Air Quality Prediction With Machine Learning Algorithms

İzzet Ahmet , İlbey Gülmez

Abstract

In the existing literature on air quality, there are numerous studies focused on developing new formulas for calculating Air Quality Index (AQI). The study tries to help calculate AQI at a lower cost by focusing on estimating CO values related to air quality without the need for measurement tools and at minimum cost. The primary aim is to quickly obtain CO data using machine learning methods such as Support Vector Regression, KNN Regression, Multiple Linear Regression, DTR and RFR rather than calculating the Air Quality Index (AQI). The study compares the effectiveness of reference analyzers (GT) and nominally targeted (PT08) meters. The success rate of the study was reported as 69% and the mean square error (MSE) was reported as 29%.

1. Introduction

Air pollution has emerged as a worldwide concern, causing diverse health and environmental challenges. The rise in industrialization, agricultural practices, urban development, and heightened fuel consumption driven by recent technological progress has resulted in environmental issues like water, noise, and air pollution. Air pollution, in particular, directly affects human health through pollutants and particulate matter. Consequently, there is a growing interest within the scientific community to research air pollution and its consequential effects. [1].

Researchers have recognized that historical monuments face potential deterioration due to the adverse effects of air pollution [2]. The presence of greenhouse gases has a detrimental impact on climate conditions, influencing plant growth negatively [3]. Emissions of inorganic carbons and greenhouse gases further disrupt plant-soil interactions [4]. Climatic fluctuations not only impact human health and animal life but also exert significant influence on agricultural aspects and productivity [5]. This situation leads to economic losses, with estimates from the American Lung Association indicating an annual cost of approximately 37 billion dollars in the US related to illnesses associated with air pollution. California alone accounts for 15 billion dollars of this total [6].

Conventional approaches relying on probability and statistics are intricate and comparatively less efficient. The effectiveness of Machine Learning models in predicting the Air Quality Index (AQI) has been demonstrated to be more dependable and consistent. The advent of advanced technologies and sensors has simplified and enhanced the precision of data collection. Achieving accurate and reliable predictions with extensive environmental data demands thorough analysis, a task efficiently managed by ML algorithms [7].

This research aimed to develop predictive models for the concentration of CO in the air, utilizing the Air Quality Dataset specific to the city of Cassino in Italy. Various commonly employed regression analysis techniques, such as KNN Regression, Support Vector Regression (SVR), and Multilinear Regression, were utilized. Our approach involved constructing hourly models for each measurement related to pollutants and particulates, in addition to a model predicting the hourly Air Quality Index (AQI) for Cassino. The primary objective of these models was to forecast air pollution levels in the region with minimal reliance on measurement instruments, ensuring more cost-effective and accurate predictions. This facilitates the anticipation of potential adverse effects on community health and the timely detection of such impacts, duly reported.

The second section of the study provides an analysis of machine learning methods employed in the literature for air quality prediction, accompanied by a literature review. Section 3 details the dataset under examination, outlining preprocessing steps and feature selection techniques. Section 4 includes a comparison of data resulting from experiments conducted with different models and parameters. Finally, in the concluding section, the obtained results are thoroughly discussed, leading to conclusive remarks.

2. Related Works

Various methodologies have been explored in the literature about the Air Quality Index (AQI), and these can be broadly classified into three main categories. The first involves the analysis of sensor data through techniques like time

series analysis. The second pertains to the examination of visual data, while the third encompasses the study of data generated by a fusion of both sensor and visual data. These approaches and their interconnections are visually represented in Figure 1 [8].

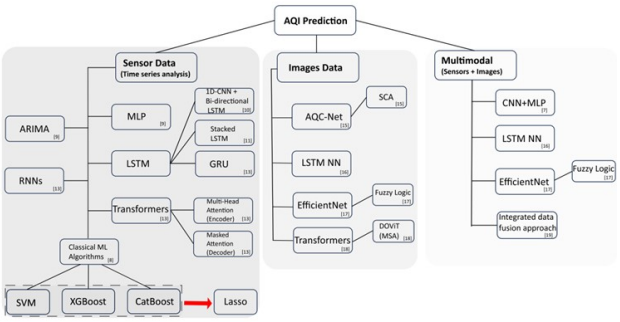

Figure 1. Types of data used in AQI estimation

In our research, we utilize sensor data, aligning with the prevailing trend in existing literature where this type of data is predominantly employed for air quality prediction. For instance, Sanjeev et al. [9] analyzed a dataset incorporating pollutant concentrations and meteorological factors. The authors trained machine learning models, including the Random Forest classifier, and emphasized that Random Forest demonstrated superior performance with reduced susceptibility to overfitting. Their models were trained on sensor data, encompassing timestamp and location details, humidity and temperature readings, along with comprehensive weather data.

Castelli et al. [10] aimed to forecast levels of pollutants and particulates in California by employing the Support Vector Regression machine learning algorithm. They asserted the introduction of a novel method for modeling hourly atmospheric pollution.

In a study by Madhuri et al. [11], the critical influence of wind speed, wind direction, humidity, and temperature on air pollutant concentration was emphasized. Utilizing supervised Machine Learning techniques, they predicted the AQI, highlighting the RF algorithm's minimal classification errors.

Liu et al. [12] analyzed Beijing's AQI from 2019 to 2021, relying on sensor data. They trained two distinct models, ARIMA and Artificial Neural Networks, on this sensor data and found that NNs outperformed ARIMA. The authors suggested the suitability of sensor data, noting that the ARIMA model faced challenges in handling complex, non-linear relationships, extracting pertinent features, modeling temporal dependencies, and adapting to changing conditions.

3. Methodology

3.1. Dataset

The dataset [15] encompasses 9358 instances of hourly averaged responses recorded by 5 metal oxide chemical sensors in an Air Quality Chemical Multisensor Device. Deployed in an environmentally challenged Italian city from March 2004 to February 2005, this dataset stands as the lengthiest collection of freely accessible recordings captured by on-field air quality chemical sensor devices. A co-located certified analyzer supplied Ground Truth concentrations for CO, Non-Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx), and Nitrogen Dioxide (NO2). This dataset includes information from both a reference analyzer and sensors tailored for specific parameters, providing a comprehensive resource for the analysis of air quality.

3.2. Preprocess

In our dataset, missing values are denoted as -200, and we opted to replace these values with the median to address the gaps in the data. This choice was motivated by the specific advantages offered by using the median in comparison to the mean or other statistical measures when handling missing values. Notably, in datasets containing outliers, utilizing the mean may result in increased sensitivity to the influence of these outliers. The median, representing the middle value in an ordered data sequence, exhibits greater resistance to the impact of outliers. Furthermore, in cases where the dataset's distribution is asymmetrical or when working with ordinal data, the median proves to be a more reliable measure. However, the selection of the most appropriate measure may vary depending on the dataset's characteristics. Hence, we conducted a comparison between using the mean and median, revealing no significant variation in prediction outcomes (Fill nan values with median approximately increases our model [3-10]% better than mean) [16].

Z-Score Normalization is a technique employed to standardize values in a dataset by utilizing the standard deviation and mean. One notable advantage of adopting this normalization method is its robustness in handling outliers. By mitigating the impact of outliers, Z-Score contributes to greater consistency and balance in the dataset. Furthermore, with Z-Score Normalization, the dataset values are scaled to achieve a mean of 0 and a standard deviation of 1, facilitating more meaningful comparisons and analyses. Particularly suitable for certain machine learning algorithms, Z-Score normalization, especially in regression algorithms, demonstrates improved performance. Thus, the decision to employ Z-Score Normalization in this dataset not only addressed imbalances but also enhanced the effectiveness of regression models.

3.3. Examining Correlation

Upon examining the correlations of features with CO, it is apparent that each of them displays a significant correlation with the levels of CO. However, it is crucial to emphasize that correlation does not inherently imply causation.

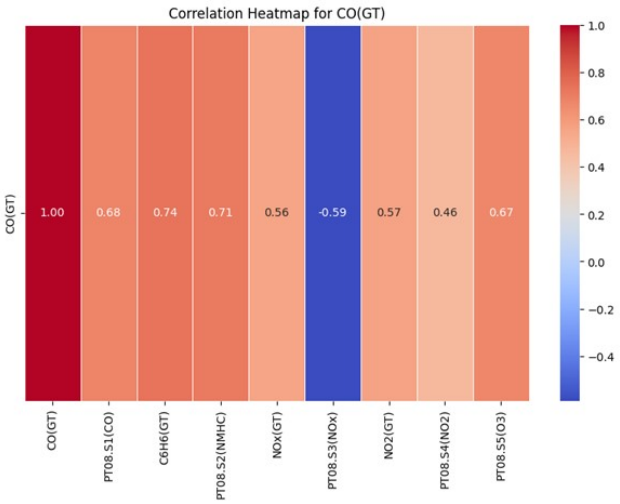

Figure 2. Correlations of other futures with CO

3.4. Evaluation Metric

In our analysis, we utilized R2 (R-squared), Mean Squared Error (MSE), and Mean Absolute Percentage Error (MAPE) as evaluation metrics to gauge the effectiveness of our models. R2, a commonly employed metric in regression tasks, offers insights into the extent to which the model explains variance in the dependent variable. With an interpretation ranging from 0 to 1, it aids in assessing the goodness of fit. Conversely, MSE calculates the average squared differences between predicted and actual values, providing a comprehensive measure of predictive accuracy. Meanwhile, MAPE, expressed as a percentage, quantifies the average percentage difference between predicted and actual values, making it particularly valuable for assessing errors in terms of relative accuracy. By incorporating this array of metrics, our goal is to comprehensively evaluate the predictive power and accuracy of our models, considering both absolute and relative performance aspects.

3.5. Approach

In the upcoming section, our initial objective is to partition our dataset into two segments: GT and PT08, and conduct distinct analyses for each. As we scrutinize the outcomes, we aim to discern which dataset can contribute more effectively to the development of an improved model. By drawing insights from the results obtained, we intend to pinpoint variables that hold greater significance for subsequent research endeavors. This strategic approach allows us to achieve more precise results in future studies while utilizing fewer measurement tools. To execute this strategy, we employed commonly used machine learning models for regression, optimizing parameters within specified intervals (This process will be elucidated in the subsequent section). Another facet of our analysis involves examining prediction results on a seasonal basis and providing interpretations based on these outcomes.

4. Experimental Result

Within this section, our objective is to ascertain whether nominally targeted sensors or reference analyzers yield superior results in predicting the quantity of CO. To accomplish this, we employ regression analysis methods, given that our model is constructed from data collected on an hourly average basis over one year. Before initiating the model creation process, the dataset has undergone a division into training and testing sets, with 80% of the data allocated for training and 20% for testing purposes.(Due to the insufficient size of our dataset, we opted not to use a validation set. Instead, by splitting the data into an 80-20 ratio, we achieved higher accuracy values, ranging up to 5%, depending on the model. We made a trade-off by obtaining a slightly less generalized model to achieve a reasonably high level of model accuracy.)

4.1. Hypertuning On Annual Data

We conducted an empirical evaluation of our model using varying k values on the reference analyzer data, systematically assessing the resultant accuracy and Mean Squared Error (MSE) outcomes.

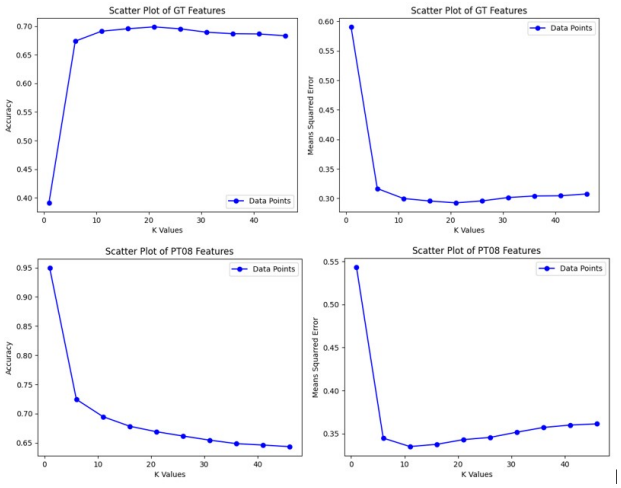

Figure 3. Accuracy and MSE values for different K

As depicted in the graphics illustrated in Figure 3, the optimal k value is determined to be 21, resulting in an R2 score of 0.69 and an MSE of 0.29 for GT features. With an increase in the k value, there is a reduction in underfitting, leading to an improvement in the model's performance. Conversely, for PT08 features, the optimal k value is found to be 1, yielding an accuracy of 0.95 and an MSE of 0.54, indicating potential overfitting at this k value. The identification of the lowest MSE value at k = 11 (selected by considering minimum MSE) for both GT and PT08 features results in an accuracy of 0.69 and an MSE of 0.34.

An R2 score of 0.69 suggests a commendable level of predictive performance, and with an MSE of 0.34 in the context of regression models, a lower MSE is preferable, and 0.34 is considered moderately low. This signifies that, on average, the model's predictions exhibit a relatively small squared difference from the actual values. Therefore, the KNN model demonstrates a reasonably good ability to make accurate predictions, effectively capturing the underlying patterns in the data with a moderate level of precision.

4.1.1. Multilinear Regression Analysis

The R2 score of Linear Regression with GT features stands at 0.62, accompanied by an MSE of 0.37. For PT08 features, these metrics are slightly inferior, with accuracy values ranging from 0.52 to 0.45, presenting worse results compared to KNN. Nevertheless, it can be asserted that GT features yield better outcomes compared to PT08 features.

In the dataset analysis, the observed outcome indicating Multilinear Regression's suboptimal performance relative to KNN suggests potential intricacies in the relationships within the dataset. Multilinear Regression predominantly focuses on capturing linear relationships between independent and dependent variables. In contrast, KNN, being a more flexible model, can discern intricate, non-linear patterns. If the relationships in the dataset deviate from linearity or involve non-linear features like interactions and changing slopes among independent variables, non-linear models such as KNN may exhibit superior performance.

Moreover, the presence of outliers or instances where the assumptions of the regression model are violated can impact the effectiveness of Multilinear Regression. Hence, the indication of Multilinear Regression performing less favorably than KNN underscores the possibility that the dataset's complexity might be better explained by non-linear models.

4.1.2. Decision Tree

We systematically investigated the spectrum of R2 score values employing 500 distinct randomly generated trees within the Decision Tree model. The objective of this analysis was to discern the potential of the Decision Tree in attaining a satisfactory R2 score for the given dataset. Consequently, the R2 scores for PT08 features were found to lie within the interval [0.34, 0.39], whereas those for GT features fell within [0.48, 0.51]. This observation suggests that a randomly generated Decision Tree may not be deemed suitable for effectively modeling PT08 or GT features in this context.

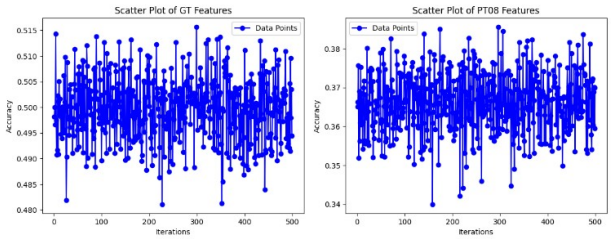

Figure 4. Accuracy and MSE values for randomly generated 500 of Decision Tree Regressor model for GT and PT08 features

4.1.3. Random Forest

Random Forest operates by using multiple decision trees. Decision trees can be considered as tree-like structures created based on specific conditions on input features. With its randomization, voting, and resistance to overfitting, we expect Random Forest to be more generalizable and to have a higher R2 score compared to individual decision trees. To observe this, we individually experimented with the number of estimators in the range [0, 100], expecting higher R2 scores. For GT features, the R2 score becomes greater than 0.66 when the number of estimators is greater than 15, while for PT08, it becomes greater than 0.61 when the number of estimators is greater than 15.

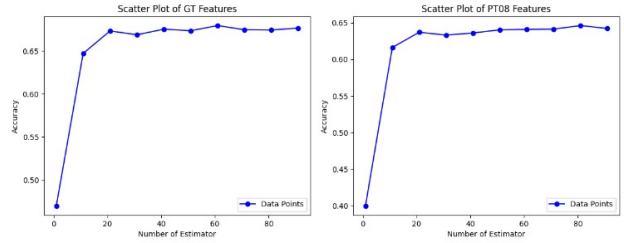

Figure 5. Accuracy and MSE values for different estimator numbers of Random Forest model for GT and PT08 features

4.1.4. SVR

Following the graph presented in Figure 6, SVR for PT08 features has a maximum accuracy value (which is 0.625) with C = 75 and epsilon = 0.35 values and has the minimum MSE value (which is 0.356) for those values. Which is still smaller than KNN Regression model with k = 11

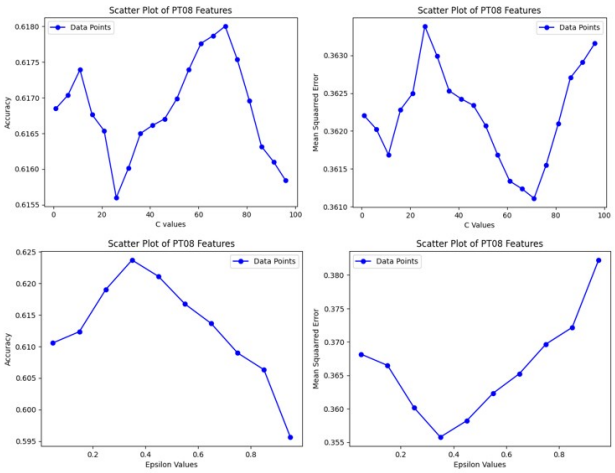

Figure 6. Accuracy and MSE values for different C and epsilon parameters of SVR model for PT08 features

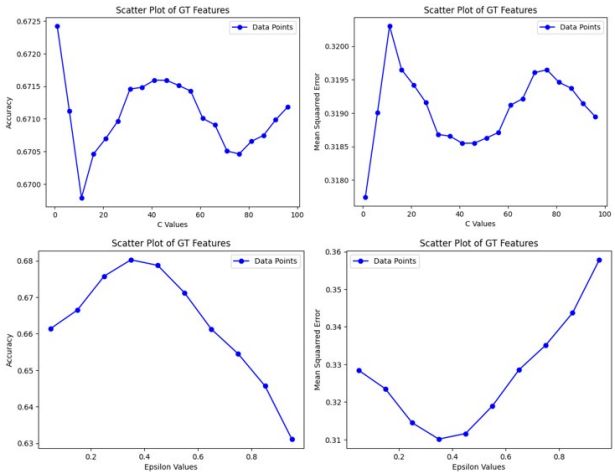

Figure 7. Accuracy and MSE values for different C and epsilon parameters of SVR model for GT features

By the data presented in Figure 7, SVR for GT features has a maximum accuracy value (which is 0.68) with C = 40 and epsilon = 0.35 values and has the minimum MSE value (which is 0.356) for those values. Which is still smaller than KNN Regression model with k = 21

4.2. Test on Seasonal Data

To assess the seasonal performance of the five machine learning models trained on annual data, we conducted rigorous tests using randomly selected seasonal datasets. Rather than training the models separately for each season, we opted for a holistic approach. Dividing the data into four seasonal segments could potentially lead to a reduction in dataset size, causing the model's predictive values to deviate. To mitigate this, we chose to test the models comprehensively on annual data, aiming to understand their responses to varying environmental conditions. To do this better, we first examined the distributions of the data. These distributions are given in Table 1 and Figure 8.

Table 1. Metrics of season data

SEASON	VARIANCE	MEAN	MEDIAN
SPRING	1.00	0.063	-0.09
SUMMER	0.73	0.22	-0.09
FALL	1.038	0.13	-0.09
WINTER	1.1	0.09	-0.09

When we look at the median values of the seasons, we observe that they all have exactly the same value, but the variance and mean values vary. This shows us that the data is distributed across balanced seasons but that outliers are effective. Variance values being equal to 1 (except for summer) is another indication that the data is evenly distributed. We can also see these results in Figure 8. While there is a balanced distribution for other seasons, there are few positive values for summer.

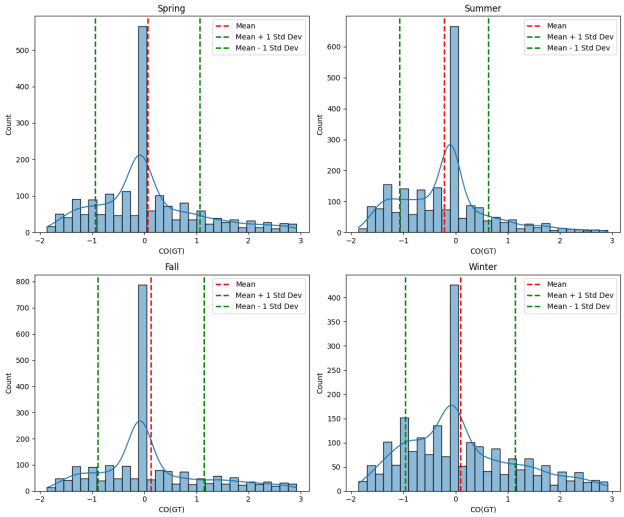

Figure 8. Distribution of data for each season

Our objective was not only to prevent potential biases induced by seasonal divisions but also to observe how the models, originally trained on annual data, would adapt to different environmental changes. The R2 score values, a key metric for model evaluation, were obtained through seasonal testing for the five distinct models trained on annual data, leveraging GT features. These scores are visually rep-

resented in Figure 9. Furthermore, Figure 10 showcases the R2 score values corresponding to PT08 features.

This strategic testing methodology provides valuable insights into the models' robustness across seasons and their adaptability to diverse environmental scenarios.

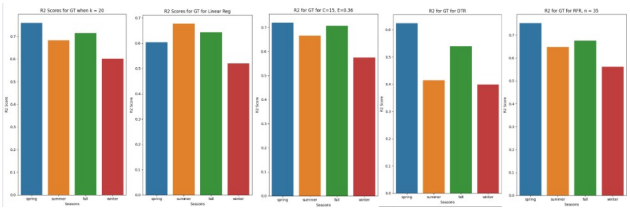

Figure 9. GT features R2 scores for each season for each model (Spring, Summer, Fall, and Summer plotted respectively)

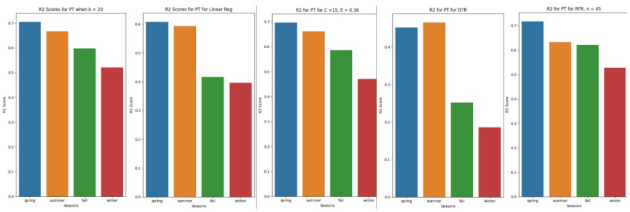

Figure 10. PT features R2 scores for each season for each model (Spring, Summer, Fall, and Summer plotted respectively)

When we examine the R2 score distribution, our models work with very high accuracy for spring, but with low accuracy for winter. The reason may be better explained by someone who knows the behavior of particles in the air better, but since we do not have knowledge on the subject, we think it is not right to make a definitive comment, as a data scientist, we can say that the correlations of seasonal features with CO may be an effective factor here.

Table 2. Correlation based on seasons for PT08 features

SEASON	TOTAL CORRELATION
SPRING	0.75
SUMMER	0.75
FALL	0.67
WINTER	0.68

While the outlined rationale holds its validity admirably for PT08 features, its precision wavers when applied to GT features. Intriguingly, the correlation-based analysis, while indicating a lower correlation during the Fall season, did not manifest as distinctly as anticipated when juxtaposed with Summer. This nuanced observation highlights the intricate nature of the relationship between GT features and seasonal

Table 3. Correlation based on seasons for GT features

SEASON	TOTAL CORRELATION
SPRING	0.78
SUMMER	0.81
FALL	0.71
WINTER	0.66

variations. The discrepancy in the expected and observed outcomes prompts a deeper exploration into the underlying dynamics of GT features across different seasons, necessitating a more refined analytical approach.

To conduct a more detailed analysis of seasonal data, we selected the top two models with the highest R2 score and the lowest error values from our annual data. These models happened to be the K-Nearest Neighbors (KNN) and Random Forest Regression models. Below are the tables associated with these models:

K-Nearest Neighbors (KNN) Model:

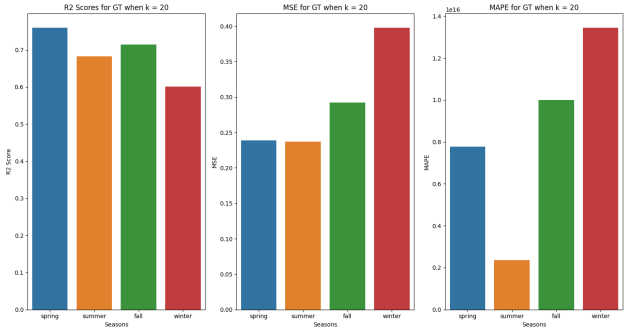

Figure 11. KNN results for each season (R2, MSE, MAPE respectively)

In our in-depth scrutiny of the KNN model, informed by the extensive assessments detailed earlier, our expectations align harmoniously with the empirical outcomes. We foresaw a notable uptick in accuracy accompanied by minimal error rates for the spring and summer seasons. In stark contrast, a subdued accuracy level and heightened error rates were anticipated for the winter season. Encouragingly, the actual results impeccably mirror our prognostications.

Random Forest Regression Model:

The trend predicted in the correlation part is observed more consistently in the Random Forest Regression (RFR) model rather than KNN. Notably, the RFR model exhibits better precision in predictions during the spring and summer seasons, demonstrating a closer alignment with actual values. Conversely, the predictive efficacy tends to diminish for

the fall and winter seasons, manifesting as more distanced predictions.

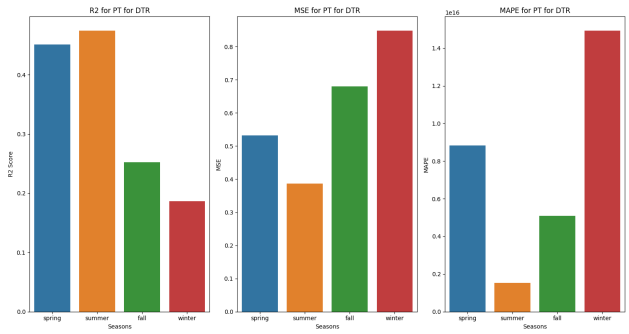

Figure 12. Random Forest results for each season (R2, MSE, MAPE respectively)

This parallel behavior in both models, KNN and RFR, accentuates the seasonally-dependent dynamics influencing their predictive capabilities. The concurrence in the observed patterns underscores the robustness of the identified trend, offering valuable insights into the nuanced interplay between machine learning models and distinct seasonal variations.

5. Conclusion

In our comprehensive evaluation of the performance of five distinct machine learning models applied to GT and PT08 futures, it is discerned that GT futures consistently exhibit superior efficacy in the CO calculation process across all models. This observed proficiency in GT data may be attributed to its direct referencing of an elemental entity and the concomitant precision in measurements. Leveraging GT data for predictive modeling in CO calculations thus yields outcomes characterized by heightened accuracy.

Upon scrutinizing the outcomes through a seasonal lens, a discernible pattern emerges. Forecasts generated during the spring exhibit a heightened R2 score and diminished error values, in stark contrast to those formulated during the winter season, which display a notably diminished R2 score coupled with elevated error values. The rationale behind these seasonal variations may stem from the dynamic correlation between CO gas and other atmospheric constituents, which exhibits fluctuations contingent on seasonal changes. While the precise cause necessitates further investigation, we postulate that the observed divergence may be indicative of divergent air pollution levels in winter and summer.

5.1. Suggestions For Further Results

Proposing a prospective avenue for research, we advocate the determination of seasonal variations in air pollution through AQI measurements, positing that such an inquiry would yield invaluable insights for subsequent studies. Implicit in this proposal is the contention that crafting predictive models contingent on seasonal stratification holds the promise of enhancing forecast accuracy and efficacy. This methodological approach underscores a nuanced understanding of air pollution dynamics and accentuates the necessity of accounting for seasonal intricacies in predictive modeling endeavors.

Furthermore, it is noteworthy to highlight that the incorporation of Temperature, Relative Humidity, and Absolute Humidity features has resulted in a discernible reduction in the model's accuracy. Through empirical measurements, we have observed an average decrease of approximately 10 percent in accuracy rates.

Table 4. Annual model results for GT features

MODEL	PARAMETERS	R2	MSE	MAPE
KNN	K = 20	.7	.29	8.3
MLR	-	.62	.37	9.4
SVR	C = 20,E = .35	.68	.31	.65
DTR	-	.49±.02	.48±.02	.48± .02
RFR	N = 20	.67	.325	7.1

Table 5. Annual model results for PT08 features

MODEL	PARAMETERS	R2	MSE	MAPE
KNN	K = 10	.69	.34	7.2
MLR	-	.52	.44	1.2
SVR	C = 75, E = .35	.622	.355	7.25
DTR	-	.36± .02	.6± .02	.6± .02
RFR	N = 20	.64	.35	7.0

Acknowledgements

We would like to express our sincere gratitude for the invaluable support received from the ChatGPT language model developed by OpenAI during the composition of this article. The assistance provided in refining the language and ensuring a more academic tone has been instrumental in enhancing the overall quality of our work. We extend our appreciation to the developers and contributors involved in the creation of this advanced language model

6. References

[1] U. A. Hvidtfeldt, M. Ketzel, M. Sørensen et al., "Evaluation of the Danish AirGIS air pollution modeling system against measured concentrations of PM2.5, PM10, and black

carbon," Environmental Epidemiology, vol. 2, no. 2, 2018.

[2] Rogers CD (2019) Pollution's impact on historical monuments pollution's impact on historical monuments. SCIENCING. https://scien cing.com/about-6372037-pollution-s-impact-historical-monum ents.html

[3] Fahad S, Sönmez O, Saud S, Wang D, Wu C, Adnan M, Turan, V (2021a) Plant growth regulators for climate-smart agriculture (1st ed.). CRC Press. https://doi.org/10.1201/9781003109013

[4] Fahad, S, Sönmez O, Saud S, Wang D, Wu C, Adnan M, Turan V (2021b) Sustainable soil and land management and climate change (1st ed.). CRC Press. https://doi.org/10.1201/9781003108894

[5] Sönmez O, Saud S, Wang D, Wu C, Adnan M, Turan V (2021) Climate change and plants: biodiversity, growth and interactions (S. Fahad, Ed.) (1st ed.). CRC Press. https://doi.org/10.1201/978100310893

[6] B. Holmes-gen and W. Barrett, Clean Air Future, Health and Climate Benefits of Zero Emission Vehicles, American Lung Association, Chicago, IL, USA, 2016

[7] Air pollution prediction with machine learning: a case study of Indian cities K. Kumar, B. P. Pande, https://doi.org/10.1007/s13762-022-04241-5

[8] Deep learning based multimodal urban air quality prediction and trafc analytics Saad Hameed , Ashadul Islam , KashifAhmad , Samir Brahim Belhaouari , Junaid Qadir AlaAl Fuqaha

[9] Sanjeev D (2021) Implementation of machine learning algorithms for analysis and prediction of air quality. Int. J. Eng. Res. Technol. 10(3):533–538

[10] Castelli M, Clemente FM, Popovi ˇc A, Silva S, Vanneschi L (2020) A machine learning approach to predict air quality in California. Complexity 2020(8049504):1–23. https://doi.org/10.1155/ 2020/8049504

[11] Madhuri VM, Samyama GGH, Kamalapurkar S (2020) Air pollution prediction using machine learning supervised learning approach. Int J Sci Technol Res 9(4):118–123

[12] 9. Liu, T. You, S. Analysis and forecast of beijing's air quality index based on arima model and neural network model. Atmosphere 13, 512 (2022)

[13] https://archive.ics.uci.edu/dataset/360/air+quality