Single Page Aplikasi Website Prediksi Kualitas Udara *What The Air*

1st Muhammad Abdurrahman Al - Jauzy
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
aljauzy@student.telkomuniversity.ac.id

2nd Sri Suryani Prasetyowati
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
srisuryani@telkomuniversity.ac.id

3rd Yuliant Sibaroni
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
yuliant@telkomuniversity.ac.id

Abstrak-Polusi udara biasa diartikan sebagai pencemaran udara dimana jumlah bahan pencemar berada diluar batas. Kualitas udara saat ini adalah salah satu faktor penting dalam kehidupan sehari - hari. Terlalu banyak menghirup udara dengan kualitas yang rendah dapat berdampak buruk pada kesehatan. Dengan menggunakan alat pengukur kualitas udara kita bisa mengukur tingkat indeks kualitas udara, namun kenapa hanya berhenti disitu jika kita bisa menggunakan *Machine Learning* untuk melakukan Prediksi dalam beberapa tahun kedepan. Di studi ini digunakan metode *Support Vector Machine* yangakan melakukan klasifikasi terhadap data yang didapat dari sensor. SVM dipilih karena dinilai baik dalam mengklasifikasikan data yang berupa kelas - kelas. Data yang diolah adalah SO_2, NO_2, *CO*, PM_{10}, PM_{25} dan O_3. Kemudian data hasil klasifikasi akan diproses untuk prediksi dengan teknik perluasan model. Penelitian ini akan menghasilkan pemetaan prediksi polusi udara di provinsi jakarta untuk tahun 2022, diharapkan penelitian ini dapat membantu masyarakat untuk mengetahui tentang kondisi udara.

Kata Kunci— kualitas udara, *support vector machine*, klasifikasi, *machine learning*.

Abstract-Air pollution is interpreted as air pollution where the amount of pollutants is outside the limit. Air quality today is one of the important factors in everyday life. Breathing too much air of low quality can harm health. Using air quality gauges, we can measure the level of the air quality index, but why stop there if we can use Machine Learning to make predictions in the next few years. In this study, use the Support Vector Machine method will classify the data obtained from the sensor. SVM was elected because it was considered good at classifying data in the form of classes. The processed data are SO2, NO2, CO, PM10, PM25, and O3. Then, the data from the classification will be processed for prediction with the model extension technique. This research will produce a mapping of prediction of air pollution in the province of Jakarta for 2022, hope that this research can help the public to know about air conditions.

Keywords— air pollution, support vector machine, classification, machine learning.

I. PENDAHULUAN

A. Latar Belakang

Pencemaran udara merupakan salah satu masalah lingkungan yang paling serius bagi banyak negara. Pencemaran udara diartikan sebagai peristiwa berupa gas pencemar udara menggantikan fungsi udara, yang dapat mengganggu aktivitas makhluk hidup. Terdapat beberapa gas pencemar udara yang sangat berbahaya seperti karbon, NO_2, SO_2, dan partikulat (PM) yang dapat mempengaruhi kualitas hidup, gangguan Kesehatan dan pencemaran lingkungan[3]. Berdasarkan informasi ditppu.menlhk.go.id terdapat 5 kota di Indonesia yang dikategorikan pencemaran udara yang berbahaya yaitu Jambi, Palembang, Palangkaraya, Pekanbaru dan Pontianak yang penyebab utama adalah kebakaran hutan. Tahun 2020 adanya penerapan Pembatasan Sosial Berskala Besar di Kota Jakarta tidak mempengaruhi tingkat polusi PM2.5 yang tetap tinggi, meskipun nilai konsentrasi NO2 turun sebanyak 33% namun Kota Jakarta tetap masuk kedalam 5 Kota Besar dengan kualitas udara terburuk[12].

Efek kesehatan yang dapat ditimbulkan dari buruknya kualitas udara jika terpapar dalam konsentrasi yang tinggi bisa beragam dari yang ringan seperti iritasi, hingga ekstrim seperti kanker atau kematian [2]. Sehingga perlu dilakukan pemantauan kualitas udara untuk mengevaluasi kualitas udara dan memprediksi konsentrasi pencemaran diwaktu mendatang secara akurat.

Pengklasifikasian kualitas udara tersebut dapat menggunakan metode Support Vector Machine dengan menggunakan data yang berasal dari pembacaan sensor. Sensor yang digunakan dalam pendeteksian CO, CO2, HC, debu/PM10 dan temperatur yaitu TGS-2442, TGS-2611, MG-811, GP2Y1010AU0F dan DHT-11. Setelah dilakukan pengujian, diperoleh hasil dengan akurasi keakuratan klasifikasi sebesar 95,02% [3]. Penggunaan SVM bertujuan untuk mengolah dataset yang memiliki banyak dimensi [13] sehingga akan cocok digunakan dalam melakukan prediksi kualitas udara. Beberapa penelitian tentang polusi udara telah dilakukan dengan berbagai metode pendekatan *machine learning*. Seperti *Random Forest, Decision Tree, Support Vector Machine, KNN, Naive Bayes*. Dalam penelitian yang dilakukan oleh Bingchun Liu dan timnya pada data kualitas udara 4 kota besar china tahun 2019, didapat bahwa metode SVM dengan kernel BRF menghasilkan performa dan hasil yang lebih baik dibandingkan dengan ANN dan KNN dalam mengklasifikasikan data non-linier. Akurasi yang didapat akan meningkat jika digunakan model hybrid SVM dengan *Information Gain*. Akurasi yang didapat adalah rata-rata masing-masing SVM sebesar 86,25%, ANN sebesar 82%, KNN sebesar 78,25% [1]. Ini dapat dikonfirmasi dari penelitian yang dilakukan Elia Georgiana D. yang menggunakan K-NN sebagai metode klasifikasinya dalam mengklasifikasikan kualitas udara yang menghasilkan akurasi hanya 65,51%. Namun menurut Elia, ini diklaim sudah cukup baik karena data yang digunakan hanya sebanyak 29 data [7]. Dalam penelitian lain yang dilakukan oleh Ade Silvia Handayani dan tim menggunakan metode SVM untuk klasifikasi polusi udara di indonesia bahkan menghasilkan akurasi yang sangat besar yaitu 95.02% [3]. Dalam ketiga penelitian tentang polusi udara ini berfokus pada pembangunan model klasifikasi tanpa prediksi dan pemetaan. Dalam penelitian yang dilakukan oleh S. V. Kottur and D. S. S. Mantha, dalam penelitiannya tentang model integrasi antara ANN dan Kriging untuk prediksi polusi udara mampu memprediksi nilai polutan untuk lokasi yang tidak diketahui sehingga memungkinkan untuk memperkirakan nilai polutan udara pada tingkat yang tidak terpantau lokasi [10]. Namun dalam penelitian ini, prediksi yang dilakukan hanya dapat memprediksi 3 hari kedepan dan tidak menghasilkan peta prediksi. Xiaoxiao zhang dkk juga melakukan penelitian dengan prediksi yaitu dengan menggunakan SVM dan IDW untuk meneliti data curah hujan tahunan. Berdasarkan penelitian [17] menggunakan *machine learning* untuk melakukan peramalan menggunakan data *time series.* Pada penelitian lainnya [18] *machine learning* digunakan untuk memprediksi tekanan air menggunakan 3 jenis metode *machine learning* yaitu RNN, LSTM, dan GRU dengan menggunakan data masukan non-linear time series.

Berdasakan hasil penelitian sebelumnya, disimpulkan bahwa *Support Vector Machine* efektif untuk mengklasifikasikan data spatial time kualitas udara. Juga dengan mengimplementasikan prediksi sangat membantu memperkirakan polusi untuk waktu kedepannya. Namun dalam penelitian - penelitian yang sudah dilakukan tersebut belum ada penerapan prediksi klasifikasi dengan menggunakan teknik perluasan model dan untuk tahun - tahun berikutnya. Oleh karena itu dalam penelitian ini berdasarkan koordinat dan waktu digunakan teknik klasifikasi perluasan model untuk memprediksi kualitas udara beberapa tahun ke depan dibarengi dengan metode *Support Vector Machine* sebagai metode pengklasifikasi yang kemudian akan di visualisasikan dalam bentuk aplikasi *What The Air.*

B. Rumusan Masalah

Berdasarkan latar belakang diatas, terdapat beberapa rumusan masalah yang diteliti, yaitu:

1. Bagaimana mengklasifikasikan kualitas udara menggunakan metode SVM?
2. Apakah SVM merupakan metode Machine Learning yang tepat untuk memprediksi kualitas udara ?
3. Bagaimana hasil prediksi kualitas udara dengan teknik perluasan model ?

C. Tujuan

Berdasarkan rumusan masalah yang telah disampaikan, terdapat beberapa tujuan yang hendak dicapai dalam penelitian ini adalah :

1. Untuk mengklasifikasikan kualitas udara menggunakan menggunakan metode SVM.
2. Untuk mengetahui apakah SVM merupakan metode Machine Learning yang tepat untuk memprediksi kualitas udara.
3. Mengetahui kualitas teknik perluasan

model untuk prediksi data.

D. Rencana Kegiatan

1. Pengumpulan Data

Melakukan *gathering data* yang diperlukan dengan mencari kedalam sumber-sumber terpercaya seperti ISPU, Kaggle, dan lain-lain. Data yang dicari berupa data kadar SO_2, NO_2, *CO*, PM_{10}, PM_{25} dan O_3 untuk tahun 2016 - 2021, dan tahun-tahun sebelumnya untuk beberapa daerah wilayah jawa.

2. Kajian Pustaka

Mengumpulkan referensi pendukung untuk penelitian yang berupa rumus - rumus, definisi, metode, dan hasil dari jurnal penelitian, buku maupun internet.

3. Rancangan Penelitian

Membuat rancangan model yang terbaik dengan data yang telah dikumpulkan untuk memprediksi peta prediksi polusi udara tahun 2022 dan seterusnya dengan metode *Support Vector Machine* dan *Inverse Distance Weighted*.

4. Perancangan Sistem

Setelah data diperoleh akan dilakukan persiapan data yang kemudian akan diteruskan dengan *training model* dengan menggunakan metode *Support Vector Machine* (SVM). Setelah model terbaik didapat, maka akan dilanjutkan dengan mencari prediksi yang merupakan hasil yang diharapkan. Terakhir, implemetasi model *machine learning* kedalam aplikasi *What The Air*.

Hasil yang didapat dari model akan dibandingkan dengan penelitian serupa yang menggunakan metode lainnya yang kemudian akan disimpulkan tingkat efisiensi dan keakuratan model dan metode dalam laporan evaluasi penelitian.

5. Hipotesis

Dalam penelitian ini digunakan metode *Support Vector Machine* (SVM) karena dinilai efektif untuk memprediksi kualitas udara dari data spasial kualitas udara. SVM memiliki kelebihan generalisasi, yaitu mampu mengklasifikasikan suatu *pattern* yang terdapat pada dataset. Vapnik menjelaskan bahwa *generalization error* dipengaruhi oleh dua faktor: error terhadap *training set*, dan satu faktor lagi yang dipengaruhi oleh dimensi VC (Vapnik-Chervokinensis). Strategi dalam implementasi *Machine Learning* umumnya dilakukan dengan pendekatan meminimalisir error pada training set. Strategi ini dinamakan *Empirical Risk Minimization* (ERM). Selain ERM ada juga SRM (*Structural Risk Minimization*), yaitu dengan memilih *hyperlane* dengan margin terbesar [4].

SVM juga memiliki kelebihan *curse of dimensionality* yang dapat didefinisikan sebagai masalah yang dihadapi suatu metode *pattern* recognition dalam menentukan parameter yang berkaitan dikarenakan jumlah data *training* yang sedikit dibandingkan dimensi ruang vektornya. Semakin tinggi dimensi ruang vektornya maka akan diperlukan lebih banyak lagi data sampel yang akan digunakan sebagai training-set [4]. Pada kenyataannya seringkali terjadi, data yang diolah berjumlah terbatas, dan untuk mengumpulkan data yang lebih banyak tidak mungkin dilakukan karena kendala biaya dan kesulitan teknis. Vapnik membuktikan bahwa tingkat generalisasi yang diperoleh oleh SVM tidak dipengaruhi oleh dimensi dari input vector. Faktor ini yang menjadi alasan mengapa SVM yang tepat berkaitan dengan penilitian paper ini.

Dataset yang didapat dari SVM selanjutnya akan dilakukan klasifikasi prediksi untuk memperkirakan data tahun berikutnya.

II. KAJIAN TEORI

Polusi udara telah menjadi salah satu masalah lingkungan yang paling serius bagi banyak negara, termasuk di Indonesia. Masalah ini akan berdampak pada perkembangan sosial dan ekonomi negara mana pun di dunia. Penyebabnya karena masuknya berbagai polutan berbahaya di udara yang menyebabkan ketidaknyama -nan bagi spesies hidup dan merusak lingkungan dan iklim kita. Zat-zat yang tidak diinginkan yang ditambahkan ke atmosfer melalui operasi industri dan manufaktur, pembakaran bahan bakar fosil, emisi mobil dan beberapa proses alami disebut polutan udara. Akibat drastis dan dampak buruk pencemaran udara ini telah mengalihkan perhatian pihak berwenang, peneliti dan masyarakat umum terhadap bidang kualitas udara. Oleh karena itu, ada kebutuhan mendesak untuk pemodelan, perencanaan dan peramalan kualitas udara. Prediksi dan klasifikasi

berfungsi sebagai dua komponen utama yang memberikan informasi kualitas udara kepada pihak berwenang terlebih dahulu agar dapat segera mengambil tindakan yang diperlukan untuk kesejahteraan masyarakat [1].

Agar dapat diketahui jumlah kadungan polutan di udara, stasiun cuaca didirikan di titik-titik tertentu di beberapa lokasi secara teratur untuk memperoleh data dan menginformasikan orang-orang tentang kualitas udara. Dalam aplikasi Smart City, ini bertujuan untuk melakukan proses dengan kecepatan dan akurasi yang lebih tinggi dengan mengumpulkan data dengan ribuan sensor berbasis Internet of Things (IoT). Pada tahap ini, kecerdasan buatan dan *Machine Learning* memainkan peran penting dalam menganalisis data yang akan diperoleh. Dalam penelitian ini, terdapat enam konsentrasi polutan yang digunakan untuk proses pengklasifikasian; partikel ($PM_{2.5}$ dan PM_{10}), nitrogen dioksida (NO_2), belerang dioksida (SO_2), Ozon (O_3), dan karbon monoksida (CO) [5]. Belerang dioksida (SO_2) merupakan gas tidak bewarna yang termasuk kedalam gas pencemar udara, di udara kadarnya mencapai 18%. Gas ini memiliki bau yang menyengat dang berbahaya bagi makhluk hidup. Apabila SO_2 dan NO_2 bertemu di udara ini akan membentuk senyawa H_2SO_4 yang dapat membentuk hujan asam, yang bersifat korosif terhadap metal [14]. Sumber penyebabnya yaitu kendaraan bermotor (1%), pabrik, generator, dan pemanas (99%) [15]. Pengaruh konsentrasi SO_2 yang melebihi batas yang diperbolehkan akan mempengaruhi kesehatan makhluk hidup. SO_2 dapat mengiritasi dan lebih dari 95% SO_2 akan terhirup selama proses respirasi. Nitrogen dioksida (NO_2) digunakan sebagai bahan baku sintetis untuk menghasilkan asam nitrat, yang outputnya berjumlah jutaan ton per tahun. Gas ini berwarna coklat kemerahan dan bersifat racun, berbau menyengat, dan merupakan salah satu pencemar udara utama [16]. Dari penelitian tersebut perlu segera dilakukan pengklasifikasian kualitas udara.

Perlunya diketahui kualitas udara di suatu tempat agar dapat dilakukan klasifikasi kualitas udara sehingga dapat diketahui berapa jumlah polutan di udara daerah tersebut. Umumnya kualitas udara dinilai berdasarkan konsentrasi parameter pencemaran udara yang terukur, apakah melebihi nilai Baku Mutu Udara Ambien Nasional atau dibawah nilai tersebut. Baku Mutu Udara merupakan nilai batas kadar pencemaran udara yang dapat diterima keberadaanya dalam udara ambien. Pengertian dari udara ambien adalah udara bebas yang berada di permukaan bumi pada lapisan troposfer yang berada di area yurisdiksi Republik Indonesia dan mempengaruhi kesehatan makhluk hidup didalamnya (Kurniawan, 2017).

Komposisi udara normal terdiri dari N_2 dengan konsentrasi 78,09%, O_2 dengan konsentrasi 20,95%, CO_2 dengan konsentrasi 0,93%, dan selebihnya adalah gas Ar, Ne, He, Kr, Xe sebanyak 0,03% (BLH Prop.Sumut, 2010).

Penetapan Baku mutu udara ambien dalam PP No 41 Tahun 1999 sebagai pencegah terjadinya pencemaran udara serta melindungi kesehatan dan kenyaman masyarakat. Nilai baku mutu udara ambien seperti pada table 1 berikut.

TABEL 2-1.
BAKU MUTU UDARA AMBIEN BERDASARKAN PP NO 41 TAHUN 1999

No.	Parameter	Waktu	Baku Mutu
1.	Aerosol (PM_{10})	24 Jam	150 $\mu g/m^3$
2.	Karbonmonoksida (CO)	1 Jam	30000 $\mu g/m^3$
		24 Jam	10000 $\mu g/m^3$
3.	Ozon (O_3)	1 Jam	235 $\mu g/m^3$
		1 Tahun	50 $\mu g/m^3$
4.	Sulfurdioksida (SO_2)	24 Jam	365 $\mu g/m^3$
		1 Tahun	80 $\mu g/m^3$
5.	Nitrogendioksida	1 Jam	0,25 $\mu g/m^3$
		1 Tahun	100 $\mu g/m^3$

Menurut Kurniawan (2017) ISPU atau Indeks Standar Pencemaran Udara diartikan sebagai angka yang mendeskripsikan kondisi mutu udara ambien suatu wilayah, berdasarkan dampak terhadap Kesehatan manusia, nilai estetika dan makhluk hidup lainya. Parameter nilai ISPU mengikuti Keputusan Menteri Negara Lingkungan Hidup No. 45 Tahun 1997

Tentang Indeks Standar Pencemaran Udara, seperti pada tabel berikut.

TABEL 2-2. INDEKS STANDAR PENCEMARAN UDARA

Kategori	Rentang	Penjelasan
Baik	0-50	Tingkat kulitas udara yang tidak memberikan efek bagi Kesehatan manusia atau hewan dan tidak berpengaruh pada tumbuhan, bangunan atau nilai estetika.
Sedang	51-100	Tingkat kualitas udara yang tidak berpengaruh pada Kesehatan manusia ataupun hewan tetapi berpengaruh pada tumbuhan yang sensitive, dan nilai estetika.
Kategori	**Rentang**	**Penjelasan**
Tidak Sehat	101-199	Tingkat kualitas udara yang bersifat merugikan pada manusia ataupun kelompok hewan yang sensitive atau bisa menimbulkan kerusakan pada tumbuhan ataupun nilai estetika.
Sangat Tidak Sehat	200-299	Tingkat kualitas udara yang dapat merugikan Kesehatan pada sejumlah segmen populasi yang terpapar.
Berbahaya	300-lebih	Tingkat kualitas udara berbahaya yang secara umum dapat merugikan kesehatan yang serius.

Metode penghitungan indeks polusi udara yang ada saat ini rumit dan memakan waktu yang cukup panjang. Oleh karena itu diperlukan teknik pemodelan baru yang akurat dan efisien perlu diusulkan seperti pada penelitian [11] dengan demikian support vector machine diusulkan dalam penelitian ini untuk memodelkan indeks polusi udara. Ada tiga parameter utama yang mempengaruhi kinerja support vector machine: faktor penalti (C), parameter regularisasi (ε) dan jenis fungsi kernel yang digunakan. Namun, dalam penelitian ini, hanya parameter model fungsi kernel yang diselidiki. Hasil model dianalisis dengan menggunakan sum of squares error (SSE), mean of sum of squares error (MSSE) dan koefisien determinasi (R2). Ditemukan bahwa model yang diusulkan menggunakan fungsi kernel radial basis function (RBF) secara efektif dan akurat mampu menyelesaikan masalah pemodelan indeks polusi udara yang kompleks dengan kesalahan sum square (SSE), mean sum square error (MSSE) dan koefisien determinasi (R2) tahun 2008, masing-masing 3.1.4440 dan 0.9843. Studi ini telah berhasil

mengembangkan model SVM yang cocok dan akurat untuk prediksi API hanya menggunakan prediktor tanpa perhitungan kompleks yang sebelumnya digunakan. Perbandingan hasil pemodelan berdasarkan fungsi kernel menunjukan pengaturan parameter SVM yang berbeda dapat memberikan hasil yang berbeda, di mana fungsi kernel terakhir digunakan di dalam penelitian ini adalah RBF. Pengaturan terbaik untuk model API SVM menggunakan fungsi kernel RBF dengan nilai R2 yaitu 0,9843 dan nilai SSE dan MSSE masing-masing 2008 dan 1,444.

Pada penelitian [1], memperkenalkan model klasifikasi hybrid baru berdasarkan teori informasi dan support vector machine (SVM) menggunakan data kualitas udara dari 4 kota di China yaitu Beijing, Guangzhou, Shanghai dan Tianjin dari 1 Januari 2014 hingga 30 April 2016. Kementerian Perlindungan Lingkungan China telah mengklasifikasikan kualitas udara harian menjadi 6 tingkat, yaitu, polusi berat, polusi berat, polusi sedang, polusi ringan, baik dan sangat baik berdasarkan nilai indeks kualitas udara (AQI) masing-masing. Menggunakan teori informasi, perolehan informasi (IG) dihitung dan pemilihan fitur dilakukan untuk fitur kategorikal dan fitur numerik berkelanjutan. Kemudian algoritma pembelajaran mesin SVM diimplementasikan pada fitur-fitur yang dipilih dengan validasi silang. Evaluasi akhir mengungkapkan bahwa model hybrid IG dan SVM berkinerja lebih baik daripada model SVM (alone), Model using Artificial Neural Network (ANN) dan Knearest neighbor (KNN) dalam hal akurasi dan juga kompleksitas.

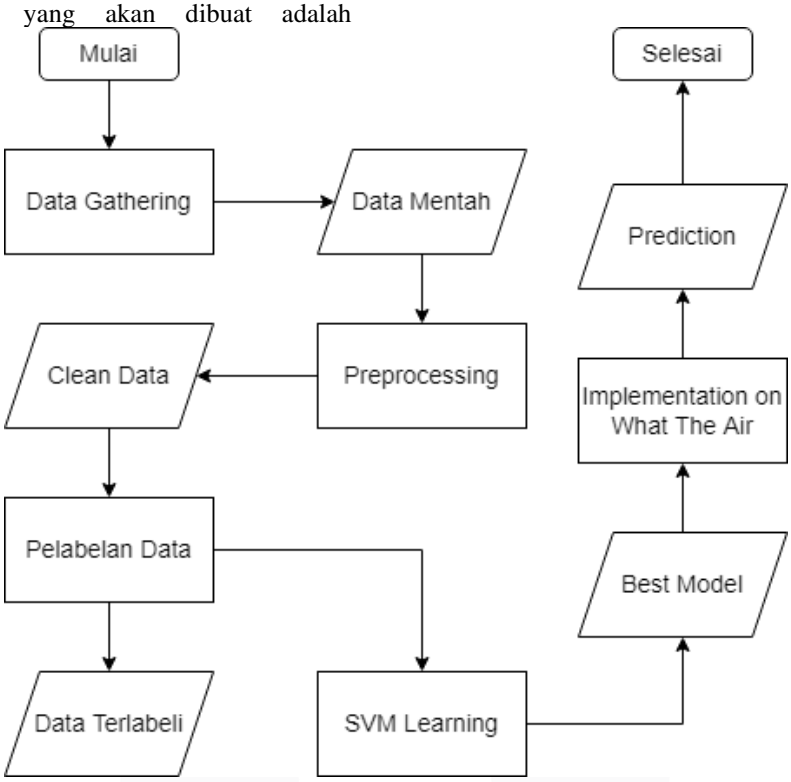
Berdasarkan penelitian [3], menyimpulkan bahwa metode Support Vector Machine telah berhasil mengklasifikasikan kualitas udara dari pembacaan sensor. Kestabilan sensor dapat mempengaruhi keberhasilan dalam klasifikasi data. Kinerja akurasi klasifikasi yang diperoleh sebesar 95,02%. Hal ini dibuktikan dengan hasil simulasi data uji yang mampu memisahkan dua kelas data yang berbeda. Namun, Berdasarkan penelitian, penggunaan sensor yang tidak dikondisikan dengan baik dapat menyebabkan sistem menghasilkan

pembacaan sensor yang salah sehingga dapat mengakibatkan kesalahan klasifikasi dalam pengujian.

III. METODE

Sistem yang akan dibuat adalah klasifikasi dan prediksi kualitas udara di beberapa wilayah jawa menggunakan metode *Support Vector Machine (SVM)* berikut adalah flowchart dari proses pembuatanya:

Gambar 3.1 Gambaran Umum Sistem



A. Data Gathering

Penelitian ini menggunakan data sekunder sebagai Dataset. Data Sekunder merupakan data yang telah diolah oleh suatu instansi atau berasal dari penelitian sebelumnya. Data sekunder yang digunakan pada penelitian ini adalah SO_2, NO_2, *CO*, PM_{10}, PM_{25} dan O_3. Data ini masih disebut sebagai data mentah karena perlu dilakukan pengolahan pada data sebelum dapat digunakan lebih lanjut.

B. Preprocessing

Pada SVM Learning, data yang akan digunakan sebelumnya harus dilakukan pengubahan bentuk menjadi data yang terstruktur untuk kebutuhan modeling. Data hasil preprocessing disebut sebagai data bersih atau *clean data* karena sudah memiliki struktur yang dapat digunakan. Proses preprocessing ini tersusun dalam beberapa tahap, yaitu:

C. Filling Missing Data

Tahap ini berguna untuk menghapus baris atau mengisi kolom yang memiliki value kosong. Data yang digunakan untuk mengisi kolom data yang kosong adalah dengan mengambil nilai mean dari kolom tersebut.

D. Normalization

Tahap normalisasi berguna untuk menyamakan range data sehingga seragam dalam rentan 0 - 1. Sehingga tidak akan ada variabel yang menjadi kurang berpengaruh karena range data yang berbeda. Rumus yang digunakan adalah :

$$norm(x) = \frac{x - min}{max - min} \quad (1)$$

Keterangan :
x = nilai data pada kolom
min = nilai minimal kolom
max = nilai maksimal kolom

E. Feature Optimization

Tahap ini berguna untuk menyeleksi atribut yang akan digunakan dalam proses SVM *learning*. Penyeleksian atribut ini dilakukan dengan tujuan untuk mengurangi dimensi data untuk mengurangi tingkat kesulitan dan hasil yang optimal. Metode

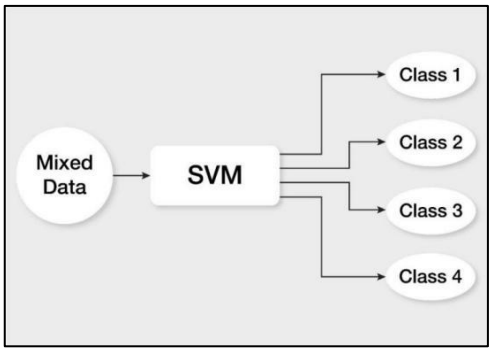
yang dapat digunakan adalah dengan menggunakan matriks korelasi.

F. Pelabelan Data

Hasil clean data ini yang akan dilakukan proses lebih lanjut. Aspek dari setiap dataset tersebut, kemudian dilabeli secara manual. Pelabelan ini di klasifikasikan menjadi 5 berdasarkan Keputusan Menteri Negara Lingkungan Hidup No. 45 Tahun 1997 Tentang Indeks Standar Pencemaran Udara yaitu baik, sedang, tidak sehat, sangat tidak sehat, dan berbahaya.

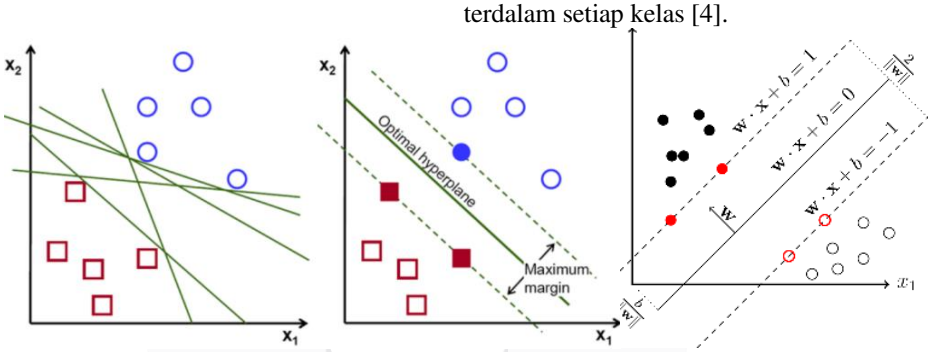
G. SVM Learning

SVM adalah algoritma *machine learning* yang biasa digunakan untuk menyelesaikan permasalahan klasifikasi big data. Pertama kali diperkenalkan oleh Vapnik tahun 1979 [8]. Dalam pemprosesannya, model terbentuk ditengah proses seperti pada gambar 3.2. Sebelum masuk ke dalam SVM, akan dilakukan data preprocessing terlebih dahulu.



GAMBAR 3.2.
PROSES KLASIFIKASI PREDIKSI SVM

Dalam SVM diperkenalkan *hyperplane* sebagai pembatas antar kelas. Dalam gambar 3.3, ada beberapa *hyperlane* yang memungkinkan sebagai pembatas dua kelas. Namun, dalam SVM tujuan yang dicari adalah *hyperplane* paling optimal sebagai pembatas dua kelas agar didapat akurasi maksimal. *Hyperplane* optimal ini akan memiliki margin pembatas yang sama besarnya antara titik terdalam setiap kelas [4].



GAMBAR 3.3.
HYPERLANE SVM

Berdasarkan illustrasi gambar 3.3, misal kita definisikan kelas (+1) adalah titik berwarna merah dan kelas (-1) adalah titik berwarna biru. Maka akan didapat :

$(x_i, y_i) \in \{\pm 1\}, 1 \leq i \leq N.$ (3)

Dua kelas yang ada (+1 dan -1) akan didapat dari fungsi berikut setelah proses training :

$f\, w, b(x) = \text{bip_sign}(w.x + b)$ (4)

Dengan variabel w sebagai koefisien vektor dan b sebagai bias. Fungsi bip_sign merupakan fungsi *bipolar sign activation* yang akan menentukan kelas. Fungsi bipolar dapat direpresentasikan seperti berikut :

$f(n) = 1$ saat $n > 0$, 0 saat $n = 0$, dan -1 saat $n < 0$ (5)

Seperti dalam gambar 2, Agar *hyperplane* optimal kondisi berikut harus dipenuhi oleh hyperplane :

$yi[w.xi + b] \geq 1, i = 1,2, \ldots\ldots, N$ (6)

Dalam implementasi SVM ada beberapa kernel yang menjadi patokan dalam penentuan kelayakan implementasi model ini pada suatu masalah. Kernel adalah fungsi yang mengambil data input dan merubahnya menjadi bentuk yang diperlukan untuk proses. Fungsi kernel ini menyediakan *shortcut* untuk menghindari perhitungan yang kompleks. Kelayakan implementasi SVM umumnya dapat diambil dari perbandingan tiga kernel yaitu linear, polinomial, dan RBF. Penggunaan metode SVM dipenelitian ini ditujukan untuk mengklasifikasikan data polusi menjadi lima bagian yaitu baik, sedang, tidak sehat, sangat tidak sehat, dan berbahaya.

Untuk mendapatkan best model, dilakukan evaluasi model dengan metode *cross-valdiation* dan menggunakan beberapa ukuran statistik seperti akurasi dan *f-measure* sebagai alat ukurnya. Ukuran statistik ini diukur menggunakan *multi-class confussion matrix.*

Berdasarkan *confusion matrix* yang ada dapat dihitung performansi klasifikasi yang dibangun antara lain:

1. Akurasi

Akurasi adalah persentase dari semua sampel yang diklasifikasikan dengan benar oleh pengklasifikasi. Akurasi dapat dirumuskan sebagai berikut:

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN}$$

7)

2. Precision

Precision adalah rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positf. *Precision* dapat dirumuskan sebagai berikut:

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

3. Recall

Recall adalah rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif. *Recall* dapat dirumuskan sebagai berikut:

$$Recall = \frac{TP}{TP+FN} \quad (9)$$

4. F1-Score

F1-Score adalah merupakan perbandingan rata-rata presisi dan recall yang dibobotkan. *F1-Score* dapat dirumuskan sebagai berikut:

$$F1 - Score = \frac{2(Precision\ x\ Recall)}{Precision+Recall}$$

(10)

IV. HASIL DAN PEMBAHASAN

Pada bab ini dilakukan akan dilakukan implementasi machine learning dengan menggunakan analis perancangan yang dibuat pada bab sebelumnya.

A. Skenario Uji

Dataset ISPU DKI Jakarta yang didapat digabung dengan data jumlah pohon, kendaraan, populasi, dan curah hujan untuk mendapatkan data baik untuk pemodelan.

Dibuat dua model dari pengerjaan Tugas Akhir ini untuk perbandingan menentukan hasil yang baik. Model pengujian menggunakan dataset yang telah dibagi menjadi dua bagian, yaitu data training dan data testing, dengan rasio keduanya secara berurutan 80% dan 20%. Kedua model yang dibuat masing - masing berupa model dengan atribut tambahan dan model tanpa atribut tambahan. Model tanpa atribut tambahan hanya menggunakan data perluasan model dari tahun sebelumnya.

Perluasan model dibentuk dengan menggunakan 3 hari sebelumnya dari tahun sebelumnya sebagai data acuan, sehingga dapat diprediksi data untuk tahun berikutnya.

Model tanpa atribut tambahan akan digunakan jika user mengirimkan data prediksi tanpa menspesifikasikan atribut tambahan seperti jumlah pohon, kendaraan, dan lain - lain. Model dengan atribut tambahan akan digunakan jika user menspesifikan nilai atribut tambahan sebagai parameter prediksi.

Dari kedua model yang terbentuk, dibuat variasi pelatihan dengan menggunakan perluasan model dengan 3 hari dan 2 hari sebagai perbandingan model terbaik.

B. Analisis Hasil Pengujian

Hasil implementasi dari penerapan klasifikasi pada data ISPU tahun 2016 - 2021 dengan metode Support Vector Machine untuk perluasan model 3 hari didapat hasil rata - rata dari kedua model seperti pada tabel 4.1.

TABEL 4.1
HASIL UJI PERLUASAN MODEL 3 HARI

	Precision	Recall	F1-Score	Support
Sedang	76%	94%	84%	6936
Tidak Sehat	68%	48%	66%	1040
Baik	74%	53%	61%	2024
Accuracy			76%	10000
Macro avg	38%	37%	36%	10000
Weighted avg	68%	76%	71%	10000

Berdasarkan tabel 4.1 didapat hasil akurasi 76% dimana nilai dari kategori “sedang” menghasilkan precision 76%, recall 94%, f1-score 84%. Kategori “tidak sehat” menghasilkan precision 68%, recall 48%, f1-score 66%. Kategori “baik” menghasilkan precision 74%, recall 48%, f1-score 61%.

Sedangkan hasil dari penerapan klasifikasi dengan perluasan model 2 hari didapat hasil rata - rata seperti pada tabel 4.2.

TABEL 4.2
HASIL UJI PERLUASAN MODEL 2 HARI

	Precision	Recall	F1-Score	Support
Sedang	64%	79%	74%	6936
Tidak Sehat	66%	53%	62%	1040
Baik	54%	50%	49%	2024
Accuracy			61%	10000
Macro avg	33%	34%	32%	10000
Weighted avg	56%	66%	69%	10000

Berdasarkan tabel 4.2 didapat hasil akurasi 61% dimana nilai dari kategori “sedang” menghasilkan precision 64%, recall 79%, f1-score 74%. Kategori “tidak sehat” menghasilkan precision 66%, recall 53%, f1-score 62%. Kategori “baik” menghasilkan precision 54%, recall 50%, f1-score 49%.

V. KESIMPULAN

Berdasarkan hasil dari model - model yang sudah terbentuk terhadap data ISPU DKI Jakarta dengan data tambahan lainnya (curah hujan, populasi, jumlah pohon dan kendaraan) dengan metode Support Vector Machine menghasilkan hasil yang cukup baik dengan nilai rata - rata akurasi 75%, precision 76%, recall , f1-score untuk perluasan model 3 hari. Dapat disimpulkan, penggunaan metode SVM dengan tambahan attribut pada data ISPU DKI Jakarta dan penggunaan perluasan model 3 hari sangat berperan dalam pemodelan terhadap akurasi dari metode Support Vector Machine. Sehingga dalam penerapannya untuk aplikasi What The Air digunakan metode SVM dengan perluasan model 3 hari.

REFERENSI

[1] Bingchun Liu, Hui Wang, Arihant Binaykia, Chuanchuan Fu, dan Bingpeng Xiang. 2019. Multi-Level Air Quality Classification in China Using Information Gain and Support Vector Machine Hybrid Model. Nature Environment and Pollution Technology. International Quarterly Scientific Journal.

[2] Deval L. Patrick, Timothy P. Murray, Richard K. Sullivan Jr., dan Kenneth L. Kimmel. Health Environmental Effects of Air Pollution. Commonwealth of Massachusetts Departement of Environmental Protection. MassDepp.

[3] Ade Silvia Handayani, Sopian Soim, Theresia Enim Agusdi, dan Nyayu Latifah Husni. 2021. Air Quality Classification Using Support Vector Machine. Computer Engineering and Applications. State Polytechnic of Sriwijaya.

[4] AdeAnto Satriyo Nugroho, Arief Budi Witarto, dan Dwi Handoko. 2003. Support Vector Machine Teori dan Aplikasinya dalam Bioinformatika. IlmuKomputer.Com, IlmuKomputer.

[5] K. Alpan, dan B. Sekeroglu. 2020. Prediction Of Pollutant Concentration By Meteorological Data Using Machine Learning Algorithms. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. NEU, Information Systems Engineering, Turkey.

[6] Nils J. Nilsson. Introduction To Machine Learning. Stanford University. Robotic Laboratory. 1-2.

[7] Elia Georgiana Dragomir. 2010. Air Quality Index Prediction using K-Nearest Neighbor Technique. Informatics Departement. Petroleum-Gas University of Ploiesti, Romania.

[8] Vapnik V. N., The Nature of Statistical Learning Theory(2ed). .

[9] Kurniawan. 2017. Pengukuran Parameter Kualitas Udara (CO, NO2,SO2,O3, dan PM10) di Bukit Kototabang Berbasis ISPU. Jurnal Tekno Sains. 7(1-82).

[10] S. V. Kottur and D. S. S. Mantha, "An Integrated Model using Artificial Neural Network (ANN) and Kriging for Forecasting Air Pollutants using Meteorological Data," in International Journal of Advanced Research in Computer and Communication Engineering, India, 2015.

[11] A. Faudzan, S. Suryani and T. Budiawati. 2015. "PERBANDINGAN METODE INVERSE DISTANCE WEIGHTED (IDW) DENGAN METODE ORDINARY KRIGING UNTUK ESTIMASI SEBARAN POLUSI UDARA DI BANDUNG," e-Proceeding *of Engineering,* vol. 2, no. 2.

[12] Greenpeace Indonesia. 2022. Polusi udara memakan biaya Rp 21 triliun di Jakarta pada tahun 2020. https://www.greenpeace.org/indonesia/siaran-pers/5389/polusi-udara-memakan-biaya-rp-21-triliun-di-jakarta-pada-tahun-2020/

[13] A. Holleman and E. Wiberg. 2001. Inorganic Chemistry. San Diego: Academic Press.

[14] Z. Arifin.2009. Pengendalian Polusi Kendaraan. Yogyakarta: Afabeta.

[15] F. Aditya, S. Sri, B. Tuti. 2015. Perbandingan Metode Inverse Distance Weighted (idw) Dengan Metode Ordinary Kriging Untuk Estimasi Sebaran Polusi Udara Di Bandung. eProceedings of Engineering, 2(2).

[16] B. Gianluca, B.T. Souhaib, Yann. 2012. Machine Learning Strategies for Time Series Forecasting. European business intelligence summer school. Springer, Berlin, Heidelberg, 2012. p. 62-77.

[17] X. Wei, L. Zhang, H.Q. Yang, L. Zhang & Y.P. Yao. 2021. Machine learning for pore-water pressure time-series prediction: Application of recurrent neural networks. Geoscience Frontiers, 12(1), 453-467.

[18] Xiaoxiao Zhang, Guodong Liu, Hantao Wang, dan Xiaodong Li. 2017. Application of a Hybrid Interpolation Method Based on Support Vector Machine in the Precipitation Spatial Interpolation of Basins. MDPI. China.