Klasifikasi Kategori Hasil Perhitungan Indeks Standar Pencemaran Udara dengan Gausian Naïve Bayes (Studi Kasus: ISPU DKI Jakarta 2020)

Devi Dwi Purwanto[1], **Eric Sugiharto Honggara**[1]
[1]Departemen Sistem Informasi, Fakultas Sains dan Teknologi, Institut Sains dan Teknologi Terpadu Surabaya, Surabaya, Indonesia

Corresponding author: Devi Dwi Purwanto (e-mail: devi@stts.edu).

ABSTRACT Air pollution is a problem that endangers humans, especially for the respiratory system. At present, air pollution always occurs due to several reasons such as vehicle, power plants and others. One of the places where air pollution occurs most is in big cities where many people gather. One of the places of concern is the station which is in the special area of the capital city of Jakarta. A station is a place where many people gather and wait to travel. Therefore, DKI Jakarta's environmental service open their data on air pollution that occurs at stations so that it can be used by the public for processing. The data will be preprocessed first by handling the missing values, then through data normalization and also used one hot encoding to uniform the data. The data will then be classified using the Gausian Naïve Bayes algorithm. After obtaining the results of the classification, it can be concluded that the max and critical attributes in the dataset have no effect on the clasification results for the ISPU category. The attributes of the data that influence the clasification of the ISPU category are PM10, SO_2, CO, O_3, and NO_2. By using 5 attributes and gausian naïve Bayes, the system can provide clasifications with an accuracy of 91.16% and an error rate of 8.84%. While the value of Weighted Average Recall 93,36%, Weighted Average Precision 93,92%, and Weighted Average F1-Score sebesar 93,68%.

KEYWORDS Gausian Naïve Bayes, Air Pollution, One Hot Encoding, Cross validation k-folds

ABSTRAK Pencemaran udara adalah masalah yang membahayakan manusia terutama untuk sistem pernafasan. Saat ini pencemaran udara selalu terjadi akibat beberapa hal seperti asap kendaraan, pembangkit listrik dan lainnya. Salah satu tempat di mana pencemaran udara terjadi adalah di kota besar di mana banyak orang berkumpul. Salah satu tempat yang menjadi perhatian adalah stasiun yang berada di daerah khusus ibukota jakarta. Stasiun adalah tempat di mana banyak orang berkumpul dan menunggu untuk melakukan perjalanan. Maka dari itu dinas lingkungan hidup DKI Jakarta membuka data pencemaran udara yang terjadi di stasiun agar dapat digunakan oleh masyarakat untuk diolah. Data tersebut akan dilakukan preprocessing yaitu penanganan missing value, normalisasi data, dan menggunakan one hot encoding. Data tersebut kemudian akan diklasifikasi dengan menggunakan algoritma Gausian Naïve Bayes. Setelah memperoleh hasil dari klasifikasi dapat disimpulkan bahwa atribut max dan critical yang berada dalam dataset tidak memiliki pengaruh terhadap hasil klasifikasi kategori ISPU. Atribut-atribut dari data yang berpengaruh terhadap klasifikasi kategori ISPU adalah PM10, SO_2, CO, O_3, dan NO_2. Dengan menggunakan 5 atribut dan gausian naïve bayes, sistem dapat memberikan klasifikasi dengan akurasi sebesar 91,16% dan memiliki error rate sebesar 8,84%. Sedangkan nilai Weighted Average Recall 93,36%, Weighted Average Precision 93,92% , dan Weighted Average F1-Score sebesar 93,68%.

KATA KUNCI Gausian Naïve Bayes, Polusi Udara, One Hot Encoding, Cross validation k-folds

I. PENDAHULUAN

Pencemaran udara adalah kehadiran sebuah subtansi fisik, kimia atau biologi dalam udara yang dapat menganggu kesehatan manusia. Secara spesifik, pencemaran ini dapat membuat manusia memiliki gangguan pernapasan, seperti asma, ISPA dan bahkan kanker paru-paru. Dinas lingkungan hidup Daerah Khusus Ibukota(DKI) Jakarta memiliki data yang telah dikumpulkan selama beberapa tahun dan ingin menggunakan data tersebut untuk melakukan klasifikasi kategori pencemaran udara yang terjadi pada ke-5 stasiun yang berada di kawasannya. Pencemaran udara yang terjadi di stasiun dikarenakan banyak faktor. Beberapa diantaranya adalah asap kendaraan bermotor [1], asap yang dihasilkan industri, pembangkit listrik dan rumah tangga, termasuk pencemaran udara karena debu yang berterbangan akibat aktifitas manusia yang berada di daerah sekitar stasiun. Maka pencemaran udara adalah sebuah masalah besar namun tidak kasat mata yang dapat membahayakan kesehatan, bahkan jiwa manusia dan hal ini tidak disadari oleh masyarakat yang tinggal dan menjalankan kehidupan mereka di kota besar [2]. Adapun pencemaran yang terjadi di dalam stasiun yang hari ini dapat terdeteksi adalah pecemaran PM10, SO_2, CO, O_3 dan NO_2. PM10, sebuah partikel di udara dengan ukuran lebih kecil dari 10 mikron dan dapat menyebabkan resiko karsinogenik [3]. SO_2 yang merupakan sulfur dioksida, sebuah senyawa kimia yang beracun bagi manusia dan biasanya berasa dari gunung atau hasil pemprosesan industri. SO_2 juga dapat menyebakan penurunan fungsi paru [4]. CO yang dikenal sebagai karbon monoksida, berupa gas. Gas CO ini tidak berwarna, tidak berbau, tidak berasa dan tidak memberikan rangsangan, oleh karena itu gas ini adalah gas yang susah dideteksi manusia dan berbahaya bagi kesehatan manusia [5]. O_3 adalah ozone, gas ini berbahaya bagi manusia karena dapat menyebabkan iritasi paru-paru dan tenggorokan, batuk dan memperburuk gejala asma [6]. Berikutnya adalah NO_2 atau Nitrogen Dioksida. Gas ini juga berbahaya bagi manusia karena dapat menyebabkan gangguan pernapasan seperti batuk, kemudian mata merah dan perih yang terjadi pada mata [7] [4]. Data yang dimiliki oleh dinas lingkungan hidup DKI Jakarta juga memiliki parameter max yang digunakan untuk melihat nilai ukur paling tinggi dalam waktu yang sama dan parameter critical yang digunakan untuk melihat mana parameter yang pengukurannya paling tinggi. Data ini belum dimanfaatkan sepenuhnya oleh dinas lingkungan hidup, oleh karena itu data tersebut dibuka kepada masyarakat agar bisa dimanfaatkan. Sehingga siapapun dapat melakukan pengolahan pada data yang ada agar dapat membantu Dinas Lingkungan Hidup DKI Jakarta untuk memprediksi pencemaran udara yang terjadi pada kelima statsiunnya.

II. LANDASAN TEORI

Berikut adalah landasan teori yang digunakan dalam melakukan pemprosesan data hingga memperoleh hasil sesuai tujuan.

A. ISPU

ISPU merupakan angka tanpa satuan yang digunakan untuk menggambarkan kondisi mutu udara ambien pada suatu lokasi yang didasarkan pada dampak terhadap kesehatan manusia dan makhluk hidup lainnya [8]. Menurut Peraturan Menteri Lingkungan Hidup dan Kehutanan no 45 tahun 1997 tentang Indeks Standar Pencemaran Udara ditetapkan bahwa ada 5 parameter yang mempengaruhi perhitungan ISPU yaitu PM10, NO2, SO2, CO, dan O3. Dari 5 parameter tersebut didapatkan nilai konversi parameter ISPU dan dikategorikan menjadi 5 kategori yaitu baik, sedang, tidak sehat, sangat tidak sehat, dan berbahaya. Tabel Kategori tersebut dapat dilihat pada tabel 1.

TABEL I
KATEGORI INDEKS STANDAR PENCEMARAN UDARA (ISPU)

RENTANG	KATEGORI	PENJELASAN
1-50	Baik	Tingkat mutu udara yang sangat baik, tidak memberikan efek negative terhadap manusia, hewan, dan tumbuhan.
51-100	Sedang	Tingkat mutu udara masih dapat diterima pada kesehatan manusia, hewan, dan tumbuhan.
101-200	Tidak Sehat	Tingkat mutu udara yang bersifat merugikan pada kesehatan manusia, hewan, dan tumbuhan.
201-300	Sangat Tidak Sehat	Tingkat mutu udara yang dapat meningkatkan resiko kesehatan pada sejumlah segmen populasi yang terpapar.
301+	Berbahaya	Tingkat mutu udara yang dapat merugikan kesehatan serius pada populasi dan perlu penanganan cepat.

Data pengukuran tersebut dilakukan selama 24 jam secara terus-menerus untuk mendapatkan 5 parameter tersebut dan nantinya ISPU yang diambil adalah perhitungan ISPU berdasarkan ISPU batas atas, batas bawah, ambien batas atas, ambien batas bawah, dan konsentrasi ambien hasil pengukuran. Persamaan perhitungan ISPU tersebut adalah sebagai berikut: [13]

$$I = \frac{I_a - I_b}{X_a - X_b}(X_x - X_b) + I_b \quad (1)$$

Dimana:
I = ISPU terhitung
I_a = ISPU batas atas
I_b = ISPU batas bawah
X_a = konsentrasi ambien batas atas (µg/m^3)
X_b = konsentrasi ambien batas bawah (µg/m^3)
X_x = konsentrasi ambien nyata hasil pengukuran (µg/m^3)

B. Data Preprocessing

Data preprocessing adalah proses mengubah data mentah ke dalam bentuk yang mudah dipahami dan siap untuk digunakan untuk proses berikutnya, karena data yang berkualitas akan berdampak pada keberhasilan terhadap proyek yang melibatkan analisa data. Proses dari Data

Preprocessing meliputi data cleaning, data integration, data transformation, dan data reduction. Proses ini juga disebut proses Extract, Transform, Load atau ETL yang juga dilakukan pada Data Warehouse. Fungsi utama dari ETL ini adalah mengurangi waktu reposn dan meningkatkan performa [9]. Tujuan akhir dari datawarehouse adalah menyiapkan sebuah data yang siap diproses.

Data cleaning adalah proses untuk membersihkan data yang missing value, menghaluskan data yang tidak pada umumnya, dan menyelesaikan data yang tidak konsisten yang terdapat di dalam dataset. Untuk penanganan missing value dapat dilakukan dengan beberapa cara yaitu menghilangkan data tersebut, mengganti dengan variable tertentu, mengisi dengan rata-rata atribut tersebut, mengisi dengan rata-rata atribut pada kelas yang sama, ataupun melakukan regresi untuk mengganti isi data yang kosong.

Data integration adalah tahap yang menggabungkan data dari berbagai sumber menjadi satu kesatuan data, dimana perlu diperhatikan bahwa saat digabungkan data harus memiliki format yang sama, melakukan deteksi nilai data yang konflik, dan menghapus atribut yang tidak dibutuhkan pada proses selanjutnya. Penghapusan atribut tersebut didasarkan pada tujuan melakukan mining.

Data transformation adalah proses untuk melakukan normalisasi dan generalisasi untuk memastikan bahwa tidak ada data yang berada di luar range dan menyeragamkan range data agar tidak terjadi ketimpangan. Hal ini dikarenakan data yang memiliki range yang timpang akan menyebabkan impact yang berbeda, dimana semakin besar valuenya akan semakin besar impact atribut tersebut.

Data reduction adalah proses pengurangan jumlah data, pengurangan dimensi, ataupun melakukan kompresi data sehingga data yang akan digunakan nantinya tidak menyebabkan akurasi menjadi rendah.

C. *One Hot Encoding*

One hot encoding adalah sebuah proses di mana variabel-variabel yang ada diubah menjadi sebuah bentuk yang bisa digunakan algoritma *machine learning* untuk melakukan klasifikasi dengan lebih baik. Bentuk dari hasil encoding ini adalah 1 dan 0 [10]. One hot encoding digunakan untuk data yang tidak memiliki relasi satu sama lainnya dan keunggulan utamanya adalah kemudahan dalam melakukan skala data. Untuk data dari dinas lingkungan hidup, data yang diubah dengan one hot encoding adalah parameter stasiun yang isinya hanya terdiri dari salah satu dari 5 stasiun di DKI Jakarta yakni "Stasiun_DKI1 (Bundaran HI)", "Stasiun_DKI2 (Kelapa Gading)", "Stasiun_DKI3 (Jagakarsa)", "Stasiun_DKI4 (Lubang Buaya)", dan "Stasiun_DKI5 (Kebon Jeruk)". Data tersebut akan diubah menjadi bentuk biner 1 untuk kategori yang muncul dan nilai 0 untuk kategori lain.

S_DKI1	S_DKI2	S_DKI3	S_DKI4	S_DKI5
1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	0	1

GAMBAR 1. Contoh Perubahan One Hot Encoding

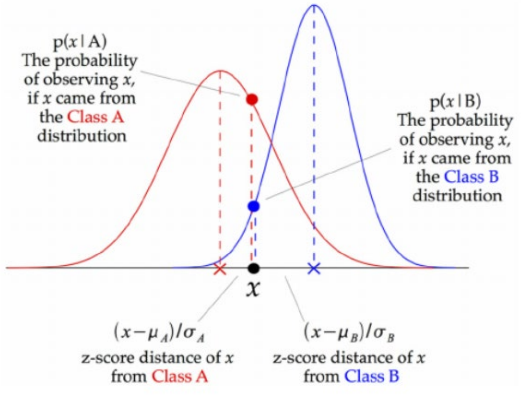
D. *Gausian Naïve Bayes*

Naïve Bayes adalah algoritma sederhana yang memanfaatkan aturan yang ada dalam bayes. Algoritma ini digunakan dengan asumsi bahwa atribut-atribut yang ada dalam data yang dimiliki tidak bergantung satu dengan yang lain. Algoritma Naïve Bayes memiliki peforma yang sangat baik untuk melakukan klasifikasi terutama dalam akurasi dari klasifikasi yang diberikan **[11] [12]**.

Gausian Naïve Bayes merupakan salah satu dari varian Naïve Bayes yang mengikuti distribusi normal gausian dan digunakan untuk data kontinu dengan rumus:

$$\boldsymbol{P(y) = \frac{1}{\sqrt{2\pi\sigma^2 y}} exp^{\left(-\frac{(x_i-\mu_y)^2}{2\sigma^2 y}\right)}} \quad (2)$$

Pendekatan untuk membuat model sederhana adalah dengan mengasumsikan bahwa data dideskripsikan oleh distribusi Gaussian tanpa co-variance antar dimensi. Dengan menggunakan model ini dapat dicocokkan hanya dengan menemukan rata-rata dan standar deviasi dari titik-titik dalam setiap label.



GAMBAR 2. Ilustrasi Klasifikasi dengan Gausian Naïve Bayes

Ilustrasi pada gambar 2 menunjukkan cara kerja pengklasifikasi Gaussian Naive Bayes (GNB) di mana pada setiap titik data dihitung jarak z-score antara titik tersebut dengan rata-rata kelas, yaitu jarak dari rata-rata kelas dibagi dengan standar deviasi kelas tersebut. Sehingga dapat dikatakan Gaussian Naive Bayes memiliki pendekatan yang sedikit berbeda dan dapat digunakan secara efisien pada data kontinu. Langkah selanjutnya untuk menentukan klasifikasi kelas akan dilakukan perhitungan probabilitas untuk masing-

masing atribut dengan memperhatikan teorema bayes, dimana untuk mendapatkan pengetahuan baru digunakan peluang statistika dari masing-masing atribut. Jika diasumsikan suatu kelas dengan C, maka dapat diketahui Probabilitas Class Prior P(C_i) dengan C_i adalah label kelas ke-i dan i = 1, 2, ..., m, sehingga menjadi:

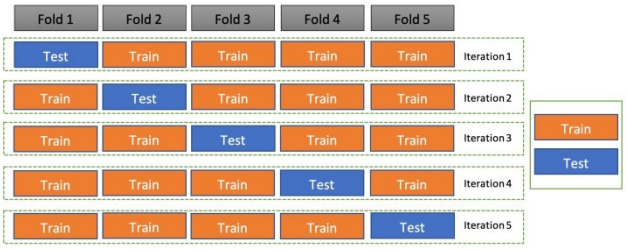
$$P(C_i) = \frac{N_c}{N} \quad (3)$$

Setelah mendapatkan probabilitas dari masing-masing kelas tersebut dan probabilitas class prior akan dilakukan perhitungan probabilitas akhir antara data testing dengan kelas target, hasil probabilitas yang akan diambil adalah probabilitas yang terbesar dan value dari kelas tersebutlah yang akan digunakan untuk klasifikasi kelas targetnya.

$$Pred = argmax_{C_i \in C} \frac{P(C_i)x\ P(C_i)}{P(x_1, x_{2,\ldots,} x_n)} \quad (4)$$

E. Cross Validation

Cross validation merupakan salah satu metode evaluasi untuk membandingkan antara data asli dengan data hasil klasifikasi dengan membagi menjadi 2 segmen yaitu data training dan data testing dan dilakukan secara acak. Metode ini merupakan salah satu metode yang biasanya digunakan jika terdapat jumlah value kelas target dari dataset yang inbalance. Metode ini dikenal dengan sebutan k-fold cross validation, dimana k adalah banyak iterasi untuk membagi data training dan testing. Ilustrasi k-fold cross validation dapat dilihat pada gambar 3.



GAMBAR 3. K-fold cross validation

III. Hasil

Pada bagian ini akan dijelaskan mengenakan tahapan-tahapan yang akan dilakukan untuk dapat melakukan klasifikasi pada dataset yang digunakan sebagai studi kasus. Tahapan tersebut adalah pengumpulan data, preprocessing, Data Mining, dan Evaluasi.

A. *Pengumpulan Data*

Dataset yang digunakan pada penelitian ini adalah open dataset dari Dinas Lingkungan Hidup Provinsi DKI Jakarta pada tahun 2020, dan diterbitkan dengan frekuensi penerbitan 1 bulan sekali yang diambil dari web https://data.jakarta.go.id/dataset/indeks-standar-pencemaran-udara-ispu-tahun-2020. Dataset ini mengambil sampel 5 stasiun besar yang ada di DKI Jakarta dengan atribut yang dicatat adalah tanggal pengambilan data, nama stasiun pengambilan data, partikulat salah satu parameter yang diukur(pm10), Sulfida dalam bentuk SO_2, carbon monoksida, ozon (O_3), nitrogen (NO_2), Nilai ukur paling tinggi dari seluruh parameter yang diukur dalam waktu yang sama(Max), Parameter yang hasil pengukurannya paling tinggi (critical), dan kategori. Dataset ini nantinya akan dibedakan menjadi data training dan data testing, dimana data training diambil dari ISPU bulan Januari hingga Agustus 2020 namun data pada bulan ke-3 tidak digunakan. Hal ini dikarenakan pada bulan ke-3 data yang diperoleh tidak ada nama stasiun tempat data tersebut dikumpulkan, sehingga data tidak dapat digunakan untuk melakukan klasifikasi pencemaran udara pada stasiun tersebut. Pada dataset tersebut hanya menggunakan 3 value dari 5 value kategori ISPU yang ditetapkan oleh Dinas Lingkungan Hidup dan Kehutanan. Tiga value tersebut yaitu baik, sedang, dan tidak sehat.

B. *Preprocessing*

Data preprocessing merupakan langkah awal yang perlu dilakukan dalam proses data mining agar data yang akan digunakan nantinya tidak bernoise, lengkap, dan formatnya terstruktur. Atribut yang akan digunakan nantinya adalah stasiun, PM10, NO2, SO2, CO, dan O3. Data preprocessing yang akan dilakukan pada bagian ini meliputi penanganan missing value pada semua atribut yang digunakan, normalization, dan one hot encoding.

GAMBAR 4. Preprocessing Raw Data

Sebelum data diproses maka data akan dilakukan pemeriksaan untuk meningkatkan akurasi. Dataset yang digunakan memiliki sebuah kekurangan, yaitu dataset tersebut terdapat missing value pada gambar 5 yang ditandai dengan '---'.

```
2020-08-13,DKI4 (Lubang Buaya),81,24,16,37,4,81,PM10,SEDANG
2020-08-14,DKI4 (Lubang Buaya),---,15,14,36,8,36,O3,BAIK
2020-08-15,DKI4 (Lubang Buaya),---,23,21,35,9,35,O3,BAIK
2020-08-16,DKI4 (Lubang Buaya),---,24,11,54,8,54,O3,SEDANG
2020-08-17,DKI4 (Lubang Buaya),---,22,17,50,7,50,O3,BAIK
```

GAMBAR 5. Bentuk Data Pencemaran Udara Dalam Stasiun

Penanganan missing value untuk 1 atribut yang kosong ini akan diganti dengan menghitung nilai rata-rata dari value yang ada dalam dataset untuk menggantikan missing value tersebut. Selain missing value, masalah kedua yang ada dalam dataset adalah adanya hari di mana data tersebut tidak ada. Untuk penanganan data yang tidak ada sama sekali dalam satu hari ini akan dilakukan penghapusan terhadap data tersebut. Sehingga dataset dapat diproses tanpa ada kendala.

```
2020-08-01,DKI4 (Lubang Buaya),---,---,---,---,---,0,,TIDAK ADA DATA
```

GAMBAR 6. Bentuk Data dengan Missing Value

Dapat dilihat pada gambar 6, karena seluruh data pada hari tersebut berupa garis maka data tersebut akan dihapus. Total data training yang digunakan ada sebanyak 1215 data setelah dilakukan pembersihan dan menghapus data yang "Tidak ada data" maka jumlah total akhir data sebanyak 1206 data, dimana detail perbandingan data training dapat dilihat pada gambar 7.

stasiun	categori	
DKI1 (Bunderan HI)	BAIK	81
	SEDANG	159
	TIDAK SEHAT	3
DKI2 (Kelapa Gading)	BAIK	45
	SEDANG	177
	TIDAK SEHAT	19
DKI3 (Jagakarsa)	BAIK	36
	SEDANG	189
	TIDAK SEHAT	18
DKI4 (Lubang Buaya)	BAIK	28
	SEDANG	198
	TIDAK SEHAT	10
DKI5 (Kebon Jeruk) Jakarta Barat	BAIK	21
	SEDANG	181
	TIDAK SEHAT	41

GAMBAR 7. Detail Perbandingan Data Training

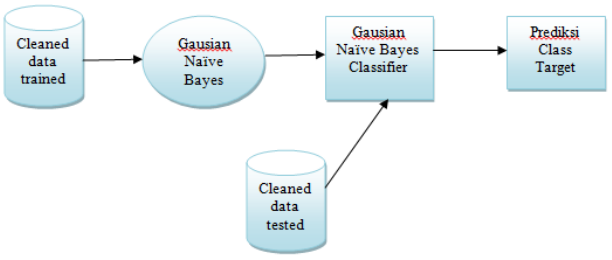
Langkah berikutnya adalah melakukan normalisasi data pada range yang sama. Setiap nilai dalam dataset memiliki value yang bervariatif dan memiliki batas atas dan bawah yang berbeda. Sehingga perlu dilakukan normalisasi agar data tersebut berada pada range yang sama. Untuk data PM10, SO_2, CO, NO_2, O_3 dan max akan dilakukan normalisasi dengan menggunakan min-max normalization dengan range nilai yang baru setelah dilakukan normalisasi yaitu 0-1. Rumus min max normalization yang digunakan adalah sebagai berikut

$$v' = \frac{v - (A)}{(A) - (A)}(new\_min\ (A) - new\_min\ (A)) + new\_min\ (A) \quad (5)$$

Sedangkan pada nama stasiun akan dilakukan one hot encoding sehingga data tersebut berubah menjadi bentuk yang bisa diolah dengan lebih mudah dengan menggunakan Gausian Naïve Bayes.

C. *Klasifikasi*

Tahapan berikutnya setelah melakukan preprocessing adalah klasifikasi. Tahapan klasifikasi tersebut dapat dilihat pada gambar 8. Dimana setelah melakukan data preprocessing dihasilkan cleaned Data, dimana cleaned data tersebut kemudian diproses dengan Gausian Naïve Bayes dan akan menghasilkan Gausian Naïve Bayes Classifier. Gausian Naïve Bayes Classifier tersebut nantinya akan digunakan untuk melakukan klasifikasi terhadap data testing yang diberikan dengan memperhatikan stasiun dan 5 parameterISPU.



GAMBAR 8. Tahapan Klasifikasi

Gausian Naïve Bayes Classifier yang dihasilkan disini berupa rata-rata tiap kelas, dan standar deviasi tiap kelas. Hal ini dikarenakan semua atribut yang digunakan berupa data kontinu, hanya 1 atribut yang bukan data kontinu yaitu atribut stasiun yang akan langsung dihitung probabilitasnya setelah dilakukan one hot encoding. Dari classifier tersebut nantinya digunakan untuk melakukan klasifikasi pada data testing yang diberikan. Dari data testing yang diberikan tersebut akan dihitung distribusi normal gausian untuk masing-masing atribut terhadap kelas untu setiap data testing. Setelah itu akan dilakukan perhitungan probabilitasnya untuk menentukan klasifikasi dengan menggunakan Gausian Naïve Bayes tersebut. Nilai probabilitas yang tertinggi tersebut yang akan diambil dan digunakan sebagai klasifikasi kelasnya. Hasil klasifikasi dengan menggunakan Gausian Naïve Bayes tersebut dapat dilihat pada gambar 9.

	stasiun	pm10	so2	co	o3	no2	max	critical	categori	prediksi
DKI1	(Bunderan HI)	45	18	4	49	10	68	PM25	SEDANG	SEDANG
DKI1	(Bunderan HI)	52	19	4	42	10	71	PM25	SEDANG	SEDANG
DKI1	(Bunderan HI)	60	22	5	68	11	86	PM25	SEDANG	SEDANG
DKI1	(Bunderan HI)	44	19	3	35	6	56	PM25	SEDANG	SEDANG
DKI1	(Bunderan HI)	46	18	8	64	9	64	O3	SEDANG	SEDANG
DKI1	(Bunderan HI)	37	15	6	82	11	82	O3	SEDANG	SEDANG
DKI1	(Bunderan HI)	54	20	7	78	12	78	O3	SEDANG	SEDANG
DKI1	(Bunderan HI)	57	17	6	82	14	82	O3	SEDANG	SEDANG
DKI1	(Bunderan HI)	62	19	6	65	11	91	PM25	SEDANG	SEDANG
DKI1	(Bunderan HI)	69	19	10	91	14	103	PM25	TIDAK SEHAT	TIDAK SEHAT
DKI1	(Bunderan HI)	61	17	11	77	13	87	PM25	SEDANG	SEDANG
DKI1	(Bunderan HI)	42	16	5	70	10	70	O3	SEDANG	SEDANG
DKI1	(Bunderan HI)	52	19	6	66	11	72	PM25	SEDANG	SEDANG
DKI1	(Bunderan HI)	55	19	6	47	11	76	PM25	SEDANG	SEDANG
DKI2	(Kelapa Gading)	69	25	8	70	13	96	PM25	SEDANG	SEDANG
DKI2	(Kelapa Gading)	72	21	10	122	11	122	O3	TIDAK SEHAT	TIDAK SEHAT
DKI2	(Kelapa Gading)	80	21	12	94	12	94	O3	SEDANG	SEDANG
DKI2	(Kelapa Gading)	56	21	6	51	9	73	PM25	SEDANG	SEDANG
DKI2	(Kelapa Gading)	54	22	6	118	7	118	O3	TIDAK SEHAT	TIDAK SEHAT
DKI2	(Kelapa Gading)	67	21	12	121	11	121	O3	TIDAK SEHAT	TIDAK SEHAT
DKI2	(Kelapa Gading)	69	26	8	84	13	101	PM25	TIDAK SEHAT	SEDANG

GAMBAR 9. Contoh Hasil Klasifikasi dengan Gausian Naïve Bayes

IV. Uji Coba

Setelah melakukan klasifikasi, proses berikutnya adalah pembentukan summary hasil klasifikasi terhadap data aktual yang mana nantinya digunakan untuk menghitung akurasi dari hasil klasifikasi terhadap data testing yang diujicobakan. Untuk mengatasi inbalanced data akan digunakan cross validation dengan k-fold=10. Hasil evaluasi model dapat dilihat pada gambar 10. Dari situ dapat dihitung pula akurasi yang diperoleh dengan menggunakan Gaussian Naïve Bayes

pada data lingkungan hidup DKI Jakarta untuk 5 stasiun yaitu sebesar 93,8%.

```
akurasi k-0=0.8760330578512396
akurasi k-1=0.9008264462809917
akurasi k-2=0.9421487603305785
akurasi k-3=0.9256198347107438
akurasi k-4=0.9173553719008265
akurasi k-5=0.9752066115702479
akurasi k-6=0.975
akurasi k-7=0.9416666666666667
akurasi k-8=0.95
akurasi k-9=0.975


Cross-Validation accuracy: 0.938
```

GAMBAR 10. Hasil Evaluasi dan Confusion Matrix

Selain melakukan klasifikasi, juga dilakukan uji coba terhadap atribut mana yang berpengaruh dan tidak terhadap penetuan kategori ISPU DKI Jakarta. Dimana data mentah yang diperoleh dari dataset memiliki 9 atribut dan 1 kelas target. Namun dalam datase tersebut terdapat beberapa atribut yang tidak bisa digunakan pada penelitian ini yaitu atribut max, critical, dan tanggal. Dimana pada pengujian jika algoritma Gaussian Naïve Bayes menggunakan atribut max dan critical hasil klasifikasi menjadi kurang akurat dan tingkat akurasi yang dihasilkan menurun hingga 66,7%. Sedangkan tanggal tidak memberikan pengaruh apapun pada tingkat akurasi. Confusion matrix dan akurasi tersebut dapat dilihat pada gambar 11.

```
akurasi k-0=0.5950413223140496
akurasi k-1=0.7107438016528925
akurasi k-2=0.4297520661157025
akurasi k-3=0.8760330578512396
akurasi k-4=0.7107438016528925
akurasi k-5=0.5950413223140496
akurasi k-6=0.725
akurasi k-7=0.6583333333333333
akurasi k-8=0.7
akurasi k-9=0.7666666666666667


Cross-Validation accuracy: 0.677
```

GAMBAR 11. Hasil Evaluasi dan Confusion Matrix dengan semua atribut

Dari hasil uji coba dengan menggunakan k-fold=10 untuk 5 atribut dilakukan perhitungan Recall, Precision, dan F1 Score. Hasil Recall, Precision dan F1 score dapat dilihat pada Tabel II. Didapatkan hasil WA Recall, Precision, dan F1 Score dengan 5 atribut sebesar 91%, sedangkan dengan menggunakan 7 atribut didapatkan WA Recall sebesar 65%, WA Precision 77%, dan WA F1 Score 62%.

TABEL II
PERHITUNGAN RECALL, PRECISION, DAN F1 SCORE

Iterasi		Baik	Sedang	Tidak Sehat	Weighted Average
1	Recall	0,86	0,89	1	0,876
	Precision	0,90	0,84	1	0,876
	F1 Score	0,88	0,86	1	0,874
2	Recall	0,88	0,96	0,6	0,901
	Precision	0,97	0,83	1	0,913
	F1-Score	0,92	0,89	0,75	0,9
3	Recall	0,88	0,94	0,90	0,928
	Precision	0,78	0,97	0,82	0,931
	F1-Score	0,82	0,95	0,86	0,925
4	Recall	1	0,95	0,75	0,924
	Precision	1	0,94	0,79	0,924
	F1-Score	1	0,95	0,77	0,927
5	Recall	0,65	0,97	0,92	0,92
	Precision	1	0,93	0,8	0,926
	F1-Score	0,79	0,95	0,86	0,918
6	Recall	0,92	0,99	1	0,976
	Precision	0,96	0,98	1	0,976
	F1-Score	0,94	0,98	1	0,972
7	Recall	0,91	0,98	1	0,974
	Precision	0,83	0,99	1	0,976
	F1-Score	0,87	0,99	1	0,979
8	Recall	1	0,98	0,55	0,941
	Precision	0,75	0,95	0,86	0,937
	F1-Score	0,86	0,97	0,67	0,94
9	Recall	0,83	0,99	0,71	0,949
	Precision	0,83	0,95	1	0,95
	F1-Score	0,83	0,97	0,83	0,947
10	Recall	1	0,99	0,71	0,982
	Precision	0,83	0,98	1	0,983
	F1-Score	0,91	0,99	0,83	0,986
Weighted Average	Recall	0,9336			
	Precision	0,9392			
	F1-Score	0,9368			

V. KESIMPULAN

Dari hasil uji coba yang telah dilakukan dalam melakukan klasifikasi kategori ISPU dengan menggunakan Gausian Naïve Bayes pada data lingkungan hidup 5 stasiun di DKI Jakarta dapat disimpukan:

1. Atribut max dan critical yang berada dalam dataset tidak memiliki pengaruh terhadap hasil klasifikasi kategori ISPU, terbukti dengan akurasi yang didapatkan bila mengikutkan semua atribut adalah 67,7%.
2. Atribut-atribut dari data yang berpengaruh terhadap klasifikasi kategori ISPU adalah PM10, SO2, CO, O3, dan NO2.
3. Dengan menggunakan 5 atribut dan gausian naïve bayes, sistem dapat memberikan klasifikasi dengan akurasi sebesar 93,8% dan memiliki error rate sebesar 6,2%. Sedangkan nilai WA Recall 93,36%, WA Precision 93,92% , dan WA F1 Score sebesar 93,68%.

PERAN PENULIS

Devi Dwi Purwanto: Konseptualisasi, metodologi, perangkat lunak, validasi, investigasi, kurasi data, penyusunan draft asli.

Eric Sugiharto: Investigasi, validasi, Analisis Formal, Investigasi, peninjauan dan penyuntingan

COPYRIGHT

DAFTAR PUSTAKA

[1] Indrayani and S. Asfiati, "Pencemaran Udara Akibat Kinerja Lalulintas Kendaraan Bermotor di Kota Medan," *Jurnal Pemukiman,* pp. 108-112, 2018.

[2] I. Ma'rufi, "Analisis Risiko Kesehatan Lingkungan(SO2 , H2S, NO2, dan TSP) Akibat Transportasi Kendaraan Bermotor di Kota Surabaya," *Media Pharmaceutica Indonesiana,* pp. 189-196, 2017.

[3] R. A. Lestari, A. R. Handik and S. I. Purwaningrum, "Analisis Risiko Karsinogenik Paparan PM10 Terhadap Pedagang di Kelurahan Pasar Jambi," *Jurnal Dampak,* pp. 59-65, 2019.

[4] A. Masito, "ANALISIS RISIKO KUALITAS UDARA AMBIEN (NO2 DAN SO2) DAN GANGGUAN PERNAPASAN PADA MASYARAKAT DI WILAYAH KALIANAK SURABAYA," *Jurnal Kesehatan Lingkungan,* pp. 394-401, 2018.

[5] L. M. Saleh, Keselamatan dan Kesehatan Kerja Kelautan : (Kajian Keselamatan dan Kesehatan Kerja Sektor Maritim), Deepublish Publisher, 2018.

[6] D. Nuvolone , D. Petri and F. Voller, "The Effects of Ozone on Human Health," *Enviromental Science and Pollution Research,* pp. 8074-8088, 2017.

[7] R. Darmawan, "ANALISIS RISIKO KESEHATAN LINGKUNGAN KADAR NO2 SERTA KELUHAN KESEHATAN PETUGAS PEMUNGUT KARCIS TOL," *Jurnal Kesehatan Lingkungan,* pp. 116-126, 2018.

[8] J. Abidin and F. A. Hasibuan, "Pengaruh Dampak Pencemaran Udara Terhadap Kesehatan untuk Menambah Pemahaman Masyarakat Awam Tentang Bahaya Dari Polusi Udara," in *Prosiding Seminar Nasional Fisika Universitas Riau IV*, Pekanbaru, 2019.

[9] V. Goar, S. S. Sarangdevot, G. Tanwar and A. Sharma, "Improve Performance of Extract, Transform and Load(ETL) in Data Warehouse," *International Journal on Computer Science and Engineering,* pp. 786-789, 2010.

[10] D. Harris and S. Harris, Digital Design and Computer Architecture, San Francisco: Morgan Kaufmann, 2012.

[11] M. M. Saritas and A. Yasar, "Perfromance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification," *International Journal of Intelligent Systems and Applications in Engineering,* pp. 88-91, 2019.

[12] A. A. Mahran, R. K. Hapsari and H. Nugroho, "Penerapan Naive Bayes Gaussian Pada Klasifikasi Jenis Jamur Bedasarkan Ciri Statistik Orde Pertama," *Networking Engineering Research Operation,* pp. 91-99, 2020.

[13] Kementerian Lingkungan Hidup dan Kehutanan, "WEB PORTAL DIREKTORAT PENGENDALIAN PENCEMARANUDARA," [Online]. Available: https://ditppu.menlhk.go.id/portal/read/indeks-standar-pencemar-udara-ispu-sebagai-informasi-mutu-udara-ambien-di-indonesia. [Accessed 20 December 2022].