TESIS – SS142501

ENSEMBLE FUZZY, K-PROTOTYPES & DENSITY PEAKS CLUSTERING MIXED PADA PENGELOMPOKAN DATA PELAMAR BIDIKMISI SEJAWA-TIMUR TAHUN 2016-2017

Laila Qadrini
NRP. 06211650010043

DOSEN PEMBIMBING
Dr. Kartika Fithriasari, M.Si
Prof. Drs. Nur Iriawan, M.Ikomp, Ph.D

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

THESIS – SS142501

ENSEMBLE FUZZY, K-PROTOTYPES & DENSITY PEAKS CLUSTERING MIXED FOR CLUSTERING THE APPLICANT OF BIDIKMISI ON EAST JAVA IN 2016-2017

Laila Qadrini
NRP. 06211650010043

SUPERVISOR
Dr. Kartika Fithriasari, M.Si
Prof. Drs. Nur Iriawan, M.Ikomp, Ph.D

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

***ENSEMBLE FUZZY, K-PROTOTYPES, & DENSITY PEAKS CLUSTERING MIXED* PADA PENGELOMPOKAN DATA PELAMAR BIDIKMISI SEJAWA-TIMUR TAHUN 2016-2017**

Tesis disusun untuk memenuhi salah satu syarat memperoleh gelar
Master Sains (M.Si)
di
Institut Teknologi Sepuluh Nopember

oleh :

LAILA QADRINI
NRP. 06211650010043

Tanggal Ujian : 11 Juli 2018
Periode Wisuda : September 2018

Disetujui oleh:

1. **Dr. Kartika Fithriasari, M.Si.** **(Pembimbing I)**
   NIP: 19691212 199303 2 002
2. **Prof. Drs. Nur Iriawan, MIKomp., Ph.D.** **(Pembimbing II)**
   NIP: 19621015 198803 1 002
3. **Irhamah, M.Si., Ph.D.** **(Penguji)**
   NIP: 19780406 200112 2 002
4. **Dr. Vita Ratnasari, S.Si., M.Si** **(Penguji)**
   NIP: 19700910 199702 1 001

Dekan
Fakultas Matematika, Komputasi dan Sains Data
Institut Teknologi Sepuluh Nopember

Prof. Dr. Basuki Widodo, M.Sc.
NIP: 19650605 198903 1 002

ENSEMBLE FUZZY, K-PROTOTYPES & DENSITY PEAKS CLUSTERING MIXED) PADA PENGELOMPOKAN DATA PELAMAR BIDIKMISI SEJAWA-TIMUR TAHUN 2016-2017

Nama Mahasiswa	**: Laila Qadrini**
NRP	**: 06211650010043**
Pembimbing	**: Dr. Kartika Fitriasari, M.Si**
Co-Pembimbing	**: Prof. Drs. Nur Iriawan, M.Ikom, Ph.D**

ABSTRAK

Metode Pengelompokan pada data mining berbeda dengan metode konvensional yang biasa digunakan untuk pengelompokkan. 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 kategorik. Permasalahan yang sering ditemui dalam analisis pengelompokan adalah data yang berskala campuran numerik dan kategorik. Penelitian ini bertujuan untuk membandingkan hasil pengelompokan dari. Ensembel *Fuzzy*, K-*Prototypes* dan DPC-M. Ketiga Algoritma ini diterapkan untuk mengelompokkan pelamar beasiswa Bidikmisi di Jawa Timur selama tahun 2016-2017. Secara umum, validasi pengelompokan dapat dikategorikan ke dalam tiga kelas, yaitu validasi pengelompokan internal, validasi pengelompokan eksternal, dan validasi relatif. Pada penelitian ini kita fokus pada indeks validitas internal dan eksternal kelompok, berdasarkan hasil penelitian menunjukkan bahwa, secara keseluruhan, Algoritma Ensembel *Fuzzy* memiliki hasil pengelompokan yang lebih baik daripada Algoritma K-*Prototypes* dan DPC-M.

Kata Kunci :*Ensemble Fuzzy*, K-*Prototypes*, *Density Peaks Clustering Mixed* (DPC-M)

“*Halaman ini sengaja dikosongkan*”

ENSEMBLE FUZZY, K-PROTOTYPES & DENSITY PEAKS CLUSTERING MIXED) FOR CLUSTERING THE APPLICANT OF BIDIKMISI ON EAST JAVA IN 2016-2017

Name	**: Laila Qadrini**
Student Identity Number	**: 06211650010043**
Supervisor	**: Dr. Kartika Fitriasari, M.Si**
Co Supervisor	**: Prof. Drs. Nur Iriawan, M.Ikomp, Ph.D**

ABSTRACT

The Clustering method in data mining differs from the conventional method commonly used for clustering. The difference is that data mining has a high data dimension that can consist of tens of thousands or millions of records with tens or hundreds of attributes. In addition to data mining data can consist of mixed data types such as numerical and categorical data. The problems that are often encountered in clustering analysis are numerical and categorical mixed data. This study aims to compare the results of clustering from. Fuzzy Ensembles, K-Prototypes and DPC-M. These three algorithms are applied to classify Bidikmisi scholarship applicants in East Java during 2016-2017. In general, clustering validation can be categorized into three classes, which are internal clustering validation, external clustering validation, and relative validation. In this study we focus on internal and external group validity indexes, based on the results of the research indicating that, overall, The Fuzzy Ensemble Algorithm has better clustering results than K-Prototypes Algorithm and DPC-M.

Keywords :Ensemble Fuzzy, K-Prototypes, Density Peaks Clustering Mixed (DPC-M)

“*Halaman ini sengaja dikosongkan*”

KATA PENGANTAR

Syukur Alhamdulillah penulis panjatkan kehadirat Allah SWT yang maha menguasai segala ilmu dan alam. Atas rahmat, ridho dan hidayah-Nya sehingga pengerjaan serta penulisan Tesis dengan judul "*Ensemble Fuzzy*, *K-Prototypes* & *Density Peaks Clustering Mixed* pada Pengelompokan Data Pelamar Bidikmisi Sejawa-Timur Tahun 2016-2017" dapat terselesaikan dengan baik dan lancar.

Penulisan Tesis ini adalah salah satu syarat yang harus dipenuhi dalam memperoleh gelar Magister sesuai dengan kurikulum Departemen Statistika FMKSD-ITS Surabaya. Dalam penyelesaian Tesis serta laporan ini penulis tidak terlepas dari bantuan serta dukungan dari berbagai pihak. Oleh karena itu penulis ingin mengucapkan terima kasih sebesar-besarnya kepada:

1. Suamiku Muhammad Syahrir Fanzuri, Ibuku Hj. Hawara, Bapak mertua dan ibu mertuaku serta keluarga besar penulis atas segala doa, dukungan materi, motivasi, kepercayaan dan rasa kasih sayang.
2. Bapak Dr. Suhartono, M.Sc. selaku Ketua Departemen Statistika ITS yang telah banyak memberikan inspirasi kepada mahasiswa untuk senantiasa berkarya.
3. Ibu Dr. Kartika Fithriasari, M.Si dan Bapak Prof. Drs. Nur Iriawan, M.Ikomp, Ph.D selaku dosen pembimbing yang dengan sabar memberikan bimbingan, arahan, dan masukan selama pengerjaan Tesis.
4. Ibu Irhamah, M.Si., Ph.D dan Dr. Vita Ratnasari, S.Si, M.Si selaku dosen penguji yang telah memberikan banyak tambahan ilmu selama proses perbaikan laporan Tesis.
5. Bapak Dr. rer.pol. Heri Kuswanto, M.Si selaku Ketua Program Studi Pascasarjana Statistika ITS yang memberikan motivasi dalam pendidikan.
6. Seluruh dosen pengajar serta karyawan di departemen Statistika ITS, yang telah memberikan bantuan dan ilmunya selama masa perkuliahan dan bekal dalam pengerjaan Tesis.

7. Teman-teman S2 Statistika ITS angkatan 2016, khususnya teman seperjuangan Nita, Sinta, Ai, Nurma, Arip, Ihsan, Nendy, Febri, Ghazali, Mawanda, dan teman-teman lainnya yang tidak bisa diketik satu-persatu yang telah membantu dalam penyelesaian laporan.
8. Pihak-pihak lain yang telah mendukung dan membantu dalam penyusunan Tesis ini yang tidak mungkin penulis sebutkan satu per satu. Terima kasih.

Penulis menyadari bahwa penyusunan Tesis ini masih jauh dari sempurna, maka kritik dan saran yang membangun akan senantiasa penulis harapkan demi kesempurnaan di masa mendatang. Semoga laporan ini dapat memberikan sumbangan yang bermanfaat bagi semua pihak.

Surabaya, Juli 2018

Penulis

DAFTAR ISI

Halaman

“*Halaman ini sengaja dikosongkan*”

DAFTAR TABEL

Halaman

"*Halaman ini sengaja dikosongkan*"

DAFTAR GAMBAR

Halaman

“*Halaman ini sengaja dikosongkan*”

DAFTAR LAMPIRAN

Halaman

“*Halaman ini sengaja dikosongkan*”

BAB I
PENDAHULUAN

1.1 Latar Belakang

Pendidikan menjadi garda terdepan pada pengembangan sumber daya manusia dan menjadi momentum untuk melakukan perbaikan di sektor pendidikan Indonesia agar mampu menghasilkan sumber daya manusia yang memiliki daya saing tinggi. Pemerintah dalam rangka meningkatkan produktivitas dan daya saing masyarakat di pasar internasional membuat Program Bantuan Biaya Pendidikan Bidik-Misi (Iriawan, Fithriasari, Ulama, Suryaningtyas, Susanto, dan Pravitasari, 2015). Program Bidikmisi merupakan salah satu program unggulan pemerintah yang bertujuan untuk meningkatkan akses dan kesempatan belajar di Perguruan Tinggi bagi peserta didik yang tidak mampu secara ekonomi dan memiliki prestasi akademik yang baik. Sampai dengan tahun 2017 ini tercatat lebih dari 432.409 mahasiswa yang telah memperoleh bantuan biaya pendidikan Bidikmisi, dari jumlah tersebut sebanyak 145.000 telah menyelesaikan pendidikannya. Jumlah peminat Program Bidikmisi menunjukkan peningkatan yang sangat signifikan dari tahun ke tahun, untuk tahun 2017 tercatat sebanyak 520.688 pelamar tetapi hanya sekitar 80.000 saja yang bisa diakomodir karena keterbatasan anggaran pemerintah. Secara umum pelaksanaan Program Bidikmisi telah berjalan dengan baik, sehingga mampu meningkatkan akses dan kesempatan belajar di Perguruan Tinggi bagi peserta didik yang kurang mampu secara ekonomi akan tetapi mempunyai potensi akademik yang baik. Dari segi prestasi, para mahasiswa Bidikmisi juga menunjukkan kemampuan akademik yang luar biasa dengan capaian IPK lebih dari 87% mahasiswa Bidikmisi memperoleh IPK di atas antara 3,0.

Dengan demikian para mahasiswa Bidikmisi turut berkontribusi terhadap peningkatan mutu pendidikan disetiap perguruan tinggi. Sementara itu hasil tracer study terhadap lulusan Bidikmisi di beberapa perguruan tinggi didapatkan gambaran profil lulusan Bidikmisi seperti berikut: Guru yang masih dalam proses mengikuti PPG 39%, Pegawai Negeri/Swasta/BUMN 26%,

Wirausaha 29% dan studi lanjut ke jenjang pascasarjana di dalam dan di luar negeri 6%. Dari hasil *tracer study* tersebut terlihat lulusan Bidikmisi yang berprofesi sebagai wirausaha cukup besar, hal ini tentu sangat menggembirakan karena untuk meningkatkan daya saing kita masih perlu meningkatkan jumlah wirausaha, yang saat ini jumlahnya baru sekitar 3,1% dari populasi penduduk, sementara negara tetangga kita Malaysia sudah mencapai angka 6% dan Singapura 7%. Proses pemilihan penerima Bidikmisi agar tepat sasaran dan transparan dalam proses penyeleksiannya diperlukan sebuah sistem yang mampu mengetahui siapa saja mahasiswa yang pantas mendapatkan Bidikmisi, dalam sistem tersebut diperlukan sebuah proses data mining yang digunakan untuk melakukan perhitungan nilai-nilai kriteria yang dimiliki oleh mahasiswa. Jumlah mahasiswa aktif yang tercatat sekitar 1736 mahasiswa pada tahun 2016-2017, sistem pemilihan penerimaan Bidikmisi menggunakan konsep data mining membantu dalam pemilihan mahasiswa penerima Bidikmisi. Selain itu hal tersebut juga mampu memberikan kemudahan kepada pihak pemberi Bidikmisi dalam menentukan penerima Bidikmisi yang sesuai dengan ketentuannya dan memberikan rasa transparansi kepada masyarakat terhadap proses pemilihan penerimaan Bidikmisi dengan menuangkan sistem tersebut dalam bentuk perangkat lunak. Sehingga teridentifikasi permasalahan yang timbul dikarenakan pemberian Bidikmisi yang tidak tepat sasaran disebabkan oleh belum adanya sistem verifikasi Bidikmisi yang optimal sehingga mengakibatkan pelamar dengan prestasi terbaik belum tentu terpilih sebagai penerima Bidikmisi.

Metode pengelompokan dalam data mining berbeda dengan metode konvensional yang biasa digunakan untuk pengelompokkan. 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 ensembel untuk data campuran oleh He, Xu dan Deng (2005). Pengelompokan ensembel (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 ensembel untuk mengelompokkan penyakit hati dan persetujuan pengajuan kartu kredit dengan Kelompok *Ensemble Based Mixed Data Clustering* (CEBMDC) dan mengombinasikan tahapan ensembel 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 ensembel ROCK dan ensembel SWFM. Kedua metode digunakan pada studi kasus pengelompokkan aksesi 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. Hasil tersebut menunjukkan bahwa metode ensembel ROCK memberikan hasil pengelompokan lebih baik daripada metode ensembel 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* dan K-*Modes*) dan algoritma pengelompokan *Fuzzy*

(*Fuzzy* C-*Means* dan *Fuzzy* C-*Modes*). Kelompok yang dihasilkan dikombinasikan dengan metode pengelompokan ensembel dan dievaluasi dengan ukuran f-*measure* dan *entropy*. Ditemukan bahwa pembagian data lebih menguntungkan dan penerapan 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. 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 K-*Prototype*.

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 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).

1.2 Rumusan masalah

Berdasarkan latar belakang yang telah dipaparkan, permasalahan penelitian ini adalah sebagai berikut&

1. Bagaimana penerapan algoritma pengelompokan ensembel *Fuzzy*, K-*Prototypes* dan *Density Peaks Clustering Mixed* (DPC-M) pada data dengan variabel berskala campuran numerik dan kategorik ?
2. Bagaimana perbandingan hasil metode ensembel *Fuzzy*, *Krototypes* dan *Density Peaks Clustering Mixed* (DPC-M) berdasarkan indeks validitas internal kelompok dan indeks validitas eksternal kelompok pada data pelamar Bidikmisi seJawa-Timur tahun 2016-2017 ?

1.3 Tujuan Penelitian

Untuk menjawab rumusan permasalahan yang telah dipaparkan, maka tujuan yang ingin dicapai dari penelitian ini adalah sebagai berikut,

1. Menggunakan Analisis Kelompok dengan metode Ensembel *Fuzzy*, *Krototypes* dan *Density Peaks Clustering Mixed* (DPC-M) pada data pelamar Bidikmisi Sejawa-timur 2016-2017
2. Membandingkan hasil pengelompokan metode Ensembel *Fuzzy*, *Krototypes* dan *Density Peaks Clustering Mixed* (DPC-M) pada data pelamar Bidikmisi seJawa-Timur tahun 2016-2017 berdasarkan ukuran indeks validitas internal dan eksternal hasil pengelompokan

1.4 Manfaat penelitian

Penelitian ini diharapkan bermanfaat bagi,

1. Manfaat bagi Perguruan Tinggi Negeri/Swasta

   Hasil penelitian ini dapat digunakan untuk melakukan penyeleksian pelamar Bidikmisi yang lebih efisien.
2. Manfaat bagi ilmu pengetahuan

   Penelitian ini diharapkan dapat mengembangkan wawasan keilmuan data mining terutama dalam melakukan pengelompokan menggunakan metode

Ensembel *Fuzzy*, K-*Prototypes* dan *Density Peaks Clustering Mixed* (DPC--M) pada data berskala campuran numerik dan kategorik.

1.5 Batasan Masalah

Penelitian ini membatasi bahwa pengelompokan pelamar Bidikmisi dilakukan berdasarkan beberapa variabel yang ada pada *User Manual* Bidikmisi. Metode-metode yang digunakan untuk pengelompokan pelamar Bidikmisi adalah metode ensemble *Fuzzy*, K-*Prototypes* dan *Density Peaks Clustering Mixed* (DPC-M). Kriteria perbandingan hasil antar metode menggunakan Indeks validitas internal yaitu *Sum Square Within*, rata-rata koefisien *Silhouette*, dan indeks *Dunn*. Dan uji validitas eksternal yaitu *Purity* dan *Entropy*.

BAB II
TINJAUAN PUSTAKA

Pada bab ini dijelaskan mengenai beberapa kajian teori yang digunakan dalam melakukan analisis pengelompokan untuk menyelesaikan studi kasus yang merupakan data mining berskala campuran numerik dan kategorik pada data pelamar Bidikmisi se-Jawa timur tahun 2016-2017.

2.1 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 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.

2.2 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.

2.3 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

2.4 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.

2.5 Hak dan Kewajiban Penerima Bidikmisi

Hak dan kewajiban penerima Bidikmisi adalah sebagai berikut.

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.

Kewajiban

1. Menjunjung tinggi negara kesatuan Republik Indonesia dengan dasar negara

Pancasila dan UUD 1945.

2. Memenuhi kontrak hasil Bidikmisi dengan Perguruan Tinggi Penyelenggara,
   termasuk namun tidak terbatas pada kewajiban akademis dan administratif.
3. Berperan aktif dan berkontribusi dalam pelaksanaan Tridarma Perguruan Tinggi.

2.6 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*) (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.

2.7 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.

2.8 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*”.

2.9 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* $\left(d_{ij}\right)$ 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* dengan objek ke-*j*, berdasarkan karakteristik variabel yang dikelompokkan.

a. 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 2.1.

Tabel 2.1 Tabel Kontingensi Data Biner

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

adapun perhitungan ukuran jarak antara variabel x_i dan x_j untuk pengukuran data biner dapat menggunakan beberapa ukuran yang disajikan pada Tabel 2.2.

Tabel 2.2 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	$DICE(x_i, x_j) = \frac{2a}{2a+b+c}$

b. Metode pengukuran jarak untuk variabel kategorik nominal

Pada pengamatan dengan variabel nominal maka pengukuran memiliki konsep yang sama dengan *simple matcing 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),

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

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

c. 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

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

d. 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* yaitu $x_i = \left[x_1, x_2, ..., x_m\right]^T$ dan $x_j = \left[x_1, x_2, ..., x_m\right]^T$. Konsep jarak *euclidean* yang mengukur jarak antara observasi x_i dan x_j adalah sebagai berikut,

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

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).

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

2.10 Metode *Fuzzy C-Means*

Fuzzy C-Means (FCM) adalah teknik pengelompokkan data yang memungkinkan suatu data menjadi dua atau lebih kelompok. 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

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

dimana, $\left\|x_i - v_j\right\|$ adalah jarak *Euclidean* antara data ke *i* dan pusat kelompok ke *j*. Partisi *Fuzzy* dilakukan melalui iterasi optimalisasi fungsi objektif yang ditunjukkan di atas, dengan memperbarui keanggotaan μ_{ij} dan pusat kelompok v_j oleh:

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

$$v_j = \left(\sum_{i=1}^{n}(\mu_{ij})^m x_i\right)/\left(\sum_{i=1}^{n}(\mu_{ij})^m\right), \forall j = 1,2,...,c \quad (2.6)$$

dimana $\{\mu_1,...,\mu_n\}$ adalah partisi *Fuzzy c* dan $\{v_1,...,v_n\}$ adalah himpunan *centroid* '*m*' adalah sebuah bilangan riil yang lebih dari 1 yang menyatakan derajat kekaburan (*degree of Fuzzyness*), indeks m ∈ [1, ∞], '*c*' mewakili jumlah kelompok pusat, v_j mewakili pusat kelompok, μ_{ij} mewakili keanggotaan kelompok ke *i* ke kelompok *j* dan d_{ij} mewakili jarak *Euclidean* antara data ke *i* dan pusat kelompok ke *j*. dalam metode pengelompokan *Fuzzy* suatu obyek dapat menjadi anggota lebih dari satu cluster (beberapa cluster) secara bersamaan tetapi dengan derajat keanggotaan yang berbeda. Metode *fuzzy* ini lebih natural dibandingkan dengan metode *hard clustering* (Purnomo & Iriawan, 2012).

2.11 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 centroid* 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 *Fuzzy C-Means*. Algoritma ini memperbarui pusat kelompok pada setiap iterasi dengan mengukur jarak antar masing-masing sentroid kelompok dan masing-masing objek. Misalkan $X = \{X_1, X_2, ..., X_n\}$ menjadi himpunan *n* objek. Objek X_i diwakili sebagai $[x_{i1}, x_{i2}, ..., x_{im}]$ dan $X_i = X_k$ jika $x_{i,j} = x_{k,j}, 1 \le j \le m.$ Algoritma *Fuzzy C-Modes* mengelompokkan data *X* ke dalam *k* kelompok dengan meminimalkan fungsi objektif berikut:

$$F(W, Z) = \sum_{l=1}^{k} \sum_{i=1}^{n} W_h^{\alpha} d\left(Z_1, X_i\right) \tag{2.7}$$

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

Sedangkan $*_{li}$ adalah derajat keanggotaan data X_i ke *l* kelompok, dan merupakan elemen matriks partisi $(k \times n)$. $W = [*_{li}].C* = [C*_1, C*_2, ..., C*_l, ..., C*_k]$ dan $C*_l$ adalah pusat kelompok ke *l* 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 $[x_1, x_2, ..., x_m]$ dan $[y_1, y_2, ..., y_m]$ maka $d(X, Y) = \sum_{j=1}^{m} \delta\left(x_j, y_j\right)$ dimana

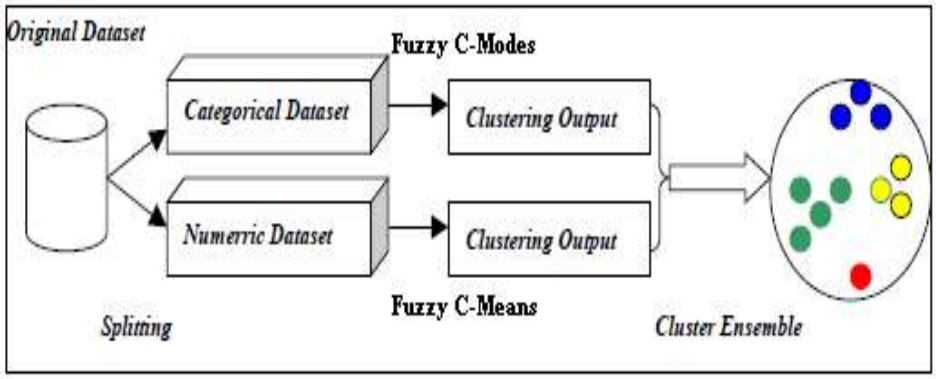
$\delta\left(x_j, y_j\right) = \begin{cases} 0, & x_j = y_j \\ 1, & x_j \ne y_j \end{cases}$. Ukuran δ memenuhi ruang metrik objek-objek kategorik.

2.12 Pengelompokan Ensembel

Pengelompokan ensembel 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 ensembel 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 ensembel terdiri atas dua tahap algoritma. Tahap pertama adalah melakukan pengelompokan dengan beberapa algoritma dan menyimpan hasil pengelompokan tersebut. Kedua, menggunakan fungsi konsensus untuk menentukan *final* kelompok dari kelompok-kelompok hasil tahap pertama. Langkah-langkah dalam analisis data campuran menggunakan metode pengelompokan ensembel yang disebut Algoritma CEBMDC memiliki tahapan sebagai berikut, (He, Xu, dan Deng, 2005b)

a. Membagi data menjadi dua subdata, yaitu murni numerik dan murni kategorik.
b. Melakukan pengelompokan objek dengan variabel numerik dengan algoritma pengelompokan data numerik, serta melakukan pengelompokan objek dengan variabel kategorik dengan algoritma pengelompokan data kategorik.
c. Menggabungkan (*combining*) hasil pengelompokan dari variabel numerik dan kategorik, yang disebut proses ensembel.
d. Melakukan pengelompokan ensembel menggunakan algoritma pengelompokan data kategorik untuk mendapatkan kelompok akhir (*final* kelompok).



Gambar 2.1 *Overview* Pengelompokan Ensembel

2.13 Metode K-*Prototypes*

K-*Prototypes* adalah metode pengelompokan pada data-data dengan tipe campuran numerik dan kategorikal. Algoritma ini dipilih karena sangat sederhana dari sisi kompleksitas algoritma dan mampu menangani data dengan ukuran yang sangat besar. K-*Prototype* adalah salah satu metode Pengelompokan yang berbasis parti. Algoritma ini adalah hasil pengembangan atau kombinasi antara K-*Means* dan K-*Modes* Pengelompokan untuk menangani Pengelompokan pada data dengan campuran atribut bertipe numerik dan kategorik (Huang, 1997). *K-Means Cluster Analysis* merupakan salah satu metode analisis klaster non hirarki yang dapat digunakan untuk mempartisi objek ke dalam kelompok-kelompok berdasarkan kedekatan karakteristik, sehingga objek yang mempunyai karakteristik yang sama dikelompokan dalam satu klaster yang sama dan objek yang mempunyai karakteristik yang berbeda dikelompokan ke dalam klaster yang lain (Marina dan Kartika, 2015). Jika perubahan sentroid pada K-*Means Cluster Analysis* menggunakan rata-rata, maka K-*Modes Cluster Analysis* pada perubahan sentroidnya menggunakan modus. K-*Modes Cluster Analysis* juga sama tahapan analisisnya dengan K-*Means Cluster Analysis*. Perubahan yang mendasar terdapat pada pengukuran kesamaan (*similarity measure*) antara objek dengan sentroid (*Prototype*)-nya. Pada proses pengelompokan dengan K-*Prototype* disini dilakukan beberapa proses utama yaitu:

1. Inisialiasi awal *prototype*

   Pada proses ini akan dilakukan pemilihan sejumlah K-*Prototype* dari *dataset X* sesuai dengan jumlah kelompok yang ditentukan. Pemilihan ini biasanya dilakukan secara acak. Proses ini akan menghasilkan matrik *prototype* berukuran *k* kali panjang atribut *X*.

2. Alokasi *object* di dalam X ke kelompok dengan *prototype* terdekat

   Pada proses ini akan dilakukan pengalokasian semua *object* di dalam dataset *X* ke kelompok yang memiliki jarak *prototype* terdekat dengan *object* yang diukur. Untuk setiap kali objek *x* selesai dialokasikan, maka selanjutnya

akan dilakukan penghitungan (*update*) terhadap *prototype* kelompok yang berkaitan.

3. Realokasi objek jika terjadi perubahan *prototype*

   Setelah semua objek dalam *X* selesai dialokasikan, selanjutnya akan dilakukan pengukuran ulang jarak antara semua *object* di dalam *X* terhadap semua *prototype* yang ada. Jika ditemukan adanya *object* yang ternyata lebih dekat ke *prototype* yang lain, maka akan dilakukan pemindahan keanggotaan dan kemudian akan dilakukan update terhadap *prototype* kelompok lama dan *prototype* kelompok baru. Proses ini akan terus dilakukan sampai tidak ada lagi perubahan *prototype*. Misalkan_ $X = \{X_1, X_2, ..., X_n\}$ adalah kumpulan *n* objek dan $X_i = [X_{i1}, X_{i2}, ..., X_{im}]^T$, dimana *m* menunjukan atribut dan *i* menunjukkan kelompok ke-*i*.

2.14 Ukuran Kesamaan (*Similarity Measure)*

Bentuk umum ukuran kesamaan dinyatakan sebagai berikut

$$d(X_j, Z_i) = \sum_{j=1}^{m} \delta(x_{jl}, z_{il}) \tag{2.8}$$

$z_i = [z_{i1}, z_{i2}, ..., z_{im}]^T$ adalah *prototype* untuk kelompok *i*. Ukuran kesamaan untuk atribut numerik dikenal dengan jarak euclidean ditunjukkan dalam persamaan (2.9) berikut ini

$$d(X_j, Z_i) = \left(\sum_{l=1}^{m_r} (x_{jl}{}^r - z_{il}{}^r)^2 \right)^{1/2} \tag{2.9}$$

$x_{jl}{}^r$ adalah nilai pada atribut numeric *l*, $z_{il}{}^r$ adalah rata-rata atau *prototype* atribut numerik ke *l* kelompok *i*. m_r adalah jumlah atribut numerik. Sedangkan ukuran kesamaan untuk data kategorikal adalah

$$d(X_j, Z_i) = \gamma_i \sum_{l=l+1}^{m_c} \delta(x_{jl}{}^c, z_{il}{}^c) \tag{2.10}$$

dan *simple matching similarity measure* untuk data kategorik adalah

$$\delta(x_{jl}{}^{c}, z_{il}{}^{c}) = \begin{cases} 0 & \left(x_{jl}{}^{c} - z_{il}{}^{c}\right) \\ 1 & \left(x_{jl}{}^{c} \neq z_{il}{}^{c}\right) \end{cases} \quad (2.11)$$

γ_i adalah bobot untuk atribut kategori pada kelompok *i* yang nilainya merupakan nilai standar deviasi untuk atribut numerik pada masing-masing kelompok. Ketika $x_{jl}{}^{c}$ adalah nilai atribut kategorik $z_{il}{}^{c}$ adalah modus atribut ke *l* kelompok *i*. m_c adalah jumlah atribut kategorik. He memodifikasi *simple matching similarity measure* menjadi persamaan (2.12) untuk meningkatkan kemiripan objek dalam kelompok dengan atribut kategorik sehingga hasil pengelompokan menjadi lebih baik (Amir & Lipika, 2007).

$$\delta(x_{jl}{}^{c}, z_{il}{}^{c}) = \begin{cases} 1 - x_{jl}^{c} & (x_{jl}{}^{c} - z_{il}{}^{c}) \\ 1 & (x_{jl}{}^{c} \neq z_{il}{}^{c}) \end{cases} \quad (2.12)$$

$\omega(x_{jl}^{c}, i)$ adalah nilai penimbang untuk x_{jl}^{c} dimana nilai $\omega(x_{jl}^{c}, i)$ adalah

$$\omega(x_{jl}^{c}, i) - \frac{f\left(x_l^c \mid C_l\right)}{\left|C_i\right| f(x_l^c \mid D)} \quad (2.13)$$

$f\left(x_l^c \mid C_l\right)$ adalah frekuensi nilai x_l^c dalam kluster *i*, dan $\left|C_i\right|$ adalah jumlah objek dalam kluster *i*, dan $f(x_l^c \mid D)$ adalah frekuensi nilai pada keseluruhan dataset. *Matching Similarity Measure* yang digunakan untuk data kategorikal menggunakan formula He. Berdasarkan persamaan (2.8) hingga persamaan (2.11), maka ukuran kesamaan untuk data yang memiliki atribut numerik dan atribut kategorikal adalah

$$d\left(X_j, Z_i\right) = \left(\sum_{l=1}^{m_r} \left(x_{jl}{}^{r} - z_{jl}{}^{r}\right)^2 + \gamma_i \sum_{l=l+1}^{m_c} \delta(x_{jl}{}^{c}, z_{il}{}^{c}) \right)^{\frac{1}{2}} \quad (2.14)$$

2.15 Pengelompokan data campuran berdasarkan Puncak Densitas (DPC_M)

A. Algoritma DPC

Algoritma DPC didasarkan pada dua asumsi mendasar: yaitu titik pusat kelompok memiliki kepadatan lokal yang tinggi dan dikelilingi oleh titik

dengan kerapatan lokal yang lebih rendah, dan titik pusat kelompok relatif jauh dari titik tetangganya dengan kepadatan lebih tinggi. Oleh karena itu, algoritma DPC menyusun sebuah *Decision Graph* dengan menghitung kerapatan local ρ_i dan jarak relatif δi untuk menemukan pusat kelompok dalam kumpulan data. Sisa titik data dalam kumpulan data dialokasikan sekaligus ke kluster pusat terdekat, Misalkan $S = \{X_1, X_2, ..., X_n\}$ adalah kumpulan data untuk pengelompokan dan $d_{ij} = dist(X_i, X_j)$ adalah jarak antara titik data X_i dan X_j. Algoritma DPC mendefinisikan jarak *cutoff* d_c dan kepadatan lokal ρ_i sebagai formula (2.15) dan jarak δ_i sebagai formula (2.16) untuk titik data, dimana $\chi(x) = 1$ jika $x < 0$ dan $\chi(x) = 0$, jika tidak maka

$$\rho_i = \sum_j \chi(d_{ij} - d_c) \tag{2.15}$$

$$\delta_i = \begin{cases} \min\limits_{j:\rho_j > \rho_i}(d_{ij}), \rho_i < \max\limits_k(\rho_k) \\ \max\limits_j(d_{ij}), \rho_i < \max\limits_k(\rho_k) \end{cases} \tag{2.16}$$

disini, jarak δ_i didefinisikan sebagai jarak yang sesuai ke titik data X_i bila kepadatan lokal tidak maksimum namun memiliki nilai minimum dari jarak tersebut dari titik ke titik di mana semua kerapatannya lebih besar daripada itu, atau butuh jarak maksimum ke titik yang lain. Bila jumlah titik data dalam dataset sedikit, Efek penghitungan kepadatan lokal ρ_i dengan rumus (2.15) tidak ideal. Oleh karena itu, (Rodriguez & Laio,2014) memberi fungsi kernel Gaussian untuk dataset dengan data lebih sedikit sebagai berikut&

$$\rho_i = \sum_{j \neq i} \exp\left(-\left(\frac{d_{ij}}{d_c}\right)^2\right) \tag{2.17}$$

Berdasarkan kepadatan lokal ρ_i dan jarak δ_i untuk setiap data titik, dapat secara eksplisit memilih jumlah dan titik pusatnya dari kelompok pada

Decision Graph, ketika titik tengahnya ditentukan, setiap titik data yang tersisa dapat diklasifikasikan menjadi kluster yang sama dengan tetangga terdekatnya dengan kepadatan tinggi.

B. Ukuran Jarak Terpadu.

Anggap $S = \{X_1, X_2, ..., X_n\}$ menjadi kumpulan data campuran dengan dimensi d dan n contoh, di mana atribut numeriknya d_r dimensi dan atribut kategorik memiliki ukuran $d_c = d - d_r$. Untuk dua titik data X_i dan X_j, jarak kedua titik data tersebut didefinisikan seperti ditunjukkan pada rumus berikut&

$$D(X_i, X_j) = d(X_i, X_j)_r + d(X_i, X_j)_c \quad (2.18)$$

Persamaan (2.15) menggambarkan jarak perhitungan dari atribut numerik $d(X_i, X_j)_r$ dan atribut kategorik $d(X_i, X_j)_c$ masing-masing:

$$d\left(X_j, Z_i\right) = \left(\sum_{l=1}^{m_r} (x_{jl}{}^r - z_{il}{}^r)^2\right)^{1/2} \quad (2.19)$$

dimana $d(X_i, X_j)_r$ adalah jarak *Euclidean* sesuai persamaan 2.19 dari atribut numerik titik data X_i, X_j. Sedangkan untuk jarak atribut kategorik, digunakan jarak yang sesuai dari titik data (X_i, X_j) pada ke t atribut kategorik dihitung dengan rumus sebagai berikut:

$$\delta\left(x_{it}, x_{jt}\right) = \begin{cases} 0, (x_{it} = x_{jt}) \\ 1, (x_{it} \neq x_{jt}) \end{cases} \quad (2.20)$$

2.16 *Preprocessing* Data Mining

A. Karakteristik dari Data Mentah

Pada data mentah sering ditemukan banyaknya nilai yang hilang (*missing value),* distorsi nilai, tidak tersimpannya nilai (*misrecording),* sampling yang tidak cukup bagus dan sebagainya. Untuk itu perlu ditingkatkan

kualitasnya dengan melakukan penyiapan data (*preprocessing).*Penyebab kurang baiknya kualitas data mentah adalah karena adanya kesalahan dalam penyimpanan dan pengukuran, tapi bisa juga karena tidak adanya nilai mewakili yang tersedia. *Outlier* atau adanya nilai yang tidak biasa (lain dari umumnya) muncul karena banyak hal, antara lain kesalahan pada entri data dan adanya data yang tidak tersimpan sehingga nilai default otomatis tersimpan.

B. Transformasi Data

Data mentah perlu dilakukan proses transformasi untuk meningkatkan performanya. Salah satu transformasi yang umum digunakan adalah dengan melakukan normalisasi.

C. Penanganan Terhadap Data yang Hilang

Metode data mining seringkali mensyaratkan semua nilai data lengkap atau tidak ada yang hilang. Padahal pada kenyataannya banyak atribut atau field dari beberapa record yang tidak diketahui nilainya. Solusi paling sederhana adalah dengan menghapus semua record yang berisi nilai yang kosong. Untuk data yang besar mungkin cara ini tidak berpengaruh terhadap model data mining yang dihasilkannya. Lain hasilnya jika data-data yang dihapus ini memiliki potensi yang sangat besar. Solusi untuk menangani data yang hilang adalah data miner bersama-sama dengan pakar domain secara manual menguji data-data yang kosong kemudian memperkirakan nilai yang tepat untuk data tersebut. Akan tetapi metode ini akan membutuhkan waktu yang lama apalagi jika data yang ditangani berukuran besar dan berdimensi banyak. Pendekatan kedua dilakukan dengan cara penggantian suatu nilai konstanta terhadap nilai yang hilang tersebut. Selain itu ada lagi cara yang bisa dilakukan, yaitu dengan mengintepretasikan nilai yang hilang sebagai nilai ”*don’t care*”. Dengan cara ini, suatu sample data dengan nilai yang kosong akan digantikan oleh beberapa data dari himpunan sampel buatan yang berisi semua kemungkinan yang ada dari domain nilai tersebut.

D. Analisa *Outlier*

Seringkali pada data set, terdapat suatu nilai yang berbeda dari biasanya dan tidak mencerminkan karakteristik data secara umum. Nilai yang tidak konsisten itu dinamakan *outlie*r. Berikut ini metode untuk melakukan deteksi terhadap outlier:

1. Deteksi outlier berdasarkan teknik statistik

   Cara paling sederhana adalah dengan cara statistik. Perlu dilakukan perhitungan rata-rata dan standar deviasi. Kemudian berdasarkan nilai tersebut dibuat fungsi threshold berpotensi untuk dinyatakan sebagai outlier.

2. *Distance Based Outlier Detection*

   Metode yang kedua ini berusaha mengeliminasi keterbatasan dari pendeteksian berdasarkan teknik statistik. Metode ini cocok digunakan untuk data yang multidimensi. Cara yang dilakukan adalah dengan mengevaluasi nilai jarak diantara semua sampel data set yang berukuran *n*-dimensi.

2.17 Evaluasi Pengelompokan

Untuk mengetahui kelayakan hasil pengelompokan dapat melakukan evaluasi yang meliputi dua hal yaitu:

2.17.1 Uji Validitas Kelompok

Permasalahan utama dalam analisis kelompok adalah jumlah kelompok yang harus ditentukan oleh peneliti karena belum ada dasar yang kuat mengenai jumlah kelompok terbaik. Langkah selanjutnya yaitu melakukan uji validitas kelompok untuk mengevaluasi hasil dari Analisis kelompok secara kuantitatif sehingga dihasilkan kelompok optimum. Kelompok optimum adalah kelompok yang mempunyai jarak yang padat antar individu dalam kelompok dan terisolasi dari kelompok lain dengan baik (Dubes & Jain, 1988). Setelah kelompok-kelompok terbentuk, timbul pertanyaan apakah kelompok-kelompok yang terbentuk tersebut valid atau tidak. Ada dua tipe untuk megukur

validitas kelompok, yaitu ukuran eksternal dan ukuran internal (Steinbach, Karypis, dan Kumar, 2000).

A. Ukuran Internal

Ukuran ini digunakan untuk mengukur struktur kelompok yang terbentuk tanpa pertimbangan informasi dari luar, (Steinbach, Karypis, dan Kumar, 2000).

1. Kelompok *cohesion* dan kelompok *separation* adalah dua ukuran internal yang dapat digunakan dalam menentukan validitas kelompok (Steinbach, Karypis, dan Kumar, 2004). Kelompok *cohesion* mengukur seberapa dekat obyek-obyek yang berada dalam satu kelompok, ukuran ini diukur dengan menggunakan jumlah galat kuadrat didalam kelompok,

   $$WSS = \sum_k \sum_{x \in C_k} (x - \bar{x}_k)^2 \qquad (2.21)$$

   Kelompok *separation* digunakan untuk mengukur seberapa berbeda kelompok-kelompok yang terbentuk. Ukuran ini diukur bedasarkan jumlah kuadrat kesalahan error kelompok (*between cluster sum squares of error*)

   $$BSS = \sum_k |C_k| (x - \bar{x}_k)^2 \qquad (2.22)$$

   Dimana $|C_k|$adalah kelompok ke-k, dan $\bar{x}$ adalah rata-rata objek pengamatan.

2. Indeks *Dunn* (*D*)

   Indeks validasi *Dunn* dilambangkan dengan *D* dihitung dengan rumusan berikut:

   $$D = \min_{1 \le l \le n} \left\{ \min_{1 \le j \le n, i \ne j} \left\{ \frac{d\left(c_i, c_j\right)}{\max_{1 \le k \le n} (d'(c_k))} \right\} \right\} \qquad (2.23)$$

   dengan $d\left(c_i, c_j\right)$ = jarak antar kelompok c_i dan c_j,

   $d'\left(c_k\right)$ = jarak dalam kelompok c_k .

Nilai terbesar dari D diambil sebagai jumlah optimum kelompok (Azuaje & Bolshakova, 2001).

3. Indeks *Davies-Bouldin* (DB)

   Rumus indeks *Davied-Bouldin* dapat ditulis sebagai:

$$DB = \frac{1}{n}\sum_{i=1}^{n} \max_{i \neq j} \left[\frac{d'(c_i) + d'(c_j)}{d(c_i, c_j)} \right] \tag{2.24}$$

   dengan $d(c_i, c_j)$ = jarak antar kelompok c_i dan c_j, $d'(c_k)$ = jarak dalam kelompok c_k. Nilai indeks Davies-Bouldin yang kecil menunjukkan kelompok yang baik (Su, 2003).

4. Indeks *Global Silhouette*

   Untuk mendapatkan indeks kelompok $S(i)$ digunakan rumus berikut:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{2.25}$$

   dengan $a(i)$ = rata-rata perbedaan dari i-obyek dengan semua obyek lain pada kelompok yang sama. $b(i)$ = obyek pada kelompok lain (di kelompok terdekat). Nilai yang paling besar dari Indeks *Global Silhoutte* menandai jumlah kelompok terbaik yang kemudian diambil sebagai kelompok optimum.

   Rumusan *Global Silhoutte* diberikan oleh:

$$GS_n = \frac{1}{n}\sum_{i=1}^{n} S(i) \tag{2.26}$$

   dengan

   $S(i)$ = Silhoutte kelompok ke-i,

   n = Jumlah kelompok.

B. Ukuran Eksternal

Ukuran ini berdasarkan pada informasi kelas data yang diketahui peneliti. Digunakan untuk mengukur sejauh mana kecocokan kelompok yang telah terbentuk dengan informasi kelas data.

1. *Purity Measure*

Suwarsa (2013) mengemukakan bahwa kelompok dikatakan murni (*pure*) semua objek dengan *class* yang sama berada pada kelompok yang sama. Untuk mengukur tingkat akurasi pengelompokan atau '*r*', pengukuran nilai '*r*' ini menggunakan persamaan berikut ini:

$$r = \frac{1}{n}\sum_{i=1}^{k} a_i \qquad (2.27)$$

Dimana:

r : Tingkat akurasi pengelompokan

k : Jumlah kelompok

a_i : Objek yang muncul didalam kelompok C_i dan pada label *class* yang sesuai. Semakin tinggi nilai *r* (semakin mendekati 1), semakin baik kualitas kelompok. sedangkan untuk menghitung *error* kelompok atau '*e*' seperti persamaan berikut ini:

$$e = 1 - r \qquad (2.28)$$

r adalah nilai tingkat kemurnian kelompok (Suwarsa, 2013).

2. *Entropy*

Salah satu ukuran kualitas eksternal adalah *entropy* (Steinbach, Karypis, dan Kumar, 2000). Langkah pertama dalam perhitungan *entropy* adalah menghitung kelas distribusi untuk masing-masing kelompok. Kemudian hitung p_{ik} peluang bahwa kelompok ke *k* memuat anggota kelas ke *i* , dengan $p_{ik} = \frac{n_{ik}}{n_k}$ dimana n_{ik} adalah banyak anggota kelas *i* yang

berada dikelompok ke *k* dan n_k adalah banyaknya anggota kelompok ke *k*. Nilai *entropy* untuk setiap kelompok adalah:

$$E_k = -\sum_i p_{ik} \log(p_{ik}) \quad (2.29)$$

Jumlah total *entropy*nya adalah&

$$E = \sum_{k=1}^{g} \frac{n_k * E_k}{n} \quad (2.30)$$

dimana g menunjukkan banyaknya kelompok, dan *n* adalah banyaknya obyek pengamatan.

“*Halaman ini sengaja dikosongkan*”

BAB III
METODE PENELITIAN

Pada bab ini dijelaskan mengenai urutan kerja analisis secara umum dalam menyelesaikan algoritma pengelompokan untuk menyelesaikan studi kasus yang merupakan data berskala campuran numerik dan kategorik.

3.1 Data Penelitian

Data yang digunakan dalam penelitian ini adalah data sekunder pelamar Bidikmisi se-Jawatimur Tahun 2016-2017 dari Kementerian Riset, Teknologi, dan Pendidikan Tinggi Republik Indonesia (Kemenristekdikti) data yang digunakan merupakan data-data internal pelamar Bidikmisi yang dibedakan antara data numerik dan data kategorik

3.2 Identifikasi Variabel Penelitian

Variabel penelitian dalam penelitian ini menggunakan data karakteristik pelamar Bidikmisi dan kelengkapan persyaratan pendaftaran yang telah ditentukan oleh Pedoman Bidikmisi 2018. Variabel yang menunjukkan data numerik pelamar Bidikmisi ditampilkan pada Tabel 3.1, dan variabel yang menunjukkan data kategorik pelamar Bidikmisi ditampilkan pada Tabel 3.2.

Tabel 3.1 Variabel Skala Numerik Pelamar Bidikmisi

Variabel	Satuan
Jarak kota	Kilometer
Jumlah tanggungan	Orang
Nilai Semester 4,5	Numerik

Tabel 3.2 Variabel Skala Kategorik Pelamar Bidikmisi

Variabel	Kategori
Pekerjaan Ayah	1 = PNS 2 = Peg. swasta 3 = Wirausaha 4 = TNI&POLRI 5 = Petani 6 = Nelayan 7 = Lainnya
Pekerjaan Ibu	1 = PNS 2 =Peg. swasta 3 = Wirausaha 4 = TNI&POLRI 5 =Petani 6 = Nelayan 7 = Lainnya 8 = Tidak bekerja
Kode MCK	1 = Milik sendiri didalam 2 = Milik sendiri diluar 3 = Berbagai pakai
Sumber Listrik	1 = PLN 2 = Genset&mandiri 3 = Tenaga Surya 4 = PLN & Genset 5 = Tidak ada
Kode Kepemilikan Rumah	1 = Sendiri 2 = Sewa Tahunan 3 = Sewa Bulanan 4 = Menumpang 5 = Tidak memiliki
Sumber Air	1 = Sumur 2 = Sungai&Mata air 3 = PDAM 4 = Kemasan

Tabel 3.2 Lanjutan Variabel Skala Kategorik Pelamar Bidikmisi

Pendidikan Ayah	1 = Tidak sekolah 2 = SD&MI 3 = SMP&MTs 4 = SMA&MA 5 = D1 6 = D2&D3 7 = S1&D4
Pendidikan Ibu	1 = Tidak sekolah 2 = SD&MI 3 = SMP&MTs

	4 = SMA&MA 5 = D1 6 = D2&D3 7 = S1&D4

Tabel 3.2 Lanjutan Variabel Skala Kategorik Pelamar Bidikmisi

Penghasilan Ayah	1	Tidak Berpenghasilan
	2	< Rp. 250.000
	3	Rp. 250.001 - Rp. 500.000
	4	Rp. 500.001 - Rp. 750.000
	5	Rp. 750.001 - Rp. 1.000.000
	6	Rp. 1.000.001 - Rp. 1.250.000
	7	Rp. 1.250.001 - Rp. 1.500.000
	42	Rp. 1.500.001 - Rp. 1.750.000
	43	Rp. 1.750.001 - Rp. 2.000.000
	44	Rp. 2.000.001 - Rp. 2.250.000
	45	Rp. 2.250.001 - Rp. 2.500.000
	46	Rp. 2.500.001 - Rp. 2.750.000
	47	Rp. 2.750.001 - Rp. 3.000.000
	48	Rp. 3.000.001 - Rp. 3.250.000
	49	Rp. 3.250.001 - Rp. 3.500.000
	50	Rp. 3.500.001 - Rp. 3.750.000
	51	Rp. 3.750.001 - Rp. 4.000.000
	52	Rp. 4.000.001 - Rp. 4.250.000
	53	Rp. 4.250.001 - Rp. 4.500.000
	54	Rp. 4.500.001 - Rp. 4.750.000
	55	Rp. 4.750.001 - Rp. 5.000.000
	56	Rp. 5.000.001 - Rp. 5.250.000
	57	Rp. 5.250.001 - Rp. 5.500.000
	58	Rp. 5.500.001 - Rp. 5.750.000
	59	Rp. 5.750.001 - Rp. 6.000.000
	60	Rp. 6.000.001 - Rp. 6.250.000
	61	Rp. 6.250.001 - Rp. 6.500.000
	62	Rp. 6.500.001 - Rp. 6.750.000
	63	Rp. 6.750.001 - Rp. 7.000.000
	64	Rp. 7.000.001 - Rp. 7.250.000
	65	Rp. 7.250.001 - Rp. 7.500.000
	66	Rp. 7.500.001 - Rp. 7.750.000
	67	Rp. 7.750.001 - Rp. 8.000.000

	68	Rp. 8.000.001 - Rp. 8.250.000
	69	Rp. 8.250.001 - Rp. 8.500.000
	70	Rp. 8.500.001 - Rp. 8.750.000
	71	Rp. 8.750.001 - Rp. 9.000.000
	72	Rp. 9.000.001 - Rp. 9.250.000
	73	Rp. 9.250.001 - Rp. 9.500.000
	74	Rp. 9.500.001 - Rp. 9.750.000
	75	Rp. 9.750.001 - Rp. 10.000.000
	77	> Rp. 10.000.000

Tabel 3.2 Lanjutan Variabel Skala Kategorik Pelamar Bidikmisi

Penghasilan Ibu	1	Tidak Berpenghasilan
	2	< Rp. 250.000
	3	Rp. 250.001 - Rp. 500.000
	4	Rp. 500.001 - Rp. 750.000
	5	Rp. 750.001 - Rp. 1.000.000
	6	Rp. 1.000.001 - Rp. 1.250.000
	7	Rp. 1.250.001 - Rp. 1.500.000
	42	Rp. 1.500.001 - Rp. 1.750.000
	43	Rp. 1.750.001 - Rp. 2.000.000
	44	Rp. 2.000.001 - Rp. 2.250.000
	45	Rp. 2.250.001 - Rp. 2.500.000
	46	Rp. 2.500.001 - Rp. 2.750.000
	47	Rp. 2.750.001 - Rp. 3.000.000
	48	Rp. 3.000.001 - Rp. 3.250.000
	49	Rp. 3.250.001 - Rp. 3.500.000
	50	Rp. 3.500.001 - Rp. 3.750.000
	51	Rp. 3.750.001 - Rp. 4.000.000
	52	Rp. 4.000.001 - Rp. 4.250.000
	53	Rp. 4.250.001 - Rp. 4.500.000
	54	Rp. 4.500.001 - Rp. 4.750.000
	55	Rp. 4.750.001 - Rp. 5.000.000
	56	Rp. 5.000.001 - Rp. 5.250.000
	57	Rp. 5.250.001 - Rp. 5.500.000
	58	Rp. 5.500.001 - Rp. 5.750.000
	59	Rp. 5.750.001 - Rp. 6.000.000
	60	Rp. 6.000.001 - Rp. 6.250.000
	61	Rp. 6.250.001 - Rp. 6.500.000

	62	Rp. 6.500.001 - Rp. 6.750.000
	63	Rp. 6.750.001 - Rp. 7.000.000
	64	Rp. 7.000.001 - Rp. 7.250.000
	65	Rp. 7.250.001 - Rp. 7.500.000
	66	Rp. 7.500.001 - Rp. 7.750.000
	67	Rp. 7.750.001 - Rp. 8.000.000
	68	Rp. 8.000.001 - Rp. 8.250.000
	69	Rp. 8.250.001 - Rp. 8.500.000
	70	Rp. 8.500.001 - Rp. 8.750.000
	71	Rp. 8.750.001 - Rp. 9.000.000
	72	Rp. 9.000.001 - Rp. 9.250.000
	73	Rp. 9.250.001 - Rp. 9.500.000
	74	Rp. 9.500.001 - Rp. 9.750.000
	75	Rp. 9.750.001 - Rp. 10.000.000
	77	> Rp. 10.000.000

Tabel 3.2 Lanjutan Variabel Skala Kategorik Pelamar Bidikmisi

Luas bangunan	1 = > 200 m^2 2 = 100-200 m^2 3 = 50-99 m^2 4 = 25-50 m^2 5 = < 25 m^2
Luas Tanah	1 = > 200 m^2 2 = 100-200 m^2 3 = 50-99 m^2 4 = 25-50 m^2 5 = < 25 m^2

Variabel-variabel yang digunakan pada penelitian ini terdiri dari variabel data numerik dan kategorik, dengan deskripsi variabel data numerik sebagai berikut berdasarkan Petunjuk Teknis Bidikmisi Siswa 2018:

Nilai Semester 4 dan 5 merupakan Nilai raport siswa selama semester 4 dan 5, Jarak Kota merupakan jarak rumah tinggal dari pusat kabupaten/kota dalam satuan Kilometer. Jumlah tanggungan merupakan Jumlah orang yang ditanggung keluarga, tidak termasuk ayah/ibu sendiri. Luas tanah adalah luas tanah rumah tinggal siswa, Luas bangunan adalah luas rumah tinggal siswa,

Kode MCK adalah fasilitas mandi cuci kakus di rumah siswa, Sumber air adalah sumber air yang digunakan di rumah tinggal siswa, Penghasilan ayah adalah rata-rata penghasilan kotor per bulan yang diterima oleh Ayah/Wali dalam empat bulan terakhir, Penghasilan ibu adalah rata-rata penghasilan kotor per bulan yang diterima oleh Ibu dalam empat bulan terakhir, Kerja Ayah merupakan pekerjaan yang saat ini sedang ditekuni oleh Ayah siswa, Kerja ibu merupakan pekerjaan yang saat ini sedang ditekuni oleh ibu siswa, Pendidikan ayah merupakan pendidikan terakhir yang pernah ditempuh oleh Ayah. Pendidikan ibu merupakan pendidikan terakhir yang pernah ditempuh oleh Ibu, Kode Kepemilikan adalah Kepemilikan rumah yang ditempati keluarga sekarang, Kode Listrik adalah Daya Listrik yang digunakan dirumah tinggal siswa.

3.3 Struktur Data Penelitian

Struktur data dari penelitian ini berdasarkan variabel-variabel yang telah disebutkan sebelumnya disajikan dalam Tabel 3.3

Tabel 3.3 Struktur Data Penelitian

Subjek ke *i*	**Variabel ke-*j***					
	X_1	X_2	X_3	X_4	…	X_{16}
$X_{1.j}$	$X_{1.1}$	$X_{1.2}$	$X_{1.3}$	$X_{1.4}$	…	$X_{1.16}$

Tabel 3.3 Lanjutan Struktur Data Penelitian

$X_{2.j}$	$X_{2.1}$	$X_{2.2}$	$X_{2.3}$	$X_{2.4}$	…	$X_{2.16}$
$X_{3.j}$	$X_{3.1}$	$X_{3.2}$	$X_{3.3}$	$X_{3.4}$	…	$X_{3.16}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$X_{i.j}$	$X_{i.1}$	$X_{i.2}$	$X_{i.3}$	$X_{i.4}$	…	$X_{i.16}$

3.4 Menyusun Algoritma Metode Ensembel *Fuzzy*

Metode ensembel *Fuzzy* yang digunakan dengan langkah-langkah berikut ini&

A. Langkah-langkah algoritma *Fuzzy* C-Means

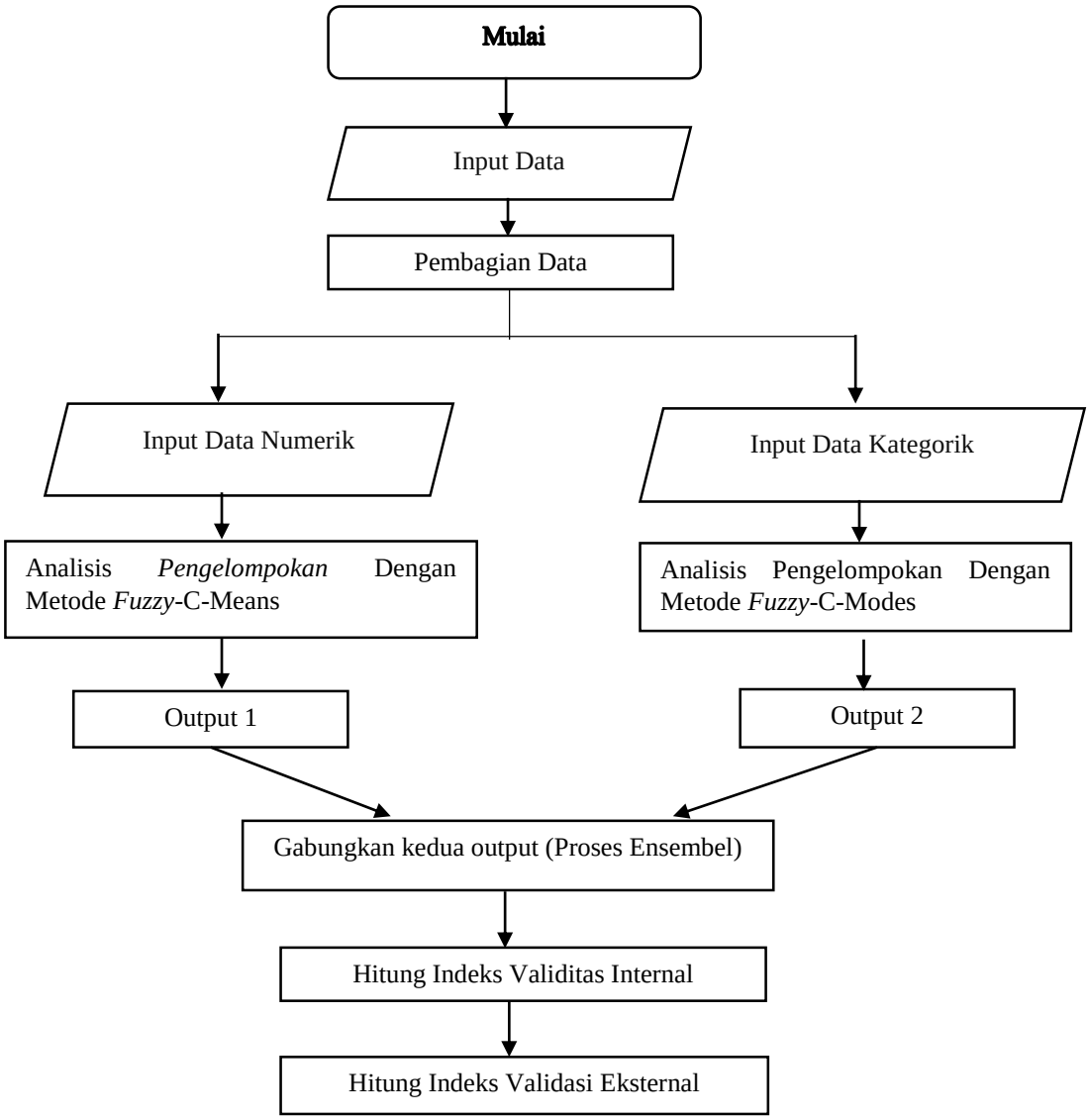
1. Menetapkan nilai dari tiap $\mu_{ij,}$ dimana $\mu_{ij,} \geq 0$.
2. Melakukan iterasi pada Langkah 1.
3. Menghitung sentroid kelompok v_j menggunakan Persamaan 2.6.
4. Menggunakan v_j yang terbaru, perbarui nilai μ_{ij} oleh Persamaan 2.5.

B. Langkah-langkah algoritma *Fuzzy C-Modes*

1. Secara acak menetapkan label kelompok untuk setiap objek, yaitu menginisialisasi keanggotaan kelompok $W^{(1)}$ Tentukan $C^{*(1)}$ sehingga meminimalkan $F\left(W^{(1)}, C^{*(1)}\right)$ jika t =1.
2. Menentukan $W^{(t+1)}$ sedemikian hingga $F\left(W^{(t+1)}, C^{*(t)}\right)$ paling minimal. Jika $F\left(W^{(t+1)}, C*^{(t)}\right) = F\left(W^{(t)}, C*^{(t)}\right)$ maka berhenti. Jika tidak $t = t+1$ dan menuju pada Langkah 3.
3. Menentukan $C*^{(t+1)}$ sehingga meminimalkan $F\left(W^{(t+1)}, C*^{(t+1)}\right)$. Jika $F\left(W^{(t+1)}, C*^{(t+1)}\right) = F\left(W^{(t+1)}, C*^{(t)}\right)$ maka berhenti, jika tidak kembali pada Langkah 2.

3.4.1 Tahapan Analisis Data Metode Ensembel *Fuzzy*

Pada Tahapan Analisis Data Metode Ensembel *Fuzzy* memiliki prosedur pengelompokan seperti disajikan pada Gambar 3.1



Gambar 3.1 *Flowchart* Prosedur Pengelompokan Pada Metode Ensembel *Fuzzy*

3.4.2 Tahapan Algoritma *K-Prototypes*

Sebelum masuk proses algoritma *K-Prototypes* tentukan jumlah *k* yang akan dibentuk batasannya minimal 2 dan maksimal $\sqrt{n}$ dimana *n* adalah jumlah data poin atau obyek. (Lin et al, 2005)

Tahap 1: Menentukan *k* dengan inisial kluster $z_1, z_2, ..., z_k$ secara acak dari *n* buah titik $\{x_1, x_2, ..., x_n\}$.

Tahap 2: Menghitung jarak seluruh data *point* pada dataset terhadap inisial kluster awal, alokasikan data point ke dalam kelompok yang memiliki jarak *prototype* terdekat dengan objek yang diukur.

Tahap 3: Menghitung titik pusat kelompok yang baru setelah semua objek dialokasikan. Lalu realokasikan semua titik data pada dataset terhadap *prototype* yang baru.

Tahap 4: Jika titik pusat kelompok tidak berubah atau sudah konvergen maka proses algoritma berhenti tetapi jika titik pusat masih berubah-ubah secara signifikan maka proses kembali ke tahap 2 dan 3 hingga iterasi maksimum tercapai atau sudah tidak ada perpindahan objek.

3.4.3 Tahapan algoritma *Density Peaks Clustering Mixed* (DPC-M)

Tahap-Tahap Algoritma Pengelompokan Metode DPC-M adalah:

Tahap 1: Menggunakan persamaan (2.18) untuk menghitung jarak masing-masing dua titik dalam kumpulan data.

Tahap 2: Menghitung kepadatan lokal setiap titik data ρ_i berdasarkan persamaan (2.17) dan *p*, dan kemudian jarak δ_i dihitung dengan persamaan (2.16) dan menghitung $\gamma_i = \rho_i * \delta_i$.

Tahap 3: Mengurutkan $\gamma_i = \rho_i * \delta_i$ pada urutan menurun, menghitung infleksi titik yang digunakan untuk menentukan pusat kelompok dan mengatur pusat label kelas.

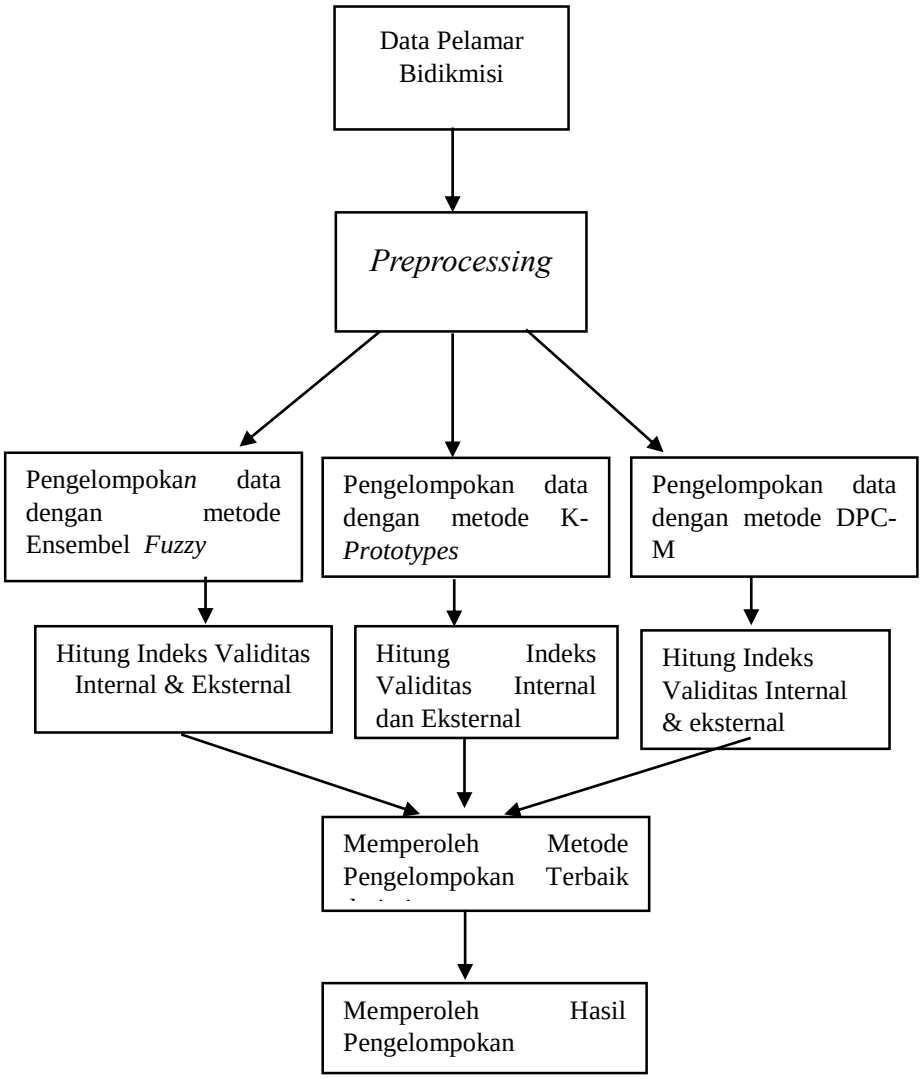
Tahap 4: Titik yang tersisa ditugaskan ke kelas yang labelnya sama satu persatu sebagai tetangga terdekat dengan kepadatan lebih tinggi dan hasil kelompok tercapai.

3.5 Membentuk Rancangan Program Ensembel *Fuzzy* pada *Software R*

Perancangan program pada *software R* dilakukan dengan membentuk algoritma menggunakan bahasa pemrograman dan fungsi-fungsi analisis yang terdapat dalam *package* pada software *R* yang telah tersedia.

3.6 Tahapan Analisis Data

Tahap-tahap Analisis Data menggunakan ketiga metode yaitu: Metode Ensembel *Fuzzy*, K-*Prototypes*, DPC-M seperti disajikan pada Gambar 3.2



Gambar 3.2 Diagram Alir Analisis Data Menggunakan Metode Ensembel *Fuzzy*, K-*Prototypes*, DPC-M

BAB IV
ANALISIS DAN PEMBAHASAN

4.1 Deskripsi Karakteristik Variabel

Analisis deskriptif ini dilakukan terhadap masing-masing variabel yang digunakan dalam penelitian yang meliputi: Nilai Semester 4 dan 5, Jarak Kota dan Jumlah Tanggungan.Nilai Semester 4 dan 5 merupakan Nilai raport siswa selama semester 4 dengan nilai maksimal sebesar 19540, dan minimal 0 dengan rata-rata 1177, dan nilai semester 5, dengan nilai maksimal sebesar 17917, dan minimal 0, dan rata-rata 1166, Jarak Kota merupakan jarak rumah tinggal dari pusat kabupaten & kota dalam satuan Kilometer minimal 0,001 Km dan maksimal 700 Km.

Tabel 4.1 Deskripsi Karakteristik Data Numerik

Variabel	Minimal	Maksimal	Rata-rata
Nilai Semester 4	0	19540	1177
Nilai Semester 5	0	17917	1166
Jarak Kota	0,001	700	19,842
Jumlah Tanggungan	2	65	3

4.2 Tahapan Analisis Metode Ensembel *Fuzzy*, K-*Prototypes* dan DPC-M

Pada subbab ini dijelaskan mengenai algoritma dan program aplikasi yang dilakukan untuk menjawab tujuan pertama dalam penelitian. Penyusunan algoritma dilakukan berdasarkan metode penelitian yang telah dibahas pada Bab III, algoritma tersebut kemudian dijadikan dasar dalam penyusunan program aplikasi pada *Software R project*.

4.3 Tahapan Analisis Pengelompokan Ensembel *Fuzzy*

Algoritma pada analisis pengelompokan ensembel *Fuzzy* dapat dijelaskan sebagai berikut:

Tahap 1: Input yang dimasukkan dalam Analisis Pengelompokan Ensembel *Fuzzy* merupakan matriks *D* sebagai berikut:

$$D = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m_{numerik}} \\ x_{21} & x_{22} & & x_{2m_{numerik}} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm_{numerik}} \end{pmatrix} \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m_{kategorik}} \\ x_{21} & x_{22} & & x_{2m_{kategorik}} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm_{kategorik}} \end{pmatrix}$$

Tahap 2: Input data numerik pada analisis pengelompokan *Fuzzy C-Means* merupakan matriks *DN* sebagai berikut:

$$DN = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m_{numerik}} \\ x_{21} & x_{22} & & x_{2m_{numerik}} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm_{numerik}} \end{pmatrix}$$

Tahap 3: Menghitung jarak antar objek $d\left(X_j, Z_i\right) = \left(\sum_{l=1}^{m_r} (x_{jl}{}^r - z_{il}{}^r)^2 \right)^{1/2}$

Tahap 4: Melakukan pengelompokan data numerik dengan algoritma *Fuzzy* C-Means Metode ensembel *Fuzzy* yang digunakan dengan langkah-langkah berikut ini:

A. Langkah-langkah algoritma *Fuzzy* C-Means

1. Menetapkan nilai dari tiap $\mu_{ij,}$ sesuai persamaan (2.5) yaitu $\mu_{ij} = 1 / \sum_{k=1}^{c} \left(d_{ij} / d_{ik} \right)^{(2/m-1)}$ dimana $\mu_{ij,} \geq 0$
2. Melakukan iterasi pada Langkah 1.
3. Menghitung sentroid kelompok v_j menggunakan Persamaan(2.6).

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

4. Menggunakan v_j yang terbaru, perbarui nilai μ_{ij} oleh Persamaan (2.5).

Tahap 5: Input data numerik pada analisis pengelompokan *Fuzzy* C-*Modes* merupakan matriks *DK* sebagai berikut:

$$DK = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m_{kategorik}} \\ x_{21} & x_{22} & & x_{2m_{kategorik}} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm_{kategorik}} \end{pmatrix}$$

Fuzzy C-modes menggunakan ukuran ketidakmiripan yang sederhana untuk objek yang kategorik yaitu $\delta\left(x_j, y_j\right) = \begin{cases} 0, & x_j = y_j \\ 1, & x_j \neq y_j \end{cases}$.

Tahap 6: Melakukan Pengelompokan dengan Metode K-*Modes*

Langkah-langkah algoritma *Fuzzy C-Modes*

1. Secara acak menetapkan label kelompok untuk setiap objek, yaitu menginisialisasi keanggotaan kelompok $W^{(1)}$ Tentukan $C^{*(1)}$ sehingga meminimalkan $F\left(W^{(1)}, C^{*(1)}\right)$ jika t =1.
2. Menentukan $W^{(t+1)}$ sedemikian hingga $F\left(W^{(t+1)}, C^{*(t)}\right)$ paling minimal. Jika $F\left(W^{(t+1)}, C*^{(t)}\right) = F\left(W^{(t)}, C*^{(t)}\right)$ maka berhenti. Jika tidak $t = t+1$ dan menuju pada Langkah 3.
3. Menentukan $C*^{(t+1)}$ sehingga meminimalkan $F\left(W^{(t+1)}, C*^{(t+1)}\right)$ .Jika $F\left(W^{(t+1)}, C*^{(t+1)}\right) = F\left(W^{(t+1)}, C*^{(t)}\right)$ maka berhenti, jika tidak kembali pada Langkah 2.

Tahap 7: Mengabungkan hasil dari Tahap 2 dan Tahap 3

Tahap 8: Menghitung Indeks Validasi Internal hasil pengelompokan yaitu: SSW sesuai persamaan (2.21) $WSS = \sum_k \sum_{x \in C_k} (x - \bar{x}_k)^2$, Rata-rata Koefisien *Silhouette* sesuai persaman (2.25) $s(i) = \dfrac{b(i) - a(i)}{\max\left(a(i), b(i)\right)}$ dan Indeks *Dunn* sesuai persamaan (2.23)

$$D = \min_{1 \le l \le n} \left\{ \min_{1 \le j \le n, i \ne j} \left\{ \frac{d\left(c_i, c_j\right)}{\max_{1 \le k \le n}\left(d'\left(c_k\right)\right)} \right\} \right\}$$

Tahap 9: Menghitung indeks validasi eksternal yaitu *Purity* berdasarkan (2.27) $r = \frac{1}{n}\sum_{i=1}^{k} a_i$ dan Entropi berdasarkan persamaan (2.30) $E = \sum_{k=1}^{g} \frac{n_k * E_k}{n}$

4.4 Algoritma Pemrograman Metode Ensembel *Fuzzy*

Untuk mempermudah dilakukannya pengelompokan, dibangun rumusan algoritma dalam bahasa pemograman (*syntax*) yang menangani pengelompokan ensembel *Fuzzy* untuk pengelompokan data campuran. Dengan menggunakan program statistika *R Project* (paket R) versi 3.5.0, pemograman untuk pengelompokan ensembel *Fuzzy* menggunakan *syntax* yang dilampirkan pada Lampiran 2. Program Ensembel *Fuzzy* memiliki langkah-langkah utama sebagai berikut:

Input : Data Campuran

Output : Hasil pengelompokan dimana setiap objek pengamatan masuk dalam kelompok tertentu, serta nilai indeks validasi internal kelompok.

Tahap 1 : Memisahkan data menjadi dua kelompok data yaitu Data Numerik dan Data Kategorik.

Tahap 2 : Melakukan pengelompokan Data Numerik dengan menggunakan algoritma pengelompokan metode *Fuzzy* C-*Means*.

Tahap 3 : Melakukan pengelompokan data kategorik dengan menggunakan algoritma pengelompokan metode K-*Modes*.

Tahap 4 : Mengabungkan hasil dari Tahap 2 dan Tahap 3

Tahap 5 : Menghitung Indeks Validasi Internal yaitu: SSW, Rata-rata Koefisien *Silhouette*, dan Indeks *Dunn*.

Tahap 6: Menghitung indeks validasi eksternal yaitu *Purity* dan Entropi.

Adapun Penjelasan pada tiap tahap program Ensembel *Fuzzy* adalah:

Tahap 1: Pemisahan dilakukan dengan cara menggabungkan data-data numerik (X_1, X_4) menjadi satu kelompok data sendiri dan data kategorik (X_5, X_{12}) menjadi satu kelompok data sendiri. Analisis dilakukan dengan menggunakan fungsi data.frame() untuk menggabungkan data.

Tahap 2: Pengelompokan data numerik dilakukan dengan *syntax* yang dilampirkan pada Lampiran 2. Analisis tersebut dilakukan dengan bantuan *package* e1071 dan fungsi-fungsi berikut ini:

a. `dist()`, untuk membentuk matriks jarak.

b. `cmeans()`, untuk melakukan pengelompokan metode *Fuzzy*.

Tahap 3: Pengelompokan data kategorik dilakukan dengan *syntax* yang dilampirkan pada Lampiran 2. Analisis pengelompokan metode Ensembel *Fuzzy* pada *software R* dilakukan dengan bantuan *package* **kLar**, fungsi-fungsi yang digunakan adalah sebagai berikut,

a. `fcmodes()`, untuk melakukan pengelompokan data kategorik

Tahap 4: Melakukan penggabungan tahap 2 dan 3 dengan fungsi data.frame()

Tahap 5: Menghitung indeks validasi internal hasil pengelompokan

a. `cluster.stats()` untuk menghitung indeks validasi internal hasil pengelompokan yaitu SSW, *Average Silhouette coefficient*, dan Index *Dunn*

b. `fviz_Silhouette()` untuk visualisasi hasil pengelompokan

Tahap 6: Menghitung indeks validasi eksternal hasil pengelompokan menggunakan *package* (*FunTime*) dan *Package* (*Entropy*)

a. `Purity()` untuk menghitung kemurnian kelompok

b. `Entropy()` untuk menghitung entropi

4.5 Karakteristik Hasil Pengelompokan Metode Ensembel *Fuzzy*

Berdasarkan Lampiran 2 diperolah Hasil Pengelompokan yang tersaji pada Tabel 4.2

Tabel 4.2 Hasil Pengelompokan Metode Ensembel *Fuzzy*, K-*Prototypes* dan DPC-M

Metode Pengelompokan	Jumlah Kelompok	Pembagian Kelompok	Jumlah Subjek
Ensembel *Fuzzy*	2	1	4029
		2	8129
	3	1	2358
		2	8781
		3	1019
	4	1	2788
		2	7479
		3	1014
		4	877
K-Prototypes	2	1	11138
		2	1020
	3	1	6226
		2	4916
		3	1016
	4	1	5189
		2	4456
		3	1016
		4	1497
DPC-M	2	1	11137
		2	1021
	3	1	8582
		2	1021
		3	2555
	4	1	8582
		2	1021
		3	1265
		4	1290

Setelah melakukan seluruh tahap pengelompokan menggunakan metode Ensembel *Fuzzy* maka karakteristik hasil pengelompokannya disajikan pada Tabel 4.3.

Tabel 4.3 Karakteristik Hasil Pengelompokan Metode Ensembel *Fuzzy*

Metode Pengelompokan	**Ensembel *Fuzzy***								
Jumlah Kelompok	**2**		**3**			**4**			
Variabel Penelitian	**Karakteristik Kelompok**								
	1	**2**	**1**	**2**	**3**	**1**	**2**	**3**	**4**
Rata-rata Nilai Semester 4	1504	967,9	1646	1130	123,1	1405	1096	120,6	1931
Rata-rata Nilai Semester 5	1379	976,5	1592	1128	115	1360	1105	112,7	1834
Rata-rata Jarak Kota	21,44	21,65	21,64	21,42	22,88	20,57	21,43	22,81	24,74
Rata-rata Jumlah Tanggungan	3	3	3	3	3	3	3	3	3
Mayoritas Luas Tanah & Jumlah Subjek	1&1077	1&2718	1&643	1&3002	3&355	2&764	1&2677	3&353	1&293
Mayoritas Luas Bangunan & Jumlah Subjek	3&1674	3&3758	3&974	3&4031	3&427	3&1135	3&3493	3&427	3&377
Mayoritas Kode MCK & Jumlah Subjek	1&2785	1&5628	1&1590	1&6110	1&713	1&1898	1&5173	1&710	1&632
Mayoritas Sumber Air & Jumlah Subjek	1&3036	2&1866	1&1807	1&6498	1&808	1&2125	1&5535	1&805	1&648
Mayoritas Penghasilan Ayah & Jumlah Subjek	3&928	1&6077	3&525	3&1885	5&232	3&624	3&1572	5&231	3&214
Mayoritas Penghasilan Ibu & Jumlah Subjek	1&1831	3&1681	1&1057	1&4381	1&428	1&1274	1&3761	1&426	1&405
Mayoritas Kerja Ayah & Jumlah Subjek	7& 1397	1&4035	7&800	7&2982	7&357	7&1025	7&2502	7&355	5&298
Mayoritas Kerja Ibu & Jumlah Subjek	8&1422	7&2742	8&833	8&3558	7&394	8&988	8&3085	7&393	8&319
Mayoritas Pendidikan Ayah & Jumlah Subjek	2&1745	8&3327	2&1027	2&3560	2&379	2&1245	2&3001	2&378	2&342
Mayoritas Pendidikan Ibu & Jumlah Subjek	2&1866	2&3221	2&1092	2&3720	2&396	2&1311	2&3130	2&394	2&373
Mayoritas Kode Kepemilikan Tempat tinggal & Jumlah Subjek	1&3150	2&3342	1&1836	1&6580	1&834	1&2181	1&5554	1&830	1&685
Mayoritas Kode Listrik & Jumlah Subjek	1&3963	1&6100	1&2302	1&8680	1&1012	1&2735	1&7386	1&1007	1&866

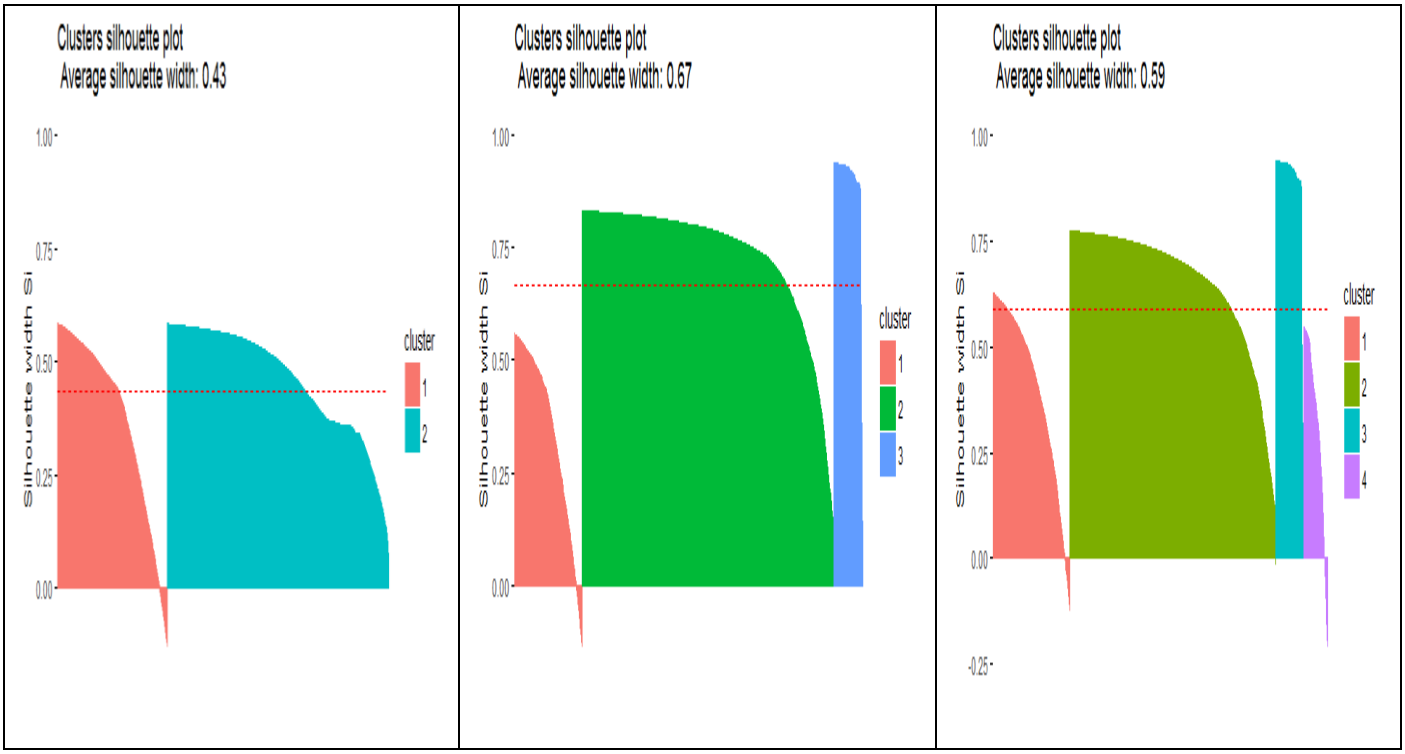
4.6 Perhitungan Indeks Validasi Internal Metode Ensembel *Fuzzy*

Berdasarkan Lampiran 2 diperoleh indeks validasi internal sebagai berikut:

Tabel 4.4 Hasil Perhitungan Indeks Validasi Internal Kelompok Metode Ensembel *Fuzzy*

Metode Ensembel *Fuzzy*			
Kelompok / **Indeks Validasi Eksternal**	**2**	**3**	**4**
SSW	**2.539.938**	7.074.771	4.705.602
Rata-rata Koefisien *Silhouette*	0,4331465	**0,6689594**	0,5897467
Indeks *Dunn*	0,0008684	**0,0010923**	0,0010845

Tabel 4.4 menjelaskan bahwa metode Ensembel *Fuzzy* menghasilkan dua, tiga dan empat kelompok, diperolah nilai SSW untuk dua kelompok sebesar 2.539.938, 3 kelompok sebesar 7.074.771, dan 4 cluster sebesar 4.705.602 Rata-rata koefisien *Silhouette* terletak di antara -1 (pengamatan yang tidak bergerombol) menjadi 1 (observasi yang terkumpul dengan baik). nilai rata-rata koefisien *Silhouette* dua kelompok terbentuk sebesar 0,4331465, tiga kelompok sebesar 0,6689594 dan empat kelompok sebesar 0,5897467, dari dua, tiga dan empat kelompok rata-rata koefisien *Silhouette* dari tiga kelompoklah sebesar 0,6689594 yang mendekati 1 artinya sebesar 67% data campuran pada penelitian telah terkumpul dengan baik. Nilai Indeks *Dunn* sebesar 0,0010923 artinya jarak terkecil antar subjek pengamatan tidak dalam kelompok yang sama dengan jarak intra-kelompok terbesar, indeks ini bernilai lebih maksimal lebih baik, artinya hasil pengelompokan antar anggota dalam kelompok hasilnya baik. Berikut adalah Plot *Silhouette* hasil pengelompokan menggunakan Metode Ensembel *Fuzzy* disajikan pada Gambar 4.1.



Gambar 4.1 Plot *Silhouette* Dua, Tiga Dan Empat Kelompok Metode Ensembel *Fuzzy*

4.7 Perhitungan Indeks Validasi Eksternal Metode Ensembel *Fuzzy*

Tabel 4.5 Hasil Perhitungan Indeks Validasi Eksternal Metode Ensembel *Fuzzy*

Metode Ensembel *Fuzzy*			
Indeks Validasi Eksternal	**2**	**3**	**4**
Purity	0,53	**0,57**	0,48
Entropi	9,26	9,26	9,26

Berdasarkan Lampiran 5 diperoleh indeks validasi eksternal adalah sebagai berikut: Nilai *Purity* untuk dua kelompok sebesar 0,53, untuk tiga kelompok

0,57 dan untuk empat kelompok sebesar 0,48, artinya hasil pengelompokan dikatakan murni sebesar 53%, 57% dan 48% untuk semua subjek pengamatan dalam penelitian ini adalah pelamar Bidikmisi dengan *class* yang sama berada pada kelompok yang sama dan *error* yang dihasilkan sebesar 47%, 43% ,52% dan nilai entropi sebesar 9,26 yang cenderung sama untuk tiap cluster yang dibentuk adalah nilai yang kecil mengindikasikan data campuran telah mengelompok dengan baik.

4.8 Tahapan Analisis Pengelompokan K-*Prototypes*

Tahap 1: Menentukan *k* dengan inisial kluster $z_1, z_2, ..., z_k$ secara acak dari *n* buah titik $\{x_1, x_2, ..., x_n\}$

Tahap 2: Menghitung jarak seluruh data *point* pada dataset terhadap inisial kluster awal sesuai persamaan 2.19 yaitu

$d\left(X_j, Z_i\right) = \left(\sum_{l=1}^{m_r} (x_{jl}{}^r - z_{il}{}^r)^2\right)^{1/2}$, alokasikan titik data ke dalam kelompok yang memiliki jarak *prototype* terdekat dengan objek yang diukur.

Tahap 3: Menghitung titik pusat kelompok yang baru setelah semua objek dialokasikan. Lalu realokasikan semua titik data pada dataset terhadap *prototype* yang baru

Tahap 4: Jika titik pusat kelompok tidak berubah atau sudah konvergen maka proses algoritma berhenti tetapi jika titik pusat masih berubah-ubah secara signifikan maka proses kembali ke tahap 2 dan 3 hingga iterasi maksimum tercapai atau sudah tidak ada perpindahan objek.

4.9 Algoritma Pemrograman Metode K-*Prototypes*

Pemrograman untuk metode pengelompokan K-*Prototypes* menggunakan *sintax* seperti yang dilampirkan pada Lampiran 3. Pemrograman K-*Prototypes* memiliki langkah-langkah utama sebagai berikut:

Input : Data D

Output : Hasil pengelompokan dimana setiap subjek pengamatan masuk dalam kelompok tertentu, serta nilai SSW, indeks rata-rata koefisien *Silhouette*, nilai index *Dunn*, nilai *Purity* dan Entropi.

Tahap 1 : Menyiapkan data campuran yang telah lolos *preprocessing* terlebih dahulu

Tahap 2 : Melakukan pengelompokan data campuran menggunakan algoritma K-*Prototypes* menggunakan *package* (*ClustMixType*)

Tahap 3 : Memperoleh nilai *Estimated* Lambda

Tahap 4 : Menghitung indeks validasi internal yaitu& nilai SSW, indeks rata-rata koefisien *Silhouette*, dan nilai index *Dunn*

Tahap 5 : Menghitung indeks validasi eksternal yaitu& *Purity* dan *Entropy*

Adapun penjelasan teknis tiap tahap pemrograman adalah sebagai berikut:

Tahap 1 : Membaca data campuran numerik (X_1, X_4) dan kategorik (X_5, X_{12}) yang telah dipersiapkan dalam bentuk file *csv* di *R.*

Tahap 2 : Pengelompokan data campuran dengan metode K-*Prototypes* dilakukan dengan bantuan *package* (*ClustMixType*) dan fungsi-fungsi yang ada pada *package* (*ClustMixType*) berikut ini&

a. `Dist()` untuk membentuk matriks jarak
b. `kproto()` untuk membentuk membentuk pengelompokan dengan metode K-*prototypes*
c. `split()` untuk memisahkan hasil pengelompokan

Tahap 3 : Menghitung indeks validasi internal hasil pengelompokan

a. `cluster.stats()` untuk menghitung indeks validasi internal hasil pengelompokan yaitu SSW, *Average Silhouette coefficient*, dan Index *Dunn*
b. `fviz_silhouette()` untuk visualisasi hasil pengelompokan

Tahap 4 : Menghitung indeks validasi eksternal hasil pengelompokan menggunakan *package* (*FunTime*) dan *Package* (*Entropy*)

a. `Purity()` untuk menghitung kemurnian kelompok
b. `Entropy()` untuk menghitung entropi

4.10 Karakteristik Hasil Pengelompokan Metode K-*Prototypes*

Tabel 4.6 Karakteristik Hasil Pengelompokan Metode K-*Prototypes*

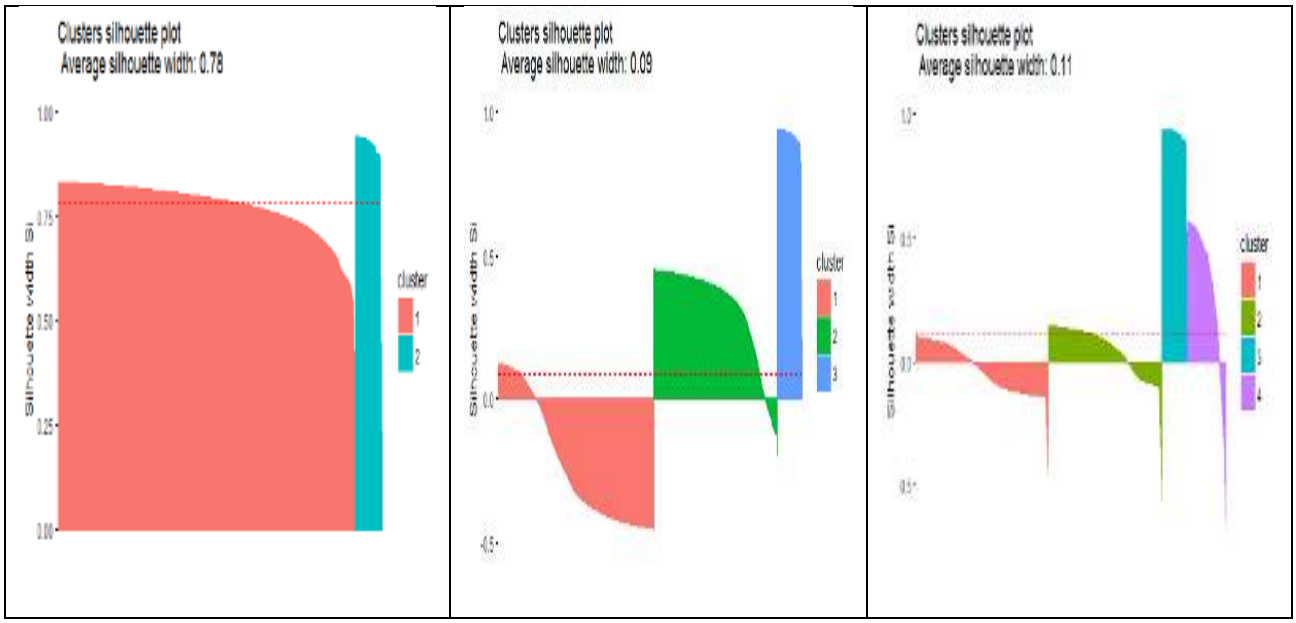
Metode Pengelompokan	**K-*Prototypes***								
Jumlah Kelompok	**2**		**3**			**4**			
Variabel Penelitian	**Karakteristik Kelompok**								
	1	**2**	**1**	**2**	**3**	**1**	**2**	**3**	**4**
Rata-rata Nilai Semester 4	1239	123,1	1294	1169	122,7	1173	1145	122,6	1748
Rata-rata Nilai Semester 5	1227	116,1	1275	1165	112,7	1167	1144	112,6	1679
Rata-rata Jarak Kota	21,47	22,87	23,72	18,62	22,88	23,22	18,65	22,86	23,79
Rata-rata Jumlah Tanggungan	3	3	3	3	3	3	3	3	3
Mayoritas Luas Tanah & Jumlah Subjek	1&3645	3&355	1&2608	2&1939	3&355	1&2309	2&1821	3&355	1&435
Mayoritas Luas Bangunan & Jumlah Subjek	3&5005	3&427	3&2689	3&2317	3&426	3&2380	3&2143	3&426	4&528
Mayoritas Kode MCK & Jumlah Subjek	1&7699	1&714	1&3995	1&3708	1&710	1&3401	1&3332	1&710	1&970
Mayoritas Sumber Air & Jumlah Subjek	1&8305	1&808	1&4656	1&3650	1&807	1&3850	1&3334	1&807	1&1122
Mayoritas Penghasilan Ayah & Jumlah Subjek	3&2410	5&232	4&1718	5&1348	5&232	4&1548	5&1261	5&232	3&452
Mayoritas Penghasilan Ibu & Jumlah Subjek	1&5438	1&428	1&3284	1&2154	1&428	1&2988	1&1979	1&428	2&517
Mayoritas Kerja Ayah & Jumlah Subjek	7&3781	7&358	5&3027	7&2460	7&355	5&2455	7&2266	7&356	5&596
Mayoritas Kerja Ibu & Jumlah Subjek	8&4391	7&394	8&2852	7&2288	7&392	8&2651	7&2094	7&393	7&542
Mayoritas Pendidikan Ayah & Jumlah Subjek	2&4587	2&379	2&3776	4&2697	2&379	2&3072	4&2417	2&379	2&737
Mayoritas Pendidikan Ibu & Jumlah Subjek	2&4812	2&396	2&3893	4&2297	2&396	2&3244	3&1692	2&396	2&784
Mayoritas Kode Kepemilikan Tempat tinggal & Jumlah Subjek	1&8416	1&834	1&5062	1&3355	1&833	1&4133	1&3082	1&833	1&1202
Mayoritas Kode Listrik & Jumlah Subjek	1&10981	1&1013	1&6118	1&4867	1&1009	1&5125	1&4408	1&1009	1&1452

4.11 Perhitungan Indeks Validasi Internal Metode K-*Prototypes*

Tabel 4.7 Hasil Perhitungan Indeks Validasi Internal Kelompok Metode K-*Prototypes*

Metode K-Prototypes			
Kelompok / **Indeks Validasi Eksternal**	**2**	**3**	**4**
SSW	1.604.825	**1.529.382**	8.011.886
Rata-rata Koefisien *Silhouette*	**0,7844612**	0,0862734	0,1135709
Indeks *Dunn*	**0,0426929**	0,0002147	0,0005094

Berdasarkan Lampiran 3 diperoleh indeks validasi internal sebagai berikut bahwa nilai *Estimated* lambda sebesar 135653,3, menghasilkan dua, tiga, dan empat kelompok, diperoleh nilai SSW sebesar 1.604.825 untuk dua kelompok, 1.529.382 untuk tiga kelompok, dan 8.011.886 untuk empat kelompok, nilai rata-rata koefisien *Silhouette* dua kelompok sebesar 0,78 mendekati 1 artinya sebesar 78% data campuran telah terkumpul dengan baik. rata-rata koefisien *Silhouette* tiga kelompok sebesar 0,0862734 dan empat kelompok sebesar 0,1135709, Nilai Indeks *Dunn* sebesar 0,0426929 artinya jarak terkecil antar subjek pengamatan tidak dalam kelompok yang sama dengan jarak intrakelompok terbesar, indeks ini bernilai lebih maksimal lebih baik, artinya hasil pengelompokan antar anggota dalam kelompok hasilnya baik, namun hasil analisis menjelaskan bahwa nilai indeks *Dunn* untuk 3 kelompok dan empat kelompok yang dihasilkan hanya 0,02 % .dan 0,05 %. Pada Gambar 4.2 adalah Plot *Silhouette* hasil pengelompokan menggunakan Metode K-*Prototypes*.



Gambar 4.2 Plot *Silhouette* Dua, Tiga Dan Empat Kelompok Metode K-*Prototypes*

4.12 Perhitungan Indeks Validasi Eksternal Metode K-*Prototypes*

Hasil Perhitungan Indeks Validasi Eksternal Metode K-*Prototypes* tersaji pada Tabel 4.8.

Tabel 4.8 Hasil Perhitungan Indeks Validasi Eksternal Metode Ensembel K-Prototypes

Metode Ensembel K-Prototypes			
Indeks Validasi Eksternal	**2**	**3**	**4**
Purity	**0,73**	0,50	0,42
Entropi	9,37	9,32	**9,27**

Berdasarkan Lampiran 5 diperoleh indeks validasi eksternal adalah sebagai berikut: Nilai *Purity* untuk dua kelompok sebesar 0,73, untuk tiga kelompok 0,50 dan untuk empat kelompok sebesar 0,42, artinya hasil pengelompokan dikatakan murni sebesar 73%, 50% dan 42% untuk semua subjek

pengamatan dalam penelitian ini adalah pelamar Bidikmisi dengan *class* yang sama berada pada kelompok yang sama dan *error* yang dihasilkan sebesar 27%, 50% ,58% dan nilai Entropi yang cenderung berubah-ubah untuk tiap kelompok yang dibentuk namun hanya selisih sedikit tiap kelompok mengindikasikan data campuran telah mengelompok dengan baik.

4.13 Tahapan Analisis Pengelompokan DPC-M

Tahapan Analisis Pengelompokan DPC-M adalah:

Tahap 1: Menggunakan persamaan (2.18) $D(X_i, X_j) = d(X_i, X_j)_r + d(X_i, X_j)_c$ untuk menghitung jarak masing-masing dua titik dalam kumpulan data.

Tahap 2: Menghitung kepadatan lokal setiap titik data ρ_i berdasarkan persamaan (2.15) yaitu $\rho_i = \sum_j \chi\left(d_{ij} - d_c\right)$ dan p, kemudian jarak δ_i dihitung dengan persamaan 2.16 $\delta_i = \begin{cases} \min\limits_{j:\rho_j > \rho_i}\left(d_{ij}\right), \rho_i < \max\limits_k(\rho_k) \\ \max\limits_j\left(d_{ij}\right), \rho_i < \max\limits_k(\rho_k) \end{cases}$, dan menghitung $\gamma_i = \rho_i * \delta_i$

Tahap 3: Mengurutkan $\gamma_i = \rho_i * \delta_i$ pada urutan menurun, menghitung infleksi titik yang digunakan untuk menentukan pusat kelompok dan mengatur pusat label kelas.

Tahap 4: Titik yang tersisa ditugaskan ke kelas yang labelnya sama satu persatu sebagai tetangga terdekat dengan kepadatan lebih tinggi dan hasil kelompok tercapai.

4.14 Algoritma Metode Pengelompokan DPC-M

Rumusan algoritma dalam Bahasa pemrograman (Sintaks) yang memuat metode DPC-M untuk pengelompokan data campuran, dengan menggunakan

program statistika R-Project versi 3.5.0, pemrograman untuk metode pengelompokan DPC-M menggunakan sintax seperti yang dilampirkan pada Lampiran 4. Pemrograman DPC-M memiliki langkah-langkah utama sebagai berikut:

Input : Data D

Output : Hasil pengelompokan dimana setiap subjek pengamatan masuk dalam kelompok tertentu, serta nilai SSW, indeks rata-rata koefisien *Silhouette*, nilai index *Dunn*, nilai *Purity* dan Entropy.

Tahap 1 : Menyiapkan data campuran yang telah lolos preprocessing terlebih dahulu

Tahap 2 : Melakukan pengelompokan data campuran menggunakan algoritma DPC-M menggunakan *package* (*Density Clust*)

Tahap 3 : Memperoleh nilai *Distance cut-off*

Tahap 4 : Menghitung indeks validasi internal yaitu& nilai SSW, indeks rata-rata koefisien *Silhouette*, dan nilai index *Dunn*

Tahap 5 : Menghitung indeks validasi eksternal yaitu: *Purity* dan *Entropy*

Adapun penjelasan teknis tiap tahap pemrograman DPC-M adalah sebagai berikut:

Tahap 1 : Membaca data campuran numerik (X_1, X_4) dan kategorik (X_5, X_{12}) yang telah dipersiapkan dalam bentuk file *csv* di *R.*

Tahap 2 : Pengelompokan data campuran dengan metode DPC-M dilakukan dengan bantuan *package* (DensityClust) dan fungsi-fungsi yang ada pada *package* (DensityClust) berikut ini:

a. `Dist()` untuk membentuk matriks jarak
b. `densityClust()` untuk membentuk *distance cut-off*
c. `findClusters()`untuk membentuk pengelompokan metode DPC-M
d. `split()` untuk memisahkan hasil pengelompokan

Tahap 3 : Menghitung indeks validasi internal hasil pengelompokan

a. `cluster.stats()` untuk menghitung indeks validasi

internal hasil pengelompokan yaitu SSW, *Average Silhouette coefficient*, dan Index *Dunn*.

e. `fviz_silhouette()` untuk visualisasi hasil pengelompokan

Tahap 4 : Menghitung indeks validasi eksternal hasil pengelompokan menggunakan *package* (FunTime) dan *Package* (Entropy)

a. `Purity()` untuk menghitung kemurnian kelompok

b. `Entropy()` untuk menghitung entropi

4.15 Karakteristik Hasil Pengelompokan Metode DPC-M

Berdasarkan Lampiran 3 diperoleh karakteristik hasil pengelompokan tersaji pada Tabel 4.9:

Tabel 4.9 Karakteristik Hasil Pengelompokan Metode DPC-M

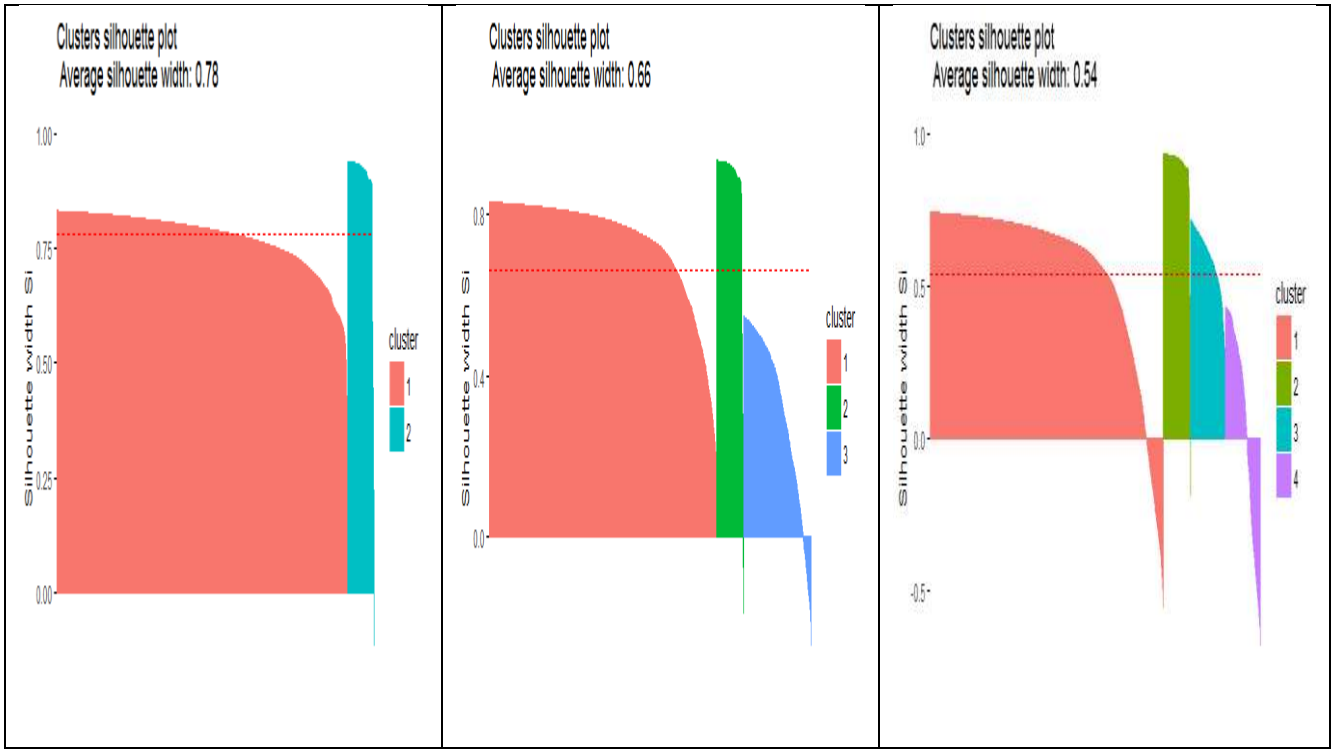
Metode Pengelompokan	**DPC-M**								
Jumlah Kelompok	**2**		**3**			**4**			
Variabel Penelitian	**Karakteristik Kelompok**								
	1	**2**	**1**	**2**	**3**	**1**	**2**	**3**	**4**
Rata-rata Nilai Semester 4	1239	124,4	1124	124,4	1627	1124	124,4	1437	1814
Rata-rata Nilai Semester 5	1227	116,1	1125	116,1	1570	1125	116,1	1400	1736
Rata-rata Jarak Kota	21,47	22,86	21,44	22,86	21,58	21,44	22,86	19,54	23,58
Rata-rata Jumlah Tanggungan	3	3	3	3	3	3	3	3	3
Mayoritas Luas Tanah & Jumlah Subjek	1&3645	3&356	1&2970	3&356	1&675	1&2970	3&356	2&371	1&382
Mayoritas Luas Bangunan & Jumlah Subjek	3&5005	3&427	3&3948	3&427	3&1057	3&3948	3&427	3&531	3&526
Mayoritas Kode MCK & Jumlah Subjek	1&7698	1&715	1&5980	1&715	1&1718	1&5980	1&715	1&835	1&883
Mayoritas Sumber Air & Jumlah Subjek	1&8305	1&808	1&6352	1&808	1&1953	1&6352	1&808	1&975	1&978
Mayoritas Penghasilan Ayah & Jumlah Subjek	3&2410	5&232	3&1841	5&232	3&569	3&1841	5&232	3&275	3&294
Mayoritas Penghasilan Ibu & Jumlah Subjek	1&5437	1&429	1&4281	1&429	1&1156	1&4281	1&429	1&579	1&577
Mayoritas Kerja Ayah & Jumlah Subjek	7&3781	7&358	7&2903	7&358	7&878	7&2903	7&358	7&491	5&455

Mayoritas Kerja Ibu & Jumlah Subjek	8&4391	7&394	8&3491	7&394	8&900	8&3491	7&394	7&445	8&459

Tabel 4.9 Lanjutan Karakteristik Hasil Pengelompokan Metode DPC-M

Mayoritas Pendidikan Ayah & Jumlah Subjek	2&4586	2&380	2&3473	2&380	2&2113	2&3473	2&380	2&573	2&540
Mayoritas Pendidikan Ibu & Jumlah Subjek	2&4811	2&397	2&3631	2&397	2&1180	2&3631	2&397	2&615	2&565
Mayoritas Kode Kepemilikan Tempat tinggal & Jumlah Subjek	1&8415	1&835	1&6419	1&835	1&1996	1&6419	1&835	1&987	1&1009
Mayoritas Kode Listrik & Jumlah Subjek	1&10980	1&1014	1&8482	1&1014	1&2498	1&8482	1&1014	1&1245	1&1253

Plot *Silhouette* hasil pengelompokan menggunakan Metode DPC-M disajikan pada Gambar 4.3



Gambar 4.3 Plot *Silhouette* Dua, Tiga Dan Empat Kelompok Metode DPC-M

4.16 Perhitungan Indeks Validasi Internal Metode DPC-M

Hasil Perhitungan Indeks Validasi Internal Metode DPC-M tersaji pada Tabel 4.10

Tabel 4.10 Hasil Perhitungan Indeks Validasi Internal Kelompok Metode DPC-M

Metode DPC-M			
Kelompok / **Indeks Validasi Eksternal**	**2**	**3**	**4**
SSW	**1.605.250**	7.161.342	5.531.412
Rata-rata Koefisien *Silhouette*	**0,7843734**	0,6605946	0,5386398
Indeks *Dunn*	**0,0426929**	0,0025307	0,0014568

Berdasarkan Lampiran 4 diperoleh indeks validasi internal sebagai berikut bahwa nilai *distance cut-off* sebesar 24,91508, menghasilkan dua, tiga, dan empat kelompok, diperolah nilai SSW sebesar 1.605.250 untuk dua kelompok, 7.161.342 untuk tiga kelompok, dan 5.531.412 untuk empat kelompok, nilai rata-rata koefisien *Silhouette* dua kelompok sebesar 0,7843734 mendekati 1 artinya sebesar 78% data campuran telah terkumpul dengan baik. rata-rata koefisien *Silhouette* tiga kelompok sebesar 0,6605946 dan empat kelompok sebesar 0,5386398, Nilai Indeks *Dunn* sebesar 0,0426929 artinya jarak terkecil antar subjek pengamatan tidak dalam kelompok yang sama dengan jarak intra kelompok terbesar, indeks ini bernilai lebih maksimal lebih baik, artinya hasil pengelompokan antar anggota dalam kelompok hasilnya baik, namun hasil analisis menjelaskan bahwa nilai indeks *Dunn* untuk tiga kelompok dan empat kelompok yang dihasilkan hanya 0,02 % .dan 0,01 %.

4.17 Perhitungan Indeks Validasi Eksternal Metode DPC-M

Hasil perhitungan Indeks Validasi Eksternal DPC-M tersaji pada Tabel 4.11.

Tabel 4.11 Hasil Perhitungan Indeks Validasi Eksternal Metode Ensembel DPC-M

Metode DPC-M			
Indeks Validasi Eksternal	**2**	**3**	**4**
Purity	**0,73**	0,56	0,56
Entropi	9,26	9,26	9,26

Berdasarkan Lampiran 5 diperoleh indeks validasi eksternal adalah sebagai berikut: Nilai *Purity* untuk dua kelompok sebesar 0,73, untuk tiga kelompok 0,56 dan untuk empat kelompok sebesar 0,56, artinya hasil pengelompokan dikatakan murni sebesar 73%, 56% dan 56% untuk semua subjek pengamatan dalam penelitian ini adalah pelamar Bidikmisi dengan *class* yang sama berada pada kelompok yang sama dan *error* yang dihasilkan sebesar 27%, 44% ,44% dan nilai entropi yang cenderung sama untuk tiap kelompok yang dibentuk mengindikasikan data campuran telah mengelompok dengan baik.

4.18 Perbandingan Hasil Pengelompokan Ensembel *Fuzzy*, Metode K-*Prototypes*, DPC-M

Pada subbab ini dijelaskan mengenai perbandingan hasil pengelompokan Ensembel *Fuzzy*, K-*Prototypes* dan DPC-M yang dilakukan untuk menjawab tujuan kedua penelitian. Berdasarkan Lampiran 5 diperoleh hasil perbandingan metode pengelompokan pada penelitian ini seperti disajikan pada Tabel 4.12.

Tabel 4.12 Perbandingan Nilai Indeks Validasi Internal dan Eksternal untuk Dua Kelompok

Indeks Validasi Internal	**Ensembel *Fuzzy***	**K-*Prototypes***	**DPC-M**
	2		
SSW	2.539.938	**1.604.825**	1.605.250
Rata-rata Koefisien *Silhouette*	0,4331465	**0,7844612**	0,7843734
Indeks *Dunn*	0,0008684	**0,0426929**	0,0426929
Indeks Validasi Eksternal	**2**		
Purity	0,5261556	**0,7344136**	0,7343313
Entropi	9,2616320	9,3790020	9,2616320

Berdasarkan Tabel 4.1 diperoleh bahwa metode K-*Prototypes* adalah metode pengelompokan yang terbaik untuk pengelompokan dua kelompok dibandingkan metode Ensembel *Fuzzy* dan metode DPC-M karena memiliki nilai SSW paling rendah, Rata-rata koefisien *Silhouette* sebesar 0,7844612 mendekati 1 dan nilai Indeks *Dunn* lebih besar sebesar 0,0426929. Dan berdasarkan nilai Indeks validasi eksternal kelompok metode K-protoypes adalah yang terbaik dengan nilai *Purity* 0,7344136 yang mendekati 1. Perbandingan Nilai Indeks Validasi Internal dan Eksternal untuk Tiga Kelompok disajikan pada Tabel 4.13

Tabel 4.13 Perbandingan Nilai Indeks Validasi Internal dan Eksternal untuk Tiga Kelompok

Indeks Validasi Internal	**Metode Ensembel *Fuzzy***	**K-Prototypes**	**DPC-M**
	3		
SSW	7.074.771	**1.529.382**	7.161.342
Rata-rata Koefisien *Silhouette*	**0,6689594**	0,0862733	0,6605946
Indeks *Dunn*	**0,0010923**	0,0002147	0,0022530
Indeks Validasi Eksternal	**3**		
Purity	**0,5718046**	0,5090475	0,5577398
Entropy	9,2616320	9,2616320	9,2616320

Berdasarkan Tabel 4.13 diketahui bahwa metode pengelompokan yang terbaik untuk pengelompokan tiga kelompok adalah metode Ensembel *Fuzzy* dibandingkan metode K-*Prototypes* dan DPC-M karena memiliki nilai SSW paling rendah, Rata-rata koefisien *Silhouette* sebesar 0,6689594 mendekati 1 dan nilai Indeks *Dunn* lebih kecil sebesar 0,0010923. Dan berdasarkan nilai Indeks validasi eksternal kelompok metode Ensembel *Fuzzy* adalah yang terbaik dengan nilai *Purity* 0,5718046 yang mendekati 1. Dan Perbandingan Nilai Indeks Validasi Internal dan Eksternal untuk Empat Kelompok disajikan pada Tabel 4.14.

Tabel 4.14 Perbandingan Nilai Indeks Validasi Internal dan Eksternal untuk Empat Kelompok

Indeks Validasi Internal	**Metode Ensembel *Fuzzy***	**K-Prototypes**	**DPC-M**
	4		
SSW	**4.705.602**	8.011.886	5.531.412
Rata-rata Koefisien *Silhouette*	**0,5897467**	0,1135709	0,5386398
Indeks *Dunn*	**0,0010845**	0,0005094	0,0014568
Indeks Validasi Eksternal	**4**		
Purity	**0,4839612**	0,4262214	0,5576575
Entropy	**9,2616320**	9,2786900	9,2616320

Berdasarkan Tabel 4.14 dapat diketahui bahwa metode pengelompokan yang terbaik untuk pengelompokan empat kelompok adalah metode Ensembel *Fuzzy* dibandingkan metode K-*Prototypes* dan DPC-M karena memiliki nilai SSW paling rendah, Rata-rata koefisien *Silhouette* sebesar 0,5897467 mendekati 1 dan nilai Indeks *Dunn* lebih besar sebesar 0,0010845.

BAB V
KESIMPULAN DAN SARAN

Pada bab ini dijabarkan kesimpulan dan saran yang diperoleh dari analisis dan pembahasan yang diperoleh dari BAB IV. Pada bagian awal bab ini menjabarkan beberapa kesimpulan yang diperoleh berdasarkan penjabaran algoritma dan pemrograman, serta berdasarkan hasil analisis pada penelitian yang dimiliki. Bagian akhir pada bab ini menjabarkan beberapa saran yang dapat diberikan peneliti terkait perbaikan maupun pengembangan yang lebih lanjut.

5.1. Kesimpulan

Berdasarkan hasil analisis dan pembahasan yang telah dilakukan, maka diperoleh beberapa kesimpulan antara lain,

1. Algoritma yang dibangun untuk melakukan analisis metode dapat digunakan untuk menangani permasalahan dalam melakukan pengelompokan data berskala campuran numerik dan kategorik. Pemrograman pada *software R Project* yang dibangun dapat mempermudah dilakukannya analisis dan perbandingan dari kedua metode tersebut.
2. Pengelompokan data campuran numerik dan kategorik menghasilkan dua, tiga dan empat kelompok, baik menggunakan metode Ensembel *Fuzzy*, metode K-*Prototypes* dan DPC-M
3. Berdasarkan dengan perbandingan indeks validasi internal dan eksternal dapat dikatakan bahwa metode Ensembel *Fuzzy* lebih tepat digunakan pada data penelitian ini karena memiliki nilai SSW yang rendah, rata-rata Koefisien *Silhouette* mendekati 1, dan nilai Indeks *Dunn* yang makin besar, dibandingkan metode klasik pengelompokan data campuran yaitu K-*Prototypes* dan DPC-M

5.2. Saran

Penelitian yang telah dilakukan ini masih terdapat beberapa perbaikan dan pengembangan selanjutnya, diantaranya&

1. Penggunaan jarak numerik pada penelitian ini hanya sebatas menggunakan metode *Euclidean* jadi diharapkan untuk penelitian kedepannya bisa menggunakan modifikasi jarak Euclidean sebab penggunaan jarak dapat menentukan kualitas hasil pengelompokan dan hasil pengelompokan yang berbeda dan membandingkan hasil pengelompokan dengan menggunakan jarak yang tidak dimodifikasi
2. Pendekatan pengelompokkan data numerik pada penelitian ini adalah dengan metode Ensembel *Fuzzy*, K-*Prototypes* dan DPC-M sehingga masih terdapat beberapa metode pengelompokan data campuran lain, missal membandingkan hasil pengelompokan dengan metode *two step cluster* untuk data campuran dan biss menggunakan jenis indeks validitas pengelompokan lainnya.

DAFTAR PUSTAKA

Ahmad, A. dan Dey, L. (2007), "A K-Mean Clustering Algoritm for Mixed Numeric And Categorical Data," *Data & Knowledge Engineering*, vol 63, hal 503-527.

Alvionita. (2017), *Metode Ensembel Rock dan SWFM untuk Pengelompokan Data Campuran Kategorik Dan Numerik Pada Kasus Aksesi Jeruk*, Tesis Program Magister FMIPA, Statistika, Institut Teknologi Sepuluh Nopember, Surabaya.

Azuaje, F. dan Nadia, B. (2001), "Improving Expression Data Mining through Kelompok Validity", Departement of Computer Science. Trinity College Dublin, http///www.cs.tcd.ie/publications/techreports/reports.02&TCD-CS-2002-.pdf. Tanggal akses: 28 Februari 2018, Irlandia.

Bolshakova, N. (2003), "Cluster Validity Algorithms", Departement of Computer Science.TrinityCollege Dublin http:///www.cs.tcd.ie/Nadia.Bolshakova/validation_algorithms.html., Tanggal akses: 28 Februari 2018, Irlandia.

Dewangan, R. R. Sharma, L. K. dan Akasapu, A. K. (2010), "*Fuzzy* Clustering Technique for Numerical and Categorical Dataset", *International Journal on Computer Science and Engineering*, hal 75-80.

Dubes. dan Jain. A.K.(1988), "Algorithm for Clustering Data". Prentice Hall. New Jersey.

Fernandes. Rinaldo, A.A. dan Soehono A.L. (2006), "Kajian Analisis Cluster pada Data Berskala Campuran", Laporan Penelitian, FMIPA, Universitas Brawijaya, Malang.

Guha, S. Rastogi, R. dan Shim, K. (2000), "ROCK& A Robust Clustering Algorithm for Categorical Attributes", *Proceedings of the 15th International Conference on Data Engineering.*

Hair, J. F. Black, W. C. Babin, J. B. dan Anderson, E. R. (2009), *Multivariate Data Analysis* (seventh ed.), Prentice Hall Inc, New Jersey.

He, Z. Xu, X. dan Deng, S. (2005a), "A Cluster Ensemble Method For Clustering Categorical Data", *Information Fusion*, 6, hal 143-151.

He, Z. Xu, X. dan Deng, S. (2005a), "A Cluster Ensemble Method For Clustering Categorical Data", *Information Fusion*, 6, hal 143-151.

Huang, Z.X, "Clustering Large Data Sets with Mixed and Numeric and Categorical values", *Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD '97)*, hal. 21–34, 1997.

Iriawan, N, Fithriasari, K, Ulama, S.S.B, Suryaningtyas, W. Susanto, I. dan Pravitasari, A.A. (2018), "Bayesian Bernoulli Mixture Regression Model For Bidikmisi Scholarship Classification", *Jurnal Ilmu Komputer dan Informasi*, hal 67-76, DOI: http://dx.doi.org/10.21609/jiki.v11i2.536.

Iriawan, N, dan Purnomo H.D. (2012), "Penerapan Metode Gustafson-Kessel Clustering untuk Menentukan Segmentasi Debitur pada Bank CIMB Niaga", *Prosiding Seminar Nasional Manajemen Teknologi XV*, Program Studi MMT-ITS Surabaya.

Johnson, R. A., dan Winchern, D. W. (2007), *Applied Multivariate Statistical Analysis* (sixth ed.), Pearson Education, Inc, New Jersey.

Liu, S., Zhou, B, Decai, H., dan Shen, L. (2017), "Clustering Mixed Data by Fast Search and Find of Density Peaks", Journal of Mathematical Problems in Engineering, Hindawi, Vol 2017.

Marina, M.P. dan Fithriahsari, K. (2015), "Pengelompokan Kabupaten/Kota di Jawa Timur Berdasarkan Indikator Kesehatan Masyarakat Menggunakan Metode Kohonen SOM dan *K-Means*", *Jurnal Sains dan Seni ITS*. Vol 4. No. 1.

Muhammad, F. A. (2015). *Klasterisasi Proses Seleksi Pemain Menggunakan Algoritma K-Means* (*Studi Kasus Tim Hockey Kabupaten Kendal*), Skripsi, Teknik Informatika, Universitas Dian Nuswantoro, Semarang.

Nooraeni., R. (2015), "K-*Prototype* untuk Pengelompokan Data Campuran", *Jurnal Sains*, Jurusan Statistika FMIPA, Universitas Padjajaran. Bandung.

Reddy, M. V. J. dan Kavitha, B. (2010), "Efficient Ensemble Algorithm for Mixed Numeric and Categorical Data", *Computational Intelligence and Computing research (ICCIC)*, IEEE International Conference.

Reddy, M. V. J. dan Kavitha, B. (2012), "Clustering the Mixed Numerical and Categorical Dataset using Similarity Weight and Filter Method", *International Journal of Database Theory and Application*, Vol 5, No 1.

Rodriguez, A. dan Laio, A. (2014), "Clustering by Fast Search And Find Of Density Peaks", *Science*, vol. 344, no. 6191, hal. 1492–1496.

Rohmawati, N. (2017), "Implementasi Algoritma K-Means Pada Pengelompokan Pelamar Beasiswa", *Jurnal Ilmiah Teknologi Informasi Terapan*, Universitas Widyatama.

Sharma, S. (1996), *Applied Multivariate Technique,* John Wiley and Sons, Inc, New York.

Steinbach, M. Karypis, G. dan Kumar, V. (2000). "*A Comparison of Document Clustering Techniques*".[Home page of University of Minnesota, Technical Report #00-034][Online]. Available at: http///www.cs.umn.edu&tech_reports. Diakses pada tanggal 3 Juni 2018.

Su, M.C. (2003), "A New Index of Cluster Validity", http///www.cs.missouri.edu/-skubic/8820/KelompokValid.pdf, Tanggal akses: 28 Februari 2018.

Suguna, J. dan Selvi, M. A. (2012), "Ensemble *Fuzzy* Clustering for Mixed Numerical and Categorical Data", *International Journal of Computer Application*, vol 42, no 3.

Suwarsa, R. D. (2013), "Imlementasi K-Modes Pada Pengelompokan Data Kategori menggunakan *New Dissimilarity Measure*", Universitas Brawijaya, Malang.

Suyanto. (2017), *Data Mining untuk Klasifikasi dan Klasterisasi data*, Informatika Bandung, Bandung

Velmurugan, T. dan Santhanam, T. (2010), "Clustering Mixed Data Points using *Fuzzy* C-Means Clustering Algorithm for Performance Analysis", *International Journal on Computer Science and Engineering*, vol. 2, no. 9.

Yoon, H. S. Ahn, S. Y. Lee, S. H. Cho, S. B. dan Kim, J. H. (2005), "Heterogeneous Clustering Ensemble Method For Combining Different Cluster Results". *BioDM 2006,* hal 82-91.

“Halaman ini sengaja dikosongkan”

LAMPIRAN-LAMPIRAN

Lampiran 1 : *Preprocessing* Data

```
## First, Reading my original data namely: MY_DATA (on Downloads)
ori<-read.csv(file.choose(),header=TRUE)
str(ori)

## Replace missing value as #N/A
is.na(ori) <- ori == "#N/A"

## Periksa missing value pada dataset
sapply(ori, function(x) sum(is.na(x)))

## Creating my dataframe of numerical data
## PERIKSA NA dari dataset yang baru
## Membuat dataframe untuk variabel numerik (kolom 1:4 numerik)
ori[1:4]<- lapply(ori[1:4], as.numeric)
numerik1<-ori[1:4]
str(numerik1)
sapply(numerik1, function(x) sum(is.na(x)))
write.csv(numerik1,"E:/numerik1.csv")
numerik2<-rbind(numerik1[1:12158,])
str(numerik2)
write.csv(numerik2,"E:/numerik2.csv")

## membuat data frame untuk variabel kategorik (kolom 5:17 kategorik)
ori[5:17] <- lapply(ori[5:17], factor) ## not use as
categ1<-ori[5:17]
str(categ1)

## PERIKSA NA dari dataset yang baru
sapply(categ1, function(x) sum(is.na(x)))
write.csv(categ1,"E:/categ1.csv")
categ2<-ori[5:16]
str(categ2)
kategorik2<-rbind(ori[1:12158,5:16])
str(kategorik2)
write.csv(kategorik2,"E:/kategorik2.csv")

## Creating my mixed data
campurannonsc<-data.frame(numerik2,kategorik2)
str(campurannonsc)
write.csv(campurannonsc,"E:/campurannonsc.csv")

## Uji korelasi variabel numerik
```

```
tescor<-cor(numerik2,method = "pearson")
tescor


## Checking VIF value on numerik2
library(usdm)
vif(numerik2)## (a VIF value that exceeds 5 or 10 indicates a problematic amount of
collinearity (James et al. 2014)AMAN DARI MULTIKOLINEARITAS)


### Not For Run
## Handle NA
library(data.table)
library(VIM)
ori1<-kNN(ori,k=2)
str(ori1)

ori2<-kNN(ori(1:4),k=2)
str(ori1)
library(knncat)
ori3<-knncat(ori[5:17])
str(ori3)
real<-cbind(ori1[1:17])
str(real)

## Not For Run
## Create my mixed data
campuransc<-data.frame(numerik2sc,kategorik2)
str(campuransc)
write.csv(campuran,"E:/campuransc.csv")
```

Output Lampiran 1

```
> ## First, Reading my original data namely: MY_DATA (on Downlo
ads)
> ori<-read.csv(file.choose(),header=TRUE)
> str(ori)
'data.frame':  55148 obs. of  17 variables:
 $ NILAI_SMT_4      : int  1161 1214 1256 1031 1335 1110 1106 1
264 1057 1229 ...
 $ NILAI_SMT_5      : int  1145 1179 1255 1037 1333 1091 1114 1
269 1069 1203 ...
 $ JARAK_KOTA       : num  31 24 31 24 24 31 24 27 24 24 ...
 $ JUMLAH_TANGGUNGAN: int  2 3 6 2 6 7 6 2 6 2 ...
 $ LUAS_TANAH       : int  2 1 2 2 2 4 3 2 1 1 ...
 $ LUAS_BANGUNAN    : int  3 3 2 2 3 4 3 2 3 3 ...
 $ KODE_MCK         : int  1 1 1 1 1 1 1 1 1 2 ...
 $ SUMBER_AIR       : int  1 1 1 1 1 1 1 1 1 1 ...
 $ PENGHASILAN_AYAH : int  3 1 6 3 3 3 3 5 3 1 ...
```

```
 $ PENGHASILAN_IBU  : int  1 3 1 1 1 3 3 1 2 3 ...
 $ KERJA_AYAH       : int  5 8 2 7 7 7 7 7 3 8 ...
 $ KERJA_IBU        : int  5 7 8 8 8 7 7 8 7 3 ...
 $ PENDIDIKAN_AYAH  : int  4 8 3 2 2 3 3 2 4 3 ...
 $ PENDIDIKAN_IBU   : int  2 4 3 4 2 2 2 2 4 4 ...
 $ KODE_KEPEMILIKAN : int  1 4 1 4 1 4 1 1 4 1 ...
 $ KODE_LISTRIK     : int  1 1 1 1 1 1 1 1 1 1 ...
 $ ID_TAHAPAN       : int  3 1 1 1 3 1 3 1 1 3 ...
> ## Replace missing value as #N/A
> is.na(ori) <- ori == "#N/A"
> ## Periksa missing value pada dataset
> sapply(ori, function(x) sum(is.na(x)))
      NILAI_SMT_4       NILAI_SMT_5        JARAK_KOTA JUMLAH_TA
NGGUNGAN
                0                 0                 0
0
       LUAS_TANAH     LUAS_BANGUNAN         KODE_MCK         SU
MBER_AIR
                0                 0                 0
0
 PENGHASILAN_AYAH   PENGHASILAN_IBU        KERJA_AYAH          K
ERJA_IBU
                0                 0                 0
0
  PENDIDIKAN_AYAH    PENDIDIKAN_IBU  KODE_KEPEMILIKAN       KODE
_LISTRIK
                0                 0                 0
0
       ID_TAHAPAN
                0
> ## Creating my dataframe of numerical data
> ## PERIKSA NA dari dataset yang baru
> ## Membuat dataframe untuk variabel numerik (kolom 1:4 numeri
k)
> ori[1:4]<- lapply(ori[1:4], as.numeric)
> numerik1<-ori[1:4]
> str(numerik1)
'data.frame':  55148 obs. of  4 variables:
 $ NILAI_SMT_4      : num  1161 1214 1256 1031 1335 ...
 $ NILAI_SMT_5      : num  1145 1179 1255 1037 1333 ...
 $ JARAK_KOTA       : num  31 24 31 24 24 31 24 27 24 24 ...
 $ JUMLAH_TANGGUNGAN: num  2 3 6 2 6 7 6 2 6 2 ...
> sapply(numerik1, function(x) sum(is.na(x)))
      NILAI_SMT_4       NILAI_SMT_5        JARAK_KOTA JUMLAH_TA
NGGUNGAN
                0                 0                 0
0
> write.csv(numerik1,"E:/numerik1.csv")
> numerik2<-rbind(numerik1[1:12158,])
> str(numerik2)
'data.frame':  12158 obs. of  4 variables:
 $ NILAI_SMT_4      : num  1161 1214 1256 1031 1335 ...
 $ NILAI_SMT_5      : num  1145 1179 1255 1037 1333 ...
 $ JARAK_KOTA       : num  31 24 31 24 24 31 24 27 24 24 ...
 $ JUMLAH_TANGGUNGAN: num  2 3 6 2 6 7 6 2 6 2 ...
> write.csv(numerik2,"E:/numerik2.csv")
```

```
> ## membuat data frame untuk variabel kategorik (kolom 5:17 ka
tegorik)
> ori[5:17] <- lapply(ori[5:17], factor) ## not use as
> categ1<-ori[5:17]
> str(categ1)
'data.frame':  55148 obs. of  13 variables:
 $ LUAS_TANAH      : Factor w/ 5 levels "1","2","3","4",..: 2 1
2 2 2 4 3 2 1 1 ...
 $ LUAS_BANGUNAN   : Factor w/ 5 levels "1","2","3","4",..: 3 3
2 2 3 4 3 2 3 3 ...
 $ KODE_MCK        : Factor w/ 3 levels "1","2","3": 1 1 1 1 1
1 1 1 1 2 ...
 $ SUMBER_AIR      : Factor w/ 4 levels "1","2","3","4": 1 1 1
1 1 1 1 1 1 1 ...
 $ PENGHASILAN_AYAH: Factor w/ 25 levels "1","2","3","4",..: 3
1 6 3 3 3 3 5 3 1 ...
 $ PENGHASILAN_IBU : Factor w/ 21 levels "1","2","3","4",..: 1
3 1 1 1 3 3 1 2 3 ...
 $ KERJA_AYAH      : Factor w/ 8 levels "1","2","3","4",..: 5 8
2 7 7 7 7 7 3 8 ...
 $ KERJA_IBU       : Factor w/ 8 levels "1","2","3","4",..: 5 7
8 8 8 7 7 8 7 3 ...
 $ PENDIDIKAN_AYAH : Factor w/ 10 levels "1","2","3","4",..: 4
8 3 2 2 3 3 2 4 3 ...
 $ PENDIDIKAN_IBU  : Factor w/ 10 levels "1","2","3","4",..: 2
4 3 4 2 2 2 2 4 4 ...
 $ KODE_KEPEMILIKAN: Factor w/ 6 levels "1","2","3","4",..: 1 4
1 4 1 4 1 1 4 1 ...
 $ KODE_LISTRIK    : Factor w/ 5 levels "1","2","3","4",..: 1 1
1 1 1 1 1 1 1 1 ...
 $ ID_TAHAPAN      : Factor w/ 3 levels "1","2","3": 3 1 1 1 3
1 3 1 1 3 ...
> ## PERIKSA NA dari dataset yang baru
> sapply(categ1, function(x) sum(is.na(x)))
      LUAS_TANAH    LUAS_BANGUNAN         KODE_MCK        SUMBER
_AIR
               0                0                0
0
PENGHASILAN_AYAH  PENGHASILAN_IBU       KERJA_AYAH        KERJA
_IBU
               0                0                0
0
 PENDIDIKAN_AYAH   PENDIDIKAN_IBU KODE_KEPEMILIKAN     KODE_LIS
TRIK
               0                0                0
0
      ID_TAHAPAN
               0
> write.csv(categ1,"E:/categ1.csv")
> categ2<-ori[5:16]
> str(categ2)
'data.frame':  55148 obs. of  12 variables:
 $ LUAS_TANAH      : Factor w/ 5 levels "1","2","3","4",..: 2 1
2 2 2 4 3 2 1 1 ...
 $ LUAS_BANGUNAN   : Factor w/ 5 levels "1","2","3","4",..: 3 3
2 2 3 4 3 2 3 3 ...
```

```
 $ KODE_MCK         : Factor w/ 3 levels "1","2","3": 1 1 1 1 1
1 1 1 1 2 ...
 $ SUMBER_AIR       : Factor w/ 4 levels "1","2","3","4": 1 1 1
1 1 1 1 1 1 1 ...
 $ PENGHASILAN_AYAH: Factor w/ 25 levels "1","2","3","4",..: 3
1 6 3 3 3 3 5 3 1 ...
 $ PENGHASILAN_IBU : Factor w/ 21 levels "1","2","3","4",..: 1
3 1 1 1 3 3 1 2 3 ...
 $ KERJA_AYAH       : Factor w/ 8 levels "1","2","3","4",..: 5 8
2 7 7 7 7 7 3 8 ...
 $ KERJA_IBU        : Factor w/ 8 levels "1","2","3","4",..: 5 7
8 8 8 7 7 8 7 3 ...
 $ PENDIDIKAN_AYAH : Factor w/ 10 levels "1","2","3","4",..: 4
8 3 2 2 3 3 2 4 3 ...
 $ PENDIDIKAN_IBU  : Factor w/ 10 levels "1","2","3","4",..: 2
4 3 4 2 2 2 2 4 4 ...
 $ KODE_KEPEMILIKAN: Factor w/ 6 levels "1","2","3","4",..: 1 4
1 4 1 4 1 1 4 1 ...
 $ KODE_LISTRIK     : Factor w/ 5 levels "1","2","3","4",..: 1 1
1 1 1 1 1 1 1 1 ...
> kategorik2<-rbind(ori[1:12158,5:16])
> str(kategorik2)
'data.frame':  12158 obs. of  12 variables:
 $ LUAS_TANAH       : Factor w/ 5 levels "1","2","3","4",..: 2 1
2 2 2 4 3 2 1 1 ...
 $ LUAS_BANGUNAN    : Factor w/ 5 levels "1","2","3","4",..: 3 3
2 2 3 4 3 2 3 3 ...
 $ KODE_MCK         : Factor w/ 3 levels "1","2","3": 1 1 1 1 1
1 1 1 1 2 ...
 $ SUMBER_AIR       : Factor w/ 4 levels "1","2","3","4": 1 1 1
1 1 1 1 1 1 1 ...
 $ PENGHASILAN_AYAH: Factor w/ 25 levels "1","2","3","4",..: 3
1 6 3 3 3 3 5 3 1 ...
 $ PENGHASILAN_IBU : Factor w/ 21 levels "1","2","3","4",..: 1
3 1 1 1 3 3 1 2 3 ...
 $ KERJA_AYAH       : Factor w/ 8 levels "1","2","3","4",..: 5 8
2 7 7 7 7 7 3 8 ...
 $ KERJA_IBU        : Factor w/ 8 levels "1","2","3","4",..: 5 7
8 8 8 7 7 8 7 3 ...
 $ PENDIDIKAN_AYAH : Factor w/ 10 levels "1","2","3","4",..: 4
8 3 2 2 3 3 2 4 3 ...
 $ PENDIDIKAN_IBU  : Factor w/ 10 levels "1","2","3","4",..: 2
4 3 4 2 2 2 2 4 4 ...
 $ KODE_KEPEMILIKAN: Factor w/ 6 levels "1","2","3","4",..: 1 4
1 4 1 4 1 1 4 1 ...
 $ KODE_LISTRIK     : Factor w/ 5 levels "1","2","3","4",..: 1 1
1 1 1 1 1 1 1 1 ...
> write.csv(kategorik2,"E:/kategorik2.csv")
> ## Creating my mixed data
> campurannonsc<-data.frame(numerik2,kategorik2)
> str(campurannonsc)
'data.frame':  12158 obs. of  16 variables:
 $ NILAI_SMT_4      : num  1161 1214 1256 1031 1335 ...
 $ NILAI_SMT_5      : num  1145 1179 1255 1037 1333 ...
 $ JARAK_KOTA       : num  31 24 31 24 24 31 24 27 24 24 ...
 $ JUMLAH_TANGGUNGAN: num  2 3 6 2 6 7 6 2 6 2 ...
```

```
 $ LUAS_TANAH       : Factor w/ 5 levels "1","2","3","4",..: 2
1 2 2 2 4 3 2 1 1 ...
 $ LUAS_BANGUNAN    : Factor w/ 5 levels "1","2","3","4",..: 3
3 2 2 3 4 3 2 3 3 ...
 $ KODE_MCK         : Factor w/ 3 levels "1","2","3": 1 1 1 1 1
1 1 1 1 2 ...
 $ SUMBER_AIR       : Factor w/ 4 levels "1","2","3","4": 1 1 1
1 1 1 1 1 1 1 ...
 $ PENGHASILAN_AYAH : Factor w/ 25 levels "1","2","3","4",..: 3
1 6 3 3 3 3 5 3 1 ...
 $ PENGHASILAN_IBU  : Factor w/ 21 levels "1","2","3","4",..: 1
3 1 1 1 3 3 1 2 3 ...
 $ KERJA_AYAH       : Factor w/ 8 levels "1","2","3","4",..: 5
8 2 7 7 7 7 7 3 8 ...
 $ KERJA_IBU        : Factor w/ 8 levels "1","2","3","4",..: 5
7 8 8 8 7 7 8 7 3 ...
 $ PENDIDIKAN_AYAH  : Factor w/ 10 levels "1","2","3","4",..: 4
8 3 2 2 3 3 2 4 3 ...
 $ PENDIDIKAN_IBU   : Factor w/ 10 levels "1","2","3","4",..: 2
4 3 4 2 2 2 2 4 4 ...
 $ KODE_KEPEMILIKAN : Factor w/ 6 levels "1","2","3","4",..: 1
4 1 4 1 4 1 1 4 1 ...
 $ KODE_LISTRIK     : Factor w/ 5 levels "1","2","3","4",..: 1
1 1 1 1 1 1 1 1 1 ...
> write.csv(campurannonsc,"E:/campurannonsc.csv")
> tescor<-cor(numerik2,method = "pearson")
> tescor
                  NILAI_SMT_4  NILAI_SMT_5   JARAK_KOTA JUMLAH_
TANGGUNGAN
NILAI_SMT_4       1.000000000 0.9489481559 0.0031664275
0.01971831
NILAI_SMT_5       0.948948156 1.0000000000 0.0001892177
0.02918483
JARAK_KOTA        0.003166428 0.0001892177 1.0000000000
0.01978493
JUMLAH_TANGGUNGAN 0.019718308 0.0291848332 0.0197849343
1.00000000
> ## Checking VIF value on numerik2
> library(usdm)
Loading required package: sp
Loading required package: raster

Attaching package: ‘raster’

The following object is masked from ‘package:philentropy’:

    distance

The following objects are masked from ‘package:MASS’:

    area, select

The following object is masked from ‘package:e1071’:

    interpolate

The following object is masked from ‘package:data.table’:
```

```
    Shift

> vif(numerik2)## (a VIF value that exceeds 5 or 10 indicates a
problematic amount of collinearity (James et al. 2014)AMAN DARI
MULTIKOLINEARITAS)
          Variables      VIF
1       NILAI_SMT_4 9.744578
2       NILAI_SMT_5 9.762070
3        JARAK_KOTA 1.000585
4 JUMLAH_TANGGUNGAN 1.006529
```

Lampiran 2 : Metode Ensembel *Fuzzy*

```
# Algoritma ensemble fuzzy
set.seed(1234)
library(e1071)
x<-numerik2
str(numerik2)
fcm <- cmeans(x, 3)

## Calculate distance of numerik2
library(philentropy)
euc_dist<-dist(numerik2,"euclidean")

## Calculate cluster (FCModes)
library(klaR)
fcmodes<-kmodes(kategorik2, modes=3, iter.max = 10, weighted = FALSE)


## ENSEMBEL FUZZY
library(clue)
ensembel<-as.integer(fcm$cluster,fcmodes$cluster)
ensembeldf<-data.frame(ensembel)


## Find all statistic cluster
library(fpc)
clust_statsfuzzy<- cluster.stats(euc_dist,ensembel)
clust_statsfuzzy

## Visualize cluster
# Plot Silhouette
library(factoextra)
library(cluster)
library(NbClust)
silfuzzy<-silhouette(ensembel,euc_dist)
fviz_silhouette(silfuzzy)

## deskripsi hasil cluster
hslsplitfuzzynum<-split(numerik2, ensembel)
summary(hslsplitfuzzynum$`1`)
summary(hslsplitfuzzynum$`2`)
summary(hslsplitfuzzynum$`3`)
hslsplitfuzzykat<-split(kategorik2,ensembel)
summary(hslsplitfuzzykat$`1`)
summary(hslsplitfuzzykat$`2`)
summary(hslsplitfuzzykat$`3`)

### not for run
```

```
##  buat dummy
library(cba)
kategorik2dum<-as.dummy(kategorik2)
str(kategorik2dum)

# Calculate distance
library(philentropy)
getDistMethods()
euc_dist<-dist(numerik2, method = "euclidean")
head(euc_dist)

## calculate ditsnce categorical value
str(kategorik2dum)
library(philentropy)
my_distkate <- dist(kategorik2dum, method="dice")
head(my_distkate)
```

Output Sintax Lampiran 2

```
> ## Algoritma ensemble fuzzy
> set.seed(1234)
> library(e1071)
> x<-numerik2
> str(numerik2)
'data.frame':  12158 obs. of  4 variables:
 $ NILAI_SMT_4      : num  1161 1214 1256 1031 1335 ...
 $ NILAI_SMT_5      : num  1145 1179 1255 1037 1333 ...
 $ JARAK_KOTA       : num  31 24 31 24 24 31 24 27 24 24 ...
 $ JUMLAH_TANGGUNGAN: num  2 3 6 2 6 7 6 2 6 2 ...
> fcm <- cmeans(x, 3)
> ## Calculate distance of numerik2
> library(philentropy)
> euc_dist<-dist(numerik2,"euclidean")
> ## Calculate cluster (FCModes)
> library(klaR)
> fcmodes<-kmodes(kategorik2, modes=3, iter.max = 10, weighted
= FALSE)
> ## ENSEMBEL FUZZY
> library(clue)
> ensembel<-as.integer(fcm$cluster,fcmodes$cluster)
> ensembeldf<-data.frame(ensembel)
> ## Find all statistic cluster
> library(fpc)
> clust_statsfuzzy<- cluster.stats(euc_dist,ensembel)
> clust_statsfuzzy
$`n`
[1] 12158

$cluster.number
[1] 3

$cluster.size
[1] 2358 8781 1019
```

```
$min.cluster.size
[1] 1019

$noisen
[1] 0

$diameter
[1] 3774.404 1836.457 1534.678

$average.distance
[1] 441.6805 178.6704 146.4628

$median.distance
[1] 331.40760 154.74495  83.36666

$separation
[1]  4.123106  4.123106 98.574845

$average.toother
[1]  862.2263  931.8748 1580.3914

$separation.matrix
            [,1]      [,2]       [,3]
[1,]    0.000000  4.123106 1057.48617
[2,]    4.123106  0.000000   98.57484
[3,] 1057.486170 98.574845    0.00000

$ave.between.matrix
          [,1]      [,2]     [,3]
[1,]    0.0000  715.2535 2128.731
[2,]  715.2535    0.0000 1433.144
[3,] 2128.7307 1433.1435    0.000

$average.between
[1] 1021.586

$average.within
[1] 195.7371

$n.between
[1] 32056239

$n.within
[1] 41846164

$max.diameter
[1] 3774.404

$min.separation
[1] 4.123106

$within.cluster.ss
[1] 707477124

$clus.avg.silwidths
        1         2         3
```

```
0.3375860 0.7329178 0.8846220

$avg.silwidth
[1] 0.6689594

$g2
NULL

$g3
NULL

$pearsongamma
[1] 0.7040567

$Dunn
[1] 0.001092386

$Dunn2
[1] 1.619391

$entropy
[1] 0.7609073

$wb.ratio
[1] 0.1916012

$ch
[1] 27592.36

$cwidegap
[1] 1506.0186  726.3009  525.4684

$widestgap
[1] 1506.019

$sindex
[1] 56.96284

$corrected.rand
NULL

$vi
NULL

> ## Visualize cluster
> # Plot Silhouette
> library(factoextra)
> library(cluster)
> library(NbClust)
> silfuzzy<-silhouette(ensembel,euc_dist)
> fviz_silhouette(silfuzzy)
  cluster size ave.sil.width
1       1 2358          0.34
2       2 8781          0.73
3       3 1019          0.88
> ## deskripsi hasil cluster
```

```
> hslsplitfuzzynum<-split(numerik2, ensembel)
> summary(hslsplitfuzzynum$`1`)
  NILAI_SMT_4     NILAI_SMT_5      JARAK_KOTA      JUMLAH_TANGGUNG
AN
 Min.   :1007   Min.   :1055   Min.   :  0.10   Min.   : 2.000
 1st Qu.:1439   1st Qu.:1416   1st Qu.:  8.00   1st Qu.: 2.000
 Median :1562   Median :1519   Median : 18.80   Median : 3.000
 Mean   :1646   Mean   :1592   Mean   : 21.64   Mean   : 3.103
 3rd Qu.:1775   3rd Qu.:1684   3rd Qu.: 30.00   3rd Qu.: 4.000
 Max.   :4198   Max.   :3865   Max.   :600.00   Max.   :65.000
> summary(hslsplitfuzzynum$`2`)
  NILAI_SMT_4     NILAI_SMT_5      JARAK_KOTA      JUMLAH_TANGGUNG
AN
 Min.   : 116   Min.   : 109   Min.   :  0.05   Min.   : 2.000
 1st Qu.:1050   1st Qu.:1069   1st Qu.: 10.00   1st Qu.: 2.000
 Median :1107   Median :1122   Median : 18.00   Median : 3.000
 Mean   :1130   Mean   :1128   Mean   : 21.42   Mean   : 3.075
 3rd Qu.:1193   3rd Qu.:1184   3rd Qu.: 30.00   3rd Qu.: 4.000
 Max.   :1651   Max.   :1596   Max.   :600.00   Max.   :51.000
> summary(hslsplitfuzzynum$`3`)
  NILAI_SMT_4      NILAI_SMT_5      JARAK_KOTA      JUMLAH_TANGGU
NGAN
 Min.   :   0.0   Min.   :   0   Min.   :  0.50   Min.   : 2.00
 1st Qu.:  79.0   1st Qu.:  80   1st Qu.: 12.00   1st Qu.: 2.00
 Median :  91.0   Median :  92   Median : 18.00   Median : 3.00
 Mean   : 123.1   Mean   : 115   Mean   : 22.88   Mean   : 2.99
 3rd Qu.: 137.5   3rd Qu.: 131   3rd Qu.: 30.00   3rd Qu.: 4.00
 Max.   :1105.0   Max.   :1103   Max.   :200.00   Max.   :16.00
> hslsplitfuzzykat<-split(kategorik2,ensembel)
> summary(hslsplitfuzzykat$`1`)
 LUAS_TANAH LUAS_BANGUNAN KODE_MCK SUMBER_AIR PENGHASILAN_AYAH
PENGHASILAN_IBU
 1:643      1: 74         1:1590   1:1807     3      :525
1      :1057
 2:618      2:292         2: 542   2: 195     5      :459
2      : 541
 3:537      3:974         3: 226   3: 351     4      :457
3      : 404
 4:349      4:635                  4:   5     2      :262
4      : 175
 5:211      5:383                             1      :208
5      :  96
                                              6      :172
6      :  25
                                              (Other):275
(Other):  60
   KERJA_AYAH    KERJA_IBU   PENDIDIKAN_AYAH PENDIDIKAN_IBU KOD
E_KEPEMILIKAN
 7      :800   8      :833   2      :1027    2      :1092   1:1
836
 5      :789   7      :746   4      : 695    4      : 550   2:
67
 2      :304   5      :430   3      : 455    3      : 544   3:
18
 3      :260   3      :220   1      :  83    1      :  93   4:
406
```

```
 8       :139   2       :117   8       :  74    8       :  57   5:
30
 1       : 30   1       : 12   7       :  18    7       :  13   6:
1
 (Other): 36   (Other):  0   (Other):   6    (Other):   9
 KODE_LISTRIK
 1:2302
 2:  26
 3:   0
 4:   4
 5:  26


> summary(hslsplitfuzzykat$`2`)
 LUAS_TANAH LUAS_BANGUNAN KODE_MCK SUMBER_AIR PENGHASILAN_AYAH
PENGHASILAN_IBU
 1:3002     1: 305        1:6110   1:6498     3       :1885
1       :4381
 2:2461     2:1197        2:1957   2: 932     4       :1829
2       :1605
 3:1704     3:4031        3: 714   3:1338     5       :1686
3       :1415
 4: 979     4:2125                 4:  13     1       : 914
4       : 623
 5: 635     5:1123                            6       : 598
5       : 397
                                              2       : 585
6       : 116
                                              (Other):1284
(Other): 244
   KERJA_AYAH     KERJA_IBU    PENDIDIKAN_AYAH PENDIDIKAN_IBU K
ODE_KEPEMILIKAN
 7       :2982   8       :3558   2       :3560    2       :3720   1
:6580
 5       :2663   7       :2668   4       :2814    4       :2339   2
: 251
 2       :1236   5       :1337   3       :1728    3       :2139   3
:  49
 3       :1013   3       : 783   8       : 313    1       : 263   4
:1816
 8       : 582   2       : 381   1       : 235    8       : 243   5
:  82
 1       : 189   1       :  51   7       :  91    7       :  46   6
:   3
 (Other): 116   (Other):   3   (Other):  40    (Other):  31
 KODE_LISTRIK
 1:8680
 2:   5
 3:   1
 4:   9
 5:  86
```

```
> summary(hslsplitfuzzykat$`3`)
 LUAS_TANAH LUAS_BANGUNAN KODE_MCK SUMBER_AIR PENGHASILAN_AYAH
PENGHASILAN_IBU
 1:150      1: 42         1:713    1:808      5      :232
1      :428
 2:281      2:107         2:215    2: 93      3      :199
2      :221
 3:355      3:427         3: 91    3:117      4      :187
3      :186
 4:154      4:306                  4:  1      1      : 84
4      :104
 5: 79      5:137                             6      : 80
5      : 46
                                              2      : 74
6      : 11
                                              (Other):163
(Other): 23
   KERJA_AYAH    KERJA_IBU   PENDIDIKAN_AYAH PENDIDIKAN_IBU KOD
E_KEPEMILIKAN
 7      :357   7      :394   2      :379     2      :396    1:8
34
 5      :286   8      :358   4      :327     3      :277    2:
30
 2      :190   5      :138   3      :222     4      :269    3:
12
 3      : 79   3      : 73   1      : 41     1      : 36    4:1
31
 8      : 54   2      : 46   8      : 41     8      : 32    5:
11
 6      : 34   1      :  7   7      :  5     7      :  6    6:
1
 (Other): 19   (Other):  3   (Other):  4     (Other):  3
 KODE_LISTRIK
 1:1012
 2:   2
 3:   0
 4:   1
 5:   4
```


```
hslsplitfuzzynum<-split(numerik2, ensembel)
> summary(hslsplitfuzzynum$`1`)
  NILAI_SMT_4      NILAI_SMT_5     JARAK_KOTA      JUMLAH_TANGGUNGAN
 Min.   :   0.0   Min.   :   0   Min.   :  0.50   Min.   : 2.00
 1st Qu.:  79.0   1st Qu.:  80   1st Qu.: 12.00   1st Qu.: 2.00
 Median :  91.0   Median :  92   Median : 18.00   Median : 3.00
 Mean   : 123.1   Mean   : 115   Mean   : 22.88   Mean   : 2.99
 3rd Qu.: 137.5   3rd Qu.: 131   3rd Qu.: 30.00   3rd Qu.: 4.00
 Max.   :1105.0   Max.   :1103   Max.   :200.00   Max.   :16.00
> summary(hslsplitfuzzynum$`2`)
  NILAI_SMT_4    NILAI_SMT_5     JARAK_KOTA      JUMLAH_TANGGUNGAN
 Min.   :1007   Min.   :1055   Min.   :  0.10   Min.   : 2.000
 1st Qu.:1439   1st Qu.:1416   1st Qu.:  8.00   1st Qu.: 2.000
 Median :1562   Median :1519   Median : 18.80   Median : 3.000
 Mean   :1646   Mean   :1592   Mean   : 21.64   Mean   : 3.103
 3rd Qu.:1775   3rd Qu.:1684   3rd Qu.: 30.00   3rd Qu.: 4.000
 Max.   :4198   Max.   :3865   Max.   :600.00   Max.   :65.000
> summary(hslsplitfuzzynum$`3`)
  NILAI_SMT_4    NILAI_SMT_5     JARAK_KOTA      JUMLAH_TANGGUNGAN
 Min.   : 116   Min.   : 109   Min.   :  0.05   Min.   : 2.000
 1st Qu.:1050   1st Qu.:1069   1st Qu.: 10.00   1st Qu.: 2.000
 Median :1107   Median :1122   Median : 18.00   Median : 3.000
 Mean   :1130   Mean   :1128   Mean   : 21.42   Mean   : 3.075
 3rd Qu.:1193   3rd Qu.:1184   3rd Qu.: 30.00   3rd Qu.: 4.000
 Max.   :1651   Max.   :1596   Max.   :600.00   Max.   :51.000
> hslsplitfuzzykat<-split(kategorik2,ensembel)
> summary(hslsplitfuzzykat$`1`)
 LUAS_TANAH LUAS_BANGUNAN KODE_MCK SUMBER_AIR PENGHASILAN_AYAH PENGHASILAN_IBU
 1:150      1: 42         1:713    1:808      5      :232      1      :428
 2:281      2:107         2:215    2: 93      3      :199      2      :221
 3:355      3:427         3: 91    3:117      4      :187      3      :186
 4:154      4:306                   4:  1      1      : 84      4      :104
 5: 79      5:137                              6      : 80      5      : 46
                                               2      : 74      6      : 11
```

```
                                                (Other):163      (Other): 23
   KERJA_AYAH    KERJA_IBU   PENDIDIKAN_AYAH PENDIDIKAN_IBU KODE_KEPEMILIKAN
 7       :357   7       :394   2       :379     2       :396    1:834
 5       :286   8       :358   4       :327     3       :277    2: 30
 2       :190   5       :138   3       :222     4       :269    3: 12
 3       : 79   3       : 73   1       : 41     1       : 36    4:131
 8       : 54   2       : 46   8       : 41     8       : 32    5: 11
 6       : 34   1       :  7   7       :  5     7       :  6    6:  1
 (Other): 19   (Other):  3   (Other):  4     (Other):  3
 KODE_LISTRIK
 1:1012
 2:   2
 3:   0
 4:   1
 5:   4


> summary(hslsplitfuzzykat$`2`)
 LUAS_TANAH LUAS_BANGUNAN KODE_MCK SUMBER_AIR PENGHASILAN_AYAH PENGHASILAN_IBU
 1:643      1: 74         1:1590   1:1807     3       :525      1       :1057
 2:618      2:292         2: 542   2: 195     5       :459      2       : 541
 3:537      3:974         3: 226   3: 351     4       :457      3       : 404
 4:349      4:635                  4:   5     2       :262      4       : 175
 5:211      5:383                             1       :208      5       :  96
                                              6       :172      6       :  25
                                              (Other):275       (Other):  60

KERJA_AYAH    KERJA_IBU   PENDIDIKAN_AYAH PENDIDIKAN_IBU KODE_KEPEMILIKAN
 7       :800   8       :833   2       :1027    2       :1092   1:1836
 5       :789   7       :746   4       : 695    4       : 550   2:  67
 2       :304   5       :430   3       : 455    3       : 544   3:  18
 3       :260   3       :220   1       :  83    1       :  93   4: 406
 8       :139   2       :117   8       :  74    8       :  57   5:  30
 1       : 30   1       : 12   7       :  18    7       :  13   6:   1
 (Other): 36   (Other):  0   (Other):   6     (Other):   9
 KODE_LISTRIK
 1:2302
 2:  26
 3:   0
 4:   4
 5:  26


> summary(hslsplitfuzzykat$`3`)
 LUAS_TANAH LUAS_BANGUNAN KODE_MCK SUMBER_AIR PENGHASILAN_AYAH PENGHASILAN_IBU
 1:3002     1: 305        1:6110   1:6498     3       :1885     1       :4381
 2:2461     2:1197        2:1957   2: 932     4       :1829     2       :1605
 3:1704     3:4031        3: 714   3:1338     5       :1686     3       :1415
 4: 979     4:2125                 4:  13     1       : 914     4       : 623
 5: 635     5:1123                            6       : 598     5       : 397
                                              2       : 585     6       : 116
                                              (Other):1284      (Other): 244

KERJA_AYAH     KERJA_IBU    PENDIDIKAN_AYAH PENDIDIKAN_IBU KODE_KEPEMILIKAN
 7       :2982   8       :3558   2       :3560    2       :3720   1:6580
 5       :2663   7       :2668   4       :2814    4       :2339   2: 251
```

```
  2      :1236   5      :1337   3      :1728    3      :2139   3:  49
  3      :1013   3      : 783   8      : 313    1      : 263   4:1816
  8      : 582   2      : 381   1      : 235    8      : 243   5:  82
  1      : 189   1      :  51   7      :  91    7      :  46   6:   3
  (Other): 116   (Other):   3   (Other):  40    (Other):  31
  KODE_LISTRIK
  1:8680
  2:   5
  3:   1
  4:   9
  5:  86
```

Lampiran 3: Metode K-*Prototypes*

```
> ## Algoritma K-prototypes
> ## calculate cluster using k-prototypes
> set.seed(1234)
> library(clustMixType)
> library(data.table)
data.table 1.11.4  Latest news: http://r-datatable.com
> str(campurannonsc)
'data.frame':  12158 obs. of  16 variables:
 $ NILAI_SMT_4       : num  1161 1214 1256 1031 1335 ...
 $ NILAI_SMT_5       : num  1145 1179 1255 1037 1333 ...
 $ JARAK_KOTA        : num  31 24 31 24 24 31 24 27 24 24 ...
 $ JUMLAH_TANGGUNGAN: num  2 3 6 2 6 7 6 2 6 2 ...
 $ LUAS_TANAH        : Factor w/ 5 levels "1","2","3","4",..: 2
1 2 2 2 4 3 2 1 1 ...
 $ LUAS_BANGUNAN     : Factor w/ 5 levels "1","2","3","4",..: 3
3 2 2 3 4 3 2 3 3 ...
 $ KODE_MCK          : Factor w/ 3 levels "1","2","3": 1 1 1 1 1
1 1 1 1 2 ...
 $ SUMBER_AIR        : Factor w/ 4 levels "1","2","3","4": 1 1 1
1 1 1 1 1 1 1 ...
 $ PENGHASILAN_AYAH : Factor w/ 25 levels "1","2","3","4",..: 3
1 6 3 3 3 3 5 3 1 ...
 $ PENGHASILAN_IBU  : Factor w/ 21 levels "1","2","3","4",..: 1
3 1 1 1 3 3 1 2 3 ...
 $ KERJA_AYAH        : Factor w/ 8 levels "1","2","3","4",..: 5
8 2 7 7 7 7 7 3 8 ...
 $ KERJA_IBU         : Factor w/ 8 levels "1","2","3","4",..: 5
7 8 8 8 7 7 8 7 3 ...
 $ PENDIDIKAN_AYAH  : Factor w/ 10 levels "1","2","3","4",..: 4
8 3 2 2 3 3 2 4 3 ...
 $ PENDIDIKAN_IBU   : Factor w/ 10 levels "1","2","3","4",..: 2
4 3 4 2 2 2 2 4 4 ...
 $ KODE_KEPEMILIKAN : Factor w/ 6 levels "1","2","3","4",..: 1
4 1 4 1 4 1 1 4 1 ...
 $ KODE_LISTRIK      : Factor w/ 5 levels "1","2","3","4",..: 1
1 1 1 1 1 1 1 1 1 ...
> Distkproto<- dist(campurannonsc, method = "euclidean")
> findclustkpro <- kproto(campurannonsc, k=3)
Estimated lambda: 135653.3

> resultsc<-cbind(campurannonsc,findclustkpro$cluster)
> head(findclustkpro$cluster)
[1] 2 2 1 2 1 2
> str(resultsc)
'data.frame':  12158 obs. of  17 variables:
 $ NILAI_SMT_4           : num  1161 1214 1256 1031 1335 ...
 $ NILAI_SMT_5           : num  1145 1179 1255 1037 1333 ...
 $ JARAK_KOTA            : num  31 24 31 24 24 31 24 27 24 24 ..
.
 $ JUMLAH_TANGGUNGAN     : num  2 3 6 2 6 7 6 2 6 2 ...
 $ LUAS_TANAH            : Factor w/ 5 levels "1","2","3","4",..
: 2 1 2 2 2 4 3 2 1 1 ...
 $ LUAS_BANGUNAN         : Factor w/ 5 levels "1","2","3","4",..
: 3 3 2 2 3 4 3 2 3 3 ...
 $ KODE_MCK              : Factor w/ 3 levels "1","2","3": 1 1 1
1 1 1 1 1 1 2 ...
 $ SUMBER_AIR            : Factor w/ 4 levels "1","2","3","4": 1
1 1 1 1 1 1 1 1 1 ...
 $ PENGHASILAN_AYAH      : Factor w/ 25 levels "1","2","3","4",..
.: 3 1 6 3 3 3 3 5 3 1 ...
```

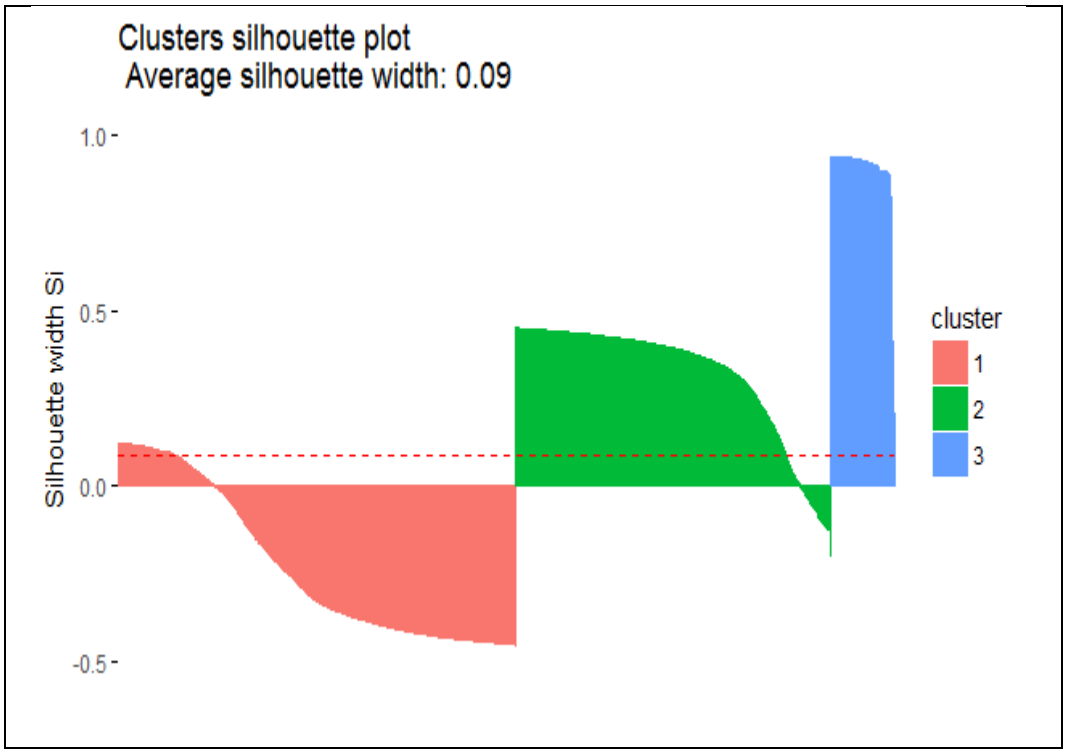
```
 $ PENGHASILAN_IBU      : Factor w/ 21 levels "1","2","3","4",..
.: 1 3 1 1 1 3 3 1 2 3 ...
 $ KERJA_AYAH           : Factor w/ 8 levels "1","2","3","4",..
: 5 8 2 7 7 7 7 3 8 ...
 $ KERJA_IBU            : Factor w/ 8 levels "1","2","3","4",..
: 5 7 8 8 8 7 7 8 7 3 ...
 $ PENDIDIKAN_AYAH      : Factor w/ 10 levels "1","2","3","4",..
.: 4 8 3 2 2 3 3 2 4 3 ...
 $ PENDIDIKAN_IBU       : Factor w/ 10 levels "1","2","3","4",..
.: 2 4 3 4 2 2 2 2 4 4 ...
 $ KODE_KEPEMILIKAN     : Factor w/ 6 levels "1","2","3","4",..
: 1 4 1 4 1 4 1 1 4 1 ...
 $ KODE_LISTRIK         : Factor w/ 5 levels "1","2","3","4",..
: 1 1 1 1 1 1 1 1 1 1 ...
 $ findclustkpro$cluster: int  2 2 1 2 1 2 2 1 2 2 ...
> ## Find all statistic cluster
> library(fpc)
> calcstatkpro <- cluster.stats(Distkproto, findclustkpro$clust
er)
> calcstatkpro
$`n`
[1] 12158

$cluster.number
[1] 3

$cluster.size
[1] 6226 4916 1016

$min.cluster.size
[1] 1016

$noisen
[1] 0

$diameter
[1] 4657.542 2262.417 1535.213

$average.distance
[1] 439.4150 261.9665 144.8399

$median.distance
[1] 310.42713 186.94652  84.88816

$separation
[1]  1.00000  1.00000 12.72792

$average.toother
[1]  590.7623  527.5198 1581.7605

$separation.matrix
          [,1]     [,2]      [,3]
[1,]   0.0000  1.00000 258.88608
[2,]   1.0000  0.00000  12.72792
[3,] 258.8861 12.72792   0.00000

$ave.between.matrix
          [,1]      [,2]     [,3]
[1,]    0.0000  370.7162 1655.474
[2,]  370.7162    0.0000 1488.405
[3,] 1655.4735 1488.4046    0.000

$average.between
```

```
[1] 697.6954

$average.within
[1] 367.6199

$n.between
[1] 41927288

$n.within
[1] 31975115

$max.diameter
[1] 4657.542

$min.separation
[1] 1

$within.cluster.ss
[1] 1529382029

$clus.avg.silwidths
         1          2          3
-0.2285604  0.3186025  0.8914161

$avg.silwidth
[1] 0.08627338

$g2
NULL

$g3
NULL

$pearsongamma
[1] 0.2816024

$Dunn
[1] 0.0002147055

$Dunn2
[1] 0.8436585

$entropy
[1] 0.916269

$wb.ratio
[1] 0.526906

$ch
[1] 9505.792

$cwidegap
[1] 1520.9970  709.2073  525.4950

$widestgap
[1] 1520.997

$sindex
[1] 5.204423

$corrected.rand
NULL
```

```
$vi
NULL

> ## Visualize cluster
> # Plot Silhouette
> library(factoextra)
Loading required package: ggplot2
Welcome! Related Books: `Practical Guide To Cluster Analysis in
R` at https://goo.gl/13EFCZ
> library(cluster)
> library(NbClust)
> sil2<-silhouette(findclustkpro$cluster, Distkproto)
> fviz_silhouette(sil2)
  cluster size ave.sil.width
1       1 6226         -0.23
2       2 4916          0.32
3       3 1016          0.89
> ## deskripsi hasil cluster
> hslsplitkproto<-split(campurannonsc, findclustkpro$cluster)
> summary(hslsplitkproto$`1`)
  NILAI_SMT_4    NILAI_SMT_5     JARAK_KOTA     JUMLAH_TANGGUNG
AN LUAS_TANAH
 Min.   : 798   Min.   : 109   Min.   :  0.05   Min.   : 2.000
1:2608
 1st Qu.:1076   1st Qu.:1097   1st Qu.: 11.00   1st Qu.: 2.000
2:1142
 Median :1188   Median :1182   Median : 20.00   Median : 3.000
3:1188
 Mean   :1294   Mean   :1275   Mean   : 23.72   Mean   : 3.083
4: 781
 3rd Qu.:1420   3rd Qu.:1377   3rd Qu.: 30.00   3rd Qu.: 4.000
5: 507
 Max.   :4198   Max.   :3865   Max.   :600.00   Max.   :25.000

 LUAS_BANGUNAN KODE_MCK SUMBER_AIR PENGHASILAN_AYAH PENGHASILAN
_IBU   KERJA_AYAH
 1: 262        1:3995   1:4656     4      :1718     1      :328
4    5      :3027
 2: 762        2:1627   2: 733     3      :1576     2      :127
4    7      :1324
 3:2689        3: 604   3: 829     5      : 797     3      : 92
9    3      : 677
 4:1578                 4:   8     2      : 611     4      : 40
0    2      : 651
 5: 935                            1      : 504     5      : 18
0    8      : 348
                                   6      : 366     6      :  5
3    1      :  94
                                   (Other): 654     (Other): 10
6    (Other): 105


KERJA_IBU    PENDIDIKAN_AYAH PENDIDIKAN_IBU KODE_KEPEMILIKAN KO
DE_LISTRIK
 8      :2852   2      :3776    2      :3893   1:5062
1:6118
 5      :1537   3      :1199    3      :1361   2:  86
2:  29
 7      :1128   4      : 814    4      : 595   3:  21
3:   0
 3      : 487   1      : 215    1      : 243   4:1001
4:   4
```

```
 2       : 204   8       : 162    8       : 102   5:  54
5:  75
 1       :  17   7       :  46    7       :  18   6:   2
 (Other):   1   (Other):  14    (Other):  14
> summary(hslsplitkproto$`2`)
  NILAI_SMT_4    NILAI_SMT_5     JARAK_KOTA      JUMLAH_TANGGUNG
AN LUAS_TANAH
 Min.   :  38   Min.   : 544   Min.   :  0.05   Min.   : 2.000
1:1037
 1st Qu.:1054   1st Qu.:1073   1st Qu.:  7.00   1st Qu.: 2.000
2:1939
 Median :1116   Median :1129   Median : 15.00   Median : 3.000
3:1053
 Mean   :1169   Mean   :1165   Mean   : 18.62   Mean   : 3.077
4: 547
 3rd Qu.:1235   3rd Qu.:1204   3rd Qu.: 25.00   3rd Qu.: 4.000
5: 340
 Max.   :2213   Max.   :2252   Max.   :600.00   Max.   :65.000

 LUAS_BANGUNAN KODE_MCK SUMBER_AIR PENGHASILAN_AYAH PENGHASILAN
_IBU   KERJA_AYAH
 1: 117        1:3708   1:3650     5       :1348     1       :215
4    7       :2460
 2: 728        2: 872   2: 394     3       : 834     3       : 89
0    2       : 889
 3:2317        3: 336   3: 862     1       : 619     2       : 87
2    3       : 596
 4:1182                 4:  10     4       : 570     4       : 40
0    5       : 425
 5: 572                            6       : 404     5       : 31
3    8       : 374
                                   7       : 398     6       :  8
8    1       : 125
                                   (Other): 743     (Other): 19
9    (Other):  47
   KERJA_IBU    PENDIDIKAN_AYAH PENDIDIKAN_IBU KODE_KEPEMILIKAN
KODE_LISTRIK
 7       :2288   4       :2697    4       :2297   1:3355
1:4867
 8       :1539   3       : 985    3       :1322   2: 232
2:   2
 3       : 517   2       : 811    2       : 919   3:  46
3:   1
 2       : 294   8       : 225    8       : 198   4:1223
4:   9
 5       : 230   1       : 103    1       : 113   5:  58
5:  37
 1       :  46   7       :  63    7       :  41   6:   2
 (Other):   2   (Other):  32    (Other):  26

> summary(hslsplitkproto$`3`)
  NILAI_SMT_4      NILAI_SMT_5       JARAK_KOTA      JUMLAH_TANG
GUNGAN LUAS_TANAH
 Min.   :   0.0   Min.   :   0.0   Min.   :  0.50   Min.   : 2.
00     1:150
 1st Qu.:  79.0   1st Qu.:  80.0   1st Qu.: 12.00   1st Qu.: 2.
00     2:279
 Median :  91.0   Median :  92.0   Median : 18.00   Median : 3.
00     3:355
 Mean   : 122.7   Mean   : 112.7   Mean   : 22.88   Mean   : 2.
99     4:154
 3rd Qu.: 137.2   3rd Qu.: 130.0   3rd Qu.: 30.00   3rd Qu.: 4.
00     5: 78
```

```
 Max.   :1105.0   Max.   :1078.0   Max.   :200.00   Max.   :16.
00

 LUAS_BANGUNAN KODE_MCK SUMBER_AIR PENGHASILAN_AYAH PENGHASILAN
_IBU   KERJA_AYAH
 1: 42         1:710    1:807      5      :232      1      :428
7      :355
 2:106         2:215    2: 93      3      :199      2      :221
5      :286
 3:426         3: 91    3:115      4      :185      3      :186
2      :190
 4:306                  4:  1      1      : 83      4      :102
3      : 79
 5:136                             6      : 80      5      : 46
8      : 53
                                   2      : 74      6      : 11
6      : 34
                                   (Other):163      (Other): 22
(Other): 19
   KERJA_IBU   PENDIDIKAN_AYAH PENDIDIKAN_IBU KODE_KEPEMILIKAN
KODE_LISTRIK
 7      :392   2      :379     2      :396    1:833
1:1009
 8      :358   4      :325     3      :277    2: 30
2:   2
 5      :138   3      :221     4      :266    3: 12
3:   0
 3      : 72   1      : 41     1      : 36    4:129
4:   1
 2      : 46   8      : 41     8      : 32    5: 11
5:   4
 1      :  7   7      :  5     7      :  6    6:  1
 (Other):  3   (Other):  4     (Other):  3
```


Clusters silhouette plot
Average silhouette width: 0.09
Silhouette width Si
1.0
0.5
0.0
-0.5
cluster
1
2
3

Lampiran 4: Metode DPC-M

```
## calculate cluster (DPC-M)
set.seed(1234)
str(campurannonsc)
library(densityClust)
setdistsc <- dist(campurannonsc, method = "euclidean")
head(setdistsc)
## Normalization Euclidean Distance
library(IntClust)
norm_eucdist<-Normalization(setdistsc,method="R")
setClustsc <- densityClust(setdistsc)
setClustsc <- findClusters(setClustsc, rho=48, delta=160)


## Find all statistic cluster
library(fpc)
calcstatdpc_m <- cluster.stats(setdistsc,setClustsc$clusters)
calcstatdpc_m

## Visualize cluster
# Plot Silhouette
library(factoextra)
library(cluster)
library(NbClust)
sil2sc<-silhouette(setClustsc$cluster, setdistsc)
fviz_silhouette(sil2sc)

## Deskripsi Hasil cluster
hslsplit<-split(campurannonsc, setClustsc$clusters)
hslsplit$`1`
summary(hslsplit$`1`)
summary(hslsplit$`2`)
summary(hslsplit$`3`)


## Not for Run
## Calculate index Dunn
library(clValid)
library(cluster)
Dunn(setdistsc,setClustsc$clusters)
plot(Dunn(setdistsc,setClustsc$clusters))

## find Silhouette Coefficient
library(cluster)
library(tools)
```

```
library(HSAUR)
library(fpc)
setClustsc$clusters
plot(silhouette(setClustsc$clusters, setdistsc))
```

Hasil Output

```
> ## calculate cluster (DPC-M)
> set.seed(1234)
> str(campurannonsc)
'data.frame':  12158 obs. of  16 variables:
 $ NILAI_SMT_4      : num  1161 1214 1256 1031 1335 ...
 $ NILAI_SMT_5      : num  1145 1179 1255 1037 1333 ...
 $ JARAK_KOTA       : num  31 24 31 24 24 31 24 27 24 24 ...
 $ JUMLAH_TANGGUNGAN: num  2 3 6 2 6 7 6 2 6 2 ...
 $ LUAS_TANAH       : Factor w/ 5 levels "1","2","3","4",..: 2 1 2
2 2 4 3 2 1 1 ...
 $ LUAS_BANGUNAN    : Factor w/ 5 levels "1","2","3","4",..: 3 3 2
2 3 4 3 2 3 3 ...
 $ KODE_MCK         : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1
1 1 1 2 ...
 $ SUMBER_AIR       : Factor w/ 4 levels "1","2","3","4": 1 1 1 1
1 1 1 1 1 1 ...
 $ PENGHASILAN_AYAH : Factor w/ 25 levels "1","2","3","4",..: 3 1
6 3 3 3 3 5 3 1 ...
 $ PENGHASILAN_IBU  : Factor w/ 21 levels "1","2","3","4",..: 1 3
1 1 1 3 3 1 2 3 ...
 $ KERJA_AYAH       : Factor w/ 8 levels "1","2","3","4",..: 5 8 2
7 7 7 7 7 3 8 ...
 $ KERJA_IBU        : Factor w/ 8 levels "1","2","3","4",..: 5 7 8
8 8 7 7 8 7 3 ...
 $ PENDIDIKAN_AYAH  : Factor w/ 10 levels "1","2","3","4",..: 4 8
3 2 2 3 3 2 4 3 ...
 $ PENDIDIKAN_IBU   : Factor w/ 10 levels "1","2","3","4",..: 2 4
3 4 2 2 2 2 4 4 ...
 $ KODE_KEPEMILIKAN : Factor w/ 6 levels "1","2","3","4",..: 1 4 1
4 1 4 1 1 4 1 ...
 $ KODE_LISTRIK     : Factor w/ 5 levels "1","2","3","4",..: 1 1 1
1 1 1 1 1 1 1 ...
> library(densityClust)
> setdistsc <- dist(campurannonsc, method = "euclidean")
> head(setdistsc)
[1]  63.76519 145.50258 169.24538 256.32401  74.62573  63.75735
> setClustsc <- densityClust(setdistsc)
Distance cutoff calculated to 24.91508
> setClustsc <- findClusters(setClustsc, rho=48, delta=160)
> ## Find all statistic cluster
> library(fpc)
> calcstatdpc_m <- cluster.stats(setdistsc,setClustsc$clusters)
> calcstatdpc_m
$`n`
[1] 12158

$cluster.number
[1] 3
```

```
$cluster.size
[1] 8582 1021 2555

$min.cluster.size
[1] 1021

$noisen
[1] 0

$diameter
[1] 1507.548 1729.062 3831.123

$average.distance
[1] 171.8998 153.3271 439.0809

$median.distance
[1] 150.11662  85.89529 326.25450

$separation
[1]   9.69536 203.59764   9.69536

$average.toother
[1]  902.3194 1579.7294  842.6907

$separation.matrix
          [,1]     [,2]      [,3]
[1,]   0.00000 203.5976   9.69536
[2,] 203.59764   0.0000 911.76532
[3,]   9.69536 911.7653   0.00000

$ave.between.matrix
          [,1]     [,2]      [,3]
[1,]    0.0000 1425.347  693.3129
[2,] 1425.3475    0.000 2098.2836
[3,]  693.3129 2098.284    0.0000

$average.between
[1] 996.0148

$average.within
[1] 193.1307

$n.between
[1] 33297887

$n.within
[1] 40604516

$max.diameter
[1] 3831.123

$min.separation
[1] 9.69536

$within.cluster.ss
[1] 716134205
```

```
$clus.avg.silwidths
        1         2         3
0.7373447 0.8786642 0.3156558

$avg.silwidth
[1] 0.6605946

$g2
NULL

$g3
NULL

$pearsongamma
[1] 0.687887

$Dunn
[1] 0.002530683

$Dunn2
[1] 1.579009

$entropy
[1] 0.7817195

$wb.ratio
[1] 0.1939034

$ch
[1] 27202.3

$cwidegap
[1]  788.4352  525.4950 1506.0378

$widestgap
[1] 1506.038

$sindex
[1] 56.52991

$corrected.rand
NULL

$vi
NULL

> ## Visualize cluster
> # Plot Silhouette
> library(factoextra)
> library(cluster)
> library(NbClust)
> sil2sc<-silhouette(setClustsc$cluster, setdistsc)
> fviz_silhouette(sil2sc)
  cluster size ave.sil.width
1       1 8582          0.74
2       2 1021          0.88
3       3 2555          0.32
```

```
> ## Deskripsi Hasil cluster
> hslsplit<-split(campurannonsc, setClustsc$clusters)
> hslsplit$`1`
      NILAI_SMT_4 NILAI_SMT_5 JARAK_KOTA JUMLAH_TANGGUNGAN LUAS_TA
NAH LUAS_BANGUNAN
1            1161        1145       31.00                   2
2             3
2            1214        1179       24.00                   3
1             3
3            1256        1255       31.00                   6
2             2
4            1031        1037       24.00                   2
2             2
5            1335        1333       24.00                   6
2             3
6            1110        1091       31.00                   7
4             4
7            1106        1114       24.00                   6
3             3
8            1264        1269       27.00                   2
2             2
9            1057        1069       24.00                   6
1             3
10           1229        1203       24.00                   2
1             3
11           1110        1147       30.00                   2
1             2
12           1150        1164       30.00                   5
3             3
13           1188        1206       23.00                   2
3             3
14           1170        1191       25.00                   5
2             2
15           1124        1144       34.00                   3
1             2
16           1123        1174       13.00                   2
2             4
17            984        1010       50.00                   3
1             4
18            954         985       50.00                   2
1             2
19            951         960       50.00                   6
1             3
20            955         973       50.00                   6
1             3
21           1007        1032       50.00                   2
1             3
22            967         990       50.00                   2
1             4
23            950         980       18.00                   5
2             3
24            989        1001       50.00                   3
1             4
25            956         985       50.00                   4
1             3
26           1017        1032       50.00                   8
1             3
```

27	969	997	50.00	3
1	3			
28	992	1001	50.00	3
2	3			
29	971	1000	36.00	5
1	4			
30	954	980	50.00	3
1	3			
31	974	1004	50.00	3
1	3			
32	980	988	50.00	4
1	3			
33	977	1008	50.00	5
1	3			
34	977	1011	35.00	3
1	2			
35	989	985	50.00	2
1	3			
36	975	1031	50.00	3
1	2			
37	971	989	50.00	4
2	4			
38	1008	1020	60.00	3
1	4			
39	986	999	42.00	2
2	3			
40	952	992	50.00	5
3	3			
41	1000	996	50.00	2
1	2			
42	967	982	50.00	5
4	3			
43	1009	1008	50.00	6
1	3			
44	949	964	50.00	2
1	3			
45	1080	1202	21.00	3
3	3			
46	1105	1327	23.00	2
2	3			
47	1210	1212	20.00	4
3	4			
48	1168	1208	21.00	6
3	3			
49	1047	1209	21.00	4
3	4			
50	1065	1196	20.00	2
3	4			
51	1051	1295	21.00	6
3	3			
52	1063	1220	22.00	3
3	4			
53	1095	1231	21.00	4
2	3			
54	1097	1328	22.00	6
3	4			

```
55              1210          1211        27.00                    3
3               3
56              1062          1206        21.00                    6
2               3
57              1072          1292        21.00                    6
3               3
58              1184          1202        20.00                    3
2               3
59              1191          1200        19.00                    2
3               3
60              1073          1167        17.00                    3
2               3
61              1075          1200        21.00                    5
2               3
62              1178          1205        18.00                    5
3               3
KODE_MCK SUMBER_AIR PENGHASILAN_AYAH PENGHASILAN_IBU KERJA_AYAH KE
RJA_IBU
1               1             1                    3                1
5          5
2               1             1                    1                3
8          7
3               1             1                    6                1
2          8
4               1             1                    3                1
7          8
5               1             1                    3                1
7          8
6               1             1                    3                3
7          7
7               1             1                    3                3
7          7
8               1             1                    5                1
7          8
9               1             1                    3                2
3          7
10              2             1                    1                3
8          3
11              1             1                    3                1
8          8
12              1             1                    7                1
7          8
13              1             1                    4                4
2          2
14              1             3                   45                1
2          8
15              1             1                    5                1
2          8
16              1             1                    3                1
7          7
17              1             1                    3                1
3          8
18              1             1                    3                3
7          7
19              3             1                    7                1
2          8
```

20	1	1	43	1
3	8			
21	1	1	3	1
5	8			
22	2	1	3	3
7	7			
23	1	3	4	4
5	7			
24	1	3	5	1
1	8			
25	1	2	3	3
5	3			
26	2	1	43	1
1	8			
27	1	1	3	2
2	7			
28	1	1	4	1
5	8			
29	1	1	5	1
5	8			
30	2	1	3	1
5	7			
31	1	1	6	1
3	8			
32	1	1	3	1
5	8			
33	1	1	2	2
3	3			
34	1	1	3	3
5	7			
35	1	1	1	3
8	8			
36	1	1	5	1
5	8			
37	1	1	4	2
3	7			
38	2	1	4	1
5	5			
39	2	3	3	3
7	3			
40	3	1	3	1
5	8			
41	3	1	3	2
5	5			
42	1	1	4	1
2	8			
43	1	3	3	48
7	1			
44	1	1	4	4
7	7			
45	3	2	4	1
5	8			
46	3	1	4	2
5	7			
47	3	1	3	2
5	5			

48	3	1	2	2	5
5					
49	3	1	2	2	5
5					
50	3	2	3	2	5
5					
51	3	1	1	2	8
5					
52	3	2	3	2	5
5					
53	3	1	4	2	5
5					
54	3	1	2	2	5
5					
55	3	1	3	1	5
8					
56	3	1	3	1	5
8					
57	3	1	2	2	5
7					
58	3	2	2	2	5
5					
59	3	1	2	2	5
5					
60	2	1	5	1	7
8					
61	3	1	3	2	7
7					
62	3	1	3	1	5
7					

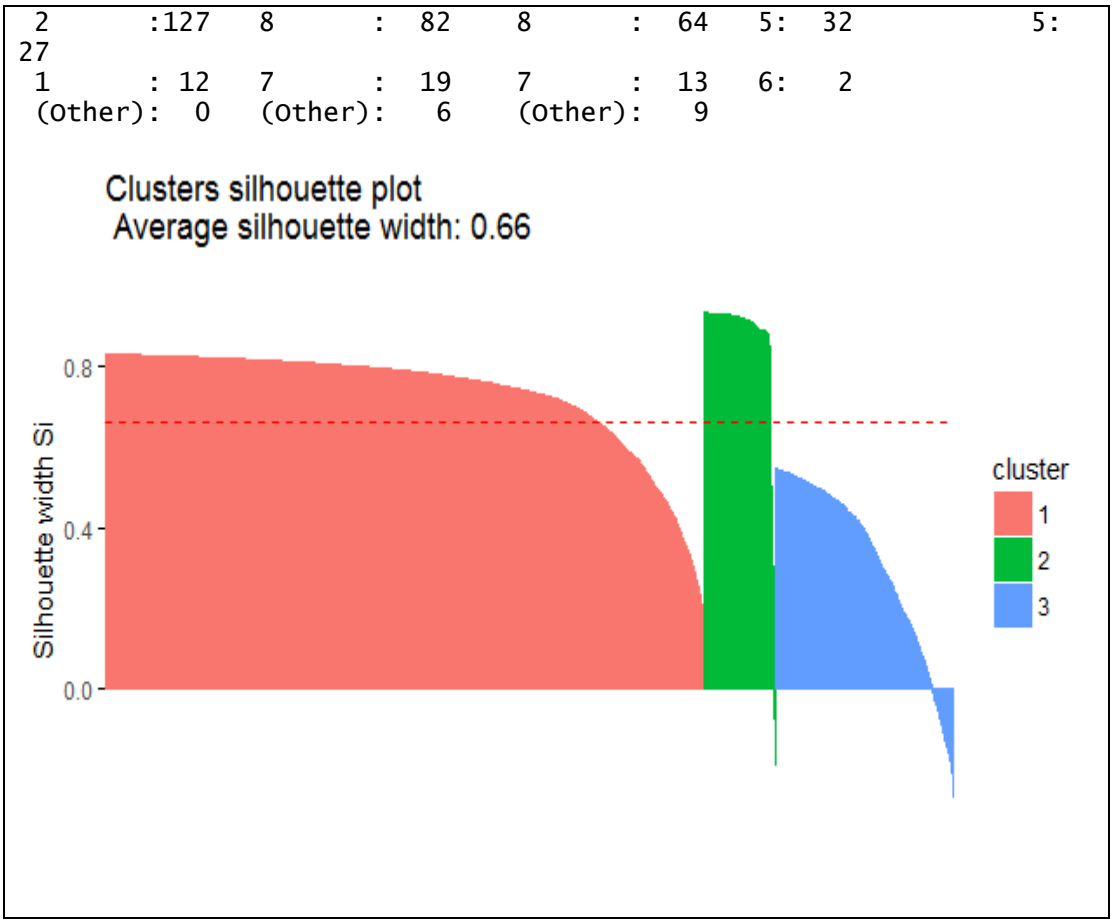
	PENDIDIKAN_AYAH	PENDIDIKAN_IBU	KODE_KEPEMILIKAN	KODE_LISTRIK
1	4	2	1	1
2	8	4	4	1
3	3	3	1	1
4	2	4	4	1
5	2	2	1	1
6	3	2	4	1
7	3	2	1	1
8	2	2	1	1
9	4	4	4	1
10	3	4	1	1
11	2	3	1	1
12	4	4	1	1
13	8	4	1	1
14	8	3	1	1
15	2	2	1	1
16	4	4	1	1
17	3	1	1	1
18	2	2	1	1
19	4	2	1	1
20	4	4	1	1
21	1	1	1	1
22	2	2	1	1
23	1	1	1	1
24	8	4	1	1
25	8	3	1	1
26	3	1	1	1

```
27                    8                8                    1                1
28                    4                4                    4                1
29                    2                1                    1                1
30                    3                2                    1                1
31                    3                4                    1                1
32                    4                2                    1                1
33                    3                3                    1                1
34                    4                8                    1                1
35                    8                2                    1                1
36                    2                2                    1                1
37                    2                2                    1                1
38                    2                2                    1                1
39                    4                3                    1                1
40                    3                2                    1                1
41                    4                2                    1                1
42                    4                4                    3                1
43                    4                4                    1                1
44                    2                2                    1                1
45                    4                3                    1                1
46                    2                2                    1                1
47                    4                2                    1                1
48                    2                2                    1                1
49                    2                2                    1                5
50                    1                2                    1                1
51                    2                2                    1                1
52                    3                3                    1                5
53                    3                2                    1                1
54                    2                2                    1                5
55                    2                2                    1                1
56                    3                3                    1                1
57                    2                3                    1                1
58                    3                2                    1                1
59                    2                2                    1                1
60                    4                3                    1                1
61                    3                2                    1                1
62                    3                2                    1                1
 [ reached getOption("max.print") -- omitted 8520 rows ]
> summary(hslsplit$`1`)
  NILAI_SMT_4    NILAI_SMT_5     JARAK_KOTA     JUMLAH_TANGGUNGAN
LUAS_TANAH
 Min.   : 116   Min.   : 580   Min.   :  0.05   Min.   : 2.000
1:2970
 1st Qu.:1049   1st Qu.:1068   1st Qu.: 10.00   1st Qu.: 2.000
2:2405
 Median :1105   Median :1119   Median : 18.00   Median : 3.000
3:1654
 Mean   :1124   Mean   :1125   Mean   : 21.44   Mean   : 3.077
4: 935
 3rd Qu.:1185   3rd Qu.:1179   3rd Qu.: 30.00   3rd Qu.: 4.000
5: 618
 Max.   :1572   Max.   :1498   Max.   :550.00   Max.   :51.000

 LUAS_BANGUNAN KODE_MCK SUMBER_AIR PENGHASILAN_AYAH PENGHASILAN_IB
U   KERJA_AYAH
 1: 301        1:5980   1:6352     3      :1841     1      :4281
7      :2903
```

```
 2:1178         2:1909    2: 906      4       :1781      2       :1571
5       :2602
 3:3948         3: 693    3:1311      5       :1656      3       :1381
2       :1208
 4:2067                   4:  13      1       : 896      4       : 603
3       : 994
 5:1088                               6       : 579      5       : 391
8       : 572
                                      2       : 569      6       : 116
1       : 187
                                      (Other):1260      (Other): 239
(Other): 116
   KERJA_IBU     PENDIDIKAN_AYAH PENDIDIKAN_IBU KODE_KEPEMILIKAN KO
DE_LISTRIK
 8       :3491   2       :3473    2       :3631   1:6419            1:
8482
 7       :2598   4       :2766    4       :2292   2: 247            2:
5
 5       :1302   3       :1685    3       :2088   3:  48            3:
1
 3       : 766   8       : 305    1       : 258   4:1786            4:
9
 2       : 371   1       : 223    8       : 236   5:  80            5:
85
 1       :  51   7       :  90    7       :  46   6:   2
 (Other):   3   (Other):  40    (Other):  31
> summary(hslsplit$`2`)
  NILAI_SMT_4      NILAI_SMT_5       JARAK_KOTA      JUMLAH_TANGGUN
GAN LUAS_TANAH
 Min.   :   0.0   Min.   :   0.0   Min.   :  0.50   Min.   : 2.000
1:150
 1st Qu.:  79.0   1st Qu.:  80.0   1st Qu.: 12.00   1st Qu.: 2.000
2:281
 Median :  91.0   Median :  92.0   Median : 18.00   Median : 3.000
3:356
 Mean   : 124.4   Mean   : 116.1   Mean   : 22.86   Mean   : 2.989
4:155
 3rd Qu.: 138.0   3rd Qu.: 131.0   3rd Qu.: 30.00   3rd Qu.: 4.000
5: 79
 Max.   :1432.0   Max.   :1192.0   Max.   :200.00   Max.   :16.000

 LUAS_BANGUNAN KODE_MCK SUMBER_AIR PENGHASILAN_AYAH PENGHASILAN_IB
U   KERJA_AYAH
 1: 42         1:715    1:808      5       :232      1       :429
7       :358
 2:107         2:215    2: 93      3       :199      2       :221
5       :287
 3:427         3: 91    3:119      4       :188      3       :186
2       :190
 4:308                  4:  1      1       : 85      4       :104
3       : 79
 5:137                               6       : 80      5       : 47
8       : 54
                                     2       : 74      6       : 11
6       : 34
                                     (Other):163      (Other): 23
(Other): 19
```

```
   KERJA_IBU   PENDIDIKAN_AYAH PENDIDIKAN_IBU KODE_KEPEMILIKAN KOD
E_LISTRIK
 7      :394   2      :380     2      :397    1:835             1:1
014
 8      :358   4      :328     3      :277    2: 31             2:
2
 5      :139   3      :222     4      :270    3: 12             3:
0
 3      : 74   1      : 41     1      : 36    4:131             4:
1
 2      : 46   8      : 41     8      : 32    5: 11             5:
4
 1      :  7   7      :  5     7      :  6    6:  1
 (Other):  3   (Other):  4     (Other):  3
> summary(hslsplit$`3`)
  NILAI_SMT_4    NILAI_SMT_5     JARAK_KOTA     JUMLAH_TANGGUNGAN
LUAS_TANAH
 Min.   :1007   Min.   : 994   Min.   :  0.10   Min.   : 2.000
1:675
 1st Qu.:1426   1st Qu.:1391   1st Qu.:  8.00   1s
t Qu.: 2.000    2:674
 Median :1544   Median :1505   Median : 18.00   Median : 3.000
3:586
 Mean   :1627   Mean   :1570   Mean   : 21.58   Mean   : 3.093
4:392
 3rd Qu.:1755   3rd Qu.:1655   3rd Qu.: 29.00   3rd Qu.: 4.000
5:228
 Max.   :4198   Max.   :3865   Max.   :600.00   Max.   :65.000



 LUAS_BANGUNAN KODE_MCK SUMBER_AIR PENGHASILAN_AYAH PENGHASILAN_IB
U   KERJA_AYAH
 1:  78        1:1718   1:1953     3      :569      1      :1156
7      :878
 2: 311        2: 590   2: 221     4      :504      2      : 575
5      :849
 3:1057        3: 247   3: 376     5      :489      3      : 438
2      :332
 4: 691                 4:   5     2      :278      4      : 195
3      :279
 5: 418                            1      :225      5      : 101
8      :149
                                   6      :191      7      :  26
1      : 32
                                   (Other):299      (Other):  64
(Other): 36
   KERJA_IBU   PENDIDIKAN_AYAH PENDIDIKAN_IBU KODE_KEPEMILIKAN KOD
E_LISTRIK
 8      :900   2      :1113    2      :1180   1:1996            1:2
498
 7      :816   4      : 742    4      : 596   2:  70            2:
26
 5      :464   3      : 498    3      : 595   3:  19            3:
0
 3      :236   1      :  95    1      :  98   4: 436            4:
4
```

```
 2       :127   8       :  82    8       :  64    5:  32              5:
27
 1       : 12   7       :  19    7       :  13    6:   2
 (Other):  0   (Other):   6    (Other):   9
```



Lampiran 5: Perhitungan Indeks Validasi Eksternal Kelompok

```
## Calculate Eksternal Validation Index
## Create Y as a class

str(campurannonsc)
class<-rbind(ori[1:12158,17])
str(class)
setClustsc$clusters
y<- as.numeric(class)
str(y)

## calculate Purity
library(funtimes)
PurityDPC_M <- Purity(y,setClustsc$clusters)
PurityDPC_M
Puritykproto <- Purity(findclustkpro$cluster,y)
Puritykproto
Purity_ensfuzzy<-Purity(y,ensembeldf)
Purity_ensfuzzy

## calculate entropy
library(entropy)
entroDPCM <-entropy(y,setClustsc$clusters)
entroDPCM
entrokproto <-entropy(findclustkpro$cluster, y)
entrokproto
entro_ensfuzzy<-entropy(y,ensembeldf)
entro_ensfuzzy

## Not for Run
## calculate manual Purity
summary(campurannonsc)
library(caret)
expected <- factor(findclustkpro$cluster)
predicted <- factor(y)
results <- confusionMatrix(data=predicted, reference=expected)
print(results)

Output
> ## Calculate Eksternal Validation Index
> ## Create Y as a class
> set.seed(1234)
> str(campurannonsc)
'data.frame':  12158 obs. of  16 variables:
$ NILAI_SMT_4      : num  1161 1214 1256 1031 1335 ...
$ NILAI_SMT_5      : num  1145 1179 1255 1037 1333 ...
```

```
$ JARAK_KOTA       : num  31 24 31 24 24 31 24 27 24 24 ...
$ JUMLAH_TANGGUNGAN: num  2 3 6 2 6 7 6 2 6 2 ...
$ LUAS_TANAH       : Factor w/ 5 levels "1","2","3","4",..: 2 1 2 2
2 4 3 2 1 1 ...
$ LUAS_BANGUNAN    : Factor w/ 5 levels "1","2","3","4",..: 3 3 2 2
3 4 3 2 3 3 ...
$ KODE_MCK         : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1
1 1 2 ...
$ SUMBER_AIR       : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1
1 1 1 1 1 ...
$ PENGHASILAN_AYAH : Factor w/ 25 levels "1","2","3","4",..: 3 1 6
3 3 3 3 5 3 1 ...
$ PENGHASILAN_IBU  : Factor w/ 21 levels "1","2","3","4",..: 1 3 1
1 1 3 3 1 2 3 ...
$ KERJA_AYAH       : Factor w/ 8 levels "1","2","3","4",..: 5 8 2 7
7 7 7 7 3 8 ...
$ KERJA_IBU        : Factor w/ 8 levels "1","2","3","4",..: 5 7 8 8
8 7 7 8 7 3 ...
$ PENDIDIKAN_AYAH  : Factor w/ 10 levels "1","2","3","4",..: 4 8 3
2 2 3 3 2 4 3 ...
$ PENDIDIKAN_IBU   : Factor w/ 10 levels "1","2","3","4",..: 2 4 3
4 2 2 2 2 4 4 ...
$ KODE_KEPEMILIKAN : Factor w/ 6 levels "1","2","3","4",..: 1 4 1 4
1 4 1 1 4 1 ...
$ KODE_LISTRIK     : Factor w/ 5 levels "1","2","3","4",..: 1 1 1 1
1 1 1 1 1 1 ...
> class<-rbind(ori[1:12158,17])
> str(class)
int [1, 1:12158] 3 1 1 1 3 1 3 1 1 3 ...
> setClustsc$clusters
[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1
[40] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1
3 3 3 3 3 1 1 1
[79] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3
1 1 3 1 3 3 1 1
[118] 3 3 3 3 3 3 3 1 3 3 1 3 3 1 1 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3
3 1 3 1 1 3 1 1
[157] 1 1 3 1 1 1 1 1 1 1 3 1 1 1 1 1 1 3 1 1 1 3 1 1 1 1 1 1 1 3
3 3 3 3 3 3 3 3
[196] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3
[235] 3 3 3 3 3 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3
3 3 3 3 3 3 3 3
[274] 3 3 3 3 3 2 2 2 2 3 3 3 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1
[313] 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3
[352] 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1
3 3 3 3 1 3 1 1
[391] 1 1 1 1 3 1 1 1 1 1 3 1 1 2 1 1 1 1 1 3 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1
[430] 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1
[469] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
2 2 2 2 2 1 1 1
```

```
[508] 2 1 1 1 1 1 1 3 1 1 1 1 1 1 1 1 1 1 3 1 1 3 3 1 1 1 3 1 1 1 1
1 1 3 3 3 3 3 3
[547] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2
[586] 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 3 1 1 1 3 1 3 1 2 1 1 1 1 3 1 3
3 1 1 1 3 1 3 3
[625] 2 3 3 3 3 3 3 1 2 3 1 1 3 3 3 3 3 3 1 1 3 1 1 1 1 1 1 3 2 1 1
1 3 3 1 3 1 1 3
[664] 1 3 3 1 2 3 1 3 1 3 2 3 1 2 2 2 2 3 3 3 3 1 1 1 1 1 2 1 1 1 1
1 1 1 1 1 1 1 1
[703] 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1 1 3 3 3 3
1 3 3 3 3 1 1 3
[742] 1 3 1 3 3 3 2 2 2 2 2 2 1 1 1 1 1 1 2 2 2 2 2 2 2 2 3 1 3 3 2
1 1 1 1 1 1 1 1
[781] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 2 1 1
1 1 1 1 1 1 1 1
[820] 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 2 1
1 1 1 1 1 1 1 2
[859] 2 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 2 1 2 1 2 1 1 1 2 1
1 1 1 1 1 1 1 1
[898] 1 1 1 2 1 1 2 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 2 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1
[937] 1 1 1 1 2 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 2 1 1 2 1 1 1
1 1 1 1 1 1 1 1
[976] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[ reached getOption("max.print") -- omitted 11158 entries ]
> y<- as.numeric(class)
> str(y)
num [1:12158] 3 1 1 1 3 1 3 1 1 3 ...
> ## calculate Purity
> library(funtimes)
> PurityDPC_M <- Purity(y,setClustsc$clusters)
> PurityDPC_M
$`pur`
[1] 0.5577398

$out
ClassLabels ClusterLabels ClusterSize
1           1             1        6321
3           2             3           6
2           3             2         454

> Puritykproto <- Purity(findclustkpro$cluster,y)
> Puritykproto
$`pur`
[1] 0.5090475

$out
ClassLabels ClusterLabels ClusterSize
1           1             1        4878
2           2             3        1308
3           3             2           3

> Purity_ensfuzzy<-Purity(y,ensembeldf)
> Purity_ensfuzzy
$`pur`
[1] 0.5718046
```

```
$out
ClassLabels ClusterLabels ClusterSize
1           1             2        6493
3           2             1           6
2           3             3         453

> ## calculate entropy
> library(entropy)
> entroDPCM <-entropy(y,setClustsc$clusters)
> entroDPCM
[1] 9.261632
> entrokproto <-entropy(findclustkpro$cluster, y)
> entrokproto
[1] 9.3258
> entro_ensfuzzy<-entropy(y,ensembeldf)
> entro_ensfuzzy
[1] 9.261632
> ## Not for Run
```

BIODATA PENULIS

Laila Qadrini atau biasa dipanggil Ila, lahir di Ujung Pandang pada Tanggal 29 April 1987. Penulis merupakan anak keempat dari tujuh bersaudara. Penulis menyelesaikan Sekolah Dasar di SDN Kelapa Tiga Tahun 1998, SMPN 8 Makassar Tahun 2002, SMAN 1 Tinambung Sulawesi Barat Tahun 2004, masuk kuliah pada jenjang S1 di Jurusan Matematika Universitas Negeri Makassar (UNM) Tahun 2005.Setelah menempuh pendidikan Sarjana selama 4 tahun, penulis melanjutkan ke jenjang Magister Statistika ITS pada tahun 2016, penulis sebelumnya pernah bekerja di PT. Pegadaian Persero. Karya yang telah dibuat oleh penulis ketika S1 dulu adalah Analisis *multivariate* data dengan *Structural Equation Modelling* (SEM), Penulis mempunyai prinsip dalam hidup, yaitu "*Dimana ada kemauan disitu ada jalan"*. Komunikasi lebih lanjut dengan penulis dapat melalui email qadrini.laila@yahoo.com.