**LITERATUR REVIEW: TEKNIK CLUSTERING DENGAN MENGGUNAKAN METODE K-MEANS**



**Fathur Yudha Sulistya**

**15.01.55.0033**

**UNIVERSITAS STIKUBANK SEMARANG**

**2018**

**ABSTRAK**

Pengelompokan objek (objek clustering) adalah salah satu proses dari objek mining yang bertujuan untuk mempartisi objek yang ada kedalam satu atau lebih cluster objek berdasarkan karakteristiknya. Objek dengan karakteristik yang sama dikelompokkan dalam satu cluster dan objek dengan karakteristik berbeda dikelompokkan kedalam cluster yang lain. Algoritma K-Means Cluster Analysis termasuk dalam kelompok metode cluster analysis non hirarki, dimana jumlah kelompok yang akan dibentuk sudah terlebih dahulu diketahui atau ditetapkan jumlahnya. Algoritma K-Means Cluster Analysis menggunakan metode perhitungan jarak (distance) untuk mengukur tingkat kedekatan antara objek dengan titik tengah (centroid). Algoritma K-Means tidak terpengaruh terhadap urutan objek yang digunakan, hal ini dibuktikan ketika penulis mencoba menentukan secara acak titik awal pusat cluster dari salah satu objek pada permulaan perhitungan. Jumlah keanggotaan cluster yang dihasilkan berjumlah sama ketika menggunakan objek yang lain sebagai titik awal pusat cluster tersebut. Namun, hal ini hanya berpengaruh pada jumlah iterasi yang dilakukan. Algoritma K-Means Cluster Analysis pada dasarnya dapat diterapkan pada permasalahan dalam memahami perilaku konsumen, mengidentifikasi peluang produk baru dipasaran dan algoritma K-Means ini juga dapat digunakan untuk meringkas objek dari jumlah besar sehingga lebih memudahkan untuk mendiskripsikan sifat-sifat atau karakteristik dari masing-masing kelompok.

*Kata Kunci : Klastering, K-means, Algoritma Klastering,* *Klaster Analisis, Euclidian distance*

**ABSTRACT**

Object grouping (object clustering) is one of the processes of object mining that aims to partition an existing object into one or more cluster objects based on its characteristics. Objects with similar characteristics are grouped in one cluster and objects with different characteristics are grouped into other clusters. The K-Means Cluster Analysis algorithm belongs to a group of non-hierarchical cluster analysis methods, in which the number of groups to be established is already known or determined in number. The K-Means Cluster Analysis algorithm uses the distance calculation method to measure the degree of closeness between the object and the midpoint (centroid). The K-Means algorithm is not affected by the order of objects used, this is proven when the author tries to randomly determine the starting point of the cluster center of one of the objects at the beginning of the calculation. The number of cluster members generated amounts to the same when using other objects as the starting point of the cluster center. However, this only affects the number of iterations performed. K-Means Cluster Analysis algorithm basically can be applied to problems in understanding consumer behavior, identifying new product opportunities in the market and K-Means algorithm can also be used to summarize objects of large quantities making it easier to describe the properties or characteristics of each- each group.

*Kata Kunci : Clustering, K-means, Clustering algorithms,* *Cluster Analysis, Euclidian distance*

**PENDAHULUAN**

Analisis Cluster merupakan teknik multivariat yang mempunyai tujuan utama untuk mengelompokan objek-objek berdasarkan karakteristik yang dimilikinya. Analisis Cluster mengklasifikasi objek sehingga setiap objek yang paling dekat kesamaannya dengan objek lain berada dalam cluster yang sama. Cluster-cluster yang terbentuk memiliki homogenitas internal yang tinggi dan heterogenitas eksternal yang tinggi. Berbeda dengan teknik multivariat lainnya, analisis ini tidak mengestimasi set variabel secara empiris sebaliknya menggunakan set variabel yang ditentukan oleh peneliti itu sendiri. Fokus dari Analisis Cluster adalah membandingkan objek berdasarkan set variabel, hal inilah yang menyebabkan para ahli mendefinisikan set variabel sebagai tahap kritis dalam analisis cluster. Set variabel cluster adalah suatu set variabel yang merpresentasikan karakteristik yang dipakai objekobjek. Solusi Analisis Cluster bersifat tidak unik, anggota cluster untuk tiap penyelesaian/solusi tergantung pada beberapa elemen prosedur dan beberapa solusi yang berbeda dapat diperoleh dengan mengubah satu elemen atau lebih. Solusi cluster secara keseluruhan bergantung pada variabel-variabel yang digunakan sebagai dasar untuk menilai kesamaan. Penambahan atau pengurangan variabelvariabel yang relevan dapat mempengaruhi substansi hasil analisis cluster. Pada tulisan ini penulis menggunakan metode K-Means Cluster Analysis sebagai solusi untuk pengklasifikasian karakteristik dari objek. Alasan penggunaan algoritma K-Means diantaranya ialah karena algoritma ini memiliki ketelitian yang cukup tinggi terhadap ukuran objek, sehingga algoritma ini relatif lebih terukur dan efisien untuk pengolahan objek dalam jumlah besar. Selain itu algoritma K-Means ini tidak terpengaruh terhadap urutan objek.

**DASAR TEORI**

**Definisi Data**

Dalam Webster’s New World’s Dictionary tertulis bahwa datum (bentuk tunggal data) merupakan suatu yang diketahui atau dianggap. Dengan demikian, data dapat memberi gambaran tentang suatu keadaan atau persoalan. Sedangkan, data menurut kamus Oxford Dictionary adalah The Facts. Jadi, dapat disimpulkan bahwa data adalah sesuatu yang nyata diketahui atau dianggap yang dipakai untuk keperluan suatu analisa, diskusi, presentasi ilmiah atau tes statistik (Supranto, 2000).

**Data Kualitatif**

Data kualitatif secara sederhana dapat disebut data yang bukan berupa angka. Ciri utama data kualitatif didapat dengan cara menghitung, sehingga tidak memiliki nilai desimal. Selain itu data kualitatif memiliki ciri-ciri tidak bisa dilakukan operasi matematika, seperti penambahan, pengurangan, perkalian dan pembagian. Contoh data kualitatif adalah data gender, data golongan darah, data tempat tinggal atau data jenis pekerjaan. Agar dapat dilakukan proses pada data kualitatif atau non metric, data tersebut harus diubah ke dalam bentuk angka, proses ini dinamakan kategorisasi. Data kualitatif dibedakan menjadi dua jenis, yaitu data nominal dan data ordinal (Santoso, 2010). Data Nominal adalah jenis data kualitatif yang digunakan mengidentifikasi, mengklasifikasi, atau membedakan objek. Data nominal merupakan data yang paling rendah dalam level pengukuran data. Semua data memiliki posisi setara dalam arti tidak ada data yang memiliki tingkat yang lebih atau kurang dibandingkan dengan data yang lain. Jenis data nominal ini tidak memiliki jarak, urutan dan titik origin (Hidayat, 2011). Data Ordinal adalah jenis data kualitatif namun memiliki level lebih tinggi dari data nominal. Data ordinal memiliki karakteristik nominal tapi terdapat perbedaan derajat, urutan, atau peringkat dalam objek tersebut (posisi data tidak setara) (Hidayat, 2011).

**Data Kuantitatif**

Data kuantitatif dapat disebut sebagai data berupa angka dalam arti sebenarnya. Jadi, berbagai jenis operasi matematika dapat dilakukan pada data kuantitatif. Data kuantitatif merupakan data yang didapat dengan jalan mengukur sehingga bisa mempunyai nilai desimal. Contoh data kuantitatif adalah tinggi badan, usia, penjualan barang, dan sebagainya. Sebagai contoh, tinggi badan seseorang bisa bernilai 165 cm atau 165.5 cm. Seperti pada jenis data kualitatif, jenis data kuantitatif juga terbagi menjadi dua, yaitu data interval dan data rasio (Santoso, 2010). Data interval menempati level pengukuran data yang lebih tinggi dari data ordinal karena selain bisa bertingkat urutannya, urutan tersebut juga bisa dikuantitatifkan serta memiliki indikator jarak. Contohnya seperti pengukuran temperatur sebuah ruangan. Interval temperature ruangan tersebut adalah: a. Cukup panas jika temperatur antara 50 0C – 80 0C b. Panas jika temperatur antara 80 0C – 110 0C c. Sangat panas jika temperatur antara 110 0C – 140 0C Dalam kasus di atas, data temperatur bisa dikatakan data interval karena data mempunyai interval (jarak) tertentu, yaitu 30 0C. Data rasio merupakan data dengan tingkat pengukuran paling tinggi diantara jenis data lainnya. Data rasio adalah data yang bersifat angka dalam arti yang sebenarnya, bukan katagori seperti data nominal dan data ordinal sehinggga dapat dilakukan operasi matematika seperti penambahan, pengurangan, perkalian, dan pembagian. Perbedaan dengan data interval adalah data rasio memiliki indikator titik origin yang tidak dapat berubah (absolute). Contoh dari data rasio adalah jumlah suatu produk, jika jumlah produk 0 (nol) berarti memang tidak ada produk atau contoh lainnya adalah berat bedan dan tinggi badan, pengukuran-pengukurannya mempunyai nilai 0 (nol) yang sebenarnya. Misalnya jika berat badan 0 berarti memang tanpa berat.

**Jenis Data Menurut Sumbernya**

Pembagian jenis data menurut sumbernya didasarkan pada sumber perolehan data tersebut, yaitu data internal dan data eksternal (Supranto, 2007). Data internal adalah data yang dikumpulkan oleh suatu organisasi untuk menggambarkan keadaan atau kegiatan organisasi yang bersangkutan serta berguna untuk keperluan kegiatan harian dan pengawasan internal. Misalnya, data penjualan, data produksi suatu perusahaan, data keuangan, data kepegawaian, data pendidikan dan lain sebagainya. Data eksternal adalah data yang dikumpulkan untuk menggambarkan suatu keadaan atau kegiatan di luar organisasi tersebut. Contoh dari data eksternal seperti data jumlah penduduk dan data pendapatan nasional yang didapat dari kantor pusat statistik setempat. Suatu perusahaan memerlukan data eksternal seperti jumlah penduduk untuk memprediksi potensi pemermintaan, sedangkan data pendapatan nasional utnuk menentukan tingkat daya beli masyarakat yang berguna untuk dasar kebijakan tingkat harga.

**Jenis Data Menurut Cara Memperolehnya**

Berdasarkan cara memperolehnya, data dapat dibedakan menjadi dua jenis, yaitu data primer dan data sekunder (Supranto, 2000). Data primer adalah data yang dikumpulkan dan diolah sendiri oleh organisasi atau perorangan langsung dari objeknya. Misalnya suatu perusahaan ingin mengetahui konsumsi rata-rata suatu produk terhadap penduduk disuatu daerah dengan cara melakukan wawancara langsung kepada penduduk setempat. Data sekunder adalah data yang diperoleh dalam bentuk jadi dan telah diolah oleh pihak lainnya. Biasanya data sekunder ini dalam bentuk publikasi.

**Jenis Data Menurut Waktu Pengumpulannya**

Berdasarkan waktu pengumpulannya, data dapat dibedakan menjadi dua jenis, yaitu data cross section dan data berkala (time series) (Supranto, 2000). Data cross section adalah data yang dikumpulkan dalam suatu periode tertentu, biasanya menggambarkan keadaan atau kegiatan dalam periode tersebut. Misalnya, hasil sensus penduduk tahun 2012 menggambarkan keadaan Indonesia pada tahun 2012 menurut, umur, jenis kelamin, agama, tingkat pendidikan, dan sebagainya. Data berkala (time series) adalah data yang dikumpulkan dari waktu ke waktu. Tujuannya adalah untuk menggambarkan perkembangan suatu kegiatan dari waktu ke waktu. Misalnya, perkembangan produksi di suatu perusahaan selama lima tahun terakhir, perkembangan penjualan produk selama lima tahun terakhir, dan lain sebagainya. Jenis data ini juga sering disebut sebagai data historis.

**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).

**Clustering**

Pada dasarnya clustering merupakan suatu metode untuk mencari dan mengelompokkan data yang memiliki kemiripan karakteriktik (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 (taining) 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).

Hierarchical clustering adalah suatu metode pengelompokan data yang dimulai dengan mengelompokkan dua atau lebih objek yang memiliki kesamaan paling dekat. Kemudian proses diteruskan ke objek lain yang memiliki kedekatan kedua. Demikian seterusnya sehingga cluster akan membentuk semacam pohon dimana ada hierarki (tingkatan) yang jelas antar objek, dari yang paling mirip sampai yang paling tidak mirip. Secara logika semua objek pada akhirnya hanya akan membentuk sebuah cluster. Dendogram biasanya digunakan untuk membantu memperjelas proses hierarki tersebut (Santoso, 2010). Berbeda dengan metode hierarchical clustering, metode non-hierarchical clustering justru dimulai dengan menentukan terlebih dahulu jumlah cluster yang diinginkan (dua cluster, tiga cluster, atau lain sebagainya). Setelah jumlah cluster diketahui, baru proses cluster dilakukan tanpa mengikuti proses hierarki. Metode ini biasa disebut dengan K-Means Clustering (Santoso, 2010).

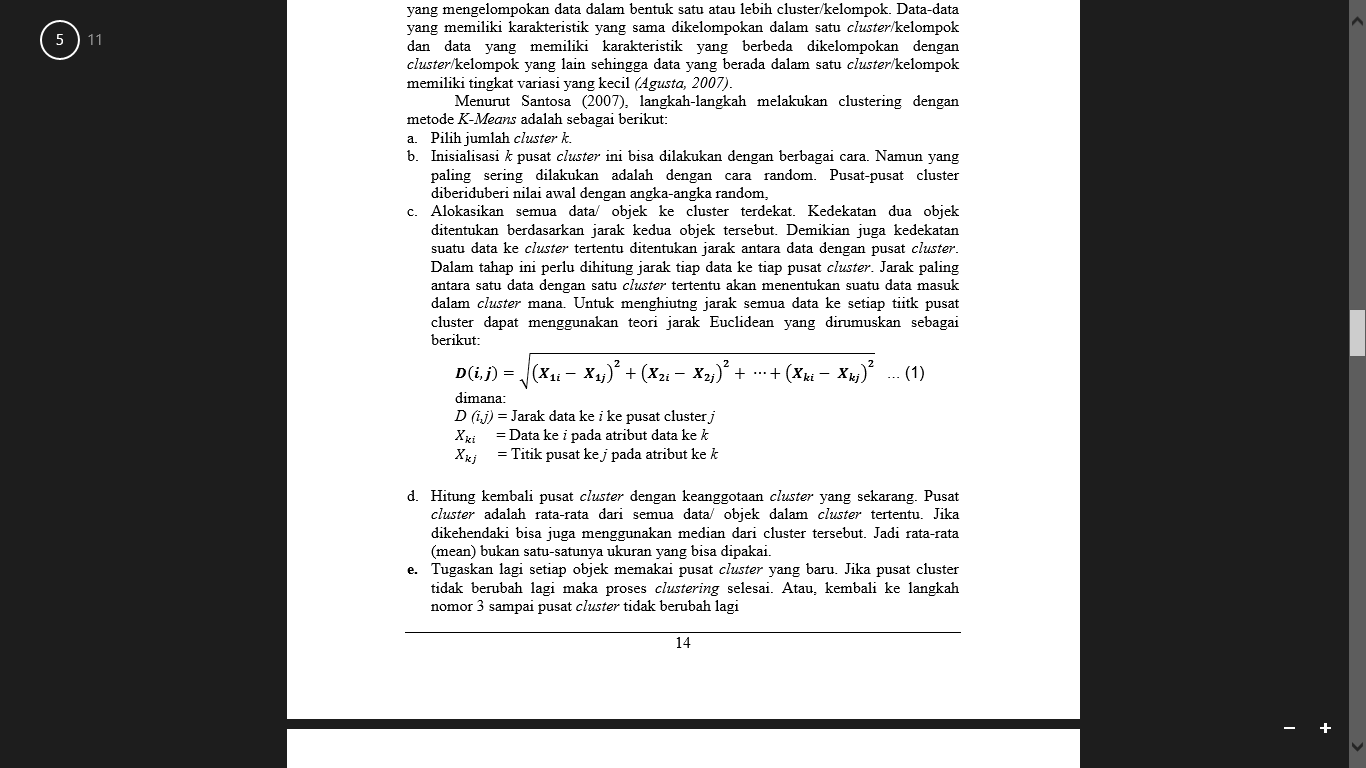
**K-means Clustering**

K-means clustering merupakan salah satu metode data clustering non-hirarki yang mengelompokan data dalam bentuk satu atau lebih cluster/kelompok. Data-data yang memiliki karakteristik yang sama dikelompokan dalam satu cluster/kelompok dan data yang memiliki karakteristik yang berbeda dikelompokan dengan cluster/kelompok yang lain sehingga data yang berada dalam satu cluster/kelompok memiliki tingkat variasi yang kecil (Agusta, 2007). Menurut Santosa (2007), langkah-langkah melakukan clustering dengan metode K-Means adalah sebagai berikut:

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 diberiduberi 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. Jarak paling antara satu data dengan satu cluster tertentu akan menentukan suatu data masuk dalam cluster mana. Untuk menghiutng jarak semua data ke setiap tiitk pusat cluster dapat menggunakan teori jarak Euclidean yang dirumuskan sebagai berikut:



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.

e. Tugaskan lagi setiap objek memakai pusat cluster yang baru. Jika pusat cluster tidak berubah lagi maka proses clustering selesai. Atau, kembali ke langkah nomor 3 sampai pusat cluster tidak berubah lagi.

**METODE**

Metode yang digunakan dalam penulisan

artikel ini adalah sebagai berikut.

1. Studi Literatur

Studi literatur merupakan langkah awal dalam penelitian ini, studi literatur dilakukan untuk melengkapi pengetahuan dasar dan teori-teori yang berasal dari buku-buku, jurnal, maupun media internet.

2. Pengumpulan Data

Data yang digunakan pada penelitian ini adalah beberapa sampel beras yang didapatkan dari pedagang beras secara acak.

3. Identifikasi Masalah

Tahap ini dilakukan setelah mendapatkan dataset yang sesuai untuk dilakukan proses clustering.

4. Pre Processing

Tahap pre processing merupakan tahap seleksi data yang bertujuan untuk mendapatkan data yang bersih. Tahap ini ditunjukkan pada proses binarisasi gambar hasil capture image.

5. Proses Clustering dengan Metode K

Means

Tahap ini merupakan proses dimana data-data yang sudah bersih dan tidak ada noise, di cluster-kan sesuai dengan ciri fisiknya yaitu panjang dan lebar beras.

6. Hasil

Hasil yang didapat setelah proses clustering adalah sejumlah cluster data yang dapat digunakan sebagai estimasi terhadap kualitas mutu beras.

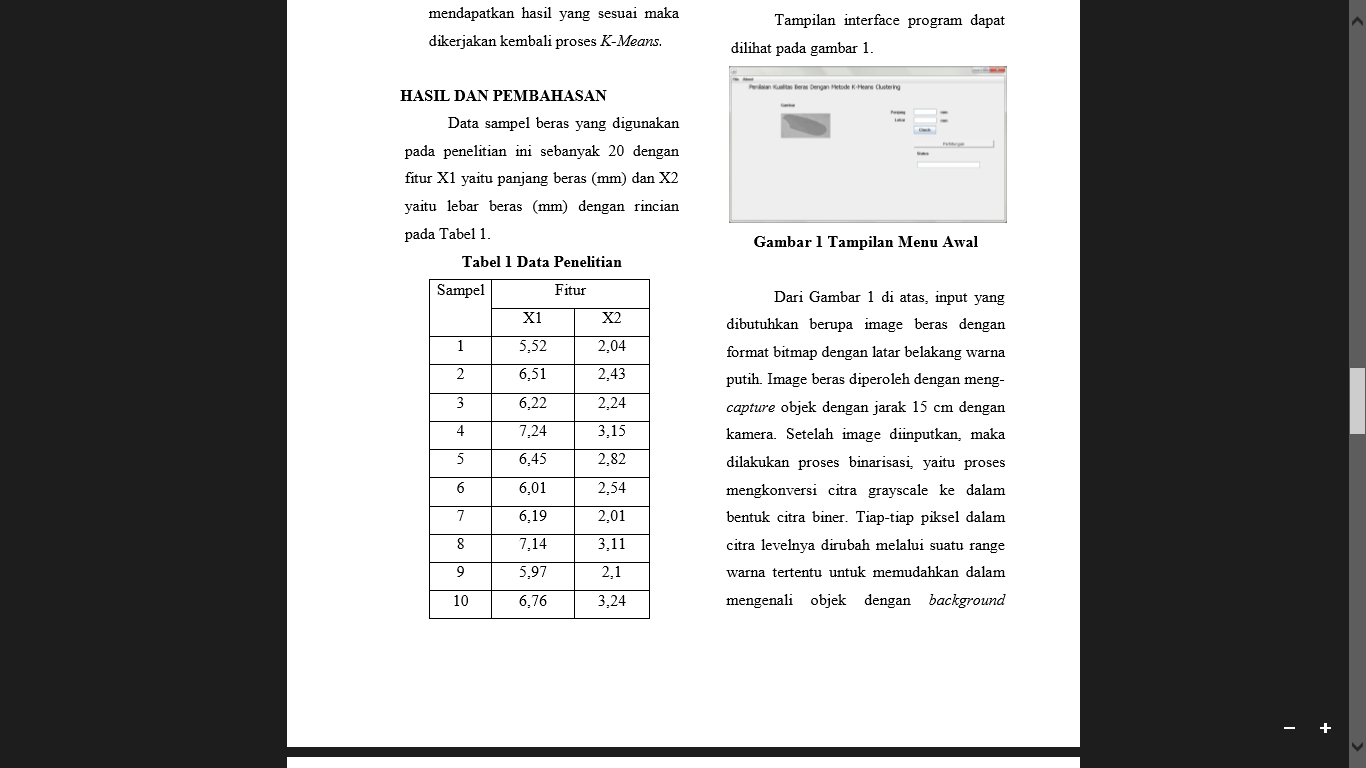
7. Evaluasi

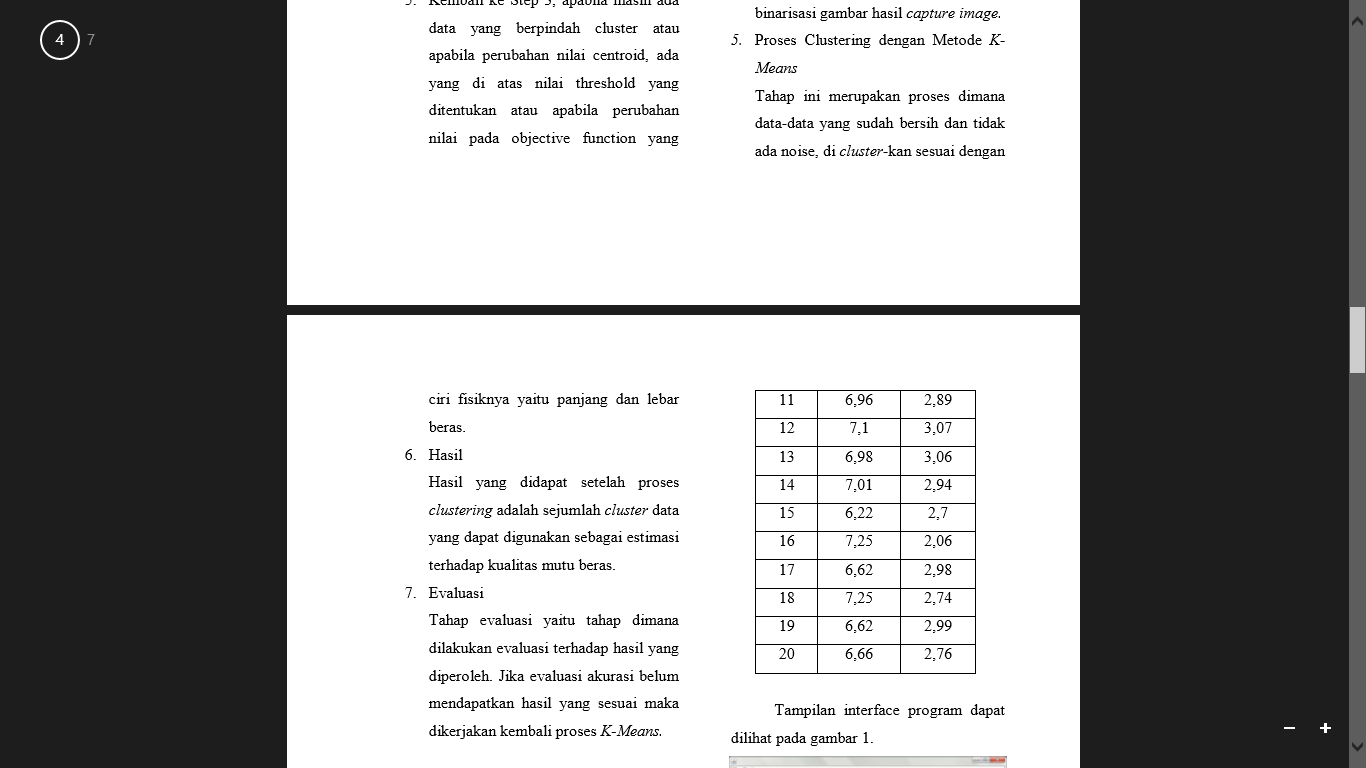
Tahap evaluasi yaitu tahap dimana dilakukan evaluasi terhadap hasil yang diperoleh. Jika evaluasi akurasi belum mendapatkan hasil yang sesuai maka dikerjakan kembali proses K-Means.

**HASIL DAN PEMBAHASAN**

Data sampel beras yang digunakan pada penelitian ini sebanyak 20 dengan fitur X1 yaitu panjang beras (mm) dan X2 yaitu lebar beras (mm) dengan rincian pada Tabel 1.

Tabel 1 Data Penelitian





**Proses K-Means Clustering**

Perhitungan pada program Klasterisasi Kualitas Beras dengan KMeans diawali dengan menentukan jumlah cluster. Penelitian ini menggunakan 3 cluster untuk menentukan kualitas beras, dengan keterangan bahwa cluster 1 adalah beras kualitas buruk, cluster 2 adalah beras kualitas sedang, dan cluster 3 adalah beras kualitas baik. Setelah menentukan jumlah cluster, kemudian menentukan 3 pusat cluster awal, yaitu pusat cluster 1 (5,5 ; 2), pusat cluster 2 (6,2 ; 2,45), dan pusat cluster 3 (7 ; 3,2). Langkah selanjutnya adalah mengalokasikan data ke dalam cluster, kemudian menghitung jarak setiap data terhadap setiap pusat cluster dengan menggunakan Persamaan 1. Suatu data akan menjadi anggota dari suatu cluster yang memiliki jarak terkecil dari pusat cluster-nya. Setelah itu menghitung pusat cluster baru sebagai acuan untuk iterasi berikutnya. Pada iterasi 1 dihasilkan 3 pusat cluster baru, pusat cluster 1 (5,89333;2,05), pusat cluster 2 (6,28199;2,546), dan pusat cluster 3 (6,96583;2,99917). Perhitungan pada iterasi selanjutnya sama dengan iterasi 1, hingga posisi data tidak mengalami perubahan. Pusat cluster yang dihasilkan pada iterasi 2 yaitu pusat cluster 1 (5,74499;2,07), pusat cluster 2 (6,26667;2,45667), dan pusat cluster 3 (6,96583;2,99917). Pusat cluster yang dihasilkan pada iterasi 3 yaitu pusat cluster 1 (6,28199;2,546), pusat cluster 2 (5,89333;2,05), dan pusat cluster 3 (6,96583;2,99917). Pusat cluster yang dihasilkan pada iterasi 4 yaitu pusat cluster 1 (5,89333;2,05), pusat cluster 2 (6,28199;2,546), dan pusat cluster 3 (6,96583;2,99917). karena pada iterasi selanjutnya posisi data tidak berubah, maka iterasi dihentikan dan hasil akhir yang diperoleh adalah 3 pusat cluster pada iterasi ke-4.

**KESIMPULAN**

K-Means Clustering merupakan metode klasterisasi berdasarkan persamaan karakteristik, dan merupakan metode yang sangat berguna karena mampu mentranslasi ukuran persamaan yang intuitif menjadi ukuran yang kuantitatif. Penelitian ini menggunakan 20 data uji, dimana ke-20 data tersebut dibagi menjadi 3 cluster dengan keterangan Cluster 1 merupakan beras kualitas buruk , Cluster 2 beras kualitas sedang, dan Cluster 3 beras kualitas baik. Dari hasil penelitian, didapatkan 3 pusat cluster akhir yaitu pusat cluster 1 (5,89333;2,05), pusat cluster 2 (6,28199;2,546), dan pusat cluster 3 (6,96583;2,999167) serta dihasilkan validasi sebesar 92,82% yang menunjukkan bahwa program ini dapat dijadikan sebagai acuan dalam klasterisasi kualitas beras.

**SARAN**

Untuk menampilkan hasil yang maksimal diperlukan lebih banyak data. Semakin banyak data yang dijadikan sampel akan semakin baik hasil yang diperoleh.

**DAFTAR PUSTAKA**

Agusta, Y. 2007. K-means - Penerapan, Permasalahan dan Metode Terkait. Jurnal Sistem dan Informatika Vol. 3 (Februari 2007): 47-60. Hidayat.

M. M. Putri dan K. Fithriasari, “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, pp. D13 - D18, 2015.

N. Atthina dan L. Iswari, “Klasterisasi Data Kesehatan Penduduk untuk Menentukan Rentang Derajat kesehatan Daerah denganMetode K-Means,” Seminar Nasional Aplikasi Teknologi Informasi

Rismawan dan Kusumadewi. Aplikasi K-Means untuk Pengelompokan Mahasiswa Berdasarkan Nilai Body Mass index (BMI) & Ukuran Kerangka. Yogyakarta: Seminar Nasional Aplikasi Teknologi Informasi 2008 (SNATI 2008) ISSN: 1907-5022; 2008.

Apriyanti, Nur Ridha. 2015. Pengolahan Citra Digital Landsat 8 dengan Algoritma K-Means Clustering (Studi Kasus: Banjarbaru, Kalimantan Selatan). Program S-1 Ilmu Komputer, Universitas Lambung Mangkurat: Banjarbaru.

Ediyanto, dkk. 2013. Pengklasifikasian Karakteristik dengan Metode K-Means Cluster Analysis. Buletin Ilmiah Mat. Stat. dan Terapannya (Bimaster) vol 02, No.2, hal 133-136.

Syakry, Sila Abdullah. 2013. Klasifikasi Citra Sidik Jari (Berminyak, Normal, Kering) Berdasarkan Nilai Pixel Menggunakan K-Means Klastering. Politeknik Negeri Lhokseumawe, Aceh.

Oscar,Johan.2013. Implementasi algoritma k-means clustering untuk menentukan strategi marketing president university. Jurnal Ilmiah Teknik Industri, Vol. 12, No. 1, Juni 2013.

Wardhani, Anindya.2016. implementasi algoritma k-means untuk pengelompokkan penyakit pasien pada puskesmas kajen pekalongan. jurnal transformatika, volume 14, nomor 1, juli 2016.

Nasari,fina dkk.2016. Penerapan Algoritma K-Means Clustering Untuk Pengelompokkan Penyebaran Diare Di Kabupaten Langkat .cogito smart journal/vol. 2/no. 2/desember 2016