

**Northeastern University**

**ALY6060: Decision Support and Business Intelligence**

**Professor Michael Cavallo**

**Module Assignment: Final Project Paper**

**Methapat Vorakamolpisit**

**05/17/2024**

Thailand, situated centrally within Southeast Asia, is bordered by Myanmar, Laos, Cambodia, and Malaysia. The country is characterized by its six distinct regions: Central, North, Northeast, East, and South, each boasting unique cultures, culinary delights, and dialects. Among these, the southern region is renowned for its stunning beaches and azure seas, while the northern part captivates with its majestic mountains, lush forests, and culinary treasures like Kao Soi

Nestled in the mountainous region of northern Thailand, Chiang Mai exudes a unique charm that effortlessly blends ancient traditions with modern influences. As the largest city in northern Thailand, Chiang Mai is renowned for its rich cultural heritage, vibrant arts scene, and stunning natural landscapes. The city is adorned with centuries-old temples, including the iconic Wat Phra That Doi Suthep, which sits atop a hill overlooking the city and offers panoramic views of the surrounding countryside. Chiang Mai's bustling markets, such as the famous Night Bazaar, offer a kaleidoscope of colors, scents, and sounds, where visitors can immerse themselves in the local culture and shop for handmade crafts, textiles, and delectable street food. Beyond the city's historic walls, verdant rice paddies, lush jungles, and cascading waterfalls beckon adventurers to explore the region's pristine beauty. With its warm hospitality, laid-back atmosphere, and abundance of attractions, Chiang Mai captivates travelers from around the globe, earning its reputation as the "Rose of the North."

However, the once idyllic weather in the northern region has been marred by the persistent presence of PM 2.5, or Particulate Matter 2.5, for nearly a decade. PM2.5 refers to tiny particles in the air that are smaller than 2.5 micrometers in diameter, originating from various sources such as vehicle emissions, industrial activities, construction, agriculture, and natural sources like wildfires and dust storms. Due to their minuscule size, PM2.5 particles can penetrate deep into the lungs and even enter the bloodstream when inhaled, posing serious health risks. Chronic exposure to PM2.5 has been linked to respiratory diseases such as asthma, bronchitis, and lung cancer, as well as cardiovascular problems and premature death. Additionally, PM2.5 can exacerbate existing health conditions and impair lung function, particularly in vulnerable populations such as children, the elderly, and individuals with pre-existing health issues. The threat of PM2.5 pollution underscores the urgent need for comprehensive measures to reduce emissions and mitigate its harmful effects on public health and the environment.

I aim to investigate the impact of PM2.5 pollution on Chiang Mai using a dataset containing various variables such as date, atmospheric pressure, temperature, humidity, precipitation, sunshine duration, evaporation, wind direction, wind speed, and PM2.5 levels. The dataset covers a time span from July 11, 2016, to June 30, 2023. By analyzing this time series data, I seek to highlight the challenges faced by Chiang Mai due to elevated PM2.5 levels, emphasizing the fluctuations and trends observed over the specified period. This investigation aims to shed light on the severity of PM2.5 pollution in Chiang Mai and its implications for public health and environmental well-being.

A graph of a graph

Description automatically generated with medium confidence

*Figure 1*

*Figure* 1 illustrates the PM2.5 levels in Chiang Mai province from July 11, 2016, to June 30, 2023. Notably, the data shows a recurring pattern of increased PM2.5 concentrations at the beginning of each year. The PM2.5 levels peak in early 2019 and again in 2023, with values surpassing 100, indicating very severe pollution, except for a notable dip in 2022. This cyclical pattern is characterized by high PM2.5 levels at the start of the year, a decline around mid-year, followed by a gradual increase towards the year's end. This seasonal trend underscores the need for targeted measures to address air quality issues in Chiang Mai, particularly during peak pollution periods.

To analyze the PM2.5 levels in Chiang Mai, the dataset is divided into features and target variables. The features (X) include all columns except 'PM25', which is the target variable (y) representing the PM2.5 levels. The dataset is then split into training and testing sets based on the date. Data prior to January 1, 2022, is used for training (X\_train and y\_train), while data from January 1, 2022, onwards is reserved for testing (X\_test and y\_test). This temporal split ensures that the model is trained on historical data and tested on more recent data, allowing for an evaluation of its predictive performance. The shapes of the training and testing sets are displayed to confirm the correct partitioning of the dataset.

To predict PM2.5 levels, the XGBoost regression model (XGBRegressor) is utilized. The model is then trained using the training data (X\_train and y\_train), with evaluation metrics applied to both the training and testing datasets to monitor performance. The evaluation metric used is Root Mean Square Error (RMSE), which measures the differences between predicted and actual values. Additionally, early stopping is implemented with a threshold of 10 rounds to prevent overfitting; training halts if the model's performance does not improve over 10 consecutive rounds, ensuring optimal predictive accuracy and efficiency.

To improve the model's predictions, we add lag features to the dataset. This means creating new columns that contain PM2.5 values from previous days. For example, if we choose 3 lags, we'll add three new columns: 'lag\_1', 'lag\_2', and 'lag\_3'. These columns will show the PM2.5 levels from one day ago, two days ago, and three days ago. This helps the model understand the relationship between past and current PM2.5 levels. After adding these new columns, we remove any rows with missing values that resulted from this process. Finally, we check the first few rows of the updated dataset to ensure the lag features are added correctly.

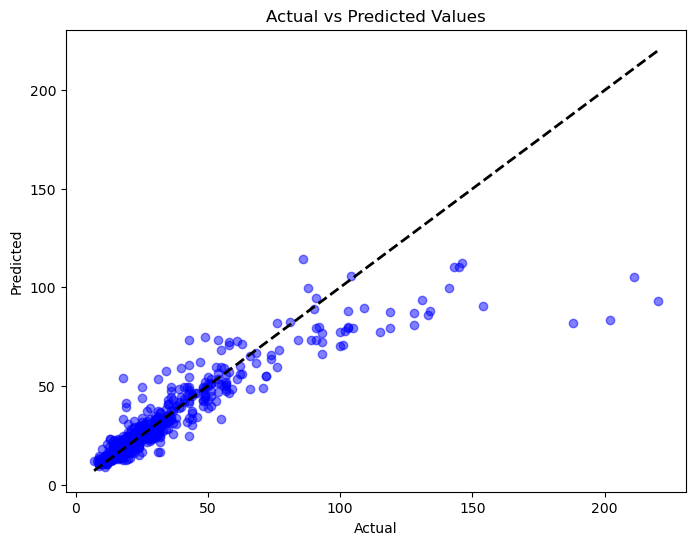
A graph with orange and blue lines

Description automatically generated

*Figure 2*

*Figure 2* illustrates the results of the PM2.5 predictions, where the orange line represents the predicted PM2.5 values and the blue line shows the actual PM2.5 measurements. Initially, the predictions closely match the actual values, indicating high accuracy. However, after February 2023, the predictions begin to slightly diverge from the actual values, generally falling below them.

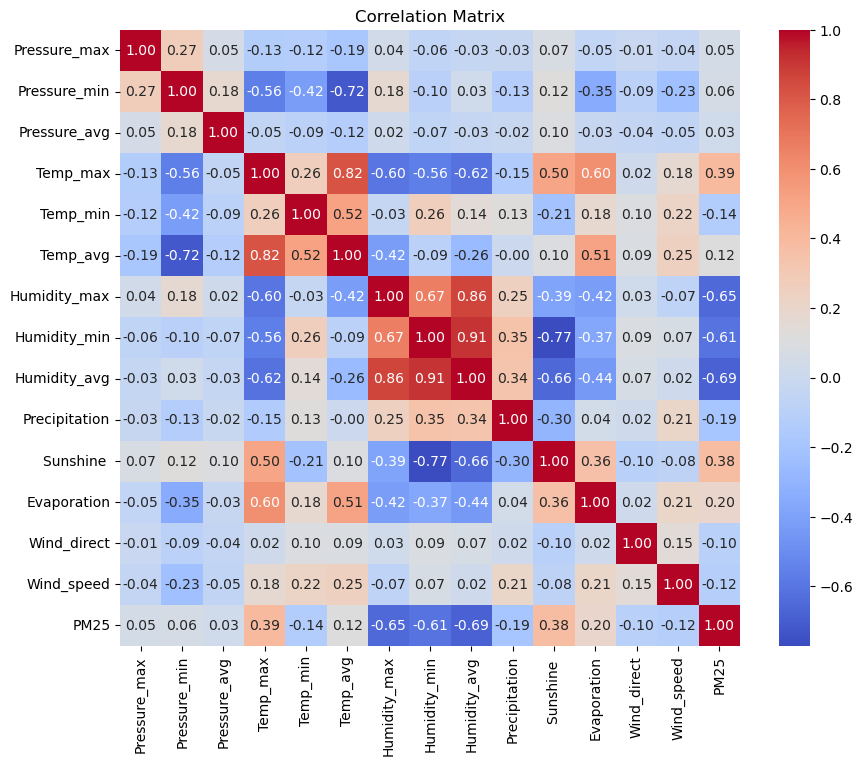
To evaluate the model's performance, several error metrics were calculated. The Root Mean Square Error (RMSE) is 13.76, the Mean Squared Error (MSE) is 189.41, the Mean Absolute Error (MAE) is 6.27, and the Mean Absolute Percentage Error (MAPE) is 16.80%. These metrics indicate that the model performs reasonably well, with the RMSE of 13.76 suggesting a relatively low overall error. This supports the model's credibility in predicting PM2.5 levels, although there is room for improvement in its accuracy, particularly for predictions made after February 2023.



*Figure 3*

*Figure 3* presents a scatter plot comparing the actual PM2.5 values with the predicted values. Each blue dot represents a data point, plotting actual PM2.5 values on the x-axis against predicted values on the y-axis. The black dashed line represents the ideal scenario where predicted values perfectly match actual values (a 45-degree line).

From the scatter plot, it is evident that many of the data points closely align with the dashed line, indicating that the model's predictions are generally accurate. However, as the actual PM2.5 values increase, the spread of the points widens, showing that the model's predictions tend to underestimate higher PM2.5 values. This observation aligns with earlier findings where predictions after February 2023 were noted to be slightly below the actual values*.*



*Figure 4*

In the provided correlation matrix (*Figure 4*), the primary focus of this report is to scrutinize the relationship between PM 2.5 and other variables. Initial observations indicate a minimal correlation between PM 2.5 and atmospheric pressure. Conversely, an intriguing finding emerges regarding the average humidity, which exhibits a negative moderate correlation with PM 2.5 levels. This suggests that as humidity increases, there is a tendency for PM 2.5 concentrations to decrease. Such insight could elucidate the seasonal fluctuations in PM 2.5 levels, particularly during the middle of the year, coinciding with the rainy season. During this period, the elevated humidity levels may contribute to a concurrent reduction in PM 2.5 concentrations. This interpretation sheds light on the intricate interplay between environmental factors and particulate matter levels, enriching our understanding of air quality dynamics.

A graph of blue and orange lines

Description automatically generated

*Figure 5*

*In Figure 5,* I observe a compelling correlation between Humidity Average and PM 2.5 levels. Notably, as humidity decreases, there is a noticeable uptick in PM 2.5 concentration, depicted on the right Y-axis. This relationship underscores the significance of humidity in air quality dynamics, with lower humidity often corresponding to heightened levels of particulate matter. Understanding this interplay is crucial for mitigating air pollution and its associated impacts on health and the environment.

In conclusion, this report highlights the problem of PM2.5 pollution in Chiang Mai, Thailand, showing how it affects the city's cultural and environmental aspects. By studying data from several years, we see a trend of high PM2.5 levels, especially at the start of each year. While our prediction model helps understand this trend, it struggles with accuracy after early 2023. Still, we find that humidity plays a crucial role in PM2.5 levels, offering insights for potential solutions. Overall, this report emphasizes the need for coordinated efforts to address PM2.5 pollution in Chiang Mai, aiming to protect the city's people, environment, and cultural heritage for generations to come.