

## PERBANDINGAN KINERJA K-MEANS DENGAN DBSCAN UNTUK METODE CLUSTERING DATA PENJUALAN *ONLINE RETAIL*

Bena Siti Ashari<sup>1)</sup>, Steven Christ Otniel<sup>2)</sup>, Rianto<sup>3)</sup>

Program Studi Informatika, Fakultas Teknik, Universitas Siliwangi

e-mail: 167006024@student.unsil.ac.id<sup>1)</sup>, 167006038@student.unsil.ac.id<sup>2)</sup>, rianto@unsil.ac.id<sup>3)</sup>

### Abstrak

Perkembangan *Online Shop* melalui media internet sudah dikenal oleh masyarakat saat ini. Beragam kemudahan dalam berbelanja dan bermacam jenis produk maupun jasa yang ditawarkan membuat masyarakat menjadikan *Online Shop* sebagai salah satu sarana yang sering digunakan. Mengetahui karakteristik dari penjualan yang terjadi merupakan suatu hal sangat bermanfaat bagi peningkatan penjualan. Terdapat berbagai metode yang dapat digunakan dalam pengelompokan data, diantaranya : K-Means dan DBSCAN. Belum diketahuinya kinerja kedua metode ini, sehingga perlu pemahaman yang baik ketika akan menggunakannya. Tujuan dari penelitian ini untuk melakukan pengelompokan data penjualan menggunakan metode K-Means dan DBSCAN. Percobaan dilakukan menggunakan dataset sebanyak 500 data dan memiliki 3 atribut: deskripsi, kuantitas barang per transaksi dan harga barang per unit. Hasilnya percobaan pada penelitian ini menunjukkan, penerapan K-Means menghasilkan 3 *cluster* dengan masing-masing berjumlah 103 anggota, 261 anggota dan 134 anggota. Sedangkan menggunakan DBSCAN dengan epsilon 1,005 dan *minimum points* 11 menghasilkan 3 *cluster* dengan masing-masing berjumlah 30 anggota, 47 anggota dan 347 anggota, serta 74 data *noise*.

**Kata Kunci :** *clustering, dbscan, k-means, online retail*

### Abstract

*The development of Online Shop through internet media has been known by the public at this time. Various convenience in shopping and various types of products and services offered make people make the Online Shop as a means that is often used. Knowing the characteristics of sales that occur is very useful for increasing sales. There are various methods that can be used in grouping data, including: K-Means and DBSCAN. The performance of these two methods is not yet known, so a good understanding is needed when using them. The purpose of this study is to group sales data using the K-Means and DBSCAN methods. The experiment was conducted using a dataset of 500 data and has 3 attributes: description, quantity of goods per transaction and price of goods per unit. The results of experiments in this study showed that K-Means management produced 3 clusters with 103 members each, 261 members and 134 members. Whereas using DBSCAN with epsilon 1,005 and minimum points 11 produced 3 clusters with 30 members each, 47 members and 347 members, and 74 data noise.*

**Keywords:** *clustering, dbscan, k-means, online retail*

### I. PENDAHULUAN

Perkembangan Online Shop melalui media internet sudah dikenal oleh masyarakat. Beragam kemudahan dalam berbelanja dan bermacam jenis produk maupun jasa yang ditawarkan membuat masyarakat menjadikan Online Shop sebagai salah satu sarana yang sering digunakan untuk belanja. Adapun banyak macam cara penjualan yang tersedia dalam Online Shop, salah satu diantaranya yaitu penjualan Online Retail. Online Retail merupakan salah satu cara pemesanan produk secara online meliputi semua aktivitas yang melibatkan penjualan barang secara langsung ke konsumen akhir untuk penggunaan pribadi.

Mengetahui karakteristik dari penjualan yang terjadi merupakan suatu hal sangat bermanfaat bagi

peningkatan penjualan. Pengelompokan data penjualan merupakan salah satu solusi untuk mengetahui karakteristik data penjualan. Salah satu masalah dalam pengelompokan data penjualan *online retail* yaitu terdapatnya kelompok data yang sangat beragam. Terdapat berbagai metode yang dapat digunakan dalam pengelompokan data, diantaranya: K-Means dan DBSCAN. Belum diketahuinya kinerja kedua metode ini, sehingga perlu pemahaman yang baik ketika akan menggunakannya.

Metode *non-hierarchical (partitioning) clustering* dalam prosesnya diawali dengan menentukan jumlah cluster yang diinginkan, selanjutnya proses pembentukan cluster dilakukan. Metode ini biasa disebut dengan K-Means Clustering[3] . Pada Metode

K-Means pengelompokan data nonhierarki (sekatan) dilakukan dengan cara mempartisi data ke dalam bentuk dua kelompok atau lebih, sehingga data yang berkarakteristik sama dimasukkan ke dalam satu kelompok yang sama. Sedangkan metode *Density-Based Spatial Clustering Algorithm with noise (DBSCAN)* merupakan suatu algoritma pengelompokan yang didasarkan pada kepadatan (*density*) data. Konsep kepadatan dalam DBSCAN menghasilkan tiga macam status dari setiap data, yaitu inti (*core*), batas (*border*), dan *noise*.

Penelitian ini bertujuan untuk mengetahui kinerja metode K-Means dengan DBSCAN dengan cara mengelompokkan data *online retail*. Hasil pengelompokan difokuskan untuk mengetahui barang yang paling laku terjual berdasarkan kelompok yang optimal, kemudian dilakukan interpretasi berdasarkan kelompok dengan interval yang ditentukan [4].

## II. TINJAUAN PUSTAKA

### 1. Data Mining

*Data Mining* adalah proses yang menggunakan teknik statistik, kecerdasan buatan, *machine learning* untuk mengekstrasi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar (Turban dkk. 2005). Kemampuan *data mining* untuk mencari informasi bisnis yang berharga dari basis data yang sangat besar, dapat dianalogikan dengan penambangan logam mulia dari lahan sumbernya, teknologi ini dipakai untuk :

- Prediksi trend dan sifat-sifat bisnis, dimana data mining mengotomatisasi proses pencarian informasi prediksi di dalam basis data yang besar.
- Penemuan pola-pola yang tidak diketahui sebelumnya, dimana data mining menyapu basis data, kemudian mengidentifikasi pola-pola yang sebelumnya tersembunyi dalam satu sapuan.
- Data mining berguna untuk membuat keputusan yang kritis, terutama dalam strategi.

Menurut [5] *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.

### 2. Pemrosesan Data Awal (*Data Preprocessing*)

Pemrosesan Data menurut George R. Terry, Phd adalah serangkaian operasi informasi yang direncanakan guna mencapai tujuan atau hasil yang diinginkan (Martin M. Lipschutz, 1990). Sedang menurut Gordon B. Davis adalah sebagai bahan mentah dari informasi yang dirumuskan sebagai kelompok lambang-lambang tidak acak yang menunjukkan jumlah atau tindakan- tindakan atau hal

(Gordon B. Davis, 1997). [1]

Kumpulan data yang akan diproses dengan metode- metode dalam data mining sering kali harus melalui pekerjaan awal dikarenakan memiliki masalah seperti populasi data yang terlalu besar, dimensi data yang terlalu tinggi, banyaknya fitur yang tidak berkontribusi besar, dan adanya perbedaan skala pada variabel.

Data transformation (transformasi data) adalah pekerjaan mengubah data ke dalam bentuk yang paling tepat atau cocok untuk proses data mining, untuk masalah data yang variabelnya memiliki skala yang berbeda dapat dilakukan normalisasi. Normalisasi min-max adalah melakukan transformasi linier pada data asli. Normalisasi ini mengubah nilai asli ke berbagai nilai data baru biasanya nilainya 0 – 1 dan dinyatakan dalam bentuk persamaan (1)[6].

$$v_1 = \frac{v_i - \min_A}{\max_A - \min_A} \cdot (\text{new\_max}_A - \text{new\_min}_A) + \text{new\_min}_A$$

### 3. Analisis Kelompok

Analisis Kelompok (*cluster analysis*) adalah pekerjaan mengelompokkan data (objek) yang didasarkan hanya pada informasi yang ditemukan dalam data yang menggambarkan objek tersebut dan hubungan diantaranya[7]. Dalam pengelompokan data, data yang akan dikelompokkan tidak mempunyai label kelas, tetapi dikelompokkan menurut karakteristiknya, barulah kelompok tersebut diberi label sesuai hasil karakteristik kelompok masing-masing. Karena alasan tersebut analisis kelompok sering disebut *segmentation* atau *partitioning*[6].

### 4. K-Means

Merupakan salah satu metode pengelompokan data nonhierarki (sekatan) yang berusaha membagi data ke dalam bentuk dua atau lebih kelompok. Metode ini membagi data ke dalam kelompok sehingga data berkarakteristik sama dimasukkan ke dalam satu kelompok yang sama dan data yang berbeda karakteristik dikelompokkan ke dalam kelompok yang lain[6]. Pengelompokan data dengan metode K- Means secara umum dilakukan dengan algoritma sebagai berikut[6] :

- Menentukan banyaknya k kelompok.
- Membagi data ke dalam k kelompok.
- Menghitung pusat kelompok (sentroid) dari data yang ada di masing-masing kelompok dan dinyatakan dalam bentuk persamaan(2).

$$C_1 = \frac{1}{M} \sum_j^M x_j$$

Dimana C adalah sentroid, M adalah banyak data, i adalah banyak kelompok.

- d. Masing-masing data dialokasikan ke sentroid terdekat. Menghitung jarak data ke setiap sentroid menggunakan jarak Euclidean dan dinyatakan dalam bentuk persamaan(3).

$$D(X_i, C_i) = \sqrt{\sum_{j=1}^q (x_{ij} - c_{ij})^2}$$

### 5. DBSCAN

*Density – Based Spatial Clustering Algorithm With Noise (DBSCAN)* adalah algoritma pengelompokan yang didasarkan pada kepadatan (*density*) data. Konsep kepadatan yang dimaksud dalam DBSCAN adalah banyaknya data (*minPts*) yang berada dalam radius *Eps* ( $\epsilon$ ) dari setiap data. Konsep kepadatan seperti ini menghasilkan tiga macam status dari setiap data, yaitu inti (*core*), batas (*border*), dan *noise*. Data inti merupakan data yang jumlah data di dalam radius *eps* lebih dari *minPts*, data *noise* merupakan data yang jumlah data di dalam radius *eps* kurang dari *minPts*, dan data batas merupakan data yang jumlah data di dalam radius *eps* kurang dari *minPts* tetapi menjadikan data tetangganya menjadi data inti. Proses pengelompokan DBSCAN adalah menghitung jarak titik pusat (*p*) ke titik yang lain menggunakan jarak Euclidean dan dinyatakan dalam persamaan (4).

$$D(x_1, p_1) = \sqrt{\sum_j^q (x_{1j} - p_{1j})^2}$$

Iterasi dilakukan pada setiap titik yang menjadi tetangga titik pusat (*p*) dan titik yang belum dikunjungi[8] [1].

### III. METODOLOGI

Dalam penelitian ini digunakan dataset yang telah diunduh dari UCI Dataset. Data set yang diperoleh yaitu *Online Retail Dataset* yang merupakan kumpulan data transnasional di UK atau negara-negara di eropa yang berisi semua transaksi yang terjadi pada 01/12/2010 sampai 09/12/2011. Perusahaan ini merupakan perusahaan retail yang menjual beberapa produk unik yang diperlukan dalam suatu acara/event secara online.

Pada dataset ini terdapat beberapa atribut, diantaranya yaitu : *VoiceNo* atau Invoice Nomor, *StockCode* atau Kode Produk per item, *Description* atau Nama Produk, *Quantity* atau kuantitas dari semua produk per transaksi, *InvoiceDate* atau tanggal dan waktu, *UnitePrice* atau harga per unit, *CustomerID* berupa nomor ID, *Country* atau negara. Namun setelah dilakukan proses cleaning hanya digunakan 3 atribut yaitu *Description*, *Quantity per Transaction* dan *Unit Price*. Data yang tersimpan sebanyak 500 data.

## IV. HASIL DAN PEMBAHASAN

### 1. Implementasi

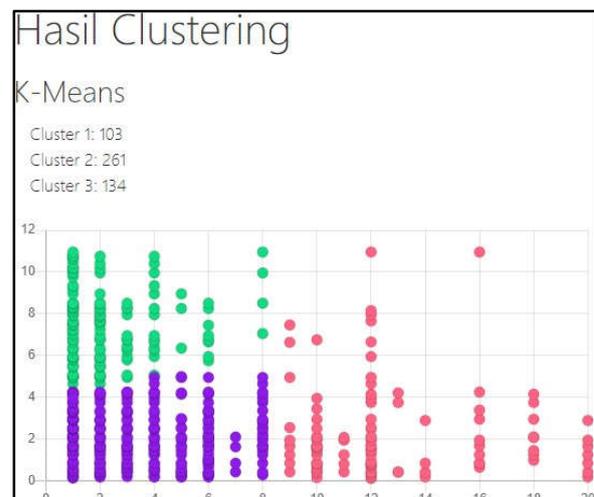
Hasil analisa diperoleh dengan cara mengimplementasikan sistem ke dalam aplikasi berbasis web. Tampilan aplikasi terlihat pada gambar 1.



Gambar 1. Tampilan Upload Dataset

Terdapat sebuah *button* untuk *upload file* data set yang telah siap digunakan. Data set yang diunggah harus berupa format json dengan 3 attribute yang bernama A, B dan C. A adalah untuk deskripsinya B dan C untuk atribut numerisnya. Selanjutnya data tersebut akan langsung di proses. Untuk penelitian kali ini digunakan data set online retail dengan 3 kolom atribut, yaitu deskripsi barang, jumlah penjualan per transaksi dan harga barang per unit.

Setelah data diproses, selanjutnya akan ditampilkan hasil clustering dalam bentuk diagram, dengan 2 hasil untuk perbandingan yaitu hasil dari metode K-Means dan metode DBSCAN.

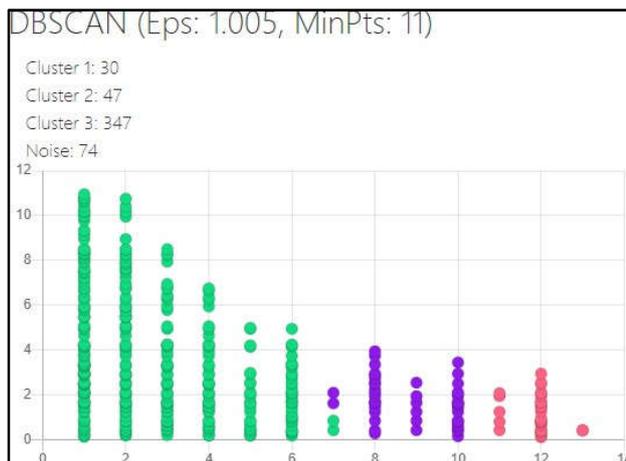


Gambar 2. Hasil Pengelompokan K-Means.

Hasil pengelompokan dari dataset Retail Online menggunakan metode K-Means dengan jumlah kelompok tersaji dalam Tabel 1.

Tabel 1. Hasil Pengelompokan K-Means

Kelompok	1	2	3
Jumlah Data	103	261	134



Gambar 3. Hasil Clustering DBSCAN.

Hasil pengelompokan dari data set Retail Online dengan menggunakan metode DBSCAN dengan nilai Eps: 1.005 dan MinPts: 11 tersaji dalam Tabel 2.

Tabel 2. Hasil Pengelompokan DBSCAN

Kelompok	1	2	3
Jumlah Data	30	47	347

## 2. Interpretasi Hasil Pengelompokan

Hasil dari K-means ada 3 cluster dengan cluster 1 yang ditandai dengan warna merah adalah data yang jumlah penjualan per transaksinya besar dan harga barang tinggi, untuk cluster 2 yang ditandai dengan warna ungu jumlah penjualan per transaksinya kecil dan harganya rendah dan untuk cluster 3 yang ditandai dengan warna hijau jumlah penjualan per transaksinya kecil dan harganya tinggi.

Sedangkan untuk hasil dari DBSCAN dengan epsilon 1,005 dan minimal pointnya 11 pada cluster 1 yang ditandai dengan warna merah adalah data dengan harga tinggi, cluster 2 dengan warna ungu adalah data dengan harga sedang dan cluster 3 dengan warna hijau

adalah data dengan harga rendah. Sisanya 74 data adalah noise yang tidak masuk kedalam cluster manapun.

## V. KESIMPULAN DAN SARAN

Pengelompokan data dari data set Retail Online dengan menggunakan K-Means dengan jumlah kelompok tiga menghasilkan kelompok 1 dengan 103 data, kelompok 2 dengan 261 data, kelompok 3 dengan 134 data. Sedangkan pengelompokan menggunakan metode DBSCAN dengan nilai Epsilon 1.005 dan Minimal Points: 11 menghasilkan kelompok 1 dengan 30 data, kelompok 2 dengan 47 data dan kelompok 3 dengan 347 data dan sisa 74 data adalah noise yang tidak masuk kedalam cluster manapun.

## DAFTAR PUSTAKA

- [1.] F. Rahadi, "Pengertian Pemrosesan Data," p. 1, 14 November 2011.
- [2.] N. P. I. S. I. Devi, "Implementasi Metode Clustering DBSCAN pada Proses Pengambilan Keputusan," vol. 6, p. 3, 2015.
- [3.] A. H. Astqolani, Online Shop di Era Multi Technology, 2013.
- [4.] A. I. G. Heni Sulastri, "PENERAPAN DATA MINING DALAM PENGELOMPOKAN PENDERITA THALASSEMIA," Jurnal Nasional Teknologi dan Sistem Informasi, vol. 03, 2017.
- [5.] B. Santosa, Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis, Yogyakarta: Graha Ilmu, 2007.
- [6.] D. S. D. I. Sisca Agustin Diani Budiman, "PERBANDINGAN METODE K- MEANS DAN METODE DBSCAN PADA PENGELOMPOKAN RUMAH KOST MAHASISWA DI KELURAHAN TEMBALANG SEMARANG," Jurnal Gaussian, vol. 5, pp. 757-762, 2016.
- [7.] S. Santoso, Statistik Multivariant, Jakarta: Alex Media Komputindo, 2010.
- [8.] E. Prasetyo, Data Mining Konsep dan Aplikasi menggunakan MATLAB, Yogyakarta: Andi, 2012.
- [9.] P. S. M. K. V. Tan, Introduction to Data Mining, New York: Pearson Addison Wesley, 2006.
- [10.] Eriyanto, Teknik Sampling Analisis Opini Publik, Yogyakarta: LKis Yogyakarta, 2007.