Journal of Information Systems and Informatics
Vol. 6, No. 2, June 2024 e-ISSN: 2656-4882 p-ISSN: 2656-5935
DOI: 10.51519/journalisi.v6i2.705 Published By DRPM-UBD

Air Quality Prediction Using the Support Vector Machine Algorithm

Liza Widyarini[1], Hindriyanto Dwi Purnomo[2]

[1,2]Informatics Department, Satya Wacana Christian University, Salatiga, Indonesia
E-mail: [1]672020170@student.uksw.edu, [2]hindriyanto.purnomo@uksw.edu

Abstract

Air quality is an important factor in maintaining the health and well-being of humans and the environment. To anticipate and manage air pollution, air quality prediction has become an important research topic. In this research, researchers propose using the Support Vector Machine (SVM) algorithm to predict air quality. SVM has proven to be an effective method in classification and regression, especially in the context of complex and non-linear data such as air quality data. Researchers utilized historical air quality datasets that include various parameters such as particulates, ozone, nitrogen dioxide and carbon monoxide. Experiments were conducted to compare the performance of SVM with other prediction methods, and the results show that SVM provides accurate and reliable predictions in modeling air quality.

Keywords: Air Quality, Prediction, SVM Algorithm, Air Pollution, Support Vector Machine.

1. INTRODUCTION

Air is a mixture of gases that envelops the Earth, consisting primarily of 78% nitrogen, 20% oxygen, 0.93% argon, and 0.30% carbon dioxide, with the remainder being various other gases. However, the air found in nature is not always clean, which can lead to a decline in air quality. In Indonesia, rapid urbanization and population growth in major cities contribute significantly to air pollution. Human activities such as industrial processes, transportation emissions, land or forest burning, and cigarette smoke are major contributors to this problem. Air pollution has serious health implications, including respiratory tract infections, shortness of breath, lung cancer, and other ailments. It also negatively impacts ecosystems and plant growth, affecting all aspects of life on Earth.

The Air Quality Index (AQI) is a vital measure that indicates the concentration of air pollutants and the associated health risks. The AQI is a numerical, unit-less value that provides information on air quality. Additionally, the Air Pollution Index (ISPU) is used in Indonesia to report air quality and its health effects. The ISPU helps in informing the public and guiding efforts to mitigate air pollution.

p-ISSN: 2656-5935 http://journal-isi.org/index.php/isi e-ISSN: 2656-4882

Given the severity of air pollution and its effects, there is a pressing need for effective methods to predict and monitor air quality. This research addresses this need through the application of data mining techniques to analyze large datasets for predicting and monitoring air quality.

Data mining, the process of extracting significant hidden information from large databases, has evolved significantly with the advancement of analytical tools, facilitating faster and deeper data analysis [1]. It involves several stages, including data collection, preprocessing, transformation, mining, pattern evaluation, and knowledge representation. These stages ensure a systematic approach to uncovering patterns and insights from vast datasets, as depicted in Figure 1. For this study, the classification process employs the Support Vector Machine (SVM) algorithm, a powerful tool in machine learning.

Support Vector Machine (SVM) is a robust learning system that uses linear functions in a high-dimensional feature space, trained through optimization-based algorithms [2]. Introduced by Vapnik in 1992, SVM is effective for both linear and non-linear classification problems. It is divided into Linear SVM, which separates data linearly using a hyperplane, and Non-Linear SVM, which applies the kernel trick to handle data in high-dimensional spaces [3]. The central concept of SVM is to find the optimal hyperplane that best separates the two predefined classes [4]. In this study, the SVM process includes inputting air quality data, calculating the SVM kernel using the Radial Basis Function (RBF), conducting sequential training and testing, and evaluating the classification results [5].

Previous research has utilized various methods for air quality analysis. For example, the study "Data Mining to Aid Policy Making in Air Pollution Management" used Self-Organizing Map Neural Networks to identify data patterns and map air quality distributions in Taiwan [6]. Another study, "Air Pollutants Concentrations Forecasting Using Back Propagation Neural Network Based on Wavelet Decomposition with Meteorological Conditions," employed neural networks to predict daily air pollution concentrations [7]. These studies highlight the importance of using advanced analytical techniques to address air quality issues.

Despite the progress made in air quality prediction, there are gaps in the accuracy and applicability of these models, especially in rapidly growing urban areas like those in Indonesia. Existing models often struggle with the complexity and variability of urban air pollution data. Therefore, this research aims to fill this gap by developing a more accurate and reliable air quality prediction model using the SVM algorithm. By leveraging historical data on air pollution, temperature, humidity, wind speed, and other factors, this study intends to train SVM models to recognize patterns and trends in air quality. The ultimate goal is to enhance the

precision of air quality predictions, providing valuable information for decision-making related to health and environmental management.

This research aims to develop an air quality prediction model using the Support Vector Machine (SVM) algorithm, focusing on improving the accuracy of predictions to assist in decision-making regarding health and environmental issues. The model will use historical data to identify patterns and predict future air quality, benefiting the government, environmental institutions, and the general public. The findings of this study are expected to have significant implications for maintaining healthy air quality and mitigating the adverse effects of air pollution on all living organisms.

2. METHODS

2.1. Research Approach

Data mining refers to the process of extracting significant hidden information from large databases, enabling companies to make informed decisions through the accumulation, analysis, and access to data [8]. Over recent years, the advancement in data mining techniques and the sophistication of analytical tools have significantly enhanced the capability to process and analyze data, leading to a surge in the number of studies utilizing these techniques. Data mining is a multi-stage process that encompasses data collection, preprocessing, transformation, mining, pattern evaluation, and knowledge representation. These stages ensure a systematic approach to uncovering patterns and insights from vast datasets, as depicted in Figure 1.

One effective technique used in the classification process within data mining is the Support Vector Machine (SVM). SVM is a learning system that operates in a high-dimensional feature space using linear functions, and it is trained using optimization-based learning algorithms [9]. Introduced by Vapnik in 1992, SVM has proven to be an efficient classification method, particularly for non-linear problems. SVM can be categorized into Linear SVM and Non-Linear SVM based on its characteristics. Linear SVM is used for data that can be separated linearly, involving the separation of two classes on a hyperplane with a soft margin. Conversely, Non-Linear SVM employs the kernel trick to handle data in high-dimensional spaces [10].

The SVM method relies on mathematical transformations to select the appropriate kernel function for linear problem-solving. The central concept of SVM is to identify the optimal hyperplane that best separates the two predefined classes [11]. The SVM process begins with the input of air quality data, followed by calculating the SVM kernel, specifically the Radial Basis Function (RBF) kernel in this study. Once the kernel value is determined, the Sequential Training SVM calculation

p-ISSN: 2656-5935 http://journal-isi.org/index.php/isi e-ISSN: 2656-4882

process is conducted, followed by SVM testing. The final step involves evaluating the classification results to ensure the model's accuracy and effectiveness [12].

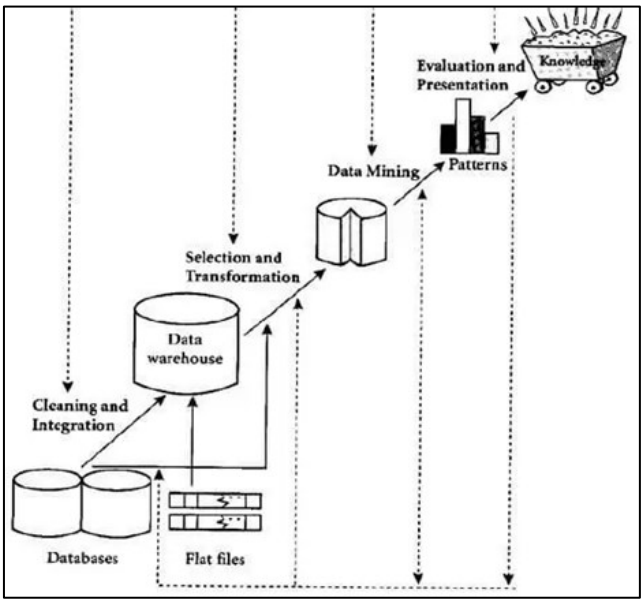

Figure 1. Data Mining Stages [2]

This systematic approach to using SVM ensures robust and reliable classification, making it a powerful tool for analyzing complex datasets and deriving meaningful insights. By following these detailed steps, this study leverages SVM to achieve high precision in classifying air quality data, demonstrating the practical application and benefits of data mining techniques in real-world scenarios.

2.2. Conceptual Model

A conceptual framework is a structure intended to provide researchers with an overview or explanation of the natural development of the phenomenon to be studied or researched. Meanwhile, that theoretically, the conceptual framework will link independent variables and dependent variables, each of which will be measured and observed during the research process. The conceptual model above contains the data needed for the research process and is used to describe the concepts of the problem to be researched so that it is easy to understand. In this research, we use a dataset from Jakarta Open Data to predict air quality in DKI Jakarta [13]. In the basic science section, data mining methods are aimed at predicting air quality data based on the parameters PM10, PM25, SO2, CO, O3, and NO2. To make accurate predictions, the algorithm used is SVM.

p-ISSN: 2656-5935 http://journal-isi.org/index.php/isi e-ISSN: 2656-4882

3. RESULTS AND DISCUSSION

In this research, the Support Vector Machine (SVM) algorithm is used as a method to predict air quality. SVM is a machine learning technique used for classification and regression. Specifically, SVM is used to build models that can learn patterns from historical data about air quality and the factors that influence it. SVM models can be developed and tested to predict air quality with a sufficient level of accuracy. This provides a basis for better decision making in an effort to maintain healthy air quality for society and the environment.

3.1. Pre-Processing

The dataset used is from the Jakarta Open Data website in 2021. The data is in CSV format and in separate files by month. This dataset contains ISPU (Air Pollution Standard Index). An explanation of the parameters or variables contained in the dataset is as follows: Date, PM10, PM25, SO2, CO, O3, NO2, Max, Critical, and Category. Overall, there are 10 attributes in the dataset and l classes. Then the data is combined into one and the same database server. The combined sample data is as in Figure 2.

1/1/2021	43	58	29	35	65	65	O3
1/2/2021	58	86	38	64	80	86	PM25
1/3/2021	64	93	25	62	86	93	PM25
1/4/2021	50	67	24	31	77	77	O3
1/5/2021	59	89	24	35	77	89	PM25
1/6/2021	73	81	29	66	85	85	O3
1/7/2021	36	52	22	55	72	72	O3
1/8/2021	38	68	26	51	71	71	O3
1/9/2021	60	77	34	42	80	80	O3
1/10/2021	24	39	16	38	59	59	O3
1/11/2021	51	72	17	57	68	72	PM25
1/12/2021	29	58	20	44	77	77	O3
1/13/2021	36	47	17	32	68	68	O3
1/14/2021	36	78	20	38	65	78	PM25
1/15/2021	52	82	20	56	65	82	PM25

Figure 2. Data Fusion

After the data merging process, the next step is data selection, the purpose of data selection is to retrieve the column under study. In this study, 6 columns were taken with 5 air quality parameter attributes such as PM10, SO2, CO, O3, NO2, and category attributes. Then enter the data cleaning stage to delete, correct and find inaccurate data, so as to produce high quality data as shown in Figure 3.

p-ISSN: 2656-5935 http://journal-isi.org/index.php/isi e-ISSN: 2656-4882

pm10	pm25	so2	co	o3	no2	max	critical
64	91	77	17	48	41	91	PM25
54	77	51	18	49	35	77	PM25
52	83	51	13	73	38	83	PM25
63	95	52	17	63	47	95	PM25
64	95	52	23	46	46	95	PM25
57	86	52	13	52	40	86	PM25
52	81	51	10	55	25	81	PM25
53	73	52	20	39	41	73	PM25
67	106	51	21	47	38	106	PM25
61	100	51	11	36	25	100	PM25
47	71	51	12	36	32	71	PM25
44	63	47	10	38	19	63	PM25
52	82	47	16	52	40	82	PM25
50	68	47	13	46	34	68	PM25
36	58	47	9	36	17	58	PM25
53	82	47	11	41	14	82	PM25
60	99	50	13	45	35	99	PM25
63	106	43	18	56	35	106	PM25
32	45	41	7	42	12	45	PM25

Figure 3. Data Cleaning Sample

After cleaning the data, you will get ready-to-use data. The following is the output table produced after preprocessing the data. This data will later be processed using the RapidMiner tool. Figure 4 is a display of ready-to-process data.

pm10	so2	co	o3	no2
64	77	17	48	41
54	51	18	49	35
52	51	13	73	38
63	52	17	63	47
64	52	23	46	46
57	52	13	52	40
52	51	10	55	25
53	52	20	39	41
67	51	21	47	38
61	51	11	36	25
47	51	12	36	32
44	47	10	38	19
52	47	16	52	40
50	47	13	46	34
36	47	9	36	17
53	47	11	41	14
60	50	13	45	35
63	43	18	56	35
32	41	7	42	12

Figure 4. Clean Data to Process

p-ISSN: 2656-5935 http://journal-isi.org/index.php/isi e-ISSN: 2656-4882

3.2. Classification Process

The data collected and analyzed in this study has been categorized and presented across several installation services, considering various influencing factors. Figure 5 illustrates the data distribution, highlighting the different air pollutants measured. As shown in Figure 5, the distribution value of air pollution reveals that PM10 (particulate matter with a diameter of 10 micrometers or less) has the highest concentration among the pollutants, with a distribution value of 1.230. This indicates that PM10 is a significant concern in air quality measurements, as it occupies the top position in terms of distribution.

Kernel Model

```
Total number of Support Vectors: 12
Bias (offset): -1.080

w[Bulan] = 0.333
w[pm10] = 1.230
w[so2] = 0.631
w[co] = -0.253
w[o3] = 0.161
w[no2] = 0.056
```

Figure 5. Data Distribution

The data results indicate that various factors influence air quality in Jakarta, with PM10 emerging as the most dominant pollutant, as depicted in Figure 6. This dominance of PM10 underscores the need for targeted strategies to reduce its levels, considering its impact on public health and the environment.

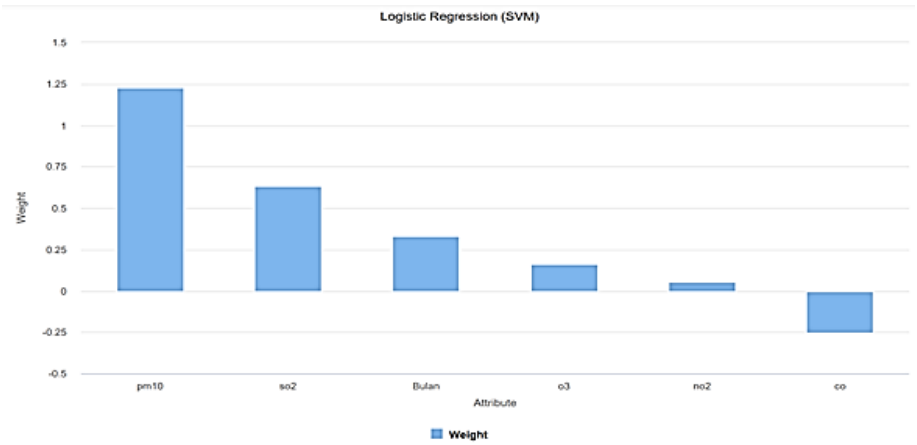

Figure 6. Results with SVM

p-ISSN: 2656-5935 http://journal-isi.org/index.php/isi e-ISSN: 2656-4882

Using the Support Vector Machine (SVM) method, the study aimed to predict air quality in DKI Jakarta. The performance of the SVM model is summarized in Figure 5, which presents the accuracy results of the predictions. The SVM algorithm achieved an accuracy of 83.33%, indicating a high level of precision in predicting air quality based on the data.

accuracy: 83.33%

	true SEDANG	true TIDAK SEHAT	class precision
pred. SEDANG	8	2	80.00%
pred. TIDAK SEHAT	0	2	100.00%
class recall	100.00%	50.00%	

Figure 7. Accuracy Results

3.3. Discussion

While the SVM model demonstrates a strong accuracy rate, it is crucial to understand the nuances of this metric. Accuracy, which considers both True Negatives and False Negatives, does not always reflect perfect model performance, even if precision and recall values are 100%. The precision of 100% means that all predicted positives were actually positive and recall of 100% indicates that all actual positives were correctly identified. However, the overall accuracy is influenced by the balance of all classes in the dataset, including True Negatives and False Negatives. Additionally, other metrics such as the F1-score, which is the harmonic mean of precision and recall, provide a more comprehensive evaluation of the model's performance. The F1-score is particularly useful in cases where there is an uneven class distribution, offering a balanced measure of accuracy.

The findings from this study have significant implications for air quality management in urban areas like Jakarta. The high levels of PM10 identified suggest that immediate and targeted interventions are necessary to mitigate its impact. The use of SVM for predicting air quality proves to be an effective approach, offering a reliable tool for policymakers and environmental agencies.

By leveraging historical data and sophisticated machine learning algorithms, this research provides a framework for improving air quality predictions. The model's accuracy and reliability can help in formulating more effective policies and actions to address air pollution. Furthermore, the insights gained from this study can be used to enhance public awareness and encourage proactive measures to maintain and improve air quality.

p-ISSN: 2656-5935 http://journal-isi.org/index.php/isi e-ISSN: 2656-4882

4. CONCLUSION

This study demonstrates the application of SVM in predicting air quality with a notable accuracy rate. The dominance of PM10 in Jakarta's air pollution highlights the need for focused mitigation strategies. While accuracy is a crucial metric, it is essential to consider other performance measures like the F1-score to fully understand the model's efficacy. The results of this study offer valuable insights for enhancing air quality management and ensuring healthier living conditions in urban environments.

REFERENCES

[1] S. T. Li and L. Y. Shue, "Data mining to aid policy making in air pollution management," *Expert Systems with Applications*, vol. 27, no. 3, pp. 331-340, 2004.

[2] Y. Bai, Y. Li, X. Wang, J. Xie, and C. Li, "Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions," *Atmospheric Pollution Research*, vol. 7, no. 3, pp. 557-566, 2016.

[3] S. Mahajan, P. Kumar, J. A. Pinto, A. Riccetti, K. Schaaf, G. Camprodon, V. Smári, A. Passani, and G. Forino, "A citizen science approach for enhancing public understanding of air pollution," *Sustainable Cities and Society*, vol. 52, p. 101800, 2020.

[4] B. Novianti, T. Rismawan, and S. Bahri, "Implementasi Data Mining Dengan Algoritma C4.5 Untuk Penjurusan Siswa (Studi Kasus: Sma Negeri 1 Pontianak)," *J. Coding, Sist. Komput. Untan*, vol. 4, no. 3, pp. 75-84, 2016.

[5] P. M. S. Tarigan, J. T. Hardinata, H. Qurniawan, M. Safii, and R. Winanjaya, "Implementasi Data Mining Menggunakan Algoritma Apriori Dalam Menentukan Persediaan Barang: Studi Kasus: Toko Sinar Harahap," *Jurnal Janitra Informatika dan Sistem Informasi*, vol. 2, no. 1, pp. 9-19, 2022.

[6] A. M., Puspitasari, D. E. Ratnawati, and A. W. Widodo, "Klasifikasi Penyakit Gigi dan Mulut Menggunakan Metode Support Vector Machine," *Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 2, 2018.

[7] W. Purnami, A. M. Regresi, and L. Ordinal, "Perbandinganl Klasifikasi Tingkat Keganasan Breast Cancer Dengan Menggunakan Regresi Logistik Ordinal Dan Support Vector Machine (SVM)," *J. Sains Dan Seni ITS*, vol. 1, no. 1, 2012.

[8] R. Tineges, T. Agung, and D. S. Ira, "Analisis Sentimen Terhadap Layanan Indihome Berdasarkan Twitter Dengan Metode Klasifikasi Support Vector Machine (SVM)," *Jurnal Media Informatika Budidarma*, vol. 4, no. 3, p. 650, 2020.

[9] S. H. Wibowo and R. Toyib, "Support Vector Machine Method for Recognizing Patterns in Signatures," *Jurnal Media Infotama*, vol. 18, no. 2, pp. 323-327, 2022.

[10] A. Laturiuw and Y. Singgalen, "Sentiment Analysis of Raja Ampat Tourism Destination Using CRISP-DM: SVM, NBC, DT, and k-NN Algorithm," *J. Inf. Syst. Informatics*, vol. 5, no. 2, pp. 518-535, May 2023.
[11] A. Darmawan, M. Al Wajieh, M. Setyawan, T. Yandi, and H. Hoiriyah, "Hoax News Analysis for the Indonesian National Capital Relocation Public Policy with the Support Vector Machine and Random Forest Algorithms," *J. Inf. Syst. Informatics*, vol. 5, no. 1, pp. 150-173, Mar. 2023.
[12] Y. Singgalen, "Sentiment Analysis on Customer Perception towards Products and Services of Restaurant in Labuan Bajo," *J. Inf. Syst. Informatics*, vol. 4, no. 3, pp. 511-523, Sep. 2022.
[13] Minister of Environment and Forestry Regulation number 14 of 2020 concerning the Air Pollution Standard Index (ISPU).