E-ISSN : 2988-1986
https://ejournal.warunayama.org/kohesi

Kohesi: Jurnal Multidisiplin Saintek
Volume 4 No 10 Tahun 2024

IMPLEMENTASI PREDIKSI KUALITAS UDARA JAKARTA MENGGUNAKAN *K-NEAREST NEIGHBOR* DAN *SUPPORT VECTOR MACHINE*

Muhammad Ridho Alfarid[1], Qonita Husnia Rahmah[2]
[1]Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Islam Indonesia, D.I.Yogyakarta, email: 22611087@students.uii.ac.id
[2]Sekolah Sains Data Matematika dan Informatika Institut Pertanian Bogor, Bogor, email: qnthusnia@apps.ipb.ac.id
Corresponding Author: Muhammad Ridho Alfarid

ABSTRAK — Udara adalah unsur esensial bagi kehidupan semua makhluk hidup di bumi. Kualitas udara merupakan isu penting yang perlu mendapatkan perhatian khusus karena kondisinya yang semakin memburuk, ditandai dengan peningkatan polusi udara. Dinas Kesehatan provinsi DKI Jakarta mencatat 638.291 kasus Infeksi Saluran Pernapasan Atas (ISPA) selama periode Januari hingga Juni 2023. Angka ini menunjukkan bahwa kualitas udara menjadi suatu tantangan yang harus segera diatasi karena dapat memberikan dampak buruk bagi kesehatan pernapasan dan dapat mengakibatkan kematian. Penelitian ini menggunakan data Indeks Standar Pencemaran Udara (ISPU) Provinsi DKI Jakarta yang diperoleh dari Dinas Lingkungan Hidup Provinsi DKI Jakarta periode 1 Januari 2021 hingga 31 Juli 2024 dengan lima lokasi amatan. Penelitian ini bertujuan untuk mengklasifikasikan kualitas udara di DKI Jakarta berdasarkan konsentrasi partikulat PM10, PM2.5, SO2, CO, O3, dan NO2 menggunakan pendekatan *data mining* berbasis algoritma *K-Nearest Neighbor* (KNN) dan *Support Vector Machine* (SVM) serta penggunaan metode *Winsorization* dan *Box-Cox Transformation* dalam penanganan *outliers* dan ketidaknormalan distribusi data. Analisis dilakukan dengan membandingkan metrik *accuracy*, *precision*, *recall*, dan *F1-Score* pada beberapa model yang dibangun berdasarkan algoritma dan metode yang dipilih. Hasil model terbaik dengan nilai metrik tertinggi akan dilakukan *deployment* menggunakan *framework* Streamlit dan diaplikasikan dalam *website* prediksi yang dapat digunakan oleh masyarakat luas. *Website* ini diberi nama "AirWise" yang dirancang untuk menampilkan hasil prediksi kondisi udara saat pengguna memasukkan data konsentrasi partikulat. Hasil prediksi ini diharapkan dapat memberikan pengetahuan dan menjadi acuan berkelanjutan bagi masyarakat dalam menentukan langkah antisipasi yang tepat saat beraktivitas di luar ruangan, sehingga mereka dapat menyesuaikan diri dengan kondisi udara di kondisi tersebut.
KATA KUNCI: Kualitas Udara, Klasifikasi, *K-Nearest Neighbor*, *Support Vector Machine*, ISPU

Article History
Received: November 2024
Reviewed: November 2024
Published: November 2024

Plagirism Checker No 234
Prefix DOI : Prefix DOI : 10.8734/Kohesi.v1i2.365

I. PENDAHULUAN

Udara adalah unsur esensial bagi kehidupan semua makhluk hidup di bumi. Udara terdiri dari campuran berbagai gas alam, yang meliputi 78% nitrogen, 20% oksigen, 0.93% argon, 0.30% karbon dioksida, dan gas-gas lainnya [1]. Kualitas udara merupakan isu penting yang mendapatkan perhatian khusus karena kondisinya yang semakin memburuk, ditandai dengan peningkatan

polusi udara. Polusi udara yang terjadi ditandai oleh keberadaan senyawa berbahaya seperti materi partikulat (PM10, PM2.5), nitrogen dioksida (NO2), sulfur dioksida (SO2), karbon monoksida (CO), dan ozon (O3). Konsentrasi tinggi dari senyawa-senyawa ini dapat menimbulkan risiko serius bagi kesehatan masyarakat, termasuk penyakit pernapasan dan potensi kematian [2].

Kualitas udara yang buruk menjadi tantangan bagi Jakarta terkait fungsinya sebagai pusat perdagangan, pusat kegiatan layanan jasa dan layanan jasa keuangan, serta pusat kegiatan bisnis nasional, regional, dan global [3]. Referensi [4] mencatat bahwa pada tanggal 22 Agustus 2023, kualitas udara di Jakarta tercatat dalam kondisi tidak sehat dengan Indeks Kualitas Udara (AQI) mencapai nilai 153. Konsentrasi polutan pada hari tersebut adalah sebagai berikut: PM2.5 sebesar 59 $\mu g/m^3$, PM10 sebesar 107.7 $\mu g/m^3$, ozon (O3) sebesar 65.2 $\mu g/m^3$, nitrogen dioksida (NO2) sebesar 74.8 $\mu g/m^3$, dan sulfur dioksida (SO2) sebesar 153 $\mu g/m^3$.

Berdasarkan data Dinas Kesehatan DKI Jakarta, terdapat 638.291 kasus Infeksi Saluran Pernapasan Atas (ISPA) selama periode Januari hingga Juni 2023. Angka ini menunjukkan bahwa Jakarta memerlukan penanganan yang lebih intensif untuk mengurangi kasus ISPA di masa depan, dengan fokus pada tingkat polusi udara.

Penelitian ini bertujuan untuk mengklasifikasikan kualitas udara di DKI Jakarta berdasarkan konsentrasi partikulat PM10, PM2.5, SO2, CO, O3, dan NO2 menggunakan pendekatan *Data Mining* dengan penggunaan algoritma *K-Nearest Neighbor* dan *Support Vector Machine* yang dituangkan dalam *website* prediksi. *Website* ini dirancang untuk menampilkan hasil prediksi kondisi udara saat pengguna memasukkan data konsentrasi partikulat. Hasil prediksi ini diharapkan dapat menjadi acuan berkelanjutan bagi masyarakat dalam menentukan langkah antisipasi yang tepat saat beraktivitas di luar ruangan, sehingga mereka dapat menyesuaikan diri dengan kondisi udara di kondisi tersebut.

II. KAJIAN TERKAIT

Klasifikasi kualitas udara di Indonesia telah menarik banyak perhatian peneliti. Referensi [5] menunjukkan klasifikasi kualitas udara berdasarkan Indeks Standar Pencemaran Udara (ISPU) menggunakan metode *Fuzzy Tsukamoto* dan didapat tingkat akurasi pelatihan sebesar 0,97 atau 97%.

Peneliti [1] melakukan perbandingan algoritma KNN dan SVM yang digunakan untuk klasifikasi kualitas udara di Jakarta. Hasil penelitian menunjukkan bahwa SVM memiliki tingkat akurasi lebih tinggi yaitu sebesar 98% dibandingkan dengan tingkat akurasi KNN sebesar 96%, sehingga algoritma SVM dapat membantu dalam mengklasifikasikan indeks standar pencemaran udara secara akurat.

Referensi [6] memprediksi kualitas udara berdasarkan indeks standar pencemaran udara menggunakan XGBoost dengan *Synthetic Minority Oversampling Technique* (SMOTE) dan diperoleh nilai rata-rata akurasi sebesar 98% yang mengindikasikan model memiliki kinerja yang baik dalam hal prediksi kualitas udara.

Peneliti [7] melakukan klasifikasi tingkat kualitas udara menggunakan algoritma *Naive Bayes* dengan fokus pada Kota Tangerang Selatan dengan tingkat akurasi mencapai 79.38% dan uji skenario yang dilakukan menunjukkan kinerja yang positif.

Berdasarkan penelitian-penelitian sebelumnya yang sebagian besar berfokus pada klasifikasi kualitas udara sebagai acuan bagi kebijakan pemerintah, diperlukan pengembangan implementasi praktis yang dapat diakses oleh masyarakat, sehingga dapat diaplikasikan secara langsung untuk manfaat sosial. Hasil penelitian ini dituangkan dalam bentuk *website* prediksi yang merupakan implementasi hasil klasifikasi kualitas udara dengan menggunakan algoritma *K-Nearest Neighbor* dan *Support Vector Machine*, serta data yang digunakan adalah data Indeks Standar Pencemaran Udara (ISPU) DKI Jakarta Tahun 2021 hingga 2024. Pemilihan dan penggunaan model *K-Nearest Neighbor* (KNN) dan *Support Vector Machine* (SVM), didasari oleh kesesuaian dengan karakteristik data yang digunakan serta tingkat akurasi model yang tinggi. Oleh karena itu, penelitian ini menggabungkan teknologi web, prediksi masa depan, dan aksesibilitas yang luas, memberikan nilai tambah yang signifikan dibandingkan dengan penelitian sebelumnya.

III. SOLUSI DAN USULAN

A. DESKRIPSI SOLUSI

AirWise adalah sebuah *website* prediksi yang ditujukan untuk masyarakat umum dan dapat digunakan sebagai acuan dalam menentukan langkah antisipasi yang tepat saat beraktivitas di luar ruangan atau ruang terbuka. Prosedur kerja AirWise memerlukan data *input* dari pengguna berupa data konsentrasi materi partikulat (PM10, PM2.5), nitrogen dioksida (NO2), sulfur dioksida (SO2), karbon monoksida (CO), dan ozon (O3) pada hari tertentu. Setelah dilakukan *input* data, AirWise akan bekerja untuk menampilkan prediksi kualitas udara. Prediksi kualitas udara yang diperoleh akan menunjukkan empat kategori kualitas udara, yaitu baik, sedang, tidak sehat, dan sangat tidak sehat.

Ketika prediksi menunjukkan hasil tidak sehat atau sangat tidak sehat, masyarakat dapat mempertimbangkan untuk meminimalisir kegiatan di luar ruangan atau ruang terbuka. Jika memang harus beraktivitas di luar ruangan, masyarakat disarankan untuk menggunakan masker sebagai salah satu langkah antisipasi pencegahan penyakit saluran pernapasan yang dapat membahayakan kesehatan, bahkan menyebabkan kematian.

B. DATASET

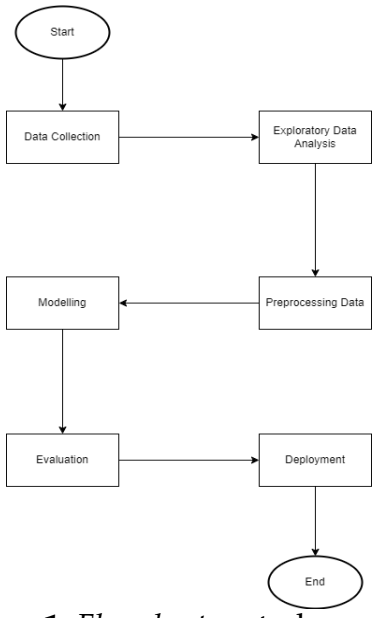
Data yang digunakan dalam penelitian ini merupakan data Indeks Standar Pencemaran Udara (ISPU) Provinsi DKI Jakarta yang diperoleh dari Dinas Lingkungan Hidup Provinsi DKI Jakarta dan dipublikasikan dalam laman Satu Data Jakarta. *Dataset* yang digunakan merupakan data ISPU pada periode 1 Januari 2021 hingga 31 Juli 2024 dengan lima lokasi amatan, yaitu Bundaran Hotel Indonesia, Kelapa Gading, Jagakarsa, Lubang Buaya, dan Kebon Jeruk. *Dataset* ini terdiri dari 5.010 data yang masih mengandung *missing value* dengan 7 atribut: PM10, PM2.5, SO2, CO, O3, NO2, dan *category* yang digunakan dalam analisis dan perancangan model.

TABEL I
SAMPLE DATA

pm10	pm2.5	so2	co	o3	no2	category
38	53	29	6	31	13	SEDANG
55	NaN	19	29	67	13	SEDANG
43	56	15	10	33	5	SEDANG
35	47	22	6	39	10	BAIK
72	108	14	43	44	20	TIDAK SEHAT

C. METODE PENELITIAN

Pada bagian ini akan ditentukan alur dalam penelitian, secara umum penelitian ini akan menggunakan algoritma *K-Nearest Neighbor* (KNN) dan *Support Vector Machine* (SVM) serta penggunaan metode *Winsorization* dan *Box-Cox Transformation* yang digunakan untuk menangani ketidaknormalan data dan penanganan *outliers*, adapun langkah-langkah rinci dari metode yang diusulkan pada Gambar 1.



Gambar 1. *Flowchart* metode penelitan

1) DATA COLLECTION

Data yang dikumpulkan merupakan data terbuka yang terjamin validitas dan reliabilitasnya dari Dinas Lingkungan Hidup Provinsi DKI Jakarta yang dipublikasikan dalam laman Satu Data Jakarta. Data Indeks Standar Pencemaran Udara (ISPU) merupakan *dataset* yang diperbarui satu bulan sekali dan dapat diunduh dalam file per tahun. Data yang digunakan dalam eksperimen merupakan data periode awal 2021 sampai dengan pertengahan 2024, sehingga terdapat empat *file* ISPU. Selanjutnya, dilakukan penggabungan empat *file* ISPU menjadi satu *dataset* dan penghapusan atribut data yang tidak digunakan. Atribut terpilih untuk digunakan pada tahap selanjutnya adalah enam atribut numerik yaitu PM10, PM2.5, SO2, CO, O3, NO2, serta satu atribut kategorik yaitu *category*.

2) EXPLORATORY DATA ANALYSIS

Exploratory data analysis (EDA) atau analisis data eksploratif berperan penting dalam membantu pengumpulan informasi dan pemahaman terkait data, meminimalkan kekeliruan, mendeteksi *outliers* atau pencilan, serta mengamati sebaran distribusi data dengan metode visualisasi [8]. *Exploratory data analysis* ini dilakukan dengan melihat dan menganalisis sebaran data menggunakan *boxplot* dan histogram, mengidentifikasi *outliers* melalui *boxplot*, serta mengamati *imbalance class* menggunakan *bar plot*.

3) PRE-PROCESSING

Fokus utama penelitian pada tahap *preprocessing* data ini adalah melakukan penanganan nilai yang hilang (*missing value*), ketidaknormalan distribusi, keberadaan *outliers* pada *dataset*, dan ketidakseimbangan kelas (*imbalance class*). Penanganan ini dilakukan untuk memaksimalkan interpretasi dan informasi dari *dataset*.

Missing value merupakan suatu kondisi di mana beberapa nilai atribut dalam kumpulan data tidak tersedia atau kosong [9]. Penanganan *missing value* dilakukan dengan metode imputasi dan penghapusan gugus data dengan mempertimbangkan kinerja model dan bias yang dihasilkan. Data yang sudah tidak terdapat *missing value* akan dilanjutkan pada tahap *encoding variable*. *Encoding variable* dilakukan pada variabel "*category*" yang semula bertipe string menjadi numerik dengan skala ordinal.

Tahap *preprocessing* dilanjutkan dengan penanganan ketidaknormalan distribusi dengan transformasi Box-Cox. Penggunaan metode transformasi Box-Cox sebagai penanganan ketidaknormalan distribusi data didasari alasan bahwa transformasi Box-Cox dapat menangani distribusi data yang cenderung miring [10]. Transformasi Box-Cox dilakukan menggunakan formula:

$$y(\lambda) = \begin{cases} \frac{y^{\lambda-1}}{\lambda}, if\ \lambda \neq 0 \\ \log y, if\ \lambda = 0 \end{cases} \qquad (1)$$

Tahapan selanjutnya yaitu penanganan keberadaan *outliers* dengan metode *Winsorization*. *Winsorization* adalah teknik transformasi statistik dengan membatasi nilai ekstrem pada data statistik dengan tujuan meminimalisir efek kemungkinan *outliers* dengan mengubah data yang dianggap *outliers* menjadi nilai pada persentil tertentu misal pada persentil ke-5 dan persentil ke-95 [11].

Imbalance class atau ketidakseimbangan kelas yang terjadi dapat diatasi menggunakan metode *resampling* dengan SMOTEENN yang diharapkan mampu menciptakan kumpulan data yang lebih seimbang dan bersih, yang pada akhirnya dapat meningkatkan generalisasi model serta memiliki performa yang baik.

4) MODELING

Pengolahan data dengan tujuan mendapatkan model terbaik dalam memprediksi kualitas udara dilakukan dengan algoritma *K-Nearest Neighbor* (KNN) dan *Support Vector Machine* (SVM).

Algoritma *K-Nearest Neighbor* (KNN) merupakan algoritma *supervised learning* yang memiliki hasil *query instance* baru yang didapat dengan klasifikasi dengan nilai mayoritas dari kategori. Algoritma KNN bertujuan untuk mengklasifikasi objek baru berdasarkan atribut data dan *training sample* [12]. Implementasi algoritma KNN dalam penelitian ini dimulai dengan melakukan *split data* pada dataset menjadi *training data* untuk pelatihan dan *testing data* untuk pengujian. Kemudian menentukan nilai k yang menunjukkan jumlah tetangga terdekat untuk digunakan dalam menentukan hasil prediksi. Dilanjutkan dengan memasukkan nilai data baru untuk pengujian (*testing*), perhitungan nilai *euclidean distance*, pengurutan (*sorting*) data dari terkecil hingga terbesar, data diambil sesuai jumlah nilai k, dan mayoritas hasil yang didapatkan diambil sebagai hasil prediksi.

Referensi [13] menyebutkan bahwa *Support Vector Machine* (SVM) merupakan teknik yang digunakan untuk memprediksi, baik dalam kasus regresi maupun klasifikasi. SVM memiliki prinsip dasar linier *classifier*, yaitu kasus klasifikasi yang secara linier dapat dipisahkan, namun SVM telah dikembangkan agar dapat bekerja pada masalah non-linier dengan memasukkan konsep kernel pada ruang kerja berdimensi tinggi. Pada ruang berdimensi tinggi, akan dicari *hyperplane* yang dapat memaksimalkan jarak antara kelas data.

5) EVALUATING

Model yang telah didapatkan akan dilakukan evaluasi dengan teknik *K-Fold Cross Validation*. *K-Fold Cross Validation* diterapkan untuk mengukur kinerja suatu algoritma dengan membagi *dataset* menjadi data pelatihan (*training*) dan data pengujian (*testing*). Setiap *fold* akan digunakan secara bergantian, yaitu satu *fold* sebagai *testing* dan *fold* lainnya digunakan sebagai *training*. Tujuan dari teknik ini adalah untuk mengidentifikasi tingkat akurasi serta menghindari masalah overfitting dan underfitting [14].

Confusion matrix atau matriks kebingungan merupakan tabel yang digunakan untuk mengevaluasi kinerja dari model klasifikasi pada *dataset* yang sebelumnya telah diketahui kelasnya. *Confusion matrix* terdiri dari 4 kombinasi hasil prediksi dengan nilai aktual yang berbeda, yaitu *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). TP adalah nilai positif yang diprediksi dengan benar, TN adalah hasil negatif yang diprediksi dengan benar, FP adalah nilai positif yang diprediksi secara salah, dan FN adalah nilai negatif yang diprediksi secara salah.

Nilai-nilai yang ada pada *confusion matrix* dapat diolah menjadi beberapa *performance metrics*, yaitu *accuracy*, *precision*, *recall*, dan *F1-Score* [15]

Accuracy diperoleh dari perhitungan dengan membagi jumlah prediksi benar (TP dan TN) dengan jumlah total data.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

Recall adalah suatu *performance metrics* yang berfungsi untuk mengukur kebaikan model dalam melihat data yang sebenarnya positif (TP) dari keseluruhan data yang sebenarnya positif (TP dan FN).

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

Precision adalah suatu *performance metrics* yang berfungsi untuk mengukur kebaikan model dalam memprediksi data positif dengan benar (TP) dari keseluruhan data yang diprediksi positif (TP dan FP).

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

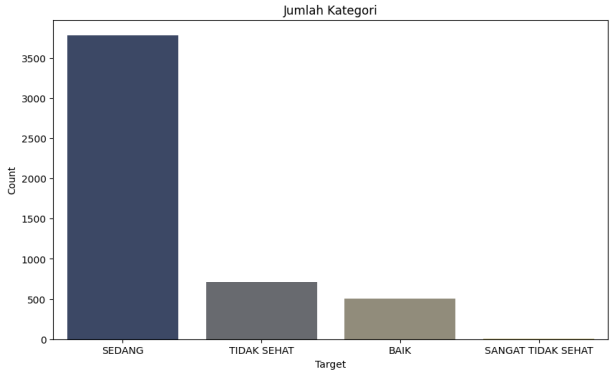
F1-Score adalah suatu *performance metrics* yang didapat dari penggabungan nilai *precision* dan nilai *recall*, yang mana digunakan untuk mengevaluasi kinerja model dengan lebih komprehensif.

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (5)$$

Deployment merupakan suatu proses penerapan atau penempatan suatu aplikasi, sistem, solusi, atau teknologi ke dalam lingkungan operasional sebenarnya setelah melalui tahap pengembangan (*development*) [16]. Pada tahap ini, hasil dari model terbaik akan diimplementasikan dalam sebuah *website* yang dapat diakses oleh masyarakat umum.

IV. HASIL EKSPERIMEN DAN PENGUJIAN

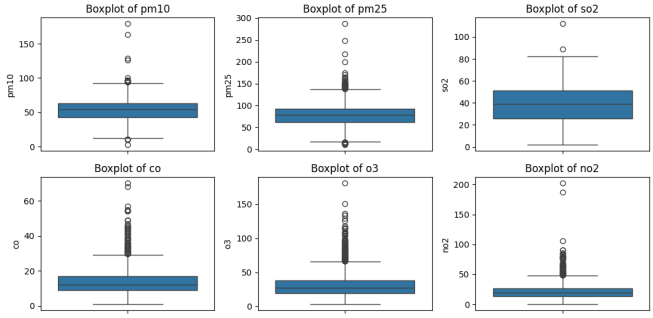
Eksperimen dan pengujian dilakukan melalui beberapa tahapan, dimulai dari *exploratory data analysis* untuk memahami karakteristik *dataset* dan meminimalisir kekeliruan. Identifikasi kelas kategori dilakukan untuk mengetahui adanya indikasi *imbalanced* pada *dataset* yang menentukan dilakukannya *resampling*.



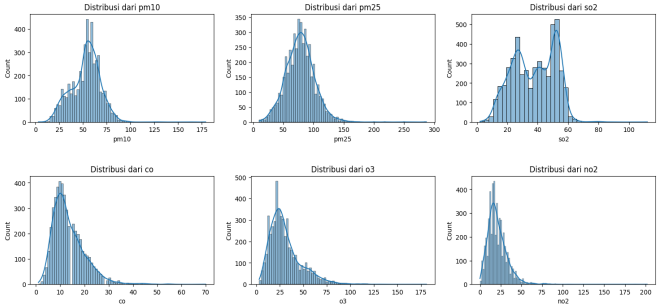
Gambar 2. *Bar plot* identifikasi kelas

Berdasarkan *bar plot* di atas, menunjukkan bahwa *dataset* memiliki *imbalanced class* atau kelas yang ketidakseimbangan sehingga perlu dilakukan *resampling* untuk menangani hal tersebut. Ini sangat berguna untuk menyeimbangkan kelas minoritas agar meningkatkan generalisasi model.

Selanjutnya, dilakukan identifikasi distribusi data dan *outliers* pada *dataset* untuk mengetahui sebaran distribusi data dan keberadaan *outliers*. Identifikasi ini digunakan untuk menentukan transformasi yang dilakukan dalam meningkatkan kinerja model saat *dataset* tidak menyebar normal dan terdapat *outliers*.



Gambar 3. *Box plot* identifikasi distribusi dan *outliers*



Gambar 4. *Histogram* identifikasi distribusi dan *outliers*

Berdasarkan Gambar 1 dan Gambar 2, diperoleh hasil bahwa hampir semua variabel memiliki *outliers* yang terletak jauh dari median data. Selain itu, distribusi juga menunjukkan hampir semua variabel memiliki distribusi tidak normal dan cenderung miring ke kanan.

Penanganan *missing value* dilakukan dengan menghapus seluruh *missing value* sehingga terjadi reduksi jumlah data yang semula 5.010 data menjadi 4.152 data. Penghapusan ini dilakukan dengan mempertimbangkan penanganan lain yaitu metode imputasi yang mana dapat memberikan bias pada *dataset* dikarenakan terlalu banyak *missing value* yang akan dilakukan imputasi. Setelah diperoleh data yang sudah tidak mengandung *missing value*, dilanjutkan dengan penanganan ketidaknormalan menggunakan transformasi Box-Cox, serta penanganan *outliers* menggunakan metode *Winsorization*.

Berdasarkan hasil *exploratory data analysis*, selanjutnya akan dilakukan klasifikasi menggunakan algoritma *K-Nearest Neighbor* (KNN) dan *Support Vector Machine* (SVM) yang mana dilakukan dua skenario eksperimen.

A. SKENARIO 1

Eksperimen yang dilakukan pada skenario 1 adalah melakukan pemodelan menggunakan algoritma KNN dan SVM dengan metode penanganan *outliers Winsorization*. Metode *Winsorization* dipilih untuk mengatasi *outliers* dengan cara mempertahankan jumlah data dan membatasi nilai ekstrem menjadi nilai persentil ke-5 dan ke-95. Penanganan *outliers* ini sangat penting untuk mengurangi dampak terhadap model yang akan dibuat.

Dalam eksperimen skenario 1 ini, algoritma KNN dikonfigurasi dengan parameter 'algorithm' yang diatur pada opsi 'auto', 'leaf size' sebesar 30, 'n_neighbors' sebanyak 5, dan 'weights' yang diatur pada opsi 'uniform'. Sementara itu, algoritma *Support Vector Machine* (SVM) menggunakan parameter 'C' yang diatur pada nilai 1.0, 'gamma' yang diatur pada opsi 'scale', dan 'kernel' yang digunakan adalah 'rbf'.

Teknik evaluasi yang digunakan untuk menilai eksperimen ini adalah *K-Fold Cross Validation* dengan jumlah *fold* sebanyak 5. Pengujian dilakukan dengan menggunakan metrik *accuracy*, *precision*, *recall*, dan F1-*Score*, yang disajikan dalam Tabel II.

TABEL II
NILAI K-FOLD CROSS VALIDATION SKENARIO 1

Model	***Accuracy***	***Precision***	***Recall***	***F1-Score***
KNN	0.960	0.959	0.960	0.959
SVM	0.972	0.971	0.972	0.970

Tabel II menyajikan nilai evaluasi *K-Fold Cross Validation*, yang menghasilkan *accuracy*, *precision*, *recall*, dan F1-*Score* untuk model *K-Nearest Neighbors* (KNN) dengan nilai berturut-turut sebesar 0.960, 0.959, 0.960, dan 0.959. Sementara itu, untuk model *Support Vector Machine* (SVM), nilai yang diperoleh secara berturut-turut adalah 0.972, 0.971, 0.972, dan 0.970.

B. SKENARIO 2

Eksperimen yang dilakukan pada skenario 2 dilakukan dengan uji coba pemodelan menggunakan algoritma KNN dan SVM dengan metode penanganan *outliers Winsorization* dan transformasi Box-Cox untuk menormalkan distribusi data. Penggunaan transformasi Box-Cox menjadi pembeda dalam skenario ini dengan skenario sebelumnya dikarenakan distribusi data

pada variabel PM10, PM2.5, SO2, CO, O3, dan NO2 yang cenderung miring ke kanan sehingga perlu dilakukan transformasi agar data menjadi lebih normal.

Setelah menangani distribusi data menggunakan transformasi Box-Cox, langkah selanjutnya adalah melakukan *handling outliers* menggunakan metode *Winsorization*. Metode *Winsorization* dilakukan dengan cara mempertahankan jumlah data dengan membatasi nilai ekstrem menjadi nilai persentil ke-5 dan ke-95.

Dalam skenario kedua, algoritma KNN dikonfigurasi dengan parameter 'algorithm' yang diatur pada opsi 'auto', 'leaf size' sebesar 30, 'n_neighbors' sebanyak 5, dan 'weights' yang diatur pada opsi 'uniform'. algoritma *Support Vector Machine* (SVM) menggunakan parameter 'C' yang diatur pada nilai 1.0, 'gamma' yang diatur pada opsi 'scale', dan 'kernel' yang digunakan adalah 'rbf'.

Metode evaluasi yang digunakan untuk menilai eksperimen ini adalah *K-Fold Cross Validation* dengan *folds* sebanyak 5. Pengujian dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, dan F1-*Score* yang disajikan dalam Tabel III.

TABEL III
NILAI K-FOLD CROSS VALIDATION SKENARIO 2

Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*
KNN	0.955	0.954	0.955	0.954
SVM	0.942	0.942	0.942	0.940

Tabel III menyajikan nilai evaluasi *K-Fold Cross Validation* pada skenario 2, dengan hasil *accuracy*, *precision*, *recall*, dan F1-*Score* untuk model *K-Nearest Neighbors* (KNN) berturut-turut sebesar 0.955, 0.954, 0.955, dan 0.954. Sementara itu, untuk model *Support Vector Machine* (SVM), nilai yang diperoleh secara berturut-turut adalah 0.942, 0.942, 0.942, dan 0.940.

Klasifikasi kualitas udara dilakukan dalam sua skenario, dengan setiap skenario menggunakan algoritma *K-Nearest Neighbor* (KNN) dan *Support Vector Machine* (SVM).

V. ANALISIS HASIL EKSPERIMEN DAN PENGUJIAN

Klasifikasi kualitas udara dilakukan dalam sua skenario, dengan setiap skenario menggunakan algoritma *K-Nearest Neighbor* (KNN) dan *Support Vector Machine* (SVM). Berdasarkan analisis skenario 1, algoritma *Support Vector Machine* (SVM) memiliki hasil yang lebih baik dibandingkan dengan algoritma *K-Nearest Neighbor* (KNN) dengan *accuracy* sebesar 0.972, *precision* sebesar 0.971, *recall* sebesar 0.972, dan F1-*Score* sebesar 0.97. Pada skema 2, algoritma *K-Nearest Neighbor* (KNN) memberikan hasil lebih baik dibandingkan dengan algoritma *Support Vector Machine* (SVM) dengan *accuracy* sebesar 0.955, *precision* sebesar 0.954, *recall* sebesar 0.955, dan F1-*Score* sebesar 0.954.

Terdapat perbedaan hasil algoritma terbaik dari kedua skenario analisis yang dilakukan, dua algoritma tersebut yaitu SVM skenario 1 dan KNN skenario 2. Dari kedua algoritma tersebut diperoleh hasil bahwa SVM skenario 1 lebih unggul dibandingkan KNN skenario 2 berdasarkan hasil *K-Fold Cross Validation*. Hasil ini berkaitan dengan karakteristik dari data dan sifat dari kedua algoritma tersebut.

Secara inheren, SVM lebih *robust* terhadap *outliers* dibandingkan dengan KNN. Pada skenario 1, metode *Winsorizing* yang membatasi nilai ekstrem di kedua ujung distribusi data memberikan keuntungan bagi SVM karena SVM terfokus pada margin terbesar antara kelas. Metode *winsorizing* juga diterapkan pada KNN, namun efeknya mungkin kurang signifikan dikarena algoritma ini sangat sensitif terhadap keberadaan dan distribusi dari tetangga terdekat.

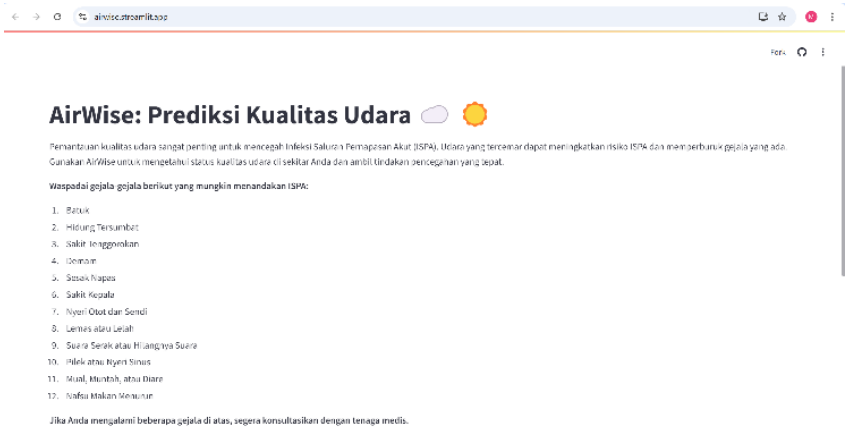
Penerapan transformasi Box-Cox model algoritma KNN dengan tujuan menormalisasi distribusi data ada kemungkinan dapat mengubah hubungan intrinsik dalam data yang lebih cocok untuk metode berbasis jarak seperti KNN. Perubahan ini mungkin tidak selalu menghasilkan peningkatan performa, terutama jika data asli sudah cukup baik untuk pemodelan dengan KNN.

SVM dirancang untuk menemukan *hyperplane* pemisah yang optimal bahkan dalam ruang berdimensi tinggi. Dalam kasus di mana transformasi Box-Cox menghasilkan perluasan dimensi atau perubahan skala yang signifikan, SVM mungkin lebih mampu mengadaptasi perubahan tersebut dibandingkan KNN yang mengandalkan jarak euclidean murni.

Konfigurasi algoritma *Support Vector Machine* (SVM) dan parameter dalam penelitian ini dilakukan berdasarkan pengalaman empiris dan eksperimen praktis. Pemilihan parameter spesifik dalam penelitian ini didasarkan pada beberapa pertimbangan. Pertama, nilai C disetel pada 1.0.

Hal ini karena nilai tersebut menyediakan keseimbangan yang rasional antara kompleksitas model dan kesalahan klasifikasi dan umum untuk digunakan. Kedua, gamma disetel pada pilihan 'scale'. Pilihan ini secara otomatis mengatur nilai gamma berdasarkan jumlah fitur (variabel) dari data, memungkinkan penyesuaian gamma yang dinamis berdasarkan data yang diberikan, sehingga menghasilkan performa yang lebih baik dibandingkan dengan nilai gamma lainnya. Terakhir, Kernel Radial Basis Function (RBF) dipilih karena fleksibilitasnya dalam menangani kasus di mana batas keputusan antar kelas tidak linier. Kernel RBF mampu memetakan sampel ke ruang dimensi yang lebih tinggi tanpa perlu menghitung dimensi yang tepat, sangat berguna dalam menemukan batas keputusan non-linear. Penyesuaian parameter ini bertujuan untuk memastikan bahwa model SVM yang dikonfigurasi dapat menghasilkan prediksi yang akurat dan efektif dalam menghadapi berbagai kondisi data.

Model terbaik yang sudah diperoleh pada eksperimen ini akan dilakukan *deployment* menggunakan *framework* streamlit sehingga menghasilkan *website* yang dapat diakses oleh masyarakat umum. *Website* ini diberi nama "AirWise" dengan tampilan seperti Gambar 5 dan Gambar 6.

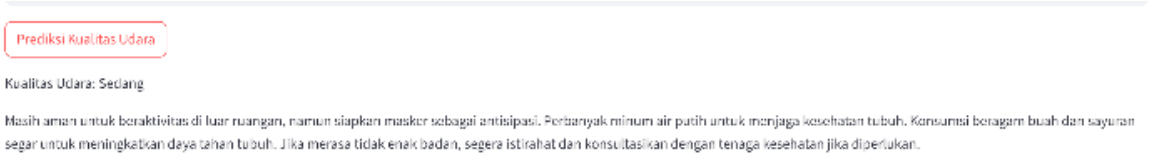


Gambar 5. Tampilan "AirWise" 1

Gambar 6. Tampilan "AirWise" 2

AirWise memiliki makna penggabungan kata "*Air*" yang artinya udara dan "*Wise*" yang artinya bijak. Penamaan tersebut menunjukkan bahwa *website* ini memberikan pengetahuan mengenai kualitas udara dan membantu masyarakat dalam pengambilan keputusan yang bijak dalam upaya pencegahan penyakit infeksi saluran pernapasan.

Penggunaan *website* ini cukup sederhana karena mempertimbangkan sasaran pengguna *website* yang tidak bergantung pada usia dan gender tertentu, yaitu dengan memberikan nilai konsentrasi materi partikulat (PM10, PM2.5), nitrogen dioksida (NO2), sulfur dioksida (SO2), karbon monoksida (CO), dan ozon (O3) pada hari tertentu. AirWise akan menampilkan prediksi kualitas udara dengan skala Baik, Sedang, Tidak Sehat, dan Sangat Tidak Sehat. AirWise juga akan menampilkan saran antisipasi yang dapat dilakukan masyarakat sesuai dengan hasil kualitas udara yang diprediksi seperti Gambar 7.



Gambar 7. Tampilan "AirWise" 3

Melihat tujuan dan aplikasi praktis *website* AirWise yang dibangun melalui penambangan data, maka penelitian ini dirasa penting untuk penggunaan jangka panjang yang sejalan dengan kondisi udara di mana kian hari kian memburuk. *Website* ini masih memerlukan penambahan fitur lain yang relevan, seperti kontak person kesehatan untuk masyarakat yang mengalami penurunan kesehatan pernapasan.

VI. KESIMPULAN

Berdasarkan hasil penambangan data yang telah dilakukan menggunakan data Indeks Standar Pencemaran Udara (ISPU) Provinsi DKI Jakarta periode 1 Januari 2021 hingga 31 Juli 2024 didapati jumlah kualitas udara terbanyak secara berturut-turut yaitu sedang, tidak sehat, baik, dan sangat tidak sehat. Hasil ini menandakan kualitas udara Provinsi DKI Jakarta yang harus segara ditangani oleh pemerintah dan masyarakat untuk menjadikan kualitas udara kembali baik.

Penggunaan model berbasis Algortima *Support Vector Machine* (SVM) dan juga metode *Winsorization* untuk menangani *outliers* merupakan model terbaik yang didapatkan selama selama proses eksperimen yang telah dilakukan. Model algoritma *Support Vector Machine* (SVM) dipadukan dengan metode *Winsorization* berhasil mencapai *accuracy* 97.2%, *precision* 97.1%, *recall* 97.2%, dan *f1-score* 97%. Hasil evaluasi ini menunjukkan efektivitas model untuk menghasilkan prediksi yang akurat dan konsisten sehingga dapat digunakan masyarakat umum.

Penggunaan *framework* Streamlit untuk menghasilkan *website* berbasis SVM dipadukan metode *Winsorization* memberikan kemudahan masyarakat untuk mengetahui kualitas udara di lingkungan sekitar. Kehadiran *website* AirWise

diharapkan membantu masyarakat untuk membantu masyarakat dalam

upaya meminimalisir penyakit saluran pernapasan. AirWise memberikan informasi terkait tindakan yang harus dilakukan berdasarkan kualitas udara yang terjadi saat pengguna memberikan nilai konsentrasi polutan ke dalam *website* AirWise.

REFERENSI

[1] B.V. Jayadi, T. Handhayani, dan M.D. Lauro, "Perbandingan KNN dan SVM untuk Klasifikasi Kualitas Udara di Jakarta," *Jurnal Ilmu Komputer dan Sistem Informasi*, Vol. 11, No. 02, 2023, doi: 10.24912/jiksi.v11i2.26006

[2] K.A. Arsyad dan Y. Priyana, "Studi Kausalitas antara Polusi Udara dan Kejadian Penyakit Saluran Pernapasan pada Penduduk Kota Bogor, Jawa Barat, Indonesia," *Jurnal Multidisiplin West Science*, vol. 2, no. 06, hal 462–472, Jun 2023, doi: 10.58812/jmws.v2i6.434.

[3] Pemerintah Indonesia. 2024. *Undang-Undang Republik Indonesia Nomor 2 Tahun 2024 tentang Provinsi Daerah Khusus Jakarta*. Lembaran RI Tahun 2024. Jakarta.

[4] IQAir, "Kualitas udara di Jakarta 22 Agustus 2023," 2023. Tanggal akses: 22-Sep-2024. [Online]. Tersedia: Indeks Kualitas Udara (AQI) Jakarta dan Polusi Udara di Indonesia | IQAir

[5] A.E. Putra dan T. Rismawan, "Klasifikasi Kualitas Udara Berdasarkan Indeks Standar Pencemaran Udara (ISPU) Menggunakan Metode *Fuzzy Tsukamoto*," *Jurnal Komputer dan Aplikasi*, vol. 11, no. 02, hal 190–196, 2023, doi: 10.26418/coding.v11i2.58704.

[6] A.A. Nababan, M. Jannah, M. Aulina, dan D. Andrian, "Prediksi Kualitas Udara Menggunakan XGBoost dengan *Synthetic Minority Oversampling Technique* (SMOTE) Berdasarkan Indeks Standar Pencemaran Udara (ISPU)," *Jurnal metode Informatika Kaputama (JTIK)*, Vol. 7, No. 1, Jan. 2023, doi: 10.59697/jtik.v7i1.66.

[7] F. Widiawati, R. Kurniawan, dan T. Suprapri, "Klasifikasi Data Tingkat Kualitas Udara di Tangerang Selatan Menggunakan Algoritma Naive Bayes," *JATI (Jurnal Mahasiswa metode Informatika)*, Vol. 7, No. 6, hal 3739–3745, Des. 2023, doi: 10.36040/jati.v7i6.8261.

[8] Ramadhani, Ramadhanu, T. Hidayat, "Exploratory Data Analysis (EDA) untuk Mengetahui Distribusi Data Kualitas Susu Sapi," *Jurnal SAINTIKOM (Jurnal Sains Manajemen Informatika dan Komputer)*, Vol. 23, No. 1, hal 68-76, Feb. 2024, doi: 10.53513/jis.v23i1.9500
[9] Farhan, "Mengatasi Keberadaan Missing Value." 2024. Tanggal akses: 03-Okt-2024. [Online]. Tersedia: Mengatasi Keberadaan Missing Value | by Farhan | Medium
[10] S. Marimuthu, T. Mani, T. D. Sudarsanam , S. George, L. Jeyaseelan, "Preferring Box-Cox transformation, instead of log transformation to convert skewed distribution of outcomes to normal in medical research," *Clinical Epidemiology and Global Health*, Vol. 15, hal 1–5, April 2022, doi: 10.1016/j.cegh.2022.10104.
[11] P. R. Sihombing, Suryadiningrat, D. A. Sunarjo, Y. P. C. Yuda, "Identifikasi Data Outlier (Pencilan) dan Kenormalan Data Pada Data Univariat serta Alternatif Penyelesaiannya," *Jurnal Ekonomi dan Statistik Indonesia*, Vol. 2, No. 3, hal 307 – 316, Des. 2022, doi: 10.11594/jesi.02.03.07
[12] A. Yudhana, Sunardi1, dan A.J.S. Hartanta, "Algoritma K-NN dengan Euclidean Distance untuk Prediksi Hasil Penggergajian Kayu Sengon," *Transmisi: Jurnal Ilmiah metode Elektro*, Vol. 22, No. 4, hal 123–129, Okt. 2020, doi: 10.14710/transmisi.22.4.107-141.
[13] R. Damasela, B. P. Tomasouw, Z. A. Leleury, "Penerapan Metode Support Vector Machine (SVM) untuk Mendeteksi Penyalahgunaan Narkoba," *PARAMETER: Jurnal Matematika, Statistika, dan Terapannya*, Vol. 1, No. 2, hal 111–122, Okt. 2022, doi: 10.30598/parameterv1i2pp111-122
[14] H. Hardianti, "Penerapan K-Fold Cross Validation untuk Menganalisis Kinerja Algoritma K-Nearest Neighbor pada Data Kasus Covid-19 di Indonesia" *Journal of Mathematics, Computations, and Statistics*, Vol. 6,
[15] No. 2, hal 161–168, Okt. 2023, doi: 10.35580/jmathcos.v6i2.53043.
[16] H. Ardikatama, "Jenis-Jenis Performance Metrics yang Wajib Kamu Ketahui!," 2021. Tanggal akses: 28-Sep-2024. [Online]. Tersedia: Jenis-Jenis Performance Metrics yang Wajib Kamu Ketahui! | by Haikal Ardikatama | Medium
[17] H. I. Abdulmalik, "Deployment: Pengertian, Tujuan, dan Jenis-jenisnya," 2024. Tanggal akses: 28-Sep-2024. [Online]. Tersedia: Deployment: Pengertian, Tujuan, dan Jenis-jenisnya (dicoding.com)