



Identifikasi *Cluster* Penduduk Usia Kerja Pada Provinsi Sumatera Selatan Menggunakan K-Modes

Fithri Selva Jumeilah¹ dan Dicky Pratama²

¹STMIK Global Informatika MDP, email: fithri.selva@mdp.ac.id

²STMIK Global Informatika MDP, email: dqpratama@mdp.ac.id

Abstrak

Semakin tinggi jumlah penduduk yang tidak produktif maka semakin tinggi rasio ketergantungan pada suatu negara. Produktifitas penduduk usia kerja akan mendorong ekonomi negara tetapi sebaliknya akan menciptakan ketidakstabilan sosial maupun politik. Indonesia diperkirakan akan menghadapi bonus demografi puncaknya pada tahun 2030. Untuk mempermudah penduduk usia kerja mendapatkan pekerjaan dibutuhkan pembangunan dan kebijakan pemerintah salah satunya provinsi Sumatera Selatan. Membuat kebijakan membutuhkan gambaran keadaan penduduk usia kerja yang diperoleh dengan mengelompokkan data sensus Provinsi Sumatera selatan tahun 2010 yang diperoleh dari BPS dan melihat pola dari setiap kelompoknya. Data dikelompokkan dengan menggunakan R-studio dengan algoritma K-modes. Dari hasil penelitian diperoleh penduduk usia kerja dikelompokkan menjadi 8 cluster, yaitu: C1 yang merupakan kumpulan laki-laki tinggal di desa dan masih bekerja, C2 sebagian besar penduduk yang berusia 25 tahun dan merupakan pelajar, C3 perempuan tinggal di desa dan tidak bekerja, C4 sebagian besar perempuan yang tinggal di desa, C5 rata-rata laki-laki tinggal di kota yang bekerja di bidang jasa kesehatan, perdagangan dan konstruksi, C6 merupakan kelompok perempuan yang tinggal di desa bekerja yang memiliki 1 anak, C7 laki-laki belum menikah <30 tahun yang bekerja di perkebunan, dan terakhir C8 kumpulan ibu rumah tangga tinggal di kota.

Kata kunci: Data Mining, Clustering, Usia Kerja, Sumatera Selatan, Bonus Demografi

Abstract

The higher the number of unproductive populations the higher the dependency ratio in a country. The productivity of the adult population will stimulate the state economy but will create social and political instability. Indonesia is predicted to experience demographic bonus in 2030. To facilitate the adult population to be easy to get job required development and policy of South Sumatera government. Making the policy requires an overview of the adult population obtained by clustering the South Sumatra census data of 2010 from BPS. The data is clustered using K-mode algorithm. From the results of study, the adult population is clustering into 8 groups: C1 is a group of men living in the village and still working, C2 is a group of populations 25 years old and students, C3 is women's group who live in the village and not work, C4 is predominantly female living in the village, C5 is the average of male living in the city, working in health services, trade and construction, C6 is women's group living and working in village with 1 child, C7 is a group of unmarried men under 30 years, working on plantations, and C8 is a group of housewives living in the city.

Keywords: Data Mining, Clustering, Unproductive Age, South Sumatera, Demographic Bonus

1. Pendahuluan

Menurut Undang-undang ketenagakerjaan No.13 Tahun 2003 [1], penduduk usia kerja adalah penduduk yang berusia 15 tahun sampai 64 tahun. Berdasarkan hasil sensus penduduk tahun 2010 jumlah penduduk Indonesia 237.641.326 jiwa dan 169 juta jiwa adalah penduduk usia kerja [2]. Di provinsi Sumatera Selatan penduduk usia kerja terdapat sebanyak 3.417.875 jiwa. Penduduk usia kerja adalah kategori penduduk yang sangat mempengaruhi keadaan negara karena penduduk usia kerja yang menanggung penduduk yang belum produktif (<15 tahun) dan penduduk tidak lagi produktif (>64 tahun) atau sering disebut dengan rasio ketergantungan. Tingginya tingkat ketergantungan mencerminkan beban tanggungan penduduk usia kerja untuk membiayai kehidupan semua anggota keluarga akan semakin sulit.

Penduduk usia kerja sangat memengaruhi negara. Jika penduduk usia kerja sebagian besar bekerja maka akan mendorong ekonomi negara semakin membaik tetapi sebaliknya jika mereka tidak bekerja maka akan menyebabkan ketidakstabilan sosial dan politik. Diperkirakan pada tahun 2030, penduduk usia kerja akan sangat meningkat (bonus demografi) dan akan mencapai puncak [3]. Bonus demografi terjadi karena menurunnya angka kematian bayi dan berhasilnya program Keluarga Berencana (KB) [4]. Dengan adanya bonus demografi, pemerintah harus memanfaatkan kesempatan tersebut untuk meningkatkan kesejahteraan rakyat dengan pembangunan dan membuat kebijakan, seperti: pemberian pelatihan kerja, sertifikasi profesi, penyediaan lapangan kerja dan masih banyak kebijakan lainnya. Untuk mempermudah pemerintah terutama pemerintah Sumatera Selatan membuat kebijakan mengenai penduduk usia kerja maka dibutuhkan klasterisasi data penduduk usia kerja untuk melihat gambaran karakteristik setiap *cluster* dari penduduk usia kerja.

Penelitian mengenai klasterisasi telah banyak dilakukan diantaranya: untuk pengelompokan kemiskinan di Jawa Timur dengan menggunakan klasterisasi hierarki [5], *clustering* koleksi perpustakaan dengan menggunakan metode k-means untuk melihat pola minat pengunjung untuk membantu penentuan kebijakan pengadaan koleksi [6], klasterisasi penduduk lanjut usia Provinsi Sumatera Selatan yang dapat digunakan sebagai landasan pengambilan kebijakan [7], *clustering* data rasio keuntungan perusahaan bidang pertambangan dan sektor industri dengan metode k-means [8]. Selain itu juga terdapat penelitian yang dilakukan untuk klasterisasi data kecelakaan lalu lintas dengan menggunakan metode k-means [9]. Dimana data yang digunakan diperoleh dari Dislantas Polres Kabupaten Sleman Daerah Istimewa Yogyakarta yang mencatat semua kejadian kecelakaan setiap kejadian. Dari hasil penelitian tersebut, diperoleh sebuah sistem yang mampu mengelolah data kecelakaan setiap bulan dengan metode k-means. Setelah diperoleh pola tersebut dapat dilakukan analisis oleh pakar untuk menentukan kategori *cluster*. Setelah data di kategorikan maka akan divisualisasikan menjadi sebuah peta yang sudah terhubung dengan data spasial.

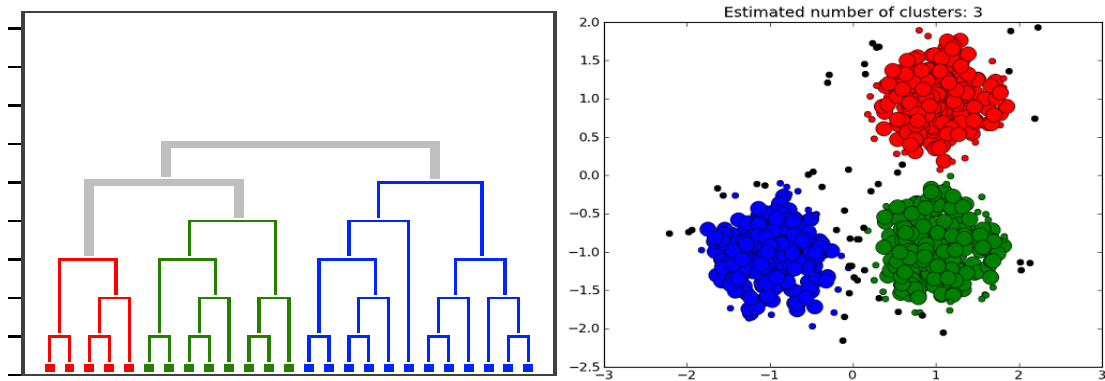
Data yang akan digunakan dalam penelitian ini adalah data sensus penduduk Sumatera selatan tahun 2010 yang diperoleh dari Badan Pusat Statistik (BPS). Data yang diperoleh hampir semuanya bersifat kategoris atau nominal seperti jenis kelamin, pendidikan terakhir, pekerjaan dan lainnya. Setiap atribut kategoris diwakili seperangkat nilai kategoris unik. Tidak seperti data numerik, nilai kategoris bersifat diskrit dan tidak teratur. Oleh sebab itu, algoritma pengelompokan data numerik tidak bisa digunakan untuk mengelompokan data kategoris [10]. Pada penelitian ini metode yang akan digunakan untuk *clustering* adalah k-modes. K-modes merupakan perbaikan dari metode k-means yang mengalami perubahan cara penentuan *cluster center* [10].

2. Tinjauan Pustaka

2.1 Clustering

Clustering adalah menganalisis sekumpulan objek dan mengelompokkan objek-objek tersebut kedalam *cluster* berdasarkan nilai kemiripan (similaritas) [10]. *Clustering* adalah salah satu metode *data mining* dengan pendekatan *unsupervised* (pembelajaran tidak terawasi) yang tidak membutuhkan target *ouput*. *Data mining* adalah proses pengekstrakan informasi yang digunakan untuk mencari pola yang belum diketahui dari sekumpulan data yang berjumlah besar [11]. Objek dalam sebuah *cluster* memiliki data yang mirip dan akan sangat berbeda dengan data pada objek *cluster* yang lain [12]. Dengan demikian, terdapat nilai kesamaan yang tinggi untuk setiap objek dalam satu *cluster* atau kelompok.

Clustering memiliki dua metode yaitu hierarki dan non-hierarki. *Cluster* dari metode hierarki yang dihasilkan akan membentuk hierarki mulai dari data yang paling mirip sampai ke data yang tidak mirip. Metode non-hierarki jumlah kelompok ditentukan di awal sedangkan metode hierarki di tentukan pada tahap akhir analisis [5]. Gambaran hasil *clustering* menggunakan metode hirarki atau non hierarki dapat dilihat pada Gambar 3.



Gambar 3. Contoh Hasil *Clustering* Hierarki dan Non-hierarki

Pada metode non-hierarki, pengelompokan objek berdasarkan nilai kemiripan. Untuk data yang berjenis numerik dapat dilakukan perhitungan jarak geometris seperti *Euclidean distance* atau *Manhattan distance* seperti yang metode k-means. Sedangkan untuk data nominal atau kategorikal dapat menggunakan *dissimilarity* seperti metode k-modes [10].

2.2 K-Modes

K-modes pertamakali dipublikasikan pada tahun 1997 yang merupakan metode pertama yang dapat mengelompokkan data kategoris [12]. Pendekatan k-modes merupakan hasil modifikasi dari k-means standar untuk mengklasterisasi data kategoris dengan menggantikan fungsi jarak Euclidean dengan jarak *mismatching* [13,14]. Pada metode k-modes, *cluster center* diwakili nilai yang paling sering muncul atau sering sebut dengan *mode*. Misalkan, *mode* himpunan $\{n, j, k, n, n, m\}$ adalah n nilai yang paling sering muncul yang akan digunakan sebagai *cluster center* [7,12]. *Cluster center* atau *mode* akan terus diperbaharui dengan nilai kategoris yang paling sering pada setiap iterasinya. Cara kerja dari metode k-modes adalah meminimalkan jarak masing-masing objek dalam sebuah *cluster* [13]. Untuk menghitung nilai fungsi objektif pada k-modes dapat menggunakan Persamaan 1.

$$F(W, Z) = \sum_{l=1}^k \sum_{i=1}^n y_{i,l} d(X_i, Q_k) \quad (1)$$

Dimana

$$y_{i,l} \in \{0,1\}, 1 \leq l \leq k, 1 \leq i \leq n,$$

$$\sum_{l=1}^k y_{li} = 1, 1 \leq i \leq n,$$

dan

$$0 < \sum_{l=1}^n y_{li} = 1, 1 \leq l \leq k$$

Dimana $y_{i,l}$ adalah elemen matrik partisi $Y_{n \times k}$. Nilai Q_k adalah nilai *mode* dari *cluster*. Nilai *mode* diambil dari nilai atribut yang paling sering muncul. Setiap objek akan dilakukan perhitungan jaraknya dengan *mode* dimana jarak tersebut sering disebut dengan nilai *dissimilarity*. Dari hasil dari Persamaan 2 setiap objek akan dikelompokkan pada *cluster* terdekat.

$$d(X, Y) = \sum_{j=1}^k \delta(x_j, y_j), \quad (2)$$

dimana:

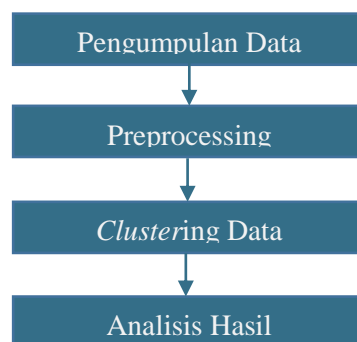
$$\delta(x_j, y_j) = \begin{cases} 0; & \text{if } x_j = y_j \\ 1; & \text{if } x_j \neq y_j \end{cases}$$

Ketidaksamaan antara objek X dan Y dapat didefinisikan dengan total *mismatching* [12]. Semakin kecil nilai *mismatching*, semakin mirip dua objek tersebut. Tahapan *clustering* dengan menggunakan metode k-modes adalah sebagai berikut [12]:

1. Pilihlah objek yang paling sering muncul sebanyak k yang akan digunakan sebagai *cluster center* atau sering disebut dengan *mode*.
2. Hitung nilai *dissimilarity* yang merupakan jarak dari objek ke setiap *mode cluster*. Kelompokkan objek berdasarkan jarak dengan pusat *cluster*.
3. Ulangi tahapan 2 sampai semua objek sudah dikelompokkan kedalam *cluster*.
4. Lakukan pengujian ulang semua objek dengan menentukan *mode* baru dan bandingkan dengan hasil *cluster mode* yang lama, jika ada perubahan ulangi tahapan 2 ; jika tidak, perulangan berhenti.

3. Metodologi

Adapun langkah-langkah penelitian ini dapat dilihat pada Gambar 4.



Gambar 4. Metodologi Penelitian

3.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah data sensus penduduk Provinsi Sumatera Selatan yang paling terbaru yaitu data tahun 2010 yang di peroleh dari Badan Pusat Statistik (BPS). Data yang digunakan hanya 10% karena sesuai dengan aturan pemerintah mengenai data mentah data sensus hanya boleh maksimal 10%. Pengumpulan data dilakukan melalui wawancara tatap muka langsung kepada seluruh penduduk Indonesia. Dari 10% data

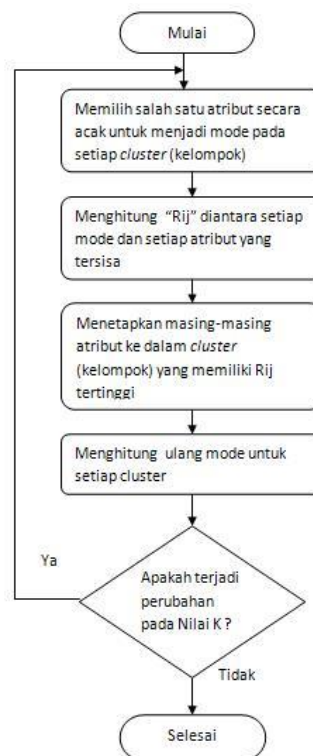
sensus penduduk hanya data penduduk usia kerja yang digunakan, yaitu penduduk dengan usia 15 tahun sampai dengan 64 tahun yaitu sebanyak 472.918 data.

3.2 Preprocessing

Data yang diperoleh dari BPS terdapat data 711.823 jiwa. Jumlah data penduduk yang usia kerja adalah 472.918 jiwa. Dari data sensus terdapat beberapa data yang kosong. Juga terjadi penghapusan beberapa atribut yang tidak dibutuhkan dalam penelitian ini seperti tanggal dan bulan lahir. Ada beberapa atribut yang ditransformasikan seperti atribut tahun lahir menjadi umur. Selain pengurangan atribut, juga terjadi pengurangan baris sebanyak 350 baris karena terlalu banyak data yang kosong.

3.3 Clustering Data

Setelah data diolah data siap digunakan untuk *clustering*. Pada penelitian ini metode *clustering* yang akan digunakan adalah k-modes dari R-studio. Diagram alir dari metode k-modes dapat dilihat pada Gambar 5. Jumlah *cluster* akan didapatkan dengan menentukan nilai *silhouette*. Jumlah *cluster* yang akan diambil jika nilai *silhouette* yang paling mendekati 1.



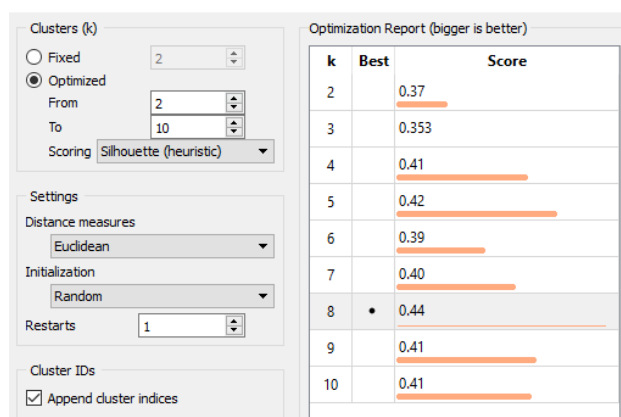
Gambar 5. Flowchart K-Modes

3.4 Analisis Hasil

Dari hasil *clustering* akan diperoleh beberapa *cluster* dan dapat diketahui atribut mana saja yang paling mempengaruhi setiap *cluster*. Setelah diketahui atribut yang paling mempengaruhi maka hasil dari *clustering* dapat divisualisasikan secara 3 dimensi. Untuk memvisualisasikan hasil *clustering* dapat menggunakan Scatter3d yang ada di R-studio. Visualisasi tersebut terbentuk berdasarkan 3 atribut yang paling mempengaruhi. Hasil analisis tersebut juga akan digunakan untuk menarik kesimpulan akhir penelitian.

4. Hasil dan Pembahasan

Data yang diperoleh dari BPS akan dilakukan proses *cleaning* terlebih dahulu baru akan digunakan pada tahap *clustering*. *Clustering* akan mengelompokkan data menjadi *k cluster*. Dimana *k* diperoleh dari nilai *silhouette* yang paling besar atau yang paling mendekati nilai 1. Plot *silhouette* menampilkan ukuran seberapa dekat setiap titik dalam satu *cluster* dari titik *cluster* yang lain. Dari hasil analisis pada Gambar 6 diperoleh nilai *k* yang paling baik yaitu 8 dimana *silhouette score*-nya adalah 0.44.

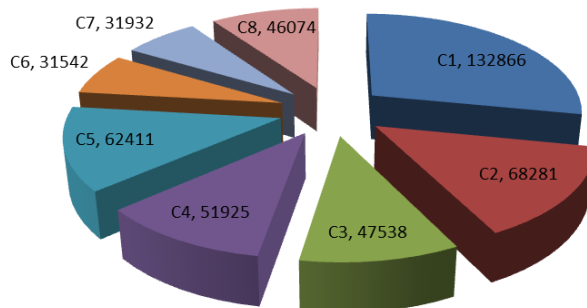


Gambar 6. Silhouette Score

Setelah diketahui jumlah *cluster*, data sudah bisa di *clustering* dengan menggunakan *tools* R-studio dan metode yang digunakan k-modes dengan perintah di bawah ini.

```
cluster.results <- kmodes(data.to.cluster[,1:20], 8, iter.max = 10, weighted = FALSE)
```

Angka 8 adalah jumlah *cluster* dan [1:20] merupakan index dari atribut yang akan digunakan pada proses *clustering*. Hasil *clustering* dapat dilihat pada Gambar 7 yang menyatakan jumlah objek dari setiap *cluster*.



Gambar 7. Pie Chart Jumlah Objek per cluster

Dari hasil analisis diperoleh pola similariti setiap *cluster*-nya. Pada C1 sebanyak 83% tinggal di desa, 91% berjenis kelamin laki-laki yang sudah menikah, 96% bekerja dimana sebagian bekerja dibidang perkebunan dan hanya 0,6% yang tidak ingin bekerja. Rata-rata objek pada *cluster* ini memiliki umur antara 25 tahun dan 55 tahun yang berperan sebagai kepala keluarga. Sebagian besar merupakan lulusan SD dan SMP dan 3,5% tidak pernah sekolah.

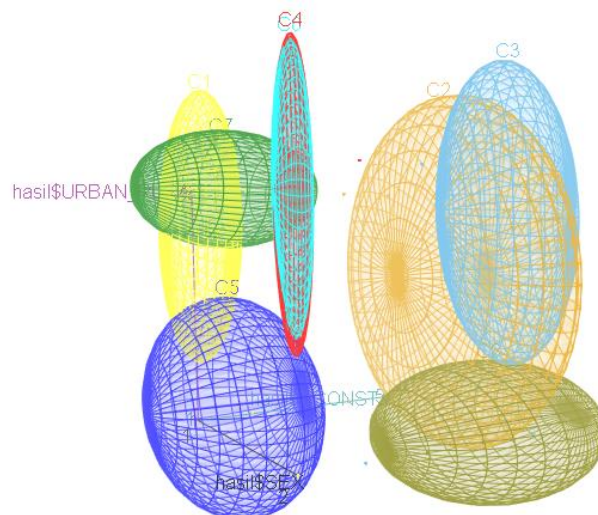
Untuk C2 sebagian besar penduduk yang tinggal di desa baik itu perempuan atau laki-laki yang merupakan anak dari kepala keluarga berumur <25 tahun yang tidak ingin bekerja karena masih mengenyam pendidikan yang sebagian besar sekolah tingkat SMA dan status belum kawin. Sedangkan untuk C3, kelompok perempuan yang merupakan istri dari kepala keluarga. Sebanyak 80% penduduk *cluster* C3 tinggal di desa dan berusia antara 20 tahun dan 37 tahun.

Pada *cluster* ini sebagian besar hanya tamatan Sekolah Dasar (SD) dan terdapat 2.049 orang yang tidak bisa membaca dan menulis, 2.711 orang yang tidak bisa berbahasa Indonesia. Sebanyak 75% dari *cluster* C3 tidak ingin bekerja.

Cluster C4 hampir semuanya adalah perempuan yang berstatus menikah dengan kisaran umur 30 sampai 40 tahun dan terdapat 87% objeknya tinggal di desa. Sebagian besar objek pada *cluster* ini tamatan SD tetapi terdapat 1136 orang yang tamatan S1 dan terdapat 9 orang yang tamatan S2 atau S3. Terdapat 3.352 orang yang tidak bisa bahasa Indonesia. *Cluster* C4 dan C3 didominasi oleh perempuan tetapi bedanya objek C4 96% bekerja di bidang perkebunan milik keluarga sendiri. *Cluster* kelima C5 yang merupakan kumpulan penduduk laki-laki yang 90% tinggal di kota. Sebagian merupakan kepala keluarga dan sebagiannya lagi merupakan anak kandung dari kepala keluarga. Sebanyak 91% penduduk *cluster* ini bekerja sebagai karyawan di berbagai bidang diantaranya: di jasa kesehatan (26%), di bekerja di bidang perdagangan (16%) dan bekerja konstruksi atau bangunan (11%). Rata-rata penduduk pada *cluster* C5 merupakan tamatan SMA dan terdapat 11% yang merupakan tamatan S1. Hanya 49% penduduk *cluster* ini yang berstatus belum menikah.

Berikutnya *cluster* C6 yang merupakan kumpulan perempuan yang tinggal di desa yang sebagian besar tamatan SD dan terdapat 89% sudah berstatus menikah. Sebanyak 97% penduduk pada *cluster* ini bekerja di perkebunan milik keluarganya sendiri dan sebanyak 72% baru memiliki 1 anak. *Cluster* C7 adalah kelompok penduduk laki-laki (85%) yang tinggal di desa (97%) dimana rata-rata umur < 30 tahun. Sebagian besar belum menikah (91%) tamatan SD tetapi sudah bekerja (86%) dibidang perkebunan milik keluarganya. *Cluster* terakhir yaitu C8 yang merupakan kumpulan dari perempuan yang sudah menikah tinggal di kota. Pendidikan terakhir rata-rata tamatan SMA dan tidak ingin bekerja.

Dari hasil pola similariti diketahui bahwa atribut yang paling mempengaruhi dalam setiap *cluster*. Visualisasi dari hasil *clustering* berdasarkan 3 atribut yang paling mempengaruhi *cluster* dapat dilihat dengan menggunakan *scatter3d*.



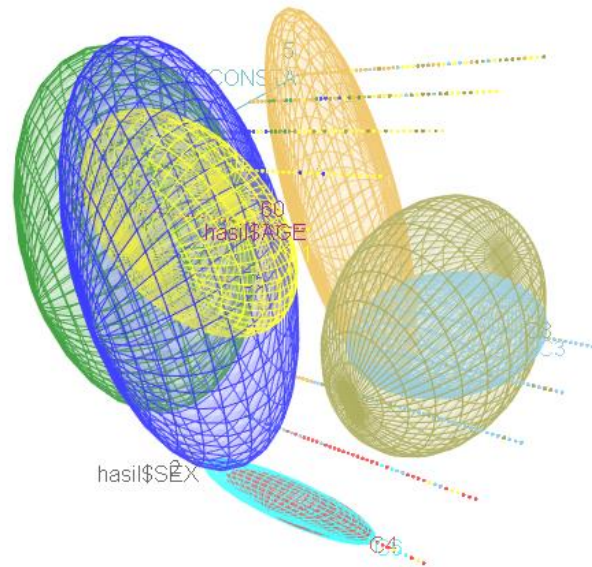
Gambar 8. Visualisasi 3D Hasil *Clustering* 1

Gambar 8 diperoleh dari perintah di bawah ini, dimana sumbu x = Urban_Rura, sumbu y = Sex (jenis kelamin) dan sumbu z = Econsta (Status bekerja).

```
scatter3d(x = hasil$URBAN_RURA, y = hasil$SEX, z = hasil$ECONSTA, groups = hasil$CLUSTER, surface=FALSE, grid = TRUE, ellipsoid= TRUE, surface.col = c("999999", "#E69F00", "#56B4E9", "#FF0000", "#0000FF", "#00FFFF", "#008000", "#808000"))
```


Sedangkan untuk visualisasi *scatter* dengan sumbu x adalah umur, sumbu y atribut jenis kelamin, dan sumbu z pekerjaan. Dengan perintah di bawah ini akan menghasilkan keluaran seperti Gambar 9.

```
scatter3d(x = hasil$AGE, y = hasil$SEX, z = hasil$ECONSTA, groups =
hasil$CLUSTER, surface=FALSE, grid = TRUE, ellipsoid= TRUE, surface.col =
c("999999", "#E69F00", "#56B4E9", "#FF0000", "#0000FF", "#00FFFF", "#008000",
"#808000"))
```



Gambar 9. Visualisasi 3D Hasil *Clustering*2

5. Kesimpulan

Setelah dilakukan analisis dan diketahui pola similaritinya maka dapat disimpulkan bahwa pola menarik dari setiap *cluster* adalah yaitu: C1 merupakan kumpulan laki-laki yang tinggal di desa dan masih bekerja, C2 *cluster* penduduk yang berumur 25 tahun yang masih sekolah, C3 *cluster* perempuan tinggal di desa dan tidak bekerja, C4 kelompok perempuan yang tinggal di desa tetapi bekerja memiliki 2 anak, C5 *cluster* laki-laki tinggal di kota yang bekerja di bidang jasa kesehatan, perdagangan dan bangunan, C6 kelompok perempuan yang tinggal di desa bekerja yang rata-rata memiliki 1 anak, C7 kumpulan laki-laki belum menikah < 30 tahun yang bekerja di perkebunan, dan terakhir C8 kumpulan ibu rumah tangga yang tinggal di kota.

Dari hasil analisis pola *cluster* penduduk usia kerja di Sumatera Selatan diharapkan dapat membantu pemerintah dalam mengambil kebijakan dan rencana pembangunan guna menghadapi ledakan penduduk usia kerja di tahun 2030 mendatang. Untuk peneliti selanjutnya diharapkan mampu analisis yang lebih dalam dan menggunakan metode lain yang lebih baik.

Daftar Pustaka

- [1] Republik Indonesia Undang-undang Nomor 13 Tahun 2003 tentang ketenagakerjaan.
- [2] Badan Pusat Statistik. "Jumlah dan Distribusi Penduduk." Internet: <https://sp2010.bps.go.id/index.php/site/index>, 2010, [Sep, 27, 2017].
- [3] Ngadi, et. al. "Kebijakan Ketenaga kerjaan menghadapi Ledakan Penduduk Usia Kerja di Indonesia". Jakarta: LIPI, 2010.

- [4] Djoko M. A.S., et. al. *Mobilitas Penduduk dan Bonus Demografi*. Bandung: Unpad Press, 2015, pp. 84-85.
- [5] Anuraga Gangga. "Hierarchical *Clustering* Multiscale Bootstrap untuk Pengelompokan Kemiskinan di Jawa Timur." *Jurnal Statistika*, Vol. 1, No. 3, pp. 28-33. 2015.
- [6] Nengsih Warnia. "*Clustering* K-Means Analisis (Studi Kasus: Koleksi Perpustakaan)." In *Seminar Nasional Teknologi Informasi Komunikasi dan Industri*, 2014.
- [7] Jumeilah Fithri Selva dan Pratama Dicky. *Klasifikasi Penduduk Lanjut Usia Sumatera Selatan Menggunakan Algoritma K-Modes*. *Jurnal Technology Acceptance Model*, Vol. 8, No. 2, pp. 85-89, 2017.
- [8] Kristanto H. N., A. L. Christopher Andreas, dan S. Budi Halim. "Implementasi K-means *Clustering* untuk Pengelompokan Analisis Rasio Profitabilitas dalam Working Capital." *Jurnal Informatika dan Sistem Informasi*, Vol. 02, No. 01, pp. 9-14, 2016.
- [9] Iswari Lizda dan Ayu Gita Ervina. "Pemanfaatan Algoritma K-means untuk Pemetaan Hasil Klasterisasi Data Kecelakaan Lalu Lintas." *Jurnal Teknologi dan Industri*, Vol. 21, No. 1, pp. 01-11, 2015.
- [10] Huda Misbachul, Hayun Rahma Dian, Sri Annisa. "Diversity-based Attribute Weighting for K-Modes *Clustering*." *Jurnal Ilmu Komputer dan Informasi*, Vol. 7, No. 2, pp. 61-66, 2014.
- [11] S.Neelamegam, Dr.E.Ramaraj. "Classification algorithm in Data mining: An Overview ". *International Journal of P2P Network Trends and Technology (IJPTT)*, Vol.3, No. 8. pp.369 – 374, 2013
- [12] Aranganayagi, S., and Thangavel, K., "Improved K-Modes for Categorical *Clustering* Using Weighted Dissimilarity Measure." *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, Vol.3, No.3, pp. 729-735, 2009.
- [13] Huang, Z., "A Fast *Clustering* Algorithm to *cluster* Very Large Categorical Datasets in Data Mining", In *Proc. SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*. 1997.
- [14] Xiang, Z., and Zahidul, M.. *Hartigan's Method for K-modes Clustering and Its Advantages*. *Proceedings of the Australasian Data Mining Conference*, Vol.158, 2014.