Buletin Ilmiah Sarjana Teknik Elektro
Vol. 4, No. 1, April 2022, pp. 12-21
ISSN: 2685-9572, DOI: 10.12928/biste.v4i1.6079

Model Prediksi Kualitas Udara dengan Support Vector Machines dengan Optimasi Hyperparameter GridSearch CV

Ahmad Toha [1], Purwono Purwono [2], Windu Gata [3]
[1,3] Program Studi Ilmu Komputer, Universitas Nusa Mandiri, Indonesia
[2] Program Studi Informatika, Universitas Harapan Bangsa, Indonesia

INFORMASI ARTIKEL

Riwayat Artikel:

Dikirimkan 16 April 2022
Direvisi 19 Mei 2022
Diterima 26 Mei 2022

Kata Kunci:

Classification;
Air Quality;
Data Science;
SVM;
Grid Search

Penulis Korespondensi:

Ahmad Toha,
Universitas Nusa Mandiri,
Jl. Kramat Raya No.18,
Kwitang, Kec. Senen, Jakarta,
Indonesia.
Email:
14210158@nusamandiri.ac.id

***ABSTRACT* / ABSTRAK**

Air pollution continues to increase in Jakarta. The city ranks 12th in the world as the capital of a country with high levels of pollution. The Jakarta Environmental Service requires processing air quality data generated by the Air Quality Monitoring Station in order to produce valuable information as a decision-making tool. This data processing can be processed with data mining techniques to seek new knowledge from the database so as to find valid, useful and easy-to-learn patterns. The SVM data mining classification model is proposed in this study. Our contribution in this research is to create a classification model with SVM with new techniques, namely improvements in data processing to perform hyperparameter tuning. We saw that previous researchers only pursued high accuracy scores. In contrast to previous studies, we used the gridsearch cv hyperparameter optimization technique on the SVM classification model. The kernel polynomial with 2 degrees is the best parameter recommendation from the grid search cv technique. The accuracy before optimization is 73,31%, while after optimization is 94,8%. This shows an increase in accuracy of 3.2% after applying the grid search cv method to the classification of air quality monitoring using the SVM model

Pencemaran udara terus meningkat di Jakarta. Kota ini menempati urutan ke 12 di dunia sebagai ibukota negara dengan tingkat polusi tinggi. Dinas Lingkungan Hidup Jakarta memerlukan pengolahan data-data kualitas udara yang dihasilkan oleh Stasiun Pemantauan Kualitas Udara agar menghasilkan informasi berharga sebagai alat pengambil keputusan. Pengolahan data ini dapat diproses dengan teknik data mining untuk mencari pengetahuan baru dari basis data sehingga menemukan pola-pola yang valid, bermanfaat dan dapat dipelajari dengan mudah. Model klasifikasi *data mining* SVM diusulkan dalam penelitian ini. Kontribusi kami dalam penelitian ini adalah membuat model klasifikasi dengan SVM dengan teknik baru yaitu perbaikan dalam pemrosesan data hingga melakukan *hyperparameter tuning*. Kami melihat para peneliti sebelumnya hanya mengejar nilai akurasi yang tinggi. Berbeda dengan penelitian sebelumnya, kami menggunakan teknik optimasi hiperparameter gridsearch cv pada model klasifikasi SVM. Polinomial kernel dengan 2 derajat merupakan rekomendasi parameter terbaik dari teknik *grid search cv*. Akurasi sebelum optimasi adalah 73,31%, sedangkan setelah optimasi adalah 94,8%. Hal ini menunjukkan peningkatan akurasi sebesar 21,5% setelah menerapkan metode *grid search cv* pada klasifikasi pemantauan kualitas udara menggunakan model SVM.

This work is licensed under a Creative Commons Attribution-Share Alike 4.0

Sitasi Dokumen ini:

A. Toha, P. Purwono, and W. Gata, "Model Prediksi Kualitas Udara dengan Support Vector Machines dengan Optimasi Hyperparameter GridSearch CV," *Buletin Ilmiah Sarjana Teknik Elektro*, vol. 4, no. 1, pp. 12-21, 2022, doi: 10.12928/biste.v4i1.6079.

1. PENDAHULUAN

Pencemaran udara di Indonesia terus meningkat akhir-akhir ini. Berdasarkan data yang diperoleh dari *katadata*, salah satu kota besar di dunia yaitu Jakarta telah menempati peringkat ke 12 sebagai ibukota negara dengan tingkat polusi tinggi. Salah satu penyebab terjadinya polusi adalah adanya peningkatan penggunaan jumlah gas buang kendaraan bermotor [1]. Polutan utama yang ditimbulkan oleh asap kendaraan adalah karbon monoksida (CO), hidrokarbon (HC), karbon dioksida (CO^2), oksigen (O^2) dan nitrogen oksida (NO_x) [2]. Polusi udara dapat berdampak buruk bagi kesehatan masyarakat di antaranya adalah gangguan pernafasan bagi orang tua maupun anak-anak [3]. Berdasarkan data yang diambil dari Badan Pusat Statistik (BPS) jumlah kenaikan penggunaan kendaraan bermotor terus meningkat tiap tahunnya. Tercatat adanya kenaikan kendaraan hingga 19% pada tahun 2018 hingga tahun 2020 di Jakarta.

Pencemaran udara menjadi masalah serius yang dialami oleh negara-negara di dunia [4]. Banyak lembaga-lembaga pemerintahan yang terus memantau data terkait pencemaran udara. DKI Jakarta sebagai ibukota negara Indonesia juga telah memiliki data-data indeks standar pencemaran udara (ISPU) yang dikelola oleh Dinas Lingkungan Hidup. ISPU digunakan sebagai parameter mengukur kualitas udara. Dinas Lingkungan Hidup memiliki 5 stasiun pemantau kualitas udara (SKPU).

Pemerintah terus berupaya untuk menangani pencemaran udara, di antara adalah peraturan tentang pengendalian pencemaran udara pada PP No. 41 Tahun 1999. Salah satu kegiatan yang dilakukan yaitu pengaturan kebijaksanaan dalam penggunaan bahan bakar yang bersih dan ramah lingkungan. Pemerintah juga telah mengembangkan kebijakan dengan membangun AQMS atau *Air Quality Monitoring System*. Berbagai upaya yang sudah dilakukan oleh pemerintah rupanya masih belum cukup efektif untuk mengatasi masalah pencemaran udara ditambah dengan minimumnya kesadaran masyarakat akan polusi udara. Pemerintah berdasarkan keputusan Badan Pengendalian Dampak Lingkungan Hiduo (BAPEDAL) Nomor KEP-107/Kabapedal/11/1997 menentukan ukuran standar kualitas udara di sebuah daerah yaitu Indeks Standar Pencemaran Udara (ISPU). Semakin tinggi level ISPU maka semakin buruk terhadap kesehatan. ISPU memiliki beberapa tingkatan yaitu baik, sedang, tidak sehat, sangat tidak sehat dan berbahaya [5].

Pencemaran udara di Jakarta menjadi masalah yang serius harus dihadapi akibat memiliki tingkat polusi yang signifikan [6]. Dinas Lingkungan Hidup Jakarta memerlukan pengolahan data-data kualitas udara yang ada agar menghasilkan informasi berharga sebagai pengambil keputusan [7]. Pengolahan data ini dapat diproses dengan teknik *data mining* atau biasa disebut dengan penambangan data. *Data mining* ialah proses pengumpulan dan pemakaian data historis untuk menemukan pengetahuan baru dalam basis data berukuran besar untuk menemukan pola-pola yang valid, bermanfaat dan dapat dipelajari dengan mudah [8][9]. *Data mining* dapat berfungsi sebagai model prediktif dari basis data [10]. Beberapa kelompok pola pada data mining antara lain yaitu model deskriptif, estimasi, prediksi, klasifikasi, klusterisasi dan asosiasi [10]. Model deskriptif akan menemukan karakteristik penting dari basis data sedangkan prediktif akan menemukan pola dari data dengan menemukan variabel lain di masa depan [11].

Salah satu jenis pola *data mining* yang digunakan dalam pengolahan data pencemaran udara di Jakarta yaitu teknik klasifikasi dengan memanfaatkan algoritma *support vector machines* (SVM). Pola ini dapat digunakan untuk berbagai jenis aktivitas seperti klasifikasi teks ataupun gambar [12]. SVM memiliki unjuk kerja yang baik untuk diimplementasikan pada bioinformatics, pengenalan, tulisan tangan, klasifikasi dan lain sebagainya [13]. Dengan klasifikasi ini kita dapat melihat pengaruh antara elemen dari data kualitas udara yang tersedia di Jakarta. Parameter yang disediakan sebagai elemen penentu kualitas udara yaitu CO, CO_2, NO_2, O_3, PM_{10}, suhu dan kelembaban. Klasifikasi ini akan memprediksi kualitas udara yang dapat menggambarkan kondisi kualitas udara saat ini berdasarkan nilai akurasi ketepatan dari model klasifikasi yang telah dibuat.

Salah satu cara untuk meningkatkan akurasi model klasifikasi adalah dengan mengatur *hyperparameter* [14]. Metode tuning yang dapat digunakan antara lain pencarian grid, pencarian acak, optimasi Bayesian, optimasi swarm partikel, dan algoritma genetika [15]. Kami memilih untuk mengoptimalkan model pembelajaran mesin klasifikasi kami menggunakan *GridSearch CrossValidation*. Ini adalah pengoptimalan parameter model pembelajaran mesin yang membangun dan mengevaluasi model untuk setiap kombinasi parameter algoritme yang ditentukan dalam kisi [16].

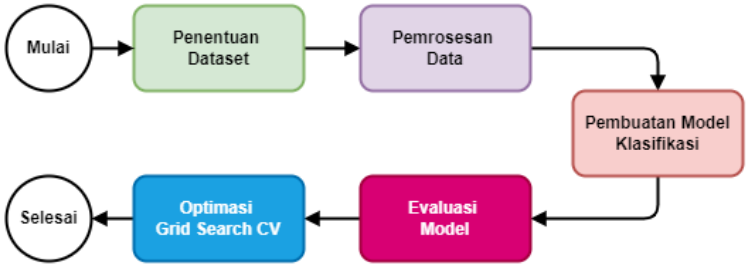
Penelitian-penelitian sebelumnya yang telah dilakukan antara lain ialah pemanfaatan SVM untuk klasifikasi kualitas udara yang dilakukan oleh Handayani [17]. SVM menghasilkan kinerja ketepatan klasifikasi dengan akurasi sebesar 99,33%. Penelitian serupa juga telah dilakukan oleh Nurjanah [7] dalam melakukan klasifikasi kualitas udara dengan memanfaatkan algoritma KNN dan telah menghasilkan kinerja ketepatan klasifikasi dengan akurasi 95.78%. Penelitian yang telah dilakukan oleh Umri [18] yaitu melakukan komparasi berbagai jenis model klasifikasi pada indeks pencemaran udara. Hasil perbandingan ini mendapatkan model dengan tingkat akurasi tertinggi yaitu *deccision tree* sebesar 99,6%. %. Penelitian yang

telah dilakukan oleh Arya [19] melakukan klasifikasi kualitas udara dengan model KNN menghasilkan nilai akurasi sebesar 94,28%.

Berdasarkan penelitian sebelumnya, nilai-nilai akurasi yang dihasilkan terlihat sangat baik dimana rata-rata angkanya di atas 90%. Nilai ini menjadi acuan kami dalam memutuskan penggunaan model klasifikasi dalam penentuan kualitas udara. Kontribusi kami dalam penelitian ini adalah melakukan kembali pembuatan model klasifikasi dengan SVM dengan teknik baru yaitu perbaikan dalam pemrosesan data hingga melakukan tuning hyperparameter. Kami melihat para peneliti sebelumnya hanya mengejar nilai akurasi yang tinggi. Atas alasan tersebut, kami mencoba menganalisis kembali penggunaan model klasifikasi kualitas udara di Jakarta dengan memanfaatkan SVM dan optimasi *hyperparameter tunning* yaitu *GridSearch CV*.

2. METODE

Metode penelitian ialah tahapan-tahapan yang dapat dilakukan untuk menyelesaikan masalah yang terjadi ketika melakukan penelitian dengan mengikuti kaidah-kaidah yang dilakukan oleh peneliti agar tujuan yang diinginkan dapat tercapai dan mendapatkan hasil yang baik [20]. Tahapan-tahapan penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. Alur Metode Penelitian

Berdasarkan Gambar 1, tahapan-tahapan penelitian pada penelitian ini terdiri dari 5 tahapan utama yaitu penentuan dataset, pemrosesan dataset, pembuatan model klasifikasi, evaluasi model dan melakukan optimasi hiperparameter dengan *grid search cv*. Tahapan-tahapan tersebut selanjutnya dijelaskan secara detail pada bagian berikut.

2.1. Dataset

Dataset yang digunakan dalam penelitian ini berasal dari situs penyedia data yaitu Jakarta Open Data dan dapat diakses pada laman https://data.jakarta.go.id/. Data diambil dari kategori lingkungan hidup dalam format *csv* yang dapat dibuka dan diolah dengan menggunakan *Microsoft Excel*. Data diambil dalam rentang waktu tahun 2020 hingga 2021 yang diambil dari 5 stasiun pemantau kualitas udata dengan total sebanyak 1829 data. Data tersebut berisi beberapa variabel yang digunakan dalam penelitian dapat dilihat pada Tabel 1 dan contoh format *dataset* dapat dilihat pada Tabel 2.

Tabel 1. ISPU Dataset Pencemaran Udara Jakarta

Variabel	Keterangan
pm10	Partikulat salah satu parameter yang diukur
so2	Sulfida (dalam bentuk SO2) salah satu parameter yang diukur
co	Carbon Monoksida salah satu parameter yang diukur
o3	Ozon salah satu parameter yang diukur
no2	Nitrogen dioksida salah satu parameter yang diukur
max	Nilai ukur paling tinggi dari seluruh parameter yang diukur dalam waktu yang sama
critical	Parameter yang hasil pengukurannya paling tinggi
categori	Kategori hasil perhitungan indeks standar pencemaran udara

Tabel 2. Format Dataset ISPU

pm10	so2	co	o3	no2	max	critical	categori
30	20	10	32	9	32	O3	BAIK
27	22	12	29	8	29	O3	BAIK
39	22	14	32	10	39	PM10	BAIK
34	22	14	38	10	38	O3	BAIK
35	22	12	31	9	35	PM10	BAIK
…	…	…	…	…	…	…	…

2.2. Pemrosesan Data

Dataset masih belum dapat digunakan dalam model klasifikasi, sehingga diperlukan upaya pengolahan data, seperti normalisasi dan pembersihan data yang tidak berarti untuk membentuk transformasi [21]. Pemrosesan data dapat dimulai dengan menangani *missing data* yaitu mengatasi beberapa data yang bersifat *nullable* atau kosong dengan cara menghapusnya. Missing data menggambarkan nilai mana yang hilang dan dapat diamati dalam kumpulan data tersebut [22].

Dataset tersebut memiliki kategori atau bisa disebut dengan class dalam klasifikasi. Pengelompokan setiap fitur yang merujuk pada suatu kelas masih bersifat string atau teks dan kita dapat mengubahnya ke dalam bentuk nominal. Sebagai contoh jika kelas adalah BAIK kita ubah menjadi kelas 1 dan seterusnya. Skema yang digunakan dalam penanganan data kategori adalah *one hot encoder*. Skema ini merupakan cara yang paling sering digunakan yang membandingkan setiap tingkat variabel kategori ke tingkat referensi tetap.

Pemrosesan data selanjutnya adalah penskalaan fitur yaitu menyatukan variabel mandiri atau rentang fitur dalam data [23]. Teknik normalization digunakan dalam penelitian ini dimana dapat melakukan normaliasi dalam kolom fitur yang dimiliki pada kisaran [0,1] dengan menggunakan *min-max scalling*. Dalam pemilihan fitur kami menggunakan *principal component analysis* (PCA) untuk mengantisipasi adanya *overfitting* akibat redudansi data. PCA berfungsi untuk mengurangi jumlah *feature* dari kumpulan data dengan mempertahankan varian sebanyak mungkin [21]. Pembagian data latih dan data uji dengan teknik *hold out* dengan membagi data dengan komposisi 75% dibanding 25% [24].

2.3. Model Klasifikasi SVM

Dalam algoritma SVM, data direpresentasikan dalam ruang *n*-dimensi di mana ia dapat memprediksi apakah instance pelatihan baru termasuk dalam kategori yang sama atau kategori yang berbeda [25]. Kumpulan data yang diubah menjadi format numerik dengan label penyandian menyederhanakan proses klasifikasi dengan SVM ini. Hal ini akan lebih jelas terlihat ketika membandingkan kategori yang sama atau berbeda yaitu antara Kelas dengan nilai 0 atau 1. Tujuan dari SVM adalah untuk menemukan *hyperplane* dalam ruang *n*-dimensi yang dapat mengklasifikasikan titik data. SVM menggunakan kernel garis lurus yang membagi dua kelas dengan persamaan linier [26] yang dapat dilihat pada persamaan 1.

$$w * x - b = 0 \tag{1}$$

Berdasarkan persamaan (1), simbol *w* adalah parameter hyperplane yang dicari, *x* untuk data input, dan *b* adalah bias. Teknik untuk menghasilkan *hyperplane* yang optimum pada SVM dapat dilakukan dengan menggunakan persamaan 2 dan 3 [27].

$$min\frac{1}{2}||\omega||^2 \tag{2}$$

$$y_i(wx_i + b) \geq 1, i = 1, \dots, \lambda \tag{3}$$

Persamaan (3) digunakan untuk memaksimalkan nilai $||\omega||^2$ dengan fokus pada pembatas $y_i(wx_i + b)$. Jika data keluaran adalah $y_i = +1$, maka pembatas menjadi $(wx_i + b) \geq 1$ dan sebaliknya jika $y_i = -1$, maka pembatas menjadi $(wx_i + b) - 1$.

2.4. Evaluasi Model

Evaluasi model SVM yang digunakan adalah *confusion matrix*. Kriteria klasifikasi dengan menggunakan confusion matrix dapat dilihat pada Tabel 3.

Tabel 3. Confusion Matrix

Kelas	Diklasifikasikan sebagai Positif	Diklasifikasikan sebagai Negatif
+	True Positive (TP)	False Positive (FP)
-	True Negative (TN)	False Negative (FN)

Tabel 3 adalah klasifikasi dari *confusion matrix*. *True positive* (TP) menunjukkan bahwa model klasifikasi dengan benar memberi label jumlah tupel positif. True Negative (TN) menunjukkan bahwa model klasifikasi dengan benar memberi label jumlah tupel negatif. False positive (FP) menunjukkan bahwa model klasifikasi memberi label yang salah untuk jumlah tupel negatif. False Negative (FN) menunjukkan bahwa model klasifikasi memberi label yang salah untuk jumlah tupel positif. Akurasi adalah ukuran kinerja model klasifikasi dan merupakan persentase jumlah data yang diprediksi dengan benar dari total data [28]. Persamaan untuk menghitung ketelitian dapat dilihat pada Persamaan 4.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (4)$$

Pengukuran *performance metrics* terdiri dari presisi, *recall* dan *f1-score* [29]. Presisi adalah rasio positif atau derajat keandalan, yaitu proporsi prediksi berlabel positif yang benar terhadap prediksi positif keseluruhan [30][12]. Persamaan untuk menghitung presisi dapat dilihat pada Persamaan 5.

$$Precission = \frac{TP}{TP + FP} \quad (5)$$

Recall juga dikenal sebagai *true positive rate* atau sensitivitas. *Recall* juga disebut sebagai derajat keandalan model dalam mendeteksi data berlabel positif dengan benar [30]. Persamaan untuk menghitung recall dapat dilihat pada persamaan 6.

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

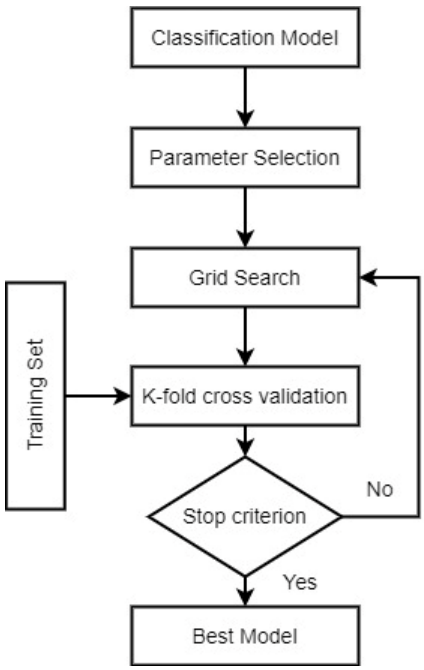
F1-score merangkum semua hasil perhitungan *precission* dan *recall* dengan membuat rata-rata harmonik [27]. Persamaan untuk menghitung *F1-score* dapat dilihat pada persamaan 7.

$$F1 - Score = \frac{2\ x\ Precission\ x\ Recall}{Precission + Recall} \quad (7)$$

2.5. Grid Search CV

Peningkatan nilai akurasi dapat dioptimalkan dengan menggunakan *cross-validation grid search* [31]. *Grid search* adalah pemilihan kombinasi model dan *hyperparameter* dengan menguji kombinasi satu per satu dan memvalidasi setiap kombinasi. Tujuan dari *grid search* adalah untuk menentukan kombinasi yang menghasilkan kinerja model terbaik yang dapat dipilih untuk digunakan sebagai model prediksi [16]. *Grid search cv* biasanya dikombinasikan dengan *k-fold cross-validation*, menciptakan indeks evaluasi untuk model klasifikasi [31].

K-fold cross-validation dapat mengulang data latih dan data uji sebanyak *k* repetisi dan pembagian 1/k dari *dataset* yang digunakan sebagai data uji [25]. Akurasi model *k* dapat diperoleh, dan kinerja model klasifikasi validasi silang *k-fold* dievaluasi berdasarkan akurasi rata-rata model *k*. Selanjutnya, parameter *classifier* diubah berdasarkan pencarian grid, dan akurasi *classifier* dihitung ulang. Proses optimasi berlanjut, seperti terlihat pada Gambar 2. Akurasi model klasifikasi dengan semua kombinasi parameter dibandingkan untuk menghasilkan nilai akurasi yang maksimal.



Gambar 2. *K-Fold Cross-Validation* dan *Grid search*

3. HASIL DAN PEMBAHASAN

3.1. Transformasi Data

Dataset telah mengalami proses transformasi data hingga layak untuk diproses dengan model klasifikasi. Dataset diubah menjadi versi numerik sehingga model klasifikasi SVM dapat memprosesnya. Pada field kategori kelas BAIK diubah menjadi angka 1, SEDANG menjadi 2 dan TIDAK SEHAT menjadi 3. Pada field critical, O3 menjadi angka 1, PM10 menjadi 2, SO_2 menjadi 3, PM_{25} menjadi 4 dan CO menjadi 5. Adapun beberapa perubahan transformasi data field categori dan critical dapat dilihat pada Tabel 4.

Tabel 4. Hasil Transformasi Dataset

index	pm10	so2	co	o3	no2	max	critical	categori
0	30	20	10	32	9	32	1	1
1	27	22	12	29	8	29	1	1
2	39	22	14	32	10	39	2	1
3	34	22	14	38	10	38	1	1
4	35	22	12	31	9	35	2	1
	…	…	…	…	…	…	…	…

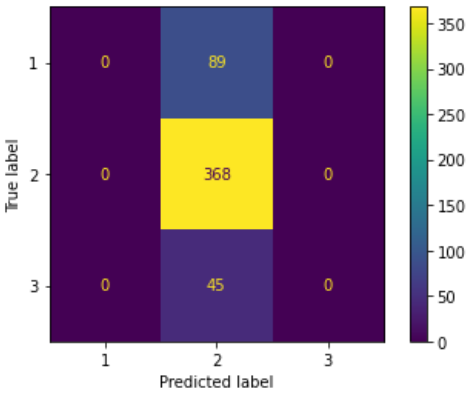
3.2. Performa SVM sebelum Optimasi

Modul seleksi memisahkan set data menjadi set data pelatihan dan pengujian. Komposisi perbandingan dataset antara data latih dan data uji adalah 75% berbanding 25%. Penerapan algoritma klasifikasi SVM kemudian diuji pada *dataset* yang telah diolah dan dipisahkan sebelumnya. Tabel 5 merupakan hasil performa pengujian *(testing)* dari 502 data uji klasifikasi algoritma SVM dengan jumlah *split k-fold* yaitu *n*=10.

Tabel 5. Performa Model SVM

Params	Class			Accuracy
	Baik	Sedang	Tidak Sehat	
Precission	0,0	0,733068	0,0	0,733068
Recall	0,0	1,000000	0,0	0,733068
F1-Score	0,0	0,845977	0,0	0,733068
Support	89,000000	368,000000	45,000000	0,733068

Performa model klasifikasi ini juga menghasilkan *confusion matrix*. Blok pada posisi diagonal merupakan nilai TP dan TN. Matrik ini dapat dilihat pada Gambar 3. Pada posisi diagonal kita bisa melihat angka 77, 358 dan 41. Matrik ini bisa kita gunakan sebagai perhitungan nilai akurasi sesuai dengan persamaan 4. Untuk mempermudah perhitungan ini, hasil confusion matrik disajikan pada Tabel 6.



Gambar 3. *Confusion Matrix*

Tabel 6. Hasil *Confusion Matrix Testing*

BAIK	SEDANG	TIDAK SEHAT	JUMLAH DATA
0	89	0	89
0	368	0	368
0	45	0	45
Akurasi : 94,8%			502

Berdasarkan Tabel 6, kita dapat menghitung nilai akurasi pengujian dengan persamaan 4 dimana nilai secara diagonal yang dijumlahkan dibagi total data uji sehingga (0 + 368 + 41)/502 = 73,31%. Tabel *confusion matrix* model SVM ini dapat dijelaskan secara detail sebagai berikut:

1. 89 data *test* untuk kelas BAIK, sistem tersebut memprediksi 0 BAIK, 89 SEDANG, dan 0 TIDAK SEHAT.
2. 368 data *test* untuk kelas SEDANG, sistem tersebut memprediksi 0 BAIK, 368 SEDANG, dan 0 TIDAK SEHAT.
3. 45 data *test* untuk kelas TIDAK SEHAT, sistem tersebut memprediksi 0 BAIK, 45 SEDANG, dan 0 TIDAK SEHAT.

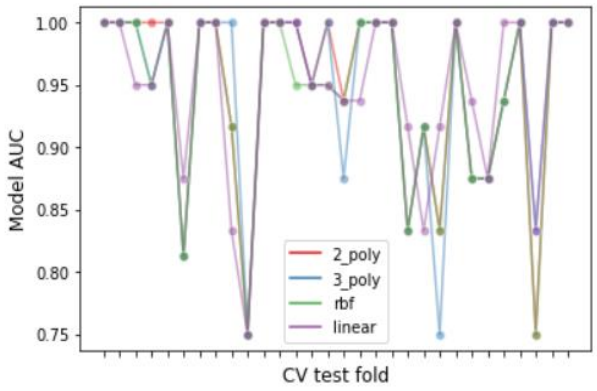
3.3. Optimasi Grid Search CV

Berdasarkan Tabel 5, akurasi yang dihasilkan oleh model klasifikasi SVM adalah 94,8%. Nilai akurasi ini masih bisa dioptimalkan untuk menghasilkan nilai yang lebih baik. Optimalisasi dimulai dengan memilih jenis *kernel* yang dapat diterapkan pada model SVM. Parameter grid yang digunakan adalah kernel linier, poli dengan derajat 2 dan 3 dan *rbf*. Validasi silang *K-fold* dijalankan pada *n_splits=10* dengan *n_repeats*=10. Pustaka pandas membuat bingkai data dari perbandingan antara *kernel* model SVM. Hasilnya adalah nilai *params*, *rank score*, dan *std score*. Tabel 7 merupakan hasil perbandingan antar *kernel* SVM.

Tabel 7. Komparasi Hyperparameter Kernel SVM

Kernel	params	rank score	std score
2 poly	{'degree': 2, 'kernel': 'poly'}	1	0,962743
3 poly	{'degree': 3, 'kernel': 'poly'}	2	0,961477
rbf	{'kernel': 'rbf'}	3	0,960190
linear	{'kernel': 'linear'}	4	0,953101

Berdasarkan Tabel 6, parameter *2 poly* menjadi rekomendasi untuk model klasifikasi SVM. Grafik pada Gambar 4 hasil perbandingan antar kernel pada model SVM dimana kedua parameter *poly* berada di atas rekomendasi.



Gambar 4. Hasil Komparasi Kernel SVM

Kami telah mendapatkan rekomendasi kernel terbaik yaitu *2 poly*. Tahap selanjutnya adalah mencari parameter terbaik dengan *grid search* cv. Tabel 8 adalah *parameter* dengan nilai param dan estimator terbaik dari model SVM.

Tabel 8. Parameter dan Estimator Terbaik SVM

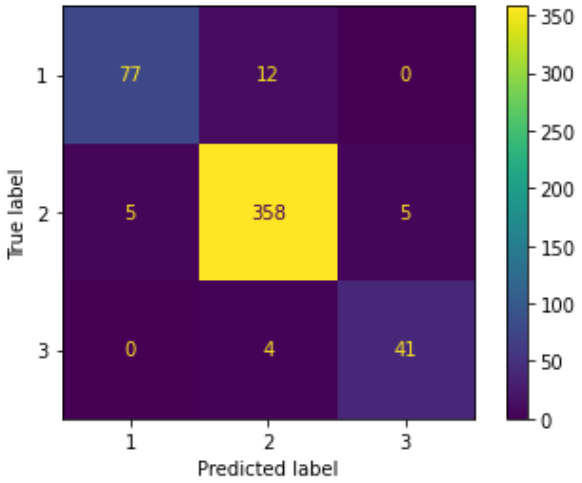
Params	Estimator
{'C': 0.1, 'gamma': 1, 'kernel': 'poly'}	SVC(C=0.1, cache_size=200, class_weight=None, coef0=0.0, decision_function_shape='ovr', degree=3, gamma=1, kernel='poly', max_iter=-1, probability=False, random_state=None, shrinking=True, tol=0.001, verbose=False)

Kami kemudian menerapkan estimator terbaik untuk model klasifikasi SVM. Hasil penerapan estimator ini dapat dilihat pada Tabel 9. Berdasarkan perbandingan antara Tabel 5 dan Tabel 9, terdapat pengaruh yang

signifikan ketika menggunakan metode optimasi *grid search cv*. Hasil akurasi meningkat 21,49%, dengan akurasi baru 94,8%. Sedangkan untuk hasil *confusion matrix* dapat dilihat pada Gambar 5 dan Tabel 10.

Tabel 9. Performa Model SVM

Params	Class			Accuracy
	Baik	Sedang	Tidak Sehat	
Precission	0,939024	0,957219	0,891304	0,948207
Recall	0,865169	0,972826	0,911111	0,948207
F1-Score	0,900585	0,964960	0,901099	0,948207
Support	89,000000	368,000000	45,000000	0,948207



Gambar 5. *Confusion Matrix* Setelah Optimasi

Tabel 10. Hasil *Confusion Matrix Testing* Setelah Optimasi

BAIK	SEDANG	TIDAK SEHAT	JUMLAH DATA
77	12	0	89
5	358	5	368
0	4	41	45
Akurasi : 94,8%			502

Berdasarkan Tabel 10, kita dapat menghitung nilai akurasi baru pada pengujian dengan persamaan 4 dimana nilai secara diagonal yang dijumlahkan dibagi total data uji sehingga (77 + 358 + 41)/502 = 94,8%. Tabel *confusion matrix* model SVM ini dapat dijelaskan secara detail sebagai berikut:

1. 89 data *test* untuk kelas BAIK, sistem tersebut memprediksi 77 BAIK, 12 SEDANG, dan 0 TIDAK SEHAT.
2. 368 data *test* untuk kelas SEDANG, sistem tersebut memprediksi 5 BAIK, 358 SEDANG, dan 5 TIDAK SEHAT.
3. 45 data *test* untuk kelas TIDAK SEHAT, sistem tersebut memprediksi 0 BAIK, 4 SEDANG, dan 41 TIDAK SEHAT.

Berdasarkan Tabel 9, nilai akurasi yang dihasilkan sudah cukup baik yaitu menghasilkan angka 94,8%, namun jika dibandingkan dengan penelitian sebelumnya maka nilai akurasi setelah dilakukan optimasi ternyata masih lebih rendah dibandingkan dengan penelitian Handayani yang mendapatkan nilai akurasi 99,33%, lebih rendah dari penelitian Umri dengan akurasi 99,6%, namun lebih tinggi dibandingkan dengan penelitian Arya yang menghasilkan nilai akurasi sebesar 94,28%.

4. KESIMPULAN

Berdasarkan hasil penelitian, ditemukan bahwa model SVM dapat digunakan sebagai alat klasifikasi untuk prediksi kualitas udara di Jakarta. Pengolahan data diperlukan dalam penelitian ini karena model SVM membutuhkan transformasi *dataset* dalam format numerik. Hasil akurasi sebelum optimasi adalah 73,31%. Optimasi dilakukan dengan membandingkan beberapa kernel yang mendukung model SVM yaitu *rbf, linear* dan *polynomial*. Perbandingan dengan parameter metode *k-fold cross-validation* yang dijalankan pada *n_splits*=10 dengan *n_repeats*=10 menghasilkan data rekomendasi kernel terbaik yaitu *poly* dengan 2 derajat.

Implementasi *kernel* terbaik menghasilkan parameter terbaik untuk model SVM yaitu *{'C': 0.1, 'gamma': 1, 'kernel': 'poly'}*. Parameter ini kemudian diterapkan kembali pada tahap klasifikasi pemantauan kualitas udara. Hasilnya adalah peningkatan nilai akurasi hingga 94,8%. Optimasi *gridsearch cv* ternyata memberikan dampak yang signifikan terhadap klasifikasi SVM, dimana terjadi peningkatan akurasi sebesar 21,49% antara sebelum dan sesudah menggunakan teknik ini. Penelitian selanjutnya adalah mencoba mengimplementasikannya dalam algoritma klasifikasi lain dan menganalisis hasilnya.

REFERENSI

[1] H. Haruna, L. Lahming, F. Amir, and A. R. Asrib, "Pencemaran Udara Akibat Gas Buang Kendaraan Bermotor Dan Dampaknya Terhadap Kesehatan," *UNM Environ. Journals*, vol. 2, no. 2, p. 57, 2019, https://doi.org/10.26858/uej.v2i2.10092.

[2] S. Machmud, "Analisis Pengaruh Tahun Perakitan Terhadap Emisi Gas Buang Kendaraan Bermotor," *J. Mesin Nusant.*, vol. 4, no. 1, pp. 21–29, 2021, https://doi.org/10.29407/jmn.v4i1.16038.

[3] A. H. R. Inaku and C. Novianus, "Pengaruh Pencemaran Udara PM 2,5 dan PM 10 Terhadap Keluhan Pernapasan Anak di Ruang Terbuka Anak di DKI Jakarta," *ARKESMAS (Arsip Kesehat. Masyarakat)*, vol. 5, no. 2, pp. 9–16, 2020, https://doi.org/10.22236/arkesmas.v5i2.4990.

[4] H. Zheng, Y. Cheng, and H. Li, "Investigation of model ensemble for fine-grained air quality prediction," *China Commun.*, vol. 17, no. 7, pp. 207–223, 2020, https://doi.org/10.23919/J.CC.2020.07.015.

[5] Badan Pengendalian Dampak Lingkungan, "Keputusan Badan pengendalian dampak lingkungan (KABAPEDAL)." pp. 13–36, 1997, https://luk.staff.ugm.ac.id/atur/sda/KEP-107-KABAPEDAL-11-1997ISPU.pdf.

[6] A. Agus, M. Ahmad, S. D. A. Kusumaningtyas, H. Nurhayati, A. N. U. Khoir, C. Sucianingsih, "Analisis Dampak Diterapkannya Kebijakan Working From Home Saat Pandemi Covid-19 Terhadap Kondisi Kualitas Udara Di Jakarta," *J. Meteorol. Klimatologi dan Geofis. Vol.6*, vol. 6, no. 3, pp. 6–14, 2019, https://jurnal.stmkg.ac.id/index.php/jmkg/article/view/141.

[7] S. Nurjanah, A. M. Siregar, and D. S. Kusumaningrum, "Penerapan Algoritma K – Nearest Neighbor (KNN) Untuk Klasifikasi Pencemaran Udara Di Kota Jakarta," *Sci. Student J. Information, Technol. Sci.*, vol. 1, no. 2, pp. 71–76, 2020, https://journal.ubpkarawang.ac.id/mahasiswa/index.php/ssj/article/view/14.

[8] S. Handoko, F. Fauziah, and E. T. E. Handayani, "Implementasi Data Mining Untuk Menentukan Tingkat Penjualan Paket Data Telkomsel Menggunakan Metode K-Means Clustering," *J. Ilm. Teknol. dan Rekayasa*, vol. 25, no. 1, pp. 76–88, 2020, https://doi.org/10.35760/tr.2020.v25i1.2677.

[9] I. S. Mangku Negara, Purwono, Purwono, and I. A. Ashari, "Analisa Cluster Data Transaksi Penjualan Minimarket Selama Pandemi," *J. Inf. Technol. Comput. Sci.*, vol. 3, no. 28, pp. 153–160, 2020, https://doi.org/10.31328/jointecs.v6i3.2693.

[10] K. Setiyanto, "Analisis Proses Data Mining Dalam Sistem Pembelajaran Berbantuan Komputer Pada Praktikum Laboratorium Sistem Informasi Universitas Gunadarma Dengan Pendekatan Machine Learning," *J. Ilm. Inform. dan Komput.*, vol. 22, no. 2, pp. 145–157, 2017, https://ejournal.gunadarma.ac.id/index.php/infokom/article/view/1735.

[11] N. Noviyanto, "Penerapan Data Mining dalam Mengelompokkan Jumlah Kematian Penderita COVID-19 Berdasarkan Negara di Benua Asia," *Paradig. - J. Komput. dan Inform.*, vol. 22, no. 2, pp. 183–188, 2020, https://doi.org/10.31294/p.v22i2.8808.

[12] R. Umar, I. Riadi, and P. Purwono, "Klasifikasi Kinerja Programmer pada Aktivitas Media Sosial dengan Metode Support Vector Machines," *CYBERNETICS*, vol. 4, no. 1, pp. 32–40, 2020, https://doi.org/10.29406/cbn.v4i01.2042.

[13] M. Ichwan, I. A. Dewi, and Z. M. S, "Klasifikasi Support Vector Machine (SVM) Untuk Menentukan TingkatKemanisan Mangga Berdasarkan Fitur Warna," *MIND J.*, vol. 3, no. 2, pp. 16–23, 2019, https://doi.org/10.26760/mindjournal.v3i2.16-23.

[14] C. G. Siji George and B. Sumathi, "Grid search tuning of hyperparameters in random forest classifier for customer feedback sentiment prediction," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 9, pp. 173–178, 2020, https://doi.org/10.14569/IJACSA.2020.0110920.

[15] E. Elgeldawi, A. Sayed, A. R. Galal, and A. M. Zaki, "Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis," *Informatics*, vol. 8, no. 4, pp. 1–21, 2021, https://doi.org/10.3390/informatics8040079.

[16] G. S. K. Ranjan, A. Kumar Verma, and S. 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, I2CT 2019*, 2019, no. March, https://doi.org/10.1109/I2CT45611.2019.9033691.

[17] A. S. Handayani, S. Soim, T. E. Agusdi, Rumiasih, and A. Nurdin, "Klasifikasi Kualitas Udara Dengan Metode Support Vector Machine," *JIRE (Jurnal Inform. Rekayasa Elektron.*, vol. 3, no. 2, pp. 187–199, 2020, http://e-journal.stmiklombok.ac.id/index.php/jire/article/view/303.

[18] S. Syihabuddin Azmil Umri, "Analisis Dan Komparasi Algoritma Klasifikasi Dalam Indeks Pencemaran Udara Di Dki Jakarta," *JIKO (Jurnal Inform. dan Komputer)*, vol. 4, no. 2, pp. 98–104, 2021, https://doi.org/10.33387/jiko.v4i2.2871.

[19] T. F. Arya, M. Faiqurahman, and Y. Azhar, "Aplikasi Wireless Sensor Network Untuk Sistem Monitoring Dan Klasifikasi Kualitas Udara," *Sistemasi*, vol. 7, no. 3, p. 281, 2018, https://doi.org/10.32520/stmsi.v7i3.312.

[20] D. N. Triwibowo, P. Purwono, I. A. Ashari, A. S. Sandi, and Y. Fadlila, "Enkripsi Pesan Menggunakan Algoritma Linear Congruential Generator (LCG) dan Konversi Kode Morse," *Bul. Ilm. Sarj. Tek. Elektro*, vol. 3, no. 3, pp. 194–201, 2022, http://journal2.uad.ac.id/index.php/biste/article/view/5546.

[21] P. Purwono, A. Wirasto, and K. Nisa, "Comparison of Machine Learning Algorithms for Classification of Drug

Groups," *Sisfotenika*, vol. 11, no. 2, p. 196, 2021, https://doi.org/10.30700/jst.v11i2.1134.
[22] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A survey on missing data in machine learning," *J. Big Data*, vol. 8, no. 1, 2021, https://doi.org/10.1186/s40537-021-00516-9.
[23] X. Wan, "Influence of feature scaling on convergence of gradient iterative algorithm," in *International Conference on Advanced Algorithms and Control Engineering*, 2019, vol. 1213, no. 3, https://doi.org/10.1088/1742-6596/1213/3/032021.
[24] P. Purwono, A. Ma'arif, I. S. Mangku Negara, W. Rahmaniar, and J. Rahmawan, "Linkage Detection of Features that Cause Stroke using Feyn Qlattice Machine Learning Model," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 7, no. 3, p. 423, 2021, https://doi.org/10.26555/jiteki.v7i3.22237.
[25] K. R. Singh, K. P. Neethu, K. Madhurekaa, A. Harita, and P. Mohan, "Parallel SVM model for forest fire prediction," *Soft Comput. Lett.*, vol. 3, no. June, p. 100014, 2021, https://doi.org/10.1016/j.socl.2021.100014.
[26] R. Umar, I. Riadi, and Purwono, "Comparison of SVM, RF and SGD Methods for Determination of Programmer's Performance Classification Model in Social Media Activities," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 2, pp. 329–335, 2020, https://doi.org/10.29207/resti.v4i2.1770.
[27] A. S. Ritonga and E. S. Purwaningsih, "Penerapan Metode Support Vector Machine (SVM) Dalam Klasifikasi Kualitas Pengelasan Smaw (Shield Metal Arc Welding)," *Ilm. Edutic*, vol. 5, no. 1, pp. 17–25, 2018, https://journal.trunojoyo.ac.id/edutic/article/view/4382.
[28] S. Katoch, V. Singh, and U. S. Tiwary, "Indian Sign Language Recognition System using SURF with SVM and CNN," *Array*, p. 100141, 2022, https://doi.org/10.1016/j.array.2022.100141.
[29] X. Xiong, S. Hu, D. Sun, S. Hao, H. Li, and G. Lin, "Detection of false data injection attack in power information physical system based on SVM–GAB algorithm," *Energy Reports*, vol. 8, pp. 1156–1164, 2022, https://doi.org/10.1016/j.egyr.2022.02.290.
[30] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognit.*, vol. 91, pp. 216–231, 2019, https://doi.org/10.1016/j.patcog.2019.02.023.
[31] T. Yan, S. L. Shen, A. Zhou, and X.-S. Chen, "Prediction of geological characteristics from shield operational parameters using integrating grid search and K-fold cross validation into stacking classification algorithm," *J. Rock Mech. Geotech. Eng.*, p. 100310, 2022, https://doi.org/10.1016/j.jrmge.2022.03.002.

BIOGRAFI PENULIS

Ahmad Toha adalah mahasiswa pasca sarjana program studi Ilmu Komputer STMIK Nusa Mandiri. Dia menyelesaikan pendidikan S1 dari STT Pelita Bangsa di jurusan Teknik Informatika dengan gelar Sarjana Teknik (S.T.). Penulis 1 memiliki peminatan pada bidang Information Technology, Computer Science, IOT, Data Science.

Purwono lahir pada 16 Mei 1989 di Banyumas Indonesia. Dia adalah lulusan Sistem Informasi Sekolah Tinggi Ilmu Komputer (STIKOM) Yos Sudarso tahun 2019. Pendidikan pasca sarjananya adalah program magister di Teknik Informatika Universitas Ahmad Dahlan (UAD). Saat ini dia sebagai dosen program studi informatika di Universitas Harapan Bangsa (UHB) Purwokerto. Bidang yang diminati adalah Data Science, Blockchain, Internet of Things

Windu Gata adalah Dosen program studi Ilmu Komputer STMIK Nusa Mandiri. Dia menyelesaikan pendidikan S1 dan S2 dari Universitas Budi Luhur dan menyelesaikan pendidikan S3 dari Universitas Negeri Jakarta. Dia memiliki peminatan pada bidang Education, Information Technology, Computer Science, IOT, Data Science. Saat ini penulis 3 menjabat sebagai Dosen Tetap Magister Ilmu Komputer STMIK Nusa Mandiri.