**Estimation Concentrations of Air Pollutants Using Machine Learning**

Mr. Shubham Sonake1, ,Dr. A.D. Sawarkar2

*1*Department of Information Technology, Shri Guru Gobind Singhji Institute of Engineering and Technology (SGGSIET), Nanded.

[*sonakeshubham1817@gmail.com*](mailto:sonakeshubham1817@gmail.com)

*2*Department of Information Technology, Shri Guru Gobind Singhji Institute of Engineering and Technology (SGGSIET), Nanded.

[*adsawarkar@sggs.ac.in*](mailto:adsawarkar@sggs.ac.in)

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

**Abstract:** Air pollution has become a growing concern globally, impacting human health and environmental well-being. Accurately predicting air quality is key in implementing preventive measures and mitigating its adverse effects. This study investigates the effectiveness of regression models for machine learning in forecasting air pollutant concentrations. Our research utilizes a dataset containing various air quality parameters, including CO (carbon monoxide), C6H6 (benzene), NOx (nitrogen oxides), and O3 (ozone), alongside meteorological elements such as temperature, relative humidity, and atmospheric pressure. For instance, CO exposure can disrupt oxygen delivery in the bloodstream, leading to headaches, dizziness, and nausea at high concentrations. Benzene, a known carcinogen, is linked to various health problems, including leukemia, anemia, and respiratory issues.

Models of regression for machine learning are employed to create connections among these variables and predict future air pollutant levels. The chosen model will be evaluated based on its accuracy in replicating real-world data. This research aims to demonstrate the capability regression in machine learning in air quality prediction. By developing a reliable prediction model, we can contribute to proactive air quality management strategies and minimize the air pollution dangers to health pollution.

**Keywords**: Machine Learning, Linear Regression, Prediction, Random Forest, Decision tree.

**1. Introduction**

One big worry is air pollution ,with negative impacts on human health and the environment. The air quality forecast fluctuations can enable individuals and authorities to take countermeasures [[1]](#first). particulate matter (PM), Ozone (O3), carbon monoxide (CO), nitrogen oxides (NOx), sulfur dioxide (SO2), volatile organic compounds (VOCs), metals, and pesticides are among the main outdoor air pollutants in metropolitan regions [[2]](#first). Vehicles release carbon monoxide (CO), which is created when wood and coal are burned. This poisonous, colorless, and odorless gas is major source of air pollution since it helps to create smog. Effective air pollution management requires an understanding of CO levels and the ability to use algorithms based on machine learning for their prediction. [[3]](#first).

Because of different generation and removal methods, high humidity is typically linked to high concentrations of some air pollutants (such CO, PM, or SO2) but low concentrations of other air pollutants (like O3 and NO2) [[4]](#first).Predictive modelling frequently uses Random Forests, Decision Trees, and Linear Regression are examples of machine learning regression techniques. [[5]](#first). However, the majority of statistical modelling research to date has been limited to using traditional regression or classification models, ignoring the correlation between sub-models in different time periods or the nature of the problem itself. However, machine learning techniques have been evolving for more than 60 years and have had great success in a number of fields [[6-7]](#first).

These models are evaluated by means of a variety of performance metrics. Common metrics include the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R²) [[8]](#first). While feature engineering entails constructing fresh variables that are able to improve model performance, feature selection focuses on determining the most crucial factors influencing CO levels [[9]](#first). To optimize model parameters and raise overall performance and efficiency, strategies like hyperparameter tweaking with Grid SearchCV can be applied [[10]](#second).

The literature on machine learning-based air quality forecasting will be reviewed in the ensuing parts, which will additionally contain information regarding methodologies, outcomes, experimental design, and discussion of conclusions and future directions. Research proves that machine learning accurately predict air quality for a range of pollutants and geographical locations. Research shows that the accuracy of predictions is much increased when meteorological data and other elements are included [[11]](#second). For example, studies conducted in heavily trafficked urban areas discovered that combining meteorological and traffic data improved the accuracy of CO estimates. Production schedules and emission data are also beneficial to industrial locations. The experimental design of study entails tuning the hyperparameters of several machine learning models, training them on pre-processed data, and assessing them with selected metrics. Cross-validation will ensure the robustness and generalizability of the model. The best algorithms and important variables affecting CO levels will be determined by the results. Ultimately, the conversation will analyze the results within the framework of current research, highlighting useful ramifications for public health and policy. The study attempts to offer information so that others can reduce their exposure to dangerous CO levels and authorities can put preventive measures into place [4]. Subsequent investigations could delve into sophisticated algorithms, integrate real-time data for dynamic forecasts, and expand the structure to anticipate other contaminants.

The following are the study's objectives:

* Examine the Regression in Machine Learning Models' Predictive Power for CO Concentrations:
* Assess the extent to which different regression models for machine learning can learn from the past.
* Analyze the models' precision and dependability for predicting future CO levels.
* Examine and Dissect the outcomes of Different Regression Algorithms for CO Prediction:
* Examine and contrast several machine learning regression methods, Decision Trees, Random Forests and Linear Regression [[12]](#second).
* Determine each algorithm's advantages in relation to CO prediction.
* Analyze each algorithm's applicability in light of variables like robustness, accuracy, and computational economy.

This work aims to support ongoing efforts to control air pollution and safeguard the environment and public health by utilizing machine learning regression models. Accurate CO level prediction has the potential to improve treatments and create a cleaner, healthier environment.

**2. Literature Review**

Veljanowska and Dimosky (2018) [[13]](#second) compared unsupervised neural network algorithms with established supervised methods such as K-nearest neighbour, decision trees support vector machine, etc. While the neural networks appeared to work well, they struggled to predict hourly pollution levels. Similarly, Zhao et al. (2018) [[14]](#second) used recurrent neural networks (RNNs) to quantify pollution at each point in time, benefited from RNN memory capabilities but face limitations in memory-free performance Moharle, Purohit, Patil (2018). [[15]](#second) applied Fuzzy Logic to forecast PM2 and PM10 concentrations, which handle irregularities well but encountered problems with aggregating data resulting in inaccurate information Furthermore, C.R Deshmukh et al. (2018) [[16]](#second) used Autoregression to identify pollution events and linear regression to predict PM2.5 levels, and highlighted the difficulties of adapting to a changing climate and thus, Zhang et al colleagues. (2018) [[17]](#second) proposed Wavelet Neural Networks but struggled in choosing the appropriate parameters, which affected the prediction accuracy. Amado and Dela Cruz (2018) [[18]](#second) combined sensors and neural networks to estimate accurate pollution levels, although there were challenges encountered in optimizing incomplete data In a paradoxical manner Kang and his colleagues. (2018) [[19]](#second) were inspired by the hourly forecasting capabilities of Deep Belief Networks but found issues with sensor data quality. Mejia et al. (2018) [[20]](#second) discovered that the Random Forest functions well for PM10 concentrations but is inaccurate for hazardous pollutants and incomplete data. Together, these studies illustrate the changing nature of air quality forecasting, highlighting the fine trend

3. **Methodology**

3.1 Dataset Description

This study utilizes a publicly available dataset via the Machine Learning Repository at UCI <https://archive.ics.uci.edu/>. The dataset contains air quality measurements collected over one year (March 2004 - February 2005) in a polluted area within an Italian city [[21]](#second).

Source:

* UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/index.php

Characteristics:

* Size: 9357 instances (data points)
* Time Resolution: Hourly
* Location: Field deployment within a polluted area, at road level, in an Italian city
* Target Variables (Primary Focus):
* CO concentration (mg/m^3) - This research will focus on predicting CO levels.
* Possible Additional Target Variables (Future Exploration):
* Non-methane hydrocarbon (NMHC) concentration (ug/m^3)
* Benzene concentration (ug/m^3)
* Total nitrogen oxide (NOx) concentration (ppb)
* Nitrogen dioxide (NO2) concentration (ug/m^3)

While the study's primary focus is predicting CO levels, the dataset offers the possibility to explore additional target variables in future research. This expands the potential applications of the developed models for comprehensive air quality monitoring.

* Sensor Data: Responses from Five chemical sensors made of metal oxide
* Additional Features:
  + Date
  + Time
  + Temperature (°C)
  + Relative Humidity (%)
  + Absolute Humidity
* Data Quality Considerations:
* Missing values are indicated by “-200”.
* Potential for sensor drift and cross-sensitivity between sensors may affect accuracy.

This dataset provides insightful information about the relationship between sensor responses and actual air pollutant concentrations. The inclusion of different environmental elements, such as temperature and humidity allow for creation of further robust models last to estimate the quality of the air utilizing machine learning techniques.

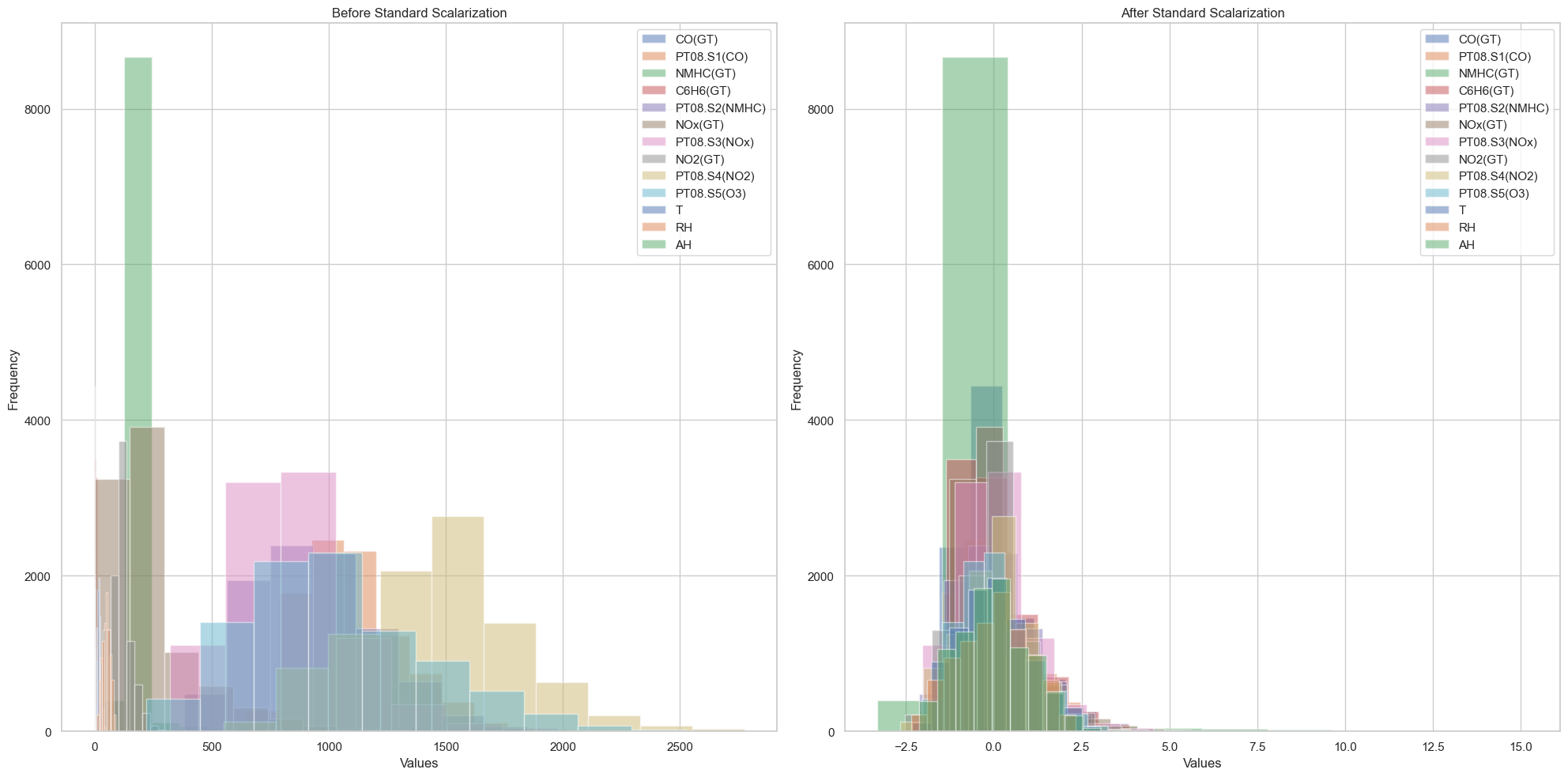
* Partitioning for Testing: A division of dataset was conducted, with 70% allocated for training purposes and 30% reserved for testing, facilitating a thorough assessment of the model's performance [[22]](#second).

3.2 Data Preprocessing:

Before the air quality dataset is used for machine learning, important preprocessing steps are performed to ensure data quality and improve model performance Data cleaning uses annotations marked with

“-200" mark and handles missing values ​​with the mean value of corresponding item. This approach minimizes the impact of missing data while preserving trends in the data structure. Additionally, entries containing completely missing values ​​or irrelevant information can be eliminated after careful consideration.

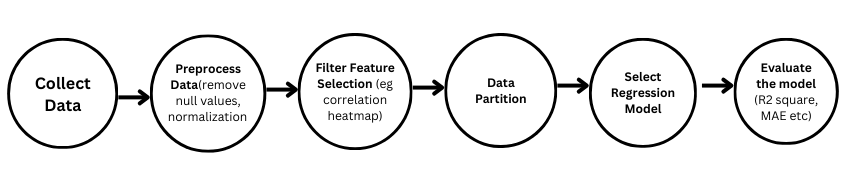
After data cleaning, standard scaling is accustomed to normalize the data. This method changes each item to a Standard Deviation of 0 to 1 [[23]](#second). This guarantees that all items contribute equally to the learning process in model, regardless of their original unit or scale.



*Fig 1.* *Feature Distribution Before and After Scalarization*

The effectiveness of this normalization will be visually demonstrated through histograms, comparing the distribution of features before and after standard scaling. This visualization will highlights impact of scaling on the data spread and how it creates a more balanced foundation for machine learning analysis.

3.3 Workflow Representation



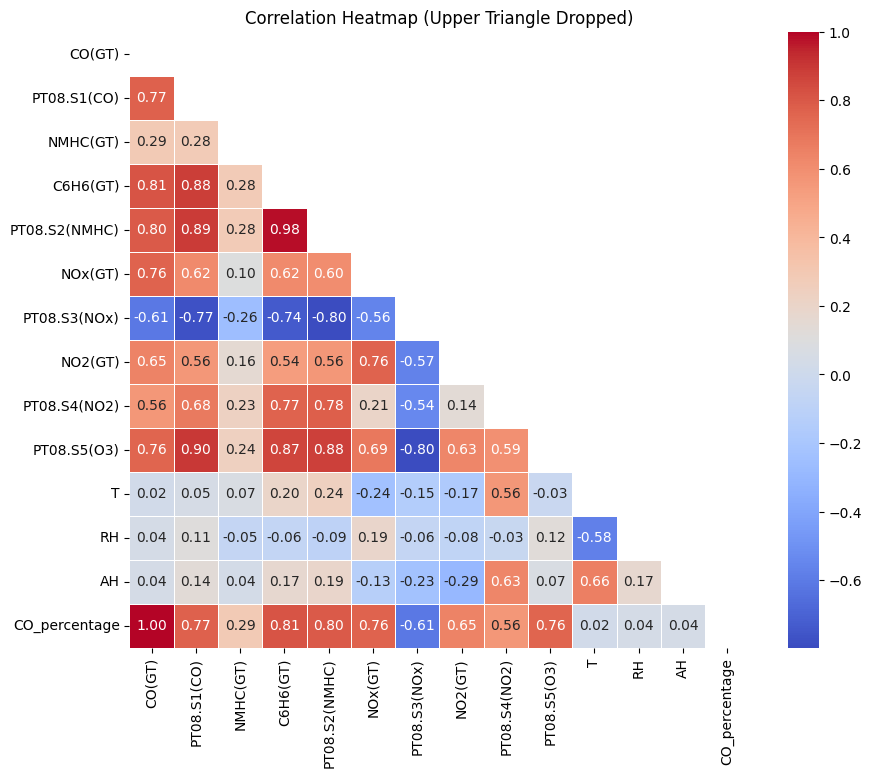
Estimating air quality requires a workflow with multiple important components. First, information is gathered from pertinent sources. After that, preparation is done on the unprocessed information to clean and arrange it for analysis. After that, feature selection is carried out, often using methods such as correlation heatmaps to determine which variables are most significant. Following the selection of the important features, the data is divided as sets for training and for testing. Next, a few regression techniques are chosen for estimating air quality. Lastly, an assessment is conducted on these models to Figure out how they forecast the future. methodical technique guarantees thorough investigation and precise air quality forecasts.

3.4 Model Selection

Here, it is important to select appropriate regression algorithms for accurate prediction of CO concentrations. This choice is influenced by several variables in addition to knowledge obtained via correlation heat map analysis:

* Model interpretability: It's critical to figure out relationship between input characteristics and CO concentrations. Linear regression provides a clear definition of how each component affects the predicted CO concentrations.
* Nonlinearity and Flexibility: Real world relationships may not always be perfectly linear. Decision trees and random forests can capture nonlinear patterns in the data [[24]](#third), Potentially, this could pave the way for more precise forecasts regarding the substantial relationship between air quality and variables and concentrations of CO.
* Generalizability and Overfitting: Our goal is to develop a model that performs well on unobserved data. Linear regression can be prone to overfitting if too many variables exist in the data [[25]](#third). Decision trees and random forests can reduce overfitting by introducing randomness during model construction.

Correlation analysis: Correlation heat maps provide valuable insights into attribute relationships. Significant positive associations among variables may indicate relative importance, potentially affecting model performance. based on heat map observations, available selection methods too many variables exist in this. In addition, the temperature map may reveal weakly negatively correlated factors with CO, which may be less predictively effective and should be considered for removal during model development.



*Fig 2. Correlation Heatmap of Pollutants*

The heat map shows the connection between components (Fig 2), where red indicates strong positive correlations (components together), blue represents strong negative correlations (opposites), and white show weak or no relationships This diagram helps identify redundant and potentially less effective features for prediction CO and model refinement methods Simplifies

Considering these factors, and the detection from the correlation heat map, a mix of algorithms are used in this study:

1) Linear regression: provides a baseline model and simplifies the interpretation of relationships between factors and CO concentrations.

2) Decision tree: captures nonlinear models and provides improved prediction accuracy for complex relationships.

3) Random forest: Combining many decision trees increases generalizability and reduces overfitting [[26]](#third), resulting in the most robust possible forecast of CO concentrations

By analysing these systems and using information from the correlation heat map, our goal is to determine the model that provides the most precise and interpretable results for estimating CO concentrations available in air quality database.

3.4 Evaluation Metrics

To evaluate efficiency of selected regression models which predicts CO concentration, we will utilize set of evaluation metrics. These measures measure how much actual CO values detected in the air quality dataset depart from the projected CO values.

1. R-squared (R²) Score: This metric represents the share of variance in the actual CO concentration that described by the model's predictions [[27]](#third). A higher R² score (closer to 1) indicates a better fit between model and data.

Formula: R² = 1 -

Where:

yi - actual CO value for data point i

ŷi - predicted CO value for data point i by the model

ȳ - average of total actual CO values

1. Mean Absolute Error (MAE): MAE calculates the mean absolute disparity between anticipated CO values and the actual CO values [[28]](#third). A lower MAE signifies a smaller average prediction error.

Formula: MAE =

Where:

n – total numbered data points

yi - actual CO value for data point i

ŷi - predicted CO value for data point i by the model

1. Mean Squared Error (MSE): MSE squares the individual differences between the anticipated and real CO values before calculating the average [[29]](#last). While sensitive to outliers, MSE offers perceptions into the average magnitude of prediction errors.

Formula: MSE =

Where:

n – total numbered data points

yi - actual CO value for data point i

ŷi - predicted CO value for data point i by the model

4)    Root Mean Squared Error (RMSE): RMSE is the square root of Mean Square Error

Formula: RMSE =

Where:

n - total numbered data points

yi - actual CO value for data point i

ŷi - predicted CO value for data point i by the model

4. **Statistical Examination**

This section describes in detail the selected regression models - random forest, linear regression, decision tree - to be applied for CO concentration forecasting The Python programming language formed the basis, using three powerful libraries: panda, scikit-learn, and matplotlib of seaborn extensions included

This study examines the Regression models' predictive efficiency CO concentrations in each dataset.

* Linear regression: This basic method establishes a linear relationship to model the dependence of CO concentrations on other air quality characteristics [[30]](#third). It provides a clear description of how each factor affects the predicted CO concentrations, making it valuable to understand the underlying relationships. Although, linear regression cannot capture complex, nonlinear patterns in the data.
* Decision Tree: This model works by splitting the information into smaller segments based on specific feature values, and ultimately creates a tree-like structure to make predictions. Decision trees are adept at handling nonlinear relationships and are able to identify intricate relationships between multiple factors that may influence CO concentrations. Although definable to some extent, it can be difficult for large data sets.
* Random forest: This clustering method combines the strengths of multiple decision trees. By training a collection of decision trees on small random features and comparing their predictions, random forests provide increased robustness and potentially greater accuracy compared to decision trees for individuals [[26]](#third). They are not as easily understandable as. linear regression but often more interpretable than complex decision trees.

5. **Results and Discussions**

In this investigation, we applied machine learning regression methods to forecast pollutant levels with notable achievement. Our study centered on utilizing these models to anticipate pollutant concentrations using diverse environmental and meteorological variables. The outcomes in table 1 highlighted the effectiveness of our methodology, showing high accuracy in prediction.

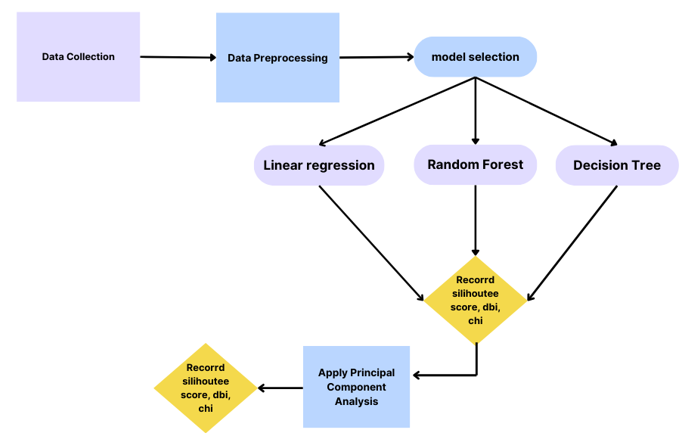
The formulas used to evaluate scores are

* R2 Scores: R² = 1 –
* Mean Absolute Error : MAE =
* Mean Square Error : MSE =
* Root Mean Square Error : RMSE =

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Models\Evaluation | R² Score | MAE | MSE | RMSE |
| Linear Regression | 0.78 | 0.40 | 0.38 | 0.61 |
| Decision Tree | 0.70 | 0.45 | 0.52 | 0.72 |
| Random Forest | 0.84 | 0.33 | 0.28 | 0.52 |

*Table 1. Results of each models*

To Improve Accuracy:

* Dimension Reduction: Principal Component Analysis (PCA) reduces the dimensions of complex data by identifying and retaining the most important patterns or features. It simplifies visualization, aids in feature selection, and enhances model performance in research settings.

After applying dimension reduction to linear regression.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model\Evaluation | R² Score | MAE | MSE | RMSE |
| Linear Regression | 0.84 | 0.33 | 0.28 | 0.52 |

* Hyperparameter Tuning: The process of hyperparameter tuning is methodically determining which set of hyperparameters will optimize a machine learning model's performance. This search is carried out by methods like as GridSearchCV and RandomizedSearchCV, which use cross-validation to assess model performance over a given grid or random distribution of hyperparameter values.

After applying Hyperparameter tuning with Grid SearchCV to random forest.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model\Evaluation | R² Score | MAE | MSE | RMSE |
| Linear Regression | 0.84 | 0.32 | 0.27 | 0.52 |

Therefore, utilizing dimension reduction approaches, the random forest maintained the same R2 score for linear regression but outperformed other models in terms of R2 scores

As a result, random forest model, the best one to employ to forecast future CO levels.

**6. Conclusion:**

The outcomes from various models like the random forest, the decision tree, and linear regression reveal that certain regions are experiencing concerning levels of CO pollution that demand immediate attention.

The average recommended safe level for CO levels is 0.20 ppm per hour. Monitoring key attributes such as SO2 to improve future air quality forecasts is required because they have substantial health implications, with asthma, bronchitis, and other respiratory issues, which data are not sequentially dated, preventing nationwide estimates of CO levels.

Going forward, the focus will shift to forecasting future SO2 levels. , the effort will prioritize the calculation of Air Quality Index (AQI) and the use of classification models to improve air quality monitoring and better address emerging environmental challenges. This review highlights important research areas such as SO2 monitoring and AQI development to guide sustainable environmental management.

**7. References:**

[1] Curtis, L.; Rea, W.; Smith-Willis, P.; Fenyves, E.; Pan, Y. Adverse health effects of outdoor air pollutants. Environ. Int. 2006, 32, 815–830.

[2] Mayer, H. Cities and air pollution. (1999) Atmos. Environ. 33, 4029–4037

[3] 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 2020, pp. 140-145, doi: 10.1109/ICACCCN51052.2020.9362912, T. Madan, S. Sagar, and D. Virmani, "Air Quality Prediction using Machine Learning Algorithms –A Review,"

[4] Dependency of urban air pollution on meteorological, Elminir, H.K. Academia Sinica 2005, 350, 225–237.

[5] Journal of Cleaner Production, Volume 203, 2018, Muhammad Waseem Ahmad, Jonathan Reynolds, Yacine Rezgui, Predictive modeling for solar thermal energy systems: A comparison of support vector regression, random forest, additional trees, and regression trees

[6] Yuan, Z.; Zhou, X.; Yang, T.; Tamerius, J.; Mantilla, R. Predicting Traffic Accidents Through Heterogeneous Urban Data: A Case Study. In Proceedings of the 6th International Workshop on Urban Computing (UrbComp 2017), Halifax, NS, Canada, 14 August 2017.

[7] Collobert, R.; Weston, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 5–9 July 2008.

[8] M Goyal and M Pandey conducted a systematic analysis using ensemble learning to predict energy performance in residential buildings. Arab J Sci Eng 46 (2021), 3155–3168. The doi: 10.1007/s13369-020-05069-2

[9] Guyon, Isabelle, and André Elisseeff. "An introduction to variable and feature selection." Journal of machine learning research 3.Mar (2003): 1157-1182.

[10] Shanthi, D.L., Chethan, N. Genetic Algorithm Based Hyper-Parameter Tuning to Improve the Performance of Machine Learning Models. SN COMPUT. SCI. 4, 119 (2023). https://doi.org/10.1007/s42979-022-01537-8

[11] S. Salcedo-Sanz, C. Casanova-Mateo, and E. G. Ortiz-García. "Accurate precipitation prediction with support vector classifiers: A study including novel predictive variables and observational data." (2014) Atmospheric research 139: 128–136.

[12] Sergio Trilles, Iskandaryan, Ditsuhi, and Francisco Ramos. "Air quality prediction in smart cities using machine learning technologies based on sensor data: a review." 2017. 10.7 (2020): 2401 in Applied Sciences.

[13] Angel Dimoski and Kostandina Veljanovska1 (2018) Predicting the Fair Wind Index with basic machine learning algorithms: an International Journal of Emerging Techniques & Technology in Computer Science (IJETTCS).

[14] Deep vein recurrence network for air quality classification, Zheng-Long Wu, Pei-Chan Chang, Xiaosong Zhao, Rui Zhang, and Yuan Ze University, 2018, Journal of information storage and multimedia signal processing.

[15] Savita Vivek Mohurle, Dr. Richa Purohit, and Manisha Patil evaluated the fuzzy clustering idea for wind measurement. International Journal of Growth Science and Research, Pollution Index, 2018.

[16] Research and Forecast of Air pollution using machine learning modelsNayana DK, Aditya CR, Chandana R Deshmukh, and Praveen Gandhi Vidyavastu, 2018, international Journal of Engineering Trends & Technology (IJETT).

[17] based Jianxiang Mei, Shan Zhang, Yang Li, and Xiaoli Li on estimates of PM2.5 concentrations in citiesIEEE, 2018; Wavelet Neural Networks.

[18] Jennifer C. Dela Cruz and Timothy M. Amado, Improvement of winds quality control and quality by machine learning-based predictive modeling, 2018, IEEE

[19] Benzene Air Pollution Analysis Modeling ANN and SVM, Ali Rodan, Maher, Salem, Feda Al-Bekain Arwa Shawabkeh, IEEE, 2018.

[s]He gave her a sly smile, then another. Air quality forecasting: Big Data and machine learning methods, Gang Zi and Shengqiang Lu, International Journal of Environmental Science and Development, 2018.

[20] Laura Melissa Montes, Juan Felipe Franco, Ivan Mura and Nicholas Mejia Martinez for machine learning Predicting the levels of PM10 Bogotá, 2018; IEEE

[21] Sergio Trilles, Francisco Ramos, Iskandaryan, Ditsuhi, and "Air quality prediction in smart cities using machine learning technologies based on sensor data: a review." In 2020, Applied Sciences 10.7: 2401.

[22] Mao, Yimin, et al. "Azarshahr travertine compression strength prediction based on point-load index (Is) data using multilayer perceptron." Scientific Reports 13.1 (2023): 20807.

[23] Robert A. Van den Berg and colleagues. "Centering, scaling, and transformations: improving the biological information content of metabolomics data." 1–15 in BMC Genomic 7 (2006).

[24] David, Steyrl et al. "Random forests in non-invasive sensorimotor rhythm brain-computer interfaces: a practical and convenient non-linear classifier." Biomedizinische Technik/Biomedical Engineering 61.1 (2016): 77–86.

[25] Daniel Russo and James Zou. "How much overfitting is there in your data exploration? use information to control bias." 56.1 (2019) IEEE Transactions on Information Theory: 302-323.

[26] Kunpeng, Shi, et al., "An improved random forest model of short‐term wind‐power forecasting to enhance accuracy, efficiency, and robustness." Energy from wind 21.12 (2018): 1383–1394.

[27] Yafouz, Ayman, et al "Comprehensive comparison of various machine learning algorithms for short-term ozone concentration prediction." 4607–4622 in Alexandria Engineering Journal 61.6 (2022).

[28] Yanmin Lu, Weijie Wang, and others. "Analysis of the root mean square error (RMSE) and the mean absolute error (MAE) in assessing rounding model." Materials science and engineering conference series, IOP, vol. 324. IOP Books, 2018.

[29] Scheinost, Dustin, et al. "Ten simple rules for predictive modeling of individual differences in neuroimaging." NeuroImage 193 (2019): 35-45.

[30] G. Fathima and K. Nandini. "Urban air quality analysis and prediction using machine learning." The first international conference on advanced technologies in computing, communication, environment, and intelligent control (ICATIECE) will take place in 2019. 2019 IEEE.