

## **BAB III**

### **LANDASAN TEORI**

#### **3.1. Puskesmas**

##### **3.1.1. Pengertian Puskesmas**

Berdasarkan Peraturan Menteri Kesehatan Republik Indonesia Nomor 75 Tahun 2014 tentang Puskesmas, Puskesmas adalah fasilitas pelayanan kesehatan yang menyelenggarakan upaya kesehatan masyarakat dan upaya kesehatan perseorangan tingkat pertama, dengan lebih mengutamakan upaya promotif dan preventif, untuk mencapai derajat kesehatan masyarakat yang setinggi-tingginya di wilayah kerjanya.

##### **3.1.2. SDM Puskesmas**

SDM Puskesmas terdiri atas tenaga kesehatan dan non kesehatan. Jenis dan jumlah tenaga kesehatan dan tenaga non kesehatan dihitung berdasarkan analisis beban kerja, dengan mempertimbangkan jumlah pelayanan yang diselenggarakan, jumlah penduduk dan persebarannya, karakteristik wilayah kerja, luas wilayah kerja, ketersediaan fasilitas pelayanan kesehatan tingkat pertama lainnya di wilayah kerja, dan pembagian waktu kerja (Kementerian Kesehatan Republik Indonesia, 2014).

Tenaga kesehatan adalah setiap orang yang mengabdikan diri dalam bidang kesehatan serta memiliki pengetahuan dan atau keterampilan melalui pendidikan di bidang kesehatan yang untuk jenis tertentu memerlukan kewenangan untuk melakukan upaya kesehatan. Jenis tenaga kesehatan yang dimaksud adalah (Kementerian Kesehatan Republik Indonesia, 2014):

- a. Dokter atau dokter layanan primer
- b. Dokter gigi
- c. Perawat
- d. Bidan
- e. Tenaga kesehatan masyarakat

- f. Tenaga kesehatan lingkungan
- g. Ahli teknologi laboratorium medik
- h. Tenaga gizi
- i. Tenaga kefarmasian

Sedangkan tenaga non kesehatan, yang kemudian disebut dengan tenaga penunjang kesehatan, adalah setiap orang yang mengabdikan diri dalam bidang kesehatan tetapi tidak melakukan upaya kesehatan. Tenaga non kesehatan meliputi Pejabat Struktural Eselon 1, Pejabat Struktural Eselon 2, Pejabat Struktural Eselon 3, Pejabat Struktural Eselon 4, Staf Pencatatan dan Pelaporan, Staf Administrasi Keuangan, Staf Teknis Program Kesehatan, Staf Pengelola TIK (Teknologi Informasi), Staf Kehumasan, Staf Bidang Hukum, Staf Perencana, Staf Pegawai Kesehatan, Staf Pespustakaan, Staf Psikolog, Staf Dosen Pengajar, Staf Widyaiswara, Staf Juru Mudi, serta tenaga lainnya yang tidak tercantum (Kementerian Kesehatan Republik Indonesia, 2016)

### **3.1.3. Sarana dan Prasarana Puskesmas**

Beberapa sarana prasarana yang dimiliki oleh Puskesmas berdasarkan Data Dasar Puskesmas Tahun 2015 adalah sebagai berikut.

- a. Rumah medis, yang secara fisik dibagi menjadi tiga kondisi yaitu:
  - 1. Baik; apabila bangunan rumah medis dalam kondisi baik atau tidak mengalami kerusakan.
  - 2. Rusak ringan; apabila bangunan rumah medis terjadi kerusakan pada komponen pintu, jendela, kaca, penggantung, pengunci, cat, dan sebagainya.
  - 3. Rusak berat; apabila bangunan rumah medis terjadi kerusakan pada komponen pokok dari bangunan seperti pilar, pondasi, *sloope*, *ring balk*.
- b. Tempat tidur, dimiliki oleh Puskesmas rawat inap.
- c. Sumber air, dibagi berdasarkan:
  - 1. PAM
  - 2. Sumur
  - 3. Lainnya.
- d. Sumber listrik, dibagi berdasarkan:

1. PLN
  2. Generator
  3. Lainnya.
- e. Jalan menuju puskesmas, dibagi menjadi beberapa kondisi berdasarkan:
1. Aspal
  2. Tanah
  3. Lainnya.
- f. *Ambulance*, merupakan kendaraan transportasi gawat darurat medis khusus orang sakit atau cedera, dari satu tempat ke tempat lain guna perawatan medis. Kondisi *ambulance* yang dimiliki Puskesmas dibagi menjadi empat, yaitu:
1. Baik; apabila *ambulance* dalam kondisi baik dan dimanfaatkan sesuai dengan peruntukannya.
  2. Rusak ringan; apabila *ambulance* dalam kondisi baik namun tidak dimanfaatkan sesuai dengan peruntukannya.
  3. Rusak sedang; apabila *ambulance* dalam kondisi tidak baik namun masih bisa dimanfaatkan sesuai dengan peruntukannya.
  4. Rusak berat; apabila *ambulance* dalam kondisi tidak baik dan tidak dapat difungsikan atau tidak dapat dimanfaatkan.
- g. Sepeda motor, dibagi menjadi empat kondisi yaitu:
1. Baik; apabila sepeda motor dalam kondisi baik dan dimanfaatkan sesuai dengan peruntukannya.
  2. Rusak ringan; apabila sepeda motor dalam kondisi baik namun tidak dimanfaatkan sesuai dengan peruntukannya.
  3. Rusak sedang; apabila sepeda motor dalam kondisi tidak baik namun masih bisa dimanfaatkan sesuai dengan peruntukannya.
  4. Rusak berat; apabila sepeda motor dalam kondisi tidak baik dan tidak dapat difungsikan atau tidak dapat dimanfaatkan.
- h. Puskesmas keliling roda empat (Pusling roda empat), merupakan unit pelayanan kesehatan kepada masyarakat di daerah terpencil berupa kendaraan bermotor roda empat dan peralatan kesehatan, komunikasi, serta seperangkat tenaga yang berasal dari Puskesmas. Pusling roda empat berfungsi untuk

menunjang dan membantu melaksanakan kegiatan-kegiatan Puskesmas dalam wilayah kerja yang belum terjangkau oleh pelayanan kesehatan karena letaknya jauh dan terpencil. Pusling roda empat dibedakan menjadi empat kondisi, yaitu:

1. Baik; apabila Pusling roda empat tersebut dalam kondisi baik dan dapat dimanfaatkan sesuai dengan peruntukannya.
  2. Rusak ringan; apabila Pusling roda empat tersebut dalam kondisi tidak baik namun masih bisa dimanfaatkan sesuai dengan peruntukannya
  3. Rusak sedang; apabila Pusling roda empat tersebut dalam kondisi rusak dan masih dapat difungsikan jika ada beberapa komponennya diganti/diperbaiki.
  4. Rusak berat; apabila Pusling roda empat tersebut dalam kondisi tidak baik dan tidak dapat difungsikan atau tidak dapat dimanfaatkan.
- i. Puskesmas keliling perairan (Perahu), merupakan unit pelayanan kesehatan kepada masyarakat di daerah terpencil berupa perahu/perahu bermotor dan peralatan kesehatan, komunikasi serta seperangkat tenaga yang berasal dari Puskesmas. Pusling ini berfungsi menunjang dan membantu melaksanakan kegiatan-kegiatan Puskesmas dalam wilayah kerjanya yang belum terjangkau oleh pelayanan kesehatan karena letaknya jauh dan terpencil. Pusling perairan (perahu) dibagi menjadi beberapa kondisi, yaitu:
1. Baik; apabila Pusling perairan tersebut dalam kondisi baik dan dimanfaatkan sesuai dengan peruntukannya
  2. Rusak ringan; apabila Pusling perairan tersebut dalam kondisi baik namun tidak dimanfaatkan sesuai dengan peruntukannya
  3. Rusak sedang; apabila Pusling perairan tersebut dalam kondisi tidak baik namun masih bisa dimanfaatkan sesuai dengan peruntukannya
  4. Rusak berat; apabila Pusling perairan tersebut dalam kondisi tidak baik dan tidak dapat difungsikan atau tidak dapat dimanfaatkan.
- j. Puskesmas pembantu (Pustu), merupakan unit pelayanan kesehatan yang sederhana dan berfungsi menunjang dan membantu memperluas jangkauan Puskesmas dengan melaksanakan kegiatan-kegiatan yang dilakukan Puskesmas dalam ruang lingkup wilayah yang lebih kecil serta jenis dari

kompetensi pelayanan yang disesuaikan dengan kemampuan tenaga dan sarana yang tersedia. Pustu yang dimiliki oleh Puskesmas dibedakan menjadi beberapa kondisi, yaitu:

1. Baik; apabila bangunan Pustu yang bersangkutan dalam kondisi baik atau tidak mengalami kerusakan.
  2. Rusak ringan; apabila bangunan Pustu yang bersangkutan terjadi kerusakan pada komponen pintu, jendela, kaca, penggantung, pengunci, cat, dan sebagainya.
  3. Rusak sedang; apabila bangunan Pustu yang bersangkutan terjadi kerusakan pada komponen pokok dari bangunan seperti pilar, pondasi, *sloope*, *ring balk*.
  4. Rusak berat; apabila bangunan Pustu yang bersangkutan sudah tidak dapat digunakan/dimanfaatkan lagi.
- k. Pos bersalin desa (Polindes), merupakan bangunan yang dibangun dengan bantuan dana pemerintah dan partisipasi masyarakat desa untuk tempat pertolongan persalinan dan pemondokan ibu bersalin sekaligus tempat pertolongan persalinan dan pemondokan ibu bersalin sekaligus tempat tinggal bidan di desa. Di samping pertolongan persalinan juga dilakukan pelayanan antenatal dan pelayanan kesehatan lain sesuai kebutuhan masyarakat dan kompetensi teknis bidan tersebut.
- l. Pos kesehatan desa (Poskesdes), merupakan wujud upaya kesehatan bersumber daya masyarakat yang dibentuk oleh, untuk dan bersama masyarakat setempat atas dasar musyawarah, dengan bantuan dari tenaga profesional kesehatan dan dukungan sektor terkait termasuk swasta dalam kerangka desa siaga demi terwujudnya desa sehat.
- m. Poskestren, merupakan salah satu wujud upaya kesehatan bersumber daya masyarakat di lingkungan pondok pesantren dengan prinsip dari, oleh, dan untuk warga pondok pesantren, yang mengutamakan pelayanan promotif (peningkatan kesehatan) dan preventif (pencegahan) tanpa mengabaikan aspek kuratif (pengobatan) dan rehabilitatif (pemulihan kesehatan) dengan binaan Puskesmas setempat.

- n. Desa siaga, merupakan desa yang penduduknya memiliki kesiapan sumber daya dan kemampuan serta kemauan untuk mencegah dan mengatasi masalah-masalah kesehatan, bencana, dan kegawatdaruratan kesehatan secara mandiri menuju desa sehat.
- o. Pos pelayanan terpadu (Posyandu), merupakan salah satu wadah peran serta masyarakat yang dikelola dan diselenggarakan dari, oleh, untuk, dan diselenggarakan dari, oleh, untuk, dan bersama masyarakat guna memperoleh pelayanan kesehatan dasar dan memantau pertumbuhan balita dalam rangka meningkatkan kualitas sumber daya manusia secara dini. Posyandu dibedakan atas:
  - 1. Posyandu pratama, adalah posyandu yang masih belum mantap, kegiatannya belum bisa rutin setiap bulan dan kader aktifnya terbatas.
  - 2. Posyandu madya, adalah posyandu yang sudah dapat melaksanakan kegiatan lebih dari 8 kali per tahun dengan rata-rata jumlah kader tugas 5 orang atau lebih. Akan tetapi cakupan program utamanya (KB, KIA, Gizi, dan imunisasi) masih rendah yaitu kurang dari 50 persen.
  - 3. Posyandu purnama, adalah posyandu yang frekuensinya lebih dari 8 kali per tahun, rata-rata jumlah kader tugas 5 orang atau lebih, dan cakupan 5 program utamanya (KB, KIA, Gizi, dan imunisasi) lebih dari 50 persen.
  - 4. Posyandu mandiri, adalah posyandu yang sudah dapat melakukan kegiatan secara teratur, cakupan 5 program utama sudah bagus, ada program tambahan dan dana sehat telah menjangkau lebih dari 50 persen KK. Intervensinya adalah pembinaan dana sehat, yaitu diarahkan agar dana sehat tersebut menggunakan prinsip JPKM.

### **3.2. Analisis Deskriptif**

Dalam arti sempit, statistik dapat diartikan sebagai data, tetapi dalam arti luas statistik dapat diartikan sebagai alat. Alat untuk analisis, dan alat untuk membuat keputusan (Sugiyono, 2007).

Ilmu statistik atau statistika adalah sebuah ilmu yang mempelajari teknik-teknik pengumpulan, pengorganisasian (pengaturan), analisis, dan interpretasi atas informasi data (Hakim, 2004)

Statistika sendiri dibedakan menjadi dua, yaitu statistika deskriptif dan statistika inferensial. Menurut Hasan (2004), statistika deskriptif adalah ilmu yang berisi metode-metode pengumpulan, penyajian, dan pengaturan data guna membuat gambaran yang jelas tentang variasi sifat data, yang pada akhirnya akan mempermudah proses analisis interpretasi data.

Sedangkan menurut Sugiyono (2007) statistika deskriptif adalah statistik yang berfungsi untuk mendiskripsikan atau memberi gambaran terhadap obyek yang diteliti melalui data sampel atau populasi sebagaimana adanya, tanpa melakukan analisis dan membuat kesimpulan yang berlaku untuk umum. Pada statistika deskriptif ini, digunakan beberapa cara penyajian data dengan tabel biasa maupun distribusi frekuensi, diagram garis maupun batang, diagram lingkaran, pictogram, penjelasan kelompok melalui modus, median, dan variasi kelompok.

### **3.2.1. Penyajian Data**

Setiap peneliti harus dapat menyajikan data yang telah diperoleh, baik yang diperoleh melalui observasi, wawancara, kuesioner (angket) maupun dokumentasi. Prinsip dasar penyajian data adalah komunikatif dan lengkap, dalam arti data yang disajikan dapat menarik perhatian pihak lain untuk membaca dan mudah memahami isi dari data tersebut (Sugiyono, 2007).

Pada penelitian ini, penyajian data dilakukan dengan dua cara, yaitu dengan menggunakan tabel dan diagram batang. Adapun penjelasan untuk tabel dan diagram batang dijelaskan sebagai berikut.

#### **a. Tabel**

Penyajian data hasil penelitian dengan menggunakan tabel merupakan penyajian yang banyak digunakan, karena lebih efisien dan cukup komunikatif. Terdapat dua macam tabel, yaitu tabel biasa dan tabel distribusi frekuensi. Setiap tabel berisi judul tabel, judul setiap kolom, nilai data dalam setiap kolom, dan sumber data dari mana data tersebut diperoleh (Sugiyono, 2007).

Contoh:

**Tabel 3.1** *Komposisi pendidikan pegawai di PT. Lodaya*

No	Bagian	Tingkat Pendidikan								Jumlah
		S3	S2	S1	SM	SMU	SMK	SMP	SD	
1	Keuangan	-	-	25	90	45	156	12	3	331
2	Umum	-	-	5	6	6	8	4	1	30
3	Penjualan	-	-	7	-	-	65	37	5	114
4	Litbang	1	8	35	-	-	-	-	-	44
		1	8	72	96	51	229	53	9	519

Sumber: Sugiyono, 2007

### b. Diagram Batang

Diagram batang digunakan untuk distribusi frekuensi. Pembatas-pembatas kelas diletakkan sepanjang sumbu horisontal dari grafik, sementara jumlah observasi, atau frekuensi, atau persentase, atau proporsi observasi diletakkan sepanjang sumbu vertikal. Tetapi dapat juga sumbu horisontal diisi dengan nilai tengah kelas. Lebar dari setiap batang menyatakan interval kelas sementara tinggi dari tiap batang menyatakan frekuensi dari observasi atau pengukuran di dalam kelas (Hakim, 2004).

Contoh:



**Gambar 3.1** *Jumlah penerima beasiswa Fakultas MIPA UII tahun 2017*

Sumber: Data simulasi

### 3.2.2. Pengukuran Gejala Pusat (*Central Tendency*)

Beberapa teknik penjelasan kelompok yang telah diobservasi dengan data kuantitatif, selain dapat dijelaskan dengan menggunakan tabel dan gambar, dapat pula dijelaskan menggunakan teknik statistik yang di dasarkan atas gejala pusat

(*tendency central*) dari kelompok tersebut. Salah satu alat pengukuran gejala pusat adalah *mean* atau rata-rata (Sugiyono, 2007).

#### a. *Mean*

*Mean* merupakan teknik penjelasan kelompok yang didasarkan atas nilai rata-rata dari suatu kelompok. *Mean* didapat dengan menjumlahkan data seluruh individu dalam kelompok, kemudian dibagi dengan jumlah individu yang ada pada kelompok tersebut. Adapun persamaan untuk menghitung *mean* adalah sebagai berikut (Sugiyono, 2007):

$$\mu = \frac{\sum_{i=1}^n X_i}{n} \dots\dots\dots(1)$$

dengan

$\mu$  = *mean*/rata-rata

$X_i$  = nilai objek ke  $i$  sampai  $n$

$n$  = banyaknya objek.

Contoh:

**Tabel 3.2** Penghasilan keluarga Bonar dalam 1 bulan

Anggota Keluarga	Penghasilan (Rp)
Ayah	5.000.000
Ibu	4.500.000
Anak pertama	4.000.000
Anak ke dua	4.200.000
Anak ke tiga	3.000.000

Sumber: Data simulasi

Berdasarkan data pada tabel 3.2 maka perhitungan rata-rata penghasilan keluarga Bonar dalam 1 bulan adalah sebagai berikut:

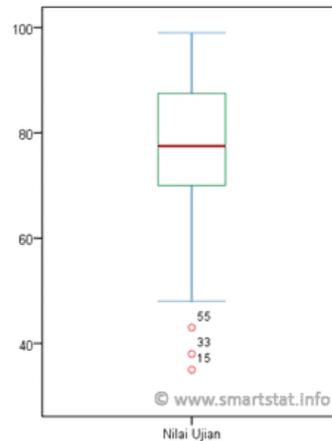
$$\mu = \frac{5.000.000 + 4.500.000 + 4.000.000 + 4.200.000 + 3.000.000}{5} = 3.940.000$$

Jadi, rata-rata penghasilan keluarga Bonar dalam 1 bulan adalah Rp3.940.000,00.

### 3.3. Deteksi Data *Outlier*

Data *outlier* adalah data yang secara nyata berbeda dengan data-data yang lain (Nugroho, 2008). Sedangkan menurut Prasetyo (2012) data *outlier* merupakan data-data yang karakteristiknya menyimpang (berbeda) dari data yang lain.

Pendeteksian data *outlier* dapat dilakukan dengan menggunakan grafik seperti *scatterplot* dan *boxplot* (Hair, dkk, 1998). Adapun contoh *boxplot* dapat dilihat pada gambar 3.2 berikut.



**Gambar 3.2** *Boxplot*

Sumber: NN, 2010

Data *outlier* dapat terdeteksi apabila terdapat data yang letaknya lebih dari 1,5 kali panjang kotak jika diukur dari atas kotak atau bawah kotak (Setiawan, 2010).

Jika data diketahui terdapat satu atau lebih data *outlier*, pada *outlier* tersebut dapat dilakukan penanganan sebagai berikut (Yulianto dan Hidayatullah, 2014):

- a. Data *outlier* dihilangkan, jika data *outlier* tersebut didapat karena kesalahan pengambilan data, kesalahan *inputing* pada komputer, dan sebagainya.
- b. Data *outlier* tetap dipertahankan dan tidak perlu dihilangkan, jika tidak terdapat kesalahan pada proses *sampling* maupun *inputing* data.

### 3.4. *Data Mining*

*Data mining* adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam *database*. *Data mining* adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat untuk pengetahuan yang terakit dari berbagai *database* besar (Turban, dkk, 2005).

Menurut Gartner Group, *data mining* adalah suatu proses menemukan hubungan yang berarti, pola, dan kecenderungan dengan memeriksa dalam sekumpulan besar data yang tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola seperti teknik statistik dan matematika (Larose, 2005).

Menurut David Hand, Heikki Mannila, dan Padhraic Smyth dari MIT, *data mining* adalah analisa terhadap data (biasanya data yang berukuran besar) untuk menemukan hubungan yang jelas serta menyimpulkannya yang belum diketahui sebelumnya dengan cara terkini dipahami dan berguna bagi pemilik data tersebut (Larose, 2006).

Tan (2006) mendefinisikan *data mining* sebagai proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar. *Data mining* juga dapat diartikan sebagai pengekstrakan informasi baru yang diambil dari bongkahan data besar yang membantu dalam pengambilan keputusan. Istilah data mining kadang disebut dengan *knowledge discovery*.

Dari definisi-definisi yang telah disampaikan, hal penting yang terkait dengan *data mining* adalah (Kusrini dan Luthfi, 2009):

- a. *Data mining* merupakan suatu proses otomatis terhadap data yang sudah ada.
- b. Data yang akan diproses berupa data yang sangat besar.
- c. Tujuan *data mining* adalah mendapatkan hubungan atau pola yang mungkin memberikan indikasi yang bermanfaat.

Adapun beberapa fungsi dalam *data mining* menurut Larose (2005) adalah:

- a. Fungsi deskripsi (*description*), digunakan untuk mencari pola dan kecenderungan yang terdapat dalam data.
- b. Fungsi estimasi (*estimation*), hampir sama dengan klasifikasi, namun variabel target dalam estimasi lebih ke arah numerik daripada kategorik.
- c. Fungsi prediksi (*prediction*), hampir sama dengan klasifikasi, namun dalam prediksi nilai dari hasil akan ada di masa mendatang.
- d. Fungsi klasifikasi (*classification*), adalah fungsi pembelajaran yang memetakan (mengklasifikasi) sebuah unsur (*item*) data ke dalam salah satu dari beberapa kelas yang sudah didefinisikan.

- e. Fungsi pengklasteran (*clustering*), merupakan tugas deskripsi yang banyak digunakan dalam mengidentifikasi sebuah hubungan terbatas pada kategori atau klaster untuk mendeskripsikan data yang ditelaah.
- f. Fungsi asosiasi (*association*), dalam *data mining* fungsi ini bertugas untuk menemukan atribut yang muncul dalam satu waktu.

Sedangkan menurut Berry dan Browne (2006), keenam fungsi *data mining* di atas dapat dipilah menjadi:

- a. Fungsi minor atau fungsi tambahan, yang meliputi ketiga fungsi pertama, yaitu deskripsi, estimasi, dan prediksi.
- b. Fungsi mayor atau fungsi utama, yang meliputi ketiga fungsi berikutnya, yaitu klasifikasi, pengklasteran, dan asosiasi.

Prasetyo (2012) menjelaskan bahwa permasalahan yang berkaitan dengan *data mining* dibagi menjadi empat kelompok, yaitu:

- a. Model prediksi, berkaitan dengan pembuatan sebuah model yang dapat melakukan pemetaan dari setiap himpunan variabel ke setiap targetnya, kemudian menggunakan model tersebut untuk memberikan nilai target pada himpunan baru yang didapat. Ada dua jenis model prediksi, yaitu klasifikasi dan regresi. Klasifikasi digunakan untuk variabel target diskret, sedangkan regresi untuk variabel target kontinu.
- b. Analisis klaster (pengklasteran), di mana analisis ini melakukan pengklasteran data-data ke dalam sejumlah klaster berdasarkan kesamaan karakteristik masing-masing data pada klaster-klaster yang ada. Data-data yang masuk dalam batas kesamaan dengan klasternya akan bergabung dalam klaster tersebut, dan akan terpisah dalam klaster yang berbeda jika keluar dari batas kesamaan dengan klaster tersebut.
- c. Analisis asosiasi, atau disebut dengan *association analysis* digunakan untuk menemukan pola yang menggambarkan kekuatan hubungan fitur dalam data. Pola yang ditemukan biasanya mempresentasikan bentuk aturan implikasi atau subset fitur. Tujuannya adalah untuk menemukan pola yang menarik dengan cara yang efisien.

- d. Deteksi anomali, atau *anomaly detection* berkaitan dengan pengamatan sebuah data dari sejumlah data yang secara signifikan mempunyai karakteristik yang berbeda dari sisa data yang lain (*outlier*). Algoritma deteksi anomali yang baik harus mempunyai laju deteksi yang tinggi dan laju *error* yang rendah. Deteksi anomali dapat diterapkan pada sistem jaringan untuk mengetahui pola data yang memasuki jaringan sehingga penyusupan bisa ditemukan jika pola kerja data yang datang berbeda.

Secara skematis, Gorunescu (2011) membagi langkah proses pelaksanaan *data mining* dalam tiga aktivitas, yaitu:

- a. Eksplorasi data, terdiri dari aktivitas pembersihan data, transformasi data, pengurangan dimensi, pemilihan ciri, dan lain-lain.
- b. Membuat model dan pengujian validitas model, merupakan pemilihan terhadap model-model yang sudah dikembangkan yang cocok dengan kasus yang dihadapi. Dengan kata lain, pemilihan model secara kompetitif.
- c. Penerapan model dengan data baru untuk menghasilkan perkiraan dari kasus yang ada. Tahap ini merupakan tahap yang menentukan apakah model yang telah dibangun dapat menjawab permasalahan yang dihadapi.

Kemajuan luar biasa yang terus berlanjut dalam bidang *data mining* didorong oleh beberapa faktor, antara lain (Larose, 2005):

- a. Pertumbuhan yang cepat dalam kumpulan data.
- b. Penyimpanan data dalam *data warehouse*, sehingga seluruh perusahaan memiliki akses ke dalam *database* yang andal.
- c. Adanya peningkatan akses data melalui navigasi web dan intranet.
- d. Tekanan kompetisi bisnis untuk meningkatkan penguasaan pasar dalam globalisasi ekonomi.
- e. Perkembangan teknologi perangkat lunak untuk *data mining* (ketersediaan teknologi).
- f. Perkembangan yang hebat dalam kemampuan komputasi dan pengembangan kapasitas media penyimpanan.

### 3.4.1. Pengklasteran

#### a. Pengertian Pengklasteran

Analisis kluster atau pengklasteran merupakan proses membagi data dalam suatu himpunan ke dalam beberapa kluster yang mana kesamaan data dalam suatu kluster lebih besar dibandingkan kesamaan data tersebut dengan data dalam kluster lain (Jang dkk, 1997).

Menurut Supranto (2004), analisis kluster merupakan suatu kelas teknik, dipergunakan untuk mengklasifikasi objek atau kasus (responden) ke dalam kelompok yang relatif homogen, yang disebut kluster. Objek/kasus dalam setiap kluster cenderung mirip satu sama lain dan berbeda jauh (tidak sama) dengan objek dari kluster lainnya. Analisis kluster juga disebut analisis klasifikasi atau taksonomi numerik (*numerical taxonomy*).

Kuncoro (2003) menyatakan bahwa pengklasteran adalah teknik yang digunakan untuk mengidentifikasi objek atau individu yang serupa dengan memperhatikan beberapa kriteria. Pengklasteran yaitu analisis untuk mengklasterkan elemen yang mirip sebagai objek penelitian menjadi kelompok (kluster) yang berbeda dan *mutually exclusive*.

Menurut Han dan Kamber (2011), pengklasteran adalah proses pengelompokkan kumpulan data menjadi beberapa kluster sehingga objek di dalam suatu kluster memiliki banyak kesamaan dan memiliki banyak perbedaan dengan objek di kluster lain. Perbedaan dan persamaannya biasanya berdasarkan nilai atribut dari objek tersebut dan dapat juga berupa perhitungan jarak. Pengklasteran sendiri juga disebut *unsupervised classification*, karena hitungan pengklasteran lebih bersifat untuk dipelajari dan diperhatikan. Analisis kluster merupakan proses partisi satu set objek data ke dalam himpunan bagian. Setiap himpunan bagian adalah kluster, sehingga objek yang di dalam kluster mirip satu sama dengan yang lainnya, dan mempunyai perbedaan dengan objek dari kluster yang lain. Partisi tidak dilakukan dengan manual tetapi dengan algoritma pengklasteran. Oleh karena itu, pengklasteran sangat berguna dan bisa menemukan kluster yang tidak dikenal dalam data.

Dalam *data mining*, pengklasteran adalah metode yang digunakan untuk membagi rangkaian data menjadi beberapa klaster berdasarkan kesamaan-kesamaan yang telah ditentukan (Gorunescu, 2011).

Tan (2006) dalam Prasetyo (2012) menjelaskan bahwa analisis klaster adalah pekerjaan mengklasterkan data (objek) yang didasarkan hanya pada informasi yang ditemukan dalam data yang menggambarkan objek tersebut dan hubungan di antaranya. Tujuannya adalah agar objek-objek yang bergabung dalam sebuah klaster merupakan objek-objek yang mirip (atau berhubungan) satu sama lain dan berbeda (atau tidak berhubungan) dengan objek dalam klaster yang lain. Lebih besar kemiripannya (homogenitas) dalam klaster dan lebih besar perbedaannya di antara klaster yang lain, konsep inilah yang terdapat dalam pengklasteran.

Adapun ciri-ciri dari pengklasteran adalah (Sitepu dkk, 2011):

1. Homogenitas (kesamaan) yang tinggi antar anggota dalam satu klaster.
2. Heterogenitas (perbedaan) yang tinggi antar klaster satu dengan yang lain.

#### **b. Ukuran Jarak dalam Pengklasteran**

Berdasarkan tujuan pengklasteran yang telah dijelaskan, maka menurut Supranto (2004) terdapat beberapa ukuran yang diperlukan untuk mengakses seberapa mirip atau berbeda objek-objek yang diklasterkan. Pendekatan yang paling biasa ialah mengukur kemiripan dinyatakan dengan jarak (*distance*) antara pasangan objek. Objek dengan jarak lebih pendek antara mereka akan lebih mirip satu sama lain dibandingkan dengan pasangan dengan jarak antara dua objek.

Ukuran jarak yang paling banyak digunakan adalah *Euclidean* dan *Mahattan*. *Euclidean* digunakan ketika ingin memberikan jarak terpendek antara dua titik (jarak lurus), sedangkan *Mahattan* memberikan jarak terjauh pada dua data. *Mahattan* juga sering digunakan karena kemampuannya dalam mendeteksi keadaan khusus dengan lebih baik (Augusta, 2005 dalam Prasetyo, 2012).

##### **1. Jarak *Euclidean***

Jarak *Euclidean* ialah akar dari jumlah kuadrat perbedaan/deviasi di dalam nilai untuk setiap variabel (Supranto, 2004). Jarak *Euclidean* juga biasa disebut sebagai metode perhitungan jarak yang didasarkan pada ruang berdimensi terbatas

bernilai riil (Kumari dan Bhagat, 2013). Adapun persamaan untuk menghitung jarak *Euclidean* adalah sebagai berikut (Prasetyo, 2014):

$$d_{ij} = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2} \dots\dots\dots(2)$$

dengan

$d_{ij}$  = jarak antara objek  $i$  dan objek  $j$

$x_{ik}$  = nilai objek  $i$  pada variabel ke  $k$

$x_{jk}$  = nilai objek  $j$  pada variabel ke  $k$

$p$  = banyak variabel yang diamati.

Contoh:

**Tabel 3.3** *Data perhitungan jarak Euclidean*

Objek	$X_1$	$X_2$
A	4	1
B	1	2
C	5	3

Sumber: Prasetyo, 2014

Perhitungan jarak *Euclidean* untuk data pada tabel 3.3 adalah:

$$d_{AB} = \sqrt{(4-1)^2 + (1-2)^2} = 3,16$$

$$d_{AC} = \sqrt{(4-5)^2 + (1-3)^2} = 2,23$$

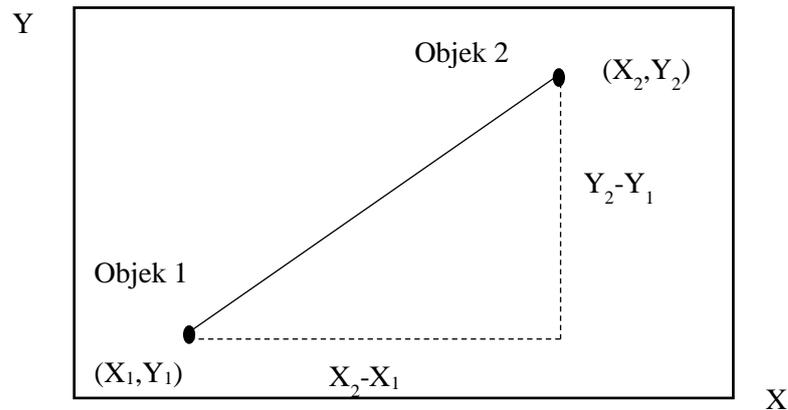
$$d_{BC} = \sqrt{(1-5)^2 + (2-3)^2} = 4,12$$

Sehingga jarak antar objek dapat dilihat secara keseluruhan pada tabel 3.4 berikut.

**Tabel 3.4** *Matriks jarak Euclidean*

	A	B	C
A	0	3,16	2,23
B	3,16	0	4,12
C	2,23	4,12	0

Secara geometris, dengan menggunakan dua fitur maka jarak *Euclidean* akan sama dengan persamaan *pythagoras* dalam menghitung sisi miring (Prasetyo, 2014). Adapun visualisasi jarak *Euclidean* adalah sebagai berikut.



**Gambar 3.3** Jarak Euclidean dalam dua dimensi

Sumber: Hair, dkk, 1998

## 2. Jarak Mahattan

Jarak *Mahattan* atau *the city block distance* merupakan jarak antara dua objek yang merupakan jumlah perbedaan mutlak/absolut di dalam nilai untuk setiap variabel (Supranto, 2004). Adapun persamaan untuk jarak *Mahattan* adalah sebagai berikut (Prasetyo, 2014):

$$d_{ij} = \sum_{k=1}^p |x_{ik} - x_{jk}| \dots\dots\dots(3)$$

dengan

$d_{ij}$  = jarak antara objek  $i$  dan objek  $j$

$x_{ik}$  = nilai objek  $i$  pada variabel ke  $k$

$x_{jk}$  = nilai objek  $j$  pada variabel ke  $k$

$p$  = banyak variabel yang diamati.

Contoh:

Dengan menggunakan data pada tabel 3.3 maka perhitungan jarak *Mahattan* adalah sebagai berikut:

$$d_{AB} = |4 - 1| + |1 - 2| = 4$$

$$d_{AC} = |4 - 5| + |1 - 3| = 3$$

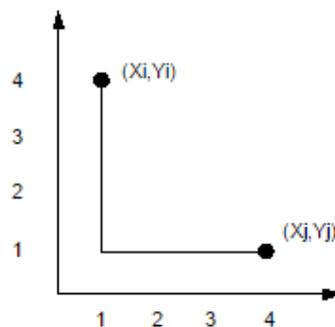
$$d_{BC} = |1 - 5| + |2 - 3| = 5$$

Sehingga jarak antar objek dapat dilihat secara keseluruhan pada tabel 3.5 berikut.

**Tabel 3.5** *Matriks jarak Mahattan*

	A	B	C
A	0	4	3
B	4	0	5
C	3	5	0

Secara geometris, dengan menggunakan dua fitur maka jarak *Mahattan* akan sama dengan jarak terjauh antara dua vektor (Prasetyo, 2014). Adapun visualisasi jarak *Mahattan* adalah sebagai berikut.



**Gambar 3.4** *Jarak Mahattan dalam dua dimensi*

Sumber: NN, 2007

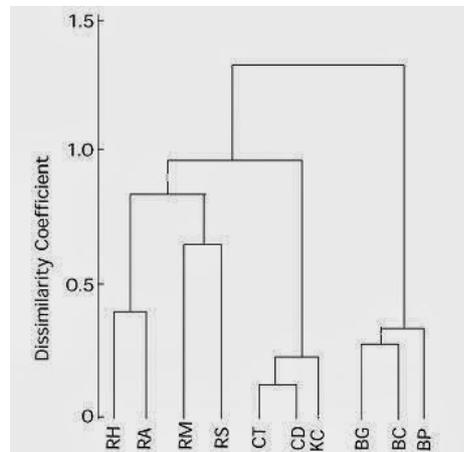
### c. Jenis Pengklasteran

Menurut struktur, pengklasteran dibagi menjadi dua, yaitu pengklasteran hierarki dan sekatan (*partitioning*) atau *non* hierarki. Dalam pengklasteran hierarki, satu data tunggal dapat dianggap sebuah klaster, dua atau lebih klaster kecil dapat bergabung menjadi sebuah klaster besar dan begitu seterusnya hingga semua data dapat bergabung menjadi sebuah klaster. Sedangkan metode pengklasteran sekatan membagi set data ke dalam sejumlah klaster yang tidak tumpang tindih antara satu klaster dengan klaster lain, yang artinya setiap data hanya menjadi anggota satu klaster (Prasetyo, 2012).

#### 1. Pengklasteran Hierarki

Pengklasteran hierarki sering ditampilkan dalam bentuk grafis menggunakan diagram yang mirip pohon (*tree*) yang disebut dengan dendogram. Dendogram merupakan diagram yang menampilkan klaster dan subklasternya dalam urutan, di

mana klasternya digabung (*agglomerative view*) atau dipisah (*divisive view*) (Prasetyo (2012)).



**Gambar 3.5** Contoh dendrogram

Sumber: Dina, 2014

Dalam statistik, pengklasteran hierarki dibagi menjadi dua jenis, yaitu aglomeratif dan divisif (Prasetyo, 2012).

- a. Pengklasteran hierarki aglomeratif, yaitu metode pengklasteran hierarki dengan pendekatan bawah-atas (*bottom up*). Proses pengklasteran dimulai dari masing-masing data sebagai satu buah klaster, kemudian secara rekursif mencari klaster terdekat sebagai pasangan untuk bergabung sebagai salah satu klaster yang lebih besar. Proses tersebut diulang terus sehingga tampak bergerak ke atas membentuk jenjang (hierarki). Cara ini membutuhkan parameter kedekatan klaster (*cluster proximity*).
- b. Pengklasteran hierarki divisif, merupakan metode pengklasteran hierarki dengan pendekatan atas-bawah (*top down*). Proses pengklasteran dimulai dari satu klaster yang berisi semua data, kemudian secara rekursif memecah klaster menjadi dua klaster hingga setiap klaster hanya berisi satu data tunggal (data itu sendiri). Hal yang dibutuhkan dalam pengklasteran ini adalah keputusan klaster yang manakah yang akan dipecah pada setiap langkah dan bagaimana cara memecahnya.

Beberapa metode pengklasteran hierarki antara lain.

1. *Single linkage* (jarak terdekat atau tautan tunggal), pengklasteran ini memberikan hasil bila klaster-klaster digabungkan menurut jarak antara anggota-anggota yang terdekat di antara dua kelompok (Prasetyo, 2012).
2. *Complete linkage* (jarak terjauh atau tautan lengkap), pengklasteran ini terjadi bila klaster-klaster digabungkan menurut jarak antara anggota-anggota yang terjauh di antara dua klaster (Prasetyo, 2012).
3. *Average linkage* (jarak rata-rata atau tautan rata-rata), pengklasteran ini menggabungkan objek menurut jarak rata-rata pasangan-pasangan anggota masing-masing pada himpunan antara dua klaster (Prasetyo, 2012).
4. *Ward's method*, metode ini menggunakan perhitungan yang lengkap dan memaksimalkan homogenitas di dalam satu klaster (Dillon dan Goldstein, 1984).

## 2. Pengklasteran Sekatan (*Partitioning*) atau *Non Hierarki*

Pengklasteran berbasis *partitioning* menghasilkan partisi dari data sehingga objek dalam klaster lebih mirip satu sama lain daripada objek yang ada dalam klaster lain (Triyanto, 2015). Berbeda dengan pengklasteran hierarki, prosedur pengklasteran sekatan tidak dilakukan secara bertahap, dan jumlah klasternya juga ditentukan terlebih dahulu (Machfudhoh dan Wahyuningsih, 2013). Beberapa metode pengklasteran sekatan yang digunakan dalam penelitian ini antara lain.

### a. *K-Means Clustering*

*K-Means Clustering* merupakan algoritma pengklasteran yang paling sederhana dibanding algoritma pengklasteran yang lain. *K-Means Clustering* mempunyai kelebihan mudah diterapkan dan dijalankan, relatif cepat, mudah untuk diadaptasi, dan paling banyak dipraktikkan dalam tugas *data mining*. Pengklasteran merupakan suatu metode untuk pengklasteran dokumen di mana dokumen diklasterkan dengan konten untuk mengurangi ruang pencarian yang diperlukan dalam merespon suatu *query* (Grossman dan Frieder, 2004).

*K-Means Clustering* merupakan algoritma pengklasteran iteratif yang melakukan partisi set data ke dalam sejumlah  $k$  klaster yang sudah ditetapkan di awal. *K-Means Clustering* sederhana untuk diimplementasikan dan dijalankan, relatif cepat, mudah beradaptasi, umum penggunaannya dalam praktek. Secara

historis, *K-Means Clustering* menjadi salah satu algoritma yang paling penting dalam bidang *data mining* (Wu dan Kumar, 2009 dalam Prasetyo, 2014).

*K-Means Clustering* diperkenalkan oleh MacQueen pada tahun 1967 (Ibrahim, dalam Chrisnanto dan Abdillah, 2015). *K-Means* merupakan metode pengklasteran secara *partitioning* yang memisahkan data ke dalam kluster yang berbeda. Dengan *partitioning* secara iteratif, *K-Means* mampu meminimalkan rata-rata jarak setiap data ke klasternya (MacQueen, 1967).

*K-Means* merupakan salah satu metode pengklasteran data *non* hierarki (sekatan) yang berusaha mempartisi data yang ada ke dalam bentuk dua atau lebih kluster. Metode ini mempartisi data ke dalam kluster sehingga data berkarakteristik berbeda diklasterkan ke dalam kluster yang lain. Adapun tujuan pengklasteran data ini adalah untuk meminimalkan fungsi objektif yang diset dalam proses pengklasteran, yang pada umumnya berusaha meminimalkan variasi di dalam suatu kluster dan memaksimalkan variasi antar kluster. Adapun langkah-langkah untuk *K-Means Clustering* adalah sebagai berikut (Prasetyo, 2014):

1. Inisialisasi: tentukan nilai  $k$  sebagai jumlah kluster yang diinginkan dan matriks jarak yang diinginkan.
2. Pilih  $k$  data dari set data  $X$  sebagai *centroid*.
3. Alokasikan semua data ke *centroid* terdekat dengan matriks jarak yang sudah ditetapkan (memperbarui kluster *ID* pada setiap data).
4. Hitung kembali *centroid* berdasarkan data yang mengikuti kluster masing-masing.
5. Ulangi langkah 3 dan 4 hingga kondisi konvergen tercapai, yaitu tidak ada data yang berpindah kluster.

*K-Means Clustering* digunakan sebagai alternatif metode kluster untuk data dengan ukuran yang besar karena memiliki kecepatan yang lebih tinggi dibandingkan dengan metode hierarki (Sitepu dkk, 2011).

Contoh penerapan *K-Means Clustering* (Prasetyo, 2014):

**Tabel 3.6** Kondisi awal data

Data ke $i$	$X$	$Y$
1	1	1
2	4	1

Data ke $i$	$X$	$Y$
3	6	1
4	1	2
5	2	3
6	5	3
7	2	5
8	3	5
9	2	6
10	3	8

1. Inisialisasi, ditentukan  $k$  (jumlah kluster) dari  $n$  objek adalah 3, kemudian dipilih  $k$  data sebagai *centroid* awal, misalnya dipilih data ke 2, 4, dan 6.

**Tabel 3.7** *Centroid awal*

<i>Centroid</i>	$x$	$y$
1	4	1
2	1	2
3	5	3

2. Iterasi 1, hitung jarak setiap data ke *centroid* terdekat. *Centroid* terdekat akan menjadi kluster yang diikuti oleh data tersebut. Dalam penelitian ini, jarak yang digunakan adalah jarak *Euclidean*. Berikut contoh perhitungan jarak ke setiap *centroid* pada data ke 1:

$$d_{x_1, c_1} = \sqrt{(1-4)^2 + (1-1)^2} = 3$$

$$d_{x_1, c_2} = \sqrt{(1-1)^2 + (1-2)^2} = 1$$

$$d_{x_1, c_3} = \sqrt{(1-5)^2 + (1-3)^2} = 4,47$$

Cara yang sama dapat dilakukan untuk perhitungan jarak ke setiap *centroid* pada data ke 2 sampai data ke 10, sehingga diperoleh tabel 3.8 berikut.

**Tabel 3.8** *Hasil perhitungan jarak ke setiap centroid*

Data ke $i$	Jarak ke <i>centroid</i>			Terdekat	Kluster yang diikuti
	1	2	3		
1	3	1	4,47	1	2
2	0	3,16	2,24	0	1
3	2	5,10	2,24	2	1
4	3,16	0	4,12	0	2
5	2,83	1,41	3	1,41	2
6	2,24	4,12	0	0	3
7	4,47	3,15	3,61	3,15	2
8	4,12	3,61	2,83	2,83	3
9	5,39	4,12	4,24	4,12	2
10	7,07	6,32	5,39	5,39	3

Selanjutnya hitung nilai *centroid* yang baru untuk setiap kluster berdasarkan data yang bergabung pada setiap klasternya. Adapun persamaan untuk perhitungan nilai *centroid* adalah sebagai berikut:

$$c_{kj} = \frac{1}{Nk} \sum_{i=1}^{Nk} x_{ij} \dots\dots\dots(4)$$

dengan

$c_{kj}$  = *centroid* kluster ke  $k$  pada variabel ke  $j$

$Nk$  = jumlah data yang tergabung dalam kluster  $k$

$x_{ij}$  = nilai objek ke  $i$  pada variabel ke  $j$ .

Sehingga perhitungan *centroid* baru untuk kluster 1 adalah:

**Tabel 3.9 Anggota kluster 1**

Data anggota	X	Y
2	4	1
3	6	1

$$c_{11} = \frac{4 + 6}{2} = 5$$

$$c_{12} = \frac{1 + 1}{2} = 1$$

Perhitungan *centroid* baru untuk kluster 2 adalah:

**Tabel 3.10 Anggota kluster 2**

Data anggota	X	Y
1	1	1
4	1	2
5	2	3
7	2	5
9	2	6

$$c_{21} = \frac{1 + 1 + 2 + 2 + 2}{5} = 1,60$$

$$c_{22} = \frac{1 + 2 + 3 + 5 + 6}{5} = 3,40$$

Perhitungan *centroid* baru untuk kluster 3 adalah:

**Tabel 3.11 Anggota kluster 3**

Data anggota	X	Y
6	5	3

Data anggota	X	Y
8	3	5
10	3	8

$$c_{31} = \frac{5 + 3 + 3}{3} = 3,67$$

$$c_{32} = \frac{3 + 5 + 8}{3} = 5,34$$

Sehingga diperoleh *centroid* baru pada tabel berikut.

**Tabel 3.12** *Centroid baru hasil iterasi 1*

Centroid	x	y
1	5	1
2	1,60	3,40
3	3,67	5,34

3. Iterasi 2, hitung kembali jarak setiap data ke *centroid* baru hasil iterasi 1. Berikut contoh perhitungan jarak ke setiap *centroid* baru hasil iterasi 1 pada data ke 1:

$$d_{x_1, c_1} = \sqrt{(1-5)^2 + (1-1)^2} = 4$$

$$d_{x_1, c_2} = \sqrt{(1-1,60)^2 + (1-3,40)^2} = 2,47$$

$$d_{x_1, c_3} = \sqrt{(1-3,67)^2 + (1-5,34)^2} = 5,02$$

Cara yang sama dapat dilakukan untuk perhitungan jarak ke setiap *centroid* baru hasil iterasi 1 pada data ke 2 sampai data ke 10, sehingga diperoleh tabel 3.13 berikut.

**Tabel 3.13** *Hasil perhitungan jarak ke setiap centroid baru hasil iterasi 1*

Data ke <i>i</i>	Jarak ke <i>centroid</i>			Terdekat	Klaster yang diikuti
	1	2	3		
1	4	2,47	5,02	2,47	2
2	1	3,39	4,32	1	1
3	1	5,01	4,92	1	1
4	4,12	1,52	4,20	1,52	2
5	3,61	0,57	2,80	0,57	2
6	2	3,42	2,69	2	1
7	5	1,65	1,63	1,63	3
8	4,47	2,13	0,67	0,67	3
9	5,83	2,63	1,75	1,75	3
10	7,28	4,81	2,77	2,77	3

Karena masih terdapat data yang berpindah kluster, maka langkah selanjutnya adalah menghitung *centroid* hasil iterasi 2 yang dilanjutkan dengan iterasi 3 hingga kondisi konvergen tercapai.

**b. *K-Medoids Clustering***

*K-Medoids Clustering*, juga dikenal sebagai *Partitioning Around Medoids (PAM)*, adalah varian dari metode *K-Means*. Hal ini didasarkan pada penggunaan *medoids* bukan dari pengamatan *mean* yang dimiliki oleh setiap kluster, dengan tujuan mengurangi sensitivitas dari partisi sehubungan dengan nilai ekstrim yang ada dalam dataset (Vercellis, 2009).

*K-Medoids Clustering* hadir untuk mengatasi kelemahan *K-Means Clustering* yang sensitif terhadap *outlier* karena suatu objek dengan suatu nilai yang besar mungkin secara substansial menyimpang dari distribusi data (Han dan Kamber, 2006). Metode ini diperkenalkan oleh Kaufman dan Rousseeuw pada tahun 1990 (Ibrahim, dalam Chrisnanto dan Abdillah, 2015)

*K-Medoids Clustering* menggunakan metode pengklasteran partisi untuk mengklusterkan sekumpulan  $n$  objek menjadi sejumlah  $k$  kluster. Algoritma ini menggunakan objek pada kumpulan objek yang mewakili sebuah kluster. Objek yang mewakili sebuah kluster disebut dengan *medoids*. Kluster dibangun dengan menghitung kedekatan yang dimiliki antara *medoids* dengan objek *non medoids* (Setyawati, 2017).

*Medoids* juga dapat diartikan sebagai sebuah objek dari sebuah gerombol yang mempunyai rata-rata jarak terkecil ke objek lainnya, dengan kata lain yaitu objek yang terletak di tengah-tengah gugus data (Flowrensia, 2010).

Menurut Han dan Kamber (2006), tahapan *K-Medoids Clustering* adalah sebagai berikut.

1. Secara acak pilih  $k$  objek pada sekumpulan  $n$  objek sebagai *medoids*.
2. Ulangi.
3. Tempatkan objek *non medoids* ke dalam kluster yang paling dekat dengan *medoids*.
4. Secara acak pilih  $O_{random}$  (sebuah objek *non medoids*).
5. Hitung total *cost* dan  $S$  dari pertukaran *medoids*  $O_j$  dengan  $O_{random}$ .

6. Jika  $S < 0$  maka tukar  $O_j$  dengan  $O_{random}$ , untuk membentuk sekumpulan  $k$  objek baru sebagai *medoids*.
7. Hingga tidak ada perubahan.

Contoh penerapan *K-Medoids Clustering*:

**Tabel 3.14** Kondisi awal data

Data ke $i$	$X$	$Y$
1	66	402
2	31	182
3	49	258
4	50	289
5	51	281
6	65	464
7	75	387
8	162	964
9	113	706
10	61	329
11	48	290
12	59	311

Sumber: Data simulasi

1. Ditentukan  $k$  (jumlah kluster) dari  $n$  objek adalah 3.
2. Tentukan *centroid* awal sebagai *medoids* dengan asumsi seperti pada tabel 3.15 berikut.

**Tabel 3.15** *Medoids*

Nama	Keterangan	$x$	$y$
$C1$	Diambil data ke 8 sebagai pusat kluster ke 1	162	964
$C2$	Diambil data ke 7 sebagai pusat kluster ke 2	75	387
$C3$	Diambil data ke 2 sebagai pusat kluster ke 3	31	182

3. Tempatkan objek-objek *non medoids* ke dalam kluster yang paling dekat dengan *medois* berdasarkan jarak *Euclidean*. Berikut contoh perhitungan jarak pada data ke 1:

$$d_{x_1, c_1} = \sqrt{(66 - 162)^2 + (402 - 964)^2} = 570,14$$

$$d_{x_1, c_2} = \sqrt{(66 - 75)^2 + (402 - 387)^2} = 16,16$$

$$d_{x_1, c_3} = \sqrt{(66 - 31)^2 + (402 - 182)^2} = 329,28$$

Cara yang sama dapat dilakukan untuk perhitungan jarak data ke 2 sampai data ke 10, sehingga diperoleh tabel 3.16 berikut.

**Tabel 3.16** Hasil perhitungan jarak ke setiap medoids

Data ke <i>i</i>	Jarak ke medoids			Terdekat	Kluster yang diikuti
	<i>C1</i>	<i>C2</i>	<i>C3</i>		
1	570,14	16,16	329,28	16,16	2
2	796,07	314,87	0	0	3
3	714,00	130,05	273,08	130,05	2
4	684,23	100,44	282,08	100,44	2
5	691,96	108,06	278,21	108,06	2
6	509,32	77,32	374,22	77,32	2
7	583,98	0	314,87	0	2
8	0	583,98	796,07	0	1
9	262,61	321,62	560,16	262,61	1
10	642,98	59,03	290,02	59,03	2
11	683,57	99,92	284,31	99,92	2
12	661,07	77,10	283,10	77,10	2
Jumlah	6.799,94	1.888,56	4.065,40		
Total Cost				12.753,90	

4. Tentukan objek *non medoids* dengan asumsi sebagai berikut.

**Tabel 3.17** Non Medoids

Nama	Keterangan	<i>x</i>	<i>y</i>
<i>D1</i>	Diambil data ke 3 sebagai pusat kluster ke 1	49	259
<i>D2</i>	Diambil data ke 5 sebagai pusat kluster ke 2	51	281
<i>D3</i>	Diambil data ke 9 sebagai pusat kluster ke 3	113	706

5. Ulangi langkah 3 untuk objek *non medoids* sehingga diperoleh tabel 3.18 berikut.

**Tabel 3.18** Hasil perhitungan jarak ke setiap non medoids

Data ke <i>i</i>	Jarak ke medoids			Terdekat	Kluster yang diikuti
	<i>D1</i>	<i>D2</i>	<i>D3</i>		
1	144,01	121,93	307,61	121,93	2
2	273,08	278,21	560,16	273,08	1
3	0	22,09	451,56	0	1
4	30,02	8,06	421,73	8,06	2
5	22,09	0	429,50	0	2
6	205,62	183,53	246,71	183,53	2
7	130,05	108,06	321,62	108,06	2
8	714	691,96	262,61	262,61	3
9	451,56	429,50	0	0	3
10	71,02	49,03	380,57	49,03	2
11	31,02	9,49	421,05	9,49	2
12	52,95	31,05	398,67	31,05	1
Jumlah	2.125,42	1.932,91	4.20,80		
Total Cost				8.260,13	

6. Hitung nilai *S* dengan persamaan sebagai berikut (Setyawati, 2017):

$$S = \text{total cost baru} - \text{total cost lama} \dots \dots \dots (5)$$

dengan

$S$  = selisih

Total *cost* baru = total *cost* untuk *non medoids*

Total *cost* lama = total *cost* untuk *medoids*.

Sehingga diperoleh:

$$S = 8.260,13 - 12.752,90 = -4.493,77$$

7. Karena nilai  $S < 0$  maka tukar *non medoids* dengan *medoids*, sehingga  $D1$ ,  $D2$ , dan  $D3$  menjadi objek *medoids*.
8. Tentukan objek *non medoids* baru dengan asumsi sebagai berikut.

**Tabel 3.19** *Non Medoids baru*

Nama	Keterangan	$x$	$y$
$E1$	Diambil data ke 8 sebagai pusat klaster ke 1	162	964
$E2$	Diambil data ke 10 sebagai pusat klaster ke 2	61	329
$E3$	Diambil data ke 12 sebagai pusat klaster ke 3	59	311

9. Ulangi langkah 3 untuk objek *non medoids* baru sehingga diperoleh tabel 3.20 berikut.

**Tabel 3.20** *Hasil perhitungan jarak ke setiap non medoids*

Data ke $i$	Jarak ke <i>medoids</i>			Terdekat	Klaster yang diikuti
	$E1$	$E2$	$E3$		
1	570,14	73,71	91,27	73,71	2
2	796,07	290,02	283,10	283,10	3
3	714	71,02	52,95	52,95	3
4	684,23	41,48	23,77	23,77	3
5	691,96	49,03	31,05	31,05	3
6	509,32	135,06	153,12	135,06	2
7	583,98	59,03	77,10	59,03	2
8	0	642,98	661,07	0	1
9	262,61	380,57	398,67	262,61	1
10	642,98	0	18,11	0	2
11	683,57	41,11	23,71	23,71	3
12	661,07	18,11	0	0	3
Jumlah	6.799,94	1.801,59	1.813,92		
Total <i>Cost</i>	10.415,45				

10. Hitung kembali nilai  $S$  sebagai berikut:

$$S = 10.415,45 - 8.260,13 = 2.155,32$$

11. Karena nilai  $S > 0$  maka proses pengklasteran dihentikan. Sehingga diperoleh anggota tiap klaster sebagai berikut.

**Tabel 3.21 Hasil pengklasteran data simulasi dengan K-Medoids Clustering**

Data ke <i>i</i>	<i>X</i>	<i>Y</i>	Klaster yang diikuti
1	66	402	2
2	31	182	1
3	49	258	1
4	50	289	2
5	51	281	2
6	65	464	2
7	75	387	2
8	162	964	3
9	113	706	3
10	61	329	2
11	48	290	2
12	59	311	1

#### d. Profiling Cluster

*Profiling cluster* digunakan untuk melihat nilai rata-rata dari anggota masing-masing variabel yang ada pada setiap klaster yang kemudian akan didapatkan karakteristik setiap klaster (Pratiwi, 2016).

*Profiling cluster* merupakan tahap interpretasi meliputi pengujian pada masing-masing klaster yang terbentuk untuk memberikan nama atau keterangan secara tepat sebagai gambaran sifat klaster serta menjelaskan bagaimana setiap klaster dapat berbeda secara relevan pada tiap dimensi. *Profiling cluster* dihitung menggunakan rata-rata setiap klaster pada setiap variabel (Laeli, 2014).

Contoh:

Dengan menggunakan hasil pengklasteran data simulasi pada tabel 3.21, maka perhitungan *profiling cluster* adalah sebagai berikut.

1. Hitung *profiling cluster* untuk klaster 1:

$$\text{Profiling cluster variabel } X = \frac{31 + 49 + 59}{3} = 46,33$$

$$\text{Profiling cluster variabel } Y = \frac{182 + 259 + 311}{3} = 250,67$$

Sehingga diperoleh tabel 3.22 berikut.

**Tabel 3.22 Profiling cluster untuk klaster 1**

Data ke <i>i</i>	<i>X</i>	<i>Y</i>
2	31	182
3	49	259
12	59	311
Rata-rata	46,33	250,67

2. Cara yang sama dapat dilakukan untuk perhitungan *profiling cluster* pada klaster 2 dan klaster 3 sehingga diperoleh hasil sebagai berikut.

**Tabel 3.23** *Profiling cluster untuk klaster 2*

Data ke <i>i</i>	X	Y
1	66	402
4	50	289
5	51	281
6	65	464
7	75	387
10	61	329
11	48	290
Rata-rata	59,43	348,86

**Tabel 3.24** *Profiling cluster untuk klaster 3*

Data ke <i>i</i>	X	Y
8	162	964
9	113	706
Rata-rata	137,5	835

3. Dari hasil perhitungan ketiga tabel tersebut maka hasil *profiling cluster* untuk pengklasteran data simulasi pada tabel 3.25 adalah sebagai berikut.

**Tabel 3.25** *Profiling cluster keseluruhan*

Variabel	Klaster 1	Klaster 2	Klaster 3
X	46,33	59,43	137,5
Y	250,67	348,86	835

Hasil *profiling cluster* pada tabel 3.25 menunjukkan bahwa klaster 3 unggul pada variabel X dan Y jika dibandingkan dengan klaster lain, sedangkan klaster 1 dan klaster 2 tidak memiliki variabel yang lebih unggul dibanding klaster lain.

### 3.5. Simpangan Baku

Untuk mengetahui metode pengklasteran yang mempunyai kinerja terbaik, dapat digunakan rata-rata simpangan baku dalam klaster dan simpangan baku antar klaster. Metode yang mempunyai rasio simpangan baku terkecil merupakan metode pengklasteran terbaik, di mana klaster yang baik adalah klaster yang mempunyai homogenitas yang tinggi antar anggota dalam satu klaster (*within cluster*) dan heterogenitas yang tinggi antar klaster (*between cluster*) (Laeli, 2014).

Contoh:

Dengan menggunakan hasil pengklasteran data simulasi pada tabel 3.21, maka langkah-langkah untuk menghitung rasio simpangan baku adalah sebagai berikut.

1. Lakukan perhitungan simpangan baku kluster ke  $k$  dengan persamaan sebagai berikut (Fithriyyah, 2017):

$$\sigma_k = \sqrt{\frac{1}{N} \sum_{i=1}^N (\bar{X}_i - \mu_k)^2} \dots\dots\dots(6)$$

dengan

$\sigma_k$  = simpangan baku kluster ke  $k$

$N$  = jumlah anggota kluster ke  $k$

$\bar{X}_i$  = rata-rata variabel data ke  $i$

$\mu_k$  = rata-rata kluster ke  $k$ .

Sehingga:

**Tabel 3.26** *Kluster 1*

Data ke $i$	$X$	$Y$	$\bar{X}_i$
2	31	182	106,5
3	49	259	154
12	59	311	185
Jumlah			445,5
$\mu_1$			148,5

$$\sigma_1 = \sqrt{\frac{1}{3} (106,5 - 148,5)^2 + (154 - 148,5)^2 + (185 - 148,5)^2} = 24,46$$

Cara yang sama digunakan untuk menghitung simpangan baku kluster ke 2 dan simpangan baku kluster ke 3 sehingga diperoleh:

$$\sigma_2 = 36,31$$

$$\sigma_3 = 76,75$$

2. Hitung simpangan baku dalam kluster dengan persamaan sebagai berikut (Fithriyyah, 2017):

$$\sigma_w = K^{-1} \sum_{k=1}^K \sigma_k \dots\dots\dots(7)$$

dengan

$\sigma_w$  = simpangan baku dalam klaster

$K$  = banyaknya klaster yang terbentuk

$\sigma_k$  = simpangan baku klaster ke  $k$ .

Sehingga:

$$\sigma_w = \frac{24,46 + 36,31 + 76,75}{3} = 45,84$$

3. Hitung simpangan baku antar klaster dengan persamaan berikut (Fithriyyah, 2017):

$$\sigma_b = \left[ (K)^{-1} \sum_{k=1}^K (\mu_k - \mu)^2 \right]^{\frac{1}{2}} \dots\dots\dots(8)$$

dengan

$\sigma_b$  = simpangan baku antar klaster

$K$  = banyaknya klaster yang terbentuk

$\mu_k$  = rata-rata klaster ke  $k$

$\mu$  = rata-rata keseluruhan klaster.

di mana

$$\mu = \frac{\sum \mu_k}{K} \dots\dots\dots(9)$$

Sehingga:

$$\mu = \frac{148,5 + 204,14 + 486,25}{3} = 279,63$$

Maka:

$$\sigma_b = \left[ (3)^{-1} \left( (148,5 - 279,63)^2 + (204,14 - 279,63)^2 + (486,25 - 279,63)^2 \right) \right]^{\frac{1}{2}} = 147,86$$

4. Langkah terakhir adalah perhitungan nilai rasio simpangan baku dengan persamaan berikut (Fithriyyah, 2017):

$$\sigma = \frac{\sigma_w}{\sigma_b} \times 100\% \dots\dots\dots(10)$$

dengan

$\sigma$  = rasio simpangan baku

$\sigma_w$  = simpangan baku dalam kluster

$\sigma_b$  = simpangan baku antar kluster.

Sehingga:

$$\sigma = \frac{45,84}{147,86} \times 100\% = 31,01\%$$