ANALISIS *DATA MINING* PENGELOMPOKAN DATA WARGA MENGGUNAKAN METODE *STATISTIK K-MEANS* UNTUK MEREKOMENDASIKAN PEKERJAAN SAMPINGAN

Lisna Zahrotun [1], Utaminingsih Linarti [2]
[1,2] Program Studi Teknik Informatika,
Fakultas Teknologi Industri,
Universitas Ahmad Dahlan
[1]Lisna.zahrotun@tif.uad.ac.id, [2]utaminingsih.linarti@ie.uad.ac.id

ABSTRACT

The Indonesian economy is growing such that the government needs to pay attention to the level of social walfare. Improvement of social welfare is prioritized for groups of people with the lowest levels of the economy. Grouping level of the existing economic community has not well arranged therefore sometimes distribution of government social aid is wrong targeted, it does not match the level of the economy.In this research the method used is k-Means clustering where k-means clustering is a method of grouping a convenient, efficient and effective. In addition to k-means clustering method, a statistical methods in determining the standard deviation is also performed. The case study is conducted in Babadan village, Bantul, Indonesia.Results from this study is the formation of three clusters based on attributes of age, education, and occupation. The first generated cluster is with cluster center of 43.16 years of age, last education of high school, and profession as an entrepreneur. The second generated cluster is with cluster center of 44.83 years of age, last education of high school, and profession as a labor. The third generated cluster is with cluster center of 43.16 years of age, last education of high school, and profession as a civil servant. Hence the second cluster is the priority in improving the welfare by providing additional occupation.
Keywords: K-Means Clustering, Economic Community, Welfare, Additional Occupation

INTISARI

Perekonomian Indonesia yang semakin berkembang menjadikan pemerintah perlu memperhatikan tingkat kesejahteraan masyarakat. Peningkatan kesejahteraan masyarakat diprioritaskan untuk kelompok masyarakat dengan tingkat perekonomian paling rendah. Pengelompokan tingkat perekonomian masyarakat yang sudah ada belum tertata sehingga terkadang terjadi kesalahan sasaran, tidak sesuai dengan tingkat perekonomian.Dalam penelitian ini metode yang digunakan adalah *k-Mean clustering*. Dimana *k-means clustering* merupakan metode pengelompokan yang mudah, efisien dan efektif. Selain metode *k-means* juga dilakukan metode statistik dalam menentukan *standart deviasi*. Studi kasus dilakukan di desa Babadan, Bantul, Indonesia. Hasil dari penelitian ini adalah terbentuknya tiga *cluster* berdasarkan atribut usia, pendidikan, dan jenis pekerjaan. *Cluster* yang pertama dihasilkan *cluster* dengan usia 43,16 tahun dan pendidikan terakhir SLTA memiliki pekerjaan sebagai wirausaha. *Cluster* kedua dihasilkan *cluster* dengan usia 44,83 tahun dan pendidikan terakhir SLTA memiliki pekerjaan sebagai buruh. *Cluster* yang ketiga dengan usia 43,16 tahun dan pendidikan terakhir Srata 1 memiliki pekerjaan sebagai PNS. Sehingga pada *cluster* kedua yang menjadi prioritas dalam peningkatan kesejahteraan dengan memberikan tambahan pekerjaan sampingan
Kata Kunci: *K-Means Clustering*, Kelompok Ekonomi, Kesejahteraan, Pekerjaan Tambahan

PENDAHULUAN

Badan Pusat Pengelolaan Statistik setiap satu tahun melakukan pendataan social ekonomi nasional atau sering disebut susenas. Pendataan ini dilakukan empat kali dalam satu tahun (N et al. 2012). Pendataan ini dilakukan guna mengetahui keadaan ekonomi masyarakat. Jika tingkat ekonomi masyarakat rendah maka pemerintah perlu melakukan suatu tindakan untuk meningkatkan perekonomian masyarakat tersebut. Salah satu yang dilakukan oleh pemerintah adalah pemberian bantuan kepada rumah tangga ekonomi rendah. Namun masyarakat tidak dapat terus bergantung kepada bantuan pemerintah.

Kesejahteraan mengandung pengertian yang relatif, dinamis, dan kuantitatif. Rumusannya tidak pernah final karena akan terus berkembang seiring dengan perkembangan kebutuhan hidup manusia. Secara umum kesejahteraan dapat diartikan sebagai suatu keadaaan dimana segenap warga negara selalu berada dalam kondisi

serba kecukupan segala kebutuhannya, baik material maupun spiritual (Roestam, 1993)

Clustering merupakan salah satu metode pengelompokkan data. Data yang dikelompokkan adalah data yang memiliki karakteristik yang mirip. Salah satu metode yang sering digunakan adalah *clustering K-Means.* Algoritma K-means ini juga memiliki kelebihan yaitu dinilai cukup efisien, yang ditunjukkan dengan kompleksitasnya O(tkn), dengan catatan n adalah banyaknya obyek data, k adalah jumlah cluster yang dibentuk, dan t banyaknya iterasi. Biasanya, nilai k dan t jauh lebih kecil daripada nilai n. Selain itu, dalam iterasinya, algoritma ini akan berhenti dalam kondisi optimum local (Tang dkk, 2005)

Metode ini digunakan dalam bidang ekonomi untuk mengelompokkan data penduduk dengan perekonomian rendah sedang dan menengah (N et al. 2012), dalam mengelompokkan Kelompok Swadaya Masyarakat (Nugroho et al. 2012), dalam pengelompokkan provinsi din indonesia (Ramdhani et al. 2015). Selain itu juga digunakan untuk melakukan strategi marketing (Ong 2013), dan juga melakukan pemetakan tanaman padi (Felicia n.d.). Di bidang pendidikan metode *K-Means* ini juga digunakan dalam mengetahui potensi akademik mahasiswa (Syafrianto 2012). Dibidang transportasi digunakan untuk mengelompokkan jumlah penumpang bus trans jogja (Zahrotun 2015). Penggabungan statistic k-means juga pernah dilakukan (Suryana 2011)

Penelitian tentang pengelompokkan rumah tangga pernah dilakukan yaitu untuk mengelompokkan rumah tangga dalam tiga kondisi ekonomi rendah, sedang ataupun menengah. Akan tetapi pada proses pengelompokkan kondisi ekonomi rendah, sedang ataupun menengah didasarkan pada tingkat pendidikan dan pekerjaan warga. (N et al. 2012).

Dari melihat kelebihan metode K-Means dan beberapa penelitian sebelumnya maka dalam penelitian dilakukan pengelompokkan warga berdasarkan pada usia, pekerjaan dan tingkat pendidikan terakhir dengan metode statistik *K-means.* Dimana *K-means* digunakan untuk proses pengelompokkan data warga dan statistic khususnya nilai deviasi rata-rata mengetahui simpangan usia warga terhadap rata-ratanya. Hasil dari Statistik dan K-means ini digunakan untuk memberikan rekomendasi pekerjaan sampingan kepada warga keluarga yang memiliki ekonomi menengah dan bawah.

Teori Pendukung

Data Mining

Data mining adalah suatu metode pengolahan data untuk menemukan pola yang tersembunyi dari data tersebut. Hasil dari pengolahan data dengan metode *data mining* ini dapat digunakan untuk mengambil keputusan di masa depan. *Data mining* ini juga dikenal dengan istilah *pattern recognition* (Santosa 2007).

Data mining merupakan metode pengolahan data berskala besar oleh karena itu *data mining* ini memiliki peranan penting dalam bidang industri, keuangan, cuaca, ilmu dan teknologi. Secara umum kajian data mining membahas metode-metode seperti, *clustering*, klasifikasi, regresi, seleksi variable, dan market basket analisis (Santosa 2007). Tahapan data mining adalah sebagai berikut: (Han & Kamber 2001)

1. Pembersihan data
   Proses ini digunakan untuk membuang data yang tidak konsisten dan bersifat *noise* dari data awal yang ada.
2. Integrasi Data
   Menyatukan data yang terdapat di berbagai basisdata yang mungkin berbeda format maupun platform yang kemudian diintegrasikan dalam satu database *data warehouse*.
3. Seleksi Data
   Data yang terdapat dalam database *datawarehouse* kemudian direduksi dengan berbagai teknik. Proses reduksi diperlukan untuk mendapatkan hasil yang lebih akurat dan mengurangi waktu komputasi terutama utuk masalah dengan skala besar (*large scale problem*).
4. *Transformasi Data*
   Transformasi data diperlukan sebagai tahap *pre-procecing*, dimana data yang diolah siap untuk ditambang.
5. *Data mining*
   Data-data yang telah diseleksi dan ditransformasi ditambang dengan berbagai teknik. Proses *data mining* adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan fungsi-fungsi tertentu. Fungsi atau algoritma dalam *data mining* sangat bervariasi. Pemilihan fungsi atau algoritma yang tepat sangat bergantung pada tujuan dan proses pencaraian pengetahuan secara keseluruhan
6. Evaluasi pola
   Tahap ini merupakan bagian dari proses pencarian pengetahuan yang mencakup pemeriksaan apakah pola atau informasi

yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya.

7. Representasi Pengetahuan
   Langkah terakhir KDD adalah mempresentasikan pengetahuan dalam bentuk yang mudah dipahami oleh pengguna

Clustering

Pada dasarnya *clustering* merupakan suatu metode untuk mencari dan mengelompokkan data yang memiliki kemiripan karakteristik (*similarity*) antara satu data dengan data yang lain. *Clustering* merupakan salah satu metode data mining yang bersifat tanpa arahan (*unsupervised*), maksudnya metode ini diterapkan tanpa adanya latihan (*training*) dan tanpa ada guru (*teacher*) serta tidak memerlukan *target output*. Dalam data mining ada dua jenis metode *clustering* yang digunakan dalam pengelompokan data, yaitu *hierarchical clustering* dan *non-hierarchical clustering* (Santosa 2007).

K-Means

K-means merupakan salah satu metode *clustering non hirarki* yang berusaha mempartisi data ke dalam satu atau lebih *cluster* / kelompok berdasarkan jarak minimal data ke *centroid*. Metode ini mempartisi data, dimana data yang memiliki karakteristik yang mirip dikelompokkan ke dalam *cluster* yang sama (Agusta 2007)(Santosa 2007).

K-means merupakan metode *cluster* berbasis jarak yang membagi data ke dalam *k-cluster*, dan algoritma ini hanya bekerja pada data numerik. Pada awalnya algoritma ini mengambil sebanyak *k- centroid* secara random dari data, namun dalam penelitian ini penentuan *centroid* pertama kali diambil dari mean data sebanyak *k-centroid*. Hitung jarak setiap data terhadap masing-masing centroid, dalam hal ini penghitungan jarak digunakan rumus *euclidean*. Alokasikan data ke *cluster* yang memiliki jarak minimum ke *centroid*. Lakukan langkah tersebut hingga *cluster* stabil / tidak berubah.

Langkah-langkah melakukan *clustering* dengan metode *K-Means* adalah sebagai berikut (Santosa 2007):

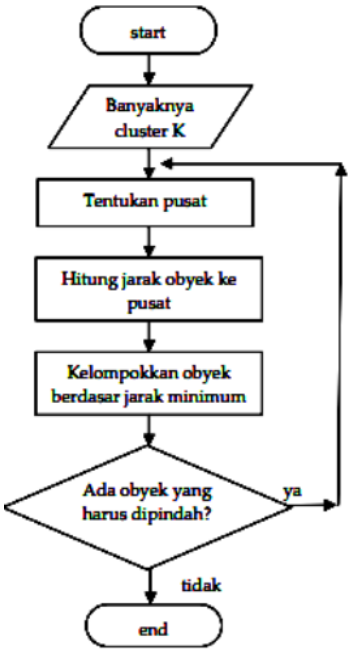
a. Pilih jumlah *cluster* k.
b. Inisialisasi k pusat *cluster* ini bisa dilakukan dengan berbagai cara. Namun yang paling sering dilakukan adalah dengan cara random. Pusat-pusat *cluster* nilai awal dengan angka-angka random,
c. Alokasikan semua data/ objek ke *cluster* terdekat. Kedekatan dua objek ditentukan berdasarkan jarak kedua objek tersebut. Demikian juga kedekatan suatu data ke *cluster* tertentu ditentukan jarak antara data dengan pusat *cluster*. Dalam tahap ini perlu dihitung jarak tiap data ke tiap pusat *cluster.*
d. Hitung kembali pusat *cluster* dengan keanggotaan *cluster* yang sekarang. Pusat *cluster* adalah rata-rata dari semua data/ objek dalam *cluster* tertentu. Jika dikehendaki bisa juga menggunakan median dari *cluster* tersebut. Jadi rata-rata (mean) bukan satu-satunya ukuran yang bisa dipakai.

$$C_k = \left(\frac{1}{n_k}\right) \sum \mathrm{d_i} \text{ (Handoyo et al. 2014)} (1)$$

Dimana $\mathbf{n_k}$ adalah jumlah dokumen dalam *cluster* **k** dan $\mathbf{d_i}$ adalah dokumen dalam *cluster* k

e. Tugaskan lagi setiap objek memakai pusat *cluster* yang baru. Jika pusat *cluster* tidak berubah lagi maka proses *clustering* selesai. Atau, kembali ke langkah c.

Flowchart proses *clustering K-Means* ditampilkan dalam Gambar 1.



Gambar 1. *Flowchart* algoritma metode *K-Means* (Syafrianto 2012)

Euclidean Distance

Ada beberapa cara yang dapat digunakan untuk mengukur jarak data ke pusat kelomok, di antaranya *Euclidean* menggunakan persamaan 1 (Handoyo et al. 2014):

$$d(i,j) = \sqrt{|xi1 - xj1|2 + |xi2 - xj2|2 + \ldots + |xin - xjn|2} \quad (2)$$

d(i,j) = jarak antara data ke i dan data ke j
xi1 = nilai atribut ke satu dari data ke i
xj1 = nilai atribut ke satu dari data ke j
n = jumlah atribut yang digunakan

METODOLOGI PENELITIAN

A. Data

Data yang digunakan dalam penelitian ini adalah data warga desa Babadan Bantul. Dengan jumlah 232 data warga.

B. Tahapan Data Mining

Dari 232 data, tidak semua digunakan dalam pengelompokan ini akan tetapi dilakukan beberapa proses dalam tahapan *data mining*.

1. Pembersihan data

Pembersihan data dilakukan dengan menghapus data yang tidak lengkap sehingga dari 232 data setelah dilakukan pembersihan data menjadi 152 data.

2. *Integrasi data*

Integrasi data dalam pengelompokan ini tidak dilakukan, karena data hanya berasal dari satu sumber saja, sehingga tidak perlu dilakukan integrasi

3. Seleksi Data

Dalam seleksi data ini, dari 11 atribut data awal maka setelah dilakukan seleksi data yang digunakan terdiri dari atribut no_id, jenis_kelamin, pekerjaan, usia dan pendidikan terakhir.

4. *Transformasi* Data

Transfomasi data di sini dilakukan dengan menyamakan beberapa data yaitu untuk atribut jenis kelamin maka digunakan data yaitu laki-laki dan perempuan. Untuk data atribut pekerjaan ada dua pekerjaan yang di jadikan satu yaitu wiraswasta dan wirausaha menjadi wirausaha.

Dalam transformasi data ini juga di gunakan deviasi rata-rata untuk mengetahui Penyebaran berdasarkan harga mutlak simpangan bilangan-bilangan terhadap rata-ratanya. Untuk menghitung deviasi rata-rata digunakan persamaan 3

$$DR = \sum_{i=1}^{n} \frac{|Xi - \overline{X}|}{n} * fi \quad (3)$$

Dimana :
DR = Deviasi Rata-rata
Xi = Data ke i
$\overline{X}$ = Nilai rata-rata
n = Jumlah data
fi = frekuensi kemunculan data ke i

5. Proses *Data Mining*

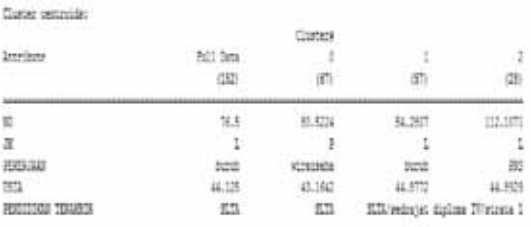
Dalam tahap ini dilakukan pengelompokkan data menggunakan metode *K-Means,* dimana setelah didapatkan data pengelompokkan maka akan dilakukan evaluasi pola dan representasi pengetahuan.

HASIL DAN PEMBAHASAN

Dalam penelitian ini setelah dilakukan tahapan *data mining* data yang siap dilakukan dalam pengelompokkan berjumlah 152 data. Dengan jumlah data yang memiliki pekerjaan wirausaha 26 warga, PNS sejumlah 15 warga, sedangkan sisanya adalah pekerjaan buruh lepas. Untuk pendidikan terakhir S1 berjumlah 14 warga, pendidikan terakhir Diploma berjumlah 11 warga, pendidikan terakhir SLTA berjumlah 70 warga dan sisanya adalah warga dengan pendidikan SLTP dan SD.

1. Pengelompokkan *K-Means*

Dengan hasil pengelompokkan menggunakan weka 3.6.11 di hasilkan titik pusat sebagai berikut :



Dari hasil weka 3.6.11 tersebut titik pusat yang terbentuk dapat dituliskan dalam table 1.

Tabel 1. Data titik pusat hasil pengelompokkan

Cluster	Titik Pusat	Jumlah data
1	usia 43,16 tahun dan pendidikan SLTA, Pekerjaan wirausaha	67
2	usia 44,83 tahun, dan pendidikan SLTA, Pekerjaan Buruh	57
3	usia 44,89, dan pendidikan Stata1, Pekerjaan PNS	28

Dengan melihat dari hasil pengelompokkan pada table 1. Maka tiga cluster tersebut adalah

a. *Cluster* yang pertama dapat diartikan bahwa kelompok pertama adalah warga, dengan usia 43,16 tahun dan pendidikan terakhir SLTA memiliki pekerjaan sebagai wirausaha.
b. *Cluster* yang kedua dapat diartikan bahwa kelompok kedua adalah warga, dengan usia 44,83 tahun dan pendidikan terakhir SLTA memiliki pekerjaan sebagai buruh.
c. *Cluster* yang ketiga dapat diartikan bahwa kelompok ketiga adalah warga, dengan usia 43,16 tahun dan pendidikan terakhir Srata 1 memiliki pekerjaan sebagai PNS.

2. Penghitungan Deviasi Rata-rata

Hasil Deviasi rata-rata dalam penelitian hanya dilakukan untuk data usia warga. Dihasilkan nilai rata-rata usia 44, 12 dan deviasi rata-rata 11.27. ini arinya adalah bahwa simpangan data usia terhadap nilai rata-rata usia semua warga adalah 11,27 tahun terhadap rata-rata usia 44.12 tahun.

Dengan melihat hasil *clustering k-means* maka dapat disimpulkan bahwa *cluster* ke dua dengan usia titik pusat usia 44.83 pendidikaan terakhir SLTA dan pekerjaannya buruh. Dan dari hasil penghitungan deviasi rata-rata maka warga pada cluster ke dua masih bisa diberikan pekerjaan sampingan. Hal ini dilihat dari simpangan dan titik pusatnya maka usia termuda adalah 33 tahun dan tertua adalah 45 tahun.

KESIMPULAN

Berdasarkan penelitian yang dilakukan, dapat disimpulkan bahwa algoritma *K-Means* bisa digunakan untuk mengelompokkan data warga berdasarkan jenis kelamin, usia, pekerjaan dan pendidikan terakhir. Dari data yang dilatih, didapatkan 3 kelompok yaitu :

1. *Cluster* yang pertama dengan usia 43,16 tahun dan pendidikan terakhir SLTA memiliki pekerjaan sebagai wirausaha.
2. *Cluster* yang kedua dengan usia 44,83 tahun dan pendidikan terakhir SLTA memiliki pekerjaan sebagai buruh.
3. *Cluster* yang ketiga dengan usia 43,16 tahun dan pendidikan terakhir Srata 1 memiliki pekerjaan sebagai PNS.

Peningkatan kesejahteraan dapat dilakukan dengan memberikan beberapa jenis pekerjaan tambahan pada data warga pada *cluster* kedua yang disesuaikan dengan usia warga yaitu antara 33 tahun dan 55 tahu Untuk penelitian selanjutnya sebaiknya

Dalam penelitian berikutnya diberikan rekomendasi jenis pekerjaan yang cocok ntuk warga berdasarkan dari hasil cluster k-means dari penelitian ini.

DAFTAR PUSTAKA

Agusta, Y., 2007. K-Means – Penerapan, Permasalahan dan Metode Terkait. *Jurnal Sistem dan informatika*, 3(Februari), pp.47–60.

Felicia, L., Penerapan Metode Clustering Dengan K-Means Untuk Memetakan Potensi Tanaman Padi Di Kota Semarang. , pp.1–5.

Han, J. & Kamber, M., 2001. *Data Mining Concepts and Techniques*, San Diego: Morgan Kaufmann.

Handoyo, R. et al., 2014. Perbandingan Metode Clustering Mengggunakan metode Single Linkage dan K-Means Pada Pengelompokkan Dokumen. *JSM STMIK Mikroskil*, 15(2), pp.73–82.

N et al., 2012. Aplikasi K-Means Untuk pengelompokkan Rumah tangga di Salatiga berdasarkan Data Susenas 2011. In *Pekan Imiah Dosen FEB-UKSW*. pp. 353–372.

Nugroho, C.A., Hendrawan, R. & Hafidz, I., 2012. Clustering Kelompok Swadaya Masyarakat (KSM) dalam Menentukan Kebijakan Bantuan Badan Pemberdayaan Masyrakat di Kota Surabaya dengan. *Jurnal Teknik ITS*, 1(1), pp.A368–A373.

Ong, J.O., 2013. Implementasi Algoritma K-Means Clustering Untuk Menetukan Strategi Marketing. *Jurnal Ilmiah Teknik Industri*, 12(Juni), pp.10–20.

Ramdhani, F., Hoyyi, A. & Mukid, M. abdul, 2015. Pengelompokkan Provinsi di Indonesia Berdasarkan Karakteristik Kesejahteraan Rakyat Menggunakan Metode K-Means cluster. *Jurnal Gaussian*, 4(4), pp.875–884.

Roestam, S. 1993. Pembangunan Nasional untuk Kesejahteraan Rakyat. Jakarta: Kantor Menteri Koordinator Bidang Kesejahteraan Rakyat Republik Indonesia.

Santosa, B., 2007. *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*, Yogyakarta: Graha Ilmu.

Suryana, N., 2011. Penggunaan Metode Statistik K-Means Clustering pada Analisis Peruntukan Lahan Usaha Tambang Berbasis Sistem Informasi Geografi. *Jurnal Statistik*, 11(1), pp.7–20.

Syafrianto, A., 2012. Perancangan aplikasi k-means untuk pengelompokan mahasiswa stmik elrahma yogyakarta berdasarkan frekuensi kunjungan ke perpustakaan dan ipk. *Jurnal Teknologi Informasi dan ilmu komputer (FAHMA)*.

Tang, ZhaoHui; MacLennan, Jamie. 2005. Data Mining with SQL Server 2005. Indiana Polis : Wiley Publishing

Zahrotun, L., 2015. Analisis PengelompokkanJumlah Penumpang Bus Trans Jogja Menggunakan metode Clustering K-Means dan Agglomerative Hierarchical Clustering (AHC). *jurnal Informatika*, 9(1), pp.1039–1047.