**LAPORAN DATA MINING METODE KLASTERING**

**MENGGUNAKAN ALGORITMA K-MEANS**

****

**Disusun oleh:**

**1. Dwi Agung Sugiharto 14.01.53.0073**

**2. Setya Aji Prabawa 14.01.53.0086**

**3. Bagas Brawijaya 14.01.53.0087**

**FAKULTAS TEKNOLOGI INFORMASI**

**UNIVERSITAS STIKUBANK (UNISBANK)**

**SEMARANG**

**2018**

**BAB I**

**PENDAHULUAN**

**1.1 Deskripsi Permasalahan**

Teknologi informasi dewasa ini semakin berkembang pesat disegala bidang. Berbagai macam data dihasilkan oleh teknologi informasi yang canggih di segala bidang, mulai dari bidang pendidikan, industri, ekonomi, ilmu dan teknologi serta berbagai bidang kehidupan lainnya. Penerapan teknologi informasi dalam dunia pendidikan juga dapat menghasilkan data yang berlimpah mengenai siswa dan proses pembelajaran yang dihasilkan.

Contoh data pada bidang pendidikan adalah data sekolah di setiap daerah. Berapa jumlah murid, guru, dan sekolah dalam daerah. Dalam keperluan tertentu diperlukan pengelompokan data yang memiliki karakterstik yang sama. Misalnya menentukan daerah mana saja yang memiliki jumlah siswa yang termasuk dalam kategori banyak dengan guru yang sedikit, daerah yang memiliki siswa yang sedikit bila dibandingkan dengan sekolahnya, dan lain-lain.

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

**1.2 Rumusan Masalah**

Pengelompokan data dalam dunia pendidikan secara manual tentu saja memakan waktu cukup lama. Keakuratan data bila dihitung manual memang lebih tepat, namun bila jumlah data yang akan dikelola dalam jumlah banyak, diperlukan sebuah system *data mining* untuk menunjang proses tersebut.

**1.3 Batasan Masalah**

Agar pembahasan tidak menyimpang dari ruang lingkup pembahasan, maka perlu dibuat suatu batasan masalah sebagai berikut :

1. Data yang dianalisis adalah data pendidikan yang didapat dari data provinsi Jawa Tengah
2. Algoritma yang digunakan adalah *K-Means Clustering.*
3. Pengelompokan data menjadi empat *Cluster.*

**1.4 Tujuan Pembuatan Makalah**

Tujuan pembuatan makalah ini adalah :

1. Membuat sebuah model untuk mengelompokan data murid yang memiliki karakteristik jumlah yang sama
2. Menerapkan *K-Means Clustering* untuk memudahkan dalam menentukan data yang memiliki karakteristik sama.

**BAB II**

**TINJAUAN PUSTAKA**

**2.1 Kajian Deduktif**

Data *Clustering* merupakan salah satu metode *Data Mining* yang bersifat tanpa arahan (unsupervised). Ada dua jenis data clustering yang sering dipergunakan dalam proses pengelompokan data yaitu hierarchical (hirarki) data clustering dan non-hierarchical (non hirarki) data clustering. K-Means merupakan salah satu metode data clustering non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih cluster/kelompok. Metode ini mempartisi data ke dalam cluster/kelompok sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu cluster yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam kelompok yang lain. Adapun tujuan dari data clustering ini adalah untuk meminimalisasikan objective function yang diset dalam proses clustering, yang pada umumnya berusaha meminimalisasikan variasi di dalam suatu cluster dan memaksimalisasikan variasi antar cluster.

A. Algoritma K-means

K-means merupakan salah satu algoritma clustering .Tujuan algoritma ini yaitu untuk membagi data menjadi beberapa kelompok. Algoritma ini menerima masukan berupa data tanpa label kelas. Hal ini berbeda dengan supervised learning yang menerima masukan berupa vektor (­x­1 , y1) , (­x­2 , y2) , …, (­x­i , yi), di mana xi merupakan data dari suatu data pelatihan dan yi merupakan label kelas untuk xi.

Pada algoritma pembelajaran ini, komputer mengelompokkan sendiri data-data yang menjadi masukannya tanpa mengetahui terlebih dulu target kelasnya. Pembelajaran ini termasuk dalam unsupervised learning. Masukan yang diterima adalah data atau objek dan k buah kelompok (cluster) yang diinginkan. Algoritma ini akan mengelompokkan data atau objek ke dalam k buah kelompok tersebut. Pada setiap cluster terdapat titik pusat (centroid) yang merepresentasikan cluster tersebut.

Data clustering menggunakan metode K-Means ini secara umum dilakukan dengan

algoritma dasar sebagai berikut (Yudi Agusta, 2007) :

1. Tentukan jumlah cluster

2. Alokasikan data ke dalam cluster secara random

3. Hitung centroid/ rata-rata dari data yang ada di masing-masing cluster

4. Alokasikan masing-masing data ke centroid/ rata-rata terdekat

5. Kembali ke Step 3, apabila masih ada data yang berpindah cluster atau apabila

perubahan nilai centroid, ada yang di atas nilai threshold yang ditentukan atau

apabila perubahan nilai pada objective function yang digunakan di atas nilai

threshold yang ditentukan.

*Distance Space* Untuk Menghitung Jarak Antara Data dan *Centroid:*

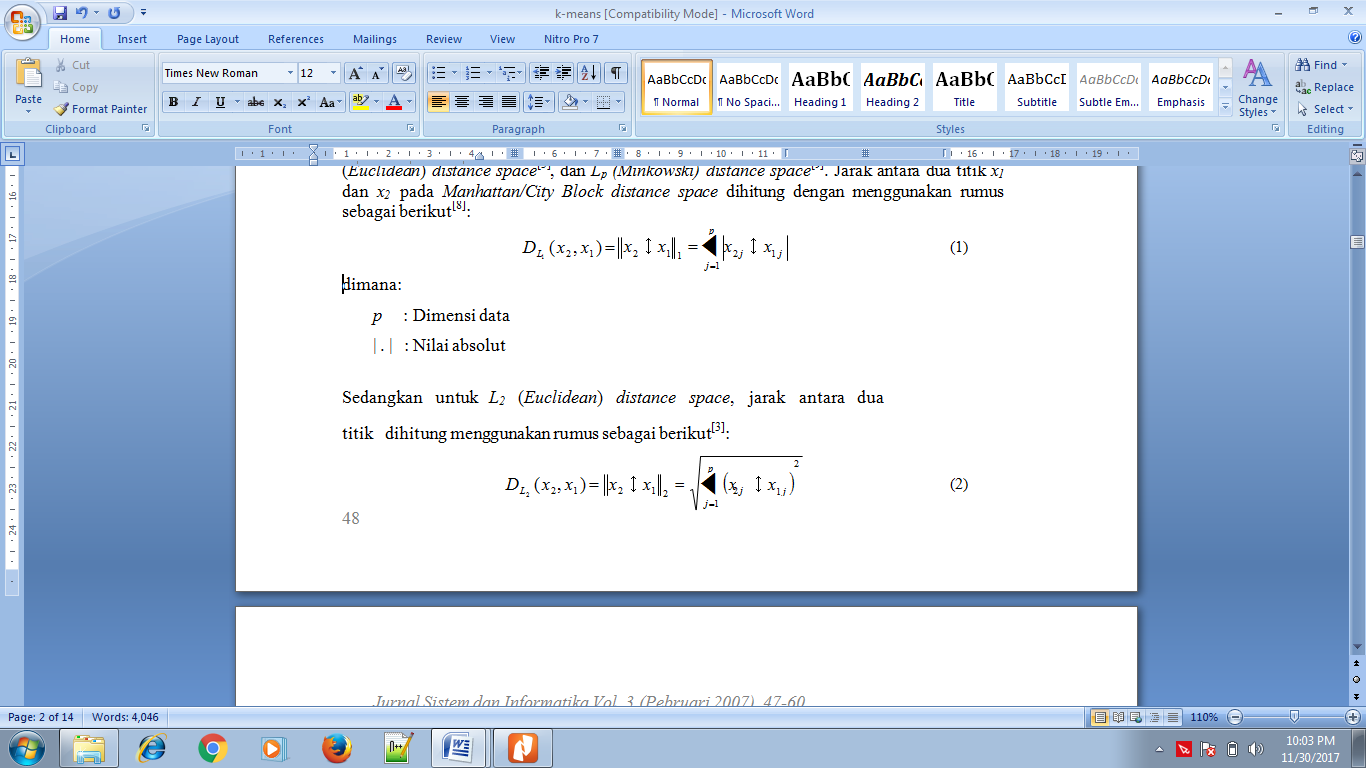
Beberapa distance space telah diimplementasikan dalam menghitung jarak (distance) antara data dan centroid termasuk di antaranya

L1 (Manhattan/City Block) distance space,

L2 (Euclidean) distance space, dan

Lp (Minkowski) distance space.

Jarak antara dua titik x1 dan x2 pada Manhattan/City Block distance space dihitung dengan menggunakan rumus sebagai berikut:

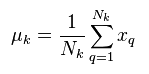


dimana:

p : Dimensi data

Lp (Minkowski) distance space yang merupakan generalisasi dari beberapa distance space yang ada seperti L1 (Manhattan/City Block) dan L2 (Euclidean), juga telah diimplementasikan[9]. Tetapi secara umum distance space yang sering digunakan adalah Manhattan dan Euclidean. Euclidean sering digunakan karena penghitungan jarak dalam distance space ini merupakan jarak terpendek yang bisa didapatkan antara dua titik yang diperhitungkan, sedangkan Manhattan sering digunakan karena kemampuannya dalam mendeteksi keadaan khusus seperti keberadaaan outliers dengan lebih baik.

Pembaharuan suatu titik centroid dapat dilakukan dengan rumus berikut:



Di mana:

µk = titik centroid dari cluster ke-K

Nk = banyaknya data pada cluster ke-K

xq = data ke-q pada cluster ke-K

* **Kekurangan dan Kelebihan K-Means**

Ada beberapa kelebihan pada algoritma k-means, yaitu:

1. Mudah untuk diimplementasikan dan dijalankan.
2. Waktu yang dibutuhkan untuk menjalankan pembelajaran ini relatif cepat.
3. Mudah untuk diadaptasi.
4. Umum digunakan.

Algoritma k-means memiliki beberapa kelebihan, namun ada kekurangannya juga. Kekurangan dari algoritma tersebut yaitu :

1. Sebelum algoritma dijalankan, k buah titik diinisialisasi secara random sehingga pengelompokkan data yang dihasilkan dapat berbeda-beda. Jika nilai random untuk inisialisasi kurang baik, maka pengelompokkan yang dihasilkan pun menjadi kurang optimal.
2. Dapat terjebak dalam masalah yang disebut curse of dimensionality. Hal ini dapat terjadi jika data pelatihan memiliki dimensi yang sangat tinggi (Contoh jika data pelatihan terdiri dari 2 atribut maka dimensinya adalah 2 dimensi. Namun jika ada 20 atribut, maka akan ada 20 dimensi). Salah satu cara kerja algoritma ini adalah mencari jarak terdekat antara k buah titik dengan titik lainnya. Jika mencari jarak antar titik pada 2 dimensi, masih mudah dilakukan. Namun bagaimana mencari jarak antar titik jika terdapat 20 dimensi. Hal ini akan menjadi sulit.
3. Jika hanya terdapat beberapa titik sampel data, maka cukup mudah untuk menghitung dan mencari titik terdekat dengan k titik yang diinisialisasi secara random. Namun jika terdapat banyak sekali titik data (misalnya satu milyar buah data), maka perhitungan dan pencarian titik terdekat akan membutuhkan waktu yang lama. Proses tersebut dapat dipercepat, namun dibutuhkan struktur data yang lebih rumit seperti kD-Tree atau hashing.

**2.2 Kajian Induktif**

Tinjauan pustaka tersebut adalah hasil penelitian terdahulu tentang informasi hasil penelitian yang telah dilakukan sebelumnya dan menghubungkan dengan masalah yang diteliti.

* Dalam rangka menghadapau persaingan bisnis dan menigkatkan pendapatan mini market swalayan, pihak terkait berhak mengambil keputusan yang tepat dalam menentukan stratregi pemasaran produk yang dijual di swalayan. Ketersediaan data yang akurat, kebutuhan akan informasi sebagai pendukung pengambilan keputusan untuk membuat solusi bisnis dan dukungan infraktuktur di bidang teknologi informas merupakan cikal-bakal dari lahirnya tekonologi data mining.

Dalam penerapan data mining ini, digunakan penerapan *clustering* dengan menggunakan algoritma *K-Means.* Dari data yang diolah dengan sampel data yang diambil di swalayan maka akan menghasilkan data penjualan rendah dan data penjualan tinggi. Sehingga dengan adanya pengelompokkan data ini pihak swalayan dapat mengetahui jenis barang yang laris terjual dan tidak. Sehingga barang yang ada di gudang tidak menumpuk.

**BAB III**

**METODOLOGI PENELITIAN**

**3.1 Metode Analisa Sistem**

**3.1.1 Analisa Sistem Aktual**

Dari hasil pra penelitian yang dilakukan di dinas pendidikan, didapatkan sistem yang sudah berjalan dan digunakan saat ini masih manual. Disamping itu, dinas terkait mempunyai kesulitan dalam pengelompokan data sekolah yang memiliki karakteristik sama untuk keperluan program kerja selanjutnya.

**3.1.2 Analisa Sistem Baru**

Dari kekurangan sistem yang sedang berjalan tersebut maka penelitian ini diterapkanlah menggunakan algortitma *K-Means* dan dengan pemrosesan software data mining yaitu Bahasa R.

**BAB IV**

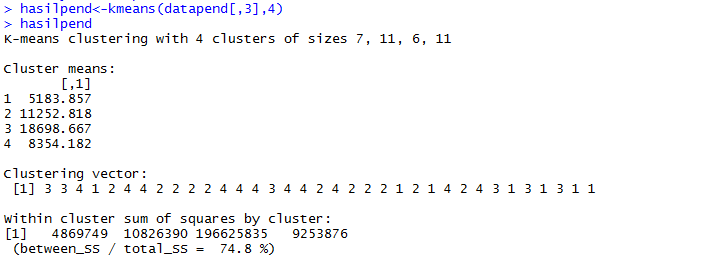
**HASIL DAN PEMBAHASAN**

**4.1 Hasil Uji Program**

Hasil dari pengujian klastering yang telah diuji di program bahasa R pada data pendidikan dibawah.

|  |  |  |  |
| --- | --- | --- | --- |
| **Kabupaten\_Kota** | **Sekolah** | **Murid** | **Guru** |
| Cilacap | 41 | 16 440 | 1 300 |
| Banyumas | 37 | 16 264 | 1 237 |
| Purbalingga | 16 | 7 173 | 590 |
| Banjarnegara | 13 | 6 456 | 492 |
| Kebumen | 25 | 10 025 | 791 |
| Purworejo | 23 | 8 112 | 778 |
| Wonosobo | 17 | 6 961 | 588 |
| Magelang | 35 | 10 826 | 1 042 |
| Boyolali | 32 | 10 338 | 989 |
| Klaten | 28 | 10 692 | 1 188 |
| Sukoharjo | 27 | 11 371 | 1 027 |
| Wonogiri | 21 | 8 911 | 805 |
| Karanganyar | 14 | 9 432 | 568 |
| Sragen | 25 | 9 053 | 859 |
| Grobogan | 36 | 15 348 | 668 |
| Blora | 22 | 8 019 | 695 |
| Rembang | 14 | 7 287 | 496 |
| Pati | 26 | 12 430 | 1 060 |
| Kudus | 17 | 9 592 | 697 |
| Jepara | 22 | 9 888 | 650 |
| Demak | 34 | 12 775 | 791 |
| Semarang | 25 | 10 841 | 750 |
| Temanggung | 15 | 5 602 | 457 |
| Kendal | 32 | 12 134 | 875 |
| Batang | 14 | 5 786 | 454 |
| Pekalongan | 17 | 8 005 | 585 |
| Pemalang | 23 | 12 461 | 869 |
| Tegal | 25 | 9 351 | 585 |
| Brebes | 32 | 16 013 | 1 084 |
| Kota Magelang | 12 | 4 914 | 454 |
| Kota Surakarta | 34 | 16 661 | 1 424 |
| Kota Salatiga | 8 | 4 154 | 336 |
| Kota Semarang | 73 | 31 466 | 7 447 |
| Kota Pekalongan | 9 | 3 950 | 284 |
| Kota Tegal | 12 | 5 425 | 378 |

Hasil pengujian adalah sebagai berikut :



Gambar 4.1 Hasil Uji Data Pendidikan

Diatas adalah hasil proses klastering dari data pendidikan, terlampir juga titik pusat, ukuran data tiap klaster dan keanggotaan klaster pada masing-masing data.

**BAB V**

**KESIMPULAN DAN saran**

**5.1 Kesimpulan**

Konsep data mining mengunakan K-Means guna mengelompokkan data pendidikan untuk mengetahui data yang memiliki kesamaan sifat (dalam hal ini adalah murid) dengan melakukan proses pengelompokan data menjadi beberapa bagian / klaster.

Hasil ini dapat digunakan untuk memberikan saran pertimbangan dalam menentukan stratregi pendidikan selanjutnya mengenai pembagian buku, menentukan anggaran daerah untuk pendidikan, dan sebagainya

**5.2 Saran**

Konsep data mining *clustering K-Means* dapat dilakukan pengembangan lebih lanjut dengan mengcluster produk berdasarkan varian dari tiap-tiap modelnya dan cluster pelanggan berdasarkan wilayahnya.

**DAFTAR PUSTAKA**

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996).From Data Mining To Knowledge Discovery In Databases.AI Magazine, 17(3), 37.

Freitas, A. A. (2013). Data Mining And Knowledge Discovery With Evolutionary Algorithms.Springer Science & Business Media.

Liautaud, B., & Hammond, M. (2000).e-Business intelligence: turning information into knowledge into profit. McGraw-Hill, Inc..

Ong, J. O. (2013). Implementasi Algoritma K-Means Clustering untuk Menentukan Strategi Marketing President University.

Ramadhani, Rima Dias (2016), Data Mining Menggunakan Algoritma K-Means Clustering Untuk Menentukan Strategi Promosi. Universitas Dian Nuswantoro.