Comparison of Decision Tree and Support Vector Machine for Predicting Jakarta Air Quality Index

1st Musyaffa Ayman Rafif
Data Science Program, School of Computer Science
Bina Nusantara University
Jakarta, Indonesia, 11480
musyaffa.rafif@binus.ac.id

2nd Gede Sanjaya Indrajaya
Data Science Program, School of Computer Science
Bina Nusantara University
Jakarta, Indonesia, 11480
gede.indrajaya@binus.ac.id

3rd Muhammad Kent Al-Ghazi
Data Science Program, School of Computer Science
Bina Nusantara University
Jakarta, Indonesia, 11480
muhammad.alghazi@binus.ac.id

4th Johnny
Data Science Program, School of Computer Sciences
Bina Nusantara University
Jakarta, Indonesia, 11480
johnny@binus.ac.id

5th Noviyanti Tri Sagala
Statistics Department, School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
noviyanti.sagala@binus.edu

***Abstract*— The people in the capital city of the Republic of Indonesia, Jakarta, have been living for years with air pollution. Air quality has been a concern for a long time due to the health risks it has on people, especially those at risk. Using classification algorithms, we would like to implement data mining for the air quality index and make predictions from our capital city, DKI Jakarta. The objective of this research is to predict the upcoming air quality in Jakarta regions, and we will use a dataset from Jakarta Open Data Website. We will be focusing on the air quality index. Some methods can be implemented to work on this dataset. However, the research shown in this paper uses Decision Trees and SVM (Support Vector Machine) to make a model of the data. We use Grid search CV to evaluate the model and give the best accuracy of 91.45%.**

Keywords*— *Decision Tree, Jakarta Air Quality, Grid Search CV, Predictive model, Support Vector Machines.

I. Introduction

The capital of Indonesia, Jakarta, is divided into five districts: Central Jakarta, West Jakarta, East Jakarta, South Jakarta, and North Jakarta. The growth of the population in Jakarta is increased rapidly. As the population increased, the number of vehicles also climbed. Along with the density of motor vehicles, infrastructural development is also accelerating. As a result, there is more air pollution, including various pollutants from burning manufacturing smoke, car exhaust, and other pollutants from the impact of infrastructure development.

According to IQAir, a Swiss-based air filtration company that documents air quality worldwide, Jakarta has reached the highest index for the worst air quality in the world. DKI Jakarta has the worst air quality in the world, with a score of 163 on the air quality index; it was declared on June 22, 2022. [1]. Unhealthy air quality can harm and even complicate the health of people with breathing illnesses [2]. Another effect can be shown with the name "sick building syndrome" (SBS).

A collection of mucosal, cutaneous, and general symptoms temporally connected to working in specific buildings make up the sick building syndrome (SBS) [3]. According to a study conducted on 350 employees from 18 offices in Jakarta over six months (July–December 2008), SBS tends to affect 50% of office workers. Building equipment/materials, outdoor pollution, microorganisms, building materials/office equipment, and poor ventilation are the leading causes of indoor air pollution in buildings [4]. The symptoms include rhinitis, coughing, shortness of breath, irritation of the eyes and nasopharynx, headaches, mucous membrane irritation, and others. However, neither the disease nor the origin of these symptoms is known with certainty.

Many studies have been conducted to examine the air quality in Jakarta, and those studies need to use datasets and algorithmic methods to analyze the data. Some studies use data from measuring stations scattered in Jakarta, while others use satellites or mathematical models to predict pollution [5]. The analysis methods are diverse, ranging from simple statistics to complex models that use artificial intelligence. This research gives us important information on air quality in Jakarta and can be used to effectively develop a strategy for controlling air pollution.

The Naive Bayes algorithm is used in previous work to classify the simple probabilistic that counts a cluster of probabilities by summing the frequency and combination from the value of the given dataset [4]. KNN (K-nearest neighbor) is also used to classify new objects based on attributes and training samples. The Naive Bayes and KNN methods show accuracies of up to 89.22 percent and 94.98, respectively [6]. This research proposes air quality forecasting using an artificial neural network to predict air quality in years to come [7]. The artificial neural network (ANN) is one of the artificial representations of the human brain that always tries to simulate the learning process in the human brain [8]. In this study, we used a dataset from Jakarta's open data. This data was taken every day from 2012 until 2020 from five air observer stations in DKI Jakarta. This dataset contains several pollutants with their

index, worst index point on the critical side, and air quality on that day [7].

After testing the data on air quality forecasting in the Jakarta region using the Neural Network, it was found that the setting of learning rate 0.1, momentum 0.2, and neuron 12 resulted in a higher accuracy rate of 88.86% and an error rate of 0.320, which was lower than other testing parameters. The predicted and actual results produced the same value, which means that the method of the Neural Network can correctly predict the class for the next day [7].

From the previous work, the Jakarta air quality data for 2020 has been used for developing different predictive models. The method used was KNN (K-Nearest Neighbor). Our research aims to compare the performance of various machine learning models in predicting air quality. Specifically, we proposed support vector machines (SVM) and decision trees to build predictive models. Furthermore, the parameters of these algorithms were also tuned to find the hyperparameters and obtain the best possible model for this dataset.

II. METHODOLOGY

A. Research Scheme

This research classifies ISPU levels to determine air quality in DKI Jakarta using the Data Mining (DM) methodology of Knowledge Discovery in Databases (KDD). KDD is a structured method for extracting meaningful patterns from vast and complicated data sets that are valid, novel, useful, and understandable. Data mining is the core of the KDD process, involving the inference of algorithms that explore the data, develop the model, and discover previously unknown patterns [9]. The model is used to analyze, predict, and interpret phenomena from the data. This methodology is used to analyze and classify the ISPU levels using data collected from various air quality monitoring stations in DKI Jakarta, which can be researched further and allow us to predict the incoming data and patterns that might be generated. The research methodology can be seen in Figure 1.

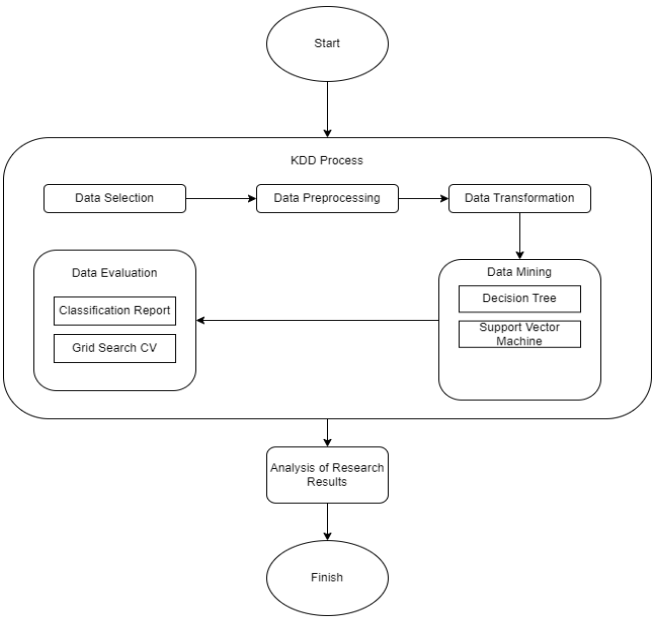

Fig 1. Research flow scheme

The KDD process begins with data selection, data preparation, data transformation, data mining, and data evaluation using the evaluation value of classification report accuracy.

B. Research Object

The object of study in this research is air pollution data from the Standard Air Pollution Index at DKI Jakarta. The study compares classification algorithms based on the data to determine which algorithm is the most accurate in predicting good, moderate, and poor air quality.

C. Data Selection

The Jakarta Standard Air Pollution Index dataset is collected and selected from the open government website of DKI Jakarta (http://www.data.jakarta.go.id/) in the form of CSV files, taken per SPKU and area, namely Central Jakarta (DKI1), North Jakarta (DKI2), South Jakarta (DKI3), East Jakarta (DKI4), and West Jakarta (DKI5) in 2020. The data consists of 'Tanggal,' 'stasiun,' PM10, SO2, CO, O3, NO2, max, critical, and 'categori', as shown in Table I. After combining each dataset into one singular yearly data, the data consists of 1675 rows.

TABLE I. ATTRIBUTES OT THE DATASET

Column	Description
Tanggal (Date)	The date of when the data is collected
Stasiun (Station)	The location of the station of where the data is collected
PM10	Particles in the air that are smaller than 10 micrometers
SO2	The value of sulfide in the air
CO	The value of carbon monoxide in the air
O3	The value of ozone in the air
NO2	The value of nitrogen dioxide in the air
Max	The highest value of the gasses measured at the time
Critical	The parameter which produces the highest value
Categori (Category)	The quality of the air based on the parameters

D. Data Preparation

The data that has been collected and selected will enter the data preparation stage. In this stage, data will be reduced by removing missing values and duplicated data. To handle this, we use resampling with the SMOTE-ENN method to handle imbalanced data. The SMOTE-ENN resampling technique combines oversampling and under sampling. The basic idea of the SMOTE method is to perform linear interpolation between neighboring minority class samples to synthesize new minority class samples, solving the problem of significant data overlap compared with random oversampling [10]. Focusing on majority class samples, the ENN (Edited Nearest Neighbor) algorithm deletes the

sample if there are two or more in the nearest three neighboring samples different from it. However, most samples are near each other, which causes limited sample removal [11]. The technique applies oversampling on the minority class samples and under sampling on the majority class samples. Then, because classification falls under supervised learning, labeling is needed on the attributes that will be classified.

E. Data Transformation

At this stage, the dataset, which is still separate between CSV files, is integrated and adjusted to the existing attributes to become a unified dataset to make it easier in the next process.

F. Data Mining

The core stage of this research is to create a classification model using Decision Tree and Support Vector Machine algorithms. *A Decision Tree* is an algorithm that classifies a population into branch-like segments that construct an inverted tree with a root node, internal nodes, and leaf nodes. The algorithm is non-parametric and can efficiently deal with large, complicated datasets without imposing a complicated parametric structure [12]. Support Vector Machine (SVM) is an algorithm that learns by example to assign labels to objects [13]. Both algorithms are chosen because they can work effectively on a high variance data.

G. Data Evaluation

In this step, we evaluated the two models using a classification report so that we could compare the two models. A *classification report* is a report that provides precision, recall, f1-score, and support scores for each class in a multi-class classification. The report helps evaluate the classification model's performance and identify areas that need improvement [14]. The results of the two classification reports are as shown in Table II.

TABLE II. CLASSIFICATION REPORT ON DECISION TREE AND SVM MODEL

Methods	Accuracy	Precision	Recall	F1
Decision Tree	87.86%	86%	86%	86%
Support Vector Machine	90.56%	90%	92%	90%

From the results of Table II, it can be seen that using Support Vector Machine is better than using Decision Tree from accuracy, precision, recall, and F1 metrics. Accuracy is the ratio of correct predictions to the number of predictions made. Precision is the ratio of the number of true positive predictions to the number of positive predictions. The recall is the ratio of the number of true positive predictions to the number of actual positive cases. F1 is the harmonic mean of precision and recall [15].

After comparing two of the models, we evaluated the model using Grid Search CV to find the best combination of parameters of the model. Grid search is an approach to parameter tuning that will methodically build and evaluate a model for each combination of algorithm parameters specified in a grid [16].

We found the best parameter for the Decision tree from the second parameter tuning algorithm using Grid Search CV, as seen in Table III.

TABLE III. BEST PARAMETERS FOR DECISION TREE USING GRID SEARCH CV

Parameters	Values
‘criterion’	‘gini’
‘max_depth’	3
'min_samples_leaf'	3
‘min_samples_split	10

.

- `Criterion` is a decision tree algorithm parameter that determines a split's quality. It measures the impurity of the input set. Commonly used criteria are "Gini" and "entropy"[17].
- `Max_depth` is a parameter that defines the maximum depth of the tree. It prevents overfitting by stopping the tree from growing too deep [17].
- `Min_samples_leaf` is a parameter that defines the minimum number of samples required to be at a leaf node. It prevents overfitting by ensuring that every leaf node has enough samples [17].
- `Min_samples_split` is a parameter that defines the minimum number of samples required to split an internal node. It prevents overfitting by ensuring that every internal node has enough samples before splitting [17].

The results of the decision tree model using the best parameters from the Grid Search CV can be observed in Table IV.

TABLE IV. CLASSIFICATION REPORT ON TUNED DECISION TREE MODEL

Methods	Accuracy	Precision	Recall	F1
DT with Grid Search CV	90.94%	90%	92%	90%

An improvement has been shown when we use the best parameter from Grid Search CV, and there are changes from all the metrics in the classification report. Furthermore, we use Grid Search CV to find the best parameters for SVM, as seen in Table V.

TABLE V. BEST PARAMETERS FOR SVM USING GRID SEARCH CV

Parameters	Values
‘C’	100
‘kernel’	‘rbf’

`C` and `kernel` are two important parameters in the support vector machine (SVM) algorithm.

- `C` is a parameter that controls the trade-off between maximizing the margin and minimizing the classification error. A lower value of C will result in a wider margin but more misclassification, while a higher value of C will result in a narrower margin but less misclassification [18].
- `Kernel` is a parameter that determines the type of function used to transform the input data into a higher dimensional space where a linear boundary can be found. Commonly used kernels are "linear," "poly," "RBF," and "sigmoid" [18].

The results of the support vector machine model using the best parameters from the Grid Search CV can be observed in Table VI.

TABLE VI. CLASSIFICATION REPORT ON TUNED SVM MODEL

Methods	Accuracy	Precision	Recall	F1
SVM with Grid Search CV	91.40%	90%	92%	90%

Table VI shows only small improvements from the SVM model without optimization with Grid Search CV in accuracy metrics. Regardless, the best model for accuracy is still SVM with the best parameters from Grid Search CV.

III. RESULT ANALYSIS

In this study, we evaluated the performance of four different models: a decision tree model, a support vector machine (SVM) model, and both the decision tree and SVM models using both default parameters and optimized hyper-parameters obtained through grid search cross-validation. We use training data for 80% of the total database. Following the training, each method's performance will be measured based on its respective classification report, although the main value of measurements will be accurate.

It is shown that a decision tree performs the worst in all parameters with 87.86% accuracy, 86% precision, 86% recall, and 86% F1. DT with Grid Search CV, Support Vector Machine and SVM with Grid Search CV have the same value of 90% precision, 92% recall, and 90% F1. The differences between the three methods are in accuracy, where SVM with 90.56%, DT with Grid Search CV with 90.54%, and SVM with Grid Search CV with 91.40%. The results above show that SVM with Grid Search CV has the best performance compared to the other measurement methods.

IV. CONCLUSION

The research we have been working on is important to the people living in the capital of Indonesia, DKI Jakarta. The best model we have worked on is the SVM (Support Vector Machine) that has been tuned with Grid Search CV compared to its counterparts, with a final accuracy of 91.40%. The work on this dataset before has an accuracy of 89.226% using KNN (K-Nearest Neighbor), and it has shown that our research has more accuracy than the previous research. For future work, we will use more algorithms in the research to explore more of the dataset, and if needed, we would like to add another dataset. The research we have been working on is purposed of predicting air quality for the next periods in Jakarta after 2020. After working on two algorithms being compared and tuned using hyperparameter tuning, we conclude that Support Vector Machine with Grid Search CV provides the best accuracy compared to its counterparts.

REFERENCES

[1] Page Title. Kualitas Udara di Jakarta. https://www.iqair.com/id/indonesia/jakarta. 2020.

[2] Dean E. Schraufnagel, MD., etc. Air Pollution and Noncommunicable Diseases. International Respiratory Societies' Environmental Committee, p.2, pp.417-418, 2018.

[3] Burge, P.S. Sick Building Syndrome. Occup Environ Med, vol.61, pp.185-190, 2004.

[4] Ridwan, A.M., etc. Analisis Gejala Sick Building Syndrome Pada Pegawai di Unit OK Rumah Sakit Marinir Cilandak Jakarta Selatan. Jurnal Kesehatan Masyarakat, vol.2, pp.117-119, 2018.

[5] Santoso, M., etc. Multiple Air Quality Monitoring Evidence of the Impacts of Large-scale Social Restrictions during the COVID-19 Pandemic in Jakarta, Indonesia. Aerosol and Air Quality Research, vol.21, pp.1-3, 2021.

[6] Sodiq, M. Ja'far. Perbandingan Metode Naive Bayes dan K-Nearest Neighbor Pada Klasifikasi Kualitas Udara di DKI Jakarta, pp.4-5, 2019.

[7] Kristiyanti, D.A., etc. Implementation of Neural Network Method for Air Quality Forecasting in Jakarta Region. Journal of Physics: Conference Series, pp.1-5, 2020.

[8] Krogh, Anders. What are Artificial Neural Networks?. Nature Biotechnology, vol.26, pp.195-196, 2008.

[9] Maimon, Oded and Lior Rokach. Data Mining and Knowledge Discovery Handbook. Department of Industrial Engineering Tel-Aviv University, cp.1, pp.1-2, 2005.

[10] Chawla, N.V., etc. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, vol.6, pp.321-357, 2002.

[11] Wilson, D.L. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data.IEEE Trans. Syst. Man Cybern, vol.SMC-2, pp. 408-421. 1972.

[12] Song, Yan-yan and Ying Lu. Decision Tree Methods: Applications for Classification and Predictions. Shanghai Arch Psychiatry, vol.27(2), pp.130-135, 2015.

[13] Boser, B.E., etc. A Training Algorithm for OptimalMargin Classifiers. In 5th Annual ACM Workshop on COLT (ed. Hausslee, D.), pp.144-152, 1992.

[14] Müller, Andreas and Sarah Guido. Introduction to Machine Learning with Python. O'Reilly, pp. 220-223, 2016.

[15] Manning, C.D., Prabhakar Raghavan, and Hinrich Schütze. An Introduction to Information Retrieval. Cambridge University Press, pp.155-195, 2008.

[16] Ranjan, G.S.K, Amar Kumar Verma, and Sudha Radhika. K-nearest Neighbors and Grid Search CV Based Real Time Fault Monitoring System for Industries. In 2019 IEEE 5th International Conference for Convergence in Technology (12CT), pp.1-5, IEEE, 2019.

[17] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learning. Springer, pp.250-256, 2009.

[18] Bishop, C.M. Pattern Recognition and Machine Learning. Springer, pp.312-352, 2006.