BAB I
PENDAHULUAN

1.1 Latar belakang masalah

Data mining merupakan bidang keilmuan yang biasanya terintegrasi dengan bidang keilmuan lain seperti ilmu statistik, *artificial intellegence*, sistem Basis Data serta yang lainnya [4]. *Data mining* merupakan sebuah proses untuk menggali informasi yang tersembunyi atau menemukan pola yang bermanfaat dalam sekumpulan data yang besar. Untuk mendukung proses penggalian informasi ini terdapat task-task (proses) dalam data mining, yaitu : klasifikasi, *clustering*, dan asosiasi.

Tugas akhir ini membahas mengenai *clustering*. *Clustering* merupakan pendekatan dalam *data mining* yang cukup populer untuk mengelompokkan data [1,2]. *Clustering* dilakukan dengan cara mengelompokkan data ke dalam beberapa kelompok/*cluster* berdasarkan karakteristiknya, dimana data dalam satu *cluster* akan memiliki kemiripan karakteristik antara satu dan yang lainnya, dan sangat berbeda karakteristiknya dengan data pada *cluster* yang lainnya.

Terdapat banyak metoda yang telah dikembangkan untuk melakukan *clustering* pada data. K-means merupakan salah satu metode yang efisien dalam melakukan pengelompokan data[1,7]. Algoritma ini mengelompokkan data dengan 4 operasi dasar : (1) menentukan jumlah *cluster*/ nilai k sebagai inisiasi awal, (2) menghitung jarak antar objek dan rata-rata/centroid dari sebuah *cluster*, (3) mengalokasikan masing-masing data ke centroid yang terdekat, (4) menghitung kembali rata-rata setiap *cluster* dari objek-objek yang sudah dialokasikan pada *cluster* tersebut hingga jarak *intra cluster* sangat dekat [7]. Algoritma k-means dinilai sangat mudah dan cukup efisien dalam *data mining*. Namun, yang menjadi masalah adalah k-means ini hanya terbatas untuk data yang bertipe numerik, karena pengelompokan dilakukan dengan menghitung means/rata-rata dari suatu data dengan data yang lain, sedangkan untuk data kategorik tidak dapat dihitung means-nya, dalam kenyataannya di dunia nyata sering juga ditemukan data yang bertipe kategorik.

Untuk menangani permasalahan diatas, k-modes, sebuah algoritma baru yang merupakan variasi dari algoritma k-means, mencoba memberikan solusi untuk pengelompokan pada data bertipe kategorik [5,6]. Seperti yang juga terdapat pada algoritma k-means, terdapat sebuah permasalahan dalam algoritma k-modes ini yaitu saat menentukan k inisialisasi awal. Ada 2 metode pemilihan inisialisasi awal yang digunakan, yang pertama adalah nilai k diambil dari k *record* pertama pada data. Kedua, menentukan k inisialisasi dengan menggunakan *frequency based method* yang dibahas lebih detil pada Bab II.

1.2 Perumusan masalah

Berdasarkan latar belakang tersebut, masalah yang dikaji pada TA ini adalah :

1. Bagaimana cara melakukan *clustering* data dengan menggunakan algoritma k-modes pada data kategorik.
2. Bagaimana kualitas hasil *clustering* yang dihasilkan algoritma k-modes dengan menggunakan metode inisialisasi random dibandingkan dengan menggunakan metode *frequency based.*
3. Bagaimana pengaruh perubahan nilai k terhadap akurasi hasil *clustering* pada algoritma k-modes dengan inisialisasi menggunakan metode inisialisasi random dan *frequency based.*

Adapun batasan masalah tugas akhir ini adalah sebagai berikut :

1. Data yang digunakan adalah data dengan tipe kategorik yang sudah punya kelas (untuk keperluan evaluasi hasil *clustering*).
2. Evaluasi kualitas hasil *clustering* dilakukan dengan menghitung akurasi suatu data diklasterkan ke dalam kelas yang benar.

1.3 Tujuan

Tujuan pengerjaan Tugas Akhir ini berdasarkan rumusan masalah di atas adalah :

1. Mengimplementasikan metode inisialisasi random dan *frequency based* pada algoritma k-modes.
2. Menganalisis parameter k-modes (jumlah k, metode penentuan centroid awal random dan *frequency based*) terkait dengan performansi hasil clustering.

1.4 Metodologi penyelesaian masalah

Metode yang digunakan dalam penyelesaian tugas akhir ini adalah menggunakan metode studi pustaka atau studi literatur dan analisis dengan langkah kerja sebagai berikut :

1. Studi Literatur :
   a. Pencarian referensi, mencari referensi dan sumber-sumber lain yang layak yang berhubungan dengan *data mining*, *clustering,* algoritma K-modes
   b. Pendalaman materi, mempelajari dan memahami materi yang berhubungan dengan tugas akhir
2. Pengumpulan data
   Mencari data kategorik untuk keperluan analisis algoritma K-modes.
3. Implementasi perangkat lunak
   a. Analisis dan design perangkat lunak
      Melakukan analisis dan desain perangkat lunak, mengenai kebutuhan sistem serta fungsionalitas – fungsionalitas yang dibutuhkan dalam sistem.
   b. Pengkodean

Pembuatan program berdasarkan analisis dan desain program yang telah ditentukan pada tahap sebelumnya menggunakan teknik pemrograman berorientasi objek.

c. Pengujian
   Menguji aplikasi yang telah dibuat.

4. Analisis hasil
   Menganalisis hasil pengelompokan yang telah dilakukan, terhadap jumlah klaster, dan kedua metode pemilihan centroid awal terkait dengan performansi.

5. Pembuatan laporan Tugas Akhir
   Mengambil kesimpulan dari hasil analisis dan pembuatan laporan tugas akhir.