# Machine Learning Models for Predicting the Air Quality Index in Hanoi, Jakarta, and Bangkok

1st Emillio Hezekiah Lucas
*Information Systems Departement*
*Faculty of Technology and Informatics*
*Universitas Multimedia Nusantara*
Tangerang, Banten, Indonesia
emillio.hezekiah@student.umn.ac.id

2nd Raymond Sunardi Oetama
*Information Systems Departement*
*Faculty of Technology and Informatics*
*Universitas Multimedia Nusantara*
Tangerang, Banten, Indonesia
raymond@umn.ac.id

***Abstract*—The environment plays a crucial role in the survival of living beings, with air quality being a key factor affecting health and ecosystems. Clean air must be free from pollution that can disrupt the environmental balance. In Southeast Asia, the three cities with the highest levels of air pollution are Hanoi, Jakarta, and Bangkok. This study aims to evaluate the accuracy of air quality prediction models in these cities using two machine learning algorithms, namely Support Vector Machine and eXtreme Gradient Boosting, optimized with GridSearch Cross-Validation and the Synthetic Minority Over-sampling Technique. The primary focus of this research is to measure model accuracy and identify the most influential pollutant compounds. The results indicate that eXtreme Gradient Boosting achieved 100% accuracy in predicting air quality even without hyperparameter optimization and oversampling.**

***Keywords— Air Quality Index, eXtreme Gradient Boosting, GridSearch Cross-Validation, Synthetic Minority Over-sampling Technique Support Vector Machine***

## I. INTRODUCTION

The environment is one of the factors essential for the survival of living beings [1]. It influences various aspects such as nature, life, and well-being, making it crucial for sustaining life. A healthy environment is clean and free from pollution or contamination. Air pollution is an undesirable alteration in the physical, chemical, and biological properties of the air [2], which has the potential to endanger life and affect various activities of living beings [3].

Air Quality Index is a standardized measure of air quality that can be used to indicate health levels based on air pollution [4]. Based on data from the Air Quality Index in February 2025, the cities of Hanoi, Jakarta, and Bangkok are three cities from Southeast Asian countries that have poor air pollution indexes. Hanoi, Vietnam has the worst air pollution index in Southeast Asia and the second worst in the world, with an air quality index level of 212, which falls into the very unhealthy category. Jakarta, Indonesia ranks as the second worst air pollution index in Southeast Asia and the 18th worst in the world, with an air quality index level of 108, which falls into the unhealthy category for sensitive groups. Bangkok, Thailand ranks as the third worst air pollution index in Southeast Asia and the 36th worst in the world, with an air quality index level of 84, which falls into the moderate category [5].

Based on the air quality issues in the three capital cities of three Southeast Asian countries, a trial was conducted to process air quality data from these cities using machine learning algorithms. Machine learning has provided good quality in predicting air quality [6], [28]. Machine learning is a part of artificial intelligence that enables computers to learn the ability to process various statistical methods for classification [7]. Machine learning algorithms are used because they have the advantage of optimizing decision-making processes by automatically providing predictive insights [8].

Several previous studies have used various machine learning models, such as Support Vector Machine with hyperparameter optimization using GridSearch Cross-Validation and eXtreme Gradient Boosting with oversampling techniques using the Synthetic Minority Over-sampling Technique. Support Vector Machine with hyperparameter optimization using GridSearch Cross-Validation was able to produce a model with an accuracy of 94.9% in predicting air quality data in Jakarta for the 2020-2021 period. eXtreme Gradient Boosting with the Synthetic Minority Over-sampling Technique was able to produce a model with an accuracy of 98.9% in predicting air quality data in Jakarta for the 2016-2021 period [9], [10].

Based on these results, this study will compare the performance of Support Vector Machine and eXtreme Gradient Boosting in predicting air quality in the three cities, with and without hyperparameter optimization using GridSearch Cross-Validation and oversampling using the Synthetic Minority Over-sampling Technique. The goal of comparing these two models is to determine which model has higher accuracy based on training accuracy, testing accuracy, Stratified k-folds Cross-Validation (SKFCV), mean precision, mean recall, and mean F1-score. Additionally, this study will also identify Feature Importance (FI), which refers to the variables (pollutant compounds) that have the most significant influence on the model and determine their percentage impact on the model.

## II. METHODS

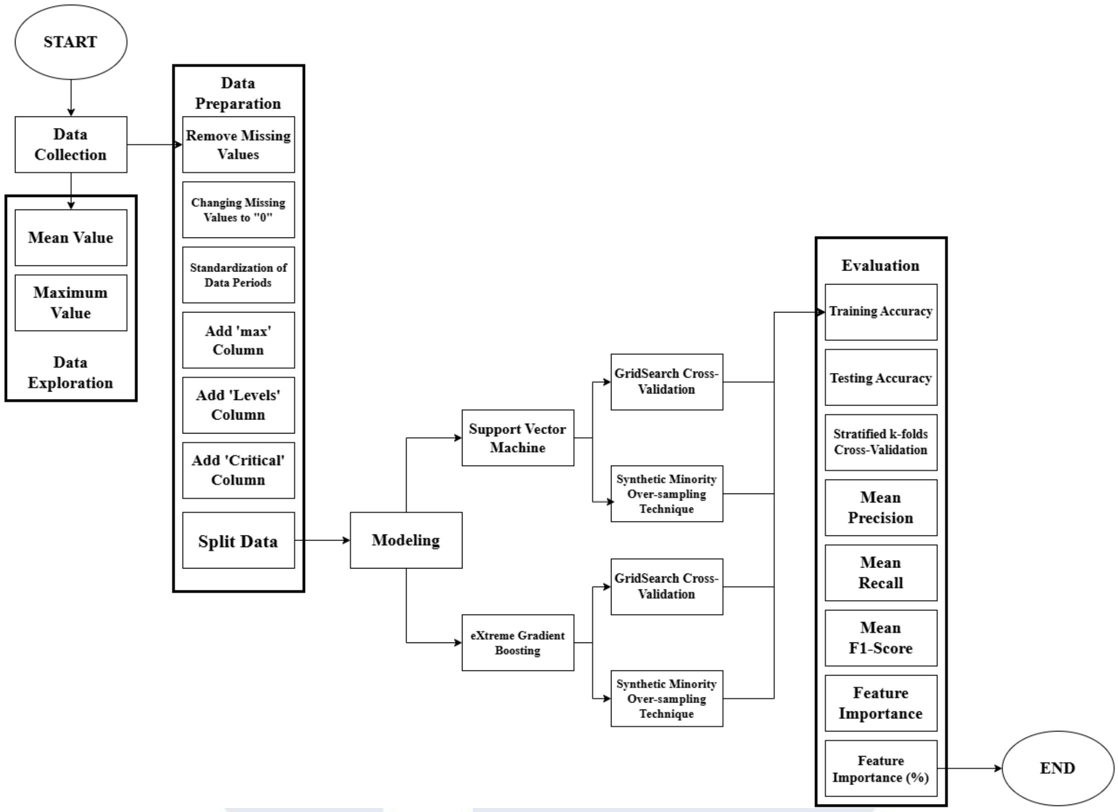


FIG. 1 RESEARCH FLOW

Fig. 1 shows the research flow. There are five stages in the research flow. The following is the implementation of these stages.

*A. Data Collection*

The dataset includes air pollution parameters, including $PM_{2.5}$ (fine particulate matter ⩽2.5 µm), $PM_{10}$ (coarse particulate matter ⩽10 µm), $O_3$ (ozone), $NO_2$ (nitrogen dioxide), $SO_2$ (sulfur dioxide), and CO (carbon monoxide). Data for Hanoi and Bangkok were obtained from (https://aqicn.org/), while data for Jakarta were sourced from (https://satudata.jakarta.go.id/). The Hanoi and Bangkok datasets only include the 'date' and pollutant columns. In contrast, the Jakarta dataset contains additional columns: 'max' (daily highest pollutant value), 'critical' (dominant daily pollutant), 'kategori' (air quality category), and 'lokasi_spku' (monitoring station location). Additionally, since the Jakarta dataset was separated by month and year, all data were merged into a unified dataset. The Hanoi dataset consists of 3.101 rows, Jakarta 10.386 rows, and Bangkok 4.068 rows.

*B. Data Exploration*

Table I. presents the three cities' average and maximum values of pollutant compounds. The highest average for Hanoi is $PM_{2.5,}$ which is 73,48, while the highest is O3, which is 498. The highest average for Jakarta is $PM_{10,}$ with a value of 55,49, while the highest value is $O_{3,}$ with a value of 243. The highest average for Bangkok is $PM_{2.5,}$ with a value of 81,49, while the highest value is $PM_{10,}$ with a value of 528.

TABLE I. POLLUTANT MEAN AND MAXIMUM VALUES FOR EACH CITY

| City | Statistic | $PM_{2.5}$ | $PM_{10}$ | $O_3$ | $NO_2$ | $SO_2$ | CO |
|---|---|---|---|---|---|---|---|
| Hanoi | Mean | **73,48** | 40,06 | 17,78 | 17,58 | 12,11 | 11,56 |
| | Maximum | 299 | 187 | **498** | 75 | 52 | 102 |
| Jakarta | Mean | - | **55,49** | 55,22 | 19,19 | 33,61 | 20,31 |
| | Maximum | - | 187 | **243** | 202 | 126 | 135 |
| Bangkok | Mean | **81,49** | 39,34 | 18,22 | 12,14 | 3,55 | 9,29 |
| | Maximum | 187 | **528** | 119 | 72 | 64 | 37 |

### C. Data Preparation

The data preprocessing stage includes various steps to ensure consistency and feasibility for analysis. This process begins with handling null values by either removing them or replacing them with "0." Next, data period standardization is performed. The Hanoi dataset covers 2014–2025, Jakarta 2010–2024, and Bangkok 2013–2025. Therefore, the analysis period is standardized to 2014–2024 to ensure a more structured and objectively comparable analysis. Subsequently, the $PM_{2.5}$ column in the Jakarta dataset is removed due to its inconsistency over several years. Conversely, additional columns, "max", "critical", and "Levels," are added to the Hanoi and Bangkok datasets to determine air quality categories. These categories are defined based on the daily maximum value of pollutant parameters and classified according to the Air Quality Index (AQI): 0–50 (Good), 51–100 (Moderate), 101–150 (Unhealthy for Sensitive Groups), 151–200 (Unhealthy), 201–300 (Very Unhealthy), and 301–500 (Hazardous) [11]. Finally, the dataset is split into training and testing data with a 90-10 ratio to build the model.

### D. Modeling

At the modeling stage, the formation of two models is carried out: Support Vector Machine and eXtreme Gradient Boosting. Both models will be optimized using GridSearch Cross-Validation for hyperparameter tuning and the Synthetic Minority Over-sampling Technique for oversampling. The modeling process is applied individually to three distinct datasets to ensure tailored optimization and performance evaluation for each. Support Vector Machine is a machine learning algorithm that seeks an optimal hyperplane with maximum margin to separate two classes [12], using kernel tricks (linear, polynomial, Radial Basis Function (RBF), sigmoid [13]) to handle nonlinear problems, leveraging a hypothesis space based on linear functions in high dimensions, and employing an optimization theory-based learning algorithm for training [14]. EXtreme Gradient Boosting is an ensemble learning method based on gradient boosting, which combines multiple decision trees with loss optimization, regularization, and pruning to improve accuracy and prevent overfitting [15]. GridSearch Cross-Validation is a procedure for searching optimal hyperparameters to enhance the model's predictive performance [16]. It evaluates the parameters used in the model and selects the best one [17]. The Synthetic Minority Over-sampling Technique is an oversampling method used to balance class distribution in a dataset by creating synthetic minority samples [16], [29]. It works by adding minority class samples to the dataset so that they are balanced with the majority class by generating new artificial data [18]. The model development consists of three main stages. First, building a baseline model without hyperparameter optimization using GridSearch Cross-Validation or the Synthetic Minority Over-sampling Technique. Second, applying hyperparameter optimization using GridSearch Cross-Validation for both models. Third, integrating the Synthetic Minority Over-sampling Technique to address data imbalance. Specifically for the Support Vector Machine model, a comparison is conducted using four types of kernels.

### E. Evaluation

Evaluation is based on training accuracy, testing accuracy, Stratified k-folds Cross-Validation, mean precision, mean recall, and mean F1-score. In addition, this study identifies the best model based on the results of training accuracy, testing accuracy, Stratified k-folds Cross-Validation, mean precision, mean recall, and mean F1-score. Also, the most influential pollutant variable on the model is also determined using feature importance (FI).

Accuracy is a metric that measures the proportion of correct predictions to the total data to assess the precision of a classification model [19]. The following is the formula for calculating accuracy (1).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

Note:

- TP: True Positive
- TN: True Negative
- FP: False Positive
- FN: False Negative

Stratified k-folds Cross-Validation is a validation technique that divides the dataset into k-blocks (folds) in a stratified manner [20]. One of the k-blocks is selected as the validation set, while the remaining k–1 blocks are used as the training set [20]. Stratified K-Fold Cross-Validation is used to determine whether the model is overfitting or underfitting [27].

Precision measures the proportion of correct positive predictions against all positive predictions to assess the relevance of the classification model [19]. The formula is as follows (2).

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

Recall measures the proportion of correct positive predictions to the total positive data, assessing the model's ability to identify relevant information [19]. The formula is as follows (3).

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

F1-score is the balanced average between precision and recall values [21]. The formula is as follows (4).

$$F1 - Score = \frac{2\,(recall\ x\ precision)}{recall+precision} \tag{4}$$

Then, an analysis was also conducted to determine which variable had the most influence on the model using feature importance. Feature importance is a technique to determine the influence of each feature on the model output [22] and to understand the factors shaping the model to avoid overfitting [23].

## III. EXPERIMENT

Table II. presents the performance of four Support Vector Machine kernels without hyperparameter optimization and oversampling techniques. The Linear kernel performs best for Hanoi, achieving 90.82% training and 92.69% testing accuracy, with the highest mean precision, recall, and F1-score are above 92%. For Jakarta and Bangkok, the Radial Basis Function kernel performs best, achieving 96.85% training and 95.43% testing accuracy for Jakarta, and 97.81% training and 95.76% testing accuracy for Bangkok. In both cities, mean precision, recall, and F1-score exceed 95%. Feature importance analysis highlights $PM_{2.5}$ as the dominant predictor for Hanoi (57.94%) and Bangkok (52.47%), while $O_3$ is most significant for Jakarta (57.23%).

Table III. shows the results of Support Vector Machine kernels optimized using GridSearch Cross-Validation. The Radial Basis Function kernel performs best for all cities, with varying optimal parameters. For Hanoi, C = 10 and Gamma = 0.1 achieve 99.14% training and 95.89% testing accuracy, with mean precision, recall, and F1-score exceeding 88%. In Jakarta, C = 100 and Gamma = scale result in 99.19% training and 97.84% testing accuracy, with mean precision, recall, and F1-score surpassing 97%. For Bangkok, C = 100 and Gamma = 0.1 lead to 99.42% training and 98.25% testing accuracy, with precision, recall, and F1-score all above 98%. Feature importance highlights $PM_{2.5}$ as dominant for Hanoi (57.21%) and Bangkok (54.81%), while $O_3$ is most significant for Jakarta (54.38%).

The results of Support Vector Machine kernels using the Synthetic Minority Over-sampling Technique is shown in Table IV. The Radial Basis Function kernel performs best for all cities with different parameters. For Hanoi, k_neighbors = 1 and random_state = 42 achieve 94.62% training and 94.46% testing accuracy, with mean precision, recall, and F1-score all above 94%. In Jakarta, k_neighbors = 5 and random_state = 42 result in 98.08% training and 93.52% testing accuracy, mean precision, recall, and F1-score exceeding 93%. For Bangkok, k_neighbors = 5 and random_state = 42 yield 97.19% training and 97.26% testing accuracy, with precision, recall, and F1-score all above 97%. Feature importance highlights $PM_{2.5}$ for Hanoi (59.02%), $O_3$ for Jakarta (58.46%), and $PM_{10}$ for Bangkok (53.47%).

TABLE II. VANILLA SUPPORT VECTOR MACHINE RESULT

| City | Kernel | Training Accuracy | Testing Accuracy | SKFCV | Mean Precision | Mean Recall | Mean F1-Score | FI | FI (%) |
|---|---|---|---|---|---|---|---|---|---|
| Hanoi | **Linear** | **90.82%** | **92.69%** | **84.92%** | **92.55%** | **93.69%** | **93.09%** | **$PM_{2.5}$** | **57.94%** |
| | Polynomial | 87.82% | 86.71% | 93.91% | 91.07% | 89.47% | 89.92% | $PM_{2.5}$ | 44.92% |
| | RBF | 95.19% | 93.02% | 93.71% | 71.96% | 72.82% | 72.37% | $PM_{2.5}$ | 55.08% |
| | Sigmoid | 70.69% | 67.44% | 42.20% | 41.73% | 41.73% | 41.77% | $PM_{2.5}$ | 31.69% |
| Jakarta | Linear | 89.72% | 89.83% | 88.16% | 82.62% | 86.61% | 82.71% | $O_3$ | 61.45% |
| | Polynomial | 92.73% | 91.49% | 89.90% | 94.52% | 89.49% | 91.55% | $O_3$ | 60.68% |
| | **RBF** | **96.85%** | **95.43%** | **94.36%** | **96.47%** | **93.64%** | **95.00%** | **$O_3$** | **57.23%** |
| | Sigmoid | 76.15% | 75.48% | 75.47% | 38.84% | 38.21% | 38.32% | $O_3$ | 76.02% |
| Bangkok | Linear | 94.40% | 93.02% | 88.25% | 93.06% | 93.02% | 93.00% | $PM_{2.5}$ | 51.82% |
| | Polynomial | 92.10% | 91.02% | 83.09% | 91.54% | 91.02% | 90.81% | $PM_{2.5}$ | 46.73% |
| | **RBF** | **97.81%** | **95.76%** | **88.75%** | **95.63%** | **95.76%** | **95.61%** | **$PM_{2.5}$** | **52.47%** |
| | Sigmoid | 76.09% | 75.31% | 75.26% | 73.75% | 75.31% | 74.51% | $PM_{2.5}$ | 30.87% |

TABLE III. SUPPORT VECTOR MACHINE USING GRIDSEARCH CROSS-VALIDATION RESULT

| City | Kernel | Parameters | Training Accuracy | Testing Accuracy | SKFCV | Mean Precision | Mean Recall | Mean F1-Score | FI | FI (%) |
|---|---|---|---|---|---|---|---|---|---|---|
| Hanoi | Linear | 'C': 100 | 91.01% | 91.63% | 90.31% | 88.77% | 92.56% | 89.94% | $PM_{2.5}$ | 55.63% |
| | Polynomial | 'C': 10, 'degree': 3 | 93.59% | 91.46% | 89.28% | 94.78% | 94.16% | 94.35% | $PM_{2.5}$ | 50.97% |
| | **RBF** | **'C': 10, 'gamma': 0.1** | **99.14%** | **95.89%** | **94.54%** | **95.75%** | **88.27%** | **90.60%** | **$PM_{2.5}$** | **57.21%** |
| | Sigmoid | 'C': 10, 'gamma': 0.01 | 88.75% | 89.00% | 87.72% | 58.52% | 57.38% | 57.75% | $PM_{2.5}$ | 51.43% |
| Jakarta | Linear | 'C': 100 | 89.68% | 89.71% | 88.26% | 89.66% | 89.71% | 89.51% | $O_3$ | 61.92% |
| | Polynomial | 'C': 10, 'degree': 3 | 93.79% | 92.50% | 90.57% | 92.40% | 92.50% | 92.22% | $O_3$ | 59.45% |
| | **RBF** | **'C': 100, 'gamma': 'scale'** | **99.19%** | **97.84%** | **96.77%** | **97.93%** | **97.84%** | **97.86%** | **$O_3$** | **54.38%** |
| | Sigmoid | 'C': 10, 'gamma': 0.01 | 89.03% | 88.69% | 87.01% | 88.40% | 88.69% | 88.36% | $O_3$ | 65.78% |
| Bangkok | Linear | 'C': 1 | 94.40% | 93.02% | 94.07% | 93.06% | 93.02% | 93.00% | $PM_{2.5}$ | 51.82% |
| | Polynomial | 'C': 10, 'degree': 3 | 96.34% | 95.26% | 95.29% | 95.36% | 95.26% | 95.22% | $PM_{2.5}$ | 53.99% |
| | **RBF** | **'C': 100, 'gamma': '0.1'** | **99.42%** | **98.25%** | **98.56%** | **98.26%** | **98.25%** | **98.26%** | **$PM_{2.5}$** | **54.81%** |
| | Sigmoid | 'C': 100, 'gamma': 0.01 | 92.91% | 91.02% | 92.55% | 91.05% | 91.02% | 91.00% | $PM_{2.5}$ | 48.98% |

TABLE IV. SUPPORT VECTOR MACHINE USING SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE RESULT

| City | Kernel | Parameters | Training Accuracy | Testing Accuracy | SKFCV | Mean Precision | Mean Recall | Mean F1-Score | FI | FI (%) |
|---|---|---|---|---|---|---|---|---|---|---|
| Hanoi | Linear | K_neighbors: 1, random_state:42 | 94.47% | 93.70% | 92.67% | 93.90% | 93.69% | 93.61% | $PM_{2.5}$ | 61.32% |
| | Polynomial | | 63.91% | 62.85% | 62.55% | 68.77% | 62.90% | 57.43% | $PM_{2.5}$ | 22.98% |
| | **RBF** | | **94.62%** | **94.46%** | **92.56%** | **94.44%** | **94.46%** | **94.39%** | **$PM_{2.5}$** | **59.02%** |
| | Sigmoid | | 87.73% | 87.15% | 85.32% | 87.24% | 87.16% | 86.87% | $PM_{2.5}$ | 53.58% |
| Jakarta | Linear | K_neighbors: 5, random_state:42 | 91.42% | 82.59% | 88.16% | 87.34% | 82.59% | 83.59% | $O_3$ | 64.33% |
| | Polynomial | | 94.39% | 90.60% | 89.90% | 91.00% | 90.60% | 90.73% | $O_3$ | 64.11% |
| | **RBF** | | **98.08%** | **93.52%** | **94.36%** | **94.63%** | **93.52%** | **93.75%** | **$O_3$** | **58.46%** |
| | Sigmoid | | 37.44% | 55.65% | 75.47% | 73.25% | 55.65% | 60.86% | N/A | N/A |
| Bangkok | Linear | K_neighbors: 5, random_state:42 | 96.38% | 96.35% | 94.91% | 96.38% | 96.35% | 96.34% | $PM_{10}$ | 56.05% |
| | Polynomial | | 74.94% | 74.62% | 74.41% | 83.83% | 74.64% | 73.17% | $PM_{10}$ | 47.42% |
| | **RBF** | | **97.19%** | **97.26%** | **95.42%** | **97.29%** | **97.26%** | **97.24%** | **$PM_{10}$** | **53.47%** |
| | Sigmoid | | 69.20% | 69.00% | 67.90% | 63.87% | 68.98% | 65.00% | $PM_{2.5}$ | 39.76% |

Table V. presents the results of eXtreme Gradient Boosting without hyperparameter tuning or oversampling. The model achieves very high performance across all cities, with 100% training and testing accuracy although Jakarta scores slightly lower at 99.87% compared to Hanoi and Bangkok. Mean precision, recall, and F1-score all reach 100% for Hanoi and Bangkok, while Jakarta scores slightly lower at 99.87%. Feature importance highlights $PM_{2.5}$ for Hanoi (88.71%) and Bangkok (88.71%), while $O_3$ is the key factor in Jakarta (36.52%).

Table VI shows the results of eXtreme Gradient Boosting with hyperparameter tuning using GridSearch Cross-Validation. Performance decreased slightly despite using the best parameters: colsample_bytree: 1.0, learning_rate: 0.01, max_depth: 5, min_child_weight: 1, n_estimators: 100, and subsample: 0.8. The model achieved 99.97% training accuracy, 100% testing accuracy, mean precision, recall, and F1-score exceeding 99%. Feature importance highlights $PM_{2.5}$ for Hanoi (80.72%) and Bangkok (88.36%), while $O_3$ is most important for Jakarta (40.60%).

The results of the eXtreme Gradient Boosting model enhanced with the Synthetic Minority Over-sampling Technique presented in Table VII. While the model achieved perfect training accuracy 100% using the best parameters *k_neighbors: n-1*, *random_state: 42* for Hanoi, and *k_neighbors: 5*, *random_state: 42* for Jakarta and Bangkok, its performance slightly decreased compared to the non-optimized version. The model achieved 100% training accuracy, 99.84% testing accuracy, mean precision, recall, and F1-score exceeding 99%. Feature importance highlights $PM_{2.5}$ for Hanoi (53.40%) and Bangkok (64.18%), while $O_3$ is most important for Jakarta (68.88%).

TABLE V. VANILLA EXTREME GRADIENT BOOSTING RESULT

| City | Parameters | Training Accuracy | Testing Accuracy | SKFCV | Mean Precision | Mean Recall | Mean F1-Score | FI | FI (%) |
|---|---|---|---|---|---|---|---|---|---|
| Hanoi | use_label_encoder=False, eval_metric="mlogloss", random_state=42 | **100%** | **100%** | 99.83% | **100%** | **100%** | **100%** | $PM_{2.5}$ | 88.71% |
| Jakarta | | **100%** | 99.87% | **99.90%** | 99.87% | 99.87% | 99.87% | $O_3$ | 36.52% |
| Bangkok | | **100%** | **100%** | 99.83% | **100%** | **100%** | **100%** | $PM_{2.5}$ | 88.71% |

TABLE VI. EXTREME GRADIENT BOOSTING USING GRIDSEARCH CROSS-VALIDATION RESULT

| City | Parameters | Training Accuracy | Testing Accuracy | SKFCV | Mean Precision | Mean Recall | Mean F1-Score | FI | FI (%) |
|---|---|---|---|---|---|---|---|---|---|
| Hanoi | colsample_bytree: 1.0<br>learning_rate: 0.01<br>max_depth: 5<br>min_child_weight: 1<br>n_estimators: 100<br>subsample: 0.8 | 99.75% | 99.18% | 99.51% | 99.63% | 99.56% | 99.59% | $PM_{2.5}$ | 80.72% |
| Jakarta | 'colsample_bytree': 1.0,<br>'learning_rate': 0.01,<br>'max_depth': 5,<br>'min_child_weight': 1,<br>'n_estimators': 100,<br>'subsample': 0.8 | **99.97%** | 99.87% | 99.82% | 99.87% | 99.87% | 99.87% | $O_3$ | 40.60% |
| Bangkok | 'colsample_bytree': 1.0,<br>'learning_rate': 0.01,<br>'max_depth': 5,<br>'min_child_weight': 1,<br>'n_estimators': 100,<br>'subsample': 0.8 | **99.97%** | **100%** | **99.85%** | **100%** | **100%** | **100%** | $PM_{2.5}$ | 88.36% |

TABLE VII. EXTREME GRADIENT BOOSTING USING SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE RESULT

| City | Parameters | Training Accuracy | Testing Accuracy | SKFCV | Mean Precision | Mean Recall | Mean F1-Score | FI | FI (%) |
|---|---|---|---|---|---|---|---|---|---|
| Hanoi | K_neighbors: 5, random_state:42 | **100%** | 99.81% | 99.68% | 99.81% | 99.81% | 99.81% | $PM_{2.5}$ | 53.40% |
| Jakarta | K_neighbors: 5, random_state:42 | **100%** | 99.72% | 99.81% | 99.72% | 99.72% | 99.72% | $O_3$ | 68.88% |
| Bangkok | K_neighbors: 5, random_state:42 | **100%** | **99.84%** | **99.82%** | **99.85%** | **99.85%** | **99.85%** | $PM_{2.5}$ | 64.18% |

The best model is eXtreme Gradient Boosting without hyperparameter tuning or oversampling, as it reaching 100% and above 99.87% for training and testing accuracy with stratified k-folds cross validation score above 99% which means the model is not overfitting and underfitting. $PM_{2.5}$ is the most influential pollutant in Hanoi and Bangkok, linked to respiratory issues [24], such as asthma, bronchitis, stroke, and lung cancer [25]. In Jakarta, $O_3$ is the key factor, increasing the risk of respiratory diseases, cardiovascular disorders, and mortality [26].

## IV. Conclusion and Future Work

### A. Conclusion

This study has successfully found which model is the best. EXtreme Gradient Boosting without any optimization is the best model in predicting air quality achieving 100% and above 99.87% for training and testing accuracy. Feature importance analysis shows that the key pollutant influencing air quality in Hanoi and Bangkok was $PM_{2.5}$, while in Jakarta, it was $O_3$.

### B. Limitation and Future Work

This research limited to only three cities. Future research can include more cities using the same model.


## ACKNOWLEDGEMENT

This study was supported by Universitas Multimedia Nusantara, whose collaboration plays crucial role in this success of this research.


## REFERENCES


[1] D. D. Sompotan and J. Sinaga, "Prevention of Environmental Pollution," SAINTEKES: Jurnal Sains, Teknologi Dan Kesehatan, vol. 1, no. 1, pp. 6–13, Jul. 2022, doi: 10.55681/saintekes.v1i1.2.

[2] P. P. Indonesia, "Law (UU) No. 32 of 2009 on Environmental Protection and Management."

[3] D. A. Kusumaningtiar, "Analysis of Soil Pollution Environmental Quality." *Journal of Reasearch and Education Chemistry,* vol. 7, pp. 158-170, April 2025.

[4] Joko Ade Nursiyono and Ima Sartika Dewi, "Utilization of Air Quality Index (AQI) as an Indicator of East Java's Economic Movement Amid the COVID-19 Pandemic," *Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika*, vol. 1, pp. 20–29, Jan. 2022.

[5] IQAir, "Ranking of the Most Polluted Major Cities."

[6] T. M. T. Lei, S. W. I. Siu, J. Monjardino, L. Mendes, and F. Ferreira, "Using Machine Learning Methods to Forecast Air Quality: A Case Study in Macao," *Atmosphere (Basel)*, vol. 13, no. 9, p. 1412, Sep. 2022, doi: 10.3390/atmos13091412.

[7] D. R. M. Nainggolan, "Data Science, Big Data, and Predictive Analytics: A Foundation for Cybersecurity Intelligence," *Jurnal Pertahanan & Bela Negara*, vol. 7, no. 2, Oct. 2017, doi: 10.33172/jpbh.v7i2.187.

[8] R. G. Wardhana, G. Wang, and F. Sibuea, "Application of Machine Learning in Predicting Disease Case Levels in Indonesia," *Journal of Information System Management (JOISM)*, vol. 5, no. 1, pp. 40–45, Jul. 2023, doi: 10.24076/joism.2023v5i1.1136.

[9] A. Toha, P. Purwono, and W. Gata, "Air Quality Prediction Model with Support Vector Machines Using Hyperparameter Optimization GridSearch CV," *Buletin Ilmiah Sarjana Teknik Elektro*, vol. 4, no. 1, pp. 12–21, May 2022, doi: 10.12928/biste.v4i1.6079.

[10] A. A. Nababan, M. Jannah, M. Aulina, and D. Andrian, "Air Quality Prediction Using XGBoost with Synthetic Minority Oversampling Technique (SMOTE) Based on the Air Pollution Standard Index (ISPU)," *Jurnal Teknik Informatika Kaputama (JTIK)*, vol. 7, no. 1, 2023.

[11] United States Environmental Protection Agency, "Air Data Basic Information."

[12] R. Tineges, A. Triayudi, and I. D. Sholihati, "Sentiment Analysis of Indihome Services Based on Twitter Using Support Vector Machine (SVM) Classification Method," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 4, no. 3, p. 650, Jul. 2020, doi: 10.30865/mib.v4i3.2181.

[13] Nanda Lutfi Syahputra, "Sentiment Analysis of Online Loans on Twitter Using Support Vector Machine and Logistic Regression," *Universitas Multimedia Nusantara*, Jul. 2023.

[14] …… Yoga, V. Wijaya, Y. Vikriansyah Wijaya, A. Erfina, and C. Warman, "Sentiment Analysis on the Electronic Information and Transactions Law (UU ITE) Using Support Vector Machine Algorithm", [Online]. Available: https://databoks.katadata.co.id/

[15] A. Moore and M. Bell, "XGBoost, A Novel Explainable AI Technique, in the Prediction of Myocardial Infarction: A UK Biobank Cohort Study," *Clin Med Insights Cardiol*, vol. 16, Jan. 2022, doi: 10.1177/11795468221133611.

[16] S. Bej, N. Davtyan, M. Wolfien, M. Nassar, and O. Wolkenhauer, "LoRAS: an oversampling approach for imbalanced datasets," *Mach Learn*, vol. 110, no. 2, pp. 279–301, Feb. 2021, doi: 10.1007/s10994-020-05913-4.

[17] R. Rofik, R. A. Hakim, J. Unjung, B. Prasetiyo, and M. A. Muslim, "Optimization of SVM and Gradient Boosting Models Using GridSearchCV in Detecting Fake Job Postings," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 23, no. 2, pp. 419–430, Mar. 2024, doi: 10.30812/matrik.v23i2.3566.

[18] N. Matondang and N. Surantha, "Effects of Oversampling SMOTE in the Classification of Hypertensive Dataset," *Advances in Science, Technology and Engineering Systems Journal*, vol. 5, no. 4, pp. 432–437, Aug. 2020, doi: 10.25046/aj050451.

[19] Serafim Clara, Dhea Laksmi Prianto, Rizal Al Habsi, Ester Friscila Lumbantobing, and Nurul Chamidah, "Implementation of Feature Selection in Machine Learning Classification Algorithms for Income Prediction on the Adult Income Dataset," *SENAMIKA*, Apr. 2021.

[20] I. Tougui, A. Jilbab, and J. El Mhamdi, "Impact of the Choice of Cross-Validation Techniques on the Results of Machine Learning-Based Diagnostic Applications," *Healthc Inform Res*, vol. 27, no. 3, pp. 189–199, Jul. 2021, doi: 10.4258/hir.2021.27.3.189.

[21] N. M. S. Iswari, "Enhancing Aspect-based Sentiment Analysis in Visitor Review using Semantic Similarity," *Journal of Applied Data Sciences*, vol. 5, no. 2, pp. 724–735, May 2024, doi: 10.47738/jads.v5i2.249.

[22] D. Rengasamy *et al.*, "Feature importance in machine learning models: A fuzzy information fusion approach," *Neurocomputing*, vol. 511, pp. 163–174, Oct. 2022, doi: 10.1016/j.neucom.2022.09.053.

[23] Faisal Ardiansyah, "Rental Price Prediction System Using Random Forest Analytics (Case Study: Exclusive Boarding Houses in the Special Region of Yogyakarta)," *UNIVERSITAS ISLAM INDONESIA*, 2020.

[24] Elsa Try Julita Sembiring, "Health Risks of PM2.5 Exposure in Ambient Air for Street Vendors Under the Pasar Pagi Asemka Flyover, Jakarta," *Jurnal Teknik Lingkungan*, vol. 26, no. 1, pp. 101–120, Apr. 2020.

[25] Kartika Dian Pertiwi, Ita Puji Lestari, and Alfan Afandi, "Environmental Health Risk Analysis of PM10 and PM2.5 Dust Exposure on Traffic

Volunteers on Diponegoro Street, Ungaran," *Environmental Health Risk Analysis of PM10 and PM2.5 Dust Exposure on Traffic Volunteers on Diponegoro Street, Ungaran,*vol. 6, no. 2, pp. 85–91, Jul. 2023.

[26] Rosatul Umah and Eva Gusmira, "Impact of Air Pollution on Public Health in Urban Areas," *Profit: Jurnal Manajemen, Bisnis dan Akuntansi*" *Profit: Jurnal Manajemen, Bisnis dan Akuntansi*, vol. 3, no. 3, Aug. 2024.

[27] W. Wijiyanto, A. I. Pradana, S. Sopingi, and V. Atina, "K-Fold Cross-Validation Technique for Evaluating Student Performance," *Jurnal Algoritma*, vol. 21, no. 1, May 2024, doi: 10.33364/algoritma/v.21-1.1618.

[28] E. Karin, F. Ernando, Winanti, "Comparative Analysis of KNN and Decision Tree Classification Algorithms for Early Stroke Prediction: A Machine Learning Approach," *Journal of Information Systems and Informatics,* vol. 6, no.1, March 2024, doi: https://doi.org/10.51519/journalisi.v6i1.661

[29] A. K. Dinar, A. S. Samuel, C.T. Vincencius, S. Jason, "Dealing Imbalance Dataset Problem in Sentiment Analysis of Recession in Indonesia," *IAES International Journal of Artificial Intelligence,* vol. 13, no. 2, June 2024, doi: 10.11591/ijai.v13.i2.pp2060-2072