Beijing Air Pollution

Shyam Bhaskar (Roll No:S20210010035)
Yalakanti Eswar (Roll No:S20210020332)

April 16, 2024

Abstract

Air pollution, particularly PM2.5, presents significant health risks in urban areas globally. Beijing, a populous city in China, confronts severe air pollution issues. Understanding the factors influencing PM2.5 concentrations is crucial for effective pollution management.

We analyze hourly PM2.5 data from the US Embassy in Beijing alongside meteorological observations from Beijing Capital International Airport. By examining the relationship between PM2.5 levels and meteorological parameters such as **dew point, temperature, pressure, combined wind direction, cumulated wind speed, cumulated hours of snow, and cumulated hours of rain**, we aim to develop predictive models.

Our analysis utilizes linear regression and support vector regression (SVR) models to forecast PM2.5 concentrations based on meteorological variables.The models exhibit moderate predictive capabilities. Specifically, employing Support Vector Regression (SVR) yielded an R-squared value of 0.6101, whereas utilizing an Artificial Neural Network (ANN) alongside SVR resulted in an improved R-squared value of 0.64134 for daily data. This study contributes to ongoing efforts to mitigate air pollution and safeguard public health in Beijing and similar urban environments.

Keywords
PM2.5(Particulate Matter), Linear Regression Model, SVR(Support Vector Regression),ANN(Artificial Neural Networks), RSquare

1 Introduction

Air pollution, particularly the presence of fine particulate matter (PM2.5), has emerged as one of the most pressing environmental challenges in contemporary times, with profound implications for public health and the sustainability of urban environments worldwide. Among the many urban areas grappling with these issues, Beijing, as a densely populated and rapidly developing metropolis, has been at the forefront of global attention due to its notorious air quality problems in recent years. The prevalence of high levels of PM2.5 pollution in Beijing has raised significant concerns regarding the health and well-being of its inhabitants, as well as the broader environmental and economic impacts.

The rise in urbanization, industrialization, and vehicular emissions has contributed to the exacerbation of air pollution, particularly in densely populated areas like Beijing. PM2.5, defined as particulate matter with a diameter of 2.5 micrometers or less, poses a substantial risk to public health due to its ability to penetrate deep into the respiratory system, causing a range of adverse health effects including respiratory diseases, cardiovascular problems, and even premature death. In addition to its direct health impacts, PM2.5 pollution also has far-reaching consequences for environmental quality, economic productivity, and social well-being, making it a complex and multifaceted issue that requires urgent attention and effective management strategies.

To address the challenge of PM2.5 pollution in urban environments like Beijing, it is essential to understand the factors that influence the concentration and distribution of PM2.5 particles in the atmosphere. While emissions from sources such as industrial facilities, vehicles, and construction activities play a significant role in PM2.5 pollution, meteorological conditions also exert a considerable influence on the dispersion and accumulation of pollutants in the air. Factors such as temperature, humidity, wind speed, atmospheric pressure, and precipitation can all affect the transport and diffusion of PM2.5 particles, shaping their spatial and temporal distribution within urban areas.

In light of these considerations, this report presents a comprehensive analysis of the Beijing PM2.5 Data Set, which provides hourly measurements of PM2.5 concentration recorded by the US Embassy in Beijing, along with accompanying meteorological data from Beijing Capital

International Airport. Spanning from January 1st, 2010, to December 31st, 2014, this dataset offers a valuable resource for examining the relationship between PM2.5 levels and meteorological variables over an extended period.

By leveraging the rich temporal and multivariate nature of the dataset, our analysis aims to uncover patterns, trends, and correlations that can shed light on the complex interactions between air pollution and meteorological conditions in Beijing. Through the application of statistical techniques and regression models, we seek to develop predictive models capable of forecasting PM2.5 concentrations accurately based on meteorological parameters, thereby providing valuable insights for policymakers, urban planners, and public health officials involved in air quality management and pollution control efforts in Beijing and similar environments around the world.

2 Data Preprocessing

2.1 Removing Duplicates

Duplicate rows in the dataset are identified and eliminated to ensure that each observation is unique. This step prevents redundancy in the data, which could skew analysis results and model performance.

2.2 Handling Missing Values

Missing values, specifically in the "pm2.5" variable, are filled by replacing them with the mean PM2.5 concentration for the respective day. This approach retains the temporal context of the data while imputing missing values. If any missing values persist after this step, the corresponding rows are removed from the dataset to maintain data completeness and integrity for further analysis. This method ensures that the dataset remains representative and reliable for subsequent modeling and interpretation.

2.3 Removing Noise and Outliers

Noise, which refers to random variation or errors in the data, and outliers, which are data points significantly different from the majority of observations, are identified and addressed. This process involves applying statistical methods or visual inspection to detect and remove or correct noisy or outlier data points, ensuring that the dataset accurately represents the underlying phenomena being studied.

2.4 One-Hot-Encoding and Standard Scaling

We translated categorical variables into numerical representations using the one-hot-encoding technique. To further guarantee that the numerical data values fit within a standardised range of -3 to 3, we additionally perform standard scaling; however, the categorical columns are not included in this transformation.

2.5 Dimensionality Reduction

To simplify the dataset and enhance computational efficiency, dimensionality reduction techniques are employed, including the application of Principal Component Analysis (PCA). The following steps are taken:

1. We have transformed hourly data into daily intervals for the purpose of visualization and prediction.

2. Removal of Unnecessary Columns: The "no" column is removed as it does not contribute to the analysis. Additionally, the "year," "month," "day," and "hour" columns are combined into a single datetime variable to reduce the number of separate time-related features.

3. Standardization: The remaining attributes are standardized to have a mean of 0 and a standard deviation of 1 to ensure uniform scaling.

4. PCA: PCA is applied to the standardized dataset to transform the original variables into a set of linearly uncorrelated variables called principal components.

5. Scree Plot Analysis: A Scree Plot is utilized to determine the number of principal components to retain. This graphical method helps identify the point at which the eigenvalues of the principal components begin to level off, indicating diminishing returns in terms of explained variance.

6. Retention of Principal Components: Based on the Scree Plot analysis, the optimal number of principal components to retain is determined, ensuring a balance between maintaining sufficient variance and reducing dimensionality.

By applying PCA and determining the appropriate number of principal components to retain through Scree Plot analysis, the dimensionality of the dataset is effectively reduced while preserving the most significant variance in the data.

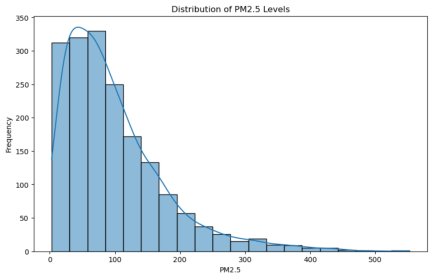

Figure 1: Distribution of PM2.5 Levels

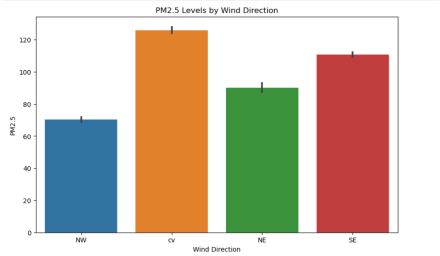

Figure 2: PM2.5 vs Wind Direction

This facilitates more efficient analysis and modeling while retaining essential information from the original dataset.

3 Explanatory Data Analysis(EDA):

Figure 1 This plot shows the distribution of PM2.5 value in the Dataset. (Right Skewed Data)

Figure 2 The plot showcases the relationship between wind direction and PM2.5 values.

Figure 3 shows the Monthly Distribution of PM2.5 values in the dataset

Figure 4 shows the Seasonal Distribution of PM2.5 values in the dataset

Figure 6 The plot displays the yearly change in PM2.5 values from 2010 to 2014

Figure 6 The plot displays the yearly distri-

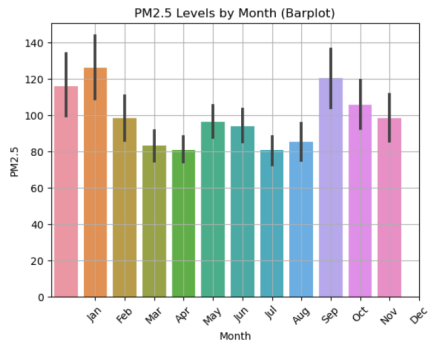

Figure 3: Monthly Distribution of PM2.5 Levels

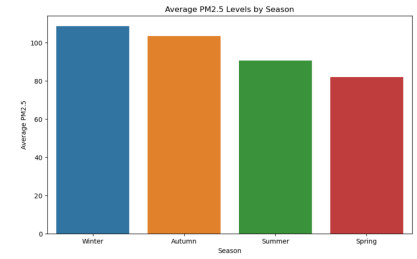

Figure 4: Seasonal Distribution of PM2.5 Levels

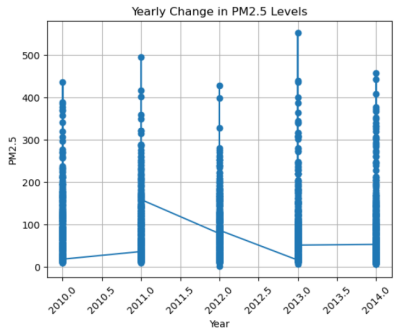

Figure 5: Yearly Change in Distribution of PM2.5 Levels

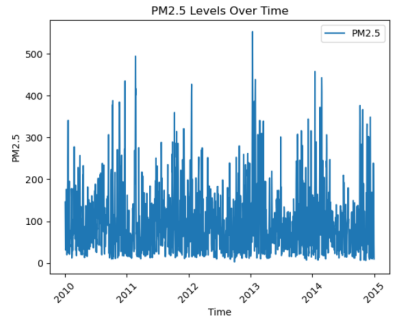

Figure 6: Yearly Distribution of PM2.5 Levels

bution of the dataset's PM2.5 values from 2010 to 2014

4 Model: Linear Regression

After conducting thorough data preprocessing steps, including removing duplicates, handling missing values, removing noise and outliers, and performing dimensionality reduction using PCA, a Linear Regression model is applied to the preprocessed dataset. Here's a detailed explanation of each step:

1. **Data Preparation:**
   We organized the cleaned-up data into two parts: features related to weather conditions (Dew Point, Temperature, Pressure, Combined Wind Direction, Cumulated Wind Speed, Cumulated Hours of Snow, Cumulated Hours of Rain) and the variable we aim to predict, which is PM2.5 concentration. We chose to focus on PM2.5 because our research indicated that it's closely associated with pollution compared to other factors.

2. **Train-Test Split:**
   The dataset is split into training and testing sets to evaluate the performance of the model. The training set, typically comprising 80% of the data, is used to train the model, while the testing set (20%) is used to assess the model's predictive performance on unseen data.

3. **Model Training:**
   The Linear Regression model is trained on the training set. During the training process, the model learns the coefficients for each feature, aiming to minimize the difference between the actual PM2.5 concentrations and the predicted values.

4. **Model Evaluation:**
   The trained model is evaluated using the testing set to assess its performance in predicting PM2.5 concentrations. Evaluation metrics such as Mean Squared Error (MSE), R-squared (R^2) value, and Mean Absolute Error (MAE) are calculated. These metrics help measure the accuracy and goodness-of-fit of the model. A lower MSE and MAE and a higher R^2 value indicate better performance.

5. **Model Interpretation:**
   The coefficients of the linear regression model are interpreted to understand the significance and impact of each feature on predicting PM2.5 concentrations. Positive coefficients indicate a positive correlation with PM2.5 concentrations, while negative coefficients suggest a negative correlation. The magnitude of the coefficients reflects the strength of the relationship between each feature and the target variable.

By employing a Linear Regression model after thorough data preprocessing, we aim to develop a predictive model capable of accurately estimating PM2.5 concentrations based on meteorological variables and other relevant features. This model serves as a fundamental baseline for air quality prediction and can provide valuable insights for pollution management efforts in urban environments such as Beijing.

5 Model: ANN (Artificial neural network) with SVR (Support vector Regression)

5.1 Daily data:

Using the **selu** activation function (scaled exponential linear unit), an Artificial Neural Network (ANN) was implemented on the daily_data to extract features from the dataset. These features were then used to an SVR model (support vector regression) using a radial basis function ('rbf') kernel. By adjusting the hyperparameters, the best characteristics were chosen.resulting in an R-squared value of 0.64057 and a Mean Squared Error (MSE) of 0.2649.

5.2 Hour wise data:

we created an Artificial Neural Network (ANN) and used the 'tanh' (hyperbolic tangent) activation function to extract features from the hourly dataset. Subsequently, a Support Vector Regressor (SVR) with a 'rbf' (radial basis function) kernel was trained these features. I found the best features by using hyperparameter optimisation, and the result was an amazing R-squared value of 0.7259.

6 Results

After applying the Linear Regression model to the testing set, we obtained the following evaluation metrics:

The table presents a comparison of model performance metrics for daily-based data using various machine learning techniques. Four

Model	MSE	RMSE	R$\hat{2}$
Linear Regression	0.337	0.580	0.542
SVR	0.2873	0.536	0.610
ANN with SVC	0.264	0.5138	0.640
polynomial reg	0.320	0.565	0.565

Table 1: Comparison of Model Performance for daily base data

models are evaluated: Linear Regression, Support Vector Regression (SVR), Artificial Neural Network (ANN) with Support Vector Regression (SVR), and Polynomial Regression. Each model is assessed based on Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) values. Among the models, ANN with SVR demonstrates the highest performance with the lowest MSE of 0.264 and RMSE of 0.5138, indicating superior predictive accuracy compared to the other models. SVR follows closely behind, with slightly higher MSE and RMSE but a commendable R^2 value of 0.610, suggesting strong explanatory power. Linear Regression and Polynomial Regression exhibit relatively weaker performance in comparison, with higher MSE and RMSE values, although their R^2 values are still notable at 0.542 and 0.565, respectively. Overall, the table highlights the efficacy of ANN with SVR for daily-based data prediction tasks.

Model	MSE	RMSE	R$\hat{2}$
Linear Regression	0.614	0.783	0.422
SVR	0.325	0.570	0.694
ANN with SVR	0.291	0.539	0.725
polynomial reg	0.369	0.607	0.652

Table 2: Comparison of Model Performance for hourly base data

The table presents a comparative analysis of model performance metrics for hourly-based data using various machine learning algorithms. Four models are evaluated: Linear Regression, Support Vector Regression (SVR), Artificial Neural Network (ANN) with SVR, and Polynomial Regression. Each model's performance is assessed based on Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) values. Among the models, ANN with SVR demonstrates the highest performance, achieving the lowest MSE of 0.291 and RMSE of 0.539, indicating superior predictive accuracy compared to other models. Additionally, ANN with SVR exhibits a remarkable R^2 value of 0.725, suggesting strong explanatory power. SVR closely follows with competitive metrics, featuring a commendable R^2 value of 0.694. However, Polynomial Regression exhibits slightly weaker performance in comparison, with higher MSE and RMSE values, although its R^2 value remains noteworthy at 0.652. Overall, the table underscores the efficacy of ANN with SVR for hourly-based data prediction tasks.

7 Conclusion

The analysis of PM2.5 pollution data in Beijing unveils complex air quality dynamics, emphasizing variations across monitoring stations and temporal scales. Correlations between PM2.5 levels and meteorological factors like temperature and humidity highlight the multifaceted nature of pollution sources and their interactions. Machine learning models, notably the ANN with SVR, show promise in predicting PM2.5 concentrations, achieving an impressive R^2 value of 0.725 for hourly data. However, challenges such as data sparsity and model validation limitations warrant consideration. Moving forward, addressing PM2.5 pollution demands concerted efforts, including stringent emission controls and enhanced monitoring infrastructure. Collaboration across disciplines is crucial for refining predictive models and advancing understanding of pollution mechanisms. By leveraging data-driven approaches and proactive measures, Beijing can mitigate PM2.5 pollution's adverse effects, safeguard public health, and promote environmental sustainability.