Jurnal Riset Statistika (JRS)
e-ISSN 2798-6578 | p-ISSN 2808-3148
https://journals.unisba.ac.id/index.php/JRS

Tersedia secara online di
Unisba Press
https://publikasi.unisba.ac.id/

Algoritma Pemilihan Variabel untuk Klasifikasi dan Penerapannya pada Klasifikasi Desa-Kelurahan

Desi Permatasari, Suliadi Suliadi*

Prodi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Bandung, Indonesia.

ARTICLE INFO

Article history :
Received : 11/5/2024
Revised : 6/6/2024
Published : 31/7/2024

Volume : 4
No. : 1
Halaman : 49 - 56
Terbitan : **Juli 2024**

ABSTRAK

K-Nearest Neighbor (KNN) adalah salah satu metode klasifikasi non-parametrik yang prinsip pengerjaannya dengan mengklasifikasikan suatu objek dalam data testing berdasarkan kelas mayoritas dari k tetangga terdekatnya (neighbor) pada data training. Salah satu permasalahan yang utama pada KNN yaitu banyak variabel yang harus digunakan dalam pelaksanaannya, terlebih terdapat kasus dimana banyaknya variabel lebih besar dibanding dengan banyaknya pengamatan. Maka dari itu, Beuren & Anzanello (2019) membuat kerangka kerja baru untuk pemilihan variabel dengan menggabungkan beberapa metode, yaitu Mutual Information (MI), Algoritma Kennard Stone (KS), Statistik Anderson Darling (AD), Statistik Kruskal Wallis (KW), K-Nearest Neighbor (KNN), Leave One Out Cross Validation (LOOCV), dan Confusion Matrix. Pada penelitian ini menerapkan metode Beuren & Anzanello (2019) untuk memilih variabel untuk mengklasifikasikan desa/kelurahan ke dalam status perdesaan dan perkotaan di wilayah Bandung Raya. Dengan menggunakan algoritma K-Nearest Neighbor (KNN) dan k = 5 tetangga terdekat dari 24 variabel diperoleh kombinasi variable independent, yaitu Kepadatan Penduduk (X_8) dan Banyaknya Penduduk (X_5) adalah kombinasi yang paling cocok dalam mengklasifikasikan status perdesaan dan perkotaan pada desa/kelurahan yang ada di wilayah Bandung Raya, diperoleh hasil akurasi sebesar 80.72%.

Kata Kunci : K-Nearest Neighbor; Pemilihan Variabel; Statistik Non-parametrik

ABSTRACT

K-Nearest Neighbor (KNN) is a non-parametric classification method whose principle is to classify an object in the testing data based on the majority class of its k nearest neighbors in the training data. One of the main problems with KNN is that there are many variables that must be used in its implementation, especially in cases where the number of variables is greater than the number of observations. Therefore, Beuren & Anzanello (2019) created a new framework for variable selection by combining several methods, namely Mutual Information (MI), Kennard Stone Algorithm (KS), Anderson Darling Test (AD), Kruskal Wallis Test (KW), K-Nearest Neighbor (KNN), Leave One Out Cross Validation (LOOCV), and Confusion Matrix. In this study, Beuren & Anzanello (2019) applied the method to select variables to classify villages/kelurahan into rural and urban status in the Greater Bandung area. By using the K-Nearest Neighbor (KNN) algorithm and k = 5 nearest neighbors from 24 variables, a combination of independent variables is obtained, namely Population Density (X_8) and Number of Population (X_5)which is the most suitable combination in classifying rural and urban status in villages/urban villages in the Greater Bandung area, obtained accuracy results of 80.72%.

Keywords : K-Nearest Neighbor; Variable Selection; Non-parametric Tests

Corresponding Author : *suliadi@gmail.com
Indexed : Garuda, Crossref, Google Scholar
DOI : https://doi.org/10.29313/jrs.v4i1.3876

A. Pendahuluan

Klasifikasi merupakan suatu proses untuk memasukkan suatu objek ke dalam satu dari beberapa kategori, salah satu metodenya adalah *K-Nearest Neighbor* (KNN). *K-Nearest Neighbor* (KNN) adalah salah satu metode klasifikasi non-parametrik yang prinsip pengerjaannya dengan mengklasifikasikan suatu objek dalam data testing berdasarkan kelas mayoritas dari k tetangga terdekatnya (neighbor) pada data training [1].

Salah satu permasalahan yang utama pada KNN yaitu terlalu banyak variabel pengklasifikasi atau kasus dimana banyaknya variabel lebih besar dibanding dengan banyaknya pengamatan. Solusi yang dapat digunakan yaitu dengan dengan mengurangi banyaknya variabel dalam kumpulan data hingga batas terendah yang dapat ditoleransi, dalam arti variabel yang dieliminasi tersebut tidak akan menyebabkan hilangnya informasi yang penting dan berguna, salah satu metodenya adalah Mutual Information (MI). Mutual Information (MI) merupakan sebuah konsep tentang teori informasi yang digunakan untuk mengukur keeratan hubungan antara dua variabel acak [2]. Statistik non-parametrik merupakan alternatif yang efisien untuk mengidentifikasi variabel yang paling relevan, terutama yang berbasis peringkat (rank) [3].

Wilayah metropolitan Bandung Raya merupakan salah satu wilayah metropolitan yang meliputi Kota Bandung, Kota Cimahi, Kabupaten Bandung, Kabupaten Bandung Barat, dan Kabupaten Sumedang. Wilayah metropolitan Bandung Raya merupakan pusat kegiatan nasional. Wilayah ini tercipta sebagai akibat dari pertumbuhan Kota Bandung, yaitu dari jumlah dan kepadatan penduduk yang berkembang cukup pesat, sehingga menyebabkan adanya kebutuhan akan ruang terutama untuk kebutuhan untuk pemukiman [4].

Klasifikasi desa/kelurahan ke dalam status perkotaan dan perdesaan membantu mengalokasikan sumber daya secara efisien, sehingga perbedaan desa/kelurahan perdesaan dan perkotaan dapat dipahami dengan jelas, sumber daya seperti infrastruktur, pendidikan, kesehatan, perekonomian, dan lain-lain dapat digunakan secara lebih tepat sasaran dan efisien, oleh karena itu menggunakan sebanyak 24 variable independent untuk mengklasifikasikan desa/kelurahan ke dalam status perdesaan dan perkotaan [5]. Maka dari itu, diperlukan pemilihan variabel dengan memperhatikan konteks lokal maupun faktor-faktor khusus yang relevan dengan wilayah metropolitan Bandung Raya ketika memilih klasifikasi status desa/kelurahan sebagai perkotaan atau perdesaan.

Pada penelitan sebelumnya dilakukan pembuatan kerangka kerja baru untuk pemilihan variabel dengan menggabungkan beberapa metode, yaitu menggabungkan beberapa metode, yaitu Mutual Information (MI), Algoritma *Kennard Stone* (KS), Statistik *Anderson Darling* (AD), Statistik *Kruskal Wallis* (KW), *K-Nearest Neighbor* (KNN), *Leave One Out Cross Validation* (LOOCV), dan *Confusion Matrix* [6]. Dalam penelitian ini peneliti akan menerapkan metode tersebut untuk menyeleksi variabel dalam mengklasifikasikan desa/kelurahan ke dalam status perkotaan dan perdesaan. Tujuan dari penulisan skripsi ini, yaitu menyeleksi variabel berdasarkan metode yang diajukan oleh penelitian sebelumnya [6] dalam mengklasifikasikan status perdesaan dan perkotaan pada desa/kelurahan yang ada di wilayah Bandung Raya dan mengklasifikasikan status perdesaan dan perkotaan pada desa/kelurahan yang ada di wilayah Bandung raya.

B. Metode Penelitian

Algoritma yang diajukan oleh penelitian sebelumnya [6] menggabungkan beberapa metode, yaitu *Mutual Information* (MI), Algoritma *Kennard Stone* (KS), Statistik *Anderson Darling* (AD), Statistik *Kruskal Wallis* (KW), *K-Nearest Neighbor* (KNN), *Leave One Out Cross Validation* (LOOCV), dan *Confusion Matrix*. Tahapan pemilihan variabel berdasarkan algoritma yang diusulkan adalah sebagai berikut.

Mutual Information (MI)

Mutual Information (MI) merupakan sebuah konsep tentang teori informasi yang digunakan untuk mengukur keeratan hubungan antara dua variabel acak [7]. Untuk menghitung *Mutual Information* (MI) antara dua variabel numerik perlu melakukan tahap-tahap. Pertama, diskritisasi data yang menggunakan algoritma *binning* atau pengelompokan dengan frekuensi atau lebar yang sama. *Binning* digunakan untuk mereduksi banyaknya *variable independent* (*feature*) yang merupakan data numerik menjadi data kategorik dengan cara membagi rentang nilai data ke dalam interval. Untuk menentukan banyaknya *bin* dengan cara menggunakan aturan *Sturges* yaitu [2]:

$$k = 1 + \log_2 n \quad (1)$$

Kedua, menghitung *Mutual Information* (MI). Misalkan variabel Y terdiri dari C_1 kategori/kelas, yaitu $y_1, y_2, \dots, y_{C_1}$dan variabel X_q setelah didiskritisasi terdiri C_2 *bin*/kategori, yaitu $x_{q_1}, x_{q_1}, \dots, x_{qC_2}$. Secara formal, *Mutual Information* (MI) antara dua variabel acak x_q dan *y* diskrit didefinisikan sebagai:

$$MI(X_q; Y) = \sum_{j=1}^{C_2} \sum_{k=1}^{C_1} p(x_{qj}, y_k) \log\left(\frac{p(x_{qj}, y_k)}{p(x_{qj})p(y_k)}\right) \quad (2)$$

Algoritma *Kennard Stone* (KS)

Algoritma *Kennard Stone* (KS) merupakan teknik yang membagi kumpulan data menjadi dalam data *training* dan data *testing* sehingga setiap sampel baru ditempatkan sejauh mungkin dari sampel yang sudah didistribusikan [8]. Dalam pemodelan, perlu data untuk mengevaluasi apakah model yang dibuat adalah baik atau tidak sehingga model tersebut perlu dicobakan pada data baru. Untuk itu dalam praktek, data yang ada dibagi dua, yaitu data *training* digunakan untuk menyusun model dan data *testing* digunakan untuk mengevaluasi model. Pemilihan data *training* dan data *testing* harus dilakukan sedemikan sehingga kedua set data tersebut memiliki karakteristik yang tidak terlalu berbeda. Algoritma *Kennard Stone* (KS) menggunakan jarak *Euclidean*. Dalam konteks jarak *Euclidean*, $d(O_i, O_{i'})$ adalah jarak *Euclidean* antara dua objek atau titik O_i dan $O_{i'}$ dalam ruang. Jika kita memiliki dua objek, yaitu $O_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ dan $O_{i'} = (x_{i'1}, x_{i'2}, \dots, x_{i'p})$, maka rumus *Euclidean* untuk menghitung jaraknya adalah sebagai berikut:

$$d(O_i, O_{i'}) = \sqrt{\sum_{j=1}^{p} (x_{iq} - x_{i'q})^2} \quad (3)$$

Statistik *Anderson Darling* (AD)

Statistik *Anderson Darling* (AD) digunakan untuk memperkirakan tingkat kepentingan dari setiap variabel *q*. Bisa juga digunakan untuk menetapkan perbedaan dalam beberapa sampel populasi dengan sensitivitas tertentu terhadap ekor dikumpulkan sampel atau dapat digunakan untuk menilai apakah beberapa sampel cukup mirip sehingga mereka dapat dikumpulkan untuk analisis lebih lanjut [9]. Statistik *Anderson Darling* (AD) didefinisikan sebagai berikut [10]:

$$AD_q = \frac{n-1}{n^2(C-1)} \sum_{c=1}^{C} \left[\frac{1}{n_c} \sum_{j=1}^{L} h_j \frac{(nF_{ij} - n_c H_j)^2}{H_j(n - H_j) - nh_j/4} \right] \quad (4)$$

Keterangan:

$n(.)$ = banyaknya pengamatan yang memenuhi $(.)$

n = banyaknya pengamatan untuk semua kelas

n_c = banyaknya pengamatan pada kelas ke-c

h_j = $n(x_{ck} = z_j)$

z_j = kumpulan data gabungan kemudian diurutkan dari yang terkecil sampai terbesar yang unik (jika ada data yang sama, maka dicatat satu kali)

H_j = $n(x_{ck} < z_j) + \frac{1}{2} n(x_{ck} = z_j)$

F_{Cj} = H_j khusus data pada kelas ke-c

C = banyaknya kelas

Statistik *Kruskal Wallis* (KW)

Statistik *Kruskal Wallis* (KW) merupakan statistik non-parametrik yang bertujuan untuk membandingkan dua akelas atau lebih dengan setiap variabel yang masing-masing variabelnya untuk menguji apakah kelas-kelas tersebut berasal dari distribusi yang sama berdasarkan peringkat median [11]. Statistik *Kruskal Wallis* (KW) didefinisikan sebagai berikut [1]:

$$KW_q = (n-1) \frac{\sum_{c=1}^{C} n_c (\bar{R}_{c.} - \bar{R})^2}{\sum_{c=1}^{C} \sum_{k=1}^{n_C} (R_{ck} - \bar{R})^2} \quad (5)$$

Keterangan:
n_c = banyaknya pengamatan ke-k dari kelas ke-c
C = banyaknya kelas
R_{ck} = peringkat x_{ck} dari keseluruhan kelas
$\bar{R}_{c.}$ = peringkat rata-rata dari semua pengamatan dari kelas ke-c
$\bar{R}$ = rata-rata dari R_{ck}

***K-Neareast Neighbor* (KNN)**
K-Nearest Neighbor (KNN) adalah salah satu metode klasifikasi non-parametrik yang prinsip pengerjaannya dengan mengklasifikasikan suatu objek dalam data *testing* berdasarkan kelas mayoritas dari *k* tetangga terdekatnya (*neighbor*) pada data *training* [12]. Hitung jarak antara objek i' (data *testing)* dengan setiap objek $i = 1,2,3, \dots, n_1$ (data *training*) menggunakan jarak *Euclidean*. Rumus untuk mencari jarak *Euclidean* adalah sebagai berikut:

$$d(i,i')_{euclidean} = \sqrt{\sum_{q=1}^{p} \left(x_{iq} - x_{i'q}\right)^2} \quad (6)$$

Keterangan:
$d(i,i')$ = jarak antara objek ke-*i* dengan objek ke-*i'*
x_{iq} = pengamatan ke-*i* dan variabel ke-*q* pada data *training*
$x_{i'q}$ = pengamatan ke-*i*' dan variabel ke-*q* pada data *testing*

***Leave-One-Out Cross Validation* (LOOCV)**
Leave-One-Out Cross Validation (LOOCV) merupakan metode validasi silang yang bertujuan untuk mengetahui akurasi setiap iterasi, di mana satu objek digunakan menjadi data testing dan sisanya digunakan menjadi data *training* [13]. Apabila seluruh dataset mempunyai sebanyak *n* objek, maka pendekatan validasi silang yang digunakan dalam metode ini adalah menggunakan 1 objek sebagai data *testing* dan menggunakan *n* – 1 objek sisanya sebagai data *training*. Pengulangan dilakukan sebanyak *n* kali.

Confusion Matrix
Confusion Matrix merupakan salah satu teknik yang bertujuan untuk mengetahui kinerja dalam klasifikasi. *Confusion Matrix* berisi informasi yang membandingkan hasil klasifikasi yang dilakukan sistem dengan hasil klasifikasi sebenarnya [14]. Berikut ini *Confusion Matrix* yang akan ditampilkan sebagaimana dalam **Tabel 1**.

Tabel 1. *Confusion Matrix*

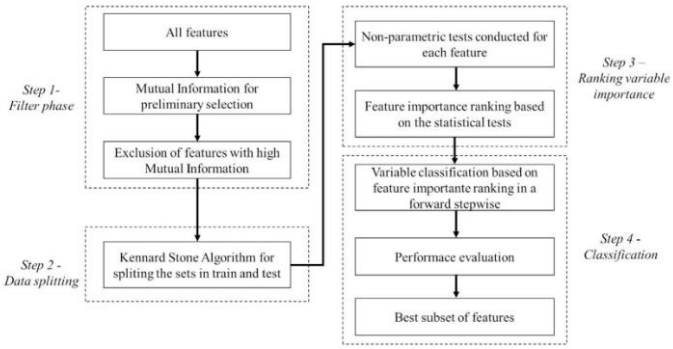
		Kelas Hasil Prediksi		**Jumlah**
		Ya	**Tidak**	
Kelas Aktual	**Ya**	TP	FN	P
	Tidak	FP	TN	N
Jumlah		P'	P'	N'

Nilai akurasi menunjukkan seberapa akurat sistem dapat menghasilkan data dengan benar atau perbandingan antara data terklasifikasi benar dengan keseluruhan data. Nilai akurasi diperoleh dengan menggunakan rumus sebagai berikut:

$$Accuracy = \frac{TP + TN}{P + N} \times 100\% \quad (7)$$

Kerangka Kerja Pemilihan Variabel
Kerangka kerja pada penelitian ini menggunakan beberapa metode, yaitu *Mutual Information* (MI) untuk mengukur seberapa erat hubungan antara dua variabel acak, Algoritma *Kennard Stone* (KS) untuk membagi data menjadi data *training* dan data *testing*, kemudian variabel yang tersisa adalah diberi peringkat berdasarkan

kepentingannya menggunakan dua statistik non-parametrik, yaitu Statistik *Anderson Darling* (AD) dan Statistik *Kruskal Wallis* (KW). Selanjutnya, pengklasifikasian menggunakan algoritma *K-Nearest Neighbor* (KNN) [6]. Berikut ini adalah diagram alur kerangka kerja untuk pemilihan variabel dalam klasifkasi ditampilkan diagram alur kerangka kerja yang akan diolah seperti pada **Gambar 1**.



Gambar 1. Langkah-langkah Kerangka Kerja [6]

Pengertian Desa-Perkotaan dan Desa-Perdesaan

Untuk mengetahui klasifikasi desa/kelurahan perkotaan perdesaan perlu dijelaskan mengenai beberapa pengertian secara statistik sebagai berikut [15]:

Daerah perkotaan merupakan status suatu wilayah administrasi pada tingkat desa/kelurahan yang memenuhi persyaratan spesifik pada hal kepadatan penduduk, banyaknya fasilitas perkotaan, sarana pendidikan formal, sarana kesehatan umum, sarana perekonomian, dan sebagainya.

Daerah perdesaan merupakan status suatu wilayah administrasi pada tingkat desa/kelurahan yang belum memenuhi persyaratan spesifik pada hal kepadatan penduduk, banyaknya fasilitas perkotaan, sarana pendidikan formal, sarana kesehatan umum, sarana perekonomian, dan sebagainya.

C. Hasil dan Pembahasan

***Mutual Information* (MI)**

Mutual Information (MI) mengasumsikan data diskrit, karena data klasifikasi status desa/kelurahan di wilayah Bandung Raya yang terdapat *variable independent* (*X*) ada sebanyak 24 adalah berupa numerik maka perlu melakukan diskritisasi data menggunakan *binning*. Setelah melakukan *binning* untuk semua *variable independent*, maka Langkah selanjutnya adalah menghitung *Mutual Information* (MI) antara masing-masing semua *variable independent* dengan *variable dependent*. Diperoleh hasil *Mutual Information* (MI) Tabel 2:

Tabel 2. *Mutual Information* (MI)

Variabel	MI	Variabel	MI
$MI(X_1;Y)$	0.007	$MI(X_{13};Y)$	0.021
$MI(X_2;Y)$	0.000	$MI(X_{14};Y)$	0.032
$MI(X_3;Y)$	0.030	$MI(X_{15};Y)$	0.067
$MI(X_4;Y)$	0.049	$MI(X_{16};Y)$	0.056
$MI(X_5;Y)$	0.087	$MI(X_{17};Y)$	0.052
$MI(X_6;Y)$	0.006	$MI(X_{18};Y)$	0.022
$MI(X_7;Y)$	0.035	$MI(X_{19};Y)$	0.021
$MI(X_8;Y)$	0.086	$MI(X_{20};Y)$	0.080
$MI(X_9;Y)$	0.000	$MI(X_{21};Y)$	0.018
$MI(X_{10};Y)$	0.032	$MI(X_{22};Y)$	0.044
$MI(X_{11};Y)$	0.028	$MI(X_{23};Y)$	0.003
$MI(X_{12};Y)$	0.025	$MI(X_{24};Y)$	0.031

Algoritma *Kennard Stone* (KS)

Membagi data menjadi data *training* dan data *testing* menggunakan Algoritma *Kennard Stone* menggunakan jarak *Euclidean* dengan menggunakan persamaan (3). Dalam penelitian ini, setiap dataset dibagi 75% pengamatan sebagai data *training* dan 25% sebagai data *testing*. Diperoleh data *training* sebanyak 664 pengamatan, yaitu terdiri dari perkotaan sebanyak 508 pengamatan dan perdesaan sebanyak 156 pengamatan. Sedangkan, data *testing* sebanyak 223 pengamatan, yaitu terdiri dari perkotaan sebanyak 170 pengamatan dan perdesaan sebanyak 53 pengamatan.

Dataset berdasarkan MI dengan *Cut Off* MI(0.04)

Terdapat delapan variable independent dengan nilai $MI > 0.04$, yaitu X_5, X_8, X_{20}, X_{15}, X_{16}, X_{17}, X_4, dan X_{22}. Kemudian variabel yang tersisa adalah diberi peringkat berdasarkan kepentingannya menggunakan dua statistik non-parametrik, yaitu Statistik *Anderson Darling* (AD) dan Statistik *Kruskal Wallis* (KW). Selanjutnya, pengklasifikasian menggunakan algoritma *K-Nearest Neighbor* (KNN) menggunakan jarak *Euclidean* dan *k* = 5.

Setelah diperoleh nilai akurasi menggunakan jarak *Euclidean* dan *k* = 5 menggunakan pemilihan variabel dan menggunakan semua variabel, selanjutnya akan dibandingkan untuk melihat apakah nilai akurasinya akan menurun jika menggunakan pemilihan variabel untuk data klasifikasi desa/kelurahan ke dalam status perkotaan dan perdesaan di wilayah Bandung Raya. Berikut merupakan tabel hasil perhitungan akurasi pada masing-masing *Confusion Matrix*:

Tabel 3. Komparasi Nilai Akurasi Pada Dataset MI(0.04)

Confusion Matrix	*Variable Independent* Terpilih	*Variable Independent* Terpilih (%)	Akurasi
AD	$\{X_8, X_{20}, X_5, X_{17}\}$	**16.67%**	**79.37%**
KW	$\{X_8, X_5, X_{17}\}$	**12.50%**	**79.37%**
Seluruh Variabel	$\{X_1 - X_{24}\}$	**100.00%**	**79.37%**

Dari ketiga *Confusion Matrix* tersebut dapat dilihat bahwa memiliki nilai akurasi yang sama, maka dapat disimpulkan bahwa pemilihan variabel tidak dapat meningkatkan nilai akurasi yang digunakan untuk klasifikasi desa/kelurahan ke dalam status perkotaan dan perdesaan di wilayah Bandung Raya berdasarkan MI dengan *cut off* $MI(0.04)$.

Dataset berdasarkan MI dengan *Cut Off* MI(0.06)

Terdapat empat variable independent dengan nilai $MI > 0.06$, yaitu X_5, X_8, X_{20}, dan X_{15}. Kemudian variabel yang tersisa adalah diberi peringkat berdasarkan kepentingannya menggunakan dua statistik non-parametrik, yaitu Statistik Anderson Darling (AD) dan Statistik *Kruskal Wallis* (KW). Selanjutnya, pengklasifikasian menggunakan algoritma *K-Nearest Neighbor* (KNN) menggunakan jarak Euclidean dan k = 5.

Setelah diperoleh nilai akurasi menggunakan jarak Euclidean dan k = 5 menggunakan pemilihan variabel dan menggunakan semua variabel, selanjutnya akan dibandingkan untuk melihat apakah nilai akurasinya akan menurun jika menggunakan pemilihan variabel untuk data klasifikasi desa/kelurahan ke dalam status perkotaan dan perdesaan di wilayah Bandung Raya. Berikut merupakan tabel hasil perhitungan akurasi pada masing-masing *Confusion Matrix*:

Tabel 4. Komparasi Nilai Akurasi Pada Dataset MI(0.06)

Confusion Matrix	*Variable Independent* Terpilih	Variabel Terpilih (%)	Akurasi
AD	$\{X_8, X_5\}$	**8.33%**	**80.72%**
KW	$\{X_8, X_5\}$	**8.33%**	**80.72%**
Seluruh Variabel	$\{X_1 - X_{24}\}$	**100.00%**	**80.27%**

Dari ketiga *Confusion Matrix* tersebut dapat dilihat bahwa pemilihan variabel memiliki nilai akurasi yang paling besar, maka dapat disimpulkan bahwa pemilihan variabel dapat meningkatkan nilai akurasi yang

digunakan untuk klasifikasi desa/kelurahan ke dalam status perkotaan dan perdesaan di wilayah Bandung Raya berdasarkan MI dengan cut off $MI(0.06)$.

Dataset berdasarkan MI dengan *Cut Off* MI(0.08)

Terdapat dua *variable independent* dengan nilai $MI > 0.08$, yaitu X_5 dan X_8. Kemudian variabel yang tersisa adalah diberi peringkat berdasarkan kepentingannya menggunakan dua statistik non-parametrik, yaitu Statistik *Anderson Darling* (AD) dan Statistik *Kruskal Wallis* (KW). Selanjutnya, pengklasifikasian menggunakan algoritma *K-Nearest Neighbor* (KNN) menggunakan jarak *Euclidean* dan *k* = 5.

Setelah diperoleh nilai akurasi menggunakan jarak *Euclidean* dan *k* = 5 menggunakan pemilihan variabel dan menggunakan semua variabel, selanjutnya akan dibandingkan untuk melihat apakah nilai akurasinya akan menurun jika menggunakan pemilihan variabel untuk data klasifikasi desa/kelurahan ke dalam status perkotaan dan perdesaan di wilayah Bandung Raya. Berikut merupakan tabel hasil perhitungan akurasi pada masing-masing *Confusion Matrix*:

Tabel 5. Komparasi Nilai Akurasi Pada Dataset MI(0.08)

Confusion Matrix	*Variable Independent* Terpilih	Variabel Terpilih (%)	Akurasi
AD	$\{X_8, X_5\}$	8.33%	80.72%
KW	$\{X_8, X_5\}$	8.33%	80.72%
Seluruh Variabel	$\{X_1 - X_{24}\}$	100.00%	80.27%

Dari ketiga *Confusion Matrix* tersebut dapat dilihat bahwa pemilihan variabel memiliki nilai akurasi yang paling besar, maka dapat disimpulkan bahwa pemilihan variabel dapat meningkatkan nilai akurasi yang digunakan untuk klasifikasi desa/kelurahan ke dalam status perkotaan dan perdesaan di wilayah Bandung Raya berdasarkan MI dengan *cut off* $MI(0.08)$.

Komparasi Keenam *Confusion Matrix*

Setelah *diperoleh* nilai akurasi menggunakan jarak Euclidean dan k = 5 masing-masing dataset, selanjutnya akan dibandingkan untuk melihat mana yang terbaik untuk data klasifikasi desa/kelurahan ke dalam status perkotaan dan perdesaan di wilayah Bandung Raya. Berikut merupakan tabel hasil perhitungan akurasi pada masing-masing *Confusion Matrix*:

Tabel 6. Komparasi Keenam *Confusion Matrix* dengan Nilai Akurasi

Confusion Matrix	Variable Independent Terpilih	Variable Independent Terpilih (%)	Akurasi
AD MI004	$\{\boldsymbol{X_8, X_{20}, X_5, X_{17}}\}$	**16.67%**	**79.37%**
KW MI004	$\{\boldsymbol{X_8, X_5, X_{17}}\}$	**12.50%**	**79.37%**
AD MI006	$\{\boldsymbol{X_8, X_5}\}$	**8.33%**	**80.72%**
KW MI006	$\{\boldsymbol{X_8, X_5}\}$	**8.33%**	**80.72%**
AD MI008	$\{\boldsymbol{X_8, X_5}\}$	**8.33%**	**80.72%**
KW MI008	$\{\boldsymbol{X_8, X_5}\}$	**8.33%**	**80.72%**

Dalam menentukan pemilihan model terbaik berdasarkan nilai akurasi yang paling besar. Dari keenam *Confusion Matrix* tersebut dapat dilihat bahwa *AD* $MI(0.06)$, *KW* $MI(0.06)$, *AD* $MI(0.08)$, dan *KW MI* (0.08) memiliki nilai akurasi yang paling besar. Maka dapat disimpulkan bahwa kombinasi *variable independent*, yaitu Kepadatan Penduduk (X_8) dan Banyaknya Penduduk (X_5) adalah kombinasi yang paling cocok untuk data klasifikasi perkotaan/perdesaan.

D. Kesimpulan

Berdasarkan pembahasan dalam penelitian ini, maka dapat disimpulkan bahwa kombinasi *variable independent*, yaitu Kepadatan Penduduk (X_8) dan Banyaknya Penduduk (X_5) adalah kombinasi yang paling cocok dalam mengklasifikasikan status perdesaan dan perkotaan pada desa/kelurahan yang ada di wilayah Bandung Raya dan penerapan algoritma *K-Nearest Neighbor* (KNN) menggunakan *variable independent*,

yaitu Kepadatan Penduduk (X_8) dan Banyaknya Penduduk (X_5) dengan menggunakan fungsi jarak *Euclidean* dan k = 5 tetangga terdekat, diperoleh hasil akurasi sebesar 80.72%. Maka, dapat disimpukan bahwa model baik dalam melakukan klasifikasi (*good classification*) dalam mengklasifikasikan status perdesaan dan perkotaan pada desa/kelurahan yang ada di wilayah Bandung Raya.

Daftar Pustaka

[1] S. Pramana, B. Yuniarto, S. Mariyah, I. Santoso, and R. Nooraeni, *Data Mining dengan R Konsep Serta Impletensi*. Bogor : IN MEDIA, 2018.

[2] R. Intan, O. Y. Yuliana, and D. Kristanto, "Bayesian Belief Network untuk Menghasilkan Fuzzy Association Rules," *Jurnal Teknik Industri*, vol. 12, no. 1, pp. 55–60, May 2010, doi: 10.9744/jti.12.1.55-60.

[3] T. P. Hettmansperger, J. W. McKean, and S. J. Sheather, "Robust Nonparametric Methods," *J Am Stat Assoc*, vol. 95, no. 452, pp. 1308–1312, Dec. 2000, doi: 10.2307/2669777.

[4] N. Suhartina and I. Sukarsih, "Model SVEIR Penyebaran Penyakit Rabies Terhadap Anjing dengan Vaksinasi," *DataMath: Journal of Statistics and Mathematics*, vol. 2, no. 1, pp. 25–32, 2024.

[5] A. D. Sofia and A. Kudus, "Pengelompokan Kabupaten/Kota berdasarkan Indikator Indeks Pembangunan Manusia 2022 Menggunakan K-Harmonic Means Clustering," *Jurnal Riset Statistika*, vol. 3, no. 2, pp. 163–172, Dec. 2023, doi: 10.29313/jrs.v3i2.3130.

[6] G. M. Beuren and M. J. Anzanello, "Variable selection using statistical non-parametric tests for classifying production batches into multiple classes," *Chemometrics and Intelligent Laboratory Systems*, vol. 193, no. 1, p. 103830, Oct. 2019, doi: 10.1016/j.chemolab.2019.103830.

[7] G. W. Corder and D. J. Foreman, *Nonparametric Statistics for Non-Statisticians*. John Wiley & Sons, Inc, 2009.

[8] R. W. Kennard and L. A. Stone, "Computer Aided Design of Experiments," *Technometrics*, vol. 11, no. 1, pp. 137–148, Feb. 1969, doi: 10.1080/00401706.1969.10490666.

[9] R. Kohavi, "Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," in *International Joint Conference on Articial Intelligence (IJCAI)*, 2001.

[10] W. H. Kruskal, "A Nonparametric test for the Several Sample Problem," *The Annals of Mathematical Statistics*, vol. 23, no. 4, pp. 525–40, 1952.

[11] B. Yu. Lemeshko and I. V. Veretel'nikova, "Power of k-Sample Tests Aimed at Checking the Homogeneity of Laws," *Measurement Techniques*, vol. 61, no. 7, pp. 647–654, Oct. 2018, doi: 10.1007/s11018-018-1479-1.

[12] BPS, *Head of BPS Regulation No. 120 of 2020 Concerning Classification of Villages, Urban and Rural Areas in Indonesia 2020, Book 2*. Badan Pusat Statitsik, 2020.

[13] F. W. Scholz and M. A. Stephens, "K-Sample Anderson-Darling Tests," *J Am Stat Assoc*, vol. 82, no. 399, pp. 918–924, 1987.

[14] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*, 2nd Edition. Wiley, 2015.

[15] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf Process Manag*, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.