Ratchathewi Station
10T	Ladkrabang Station
11T	Taling Chan Station
12T	Bang Sue Station
15T	Bang Khae Station
54T	Pathum Wan Station
61T	Bang Na Station

Table 2: 7 Monitoring Station

4.3 Implementation Stage

This research utilized several specialized Python libraries, notably Statsmodels and Scikit-learn, to perform data analysis and model training. All computational experiments were executed on a workstation equipped with an Intel Corei7 13700HX CPU, RTX 4060 laptop 8GB VRAM GPU, 24 GB RAM.

The dataset used in this research consisted of X entries, which were randomly divided into training and validation sets. Specifically, 80% of the data (X samples) were allocated for training the models, while the remaining 20% (X samples) constituted the validation set used for model testing. This study aimed to predict air quality indicators by leveraging variables listed in Table 1.

The training phase involved the application of various regression machine learning algorithms, **Linear Regression, Ridge Regression, Lasso Regression, Decision Trees, Random Forest and XGBoost.** These models were configured to optimize performance metrics appropriate to the type of model trained.

To assess the regression models, evaluation metrics such as the R^2 score, Mean Absolute Error (MAE), and Mean Squared Error (MSE) were employed.

4.4 Deployment Stage

To improve public access to air quality information, we developed a Flask-based web application that provides seven-day air quality forecasts using the X model. The application features a user-friendly interface, allowing users to view predicted $PM_{2.5}$ levels by selecting a specific day, time, and region. The model is trained on collected air quality data from various monitoring stations in Bangkok (as shown in Figure X) to generate predictions. This setup enables users to interact with the predictive model effectively, facilitating informed decision-making about air quality.

5. Method

Previous studies on PM2.5 prediction have predominantly utilized time-series deep learning models such as LSTM and GRU or statistical methods like ARIMA and SARIMA (Ganguli et al., 2025; Mohammadi et al., 2024). While these models have proven effective for temporal pattern recognition, comparative evaluations of regression-based approaches remain underexplored. Regression models, particularly polynomial, ridge, lasso, and elastic net regression, offer advantages in interpretability, feature selection, and computational efficiency, making them suitable for air quality forecasting. However, most prior studies have overlooked these methods, limiting insights into their effectiveness for PM2.5 prediction. To address this gap, our study systematically evaluates various regression techniques, assessing their performance based on accuracy, generalization, and model complexity.

Selected Model :

To develop a robust PM2.5 prediction model, we evaluate multiple regression-based machine learning algorithms, considering both parametric and non parametric regression techniques. The selected models include:

- **Linear Regression** - A fundamental regression approach for capturing linear relationships in AQI data.
- **Ridge Regression** - A regularized regression technique that introduces an L2 penalty to reduce overfitting and improve model generalization, especially when dealing with multicollinearity.
- **Lasso Regression** - A regression method that applies an L1 penalty, promoting feature selection by shrinking less important coefficients to zero, making it useful for high-dimensional data.
- **Random Forest** - An ensemble learning method that builds multiple decision trees and combines their outputs to improve accuracy and reduce overfitting.
- **Decision Trees** - A non-linear model that splits data hierarchically based on feature importance, making it interpretable and effective for capturing complex relationships.
- **XGBoost** - A gradient boosting algorithm that enhances predictive performance by iteratively improving weak learners and minimizing errors through optimization techniques.

By systematically comparing these regression models, we aim to identify the most effective technique for PM2.5 forecasting, ensuring a balance between model complexity, interpretability, and predictive accuracy.

Evaluation Metric :

To assess the performance of our PM2.5 prediction models, we employ multiple evaluation metrics covering regression metrics.

- **R^2** - Measures how well the independent variables explain the variance in the dependent variable. A higher R^2 value (closer to 1) indicates better model performance
- **Mean Square Error (MSE)** - Evaluates the average squared differences between actual and predicted values. Lower MSE values indicate better performance.
- **Root Mean absolute error (MAE)** - Measures the average absolute differences between actual and predicted values, providing a more interpretable measure of error.
- **Root Mean Square error (RMSE)** - Computes the square root of the average squared differences between actual and predicted values, penalizing larger errors more heavily than MAE. A lower RMSE indicates better predictive accuracy.

$$\mathbf{R^2} = 1 - \frac{\Sigma(yi-\widehat{yi})^2}{\Sigma(yi-\bar{y})^2}$$

where,

- yi = actual value
- $\widehat{yi}$ = predicted value
- $\bar{y}$= mean of actual values

$$\mathbf{MSE} = \frac{1}{n}\sum_{i=1}^{n}(yi - \widehat{yi})^2$$

where,

- n = number of observations
- yi = actual value
- $\widehat{yi}$ = predicted value

$$\mathbf{MAE} = \frac{1}{n}\sum_{i=1}^{n}|yi - \widehat{yi}|$$

where,

- yi = actual value
- $\widehat{yi}$ = predicted value
- n = number of observations

$$\mathbf{RMSE} = \sqrt{\frac{\Sigma(yi-\widehat{yi})^2}{n}}$$

where,

- yi = actual value
- $\widehat{yi}$ = predicted value
- n = number of observations

6. Result and Discussion

6.1 Variable importance

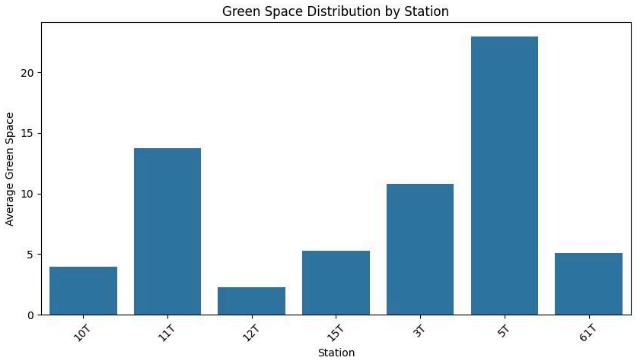
The analysis of variable importance for the X types of prediction models, conducted using ‘station’, ‘pm2_5', 'hour', 'month', 'day_of_week', 'population', 'population_density', 'household', 'green_space', 'factory_num', 'factory_area’, ‘season’ and ‘time_step’ offers crucial insights into the factors influencing PM2.5 predictions. The significance of each feature varies with the model configuration, as illustrated in visualization 7 through headmap correlation.

Notably, for hourly predictions, X is identified as the most crucial feature, highlighting its substantial impact on the accuracy of AQI values forecasted for the next seven days. This insight emphasizes the importance of considering X in predictive models to enhance their predictive reliability and accuracy.

6.2 Visualization and Analysis

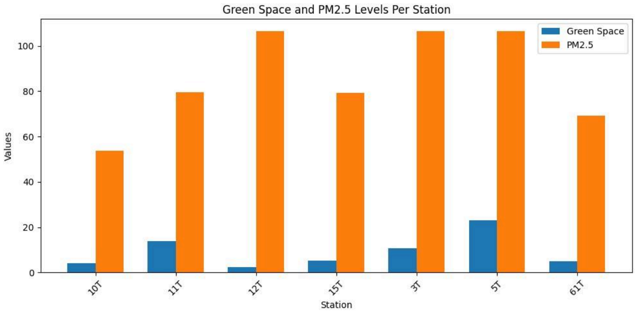
To gain deeper insights into the relationships between environmental, demographic, and industrial factors influencing $PM_{2.5}$ levels, various visualization techniques are employed. These visualizations help in identifying trends, correlations, and distribution patterns across different variables such as population density, factory numbers, green space, and PM2.5 levels. Key techniques include heatmaps for correlation analysis, bar charts for categorical comparisons, boxplots for distribution insights, and line graphs for time-series trends. By leveraging these visual tools, we can better interpret the underlying patterns in the data and assess their impact on PM2.5 on Bangkok.

Graphs may change dynamically as the data retrieved from the API updates each time the model is run. This were the result after the data was visualized



Visualization 1: Green space Distribution by Station

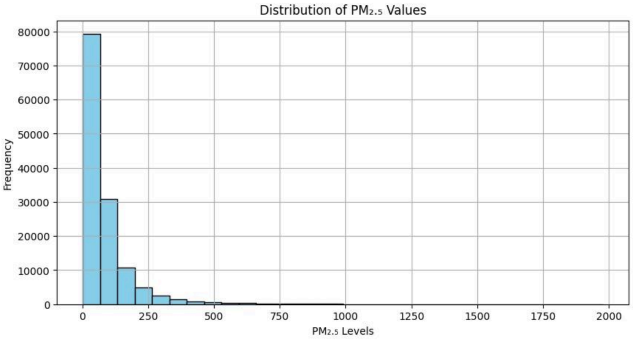
Pathum Wan Station (5T) has the highest green space availability, indicating that this area likely contains significant parks, vegetation, or open spaces. Taling Chan Station (11T) also has a relatively high level of green space, suggesting a well-vegetated environment. In contrast, Bang Sue Station (12T) has the lowest green space, likely due to urbanization and industrial development. Bang Khae Station (15T) and Bang Na Station (61T) have moderate levels of green space, while Ratchathewi Station (3T) and Ladkrabang Station (10T) show variations that may be influenced by urban planning and land use. This variation in greenery across different transit stations may have implications for air quality, urban heat island effects, and environmental planning. Areas with lower green space may experience higher pollution levels, while stations with greater green space could benefit from improved air quality and environmental sustainability.



Visualization 8: Green space and PM2.5 level per station

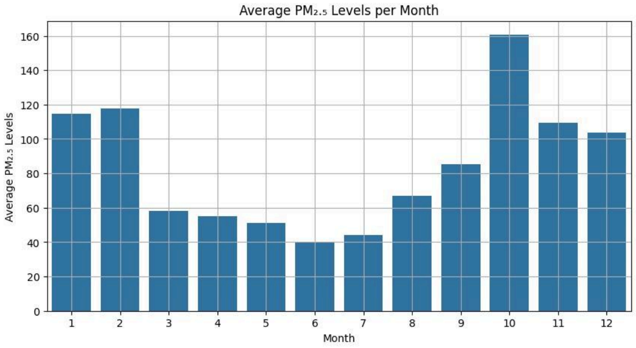
Pathum Wan Station (5T) has the highest green space availability, yet it also exhibits one of the highest $PM_{2.5}$ levels, suggesting that while vegetation may help mitigate pollution, other factors such as traffic density and industrial emissions likely contribute significantly to air quality. On the other hand, Taling Chan Station (11T), which also has substantial green space, shows moderately lower $PM_{2.5}$ levels, reinforcing the idea that vegetation may play a role in reducing air pollution. Similarly, Bang Khae Station (15T) and Bang Na Station (61T) follow a comparable trend, where increased green space corresponds to relatively lower $PM_{2.5}$ levels. However, there are exceptions to this pattern. Ratchathewi Station (3T), despite having moderate green space, still experiences high $PM_{2.5}$ levels, indicating that other environmental and human activity factors—such as vehicle emissions, industrial activity, and wind patterns—may also influence pollution levels. Additionally, Bang Sue Station (12T) has the lowest green space coverage and one of the highest $PM_{2.5}$ levels, suggesting that areas with limited vegetation may be more vulnerable to poor air quality. This trend highlights that while green space can contribute to improved air quality, it is not the sole determining factor of $PM_{2.5}$ concentrations

This pattern aligns with population density, as Bang Sue Station (12T) has the highest population density with 10,275, which correlates with its high $PM_{2.5}$ levels. Similarly, Ratchathewi Station (3T), which has the second-highest population density with 9,046, also exhibits elevated $PM_{2.5}$ levels, reinforcing the link between urban density and air pollution. Following this trend, Pathum Wan Station (5T) ranks third in population density of 4,788 and $PM_{2.5}$ levels, further supporting the idea that denser areas experience higher pollution due to increased human activity, vehicle emissions, and industrial operations.



Visualization 2: Distribution of PM2.5 values

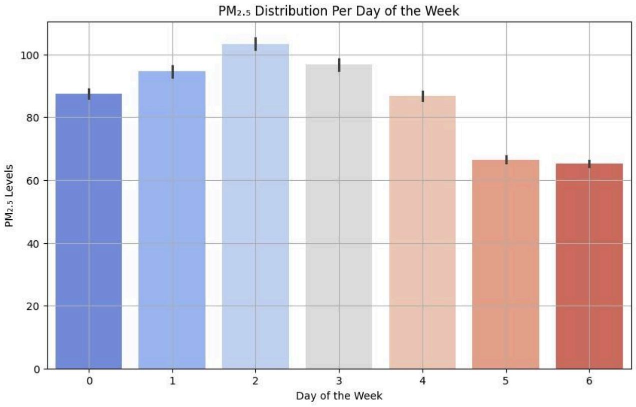
Most $PM_{2.5}$ values are concentrated at lower levels, with the highest frequency occurring below 100 μg/m3. The distribution exhibits a right-skewed pattern meaning that while the majority of $PM_{2.5}$ values are relatively low, there are some instances of extreme pollution levels, though they occur less frequently. A sharp drop-off in frequency is observed as $PM_{2.5}$ values increase, suggesting that extremely high pollution events are rare but still present in Bangkok.



Visualization 3: Average PM2.5 level per months

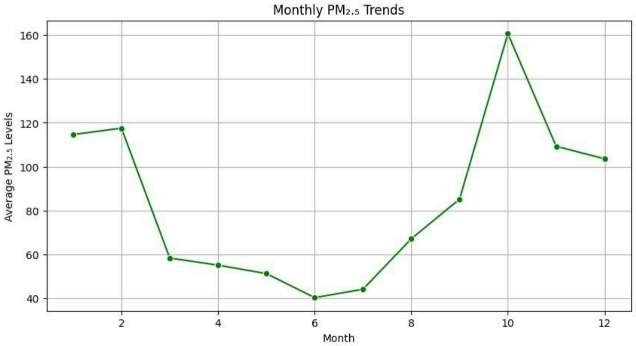
The bar chart displays the average $PM_{2.5}$ levels per month, highlighting significant seasonal variations in air pollution. The highest $PM_{2.5}$ levels occur in October, followed by February, January November and December, suggesting that pollution is most severe during 5 months stretch. This pattern may be influenced by cool seasonal weather conditions, such as temperature inversion, lower wind speeds, and reduced atmospheric dispersion, as well as increased human activities like crop burning, industrial production, and heating emissions. In contrast, March through July show the lowest $PM_{2.5}$ concentrations, likely due to warmer temperatures, stronger atmospheric mixing, and increased rainfall, which help remove airborne pollutants. However, from August onwards, $PM_{2.5}$ levels start rising again, peaking sharply in October, before slightly decreasing in November and December. This trend suggests a strong seasonal influence on air quality, where cooler months contribute to higher pollution levels, while warmer months promote better air dispersion and pollutant removal.

#October spike due to close breeze from north east (find source)



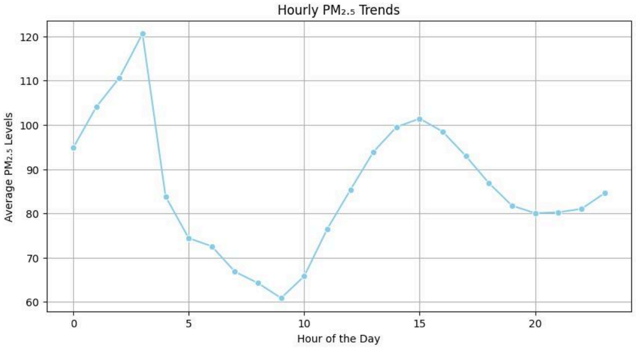
Visualization 4:

Next, we analyze $PM_{2.5}$ levels per day of the week. The bar chart illustrates the distribution of $PM_{2.5}$ levels across different days, where 0 represents Monday and 6 represents Sunday. The data reveals that $PM_{2.5}$ levels peak on Tuesday (2) and Wednesday (3), followed closely by Monday (0) and Thursday (4). In contrast, Saturday (5) and Sunday (6) show the lowest $PM_{2.5}$ levels, indicating a decline in air pollution during weekends. This trend likely reflects weekday traffic congestion, industrial activity, and urban emissions, which contribute to higher $PM_{2.5}$ concentrations during working days. The drop in pollution levels on weekends may be attributed to reduced commuting, lower industrial operations, and decreased human activity overall. Additionally, shifting lifestyle patterns post-pandemic and economic changes may have influenced this trend, as more people stay home rather than engage in shopping or leisure activities. The gradual decline from Monday to Sunday supports the idea that air pollution is strongly linked to workweek patterns and human mobility, emphasizing the role of traffic and industrial activity in shaping urban air quality.



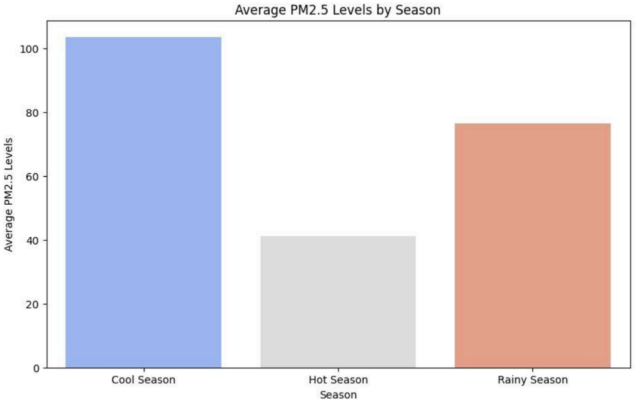
Visualization 5: PM2.5 Trend per Month

Looking the the trend of above, $PM_{2.5}$ levels appear to be relatively high at the beginning of the year, peaking around January and February, before experiencing a significant decline from March to June. During this period, the levels remain relatively stable and low, suggesting that seasonal conditions such as increased rainfall or improved air circulation might be helping to reduce pollution levels. However, from July onwards, $PM_{2.5}$ levels start to rise again, with a notable sharp increase in October, reaching the highest pollution level of the year. After this peak, the levels drop slightly in November and December but remain higher than mid-year values. The spike in October could be attributed to seasonal weather patterns, increased industrial activity, or agricultural burning, which is common in some regions during this time. The lower $PM_{2.5}$ levels between March and June might be associated with seasonal monsoons, which help clear particulate matter from the air.



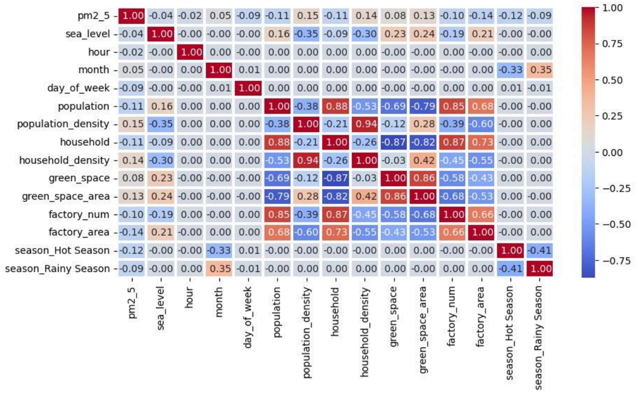
Visualization 6: PM2.5 Trend per Hour

$PM_{2.5}$ levels are highest during the early hours of the day, peeking around midnight to 2 AM, with values exceeding 120 μg/m3. After this peak, there is a sharp decline between 3 AM and 6 AM, reaching the lowest levels of the day around 8-9 AM. This drop could be attributed to decreased nighttime activity, reduced emissions from traffic and industrial operations, and possible atmospheric dispersion effects. However, starting from 9 AM onwards, $PM_{2.5}$ levels gradually increase again, reaching another peak around 3-4 PM. This secondary rise could be associated with increased human activity, rush hour traffic, and industrial operations resuming in the afternoon. Following this peak, $PM_{2.5}$ levels begin to decline after 5 PM, but they remain elevated throughout the evening, suggesting continued emissions from vehicles, businesses, and urban activities. This pattern indicates a bimodal distribution of $PM_{2.5}$ levels, with higher pollution during late night/early morning and afternoon periods, while the cleanest air quality is observed in the mid-morning hours.



Visualization 7: Average PM2.5 per seasons

The bar chart illustrates the average $PM_{2.5}$ levels across different seasons—Cool Season, Hot Season, and Rainy Season. The Cool Season exhibits the highest $PM_{2.5}$ concentration, suggesting that air pollution tends to peak during this period. This could be due to temperature inversion, lower wind speeds, or increased emissions from heating and industrial activities, as pollutants become trapped in the atmosphere. In contrast, the Hot Season shows the lowest $PM_{2.5}$ levels, likely due to higher atmospheric mixing, increased wind dispersion, and potentially reduced human activity. Meanwhile, the Rainy Season has moderate $PM_{2.5}$ levels, lower than the Cool Season but higher than the Hot Season. This is likely because rainfall helps wash out airborne pollutants, though light rain has minimal impact on $PM_{2.5}$ reduction—only moderate to heavy rainfall significantly lowers pollution levels. This trend suggests that seasonal variations play a crucial role in air quality, with meteorological conditions significantly influencing $PM_{2.5}$ concentrations. Understanding these seasonal patterns can help guide air quality management strategies, such as implementing stricter pollution controls during the Cool Season or leveraging the natural cleansing effects of rainfall to improve urban air quality.



Visualization 8: Heatmap Correlation across feature

The heatmap provides the **Pearson correlation coefficients** between **PM2.5 levels** and other features. Population Density with a positive correlation (0.17) suggests higher air pollution in densely populated areas, mainly due to traffic and human activities. Hot season with a negative correlation (-0.17) indicates a weak negative relationship meaning that during the hot season, PM2.5 levels slightly decrease, but the effect is not strong.

6.3 Evaluation of the prediction methods

The evaluation of prediction methods has demonstrated varied levels of regression and performance across different models.

For regression tasks, the X model outperformed others regression models , achieving the highest R² score and the lowest Mean Absolute Error (MAE) (X) and Mean Squared Error (MSE) (X), as shown in Table X.

These results underscore the effectiveness of the X model and X model, particularly when incorporating key variables listed in Table X, for air quality prediction tasks.

7. Summary Progress

So far, the project has successfully defined its scope, identified the input features required for model training, and determined the target variable to be displayed on the website. A literature review on related work in air quality prediction has been conducted, and data have been acquired from relevant sources via API. Additionally, the machine learning models have been selected, with a focus on regression-based approaches. The methodological approach has been outlined, detailing feature selection, model training, and evaluation strategies.

Moving forward, the focus will be on the exploratory data analysis (EDA) process, examining both univariate and multivariate distributions. Following this, different regression-based machine learning models will be trained and evaluated. To enhance performance, hyperparameter optimization will be conducted, ensuring the best possible outcomes. Once trained, the models will be assessed using R², MSE, and MAE to compare their effectiveness and determine the most suitable approach.

Despite this progress, several challenges may impact the project's completion. Data availability and quality remain a concern, as missing or inaccessible primary data could necessitate reliance on alternative sources that may not fully reflect current trends. Additionally, computational limitations could pose challenges in training complex regression models on large datasets. Another concern is feature selection and interpretation, as accurately identifying the most influential environmental factors is crucial for model reliability. Lastly, ensuring the model's generalizability—its ability to perform well on unseen data and future predictions—remains a critical consideration. Addressing these challenges will be essential for the project's success.

8. REFERENCES

[1] Ganguli, I., Nakum, M., Das, B., & Kshetrimayum, N. (2025). Comprehensive analysis of air quality trends in India using machine learning and deep learning models. *Proceedings of the 26th International Conference on Distributed Computing and Networking*, 313–318. https://doi.org/10.1145/3700838.3703681

[2] Mohammadi, F., Teiri, H., Hajizadeh, Y., Abdolahnejad, A., & Ebrahimi, A. (2024). Prediction of atmospheric $PM_{2.5}$ level by machine learning techniques in Isfahan, Iran. *Scientific Reports, 14*(1), Article 2109. https://doi.org/10.1038/s41598-024-52617-z

[3] Lin, X. (2021). *The application of machine learning models in the prediction of $PM_{2.5}/PM_{10}$ concentration*. In Proceedings of the 2021 4th International Conference on Computers in Management and Business (ICCMB '21) (pp. 94–101). Association for Computing Machinery. https://doi.org/10.1145/3450588.3450605

[4] Alam, B., Hussain, A., & Fayaz, M. (2023). *An effective approach for air quality prediction in Bishkek based on machine learning techniques*. In *Proceedings of the 2023 7th International Conference on Advances in Artificial Intelligence* (pp. 42–47). Association for Computing Machinery. https://doi.org/10.1145/3633598.3633606

[5] https://openweathermap.org/api/air-pollution

[6] https://www.nationthailand.com/thailand/general/40033881

[7] Kermani, M., Hassani, G., & Ghaffari, H. (2024). Prediction of atmospheric $PM_{2.5}$ level by machine learning approaches. *Scientific Reports, 14*, 52617. https://doi.org/10.1038/s41598-024-52617-z

[8] Lin, X., & Li, Y. (2021). A deep learning approach for air quality prediction using LSTM neural network. Proceedings of the IEEE Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 1016–1020. https://doi.org/10.1109/IAEAC50856.2021.9390642

[9] Alam, F., & Qazi, E. U. (2023). Air pollution prediction using machine learning models with hyperparameter optimization. Proceedings of the ACM SIGKDD Workshop on Data Mining for Environmental and Earth Sciences (DMEES), 45–52. https://doi.org/10.1145/3490099.3511150

[10] Sookchaiya, T., & Phimoltares, S. (2022). Prediction of $PM_{2.5}$ and PM_{10} in Chiang Mai Province: A comparison of

machine learning models. Proceedings of the 14th International Conference on Knowledge and Smart Technology (KST), 34–39. https://doi.org/10.1109/KST55814.2022.9894884

[11] Zhang, Y., & Wang, J. (2022). Air quality prediction using machine learning algorithms: A case study in China. Proceedings of the IEEE International Conference on Big Data (Big Data), 1234–1240. https://doi.org/10.1109/BigData52589.2022.10067262

[12] Kumar, A., & Singh, P. (2023). A comparative study of machine learning models for air quality prediction in smart cities. Proceedings of the IEEE International Conference on Smart City Innovations (SCI), 89–95. https://doi.org/10.1109/SCI54681.2023.10662123

[13] Di, Q., Kloog, I., Koutrakis, P., Lyapustin, A., Wang, Y., & Schwartz, J. (2016). Assessing PM2.5 Exposures with High Spatiotemporal Resolution across the Continental United States. *Environmental Science & Technology*, 50(9), 4712–4721. https://doi.org/10.1021/acs.est.5b06121

[14] Yin Long, Yazheng Wu, Yang Xie, Liqiao Huang, Wentao Wang, Xiaorui Liu, Ziqiao Zhou, Yuqiang Zhang, Tatsuya Hanaoka, Yiyi Ju, Yuan Li, Bin Chen, Yoshikuni Yoshida, PM2.5 and ozone pollution-related health challenges in Japan with regards to climate change, Global Environmental Change, Volume 79, 2023, 102640, ISSN 0959-3780,https://doi.org/10.1016/j.gloenvcha.2023.102640. (https://www.sciencedirect.com/science/article/pii/S0959378023000067).

[15] Yamagami, M.; Ikemori, F.; Nakashima, H.; Hisatsune, K.; Ueda, K.; Wakamatsu, S.; Osada, K. Trends in PM2.5 Concentration in Nagoya, Japan, from 2003 to 2018 and Impacts of PM2.5 Countermeasures. Atmosphere 2021, 12, 590. https://doi.org/10.3390/ atmos12050590

[16] Ito, A.; Wakamatsu, S.; Morikawa, T.; Kobayashi, S. 30 Years of Air Quality Trends in Japan. Atmosphere 2021, 12, 1072. https:// doi.org/10.3390/atmos12081072

[17] Kunugi, Y., Arimura, T. H., Iwata, K., Komatsu, E., & Hirayama, Y. (2018). Cost-efficient strategy for reducing $PM_{2.5}$ levels in the Tokyo metropolitan area: An integrated approach with air quality and economic models. *PLoS ONE, 13*(11), e0207623. https://doi.org/10.1371/journal.pone.0207623

[18] Geng et al. (2021): "Tracking Air Pollution in China: Near Real-Time $PM_{2.5}$ Retrievals from Multiple Data Sources" Environmental Science & Technology, 55(12), 12106–12115. https://doi.org/10.1021/acs.est.1c01863

[19] Mendez et al. (2022): "Using Low-Cost Sensors to Assess $PM_{2.5}$ Concentrations at Four South Texan Cities on the U.S.–Mexico Border" Atmosphere, 13(10), 1554. https://doi.org/10.3390/atmos13101554

[20] https://en-gb.topographic-map.com/map-63pz4/Thailand/?center=13.34057%2C100.77554&zoom=9

[21] https://www.researchgate.net/figure/Map-of-Thailand-and-the-Bangkok-metropolitan-area-AMSL-altitude-above-mean-sea-level_fig1_344854116

[22] https://webportal.bangkok.go.th/public/user_files_editor/299/3%20Pdf%2068/1%20%E0%B8%AA%E0%B8%96%E0%B8%B4%E0%B8%95%E0%B8%B4%E0%B8%9B%E0%B8%B5%202566.pdf

[23] https://webportal.bangkok.go.th/public/user_files_editor/354/aboutcpud/study%20report/2563/%E0%B8%9B%E0%B8%B5%2063/8%E0%B8%A3%E0%B8%B2%E0%B8%A2%E0%B8%87%E0%B8%B2%E0%B8%99%E0%B8%AA%E0%B8%96%E0%B8%B4%E0%B8%95%E0%B8%B4%E0%B8%AD%E0%B8%B8%E0%B8%95%E0%B8%AA%E0%B8%B2%E0%B8%AB%E0%B8%81%E0%B8%A3%E0%B8%A3%E0%B8%A1%E0%B9%83%E0%B8%99%E0%B9%80%E0%B8%82%E0%B8%95%E0%B8%81%E0%B8%A3%E0%B8%B8%E0%B8%87%E0%B9%80%E0%B8%97%E0%B8%9E%E0%B8%A1%E0%B8%AB%E0%B8%B2%E0%B8%99%E0%B8%84%E0%B8%A3%20%E0%B8%9B%E0%B8%B5%2062.pdf

[24]

About the authors:

John Q. Miner is the

Author 2 is blah blah balh…

Columns on Last Page Should Be Made Equal Length