Applied Computer Science, vol. 20, no. 1, pp. 173–188
doi: 10.35784/acs-2024-11

Submitted: 2024-01-12 | Revised: 2024-02-18 | Accepted: 2024-02-27

Keywords: knowledge management, air pollution, naïve bayes, decision tree, random forest

Siti ROHAJAWATI [0000-0002-6775-8997]*, *Hutanti SETYODEWI* [0009-0008-3937-652X]*,
Ferryansyah Muji Agustian TRESNANTO [0009-0007-7830-7415]*,
Debora MARIANTHI [0009-0002-8610-1357]*,
Maruli Tua Baja SIHOTANG [0000-0002-9348-5400]**

KNOWLEDGE MANAGEMENT APPROACH IN COMPARATIVE STUDY OF AIR POLLUTION PREDICTION MODEL

Abstract

This study utilizes knowledge management (KM) to highlight a documentation-centric approach that is enhanced through artificial intelligence. Knowledge management can improve the decision-making process for predicting models that involved datasets, such as air pollution. Currently, air pollution has become a serious global issue, impacting almost every major city worldwide. As the capital and a central hub for various activities, Jakarta experiences heightened levels of activity, resulting in increased vehicular traffic and elevated air pollution levels. The comparative study aims to measure the accuracy levels of the naïve bayes, decision trees, and random forest prediction models. Additionally, the study uses evaluation measurements to assess how well the machine learning performs, utilizing a confusion matrix. The dataset's duration is three years, from 2019 until 2021, obtained through Jakarta Open Data. The study found that the random forest achieved the best results with an accuracy rate of 94%, followed by the decision tree at 93%, and the naïve bayes had the lowest at 81%. Hence, the random forest emerges as a reliable predictive model for prediction of air pollution.

1. INTRODUCTION

Knowledge Management (KM) has evolved over time, with its primary focus on documentation. These challenges are expected to be addressed through the artificial intelligence (AI) execution, this will also change the way in which KM undergoes changes in handling the knowledge transition process (Taherdoost & Madanchian, 2023). Knowledge Management involves gathering, processing, and utilizing both tacit and explicit knowledge to improve decision-making processes and outcomes. For instance, in a corporate setting,

* Bakrie University, Faculty of Engineering and Computer Science, Information System, Indonesia, siti.rohajawati@bakrie.ac.id, hutanti.setyodewi@gmail.com, 1232912003@student.bakrie.ac.id, 1202722003@student.bakrie.ac.id
** PT. Festino Indonesia. IT Solution Architect, Indonesia, mtbsihotang@gmail.com

KM assists in fine-tuning prediction models by providing comprehensive insights and data analysis (Bilquise & Shaalan, 2022). Knowledge Management refers to the systematic process of capturing, organizing, and analyzing information within an organization. This process involves the use of cognitive information systems, collaborative tools, and knowledge-based agents to facilitate effective decision-making. By employing advanced techniques such as deep learning and ML, organizations are able to process and analyze large sets of data, thereby generating valuable insights and information that are crucial for informed decision-making (Pisoni et. al., 2023).

In study Anshari et. al., (2023), identifies a gap in the application of ML in KM in the business management sector, highlighting a need for more research. The principal conclusion is that KM systems need to utilize ML to transform extensive data into valuable assets for the organization, supporting advanced decision-making processes. Technology such as cloud computing, ML, and statistical models significantly supports the dependency on Big Data. In its essence, it increasingly relies on human qualities. Therefore, human knowledge forms the foundation for KM and Big Data, which are crucial elements in data analysis (Schaefer & Makatsaria, 2021). Meanwhile, AI is a technology that provides opportunities for computers, machines, and various statistical tools to engineer applications that replicate human skills. In the application of AI, there are three processes that can undergo self-learning activities. The first is Machine Learning (ML), followed by machine intelligence, and finally machine consciousness. In this context, ML is one of the advantages of AI, where the machine can learn on its own (Tangwannawit & Tangwannawit, 2022). The conventional ML approach is developed based on several assumptions, including the belief that the dataset can be fully stored in memory. Unfortunately, some of these assumptions no longer align with the current context, along with the characteristics of Big Data, posing challenges for conventional techniques (L'Heureux et al., 2017).

Various research initiatives have incorporated the use of ML, including study on air pollution. In this context, it is applied to predict emission levels and conduct comparative analysis (Simu et al., 2020). The issue of air pollution has been considered serious in various parts of the world. Almost every major city in every country has been affected by air pollution in recent years, including Jakarta, Indonesia. Jakarta serves as the capital of Indonesia and is a central hub for economic, political, and other aspects. Being the capital of a country, there is a significant amount of activity taking place in Jakarta, and the high level of activity contributes to an increased number of vehicles, leading to elevated air pollution amount (Anggraini et al., 2022). The study of Yarragunta et al. (2021), analyzing the pollutant of air employs a ML approach, which utilizes tools or devices and sensors for learning actions. Among these factors, the study has been conducted to understand and predict the Air Quality Index (AQI) using adaptive opportunities of ML algorithms. It applies various strategies to predict the AQI using supervised ML. The use of supervised ML is employed for dataset analysis and collecting diverse information with the aim of analyzing the AQI through forecasting the optimal outcomes of different ML models evaluated on accuracy levels. Another study Aini and Mustafa (2020), the study employed the K-Nearest Neighbour (KNN) algorithm to forecast air contamination in the city of Makassar. Study achieved the rate of accuracy 96%, score of precision 97% and score of record 100%, relying on a dataset comprising 646 rows of data.

According to Alamsyah and Salma (2018), their study has been conducted to validate the best prediction model using three well-known classification algorithms: Naïve Bayes (NB),

Decision Tree (DT), and Random Forest (RF). The study utilized an employee churn dataset covering two years from 2015 to 2017 with a dataset allocation of 70% for total 11,655 samples of training data and 30% for total 4,994 samples of testing data. The study's results revealed that the best classification model was achieved by RF with the rate of accuracy 97.5%, the runner-up among the subsequent classification models was naive bayes, with an accuracy rate of 96.6%, and the lowest accuracy was attained by the decision tree classification model with a precision rate of 88.7%. Referring to Anggraini et al. (2022), forecasting to predict air pollution quality to the five parameters (pm10, so2, co, o3, and no2) using Artificial Neural Network (ANN) method. The study's results indicate varying Root Mean Square Error (RMSE) values across the analyzed pollutants: pm10 is 9.477, so2 is 5.474, co is 8.392, o3 is 18.250, and no2 is 5.171. Based on previous study findings, the RF classification model has demonstrated remarkable effectiveness in anticipating employee turnover (Alamsyah & Salma, 2018). In this study, data sourced from Jakarta Open Data spanning from 2019 to 2021, as illustrated in Figure 1. The figure emphasizes the utilization of the ISPU (Air Pollution Standard Index) keyword. Consequently, the main aim of this research is to reevaluate the accuracy levels of three predictive models NB, DT, and RF by leveraging air pollution data as the primary dataset.

To achieve this goal, the dataset is divided into two subsets: 80% for training purposes and 20% for testing. The training data is harnessed to familiarize the system with specific algorithms, enabling it to absorb insights from the data. Following this, the testing dataset is employed to introduce novel inputs to the system, thereby evaluating its precision and efficacy. This testing phase holds significant importance as it verifies the model's capability to apply learned knowledge to novel or previously unseen data (Ameer et al., 2019; Simu et al., 2020; Yarragunta et al., 2021; Gupta et al., 2023). Through this study endeavour, it is expected to identify the most accurate predictive model for forecasting employee turnover using air pollution data, while also contributing valuable insights for future research endeavours in this field. KM and ML can reinforce each other. KM can help improve data quality, increase model accuracy, and enhance understanding and decision-making related to air pollution prediction using ML.

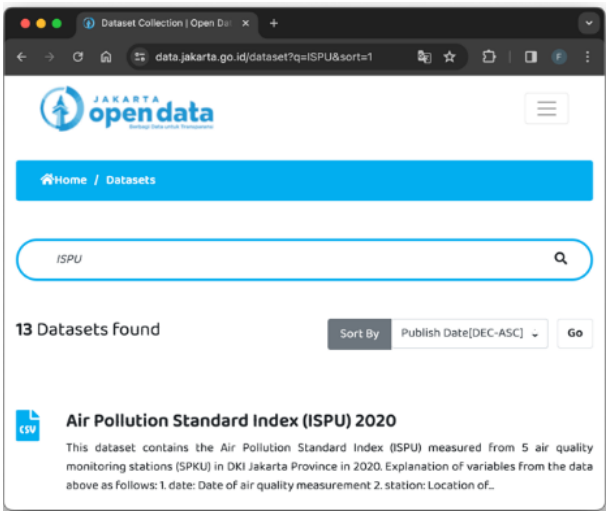

Fig. 1. Jakarta Open Data Website

2. LITERATURE REVIEW

The previous study Alamsyah and Salma (2018) explains that a comparative study provides an overview resulting in the accuracy level of three prediction models that have been conducted. This study presenting literature review of the approach employed in this study. Following that, a concise summary is provided on the ML approach techniques using air pollution data, certainly with the integration of KM presented in the framework KM model.

2.1. Framework KM model

This study utilizes the KM Model framework as the foundation for a more in-depth analysis (Schaefer & Makatsaria, 2021), as illustrated in Figure 2.

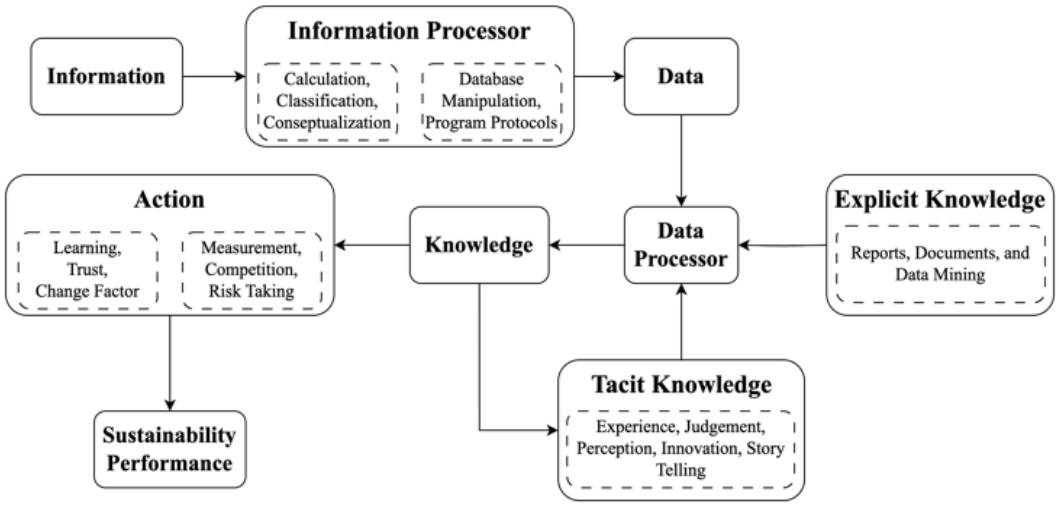

Fig. 2. Framework KM model (Schaefer & Makatsaria, 2021)

Information, initially in a raw form, undergoes a process of organization and management through information processors. Subsequently, this information is transformed into structured data, with data processors playing a role in collecting, organizing, and processing it. The outcome of this data processing forms deep and comprehensible knowledge. This knowledge serves as the foundation for targeted decision-making, contributing to the implementation of specific policies or actions. Through this series of processes, organizations can achieve sustainable performance by assessing its positive impact on long-term sustainability goals, creating an information cycle that supports sustainable actions.

2.2. Air Quality Index (AQI)

The Standard AQI is a dimensionless number that depicts quality of air ambient conditions at specific site. Its foundation lies in the effects on human well-being, visual qualities, and various life forms, as shown in Table 1.

Tab. 1. AQI index (Minister of Environment and Forestry, 2020)

Category	Index Number
Good	1 - 50
Moderate	51 - 100
Unhealthy	101 - 200
Very Unhealthy	201 - 300
Dangerous	>300

As shown above, Table 1 categorizes a range of index numbers into five distinct health-related classifications. The category "Good" corresponds to index numbers between 1 and 50, indicating a safe and healthy status. "Moderate" covers index numbers from 51 to 100, suggesting a fair but acceptable condition. The "Unhealthy" classification spans index numbers from 101 to 200, reflecting a level that may start to impact health. "Very Unhealthy" is the label given to index numbers ranging from 201 to 300, implying a serious concern for health effects. Lastly, any index number above 300 falls into the "Dangerous" category, signifying a potentially hazardous situation that could pose an emergency health risk. The AQI is like a scorecard for air pollution. The higher the number, the worse the air quality and the more danger it poses to our health. Being able to predict and understand changes in air quality is important for making informed decisions about our activities (Imam et al., 2024).

2.3. Machine Learning (ML)

Machine Learning encompasses a range of algorithms that can be broadly classified into three categories: supervised, semi-supervised, and unsupervised learning. In supervised learning, the training data must include labeled examples. These labels guide the algorithm in learning how to make predictions or categorize new, unseen data based on the patterns it has learned from the labeled training data. Semi-supervised and unsupervised learning, on the other hand, deal with partially labeled or unlabeled data, respectively (Somashekar & Boraiah, 2023). Machine Learning has attracted substantial focus across industries, spanning from emerging startups to influential platform providers. On a daily basis, the quantified aspects of air quality exceed the ideal values, presuming appropriate public interventions. Rather than merely issuing conventional commands, the AI theory, emphasizing machines autonomously making decisions, permeates various aspects of our community (Yarragunta et al., 2021). Machine learning is getting really important. It's part of artificial intelligence where computers use algorithms and statistics to learn on their own. Meanwhile, machines can predict things or make choices without needing step-by-step instructions (Barid et al., 2024).

2.3.1. Naïve Bayes (NB)

Naïve Bayes is the statistical classification algorithm commonly used for predicting the likelihood of belonging to a specific class. It relies on Bayesian theorem, sharing similarities with decision trees and neural networks. Its classification effectiveness has been demonstrated through high accuracy fast, speed, and reliable algorithm, particularly when handling large datasets. To define the NB technique, it's crucial to comprehend the

categorization procedure necessitates set indicators for ascertain the suitable category for the examined sample (Alamsyah & Salma, 2018; Aini & Mustafa, 2020; Yarragunta et al., 2021; Tangwannawit & Tangwannawit, 2022; Elvin, 2024). On the other hand, NB algorithm is a probabilistic classifier. It works by making simple calculations with probabilities, how often features appear, and how different values in the data are combined. A key assumption is that features within a class are independent (Afdhaluzzikri et al., 2022). The NB algorithm is a type of machine learning where the computer learns from labeled examples. It's used to solve problems where things need to be sorted into categories, and it's based on a statistical idea called Bayes' theorem. It works especially well for classifying text when there are lots of different features to consider. The NB is popular because it's simple, effective and helps build fast prediction models (Imam et al., 2024).

2.3.2. Decision Tree (DT)

Decision Tree is a probability method that aids in decision-making with relevance to various issues. This algorithm utilizes supervised learning to address classification and regression problems. In previous study, the method of DT used for both classification and predictions, is structured with leaf, root, and decision nodes (Alamsyah & Salma, 2018; Krishna et al., 2023; Simu et al., 2020; Benifa et al., 2022; Aram et al., 2024; Elvin, 2024; Yarragunta et al., 2021; Kang et al., 2018). Positioned at the top are the root nodes, while the leaf nodes are in the middle and decision nodes at the bottom. This focuses on identifying features that aid in classification and prediction. Decision Tree model recommended as potential approach for prediction. However, DT, which is also a tree modeling, has each branch node representing a choice among several alternatives, and at each leaf node representation, a decision is symbolically represented, as illustrated in Figure 3.

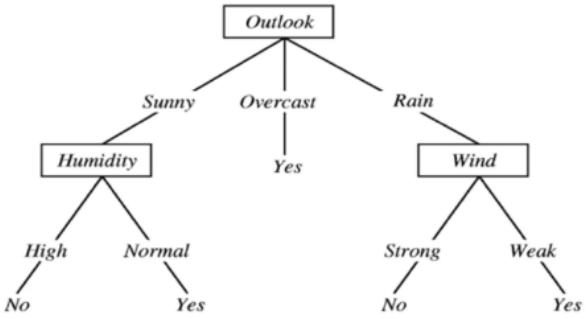

Fig. 3. Decision tree (Kang et al., 2018)

2.3.3. Random Forest (RF)

The RF algorithm, employed in supervised learning, is versatile for both classification and regression tasks. By utilizing averaging methods, the RF algorithm enhances prediction accuracy and helps mitigate the risk of overfitting (Elvin, 2024). Random Forest is a collection of trees built independently using different bootstrap samples from a dataset. It also constructs several decision trees based on the data used, for example, by sampling subsets of various attributes. In RF, predictions are made by taking the consensus decision. In the RF, each node is split using the optimal splitter chosen from a subset of predictors. At

every node, random predictors are utilized, and this element of randomness offers overfit protection (Alamsyah & Salma, 2018; Schonlau & Zou, 2020; Yarragunta et al., 2021; Hai et al., 2022; Benifa et al., 2022; Ravindiran et al., 2023; Baladjay et al., 2023; Gupta et al., 2023; Elvin, 2024; Aram et al., 2024). When presented with new data, each DT makes its own prediction. For classification, the RF final prediction is based on the majority vote of the trees. For regression, the RF averages the predictions of each tree. The power of RF lies in this randomness, as it prevents overfitting and makes the model more adaptable (Liu et al., 2023). However, a RF is a decision built based on subsets of data from multiple trees and utilizes aggregation as the final prediction, as illustrated in Figure 4.

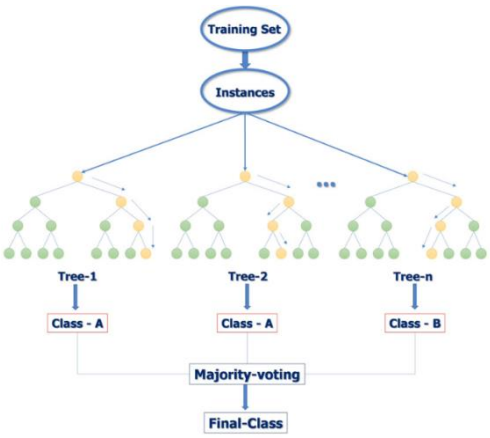

Fig. 4. Random Forest (Hai et al., 2022)

2.4. Evaluation Measurement

This model for classification computes four distinct measures that are extracted. Assess of these models is performed through the application of the confusion matrix, as shown in Table 2.

Tab. 2. Confusion matrix (Alamsyah & Salma, 2018)

		Predicted	
		Positive	**Negative**
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (FN)

Comprehensive explanations are provided:

1. TP : True Positive is positive score predicted correctly.
2. FN : False Negative is positive score predicted as negative.
3. FP : False Positive is negative score predicted as positive.
4. TN : True Negative is negative score predicted correctly.

There are several key parameters to evaluate classification models, such as accuracy, precision, recall, and F1-score (Alamsyah & Salma, 2018). These parameters have effectively yielded the performance outcomes for each predictive model, are shown in Table 3.

Tab. 3. Parameter measurement (Baladjay et al., 2023)

Parameter	Description
Accuracy	Accuracy classification score
Precision	Ratio of measurement o how appropriate the model in predicting the class
Recall	The proportion of positives that are correctly identified
F1-score	Weighted average of the precision and recall

Based on these parameters, it produces numbers calculated by the confusion matrix. These numbers include TP, FP, FN and TN. The explanation is provided below:

1. Accuracy is the score of all the correct prediction, as equation 1:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

2. Precision is the allocation of predicted category which are correct, as equation 2:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

3. Recall is the ratio real category which are correctly identified, as equation 3:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

4. F1-score is the average of a precision, as equation 4:

$$F1 - score = \left(\frac{2}{precision^{-1}+recall^{-1}}\right) = 2 \cdot \left(\frac{precision*call}{precision+recall}\right) \quad (4)$$

3. RESEARCH METHODOLOGY

This study is divided into 5 processes. First one is data collection from Jakarta open data website. Second process is preprocessing the data to involve attributes of selection. The third process is data construction tools, which uses three model classifications, NB, DT and RF. Then, the fourth process is evaluation measurement, and the last process is accuracy. For data processing, this study analyzes using the python languages and google colab as the tool, as illustrated in Figure 5.

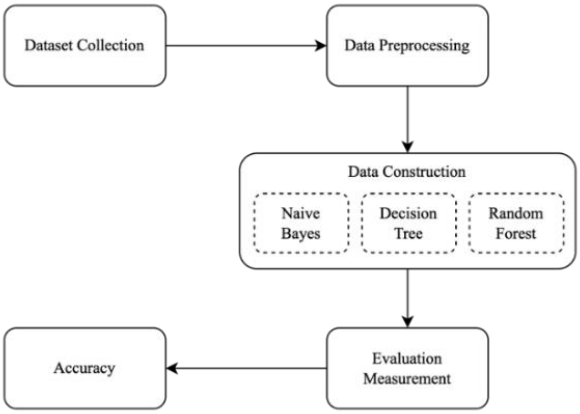

Fig. 5. Research stages

As shown in Figure 5, information, initially in unprocessed form, passes through a series of organization and management processes via information processors. This process involves transforming the information into structured data, with data processors playing a crucial role in gathering, organizing, and processing it. The outcome of this data processing forms profound and comprehensible knowledge, which then serves as the foundation for targeted decision-making. This knowledge contributes to the implementation of specific policies or actions.

In the context of the described information process, its relationship with a series of stages related to dataset collection, data preprocessing, data construction, evaluation measurement, and accuracy can be explained as an integrated information cycle. Starting with dataset collection as the initial step in the information cycle, followed by the preprocessing stage that enhances data effectiveness. Afterward, data is reconstructed into a more structured form, ready for further processing. The evaluation measurement stage is then conducted to assess the results of data processing and data construction. Finally, the measure of accuracy as primary assessment metric to determine how well information or model aligns with specific expectations or goals.

3.1. Dataset Collection

The dataset of air pollution contained 10 related attributes from air pollution information. This dataset serves as the initial foundation for subsequent comparative analysis in this study, as shown in the Table 4.

Tab. 4. Dataset attributes before preprocessing

No	Attribute name	Description
1	Date	Date of air quality measurement
2	Location	Measurement station location
3	pm10	Particulate matter measurement
4	so2	Sulfur dioxide measurement
5	co	Carbon monoxide measurement
6	o3	Ozone measurement
7	no2	Nitrogen dioxide measurement
8	Max	Highest value among all pollutant measured simultaneously
9	Critical	Highest measured parameter
10	Category	Air pollution standard index category calculation

The dataset includes the following attributes: date, location, pm10, so2, co, o3, no2, max, critical, and category. Pm10, so2, co, o3, and no2 represent the monitored pollutants. Data are collected daily to anticipate the behavior of contaminants in the upcoming days.

3.2. Data Preprocessing

Pre-processing is mandatory to make the data ready for processing, reducing data noise as well. In this phase of the study, data were obtained in the form of monthly data over a period of 3 years. Previous study (Anggraini et al., 2022) selected 5 parameters to perform data normalization or transformation, the parameter being pm10, s02, co, o3, and no2. On

the other hand, this study eliminated 4 attributes such as date, location, max, and critical, leaving 6 attributes that are relevant to this study, as shown in Table 5.

Tab. 5. Dataset attributes after preprocessing

No	Attribute Name	Description
1	pm10	Particulate matter measurement
2	so2	Sulfur dioxide measurement
3	co	Carbon monoxide measurement
4	o3	Ozone measurement
5	no2	Nitrogen dioxide measurement
6	Category	Air pollution standard index category calculation

3.3. Data Construction

After conducting data preprocessing, this study utilizes a NB, DT, and RF approach for data construction. To perform this data construction, training is conducted using a preprocessed air pollution dataset referred to as the training data. Later, testing is also carried out using another preprocessed air pollution dataset, referred to as the testing data. The allocation of the training and testing data are set at 80% for training and 20% for testing.

3.4. Evaluation Measurement

The upcoming process is evaluation measurement, which represents the accuracy score of each model from preprocessed data. This is then reevaluated for greater accuracy. This study incorporates KM in the form of Table 6.

Tab. 6. Parameter comparison

Parameter	NB	DT	RF
Accuracy	0,81	0,93	0,94
Precision	0,77	0,93	0,95
Recall	0,77	0,92	0,93
F1-score	0,74	0,93	0,94

3.5. Accuracy

The final process is accuracy, which serves as the ultimate outcome of this study. Accuracy in this context, is the final score obtained from each prediction model, such as NB, DT, and RF. The result is to determine the accuracy of the comparison among the prediction models in the form of Table 7, presenting percentages to facilitate a clearer understanding of this study.

Tab. 7. Accuracy confusion matrix

	NB	DT	RF
True Positive (TP)	36%	78%	78%
False Negative (FN)	64%	22%	22%
False Positive (FP)	72%	81%	80%
True Negative (TN)	28%	19%	20%

4. RESULT

The results of this study reveal an examination of the accuracy of the implemented ML models using NB, DT, and RF. Detailed explanations in several sections, providing an overview of the performance and accuracy of the predictive models utilized in this study.

4.1. Dataset Allocation

The air pollution dataset was split into training and testing subsets, with 80% designated for training and the remaining 20% set aside for testing, as shown in Table 6.

Tab. 6. Dataset allocation

Training/Testing	Proportion	Number of Samples
Training	80%	4,233
Testing	20%	1,059

Total dataset of this study is 5,292 with 80% of 4,233 samples data for training and 20 % of 1,059 for data testing.

4.2. Confusion Matrix

After obtaining results from the training and testing samples, this study proceeded with the confusion matrix process. The results of the confusion matrix scores generated by NB, DT, and RF forest, as illustrate in Figure 6.

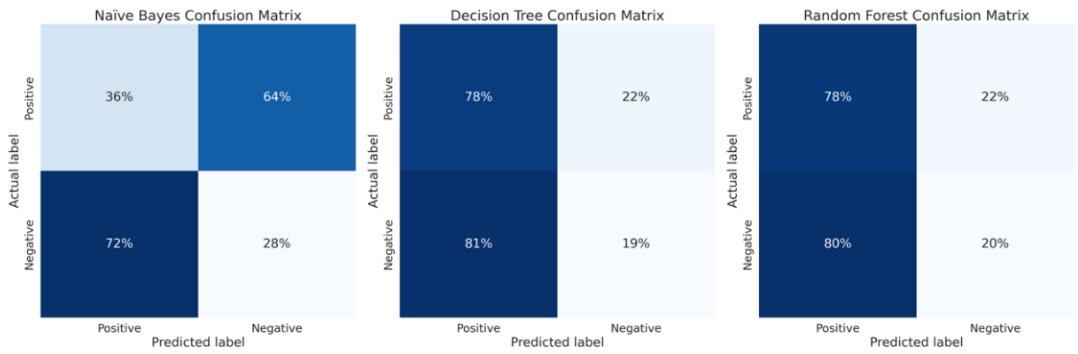

Fig. 6. Confusion matrix output

For the NB model, 36% of cases were correctly classified as positive (True Positive), while 64% were incorrect classified as negative (False Negative). Additionally, 72% were incorrectly labeled as positive when they were negative (False Positive), and 28% were accurately identified as negative (True Negative). In the case of the DT model, 78% of instances were correctly predicted as positive, with 22% being erroneously classified as negative. Furthermore, 81% were incorrectly labeled as positive, and 19% were correctly identified as negative. For the RF model, the results indicate a 78% accuracy in identifying positive cases, with a 22% misclassification rate as negative. Additionally, 80% were false positives, and 20% were true negatives. The NB model while known for speed and handling

large datasets, showed moderate accuracy but struggled with precision, possibly due to its feature independence assumption not fully aligning within the dataset. Decision Tree achieved the highest accuracy, but its low precision suggests overfitting, a known issue with this type of model. The RF is designed to combat overfitting, exhibited slightly better precision than DT, though accuracy remained similar, indicating some improvement but continued challenges with false positives.

4.3. Cross-validation

Additionally, a thorough model evaluation should be conducted using the metrics outlined in Table 2, ensuring the identification of the most proficient model in terms of performance, as illustrated in Figure 7.

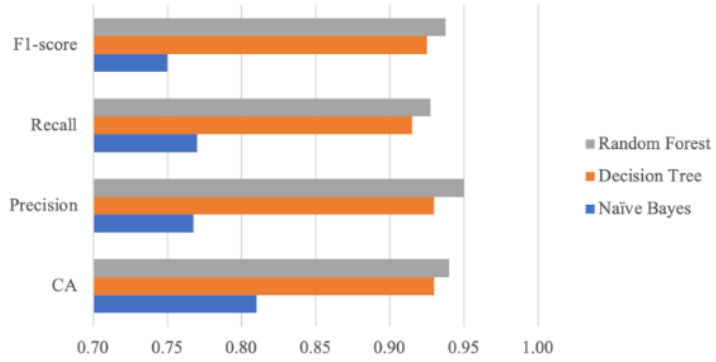

Fig. 7. Cross-validation output

Based on Figure 7, the RF algorithm consistently outperforms the NB and DT across all evaluated metrics. Boasting a classification accuracy (CA) of 94%, RF shows the highest probability of correct instance classification. This is further reinforced by a precision rate of 95%, indicating that when this model predicts an instance to be of the positive class, there is a 95% chance of it being actually positive. In terms of recall, the RF records a figure of 93%, meaning it successfully identifies 93% of all actual positive instances. The highest F1-score for RF, at 94%, affirms that this model maintains an excellent balance between precision and recall. On the other hand, while the DT is not far behind with impressive scores across all metrics, the NB appears to be the model with the lowest performance in this comparison, with CA, precision, recall, and F1-score values at 81%, 77%, 77%, and 75%, respectively. In conclusion, the RF stands out as the algorithm with the most stable and high performance for the analyzed dataset. The finding of the cross-validation output is demonstrate the expected strengths and weaknesses of RF, DT, and NB algorithms. RF outperforms DT and NB in f1-score, recall, precision, and CA due to its ensemble nature, which reduces overfitting. While less interpretable and potentially more computationally expensive than NB, RF accuracy often justifies these tradeoffs. DT performs competitively but can be prone to overfitting, making RF often preferable. NB, despite its simplicity and speed, suffers when its assumption of feature independence is violated, leading to its lower performance in this case.

4.4. Evaluation

The summarized outcomes resulting from the implementation of three predictive models in this study. The spatial distribution of these markers across the plot not only highlights the absolute numbers but also paints a comparative narrative of model efficiency, allowing this study to digest at a glance the efficacy of each model in its predictive capabilities.as illustrated in Figure 8.

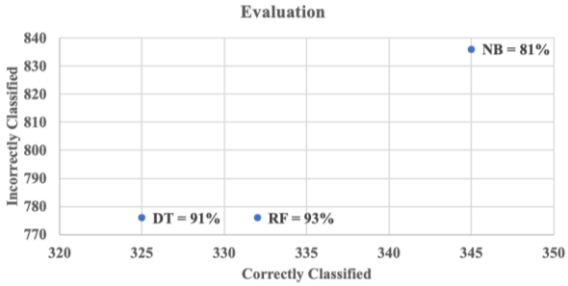

Fig. 8. Evaluation

Based on Figure 8, this study compares the performance of the three ML models, the NB model, DT and RF, based on their classification accuracy using cross-validation. The NB model calculates correctly classified 345 cases and incorrectly classified 836 cases, resulting in an accuracy of 81%. The NB may be less effective in dealing with the complexity of this dataset, which is a disadvantage, although its simplicity and speed of model learning and prediction remain advantages. Meanwhile, the DT model shows improvement, calculate correctly classifying 325 cases and incorrectly classified 776, achieving a higher accuracy of 93%. This indicates its better capability in mapping data relationships an advantage over NB, but there's a disadvantage in the potential for overfitting. In the other hand, the RF model achieve the best performance, calculate correctly classifying 332 cases and incorrectly classified 776, but achieving the highest accuracy among the three at 94%. This underscores the effectiveness of RF, particularly in dealing with complex data, which is a significant advantage. However, the primary disadvantage of the RF model is its greater computational complexity and the potential for requiring more resources, a consequence of its methodology that involves aggregating multiple DT. The findings of this evaluation demonstrate that RF achieves the highest accuracy among the tested algorithms, at 93%. This is followed by DT at 91% and NB at 81%. These results align with existing literature. While NB offers simplicity and speed, its accuracy may be lower than algorithms like RF, especially when the dataset violates its assumptions about feature independence. Decision Tree, though effective for classification, has a tendency to overfit the data. RF is superior performance underscores its effectiveness in both classification and regression tasks while mitigating the risk of overfitting.

5. CONCLUSION

This study introduces a comparative analysis of prediction models for air pollution in Jakarta, Indonesia. With using KM, this study has successfully compared three models to

predict air pollution. Evaluation measurements, including a confusion matrix are utilized to assess the performance of these ML models. In terms of classification accuracy using cross-validation on an air pollution dataset. By allocating 80% of the data for training and 20% for testing with a total of 5,292 data in the dataset, this study established a solid foundation for model evaluation.

It was observed that while NB had difficulties in accurately classifying positive and negative cases, DT and RF performed better in identifying positive cases, though they faced challenges with a high incidence of false positives. A more in-depth cross-validation revealed that RF achieved accuracy of 94%, closely followed by DT at 93%, and NB lagging at 81%. Despite its speed and simplicity, effectiveness of NB was limited in handling the dataset's complexity, as indicated by its lower accuracy. Decision Tree showed improved accuracy and was adept at mapping data relationships, yet there was a concern about it is tendency to overfit. RF, however, proved to be the most efficient model, adeptly managing complex data but requiring more computational resources. The conclusion of the study is that RF is the most suitable model for this dataset, offering high accuracy and robust data analysis capabilities. Further study to derive results for various scenarios from more diverse models, combine the RF models with other models in an ensemble to potentially boost accuracy further, and further study should also consider employing data analysis techniques to enhance the understanding and prediction of air pollution.

Author Contributions

Siti ROHAJAWATI – the main research, knowledge management design,
Hutanti SETYODEWI – research planning and progress monitoring,
Ferryansyah M. A. TRESNANTO, - data analysis using ML
Debora MARIANTHI, data visualisation using ML
Maruli T. B. SIHOTANG – content correction, consistency of content.

Acknowledgement

The authors gratefully acknowledge to Bakrie University due to institutional research fund. Special thanks to Agus Masdar who helpfully collected data on Jakarta Provincial Government, and to all for supporting this research.

REFERENCES

Afdhaluzzikri, A., Mawengkang, H., & Sitompul, O. S. (2022). Perfomance analysis of Naive Bayes method with data weighting. *Sinkron : Jurnal Dan Penelitian Teknik Informatika*, *6*(3), 817-821. https://doi.org/10.33395/sinkron.v7i3.11516

Aini, N., & Mustafa, M. S. (2020). Data mining approach to predict air pollution in makassar. *2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS)* (pp. 1-5). IEEE. https://doi.org/10.1109/ICORIS50180.2020.9320800

Alamsyah, A., & Salma, N. (2018). A comparative study of employee churn prediction model. *4th International Conference on Science and Technology (ICST)* (pp. 3–6). IEEE. https://doi.org/10.1109/ICSTC.2018.8528586

Ameer, S., Shah, M. A., Khan, A., Song, H., Maple, C., Islam, S. U., & Asghar, M. N. (2019). Comparative analysis of machine learning techniques for predicting air quality in smart cities. *IEEE Access*, *7*, 128325-128338. https://doi.org/10.1109/ACCESS.2019.2925082

Anggraini, A. N., Ummah, N. K., Fatmasari, Y., & Hayati Holle, K. F. (2022). Air quality forecasting in DKI Jakarta using Artificial Neural Network. *MATICS: Jurnal Ilmu Komputer dan Teknologi Informasi (Journal of Computer Science and Information Technology)*, *14*(1), 1-5. https://doi.org/10.18860/mat.v14i1.13863

Anshari, M., Syafrudin, M., Tan, A., Fitriyani, N. L., & Alas, Y. (2023). Optimisation of knowledge management (KM) with Machine Learning (ML) enabled. *Information*, *14*(1), 35. https://doi.org/10.3390/info14010035

Aram, S. A., Nketiah, E. A., Saalidong, B. M., Wang, H., Afitiri, A.-R., Akoto, A. B., & Lartey, P. O. (2024). Machine learning-based prediction of air quality index and air quality grade: a comparative analysis. *International Journal of Environmental Science and Technology*, *21*, 1345-1360. https://doi.org/10.1007/s13762-023-05016-2

Barid, A. J., Hadiyanto, H., & Wibowo, A. (2024). Optimization of the algorithms use ensemble and synthetic minority oversampling technique for air quality classification. *Indonesian Journal of Electrical Engineering and Computer Science*, *33*(3), 1632–1640. https://doi.org/10.11591/ijeecs.v33.i3.pp1632-1640

Benifa, J. V. B., Kumar, P. D., & Rose, J. B. R. (2022). Prediction of air quality index using machine learning techniques and the study of its influence on the health hazards at urban environment. In M. Lahby, A. Al-Fuqaha, & Y. Maleh (Eds.), *Computational intelligence techniques for green smart cities* (pp. 249-269). Springer International Publishing. https://doi.org/10.1007/978-3-030-96429-0_12

Bilquise, G., & Shaalan, K. (2022). AI-based academic advising framework: A knowledge management perspective. *International Journal of Advanced Computer Science and Applications*, *13*(8). https://doi.org/10.14569/IJACSA.2022.0130823

Elvin, W. A. (2024). Forecasting water quality through machine learning and hyperparameter optimization. *Indonesian Journal of Electrical Engineering and Computer Science*, *33*(1), 496-506. https://doi.org/10.11591/ijeecs.v33.i1.pp496-506

Minister of Environment and Forestry. (2020). *Regulation Number 14 of 2020 concerning Air Pollution Standard Index.* https://peraturan.bpk.go.id/Download/156214/Permen%20LHK%20Nomor%2014%20Tahun%202020.pdf

Gupta, N. S., Mohta, Y., Heda, K., Armaan, R., Valarmathi, B., & Arulkumaran, G. (2023). Prediction of air quality index using Machine Learning techniques: A comparative analysis. *Journal of Environmental and Public Health, 2023*, 4916267. https://doi.org/10.1155/2023/4916267

Hai, P. M., Tinh, P. H., Son, N. P., Van Thuy, T., Hanh, N. T. H., Sharma, S., Hoai, D. T., & Duy, V. C. (2022). Mangrove health assessment using spatial metrics and multi-temporal remote sensing data. *PLoS ONE*, *17*(12), e0275928. https://doi.org/10.1371/journal.pone.0275928

Imam, M., Adam, S., Dev, S., & Nesa, N. (2024). Air quality monitoring using statistical learning models for sustainable environment. *Intelligent Systems with Applications*, *22*, 200333. https://doi.org/10.1016/j.iswa.2024.200333

Baladjay, J. M., Riva, N., Santos, L. A., Cortez, D. M., Centeno, C., & Sison, A. A. R. (2023). Performance evaluation of random forest algorithm for automating classification of mathematics question items. *World Journal of Advanced Research and Reviews*, *18*(2), 034–043. https://doi.org/10.30574/wjarr.2023.18.2.0762

Kang, G. K., Gao, J. Z., Chiao, S., Lu, S., & Xie, G. (2018). Air quality prediction: Big Data and Machine Learning approaches. *International Journal of Environmental Science and Development*, *9*(1), 8–16. https://doi.org/10.18178/ijesd.2018.9.1.1066

Krishna, V. A., Koganti, H., Madhumathi, M., & Dharani, V. (2023). Air quality prediction using machine learning algorithm. *International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)* (pp. 2201–2206). IEEE. https://doi.org/10.1109/ICSCDS56580.2023.10105063

L'Heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. A. M. (2017). Machine Learning with Big Data: challenges and approaches. *IEEE Access*, *5*, 7776–7797. https://doi.org/10.1109/ACCESS.2017.2696365

Pisoni, G., Molnár, B., & Tarcsi, Á. (2023). Knowledge management and data analysis techniques for data-driven Financial Companies. *Journal of the Knowledge Economy*. https://doi.org/10.1007/s13132-023-01607-z

Ravindiran, G., Hayder, G., Kanagarathinam, K., Alagumalai, A., & Sonne, C. (2023). Air quality prediction by machine learning models: A predictive study on the indian coastal city of Visakhapatnam. *Chemosphere*, *338*, 139518. https://doi.org/10.1016/j.chemosphere.2023.139518

Schaefer, C., & Makatsaria, A. (2021). Framework of data analytics and integrating knowledge management. *International Journal of Intelligent Networks*, *2*, 156–165. https://doi.org/https://doi.org/10.1016/j.ijin.2021.09.004

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *Stata Journal*, *20*(1), 3-29. https://doi.org/10.1177/1536867X20909688

Simu, S., Turkar, V., Martires, R., Asolkar, V., Monteiro, S., Fernandes, V., & Salgaoncary, V. (2020). Air pollution prediction using machine learning. *IEEE Bombay Section Signature Conference (IBSSC)* (pp. 231-236). IEEE. https://doi.org/10.1109/IBSSC51096.2020.9332184

Somashekar, H., & Boraiah, R. (2023). Network intrusion detection and classification using machine learning predictions fusion. *Indonesian Journal of Electrical Engineering and Computer Science*, *31*(2), 1147-1153. https://doi.org/10.11591/ijeecs.v31.i2.pp1147-1153

Taherdoost, H., & Madanchian, M. (2023). Artificial intelligence and knowledge management: Impacts, benefits, and implementation. *Computers*, *12*(4), 72. https://doi.org/10.3390/computers12040072

Tangwannawit, S., & Tangwannawit, P. (2022). An optimization clustering and classification based on artificial intelligence approach for internet of things in agriculture. *IAES International Journal of Artificial Intelligence*, *11*(1), 201–209. https://doi.org/10.11591/ijai.v11.i1.pp201-209

Yarragunta, S., Nabi, M. A., Jeyanthi, P., & Revathy, S. (2021). Prediction of air pollutants using supervised machine learning. *5th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1633-1640). https://doi.org/10.1109/ICICCS51141.2021.9432078