CLUSTERING DATA INDEKS PEMBANGUNAN MANUSIA DAN PRODUK DOMESTIK BRUTO UNTUK INDENTIFIKASI PEMERATAAN PEMBANGUNAN DI INDONESIA

CLUSTERING HUMAN DEVELOPMENT INDEX AND GROSS DOMESTIC PRODUCT TO IDENTIFY EQUITABLE DEVELOPMENT IN INDONESIA

Dwinanda Septiafani[1], Andry Alamsyah, S.Si, M.Sc[2]

[1,2]Prodi S1 Manajemen Bisnis Telekomunikasi dan Infornatika, Fakultas Ekonomi dan Bisnis, Universitas Telkom

[1]dwinandaa@gmail.com, [2]andrya@telkomuniversity.ac.id

Abstrak

Indeks Pembangunan Manusia (IPM) merupakan ukuran dalam menilai seberapa baik pencapaian sebuah negara. Menurut UNDP, tidak hanya IPM tetapi terdapat pengukuran lain sebagai alternatif penilaian murni kemajuan nasional yaitu Produk Domestik Bruto (PDB). Namun berdasarakan nilai tersebut masih terdapat wilayah yang berada di bawah nilai rata-rata. Hal ini menunjukkan belum meratanya pembangunan di Indonesia. Clustering merupakan salah satu dari teknik data mining yang bertujuan untuk mengelompokkan objek yang memiliki kesamaan/kemiripan ke dalam cluster dan objek yang berbeda ke dalam cluster lain. Algoritma clustering yang digunakan untuk mengolah data IPM dan GDP dalam penelitian ini adalah DBSCAN yang menghasilkan cluster berdasarkan kerapatan data. Berdasarkan hasil clustering dari data IPM dan GDP kabupaten/kota di Indonesia diperoleh sebanyak 2 cluster dan menghasilkan 7 noise. Disimpulkan bahwa berdasarkan hasil clustering, pembangunan di Indonesia masih belum merata dikarenakan terdapat kelompok wilayah lainnya yang memiliki nilai lebih tinggi. Selain itu noise yang dihasilkan merupakan wilayah yang memiliki nilai paling tinggi diantara cluster yang telah terbentuk. Dalam mewujudkan pemerataan pembangunan, pemerintah perlu melakukan pembangunan sesuai dengan manajemen pembangunan yang dimulai dengan perencanaan, pengarahan sumber daya, menggerakkan partisipasi masyarakat, koordinasi, pemantauan dan evaluasi serta pengawasan pelaksana pembangunan. Tidak hanya itu pemerintah kabupaten/kota harus membangun sumber daya manusia yang memiliki kompetensi diantaranya dalam hal perencanaan, pelaksanaan rencana, pengorganisasian, kepemimpinan, menajemen sumber daya manusia dan teknologi, kemampuan membangun kerja sama, memberdayakan peran serta masyarakat/swasta, pengawasan dan pengendalian pembangunan dan sebagainya.

Kata Kunci: ***Data Mining, clustering*, pembangunan, DBSCAN**

Abstract

Human Development Index (HDI) is a measure in assessing how well the achievement of a country. According to UNDP, not only of the HDI but there are other measurement as an alternative to purely national progress assessment i.e. Gross Domestic Product (GDP). But the investigation by those values there are still areas that are below average. It shows the development in Indonesia is still not evenly distributed.

Clustering is one of the data mining technique which aims to classify the objects have in common/similarities into clusters and different objects into another cluster. Clustering algorithms used to process data of the HDI and GDP in this study is that produces clusters of DBSCAN based on density data.

Based on the clustering of the data of the HDI and GDP of kabupaten/kota in Indonesia gained as much as 2 cluster and generates 7 noise. It was concluded that based on the results of clustering, development in Indonesia is still not equitable because there are other area groups that have a higher value. Besides the resulting noise is an area that has the highest value among the cluster has been formed.

In realizing equitable development, Governments need to do development in accordance with the management development that started with the planning, direction, moving the resources public participation, coordination, monitoring and evaluation as well as the supervision of implementing sustainable development. Not only is it the Government's kabupaten/kota must build human resources competencies among them in terms of planning, implementation plans, organizing, leadership, human resources and menajemen technology, the ability to build teamwork, empowering the public/private participation, monitoring and control of development and so on.

Keywords: Data Mining, Clustering, Development, **DBSCAN**

1. Pendahuluan

Grafik yang terdapat pada BPS menunjukan bahwa IPM dan GDP secara garis besar mengalami kenaikan setiap tahunnya. Hal ini artinya Indonesia terus mengalami kemajuan. Namun jika diperhatikan pada pada penyebarannya terbagi ke dalam Provinsi yang tidak seluruhnya memiliki nilai IPM yang tinggi. Masih terdapat beberapa kelompok Provinsi dengan nilai IPM yang masih rendah.Ini artinya masih belum merata tingkat kemajuan di seluruh wilayah Indonesia. Data mining adalah proses menemukan pengetahuan dan pola yang menarik dari data dalam jumlah yang besar. Data mining terdiri dari beberapa metode yaitu yang umum digunakan adalah clustering. Clustering merupakan salah satu dari metode data mining yang digunakan untuk mengelompokkan objek-objek sedemikian rupa sehingga objek dalam satu cluster yang sangat mirip dan objek di berbagai cluster cukup berbeda. Clustering juga dapat melakukan analisis pengelompokan wilayah.Beberapa wilayah yang masih memiliki nilai IPM dan GDP yang rendah perlu mendapat perhatian dengan tujuan kemerataan tingkat kemajuan di seluruh Indonesia. Merata atau tidaknya pembangunan di suatu wilayah dapat dilihat dari nilai pembangunan yang memiliki kesamaan atau kemiripan. Jika pada umumnya memiliki banyak kemiripan atau rata-rata nilainya sama maka pembangunan di wilayah tersebut sudah merata. Tingkat pemerataan suatu wilayah dapat diketahui dengan memanfaatkan clustering karena tujuan dari clustering itu sendiri dapat menghasilkan jumlah kelompok berdasarkan kemiripan.

2. Dasar Teori dan Metodologi

2.1. Dasar Teori

Big Data

Big Data adalah bentuk efektif dan inovatif yang digunakan untuk menggambarkan pertumbuhan eksponensial dari ketersediaan data terstruktur dan tidak terstruktur untuk meningkatkan proses pengambilan keputusan [1]. Ada sejumlah karakteristik utama dalam Big Data atau yang disebut 6 V's yaitu:

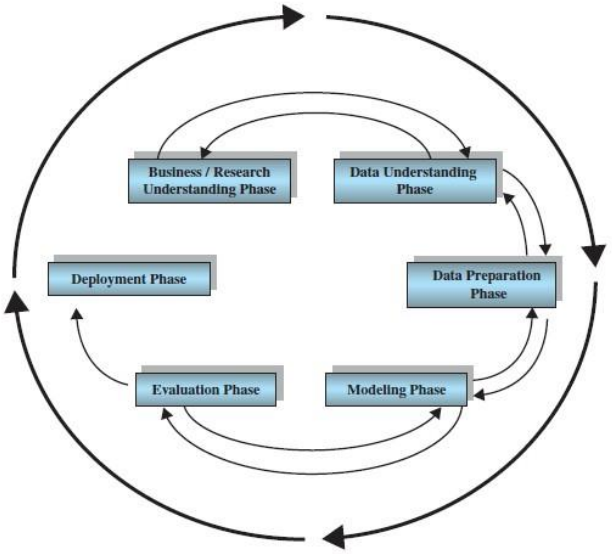
1. Ukuran (*Volume*): Karateristik ini merujuk kepada tingginya volume data yang memerlukan ruang penyimpanan yang besar atau terdiri dari sejumlah catatan besar.
2. Kecepatan (*Velocity*): Ini menunjukan tingkat aliran data pada kecepatan yang belum pernah terjadi sebelumnya (atau peristiwa kompleks pengolahan) sampai respon terhadap data tersebut.
3. Beragam (*Variety*): Inilah ciri menggambarkan field multidimensi data yang dikumpulkan dari berbagai sumber dengan keragaman format.
4. Kebenaran (*Veracity*): Mengacu pada membersihkan bias, noise dan kelainan data yang besar.
5. Kelangsungan hidup (*Viability*): Kombinasi dari semua informasi yang relevan untuk melakukan berbagai ramalan masa depan.
6. Nilai (*Value*): Karakteristik yang menggambarkan tujuan utama mengumpulkan data yang begitu besar yaitu menemukan hubungan yang baik secara eksplisit atau tersembunyi dalam data untuk mengubah nilai yang layak.

Data Mining

Data mining adalah proses menemukan pengetahuan dan pola yang menarik dari data dalam jumlah yang besar. Sumber data dapat mencakup database, gudang data, web, repositori informasi lainnya atau data yang dialirkan ke dalam sistem dinamis[2]. Selanjutnya terdapat beberapa fungsi dan tugas yang dilakukan *data mining* yang biasanya dikerjakan dalam proyek *data mining* [3] yaitu:

1. *Description*, mencari cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data.
2. *Estimation*, memperkirakan nilai dari target variabel yang berbentuk numerik dengan menggunakan variabel prediksi.
3. *Classification*, mirip dengan *estimation*, yang membedakannya adalah target variabelnya berbentuk kategorikal.
4. *Prediction*, hampir sama dengan *classification* dan *estimation*, yang membedakan *prediction* adalah hasilnya terdapat pada masa depan. Contoh: memprediksi harga dari saham 3 bulan kedepan.
5. *Clustering*, yang dimaksud dengan *clustering* adalah pengelompokkan catatan, pengamatan atau kasus kedalam kelas yang sama.
6. *Association*, adalah menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis dikenal dengan *market based analysis.*

***Cross Industry Standard Process for Data Mining* (CRISP-DM)**



Gambar 1. Proses CRISP-DM

CRIPS-DM merupakan standarisasi dari data mining yang disusun oleh tiga penggagas data mining market yaitu Daimler Chrysler (Daimler-Benz), SPSS (ISL) dan NCR. Kemudian dikembangkan di berbagai workshops (antara 1997-1999). Lebih dari 300 organisasi yang berkontribusi dalam proses modeling ini dan akhirnya CRIPS-DM 1.0 dipublikasikan pada 1999. CRIPS-DM merupakan proses standar yang generik dan tersedia secara bebas untuk proses pengalian data ke dalam strategi penyelesaian masalah bisnis atau penelitian umum lainnnya[3].

Clustering

Clustering mengacu pada pengelompokan catatan, pengamatan atau kasus ke kelas objek yang sama. *Cluster* adalah sebuah koleksi catatan yang mirip satu sama lain dan berbeda catatan dari *cluster* yang lainnya. Algoritma *clustering* mencari segmen data yang sama ke dalam sub-kelompok atau *cluster* [3]

Algoritma DBSCAN

DBSCAN (*Density-based Spatial Clustering of Application with Noise*) merupakan algoritma yang masuk dalam kategori density-based clustering, yaitu proses pembentukan cluster dilakukan berdasarkan tingkat kedekatan/kerapatan jarak antar obyek dalam dataset tersebut [4]. Konsep dasar dari algoritma DBSCAN adalah density reachability dan density connected. Kedua konsep ini bergantung dengan dua parameter input yaitu jarak maksimum dari sebuah ketetanggaan/cluster (*eps*) dan jumlah minimum objek dalam cluster (*minPts*). DBSCAN menyatakan bahwa sebuah cluster dapat dibentuk jika untuk setiap titik data, pada dalam radius tertentu eps dari titik data tersebut terdapat minial minPts titik obyek [5].

2.2. Metodologi

Knowledge Discovery in Database (KDD) menurut Fayyad et al. [5] sebagai proses nontrivial dari mengidentifikasi hubungan yang valid, baru, bermanfaat dan memiliki pola (pattern) yang dapat dimengerti dalam data. KDD sering digunakan untuk menjelaskan proses penggalian informasi tersembunyi dalam suatu *database* yang besar. Tahap-tahap KDD menurut Han et al., 2012[7] adalah sebagai berikut:

1. Pembersihan data (*data cleaning*), merupakan proses yang dilakukan untuk menghilangkan noise dan data yang tidak konsisten atau tidak relevan.
2. Integrasi data (*data integration*), Integrasi data merupakan menggabungkan data dari beberapa sumber yang berbeda.
3. Seleksi data (*data selection*), Data yang diperoleh seringkali tidak semuanya dipakai, maka dilakukan pemilihan data yang sesuai untuk dianalisis.
4. Transformasi data (*data transformation*), Data digabungkan dan diubah ke dalam format yang sesuai sehingga dapat diproses dalam data mining.
5. Data Mining, Proses penting dimana metode diterapkan untuk menemukan pola dan pengetahuan dari data.

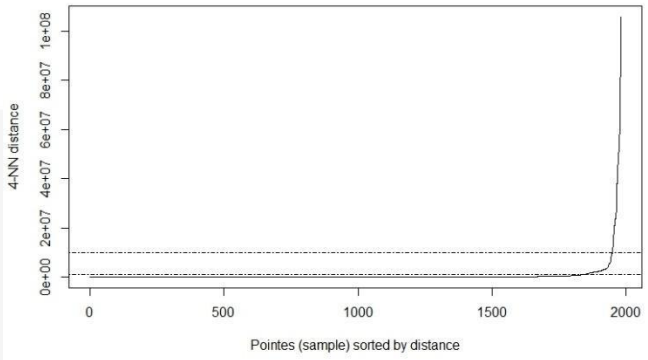
6. Evaluasi pola (*pattern evaluation*), Mengidentifikasikan pola-pola menarik yang di dapat dari proses mining ke dalam *knowledge based* yang ditemukan.
7. Presentasi pengetahuan (*knowledge presentation*), Merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan kepada pengguna.

Data yang digunakan peneliti adalah data Indeks Pembangunan Manusia (IPM) dan Produk Domestik Bruto (PDB) kabupaten/kota di Indonesia tahun 2012 sejumlah 495 data. Atribut data yang dipakai oleh peneliti yaitu nama kabupaten/kota, nilai IPM dan PDB serta latitude dan longitude. Teknik analisis data menggunakan metode *data mining clustering* dengan perangkat lunak Rstudio dan Tableau. Rstudio merupakan alat pengolah data untuk melakukan clustering algoritma DBSCAN. Sedangkan Tableau digunakan untuk memvisualisasikan hasil clustering ke dalam bentuk pemetaan . Algoritma clustering yang digunakan adalah DBSCAN.

3. Pembahasan

3.1. Penentuan nilai *eps* dan *minPts*

DBSCAN dalam melakukan clustering data memerlukan 2 nilai input parameter yaitu *eps* dan *minPts*. Penentuan kedua parameter ini sangat mempengaruhi hasil akhir dari suatu data tergolong dalam bagian suatu *cluster* atau sebuah *noise*. Langkah awal nilai k ditentukan oleh pengguna untuk menentukan k-nearest neihgbor yang digunakan sebagai nilai minPts.

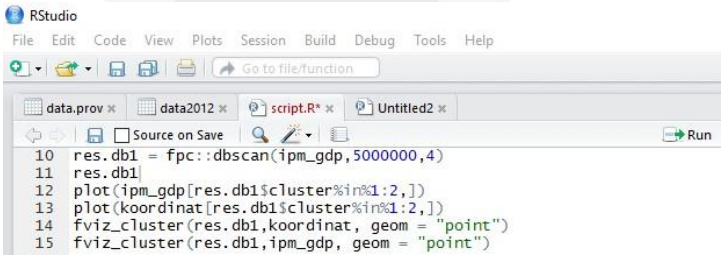


Gambar 2. Ilustrasi Sorted K-Dist Graph Data Tahun 2012

Dalam grafik k-dist terbentuk dalam garis urutan menaik maka dapat ditentukan estimasi nilai eps adalah nilai ketika terjadi peningkatan mencolok pada grafik tersebut dan dapat dilihat bahwa nilai optimal eps berada pada kisaran angka 1000000 dengan minPts bernilai 4.

3.2. Hasil *clustering*

Setelah mengetahui nilai eps dan minPts maka langkah selanjutnya adalah dengan menjalankan algoritma DBSCAN guna menghasilkan klasterisasi. Berikut adalah syntag untuk mengeluarkan hasil cluster dengan bantuan RStudio.



Gambar 3. Syntag Membentuk Cluster dengan Metode DBSCAN

3.3. Validasi Hasil Clustering

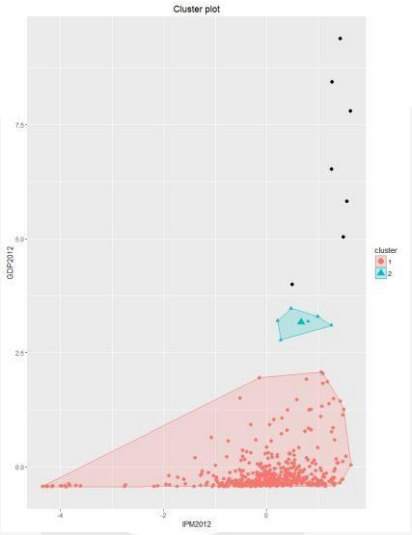
Selanjutnya, untuk mengetahui kualitas cluster yang terbaik dilakukan validasi. Dalam penelitian ini validasi dilakukan dengan menggunakan 2 metode, yaitu metode Silhoutte Index dan Entropy. Berdasarkan hasil metode

validasi Entropy, cluster yang baik adalah dengan nilai Entropy yang lebih kecil[8]. Sedangkan berdasarkan Silhoutte Index, cluster yang baik adalah dengan nilai indeks mendekati 1[9].

eps	entropy	silhoutte index	noise	cluster
1.000.000	0.41888	0.63506	21	4
2.000.000	0.46172	0.67335	21	6
3.000.000	0.47308	0.73259	10	6
4.000.000	0.20292	0.80165	9	3
5.000.000	0.20292	0.80165	9	3
6.000.000	0.20421	0.80499	8	3
7.000.000	0.20421	0.80499	8	3
8.000.000	0.20421	0.80499	8	3
9.000.000	0.20421	0.80499	8	3
10.000.000	0.13963	0.86137	7	2

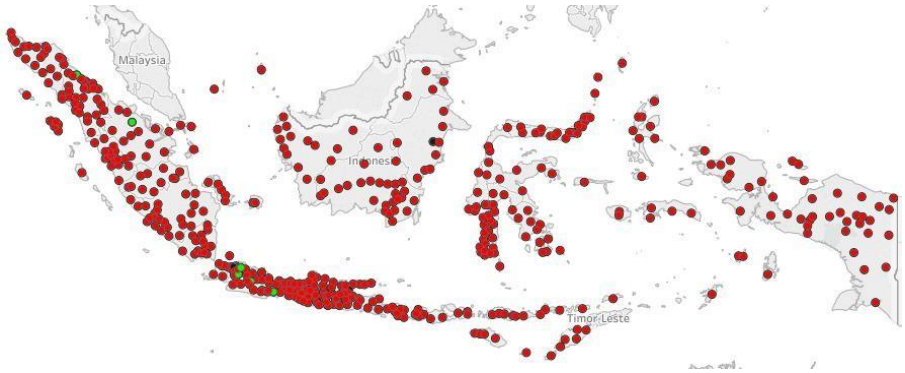
Table 1. Hasil Perhitungan Entropy dan Silhoutte Index (minPts = 4)

Tabel 4.2 menunjukkan bahwa dengan menggunakan nilai minPts = 4 dan kombinasi eps = 1000000 hingga eps = 10000000 di peroleh hasil terbaik dengan nilai eps = 10000000. Kombinasi nilai minPts = 4 dan eps = 10000000 diperoleh entropy 0.13963 dan silhoutte index 0.86137 dengan noise 7. Hasil ini merupakan yang paling baik karena silhoutte index nya sudah mendekati 1 dan noise yang terbentuk tidak banyak.



Gambar 4. Hasil plot minPts = 4 dan eps = 10000000

Berdasarkan visualisasi diatas dengan kombinasi minPts = 4 dan eps = 10000000 terbentuk 2 cluster dari data IPM dan GDP tahun 2012. Hasil plotting juga terlihat cluster 1 memiliki jumlah anggota paling banyak dibandingkan dengan cluster lainnya yaitu sebanyak 482. Hal ini menunjukkan bahwa nilai IPM dan GDP tahun 2012 di seluruh kota/kabupaten di Indonesia memiliki keragaman yang cukup kecil. Sedangkan cluster 2 dengan jumlah anggota yang tidak terlalu banyak yaitu 6 yang terdiri dari Kab. Bekasi, Kab. Bengkalis, Kab. Bogor, Kab. Cilacap, Kota Bandung, dan Kota Medan. Kota/kabupaten yang tidak termasuk dalam cluster manapun ditandai dengan titik hitam yaitu lima wilayah bagian di Jakarta, Kota Surabaya yang terletak di Pulau Jawa dan Kab.Kutai Kartanegara yang terletak di Pulau Kalimantan.



Gambar 5. Hasil pemetaan minPts = 4 dan eps = 10000000

Cluster 1 (titik merah) merupakan kelompok wilayah dengan nilai IPM dan PDB rendah, sepuluh wilayah terendah adalah kabupaten yang berada di wilayah Indonesia bagian timur yang merupakan kategori wilayah terendah dengan rata-rata IPM sebesar 50.05 yang termasuk ke dalam kategori IPM rendah. Sedangkan cluster 2 (titik hijau) merupakan kelompok wilayah dengan rata-rata IPM 75.15 termasuk ke dalam kategori IPM menengah ke atas.

4. Kesimpulan

Berdasarkan hasil yang telah didapatkan dalam bab sebelumnya, dan untuk menjawab pertanyaan penelitian dapat disimpulkan bahwa berdasarkan hasil *clustering* data IPM dan GDP, pembangunan di Indonesia masih belum merata karena terdapat kelompok wilayah yang memiliki nilai tinggi dan kelompok lainnya memiliki nilai rendah. Dan juga terdapat kelompok yang nilainya mengungguli jauh di atas atau dalam cluster terlihat sebagai noise merupakan kota besar di Indonesia.

Dalam mewujudkan pemerataan pembangunan, pemerintah perlu melakukan pembangunan sesuai dengan manajemen pembangunan yang dimulai dengan perencanaan, pengarahan sumber daya, menggerakkan partisipasi masyarakat, koordinasi, pemantauan dan evaluasi serta pengawasan pelaksana pembangunan. Tidak hanya itu pemerintah kabupaten/kota harus membangun sumber daya manusia yang memiliki kompetensi diantaranya dalam hal perencanaan, pelaksanaan rencana, pengorganisasian, kepemimpinan, menajemen sumber daya manusia dan teknologi, kemampuan membangun kerja sama, memberdayakan peran serta masyarakat/swasta, pengawasan dan pengendalian pembangunan dan sebagainya.

DAFTAR PUSTAKA

[1] Dumbill, Edd. (2012). *Big Data Now* (2012 ed.). California, USA: O'Reilly Media, Inc.

[2] Han, Jiawei., Kamber, Micheline., & Pei, Jian. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Waltham, MA, USA: Morgan Kaufmann.

[3] Larose, Daniel T & Larose Chantal D. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining* (2nd ed.). Canada: John Wiley & Sons, Inc.

[4] Furqon, Muhammad Tanzil & Muflikhah, Lailil. (2016). *Clustering The Potential Risk of Tsunami Using Density-Based Saptial Clustering of Application with Noise* (DBSCAN). Indonesia: JEEST.

[5] Fayyad, Usama., et al. (1996). *From Data Mining to Knowledge Discovery in Batabases*. AAAI/Al magazine.

[6] Kantardzic, Medmed. (2011). *Data Mining: Concepts, Models, Methods, and Algorithms.* West Sussex: John Wiley & Sons.

[7] Han, Jiawei., Kamber, Micheline., & Pei, Jian. (2012*). Data Mining: Concepts and Techniques* (3rd ed.). Waltham, MA, USA: Morgan Kaufmann.

[8] Liu, Tao., Liu, S., & Chen, Zheng. (2003). *An Evaluation of Feature Selection for Clustering*. Cina: ICML Conference.

[9] Rousseeuw, Peter J. (1987). *Silhoutte: a graphical aid to the interpretation and validation of cluster analysis.* North-Holland: Journal of Computational and Applied Mathematic.