Machine Learning-Based Binary Classification of Air Quality Using Pollutant Concentrations

Raya Basem Mahmood, Sadiq Dawood Hassan, Ameen Mohammed, Haider Sadiq

University of Information and Communications Technology

Institute of Informatics for Postgraduate Studies

Corresponding author: raya.basem.student@uoitc.edu.iq

Abstract

The primary means to protect the health of people and aid in making environmental decisions is air quality monitoring. This study creates an air quality machine learning model that analyses the concentration of pollutants to form binary air quality models. The target variable was formulated by the researchers as the categories of the original AQI_Bucket were converted into two classes: Safe (0) and Unhealthy (1). The preprocessing pipeline addressed the missing data by using numerical data processing, scaling algorithms, and converting categorical data to other forms. The authors developed three new characteristics PM_ratio, NO ratio and Gas Total that enhanced their ability to quantify the association of pollutants with each other. The researchers developed training and testing sets, using an 80:20 data split, to evaluate three machine learning algorithms, namely, Logistic Regression, Random Forest, and XGBoost. The team was able to assess the performance of the models using six measures: Accuracy, Precision, Recall, F1-score, ROC-AUC and confusion matrixes. The results demonstrated that ensemble models provided better performance results than Logistic Regression. The accuracy of Logistic Regression was 0.887 and the AUC of 0.957, whereas the accuracy of Random Forest was 0.918 and its AUC was 0.976. XGBoost produced the best overall performance with an accuracy of 0.919, precision of 0.938, recall of 0.930, F1-score of 0.934 and AUC of 0.975. The study findings are valid in showing that ensemble learning methods are effective in binary air quality classification and simplify environmental monitoring activities.

Keywords

Air Quality Classification, Machine Learning, Logistic Regression, Random Forest, XGBoost, Binary Classification, Environmental Monitoring, Pollutant Prediction

1. INTRODUCTION

Air pollution is a critical societal and environmental issue since it negatively affects human health and revolutionizes the sustainability of the urban environment. In many

cities, problems with air quality have been caused by industrial activities and transportation emissions and increased population numbers. The precision of air quality surveillance in conjunction with predictive services are now essential requirements that safeguard the wellbeing of people and enable environmental decisions making requirements.

Machine learning is an effective technique to analyze air quality due to its ability to show better results in decoding intricate interrelations between pollutant variables than the conventional approaches. The measurement of PM2.5 PM10 NO NO2 NOx NH3 CO SO2 O3 Benzene Toluene and Xylene are all pollutants that need to be measured during air quality assessment. Monitoring systems and early warning applications can achieve better results through the practice of transforming air quality prediction into two basic categories which result in simpler and more useful information.

This paper develops a machine learning framework that utilizes the data on concentration of pollutants to classify air quality based on two different levels. The researchers tested three models that comprised Logistic Regression and Random Forest and XGBoost to categorize air quality conditions into Safe and Unhealthy. The main objective of the research was model identification that would be successful in binary classification using preprocessing and feature engineering and other evaluation approaches.

2. Related Work

The research of air quality prediction and classification involving machine learning techniques has been considered due to the critical roles of these methods in the monitoring of the environment and safeguarding human health. The study proved that the conventional machine learning approaches as well as the sophisticated ensemble techniques might effectively analyze the pollutant data.

Janarthanan et al. [1] used deep learning methods to forecast Air Quality Index (AQI) measurements throughout a metropolitan area. Their study showed that nonlinear models can effectively forecast the patterns of air pollution due to their capability of managing complex environmental data. The results demonstrate the assumption that the data of air quality measurements has nonlinear correlations among different pollutants.

Kothandaraman et al. [2] did a study on intelligent air quality forecasting based on machine learning methods and showed that preprocessing data methods and feature manipulation techniques are key to building accurate predictive systems. Their studies indicated that model development must be done with a certain precaution on both missing data and other environmental variables exhibiting different characteristics.

Gupta et al. [3] compared different machine learning methods of predicting AQI that revealed the ensemble methods yielded better results compared to the linear methods which employed more simple methods. The findings suggest that both boosting-based algorithms and the random forest algorithms are effective when using environmental datasets, which contain compound interactions of features.

On the same note, Aram et al. [4], contrasted various machine learning models in predicting and classifying air quality, and it was found that the multiple evaluation metrics used give a more valid insight into the model performance. Their results proved that accuracy is not enough to evaluate environmental classification models.

The recent research also paid attention to the advanced ensemble learning. Ozupak et al. [5] have described a high level of results with the methods of ensemble based on air quality forecasting, and Liu et al. [6] have shown the effectiveness of stacking-based methods in urban air quality

analysis. All these studies point to the fact that ensemble models still have high predictive power in the application of air quality.

These findings have led to the hypothesis that this study is concerned with binary air quality classification and not with AQI regression and multi-class prediction. The suggested model is used to compare Logistic Regression, Random Forest, and XGBoost in order to determine the most appropriate model to classify air quality into the Unhealthy and Safe categories.

3. Methodology

3.1 Dataset Description

The data in this study was accessed at Kaggle and initially had 29,531 records and 16 attributes which were air quality related. The key variables were City, Date, pollutant concentrations of PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, and Xylene and AQI and AQI_Bucket. The dataset was chosen due to diverse measurements of the pollutants that can be used to classify air quality.

3.2 Target Variable Construction

The initial AQI_Bucket classes were transformed into binary target variable called AQI_binary to render prediction task easier. Good and Satisfactory were coded into Safe (0), whereas Moderate, Poor, Very Poor and Severe were coded into Unhealthy (1). The amount of samples in the dataset was 24,850, with scientists having removed samples with missing target labels.

3.3 Data Preprocessing

Before beginning model training, the researchers performed a number of preprocessing steps. To begin with, any columns that had the potential of leakage of data due to the fact that they contained AQI and AQI_Bucket and Date and AQI_binary were eliminated in the feature set. Numerical values that were missing were imputed with median and the most common category imputed the categorical values. The researchers used StandardScaler to normalize the numerical features and OneHotEncoder to encode the categorical features. The researchers applied a preprocessing

pipeline that enabled them to perform these steps that permitted them to turn data in a consistent way.

3.4 Feature Engineering

To improve representation of pollutant relationships, three additional features were generated:

- **PM_ratio = PM2.5 / (PM10 + 1)**
- **NO_ratio = NO / (NOx + 1)**
- **Gas_Total = NO2 + SO2 + CO**

After feature engineering and encoding, the final feature matrix contained 43 input features.

3.5 Train–Test Split

The data were split into training and testing data in 80:20 with random_state = 42. To maintain the same class distribution in the two subsets, the parameter stratify = y was adopted. The training data set included 19 880 samples and the testing data set included 4970 samples.

3.6 Model Development

The research used three machine learning algorithms that comprise of the Logistic Regression, random Forest and XGBoost. The researchers set the baseline of a linear classifier, which was the Logistic Regression. Random Forest and XGBoost were chosen as the ensemble methods used by the researchers as these algorithms can adequately describe nonlinear relationships and interactions of features in relation to their study of air pollution.

3.7 Hyperparameter Tuning

To improve performance, the XGBoost model was optimized using GridSearchCV with 3-fold cross-validation and accuracy as the scoring metric.

The results of the research team were based on the optimization of the XGBoost using GridSearchCV with a threefold cross-validation and the accuracy as an evaluation measure. The search by the researchers used combinations of several n_estimators and max_depth and

learning_rate. The best parameter settings were learning_rate of 0.05 and max_depth of 6 and 300 n_estimators.

3.8 Performance Evaluation

Accuracy, Precision, Recall, F1-score, and **ROC-AUC** were used to evaluate the models. The researchers applied the confusion matrix to analyse the results of their classification that assisted them to define unhealthy air quality cases.

In the proposed methodology workflow diagram, all the steps of the methodology are presented that involve data preprocessing and feature engineering and model

development and hyperparameter tuning and performance evaluation were illustrated in figure 1.

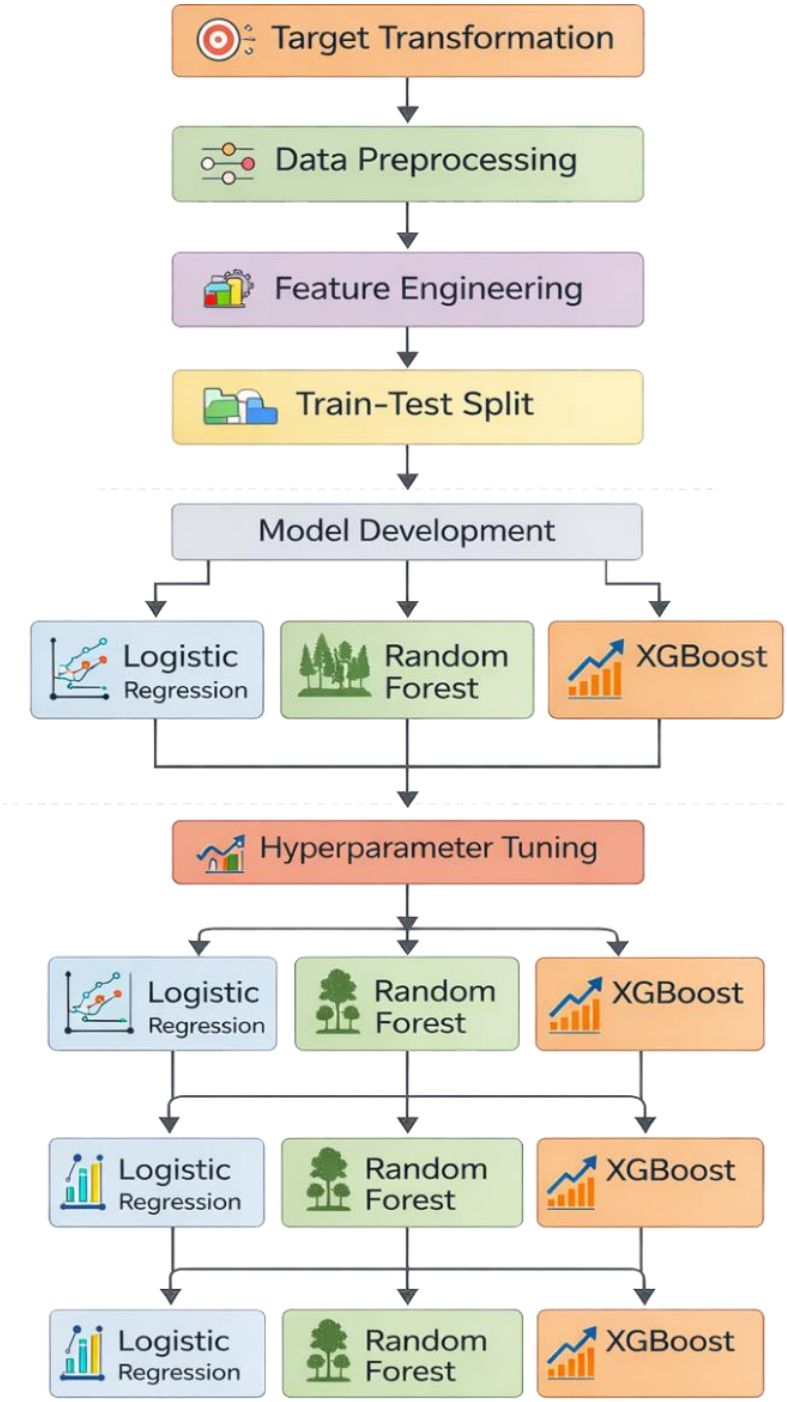

Figure 1. Overall workflow of the proposed methodology for binary air quality classification.

4. Results and Discussion

4.1 Model Performance Comparison

The three machine learning models achieved good performance in binary air quality classification. Table 1 presents the overall comparison results.

Table 1. Performance comparison of the evaluated models.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.887	0.933	0.880	0.905	0.957
Random Forest	0.918	0.932	0.935	0.933	0.976
XGBoost	0.919	0.938	0.930	0.934	0.975

The findings indicate that the performance of Logistic Regression was the lowest amongst the three models though it still gave reasonable results. Conversely, the performance of Random Forest and XGBoost was higher in most of the evaluation measures, which means that ensemble tools are more appropriate in this type of classification.

4.2 Discussion of the Best Models

XGBoost was the most successful model with an overall accuracy of 0.919, precision of 0.938, recall of 0.930, F1-score of 0.934, and AUC of 0.975. This shows that XGBoost was the most balanced in terms of classification.

Random Forest was also performing well with a slightly higher recall (0.935) and AUC (0.976) than XGBoost. This indicates that Random Forest was very effective to identify cases of unhealthy air quality. Nevertheless, XGBoost recorded a marginally higher overall precision-recall balance.

4.3 Confusion Matrix Analysis

The tuned XGBoost model confusion matrix presented the following results:

Table 2: Confusion Matrix Analysis

	Predicted Safe	Predicted Unhealthy
Actual Safe	TN = 1724	FP = 189
Actual Unhealthy	FN = 213	TP = 2844

These findings mean that the model was able to classify the majority of safe and unhealthy cases in the right direction. In particular, the number of correctly identified unhealthy cases was high, which is important for practical monitoring applications. The cases of false negative were also not that many, although it would be useful to decrease them as well since even healthy air that can be called safe can diminish the quality of warning mechanisms. Figure 2 shows the comparison of the Logistic Regression, Random Forest, and XGBoost according to several evaluation metrics. The findings show that ensemble models especially XGBoost perform better and have a higher

accuracy, precision, recall and AUC with better performance when compared to the baseline model.

Figure 2 compares model performance, showing that XGBoost achieves the best overall results.

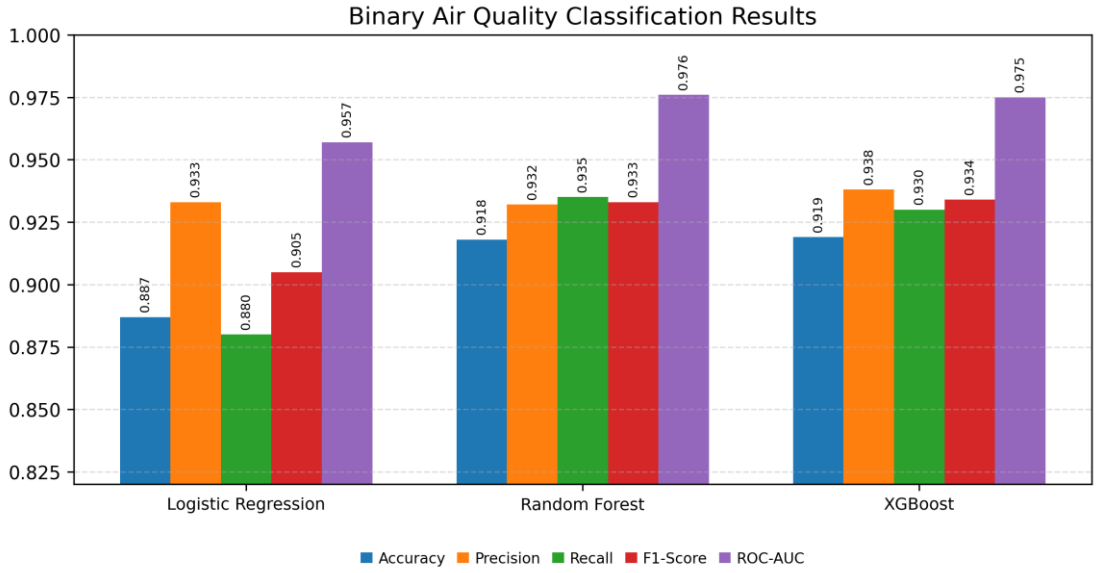

Figure 2. Comparison of model performance across the main evaluation metrics.

4.4 Interpretation of the Results

The fact that Random Forest and XGBoost perform better can be attributed to the fact that they include nonlinear interactions of the pollutant variables. There are many interacting factors that affect air quality and these interactions are usually too complicated to be modeled based on a linear model like the Logistic Regression. Moreover, the engineered features such as PM ratio, NO ratio, and Gas Total must have enhanced the modeling of the pollutant interactions, as well as the performance of the classification.

On the whole, the findings validate the idea that ensemble learning approaches are more useful in binary classification of air quality and that in the present case, the XGBoost model is the most appropriate.

5. Conclusion and Future Work

This paper introduced a machine learning system of binary classification of air quality through pollutants concentration data. The data have been preprocessed, converted into safe and unhealthy categories and were utilized to test Logistic Regression, Random Forest and XGBoost. The findings revealed that ensemble techniques performed better than Logistic Regression, and XGBoost had the highest overall performance.

The results prove that machine learning can be successfully used to classify air quality situations based on pollutant data and can be utilized to assist in the practical monitoring and warning systems. Other advanced models can be tried and more features like

meteorological variables can be added to the work in the future to further enhance the predictive performance.

REFERENCES

[1] R. Janarthanan, P. Partheeban, K. Somasundaram, and P. NavinElamparithi, "A deep learning approach for prediction of air quality index in a metropolitan city," Sustainable Cities and Society, vol. 67, p. 102720, 2021.

[2] D. Kothandaraman, S. K. Lakshmanaprabu, S. Mohanty, and A. Shankar, "Intelligent air quality forecasting using machine learning techniques," Journal of Environmental Management, vol. 265, p. 110555, 2020.

[3] P. Gupta, A. Singh, and R. K. Sharma, "Comparative analysis of machine learning algorithms for air quality prediction," Procedia Computer Science, vol. 167, pp. 209-216, 2020.

[4] F. Aram, A. Garcia, E. Solgi, and S. Mansournia, "Urban air pollution monitoring using machine learning techniques," Sustainable Cities and Society, vol. 60, p. 102199, 2020.

[5] Y. Ozupak, M. Cakmak, and A. Yilmaz, "Air quality prediction using ensemble learning methods," Environmental Monitoring and Assessment, vol. 193, no. 6, pp. 1-12, 2021.

[6] H. Liu, Y. Tian, Y. Li, and L. Zhang, "A stacking ensemble learning approach for air quality prediction," Atmospheric Pollution Research, vol. 11, no. 4, pp. 738-745, 2020.