Perbandingan antara Metode KNN dan Metode Naive Bayes dalam Memprediksi Kualitas Udara di Jakarta Tahun 2021

Angelica

2540133915

Statistics Department

*School of Computer Science*Jakarta, 11530  
[angelica008@binus.ac.id](mailto:angelica008@binus.ac.id)

Sherryl Kurniawan

2540119570

Statistics Department  
*School of Computer Science*Jakarta, 11530  
[sherryl.kurniawan@binus.ac.id](mailto:angelica008@binus.ac.id)Gabriel Matthew Mintana

2502006253

Statistics Department

*School of Computer Science*

Jakarta, 11530  
[gabriel.mintana001@binus.ac.id](mailto:hendi.sulistio@binus.ac.id)

***Abstract*— Kualitas udara merupakan suatu hal yang penting untuk kesehatan masyarakat Indonesia, namun kualitas udara di daerah Indonesia khususnya Jakarta mengalami penurunan kualitas udara yang dapat mengganggu kesehatan manusia. Penelitian ini dilakukan dengan tujuan untuk memprediksi tingkat kualitas udara di 5 lokasi (Bundaran HI, Kelapa Gading, Jagakarsa, Lubang Buaya, dan Kebon Jeruk) di Jakarta berdasarkan SPKU. Penelitian ini menggunakan metode KNN dan metode Naive Bayes untuk melakukan prediksi. Keakuratan data dalam pengujian model dalam metode KNN adalah 97% dan keakuratan data dalam pengujian model dalam metode Naive Bayes adalah 96%. Berdasarkan hasil perbandingan maka menunjukkan bahwa metode KNN memiliki hasil yang lebih baik dibandingkan metode Naive Bayes yang mengacu kepada 4 indikator pengujian yaitu *accuracy*, *precision*, *recall* dan f1-*score*.**

***Keywords—* pencemaran udara, KNN, naive bayes, jakarta**

# Introduction

Proses bercampurnya unsur berbahaya ke dalam atmosfer yang dapat menyebabkan terjadinya kerusakan lingkungan dan menurunkan kualitas lingkungan disebut sebagai pencemaran udara [1]. Penyebab pencemaran udara terbagi menjadi 2 yaitu pencemaran yang bersifat alami (berasal dari alam) dan pencemaran yang berasal dari kehidupan manusia sehari-hari misalnya asap dari kendaraan bermotor [1]. Pencemaran udara dapat menyebabkan dampak kesehatan pada manusia memburuk yang dapat mengakibatkan iritasi saluran pernafasan, iritasi mata, dan alergi kulit hingga timbulnya tumbuhan atau kanker paru [2]. Gangguan kesehatan ini dapat berpengaruh pada kehidupan manusia seperti menurunnya nilai produktivitas dan menyebabkan kerugian ekonomis dalam jangka panjang bahkan bisa menimbulkan permasalahan sosial ekonomi keluarga dan masyarakat [2].

Saat ini masalah pencemaran udara masih seringkali terjadi, namun tidak bisa kita pungkiri bahwa kita tidak bisa mengatasi masalah ini dengan mudah. Bahkan Indonesia menempati peringkat ke-17 sebagai negara paling berpolusi di dunia [3]. Polutan udara ini telah ditemukan menjadi faktor utama yang berkontribusi terhadap efek kesehatan manusia, seperti asma, stroke, penyakit jantung, dan paru-paru. Selama tahun 2021 tepatnya pada tahun kedua setelah pandemi COVID 19, banyak masyarakat yang harus *work from home* (bekerja dari rumah), dan pelajar harus belajar dari rumah. Pencemaran udara mulai berkurang dengan adanya COVID 19 dikarenakan berkurangnya transportasi di jalanan dan juga di tempat-tempat umum. Pencemaran udara terjadi di berbagai daerah di Indonesia termasuk Jakarta. Seperti yang kita ketahui keadaan di Industri di Jakarta setiap tahunnya terus meningkat sehingga kualitas udara di Jakarta dapat dikategorikan kurang baik. Untuk memprediksi kualitas udara tepatnya di tempat-tempat umum seperti di stasiun akan menggunakan metode klasifikasi. Dengan tujuan mengetahui kualitas udara di Jakarta baik di Jakarta Barat, Jakarta Utara, Jakarta Selatan, dan Jakarta Timur dengan menganalisis apa saja yang berpengaruh dalam tingkatan pencemaran udara yang diukur berdasarkan Stasiun Pemantau Kualitas Udara (SPKU) untuk mengategorikan apakah tingkatan udaranya baik, sedang ataupun tidak sehat.

Perbandingan metode KNN dan metode Naive Bayes sudah pernah dilakukan oleh peneliti-peneliti sebelumnya. Pada tahun 2022, Kirono dkk melakukan prediksi kualitas udara di Jakarta dari bulan Februari 2021-Oktober 2021 dengan metode naive bayes. Hasilnya menunjukkan nilai rata-rata *accuracy* 88%, *precision* 85%, *recall* 96%, f1-*score* 90%. Pada tahun 2023, Nurdalia dkk melakukan penelitian untuk klasifikasi Naive Bayes dan KNN untuk memprediksi pencemaran udara di Jakarta tahun 2021. Penelitian tersebut memperoleh hasil bahwa metode Naive Bayes lebih unggul dengan nilai *accuracy*, *specificity*, dan *sensitivity* berturut-turut adalah 92%, 94%, dan 94% dibandingkan metode KNN dengan nilai *accuracy*, *specificity*, dan *sensitivity* berturut-turut adalah 88%, 91%, dan 57%.

# KNN (K-Nearest Neighbors)

K-*Nearest* *Neighbors* atau yang biasa disebut KNN adalah salah satu metode *supervised learning* yang digunakan untuk mengklasifikasikan data. Algoritma KNN merupakan salah satu metode yang mudah untuk diimplementasikan untuk dataset yang jumlahnya kecil sementara jika dataset yang diolah cukup banyak dan kompleks, algoritma KNN ini tidak cocok digunakan karena waktu yang tidak efisien [4]. Algoritma KNN memiliki kelebihan yaitu dapat digunakan untuk data training yang memiliki banyak *noise* dan jumlah data yang besar. Salah satu kelemahan KNN adalah harus menentukan jumlah *nearest neighbors* dari target data yang disimbolkan dengan nilai k [5]. Namun karena menentukan jumlah yang paling terdekat, menentukan jenis jarak yang digunakan dan biaya komputasi yang tinggi, keakuratan data pada data *training* sangat tinggi [5].

K-*Nearest* *Neighbors* merupakan metode yang cukup sering digunakan oleh peneliti-peneliti sebelumnya dikarenakan teknik dari metode ini sangat sederhana, efektif dan efisien dalam bidang pengelolaan pola, kategori teks, pengolahan objek [6]. Dalam pemodelan, K-*Nearest* *Neighbors* memiliki beberapa tahapan [7], yaitu:

1. Menentukan jumlah tetangga terdekat (*nearest* *neighbors*)
2. Menghitung jarak data baru ke jumlah tetangga dengan berbagai penghitungan, salah satunya adalah *euclidean distance*.
3. Mengambil k tetangga terdekat untuk dibuat sebagai suatu keputusan dalam memprediksi berdasarkan hasil perhitungan jaraknya.

# Naive Bayes

Naive Bayes adalah metode klasifikasi probabilistik sederhana yang dimaksudkan untuk digunakan dengan asumsi bahwa variabel penjelas bersifat independen. Algoritma ini lebih menekankan pembelajaran dalam estimasi probabilistik [8].

Algoritma ini bekerja berdasarkan prinsip probabilitas bersyarat menurut teorema Bayes. Teorema Bayes menentukan probabilitas atau kemungkinan suatu peristiwa yang memberikan kemungkinan terjadinya peristiwa lain. Dalam istilah yang lebih sederhana, teorema Bayes adalah metode untuk menemukan probabilitas ketika kita mengetahui probabilitas tertentu lainnya.

Teorema Bayes dinyatakan secara matematis dalam persamaan berikut [9]:

Ada beberapa jenis algoritma ini, antara lain sebagai berikut [8]:

1. Bernoulli Naive Bayes

Di mana prediktor adalah variabel boolean. Oleh karena itu, satu-satunya yang ada adalah benar atau salah. Secara umum, algoritma ini digunakan ketika data sesuai dengan distribusi Bernoulli multivariat.

1. Naive Bayes Multinomial

Algoritma ini sering digunakan untuk menyelesaikan masalah klasifikasi dokumen. Misalnya, jika Anda ingin menentukan apakah suatu dokumen termasuk dalam suatu kategori, algoritma ini digunakan untuk menyortir. Naive Bayes menggunakan frekuensi kata saat ini sebagai fitur.

1. Gaussian Naive Bayes

Digunakan ketika prediktor tidak diskrit tetapi memiliki nilai kontinu dan prediktor diasumsikan berasal dari distribusi Gaussian.

Keunggulan utama Naive Bayes adalah modelnya sederhana namun dapat bersaing dengan model algoritmik lainnya. Implementasinya juga tidak terlalu rumit dan cocok untuk memperkirakan probabilitas bersyarat. Selain itu, pelaksanaannya sangat cepat dan ada pengulangan. Karena probabilitas dapat dihitung secara langsung. Jadi algoritma ini berguna saat Anda membutuhkan kecepatan pelatihan yang tinggi. Secara khusus, ketika asumsi independensi bersyarat benar, Anda dapat yakin bahwa itu akan memberikan hasil yang baik [10].

Sedangkan, kelemahan Naive Bayes menerapkan asumsi independensi bersyarat, sayangnya tidak selalu demikian. Dalam kebanyakan situasi, properti yang ada memiliki semacam ketergantungan. Kelemahan lain adalah masalah probabilitas. Naive Bayes memiliki masalah probabilitas nol, terutama ketika Anda menemukan kata-kata dalam data pengujian dari kelas tertentu yang tidak ada dalam data pelatihan. Anda mungkin akan berakhir dengan probabilitas "frekuensi nol". Namun, kemungkinan nol ini dapat diatasi dengan pemulusan (smoothing technique). Tambahkan faktor pemulusan ke pembilang dan penyebut dari setiap probabilitas sehingga nol tidak muncul [10].

# Evaluation Metrics

Pada tahap ini akan dilakukan pengukuran dan keakuratan kinerja model dengan membandingkan nilai sebenarnya (nilai aktual) dan nilai prediksi. Pada machine learning dalam kasus klasifikasi khususnya *supervised learning*, salah satu cara pengukuran kinerja model adalah dengan *confusion matrix* [11]. *Confusion matrix* terbagi menjadi 4 macam nilai yaitu *True Negative* (TN), *True Positive* (TP), *False Negative* (FN), dan *False Positive* (FP). Hal tersebut bisa digambarkan dalam bentuk tabel sebagai berikut ini.

Tabel 1. *Confusion Matrix* untuk Mengevaluasi Model

|  |  |  |  |
| --- | --- | --- | --- |
| Confusion Matrix | | Nilai Prediksi | |
| Positif | Negatif |
| Nilai Aktual | Positif | *True*  *Positive (TP)* | *False*  *Negative (FN)* |
| Negatif | *False*  *Positive (FP)* | *True*  *Negative (TN)* |

Keempat nilai tersebut dijelaskan antara lain sebagai berikut [12].

1. *True Negative* (TN) merupakan jumlah data yang nilai aktualnya negatif dan diprediksi pada model juga negatif.
2. *True Positive* (TP) merupakan jumlah data yang nilai aktualnya positif dan diprediksi pada model juga positif.
3. *False Negative* (FN) merupakan jumlah data yang nilai aktualnya positif dan diprediksi pada model negatif.
4. *False Positive* (FP) merupakan jumlah data yang nilai aktualnya negatif dan diprediksi pada model positif.

Berdasarkan data tersebut dapat diperoleh nilai *accuracy*, *precision*, *recall* dan f1-*score* [9].

1. Accuracy adalah nilai keseluruhan bagaimana model mengklasifikasikan dengan benar. Rumusnya dituliskan pada persamaan 1.

(1)

1. Precision adalah seberapa sering model prediksinya benar ketika model memprediksi positif. Rumusnya dituliskan pada persamaan 2.

(2)

1. Recall adalah seberapa sering model memprediksi positif ketika nilai aktualnya positif. Rumusnya dituliskan pada persamaan 3.

(3)

1. F1-Score adalah nilai rata-rata harmonik precision dan recall. Rumusnya dituliskan pada persamaan 4.

(4)

# Data and Research Methods

1. *Data*

Data yang akan diteliti dalam penelitian ini adalah pencemaran udara di beberapa stasiun di Jakarta. Data ini merupakan data sekunder dimana diambil dari sumber yang sudah ada yaitu melalui open data. Data ini berisi data indeks standar pencemaran udara di SPKU tahun 2021 di Jakarta. Jumlah data ini terdiri dari 1820 baris.

Dalam penelitian ini, data berisikan variabel kategori dan variabel numerik yaitu terdapat 3 variabel kategori dan 7 variabel numerik. Penelitian ini menggunakan 2 jenis variabel yaitu variabel dependen (variabel y) dan variabel independen (variabel x). Variabel dependen yang digunakan dalam kasus ini adalah categori. Variabel ini adalah variabel kategori yang terdiri dari baik, sedang dan tidak sehat. Selain itu variabel independennya dalam penelitian ini akan ada sebanyak 6 variabel yaitu PM10, PM25, SO2, CO, O3, NO2, max. Variabel tersebut akan dijelaskan dalam tabel 2.

Tabel 2. *Research Variable*

|  |  |  |
| --- | --- | --- |
| Jenis variabel | Nama Variabel | Deskripsi |
| Y | Categori | Tingkat kualitas udara: baik, sedang dan tidak sehat |
|  | PM10 | Perhitungan partikel udara yang lebih kecil dibandingkan 10 milimeter |
|  | PM25 | Perhitungan partikel udara yang lebih kecil dibandingkan 25 milimeter |
|  | SO2 | Nilai sulfida yang dihitung |
|  | CO2 | Kadar karbon monoksida yang dihitung |
|  | O3 | Perhitungan ozon |
|  | NO2 | Perhitungan nitrogen oksida |
|  | Max | Nilai perhitungan paling tinggi dari parameter yang dihitung |

1. *Research Methods*

Terdapat beberapa tahapan penelitian yang dilakukan yaitu *problem definition*, data *preprocessing*, *exploratory* data *analysis* (EDA), data *preparation*, KNN *modeling*, naive bayes *modeling*, dan *comparison* (perbandingan).

1. *Problem Definition*

Tujuan dari penelitian ini adalah untuk memprediksi dan menganalisis apa saja yang berpengaruh dalam tingkatan pencemaran udara berdasarkan standar Stasiun Pemantau Kualitas Udara (SPKU).

1. *Data Preprocessing*

*Preprocessing* adalah suatu tahapan proses untuk mengklasifikasikan dengan cara membersihkan, mengubah atau menghilangkan data yang tidak diperlukan [13]. Hal ini dilakukan agar data yang digunakan dapat bekerja secara optimal. Pertama-tama, melakukan pemeriksaan data untuk mengidentifikasi missing value. Jika terdapat missing value maka hapus missing value tersebut kemudian lakukan pemeriksaan kembali untuk memastikan bahwa missing value sudah terhapus. Setelah itu untuk mempermudah dalam proses pemodelan, klasifikasikan variabel kategori menjadi variabel numerik (0,1,2).

1. *Exploratory Data Analysis* (EDA)

*Exploratory Data Analysis* atau EDA adalah suatu proses yang dilakukan dengan tujuan untuk melakukan pengurangan dimensi data, memaksimalkan pengetahuan mengenai data, mengetahui variabel apa saja yang penting, mencari outlier dan anomali pada data [14]. Pada penelitian ini, dilakukan pemeriksaan *imbalance* data, memeriksa *outlier* dan pemeriksaan korelasi antar variabel.

1. *Data Preparation*

Data *Preparation* adalah sebuah tahapan yang dilakukan untuk menyeleksi data yang telah dikumpulkan dan diolah menjadi bentuk yang baru untuk digunakan dalam pemodelan [15]. Dalam penelitian ini, data dibagi menjadi dua yaitu data training dan data testing. Data *training* digunakan dalam pembuatan model sementara data *testing* digunakan untuk evaluasi model. Data dibagi dengan komposisi 80% data *training* dan 20% data *testing* dengan menggunakan *library* sklearn. Komposisi data ini dipilih dikarenakan menghasilkan *accuracy*, *precision*, *recall*, dan f1-*score* tertinggi.

1. KNN *Modeling*

Pada tahap ini dilakukan pemodelan menggunakan metode K-*Nearest* *Neighbors* (KNN). KNN digunakan untuk memprediksi variabel dependen kategorik. Dalam penelitian ini akan dilakukan pemodelan menggunakan *outlier* dan tanpa *outlier*. Tahapan yang dilakukan dalam analisis KNN baik menggunakan maupun tanpa *outlier* adalah sebagai berikut.

* Melakukan estimasi parameter data training dengan menggunakan *grid search*
* Membentuk model dengan hasil estimasi parameter terbaik
* Membuat prediksi baik menggunakan model data training dan data testing
* Membuat *confusion matrix* untuk data *training* dan data *testing*
* Mengevaluasi model dengan menampilkan nilai *accuracy*, *precision*, *recall* dan f1-*score*
* Membentuk model dengan estimasi parameter +1 dan -1 untuk melakukan pengecekan bahwa parameter yang digunakan merupakan parameter terbaik.
* Melakukan perbandingan model antara yang menggunakan *outlier* dan tanpa *outlier*

1. Naive Bayes Modeling

Pada tahap ini dilakukan pemodelan menggunakan metode Naive Bayes. Naive Bayes digunakan dalam memprediksi variabel dependen kategorik. Dalam penelitian ini akan dilakukan pemodelan menggunakan *outlier* dan tanpa *outlier*. Tahapan yang dilakukan dalam analisis Naive Bayes baik menggunakan maupun tanpa *outlier* adalah sebagai berikut.

* Melakukan estimasi parameter data training dengan menggunakan *grid search*
* Membentuk model berdasarkan hasil estimasi parameter terbaik
* Membuat prediksi baik menggunakan model data *training* dan data *testing*
* Membuat *confusion matrix* untuk data *training* dan data *testing*
* Mengevaluasi model dengan menampilkan nilai *accuracy*, *precision*, *recall* dan f1-*score*
* Membentuk model dengan estimasi parameter +1 dan -1 untuk melakukan pengecekan bahwa parameter yang digunakan merupakan parameter terbaik.
* Melakukan perbandingan model antara yang menggunakan *outlier* dan tanpa *outlier*

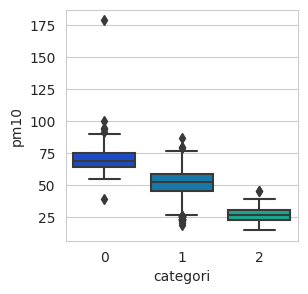
1. *Comparison*

Pada tahap ini yang akan dilakukan perbandingan dari hasil evaluasi tiap model dengan menggunakan nilai *accuracy*, *precision*, *recall* dan f1-*score* antara kedua model yaitu model KNN dan model Naive Bayes untuk memperoleh model terbaik.

# Results and Discussions

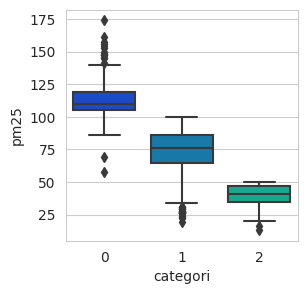
1. *Descriptive Analysis*

Bagian ini akan menjelaskan tentang deskripsi data penelitian untuk menggambarkan variabel apa saja yang dapat mempengaruhi pencemaran udara. Variabel numerik akan dideskripsikan dengan menggunakan *boxplot* agar dapat mengetahui *outlier* pada setiap variabel.



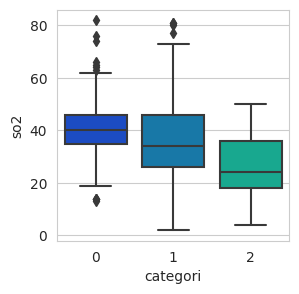
Gambar 1 Boxplot variabel PM10

Variabel PM10 menunjukkan perhitungan partikel udara lebih kecil dari 10 mm. Fig.1. terlihat untuk kategori 0, memiliki data *outlier* dibawah 45 dan diatas 90, sedangkan kategori 1, memiliki data *outlier* dibawah 25 dan diatas 75, dan terakhir untuk kategori 2, memiliki data *outlier* diatas 35.



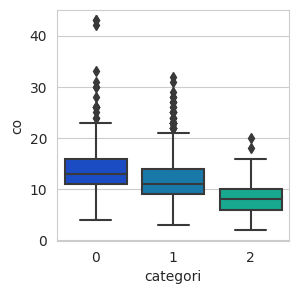
Gambar 2. Boxplot variabel PM25

Variabel PM25 menunjukkan perhitungan partikel udara lebih kecil dari 25 mm. Fig.2. terlihat untuk kategori 0, memiliki data *outlier* dibawah 80 dan diatas 140, sedangkan untuk kategori 1, hanya memiliki data *outlier* dibawah 30, dan terakhir untuk kategori 2, hanya memiliki data *outlier* di bawah 23.



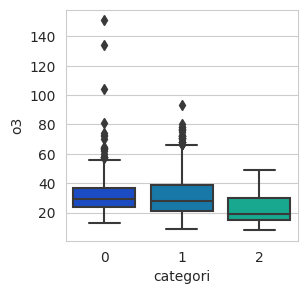
Gambar 3 Boxplot variabel SO2

Variabel SO2 menunjukkan perhitungan partikel udara yang mengandung kadar sulfida. Fig.3. terlihat untuk kategori 0, memiliki data *outlier* di bawah 19 dan diatas 62, sedangkan untuk kategori 1, hanya memiliki data *outlier* diatas 75, dan terakhir untuk kategori 2, tidak memiliki data *outlier*.



Gambar 4 Boxplot variabel CO

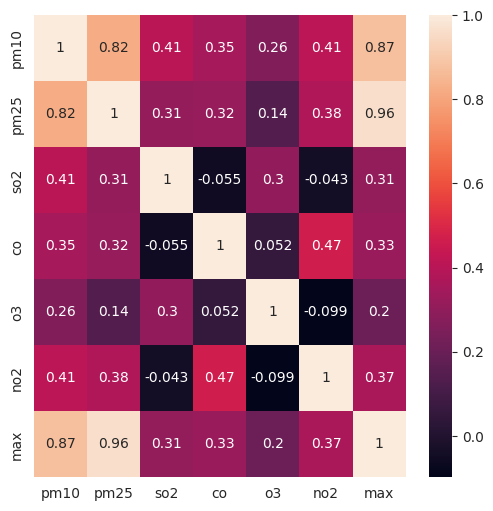
Variabel CO menunjukkan perhitungan partikel udara yang mengandung kadar karbon monoksida. Fig.4. terlihat untuk kategori 0, hanya memiliki data *outlier* diatas 25, sedangkan untuk kategori 1, hanya memiliki data *outlier* diatas 21, dan terakhir untuk kategori 2, hanya memiliki data *outlier* diatas 16.



Gambar 5 Boxplot variabel O3

Variabel O3 menunjukkan perhitungan ozon di udara. Fig.5. terlihat untuk kategori 0, memiliki data *outlier* datas 58, sedangkan untuk kategori 1, hanya memiliki data *outlier* diatas 63, dan terakhir untuk kategori 2, tidak memiliki data *outlier*.

Selain menampilkan *outlier*, bagian ini juga akan menampilkan hasil pengujian korelasi antar variabel. Untuk variabel numerik dengan variabel numerik menggunakan heatmap korelasi. Terdapat beberapa cara yang dapat digunakan untuk melihat korelasi antar variabel. Salah satu caranya adalah dengan menggunakan heatmap korelasi karena cara ini cukup efektif dalam memvisualisasikan hubungan antara variabel satu dengan variabel lainnya dalam suatu dataset [16].



Gambar 6 Heatmap Korelasi

Dari hasil koefisien korelasi menunjukkan bahwa korelasi yang paling kuat adalah antara variabel NO3 dan O3. Namun korelasinya adalah korelasi negatif. Hal ini menunjukkan adanya hubungan negatif yang sangat kuat antara variabel NO2 dengan variabel O3 dalam analisis. Selain itu juga terdapat korelasi variabel PM25 dengan variabel max yang nilainya 0.96, maka korelasinya sangat kuat dan bersifat positif.

Selain itu juga dilakukan pengujian korelasi untuk variabel kategori dengan variabel kategori yang menggunakan *chi*-*square*. Hipotesis yang diuji dengan uji *chi*-*square* adalah sebagai berikut.

tidak ada hubungan antara stasiun dan critical

ada hubungan antara stasiun dan critical

Tolak jika nilai *p*-*value* lebih kecil dari

Uji *chi*-*square* tersebut memperoleh nilai *p*-*value* sebesar . Dikarenakan lebih kecil dari 0.05 maka tolak . Sehingga dapat disimpulkan bahwa terdapat hubungan atau korelasi antara variabel stasiun dan variabel *critical*.

1. *K-Nearest Neighbors (KNN) Model*

Membuat pemodelan menggunakan metode K-*Nearest* *Neighbors* (KNN) di awali dengan pemilihan nilai N, sebagai jarak terdekat dengan tetangganya untuk menentukan kelayakan suatu model.

Untuk mendapatkan nilai N terbaik, maka menggunakan salah satu metode *tuning parameter* yaitu *grid search*. *Grid Search* CV melakukan banyak percobaan untuk setiap kombinasi dari nilai *hyperparameter* sampai mendapatkan nilai *hyperparameter* yang terbaik untuk model [17].

Tabel 3. Tuning Parameter KNN

|  |  |  |
| --- | --- | --- |
| Parameter *range* n-*neighbors* | *Cross validation* | *Best parameter* (*Result*) |
| 1-50 | 4 | 7 |

Melakukan pemodelan sebanyak 3 kali pada sampel data yang memiliki data outlier dengan N = 6, 7, 8 untuk memastikan bahwa n = 7 adalah parameter terbaik. Pemodelan dibagi menjadi data *testing* dan data *training*, dengan data *testing* sebesar 20% dan data *training* sebesar 80%.

Melakukan perhitungan prediksi data dengan metode KNN untuk N = 6.

Tabel 4A. *Confusion Matrix data testing* (N = 6)

|  |  |  |
| --- | --- | --- |
| 46 | 3 | 0 |
| 3 | 225 | 2 |
| 0 | 1 | 24 |

Tabel 4B. *Classification Report data testing* (N = 6)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** | **support** |
| 0 | 0.94 | 0.94 | 0.94 | 49 |
| 1 | 0.98 | 0.98 | 0.98 | 230 |
| 2 | 0.92 | 0.96 | 0.94 | 25 |
| Accuracy |  |  | 0.97 | 304 |
| Macro avg | 0.95 | 0.96 | 0.95 | 304 |
| Weighted avg | 0.97 | 0.97 | 0.97 | 304 |

Melakukan perhitungan prediksi data dengan metode KNN untuk N = 7.

Tabel 5A. *Confusion Matrix data testing* (N = 7)

|  |  |  |
| --- | --- | --- |
| 44 | 5 | 0 |
| 1 | 227 | 2 |
| 0 | 0 | 25 |

Tabel 5B. *Classification Report data testing* (N = 7)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** | **support** |
| 0 | 0.98 | 0.90 | 0.94 | 49 |
| 1 | 0.98 | 0.99 | 0.98 | 230 |
| 2 | 0.93 | 1 | 0.96 | 25 |
| Accuracy |  |  | 0.97 | 304 |
| Macro avg | 0.96 | 0.96 | 0.96 | 304 |
| Weighted avg | 0.97 | 0.97 | 0.97 | 304 |

Melakukan perhitungan prediksi data dengan metode KNN untuk N = 8.

Tabel 6A. *Confusion Matrix data testing* (N = 8)

|  |  |  |
| --- | --- | --- |
| 47 | 2 | 0 |
| 4 | 224 | 2 |
| 0 | 1 | 24 |

Tabel 6B. *Classification Report data testing* (N = 8)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** | **Support** |
| 0 | 0.92 | 0.96 | 0.94 | 49 |
| 1 | 0.99 | 0.97 | 0.98 | 230 |
| 2 | 0.92 | 0.96 | 0.94 | 25 |
| Accuracy |  |  | 0.97 | 304 |
| Macro avg | 0.94 | 0.96 | 0.95 | 304 |
| Weighted avg | 0.97 | 0.97 | 0.97 | 304 |

Berdasarkan hasil tersebut akan dilakukan perbandingan untuk membuktikan bahwa N = 7 merupakan nilai N terbaik untuk pemodelan dengan metode KNN.

Tabel 7. Perbandingan N parameter

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | N = 6 | | **N = 7** | |
| Training | Testing | **Training** | **Testing** |
| Accuracy | 0.98 | 0.97 | **0.99** | **0.97** |
| Precision | 0.98 | 0.95 | **0.99** | **0.96** |
| Recall | 0.95 | 0.96 | **0.97** | **0.96** |
| F1-Score | 0.96 | 0.95 | **0.98** | **0.96** |

|  |  |  |
| --- | --- | --- |
|  | N = 8 | |
| Training | Testing |
| Accuracy | 0.98 | 0.97 |
| Precision | 0.98 | 0.94 |
| Recall | 0.95 | 0.96 |
| F1-Score | 0.97 | 0.95 |

Dilakukan pengujian model dengan menghilangkan data *outlier*, lakukan kembali prediksi data dengan metode KNN untuk N = 7, lalu membandingkan metode KNN untuk N=7 dengan data tanpa *outlier* dan data memiliki *outlier*.

Tabel 8. Perbandingan KNN (N = 7)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Data Outlier** | | Tanpa Outlier | |
| **Training** | **Testing** | Training | Testing |
| Accuracy | **0.99** | **0.97** | 0.98 | 0.97 |
| Precision | **0.99** | **0.96** | 0.98 | 0.95 |
| Recall | **0.97** | **0.96** | 0.94 | 0.97 |
| F1-Score | **0.98** | **0.96** | 0.96 | 0.96 |

Berdasarkan hasil perbandingan tersebut dapat disimpulkan bahwa pemodelan dengan metode KNN menggunakan *outlier* lebih baik dibandingkan pemodelan tanpa *outlier*, sehingga dapat disimpulkan *outlier* berpengaruh pada pemodelan KNN.

1. *Naive Bayes Model*

Pemodelan Naive Bayes dilakukan dengan menggunakan 7 variabel independen. Untuk mendapatkan nilai parameter terbaik, maka menggunakan salah satu metode *tuning parameter* yaitu *grid search*.

Tabel 9. *Tuning Parameter* Naive Bayes

|  |  |  |
| --- | --- | --- |
| Parameter Range Var Smoothing | Cross Validation | Best Parameter (Result) |
| 1e-9 hingga 1 | 4 | 0.01 |

Melakukan perhitungan prediksi data dengan metode Naive Bayes, dengan parameter sebesar 0.1

Tabel 10A. *Confusion Matrix data testing*

(*var\_smoothing* = 0.1)

|  |  |  |
| --- | --- | --- |
| 42 | 7 | 0 |
| 1 | 226 | 3 |
| 0 | 2 | 23 |

Tabel 10B. *Classification Report data testing*

(*var\_smoothing* = 0.1)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** | **Support** |
| 0 | 0.98 | 0.86 | 0.91 | 49 |
| 1 | 0.96 | 0.98 | 0.97 | 230 |
| 2 | 0.88 | 0.92 | 0.90 | 25 |
| Accuracy |  |  | 0.96 | 304 |
| Macro avg | 0.94 | 0.92 | 0.93 | 304 |
| Weighted avg | 0.96 | 0.96 | 0.96 | 304 |

Melakukan perhitungan prediksi data dengan metode Naive Bayes, dengan parameter sebesar 0.01

Tabel 11A. *Confusion Matrix data testing*

(*var\_smoothing* = 0.01)

|  |  |  |
| --- | --- | --- |
| 45 | 4 | 0 |
| 2 | 224 | 4 |
| 0 | 2 | 23 |

Tabel 11B. Classification Report data testing

(var\_smoothing = 0.01)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** | **Support** |
| 0 | 0.96 | 0.92 | 0.94 | 49 |
| 1 | 0.97 | 0.97 | 0.97 | 230 |
| 2 | 0.85 | 0.92 | 0.88 | 25 |
| Accuracy |  |  | 0.96 | 304 |
| Macro avg | 0.93 | 0.94 | 0.93 | 304 |
| Weighted avg | 0.96 | 0.96 | 0.96 | 304 |

Melakukan perhitungan prediksi data dengan metode Naive Bayes, dengan parameter sebesar 0.001

Tabel 12A. *Confusion Matrix data testing*

(*var\_smoothing* = 0.001)

|  |  |  |
| --- | --- | --- |
| 48 | 1 | 0 |
| 11 | 208 | 11 |
| 0 | 2 | 23 |

Tabel 12B. *Classification Report data testing*

(*var\_smoothing* = 0.001)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** | **Support** |
| 0 | 0.81 | 0.98 | 0.89 | 49 |
| 1 | 0.99 | 0.90 | 0.94 | 230 |
| 2 | 0.68 | 0.92 | 0.78 | 25 |
| Accuracy |  |  | 0.92 | 304 |
| Macro avg | 0.83 | 0.93 | 0.87 | 304 |
| Weighted avg | 0.93 | 0.92 | 0.92 | 304 |

Berdasarkan hasil tersebut akan dilakukan perbandingan untuk membuktikan bahwa *var\_smoothing* = 0.01 merupakan nilai parameter terbaik untuk pemodelan menggunakan metode Naive Bayes.

Tabel 13. Perbandingan parameter *var\_smoothing*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 0.1 | | **0.01** | |
| Training | Testing | **Training** | **Testing** |
| Accuracy | 0.94 | 0.96 | **0.94** | **0.96** |
| Precision | 0.91 | 0.94 | **0.89** | **0.93** |
| Recall | 0.90 | 0.92 | **0.92** | **0.94** |
| F1-Score | 0.90 | 0.93 | **0.90** | **0.93** |

|  |  |  |
| --- | --- | --- |
|  | 0.001 | |
| Training | Testing |
| Accuracy | 0.94 | 0.95 |
| Precision | 0.89 | 0.91 |
| Recall | 0.92 | 0.93 |
| F1-Score | 0.90 | 0.92 |

Dikarenakan nilai accuracy antara parameter *var\_smoothing* = 0.1 dan parameter *var\_smoothing* = 0.01 memiliki kesamaan nilai, maka untuk memastikan bahwa *var\_smoothing* = 0.01 adalah yang terbaik, dilakukan penghitungan *accuracy* secara manual menggunakan persamaan (1) sebagai berikut.

* **Var\_smoothing = 0.1**

Untuk data *training*:1140/1213 = 0.9401

Untuk data *testing*: 291/304 = 0.9572

* **Var\_smoothing = 0.01**

Untuk data *training*: 1141/1213 = 0.9409

Untuk data *testing*: 292/304 = 0.9605

Berdasarkan penghitungan tersebut, maka diperoleh hasil bahwa model terbaik menggunakan parameter *var\_smoothing* = 0.01. Dilakukan pengujian model dengan menghilangkan data *outlier*, lakukan kembali prediksi data dengan metode Naive Bayes, lalu membandingkan metode Naive Bayes dengan data tanpa outlier dan data memiliki outlier.

Tabel 14. Perbandingan Naive Bayes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Data Outlier** | | Tanpa Outlier | |
| **Training** | **Testing** | Training | Testing |
| Accuracy | **0.94** | **0.96** | 0.94 | 0.96 |
| Precision | **0.89** | **0.93** | 0.89 | 0.93 |
| Recall | **0.92** | **0.94** | 0.92 | 0.94 |
| F1-Score | **0.90** | **0.93** | 0.90 | 0.93 |

Tabel 14. Menunjukkan bahwa baik data outlier maupun data outlier memiliki nilai *accuracy, precision, recall,* dan *F1-Score* yang sama. Maka dapat dikatakan bahwa outlier pada pemodelan Naive Bayes tidak berpengaruh. Oleh karena itu, pemodelan Naive Bayes akan menggunakan data yang memiliki *outlier* karena pada pemodelan KNN juga menggunakan data yang memiliki *outlier*.

1. *Perbandingan Model KNN dan Naive Bayes*

Model terbaik merupakan model yang memiliki nilai yang tinggi pada *accuracy, precision, recall,* dan *F1-Score.* Oleh karena itu, akan dilakukan perbandingan antara kedua metode untuk mengevaluasi model terbaik.

Tabel 15. Perbandingan metode KNN dan Naive Bayes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **KNN** | | Naive Bayes | |
| **Training** | **Testing** | Training | Testing |
| Accuracy | **0.99** | **0.97** | 0.94 | 0.96 |
| Precision | **0.99** | **0.96** | 0.89 | 0.93 |
| Recall | **0.97** | **0.96** | 0.92 | 0.94 |
| F1-Score | **0.98** | **0.96** | 0.90 | 0.93 |

Tabel 15 menunjukkan bahwa pemodelan dengan menggunakan metode KNN menghasilkan nilai *accuracy, precision, recall* dan f1-*score* lebih tinggi dibandingkan dengan pemodelan menggunakan metode Naive Bayes.

# Conclusion

Berdasarkan hasil pengujian kedua model yang telah dilakukan, dapat disimpulkan bahwa model KNN adalah model terbaik dengan parameter N = 7 yang memiliki nilai *accuracy*, *precision*, *recall* dan f1-*score* tertinggi yaitu masing-masing bernilai 97%, 96%, 96%, dan 96%. Sementara itu, untuk pemodelan Naive Bayes masing-masing memiliki nilai *accuracy, precision, recall* dan f1-*score* dengan nilai 96%, 93%, 94%, dan 93%.

##### References

1. Simandjuntak, A. G. (2007). Pencemaran Udara. Buletin Limbah. <https://jurnal.batan.go.id/index.php/bl/article/view/785>
2. Budiyono, A. (2001). Pencemaran Udara: Dampak Pencemaran Udara pada Lingkungan. Berita Dirgantara Vol.2(1). <https://jurnal.lapan.go.id/index.php/berita_dirgantara/article/viewFile/687/605>
3. Fajrian, H. (2022). RI Tempati Peringkat 17 Negara Dengan Polusi Udara Terburuk di Dunia. Katadata. <https://katadata.co.id/happyfajrian/berita/62396fa3bc821/ri-tempati-peringkat-17-negara-dengan-polusi-udara-terburuk-di-dunia>
4. Yahya, & Hidayanti, W. P. (2020). Penerapan Algoritma K-Nearest Neigbor Untuk Klasifikasi Efektivitas Penjualan Vape (Rokok Elektrik) pada Lombok Vape On. Jurnal Informatika dan Teknologi. Vol 3 No.2. 110-114. <https://e-journal.hamzanwadi.ac.id/index.php/infotek/article/view/2279>
5. Saifur Rohman Cholil, dkk. (2020). Implementasi Algoritma Klasifikasi K-Nearest Neighbor (KNN) Untuk Klasifikasi Seleksi Penerima Beasiswa. <https://ejournal.bsi.ac.id/ejurnal/index.php/ijcit/article/view/10438>
6. Bhatia, M., Vandana., 2010. Survey of nearest neighbor techniques. International Journal of Computer Science and Information Security 8, 1947-5500
7. Said, H., Matondang, N.H., & Irmanda H.N. (2022). Penerapan Algoritma K-Nearest Neighbor Untuk Memprediksi Kualitas Air yang Dapat Dikonsumsi. Jurnal Teknologi Informasi Techno.Com. <http://publikasi.dinus.ac.id/index.php/technoc/article/view/5901/2927>
8. Hans, R. (2022, May 23). Mengenal naive Bayes Sebagai Salah Satu Algoritma data science. <https://dqlab.id/mengenal-naive-bayes-sebagai-salah-satu-algoritma-data-science>
9. Trivusi. (2022). Pengertian Dan Contoh Algoritma naive Bayes classifier. Trivusi. <https://www.trivusi.web.id/2022/07/algoritma-naive-bayes.html>
10. Laraswati, B. D. (2022, July 21). Mengenal Kelemahan Dan Kelebihan naive Bayes. Algoritma Data Science School.
11. Nugroho, Kuncahyo Setyo, 2019, Confusion Matrix untuk Evaluasi Model pada Supervised Learning. <https://ksnugroho.medium.com/confusion-matrix-untuk-evaluasi-model-padaunsupervised-machine-learning-bc4b1ae9ae3f>
12. Saputro, I. W., & Sari, B. W. (2019). Uji Performa Algoritma Naive Bayes untuk Prediksi Masa Studi Mahasiswa. Citec Journal, vol. 6, no. 1. <https://ejurnal.tunasbangsa.ac.id/index.php/jsakti/article/view/369>
13. Muttaqin, F.A., & Bachtiar, A.M. (2013). Implementasi Teks Mining pada Aplikasi Pengawasan Penggunaan Internet Anak “Dodo Kids Browser”. Jurnal Ilmiah Komputer dan Informatika. <https://elib.unikom.ac.id/files/disk1/672/jbptunikompp-gdl-firdausakh-33590-8-unikom_f-a.pdf>
14. Wahyuni, E.D., Arifiyanti, A.A., & Kustyani, M. (2019). Exploratory Data Analysis dalam Konteks Klasifikasi Data Mining. <https://journal.itny.ac.id/index.php/ReTII/article/view/1530/908>
15. Kennardi., & Budi, A. (2015). Penerapan Data Mining untuk Prediksi Nilai Tukar Petani Tanaman Pangan di Indonesia dengan Metode Linear Regression dan Support Vector Machine. <https://jurnal.kwikkiangie.ac.id/index.php/JIB/article/view/505/277>
16. Leni, D., dkk. (2023). Analisis Heatmap Korelasi dan Scatterplot untuk Mengidentifikasi Faktor-Faktor yang Mempengaruhi Pelabelan AC efisiensi Energi. Jurnal Rekayasa Material, Manufaktur dan Energi, vol.6, no.1. <https://jurnal.umsu.ac.id/index.php/RMME/article/view/13133>
17. Fitriani, E.E, & Yustanti, W. (2022). Perbandingan Kinerja Metode Problem Transformation-KNN dan Algorithm Adaptation-KNN pada Klasifikasi Multi-Label Pertanyaan Kotakode. Journal of Emerging Information Systems and Business Intelligence, vol.3, no.3.<https://ejournal.unesa.ac.id/index.php/JEISBI/article/view/47510/39666>
18. Kirono, A.A.H., dkk. (2022). Klasifikasi Tingkat Kualitas Udara Dki Jakarta Menggunakan Algoritma Naive Bayes. <https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/18002>
19. Nurdalia, dkk. (2023). Comparison Between Naïve Bayes and K-Nearest Neighbor for DKI Jakarta Air Pollution Standard Index Classification. <https://ujsds.ppj.unp.ac.id/index.php/ujsds/article/download/29/23>