IMPLEMENTASI DATA MINING K-MEANS UNTUK MENGUKUR KEMAMPUAN LOGIKA MAHASISWA (STUDI KASUS : AMIK LABUHAN BATU)

Oleh :

Marnis Nasution
Dosen AMIK Labuhan Batu Program Studi Manajemen Informatika

Abstract

K-Means algorithm is an algorithm that is popular in the world clusterering insdustri. This algorithm is compiled on the basis of a simple idea. Was originally defined several clusters to be formed. Any object or the first element in the cluster can be selected to serve as the center (centroid point) cluster. Algorita K-Means will further repetition of the following steps until there is stability, based on the origin of school students learning achievement can be predicted so as to assist the college in the new admissions process and conduct class divisions.
Keywords: Data Mining, Clustering, K-Means, Kemampuan Logika

1. PENDAHULUAN

Clustering adalah metode yang digunakan dalam data mining yang cara kerjanya mencari dan menglompokkan data yang mempunyai kemiripan karakteristik antara data satu dengan data lainnya yang telah diperoleh. Ciri khas dari teknik data mining ini adalah mempunyai sifat tanpa arahan (*unsupervised*), yang dimaksud adalah teknik ini diterapkankan tanpa perlunya data *training* dan tanpa ada *teacher* serta tidak memerlukan target *output*

Algoritma K-Means adalah algoritma *clustering* yang popular dan banyak digunakan dalam dunia industri. Algoritma ini disusun atas dasar ide yang sederhana. Pada awalnya ditentukan berapa cluster yang akan dibentuk. Sebarang obyek atau elemen pertama dalam cluster dapat dipilih untuk dijadikan sebagai titik tengah (*centroid point*) cluster. Algoritma K-Means selanjutnya akan melakukan pengulangan langkah-langkah berikut sampai terjadi kestabilan (tidak ada obyek yang dapat dipindahkan)

Seorang mahasiswa dituntut untuk memiliki kemampuan logika yang baik, terlebih lagi bagi seorang mahasiswa dijurusan komputer.

Kemampuan logika seseoran dapat dilatih secara terus menerut untuk itulah salah satu indikator yang dijadikan penentu tingkat kemampuan logika seseorang adalah asal sekolah.

SMK memiliki sistem pendidikan yang lebih ketat dan lebih sulit dibandingkan dengan SMA, namun hal itu tidak menjadi patokan bahwa semua mahasiswa yang berasal dari SMK memiliki logika yang lebih baik dari mahasiswa yang berasal dari SMK. Sebab tu ditambahkan satu variable tambahan sebagai indikator lainnya.

Indikator yang dijadikan variabel tambahan adalah nilai tpa mahasiswa ketika mengikuti ujian masuk perguruan tinggi

2. LANDASAN TEORI

Data Mining

Data mining yang juga dikenal dengan istilah *pattern recognition* erupakan suatu metode yang digunakan untuk pengolahan data guna menemukan pola yang tersembunyi dari data yang diolah. Data yang diolah dengan teknik data mining ini kemudian menghasilkan suatu pengetahuan baru yang bersumber dari data lama, hasil dari pengolahan data tersebut dapat digunakan dalam menentukan keputusan di masa depan.

Data mining juga merupakan metode yang digunakan dalam pengolahan data berskala besar oleh karena itu data mining memiliki peranan yang sangat penting dalam beberapa bidang kehidupan diantaranya yaitu bidang industri, bidang keuangan, cuaca, ilmu dan teknologi. Dalam data mining juga terdapat metode – metode yang dapat digunakan seperti klasifikasi, clustering, regresi, seleksi variabel, dan market basket analisis.

Data mining juga bisa diartikan sebagai rangkaian kegiatan untuk menemukan pola yang

menarik dari data dalam jumlah besar, kemudian data – data tersebut dapat disimpan dalam database, data *warehouse* atau penyimpanan informasi. Ada beberapa ilmu yang mendukung teknik data mining diantaranya adalah data analisis, *signal processing*, *neural network* dan pengenalan pola.

CLUSTERING

Clustering atau pengklasteran adalah suatu teknik data mining yang digunakan untuk menganalisis data untuk memecahkan permasalahan dalam pengelompokkan data atau lebih tepatnya mempartisi dari dataset ke dalam subset. Pada teknik *clustering* targetnya adalah untuk kasus pendistribusian (objek, orang, peristiwa dan lainnya) ke dalam suatu kelompok, hingga derajat tingkat keterhubungan antar anggota *cluster* yang sama adalah kuat dan lemah antara angota *cluster* yang berbeda.

Teknik *cluster* mempunyai dua metode dalam pengelompokkannya yaitu *hierarchical clustering* dan *non-hierarchical clustering*. *hierarchical clustering* merupakan suatu metode pengelompokkan data yang cara kerjanya dengan mengelompokkan dua data atau lebih yang mempunyai kesamaan atau kemiripan, kemudian proses dilanjutkan ke objek lain yang mimiliki kedekatan dua, proses ini terus berlangsung hingga *cluster* membentuk semacam *tree* dimana ada hirarki atau tingkatan yang jelas antar objek dari yang paling mirip hingga yang paling tidak mirip. Namun secara logika semua objek pada akhirnya hanya akan membentuk sebuah *cluster*.

K-Means

K-Means merupakan suatu algoritma yang digunakan dalam pengelompokkan secara pertisi yang memisahkan data ke dalam kelompok yang berbeda – berda. Algoritma ini mampu meminimalkan jarak antara data ke *cluster*nya. Pada dasarnya penggunaan algoritma ini dalam proses *clustering* tergantung pada data yang didapatkan dan konklusi yang ingin dicapai di akhir proses.
Sehingga dalam penggunaan algoritma k-means terdapat aturan sebagai berikut :

a) Berapa jumlah *cluster*yang perlu dimasukkan
b) Hanya memiliki atribut bertipe numeric

Pada dasarnya algoritma k-means hanya mengambil sebagian dari banyaknya komponen yang didapatkan untuk kemudian dijadikan pusat *cluster*awal, pada penentuan pusat *cluster*ini dipilih secara acak dari populasi data. Kemudian algoritma k-means akan menguji masing – masing dari setiap komponen dalam populasi data tersebut dan menandai komponen tersebut ke dalam salah satu pusat *cluster* yang telah didefinisikan sebelumnya tergantung dari jarak minimum antar komponen dengan tiap – tiap pusat *cluster*. Selanjutnya posisi pusat *cluster*akan dihitung kembeli samapi semua komponen data digolongkan ke dalam tiap – tiap *cluster*dan terakhir akan terbentuk *cluster*baru.

3. METODE PENELITIAN

Konsep kesamaan adalah hal yang fundamental dalam analisis *cluster*. Kesamaan antar objek merupakan ukuran korespondensi antar objek. Ada tiga metode yang dapat diterapkan, yaitu ukuran korelasi, ukuran jarak, dan ukuran asosiasi. Dengan menggunakan ukuran jarak, ukuran kemiripan yang dapat digunakan adalah jarak *dEeculidean* dan *dManhattan* City. Jika objek pertama yang diamati adalah X=[X1,X2..Xp] dan Y=[Y1,Y2…Yp] antara 2 objek dari p dimensi maka

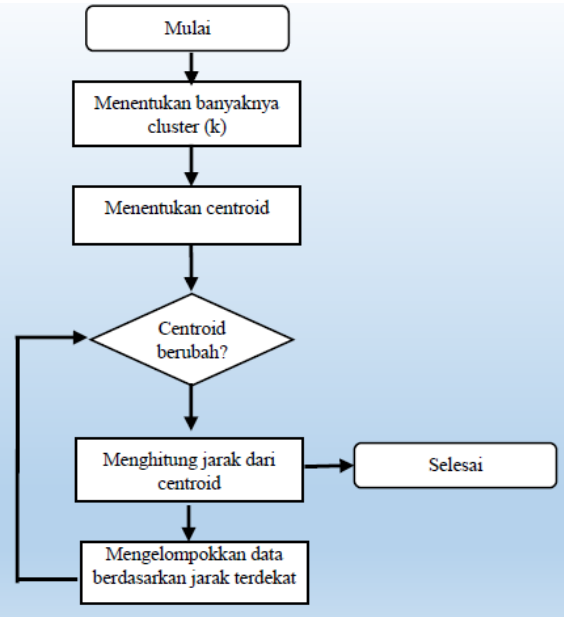
$$\text{dEculidean: } X, Y = \sqrt{\sum_i (X - Y)^2}$$

$$\text{dManhattan: } X, Y = \sqrt{\sum_i |X - Y|}$$

Adapun pun langkah-langkahnya dengan menggunakan algoritma K-Means
sebagai berikut

1. Tentukan jumlah cluster
2. Menentukan *centroid* (koordinat titik tengah setiap *cluster*), untuk iterasi pertama diambil secara random
3. Menghitung jarak obyek ke *centroid* dengan menggunakan rumus Jarak *Euclidean* atau *Manhattan*.
4. Menentukan jarak setiap obyek terhadap koordinat titik tengah,
5. mengelompokkan obyek-obyek tersebut berdasarkan pada jarak terdekat

Berikut ditampilkan diagram alir dari algoritma K-Means.



4. HASIL DAN PEMBAHASAN

Data penelitian yang sedang dilakukan merupakan data nilai mahasiswa yang akan dikelompokkan kedalam kemempuan logika "baik" dan "kurang. Pengelompokan tersebut berdasarkan atribut asal sekolah, nilai TPA dan nilai matakuliah Logika dan Algoritma, yang kemudian akan ditentukan nilai *k*=2.

Tabel Data Sampel

asal sekolah	nilai tpa	Nilai logika & algoritma
1	36	2
2	48	1
2	40	1
2	38	1
1	48	2
1	40	2
2	38	1
2	32	1
1	20	5
2	28	2
1	36	1

asal sekolah	nilai tpa	Nilai logika & algoritma
1	56	1
1	36	1
1	30	5
1	46	1
1	40	3
2	38	2
1	40	1
2	30	3
2	48	1
1	36	3
2	30	2
2	38	1
2	28	5
2	52	1
2	22	1
1	22	3

Tabel data sampel di atas telah ditransformasikan terlebih dahulu dinama asal sekolah dan nilai logika & algoritma di ubah kedalam bentuk angka

Asal Sek0lah	Transformasi
SMA	1
SMK	2

Nilai Logika & Algoritma	Transformasi
A	1
B	2
C	3
D	4
E	5

Iterasi ke-1

1. Penentuan Pusat Awal *Cluster*

Asal Sekolah	Nilai TPA	Nilai Logika & Algoritma

1	36	2
2	52	1

2. Perhitungan Jarak Pusat *Cluster*

Untuk menghitung jarak antara data dengan pusat awal cluster menggunakan persaan *Euclidean distance* sebagai berikut:

$$d\ (i,j) = \sqrt{\sum_{j}^{m} (Cij - Cjk)^2}$$

Dimana:
Cik : Pusat *Cluster*
Ckj : Data
Maka akan dapat didapat nilai matriks jarak sebagai berikut

Jarak data ke-1 pusat *cluster*

$d(x1, c1) = \sqrt{(1-1)^2 + (36-36)^2 + (2-2)^2}$
= 0

$d(x1, c2) = \sqrt{(1-2)^2 + (36-52)^2 + (2-1)^2}$
= 16,06

Jarak data ke-2 pusat *cluster*

$d(x2, c1) = \sqrt{(2-1)^2 + (48-36)^2 + (1-2)^2}$
= 12,08

$d(x2, c2) = \sqrt{(2-1)^2 + (48-36)^2 + (1-2)^2}$
= 4

Jarak data ke-3 ke pusat *cluster*

$d(x3, c1) = \sqrt{(2-1)^2 + (40-36)^2 + (1-2)^2}$
= 4,24

$d(x3, c2) = \sqrt{(2-1)^2 + (40-36)^2 + (1-2)^2}$
= 12

Dan seterusnya diajutkan menghitung untuk data ke-5… *N* terhadap pusat awal *cluster* hingga didapat matrik jarak.

3. Pengelompokan data

Jarak hasil perhitungan pada poin ke-2 akan dilakukan perbandingan dan dipilih jarak yang paling dekat antara data dengan pusat *cluster*, jarak ini akan menunjukkan bahwa data yang memiliki jarak terdekat berada dalam satu kelompok dengan pusat cluster terdekat, pengelompokkan data tersebut dapat dilihat pada tabel berikut

DC1	DC2	C1	C2
0	16.06	1	0
12.08	4	0	1
4.243	12	1	0
2.449	14	1	0
12	4.243	0	1
4	12.08	1	0
2.449	14	1	0
4.243	20	1	0
16.28	32.26	0	0
8.062	24.02	1	0
1	16.03	1	0
20.02	4.123	0	1
1	16.03	1	0
6.708	22.38	1	0
10.05	6.083	0	1
4.123	12.21	1	0
2.236	14.04	1	0
4.123	12.04	1	0
6.164	22.09	1	0
12.08	4	0	1
1	16.16	1	0
6.083	22.02	1	0
2.449	14	1	0
8.602	24.33	1	0
16.06	0	0	1
14.07	30	1	0
14.04	30.08	1	0

4. Penentuan pusat *cluster* baru

Setelah didapt anggota dari setiap cluster kemudian pusat *cluster* baru dihitung berdasarkan data anggot tiap-tiap *cluster* yang sudah didapatkan mengunakan rumus yang sesuai dengan pusat anggota cluster sebagai berikut:

$$c1 = \left(\begin{array}{c} \frac{1+2+2+1+2+2+2+1+1+1+1+2+1+2+1+2+2+2+2+1}{20} : \\ \frac{36+40+38+40+38+32+28+36+36+30+40+38+40+30+36+30+38+28+22+22}{20} : \\ \frac{2+1+1+2+1+1+2+1+1+5+3+2+1+3+3+2+1+5+1+3}{20} \end{array} \right)$$

C1= (1,6:34,9:2,3)

$$c2 = \left(\begin{array}{c} \frac{2+1+1+1+2+2}{6} : \\ \frac{48+48+56+46+48+52}{6} : \\ \frac{1+2+1+1+1+1}{6} \end{array} \right)$$

C2= (1,5:49,66:1,16)

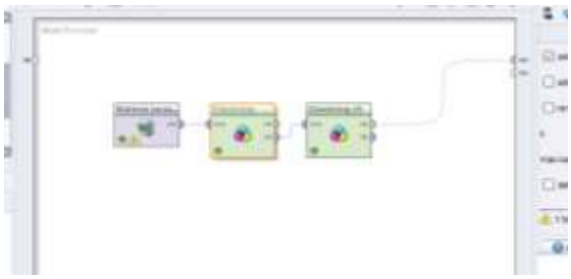
Dari perhitungan di atas maka didapatkan pusat *cluster* baru dalam matrik tabel sebagai berikut

C1	1,6	34,9	2,3
C2	1,5	49,66	1,16

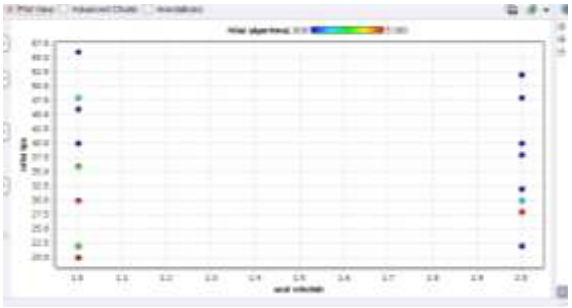
Iterasi selanjutnya dilakukan dengan cara yang sama hingga tidak ada perubahan dalam suatu *cluster*

5. IMPLEMENTASI RAPID MINER

Proses algoritma K-Means menggunakan *Rapid miner* dapat dilihat berikut ini:



Dengan menggunakan pemodelan k-means *clustering* seperti gambar diatas, dengan inisialisasi jumlah *cluster* sebanyak 2 buah, maka didapatkan hasil *cluster* yang terbentuk adalah 2, sesuai denga pendefinisian nilai k dengan jumlah cluster_0 ada 13 item, cluster_1 ada 14 item dengan total jumlah data 27.



6. KESIMPULAN

Dari hasil *clustering* yang ditampilkan dengan diagram scatter dapat terlihat jelas pengelompokannya bahwa mahasiswa yang berasal dari SMK jauh lebih unggul dibandingkan dengan mahasiswa yang berasal dari SMK, bahkan nilai TPA mahasiswa yang berasal dari SMK tidak terlalu berpengaruh terhadap kemampuan logikanya.

Sedangkan mahasiswa yang berasal dari SMK memiliki logika yang baik apabila memiliki nilai TPA yang cukup.

Jadi dapat disimpulkan bahwa mahasiswa dari SMK denangan nilai TPA beragam lebih unggul disbanding mahasiswa dari SMK untuk kemampuan logikanya.

DAFTAR PUSTAKA

Andayani, Sri. "*Pembentukan Cluster Dalam Knowledge Discovery In Database Dengan Algoritma K-Means*". SEMNAS Matematika dan Ped. Matematika. 2007.

Asroni dan Ronald Adrian. "*Penerapan Metode K-Means Untuk Clustering Mahasiswa Berdasarkan Nilai Akademik Dengan Menggunakan WEKA Interface Studi Kasus pada Jurusan teknik Informatika UMM Magelang*". Jurnal Ilmiah Semesta tekni, Vol.18, No.1. 2015.

Metisen, Benri Melpa dan Herlina Latipa Sari. "*Analisis Clustering Menggunakan Metode K-Means dalam pengelompokan Penjualan Produk Pada Swalayan Fadhila*". Jurnal Media Informatika, Vol.11, no.2. 2015.

Ong, Johan Oscar. "*Implementasi Agoritma K-Means Clustering Untuk Menentukan Strategi Marketing Presidend University*". Jurnal Ilmiah teknik Industri, vol.12, no.1. 2013.

Rismawan, Tedy dan Sri Kusuma Dewi. "*Aplikasi K-Means Untuk Pengelompokan Mahasiswa Berdasarkan Nilai Body Masa Indek (BMI)*

dan Ukuran Kerangka". Seminar Nasional Aplikasi Teknologi Informasi 2008 (SNATI 2008). 2008.

Wardhani, Anindya Khrisna. "*Implementasi Algoritma K-Means Untuk Pengelompokan Pasien Pada Puskesmas Kajen Pekalongan*". Jurnal Transformatika, vol.14, no.1. 2016.