PERBANDINGAN ALGORITMA NAÏVE BAYES DAN K-NEAREST NEIGHBOR (KNN) UNTUK KLASIFIKASI INDEKS STANDAR PENCEMARAN UDARA DI DKI JAKARTA TAHUN 2021

SKRIPSI

Diajukan sebagai salah satu persyaratan untuk memperoleh gelar sarjana statistika

Oleh
NURDALIA
18337004

PROGRAM STUDI SARJANA STATISTIKA
DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS NEGERI PADANG

2023

PERSETUJUAN SKRIPSI

PERBANDINGAN ALGORITMA NAÏVE BAYES DAN K-NEAREST NEIGHBOR (KNN) UNTUK KLASIFIKASI INDEKS STANDAR PENCEMARAN UDARA DI DKI JAKARTA TAHUN 2021

Nama	: Nurdalia
NIM	: 18337004
Program Studi	: S1 Statistika
Departemen	: Statistika
Fakultas	: Matematika dan Ilmu Pengetahuan Alam

Padang, 21 Februari 2023

Mengetahui:
Kepala Departemen Statistika

Dodi Vionanda, Ph.D
NIP : 197806112005011002

Disetujui Oleh:
Pembimbing

Zilrahmi, S.Pd., M.Si
NIP. 198911062019032009

PENGESAHAN LULUS UJIAN SKRIPSI

Nama	: Nurdalia
NIM	: 18337004
Program Studi	: S1 Statistika
Departemen	: Statistika
Fakultas	: Matematika dan Ilmu Pengetahuan Alam

PERBANDINGAN ALGORITMA NAÏVE BAYES DAN K-NEAREST NEIGHBOR (KNN) UNTUK KLASIFIKASI INDEKS STANDAR PENCEMARAN UDARA DI DKI JAKARTA TAHUN 2021

Dinyatakan lulus setelah dipertahankan di depan Tim Penguji Skripsi
Program Studi S1 Statistika Departemen Statistika Fakultas Matematika
dan Ilmu Pengetahuan Alam Universitas Negeri Padang

Padang, 21 Februari 2023

Tim Penguji

	Nama	Tanda Tangan
1. Ketua	: Zilrahmi, S.Pd, M.Si	1.
2. Anggota	: Dr. Dony Permana, M.Si	2.
3. Anggota	: Admi Salma, S.Pd, M.Si	3.

SURAT PERNYATAAN TIDAK PLAGIAT

Saya yang bertandatangan di bawah ini:

Nama	: Nurdalia
NIM	: 18337004
Program Studi	: S1 Statistika
Departemen	: Statistika
Fakultas	: Matematika dan Ilmu Pengetahuan Alam

Dengan ini menyatakan bahwa, skripsi saya dengan judul **"Perbandingan Algoritma Naive Bayes dan K-Nearest Neighbor (KNN) Untuk Klasifikasi Indeks Standar Pencemaran Udara di DKI Jakarta Tahun 2021"** adalah benar merupakan hasil karya saya dan bukan merupakan plagiat dari karya orang lain atau pengutipan dengan cara-cara yang tidak sesuai dengan etika yang berlaku dalam tradisi keilmuan.

Apabila suatu saat terbukti saya melakukan plagiat maka saya bersedia diproses dan menerima sanksi akademis maupun hukum sesuai dengan hukum dan ketentuan yang berlaku, baik di institusi UNP maupun di masyarakat dan negara.

Demikian pernyataan ini saya buat dengan penuh kesadaran rasa tanggung jawab sebagai anggota masyarakat ilmiah.

Diketahui oleh,
Ketua Departemen Statistika,

Dodi Vionanda, M.Si., Ph.D
NIP. 197906112005011002

Saya yang menyatakan,

Nurdalia
NIM. 18337004

PERBANDINGAN ALGORITMA *NAÏVE BAYES* DAN *K-NEAREST NEIGHBOR* (KNN) UNTUK KLASIFIKASI INDEKS STANDAR PENCEMARAN UDARA DI DKI JAKARTA TAHUN 2021

Nurdalia

ABSTRAK

Pencemaran udara merupakan masalah yang berdampak bagi kehidupan makhluk hidup. Udara yang tercemar akan menimbulkan berbagai macam penyakit, sehingga perlu dilakukan pengamatan tingkat pencemaran udara di lingkungan masyarakat. Untuk menentukan tingkat pencemaran udara dapat dipermudah dengan menggunakan proses klasifikasi data mining. Data mining adalah upaya untuk mendapatkan informasi dari sekumpulan data, yang nantinya informasi tersebut digunakan untuk pengambilan keputusan. Maka tujuan penelitian ini adalah untuk mengetahui model klasifikasi terbaik dalam memprediksi pencemaran udara di DKI Jakarta Tahun 2021, serta untuk mengetahui tingkat evaluasi terbaik yang dihasilkan oleh metode tersebut.

Penelitian ini merupakan penelitian terapan. Data yang digunakan adalah data indeks standar pencemaran udara DKI Jakarta Tahun 2021 yang diperoleh dari *website* kaggle berdasarkan enam kandungan udara, yaitu *partikel debu* (PM_{10}), *partikel debu* ($PM_{2,5}$), *sulfur dioksida* (SO_2), *karbon monoksida*. (CO), *ozon* (O_3), dan *nitrogen dioksida* (NO_2), dengan menggunakan metode *Naïve Bayes* dan *K-Nearest Neighbor*.

Hasil pengklasifikasian yang diperoleh dengan menggunakan algoritma *Naïve Bayes* dan *K-Nearest Neighbor* menunjukkan bahwa evaluasi terbaik dari kedua metode tersebut terdapat pada algoritma *Naïve Bayes* dengan tingkat *accuracy* yang tinggi sebesar 91%. Hal ini berlaku untuk keseluruhan kategori, meskipun terdapat data kategori tertentu yang memiliki frekuensi sedikit lebih kecil dibandingkan kategori lainnya.

Kata Kunci: *Naïve Bayes*, *K-Nearest Neighbor*, ISPU, *Confusion Matrix*

COMPARISON OF *NAÏVE BAYES* AND *K-NEAREST NEIGHBOR* (KNN) ALGORITHMS FOR AIR POLLUTION STANDARD INDEX CLASSIFICATION IN DKI JAKARTA 2021

Nurdalia

ABSTRACT

Air pollution is a problem that affects the lives of living things. Polluted air will cause various kinds of diseases, so it is necessary to observe the level of air pollution in the community environment. To determine the level of air pollution can be made easier by using a data mining classification process. Data mining is an attempt to obtain information from a set of data, which will later use this information for decision making. So the purpose of this research is to find out the best classification model for predicting air pollution in DKI Jakarta in 2021, as well as to find out the best level of evaluation produced by this method.

This research is applied research. The data used is DKI Jakarta air pollution standard index data for 2021 obtained from the kaggle website based on six air contents, namely dust particles (PM10), dust particles (PM2.5), sulfur dioxide (SO2), carbon monoxide. (CO), ozone (O3), and nitrogen dioxide (NO2), using the Naïve Bayes and K-Nearest Neighbor methods.

The classification results obtained using the Naïve Bayes and K-Nearest Neighbor algorithms show that the best evaluation of the two methods is found in the Naïve Bayes algorithm with a high accuracy rate of 91%. This applies to all categories, although there are data for certain categories that have a slightly lower frequency than other categories.

Keywords: *Naïve Bayes*, *K-Nearest Neighbor*, ISPU, *Confusion Matrix*

KATA PENGANTAR

Puji dan syukur penulis panjatkan kepada Allah Subhanaahu Wa Ta'ala atas berkat dan rahmat-Nya penulis dapat menyelesaikan Skripsi yang berjudul "Perbandingan Algoritma *Naïve Bayes* dan *K-Nearest Neighbor* Untuk Klasifikasi Indeks Standar Pencemaran Udara Di DKI Jakarta Tahun 2021". Skripsi ini disusun sebagai salah satu syarat untuk menyelesaikan Program Studi Statistika, Departemen Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Padang. Dalam menyelesaikan Skripsi ini penulis mendapatkan banyak bantuan dan dukungan dari berbagai pihak. Untuk itu penulis mengucapkan terima kasih kepada:

1. Bapak Dodi Vionanda, Ph.D. Kepala Departemen Statistika sekaligus Koordinator Program Studi S1 Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Padang.
2. Ibu Zilrahmi, S.Pd., M.Si. Dosen pembimbing Skripsi sekaligus Dosen Pembimbing Akademik yang telah memberikan arahan, bimbingan, dukungan dan motivasi dari awal sampai proses penyusunan Skripsi.
3. Bapak Dr. Dony Permana, M.Si dan Ibu Admi Salma, S.Pd., M.Si. Dosen Penguji Skripsi yang telah memberikan kontribusi terhadap Skripsi ini.
4. Bapak dan Ibu Dosen serta Staf Departemen Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Padang yang telah membantu penulis dalam menimba ilmu di Program Studi Statistika.
5. Terkhusus orang tua penulis, Bapak Bustami dan Ibu Halilah yang telah berjuang, mendukung secara materil dan non materil serta mendo'akan penulis tanpa henti dalam proses menimba ilmu pada Program Studi

Statistika, Jurusan Statistika, Fakultas Matematika dan Ilmu Pengetahuan, Universitas Negeri Padang.

6. Rekan-rekan Program Studi S1 Statistika yang telah berjuang bersama selama perkuliahan.

Semoga Skripsi ini memberikan manfaat untuk penulis sendiri, bermanfaat untuk semua pihak, dan bernilai ibadah di sisi Allah SWT. Skripsi ini tidak terlepas dari kesalahan dan kekeliruan, oleh karena itu penulis mengharapkan saran dan kritik yang bersifat membangun. Akhir kata penulis ucapkan terima kasih.

Padang, 6 Maret 2023

Penulis

DAFTAR ISI

DAFTAR GAMBAR

Gambar **Halaman**

DAFTAR TABEL

Tabel **Halaman**

DAFTAR LAMPIRAN

Lampiran **Halaman**

BAB I
PENDAHULUAN

A. Latar Belakang

Data mining adalah proses penggalian dan pencarian pengetahuan dan informasi yang bermanfaat dengan menggunakan algoritma atau metode tertentu sesuai dengan pengetahuan atau informasi (Buulolo, 2020:3). Informasi yang biasanya dikumpulkan adalah pola-pola tersembunyi pada data, hubungan antar elemen-elemen data, ataupun pembuatan model untuk keperluan peramalan data (Adinugroho, 2018:3).

Data mining dapat dikelompokkan menjadi dua yaitu metode deskriptif dan metode prediktif. Metode deskriptif bertujuan untuk menemukan pola, relasi dalam data. Contoh metode deskriptif yaitu klastering dan assosiasi Sedangkan metode prediktif bertujuan untuk memperkirakan nilai suatu variabel berdasarkan nilai variabel-variabel lainnya. Contoh metode prediktif yaitu klasifikasi dan regresi.

Klasifikasi merupakan metode pengelompokan data menjadi beberapa kategori dengan menggunakan bantuan *data testing* yang telah dikategorikan terlebih dahulu. Contoh algoritma klasifikasi yang digunakan adalah *Naïve Bayes, K-Nearest Neighbor* (KNN).

Naïve Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukan oleh ilmuan Inggris Thomas Bayes, yaitu memprediksi peluang dimasa depan berdasarkan pengalaman dimasa sebelumnya. Kelebihan *Naive Bayes* adalah mudah diimplementasikan dan

cepat. Adapun kelemahan *Naïve Bayes* adalah tidak berlaku apabila probabilitas kondisionalnya bernilai nol (Bustami, 2013).

KNN merupakan metode yang mengklasifikasikan suatu objek dengan mempertimbangkan kelas terdekat dari objek tersebut. KNN juga merupakan metode berbasis *Nearest Neighbor* (NN) yang paling tua dan paling popular (Prasetyo, 2014). KNN memiliki kelebihan yaitu bahwa efektif apabila *data testing* yang besar. Kekurangan metode KNN yaitu menentukan nilai parameter *K* (jumlah dari tetangga terdekat) (Siregar, 2020)

Pada Tahun 2014, Putri *et al* melakukan klasifikasi *Naïve Bayes* dan KNN pada analisis data status kerja di Kabupaten Demak. Dari penelitian ini, diketahui nilai akurasi pada klasifikasi *Naïve Bayes* yaitu sebesar 94%. Sedangkan nilai akurasi pada klasifikasi KNN dengan menggunakan nilai parameter *K* adalah 7 yaitu sebesar 96%. Selanjutnya Yusra *et al* (2016) membandingkan metode klasifikasi *Naïve Bayes* dan KNN pada data tugas akhir mahasiswa jurusan Teknik Informatika. Dari penelitian ini, diketahui nilai akurasi pada klasifikasi *Naïve Bayes* yaitu sebesar 87%. Sedangkan nilai akurasi pada klasifikasi KNN dengan menggunakan nilai parameter *K* adalah 5 yaitu sebesar 84%.

Berdasarkan penelitian sebelumnya, hanya melakukan evaluasi berdasarkan nilai *accurasy* saja tanpa melihat nilai evaluasi lainnya sebagai bahan pertimbangan untuk menentukan model algoritma terbaik, maka pada penelitian ini peneliti bukan hanya membandingkan nilai evaluasi *accurasy* tetapi juga membandingkan nilai evaluasi lainnya seperti *specificity* dan

sensitivity untuk mendapatkan algoritma terbaik terhadap indeks standar pencemaran udara (ISPU) DKI Jakarta Tahun 2021.

Kementerian Lingkungan Hidup dan Kehutanan (KLHK) mengatakan bahwa ISPU merupakan angka tanpa satuan, digunakan untuk menggambarkan kondisi mutu udara ambien di lokasi tertentu berdasarkan dampaknya terhadap kesehatan manusia dan makhluk hidup lainnya. Tujuan disusunnya ISPU agar memberikan kemudahan dari keseragaman informasi mutu udara ambien kepada masyarakat di lokasi dan waktu tertentu serta sebagai bahan pertimbangan dalam melakukan upaya pengendalian pencemaran udara baik bagi pemerintah pusat maupun daerah.

Pada Tahun 2020 KLHK telah mengeluarkan peraturan Nomor 14 Tahun 2020 tentang ISPU bahwa terdapat tujuh kandungan udara yaitu *Partikulat* (PM_{10}), *Partikulat* ($PM_{2,5}$), *Nitrogen Dioksida* (NO_2), *Sulfur Dioksida* (SO_2), *Karbon Monoksida* (CO), *Ozon* (O_3). Pengkategorian ISPU dapat dilihat di Tabel 1.

Tabel 1. Kategori Indeks Standar Pencemaran Udara

Rentang	Kategori
1-50	Baik
51-100	Sedang
101-200	Tidak Sehat
201-300	Sangat Tidak Sehat
≥301	Berbahaya

Pencemaran udara merupakan suatu masalah yang berdampak bagi kehidupan makhluk hidup. Udara yang tercemar akan menimbulkan berbagai macam penyakit, sehingga perlu dilakukan pengamatan tingkat pencemaran udara pada lingkungan masyarakat. Menurunnya kualitas udara di suatu daerah dapat ditentukan berdasarkan kandungan mineral pada udara

(Satra & Rachman, 2016). Menurut Apriawati dan Kiswandono (2017) kandungan udara yang dibutuhkan untuk dapat mendukung kehidupan manusia terdiri dari 78% *nitrogen*, 20% *oksigen*, 0,93% *argon*, 0,03%, *karbon dioksida* (CO_2). Apabila terjadi penambahan gas-gas lain yang menimbulkan gangguan serta perubahan maka dapat dikatakan udara sudah tercemar atau terpolusi.

Menurut Amalia *et al* (2022) pada kenyataannya, udara yang terdapat di alam tidak selalu bersih. Hal ini dapat menimbulkan penurunan kualitas udara. Salah satu Provinsi yang menyebabkan pencemaran udara adalah DKI Jakarta. DKI Jakarta merupakan penyumbang emisi udara dan penyebab menurunnya kualitas udara di Indonesia melalui kegiatan penduduk, kegiatan perindustrian dan transportasi. Berdasarkan pemantauan kualitas udara yang dilakukan oleh Amerika Serikat *Air Quality Index* (AQI) pada Juni 2022 DKI Jakarta menduduki ranking pertama kota paling berpolusi di Indonesia dan juga dunia dengan angka rentang mencapai 196 kategori tidak sehat.

Berdasarkan paparan yang telah dibahas, maka penulis tertarik melakukan penelitian yang berjudul **"Perbandingan Algoritma *Naïve Bayes* Dan *K-Nearest Neighbor* (KNN) untuk Klasifikasi Indeks Standar Pencemaran Udara di DKI Jakarta Tahun 2021.**

B. Batasan Masalah Penelitian

Adapun batasan masalah pada penelitian ini yaitu:

1. Data yang digunakan adalah data ISPU DKI Jakarta Tahun 2021.
2. Variabel yang digunakan dalam perhitungan ISPU berjumlah 7 yaitu: *Partikulat* (PM_{10}), *Partikulat* ($PM_{2,5}$), *Sulfur Dioksida* (SO_2), *Karbon Monoksida* (CO), *Ozon* (O_3), *Nitrogen Dioksida* (NO_2), Kategori Udara.

C. Rumusan Masalah Penelitian

Rumusan masalah pada penelitian ini adalah apa model klasifikasi terbaik untuk digunakan untuk mengklasifikasikan ISPU DKI Jakarta Tahun 2021?.

D. Tujuan Penelitian

Adapun tujuan dari penelitian ini mengetahui model klasifikasi yang terbaik dalam mengklasifikasikan ISPU DKI Jakarta Tahun 2021.

E. Manfaat Penelitian

Adapun manfaat yang diharapkan dari penelitian ini adalah sebagai berikut:

1. Bagi penulis, dapat menambah wawasan dan pengetahuan tentang algoritma *Naive Bayes* atau KNN.
2. Bagi instansi terkait, sebagai masukan dan pertimbangan dalam menentukan kebijakan kepada Dinas Lingkungan Hidup dan Kebersihan (DLHK) Provinsi DKI Jakarta sebagai informasi untuk mengetahui kualitas udara yang terjadi di lingkungan.
3. Bagi pembaca, dapat menjadi bahan referensi bagi penelitian selanjutnya.