

BAB II

KAJIAN PUSTAKA

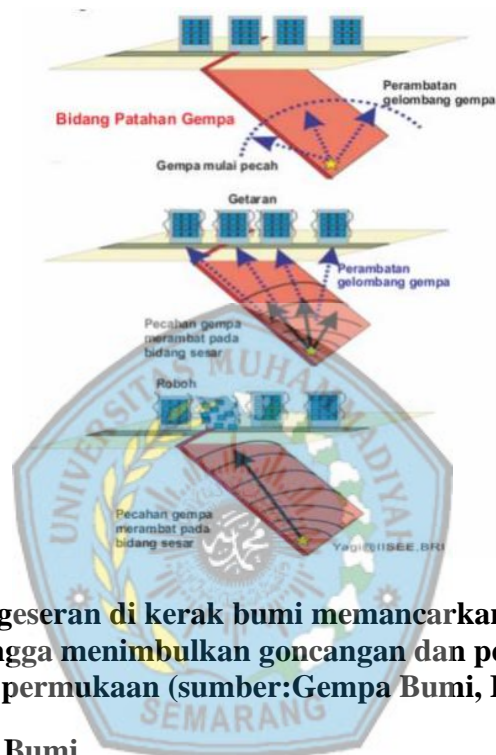
2.1 Pengertian Gempa Bumi

Gempabumi (*earthquake*) adalah peristiwa bergetar atau bergoncangnya bumi karena pergerakan/pergeseran lapisan batuan pada kulit bumi secara tiba-tiba akibat pergerakan lempeng-lempeng tektonik. Gempabumi yang disebabkan oleh aktivitas pergerakan lempeng tektonik disebut gempabumi tektonik. Namun selain itu, gempabumi bisa saja terjadi akibat aktifitas gunung berapi yang disebut sebagai gempabumi vulkanik.

Pergerakan tiba-tiba dari lapisan batuan di dalam bumi menghasilkan energi yang dipancarkan ke segala arah berupa gelombang gempabumi atau gelombang seismik. Ketika gelombang ini mencapai permukaan bumi, getarannya dapat merusak segala sesuatu di permukaan bumi seperti bangunan dan infrastruktur lainnya sehingga dapat menimbulkan korban jiwa dan harta benda. Berbeda dengan letusan gunung api dan bencana alam lain yang didahului dengan tanda-tanda atau gejala-gejala yang muncul sebelum kejadian, gempabumi selalu datang secara mendadak dan mengejutkan sehingga menimbulkan kepanikan umum yang luar biasa karena sama sekali tidak terduga sehingga tidak ada seorang pun yang sempat mempersiapkan diri.

Akibat yang ditimbulkan gempabumi luar biasa dahsyat karena mencakup wilayah yang sangat luas, menembus batas teritorial negara, bahkan antar-benua. Sifat getaran gempabumi yang sangat kuat dan merambat ke segala arah, mampu

menghancurkan bangunan-bangunan sipil yang terkuat sekalipun, sehingga tak ayal lagi sangat banyak memakan korban nyawa manusia. Bahkan gempabumi sering kali diikuti oleh bencana alam lanjutan yang jauh lebih dahsyat berupa tanah longsor dan gelombang tsunami yang akan di tunjukan pada gambar berikut:



Gambar 2. 1 Pergeseran di kerak bumi memancarkan radiasi gelombang gempabumi hingga menimbulkan guncangan dan perubahan struktur batuan di permukaan (sumber:Gempa Bumi, BMKG 2012)

2.2 Jalur Gempa Bumi

Gempabumi dapat terjadi kapan dan di mana saja. Meskipun demikian, konsentrasi gempabumi cenderung terjadi di tempat-tempat tertentu saja di lapisan kulit bumi. Lapisan kulit bumi terluar atau litosfer terdiri atas lempeng-lempeng tektonik yang kaku dan terapung di atas batuan yang relatif tidak kaku serta bergerak satu sama lain. Daerah pertemuan dua lempeng disebut sebagai plate margin atau batas lempeng, yang bisa berupa zona subduksi, pemekaran dasar samudra, atau pengangkatan, pelipatan, dll. di zona tumbukan. Gempabumi tidak

dapat terjadi di sembarang tempat, tetapi umumnya gempa bumi terjadi di sekitar batas lempeng, yang membentuk jalur gempa bumi dunia, dan sekitar sesar.

Jika dua lempeng bertemu pada suatu sesar, keduanya dapat bergerak saling menjauhi, saling mendekati, atau saling bergeser (Gambar 2.2). Umumnya, gerakan ini berlangsung lambat dan tidak dapat dirasakan oleh manusia namun terukur sebesar 0-15 cm per tahun.



Gambar 2. 2 Jenis Pergerakan Lempeng, yaitu: Saling menjauhi (A), Saling Mendekati(B) dan Saling Geser (C) (sumber : Gempa Bumi, BMKG 2012)

Terkadang gerakan lempeng ini macet dan saling mengunci sehingga terjadi pengumpulan energi yang berlangsung terus sampai pada suatu saat tidak mampu lagi menahan stress tersebut sehingga patah secara mendadak dan melepaskan energi dalam bentuk getaran yang kita kenal sebagai gempa bumi. Ada tiga kemungkinan pergerakan satu lempeng tektonik relatif terhadap lempeng lainnya, yaitu apabila kedua lempeng saling menjauhi (*spreading*), saling mendekati (*collision*) dan saling geser (*transform*).

Pada bawah lapisan kulit bumi terdapat lapisan mantel (selubung), yang suhunya jauh lebih panas. Lapisan ini sedemikian panasnya sehingga senantiasa dalam keadaan tidak kaku, dapat bergerak sesuai dengan proses pendistribusian panas yang kita kenal sebagai aliran konveksi. Aktivitas magma dalam mantel bisa juga mendesak sampai ke permukaan hingga membentuk rangkaian gunung berapi, yang dikenal dengan lingkaran api (*ring of fire*).

2.3 Penyebab Gempabumi

Berdasarkan atas penyebabnya gempabumi dapat dikelompokkan menjadi beberapa macam diantaranya: tektonik, vulkanik, runtuhan, jatuhnya meteor, dan gempabumi buatan manusia. Gempabumi tektonik adalah gempabumi yang disebabkan oleh pelepasan energi elastis yang tersimpan dalam lempeng tektonik. Karena adanya dinamika yang terjadi pada lapisan mantel bumi, lempeng tektonik bumi kita ini terus menerima energi dari lapisan tersebut. Lempeng tektonik adalah batuan yang bersifat elastis, sehingga energi yang diterima dari lapisan mantel tersimpan dalam bentuk energi elastis. Bila energi yang diterima sudah melebihi batas elastisitas lempeng tektonik, maka energi akan terlepas dalam bentuk deformasi plastis dan gelombang elastis. Daerah yang melepaskan energi elastis umumnya daerah yang lemah sehingga di daerah tersebut akan mengalami deformasi plastis, sedangkan daerah yang jauh dari sumber tersebut akan mengalami deformasi elastis dalam bentuk gelombang seismik.

Pada saat terjadi deformasi plastis di sekitar sumber gempabumi, fenomena yang dapat diamati dalam jangka waktu panjang adalah terjadi pergerakan dari lempeng tektonik dengan jenis pergerakan antara lain: penunjaman antara lempeng

samudra dan lempeng benua, tumbukan antara kedua lempeng benua, dan pergerakan lempeng samudera yang saling menjauh, serta pergerakan lempeng yang saling bergeser. Dikarenakan tepian lempeng yang tidak rata maka jika bergesekan maka, timbullah friksi. Friksi inilah yang kemudian melepaskan energi guncangan gempa bumi.

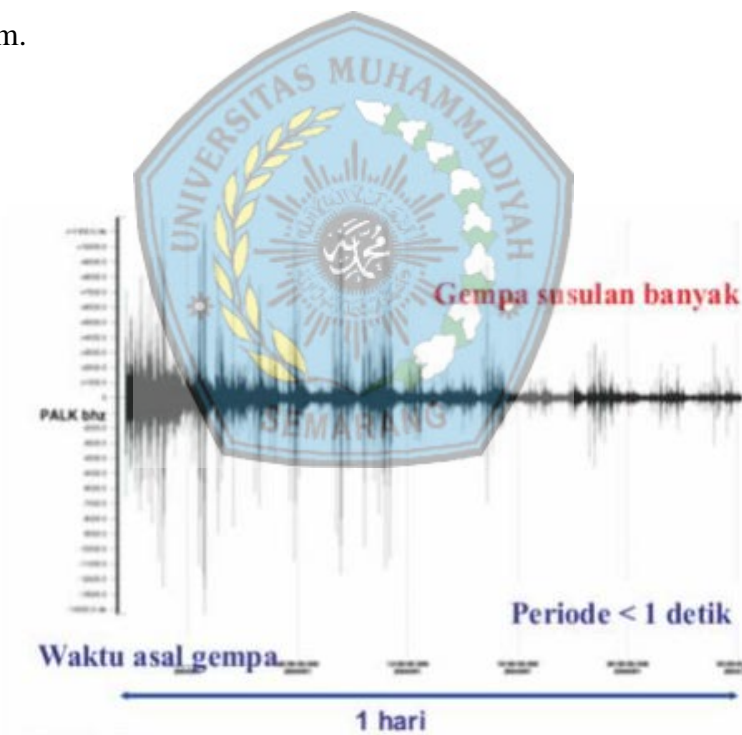
Gempabumi vulkanik adalah gempa bumi yang disebabkan oleh kegiatan gunung api. Magma yang berada pada kantong di bawah gunung tersebut mendapat tekanan dan melepaskan energinya secara tiba-tiba sehingga menimbulkan getaran tanah. Selain itu, pelepasan energi stress tersebut juga menyebabkan gerakan magma secara perlahan. Aktivitas gempa bumi tektonik dapat memicu aktivitas gempa bumi vulkanik. Naiknya magma ke permukaan dapat dipicu oleh pergeseran lempeng tektonik pada sesar bumi.

Biasanya ini terjadi pada batas lempeng tektonik yang bersifat konvergen (saling mendesak). Hanya saja pada gempa bumi vulkanik, efek guncangan lebih ditimbulkan karena desakan magma, sedangkan pada gempa bumi tektonik efek guncangan langsung ditimbulkan oleh benturan kedua lempeng tektonik. Bila lempeng tektonik yang terlibat adalah lempeng benua dengan lempeng samudera, maka akan terjadi deformasi di dasar laut yang kemudian menimbulkan tsunami karena batas lempengnya umumnya berada di dasar laut. Gempabumi runtuh adalah gempa bumi lokal yang terjadi apabila suatu gua di daerah batuan karst atau lokasi pertambangan runtuh. Sedangkan gempa bumi jatuhnya meteor akibat kejatuhan meteorit atau benda langit ke permukaan bumi. Hal ini pernah terjadi di kawasan Arizona, Amerika hingga meninggalkan bekas berupa lekukan tanah yang

cukup lebar seperti membentuk sebuah kawah. Gempabumi yang disebabkan oleh aktivitas dari manusia, yakni seperti peledakan dinamit, nuklir, ledakan bom, atau palu yang dipukulkan ke permukaan bumi.

2.4 Kedalaman dan Kekuatan Gempabumi

Kedalaman dan Kekuatan Gempabumi Fowler (1990) mengklasifikasikan gempabumi berdasarkan kedalaman fokus (*hypocentre*) sebagai berikut: gempabumi dangkal (*shallow*) kurang dari 70 km, gempabumi menengah (*intermediate*) kurang dari 300 km, dan gempabumi dalam (*deep*) lebih dari 300 km atau 450 km.



Gambar 2. 3 Seismogram digital vertikal di PALK Srilanka mencatat gempabumi susulan (sumber : Gempa Bumi, BMKG 2012)

Gempabumi dangkal menimbulkan efek guncangan dan kehancuran yang lebih dahsyat dibanding gempabumi dalam. Ini karena sumber gempabumi lebih dekat ke permukaan bumi sehingga energi gelombangnya lebih besar. Karena

pelemahan energi gelombang akibat perbedaan jarak sumber ke permukaan relatif kecil. Berdasarkan kekuatannya atau magnitudo (M) berskala Richter (SR) dapat dibedakan atas :

- a. Gempabumi sangat besar $M > 8$ SR
- b. Gempabumi besar $M 7 - 8$ SR
- c. Gempabumi merusak $M 5 - 6$ SR
- d. Gempabumi sedang $M 4 - 5$ SR.
- e. Gempabumi kecil $M 3 - 4$ SR
- f. Gempabumi mikro $M 1 - 3$ SR
- g. Gempabumi ultra mikro $M < 1$ SR.

Skala yang biasa digunakan oleh BMKG adalah Skala MMI (*Modified Mercalli Intensity*). Skala Mercalli adalah satuan untuk mengukur kekuatan gempa bumi. Satuan ini diciptakan oleh seorang vulkanologis dari Italia yang bernama Giuseppe Mercalli pada tahun 1902. Skala Mercalli terbagi menjadi 12 pecahan berdasarkan informasi dari orang-orang yang selamat dari gempa tersebut dan juga dengan melihat serta membandingkan tingkat kerusakan akibat gempa bumi tersebut. Oleh itu skala Mercalli adalah sangat subjektif dan kurang tepat dibanding dengan perhitungan magnitudo gempa yang lain. Oleh karena itu, saat ini penggunaan Skala Richter lebih luas digunakan untuk mengukur kekuatan gempa bumi. Tetapi skala Mercalli yang dimodifikasi, pada tahun 1931 oleh ahli seismologi Harry Wood dan Frank Neumann masih sering digunakan terutama apabila tidak terdapat peralatan seismometer yang dapat mengukur kekuatan gempa bumi di tempat kejadian.

Tabel 2. 1 Skala Intensitas Gempa Bumi BMKG

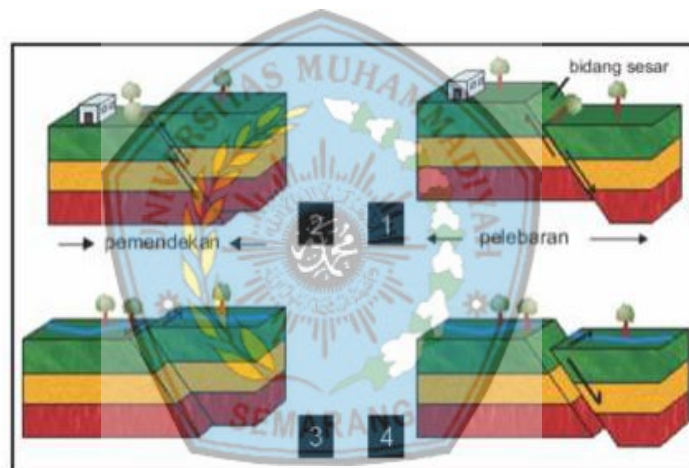
Skala SIG BMKG	Deskripsi Sederhana	Deskripsi Rinci	Skala MMI	PGA (gal)
I	TIDAK DIRASAKAN (<i>Non Felt</i>)	Tidak dirasakan atau dirasakan hanya oleh beberapa orang tetapi terekam oleh alat	I-III	< 2.9
II	DIRASAKAN (<i>Felt</i>)	Dirasakan oleh orang banyak tetapi tidak menimbulkan kerusakan. Benda-benda ringan yang digantung bergoyang dan jendela kaca bergetar	III-VI	2.9 - 88
III	KERUSAKAN RINGAN (<i>Slight Damage</i>)	Bagian non struktur bangunan mengalami kerusakan ringan, seperti retak rambut pada dinding, genteng bergeser ke bawah dan sebagian berjatuhan	VI	89- 167
IV	KERUSAKAN SEDANG (<i>Moderate Damage</i>)	Banyak retakan terjadi pada dinding bangunan sederhana, sebagian roboh, kaca pecah. Sebagian plester dinding lepas. Hampir sebagian besar genteng bergeser ke bawah atau jatuh. Struktur bangunan mengalami kerusakan ringan sampai sedang.	VII-VIII	168-564
V	KERUSAKAN BERAT (<i>Heavy Damage</i>)	Sebagian besar dinding bangunan permanen roboh. Struktur bangunan mengalami kerusakan berat. Rel kereta api melengkung	IX-XII	> 564

2.5 Jenis Gempabumi dan Urutan Kejadiannya

Berdasarkan proses kemunculan dan kesudahannya, Mogi membedakan gempabumi atas beberapa jenis, di antaranya:

- a. Gempabumi utama (*main shock*) langsung diikuti gempabumi susulan tanpa gempabumi pendahuluan (*fore shock*).

- b. Gempabumi sebelum terjadi gempa bumi utama diawali dengan adanya gempabumi pendahuluan dan selanjutnya diikuti oleh gempabumi susulan.
- c. Gempabumi terus-menerus dan dengan tidak terdapat gempa bumi utama yang signifikan disebut gempa bumi swarm. Biasanya dapat berlangsung cukup lama dan bisa mencapai 3 bulan atau lebih. Terjadi pada daerah vulkanik seperti di Gunung Lawu 1979, dan Kemiling, Bandar Lampung 2006. Berikut gambaran ilustrasi terjadinya gempa bumi menurut pergeseran sesar yang ditunjukkan pada gambar 2.4



Gambar 2. 4 Tipe pergerakan sesar gempa bumi (1) Turun, (2) Naik, (3) Mendatar, dan (4) Sesar Oblique (sumber : Gempa Bumi, BMKG 2012)

2.6 Definisi Data Mining

Data mining dikenal sejak tahun 1990-an, ketika adanya suatu pekerjaan yang memanfaatkan data menjadi suatu hal yang lebih penting dalam berbagai bidang, seperti marketing dan bisnis, sains dan teknik, serta seni dan hiburan. Sebagian ahli menyatakan bahwa data mining merupakan suatu langkah untuk menganalisis pengetahuan dalam basis data atau biasa disebut *Knowledge Discovery in Database (KDD)*. *Data mining* merupakan proses untuk menemukan

pola data dan pengetahuan yang menarik dari kumpulan data yang sangat besar. Sumber data dapat mencakup *database*, *data warehouse*, *web*, *repository*, atau data yang dialirkan ke dalam sistem dinamis.

Data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu (Larose, 2005):

1. Deskripsi

Terkadang peneliti dan analis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data. Sebagai contoh, petugas pengumpulan suara mungkin tidak dapat menemukan keterangan atau fakta bahwa siapa yang tidak cukup profesional akan sedikit didukung dalam pemilihan presiden. Deskripsi dari pola dan kecenderungan sering memberikan kemungkinan penjelasan untuk suatu pola atau kecenderungan.

2. Estimasi

Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah numerik daripada ke arah kategori. Model dibangun menggunakan record lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi. Selanjutnya pada peninjauan berikutnya, estimasi nilai dari variabel target dibuat berdasar nilai variabel prediksi. Sebagai contoh, akan dilakukan estimasi tekanan darah systolic dari pasien rumah sakit berdasarkan umur pasien, jenis kelamin, indeks berat badan dan level sodium darah. Hubungan antara tekanan darah systolic dan nilai variabel prediksi dalam proses pembelajaran akan menghasilkan model estimasi. Model estimasi yang dihasilkan dapat digunakan untuk kasus baru lainnya.

3. Prediksi

Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada di masa mendatang. Contoh dari prediksi dalam bisnis dan penelitian adalah:

- a. Prediksi harga beras dalam tiga bulan yang akan datang.
- b. Prediksi persentase kenaikan kecelakaan lalu lintas tahun depan jika batas bawah kecepatan dinaikan. Beberapa metode dan teknik yang digunakan dalam klasifikasi dan estimasi dapat pula digunakan (untuk keadaan yang tepat) untuk prediksi

4. Klasifikasi

Dalam klasifikasi, terdapat target variabel kategori. Contoh penggolongan pendapatan, dapat dipisahkan dalam tiga kategori yaitu pendapatan tinggi, pendapatan sedang dan pendapatan rendah. Contoh lain klasifikasi dalam bisnis dan penelitian adalah:

- a. Menentukan apakah suatu transaksi kartu kredit merupakan transaksi yang curang atau bukan.
- b. Memperkirakan apakah suatu pengajuan hipotek oleh nasabah merupakan suatu kredit yang baik atau buruk.
- c. Mendiagnosa penyakit seorang pasien untuk mendapatkan termasuk kategori penyakit apa.

5. Pengklusteran

Pengklusteran merupakan pengelompokan *record*, pengamatan, atau memperhatikan dan membentuk kelas obyek-obyek yang memiliki kemiripan. *Cluster* adalah kumpulan dari *record* yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidakmiripan dengan *record* dalam *cluster* lain. Pengklusteran berbeda dengan klasifikasi yaitu tidak adanya variabel target dalam pengklusteran. Pengklusteran tidak mencoba untuk melakukan klasifikasi, mengestimasi atau memprediksi nilai dari variabel target. Akan tetapi algoritma pengklusteran mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan, yang mana kemiripan *record* dalam satu kelompok akan bernilai maksimal sedangkan kemiripan dengan *record* dalam kelompok lain akan bernilai minimal. Contoh pengklusteran dalam bisnis dan penelitian adalah:

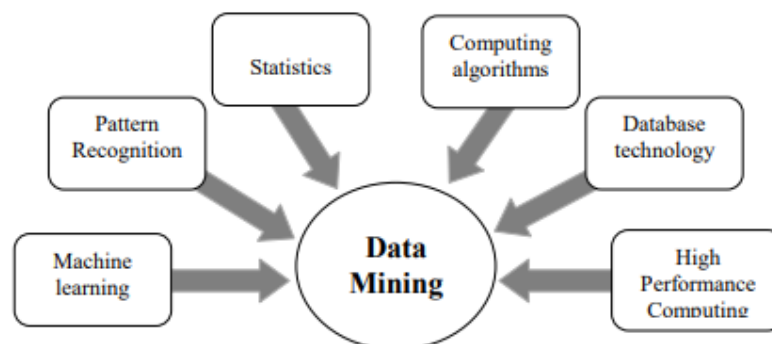
- a. Mendapatkan kelompok-kelompok konsumen untuk target pemasaran dari suatu produk bagi perusahaan yang tidak memiliki dana pemasaran yang besar.
- b. Untuk tujuan audit akuntansi, yaitu melakukan pemisahan terhadap perilaku financial dalam baik dan mencurigakan.
- c. Melakukan pengklusteran terhadap ekspresi dari gen, untuk mendapatkan kemiripan perilaku dari gen dalam jumlah besar.

6. Asosiasi

Tugas asosiasi dalam data mining adalah menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut sebagai analisis keranjang belanja. Contoh asosiasi dalam bisnis dan penelitian adalah:

- a. Meneliti jumlah pelanggan dari perusahaan telekomunikasi seluler yang diharapkan untuk memberikan respon positif terhadap penawaran upgrade layanan yang diberikan.
- b. Menemukan barang dalam supermarket yang dibeli secara bersamaan dan barang yang tidak pernah dibeli secara bersamaan

Data mining, secara sederhana merupakan suatu langkah ekstraksi untuk mendapatkan informasi penting yang sifatnya implisit dan belum diketahui. Selain itu, data mining mempunyai hubungan dengan berbagai bidang diantaranya statistik, *machine learning* (pembelajaran mesin), *pattern recognition*, *computing algorithms*, *database technology*, dan *high performance computing*. Diagram hubungan data mining disajikan pada Gambar 2.5



Gambar 2. 5 Diagram Hubungan Data Mining(Sumber: Helilintar Rika, dkk;2017)

Secara sistematis, langkah utama untuk melakukan data mining terdiri dari tiga tahap, yaitu sebagai berikut (Gonunescu, 2011);

1. Eksplorasi atau pemrosesan awal data Eksplorasi atau pemrosesan awal data terdiri dari pembersihan data, normalisasi data, transformasi data, penanganan missing value, reduksi dimensi, pemilihan subset fitur, dan sebagainya.
2. Membangun model dan validasi Membangun model dan validasi, yaitu melakukan analisis dari berbagai model dan memilih model sehingga menghasilkan kinerja yang terbaik. Pembangunan model dilakukan menggunakan metode-metode seperti klasifikasi, regresi, analisis cluster, dan asosiasi.
3. Penerapan Penerapan dilakukan dengan menerapkan model yang dipilih pada data yang baru untuk menghasilkan kinerja yang baik pada masalah yang diinvestigasi. Tahapan proses data mining ada beberapa yang sesuai dengan proses KDD adalah sebagai berikut:

a. Data Cleaning (Pembersihan Data)

Data cleaning (Pembersihan data) adalah proses yang dilakukan untuk menghilangkan noise pada data yang tidak konsisten atau bisa disebut tidak relevan. Data yang diperoleh dari database suatu perusahaan maupun hasil eksperimen yang sudah ada, tidak semuanya memiliki isian yang sempurna misalnya data yang hilang, data yang tidak valid, atau bisa juga hanya sekedar salah ketik. Data yang

tidak relevan itu dapat ditangani dengan cara dibuang atau sering disebut dengan proses *cleaning*. Proses *cleaning* dapat berpengaruh terhadap performa dari teknik *data mining*.

b. Data Integration (Integrasi data)

Integrasi data merupakan proses penggabungan data dari berbagai database sehingga menjadi satu database baru. Data yang diperlukan pada proses *data mining* tidak hanya berasal dari satu *database* tetapi juga dapat berasal dari beberapa *database*.

c. Data Selection (Seleksi Data)

Tidak semua data yang terdapat dalam database akan dipakai, karena hanya data yang sesuai saja yang akan dianalisis dan diambil dari database. Misalnya pada sebuah kasus *market basket analysis* yang akan meneliti faktor kecenderungan pelanggan, maka tidak perlu mengambil nama pelanggan, cukup dengan id pelanggan saja.

d. Data Transformation (Transformasi Data)

Transformasi data merupakan proses perubahan data dan penggabungan data ke dalam format tertentu. Data mining membutuhkan format data khusus sebelum diaplikasikan. Misalnya metode standar seperti analisis asosiasi dan *clustering* hanya bisa menerima input data yang bersifat kategorikal. Karenanya data yang berupa angka numerik apabila mempunyai sifat kontinyu perlu dibagi-bagi menjadi beberapa interval. Proses ini sering disebut dengan transformasi data.

Untuk melakukan normalisasi, salah satu rumus yang dapat digunakan yaitu

$$X = \frac{x - \min(x_i, x_j)}{\max(x_i, x_j) - \min(x_i, x_j)} \quad (2.1)$$

X = nilai setelah normalisasi

x = nilai pada atribut y

$\min(x_i, x_j)$ = nilai terkecil pada atribut y

$\max(x_i, x_j)$ = nilai terbesar pada atribut y

e. *Pattern Evaluation* (Evaluasi Pola)

Evaluasi pola bertugas untuk mengidentifikasi pola-pola yang menarik ke dalam *knowledge based* yang ditemukan. Pada tahap ini dihasilkan polapola yang khas dari model klasifikasi yang dievaluasi untuk menilai apakah hipotesa yang ada memang tercapai. Bila ternyata hasil yang diperoleh tidak sesuai dengan hipotesa, terdapat beberapa alternatif yang bisa diambil seperti menjadikannya umpan balik untuk memperbaiki proses data mining, atau mencoba metode data mining lain yang lebih sesuai.

f. *Knowledge Presentation* (Presentasi Pengetahuan)

Knowledge presentation merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan atau informasi yang telah digali oleh pengguna. Tahap terakhir dari proses data mining adalah memformulasikan keputusan dari hasil analisis yang didapat.

2.7 Ukuran Kemiripan

Untuk mengukur kemiripan antara dua objek digunakan ukuran jarak. Semakin besar nilai jarak maka semakin besar perbedaan antar kedua objek.

Menurut Johnson and Wichern (2002), ukuran kemiripan yang digunakan dalam *clustering* adalah jarak Euclidean dan jarak manhattan. Jarak Euclidean merupakan jarak terpendek antara dua titik. Jarak Euclidean ditulis sebagai

$$d_{\text{euc}}(X_{ij}, C_{kj}) = \sqrt{\sum_{j=1}^p \sum_{i=1}^n (X_{ij} - C_{kj})^2} \quad (2.2)$$

X_{ij} = Objek pengamatan ke-i pada variable ke-j

C_{kj} = Pusat Cluster ke- k pada variable ke-j

p = Banyaknya variabel yang diamarti

n = Banyaknya pengamatan

2.8 CLARA (*Clustering Large Application*)

Pada kasus dengan dataset yang besar, PAM atau *K-Medoids* tidak akan berjalan dengan baik. Maka metode berbasis sampling yang disebut dengan CLARA (*Clustering Large Application*) yang di perkenalkan oleh (Kaufman and Rousseeuw,1990) bisa di gunakan. Dalam hal ini CLARA menggunakan sample dataset secara acak. Dalam banyak kasus , sample yang semakin besar akan bekerja dengan baik, sehingga setiap objek memiliki probabilitas yang sama untuk dipilih sebagai sample. Objek-objek yang dipilih menjadi pusat *cluster (medoids)* akan cenderung mirip dengan yang sudah dipilih dari seluruh dataset.

K-medoid adalah teknik partisi klasik *clustering* yang mengcluster kumpulan data objek n ke dalam kelompok k dikenal dengan apriori. Sebuah alat yang berguna untuk menentukan k adalah *silhoutte*. Hal ini lebih kuat *noise* dan *outliernya* dibandingkan dengan *k-means* karena meminimalkan sejumlah

dissimilarities berpasangan bukannya jumlah kuadrat jarak Euclidean seperti persamaan (2.2).

CLARA membuat clustering dari banyak sampel secara acak dan menghasilkan clustering terbaik sebagai outputnya. Kompleksitas dalam menghitung *medoids* pada sample acak adalah sebagai berikut

$$O(k^2s + k(n-k)) \quad (2.3)$$

s = Ukuran Sampel

k = Jumlah *Cluster*

n = Total Objek.

Pada algoritma PAM mencari *K-Medoids* terbaik di antara dataset, tetapi CLARA mencari *k-medoids* terbaik di antara sample dataset yang terpilih. CLARA tidak bisa menghasilkan *clustering* yang baik jika *medoids* terbaik yang di disampelkan sangat jauh dari *k-medoids* terbaik. Jika objek merupakan salah satu dari *k-medoids* terbaik tetapi tidak dipilih *k-medoids* selama proses sampling maka CLARA tidak akan pernah menghasilkan *clustering* terbaik (Han , Kamber and Pei, 2012).

Adapun algoritma CLARA untuk penentuan *cluster* adalah sebagai berikut:

1. Menentukan jumlah *cluster* k yang ingin dibentuk.
2. Melakukan pengelompokan data dengan algoritma CLARA:
 - a. Memilih secara acak pusat *cluster* awal (*medoid*) sebanyak k , pada data subsample dengan ketentuan seperti persamaan 2.3
 - b. Menghitung jarak setiap objek terhadap masing-masing *medoid* pada tiap *cluster* dan menempatkan tiap objek tersebut ke *medoid* terdekat,

- c. Menghitung total jarak dengan menggunakan jarak euclidean seperti persamaan 2.2.
- d. Memilih kandidat *medoid* baru,
- e. Menghitung jarak setiap objek terhadap kandidat *medoid* baru dan menempatkan tiap objek tersebut ke kandidat *medoid* terdekat,
- f. Menghitung total jarak euclidean seperti persamaan 2.2
- g. Menghitung selisih total jarak (s), yaitu selisih dari total jarak pada kandidat *medoid* baru dengan total jarak pada *medoid* lama,
- h. Jika diperoleh nilai $s < 0$, maka kandidat *medoid* tersebut menjadi *medoid* baru,
- i. Ulangi langkah d sampai h hingga tidak terjadi perubahan *medoid*.

2.9 DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*)

DBSCAN diperkenalkan oleh Ester, algoritma ini merupakan teknik *clustering* non-parametrik (Tran et al, 2012). DBSCAN merupakan salah satu algoritma *clustering* yang mengelompokkan obyek ke dalam *cluster* dimana kategori atau *class* data ujinya belum diketahui (*unsupervised learning*). Algoritma ini cukup populer dan mendapatkan penghargaan *the test of the time award*, yaitu penghargaan yang diberikan untuk algoritma yang memperoleh perhatian khusus dalam hal teori maupun prakteknya, pada konferensi tentang data mining yang diselenggarakan oleh *Assosiacion for Computing Machinery (ACM)* pada tahun 2014 (ACM, 2014 dalam Furqon dan Muflikhah, 2016).

DBSCAN adalah algoritma pengelompokan yang didasarkan pada kepadatan (*density*) data. Konsep kepadatan yang dimaksud dalam DBSCAN

adalah jumlah data yang berada dalam radius Eps (ϵ) dari setiap data. Jika jumlah data dalam radius ϵ lebih dari atau sama dengan $MinPts$ (jumlah minimal data dalam radius ϵ), data tersebut masuk dalam kategori kepadatan yang diinginkan, jumlah data dalam radius tersebut termasuk data itu sendiri (Prasetyo, 2012).

Metode ini melakukan pengelompokan dengan baik pada data berkepadatan tinggi, dan bisa menemukan bentuk sembarang kelompok dengan baik. DBSCAN dapat memisahkan data berkepadatan tinggi dan data berkepadatan rendah karena algoritma ini berbasis kepadatan data. Dalam Algoritma ini dikenal beberapa istilah seperti berikut (Arsih et al, 2016):

1. *Core*: Titik pusat dalam cluster didasarkan pada densitas dimana ada sejumlah titik yang harus berada dalam Eps (radius atau nilai ambang batas). $MinPts$ (minimal titik dalam *cluster*) yang ditentukan pengguna.
2. *Border*: Titik yang menjadi batasan dalam kawasan titik pusat (*core*).
3. *Noise*: Titik yang tidak dapat dijangkau oleh *core* dan bukan merupakan *border*.

$$Noise = \{ x \in X \mid \forall i : x \notin C_i \} \quad (2.4)$$

Dimana X merupakan gugus data, dan C_i merupakan *cluster* ke-1.

4. Densitas terjangkau langsung: Sebuah titik dikatakan titik terjangkau langsung apabila titik tersebut terhubung secara langsung dengan titik pusat (*core*)

$$x \in N_{Eps}(y) \wedge |N_{Eps}(y)| \geq MinPts \quad (2.5)$$

$N_{Eps}(y)$ = titik sekitar y dalam radius Eps

$MinPts$ = minimal titik dalam cluster

5. Densitas terjangkau: Sebuah titik dikatakan titik terjangkau apabila titik tersebut terhubung secara tidak langsung dengan titik pusat (*core*).
6. Densitas terhubung: Sebuah titik dikatakan saling terhubung satu sama lain oleh titik lain. DBSCAN memerlukan dua parameter input yaitu *epsilon* (*Eps*) dan titik minimum (*MinPts*). Eps-titik sekitar didefinisikan sebagai:

$$N_{Eps}(x) = \{y \in D \mid dist(x, y) \leq Eps\} \quad (2.6)$$

$Eps(x)$ = titik sekitar dari x dalam radius Eps

D = gugus data dist

$dist(x, y)$ = jarak *Euclidean* dari objek x dan y

Eps = radius atau ambang batas.

Untuk mengelompokkan data, hal yang harus dilakukan adalah mendefinisikan nilai *minpts* dan *epsilon*. DBSCAN secara umum dimulai dengan titik awal secara acak.

Langkah-langkah metode DBSCAN:

1. Tentukan titik awal atau p secara acak.
2. Hitung semua jarak titik menggunakan persamaan 2.2

$$d_{\text{euc}}(X_{ij}, C_{kj}) = \sqrt{\sum_{j=1}^p \sum_{i=1}^n (X_{ij} - C_{kj})^2}$$

3. Tentukan nilai *epsilon* dan *minimum points*. Dalam menentukan nilai *epsilon* dan *minpts* terbaik harus dilakukan berkali-kali percobaan. Dari percobaan tersebut nilai *epsilon* dan *minpts* yang dipilih adalah nilai yang menghasilkan nilai *silhouette index* yang tinggi, dan jumlah *cluster* yang banyak.

Keuntungan dari algoritma ini diantaranya (Mumtaz, K, and Duraiswamy, 2010)

1. DBSCAN tidak memerlukan informasi tentang banyaknya kelompok yang akan dibentuk.
2. DBSCAN hanya membutuhkan dua parameter yang sebagian besar tidak sensitif terhadap basis data.

2.10 *Silhouette Index (SI/ Coeficient Silhouette)*

SI dapat digunakan untuk memvalidasi baik sebuah data, *cluster* tunggal (satu *cluster* dari sejumlah *cluster*), atau bahkan keseluruhan *cluster*. Untuk menghitung nilai SI dari sebuah data ke-*i*, ada 2 komponen yaitu a_i dan b_i . a_i adalah rata-rata jarak data ke-*i* terhadap semua data lainnya dalam satu *cluster*, sedangkan b_i^j didapatkan dengan menghitung rata-rata jarak data ke-*i* terhadap semua data dari *cluster* yang lain tidak dalam satu *cluster* dengan data ke-*i*, kemudian diambil yang terkecil (Tan et al, 2006, dalam Prasetyo, 2014). Berikut rumus untuk menghitung nilai a_j^i :

$$a_j^i = \frac{1}{m_j - 1} \sum_{\substack{r=1 \\ r \neq i}}^{m_j} d(X_j^i, X_r^i) \quad i = 1, 2, \dots, m_j \quad (2.7)$$

$d(X_j^i, X_r^i)$ adalah jarak data ke-*i* dengan data ke-*r* dalam satu *cluster* *j*, sedangkan m_j adalah jumlah data dalam *cluster* ke-*j*. Berikut rumus untuk menghitung nilai b_i^j :

$$b_i^j = \min_{\substack{n=1, \dots, k \\ n \neq j}} \left\{ \frac{1}{m_n} \sum_{\substack{r=1 \\ r \neq i}}^{m_n} d(X_i^j, X_r^n) \right\} \quad i = 1, 2, \dots, m_n \quad (2.8)$$

Untuk mendapatkan *Silhouette Index* (SI) data ke-*i* menggunakan persamaan 2.9

$$SI_j^i = \frac{b_i^j - a_i^j}{\max\{a_i^j, b_1^j\}} \quad (2.9)$$

Nilai a_i mengukur seberapa tidak mirip sebuah data dengan *cluster* yang diikutinya, nilai yang semakin kecil menunjukkan semakin tepat data tersebut berada di *cluster* tersebut. Nilai b_i yang besar menunjukkan seberapa jeleknya data terhadap *cluster* lain. Nilai SI yang didapat dalam rentang $[-1,+1]$. Nilai SI yang mendekati 1 menunjukkan bahwa data tersebut semakin tepat berada dalam *cluster* tersebut. Nilai SI negatif menunjukkan bahwa data tersebut tidak tepat berada dalam *cluster* tersebut. SI bernilai 0 (atau mendekati nol) menunjukkan data tersebut posisinya berada di perbatasan di antara dua *cluster*. Berikut persamaan untuk menghitung nilai SI dari sebuah *cluster*:

$$SI_j = \frac{1}{m_j} \sum_{i=1}^{m_j} SI_i^j \quad (2.10)$$

Nilai SI global diperoleh dengan persamaan 2.11

$$SI = \frac{1}{k} \sum_{j=1}^k SI_j \quad (2.11)$$

k adalah jumlah *cluster*.