GREBUCI
Global Research on Economy, Business, Communication and Information

Analisis Komparasi Algoritma *Support Vector Machine* dan *K-Nearest Neighbor* pada Klasifikasi Kualitas Udara Kota Jakarta

Yudha Aryadi Sani[1*], Budi Wasito[2]

[1,2]Departemen Informatika, Institut Bisnis dan Informatika Kwik Kian Gie
Jalan Yos Sudarso Kav 87, Sunter, Jakarta, 14350, Indonesia.
[1]Alamat email: yudhasani02@gmail.com
[2]Alamat email: budiwasito@kwikkiangie.ac.id
*Penulis korespondensi

Abstract **:** *Air pollution has become one of the biggest environmental challenges worldwide, and Jakarta, the capital city of Indonesia, is no exception. Jakarta, as one of the most populous cities in the world, faces serious problems in ensuring healthy air quality for its residents. The impact of this air pollution is not only limited to human health, but also damages the environment as a whole, including plants and aquatic ecosystems. In an effort to address this issue, many studies have been conducted to develop prediction models that can provide estimates of air quality levels in Jakarta City. Two modeling models that can be used in this context are Support Vector Machine (SVM) and K-Nearest Neighbors (KNN). This study aims to compare the performance of the SVM algorithm model with the KNN algorithm in the classification of air quality in Jakarta City. Data mining is the art and science of discovering knowledge, insights, and patterns in data based on the CRISP -DM (Cross Industry Standard Process For Data Mining) methodology. The data source in this study is the Jakarta air pollution standard index (ISPU) data obtained on the satudata.jakarta.go.id website. This study compares the accuracy of the Jakarta City air quality classification on the Support vector Machine (SVM) and K-Nearest Neighbor (KNN) algorithms using python. Comparison results are determined by the level of accuracy and score of the confusion matrix. Classification results are presented in the form of a Graphic User Interface with media interface. The comparison of classification algorithms on two models in data mining shows that the Support Vector Machine (SVM) algorithm is superior to K-Nearest Neighbor (KNN). This is evident from the higher accuracy rate in SVM, especially with the use of the Rbf kernel which reaches 97.05%, compared to KNN which has an accuracy of 94.74% with parameters p = 1 and k = 5. In addition, SVM also shows a higher correct prediction value compared to KNN. In a 1-year time span the overall quality may be quite good, with moderate quality at more than equal to 50 to 100. However, control is still needed for the unhealthy category areas.*

Keywords **:** *classification, data mining, support vector machine, k-nearest neighbor, air pollution*

Cite **:** Sani, Y. A., & Wasito, B. (2024). Analisis Komparasi Algoritma Support Vector Machine dan K-Nearest Neighbor pada Klasifikasi Kualitas Udara Kota Jakarta. *Global Research on Economy, Business, Communication, and Information*, *2*(1), 55-72. https://doi.org/10.46806/grebuci.v2i1.1757

1. PENDAHULUAN

Polusi udara telah menjadi salah satu tantangan lingkungan terbesar di seluruh dunia, dan tidak terkecuali bagi Kota Jakarta sebagai ibu kota Indonesia. Jakarta, sebagai salah satu kota terpadat di dunia, menghadapi masalah serius dalam memastikan kualitas udara yang sehat bagi penduduknya. Polusi udara dapat berasal dari berbagai sumber, termasuk kegiatan industri, transportasi bermotor, dan aktivitas manusia lainnya. Dampak dari polusi udara ini tidak hanya terbatas pada kesehatan manusia, tetapi juga merusak lingkungan secara keseluruhan, termasuk tanaman dan ekosistem air.

Pentingnya menjaga kualitas udara di Kota Jakarta menjadi semakin mendesak, mengingat dampak kesehatan yang serius yang dapat ditimbulkannya. Pencemaran udara telah terbukti dapat menyebabkan berbagai penyakit pernapasan, penyakit jantung, dan bahkan dapat berkontribusi pada tingkat kematian yang lebih tinggi. Oleh karena itu, perlu dilakukan upaya maksimal untuk memantau dan mengatasi masalah polusi udara ini.

Dalam konteks ini, banyak penelitian telah dilakukan untuk mengembangkan model klasifikasi yang dapat memberikan perkiraan tingkat kualitas udara di Kota Jakarta. Namun, Penting untuk tidak hanya fokus pada penggunaan model, tetapi juga menentukan perbedaan kualitas dari setiap kategori pada kedua algoritma. Kategori ini dapat mencakup tingkat pencemaran sehat, sedang, tidak sehat, sehingga memberikan gambaran yang lebih jelas tentang tingkat bahaya bagi kesehatan masyarakat.

Tujuan mengkomparasi antara SVM dan KNN dalam konteks klasifikasi kualitas udara di Jakarta bukan hanya sebatas untuk membandingkan performa keduanya, tetapi juga untuk mengevaluasi perbedaan kualitas yang dihasilkan. Dengan demikian, analisis tidak hanya difokuskan pada akurasi dan konfusi matriks, tetapi juga pada kemampuan model dalam mengklasifikasikan tingkat pencemaran secara tepat dan relevan.

Tingkat pencemaran udara di Jakarta sangat bervariasi dan dinamis, tergantung pada faktor- faktor seperti aktivitas industri, lalu lintas, cuaca, dan pola pemukiman. Oleh karena itu, dibutuhkan pendekatan yang sensitif dan responsif untuk memantau dan mengklasifikasikan kualitas udara secara akurat. Dengan menggunakan algoritma seperti SVM dan KNN, diharapkan dapat memberikan wawasan yang lebih dalam tentang karakteristik pola pencemaran udara di berbagai wilayah Jakarta.

Masyarakat perlu mengetahui hasil komparasi antara SVM dan KNN dalam klasifikasi kualitas udara di Jakarta agar dapat lebih memahami tingkat pencemaran udara di sekitar mereka. Informasi ini dapat membantu mereka untuk mengambil langkah-langkah perlindungan yang sesuai, seperti menggunakan masker anti polusi atau menghindari aktivitas luar ruangan saat tingkat pencemaran tinggi. Dengan demikian, hasil penelitian ini memiliki dampak yang signifikan bagi kesehatan dan kesejahteraan masyarakat Jakarta secara keseluruhan.

Penelitian ini bertujuan untuk mengkaji hasil komparasi model antara SVM dan KNN dalam mengklasifikasikan kualitas udara di Kota Jakarta, dengan memprioritaskan akurasi dan confusion matrix menggunakan data indeks pencemaran udara kota jakarta. Hasil komparasi tersebut diharapkan dapat mendukung pengambilan keputusan dan implementasi langkah-langkah pencegahan yang efektif.

2. TINJAUAN PUSTAKA

2.1. Data

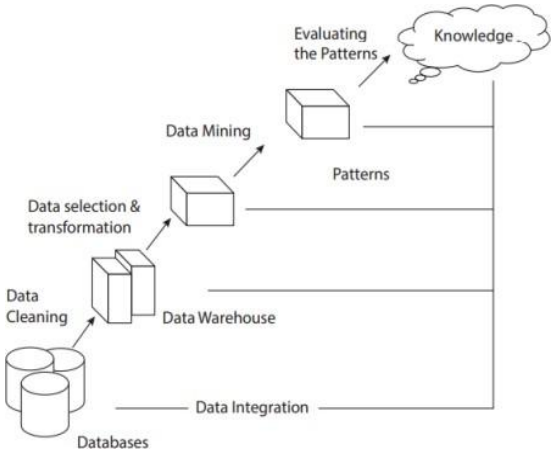
Menurut Kenneth C. Laudon dan Jane P. Laudon (2017:44), data didefinisikan sebagai "aliran fakta mentah yang mewakili peristiwa yang terjadi di dalam organisasi atau lingkungan fisik sebelum mereka memiliki makna dan berguna bagi manusia".

1. Menurut Syafrizal Helmi Situmorang (2014:3) cara memperoleh data dapat dibagi menjadi dua yaitu: 1. Data Primer (primary data) yaitu data yang dikumpulkan sendiri oleh perorangan/ suatu organisasi secara langsung dari objek yang diteliti dan untuk kepentingan studi yang bersangkutan yang dapat berupa interview, observasi.
2. Data Sekunder (secondary data) yaitu data yang diperoleh/ dikumpulkan dan disatukan oleh studi – studi sebelumnya atau yang diterbitkan oleh berbagai instansi lain. Biasanya sumber tidak langsung berupa data dokumentasi dan arsip – arsip resmi.

2.2. Data Mining

Menurut Pan Ning-Tan (2019:24), “Data Mining adalah proses menemukan informasi yang berguna secara otomatis dalam repositori data besar. Teknik data mining digunakan untuk menyelidiki set data besar dengan tujuan menemukan pola yang baru dan bermanfaat yang mungkin tidak akan diketahui sebaliknya. Mereka juga memberikan kemampuan untuk memprediksi hasil dari observasi di masa depan, seperti jumlah yang akan dihabiskan oleh pelanggan di toko online atau toko fisik.”

Data mining adalah bagian integral dari penemuan pengetahuan dalam basis data (KDD), yang merupakan proses keseluruhan untuk mengubah data mentah menjadi informasi yang berguna, seperti yang ditunjukkan dalam Gambar 1. Proses ini terdiri dari serangkaian langkah, mulai dari pra-pemrosesan data hingga pemrosesan hasil data mining.



Gambar 1 Proses data mining

Proses Data Mining:

1. *Pembersihan data.* Langkah ini dapat didefinisikan sebagai menghapus data yang tidak relevan. Menghapus data yang tidak relevan tidak lain adalah data yang tidak diinginkan dan dapat dihapus.

2. *Integrasi data.* Data dikumpulkan dari sumber yang heterogen dan diintegrasikan ke dalam sumber yang sama seperti gudang data. 9 c. Pemilihan & transformasi data Setelah data dipilih, tugas selanjutnya adalah transformasi data. d. Evaluasi pola Evaluasi didasarkan pada beberapa ukuran; setelah langkah-langkah ini diterapkan, hasil yang diambil dibandingkan / dievaluasi secara ketat berdasarkan pola yang disimpan. e. Representasi pengetahuan Merepresentasikan data yang telah diproses ke dalam format yang dibutuhkan seperti tabel dan laporan.

2.3. *Machine Learning*

Menurut Ian Goodfellow (2017 : 98), "Machine learning adalah suatu bentuk statistik terapan terapan dengan penekanan yang lebih besar pada penggunaan komputer untuk memperkirakan fungsi-fungsi yang rumit secara statistik" Tujuan machine learning biasanya dijelaskan dalam bentuk bagaimana sistem machine learning harus memproses sebuah contoh. Contohnya adalah kumpulan fitur yang telah diukur secara kuantitatif dari beberapa objek atau peristiwa yang kita ingin proses oleh sistem pembelajaran mesin.

2.4. Pencemaran Udara

Menurut Abhishek Tiwary dan Ian Williams (2019:1) Polusi udara didefinisikan sebagai "Keberadaan zat di atmosfer yang dapat menyebabkan efek buruk bagi manusia dan lingkungan". Polusi (dalam pengertian umum) didefinisikan sebagai masuknya zat atau energi ke dalam lingkungan oleh manusia yang dapat menyebabkan bahaya bagi kesehatan manusia, kerusakan sumber daya hidup dan sistem ekologi, kerusakan struktur atau fasilitas, atau gangguan terhadap penggunaan lingkungan yang sah.

Sumber Pencemaran Udara: (1) Pencemar primer adalah polutan yang dipancarkan langsung ke atmosfer - misalnya, CO berasal langsung dari pembakaran bahan bakar fosil yang tidak sempurna pada kendaraan bermotor, dan SO2 dipancarkan dari pembangkit listrik dan pabrik-pabrik industri. (2) Pencemar sekunder terbentuk di udara sebagai hasil 16 dari reaksi kimia dengan polutan lain dan gas-gas di atmosfer - misalnya, ozon dihasilkan oleh reaksi fotokimia di atmosfer (Abhishek Tiwary dan Ian Williams, 2019).

Sumber bahan pencemar udara primer dapat dibagi lagi menjadi dua golongan besar, yaitu (Abhishek Tiwary dan Ian Williams, 2019): (1) Sumber Alamiah (Natural Sources) Beberapa kegiatan alam yang bisa menyebabkan pencemaran udara adalah aktivitas gunung berapi, kebakaran hutan, badai pasir, petir dan lain-lain. (2) Sumber Buatan Manusia (Anthropogenic Sources) Sumber utamanya adalah pembakaran bahan bakar fosil untuk energi, terutama di pembangkit listrik dan kendaraan bermotor. Namun, ada banyak sumber yang tidak terkait dengan pembakaran, termasuk proses industri, pertambangan batubara, penggunaan pelarut rumah tangga dan industri, kebocoran gas alam di jaringan distribusi nasional, dan tempat pembuangan sampah. Sumber non-pembakaran sangat penting untuk VOC dan metana.

2.5. Support Vector Machine

Menurut Pang-Ning Tan, et al. (2019:478), "Support Vector Machine (SVM) adalah model klasifikasi diskriminatif yang mempelajari batas keputusan linear atau nonlinear pada ruang atribut untuk memisahkan kelas-kelas. Selain memaksimalkan

pemisahan dua kelas, SVM menawarkan kemampuan regularisasi yang kuat, yaitu mampu mengontrol kompleksitas model untuk memastikan kinerja generalisasi yang baik".

Karena kemampuan uniknya untuk secara naluriah meregulasi pembelajarannya, SVM mampu mempelajari model yang sangat ekspresif tanpa menderita dari overfitting. Oleh karena itu, SVM telah menerima perhatian yang besar di komunitas pembelajaran mesin dan umumnya digunakan dalam beberapa aplikasi praktis, mulai dari pengenalan digit tulisan tangan hingga kategorisasi teks.

2.6. K-Nearest Neighbor

Menurut Muhammad Arhami dan Muhammad Nasir (2020:96-98) "KNN merupakan salah satu algoritma untuk klasifikasi yang biasa digunakan dalam data mining. KNN juga masuk dalam kategori regresi yang juga dapat digunakan untuk memprediksi seperti halnya regresi".

Ide dari metode k-nearest-neighbors adalah untuk mengidentifikasi k records dalam dataset pelatihan yang mirip dengan rekaman baru yang ingin kita klasifikasikan. Kami kemudian menggunakan catatan-catatan yang mirip (bertetangga) ini untuk mengklasifikasikan catatan baru ke dalam sebuah kelas, menetapkan catatan baru ke kelas yang dominan di antara tetangga-tetangga ini (Galit Shmueli 2018).

Nilai K merupakan suatu parameter yang merujuk kepada jumlah tetangga yang paling dekat dengan objek yang diprediksi kelasnya sehingga dapat ditentukan tetangga yang 19 mayoritas bagi suatu objek K memainkan peranan penting dalam menentukan keberhasilan model dan akurasi yang lebih baik. K juga merupakan batasan dalam setiap kelas.

Berikut adalah langkah-langkah dari algoritma K-Nearest Neighbors (KNN): Tentukan nilai k: Langkah pertama dalam algoritma KNN adalah menentukan nilai k, yaitu jumlah tetangga terdekat yang akan dipertimbangkan saat mengklasifikasikan suatu data baru.

1. Nilai k disesuaikan juga dengan jumlah data training yang ada dan sebaiknya nilai k diambil dalam jumlah ganjil seperti 1,3,5,7, … dan sebagainya, karena jika mengambil ganjil akan mudah dalam menentukan mayoritas dan minoritas kedekatan jarak antara record data uji yang diprediksi dengan record data latih.
2. Hitung jarak antara data baru dengan semua data pada dataset. Jarak yang paling umum digunakan adalah jarak Euclidean.
3. Urutkan jarak yang telah dihitung dan tentukan tetangga terdekatnya sesuai dengan k yang dipilih dan jarak minimumnya (dari yang terkecil ke terbesar) sehingga berdasarkan urutan tersebut akan didapatkan kategori dari data yang diprediksi.
4. Kategori data baru yang diprediksi akan masuk dalam kelas mayoritas sesuai dengan nilai k yang ditentukan sebelumnya.
5. Evaluasi model: Evaluasi performa model KNN menggunakan metrik seperti akurasi, presisi, recall, dan F1 score.

Algoritma KNN dapat digunakan untuk berbagai tugas seperti klasifikasi dan regresi. Namun, algoritma ini memiliki beberapa kelemahan seperti sensitivitas terhadap nilai k dan jarak yang digunakan, serta membutuhkan memori yang besar untuk menyimpan data latih.

3. METODE PENELITIAN

3.1. Teknik Pengumpulan Data

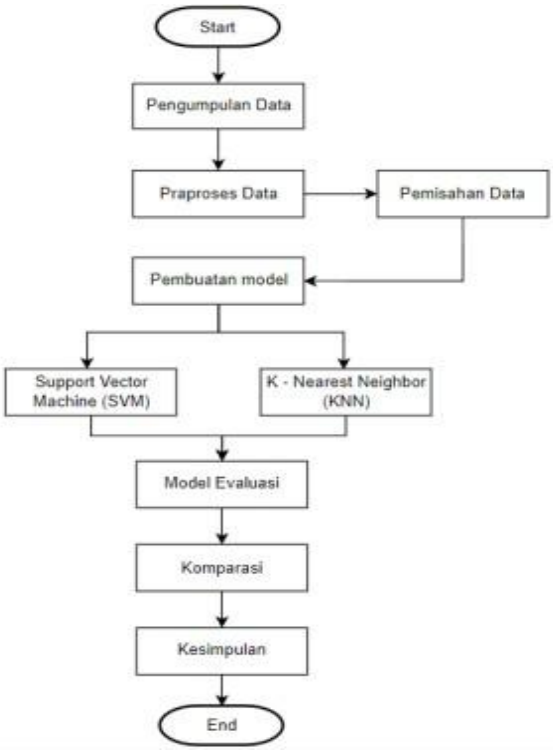
Pada penelitian ini yang digunakan adalah data sekunder yang diambil dari satudata.jakarta.go.id. Teknik pengumpulan data memiliki arti cara umum dalam mengumpulkan data, sedangkan instrumen pengumpul data merupakan alat untuk mengumpulkan data berdasarkan teknik yang digunakan. Alasan peneliti menggunakan data sekunder adalah untuk menghemat waktu dan sumber daya yang diperlukan untuk mengumpulkan data baru. Dalam mengumpulkan data, penulis menggunakan metode kuantitatif karena penelitian ini menggunakan data numerik dan 25 statistik untuk membandingkan performa dua algoritma dalam klasifikasi kualitas udara di Jakarta.

3.2. Teknik Analisis Data

Dalam penelitian ini, peneliti menggunakan teknik analisis yang dilakukan berdasarkan metode CRISP-DM (Cross-Industry Standard Process for Data Mining). CRISP-DM menyediakan proses standar yang tidak berpemilik dan tersedia secara bebas agar dapat menyesuaikan hasil data mining ke dalam strategi pemecahan masalah umum bisnis atau unit penelitian. CRISP- DM memiliki 6 siklus yang fasenya, antara lain: *business understanding phase, data understanding phase, data preparation phase, modeling phase, evaluation phase,* dan *development phase.*

3.3. Penerapan Algoritma

Pada penelitian ini akan dibahas bagaimana penggunaan komparasi algoritma Support Vector Machine (SVM) dan K-Nearest Neighbor (KNN) untuk klasifikasi kualitas udara Kota Jakarta dapat digambarkan pada diagram sebagai berikut:



Gambar 2 Workflow Penerapan Algoritma

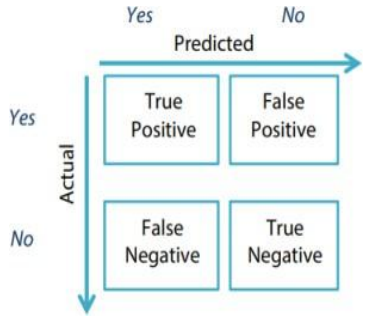
Gambar 2 merupakan workflow yang digunakan oleh penulis untuk melakukan penelitian ini. Berikut adalah penjelasannya:

1. Pengumpulan data diambil melalui website satudata.jakarta.go.id.
2. Praproses Data meliputi pembersihan data dan transformasi data menjadi data yang berkualitas untuk dilakukan tahapan data mining.
3. Pemisahan Data yaitu membagi data menjadi dua set, yaitu set pelatihan dan set pengujian. Set pelatihan digunakan untuk melatih model algoritma. Set pengujian digunakan untuk mengevaluasi kinerja model algoritma.
4. Membentuk model algoritma data mining menggunakan set pelatihan.
5. Data diolah menggunakan model Support Vector Machine dengan bahasa pemrograman Python.
6. Data diolah menggunakan model K- Nearest Neighbor dengan bahasa pemrograman Python.
7. Melakukan Evaluasi dan mengkomparasi skor hasil pengolahan.
8. Melakukan komparasi keakuratan algoritma SVM dan KNN menggunakan Python pada klasifikasi kualitas udara Kota Jakarta.
9. Kesimpulan dari hasil komparasi pada masing-masing algoritma.

3.4. Teknik Pengukuran Data

3.4.1. Menghitung Nilai Accuracy

Secara umum Accuracy dapat dirumuskan sebagai berikut:



Gambar 3 Confusion Matrix

Cara menghitung nilai Accuracy, Accuracy Akurasi didefinisikan sebagai tingkat kedekatan antara nilai dan prediksi dengan nilai aktual.

$$\text{Accuracy} : \frac{TP}{TP + TN + FN + FP}$$

3.4.2. Algoritma Support Vector Machine

Pada algoritma SVM terdapat trik kernel yaitu metode untuk menghitung kesamaan ini sebagai fungsi dari atribut aslinya. Dalam penelitian ini, fungsi kernel digunakan untuk perhitungan. Fungsi kernel yang umum digunakan adalah sebagai berikut:

Linear Kernel

$$K(x, y) = x\ y$$

Keterangan:
K = fungsi kernel x, y = Vektor

Polynomial Kernel

$$K(x1, y1) = (x1\ y1 + 1)d$$

Keterangan:
K(x,y) = hasil kernel antara dua vektor input x dan y.
d = derajat polinomial

Radial Basis Function Kernel.

$$K(x, y) = exp(-\frac{||x - y||^2}{2\sigma^2})$$

K(x,y) adalah hasil kernel antara dua vektor input x dan y, exp adalah fungsi eksponensial, yang merupakan fungsi dasar dari operasi pangkat e (bilangan Euler), $\|x - y\|$ adalah norma Euclidean (jarak Euclidean) antara vektor x dan y yang dihitung sebagai akar kuadrat dari jumlah kuadrat perbedaan elemen-elemen vektor, dan σ adalah parameter bandwidth yang dapat diatur untuk mengontrol seberapa luas distribusi kernel.

Sigmoid Kernel

$$K(x, y) = tanh(ax^T y + c)$$

K(x,y) adalah hasil kernel antara vektor input x dan y, tanh adalah fungsi 31 tangen hiperbolik, yang menghasilkan nilai antara -1 dan 1, a dan c adalah parameter yang dapat diatur untuk mengontrol bentuk dan sifat dari kernel.

3.4.3. Algoritma K-Nearest Neighbor

Pada langkah dari algoritma K-Nearest Neighbors (KNN) yaitu hitung jarak antara data baru dengan semua data pada dataset. Jarak yang digunakan dalam penelitian ini adalah jarak Euclidean dan jarak Manhattan. Berikut adalah formula dari setiap jarak:

Euclidean distance

Menurut Parteek Bhatia (2019:157) Jarak Euclidean paling umum digunakan untuk menghitung jarak. Berikut adalah formula untuk euclidean distance:

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

istance(x, y) = jarak antara titik pada data training x dan titik data testing y yang akan diklasifikasi. $(xi - yi)$ = Merupakan kuadrat dari selisih antara komponen ke- i dari 2 vektor x dan y. Ini mengukur perbedaan pada setiap dimensi.

Manhattan distance

Menurut Parteek Bhatia (2019:193) Jarak Manhattan didefinisikan sebagai jumlah panjang proyeksi ruas garis antara dua titik pada sumbu koordinat. Berikut adalah formula untuk manhattan distance:

$$\text{distance}(x, y) = \sum_{i=1}^{n} |xi - yi|$$

distance(x, y) = jarak antara titik pada data training x dan titik data testing y yang akan diklasifikasi. $|xi - yi|$: Menunjukkan nilai absolut dari selisih antara komponen ke-i dari vektor x dan y. Ini mencerminkan panjang proyeksi ruas garis (jarak) antara titik-titik pada sumbu koordinat pada dimensi tertentu.

3.4.4. Indeks Standar Pencemaran Udara

Menurut Peraturan Menteri Lingkungan Hidup Dan Kehutanan Republik Indonesia Nomor P.14/Menlhk/Setjen/Kum.1/7/2020 Tentang Indeks Standar Pencemar Udara Pasal 1 menyatakan bahwa Indeks Standar Pencemar Udara yang selanjutnya disingkat ISPU adalah angka yang tidak mempunyai satuan yang menggambarkan kondisi mutu udara ambien di lokasi tertentu, yang didasarkan kepada dampak terhadap kesehatan manusia, nilai estetika dan makhluk hidup lainnya. Terdapat penentuan kategori dalam ISPU antara lain:

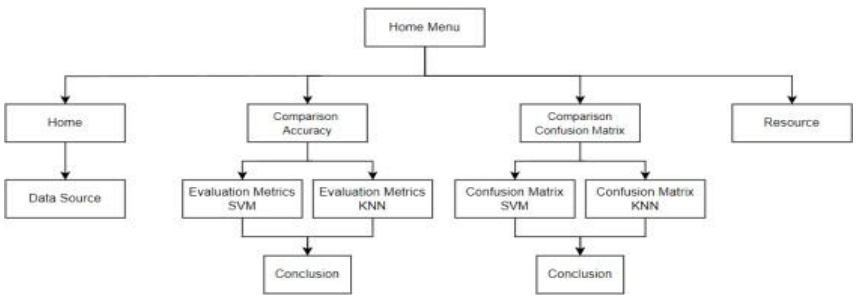
Tabel 1

Kategori Angka Rentang ISPU

Kategori	Angka Rentang	Keterangan
Baik	1 - 50	Tingkat kualitas udara yang sangat baik, tidak memberikan efek negatif terhadap manusia, hewan, tumbuhan.
Sedang	51- 100	Tingkat kualitas udara masih dapat diterima pada kesehatan manusia, hewan dan tumbuhan.
Tidak Sehat	101 - 200	Tingkat kualitas udarayang bersifat merugikan pada manusia, hewan dan tumbuhan.

3.5. Teknik Perancangan Graphic User Interface (GUI)

Pada perancangan Graphic User Interface (GUI), peneliti menggunakan Google Sites untuk menyajikan hasil proses data mining dalam bentuk dashboard. Pada halam website terdapat page home, Page Prediction, Page Comparison, dan Page Resource. Detail Rancangan GUI yang ada meliputi Home, Prediction, Comparison, Resource. Berikut gambar hirarki dari rancangan GUI tersebut:



Gambar 4 Hierarki Rancangan GUI

4. HASIL DAN PEMBAHASAN

4.1. Architectural algorithm

Peneliti menggunakan arsitektur dari setiap algoritma yang ingin digunakan sebagai mekanisme dasar yang mendasari suatu algoritma termasuk langkah-langkah yang diambil untuk memecahkan masalah, dan detail implementasi, serta alur kerja algoritma.

4.1.1. Support Vector Machine Architecture

Dalam klasifikasi kualitas udara di Kota Jakarta, arsitektur algoritma SVM melibatkan beberapa langkah:

a. Konversi Data ke Ruang Fitur Tinggi Data masukan, seperti tingkat polutan udara dan faktor- faktor lingkungan lainnya, dikonversi ke dalam ruang fitur yang lebih tinggi menggunakan fungsi kernel. Proses ini memungkinkan SVM untuk menangani masalah klasifikasi non-linear dengan menciptakan transformasi yang memungkinkan pemisahan data yang kompleks.
b. Pemisahan Data dengan Hyperplane SVM mencari hyperplane terbaik yang dapat memisahkan kelas-kelas data. hyperplane tersebut berusaha memisahkan kelas-kelas berdasarkan tingkat pencemaran udara.
c. Penentuan Vektor Pendukung Setelah hyperplane terbaik ditemukan, SVM mengidentifikasi vektor pendukung. Vektor pendukung menggambarkan titik-titik data yang berada paling dekat dengan hyperplane pemisah, yang disebut juga sebagai batas keputusan vektor pendukung mungkin mencakup data-data yang mewakili kondisi-kondisi tertentu yang memiliki dampak signifikan terhadap kualitas udara, vektor pendukung membantu memastikan bahwa hyperplane pemisah yang dihasilkan oleh SVM stabil dan mampu memisahkan dengan baik antara kelas-kelas kualitas udara yang berbeda.

4.1.2. K-Nearest Neighbor Architecture

a. Dalam konteks penggunaan algoritma K- Nearest Neighbors (KNN) untuk klasifikasi kualitas udara di Kota Jakarta, langkah- langkah algoritma KNN yang dijelaskan di atas diimplementasikan sebagai berikut:
b. Tentukan Nilai K: Langkah pertama adalah menentukan nilai K, yaitu jumlah tetangga terdekat yang akan dipertimbangkan saat mengklasifikasikan suatu data baru. Dalam kode di atas, nilai K yang diuji adalah 1, 3, 5, 7.
c. Hitung Jarak: Setelah nilai K ditentukan, algoritma KNN menghitung jarak antara data baru yang akan diprediksi dengan setiap titik data pada dataset latih. Jarak yang umum digunakan adalah jarak Euclidean dan jarak Manhattan.
d. Urutkan dan Tentukan Tetangga Terdekat: Setelah perhitungan jarak selesai, langkah selanjutnya adalah menentukan tetangga terdekat dari data baru berdasarkan nilai K yang telah ditentukan sebelumnya. Kategori dari tetangga-tetangga ini akan digunakan untuk melakukan pemilihan mayoritas kategori.
e. Prediksi Kelas Mayoritas: Setelah menentukan tetangga terdekat, langkah selanjutnya adalah menentukan mayoritas kategori dari tetangga-tetangga tersebut. Kategori yang paling sering muncul akan dijadikan prediksi untuk data baru yang sedang diproses.

f. Evaluasi Model: Setelah prediksi dilakukan, langkah terakhir adalah melakukan evaluasi performa model. Evaluasi ini melibatkan perhitungan matrik evaluasi akurasi dan konfusi matriks untuk mengukur seberapa baik model KNN dalam melakukan klasifikasi kualitas udara.

4.2. Exploratory Data Analysis (EDA)

4.2.1. Raw Dataset

Pada Gambar 5 dapat dilihat table excel pada data aktual pada pencemaran udara kota jakarta yang dimulai dari 1 Desember 2022 - 21 November 2023 dengan jumlah record sebanyak 1826.

periode_c	tanggal	stasiun	pm_sepul	pm_duakc	sulfur_dic	karbon_m	ozon	nitrogen_	max	paramete	kategori
202311	11/21/2023	DKI3 Jagak	50	73	55	12	25	13	73	PM25	SEDANG
202311	11/22/2023	DKI3 Jagak	44	63	55	13	32	10	63	PM25	SEDANG
202311	11/23/2023	DKI3 Jagak	51	82	56	14	31	16	82	PM25	SEDANG
202311	11/1/2023	DKI1 Bunc	61	86	42	8	38	41	86	PM25	SEDANG
202311	11/2/2023	DKI1 Bunc	58	84	42	6	35	38	84	PM25	SEDANG
202311	11/3/2023	DKI1 Bunc	64	89	43	9	39	37	89	PM25	SEDANG
202311	11/4/2023	DKI1 Bunc	68	97	45	8	33	41	97	PM25	SEDANG
202311	11/5/2023	DKI1 Bunc	51	63	44	8	40	29	63	PM25	SEDANG
202311	11/6/2023	DKI1 Bunc	53	73	43	8	34	29	73	PM25	SEDANG
202311	11/7/2023	DKI1 Bunc	60	85	43	7	37	30	85	PM25	SEDANG
202311	11/8/2023	DKI1 Bunc	55	74	41	6	40	25	74	PM25	SEDANG
202311	11/9/2023	DKI1 Bunc	61	89	44	6	46	30	89	PM25	SEDANG
202311	11/10/2023	DKI1 Bunc	64	107	44	8	41	33	107	PM25	TIDAK SEHAT
202311	11/11/2023	DKI1 Bunc	64	89	39	15	35	30	89	PM25	SEDANG
202311	11/12/2023	DKI1 Bunc	57	79	42	16	45	25	79	PM25	SEDANG

Gambar 5 Contoh Raw Dataset

4.2.2. Praproses Data

Pada tahap proses data peneliti memilih Python sebagai bahasa pemrograman utama untuk melakukan serangkaian operasi proses data yang diperlukan dengan Jupyter Notebook sebagai Integrated Development Environment (IDE). Penggunaan Python memudahkan peneliti dalam melakukan berbagai tugas, termasuk membersihkan data dari nilai yang hilang, mentransformasi data, mencegah overfitting dan mempersiapkan dataset agar sesuai dengan kebutuhan algoritma. Terakhir peneliti menyimpan kembali dataset tersebut dengan format .csv. Gambar 6 merupakan contoh data dari data yang sudah dibersihkan.

tanggal	pm_sepuluh	pm_duakomalima	sulfur_dioksida	karbon_monoksida	ozon	nitrogen_dioksida	kategori
12/1/2022	53	81	42	9	24	12	SEDANG
12/1/2022	52	66	18	17	27	5	SEDANG
12/1/2022	59	74	46	15	50	31	SEDANG
12/1/2022	54	73	36	12	22	13	SEDANG
12/1/2022	64	93	52	7	27	19	SEDANG
12/2/2022	55	92	43	11	24	13	SEDANG
12/2/2022	48	66	20	18	21	5	BAIK
12/2/2022	55	67	51	14	42	30	SEDANG
12/2/2022	60	94	52	9	20	18	SEDANG
12/2/2022	53	67	39	11	14	12	SEDANG
12/3/2022	49	66	20	16	32	5	BAIK
12/3/2022	60	76	38	10	22	12	SEDANG
12/3/2022	56	73	50	15	49	31	SEDANG
12/3/2022	54	74	47	8	29	11	SEDANG
12/3/2022	55	100	52	7	26	17	SEDANG
12/4/2022	65	95	52	9	18	14	SEDANG

Gambar 6 Contoh Cleaned Dataset

4.2.3. Pemisahan Data

Peneliti memisahkan data historis Pencemaran Udara Kota Jakarta menjadi dua file berformat .csv, yaitu data train dan data test. Data train digunakan untuk melatih model, sementara test set digunakan untuk menguji sejauh mana model dapat memprediksi data yang belum pernah dilihat sebelumnya. Peneliti membagi data dengan rasio 80:20 sehingga data train memiliki 1441 record dan data test mempunyai 362 record.

4.2.4. Evaluasi Matriks

Peneliti melakukan perhitungan evaluasi matriks untuk mengetahui scoring atau tingkat akurasi pada hasil prediksi. Hasil dari evaluasi matriks dapat dilihat pada tabel 2 untuk SVM dan Tabel 3 (Gambar table hasil evaluasi) untuk KNN.

Tabel 2

Hasil Evaluasi Matriks Model SVM

No	Kernel	Kategori			Mean
		Baik	Sedang	Tidak Sehat	
1	Linear	92.80%	92.80%	100%	95.20%
2	Poly	94.46%	93.91%	99.45%	95.94%
3	Rbf	96.12%	95.57%	99.45%	97.05%
4	Sigmoid	65.10%	75.07%	63.99%	68.05%

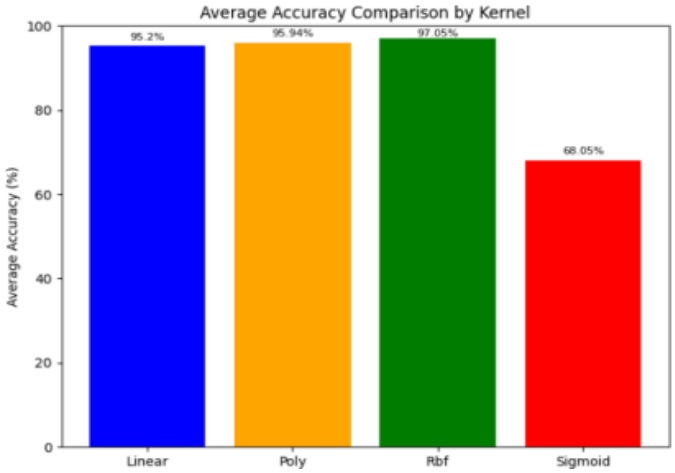
Tabel 3

Hasil Evaluasi Matriks Model KNN

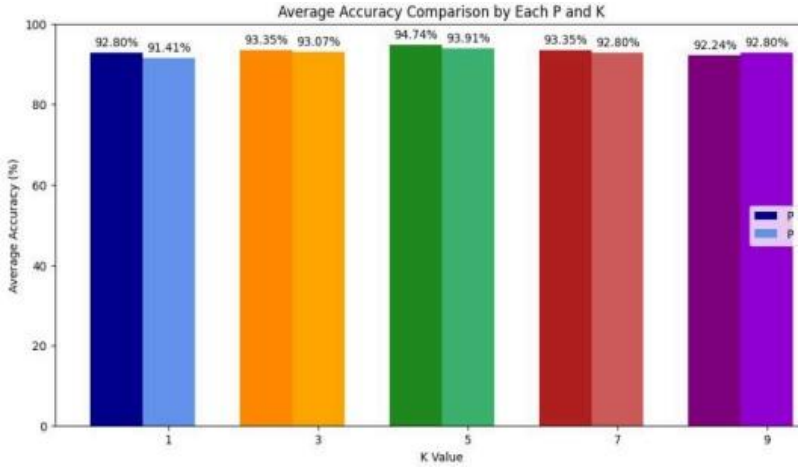
Nilai K	Hasil Akuriasi							
	p=1 (Manhattan)			Mean	p=2 (Euclidean)			Mean
	Baik	Sedang	Tidak Sehat		Baik	Sedang	Tidak Sehat	
1	78.57%	95.80%	93.02%	**92.80%**	75.00%	95.04%	90.70%	**91.41%**
3	76.79%	97.71%	88.37%	**93.35%**	75.00%	96.95%	93.02%	**93.07%**
5	80.36%	97.71%	95.35%	**94.74%**	78.57%	97.33%	93.02%	**93.91%**
7	78.57%	97.33%	88.37%	**93.35%**	75.00%	96.95%	90.70%	**92.80%**
9	73.21%	96.95%	88.37%	**92.24%**	73.21%	97.71%	88.37%	**92.80%**

4.2.5. Visualisai Data

Data visualization adalah suatu bentuk representasi grafis dari informasi dan data untuk mempermudah pemahaman pola, trend, dan relasi dalam data melalui elemen visual. Gambar 6 merupakan visualisasi untuk SVM dan Gmabar 7 merupakan visualisasi untuk KNN.



Gambar 7 Komparasi Rata-rata Akurasi Setiap Kernel Model SVM



Gambar 8 Komparasi Akurasi Nilai K dan P

4.2.6. *Confusion Matrix*

Untuk mengukur performa model, *confusion matrix* dihitung menggunakan fungsi confusion_matrix dari library scikit- learn. *Confusion matrix* memberikan informasi tentang seberapa baik model dapat membedakan antara kelas positif dan negatif, serta seberapa sering model membuat kesalahan. Gambar 9 merupakan *confusion matrix* untuk SVM dan Gambar 10 merupakan *confusion matrix* untuk KNN.

Gambar 9 Confusion Matrix Model KNN

Pada confusion matrix tersebut, model SVM dapat memprediksi nilai benar yang cukup tinggi untuk setiap kelas, dimana pada kelas 1 model berhasil memprediksi 47 nilai benar dari 56, pada kelas 2 model berhasil memprediksi 255 nilai benar dari 262, dan pada kelas 3 model berhasil memprediksi semua 43 nilai dengan benar.

Gambar 10 Confusion Matrix Model KNN

Pada confusion matrix tersebut, model KNN dapat memprediksi nilai benar yang cukup tinggi untuk setiap kelas, dimana pada kelas 1 model berhasil memprediksi 44 nilai benar dari 56, pada kelas 2 model berhasil memprediksi 256 nilai benar dari 262, dan pada kelas 3 model berhasil memprediksi 41 nilai benar dari 43.

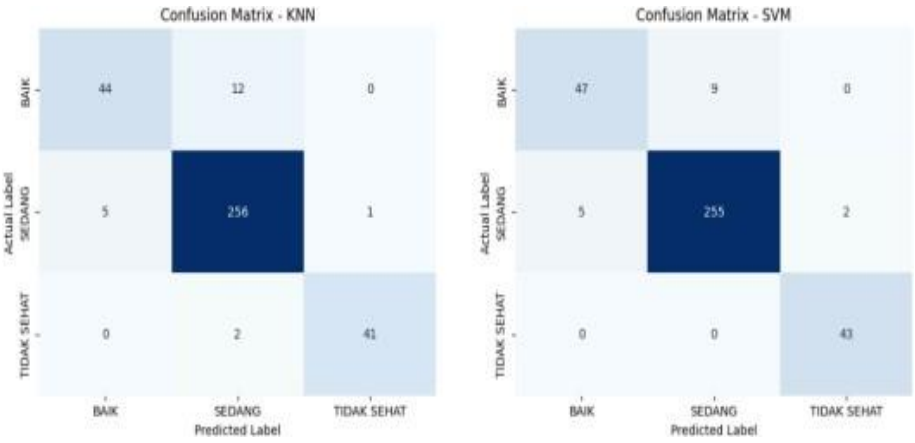
4.3. Hasil Komparasi

Peneliti kemudian menempatkan hasil klasifikasi kualitas udara menggunakan model Support Vector Machine (SVM) dan K-Nearest Neighbor (KNN) dalam bentuk tabel yang menunjukan perbandingan scoring dan perbandingan confusion matrix. Tabel 4 merupakan hasil komparasi evaluasi matriks tiap model dan gambar 11 merupakan hasil komparasi confusion matrix tiap model:

Tabel 4

Hasil Komparasi Scoring

	Model Algoritma	
Scoring	Support Vector Machine (Kernel = Rbf)	K-Nearest Neighbor (p=1, k=5)
Akurasi	97.05%	94.74%

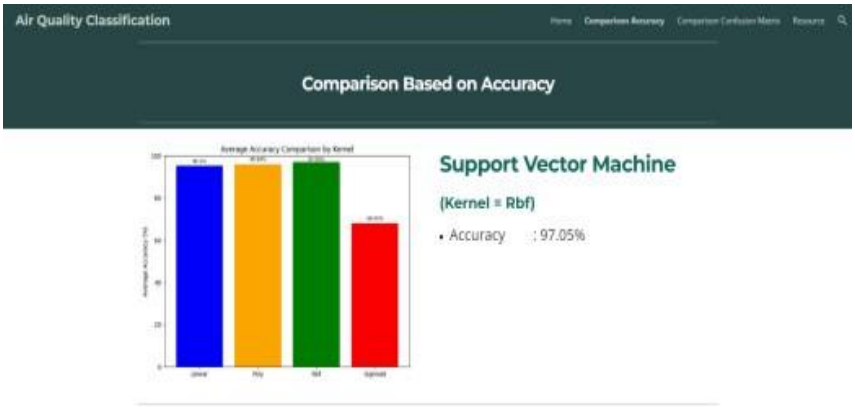


Gambar 11 Hasil Komparasi Confusion Matrix

4.4. Rancangan Graphic User Interface (GUI)

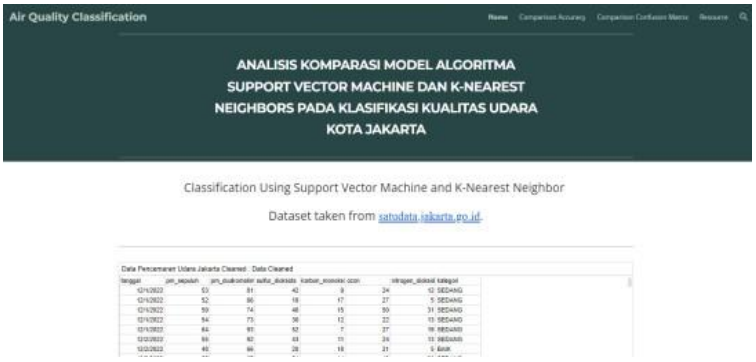
Peneliti melakukan perancangan Graphic User Interface (GUI) menggunakan Google Sites untuk mempermudah proses analisis, dan memberikan informasi mendetail mengenai hasil prediksi yang telah digunakan menggunakan algoritma Support Vector Machine (SVM) dan K- Nearest Neighbor. Pembaca dapat melihat hasil desain GUI dengan mengunjungi tautan berikut: bit.ly/Air_Quality_Classification.

Peneliti merancang 4 menu yang dapat pembaca kunjungi, yaitu menu Home, Comparison Accuracy, Comparison Confusion Matrix, dan Resource. Berikut penjelasan dan tampilan rancangan pada setiap menu.



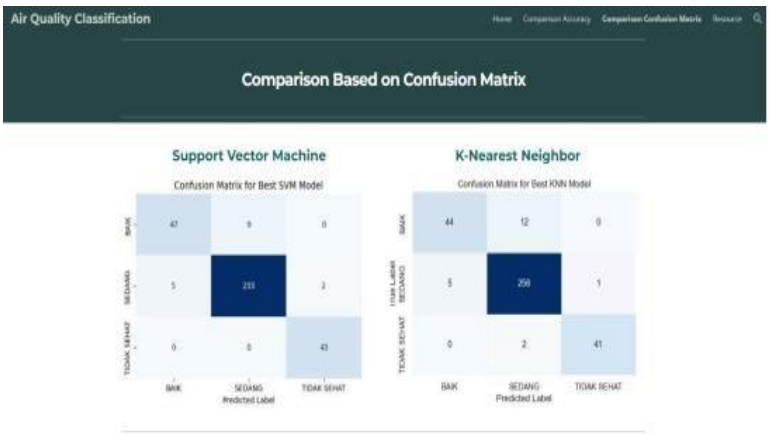
Gambar 12 Tampilan Menu Home

Gambar 12 menunjukan menu home yang berisikan penjelasan singkat yang peneliti lakukan dan dataset yang sudah di proses. Di sisi kanan atas, terdapat sub-menu yang memungkinkan pengguna untuk memilih tiga halaman berbeda, yaitu Comparison Accuracy, Comparison Confusion Matrix, dan Resource. Halaman Comparison Accuracy menampilkan perbandingan evaluasi matriks dan akurasi beserta kesimpulan pada algoritma SVM dan KNN, halaman Confusion Matrix menampilkan perbandingan konfusi matriks beserta kesimpulan pada algoritma SVM dan KNN, dan halaman Resource akan menampilkan semua source code yang digunakan peneliti dalam penelitian.



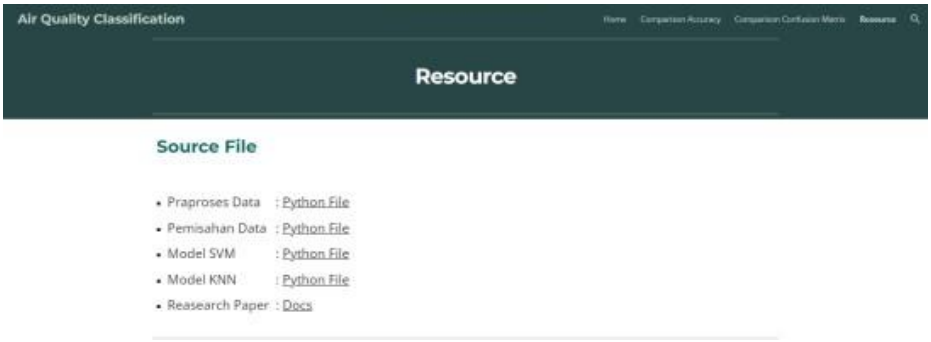
Gambar 13 Tampilan Menu Comparison Accuracy

Gambar 13 menunjukan menu comparison accuracy yang berisi perbandingan tingkat akurasi dari klasifikasi algoritma SVM dan KNN berupa visualisasi data, evaluasi matriks beserta kesimpulan yang didapatkan. Evaluasi matriks didapatkan dari kode python dan disajikan dalam bentuk tabel di google sheet.



Gambar 14 Tampilan Menu Comparison Confusion Matrix

Gambar 14 menunjukan menu comparison confusion matrix yang berisi perbandingan confusion matriks beserta dengan kesimpulan yang didapatkan antara SVM dan KNN.



Gambar 15 Tampilan Menu Resource

Gambar 15 menunjukan menu resource yang berisi source file dan kode yang digunakan peneliti untuk memprediksi klasifikasi pencemaran udara, dan karya ilmiah yang dilakukan peneliti.

5. KESIMPULAN

Hasil yang diperoleh pada penggunaan model support vector machine dan k-nearest neighbor untuk melakukan komparasi pada klasifikasi kualitas udara kota jakarta dapat disimpulkan sebagai berikut:

1. Support Vector Machine secara konsisten menunjukkan kinerja yang lebih baik dalam hal akurasi, dengan kernel Rbf yaitu 97,05% dibandingkan dengan K-Nearest Neighbor dengan nilai p=1, dan k=5 yaitu 94,74%, sehingga algoritma Support Vector Machine dalam melakukan klasifikasi kualitas udara lebih akurat dibandingkan dengan algoritma K-Nearest Neighbor.
2. Berdasarkan analisis confusion matrix untuk model Support Vector Machine (SVM) dan K-Nearest Neighbors (KNN) dalam konteks klasifikasi kualitas udara kota Jakarta, dapat disimpulkan bahwa SVM menunjukkan kinerja yang lebih baik. SVM memiliki jumlah prediksi yang benar yang lebih tinggi untuk setiap kelas, dimana pada kelas 1 model berhasil memprediksi 47 nilai benar dari 56, pada kelas 2 model berhasil memprediksi 255 nilai benar dari 262, dan pada kelas 3 model berhasil memprediksi semua 43 nilai dengan benar jika dibandingkan dengan KNN, pada kelas 1 model berhasil memprediksi 44 nilai benar dari 56, pada kelas 2 model berhasil memprediksi 256 nilai benar dari 262, dan pada kelas 3 model berhasil memprediksi 41 nilai benar dari 43. Secara spesifik, SVM memiliki nilai yang lebih tinggi untuk kelas 0, dan 2, walaupun KNN memiliki nilai lebih tinggi untuk kelas 1, nilai perbedaan tersebut tidak signifikan sehingga secara keseluruhan SVM 62 menunjukkan kemampuan yang lebih baik dalam mengidentifikasi kelas-kelas tersebut dengan akurat. Oleh karena itu, berdasarkan evaluasi confusion matrix, SVM lebih diindikasikan sebagai pilihan yang lebih baik untuk prediksi klasifikasi kualitas udara kota Jakarta dibandingkan dengan model KNN. Oleh karena itu, SVM lebih diindikasikan sebagai pilihan yang lebih baik untuk klasifikasi kualitas udara kota Jakarta dibandingkan dengan model KNN.
3. Secara keseluruhan berdasarkan prediksi model untuk setiap kelas dapat disimpulkan bahwa kualitas udara pada rentang waktu 1 tahun secara keseluruhan mungkin cukup baik, dengan kualitas sedang di angka lebih dari sama dengan 50 sampai 100. Namun, tetap diperlukan pengendalian untuk daerah kategori tidak sehat.

DAFTAR PUSTAKA

Arhami, M, dan Muhammad Nasir (2020), Data Mining Algoritma dan Implementasinya, Edisi ke -1, Yogyakarta: ANDI (Anggota IKAPI).

Bhatia, P. (2019). Data Mining and Data Warehousing: Principles and Practical Techniques. Cambridge University Press.

Goodfellow, I,. Bengio, Y, dan Courville Aaron (2016), Deep Learning , MIT Press book.

Laudon, K. C., & Laudon, J. P. (2017). Management Information Systems: Managing the Digital Firm. Pearson.

N. R., & Lichtendahl, K. C. (2017). Data Mining for Business Analytics: Concepts, Techniques,

and Applications in R. Wiley.
Ning-Tan, Pang. et al (2019), Introduction to Data Mining Second Edition, United Kingdom: Pearson Education Limited.
Raja, R. et al (2022). Data Mining and Machine Learning Applications. Wiley.
Shmueli, G., Bruce, P. C., Yahav, I., Patel,
Situmorang, Syafrizal H. et al (2014). Analisis Data Untuk Riset Manajemen dan Bisnis, Edisi 3, Medan: USU Press.
Tiwary, A., & Williams, I. (2019). Air Pollution: Measurement, Modelling and Mitigation. CRC Press, Taylor & Francis Group.
Zaki, J. M. dan Wagner Meira JR (2020), Data Mining and Machine Learning Fundamental Concepts and Algorithms, University Printing House, Cambridge CB2 8BS, United Kingdom.