BAB I

PENDAHULUAN

1.1 Latar Belakang

Analisis klaster merupakan salah satu metode statistika multivariat yang melakukan sebuah usaha untuk menggabungkan objek ke dalam beberapa kelompok (klaster), di mana anggota klaster itu tidak diketahui sebelumnya. Dengan kata lain, analisis klaster merupakan analisis statistik yang digunakan untuk mengelompokkan *n* objek ke dalam *k* buah klaster dengan setiap objek dalam klaster memiliki kemiripan yang tinggi dibandingkan objek antar klaster (Wicaksono, 2017:1).

Prinsip dasar dalam analisis klaster adalah mengelompokkan objek pada suatu klaster yang memiliki kemiripan sangat tinggi dengan objek lain dalam klaster yang sama, tetapi tidak mirip dengan objek lain pada klaster yang berbeda. Hal ini berarti susunan klaster yang baik akan mempunyai homogenitas yang tinggi antar anggota dalam satu klaster dan heterogenitas yang tinggi antar klaster yang satu dengan yang lainnya (Nuningsih, 2010:2).

Analisis klaster mempunyai beberapa asumsi yang harus dipenuhi, yaitu data bebas dari pencilan (*outlier*) dan bebas dari masalah multikolineritas. Pencilan merupakan data yang memiliki karateristik yang berbeda dengan data lainnya. Adanya pencilan dapat mengubah struktur sebenarnya dari populasi sehingga klaster-klaster yang terbentuk menjadi kurang sesuai dengan struktur sebenarnya. Sedangkan multikolinearitas adalah keberadaan hubungan linear yang sempurna atau tepat di antara sebagian atau seluruh variabel bebas. Adanya multikolinearitas mengakibatkan himpunan data yang akan diolah ke dalam beberapa klaster menjadi tidak akurat. Maka dari itu, kedua hal ini harus dihindari dari data yang akan diolah. Cara mengecek pencilan menggunakan jarak Euclid sedangkan untuk mengatasi masalah multikolinearitas menggunakan nilai *z-score* yang diperoleh setelah mentransformasikan data secara linier sehingga terbentuk sistem koordinat baru dengan varians maksimum atau biasa disebut dengan Analisis Komponen

Utama (AKU). AKU digunakan untuk meringkas data tanpa mengurangi karaterisktik data tersebut secara signifikan.

Metode pengelompokkan dalam analisis klaster dibagi dua, yaitu metode hirarki dan metode non-hirarki. Metode hirarki digunakan apabila belum ada jumlah klaster yang dipilih. Metode hirarki dibedakan menjadi dua pengelompokkan, yaitu aglomeratif dan divisif. Pada metode aglomeratif, proses pengelompokkan dimulai dari *n* klaster sehingga masing-masing objek dipandang sebagai sebuah klaster, kemudian dua klaster terdekat digabungkan yang kemudian membentuk sebuah klaster baru. Proses penggabungan terus dilakukan sampai terbentuk menjadi satu klaster yang memuat semua himpunan data. Beberapa metode aglomeratif antara lain *Single Linkage, Average Lingkage, Complete Lingkage,* dan *Ward's Method*. Sedangkan pada metode divisif, proses pengelompokkan dimulai dengan *n* objek yang digabungkan ke dalam satu klaster, kemudian klaster tersebut dipartisi ke dalam dua klaster, seterusnya sampai terbentuk menjadi *n* klaster dengan tiap klasternya beranggotakan satu objek. Beberapa metode divisif antara lain *monothetic divisive clustering* dan *polythetic divisive clustering*. Metode non-hirarki digunakan untuk mengelompokkan *n* objek ke dalam *k* klaster dimana $k < n$ dan nilai *k* sudah ditentukan sebelumnya. Beberapa metode non-hirarki antara lain *Fuzzy C-Means,* K-Means, K-Medoids, dan CLARA (Wicaksono, 2017:2).

K-Means merupakan metode pengklasteran secara *partitioning* yang memisahkan data ke dalam kelompok yang berbeda. Metode ini dikembangkan oleh James B Mac-Queen pada tahun 1967. K-Means merupakan metode pengelompokkan yang paling terkenal karena sederhana dan dapat digunakan dengan mudah di berbagai bidang. Dasar pengelompokkan dalam metode ini adalah menempatkan objek berdasarkan rata-rata (mean) klaster terdekat. Sehingga terbentuk suatu kelompok yang antar objeknya memiliki kesamaan karateristik atau homogenitas yang tinggi.

Pada dasarnya, mean adalah pengukuran yang sangat rentan terhadap pencilan. Sebuah pencilan yang bernilai ekstrim dapat menggeser rata-rata dari sebagian besar data yang kemudian menjadi tidak seimbang. Menurut Kaufmann

dan Rosseuw pada tahun 1990, metode K-Means akan lebih sensitif terhadap data yang mengandung pencilan karena menggunakan mean sebagai ukuran nilai tengahnya. Oleh karena itu, kajian tentang metode pengelompokkan yang tahan terhadap pencilan diperlukan karena keberadaan pencilan dalam sebuah data terkadang tidak dapat dihindarkan.

Di sisi lain, median adalah statistik deskriptif yang cenderung lebih tahan terhadap pencilan sehingga berkembanglah metode yang dapat mengelompokkan data yang mengandung pencilan yaitu metode K-Medoids yang merupakan salah satu dari variansi metode K-Means. Pada metode K-Means, pengelompokkan didasarkan pada nilai mean klaster terdekat sedangkan dasar pengelompokkan dalam metode K-Medoids adalah menempatkan objek berdasarkan nilai tengah (median) klaster terdekat. Oleh karena itu, penggunaan metode K-Medoids akan meminimalkan *error* pada klaster.

K-Medoids atau biasa disebut algoritma PAM (*Partitioning Around Medoids*) dikembangkan oleh Leonard Kaufman dan Peter J. Rousseeuw pada tahun 1987, dan algoritma ini sangat mirip dengan K-means terutama karena keduanya algoritma *partitioning* atau keduanya memecah dataset menjadi kelompok-kelompok, dan keduanya bekerja berusaha untuk meminimalkan kesalahan. Algoritma PAM (*Partitioning Around Medoids*) menggunakan metode partisi pada analisis klaster untuk mengelompokkan sekumpulan *n* objek menjadi sejumlah *k* klaster. Algoritma ini menggunakan objek pada kumpulan objek untuk mewakili sebuah klaster. Objek yang terpilih untuk mewakili sebuah klaster disebut *medoid*. Klaster dibangun dengan menghitung kedekatan yang dimiliki antara *medoid* dengan objek *non-medoid* (Kaufmann dan Rouseeuw, 1990:68).

Namun pada metode PAM (*Partitioning Around Medoids*) bekerja efektif untuk himpunan data yang kecil, tetapi tidak berjalan baik untuk himpunan data yang besar. Untuk bekerja dengan himpunan data yang besar, maka dibentuk metode baru yaitu metode berbasis sampling yang disebut dengan CLARA (*Clustering Large Applications*).

CLARA menggunakan himpunan data sampel secara random atau acak. Algoritma PAM kemudian diterapkan untuk menghitung medoid terbaik dari

sampel tersebut. Idealnya, sampel seharusnya menyajikan data yang sangat mirip dengan himpunan data asli. Maka dari itu, semakin besar data yang akan diolah menjadi beberapa klaster akan semakin baik karena dengan proses sampling akan memberikan probabilitas yang sama kepada setiap objek untuk dipilih ke dalam sampel. Objek-objek yang dipilih menjadi pusat klaster (medoid) akan cenderung mirip dengan yang sudah dipilih dari seluruh himpunan data. CLARA akan melakukan analisis klaster dari banyak sampel secara acak dan menghasilkan analisis klaster terbaik sebagai hasilnya.

Tingkat keefektifan CLARA bergantung pada ukuran sampel. Dalam algoritma PAM *(Partitioning Around Medoids)* mencari K-Medoid terbaik di antara himpunan data, tetapi CLARA *(Clustering Large Applications)* mencari K-Medoid terbaik di antara himpunan data sampel yang terpilih. CLARA tidak bisa menghasilkan analisis klaster yang baik jika medoid yang diperoleh dari sampel terbaik sangat jauh dari K-Medoid terbaik. Dengan demikian, berdasarkan pemaparan di atas, penulis tertarik untuk mengaji analisis klaster melalui metode CLARA secara mendalam pada himpunan data yang besar. Oleh karena itu, penelitian ini berjudul **"CLARA (*Clustering Large Application*) pada Data Simulasi Trivariat 1000 Objek"**

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah di atas, rumusan permasalahan dalam penelitian ini adalah bagaimana hasil penerapan metode CLARA (*Clustering Large Application*) dalam pembentukan klaster pada data simulasi trivariat 1000 objek?

1.3 Tujuan Penulisan

Kajian dan pemaparan terhadap permasalahan di atas bertujuan untuk menerapkan metode CLARA (*Clustering Large Application*) dalam pembentukan klaster dan hasilnya pada data simulasi trivariat 1000 objek.

1.4 Manfaat Penulisan

Adapun manfaat dari penulisan penelitian ini adalah:

1. Teoritis

   Secara teoritis manfaat penulisan ini adalah untuk memperdalam dan memperkaya pengetahuan tentang analisis statistik multivariat khususnya analisis klaster dengan metode CLARA (*Clustering Large Application*).

2. Praktis

   Secara praktis manfaat penulisan ini adalah sebagai bahan untuk pertimbangan dan masukan bagi pihak yang berkepentingan serta dapat menjadi informasi yang mendukung terlaksananya tujuan dari pihak yang berkepentingan.

1.5 Sistematika Penulisan

Adapun sistematika penulisan sebagai berikut:

BAB I : PENDAHULUAN

Bab ini membahas tentang latar belakang, rumusan masalah, tujuan penulisan, manfaat penulisan, dan sistematika penulisan.

BAB II : TINJAUAN PUSTAKA

Bab ini membahas tentang analisis klaster, CLARA, validasi klaster, dan interpretasi klaster.

BAB III : METODOLOGI PENELITIAN

Bab ini membahas tentang jenis dan sumber data, analisis data dengan bahasa R, dan langkah-lankah penelitian..

BAB IV : HASIL DAN PEMBAHASAN

Bab ini membahas tentang deskripsi data, tahapan pengolahan data, pengujian pencilan dan multikolinearitas, hasil analisis klater, interpretasi klaster dan validasi klaster.

BAB V : KESIMPULAN DAN SARAN

Bab ini membahas tentang kesimpulan mengenai keseluruhan isi penulisan dan saran untuk penelitian selanjutnya.