# BAB I

**PENDAHULUAN**

* 1. **Latar Belakang Masalah**

Teknik *Clustering* merupakan salah satu metode Data Mining yang bersifat tanpa arahan (*unsupervised learning*) dimana tidak perlu ada fase pelatihan(*learning*) pada metode tersebut. Tujuan utama dari metode *cluster* adalah pengelompokan sejumlah data/objek ke dalam kelompok – kelompok data sehingga dalam setiap kelompok akan berisi data yang semirip mungkin (Budi Santoso, 2007). Ada dua pendekatan dalam *clustering* yang sering digunakan dalam proses pengelompokan data yaitu *hierarchical* (hirarki) data *clustering* dan *partitioning* (non hirarki) data *clustering*. K-Means merupakan salah satu metode data *clustering* non hirarki yang paling sederhana. Menurut yudi Agusta (2007) Metode ini mempartisi data ke dalam *cluster* sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu *cluster* yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam kelompok yang lain dengan cara menghitung jarak terdekat data dengan titik pusat data/*centroid* yaitu dengan rumus *distance space*.

Beberapa *distance space* telah diimplementasikan dalam menghitung jarak (*distance*) antara data dan *centroid* termasuk di antaranya adalah L1 (*Manhattan / City Block*) *distance space*, L2 (*Euclidean*) *distance space*, dan Lp (*Minkowski*) *distance space*. Tetapi secara umum *distance space* yang sering digunakan adalah *Manhattan/City Block* dan *Euclidean*. *Euclidean* sering digunakan karena penghitungan jarak dalam *distance space* ini merupakan jarak terpendek yang bisa didapatkan antara dua titik yang diperhitungkan, sedangkan  *Manhattan(City Block)* sering digunakan karena kemampuannya dalam mendeteksi keadaan khusus seperti keberadaaan *outliers* dengan lebih baik (Yudi Agusta, 2007).

Pada penelitian ini, akan membandingan dua *distance space* yaitu antara *Manhattan(City Block)* dan *Euclidean* dengan objek penelitian yaitu data anak balita usia dibawah lima tahun. Dari dua rumus tersebut akan dipelajari hasil pengelompokan data yang terbentuk.

* 1. **Rumusan Masalah**

Dari uraian diatas, beberapa masalah kemudian dirumuskan sebagai berikut :

* 1. Metode K-Means mengenal beberapa rumus *distance space* yaitu *Manhattan(City Block)*, *Euclidean* dan *Minkowski metric*. Pada dasarnya belum ada yang memastikan mana yang lebih baik dari beberapa rumus tersebut sehingga perlu adanya penelitian.
  2. Hasil cluster berupa visualisasi data pada grafik, merupakan wujud dari penyajian data agar lebih mudah dianalisa dan digali informasi secara lebih mendalam.
  3. **Ruang Lingkup**

Ruang lingkup penelitian ini mencakup beberapa hal sebagai berikut :

1. Perbandingan antara dua rumus penghitungan distance space pada K-Means yaitu *Manhattan(City Block)* dengan *Euclidean* dari sisi hasil akhir cluster dan pola pengelompokan data serta informasi jumlah iterasi yang dihasilkan oleh titik awal yang di generate secara random pada masing – masing rumus dan waktu yang dibutuhkan untuk iterasi. Jumlah cluster yang diteliti adalah 4 cluster yang menggambarkan status gizi balita
2. Data yang digunakan untuk digali informasinya adalah data balita di wilayah puskesmas mlati sleman. Data akan dipilih sehingga dapat mewakili dari kumpulan data – data balita di wilayah tersebut.
   1. **Tujuan**

Adapun tujuan dari dibuatnya penelitian ini yaitu :

1. Mengetahui hasil clustering data menggunakan metode K-Means dengan rumus distance space antara *Manhattan(City Block)* dan *Euclidean*. Sehingga pada hasil akhir *cluster* diharapkan akan terdapat perbedaan pola antara kedua *distance space* tersebut dalam melakukan pengelompokan data sehingga dapat dipelajari pola pengelompokan data untuk masing-masing rumus *distance space*.
2. Menggali informasi data dari objek penelitian data balita di wilayah Mlati Sleman.