INSECT: Jurnal Teknik Informatika
INFORMATICS AND SECURITY e-ISSN. 2614-431X

Pengaruh Seleksi Fitur Terhadap Akurasi Klasifikasi Indeks Standar Pencemar Udara Menggunakan Naïve Bayes

Rizky Caesar Irjayana*[1], Abdul Fadlil[2], Rusydi Umar[3]
[1]Program Studi Magister Informatika, Universitas Ahmad Dahlan
[2]Program Studi Teknik Elektro, Universitas Ahmad Dahlan
[3]Program Studi Informatika, Universitas Ahmad Dahlan
E-mail: ***[1]caesar.sy23@gmail.com**, [2]fadlil@uad.ac.id, [3]rusydi.umar@tif.uad.ac.id

Abstrak

Tujuan penelitian ini adalah untuk mengklasifikasikan kualitas udara sesuai dengan Indeks Standar Pencemar Udara (ISPU) menggunakan algoritma Naïve Bayes serta mengevaluasi dampak penambahan fitur $PM_{2.5}$ terhadap akurasi model dengan membagi dataset kedalam tiga kategori diantaranya BAIK, SEDANG dan TIDAK SEHAT. ISPU merupakan indikator penting dalam mengukur kualitas udara berdasarkan konsentrasi polutan seperti PM_{10}, $PM_{2.5}$, SO_2, CO, O_3, NO_2 dan HC. Dengan tingginya volume data yang dikumpulkan setiap hari, diperlukan metode klasifikasi yang efektif untuk menyampaikan data kualitas udara dengan tepat. Penelitian ini mengusulkan dua skenario klasifikasi berdasarkan Peraturan NOMOR P.14/MENLHK/SETJEN/KUM.1/7/2020, yaitu: tanpa fitur $PM_{2.5}$ dan dengan fitur $PM_{2.5}$. Evaluasi dilakukan menggunakan K-Fold Cross Validation dengan K = 2, 3, 4, dan 5, dimana K = 5 menghasilkan akurasi tertinggi. Hasil penelitian menunjukkan bahwa penambahan fitur PM2.5 meningkatkan akurasi model dari 82,89% menjadi 93% dan F1-score dari 82,6% menjadi 92,8%, menunjukkan peningkatan sekitar 10%. Kontribusi utama penelitian ini adalah analisis komprehensif terhadap dampak fitur PM2.5 serta evaluasi berbagai nilai K dalam K-Fold Cross Validation. Dengan demikian, penemuan ini dapat menjadi sumbangsih ilmu pengetahuan pada ranah pengembangan sistem pemantauan kualitas udara yang lebih akurat untuk mendukung kebijakan lingkungan dan kesehatan masyarakat.

***Kata kunci**—ISPU, Naïve Bayes, K-Fold Cross Validation, Klasifikasi, Data Mining*

1. PENDAHULUAN

Kualitas udara yang terkandung di setiap lokasi menentukan dampak kondisi kesehatan maupun lingkungan. (Indeks Standar Pencemar Udara) ISPU digunakan sebagai patokan untuk mengukur kadar polutan yang terkandung pada udara. Kualitas udara di suatu tempat selalu mengalami perubahan dari hari ke hari terutama pada era pertumbuhan teknologi dan aktivitas industri yang semakin pesat kini [1]. Berdasarkan data tahun 2021 yang dikeluarkan oleh *Institute For Health Metrics and Evaluation*, polusi udara menempati urutan ke-dua faktor resiko kematian global setelah tekanan darah tinggi. Tercatat dari setiap 100.000 populasi dunia, sekitar 102 orang meninggal dunia akibat terpapar polusi udara. Sedangkan untuk data di kota Jakarta sendiri tercatat pada tahun 2021, kematian disebabkan oleh polusi udara berada di peringkat ke-empat dengan jumlah kematian sekitar 66 orang per 100.000 populasi [2]. Terkait dengan pentingnya dampak kualitas udara terhadap kesehatan, Menteri Dalam Negeri mengeluarkan Inmendagri Nomor 2 Tahun 2023 tentang Pengendalian Pencemaran Udara pada Wilayah JABODETABEK

terhadap respon buruknya kualitas udara pada wilayah tersebut. Faktor kesehatan yang diakibatkan oleh paparan kualitas udara dapat berdampak pada kesiapan pertahanan nasional. Seperti halnya ketika anggota militer terkena dampak langsung dari kualitas udara yang buruk, mental dan imunitas tubuh ikut terpengaruh bahkan partikel-partikel korosif yang terkandung dapat merusak infrastruktur militer [3].

Seiring dengan terus bertambahnya data ISPU tiap harinya melalui titik-titik pantauan di setiap lokasi, maka dibutuhkan kemampuan pengambil keputusan yang dapat menentukan kategori udara secara cepat dan akurat. Metode klasifikasi adalah salah satu solusi yang ditawarkan dalam permasalahan pengolahan data. Namun terdapat beberapa tantangan utama dalam pengolahan data ISPU yaitu, volume data yang besar sehingga dibutukan teknik klasifikasi yang efisien mengingat setiap lokasi menghasilkan data ISPU harian, kecepatan dan akurasi klasifikasi dalam mendukung pengambilan keputusan yang baik dan juga pengaruh variasi fitur yang dapat mempengaruhi performa model. Dimana dalam pembelajaran mesin untuk pemodelan klasifikasi terdapat dua fase yaitu *training* dan *validation*, fase *training* untuk membangun model dan menghasilkan *output* yang diharapkan dari *dataset* pelatihan dan fase *validation* untuk memastikan seberapa baik model yang telah dilatih dalam menghasilkan *output* [4].

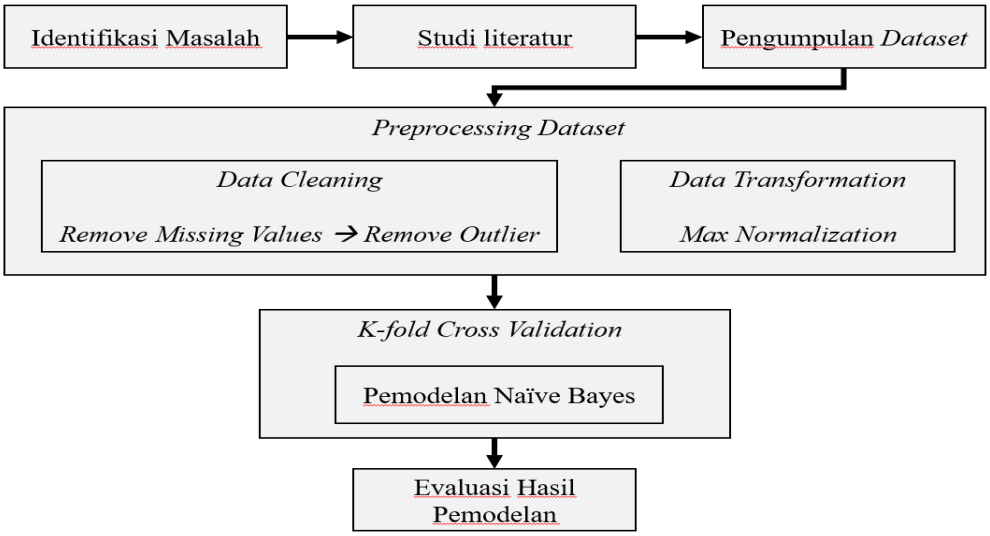
Adapun penelitian serumpun terdahulu mengenai klasifikasi kualitas udara diantaranya, Avira, dkk. (2024) melakukan komparasi metode pengklasifikasian dengan menggunakan K-NN dan *Naïve Bayes* terhadap ISPU di Kota Tanggerang Selatan untuk data tahun 2022 dengan jumlah *dataset* sebanyak 365, didapat akurasi algoritma K-NN sebesar 94,44% dan *Naïve Bayes* sebesar 86,11% untuk masing-masing penggunaan rasio 90:10 [1]. Sebelumnya, Fitri, dkk. (2023) telah menggunakan *Naïve Bayes* untuk klasifikasi ISPU Tanggerang Selatan dengan memanfaatkan *tools* RapidMiner untuk 1096 *dataset*, diperoleh akurasi sebesar 79,38% [5]. Penelitian serupa dilakukan oleh Devi, dkk (2022) melakukan klasifikasi *Gaussian Naïve Bayes* dan *k-fold cross validation* k=10 untuk ISPU DKI Jakarta tahun 2020 dengan mempertimbangkan fitur *max* dan *critical* pada *dataset*, mengakibatkan penurunan akurasi sebesar 66,7% dibandingkan tanpa menggunakan fitur *max* dan *critical* yaitu sebesar 91% dengan *recall* 93,36%, presisi 93,92% serta *f1-score* 93,68% [6]. Di sisi lain, Syekh, dkk. (2021) meneliti komparasi algoritma klasifikasi untuk ISPU DKI Jakarta tahun 2017 hingga Juni tahun 2020 dengan total 6.536 *dataset* yang dipakai, memanfaatkan *feature selection* berupa *backward elimination* dan *k-fold cross validation* k=10 maka didapati akurasi yaitu untuk SVM = 96,60% dengan waktu komputasi 11 detik, Decision Tree = 99,80% dengan waktu komputasi 0,8 detik, KNN = 97,55 dengan waktu komputasi 3 detik, *Naïve Bayes* = 91,89% dengan waktu komputasi 0,2 detik dan *Neural Network* = 98,01% dengan waktu komputasi 2 menit 28 detik [7].

Merujuk dari kasus serupa dan penelitian terdahulu mengenai klasifikasi dan topik ISPU maka, dipilihlah metode klasifikasi *Naïve Bayes* didukung dengan pertimbangan kecepatan komputasi dan kesesuaian dengan *dataset* penelitian serta dirasa mampu memberikan performa yang baik. Dengan *supervised learning* dimana *dataset* terbagi menjadi 3 kelas kategori BAIK, SEDANG dan TIDAK SEHAT pengklasifikasian *dataset* ISPU menggunakan *Naïve Bayes* diharapkan dapat memberikan tingkat akurasi dan *f1-score* yang tinggi. Adapun fitur yang digunakan dalam penelitian ini diantaranya, *partikulat matter* (PM_{10} dan $PM_{2.5}$), *sulfur dioksida* (SO_2), *karbon moniksida* (CO), *ozon* (O_3) dan *nitrogen dioksida* (NO_2). Diharapakan dengan penelitian ini dapat memberikan kontribusi dalam pengaruh seleksi fitur $PM_{2.5}$ terhadap klasifikasi ISPU yang belum dibahas oleh peneliti terdahulu.

2. METODE PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini disusun secara bertahap dengan tujuan dapat tercapainya *output* yang diharapkan oleh penulis, disusun secara sistematis dan mengikuti kaidah keilmuan. Adapun alur penelitian yang dibentuk untuk menggambarkan garis besar urutan pengerjaan dapat ditemukan pada Gambar 1.



Gambar 1 Alur Penelitian

2. 2 Metode Pengumpulan Data

Berangkat dari permasalahan penting yang sering terjadi, penelitian ini mengambil fokus pada klasifikasi kualitas udara mengingat kebutuhan akan informasi ini sangat krusial. Adapun tahapan yang dijalani dalan pengumpulan data penelitian mencakup :

1. Identifikasi Masalah
   Sebelum menentukan kerangka pengerjaan, tahap paling utama adalah mengidentifikasi permasalahan yang sedang berlangsung atau urgensi dalam penelitian.
2. Studi Literatur
   Studi literatur diperuntukan sebagai dukungan ilmiah dalam penelitian yang bersumber dari jurnal, artikel, buku, laporan maupun situs resmi mengenai informasi yang relevan
3. Pengumpulan Data
   Setelah melalui beberapa tahapan diawal, barulah data *primer* diperoleh dan sebagai dasar kerangka penelitian dengan merumuskan maksud dan tujuan akhir penelitian serta tahapan pengerjaan dan evaluasi hasil.

2. 3 (Indeks Standar Pencemar Udara) ISPU

ISPU atau penamaan secara internasional biasa dikenal dengan *Air Quality Index* (AQI) adalah indeks yang dipakai untuk mengukur kualitas udara di suatu wilayah dengan mempertimbangkan indikator *partikulat matter* (PM_{10} dan $PM_{2.5}$), *sulfur dioksida* (SO_2), *karbon monoksida* (CO), *ozon* (O_3), *nitrogen dioksida* (NO_2) dan *hidrokarbon* (HC). ISPU sendiri adalah penyebutan yang digunakan di Indonesia dan diawasi oleh Kementerian Lingkungan Hidup dan Kehutanan. Pencatatan ISPU dilakukan di titik-titik lokasi yang sudah ditentukan, dalam studi ini, data ISPU yang dianalisis yaitu ISPU wilayah DKI Jakarta yang diakses melalui situs https://satudata.jakarta.go.id/home dengan rentang data mulai tahun 2021 hingga Maret 2024 dengan total data mentah sebanyak 4620 *record dataset* yang diambil dari lima stasiun pemantauan yaitu Bunderan HI, Kelapa Gading, Jagakarsa, Lubang Buaya dan Kebon Jeruk. Ketentuan tentang ISPU ini diatur pada peraturan NOMOR P.14/MENLHK/SETJEN/KUM.1/7/2020 oleh Menteri Lingkungan Hidup dan Kehutanan tentang Indeks Standar Pencemar Udara, adapun kategori angka skala ISPU tersaji dalam Tabel 1.

Tabel 1 Kategori Angka Rentang ISPU

Kategori	Status Warna	Angka Rentang
Baik	Hijau	1 – 50
Sedang	Biru	51 - 100
Tidak Sehat	Kuning	101 – 200
Sangat Tidak Sehat	Merah	201 – 300
Berbahaya	Hitam	≥301

Peraturan terbaru ini secara langsung menggantikan peraturan sebelumnya yaitu peraturan No. 45 Tahun 1997 dimana terdapat penambahan dua parameter yaitu *partikulat matter* ($PM_{2.5}$) dan *hidrokarbon* (HC). Penambahan parameter ini didasarkan pada pentingnya HC dan $PM_{2.5}$ pada dampak kesehatan, sehingga dalam penelitian ini bertujuan untuk melakukan komparasi akurasi dan *f1-score* terkait skenario model pelatihan dengan menambahkan parameter $PM_{2.5}$ dan tanpa menambahkan parameter $PM_{2.5}$ pada *dataset* yang diambil sekaligus membuktikan apakah hasil klasifikasi berbanding lurus dengan urgensi $PM_{2.5}$ terhadap kesehatan. Lingkup pengerjaan hanya berfokus terhadap fitur $PM_{2.5}$ dikarenakan *dataset* yang diperoleh hanya menyertakan penambahan parameter $PM_{2.5}$ dan belum memiliki parameter HC.

2. 4 Data Mining

Data mining dapat dimanfaatkan untuk mengidentifikasi pola, hubungan atau informasi menarik di dalam data, yang selanjutnya bisa dikembangkan menjadi sebuah pengetahuan baru. Tujuan utama *data mining* adalah mengolah data mentah agar dapat diubah menjadi informasi yang bernilai dan dapat mendukung proses pengambilan keputusan. Dalam beberapa tahun terakhir, teknik ini mulai banyak diterapkan di sektor pariwisata karena kemampuannya mengungkap pola-pola tersembunyi dalam himpunan data berukuran besar. Berbeda dengan metode statistik tradisional, data mining juga mampu menganalisis serta mengungkap hubungan non-linier di antara data yang diteliti [8]. Secara umum *data mining* memiliki runtunan proses pengerjaan meliputi, pengumpulan, pembersihan *(data cleaning)*, transformasi, pemilihan *(data selection)*, penerapan algoritma *(pattern discovery)*, evaluasi dan visualisasi hasil. Seluruh runtunan ini tentu dapat disesuaikan dengan kebutuhan tiap peneliti dikarenakan setiap *dataset* memiliki perlakuan masing-masing berdasarkan tujuan dan hasil yang diharapkan.

2. 5 Pre-processing

Pre-processing adalah proses awal dalam mengolah data mentah agar bisa digunakan oleh model atau analisis lanjutan. Data mentah biasanya tidak dapat langsung digunakan dikarenakan mengandung *noise*, data kosong, format yang tidak konsisten dan aspek lain yang dapat mempengaruhi hasil dari performa model, sehingga diperlukan tindakan yang dapat memaksimalkan data tersebut. *Data selection*, *data cleaning* dan *data transformation* merupakan tahapan *pre-processing* yang sering digunakan dalam pengolahan data mentah. Adapun rangkaian ini tergantung pada kebutuhan dan tujuan data masing-masing.

Data selection adalah tahap dimana memilih data atau fitur yang tidak memiliki korelasi, dimana kehadiran fitur tersebut tidak diperlukan atau bahkan dapat mengganggu jika digunakan.

Data cleaning seperti *remove missing value* dan penghapusan *outlier* merupakan salah satu solusi dalam mengurangi *error* pada data, ketidakkonsistenan dan tidak relevan. Dengan mengeliminasi kumpulan data yang memiliki *value* kosong seperti nol, *blank*, atau tidak sesuai format dapat dikecualikan dari pemodelan menggunakan *remove missing value*. Deteksi *outlier* juga sangat penting dalam rangkaian *pre-processing*, kehadirannya dapat menyebabkan ketidakseimbangan dikarenakan data akan cenderung untuk berpusat di range tertentu dan tidak

dapat tersebar secara merata. *Mean-standard deviation* adalah metode pendeteksi *outlier* dengan menganggap data yang berada diluar range sebagai anomali [9]. Metode ini mampu mengeliminasi kumpulan data yang dianggap merugikan dalam pemodelan klasifikasi, *mean-standard deviation* dapat dilihat pada persamaan (1).

$$Mean - stdev = mean \pm 3\ x\ standard\ deviation \quad (1)$$

Data transformation bertujuan untuk mengubah format data mentah menjadi lebih konsisten, relevan dan lebih mudah diproses oleh algoritma *data mining* atau *machine learning*. *Normalization* merupakan salah satu pendekatan yang umum digunakan. Metode *normalization* berguna untuk memperkecil perbedaan skala antar data sekaligus mempercepat proses komputasi. Perbedaan skala yang lebih besar cenderung mendominasi fitur lain dengan skala yang lebih kecil sehingga dapat mennggangu keseimbangan dalam pengukuran jarak [10]. *Normalization* memiliki beberapa variasi, diantaranya adalah *Max Normalization*. Metode ini menormalisasikan setiap nilai dalam suatu fitur dibagi dengan nilai maksimum dari fitur tersebut sehingga semua nilai akan dipetakan ke rentang 0 sampai 1. *Max normalization* dapat dilihat pada persamaan (2).

$$x_i = \frac{x_i}{x_{max}} \quad (2)$$

Metode ini dianggap mudah dan sederhana sehingga dapat memudahkan dalam proses komputasi. Performa ini didukung oleh proses indentifikasi *outlier* sebelumnya, dimana performansi *max normalization* dapat bekerja secara baik apabila kategori *outlier* pada data telah dieliminasi sehingga rentang data dapat terdistribusikan secara merata.

2. 6 Klasifikasi Naïve Bayes

Dalam Teorema Bayes terdapat penerapan konsep tentang probabilitas awal (*prior probability*) serta probabilitas akhir (*likelihood probability*) sebagai dasar informasi dan bukti guna memperbarui pengetahuan terkait suatu hipotesis setelah mempertimbangkan data yang tersedia [11]. Teorema Bayes dinyatakan dalam persamaan (3).

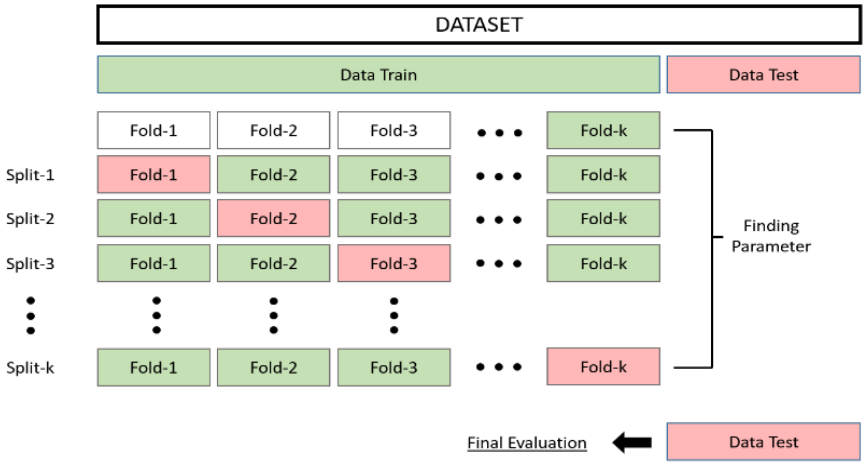
$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (3)$$

Dimana P(H|X) adalah probabilitas *posterior* peluang kejadian H setelah bukti kemunculan X, P(X|H) adalah *likelihood* peluang bukti kemuculan X jika H benar, P(H) adalah *prior probability* peluang awal terjadinya H sebelum melihat bukti, P(X) *marginal* peluang keseluruuhan dari bukti X terlepas dari hipotesis.

Naïve Bayes sendiri merujuk pada asumsi yang sangat sederhana dimana setiap fitur diasumsikan saling bebas satu sama lain dan mengabaikan korelasi antar fitur. Dalam penelitian ini, kata *naïve* merujuk pada contoh jika kenyataan hubungan kandungan antara PM 10 tinggi, kemungkinan NO2 juga tinggi dikarenakan berasal dari sumber polusi yang sama, tetapi *Naïve Bayes* memandang ini adalah fitur independen.

2. 7 K-fold Cross Validation

Pada tahap evaluasi model, *k-fold Cross Validation* memberikan solusi terkait skenario pengujian. Pada proses ini, *dataset* terbagi menjadi data latih dan uji. Penggunaan cara ini seringkali menghasilkan akurasi yang berbeda berdasarkan urutan atau pembagian data yang digunakan selama pelatihan. Dengan *k-fold cross validation,* kedua data tersebut dapat didistribusikan merata berdasarkan jumlah lipatan (*k-fold)* yang dipilih, sehingga setiap *fold* digunakan sebagai set pengujian [12]. Ilustrasi cara kerja *k-fold cross validation* dapat dilihat pada Gambar 2.



Gambar 2 Ilustrasi *K-fold Cross Validation*

Jumlah iterasi yang berlangsung tergantung pada jumlah *fold* yang ditentukan, *fold* ini yang akan menentukan jumlah pembagian seluruh *dataset* menjadi lipatan-lipatan, dimana iterasi awal dimulai dari *fold* pertama sebagai data uji dan lainnya menjadi data latih. Untuk iterasi kedua, *fold* kedua dijadikan data uji dan sisanya sebagai data latih, begitupun seterusnya iterasi ini berulang hingga mencapai *k-fold* yang ditentukan diawal. Dalam banyak studi, Secara umum pemilihan *fold* cenderung menggunakan k=5 dan k=10. Pemilihan ini didasari pada faktor keseimbangan antara bias dan varian serta pertimbangan komputasi, dimana dengan nilai k yang lebih kecil seperti 2 atau 3 memiliki peluang menghasilkan ketidakseimbangan antara bias dan varian dikarenakan data latih yang lebih sedikit sehingga pemodelan menjadi tidak stabil. Untuk nilai k yang lebih besar seperti 10 dapat mengurangi ketidakseimbangan pemodelan dengan banyaknya data latih tetapi lebih ditekankan pada proses komputasi, dikarenakan dengan jumlah lipatan (*fold*) yang banyak maka iterasi yang harus dijalankan semakin banyak dan dapat mempengaruhi efisiensi komputasi. Maka k=5 adalah solusi yang ditawarkan dimana pemodelan memiliki jumlah data latih yang cukup dan dapat merepresentasikan hasil yang tidak berbeda jauh didapat oleh k lebih besar sehingga menghemat biaya dalam komputasi.

2. 8 Evaluasi Hasil

Baik atau tidaknya sebuah pemodelan klasifikasi tergantung pada hasil yang diperoleh. Dalam hal ini, parameter yang digunakan secara umum untuk mengukur performa pemodelan klasifikasi adalah *accuracy* (4) dan *f1-score* (5) [13,14].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$F1 = \left(\frac{2}{precision^{-1} + recall^{-1}}\right) = 2.\left(\frac{precision\,.recall}{precision + recall}\right) \tag{5}$$

Dimana *accuracy* untuk menngukur persentase jumlah prediksi yang benar dari keseluruhan data, *f1-score* merupakan *harmonic mean* dari presisi (6) dan *recall* (7) yang dapat memberikan keseimbangan pada kedua parameter ini.

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

Precision digunakan sebagai alat ukur seberapa akurat prediksi positif yang dibuat oleh model sedangkan *recall* digunakan untuk mengukur seberapa banyak data positif yang berhasil ditemukan oleh model.

2. 9 Weka Software

Waikato Environment for Knowledge Analysis (Weka) adalah tools *machine learning* yang dikembangkan oleh Universitas Waikato, Selandia Baru. Weka menawarkan beragam *tools* visual dan algoritma yang mendukung proses analisis data serta dilengkapi dengan *interface* yang mudah digunakan sehingga bisa menjadi solusi dalam mengolah algoritma untuk *data mining* dan *machine learning.* Sebagai perangkat lunak *open source*, Weka dapat digunakan secara gratis dan kompatibel dengan berbagai sistem operasi [15].

3. HASIL DAN PEMBAHASAN

3. 1 Pre-processing

Dalam studi ini data diambil sebanyak 4620 *record* dimana masih banyak terdapat *missing value*, format yang tidak konsisten dan juga fitur yang tidak perlu digunakan dalam penelitian. Adapun deskripsi fitur *dataset* ISPU mentah yang ditampilkan pada Tabel 2.

Tabel 2 Deskripsi Fitur *Dataset* ISPU

No.	Fitur	Tipe Data
1	periode_data	Integer
2	tanggal	String
3	pm_10	Integer
4	pm_2,5	Integer
5	so2	Integer
6	co	Integer
7	o3	Integer
8	no2	Integer
9	max	Integer
10	critical	String
11	categori	String
12	lokasi_spku	String

Dataset awal berbentuk *file* .csv lalu diubah menjadi .xlxs untuk dikompilasi seluruh rekap data menjadi satu file utuh agar memudahkan dalam pengerjaan menggunakan Microsoft Excel. Dari urutan *dataset* mentah yang diambil, hanya fitur pm_10, pm_2,5, so2, co, o3 dan no2 saja yang dijadikan objek penelitian terhadap fitur categori dan fitur lainnya di eliminasi dikarenakan tidak memiliki korelasi dalam klasifikasi kategori ISPU.

3. 1.1 Data Cleaning

Pembersihan *dataset* menggunakan *remove missing value* dan penghapusan *outlier* yang dapat mengurangi performa saat melakukan pelatihan model. *Remove missing value* disini adalah penghapusan *record* data yang tidak memiliki *value* di salah satu fitur. Total *record* yang dihapus sebanyak 879 data. Contoh *missing value* yang dimaksud dapat dilihat pada Tabel 3.

Tabel 3 Contoh *Dataset Missing Value*

No.	pm_10	pm_2,5	so2	co	o3	no2
1	45		21	13	40	15
2	80		22	44	44	22
3	27		14	9	29	---
…						
879	N/A	37	21	9	45	11

Outlier dapat dikategorikan sebagai anomali pada kumpulan *dataset*, sehingga digunakan metode *mean-standard deviation* untuk mendeteksi batas atas dan bawah *range* untuk setiap fitur, data *range* fitur ini tertera pada Tabel 4.

Tabel 4 *Range* Atas dan Bawah *Outlier*

Fitur	Atas Mean+3 x stdev	Bawah Mean-3 x stdev
pm_10	98,806	7,280
pm_2,5	146,71	5,898
so2	78,819	-0,828
co	31,800	-6,343
o3	84,368	-21,839
no2	57,292	-14,753

Dengan mendeteksi batas atas dan bawah pada masing-masing fitur, diperoleh 155 *record* yang dikategorikan sebagai *outlier* dan dilakukan eliminasi. Sampai pada tahap *data cleaning*, didapat jumlah data yaitu 3586 *record* dengan distribusi tiap-tiap kelasnya 304 *record* untuk kelas BAIK, 2857 *record* untuk kelas SEDANG dan 425 *record* untuk kelas TIDAK SEHAT.

3. 1.2 Data Transformation

Setelah melalu tahap *data cleaning*, penerapan *max normalization* untuk mengubah seluruh skala data menjadi rentang 0 sampai 1 dengan membagikan seluruh data fitur berdasarkan *max value* masing-masing fitur. Hasil penerapan *max normalization* dapat diamati pada Tabel 3.

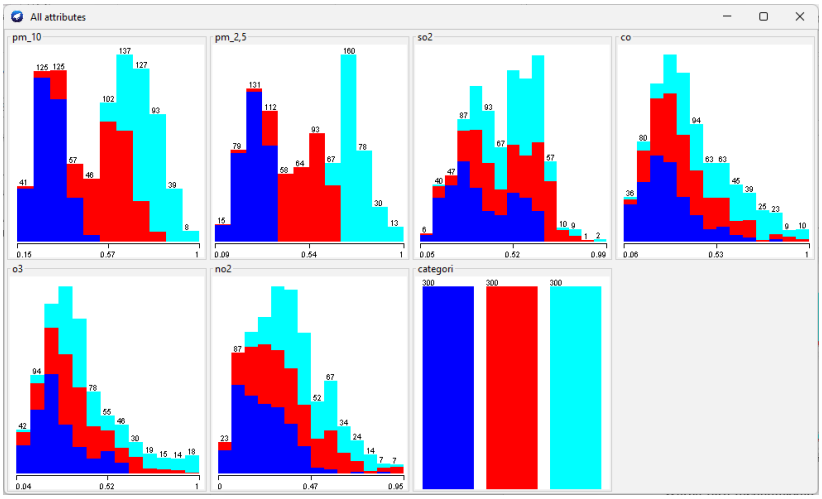
Tabel 3 Implementasi *Max Normalization*

No.	pm_10	pm_2,5	so2	co	o3	no2
1	0,2234	0,2260	0,6494	0,1290	0,1905	0,0526
2	0,2766	0,2534	0,2987	0,4194	0,1071	0,2982
3	0,2979	0,1096	0,2208	0,2903	0,5714	0,0702
…						
3586	0,7128	0,8151	0,5455	0,4194	0,3333	0,4912

Dari total 3586 *record* data bersih, hanya diambil 900 *record* untuk mewakili tiap kategori yaitu 300 kategori BAIK, 300 kategori SEDANG dan 300 kategori TIDAK SEHAT. Pemilihan *dataset* yang seimbang bertujuan untuk meningkatkan akurasi klasifikasi disemua kelas serta menghindari bias pada proses pemodelan. Selanjutnya 900 *dataset* dikonversi menjadi *file* .csv untuk memudahkan dalam pengoperasian menggunakan Weka.

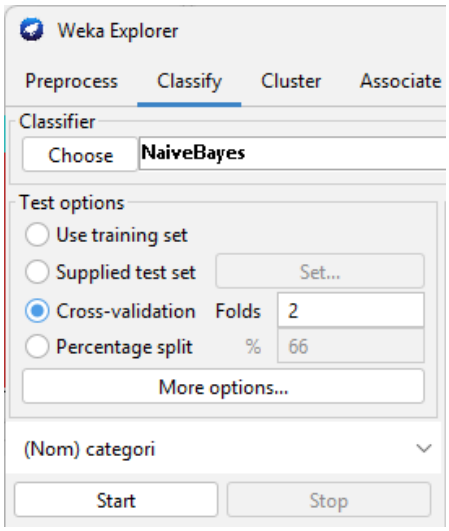
3. 2 Pemodelan Menggunakan Weka

Adapun skenario pemodelan yang dijalankan adalah dengan melibatkan fitur pm_2,5 dan tanpa fitur pm_2,5. Hal ini bertujuan untuk mengetahui komparasi akurasi maupun *f1-score* terhadap hasil yang didapat. Selanjutnya *dataset* dibuka menggunakan Weka, Gambar 3 menampilkan histogram yang merepresentasikan distribusi data.



Gambar 3 Histogram Distribusi *Dataset*

Warna biru menunjukkan kategori BAIK, merah menunjukkan kategori SEDANG dan *cyan* (biru muda) menunjukkan kategori TIDAK SEHAT. Distribusi nilai pada setiap fitur menempati *range* 0 sampai 1. Pemilihan algoritma Naïve Bayes dan *cross validation* k = 1 sampai dengan 5 untuk mengetahui akurasi mana yang terbaik. Tampilan pemilihan ini tertera pada Gambar 4.



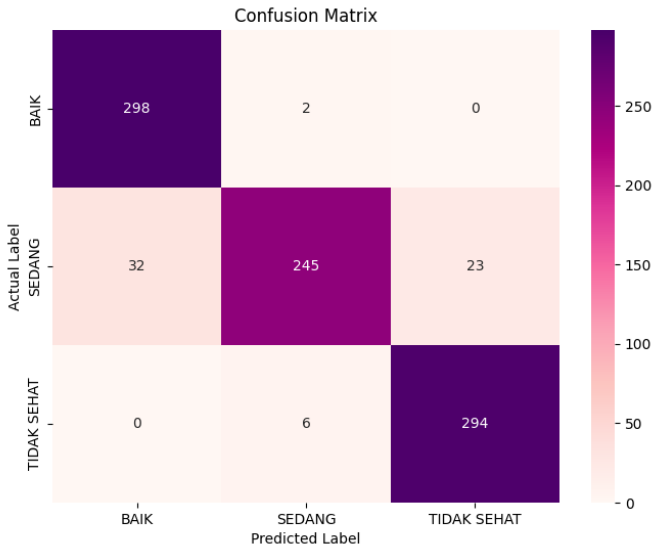
Gambar 4 Tampilan Pemilihan Model

Langkah ini diulangi tiap penambahan jumlah *k* hingga *k* = 5 dan diterapkan pada skenario selanjutnya dimana tanpa melibatkan fitur pm_2,5. Penentuan baik tidaknya model berdasarkan hasil akurasi tertinggi dan *f1-score* yang didapat setelah menjalankan seluruh runtutan skenario. Dari urutan pemodelan ini, maka didapatlah *summary* dari rekap keseluruhan yang dapat dilihat pada Tabel 4.

Tabel 4 *Summary* Hasil Model

Skenario	Fitur	*K-fold Cross Validation*	Akurasi (%)	*F1-score* (%)
Eksperimen 1	pm_10, pm_2,5, so2, co, o3, no2	2	92,1111	91,9
Eksperimen 2	pm_10, pm_2,5, so2, co, o3, no2	3	92,3333	92,2
Eksperimen 3	pm_10, pm_2,5, so2, co, o3, no2	4	92,6667	92,5
Eksperimen 4	pm_10, pm_2,5, so2, co, o3, no2	5	93	92,8
Eksperimen 5	pm_10, so2, co, o3, no2	2	82,8889	82,7
Eksperimen 6	pm_10, so2, co, o3, no2	3	83	82,8
Eksperimen 7	pm_10, so2, co, o3, no2	4	83,7778	82,6
Eksperimen 8	pm_10, so2, co, o3, no2	5	82,889	82,6

Dari Tabel 4 diatas menunjukkan akurasi dan *f1-score* tertinggi yaitu pada skenario eksperimen 4 dengan penambahan fitur pm_2,5 dan k = 5, dimana diperoleh akurasi sebesar 93% dan *f1-score* sebesar 92,8%. Visualisasi *confusion matrix* dari skenario ini ditampilkan pada Gambar 5.



Gambar 5 *Confusion Matrix*

Adapun ringkasan komparasi hasil penelitian ini terhadap studi terdahulu dapat dilihat pada Tabel 5.

Tabel 5 Komparasi Hasil Penelitian

Aspek	Penelitian Ini	Avira, dkk. (2024)	Fitri, dkk. (2023)	Devi, dkk. (2022)	Syekh, dkk. (2021)
Metode Klasifikasi	*Naïve Bayes*	KNN dan *Naïve Bayes*	*Naïve Bayes*	*Naïve Bayes*	SVM, *Decision Tree*, KNN, *Naïve Bayes* dan *Neural Network*
Dataset	900 *record* ISPU DKI Jakarta 2021-Maret 2024	365 *record* ISPU Kota Tanggerang Selatan 2022	1.096 *record* ISPU Tanggerang Selatan	1.206 *record* ISPU DKI Jakarta 2020	6.536 *record* ISPU DKI Jakarta 2017-Juni 2020
Fitur	PM_{10}, $PM_{2.5}$, SO_2, CO, O_3 dan NO_2	PM_{10}, $PM_{2.5}$, SO_2, CO, O_3 dan NO_2	PM_{10}, $PM_{2.5}$, SO_2, CO, O_3 dan NO_2	PM_{10}, SO_2, CO, O_3 dan NO_2	PM_{10}, SO_2, CO, O_3 dan NO_2

Teknik Validasi	*K-fold Cross Validation* K=5	*Split Data* 90:10	Tidak Didefinisikan	*K-fold Cross Validation* K=10	*K-fold Cross Validation* K=10
Akurasi	93% (*Naïve Bayes*)	94,44% (KNN), 86,11% (*Naïve Bayes*)	79,38% (*Naïve Bayes*)	91% (*Naïve Bayes*)	96,60% (SVM), 99,80% (*Decision Tree*), 97,55% (KNN), 91,89% (*Naïve Bayes*) 98,01% (*Neural Network*)
Keterangan	Hasil ini adalah skenario terbaik pada Tabel 4 dengan akurasi tertinggi dan penambahan fitur $PM_{2.5}$	Penggunaan jumlah fitur yang sama tetapi perbedaan teknik validasi yang digunakan	Pengguunaan jumlah fitur yang sama tetapi tidak didefinisikan dalam teknik validasi yang digunakan	Penggunaann jumlah fitur yang berbeda dan teknik validasi yang sama tetapi dengan K=10	Penggunaann jumlah fitur yang berbeda dan teknik validasi yang sama tetapi dengan K=10

4. KESIMPULAN

Paket fitur yang sesuai dipilih dengan akurasi terbaik adalah dengan menggunakan fitur $PM_{2.5}$ dan *K-fold* = 5 yaitu akurasi sebesar 93% dan *f1-score* sebesar 92,8%. Skenario serupa tanpa menggunakan fitur $PM_{2,5}$ didapat akurasi sebesar 82,8889% dan *f1-score* sebesar 82,6%. Dengan selisi kurang lebih 10% pada hasil, bisa dikatakan bahwa temuan ini berbanding lurus dengan penambahan fitur $PM_{2,5}$ pada peraturan ISPU terbaru, tidak hanya berpengaruh pada kesehatan tetapi juga dapat meningkatkan nilai akurasi dan *f1-score* dalam metode klasifikasi.

5. SARAN

Penelitian selanjutnya diharapkan dapat menambahkan fitur *Hidrokarbon* (HC) pada *dataset* pelatihan serta penambahan kategori kelas lebih dari tiga yaitu SANGAT TIDAK SEHAT dan BERBAHAYA yang merujuk pada peraturan ISPU terakhir. Dapat menguji menggunakan *dataset* yang tidak imbang dari distribusi masing-masing kelasnya dan pemilihan model serta algoritma klasifikasi berbeda yang dapat meninggkatkan akurasi maupun parameter ukur lainnya.

DAFTAR PUSTAKA

[1] A. Budiandita, N. Iman, F. M. Hana and C. B. Hakim, "Komparasi Algoritma K-Nearest Neighbor dan Naive Bayes pada Klasifikasi Tingkat Kualitas Udara Kota Tanggerang Selatan," *Jurnal Informatika dan Rekayasa Perangkat Lunak*, vol. 6, no. 1, pp. 320-327, 2024. [Online]. Available: https://doi.org/10.36499/jinrpl.v6i1.10956. [Accessed: 01-Mar-2025].

[2] "Air pollution," Institute For Health Metrics and Evaluation. [Online]. Available: https://www.healthdata.org/research-analysis/health-risks-issues/air-pollution. [Accessed: 01-Mar-2025].

[3] A. A. Anandari, A. F. Wadjdi and G. Harsono, "Dampak Polusi Udara terhadap Kesehatan dan Kesiapan Pertahanan Negara di Provinsi DKI Jakarta," *Journal on Education*, vol. 6, no. 2, pp. 10868-10884, 2024. [Online]. Available: https://doi.org/10.31004/joe.v6i2.4880. [Accessed: 01-Mar-2025].

[4] E. H. A. Rady and A. S. Anwar, "Prediction of kidney disease stages using data mining algorithms," *Informatics in Medicine Unlocked*, vol. 15, 2019, Art. no. 100178. [Online]. Available: https://doi.org/10.1016/j.imu.2019.100178. [Accessed: 01-Mar-2025].

[5] F. Widiawati, R. Kurniawan and T. Suprapti, "Klasifikasi Data Tingkat Kualitas Udara di Tanggerang Selatan Menggunakan Algoritma Naive Bayes," *Jurnal Mahasiswa Teknik Informatika*, vol. 7, no. 6, pp. 3739-3745, 2023. [Online]. Available: https://doi.org/10.36040/jati.v7i6.8261. [Accessed: 01-Mar-2025].

[6] D. D. Purwanto and E. S. Honggara, "Klasifikasi Kategori Hasil Perhitungan Indeks

Standar Pencemaran Udara Dengan Gausian Naïve Bayes (Studi Kasus: Ispu Dki Jakarta 2020)," INSYST, vol. 4, no. 2, pp. 102-108, 2022. [Online]. Available: https://doi.org/10.52985/insyst.v4i2.259. [Accessed: 01-Mar-2025].

[7] S. S. A. Umri, M. S. Firdaus and A. Primajaya, "Analisis dan Komparasi Algoritma Klasifikasi dalam Indeks Pencemaran Udara di DKI Jakarta," *Jurnal Informatika dan Komputer*, vol. 4, no. 2, pp. 98-104, 2019. [Online]. Available: https://doi.org/10.33387/jiko.v4i2.2871. [Accessed: 01-Mar-2025].

[8] Sunardi, A. Fadlil and N. M. P. Kusuma, "Implementasi Data Mining dengan Algoritma Naïve Bayes untuk Profiling Korban Penipuan Online di Indonesia," *Jurnal Media Informatika Budidarma*, vol. 6, no. 3, pp. 1562-1572, 2022. [Online]. Available: https://doi.org/10.30865/mib.v6i3.3999. [Accessed: 01-Mar-2025].

[9] H. A. Ahmed, P. J. M. Ali, A. K. Faeq and S. M. Abdullah, "An Investigation on Disparity Responds of Machine Learning Algorithms to Data Normalization Method," *ARO-THE SCIENTIFIC JOURNAL OF KOYA UNIVERSITY*, vol. 10, no. 2, pp. 29-37, 2022. [Online]. Available: https://doi.org/10.14500/aro.10970. [Accessed: 01-Mar-2025].

[10] W. Li and Z. Liu, "A method of SVM with Normalization in Intrusion Detection," *Procedia Environmental Sciences*, vol. 11, pp. 256-262, 2011. [Online]. Available: https://doi.org/10.1016/j.proenv.2011.12.040. [Accessed: 01-Mar-2025].

[11] I. Riadi, R. Umar and R. Anggara, "Prediksi Kelulusan Tepat Waktu Berdasarkan Riwayat Akademik Menggunakan Metode Naïve Bayes," *Decode: Jurnal Pendidikan Teknologi Informatika*, vol. 4, no. 1, pp. 191-203, 2024. [Online]. Available: https://doi.org/10.51454/decode.v4i1.308. [Accessed: 01-Mar-2025].

[12] A. Peryanto, A. Yudhana and R. Umar, "Klasifikasi Citra Menggunakan Convolutional Neural Network dan K Fold Cross Validation," *JAIC*, vol. 4, no. 1, pp. 45-51, 2020. [Online]. Available: https://doi.org/10.30871/jaic.v4i1.2017 . [Accessed: 01-Mar-2025].

[13] I. Riadi, A. Fadlil and B. A. Prabowo, "MAC Address Classification in Privacy Issue Using Gaussian Naïve Bayes," *JUITA: Jurnal Informatika*, vol. 12, no. 2, pp. 235-242, 2024. [Online]. Available: https://doi.org/10.30595/juita.v12i2.22571. [Accessed: 01-Mar-2025].

[14] M. Grandini, E. Bagli and G. Visani, "Metrics for Multi-Class Classification: an Overview," *arXiv*, 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2008.05756. [Accessed: 01-Mar-2025].

[15] S. Singhal and M. Jena, "A Study on WEKA Tool for Data Preprocessing, Classification and Clustering," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 2, no. 6, pp. 250-253, 2013. [Online]. Available: https://www.ijitee.org/portfolio-item/f0843052613/. [Accessed: 01-Mar-2025].