ANALISA DAN PENERAPAN DATA MINING UNTUK MENENTUKAN KUBIKASI AIR TERJUAL BERDASARKAN PENGELOMPOKAN PELANGGAN MENGGUNAKAN ALGORITMA K-MEANS CLUSTERING

Sri Tria Siska[1]

ABSTRACT

This study aims to determine how cubication water sold from water sales data subscribers to obtain a useful information for employees in grouping customers by type. In managing the data we are using Data Mining with k-means clustering algorithm. Data Mining is extracting information from large amounts of data. The information generated in the form of a group of customers who use the name of its relatively wasteful water, moderate and frugal, so the PDAM Kab.50 City can know how many customers who use its water wasteful so nantiknya be followed up. Tests were performed using RapidMiner application, so it can be determined clusters-clusters of customers who use its water wasteful, moderate and frugal.

Keywords: *Data Mining, Clustering Methods, K-means algorithm, cubication Water, RapidMiner*

INTISARI

Penelitian ini bertujuan untuk menentukan berapa kubikasi air yang terjual dari data penjualan air pelanggan untuk mendapatkan suatu informasi yang bermanfaat bagi karyawan dalam pengelompokan pelanggan berdasarkan jenisnya. Dalam mengelola data tersebut kita menggunakan metode *Data Mining* dengan algoritma *k-means clustering. Data Mining* merupakan penggalian informasi dari sejumlah data yang besar. Informasi yang dihasilkan berupa kelompok nama pelanggan yang penggunaan air nya tergolong boros, sedang dan hemat, jadi pihak PDAM Kab.50 Kota dapat mengetahui berapa banyak pelanggan yang penggunaan air nya boros sehingga nantiknya bisa di tindak lanjuti. Pengujian yang dilakukan menggunakan aplikasi *RapidMiner,* sehingga dapat ditentukan *cluster - cluster* pelanggan yang penggunaan air nya boros, sedang dan hemat.

Kata Kunci : *Data Mining, Metode Clustering, Algoritma K-means, Kubikasi Air, RapidMiner.*

[1] Dosen STMIK-AMIK Riau

PENDAHULUAN

Air bersih adalah air yang dapat dipergunakan untuk berbagai keperluan pada sektor rumah tangga seperti untuk mandi, mencuci dan lain-lain. Pada kawasan daerah Kab.50 Kota dimana tingkat ekonomi, pekerjaan dan status sosialnya yang bermacam-macam. Pengelompokan tersebut didasarkan pada m^3 air yang terpakai selama satu bulan. Untuk mengelola data tersebut, dibutuhkan metode yang bisa digunakan untuk menggali informasi dari data tersebut. Metode tersebut dikenal dengan *Data Mining*.

Data Mining merupakan penggalian informasi dari sejumlah data yang besar. Teknik pengelompokan data dapat menggunakan *K-Means Clustering.* Menurut algoritma ini kita terlebih dahulu kita harus memilih data nilai *k* sebagai pusat *cluster* awal, kemudian menghitung jarak antara setiap nilai data pada pusat *cluster* dan menetapkan *cluster* terdekat, selanjutnya memperbarui rata-rata dari semua kelompok, ulangi proses ini sampai kriteria tersebut tidak ada pertandingan.

PENDEKATAN PEMECAHAN MASALAH

Knowledge Discovery in Databases (KDD)

Data Mining, sering juga disebut *knowledge discovery in database* (KDD), adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam *set* data berukuran besar. Keluaran dari data mining bisa dipakai untuk memperbaikin pengambilan keputusan dimasa depan (Buulolo, 2013).

1. *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. 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 (Ong, 2013).

2. Clustering

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.

a. K-Means

K-Means merupakan algoritma *clustering* yang berulang-ulang. Algoritma *K-Means* dimulai Dengan pemilihan secara acak *K*, *K* di sini merupakan banyaknya *cluster* yang ingin dibentuk. Kemudian tetapkan nilai-nilai *K* secara random, untuk sementara nilai tersebut menjadi pusat dari *cluster* atau biasa disebut dengan *centroid*, mean atau "*means*". Hitung jarak setiap data yang ada terhadap masing-masing *centroid* menggunakan rumus *Euclidian* hingga ditemukan jarak yang paling dekat dari setiap data dengan *centroid*. Klasifikasikan setiap data berdasarkan kedekatannya dengan *centroid*. Lakukan langkah tersebut hingga nilai *centroid* tidak berubah (stabil) (Rismawan dan Sri Kusumadewi, 2008).

K-Means Clustering merupakan metode yang termasuk ke dalam golongan algoritma *Partitioning Clustering* (Handoyo, 2014). Langkah-langkah dari metode *K-Means* adalah sebagi berikut:

1. Tentukan *k* sebagai jumlah *cluster* yang di bentuk.
   Untuk menentukan banyaknya *cluster k* dilakukan dengan beberapa pertimbangan seperti pertimbangan teoritis dan konseptual yang mungkin diusulkan untuk menentukan berapa banyak *cluster*.
2. Bangkitkan *k Centroid* (titik pusat *cluster*) awal secara *random.*
   Penentuan *centroid* awal dilakukan secara *random*/acak dari objek-objek yang tersedia sebanyak *k cluster*, kemudian untuk menghitung *centroid cluster* ke-*i* berikutnya, digunakan rumus sebagai berikut :

$$v = \frac{\sum_{i=1}^{n} x_i}{n} \quad ;i = 1,2,3,...n$$

Di mana: *v* : *centroid* pada *cluster*
x_i : objek ke-*i*
n : banyaknya objek/jumlah objek yang menjadi anggota *cluster*

3. Hitung jarak setiap objek ke masing-masing *centroid* dari masing-masing *cluster.* Untuk menghitung jarak antara objek dengan *centroid* penulis menggunakan *Euclidian Distance.*

$$d(x,y) = \|x - y\| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad ;i = 1,2,3,...,n$$

Di mana : *xi* : objek *x* ke-*i*
yi : daya *y* ke-*i*
n : banyaknya objek

4. Alokasikan masing-masing objek ke dalam *centroid* yang paling terdekat.
   Untuk melakukan pengalokasian objek ke dalam masing-masing *cluster* pada saat iterasi secara umum dapat dilakukan dengan dua cara yaitu dengan *hard k-means,* dimana secara tegas setiap objek dinyatakan sebagai anggota *cluster* dengan mengukur jarak kedekatan sifatnya terhadap titik pusat *cluster* tersebut, cara lain dapat dilakukan dengan *fuzzy C-Means.*
5. Lakukan iterasi, kemudian tentukan posisi *centroid* baru dengan menggunakan persamaan. (1)
6. Ulangi langkah 3 jika posisi *centroid* baru tidak sama.
   Pengecekan *konvergensi* dilakukan dengan membandingkan matriks *group assignment* pada iterasi sebelumnya dengan matrik *group assignment* pada iterasi yang sedang berjalan. Jika hasilnya sama maka *algoritma k-means cluster analysis* sudah *konvergen*, tetapi jika berbeda maka belum *konvergen*

sehingga perlu dilakukan iterasi berikutnya.

Pada penerapan metode *K-Means Cluster Analysis*, data yang bisa diolah dalam perhitungan adalah data numerik yang berbentuk angka. Sedangkan data selain angka juga bisa diterapkan tetapi terlebih dahulu harus dilakukan pengkodean untuk mempermudah perhitungan jarak/kesamaan karakteristik yang dimiliki dari setiap objek. Setiap objek dihitung kedekatan jaraknya berdasarkan karakter yang dimiliki dengan pusat *cluster* yang sudah ditentukan sebelumnya, jarak terkecil antara objek dengan masing-masing *cluster* merupakan anggota *cluster* yang terdekat. Setelah jumlah *cluster* ditentukan, selanjutnya dipilih sebanyak 3 objek secara acak sesuai jumlah *cluster* yang dibentuk sebagai pusat *cluster* awal untuk dihitung jarak kedekatannya terhadap semua objek yang ada (Ediyanto *et al*, 2013).

3. Pengelompokan dengan *K-Means Clustering*

Salah satu metode dalam pengelompokan dokumen adalah *K-Means Clustering*. *K-Means Clustering* merupakan metode pengelompokan paling sederhana yang mengelompokkan data ke dalam *k* kelompok berdasar pada *centroid* masing-masing kelompok. Hanya saja hasil dari *K-Means* sangat dipengaruhi parameter *k* dan inisialisasi *centroid*. Pada umumnya *K-Means* menginisialisasi *centroid* secara acak. Namun metode yang diusulkan akan memodifikasi *K-Means* dalam inisialisasi *centroid* khususnya dalam memperbaiki performa dalam pengelompokan dokumen.

HASIL DAN PEMBAHASAN

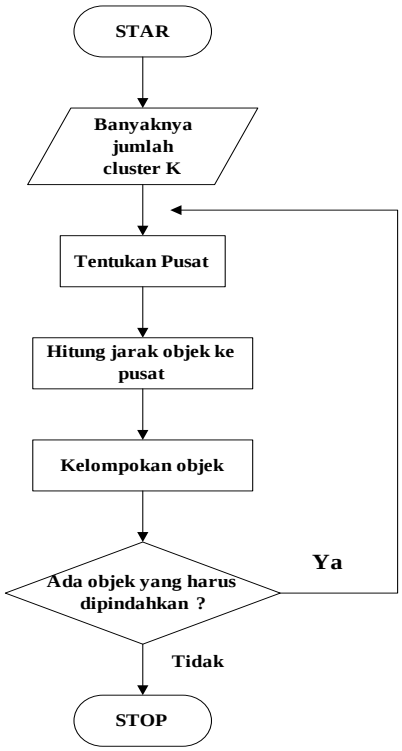
Analisa Data

Pada PDAM Kab.50 Kota memiliki data – data yang berkaitan dengan aktifitas penjualan air. Salah satunya adalah data yang digunakan untuk mencatat transaksi penjualan air pada PDAM. Data tersebut terdiri dari beberapa atribut misalnya kode pelanggan, no pelanggan, nama pelanggan, alamat, jenis pelanggan, meter bulan lalu, meter bulan sekarang, pemakaian sekarang dan keterangan. Dari data tersebut yang dijadikan sebagai atribut untuk melakukan pengolahan data untuk menentukan klasifikasi pelanggan.

1. Analisa *Clustering* Dengan Algoritma *K-Means*

K-Means termasuk dalam metode *data mining partitioning clustering* yaitu setiap data harus masuk dalam *cluster* tertentu dan memungkinkan bagi setiap data yang termasuk dalam *cluster* tertentu pada suatu tahapan proses, pada tahapan berikutnya berpindah ke *cluster* lain. *K-Means* memisahkan data ke *K* daerah bagian terkenal karena kemudian dan kemampuannya untuk mengklasifikasi data besar dan *outlier* dengan sangat cepat.

Berikut ini merupakan diagram *flowchart* dari algoritma *K-Means* dan menggambarkan langkah – langkah dalam algoritma *K-Means* dengan asumsi bahwa parameter input adalah jumlah data set sebanyak *n* data dan jumlah inisialisasi *centoid K=2* sesuai dengan penelitian.



Gambar 1. Flowchart Proses K-Means

Pengumpulan Data

Pada tahap pengumpulan data yang digunakan adalah dengan mewawancara langsung pimpinan PDAM Kab.50 Kota serta pengambilan data yang digunakan langsung oleh admininistrasi pada PDAM Kab.50 Kota. Data tersebut merupakan data rekapitulasi DRD air pada PDAM Kab.50 Kota dengan 93 nama pelanggan.

Tabel 1. Sampel Rekapitulasi DRD Air Unit Tanjung Pati Bulan : Mei 2015

NO	NAMA PELANGGAN	M3	TARIF	R.A.P
A1	ADE ASRUL	246	2.700	2.460
A2	DALIMA	22	1.800	220
A3	NIMA	15	1.350	150
A4	MELI	154	2.700	1.540
A5	SAMSINAR	47	2.700	470
A6	DEVI ANDRIANI	33	2.700	330
A7	BASTIAN	17	1.350	170
A8	EMI YUSNI	9	900	90
A9	ROSDIANA	59	2.700	590
A10	NURHAYATI	24	1.800	240
A11	H.NISAM II	7	900	70
A12	AMRI	10	900	100
A13	ROHANA	76	2.700	760
A14	DRS. SAFRUDIN D	10	900	100
A15	MAWARDI II	11	1.350	110
A16	SARMAN	8	900	80
A17	IRIANTI	51	2.700	510
A18	MAHDINAR	36	2.700	360
A19	ELI ANDRA	74	2.700	740
A20	REZI	2	900	20

Analisa Pengelompokkan Data

Pada penelitian ini penentuan jumlah kelompok atau *cluster* ditentukan berdasarkan pada pengelompokkan data rekapitulasi DRD air pada PDAM Kab.50 Kota untuk unik tanjung pati, yaitu M3 dan pembayaran air. Jadi pada penelitian ini *cluster* yang akan dibentuk adalah sebanyak tiga kelompok atau nilai *k=3*. Dimana atribut yang digunakan adalah sebanyak 2 buah atribut.

nalisa Proses Algoritma *K-Means*

Adapun prosesnya adalah sebagai berikut :

1. menentukan jumlah *cluster.*

Sebagaimana telah dijelaskan pada sub bab sebelumnya bahwa untuk mengelompokkan data pada pengujian yang pertama ini adalah sebanyak 3 *cluster,* sehingga dapat ditentukan untuk nilai *k* adalah *k=3.*

2. Menentukan titik pusat *cluster.*

Menentukan *centroid* awal dilakukan secara acak dari data/objek yang tersedia sebanyak jumlah *cluster k*. Nilai *centroid* awal pada penelitian ini dilakukan pemilihan secara acak, dimana jumlah *centroid* awal dilakukan sebanyak tiga *centroid* awal, nilai untuk C1 diambil dari baris data ke-14, nilai C2 diambil dari baris ke-91, nilai C3 diambil dari baris data ke-47. Berikut ini nilai *centroid* awal pada penelitian :

C1 = (154, 2.700, 1.540)
C2 = (2, 900, 20)
C3 = (51, 2.700, 510)

3. Hitung jarak setiap data yang ada terhadap setiap pusat *cluster.*

Untuk menghitung jarak setiap data yang ada terhadap pusat *cluster* terdapat beberapa cara, yaitu dengan menggunakan rumus *Manhattan/City Block*, dan *Euclidean Distance*. Sedangkan dalam penelitian ini penulis menggunakan rumus *Euclidean Distance* untuk melakukan perhitungan jarak setiap data terhadap titik pusat *cluster*. Berikut ini adalah contoh

perhitungan jarak dengan *Euclidean Distance* untuk iterasi 0.

Perhitungan jarak dari data ke-1 terhadap pusat *cluster* :

A(C1,A1) =

$$\sqrt{(154-246)^2+(2.700-2.700)^2+(1.540-2.460)^2}$$

= 924,59

A(C2,A1) =

$$\sqrt{(2-246)^2+(900-2.700)^2+(20-2.460)^2}$$

= 3041,9

A(C3,A1) =

$$\sqrt{(51-246)^2+(2.700-2.700)^2+(510-2.460)^2}$$

= 1959,7

4. Alokasikan masing-masing data ke dalam *centroid* yang paling terdekat.
   Dalam mengalokasikan kembali objek ke dalam masing-masing *cluster* didasarkan pada perbandingan jarak antara data dengan *centroid* setiap *cluster* yang ada, objek dialokasikan secara tegas ke dalam *cluster* yang mempunyai jarak ke *centroid* terdekat dengan data tersebut. Berikut ini adalah merupakan hal perbandingan jarak antara data dengan *centroid* setiap *cluster* yang ada. Perhitungan dilakukan terus sampai data ke-10 terhadap pusat *cluster.* Setelah dilakukan proses perhitungan maka akan didapatkan data selengkapnya adalah :
5. Lakukan iterasi, kemudian tentukan posisi *centroid* baru dengan cara menghitung rata-rata dari data-data yang berada pada *centroid* yang sama.
   Kemudian kita tentukan lagi pusat *cluster* dari data yang baru, caranya dengan menjumlahkan semua nilai M3 dan pembayaran air yang merupakan anggota dari *cluster* dan dibagi total jumlah anggota *cluster.*
6. Ulangi langkah ke-3 hingga nilai *centroid* tidak mengalami perubahan.

Karena proses yang dilakukan baru pada iterasi ke-0, maka perlu dilakukan beberapa iterasi lagi untuk dapat membandingkan nilai dari dua iterasi terakhir. Jika nilai dua iterasi terakhir tersebut telah sama, maka proses iterasi telah selesai, dan jika tidak maka ulangi lagi langkah berikutnya.

Tabel 2. Hasil Pengelompokan DataIterasi ke -0 Sampai Iterasi ke-2

ITERASI - 0			ITERASI - 1			ITERASI - 2		
C1	C2	C3	C1	C2	C3	C1	C2	C3
0	0	1	1	0	0	1	0	0
0	0	1	0	1	0	0	1	0
0	1	0	0	1	0	0	1	0
1	0	0	1	0	0	1	0	0
0	0	1	0	0	1	0	0	1
0	0	1	0	0	1	0	0	1
0	1	0	0	1	0	0	1	0
0	1	0	0	1	0	0	1	0
0	0	1	0	0	1	0	0	1
0	0	1	0	1	0	0	1	0
0	1	0	0	1	0	0	1	0
0	1	0	0	1	0	0	1	0
0	0	1	0	0	1	0	0	1
0	1	0	0	1	0	0	1	0
0	1	0	0	1	0	0	1	0
0	1	0	0	1	0	0	1	0
0	0	1	0	0	1	0	0	1
0	0	1	0	0	1	0	0	1
0	0	1	0	0	1	0	0	1
0	1	0	0	1	0	0	1	0

Dapat terlihat pada iterasi ke -1 dan iterasi ke - 2 tidak lagi mengalami perubahan pada titik *cluster*, sehingga dapat disimpulkan bahwa iterasi dapat dihentikan pada iterasi ke-2 dengan hasil :

Cluster pertama = 2 orang pelanggan
Cluster kedua = 11 orang pelanggan
Cluster ketiga = 7 orang pelanggan

Implementasi dan Pengujian

Pada bab ini merupakan tahapan tentang pembahasan metode

menggunakan *software RapidMiner.* Pada tahap ini akan digambarkan menggunakan *software RapidMiner* dan disamakan dengan pembuktian dari analisa metode terhadap permasalahan yang ada pada bab yang sebelumnya.

Tabel 3. Kelompok Masing-Masing *Cluster*

Kelompok (cluster)	Kelompok Mahasiswa	Jumlah Pelanggan
Cluster - 1	(3,5,8,9,10,19,22,23,24,30,33,36,43,47,48,49,52,53,56,59,60,65,80,82,84,85)	26 Pelanggan
Cluster – 2	(4,14,25)	3 Pelanggan
Cluster - 3	:(1,2,6,7,11,12,13,15,16,17,18,20,21,26,27,28,29,31,32,34,35,37,38,39,40,41,42,44,45,46,50,51,54,55,57,58,61,62,63,64,66,68,69,70,71,72,73,74,75,76,77,78,79,81,83,86,87,89,90,91,92,93)	64 Pelanggan

Berdasarkan tabel di atas dapat kita simpulkan bahwa dari pengelompokan data dapat diketahui kelompok nama pelanggan PDAM Kab.50 Kota yang pemakaian M3 airnya sedang terdapat pada *cluster* 1. Sedangkan pelanggan dengan pemakaian M3 air boros terdapat pada *cluster* 2 dan nama pelanggan yang berada dalam *cluster* 3 tersebut dikelompokkan pada pelanggan yang pemakaian M3 airnya hemat. Jadi pihak PDAM Kab.50 Kota dapat mengetahui siapa nama pelanggan yang pemakaian air nya boros sehingga pihak PDAM Kab.50 Kota bisa menindak lanjuti apa kesalahan yang terjadi dari pelanggan yang pemakaian air nya boros.

KESIMPULAN

Dari uraian yang telah ada pada bab – bab sebelumnya maka dapat ditarik kesimpulan sebagai berikut :

1. Metode *clustering* algoritma *k-means* dapat diterapkan pada kubikasi air terjual berdasarkan pengelompokan pelanggan di PDAM Kab.50 Kota, sehingga metode ini sangat membantu pihak PDAM Kab.50 Kota dalam menentukan pelanggan yang pemakaian air boros, sedang dan hemat.
2. Berdasarkan 3 *cluster* yang telah dilakukan pengujian menggunakan *RapidMiner* bahwa pelanggan terbanyak terdapat pada *cluster* ke-3 yang tergolong pada pemakaian air hemat.
3. Pengujian manual menggunakan sampel 20 *record* data pelanggan dan di *software RapidMiner* dengan menggunakan 20 *record* dan 93 *record* juga mendapatkan hasil yang sama. Di mana hasil *cluster* dari sampel 20 *record* adalah *cluster* 1 terdiri dari 7 pelanggan yang pemakaian air nya sedang, *cluster* 2 terdiri dari 2 pelanggan yang pemakaian air nya boros sedangkan *cluster* 3 terdiri dari 11 pelanggan yang pemakaian air nya hemat. Hasil dari sampel 20 *record* masuk ke dalam *cluster* yang sama pada jumlah data 93 *record*. Sehingga *software RapidMiner* ini sangat mempermudah pengelompokan pelanggan dengan menggunakan data sedikit maupun data yang lebih banyak dari sebelumnya.

DAFTAR PUSTAKA

[1] Buulolo, Efori. 2013. *Implementasi Algoritma Apriori Pada Sistem Persediaan Obat (Studi Kasus : Apotik Rumah Sakit Estomihi Medan)*. Pelita Informatika Budi Darma. Vol.IV, No 1.

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

[3] Handoyo, R, dkk. 2014. *Perbandingan Metode Clustering Menggunakan Metode Single Linkage Dan K-Means Pada Pengelompokan Dokumen.* ISSN. 1412-0100 Vol 15, No, 2, OKTOBER 2014.

[4] Lindawati. 2008. *Data Mining Dengan Teknik Clustering Dalam Pengklasifikasian Data Mahasiswa Studi Kasus Prediksi Lama Studi Mahasiswa Universitas Bina Nusantara.* Seminar Nasional Informatika 2008.

[5] Ong, Oscar. 2013*. Implementasi Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing President University*. Jurnal Ilmiah Teknik Industri. Vol. 12, No 1, 2010. ISSN 1412-6869.

[6] Rismawan, T dan Kusumadewi, S. 2008. *Aplikasi K-Means Untuk Pengelompokan Mahasiswa Berdasarkan Body Mass Index (BMI) & Ukuran Kerangka.* 21 Juni 2008. ISSN 1907-5022.

[7] Sijabat, Alimancon. 2015. *Penerapan Data Mining untuk Pengolahan Data Siswa dengan Menggunakan Metode Decision Tree.* Jurnal Informasi dan Teknologi Ilmiah. Volume 5 No 3. ISSN : 2339-210X.

[8] Tampubolon Kennedi, dkk. 2013. *Implementasi Data Mining Algoritma Apriori Pada Sistem Persedian Alat – Alat Kesehatan*. Informasi dan Teknologi Ilmiah(INTI).Volume 1, Nomor 1, Oktober 2013. ISSN : 2339-210X.

[9] Yusuf, A, dkk. 2013. *Support Vector Machines Yang Didukung K-Means Clustering Dalam Klasifikasi Dokumen*. Volume 11, Nomor 1, Januari 2013 : 13-16.