BAB I
PENDAHULUAN

1.1. Latar Belakang

Teknologi yang semakin maju berpengaruh besar dalam berbagai segi kehidupan manusia. Seluruh kegiatan yang dilakukan mengandung berbagai macam informasi, sehingga jumlah informasi akan terus mengalami peningkatan dari waktu ke waktu. Peningkatan informasi tersebut harus diikuti dengan penyimpanan yang baik tentunya. Informasi-informasi tersebut dapat disimpan dalam sebuah *database*. Kumpulan berbagai informasi tersebut disebut sebagai *dataset.* Namun, tanpa adanya pengorganisasian dokumen tersebut secara berkala mengakibatkan penumpukan dan menyebabkan proses pencarian menjadi sulit dan membutuhkan waktu lama. Bahkan ketika mencari dokumen yang memiliki tema yang sama akan membutuhkan waktu yang lama sehingga tidak efisien. Salah satu cara untuk menyelesaikan masalah tersebut menggunakan teknik *Clustering* dengan mengelompokkan dokumen yang memiliki tingkat kesamaan menjadi satu kelompok. Alasan penggunaan Teknik *Clustering* adalah sesuai dengan tujuan dari analisis *cluster* yaitu, untuk mengelompokkan data berdasarkan tingkat kesamaan pada data sehingga akan lebih mudah dalam melakukan analisis pada studi kasus. Analisis *cluster* menurut Anderberg dalam tulisan Mukti (2012) adalah suatu analisis *statistic multivariate* yang bertujuan untuk mengetahui struktur data dengan menempatkan kesamaan objek observasi ke dalam satu kelompok data sehingga dapat dibedakan antara kelompok satu dengan kelompok yang lain atau dengan cara memisahkan kasus atau objek ke dalam beberapa kelompok yang mempunyai sifat berbeda antar kelompok yang satu dengan yang lain.

Ada banyak metode *Clustering* yang dapat digunakan, diantaranya metode *Single Linkage Clustering*, *Complete Linkage Clustering*, *Group Average Agglomerative Clustering, K-Means* (MacQueen, 1967) , dan *K-Medoids.*

Berdasarkan metode-metode tersebut, metode yang sering digunakan adalah metode *K-Means.* Kelebihan dari metode yang dikembangkan oleh McQueen ini yaitu, mampu mengelompokkan dokumen dalam jumlah yang besar dengan waktu komputasi yang cepat. Prinsip dari metode ini adalah mempartisi suatu koleksi dokumen menjadi beberapa *cluster* dan menentukan *centroid* (titik pusat awal) secara acak. Tetapi, metode *K-Means* ini juga memiliki kelemahan yaitu sensitif terhadap *outlier.*

Selain *K-Means,* terdapat metode lain untuk mengelompokkan yaitu, *K-Medoids.* Prinsip dari metode ini hampir sama dengan *K-Means* hanya saja *centroid* yang digunakan tidak mengambil nilai rata-rata dari objek dalam sebuah *cluster* sebagai titik acuan, namun menggunakan *medoid* sebagai titik acuan yang merupakan objek dalam sebuah *cluster* yang paling terpusat. Kelebihan dari metode ini adalah termasuk metode yang fleksibel karena hampir dapat bekerja pada setiap jenis data matriks dan mampu mengelompokkan dokumen dalam jumlah yang besar dengan waktu komputasi yang cepat dan mampu mengatasi *outlier.*

Metode *K-Medoids* yang telah dimodifikasi merupakan penelitian yang dilakukan oleh Gandhi dan Sarivastava (2014) dalam jurnal dengan judul "*Analysis and Implementation of Modified K-Medoids Algorithm to Increase Scability and Efficiency for Large Dataset*". Jurnal ini membahas tentang analisis peningkatan skalabilitas dan efisiensi pada kumpulan data besar dengan menggunakan modifikasi algoritma *K-Medoids.* Algoritma yang dibahas dalam jurnal ini yaitu, algoritma *K-Medoids* yang telah dimodifikasi. Berdasarkan hasil penelitian dimana algoritma *K-Means* dan *K-Medoids* dibandingkan dengan algoritma *K-Medoids* yang sudah dimodifikasi sehingga diperoleh hasil bahwa algoritma Gandhi dan Sarivastava (2014) dalam mengelompokkan data dengan skalabilitas dan efisiensi yang lebih baik dari metode *K-Means* dan *K-Medoids.* Selain itu, algoritma ini juga menghasilkan kualitas *cluster* dan waktu komputasi yang lebih efektif.

Berdasarkan hasil penelitian yang ditulis oleh Gandhi dan Sarivastava (2014) dalam jurnal dengan judul **"*Analysis and Implementation of Modified K-***

***Medoids Algorithm to Increase Scability and Efficiency for Large Dataset*"** dengan modifikasi algoritma *K-Medoids* yang diusulkan lebih baik daripada *K-Means* dan *K-Medoids.* Untuk mengetahui bagaimana perbandingan hasil dari ketiga metode tersebut maka, penulis ingin membahas tentang efektivitas pengelompokan kumpulan data besar menggunakan algoritma *K-Medoids* yang sudah dimodifikasi dibandingkan dengan menggunakan algoritma *K-Means* dan *K-Medoids* dalam penelitian yang berjudul **"Efektivitas Algoritma *K-Means, K-Medoids,* dan Optimal Jumlah *Cluster* Untuk Modifikasi Algoritma *K-Medoids* Pada Kumpulan Data Besar"**.

1.2. Rumusan Masalah

Berdasarkan penjelasan pada latar belakang, maka permasalahan yang dapat diidentifikasi dalam penelitian ini antara lain adalah:

1. Bagaimana hasil perbandingan efektivitas Algoritma *K-Means, K-Medoids,* dan modifikasi algoritma *K-Medoids* dari segi kecepatan waktu komputasi?
2. Pada *cluster* berapa pengelompokkan menggunakan modifikasi algoritma *K-Medoids* akan optimal?

1.3. Batasan Masalah

Batasan masalah dalam penelitian ini adalah:

1. Data yang digunakan dalam penelitian ini yaitu berupa data sekunder yang didapatkan dari *website* Kaggle yaitu IMDB 5000 *Movie Dataset.*
2. Data-data yang diolah yaitu sampel data sebanyak 1000 data film pada tahun 2007 – 2016 dengan data yang digunakan adalah *Movie Title, Content Rating,* IMDB (*Internet Movie* Database) *Score, Gross, dan Budget*.
3. Metode yang digunakan dalam penelitian ini adalah Metode *K-Means, K-Medoids,* dan Modifikasi algoritma *K-Medoids* pada kumpulan data besar.
4. Efektivitas yang dilihat untuk algoritma *K-Means, K-Medoids,* dan modifikasi algoritma *K-Medoids* adalah kecepatan waktu komputasi.

5. Alat analisis yang digunakan adalah analisis waktu komputasi dan Uji Mann-Whitney.
6. Data diolah dengan menggunakan bantuan *software* MATLAB, IBM SPSS, dan *Microsoft Excel.*

1.4. Jenis Penelitian dan Metode Analisis

Jenis penelitian dalam tugas akhir ini adalah penelitian teoritis, yang mengacu pada topik yang berjudul *Analysis and Implementation of Modified K-Medoids Algorithm to Increase Scability and Efficiency for Large Dataset* dalam penelitian Gandhi dan Sarivastava (2014). Metode analisis yang akan digunakan adalah Modifikasi Algoritma *K-Medoids* yang hasilnya akan dibandingkan dengan Algoritma *K-Means* dan *K-Medoids* untuk kumpulan data besar.

1.5. Tujuan Penelitian

Tujuan penelitian dalam tugas akhir ini adalah:

1. Mengetahui efektivitas Algoritma *K-Means*, *K-Medoids*, serta modifikasi algoritma *K-Medoids* dari segi kecepatan waktu komputasi.
2. Mengetahui optimal jumlah *cluster* pada modifikasi algoritma *K-Medoids.*

1.6. Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini adalah mempermudah dan mempercepat penyelesaian kasus untuk mengelompokkan data dalam jumlah besar dengan Modifikasi Algoritma *K-Medoids*.