

Laila Qadrini

ALGORITMA PENGELOMPOKAN

ENSEMBEL FUZZY, K-PROTOTYPES
DAN
DENSITY PEAKS CLUSTERING MIXED (DPC-M)



**ALGORITMA PENGELOMPOKAN ENSEMBEL
*FUZZY, K-PROTOTYPES DAN DENSITY PEAKS
CLUSTERING MIXED (DPC-M)***

**ALGORITMA PENGELOMPOKAN ENSEMBEL *FUZZY*, K-
PROTOTYPES DAN *DENSITY PEAKS CLUSTERING MIXED*
(*DPC-M*)**

LAILA QADRINI



ALGORITMA PENGELOMPOKAN ENSEMBEL *FUZZY*, K- PROTOTYPES DAN DENSITY PEAKS CLUSTERING MIXED (DPC-M)

© Penerbit Perkumpulan Rumah Cemerlang Indonesia (PRCI)

Penulis:

LAILA QADRINI

Editor:

Erik Santoso

Cetakan Pertama : Juli 2021

Cover:

Dani Kusuma

Tata Letak : Tim Kreatif PRCI

Hak Cipta 2021, pada Penulis. Diterbitkan pertama kali oleh:

**Perkumpulan Rumah Cemerlang Indonesia
ANGGOTA IKAPI JAWA BARAT**

Pondok Karisma Residence Jalan Raflesia VI D.151
Panglayungan, Cipedes Tasikmalaya – 085223186009

Website : www.rcipress.rcipublisher.org

E-mail : rumahcemerlangindonesia@gmail.com

Copyright © 2021 by Perkumpulan Rumah Cemerlang Indonesia
All Right Reserved

- Cet. I – : Perkumpulan Rumah Cemerlang Indonesia, 2021
Vi+63; 15,5 x 23 cm
ISBN : 978-623-6535-84-4

Hak cipta dilindungi undang-undang
Dilarang memperbanyak buku ini dalam bentuk dan dengan
cara apapun tanpa izin tertulis dari penulis dan penerbit

Isi diluar tanggung jawab Penerbit
Undang-undang No.19 Tahun 2002 Tentang
Hak Cipta Pasal 72

Undang-undang No.19 Tahun 2002 Tentang Hak Cipta
Pasal 72

Barang siapa dengan sengaja melanggar dan tanpa hak melakukan perbuatan sebagaimana dimaksud dalam pasal ayat (1) atau pasal 49 ayat (1) dan ayat (2) dipidana dengan pidana penjara masing-masing paling sedikit 1 (satu) bulan dan/atau denda paling sedikit Rp.1.000.000,00 (satu juta rupiah), atau pidana penjara paling lama 7 (tujuh) tahun dan/atau denda paling banyak Rp.5.000.000.000,00 (lima miliar rupiah).

Barang siapa dengan sengaja menyiarkan, memamerkan, mengedarkan, atau menjual kepada umum suatu ciptaan atau barang hasil pelanggaran hak cipta terkait sebagai dimaksud pada ayat (1) dipidana dengan pidana penjara paling lama 5 (lima) tahun dan/atau denda paling banyak Rp.500.000.000,00 (lima ratus juta rupiah)

KATA PENGANTAR

Puji syukur kami panjatkan kehadiran Allah SWT yang telah memberikan rahmat serta karunia-Nya kepada kami sehingga kami berhasil menyelesaikan Buku dengan judul *ALGORITMA PENGELOMPOKAN ENSEMBEL FUZZY, K-PROTOTYPES DAN DENSITY PEAKS CLUSTERING MIXED (DPC-M)* sesuai yang ditargetkan. Buku ini berisikan bagaimana menganalisis data peserta didik misi menggunakan Algoritma Pengelompokan Ensemble *Fuzzy, K-Prototypes* dan *Density Peaks Clustering Mixed (DPC-M)*. Kami menyadari bahwa Buku ini masih jauh dari sempurna, oleh karena itu kritik dan saran dari semua pihak yang bersifat membangun selalu kami harapkan demi kesempurnaan buku ini.

Akhir kata, kami sampaikan terima kasih kepada semua pihak yang telah berperan serta dalam penyusunan Buku ini dari awal sampai akhir. Semoga Allah SWT senantiasa meridhoi segala usaha kita. Amin.

Juni 2021, Penulis

DAFTAR ISI

KATA PENGANTAR	i
DAFTAR ISI.....	ii
DAFTAR TABEL.....	v
DAFTAR GAMBAR.....	vi
BAB I	
PENGANTAR ALGORITMA PENGELOMPOKAN ENSEMBEL FUZZY, K-PROTOTYPES DAN DENSITY PEAKS CLUSTERING MIXED (DPC-M).....	
	1
BAB II	
DATA MINING PADA PENGELOLAAN BEASISWA BIDIK MISI	
	5
A. Tujuan Bidikmisi	6
B. Syarat penerima Bidik misi.....	6
C. Jangka Waktu Pemberian Bantuan.....	6
1. Hak	7
2. Kewajiban	7
D. Mekanisme Pendaftaran Bidikmisi.....	8
1. Pendaftaran Daring (<i>On line</i>).....	8
E. Mekanisme Penetapan.....	9
F. Data Mining.....	9
G. Data Numerik dan Data Kategorik.....	11
1. Numerik.....	11
2. Kategorik.....	11
BAB III	
ANALISIS DATA MINING	
	12
A. Analisis Kelompok.....	13
B. Metode <i>Fuzzy C-Means</i>	16
C. Metode <i>Fuzzy C-Modes</i>	18
D. Pengelompokan Ensemble	19
E. Metode <i>Kprototypes</i>	21
1. Inisialisasi awal <i>prototype</i>	22
2. Alokasi <i>object</i> di dalam X ke Kelompok dengan <i>prototype</i> terdekat	22
3. Realokasi <i>object</i> jika terjadi perubahan <i>prototype</i> ..	22

F. Ukuran Kesamaan (<i>Similarity Measure</i>)	23
G. Pengelompokan data campuran berdasarkan Puncak Densitas (DPC_M).....	25
1. Algoritma DPC	25
2. Ukuran Jarak Terpadu.....	27

BAB IV

<i>Preprocessing Data Mining</i>	29
A. Karakteristik dari data mentah.....	30
B. Transformasi Data.....	30
C. Penanganan terhadap data yang hilang.....	30
D. Analisa <i>outlier</i>	31
1. Deteksi outlier berdasarkan teknik statistik.....	31
2. <i>Distance Based Outlier Detection</i>	31
E. Evaluasi Pengelompokan.....	32
1. Uji Validitas kelompok	32
2. Ukuran internal	32
F. Ukuran Eksternal.....	34
1. Purity Measure.....	34
2. Entropy	35

BAB V

HASIL PENGOLAHAN ALGORITMA DATA MINING.....	37
A. Jenis Skala Data.....	38
B. Deskripsi Karakteristik Variabel.....	45
C. Algoritma Pemrograman Metode DPC-M	45
D. Karakteristik Hasil Pengelompokan Metode <i>Ensembl</i> <i>Fuzzy</i>	47
1. Kelompok 1.....	47
2. Kelompok 2.....	47
3. Kelompok 3.....	48
4. Metode Ensembl Fuzzy.....	49
5. Kelompok 2.....	50
6. Kelompok 3.....	50
E. Perhitungan Indeks Validasi Internal Metode DPC-M	51
F. Perhitungan Indeks Validasi Eksternal Metode DPC-M	52
G. Algoritma Pemrograman Metode K-prototypes.....	52

H. Karakteristik Hasil Pengelompokan Metode K-prototypes	54
1. Kelompok 1	54
2. Kelompok 2	54
3. Kelompok 3	55
I. Perhitungan Indeks Validasi Internal Metode K-Prototypes	56
J. Perhitungan Indeks Validasi Eksternal Metode K-Prototypes	56
K. Perbandingan Kinerja Metode K-Prototypes, DPC-M.	57
DAFTAR PUSTAKA.....	59

DAFTAR TABEL

Tabel 1 Tabel kontingensi data biner	14
Tabel 2 Ukuran jarak data biner.....	15
Tabel 3 Variabel skala numerik pelamar Bidikmisi	38
Tabel 4 Variabel skala kategorik pelamar Bidikmisi.....	38
Tabel 5 Deskripsi Karakteristik Data Numerik	45
Tabel 6 Perbandingan Hasil Analisis DPC-M dan K-Prototypes	57

DAFTAR GAMBAR

Gambar 1 <i>Overview</i> Pengelompokan Ensembel	20
Gambar 2 Hasil Pengelompokan Data dengan Metode DPC-M	51
Gambar 3 Hasil Pengelompokan Data dengan Metode K- Prototypes	56

BAB I

PENGANTAR ALGORITMA PENGELOMPOKAN ENSEMBEL FUZZY, K-PROTOTYPES DAN DENSITY PEAKS CLUSTERING MIXED (DPC-M)

Analisis pengelompokan (Kelompok *Analysis*) merupakan salah satu metode dalam data Mining. Metode kelompok dalam data mining berbeda dengan metode konvensional yang biasa digunakan untuk pengelompokan. Perbedaannya adalah data mining memiliki dimensi data yang tinggi yaitu bisa terdiri dari puluhan ribu atau jutaan *record* dengan puluhan ataupun ratusan atribut. Selain itu pada data mining data bisa terdiri dari tipe data campuran seperti data numerik dan kategorikal. Permasalahan yang sering ditemui dalam analisis pengelompokan adalah data yang berskala campuran numerik dan kategorik. Metode yang seringkali dilakukan untuk pengelompokan data berskala campuran adalah dengan mentransformasi data kategorik menjadi data numerik dan sebaliknya. Dewangan, Sharma, dan Akasapu (2010) melakukan transformasi variabel kategorik ke dalam bentuk numerik, kemudian pengelompokan objek dilakukan dengan metode pengelompokan data numerik. Kelebihan metode transformasi adalah dapat mengurangi kompleksitas dalam komputasi. Akan tetapi, metode tersebut memiliki kelemahan dalam menentukan transformasi yang tepat agar tidak kehilangan banyak informasi dari data aslinya.

Selain pengelompokan dengan metode transformasi tersebut, dikembangkan sebuah metode pengelompokan ensemble untuk data campuran oleh He, Xu dan Deng (2005). Pengelompokan ensemble (pengelompokan *ensemble*) adalah teknik pengelompokan untuk menggabungkan hasil pengelompokan beberapa algoritma pengelompokan untuk mendapatkan kelompok yang lebih baik (He, Xu, dan Deng, 2005a). He Xu, dan Deng (2005b) menerapkan teknik pengelompokan ensemble untuk mengelompokkan penyakit hati dan persetujuan pengajuan kartu kredit dengan Kelompok *Ensemble Based Mixed Data Clustering* (CEBMDC) dan mengombinasikan tahapan ensemble ke dalam algoritma pengelompokan untuk data kategorik, yaitu algoritma *Squeezer*. Algoritma *Squeezer* merupakan metode pengelompokan yang efektif digunakan

untuk data yang berjumlah besar (Reddy & Kavitha, 2010). Alvionita (2017) melakukan perbandingan hasil antara metode ensemble ROCK dan ensemble SWFM. Kedua metode digunakan pada studi kasus pengelompokkan aksesori jeruk hasil fusi protoplasma yang merupakan data campuran numerik dan kategorik. Metode pengelompokan terbaik ditentukan dengan kriteria rasio antara simpangan baku di dalam kelompok (*SW*) dan simpangan baku antar kelompok (*SB*) terkecil.

Berdasarkan 25 objek pengamatan pada studi kasus, metode ensemble ROCK dengan nilai θ sebesar 0,27 menghasilkan tiga kelompok dengan nilai rasio sebesar 0,1358, sedangkan metode ensemble SWFM menghasilkan dua kelompok dengan nilai rasio sebesar 0,3059. Hasil tersebut menunjukkan bahwa metode ensemble ROCK memberikan hasil pengelompokan lebih baik daripada metode ensemble SWFM, J. Suguna & M. Arul Selvi (2015) membagi dataset campuran asli menjadi kumpulan data numerik dan kumpulan data kategorik dan dikelompokkan menggunakan algoritma pengelompokan tradisional (*K-Means* and *K-Modes*) dan algoritma pengelompokan fuzzy (*Fuzzy C-Means* dan *Fuzzy C-Modes*). Kelompok yang dihasilkan dikombinasikan dengan metode ensemble pengelompokan dan dievaluasi dengan ukuran *f-measure* dan *entropy*. Ditemukan bahwa pembagian data lebih menguntungkan dan menerapkan algoritma pengelompokan *fuzzy* menghasilkan hasil yang lebih baik daripada algoritma pengelompokan tradisional. Algoritma *K-Prototypes* diajukan oleh Huang pada tahun 1997, dan didasarkan pada gagasan algoritma *k-means*. Metode ini menggabungkan pusat kelompok dari atribut numerik dan atribut modus kategorik untuk dibangun sebuah pusat campuran baru. Pusat data adalah prototipe, dan ukuran jarak jauh proses pengelompokan *k-means* digunakan untuk mengelompokkan data campuran secara langsung. Rani Nooraeni (2015) telah melakukan simulasi yang hasilnya, metode *K-prototype* dapat meningkatkan kehomogenan dalam kelompok dan ketidakmiripan maksimal antar kelompok, data yang digunakan

adalah Dataset Podes 2011.

Ukuran data yang digunakan pada simulasi ini sebanyak 77.961 objek atau desa dan 37 atribut yang terdiri dari 17 atribut numerik dan 20 atribut kategorikal. Pada data numerik dinormalisasi terlebih dahulu. Jumlah kluster yang dibentuk adalah 10 kelompok. Reddy & Kavitha (2012) juga membandingkan pengelompokan antara metode ensembel dan metode *K-Prototype*. Hasil yang diperoleh adalah metode ensembel memberikan rata-rata kesalahan lebih rendah daripada metode *KPrototype*. Liu et al (2017) melakukan penelitian dengan mengacu pada data campuran yang terdiri dari atribut numerik dan kategorik, mengajukan ukuran disimilaritas baru dan algoritma pengelompokan baru. Hasil penelitian menunjukkan bahwa metode baru pengelompokan ini digabung data dengan pencarian cepat dan menemukan puncak kepadatan (DPC-M) terbukti layak dan efektif pada dataset UCI. Namun, penelitian-penelitian tersebut tidak dilakukan perbandingan hasil pengelompokan antara ketiga metode. Oleh karena itu, pada penelitian ini dilakukan perbandingan hasil pengelompokan antara metode pengelompokan ensembel *Fuzzy*, *K-prototypes* dan DPC-M. Hasil pengelompokan ketiga metode tersebut dilihat berdasarkan dua tipe untuk mengukur validitas kelompok, yaitu ukuran eksternal dan ukuran internal (Steinbach, Karypis, dan Kumar, 2000).

BAB II

DATA MINING PADA PENGELOLAAN BEASISWA BIDIK MISI

A. Tujuan Bidikmisi

1. Meningkatkan akses dan kesempatan belajar di perguruan tinggi bagi peserta didik yang tidak mampu secara ekonomi namun memiliki prestasi akademik yang baik
2. Meningkatkan prestasi mahasiswa, baik pada bidang kurikuler, kokurikuler, maupun ekstrakurikuler
3. Menimbulkan dampak iring bagi mahasiswa dan calon mahasiswa lain untuk berkarakter dan selalu meningkatkan prestasi; melahirkan lulusan yang mandiri, produktif, dan memiliki kepedulian sosial sehingga mampu berperan dalam upaya pemutusan mata rantai kemiskinan dan pemberdayaan masyarakat.

B. Syarat penerima Bidik misi

Penerima Bidikmisi memiliki syarat sebagai berikut.

1. Pendapatan kotor orang tua/wali gabungan (suami dan istri) setinggi-tingginya Rp4.000.000,00 (empat juta rupiah) atau pendapatan kotor gabungan orang tua/wali dibagi jumlah anggota keluarga maksimal Rp750.000,00 (tujuh ratus lima puluh ribu rupiah)
2. Ditetapkan oleh perguruan tinggi setiap tahun akademik
3. Mahasiswa aktif dan sedang menjalani perkuliahan pada semester normal

C. Jangka Waktu Pemberian Bantuan

Jangka waktu pemberian Bidikmisi diberikan dengan ketentuan sebagai berikut.

1. Sampai dengan semester 8 (delapan) untuk S1/D4
2. Sampai dengan semester 6 (enam) untuk D3
3. Sampai dengan semester 4 (empat) untuk D2
4. Sampai dengan semester 2 (dua) untuk D1

Penerima Bidikmisi yang cuti diberhentikan bantuannya. Pengelola perguruan tinggi dapat merekomendasikan yang

bersangkutan menerima Bidikmisi pada saat aktif kembali. Keputusan akhir pengaktifan diputuskan oleh Pengelola Pusat.

D. Hak dan Kewajiban Penerima Bidikmisi

Hak dan kewajiban penerima Bidikmisi adalah sebagai berikut.

1. Hak

1. Mendapatkan akses dan kesempatan mendapatkan pendidikan yang berkualitas sama dengan peserta didik lain di Perguruan Tinggi Penyelenggara Bidikmisi.
2. Wajib mendapatkan pembebasan biaya yang terdiri atas:
 - a. UKT/SPP atau sejenisnya yang bersifat operasional pendidikan
 - b. Biaya awal pendidikan yang mencakup biaya gedung, pembinaan, investasi, infak atau sejenisnya
 - c. Biaya praktikum di laboratorium, bahan, atau biaya pendidikan lain yang belum dicakup UKT/SPP
 - d. Biaya yudisium
3. Mendapatkan pembebasan biaya pendidikan sesuai jangka waktu pemberian bantuan
4. Mendapatkan biaya hidup sekecil kecilnya Rp650.000,00 (enam ratus lima puluh ribu rupiah) per bulan yang akan dibayarkan 6 (enam) bulan sekali
5. Mendapatkan pembinaan dan fasilitasi dari perguruan tinggi pengelola untuk menunjang kegiatan akademik dan kemahasiswaan untuk mewujudkan misi program.

2. Kewajiban

- a. Menjunjung tinggi negara kesatuan Republik Indonesia dengan dasar negara Pancasila dan UUD 1945.
- b. Memenuhi kontrak hasil Bidikmisi dengan Perguruan Tinggi Penyelenggara, termasuk namun tidak terbatas pada kewajiban akademis dan administratif.
- c. Berperan aktif dan berkontribusi dalam pelaksanaan Tridarma Perguruan Tinggi.

E. Mekanisme Pendaftaran Bidikmisi

1. Pendaftaran Daring (*On line*)

Tata cara pendaftaran Bidikmisi melalui SNMPTN, SBMPTN, PMDK Politeknik atau Seleksi Mandiri Perguruan Tinggi secara *online* pada laman Bidikmisi (<http://bidikmisi.belmawa.ristekdikti.go.id/>) adalah sebagai berikut.

- a. Tahapan pendaftaran Bidikmisi
 1. Sekolah mendaftarkan diri sebagai institusi pemberi rekomendasi ke laman Bidikmisi dengan melampirkan hasil pindaian (*scan*) (Lampiran 1 bagian persetujuan dan tanda tangan) untuk mendapatkan nomor Kode Akses Sekolah
 2. Direktorat Jenderal Pembelajaran dan Kemahasiswaan memverifikasi pendaftaran dalam kurun waktu 1 x 24 jam pada hari dan jam kerja
 3. Sekolah merekomendasikan masing-masing siswa melalui laman Bidikmisi menggunakan kombinasi NPSN dan kode akses yang telah diverifikasi. Sekolah memberikan nomor pendaftaran dan kode akses kepada masing-masing siswa yang sudah direkomendasikan
 4. Siswa mendaftar melalui laman Bidikmisi dan menyelesaikan semua tahapan yang diminta di dalam sistem pendaftaran.
- b. Siswa yang sudah menyelesaikan pendaftaran Bidikmisi mendaftar seleksi nasional atau mandiri yang telah diperoleh sesuai ketentuan masing-masing pola seleksi melalui alamat berikut:
 1. SNMPTN melalui <http://www.snmptn.ac.id>
 2. SBMPTN melalui <http://www.sbmptn.ac.id>
 3. PMDK Politeknik melalui <http://pmdk.politeknik.or.id>
 4. Seleksi Mandiri PTN sesuai ketentuan masing-masing PTN
 5. Seleksi Mandiri PTS sesuai ketentuan masing masing PTS.
 6. Siswa yang mendaftar dan ditentukan lolos melalui seleksi masuk, melengkapi berkas, dan berkas dibawa pada saat pendaftaran ulang, yaitu:

- a. Kartu peserta dan formulir pendaftaran program Bidikmisi yang dicetak dari laman Bidikmisi
- b. Kartu Indonesia Pintar (KIP), atau bantuan pemerintah sejenis lainnya (jika ada)
- c. Siswa yang belum memenuhi syarat butir (b) di atas, harus membawa Surat Keterangan Penghasilan Orang Tua/Wali atau Surat Keterangan Tidak Mampu yang dapat dibuktikan kebenarannya, yang dikeluarkan oleh Kepala Desa/Kepala Dusun/Instansi tempat orang tua bekerja/tokoh masyarakat
- d. Fotokopi Kartu Keluarga atau Surat Keterangan tentang Susunan Keluarga
- e. Fotokopi rekening listrik bulan terakhir (apabila tersedia aliran listrik) dan/ atau bukti pembayaran PBB (apabila mempunyai bukti pembayaran) dari orang tua/wali-nya
- f. Berkas pendukung lainnya yang diminta oleh perguruan tinggi dan Kopertis.

F. Mekanisme Penetapan

Bagi calon mahasiswa penerima Bidikmisi yang telah dinyatakan diterima di Perguruan Tinggi, akan dilakukan hal-hal sebagai berikut:

1. Verifikasi kelayakan penerima Bidikmisi oleh perguruan tinggi dan Kopertis;
2. Penetapan mahasiswa penerima Bidikmisi oleh perguruan tinggi dan Kopertis.

G. Data Mining

Data mining berisi pencarian *trend* atau pola yang diinginkan dalam database yang besar untuk membantu pengambilan keputusan diwaktu yang akan datang. Harapannya, perangkat data mining mampu mengenali pola-pola ini dalam data dengan masukan yang minimal. Pola-pola ini dikenali oleh perangkat tertentu yang dapat memberikan suatu analisis data yang

berguna dan berwawasan yang kemudian dapat dipelajari lebih teliti, yang mungkin saja menggunakan perangkat pendukung keputusan yang lainnya (Alith, 2016). Definisi sederhana dari data mining adalah ekstraksi informasi atau pola yang penting atau menarik dari data yang ada di *database* yang besar. Data mining adalah penambangan atau penemuan informasi baru dengan mencari pola atau aturan tertentu dari sejumlah data yang sangat besar. Hasil dari pengolahan data dengan menggunakan metode data mining ini dapat digunakan untuk mengambil keputusan di masa depan.

Berdasarkan penjelasan dari beberapa penelitian yang telah dilakukan sebelumnya, maka dapat disimpulkan bahwa hasil dari proses *data mining* haruslah informasi yang baru dan bermanfaat untuk keperluan kedepannya serta mudah dimengerti. Bidikmisi merupakan program pemerintah untuk memberikan akses pendidikan tinggi kepada masyarakat miskin untuk dapat memutus mata rantai kemiskinan. Sampai saat ini jumlah penerima Bidikmisi sudah mencapai angka 432.409 mahasiswa, sehingga berkontribusi untuk meningkatkan Angka Partisipasi Kasar (APK) Pendidikan Tinggi. Bidikmisi juga memiliki skema yang berbeda dengan bantuan biaya pendidikan lain, dengan filosofinya untuk menjemput penerima.

Bidikmisi memberikan jaminan pembiayaan mulai dari pendaftaran sampai penerima Bidikmisi menuntaskan pendidikan tinggi. Bidikmisi adalah bantuan biaya pendidikan dari Kementerian Riset Teknologi dan Pendidikan Tinggi Republik Indonesia yang memberikan fasilitas pembebasan biaya pendidikan dan subsidi biaya hidup. Bidikmisi diberikan kepada penerima selama 8 (delapan) semester untuk S1 / D4, 6 (enam) semester untuk D3, 4 (empat) semester untuk D2, dan 2 (dua) semester untuk D1. Besaran subsidi biaya hidup yang diberikan serendah-rendahnya Rp650.000,00 per bulan diberikan setiap 6 bulan. Adapun pembebasan biaya pendidikan mencakup semua biaya yang dibayarkan ke Perguruan Tinggi untuk kepentingan pendidikan.

H. Data Numerik dan Data Kategorik

1. Numerik

Nilai numerik termasuk nilai real (pecahan) dan integer (bilangan bulat). Fitur dengan nilai numerik memiliki 2 properti penting, yaitu: setiap nilai memiliki urutan dan memiliki relasi jarak.

2. Kategorik

Dinyatakan dengan sama dengan atau tidak sama dengan, variabel kategori yang memiliki 2 nilai dapat dikonversi menjadi variabel numerik dengan 2 nilai values (0 atau 1). Variabel pengkodean dengan N buah nilai dapat dikonversikan ke dalam N buah variabel bertipe numerik yang memiliki nilai biner untuk setiap kategorikal. Pengkodean ini disebut "*dummy variabels*".

BAB III

ANALISIS DATA MINING

A. Analisis Kelompok

Analisis kelompok merupakan salah satu teknik data mining yang bertujuan untuk mengidentifikasi sekelompok obyek yang mempunyai kemiripan karakteristik tertentu yang dapat dipisahkan dengan kelompok obyek lainnya, sehingga obyek yang berada dalam kelompok yang sama relatif lebih homogen daripada obyek yang berada pada kelompok yang berbeda. Jumlah kelompok yang dapat diidentifikasi tergantung pada banyak dan variasi data obyek.

Analisis ini mempunyai tujuan utama untuk mengelompokkan objek-objek pengamatan menjadi beberapa kelompok berdasarkan karakteristik yang dimilikinya. Analisis kelompok mengelompokkan objek-objek sehingga setiap objek yang paling dekat kesamaannya dengan objek lain berada dalam kelompok yang sama, serta mempunyai kemiripan satu dengan yang lain. (Johnson & Wichern, 2007). Beberapa manfaat analisis kelompok adalah eksplorasi data variabel ganda, reduksi data, dan prediksi keadaan objek. Hasil analisis kelompok dipengaruhi oleh objek yang dikelompokkan, variabel yang diamati, ukuran kemiripan atau ketakmiripan yang digunakan, skala ukuran yang digunakan, serta metode pengelompokan yang digunakan.

Ukuran kemiripan dan ketidakmiripan merupakan hal yang sangat mendasar dalam kelompok analisis. Algoritma pengelompokan menggunakan ukuran kemiripan atau ketidakmiripan digunakan untuk menggabungkan atau memisahkan data objek dari suatu data. Ukuran kemiripan biasanya digunakan oleh algoritma pengelompokan untuk menganalisis data kategori, sedangkan ukuran ketidakmiripan digunakan oleh algoritma pengelompokan untuk menganalisis data numerik. Ukuran ketakmiripan antara objek ke- i dengan objek ke- j (d_{ij}) merupakan fungsi yang memiliki sifat-sifat sebagai berikut $d_{ij} \geq 0, d_{ii} = 0, d_{ij} = d_{ji}$, dan $d_{ik} + d_{jk} \geq d_{ij}$, untuk setiap i, j dan k . Semakin besar nilai ukuran ketakmiripan antara

dua objek maka semakin besar pula perbedaan antara kedua objek tersebut, sehingga makin cenderung untuk tidak berada dalam kelompok yang sama. (Johnson & Wichern, 2007).

Ukuran kemiripan dan ketakmiripan pada umumnya diukur berdasarkan jarak. Salah satu faktor yang sangat berpengaruh terhadap hasil dari kelompok yang dibentuk adalah jarak antar objek pengamatan (Sharma, 1996). Oleh karena itu, dibutuhkan suatu alat ukur untuk menentukan jarak antar objek pengamatan. Berikut ini merupakan metode-metode pengukuran jarak antara objek ke- i (x_i) dengan objek ke- j (x_j) berdasarkan karakteristik variabel yang dikelompokkan.

1. Metode pengukuran jarak untuk variabel kategorik biner bila variabel yang diamati berupa variabel biner yang hanya memiliki dua macam karakter yang berbeda (0,1), maka variabel yang diamati dapat dibentuk suatu tabel kontingensi seperti ditunjukkan pada Tabel 1 Perhitungan ukuran jarak antara variabel x_i dan x_j untuk pengukuran data biner dapat menggunakan beberapa ukuran yang disajikan pada Tabel 2

Tabel 1 Tabel kontingensi data biner

Kategori x_i	Kategori x_j		Total
	1	0	
1	a	b	a+b
0	c	d	c+d
Total			

Tabel 2 Ukuran jarak data biner

Jenis	Rumus
Russel dan Rao	$RR(x_i, x_j) = \frac{a}{a+b+c+d}$
Simple matching	$SM(x_i, x_j) = \frac{a+d}{a+b+c+d}$
Jaccard	$JACCARD(x_i, x_j) = \frac{a}{a+b+c}$
Dice Czekanowski Sorensen	$DICE(x_i, x_j) = \frac{2a}{2a+b+c}$

2. Metode pengukuran jarak untuk variabel kategorik nominal
 Pada pengamatan dengan variabel nominal maka pengukuran memiliki konsep yang sama dengan *simple matching coefficient* maupun *dice*, dimana kategorinya dapat lebih dari dua macam. Dengan jumlah variabel sebanyak m , maka rumus untuk pengukuran jarak variabel nominal antara x_i dan x_j ditunjukkan pada Persamaan (2.1),

$$sim(x_i, x_j) = \frac{1}{m} \sum_{l=1}^m S_{ijl} \quad (2.1)$$

dimana $S_{ijl} = 1$ jika $x_{il} = x_{jl}$ dan $S_{ijl} = 0$ jika $x_{il} \neq x_{jl}$.

3. Metode pengukuran jarak untuk variabel kategorik ordinal

Pada pengamatan dengan variabel ordinal maka pengukuran memiliki konsep yang digunakan sama dengan metode untuk data numerik, dimana kategorinya dinyatakan sebagai suatu bilangan bulat. Salah satu metode yang dapat digunakan untuk variabel ordinal adalah jarak *manhattan*. Dengan jumlah variabel

sebanyak m , maka rumus untuk pengukuran jarak x_i dan x_j pada variabel nominal ditunjukkan pada persamaan

$$sim(x_i, x_j) = \sum_{l=1}^m |x_{il} - x_{jl}| \quad (2.2)$$

4. Metode pengukuran jarak untuk variabel numerik

Pada variabel yang memiliki jenis data numerik maka jarak yang dapat digunakan adalah jarak *euclidean*. Misalkan terdapat dua observasi dengan variabel-variabel berdimensi m

$$\text{yaitu } x_i = [x_{i1}, x_{i2}, \dots, x_{im}]^T \text{ dan } x_j = [x_{j1}, x_{j2}, \dots, x_{jm}]^T .$$

Konsep jarak *euclidean* yang mengukur jarak antara observasi x_i dan x_j adalah sebagai berikut,

$$d_{ij} = \sqrt{(x_i - x_j)^T (x_i - x_j)} \quad (2.3)$$

Terkait dengan pengertian dan tujuan dilakukannya analisis kelompok, dapat dinyatakan bahwa suatu kelompok (kelompok) yang baik adalah kelompok yang mempunyai. (Hair, Black, Babin, & Anderson, 2009).

- Homogenitas (kesamaan) yang tinggi antara anggota dalam satu kelompok (*within*-kelompok),
- Heterogenitas (perbedaan) yang tinggi antara kelompok yang satu dengan kelompok yang lain (*between* kelompok)

B. Metode *Fuzzy C-Means*

Fuzzy C-Means (FCM) adalah teknik pengelompokan data yang memungkinkan suatu data menjadi dua atau lebih kelompok. Fuzzy C-means cluster pertama kali dikemukakan oleh Dunn (1973) dan kemudian dikembangkan oleh Bezdek (1981) yang banyak digunakan dalam *pattern recognition*.

Metode ini merupakan pengembangan dari metode non hierarki K-means Cluster, karena pada awalnya ditentukan dulu jumlah kelompok atau cluster yang akan dibentuk. Kemudian dilakukan iterasi sampai mendapatkan keanggotaan kelompok tersebut. Metode ini adalah metode yang paling digemari karena merupakan metode yang paling robust ((Klawonn dan Höppner, 2001) dan (Klawonn, 2000)) dan memberikan hasil yang smooth (halus) dengan toleransi relatif (Shihab, 2000).

Suatu titik bisa memiliki keanggotaan parsial lebih banyak dari satu kelas Tidak boleh ada kelas kosong dan tidak ada kelas yang tidak berisi titik data FCM berusaha untuk menemukan karakteristik titik terbaik pada setiap kelompok, yang dapat dipertimbangkan sebagai "pusat" kelompok. Algoritma ini bekerja dengan menugaskan keanggotaan ke setiap titik data yang sesuai setiap pusat kelompok berdasarkan jarak antar pusat kelompok dan titik data. Algoritma *Fuzzy C-Means* didasarkan pada minimisasi fungsi objektif berikut (Velmurugan & Santhanam, 2010) adalah

$$J(U,V) = \sum_{i=1}^n \sum_{j=1}^c (\mu_{ij})^m \|x_i - v_j\|^2 \quad (2.4)$$

dimana, $\|x_i - v_j\|$ adalah jarak *Euclidean* antara data ke i^{th}

dan pusat kelompok j^{th} . Partisi *Fuzzy* dilakukan melalui iterasi optimalisasi fungsi objektif yang ditunjukkan di atas, dengan memperbarui keanggotaan μ_{ij} dan pusat kelompok v_j oleh:

$$\mu_{ij} = 1 / \sum_{k=1}^c (d_{ij} / d_{ik})^{(2/m-1)} \quad (2.5)$$

$$v_j = \left(\sum_{i=1}^n (\mu_{ij})^m x_i \right) / \left(\sum_{i=1}^n (\mu_{ij})^m \right), \forall j = 1, 2, \dots, c$$

(2.6)

dimana $\{\mu_1, \dots, \mu_n\}$ adalah partisi *Fuzzy c* dan $\{v_1, \dots, v_n\}$ adalah himpunan *sentroid* 'm' adalah sebuah bilangan riil yang lebih dari 1 yang menyatakan derajat kekaburan (*degree of Fuzzyness*), indeks $m \in [1, \infty]$, 'c' mewakili jumlah kelompok pusat, v_j mewakili pusat kelompok, μ_{ij} mewakili keanggotaan kelompok i^{th} ke kelompok j^{th} , d_{ij} mewakili jarak Euclidean antara data i^{th} dan pusat kelompok j^{th} .

C. Metode Fuzzy C-Modes

Pada sebagian besar algoritma pengelompokan *Fuzzy*, yang ditugaskan anggota data ke kelompok tidak jelas, tapi sentroid itu sendiri tidak *Fuzzy*. terdapat *Fuzzy K-Modes* Algoritma dimodifikasi sebagai algoritma *Fuzzy C-Modes* yang digunakan sentroid *Fuzzy* untuk mengelompokkan data kategorik. *Fuzzy sentroid* adalah seperangkat nilai *Fuzzy* yang mengandung nilai kategori setiap atribut. Untuk mengelompokkan data kategorik, diusulkan Algoritma *Fuzzy C-Modes* Algoritma memperluas algoritma *K-Modes* berdasarkan Prosedur tipe *C-Means Fuzzy*. Algoritma ini memperbarui pusat kelompok pada setiap iterasi dengan mengukur jarak antar masing-masing sentroid kelompok sentroid dan masing-masing objek. Misalkan $X = \{X_1, X_2, \dots, X_n\}$ menjadi himpunan n objek. Objek X_i diwakili sebagai $[x_{i1}, x_{i2}, \dots, x_{im}]$ dan $X_i = X_k$ jika $x_{i,j} = x_{k,j}$, $1 \leq j \leq m$. Algoritma *Fuzzy C-Modes* mengelompokkan data X ke dalam k kelompok dengan meminimalkan fungsi objektif

$$F(W, Z) = \sum_{l=1}^k \sum_{i=1}^n W_h^\alpha d(Z_l, X_i)$$

(2.7)

$$0 \leq W_h \leq 1; 1 \leq l \leq k; 1 \leq i \leq n, \quad \sum_{l=1}^k W_{il} = 1, \quad 1 \leq i \leq n, \quad \text{dan}$$

$$0 < \sum_{i=1}^n W_h < n, 1 \leq l \leq k. \quad \text{Sedangkan } *_{li} \text{ adalah derajat}$$

keanggotaan data X_i ke l^{th} kelompok, dan merupakan elemen matriks partisi $(k \times n)$.

$W = [*_{li}]$. $C^* = [C^*_1, C^*_2, \dots, C^*_l, \dots, C^*_k]$ dan C^*_l adalah pusat

kelompok l^{th} dan parameter α mengontrol kekaburan dari tiap anggota objek. Algoritma *Fuzzy C-Modes* menggunakan ukuran ketidakmiripan yang sederhana untuk objek yang kategorik, anggap X dan Y adalah dua objek yang diwakili oleh

$$[x_1, x_2, \dots, x_m] \text{ dan } [y_1, y_2, \dots, y_m] \text{ maka } d(X, Y) = \sum_{j=1}^m \delta(x_j, y_j)$$

$$\text{dimana } \delta(x_j, y_j) = \begin{cases} 0, & x_j = y_j \\ 1, & x_j \neq y_j \end{cases}. \text{ Ukuran } \delta \text{ memenuhi ruang}$$

metrik objek-objek kategorik.

D. Pengelompokan Ensemble

Pengelompokan ensemble merupakan metode yang menggabungkan beberapa algoritma yang berbeda untuk mendapatkan partisi umum dari data, yang bertujuan untuk konsolidasi hasil pengelompokan individu (Suguna & Selvi, 2012). Tujuan pengelompokan ensemble adalah untuk menggabungkan hasil pengelompokan dari beberapa algoritma pengelompokan untuk mendapatkan hasil pengelompokan yang lebih baik dan *robust* (Yoon, Ahn, Lee, Cho, & Kim, 2006). Pengelompokan ensemble terdiri atas dua tahap algoritma. Tahap pertama adalah melakukan pengelompokan dengan beberapa algoritma dan menyimpan hasil pengelompokan