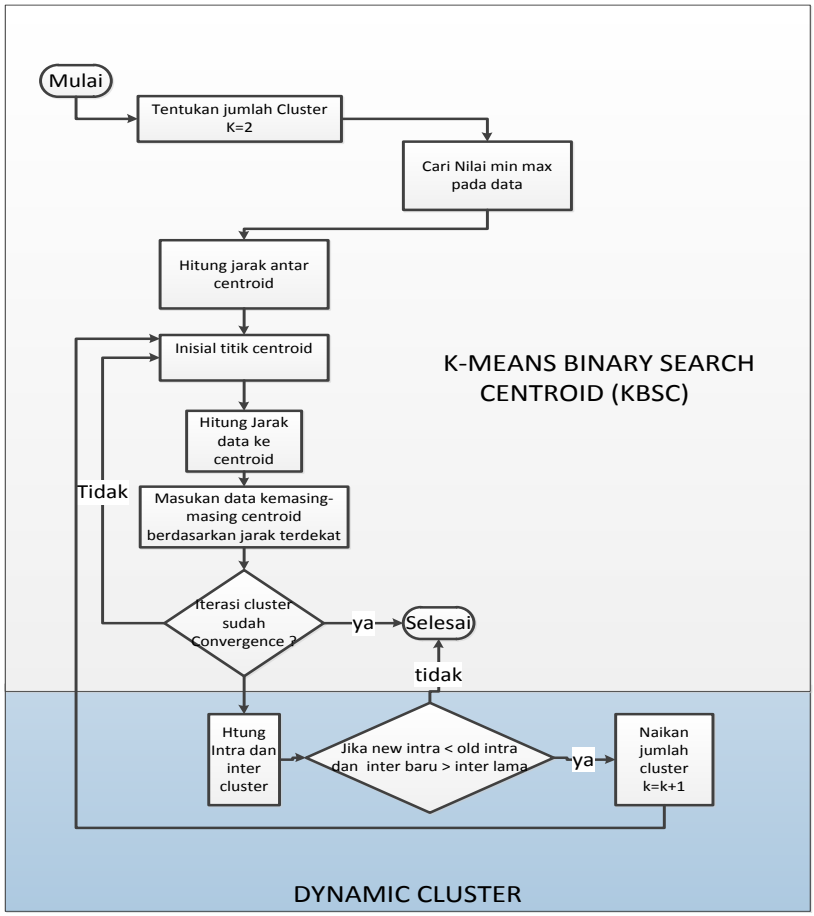
BAB 3
METODE PENELITIAN

3.1 Gabungan Algoritma Dynamic K-means dan KBSC

DK KBSC merupakan gabungan dari algoritma Dynamic K-means dan KBSC yang menggabungkan kelebihan masing-masing algoritma untuk mencari jumlah *cluster* terbaik, dan menentukan titik centroid awal pada setiap pembentukan jumlah *cluster*. Berikut ini gambar mengenai penggabungan kedua algortima :



Gambar 3.1 Alur Penggabungan Kedua Algoritma

Pada Gambar 3.1 menjelaskan alur penggabungan dari kedua algoritma. *Flowchart.* Titik *centroid* ini akan menggunakan teknik pencarian data dengan pendekatan *Binary Search* (Pencarian bagi dua terhadap data). Proses tersebut akan membagi area data berdasarkan jumlah *cluster* yang dimasukan. Berikut ini penjelasan pada alur *flowchart* :

1. Pada awalnya dilakukan penentuan jumlah cl*uster (k)* minimal 2,
2. Setelah itu proses dilanjutkan ke proses penentuan titik *centroid* menggunakan pendekatan *binary search centroid*
3. Kemudian lakukan pengelompokan dengan cara menghitung jarak *minimum* antar data dengan *titik centroid* menggunakan *Euclidiance Distance* sehingga tidak ada lagi titik data yang berpindah *cluster*. Di rumus 2.1 menjelaskan perhitungan jarak *Euclidiance Distance*
4. Setelah *cluster* tidak berubah, tahap selanjutnya masuk pada tahap proses *Dynamic K-means* yang mana prosesnya menghitung nilai *inter* dan nilai *intra*. Dan akan ditambahkan jumlah k=k+1 jika nilai *intra* baru < *intra* lama, dan *inter* baru > dari *inter* lama. Dan akan kembali ke proses penentuan titik *centroid* kembali berdasarkan *cluster* yang terbentuk sebelumnya

3.2 Macam-macam Data Penelitian

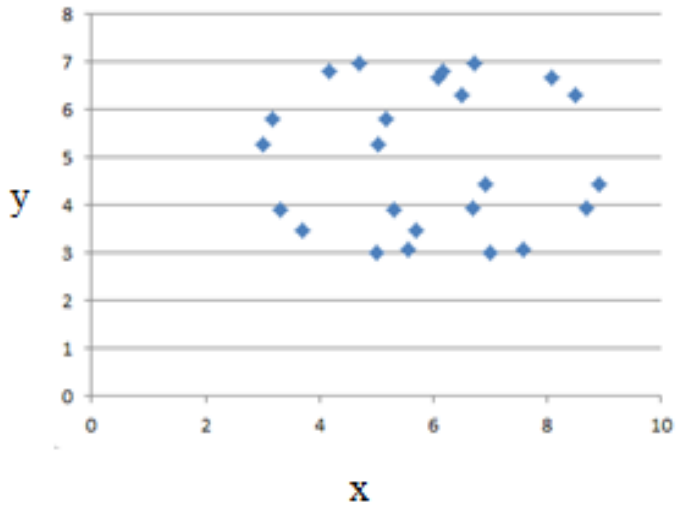
Bentuk sebaran data yang akan diuji adalah data model yang dibuat, dan yang sudah digunakan oleh penelitian pada area *clustering*. Untuk data testing merupakan data studi kasus segmentasi pelanggan.

3.2.1 Data Model

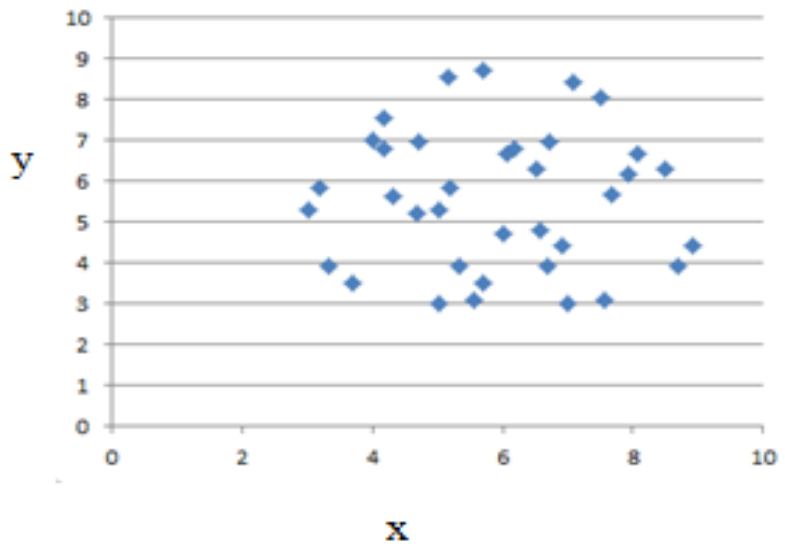
Data model uji merupakan data yang dibuat dengan bentuk yang memiliki sebaran dengan nilai kemiripan intra dan inter yang telah diatur, dan bentuk datanya dibuat dalam beberapa sebaran bentuk *cluster*. Sebaran bentuk *cluster* tersebut dilihat berdasarkan jarak intra dan inter. Pada penelitian ini, sebaran untuk data model akan dibuat sebagai berikut :

1. Sebaran Data Bentuk Lingkaran

Sebaran data bentuk lingkaran memiliki variasi nilai *intra* = 2 dan *inter = 2* , dan jumlah variasi bentuk *cluster* yang berbeda mulai dari 2 *cluster* sampai 3 *cluster* yang akan ditetapkan. Berikut ini data model yang akan diuji pada penelitian ini :



Gambar 3.2 Bentuk Sebaran Data 2 Lingkaran 2 Variabel (x,y)



Gambar 3.3 Bentuk Sebaran Data 2 *Cluster* 2 Variabel (x,y)

Bentuk sebaran data yatu kombinasi dua lingkaran dekat, dan tiga lingkaran dekat. Kedekatan sebaran data ini diatur dengan jarak *intra cluster* dan *inter cluster* dengan jarak intra =2 dan inter = 2. Jika jarak *inter cluster* semakin besar maka kedua lingkaran tersebut akan semakin berjauhan, yang menandakan bahwa bentuk sebaran sudah dipastikan berjumlah 2 *cluster*, namun pada Gambar 3.2 dan Gambar 3.3 sebaran kedua lingkaran sangat berdekatan sehingga kita tidak dapat menilai berapa *cluster* yang dapat dibentuk.

Tabel 3.1 Data Model Sebaran Bentuk Lingkaran

Sebaran *Cluster*	Jarak Intra / inter *Cluster*	Jumlah Variabel	Jumlah Data
2 lingkaran	2/2	<=13	<=768
3 lingkaran	2/2	<13	<=1152

Pada penelitian ini hanya akan berfokus pada data yang jarak intra dan inter *cluster*nya berdekatan, karena jika intra dan inter nya berdekatan belum dapat diketahui jumlah *cluster* yang dapat terbentuk seharusnya.

2. Data Model Dari *Machine Learning* Dataset

Data yang akan digunakan dalam penelitian adalah *iris, diabetes dan wine*. Untuk proses *clustering* menggunakan penggabungan algoritma Dyanamic K-means dan KBSC, pada ketiga dataset yang dicoba ini memiliki kesamaan pada tipe datanya adalah *numeric.* Data testing diperoleh dari UCI *repository of machine learning databases*, yang disediakan oleh Nationan Science Foundation yang berisi kumpulan *dataset* untuk keperluan penelitian di bidang *data mining.*

Data testing ini dipilih karena memiliki tipe data yang sesuai untuk proses *Clustering,* dan memiliki keragaman pada atribut dan jumah datanya yang akan dijadikan acuan pada proses *cluster*. Pertimbangan dalam memilih data testing, karena telah banyak digunkan dan diuji oleh beberapa penelitian sebelumnya mengenai kasus *clustering*. Berikut ini tabel mengenai karakteristik dari masing-masing dataset yang akan digunakan :

Tabel 3.2 Karakteristik Dari Data Set Yang Akan Digunakan

Data Set	Variabel (Atribut Data)	Jumlah Data	TIpe Data
Iris	4	<= 768	*Numeric*
Diabetes	8	<= 1152	
Wine	13	<= 768	

Penjelasan dari Tabel 3.2 sebagai berikut ini :

a. *Dataset* Iris

Dataset Iris berisi pengelompokan bunga iris berdasarkan empat atribut yaitu *sepal length*, *sepal width*, *petal length* dan *petal width* ke dalam 3 *Cluster* yaitu iris setosa, iris versicolour dan iris virginica.

b. *Dataset* Diabetes

Dataset Diabetes berisi pengelompokan pasien wanita berdasarkan 8 atribut yaitu *number of times pregnant, plasma glucose concentration, diastolic blood pressure, triceps skin fold thickness,* 2*-hour serum insulin, body mass index, diabetes pedigree function*, dan *age* ke dalam 2 *Cluster* yaitu 0 yang merepresentasikan bahwa pasien tidak memiliki kemungkinan terkena diabetes dan 1 yang merepresentasikan bahwa pasien memiliki kemungkinan terkena diabetes.

c. *Dataset* Wine

Dataset Wine berisi pengelompokan anggur berdasarkan 13 atribut yaitu *alcohol, malic acid, ash, alcalinity of ash, magnesium, total phenols, flavanoids, Nonflavanoid phenols, proanthocyanins, color intensity, hue,* dan *proline* ke dalam 3 *Cluster* yaitu 1, 2 dan 3.

3.2.2 Data Uji Studi Kasus

Data uji studi kasus memiliki karakteristik data dengan jumlah atribut yaitu 3 atribut, dimana ketiga atribut tersebut masing-masing memiliki nilai numeric.Jumlah data yang akan diuji yaitu 500 pelanggan dengan nilai atibut masing-masing yaitu *Recency*, *Frequency* dan *Monetary*.

Pada kasus ini ingin mengetahui segmentasi pelanggan dengan mengacu pada nilai RFM yang didapat dari data historis transaksi pelanggan, sehingga pihak perusaahaan khususnya manager pemasaran dapat membuat layanan pemasaran yang berbeda agar sesuai dengan kebutuhan dan keinginan setiap segmen pelanggan.

Berdasarkan dari data historis transaksi, akan diambil transaksi selama satu tahun dimulai tanggal 1 januari 2016 sampai 1 Januari 2017, kemudian diambil nilai RFM pelanggan tersebut. Jumlah sampel data pelanggan yang akan diproses *cluster* sebanyak 500 pelanggan. Di Tabel 19 merupakan data yang akan di proses *cluster* untuk menghasilkan segmentasi pelanggan.

Tabel 3.3 Sampel Data RFM Pelanggan

Id_pelanggan	Recency	Frequency	Monetary
1	56	20	10235900
2	84	20	15820618
3	75	15	11073303
4	92	16	15681458
5	70	20	14332580
6	72	18	16714173
7	82	19	17513733
8	60	17	16418762
9	95	19	18116761
10	96	18	12295227

Dalam RFM, Recency menjelaskan mengenai resensi yang , yang berarti adalah jarak antara waktu terakhir pemakaian atau pembelian dengan waktu sekarang, sehingga dapat mengetahui pelanggan yang lama maupun pelanggan yang baru, F merupakan Frequency yang merupakan jumlah transaksi pada periode waktu tertentu. M merupakan Monetary yang artinya jumlah uang yang digunakan untuk pembelian pada periode waktu tertentu.

3.3 Pertimbangan Dalam Pemilihan Data

Data yang dipilih untuk diuji pada proses *cluster* dipertimbangkan pada hal tertentu, bertujuan agar dapat mengetahui karakteristik kinerja algoritma. Berikut ini hal yang dipertimbangkan sebagai berikut :

1. Pertimbangan mengenai tipe data yang digunakan

   Tipe data yang digunakan adalah numeric, untuk proses *clustering* sangat cocok digunakan untuk data bersifat *numeric*, jika terdapat data yang bukan *numeric* maka dijadikan suatu range nilai berupa angka.

2. Pertimbangan Mengenai Jumlah Variabel

   Variabel dalam data uji merupakan atribut data, Jumlah variabel tersebut untuk mengetahui apakah data yang digunakan untuk diselesaikan dengan metode *Clustering*, semakin banyak dimensi atau variabel yang digunakan maka akan data tersebut semakin semakin beragam dalam pengelompokannya.

3. Pertimbangan Jumah record

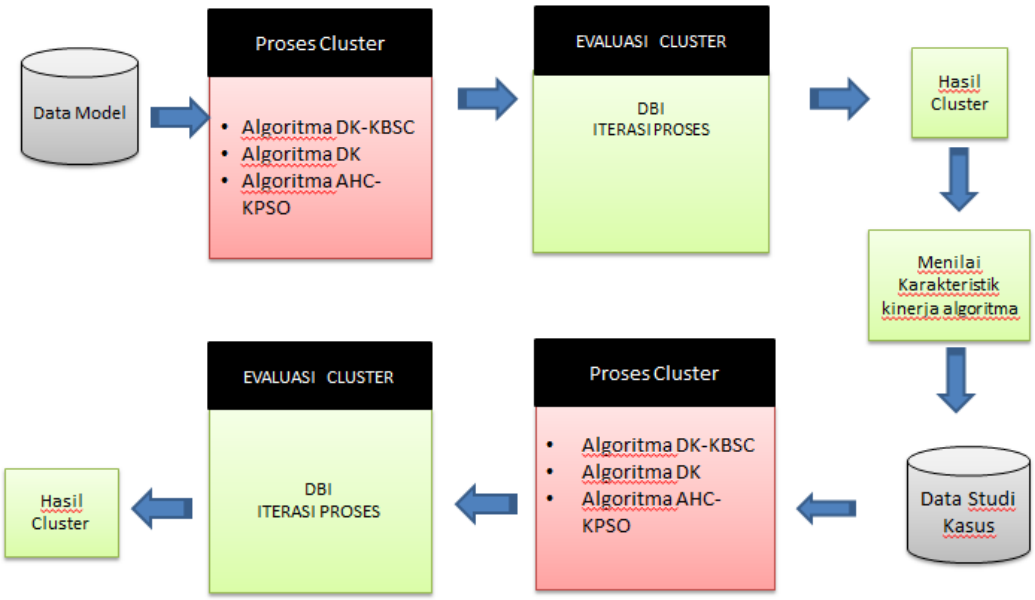
   Pertimbangan inin dilakukan agar dapat mengetahui bagaimana algoritma yang akan diuji dapat digunakan pada beragam jumlah data yang berbeda.

3.4 Langkah Penelitian

Implementasi algoritma akan dilakukan dengan cara menguji algoritma terhadap data, pengujuannya sebagai berikut :

1. Pengujian terhadap data model berdasarkan bentuk sebaran dengan pertimbangan pengujian sebagai berikut :
    a. Pengujian *cluster* berdasarkan jumlah data
    b. Pengujian *cluster* berdasarkan jumlah variabel.
2. Pengujian terhadap dataset uji, dengan pertimbangan pengujiannya adalah pengujian berdasarkan tingkatan jumlah data.

Hasil pengujian digunakan suatu penilaian secara *kuantitatif*, yang mana penilaian tersebut merupakan hasil perbandingan antara gabungan algoritma yang diajukan dengan algoritma lain, dengan tujuan agar dapat mengetahui kemampuan algoritma dalam menyelesaikan kasus *clustering* pada bentuk data tertentu. Berikut ini alur pengujian terhadap penggabungan algoritma pada data yang digunakan :



Gambar 3.4 Tahapan Pengujian dan Proses Perbandingan Kinerja Algoritma

Penjelasan berdasarkan Gambar 3.4 adalah setelah data diproses oleh algoritma yang diusulkan, maka akan muncul suatu nilai evaluasi *Clustering* yang digambarkan menggunakan grafik perbandingan. Sehingga dapat mengetahui kelebihan maupun kekurangan algoritma terhadap data yang diuji.