Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
Vol. 25, No. 2, March 2026, pp. 421∼432
ISSN: 2476-9843, accredited by Kemenristekdikti, Decree No: 10/C/C3/DT.05.00/2025
DOI: 10.30812/matrik.v25i2.5121

Performance Comparison of Decision Tree, KNN, and Naive Bayes for Air Quality Classification

Yan Yang Thanri, Juli Iriani, Lili Tanti, Luthfi Zaidi
Universitas Potensi Utama, Medan, Indonesia

Article Info

Article history:

Received January 07, 2026
Revised February 27, 2026
Accepted March 25, 2026

Keywords:

Air Quality Classification;
Decision Tree;
k-Nearest Neighbor;
Naive Bayes;
Stochastic Gradient Descent.

ABSTRACT

Air quality degradation has become a critical environmental and public health issue, necessitating accurate and reliable classification models to support effective monitoring systems. This study aims to conduct a comparative analysis of four machine learning algorithms-Decision Tree, k-Nearest Neighbor (kNN), Naive Bayes, and Stochastic Gradient Descent-for classifying air quality using environmental parameters, including particulate matter $\leq$ 2.5 μm (PM2.5), carbon monoxide, temperature, humidity, nitrogen dioxide (NO_2), and sulfur dioxide (SO_2). The methodology employs supervised learning, where each model is trained and evaluated using classification accuracy, area under the receiver operating characteristic curve, F1-Score, precision, recall, and Matthews Correlation Coefficient, supported by ROC curve and confusion matrix analyses. The results show that the Decision Tree algorithm achieves the best overall performance, attaining a classification accuracy of 93.8% with a balanced precision, recall, and F1-Score, indicating strong and consistent predictive capability. The kNN and Naive Bayes models record the highest AUC values (0.980 and 0.982, respectively), demonstrating excellent class separability, although their accuracy and F1-Score are lower than those of the Decision Tree. In addition, the SGD model, implemented with a modified Huber loss function and L2 regularization, provides interpretable feature-weight analysis, identifying PM2.5 and CO as dominant indicators of the Hazardous air quality class, while temperature and humidity significantly influence the Fair and Good classes. Based on the comprehensive evaluation, the Decision Tree algorithm is recommended as the most reliable model for accurate air quality classification, whereas the SGD model is particularly suitable for feature contribution analysis to enhance interpretability. These findings offer practical insights for selecting appropriate machine learning models in air quality monitoring and decision-support systems.

Corresponding Author:

Yan Yang Thanri, +62812-6230-2925,
Faculty of Engineering and Computer Science,
Universitas Potensi Utama, Medan, Indonesia,
Email: ythanri@gmail.com.

How to Cite:
Y. Y. Thanri, J. I. Iriani, L. T. Tanti, and L. Z. Zaidi, "Performance Comparison of Decision Tree, KNN, and Naive Bayes for Air Quality Classification", *MATRIK: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, Vol. 25, No. 2, pp. 421-432, March, 2026.

1. INTRODUCTION

Air quality is a crucial indicator in determining the environmental and public health conditions of a region [1]. With the rise of human activities, urbanization, and industrialization, air pollution has become a global issue affecting daily life [2]. To understand and mitigate the negative impacts of air pollution, classifying air quality has become an essential aspect of data-driven decision-making [3]. Therefore, accurate and efficient methods are required to classify air quality based on various environmental parameters such as PM2.5, PM10, SO_2, and NO_2 [4–6].

Various studies have developed predictive models and air quality classification using diverse approaches. This study employs machine learning algorithms to analyze air quality parameters such as PM2.5 and NO_2, aiming to enhance accuracy in calculating the air quality index. The results indicate that machine learning-based approaches show promising outcomes in supporting effective air quality management [7]. Meanwhile, the Logistic Regression method has proven effective in capturing air pollution patterns based on historical PM2.5 data [8]. Additionally, a six-year meteorological data-based study demonstrated that PM2.5 classification [9] can be reliably performed by considering weather variables such as humidity and temperature [10]. Furthermore, ensemble techniques such as Boosting and Bagging [11] have shown superiority in improving model stability and generalization compared to Support Vector Machine in air quality classification [12]. With these various approaches, predictive models and classifications continue to evolve to enhance accuracy in air quality analysis and environmental pollution mitigation.

Although numerous methods have been developed, uncertainty remains regarding which algorithm is most effective for air quality classification. This study aims to compare the performance of Decision Tree, K-Nearest Neighbor, and Naive Bayes algorithms in classifying air quality using relevant environmental parameters. Through comprehensive evaluation, this study is expected to provide new insights and useful recommendations for practitioners and policymakers in selecting the best algorithm for air quality monitoring systems.

This article is organized as follows: after the Introduction, the second section explains the research methodology used, including data collection and processing as well as the algorithms tested. The third section presents the results and analysis of model performance evaluation. Subsequently, the fourth section discusses the implications of the findings and recommendations. Finally, the article concludes with conclusions and suggestions for future research.

2. RESEARCH METHOD

The following is the research flow used to compare the performance of the Decision Tree, K-Nearest Neighbor (k-NN), and Naive Bayes algorithms in air quality classification. Each step in the diagram outlines the systematic procedures undertaken, starting from data collection to model performance evaluation. The diagram in Figure 1 provides an overview of the process designed to ensure accurate and relevant research results.

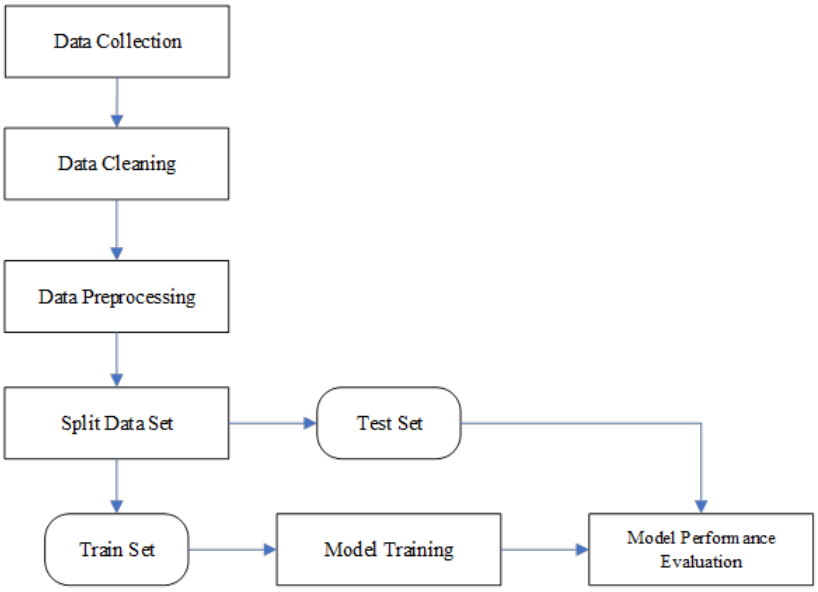


Figure 1. Research workflow for air quality classification

Figure 1 illustrates the research process for the study titled Comparative Analysis of Decision Tree, K-Nearest Neighbor, and Naive Bayes Algorithms in Air Quality Classification. The flow outlines the key stages in developing the air quality classification model. The process begins with Data Collection, where raw data containing essential attributes related to air quality is gathered. The collected data then undergoes a Data Cleaning stage to remove missing, duplicate, or irrelevant entries. Next, the data is further refined in the Data Preprocessing phase, where transformation and standardization techniques are applied to ensure the data is ready for model training. After preprocessing, the dataset is divided into two subsets during the Split Dataset stage: a Train Set and a Test Set. The Train Set is utilized in the Model Training stage, where algorithms such as Decision Tree, K-Nearest Neighbor (k-NN), and Naive Bayes are trained to identify patterns in the data. Concurrently, the Test Set is used to evaluate model performance in the Model Performance Evaluation stage. In the evaluation stage, the models are assessed using metrics such as accuracy, precision, recall, F1-score, and AUC to determine which algorithm performs best for air quality classification. This structured process ensures that the developed model is reliable and capable of making accurate predictions, making it suitable for real-world applications.

2.1. Data collection

Data collection for developing the air quality classification model based on the Naive Bayes algorithm was carried out with a total of 5,000 samples. These data include various air quality categories as target labels, namely Moderate, Good, Fair, and Hazardous. Each sample contains information about pollutant concentrations (NO_2, SO_2, CO), proximity to industrial areas, and population density in specific regions.

The data distribution across the air quality categories was designed to ensure a balanced representation or to reflect patterns observed in real-world conditions. With a sufficient and diverse dataset, the model is expected to effectively learn patterns from each category, resulting in high accuracy and reliability in air quality predictions. The data then underwent further processing through preprocessing steps such as normalization and feature selection to ensure optimal results in model development.

2.2. Data Preprocessing

The data preprocessing stage for developing the air quality classification model based on the Naive Bayes algorithm begins with collecting raw data containing information on concentrations of NO_2, SO_2, CO, proximity to industrial areas, population density, and air quality labels. This dataset consists of 5,000 samples classified into air quality categories such as Moderate, Good, Fair, and Hazardous. After data collection, data cleaning is performed to ensure there are no missing values, duplicates, or anomalies. Next, numerical data, such as pollutant concentrations and proximity to industrial areas, are standardized using the standardization method as described in Equation 1. This process transforms feature values into a scale with a mean of zero and a standard deviation of one, ensuring all features are on the same scale. Standardization aims to prevent any single feature from dominating the model and to improve the stability of the analysis [13].

After standardization, the data is split into two parts: training data and testing data. Commonly applied ratios such as 80:20 or 70:30 are used for this split. This step ensures that the data is ready for model training and evaluation, enabling accurate and reliable air quality predictions. Here, x represents the original value. The symbol μ denotes the feature mean. The symbol σ stands for the standard deviation [14].

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

2.3. Split Dataset

The dataset is divided into two parts: training data and testing data. Typically, 70–80% of the data is used for training the model, while the remaining 20–30% is used to test the model's performance. This division is done randomly so that both subsets represent the data comprehensively, allowing the model to be well-trained and accurately evaluated.

2.4. Model Trainning

At the Model Training stage, the training process is carried out to develop classification algorithms capable of predicting air quality based on the processed data. In this study, three classification algorithms are used: Decision Tree, K-Nearest Neighbor (k-NN), and Naive Bayes. The selection of these algorithms is based on their respective strengths in handling data with different

characteristics, such as complex data, neighbor-based data, and probabilistic data. By using these three algorithms, the study aims to compare the performance of each model in classifying air quality to determine the most effective and reliable method.

2.5. Decision Trees (DTs)

Decision Trees are classification models that operate based on "if-then-or" rules, with nodes representing the main classes, branches as attributes, and leaves as the final outcomes. The data is split into two subsets: training (80%) and testing (20%). The model uses the training data to build hypotheses, validates them on validation data, and tests accuracy on the testing data.

To prevent overfitting, the tree can be pruned through pre-pruning (stopping growth before becoming too deep) or post-pruning (cutting branches after overfitting occurs). This study uses Binary Decision Trees (BDTs) because they are simple and computationally efficient. The evaluation process is carried out using cross-validation, where the data is divided into several folds to alternately train and test the model, resulting in the average performance across all iterations [15]. The formulas used in Decision Trees involve selecting the best attribute to split the data based on certain measures using Entropy in equation (2) and Information Gain in equation (3) [16].

$$Entropy(S) = \sum_{i=1}^{n} -pi * log_2 pi \tag{2}$$

Where, S is the set of cases that serves as the object of analysis in this study. The feature used in dividing the set is denoted by A. Furthermore, the set S is divided into n partitions, where each partition S_i has a proportion p_i relative to the entire set S [16].

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{\S_i}{S} * Entropy(S_i) \tag{3}$$

Where, S is the set of cases under study. A represents an attribute used to divide the set into partitions. The attribute A is divided into n partitions, where $|S_i|$ denotes the number of cases in the i-th partition and $|S|$ represents the total number of cases in S [16].

2.6. Naïve Bayes (NB)

The Bayesian learning approach describes the learning process based on reasoning using conditional probabilities. The Naïve Bayes (NB) classification model is a probabilistic method that determines a set of probabilities by analyzing the frequency and distribution of values in the dataset, assuming that each feature is independent [17]. By utilizing Bayesian theory, NB is used to estimate the likelihood of an event occurring based on other events that have previously occurred. This method ensures that each feature contributes independently and objectively to the final decision [18].

The conditional probability of event X occurring given that event Y has occurred is formulated in Equation 4 [19]. X and Y are deemed independent when the condition in equation (6) is satisfied [19]. X and Y are deemed to be independent when the condition in Equation 7 is satisfied [19].

$$P(X|Y) = \frac{P(X \wedge) Y}{P(Y)} \tag{4}$$

$$P(X|Y) = \frac{P(X \wedge) Y}{P(Y)} = \frac{P(X \wedge Y) \times P(X)}{P(Y) \times P(X)} = \frac{P(X \wedge Y)}{P(Y)} \times \frac{P(X)}{P(Y)} = P(Y|X) \times \frac{P(X)}{P(Y)} \tag{5}$$

$$P(X \wedge Y) = P(X|Y)P(Y) = P(Y|X)P(X) \tag{6}$$

$$P(X \wedge Y) = P(X) \times P(Y) \tag{7}$$

This implies that there is no statistical connection between X and Y, and having knowledge of one does not assist in predicting the value of the other. It is important to note that if X and Y are independent, this relationship is expressed in Equation 8 [19]. Conditional independence explores whether a statistical relationship exists between two variables when a third variable is held at a specific value. X is considered conditionally independent of Y given Z if the condition in Equation 9 is satisfied [19]. where x_i, y_j, and z_k are possible values of X, Y, and Z, respectively, as shown in Equation 8 [19].

$$P(X|Y) = P(X) \tag{8}$$

$$\forall x_i, y_j, z_k (P(X = x_i | Y = y_j \wedge Z = z_k) = P(X = x_i | Z = z_k)) \tag{9}$$

$$P(X|Y \wedge Z) = P(X|Z) \tag{10}$$

2.7. K-Nearest Neighbor (K-NN)

K-nearest neighbor (k-NN) is a simple supervised learning algorithm that classifies new data based on similarity to existing data [17]. Used for both regression and classification, k-NN is called a "lazy learner" because it stores the training data and only performs classification when new data appears, selecting the most similar category [20].

Euclidean Distance:

In an n-dimensional space, the Euclidean distance between two points P = $(p_1, p_2, \ldots, p_n)$ and Q = $(q_1, q_2, \ldots, q_n)$ is calculated using the formula in equations (11) and (12) [21].

$$d(P, Q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{11}$$

$$d(P, Q) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{12}$$

2.8. Model Performance Evaluation

Model performance evaluation is crucial to determine how effectively machine learning algorithms classify air quality. This study presents a comparative analysis of three popular classification algorithms: Decision Tree, K-Nearest Neighbor (KNN), and Naive Bayes. Each model is evaluated based on key metrics such as accuracy, precision, recall, F1 Score, and AUC (Area Under the Curve) to measure the model's ability to accurately classify various levels of air quality [22]. The F1 Score is used to balance precision and recall, providing a more comprehensive overview especially for imbalanced data [23]. Meanwhile, AUC measures the model's capability to distinguish between positive and negative classes by considering different decision thresholds. By analyzing the strengths and weaknesses of each model in handling the dataset, this comparison aims to identify the most suitable algorithm for reliable air quality prediction. The results of this study offer important insights into the performance of these models under various conditions, which can serve as guidance for future environmental monitoring system implementations.

3. RESULT AND ANALYSIS

3.1. Data Analysis

This section presents an analysis of the characteristics of the data used in the study. The analysis includes statistical descriptions of various environmental and demographic features that play a role in air quality classification. Table 1 provides a summary of the descriptive statistics for each variable in the dataset, which will serve as the basis for subsequent modeling and evaluation processes.

Table 1. Descriptive Statistics

Feature Name	Mean	Mode	Median	Dispersion	Min	Max	Missing Data
Temperature (^oC)	30.029	26.8	29.0	0.224	13.4	58.6	0 (0%)
Humidity (%)	70.056	73.0	69.8	0.226	36.0	128.1	0 (0%)
PM2.5 Concentration (μ/m^3)	20.142	1.5	12.0	1.219	0.0	295.0	0 (0%)
PM10 Concentration (μ/m^3)	30.218	8.1	21.7	0.905	-0.2	315.8	0 (0%)
NO2 Concentration (ppb)	26.412	24.2	25.3	0.337	7.4	64.9	0 (0%)
SO2 Concentration (ppb)	10.015	5.7	8.0	0.674	-6.2	44.9	0 (0%)
CO Concentration (ppm)	15.004	0.98	1.41	0.3639	0.65	3.72	0 (0%)
Distance to Industrial Area (km)	8.425	5.1	7.9	0.429	2.5	25.8	0 (0%)
Population Density (people/km^2)	497.42	494	494	0.31	188	957	0 (0%)

The descriptive statistics of the dataset in Table 1 highlight important characteristics of the environmental and demographic features related to air quality. Although there is considerable variation in some features, such as temperature ranging from 13.4^{o}C to 58.6^{o}C and humidity between 36% and 128%, these values are still considered valid within the context of dynamic environmental measurements and the variability of data collection locations. Furthermore, there are no missing data or extreme outliers that are highly suspicious, so extensive data cleaning is not significantly required. However, given the differing value ranges and varying feature scales—such as pollutant concentrations, distance to industrial areas, and population density—normalization or standardization is necessary to optimize the classification model's performance. This normalization or standardization aims to align the scales of the features so that algorithms sensitive to distance and scale, like K-Nearest Neighbor, can produce more accurate and stable predictions.

3.2. Data Preprocessing

The data preprocessing stage is carried out to ensure that the dataset is ready for modeling and analysis. Considering that the features in this dataset have varying scales and value ranges, the standardization step becomes crucial. Standardization is performed to bring all features onto the same scale so that algorithms sensitive to scale differences, such as K-Nearest Neighbor and Naive Bayes, can operate optimally using equation 1. This process aims to improve the accuracy and consistency of classification results in the applied models, as illustrated in Figure 2.

Dataset After Normalization

	Air Quality	Temperature (°C)	Humidity (%)	Concentration (µg/	5 Concentration (µg/m	oncentration of NO2 (ppb	ncentration of SO2 (p	centration of CO (p	ity to Industrial Area	ulation Density (people/k
1	Fair	-0.0341	-0.6907	-0.6086	-0.4505	-0.8446	-0.1207	0.4023	-0.5887	-1.1682
2	Fair	-0.2573	0.3495	-0.7267	-0.6589	0.4933	-0.0466	0.2558	-0.6717	0.7436
3	Fair	-1.0311	0.2928	0.2671	0.1310	-0.2262	0.3830	0.2375	-0.8933	0.7960
4	Good	-0.4359	-1.9516	-0.5719	-0.8746	-1.4517	-0.6985	-0.6417	0.7408	0.3508
5	Good	-0.5252	0.0406	-0.5393	-0.5199	-0.5073	-0.6541	-0.8981	1.1839	-1.2729
6	Hazardous	1.3945	1.6734	-0.2257	0.1931	1.8537	1.1682	0.5855	-1.4749	1.1561
7	Moderate	1.7368	0.7845	-0.7511	-0.5272	0.5271	0.3978	0.5488	-1.0595	1.5554
8	Fair	0.1445	-0.6592	-0.6167	-0.4907	-0.2487	0.5312	-0.2204	-0.5887	-0.3563
9	Moderate	-0.0936	1.4969	-0.4009	-0.2749	2.1011	0.2645	0.9701	-0.8379	-0.0748
10	Moderate	0.4719	0.6584	-0.3683	-0.2128	0.6282	0.7830	0.3474	-0.9764	0.2460

Figure 2. Dataset after normalization

Based on Figure 2 showing the results of the standard normalization above, the dataset has been transformed to a uniform scale with a mean of 0 and a standard deviation of 1 for each feature. This normalization ensures that each variable contributes equally in the analysis and predictive models. Values such as temperature, humidity, and pollutant concentrations (PM2.5, PM10, NO2, SO2, and CO) exhibit a normal distribution with some variability that reflects the dynamics of the original data. Features like proximity to industrial areas and population density also show significant value spread after normalization, reflecting important differences in demographic and environmental characteristics across the dataset. Air quality has been classified into categories such as "Good," "Fair," "Moderate," and "Hazardous," representing different environmental conditions based on these measurement results.

3.3. Model Training

Model training is a crucial stage in the development of an air quality classification system, where machine learning algorithms are trained using historical data to recognize patterns and characteristics of each air quality class. In this process, the model learns from the available features to make accurate predictions on new data. This training phase forms the foundation for the overall performance of the model in classifying air conditions, so it must be conducted with appropriate methods and parameters to achieve optimal results. The training process in this study uses three popular algorithms—Decision Tree, k-Nearest Neighbor (k-NN), and Naive Bayes—each with its own advantages in handling classification data, as shown in Figure 3.

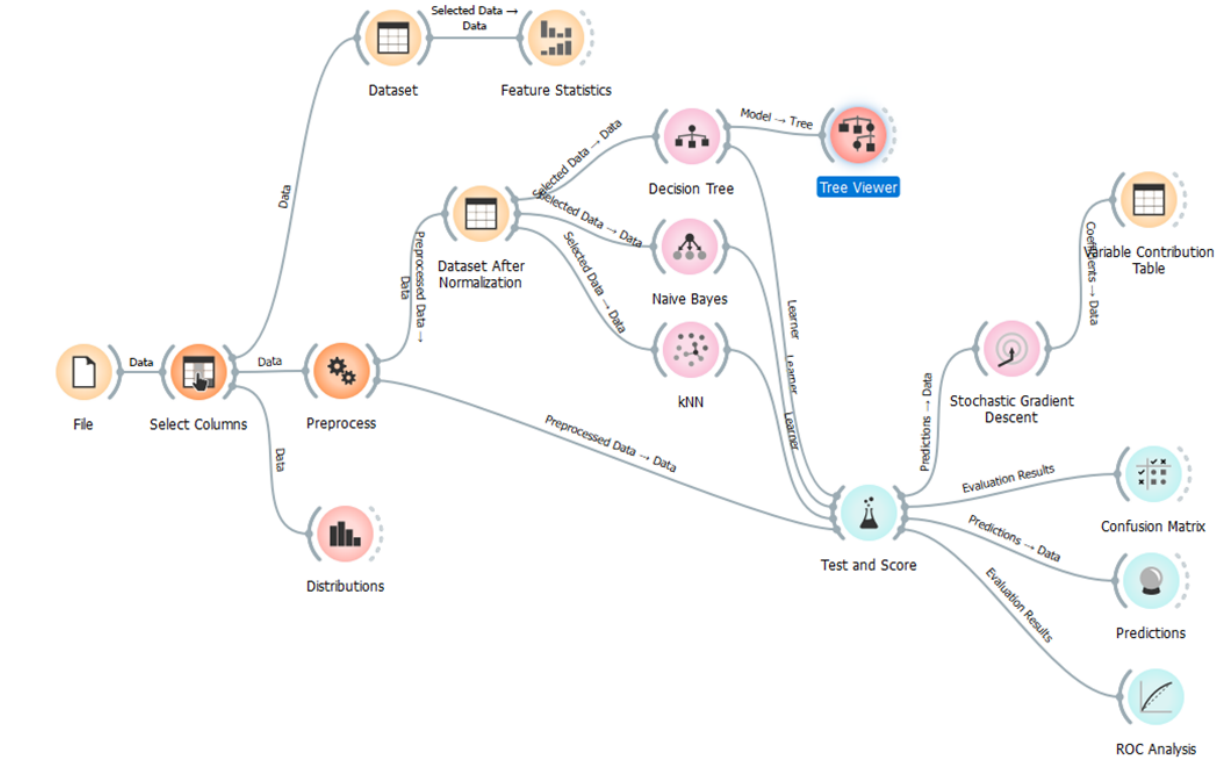


Figure 3. Workflow of comparative analysis of decision tree, k-nearest neighbor, and naive bayes algorithms for air quality classification

Figure 3 shows the workflow for performing a comparative analysis of the Decision Tree, K-Nearest Neighbor (k-NN), and Naive Bayes algorithms in air quality classification using Orange Data Mining software. The process begins by importing the dataset through the File Input widget, where data is selected and filtered using Select Columns to ensure that only relevant attributes are used. Next, the data is processed through the Preprocess widget, which includes normalization and other steps to ensure consistent data quality. Initial analysis is performed using the Distributions and Feature Statistics widgets to understand the data distribution and the contribution of features within the dataset.

Once the data is prepared, the three main algorithms are applied. Decision Tree is used to build an easy-to-interpret rule-based model. Naive Bayes, a probabilistic classifier, is applied for category-based analysis, while k-Nearest Neighbor (k-NN) classifies data based on proximity to the nearest neighbors. As a comparison, the Stochastic Gradient Descent (SGD) algorithm is also used. Model evaluation is carried out through the Test and Score widget, which calculates performance metrics such as accuracy, precision, and recall. Additionally, the Confusion Matrix widget provides detailed classification results, while ROC Analysis visualizes model performance in terms of sensitivity and specificity.

For further interpretation, the Tree Viewer widget is used to visualize the Decision Tree model, and the Variable Contribution Table highlights the most significant features in the prediction. This workflow is designed to provide an in-depth analysis of the strengths and weaknesses of each algorithm in air quality classification, enabling users to select the most appropriate model based on their analytical needs.

3.4. Model Performance Evaluation

Model performance evaluation is conducted by measuring accuracy, AUC, precision, recall, F1-score, and MCC to assess the capability of each algorithm in classifying air quality. These evaluation results help determine the most effective and suitable model for environmental monitoring applications.

Table 2. Classifiers Performance Assessment

Model	CA	AUC	F1-Score	Precision	Recall	MCC
Decision Tree	0.938	0.958	0.938	0.938	0.938	0.911
kNN	0.929	0.980	0.927	0.929	0.929	0.899
Naïve Bayes	0.888	0.982	0.887	0.886	0.888	0.840

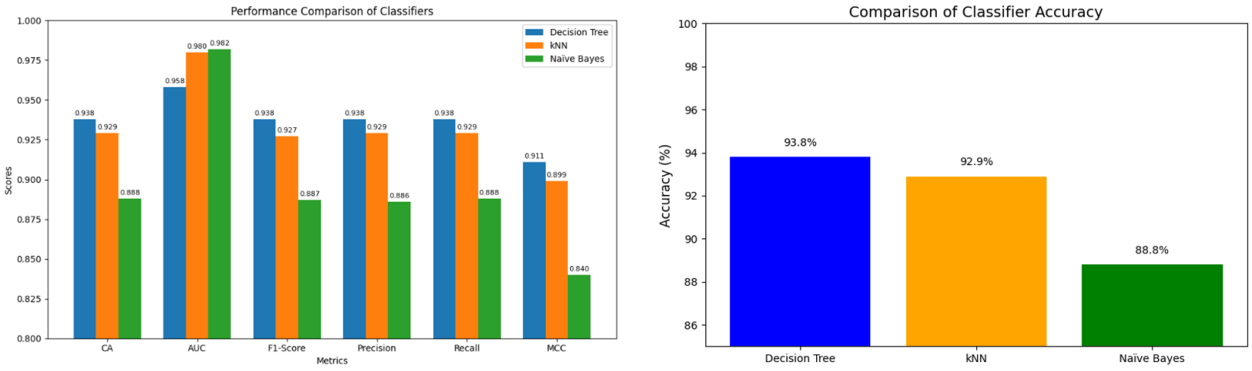


Figure 4. Classifiers overall performance assessment

Table 2 and Figure 4 present the performance evaluation results of three classification models: Decision Tree, k-Nearest Neighbor (kNN), and Naïve Bayes, based on several key metrics such as Classification Accuracy (CA), Area Under the Curve (AUC), F1-Score, Precision, Recall, and Matthews Correlation Coefficient (MCC). The table shows that the Decision Tree achieves a high classification accuracy of 93.8% with an AUC of 0.958, indicating a strong ability to distinguish air quality classes. The F1-Score, Precision, and Recall values for the Decision Tree are all at 0.938, reflecting a good balance between precision and sensitivity of the model. kNN demonstrates competitive performance with an accuracy of 92.9% and the highest AUC among the three models at 0.980, indicating excellent predictive capability. However, the kNN's F1-Score is slightly lower at 0.927. Naïve Bayes has the lowest accuracy among the three, at 88.8%, but still maintains a high AUC of 0.982, showing that the model remains effective at separating classes despite the lower overall accuracy. Overall, Decision Tree and kNN show superior performance, while Naïve Bayes offers a probabilistic approach with a fairly good MCC value of 0.840, indicating a positive correlation between predictions and actual classes.

Variable Contribution Table

	name	Fair	Good	Hazardous	Moderate
1	intercept	-16.6898	-387.932	-1286.21	-65.3268
2	Temperature (°C)	7.64802	-127.004	120.512	-20.0631
3	Humidity (%)	40.2752	-18.091	114.872	-8.4505
4	PM2.5 Concentration (µg/m³) (1)	37.3452	101.014	-115.617	-27.2109
5	PM2.5 Concentration (µg/m³) (2)	-10.596	-129.496	225.953	38.9413
6	Concentration of NO2 (ppb)	-29.7369	-95.5928	177.551	-2.60165
7	Concentration of SO2 (ppb)	-39.2441	-56.6136	92.5121	0.93474
8	Concentration of CO (ppm)	-24.9745	-299.414	276.115	-28.1166
9	Proximity to Industrial Areas (km)	-34.7097	119.37	-183.835	6.6598
10	Population Density (people/km²)	-1.54571	-69.3689	155.193	-6.11914

Figure 5. Variable contributions of the stochastic gradient descent Model

Figure 5 shows the variable contribution table from the Stochastic Gradient Descent (SGD) model classifying air quality into the classes Fair, Good, Hazardous, and Moderate. The coefficients in the table reflect the influence of each feature on the respective classes; PM2.5 and CO have strong positive contributions to the Hazardous class, while temperature and humidity affect the Fair and Good classes. Negative coefficients indicate an inverse relationship. The model uses the Modified Huber loss function with Ridge (L2) regularization at $\alpha = 0.00001$. Training was performed over 1000 iterations with a tolerance of 0.001. After training, evaluation was conducted using the Confusion Matrix (Figure 7), ROC Analysis (Figure 6), and Variable Contribution Table (Figure 5), forming a structured and comprehensive pipeline for air quality prediction.

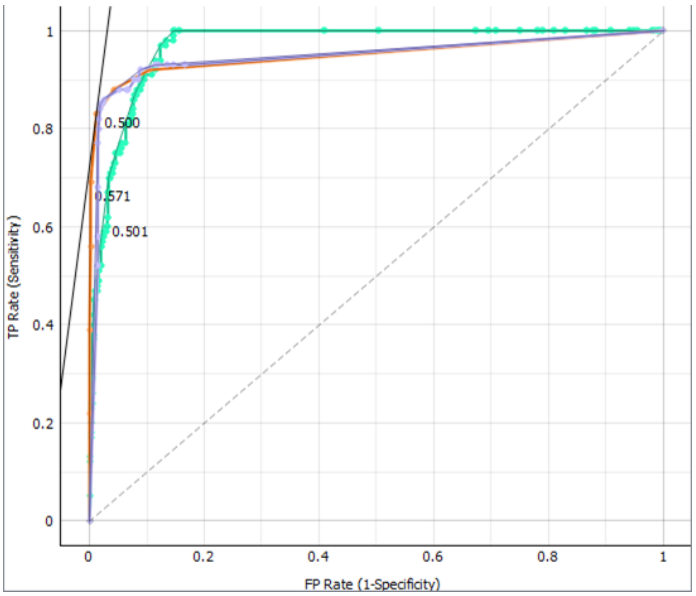
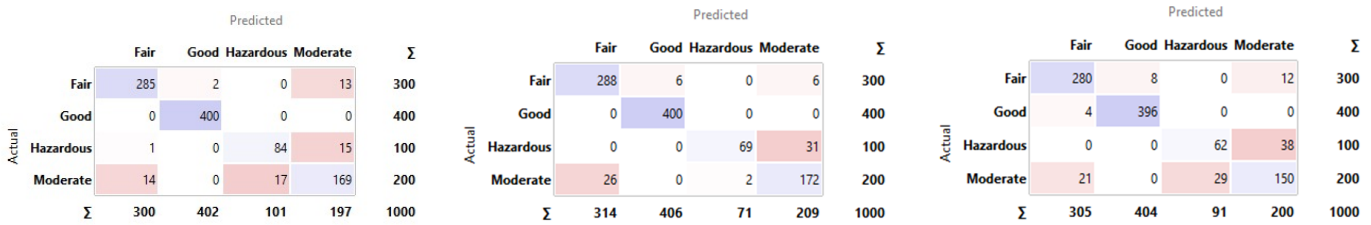


Figure 6. Classifiers ROC curves

Figure 5, the previously shown ROC curve, provides a visualization of the classification performance of three algorithms Decision Tree, k-Nearest Neighbor (kNN), and Naïve Bayes-in the context of air quality classification. The ROC curve illustrates the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR) across various classification thresholds. All three curves in the figure are well above the diagonal reference line (random guess), indicating that all models perform well. However, when compared with the data in the model evaluation table, there are notable differences between the shapes of the ROC curves and other metrics.

Although Naïve Bayes shows the highest and smoothest ROC curve (evidenced by the highest AUC value of 0.982), its overall performance is lower compared to the other two models based on key metrics such as accuracy (CA = 0.888) and F1-Score (0.887). This indicates that while Naïve Bayes excels at distinguishing classes generally (high AUC), it makes more errors in actual classification compared to Decision Tree and kNN. Meanwhile, kNN also has a high AUC (0.980), demonstrating strong class separation ability. However, in terms of accuracy and F1-Score, this model is slightly below Decision Tree (CA = 0.929, F1 = 0.927). The relatively steep initial slope of the kNN ROC curve supports these results. As for Decision Tree, despite having a lower AUC (0.958), it overall shows the best performance across almost all metrics: accuracy, precision, recall, F1-Score, and MCC. This means the Decision Tree consistently provides correct predictions across various thresholds, even though it does not have as high an AUC as the other two models. Thus, ROC curves and AUC values offer a comprehensive view of class separation ability but need to be combined with other metrics to assess classification consistency and accuracy. Based on the combined information from the curves and the table, the Decision Tree can be concluded as the best overall model for this air quality classification task.

Figure 7 shows the confusion matrix for three classification algorithms: Decision Tree, k-Nearest Neighbor (kNN), and Naïve Bayes in classifying air quality into four classes: Fair, Good, Hazardous, and Moderate. The Decision Tree demonstrates high accuracy with predominantly correct predictions, especially for the Good and Fair classes. The kNN model also shows good performance with a high number of correct predictions, although it makes slightly more errors in the Moderate class. Meanwhile, Naïve Bayes exhibits more classification errors, particularly between the Moderate and Hazardous classes, indicating less stable performance compared to the other two models.



(a) Decision Tree Confusion Matrix (b) kNN Confusion Matrix (c) Naïve Bayes Confusion Matrix

Figure 7. Classifiers performance confusion matrices

4. CONCLUSION

This study successfully conducted a comprehensive comparison between the Decision Tree, k-Nearest Neighbor (kNN), and Naïve Bayes algorithms for air quality classification based on environmental and demographic data. The evaluation results show that the Decision Tree achieved the highest classification accuracy of 93.8%, indicating that this model can predict air quality categories with the greatest precision compared to the other two algorithms. Meanwhile, kNN excelled in terms of Area Under the Curve (AUC) with a value of 0.980, demonstrating its excellent ability to distinguish air quality classes overall. Although Naïve Bayes had the lowest classification accuracy at 88.8%, it still showed strong classification performance based on an AUC of 0.982, indicating its effectiveness in separating classes despite having more frequent misclassifications compared to the other models. These findings highlight that each algorithm has its own strengths and weaknesses that should be considered according to the objectives and application context. For further development, integrating additional features such as meteorological data (e.g., wind speed, rainfall, and air pressure) is expected to improve the accuracy and robustness of the models. Additionally, exploring ensemble learning methods—which combine multiple algorithms to enhance prediction stability and performance can be a strategic next step. With these improvements, the air quality classification system can be developed into a more advanced and reliable tool, supporting the implementation of real-time air quality monitoring systems with high scalability. This is crucial to support effective pollution control policies and public health protection, especially in urban areas that are vulnerable to rapid and dynamic changes in air quality.

5. ACKNOWLEDGEMENTS

The author would like to express sincere gratitude to Universitas Potensi Utama for providing the facilities and academic guidance throughout this research. Appreciation is also extended to colleagues, academic advisors, and all others who offered assistance and support in the completion of this study.

6. DECLARATIONS

AUTHOR CONTIBUTION

All authors contributed significantly to the conception, design, data collection, analysis, and writing of this research. Yan Yang Thanri led the development of the concept and methodology. Juli Iriani handled data analysis and interpretation. Lili Tanti and Luthfi Zaidi contributed to drafting the initial manuscript and revising it. All authors have read and approved the final version of the manuscript.

FUNDING STATEMENT

This research was supported and funded by Universitas Potensi Utama through the Regular Research Competition scheme organized by LPPM. The authors express their gratitude for this support, which has contributed to the successful completion of this study.

COMPETING INTEREST

The authors declare no conflict of interest related to this research, its results, or its publication.

REFERENCES

[1] N. S. Gupta, Y. Mohta, K. Heda, R. Armaan, B. Valarmathi, and G. Arulkumaran, "Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis," *Journal of Environmental and Public Health*, vol. 2023, pp. 1–26, Jan. 2023, https://doi.org/10.1155/2023/4916267.

[2] E. X. Neo, K. Hasikin, K. W. Lai, M. I. Mokhtar, M. M. Azizan, H. F. Hizaddin, S. A. Razak, and Yanto, "Artificial intelligence-assisted air quality monitoring for smart city management," *PeerJ Computer Science*, vol. 9, p. e1306, May 2023, https://doi.org/10.7717/peerj-cs.1306.

[3] S. Al-Eidi, F. Amsaad, O. Darwish, Y. Tashtoush, A. Alqahtani, and N. Niveshitha, "Comparative Analysis Study for Air Quality Prediction in Smart Cities Using Regression Techniques," *IEEE Access*, vol. 11, pp. 115 140–115 149, 2023, https://doi.org/10.1109/ACCESS.2023.3323447.

[4] A. Pant, S. Sharma, M. Bansal, and M. Narang, "Comparative Analysis of Supervised Machine Learning Techniques for AQI Prediction," in *2022 International Conference on Advanced Computing Technologies and Applications (ICACTA)*. Coimbatore, India: IEEE, march(04-05) 2022, pp. 1–4, https://doi.org/10.1109/ICACTA54488.2022.9753636.

[5] S. K. Sunori, P. B. Negi, and P. Juneja, "Estimation of Air Quality Index using AI and ML Techniques," in *2023 International Conference on Sustainable Communication Networks and Application (ICSCNA)*. Theni, India: IEEE, November (15-17) 2023, pp. 1078–1082, https://doi.org/10.1109/ICSCNA58489.2023.10370690.

[6] V. Behal and R. Singh, "Personalised healthcare model for monitoring and prediction of airpollution: Machine learning approach," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 33, no. 3, pp. 425–449, May 2021, https://doi.org/10.1080/0952813X.2020.1744197.

[7] I. Dawar, M. Singal, V. Singh, S. Lamba, and S. Jain, "Predicting air quality index using machine learning: A case study of the Himalayan city of Dehradun," *Natural Hazards*, vol. 121, no. 5, pp. 5821–5847, Mar. 2025, https://doi.org/10.1007/s11069-024-07027-9.

[8] A. Rowley and O. Karakus, "Predicting air quality via multimodal AI and satellite imagery," *Remote Sensing of Environment*, vol. 293, p. 113609, Aug. 2023, https://doi.org/10.1016/j.rse.2023.113609.

[9] P. Vongruang, K. Suppoung, S. Kirtsaeng, K. Prueksakorn, P. T. B. Thao, and S. Pimonsree, "Development of Meteorological Criteria for Classifying PM2.5 Risk in a Coastal Industrial Province in Thailand," *Aerosol and Air Quality Research*, vol. 24, no. 10, p. 230321, 2024, https://doi.org/10.4209/aaqr.230321.

[10] S. Saminathan and C. Malathy, "Ensemble-based classification approach for PM2.5 concentration forecasting using meteorological data," *Frontiers in Big Data*, vol. 6, p. 1175259, Jun. 2023, https://doi.org/10.3389/fdata.2023.1175259.

[11] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," *IEEE Access*, vol. 10, pp. 99 129–99 149, 2022, https://doi.org/10.1109/ACCESS.2022.3207287.

[12] A. S. Handayani, S. Soim, T. E. Agusdi, and N. L. Husni, "Air Quality Classification Using Support Vector Machine," *Computer Engineering and Applications Journal*, vol. 10, no. 1, pp. 55–69, Feb. 2021, https://doi.org/10.18495/comengapp.v10i1.350.

[13] A. Demircioglu, "The effect of feature normalization methods in radiomics," *Insights into Imaging*, vol. 15, no. 1, p. 2, Jan. 2024, https://doi.org/10.1186/s13244-023-01575-7.

[14] T. Bikaun, T. French, M. Hodkiewicz, M. Stewart, and W. Liu, "LexiClean: An annotation tool for rapid multi-task lexical normalisation," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, November, 2021, pp. 212–219, https://doi.org/10.18653/v1/2021.emnlp-demo.25.

[15] M. Sivananda and D. G. K. Kumar, "Classification and Regression Based on Decision Tree Algorithm for Machine Learning," *Interantional Journal of Scientific Research in Engineering and Management*, vol. 08, no. 02, pp. 1–13, Feb. 2024, https://doi.org/10.55041/IJSREM28533.

[16] N. Kokash and L. Makhnist, "Using Decision Trees for Interpretable Supervised Clustering," *SN Computer Science*, vol. 5, no. 2, p. 268, Feb. 2024, https://doi.org/10.1007/s42979-023-02590-7.

[17] R. Kumar, B. Krishna Goswami, S. Motiram Mhatre, and S. Agrawal, "Naive Bayes in Focus: A Thorough Examination of its Algorithmic Foundations and Use Cases," *International Journal of Innovative Science and Research Technology (IJISRT)*, vol. 9, no. 5, pp. 2078–2081, Jun. May, 2024, https://doi.org/10.38124/ijisrt/IJISRT24MAY1438.

[18] L.-K. Foo, S.-L. Chua, and N. Ibrahim, "Attribute weighted naive bayes classifier," *Comput. Mater. Contin*, vol. 71, no. 1, pp. 1945–1957, 2022, https://doi.org/10.32604/cmc.2022.022011.

[19] P. Dilliswar Reddy and L. Rama Parvathy, "Prediction of Air Pollution Level in Particular Region Area Using Logistic Regression and Naive Bayes," in *Advances in Parallel Computing*, D. J. Hemanth, T. N. Nguyen, J. Indumathi, and S. Lakshmanan, Eds. IOS Press, Nov. 2022, vol. 41, no. 1, https://doi.org/10.3233/APC220088.

[20] M. Suyal and P. Goyal, "A Review on Analysis of K-Nearest Neighbor Classification Machine Learning Algorithms based on Supervised Learning," *International Journal of Engineering Trends and Technology*, vol. 70, no. 7, pp. 43–48, Jul. 2022, https://doi.org/10.14445/22315381/IJETT-V70I7P205.

[21] C. Feng, B. Zhao, X. Zhou, X. Ding, and Z. Shan, "An Enhanced Quantum K-Nearest Neighbor Classification Algorithm Based on Polar Distance," *Entropy*, vol. 25, no. 1, p. 127, Jan. 2023, https://doi.org/10.3390/e25010127.

[22] M. Gavidia-Calderon, D. Schuch, A. Vara-Vela, R. Inoue, E. D. Freitas, T. T. D. A. Albuquerque, Y. Zhang, M. D. F. Andrade, and M. L. Bell, "Air quality modeling in the metropolitan area of Sao Paulo, Brazil: A review," *Atmospheric Environment*, vol. 319, p. 120301, Feb. 2024, https://doi.org/10.1016/j.atmosenv.2023.120301.

[23] S. Li and S. Qu, "Fund Performance Evaluation Based on Bayesian Model and Machine Learning Algorithm," *Discrete Dynamics in Nature and Society*, vol. 2022, no. 1, p. 2467521, Jan. 2022, https://doi.org/10.1155/2022/2467521.