BAB 1

PENDAHULUAN

1.1 Latar Belakang Masalah

Data Mining merupakan proses untuk mendapatkan informasi dari basis data untuk pengekstrakan informasi baru yang diambil dari data dengan jumlah yang besar untuk keperluan pengambilan keputusan [**1**]. Ada berbagai macam karakteristik data dilihat dari jumlah data dan jumlah atribut data, pada studi kasus penelitian ini akan melakukan segmentasi pelanggan distribustor farmasi. Data yang digunakan memiliki karakteristik data dengan 3 atribut yang dilihat dari nilai *Recency* (jumlah hari yang dihitung dari tanggal terakhir transaksi ke tanggal acuan), *Frequency* (jumlah transaksi yang telah dilakukan), dan *Monetary* (total biaya yang telah dikeluarkan untuk transaksi) dengan jumlah data sebanyak 500 pelanggan. Tipe data memiliki nilai angka, yang diambil dari riwayat transaksi dalam periode satu tahun dari Januari sampai Desember di tahun 2016. Segmentasi akan dilakukan pada tanggal acuan 1 januari 2017 karena pertimbangannya jika ada pelanggan yang bertransaksi pada tanggal 31 Desember 2016 maka dapat menghitung nilai *Recency* sebanyak 1 hari.

Dalam kasus segmentasi maka terdapat cara penyelasaian yaitu dengan teknik *Clustering*. *Clustering* merupakan proses untuk mengelompokan suatu data dengan cara membandingkan kemiripan antar data, sehingga setiap data masuk ke dalam *Cluster* (grup) [**2**]. *Clustering* termasuk kedalam *unsupervised learning* (pembelajaran tidak terbimbing), karena ketika pengelompokan data didasarkan

atas kemiripan dan ketidakmiripan antar datanya. Analisis *cluster* telah banyak digunakan termasuk riset pasar, pengenalan pola, analisis data, dan pengolahan citra. Dalam bisnis, *clustering* dapat membantu pemasar untuk menemukan kepentingan pelanggan mereka berdasarkan pada pola pembelian dan ciri kelompok pelanggan [**3**], oleh karena itu dalam kasus segmentasi pelanggan, maka diperlukan algoritma *clustering* yang cocok berdasarkan karakteristik dari datanya. Untuk mengetahui algoritma *clustering* apa yang cocok digunakan pada segmentasi pelanggan, maka perlu mengetahui karakteristik dari beberapa algoritma *clustering.* Pada teknik partisi *clustering* terdapat algoritma K-means, algoritma K-means merupakan suatu algoritma *clustering* yang mempartisi dataset kedalam beberapa k *cluster,* algoritma K-means cukup mudah untuk diimplementasi dan dijalankan, relatif cepat, mudah disesuaikan dan banyak digunakan [**4**]. Proses segmentasi pelanggan terbanyak menggunakan algoritma K-means *clustering*, dan hasil menyatakan bahwa segmentasi sukses dan efektif [**5**]. Algoritma K-means digunakan untuk kasus segmentasi pelanggan karena penentuan jumlah *cluster* dapat disesuaikan dengan kebutuhan sehingga jumlah *cluster* dapat diasumsikan diawal proses *cluster*. namun algoritma K-means mempunyai kelemahan yaitu harus dapat menduga jumlah *cluster* diawal [**6**], karena asumsi jumlah *cluster* yang ditentukan diawal belum tentu menghasilkan *cluster* baik untuk hasil segmentasi. Kemudian algoritma K-means lainnya yaitu dalam penentuan titik *centroid* awal yang masih dipilih secara acak [**7**] yang akan berpengaruh pada kondisi *local minima* (kondisi hasil *cluster* belum tentu yang terbaik). Penentuan titik centroid akan berpengaruh pada iterasi proses untuk

menghasilkan *cluster convergence* yang artinya hasil *cluster* memiliki anggota *cluster* yang tidak berubah [**8**].

Pada saat kondisi dimana data segmetasi pelanggan tidak ingin ditentukan sesuai kebutuhan, namun ingin menghasilkan segmentasi dari sebaran data, maka untuk penentuan asumsi jumlah *cluster* tidak diduga diawal, namun harus dicari asumsi jumlah *cluster* yang akan terbentuk. Untuk menangani kondisi tersebut, terdapat pengembangan dari algoritma K-means, yaitu algoritma Dinamyc K-means (DK) yang memiliki proses untuk mencari jumlah *cluster* tanpa harus menduga asumsi jumlah *cluster* [**9**], Pada penelitian Ahamed Shafeeq dan Hareesha hasil *cluster* memiliki kualitas *cluster* yang baik berdasarkan nilai *intra* dan *inter cluster*, dibandingkan dengan algoritma K-means namun memiliki beberapa kekurangan yang sama dengan algoritma K-means yaitu masalah pada penentuan titik *centroid* pada proses *cluster* yang masih dipilih secara acak.

Berdasarkan penelitian dari Yugar Kumar dan G.Sahoo untuk masalah penentuan titik centroid dilakukan dengan cara mengembangkan K-means menjadi K-means Binary Search Centroid (KBSC) yang memiliki proses dalam menenentuan titik centroid menggunakan pendekatan teknik *Binary Search* [**8**]. Pada hasil penelitian oleh Yugar Kumar dan G Sahoo, algoritma KBSC diuji pada data Iris,Wine, dan Diabetes dengan masing-masing jumlah atribut (4, 13 dan 8), berdasarkan hasil percobaan Algoritma KBSC memiliki nilai *intra* dan *inter cluster* yang lebih baik dari algoritma K-mean, dan juga jika dilihat berdasarkan akurasi *cluster* algoritma KBSC baik digunakan pada data dengan jumlah atribut = 4 dan jumlah data 150 yaitu pada hasil data Iris jika dibandingkan dengan

percobaan data Wine dan Diabetes. Algoritma KBSC memiliki keterbatasan dalam menentukan asumsi jumlah *cluster* yang akan dibentuk.

Berdasarkan penjelasan masing-masing Algoritma DK dan KBSC dapat dilihat beberapa kelebihan dan kekurangannya. Berikut ini perbandingan kelebihan dan kekurangan masing-masing algoritma :

Tabel 1.1 Perbandingan Kelebihan dan Kekurangan Algoritma

Algoritma	**Penentuan Pusat *Cluster* awal**	**Penentuan Jumlah *Cluster***
K-means Binary Search Centroid [**8**]	√	-
Dynamic K-means [**9**]	-	√

Di Tabel 1.1 menjelaskan mengenai kelebihan dan kekurangan yang dimiliki masing-masing algoritma. Oleh karena itu maka diusulkan penggabungan algoritma DK dengan algoritma KBSC dapat saling melengkapi proses *cluster* pada area asumsi jumlah *cluster* dan penentuan titik *centroid cluster*, sehingga dapat menghasilkan nilai DBI yang terbaik untuk studi kasus segmentasi pelanggan distributor farmasi.

1.2 Identifikasi Masalah

Pada algoritma DK dan KBSC diimplementasikan pada data set yang sudah digunakan pada penelitian di bidang *clustering* yang diperoleh dari UCI *repository of machine learning databases*, namun tidak dijelaskan bagaimana karakteristik kinerja algoritma tersebut dan data seperti apa yang cocok untuk diproses *cluster* oleh algoritma tersebut.

Berdasarkan permasalahan tersebut maka hal berikutnya yang dilakukan dalam penelitian ini adalah meneliti karakteristik algoritma dalam menangani data

dengan variasi jumlah data, jumlah atribut dan jumlah *cluster* menggunakan *dataset* model. Sehingga dapat diketahui seperti apa karakteristik dari penggabungan algoritma DK dan KBSC, kemudian setelah mengetahui karakteristik tersebut algoritma dapat diuji pada segmentasi pelanggan distributor farmasi dengan 3 atribut dan 500 data.

1.3 Premis

Berdasarkan latar belakang masalah maka didapat premis sebagai berikut ini :

1. Algoritma Dynamic K-means memiliki kemampuan dalam mencari jumlah *cluster* dibandingkan, namun algoritma Dynamic K-means memiliki kekurangan dalam penentuana asumsi titik awal *centroid* yang berpengaruh terhadap jumlah iterasi proses untuk membentuk *cluster convergence* dan akan terjadi kondisi *local minima* (hasil *cluster* belum tentu yang terbaik).
2. Algoritma KBSC memiliki kemampuan dalam menentukan asumsi titik *centroid* dan memiliki hasil *cluster* dengan nilai *intra cluster* (kemiripan dalam satu *cluster*) dan *inter cluster* (perbedaan data antar *cluster*) yang baik, namun masih memiliki kekurangan harus bisa menduga asumsi jumlah *cluster* yang akan dibentuk.

1.4 Hipotesis

Pengabungan algoritma DK dan KBSC dapat menghasilkan DBI *cluster* yang lebih baik daripada algoritma lainnya ketika digunakan pada data segmentasi

data pelanggan dengan jumlah atribut *3 (Retency, Frequency, dan Monetery)* dan jumlah data 500 pelanggan.

1.5 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut :

1. Menggabungkan algoritma DK dengan KBSC sehingga dapat menghasilkan DBI *cluster* yang lebih baik dari algoritma lainnya yang diuji pada penelitian ini.
2. Mengetahui karakteristik kinerja algoritma DK-KBSC sehingga dapat digunakan pada studi kasus segmentasi pelanggan.

1.6 Batasan Masalah

Batasan masalah pada penelitian ini adalah :

1. Algoritma yang akan digunakan yaitu algoritma DK dan KBSC.
2. penelitian ini akan diuji menggunakan data model yang dibuat dalam bentuk sebaran 2 lingkaran dan 3 lingkaran *Iris, Wine,* dan *Diabetes.*
3. Algoritma diuji pada data testing studi kasus mengenai segmentasi pelanggan berdasarkan dan jumlah atribut 3 (*Recency, Frequency,* dan *Monetary)* dengan jumlah data 500 yang diambil dari riwayat transaksi dalam satu tahun.

1.7 Manfaat Penelitian

Manfaat yang dapat diperoleh dari penelitian ini adalah sebagai berikut:

1. Bagi sisi keilmuan, penelitian ini dapat menghasilkan algoritma yang dapat menentukan jumlah *cluster* dan pusat *cluster* awal secara otomatis sehingga hasil *cluster* memiliki kualitas hasil *cluster* yang baik berdasarkan nilai DBI.
2. Bagi bidang penelitian di bidang data mining, penelitian ini memberikan pengetahuan mengenai karakteristik algoritma sehingga dapat diimplementasikan pada studi kasus segmentasi pelanggan distributor farmasi.

1.8 Sistematika Penulisan

Sistematika penulisan disusun untuk memberikan gambaran umum tentang penelitian yang dijalankan. Sistematika penulisan tesis ini adalah sebagai berikut:

BAB 1 PENDAHULUAN

Bab ini menguraikan tentang latar belakang masalah, identifikasi masalah, premis, hipotesis, tujuan penelitian, manfaat penelitian serta sistematika penulisan.

BAB 2 TINJAUAN PUSTAKA

Bab ini menguraikan tentang dasar-dasar teori mengenai *clustering* dalam *data mining*, teknik-teknik dalam *data mining* serta evaluasi hasil *clustering*.

BAB 3 METODE PENELITIAN

Bab ini menguraikan tentang penggabungan algoritma Dynamic K-means dan K-means Binary Search Centroid, data yang digunakan dalam penelitian serta alur atau teknik pengujian dan analisis data.

BAB 4 HASIL PENELITIAN

Bab ini menguraikan tentang hasil penelitian yang disajikan dalam bentuk grafik-grafik serta analisis hasil penelitian.

BAB 5 KESIMPULAN DAN SARAN

Bab ini menguraikan tentang kesimpulan dari penelitian yang telah dilakukan serta saran-saran untuk penelitian selanjutnya.