Klasifikasi Tingkat Kualitas Udara Dki Jakarta Menggunakan Algoritma *Naive Bayes*

1st Abdul Aziiz Hendrie Kirono
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
abdulaziiz@students.telkomuniversity.ac.id

2nd Ibnu Asror
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
ibnu@telkomuniversity.ac.id

3rd Yanuar Firdaus Arie Wibowo
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
yanuar@telkomuniversity.ac.id

Abstrak
DKI Jakarta merupakan sebuah kota di Indonesia yang memiliki padat penduduk yang tinggi yang harus kita perhatikan kondisi kesehatannya. Kualitas udara yang baik bisa menunjang produktivitas masyarakat DKI Jakarta untuk lebih aktif dan menciptakan kondisi yang segar. Banyak data yang muncul tentang kualitas udara pada DKI Jakarta yang selalu menurun karena berbagai faktor, oleh karena itu dibutuhkan sistem klasifikasi dengan penggunaan algoritma naive bayes yang dapat menunjang informasi kepada masyarakat setempat. Algoritma naive bayes adalah salah satu algoritma klasifikasi berdasarkan probabilitas yang membandingkan data training dan data testing. Keduanya dibandingkan melalui beberapa tahap persamaan, yang akhirnya diperoleh probabilitas tertinggi yang ditetapkan sebagai informasi. Membuat sebuah model uji pada klasifikasi menggunakan algoritma *naive bayes* yang bertujuan mencari hasil yang baik. Hasil dari pengklasifikasian pada data Indeks Standar Pencemaran Udara pada kota DKI Jakarta menghasilkan yaitu, dengan rata-rata akurasi 88%, *precision* 85%, *recall* 96%, *f1-score* 90%.

Kata kunci : Klasifikasi, *Naive bayes*, *data mining*, Indeks Standar Pencemaran Udara, Udara

Abstract
DKI Jakarta is a city in Indonesia that has a high population density which we must pay attention to its health condition. Good air quality can support the productivity of the people of DKI Jakarta to be more active and create fresh conditions. There is a lot of data that appears about the air quality in DKI Jakarta which is always decreasing due to various factors, therefore a classfication system is needed using the naive bayes algorithm that can support information to the local communities. Naive bayes algorithm is a classification algorithm based on probability that compares training data and testing data. The two are compared through several stages of equations, finally the highest probability is obtained which is defined as information. Creating a test model on classification using the Naive Bayes algorithm that aims to find good results. The results of the classification on the Index Standar Pencemaran Udara data in the city of DKI Jakarta produce four results, with an average accuracy of 88%, precision 85%, recall 96%, f1-score 90%.

Keywords: Classificaiton, Naive bayes, data mining, Indeks Standar Pencemaran Udara, Air

I. PENDAHULUAN

A. Latar Belakang

Udara adalah suatu campuran gas alam yang mengelilingi permukaan bumi. Udara tersusun dari banyaknya gas, diantaranya nitrogen 78%, oksigen 20%, argon 0,93% dan karbon dioksida 0,30%, dan gas gas lainnya [1]. Udara memiliki peranan penting untuk segala aspek makhluk hidup di bumi, baik itu manusia, hewan atau tumbuhan. Tanpa adanya udara, makhluk hidup tidak ada yang bisa bertahan hidup. Udara berbentuk transparan, tidak bisa dilihat tapi bisa kita rasa. Banyak manfaat udara bersih bagi manusia untuk kelangsungan hidupnya, diantaranya: menyehatkan saluran pernapasan, menurunkan risiko penyakit kronis, memperpanjang usia, meningkatkan stamina dan fokus dan memperbaiki mood.

Pencemaran udara adalah kehadiran suatu atau lebih subtansi fisik, kimia, atau biologi di atmosfer dalam jumlah yang dapat membahayakan kesehatan manusia, hewan, dan tumbuhan, mengganggu estetika dan kenyamanan, atau merusak properti [2]. Pencemaran udara membuat kualitas udara disuatu daerah menjadi turun, dan kurang bagus untuk makhluk hidup, bahkan bisa membahayakan. Banyak faktor yang memicu penurunan kualitas udara disuatu daerah diantaranya aktivitas transportasi, industri, jasa, kegiatan lainnya yang meningkat, dan buangan sisa-sisa kegiatan ke udara. Di kota besar seperti DKI Jakarta, pencemaran udara merupakan salah satu masalah yang cukup serius yang perlu dihadapi oleh masyarakat DKI Jakarta.

Menurut data AirVisual yang ditampilkan oleh AQI, Indonesia menjadi negara ke-9 sebagai negara yang paling berpolusi di dunia pada tahun 2020, dan Jakarta menjadi kota ke-202 sebagai kota paling berpolusi di dunia tahun 2020 [3]. Dari data AirVisual yang berikan oleh AQI, Jakarta bisa dikategorikan menjadi kota dengan udara yang sedang – tidak sehat. Pada penelitian sebelumnya oleh Avijoy Chakma, Ben Vizena, Tingting Cao, Jerry Lin, Jing Zhang mengenai *air quality* klasifikasi menggunakan random forest mendapatkan nilai akurasi 63,62% dan CNN 68,74% [15]. Untuk *clustering* pada perubahan udara Southampton yang dilakukan oleh Jana Shafi mendapati kesimpulan perubahan cepat yang terjadi pada kualitas udara dari level paling bawah hingga mencapai level toksik tinggi di tempat yang sama akibat polutan api hanya dalam beberapa jam dengan menggunakan metode K-Means [16]

B. Topik dan Batasannya

Berdasarkan permasalahan udara yang sedang terjadi pada DKI Jakarta, dapat kita menerapkan suatu proses olah data untuk menjadi suatu nilai informasi. Data mining adalah suatu proses pengerukan atau pengumpulan informasi penting dari suatu data yang besar. Klasifikasi adalah model fungsi yang menggambarkan dan membedakan kelas atau konsep untuk prediksi masa depan, dari data yang diperoleh sebelumnya. Naive bayes merupakan suatu algoritma klasifikasi yang berakar pada probabilitas dan statistik. Dari Indeks Standar Pencemaran Udara (ISPU) tersebut dapat diimplementasikan pengklasifikasian untuk kualitas udara di DKI Jakarta. Penerapan klasifikasi tersebut menggunakan algoritma *naive bayes* dimana data dari Indeks Standar Pencemaran Udara (ISPU) yang bersifat *kontinu* cocok digunakan algoritma *naive bayes*. Hasil dari klasifikasi *dataset* Indeks Standar Pencemaran Udara (ISPU) harus diukur performansinya guna mengetahui penerapan klasifikasi pada *dataset* Indeks Standar Pencemaran Udara (ISPU) menggunakan algoritma *naive bayes* baik atau kurang baik.

C. Tujuan

Kondisi yang hendak dicapai adalah pengklasifikasian Indeks Standar Pencemaran Udara (ISPU) menggunakan algoritma naive bayes sebagai bahan analisis, dan menjadi sebuah informasi yang khususnya tentang kualitas udara di DKI Jakarta.

II. KAJIAN TEORI

A. Udara

Udara memiliki peranan yang sangat penting demi kelangsungan semua makhluk hidup terutama oksigen. Oksigen dihasilkan dari penyerapan karbondioksida (CO_2) oleh tumbuhan dan alga melewati proses fotosintesis. Bagi manusia dan hewan, oksigen digunakan sebagai zat untuk proses pernapasan. Udara merupakan suatu yang kita perlukan untuk bernapas sehari-hari [4]. Dari jenisnya udara dapat dibagi menjadi dua, yaitu udara bersih dan udara tidak bersih.

B. Pencemaran Udara

Penyebab faktor pencemaran udara oleh alam contohnya seperti kebakaran hutan, gunung meletus. Sedangkan faktor pencemaran udara oleh manusia contohnya kendaraan bermotor, limbah manusia, dan asap pabrik. Pencemaran udara dapat didefinisikan sebagai masuknya partikel pencemar ke dalam udara baik secara alamiah maupun akibat kegiatan manusia [5]. Pencemaran udara merupakan salah satu kerusakan lingkungan yang membuat penurunan kualitas udara yang tersebar bagi makhluk hidup. Faktor pencemaran udara terbagi 2 yaitu, manusia dan alam.

C. Indeks Standar Pencemaran Udara (ISPU)

Indeks standar pencemaran udara (ISPU) adalah angka yang tidak mempunyai satuan yang menggambarkan kondisi kualitas udara ambien di lokasi dan waktu tertentu yang didasarkan kepada dampak terhadap kesehatan manusia, nilai estetika dan makhluk hidup lainnya. Satuan dari nilai kosentrasi adalah (µgram/m3) yang dapat diartikan banyaknya molekul yang tersebar disuatu daerah.

TABEL 2. 1 Konvensi nilai kosentrasi

ISPU	24 Jam PM_{10} ($\mu g/m^3$)	24 Jam PM_{25} ($\mu g/m^3$)	24 Jam SO_2 ($\mu g/m^3$)	24 Jam CO ($\mu g/m^3$)	24 Jam O_3 ($\mu g/m^3$)	24 Jam NO_2 ($\mu g/m^3$)	24 Jam HC ($\mu g/m^3$)
0-50	50	15,5	52	4000	120	80	45
51-100	150	55,4	180	8000	235	200	100
101-200	350	150,4	400	15000	400	1130	215
201-300	420	250,4	800	30000	800	2260	432
>300	500	500	1200	45000	1000	300	648

Keterangan:

- Data pengukuran selama 24 jam secara terus-menerus.
- Hasil perhitungan ISPU parameter partikulat PM_{25} disampaikan tiap jam selama 24 jam.
- Hasil perhitungan ISPU parameter partikel debu (PM_{10}), Karbon monoksida (CO), sulfur dioksida (SO_2), Nitrogen dioksida (NO_2), Ozon permukaan (O_3), dan Hidrokarbon (HC), diambil nilai ISPU parameter tertinggi dan paling sedikit disampaikan setiap jam 09.00 dan jam 15.00

Perhitungan Indeks standar pencemaran udara (ISPU) ditetukan berdasarkan nilai ISPU batas atas, ISPU batas bawah, ambien batas atas, ambien batas bawah, dan konsentrasi ambien hasil pengukuran. Untuk persamaan perhitungan Indeks standar pencemaran udara (ISPU) sebagai berikut:

$$I = \frac{Ia-Ib}{Xa-Xb}(X_x - X_b) + I_b \quad (1)$$

Keterangan:

I = ISPU terhitung

Ia = ISPU batas atas
Ib = ISPU batas bawah
Xa = Konsentrasi ambien batas atas (µg/m^3)
Xb = Konsentrasi ambien batas bawah (µg/m^3)
Xx = Konsentrasi ambien nyata hasil pengukuran (µg/m^3)

Indeks Standar Pencemaran Udara (ISPU) ditetapkan dengan cara mengubah kadar pencemaran udara yang terukur menjadi suatu angka yang tidak berdimensi. Rentang nilai indeks yang ditetapkan oleh Indeks Standar Pencemaran Udara (ISPU) terdapat pada tabel 2.1.

TABEL 2. 2 Indeks Standar Pencemaran Udara(ISPU)

ISPU	Pencemaran Udara Level	Dampak Kesehatan
0-50	Baik	Tidak memberikan dampak bagi kesehatan manusia atau hewan
51-100	Sedang	Tidak berpengaruh pada kesehatan manusia ataupun hewan tetapi berpengaruh pada tumbuhan yang peka
101-199	Tidak Sehat	Bersifat merugikan pada manusia ataupun kelompok hewan yang peka atau dapat menimbulkan kerusakan pada tumbuhan atau nilai estetika
200-299	Sangat Tidak Sehat	Kualitas udara yang dapat merugikan kesehatan pada sejumlah segmen populasi yang terpapar
300-500	Berbahaya	Kualitas udara berbahaya yang secara umum dapat merugikan kesehatan yang serius pada populasi (misalnya iritasi mata, batuk, dahak, dan sakit tenggorokan).

Dimana:

a. *Range* 0-50 (Baik): Tingkat mutu udara yang sangat baik, tidak memberikan efek negatif terhadap manusia, hewan dan tumbuhan
b. *Range* 51-100 (Sedang): Tingkat mutu udara masih dapat diterima pada kesehatan manusia, hewan dan tumbuhan
c. *Range* 101-200 (Tidak Sehat): Tingkat mutu udara yang bersifat merugikan kepada manusia, hewan dan tumbuhan
d. *Range* 201-300 (Sangat Tidak Sehat): Tingkat mutu udara yang dapat meningkatkan risiko kesehatan pada sejumlah segmen populasi yang terpapar
e. *Range* 301+ (Berbahaya): Tingkat mutu udara yang dapat merugikan kesehatan serius pada populasi dan perlu penanganan cepat

D. Data Mining

Pengelolaan data pada saat ini menjadi aset yang sangat penting baik kepentingan pribadi atau perusahaan, dimana dari informasi data akan sangat berpengaruh dalam hal pengambilan sebuah kebutuhan yang akan diambil oleh pengguna. Proses pengumpulan sebuah informasi penting dari kumpulan suatu data istilah ini disebut *data mining*. Statistika, *database*, *machine learning*, *pattern recognition*, *artificial intelligence*, dan visualisasi semuanya memiliki peran dalam *data mining* [6]. Data yang akan diolah oleh *data mining* harus lah mengikuti alur *Knowledge Discovery in Databases* (KDD) dan sudah menjadi 2 bagian data yaitu, *data training* dan *data testing*. Prinsip pareto adalah sebuah prinsip yang percaya bahwa 80% hasil kinerja seseorang merupakan buah dari 20% upaya yang telah dilakukan [17]. Pembagian dari *data training* 80% dari *data set* dan *data test* 20% dari *data set*.

E. Machine Learning

Istilah *machine learning* mengacu pada pendeteksian otomatis dari sebuah data yang sifatnya bermakna [13]. *Machine learning* adalah aplikasi atau bagian dari *artificial intelligence* yang membuat sistem memiliki kemampuan belajar secara otomatis dan meningkatkan kemampuannya berdasarkan pengalaman tanpa diprogram secara eksplisit [14].

F. Klasifikasi

Klasifikasi adalah proses untuk mencari fungsi yang menjelaskan suatu kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek atau data yang label atau nilainya tidak diketahui. Untuk mencapai tujuannya tersebut proses klasifikasi membentuk model atau algoritma yang ditujukan mampu membedakan data kedalam kelas-

kelas yang berbeda berdasarkan fungsi tertentu. Model itu sendiri bisa berupa aturan "jika-maka", berupa pohon keputusan, atau formula matematis [7].

G. Naive Bayes

Naive bayes menjadi salah satu algoritma yang terdapat dalam metode klasifikasi. *Naive bayes* merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukan oleh ilmuan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman dimasa sebelumnya sehingga dikenal sebagai *teorema bayes*. Fungsi dari *Naive Bayes Gausian* adalah untuk memodelkan atau menghitung langsung atribut data yang bersifat kontinu. Dimana persamaan dari metode *Naive Bayes Gaussian* adalah [7]:

$$(X_i|C) = \frac{1}{\sqrt{2\pi\sigma_{c,i}^{2}}} \cdot e^{\left(-\frac{(xi-\mu c,i)2}{2a\frac{2}{c,i}}\right)} \quad (2)$$

Keterangan:
(Xi|C) = Likelihood
σ = Standar Deviansi dari atribut
μ = mean atau rata-rata atribut

H. Open Goverment Data

Open goverment data (OGP) telah menjadi sebuah platform bagi negara-negara yang berpartisipasi di dalamnya untuk mengembangkan tata pemerintahan yang mempromosikan keterbukaan, pelibatan masyarakat, akuntabilitas dan penggunaan tekonologi untuk memperkuat pemerintahan [8].

I. Jupyter Notebook

Pada saat ini banyak *Integrated Development Environment* atau yang kita kenal sebagai IDE yang tersedia secara *open soucre* untuk bahasa pemrograman python. *Integrated Development Environment* (IDE) sendiri bertugas sebagai *code editor* untuk menulis kode program. *Integrated Development Environment* (IDE) yang cukup populer terutama di kalangan *data scientist* untuk menulis bahasa pemrograman python adalah jupyter notebook [9].

J. Pandas

Python merupakan salah satu bahasa pemrograman yang memiliki library yang baik digunakan untuk *data scientists*. Pandas library adalah salah satu *tools* yang diberikan oleh python itu sendiri [10]. Struktur data dasar pada pandas dinamakan *data frame* yang berguna untuk memudahkan *programmer* untuk membaca sebuah file dengan banyak, jenis format seperti file .txt, .csv, dan tsv.

K. Numpy

Numpy merupakan salah satu library pada python yang berfungsi melakukan proses komputasi numerik. Numpy adalah sebuah library untuk bahasa pemrograman python, numpy memberikan dukungan untuk himpunan dan matriks multidimensi yang besar, dan dilengkapi koleksi sejumlah besar fungsi matematika untuk beroperasi pada himpunan ini [11].

L. Scikit-Learn

Scikit-learn adalah library *machine learning* khusus python yang menyediakan alat sederhana dan efisien untuk *data analysis* dan *data mining*. Scikit-learn menyediakan berbagai pilihan algoritma yang *supervised learning* dan *unsupervised learning*. Dalam *supervised learning* data dan label untuk *train* atau *test* sudah jelas, sedangkan *unsupervised learning* kebalikannya, labelnya tidak diketahui sehingga harus melakukan pelabelan berdasarkan persamaan data yang terkait.

M. Confusion Matrix

Confusion matrix adalah table matrix yang digunakan sebagai perhitungan performansi dari suatu model data atau algoritma [12]. Setiap baris dari matrix tersebut merepresentasikan kelas aktual dari data, dan setiap kolom merepresentasikan kelas prediksi dari data (atau sebaliknya).

TABEL 2. 3 Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	True Negative (TN)
Actual Negative	False Positive (FP)	False Negative (FN)

a. True Positive: Seberapa banyak data yang aktual kelasnya positif, dan model memprediksi positif
b. True Negative: Seberapa banyak data yang aktual kelasnya negatif, dan model memprediksi negatif
c. False Positive: Seberapa banyak data yang aktual kelasnya negatif, namun model memprediksi positif
d. False Negative: Seberapa banyak data yang aktual kelasnya positif, namun model memprediksi negatif

Accuracy: Total keseluruhan seberapa sering model benar mengklasifikasi. Permodelan persamaan matematika accurasy dapat dituliskan sebagai berikut

$$\frac{TP+TN}{TP+FP+FN+TN} \quad (3)$$

Precision: Ketika model memprediksi positif, seberapa sering prediksi itu benar. Permodelan persamaan matematika accurasy dapat dituliskan sebagai berikut

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

Recall: Ketika kelas aktualnya positif, seberapa sering model memprediksi positif. Permodelan persamaan matematika accurasy dapat dituliskan sebagai berikut

$$\frac{TP}{TP+FN} \quad (5)$$

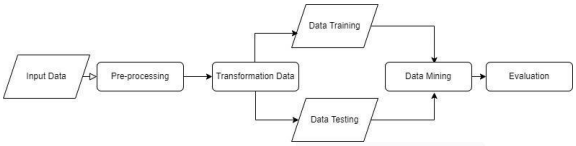
F1-Score: Merupakan rata-rata harmonik dari precision dan recall. Permodelan persamaan matematika accurasy dapat dituliskan sebagai berikut

$$\frac{2\ (Recall* Precision)}{(Recall+Precision)} \quad (6)$$

III. METODE

A. Sistematika Penelitian

Penelitian ini ditegaskan untuk mengukur tingkat kualitas udara pada DKI Jakarta dengan penerapan metode klasifikasi dengan menggunakan algoritma *naive bayes*. Adapun tahapan-tahapan dari penelitian ini:



Gambar 3. 1 Sistematika Penelitian

B. Kerangka Pemikiran

Dalam kerangka pemikiran ini terdapat beberapa tahapan selama penelitian dari mulai persiapan
penelitian hingga pembuatan laporan penelitian. Berikut adalah tabel dari kerangka pemikiran dalam proses penelitian:

TABEL 3. 1 Kerangka Penelitian

Kerangka Pemikiran		
Tahapan	Kegiatan	Hasil
Pendahuluan	1. Menentukan objek penelitian 2. Identifikasi masalah	Masalah
Pengumpulan data	1. Mencari sumber data 2. Studi litelatur	Data masalah
Pengolahan data	1. Proses pengolahan data KDD 2. Penentuan metode	Data *training* dan Data *testing*
Hasil	1. Implementasi metode 1. Perhitungan akurasi metode (confusion matrix)	*Knowledge*
Dokumentasi	Pembuatan Tugas Akhir	Tugas Akhir

C. Pengumpulan Data

Data yang digunakan dalam penelitian ini menggunakan *open goverment* yang didapatkan dari bank data portal https://data.jakarta.go.id/ . Data yang digunakan adalah data indeks standar pencemaran udara (ISPU) dari bulan februari hingga oktober dengan berformat csv. Berikut adalah contoh data yang diambil pada bulan februari 2021. Satuan dari nilai kosentrasi adalah (µgram/m^3) yang dapat diartikan banyaknya molekul yang tersebar disuatu daerah.

TABEL 3. 2 Data bulan februari

tanggal	pm10	pm25	so2	co	o3	no2	max	critical	categori	location
01/02/2021	73	126	38	26	46	34	126	PM25	TIDAK SEH	DKI5
02/02/2021	53	70	40	14	55	25	70	PM25	SEDANG	DKI3
03/02/2021	32	53	40	11	42	19	53	PM25	SEDANG	DKI3
04/02/2021	36	59	40	14	47	24	59	PM25	SEDANG	DKI5
05/02/2021	29	51	40	14	45	35	51	PM25	SEDANG	DKI3
06/02/2021	34	53	40	8	57	15	57	O3	SEDANG	DKI2
07/02/2021	33	55	40	10	57	13	57	O3	SEDANG	DKI2
08/02/2021	26	44	39	10	54	17	54	O3	SEDANG	DKI2
09/02/2021	33	57	40	13	47	22	57	PM25	SEDANG	DKI4
10/02/2021	50	64	40	13	49	16	64	PM25	SEDANG	DKI3
11/02/2021	38	57	43	13	35	17	57	PM25	SEDANG	DKI3
12/02/2021	63	98	43	16	33	42	98	PM25	SEDANG	DKI3
13/02/2021	59	89	40	12	40	16	89	PM25	SEDANG	DKI3
14/02/2021	55	73	40	11	42	19	73	PM25	SEDANG	DKI3
15/02/2021	42	66	40	13	37	25	66	PM25	SEDANG	DKI4
16/02/2021	43	63	40	11	42	26	63	PM25	SEDANG	DKI4
17/02/2021	46	71	40	25	41	37	71	PM25	SEDANG	DKI5
18/02/2021	53	70	44	13	41	29	70	PM25	SEDANG	DKI3
19/02/2021	32	52	42	20	41	33	52	PM25	SEDANG	DKI5
20/02/2021	45	63	39	13	53	20	63	PM25	SEDANG	DKI4
21/02/2021	36	52	39	10	48	17	52	PM25	SEDANG	DKI5
22/02/2021	68	103	42	22	41	40	103	PM25	TIDAK SEH	DKI3
23/02/2021	66	90	40	16	54	23	90	PM25	SEDANG	DKI3
24/02/2021	42	61	40	10	33	16	61	PM25	SEDANG	DKI4
25/02/2021	31	54	43	12	45	23	54	PM25	SEDANG	DKI3
26/02/2021	48	75	43	13	40	20	75	PM25	SEDANG	DKI4
27/02/2021	59	94	44	15	50	31	94	PM25	SEDANG	DKI4

Data yang digunakan merupakan data yang didapat dari *open goverment* yang bersifat sekunder. Data sekunder adalah data yang didapatkan secara tidak langsung dari sebuah objek peneletian. *Dataset* ini memiliki 10 atribut termasuk atribut kelas, yaitu:

a. Tanggal: Waktu pengukuran kualitas udara
b. Location: Lokasi pengukuran di *location*
c. PM_{10}: Partikulat salah satu parameter yang diukur
d. PM_{25}: Partikulat salah satu parameter yang diukur
e. SO_2: sulfida (dalam bentuk SO2) salah satu parameter yang diukur

f. CO: Carbon Monoksida salah satu parameter yang diukur
g. O_3: Ozon salah satu parameter yang diukur
h. NO_2: Nitrogen dioksida salah satu parameter yang diukur
i. Max: Nilai ukur paling tinggi dari seluruh parameter yang diukur dalam waktu yang sama
j. Critical: parameter yang hasilnya pengukuran indeks paling tinggi
k. Categori: kategori hasil perhitungan indeks standar pencemaran udara dan mencakup sebagai nilai dari kelas

D. Proses KDD

a. Input Data

Pengumpulan data untuk kasus ini bersumber dari bank data portal https://data.jakarta.go.id/ yang dimana memiliki banyak data berdasarkan bulan. Bulan yang dipilih adalah februari 2021, maret 2021, april 2021, mei 2021, juni 2021, juli 2021, agustus 2021, september 2021, oktober 2021. Terdapat total data record sebesar 3003 data yang tercatat yang ditunjukan oleh https://data.jakarta.go.id/ dari februari 2021 hingga oktober 2021. Pada proses ini akan diambil data yang relevan dengan data yang dianalis yaitu dengan mengurangi jumlah atribut dan data record yang ada dengan tujuan agar data tetap informatif. Untuk variabel input yaitu tanggal, *location*, PM_{10}, PM_{25}, SO_2, CO, O_3, NO_2, max, critical. Pemilihan atribut yang digunakan hanya tanggal, *location*, PM_{10}, PM_{25}, SO_2, CO, O_3, NO_2 saja yang mendekati pengaruh nilai dari variabel output yang akan dilakukan didalam *pre-processing*, dan untuk variabel output adalah *category*.

	tanggal	pm10	pm25	so2	co	o3	no2	location
0	Feb	73	126	38	26	46	34	DKI5
1	Feb	53	70	40	14	55	25	DKI3
2	Feb	32	53	40	11	42	19	DKI3
3	Feb	36	59	40	14	47	24	DKI5
4	Feb	29	51	40	14	45	35	DKI3
...	...	...	...	...	...	...	...	...
268	Oct	62	90	64	15	50	39	DKI4
269	Oct	54	78	67	16	56	39	DKI4
270	Oct	54	79	80	19	49	35	DKI2
271	Oct	64	103	81	15	58	40	DKI4
272	Oct	56	79	63	28	56	32	DKI4

GAMBAR 3. 2 Data Selection

Pre-Processing

Pada proses *pre-processing* akan dilakukan pembersihan data yang tidak lengkap, kosong, *noise*, duplikat, dan data yang tidak konsisten. Pada tahap ini dilakukan proses *pre-processing* dimana atribut yang tidak terpakai didalam *dataset* akan dibuang.

```
X = df.drop(['categori', 'tanggal', 'location', 'critical'],axis = 1)
X
```

	pm10	pm25	so2	co	o3	no2	max
0	73	126	38	26	46	34	126
1	53	70	40	14	55	25	70
2	32	53	40	11	42	19	53
3	36	59	40	14	47	24	59
4	29	51	40	14	45	35	51
...	...	...	...	...	...	...	...
268	62	90	64	15	50	39	90
269	54	78	67	16	56	39	78
270	54	79	80	19	49	35	80
271	64	103	81	15	58	40	103
272	56	79	63	28	56	32	79

273 rows × 7 columns

GAMBAR 3. 3 Pre-processing

b. Data Training

Sebelum proses pengklasifikasian data yang sudah diseleksi data akan dibagi menjadi dua bagian yaitu *data training* dan *data testing*. Terdapat record data dengan total 3003 data set. Dari total data set tersebut dibagi menjadi 2 bagian, *data training* dan *data testing*. Pada tahap ini dilakukan pembuatan *data training* untuk melatih algoritma *naive bayes*. Perbandingan pembuatan *data training* adalah 80% dari *dataset*.

c. Data Test

Pada tahap ini dilakukan pembuatan data testing untuk mengetahui performa *naive bayes*. Perbandingan pembuatan *data testing* adalah 20% dari *dataset*.

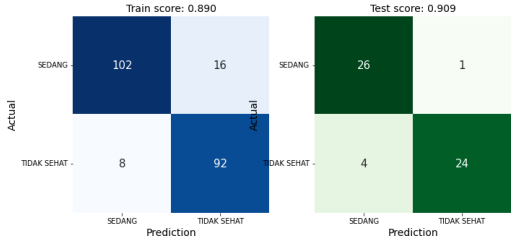
d. Data Mining

Data yang sudah siap akan di uji mengunakan motode yang telah ditentukan. Metode yang digunakan klasifikasi dengan menggunakan algoritma *naive bayes*. Tipe data yang bersifat *kontinu*, yang berbentuk *numerical* dan *categorycal* yang cocok digunakan pada *naive bayes*. Proses pengujian metode akan dikerjan pada bahasa pemrograman python menggunakan library yang tersedia. Karena data yang bersifat *kontinu*, maka *gaussian naive bayes* yang akan diuji pada permodelan ini.

e. Interpretation / Evaluation

Pada tahap *interpretation* atau *evaluation* ini dilakukan proses pembentukan keluaran yang mudah dimengerti yang bersumber pada hasil dari proses *data mining* berbentuk sebuah informasi dari pengujian *data training* dan *data testing*. Hasil evaluasi model dilakukan dipresentasikan menggunakan tabel matriks. Cara menghitung nilai dari hasil evaluasinya dengan menguji hasil model *data mining* menggunakan perhitungan *confusion matrix*. Pada tahap ini *data training* dan *data testing* yang sudah dilakukan permodelan algoritma *naive*

bayes akan diujikan performansinya.



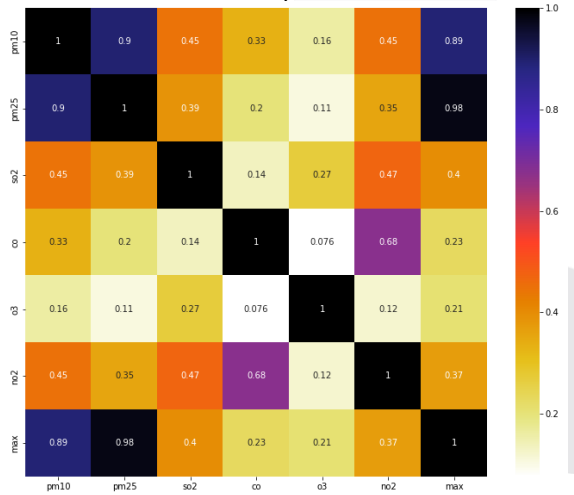
GAMBAR 3. 4 Visualisasi Performansi

IV. HASIL DAN PEMBAHASAN

A. Skenario Uji

Terdapat model pengujian dalam pengerjaan Tugas Akhir ini demi menjadi perbandingan untuk mencari hasil terbaik. Model pengujian pada pengerjaan ini dengan membagi *dataset* menjadi 2 bagian utama, yaitu *data training* dan *data testing*. Pemgbagian *dataset* terhadap *data training* dan *data testing*, yaitu 80% *data training* dan 20% *data testing*.

Selain pembagian dataset, dilakukan juga seleksi fitur berbasis pearson correlation. Penggunaan seleksi fitur ini guna untuk mengoptimasi fitur terhadap permodelan data mining. Berikut adalah penyebaran data terhadap seleksi fitur pearson correlation:



GAMBAR 4. 1 Pearson Correlation

Dari hasil seleksi fitur menggunakan *pearson correlation*, fitur yang tidak digunakan adalah 'max'.

B. Analisis Hasil Pengujian

Hasil implementasi dari penerapan klasifikasi pada data Indeks Standar Pencemaran Udara (ISPU) yang diambil pada bulan februari tahun 2021 hingga oktober 2021 dengan metode algoritma *naive bayes* dengan empat skenario uji memberikan hasil yang berbeda pada tiap modelnya. *Macro average* adalah menghitung matriks secara bebas untuk setiap kelas untuk mengambil rata-rata. *Weigthed average* adalah menghitung rata-rata dengan memperhitungkan bobot pada setiap data.

TABEL 4. 1 Hasil Uji

	Precision	Recall	F1-score	Support
Sedang	87%	96%	91%	27
Tidak Sehat	96%	86%	91%	28
Accuracy			91%	55
Macro avg	91%	91%	91%	55
Weighted avg	91%	91%	91%	55

Berdasarkan tabel 4.1 menunjukan hasil *accuracy* 91% dimana *score* dari *category* "sedang" pada menghasilkan *precision* 87%, *recall* 96%, *f1-score* 91%, dan *category* "tidak sehat" menghasilkan *precision* 96%, *recall* 86%, *f1-score* 91%. Menghasilkan nilai rata-rata dengan *macro average* menghasilkan *precision* 91%, *recall* 91%, *f1-score* 91%, sedangkan *weighted average* menghasilkan *precision* 91%, *recall* 91%, *f1-score* 91%.

V. KESIMPULAN

Berdasarkan hasil dari perhitungan yang telah dilakukan terhadap data ISPU DKI Jakarta dengen menggunakan teknik klasifikasi *data mining* menggunakan algoritma *naive bayes*, dapat disimpulkan dari permodelan uji skenario menghasil hasil yang baik. Hasil dari pengklasifikasian pada data Indeks Standar Pencemaran Udara pada kota DKI Jakarta menghasilkan yaitu, dengan rata-rata akurasi 88%, *precision* 85%, *recall* 96%, *f1-score* 90%. Dapat kita simpulkan, penggunaan selection fitur terhadap data Indeks Standar Pencemaran Udara pada kota DKI Jakarta sangat berperan penting terhadap penigkatan akurasi dari permodelan klasifikasi dengan algoritma *naive bayes*.

REFERENSI

[1] Fardiaz "Mikrobiologi Pangan I", Jakarta, Gramedia Pustaka Utama, 1992.

[2] Kementrian Lingkungan Hidup dan Kehutanan DITJEN Pengedalian Pencemaran Dan Kerusakan Lingkungan DIREKTORAT Pengendalian Pencemaran Udara "INDEKS STANDAR PENCEMARAN UDARA (ISPU) SEBAGAI INFORMASI MUTU UDARA AMBIEN DI INDONESIA", 2020 [online].
Avaliable:

https://ditppu.menlhk.go.id/portal/read/indeks-standar-pencemar-udara-ispu-sebagai-informasi-mutu-udara-ambien-di-indonesia.

[3] IQ*Air* "Kualitas udara di Indonesia", 2021 [online]. *Avaliable*: https://www.iqair.com/id/indonesia

[4] Arya Wardana, Wisnu "Dampak pencemaran lingkungan", Yogyakarta: Andi, 2001.

[5] Soedomo, Moestikahadi "Pencemaran Udara", Bandung: ITB, 2001

[6] David J. Hand, Heikki Mannila, Padhraic "Principles of Data Mining", ISBN 0-262-08290-X.

[7] Ali Haghpanah, Mohammad Taheri, "A Non-paramatric mixture of Gaussian naive Bayes classifiers based on local independent features", IEEE 17652603, 2017.

[8] Andy Heny Mulawati Nurdin, "MENUJU PEMERINTAHAN TERBUKA (*OPEN GOVERMENT*) MELALUI PENERAPAN *E GOVERMENT*", Vol. 5, No.1/ Juni 2018, Intitut Pemerintahan Dalam Negeri, 2018.

[9] Budi Ariwibowo, "Belajar Python dari Nol via Jupyter Notebook", CV Diandra Kreatif, ISBN 9786232400733.

[10] Matt Harrison, "Learning the Pandas Library Python Tools for Data Mungining, Analysis, dan Visual", CreateSpace Independent Publishing Platform, 2016.

[11] Charles R Harris, K. Jarrod Millman, Stefan J. van der Walt, dll "Array programing with Numpy", ISSN 14764687, 2020.

[12] Irkham Widhi Saputro, Bety Wulan Sari, "Uji Performa Algoritma Naive Bayes untuk Prediksi Masa Studi Mahasiswa", Citec Journal, Vol. 6, No. 1 ISSN: 2460-4259, Universitas AMIKOM Yogyakarta, 2019.

[13] Ethem Alpaydin, "Machine Learining, Revised and Updated Edition", MIT PRESS.

[14] Purba Daru Kusuma, "Machine Learning Teori, Program, Dan Studi Kasus", ISBN 9786230210839, 6230210835, 2020.

[15] Avijoy Chakma, Ben Vizena, Tingting Cao, Jerry Lin, Jing Zhang, "IMAGE-BASED AIR QUALITY ANALYSIS USING DEEP CONVOLUTIONAL NEURAL NETWORK", IEEE 978-1-5090 2175-8/17, 2017

[16] Jana Shafi, "K-Means Clustering Analysing Abrupt Change in Air Quality", IEEE 978-1-7281-6387-1/20, 2020

[17] Bunkley, Nick "Joseph Juran, 103, Pioneer in Quality Control, Dies" The New York Times, 2008.