BAB I
PENDAHULUAN

1.1 Latar belakang

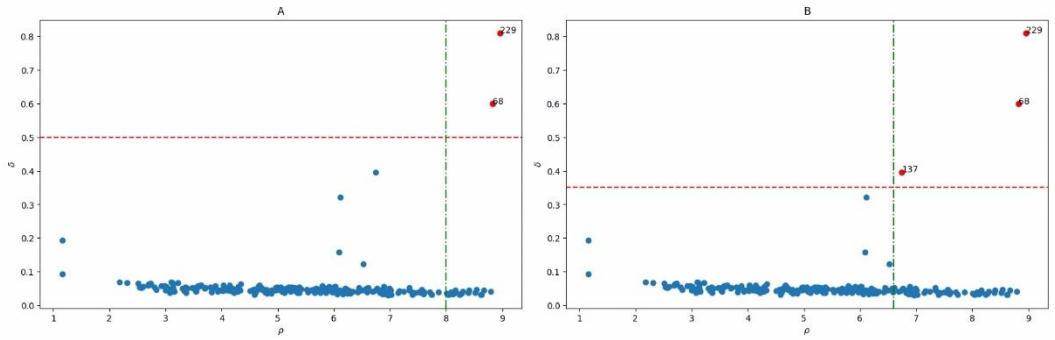
Seiring dengan pesatnya perkembangan teknologi, internet, dan banyaknya perangkat elektronik, data pun akan bertambah semakin cepat. Pada tahun 2017 saja terdapat 2.5 triliun bytes data yang dihasilkan per hari (Ghosal et al., 2019). Menurut perkiraan terbaru, saat ini data dihasilkan mencapai 328.77 juta *terabytes* per hari (Duarte, 2023). Dengan banyaknya data yang ada, kita membutuhkan metode canggih yang dapat secara otomatis melakukan analisis dan mengenali pola-pola sehingga dapat menghasilkan sebuah informasi yang berguna. Contohnya dalam konteks media sosial, pengguna akan cenderung memposting opini mengenai produk maupun orang (*content creator*), data opini ini dapat dikumpulkan dan diproses untuk mengetahui aspek apa yang banyak disukai dan tidak disukai (Flores & Garza, 2020). Dalam hal ini, metode *clustering* merupakan salah satu metode dapat digunakan untuk menganalisis sebuah data secara otomatis.

Clustering merupakan salah satu algoritma *unsupervised learnin*g yang paling penting (Xu et al., 2020). Untuk mencari informasi pada data yang kompleks, kita dapat melakukan *clustering* data terlebih dahulu, yaitu dengan mengelompokkan data yang sejenis dalam satu kelas sesuai dengan karakteristik data. Metode *clustering* telah banyak mengalami perkembangan dan telah diterapkan secara luas pada data *mining* (Batubara et al., 2020), *image processing* (Hu et al., 2021), *bioinformatics* (Y. Zhang & Kiryu, 2022), *recommendation* (Saputra et al., 2019), dan masih banyak lagi. Sampai saat ini, banyak algoritma *clustering* yang telah dikembangkan, diantaranya adalah *K-means clustering, fuzzy C-means clustering (FCM), Density-based Spatial Clustering of Application (DBSCAN)*, dan lainnya. Algoritma-algoritma ini dikategorikan menjadi lima kelompok, yaitu *Partitional Clustering (Distance-Based Clustering), Hierarchical Clustering, Density-Based Clustering, Grid-Based Clustering, dan Model-Based Clustering.*

K-means adalah salah satu algoritma klasterisasi yang paling populer dan banyak digunakan (Ahmed et al., 2020). Prinsip dasar K-means adalah membagi n objek ke dalam k klaster berdasarkan kedekatan rata-rata. Proses ini melibatkan dua langkah utama: penugasan (*assignment*) dan pembaruan (*update*). Pada tahap penugasan, setiap titik data ditempatkan ke klaster dengan *centroid* terdekat. Pada tahap pembaruan, centroid setiap klaster dihitung kembali sebagai rata-rata dari titik-titik data dalam klaster tersebut. Proses ini diulang hingga konvergensi tercapai, yaitu ketika centroid tidak berubah secara signifikan. Meskipun efektif, algoritma K-means memiliki beberapa keterbatasan, seperti keharusan menentukan jumlah klaster (k) terlebih dahulu dan ketergantungan pada pemilihan awal centroid yang dapat mempengaruhi hasil akhir (Poerwanto & Ali, 2019).

Selain itu algoritma *Density-Based Clustering* juga merupakan salah satu algoritma yang paling penting dan banyak digunakan (Z. Zhang et al., 2021). *Density-Based Clustering* bergantung pada gagasan menemukan kepadatan pada suatu wilayah dalam kumpulan data (Bhattacharjee & Mitra, 2020). Tujuannya adalah untuk menemukan klaster pada tingkat yang berbeda-beda. Dalam *Density-Based Clustering*, daerah dengan kepadatan lebih tinggi akan dipisahkan dan diidentifikasi sebagai pusat klaster, sedangkan daerah dengan kepadatan lebih rendah digunakan sebagai partisi (Ghosal et al., 2019). Algoritma *Density-Based Clustering* tidak membutuhkan jumlah klaster sebagai parameter masukan (Campello et al., 2020). Algoritma-algoritma yang termasuk ke dalam kategori *Density-Based Clustering*, yaitu *Density-based Spatial Clustering of Application (DBSCAN), Mean Shift, Spectral Method, Subtractive Method*, dan lain-lain.

Pada tahun 2014 Rodriguez dan Laio mengusulkan sebuah algoritma canggih yang Bernama *Density Peaks Clustering (DPC)* dan termasuk dalam kategori *Density-Based Clustering*. Sejak pertama kali diusulkan, algoritma DPC telah banyak diterapkan pada berbagai penelitian (Y. Wang et al., 2024). Algoritma DPC memiliki prinsip yang sederhana dan efisiensi yang tinggi (Z. Wang & Wang, 2020). Seperti algoritma *Density-Based* lainnya, algoritma DPC juga menghitung kepadatan untuk setiap titik data. Proses *clustering* dari algoritma DPC dimulai dengan menentukan puncak atau pusat klaster, kemudian data lainnya akan digabungkan dengan data yang terdekat dan memiliki nilai kepadatan lebih tinggi.



Gambar 1.1 Titik Pusat

Namun algoritma DPC masih memiliki masalah dalam penentuan pusat klaster. Penentuan pusat klaster masih dilakukan secara manual melalui grafik keputusan dengan melihat titik-titik data yang mempunyai nilai kepadatan lokal (ρ) dan jarak lokal (δ) yang relatif tinggi, atau dapat diidentifikasi dengan memilih titik-titik data yang berada pada kanan atas grafik keputusan. Namun, seberapa besar nilai yang dianggap tinggi? Tidak ada standar yang jelas terkait batasan sebuah nilai yang akan dianggap tinggi. Proses ini akan menjadi sangat subjektif tergantung dari batasan yang ditetapkan oleh masing-masing pengguna. Contoh masalah dapat dilihat pada Gambar 1.1. Titik yang berwarna merah merupakan contoh pusat klaster yang mungkin dipilih. Kedua gambar menunjukkan jumlah pusat yang berbeda, gambar (a) dipilih dua data, sedangkan gambar (b) dipilih tiga titik data.

GB-DPC (Gap Based-Density Peaks Clustering) adalah varian dari algoritma DPC yang memperkenalkan metode deteksi pusat klaster secara otomatis berdasarkan selisih (gap) antara titik data. Metode ini menggunakan perbedaan antara titik data dengan nilai kepadatan lokal dan jarak lokal yang berdekatan untuk menentukan pusat klaster secara lebih efektif. GB-DPC mengatasi beberapa kelemahan dari DPC standar, seperti ketergantungan pada parameter yang harus ditentukan secara manual dan kesulitan dalam mengidentifikasi pusat klaster yang optimal (Flores & Garza, 2020).

Oleh karena itu, dalam penelitian ini digunakan metode untuk mengatasi masalah yang telah dijelaskan sebelumnya. Untuk menyelesaikan masalah, digunakan metode *Inter Quartile Range* (IQR) untuk menetapkan batasan sebuah

titik yang dapat dinyatakan sebagai pusat klaster. Kemudian, untuk lebih meningkatkan kualitas pusat klaster yang dipilih, digunakan metode untuk menentukan konektivitas setiap sub klaster ke sub klaster lainnya, sehingga memungkinkan untuk menggabungkan sebuah subklaster ke klaster yang tepat. Metode ini diharapkan dapat membantu algoritma DPC dalam memilih pusat klaster secara otomatis. Dalam penelitian ini, metode ini disebut sebagai algoritma DPC-IQRSM.

Sebagai perbandingan, penelitian ini juga akan mengevaluasi kualitas hasil klasterisasi dari algoritma DPC-IQRSM dengan algoritma K-means dan GB-DPC. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam bidang klasterisasi data, khususnya dalam hal peningkatan akurasi dan otomatisasi penentuan pusat klaster.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijabarkan, maka rumusan masalah yang akan dibahas adalah sebagai berikut:

1. Bagaimana cara algoritma *Density Peak Clustering* dapat mengklasterkan sebuah data?
2. Bagaimana cara menentukan pusat klaster pada algoritma *Density Peak Clustering* secara otomatis?
3. Bagaimana hasil *clustering* dari algoritma *Density Peak Clustering* setelah diterapkan otomatisasi pada proses pemilihan pusat klaster?

1.3 Tujuan

Berdasarkan rumusan penelitian yang telah dirumuskan di atas, maka tujuan penelitian ini adalah sebagai berikut:

1. Untuk mengetahui cara algoritma *Density Peak Clustering* dalam mengklasterkan sebuah data
2. Untuk mengetahui salah satu cara menentukan pusat klaster pada algoritma *Density Peaks Clustering* secara otomatis.
3. Untuk mengetahui hasil *clustering* dari algoritma *Density Peaks Clustering*.

1.4 Manfaat

Berdasarkan tujuan penelitian yang telah dirumuskan di atas, manfaat yang diharapkan dari penelitian ini adalah sebagai berikut:

1. Memberikan informasi tentang proses algoritma *Density Peak Clustering* dalam mengklasterkan sebuah data.
2. Memberikan informasi tentang cara memilih pusat klaster secara otomatis algoritma *Density Peaks Clustering*.
3. Dapat digunakan sebagai bahan evaluasi maupun referensi pada penelitian selanjutnya yang berkaitan dengan algoritma *Density Peaks Clustering* maupun algoritma *clustering* lainnya.

1.5 Batasan Masalah

Batasan masalah yang dapat ditentukan dalam penelitian ini adalah sebagai berikut:

1. Penelitian ini menggunakan parameter masukan *p* sebesar 2.
2. Data yang digunakan dalam penelitian ini merupakan data sekunder (data yang diambil secara tidak langsung oleh peneliti) yang diperoleh melalui penelitian terdahulu dan memiliki dimensi yang rendah.