Klasifikasi Kualitas Udara menggunakan Adaptive KNN dan Weighted KNN dengan Penggunaan SMOTE-Tomek Links dan Pendekatan Bagging

1st Farrel Rassya
Fakultas Teknik Elektro
Universitas Telkom
Bandung, Indonesia
farrelrassya@student.telkomuniversity.ac.id

2nd Meta Kallista
Fakultas Teknik Elektro
Universitas Telkom
Bandung, Indonesia
metakallista@telkomuniversity.ac.id

3rd Ig. Prasetya D. Wibawa
Fakultas Teknik Elektro
Universitas Telkom
Bandung, Indonesia
prasdwibawa@telkomuniversity.ac.id

***Abstrak* — Standar Indeks Pencemaran Udara adalah angka yang menggambarkan kondisi kualitas udara pada suatu lokasi dan waktu tertentu di suatu wilayah. Parameter indeks pencemaran udara meliputi partikel (PM10), karbon monoksida (CO), sulfur dioksida (SO2), nitrogen dioksida (NO2) dan ozon (O3). Berdasarkan permasalahan yang dihadapi maka dilakukan penelitian untuk mengklasifikasikan kualitas udara untuk memahami tingkat kualitas udara. Klasifikasi kualitas udara menggunakan KNN adaptif dan KNN tertimbang dengan menggunakan metode SMOTE-Tomek link and bagging yang menggunakan penerapan teknik SMOTE-Tomek link untuk menangani masalah ketidakseimbangan berdasarkan kelas pada data. Selain itu, metode Bagging juga diterapkan untuk mengoptimalkan performa model secara keseluruhan. Dalam penelitian ini, kami membandingkan hasil pembelajaran mesin KNN adaptif dan KNN tertimbang dengan Bagging. KNN adaptif mencapai nilai presisi 85%, presisi 85%, recall 83%, dan skor F1 83%, sedangkan KNN tertimbang dan mengantongi mencapai presisi 95%, presisi 96%, recall rate 92%, dan skor F1 sebesar 93%, dan nilai rata-rata G sebesar 0,97, stratified K-fold 0,97, dan cross-validation 0,84**

***Kata kunci*— Air Classification, Adaptive KNN, Weighted KNN.**

I. PENDAHULUAN

Kualitas udara adalah suatu ukuran atau derajat mutu atau kualitas campuran gas-gas yang ada di troposfer yang diperlukan dan mempengaruhi kesehatan manusia, organisme hidup dan faktor lingkungan lain yang komposisinya tidak selalu tetap [5]. Oleh karena itu, memahami data kualitas udara sangat penting untuk menjaga kesehatan manusia dan meminimalkan upaya pengendalian polusi udara di wilayah tersebut.

Penurunan kualitas udara dapat terjadi karena beberapa aktivitas manusia seperti asap rokok, aktivitas industri, transportasi, pembakaran lahan atau hutan, dan lain-lain [1]. Faktor-faktor ini bisa menjadi penyebab terjadinya polusi udara. Polusi udara dapat menyebabkan penyakit pada manusia seperti kesulitan bernapas, kanker paru-paru, penyakit jantung, infeksi saluran pernafasan bahkan kematian [2].

Berdasarkan Air Quality Live Index (AQLI), per April 2021, DKI Jakarta menduduki peringkat ke-6 (enam) kota dengan kualitas udara paling buruk. Hal ini ditunjukkan dengan nilai AQI DKI Jakarta sebesar 156 yang masuk dalam kategori tidak sehat. Polutan utama penyebab penurunan kualitas udara adalah PM2.5 yang jumlahnya tidak boleh melebihi 10 mikron saat berada di udara. Di DKI Jakarta, polutan ini tercatat sebesar 57 mikron per meter kubik, yang menunjukkan kualitas udara di DKI Jakarta sangat buruk [6]. Oleh karena itu, dilakukan penelitian dengan tujuan untuk mengklasifikasikan tingkat kualitas udara di wilayah tersebut. .

Pada penelitian sebelumnya, prakiraan kualitas udara dilakukan menggunakan XGBoost dengan tingkat akurasi 98% dan akurasi 78%. serta prakiraan kualitas udara dengan KNN dengan akurasi 92%. Dan klasifikasi kualitas udara menggunakan Naive Bayes dan pohon keputusan yang dilakukan pada tahun 2022 menghasilkan akurasi sebesar 80% dan 63%. Sedangkan pencarian menggunakan algoritma Gaussian Naive Bayes memberikan akurasi 91%. Berdasarkan penelitian terdahulu, maka akan dilakukan penelitian tentang “Klasifikasi Kualitas Udara Menggunakan SMOTE-Tomek Linkage dan Weighted KNN serta Algoritma KNN Adaptif”.

II. KAJIAN TEORI

A. Kualitas Udara

Kualitas udara adalah ukuran atau tingkat baik buruknya suatu campuran gas yang terdapat pada lapisan troposfer yang dibutuhkan dan mempengaruhi kesehatan manusia, makhluk hidup, dan unsur lingkungan hidup lainnya yang komposisinya tidak selalu konstan [5].

B. Pencemaran Udara

Pencemaran udara adalah masuknya atau dimasukkannya zat, energi dan/atau komponen lain ke dalam udara ambien oleh kegiatan manusia, sehingga mutu udara ambien turun sampai ke tingkat tertentu yang menyebabkan udara ambien tidak dapat memenuhi fungsinya [6]. Ada berbagai macam jenis zat pencemar udara terhadap penurunan kualitas udara seperti gas pencemar yang terdiri dari:

a. Nitrogen Dioksida (NO2)
b. Karbon Monoksida (CO)
c. Sulfur Dioksida (SO2)
d. Ozon (O3)
e. Partikulat Debu (PM10)

C. Indeks Pencemaran Udara

Kualitas udara sering kali dinilai dari konsentrasi parameter polusi udara yang diukur di atas atau di bawah nilai standar kualitas udara ambien nasional. Baku mutu udara merupakan ukuran batas atau kadar zat pencemar udara yang dapat ditoleransi di udara sekitar. Udara ambien adalah udara bebas yang berada di atas permukaan bumi pada lapisan troposfer (lapisan udara setebal 16 km dari permukaan bumi) yang berada di bawah pengelolaan Negara Republik Indonesia, yang sangat diperlukan dan mempengaruhi kesehatan manusia dan hewan. dan faktor lingkungan lainnya. Baku Mutu Udara Ambien Nasional ditetapkan sebagai batas maksimum mutu udara ambien untuk mencegah pencemaran udara, sebagaimana diatur dalam PP No. 41 Tahun 1999. Pemerintah menetapkan baku mutu baku mutu udara ambien nasional untuk melindungi kesehatan dan kenyamanan masyarakat [1]. Standar kualitas udara ambien nasional didefinisikan sebagai angka bebas satuan yang menggambarkan keadaan kualitas udara ambien saat ini di lokasi tertentu, berdasarkan dampak terhadap kesehatan manusia, perawatan kosmetik, dan organisme lainnya. Meskipun nilai ISPU lebih cocok untuk perkotaan, namun pada prinsipnya dapat diterapkan pada semua jenis kawasan. Parameter yang digunakan untuk menentukan nilai ISPU dijelaskan lebih rinci pada lampiran yang dilampirkan pada Keputusan Kepala Badan Pengendalian Dampak Lingkungan Nomor 12/2014/TT-BTC. 107 Tahun 1997 tentang perhitungan, pelaporan dan pelaporan standar indeks pencemaran udara [1].

D. Machine Learning

Machine learning, yang termasuk cabang dari kecerdasan buatan, adalah adalah ilmu dan seni memprogram komputer agar dapat mempelajari pola dari data. pembelajaran mesin adalah bidang studi yang memberi komputer kemampuan untuk belajar tanpa diprogram secara eksplisit [8]. Ada banyak jenis dari *machine learning* sehingga berguna untuk mengklasifikasikannya ke dalam kategori yang luas, termasuk salah satunya adalah *unsupervised* and *supervised learning* [8].

E. K- Nearest Neighbors

Algoritma *Machine learning K-Nearest Neighbor* merupakan salah satu metode algoritma untuk klasifikasi objek berdasarkan data yang jaraknya paling dekat dengan objek tersebut. Secara umum digunakan untuk mendefinisikan jarak antara dua objek menggunakan persamaan *Euclidean Distance* [17]. Persamaan *Euclidean Distance* dilihat pada persamaan (1).

$$d(x,y) = \sqrt{\sum_{i=1}^{p}(Xi - Yi)^2} \qquad (1)$$

dimana,

Xi = Sampel Data
yi = Data Uji
i = Data Variable
d = Jarak
p = Dimensi pada Data

F. Weighted K-Nearest Neighbors

Weighted K-Nearest Neighbors adalah variasi dari algoritma tradisional *K-Nearest Neighbors* yang menetapkan *weight* berbeda ke tetangga berdasarkan jarak mereka dari titik kueri. KNN adalah algoritma klasifikasi sederhana yang mengklasifikasikan titik data berdasarkan kelas mayoritas dari k tetangga terdekatnya. Di *Weighted KNN*, pengaruh suara setiap tetangga pada klasifikasi akhir disesuaikan menurut kedekatannya dengan titik kueri [17]. Dalam kasus paling sederhana, bobotnya bisa proporsional kebalikan dari jarak antara dua vektor seperti pada persamaan (2).

$$\check{y} = arg\,max \sum_{i=1}^{k} \frac{Wi}{di} \circ (y1 = y) \qquad (2)$$

G. Adaptive K- Nearest Neighbors

Adaptive K-Nearest Neighbors atau Adaptive KNN adalah variasi dari algoritma K-Nearest Neighbors tradisional. Ide dibalik algoritma Adaptive KNN adalah untuk memperkirakan jarak secara adaptif. Kemudian, menemukan superset yang berisi k tetangga terdekat dan direduksi menjadi masalah "multi-armed bandit" dengan tujuan mengidentifikasi satu set ukuran *k+h* yang berisi k lengan terkecil. Kompleksitas sampel dari algoritma ini mendekati kompleksitas komputasionalnya, yang memungkinkan untuk memberikan batasan atas pada kompleksitas komputasional algoritma ini dengan membuktikan batasan atas pada kompleksitas sampel. Dalam hal yang sama, juga bisa membuktikan batasan bawah bagi semua algoritma yang menggunakan estimasi jarak yang adaptif [18].

H. Synthetic Miniority Oversampling Technique

Synthetic Minority Oversampling Technique (SMOTE) merupakan metode oversampling kelas membuat data buatan (sintetis) untuk memecahkan masalah ketidakseimbangan kelas data[12]. Permasalahan ketimpangan kelas menjadi permasalahan penting untuk diatasi. Ketidakseimbangan kelas kondisi dimana jumlah instance kelas mayoritas lebih banyak dibandingkan dengan jumlah instance kelas minoritas. SMOTE adalah teknik oversampling yang digunakan untuk menghindari penurunan kinerja pengklasifikasi yang disebabkan oleh ketidakseimbangan kelas dalam kumpulan data[14]. SMOTE bekerja dengan "secara sintetis" membuat instance baru dari kelas minoritas. Sebelum melakukan *SMOTE*, data dari kelas minoritas dipilih dan ditentukan jumlah k tetangga terdekatnya.

Kemudian nearest neighbor dipilih berdasarkan jarak euclidean antara kedua data. maka jarak *euclidean* $d(x, y)$ adalah pada persamaan (3) berikut

$$d(x,y) = \sqrt{(\mathrm{x1} - \mathrm{y1})^2 + \ldots + (\mathrm{xn} - \mathrm{yn})^2} \quad (3)$$

I. Tomek Links

Tomek Links merupakan salah satu modifikasi dari teknik *undersampling Condensed Nearest Neighbors* (CNN) yang dikembangkan oleh Tomek pada tahun 1976 [15]. *Tomek Links* digunakan untuk mengidentifikasi item data kelas mayoritas yang akan dihapus. *Tomek Links* terjadi antara dua item data yang memiliki kelas berbeda, tetapi merupakan tetangga terdekat satu sama lain [16]. Proses *Tomek Links* yang ditunjukkan pada Gambar 2.1 mengidentifikasi item data kelas mayoritas yang akan dihapus. Titik lingkaran hitam adalah kelas mayoritas dan titik lingkaran terarsir adalah kelas minoritas. *Tomek Links* menghapus data kelas mayoritas yang memiliki kesamaan karakteristik dengan kelas minoritas pada garis jarak minimum sehingga data menjadi seimbang [16].

J. SMOTE-Tomek Links

Kombinasi SMOTE dan Tomek Links pertama kali diperkenalkan oleh Batista pada tahun 2003. Metode ini menggabungkan kemampuan SMOTE untuk menghasilkan data sintetis untuk kelas minoritas dan kemampuan Tomek Links untuk membersihkan data yang dihasilkan.Diidentifikasi sebagai Tomek Links dari kelas mayoritas, yaitu data sampel dari mayoritas. kelas yang paling dekat dengan data kelas minoritas [16]. Proses penggabungan SMOTE dan Tomek Links merupakan langkah awal SMOTE yang dimulai dengan menambah jumlah observasi pada kelas minoritas dengan membuat fitur atau observasi sintetik yaitu fitur baru yang tidak ada pada dataset tetapi mirip dengan objek di dalamnya. kumpulan data. Himpunan data. . Observasi sintetik terbentuk dari dua observasi, yaitu observasi pertama yang dipilih dari data kelas minoritas dan observasi kedua dari data kelas minoritas yang dipilih secara acak dengan k tetangga terdekat dari observasi pertama kelas minoritas. Dengan observasi agregat tersebut maka jumlah observasi pada data kelas minoritas ditambah agar lebih seimbang dengan data kelas mayoritas. Kemudian, identifikasi *Tomek Links* pada data hasil *SMOTE*. Sepasang observasi disebut *Tomek Links* jika kedua observasi tersebut bertetangga terdekat tetapi mempunyai kelas yang berbeda. Pasangan observasi yang diidentifikasi sebagai Tomek Links kemudian dihapus dari kumpulan data. Identifikasi *Tomek Links* diulangi hingga dihasilkan data bebas noise [16].

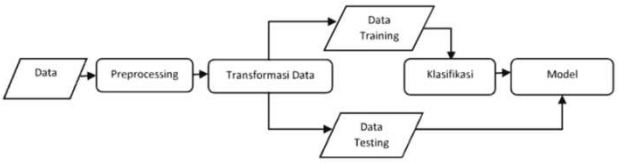
K. Bagging Aggregation

Agregasi bagging, umumnya dikenal sebagai Bagging, adalah teknik pembelajaran ansambel yang digunakan untuk meningkatkan performa model pembelajaran mesin, termasuk mengurangi varians dan overfitting. Bagging melibatkan pembuatan beberapa subset data pelatihan menggunakan pengambilan sampel acak dengan penggantian dan pelatihan model dasar pada masing-masing subset tersebut. Prediksi model dasar ini mengikuti digabungkan untuk menghasilkan prediksi akhir yang lebih akurat dan andal dibandingkan dengan model tunggal [18]

III. METODE

A. Gambaran Umum Sistem

Sistem klasifikasi prediksi Indeks Pencemaran Udara Standar (SIPU) menggunakan metode KNN tertimbang dan KNN adaptif dengan gambaran umum langkah-langkah untuk memperoleh hasil akhir. Untuk lebih jelasnya dapat dilihat diagram prosesnya pada Gambar 1.



GAMBAR 1
(Diagram alur sistem klasifkasi udara)

B. Dataset

Dataset yang didapatkan dari online website data Jakarta yang ditunjukkan oleh https://data.jakarta.go.id dari tahun 2017 sampai 2021. Dataset tersebut memiliki atribut parameter ‘Tanggal’, ‘Wilayah’, ‘PM10’, ‘SO2’, ‘C0’, ‘O3’, ‘NO2’, ‘Max’, ‘Critical’, ‘Kategori’. Dataset tersebut memiliki total 8076 baris dan 8 kolom. Pada dataset tersebut memiliki 8 atribut parameter yaitu:

1. Tanggal: Waktu pengukuran kualitas udara
2. PM10: Partikulat berukuran 10 mikron
3. SO2: Sulfur Oksida
4. CO: Karbon Monoksida
5. O3: Ozon
6. NO2: Nitrogen dioksida
7. Max: Parameter pengukuran indeks paling tinggi
8. Kategori: Kategori hasil perhitungan indeks standar pencemaran udara dan mencakup sebagai nilai dari kelas

C. Pra-pemrosesan Data

Pada tahap ini, dataset diproses menggunakan perpustakaan Pandas. Tujuan pengolahan data ini adalah untuk mengubah jenis kategori untuk setiap parameter yang tersedia. Parameter 'Area', 'Penting' dan 'Kategori' selalu bertipe objek dan akan diubah menjadi tipe Kategori. Kemudian, temukan nilai dalam bingkai data yang berisi nilai yang hilang dan hilangkan. Selanjutnya, beri label parameter “Kategori” dengan encoder label.

D. Transformasi Data

Transformasi data dimulai dengan memisahkan dua jenis data, yaitu data numerik dan data kategorikal. Pemisahan ini memungkinkan untuk menganalisis fungsi-fungsi yang akan digunakan pada parameter model pembelajaran mesin yang akan dibuat. Dalam kumpulan data ini, data numerik mencakup “PM10”, “SO2”, “CO”, “O3” dan “NO2”. sedangkan pada klasifikasi, yang digunakan untuk parameter model pembelajaran mesin adalah “Kategori”. Karena adanya ketidakseimbangan jumlah tiap kelas pada parameter “Kategori”, maka dilakukan downsampling dan upsampling menggunakan link SMOTE-Tomek. Setelah jumlah kelas pada label “Kategori” seimbang, kumpulan data akan dibagi untuk membangun model algoritma ini dengan perbandingan 10:90, 20:80, 30:70, 40:60, 50:50, 60:40, 70:30 dan 20:80.

IV. HASIL DAN PEMBAHASAN

A. Weighted KNN + Bagging

Pada model pembelajaran mesin *Weighted KNN* dan *SMOTE-Tomek Links*, dilakukan eksperimen dengan menggunakan sepuluh data test dan training 10:90, 20:80, 30:70, 40:60, 50:50, 60:40, 70:30, dan 80:20 dengan nilai k=1,2,3,4,5,6,7,8,9,10. Hasil pembuatan model pembelajaran mesin dengan *Weighted KNN* dan *Bagging* dengan data uji dan data latih 10:90 dengan nilai k = 1,2,3,4,5,6,7,8,9,10.

TABEL 1
(Matriks evaluasi *Weighted KNN dan Bagging* dengan data test dan training 10:90)

Nilai K	Akurasi	Presisi	*Recall*	*F-Measure*
1	0.932	0.941	0.932	0.931
2	0.9324	0.941	0.935	0.938
3	0.942	0.950	0.940	0.940
4	0.944	0.904	0.900	0.910
5	0.943	0.900	0.939	0.918
6	0.914	0.904	0.940	0.900
7	0.942	0.900	0.940	0.917
8	0.944	0.903	0.942	0.919
9	0.950	0.945	0.945	0.920
10	0.940	0.900	0.903	0.919

nilai K tertinggi diperoleh adalah K=9. Dengan akurasi 95%, Presisi 94%, *Recall* 94% serta *F1* bernilai 92%. Akurasi sebesar 95% menunjukkan bahwa model berhasil memprediksi sebagian besar data dengan benar. Nilai Presisi 94% membantu meminimalisir *false positif*. Nilai recall 94% -- sejauh mana model dapat meminimalkan *false negatif*. Nilai *F1 Score* 92% bahwa model mencapai keseimbangan yang baik antara presisi dan *recall*.

TABEL 2
(Matriks evaluasi *Weighted KNN dan Bagging* dengan data test dan training 20:80)

Nilai K	Akurasi	Presisi	*Recall*	*F-Measure*
1	0.939	0.920	0.937	0.932
2	0.939	0.940	0.937	0.943
3	0.944	0.954	0.942	0.940
4	0.945	0.933	0.944	0.938
5	0.942	0.930	0.942	0.936
6	0.942	0.932	0.940	0.35
7	0.940	0.930	0.941	0.930
8	0.940	0.930	0.941	0.935
9	0.950	0.947	0.949	0.944
10	0.941	0.935	0.937	0.93

TABEL 3
(Matriks evaluasi *Weighted KNN dan Bagging* dengan data test dan training 30:70)

Nilai K	Akurasi	Presisi	*Recall*	*F-Measure*
1	0.933	0.932	0.904	0.910
2	0.933	0.930	0.903	0.920
3	0.940	0.940	0.909	0.928
4	0.940	0.950	0.909	0.929
5	0.939	0.951	0.921	0.935
6	0.940	0.952	0.922	0.930
7	0.939	0.953	0.920	0.936
8	0.949	0.954	0.912	0.940
9	0.950	0.964	0.920	0.930
10	0.937	0.953	0.881	0.90

Pada tabel 3, Matriks evaluasi *Weighted KNN dan Bagging* dengan data test dan training 20:80. nilai K terbaik didapatkan adalah K=9. Dengan akurasi 95%, Presisi 94%, *Recall* 94% serta *F1* bernilai 94%. Akurasi 94% menunjukkan bahwa model berhasil memprediksi sebagian besar data dengan benar. Nilai Presisi 94%, artinya model berhasil meminimalkan nilai *false positif*. Nilai recall 94% bahwa model dapat meminimalkan *false negatif*. Nilai *F1 Score* 94% bahwa model mencapai keseimbangan yang baik antara presisi dan *recall*. Dan pada tabel 4, Matriks evaluasi *Weighted KNN dan Bagging* dengan data uji dan data latih sebanyak 30:70. nilai K terbaik diperoleh adalah K=9. Dengan akurasi 95%, Presisi 96%, Recall 92% serta F1 bernilai 93%. Akurasi 95% menunjukkan bahwa model berhasil memprediksi sebagian besar data dengan tepat. Nilai Presisi 96% meminimalkan nilai *false positif*. Nilai recall 92% bahwa model dapat meminimalkan *false negatif.* Nilai F1 Score 93% bahwa model mencapai keseimbangan yang baik antara presisi dan recall. Model dengan K = 9 memiliki kemampuan baik dalam meminimalkan *false* positif dan *false* negatif, serta memberikan keseimbangan yang baik antara presisi dan *recall*.

TABEL 4
(Matriks evaluasi *Weighted KNN dan Bagging* dengan data test dan training 40:60)

Nilai K	Akurasi	Presisi	*Recall*	*F-Measure*
1	0.933	0.931	0.920	0.925
2	0.930	0.936	0.920	0.928
3	0.930	0.95	0.913	0.930
4	0.940	0.940	0.910	0.93
5	0.940	0.95	0.88	0.92
6	0.942	0.950	0.89	0.92
7	0.941	0.95	0.900	0.92
8	0.942	0.95	0.90	0.92
9	0.944	0.96	0.90	0.93
10	0.94	0.95	0.87	0,90

Pada Tabel 4, Matriks evaluasi *Weighted KNN dan Bagging* dengan data uji dan data pelatihan 40:60. nilai K terbaik diperoleh adalah K=9. Dengan akurasi 94%, Presisi 96%, Recall 90% serta F1 bernilai 93%. Akurasi 94% menunjukkan bahwa model berhasil memprediksi sebagian besar data dengan benar. Nilai Presisi 96% meminimalkan *false positif*. Nilai recall 90% bahwa model dapat meminimalkan *false negatif*. Nilai *F1 Score* 93% bahwa

model mencapai keseimbangan yang baik antara presisi dan recall.

TABEL 5
(Matriks evaluasi *Weighted KNN dan Bagging* dengan data test dan training 50:50)

Nilai K	Akurasi	Presisi	*Recall*	*F-Measure*
1	0.932	0.933	0.920	0.920
2	0.936	0.95	0.920	0.935
3	0.941	0.955	0.925	0.934
4	0.943	0.958	0,925	0,940
5	0.944	0.96	0.925	0.941
6	0.945	0.951	0.926	0.935
7	0.946	0.961	0.908	0.933
8	0.948	0.961	0.907	0.932
9	0.950	0.965	0.910	0.933
10	0.946	0.946	0.918	0.934

Pada Tabel 5, Matriks evaluasi *Weighted KNN* dan *Bagging* dengan data uji dan data latih sebesar 50:50. nilai K terbaik didapatkan adalah K=9. Dengan akurasi 95%, Presisi 96%, *Recall* 90% serta F1 bernilai 93%. Akurasi 95% menunjukkan bahwa model berhasil memprediksi sebagian besar data dengan benar. Nilai Presisi 96% meminimalkan nilai *false positif*. Nilai *recall* 90% bahwa model dapat meminimalkan *false negatif* pada data uji yang diberikan. Nilai *F1 Score* 93% bahwa model mencapai keseimbangan yang baik antara presisi dan recall.

TABEL 6
(Matriks evaluasi *Weighted KNN dan Bagging* dengan data test dan training 60:40)

Nilai K	Akurasi	Presisi	*Recall*	*F-Measure*
1	0.929	0.921	0.904	0.913
2	0.933	0.930	0.920	0.915
3	0.939	0.937	0.907	0.921
4	0.941	0.940	0.918	0.928
5	0.945	0.96	0.925	0.941
6	0.943	0.951	0.908	0.928
7	0.944	0.952	0.900	0.926
8	0.944	0.962	0.909	0.934
9	0.953	0.970	0.909	0.939
10	0.945	0.962	0.890	0.923

Pada Tabel 6, Matriks evaluasi *Weighted KNN dan Bagging* dengan data uji dan data latih sebesar 60:40. nilai K terbaik didapatkan adalah K=9. Dengan akurasi 95%, Presisi 97%, Recall 90% serta *F1* bernilai 93%. Akurasi 95% menunjukkan bahwa model berhasil memprediksi sebagian besar data dengan benar. Nilai Presisi 97% meminimalkan hasil *false positif*.

TABEL 7
(Matriks evaluasi *Weighted KNN dan Bagging* dengan data test dan training 70:30)

Nilai K	Akurasi	Presisi	*Recall*	*F-Measure*
1	0.92	0.91	0.84	0.87
2	0.932	0.91	0.85	0.88
3	0.933	0.92	0.85	0.89
4	0.934	0.941	0.857	0.88
5	0.932	0.941	0.85	0.88
6	0.931	0.94	0.85	0.89
7	0.933	0.94	0.85	0.89
8	0.933	0.94	0.85	0.89
9	0.933	0.94	0.85	0.89
10	0.933	0.94	0.84	0.88

Pada Tabel 7, Matriks evaluasi *Weighted KNN dan Bagging* dengan *data uji dan data latih sebanyak* 70:30. nilai K terbaik didapatkan adalah K=9. Dengan akurasi 93%, Presisi 94%, Recall 85% serta *F1 score* bernilai 89%. Akurasi 94% menunjukkan bahwa model berhasil memprediksi sebagian besar data dengan benar. Nilai Presisi 94% meminimalkan nilai false positif. Nilai recall 85% bahwa model dapat meminimalkan hasil false negatif.

TABEL 8
(Matriks evaluasi *Weighted KNN dan Bagging* dengan data test dan training 80:20)

Nilai K	Akurasi	Presisi	*Recall*	*F-Measure*
1	0.920	0.900	0.85	0.87
2	0.921	0.901	0.86	0.88
3	0.92	0.91	0.85	0.88
4	0.932	0.922	0.86	0.89
5	0.932	0.939	0.864	0.89
6	0.932	0.932	0.86	0.89
7	0.932	0.942	0.86	0.89
8	0.931	0.941	0.86	0.90
9	0.931	0.950	0.86	0.90
10	0.930	0.950	0.84	0.88

Pada tabel 8, Matriks evaluasi *Weighted KNN dan Bagging* dengan data uji dan data latih sebesar 80:20. nilai K terbaik didapatkan adalah K=9. Dengan akurasi 93%, Presisi 95%, Recall 86% serta F1 bernilai 90%. Akurasi 92% menunjukkan bahwa model berhasil memprediksi sebagian besar data dengan benar. Nilai Presisi 95% meminimalkan *hasil false positif pada algoritma yang dibuat*. Nilai recall 86% bahwa model dapat meminimalkan hasil false negatif. Nilai *F1 Score* 90% bahwa model mencapai keseimbangan yang baik antara presisi dan recall. Model dengan K = 9 memiliki kemampuan baik dalam meminimalkan *false positif* dan *false negatif*, serta memberikan keseimbangan yang baik antara presisi dan recall.

TABEL 9
(*Matriks* evaluasi *Weighted KNN dan Bagging* dengan K=9)

Data Splitting	Akurasi	Presisi	*Recall*	*F-Measure*
10:90	95 %	94 %	94 %	92 %
20:80	95 %	94 %	94 %	94 %
30:70	95 %	96 %	92 %	93 %
40:60	94 %	96 %	90 %	93 %
50:50	95 %	96 %	90 %	93 %
60:40	95 %	97 %	90 %	93 %
70:30	93 %	94 %	85 %	89 %
80:20	93 %	95 %	86 %	90 %

Pada tabel 9, matriks evaluasi tersebut menunjukkan hasil kinerja performa *Weighted KNN* dan *Bagging* dengan nilai K = 9. pada berbagai skenario pembagian data yang berbeda, yang disusun berdasarkan perbandingan data uji dan data latih. Hasil pada pembagian data 10:90 menunjukkan performa yang baik dengan akurasi, presisi, dan recall yang tinggi. *F-Measure* yang cukup tinggi 92% menunjukkan bahwa model mampu mencapai keseimbangan nilai antara presisi dan recall. Performa model pada pembagian data 20:80 sangat mirip dengan hasil pada pembagian data 10:90, dengan semua metrik evaluasi yang tinggi. Ini menunjukkan konsistensi dalam performa model dengan perubahan proporsi data uji dan latih. Pada pembagian data 30:70, model memiliki tingkat presisi yang lebih tinggi, tetapi recall yang sedikit rendah. Ini mungkin menunjukkan bahwa model lebih baik dalam mengidentifikasi ragam nilai uji yang bernilai instance positif yang sebenarnya (tingkat presisi yang tinggi), tetapi beberapa instance positif dapat terlewatkan (recall yang lebih rendah).

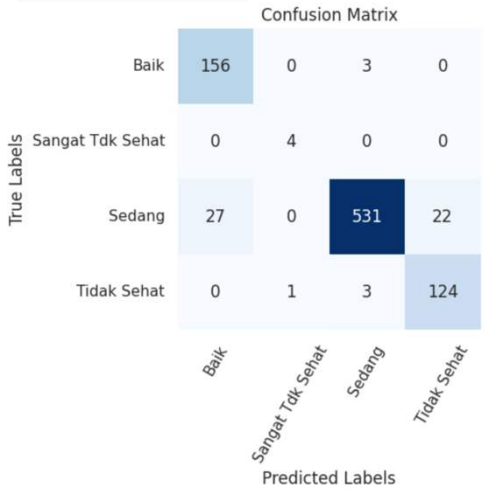
TABEL 10
(Tabel *G-Mean, Stratified KFold dan Cross Validation* pada Weighted KNN dan Bagging.)

Data Splitting	*G-Mean*	*Stratified KFold*	*Cross Validation*
10:90	0.971	0.976	0.884
20:80	0.970	0.974	0.884
30:70	0.959	0.975	0.884
40:60	0.959	0.973	0.884
50:50	0.960	0.973	0.884
60:40	0.959	0.970	0.902
70:30	0.946	0.967	0.902
80:20	0.947	0.966	0.902

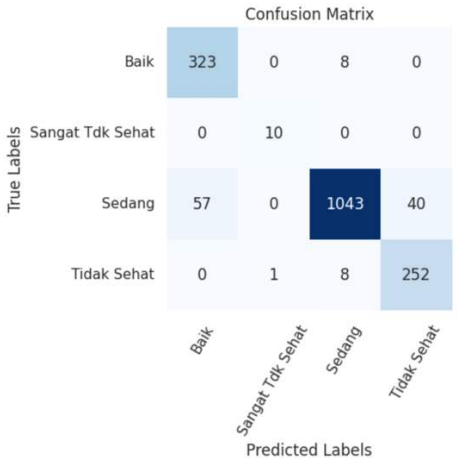
Selain matriks evaluasi berupa akurasi, presisi, recall, dan *nilai F1*. Nilai *G-Mean* dapat dijadikan bahan pertimbangan untuk menilai pembagian data mana yang terbaik. Selain *G-mean*, Nilai pada matrik *Stratified KFold* dan *Cross Validation* juga dapat digunakan untuk mengukur parameter model machine learning tersebut. *Stratified KFold* adalah metode validasi silang yang membagi kumpulan data menjadi beberapa bagian dengan tetap menjaga proporsi kelas yang seimbang. Sedangkan validasi silang adalah teknik yang digunakan untuk mengukur kinerja model yang dibyat dengan membagi kumpulan data menjadi bagian-bagian yang berbeda, mengambil satu bagian sebagai data uji, dan yang lainnya sebagai data latih. Validasi silang dilakukan untuk membantu menghindari hasil galat yang tinggi dan memberikan perkiraan yang lebih baik tentang bagaimana model akan berperforma pada data independen.

Pada Tabel 10, didapatkan tingkat pembagian data 10:90 hingga 50:50 mendapatkan nilai *G-Mean* yang stabil sekitar 0.96 hingga 0.97. nilai tersebut menunjukkan bahwa performa model pada tingkat pembagian data ini sangat baik dalam mengenali kelas positif dan negatif dengan seimbang. Ketika tingkat pembagian data meningkat menjadi 60:40 hingga 80:20, *G-Mean* memiliki sedikit penururan. Ini mungkin menunjukkan bahwa performa model sedikit menurun dalam hal keseimbangan antara sensitivitas dan spesifisitas pada tingkat pembagian data yang lebih besar. *Stratified KFold* memiliki nilai yang cukup tinggi, menunjukkan bahwa penggunaan validasi silang dengan mempertahankan proporsi kelas yang seimbang memberikan hasil yang baik. Nilai pada *Stratified KFold* yang tinggi mengindikasikan bahwa penggunaan *SMOTE-Tomek links* berhasil dalan menangani ketidakseimbangan data. *Cross Validation* memiliki nilai yang cukup stabil di seluruh tingkat pembagiand data, menunjukkan bahwa model memiliki kemampuan yang baik dalam menggeneralisasi hasilnya pada data uji yang berbeda. Berdasarkan hasil matrik evaluasi dari Tabel 4.11 dan juga tabel *G-mean, Stratified KFold* dan juga validasi silang pada Tabel 4.12 dapat disimpulkan bahwa pembagian data yang terbaik adalah 10:90, 20:80, dan juga 30:70.

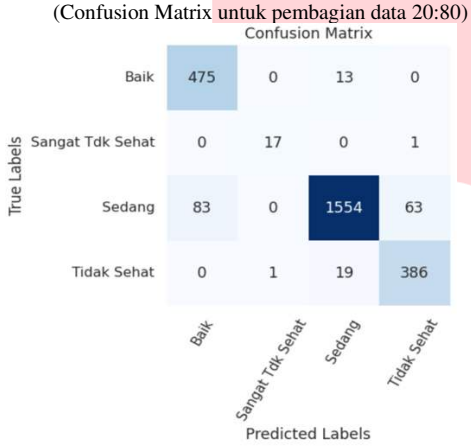
Selain mengukur model dengan matrik evaluasi dan *G-mean, Stratified KFold* dan juga *Cross Validation*, pengukuran hasil model pembelajaran mesin bisa diamati dengan *confusion matrix*. Konfusi Matrik dapat digunakan ketika menggabungkan hasil prediksi kelas dari suatu kumpulan data dengan merangkum kinerja klasifikasi menggunakan tabulasi silang antara kelas aktual dan kelas prediksi. Kemudian akan dibandingkan hasil *konfusi Matrik* dari pembagian data terbaik yaitu 10:90, 20:80, dan 30:70.



GAMBAR 2
(Confusion Matrix untuk pembagian data 10:90)



GAMBAR 3
(Confusion Matrix untuk pembagian data 20:80)



GAMBAR 4
(Confusion Matrix untuk pembagian data 30:70)

Pada model *machine learning Adaptive KNN*, dilakukan percobaan dengan dengan data test dan training 10:90, 20:80, 30:70, 40:60, 50:50, 60:40, 70:30, dan 80:20 yang sudah menggunakan SMOTE-Tomek links dengan nilai kmax = 20. Berbeda dengan model *machine learning Weighted KNN*, Model *Adaptive KNN* mencari nilai K terbaik. *Adaptive KNN* beradaptasi mendapatkan nilai K terbaik yaitu 9.

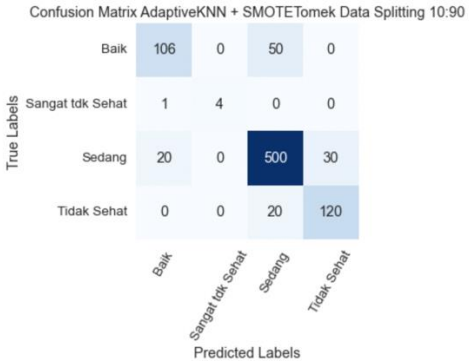
B. Adaptivve Knn dan bagging

TABEL 11
(Tabel Evaluasi Matrik model *Adaptive KNN dan Bagging*.)

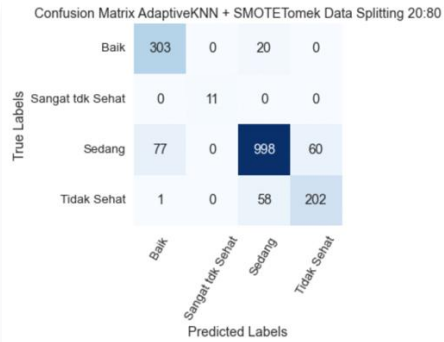
Data Splitting	Akurasi	Presisi	*Recall*	*F-Measure*
10:90	0.833	0.779	0.891	0.849
20:80	0.803	0.879	0.891	0.843
30:70	0.773	0.899	0.847	0.853
40:60	0.767	0.809	0.839	0.843
50:50	0.759	0.799	0.825	0.803
60:40	0.760	0.802	0.830	0.800
70:30	0.750	0.810	0.826	0.803
80:20	0.748	0.820	0.820	0.800

Pada Tabel 11, didapatkan nilai evaluasi matriks untuk model pembelajaran mesin *Adaptive KNN* dan *Bagging*. Model *Adaptive KNN* beradaptasi pada nilai K berdasarkan distribusi kelas dari setiap titik data, hal itu menyebabkan fluktuasi dalam performa. Dengan bantuan *SMOTE-Tomek Links*, masalah ketidakseimbangan kelas pada *dataset* yang digunakan dapat teratasi

Selain mengukur model dengan matrik evaluasi, pengukuran hasil model pembelajaran mesin dapat diamati dengan menggunakan *konfusi matriks*. Matriks konfusi digunakan ketika menyusun hasil prediksi kelas dari suatu kumpulan data yang diberikan dengan meringkas kinerja klasifikasi menggunakan tabulasi silang antara kelas aktual dan kelas prediksi. Setelah itu akan dibandingkan hasil *confusion matrix* dari pembagian data terbaik yaitu 10:90 dan 20:80.

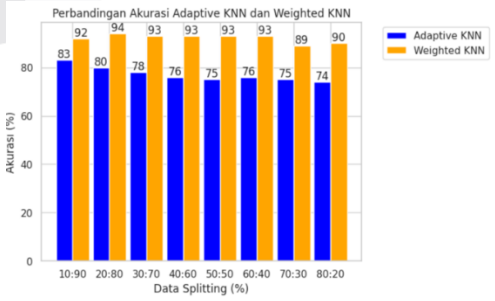


GAMBAR 5
(Confusion Matrix Adaptive KNN untuk pembagian data 10:90)

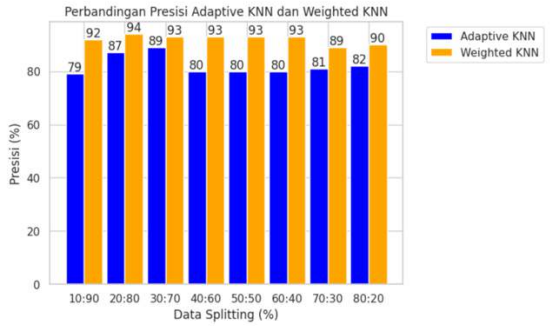


GAMBAR 6
(Confusion Matrix Adaptive KNN untuk pembagian data 20:80)

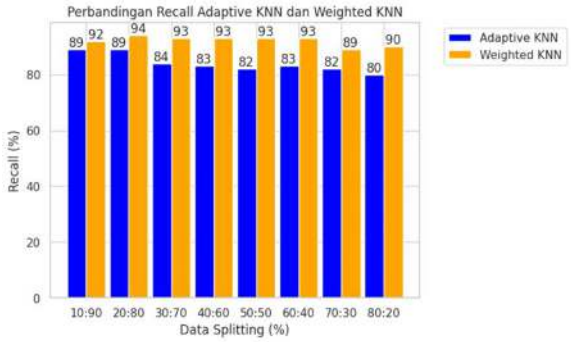
C. Perbandingan evaluasi matriks weighted knn dan adaptive knn



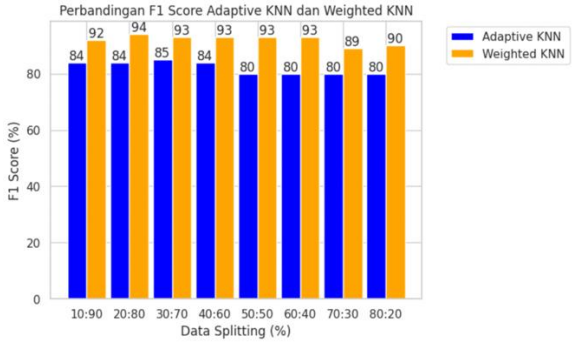
GAMBAR 7
(Perbandingan Akurasi Adaptive KNN dan Weighted KNN)



GAMBAR 8
(Perbandingan presisi Adaptive KNN dan Weighted KNN)



GAMBAR 9
(Perbandingan recall Adaptive KNN dan Weighted KNN)



GAMBAR 10
(Perbandingan F1 Score Adaptive KNN dan Weighted KNN)

V. KESIMPULAN

Klasifikasi kualitas udara menggunakan Adaptive KNN dan Weighted KNN dengan menggunakan metode SMOTE-Tomek Links dan bagging yang menggunakan penerapan teknik SMOTE-Tomek Links untuk menangani permasalahan ketidakseimbangan per kelas pada data latih dan data uji. Metode Bagging juga diterapkan untuk mengoptimalkan performa model secara keseluruhan.

Dalam penelitian ini, diperoleh hasil pada KNN Adaptif, KNN Tertimbang. KNN adaptif mencapai nilai presisi sebesar 85%, presisi sebesar 85%, recall sebesar 84%, dan f1-score sebesar 83%, sedangkan KNN tertimbang dan mengantongi mencapai presisi sebesar 95%, presisi sebesar 96%, recall 92% dan F1 skornya adalah 93% dan nilai rata-rata G adalah 0,97, stratified K-fold 0,97 dan cross-validation 0,84.

Dari hasil penelitian dapat disimpulkan bahwa kinerja KNN tertimbang dengan bagging lebih baik dalam klasifikasi kualitas udara dibandingkan model KNN adaptif. Dengan bantuan ikatan SMOTE-Tomek, ketidakseimbangan antar lapisan dapat diperbaiki.

REFERENSI

[1] Kurniawan, A. (2018). Pengukuran parameter kualitas udara (CO, NO2, SO2, O3 dan PM10) di Bukit Kototabang berbasis ISPU. *Jurnal Teknosains*, *7*(1), 1-13.

[2] Mukono, H. J. (2011). *Aspek kesehatan pencemaran udara*. Airlangga University Press.

[3] M. L. H. RI, "peraturan menteri lingkungan hidup dan kehutanan," - June 2020. [Online]. Available: https://ditppu.menlhk.go.id/portal/uploads/laporan/1601040067_P_14_2020_ISPU_menlhk.pdf.

[4] Amalia, A., Zaidiah, A., & Isnainiyah, I. N. (2022). Prediksi Kualitas Udara Menggunakan Algoritma K-Nearest Neighbor. *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, *7*(2), 496-507.

[5] Indonesia, P. R. (1999). Peraturan Pemerintah No. 41 Tahun 1999 Tentang: Pengendalian Pencemaran Udara. *Lembaran Negara RI Tahun*, *86*.

[6] J. Government, "Indeks Kualitas Udara (AQI) Jakarta dan Polusi Udara Indonesia," Government, Jakarta, - - 2021. [Online]. Available: https://www.iqair.com/id/indonesia/jakarta. [Accessed 18 December 2022].

[7] Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business horizons*, *62*(1), 15-25.

[8] Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc."

[9] Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

[10] L. A. Demidova, "Two-stage hybrid data classifiers based on svm and knn algorithms," symmetry, vol. 13, no. 4, p. 32, 2021.

[11] Mahmood, A. M. (2015). Class imbalance learning in data mining – A survey. International Journal of Communication Technology for Social Networking Services, 3(2).

[12] Townsend, J. T. (1971). Theoretical analysis of an alphabetic confusion matrix. *Perception & Psychophysics*, *9*, 40-50.

[13] Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN model-based approach in classification. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania,*

Sicily, Italy, November 3-7, 2003. Proceedings (pp. 986-996). Springer Berlin Heidelberg.

[14] Buhlmann, P., & Yu, B. (2002). Analyzing bagging. *Annals of statistics*, *30*(4), 927-961.

[15] Pereira, R. M., Costa, Y. M., & Silla Jr, C. N. (2020). MLTL: A multi-label approach for the Tomek Link undersampling algorithm. *Neurocomputing*, *383*, 95-105.

[16] Elhassan, T., & Aljurf, M. (2016). Classification of imbalance data using tomek link (t-link) combined with random under-sampling (rus) as a data reduction method. *Global J Technol Optim S*, *1*, 2016.

[17] Yigit, H. (2013, November). A weighting approach for KNN classifier. In *2013 international conference on electronics, computer and computation (ICECCO)* (pp. 228-231). IEEE.

[18] LeJeune, D., Heckel, R., & Baraniuk, R. (2019, April). Adaptive estimation for approximate $ k $-nearest-neighbor computations. In *The 22nd International Conference on Artificial Intelligence and Statistics* (pp. 3099-3107). PMLR.