TUGAS AKHIR – SS141501

PENERAPAN METODE DATA CAMPURAN *ENSEMBLE K-MODES* DAN *SIMILARITY WEIGHT AND FILTER METHOD* (SWFM) PADA PENGELOMPOKAN KABUPATEN/KOTA DI JAWA TIMUR BERDASARKAN INDIKATOR DAERAH TERTINGGAL

LARASATI DEWINTA PARAMARDINI
NRP 062114 4000 0058

Dosen Pembimbing
Dr. Santi Puteri Rahayu, M.Si.
M. Sjahid Akbar, S.Si., M.Si.

PROGRAM STUDI SARJANA
DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA, KOMPUTASI, DAN SAINS DATA
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA 2018

TUGAS AKHIR – SS141501

PENERAPAN METODE DATA CAMPURAN *ENSEMBLE K-MODES* DAN *SIMILARITY WEIGHT AND FILTER METHOD* (SWFM) PADA PENGELOMPOKAN KABUPATEN/KOTA DI JAWA TIMUR BERDASARKAN INDIKATOR DAERAH TERTINGGAL

LARASATI DEWINTA PARAMARDINI
NRP 062114 4000 0058

Dosen Pembimbing
Dr. Santi Puteri Rahayu, M.Si.
M. Sjahid Akbar, S.Si., M.Si.

PROGRAM STUDI SARJANA
DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA, KOMPUTASI, DAN SAINS DATA
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA 2018

FINAL PROJECT – SS141501

IMPLEMENTATION OF MIXED DATA METHODS ENSEMBLE K-MODES AND SIMILARITY WEIGHT AND FILTER METHOD (SWFM) TO CLUSTERING DISTRICT IN EAST JAVA BASED ON UNDERDEVELOPED DISTRICT INDICATORS

LARASATI DEWINTA PARAMARDINI
NRP. 062114 4000 0058

Supervisors
Dr. Santi Puteri Rahayu, M.Si
M. Sjahid Akbar, S.Si., M.Si

UNDERGRADUATE PROGRAMME
DEPARTMENT OF STATISTICS
FACULTY OF MATHEMATICS, COMPUTING, AND DATA SCIENCE
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA 2018

LEMBAR PENGESAHAN

PENERAPAN METODE DATA CAMPURAN *ENSEMBLE K-MODES* DAN *SIMILARITY WEIGHT AND FILTER METHOD* (SWFM) PADA PENGELOMPOKAN KABUPATEN/KOTA DI JAWA TIMUR BERDASARKAN INDIKATOR DAERAH TERTINGGAL

TUGAS AKHIR

Diajukan Untuk Memenuhi Salah Satu Syarat
Memperoleh Gelar Sarjana Sains
pada
Program Studi Sarjana Departemen Statistika
Fakultas Matematika, Komputasi, dan Sains Data
Institut Teknologi Sepuluh Nopember

Oleh :
Larasati Dewinta Paramardini
NRP. 062114 4000 0058

Disetujui oleh Pembimbing :
Dr. Santi Puteri Rahayu, M.Si ( )
NIP. 19750115 199903 2 003

M. Sjahid Akbar, S.Si., M.Si ( )
NIP. 19720705 199802 1 001

Mengetahui,
Kepala Departemen

Dr. Suhartono
NIP. 19710929 199512 1 001

SURABAYA, JULI 2018

(Halaman Ini Sengaja Dikosongkan)

PENERAPAN METODE DATA CAMPURAN *ENSEMBLE K-MODES* DAN *SIMILARITY WEIGHT AND FILTER METHOD* (SWFM) PADA PENGELOMPOKAN KABUPATEN/KOTA DI JAWA TIMUR BERDASARKAN INDIKATOR DAERAH TERTINGGAL

Nama Mahasiswa	**:**	**Larasati Dewinta Paramardini**
NRP	**:**	**062114 4000 0058**
Departemen	**:**	**Statistika**
Dosen Pembimbing	**:**	**Dr. Santi Puteri Rahayu, M.Si**
		M. Sjahid Akbar, S.Si., M.Si

Abstrak

Sebagian besar metode multivariat analisis klaster hanya berfokus pada data skala numerik, sedangkan sering kali data yang dimiliki berupa skala campuran (numerik dan kategorik). Ensemble K-Modes dan Similarity Weight and Filter Method (SWFM) merupakan metode analisis klaster yang dapat menangani data berskala campuran. Keduanya tersusun dari hasil analisis klaster data numerik dan kategorik yang digabungkan. Indikator ketertinggalan daerah di kabupaten/kota Jawa Timur terdiri atas data campuran berskala numerik dan kategorik. Oleh karena itu untuk mengetahui daerah apa saja yang tergolong tertinggal dilakukan pengelompokan dengan metode analisis klaster data campuran. Dengan membandingkan hasil pengelompokan, diperoleh hasil bahwa menggunakan metode ensemble k-modes memiliki nilai akurasi yang lebih besar dibandingkan SWFM yaitu 72.638%. Berdasarkan karakteristik pada tiap klaster terbentuk, klaster 2 yang terdiri dari 39.47% kabupaten/kota di Jawa Timur menunjukkan kondisi yang paling rendah sehingga dianggap sebagai kelompok daerah sangat tertinggal. Daerah lainnya dengan status tertinggal terdapat pada klaster 5 yang terdiri dari 28.95% kabupaten/kota di Jawa Timur.

Kata Kunci : Daerah Tertinggal, Ensembel K-Modes, Skala Campuran, SWFM

(Halaman Ini Sengaja Dikosongkan)

IMPLEMENTATION OF MIXED DATA METHODS ENSEMBLE K-MODES AND SIMILARITY WEIGHT AND FILTER METHOD (SWFM) TO CLUSTERING DISTRICT IN EAST JAVA BASED ON UNDERDEVELOPED DISTRICT INDICATORS

Name	**:**	**Larasati Dewinta Paramardini**
Student Number	**:**	**062114 4000 0058**
Department	**:**	**Statistics**
Supervisors	**:**	**Dr. Santi Puteri Rahayu, M.Si**
		M. Sjahid Akbar, S.Si., M.Si

Abstract

Most of the multivariate cluster analysis methods only focus on numerical scale data, whereas the data are often in the form of mixed scales (numerical and categorical). Ensemble K-Modes and Similarity Weight and Filter Method (SWFM) is a cluster analysis that could handle mixed data. Both are composed by combined the result of numerical and categorical group data analysis. Underdeveloped district indicator in East Java consists of numerical and categorical mixed data. Therefore, to know the districts of regions are classified as underdeveloped district is done by mixed data clustering analysis method. By comparing the results of clustering, it was found that using k-modes ensemble has a greater accuracy value than SWFM that is 72.638%. Based on the characteristics in each cluster, cluster 2 composed by 39.47% of the district in East Java shows the lowest condition so that it is considered as a very underdeveloped district group. Others area with underdeveloped status are found in cluster 5 that composed by 28.95% of the district in East Java.

Keywords : Ensemble K-Modes, Mixed scale, SWFM, Underdeveloped District

(Halaman Ini Sengaja Dikosongkan)

KATA PENGANTAR

Puji syukur yang kehadirat Allah SWT, Tuhan Yang Maha Esa. Berkat rahmat dan ridho-Nya penulis dapat menyelesaikan laporan Tugas Akhir yang berjudul **"Penerapan Metode Data Campuran *Ensemble K-Modes* dan *Similarity Weight and Filter Method* (SWFM) Pada Pengelompokan Kabupaten/Kota di Jawa Timur Berdasarkan Indikator Daerah Tertinggal"** dengan lancar.

Keberhasilan penyusunan Tugas Akhir ini tidak lepas dari banyaknya bantuan dan dukungan yang diberikan dari berbagai pihak. Penulis mengucapkan terima kasih kepada:

1. Ayah dan Ibu tercinta, adik tersayang serta keluarga yang selalu memberikan dukungan, kasih sayang, semangat, dan doa yang tidak pernah putus kepada penulis.
2. Bapak Dr. Suhartono selaku Kepala Departemen Statistika dan Bapak Dr. Sutikno, M.Si selaku Koordinator Program Studi Sarjana yang telah memberikan fasilitas untuk kelancaran penyelesaian Tugas Akhir.
3. Ibu Dr. Santi Puteri Rahayu, M.Si selaku dosen pembimbing serta dosen wali yang dengan sabar telah memberikan bimbingan, ilmu, saran, dan dukungan selama masa perkuliahan dan penyusunan Tugas Akhir.
4. Bapak M. Sjahid Akbar, S,Si., M.Si selaku co-dosen pembimbing yang dengan sabar membimbing dari awal hingga akhir penyusunan Tugas Akhir.
5. Bapak Dr.rer.pol Dedy Dwi Prastyo, S.Si., M.Si dan Bapak Dr.rer.pol Heri Kuswanto, S.Si., M.Si selaku dosen penguji yang telah memberikan banyak bantuan dan saran untuk Tugas Akhir ini.
6. Teman-teman dekat penulis Lita, Safitri, Dwita, Desi, dan Anisa yang selalu memberikan semangat, dukungan, serta doa bagi penulis.
7. Teman-teman Statistika ITS angkatan 2014, keluarga Σ25 RΣSPECT yang senantiasa membantu dan selalu berjuang bersama.

8. Semua pihak yang telah memberikan dukungan baik moril maupun materiil yang tidak dapat penulis sebutkan satu persatu.

Penulis berharap hasil Tugas Akhir ini dapat bermanfaat bagi pembaca. Penulis menyadari bahwa Tugas Akhir ini masih belum sempurna, sehingga kritik atau saran dari pembaca sangat berguna bagi penyempurnaan laporan ini.

Surabaya, Juli 2018

Penulis

DAFTAR ISI

Halaman

HALAMAN JUDUL iii
LEMBAR PENGESAHAN vi
ABSTRAK ix
ABSTRACT xi
KATA PENGANTAR xiii
DAFTAR ISI xv
DAFTAR GAMBAR xvii
DAFTAR TABEL xix
DAFTAR LAMPIRAN xxi
BAB I PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Rumusan Masalah 6
1.3 Tujuan Penelitian 7
1.4 Manfaat Penelitian 7
1.5 Batasan Masalah 7
BAB II TIJAUAN PUSTAKA 9
2.1 Statistika Deskriptif 9
2.2 Analisis Faktor dan Kelayakan 9
2.2.1 Pemeriksaan Kecukupan Variabel 10
2.2.2 Uji Korelasi Variabel 11
2.2.3 Analisis Faktor 14
2.3 Analisis Kelompok (*Cluster Analysis*) 16
2.3.1 *Fuzzy C-Means* (FCM) 16
2.3.2 *K-Modes* 27
2.3.3 *Ensemble Cluster* 32
2.4 *One Way Analysis of Variance* (ANOVA) 36
2.5 *One Way Multivariate Analysis of Variance* (MANOVA) 36
2.6 Uji *Kruskal Wallis* 39
2.7 Daerah Tertinggal 40
BAB III METODOLOGI PENELITIAN 43
3.1 Sumber Data 43
3.2 Variabel Penelitian 43
3.3 Langkah Analisis 46
BAB IV ANALISIS DAN PEMBAHASAN 51
4.1 Karakteristik Kabupaten/Kota di Jawa Timur 51

4.2 Pengelompokan Kabupaten/Kota di Jawa Timur dengan *Ensemble K-Modes* 59
4.2.1 Pereduksian Variabel Numerik dengan Analisis Faktor ... 59
4.2.2 Pengelompokan Variabel Numerik dengan *Fuzzy C-Means* 63
4.2.3 Pengelompokan Variabel Kategorik dengan *K-Modes* 65
4.2.4 Pengelompokan Variabel Campuran dengan *Ensemble K-Modes* 66
4.3 Pengelompokan Kabupaten/Kota di Jawa Timur dengan *Similarity Weighted and Filter Method* (SWFM) 68
4.4 Hasil Pengelompokan Data Campuran Terbaik Beserta Karakteristiknya 69
BAB V KESIMPULAN DAN SARAN.....81
5.1 Kesimpulan..... 81
5.2 Saran..... 82
DAFTAR PUSTAKA 85
LAMPIRAN.....89

DAFTAR GAMBAR

Halaman

Gambar 2.1 Representasi Linear Naik 18
Gambar 2.2 Representasi Linear Turun 18
Gambar 2.3 Representasi Kurva Segitiga 19
Gambar 2.4 Representasi Kurva Trapesium 20
Gambar 2.5 Skema Metode *Ensemble K-Modes* 32
Gambar 2.6 Kriteria Ketertinggalan Daerah 41
Gambar 3.1 Diagram Alir Penelitian 48
Gambar 4.1 Karakteristik Rasio Desa dengan Permukaan Jalan Terluas Aspal 52
Gambar 4.2 Karakteristik Rasio Rumah Sakit 53
Gambar 4.3 Karakteristik Rasio SMA/MA 53
Gambar 4.4 Karakteristik Persentase Rumah Tangga Pengguna Listrik 54
Gambar 4.5 Karakteristik Rasio Lahan Kritis 54
Gambar 4.6 Karakteristik Persentase Penduduk Miskin 55
Gambar 4.7 Karakteristik Pengeluaran Penduduk per Kapita 55
Gambar 4.8 Karakteristik Angka Harapan Hidup 56
Gambar 4.9 Karakteristik Angka Melek Huruf 56
Gambar 4.10 Karakteristik Persentase Pengangguran 57
Gambar 4.11 Karakteristik Keberadaan Kawasan Industri 57
Gambar 4.12 Karakteristik Tingkat Desentralisasi Fiskal 58
Gambar 4.13 Karakteristik Sumber Air Minum 58
Gambar 4.14 *Scatter Plot* Data Penelitian 60
Gambar 4.15 *Scree Plot* 62
Gambar 4.16 Pemetaan Hasil Pengelompokan dengan *Ensemble K-Modes* 67
Gambar 4.17 Pemetaan Hasil Pengelompokan SWFM 68
Gambar 4.18 Perbandingan Karakteristik Kriteria Infrastruktur 71
Gambar 4.19 Perbandingan Karakteristik Rasio Lahan Kritis 72
Gambar 4.20 Perbandingan Karakteristik Kriteria Ekonomi 72
Gambar 4.21 Perbandingan Karakteristik Kriteria Sumber Daya Manusia 73

Gambar 4.22 Karakteristik Hasil Pengelompokan pada Variabel Keberadaan Kawasan Industri 76
Gambar 4.23 Karakteristik Hasil Pengelompokan pada Variabel Tingkat Desentralisasi Fiskal..... 77
Gambar 4.24 Karakteristik Hasil Pengelompokan pada Variabel Sumber Air Minum..... 77

DAFTAR TABEL

Halaman

Tabel 2.1 Ilustrasi Perhitungan KMO 10
Tabel 2.2 Ilustrasi Data Pemeriksaan Normal Multivariat 13
Tabel 2.3 Ilustrasi Data Numerik 20
Tabel 2.4 Hasil Perhitungan Fungsi Keanggotaan Linier Naik 2 Klaster 21
Tabel 2.5 Hasil Perhitungan Fungsi Keanggotaan Linier Naik 3 Klaster 22
Tabel 2.6 Hasil Perhitungan dengan Fungsi Keanggotaan Kurva Segitiga 23
Tabel 2.7 Ilustrasi Data Kategorik 28
Tabel 2.8 Hasil Perhitungan Jarak 29
Tabel 2.9 Hasil Penentuan Pengelompokan 29
Tabel 2.10 Ilustrasi Hasil Penggabungan (*Ensemble*) 33
Tabel 3.1 Variabel Penelitian Numerik 43
Tabel 3.2 Variabel Penelitian Kategorik 44
Tabel 3.3 Struktur Data 46
Tabel 4.1 Statistika Deskriptif Variabel Numerik 51
Tabel 4.2 Uji Normal Multivariat 60
Tabel 4.3 Uji *Bartlett Sphericity* 60
Tabel 4.4 Nilai Eigen 61
Tabel 4.5 Matriks Komponen 62
Tabel 4.6 *Pseudo Fstatistics* 63
Tabel 4.7 *ICD Rate Fuzzy C-Means* Fungsi Keanggotaan Linier Turun 64
Tabel 4.8 Anggota Kelompok 5 Klaster *Fuzzy C-Means* 64
Tabel 4.9 Nilai Akurasi *K-Modes* 65
Tabel 4.10 Anggota Kelompok 5 Klaster *K-Modes* 65
Tabel 4.11 Nilai Akurasi *Ensemble K-Modes* 66
Tabel 4.12 Anggota Kelompok 5 Klaster *Ensemble K-Modes* 67
Tabel 4.13 Nilai Akurasi SWFM 68
Tabel 4.14 Anggota Kelompok 5 Klaster SWFM 69

Tabel 4.15 Perbandingan Nilai Akurasi Metode Pengelompokan Data Campuran 69
Tabel 4.16 Uji Homogenitas *Box's M Test* 70
Tabel 4.17 Hasil Pengujian ANOVA *One-way*..... 73
Tabel 4.18 Hasil Pengujian *Kruskal Wallis* 74
Tabel 4.19 Karakteristik Hasil Pengelompokan Tiap Variabel 75
Tabel 4.20 Karakteristik Ketertinggalan pada Tiap Klaster 78

DAFTAR LAMPIRAN

Halaman

Lampiran 1 Data Penelitian Variabel Numerik 89
Lampiran 2 Data Penelitian Variabel Kategorik..... 90
Lampiran 3 *Output Mshapiro test* 91
Lampiran 4 Syntax Pengelompokan Numerik *Fuzzy C-Means* . 92
Lampiran 5 Syntax Pengelompokan Kategorik *K-Modes*..... 94
Lampiran 6 Syntax Pengelompokan Campuran SWFM..... 95
Lampiran 7 Matriks Korelasi 96
Lampiran 8 Pemeriksaan KMO dan *Bartlett Test*..... 96
Lampiran 9 Persentase Variansi Kumulatif 97
Lampiran 10 *Rotated Component Matrix*..... 97
Lampiran 11 Hasil Pengelompokan Numerik *Fuzzy C-Means* 3 Variabel 5 Klaster 98
Lampiran 12 Hasil Pengelompokan Kategorik dengan *K-Modes* 5 Klaster..... 99
Lampiran 13 Data dan Hasil Pengelompokan dengan *Ensemble K-Modes* dan SWFM 100
Lampiran 14 *Output* Pengujian Beda *Mean* MANOVA..... 101
Lampiran 15 *Output* Pengujian ANOVA 102
Lampiran 16 Output Uji *Kruskal Wallis* 105
Lampiran 17 Surat Keterangan Data Penelitian 106

(Halaman Ini Sengaja Dikosongkan)

BAB I
PENDAHULUAN

1.1 Latar Belakang

Analisis kelompok (*cluster analysis*) merupakan salah satu teknik multivariat yang bertujuan untuk mengelompokan objek-objek yang memiliki kesamaan karakteristik. Sehingga sebuah kelompok yang dihasilkan akan terdiri dari objek yang memiliki kesamaan paling dekat (Johnson & Wichern, 2007). Ciri dari sebuah kelompok yang baik adalah memiliki homogenitas yang tinggi antar objek dalam satu kelompok dan memiliki heterogenitas yang tinggi dengan objek di kelompok lain. Analisis kelompok dibedakan menjadi dua, yaitu metode hirarki dan metode non hirarki. Metode yang termasuk dalam pengelompokan hirarki dan biasa digunakan adalah pautan tunggal, pautan lengkap, dan pautan rata-rata. Sedangkan metode pengelompokan non hirarki atau disebut metode partisi yang sering digunakan yaitu *C-Means*. *C-Means* melakukan pembentukan partisi sehingga setiap objek akan berada tepat pada satu partisi. Akan tetapi pada suatu kondisi hal tersebut tidak dapat diterapkan, karena sebenarnya bisa saja objek tersebut terletak diantara dua atau lebih partisi yang lain. Kekurangan lain dari *C-Means* yaitu, metode ini hanya dapat diterapkan pada kasus data berskala numerik saja.

Metode *Fuzzy C-Means* (FCM) merupakan pengembangan dari metode pengelompokan non hirarki *C-Means* yang ditambahkan konsep *fuzzy*. Konsep tersebut ditambahkan agar dapat meminimalkan kegagalan konvergen metode *C-Means* pada data berskala numerik yang memiliki keragaman rendah. FCM melakukan pengelompokan dengan mempertimbangkan derajat keanggotaan sebagai dasar pembobotan dalam pengelompokan. Derajat keanggotaan menentukan tingkat keberadaan objek pada suatu kelas atau kelompok. Sehingga objek dapat menjadi anggota dari semua kelompok terbentuk dengan derajat atau tingkat keanggotaan yang berbeda antara 0 sampai 1 (Bezdek, Ehrlich, & Full, 1984). Algoritma FCM pertama kali diperkenalkan oleh Dunn (1974) kemudian dikembangkan oleh Bezdek (1981) dan algoritma FCM dari Bezdek yang paling banyak digunakan. Sukim (2011) telah melakukan penelitian menggunakan *Fuzzy C-Means* dan

diketahui bahwa metode ini lebih kokoh untuk mempertahankan banyaknya klaster terhadap adanya data pencilan jika dibandingkan dengan metode *C-Means*. Selain itu Rizal dan Hakim (2015) juga membandingkan metode FCM dengan *C-Means* dalam studi kasus indeks pembangunan manusia di Kawasan Indonesia Timur dan dihasilkan bahwa metode FCM menghasilkan nilai rasio S_W dan S_B yang lebih kecil.

C-Means hanya mampu diterapkan pada kasus pengelompokan data berskala numerik saja, maka dilakukan pengembangan metode hingga dihasilkan metode *K-Modes.* *K-Modes* merupakan pengembangan dari *C-Means* agar dapat melakukan pengelompokan menggunakan data berskala kategorik (Huang Z. , 1998). Perbedaan metode ini dengan *C-Means* yaitu *K-Modes* menggantikan nilai *mean* menjadi nilai modus, ukuran jarak yang dihitung berupa kesesuaian suatu nilai atribut tiap dimensi terhadap titik pusat klaster, dan menggunakan metode berbasis frekuensi untuk memutakhirkan modus. Suguna (2012) dalam jurnal penelitian internasionalnya menuliskan bahwa algoritma *K-Modes* mampu efisien dalam mengelompokkan data kategorik. Selain itu metode *K-Modes* menghasilkan kelompok yang lebih rinci dan waktu pembentukan kelompok relatif lebih singkat.

Algoritma pengelompokan yang ada sebagian besar berfokus pada pengelompokan objek dengan variabel yang memiliki data berskala numerik. Sedangkan pada kenyataannya tidak semua permasalahan di dunia ini memiliki data berupa variabel berskala numerik. Seringkali data yang dimiliki berskala campuran, sehingga analisisnya tidak bisa disamakan. Terdapat dua metode yang seringkali digunakan untuk mengelompokkan data berskala campuran yaitu dengan mentransformasi data kategorik menjadi data numerik dan sebaliknya. Walaupun transformasi dapat mengurangi kompleksitas dalam komputasi, akan tetapi metode ini mendatangkan kesulitan dalam menentukan transformasi yang tepat agar tidak kehilangan banyak informasi dari data aslinya. Selain menggunakan metode transformasi, metode *clustering ensemble* dapat diterapkan untuk mengelompokkan objek-objek dengan variabel berskala campuran. Pengelompokan *ensemble* (*clustering ensemble*) merupakan metode pengelompokan yang menggabungkan hasil pengelompokan beberapa algoritma agar

diperoleh hasil akhir pengelompokan yang lebih baik. Pengelompokan menggunakan metode *ensemble cluster* memiliki beberapa macam penerapan, salah satunya menggunakan prinsip pengelompokan data kategorik (He, Xu, & Deng, 2005). *Ensemble K-Modes* merupakan salah satu metode penggabungan dengan menggunakan algoritma data kategorik yaitu *K-Modes* untuk memperoleh kelompok akhir dari sebuah pengelompokan data campuran. Metode ini menggabungkan hasil pengelompokan dari variabel kategorik dan numerik sehingga menghasilkan data berskala kategorik, karenanya algoritma pengelompokan data kategorik digunakan untuk mendapat kelompok akhir (*final cluster*). Saputra (2016) pernah menggunakan *ensemble cluster* dengan metode *K-Modes* untuk mengelompokan aksesi jeruk dan membandingkan hasil pengelompokannya dengan metode numerik murni (FCM) serta kategorik murni (*k-modes*) sehingga diketahui bahwa metode *ensemble k-modes* mampu menerangkan dan menggabungkan kedua tipe karakter menjadi satu karakter data yaitu kategorik dengan hasil yang lebih baik.

Similarity Weight and Filter Method (SWFM) merupakan pengembangan dari metode *ensemble* yang memiliki perbedaan pada tahapan pembentukan *final cluster*. Dalam metode *ensemble* pembentukan kelompok akhir (*final cluster*) dilakukan dengan menggunakan algoritma pengelompokan pada data kategorik, sedangkan SWFM terbagi menjadi dua tahapan. Tahapan *similarity weight* digunakan untuk menghitung ukuran kemiripan dengan memasukkan faktor bobot pada rumus ukuran kemiripan. Kemudian tahapan *filter* digunakan untuk memperoleh hasil akhir pengelompokan. Penghitungan tahapan *filter* ini didasarkan pada tahap *similarity weight.* Kelebihan dari metode SWFM adalah penggunaan yang efisien untuk beberapa dimensi serta mengurangi kompleksitas waktu, dan pencarian jarak yang digunakan lebih efisien bahkan jika batas-batas kelompok tidak teratur (Reddy & Khavita, 2012). Penelitian menggunakan metode SWFM pernah dilakukan oleh Prakoso (2017) untuk pengelompokan Sekolah Menengah Atas (SMA) di Sidoarjo dengan menggabungkan hasil pengelompokan metode *single linkage* untuk data numerik dan metode *k-modes* untuk data kategorik. Reddy dan Kavitha (2012) melakukan perbandingan antara algoritma *k-prototype* dengan

algoritma SWFM, diperoleh kesimpulan bahwa algoritma SWFM lebih baik karena memiliki rata-rata kesalahan pengelompokan lebih kecil dibandingkan dengan algoritma *k-prototype*.

Metode *ensemble k-modes* dan SWFM merupakan metode yang sama-sama digunakan untuk mengelompokan data yang memiliki skala data campuran. Secara umum konsep dasar kedua metode tersebut sama, hanya pembentukan *final cluster*nya yang berbeda. Sehingga kedua metode tersebut dapat digunakan untuk mengelompokkan kabupaten/kota Jawa Timur berdasarkan indikator daerah tertinggal yang memiliki data berskala campuran. Sesuai dengan Peraturan Presiden Nomor 131 Tahun 2015 tentang Penetapan Daerah Tertinggal Tahun 2015–2019, suatu daerah ditetapkan sebagai daerah tertinggal berdasarkan kriteria perekonomian masyarakat, sumber daya manusia, sarana dan prasarana, kemampuan keuangan daerah, aksesibilitas, dan karakteristik daerah. Keenam kriteria tersebut memiliki beberapa sub indikator yang umumnya memiliki data tidak hanya berskala numerik, tetapi juga berskala kategorik.

Kesenjangan antar wilayah dan ketertinggalan suatu daerah masih menjadi isu yang harus diatasi sejalan dengan tujuan pembangunan nasional. Penyebab terjadinya kesenjangan tersebut sangatlah beragam mulai dari perbedaan ketersediaan sumber daya alam, letak geografis, kualitas sumber daya manusia, kemajuan ekonomi, hingga pada aspek sosial budaya. Kesenjangan pembangunan tersebut ditunjukkan dengan masih adanya daerah-daerah yang tingkat perkembangannya masih tertinggal dibandingkan daerah lainnya dengan kata lain keberadaan daerah tertinggal sebagai indikator adanya kesenjangan dalam pembangunan. Daerah tertinggal adalah daerah kabupaten yang wilayah serta masyarakatnya kurang berkembang dibandingkan dengan daerah lain dalam skala nasional (Bappenas, 2016). Jumlah daerah tertinggal sejak tahun 2005 hingga saat ini telah mengalami penurunan, periode tahun 2005-2009 terdapat 199 daerah tertinggal lalu terus menurun hingga periode 2015-2019 dietapkan terdapat 122 kabupaten yang termasuk dalam kategori daerah tertinggal. Berdasarkan Perpres No. 131 Tahun 2015, Jawa Timur merupakan salah satu provinsi yang memiliki kabupaten/kota berstatus daerah tertinggal. Hal itu terjadi salah satunya dikarenakan Jawa Timur

memiliki Indeks Pembangunan Manusia (IPM) yang berada di bawah nilai IPM Indonesia dan paling kecil diantara Provinsi di Pulau Jawa lainnya. Pada tahun 2014, IPM Jawa Timur adalah sebesar 68.14 sedangkan IPM Indonesia sebesar 68.90. IPM merupakan indikator penting untuk mengukur keberhasilan dalam upaya membangun kualitas hidup manusia sehingga dapat menentukan peringkat atau level pembangunan suatu wilayah. Sehingga IPM dapat dikatakan termasuk ke dalam kriteria kualitas sumber daya manusia yang ikut menentukan ketertinggalan daerah.

Pemerintah berharap pada akhir periode Rencana Pembangunan Jangka Menengah Nasional (RPJMN) 2015-2019 dapat terentaskan sebanyak 80 kabupaten tertinggal. Oleh karena itu perlu diketahui oleh pemerintah daerah mana saja yang masih tergolong daerah teringgal dan indikator apa yang menyebabkan daerah tersebut tertinggal. Sebelumnya Kementrian Desa, Pembangunan Daerah Tertinggal, dan Transmigrasi (KDPDTT) juga melakukan penentuan sebuah kabupaten termasuk daerah tertinggal atau tidak tertinggal, akan tetapi penentuan tersebut hanya mengakomodasi variabel numerik saja. Pada penelitian kali ini akan dilakukan pengelompokan kabupaten/kota di Jawa Timur untuk mengetahui kabupaten/kota yang termasuk dalam daerah tertinggal serta karakteristiknya, agar kelompok daerah tertinggal tersebut dapat menjadi perhatian utama pemerintah Jawa Timur. Pengelompokan akan dilakukan menggunakan metode *ensemble k-modes* dan SWFM yang menggabungkan hasil pengelompokan numerik menggunakan FCM dan pengelompokan kategorik dengan *K-Modes* untuk memperoleh *final cluster*. Diharapkan hasil penelitian juga dapat membantu pemerintah dalam menentukan kebijakan yang sesuai dengan karakteristik kelompok sehingga dapat mengeluarkan kabupaten/kota tersebut dari status daerah tertinggal.

Penelitian mengenai daerah tertinggal dengan data berskala campuran pernah dilakukan oleh Dewi (2012) yang mengelompokkan desa perdesaan di Provinsi Riau dengan membandingkan tiga metode. Ketiga metode tersebut yaitu tansformasi data menjadi numerik dan dikelompokkan menggunakan metode *average linkage*, transformasi data menjadi kategorik kemudian dikelompokkan menggunakan algoritma *squeezer*, dan menggunakan *ensemble cluster* yang menggabungkan

hasil pengelompokan *average linkage* dan *squeezer*. Perbandingan ketiga metode tersebut menghasilkan kesimpulan bahwa metode *ensemble cluster* dengan pembentukan *final cluster* menggunakan algoritma *squeezer* memiliki nilai rasio S_W dan S_B yang lebih kecil dibandingkan dua metode lainnya. Selain itu Ramdhany (2018) melakukan pengelompokan desa tertinggal di Kabupaten Bondowoso dengan *ensemble* ROCK (*RObust Clustering using linKs*). Pengelompokan data numerik dilakukan dengan membandingkan metode *single linkage*, *complete linkage*, dan *average linkage* untuk mendapatkan metode terbaik, diperoleh hasil bahwa ketiga metode tersebut memiliki nilai ICDrate yang sama. Sedangkan pengelompokan data kategorik dilakukan menggunakan algoritma ROCK yang kemudian kedua hasil pengelompokan tersebut digabungkan untuk memperoleh *final cluster* menggunakan *ensemble* ROCK. Sehingga diketahui bahwa dengan dengan metode *ensemble* ROCK, desa-desa di Kabupaten Bondowoso dapat dikelompokkan dalam 2 kelompok dengan status desa yang berbeda.

1.2 Rumusan Masalah

Dari periode ke periode jumlah daerah tertinggal di Indonesia terus berkurang, akan tetapi masih sangat diperlukan upaya pemerintah untuk mewujudkan target bahwa akan terentaskan 80 kabupaten tertinggal pada akhir periode RPJMN 2015-2019. Jawa Timur merupakan salah satu provinsi yang memiliki daerah dengan status tertinggal. Secara administratif, Jawa Timur terdiri dari 29 kabupaten dan 9 kota yang memiliki karakteristik berbeda-beda. Agar status ketertinggalan dapat diatasi, maka perlu diketahui karakteristik kabupaten/kota di Jawa Timur berdasarkan indikator ketertinggalan daerah. Kementrian Desa, Pembangunan Daerah Tertinggal, dan Transmigrasi (KDPDTT) melakukan penentuan sebuah kabupaten termasuk daerah tertinggal atau tidak, akan tetapi penentuan tersebut hanya mengakomodasi variabel numerik saja. Sehingga penelitian ini akan mengelompokkan kabupaten/kota di Jawa Timur berdasarkan indikator daerah tertinggal dengan menggunakan metode pengelompokan data campuran *ensemble k-modes* dan SWFM untuk mengetahui pengelompokan daerah serta karakteristik hasil pengelompokan kabupaten/kota di Jawa Timur.

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang ada, tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut.

1. Mendeskripsikan karakteristik kabupaten/kota di Jawa Timur berdasarkan indikator daerah tertinggal.
2. Mengelompokkan kabupaten/kota di Jawa Timur berdasarkan indikator daerah tertinggal dengan menggunakan metode pengelompokan data campuran *ensemble k-modes*.
3. Mengelompokkan kabupaten/kota di Jawa Timur berdasarkan indikator daerah tertinggal dengan menggunakan metode pengelompokan data campuran SWFM.
4. Mendapatkan hasil kelompok akhir terbaik dan karakteristik hasil pengelompokan kabupaten/kota di Jawa Timur.

1.4 Manfaat Penelitian

Hasil dari penelitian ini diharapkan mampu memberikan manfaat sebagai berikut.

1. Memberikan rekomendasi atau pertimbangan bagi pemerintah dalam menyusun kebijakan sehingga dapat menyelesaikan permasalahan daerah tertinggal di Jawa Timur
2. Mengaplikasikan ilmu statistika terutama analisis pengelompokan dengan data multivariat yang memiliki skala campuran menggunakan metode *ensemble k-modes* dan SWFM sehingga dapat dijadikan referensi untuk penelitian selanjutnya.

1.5 Batasan Masalah

Batasan masalah dalam penelitian ini adalah data yang digunakan merupakan data dari Potensi Desa dan Susenas pada tahun 2014. Karena dalam penelitian ini menggunakan skala data campuran, maka analisis faktor hanya akan digunakan pada variabel berskala numerik. Data dalam penelitian ini diasumsikan berdistribusi normal multivariat dan matriks kovarian bersifat homogen.

(Halaman Ini Sengaja Dikosongkan)

BAB II
TINJAUAN PUSTAKA

Pada tinjauan pustaka akan dijelaskan mengenai teori-teori yang digunakan dalam melakukan analisis pengelompokan dengan skala data campuran. Bab ini akan menguraikan mengenai statistika deskriptif, analisis faktor, dan asumsinya, analisis klaster yang meliputi *fuzzy c-means* (FCM), *k-modes*, *ensemble cluster, similarity weight and filter method* (SWFM), penentuan kelompok optimum, kinerja hasil pengelompokan, *Multivariate Analysis of Variance* (MANOVA), Uji Kruskal Wallis, dan daerah tertinggal.

2.1 Statistika Deskriptif

Statistika deskriptif adalah salah satu metode statistika yang berkaitan dengan pengumpulan dan penyajian suatu gugus data sehingga memberikan informasi yang berguna. Statistika deskriptif hanya memberikan informasi mengenai data yang dipunyai dan sama sekali tidak menarik inferensia atau kesimpulan apapun. Dengan statistika deskriptif, kumpulan data yang diperoleh akan tersaji dengan ringkas dan rapi serta dapat memberikan informasi inti dari kumpulan data yang ada. Informasi yang dapat diperoleh dari statistika deskriptif ini antara lain ukuran pemusatan data dan ukuran penyebaran data (Walpole, Myers, Myers, & Ye, 2007).

Pada variabel berskala numerik digunakan ukuran pemusatan data rata-rata (*mean*) dan median. Sedangkan untuk variabel berskala kategorik menggunakan ukuran pemusatan data modus atau nilai yang sering muncul pada suatu kelompok data. Data dengan frekuensi terbanyak menunjukkan modus. Selain itu juga statistika deskriptif bisa disajikan dalam bentuk tabel, *boxplot*, dan *scatter plot* sehingga informasi yang disampaikan akan lebih mudah dipahami oleh pembaca.

2.2 Analisis Faktor dan Kelayakan

Analisis faktor adalah metode multivariat untuk menganalisis independensi variabel dengan dimensi yang cukup tinggi, dengan tujuan faktorisasi. Tujuan dari analisis faktor adalah untuk menggambarkan hubungan-hubungan kovarian antara beberapa variabel yang mendasari tetapi tidak teramati, kuantitas random yang

disebut faktor. Dalam penerapannya analisis faktor hanya dapat dilakukan pada data berskala numerik saja. Namun sebelum melakukan analisis ini, perlu dilakukan pengujian dan pemeriksaan asumsi yang harus dipenuhi terlebih dahulu.

2.2.1 Pemeriksaan Kecukupan Variabel

Kecukupan variabel yang digunakan dalam penelitian dapat diukur dengan menggunakan nilai *Kaiser-Meyer-Olkin* (KMO). Metode ini mengukur kecukupan sampling secara menyeluruh dan mengukur kecukupan sampling untuk tiap indikator dengan menggunakan nilai korelasi antar variabel (Rencher, 1998).

Pemeriksaan kecukupan variabel dengan KMO dilakukan dengan rumus sebagai berikut.

$$KMO = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2}{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2 + \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}^2} \qquad (2.1)$$

Keterangan:
i : 1, 2, 3, ..., p
j : 1, 2, 3, ..., p
r_{ij} : Koefisien korelasi antara variabel i dan j
a_{ij} : Koefisien korelasi parsial antara variabel i dan j

Berikut ini akan diberikan ilustrasi untuk menghitung nilai KMO. Apabila terdapat 10 objek dengan 3 variabel maka cara mendapatkan nilai KMO adalah sebagai berikut.

Tabel 2.1 Ilustrasi Perhitungan KMO

Obj	X_1	X_2	X_3	X_1^2	X_2^2	X_3^2	$X_1 X_2$	$X_1 X_3$	X_2X_3
1	2.5	4	3	6.25	16	9	10	7.5	12
2	1.8	2	2.3	3.24	4	5.29	3.6	4.14	4.6
3	4	3	2	16	9	4	12	8	6
4	2	4	5	4	16	25	8	10	20
5	1	8	8	1	64	64	8	8	64
6	6.3	2	4	39.69	4	16	12.6	25.2	8
7	0.8	3	3.3	0.64	9	10.89	2.4	2.64	9.9
8	1.2	3	6	1.44	9	36	3.6	7.2	18
9	2.9	4	2	8.41	16	4	11.6	5.8	8
10	4.3	4.5	3	18.49	20.25	9	19.35	12.9	13.5
Total	26.8	37.5	38.6	99.16	167.25	183.18	91.15	91.38	164

Sebelum menghitung nilai KMO, harus dihitung terlebih dahulu nilai korelasi dan korelasi parsial antar variabel. Sebagai contoh akan diberikan perhitungan nilai korelasi antar variabel X_1 dan X_2 serta nilai korelasi parsial antara X_1 terhadap X_2X_3.

$$r_{x_1x_2} = \frac{n\sum x_1x_2 - (\sum x_1)(\sum x_2)}{\sqrt{n\sum x_1^2 - (\sum x_1)^2 \text{ x } n\sum x_2^2 - (\sum x_2)^2}}$$

$$= \frac{10(91.15) - (26.8)(37.5)}{\sqrt{10(99.16) - (26.8)^2 \text{ x } 10(167.25) - (37.5)^2}}$$

$$= \frac{-93.5}{269.782} = -0.34658$$

Perhitungan korelasi diatas digunakan untuk mencari nilai r_{ij}, sehingga akan diperoleh pula nilai r_{x1x3} sebesar -0.39478 dan r_{x2x3} sebesar 0.63808. Selanjutnya untuk mencari nilai korelasi parsial dilakukan perhitungan berikut.

$$a_{x_1(x_2x_3)} = \frac{r_{x_2x_3} - (r_{x_1x_2})(r_{x_1x_3})}{\sqrt{(1 - r^2_{x_1x_2})(1 - r^2_{x_1x_3})}}$$

$$= \frac{0.63808 - (-0.34658)(-0.39478)}{\sqrt{(1 - (-0.34658)^2)(1 - (-0.39478)^2)}}$$

$$= \frac{0.50126}{0.86183} = 0.58162$$

Sehingga diperoleh pula korelasi parsial $a_{x2(x1x3)}$ sebesar -0.24041 dan $a_{x3(x1x2)}$ sebesar -0.13383. Kemudian dari korelasi-korelasi tersebut dapat diperoleh nilai KMO yaitu sebesar 0.62265.

Sampel akan dikatakan cukup atau layak untuk dilakukan analisis faktor apabila nilai KMO > 0.5 (Hair, Black, Babin, & Anderson, 2010).

2.2.2 Uji Korelasi Variabel

Asumsi selanjutnya yang harus dipenuhi yaitu independensi antar variabel. Variabel $X_1, X_2, \ldots, X_p$ dikatakan bersifat saling bebas (*independent*) jika matriks korelasi antar variabel membentuk matriks indentitas. Asumsi ini dapat diuji menggunakan *Bartlett test of Spericity* dengan hipotesis sebagai berikut.

H_0: $\boldsymbol{\rho} = \mathbf{I}$ (Variabel bersifat saling bebas)

H_1: $\boldsymbol{\rho} \neq \mathbf{I}$ (Variabel bersifat tidak saling bebas)

Statistik uji untuk *Bartlett test* yaitu sebagai berikut.

$$\chi^2_{hitung} = -\left[(n-1) - \frac{(2p+5)}{6}\right] \ln|\mathbf{R}| \qquad (2.2)$$

dengan

n : jumlah observasi

p : jumlah variabel

$|\mathbf{R}|$: determinan dari matriks korelasi

Berikut ini akan diberikan ilustrasi dalam pengujian korelasi variabel. Diketahui terdapat 10 objek dengan 3 variabel yang sama seperti Tabel 2.1 maka untuk memperoleh nilai *chi-square* hitung sebelumnya perlu dilakukan perhitungan matriks korelasi dahulu. Sebelumnya pada subbab 2.2.1 telah dilakukan perhitungan korelasi antar variabel, sehingga apabila dibentuk ke dalam matriks memiliki bentuk sebagai berikut.

$$\mathbf{R} = \begin{pmatrix} 1 & -0.34658 & -0.39478 \\ -0.34658 & 1 & 0.63808 \\ -0.39478 & 0.63808 & 1 \end{pmatrix}$$

Kemudian dihitung determinan dari matriks korelasih tersebut, sehingga det$|\mathbf{R}|$ = 0.49149. Setelah diperoleh semua nilai yang diperlukan dalam perhitungan maka diperoleh nilai *chi-square* hitung sebesar 5.09052.

Pengujian ini akan mendapat keputusan Tolak H_0 ketika $\chi^2_{hitung} > \chi^2_{a,p(p-1)/2}$, artinya antar variabel bersifat dependen. Selain itu, suatu matriks dikatakan menyerupai matriks identitas apabila nilai determinannya mendekati 1 (Hair, Black, Babin, & Anderson, 2010).

Dalam pengujian korelasi variabel, perlu dilakukan pula pemeriksaan asumsi distribusi normal multivariat. Pemeriksaan distribusi normal multivariat digunakan untuk mengetahui apakah data mengikuti distribusi normal (Johnson & Wichern, 2007). Pemeriksaan distribusi normal multivariat dapat dilakukan salah satunya menggunakan uji *Shapiro Wilk* dengan hipotesis sebagai berikut.

H_0: Data berdistribusi normal multivariat

H_1: Data tidak berdistribusi normal multivariat

Statistik uji untuk pemeriksaan asumsi berdistribusi normal multivariat ditunjukkan pada persamaan (2.3)

$$W = \frac{1}{p}\sum_{k=1}^{P} W_k \tag{2.3}$$

Dimana

$$W_k = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \frac{b^2}{s^2} \tag{2.4}$$

$$a_i = (a_1, a_2, \ldots, a_n) = \frac{m' V^{-1}}{(m' V^{-1} V^{-1} m)^{1/2}}$$

Sehingga variabel dikatakan mengikuti distribusi normal multivariat jika nilai statistik uji W_k mendekati 1, namun jika nilai statistik uji W_k kecil atau jauh dari 1 maka dikatakan tidak mengikuti distribusi normal multivariat. Daerah penolakan H_0 yaitu saat $W > C_{(\alpha,n)}$ (Alva & Estrada, 2009).

Berikut ini merupakan algoritma serta ilustrasi dalam menghitung normal multivariat dengan *shapiro-wilk test*.

1. Mengurutkan data pengamatan
   Diurutkan dari nilai terendah hingga tertinggi. Menggunakan data yang sama pada Tabel 2.1 maka hasil pengurutan akan sebagai berikut.

Tabel 2.2 Ilustrasi Data Pemeriksaan Normal Multivariat

Obj	X_1	X_2	X_3
1	0.8	2	2
2	1	2	2
3	1.2	3	2.3
4	1.8	3	3
5	2	3	3
6	2.5	4	3.3
7	2.9	4	4
8	4	4	5
9	4.3	4.5	6
10	6.3	8	8
Total	26.8	37.5	38.6
$\mathbf{s^2}$	27.336	26.625	34.184

2. Menghitung nilai *b*
   a. Jika *n* genap maka *m = n/2*
   b. Jika *n* ganjil maka *m = (n-1)/2*

$$b = \begin{cases} \sum_{i=1}^{m} a_{n-i+1}(x_{n-i+1} - x_i) & ; \text{genap} \\ b = a_n(x_n - x_1) + \ldots + a_{m+2}(x_{m+2} - x_m) & ; \text{ganjil} \end{cases} \quad (2.5)$$

   Dimana nilai x_{m+1} adalah median dan nilai *a* merupakan nilai koefisien tabel *shapiro*. Karena pada ilustrasi digunakan data genap, maka diketahui bahwa *m* adalah sebesar 5. Setelah dilakukan perhitungan diperoleh nilai *b* pada masing-masing variabel adalah sebesar 4.99655, 4.64255, dan 5.47224.
3. Menghitung nilai W_k pada masing-masing variabel.
   W_1 = 0.913283
   W_2 = 0.809513
   W_3 = 0.876007
4. Menghitung nilai *W* dengan menghitung rata-rata hasil dari langkah ketiga. Sehingga diperoleh nilai *W* sebesar 0.866267.

2.2.3 Analisis Faktor

Analisis faktor adalah metode multivariat untuk menganalisis independensi variabel dengan dimensi yang cukup tinggi, dengan tujuan faktorisasi. Tujuan dari analisis faktor adalah untuk menggambarkan hubungan-hubungan kovarian antara beberapa variabel yang mendasari tetapi tidak teramati, kuantitas random yang disebut faktor (Rencher, 1998).

Vektor random teramati *X* dengan *p* komponen, memiliki rata-rata μ dan matrik kovarian $\sum$. Model analisis faktor adalah sebagai berikut

$$x_1 - \mu_1 = l_{11}F_1 + l_{12}F_2 + \cdots + l_{1m}F_m + \varepsilon_1$$
$$\vdots$$
$$x_p - \mu_p = l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pm}F_m + \varepsilon_p$$

Atau dapat ditulis dalam notasi matriks seperti pada persamaan berikut ini.

$$\mathbf{X}_{(pxl)} = \boldsymbol{\mu}_{(pxl)} + \mathbf{L}_{(pxm)} \times \mathbf{F}_{(mxl)} + \boldsymbol{\varepsilon}_{(pxl)} \quad (2.6)$$

keterangan :
μ_i : rata-rata variabel ke-*i*; *i* = 1,2,...,*p*
ε_i : faktor spesifik ke-*i*
F_j : *common faktor* ke-*j*; *j* = 1,2,...*m*
ℓ_{ij} : *loading* dari variabel ke-*i* pada faktor ke-*j*.

Bagian dari varian variabel ke–*i* dari *mcommon faktor* disebut komunalitas ke–*i* yang merupakan jumlah kuadrat dari *loading* variabel ke–*i* pada *mcommon factor* dengan rumus,

$$h_i^2 = l_{i1}^2 + l_{i2}^2 + \ldots + l_{im}^2 \qquad (2.7)$$

Penentuan banyaknya faktor terbentuk adalah berdasarkan faktor yang memiliki nilai eigen lebih dari satu dan dengan melihat persentase kumulatif nilai eigen pada faktor terbentuk. Menurut Hair (2010), persentase kumulatif nilai eigen yang ditoleransi adalah sebesar 60%. Secara visual, penentuan banyaknya faktor terbentuk dapat dilakukan dengan melihat *scree plot* yaitu melihat letak terjadinya belokan.

Nilai eigen atau total variansi yang dapat dijelaskan oleh masing-masing faktor dapat diperoleh dari matriks korelasinya, seperti pada persamaan berikut.

$$\det(\mathbf{R} - \lambda\mathbf{I}) = 0 \qquad (2.8)$$

Keterangan
R : matriks korelasi
λ : nilai eigen
I : matriks identitas

Analisis faktor memiliki beberapa tujuan yaitu menggunakan matriks korelasi hitungan untuk mengidentifikasi jumlah terkecil dari faktor umum (model faktor yang paling parsimoni) yang mempunyai penjelasan terbaik atau menghubungkan korelasi diantara variabel indikator, mengidentifikasi melalui faktor rotasi, estimasi bentuk dan struktur loading, komunaliti dan varian unik dari indikator, melakukan interpretasi dari faktor umum dan jika perlu akan dilakukan estimasi faktor skor (Johnson & Wichern, 2007). Apabila analisis berikutnya adalah analisis klaster maka dapat menggunakan nilai komponen utama atau variabel representatif dari masing-masing komponen. Analisis faktor perlu dilakukan jika banyaknya variabel di tiap komponen berbeda. Karena akan berpengaruh terhadap performansi dari hasil klaster.

2.3 Analisis Kelompok (*Cluster Analysis*)

Analisis kelompok merupakan salah satu metode dalam hubungan multivariat yang digunakan untuk mengelompokkan objek pengamatan berdasarkan karakteristik-karakteristik yang dimiliki (Johnson & Wichern, 2007). Pengelompokan dilakukan dengan memaksimalkan kehomogenan objek pengamatan dalam satu kelompok sekaligus memaksimalkan keheterogenan antar kelompok. Objek-objek yang telah diklasifikasikan dalam satu kelompok merupakan objek-objek yang memiliki kedekatan jarak relatif sama dengan objek lainnya (Narimawati, 2008).

Terdapat dua metode dalam analisis kelompok yaitu meode hirarki dan metode non hirarki. Perbedaan dari kedua metode tersebut terletak pada prosedur pengelompokannya, metode hirarki mengelompokkan suatu pengamatan secara bertahap, sementara pada metode nonhirarki dilakukan partisi pada ruang sampelnya. Salah satu metode non hirarki yang sering digunakan yaitu metode *C-Means*. Akan tetapi metode *C-Means* hanya mampu mengakomodasi pengelompokan untuk data berskala numerik saja, selain itu metode ini mengelompokan suatu objek pengamatan ke dalam satu kelompok pasti. Metode FCM merupakan pengembangan dari metode *C-Means* dengan menambahkan konsep fuzzy sehingga mampu mengatasi kekonvergenan yang terjadi saat menggunakan metode *C-Means*. *K-Modes* merupakan pengembangan dari metode *C-Means* yang mampu mengelompokkan data berskala kategorik.

Data dalam penelitian ini berskala numerik dan kategorik, sehingga akan digunakan metode *Ensemble Cluster* yang mampu mengelompokkan data berskala campuran. Konsep dari metode ini adalah menggabungkan beberapa algoritma berbeda untuk mendapatkan kelompok dari hasil pengelompokan indvidu. Dalam penelitian ini digunakan metode *Ensemble K-Modes* dan SWFM, yang menggabungkan hasil pengelompokan numerik dengan metode FCM dan pengelompokan kategorik dengan metode *K-Modes*.

2.3.1 *Fuzzy C-Means* (FCM)

Metode FCM merupakan pengembangan dari metode nonhirarki *C-Means*. Konsep *fuzzy* diharapkan mampu untuk meminimalkan kejadian konvergen yang dialami metode

pengelompokan biasanya. FCM pertama kali ditemukan oleh Dunn (1973) dan selanjutnya dikembangkan oleh Bezdek (1981), pengembangan oleh Bezdek yang sampai saat ini banyak digunakan. Algoritma metode ini secara umum sama dengan *C-Means*, karena pada awalnya menentukan jumlah kelompok yang akan dibentuk kemudian melakukan iterasi sampai terbentuk anggota tiap kelompok. Perbedaan terletak pada hasil pengelompokan, pada metode FCM objek pengamatan tidak mutlak untuk menjadi anggota satu kelompok saja tetapi memungkinkan untuk menjadi anggota kelompok lain dengan ukuran tingkat keanggotaan yang berbeda. Pemilihan anggota kelompok dilakukan dengan melihat derajat keanggotaan objek dalam kelompok itu paling besar dibandingkan dengan derajat keanggotaan objek tersebut di kelompok lain.

Dalam mendefinisikan keanggotaan himpunan *fuzzy* dapat dilakukan secara numerik dan fungsional. Metode *fuzzy clustering*, keanggotaan himpunan *fuzzy* dilakukan secara fungsional atau pendekatan fungsi. Derajat keanggotaan (*membership function*) memiliki arti sebagai suatu kurva yang menunjukkan pemetaan titik-titik input data ke dalam nilai keanggotaan yang memiliki interval antara 0 sampai 1. Beberapa fungsi keanggotaan yang paling sederhana dan bisa digunakan adalah representasi linear, kurva segitiga, dan kurva trapesium (Kusumadewi & Hari, 2004).

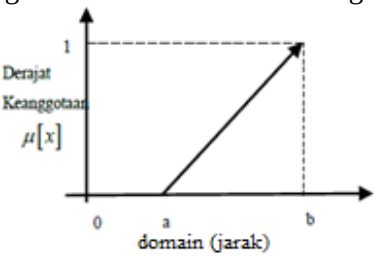
a. Representasi Linear

Pemetaan input derajat keanggotaan pada representasi linear digambarkan sebagai sebuah garis lurus. Bentuk ini merupakan bentuk paling sederhana dan menjadi pilihan yang baik untuk mendekati suatu konsep yang kurang jelas. Terdapat dua bentuk fungsi keanggotaan linear. Pertama adalah representasi fungsi linear naik, yaitu kenaikan himpunan dimulai nilai *domain* dengan derajat keanggotaan nol dan bergerak ke kanan menuju ke nilai *domain* dengan derajat keanggotaan yang leih tinggi. Rumus representasi linear naik ditampilkan pada persamaan 2.9.

$$\mu[x,a,b] = \begin{cases} 0; x \le a \\ \dfrac{(x-a)}{(b-a)}; a \le x \le b \\ 1; x \ge b \end{cases} \qquad (2.9)$$

Keterangan:

a : nilai domain yang mempunyai derajat keanggotaan nol
b : nilai domain yang mempunyai derajat keanggotaan satu
x : nilai input yang akan diubah ke dalam bilangan *fuzzy*

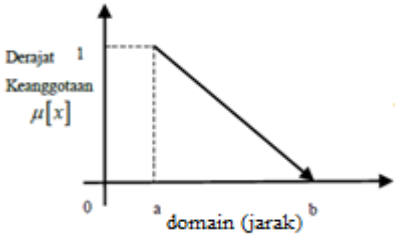


Gambar 2.1 Representasi Linear Naik

Kedua yaitu representasi fungsi linear turun yang mana garis lurus dimulai dari nilai domain dengan derajat keanggotaan tertinggi pada sisi kiri, kemudian menurun menuju nilai domain dengan derajat keanggotaan yang lebih rendah. Rumus representasi linear turun adalah sebagai berikut

$$\mu[x,a,b] = \begin{cases} \dfrac{(b-x)}{(b-a)}; & a \le x \le b \\ 0; & x \ge b \end{cases} \qquad (2.10)$$

Keterangan:
a : nilai domain yang mempunyai derajat keanggotaan satu
b : nilai domain yang mempunyai derajat keanggotaan nol
x : nilai input yang akan diubah ke dalam bilangan *fuzzy*



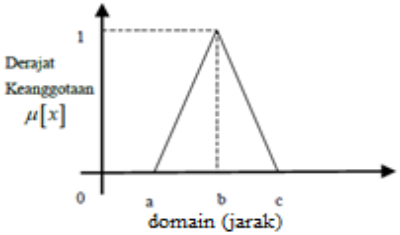
Gambar 2.2 Representasi Linear Turun

b. Representasi Kurva Segitiga

Kurva segitiga merupakan gabungan antara dua garis linear dengan nilai-nilai disekitar *b* memiliki derajat keanggotaan turun cukup tajam (menjauhi 1). Rumus representasi kurva segitiga ditunjukkan pada persamaan 2.11.

$$\mu[x,a,b,c] = \begin{cases} 0; x \le a \text{ atau } x \ge c \\ (x-a)/(b-a); a \le x \le b \\ (c-x)/(c-b); b \le x \le c \end{cases} \quad (2.11)$$

Keterangan:
a : nilai domain terkecil yang mempunyai derajat keanggotaan nol
b : nilai domain yang mempunyai derajat keanggotaan satu
c : nilai domain terbesar yang mempunyai derajat keanggotaan nol



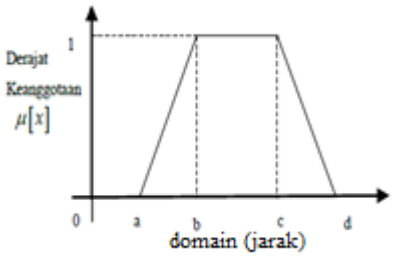
Gambar 2.3 Representasi Kurva Segitiga

c. Representasi Kurva Trapesium

Kurva trapesium pada dasarnya seperti bentuk representasi kurva segitiga, yang membedakan adalah terdapat beberapa titik yang memiliki keanggotaan bernilai 1. Rumus representasi kurva trapesium adalah sebagai berikut.

$$\mu[x,a,b,c,d] = \begin{cases} 0; x \le a \text{ atau } x \ge d \\ (x-a)/(b-a); a \le x \le b \\ 1; b \le x \le c \\ (d-x)/(d-c); c \le x \le d \end{cases} \quad (2.12)$$

Keterangan:
a : nilai domain terkecil yang mempunyai derajat keanggotaan nol
b : nilai domain terkecil yang mempunyai derajat keanggotaan satu
c : nilai domain terbesar yang mempunyai derajat keanggotaan satu
d : nilai domain terbesar yang mempunyai derajat keanggotaan nol



Gambar 2.4 Representasi Kurva Trapesium

Berikut ini diberikan ilustrasi untuk menghitung derajat keanggotaan yaitu fungsi keanggotaan linier naik dan kurva segitiga. Apabila terdapat matriks **X** dengan ukuran 10x3, maka dalam mendapatkan derajat keanggotaannya dilakukan langkah berikut.

Tabel 2.3 Ilustrasi Data Numerik

Objek	X_1	X_2	X_3	Jarak
1	2.5	4	3	0.91351
2	1.8	2	2.3	2.504097
3	4	3	2	2.400937
4	2	4	5	1.350741
5	1	8	8	6.166401
6	6.3	2	4	4.023245
7	0.8	3	3.3	2.100119
8	1.2	3	6	2.707859
9	2.9	4	2	1.889577
10	4.3	4.5	3	1.98154
Mean	**2.68**	**3.75**	**3.86**	

Langkah pertama yang harus dilakukan adalah menghitung jarak antara objek dengan rata-ratanya. Perhitungan jarak yang digunakan adalah jarak euclidean kuadrat dengan rumus sebagai berikut dengan x_{ij} adalah objek pada pengamatan ke-i variabel ke-j dan $\bar{x}_j$ adalah nilai mean dari variabel ke-j.

$$d_i(x_i, \bar{x}) = \sqrt{\sum_{j=1}^{p} (x_{ij} - \bar{x}_j)^2}$$

Setelah diperoleh jarak dari masing-masing objek terhadap rata-ratanya, selanjutnya ditentukan domain untuk masing-masing fungsi keanggotaan.

a. Representasi Linier Naik

Pada fungsi keanggotaan linier naik terdapat tiga nilai yang harus ditentukan nilainya sebelum menghitung derajat keanggotaan.

a : domain yang berupa nilai minimum dari jarak (0.91351)

b : domain yang berupa nilai maksimum dari jarak (6.166401)

x : nilai jarak objek terhadap rata-rata

Setelah ditentukan nilai-nilai tersebut maka dilanjutkan untuk menghitung derajat keanggotaan berdasarkan fungsi pada persamaan 2.9.

Tabel 2.4 Hasil Perhitungan Fungsi Keanggotaan Linier Naik 2 Klaster

Objek	Jarak (x)	Derajat Keanggotaan 1	Derajat Keanggotaan 2
1	0.91351	0	1
2	2.504097	0.302802	0.697198
3	2.400937	0.283164	0.716836
4	1.350741	0.083236	0.916764
5	6.166401	1	0
6	4.023245	0.592005	0.407995
7	2.100119	0.225896	0.774104
8	2.707859	0.341593	0.658407
9	1.889577	0.185815	0.814185
10	1.98154	0.203322	0.796678

Derajat keanggotaan sebuah objek apabila dijumlahkan harus sama dengan 1. Sehingga jika ingin membentuk 2 kelompok maka derajat keanggotaan pada kelompok 2 dihitung dengan cara satu dikurangi derajat keanggotaan pada kelompok 1.

Akan ditampilkan pula cara menghitung keanggotaan apabila ingin dibentuk 3 klaster. Pertama kali yang harus dilakukan adalah menghitung sesuai fungsi keanggotaan yang digunakan, hasilnya seperti pada kolom A di Tabel 2.4. Kolom B diperoleh dengan cara satu dikurangi kolom A lalu dibagi 2, hal itu dilakukan karena tersisa 2 klaster yang belum memiliki derajat keanggotaan. Derajat keanggotaan setiap objek apabila dijumlahkan harus sama dengan satu. Sehingga untuk menentukan suatu objek termasuk pada kelompok yang mana dilakukan dengan membagi angka 1 terhadap banyaknya klaster yang ingin dibentuk. Kemudian hasil bagi tersebut dianggap sebagai batasan dalam penentuan nilai derajat keanggotaan. Apabila akan membentuk 3 klaster, maka diperoleh

batasan derajat keanggotaan pada klaster 1 adalah 0 - 0.332, klaster 2 yaitu 0.333 – 0.666, dan batasan klaster 3 adalah 0.667 – 1.

Tabel 2.5 Hasil Perhitungan Fungsi Keanggotaan Linier Naik 3 Klaster

A	B	C			D.Anggota 1	D.Anggota 2	D.Anggota 3
		1	2	3			
0	0.5	1	0	0	0	0.5	0.5
0.302802	0.348599	1	0	0	0.302802	0.348599	0.348599
0.283164	0.358418	1	0	0	0.283164	0.358418	0.358418
0.083236	0.458382	1	0	0	0.083236	0.458382	0.458382
1	0	0	0	1	0	0	1
0.592005	0.2039975	0	1	0	0.203998	0.592005	0.203998
0.225896	0.387052	1	0	0	0.225896	0.387052	0.387052
0.341593	0.3292035	0	1	0	0.329204	0.341593	0.329204
0.185815	0.4070925	1	0	0	0.185815	0.407093	0.407093
0.203322	0.398339	1	0	0	0.203322	0.398339	0.398339

Apabila nilai pada kolom A berada diantara batasan-batasan yang telah disebutkan sebelumnya, maka akan dinotasikan dengan angka 1 di klaster yang sesuai. Misalnya pada objek 1, diketahui pada kolom A memiliki nilai 0 yang artinya nilai tersebut berada pada batasan klaster 1 sehingga akan diberikan notasi 1 pada kolom C yang pertama. Sedangkan untuk memperoleh derajat keanggotaannya memperhatikan notasi-notasi pada kolom C. Apabila kolom C tertulis 1, maka derajat keanggotaannya adalah nilai pada kolom A. Sebaliknya jika tertulis 0 maka derajat keanggotaan pada klaster tersebut bernilai sama dengan kolom B.

b. Representasi Kurva Segitiga

Beberapa nilai yang harus ditentukan sebelum menghitung derajat keanggotaan kurva segitiga adalah sebagai berikut

a : domain yang berupa nilai minimum dari jarak (0.91351)

b : domain yang berupa nilai tengah dari jarak (2.250528)

c : domain yang berupa nilai maksimum dari jarak (6.166401)

x : nilai jarak objek terhadap rata-rata

Perhitungan derajat keanggotaan kurva segitiga dilakukan berdasarkan fungsi pada persamaan 2.11.

Tabel 2.6 Hasil Perhitungan Fungsi Keanggotaan Kurva Segitiga

Objek	Jarak (x)	a≤ x≤ b	b≤ x≤ c	Derajat Keanggotaan 1	Derajat Keanggotaan 2
1	0.91351	0	1.341436	0	1
2	2.504097	1.189652	0.935246	0.935246	0.064754
3	2.400937	1.112496	0.96159	0.96159	0.03841
4	1.350741	0.327019	1.22978	0.327019	0.672981
5	6.166401	3.92881	0	0	1
6	4.023245	2.325873	0.5473	0.5473	0.4527
7	2.100119	0.887504	1.03841	0.887504	0.112496
8	2.707859	1.342053	0.883211	0.883211	0.116789
9	1.889577	0.730033	1.092177	0.730033	0.269967
10	1.98154	0.798815	1.068692	0.798815	0.201185

Perlu diketahui bahwa perhitungan derajat keanggotaan ini hanya digunakan untuk mencari nilai elemen matriks partisi awal (matriks **U**). Bila ingin memperoleh hasil sampai dengan kelompok akhir, maka derajat keanggotaan akan diperbaharui sesuai dengan rumus yang terdapat pada algoritma FCM.

Pengelompokan menggunakan metode *Fuzzy C-Means* dapat dilakukan menggunakan algoritma FCM sebagai berikut:

1. Input data yang akan dikelompokan, berupa matriks **X** berukuran *n* x *m* dimana *n* adalah banyaknya objek pengamatan dan *m* adalah banyaknya variabel atau atribut berskala numerik. (*i* = 1,2,...,*n* dan *j* = 1,2,...,*m*)

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1m} \\ X_{21} & X_{22} & \cdots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nm} \end{bmatrix} \quad (2.13)$$

2. Menentukan beberapa input yang dibutuhkan dalam perhitungan FCM, yaitu:

a. Jumlah kelompok (*c*) merupakan banyaknya kelompok yang akan dibentuk

b. *Weighting exponent* (*w*) merupakan peubah besar pengaruh dari *membership function* dan memiliki pengaruh terhadap banyaknya interasi dalam *objective function*. Dalam FCM *w* memiliki nilai yang lebih besar dari 1 (*w*>1), tetapi belum ada

ketentuan yang jelas berapa nilai yang optimal dalam proses optimasi permasalahan pengelompokan. Umumnya digunakan nilai *w* sebesar 2 (*w*=2).

c. Maksimum iterasi merupakan batas perulangan (*looping*), perulangan akan berhenti jika nilai iterasi sudah tercapai
d. Error terkecil (*e*) berupa batasan nilai yang membuat perulangan akan berakhir setelah diperoleh nilai *error* yang diharapkan (e= 10^{-6})
e. Fungsi objektif awal (p_0=0) adalah suatu fungsi yang akan dioptimumkan, nilai 0 berarti untuk memperoleh nilai minimum

3. Membangkitkan bilangan acak μ_{ik} (i= 1,2,...,n; k= 1,2,...,c) sebagai elemen matriks partisi awal **U**. Elemen setiap baris (kelompok/klaster) mempunyai jumlahan sama dengan satu.

$$\mathbf{U} = \begin{bmatrix} \mu_{11}(x_1) & \mu_{12}(x_1) & \cdots & \mu_{1c}(x_1) \\ \mu_{21}(x_2) & \mu_{22}(x_2) & \cdots & \mu_{2c}(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{n1}(x_n) & \mu_{n2}(x_n) & \cdots & \mu_{nc}(x_n) \end{bmatrix} \qquad (2.14)$$

4. Menghitung sentroid dari masing-masing kelompok.

$$v_{kj} = \frac{\sum_{i=1}^{n} (\mu_{ik})^w x_{ij}}{\sum_{i=1}^{n} (\mu_{ik})^w} \qquad (2.15)$$

Keterangan:

n : banyaknya pengamatan
μ_{ik} : keanggotaan objek ke-*i* dan klaster ke-*k*
v_{kj} : sentroid atau rata-rata klaster ke-k untuk variabel ke-*j*
w : *weighting exponent*
x_{ij} : nilai objek ke-*i* yang ada di dalam klaster tersebut untuk variabel ke-*j*

5. Menentukan kriteria penghentian iterasi dengan melihat perubahan partisi matriks pada iterasi sekarang dan iterasi sebelumnya. Bila $|U^l - U^{(l-1)}| < \varepsilon$ maka proses iterasi berhenti.

$$U^l = \sum_{i=1}^{n} \sum_{k=1}^{c} \left(\left[\sum_{j=1}^{m} (x_{ij} - v_{kj})^2 \right] (\mu_{ik})^w \right) \qquad (2.16)$$

Keterangan:

m : banyaknya atribut
l : iterasi ke-*t*
U : derajat keanggotaan (Bezdek, Ehrlich, & Full, 1984).

6. Menghitung derajat keanggotaan setiap pengamatan pada setiap kelompok.

$$\mu_{ik} = \frac{[\sum_{j=1}^{m}(x_{ij} - v_{kj})^2]^{-\frac{1}{w-1}}}{\sum_{k=1}^{c}[\sum_{j=1}^{m}(x_{ij} - v_{kj})^2]^{-\frac{1}{w-1}}} \quad (2.17)$$

Penentuan jumlah kelompok optimum pada pengelompokan data berskala numerik dapat dilakukan dengan menggunakan nilai *pseudo-f statistic statistic* (Orpin & Kostylev, 2006). Nilai *pseudo-f* tertinggi menunjukkan bahwa jumlah kelompok telah optimal, dimana keragaman antar anggota dalam kelompok sangat homogen dan keragaman antar anggota kelompok lain sangat heterogen. Rumus untuk menghitung nilai *pseudo-f statistic* ditampilkan pada persamaan 2.18.

$$Pseudo-f = \frac{\left(\frac{R^2}{c-1}\right)}{\left(\frac{1-R^2}{n-c}\right)} \quad (2.18)$$

dengan

$$R^2 = \frac{(SST - SSW)}{SST} \quad (2.19)$$

$$SST = \sum_{l=1}^{m_{num}} \sum_{i=1}^{n} (x_{il} - \bar{x}_l)^2 \quad (2.20)$$

$$SSW = \sum_{c=1}^{C} \sum_{l=1}^{m_{num}} \sum_{i=1}^{n_c} (x_{ilc} - \bar{x}_{lc})^2 \quad (2.21)$$

Keterangan:
SST : total jumlah dari kuadrat jarak objek terhadap rata-rata keseluruhan
SSW : total jumlah dari kuadrat jarak objek terhadap rata-rata kelompoknya
n : banyak objek
n_c : jumlah anggota pada kelompok ke-*c* untuk *c*= 1,2,..., *C*

c : banyak klaster

m_{num} : banyak variabel numerik dalam pengamatan

$\bar{x}_l$: rata-rata seluruh objek pada variabel ke-*l* ; *l*= 1,2,..., m_{num}

$\bar{x}_{lc}$: rata-rata variabel ke-*l* pada kelompok ke-*c*

Setelah diperoleh jumlah kelompok optimum dalam setiap bentuk fungsi keanggotaan, selanjutnya dapat ditentukan bentuk fungsi keanggotaan yang optimum berdasarkan nilai *Internal Cluster Dispersion (ICD) Rate*. Berikut adalah rumus perhitungan *ICD Rate.*

$$ICD\ Rate = 1 - \frac{(SST - SSW)}{SST} = 1 - R^2 \qquad (2.22)$$

Berikut ini akan diberikan ilustrasi dalam perhitungan nilai *pseudo-f statistics* dan juga nilai ICD *Rate*. Data yang digunakan dalam ilustrasi merupakan data yang sama pada Tabel 2.1 dan diketahui terbentuk 2 kelompok dari data tersebut.

1. Menghitung jumlahan dari kuadrat jarak objek terhadap rata-rata keseluruhannya. Perhitungan dilakukan berdasarkan persamaan 2.20.

Klaster	**X_1**	**X_2**	**X_3**	$(X_{i1} - \bar{X}_1)^2$	$(X_{i2} - \bar{X}_2)^2$	$(X_{i3} - \bar{X}_3)^2$
1	2.5	4	3	0.0324	0.0625	0.7396
1	1.8	2	2.3	0.7744	3.0625	2.4336
2	4	3	2	1.7424	0.5625	3.4596
2	2	4	5	0.4624	0.0625	1.2996
2	1	8	8	2.8224	18.0625	17.1396
2	6.3	2	4	13.1044	3.0625	0.0196
1	0.8	3	3.3	3.5344	0.5625	0.3136
2	1.2	3	6	2.1904	0.5625	4.5796
1	2.9	4	2	0.0484	0.0625	3.4596
2	4.3	4.5	3	2.6244	0.5625	0.7396
$\bar{X}_l$	**2.68**	**3.75**	**3.86**		**SST**	**88.145**

Nilai SST diperoleh dengan menjumlahkan seluruh kuadrat jarak objek terhadap rata-ratanya. Sehingga diperoleh nilai SST sebesar 88.145.

2. Menghitung jumlahan kuadrat dari jarak objek terhadap rata-rata kelompoknya sesuai dengan persamaan 2.21. Jadi perhitungan dilakukan pada masing-masing kelompok.

Klaster	**X_1**	**X_2**	**X_3**	$(X_{i11} - \bar{X}_{11})^2$	$(X_{i21} - \bar{X}_{21})^2$	$(X_{i31} - \bar{X}_{31})^2$

1	2.5	4	3	0.25	0.5625	0.1225
1	1.8	2	2.3	0.04	1.5625	0.1225
1	0.8	3	3.3	1.44	0.0625	0.4225
1	2.9	4	2	0.81	0.5625	0.4225
$\bar{X}_{l1}$	**2**	**3.25**	**2.65**			

Klaster	X_1	X_2	X_3	$(X_{i12} - \bar{X}_{12})^2$	$(X_{i22} - \bar{X}_{22})^2$	$(X_{i32} - \bar{X}_{32})^2$
2	4	3	2	0.751111	1.173611	7.111111
2	2	4	5	1.284444	0.006944	0.111111
2	1	8	8	4.551111	15.34028	11.11111
2	6.3	2	4	10.02778	4.340278	0.444444
2	1.2	3	6	3.737778	1.173611	1.777778
2	4.3	4.5	3	1.361111	0.173611	2.777778
$\bar{X}_{l2}$	**3.13**	**4.08**	**4.67**			

Nilai SSW diperoleh dengan menjumlahkan seluruh kuadrat jarak objek terhadap rata-rata kelompoknya. Sehingga diperoleh nilai SSW dari klaster 1 dan klaster 2 sebesar 73.635.

3. Selanjutnya menghitung nilai R^2 berdasarkan persamaan 2.19.

$$R^2 = \frac{(SST - SSW)}{SST} = \frac{(88.145 - 73.635)}{88.145} = 0.164615$$

4. Diketahui dari data terdapat 10 objek dan 2 kelompok terbentuk. Setelah diperoleh nilai R^2, maka nilai *pseudo-f* dapat dihitung.

$$pseudo - f = \frac{\left(\frac{R^2}{c-1}\right)}{\left(\frac{1-R^2}{n-c}\right)} = \frac{\left(\frac{0.164615}{2-1}\right)}{\left(\frac{1-0.164615}{10-2}\right)} = 1.576424$$

5. Sedangkan ICD *Rate* dapat dihitung berdasarkan persamaan 2.22.

$$ICD\ rate = 1 - R^2 = 1 - 0.164615 = 0.835385$$

2.3.2 K-Modes

K-Modes merupakan pengembangan dari metode *C-Means* agar dapat digunakan untuk mengelompokkan data berskala kategorik. Metode ini pertama kali dikenalkan oleh Huang (1998). Modifikasi yang dilakukan pada *C-Means* sehingga menghasilkan metode *K-Modes* antara lain yaitu mengganti fungsi jarak *euclidean* dengan jarak *simple matching*. Selain itu menggantikan nilai *mean* dalam *C-Means* menjadi nilai modus untuk menentukan pusat klaster. Setiap objek pengamatan akan dibandingkan dengan pusat

klaster menggunakan jarak *simple matching* lalu setiap objek akan dialokasikan ke kelompok terdekat. Berikut merupakan perhitungan jarak *simple matching*.

$$d(X,Y) = \sum_{j=1}^{m} \delta(x_j, y_j) \tag{2.23}$$

Dimana

$$\delta(x_j, y_j) = \begin{cases} 0, & x_j = y_j \\ 1, & x_j \neq y_j \end{cases}$$

Ukuran ketidaksamaan antara kedua objek X dan Y dapat diartikan oleh total ketidakcocokan dari X dan Y. Semakin kecil nilai ketidakcocokan, semakin mirip kedua objek tersebut. Adapun algoritma dari metode *K-Modes* adalah sebagai berikut:

1. Menentukan *k* sebagai pusat awal klaster untuk masing-masing kelompok
2. Menghitung jarak antara masing-masing objek dengan pusat awal klaster menggunakan jarak *simple matching* pada persamaan (2.23).
3. Mengalokasikan setiap objek ke suatu klaster dengan melihat jarak pusat klaster terdekat.
4. Memperbarui pusat klaster dari masing-masing kelompok dengan nilai yang sering muncul di tiap kelompok tersebut.
5. Apabila masih ada objek yang berpindah kelompok atau ada perubahan nilai fungsi objektif yang berada di bawah ambang batas yang ditentukan, maka diulangi mulai langkah (2) hingga tidak ada objek yang berpindah kelompok.

Berikut ini merupakan ilustrasi untuk pengelompokan data kategorik dengan menggunakan algoritma *K-Modes*. Apabila terdapat 5 objek dengan 3 variabel, maka dalam mengelompokannya menjadi 2 kelompok dilakukan langkah-langkah sebagai berikut.

Tabel 2.7 Ilustrasi Data Kategorik

Objek	**X_1**	**X_2**	**X_3**
1	1	1	2
2	2	2	2
3	2	1	3
4	3	2	3
5	2	1	3

Pertama kali perlu ditentukan pusat awal klaster untuk kedua kelompok yang akan dibentuk. Penentuan pusat awal klaster dilakukan secara acak. Pada ilustrasi ini objek 3 dan objek 4 dianggap sebagai pusat awal klaster 1 dan klaster 2. Kemudian dihitung jarak antara tiap objek terhadap masing-masing pusat awal klaster sesuai dengan persamaan 2.23.

Tabel 2.8 Hasil Perhitungan Jarak

Objek	X_1	X_2	Jarak	
			1	**2**
1	1	2	2	2
2	2	2	1	2
3	**2**	**3**	0	1
4	**3**	**3**	1	0
5	3	3	1	0
6	1	1	2	2
7	3	1	2	1
8	2	1	1	2
9	1	2	2	2
10	3	2	2	1

Misalnya akan dihitung jarak antara objek 1 (X) terhadap pusat awal klaster 1 yaitu objek 3 (Y).

$$d(1,3) = \delta(1_{X_1}, 3_{X_1}) + \delta(1_{X_2}, 3_{X_2})$$
$$= \delta(1,2) + \delta(2,3)$$
$$= 1 + 1 = 2$$

Apabila nilai x_j sama dengan y_j maka delta sama dengan nol dan ketika nilai x_j tidak sama dengan y_j maka bernilai 1. Diketahui variabel 1 pada objek 1 bernilai 1 dan pada objek 3 bernilai 2, karena memiliki nilai yang berbeda maka delta bernilai 1. Sehingga diperoleh nilai jarak antara objek 1 terhadap pusat awal klaster 1 bernilai 2.

Tabel 2.9 Hasil Penentuan Pengelompokan

Objek	X_1	X_2	Jarak		Klaster
			1	**2**	
1	1	2	2	2	1
2	2	2	1	2	1
3	**2**	**3**	0	1	1
4	**3**	**3**	1	0	2

Tabel 2.9 Hasil Penentuan Pengelompokan (Lanjutan)

Objek	X_1	X_2	Jarak		Klaster
			1	**2**	
5	3	3	1	0	2
6	1	1	2	2	1
7	3	1	2	1	2
8	2	1	1	2	1
9	1	2	2	2	1
10	3	2	2	1	2

Penentuan suatu objek termasuk ke dalam kelompok yang mana dilakukan dengan melihat nilai jarak terdekat. Misalnya objek 2 memiliki jarak terhadap pusat awal klaster 1 sebesar 1 dan terhadap pusat awal klaster 2 sebesar 2, nilai jarak terdekat adalah terhadap klaster 1 maka objek dialokasikan pada klaster 1. Apabila terdapat nilai jarak yang sama, maka pengalokasian klaster dilakukan secara bebas. Selanjutnya dilakukan penentuan pusat klaster dengan melihat nilai yang sering muncul di tiap klaster tersebut. Algoritma tersebut diulangi hingga tidak ada objek yang berpindah kelompok. Berikut merupakan hasil akhir dari pengelompokan tersebut.

Objek	X_1	X_2	Jarak		Klaster
			1	**2**	
1	1	2	2	0	2
2	2	2	2	1	2
3	2	3	1	2	1
4	**3**	**3**	0	2	1
5	3	3	0	2	1
6	1	1	2	1	2
7	3	1	1	2	1
8	2	1	2	2	1
9	**1**	**2**	2	0	2
10	3	2	1	1	1

Pada metode ini untuk menentukan kelompok optimum, digunakan tingkat akurasi dari hasil pengelompokan. Berikut adalah persamaan yang dapat digunakan (Huang & Ng, 1999).

$$r = \frac{1}{n}\sum_{l=1}^{k} a_l \tag{2.24}$$

Keterangan:

n : banyaknya pengamatan

a_l :banyaknya objek yang mendominasi pada kelompok ke-l (l= 1,2,...,k).

Nilai a_l diperoleh dari hasil pengelompokan yang sebelumnya dikelompokkan dengan memperhatikan persamaan 2.23. Dihitung dengan melihat berapa banyak kategori yang sering muncul pada tiap kelompok di masing-masing variabel.

Berikut ini diberikan ilustrasi dalam menghitung akurasi pada hasil pengelompokan. Perhitungan akurasi menggunakan hasil akhir dari ilustrasi pengelompokan *k-modes*.

Klaster 1			**Klaster 2**		
Objek	**X_1**	**X_2**	**Objek**	**X_1**	**X_2**
3	2	3	1	1	2
4	3	3	2	2	2
5	3	3	6	1	1
7	3	1	9	1	2
8	2	1			
10	3	2			

Pada klaster 1 diketahui terdapat 6 objek dan pada klaster 2 terdapat 4 objek. Kemudian pada tiap klaster dilihat nilai modus pada masing-masing variabelnya dan dihitung berapa banyaknya. Misalnya pada klaster 1 variabel X_1 diketahui bahwa nilai 3 merupakan modusnya dengan frekuensi sebesar 4 dan pada variabel X_2 nilai modusnya adalah 3 dengan frekuensi sebanyak 3. Sedangkan pada klaster 2 variabel X_1 memiliki nilai modus 1 dengan frekuensi sebesar 3 dan pada variabel 2 nilai modusnya adalah 2 dengan frekuensi sebanyak 3.

Klaster	**X_1**	**X_2**
1	4	3
2	2	2
Jumlah	6	5
Rata-rata	6/10	5/10

Tabel tersebut merupakan banyaknya objek yang mendominasi pada masing-masing variabel di tiap klaster. Selanjutnya untuk memperoleh nilai akurasi nilai rata-rata pada

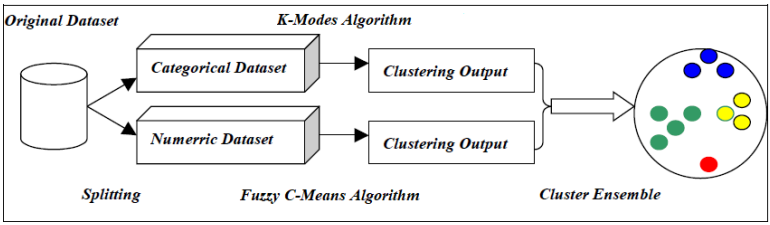
Tabel tersebut dirata-rata, sehingga diperoleh nilai akurasi sebesar 0.55 atau 55%.

2.3.3 *Ensemble Cluster*

Ensemble cluster merupakan metode yang menggabungkan beberapa algoritma berbeda untuk mendapatkan partisi umum dari hasil pengelompokan individu (He, Xu, & Deng, 2005). Tujuan dari metode ini adalah menggabungkan hasil pengelompokan dari beberapa algoritma pengelompokan untuk mendapatkan hasil pengelompokan yang lebih baik. Dengan menggunakan *ensemble cluster* dapat meningkatkan akurasi dan stabilitas dari pengklasifikasian pada *data mining* (Yoon, Ahn, Lee, Cho, & Kim, 2006). *Ensemble cluster* merupakan alternatif lain untuk mengatasi kasus pengelompokan yang memiliki 2 tipe data (numerik dan kategorik). Pada penelitian ini digunakan 2 metode pengelompokan data campuran yaitu *ensemble k-modes* dan SWFM.

2.3.3.1 *Ensemble K-Modes*

Pengelompokan menggunakan metode *ensemble cluster* ada beberapa macam penerapannya, salah satunya menggunakan prinsip pengelompokan data kategorik (He, Xu, & Deng, 2005). *Ensemble K-Modes* merupakan metode penggabungan dengan menggunakan algoritma data kategorik yaitu *K-Modes* untuk memperoleh kelompok akhir dari sebuah pengelompokan data campuran. Pada penelitian ini digunakan hasil penggabungan dari FCM untuk data numerik dan *K-Modes* pada data kategorik. Berikut merupakan skema pengelompokan menggunakan algoritma *ensemble K-Modes*.



Gambar 2.5 Skema Metode Ensemble K-Modes

Langkah-langkah yang dilakukan dalam metode pengelompokan *ensemble k-modes* pada penelitian ini adalah sebagai berikut:

1. Data awal yang terdiri dari variabel campuran (numerik dan kategorik) dibagi menjadi dua sub data, yaitu variabel murni numerik dan variabel murni kategorik.
2. Melakukan pengelompokan objek penelitian sesuai dengan skala data masing-masing. Dalam penelitian ini, variabel dengan skala data numerik dikelompokkan menggunakan algoritma FCM dan hasilnya dinyatakan sebagai *output* 1. Sementara pada variabel berskala data kategorik dikelompokkan dengan algoritma *K-Modes* hasilnya dinyatakan sebagai *output 2*.
3. Menggabungkan hasil pengelompokan dari langkah 2, proses ini disebut *ensemble*. Pada langkah ini, *output* 1 dan *output* 2 diperlakukan sebagai variabel kategorik yang akan digunakan untuk mendapatkan kelompok akhir. Penggabungan dilakukan dengan *output* 1 diletakkan pada kolom 1 dan dianggap sebagai variabel baru 1, sedangkan *output* 2 diletakkan pada kolom 2 dianggap sebagai variabel baru 2.

Tabel 2.10 Ilustrasi Hasil Penggabungan (*Ensemble*)

Objek	***Output* 1 (*Fuzzy C-Means*)**	***Output* 2 (*K-Modes*)**
1	1	1
2	2	1
3	3	3
⋮	⋮	⋮
n	2	1

4. Melakukan pengelompokan menggunakan data dari hasil langkah 3 dengan algoritma variabel berskala kategorik yaitu *K-Modes*.

2.3.3.2 *Similarity Weight and Filter Method* (SWFM)

Metode pengelompokan menggunakan SWFM memiliki konsep analisis yang sama dengan pengelompokan *ensemble cluster* pada umumnya. Metode ini merupakan perkembangan dari pengelompokan *ensemble* yang memiliki perbedaan pada tahapan pembentukan *final cluster*. Dalam metode *ensemble* SWFM, pembentukan *final* cluster dilakukan dengan menggunakan algoritma yang berdasarkan dengan *similarity weight* dan *filter method*.

Tahapan metode *similarity weight* menggunakan ukuran kemiripan dengan memasukkan faktor bobot pada rumus ukuran kemiripan. Bobot yang diberikan didasarkan dengan jumlah anggota pengamatan (n_i atau n_j). Rumus yang dapat digunakan untuk menghitung ukuran kemiripan antar pasangan objek ke-*i* dan objek ke-*j* ditunjukkan pada persamaan (2.25) (Reddy & Khavita, 2012).

$$sim(X_i, X_j) = \sum_{i \le n_i, j \le n_j} \frac{S_{ij}}{\max(n_i, n_j)}, i \neq j \qquad (2.25)$$

Keterangan:

$S_{ij} = \left| \frac{X_i \cap X_j}{X_i \cup X_j} \right|$

X_i : himpunan pengamatan ke-*i* dengan $X_i = \{x_{1i}, x_{2i}, x_{3i}, \ldots, x_{ki}\}$

X_j : himpunan pengamatan ke-*j* dengan $X_j = \{x_{1j}, x_{2j}, x_{3j}, \ldots, x_{kj}\}$

n_i : jumlah anggota dalam kelompok ke-*i* dengan $i = 1,2,...,n_i$

n_j : jumlah anggota dalam kelompok ke-*j* dengan $j = 1,2,...,n_j$

Untuk mengelompokkan data berskala campuran digunakan metode *filter* yang didasarkan pada *similarity weight.* Algoritmanya yaitu pertama membagi data menjadi dua sub-data yaitu numerik murni dan kategorik murni. Pada masing-masing skala data dilakukan pengelompokan sesuai jenis data untuk mendapatkan kelompok yang sesuai. Selanjutnya hasil pengelompokan dari kedua jenis data digabungkan untuk mendapatkan kelompok akhir (*final cluster*) menggunakan metode *filter*. Untuk metode *filter* digunakan persamaan sebagai berikut.

$$F(X_i, X_j) = \sum_{i=1}^{n_i} \sum_{j=1}^{n_j} w_{ij} d(X_i, X_j) \qquad (2.26)$$

dimana

$d(X_i, X_j) = 1 - sim(X_i, X_j)$

W_{ij} : bobot antara kelompok ke-*i* dan ke-*j* (umumnya bernilai 0.5)

Berikut ini akan diberikan ilustrasi pengelompokan menggunakan metode SWFM.

1. Menggabungkan hasil pengelompokan variabel numerik dan variabel kategorik yang dianggap sebagai variabel baru dengan skala data kategorik. Berikut merupakan data yang digunakan pada ilustrasi.

Objek	A	B	C	D	E
X1	1	2	3	3	2
X2	2	2	1	2	2

2. Menghitung similaritas atau kesamaan antar objek menggunakan persamaan 2.25.

$$sim(A,B) = \frac{A \cap B}{A \cup B} = \frac{n(A \cap B)}{n(A) + n(B) - n(A \cap B)}$$
$$= \frac{1}{2+2-1} = 0.333$$

Sehingga diketahui similaritas antar objek A dan B adalah 0.333. Perhitungan similaritas dilakukan hingga diperoleh seluruh similaritas antar objek. Matriks similaritas disajikan dalam matriks berikut.

$$\mathbf{sim} = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{pmatrix} 1 & & & & \\ 0,33 & 1 & & & \\ 0 & 0 & 1 & & \\ 0 & 0,33 & 0,33 & 1 & \\ 0,33 & 1 & 0 & 0,33 & 1 \end{pmatrix}$$

3. Metode *filter* dihitung menggunakan persamaan 2.26. Bobot yang digunakan (*w*) pada penilitian ini yaitu sebesar 0.5. Diperoleh hasil perhitungan metode *filter* antar objek ditampilan pada matriks **F** berikut.

$$\mathbf{F} = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{pmatrix} 0 & & & & \\ 0,33 & 0 & & & \\ 0,5 & 0,5 & 0 & & \\ 0,5 & 0,33 & 0,33 & 0 & \\ 0,33 & 0 & 0,5 & 0,33 & 0 \end{pmatrix}$$

4. Output dari metode *filter* ini berupa matriks jarak (matriks **F**). Sehingga untuk mengelompokkan data yang berupa matriks jarak, digunakan metode pengelompokan hirarki. Dalam penelitian ini digunakan metode pengelompokan *single linkage* karena memperhatikan jarak minimum antar objek yang akan dikelompokkan. Pengelompokan dilakukan hingga diperoleh jumlah kelompok yang diinginkan.

Kelebihan dari metode SWFM adalah penggunaan yang efisien untuk beberapa dimensi serta mengurangi kompleksitas waktu, dan pencarian jarak yang digunakan lebih efisien bahkan jika batas-batas kelompok tidak teratur (Reddy & Khavita, 2012).

2.4 *One Way Analysis of Variance* (ANOVA)

ANOVA dilakukan untuk mengetahui perbedaan rata-rata antar klaster atau kelompok yang terbentuk. Teknik ANOVA merupakan teknik univariat, dimana perbedaan rata-rata antar kelompok terbentuk dilihat dari masing-masing variabelnya. Hipotesis yang digunakan dalam pengujian ANOVA adalah sebagai berikut

H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_c = \boldsymbol{\mu}$

H_1: minimal ada satu $\boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j$

Dengan statistik uji yang digunakan yaitu

$$F = \frac{MSB}{MSW} \tag{2.27}$$

Dengan

$$MSB = \frac{(SST - SSW)}{k-1} \tag{2.28}$$

$$MSW = \frac{SSW}{N-k} \tag{2.29}$$

Adapun SST dan SSW diperoleh dari persamaan 2.20 dan 2.21. Pengambilan keputusan pada pengujian ini adalah Tolak H_0 jika $F > F_{\alpha;k-1;N-k}$.

2.5 *One Way Multivariate Analysis of Variance* (MANOVA)

Pengujian perbedaan rata-rata dilakukan setelah didapatkan hasil pengelompokan yang optimum. Pengujian rata-rata pada data multivariat dilakukan menggunakan *Multivariate Analysis of Variance* (MANOVA). MANOVA merupakan perluasan dari teknik univariat *Analysis of Variance* (ANOVA) yang melibatkan lebih dari satu variabel (Johnson & Wichern, 2007). Hipotesis yang digunakan dalam pengujian MANOVA adalah sebagai berikut

H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_c = \boldsymbol{\mu}$

H_1: minimal ada satu $\boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j$

Statistik uji dalam pengambilan keputusan untuk melihat perbedaan antar kelompok adalah *wilk's lambda* yang bernilai antara 0 sampai 1. Semakin rendah nilai *wilk's lambda*, maka perbedaan kelompok semakin signifikan. Berikut adalah rumusan untuk statistik uji *wilk's lambda*.

$$\Lambda^* = \frac{|\mathbf{W}|}{|\mathbf{W}+\mathbf{B}|} \tag{2.30}$$

Dimana

$$\mathbf{W} = \sum_{l=1}^{g}\sum_{j=1}^{n_l}\left(x_{lj}-\bar{x}_l\right)\left(x_{lj}-\bar{x}_l\right)' \tag{2.31}$$

$$\mathbf{B} = \sum_{l=1}^{g} n_l\left(\bar{x}_l-\bar{x}\right)\left(\bar{x}_l-\bar{x}\right)' \tag{2.32}$$

Keterangan:

x_{lj} : objek pengamatan ke-*j* pada kelompok ke-*l*

$\bar{x}_l$: rata-rata seluruh objek pada kelompok ke-*l*

$\bar{x}$: rata-rata keseluruhan objek pengamatan

Daerah penolakan dari uji *Wilks Lambda* yaitu tolak H_0 jika nilai $\Lambda^* > U_{\alpha;(g-1,\sum n-g)}$.

Penggunaan *Wilks Lambda* tidak akan menjadi lebih baik bila asumsi matriks kovarian homogen tidak terpenuhi. Oleh karena itu, statistik uji yang tepat digunakan adalah *Pillai's Trace. Pillai's Trace* sedikit lebih *robust* terhadap data tidak normal dan juga matriks kovarian tidak homogen (Johnson & Wichern, 2007). Berikut merupakan rumus dari *Pillai's Trace*

$$V = tr[(\mathbf{W}+\mathbf{B})^{-1}\mathbf{B}] \tag{2.33}$$

Nilai statistik *Pillai's Trace* dapat diketahui dengan statistik uji F. Diperoleh keputusan tolak H_0 jika nilai $V^* > F_{\alpha;c-1;n-c}$.

Sebelum melakukan analisis dengan MANOVA, data yang dimiliki harus berdistribusi normal mutivariat dan memiliki matriks kovarian yang sama (homogen). Salah satu nilai yang dapat digunakan untuk menguji apakah data berdistribusi normal multivariat adalah dengan menggunakan nilai *saphiro wilk* menggunakan hipotesis dan statistik uji seperti pada persamaan (2.3)

Uji homogenitas matriks varian kovarian dilakukan untuk mendeteksi dua atau lebih kelompok data sampel dari populasi memiliki matriks varian kovarian yang homogen. Pengujian dilakukan dengan menggunakan *Box's M Test*, dengan hipotesis pengujian sebagai berikut (Rencher, 1998).

H_0: $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \ldots = \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$ (matriks varians kovarians bersifat homogen)

H_1: minimal ada satu $\boldsymbol{\Sigma}_i \neq \boldsymbol{\Sigma}_j$ (matriks varians kovarians tidak bersifat homogen)

Statistik uji pada *Box's M test* adalah sebagai berikut.

$$c_1 = \frac{(k+1)(2p^2+3p-1)}{6kv(p+1)} \tag{2.34}$$

$$c_2 = \frac{(p-1)(p+2)}{6(k-1)}\left[\sum_{i=1}^{k}\frac{1}{v_i^2} - \frac{1}{\left(\sum_{i=1}^{k} v_i\right)^2}\right] \tag{2.35}$$

$$a_1 = \frac{1}{2}(k-1)p(p+1), \qquad a_2 = \frac{a_1+2}{|c_2 - c_1^2|}$$
$$b_1 = \frac{1-c_1-a_1/a_2}{a_1}, \qquad b_2 = \frac{1-c_1+2/a_2}{a_2} \tag{2.36}$$

Dengan kriteria

Jika $c_2 > c_1^2$ maka,

$$F = -2b_1 \ln \mathbf{M} \tag{2.37}$$

Jika $c_2 < c_1^2$ maka,

$$F = \frac{2a_2 b_2 \ln \mathbf{M}}{a_1(1+2b_2 \ln \mathbf{M})} \tag{2.38}$$

Dimana

$$\ln \mathbf{M} = \frac{1}{2}\sum_{i=1}^{k} v_i \ln|\mathbf{S}_i| - \frac{1}{2}\left(\sum_{i=1}^{k} v_i\right)\ln|\mathbf{S}_{pl}| \tag{2.39}$$

$$\mathbf{S}_{pl} = \frac{\sum_{i=1}^{k} v_i \mathbf{S}_i}{\sum_{i=1}^{k} v_i} \tag{2.40}$$

dengan

$$\mathbf{S}_i = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{bmatrix}$$

Keterangan

v_i $= n_i - 1$

n_i : banyaknya data pada kelompok ke-*i* dengan i=1,2,...,k

k : banyaknya kelompok

$\boldsymbol{S}_{pl}$: matriks varians kovarians kelompok gabungan

p : jumlah variabel independen

$\boldsymbol{S}_i$: matriks varians kovarians kelompok ke-*i*

Daerah penolakan pada uji Box's M yaitu H_0 akan ditolak apabila $F > F_{\alpha(a_1,a_2)}$.

2.6 Uji Kruskal Wallis

Uji Kruskal Wallis digunakan untuk melihat apakah sampel berasala dari populasi yang sama atau identik. Pengujian ini memerlukan beberapa asumsi yaitu data terdiri dari sampel acak, pengamatan-pengamatan bebas baik di dalam maupun di antara sampel-sampel, dan skala pengukuran yang digunakan setidaknya ordinal (Daniel, 1989).

Hipotesis yang digunakan dalam pengujian Kruskal Wallis adalah

H_0: Kedua populasi identik

H_1: Tidak semua populasi identik

Dengan statistik uji dalam pengujian ini diperoleh dari perhitungan berikut

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{g} \frac{R_i^2}{n_i} - 3(N+1) \qquad (2.41)$$

Keterangan:

R_i : jumlah peringkat-peringkat yang ditetapkan bagi hasil-hasil pengamatan di kelompok ke-*i*

n_i : banyaknya pengamatan di tiap kelompok

N : banyaknya seluruh pengamatan

g : banyaknya kelompok

Dalam pengambilan keputusan digunakan nilai *H* yang dibandingkan dengan nilai $\chi^2_{(\alpha,g-1)}$. Sehingga diperoleh keputusan Tolak H_0 apabila nilai *H* lebih besar dari $\chi^2_{(\alpha,g-1)}$ atau nilai χ^2_{hitung} lebih besar daripada $\chi^2_{(\alpha,g-1)}$.

2.7 Daerah Tertinggal

Daerah tertinggal merupakan suatu daerah dengan kabupaten yang masyarakat dan wilayahnya relatif kurang berkembang dibandingkan daerah lain dalam skala nasional. Berdasarkan Peraturan Menteri Desa, Pembangunan Daerah Tertinggal, dan Transmigrasi Republik Indonesia nomor 2 Tahun 2016 tentang Indeks Desa Membangun, terdapat 5 derajat ketertinggalan daerah yang ditetapkan yaitu desa mandiri, desa maju, desa berkembang, desa tertinggal, dan desa sangat tertinggal. Sehingga dalam penelitian ini derajat ketertinggalan kabupaten/kota akan dibagi menjadi 5 sesuai dengan pembagian oleh pemerintah agar hasil dari pengelompokan dapat disesuaikan dengan penentuan status ketertinggalan pemerintah. Ketertinggalan daerah tersebut dapat diukur berdasarkan enam kriteria utama yaitu ekonomi, sumber daya manusia, infrastruktur, kapasitas keuangan daerah, aksesibilitas, dan karakteristik daerah. Oleh karena itu, diperlukan upaya pembangunan daerah yang terencana dan sistematis agar daerah tertinggal tersebut pada akhirnya setara dengan daerah lainnya di Indonesia yang telah maju terlebih dahulu (Bappenas, 2016). Berikut diberikan variabel-variabel yang mewakili 6 kriteria ketertinggalan daerah.

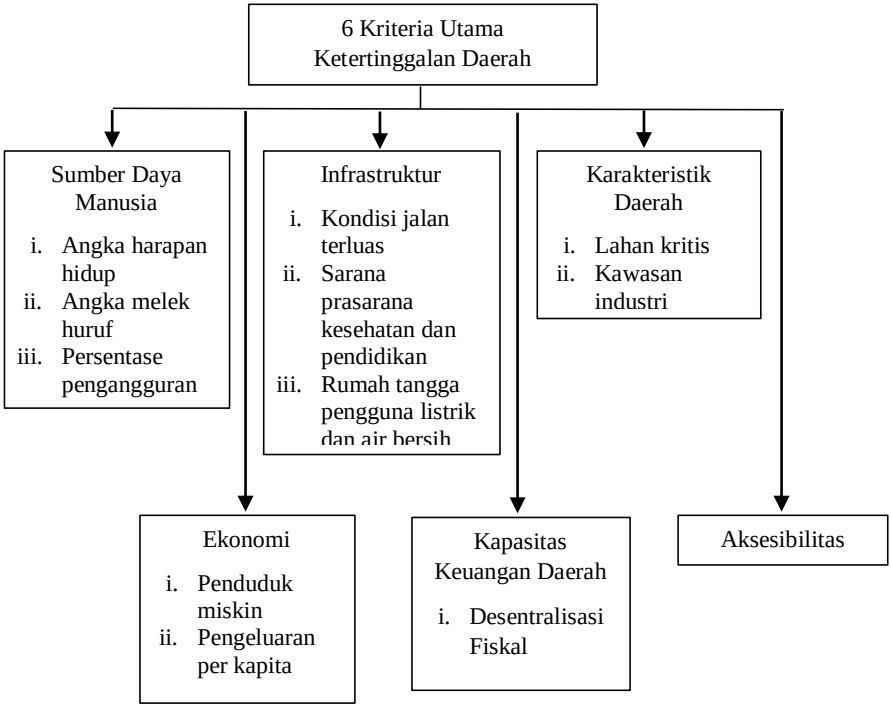
a. Ekonomi

Kriteria ini berhubungan dengan kondisi perekonomian penduduk di suatu daerah. Indikator-indikator yang mewakilinya adalah persentase penduduk miskin dan pengeluaran per kapita penduduk.

b. Sumber Daya Manusia

Kriteria sumber daya manusia menggambarkan bagaimana kualitas sumber daya manusia yang dimiliki masing-maisng daerah. Variabel yang mewakili kriteria sumber daya manusia antara lain angka harapan hidup, angka melek huruf, dan persentase pengangguran. Angka harapan hidup adalah alat untuk

mengevaluasi kinerja pemerintah dalam meningkatkan kesejahteraan penduduk pada umumnya dan meningkatkan derajat kesehatan pada khususnya.



Gambar 2.6 Kriteria Ketertinggalan Daerah

c. Infrastruktur

Kriteria ini mewakili kondisi infrastruktur dan fasilitas pendidikan, perdagangan, dan kesehatan di suatu daerah. Variabel yang mewakili yaitu kondisi jalan terluas, jumlah unit fasilitas kesehatan dan pendidikan, serta rumah tangga pengguna listrik dan air bersih.

d. Kapasitas Keuangan Daerah

Kapasistas keuangan daerah menunjukkan mengenai pengelolaan keuangan daerah. Desentralisasi fiskal dapat mewakili kemampuan keuangan daerah. Tingkat desentralisasi fiskal dihitung

dengan perbandingan antara Pendapatan Asli Daerah (PAD) dengan total penerimaan daerah. Nilai ini menunjukkan derajat kontribusi PAD terhadap total penerimaan daerah.

e. Aksesibilitas

Kriteria aksesibilitas menggambarkan kemudahan akses dalam menjangkau suatu daerah.

f. Karakteristik Daerah

Kriteria karakteristik daerah diwakili oleh beberapa indikator antara lain keberadaan kawasan industri dan lahan kritis.

BAB III
METODOLOGI PENELITIAN

3.1 Sumber Data

Sumber data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari Badan Pusat Statistik (BPS) Jawa Timur. Data yang digunakan berupa publikasi hasil Potensi Desa (PODES) dan Survei Sosial dan Ekonomi Nasional (Susenas) periode 2014 dengan unit pengamatan kabupaten/kota yang terbagi menjadi 29 kabupaten dan 9 kota. Pada penelitian ini data yang digunakan memiliki skala numerik dan kategorik.

Susenas merupakan salah satu survei yang dilaksanakan setiap tahun oleh BPS yang dirancang untuk dapat memenuhi kebutuhan data yang berkaitan dengan kualitas sumber daya manusia, khususnya yang berhubungan dengan karakteristik sosial ekonomi. Sehingga hasil dari Susenas dapat memberikan gambaran tingkat kesejahteraan masyarakat. Pengumpulan data PODES dilakukan sebanyak 3 (tiga) kali dalam kurun waktu 10 tahun, sebagai bagian dari siklus 10 tahunan kegiatan sensus yang dilakukan oleh BPS. Pendataan PODES dilakukan pada tahun-tahun berakhiran 1, 4, dan 8. Sehingga periode pendataan PODES yang terakhir adalah PODES 2014.

3.2 Variabel Penelitian

Penelitian ini menggunakan data campuran, yaitu data berskala numerik dan kategorik. Terdapat 33 variabel yang merupakan indikator-indikator ketertinggalan daerah. Berikut merupakan variabel yang digunakan dalam penelitian.

Tabel 3.1 Variabel Penelitian Numerik

No	Kriteria	Variabel	Keterangan
1	Infrastruktur	X_1	Rasio Desa dengan Permukaan Jalan Terluas Aspal
2		X_2	Rasio Rumah Sakit terhadap Jumlah Penduduk
3		X_3	Rasio SMA/MA terhadap Jumlah Penduduk Usia SMA sederajat
4		X_4	Persentase Rumah Tangga Pengguna Listrik

Tabel 3.1 Variabel Penelitian Numerik (Lanjutan)

No	Kriteria	Variabel	Keterangan
5	Karakteristik Daerah	X_5	Rasio Luas Lahan Kritis
6	Ekonomi	X_6	Persentase Penduduk Miskin
7		X_7	Pengeluaran Penduduk Perkapita
8	Sumber	X_8	Angka Harapan Hidup
9	Daya	X_9	Angka Melek Huruf
10	Manusia	X_{10}	Persentase Pengangguran

Tabel 3.2 Variabel Penelitian Kategorik

No	Kriteria	Variabel	Keterangan	Kategori
11	Karakteristik Daerah	X_{11}	Keberadaan Kawasan Industri	1: Ada 2: Tidak Ada
12	Kapasitas Keuangan Daerah	X_{12}	Tingkat Desentralisasi Fiskal	0: Sangat Kurang 1: Kurang 2: Sedang 3: Cukup 4: Baik 5: Sangat Baik
13	Infrastruktur	X_{13}	Sumber Air Minum	1: Air kemasan 2: Ledeng meteran 3: Ledeng tanpa meteran 4: Sumur bor/pompa 5: Sumur 6: Mata air

Berikut merupakan penjelasan masing-masing variabel yang digunakan pada penelitian ini.

1. Rasio desa dengan permukaan jalan terluas aspal adalah jumlah desa yang memiliki permukaan jalan terluas berupa aspal dibagi dengan banyaknya desa di masing-masing kabupaten/kota
2. Rasio rumah sakit terhadap jumlah penduduk merupakan ketersediaan rumah sakit dibagi dengan jumlah penduduk di masing-masing kabupaten/kota.
3. Rasio SMA/MA terhadap jumlah penduduk usia SMA sederajat adalah banyaknya SMA/MA dibagi dengan jumlah penduduk berusia 16-18 tahun.

4. Persentase rumah tangga pengguna listrik adalah rumah tangga yang terdaftar sebagai pengguna atau konsumen dari listrik PLN
5. Rasio luas lahan kritis merupakan luasan dari lahan di dalam maupun di luar kawasan hutan yang telah mengalami kerusakan, yang diakibatkan oleh berbagai faktor degradasi fisik, kimia, dan biologi tanah sehingga kehilangan atau berkurang fungsinya dibagi dengan luasan lahan total.
6. Penduduk miskin adalah penduduk dengan rata-rata pengeluaran per kapita per bulan dibawah garis kemiskinan
7. Pengeluaran penduduk per kapita adalah biaya yang dikeluarkan untuk konsumsi semua anggota rumah tangga selama sebulan dibagi dengan banyaknya anggota rumah tangga
8. Angka harapan hidup merupakan rata-rata tahun hidup yang masih akan dijalani oleh seseorang yang telah berhasil mencapai umur x, pada suatu tahun tertentu, dalam situasi mortalitas yang berlaku di lingkngan masyarakatnya.
9. Angka melek huruf berupa proporsi penduduk berusia 15 tahun ke atas yang memiliki kemampuan membaca dan menulis kalimat sederhana dalam huruf latin, huruf arab, dan huruf lainnya (seperti huruf jawa, kanji, dll) terhadap penduduk usia 15 tahun ke atas
10. Persentase pengangguran berupa tingkat pengangguran terbuka, yaitu persentase jumlah pengangguran terhadap jumlah angkatan kerja
11. Keberadaan kawasan industri adalah ada atau tidaknya kawasan industri di suatu kabupaten/kota. Kawasan industri merupakan kawasan tempat pemusatan kegiatan industri yang dilengkapi dengan prasarana dan sarana penunjang yang dikembangkan dan dikelola oleh perusahaan kawasan industri yang telah memiliki izin usaha kawasan industri.
12. Tingkat desentralisasi fiskal dihitung berdasarkan perbandingan antara pendapatan asli daerah (PAD) dengan total penerimaan daerah. Rasio ini menunjukkan derajat kontribusi PAD terhadap total penerimaan daerah.
13. Sumber air minum menjelaskan mengenai darimana asal air minum yang dikonsumsi masyarakat pada tiap daerah. Dilihat

dari mayoritas sumber air minum di masing-masing desa pada tiap daerah.

Struktur data dari variabel penelitian ditampilkan pada Tabel 3.3 berikut.

Tabel 3.3 Struktur Data

Kab/Kota	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$	**...**	$\mathbf{X_{13}}$
1	$X_{1,1}$	$X_{2,1}$	$X_{3,1}$	$X_{4,1}$	...	$X_{13,1}$
2	$X_{1,2}$	$X_{2,2}$	$X_{3,2}$	$X_{4,2}$	...	$X_{13,2}$
3	$X_{1,3}$	$X_{2,3}$	$X_{3,3}$	$X_{4,3}$	...	$X_{13,3}$
4	$X_{1,4}$	$X_{2,4}$	$X_{3,4}$	$X_{4,4}$	...	$X_{13,4}$
5	$X_{1,5}$	$X_{2,5}$	$X_{3,5}$	$X_{4,5}$	...	$X_{13,5}$
⋮	⋮	⋮	⋮	⋮	⋱	⋮
38	$X_{1,38}$	$X_{2,38}$	$X_{3,38}$	$X_{4,38}$	...	$X_{13,38}$

3.3 Langkah Analisis

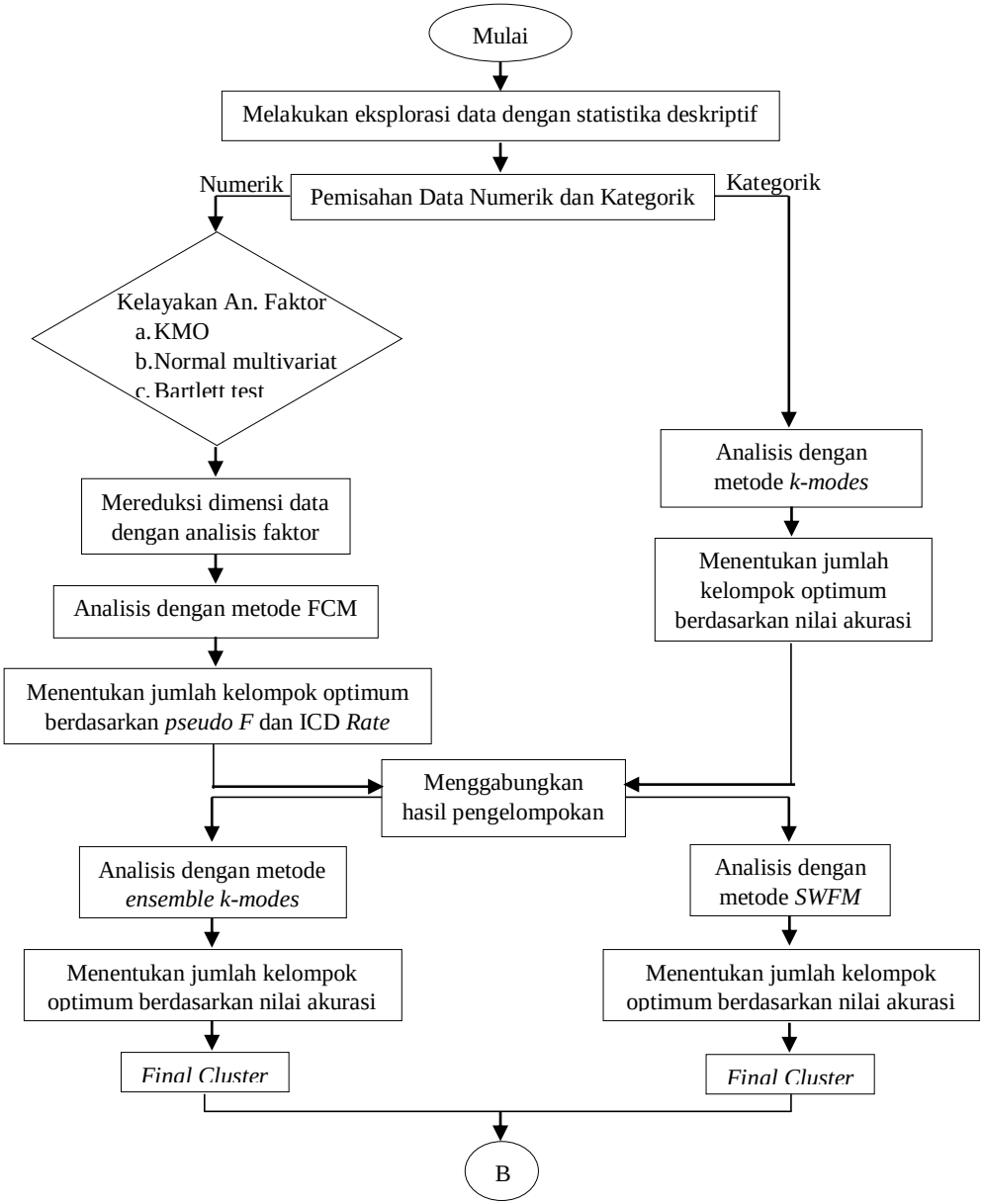
Langkah analisis yang dilakukan dalam penelitian ini yaitu sebagai berikut.

1. Mendeskripsikan karakteristik kabupaten/kota berdasarkan indikator daerah tertinggal sesuai dengan subbab 2.1.
2. Mengelompokkan kabupaten/kota berdasarkan indikator daerah tertinggal dengan metode data campuran *ensemble k-modes*, dilakukan langkah-langkah sebagai berikut.
   a. Melakukan pemisahan data numerik dan kategorik.
   b. Mereduksi variabel berskala data numerik menggunakan analisis faktor beserta asumsinya sesuai subbab 2.2. Hasil dari analisis faktor akan dipilih variabel dengan nilai *loading factor* tertinggi pada masing-masing faktor terbentuk. Variabel-variabel terpilih itulah yang selanjutnya akan digunakan pada analisis pengelompokan data numerik.
   c. Melakukan pengelompokan data numerik dengan metode *fuzzy c-means* menggunakan algoritma pada subbab 2.3.1. Penentuan derajat keanggotaan awal dilakukan dengan menghitung jarak antara objek dengan rata-rata tiap variabelnya. Jarak yang digunakan yaitu *euclidean* kuadrat. Jumlah klaster optimum ditentukan dengan melihat nilai *pseudo-f* terbesar sesuai persamaan 2.18 dan nilai *ICD Rate* pada persamaan 2.22 sehingga menghasilkan *output* 1 dengan skala data kategorik.

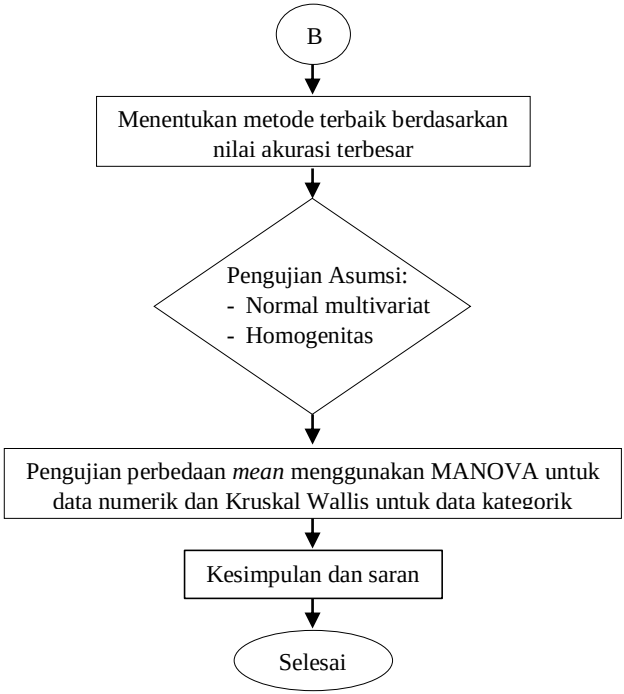
d. Melakukan pengelompokan data kategorik dengan metode *k-modes* menggunakan algoritma pada subbab 2.3.2 dan menentukan jumlah klaster optimum menggunakan nilai akurasi untuk data kategori sesuai persamaan 2.24 sehingga menghasilkan *output* 2 dengan skala data kategorik.
e. Menggabungkan hasil pengelompokan dari langkah (c) dan (d) dengan *output* 1 diletakkan pada kolom 1 dan *output* 2 pada kolom 2 seperti pada Tabel 2.10
f. Melakukan pengelompokan akhir dengan data dari langkah (e) menggunakan algoritma pengelompokan data kategorik *K-Modes*.

3. Mengelompokkan kabupaten/kota berdasarkan indikator daerah tertinggal dengan metode data campuran SWFM, dilakukan langkah-langkah sebagai berikut.
   a. Menyiapkan data yang diperoleh dari langkah (2.e).
   b. Melakukan pengelompokan menggunakan *similarity weighted* pada persamaan 2.25 serta algoritma *filter* pada persamaan 2.26.
   c. Matriks jarak yang dihasilkan oleh algoritma *filter* dikelompokkan menggunakan metode pengelompokan hirarki yaitu *single linkage*.
   d. Menentukan jumlah kelompok optimum dari langkah b dengan melihat akurasi.
4. Mendapatkan kelompok akhir dari pengelompokan dan karakteristiknya dilakukan dengan langkah sebagai berikut.
   a. Membandingkan hasil pengelompokan terbaik dari metode *ensemble k-modes* dan SWFM menggunakan nilai akurasi terbesar.
   b. Menggunakan metode MANOVA dan Kruskal Wallis sesuai dengan subbab 2.5 serta subbab 2.6 untuk mengetahui karakteristik kelompok yang dihasilkan berbeda atau tidak.

Langkah penelitian sebagaimana yang telah dijelaskan di atas dapat disajikan dalam diagram alir pada Gambar 3.1.



Gambar 3.1 Diagram Alir Penelitian



Gambar 3.1 Diagram Alir Penelitian (Lanjutan)

(Halaman Ini Sengaja Dikosongkan)

BAB IV
ANALISIS DAN PEMBAHASAN

Pada bab ini akan dijelaskan mengenai hasil analisis dan pembahasan. Berdasarkan hasil analisis akan diperoleh karakteristik indikator daerah tertinggal, pengelompokan Kabupaten/Kota di Jawa Timur, perbandingan hasil pengelompokan, dan karakteristik dari kelompok yang terbentuk.

4.1 Karakteristik Kabupaten/Kota di Jawa Timur

Pada subbab ini karakteristik Kabupaten/Kota di Jawa Timur diperoleh dari hasil analisis statistika deskriptif berdasarkan indikator daerah tertinggal. Analisis deskriptif dilakukan pada masing-masing variabel numerik dan variabel kategorik. Penelitian ini menggunakan 10 variabel numerik dan 3 variabel kategorik.

Tabel 4.1 Statistika Deskriptif Variabel Numerik

Variabel	Mean	Median	Min	Max
Rasio Jalan Aspal	0.9223	0.9493	0.4608	1.000
Rasio Rumah Sakit per 1000 penduduk	0.0056	0.0040	0.0010	0.0300
Rasio SMA/MA terhadap penduduk usia SMA	0.0025	0.0023	0.0013	0.0044
Presentase Pengguna Listrik PLN	98.953	99.730	81.040	100.00
Rasio Lahan Kritis	0.036	0.026	0.000	0.254
Presentase Penduduk Miskin	12.540	12.030	4.770	27.080
Pengeluaran per Kapita Penduduk	6.719	5.936	4.445	13.888
Angka Harapan Hidup	70.728	71.350	65.430	73.850
Angka Melek Huruf	92.301	94.005	77.930	98.520
Presentase Pengangguran	4.225	4.240	1.010	7.660

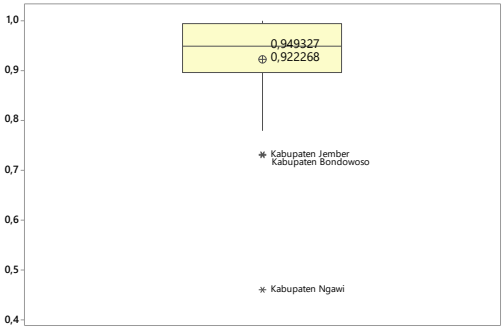
Tabel 4.1 menunjukkan rasio desa dengan permukaan jalan aspal di Jawa Timur memiliki rata-rata cukup tinggi yaitu sebesar 0.9223. Pada fasilitas kesehatan, terdapat rata-rata 6 rumah sakit yang mampu menampung 1000 penduduknya. Fasilitas pendidikan di Jawa Timur memiliki rata-rata 3 SMA yang mampu menampung 1000 penduduk usia SMA sederajat. Diketahui rata-rata persentase rumah tangga yang menggunakan listrik PLN adalah sebesar

98.953%. Rata-rata rasio lahan kritis di Jawa Timur bernilai cukup kecil yaitu 0.036 dengan 9 daerah tidak memiliki lahan kritis sama sekali.

Pada kriteria ekonomi diketahui Jawa Timur memiliki rata-rata penduduk miskin mencapai 12.54%. Persentase penduduk miskin tertinggi terdapat pada Kabupaten Sampang yaitu sebesar 27.08%. Sedangkan persentase penduduk miskin di Kota Batu adalah yang terendah yaitu 4.77%, hak tersebut menyebabkan adanya ketimpangan pada persentase penduduk miskin. Penduduk Jawa Timur memiliki rata-rata pengeluaran per kapita sebesar Rp 671.900,00. Kota Surabaya adalah daerah dengan pengeluaran per kapita tertinggi dan Kabupaten Lumajang memiliki pengeluaran per kapita terendah.

Kriteria sumber daya manusia menunjukkan Jawa Timur memiliki rata-rata angka harapan hidup sebesar 70.728 dan rata-rata angka melek huruf sebesar 92.301. Pengangguran di Jawa Timur memiliki rata-rata mencapai 4.225%. Kota Kediri merupakan daerah dengan persentase pengangguran tertinggi di Jawa Timur sebesar 7.66%.

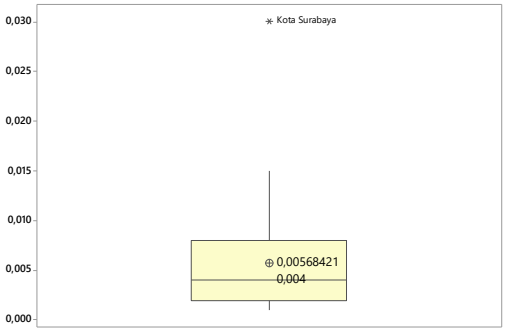
Secara visual, karakteristik masing-masing variabel numerik akan ditampilkan menggunakan boxplot pada Gambar 4.1 sampai dengan Gambar 4.10.



Gambar 4.1 Karakteristik Rasio Desa dengan Permukaan Jalan Terluas Aspal

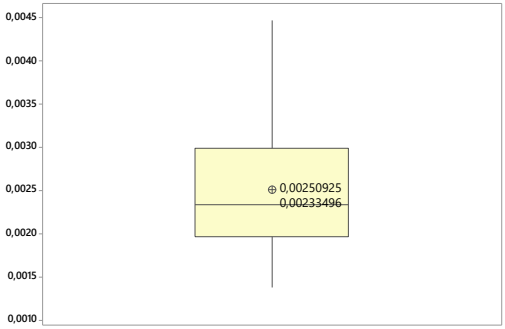
Diketahui dari Gambar 4.1 bahwa Kabupaten Ngawi merupakan data ekstrim, hal itu dikarenakan daerah tersebut memang sebagian besar permukaan jalannya belum berupa aspal melainkan masih berupa jalan tanah yang diperkeras. Selain itu

Kabupaten Jember dan Kabupaten Bondowoso merupakan data *outlier* karena daerah tersebut memiliki rasio lebih kecil dibandingkan dengan daerah lainnya. Rasio desa dengan permukaan jalan aspal maksimum dimiliki oleh seluruh wilayah Kota di Jawa Timur.



Gambar 4.2 Karakteristik Rasio Rumah Sakit

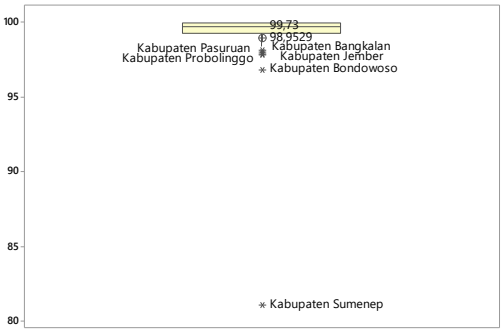
Gambar 4.2 menunjukkan Kota Surabaya merupakan data ekstrim yang berarti rasio rumah sakit pada daerah ini jauh lebih tinggi apabila dibandingkan dengan daerah lainnya. Kabupaten Magetan, Kabupaten Sampang, Kabupaten Sumenep, dan Kota Pasuruan memiliki rasio rumah sakit paling rendah dengan nilai sebesar 0.0010.



Gambar 4.3 Karakteristik Rasio SMA/MA

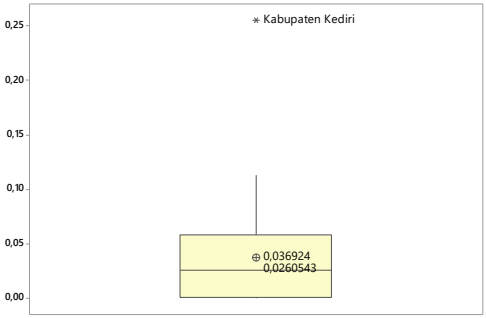
Gambar 4.3 menunjukkan karakteristik rasio SMA/MA. Diketahui dari boxplot bahwa variabel ini tidak memiliki data *outlier*. Rasio SMA/MA tertinggi terdapat pada Kabupaten

Sumenep. Sedangkan rasio ketersediaan SMA/MA terendah terdapat pada Kota Blitar.



Gambar 4.4 Karakteristik Persentase Rumah Tangga Pengguna Listrik

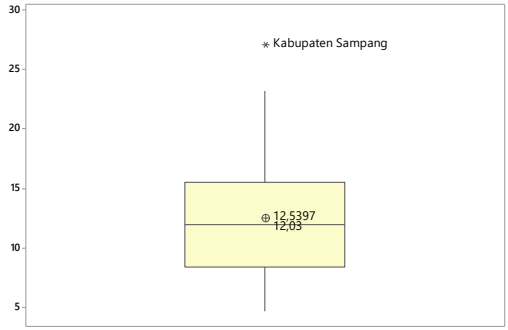
Karakteristik persentase rumah tangga pengguna listrik di Jawa Timur ditunjukkan oleh Gambar 4.4. Diketahui bahwa Kabupaten Sumenep merupakan data ekstrim bawah, hal itu dikarenakan daerah ini memiliki persentase pengguna listrik yang jauh lebih rendah dibandingkan dengan Kabupaten/Kota lain. Selain itu terdapat 4 daerah yang merupakan data *outlier* yaitu Kabupaten Bangkalan, Kabupaten Pasuruan, Kabupaten Probolinggo, dan Kabupaten Jember.



Gambar 4.5 Karakteristik Rasio Lahan Kritis

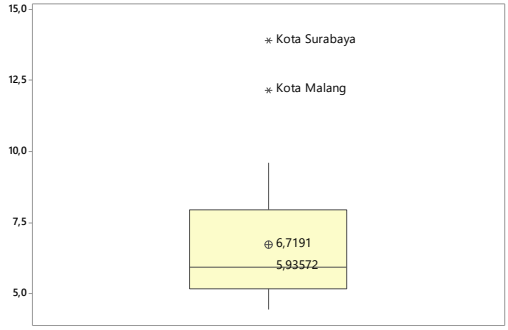
Boxplot pada Gambar 4.5 menggambarkan karakteristik rasio lahan kritis di Jawa Timur. Rasio lahan kritis pada Kabupaten Kediri memiliki nilai yang jauh lebih tinggi sehingga menyebabkan daerah tersebut sebagai data ekstrim. Tingginya rasio lahan kritis pada

Kabupaten Kediri menunjukkan bahwa banyak lahan produktif yang mengalami kerusakan dan berkurang fungsinya di daerah tersebut. Nilai rasio lahan kritis terendah dimiliki oleh 9 Kabupaten/Kota, salah satunya yaitu Kota Malang.



Gambar 4.6 Karakteristik Persentase Penduduk Miskin

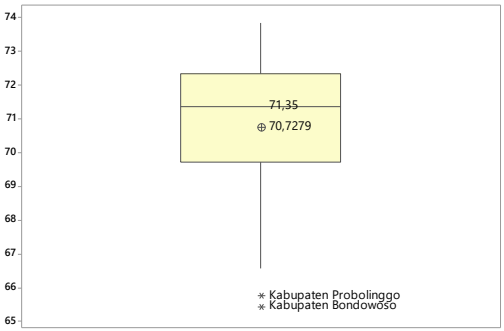
Gambar 4.6 menunjukkan bahwa Kabupaten Sampang merupakan data *outlier*. Karena daerah tersebut memiliki persentase penduduk miskin yang lebih tinggi dibandingkan dengan Kabupaten/Kota di Jawa Timur yang lain. Sedangkan rasio terendah persentase penduduk miskin dimiliki oleh Kota Batu.



Gambar 4.7 Karakteristik Pengeluaran Penduduk per Kapita

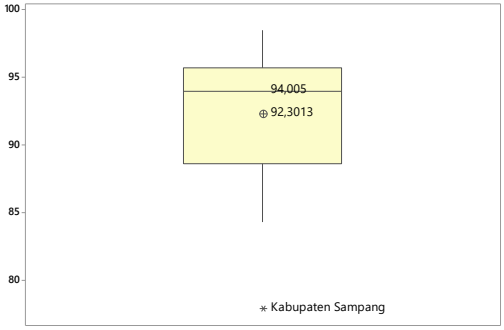
Diketahui dari Gambar 4.7 Kota Surabaya dan Kota Malang memiliki pengeluaran penduduk per kapita yang lebih tinggi dibandingkan dengan daerah lain. Pengeluaran penduduk per kapita paling rendah terdapat pada Kabupaten Lumajang. Rata-rata

pengeluaran penduduk per kapita di Provinsi Jawa Timur adalah sebesar 6.719 dalam ratusan ribu.



Gambar 4.8 Karakteristik Angka Harapan Hidup

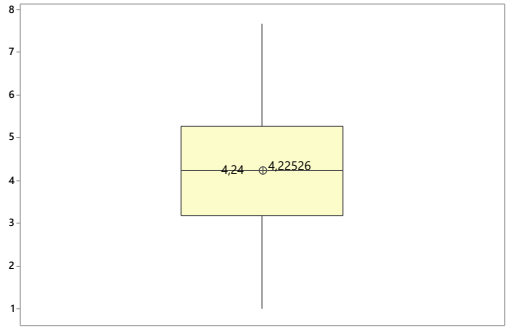
Karakteristik angka harapan hidup di Provinsi Jawa Timur ditunjukkan pada Gambar 4.8. Berdasarkan gambar tersebut diketahui bahwa Kabupaten Bondowoso merupakan daerah dengan angka harapan hidup yang jauh lebih rendah dibandingkan daerah lainnya. Kota Surabaya memiliki angka harapan hidup tertinggi yaitu sebesar 73.85. Rata-rata angka harapan hidup di Provinsi Jawa Timur adalah sebesar 70.7279.



Gambar 4.9 Karakteristik Angka Melek Huruf

Ditunjukkan pada Gambar 4.9 bahwa data *outlier* dimiliki oleh Kabupaten Sampang. Artinya Kabupaten Sampang memiliki nilai angka melek huruf yang jauh lebih rendah. Angka melek huruf tertinggi terdapat pada Kota Pasuruan dengan nilai sebesar 98.52.

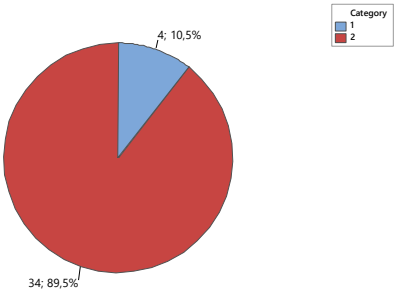
Rata-rata angka melek huruf pada Provinsi Jawa Timur adalah sebesar 92.3013.



Gambar 4.10 Karakteristik Persentase Pengangguran

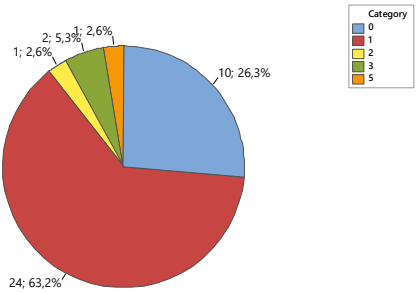
Boxplot pada Gambar 4.10 menggambarkan karaktersitik persentase pengangguran di Provinsi Jawa Timur. Diketahui dari boxplot tersebut bahwa tidak terdapat data *outlier* maupun data ekstrim. Persentase pengangguran tertinggi terdapat pada Kota Kediri yaitu sebesar 7.66%, sedangkan Kabupaten Sumenep merupakan daerah dengan persentase pengangguran terendah. Rata-rata persentase pengangguran di Provinsi Jawa Timur adalah sebesar 4.2253%.

Variabel kategorik yang digunakan pada penelitian ini yaitu meliputi keberadaan kawasan industri, tingkat desentralisasi fiskal, dan sumber air minum masyarakat. Karakteristik Kabupaten/Kota di Jawa Timur berdasarkan variabel kategorik akan ditampilkan pada diagram lingkaran pada Gambar 4.11 hingga Gambar 4.13 berikut ini.



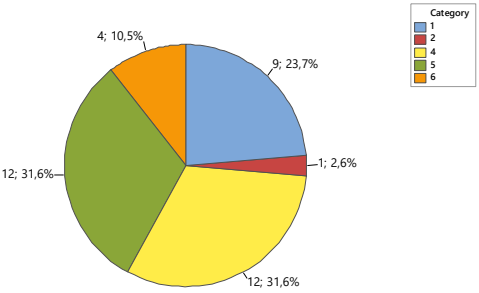
Gambar 4.11 Karakteristik Keberadaan Kawasan Industri

Keberadaaan kawasan industri menggambarkan kondisi ada atau tidaknya sektor industri besar seperti pabrik di suatu daerah. Dari Gambar 4.11 diketahui bahwa hanya terdapat 4 daerah atau sebanyak 10.5% daerah di Jawa Timur yang memiliki kawasan industri. Sedangkan mayoritas kabupaten/kota di Jawa Timur tidak memiliki kawasan industri di daerahnya.



Gambar 4.12 Karakteristik Tingkat Desentralisasi Fiskal

Tingkat desentralisasi fiskal menggambarkan bagaimana tingkat keamampuan keuangan suatu daerah. Gambar 4.12 menunjukkan bahwa sebagian besar (63.2%) daerah di Jawa Timur memiliki tingkat desentralisasi fiskal pada kondisi kurang. Daerah yang memiliki tingkat desentralisasi fiskal pada kondisi sangat kurang juga cukup banyak yaitu sebesar 26.3%. Sementara hanya terdapat masing-masing satu daerah saja yang memiliki tingkat desentralisasi fiskal pada kondisi cukup dan sangat baik.



Gambar 4.13 Karakteristik Sumber Air Minum

Gambar 4.13 menunjukkan karakteristik dari sumber air minum yang dikonsumsi mayoritas penduduk di suatu kabupaten/kota. Terdapat 9 daerah di Jawa Timur yang menggunakan air minum dengan sumber air kemasan. Hanya ada satu daerah yang mengkonsumsi air minum bersumber dari ledeng meteran. Sementara masing-masing 12 daerah mengkonsumsi air minum dengan yang bersumber dari air pompa dan air sumur. Sedangkan 4 kabupaten/kota sisanya menggunakan sumber air minum berupa mata air.

4.2 Pengelompokan Kabupaten/Kota di Jawa Timur dengan *Ensemble K-Modes*

Sebelum dilakukan pengelompokan, terlebih dahulu akan dipilih beberapa variabel yang dapat mewakili pengelompokan kabupaten/kota di Jawa Timur menggunakan analisis faktor. Dalam analisis faktor perlu dilakukan beberapa pengujian asumsi terlebih dahulu yaitu pemeriksaan asumsi kecukupan korelasi, asumsi normal multivariat, dan asumsi dependensi variabel.

4.2.1 Pereduksian Variabel Numerik dengan Analisis Faktor

Sebelum melakukan analisis faktor perlu dilakukan pemeriksaan serta pengujian asumsinya. Terdapat tiga asumsi yang perlu dipenuhi yaitu asumsi kecukupan korelasi variabel, asumsi normal multivariat, dan asumsi dependensi variabel.

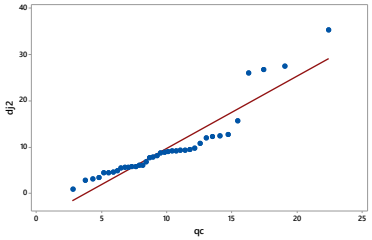
Pemeriksaan asumsi kecukupan korelasi variabel dilakukan untuk mengetahui apakah variabel cukup berkorelasi untuk dilakukan analisis faktor atau tidak. Kecukupan korelasi variabel dapat dilihat dari nilai KMO, melalui perhitungan menggunakan bantuan diperoleh nilai KMO sebesar 0.705. Karena nilai KMO yang diperoleh lebih dari 0.5 maka variabel dikatakan cukup dan mampu dilakukan analisis faktor.

Selanjutnya dilakukan pengujian asumsi dependensi variabel dengan melihat nilai *Bartlett Sphrecity*. Namun sebelum melakukan pengujian asumsi tersebut, perlu dilakukan pemeriksaan normal multivariat pada data. Pemeriksaan dilakukan untuk mengetahui apakah data telah berdistribusi normal secara multivariat atau tidak, digunakan uji *Shapiro Wilk*.

Tabel 4.2 Uji Normal Multivariat

W	0.26891
P-Value	1.537×10^{-12}
Keputusan	Data Tidak Normal Multivariat

Informasi yang diperoleh dari Tabel 4.2 yaitu data yang digunakan dalam penelitian tidak normal secara multivariat. Hal itu ditunjukkan oleh nilai W yang tidak mendekati 1 dan *p-value* yang kurang dari α (0.05). Diduga data tidak normal disebabkan adanya data *outlier*. Berikut merupakan grafis *scatter plot* dari data penelitian.



Gambar 4.14 *Scatter Plot* Data Penelitian

Dapat dilihat dari Gambar 4.14 bahwa persebaran data penelitian tidak mengikuti garis merah, dan diketahui terdapat empat titik yang *outlier* yaitu berada jauh dari titik yang lain. Sehingga dapat dikatakan bahwa data tidak normal disebabkan adanya data ekstrim maupun *outlier* pada data penelitian. Dalam penelitian ini ketidaknormalan akibat *outlier* pada data tidak diatasi, karena pengelompokan akan dilakukan menggunakan metode *Fuzzy C-Means* yang *robust* terhadap *outlier*. Namun dalam penelitian ini data diasumsikan berdistribusi normal multivariat.

Pengujian dependensi variabel dilakukan untuk mengetahui apakah variabel dependen (saling berkorelasi) atau tidak, dilakukan dengan menggunakan uji *Bartlett Sphericity*. Hasil pemeriksaan asumsi ini ditampilkan pada Tabel 4.3 berikut.

Tabel 4.3 Uji *Bartlett Sphericity*

Chi-Square	173.302
P-Value	0.000
Keputusan	Variabel Dependen (Saling Berkorelasi)

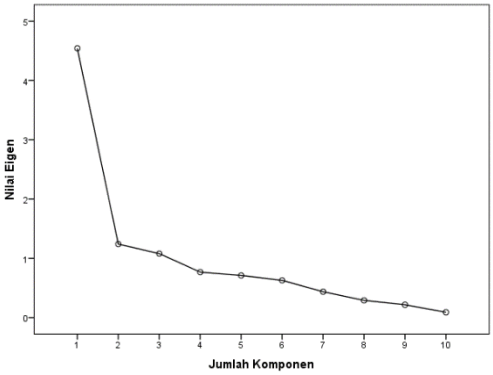
Berdasarkan Tabel 4.3 diketahui bahwa variabel yang digunakan pada penelitian saling berkorelasi (dependen), karena memiliki nilai *p-value* yang lebih kecil dari α (0.05). Sehingga dapat dilanjutkan ke analisis faktor.

Analisis faktor digunakan untuk membentuk faktor-faktor yang mampu mewakili seluruh variabel yang memiliki kemiripan karakteristik pada masing-masing faktor terbentuk. Akan ditentukan berapa komponen terbentuk dari sepuluh variabel numerik yang ada. Pada penelitian ini, variabel terpilih ditentukan dari *loading factor* tertinggi pada masing-masing faktor.

Tabel 4.4 Nilai Eigen

Komponen	Initial Eigenvalues		
	Total	**Varians (%)**	**Kumulatif (%)**
1	**4.542**	45.42	45.42
2	**1.24**	12.397	57.817
3	**1.079**	10.794	**68.611**
4	0.767	7.673	76.284
5	0.711	7.107	83.392
6	0.628	6.28	89.671
7	0.436	4.364	94.035
8	0.291	2.908	96.943
9	0.216	2.165	99.108
10	0.089	0.892	100

Tabel 4.4 menunjukkan nilai eigen yang dihasilkan bervariasi dan nilai eigen yang lebih dari nilai 1 ada sebanyak 3 komponen. Ketiga komponen tersebut memberikan kumulasi keragaman sebesar 68.611%. Sesuai dengan teori Hair (2010) yaitu persentase kumulatif nilai eigen yang ditoleransi sebesar 60%, sehingga dengan terbentuknya 3 komponen tersebut sudah dikatakan mampu mewakili sepuluh variabel yang ada. Karena nilai eigen yang bervariasi tersebut, maka dugaan data tidak berdistribusi normal multivariat tidak terbukti. Sehingga uji *Bartlett* valid untuk digunakan memutuskan bahwa variabel saling berkorelasi. Selain menggunakan nilai eigen, secara visual jumlah komponen terbentuk dapat dilihat dari grafik *scree plot* pada Gambar 4.15.



Gambar 4.15 *Scree Plot*

Dari Gambar 4.15 dapat dilihat bahwa komponen 1, 2, dan 3 memiliki nilai *eigen* lebih dari 1. Sedangkan komponen 4 sampai dengan komponen 10 memiliki nilai *eigen* kurang dari 1. Setelah diperoleh jumlah komponen terbentuk, selanjutnya dilihat matriks komponen untuk mengetahui anggota masing-masing komponen terbentuk.

Tabel 4.5 Matriks Komponen

Variabel	Komponen		
	1	**2**	**3**
X_1	**0.650**	0.104	0.527
X_2	**0.679**	-0.095	-0.163
X_3	-0.149	**-0.857***	-0.071
X_4	0.090	**0.764**	-0.065
X_5	-0.248	-0.047	**0.805***
X_6	**-0.755**	-0.436	0.054
X_7	**0.852***	0.106	-0.284
X_8	**0.725**	0.354	-0.031
X_9	**0.780**	0.414	-0.122
X_{10}	0.437	**0.473**	-0.376

Tabel 4.5 menunjukkan matriks komponen yang berisikan anggota-anggota tiap komponen yang ditentukan dari nilai komponen terbesar. Komponen 1 terdiri dari jalan aspal (X_1), rumah sakit (X_2), persentase penduduk miskin (X_6), pengeluaran penduduk per kapita (X_7), angka harapan hidup (X_8), dan angka melek huruf (X_9). Komponen 2 terdiri dari SMA/MA (X_3), rumah tangga

pengguna listrik (X_4), dan persentase pengangguran (X_{10}). Sedangkan komponen ketiga beranggotakan lahan kritis (X_5). Pengelompokan variabel ke dalam faktor dilihat pada besarnya korelasi terhadap faktor, matrik korelasi dapat dilihat pada Lampiran 7. Misalnya, variabel X_1 memiliki korelasi yang kuat dengan variabel-variabel lainnya yang juga masuk dalam faktor 1. Tiap komponen memiliki jumlah anggota (variabel) yang berbeda, sehingga hal tersebut memenuhi teori Hair yang menyatakan bahwa kondisi sepeprti ini layak untuk dilakukan analisis faktor.

Dari masing-masing komponen tersebut dipilih satu variabel yang memiliki nilai matriks komponen tertinggi untuk selanjutnya digunakan dalam analisis pengelompokan variabel numerik. Variabel terpilih tersebut adalah pengeluaran penduduk per kapita (X_7), SMA/MA (X_3), dan lahan kritis (X_5).

4.2.2 Pengelompokan Variabel Numerik dengan *Fuzzy C-Means*

Pengelompokan data numerik pada penelitian ini dilakukan menggunakan metode *fuzzy c-means.* FCM dilakukan sebanyak dua kali. Analisis pertama menggunakan sepuluh variabel keseluruhan, sedangkan analisis kedua digunakan tiga variabel yang merupakan variabel terpilih hasil dari analisis faktor. Hal ini dilakukan untuk mengetahui apakah hasil pengelompokan lebih baik menggunakan variabel keseluruhan atau menggunakan variabel terpilih.

Tabel 4.6 *Pseudo Fstatistics*

	Jumlah Klaster	**Fungsi Keanggotaan**			
		Linier Naik	**Linier Turun**	**Kurva Segitiga**	**Kurva Trapesium**
Tanpa Analisis Faktor	**2**	**34.49104760**	**34.49104770***	**34.49104585**	**34.49104584**
	3	26.56911379	26.56910961	26.56910946	26.56911381
	4	29.52130580	29.52130575	29.52130616	29.52130577
	5	22.01197152	22.01197129	22.01197162	22.01194871
Analisis Faktor	2	85.07968288	85.07967774	85.07973289	85.07967852
	3	151.86835628	151.86835888	151.86835971	151.86836054
	4	185.74811745	185.74812571	185.74812570	185.74812679
	5	**249.55825373**	**249.5583178***	**249.55825359**	**249.55825170**

Setelah dilakukan pengelompokan diperoleh nilai *Pseudo Fstatistics* seperti yang ditampilkan pada Tabel 4.6. Nilai tersebut digunakan untuk mengetahui jumlah klaster yang paling optimum, semakin tinggi nilai *Pseudo Fstatistics* maka semakin optimum klaster tersebut. Pada metode FCM tanpa analisis faktor diperoleh nilai *Pseudo Fstatistics* tertinggi berada pada 2 klaster. Sedangkan pada FCM dengan analisis faktor diperoleh nilai *Pseudo Fstatistics* tertinggi pada jumlah klaster 5. Jumlah klaster optimum keduanya terdapat pada fungsi keanggotaan linier turun.

Tabel 4.7 *ICD Rate Fuzzy C-Means* Fungsi Keanggotaan Linier Turun

	Jumlah Klaster	***ICD Rate***
Tanpa Analisis Faktor	2	0.510703148
Analisis Faktor	**5**	**0.032000519**

Tabel 4.7 memberikan informasi mengenai nilai *icd rate* yang digunakan untuk menentukan metode terbaik dalam analisis kelompok. Semakin kecil nilai *icd rate* maka suatu metode dikatakan baik karena semakin kecil perbedaan di dalam kelompok. Diperoleh hasil untuk metode FCM dengan menggunakan analisis faktor memiliki nilai *icd rate* yang lebih kecil daripada tanpa analisis faktor. Sehingga dalam pengelompokan dengan metode FCM akan digunakan analisis faktor untuk mereduksi variabel, digunakan fungsi keanggotaan linier turun, dan dipilih jumlah klaster sebanyak lima. Hasil dari pengelompokan tersebut adalah kelompok 1 memiliki 4 daerah, kelompok 2 dengan 17 daerah, kelompok 3 terdiri dari 6 daerah, kelompok 4 memiliki 2 daerah, dan daerah sisanya berada pada kelompok 5.

Tabel 4.8 Anggota Kelompok 5 Klaster *Fuzzy C-Means*

Klaster	**Anggota**
1	Kab. Sidoarjo, Kota Blitar, Kota Probolinggo, dan Kota Madiun
2	Kab. Pacitan, Kab. Ponorogo, Kab. Trenggalek, Kab. Kediri, Kab. Lumajang, Kab. Jember, Kab. Bondowoso, Kab. Situbondo, Kab. Probolinggo, Kab. Pasuruan, Kab. Nganjuk, Kab. Bojonegoro, Kab. Tuban, Kab. Bangkalan, Kab. Sampang, Kab. Pamekasan, dan Kab. Sumenep
3	Kab. Mojokerto, Kab. Gresik, Kota Kediri, Kota Pasuruan, Kota Mojokerto, dan Kota Batu

Tabel 4.8 Anggota Kelompok 5 Klaster *Fuzzy C-Means* (Lanjutan)

Klaster	Anggota
4	Kota Malang dan Kota Surabaya
5	Kab. Tulungagung, Kab. Blitar, Kab. Malang, Kab. Banyuwangi, Kab. Jombang, Kab. Madiun, Kab. Magetan, Kab. Ngawi, dan Kab. Lamongan

4.2.3 Pengelompokan Variabel Kategorik dengan *K-Modes*

Metode *K-Modes* akan digunakan untuk mengelompokkan variabel dengan skala data kategorik. Metode ini menempatkan sebuah objek pengamatan tepat menjadi anggota suatu kelompok. Ukuran kemiripan pada metode ini diukur menggunakan nilai modus. Dalam penelitian ini jumlah kelompok yang digunakan yaitu 2, 3, 4, dan 5. Untuk menentukan jumlah kelompok yang optimum, digunakan nilai akurasi yang diperoleh dari hasil masing-masing jumlah kelompok. Tabel 4.9 menunjukkan perbandingan nilai akurasi pengelompokan.

Tabel 4.9 Nilai Akurasi *K-Modes*

Jumlah Klaster	Nilai Akurasi
2	0.66667
3	0.73684
4	0.84211
5	**0.85965**

Hasil perbandingan nilai akurasi pada Tabel 4.9 diperoleh hasil bahwa pengelompokan *k-modes* dengan jumlah klaster sebanyak 5 memiliki nilai akurasi yang lebih besar dibandingkan dengan jumlah klaster lainnya. Pengelompokan tersebut membagi kabupaten/kota di Jawa Timur menjadi 5 kelompok dengan masing-masing anggota sebanyak 3 daerah, 19 daerah, 8 daerah, 7 daerah, dan 1 daerah.

Tabel 4.10 Anggota Kelompok 5 Klaster *K-Modes*

Klaster	Anggota
1	Kab. Sidoarjo, Kab. Gresik, dan Kota Surabaya
2	Kab. Ponorogo, Kab. Kediri, Kab. Malang, Kab. Situbondo, Kab. Probolinggo, Kab. Pasuruan, Kab. Mojokerto, Kab. Jombang, Kab. Nganjuk, Kab. Madiun, Kab. Bojonegoro,

Tabel 4.10 Anggota Kelompok 5 Klaster *K-Modes* (Lanjutan)

Klaster	Anggota
2	Kab. Tuban, Kab. Lamongan, Kota Kediri, Kota Probolinggo, Kota Pasuruan, Kota Mojokerto, Kota Madiun, dan Kota Batu
3	Kab. Pacitan, Kab. Trenggalek, Kab. Blitar, Kab.Bondowoso, Kab. Magetan, Kab. Bangkalan, Kab. Sampang, dan Kab. Sumenep
4	Kab. Tulungagung, Kab. Lumajang, Kab. Jember, Kab. Banyuwangi, Kab. Ngawi, Kab. Pamekasan, dan Kota Blitar
5	Kota Malang

4.2.4 Pengelompokan Variabel Campuran dengan *Ensemble K-Modes*

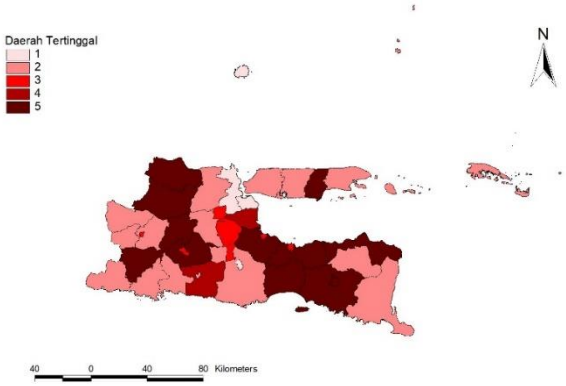
Sebelumnya telah diperoleh hasil pengelompokan pada variabel numerik dengan jumlah kelompok optimum sebanyak 5 kelompok. Hasil dari pengelompokan optimum variabel kategorik sebanyak 5 kelompok. Setelah dilakukan pengelompokan pada masing-masing variabel numerik dan variabel kategorik maka selanjutnya dilakukan pengelompokan gabungan (*ensemble*) untuk memperoleh hasil kelompok akhir dari sebuah pengelompokan data campuran.

Pada pengelompokan gabungan, data yang digunakan merupakan hasil dari pengelompokan pada masing-masing skala variabel yang nantinnya menjadi variabel baru. Hasil dari pengelompokan numerik dituliskan sebagai variabel baru X_1, sedangkan hasil pengelompokan kategorik dianggap sebagai variabel baru X_2. Penelitian ini menggunakan penerapan metode pengelompokan gabungan dengan prinsip pengelompokan data kategorik, sehingga untuk mendapatkan hasil pengelompokan akhir digunakan metode *k-modes*. Berikut ini merupakan nilai akurasi dari pengelompokan metode *ensemble k-modes*.

Tabel 4.11 Nilai Akurasi *Ensemble K-Modes*

Jumlah Klaster	Nilai Akurasi
2	0.59211
3	0.64474
4	0.67105
5	**0.72368**

Berdasarkan nilai akurasi pada Tabel 4.11 diketahui bahwa nilai akurasi tertinggi terdapat pada jumlah klaster sebanyak 5. Maka pada penelitian ini dipilih jumlah klaster optimum sebesar 5 dengan nilai akurasi sebesar 0.72368 atau sebesar 72.368%.



Gambar 4.16 Pemetaan Hasil Pengelompokan *Ensemble K-Modes*

Secara visual hasil pengelompokan dari metode *ensemble k-modes* ditampilkan pada peta Jawa Timur Gambar 4.16. Sedangkan rincian anggota dari hasil pengelompokan dengan jumlah klaster 5 akan ditampilkan pada Tabel 4.12.

Tabel 4.12 Anggota Kelompok 5 Klaster *Ensemble K-Modes*

Klaster	Anggota
1	Kab. Gresik, Kota Malang, dan Kota Surabaya
2	Kab. Pacitan, Kab. Trenggalek, Kab. Tulungagung, Kab. Blitar, Kab. Malang, Kab. Banyuwangi, Kab. Bondowoso, Kab. Jombang, Kab. Madiun, Kab. Magetan, Kab. Ngawi, Kab. Lamongan, Kab. Bangkalan, Kab. Sampang, dan Kab. Sumenep
3	Kab. Mojokerto, Kota Kediri, Kota Probolinggo, Kota Pasuruan, Kota Mojokerto, Kota Madiun, dan Kota Batu
4	Kab. Sidoarjo dan Kota Blitar
5	Kab. Ponorogo, Kab. Kediri, Kab. Lumajang, Kab. Jember, Kab. Situbondo, Kab. Probolinggo, Kab. Pasuruan, Kab. Nganjuk, Kab. Bojonegoro, Kab. Tuban, dan Kab. Pamekasan

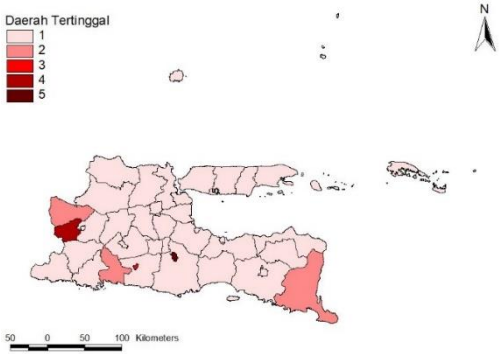
4.3 Pengelompokan Kabupaten/Kota di Jawa Timur dengan *Similarity Weighted and Filter Method* (SWFM)

Selain menggunakan *ensemble cluster*, pengelompokan data campuran juga dapat dilakukan menggunakan metode SWFM. Tahap pertama yaitu mengelompokkan masing-masing skala data dengan masing-masing metodenya sesuai dengan subbab 4.2.2 untuk data numerik dan 4.2.3 untuk data kategorik. Hasil dari masing-masing pengelompokan tersebut dianggap sebagai variabel baru dengan skala data kategorik dan kemudian variabel tersebut dianalisis menggunakan metode SWFM.

Tabel 4.13 Nilai Akurasi SWFM

Jumlah Klaster	Akurasi
2	0.526316
3	0.605263
4	0.605263
5	**0.631579**

Tabel 4.13 menjelaskan mengenai nilai akurasi yang dihasilkan oleh masing-masing jumlah klaster. Berdasarkan tabel tersebut diketahui bahwa jumlah klaster 3, dan 4 memiliki nilai akurasi yang sama. Jumlah klaster yang terpilih yaitu 5 klaster dengan nilai akurasi yaitu sebesar 0.631579 atau 63.15%.



Gambar 4.17 Pemetaan Hasil Pengelompokan SWFM

Secara visual hasil pengelompokan dari metode SWFM ditampilkan pada peta Jawa Timur Gambar 4.17. Rincian hasil pengelompokan ditampilkan pada Tabel 4.14.

Tabel 4.14 Anggota Kelompok 5 Klaster SWFM

Klaster	Anggota
1	Kab. Pacitan, Kab. Ponorogo, Kab. Trenggalek, Kab. Kediri, Kab. Malang, Kab. Lumajang, Kab. Jember, Kab. Bondowoso, Kab. Situbondo, Kab. Probolinggo, Kab. Pasuruan, Kab. Sidoarjo, Kab. Mojokerto, Kab. Jombang, Kab. Nganjuk, Kab. Madiun, Kab. Bojonegoro, Kab. Tuban, Kab. Lamongan, Kab. Gresik, Kab. Bangkalan, Kab. Sampang, Kab. Pamekasan, Kab. Sumenep, Kota Kediri, Kota Blitar, Kota Probolinggo, Kota Pasuruan, Kota Mojokerto, Kota Madiun, Kota Surabaya, dan Kota Batu
2	Kab. Tulungagung, Kab. Banyuwangi, dan Kab. Ngawi
3	Kab. Blitar
4	Kab. Magetan
5	Kota Malang

4.4 Hasil Pengelompokan Data Campuran Terbaik Beserta Karakteristiknya

Hasil pengelompokan terbaik diperoleh dengan membandingkan hasil pengelompokan optimum data campuran menggunakan metode *ensemble k-modes* dan metode SWFM. Perbandingan dilakukan dengan melihat nilai akurasi hasil pengelompokan, semakin tinggi nilai akurasi maka kinerja pengelompokan akan semakin baik. Berikut merupakan nilai perbandingan untuk kedua metode pengelompokan data campuran.

Tabel 4.15 Perbandingan Nilai Akurasi Metode Pengelompokan Data Campuran

Metode	Jumlah Klaster	Akurasi
Ensemble K-Modes	**5**	**0.72368**
SWFM	5	0.631579

Berdasarkan Tabel 4.15 metode pengelompokan data campuran *ensemble k-modes* dengan jumlah klaster optimum sebanyak 5 memiliki nilai akurasi terbesar yaitu sebesar 0.72368. Sehingga dapat diketahui bahwa metode pengelompokan data campuran terbaik untuk mengelompokkan Kabupaten/Kota di Jawa Timur

berdasarkan indikator daerah tertinggal adalah dengan menggunakan metode *ensemble k-modes*.

Setelah diperoleh metode dengan hasil pengelompokan optimum, selanjutnya dilakukan pengujian beda rata-rata antar kelompok untuk mengetahui kelompok yang dihasilkan berbeda secara signifikan atau tidak. Pengujian beda rata-rata secara multivariat dilakukan dengan menggunakan metode MANOVA.

MANOVA hanya digunakan pada variabel numerik saja. Sebelum dilakukan analisis dengan MANOVA, dilakukan terlebih dahulu pemeriksaan asumsi normal multivariat dari data menggunakan uji *Shapiro Wilks.* Nilai yang dihasilkan oleh *mshapiro test* adalah *p-value* sebesar $1.537x10^{-12}$. Karena nilai tersebut lebih kecil dari α (0.05), maka asumsi data berdistribusi normal multivariat tidak terpenuhi. Hal itu disebabkan oleh adanya data outlier.

Dilakukan pula pemeriksaan homogenitas varian kovarian. Uji homogenitas matriks varian kovarian dilakukan untuk mendeteksi dua atau lebih kelompok data sampel dari populasi memiliki matriks varian kovarian yang homogen. Uji homogenitas dilakukan dengan *Box's M Test* yang hasilnya akan ditampilkan pada Tabel 4.16.

Tabel 4.16 Uji Homogenitas Box's M Test

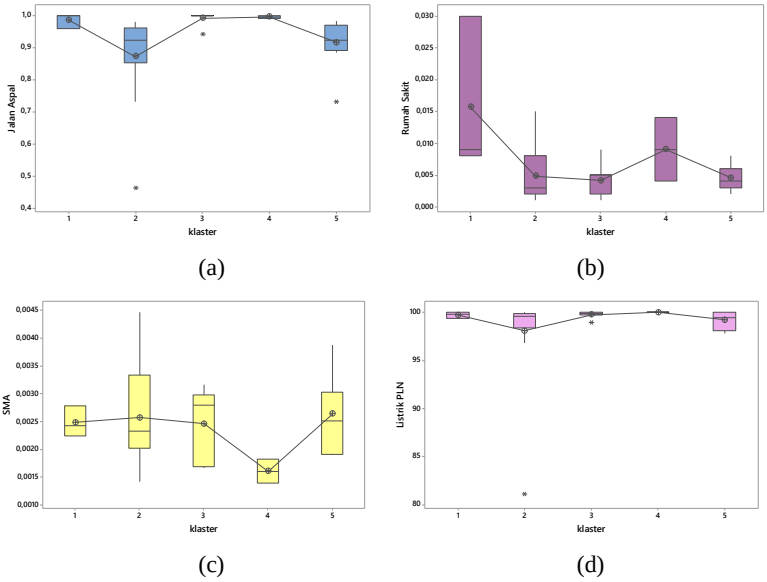
Statistik Uji	Nilai
Box's M	210.342
F	1.963
df1	55
df2	1504.925
Sig.	0.000

Tabel 4.16 memberikan informasi mengenai hasil uji homogenitas, dengan taraf signifikansi (α) sebesar 0.05 maka dapat diambil keputusan Tolak H_0 karena nilai signifikansi sebesar 0.000 lebih kecil dari α. Selain itu dapat dilihat dari nilai F hitung yang dihasilkan lebih besar dari F tabel sebesar 1.3419. Artinya matriks varian kovarian yang dihasilkan dari data tidak bersifat homogen.

Karena asumsi normal multivariat dan homogenitas tidak terpenuhi, maka pengujian beda rata-rata antar kelompok dilakukan dengan MANOVA mengggunakan statistik uji *Pillai's Trace* (Johnson & Wichern, 2007). Diperoleh nilai signifikansi statistik

Pillai's Trace sebesar 1.642 dan *p-value* sebesar 0.005. Artinya bahwa dengan taraf signifikansi 0.05, maka akan diperoleh keputusan tolak H_0 sehingga diketahui minimal terdapat satu nilai rata-rata yang berbeda.

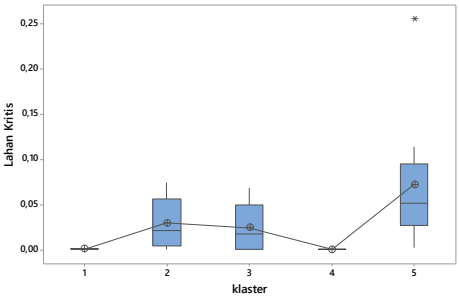
Guna mengetahui karakteristik pada masing-masing variabel terhadap klaster yang terbentuk, secara deskriptif akan ditampilkan pada Gambar 4.18 sampai dengan Gambar 4.21 berikut ini.



Gambar 4.18 Perbandingan Karakteristik Kriteria Infrastruktur
(a) Jalan Aspal, (b) Rumah Sakit, (c) SMA, (d) Listrik

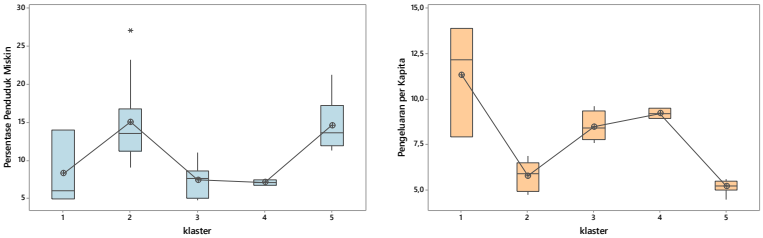
Perbandingan karakteristik kelompok berdasarkan kriteria infrastruktur ditampilkan pada Gambar 4.18. Diketahui bahwa pada variabel jalan aspal, klaster 2 merupakan kelompok dengan rata-rata terendah yaitu sebesar 0.8722. Begitu pula pada variabel pengguna listrik PLN klaster 4 memiliki rata-rata terendah dengan nilai yaitu 98.07. Pada variabel rumah sakit, rata-rata terendah terdapat pada klaster 3 sebesar 0.0041. Sedangkan klaster 4 memiliki rata-rata terendah pada variabel rasio SMA yaitu sebesar 0.0016.

Kriteria karakteristik daerah ditunjukkan oleh variabel rasio luas lahan kritis. Perbandingan karakteristik kelompok berdasarkan variabel tersebut akan ditampilkan pada Gambar 4.19 berikut ini.



Gambar 4.19 Perbandingan Karakteristik Rasio Lahan Kritis

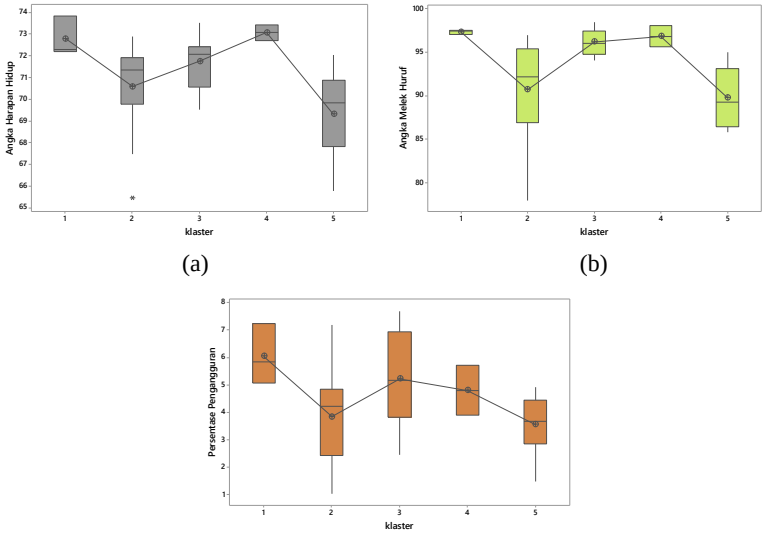
Berdasarkan hasil pengelompokan, diketahui dari Gambar 4.19 bahwa kelompok 4 memiliki rata-rata rasio lahan kritis terendah yaitu sebesar 0. Artinya kabupaten/kota yang terkelompok pada kelompok 4 tidak ada yang memiliki lahan kritis. Sedangkan rata-rata tertinggi terdapat pada kelompok 5 yaitu sebesar 0.072, namun memiliki nilai *outlier*.



(a) (b)

Gambar 4.20 Perbandingan Karakteristik Kriteria Ekonomi
(a) Penduduk Miskin dan (b) Pengeluaran per Kapita

Dari segi ekonomi, kriteria dapat dilihat melalui variabel persentase penduduk miskin dan pengeluaran per kapita. Kelompok 4 memiliki rata-rata penduduk miskin paling sedikit yaitu sebesar 7.07%. Kelompok yang memiliki penduduk miskin terbanyak adalah kelompok 2. Sedangkan pengeluaran per kapita tertinggi terdapat pada kelompok 1 yaitu sebesar 11.3266. Kelompok 5 memiliki rata-rata pengeluaran per kapita terendah.



Gambar 4.21 Perbandingan Karakteristik Kriteria Sumber Daya Manusia (a) Angka Harapan Hidup, (b) Angka Melek Huruf, (c) Persentase Pengangguran

Dilihat dari Gambar 4.21, variabel angka harapan hidup memiliki nilai rata-rata sebesar 69.29 yaitu nilai terendah pada klaster 5. Begitu pula variabel angka melek huruf bernilai rata-rata terendah 88.77 pada klaster 5. Variabel persentase pengangguran rata-rata tertinggi terdapat pada klaster 1 sebesar 6.033%.

ANOVA *one-way* dilakukan untuk mengetahui variabel apa yang berbeda diantara kelima kelompok yang terbentuk. Tabel 4.17 akan menunjukkan hasil dari ANOVA *one-way* pada kesepuluh variabel numeriknya.

Tabel 4.17 Hasil Pengujian ANOVA *One-way*

Variabel	*P-Value*
Rasio Desa dengan Permukaan Jalan Terluas Aspal	0.058*
Rasio Rumah Sakit terhadap Jumlah Penduduk	0.006
Rasio SMA/MA terhadap Jumlah Penduduk Usia SMA Sederajat	0.453*
Persentase Rumah Tangga Pengguna Listrik	0.718*
Luas Lahan Kritis	0.031
Persentase Penduduk Miskin	0.001

Tabel 4.17 Hasil Pengujian ANOVA *One-way* (Lanjutan)

Variabel	*P-Value*
Pengeluaran Penduduk Perkapita	0.000
Angka Harapan Hidup	0.010
Angka Melek Huruf	0.002
Persentase Pengangguran	0.048

Berdasarkan Tabel 4.17 diketahui bahwa dari kesepuluh variabel numerik yang digunakan, dengan menggunakan taraf signikansi α sebesar 0.05 terdapat tujuh variabel yang signifikan berbeda antar 5 klaster terbentuk. Sedangkan 3 variabel lainnya memiliki *p-value* lebih dari α. Artinya ketiga variabel tersebut hanya memiliki sedikit perbedaan antara anggota yang ada di klaster 1 hingga klaster 5. Hal tersebut dapat ditunjukkan pada karakteristik secara deskriptifnya.

Untuk mengetahui variabel kategorik yang berbeda diantara kelompok yang terbentuk, digunakan uji Kruskal Wallis. Tabel 4.18 menunjukkan hasil pengujian Kruskal Wallis pada 3 variabel terhadap 5 kelompok yang terbentuk.

Tabel 4.18 Hasil Pengujian Kruskal Wallis

Variabel	*Chi-Square*	*P-value*
Keberadaan Kawasan Industri	11.848	0.019
Tingkat Desentralisasi Fiskal	21.412	0.000
Sumber Air Minum	12.359	0.015

Tabel 4.18 menunjukkan hasil pengujian perbedaan pengelompokan yang dihasilkan terhadap variabel kategorik. Diketahui dari hasil tersebut bahwa seluruh variabel kategorik yang digunakan dalam analisis memiliki nilai χ^2_{hitung} lebih besar dari nilai $\chi^2_{(0.05,5-1)}$ yaitu sebesar 9.488. Sehingga variabel keberadaan kawasan industri, tingkat desentralisasi fiskal, dan sumber air minum signifikan berbeda antara kelima kelompok yang terbentuk.

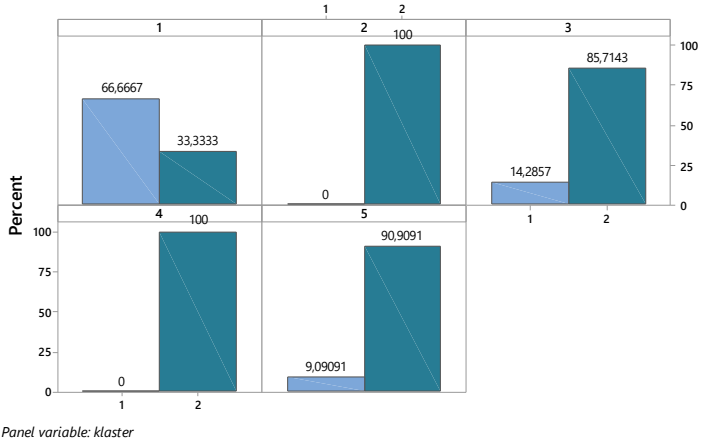
Untuk membandingkan karakteristik pengelompokan pada tiap klaster berdasarkan variabel numerik dapat dilakukan dengan melihat nilai *mean*. Berikut ini akan ditampilkan tabel yang berisikan nilai *mean* masing-masing variabel pada tiap klaster.

Tabel 4.19 Karakteristik Hasil Pengelompokan Tiap Variabel

Variabel	Klaster				
	1	2	3	4	5
Rasio Jalan Aspal	0.9859	0.8722*	0.9915	**0.9957**	0.9157
Rasio Rumah Sakit per 1000 penduduk	**0.1566**	0.0048	0.0041*	0.0090	0.0045
Rasio SMA/MA terhadap penduduk usia SMA	0.00248	0.00256	0.00245	0.00160*	**0.00263**
Presentase Pengguna Listrik PLN	99.74	98.07*	99.77	**100**	99.23
Rasio Lahan Kritis	0.0005	0.0294	0.0241	0*	**0.7201**
Presentase Penduduk Miskin	8.27	**15.019**	7.401	7.07*	14.588
Pengeluaran per Kapita	**11.327**	5.7682	8.4649	9.2123	5.1949*
Angka Harapan Hidup	72.783	70.583	71.743	**73.065**	69.295*
Angka Melek Huruf	**97.343**	90.715	96.214	96.860	89.771*
Presentase Pengangguran	**6.0333**	3.8247	5.2143	4.7950	3.5455*

Dari Tabel 4.19 diketahui bahwa klaster 1 memiliki karakteristik berupa rasio rumah sakit tertinggi, pengeluaran per kapita tertinggi, angka melek huruf tertinggi, dan persentase pengangguran tertinggi. Klaster 2 memiliki karakteristik rasio jalan aspal terendah, pengguna listrik PLN terendah, dan persentase penduduk miskin tertinggi. Klaster 3 memiliki karakteristik rasio rumah sakit terendah. Klaster 4 memiliki karakteristik rasio jalan aspal tertinggi, rasio SMA terendah, pengguna listrik tertinggi, rasio lahan kritis terendah, persentase penduduk miskin terendah, dan angka harapan hidup tertinggi. Sedangkan pada klaster 5 memiliki karakteristik yaitu rasio SMA tertinggi, rasio lahan kritis tertinggi, pengeluaran per kapita terendah, angka harapan hidup dan angka melek huruf terendah, dan persentase pengangguran terendah.

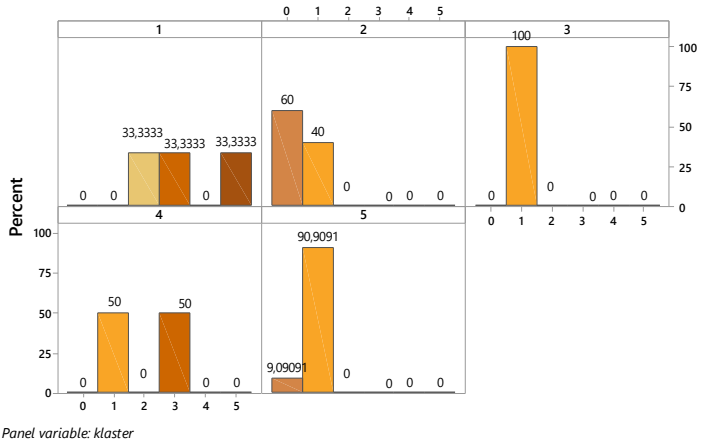
Karakteristik hasil pengelompokan kabupaten/kota di Jawa Timur pada tiap klaster berdasarkan variabel berskala kategorik dapat ditunjukkan melalui nilai persentase dari datanya. Karakteristik masing-masing klaster akan ditampilkan pada Gambar 4.22 sampai Gambar 4.24.



Gambar 4.22 Karakteristik Pengelompokan pada Variabel Keberadaan Kawasan Industri

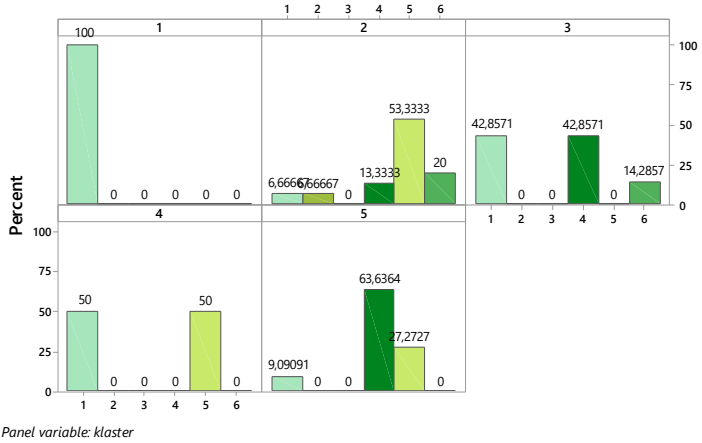
Gambar 4.22 merupakan karakteristik variabel keberadaan kawasan industri berdasarkan hasil pengelompokan. Diketahui terdapat sebanyak 66.67% daerah di kelompok 1 memiliki kawasan industri. Sedangkan pada keempat kelompok lainnya, sebagian besar daerahnya tidak memiliki kawasan industri. Bahkan dapat dilihat pada klaster 2 dan 4 seluruh daerahnya tidak memiliki kawasan industri. Pada klaster 3 terdapat 85.71% daerah yang tidak memiliki kawasan industri. Begitu halnya dengan klaster 5 yang 90.91% daerahnya tidak terdapat kawasan industri.

Pada variabel tingkat desentralisasi fiskal di Gambar 4.23, diketahui bahwa klaster 2 memiliki 60% daerah dengan kategori nol atau sangat kurang. Klaster 3 beranggotakan daerah yang seluruhnya memiliki tingkat desentralisasi fiskal pada kategori satu atau kurang, begitu pula dengan 90.91% daerah di klaster 5. Sedangkan klaster 1 dan 4 tidak memiliki nilai modus karena kategori yang dimiliki berbeda-beda.



Gambar 4.23 Karakteristik Pengelompokan pada Variabel Tingkat Desentralisasi Fiskal

Pada variabel sumber air minum berdasarkan Gambar 4.24 diperoleh 100% daerah di kelompok 1 mengonsumsi air minum yang bersumber dari air kemasan, kelompok 2 mengonsumsi air besumber air sumur, kelompok 3 dan 5 air minumnya bersumber dari air pompa, sedangkan kelompok 4 tidak memiliki modus karena memiliki kategori yang berbeda tiap anggotanya.



Gambar 4.24 Karakteristik Pengelompokan pada Variabel Sumber Air Minum

Apabila dilihat dari indikator ketertinggalan daerah, karakteristik tiap klaster yang menunjukkan status ketertinggalan akan dirangkum dan ditampilkan pada Tabel 4.20.

Tabel 4.20 Karakteristik Ketertinggalan pada Tiap Klaster

Klaster	Karakteristik Ketertinggalan	Status Ketertinggalan
1	Persentase pengangguran tertinggi	Daerah Mandiri
2	Rasio jalan aspal terendah, persentase pengguna listrik terendah, persentase penduduk miskin tertinggi, tidak memiliki kawasan industri, dan sebagian besar daerahnya memiliki tingkat desentralisasi fiskal yang sangat kurang.	Daerah Sangat Tertinggal
3	Rasio rumah sakit terendah, sebagian besar daerahnya tidak memiliki kawasan industri, dan memiliki tingkat desentralisasi fiskal pada kategori kurang.	Daerah Berkembang
4	Rasio SMA/MA terendah dan tidak memiliki kawasan industri	Daerah Maju
5	Pengeluaran per kapita terendah, Angka Harapan Hidup dan Angka Melek Huruf terendah, sebagian besar daerahnya tidak memiliki kawasan industri dan memiliki tingkat desentralisasi fiskal pada kategori kurang.	Daerah Tertinggal

Dari hasil karakteristik tersebut diketahui bahwa klaster 2 memiliki kondisi terendah. Hal itu ditunjukkan dari karakteristik pada variabel numeriknya dimana kelompok ini memiliki rata-rata jalan aspal terendah, persentase rumah tangga pengguna listrik terendah, memiliki penduduk miskin yang tinggi, serta angka harapan hidup dan angka melek huruf yang cukup rendah. Selain itu dilihat dari karakteristik variabel kategoriknya, diketahui bahwa 100% anggota klaster 2 tidak memiliki kawasan industri dan 60% anggotanya memiliki tingkat desentralisasi fiskal yang sangat

kurang. Klaster 2 sendiri terdiri dari 39.47% kabupaten/kota di Jawa Timur.

Selain klaster 2 yang memiliki status daerah sangat tertinggal, perlu diketahui pula bahwa klaster 5 memiliki status daerah tertinggal. Hal itu ditunjukkan dari karakteristik yang dimiliki yaitu memiliki pengeluaran per kapita terendah, Angka Harapan Hidup dan Angka Melek Huruf terendah, serta 91% anggota klaster 5 tidak memiliki kawasan industri dan memiliki tingkat desentralisasi fiskal yang kurang. Klaster 5 terdiri dari 28.95% kabupaten/kota di Jawa Timur.

(Halaman Ini Sengaja Dikosongkan)

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan analisis yang telah dilakukan pada Bab IV, maka dapat ditarik kesimpulan sebagai berikut.

1. Indikator ketertinggalan daerah pada Provinsi Jawa Timur ditinjau dari kriteria infrastruktur, karakteristik daerah, ekonomi, sumber daya manusia, dan kapasitas keuangan daerah memiliki karakteristik yaitu
   a. Pada kriteria infrastruktur yaitu permukaan jalan terluas, ketersediaan rumah sakit, SMA, dan listrik PLN masing-masing memiliki data *outlier*. Apabila dilihat dari sumber air minum yang digunakan, sebagian besar masyarakat mengkonsumsi air minum yang bersumber dari air pompa dan air sumur
   b. Pada kriteria karakteristik daerah variabel luas lahan kritis, terdapat data *outlier* yaitu Kota Kediri. Sedangkan apabila ditinjau dari variabel keberadaan kawasan industri, 89.5% wilayah di Jawa Timur tidak memiliki kawasan industri.
   c. Pada kriteria ekonomi diketahui bahwa Kabupaten Sampang memiliki persentase penduduk miskin yang jauh lebih tinggi dibandingkan daerah lain. Sedangkan pengeluaran per kapita tertinggi dimiliki oleh Kota Surabaya.
   d. Pada kriteria kapasitas keuangan daerah diketahui terdapat 24 kabupaten/kota di Jawa Timur yang tingkat desentralisasi fisklanya berada pada kategori kurang.
   e. Pada kriteria sumber daya manusia, variabel angka harapan hidup dan angka melek huruf memiliki data *outlier* masing-masing adalah Kabupaten Bondowoso dan Kabupaten Sampang. Variabel persentase pengangguran tidak memiliki data *outlier* dengan nilai rata-rata pengangguran di Jawa Timur sebesar 4.2253%.
2. Pengelompokan kabupaten/kota di Jawa Timur berdasarkan indikator daerah tertinggal pada variabel numerik dengan metode FCM diperoleh jumah klaster optimum sebanyak 5 dengan *pseudo-f statistics* sebesar 249.5583 dan menggunakan

fungsi keanggotaan linier turun dengan ICD *rate* sebesar 0.032 yang sebelumnya dilakukan analisis faktor terlebih dahulu. Pengelompokan data kategorik menggunakan metode *k-modes* menghasilkan jumlah klaster optimum sebanyak 5 dengan akurasi sebesar 85.97%. Sedangkan pengelompokan data campuran dengan *ensemble k-modes* menghasilkan 5 kelompok optimum bernilai akurasi 72.368% dengan masing-masing beranggotakan 3 daerah, 15 daerah, 7 daerah, 2 daerah, dan 11 daerah.

3. Hasil pengelompokan kabupaten/kota di Jawa Timur dengan menggunakan SWFM adalah terdiri dari 5 klaster dengan nilai akurasi 63.16%. Anggota dari tiap klaster yaitu masing-masing 32, 3, dan sisanya masing-masing 1 daerah.
4. Metode pengelompokan data campuran terbaik menunjukkan bahwa *ensemble k-modes* menghasilkan pengelompokan yang lebih baik dibandingkan dengan metode SWFM karena nilai akurasi pengelompokan yang dimiliki *ensemble k-modes* lebih besar yaitu 72.368%. Dengan demikian pengelompokan kabupaten/kota di Jawa Timur berdasarkan indikator daerah tertinggal menghasilkan 5 kelompok. Berdasarkan karakteristik masing-masing kelompok, diketahui bahwa kabupaten/kota di klaster 2 yang terdiri dari 39.47% kabupaten/kota di Jawa Timur menunjukkan kondisi yang paling rendah sehingga dianggap sebagai kelompok daerah sangat tertinggal. Daerah lainnya dengan status tertinggal terdapat pada klaster 5 yang terdiri dari 28.95% kabupaten/kota di Jawa Timur.

5.2 Saran

Saran yang dapat diberikan berdasarkan hasil penelitian ini adalah dengan diketahuinya kabupaten/kota di Jawa Timur yang termasuk dalam kelompok daerah tertingal maka pemerintah harus memfokuskan perhatiannya agar dapat memajukan daerah-daerah tersebut. Pemerintah perlu fokus untuk meningkatkan perekonomian serta kualitas SDM di daerah sehingga dapat mengentaskan daerah-daerah tersebut dari status ketertinggalan.

Pada variabel numerik diketahui bahwa dengan melakukan analisis faktor menghasilkan pengelompokan yang lebih baik, maka

pada penelitian selanjutnya disarankan untuk melakukan reduksi variabel yang dapat mengakomodasi variabel kategorik. Selain itu digunakan juga unit sampling berupa desa di Jawa Timur agar lebih memberikan informasi mengenai desa tertinggal, sehingga target pembangunan nasional lebih cepat tercapai. Untuk meningkatkan kebaikan hasil pengelompokan, dapat dilakukan pengelompokan data campuran dengan mencoba metode pengelompokan ensembel lainnya.

(Halaman Ini Sengaja Dikosongkan)

DAFTAR PUSTAKA

Alva, J., & Estrada, E. (2009). A Generalization of Shapiro-Wilk's Test for Multivariate-Normality. *Communication in Statistics-Theory and Methods, 38*(11), 1870-1883.

Bappenas. (2016). *Laporan Akhir Koordinasi Strategis Percepatan Pelaksanaan Program Pembangunan Daerah Tertinggal untuk Mendukung PP No. 78 Tahun 2014 dan Perpres No. 131 Tahun 2015*. Jakarta: Direktorat Daerah Tertinggal, Transmigrasi, dan Perdesaan.

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The Fuzzy C-Means Clustering Algorithm. *Computer and Geoscience, 10*, pp. 191-203.

Daniel, W. (1989). *Statistika Nonparametrik Terapan.* (Alex Tri Kantjono W., Trans.) Jakarta: PT. Gramedia.

Dewi, A. (2012). *Metode Cluster Ensemble untuk Pengelompokan Desa Perdesaan di Provinsi Riau.* Thesis, Institut Teknologi Sepuluh Nopember, Statistika, Surabaya.

Hair, J., Black, W., Babin, B., & Anderson, R. (2010). *Multivariate Data Analysis* (7th ed.). London: Pearson Prentice Hall.

He, Z., Xu, X., & Deng, S. (2005). *Clustering Mixed Numeric and Categorical Data: A Cluster Ensemble Approach.* Harbin Institute of Technology, Department of Computer Science and Engineering.

Huang, Z. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining Knowledge Discovery, 2*, 283-304.

Huang, Z., & Ng, M. K. (1999). A Fuzzy K-Modes Algorithm for Clustering Categorical Data. *IEEE Transactions On Fuzzy Systems, 7*(4).

Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis* (6th ed.). New Jersey: Prentice Hall, Englewood Cliffs.

Kusumadewi, S., & Hari, P. (2004). *Aplikasi Logika Fuzzy untuk Pendukung Keputusan.* Yogyakarta: Graha Ilmu.

Narimawati, U. (2008). *Metodologi Penelitian Kualitatif dan Kuantitatif, Teori dan Aplikasi.* Bandung: Agung Medai.

Orpin, A., & Kostylev, V. (2006). *Towards a Statistically Valid Method of Textural Sea Floor Characterization of Benthic Habitats.* Marine Geology.

Prakoso, A. J. (2017). *Pengelompokan Sekolah Menengah Atas (SMA) di Sidoarjo Menggunakan Similarity Weighted and Filter Method (SWFM).* Skripsi, Institut Teknologi Sepuluh Nopember, Statistika, Surabaya.

Ramdhany, R. (2018). *Pengelompokan Desa di Kabupaten Bondowoso Berdasarkan Data Campuran Numerik dan Kategorik Menggunakan Metode Ensembel ROCK.* Skripsi, Institut Teknologi Sepuluh Nopember, Statistika, Surabaya.

Reddy, M., & Khavita, B. (2012, March). Clustering the mixed Numerical and Categorical Dataset using Similarity Weight and Filter Method. *International Journal of Database Theory and Application, 5*(1).

Rencher, A. (1998). *Multivariate Statistical Inference and Application.* Brigham: Wiley-Interscience Publication.

Rizal, A. S., & Hakim, R. B. (2015). Metode K-Means Cluster dan Fuzzy C-Means Cluster (Studi Kasus: Indeks Pembangunan Manusia di Kawasan Indonesia Timur tahun 2012). *Prosiding Seminar Nasional Matematika dan Pendidikan Matematika UMS.*

Saputra, C. W. (2016). *Pengelompokan Aksesi Jeruk Persilangan Berdasarkan Karakter Kuantitatif dan Kualitatif Menggunakan Fuzzy C-Menas dan K-Modes.* Skripsi, Institut Teknologi Sepuluh Nopember, Statistika.

Suguna, J. (2012). Ensemble Fuzzy Clustering for Mixed Numeric and Categorical Data. *International Journal od Computer Application, 42*(3).

Sukim. (2011). *Metode C-Means Cluster dan Fuzzy C-Means Cluster Serta Aplikasinya pada Kasus Pengelompokkan Desa/Kelurahan Berdasarkan Status Ketertinggalan.* Thesis, Institut Teknologi Sepuluh Nopember, Jurusan Statistika FMIPA.

Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2007). *Probability & Statistics for Engineers & Scientist.* New Jersey: Pearson Prentice Hall.

Yoon, H., Ahn, S., Lee, S., Cho, S., & Kim, J. (2006). Heterogeneous Clustering Ensemble Method for Combining Different Cluster Results. *BioDM 2006. Lecture Notes in Computer Science, 3916*.

(Halaman Ini Sengaja Dikosongkan)

LAMPIRAN

Lampiran 1 Data Penelitian Variabel Numerik

Kab/Kota	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Pacitan	0.94	0.002	0.0023	99.61	0.056	16.73	4.93	70.75	89.57	1.08
Ponorogo	0.90	0.006	0.0029	100	0.027	11.92	4.93	71.88	95.02	3.66
Trenggalek	0.87	0.003	0.0023	99.55	0.021	13.56	5.37	72.51	95.4	4.2
Tulungagung	0.97	0.007	0.0014	100	0.018	9.07	5.88	72.88	96.97	2.42
Blitar	0.98	0.004	0.0018	99.49	0.031	10.57	6.14	72.5	92.21	3.08
Kediri	0.98	0.006	0.0019	99.82	0.255	13.23	5.45	72.04	93.16	4.91
Malang	0.97	0.015	0.0022	99.68	0.074	11.48	5.99	71.78	93.27	4.83
Lumajang	0.92	0.003	0.0023	100	0.113	12.14	4.44	69.07	87.03	2.83
Jember	0.73	0.006	0.0025	97.82	0.036	11.68	4.98	67.8	89.77	4.64
Banyuwangi	0.93	0.008	0.0020	98.4	0.019	9.61	6.48	69.93	94.99	7.17
Bondowoso	0.73	0.003	0.0035	96.82	0.063	15.29	4.80	65.43	86.91	3.72
Situbondo	0.97	0.004	0.0030	99.15	0.095	13.65	5.08	68.08	85.77	4.15
Probolinggo	0.96	0.004	0.0034	98.09	0.027	21.21	5.22	65.75	86.41	1.47
Pasuruan	0.89	0.003	0.0019	97.99	0.051	11.26	5.41	69.83	94.8	4.43
Sidoarjo	0.99	0.014	0.0018	100	0.000	6.72	8.93	73.43	98.05	3.88
…	…	…	…	…	…	…	…	…	…	…
Magetan	0.96	0.001	0.0022	99.73	0.003	12.19	6.48	71.91	95.57	4.28
Ngawi	0.46	0.003	0.0020	99.9	0.004	15.45	5.75	71.33	89.89	5.61
Bojonegoro	0.92	0.008	0.0030	99.29	0.003	16.02	5.53	70.11	89.24	3.21
Tuban	0.98	0.003	0.0019	99.92	0.078	17.23	5.59	70.25	86.3	3.63
Lamongan	0.91	0.009	0.0037	99.88	0.000	16.18	6.80	71.47	93.92	4.3
Gresik	0.96	0.009	0.0028	99.8	0.001	13.94	7.94	72.2	97.54	5.06
Bangkalan	0.92	0.002	0.0025	97.87	0.038	23.23	5.43	69.62	85.27	5.68
Sampang	0.78	0.001	0.0027	99.6	0.025	27.08	4.75	67.48	77.93	2.22
Pamekasan	0.88	0.002	0.0039	100	0.063	18.53	5.05	66.56	88.27	2.14
Sumenep	0.85	0.001	0.0045	81.04	0.072	21.22	4.73	70.02	84.37	1.01
Kota Kediri	1.00	0.005	0.0017	99.81	0.000	8.23	8.07	73.52	97.04	7.66
Kota Blitar	1.00	0.004	0.0014	100	0.000	7.42	9.49	72.7	95.67	5.71
Kota Malang	1.00	0.008	0.0024	100	0.000	4.87	12.16	72.3	97.45	7.22
Kota Probolinggo	1.00	0.002	0.0030	100	0.049	8.55	9.35	69.52	95.57	5.16
Kota Pasuruan	1.00	0.001	0.0017	98.91	0.017	7.6	7.57	70.54	98.52	6.09
Kota Mojokerto	1.00	0.005	0.0020	99.73	0.000	6.65	8.50	72.39	97.44	4.42
Kota Madiun	1.00	0.005	0.0032	100	0.000	5.02	9.61	72.41	96.05	6.93
Kota Surabaya	1.00	0.03	0.0022	99.42	0.000	6	13.89	73.85	97.04	5.82
Kota Batu	1.00	0.002	0.0028	100	0.034	4.77	8.40	72.06	94.79	2.43

Lampiran 2 Data Penelitian Variabel Kategorik

Kab/Kota	X11	X12	X13
Pacitan	2	0	6
Ponorogo	2	1	4
Trenggalek	2	0	5
Tulungagung	2	1	5
Blitar	2	0	5
Kediri	2	1	4
Malang	2	1	6
Lumajang	2	1	5
Jember	2	1	5
Banyuwangi	2	1	5
Bondowoso	2	0	6
Situbondo	2	0	4
Probolinggo	2	1	4
Pasuruan	1	1	4
Sidoarjo	2	3	1
…	…	…	…
Magetan	2	1	4
Ngawi	2	1	4
Bojonegoro	2	0	4
Tuban	2	0	2
Lamongan	2	1	5
Gresik	2	1	4
Bangkalan	2	1	1
Sampang	2	1	1
Pamekasan	1	3	1
Sumenep	2	0	5
Kota Kediri	2	0	5
Kota Blitar	2	1	5
Kota Malang	2	0	5
Kota Probolinggo	2	1	4
Kota Pasuruan	2	1	5
Kota Mojokerto	2	2	1
Kota Madiun	2	1	4
Kota Surabaya	2	1	1
Kota Batu	2	1	1

Lampiran 3 Output Mshapiro test

```
> data= read.csv("D:/KULIAH/SEMESTER 8/TUGAS AKHI
R/data/Book1.csv", sep=';', header=FALSE)
> a=t(data[1:38,1:10])
> library(mvnormtest)
> mshapiro.test(a)
        Shapiro-Wilk normality test
data:  Z
W = 0.26891, p-value = 1.537e-12
```

Lampiran 4 Syntax Pengelompokan Numerik *Fuzzy C-Means*

```
#memanggil package untuk FCM#
library (ppclust)

#memanggil data yang digunakan dalam format txt#
data=read.table("D:/LARAS/SEMESTER 8/TUGAS
AKHIR/data/FCM/data3.txt",header=FALSE)

#mendefinisikan derajat keanggotaan awal yang telah dihitung seperti
ilustrasi pada subbab 2.3.1#
u2_naik=read.table("D:/LARAS/SEMESTER 8/TUGAS
AKHIR/data/FCM/u2_naik3.txt",header=FALSE)
u2_turun=read.table("D:/LARAS/SEMESTER 8/TUGAS
AKHIR/data/FCM/u2_turun3.txt",header=FALSE)
u2_segitiga=read.table("D:/LARAS/SEMESTER 8/TUGAS
AKHIR/data/FCM/u2_segitiga3.txt",header=FALSE)
u2_trapesium=read.table("D:/LARAS/SEMESTER 8/TUGAS
AKHIR/data/FCM/u2_trapesium3.txt",header=FALSE)

#melakukan pengelompokan dengan masing-masing derajat keanggotaan#
#derajat keanggotaan linier naik
k2_naik=fcm(data, centers=2, memberships=u2_naik, iter.max = 100,
con.val = 10^-6)
SST=k2_naik$sumsqrs$tot.ss
SSW=k2_naik$sumsqrs$tot.within.ss
Rsq_k2naik=(SST-SSW)/SST
cluster2naik=k2_naik$cluster
membership2naik=k2_naik$u

#derajat keanggotaan linier turun
k2_turun=fcm(data, centers=2, memberships=u2_turun, iter.max = 100,
con.val = 10^-6)
SST=k2_turun$sumsqrs$tot.ss
SSW=k2_turun$sumsqrs$tot.within.ss
Rsq_k2turun=(SST-SSW)/SST
cluster2turun=k2_turun$cluster
membership2turun=k2_turun$u
```

Lampiran 4 *Syntax Pengelompokan Numerik* Fuzzy C-Means (Lanjutan)

```
#derajat keanggotaan kurva segitiga
k2_segitiga=fcm(data, centers=2, memberships=u2_segitiga,
iter.max = 100, con.val = 10^-6)
SST=k2_segitiga$sumsqrs$tot.ss
SSW=k2_segitiga$sumsqrs$tot.within.ss
Rsq_k2segitiga=(SST-SSW)/SST
cluster2segitiga=k2_segitiga$cluster
membership2segitiga=k2_segitiga$u

#derajat keanggotaan kurva trapesium
k2_trapesium=fcm(data, centers=2,
memberships=u2_trapesium, iter.max = 100, con.val = 10^-6)
SST=k2_trapesium$sumsqrs$tot.ss
SSW=k2_trapesium$sumsqrs$tot.within.ss
Rsq_k2trapesium=(SST-SSW)/SST
cluster2trapesium=k2_trapesium$cluster
membership2trapesium=k2_trapesium$u

#menyimpan nilai kebaikan dan anggota hasil pengelompokan#
Rsq2=data.frame(Rsq_k2naik, Rsq_k2turun, Rsq_k2segitiga,
Rsq_k2trapesium)
Hasil2cluster=data.frame(cluster2naik, membership2naik,
cluster2turun, membership2turun, cluster2segitiga,
membership2segitiga, cluster2trapesium,
membership2trapesium)

#disimpan dalam format csv#
write.csv(Hasil2cluster, "D:/Hasil2cluster.csv")
write.table(Rsq2, "D:/Rsq2.txt")
```

Lampiran 5 Syntax Pengelompokan Kategorik *K-Modes*

```
#memanggil package untuk K-Modes#
library(klaR)

#memanggil data yang digunakan dalam format txt#
data.to.cluster <- read.csv("D:/KULIAH/SEMESTER
8/TUGAS AKHIR/data/KATEGORIK.csv", header = TRUE,
sep = ';')

#melakukan pengelompokan dengan metode K-Modes#
cluster.results <-kmodes(data.to.cluster[-1],5, iter.max = 100,
weighted = FALSE ) #iter.max dan jumlah cluster
cluster.results
summary(cluster.results)

#menyimpan hasil anggota pengelompokan#
cluster.output <- cbind(data.to.cluster,cluster.results$cluster)
write.csv(cluster.output, file = "d:/kmodes5.csv", row.names =
TRUE)
```

Lampiran 6 Syntax Pengelompokan Campuran SWFM

```
data<- read.csv("D:/KULIAH/SEMESTER 8/TUGAS
AKHIR/data ensemble2.csv", header = TRUE, sep = ';')
data_ensemble<- data[-1]
sij = function(x,y)
{
  sij = length(intersect(x,y))/length(union(x,y))
  p = length(unique(x))
  q = length(unique(y))
  m = max (p,q)
  sm = sum (sij/m)
  return(sm)
}
  sim = matrix (1, nrow = 38, ncol = 38)
  rownames(sim) = c(1:38)
  colnames(sim) = c(1:38)
  for (i in 1:(nrow(sim)-1))
  {
    for (j in (i+1):nrow(sim))
    {
      x = data_ensemble[i,]
      y = data_ensemble[j,]
      sim [i,j] = sij(x,y)
      sim [j,i] = sim[i,j]
    }
  }
wi = 0.5
F = wi*(1-sim)
jarak = as.dist(F)
fit.sin = hclust(jarak,method = "single")
hasil.k2 = cutree(fit.sin, k=2)
hasil.k3 = cutree(fit.sin, k=3)
hasil.k4 = cutree(fit.sin, k=4)
hasil.k5 = cutree(fit.sin, k=5)
hasil.kelompok = data.frame(hasil.k2,hasil.k3,hasil.k4,hasil.k5)
SWFM = data.frame(data, hasil.kelompok)
write.csv(SWFM, "D:/KULIAH/SEMESTER 8/TUGAS
AKHIR/SWFM.csv")
```

Lampiran 7 Output Analisis Faktor

a) Matriks Korelasi

		X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Corre-lation	X1	1.000	.210	-.200	.182	.036	-.436	.442	.385	.435	.120
	X2	.210	1.000	-.084	.172	-.151	-.361	.578	.432	.399	.281
	X3	-.200	-.084	1.000	-.478	.068	.433	-.197	-.486	-.392	-.379
	X4	.182	.172	-.478	1.000	-.129	-.343	.232	.208	.345	.310
	X5	.036	-.151	.068	-.129	1.000	.225	-.406	-.241	-.315	-.226
	X6	-.436	-.361	.433	-.343	.225	1.000	-.691	-.619	-.844	-.527
	X7	.442	.578	-.197	.232	-.406	-.691	1.000	.573	.641	.562
	X8	.385	.432	-.486	.208	-.241	-.619	.573	1.000	.738	.401
	X9	.435	.399	-.392	.345	-.315	-.844	.641	.738	1.000	.545
	X10	.120	.281	-.379	.310	-.226	-.527	.562	.401	.545	1.000

b) Pemeriksaan KMO dan Uji Bartlett

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.705
Bartlett's Test of Sphericity	Approx. Chi-Square	173.302
	df	45
	Sig.	.000

Lampiran 7 Output Analisis Faktor (Lanjutan)

c) Persentase Variansi Kumulatif

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	4.542	45.421	45.421
2	1.240	12.399	57.819
3	1.079	10.792	68.611
4	.767	7.672	76.283
5	.711	7.109	83.392
6	.628	6.279	89.671
7	.436	4.364	94.035
8	.291	2.908	96.943
9	.216	2.164	99.108
10	.089	.892	100.000

d) Rotated Component Matrix

Rotated Component Matrix[a]

	Component		
	1	2	3
Aspal	.651	.104	.526
RS	.678	-.095	-.164
SMA	-.149	-.857	-.071
Listrik PLN	.089	.764	-.066
lahan kritis	-.248	-.047	.805
Persentase penduduk miskin	-.755	-.436	.055
pengeluaran per kapita	.852	.106	-.285
AHH	.725	.354	-.031
AMH	.780	.414	-.122
Persentase pengangguran	.436	.473	-.376

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 4 iterations.

Lampiran 8 Hasil Pengelompokan Numerik *Fuzzy C-Means* 3 Variabel 5 Klaster

Kab/Kota	Linier Naik	Linier Turun	Kurva Segitiga	Kurva Trapesium
Pacitan	4	2	3	1
Ponorogo	4	2	3	1
Trenggalek	4	2	3	1
Tulungagung	1	5	2	3
Blitar	1	5	2	3
Kediri	4	2	3	1
Malang	1	5	2	3
Lumajang	4	2	3	1
Jember	4	2	3	1
Banyuwangi	1	5	2	3
Bondowoso	4	2	3	1
Situbondo	4	2	3	1
Probolinggo	4	2	3	1
Pasuruan	4	2	3	1
Sidoarjo	3	1	5	2
Mojokerto	5	3	1	4
Jombang	1	5	2	3
Nganjuk	4	2	3	1
Madiun	1	5	2	3
Magetan	1	5	2	3
Ngawi	1	5	2	3
Bojonegoro	4	2	3	1
Tuban	4	2	3	1
Lamongan	1	5	2	3
Gresik	5	3	1	4
Bangkalan	4	2	3	1
Sampang	4	2	3	1
Pamekasan	4	2	3	1
Sumenep	4	2	3	1
Kota Kediri	5	3	1	4
Kota Blitar	3	1	5	2
Kota Malang	2	4	4	5
Kota Probolinggo	3	1	5	2
Kota Pasuruan	5	3	1	4
Kota Mojokerto	5	3	1	4
Kota Madiun	3	1	5	2
Kota Surabaya	2	4	4	5
Kota Batu	5	3	1	4

Lampiran 9 Hasil Pengelompokan Kategorik *K-Modes* 5 Klaster

Kab/Kota	2 Klaster	3 Klaster	4 Klaster	5 Klaster
Pacitan	1	1	3	3
Ponorogo	1	2	2	2
Trenggalek	1	1	3	3
Tulungagung	1	1	1	4
Blitar	1	1	3	3
Kediri	1	2	2	2
Malang	1	2	1	2
Lumajang	1	1	1	4
Jember	1	1	1	4
Banyuwangi	1	1	1	4
Bondowoso	1	1	3	3
Situbondo	1	1	2	2
Probolinggo	1	2	2	2
Pasuruan	2	2	2	2
Sidoarjo	1	1	4	1
Mojokerto	2	2	2	2
Jombang	1	2	2	2
Nganjuk	1	2	2	2
Madiun	1	1	2	2
Magetan	1	3	3	3
Ngawi	1	1	1	4
Bojonegoro	1	2	2	2
Tuban	1	2	1	2
Lamongan	1	2	1	2
Gresik	2	1	4	1
Bangkalan	1	1	3	3
Sampang	1	1	3	3
Pamekasan	1	1	1	4
Sumenep	1	1	3	3
Kota Kediri	1	2	2	2
Kota Blitar	1	1	1	4
Kota Malang	1	1	1	5
Kota Probolinggo	1	2	2	2
Kota Pasuruan	1	2	1	2
Kota Mojokerto	1	2	1	2
Kota Madiun	1	2	1	2
Kota Surabaya	2	1	4	1
Kota Batu	1	2	1	2

Lampiran 10 Data dan Hasil Pengelompokan *Ensemble K-Modes* dan SWFM

Kab/Kota	Hasil FCM	Hasil K-Modes	Ensemble	SWFM
Pacitan	2	3	2	1
Ponorogo	2	2	5	1
Trenggalek	2	3	2	1
Tulungagung	5	4	2	2
Blitar	5	3	2	3
Kediri	2	2	5	1
Malang	5	2	2	1
Lumajang	2	4	5	1
Jember	2	4	5	1
Banyuwangi	5	4	2	2
Bondowoso	2	3	2	1
Situbondo	2	2	5	1
Probolinggo	2	2	5	1
Pasuruan	2	2	5	1
Sidoarjo	1	1	4	1
Mojokerto	3	2	3	1
Jombang	5	2	2	1
Nganjuk	2	2	5	1
Madiun	5	2	2	1
Magetan	5	3	2	4
Ngawi	5	4	2	2
Bojonegoro	2	2	5	1
Tuban	2	2	5	1
Lamongan	5	2	2	1
Gresik	3	1	1	1
Bangkalan	2	3	2	1
Sampang	2	3	2	1
Pamekasan	2	4	5	1
Sumenep	2	3	2	1
Kota Kediri	3	2	3	1
Kota Blitar	1	4	4	1
Kota Malang	4	5	1	5
Kota Probolinggo	1	2	3	1
Kota Pasuruan	3	2	3	1
Kota Mojokerto	3	2	3	1
Kota Madiun	1	2	3	1
Kota Surabaya	4	1	1	1
Kota Batu	3	2	3	1

Lampiran 11 *Output* Pengujian Beda Mean MANOVA

a) Uji Box's M

Box's Test of Equality of Covariance Matrices[a]

Box's M	210.342
F	1.963
df1	55
df2	1504.925
Sig.	.000

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.
a. Design: Intercept + klaster

b) Multivariate Tests

Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	1.000	7902.203[b]	10.000	24.000	.000
	Wilks' Lambda	.000	7902.203[b]	10.000	24.000	.000
	Hotelling's Trace	3292.584	7902.203[b]	10.000	24.000	.000
	Roy's Largest Root	3292.584	7902.203[b]	10.000	24.000	.000
Klaster	Pillai's Trace	1.642	1.880	40.000	108.000	.005
	Wilks' Lambda	.059	2.571	40.000	92.861	.000
	Hotelling's Trace	6.604	3.715	40.000	90.000	.000
	Roy's Largest Root	5.395	14.567[c]	10.000	27.000	.000

a. Design: Intercept + Klaster
b. Exact statistic
c. The statistic is an upper bound on F that yields a lower bound on the significance level.

Lampiran 12 *Output* Pengujian ANOVA

One-way ANOVA: Jalan Aspal versus Klaster

Source	DF	Adj SS	Adj MS	F-Value	P-Value
klaster	4	0,09462	0,023655	2,55	0,058
Error	33	0,30648	0,009287		
Total	37	0,40110			

One-way ANOVA: Rumah Sakit versus Klaster

Source	DF	Adj SS	Adj MS	F-Value	P-Value
klaster	4	0,000364	0,000091	4,36	0,006
Error	33	0,000689	0,000021		
Total	37	0,001052			

One-way ANOVA: SMA versus Klaster

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Klaster	4	0,000002	0,000000	0,94	0,453
Error	33	0,000017	0,000001		
Total	37	0,000018			

One-way ANOVA: Listrik PLN versus Klaster

Source	DF	Adj SS	Adj MS	F-Value	P-Value
klaster	4	21,08	5,270	0,53	0,718
Error	33	330,78	10,024		
Total	37	351,86			

Lampiran 15 *Output* Pengujian ANOVA (Lanjutan)

One-way ANOVA: Lahan Kritis versus Klaster

Source	DF	Adj SS	Adj MS	F-Value	P-Value
klaster	4	0,02231	0,005577	3,03	0,031
Error	33	0,06075	0,001841		
Total	37	0,08306			

One-way ANOVA: Persentase penduduk miskin versus Klaster

Source	DF	Adj SS	Adj MS	F-Value	P-Value
klaster	4	437,7	109,42	6,38	0,001
Error	33	566,3	17,16		
Total	37	1004,0			

One-way ANOVA: Pengeluaran per Kapita versus Klaster

Source	DF	Adj SS	Adj MS	F-Value	P-Value
klaster	4	136,57	34,1429	36,01	0,000
Error	33	31,29	0,9481		
Total	37	167,86			

One-way ANOVA: AHH versus Klaster

Source	DF	Adj SS	Adj MS	F-Value	P-Value
klaster	4	53,73	13,431	3,98	0,010
Error	33	111,24	3,371		
Total	37	164,96			

Lampiran 15 *Output* Pengujian ANOVA (Lanjutan)

One-way ANOVA: AMH versus Klaster

Source	DF	Adj SS	Adj MS	F-Value	P-Value
klaster	4	333,2	83,30	5,14	0,002
Error	33	534,9	16,21		
Total	37	868,1			

One-way ANOVA: Persentase Pengangguran versus Klaster

Source	DF	Adj SS	Adj MS	F-Value	P-Value
klaster	4	24,79	6,199	2,70	0,048
Error	33	75,90	2,300		
Total	37	100,70			

Lampiran 13 *Output* Uji *Kruskal Wallis*

Test Statistics[a,b]

	Keberadaan Kawasan Industri	Tingkat Desentralisasi Fiskal	Sumber Air Minum
Chi-Square	11.848	21.412	12.359
df	4	4	4
Asymp. Sig.	.019	.000	.015

a. Kruskal Wallis Test
b. Grouping Variable: klaster

Lampiran 14 Surat Keterangan Data Penelitian

BADAN PUSAT STATISTIK
PROVINSI JAWA TIMUR

SENSUS EKONOMI

SURAT KETERANGAN

Yang bertanda tangan dibawah ini :

N a m a	:	Thomas Wunang Tjahjo, M.Sc, M.Eng.
N I P	:	19700329 1992 11 1 001
Jabatan	:	Kepala Bidang Integrasi Pengolahan dan Diseminasi Statistik

Dengan ini menerangkan bahwa :

N a m a	:	Larasati Dewinta Paramardini
Fakultas/Program Studi	:	Fakultas Matematika, Komputasi dan Sains Data / Statistika
N.R.P	:	06211440000058
Alamat Rumah	:	Jalan Keputih Makam, No. 7 Kecamatan Sukolilo, Surabaya
Akademi / Universitas	:	Institut Teknologi Sepuluh Nopember (ITS) Telp (031) 594 3352, (031) 599 4251-55 Fax (031) 592 2940

Benar-benar telah mencari data di Kantor Badan Pusat Statistik (BPS) Provinsi Jawa Timur dalam rangka menyusun Tugas Akhir / Skripsi dengan judul :

"Penerapan Metode Data Campuran Ensemble K-Modes dan Similarity Weight and Filter Method (SWFM) pada Pengelompokan Kabupaten/Kota di Jawa Timur Berdasarkan Indikator Daerah Tertinggal"

Demikian surat keterangan ini dibuat dan agar dipergunakan sebagaimana mestinya

Surabaya, 2 Mei 2018

An. Kepala BPS Provinsi Jawa Timur
Kepala Bidang IPDS

Thomas Wunang Tjahjo, M.Sc, M.Eng.

Jalan Raya Kendangsari Industri No. 43 - 44, Surabaya - 60292
Telp. 031 - 8439343 Fax. 031 - 8494007. Homepage: http://jatim.bps.go.id E-mail: bps3500@bps.go.id

BIODATA PENULIS

Penulis memiliki nama lengkap Larasati Dewinta Paramardini atau akrab dengan panggilan Laras. Lahir di Kota Dili pada tanggal 10 Desember 1996 dan merupakan anak pertama dari dua bersaudara. Penulis tumbuh dan besar di Kota Madiun. Selama di Madiun, penulis pernah menempuh pendidikan di SDN 02 Mojorejo, SMPN 2 Madiun, dan SMAN 3 Madiun. Kemudian melanjutkan pendidikan pada jenjang perguruan tinggi di Institut Teknologi Sepuluh Nopember Departemen Statistika melalui jalur SNMPTN. Selama masa perkuliahan, penulis aktif dalam bidang organisasi yaitu Himpunan Mahasiswa Statistika ITS (HIMASTA-ITS). Pada tahun kedua sebagai staff Departemen Kewirausahaan serta staff Divisi PERS. Tahun ketiga penulis berkesempatan menjadi Ketua Biro Departemen Kewirausahaan HIMASTA-ITS.

Penulis terbuka dalam menerima kritik dan saran serta diskusi mengenai Tugas Akhir ini. Apabila ada yang ditanyakan serta ingin memberikan masukan terhadap penelitian ini, dapat dikirimkan melalui *email* larasdwnt@gmail.com.