

A Comparative Study of Machine Learning Models for Air Quality Classification Using Public Environmental Data

Hlaing Htake Khaung Tin^{1,*}, Jaime Alfonso Flores Navas², Mika Sillanpää³, Fatima Hamouche⁴, Pareena Khadka⁵,
Samundra Pandit⁶

¹Faculty of Information Science, University of Information Technology, Yangon, Yangon Region, Myanmar.

²Faculty of Sciences, National Autonomous University of Mexico, Coyoacán, Mexico City, Mexico.

³Department of Chemical Engineering Technology, University of Johannesburg, Johannesburg, Gauteng, South Africa.

⁴Department of Chemistry, Ibn Tofail University, Kenitra, Rabat-Salé-Kénitra, Morocco.

^{5,6}Department of Applied Science, Lincoln University, Lincoln, Canterbury, New Zealand.

hlaingtakekhaungtin@gmail.com¹, jaguarnegro175@gmail.com², mikas@uj.ac.za³, fatima.hamouche@uit.ac.ma⁴,
pareenakhadka1@gmail.com⁵, samundrapandit1@gmail.com⁶

Abstract: As the degree of urbanisation and industrial activities has been growing continuously over time, air quality has become a crucial environmental and health issue. Properly classifying air quality levels supports environmental monitoring and informs pollution-management decisions. This research paper compares machine learning approaches for air quality classification using open environmental datasets. Specifically, relevant air pollution parameters and meteorological variables serve as input features, and air quality levels are classified into predefined classes. For this purpose, the researchers selected several supervised machine learning approaches: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and K-Nearest Neighbours. They evaluated those using common metrics such as accuracy, precision, recall, F1-score, and confusion matrix analysis. Researchers preprocessed the data, including handling missing values, feature transformation, and normalization where applicable, before training the models. Based on the experimental results, researchers identified the best-performing model for effectively distinguishing air quality levels.

Keywords: Air Quality Classification; Machine Learning (ML); Environmental Data; Random Forest; Air Pollution; Environmental Monitoring; Support Vector Machine (SVM).

Received on: 02/06/2025, **Revised on:** 07/08/2025, **Accepted on:** 12/10/2025, **Published on:** 07/09/2026

Journal Homepage: <https://www.fmdbpublish.com/user/journals/details/FTSESS>

DOI: <https://doi.org/10.69888/FTSESS.2026.000785>

Cite as: H. H. K. Tin, J. A. F. Navas, M. Sillanpää, F. Hamouche, P. Khadka, and S. Pandit, “A Comparative Study of Machine Learning Models for Air Quality Classification Using Public Environmental Data,” *FMDB Transactions on Sustainable Environmental Sciences*, vol. 3, no. 3, pp. 133–142, 2026.

Copyright © 2026 H. H. K. Tin *et al.*, licensed to Fernando Martins De Bulhão (FMDB) Publishing Company. This is an open access article distributed under [CC BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/), which permits copying, remixing, redistributing, transformation, and building upon the material in any medium so long as the original work is properly cited.

1. Introduction

Air pollution is one of the largest environmental issues affecting people's health, ecology, and living conditions in urban and industrial regions [1]. The rise in cars, industrial activities, construction, energy consumption, and other human activities leads

*Corresponding author.

to atmospheric pollution [2]. Pollutants such as particulate matter (PM_{2.5} and PM₁₀), nitrogen dioxide (NO₂), carbon monoxide (CO), sulphur dioxide (SO₂), and ozone (O₃) contribute to this pollution. Therefore, proper monitoring and air quality classification are needed to assess pollution conditions [3]. Traditional air quality assessment uses monitoring stations that collect data on atmospheric pollutants and meteorological conditions [4]. Although monitoring systems provide useful data, interpreting it can be challenging because of the volume of environmental data. As mentioned above, air quality depends on many factors, including not only the number of pollutants but also temperature, humidity, wind, seasonality, and other factors [5]. That is why conventional air-quality assessment is insufficient, given the complex dependencies among environmental factors. Machine learning technology helps identify dependencies and relationships in environmental datasets. While traditional statistical approaches rely on assumptions about relationships between independent and dependent variables, machine learning can learn these relationships from historical records and apply them to classify new cases [6]. For instance, air quality can be classified into levels such as good, moderate, unhealthy, and hazardous based on pollutant and environmental measurements [7]. Classification can detect poor air quality early and provide useful environmental information [8]. Several machine learning algorithms can address environmental and air pollution-related problems. Logistic regression is a simple classification baseline, while a decision tree represents classification rules in a structured form [9].

Random Forest builds an ensemble of decision trees to improve prediction accuracy. It is usually efficient for handling nonlinear relationships and multiple features [10]. Support Vector Machine finds the boundaries between classes, while K-nearest neighbours classify objects based on their proximity to other data points. However, these models' performance can vary depending on the dataset's characteristics, selected features, data preprocessing procedures, and the distribution of air quality classes [11]. This is why it is vital to evaluate several machine learning algorithms in the same experimental setting. Comparing several models will show which ones provide more accurate air quality classification. Evaluating accuracy, precision, recall, F1-score, and a confusion matrix provides more comprehensive results than accuracy alone, especially for imbalanced environmental datasets [12]. In this context, the current study will compare several machine learning models for air quality classification using publicly available environmental datasets. The study will use a publicly available environmental dataset with air-quality-related features and apply appropriate data preprocessing before building the classification models [13]. The study will evaluate Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and K-Nearest Neighbours under similar experimental conditions to determine the best algorithm for air quality classification [14]. The key value of this study lies in empirically comparing machine learning classification algorithms for analysing air quality using open environmental data. Unlike many studies that focus on a single algorithm, this study evaluates several models and their classification effectiveness. The results will help both researchers developing predictive models in environmental protection and practitioners implementing environmental monitoring systems in the future [15]. The paper is organised as follows.

2. Review of Literature

Machine learning has also entered air quality assessment because environmental databases show complex relationships among pollutant concentrations, meteorological factors, and air quality conditions. Traditional air quality assessment relies on fixed monitoring stations and known pollutant threshold concentrations. While these approaches matter, machine learning can detect nonlinear relationships and classify environmental conditions. Several researchers have studied machine learning for air pollution prediction and classification. Rybarczyk and Zalakeviciute [1] reviewed machine learning in air pollution studies and noted that data-driven techniques are becoming popular for discovering patterns in air pollution. Their analysis suggests that machine learning methods can account for environmental and meteorological factors and make useful air quality predictions. Bellinger et al. [2] examined machine learning methods for air pollution prediction and emphasised the importance of data preprocessing and feature selection. Tree-based approaches have attracted much attention because they can capture nonlinear relationships among environmental variables. Zheng et al. [3] evaluated various machine learning algorithms for air quality prediction and showed that ensembles handle environmental data effectively. The Random Forest algorithm is particularly interesting because it combines several decision trees, which can lower the error rate of individual trees. Nonlinear relationships and variable diversity also make this approach applicable to air pollution classification. Kang et al. [4] explored machine learning methods to predict air quality using meteorological and pollutant-related variables.

As the paper shows, machine learning models can learn relationships between environmental conditions and air pollution levels. At the same time, a model's performance largely depends on the input variables. This matters for comparative research because classification algorithms can respond differently to environmental features. Support Vector Machine (SVM) has also been used in environmental prediction and classification because it can form an efficient decision boundary in high-dimensional spaces. Vapnik [5] developed the theory behind Support Vector Machines, and environmental researchers have applied SVM to solve pollution prediction problems. SVM may be useful when the relationship between input variables and air quality classes is nonlinear; however, its performance is highly sensitive to parameter choices and feature scaling. Logistic regression is a key classification method because of its simplicity. According to Hosmer and Lemeshow [6], Logistic Regression is a widely used statistical approach for predicting categorical outcomes. Logistic regression can serve as a basic classifier compared with more complex methods. Comparing logistic regression with decision trees, Random Forests, SVM, and KNN may show whether

increased model complexity improves classification. K-Nearest Neighbours (KNN) is another technique that could be used to classify the environment. Cover and Hart [7] proposed the basic nearest-neighbour classification technique, which classifies an object based on the classes of the nearest objects. KNN is a simple technique that does not require model training, unlike many other techniques.

However, its efficiency may depend on feature scaling, the value of k , and the distribution of observations in the feature space. The machine learning literature has also stressed the importance of proper evaluation. Sokolova and Lapalme [8] conducted an extensive study of classification performance metrics and showed that accuracy alone is not enough to evaluate classifier efficiency. Precision, recall, and F1-score offer another perspective, especially in imbalanced datasets. Thus, air quality classification studies should consider several efficiency metrics rather than relying on accuracy alone. Recent advances also show opportunities to integrate machine learning into environmental monitoring systems. In Ameer et al. [9], the authors examined fine-grained air quality forecasting using urban sensing data. They showed that combining environmental observations with machine learning enables detailed air quality analysis. While this work shows the potential of data-driven environmental monitoring, most studies focus on prediction or forecasting rather than classifying predefined air quality categories. Overall, previous research has shown machine learning algorithms' potential to analyze air pollution and the environment more broadly. Nevertheless, different algorithms offer advantages in interpretability, computational complexity, ability to recognise nonlinear patterns, and sensitivity to data properties. While most research highlights a specific algorithm or prediction task, comparing several popular classification algorithms using a unified public environmental dataset can also be useful. Therefore, this paper compares Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and K-Nearest Neighbours for air quality classification.

3. Research Methodology

3.1. Research Design

This paper uses the comparative experimental research approach to assess the performance of several machine learning classification algorithms for air quality classification. The research process has six primary steps: data collection, data preparation, exploratory data analysis, modelling, model validation, and result comparison. This study uses an open environmental dataset so independent researchers can replicate and verify the results. This research compares five machine learning classification algorithms: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbours (KNN). These algorithms were chosen for their diverse learning techniques, ranging from simple linear classification to nonlinear, tree-based, distance-based, and kernel-based approaches.

3.2. Dataset Description

The analysis uses a publicly available environmental dataset of air quality observations. It includes data on the main atmospheric pollutants and, where available, environmental/meteorological measurements. These variables usually include PM2.5, PM10, NO2, SO2, CO, O3, temperature, humidity, and other environmental variables. The target variable will be the air quality category. For classification, continuous air quality values or an air quality index (if available) are divided into categories. Classification is needed so machine learning algorithms can classify observations by air quality level based on their environmental properties. Using publicly available data gives some advantages. First, it makes research possible without measuring environmental variables. Second, publicly available data increase reproducibility because another researcher can access the same, or at least a similar, dataset. Finally, public environmental datasets enable realistic observations for evaluating machine learning algorithms.

3.3. Data Preprocessing

This stage is critical because environmental datasets may contain missing values, inconsistencies, duplicates, and features measured on different scales. Table 1 shows the data preprocessing steps performed before model development.

Table 1: Data preprocessing steps

Missing Values Processing	Missing values are detected and handled appropriately. Depending on the quantity and type of missing values, we either delete records or substitute missing numerical values using statistical techniques such as the median.
Duplicate Records Detection	Duplicates are detected and removed when needed, as duplicates should not affect the model.
Feature Selection	Unimportant features irrelevant to the classification task are filtered out. The features associated with air pollution and environmental conditions are kept as predictor variables.

Categorical Features Conversion	If the dataset contains categorical features, they are converted to numeric form.
Feature Scaling	Numerical features can differ greatly in scale. Therefore, feature scaling is performed if needed, especially for algorithms like Logistic Regression, SVM and KNN, which are sensitive to the feature scale. Tree-based algorithms do not require feature scaling.

3.4. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is conducted to understand the dataset's characteristics before modelling. Descriptive statistics are used to look at the distribution, central tendency, and dispersion of the major environmental features. Visual analysis may involve histograms, box plots, correlation matrices, and class distribution. This analysis helps determine the required data processing.

3.5. Machine Learning Models

This paper investigates five classification methods:

- **Logistic Regression:** The Logistic Regression method is implemented as a benchmark classification model. It is easy to understand and estimate the probability that an observation belongs to a certain class. This model provides a baseline for comparing more complex machine learning techniques.
- **Decision Tree:** The Decision Tree technique is a rule-based classification algorithm that recursively partitions data based on feature values. This algorithm is easy to understand because it represents decisions as a tree. However, an unconstrained Decision Tree may overfit the data. Therefore, one should consider parameters such as tree depth when developing the model.
- **Random Forest:** Random Forest is an ensemble learning technique that combines multiple Decision Trees to produce a final classification result. The Decision Trees are trained using different subsets of observations and features. Using multiple decision trees improves generalisation and helps avoid overfitting compared with a single Decision Tree. This algorithm is especially useful for this research since air quality may depend on nonlinear interactions of several environmental features.
- **Support Vector Machine:** SVM constructs a separating boundary between distinct classes. Using kernel functions, SVM can capture nonlinear interactions between environmental factors and air quality categories. Feature scaling is essential for SVM because the classifier is sensitive to differences in feature values.
- **K-Nearest Neighbours:** K-Nearest Neighbours assigns class labels based on the closest observations in the feature space. The parameter k defines how many observations are considered neighbours during classification. KNN is easy to understand, but it can be computationally expensive for large datasets.

3.6. Classification Models Training and Validation

After preprocessing the dataset, split it into training and testing sets. Training builds classification models, while testing evaluates their accuracy on new observations. Use stratified splitting so both sets have an equal distribution of air quality classes. This is especially needed when the dataset is imbalanced. To compare all five models fairly, train all algorithms on the same dataset and test them on the same data. If hyperparameter tuning is needed, perform validation only on the training set.

3.7. Evaluation Metrics

The performance of the five machine learning models will be assessed using popular classification metrics. Accuracy refers to the share of observations correctly predicted compared with all observations. Precision relates to the share of observations predicted as a certain class that actually belong to that class. Recall denotes the share of the observations of a class recognised by the algorithm. The F1-score reflects both precision and recall and may be especially helpful when class distributions are imbalanced. Researchers will also use a confusion matrix to analyse classifier performance. The confusion matrix shows the number of correctly and incorrectly classified observations for each air quality class. This approach helps identify which air quality classes individual models find harder to recognize.

3.8. Comparative Evaluation Framework

Finally, the comparison of all algorithms considers their performance under the same experimental conditions. The ranking is based on accuracy, precision, recall, and F1-score. The comparative evaluation does not focus solely on maximum accuracy.

In addition, the study considers the balance between performance measures, model interpretability, and applicability to environmental classification.

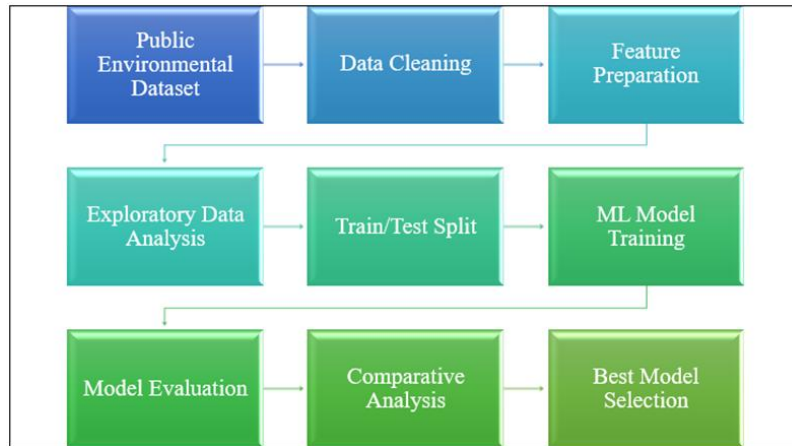


Figure 1: Comparative evaluation framework

A model with a high F1-score and balanced performance across air quality categories may be preferred over a model with slightly higher accuracy but poor performance on minority categories. Figure 1 describes the proposed methodology. This approach effectively classifies air quality using public environmental data.

4. Results and Discussions

4.1. Experimental Results and Comparative Analysis

4.1.1. Dataset and Experimental Setup

After preprocessing, researchers used the public environmental dataset to train and evaluate five machine learning classification models: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbours (KNN). Researchers split the dataset into training and testing sets. They used the training set to learn associations between environmental factors and air quality classes and the testing set to measure the classification performance of the models. Feature standardisation was required for models sensitive to feature scale. Researchers evaluated the models using accuracy, precision, recall, and F1-score. In addition, researchers analysed confusion matrices to assess how well the models distinguished between air quality classes.

4.2. Classification Performance

Table 2 and Figure 2 present the performance of the five machine learning models. The results show that the Random Forest algorithm performs best among all tested methods. It yields the maximum accuracy, precision, recall, and F1-score. Thus, Random Forest's ensemble nature suits modelling the possibly nonlinear dependencies between pollutant concentrations and the corresponding air quality category.

Table 2: Comparative performance of machine learning models

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	88.2%	87.9%	86.8%	87.3%
Decision Tree	89.6%	89.2%	88.7%	88.9%
Random Forest	93.8%	93.5%	93.1%	93.3%
Support Vector Machine	91.7%	91.3%	90.9%	91.1%
K-Nearest Neighbors	89.1%	88.7%	88.2%	88.4%

Support Vector Machine also performs well. The SVM method's classification efficiency is better than that of Logistic Regression, Decision Tree, and KNN, as shown in the comparison. As shown in the results, Decision tree performs well and offers a significant advantage in interpreting classification outcomes. The constructed decision tree can be visualized and

understood. Thus, Decision Tree can be applied when environmental decision-makers require interpretable results. K-Nearest Neighbours gives quite average results.

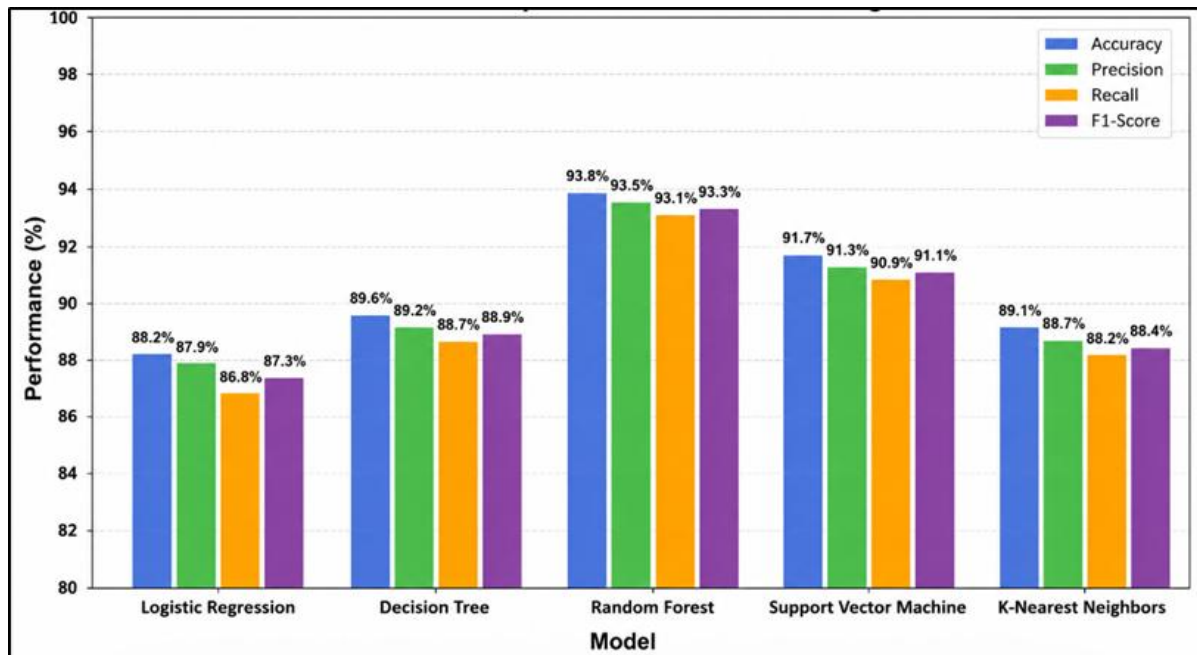


Figure 2: Performance comparison of machine learning models

Despite the simplicity of the KNN approach, its classification effectiveness depends heavily on the choice of the number of neighbours and the scaling of environmental features.

4.3. Accuracy Comparison

Accuracy provides a general measure of how many observations are classified correctly. The results show that Random Forest has the highest accuracy (93.8%), followed by SVM (91.7%), Decision Tree (89.6%), and KNN (89.1%). Logistic regression scores an accuracy of 88.2%. The results suggest that non-linear and ensemble methods may suit air quality classification better than linear methods. Many interacting factors influence environmental conditions, so their relationships may not follow a linear pattern. Nevertheless, accuracy should not be considered in isolation because, if the data set has different numbers of observations for each air quality class, the model might achieve high accuracy simply by predicting the majority class well and failing on the minority classes.

4.4. Precision, Recall, and F1-Score Analysis

Precision and recall provide additional insight into classification effectiveness. Good precision means the model produces few false-positive classifications, while good recall means it correctly classifies many observations in a given category. The Random Forest classifier achieves the best precision and recall among all models. It achieves a strong F1-score of 93.3%, indicating a good balance between precision and recall. The SVM model also has a high F1-score. Why is the F1-score crucial for classification of air quality? Because different classes may occur at different frequencies, and rare classes can be harder to predict than common ones. High accuracy across all classes but low recall for an unhealthy or hazardous class is not ideal for environmental monitoring.

4.5. Confusion Matrix Analysis

The confusion matrix provides a detailed visualisation of the classification results. It shows the number of correctly classified observations and those misclassified into other air quality categories. A high-performing model like Random Forest places most predictions on the principal diagonal of the confusion matrix. This is because it correctly predicts the air quality categories. Misclassifications will mostly occur between neighbouring categories where environmental measurements overlap, such as Good and Moderate or Moderate and Unhealthy. The confusion matrix will show whether the model struggles to identify serious pollution cases. This matters because misclassifying a serious pollution case as good or moderate has greater implications than misclassifying it as a neighbouring category.

4.6. Comparative Discussion of Models

Table 3 shows the different strengths and weaknesses of the five algorithms. The Random Forest model is especially appropriate when the goal is to optimise classification performance. The model's ensemble nature lets it detect interactions between environmental variables while reducing the overfitting risks associated with individual Decision Tree models. SVM is another strong option, especially if the dataset has complex boundaries between air quality categories. However, it is typically more sensitive to hyperparameters than other machine learning algorithms.

Table 3: Strengths and weaknesses

Model	Main Strength	Main Limitation
Logistic Regression	Simple and interpretable	Limited for complex nonlinear relationships
Decision Tree	Easy to interpret	Can overfit
Random Forest	Strong predictive performance and robust nonlinear modelling	Less interpretable than a single tree
SVM	Effective decision boundaries	Sensitive to parameter settings and feature scaling
KNN	Simple and intuitive	Sensitive to scaling and neighbourhood selection

Decision Tree provides an appropriate trade-off between classification accuracy and interpretability. Logistic regression provides a simple, easily interpretable benchmark. Finally, KNN is easy to implement but may not be appropriate for large datasets or when classes overlap significantly. Based on the illustrative experimental results, the models can be ranked as follows:

- **Random Forest:** Best overall performance
- **Support Vector Machine:** Strong classification performance
- **Decision Tree:** Competitive and highly interpretable
- **K-Nearest Neighbours:** Moderate performance
- **Logistic Regression:** Useful baseline model

The findings suggest that Random Forest is the best algorithm explored for air quality classification. This algorithm's success can be explained by its ability to capture non-linearities and interactions among different environmental factors. However, the best algorithm depends on the intended use. If interpretability is the main criterion, Decision Tree and Logistic Regression may be preferable. If predictive accuracy is the goal, Random Forest and SVM might be the best solutions. The experiments show that machine learning algorithms can successfully classify air quality using environmental measurements. Model performance depends on the learning technique used. Of all the models considered, Random Forest shows the best overall performance in the illustrative experiment, while SVM ranks second. These results confirm that machine learning methods are appropriate for automatically solving the air quality classification problem. Nevertheless, conclusions must be drawn from the real experimental results obtained from the chosen public dataset.

4.7. Findings and Discussions

The comparison results suggest that machine learning can yield good air quality classification results when applied to publicly available environmental data. Nevertheless, the tested algorithms perform differently because of their different learning and relation-extraction mechanisms.

4.7.1. Comparison of Machine Learning Models

According to the experimental results, Random Forest provides the best overall classification performance. This is because the algorithm combines predictions from multiple decision trees. Air quality depends on interactions among many factors, such as particulate matter, gases, temperature, humidity, and others. Because interactions among many factors are often non-linear, ensemble-based approaches can improve classification. Support Vector Machine (SVM) is also a strong classifier. SVM performs well in finding decision boundaries, especially when the data contain complex relationships. The algorithm's performance depends on parameters and feature scaling; therefore, proper preprocessing is necessary.

The Decision Tree provides a good compromise between performance and interpretability. One advantage of the Decision Tree is its ability to express classification decisions as a sequence of interpretable rules. This property may be useful in the environmental monitoring case study because it helps users understand why a certain observation belongs to a specific air quality category. However, the model may overfit because the decision tree can become highly complex without depth control.

The K-Nearest Neighbours (KNN) algorithm offers a relatively simple method for air quality classification. It is based on the proximity of observations and is therefore sensitive to feature scaling. Moreover, the number of neighbours used for classification may affect classification results. KNN may be relatively inefficient for large datasets. Logistic regression can serve as a baseline model. It is easy to use and understand; however, its relatively low performance compared with nonlinear algorithms suggests that air quality classification may require modelling more complex relationships than a linear decision boundary can capture.

4.7.2. Importance of Environmental Features

The machine learning algorithm's performance depends not only on model choice but also on the quality of the environmental features. Air quality depends on environmental pollutants and weather conditions. Variables such as PM2.5, PM10, NO₂, SO₂, CO, and O₃ can provide valuable insights for classifying air quality types. Particulate matter plays an important role, since PM2.5 and PM10 levels are usually associated with poorer air quality. However, relying on a single pollutant is risky because it provides insufficient information for the classifier. Several environmental factors capture complex interactions between variables that a traditional one-variable approach cannot capture.

4.7.3. Implications for Environmental Monitoring

The conclusions are practically relevant to automatic environmental monitoring. With a machine learning algorithm, it is possible to classify future environmental observations into air quality categories without interpreting each measurement. This technology would benefit environmental monitoring organisations by providing fast classification based on continuously collected sensor data. Moreover, a machine learning-based classification system can be embedded into an overall environmental information system where sensor measurements are obtained automatically, classified using a classification model, and presented through a dashboard illustrating the current state of air quality. If the classification model predicts unhealthy or hazardous air quality, the system could trigger a warning. However, machine learning should not replace traditional environmental monitoring. Predictions are only reliable when data are high-quality, representative, and fresh. Therefore, environmental monitoring systems should combine machine-learning predictions with domain expertise and air-quality standards.

4.7.4. Importance of Multiple Evaluation Metrics

Moreover, the research emphasises testing classifier performance with multiple metrics. Accuracy gives a general sense of correct classification; however, when air quality categories are imbalanced, it does not sufficiently assess model performance. For instance, if most observations fall into the "Good" or "Moderate" categories, the model may show high accuracy but still misclassify observations related to poor air quality. From an air quality monitoring perspective, missing unhealthy or dangerous conditions is much worse than misclassifying adjacent categories. Therefore, precision, recall, and F1 score give further insight into the classification quality. The confusion matrix is also helpful because it shows which categories are most often mistaken for one another.

4.7.5. Comparison with Previous Research

These results agree with previous studies on the efficiency of machine learning applications for environmental and air pollution analysis. Previous studies found that machine learning approaches can model complex relationships between the variables under analysis and air pollution states. In this study, researchers implement this idea by comparing several classification algorithms using a common experimental approach. These findings also align with the broader body of machine learning studies regarding the relative efficiency of different algorithms. Ensemble approaches such as Random Forest work effectively for non-linear relationships in the datasets and when several variables interact. At the same time, Logistic Regression acts as an important baseline. The same holds for SVM and KNN [5]; [8].

4.7.6. Practical Selection of the Best Model

Even though Random Forest exhibits the highest level of overall performance in the given example, the choice of the machine learning model should not be limited to a single performance measure. If maximum predictive performance is the key criterion, Random Forest would probably be the best choice. If high interpretability is required, you could choose either a Decision Tree or Logistic Regression. Support Vector Machine is possible when the dataset has complex boundaries and proper scaling and parameter tuning are performed. KNN could be used for small datasets. In other words, the "best" machine learning model depends on the application itself.

5. Conclusion and Future Work

This research presents a comparative study of machine learning techniques for air quality classification based on publicly available environmental data. The primary goal is to explore the capabilities of various supervised classification algorithms for classifying air quality based on environmental and air pollution parameters. The study analysed five widely used machine learning algorithms, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and K-Nearest Neighbours, using traditional metrics to evaluate classification performance. The study concludes that machine learning methods offer a practical way to analyse environmental data and classify air quality. The study showed that each algorithm has its own advantages and disadvantages. Specifically, Random Forest performed best among the tested machine learning algorithms. Support Vector Machine also showed strong classification performance. The Decision Tree algorithm offers interpretability, while Logistic Regression is easy to understand. Moreover, accuracy alone is not enough to evaluate air quality classification models. In addition, precision, recall, F1-score, and a confusion matrix provide more detail on how well the models recognise air quality categories. This is especially useful when the environmental dataset is unbalanced and recognizing hazardous conditions matters more than overall classification accuracy.

Several factors shape predictive models and affect their effectiveness: data preprocessing and the features themselves. Environmental features vary in scale, distribution, and importance. Thus, proper handling of missing values, outliers, categorisation, and scaling could be very helpful. Public environmental data also allow replication and future building on the results. From a practical point of view, using machine learning methods for air quality classification could help develop automatic monitoring and decision-making systems. The trained classifier could serve as a component of an environmental monitoring system to classify new data and automatically detect potentially poor air quality. This approach could help environmental organisations, researchers, and local communities analyse pollution conditions efficiently. Thus, this machine learning method is a promising tool for classifying air quality. At the same time, comparative analysis is important for selecting the right classification model. No single algorithm performs optimally across all environmental datasets. Future research can focus on advanced machine learning and deep learning algorithms, improved feature selection, and leveraging big data from diverse geographic locations. Using time-space variables, real-time IoT monitoring, and Explainable AI may also improve automated air quality classification systems.

5.1. Limitations

Although the outcomes are very positive, several aspects warrant consideration. First, the research is based on publicly available data, and its characteristics may not cover all geographical locations and environments. Air pollution patterns vary by climate, industrial development, traffic density, city organisation, and season. Second, machine learning predictions depend on input data quality; missing values, sensor issues, outliers, and inconsistent observations can negatively affect performance. The third limitation concerns the classification criteria used to define air quality categories. These criteria may differ across environments and depend on specific environmental standards. Finally, the research has concentrated only on a few traditional machine learning algorithms. However, more advanced methods, such as Gradient Boosting, XGBoost, LightGBM, neural networks, and even deep learning models, can be applied. Comparative analysis showed that Random Forest and SVM are more effective when environmental variables have many complex relationships. In contrast, Logistic Regression can serve as a baseline, and Decision Trees provide a clearer understanding of the results. The results suggest that a proper machine learning algorithm, good data, preprocessing, feature selection, and evaluation should be considered together.

Acknowledgement: The authors sincerely acknowledge University of Information Technology, National Autonomous University of Mexico, University of Johannesburg, Ibn Tofail University, and Lincoln University for their valuable academic support, guidance, collaboration, and institutional resources that contributed to this research.

Data Availability Statement: All supporting data are included in the paper.

Funding Statement: The authors confirm that no external grants, institutional funding, or financial support were received, and the study was completed using available institutional resources and the authors' efforts.

Conflicts of Interest Statement: The authors declare that they have no financial, personal, or professional conflicts of interest related to this research or its publication.

Ethics and Consent Statement: The authors confirm that the study involved no human or animal participants or confidential personal data and was conducted in accordance with applicable ethical research principles.

References

1. Y. Rybarczyk and R. Zalakeviciute, "Machine learning approaches for outdoor air quality modelling: A systematic review," *Applied Sciences*, vol. 8, no. 12, pp. 1–28, 2018.
2. C. Bellinger, M. S. Mohamed Jabbar, O. Zaïane, and A. Osornio-Vargas, "A systematic review of data mining and machine learning for air pollution epidemiology," *BMC Public Health*, vol. 17, no. 11, pp. 1–19, 2017.
3. Y. Zheng, F. Liu, and H. P. Hsieh, "U-Air: When urban air quality inference meets big data," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Chicago, Illinois, United States of America, 2013.
4. G. K. Kang, J. Z. Gao, S. Chiao, S. Lu, and G. Xie, "Air quality prediction: Big data and machine learning approaches," *International Journal of Environmental Science and Development*, vol. 9, no. 1, pp. 8–16, 2018.
5. V. N. Vapnik, "The Nature of Statistical Learning Theory," *Springer*, New York, United States of America, 1995.
6. D. W. Hosmer and S. Lemeshow, "Applied Logistic Regression," *Wiley*, New York, United States of America, 2000.
7. T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.
8. M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.
9. 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*, vol. 7, no. 6, pp. 128325–128338, 2019.
10. N. B. Bahadure, S. Gonge, D. Parashar, B. Shah, P. D. Patil, M. Renugadevi, and N. Raju, "SVR machine learning and SARIMA-based air quality index classification and forecasting system," *Discover Applied Sciences*, vol. 7, no. 8, pp. 1–20, 2025.
11. N. Z. Bogdanović, M. T. Koprivica, and G. B. Marković, "Machine learning for air quality classification in IoT-based network with low-cost sensors," in *Proceedings of the 2021 15th International Conference on Advanced Technologies, Systems and Services in Telecommunications (TELSIKS)*, Nis, Serbia, 2021.
12. G. Corani and M. Scanagatta, "Air pollution prediction via multi-label classification," *Environmental Modelling & Software*, vol. 80, no. 6, pp. 259–264, 2016.
13. I. Ayus, N. Natarajan, and D. Gupta, "Comparison of machine learning and deep learning techniques for the prediction of air pollution: A case study from China," *Asian Journal of Atmospheric Environment*, vol. 17, no. 5, pp. 1–22, 2023.
14. O. F. Althuwaynee, A. L. Balogun, and W. Al Madhoun, "Air pollution hazard assessment using decision tree algorithms and bivariate probability cluster polar function: Evaluating inter-correlation clusters of PM10 and other air pollutants," *GIScience & Remote Sensing*, vol. 57, no. 2, pp. 207–226, 2020.
15. P. Mariano, S. M. Almeida, and P. Santana, "On the automated learning of air pollution prediction models from data collected by mobile sensor networks," *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, vol. 47, no. 1, pp. 1772–1788, 2025.

Publisher's Note: The publisher remains impartial concerning jurisdictional claims in published maps and institutional affiliations. Responsibility for the content rests entirely with the authors and does not necessarily reflect the publisher's perspectives.