PERBANDINGAN KNN DAN SVM UNTUK KLASIFIKASI KUALITAS UDARA DI JAKARTA

Bryan Valentino Jayadi [1] Teny Handhayani, PhD. [2] Manatap Dolok Lauro, S.Kom.,M.M.S.I [3]

[1) 2)3)] Teknik Informatika, FTI, Universitas Tarumanagara
Jl. Letjen S. Parman No. 1, Jakarta 11440 Indonesia
[1)]email : bryan.535190019@stu.untra.ac.id [2)]email : tenyh@fti.untar.ac.id [3)]email: manataps@fti.untar.ac.id

ABSTRACT

The growth and economic development of a city is one of the factors causing air pollution because air quality has been mixed with various components of chemical compounds such as motor vehicle exhaust gases and factory smoke waste. Data mining is a method to find out information about air pollution in the city of Jakarta. The data mining method used is classification because this method can process air pollution standard index (AQI) parameter data into information that can show the level of air quality per day using the K-Nearest Neighbor algorithm and Support Vector Machine.

The result of the application of data mining for air quality classification in Jakarta is that the Support Vector Machine algorithm has better accuracy performance compared to the K-Nearest Neighbor algorithm. The Support Vector Machine algorithm uses the RBF kernel and 100 kernel parameter gets an accuracy value of 98%, precission of 97%, recall of 97%, and F1-Score of 97% while the K-Nearest Neighbor algorithm uses the number of K as much as 6 gets an accuracy value of 96%, precission of 96%, recall of 93%, and F1-Score of 94%.

Key words

Classification; AQI; K-Nearest Neighbor; Support Vector Machine

1. Pendahuluan

Pertumbuhan ekonomi yang mulai berkembang di DKI Jakarta membuat maraknya masyarakat dengan berkendara roda dua maupun roda empat, sehingga hal ini dapat menimbulkan dampak buruk bagi lingkungan. Salah satunya yaitu pencemaran udara akibat dari gas buang yang dihasilkan oleh kendaraan bermotor. Udara tersebut terbentuk dari suatu campuran gas alam yang tersusun dari banyaknya gas, diantaranya nitrogen 78%, oksigen 20%, argon 0,93% dan karbon dioksida 0,30%, dan gas gas lainnya [1].

Pencemaran udara dapat disebabkan karena alam maupun kegiatan manusia seperti aktivitas pabrik dan aktivitas kendaraan bermotor. Umumnya pada kota-kota besar memiliki tingkat kualitas udara yang lebih buruk daripada di desa, karena ramainya penduduk yang beraktifitas menggunakan kendaraan bermotor dan juga banyaknya pembangunan pabrik [2]. Urgensi dalam melakukan klasifikasi yang merupakan suatu proses untuk menemukan persamaan definisi karakteristik pada sebuah kelompok atau kelas, pada klasifikasi memilki berbagai algoritma seperti K-Nearest Neighbor dan Support Vector Machine.

Algoritma Support Machine menghasilkan tingkat akurasi peralihan yang berpegang pada fungsi kernel dan parameter yang digunakan. Terlebih lagi, metode Support Vector Machine terbagi atas dua jenis yang didasari oleh karakteristik yaitu Support Vector Machine Linier dan Support Vector Machine Non-Linier [3].

Sedangkan untuk algoritma K-Nearest Neighbor merupakan suatu metode dengan fungsi mengelompokkan objek berdasarkan data pembelajaran yaitu jarak terdekat dengan objek yang diuji [4]. Metode ini umumnya digunakan untuk menyelesaikan berbagai kasus klasifikasi seperti text categorization, pengenalan pola, peramalan, image-similarity, data visualization, pengklasifikasian hingga estimasi posisi maupun lainnya [5].

Berdasarkan penjabaran di atas dapat di simpulkan bahwa rumusan masalah penelitian adalah membandingkan kinerja algoritma klasifikasi K-Nearest Neighbor dan Support Vector Machine untuk klasifikasi data Indeks Standar Pencemaran Udara (ISPU) di DKI Jakarta serta diharapkan penelitian ini dapat memberikan informasi lebih dalam dari penelitian lain untuk menghasilkan akurasi yang paling baik di antara kedua metode tersebut. Dari hasil penelitian yang dilakukan penulis, maka dapat di simpulkan bahwa metode Support Vector Machine memiliki akurasi yang lebih baik dalam melakukan klasifikasi dibandingkan dengan K-Nearest Neighbor dengan nilai akurasi pada SVM sebesar 98% sedangkan KNN sebesar 96%.

2. Tinjauan Pustaka

2.1 Data Mining

Data mining merupakan penggalian atau pengumpulan informasi yang berguna dari kumpulan

data. Informasi yang dikumpulkan adalah pola-pola tersembunyi pada data, hubungan antar elemen data, atau pembuatan model untuk keperluan prediksi kata. Operasi dalam data mining dikelompokkan menjadi dua kategori yaitu metode deskriptif dan metode prediktif. Metode deskriptif bertujuan untuk menemukan pola, relasi, atau anomali data yang mudah di pahami oleh manusia. Sedangkan metode prediktif bertujuan untuk memperkirakan nilai suatu variabel berdasarkan nilai variabel lainnya [6].

Machine learning menyediakan teknik dasar dari data mining. Machine Learning dilakukan untuk memperoleh informasi dari data mentah dalam basis data yang dapat diwujudkan dalam bentuk yang dapat dipahami serta dapat digunakan untuk berbagai tujuan [7].

2.2 Klasifikasi

Proses klasifikasi sendiri merupakan proses untuk menemukan model atau membedakan kelas atau data yang bisa digunakan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui [8]. Dalam melakukan klasifikasi dimana dilakukan data training dan data testing menggunakan Algoritma K-Nearest Neighbor dan Support Vector Machine dengan library Python, yaitu dengan SVC Classifier dan KNN Classifier pada modul Scikit-learn. SVC Classifier merupakan library untuk melakukan klasifikasi dengan Algoritma Support Vector Machine. Sedangkan untuk KNN Classifier adalah library untuk melakukan klasifikasi dengan K-Nearest Neighbor [9].

Hasil training memberikan pengetahuan terhadap model dalam mengklasifikasikan data ke dalam kategori tingkat kualitas udara yang diuji pada 3,506 data latih dan 877 data uji. Pengujian pada kedua data latih dan data uji dilakukan untuk dapat mengetahui seberapa tingkat kualitas udara yang baik pada klasifikasi kategori tingkat kualitas udara yang mana menghasilkan klasifikasi dari tiap algoritma yang digunakan untuk kemudian dilakukan analisis dan evaluasi menggunakan Confusion Matrix.

2.3 Prepocessing Data

Data preprocessing adalah proses mengubah data mentah ke dalam bentuk yang mudah dipahami dan siap untuk digunakan untuk proses berikutnya, karena data yang berkualitas akan berdampak pada keberhasilan terhadap proyek yang melibatkan analisa data [10]. Preprocessing data yang pertama dilakukan adalah proses *data cleaning* yaitu dengan menghilangkan simbol --- pada data dan digantikan dengan kolom kosong. Setelah proses *data cleaning*, selanjutnya yaitu proses Validasi data untuk mengidentifikasi dan menghilangkan data ganjil (outlier/noise), data tidak konsisten, dan data tidak lengkap (missing value) untuk penanganan missing value menggunakan metode interpolasi [11]. *Data cleaning* adalah proses untuk membersihkan data yang missing value, menghaluskan data yang tidak pada umumnya, dan menyelesaikan data yang tidak konsisten yang terdapat di dalam dataset.

Untuk penanganan missing value dapat dilakukan dengan beberapa cara yaitu menghilangkan data tersebut, mengganti dengan variable tertentu, mengisi dengan rata-rata atribut tersebut, mengisi dengan rata-rata atribut pada kelas yang sama, ataupun melakukan regresi untuk mengganti isi data yang kosong [10].

2.4 K-Nearest Neighbor

K-Nearest Neighbor Classifier adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Data pembelajaran diproyeksikan ke ruang berdimensi banyak, yang masing-masing dimensi merepresentasikan fitur dari data. Ruang dimensi dibagi menjadi bagian-bagian berdasarkan klasifikasi data pembelajaran [12].

Nilai k yang terbaik untuk algoritma ini tergantung pada data, secara umum nilai k yang tinggi akan mengurangi efek noise pada klasifikasi, Kasus khusus untuk klasifikasi diprediksikan berdasarkan data pembelajaran yang paling dekat (dengan kata lain, k = 1) yang biasanya disebut algoritma nearest neighbor. Metode Euclidean Distance merupakan penghitungan jarak pada algoritma KNN yang paling banyak digunakan oleh peneliti [2]. Rumus Euclidean Distance dapat dilihat pada persamaan N. Rumus Euclidean Distance dapat dilihat pada Persamaan 1 [13]:

$$d(x,y)= \sqrt{\sum_{i=1}^{n}(xi - yi)^2} \quad (1)$$

Yang dimana,

D(x,y) adalah jarak antara data x ke data y
xi adalah data testing ke-i
yi adalah data training ke-i
n adalah dimensi data

Berikut ini langkah-langkah perhitungan secara manual algoritma K-Nearest Neighbor dengan rumus Euclidean Distance [14]:

Menentukan nilai K sebagai parameter banyaknya jumlah tetangga terdekat dengan objek yang akan diuji.

Menghitung jarak antara objek yang baru terhadap semua objek data yang ada di data latih. Perhitungan jarak dilakukan pada setiap baris data dengan memasukan nilai-nilai yang ada di data latih dan data uji seperti dalam Persamaan 5 [15]:

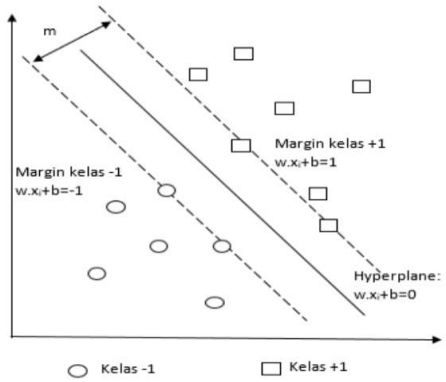
1. Melakukan pengurutan hasil perhitungan Euclidean Distance dari jarak yang terkecil sampai jarak yang terbesar.
2. Menentukan tetangga terdekat berdasarkan nilai K yang sudah ditentukan sebelumnya.

3. Menentukan kategori dari tetangga terdekat dari objek baru yang diuji.

2.5 Support Vector Machine

Support Vector Machine merupakan salah satu metode klasifikasi dengan menerapkan metode machine learning dalam memprediksi kelas berdasarkan model dari hasil proses training yang menggunakan prinsip Structural Risk Minimization. Klasifikasi umumnya digunakan pada bidang pembatas yang kemudian menggolongkan berdasarkan kelas positive dan kelas negative. Algoritma Support Vector Machine bertujuan untuk menemukan bidang pembatas terbaik untuk menjadi garis pembatas antara dua buah kelas. Pada penelitian yang dilakukan ini akan menerapkan Algoritma Support Vector Machine kernel linear selaku model prediksi [16]. Berdasarkan dari karakteristiknya, metode Support Vector Machine (SVM) dibagi menjadi dua, yaitu Support Vector Machine linier merupakan data yang dipisahkan secara linier, yaitu memisahkan kedua class pada hyperplane dengan soft margin. Sedangakan Support Vector Machine Non-Linier yaitu menerapkan fungsi dari kernel trick terhadap ruang yang berdimensi tinggi [3]. Beberapa fungsi kernel yang umum digunakan adalah kernel Linear, Sigmoid, Polynomial dan juga Radial Basic Function (RBF) [17].

Pada SVM linear menyatakan bahwa tiap data training dideklarasikan dengan (, yi), dengan i = 1,2, … N, serta x = {$xi1$, $xi2$, … . xiq} T adalah data latih ke – i yang merupakan atribut (fitur) set dan $yi \in \{-1, +1\}$ menunjukkan label pada kelas. Dapat di lihat pada Gambar 1 adalah untuk mencari hyperplane klasifikasi linear SVM [18].



Gambar 1. Margin Hyperplane

Untuk mencari hyperplane pada klasifikasi Support Vector Machine Linear dihitung menggunakan rumus Persamaan (2) [18]:

$$w.x_i + b = 0 \quad (2)$$

Pada rumus diatas, dapat dijelaskan bahwa w serta b merupakan parameter model, dimana $w.\ xi$ adalah inner-product dalam antara w dan xi , sedangkan data xi yang ada di dalam kelas -1 merupakan data yang sesuai dengan pertidaksamaan sebagai berikut seperti pada Persamaan (3) [18]:

$$w.x_i + b \leq -1 \quad (3)$$

Sedangkan data xi yang ada di dalam kelas +1 adalah data yang sesuai dengan pertidaksamaan sebagai berikut seperti pada Persamaan (4) [18]:

$$w.x_i + b \leq +1 \quad (4)$$

Dengan memberi label -1 pada kelas pertama dan +1 pada kelas kedua, sedangkan prediksi semua data uji akan memakai formula seperti pada Persamaan (5) [18]:

$$y = \begin{cases} +1, Jika\ w.z + b > 0 \\ -1, Jika\ w.z + b < 0 \end{cases} \quad (5)$$

Klasifikasi pada kelas data SVM di persamaan (1) dan (2) dapat dijadikan satu dengan notasi sebagai berikut seperti pada Persamaan (6) [18]:

$$y_i(w.x_i + b) \geq 1, i = 1,2,...N \quad (6)$$

2.6 Indeks Standar Pencemaran Udara

Indeks Standar Pencemaran Udara (ISPU) merupakan angka yang tidak mempunyai satuan dalam menggambarkan kondisi kualitas udara ambien di lokasi dan waktu tertentu untuk dapat melihat dampak terhadap Kesehatan manusia, nilai estetika dan makhluk hidup lainnya [18]. Berikut adalah tabel indeks standar pencemaran udara yang dapat di lihat pada Tabel 1 [19].

Tabel 1. Indeks Standar Pencemaran Udara

ISPU	Pencemaran Udara Level	Dampak Kesehatan
0-50	Baik	Tidak memberikan dampak bagi Kesehatan manusia atau hewan
51-100	Sedang	Tidak berpengaruh pada Kesehatan manusia ataupun hewan tetapi berpengaruh pada tumbuhan yang peka
101-199	Tidak Sehat	Bersifat merugikan pada manusia ataupun kelompok hewan yang peka atau dapat menimbulkan kerusakan pada tumbuhan dan nilai estetika
200-299	Sangat Tidak Sehat	Kualitas udara yang dapat merugikan Kesehatan pada sejumlah segmen populasi yang terpapar
300-500	Berbahaya	Kualitas udara berbahaya yang secara umum dapat merugikan Kesehatan yang serius pada populasi (misalnya iritasi mata, batuk, dahak, dan sakit tenggorokan)

Perhitungan ISPU dilakukan berdasarkan nilai ISPU batas atas, ISPU batas bawah, ambien batas atas, ambien batas bawah, dan konsentrasi ambien hasil pengukuran. Persamaan matematika perhitungan ISPU sebagai berikut seperti pada Persamaan (7) [19] :

$$I = \frac{I_a + I_b}{X_a - X_b}(X_x + X_b) + I_b \quad (7)$$

Yang dimana,

I = ISPU terhitung
Ia = ISPU batas atas
Ib = ISPU batas bawah
Xa = Konsentrasi ambien batas atas (µg/m3)
Xb = Konsentrasi ambien batas bawah (µg/m3)
Xx = Konsentrasi ambien nyata hasil pengukuran (µg/m3).

2.7 Euclidean Distance

Euclidean distance merupakan perhitungan jarak dari dua buah titik dalam Euclidean space. Euclidean space telah diperkenalkan sebelumnya oleh Euclide yang merupakan seorang matematikawan asal Yunani pada tahun 300 sebelum masehi yang mempelajari mengenai hubungan antara sudut dan jarak. Euclidean memiliki kaitan dengan Teorema Phytagoras yang umumnya digunakan pada satu, dua dan tiga dimensi yang juga secara sederhana digunakan untuk dimensi yang lebih tinggi [20]. Rumus *Euclidean Distance* dapat dilihat pada persamaan (8) [21].

$$K = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (8)$$

Keterangan :
x_1 = Sampel Data
x_2 = Data Uji/Testing
K = Jarak Data
y_1 = Sampel Data
y_2 = Data Uji/Testing

2.8 Confusion Matrix

Algoritma K-Nearest Neighbor dan Support Vector Machine merupakan salah satu algoritma *Supervised Learning* yang digunakan dalam melakukan klasifikasi. Kinerja dari *Machine Learning* di analisis dengan menggunakan *Confusion Matrix*. *Confusion Matrix* merupakan table matrix yang dimanfaatkan sebagai perhitungan performansi dari suatu model data atau algorima [2]. *Confusion matrix* yang digunakan terdapat pada Tabel 2.

Tabel 2. *Confusion Matrix* pada *Supervised Learning*

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

Pada Tabel 3 memperlihatkan *Confusion Matrix* yang digunakan dengan memiliki 4 kondisi [22], sebagai berikut:

1. True Positive: Banyaknya data yang aktual kelasnya positif dan model memprediksi positif.
2. True Negative: Banyaknya data yang aktual kelasnya negative, dan model memprediksi negative.
3. False Positive: Banyaknya data yang aktual kelasnya negative, namun model memprediksi positif.
4. False Negative: Banyaknya data yang aktual kelasnya positif, namun model memprediksi negative

Terlebih lagi, pada *confusion matrix* memiliki empat turunan jenis metode yaitu:

1. Accuracy: Total dari keseluruhan seberapa sering model benar mengklasifikasi. Permodelan persamaan matematika accuracy dapat dituliskan sebagai berikut seperti pada Persamaan (9) [2]:

$$\frac{TP + TN}{TP + FP + FN + TN} \quad (9)$$

2. Precision: Yaitu Ketika model memprediksi positif, dimana seberapa sering prediksi itu benar. Permodelan persamaan matematika accuracy dapat dituliskan sebagai berikut seperti pada Persamaan (10) [2]:

$$\frac{TP}{TP + FP} \quad (10)$$

3. Recall: Ketika kelas aktualnya positif, seberapa sering model memprediksi positif. Permodelan persamaan matematika accuracy dapat ditulisakan sebagai berikut seperti pada Persamaan (11) [2]:

$$\frac{TP}{TP + FN} \quad (11)$$

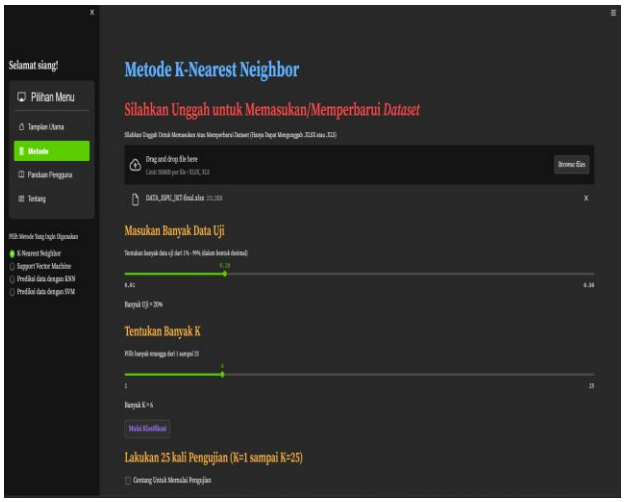
4. F1-Score: Adalah suatu rata-rata harmonic dari precision dan recall. Permodelan persamaan matematika accuracy dapat dituliskan sebagai berikut seperti pada Persamaan (12) [2]:

$$\frac{2(Recall * Precision)}{(Recall + Precision)} \quad (12)$$

3. Hasil Percobaan

Proses klasifikasi data menggunakan aplikasi berbasis web yang Klasifikasi Indeks Pencemaran Udara di Jakarta Dengan Metode K-Nearest Neighbor dan Support Vector Machine, dengan melakukan pengujian pada data indeks standar pencemaran udara yang bersumber dari Open Data dengan jumlah yang digunakan sebanyak 4383 data. https://data.jakarta.go.id/dataset?q=Indeks+Standar+Pencemaran+Udara+ISPU&sort=1
yang telah didapatkan pada bulan Januari 2010 sampai Desember 2021. Berikut merupakan langkah – langkah yang dilakukan dalam pengujian klasifikasi dengan metode K-Nearest Neighbor, yaitu sebagai berikut seperti pada Gambar 2:

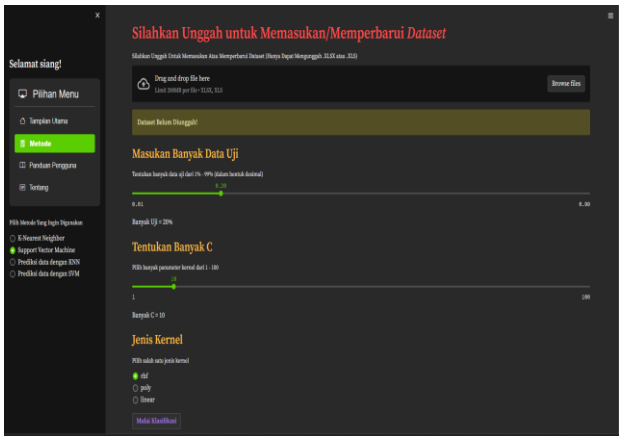
1. Upload data yang akan di gunakan dalam bentuk excel (.XLSX atau .XLS)
2. Tentukan banyak data uji
3. Tentukan banyak K (banyak tetangga)
4. Tekan tombol “Mulai Klasifikasi”



Gambar 2. Pengujian Klasifikasi Dengan Metode K-Nearest Neighbor

Selanjutnya merupakan langkah – langkah yang dilakukan dalam pengujian klasifikasi dengan metode Support Vector Machine, yaitu sebagai berikut seperti pada Gambar 3:

1. Upload data yang akan di gunakan dalam bentuk excel (.XLSX atau .XLS)
2. Tentukan banyak data uji
3. Tentukan banyak C (parameter kernel)
4. Pilih jenis kernel
5. Tekan tombol “Mulai Klasifikasi”



Gambar 3. Pengujian Klasifikasi Dengan Metode Support Vector Machine

Dataset Indeks Standar Pencemaran Udara yang terdiri dari 4383 data dalam format XLSX yang digunakan sebagai data untuk memodelkan K-Nearest Neighbor dan Support Vector Machine. Dalam membaca *dataset* menggunakan Phyton dengan *library* Pandas yaitu *library* yang digunakan untuk memproses data.

melakukan pembersihan data, memanipulasi data sampai dengan melakukan analisis data. *Dataset* ISPU yang sudah diproses dengan menggunakan Phyton dengan *library* Pandas.

Tabel 3. Hasil Percobaan Skor Akurasi K-Nearest Neighbor

No.	Jumlah K	Akurasi	Precision	Recall	F1-Score
1	2	95%	95%	90%	92%
2	3	96%	94%	92%	93%
3	4	96%	94%	92%	93%
4	5	96%	95%	92%	93%
5	6	96%	96%	93%	94%
6	7	96%	96%	92%	94%
7	8	96%	96%	92%	94%
8	9	96%	96%	90%	93%
9	10	96%	96%	91%	93%

Tabel 3 merupakan hasil Percobaan skor dari akurasi algoritma K-Nearest Neighbor mulai dari jumlah K sebanyak 2 hingga K sebanyak 10 menggunakan 3506 data latih dan 877 data uji, maka hasil akurasi tertinggi di dapatkan dengan jumlah K sebanyak 6 yaitu sebesar 96%. Hasil *confussion matrix* Algoritma KNN dengan K = 6 dapat ditampilkan pada Tabel 4.

Tabel 4. Confusion Matrix Algortima K-Nearest Neighbor Dengan K = 6

Categori	Baik	Sedang	Tidak Sehat	Total
Baik	213	15	0	228
Sedang	14	601	1	616
Tidak Sehat	0	4	29	33
Total	227	620	30	877

Tabel 4 memperlihatkan jumlah data uji untuk categori baik yaitu sebanyak 228 data, categori sedang yaitu sebanyak 616 data, categori tidak sehat yaiut sebanyak 16 data, sehingga total data uji yaitu sebesar 877 data. Adapun berdasarkan hasil klasifikasi pada Tabel 4, warna biru menyatakan bahwa algoritma KNN memprediksi secara tepat categori tingkat kualitas udara, sedangkan warna merah menunjukkan hasil prediksi yang salah. Berdasarkan Tabel 3 pengujian dilakukan dengan algoritma K-Nearest Neighbor dengan K = 6 dan data uji sebanyak 877 data, maka didapatkan nilai akurasi, *precission, recall, dan F1-Score* dengan algoritma Support Vector Machine berturut-turut menunjukkan nilai yang baik, yaitu sebesar 96%; 96%; 93%; dan 94%.

Tabel 5. Hasil Percobaan Skor Akurasi Support Vector Machine

No	Jenis Kernel	Parameter Kernel	Akurasi	Precision	Recall	F1-Score
1.	Polynomial	1.0	97%	97%	96%	97%
		10.0	97%	96%	97%	97%
		100.0	97%	96%	97%	96%
2.	RBF	1.0	96%	98%	94%	96%
		10.0	97%	96%	95%	96%
		100.0	**98%**	**97%**	**97%**	**97%**
3.	Linear	1.0	92%	93%	89%	91%
		10.0	92%	93%	89%	91%
		100.0	92%	93%	89%	91%

Tabel 5 merupakan hasil Percobaan skor dari akurasi algoritma Support Vector Machine, dan hasil akurasi tertinggi di dapatkan dengan kernel RBF dan parameter 100 menggunakan 3506 data latih dan 877 data uji didapatkan akurasi 98%. Hasil *confussion matrix* Algoritma Support Vector Machine dengan menggunakan kernel RBF dan parameter 100 dapat dilihat pada Tabel 6.

Tabel 6. *Confusion Matrix* Algoritma Support Vector Machine

Categori	Baik	Sedang	Tidak Sehat	Total
Baik	**217**	11	0	228
Sedang	5	**609**	2	616
Tidak Sehat	0	1	**32**	33
Total	227	620	30	877

Tabel 6 memperlihatkan jumlah data uji untuk categori baik yaitu sebanyak 228 data, categori sedang yaitu sebanyak 616 data, categori tidak sehat yaiut sebanyak 16 data, sehingga total data uji yaitu sebesar 877 data. Adapun berdasarkan hasil klasifikasi pada Tabel 6, warna biru menyatakan bahwa algoritma SVM memprediksi secara tepat categori tingkat kualitas udara, sedangkan warna merah menunjukkan hasil prediksi yang salah. Berdasarkan Tabel 6 pengujian dilakukan dengan algoritma Support Vector Machine dengan menggunakan kernel RBF dan parameter 100 dan data uji sebanyak 877 data, maka didapatkan nilai akurasi, *precission, recall, dan F1-Score* dengan algoritma Support Vector Machine berturut-turut menunjukkan nilai yang baik, yaitu sebesar 98%; 97%; 97%; dan 97%.

4. Kesimpulan

Penelitian ini menggunakan *dataset* indeks standar pencemaran udara yang diperoleh melalui *website* Open Data Jakarta yang tersedia secara publik sebanyak 4384 data. Dilakukan beberapa kali pengujian klasifikasi categori tingkat kualitas udara dengan menggunakan algortima K-Nearest Neighbor dan Support Vector Machine dengan 3506 data latih dan 877 data uji, untuk KNN dilakukan pengujian mulai dari K = 2 hingga K = 10 didapatkan hasil terbaik yaitu pada K = 6 dengan nilai akurasi sebesar 96%, *precision* sebesar 96% , *recall* sebesar 93% , dan *F1-Score* sebesar 94%. Untuk SVM dilakukan pengujian dengan jenis kernel polynomial, rbf, hingga linear dan masing-masing menggunakan jumlah paramater kernel 1, 10, dan 100, maka didapatkan hasil terbaik yaitu dengan jenis kernel rbf dan jumlah parameter kernel 100 memperoleh hasil akurasi sebesar 98%, *precision* sebesar 97% , *recall* sebesar 97% , dan *F1-Score* sebesar 97%.

Hal ini menunjukan bahwa klasifikasi dan prediksi model algoritma semakin baik pula. Berdasarkan hasil dari nilai akurasi, *precision*, recall, dan F1-Score dapat disimpulkan bahwa Algoritma SVM memiliki kinerja yang lebih baik daripada Algoritma KNN dalam mengklasifikasikan categori tingkat pencemaran kualitas udara. Algoritma SVM dapat membantu dalam mengklasifikasikan indeks standar pencemaran udara berdasarkan tingkat pencemaran kualiatas udara secara akurat yang dapat digunakan pada penelitian selanjutnya dan pengembangan selanjutnya.

REFERENSI

[1] Arief, Abdullah. 2014, Klasifikasi Kualitas Udara Pekanbaru Menggunakan Algoritma K-NN Dengan Euclidean Distance berdasarkan Kategori Indeks Standar Pencemaran Udara (Ispu), https://repository.uin-suska.ac.id/3462/., Tanggal akses 5 April 2023.

[2] M. Ja'far, Sodiq,. 2020, Perbandingan Metode Naive Bayes Dan K-Nearest Neighbor Pada Klasifikasi Kualitas Udara Di Dki Jakarta, http://eprints.uty.ac.id/4903/., Tanggal akses 16 Februari 2023.

[3] Sang, Adinda Inez., Sutoyo, Edi., dan Darmawan, Irfan., 2021, "Analisis Data Mining Untuk Klasifikasi Data Kualitas Udara DKI Jakarta Menggunakan Algoritma Decision Tree Dan Support Vector Machine", eProceedings of Engineering, vol. 8, no. 5.

[4] Permana, Zulfia Sari. 2021, Perbandingan Akurasi Algotima Klasifikasi Decision Tree dan K-Nearest Neighbor Pada Data Indeks Standar Pencemaran Udara (ISPU), http://repositori.unsil.ac.id/4179/., Tanggal akses 17 Februari 2023.
[5] Yuliska dan Syaliman, Khairul Umam. 2020, "Peningkatan Akurasi K-Nearest Neighbor Pada Data Index Standar Pencemaran Udara Kota Pekanbaru", IT Journal Research and Development (ITJRD) vol. 5, no. 1.
[6] Adinugroho, Sigit dan Sari, Yuita Arum. 2018, "Implementasi Data Mining Menggunakan WEKA", Edisi 1, UB Press.
[7] Written, Ian H., Frank, Eibe dan Hall, Mark A. 2011, "Practical Machine Learning Tools And Techniques", Ed. 3, Morgan Kaufmann Publishers.
[8] Yudha, Bayu Laksana., Muflikhah, Lailil., dan Wihandika, Randy. 2018, "Klasifikasi Risiko Hipertensi Menggunakan Metode Neighbor Weighted K-Nearest Neighbor (NWKNN)", Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 2, no. 2.
[9] Rindri, Yang Agita dan Fitriyani, Agus. 2023, "Analisis Perbandingan Kinerja Algoritma Multilayer Perceptron dan K-Nearest Neighbor pada Klasifikasi Tipe Migrain", Jurnal Teknologi dan Informasi (JATI) vol. 13, no. 1.
[10] Purwanto, Devi dan Honggara, Eric. 2022, "Klasifikasi Kategori Hasil Perhitungan Indeks Standar Pencemaran Udara dengan Gausian Naïve Bayes (Studi Kasus: ISPU DKI Jakarta 2020)," INSYST: Journal of Intelligent System and Computation, vol. 4, no. 2.
[11] Ridho, Ihda Innar dan Mahalisa, Galih. 2023, "Analisis Klasifikasi Dataset Indeks Standar Pencemaran Udara (ISPU) Di Masa Pandemi Menggunakan Algoritma Support Vector Machine (SVM)," Technologia Jurnal Ilmiah, vol. 14, no. 1.
[12] Liantoni, Febri. 2015, "Klasifikasi Daun Dengan Perbaikan Fitur Citra Menggunakan Metode K-Nearest Neighbor", Ultimatics: Jurnal Teknik Informatika, vol. 7, no. 2.
[13] Wijaya, Chandra., Irsyad, dan Widhiarso, 2020, "Klasifikasi Pneumonia Menggunakan Metode K-Nearest Neighbor dengan Ekstraksi GLCM", Jurnal Algoritme, vol. 1, no. 1.
[14] Umam Syaliman, Khairul., Yuliska, dan Nina Fadilah Najwa, 2022, "Seleksi Fitur Menggunakan Pendekatan K-Nearest Neighbor (K-NN)", Jurnal Sistem Informasi dan Teknologi Jaringan (SISFOTEKJAR), vol. 3, no. 1.
[15] Amalia, Adinda., Zaidiah, Ati., dan Isnainiyah, Ika Nurlaili. 2022, "Prediksi Kualitas Udara Menggunakan Algoritma K-Nearest Neighbor", Jurnal Ilmiah Penelitian dan Pembelajaran Informatika (JIPI) vol. 7, no. 2.
[16] Filemon, Bryan., Mawardi, dan Perdana, Jaya. 2022, "Penggunaan Metode Support Vector Machine Untuk Klasifikasi Sentimen E-Wallet", Jurnal Ilmu Komputer dan Sistem Informasi, vol. 10, no. 1.
[17] Rahman, Oryza Habibie., Abdillah, Gunawan., dan Komarudin, Agus. 2021, "Klasifikasi Ujaran Kebencian pada Media Sosial Twitter Menggunakan Support Vector Machine", Jurnal Resti (Rekayasa Sistem dan Teknologi Informasi), vol. 5, no. 1.
[18] Qosim, Ahmad. 2021, "Perbandingan Metode Klasifikasi Support Vector Machine (SVM) dan Naïve Bayes Classifier (NBC) Untuk Menentukan Kualitas Udara", Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim.
[19] Chaniag, Zahara, dan Ramadhani, 2020, "Indeks Standar Pencemar Udara (ISPU) Sebagai Informasi Mutu Udara Ambien di Indonesia", Kementerian Lingkungan Hidup dan Kehutanan (KLHK), Tanggal akses 15 April 2023.
[20] Setiawan, Supriyadin, Santoso, dan Buana, 2018, "Menghitung Rute Terpendek Menggunakan Algoritma A* Dengan Fungsi Euclidean Distance", Seminar Nasional Teknologi Informasi dan Komunikasi 2018 (SENTIKA 2018).
[21] Wijaya, Andy., Arisandi, Desi., dan Mulyawan, Bagus. 2020, "Pemilihan Lapangan Basket Wilayah Jakarta dengan Menggunakan Metode K-Nearest Neighbor", Jurnal Ilmu Komputer dan Sistem Informasi vol. 8, no. 1.
[22] Hadianto, Nur., Novitasari, Hafifah Bella., dan Rahmawati, Ami. 2019, "Klasifikasi Peminjaman Nasabah Bank Menggunakan Metode Neural Network", PILAR Nusa Mandiri Journal of Computing and Information System, vol. 15, no. 2.

Bryan Valentino Jayadi, mahasiswa S1, program studi Teknik Informatika Universitas Tarumanagara.