JNATIA Volume 1, Nomor 1, November 2022
Jurnal Nasional Teknologi Informasi dan Aplikasinya

Implementasi Metode K-Nearest Neighbors Pada Sistem Klasifikasi Kualitas Udara Berdasarkan Partikulat Berbahaya Yang Terkandung

I Komang Roni Sudarmawan[a1], I Gusti Agung Gede Arya Kadyanan[a2], I Wayan Supriana[a3]

[a]Informatics Department, Udayana University
South Kuta, Badung, Bali, Indonesia
[1]ronnysudarmawan17@gmail.com
[2]gungde@unud.ac.id
[3]wayan.supriana@unud.ac.id

Abstract

Air pollution is one of the most dangerous issues for health in the current industrial 4.0 era. Air quality, especially in big cities such as Jakarta, which is the capital of Indonesia as well as the center of government, is an issue that is being eradicated by the government. With a very high percentage of the productive age of the population in the city of Jakarta, motor vehicle activity in the city of Jakarta will also be equivalent to the productive age of the population of the city of Jakarta. This study will design an air quality classification system using the K-Nearest Neighbors method. Before being classified, a data preprocessing process will be carried out, such as handling missing values and handling outliers. In addition, because the data obtained is quite small, a K-Fold Validation process will be carried out for the model selection process. Finally, the performance evaluation of the model will be carried out using the confusion matrix method.

Keywords: *K-Nearest Neighbors, Pre-Processing, Missing Value, Outlier, K-fold Validation, Confusion Matrix*

1. Introduction

Kota Jakarta merupakan pusat pemerintahan negara Indonesia dengan jumlah penduduk mencapai 10,56 juta dan tingkat kepadatan penduduk mencapai 16.937 jiwa/km2. Dengan data dari Direktorat Jenderal Kependudukan dan Pencatatan Sipil (Dukcapil) Kementerian Dalam Negeri menyebutkan bahwa 70% dari penduduk kota Jakarta merupakan penduduk usia produktif, hal tersebut menandakan bahwa aktivitas seperti kendaraan bermotor maupun kendaraan umum di Jakarta sangat tinggi [1]. Hal tersebut menyebabkan tingkat pencemaran udara di kota Jakarta menjadi sangat tinggi. Berdasarkan Situs IQAir, konsentrasi polutan PM 2,5 di Jakarta mencapai 96 µg/m³. Jumlah konsentrasi PM 2.5 tersebut 4 kali lipat di atas ambang panduan Badan Kesehatan Dunia (*World Health Organization*/WHO). Hal tersebut menandakan bahwa kualitas udara di kota Jakarta kurang layak untuk dihirup.

Dengan pencemaran udara yang semakin meningkat di kota Jakarta, maka pada penelitian ini akan dilakukan teknik data mining untuk dapat mengklasifikasikan kualitas udara dari beberapa parameter indeks standar pencemaran udara (ISPU) yang diukur dari 5 stasiun pemantau kualitas udara (SPKU) di provinsi DKI Jakarta oleh Dinas Lingkungan Hidup Provinsi DKI Jakarta. ISPU merupakan angka tanpa satuan, digunakan untuk menggambarkan kondisi mutu udara ambien di lokasi tertentu dan didasarkan kepada dampak terhadap kesehatan manusia, nilai estetika dan makhluk hidup lainnya [2]. Adapun parameter ISPU meliputi Hidrokarbon (HC), Karbon monoksida (CO), Sulfur dioksida (SO2), Nitrogen dioksida (NO2), Ozon (O3), dan Partikulat (PM10 dan PM2,5). [3]

Data mining adalah proses pencarian informasi yang berguna secara otomatis dalam data yang besar [4]. Teknik *Data mining* yang akan dimanfaatkan pada sistem klasifikasi ini adalah algoritma K-Nearest Neighbors yang merupakan salah satu teknik pengklasifikasian sederhana. Sistem klasifikasi ini akan mengkategorikan kualitas udara berdasarkan data kandungan partikel-partikel yang ada dengan hasil menjadi 3 kategori, yaitu baik, sedang dan tidak sehat. Dengan adanya sistem ini diharapkan semakin mempermudah dalam pengklasifikasian kualitas udara

2. Research Method

2.1. Teknik Pengumpulan Data

Teknik pengumpulan data yang digunakan pada penelitian ini adalah dengan teknik pengumpulan data sekunder. Data yang digunakan pada penelitian ini berasal dari portal data Jakarta yang dapat diakses pada link https://data.jakarta.go.id/. Data tersebut bersumber dari dinas lingkungan hidup DKI Jakarta dengan nama dataset Indeks Standar Pencemaran Udara (ISPU) Tahun 2021. Dataset tersebut berisi Indeks Standar Pencemar Udara (ISPU) yang diukur dari 5 stasiun pemantau kualitas udara (SPKU) yang ada di Provinsi DKI Jakarta pada Tahun 2021 setiap bulannya.

Dataset ini digunakan untuk dapat mengklasifikasikan kualitas udara berdasarkan kandungan partikel-partikel yang ada di udara, mulai dari kandungan partikulat meter pm10, pm25, sulfida, karbon monoksida, ozon dan nitrogen sehingga sistem dapat menarik kesimpulan atas klasifikasi dari kualitas udara dengan parameter masukkan pengguna.

2.2. K-Nearest Neighbors

K-*Nearest Neighbor* (k-NN atau KNN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran (*neighbor*) yang jaraknya paling dekat dengan objek tersebut [5]. K-NN dilakukan dengan mencari kelompok k objek dalam data training yang paling dekat (mirip) dengan objek pada data baru atau data *testing* [6]. Algoritma K-NN menggunakan klasifikasi ketetanggaan sebagai nilai prediksi dari sampel uji yang baru. Dekat atau jauhnya tetangga biasanya dihitung berdasarkan jarak *Euclidean* [7].

Langkah-langkah untuk menghitung algoritma K Nearest Neighbor [8]:

- Menentukan parameter K (Jumlah tetangga paling dekat).
- Menghitung kuadrat jarak Euclid (query instance) masing-masing objek terhadap data sampel yang diberikan.
- Mengurutkan objek-objek tersebut ke dalam kelompok yang mempunyai jarak Euclid terkecil.
- Mengumpulkan kategori Y (Klasifikasi *Nearest Neighbor*) Pada Penelitian ini, algoritma KNN digunakan dengan memanfaatkan library scikit learn.

Scikit-learn atau sklearn adalah modul untuk bahasa pemrograman *python* atau dapat disebut juga sebagai *machine learning library*. Dengan memanfaatkan *library* ini, maka akan mempermudah proses klasifikasi dengan algoritma KNN.

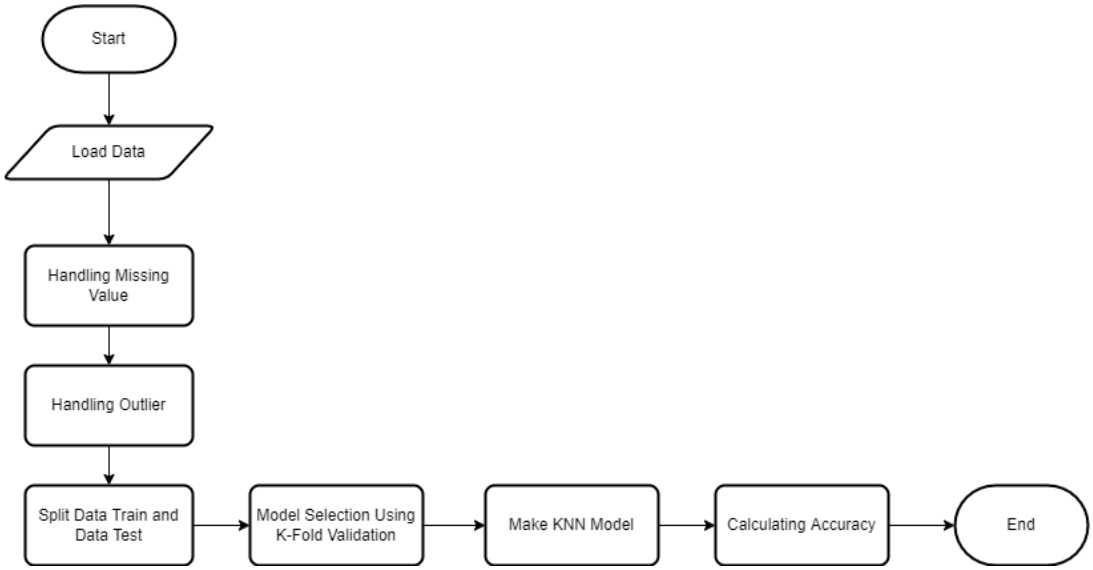
2.3. Confusion Matrix

Confusion matrix merupakan salah satu metode yang dapat digunakan untuk mengukur kinerja suatu metode klasifikasi. Pada dasarnya *confusion matrix* mengandung informasi yang membandingkan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi yang seharusnya. Pada pengukuran kinerja menggunakan *confusion matrix*, terdapat 4 (empat) istilah sebagai representasi hasil proses klasifikasi. Keempat istilah tersebut adalah *True Positive* (TP), *True Negative* (TN), *False Positive* (FP) dan *False Negative* (FN). Nilai True Negative (TN) merupakan jumlah data negatif yang terdeteksi dengan benar, sedangkan False Positive (FP) merupakan data negatif namun terdeteksi sebagai data positif [10].

Pada Penelitian ini, pengukuran tingkat akurasi dengan Confusion Matrix digunakan dengan memanfaatkan *library* scikit learn. Scikit-learn atau sklearn adalah modul untuk bahasa pemrograman python atau dapat disebut juga sebagai machine learning library. Dengan memanfaatkan library ini, maka akan mempermudah proses pengukuran tingkat akurasi dengan metode *Confusion Matrix*.

3. Result and Discussion

3.1. Alur Klasifikasi



Gambar 1. Alur Klasifikasi

Alur klasifikasi dari sistem ini bisa dilihat pada Gambar 1.Pertama, akan dilakukan proses memuat data dari file dengan format .csv yang diunduh dari situs portal data Jakarta. Kemudian dilakukan preprocessing data dengan melakukan penanganan terhadap missing value dengan cara menghapus baris yang nilai *missing value*. Setelah itu, dilakukan penanganan terhadap outlier pada setiap kolom yang ada dengan menggunakan metode winsorize. Metode winsorize akan melakukan transformasi nilai dengan membatasi nilai ekstrim dalam data untuk mengurangi efek *outlier* yang ada pada data.

Setelah proses pra-pemrosesan data, dilakukan algoritma KNN yang dimulai dari menyiapkan data uji dan data latih. Perbandingan dari data uji dan data latih pada penelitian ini adalah 80% data latih dan 20% data uji. Setelah menyiapkan data uji dan data latih, maka dilakukan proses seleksi model dengan menggunakan K-Fold Validation untuk menghitung akurasi dari setiap model yang tersedia. Setelah itu, akan dibuat model KNN dengan akurasi tertinggi menggunakan data uji dan data latih. Penerapan algoritma KNN tersebut pada penelitian ini memanfaatkan *library* scikit learn sehingga akan mempermudah pembuatan model klasifikasi. Terakhir, dilakukan perhitungan akurasi dari model yang telah dibuat menggunakan confusion matrix.

3.2. Seleksi Model

Adapun proses seleksi model menggunakan K-*Fold Validation* dengan fold = 4 dari model KNN dengan K=3, K=5, K=7 dan K=9. K-*Fold Validation* akan membagi data latih menjadi sebesar nilai K yang pada penelitian kali ini dibagi menjadi 4 bagian untuk digunakan sebagai data uji dan data latih sementara. Hasil dari K-*Fold Validation* ini adalah model yang akan digunakan pada penelitian ini dimana pemilihan model tersebut didapatkan dari model yang menghasilkan nilai akurasi paling tinggi pada proses seleksi model dengan K-*Fold Validation*. Hasil skor yang

didapat tiap fold dan rata-rata dari akurasi tiap fold setiap model yang tersedia bisa dilihat pada Gambar 2, Gambar 3, Gambar 4, dan Gambar 5.

```
Akurasi model KNN dengan K=3 untuk tiap fold :  [0.87037037 0.98113208 0.96226415 0.94339623]
Akurasi model KNN dengan K=3 dengan 4-fold Cross Validation :  0.9392907058001398
```

Gambar 2. Akurasi K-*Fold Validation* Pada Model KNN dengan K=3

```
Akurasi model KNN dengan K=5 untuk tiap fold :  [0.90740741 0.96226415 0.94339623 0.94339623]
Akurasi model KNN dengan K=5 dengan 4-fold Cross Validation :  0.9391160027952481
```

Gambar 3. Akurasi K-*Fold Validation* Pada Model KNN dengan K=5

```
Akurasi model KNN dengan K=7 untuk tiap fold :  [0.92592593 0.94339623 0.94339623 0.94339623]
Akurasi model KNN dengan K=7 dengan 4-fold Cross Validation :  0.9390286512928023
```

Gambar 4. Akurasi K-*Fold Validation* Pada Model KNN dengan K=7

```
Akurasi model KNN dengan K=9 untuk tiap fold :  [0.92592593 0.94339623 0.94339623 0.94339623]
Akurasi model KNN dengan K=9 dengan 4-fold Cross Validation :  0.9390286512928023
```

Gambar 5. Akurasi K-*Fold Validation* Pada Model KNN dengan K=9

Berdasarkan akurasi K-*Fold Validation* dari setiap model tersebut, maka akan digunakan model KNN dengan nilai K=3 yang menghasilkan akurasi tertinggi dari proses K-*Fold Validation* tersebut sebagai model klasifikasi pada sistem ini.

3.3. Evaluasi Sistem

Evaluasi dari performa model KNN dengan nilai K=3 pada penelitian ini dapat dilihat pada *confusion matrix* di Tabel 1. Dari confusion matriks tersebut, bisa didapatkan nilai dari *Precision, Recall, F-Measure* dan *Accuracy* yang didapatkan oleh model klasifikasi yang telah dibuat. Nilai dari *Precision, Recall, F-Measure* dan *Accuracy* bisa dilihat pada Gambar 6.

Tabel 1. Confusion Matriks

		Nilai Sebenarnya		
Prediksi Sistem		Baik	Sedang	Tidak Sehat
	Baik	12	1	0
	Sedang	0	73	0
	Tidak Sehat	0	5	1

```
Accuracy using K-NN with K = 3 : 93.47826086956522%

Precision using K-NN with K = 3 : 93.97358282883876%

Recall using K-NN with K = 3 : 93.47826086956522%

F-Measure using K-NN with K = 3 : 91.64424648577966%
```

Gambar 6. Nilai dari Precision, Recall, F-Measure dan Accuracy dari model KNN

4. Conclusion

Berdasarkan hasil penelitian mengenai perancangan sistem klasifikasi udara dengan metode k-nearest neighbors berdasarkan partikulat udara yang terkandung, maka penulis dapat mengambil kesimpulan sebagai berikut:

1) Teknik Preprocessing pada penelitian ini adalah penanganan terhadap missing value dan penanganan terhadap outlier dengan metode winsorize.
2) Model K-*Nearest Neighbors* yang menghasilkan akurasi terbaik diantara nilai K = 3, 5, 7, dan 9 adalah model K-Nearest Neighbors dengan nilai K = 3 dengan nilai akurasinya 93.47826086956522%.

References

[1] V. B. Kusnandar, "*Lebih dari 70% Penduduk Jakarta Merupakan Usia Produktif*", 22 November 2021. [Online]. Available: https://databoks.katadata.co.id/datapublish/2021/11/22/lebih-dari-70-penduduk-jakarta-merupakan-usia-produktif#:~:text=Jumlah%20Penduduk%20DKI%20Jakarta%20sebanyak,(15%2D64%20tahun. [15 June 2022]

[2] D. Chaniago, A. Zahara, I. S. Ramadhani, "INDEKS STANDAR PENCEMAR UDARA (ISPU) SEBAGAI INFORMASI MUTU UDARA AMBIEN DI INDONESIA", 24 September 2020. [Online]. Available: https://ditppu.menlhk.go.id/portal/read/indeks-standar-pencemar-udara-ispu-sebagai-informasi-mutu-udara-ambien-di-indonesia. [9 March 2022]

[3] R. Ramadhan, "Indeks Standar Pencemar Udara (ISPU) Berbasis Android : "ISPU Net"", 2 August 2021. [Online]. Available: https://lingkunganhidup.jogjakota.go.id/detail/index/330. [9 March 2021]

[4] P. M. A. Putra, and I. G. A. G. A. Kadyanan, "Implementation of K-Means Clustering Algorithm in Determining Classification of the Spread of the COVID19 Virus in Bali" *Jurnal Elektronik Ilmu Komputer Udayana,* vol. 10, no. 1, p. 21-28, 2021.

[5] Y. Wiyli, "Algoritma K-Nearest Neighbour untuk Memprediksi Harga Jual Tanah" *Jurnal Matematika, Statistika, & Komputasi,* vol. 9, no. 1, p. 57-68, 2012.

[6] Suhartini, H. Bahtiar, "Klasifikasi Algoritma K-Nearest Neighbor Berbasis Particle Swarm Optimization Untuk Kelayakan Bantuan Rehabilitasi Rumah Tidak Layak Huni Pada Desa Lenek Duren Kecamatan Aikmel Kabupaten Lombok Timu" *Infotek : Jurnal Informatika dan Teknologi,* vol. 2, no. 2, p. 79-85, 2019.

[7] Y. Yahya and W. P. Hidayanti, "Penerapan Algoritma K-Nearest Neighbor Untuk Klasifikasi Efektivitas Penjualan Vape (Rokok Elektrik) pada "Lombok Vape On"" *Infotek : Jurnal Informatika dan Teknologi*, vol. 3, no. 2, p. 104-114, 2020.

[8] Y. Yahya and R. Zuliana, "Prediksi Jumlah Penggunaan BBM Perbulan Menggunakan Algoritma Decition Tree(C4.5)" *Infotek : Jurnal Informatika dan Teknologi*, vol. 1, no. 1, pp. 56–63, 2018.

[9] A. Rizal, "K-Nearest Neighbor (K-NN)", 26 July 2011. [Online]. Available: https://achmadrizal.staff.telkomuniversity.ac.id/k-nearest-neighbor-k-nn/. [15 June 2022].

[10] A. Solichin, "*Mengukur Kinerja Algoritma Klasifikasi dengan Confusion Matrix", 19 March 2017.* [Online]. Available: https://achmatim.net/2017/03/19/mengukur-kinerja-algoritma-klasifikasi-dengan-confusion-matrix/. [15 June 2022].

Halaman ini sengaja dibiarkan kosong