BAB I
PENDAHULUAN

1.1 Latar Belakang

Clustering bekerja dengan mengelompokkan objek-objek data (pola, entitas, kejadian, unit, hasil observasi) ke dalam sejumlah *cluster* tertentu (Xu and Wunsch, 2009). Dengan kata lain algoritma Clustering melakukan pemisahan, pemecahan atau segmentasi data menjadi kelompok (*cluster*) yang dipisahkan dari kelompok lain dalam dua cara yaitu pengamatan dalam sebuah kluster mirip satu sama lain dan berbeda dengan pengamatan di kelompok lain(Hämäläinen et al., 2017) ke dalam sejumlah kelompok (*cluster*) menurut karakteristik tertentu. Penerapannya dapat ditemukan di bidang dan disiplin ilmu yang sangat berbeda seperti kedokteran, pembelajaran mesin, pengenalan pola, analisis gambar. Ada banyak metode pengelompokan untuk membuat cluster secara otomatis. Masing-masing menggunakan strategi dan fungsi tujuan khusus untuk memperoleh kelompok ini.

Halkidi et al., (2001) memperkenalkan konsep dasar pengelompokan dengan mensurvei algoritma pengelompokan yang dikenal secara luas dengan cara komparatif atau perbandingan. Selain itu, penelitiannya juga membahas masalah penting tentang proses pengelompokan mengenai penilaian kualitas hasil pengelompokan, yang mana salah satu teknik yang paling populer untuk pengelompokan adalah K-means karena teknik ini termasuk yang paling baik dan sederhana diantar yang lainnya.

Dalam literatur terdapat banyak algoritma pengelompokan, dan sangat sulit untuk memberikan jenis yang baik tentang metode clustering, karena jenis ini mungkin akan saling tumpang tindih, sehingga beberapa metode akan memiliki sifat atau karateristik yang hamper sama dari beberapa jenis. Walaupun demikian, ada pentingnya juga mempresentasikan gambaran yang relative terorganisir dari beberapa metode pengelompokan. Secara umum, metode pengelompokan utama dapat diklasifikasikan sebagai beriku dimana Han, *et. al* (2012) memperkenalkan Metode clustering

yang terdiri dari beberapa jenis, yaitu: metode partisi, metode hirarki, metode berbasis kepadatan, metode berbasis grid, metode berbasis model.

Dalam *clustering* menggunakan konsep partisi terdapat tiga konsep yang dapat digunakan, yakni partisi klasik, partisi *fuzzy* dan partisi probabilistik. Pada partisi klasik, suatu data secara eksklusif menjadi anggota hanya pada satu cluster saja. Pada partisi fuzzy, nilai keanggotaan suatu data pada suatu cluster terletak pada interval [0,1]. Jumlah derajat keanggotaan setiap data pada semua cluster adalah 1. Sementara pada partisi *possibilistics*, jumlah nilai keanggotaan suatu data pada semua *cluster* tidak harus 1. Namun begitu, untuk menjamin suatu data menjadi anggota dari paling tidak satu *cluster* maka diharuskan nilai keanggotaannya lebih dari 0 (Kusumadewi dkk. 2006). Contoh metode partitional *clustering*: *K-Means, Fuzzy C-means* dan *Mixture Modelling.*

Algoritma *clustering K-means* sensitif terhadap inisial nilai, pilihan acak dari lokasi centroid pada awal algoritma, perlakuan variabel sebagai angka dan jumlah cluster yang tidak diketahui k, dimana nilai awal yang berbeda dapat menyebabkan perbedaan hasil clustering Algoritma *K-means* sehingga memungkinkan untuk memilih dua atau lebih cluster dalam sebuah cluster(Min & Kai-fei, 2015). Seperti banyak lainnya algoritma penambangan data, K-means telah mengurangi kehebatan kala menerima data dengan dimensi tinggi karena set data hampir selalu terlalu jarang, hal ini terjadi karena penggunaan jarak Euclidean juga tidak mampu mengidentifikasi atribut-atribut penting dari atribut yang tidak relevan sehingga hasilnya, semua jarak antara elemen data tampak identik yang menjadi tidak berarti dalam ruang dimensi tinggi (Francois, 2007). Solusi yang mungkin dilakukan adalah menggabungkan K-means dengan metode ekstraksi data dimana hasilnya menunjukkan pengurangan dimensi tanpa pengawasan terkait erat dengan pembelajaran tanpa pengawasan (Ding et al., 2015).

Arbelaitz *et al.,*(2013) menyatakan fakta bahwa algoritma pengelompokan optimal tidak ada untuk setiap kasus menjadi kelemahan utama dari teknik ini. Telah ditetapkan bahwa inisialisasi yang berbeda dari algoritma yang sama menghasilkan kelompok yang berbeda di banyak lingkungan yang berbeda dengan karakteristik yang berbeda dan tidak ada yang terbaik dalam semua kasus. Jadi, dalam proses

pengelompokan yang efektif kita harus memilih partisi yang paling cocok dengan data. Banyak algoritma *clustering* perlu menyediakan parameter M awal seperti *K-Means* dan *Fuzzy C-Means*.

Fuzzy c-means (FCM) adalah metode pengelompokan fuzzy yang dikemukakan oleh Dunn pada tahun 1973 dan digeneralisasi oleh Bezdek pada tahun 1981 yang merupakan perluasan yang memungkinkan elemen-elemen menjadi milik beberapa kluster secara bersamaan (Ouchica et., al 2018), sering disebut *C-Means* sebagai c jumlah kelas atau cluster(Nayak et., al 2015). Karim (2011) menjelaskan bahwa cara kerja *Fuzzy C- Means* (FCM) melakukan metode pengelompokan yang memungkinkan satu bagian dari data untuk memiliki dua atau lebih kelompok. Pada kondisi awal, pusat pengelompokan ini masih belum akurat. Tiap-tiap data memiliki derajat keanggotaan untuk tiap-tiap kelompok. Dengan cara memperbaiki pusat kelompok dan nilai keanggotaan tiap-tiap data secara berulang, maka akan dapat dilihat bahwa pusat kelompok akan bergerak menuju lokasi yang tepat

Seleksi *centroid* awal tidak hanya mempengaruhi efisiensi algoritma tetapi juga jumlah iterasi yang diinginkan untuk menjalankan algoritma *k-means* asli dimana ketika memilih *centroid* k awal secara acak akan mengarah pada rekomendasi yang tidak akurat dan peningkatan biaya untuk pelatihan offline (Zahra et,. al 2015) serta hasil cluster terakhir tidak selalu terjamin (Nazeer et,. al 2010). Akhirnya untuk mengatasi masalah tersebut diusulkan metode heuristik untuk meningkatkan akurasi dan efisiensi algoritma klaster k-means dan fuzzy C-means. Algoritma yang dimodifikasi kemudian diterapkan untuk pengelompokan data(Zahra et., 2015)

Metode *K-Dimensional Tree* atau *KD-Tree* merupakan metode untuk mengelompopokan data. Disebut k-dimensional tree karena *kd-tree* adalah *binary tree* yang node-nya merupakan *k-dimensional* point. K dapat bernilai 2 atau lebih. Cara k-d-tree mengelompokkan data adalah sama dengan *binary tree* biasa. Suatu region dibagi menjadi dua, kemudian masing-masing dua region tadi dibagi lagi menjadi dua region, demikian seterusnya hingga tidak dapat dibagi lagi(Rudi, R hermanto, 2012).

Kd Tree K–Means Clustering (Stephen and Conor, 2007) merupakan algoritma pengoptimalan dari algoritma kkz/katsoudivinis (I. Katsavounidis dkk, 1994)yang menggunakan nilai kerapatan dan jarak antar poin/data dalam menentukan posisi awal titik tengah *cluster k means clustering*. *Kd tree k-means clustering* menggunakan struktur data *k-dimensional tree* dalam proses inisialisai titik tengah *cluster*. *Kd tree k means clustering* juga merupakan metode yang mampu menangani data set yang memiliki *noise/outlier* dengan cara menghapus 20 persen calon titik tengah *cluster* dengan nilai kerapatan terendah (Stephen and Conor, 2007).

Masalah utama dari *clustering* adalah masalah jumlah cluster yang optimal. Masalah ini dapat diselesaikan dengan proses hasil akhir dari algoritma pengelompokan, yang terdiri dari memilih jumlah cluster yang optimal dengan menggunakan indeks validitas. Indeks validitas cluster sangat penting karena dirancang dengan tujuan memperkirakan seberapa baik sebuah partisi cocok dengan struktur dasar dataset (Lianyu & Caiming, 2019). Indeks ini dibangun dengan menggabungkan ukuran kekompakan dan ukuran pemisahan. Adapun pemisahan dilakukan dengan menghitung jarak antara pusat *cluster* yang digunakan. Namun jaraknya begitu jauh tidak selalu mencerminkan kualitas partisi antara cluster dan terkadang memberikan hasil yang mengelirukan(Said *et al*., 2017).

Tantangan lainnya dalam pengelompokan adalah memperbaiki kondisi awal karena perkiraan awal dan acak dari sebagian besar algoritma pengelompokan akan mempengaruhi keandalan hasil, oleh karena itu dalam mengevaluasi algoritma pengelompokan, dua parameter digunakan untuk menentukan ukuran kinerja pengelompokan (Salem et., al 2005) yang pertama adalah ukuran validasi, yang digunakan untuk menentukan seberapa baik algoritma bekerja pada set nilai parameter tertentu, dan yang kedua adalah ukuran pengulangan, yang digunakan untuk mempelajari efek kondisi awal pada keanggotaan cluster.

Banyak teknik validasi klaster tersedia (Zhao et.,al 2012). Evaluasi ini bisa digunakan untuk menentukan jumlah cluster paling dapat diandalkan dalam satu set data. Bezdek(1998) mengusulkan beberapa indeks validitas sperti statistik Hubert,

indeks Davies-Bouldin, dan indeks Dunn dimana Dunn memberikan hasil validasi terbaik untuk simulasi yang disajikan. Halkidi et., al (2001) menyatakan Dalam banyak validitas indeks dua sifat cluster diperhitungkan, yaitu, kekompakan dan keterpisahan. Input parameter signifikan dari banyak algoritma pengelompokan adalah jumlah cluster, yang sering dipilih terlebih dahulu. Dengan demikian, masalah utama adalah bagaimana masalah hasil pengelompokan data. Melakukan partisipasi partisi Diperlukan yang diperoleh dengan menggunakan algoritma *clustering* yang berbeda nilai parameter input beberapa kali. Lalu, klaster indeks validitas digunakan untuk menemukan partisi terbaik data.

Dalam literatur terkenal validitas klaster indeks seperti, mis Dunn (Dunn, 1974), Davies-Bouldin (DB) dalam (Davies DL & Bouldin DW, 1979), Pakhira et., al (2004) indeks Davies-Bouldin, indeks Dunn dan indeks Xie-Beni, atau Silhouette (SIL) (P. J. Rousseeuw, 1987) seringkali digunakan ketika membandingkan hasil teknik pengelompokan yang berbeda. Indeks Dunn adalah rasio intercluster minimum jarak ke diameter kluster maksimum. Gantinya, indeks Davies-Bouldin (DB) adalah rasio dari jumlah penyebaran dalam cluster ke pemisahan antar-cluster. Indeks siluet (SIL) adalah rata-rata Saat melakukan tugas pengelompokan hasil harus diperlakukan dengan hati-hati (Saha S, Bandyopadhyay S, 2012) Identifikasi jumlah cluster yang tepat dan teknik partisi yang tepat adalah beberapa pertimbangan penting dalam *clustering* karena pengelompokan adalah tugas subjektif yang sulit.

Penelitian mengenai analisis kelompok dan indeks validitas telah banyak dilakukan. Seperti penelitian yang menggunakan metode *Kmeans* dan indeks validitas *Davies Bouldin* yang diimplementasikan pada data sekuens DNA untuk mengenali suatu spesies dan membedakan spesies yang satu dengan yang lainnya berdasarkan kesamaan ciri yang dimiliki (Sinurat, 2014).

penelitian ini akan membahas **PENERAPAN VALIDITY INDEKS PADA K-MEANS DAN FUZZY C-MEANS** dalam mengidentifikasi jumlah cluster yang tepat dan teknik partisi yang tepat. Pengelompokan dilakukan berdasarkan kemiripan sifat yang dimiliki. Pendekatan *probabilistic* (kemungkinan) dapat berguna dalam analisis

kelompok partisi sehingga mendapatkan strategi pengelompokan yang cocok termasuk perbedaan data dan hubungan kesamaan acak.

1.2 Masalah penelitian

Masalah penellitian yang akan dibahas dari penulisan tesis ini akan dibagi menjadi 2(dua) bagian yaitu:

1.2.1 Identifikasi Masalah

Berikut ini akan dilakukan identifikasi masalah berdasarkan latar belakang yang diuraikan antara lain:

1. Pemilihan *centroid* awal yang mempengaruhi hasil dari algoritma k-means dan fuzzy c-means.
2. Jumlah *cluster* algoritma *K-Means* dan *Fuzzy C-Means* yang tidak optimal.

1.2.2 Rumusan Masalah

Titik pusat *cluster* yang ditentukan secara acak pada algoritma *K-Means* dan *Fuzzy C-Means* mengakibatkan pengelompokan data menjadi tidak stabil, sehingga perlu adanya suatu pendekatan untuk menentukan titik pusat *cluster* untuk kemudian digunakan sebagai titik pusat *cluster* awal pada proses *clustering* algoritma *K-Means dan Fuzzy c-Means* sehingga menghasilkan k dan c optimal yang lebih akurat dalam *clustering* suatu dataset.

1.3 Tujuan dan Manfaat Penelitian

Tujuan dari penelitian ini adalah untuk meningkatkan akurasi hasil dari algoritma *K-Means* dan *Fuzzy C-Means* dengan menentukan titik awal pusat cluster menggunakan metode *KD-Tree* dan mengevaluasi hasil *cluster* menggunakan *validity index.* Manfaat dari penelitian ini untuk meningkatkan kinerja algoritma *K-Means* dan *Fuzzy C-Means* dengan metode *Kd-Tree* dalam mencari k dan c optimal.

1.4 Ruang lingkup

Berikut ini akan dijelaskan ruang lingkup penelitian dalam penulisan tesis ini antara lain :

1. Dataset yang digunakan pada penelitian ini yaitu *sales transactions dataset weekly* dari situs *archive.ics.uci.edu* yang terdiri dari 52 atribut yang tersedia di: https://archive.ics.uci.edu/ml/datasets/Sales_Transactions_Dataset_Weekly
2. Penentuan *centroid* untuk *K Means* dan *Fuzzy C-Means* menggunakan algoritma *Kd Tree*
3. Evaluasi hasil cluster menggunakan *Davies Bouldin index*

1.5 Metodologi penelitian

Metodologi yang digunakan dalam penelitian ini adalah sebagai berikut:

1. Studi literatur

   Pada tahapan ini dilakukan proses untuk memahami bagaimana tahapan cara kerja dari algoritma *K-Means* dan *Fuzzy C-Means* berdasarkan sumber atau referensi berupa jurnal, buku dan website.
2. Analisis masalah

   Pada tahap ini dilakukan analisis berdasarkan hasil studi literatur untuk mengidentifikasi masalah yang harus diselesaikan, data yang dibutuhkan dan menentukan metode yang diusulkan untuk menyelesaikan masalah.
3. Perancangan model

   Pada tahap ini dilakukan perancangan model dengan membuat *flowchart* yang menggambarkan proses pemilihan *centroid* dengan *Kd tree* dan menerapkan *validity index* pada hasil *clutering K-Means* dan *Fuzzy C-Means.*
4. Pengujian

   Pengujian menggunakan *Microsoft Office Excel* dan *Matlab*. Pengujian yang dilakukan:

   a. Melakukan pemilihan *centroid* dengan *Kd tree*

   b. Melakukan pengujian menggunakan hasil pemilihan *centroid* dengan *Kd*

tree pada *K-Means* dan *Fuzzy C-Means*

c. Mengevaluasi hasil *cluster* dari *K-Means* dan *Fuzzy C-Means* Menggunakan *validity index.*

5. Menarik kesimpulan dari hasil pengujian
6. Menyusun laporan Tesis

1.6 Sistematika penulisan

Sistematika penulisan laporan penelitian ini terdiri dari 5 Bab, dimana secara garis besar masing-masing bab membahas hal-hal berikut ini. Bab 1 pendahuluan berisi penjelasan umum berupa studi literatur dan masalah pada algoritma *K-Means* dan *Fuzzy C-Means* dan solusi yang sudah ada dan proses yang akan dilakukan. Bab 2 berisi studi litaratur tentang data mining, tahapan data mining, metode data mining, *clustering* serta *Validity Index*. Bab 3 Metodologi penelitian, berisi identifikasi masalah, langkah-langkah dari metode yang diusulkan yang digambarkan dengan *flowchart*, data yang digunakan, alat-alat penelitian dan metode analisis. Bab 4 hasil dan pengujian, berisi hasil yang diperoleh dari sistem yang dibangun dan pengujian penelitian yang dilakukan dan saran yang dapat dilakukan untuk hasil yang lebih baik dari penelitian selanjutnya. Bab 5, berisi hasil dari pengujian dengan menerapkan pemilihan centroid dengan *Kd tree* yang diproses dengan algoritma *K-Means* dan *Fuzzy C-Means* dan hasil cluster dari algoritma tersebut akan dievaluasi dengan *Validity index*, serta menarik kesimpulan dari proses algoritma tersebut.