Journal of Engineering and Sustainable Development
https://jeasd.uomustansiriyah.edu.iq/index.php/jeasd
https://doi.org/10.31272/jeasd.28.3.9

Vol. 28, No. 03, May 2024
ISSN 2520-0917

Original Research

HYBRID MODEL AND FRAMEWORK FOR PREDICTING AIR POLLUTANTS IN SMART CITIES

Qutaiba Humadi Mohammed[1*], Anupama Namburu[2]

[1]*College of Nursing, University of Baghdad, Baghdad, Iraq*
[2]*School of Engineering, Jawaharlal Nehru University, New Delhi 110067, India*
[1]https://orcid.org/0000-0001-9837-3320
[2]https://orcid.org/0000-0002-7826-0158

Received 25/10/2023 ***Revised 23/02/2024*** ***Accepted 21/03/2024***

Abstract: The pollution index of any urban area is indicated by its air quality. It also shows a fine balance is maintained between the needs of the populace and the industrial ecosystem. To mitigate such pollution in real-time, smart cities have a significant role to play. It's common knowledge that air pollution in a city severely affects the health of its dependents. More alarmingly, human health damage and disease burden are caused by phenomena like acid rain, and global warming. More precisely, lung ailments, CPOD, heart problems and skin cancer are caused by polluted air in congested urban places. Amongst the worst air pollutants, CO, C6H6, SO_2, NO_2, O_3, RSPM/PM10, and $PM_{2.5}$ cause maximum havoc. The climatic variables like atmospheric wind velocity, direction, relative humidity, and temperature control air contaminants in the air. Lately, numerous techniques have been applied by researchers and environmentalists to determine the Air Quality Index over a place. However, not a single technique has found acceptance from all quarters as being effective in every situation or scenario. Here, the main aspect relates to achieving authentic prediction in AQI levels by applying Machine Learning algorithms so worst situations can be averted by timely action. To enhance the performance of Machine Learning methods study adopted imputation and feature selection methods. When feature selection is applied, the experimental outcomes indicate a more accurate prediction over other techniques, showing promise for the application of the model in smart cities by syncing data from different monitoring stations.

Keywords: *Air Pollution; Air Quality Index; Machine Learning; Smart Cities*

1. Introduction

The most critical natural resource that ensures the survival of all species on the earth is air. Air is essential for all living things, both plants and animals alike. Hence, to exist, all living things require clean air quality free from harmful particles. The climatic variables like ambient air circulation, direction, humidity, and temperature control contaminants in the surrounding atmosphere [1]. We experience significantly greater heat as moisture doesn't escape into the air at higher humidity. Primarily, mindless urbanization causes air pollution on account of rising vehicular movement that releases more pollutants into the atmosphere. Industrialization causes massive air pollution [2-3]. Among pollutants, NO, CO, PM, and SO_2 largely pollute the air. Carbon Monoxide is created as a result of inadequate oxidation of propellants such as gas, oil, and other types of fuel. Nitrogen oxide is formed when thermal fuel burns and induces vomiting and dizziness [4-7]. Smoking produces benzene, thereby irritating the respiratory system. To reduce environmental air pollution, steps must

*Corresponding Author:
Qutaiba.h@conursing.uobaghdad.edu.iq

be taken. AQI gauges air in qualitative terms. In the past, conventional techniques like probability and statistics were applied to gauge air quality. In recent times, newer technologies easily collected data by applying sensors to measure air pollutants. More recent studies have shown the potential of AQI predictive models based on Machine Learning and found them more dependable. Vast air pollution data can be effectively handled by ML algorithms for accuracy and reliability in prediction [8-9]. Data on the most significant air pollutants, including SO_2, NOx, NO, NO_2, and CO, were gathered in airspace applying several fixed stations for the current study [10-11]. This work will make an assessment and evaluation of air pollution over the study area through ML techniques.

2. Related Works

A novel outdoor air quality monitoring device is being researched, and tested in a limited capacity, and has shown some promise. The obtained data on carbon dioxide, nitrogen dioxide, Sulphur dioxide, ozone, particle matter, temperature, and humidity are the key contributions of this work. Furthermore, we use airflow as the energy model, define the dilution and diffusion coefficients of pollution diffusion, and, by employing the technique of pollution tracing, obtain the exact pollution monitoring across the city by local stations [12].

2.1. Air Pollution Index

Most nations recognize API as the AQI. As per experts, API serves as a baseline for the state of the air in a certain location. $PM_{2.5}$, PM_{10}, SO_2, O_3, CO, NO_2, and ammonia are the seven pollutant parameters used to compute API (NH3). To enhance the quality of the air in smart cities, the air pollutants must be precisely assessed, watched over, and regulated. The average computation for each parameter is taken over a separate time to identify the API [13-15]. Because human tolerance for different pollutant exposure times in everyday living varies, measurements are made in this fashion [3]. AQI is calculated as:

$$\mathbf{Ip=[IHi-ILo/BPHi-BPLo](Cp-BPLo)+ILo}$$

where,

Ip = pollutant p index

Cp = condensed concentration of pollutant p

BPHi = concentration breakpoint i.e., > or = Cp

BPLo = concentration breakpoint i.e., < or = Cp

IHi = AQI value related to BPHi

ILo = AQI value related to BPLo

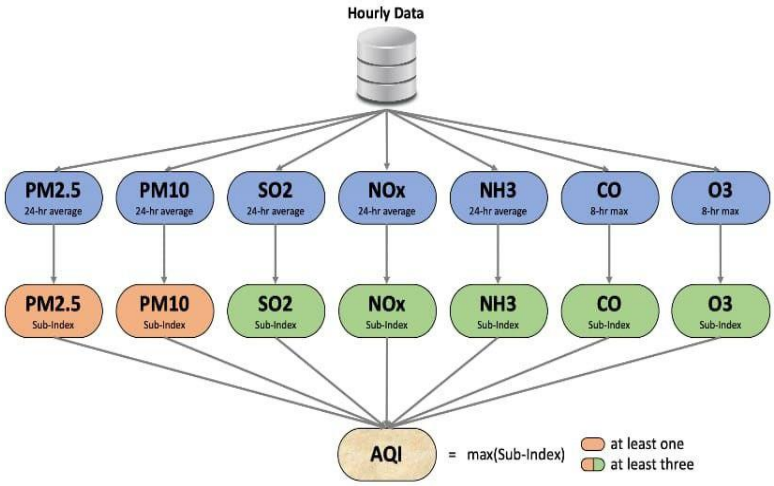

Figure 1. The air pollutant information

Fig.1 shows the government organizations containing public information about air contamination levels, both current and projected. People may experience more serious health impacts when the AQI rises as a larger populace has likely exposure [16], and Table 1 explains the classification for AQI.

Table 1. AQI Classification

AQI	Air Pollution Level
0-50	Excellent
51-100	Good
101-150	Light Pollution
151-200	Moderate Pollution
201-300	Heavy Pollution
300+	Severe Pollution

2.2. Air Quality Standards

Primary Standards: Prevents the occurrence of dangerous health conditions.

Secondary standards: Protects damages to crops, flora and fauna as well as habitats.

Standards for some contaminants' short- and long-term average durations are in place. The long-term standards safeguard from long-lasting health consequences, whereas the short-term requirements help in avoiding acute medical conditions [17-18]. The modelling of air pollution problems has, up to this point, mostly relied on dispersion models, which approximate the intricate physicochemical processes that occur, as per researchers E. Kalapanidas and N. Avouris [9]. Although the intricacy and advancement of such models have grown over time, their usage in real-time air pollution monitoring is not great in meeting input data needs, and overcoming limitations of time. Nevertheless, due to their inability to foresee catastrophic situations, such sorts of procedures have a limited degree of accuracy [19-20].

Table 2. Lists All Criteria Pollutants and Standards

Pollutant	Primary/ Secondary	Averaging Time	Level	Form
Carbon Monoxide (CO)	Primary	8 hours 1 hour	9 ppm 35 ppm	Not exceedingly more than once per year
Lead (Pb)	Primary and Secondary	Rolling 3 Month average	0.15μg/m^3	Not to exceed
Nitrogen Dioxide (NO_2)	Primary	1-hour 1 year	100ppb 53 ppb	98th percentile of daily 1-hour optimal concentration with a 3-year average. Annual Mean
Ozone (O_3)	Primary and Secondary	8 hours	0.07 ppm	Annual fourth highest daily maximum 8-hour concentration with 3-year average

Table 2 shows the criteria for pollutants and standards. The application layer and the sensing/network layer are the two main areas of attention in the literature on air quality monitoring systems. Toxic gases (NO_2, O_3, CO, SO_2, H_2S), particle matter, and climatic factors are often targeted elements [21].

Gopalakrishnan (2021) conducted ML experiments on air quality over Oakland, California by analyzing Google Street View data with a special focus on missing data locations. In this work, a web application demonstrated air quality predictions across all corners within city limits [22].

Sanjeev (2021) studied meteorological and pollutant data using a Random Forest (RF) classifier, which the author vouched as best performing due to its low susceptibility against over-fitting, was used to examine and forecast the air quality [23].

Castelli et al. (2020) used the Support Vector Regression (SVR) machine learning technique to attempt to anticipate Californian air quality based on contaminants and particles. The authors asserted to have created a brand-new approach to simulate hourly air pollution [24]. To anticipate PM concentration in the air, Doreswamy et al. (2020) looked into ML prediction models. The scientists examined Taiwanese air quality monitoring data by using known models while asserting that the measured values and projected values were fairly similar [25-26]. Depending on 11 years of data, Liang et al. (2020) examined ML classifiers' actions in forecasting Taiwan's AQI. The best methods for predicting air quality are AdaBoost and Stacking Ensemble, albeit the accuracy of these methods varies by area [26-27].

3. Materials and Methods

Here, the Knowledge Discovery from Databases (KDD) methodology is used to garner the required data for decision-making in air quality management. Primarily, this framework is designed for extraction of the implicit, hitherto unknown, and promising information from real data as well as prediction of the air quality to make effective decisions. The proposed methodology consists of (a) Data preprocessing (b) Feature Selection (c) Modeling (d) Evaluation. The next section explains all stages in this process [28].

3.1. Air Quality Data Source

Stationary weather stations track air quality over Tehran every hour. The Tehran Municipality's Air Quality Control Company (AQCC) is in charge of keeping an eye on certain air pollutants in Tehran City. There were 22 AQCC-affiliated monitoring sites in Tehran at the time of this study, in 2022. We chose data from AQCC-affiliated air quality monitoring stations in this study because they provide hourly data that is available online and accessible to the general public, as well as because they are geographically distributed throughout all districts of Tehran. At the Tehran city's air quality monitoring stations, $PM_{2.5}$ and PM_{10}, NO_2, O_3, CO, and SO_2 are analyzed through beta-attenuation (Met One BAM-1020, USA; and Environment SA, MP 101 M, France), chemiluminescence (Ecotech Serinus 40 Oxides of Nitrogen Analyzer, Australia), UV-spectrophotometry (Ecotech Serinus 10 Ozone Analyzer, Australia), non-dispersive infrared Next, individuals' 3-year residence data for six outdoor criterion air pollutants were gathered hourly from the AQCC website [29].

3.2. Air Quality Data Processing

The most crucial step in analyzing air quality data and determining health consequences is data quality management. The WHO and EPA requirements for data quality assurance were followed. Missing value detection and feature selection from monitoring stations is crucial since there are many operational and calibration issues with air pollution measurement stations. If this phase is skipped, the data will not have enough scientific validity.

3.2.1. Missing Data Imputation

Proposed Algorithm: This section through Fig. 2 introduces our approach, CMI (Correlation-based Missing Data Imputation). By creating a linear regression model, our technique imputes the missing values of an attribute.

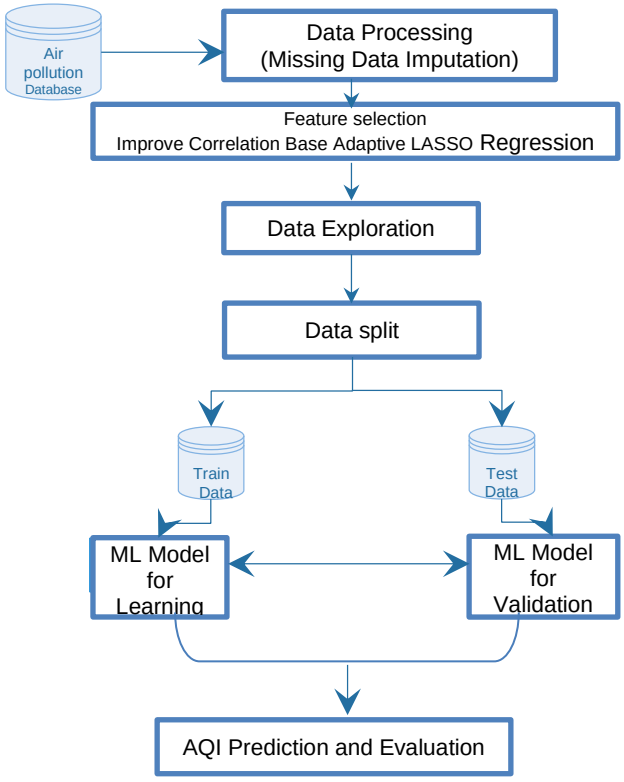

Figure2. The proposed Methodology

The top three (K=3) qualities in the dataset that have the highest association with the target column (column with missing values) are known as predictors.

3.2.2. Feature Selection

The analysis is a difficult undertaking since there are many attributes in the air quality data. An essential phase in the modelling and decision-making process is the selection of the most crucial attributes. Most feature selection techniques aim to minimize the number of

dimensions in the target dataset to facilitate and improve analysis. Since datasets used for air quality analysis include an intermediate count of features, it has been shown that feature importance or ranking approaches are more useful for determining which contaminants to pay attention to. When the overall regression coefficients must be less than a predetermined fixed value, the LASSO Regression technique conducts both regularization and feature selection. By utilizing the tuning parameter, which stands for a penalty, it decreases the dimensionality. Although LASSO has accurate subset selection for high regression coefficients, it does not have ideal prediction rates and even has wrong predictor selection while estimating maximum rates [30][31].

Algorithm 1: Correlation based Missing Data Imputation

Input: X_{miss} : Missing Values Columns
D_{obs}: Dataset without Missing Values
D_{miss}: Missing Values Dataset
Output: D_{obs}: Imputed Dataset.
1. for every Missing Column X_{miss} in $x_{\text{miss}}^{(i)}$ do :
2. for every column $x_{\text{miss}}^{(j)}$ in the dataset do :
Correlation Coefficient is calculated within $x_{\text{miss}}^{(i)}$ and $x_{\text{miss}}^{(j)}$ and then stored it in C
end
3. M = Top 3 columns X_{miss} with highest correlation X_{miss}
4. Train a regression model on predictor variables p_1, p_2, p_3 and target X_{miss}.
5. Calculate Loss Function:
$\mathbb{U}$ = a vector with correlation coefficient between predicted value λ and dataset with the only p_1, p_2, p_3 attributes u ; $u[p_1]$, λ, $u[p_2]$, λ and $u[p_2]$, λ.
Loss=E+ML$(C||\mathbb{U})$
E = cross entropy loss between observed value x predicted value λ
5. Predict the values of D_{miss} using the model trained in 4.
6. end

Adaptive LASSO removes the LASSO regression issues due to its unique properties. This approach minimizes the residual sum of squares on the basis of the sum of the absolute value of the coefficients at lesser than a constant that is identical to keeping the sum of squares with a constant $\sum|\beta_j| \leq s$ to the minimum, while certain coefficients shrink to zero. The method applies L1 regularization to the objective under optimization by imposing a penalty. The total absolute values of the coefficients determine the penalty and determine the targeted coefficients and their possible shrinking. The LASSO estimates:

$$\beta_{AL} = \min_{\beta}\left(\sum_{i=1}^{n}\left(y_i - \sum_j \beta_j X_{ij}\right)^2 + \lambda \sum_j \hat{w}_j \left|\beta_j\right|\right) \qquad (1)$$

where λ is termed as the shrinkage parameter and $w_j = \frac{1}{|\hat{\beta}_j|^{\gamma}}, (j=1,...,p)$ are the weight functions, γ is a positive constant and $\hat{\beta}_j$ are the initial estimates of β coefficients. The Adaptive LASSO for consistent variable selection is combined with Maximum Tangent Likelihood Estimation (MTE) regression in the MTE-LASSO criteria.

$$\beta = \underset{\beta \in R^d}{\mathbf{arg\ max}}\left\{\mathbf{L}(\beta) + \lambda_n \sum_{j=1}^{d} \mathbf{W}_j\left(\left|\beta_j\right|\right)\right\} \qquad (2)$$

Where $L(\beta)$ is the MTE loss function defined as:

$$\beta = \min_{\beta} \sum_{i=1}^{n} H_M\,(y_i\text{-}x_i^T\beta) + \lambda \sum_{i=1}^{p} \widehat{w}_j|\beta_j| \quad (3)$$

where $\widehat{w}_j = (w_1, w_p)$ is known as weight function.

$$\beta = \arg\max_{\beta \in R^d} \left\{ \sum_{i=1}^{n} \ln_t\left(f(z_i:\beta)\right) - n\sum_{j=1}^{d} p\lambda_{nj}\left(|\beta_j|\right) \right\} \quad (4)$$

$\lambda_n$denotes regularization parameter of L1 penalty and where $W_j = (w_1, \ldots, w_p)$ is referred to as weight vector.

3.2.3. Model Evaluation

The outcomes should be validated and compared when the strategy is put into practice and modelled. As a consequence, the findings of this study were estimated by applying the coefficient of determination and root-mean-square error.

$$R^2 = \left[\frac{1}{N}\frac{\sum_{i=1}^{N}[(P_i-\bar{P})(O_i-\bar{O})]}{\sigma_p \sigma_o}\right]^2 \quad (5)$$

$$\text{RMSE} = \left(\frac{1}{N}\sum_{i=1}^{N}[P_i - O_i]^2\right)^{\frac{1}{2}} \quad (6)$$

where Nstands for observations; $O_i$refers to observed parameter; P_i represents the calculated parameter;$\bar{O}$denotes mean of the observation parameter; $\bar{P}$ represents the average calculation parameter;$\sigma_o$refers to the standard deviation of observations; and $\sigma_p$represents calculation standard deviation [32][33][34]. In this study, a portion of the 20% test dataset was applied to evaluate the performance of different models, and validation results were compared using several validation indicators, such as accuracy, precision, sensitivity (recall) and F1-score [35][36].

$$\text{Precision} = \frac{T_p}{T_p + F_p}$$

$$\text{Recall} = \frac{T_p}{T_p + F_N}$$

$$\text{Accuracy} = \frac{T_p + T_N}{T_p + T_N + F_p + F_N}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

4. Results and discussion

4.1. Results of Missing Data Imputation

4.1.1. Data preprocessing

The Karaj, Shahryar, and Rey monitoring stations display 91,672, 94,453, and 94,145 raw data points, respectively. The preprocessing procedures help to reduce data noise, thereby speeding up processing and expanding scope of ML algorithms. Frequently, outliers and missing data affect data extraction and monitoring applications. Actions like filling in missing data (NAN), and eliminating or modifying outlier data are done during the data preparation stage. You'll see that out of all the characteristics, PM_{10} and $PM_{2.5}$ have the most missing data, whereas CO has the fewest. To resolve the missing data issue, all of the missing values are imputed using correlation-based missing data imputation (CMI) for each feature.

Main Cities=Karaj, Shahryar, Rey

Other Cities=Mahallati, Punak, Piroozi, Darous, Shadabad, Rey, Golbarg, and Massoudieh Station.

Missing data in = PM10 and PM2.5

Table 3. Results of ML Algorithms for KarajAQI Prediction.

Method	with CMI Imputation			without Imputation		
	RMSE	MAE	R^2	RMSE	MAE	R^2
SVM	6.722	5.165	0.943	7.582	5.990	0.953
Random Forest	3.022	2.185	0.982	3.224	2.386	0.975
Decision Tree	3.204	2.177	0.982	3.411	2.574	0.978
NB	3.536	2.372	0.980	3.744	2.614	0.971
MLP	3.103	2.071	0.984	3.294	2.171	0.980

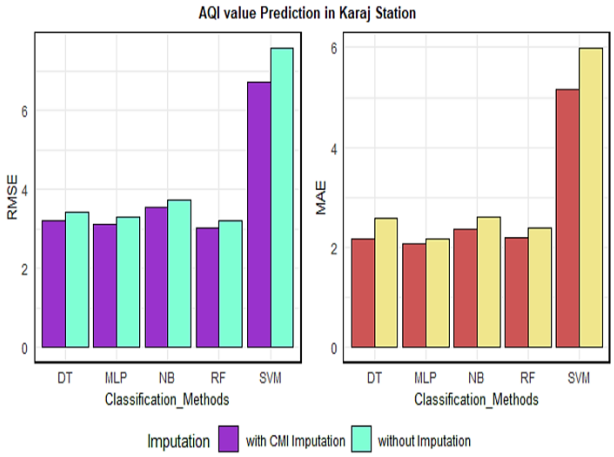

Figure.3. Results of ML Algorithms for Karaj AQI Prediction

Table 3 and Fig. 3 show evaluation outcomes related to Karaj AQI prediction employing five approaches (SVM, Random Forest, MLP, NB, and Decision Tree) with and without CMI imputation. Machine learning algorithms fared exceptionally in making prediction of AQI levels in Karaj. The MLP is proven to be the best technique for predicting AQI level, with results of 0.984 when CMI imputation is used and 0.980 when no imputation technique is used. The imputation helps RF to improve all evaluation measures. Furthermore, MLP outperforms RF and DT in terms of MAE score, whereas RF dominates those two techniques in terms of RMSE score. The MLP generates MAE scores of 2.071 and 2.171 with and without imputation, respectively. In all three-evaluation metrics SVM is producing negative results compared to all the four methods.

Table 4. Results of ML Algorithms for Shahryar AQI Prediction.

Method	with CMI Imputation			without Imputation		
	RMSE	MAE	R^2	RMSE	MAE	R^2
SVM	9.077	6.984	0.940	9.933	7.562	0.930
Random Forest	3.028	2.034	0.984	3.074	2.045	0.984
Decision Tree	3.062	2.026	0.984	3.095	2.033	0.984
NB	3.812	2.483	0.981	3.831	2.516	0.981
MLP	3.009	2.023	0.986	3.045	2.031	0.985

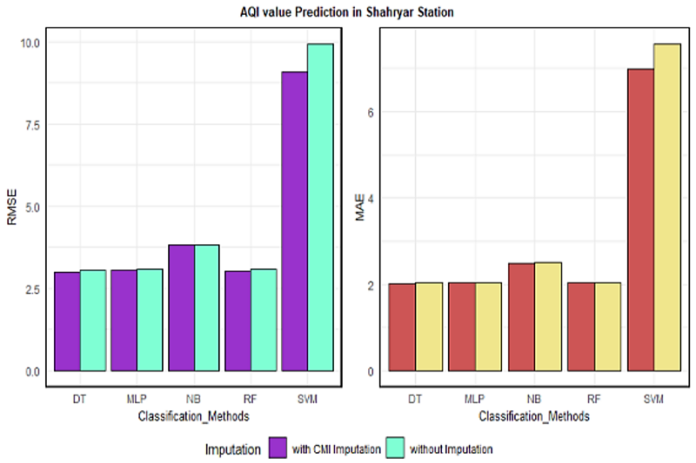

Figure 4. Results of ML Algorithms for Shahryar AQI Prediction

The AQI prediction results of Shahryar adopting the five approaches (SVM, Random Forest, MLP, NB, and Decision Tree) with and without CMI imputation are shown in Table 4 and Fig. 4 Inferring from the results; it may be concluded that machine learning algorithms did an outstanding job of predicting AQI levels in Shahryar station. With results of 0.986 with the CMI imputation and 0.985 without the use of any imputation technique, the MLP is shown to be the best technique for predicting AQI level. All evaluation measures can be improved by MLP due to imputation. Additionally, MLP outperforms than RF and DT in terms of MAE

and RMSE scores. When compared to the RF, DT, and NB, the RMSE scores for the MLP are 3.009 and 3.045. MLP generates 2.023 and 2.031of MAE score with and without imputation and is better value compared RF and DT. In all three-evaluation metrics SVM is producing negative results compared to all the four methods

Table 5. Results of ML Algorithms for Rey AQI Prediction.

Method	with CMI Imputation			without Imputation		
	RMSE	MAE	R^2	RMSE	MAE	R^2
SVM	8.421	6.326	0.963	9.434	6.555	0.953
Random Forest	2.941	1.820	0.982	2.949	1.849	0.980
Decision Tree	2.920	1.613	0.983	2.966	1.748	0.981
NB	3.526	2.019	0.982	3.783	2.559	0.982
MLP	2.496	1.405	0.986	2.892	1.796	0.983

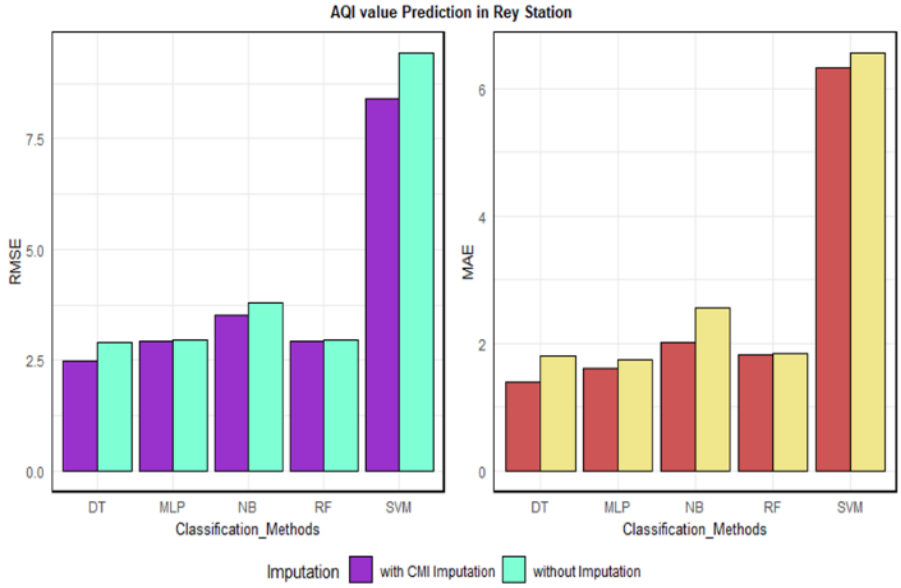

Figure 5. Results of ML Algorithms for Ray AQI Prediction

Table 5 and Fig. 5 show the assessment results of Ray AQI prediction employing five approaches (SVM, Random Forest, MLP, NB, and Decision Tree) with and without CMI imputation. In Ray station, machine learning algorithms fared better in predicting AQI levels. The MLP is shown to be the most accurate technique for predicting AQI level, with results of 0.986 when CMI imputation is used and 0.983 when no imputation technique is used. The imputation allows MLP to improve all assessment measures. Furthermore, MLP outperforms RF and DT in terms of MAE and RMSE scores. In the case of the MLP, the RMSE values of 2.496 and 2.892 are better than those of the RF, DT, and NB. The MLP generates 1.405 and 1.796 MAE scores with and without imputation and is a better value compared to RF and DT. In all three-evaluation

metrics, SVM is producing negative results compared to all the four methods.

4.1.2. AQI Prediction without Feature Selection

In the training stage, adopted ML models produced results related to accuracy, precision, recall, and F1-score as provided in Table 6 and Fig. 6 below. The recall denotes the percentage of applicable cases retrieved; precision indicates the percentage of relevant instances that are found in retrieved cases. The ratio of precisely identified attributes to the entire set of variables denotes accuracy. A weighted average of recall and precision is the F1-score. The SVM model had the lowest accuracy, whereas the MLP model had the highest accuracy.

Table 6. Comparison of model results in the training set

ML Methods	Accuracy	Precision	Recall	F1-score
SVM	80	89	86	85
RF	85	91	94	88
DT	81	90	93	88
NB	88	93	88	92
MLP	93	95	95	96

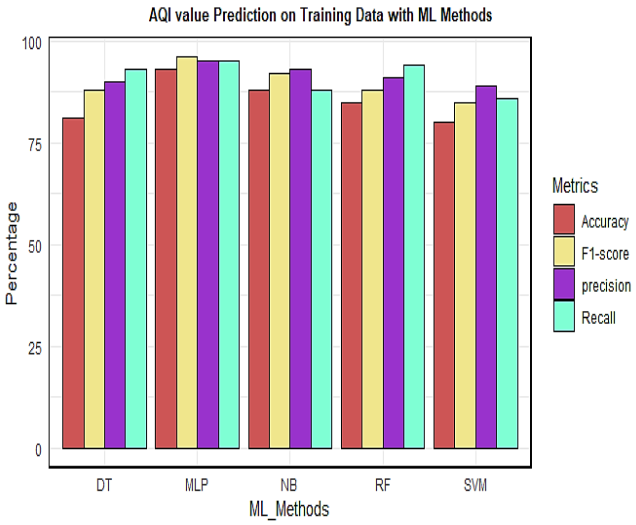

Figure.6. AQI value Prediction on Training Data without Feature Selection

The projected ML models produced results from testing stage as provided in Table 7 and Fig. 7 below. The MLP model once again outperformed the competition, while the SVM model also achieved least accuracy during testing period as exhibited in Table 7. The testing set's ML model performances are assessed using the standard performance metrics.

Table 7. Comparison of model results in the testing set

ML Methods	Accuracy	Precision	Recall	F1-score
SVM	81	86	85	81
RF	83	88	89	83
DT	78	91	90	78
NB	86	92	91	86
MLP	90	96	95	90

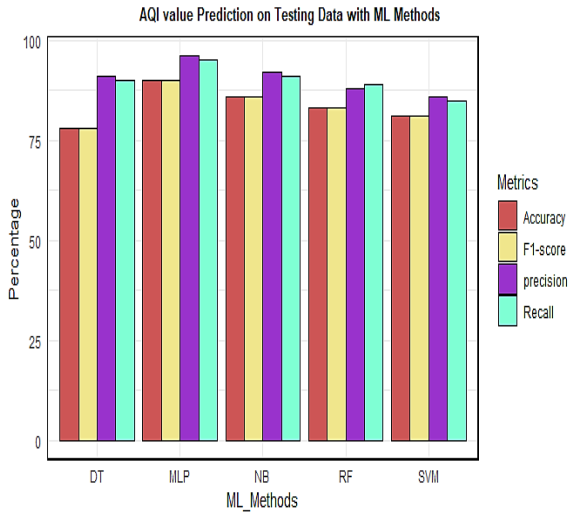

Figure 7. AQI value Prediction on Testing Data without Feature Selection

4.1.3. AQI Prediction with Feature Selection

The results of the adopted ML models with feature selection (i.e., Relief method) are shown in Table 8 and Fig. 8 in terms of accuracy, precision, recall, and F1-score in the training phase. In a given machine learning problem,

feature selection determines the collection of characteristics that determine the prediction accuracy of the target variables/class labels.

Identification of pertinent features enhances the overall performance of machine learning (ML) models and makes it easier to understand the data about the ML model. It should be noted that after applying the feature selection to the individual methods still the SVM model had the least accuracy, whereas the MLP model had the highest accuracy and this value is much better than the standard MLP.

Table 8. Comparison of model results in the training set

ML Methods	Accuracy	Precision	Recall	F1-score
SVM	91.31	96.11	92.27	98.03
RF	87.47	93.23	96.11	90.35
DT	83.62	92.27	95.15	90.35
NB	90.35	95.15	90.35	94.19
MLP	93.23	97.07	97.07	93.23

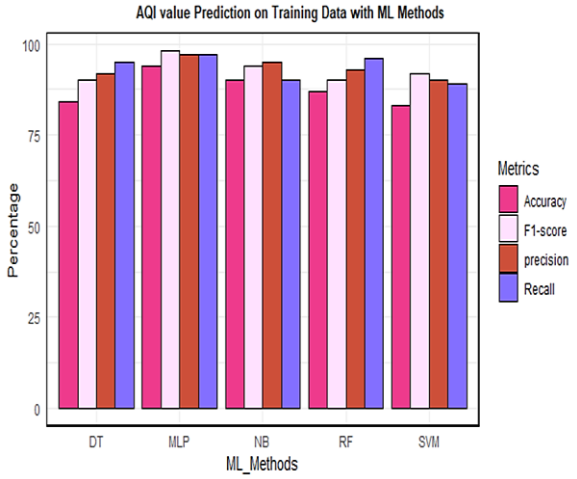

Figure 8. AQI value Prediction on Training Data with Feature Selection

The results of the tested machine learning models with feature selection are provided in Table 9 and Fig. 9 below. The MLP model once again outperformed the competition, while the SVM model also achieved least accuracy during testing period.

Table 9. Comparison of model results in the testing set

ML Methods	Accuracy	Precision	Recall	F1-score
SVM	92.08	96.84	93.04	98.74
RF	88.28	93.99	96.84	91.13
DT	84.48	93.04	95.89	91.13
NB	91.13	95.89	91.13	94.94
MLP	93.99	97.79	97.79	93.99

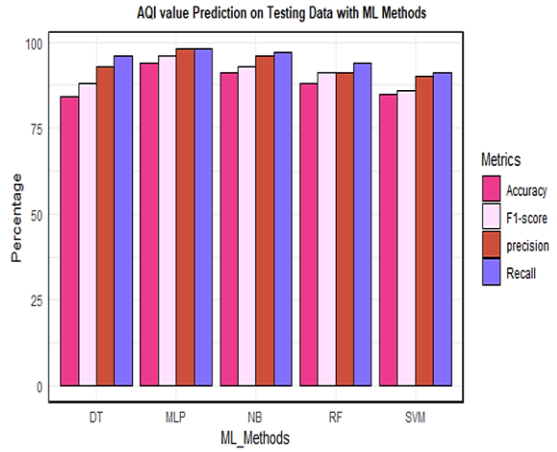

Figure 9. AQI value Prediction on Testing Data with Feature Selection

The previous tables show the results of several ML models used on the training and testing sets with feature selection techniques as well as alone. When used with feature selection, it has been seen that all ML models improved on almost all evaluation measures. In error statistics, the MLP model performed best with the most optimum performance values.

5. Conclusions

This study manages to accurately predict AQI in smart cities with the use of Machine Learning

Algorithms. To enhance Machine learning algorithms' performance levels, we have adopted a set of pre-processing techniques like imputation and feature selection methods. The experimental outcomes showed more accurate prediction with imputation and feature selection in each station of the smart city. The MLP method performed better in comparison to other methods like DT, RF and NB. It is noticed that the SVM method is performing poor performance in all aspects. The study is limited to only three monitoring stations Karaj, Shahryar, and Rey and further extended to more stations. Also, in future, it is extended today, month and year-wise analysis of air pollution in smart cities.

Acknowledgements

The author(s) would like to thank (Baghdad and Mustansiriyah) Universities in Iraq. Also, Acharya Nagarjuna University in India.

Abbreviations

AQCC	Air Quality Control Company
AQI	Air Quality Index
CMI	Conditional Mean Imputation
DT	Decision Trees
KDD	Knowledge Discovery in Databases
ML	Machine Learning
MAE	Mean Absolute Error
MLP	Multilayer Perceptron
NB	Naive Bayes
RF	Random Forest
RMSE	Root Mean Squared Error
SVM	Support Vector Machine
SVR	Support Vector Regression

Conflict of interest

The authors confirm that there is no conflict of interest.

Author Contribution Statement

Qutaiba Humadi Mohammed: proposed the research problem, developed the theory and performed the computations.

Anupama Namburu.: verified the analytical methods and investigated and supervised the findings of this work.

Both authors discussed the results and contributed to the final manuscript.

References

1. Wark, K.;Warner, C.F. Air Pollution: Its Origin and Control; Harper and Row: New York, NY, USA, 1981, Book:ISBN:9780700225347
2. Seinfeld, J.H.; Pandis, S.N. Atmospheric Chemistry and Physics: From Air Pollution to Climate Change; JohnWiley& Sons: New York, NY, USA, 1998. https://doi.org/10.1063/1.882420
3. Mlakar, P.; Boznar, M. Perceptron neural network-based model predicts air pollution. In Proceedings of the Intelligent Information Systems, Grand Bahama Island, Bahamas, 8–10 December 1997; pp. 345–349. https://doi.org/10.1109/IIS.1997.645288
4. Murtadah, I., Al-Sharify, Z. T., & Hasan, M. B. (2020, June). Atmospheric concentration saturated and aromatic hydrocarbons around Dura refinery. In *IOP Conference Series: Materials Science and Engineering* (Vol. 870, No. 1, p. 012033). IOP Publishing. https:// https://doi.org/10.1088/1757-899X/870/1/012033
5. Abdulhussein, Z. A., Al-Sharify, Z. T., Alzuraiji, M., & Onyeaka, H. (2023). Environmental significance of fouling on the crude oil flow. A comprehensive

review. *Journal of Engineering and Sustainable Development (JEASD), 27(3*). https://doi.org/10.31272/jeasd.27.3.3

6. Aalhashem, N. A., Naser, Z. A., Al-Sharify, T. A., Al-Sharify, Z. T., Al-Sharify, M. T., Al-Hamd, R. K. S., & Onyeaka, H. (2022, November). Environmental impact of using geothermal clean energy (heating and cooling systems) in economic sustainable modern buildings architecture design in Iraq: A review. In AIP Conference Proceedings (Vol. 2660, No. 1). AIP Publishing. https://doi.org/10.1063/5.0109553
7. Muhaisn, L. F., Naser, Z. A., Nayel, D. H., Al-Sharify, Z. T., & Muhaisn, F. F. (2023, July). Investigate the environmental impact of aircraft on the Earth's atmosphere and analyzing its effect on air and water pollution. In AIP Conference Proceedings (Vol. 2787, No. 1). AIP Publishing. https://doi.org/10.1063/5.0150150
8. AL-Bakri, N. F., & Hashim, S. H. (2019). A study on the accuracy of prediction in recommendation system based on similarity measures. *Baghdad Science Journal*, *16*(1 Supplement), 263-269. http://dx.doi.org/10.21123/bsj.2019.16.1(Suppl.).0263
9. Salih, N. Z., & Khalaf, W. (2021). Improving students performance prediction using machine learning and synthetic minority oversampling technique. *Journal of Engineering and Sustainable Development*, *25*(6), 56-64. https://doi.org/10.31272/jeasd.25.6.6
10. Brunelli, U.; Piazza, V.; Pignato, L.; Sorbello, F.; Vitabile, S. Three hours ahead prevision of SO2 pollutant concentration using an Elman neural based forecaster. Build. Environ. 2008, 43, 304–314. https://doi.org/10.1016/j.buildenv.2006.05.011
11. Anad, A. M., Hassoon, A. F., & Al-Jiboori, M. H. (2022). Assessment of air pollution around Durra refinery (Baghdad) from emission NO2 gas at April Month. *Baghdad Science Journal*, *19*(3), 0515-0515. http://dx.doi.org/10.21123/bsj.2022.19.3.0515
12. Singh, K.P.; Gupta, S.; Rai, P. Identifying pollution sources and predicting urban air quality using ensemble learning methods. Atmos. Environ. 2013, 80, 426–437. https://doi.org/10.1016/j.atmosenv.2013.08.023
13. Fernando, H.J.; Mammarella, M.C. Grandoni, G.; Fedele, P.; Di Marco, R.; Dimitrova, R.; Hyde, P. Forecasting PM10 in metropolitan areas: Efficacy of neural networks. Environ. Pollut. 2012, 163, 62–67. https://doi.org/10.1016/j.envpol.2011.12.018
14. Talib, A. H., & Zainab, A. (2021). Measurement of some Air Pollutantsin Printing Units and Copy Centers Within Baghdad City. Baghdad Science Journal, 18(1 (Suppl.)), 0687-0687. http://dx.doi.org/10.21123/bsj.2021.18.1(Suppl.).0687
15. Ali, S. M. (2017). A study of Land Zoning using ArcGIS. Al-Khwarizmi Engineering Journal, 13(4), 137-151. https://doi.org/10.22153/kej.2017.06.002
16. Air Quality Control Company. Tehran Air Pollution Forecasting System; MF96/05/01 (U/01); Air Quality Control Company: Tehran, Iran, 2018. (In Persian). https://doi.org/10.3390/ijgi8020099

17. Kelly, F.J.; Fuller, G.W.; Walton, H.A.; Fussell, J.C. Monitoring air pollution: Use of early warning systems for public health. Respirology 2012, 17, 7–19. https://doi.org/10.1111/j.1440-1843.2011.02065.x

18. مريم حسن احمد سليمان. (2016). practical and theoretical study of air pollution of the north gas company in the city of kirkuk. *Journal of Engineering and Sustainable Development*, *20*(3). https://jeasd.uomustansiriyah.edu.iq/index.php/jeasd/article/view/606/480

19. Pope, C.A., III; Dockery, D.W. Health effects of fine particulate air pollution: Lines that connect. J. Air Waste Manag. Assoc. 2006, 56, 709–742. https://doi.org/10.1080/10473289.2006.10464485

20. Wang, P.; Liu, Y.; Qin, Z.; Zhang, G. A novel hybrid forecasting model for PM10 and SO2 daily concentrations. Total Environ. 2015, 505, 1202–1212. https://doi.org/10.1016/j.scitotenv.2014.10.078

21. Venegas, L.E.; Mazzeo, N.A.; Dezzutti, M.C. A simple model for calculating air pollution within street canyons. Atmos. Environ. 2014, 87, 77–86. https://doi.org/10.1016/j.atmosenv.2014.01.005

22. Gopalakrishnan, V. (2021). Hyperlocal air quality prediction using machine learning. Towards data science. https://towardsdatascience.com/hyperlocal-air-quality-prediction-using-machine-learning-ed3a661b9a71

23. Sanjeev, D. (2021). Implementation of machine learning algorithms for analysis and prediction of air quality. *International Journal of Engineering Research & Technology (IJERT)*, *10*(3), 533-538. https://doi.org/10.17577/IJERTV10IS030323

24. Castelli, M., Clemente, F. M., Popovič, A., Silva, S., & Vanneschi, L. (2020). A machine learning approach to predict air quality in California. *Complexity*, *2020*. https://doi.org/10.1155/2020/8049504

25. Kumar, K., & Pande, B. P. (2023). Air pollution prediction with machine learning: a case study of Indian cities. *International Journal of Environmental Science and Technology*, *20*(5), 5333-5348. https://doi.org/10.1007/s13762-022-04241-5

26. Harishkumar, K. S., Yogesh, K. M., & Gad, I. (2020). Forecasting air pollution particulate matter (PM2. 5) using machine learning regression models. *Procedia Computer Science*, *171*, 2057-2066. https://doi.org/10.1016/j.procs.2020.04.221

27. Liang, Y. C., Maimury, Y., Chen, A. H. L., & Juarez, J. R. C. (2020). Machine learning-based prediction of air quality. *applied sciences*, *10*(24), 9151. https://doi.org/10.3390/app10249151

28. Bibri, S. E., & Bibri, S. E. (2018). Data science for urban sustainability: Data mining and data-analytic thinking in the next wave of city analytics. Smart Sustainable Cities of the Future: The Untapped Potential of Big Data Analytics and Context–Aware Computing for Advancing Sustainability,189-246. https://doi.org/10.1007/978-3-319-73981-6_4

29. Cabrera, B. A Geostatistical Method for the Analysis and Prediction of Air Quality Time

Series: Application to the Aburrá Valley Region. Master's Thesis, Technische Universität München (TUM), München, Germany, 2016.

30. Mohammed, Q. H. & Reddy, E. S. (2019, February). Exploring Missing Data using Adaptive LASSO Regression Imputation in Relation to Parkinson's disease. *International Journal of Innovative Technology and Exploring Engineering*, 8(4S), 413-421.

31. Mahdi, G. J., & Salih, O. M. (2022). Variable Selection Using aModified Gibbs Sampler Algorithm with Application on Rock Strength Dataset. *Baghdad Science Journal*, *19*(3), 0551-0551. http://dx.doi.org/10.21123/bsj.2022.19.3.0551

32. Abbas, H. K., Al-Zuky, A. A., & Mahdy, A. H. (2014). Multifocus Images Fusion Based on Homogeneity and Edges Measures. *Baghdad Science Journal*, *11*(2), 660-672. https://doi.org/10.21123/bsj.2014.11.2

33. Jabbar, R. R., & Alkhafaji, A. A. A. (2023). Analysis of Traditional and Fuzzy Quality Control Charts to Improve Short-Run Production in the Manufacturing Industry. *Journal of Engineering*, *29*(6), 159-176. https://doi.org/10.31026/j.eng.2023.06.12

34. Awad, J. H., & Majeed, B. D. (2020). Moving Objects Detection Based on Frequency Domain. *Baghdad Science Journal*, *17*(2). http://dx.doi.org/10.21123/bsj.2020.17.2.0556

35. Yusro, M. M., Ali, R., & Hitam, M. S. (2023). Comparison of Faster R-CNN and YOLOv5 for Overlapping Objects Recognition. Baghdad Science Journal, 20(3), 0893-0893. https://doi.org/10.21123/bsj.2022.7243

36. Zaki, S. M., Jaber, M. M., & Kashmoola, M. A. (2022). Diagnosing COVID-19 Infection in Chest X-Ray Images Using Neural Network. *Baghdad Science Journal*, *19*(6), 1356-1356. https://dx.doi.org/10.21123/bsj.2022.5965