

Penerapan Metode *K-Means* dan *Fuzzy C-Means* untuk Pengelompokan Sekolah Menengah Pertama (SMP) di Indonesia Berdasarkan Standar Nasional Pendidikan (SNP)

Haniza Annuril Chusna dan Agnes Tuti Rumiati
Departemen Statistika, Institut Teknologi Sepuluh Nopember (ITS)
e-mail: agnes_tuti@statistika.its.ac.id

Abstrak—Standar Nasional Pendidikan merupakan salah satu upaya pemerintah untuk mencapai pemerataan mutu pendidikan. Standar Nasional Pendidikan meliputi delapan capaian yaitu Standar Kompetensi Kelulusan, Standar Isi, Standar Proses, Standar Penilaian, Standar Pendidik dan Tenaga Pendidik, Standar Sarana dan Prasarana, Standar Pengelolaan, dan Standar Pembiayaan. Penelitian ini dilakukan untuk mengelompokkan Sekolah Menengah Pertama (SMP) di Indonesia berdasarkan SNP menggunakan metode *K-Means* dan *Fuzzy C-Means*. Sebelum dilakukan analisis kluster, dilakukan imputasi missing value menggunakan regresi. Berdasarkan hasil pengelompokan menggunakan *K-Means* dan *Fuzzy C-Means*, diperoleh metode terbaik adalah *K-Means* dengan jumlah kluster optimum sebanyak empat dengan nilai icdrate sebesar 0,4826. Kemudian dilakukan pengelompokan 514 kabupaten dan kota di Indonesia menggunakan metode *K-Means* dengan kluster optimum sebanyak 3. Dimana sebagian besar kota di Pulau Jawa berada pada kluster kategori sudah memenuhi SNP sementara kluster belum memenuhi SNP mayoritas berada di Pulau Kalimantan dan Pulau Papua.

Kata Kunci—*Fuzzy C-Means*, Imputasi, *K-Means*, SMP, SNP.

I. PENDAHULUAN

INDONESIA dalam pembukaan Undang-Undang Dasar Negara Republik Indonesia tahun 1945 memiliki beberapa tujuan salah satunya adalah untuk mencerdaskan kehidupan bangsa. Salah satu cara untuk mencapai tujuan tersebut adalah melalui pendidikan. Pendidikan adalah usaha terencana untuk mewujudkan suasana belajar dan proses pembelajaran agar peserta didik secara aktif mengembangkan potensi dirinya. Untuk memudahkan kontroling sekolah dan mewujudkan pemerataan pendidikan di Indonesia, pemerintah pusat melalui Kementerian Pendidikan dan Kebudayaan membuat sebuah standar yang diterapkan di semua sekolah di bawah naungan Kementerian Pendidikan dan Kebudayaan Republik Indonesia. Standar tersebut adalah Standar Nasional Pendidikan (SNP). Terdapat tiga komponen dalam SNP yaitu nilai standar, indikator, dan sub indikator. Dimana komponen sub indikator merupakan pembentuk nilai indikator, dan nilai indikator merupakan pembentuk nilai SNP. Terdapat delapan nilai SNP yang diatur meliputi standar kompetensi kelulusan, standar isi, standar proses, standar pendidikan dan tenaga kependidikan, standar sarana dan prasarana, standar pengelolaan, standar pembiayaan pendidikan, dan standar penilaian pendidikan. Standar ini berfungsi sebagai dasar dalam perencanaan, pelaksanaan, dan pengawasan pendidikan dalam rangka mewujudkan pendidikan nasional yang bermutu serta bertujuan untuk

menjamin mutu pendidikan nasional dalam rangka mencerdaskan kehidupan bangsa dan membentuk watak serta peradaban bangsa yang bermartabat.

Pemetaan mutu adalah proses terkait kegiatan pengumpulan, pengolahan, analisis data dan informasi tentang capaian pemenuhan Standar Nasional Pendidikan dari tingkat sekolah, kabupaten/kota, provinsi, hingga nasional. Hasil dari pemetaan mutu dapat dimanfaatkan oleh sekolah, pemerintah daerah, dan pemerintah pusat sebagai acuan dalam perencanaan perbaikan dan peningkatan mutu pendidikan sesuai kewenangan masing-masing [1]. Pada tahun 2018, capaian SNP jenjang SMP di Indonesia mencapai angka 5,43 atau menuju SNP level 4. Artinya dengan menggunakan agregasi indikator, mutu pendidikan SMP belum mencapai SNP.

Nilai SNP masing-masing sekolah ditentukan oleh komponen terkecil yaitu nilai sub indikator. Capaian sub indikator merupakan agregasi dari dua sumber data pemetaan mutu yaitu data primer dan data sekunder. Dimana sumber dari dua data tersebut adalah dari PMP dan DAPODIK yang merupakan jawaban dari kuesioner yang diberikan ke sekolah. Jika pihak sekolah memberikan jawaban dengan lengkap, maka nilai SNP yang dihasilkan juga lebih valid sehingga kebijakan yang dihasilkan tepat sasaran. Namun pada kenyataannya, banyak sekolah yang tidak menjawab pertanyaan secara lengkap sehingga mempengaruhi perhitungan nilai indikator yang berdampak pula pada nilai capaian SNP. Oleh sebab itu perlu dilakukan perbaikan kualitas data sebelum dilakukan analisis yang lebih mendalam agar memberikan hasil analisis yang baik karena hasil analisis akan mempengaruhi penentuan kebijakan.

Dalam penelitian ini akan dilakukan analisis kluster untuk mengelompokkan SMP di Indonesia berdasarkan nilai SNP. Metode yang digunakan dalam penelitian ini adalah *K-Means* dan *Fuzzy C-Means*. Sebelum melakukan analisis kluster, terlebih dahulu dilakukan *pre-processing* data. Kemudian melakukan imputasi *missing value* menggunakan regresi untuk mengisi data yang hilang dan dilanjutkan dengan statistika deskriptif untuk mengetahui karakteristik data.

II. URAIAN PENELITIAN

A. Standar Nasional Pendidikan (SNP)

Standar Nasional Pendidikan (SNP) adalah kriteria minimal tentang sistem pendidikan di seluruh wilayah hukum Negara Kesatuan Republik Indonesia yang terdiri dari Standar Kompetensi Kelulusan (SKL), Standar Isi, Standar

Proses, Standar Pendidikan dan Tenaga Kependidikan, Sandar Sarana dan Prasarana, Standar Pengelolaan, Standar Pembiayaan Pendidikan, dan Standar Penilaian Pendidikan. Adapun fungsi dan tujuan dibentuknya Standar Nasional Pendidikan adalah sebagai dasar dalam perencanaan, pelaksanaan, dan pengawasan pendidikan dalam rangka mewujudkan pendidikan nasional yang bermutu yang bertujuan untuk menjamin mutu pendidikan nasional. Ketentuan mengenai Standar Nasional Pendidikan diatur dalam Peraturan Menteri Pendidikan Nasional.

Penomoran subindikator, indikator dan standar menyesuaikan tabel data. Angka sebelum titik pada penomoran sub indikator menunjukkan pada indikator mana dari standar mana sub indikator tersebut berada (misalnya sub indikator 1.2.1 merupakan bagian dari indikator ke 2 dari standar ke 1 dan sub indikator 8.1.2 merupakan bagian dari indikator ke 1 dari standar ke 8). Perhitungan nilai pencapaian SNP merupakan nilai komposit berkelanjutan mulai dari level pertanyaan hingga keseluruhan standar.

Capaian subindikator merupakan agregasi dari dua sumber data pemetaan mutu yaitu data primer dan data sekunder. Perhitungan nilai komposit sub indikator secara matematis dituliskan pada persamaan berikut:

$$U_{t,u,v} = \sum_1^n (a_{t,u,v}(n)X_{t,u,v}(n)) + \sum_1^m (a_{t,u,v}(m)Y_{t,u,v}(m))$$

dimana,

$U_{t,u,v}$ = Capaian subindikator ke- v pada indikator ke- u pada standar ke- t .

$X_{t,u,v}(n)$ = Nilai data primer ke- n untuk subindikator ke- v pada indikator ke- u pada standar ke- t .

$Y_{t,u,v}(m)$ = Nilai data sekunder ke- m untuk subindikator ke- v pada indikator ke- u pada standar ke- t .

$a_{t,u,v}(n)$ = Bobot nilai data untuk subindikator ke- v pada indikator ke- u pada standar ke- t .

Capaian indikator pada umumnya merupakan rata-rata aritmatik dari sub indikator dalam indikator yang sama. Hal tersebut berlaku pada sebagian besar indikator. Secara matematis, perhitungan nilai komposit capaian indikator dituliskan pada persamaan berikut:

$$T_{t,u} = \sum_1^v (\beta_{t,u}(v)U_{t,u}(v))$$

Keterangan:

$T_{t,u}$ = Capaian indikator ke- u pada standar ke- t .

$\beta_{t,u}(v)$ = Bobot nilai subindikator ke- v pada indikator ke- u pada standar ke- t .

Nilai akhir capaian standar dihitung di bawah ini.

$$S_t = \sum_1^u (\gamma_{t,u}(u)T_{t,u}(u))$$

dimana,

S_t = Capaian standar ke- t .

$\gamma_{t,u}(u)$ = Bobot nilai indikator ke- u pada standar ke- t dengan $v\{1,2, \dots, k\}$, $u\{1,2, \dots, l\}$ dan $t\{1,2, \dots, 8\}$.

Angka capaian subindikator hingga standar disajikan pada skala angka 0–7. Angka capaian SNP baik pada level subindikator hingga standar dikategorikan sesuai tahapan pencapaian SNP. Penentuan rentang batas atas dan bawah mengadopsi metode fungsi mean-standar deviasi dan metode manual menyesuaikan konsep tahapan pencapaian SNP.

Tabel 1.
Batas atas dan batas bawah masing-masing level

Kategori	Tahapan Pencapaian SNP				SNP (M5)
	Menuju SNP Level 1 (M1)	Menuju SNP Level 2 (M2)	Menuju SNP Level 3 (M3)	Menuju SNP Level 4 (M4)	
Batas Bawah	0,00	2,05	3,71	5,07	6,67
Batas Atas	2,04	3,70	5,06	6,66	7,00

Angka pencapaian sub indikator hingga standar disajikan pada skala 0-7. Penentuan rentang batas atas dan batas bawah dilakukan dengan mengdopsi metode fungsi *mean* standar deviasi dan metode manual menyesuaikan konsep tahapan pencapaian SNP. Pada Tabel 1 disajikan batas atas dan batas bawah masing-masing tahap pencapaian SNP menurut Kementerian Pendidikan dan Kebudayaan Republik Indonesia.

B. Pre-Processing Data

Pre-processing data adalah proses mempersiapkan data sebelum diolah agar memiliki kualitas data yang baik. Proses ini dilakukan dengan menghilangkan data yang tidak diperlukan [2]. Pada penelitian ini, *pre-processing* data yang dilakukan meliputi penghapusan sekolah yang memiliki nilai standar nol, penghapusan sekolah luar negeri, dan mengatasi *missing value*.

Missing value secara umum adalah keadaan dimana ada *value* (nilai) dari satu atau lebih variabel yang hilang/tidak tersedia untuk analisis. *Missing value* dapat menyebabkan data menjadi bias sehingga memungkinkan hasil dari nalisis data tersebut tidak valid. Terdapat beberapa perlakuan untuk mengatasi *missing value* yaitu:

1) Missing Value Kurang < 10%

Apabila *missing value* kurang < 10% bisa diabaikan atau dengan kata lain bisa dilanjutkan ke analisis selanjutnya tanpa imputasi.

2) Missing Value Lebih Dari 15%

Apabila *missing value* lebih dari 15% bisa menjadi kandidat untuk dilakukan penghapusan variabel ataupun observasi, tetapi apabila nilai *missing value* 20% hingga 30% masih dapat diatasi (tergantung dengan peneliti).

3) Missing Value > 50%

Apabila *missing value* > 50%, dapat dilakukan penghapusan data

Terdapat beberapa cara untuk melakukan imputasi *missing value* seperti menggunakan nilai rata-rata, median, estimasi regresi, dan lain-lain [3]. Imputasi *missing value* yang digunakan pada penelitian ini dilakukan menggunakan regresi. Imputasi *missing value* dengan regresi merupakan sebuah metode imputasi dengan memprediksi nilai menggunakan informasi dari data yang lengkap untuk mengisi variabel yang tidak lengkap [4].

C. K-Means

Istilah *K-Means* digunakan untuk menggambarkan algoritma yang menetapkan setiap item ke dalam *cluster* yang memiliki *centroid* terdekat (*means*). Diawali dengan mempartisi membagi objek menjadi k cluster awal. Kemudian menetapkan objek ke keseluruhan *cluster centroid* terdekat yang biasanya dihitung berdasarkan jarak *euclidian*.

Selanjutnya adalah menghitung ulang centroid untuk cluster yang menerima bek baru dan cluster yang kehilangan objek. Sebagian besar perubahan cluster terjadi realokasi yang pertama [5]. Adapun langkah-langkah dalam *K-Means* adalah sebagai berikut:

1. Menentukan k sebagai jumlah cluster yang ingin dibentuk
2. Menentukan pusat cluster awal secara acak
3. Menentukan pusat cluster (*centroid*) dari data yang ada pada masing-masing cluster yang didapatkan dari persamaan berikut:

$$C_{kj} = \frac{x_{1j} + x_{2j} + \dots + x_{nj}}{n}$$

Keterangan:

C_{kj} = Pusat cluster ke- k pada variabel ke- j , dimana $j = 1, 2, \dots, p$.

n = Banyak data pada cluster ke- k .

4. Menentukan jarak setiap objek dengan setiap centroid menggunakan perhitungan jarak *euclidian*.
5. Menghitung fungsi objektif dengan persamaan di bawah.

$$J = \sum_{i=1}^n \sum_{j=1}^k a_{ij} d(x_i, C_{kj})^2$$

6. Mengalokasikan masing-masing data ke *centroid* atau rata-rata terdekat yang dirumuskan pada persamaan berikut

$$a_{ij} = \begin{cases} 1 & \text{if } s = \min \{d(x_i, C_{kj})\} \\ 0 & \text{lainnya} \end{cases}$$

Mengulangi kembali langkah 3-6 sampai tidak ada lagi perpindahan objek atau tidak ada perubahan pada fungsi objektifnya.

D. Fuzzy C-Means

Dari *Fuzzy C-Means* adalah salah satu metode *non hierarchical clustering* yang membuat partisi optimasi. Pada metode partisi optimasi FCM, sebuah objek dapat dikategorikan ke dalam sebuah klaster, kemudian dikeluarkan lagi karena kedekatannya dengan klaster lain. FCM adalah suatu teknik pengelompokan data, dimana setiap titik dalam satu klaster ditentukan oleh derajat keanggotaan. Konsep dasar FCM adalah sebagai berikut:

1) Menentukan Pusat Cluster

Menentukan pusat cluster, dimana pusat tersebut akan menandai lokasi rata-rata untuk masing-masing cluster.

2) Memperbaiki Pusat Cluster

Memperbaiki pusat cluster dan derajat keanggotaan setiap titik data secara berulang sehingga pusat cluster bergerak menuju lokasi yang tepat. Perulangan didasarkan pada minimalisasi fungsi objektif yang menggambarkan jarak dari titik data terhadap pusat cluster yang terboboti oleh derajat keanggotaan titik data tersebut. Fungsi keanggotaan diartikan sebagai seberapa besar kemungkinan suatu data bisa menjadi anggota dalam suatu cluster. Masing-masing data mempunyai nilai keanggotaan yaitu setiap cluster yang terbentuk secara random sebagai berikut:

$$U^{(Y)} = \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ u_{21} & u_{22} & \dots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ u_{c1} & u_{i2} & \dots & u_{cn} \end{bmatrix}, u_{ij} \in [0,1]$$

Keterangan:

u_{ij} = Nilai keanggotaan pada cluster ke- i ($i = 1, 2, \dots, c$) dan objek ke- j ($j = 1, 2, \dots, n$) dengan $\sum_{i=1}^c u_{ij} = 1, j = 1, 2, \dots, n$ dan $0 < \sum_{i=1}^c u_{ij} < 1$

$U^{(Y)}$ = Matriks fungsi keanggotaan pada iterasi ke- i .

Besarnya cluster ke- i yang akan menandai lokasi rata-rata untuk masing-masing cluster dapat diperoleh dengan persamaan berikut:

$$v_i = \frac{\sum_{j=1}^n (u_{ij})^m x_j}{\sum_{j=1}^n (u_{ij})^m}$$

Keterangan:

v_i = Pusat cluster ke- i dimana $i = 1, 2, \dots, c$.

m = Pangkat pembobot

Besarnya pangkat pembobot dapat mempengaruhi fungsi u_{ij} dengan nilai $m > 1$. Pada kondisi awal, pusat cluster belum akurat karena terbentuk secara random. Perbaikan nilai pusat cluster dan nilai keanggotaan secara berulang mengakibatkan cluster bergerak menuju tempat yang tepat, sehingga pada iterasi ke- $(t+1)$ nilai keanggotaan objek ke- j pada cluster ke- i dituliskan dengan persamaan di bawah.

$$u_{ij}^{(t+1)} = \left[\sum_{k=1}^c \left(\frac{\|x_j - v_i\|^2}{\|x_j - v_k\|^2} \right)^{1/(m-1)} \right]^{-1}$$

Setelah memperbaiki pusat cluster, langkah terakhir adalah memperbarui fungsi objektif J_m yang dapat dituliskan dengan persamaan berikut.

$$J_m(U, V) = \sum_{j=1}^n \sum_{i=1}^c u_{ij}^m \|x_j - v_i\|^2$$

Iterasi akan berhenti apabila $J_m < \epsilon$, akan tetapi jika $J_m > \epsilon$, maka kembali menghitung pusat cluster yang baru. Algoritma *Fuzzy Cluster Mean* [6].

1. Menentukan beberapa hal yang harus diperhatikan adalah sebagai berikut:
 - a. Jumlah cluster (c) dengan $2 \leq c \leq n$.
 - b. Pangkat pembobot (m) dengan $m > 1$.
 - c. Maksimum iterasi.
 - d. Iterasi awal yaitu $t = 1$.
 - e. Nilai error yang diharapkan (ϵ) adalah 0.
2. Menentukan Keanggotaan U_{ij} Dimana $\sum_{i=1}^c u_{ij} = 1$
3. Menghitung Pusat Cluster v dengan $i = 1, 2, \dots, c$.
4. Memperbarui Keanggotaan Fuzzy C-Means
5. Menghitung fungsi objektif jika $J_m < \epsilon$, maka iterasi berhenti. Akan tetapi jika $J_m > \epsilon$, maka kembali pada langkah ke-3.

E. Pseudo-F Statistics

Jumlah cluster yang optimum dapat ditentukan berdasarkan kriteria nilai *Pseudo-F*. Nilai *Pseudo-F* tertinggi menunjukkan bahwa jumlah kelompok yang digunakan untuk mengelompokkan data telah optimal [7]. Persamaan yang digunakan untuk menghitung nilai *Pseudo-F* terdapat pada persamaan berikut:

$$Pseudo F = \frac{\left(\frac{R^2}{c-1} \right)}{\left(\frac{1-R^2}{n-c} \right)}$$

dimana:

$$R^2 = \frac{SST - SSW}{SST}$$

$$SST = \sum_{i=1}^n \sum_{j=1}^c \sum_{k=1}^p (x_{ijk} - \bar{x}_j)^2$$

$$SSW = \sum_{i=1}^n \sum_{j=1}^c \sum_{k=1}^p (x_{ijk} - \bar{x}_{jk})^2$$

Keterangan:

SST = Total jumlah dari kuadrat jarak sampel terhadap rata-rata keseluruhan.

SSW = Total jumlah dari kuadrat jarak sampel terhadap rata-rata kelompoknya.

n = Banyaknya sampel

c = Banyaknya klaster

p = Banyaknya variabel

x_{ijk} = Sampel ke-*i* pada variabel ke-*j* kelompok ke-*k*

x̄_j = Rata-rata seluruh sampel pada variabel ke-*j*

x̄_{jk} = Rata-rata seluruh sampel pada variabel ke-*i* dan kelompok ke-*k*.

F. Icdrate

Penentuan metode klaster terbaik dalam penelitian ini adalah dengan melihat nilai *icd rate* (*internal cluster dispersion*) *rate*. Semakin kecil nilai *icd rate*, maka hasil pengelompokannya semakin baik. Nilai *icd rate* merupakan tingkat *disperse* dalam klaster, nilai tersebut dapat dituliskan dengan persamaan di bawah [8].

$$icd\ rate = 1 - \frac{SST - SSW}{SST} = 1 - \frac{SSB}{SST} = 1 - R^2$$

dimana:

$$SSB = \sum_{j=1}^c \sum_{k=1}^p (x_{ijk} - \bar{x}_j)^2$$

Keterangan:

SST = Total jumlah dari kuadrat jarak terhadap rata-rata keseluruhan

SSW = Total jumlah dari kuadrat jarak sampel terhadap rata-rata kelompoknya

SSB = *Sum Square Between*

c = Banyaknya variabel

p = Banyaknya kelompok

x_{ijk} = Sampel ke-*i* pada variabel ke-*j* kelompok ke-*k*

x̄_j = Rata-rata seluruh sampel pada variabel ke-*j*

III. METODOLOGI PENELITIAN

A. Sumber Data

Data yang digunakan pada penelitian ini merupakan data sekunder yang diperoleh dari Direktorat Jenderal Pendidikan Dasar dan Menengah Kementerian Pendidikan dan Kebudayaan Republik Indonesia. Data tersebut mengenai nilai Standar Nasional Pendidikan, indikator, dan sub indikator dari 38.191 SMP di seluruh Indonesia pada tahun 2018.

B. Variabel Penelitian

Variabel yang digunakan untuk penelitian ini disajikan pada Tabel 2.

C. Langkah Penelitian

Langkah analisis yang digunakan dalam melakukan penelitian ini adalah sebagai berikut:

1) Mengumpulkan Data dan Studi Literatur

Mengumpulkan data dan melakukan kajian tentang metode yang digunakan yaitu *K-Means* dan *Fuzzy C-Means*.

2) Pre-Processing dan Cleaning Data

Melakukan *pre-processing* data dan *cleaning* data sebelum diolah dengan urutan sebagai berikut:

- a. Mengurutkan sekolah berdasarkan provinsi
- b. Menghapus data sekolah yang berada di luar negeri
- c. Menghapus data sekolah yang memiliki nilai nol pada SNP, indikator, dan subindikator.
- d. Menghitung jumlah *missing value* pada masing-masing nilai standar, indikator, dan sub indikator.
- e. Menghapus variabel yang memiliki jumlah *missing value* $\geq 80\%$.
- f. Memisahkan file nilai sub indikator berdasarkan indikator.
- g. Melakukan imputasi *missing value* terhadap sub indikator yang memiliki *missing value* $< 80\%$ menggunakan regresi.
- h. Menghitung nilai indikator.
- i. Menghitung kembali nilai Standar Nasional Pendidikan berdasarkan nilai indikator dan sub indikator yang sudah diimputasi.

3) Mendeskripsikan Ketercapaian SNI

Mendeskripsikan ketercapaian delapan Standar Nasional Pendidikan pada jenjang SMP di Indonesia.

4) Pengelompokan Berdasarkan Metode K-Means

Mengelompokkan SMP berdasarkan mutu pendidikan menggunakan metode *K-Means* dengan tahapan sebagai berikut:

- a. Menentukan nilai *k*, yaitu *k* = 2,3,4, dan 5.
- b. Menghitung *centroid* pada setiap nilai *k*.
- c. Menghitung nilai jarak antar sekolah pada setiap *k*.
- d. Mengelompokkan sekolah berdasarkan nilai jarak terdekat.

5) Pengelompokan Berdasarkan Metode Fuzzy C-Means

Mengelompokkan SMP berdasarkan mutu pendidikan di Indonesia dengan metode *Fuzzy C-Means*. Tahapan analisis data dengan metode tersebut adalah sebagai berikut:

- a. Menentukan jumlah klaster *k* = 2,3,4, dan 5.
- b. Menentukan sekolah yang masuk dalam keanggotaan *U_i* pada setiap *k*.
- c. Menghitung *centroid* dari keanggotaan *U_i* pada setiap *k*.
- d. Menghitung fungsi objektif.
- e. Memperbaiki keanggotaan *U_i* dengan melihat nilai fungsi objektif yang kurang dari error.
- f. Menghentikan iterasi apabila $J_m < \epsilon$ dan mengulangi langkah ke-c apabila $J_m > \epsilon$.

6) Penentuan Jumlah Cluster Optimum

Menentukan jumlah *cluster optimum* dari masing-masing metode menggunakan nilai *pseudo-f statistics*.

Tabel 2.
Variabel penelitian

Variabel	Keterangan	Skala Pengukuran
X ₁	Standar Kompetensi Kelulusan	Interval
X ₂	Standar Isi	Interval
X ₃	Standar Proses	Interval
X ₄	Standar Penilaian	Interval
X ₅	Standar Pendidik dan Tenaga Pendidik	Interval
X ₆	Standar Sarana dan Prasarana	Interval
X ₇	Standar Pengelolaan	Interval
X ₈	Standar Pembiayaan	Interval

Tabel 3.
Perbandingan nilai *Pseudo-F Statistics* dari metode *K-Means* dan *Fuzzy C-Means*

Jumlah Kluster	<i>K-Means</i>	<i>Fuzzy C-Means</i>
4	13.160,5141	7.992,2615
5	11.177,7310	6.560,0123
6	10.112,6776	5.537,9958

Tabel 4.
Perbandingan nilai *icdrate*

Metode	Kluster Optimum	<i>Icd Rate</i>
<i>K-Means</i>	4	0,4826
<i>Fuzzy C-Means</i>	4	0,6057

7) Menentukan Metode Terbaik

Menentukan metode terbaik dengan menggunakan nilai *Icdrate*. Kemudian mendeskripsikan ketercapaian delapan SNP berdasarkan kluster

8) Pengelompokan Kabupaten/Kota Berdasarkan Rata-Rata Nilai SNP

Mengelompokkan 514 kabupaten/ kota di Indonesia berdasarkan rata-rata nilai SNP tiap kabupaten/kota menggunakan metode terbaik yang didapatkan dari langkah ke-7 dengan jumlah kluster 3 dan 4.

9) Membuat Peta Kluster

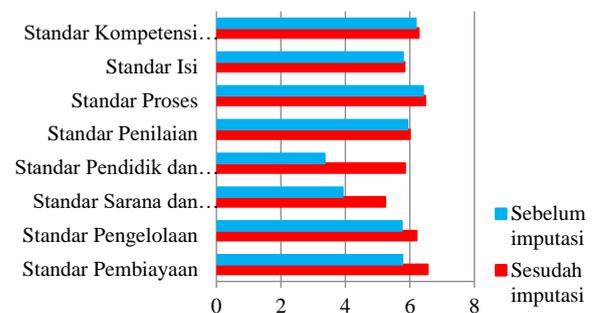
Membuat peta kluster hasil pengelompokan Kabupaten/Kota di Indonesia. Lalu menginterpretasikan hasil dan membuat kesimpulan.

IV. ANALISIS DAN PEMBAHASAN

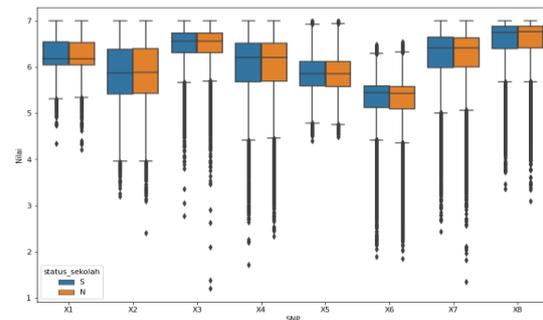
Pada bab ini akan dijelaskan hasil analisis meliputi hasil *pre-processing* data, karakteristik data mutu pendidikan SMP, pengelompokan SMP berdasarkan mutu pendidikan, penentuan jumlah kluster terbaik dan penentuan metode terbaik, serta karakteristik dari masing-masing kluster.

A. *Preprocessing* Data

Data pada penelitian ini memiliki tiga komponen utama yaitu nilai standar, indikator, dan sub indikator dari total 38.191 sekolah SMP/ sederajat di dalam dan luar negeri. Sebelum melakukan analisis kluster, dilakukan *pre-processing* data. Adapun *pre-processing* data dalam penelitian ini dilakukan dengan beberapa tahapan. Tahap pertama adalah menghapus data SMP yang berada di luar negeri sebanyak 59 sekolah dan menghapus data sekolah yang memiliki nilai nol pada nilai standar, indikator, dan sub indikator sekaligus. Jumlah Sekolah Menengah Pertama di Indonesia setelah dilakukan penghapusan adalah 36.840 sekolah. Langkah selanjutnya adalah menghitung jumlah



Gambar 1. Perbandingan rata-rata capaian snp smp sebelum dan sesudah dilakukan imputasi.



Gambar 2. Perbandingan nilai SNP berdasarkan status sekolah.



Gambar 3. Line chart rata-rata tiap variabel pada masing-masing kluster.

missing value dari masing-masing variabel nilai standar, indikator, dan sub indikator. Kemudian menghapus sub indikator dengan *missing value* $\geq 80\%$.

Missing value dalam penelitian ini diimputasi menggunakan model regresi. Penggunaan imputasi regresi dalam penelitian ini dilakukan karena variansi data hasil imputasi lebih baik daripada jika diimputasi menggunakan rata-rata maupun median. Selain itu, metode ini baik digunakan untuk data yang saling berhubungan dan tidak berkelipatan. Nilai sub indikator yang dihasilkan digunakan untuk menghitung nilai indikator dan nilai indikator yang dihasilkan digunakan untuk menghitung nilai Standar Nasional Pendidikan. Hasil perhitungan akhir nilai SNP ini akan digunakan untuk proses kluster menggunakan *K-Means* dan *Fuzzy C-Means*.

B. Karakteristik Data

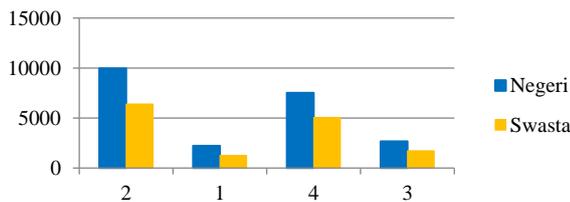
Setelah dilakukan *pre-processing* data, didapatkan data delapan nilai SNP baru dari 36.840 SMP di Indonesia yang selanjutnya akan digunakan untuk analisis kluster. Sebelum melakukan analisis kluster, dilakukan analisis statistika deskriptif untuk mengetahui karakteristik data. Berikut adalah perbandingan rata-rata capaian Standar Nasional P

Tabel 5.
Kategori klaster baru

Klaster	Kategori Klaster Baru	Jumlah Anggota	Persentase
2	Memenuhi SNP	16.371	44,44%
1	Cukup Memenuhi SNP	3.487	9,46%
4	Kurang Memenuhi SNP	12.585	34,16%
3	Belum Memenuhi SNP	4.397	11,94%

Tabel 6.
Perbandingan nilai *Pseudo-F Statistics*

Jumlah Klaster	<i>Pseudo-f</i>
3	248,4448
4	209,4643



Gambar 4. Bar chart perbandingan jumlah smp negeri dan swasta pada masing-masing klaster.

endidikan SMP di Indonesia sebelum dan sesudah imputasi *missing value*.

Berdasarkan Gambar 1, rata-rata dari delapan capaian SNP jenjang SMP mengalami kenaikan setelah dilakukan imputasi *missing value*. Standar Pendidik dan Tenaga Kependidikan mengalami kenaikan tertinggi yaitu sebesar 2,494774 atau semula berada pada tahapan Menuju SNP Level 2 (M2) menjadi tahap Menuju SNP Level 4 (M4). Standar Sarana dan Prasarana mengalami kenaikan terbesar kedua yaitu sebesar 1,322191 atau yang semula berada pada tahapan Menuju SNP Level 3 (M3) menjadi Menuju SNP Level 4 (M4). Sementara lima standar lain yaitu Standar Kompetensi Kelulusan, Standar Isi, Standar Proses, Standar Penilaian, dan Standar Pengelolaan mengalami kenaikan rata-rata capaian setelah dilakukan imputasi *missing value* namun masih berada di tahapan atau level yang sama.

Kenaikan rata-rata yang tinggi dari Standar Pendidik dan Tenaga Kependidikan serta Standar Sarana dan Prasarana disebabkan karena kedua standar tersebut memiliki *missing value* yang terbanyak jika dibandingkan dengan standar lain. Kedelapan capaian SNP mengalami kenaikan rata-rata dimana hal ini membuktikan bahwa imputasi *missing value* dapat meningkatkan nilai SNP. Meskipun demikian, rata-rata dari delapan nilai standar masih di bawah 6,66 atau belum mencapai SNP. Visualisasi perbandingan delapan nilai SNP berdasarkan status sekolah disajikan pada Gambar 2.

Berdasarkan Gambar 2, Standar Kompetensi Kelulusan, Standar Isi, Standar Proses, Standar Pengelolaan, dan Standar Pembiayaan pada SMP negeri memiliki jangkauan lebih lebar dibandingkan SMP swasta yang dibuktikan dengan keberadaan titik outlier yang lebih jauh. Secara visual, tidak ada perbedaan yang signifikan antara nilai standar SMP negeri dengan SMP swasta. Hal ini mengartikan bahwa SMP negeri maupun swasta di Indonesia memiliki kualitas yang cenderung sama.

C. Analisis Klaster Menggunakan *K-Means* dan *Fuzzy C-Means*

Setelah dilakukan imputasi *missing value* dan perhitungan nilai indikator, didapatkan nilai SNP baru dari 36.840 SMP di Indonesia. Nilai ini digunakan untuk mengelompokkan SMP menggunakan metode *K-Means* dan *Fuzzy C-Means*. Adapun jumlah klaster yang digunakan dalam penelitian ini adalah 4,5, dan 6. Perbandingan nilai *pseudo-f statistics* dari kedua metode disajikan pada Tabel 3.

Jumlah klaster optimum ditentukan dari nilai *pseudo-f statistics* yang terbesar. Tabel 3 menunjukkan perbandingan nilai *pseudo-f statistics* dari masing-masing metode dengan jumlah klaster 4,5, dan 6. Pada metode *K-Means* dan *Fuzzy C-Means*, nilai *pseudo-f statistics* yang terbesar adalah 13.160,5141 dan 7.992,2615 dengan jumlah klaster 4. Dengan demikian dapat disimpulkan bahwa jumlah klaster optimum dari metode *K-Means* dan *Fuzzy C-Means* adalah 4. Metode terbaik ditentukan berdasarkan nilai *icdrate* yang terkecil. Perbandingan nilai *icdrate* pada metode *K-Means* dan *Fuzzy C-Means* disajikan dalam Tabel 4.

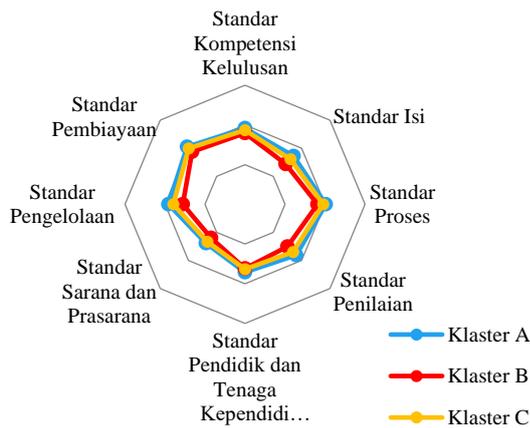
Tabel 4 menunjukkan perbandingan nilai *icdrate* dari metode *K-Means* dan *Fuzzy C-Means*. Dapat diketahui bahwa nilai *icdrate* yang terkecil adalah 0,4826 dengan metode *K-Means*. Dengan demikian dapat disimpulkan bahwa metode terbaik dalam pengelompokan SMP di Indonesia berdasarkan mutu pendidikan (SNP) adalah metode *K-Means*.

D. Karakteristik Tiap Kelompok

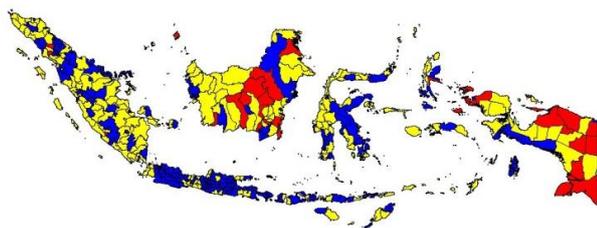
Setelah dilakukan analisis klaster menggunakan metode *K-Means* dan *Fuzzy C-Means*, diperoleh hasil bahwa metode *K-Means* lebih baik dibandingkan metode *Fuzzy C-Means* dalam pengelompokan SMP di Indonesia berdasarkan nilai SNP. Jumlah klaster optimum dari metode tersebut adalah 4. Perbandingan rata-rata dari delapan variabel pada masing-masing klaster secara visual disajikan pada Gambar 3.

Berdasarkan Gambar 3 dapat diketahui bahwa garis berwarna hijau (klaster 2) berada pada paling atas untuk semua variabel. Sedangkan garis berwarna kuning (klaster 1) berada di bawah garis hijau untuk tujuh variabel sementara satu variabel lain yaitu X₆ atau Standar Sarana dan Prasarana memiliki nilai terendah dibandingkan semua klaster. Hal ini mengindikasikan bahwa SMP yang termasuk dalam klaster 1 memiliki sarana dan prasarana yang paling buruk dibanding tiga klaster lain. Garis berwarna merah milik klaster 3 berada paling bawah pada 7 nilai SNP. Berdasarkan Gambar 4 dapat disimpulkan bahwa urutan klaster dari yang tertinggi ke terendah adalah klaster 2, klaster 1, klaster 4, dan klaster 3. Meskipun rata-rata Standar Sarana dan Prasarana pada klaster 1 memiliki nilai terendah dibanding rata-rata tiga klaster lain, namun klaster ini tetap berada di peringkat 2 karena rata-rata dari tujuh standar lain berada urutan ke-2. Selanjutnya dilakukan pemeringkatan dan penamaan klaster baru. Pemeringkatan dan penamaan klaster baru disajikan pada Tabel 5.

Berdasarkan Tabel 5, jumlah anggota klaster terbanyak adalah dari klaster 2 atau kategori memenuhi SNP dengan 44,44% atau sebanyak 16.371 sekolah. Artinya berdasarkan kategori klaster baru, sebanyak 44,44% SMP di Indonesia memiliki kualitas yang baik. Sementara 4.397 SMP atau 11,94% berada pada klaster 3 atau pada kategori belum



Gambar 5. Radar chart perbandingan rata-rata masing-masing variabel.



Gambar 6. Peta kluster kabupaten dan kota di Indonesia berdasarkan rata-rata SNP.

memenuhi SNP. Status sekolah dari masing-masing kluster disajikan pada Gambar 4.

Gambar 4 merupakan bar chart perbandingan jumlah SMP negeri dan swasta pada masing-masing kaster dengan urutan menyesuaikan dengan pemeringkatan pada Tabel 4. Jumlah sekolah yang digunakan dalam penelitian ini adalah 36.840 dimana 60,73% atau sekitar 22.374 sekolah berstatus negeri dan 39,27% atau 14.466 sekolah berstatus swasta. Berdasarkan Gambar 4 dapat diketahui bahwa jumlah sekolah negeri dari keempat kluster lebih dominan daripada sekolah swasta.

E. Pengelompokan Kabupaten dan Kota di Indonesia Berdasarkan Rata-Rata Nilai SNP Menggunakan Metode K-Means

Setelah melakukan pengelompokan SMP di Indonesia menggunakan metode *K-Means* dan *Fuzzy C-Means*, dilakukan pengelompokan kabupaten dan kota di Indonesia berdasarkan rata-rata nilai SNP menggunakan yaitu metode *K-Means*. Tujuan dari pengelompokan ini adalah untuk melihat visualisasi kluster dan persebaran mutu pendidikan dari 514 kabupaten dan kota di Indonesia. Variabel dari pengelompokan ini adalah rata-rata delapan nilai SNP dari 514 kabupaten dan kota di Indonesia. Adapun jumlah kluster yang digunakan dalam pengelompokan ini adalah 3 dan 4. Penentuan jumlah kluster optimum menggunakan nilai *pseudo-f statistics* yang terbesar. Berikut merupakan perbandingan nilai *pseudo-f statistics* dari tiap kluster.

Berdasarkan Tabel 6 dapat diketahui bahwa nilai *pseudo-f* pada kluster 3 lebih besar daripada kluster 4 sehingga dapat disimpulkan bahwa jumlah kluster optimum untuk pengelompokan 514 kabupaten dan kota di Indonesia berdasarkan rata-rata SNP menggunakan metode *K-Means* adalah 3. Setelah mendapatkan kluster optimum, selanjutnya

Tabel 7. Perbandingan rata-rata

Variabel	Kluster A	Kluster B	Kluster C
X ₁	6.3352 (M4)	6.0869 (M4)	6.2365 (M4)
X ₂	5.9336 (M4)	5.3734 (M4)	5.7163 (M4)
X ₃	6.5358 (M4)	6.1263 (M4)	6.4057 (M4)
X ₄	6.1186 (M4)	5.5042 (M4)	5.8929 (M4)
X ₅	5.9198 (M4)	5.7328 (M4)	5.7674 (M4)
X ₆	5.2677 (M4)	4.8975 (M3)	5.1677 (M4)
X ₇	6.3270 (M4)	5.6128 (M4)	6.0895 (M4)
X ₈	6.5981 (M4)	6.2478 (M4)	6.4867 (M4)

Tabel 8. Pemeringkatan dan penamaan kluster kabupaten dan kota

Kluster	Kategori	Jumlah	Persentase
A	Sudah Memenuhi SNP	257	50%
C	Cukup Memenuhi SNP	216	42%
B	Belum Memenuhi SNP	41	8%

adalah mendeskripsikan karakteristik dari masing-masing kluster. Visualisasi perbandingan rata-rata dari masing-masing variabel pada setiap kluster disajikan pada Gambar 5.

Gambar 6 merupakan radar chart dari perbandingan rata-rata masing-masing variabel pada setiap kluster dimana jika letak radar semakin di luar, maka nilainya semakin besar begitu pula sebaliknya. Berdasarkan Gambar 6 dapat diketahui bahwa radar berwarna biru (kluster A) berada paling luar dan radar merah (kluster B) berada pada bagian paling dalam. Sementara radar oranye (kluster C) berada diantara radar kluster A dan kluster B. Sehingga secara visual dapat disimpulkan urutan kluster yang memiliki rata-rata tertinggi ke terendah secara urut adalah kluster A, kluster C, dan kluster B. Perbandingan rata-rata dari ketiga kluster secara rinci disajikan pada Tabel 7.

Berdasarkan Tabel 7 dapat diketahui bahwa rata-rata dari masing-masing variabel pada setiap kluster belum ada yang mencapai SNP. Pada kluster A, Standar Pembiayaan memiliki rata-rata tertinggi sebesar 6,5981 atau berada pada tahapan Menuju SNP Level 4 (M4). Sementara Standar Sarana dan Prasarana pada kluster A memiliki nilai terendah sebesar 5,2677 atau berada pada tahap Menuju SNP Level 4 (M4). Enam standar lain juga berada di tahapan yang sama yaitu M4.

Sementara pada kluster B, Standar Sarana dan Prasarana bernilai 4,8975 atau berada pada tahapan Menuju SNP Level 3 (M3) dan tujuh nilai standar lain memiliki rata-rata di atas 5 atau berada di level M4. Pada kluster C, rata-rata kedelapan variabel berada pada tahapan Menuju SNP level 4 (M4) namun nilai rata-rata variabel pada kluster C lebih rendah dibandingkan dengan rata-rata variabel dari kluster A. Sehingga berdasarkan Gambar 5 dan Tabel 7 dapat disimpulkan bahwa urutan rata-rata kluster dari yang tertinggi ke terendah adalah kluster A, kluster C, dan kluster B. Selanjutnya dilakukan pemeringkatan dan penamaan kluster baru. Pemeringkatan dan penamaan ini disajikan pada Tabel 8.

Berdasarkan Tabel 8 dapat diketahui bahwa 50% atau 257 kabupaten dan kota di Indonesia berada pada kluster A atau kategori sudah memenuhi SNP. Hal ini mengindikasikan bahwa setengah dari kabupaten dan kota di Indonesia memiliki kualitas SMP yang baik. Kemudian terdapat 8%

atau sekitar 41 kabupaten dan kota di Indonesia yang berada pada kluster B atau pada kategori belum memenuhi SNP. Setelah dilakukan pemeringkatan kluster, dilakukan visualisasi untuk melihat persebaran dari 514 kabupaten dan kota di Indonesia. Peta kluster 514 kabupaten dan kota di Indonesia menggunakan metode *K-Means* disajikan pada Gambar 6.

Berdasarkan Gambar 6 dapat diketahui bahwa mayoritas kabupaten dan kota di Pulau Jawa termasuk dalam kluster A atau sudah mencapai SNP. Kabupaten dan kota di Pulau Sumatra mayoritas berada pada kluster C atau dalam kategori cukup memenuhi SNP. Sementara pada Pulau Sulawesi, terlihat sebagian berada pada kluster A dan dan kluster C. Kabupaten dan kota yang termasuk pada kluster B atau kategori belum memenuhi SNP banyak tersebar di Pulau Kalimantan dan Papua. Hal ini menunjukkan adanya ketimpangan kualitas pendidikan dimana SMP yang berada di Pulau Jawa mayoritas memiliki kualitas pendidikan yang baik sedangkan SMP yang berada di Pulau Papua dan Kalimantan memiliki kualitas yang masih kurang.

V. KESIMPULAN

Berdasarkan hasil *pre-processing* data, jumlah sekolah yang akan digunakan untuk analisis kluster sebanyak 36.840 sekolah. Sementara jumlah komponen SNP yang dihasilkan setelah penghapusan sub indikator adalah sebanyak 169 komponen dengan rincian 8 nilai SNP, 29 indikator, dan 132 sub indikator. Standar Pendidik dan Tenaga Pendidik mengalami kenaikan rata-rata tertinggi dan meningkatkan level pencapaian SNP dari M2 menjadi M4.

Pengelompokan menggunakan metode *K-Means* dan *Fuzzy C-Means* dengan jumlah kluster sebanyak 4,5, dan 6 memberikan hasil bahwa 4 kluster merupakan jumlah kluster optimum dari dua metode tersebut. Dari kedua metode

tersebut, metode *K-Means* memiliki nilai *icdrate* terkecil yang menunjukkan bahwa metode *K-Means* lebih baik dibandingkan dengan *Fuzzy C-Means*. Urutan pemeringkatan dari keempat kluster adalah kluster 2 (memenuhi SNP), kluster 1 (cukup memenuhi SNP), kluster 4 (kurang memenuhi SNP), dan kluster 3 (belum memenuhi SNP) dengan jumlah anggota berturut-turut adalah 16.371, 3.486, 12.585, dan 4.397 dimana jumlah sekolah negeri lebih dominan daripada sekolah swasta dari keempat kluster.

Jumlah kluster optimum pada pengelompokan kabupaten dan kota di Indonesia menggunakan *K-Means* adalah 3 dengan urutan kluster A (sudah memenuhi SNP) sebanyak 257 SMP, kluster C (kurang memenuhi SNP) sebanyak 216 SMP, dan kluster B (belum memenuhi SNP) sebanyak 41 SMP. Adapun kabupaten dan kota di Pulau Jawa mayoritas berada pada kluster A sementara kluster B tersebar paling banyak di Pulau Kalimantan dan Papua.

DAFTAR PUSTAKA

- [1] Dikdasmen, *Pemetaan Mutu Pendidikan: Profil Mutu Pendidikan Dasar dan Menengah di Indonesia 2016*. Jakarta: Kementerian Pendidikan dan Kebudayaan, 2016.
- [2] D. Pyle, *Data Preparation for Data Mining*. California: Morgan Kaufmann Publishers, 1999.
- [3] S. Santoso, *Statistik Parametrik, Konsep dan Aplikasi dengan SPSS*. Jakarta: PT. Gramedia, 2010.
- [4] C. G. Enders, *Applied Missing Data Analysis*. New York: The Guilford Press, 2010.
- [5] R. A. Johnson and D. W. Wicherin, *Applied Multivariate Statistical Analysis*, 6th ed. New Jersey: Person Prentice Hall, 2007.
- [6] W. Wang and Y. Zhang, "On fuzzy cluster validity indices," *Fuzzy sets Syst.*, vol. 158, no. 19, pp. 2095–2117, 2007, doi: <https://doi.org/10.1016/j.fss.2007.03.004>.
- [7] J. Hair, *Multivariate Data Analysis*, 7th ed. New York: Pearson Education, Inc, 2010.
- [8] S. A. Mingoti and J. O. Lima, "Comparing SOM neural network with fuzzy c-means, k-means and traditional hierarchical clustering algorithm," *Eur. Journal Oper. Res.*, vol. 174, no. 3, pp. 1742–1759, 2006.