Journal of Applied Informatics and Computing (JAIC)
Vol.8, No.2, December 2024, pp. 478~486
e-ISSN: 2548-6861

Evaluation of the Decision Tree Model for Air Condition Classification on the Global Air Pollution Dataset

Cindy Dinda Sabella [1]*, Yoga Pristyanto [2]**
* Information System, Amikom University Yogyakarta
cindydinda@students.amikom.ac.id [1], yoga.pristyanto@amikom.ac.id[2]

Article Info

Article history:

Received 2024-10-01
Revised 2024-11-04
Accepted 2024-11-09

Keyword:

Decision Tree,
Air Quality Index,
Air Pollution,
Machine Learning,
Classification.

ABSTRACT

Air pollution is an urgent global environmental problem, with significant impacts on public health and ecosystem stability. This research aims to develop an air quality classification model using the Global Air Pollution dataset from Kaggle, which consists of 23,463 rows of data and 12 features, including important variables such as Air Quality Index (AQI), PM2.5, NO2, and O3. Decision Tree, Random Forest, and Support Vector Machine (SVM) algorithms are applied to perform classification, with a focus on hyperparameter tuning to increase model accuracy. The research results show that the Decision Tree provides the best results with an accuracy of 99.89% after tuning hyperparameters using the Grid Search method. The SVM model showed an improvement of 94.89% to 99.32%, while Random Forest recorded an accuracy of 96.87% with no significant improvement after tuning. Importance feature analysis identified PM2.5 and AQI as the dominant factors in influencing air quality, with PM2.5 having the highest importance value of 0.93. This research confirms that machine learning can be an effective tool for integrating and classifying air pollution. It is hoped that the integration of this model into a real-time air quality monitoring system can help make more responsive and precise decisions in dealing with air pollution problems.

I. INTRODUCTION

The global climate crisis is one of the greatest challenges facing the world today, with significant impacts on environmental health and the quality of human life. One of the most disturbing impacts of this crisis is the worsening air quality in various cities around the world. Air pollution has become a global environmental problem that affects public health, ecosystems and climate stability. Human activities, such as industrialization, transportation, biomass burning, and high energy consumption, have resulted in increased greenhouse gas emissions and pollutant particles in the atmosphere. Emissions from various sources produce pollutants such as carbon monoxide (CO), ozone (O_3), nitrogen dioxide (NO_2), and fine particles (PM2.5), which contribute to poor air quality in various regions, especially in large cities.

The latest report from IQAir.com in 2023 shows that the city of Delhi is ranked first with the worst air quality in the world, followed by other cities such as Dhaka and Lahore. Jakarta is also recorded as having increasingly worse air quality, which is caused by increasing urbanization and industrialization that are not environmentally friendly. The impact of air pollution is very broad, including an increase in respiratory diseases such as asthma, bronchitis, and lung cancer as well as worsening global warming conditions which have influenced extreme climate change.

In facing these challenges, effective monitoring and management of air quality is essential. For this reason, data science and machine learning technology can play an important role in helping understand and overcome this problem. The use of machine learning techniques allows more accurate predictions regarding air pollution conditions based on historical and real-time data. This research aims to develop an air pollution classification model using the Global Air Pollution dataset from Kaggle, which has the following characteristics: consisting of 23,463 rows of data and 12 features which include important variables such as Air

Quality Index (AQI), PM2.5, NO2, and O3. This dataset includes air quality data from various cities around the world, providing a good representation of air pollution conditions in various regions. Using this dataset, this research aims to classify air quality levels and identify dominant factors that influence changes in air quality.

This research uses three main machine learning algorithms, namely Decision Tree, Random Forest, and Support Vector Machine (SVM), to classify air quality. The Decision Tree model was chosen because of its ability to provide clear interpretations and decision rules that are easy to understand. The Random Forest model is used to increase accuracy with an ensemble method that combines several decision trees, while SVM is chosen because it is effective in handling non-linear data.

This model is designed to analyze and classify air quality based on historical data, with the specific goal of providing insight into air quality trends over time. In addition, this research aims to identify the main factors that influence air quality, as well as provide information that can help in making policies that are better and more responsive to changing air pollution conditions. This model is also planned to be integrated into real air quality monitoring systems, which can provide real-time air quality predictions and help make more responsive decisions. Thus, this research not only focuses on historical analysis, but also seeks to provide practical solutions to the increasingly pressing problem of air pollution.

However, this research also has limitations. The Decision Tree model used has the potential to experience overfitting if it is not adjusted properly through hyperparameter tuning. These limitations may affect the generalization of the model to new data. In addition, the dataset used has a specific geographic coverage and a limited number of variables, which may influence the results and interpretation. Further research is recommended to use more diverse datasets and expand the variables analyzed, in order to improve the accuracy and reliability of the model in a broader context.

Several previous literature supports these findings, as described in research by E Sutoyo (2021), who also used Decision Tree and SVM algorithms to analyze air quality in DKI Jakarta. Although Decision Trees prove to be easier to interpret, their results are slightly inferior to SVMs, especially when dealing with datasets with high variability. Another research by Ahmad Efendi (2023) uses the Random Forest algorithm to predict forest fires in Riau with 97% accuracy. In addition, Sentinel-2 imagery and the Normalized Burn Ratio (NBR) method are used to visualize and measure the fire area. The combination of these two methods aims to support forest fire prevention through an early warning system and reduce environmental and health impacts due to fire.

In research conducted by I Irwansyah (2023), Decision Tree was compared with Naive Bayes and K-Nearest Neighbor (KNN) in air classification in Jakarta. Although Decision Trees are superior in terms of interpretability, Naive Bayes provides better accuracy on small datasets. This highlights that in larger and more varied datasets, ensemble algorithms such as Random Forest may be more effective. A further study from B Sunarko (2023) used a Decision Tree as part of a Stacking Ensemble to predict the impact of air pollution on health. Decision Trees play an important role in supporting easy-to-understand interpretation of results, although ensemble models are overall superior in prediction accuracy. Research from Rizky Fauzi Ramadhani (2022) also shows that Decision Trees are superior to Artificial Neural Networks (ANN) in several predictions of air pollution in DKI Jakarta, with accuracy reaching 99%. Nevertheless, the use of ANNs provides additional insight into algorithm performance in non-linear contexts. Finally, a study by F Widiawati (2023) shows that although Naive Bayes is superior in accuracy, Decision Tree remains competitive in predicting air pollution levels in South Tangerang, with an advantage in terms of interpretability. Although, this study does not explore in depth the role of Decision Trees, the results are still relevant for comparing model performance.

Overall, these studies reinforce the finding that Decision Trees, although sometimes less accurate than other models such as SVM or Naive Bayes, have great advantages in interpretability and the ability to produce predictions that can be clearly understood by policy makers.

II. METHOD

Research stages This research was carried out through several stages, namely Data Collection (Dataset), Preprocessing, EDA (Exploratory Data Analysis), Modeling, Hyperparameter Tuning, Evaluation.

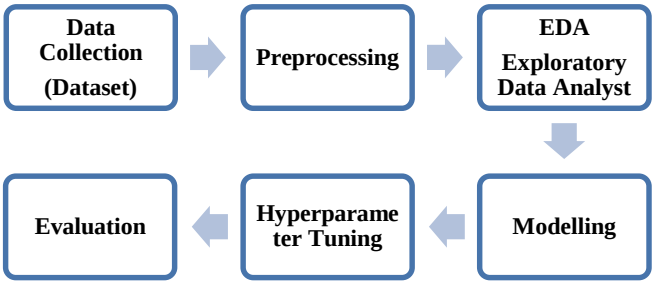

Figure 1. Research Stages

A. Data Collection (Datasets)

This research uses the "Global Air Pollution" dataset obtained from Kaggle. This dataset can be accessed via the following link.
https://www.kaggle.com/datasets/hasibalmuzdadid/global-air-pollutiondataset?select=global+air+pollution+dataset.csv
This dataset contains air pollution data from various cities in the world with main variables such as Air Quality Index (AQI), PM2.5, NO_2, and O_3. The dataset consists of 23,463 data rows and 12 columns, representing information related to

air quality in various countries and cities. The data is used to train and test air classification models.
Table 1 contains the data set used.

TABLE I.
ATTRIBUTES DATASETS

Attribute	Deskription
Country	Country name
City	City name
AQI Value	Air quality index value
AQI Category	Air quality index category
CO AQI Value	Carbon monoxide air quality index value
CO AQI Category	Carbonmonoxide category of air quality index
Ozone AQI Value	Air quality index ozone value
Ozone AQI Category	Air quality index ozone category
NO2 AQI Value	Nitrogen dioxide air quality index value
NO2 AQI Category	Nitrogen dioxide category of air quality index
PM 2.5 AQI Value	The particulate value is less than 2.5 micrometers
P.M 2.5 AQI Category	The air quality index category includes particulates measuring 2.5 or smaller micrometers

B. Preprocessing

At the data cleaning stage, the following steps are carried out:
Deleting Missing Values: Found 427 missing values in the Country column and 1 missing value in the City column. Missing values were removed from the dataset to maintain data integrity.
Split Data: The dataset is split into training data (80%) and test data (20%) to train the model and test the model's performance on never-before-seen data. The amount of data used is:
Training data: 18,428 rows
Test data: 4,607 lines
Data is split using the formula:
Train_Data = 0.8 x Total_Data
Test_Data = 0.2 x Total_Data

C. EDA (Exploratory Data Analysis)

Exploratory data analysis is performed to identify patterns and trends in the dataset. Based on EDA, the following information is obtained:
The majority of AQI values for all pollutants (CO, O_3, NO_2, PM2.5) are in the low range, which shows that air quality tends to be good in most areas.
The correlation heatmap shows that the PM2.5 AQI Value has the strongest correlation with the overall AQI value. This graph shows that fine particulate pollutants (PM2.5) have a significant influence on air quality in various cities.

D. Modelling

In this research, three machine learning algorithms are applied for air quality classification, namely Decision Tree, Random Forest, and Support Vector Machine (SVM). Each algorithm is implemented with appropriate parameters, as follows:
Decision Tree builds a decision tree based on air pollution features. Each node in the decision tree separates data based on rules that maximize the information gain or Gini index.
Gini Index is calculated as:

$$\text{Gini} = 1 - \sum_{i=1}^{n} (pi)2$$

The entropy formula used in Decision Trees to calculate the level of disorder is:
Entropy(S)=−i=1∑npilog2(pi)
Where:
S is a data set
Pi is the probability of each data class.
These formulas are used to measure node homogeneity, where the lower the entropy or Gini value, the better the separation produced by the Decision Tree.
Parameters used:
Max Depth: Sets the maximum depth of the tree to avoid overfitting.
Min Samples Split: Specifies the minimum number of samples required to split a node.
Criterion: Split evaluation method, using Gini or Entropy.
Decision Trees were chosen because of their ability to provide easy interpretation of the decision-making process.

Random Forest is a combination of several decision trees to reduce overfitting. Each tree is trained using a random subset of the data, and the final result is the average of all decision trees. Formula to calculate average prediction from Random Forest:

^y=n1i=1∑nfi(x)

Where :
fi(x) is the prediction of the ith decision tree
N is the number of trees in the forest (forest)
Parameters used:
n_estimators: The number of trees in the forest that affects the accuracy and stability of the model.
Max Features: Sets the number of randomly selected features for each tree, increasing tree variety.
Random Forest was chosen because of its ability to increase accuracy by combining multiple models and minimize overfitting by exploiting variation between trees.

Support Vector Machine (SVM) is used to separate data classes with a hyperplane that maximizes the margin between classes. The formula for calculating maximum margin is:

$w \cdot x - b = 0$

Where w is the weight vector, x is the input feature, and b is the bias. The goal of SVM is to maximize the margin 2/||w||, while minimizing classification error.
In the case of non-linear data, SVM uses a karnel trick, such as RBF (Radial Basis Function) which is defined by the formula:

$K(xi,xj) = \exp(-y||xi - xj||2)$

Parameters used:
C: A regularization parameter that controls how tight the margins of the hyperplane are.
Gamma: Controls how much influence one data point has on another.
Karnel: In this study, the RBF kernel was chosen to handle non-linear data.
SVM was chosen for its ability to handle non-linear data and its strong performance in various classification tasks.

E. Hyperparameter Tuning

Each algorithm goes through a hyperparameter tuning process using Grid Search to improve model accuracy and performance. Parameters tested in tuning include:

1) Decision Tree:
   - Max Depth: The maximum depth of the decision tree.
   - Min Samples Split: The minimum number of samples required to split a node.
   - Criterion: The method used to measure split quality (Gini or Entropy).
2) Random Forest:
   - Number of Trees (n_estimators): Number of trees in the forest
   - Max Features: The number of features considered when searching for the best split.
3) SVM
   - C: A regularization parameter that controls how tight the margins of the hyperplane are.
   - Gamma: Controls how much influence one data point has on another.
   - Karnel: Function used to project data (linear, RBF, polynomial).

F. Evaluation

Following training and fine-tuning, test data is used to assess the model's performance using a number of indicators, including:

1) Accuracy: Indicates how frequently the model predicts the future correctly.
2) Precision: Evaluates how well the model predicts favourable outcomes.
3) Recall : Evaluates the model's ability to identify every positive sample.
4) F1-Score: The precision and recall harmonic average

Formula for evaluating accuracy:

$$\text{Accuracy} = \frac{TP+TN}{TP + TN + FP + FN}$$

Where FN stands for False Negative, FP for False Positive, TN for True Negative, and TP for True Positive.
In addition, feature importance analysis is used to determine which features are most influential. The findings indicate that, with a high significance value, PM2.5 AQI is the most significant variabel.

III. Result And Discussion

A. Data Processing Result

In the initial stage, the dataset used was Global Air Pollution from Kaggle, with a total of 23,463 rows of data and 12 columns containing information about air quality in various cities in the world. After going through data cleansing, several important steps are taken:

1) Missing Values Cleanup: Found 427 missing values in the Country column and 1 missing value in the City column. Missing data is removed from the dataset to ensure the integrity of the data to be used in model training.
   To start, the number of missing values in the dataset is analyzed to identify potential problems that could affect the validity of the statistical model. The table below shows the results of calculating the number of missing values in each dataset column.

TABLE II.
Missing Values

Column	Number of Missing Values
Country	427
City	1
AQI Value	0
AQI Category	0
CO AQI Value	0
CO AQI Category	0
Ozone AQI Value	0
Ozone AQI Category	0
NO2 AQI Value	0
NO2 AQI Category	0
PM2.5 AQI Value	0
PM2.5 AQI Category	0

From the table above, it can be seen that missing values occurred mainly in the Country column with 427 entries and City with 1 entry. The presence of missing values in the Country column can have a significant impact on cross-regional analysis, because this column is an important variable used to separate data based on geographic location. Therefore, techniques for handling missing values must be applied, such as deleting entries with missing values or

nearest location-based imputation methods (e.g. using data from neighboring countries or similar cities). A condition where most of the pollutant variables and AQI do not have missing values is a good indication, which allows statistical analysis to be carried out without additional handling of missing data. This avoids inaccuracies that can arise due to inappropriate imputation methods.

2) Data Division (Train-Test Split)

Data is divided into training data (80%) and test data (20%). After division, where the size of the original dataset is 23,035 rows with 4 features, after that the training data consists of 18,428 rows and 4 features, while the test data consists of 4,607 rows with 4 features.

Results of data division showing the proportion of training data and test data:

TABLE III.
DATA SHARING RESULT

Datasets	Number of Rows	Number of Features
Total Data	23.035	4
Training Data	18.428	4
Test Data	4.607	4

The data division process is carried out to separate the dataset into two main parts, namely training data and test data. Training data is used to train the model, while test data is used to evaluate the model's performance on data that the model has never seen before. In this study, the initial dataset had 23,035 rows and 4 features. The data is divided into a proportion of 80% for training data and 20% for test data.

Based on the data division results shown in the table above, the training data consists of 18,428 rows with 4 features, while the test data consists of 4,607 rows with 4 features. This division is done to ensure that the model obtains enough information from the training data so that it can learn well, and at the same time, enough data is set aside for testing to ensure that the model evaluation is carried out objectively. This 80:20 proportion is an approach that is often used in machine learning practice because it provides a balance between the availability of data to train the model and evaluation of model performance.

This division is important to avoid overfitting, namely when the model learns too well on training data but is unable to generalize its performance on test data. By having proportional test data, this risk can be minimized and model evaluation can be better.

After processing the training data and test data using scaling techniques, the data size remains consistent, namely:

TABLE IV.
DATA SCALING RESULT

Datasets	Number of Rows	Number of Features
Training Data	18.428	4
Test Data	4.607	4

After data sharing is carried out, the next step is to apply the feature scaling process. Scaling is a technique used to normalize features so that they are on the same scale. This is very important, especially in algorithms like Support Vector Machines (SVM), where the distance between data points depends heavily on the scale of the features.

The scaling process is carried out by changing the distribution of values for each feature in the training data and test data so that they are in the same range. Based on the results shown in the table above, the dataset size after scaling remains consistent with the size before scaling, namely 18,428 rows for training data and 4,607 rows for test data, with 4 features for each part. This shows that the scaling process only affects feature values without changing the structure of the dataset.

Applying scaling helps improve model performance because the model can account for each feature proportionally. If scaling is not done, features with a larger scale can dominate the training process, which can lead to an inaccurate model. By scaling, the contribution of each feature to predictions becomes more balanced, so that the model can be more optimal in producing predictions.

B. Exploratory Data Analyst (EDA)

EDA is carried out to understand more deeply the distribution of data and the pattern of relationships between important variables. Some of the analyzes carried out are:

1) AQI and Pollutant Distribution

The distribution graph of AQI and main pollutants (PM2.5, CO, NO_2, O_3) shows that most cities have low AQI values, which means that air quality in most areas tends to be good. However, there are several areas that have high AQI values, especially those related to PM2.5 pollutants.

Distribution graph showing the distribution of AQI values in various cities:

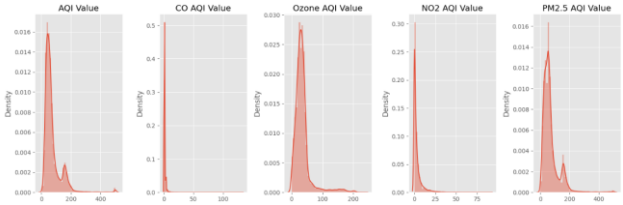

Figure 2. Distribution graph of AQI Values

Pm2.5 is seen as a pollutant with the highest value distribution, especially in dense urban areas.

The graph above shows the AQI value. CO. Ozone, NO2 and high frequency PM 2.5 are rated low which shows that air quality tends to be good while only a few are rated high. From the density distribution graph of AQI values for CO, Ozone, NO2 and PM2.5 pollutants, there are several conclusions that can be drawn:

Distribution of AQI Values: The majority of AQI values for all pollutants (CO, Ozone, NO2, PM2.5) are in the low range,

indicating that air quality tends to be good in most of the areas analyzed.
Low Frequency of High Values: Only a few AQI values reach high values, indicating that severe air pollution events are relatively rare.
Specific Pollutants: Each pollutant has a distribution that shows a similar pattern, with most values in the low range, especially for CO and NO2 which are almost all in the very low range.
Overall, this analysis shows that air quality in the areas analyzed is generally good, with a few minor exceptions where AQI values reached higher levels.

2) Heatmap of Correlation Between Features

The correlation heatmap between variables shows that the PM2.5 AQI Value feature has the highest correlation with the AQI Value value. This shows that PM2.5 is the most significant factor affecting air quality in various cities.
Heatmap of the correlation between features generated from the notebook, shows a strong correlation between PM2.5 and AQI:

Figure 3. Heatmap Correlation

The correlation between PM2.5 and AQI Value is the strongest, with a correlation value close to 1. The correlation heatmap above illustrates the correlation between various Air Quality Index (AQI) values for various pollutants. This includes the overall AQI, as well as specific AQIs for Carbon Monoxide (CO), Ozone, Nitrogen Dioxide (NO2), and PM2.5 (particulate matter with a diameter less than or equal to 2.5 micrometers). Correlation values range from -1.0 to 1.0, represented by a color scale from blue to yellow. A value of 1.0 indicates perfect positive correlation, meaning that as one feature increases, the other features also increase by the same proportion. Conversely, negative values indicate negative correlation, where one value increases while the other decreases. In this heatmap, PM2.5 shows a very high correlation with the overall AQI value (0.98), which indicates that PM2.5 has a big influence on the overall AQI calculation. Correlations between other pollutants vary from low to moderate.

C. Model Development

To predict and classify air pollution conditions, three main algorithms are used: Decision Tree, Random Forest, and Support Vector Machine (SVM).
The accuracy prediction results before tuning for each model can be seen in the following table:

TABLE V.
ACCURACY RESULT BEFORE TUNING

Model	Accuracy Before Tuning
Decision Tree	97.91%
Random Forest	96.87%
SVM	94.89%

Before tuning, Decision Tree recorded the highest accuracy of 97.91%, indicating that this model is quite effective in separating data, even without further optimization. Random Forest, which combines multiple Decision Trees, is in second place with an accuracy of 96.87%, slightly lower than the single Decision Tree model, but still performs well. SVM recorded the lowest accuracy, namely 94.89%, which although quite good, still requires improvement through tuning to achieve optimal performance.

D. Hyperparameter Tuning

After initial model training, hyperparameter tuning is carried out to improve model accuracy. The parameters tested include max depth and min samples split in Decision Tree, as well as number of trees (n_estimators) in Random Forest.

TABLE VI.
COMPARISON OF MODEL ACCURACY BEFORE AND AFTER TUNING

Model	**Accuracy Before Tuning**	**Accuracy After Tuning**	**Improved Accuracy**
Decision Tree	97.91%	99.89%	1.98%
Random Forest	96.87%	96.87%	0%
SVM	94.89%	99.32%	4.43%

The results of this research show that the Decision Tree model provides the best performance with accuracy increasing from 97.91% to 99.89% after the hyperparameter tuning process. The tuning process is carried out using the Grid Search method, which facilitates the search for optimal parameters by evaluating various combinations of the main parameters. In this case, the parameters to be adjusted are Max Depth (maximum depth of the decision tree), Min Samples Split (minimum number of samples to divide nodes, and Criterion (split evaluation method such as Gini or Entropy).
The success of this tuning shows that appropriate parameter adjustments can reduce overfitting on training data

while increasing generalization ability on test data. This 1.98% increase in accuracy is quite significant, especially in the context of very complex air quality classification, where features such as PM2.5 and AQI have a very strong correlation and directly influence prediction results. Apart from that, the Random Forest model did not experience a significant increase in accuracy even though it had gone through the tuning process. This may be due to the fact that Random Forest, which uses multiple Decision Trees, has achieved optimal performance with its initial configuration. This indicates that parameters such as the number of trees (n_estimators) do not have a significant impact on the dataset used. In the SVM model, the largest increase in accuracy reached 4.43%, which shows that tuning parameters such as C (regularization) and gamma greatly influences model performance. SVM utilizes optimal margins to separate classes, and hyperparameter adjustments make the model better at handling margins between classes that might previously have been too narrow or too wide.

Overall, this research confirms that appropriate hyperparameter tuning methods can provide substantial accuracy improvements, especially on simpler models such as Decision Trees. On the other hand, for more complex models such as Random Forest and SVM, hyperparameter tuning can result in varying performance improvements depending on the data structure and characteristics of the problem at hand.

D. Evaluation Model

After tuning, the model is evaluated using test data with metrics such as Accuracy, Precision, Recall, and F1-Score. The evaluation results show that the Decision Tree model has the best performance with the highest accuracy.

TABLE VII.
MODEL EVALUATION RESULT

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	99.89%	0.99	0.98	0.99
Random Forest	96.87%	0.62	0.65	0.63
SVM	99.32%	0.96	0.96	0.96

Based on the evaluation results in the table above, the Decision Tree model shows the best performance compared to Random Forest and SVM. With the highest accuracy of 99.89%, Decision Tree is able to classify data very well. SVM is in second place with an accuracy of 99.32%, while Random Forest has a lower accuracy, namely 96.87%.

Apart from accuracy, the precision and recall of the Decision Tree model are also superior, at 0.99 and 0.98 respectively, which shows very accurate positive predictions and high ability to detect positive data. SVM has precision and recall of 0.96, close to Decision Tree, but still slightly below it. In contrast, Random Forest shows a precision of 0.62 and a recall of 0.65, which indicates that this model is less than optimal in detecting and predicting positive classes.

In terms of F1-Score, Decision Tree is again ahead with a value of 0.99, followed by SVM with 0.96, while Random Forest only gets 0.63. Overall, Decision Tree was the best model in all evaluation metrics, followed by SVM, while Random Forest showed the lowest performance and is not recommended as a primary choice in this context.

E. Feature Importance

Feature importance analysis shows that PM2.5 AQI Value is the most influential feature in the Decision Tree model predictions, with an importance value of 0.93. Other features such as O_3 AQI Value and CO AQI Value have a lower impact.

The feature importance graph is produced from the Decision Tree model which shows the contribution of each feature.

```
Feature Importance:
           Feature  Importance
3  PM2.5 AQI Value        0.93
1  Ozone AQI Value        0.07
0     CO AQI Value        0.00
2    NO2 AQI Value        0.00
```

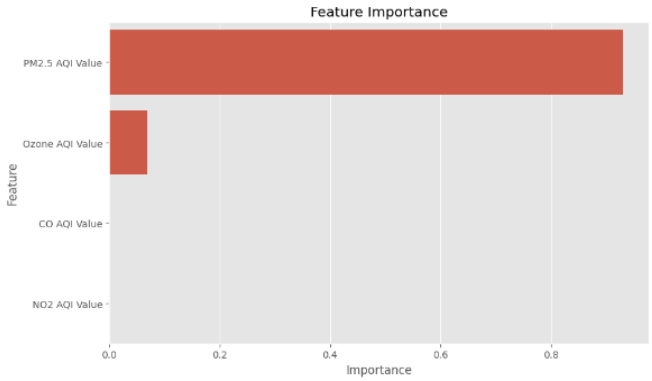

Figure 4. Feature Importance

Feature importance analysis is carried out to determine how much influence each feature has in predicting targets, in this case air quality. In the Decision Tree model, the importance of features is measured by calculating the contribution of each feature to reducing uncertainty (impurity) when the decision tree is built.

In this case, PM2.5 and AQI were identified as dominant factors, with an importance value of 0.93 each. This shows that PM2.5 is the most influential pollutant in the model classification. This measurement process is carried out by adding up the reductions carried out by adding up the impurity reductions from each node involving that feature throughout the tree. The more often features are used to divide data and the greater their influence on effective separation, the higher the effective feature importance value given.

This analysis supports the understanding that PM2.5 and AQI are important indicators in predicting air quality, and the

results are in line with findings in the literature showing that these pollutants contribute significantly to public health.

F. Evaluation Best Model (Decision Tree)

TABLE VIII.
ACCURACY BEST MODEL

Accuracy Of Best Model Decision Tree Before Tuning	Accuracy Of Best Model Decision Tree After Tuning
97.91%	99.89%

With this good classification performance, the decision tree model has proven to be very reliable for analyzing and categorizing air pollution data globally with a Test Accuracy of 97.91% before tuning and 99.89% after tuning. This model can be used to predict air quality with accuracy high levels, which is important for environmental monitoring and decision making regarding public health. Confusion Matrix Best Model (Decision Tree)

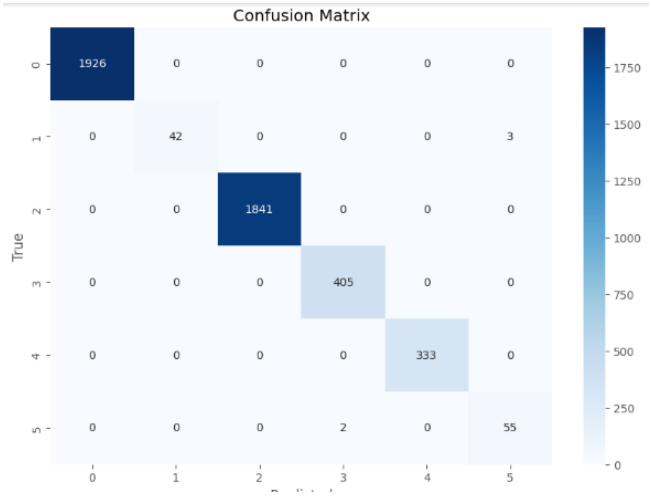

Figure 5. Confusion Matrix

The Confusion Matrix describes the performance of the Decision Tree classification model that you have trained. The main diagonal elements (1926, 1841) show the number of correct predictions for each class (Good and Moderate), where the model succeeded in predicting most of the data accurately. Elements outside the main diagonal (0, 42, 0, 405, 3) show prediction errors, with 42 misclassifications for the Good class as Moderate and 405 misclassifications for the Moderate class as Unhealthy. This could indicate that the decision boundaries between some classes may not be sharp enough, or that the data for those classes overlap.

Based on the Classification Report, the model shows excellent performance in classifying air quality. With almost perfect precision, recall, and F1-score in most categories such as Good, Moderate, Unhealthy, and Unhealthy for Sensitive Groups, this model is able to predict very accurately. However, in the Hazardous and Very Unhealthy categories, performance decreased slightly although it was still within good limits, with F1-scores of 0.97 and 0.96 respectively. The overall accuracy reached 1.00, indicating that this model is very effective in predicting the data as a whole, especially for categories with larger amounts of data

TABLE IX.
CLASSIFICATION REPORT

Category	Precision	Recall	F1-Score	Support
Good	1.00	1.00	1.00	1926
Hazardous	1.00	0.93	0.97	45
Moderate	1.00	1.00	1.00	1841
Unhealthy	1.00	1.00	1.00	405
Unhealthy for Sensitive Groups	1.00	1.00	1.00	333
Very Unhealthy	0.95	0.96	0.96	57
Accuracy			1.00	4607
Macro Average	0.99	0.98	0.99	4607
Weighted Average	1.00	1.00	1.00	4607

G. Discussion

The evaluation results show that the Decision Tree model has the best performance with the highest accuracy after the hyperparameter tuning process. This model provides an accuracy rate of 99.89% in classifying air quality based on the Air Quality Index (AQI) value. Meanwhile, the Random Forest and Support Vector Machine (SVM) models also show good performance with accuracy of 96.87% and 99.32%. Decision Trees have advantages in terms of easier interpretation and faster processing compared to ensemble models such as Random Forest and SVM models.

Feature importance analysis carried out on the Decision Tree model shows that PM2.5 pollutants play a major role in determining air quality. PM2.5 has a very strong level of correlation with AQI values, which means that this fine particulate pollution greatly influences the air classification in the categories "good", "moderate", to "hazardous". Previous research also supports these findings, where PM2.5 particles have been identified as one of the main causes of respiratory problems and other health disorders. Because of their very small size, PM2.5 particles can easily enter the respiratory tract and cause damage to the lungs and cardiovascular system. This makes PM2.5 a critical indicator in assessing air quality in various cities. Although Random Forest models provide a high level of accuracy, they tend to be more complex and require more processing time than Decision Trees. With the ensemble method, Random Forest is able to reduce overfitting by utilizing many decision trees, which leads to more stable results than a single tree.

However, Support Vector Machine (SVM), although it can separate data classes with a good margin, shows slower performance than the other two models. This is because SVM uses a carnel function that calculates the distance between

data points in a high-dimensional space, which makes its computing time longer. Despite this, SVM is still a powerful model, especially when dealing with data that has complex separation boundaries. This research reinforces the conclusion that Decision Trees not only provide accurate results, but are also easier to interpret than other models, making them a powerful tool in decision making regarding air quality management.

IV. CONCLUSION

This research successfully implemented and evaluated the Decision Tree, Random Forest, and Support Vector Machine (SVM) models for classifying air pollution conditions using the Global Air Pollution dataset. Based on the evaluation results, the Decision Tree model was proven to be the most superior with an accuracy of 99.89% after hyperparameter tuning using the Grid Search method. SVM also showed significant improvement, with accuracy increasing from 94.89% to 99.32% after tuning, while Random Forest recorded an accuracy of 96.87%. Feature importance analysis shows that PM2.5 is the most significant factor in determining air quality, which is in line with findings from previous literature. This emphasizes the importance of fine particulate pollutants in having a significant impact on air quality and human health. The application of this model in a real-time air quality monitoring system can help the government and society respond quickly to changes in air pollution conditions. The model developed in this research not only allows more accurate predictions of air quality, but also provides better insight into the main factors influencing air pollution. However, this research also has limitations. Decision Tree models have the potential to experience overfitting if not tuned properly, which can reduce the generalization ability of the model. In addition, the dataset used has limited geographic coverage, so the results do not fully represent air quality conditions in a wider area. Therefore, further research is recommended to use more diverse datasets and expand the variables analyzed to improve the accuracy and reliability of the model in a global context.

REFERENCES

[1] F. Widiawati, R. Kurniawan, and T. Suprapti, "Klasifikasi Data Tingkat Kualitas Udara Di Tangerang Selatan Menggunakan Algoritma Naive Bayes," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 6, pp. 3739–3745, 2024, doi: 10.36040/jati.v7i6.8261.

[2] A. Efendi, I. Iskandar, R. Kurniawan, and M. Affandes, "Klasifikasi Kebakaran Hutan Riau Menggunakan Random Forest dan Visualisasi Citra Sentinel-2," *Kaji. Ilm. Inform. dan Komput.*, vol. 4, no. 3, pp. 1602–1612, 2023, doi: 10.30865/klik.v4i3.1521.

[3] R. F. Ramadhani, S. S. Prasetiyowati, and Y. Sibaroni, "Performance Analysis of Air Pollution Classification Prediction Map with Decision Tree and ANN," *J. Comput. Syst. Informatics*, vol. 3, no. 4, pp. 536–543, 2022, doi: 10.47065/josyc.v3i4.2117.

[4] I. Irwansyah, A. D. Wiranata, and T. T. M, "Komparasi Algoritma Decision Tree, Naive Bayes Dan K-Nearest Neighbor Untuk Menentukan Kualitas Udara Di Provinsi Dki Jakarta," *Infotech J. Technol. Inf.*, vol. 9, no. 2, pp. 193–198, 2023, doi: 10.37365/jti.v9i2.203.

[5] B. Sunarko *et al.*, "Penerapan Stacking Ensemble Learning untuk Klasifikasi Efek Kesehatan Akibat Pencemaran Udara," *Edu Komputika J.*, vol. 10, no. 1, pp. 55–63, 2023, doi: 10.15294/edukomputika.v10i1.72080.

[6] A. I. Sang, E. Sutoyo, and I. Darmawan, "Analisis Data Mining Untuk Klasifikasi Data Kualitas Udara Dki Jakarta Menggunakan Algoritma Decision Tree Dan Support Vector Machine Data Minning Analysis for Classification of Air Quality Data Dki Jakarta Using Decision Tree Algorthm and Support Vector ," *e-Proceeding Eng.*, vol. 8, no. 5, pp. 8954–8963, 2021.

[7] R. Rofiani, L. Oktaviani, D. Vernanda, and T. Hendriawan, "Penerapan Metode Klasifikasi Decision Tree dalam Prediksi Kanker Paru-Paru Menggunakan Algoritma C4.5," *J. Tekno Kompak*, vol. 18, no. 1, p. 126, 2024, doi: 10.33365/jtk.v18i1.3525.

[8] S. Syihabuddin Azmil Umri, "Analisis Dan Komparasi Algoritma Klasifikasi Dalam Indeks Pencemaran Udara Di Dki Jakarta," *JIKO (Jurnal Inform. dan Komputer)*, vol. 4, no. 2, pp. 98–104, 2021, doi: 10.33387/jiko.v4i2.2871.

[9] R. S. Amanu, F. A. Ramadhan, and A. H. Saputra, "Perbandingan Model Prediksi Data Mining Dalam Memprediksi Konsentrasi Polutan Karbon Monoksida (Co) Di Jakarta," *J. Teknol. Inf. J. Keilmuan dan Apl. Bid. Tek. Inform.*, vol. 18, no. 1, pp. 7–21, 2024, doi: 10.47111/jti.v18i1.12451.

[10] Y. V. Sari, Z. Muallifah, and A. Fanani, "Klasifikasi Kualitas Air Menggunakan Metode Extreme Learning Machine (ELM)," *J. JUPITER*, vol. 15, no. 2, pp. 983–994, 2023, [Online]. Available: https://jurnal.polsri.ac.id/index.php/jupiter/article/view/6995

[11] T. Madan, S. Sagar and D. Virmani, "Air Quality Prediction using Machine Learning Algorithms –A Review," 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 2020, pp. 140-145, doi: 10.1109/ICACCCN51052.2020.9362912.

[12] R. Murugan and N. Palanichamy, "Smart City Air Quality Prediction using Machine Learning," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2021, pp. 1048-1054, doi: 10.1109/ICICCS51141.2021.9432074.

[13] S. Ameer et al., "Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities," in IEEE Access, vol. 7, pp. 128325-128338, 2019, doi: 10.1109/ACCESS.2019.2925082.

[14] N. Das and Asaduzzaman, "An IoT-based System for Air Pollution Data Analysis and Visualization," 2021 5th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), Dhaka, Bangladesh, 2021, pp. 1-6, doi: 10.1109/ICEEICT53905.2021.9667912

[15] S. Rani, P. Kumari and S. K. Singh, "Machine Learning-based Multiclass Classification Model for Effective Air Quality Prediction," 2023 IEEE IAS Global Conference on Emerging Technologies (GlobConET), London, United Kingdom, 2023, pp. 1-7, doi: 10.1109/GlobConET56651.2023.10149947.