OKTAL : Jurnal Ilmu Komputer dan Science
Volume 2, No. 11, November 2023
ISSN 2828-2442 (media online)
Hal 3043-3049

Komparasi Algoritma Support Vector Machine Dan CART Untuk Klasifikasi Kualitas Udara Dki Jakarta

Rizki Prayogi Widartama[1*], Maulana Fansyuri[1]

[1]Fakultas Ilmu Komputer, Teknik Informatika, Universitas Pamulang, Jl. Raya Puspiptek No. 46, Kel. Buaran, Kec. Serpong, Kota Tangerang Selatan. Banten 15310, Indonesia
Email: [1*]prayogi.rizki00@gmail.com, [2]dosen02359@unpam.ac.id
(* : coressponding author)

Abstrak− Kualitas udara atau mutu udara merupakan ukuran kondisi udara pada waktu dan tempat tertentu yang diukur dan/atau diuji berdasarkan parameter tertentu. Paparan tingkat tinggi pencemaran udara dapat menyebabkan berbagai kerugian untuk kesehatan, yaitu dapat meningkatkan resiko infeksi pernapasan, penyakit jantung dan kanker paru – paru. Jakarta menempati peringkat ke 12 sebagai ibu kota regional 2021 dengan konsentrasi PM2.5 tahunan rata – rata tertinggi. Sedangkan untuk di kawasan Asia Tenggara, Jakarta menempati peringkat ke 6 sebagai kota regional paling terpolusi. Pengklasifikasian kualitas udara yang seragam dan tepat dapat menjadi peran penting untuk perencanaan serta pengenalan kebijakan dan peraturan yang relevan untuk pengelolaan polusi udara oleh para pengambil keputusan, dalam melakukan klasifikasi dapat menggunakan teknik data mining. Algoritma Support Vector Machine (SVM) dan Classification and Regression Tree (CART) merupakan bagian dari metode klasifikasi. Pada penelitian ini dilakukan analisis dan komparasi untuk mengetahui kinerja kedua metode tersebut dalam mengklasifikasi kualitas udara di Jakarta pada tahun 2021. Dan menghasilkan klasifikasi SVM memiliki nilai akurasi sebesar 95.05% dan nilai classification error sebesar 4.95%, serta hasil klasifikasi CART yaitu dengan nilai akurasi sebesar 99.67% dan nilai classification error sebesar 0.33%. Dapat disimpulkan algoritma CART lebih baik dibandingkan SVM dalam melkakukan klasifikasi untuk mengetahui kualitas udara di DKI Jakarta.

Kata Kunci: Kualitas Udara, Data Mining, Klasifikasi, Support Vector Machine, *Classification and Regression Tree*

Abstract− *Air quality or air quality is a measure of air condition at a certain time and place that is measured and/or tested based on certain parameters. Exposure to high levels of air pollution can cause various harms to health, which can increase the risk of respiratory infections, heart disease and lung cancer. Jakarta is ranked 12th as a regional capital for 2021 with an annual average concentration of PM2.5 – the highest on average. As for the Southeast Asia region, Jakarta is ranked 6th as the most populous regional polluted city. Uniform and precise air quality classification can be an important role for planning and introduction of relevant policies and regulations for air pollution management by decision makers, in carrying out the classification can use technical data mining. The Support Vector Machine (SVM) and Classification and Regression Tree (CART) algorithms are part of the classification method. In this study, an analysis and comparison was carried out to determine the performance of the two methods in classifying air quality in Jakarta in 2021. And the resulting SVM classification has an accuracy value of 95.05% and an error classification value of 4.95%, and the results of the CART classification are with an accuracy value of 99.67% and a misclassification value of 0.33%. It can be interpreted that the CART algorithm is better than SVM in classification classification to determine air quality in DKI Jakarta.*

Keywords*: Air Quality, Data Mining, Classification, Support Vector Machine, Classification and Regression Tree*

1. PENDAHULUAN

Udara menjadi faktor yang penting bagi kesehatan manusia. Dimana era saat ini sejalan dengan pertumbuhan dan perkembangan pesat fisik bangunan, pusat industri serta transportasi, berimbas kepada kualitas udara yang mengalami perubahan akibat terjadinya pencemaran udara atau berubahnya salah satu komposisi udara yaitu masuknya zat pencemar (berbentuk gas – gas dan atau partikel kecil) kedalam udara dengan jumlah tertentu dalam jangka waktu yang cukup lama (Ismiyati et al., 2014). Berdasarkan (Gusnita, 2012) sumber pencemaran dari industri dan sarana transportasi kendaraan bermotor menjadi kontributor terbesar dalam pencemaran udara, terutama emisi transportasi di perkotaan menjadi penyumbang pencemaran udara tertinggi di Indonesia yaitu sekitar 85 persen. Menurut (WHO, 2019) paparan tingkat tinggi pencemaran udara dapat menyebabkan

OKTAL : Jurnal Ilmu Komputer dan Science
Volume 2, No. 11, November 2023
ISSN 2828-2442 (media online)
Hal 3043-3049

berbagai kerugian untuk kesehatan, yaitu dapat meningkatkan resiko infeksi pernapasan, penyakit jantung dan kanker paru – paru.

Hasil pemantauan (IQAir, 2021) sebagai *platform* informasi kualitas udara *real-time* terbesar di dunia, Jakarta menempati peringkat ke 12 sebagai ibu kota regional 2021 dengan konsentrasi PM2.5 tahunan rata – rata tertinggi. Sedangkan untuk di kawasan Asia Tenggara, Jakarta menempati peringkat ke 6 sebagai kota regional paling terpolusi dengan PM2.5 tahunan rata – rata konsentrasi tujuh kali melebihi pedoman kualitas udara WHO.

Di Indonesia analisis kualitas udara dapat dilakukan dengan Indeks Standar Pencemaran Udara (ISPU), dimana perhitungannya merujuk pada peraturan terbaru yaitu Peraturan Menteri Lingkungan Hidup dan Kehutanan Nomor P.14/MENLHK/SETJEN/KUM.1/7/2020 (DLH DKI Jakarta, 2021).

Untuk menunjukkan konsentrasi polutan maka diusulkan indeks dengan pengklasifikasi berbasis algoritma *supervised learning* (Saxena & Shekhawat, 2017). Pengklasifikasian kualitas udara yang seragam dan tepat dapat menjadi peran penting untuk perencanaan serta pengenalan kebijakan dan peraturan yang relevan untuk pengelolaan polusi udara oleh para pengambil keputusan (Arif, 2018; Hu & Li, 2017), dalam melakukan klasifikasi dapat menggunakan teknik *data mining* (Arif, 2018).

Sebelumnya, terdapat beberapa penelitian yang telah dilakukan menggunakan teknik klasifikasi *data mining* dengan algoritma *supervised learning*. Beberapa literatur menggunakan *dataset* mengenai polusi dan kualitas udara, dimana data tersebut bersifat fluktuatif, kontinu, berbentuk numerik dan memiliki topik yang sama. Misalnuya pada penelitian yang dilakukan oleh Bingchun Liu dkk (Liu et al., 2018) yang hasil penelitiannya menunjukkan algoritma SVM sebagai algoritma dengan nilai akurasi tertinggi sebesar 90.12%, sedangkan untuk ANN 88.25%, KNN 88.01%. Penelitian berikutnya yang telah dilakukan oleh Yin Zhao dan Yahya Abu Hasan (Zhao & Hasan, 2013) yang hasil penelitiannya menunjukkan algoritma Support Vector Machine memiliki tingkat akurasi tertinggi sebesar 81.1% sedangkan ANN 77.6%. Penelitian yang dilakukan oleh Aditya Hermawan (Hermawan, 2019) hasil penelitiannya menunjukkan algoritma Support Vector Machine dengan nilai akurasi sebesar 96.03%. Selanjutnya penelitian yang dilakukan Snezhana Georgieva Gocheva-Ilieva dkk (Gocheva-Ilieva et al., 2019) dengan ketepatasn algoritma CART pada data polusi partikel PM_{10} di dua kota Bulgaria yaitu Ruse dan Pernik menghasilkan nilai akurasi 89% untuk data kota Ruse dan 91% data diklasifikasikan dengan benar untuk kota Pernik. Penelitian yang dilakukan oleh Snezhana Georgieva Gocheva-Ilieva dan Maya Stoimenova (Gocheva-Ilieva & Stoimenova, 2018) hasil penelitiannya menunjukkan algoritma CART memiliki nilai akurasi sebesar 90.3%. Penelitian yang dilakukan Maya Stoimenova-Minova dkk (Stoimenova-Minova et al., 2020) hasil penelitiannya dengan algoritma CART menunjukkan nilai akurasi 86.8%.

Sebelumnya pada penelitian yang dilakukan oleh (Hermawan, 2019) telah melakukan pengklasifikasian kualitas udara yang berlokasi di DKI Jakarta, dengan lima parameter pencemar yaitu CO, SO_2, PM_{10}, O_3, NO_2. Pada penelitian ini akan melakukan pengklasifikasian kualitas udara dengan enam parameter pencemar yaitu CO, SO_2, PM_{10}, O_3, NO_2 dan $PM_{2.5}$ dalam menentukan kualitas udara di DKI Jakarta.

Berdasarkan penelitian yang telah diuraikan menunjukkan bahwa menggunakan algoritma SVM dan CART menghasilkan model dengan kinerja tertinggi. Oleh karena itu pada penelitian ini akan melakukan perbandingan algoritma *Support Vector Machine* (SVM) dan *Classification and Regression Trees* (CART) untuk klasifikasi kualitas udara di DKI Jakarta berdasarkan kategori dari dataset ISPU.

2. METODOLOGI PENELITIAN

2.1 Metode Pengumpulan Data

Data yang digunakan pada penelitian ini yaitu data Indeks Standar Pencemaran Udara (ISPU) di DKI Jakarta selama periode 1 Januari – 31 Desember 2021 yang diperoleh dari bidang pengendalian dampak lingkungan DLH Provinsi DKI Jakarta. Dengan dataset yang diperoleh berjumlah 12 yang terdiri dari 1825 *record*.

OKTAL : Jurnal Ilmu Komputer dan Science
Volume 2, No. 11, November 2023
ISSN 2828-2442 (media online)
Hal 3043-3049

2.2 Pengolahan Awal Data (*Preprocessing*)

Pengolahan awal data meliputi proses seleksi atribut, pembersihan data dan pemecahan data menjadi data training dan testing. Pada penelitian ini atribut yang akan digunakan yaitu enam atribut prediktor di antaranya pm10, pm25, so2, co, o3, no2 dan satu atribut kelas yaitu categori.

Pada atribut yang memiliki data kosong maka akan dilakukan pembersihan yaitu tidak memakainya atau menghapus data yang kosong. Sedangkan untuk atribut "categori" meskipun tidak mempunyai data kosong akan dilakukan penghapusan data yang memiliki nilai "TIDAK ADA DATA" karena data tersebut tidak memiliki makna dan tidak digunakan. Tabel 1 berikut menampilkan data kosong pada atribut yang akan digunakan.

Tabel 1. Data Kosong Pada Atribut

Atribut	Data Kosong	Proses
PM_{10}	68	Digunakan
PM_{25}	100	Digunakan
SO_2	114	Digunakan
CO	36	Digunakan
O_3	68	Digunakan
No_2	35	Digunakan
Categori	0	Digunakan

Sehingga dari keseluruahan record yang awal berjumlah 1825 *record*, setelah dilakukan pembersihan data pada atribut yang digunakan menjadi 1518 *record*. Selanjutnya untuk pemecahan data, pada penelitian ini dilakukan *split data* secara *Stratified sampling* dimana pembagian data secara acak dengan proporsi distribusi kelas yang sama. Pembagian data *training* dan *testing* masing - masing sebesar 80% dan 20%. Dari total sampel data yang telah dilakukan pembersihan berjumlah 1518, maka untuk data *training* diperoleh data sebesar 1215 sampel data, dan untuk data *testing* diperoleh data sebesar 303 sampel data.

2.3 Model Yang Diusulkan

Model yang diusulkan adalah komparasi antara algoritma Support Vector Machine (SVM) dan Classification and Regression Tree (CART) untuk klasifikasi kualitas udara DKI Jakarta. Kemudian di implementasikan menggunakan *software RapidMiner V9.6*.

2.4 Pengujian Metode

Pengujian metode penelitian ini menggunakan *software RapidMiner 9.6* untuk melakukan proses pengolahan dan penghitungan terhadap model yang telah diusulkan.

2.5 Evaluasi Dan Validasi Hasil

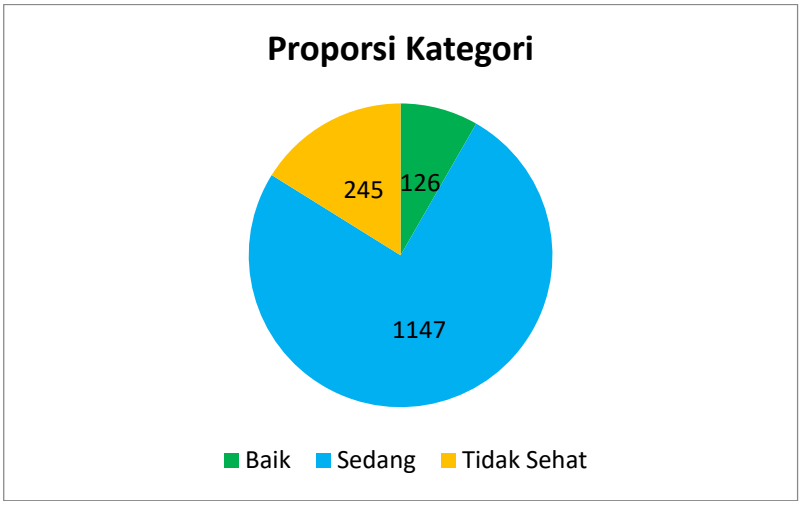
Confussion Matrix digunakan untuk melakukan evaluasi dan validasi hasil yang nantinya akan digunakan untuk komparasi kedua algoritma yang diusulkan. Dengan mengamati hasil klasifikasi menggunakan algoritma *Support Vector Machine* dan *Classification and Regression Tree*. Validasi dilakukan dengan mengukur hasil klasifikasi data uji(*Testing*). Pengukuran kinerja menggunakan *Accuracy, Recall, Precission* dan *Classification Error*. Informasi yang didapatkan kemudian akan di komparasi sehingga mendapatkan model yang lebih akurat diantara kedua algoritma yang diusulkan.

3. ANALISA DAN PEMBAHASAN

3.1 Analisa

Proses pengujian ini menggunakan 1518 *record* data yang sudah dibersihkan. Dengan variabel yang digunakan sebanyak 7, dimana 6 variebel bertipe integer yang akan diigunakan sebagai atribut prediktor dan 1 variabel bertipe kategorial/nominal yang akan digunakan sebagai atrbut kelas. Atribut kelas yang digunakan dibagi 3, yaitu 'Baik', 'Sedang' dan 'Tidak Sehat' dengan proporsi seperti pada Gambar 1 berikut.

OKTAL : Jurnal Ilmu Komputer dan Science
Volume 2, No. 11, November 2023
ISSN 2828-2442 (media online)
Hal 3043-3049

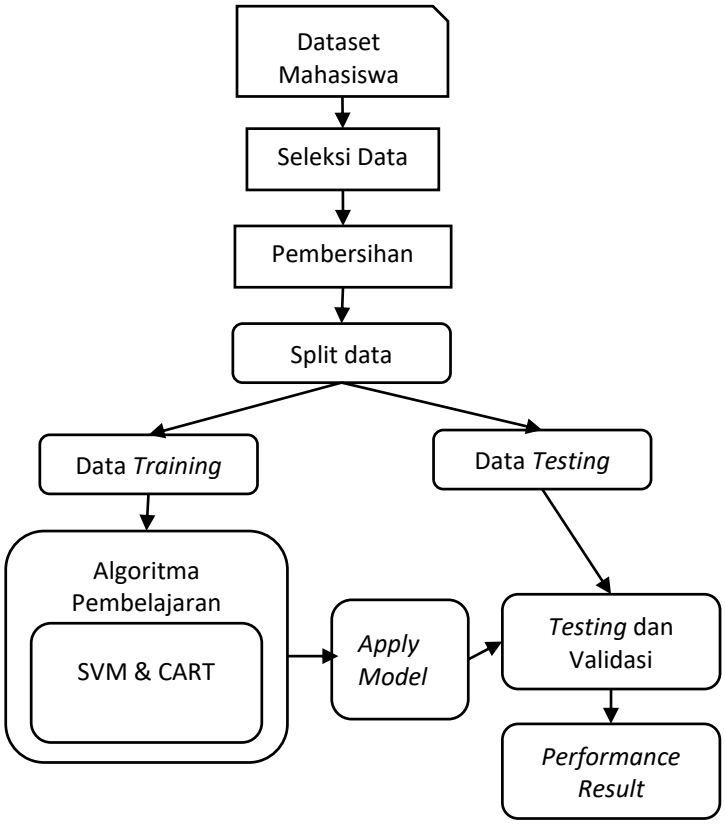


Gambar 1. Jumlah Proporsi Kategori

Dari Gambar 1 diatas dapat dilihat proporsi untuk atribut kategori dari total 1518 data terdapat kategori 'Baik' sebanyak 126 data, 'Sedang' sebanyak 1147 data dan 'Tidak Sehat' sebanyak 245 data.

3.2 Diagram Alur

Diagram alur penelitian yang dilakukan berdasarkan model yang diusulkan untuk mendapatkan hasil yang diharapkan seperti pada Gambar 2 berikut.



Gambar 2. Diagram Alur

OKTAL : Jurnal Ilmu Komputer dan Science
Volume 2, No. 11, November 2023
ISSN 2828-2442 (media online)
Hal 3043-3049

4. IMPLEMENTASI

4.1 Pengujian Algoritma *Support Vector Machine*

Tools RapidMiner Version 9.6 digunakan untuk pengolahan data mining pada pengujian klasifikasi dengan metode *Support Vector Machine rbf kernel*. Dari 1518 *record data* yang telah dibersihkan, dengan pembagian data *training* sebesar 80% dan data *testing* sebesar 20% yaitu 1215 data *training* dan 303 data *testing*.

Dari hasil pengujian dengan data testing menggunakan *tools rapidminer version 9.6* diperoleh hasil prediksi benar sebanyak 288 dan prediksi salah sebanyak 15 dari 303 data *testing*.

Diperoleh hasil *confusion matrix* pada data testing menggunakan algoritma *support vector machine rbf kernel* seperti pada Gambar 3 berikut.

	true SEDANG	true BAIK	true TIDAK SEHAT	class precision
pred. SEDANG	228	9	5	94.21%
pred. BAIK	0	16	0	100.00%
pred. TIDAK SEHAT	1	0	44	97.78%
class recall	99.56%	64.00%	89.80%	

Gambar 3. *Confusion Matrix* SVM

4.2 Pengujian Algoritma *Classification and Regression Tree*

Tools RapidMiner Version 9.6 digunakan untuk pengolahan data mining pada pengujian klasifikasi dengan metode *Classification and Regression Tree*. Dari 1518 record data yang telah dibersihkan, dengan pembagian data *training* sebesar 80% dan data *testing* sebesar 20% yaitu 1215 data *training* dan 303 data *testing*. Dari hasil pengujian dengan data *testing* menggunakan *tools rapidminer version 9.6* diperoleh hasil prediksi benar sebanyak 302 data dan prediksi salah sebanyak 1 data dari 303 data *testing*.

Diperoleh hasil *confusion matrix* pada data testing menggunakan algoritma *Classification and Regression Tree* seperti pada Gambar 4 berikut.

	true SEDANG	true BAIK	true TIDAK SEHAT	class precision
pred. SEDANG	229	0	1	99.57%
pred. BAIK	0	25	0	100.00%
pred. TIDAK SEHAT	0	0	48	100.00%
class recall	100.00%	100.00%	97.96%	

Gambar 4. *Confusion Matrix* CART

4.3 Hasil Komparasi Algoritma SVM dan CART

Dari hasil pengujian masing - masing metode dengan pembagian data 80% data *training* dan 20% data *testing*, didapatkan nilai *accuracy*, *recall*, *precision* dan *classification error* yang berbeda – beda. Berikut ini adalah komparasi dari algoritma *Support Vector Machine* dan *Classification and Regression Tree* dapat dilihat pada Tabel 1 berikut.

Tabel 1. Hasil Komparasi Algoritma SVM dan CART

Algoritma	***Accuracy***	***Recall***	***Precision***	***Classification Error***
SVM *rbf-kernel*	95.05%	84.45%	97.33%	4.95%
CART	99.67%	99.32%	99.86%	0.33%

OKTAL : Jurnal Ilmu Komputer dan Science
Volume 2, No. 11, November 2023
ISSN 2828-2442 (media online)
Hal 3043-3049

Pada Tabel 1 menjelaskan tentang akurasi algoritma support vector machine menghasilkan nilai nilai sebesar 95.05%, rata – rata nilai recall sebesar 84.45%, rata – rata nilai precision sebesar 97.33%, dan nilai classification error sebesar 4.95%. Sedangkan algoritma classification and regression tree menghasilkan nilai akurasi sebesar 99.67%, rata – rata nilai recall sebesar 99.32%, rata – rata nilai precision sebesar 99.86%, dan nilai classification error sebesar 0.33%. Pada Tabel 4.5 dapat dilihat bahwa dalam melakukan klasifikasi kualitas udara DKI Jakarta tahun 2021, nilai akurasi CART lebih tinggi dibandingkan algoritma SVM, ini menunjukkan bahwa dalam klasifikasi kualitas udara DKI Jakarta tahun 2021 algoritma CART lebih unggul. Meskipun begitu SVM juga masih dapat digunakan sebagai model klasifikasi yang baik karena menghasilkan nilai akurasi diatas 90%.

5. KESIMPULAN

Berdasarkan pada hasil penerapan dan pengujian algoritma Support Vector Machine dan Classification and Regression Tree dalam mengklasifikasi kualitas udra di DKI Jakarta tahun 2021, dapat disimpulkan sebagai berikut:

a. Penerapan algoritma Support Vector Machine dan Classification and Regression Tree untuk klasifikasi kualitas udara di DKI Jakarta dengan variabel PM2.5, CO, SO2, PM10, O3, dan NO2 melalui aplikasi RapidMiner.
b. Nilai akurasi yang dihasilkan berdasarkan pengujian menunjukkan algoritma Support Vector Machine nilai akurasinya lebih rendah dibandingkan algoritma Classification and Regression Tree. Dengan hasil klasifikasi SVM memiliki nilai akurasi sebesar 95.05% dan nilai classification error sebesar 4.95%, serta hasil klasifikasi CART yaitu dengan nilai akurasi sebesar 99.67% dan nilai classification error sebesar 0.33%. Dapat disimpulkan algoritma CART lebih baik dibandingkan SVM dalam melkakukan klasifikasi untuk mengetahui kualitas udara di DKI Jakarta.

REFERENCES

Arif, M. (2018). *Klasifikasi Kualitas Udara Menggunakan Metode Modified K-Nearest Neighbor(Mk-Nn)(Studi Kasus Kota Pekanbaru)*. 95. http://repository.uin-suska.ac.id/id/eprint/16506

DLH DKI Jakarta. (2021). Laporan Akhir Kegiatan Pemantauan Kualitas Udara Provinsi DKI Jakarta Tahun 2021. In *Dinas Lingkungan Hidup Provinsi DKI Jakarta*. https://lingkunganhidup.jakarta.go.id/files/laporan2021/Laporan_Akhir_Pemantauan_KualitasUdara_2021_final.pdf

Gocheva-Ilieva, S. G., & Stoimenova, M. P. (2018). *PM10 Prediction and Forecasting Using CART : A Case Study for Pleven, Bulgaria*. *12*(9), 572–577.

Gocheva-Ilieva, S. G., Voynikova, D. S., Stoimenova, M. P., Ivanov, A. V., & Iliev, I. P. (2019). Regression trees modeling of time series for air pollution analysis and forecasting. *Neural Computing and Applications*, *31*(12), 9023–9039. https://doi.org/10.1007/s00521-019-04432-1

Gusnita, D. (2012). Pencemaran logam berat timbal (pb) di udara dan upaya penghapusan bensin bertimbal. *Berita Dirgantara*, *13*(3), 95–101.

Hermawan, A. (2019). *SPKU: Sistem Prediksi Kualitas Udara (Studi Kasus: Dki Jakarta)*. http://eprints.uty.ac.id/3552/

Hu, N., & Li, Q. (2017). *Application of Decision Tree C4.5 Algorithm in Air Quality Evaluation*. *118*(Amcce), 1095–1099. https://doi.org/10.2991/amcce-17.2017.197

IQAir. (2021). *World Air Quality Report 2021 Region and City PM2.5 Ranking*. https://www.iqair.com/world-most-polluted-cities/world-air-quality-report-2021-en.pdf

Ismiyati, Marlita, D., & Saidah, D. (2014). Pencemaran Udara Akibat Emisi Gas Buang Kendaraan Bermotor. *Jurnal Manajemen Transportasi & Logistik (JMTransLog)*, *01*(03), 241–248.

Liu, B., Chang, P.-C., Huang, N., & Li, D. (2018). Multi-Level Air Quality Classification in China Using Information Gain and Support Vector Machine. *World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering*, *12*(12), 1092–1101.

Saxena, A., & Shekhawat, S. (2017). Ambient Air Quality Classification by Grey Wolf Optimizer Based Support Vector Machine. *Journal of Environmental and Public Health*, *2017*. https://doi.org/10.1155/2017/3131083

Stoimenova-Minova, M., Gocheva-Ilieva, S., & Ivanov, A. (2020). PM10 Prediction Using CART Method Depending on the Number of Observations. *ACM International Conference Proceeding Series*, 65–70. https://doi.org/10.1145/3409915.3409919

WHO. (2019). *Health consequences of air pollution on populations*. World Health Organisation (WHO). https://www.who.int/news-room/detail/15-11-2019-what-are-health-consequences-of-air-pollution-on-populations

Zhao, Y., & Hasan, Y. A. (2013). *Machine learning algorithms for predicting roadside fine particulate*. *3*(3), 61–73.