Perbandingan Akurasi Algoritma Naïve Bayes dan Support Vector Machine Untuk Pencemaran Udara Di DKI Jakarta

TUGAS AKHIR

Muhammad Ikhsan Haikal
41518010116

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS MERCU BUANA
JAKARTA
2022

Perbandingan Akurasi Algoritma Naïve Bayes dan Support Vector Machine Untuk Pencemaran Udara Di DKI Jakarta

Tugas Akhir

Diajukan Untuk Melengkapi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer

Oleh:
Muhammad Ikhsan Haikal
41518010116

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS MERCU BUANA
JAKARTA
2022

LEMBAR PERNYATAAN ORISINALITAS

LEMBAR PERNYATAAN ORISINALITAS

Yang bertanda tangan dibawah ini:

NIM	: 41518010116
Nama	: Muhammad Ikhsan Haikal
Judul Tugas Akhir	: Perbandingan Akurasi Algoritma Naïve Bayes dan Support Vector Machine Untuk Pencemaran Udara Di DKI Jakarta

Menyatakan bahwa Laporan Tugas Akhir saya adalah hasil karya sendiri dan bukan plagiat. Apabila ternyata ditemukan didalam laporan Tugas Akhir saya terdapat unsur plagiat, maka saya siap untuk mendapatkan sanksi akademik yang terkait dengan hal tersebut.

Jakarta, 7 Juli 2022

Muhammad Ikhsan Haikal

ii

SURAT PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR

Sebagai mahasiswa Universitas Mercu Buana, saya yang bertanda tangan di bawah ini :

Nama Mahasiswa	:	Muhammad Ikhsan Haikal
NIM	:	41518010116
Judul Tugas Akhir	:	Perbandingan Akurasi Algoritma Naïve Bayes dan Support Vector Machine Untuk Pencemaran Udara Di DKI Jakarta

Dengan ini memberikan izin dan menyetujui untuk memberikan kepada Universitas Mercu Buana **Hak Bebas Royalti Noneksklusif** (*None-exclusive Royalty Free Right*) atas karya ilmiah saya yang berjudul diatas beserta perangkat yang ada (jika diperlukan).

Dengan Hak Bebas Royalti/Noneksklusif ini Universitas Mercu Buana berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat dan mempublikasikan tugas akhir saya.

Selain itu, demi pengembangan ilmu pengetahuan di lingkungan Universitas Mercu Buana, saya memberikan izin kepada Peneliti di Lab Riset Fakultas Ilmu Komputer, Universitas Mercu Buana untuk menggunakan dan mengembangkan hasil riset yang ada dalam tugas akhir untuk kepentingan riset dan publikasi selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Jakarta, 7 Juli 2022

Muhammad Ikhsan Haikal

SURAT PERNYATAAN LUARAN TUGAS AKHIR

SURAT PERNYATAAN LUARAN TUGAS AKHIR

Sebagai mahasiswa Universitas Mercu Buana, saya yang bertanda tangan di bawah ini :

Nama Mahasiswa : Muhammad Ikhsan Haikal
NIM : 41518010116
Judul Tugas Akhir : Perbandingan Akurasi Algoritma Naïve Bayes dan Support Vector Machine Untuk Pencemaran Udara Di DKI Jakarta

Menyatakan bahwa :

1. Luaran Tugas Akhir saya adalah sebagai berikut :

No	Luaran	Jenis			Status	
1	Publikasi Ilmiah	Jurnal Nasional Tidak Terakreditasi			Diajukan	✓
		Jurnal Nasional Terakreditasi	✓			
		Jurnal International Tidak Bereputasi			Diterima	
		Jurnal International Bereputasi				
	Disubmit/dipublikasikan di :	Nama Jurnal	: Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)			
		ISSN	: 2580-0760			
		Link Jurnal	: http://jurnal.iaii.or.id/index.php/RESTI/			
		Link File Jurnal Jika Sudah di Publish	:			

2. Bersedia untuk menyelesaikan seluruh proses publikasi artikel mulai dari submit, revisi artikel sampai dengan dinyatakan dapat diterbitkan pada jurnal yang dituju.
3. Diminta untuk melampirkan scan KTP dan Surat Pernyataan (Lihat Lampiran Dokumen HKI), untuk kepentingan pendaftaran HKI apabila diperlukan

Demikian pernyataan ini saya buat dengan sebenarnya.

Jakarta, 7 Juli 2022

Muhammad Ikhsan Haikal

LEMBAR PERSETUJUAN PENGUJI

LEMBAR PERSETUJUAN PENGUJI

NIM	:	41518010116
Nama	:	Muhammad Ikhsan Haikal
Judul Tugas Akhir	:	Perbandingan Akurasi Algoritma Naïve Bayes dan Support Vector Machine Untuk Pencemaran Udara Di DKI Jakarta

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Jakarta, 29 Juli 2022

(Ir. Emil R. Kaburuan, Ph.D., IPM.)

v

LEMBAR PERSETUJUAN PENGUJI

NIM	:	41518010116
Nama	:	Muhammad Ikhsan Haikal
Judul Tugas Akhir	:	Perbandingan Akurasi Algoritma Naïve Bayes dan Support Vector Machine Untuk Pencemaran Udara Di DKI Jakarta

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Jakarta, 29 Juli 2022

(Misbahul Fajri, ST, MTI)

LEMBAR PERSETUJUAN PENGUJI

NIM	:	41518010116
Nama	:	Muhammad Ikhsan Haikal
Judul Tugas Akhir	:	Perbandingan Akurasi Algoritma Naïve Bayes dan Support Vector Machine Untuk Pencemaran Udara Di DKI Jakarta

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Jakarta, 29 Juli 2022

(Saruni Dwiasnati, ST., MM., M.Kom)

LEMBAR PENGESAHAN

NIM	:	41518010116
Nama	:	Muhammad Ikhsan Haikal
Judul Tugas Akhir	:	Perbandingan Akurasi Algoritma Naïve Bayes dan Support Vector Machine Untuk Pencemaran Udara Di DKI Jakarta

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Jakarta, 29 Juli 2022

Menyetujui,

(Raka Yusuf, ST, MTI)
Dosen Pembimbing

Mengetahui,

(Wawan Gunawan, S.Kom, MT)	(Ir. Emil R. Kaburuan, Ph.D., IPM.)
Koord. Tugas Akhir Teknik Informatika	Ka. Prodi Teknik Informatika

KATA PENGANTAR

Puji syukur kita panjatkan kepada ALLAH SWT yang telah memberikan nikmat dan rahmatnya sehingga penulis dapat menyelesaikan laporan tugas akhir ini yang berjudul "Analisa Algoritma Naïve Bayes Untuk Memprediksi Tingkat Pencemaran Udara Di Kota DKI Jakarta" dengan lancar. Tugas akhir ini disusun untuk memenuhi syarat memperoleh gelar sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer Universitas Mercu Buana. Tugas akhir ini tentunya tidak lepas dari bantuan ketersediaan data, bimbingan, masukan, dan arahan dari berbagai pihak.

Penulis menyadari bahwa tanpa bantuan dan bimbingan dosen pembimbing, orang tua serta teman-teman tidak akan terselesaikan dengan baik tugas akhir ini. Oleh karena itu, penulis mengucapkan terima kasih kepada :

1. Orang tua yang selalu memberikan dukungan secara penuh dan doa sehingga dapat menyelesaikan semua laporan tugas akhir dengan lancar.
2. Bapak Emil Robert Kaburuan, PhD selaku Kepala Program Studi Teknik Informatika.
3. Bapak Wawan Gunawan, S.Kom, MT selaku Koordinator Tugas Akhir Jurusan Teknik Informatika.
4. Ibu Harni Kusniyati, M.Kom selaku Dosen Pembimbing Akademik.
5. Bapak Raka Yusuf, ST, MTI selaku Dosen Pembimbing Tugas Akhir.
6. Teman-teman dan sahabat yang selalu memberikan dukungan dan memotivasi dalam melakukan penulisan tugas akhir ini agar dapat terselesaikan dengan baik.

Akhir kata, penulis berharap tugas akhir ini dapat menjadi bermanfaat bagi pembaca dan menambah wawasan pengetahuan semua pihak.

Jakarta, 7 Juli 2022

Muhammad Ikhsan Haikal

DAFTAR ISI

HALAMAN SAMPUL i

HALAMAN JUDUL i

LEMBAR PERNYATAAN ORISINALITAS ii

SURAT PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR ... iii

SURAT PERNYATAAN LUARAN TUGAS AKHIR iv

LEMBAR PERSETUJUAN PENGUJI v

ABSTRAK ix

ABSTRACT x

KATA PENGANTAR xi

DAFTAR ISI xii

NASKAH JURNAL 1

KERTAS KERJA 8

BAB 1. LITERATUR REVIEW 9

BAB 2. ANALISIS DAN PERANCANGAN 23

BAB 3. SOURCE CODE 32

BAB 4. DATASET 36

BAB 5. TAHAPAN EKSPERIMEN 37

BAB 6. HASIL SEMUA EKSPERIMEN 41

DAFTAR PUSTAKA 46

LAMPIRAN DOKUMEN HAKI 50

LAMPIRAN KORESPONDENSI 52

NASKAH JURNAL

Terbit online pada laman web jurnal: **http://jurnal.iaii.or.id**

JURNAL RESTI

(Rekayasa Sistem dan Teknologi Informasi)

Vol. 6 No. x (2022) x - x ISSN Media Elektronik: 2580-0760

Perbandingan Akurasi Naïve Bayes dan Support Vector Machine Untuk Pencemaran Udara Jakarta

Muhammad Ikhsan Haikal[1], Raka Yusuf[2]
[1]Teknik Informatika, Ilmu Komputer, Universitas Mercu Buana
[2]Teknik Informatika, Ilmu Komputer, Universitas Mercu Buana
[1]Ikhsanhaikal2399@gmail.com

Abstract

Air pollution is a decrease in air quality so that the air experiences a decrease in quality in its use which ultimately can no longer be used as it should according to its function. This study aims to determine the application of the Naïve Bayes algorithm and Support Vector Machine to obtain good accuracy results in predicting the level of air pollution in DKI Jakarta. With the application of the Naïve Bayes algorithm and Support Vector Machine for predicting the level of air pollution, it is hoped that it can reduce casualties and other losses caused by this air pollution. And can facilitate the DKI Jakarta provincial government in making decisions on the right steps to solve air pollution problems in the DKI Jakarta area. The Support Vector Machine algorithm has better results with an average accuracy rate of 98.5% with an average level of precision, recall and f1-score of 92.58%, 98.75%, 95%. Meanwhile, the Naïve Bayes algorithm only obtained an accuracy of 87.75% with an average level of precision, recall, and f1-score of 73.5%, 91.16%, 77.6%. The results produced by the two algorithms are also influenced by the large amount of training data and test data.

Keywords: Naïve Bayes, SVM, Classification, Air pollution

Abstrak

Pencemaran udara adalah turunnya kualitas udara sehingga udara mengalami penurunan mutu dalam penggunaannya yang akhirnya tidak dapat digunakan lagi sebagaimana mestinya sesuai fungsinya. Penelitian ini bertujuan untuk mengetahui penerapan algoritma Naïve Bayes dan Support Vector Machine untuk mendapatkan hasil akurasi yang baik dalam memprediksi tingkat pencemaran udara di DKI Jakarta. Dengan adanya penerapan algoritma Naïve Bayes dan Support Vector Machine untuk prediksi tingkat pencemaran udara, diharapkan dapat mengurangi korban jiwa dan kerugian-kerugian lain yang disebabkan oleh penecemaran udara ini. Serta dapat memudahkan pemerintah provinsi DKI Jakarta dalam pengambilan keputusan langkah yang tepat untuk menyelesaikan masalah penceraman udara di wilayah DKI Jakarta. Algoritma Support Vector Machine memiliki hasil yang lebih baik dengan tingkat akurasi rata-rata sebesar 98.5% dengan tingkat rata-rata *precision*, *recall* dan *f1-score* sebesar 92,58%, 98,75%, 95%. Sedangkan algoritma Naïve Bayes hanya memperoleh akurasi sebesar 87.75% dengan tingkat rata-rata precision, recall, dan f1-score sebesar 73.5%, 91,16%, 77,6%. Hasil yang dihasilkan oleh kedua algoritma juga di pengaruhi oleh besarnya jumlah data latih dan data uji.

Kata kunci: : Naïve Bayes, SVM, Klasifikasi, Pencemaran Udara

1. Pendahuluan

Pencemaran udara adalah turunnya kualitas udara sehingga udara mengalami penurunan mutu dalam penggunaannya yang akhirnya tidak dapat digunakan lagi sebagaimana mestinya sesuai fungsinya. Pencemaran udara yang terjadi dapat menyebabkan gangguan kesehatan manusia terutama saluran pernapasan selain itu juga akan berdampak pada lingkungan.

Kota Jakarta merupakan salah satu kota besar dan terpadat di Indonesia yang rawan akan terjadinya polusi udara. Berdasarkan data dari Badan Pusat Statistik Kota Jakarta, jumlah kendaraan terus mengalami peningkatan tiap tahun terutama jumlah

Diterima Redaksi: xx-xx-2022 | Selesai Revisi: xx-04-2022 | Diterbitkan Online: xx-04-2022

sepeda motor dengan rata-rata pertumbuhan 5,3% per tahun.

Berdasarkan penelitian yang dilakukan oleh Amri Wicahyo, Ahmad Pudoli, Dewi Kusumaningsih, Noripansyah pada tahun 2021 yang berjudul "Penggunaan Algoritma Naive Bayes dalam klasifikasi Pengaruh Pencemaran Udara" [1]. Penelitian ini mengenai menyelesaikan masalah dalam memberikan informasi berupa klasifikasi pengaruh terhadap indikator parameter, Proses pengklasifikasian pengaruh tersebut menggunakan metode algoritma Naive Bayes dengan hasil klasifikasi berupa tidak ada efek, sedikit berbau, luka pada beberapa spesies tumbuhan akibat kombinasi dengan SO2, luka pada beberapa spesies tumbuhan akibat kombinasi dengan O3.

Dalam menentukan dampak atau pengaruh pencemaran udara terdapat indeks parameter sebagai standar pengukuran dampak atau pengaruh pencemaran udara. Terdapat lima indeks parameter sebagai standar pengukuran yaitu Partikulat (PM10), Sulfur Dioksida (SO2), Karbon Monoksida (CO), Ozon (O3), dan Nitrogen Dioksida (NO2). Pada tahap pengujian performa menggunakan Confusion Matrix untuk menghasilkan nilai Accuracy, Precison dan Recall. Data tersebut dipecah menjadi 70% Data Training dan 30% Data Testing.

Hasil dari penelitian ini bahwa aplikasi klasifikasi Pengaruh Pencemaran Udara berhasil menentukan klasifikasi pengaruh pencemaran udara menggunakan metode Naive Bayes terhadap dataset bulan Januari tahun 2018 hingga bulan Juni tahun 2020. Hasil klasifikasi pengaruh pencemaran udara mendapatkan nilai akurasi sebesar 96%. Hasil akurasi cukup baik untuk menentukan 129 data testing terhadap data training berkisar yang 4000 baris data, sehingga peneliti menyimpulkan metode Naive Bayes baik dalam melakukan klasifikasi pengaruh pencemaran udara.

2. Tinjauan Pustaka

2.1. Naïve Bayes

Algoritma Naive Bayes secara umum merupakan berasal dari teorema Bayes yang memiliki pengertian suatu proses prediksi peluang untuk masa depan berdasarkan pengalaman masa lalu. Algoritma Naive Bayes sendiri merupakan salah satu teknik pengklasifikasian terhadap data yang bersifat kuantitatif dan diskrit dengan menggunakan metode probabilitas dan statistik untuk mendapatkan hasil hipotesis sebagai informasi baru dalam mengambil keputusan. Menurut peneliti lain Mustafa, mengemukakan sebagai berikut: "Algoritma Naive Bayes Classifier (NBC) dapat mengolah data kuantitatif dan data diskrit yang hanya memerlukan sejumlah kecil data pelatihan untuk perhitungan estimasi peluang yang dibutuhkan untuk klasifikasi"[2]. Persamaan teorema *Bayes*[3] dapat dilihat dibawah ini.

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

Keterangan :

X :Data dengan class yang belum diketahui.
H :Hipotesis data merupakan suatu class spesifik.
P(H|X) :Probabilitas hipotesis H berdasarkan kondisi X (posteriori probabilitas).
P(H) :Probabilitas hipotesis H (prior probabilitas).
P(X|H) :Probabilitas X berdasarkan kondisi pada hipotesis H
P(X) :Probabilitas X

2.2. Support Vector Machine

Support Vector Machine (SVM) adalah metode klasifikasi yang bertujuan untuk menemukan hyperplane terbaik yang memisahkan dua kelas di ruang input. SVM memiliki prinsip dasar yaitu linear classifier dan dikembangkan untuk bekerja pada non-linear masalah dengan memasukkan kartu trik dalam dimensi tinggi ruang kerja. Dalam data yang dapat dipisahkan secara linier, data yang ada dapat dipisahkan secara linier. Misalkan ada pola { $x1, x2, ...$ } adalah a anggota dari dua kelas, yaitu +1 dan -1. Setiap pola tergabung dalam setiap kelas memiliki ciri khas tersendiri [4].

2.3. Klasifikasi

Klasifikasi merupakan suatu proses pengelompokan suatu data menjadi beberapa kelompok yang memiliki keterkaitan dalam ruang lingkup masalah yang ada dengan tujuan akhir untuk menemukan model atau fungsi yang menggambarkan dan membedakan tiap kelompok data sehingga dapat memprediksikan suatu kelompok data dari data yang belum terdapat pada beberapa kelompok data. Berdasarkan peneliti lain mengemukakan sebagai berikut: "Klasifikasi adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui, dalam mencapai tujuan tersebut klasifikasi membentuk suatu model yang mampu membedakan data dalam kelas-kelas yang berbeda berdasarkan aturan atau fungsi tertentu, dengan model yang berupa pohon keputusan, atau formula matematis" [1].

DOI: https://doi.org/10.29207/resti.v6iX.xxx

2.4. Confusion Matrix & Classification Report

Hasil klasifikasi yang didapatkan dari kedua algoritma tersebut dianalisis dan di evaluasi menggunakan *Confusion Matrix* dan *Classification Report* sesuai dengan hasil akurasi yang telah didapatkan.

Predicted Values	Actual Values	
	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

Tabel 1. *Confusion Matrix* [3]

Keterangan :

- **TP (True Positif)** adalah jumlah data positif yang terklasifikasi dengan benar bahwa data itu positive oleh system.
- **TN (True Negative)** adalah jumlah data negative yang terklasifikasi dengan benar bahwa data itu negative oleh system.
- **FN (False Negative)** adalah jumlah data negative namun terklasifikasi salah oleh sistem bahwa sebenarnya positif.
- **FP (False Negative)** adalah jumlah data positif namun terklasifikasi salah oleh sistem bahwa sebenernya adalah negative.

Setelah melakukan evaluasi menggunakan *Confusion Matrix* selanjutnya adalah tahapan *Classification Report* yang bertujuan untuk mengetahui hasil evaluasi pada *accuracy, precision, recall,* dan *f1-score*. Tahap ini dilakukan perhitungan yang diuji menggunakan parameter pada *Confusion Matrix*[3]

Metrik	Definisi
Precision	Rasio prediksi benar positif dibandingkan dengan keseluruhan hasil (Positif benar dan Positif Palsu).
Recall	Rasio prediksi benar positif dibandingkan dengan keseluruhan data dengan jumlah positif benar dan negatif palsu.
F1 Score	Rata-rata dari presisi dan recall. Semakin mendekati nilai skor F1 dengan 1,0 maka semakin baik kinerja model algoritma.
Support	Jumlah kejadian aktual dari kelas dalam dataset.
Accuracy	rasio prediksi benar (Positif dan Negatif) dengan keseluruhan data.

Tabel 2. *Classification Report* [5]

- ***Precision***

$$Precision = \frac{TP}{(FP+TP)} \quad [5]$$

- ***Recall***

$$Recall = \frac{TP}{(FN+TP)} \quad [5]$$

- ***F1 Score***

$$F1\ Score = 2 \times \frac{(Recall \times Precision)}{(Recall+Precision)} \quad [5]$$

- ***Accuracy***

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad [5]$$

3. Metode Penelitian

Metode penelitian ini termasuk ke dalam penelitian kuantitatif karena menggunakan data berupa angka untuk dianalisa hubungan variable yang diteliti berdasarkan dataset yang digunakan. pengumpulan data dari penelitian ini yaitu dengan cara mengunduh dataset dari portal website Open Data Jakarta (https://data.jakarta.go.id/). Data yang digunakan pada penelitian ini mengenai indeks standar pencemaran udara (ISPU) pada jangka waktu 2020 - 2021. Jumlah Data pada dataset ini yaitu sebanyak 670 data dan memiliki 11 variable.

Pada penelitian ini dimulai dengan melakukan pengumpulan data. Data yang digunakan pada penelitian ini mengenai indeks standar pencemaran udara (ISPU) pada jangka waktu 2020 - 2021. Pada penelitian ini menggunakan parameter pm10, pm25, so2, co, o3, no2. Terdapat 2 Parameter pendukung yang ditampilkan agar mempermudah pada tahap visualisasi data pada penelitian ini, yaitu categori dan location. Sebelum dilakukan tahap *klasifikasi* dilakukan *pre-processing* data terlebih dahulu.

A. Pengolahan Awal Data (Pre-Processing)

Data yang telah diperoleh perlu dilakukan pre-processing. Preprocessing data adalah hal yang harus

DOI: https://doi.org/10.29207/resti.v6iX.xxx

dilakukan dalam proses data, karena tidak semua data atau atribut data dalam data digunakan dalam proses data. Proses ini dilakukan agar data yang akan digunakan sesuai dengan kebutuhan. Terdapat beberapa jenis pre-processing yang akan digunakan dalam penelitian ini, antara lain:

1. Seleksi data

Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Seperti menentukan variabel apa saja yang dipakai untuk bisa dimulai ke tahap selanjutnya.

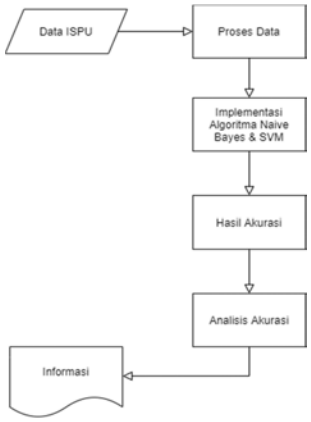
2. Data Cleaning

Pada tahap data cleaning dilakukan pembersihan data dari messing values dan duplikat yang ada pada data indeks standar pencemaran udara. Processing/Cleaning merupakan proses perbersihan yang membuang duplikat data, memerikas data yang inkosisten, dan memeperbaiki kesahan pada data.

3. Normalisasi

Proses menskalakan nilai data dalam rentang yang telah ditentukan sebelumnya. attribute selection merupakan prose yang menggunakan atribut untuk membuat data baru sehingga dapat mengatur kumpulan data dan membantu menganalisis data yang tersembunyi.

Setelah tahap *preprocessing* selanjutnya dilakukan klasifikasi menggunakan algoritma *Naïve Bayes* dan *SVM*. Data parameter polutan ini diklasifikasikan berdasarkan tingkat pencemaran udara menggunakan teknik *data mining*, sehingga dapat menghasilkan informasi dari data-data tersebut. Kemudian hasil klasifikasi yang didapatkan dari kedua algoritma tersebut dianalisis sesuai dengan hasil akurasi yang telah didapatkan. Berdasarkan analisis hasil dan pengujian data dengan menerapkan algoritma *Naïve Bayes* dan *SVM* diperoleh kesimpulan atau informasi dari penelitian yang telah dilakukan.



Gambar 1 Tahapan Penelitian

4. Hasil dan Pembahasan

Rangkaian hasil penelitian berdasarkan urutan/susunan logis untuk membentuk sebuah cerita. Isinya menunjukan fakta/data. Dapat menggunakan Tabel dan Angka tetapi tidak menguraikan secara berulang terhadap data yang sama dalam gambar, tabel dan teks. Untuk lebih memperjelas uraian, dapat mengunakan sub judul.

Pembahasan adalah penjelasan dasar, hubungan dan generalisasi yang ditunjukkan oleh hasil. Uraiannya menjawab pertanyaan penelitian. Jika ada hasil yang meragukan maka tampilkan secara objektif.

4.1. Dataset

Variabel pada dataset ini yaitu, tanggal = Tanggal pengukuran kualitas udara, stasiun = Lokasi pengukuran di stasiun, pm10 = Partikulat salah satu parameter yang diukur, pm25 = Partikulat salah satu parameter yang diukur, so2 = Sulfida (dalam bentuk SO2) salah satu parameter yang diukur, co = Carbon Monoksida salah satu parameter yand diukur, o3 = Ozon salah satu parameter yang diukur, no2 = Nitrogen dioksida salah satu parameter yang diukur, max = Nilai ukur paling tinggi dari seluruh parameter yang diukur dalam waktu yang sama, critical = Parameter yang hasil pengukurannya paling tinggi, categori = Kategori hasil perhitungan indeks standar pencemaran udara. Berikut tampilan dataset sebelum dan sesudah di normalisasi :

	tanggal	pm10	pm25	so2	co	o3	no2	max	critical	categori	location
0	1/1/2020	38	NaN	36.0	25.0	46	9	46	O3	BAIK	DKI5
1	1/2/2020	45	NaN	36.0	39.0	102	8	102	O3	TIDAK SEHAT	DKI5
2	1/3/2020	51	NaN	37.0	27.0	63	10	63	O3	SEDANG	DKI5
3	1/4/2020	51	NaN	38.0	19.0	85	10	85	O3	SEDANG	DKI5
4	1/5/2020	52	NaN	39.0	25.0	62	9	62	O3	SEDANG	DKI5
5	1/6/2020	62	NaN	37.0	39.0	64	9	64	O3	SEDANG	DKI5

Gambar 2. Dataset sebelum di normalisasi

	tanggal	pm10	pm25	so2	co	o3	no2	max	critical	categori	location
0	1/1/2020	38	98.216117	36.0	25.0	46	9	46	O3	BAIK	DKI5
1	1/2/2020	45	98.216117	36.0	39.0	102	8	102	O3	TIDAK SEHAT	DKI5
2	1/3/2020	51	98.216117	37.0	27.0	63	10	63	O3	SEDANG	DKI5
3	1/4/2020	51	98.216117	38.0	19.0	85	10	85	O3	SEDANG	DKI5
4	1/5/2020	52	98.216117	39.0	25.0	62	9	62	O3	SEDANG	DKI5
5	1/6/2020	62	98.216117	37.0	39.0	64	9	64	O3	SEDANG	DKI5

Gambar 3. Dataset sesudah di normalisasi

4.2. Tahapan Eksperimen

1. Pengumpulan Data

Pada penelitian ini dimulai dengan tahap pengumpulan data, data yang digunakan merupakan data Dinas Lingkungan Hidup Provinsi DKI Jakarta yang berasal dari website Open Data Jakarta

DOI: https://doi.org/10.29207/resti.v6iX.xxx

(https://data.jakarta.go.id/). Data yang digunakan pada penelitian ini mengenai indeks standar pencemaran udara (ISPU) pada jangka waktu 2020 - 2021. Jumlah Data pada dataset ini yaitu sebanyak 670 data dan memiliki 11 variable.

2. Preparing Data

Setelah data dikumpulkan tahap selanjutnya adalah preparing data atau mempersiapkan data yang ingin digunakan. Preparing data merupakan bagian dari Preprocessing data yang terdiri dari Preparing data dan Cleaning data. Pada tahap ini data yang sudah dikumpulkan disiapkan agar mempermudah dalam pengolahan data. Data ISPU yang mempunyai 670 record data. Data tersebut diseleksi yang nantinya akan masuk kedalam tahap Cleaning data.

3. Cleaning Data

Tahap cleaning data merupakan tahap yang penting, dimana tahap ini mempengaruhi hasil dari ouput process. Pada tahap ini dilakukan pembersihan terhadap data yang ingin digunakan, dari 4 atribut data dihilangkan menjadi 7 atribut data yang akan diimplementasikan menggunakan algoritma Support Vector Machine dan Naïve Bayes. Lalu data-data tersebut di cek apakah ada record data yang kosong, jika ada record data yang kosong maka baris data tersebut akan di dihilangkan. Pada tahap ini data dibersihkan sebaik mungkin agar saat pemrosesan data menghasilkan hasil yang maksimal.

4. Processing Data

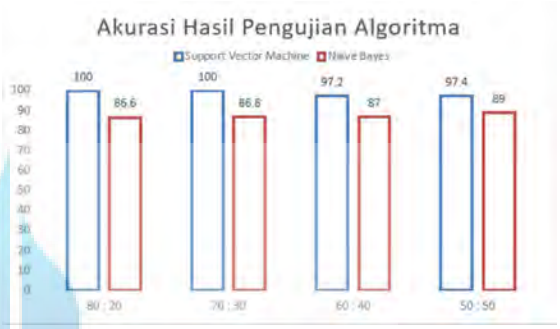
Setelah tahap cleaning data, tahap selanjutnya adalah processing data. Pada tahap ini data yang sudah siap digunakan di implementasikan kedalam algoritma klasifikasi. Algoritma klasifikasi pada tahap ini menggunakan algoritma Support Vector Machine dan Naïve Bayes dan di analisis serta divalidasi menggunakan K-Fold Cross Validation.

4.3. Hasil Eksperimen

Berikut ini merupakan hasil dari pengujian menggunakan algoritma Support Vector Machine dan Naïve Bayes menggunakan beberapa pembagian data. Hasil tersebut bisa dilihat pada gambar 4.

Perbandingan Data Latih (%)	Perbandingan Data Uji (%)	Label	Precision SVM (%)	Recall SVM (%)	F-1 Score SVM (%)	Accuracy	Precision Naive Bayes (%)	Recall Naive Bayes (%)	F-1 Score Naive Bayes (%)	Accuracy
80	20	Baik	100	100	100	100	40	100	57	97
		Sedang	100	100	100		91	87	89	
		Tidak Sehat	100	100	100		89	82	87	
70	30	Baik	88	100	93	100	33	100	50	87
		Sedang	100	99	100		91	87	89	
		Tidak Sehat	100	100	100		91	84	87	
60	40	Baik	64	100	78	97	39	100	56	83
		Sedang	97	98	97		92	88	90	
		Tidak Sehat	100	94	97		90	85	88	
50	50	Baik	69	100	82	97	41	100	58	89
		Sedang	98	98	98		93	89	91	
		Tidak Sehat	100	96	98		92	88	90	
Rata-Rata			92.58	98.75	95	98.2	73.3	91.16	77.8	87.5

Gambar 4. Hasil Pengujian Algoritma



Gambar 5. Grafik Akurasi Hasil Pengujian Algoritma

Berikut ini merupakan hasil dari validasi Support Vector Machine menggunakan 10-Cross Validation yang bertujuan untuk pengujian dan mengevaluasi kinerja algoritma bisa diliat pada tabel 6.

Support Vector Machine Cross Validation	
CV	Accuracy
1	100
2	93
3	100
4	100
5	86
6	100
7	93
8	100
9	100
10	100
Rata-rata	97.2

Gambar 6. SVM 10-Cross Validation

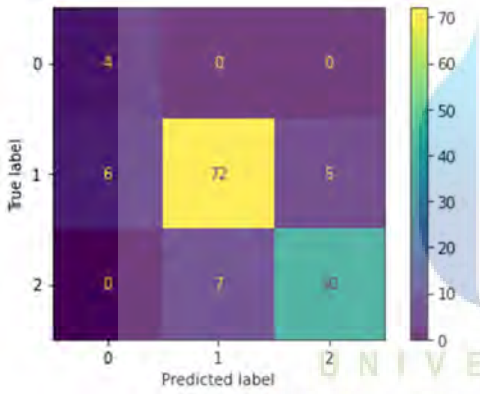
Berikut ini merupakan hasil dari validasi Naïve Bayes menggunakan 10-Cross Validation yang bertujuan untuk pengujian dan mengevaluasi kinerja algoritma bisa diliat pada tabel 7.

DOI: https://doi.org/10.29207/resti.v6iX.xxx

Naïve Bayes Cross Validation	
CV	Accuracy
1	88
2	92
3	95
4	91
5	61
6	64
7	91
8	88
9	98
10	92
Rata-rata	86

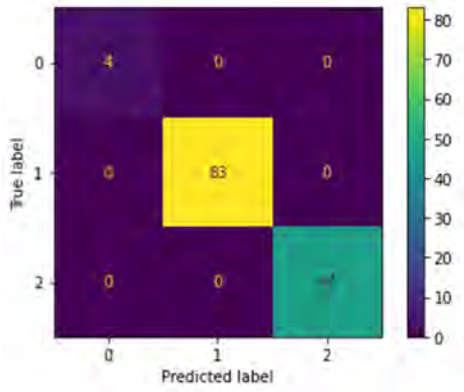
Gambar 7. Naïve Bayes 10-Cross Validation

Dari hasil kedua algoritma Support Vector Machine dan Naïve Bayes, selanjutnya dilakukan evaluasi menggunakan *Confusion Matrix* untuk mengukur performa menggunakan 4 kombinasi dari nilai prediksi dan nilai actual.



Gambar 8. Confusion Matrix Naïve Bayes

Confusion matrix algoritma Naïve Bayes memiliki *True positif* sebesar 116 yang artinya terdapat 116 data positif yang di prediksi benar positif, terdapat 18 *False positif* yang artinya terdapat 18 data negatif yang di prediksi sebagai positif (positif palsu), lalu terdapat 18 *False negatif* yang artinya sebanyak 18 data positif di prediksi sebagai negative (negative palsu).



Gambar 9. Confusion Matrix SVM

Confusion matrix algoritma Support Vector Machine memiliki *True positif* sebesar 134 yang artinya terdapat 134 data positif yang di prediksi benar positif, Tidak terdapat *False positif* yang artinya terdapat 0 data negatif yang di prediksi sebagai positif (positif palsu), lalu tidak terdapat *False negatif* yang artinya sebanyak 0 data positif di prediksi sebagai negative (negative palsu).

5. Kesimpulan

Pada tabel 6.1 dapat dilihat bahwa algoritma Support Vector Machine memiliki hasil yang lebih baik dengan tingkat akurasi rata-rata sebesar 98.5% dengan tingkat rata-rata *precision*, *recall* dan *f1-score* sebesar 92,58%, 98,75%, 95%. Sedangkan algoritma Naïve Bayes hanya memperoleh akurasi sebesar 87.75% dengan tingkat rata-rata precision, recall, dan f1-score sebesar 73.5%, 91,16%, 77,6%. Hasil yang dihasilkan oleh kedua algoritma juga di pengaruhi oleh besarnya jumlah data latih dan data uji bisa dilihat pada tabel 6.1.

Hasilnya rata-rata akurasi algoritma Support Vector Machine lebih baik dibandingkan dengan algoritma *Naïve Bayes,* namun jarak rata-rata akurasi algoritma Naïve Bayes lebih dekat dibandingkan menggunakan algoritma Support Vector Machine. Dalam penelitian ini diketahui bahwa metode Support Vector Machine dan Naïve Bayes cocok digunakan pada data yang jumlahnya banyak dan noise (berantakan) salah satunya adalah data ISPU, hal ini ditunjukan dengan nilai akurasi kedua algoritma yang cukup baik.

Hasil penelitian ini di harapkan dapat membantu dalam memprediksi kualitas udara di DKI Jakarta sesuai dengan kriteria Dinas Lingkungan Hidup Provinsi DKI Jakarta, dan diharapkan pada penelitian selanjutnya jumlah data dan variabel yang digunakan lebih banyak lagi agar informasi yang diberikan lebih akurat dan disarankan menggunakan metode lainnya

DOI: https://doi.org/10.29207/resti.v6iX.xxx

sehingga prediksi tingkat kualitas udara di DKI Jakarta lebihbaik.

Daftar Rujukan

[1] Amri Wicahyo , Ahmad Pudoli , Dewi Kusumaningsih , dan Noripansyah,"Penggunaan Algoritma Naive Bayes dalam klasifikasi Pengaruh Pencemaran Udara". *Jurnal ICT: Information Communication & Technology*., Vol. 20, N0.1, Juli 2021, pp. 103-108, p-ISSN: **2302-0261**, e-ISSN: **2303-3363**

[2] M. Ja'far Sodiq, Enny Itje Sela (2019). Perbandingan Metode Naïve Bayes dan K-Nearest Neighbor Pada Klasifikasi Kualitas Udara di DKI Jakarta. Journal of Computer Science, Diakses 13 September 2021, dari Universitas Teknologi Yogyakarta.

[3] D. Marutho, "Perbandingan Metode Naive Bayes , KNN , Decision Tree Pada Laporan Water Level Jakarta," *Manaj. Inform. AMIK JTC Semarang*, vol. 15, no. 2, pp. 90–97, 2019.

[4] Ghufran Rifaldi, M. H., & Budi Setiawan, E. (2019). "Competence Classification of Twitter Users Using Support Vector Machine (SVM) Method." *2019 7th International Conference on Information and Communication Technology (ICoICT).* doi:10.1109/icoict.2019.8835191 D. Marutho, "Perbandingan Metode Naive Bayes , KNN , Decision Tree Pada Laporan Water Level Jakarta," *Manaj. Inform. AMIK JTC Semarang*, vol. 15, no. 2, pp. 90–97, 2019.

[5] D. Pembimbing, R. P. K. S. T, M. T. Scjp, and D. S. Informasi, "Prediksi Diabetes Berdasarkan Faktor Risiko Behavioral Menggunakan Algoritma Support Vector Macihine Prediction of Diabetes Based on Behavioral Risk Factor Using Support," 2018.

DOI: https://doi.org/10.29207/resti.v6iX.xxx

KERTAS KERJA

Ringkasan

Kertas kerja ini merupakan material kelengkapan artikel jurnal dengan judul dengan judul "Analisa Algoritma Naïve Bayes Untuk Memprediksi Tingkat Pencemaran Udara Di Kota DKI Jakarta". Kertas kerja berisi semua material hasil penelitan Tugas Akhir yang tidak dimuat/atau disertakan di artikel jurnal. Di dalam kertas kerja ini disajikan: literature review, dataset yang digunakan, source code, dan hasil eksperimen secara keseluruhan.

Literatur review berisi tentang jurnal-jurnal pendukung yang terkait dengan penelitian analisa algoritma yang telah dibuat. Pada bab *Data Set* dilampirkan variabel-variabel yang digunakan sebagai landasan pengujian algoritma, Data tersebut terbagi menjadi dua yaitu data latih dan data uji.

Pada bab ini, dilampirkan pula sedikit sampel data yang digunakan dalam penelitian. Tahapan eksperimen merupakan bab yang membahas tentang eksperimen yang akan dilakukan untuk analisa kategori kualitas udara menggunakan metode *Support Vector Macihine* dan *Naïve Bayes*. Pada bab source code disajikan beberapa potongan code yang digunakan untuk melakuan pengujian algoritma dengan menggunakan Bahasa Python dan beberapa library pendukungnya. Kemudian yang terakhir yaitu bab hasil eksperimen secara keseluruhan membahas hasil yang didapat saat melakukan pengujian algoritma baik dari segi akurasi, presisi, *recall* dan *f1-score*. Selain itu, dijabarkan pula hasil dari perbandingan kedua algoritma yang digunakan tersebut dalam melakukan analisa kategori kualitas udara sampai pada penarikan kesimpulan algoritma terbaik.