ANALISIS DATA WORLD DEVELOPMENT INDICATORS MENGGUNAKAN CLUSTER DATA MINING

Sigit Kamseno[1)], **Barka Satya** [2)]

[1), 2)] *Teknik Informatika STMIK AMIKOM Yogyakarta*
[3)]*Sistem Informasi STMIK AMIKOM Yogyakarta*
Jl Ring road Utara, Condongcatur, Sleman, Yogyakarta 55281
Email : sigit.k@students.amikom.com[1)], *barka.satya@amikom.ac.id*[2)]

Abstrak

World Development Indicators merupakan sebuah database yang berisi tentang track records indikator - indikator yang mempengaruhi perkembangan suatu negara. Ada berbagai macam data yang tercatat dalam database tersebut, seperti nama negara, kode negara, sistem perdagangan yang digunakan, kategori pendapatan, survei - survei, dan masih banyak lagi. Data tersebut dihimpun oleh World Bank sebagai salah satu organisasi internasional yang berperan untuk membantu negara berkembang menjadi negara maju, khususnya dalam mengembangkan ekonomi.

Dalam penelitian ini akan dibuat klaster yang akan membagi negara - negara di dunia ke dalam sebuah klaster (kelompok) menggunakan algoritma DBSCAN. Klaster tersebut akan mewakilkan suatu negara termasuk dalam klaster negara maju, negara berkembang, atau negara yang tertinggal soal perkembangannya. Klasterisasi akan dilakukan dengan melakukan perbandingan indikator - indikator dimasing - masing negara yang kemudian indikator tersebut di transformasi menjadi nilai - nilai kuantitatif. Perbandingan nilai kuantitatif ditentukan dengan nilai eps (kedekatan), seberapa dekat nilai kuantitatif suatu negara dengan nilai kuantitatif negara lain. Setelah pebandingan nilai eps ditentukan pula minPts (minimum points) yang akan menentukan kedekatan - kedekatan points layak menjadi sebuah klaster. Klaster yang terbentuk akan mempengaruhi penamaan klaster, ditentukan berdasarkan rata - rata nilai kuantitatif negara di klaster tersebut lebih kecil dari klaster lain atau lebih besar dari klaster disekitarnya.

Evaluasi klaster dilakukan setelah klaster terbentuk dengan metode Sillhouette Index. Metode ini dilakukan dengan mencari rata - rata jarak atau kemiripan klaster.

Kata kunci:*Data mining, clustering, dbscan, sillhouette coefficient, visualisasi cluster, algoritma cluster*

1. Pendahuluan

Dataset world development indicators berisi *record* negara didunia, record tersebut berisi *system of trade*, *income group*, *currency*, dan lain - lain. *Dataset* tersebut akan diolah dengan klaster *data mining* menggunakan algoritma DBSCAN. Algoritma DBSCAN akan mengolah *record* negara di dalam *dataset* tersebut yang akan dikelompokkan berdasarkan nilai kedekatan atau kemiripan. Nilai kepadatan ditentukan dari masukkan nilai *minimum points* dan *epsilon* (jarak), seberapa padat dan mirip data tersebut akan dibuat menjadi klaster.

Clustering adalah suatu teknik dalam data mining yang digunakan untuk memasukkan data ke dalam grup yang bersesuaian tanpa pengetahuan yang mendalam tentang grup tersebut (Santosa, 2007). *Clustering* ini bertujuan untuk meminimalisasikan *objective function* yang diset dalam proses *clustering*, yang pada umumnya berusaha meminimalisasikan variasi antar klaster (Agusta, 2007). Sampai saat ini, para ilmuwan data (*data scienties*) terus melakukan berbagai usaha untuk melakukan perbaikan model klaster dan menghitung jumlah klaster yang optimal sehingga dapat di hasilkan klaster yang baik (Alfina, et al. 2012).

DBSCAN adalah algoritma *clustering* yang melihat bahwa sebuah klaster merupakan daerah yang padat obyek dan terpisah dari daerah yang memiliki tingkat kepadatan yang rendah (*noise*). DBSCAN dapat membentuk daerah dengan bentuk yang tidak beraturan di dalam ruang data. Bagian penting algoritma ini adalah kepadatan obyek dan hubungan antar obyek yang dibentuk dari obyek yang berdekatan [1]. Konsep kepadatan yang dimaksud dalam DBSCAN adalah jumlah data yang berada dalam radius *Eps* (ε) dari setiap data. Jika jumlah data dalam radius ε lebih dari atau sama dengan *MinPts* (jumlah minimal data dalam radius ε), data tersebut masuk dalam kategori kepadatan yang diinginkan. Konsep kepadatan seperti ini melahirkan tiga macam status dari setiap data, yaitu inti (*core*), batas (*border*), dan noise (*noise*) [2]. DBSCAN menyatakan bahwa sebuah kluster dapat dibentuk jika untuk setiap titik data, pada dalam radius tertentu *Eps* dari titik data tersebut terdapat minimal minPts titik obyek. [1].

Sillhouette Coefficient digunakan untuk memvalidasi baik sebuah data, klaster tunggal, atau bahkan keseluruhan. Metode validasi klaster yang menggabungkan nilai kohesi dan separasi. Menghitung nilai *sillhouette coefficient* sebuah data ada dua komponen a_i dan b_i. a_i adalah rata - rata jarak data ke-i terhadap semua data lainnya dalam satu klaster, sedangkan b_i didapatkan dengan menghitung rata - rata jarak data ke-i terhadap semua data dari klaster yang lain tidak dalam satu klaster dengan data ke-i, kemudian

diambil yang terkecil ([Tan *et al*, 2006], [Petrovic, 2003]) [3].

Berikut formula menghitung a_i^j :

$$a\,\frac{j}{i} = \frac{1}{m_j - 1}\sum_{r=1}^{m_j} d\{x_i^j, x_r^j\} \quad (1)$$

$d = (x_i^j, x_r^j)$ adalah perhitungan (*euclidean* kuadrat) jarak data ke-i dengan data ke-r dalam satu klaster j, sedangkan m_j adalah jumlah data dalam klaster ke-j.

Berikut formula untuk menghitung b_i^j :

$$b_i^j = \min_{\substack{n=1,\ldots,k \\ n!=j}} \{\frac{1}{m_n}\sum_{\substack{r=1 \\ r!=1}}^{m_n} d(x_i^j, x_r^n)\} \quad (2)$$

Untuk mendapatkan *sillhouette coefficient* data ke-i manggunakan persamaan sebagai berikut :

$$SI_i^j = \frac{b_i^j - a_i^j}{\max(a_i^j, b_i^j)} \quad (3)$$

Nilai *sillhouette coefficient* yang didapat dalam rentang [-1,+1]. Nilai *sillhouette coefficient* yang mendekati 1 menandakan bahwa data tersebut berada dalam klaster yang tepat.

Persamaan nilai *sillhouette coefficient* klaster :

$$SI_j = \frac{1}{m_j}\sum_{i=1}^{m_j} SI_i^j \quad (4)$$

Persamaan nilai *sillhouette coefficient* global :

$$SI = \frac{1}{k}\sum_{j=1}^{k} SI_j \quad (5)$$

k adalah jumlah klaster.[4]

Berikut ini adalah representasi dari nilai *sillhouette coefficient* [5] *:*

Tabel 2.*Representasi Kauffman dan Rousseeuw (1990)*

SC	Representasi
0.71 - 1.00	Baik
0.51 - 0.70	Sedang
0.26 - 0.50	Buruk
≤ 0.25	Berada di klaster lain

2. Pembahasan

2.1. Pengumpulan Data

Dataset diperoleh dari kaggle *dataset*, salah satu platform terbesar yang memiliki ratusan *dataset* apapun. Sebelum dilakukan *clustering* analisis *dataset* terlebih dahulu untuk menentukan attribut yang mempunyai relasi terkuat. *Dataset* harus melalui tahap *data selection, data cleaning, data transformation* (inisialisasi), *cluster* dan evaluasi.

Data selection dilakukan untuk menghapus attribut yang tidak memiliki relasi kuat dengan attribut yang lain. Setelah analisis didapat 3 attribut dari tabel *country* yaitu *country, income group,* dan *latest trade data.*

Tabel 3.*Data Selection*

Attribut	Keterangan
Country	Nama negara
Income Group	Kategori pendapatan
Lates Trade Data	Tahun survei pendapatan

Data cleaning menghapus atau memperbaiki attribut yang miliki masukkan yang tidak relevan, kosong, atau tidak sesuai dengan data yang lain.

Data transformation untuk meng-inisialisasi data yang sudah melalui tahap *selection* dan *cleaning.*

Tabel 4. *Inisialisasi Country*

Country	**Inisialisasi**
Afganistan	1
Albania	2
Algeria	3
dst...	dst...
Zimbabwe	247

Tabel 5. *Inisialisasi Income Group*

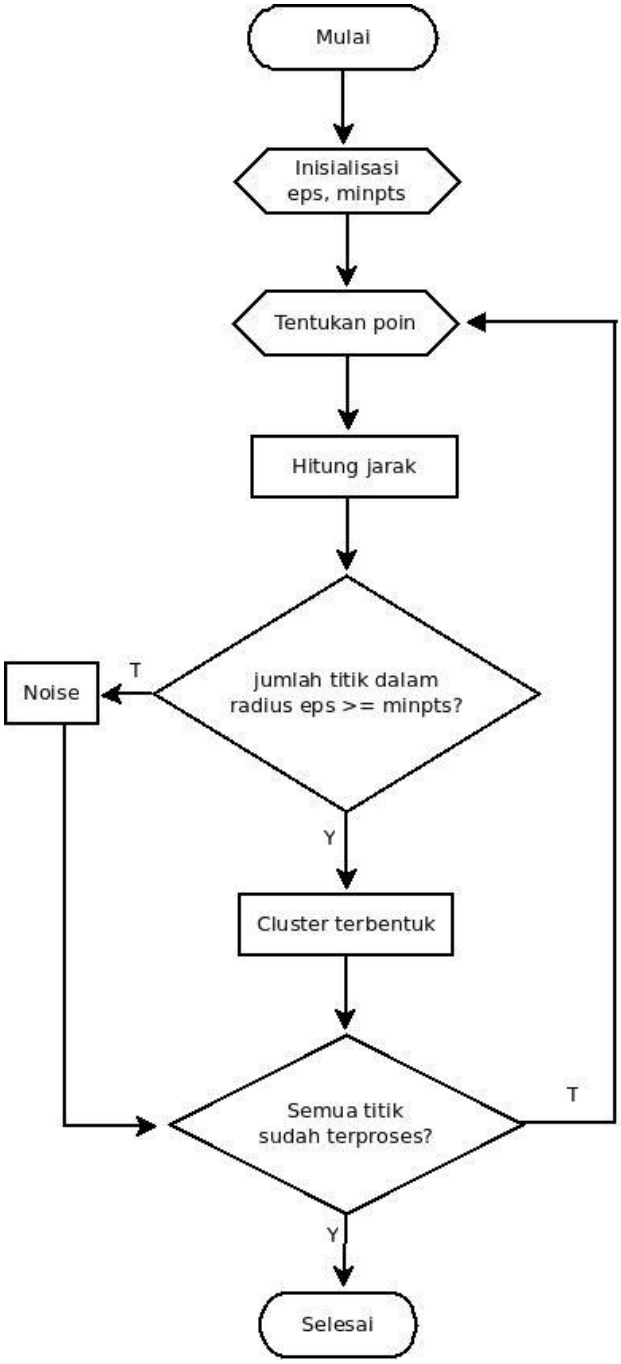
Income Group	**Frekuensi**
High income : OECD	25
High income : non OECD	34
Upper middle income	31
Lower middle income	32
Low income	15
none	110

Tabel 6.*Inisialisasi Latest Trade Data*

Latest Trade Data	**Frekuensi**
2012 - 2015	97
2008 - 2011	13
2004 - 2007	3
2000 - 2003	2
none	132

2.2. Alur Pengujian

Langkah - langkah *clustering* algoritma DBSCAN yaitu memasukkan data yang sudah melalui tahap analisis dan transformasi data, selanjutnya akan dilakukan proses. Berikut alurnya :



Gambar 1.*Alur Pengujian*

2.3. Implementasi Algoritma DBSCAN

Pengujian dilakukan dengan 10 data sampel yang diambil secara acak.

Tabel 7.*Sampel Data*

Income Group	*Latest Trade Data*	*Country*
1	1	1
4	5	2
4	5	3
4	1	4
5	3	5
1	1	6
1	1	7
1	1	8
5	5	9
5	5	10

Selanjutnya nilai - nilai sampel data akan dipadatkan dengan standard deviasi. Berikut hasil pemadatan data :

Tabel 8. *Pemadatan Nilai Data Sampel*

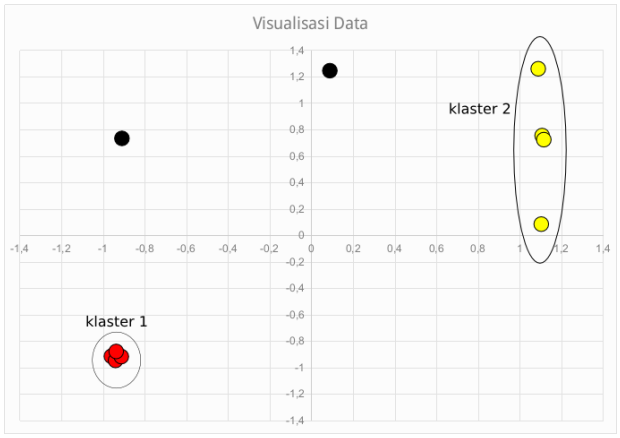
Income Group	*Latest Trade Data*	*Country*
-0.955	-0.914	-1.725
0.679	1.162	-1.711
0.679	1.162	-1.697
0.679	-0.914	-1.683
1.224	0.123	-1.669
-0.955	-0.914	-1.655
-0.955	-0.914	-1.641
-0.955	-0.914	-1.627
1.224	1.162	-1.613
0.135	1.162	-1.599

Data dalam tabel akan dilakukan proses klaster dengan nilai *epsilon* = 0.6 dan *minimum points* = 2.

$N_{0.6}(1)=\{6,7,8\}$; $N_{0.6}(2)=\{3,9,10\}$; $N_{0.6}(3)=\{2,9,10\}$; $N_{0.6}(4)=\{\}$; $N_{0.6}(5)=\{\}$; $N_{0.6}(6)=\{1,7,8\}$; $N_{0.6}(7)=\{1,6,8\}$; $N_{0.6}(8)=\{1,6,7\}$; $N_{0.6}(9)=\{2,3\}$; $N_{0.6}(10)=\{2,3\}$;

Tabel 9.*Sample Data hasil klaster*

Income Group	*Latest Trade Data*	*Country*	*Cluster*
-0.955	-0.914	-1.725	1
0.679	1.162	-1.711	2
0.679	1.162	-1.697	2
0.679	-0.914	-1.683	*noise*
1.224	0.123	-1.669	*noise*
-0.955	-0.914	-1.655	1
-0.955	-0.914	-1.641	1
-0.955	-0.914	-1.627	1
1.224	1.162	-1.613	2
0.135	1.162	-1.599	2



Gambar 2.*Hasil Klaster*

2.4. Evaluasi

Silhouette Coefficient digunakan untuk melihat kualitas dan kekuatan klaster, seberapa baik suatu objek ditempatkan dalam suatu klaster. Metode ini merupakan gabungan dari metode *cohesion* dan *separation*. Tahapan perhitungan s*ilhouette coefficient.*

Perhitungan nilai α untuk data yang berada dalam klaster 1 sebagai berikut :

$$a_1^1 = \frac{1}{m_1 - 1}\sum_{\substack{r=1 \\ r!=j}}^{m1} d(x_1^1, x_r^1) \quad (6)$$

$$a_1^1 = \frac{1}{4-1} = (d(x_1^1, x_2^1) + d(x_1^1, x_3^1) + d(x_1^1, x_4^1) \quad (7)$$

$$a_1^1 = \frac{1}{3}(0+0+0) = 0 \quad (8)$$

Perhitungan nilai *b* untuk data yang berada dalam klaster 1 sebagai berikut :

$$b_1^1 = \min\{\frac{1}{4}(2{,}659 + 2{,}659 + 2{,}659)\} \quad (9)$$

$$b_1^1 = \min(2{,}659) = 2{,}659 \quad (10)$$

Perhitungan nilai *sillhouette coefficient* untuk data yang berada dalam klaster 1 sebagai berikut :

$$SI_1^1 = \frac{b_1^1 - a_1^1}{\max\{a_1^1, b_1^1\}} = \frac{2{,}659 - 0}{\max\{(2{,}659, (0)\}} = 1 \quad (11)$$

Tabel 10.*Nilai SC untuk setiap data dalam klaster 1*

Data ke-i					
Data di Klaster 1	jarak	1	6	7	8
	1	-	0	0	0
	6	0	-	0	0
	7	0	0	-	0
	8	0	0	0	-
a		0	0	0	0
Data di Klaster 2	2	2,642	2,642	2,642	2,642
	3	2,642	2,642	2,642	2,642
	9	3,010	3,010	3,010	3,010
	10	2,345	2,345	2,345	2,345
Rata - rata		2,660	2,660	2,660	2,660
b		2,660	2,660	2,660	2,660
SI		1	1	1	1

Nilai *sillhouette coefficient* untuk keseluruhan klaster 1 :

$$SI_1 = \frac{1}{m_1}\sum_{i=1}^{m_1} SI_i^1 = \frac{1}{m_1}(SI_1^1 + SI_2^1 + SI_3^1 + SI_4^1) \quad (12)$$

$$SI_1 = \frac{1}{4}(1+1+1+1) = 1 \quad (13)$$

Tabel 11.*Nilai SC untuk setiap data dalam klaster 2*

Data ke-i					
Data di Klaster 2	jarak	2	3	9	10
	2	-	0	0,545	0,545
	3	0	-	0,545	0,545
	9	0,545	0,545	-	1,090
	10	0,545	0,545	1,090	-
a		0,363	0,363	0,727	0,727
Data di Klaster 1	1	2,642	2,642	3,010	2,345
	6	2,642	2,642	3,010	2,345
	7	2,642	2,642	3,010	2,345
	8	2,642	2,642	3,010	2,345
Rata - rata		2,642	2,642	3,010	2,345

b	2,642	2,642	3,010	2,345
SI	0,863	0,863	0,759	0,690

$$SI_2 = \frac{1}{4}(0{,}862 + 0{,}862 + 0{,}758 + 0{,}690) = 0{,}793 \quad (14)$$

$$SI = \frac{1}{k}\sum_{j=1}^{k} SI_j = \frac{1}{2}(SI_1 + SI_2) \quad (15)$$

$$SI = \frac{1}{2}(1 + 0{,}793) = 0{,}896 \quad (16)$$

3. Kesimpulan

Berdasarkan hasil uji coba dapat ditarik beberapa kesimpulan sebagai berikut :

a. Penerapan klaster dengan data 10 negara menggunakan nilai eps 0.6 dan minpts 2 menghasilkan 2 klaster. 4 negara pada klaster pertama, 4 negara pada klaster kedua dan terdeteksi 2 negara sebagai noise. Sedangkan untuk data 246 negara terbentuk 4 klaster.
b. Nilai global evaluasi yang dihasilkan dari penerapan klaster dengan data 10 negara yaitu 0,896 termasuk kategori baik. Sedangkan klaster dengan data 246 negara yaitu 0,068 termasuk kategori buruk.
c. Pemberian nama untuk setiap klaster disesuaikan dengan jumlah klaster yang dihasilkan. Klaster dengan rata - rata data dengan nilai besar maka termasuk klaster negara – negara maju. Klaster kedua termasuk klaster negara berkembang dan seterusnya.

Daftar Pustaka

[1] Kantardzic, Mehmed. "Data Mining: Concepts, Models, Methods, and Algorithms (2nd ed.)". Wiley-IEEE Press. 2011.
[2] Pasetyo, Eko. "Data Mining Konsep dan Aplikasi menggunakan MATLAB". Yogyakarta : Andi, 2012.
[3] Hermawati, Fajar Astuti. "Data Mining". Yogyakarta : Andi, 2009.
[4] Prasetyo, Eko. "Data Mining mengolah Data menjadi Informasi menggunakan Matlab". Yogyakarta : Andi, 2014
[5] Susanto, Eko Budi. "Evaluasi hasil Klaster pada Dataset Iris, Soybean-small, Wine menggunakan Algoritma Fuzzy C-means dan K-means++". 2016.

Biodata Penulis

Sigit Kamseno, mahasiswa Jurusan Teknik Informatika STMIK AMIKOM Yogyakarta

Barka Satya, menempuh D3 STMIK AMIKOM Yogyakarta Tahun 2001, S1 STMIK AMIKOM Ygyakarta Tahun 2005 MAGISTER TEKNIK INFORMATIKA STMIK AMIKOM .Saat ini menjadi Dosen di STMIK AMIKOM Yogyakarta.

ISSN : 2302-3805