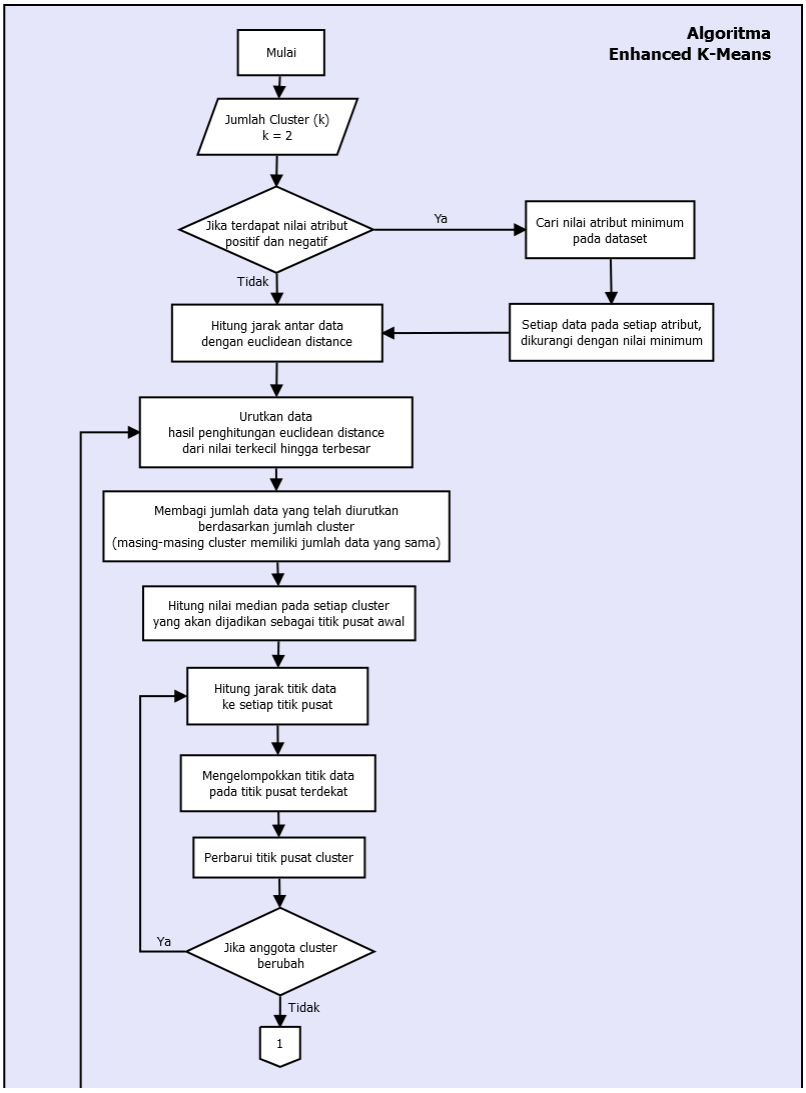
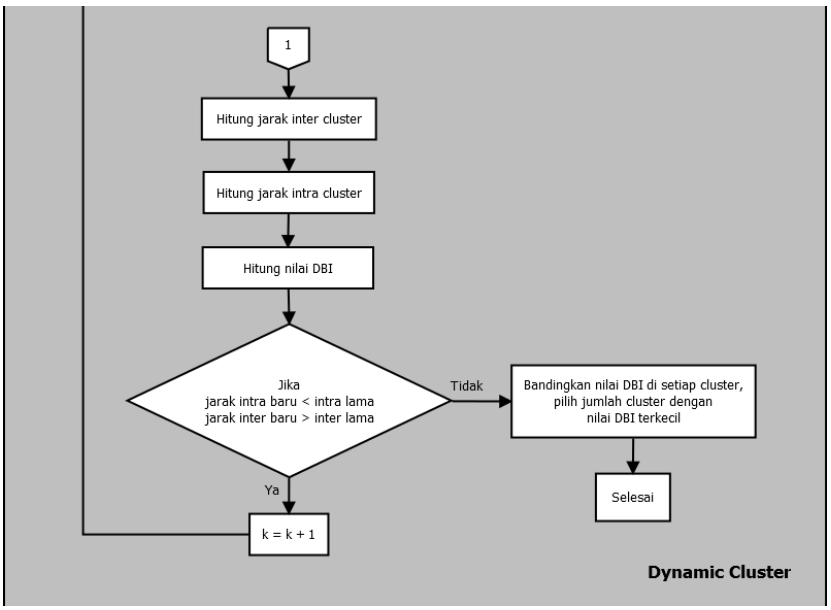
BAB 3
METODE PENELITIAN

3.1 Enhanced K-Means – Dynamic Cluster

Algoritma Enhanced K-Means – Dynamic Cluster (DEKM) adalah penggabungan antara algoritma Dynamic Cluster dengan Algoritma Enhanced K-Means, penggabungan algoritma tersebut bertujuan untuk menggabungkan kelebihan dari masing-masing algoritma, sehingga dapat menentukan jumlah *cluster* dan titik pusat terbaik. Berikut ini proses penggabungan kedua algoritma yang diperlihatkan pada Gambar 3.1.





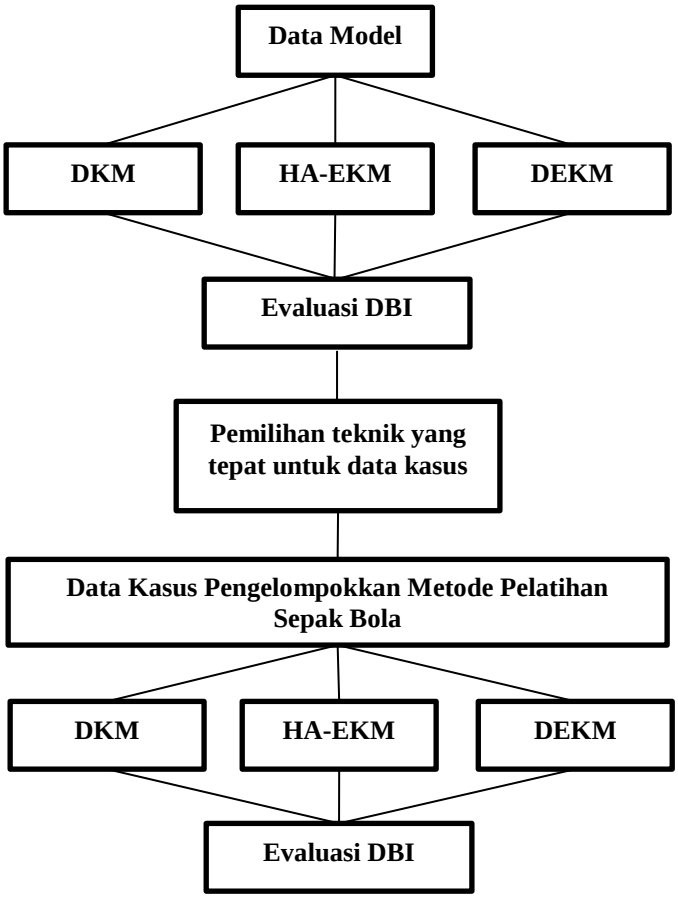
Gambar 3.1 Algoritma Enhanced K-Means dan Dynamic *Cluster*

Berdasarkan Gambar 3.1 terlihat bahwa proses *clustering* diawali dengan menduga jumlah *cluster* sebanyak 2 buah, kemudian dilanjutkan dengan proses pada algoritma Enhanced K-Means hingga titik data tidak berpindah *cluster*. Kemudian hasil *clustering* yang telah terbentuk pada pemrosesan algoritma Enhanced K-Means akan diproses oleh algoritma Dynamic Cluster yang akan menghitung jarak inter, intra dan nilai DBI, setelah mendapatkan hasil jarak intra dan inter maka akan dilakukan seleksi dengan ketentuan jika jarak intra baru lebih kecil dari jarak intra lama dan jarak inter baru lebih besar dari jarak inter lama maka jumlah cluster akan ditambah 1 dan proses clustering diulang kembali, sedangkan jika kondisi tidak sesuai maka akan dilakukan pembandingan nilai DBI dari setiap cluster, kemudian memilih jumlah cluster dengan nilai DBI terkecil sebagai jumlah cluster terbaik.

3.2 Teknik Pengujian

Pada penelitian ini, pengujian dibagi menjadi dua, yaitu pengujian terhadap data model dan data testing, kemudian pada setiap pengujian akan diimplementasikan pada algoritma Enhanced K-Means – Dynamic Cluster (DEKM) dan dua algoritma pembanding yaitu Dynamic Cluster – K-Means (DKM) dan Hierarchical Agglomerative Enhanced K-Means (HA-EKM).

Algoritma pembanding DKM dipilih berdasarkan penelitian sebelumnya yang dilakukan oleh Madu dkk (Madu, Srinivasa, & Pathakota, 2010) yang membandingkan algoritma usulan dengan algoritma pada penelitian sebelumnya. Sedangkan algoritma HA-EKM dipilih karena memiliki kesamaan dengan algoritma DEKM yang dapat menentukan titik pusat awal dan menduga jumlah cluster terbaik. Alur teknik pengujian dapat dilihat pada Gambar 3.2.



Gambar 3.2 Teknik Pengujian

Pada Gambar 3.2, data model dan data kasus akan diolah oleh algoritma DKM, HA-EKM dan DEKM, kemudian hasil *clustering* dari ketiga algoritma di evaluasi oleh DBI hingga menghasilkan nilai DBI pada masing-masing algoritma.

3.2.1 Pemilihan Data Penelitian

Data yang akan digunakan pada penelitian ini terbagi menjadi dua, yaitu data model dan data kasus.

3.2.1.1 Data Model

Data model merupakan kumpulan data yang memiliki keanekaragaman karakteristik dan digunakan untuk mengetahui karakteristik dari algoritma Enhanced K-Means – Dynamic Cluster (DEKM). Data yang digunakan pada penelitian ini yaitu data model buatan dan data UCI.

1. Data Model Buatan

Data model buatan adalah data yang sengaja dibuat untuk mengetahui karakteristik dari algoritma, karakteristik data pada data model buatan telah dimanipulasi jumlah data, jumlah *cluster*, jumlah atribut serta sebaran datanya.

Berikut ini tabel mengenai keragaman data model buatan yang telah dibuat:

Tabel 3.1 Karakteristik Data Model Buatan

Dataset	**Jumlah Data**	**Jumlah Atribut**	**Keterangan**
2 *Cluster*	24, 48, 96, 192, 384 dan 768	2 – 13	Dataset model 2 cluster dengan beragam atribut dan jumlah data
3 *Cluster*	36, 72, 144, 288, 576 dan 1152	2 - 13	Dataset model 3 cluster dengan beragam atribut dan jumlah data

Kemudian sebaran data pada data model buatan dimanipulasi dengan cara menjauhkan jarak antar *cluster*.

2. Data UCI

Data UCI pada penelitian ini adalah kumpulan *dataset* yang diperoleh dari situs UCI. Data UCI merupakan data yang umum digunakan oleh para peneliti di bidang data mining untuk menguji algoritma dalam menyelesaikan kasus *clustering*. Data UCI yang digunakan dalam penelitian ini yaitu iris, wine dan diabetes, ketiga dataset tersebut dipilih karena memiliki karakteristik yang berbeda. Berikut ini karakteristik dari data UCI.

Tabel 3.2 Karakteristik Data UCI

Dataset	Jumlah Data	Jumlah Atribut	Keterangan
Iris	150	4	Dataset mengenai pengelompokkan bunga iris berdasarkan ukuran sepal dan petal.
Wine	178	13	Dataset mengenai penentuan tipe wine berdasarkan kriteria tertentu
Diabetes	768	8	Dataset mengenai kemungkinan seseorang terkena penyakit diabetes berdasarkan kriteria tertentu.

Berikut ini atribut dari masing-masing *dataset* :

a. *Dataset* Iris

Dataset Iris merupakan *dataset* mengenai pengelompokkan bunga iris berdasarkan 4 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* Wine

Dataset Wine merupakan *dataset* mengenai pengelompokkan wine 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.

c. *Dataset* Diabetes

Dataset Diabetes merupakan *dataset* mengenai kemungkinan seseorang terkena penyakit diabetes 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.

3.2.1.2 Data Kasus

Data kasus yang digunakan pada penelitian ini adalah data pemain bola yang didapatkan dari Kaggle. Untuk mengetahui pengelompokkan metode pelatihan yang tepat bagi pemain maka atribut yang akan digunakan adalah *ball control,*

dribbling, heading, shooting, passing, composure, speed, stamina, strength, balance dan *agility*. Berdasarkan atribut tersebut diketahui metode pelatihan fisik dari atribut *speed, stamina, strength, balance* dan *agility*. Kemudian metode pelatihan skill dari atribut *ball control, dribbling, heading, shooting, passing*. Sedangkan metode pelatihan mental dari atribut *composure*. Data tersebut memiliki jumlah data sebanyak 438 dan atribut sebanyak 11.

Pengelompokkan data berdasarkan metode pelatihan sepak bola tersebut sangat tepat diimplementasikan pada algoritma Enhanced K-Means – Dynamic Cluster (DEKM) yang memilik karakteristik serupa. Model data pemain bola dapat dilihat di Tabel 3.3.

Tabel 3.3 Model Data Pemain Binaan

Id	Ball Control	Drib	Head	Shoot	Pass	Comp	Speed	Sta	Str	Bal	Agi
1	58	66	35	52	43	47	87	62	34	74	78
2	52	60	47	27	45	43	69	59	61	65	66
3	50	47	68	36	63	53	69	72	68	64	46
4	74	77	50	67	57	50	88	55	67	62	74
5	72	71	66	60	74	76	75	69	75	72	75

Tabel 3.3 memperlihatkan mengenai model data pemain binaan yang terdiri dari 11, masing-masing atribut memperlihatkan rating kemampuan dari setiap pemain binaan. Data pemain binaan tersebut bertipe numerik sehingga dapat langsung diolah.

3.2.2 Alur Pengujian

Alur pengujian yang akan dilakukan pada penelitian ini terbagi menjadi dua yaitu pengujian terhadap data model dan data kasus.

1. Pengujian Data Model

Pengujian data model dilakukan untuk mengetahui karakteristik dari algoritma Enhanced K-Means - Dynamic Cluster (DEKM) terhadap data, sehingga dapat diketahui kasus seperti apa yang tepat untuk diimplementasikan pada algoritma DEKM. Pada pengujian data model, algoritma DEKM, DKM dan

HA-EKM akan diimplementasikan pada seluruh data model, kemudian hasil *cluster* yang terbentuk dari setiap algoritma akan dievaluasi oleh Davies-Bouldin Index (DBI), selanjutnya akan dibandingkan kinerja setiap algoritma berdasarkan nilai DBI dan iterasi.

2. Pengujian Data Kasus

Pengujian data kasus dilakukan untuk membuktikan hasil implementasi algoritma pada data model terhadap kasus nyata. Pengujian data kasus dilakukan dengan cara yang sama dengan pengujian data model, yaitu setiap algoritma akan melakukan proses *clustering* pada data kasus, kemudian hasil *cluster* yang terbentuk dari setiap algoritma akan dievaluasi oleh DBI, dan akan dilakukan perbandingan hasil nilai DBI juga iterasi dari setiap algoritma.