IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Digital Object Identifier

Comparative Analysis Study for Air Quality Prediction in Smart Cities Using Regression Techniques

SHOROUQ AL-EIDI[1], FATHI AMSAAD[2], OMAR DARWISH[3],YAHYA TASHTOUSH[4], ALI ALQAHTANI[5], AND NIVESHITHA NIVESHITHA[2].

[1]Computer Science Department, Tafila Technical University, Tafila, Jordan, (saleidi@ttu.edu.jo)
[2]Computer Science and Engineering, Wright State University, Colonel, USA, (fathi.amsaad@wright.edu), (niveshitha.2@wright.edu)
[3]Information Security and Applied Computing, Eastern Michigan University, Michigan, USA, (odarwish@emich.edu)
[4]Department of Computer Science, Jordan University of Science and Technology, Irbid,Jordan, (yahya-t@just.edu.jo)
[5]Department of Networks and Communication Engineering, Najran University, Najran, Saudi Arabia, (asalqahtany@nu.edu.sa)

Corresponding author: Shorouq Al-Eidi (e-mail: saleidi@ttu.edu.jo).

ABSTRACT Air pollution has detrimental impacts on our physical health and the quality of our living environment, particularly in smart cities. Monitoring and predicting air pollution is crucial to empower individuals to make informed decisions that protect their health. Predicting air quality accurately plays an important effective action plan to mitigate air pollution and create healthier and more sustainable environments. This can be achieved by relying on the Air Quality Index, one of the most reliable indicators for air pollutant concentration levels in certain cities.

This study provides a comparative analysis study for air quality prediction using three regression techniques: Random Forest regression, Linear regression, and Decision Tree regression, employing the AQI values. This comparison aims to identify the most efficient model based on various evaluation criteria, such as Mean Absolute Error and R^2 measures. Additionally, it considers both error rate minimization and processing time efficiency of each regression model evaluated within two distinct frameworks as other important measures to determine the best-fitted model. The findings of this study showcase the superiority of the Decision Tree regression technique over the other models, demonstrating its exceptional accuracy with a high R^2 score and low error rate. Moreover, integrating cloud computing technology has yielded significant improvements in model execution time, substantially enhancing the overall efficiency of the prediction process. By leveraging distributed computing resources, real-time air quality forecasting becomes feasible, enabling timely decision-making and proactive measures to address air pollution episodes.

INDEX TERMS Air pollution, Machine learning, IoT, Smart City, Air Quality Index.

I. INTRODUCTION

Recently, the adverse effects of air pollution have garnered significant attention, with studies from the World Health Organization highlighting its impact on the global atmosphere and human health. It has been identified as a leading cause of illnesses, allergies, and premature deaths, responsible for a staggering 12% of global deaths in 2019 [6]. Moreover, air pollution introduces dangerous substances into the atmosphere, including greenhouse gases and biological compounds [9], further exacerbating our environmental challenges. Recognizing the importance of air quality in our daily lives, there is a growing demand for better air quality management in smart city environments. With their high concentration of human activities, urban areas experience rapid variations in pollutant emissions, making quantifying air quality essential. To assess pollutant concentrations in cities, the Air Quality Index (AQI) has emerged as a crucial tool [2]. It guides decision-making processes and is a key resource for air pollution analysis and warning. Given the situation of urgency, adopting sustainable solutions that effectively mitigate air pollution has become imperative, particularly for the well-being of future generations. Early prediction of AQI levels plays a vital role in environmental management and preventing potential dangers of air pollution. Various forecasting models have been proposed to predict pollution levels and concentrations in recent years. One notable approach is machine learning, which has gained prominence due to its ability to handle complex interactions between air quality parameters. Machine learning-based prediction systems are increasingly attractive for accurate forecasting in air quality management [8], [14].

This study aims to address the challenges of time and cost constraints in air pollution prediction by leveraging the efficiency of machine learning techniques with the AQI. It compares three regression approaches for predicting air quality. The performance of each technique is evaluated using established criteria such as Root Mean Square Error (RMSE), R^2 score, and Mean Absolute Error (MAE) to identify an efficient and the most suitable regression model for predicting air quality. The analysis considers multiple pollutants, such as Sulphur Dioxide, Nitrogen Dioxide, and more, to enhance the accuracy of the air pollution prediction model.

In addition to accuracy, this study recognizes the real-time processing requirements of smart cities and evaluates the processing time of each regression technique. Distributed computing techniques are employed to reduce execution time while maintaining prediction accuracy. Optimization considerations, including data size and processing time, are considered.

This study's results have practical implications for creating effective air pollution control strategies and contribute to advancements in air quality prediction methodologies. Particularly in urban areas where AQI monitoring is crucial for public health and environmental management, these findings can inform decision-making processes, aiding in the development of proactive measures to address air pollution challenges effectively.

This paper is organized as follows. Section II provides a review of the air quality and pollution prediction literature. Section III details the methodology proposed approaches, which illustrate the experimental setup pre-processing techniques and utilize machine learning algorithms to predict air pollution. Section IV presents the experiment results. Section V offers a conclusion and potential future work.

II. LITERATURE REVIEWS

The field of air pollution prediction has experienced a notable rise in machine learning techniques to address the challenges associated with forecasting air quality levels. These techniques have demonstrated their effectiveness in predicting air pollution, thus contributing significantly to developing air quality management strategies. This section comprehensively explores the most notable models utilized for calculating and predicting the Air Quality Index (AQI) and the concentration levels of various air pollutants through different machine learning algorithms, such as regression techniques. These models hold considerable relevance and find practical utility in other application domains such as cloud computing.

Patil et al. [18] extensively reviewed different methodologies and techniques to analyze the concentration level of air pollution and the prediction of AQI. This study highlighted the performance of these analytical methods and presented the importance of calculating AQI as a significant measure for assessing pollution levels and how it dramatically influences human health and the environment. Similarly, Oliveri et al. [15] reviewed air quality models while discussing the effect of air pollution concentration on human health.

A noteworthy study by Ameer et al. [1] scrutinized the efficiency of four regression methods, namely Decision Tree, Gradient Boosting, Multilayer Perceptron, and Artificial Neural Network (ANN), in predicting air quality levels. These methods were evaluated based on tracking PM2.5 levels in the air and calculating the AQI. The findings of this study concluded the Random Forest regression method outperformed the others, achieving an adjusted MAE of 16% for Beijing City. Notably, this method also significantly reduces the computation time compared to Multilayer Perceptron and Gradient Boosting, emphasizing its practicality. Similarly, HeidarMaleki et al. [12] employed the ANN algorithm to predict the AQI and concentration levels of several air pollutants such as NO2, SO2, PM10, PM2.5, CO and O3. This study encompassed various monitoring stations, including Naderi, MohiteZist, Havashenasi, Behdasht, and Iran. The authors employed a set of parameters, such as time, date, air pollutant concentration, and meteorological data, as inputs for their ANN model. This approach resulted in the development of a robust predictive model that can provide insights into air quality.

Moreover, Zhang et al. [22] proposed a deep learning model for air quality prediction utilizing the long short-term memory (LSTM) network. Their study included a series of experiments using Detrended Cross-Correlation Analysis (DCCA) to investigate the relationship between predicting levels of several air pollutants (SO2, NO2, PM10, PM2.5, O3, and CO) and meteorological data, such as temperature, barometric pressure, humidity, and wind speed. Experiment results found a negative correlation between AQI and temperature, humidity, and wind speed, while a strong positive correlation was observed between pressure and AQI. Furthermore, Bougoudis [3] developed a hybrid computational method to identify the correlation between air pollutants and weather conditions to determine the actual cause of pollution. The study employed ANN and Random Forest as ensemble learning methods, claiming increased accuracy. However, the feedforward neural network faced challenges predicting continuous values due to insufficient data.

For using classification machine learning algorithms, Gore et al. [5] proposed a classification approach to study how air pollutant levels affect the health of humans. In their process, they employed Naive Bayes and Decision Tree algorithms and achieved a high accuracy using the Decision Tree model. Moreover, Simu et al. [21] presented a comparative study to compare the performance of several machine learning algorithms, such as Random Forest and Multi-linear Regression, in analyzing air pollutants and predicting air pollution levels. The study results concluded that the Multilayer Perceptron algorithm outperformed the other.

Moreover, In [19], Peng et al. utilized Multilayer Perceptron to enhance the air quality prediction accuracy. However, they noted limitations in data extension and the high computational cost because of the seasonal update of the model.

Mahalingam et al. [10] proposed using ANN and SVM algorithms to predict the AQI in the smart city of Deldi with impressive accuracies, mainly the Medium Gaussian SVM function. To predict the AQI and air pollution levels, Sharma et al. [20] implied various algorithms, including Linear regression, ANNs, Lasso regression, and XGBoost regression. The study focused on tracking the values of several pollutants, including NO2, SO2, PM2.5, PM10, CO, and O3. The research findings indicated that the Random Forest algorithm outperformed the other algorithms, demonstrating its high performance in predicting the AQI and air pollution levels.

Nandini et al. [13] used Decision Trees and Multinomial Logistic Regression to forecast and analyze air quality pollutant levels, achieving better accuracy with Multinomial Logistic Regression compared to Decision Tree. Similarly, in a study by Mahanta et al., [11], a comprehensive comparison of several algorithms, including Linear regression, Decision Forest, XGBoost, ElasticNet, Boosted Decision Tree, KNN, Lasso regression, and Ridge regression to predict air pollutant levels. Among these algorithms, Extra Trees exhibited superior performance due to its technique of ranking the essential features to improve the accuracy of the predictions. Moreover, Pasupuleti et al. [17] conducted a study comparing Random Forest, Decision Tree, and Linear regression models for predicting air pollutants and meteorological conditions in the Arduino platform. The study found that the Random Forest model provided better performance by reducing errors caused by overfitting. However, it was noted that the Random Forest model required more memory and incurred higher costs.

For using the clustering approach, Kingsy et al. [7] enhanced the K-Means algorithm to analyze and identify the air pollution level. Their method calculates the correlation coefficient between pollutant data to determine the AQI value and find the air pollution level in a specific location. To validate their findings and evaluate the effectiveness of their approach, the authors compared their proposed algorithm with the Fuzzy C-Means algorithm. Their results demonstrated that the proposed K-Means clustering algorithm achieved higher accuracy and less execution time than the Fuzzy C-Means algorithm.

Ganeshkumar et al. [4] presented an efficient and cost-effective classification model for environmental monitoring and air pollution prediction. Their study the authors used several artificial methods with a cloud platform for data processing, leading to significant time savings, reduced labor efforts, and producing high-quality outcomes. This research highlights the importance of integrating cloud platform solutions to enhance the efficiency and accuracy of monitoring and air quality prediction models, which is beneficial for addressing environment mentoring challenges. Similarly, Park et al. [16] used their own cloud computing technique to reduce the processing time of processing and visualization of urban air pollution data.

The literature review underscores the widespread prediction of air quality and air pollution utilizing machine learning algorithms, highlighting their potential to achieve accurate results, efficient computation, and effective prediction of air quality levels. However, certain limitations need to be addressed. These include the necessity for more extensive and more comprehensive datasets, challenges in accurately predicting continuous values, and the high computational cost associated with model updates. Additionally, the review identifies a research gap in the focus on predicting the AQI based solely on PM2.5 measurements, neglecting the inclusion of other important air pollutants. Incorporating data on multiple pollutants such as O3, NO2, SO2, and PM2.5 can significantly enhance the accuracy of air pollution prediction models. These insights provide valuable guidance for future research endeavors and for developing effective air quality management strategies, particularly in smart cities.

III. METHODOLOGY

The methodology for the proposed air quality prediction model in this study is outlined in Figure 1. Initially, air quality datasets were collected and loaded for analysis. A series of preprocessing steps were carried out to ensure data quality, including handling missing values, reducing noise, and calculating the Air Quality Index (AQI). The preprocessed dataset contains pollutant information, such as CO, SO2, O3, NO2, and PM2.5, along with their corresponding AQI values.

The next step identified the air quality prediction process's most relevant and important features. This step helps reduce the dimensionality of the air dataset and focuses only on the significant variables. The dataset was then balanced to ensure equal representation of different classes, followed by splitting it into training and testing sets. In the training set, the models were trained using preprocessed data, while the testing was based on assessing the models' predictive accuracy in estimating AQI. The regression models were evaluated during the testing phase using a separate dataset designed for testing purposes. Performance metrics were computed by comparing the predicted AQI values with the observed data, allowing for the identification of a suitable and efficient model for predicting the air quality level. The subsequent sections will provide more detailed discussions on each methodology component, shedding light on this study's specific techniques and procedures.

A. DATASET DESCRIPTION

The dataset used in this study encompasses a comprehensive collection of 103,205 records, featuring data from monitoring stations situated across ten diverse locations within Pune City [1]. These areas include Bopadi Square 65, Karve Statue Square 5, Lullanagar Square 14, Hadapsar Gadital 01, PMPML Bus Depot Deccan 15, Goodluck Square Cafe

[1] https://www.kaggle.com/datasets/akshman/pune-smartcity-test-dataset

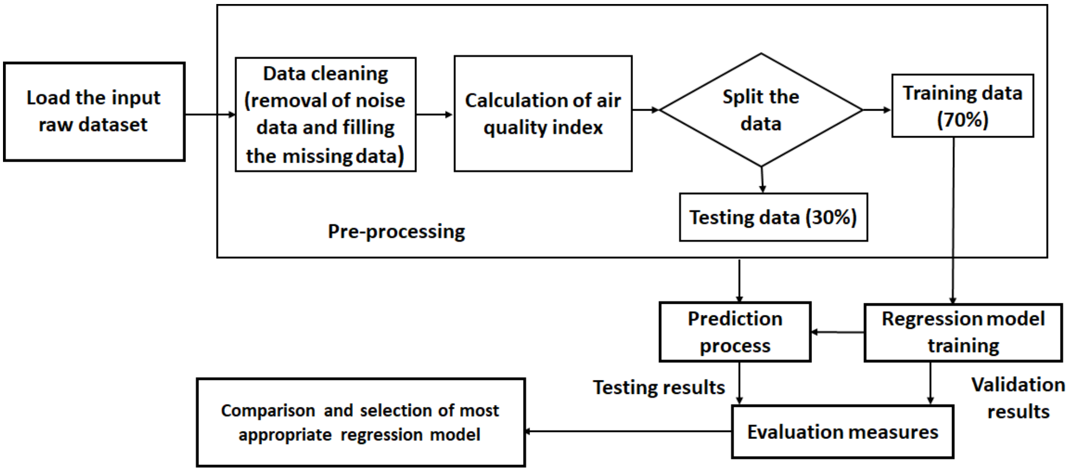

Load the input raw dataset
Data cleaning (removal of noise data and filling the missing data)
Calculation of air quality index
Split the data
Training data (70%)
Testing data (30%)
Pre-processing
Prediction process
Regression model training
Testing results
Validation results
Evaluation measures
Comparison and selection of most appropriate regression model

FIGURE 1: Air Quality Prediction Model.

23, Chitale Bandhu Corner 41, Pune Railway Station 28, Rajashri Shahu Bus Stand 19, and Dr. Baba Saheb Ambedkar Sethu Junction 60. The dataset, compiled in 2019, resulted from a collaborative effort between the Pune smart city and the Indian Institute of Science, Bangalore.

Within the dataset, we focus on 28 distinct features related to air pollution, including NO2 (Nitrogen dioxide), O3 (Ozone), PM10 (Particulates) with a diameter of less than 10 microns, PM2.5 with a diameter of less than 2.5 microns, SO2 (Sulphur dioxide), CO (Carbon monoxide), and AQI. This study considers these features as crucial indicators of the pollutant concentration in the air. They enable us to calculate and predict the air quality in Pune's smart city, enabling a comprehensive understanding of pollution patterns.

B. DATA PRE-PROCESSING

Data pre-processing is an important step in any data analysis process as it focuses on improving the quality and reliability of the dataset by reducing noise and inconsistencies. In this study, we employed several pre-processing approaches to focus on the integrity of the data. It consisted of several key stages: data cleaning, AQI calculation, dataset splitting, and data balancing.

The first step in data cleaning is handling missing values in the raw data. The dataset comprised 103,205 entries, encompassing several data types such as integers, floats, and objects. Some of these entries had null or missing values, which must be addressed. To handle this issue, missing values were replaced with the mean values for pollutant parameters. This approach helped maintain the dataset by ensuring no crucial information was lost due to missing values.

The interquartile range (IQR) method addressed duplicate observations and outliers. This method utilizes quartiles Q1 (25th), Q2 (50th), and Q3 (75th) percentiles. The IQR can be calculated as the difference between the two values, Q3 and Q1. And consider the outlier as any values falling outside the range of $(Q1 - 1.5 * IQR)$ and $(Q3 + 1.5 * IQR)$. Instead of removing the outlier values, we used the lower and upper boundary values to replace them and retain important information while mitigating the impact of data outliers on the analysis process.

For Pune city, the AQI values were calculated using the air pollutant features present in the dataset. These features, including NO2, PM10, PM2.5, O3, SO2, and CO, represent different pollutants in the air. The calculation of AQI provides a standardized measure of air quality conditions in Pune city.

Exploratory Data Analysis (EDA) is a powerful tool in this study to gain insights into the dataset and understand its characteristics. It is considered an important tool for cleaning and preparing the raw data for training purposes. We conducted descriptive statistics of the dataset during the EDA process. This involved analyzing various statistical measures such as percentiles, minimum and maximum values, standard deviation, and mean for each pollutant feature. By calculating these statistics values, we obtained a comprehensive dataset overview, enabling us to identify potential anomalies that could affect the analysis.

TABLE 1: Basic Characteristic of Dataset

Feature	Mean	Std	Min	Q1	Q2	Q3	Max
NO2	67.7	34.5	0	42.5	70.5	91.5	315.5
PM10	16.8	11.4	0	7.0	16	26	48.5
PM25	13.4	8.5	0	6.5	12.5	20	37.5
SO2	4.9	11.4	0	1.0	3	5.5	165
CO	71.2	27.8	13.5	50	75.5	89.5	144.5
OZONE	10.8	22.8	0	0	3.5	12	335

1) AQI Calculation

The Air Quality Index (AQI) is one of the most crucial parameters in assessing and monitoring the air quality in a specific area. It provides a standard measure system that quantifies air pollution and helps understand its effects on human health and the population. The AQI is a numerical value within a defined range, typically from 0 to 500. A higher value of AQI indicates poorer air quality and the existence of harmful air pollutants. Each pollutant has specific constraints and specific averaging periods to ensure accurate assessment for O3. The period is 8-hour maximum, and the 24-hour average concentrations for SO2, PM10, CO, NO2, PM2.5.

To calculate the AQI, the concentrations of these air pollutants are categorized into sub-indices. These sub-indices are then defined based on predefined ranges that indicate the level of air quality, ranging from "good" to "hazardous." The highest sub-index value among the pollutants represents the overall air quality index or air pollutant index for a specific location. The computation of the AQI is based on Equation 1, which combines the sub-indices of each pollutant [11]. This equation considers the weightage assigned to each pollutant based on its potential health impact. In the end, by incorporating multiple pollutants and their respective sub-indices, the AQI helps to assess the air quality in a particular area [1].

$$I = \frac{I_{high} - I_{low}}{C_{high} - C_{low}}(C - I_{low}) + I_{low} \tag{1}$$

where, I is Air Quality Index, C is Pollutant concentration.

C. FEATURE SELECTION

Feature selection becomes crucial in our research following the data preprocessing and exploratory data analysis step. This process involves identifying and selecting the most relevant features related to the AQI, representing the overall air quality. The features in this study based on the preprocessed dataset contain several pollutant information such as CO, SO2, O3, OZONE, NO2, PM10, and PM2.5, along with their corresponding AQI values.

We used the correlation analysis to determine the relationship between the features and AQI. Correlation analysis can be used to find the linear relationship between two variables. By calculating the correlation coefficients between each feature and the AQI, we can assess their predictive value in understanding and predicting variations in air pollutant levels. The correlation values are compiled into a correlation matrix, which provides a view of the relationships between all dataset variables and identifies features with strong positive or negative correlations with the AQI, as shown in Figure 2.

In this study, we have found that all the values of air pollutants in the dataset demonstrate a positive correlation with the AQI. This indicates that higher concentrations of these pollutants are associated with higher AQI values, reflecting poorer air quality. This highlights how these features are important in analyzing and predicting air quality variations in the study area by selecting important features representing significant correlations with the AQI. Moreover, we can eliminate irrelevant or redundant features and build a more effective predictive model for forecasting air quality levels.

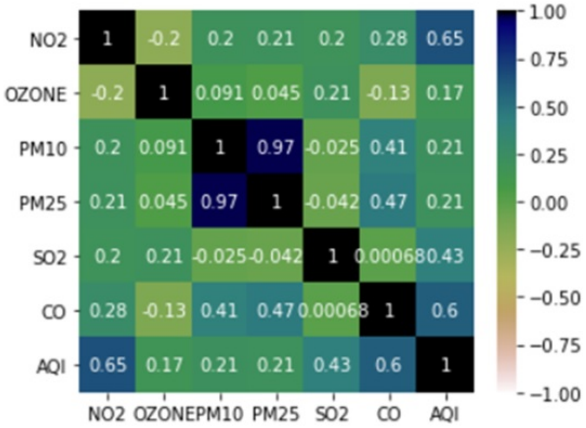


FIGURE 2: Correlation of AQI Air Pollutants.

D. SPLITTING DATA

In this step, the train-test split() method was utilized to split the data into two parts with a ratio of 70:30 for training and testing sets. This means 70% of the total dataset was chosen for training, while the remaining 30% of data was assigned for testing data. With this splitting ratio, the model is trained on a large sufficient portion of the data and evaluated on test data to assess its performance.

E. BALANCING DATA

In machine learning tasks, addressing the issue of imbalanced data is crucial to ensure reliable and accurate predictions. The distribution of AQI values exhibits an imbalance in the given dataset, with certain values occurring more frequently than others. This can be observed by categorizing the AQI values into predefined ranges, as shown in Figure 3.

Working with imbalanced data poses challenges and can significantly impact the performance of machine learning models. Biases can occur as models favor the majority class and overlook minority classes with fewer instances. In this step, the SMOTER (Synthetic Minority Over-sampling technique for Regression with Gaussian Noise) approach is used to reduce the majority class issue and improve the model's performance.

The SMOTER approach is used in the imbalanced dataset by generating a synthetic minority and under-sampling the majority class. This way helps to get a balanced dataset and ensures a more equitable representation of different AQI values. By finding synthetic samples, the minority class is amplified and can create a more balanced distribution of data points. Gaussian noise is also added to these synthetic samples, which introduces variations and helps prevent overfitting. By utilizing the SMOTER approach and balancing the dataset, the models are trained on a more representative and several sets of data points. In this step, the model's ability to capture patterns and relationships across different AQI values will be enhanced, leading to improved performance and more accurate predictions.

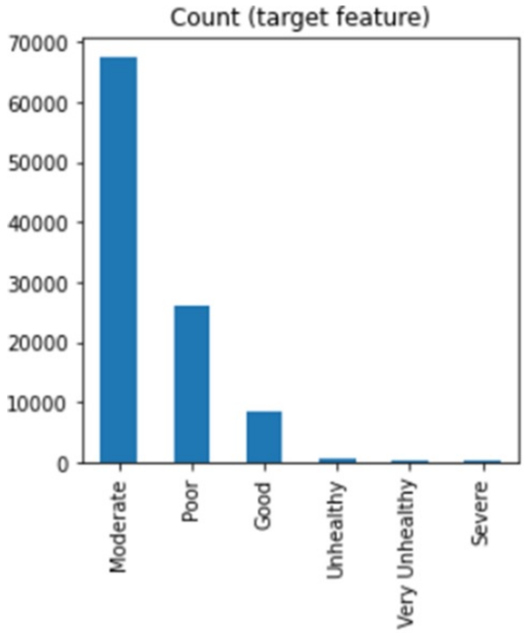


FIGURE 3: AQI Classified Categories.

F. REGRESSION MACHINE LEARNING MODELS

1) **Decision Tree regression**: is a supervised machine learning algorithm commonly used to model non-linear relationships between output variables and input features. The algorithm partitions the data into subsets based on specific rules or criteria in this regression approach. These rules are selected to minimize the difference in space between the predicted and the actual values.
   By considering several input factors and training the model using historical air pollution and AQI data, Decision Tree regression can be applied to predict the air quality. The model analyzes the relationships between the input factors and AQI to accurately predict upcoming periods.
2) **Linear regression**: is a commonly employed statistical method in several approaches for prediction and forecasting air pollution [20]. It is used for examining the relations between pollutant concentrations and the AQI. Linear regression can make reliable predictions about future air pollution levels by analyzing historical data and discerning trends and patterns. Furthermore, Linear regression aids in identifying the primary factors contributing to air pollution. By assessing the regression coefficients, it becomes possible to determine the extent to which the variable influences the AQI. This information can be crucial in formulating effective control measures to mitigate pollution and enhance air quality.
3) **Random Forest regression**: is a supervised learning technique that combines Decision Trees and can be used for regression problems. The input data goes through multiple Decision Trees, and the average of each tree is used as the model's output in the training process [1].

G. EVALUATION MEASURES

To compare the regression models performance, several evaluation measures are utilized. These measures provide quantitative insights into the accuracy and effectiveness of the models. The following measures are commonly employed:

- **Mean Absolute Error (MAE):** is a metric used to calculate the average of differences of the actual and predicted values in the testing data. It provides an indication of the average of the model errors as shown in the equation below:

$$MAE = \frac{1}{n}\sum_{j=1}^{n}(y_j - y_j^{'}) \tag{2}$$

- **Root Mean Square Error (RMSE)**: is a widely used for evaluating regression models. Its used to calculate the average deviation between predicting and the actual of model values. A lower RMSE value highlighted that the model achieved better performance. It can be calculated using the following formula:

$$RMSE = \sqrt{\frac{1}{n}(\sum_{j=1}^{n}(y_j - y_j^{'})^2)} \tag{3}$$

- R^2 **Score**: also known as the coefficient of determination, used to find the variance of target variables in the model. It ranges from 0 to 1, where a higher value representing a the model fit the dataset in good way. It is calculated using the following formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i')^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{4}$$

IV. RESULT AND DISCUSSION

This section is represented by evaluating the performance of three regression techniques, considering multiple factors. The primary focus is to assess the accuracy and reliability of each technique in predicting the air quality process based on comparing the actual and predicted values. In addition to accuracy, the performance of the regression techniques was compared based on their implementation of two different frameworks. This allows us to analyze the platform's impact on the models' efficiency and effectiveness.

Moreover, execution times for each regression technique on the selected platforms were considered an important performance measure. This analysis provides valuable insights into the computational efficiency and speed of the models. In the following subsequent sections, the detailed results of the comparisons represent and offer a comprehensive understanding of the performance of the regression techniques.

A. COMPARISON OF ACTUAL AND PREDICTED DATA

In any analysis, evaluating the accuracy and reliability of the models used is crucial. In this study, the performance of three regression models is being compared. Evaluating the accuracy and reliability of these models is crucial to assess their suitability for predicting air quality. To assess these models' goodness fit, a visualization technique was utilized to compare the actual values of the model and the prediction result values. By visually comparing the two, we can quickly evaluate the proximity of the predicted values to the actual values, providing insights into the accuracy of each model. Figure 4 displays the actual values versus the predicted values specifically for linear regression. The blue line represents the perfect regression line, and the model accuracy depends on how closely the data points align with this line. Upon examining the linear regression results, we observe that the data points are clustered at the bottom of the graph and are not closely aligned with the regression line. This suggests that linear regression may not be the most suitable model for air quality prediction in this study.

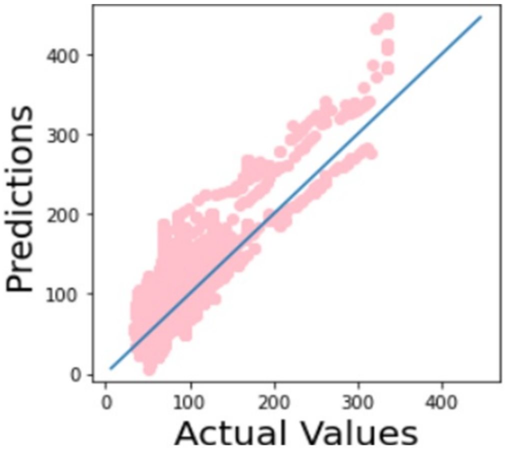


FIGURE 4: Actual vs Predicted for Linear Regression.

Continuing with the evaluation of regression models, Figure 5 compares the actual and the prediction values of the Decision Tree regression model. Analyzing the results, we observe that the data points are more evenly distributed throughout the graph and closer to the regression line than the Linear regression case. This indicates that our study's Decision Tree regression model performs better in predicting air quality.

The improved distribution and proximity of data points to the regression line in the case of Decision Tree regression signify a higher level of accuracy and reliability in predicting air quality compared to linear regression. This suggests that the Decision Tree regression model may provide more precise predictions based on the dataset.

Concluding the evaluation of regression models, Figure 6 illustrates the comparison values of the Random Forest regression model. Upon analysis, we observe that the data points are distributed and closer to the regression line, and this graph looks similar to the Decision Tree model graph. While the Random Forest model may offer advantages in handling complex relationships and reducing overfitting, the Decision Tree model's simplicity and interpret-ability make it a compelling option for understanding the factors influencing air quality. The Decision Tree model can provide valuable insight into the variables that can represent the most significant impact on air quality, aiding decision-making processes.

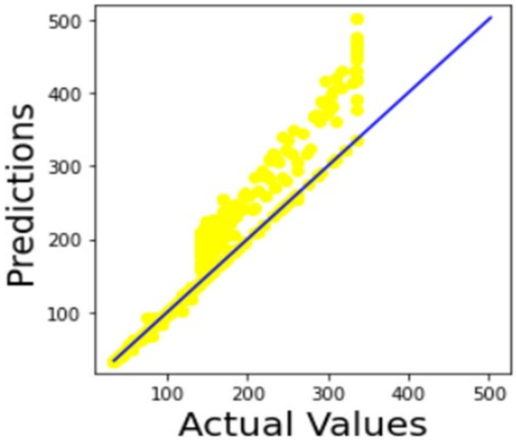


FIGURE 5: Actual vs Predicted for Decision Tree Regression.

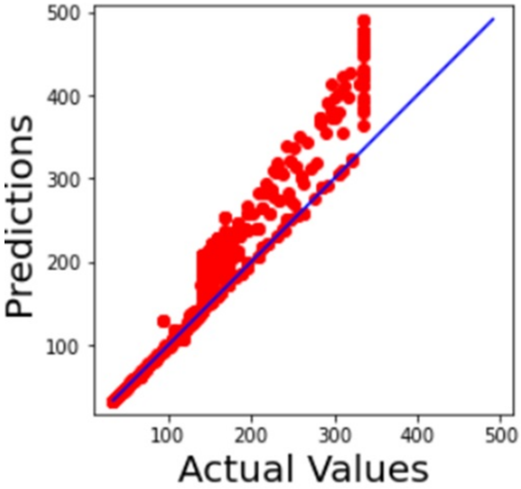


FIGURE 6: Actual vs Predicted for Random Forest Regression.

B. PERFORMANCE EVALUATION USING DIFFERENT CONFIGURATION

In this section, we evaluate the performance of various regression models for predicting air quality in different configurations. We consider two configurations: personal laptop and cloud. The evaluation is based on evaluation measures. Assessing the models' performance in different platforms is crucial to ensure reliability and gain insights into their suitability for real-world applications. Additionally, it allows us to assess the impact of computational resources and infrastructure on the models' effectiveness. The following subsections provide the evaluation results for each platform.

1) Performance Evaluation in First Configuration

In the first configuration, we evaluate the performance of the regression models using a personal laptop platform. Tables 2 and 3 represented the results obtained from the training and testing datasets, respectively. Upon analyzing the error evaluation measures (RMSE, MAE), results conclude that the Decision Tree regression model outperforms the other models. It achieved 2.02% of MAE and 10.14% of an RMSE, indicating its ability to make predictions with minimal average error and variability. On the other hand, Linear regression exhibits the poorest performance among the models, with a relatively high MAE of 32.19% and RMSE of 42.70%. This indicates a high deviation between the predicted and actual values, suggesting that Linear regression may not be the suitable model for predicting air quality accurately.

TABLE 2: Evaluation Results of Training Dataset Using Laptop Configuration

Model	MAE	RMSE	R^2
Random Forest Regression	2.46	10.39	98.80
Decision Tree Regression	2.02	10.14	98.86
Linear Regression	32.19	42.70	79.73

TABLE 3: Evaluation Results of Testing Dataset Using Laptop Configuration

Model	MAE	RMSE	R^2
Random Forest Regression	0.795	7.141	93.08
Decision Tree Regression	0.738	7.073	93.21
Linear Regression	20.137	25.562	11.36

2) Performance Evaluation in Second Configuration

In the second configuration, the regression models' performance was evaluated using a cloud platform. The evaluation results for the training and testing datasets are presented in Tables 4 and 5, respectively.

Upon analyzing the evaluation metrics, we observe that the performance of the linear regression model remains relatively consistent compared to the first configuration. The MAE and RMSE values are almost unchanged, suggesting that the platform variation does not significantly affect the model's performance. On the other hand, there is a slight improvement in the performance of the Decision Tree regression model when running on the cloud platform. The MAE and RMSE values for the training dataset show a marginal decrease, with an MAE of 1.97% and a RMSE of 9.94%. This improvement indicates a slightly higher predicting air quality results than the previous configuration.

Table 5 shows that the Random Forest performance is comparable to the Decision Tree model. Both models represent similar MAE and RMSE values with similar predictive capabilities. However, it is worth noting that the Random Forest model tends to have a longer execution time, which may limit its suitability and efficiency for certain real-world applications where time is crucial.

TABLE 4: Evaluation Results of Training Dataset Using Cloud Configuration

Model	MAE	RMSE	R^2
Random Forest Regression	2.39	10.19	98.82
Decision Tree Regression	1.97	9.94	98.88
Linear Regression	32.56	43.08	78.89

TABLE 5: Evaluation Results of Testing Dataset Using Cloud Configuration

Model	MAE	RMSE	R^2
Random Forest Regression	0.82	7.28	93.09
Decision Tree Regression	0.77	7.23	93.19
Linear Regression	20.51	26.16	10.89

C. EXECUTION TIME COMPARISON

This study compared the execution time for three regression models with the SMOTER technique on two different platforms: a personal laptop and a cloud. The goal was to evaluate the impact of cloud computing technology on the efficiency and speed-up of these models. The results presented in Table 6 demonstrate a significant reduction in execution time when the models run on the cloud compared to the personal laptop. The reduction in execution time of the model highlights the advantages of utilizing cloud computing technology for machine learning tasks. For example, the execution time for SMOTER was reduced from 1292.89 seconds on the personal laptop to 464.22 seconds on the cloud, resulting in a reduction of approximately 64%. Similarly, the execution time of Decision Tree regression decreased from 0.46 seconds on the personal laptop to 0.28 seconds on the cloud, representing a reduction significant reduction.

Additionally, the execution time for the Random Forest regression was reduced from 39.40 seconds on using the personal laptop platform to 17.27 seconds on using the cloud, indicating a reduction of approximately 56%. On the other hand, the execution time for Linear regression was already relatively low on the personal laptop, with only 0.07 seconds, and it further decreased to 0.02 seconds on the cloud. These findings demonstrate the benefits of utilizing cloud computing technology in reducing the execution time of regression models, including SMOTER. Reducing the execution time of the model helps achieve more efficient and faster performance, enabling researchers and practitioners to conduct rapid experimentation and deployment of machine learning models. Particularly for larger and more complex datasets, cloud computing frameworks enable distributed data processing and model training, providing a solution to avoid computational challenges and expedite the machine learning workflow.

TABLE 6: Model Execution Times in Seconds

Execution Time	Personal Computer	Cloud
SMOTER	1292.89	464.22
Random Forest Regression	39.40	17.27
Decision Tree Regression	0.46	0.28
Linear Regression	0.07	0.02

V. CONCLUSION

This study provides a comprehensive comparative analysis of different regression models for predicting air quality in smart cities. Notably, the Decision Tree regression model demonstrated a high performance compared to other regression models. Incorporating Exploratory Data Analysis and the SMOTER technique played a pivotal role in enhancing model accuracy by addressing data imbalances and optimizing feature selection. Moreover, the study emphasized the advantages of utilizing cloud computing in regression modeling. Utilizing cloud resources led to reduced model execution time, resulting in enhanced efficiency and scalability. This accelerated experimentation, training, and deployment of the models, enhancing their practical applicability in real-world applications.

For future work recommendations, we explore diverse machine-learning approaches for predicting air quality and air pollution in smart cities. Additionally, investigating the effect of meteorological data, including temperature, pressure, humidity, and wind speed, further enhances AQI and air pollution prediction accuracy. This endeavor provides valuable insight into identifying air quality levels and contributes to more effective air quality management approaches.

REFERENCES

[1] S. Ameer, M. A. Shah, A. Khan, H. Song, C. Maple, S. U. Islam, and M. N. Asghar. Comparative analysis of machine learning techniques for predicting air quality in smart cities. IEEE Access, 7:128325–128338, 2019.

[2] M. Batty, K. W. Axhausen, F. Giannotti, A. Pozdnoukhov, A. Bazzani, M. Wachowicz, G. Ouzounis, and Y. Portugali. Smart cities of the future. The European Physical Journal Special Topics, 214:481–518, 2012.

[3] I. Bougoudis, K. Demertzis, and L. Iliadis. Hisycol a hybrid computational intelligence system for combined machine learning: the case of air pollution modeling in athens. Neural Computing and Applications, 27:1191–1206, 2016.

[4] D. Ganeshkumar, V. Parimala, S. Santhoshkumar, T. Vignesh, and M. Surendar. Air and sound pollution monitoring system using cloud computing. Int. J. Eng. Res., 2020.

[5] R. W. Gore and D. S. Deshpande. An approach for classification of health risks based on air quality levels. In 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM), pages 58–61. IEEE, 2017.

[6] B.-J. He, L. Ding, and D. Prasad. Enhancing urban ventilation performance through the development of precinct ventilation zones: A case study based on the greater sydney, australia. Sustainable Cities and Society, 47:101472, 2019.

[7] G. R. Kingsy, R. Manimegalai, D. M. Geetha, S. Rajathi, K. Usha, and B. N. Raabiathul. Air pollution analysis using enhanced k-means clustering algorithm for real time sensor data. In 2016 IEEE Region 10 Conference (TENCON), pages 1945–1949. IEEE, 2016.

[8] C. G. Kirwan and F. Zhiyong. Smart cities and artificial intelligence: convergent systems for planning, design, and operations. Elsevier, 2020.

[9] Z. Lv, D. Chen, R. Lou, and Q. Wang. Intelligent edge computing based on machine learning for smart city. Future Generation Computer Systems, 115:90–99, 2021.

[10] U. Mahalingam, K. Elangovan, H. Dobhal, C. Valliappa, S. Shrestha, and G. Kedam. A machine learning model for air quality prediction for smart cities. In 2019 International conference on wireless communications signal processing and networking (WiSPNET), pages 452–457. IEEE, 2019.

[11] S. Mahanta, T. Ramakrishnudu, R. R. Jha, and N. Tailor. Urban air quality prediction using regression analysis. In TENCON 2019-2019 IEEE Region 10 Conference (TENCON), pages 1118–1123. IEEE, 2019.

[12] H. Maleki, A. Sorooshian, G. Goudarzi, Z. Baboli, Y. Tahmasebi Birgani, and M. Rahmati. Air pollution prediction by using an artificial neural network model. Clean technologies and environmental policy, 21:1341–1352, 2019.

[13] K. Nandini and G. Fathima. Urban air quality analysis and prediction using machine learning. In 2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE), pages 98–102. IEEE, 2019.

[14] P. J. Navarathna and V. P. Malagi. Artificial intelligence in smart city analysis. In 2018 International conference on smart systems and inventive technology (ICSSIT), pages 44–47. IEEE, 2018.

[15] G. Oliveri Conti, B. Heibati, I. Kloog, M. Fiore, and M. Ferrante. A review of airq models and their applications for forecasting the air pollution health outcomes. Environmental Science and Pollution Research, 24:6426–6445, 2017.

[16] J. W. Park, C. H. Yun, H. S. Jung, and Y. W. Lee. Visualization of urban air pollution with cloud computing. In 2011 IEEE world congress on services, pages 578–583. IEEE, 2011.

[17] V. R. Pasupuleti, P. Kalyan, H. K. Reddy, et al. Air quality prediction of data log by machine learning. In 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), pages 1395–1399. IEEE, 2020.

[18] R. M. Patil, H. Dinde, S. K. Powar, and P. M. Ganeshkhind. A literature review on prediction of air quality index and forecasting ambient air pollutants using machine learning algorithms. International Journal of Innovative Science and Research Technology, 5(8), 2020.

[19] H. Peng, A. R. Lima, A. Teakles, J. Jin, A. J. Cannon, and W. W. Hsieh. Evaluating hourly air quality forecasting in canada with nonlinear updatable machine learning methods. Air Quality, Atmosphere & Health, 10:195–211, 2017.

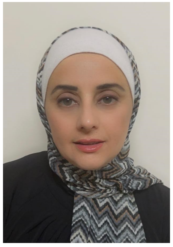
[20] R. Sharma, G. Shilimkar, and S. Pisal. Air quality prediction by machine learning. 2021.

[21] S. Simu, V. Turkar, R. Martires, V. Asolkar, S. Monteiro, V. Fernandes, and V. Salgaoncary. Air pollution prediction using machine learning. In 2020 IEEE Bombay Section Signature Conference (IBSSC), pages 231–236. IEEE, 2020.

[22] Z. Zhang, H. Chen, and X. Huang. Prediction of air quality combining wavelet transform, dcca correlation analysis and lstm model. Applied Sciences, 13(5):2796, 2023.

VI. ACKNOWLEDGEMENT

The authors are thankful to the Deanship of Scientific Research at Najran University for funding this work under the Research Groups Funding program grant code (NU/RG/SERC/12/9).



SHOROUQ AL-EIDI is an Assistant Professor at the computer science department in computer science department, at Tafila Technical University. She received here PhD degree in computer science from Memorial University of Newfoundland in Canada, and the MS degree in computer science from Jordan University of Science and Technology in Jordan. Her research interests include: cyber security, machine learning, networks, and big data analysis.

FATHI AMSAAD received the Ph.D. degree in engineering science from Toledo (UToledo), OH, USA. He is currently an Assistant Professor with the School of Information Security and Applied Computing (SISAC), Eastern Michigan University (EMU). He is also the Founder and the Director of the Cyber Security Laboratory, and the Co-Director of the Advanced Computing Research Laboratory. His research expertise include in the areas of cyber security and cyber-physical systems with special interests in hardware-oriented security and trust for device and system authentication, secure embedded architectures, VLSI/FPGA systems testing, fault tolerance hardware, detection and prevention of hardware trojans, network and mobile wireless security, and the security of IoT applicaions and smart systems. He is also an active IEEE/ACM member. He has served as a Project Adviser for several groups of senior undergraduate students and a Reviewer of high impact and peer-review conferences/journals. He was a recipient of the prestigious IEEE Best Graduate Student Award by IEEE Region 4 and the College of Engineering, UToledo. Additionally, he was the 2017 nominee for the best Ph.D. Dissertation Award. He holds MCP, MCSA, MCTS, and MCSE Professional Certificates from Microsoft

OMAR DARWISH is an Assistant Professor in the Information Security and Applied Computing department (Game Above College of Engineering and Technology) at Eastern Michigan University. He received his PhD degree in computer science from Western Michigan University in USA, and his MS degree from Jordan University of Science and Technology in Jordan. He worked as an Assistant Professor, Program Coordinator of Computer Information Systems, and Director of the IoT and Cybersecurity Lab at Ferrum College, Visiting Assistant Professor at West Virginia University Institute of Technology, a Software Engineer at MathWorks, and a programmer at Nuqul group. His research interests include cyber security, IoT, machine learning, networks, big data analysis, cloud computing, artificial intelligence, data mining, and information retrieval.

YAHYA TASHTOUSH is a Full Professor at the College of Computer and Information Technology, Jordan University of Science and Technology (JUST), Irbid, Jordan. He received his B.Sc. and M.Sc. degrees in Electrical Engineering from JUST in 1995 and 1999, respectively. He received his Ph.D. degree in Computer Engineering from the University of Alabama in Huntsville and the University of Alabama at Birmingham, AL, USA in 2006 (Joint Degree). His current research interests are IoT, Deep/Machine Learning, Wireless Networks, Robotics and Fuzzy Systems.

ALI ALQAHTANI received a Ph.D. in computer engineering from Oakland University, Rochester Hills City, MI, USA, in 2020. He is currently an Assistant Professor at Najran University (NU). His research interests include machine learning in general and deep learning in image and signal processing, wireless vehicular networks (VANETs), wireless sensor networks, and cyber-physical systems.

NIVESHITHA NIVESHITHA is a graduate student at wright state university. His research interests include artificial intelligence, machine learning, and cloud computing.