ANALISIS CLUSTERING MENGGUNAKAN METODE K-MEANS DALAM PENGELOMPOKKAN PENJUALAN PRODUK PADA SWALAYAN FADHILA

Benri Melpa Metisen, Herlina Latipa Sari

Program Studi Teknik Informatika Fakultas Ilmu Komputer Universitas Dehasen Bengkulu
Jl. Meranti Raya No. 32 Kota Bengkulu 38228 Telp. (0736) 22027, 26957 Fax. (0736) 341139

ABSTRACT

This report describes the software application Tanagra in data mining. Tanagra is a data mining software that can be used to access some of the existing methods of data mining. This application uses the input dataset. In implementing this algorithm testing data used is the data item in the supermarket Fadhilla Bengkulu. In this application, use the application clustering using K -means algorithm. Of data processed by the data sample taken in Supermarkets Fadhilla Bengkulu, it produces two types of data sets. Ie low sales data and high sales data. So with the grouping of data is the self- Fadhilla can determine the type of goods sold and selling. So that goods in the warehouse does not accumulate.
Keyword : K-Means, Clustering, Data Mining.

INTISARI

Laporan ini menjelaskan tentang aplikasi perangkat lunak Tanagra pada data mining. Tanagra adalah *software* data mining yang dapat digunakan untuk mengakses beberapa metode data mining yang ada. Aplikasi ini menggunakan *dataset input*. Dalam melaksanakan pengujian algoritma ini data yang dipakai adalah data barang di swalayan Fadhilla Bengkulu. Dalam penerapan ini, digunakan penerapan *clustering* dengan menggunakan algoritma *K-means*. Dari data yang diolah dengan sampel data yang diambil di Swalayan Fadhilla Bengkulu, maka menghasilkan dua jenis kelompok data. Yaitu data penjualan rendah dan data penjualan tinggi. Sehingga dengan adanya pengelompokkan data ini pihak swalayan Fadhilla dapat mengetahui jenis barang yang laris terjual dan tidak. Sehingga barang yang ada di gudang tidak menumpuk.
Kata kunci: *K-Means, Clustering, Data Mining*.

I. PENDAHULUAN

Dalam era globalisasi, perkembangan kecanggihan teknologi yang semakin pesat merupakan aspek yang dapat dimanfaatkan untuk mencapai kemudahan-kemudahan, tidak terkecuali dalam arus informasi. Kecanggihan teknologi tersebut terlihat semakin marak dengan penggunaan komputer yang memang sudah sangat luas diberbagai bidang kehidupan misalnya di bidang pendidikan, kesehatan, hiburan, terlebih pada bidang bisnis yang semuanya itu menuntut penggunaan dari komputer.

Dalam toko, mini market dan swalayan masih ada proses-proses yang dilakukan secara manual sehingga sering terjadi kesalahan dalam pecatatan data-data yang ada dan juga kurangnya efesiensi waktu yang diperlukan. Seperti halnya pada Swalayan Fadhilla Bengkulu. Pada saat ini swalayan Fadhilla masih memproses data penjulannya secara manual. Disamping itu swalayan Fadhilla tidak dapat mengelompokkan produk yang laris dan yang tidak laris terjual. Sehingga kesulitan yang dialami yaitu seringnya kekurangan stok produk yang laku karena penjualannya tinggi. Dan menumpuknya produk yang tidak laku di gudang karena penjualnnya rendah.

Oleh karena itu dibutuhkan sistem informasi terkomputerisasi yang menunjang arus data dan informasi sesuai dengan kebutuhan dari proses-proses tersebut.

II. TINJAUAN PUSTAKA

A) Data Mining

Menurut Widodo (2013:1) Data mining adalah analisa terhadap data untuk menemukan hubungan yang jelas serta menyimpulkannya yang belum diketahui sebelumnya dengan cara terkini dipahami dan berguna bagi pemilik data tersebut.

Secara garis besar, data mining dapat dikelompokkan menjadi 2 kategori utama, yaitu:

1) *Descriptive mining*, yaitu proses untuk menemukan karakteristik penting dari data dalam satu basis data. Teknik data mining yang termasuk descriptive mining adalah clustering, asosiation, dan sequential mining.

2) *Predictive*, yaitu proses untuk menemukan pola dari data dengan menggunakan beberapa variable

lain di masa depan. Salah satu teknik yang terdapat dalam predictive mining adalah klasifikasi.

Secara sederhana data mining biasa dikatakan sebagai proses penyaring atau "menambang" pengtahuan dari sejumlah data yang besar. Istilah lain untuk data miing adalah *Knowlegde Discoveryin Database* (KDD). Walaupun data mining sendiri adalah bagian dari tahapan proses KDD seperti yang terlihat pada Gambar 1.

1) Data Selection

Menciptakan himpunan data target, pemilihan himpunan data, atau memfokuskan pada subset variabel atau sampel data, dimana penemuan (discovery) akan dilakukan. Hasil seleksi disimpan dalam suatu berkas, terpisah dari basis data operasional.

2) Pre-processing / Cleaning

Pre-processing dan cleaning data merupakan operasi dasar yang dilakukan seperti penghapusan noise. Proses cleaning mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak. Data bisa diperkaya dengan data atau informasi ekternal yang relevan.

3) Transformation

Merupakan proses integrasi pada data yang telah dipilih, sehingga data sesuai untuk proses data mining. Merupakan proses yang sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

4) Data mining

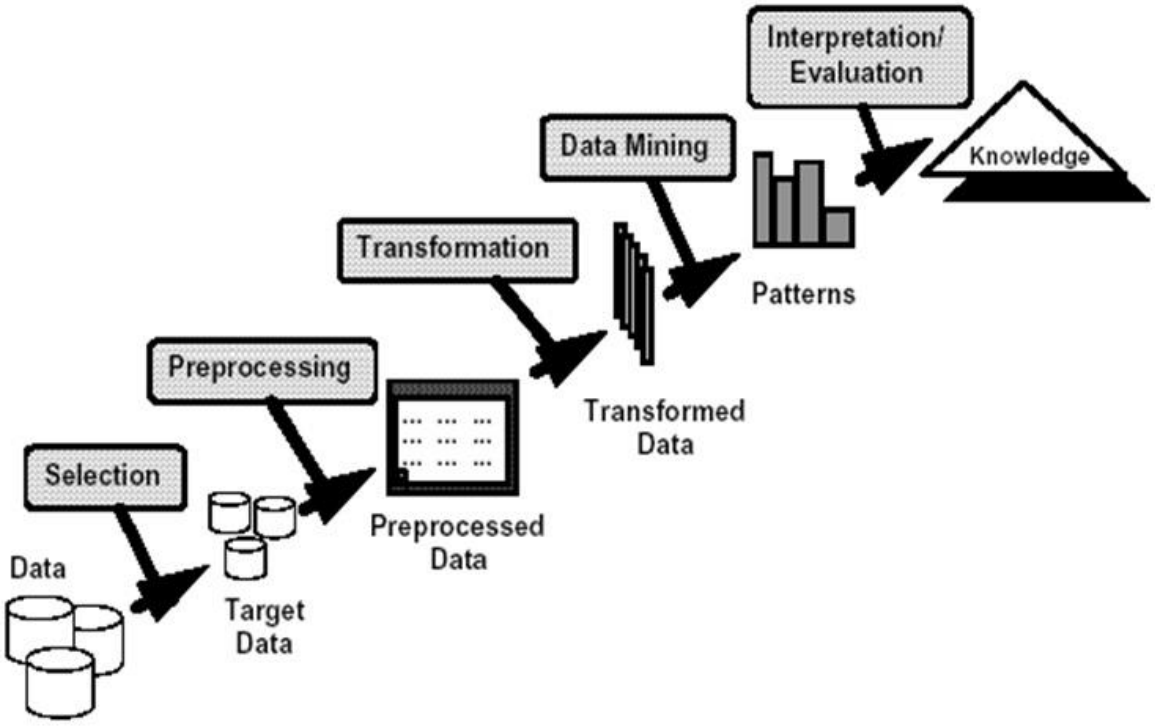
Pemilihan tugas data mining merupakan pemilihan goal dari proses KDD misalnya karakterisasi, klasifikasi, regresi, clustering, asosiasi, dan lain-lain. Pemilihan tugas data mining merupakan pemilihan goal dari proses KDD misalnya karakterisasi, klasifikasi, regresi, clustering, asosiasi, dan lain-lain. Pemilihan teknik, metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

5) Interpretation/ Evaluation

Yaitu penerjemahan pola-pola yang dihasilkan dari data mining. Pola informasi yang dihasilkan perlu ditampilkan dalam bentuk yang mudah dimengerti. Tahap ini melakukan pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya.

Tujuan dari data mining:

1) *Explonatory*, yaitu untuk menjelaskan beberapa kegiatan opservasi atau kondisi.
2) *Confirmatory*, yaitu untuk mengkonfirmasi suatu hipotesis yang telah ada.
3) *Exploratory*, yaitu untuk menganalisis data baru suatu relasi yang jangggal.



Gambar 1. KDD

B) Clustering

Menurut Widodo (2013:9) *Clustering* atau klasifikasi adalah metode yang digunakan untuk membagi rangkaian data menjadi beberapa group berdasarkan kesamaan-kesamaan yang telah ditentukan sebelumnya. Cluster adalah sekelom-pok atau sekumpulan objek-objek data yang similar satu sama lain dalam cluster yang sama dan disimilar terhadap objek-objek yang berbeda cluster. Objek akan dikelomokkan ke dalam satu atau lebih cluster sehingga objek-objek yang berada dalam satu cluster akan mempunyai kesamaan yang tinggi antara satu dengan yang lainnya.

Dengan menggunakan *clustering* ini, kita dapat mengkalsifikasikan daerah yang padat, menemu-kan pola-pola distribusi secara keseluruhan, dan menemukan keterkaitan yang menarik antara atribut data. Dalam data mining, usaha difokuskan pada metode-metode penemuan untuk cluster pada basis data berukuran besar secara efektif dan efisien. Beberapa kebutuhan clustering dalam data mining meliputi skalabilitas, kemampuan untuk menangani tipe atribut yang berbeda, mampu menangani dimensionalitas yang tinggi, mena-ngani data yang mempunyai noise, dan dapat diterjemakan dengan mudah. Pengolompokkan cluster ini disajikan pada Gambar 2.

C) Partitioning Clustering

Partitioning Clustering merupakan salah satu metode data mining yang bersifat tanpa arahan (unsupervised). Konsep dasar dari Partitioning Clustering adalah membagi n jumlah cluster ke dalam k cluster. Metode ini merupakan metode pengelompakan yang bertujuan mengelompokan objek sehingga jarak antar tiap objek ke pusat kelompok di dalam satu kelompok adalah minimum.

Cluster adalah kumpulan data dimana jika objek data yang terletak di dalam cluster harus memiliki kemiripan, sedangkan yang tida berada dalam satu cluster tidak memiliki kemiripan. Jika ada *n* objek pengamatan dengan *p* variable, maka sebelum di lakukan pengelompokan data atau objek, terlebih dahulu menentukan ukuran kedekatan sifat antar data. Ukuran data yang bisa di gunakan adalah jarak *Euclidean distance*, antar dua objek dari *p* dimensi pengamatan. Jika objek pertama yang diamati adalah:

$$X = [x_1, x_2, \ldots\ldots x_p] \text{ dan } Y = [y_1, y_2, \ldots\ldots y_p]$$

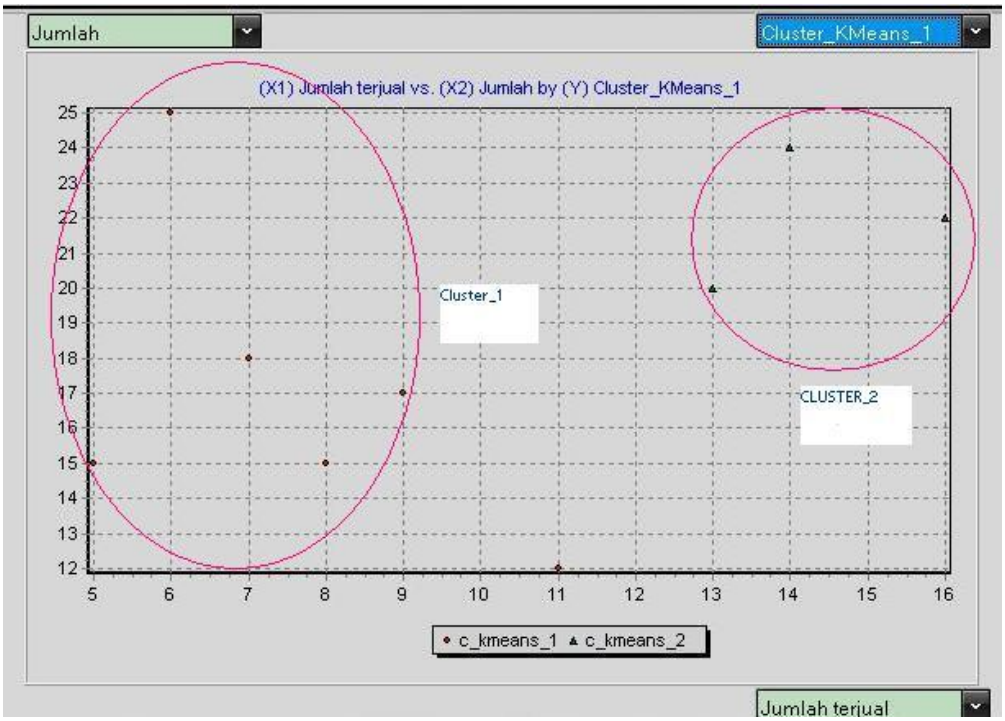
adalah :

$$\mathbf{D_{(x,y)}} = \sqrt{\sum_{j=1}^{p} (xj - yj)^2}$$

Dengan *d* alah jarak antara titik pada data *x* dan titik data *y*, dimana $x = x_1, x_2, \ldots\ldots x_i$ dan $y = y_1, y_2, \ldots\ldots y_i$ dan *j* mempresentasikan nilai atribut serta *p* merupakan dimensi atribut.

D) Metode K-means

K-Means merupakan saalah satu metode data clustering non hierarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih cluster atau kelompok sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu cluster yang sama dan data yang mempunyai



Gambar 2. Pengelompokkan Cluster

karakteristik yang berbeda dikelompokkan ke dalam kelompok yang lainnya.

K-Means adalah metode clustering berbasis jarak yang membagi data ke dalam sejumlah cluster dan algoritma ini hanya bekerja pada atribut numeric. Algoritma *K-Means* termasuk partitioning clustering yang memisahkan data ke *k* daerah bagian yang terpisah. Algoritma *K-Means* sangat terkenal karena kemudahan dan kemampuannya untuk mengcluster data yang besar dan data outlier dengan sangat cepat.Dalam algoritma *K-Means*, setiap data harus termasuk ke cluster tertentu dan bisa dimungkinkan bagi setiap data yang termasuk cluster tertentu pada suatu tahapan proses, pada tahapan berikutnya berpindah ke cluster lainnya.

Algoritma *K-Means* merupakan metode non hierarki yang pada awalnya mengambil sebagian banyaknya komponen populasi untuk dijadikan pusat cluster awal. Pada tahap ini pusat cluster dipilih secara acak dari sekumpulan populasi data. Berikutnya *K-Means* menguji masing-masing komponen di dalam populasi data dan menandai komponen tersebut ke salah satu pusat cluster yang telah didefinisikan tergantung dari jarak minimum antar komponen dengan tiap-tiap cluster. Posisi pusat cluster akan dihitung kembali sampai semua komponen data digolongkan kedalam tiap-tiap pusat cluster dan terakhir akan terbentuk posisi pusat cluster yang baru.

III. METODOLOGI PENELITIAN

A) Metode Analisa Sistem

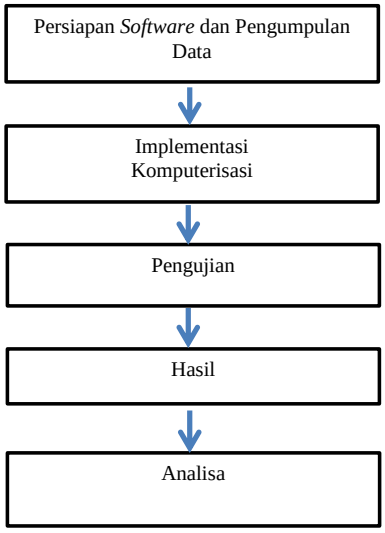
1) *Analisa Sistem Aktual*

Dari hasil pra penelitian yang dilakukan di Swalayan Fadhilla, didapatkan sistem yang sudah berjalan dan digunakan saat ini masih manual. Disamping itu, pemilik Swalayan Fadilla mempunyai kesulitan dalam pengklasifikasian penjualan produk yang laku dengan yang tidak laku.

2) *Analisa Sistem Baru*

Dari kekurang sistem yang sdang berjalan tersebut maka dalam penelitian ini diterapkanlah analisis data mining menggunakan *Clustring* menggunakan algoritma K-Means dan dengan pemrosesan menggunakan *software* data mining yaitu Tanagra.

Sehingga dengan mudah menentukan dan mengklasifikasikan penjualan produk yang laku dan kurang laku. Sehingga pemesanan barang yang kurang laku dapat dikurangi. Adapun sketsa tahap analisa pemrosesan menggunakan *software* Tanagra ditunjukkan pada Gambar 3.

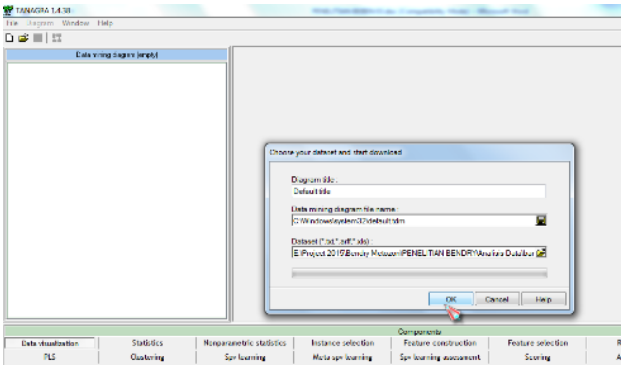


Gambar 3. Sketsa Tahap Analisa Pemrosesan Menggunakan *Software* Tanagra

IV. PEMBAHASAN

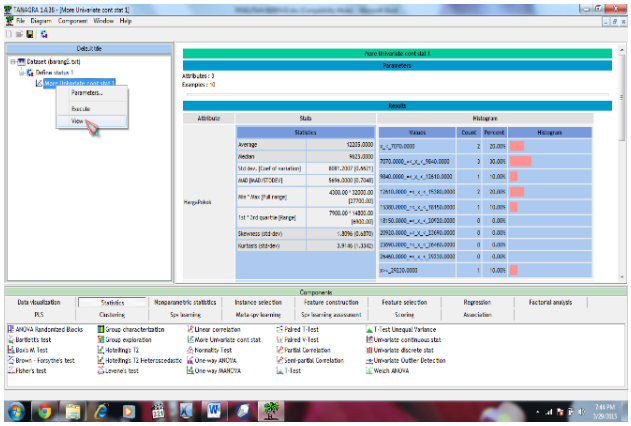
A) Hasil

Dari sketsa pemrosesan data yang ada pada bab sebelumnya, maka didapatkan hasil dari pemrosesan data menggunakan software Tanagra dan 2 buah software pembanding yaitu SPSS dan XLMiner. Dalam pengujian masing-masing software menghasilkan pengelompokkan 2 *cluster* produk.



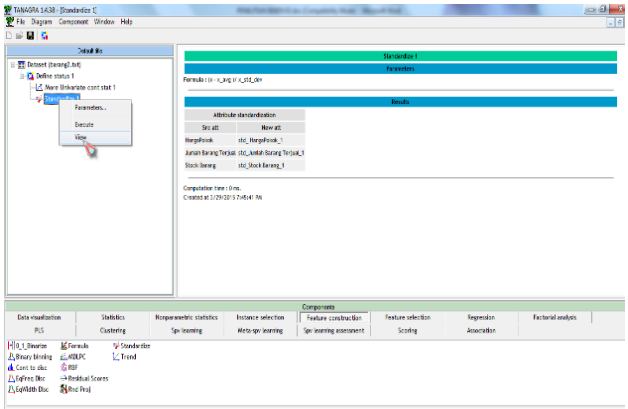
Gambar 4. *Input* Data **.txt*

Kemudian klik kanan dan VIEW. Maka hasilnya seperti Gambar 5.



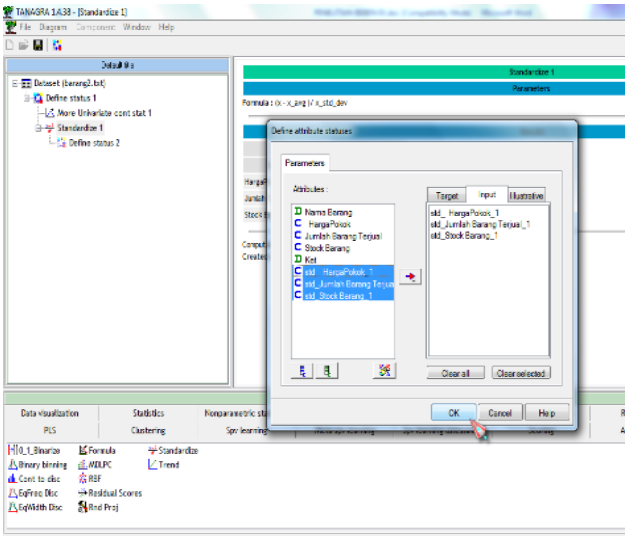
Gambar 5. *More Univariate*

Kemudian klik komponen STANDARDIZE (tab FEATURE CONSTRUCTION). Tarik ke DEFINE STATUS 1. Kemudian klik kanan dan VIEW. Seperti pada Gambar 6.



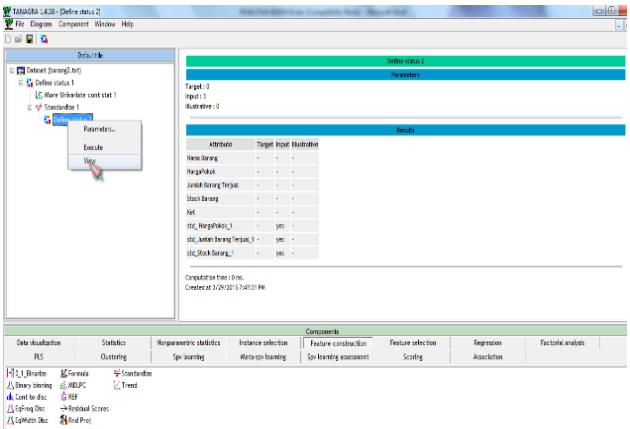
Gambar 6 Standardize di Define Status 1

Setelah itu tambahkan DEFINE STATUS baru ke dalam diagram, sehingga terbentuk DEFINE STATUS 2. Dan Atribut baru ditetapkan sebagai INPUT. Seperti pada Gambar 7.



Gambar 7. Variabel Input di Define Status 2

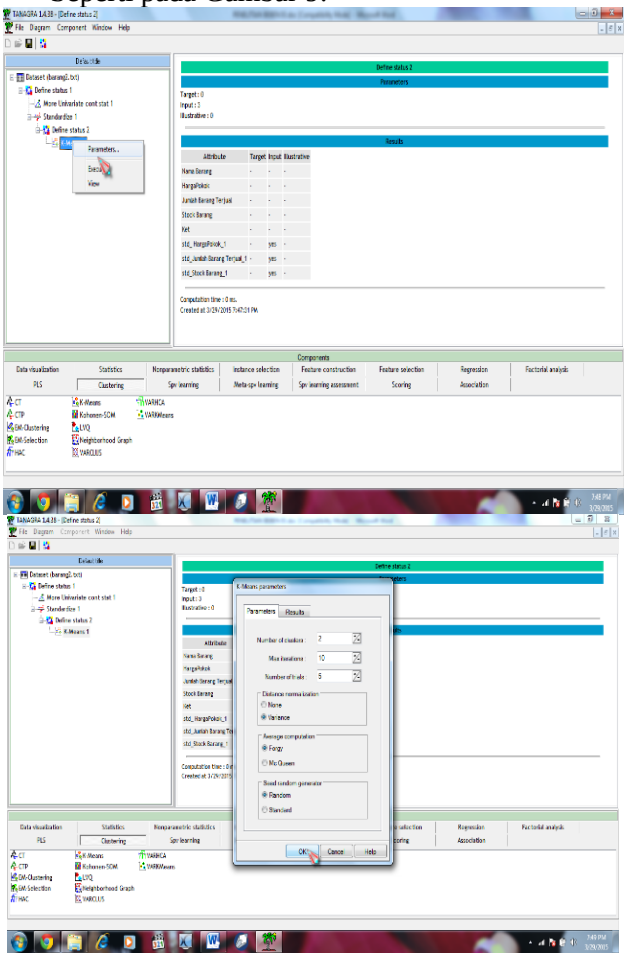
Maka hasilnya dapat dilihat sebagai berikut:



Gambar 8. View

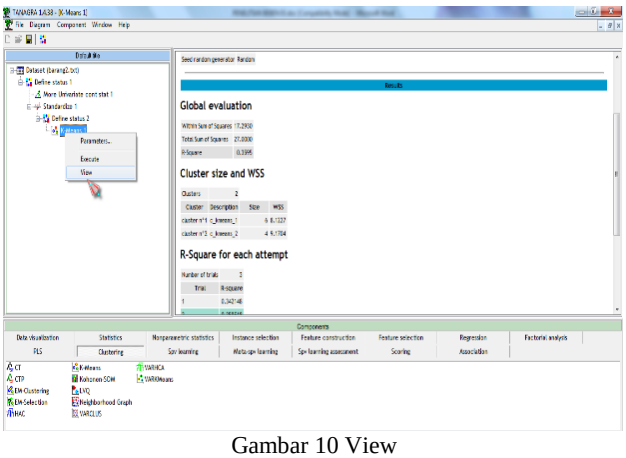
Pada tahap ini data dikelompokkan dengan menggunakan metode K-Means pada software Tanagra. Yaitu dengan tahapan sebagai berikut:

1) Insert Komponen K-means pada tab CLUSTERING.
2) Tarik ke DEFINE STATUS 2
3) Kemudian klik kanan dan klik PARAMETER. Seperti pada Gambar 9.



Gambar 9. K-Means

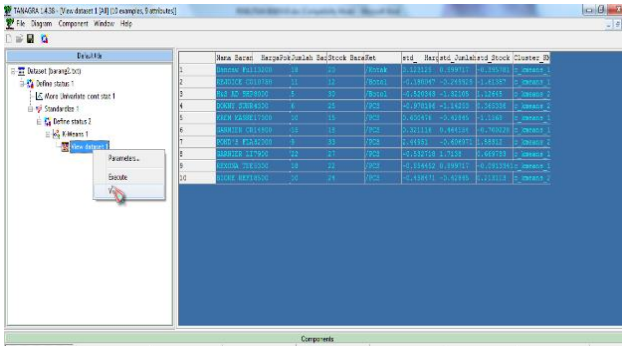
Pada tahap Custer ini data dijadikan menjadi 2 Cluster.Yang telah menggunakan standar variable.Untuk melihat hasilnya klik menu VIEW seperti gambar dibawah ini:



Gambar 10 View

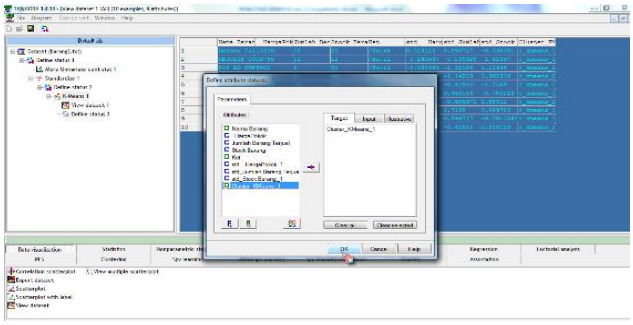
Pada tahap ini merupakan langkah awal pada proses *clustering*. Yang mana kita akan menginterpretasikan kelompok dan menentukan karakteristik setiap klaster dan membedakannya satu sama lain dengan tahapan sebagai berikut:

1) Tambahakan VIEW DATASET (tab DATVISUALIZATION)
2) Tarik ke K-means_1
3) Klik kanan dan klik VIEW. Maka didapatkan hasil sebagai berikut:



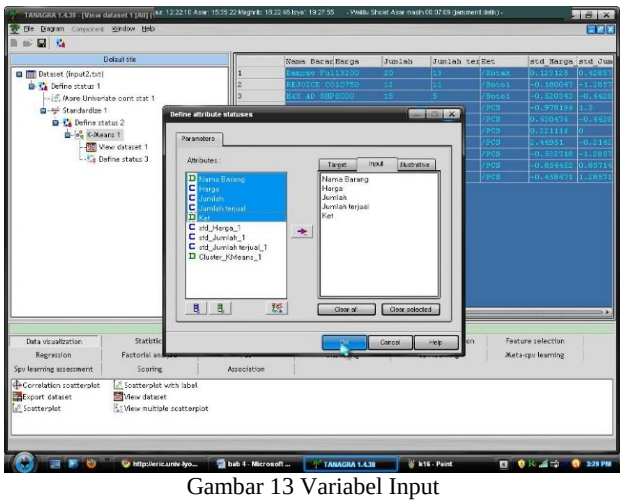
Gambar 11 View Dataset

Kemudian Tambahkan DEFINE STATUS baru ke dalam diagram,sehingga terbentuk DEFINE STATUS 3. Klik PARAMETER. Dan Data awal ditetapkan sebagai INPUT. Cluster K-Means-1 menjadi target. Seperti pada gambar berikut:



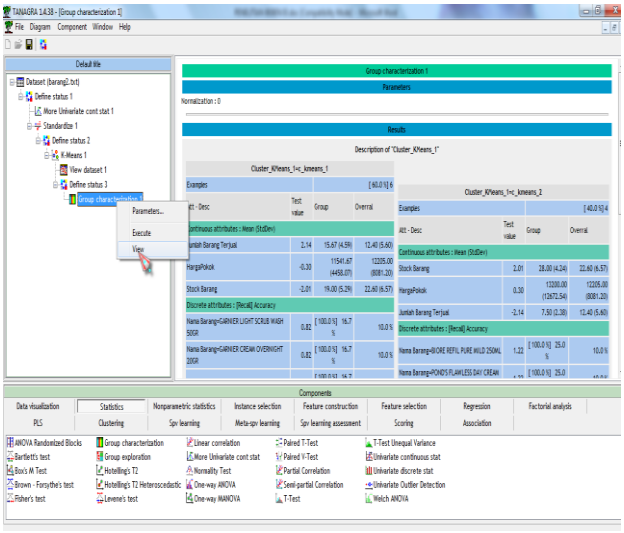
Gambar 12 Variabel Target

Berikut Variabel input



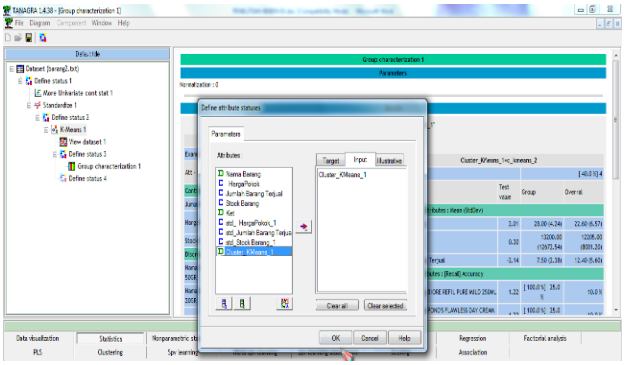
Gambar 13 Variabel Input

Kemudian Tarik Komponen GROUP CHARACTERIZATION (tab STATISTICS) ke DEFINE STATUS 3. Klik kanan dan klik VIEW

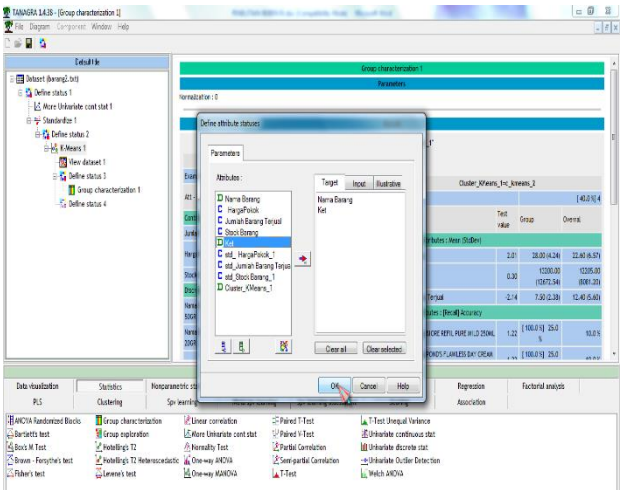


Gambar 14 Group characterization

Setelah itu tambahkan lagi DEFINE STATUS baru ke dalam diagram,sehingga terbentuk DEFINE STATUS 4. Data awal ditetapkan sebagai TARGET.Cluster K-Means-1 menjadi INPUT. Seperti gambar berikut:

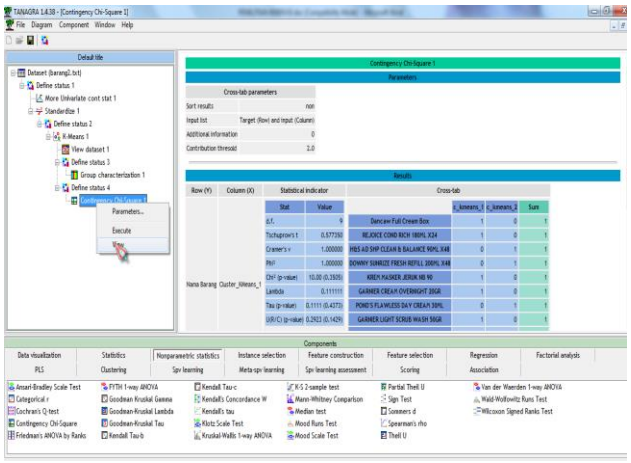


Gambar 15. Variabel Input Define Status 4



Gambar 16 Variabel Target di Define Status 4

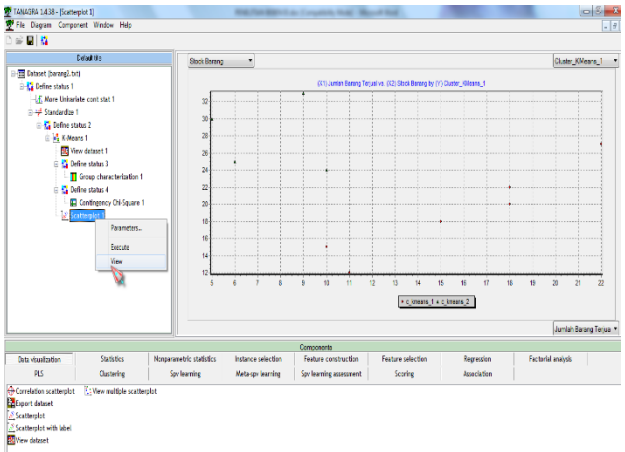
Setelah itu tambahkan *component* CONTIGENCY CHI-SQUARE (tab *nonparametric statistics*) ke dalam diagram. Tarik ke DEFINE STATUS 4. Kemudian klik kanan dan klik VIEW. Seperti pada gambar berikut:



Gambar 17 Contigency chi-squere

Kemudian *Scatter plot* atau *scatter diagrams* disebut juga diagram pencar. Diagram ini memberikan gambaran hubungan diantara 2 buah kelompok data,atau diagram pencar adalah grafik yang menunjukkan hubungan antara dua kelompok data yang jumlahnya sama, dimana untuk setiap nilai x terdapat nilai pasangannya y. Tahapan Scatter Plot adalah sebagai berikut:

1) Tambahkan SCATTERPLOT (tab DATA VISUALIZATION).
2) Tarik ke K_means_1
3) Kemudian klik kanan dan klik VIEW. Seperti pada gambar berikut:

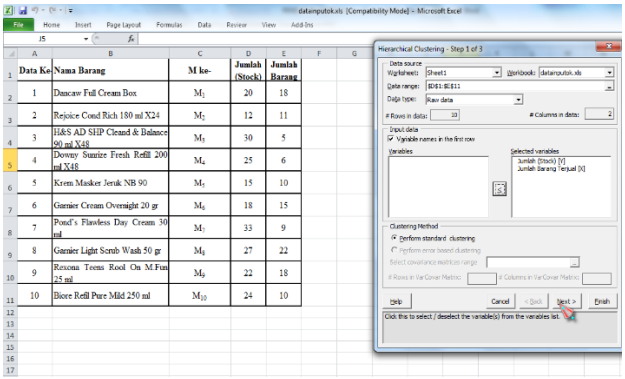


Gambar 18. Scatter Plot

Hasil dari proses *data mining* menggunakan *XLMiner* ditampilkan *worksheet* baru dengan nama *Hierarhical Clustering* seperti Gambar 19 Informasi yang dihasilkan di tampilkan dalam 1 tabel dimana tabel tersebut memberikan infomasi mengenai data yang digunakan. Untuk menentukan *Hierarhical Clustering* selanjutnya sama seperti proses yang telah dilakukan seperti langkah-langkah yang diatas karena data yang sudah di *Hierarhical Clustering* sama hasilnya.

Untuk menentukan *clustering* yang telah di *Hierarhical Clustering* maka memerlukan langkah-langkah sebagai berikut :

Data *input* yang telah di transformasikan dan siap untuk menentukan algoritma *Hierarhical Clustering*, Langkah yang selanjutnya penulis menetukan *Hirarki Clustering*, input *variables* dengan memilih jumlah dari data transaksi penjualan, lalu klik *next* seperti gambar dibawah ini.



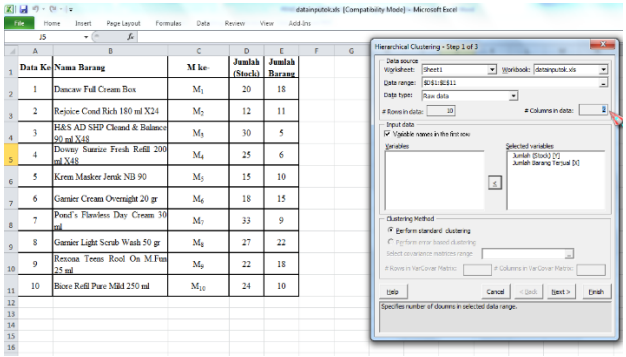
	Data Ke	Nama Barang	M ke-	Jumlah (Stock)	Jumlah Barang
2	1	Duncaw Full Cream Box	M_1	20	18
3	2	Rejoice Cond Rich 180 ml X24	M_2	12	11
4	3	H&S AD SHP Cleand & Balance 90 ml X48	M_3	30	5
5	4	Downy Sunrize Fresh Refill 200 ml X48	M_4	25	6
6	5	Krem Masker Jeruk NB 90	M_5	15	10
7	6	Garnier Cream Overnight 20 gr	M_6	18	15
8	7	Pond's Flawless Day Cream 30 ml	M_7	33	9
9	8	Garnier Light Scrub Wash 50 gr	M_8	27	22
10	9	Rexona Teens Rool On M.Fus 25 ml	M_9	22	18
11	10	Biore Refil Pure Mild 250 ml	M_{10}	24	10

Gambar 19. Hierarchical Clustering

Tahapan ini merupakan langkah untuk menentukan algoritma *hirarki algoritma* :

1) *Worksheet* : data transaksi yang di proses menggunakan algoritma *hirarki algoritma* .
2) *Data Range*: jumlah *cell* Seleksi terhadap *cell*/data yang digunakan.
3) *Variables in data source* : merupakan kelompok data yang terdapat di dalam data yang sudah di transfaromasikan untuk menentuakan algoritma *hirarki algoritma* .
4) *Input Variables* : menentukan proses pemilihan *dataset.*
5) *Rows in data* : jumlah data cell yang ada di *XLMiner*
6) *Columns in data*: jumlah kolom yang ada di *XLMiner*

Untuk menetukan jumlah *cluster* yang di inginkan, disini penulis menentukan 2 jumlah *cluster* karena jumlah *cluster* menentukan hasil dari nilai data transaksi penjualan barang, *iterations* nya 10, lalu klik *finish* lihat gambar dibawah ini:



Gambar 20. Menentukan Jumlah Cluster

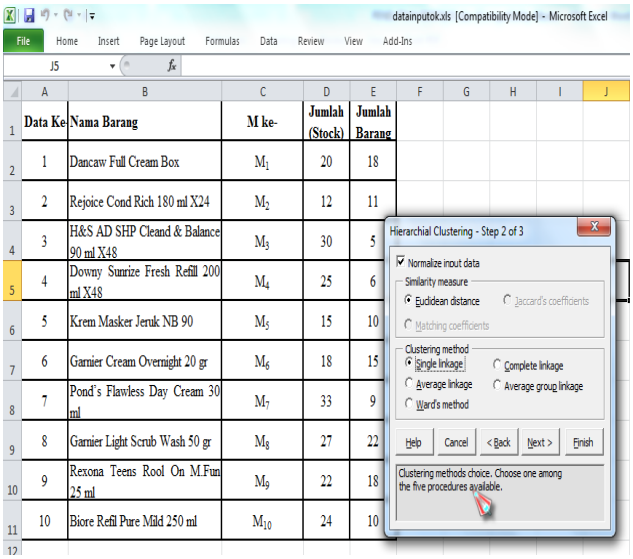
Euclidean distance : jarak kemiripan

Single linkage : jarak minimum yang diawali dengan mencari dua obyek terdekat dan keduanya membentuk cluster yang pertama

Average linkage : jarak rata-rata antar tiap pasangan obyek yang mungkin.

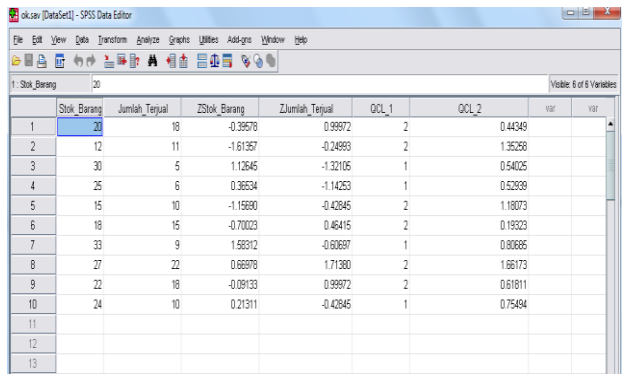
Complete linkage : jarak yang digunakan adalah jarak terjauh antar objek.

Apabila semua step *clustering* dilakukan maka pilih finish untuk menampilkan *Hirarki Clustering input*.



Data Ke-	Nama Barang	M ke-	Jumlah (Stock)	Jumlah Barang
1	Dancaw Full Cream Box	M_1	20	18
2	Rejoice Cond Rich 180 ml X24	M_2	12	11
3	H&S AD SHP Cleand & Balance 90 ml X48	M_3	30	5
4	Downy Sunrize Fresh Refill 200 ml X48	M_4	25	6
5	Krem Masker Jeruk NB 90	M_5	15	10
6	Garnier Cream Overnight 20 gr	M_6	18	15
7	Pond's Flawless Day Cream 30 ml	M_7	33	9
8	Garnier Light Scrub Wash 50 gr	M_8	27	22
9	Rexona Teens Rool On M.Fun 25 ml	M_9	22	18
10	Biore Refil Pure Mild 250 ml	M_{10}	24	10

Gambar 21 XLMiner Hirarki Clustering



Gambar 22. Hasil Pengujian SPSS

Dari hasil pengujian menggunakan *software SPSS* di atas dapat disimpulkan hasil cluster sebagai berikut:

Tabel 1 Hasil Pengujian SPSS

Data Ke-	Nama Barang	Jumlah (Stock) [Y]	Jumlah Barang Terjual [X]
Clustr K_Means 1(Tidak laris)	H&S AD SHP Cleand & Balance 90 ml X48	30	5
	Downy Sunrize Fresh Refill 200 ml X48	25	6
	Pond's Flawless Day Cream 30 ml	33	9
	Biore Refil Pure Mild 250 ml	24	10
Cluster K_Means 2(Laris)	Dancaw Full Cream Box	20	18
	Rejoice Cond Rich 180 ml X24	12	11
	Krem Masker Jeruk NB 90	15	10
	Garnier Cream Overnight 20 gr	18	15
	Garnier Light Scrub Wash 50 gr	27	22
	Rexona Teens Rool On MFun 25 ml	22	18

Dari hasil pengujian di atas, dapat disimpulkan bahwasanya Cluster 1, terdiri dari 4 produk tidak laris dan 6 produk yang laris di Cluster 2. Jika dibandingkan dengan pengujian menggunakan Tanagra dan *XLMiner*, pengujian menggunakan SPSS menghasilkan jumlah produk dalam cluster 1 dan cluster 2 yang berbeda. Namun, data produk yang diclusterkan tetap sama. Yaitu 4 produk yang tidak laris dan 6 produk yang laris.

V. PENUTUP

A) Kesimpulan

Berdasarkan pemrosesan data menggunakan beberapa *software* data mining di atas, maka penulis dapat mengemukakan beberapa kesimpulan antara lain :

Proses cluster secara hirarki dengan menggunakan metode *K-means* menghasilkan sebuah informasi gambaran penjualan terkluster atau terkelompok.

Hasil dari pemrosesan data menggunakan beberapa *software* data mining tersebut pada intinya sama. Yaitu menghasilkan kelompok data menjadi laris dan kurang laris

Hasil yang dicari secara manual eqiuvalen dengan hasil yang diproses dengan nonmanual.

B) Saran

Adapun saran yang dikemukakan dalam penelitian ini dalah sebagai berikut:

a. Setelah mengetahui tingkat pembelian produk, diharapkan para pelaku mini market lebih memperhatikan produk.
b. Untuk peneliti selanjutnya diharapkan dapat menggunakan analisis statistik yang lain dalam pengolahan data.

DAFTAR PUSTAKA

Widodo. 2004. *Psikologi Belajar*. Jakarta: RinekaCipta.

Feri *Sulianta*, Dominikus Juju, (*2010*), *Data Mining Meramalkan Bisnis*. *Perusahaan*, Jakarta : Elex Media Komputindo.

Ems, TIM. 2012. *Web Programming for Beginners*. Jakarta: PT Elex Media Komputindo.

Kurniawan, Rulianto. 2010. *PHP dan MySQL*. Palembang. Maxikom

Pahevi, Said Mirza. 2013. *Pembangunan Basis Data*. Jakarta: PT. Elex Media Komputindo