Guru Gobind Singh Indraprastha University

University School of Automation & Robotics

Machine Learning Project (ARM252)

Beijing Multi-Site Air-Quality Data

Submitted to:

Dr. Sanjay Kumar Singh
Asst. Professor, USAR

Submitted by:

Yatin Sharma
(08219011921)
AI-DS Batch 2

yatin.08219011921@ipu.ac.in

Introduction

- Air pollution has become a major concern in urban areas, with adverse effects on public health and the environment. Accurately predicting air quality is crucial for effective urban planning and public health management.
- This is a machine learning project focused on developing a regression model to predict air quality in Beijing using the "Beijing Multi-Site Air-Quality Data" dataset.
- Our goal is to build a regression model that can accurately predict air pollutant concentrations based on the available features.
- We will employ a supervised learning approach, utilizing regression algorithms such as linear regression, decision trees, etc.
- The model will be trained on the given dataset and evaluated using appropriate evaluation metrics, such as mean squared error or R-squared.
- By accurately predicting air quality, authorities can implement timely interventions to mitigate pollution levels, issue health advisories, and improve overall air quality management strategies.

Keywords: air quality, regression, machine learning, Beijing, multi-site data, public health management

Proposed Methodology

Regression

- Regression refers to a supervised learning task that involves predicting a continuous numeric output based on input variables
- It aims to find a function or model that can accurately map the input variables to the output variable.
- The regression model learns from a given dataset, adjusting its parameters to minimize the difference between predicted and actual values.
- Regression algorithms can handle various types of data and are widely used for tasks like predicting housing prices, stock market trends, or customer demand.

- Types of Regression Models used:

1. Linear Regression
   - Linear regression is a type of regression analysis that models the relationship between a dependent variable and one or more independent variables as a linear equation.
   - It assumes a linear relationship between the input variables and the output variable, meaning that the relationship can be represented by a straight line.
   - The goal of linear regression is to estimate the coefficients of the linear equation that minimize the difference between the predicted values and the actual values in the training data.
   - The coefficients represent the slope and intercept of the line and indicate the strength and direction of the relationship between the variables.
   - Linear regression can be used for both simple regression, involving a single independent variable, and multiple regression, involving multiple independent variables.

2. Random Forest Regressor
   - Random Forest Regressor is a machine learning algorithm that uses an ensemble of decision trees to perform regression tasks.
   - It combines the predictions of multiple decision trees, known as a forest, to make more accurate and robust predictions.
   - Each decision tree in the random forest is trained on a randomly selected subset of the training data and a subset of the input features.
   - During prediction, the random forest aggregates the predictions from all the trees to generate the final prediction, typically by taking the average or majority vote.
   - Random Forest Regressor is effective for handling complex relationships, handling both numerical and categorical data, and avoiding overfitting due to the randomness introduced in the training process.

3. Gradient Boosting Regressor

- Gradient Boosting Regressor is a machine learning algorithm that builds an ensemble of weak prediction models, typically decision trees, in a sequential manner.
- It starts with an initial weak model and iteratively improves it by adding new models that focus on correcting the mistakes made by the previous models.
- Each new model is trained to predict the residual errors of the ensemble up to that point.
- During prediction, the outputs of all the models are combined to generate the final prediction, with each model's contribution weighted by a learning rate.
- Gradient Boosting Regressor is known for its high predictive power and ability to handle complex relationships, but it can be prone to overfitting if not properly regularized.

4. K Neighbors Regressor

- K Neighbors Regressor is a machine learning algorithm used for regression tasks that predicts the value of a new data point based on the average of its k nearest neighbors.
- It calculates the distance between the new data point and the training data points to identify the k closest neighbors.
- The predicted value is determined by taking the average (or weighted average) of the target values of these k neighbors.
- The choice of k, the number of neighbors to consider, is a hyperparameter that can be tuned to achieve optimal performance.
- K Neighbors Regressor is a non-parametric algorithm and is useful when there is no explicit functional relationship between the input variables and the target variable.

5. Decision Tree Regressor

- Decision Tree Regressor is a machine learning algorithm that uses a tree-like structure to model the relationship between input variables and a continuous numeric output.
- It divides the input space into regions based on feature values, with each region representing a leaf node in the tree.
- The tree structure is built by recursively splitting the data based on the selected features and their corresponding thresholds, optimizing a criterion such as mean squared error or variance reduction.
- During prediction, the decision tree traverses the tree structure from the root to a leaf node, and the predicted value is typically the average of the target values in that leaf node.
- Decision Tree Regressor is interpretable, can handle both numerical and categorical features, and is capable of capturing complex relationships. However, it can be prone to overfitting and may not generalize well to unseen data without proper regularization techniques.

Dataset Description

- This data set includes hourly air pollutants data from 12 nationally-controlled air-quality monitoring sites. The air-quality data are from the Beijing Municipal Environmental Monitoring Center. The meteorological data in each air-quality site are matched with the nearest weather station from the China Meteorological Administration. The time period is from March 1st, 2013 to February 28th, 2017.

Dataset Characteristics Multivariate, Time-Series	**Subject Area** Physical	**Associated Tasks** Regression
Attribute Type Integer, Real	**Instances** 420768	**Attributes** 18

Glimpse of the data

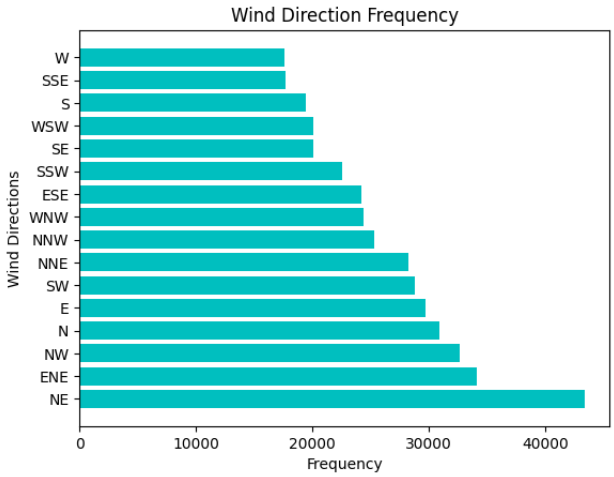
	No	year	month	day	hour	PM2.5	PM10	SO2	NO2	CO	O3	TEMP	PRES	DEWP	RAIN	wd	WSPM	station
0	1	2013	3	1	0	4.0	4.0	4.0	7.0	300.0	77.0	-0.7	1023.0	-18.8	0.0	NNW	4.4	Aotizhongxin
1	2	2013	3	1	1	8.0	8.0	4.0	7.0	300.0	77.0	-1.1	1023.2	-18.2	0.0	N	4.7	Aotizhongxin
2	3	2013	3	1	2	7.0	7.0	5.0	10.0	300.0	73.0	-1.1	1023.5	-18.2	0.0	NNW	5.6	Aotizhongxin
3	4	2013	3	1	3	6.0	6.0	11.0	11.0	300.0	72.0	-1.4	1024.5	-19.4	0.0	NW	3.1	Aotizhongxin
4	5	2013	3	1	4	3.0	3.0	12.0	12.0	300.0	72.0	-2.0	1025.2	-19.5	0.0	N	2.0	Aotizhongxin

Description of Columns

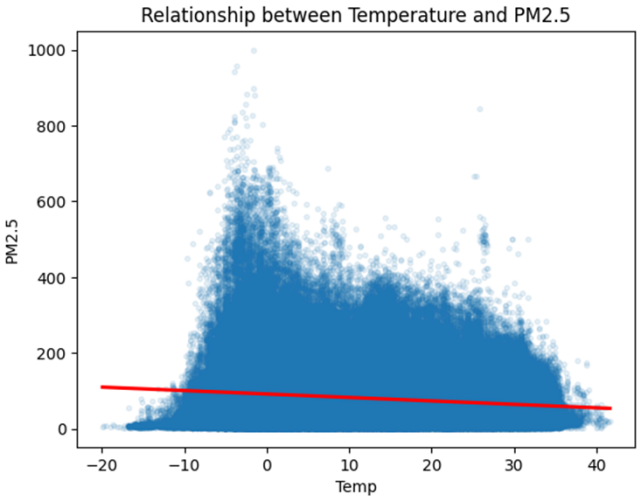
Column Names	Column Details
No	Row number
year	Year of data in this row
month	Month of data in this row
day	Day of data in this row
hour	Hour of data in this row
PM2.5	PM2.5 concentration (ug/m^3)
PM10	PM10 concentration (ug/m^3)
SO2	SO_2 concentration (ug/m^3)
NO2	NO_2 concentration (ug/m^3)
CO	CO concentration (ug/m^3)
O3	O_3 concentration (ug/m^3)
TEMP	temperature (degree Celsius)
PRES	pressure (h Pa)
DEWP	dew point temperature (degree Celsius)
RAIN	precipitation (mm)
wd	wind direction
WSPM	wind speed (m/s)
station	name of the air-quality monitoring site

Data Visualization

1. Wind Direction Frequency

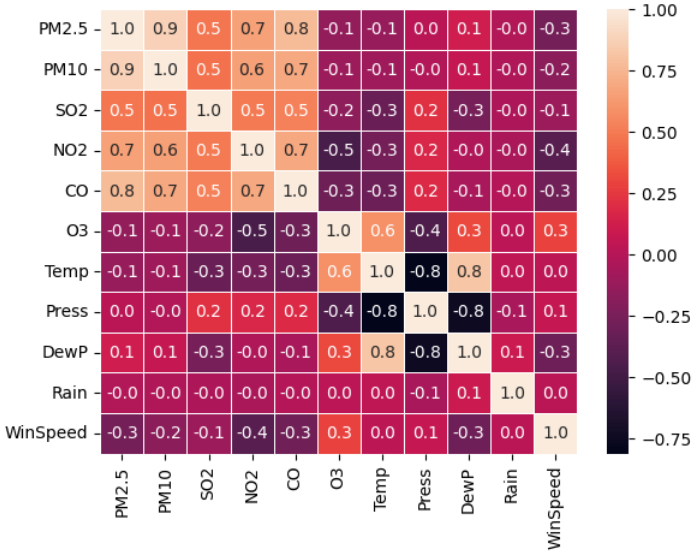


2. Relationship between PM2.5 and Temperature

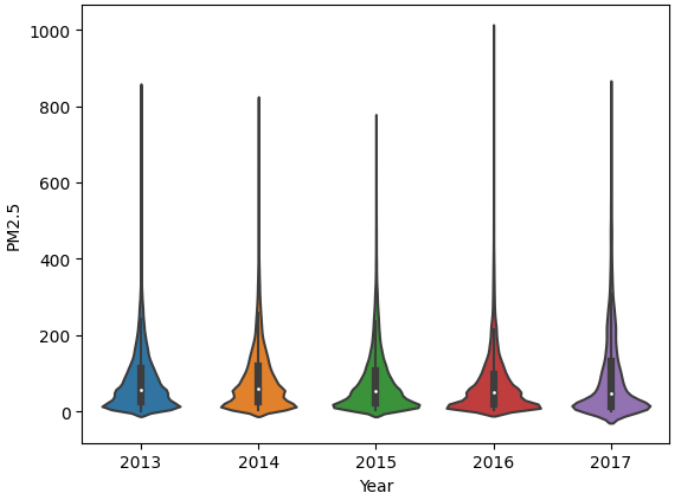


We can see that as temperature increases, PM2.5 decreases

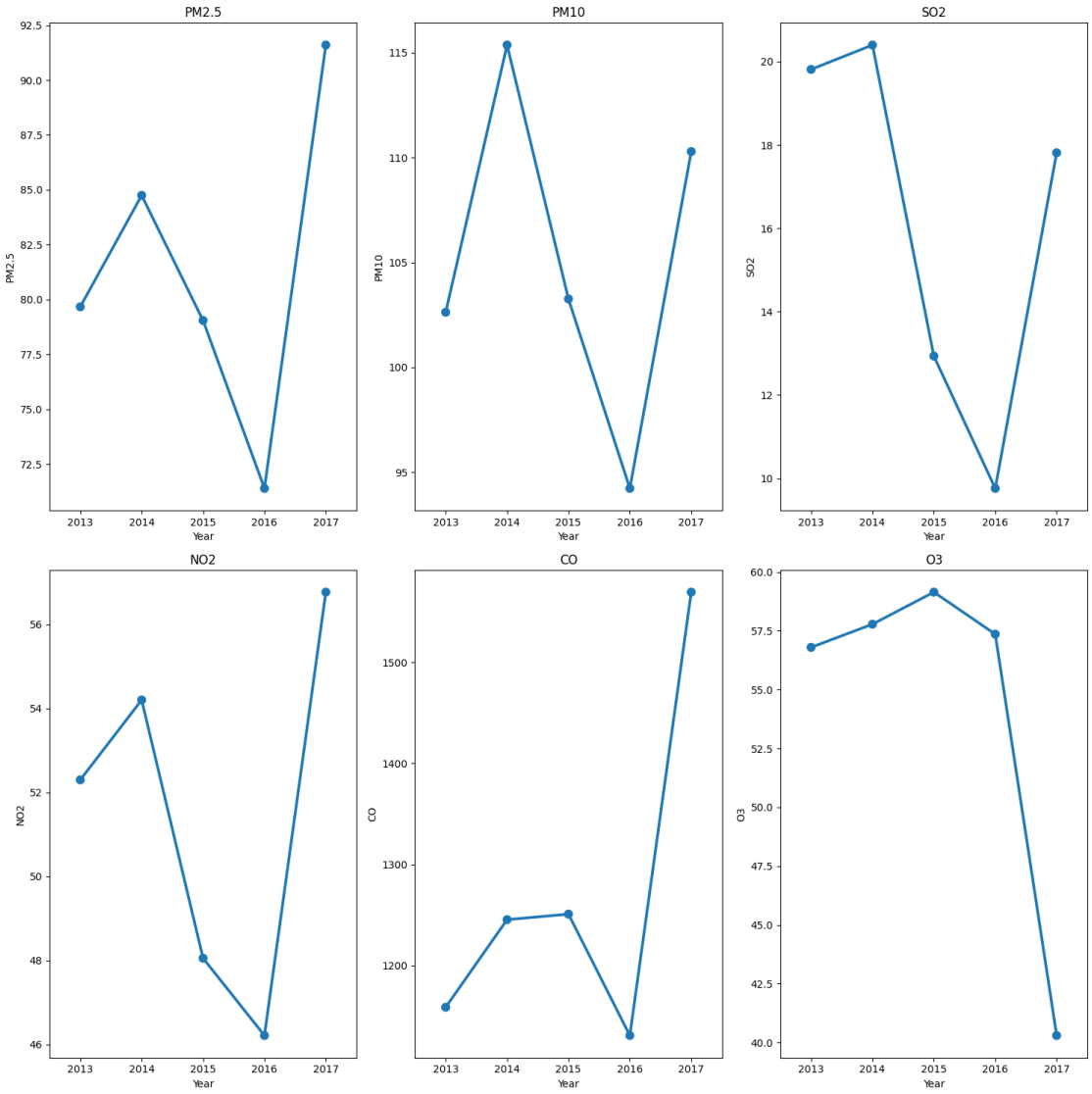
3. A Corelation Heatmap



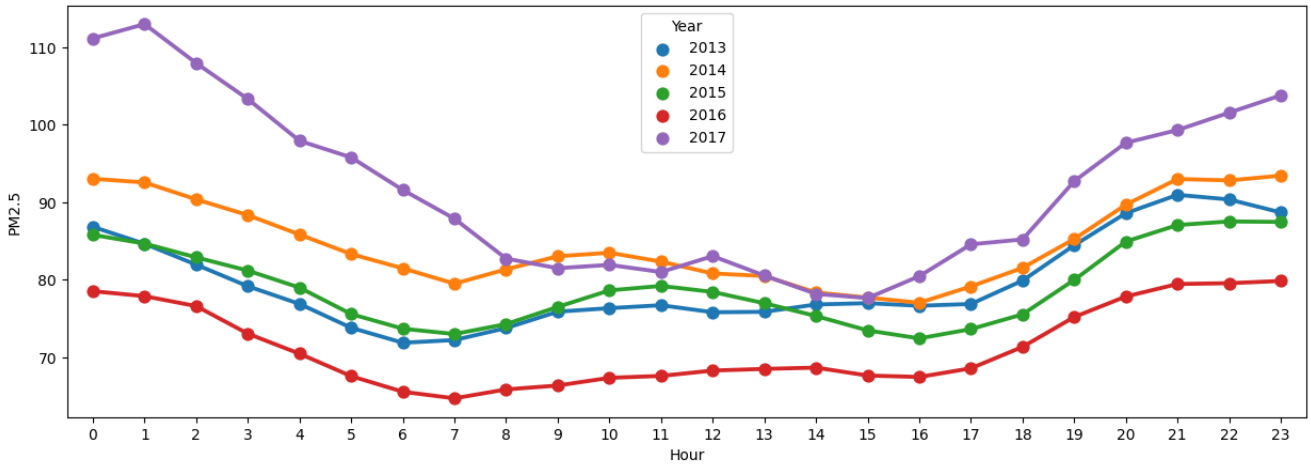
4. Yearly PM2.5 concentration

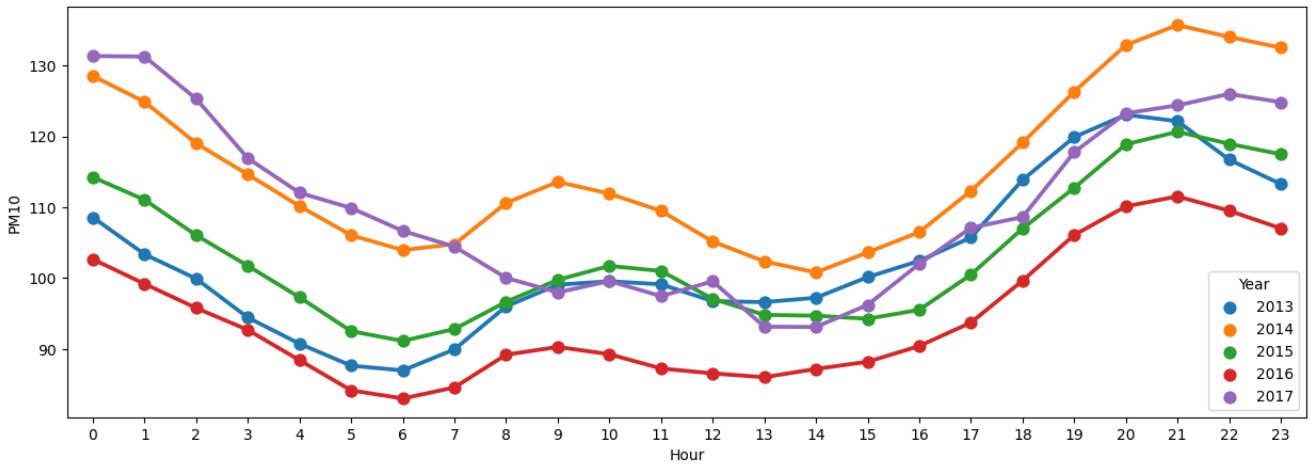


5. Yearly Analysis of Compounds



6. Hourly Analysis of PM2.5 & PM10





Applying Machine Learning Models

1. Linear Regression

```
LinearRegression Mean Absolute Error = 30.175786001352552
LinearRegression Root Mean Square Error = 45.54667403586429
LinearRegression R2 Score = 0.6777887100584677
```

2. Random Forest Regressor

```
RandomForestRegressor Mean Absolute Error = 18.656872517583782
RandomForestRegressor Root Mean Square Error = 30.362195263811937
RandomForestRegressor R2 Score = 0.8566534652602222
```

3. Gradient Boost Regressor

```
GradientBoostingRegressor Mean Absolute Error = 25.587340908851413
GradientBoostingRegressor Root Mean Square Error = 40.212722369687256
GradientBoostingRegressor R2 Score = 0.7488153137971704
```

4. K Neighbors Regressor

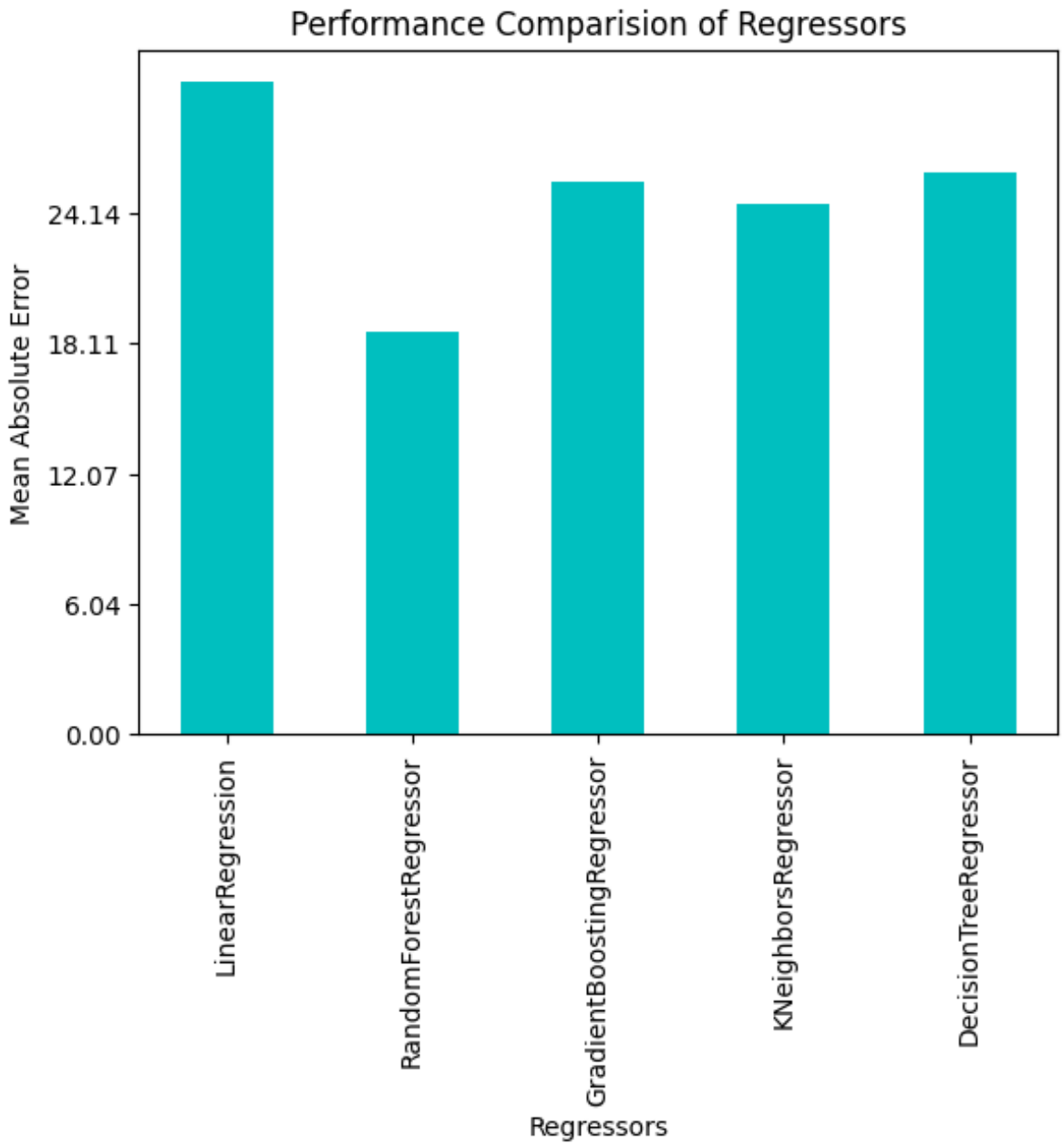
```
KNeighborsRegressor Mean Absolute Error = 24.590591642532065
KNeighborsRegressor Root Mean Square Error = 40.0439525123241
KNeighborsRegressor R2 Score = 0.7509415441654228
```

5. Decision Tree Regressor

```
DecisionTreeRegressor Mean Absolute Error = 25.942965691734827
DecisionTreeRegressor Root Mean Square Error = 44.606252092333804
DecisionTreeRegressor R2 Score = 0.6903309894412163
```

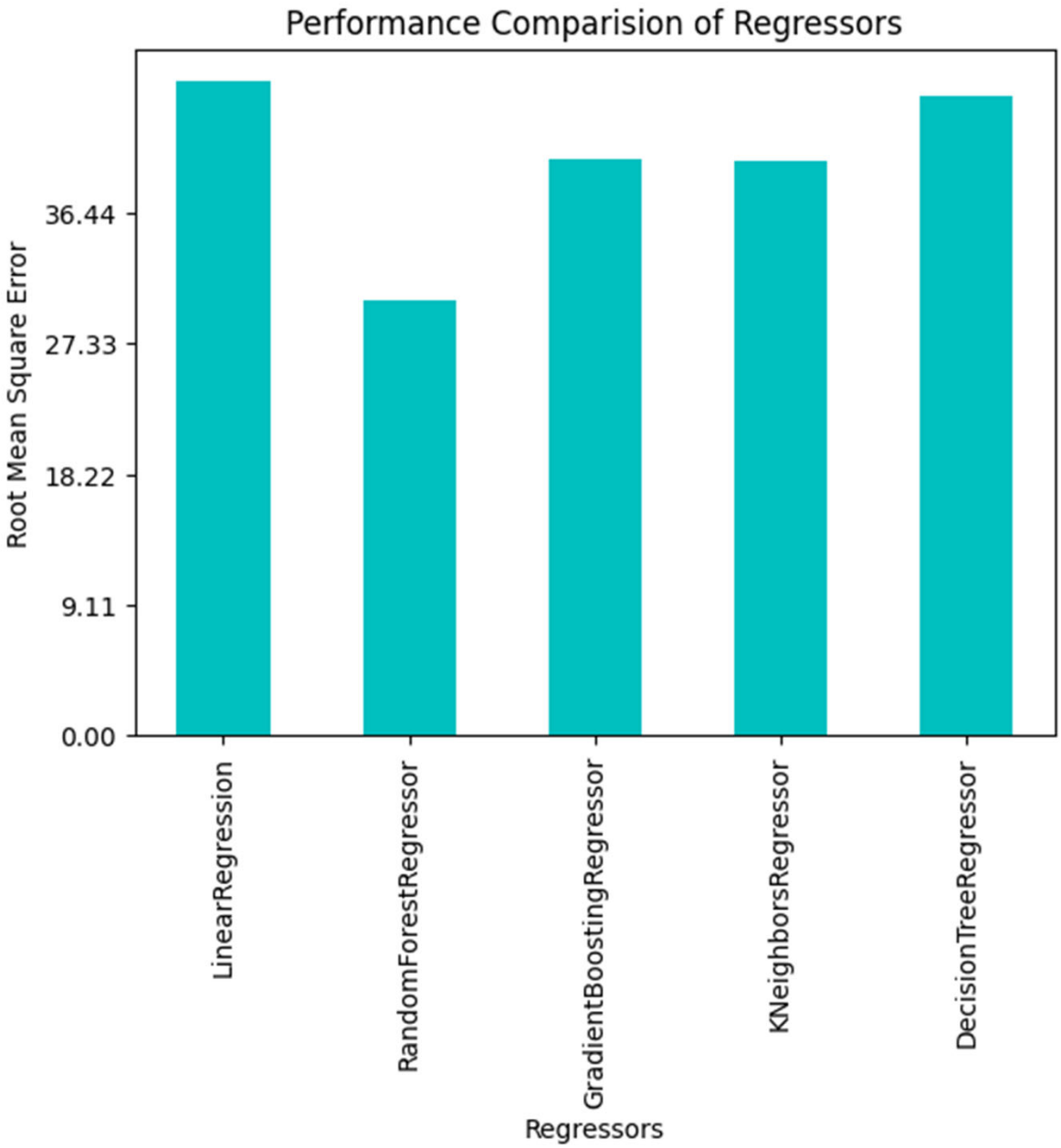
Performance Comparison of Regressors

1. Mean Absolute Error



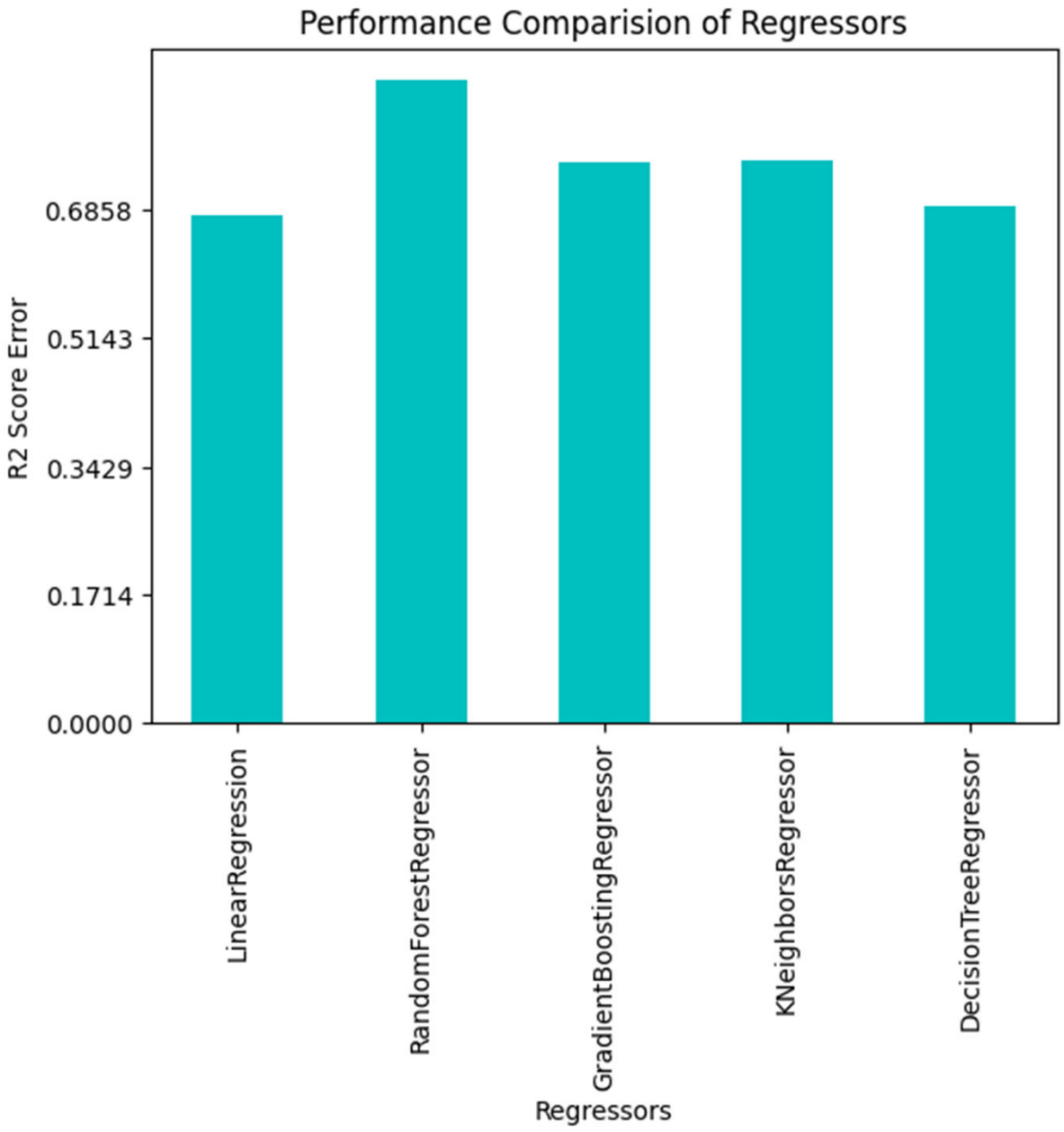
We can see that **RandomForestRegressor** is giving least Mean Absolute Error on given dataset.

2. Root Mean Square Error



We can see that **RandomForestRegressor** is giving least Root Mean Square Error on given dataset.

3. R2 Score



We can see that **RandomForestRegressor** is performing best on given dataset.

Conclusion

In this study, we focused on predicting air quality in Beijing using the "Beijing Multi-Site Air-Quality Data" dataset through a regression approach. By leveraging machine learning techniques, we aimed to develop a model capable of accurately estimating PM2.5 pollutant based on various features.

Throughout our project, we explored different regression algorithms and evaluated their performance using appropriate metrics.

Our developed regression model achieved promising performance, effectively capturing the complex relationships between different factors influencing air pollutant concentrations.