Estimating Air Pollutant Concentrations with Machine Learning

Mr. Shubham Sonake[1,] ,Dr. A.D. Sawarkar[2]

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

sonakeshubham1817@gmail.com

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

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 machine learning regression models 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.

Regression models 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 of machine learning regression 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, Random Forest, Linear Regression, Prediction, Decision tree.

1. Introduction

Air pollution is a major concern with negative impacts on human health and the environment. The prediction of air quality fluctuations can enable individuals and authorities to take countermeasures. This study examines the ability of machine learning regression models to predict carbon monoxide (CO) levels in the coming days/months.

Machine learning regression is a robust technique that studies the relationship between input variables and continuous target variable. Here, the model will be trained on historical data including various factors affecting CO concentrations such as climate and industrial activity and analysis of these relationships will enable the model to accurately predict future CO concentrations.

The following are the study's objectives:

- Analyze the Machine Learning Regression Models' Predictive Power for CO Concentrations:
  - Evaluate the degree to which various machine learning regression models are able to learn from the past.
  - Analyze the models' precision and dependability for predicting future CO levels.

- Examine and Dissect the Results of Different Regression Algorithms for CO Prediction:
  - Examine and contrast several machine learning regression methods, including Neural Networks, Decision Trees, Random Forests and Linear Regression.
  - 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.

Predictive modeling frequently uses machine learning regression methods including Neural Networks, Decision Trees, Random Forests, Support Vector Regression (SVR), and Linear Regression. Every one of these algorithms has advantages and disadvantages, and how well they work depends on the particular problem at hand as well as the characteristics of the data. Simple and easy to understand, linear regression may miss intricate patterns in the data. SVR can require a lot of processing power, but it works well with high-dimensional data. Although decision trees are simple to see and understand, overfitting is a potential problem. By averaging several trees, Random Forests, an ensemble technique, lessen overfitting; nonetheless, they can be more difficult to understand. Deep learning models in particular, which use neural networks, are excellent at identifying complex patterns, but they need a lot of data.

These models are evaluated using a variety of performance metrics. The Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) are examples of common metrics. These measures shed light on how accurate and consistent the model predictions are. The mean absolute error (MAE), the standard error (MSE), the interpretable measure (RMSE) in the same units as the target variable, and the percentage of the dependent variable's variance that can be predicted from the independent variables are all indicated by R^2.Model performance is greatly influenced by feature selection and engineering in addition to selecting the appropriate algorithm and performance measures. While feature engineering entails developing new variables that can improve model performance, feature selection focuses on determining the most pertinent variables that affect CO levels. To optimize model parameters and raise overall performance and efficiency, strategies like hyperparameter tweaking with Grid SearchCV can be applied.

The literature on machine learning-based air quality forecasting will be reviewed in the ensuing parts, which will also include details about methodologies, outcomes, experimental design, and a discussion of conclusions and future directions.Research demonstrates that machine learning can 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. 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 this study entails tuning the hyperparameters of several machine learning models, training them on preprocessed data, and assessing them with selected metrics. Cross-validation will guarantee the generalizability and robustness 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 people can lower their exposure to dangerous CO levels and authorities can put preventive measures into place. Subsequent investigations could delve into sophisticated algorithms, integrate real-time data for dynamic forecasts, and expand the structure to anticipate other contaminants.

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

Recent studies have investigated different methods for air pollution quantification, using different methods and techniques.

Veljanowska and Dimosky (2018) [1] 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) [2] 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) [2018] . [3] applied Fuzzy Logic to forecast PM2 and PM10 concentrations, which handle irregularities well but encountered problems with aggregating data resulting in inaccurate information Furthermore, CR et al. (2018) [4] used Autoregression to identify pollution events and linear regression to predict PM2.5 levels, and highlighted the challenges of adapting to a changing climate and thus, Zhang et al colleagues. (2018) [5] proposed Wavelet Neural Networks but struggled in choosing the appropriate parameters, which affected the prediction accuracy. Amado and Dela Cruz (2018) [6] 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) [8] were impressed by the hourly forecasting capabilities of Deep Belief Networks but found issues with sensor data quality. Mejia et al. (2018) [9] 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.

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 primary focus of this study 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 metal oxide chemical sensors
- 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 allows for the creation of further robust models for forecasting air quality using machine learning techniques.

- Partitioning for Testing: A division of the dataset was conducted, with 70% allocated for training purposes and 30% reserved for testing, facilitating a thorough assessment of the model's performance.

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 the 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 used to normalize the data. This method changes each item to a Standard Deviation of 0 to 1. This guarantees that all items contribute equally to the learning process in the model, regardless of their original unit or scale.

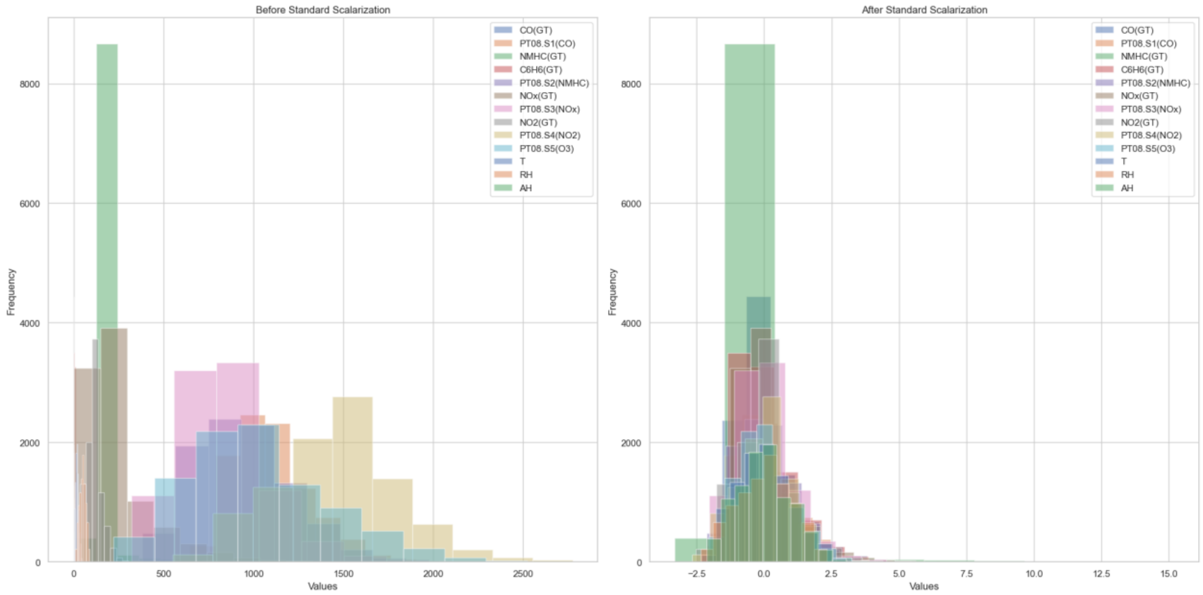

Fig 1. Feature Distribution Before and After Scalarization

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

3.3 Model Selection

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

- Model interpretability: It is important To figure out the 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,

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 there are far too many factors in the data. 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. on the basis of heat map observations, available selection methods can be used to address 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.

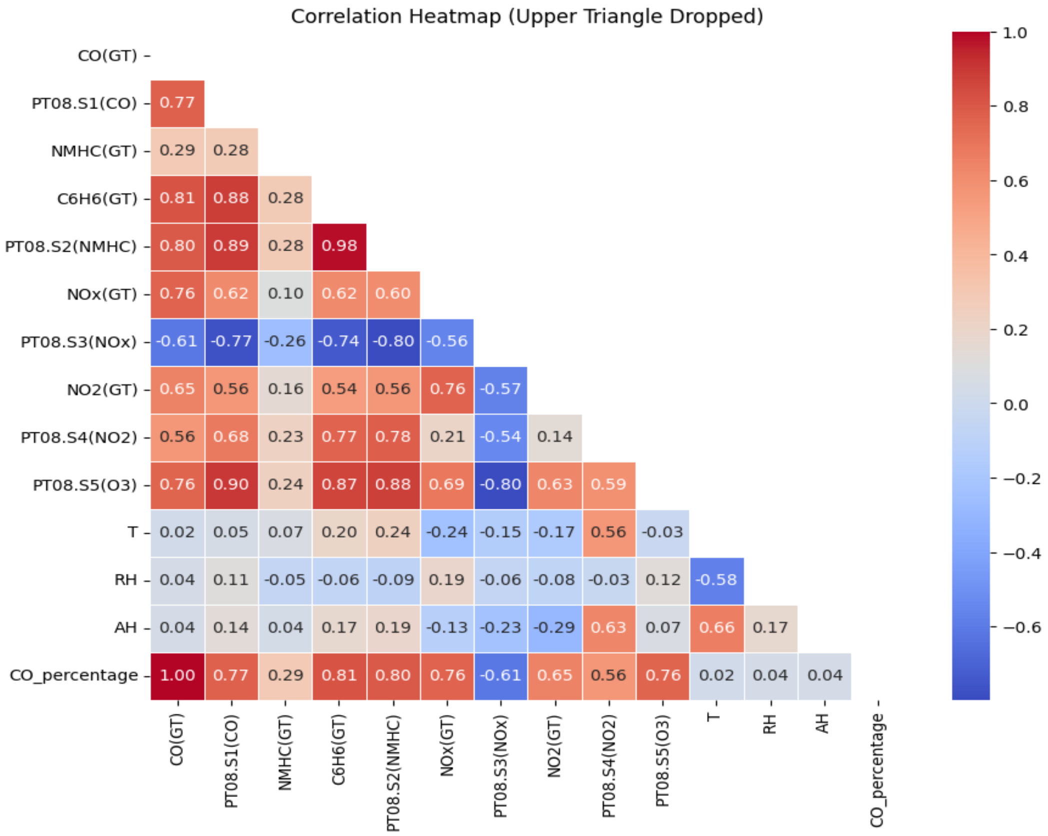

Fig 2. Correlation Heatmap of Pollutants

The heat map shows the relationship 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 combination 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 multiple decision trees increases generalizability and reduces overfitting, resulting in the most robust possible forecast of CO concentrations
By analyzing these systems and using information from the correlation heat map, we aim to identify the model that provides the most accurate and interpretable results for estimating CO concentrations available in air quality database.

3.4 Evaluation Metrics

In order to assess the performance of the selected regression models in predicting CO concentration, we will utilize a set of evaluation metrics. These measures measure how much the actual CO values detected in the air quality dataset depart from the projected CO values.

1) R-squared (R^2) Score: This metric represents The share of variance in the actual CO concentration that can be explained by the model's predictions. A higher R^2 score (closer to 1) indicates a better fit between the model and the data.

   Formula: $R^2 = 1 - \frac{(y - \bar{y})^2}{(y - \hat{y}i\ )^2}$

   Where:
   yi - actual CO value for data point i
   ŷi - predicted CO value for data point i by the model
   ȳ - average of all actual CO values

2) Mean Absolute Error (MAE): MAE calculates the average absolute difference between predicted CO values and the actual CO values. A lower MAE signifies a smaller average prediction error.

   Formula: $\text{MAE} = \frac{\sum_{i=1}^{i=n} |\text{yi} - \hat{\text{y}}\text{i}|}{n}$

   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

3) Mean Squared Error (MSE): MSE squares the individual differences between predicted and actual CO values before calculating the average. While sensitive to outliers, MSE provides insights into the average magnitude of prediction errors.

   Formula: $\text{MSE} = \frac{\sum_{i=1}^{i=n} |\text{yi} - \hat{\text{y}}\text{i}|^2}{n}$

   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) Mean Squared Error (RMSE): MSE squares the individual differences between predicted and actual CO values before calculating the average. While sensitive to outliers, MSE provides insights into the average magnitude of prediction errors.

Formula: $$\text{RMSE} = \sqrt[2]{\frac{\sum_{i=1}^{i=n} |yi - \hat{y}i|^{2}}{n}}$$

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 - linear regression , random forest, decision tree - to be used 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. 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 dividing the data 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 complex interactions among factors that can affect 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 They are not as easily understandable as. linear regression but often more interpretable than complex decision trees.

5. **Results**

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

Models\Evaluation	R^2 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

To Improve Accuracy:

1) 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	R^2 Score	MAE	MSE	RMSE
Linear Regression	0.84	0.33	0.28	0.52

6. **Conclusion:**

The outcomes from various models like the random forest, the decision tree, and the 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 use 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.

6. References:

[1] 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).

[2] 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.

[3] 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.

[4] 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).

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

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

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

[8]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.

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