BAB II
LANDASAN TEORI

2.1. Pengertian *Data mining*

Data mining merupakan suatu kegiatan yang meliputi pengumpulan, pemakaian dan historis untuk menentukan keteraturan, pola atau hubungan dalam set data berukuran besar. Salah satu tugas utama dari *data mining* adalah pengelompokan clustering dimana data yang dikelompokkan belum mempunyai contoh kelompok. *Data mining*, sering juga disebut sebagai *Knowledge Discovery in Database* (KDD). KDD adalah kegiatan yang meliputi pengumpulan, pemakaian data, historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar (Santosa, 2007).

Secara sederhana, *data mining* dapat diartikan sebagai proses mengekstrak atau "menggali" pengetahuan yang ada pada sekumpulan data. Banyak orang yang setuju bahwa *data mining* adalah sinonim dari *Knowledge-Discovery in Database* atau yang biasa disebut KDD. Dari sudut pandang yang lain, *data mining* dianggap sebagai satu langkah yang penting didalam proses KDD.

Menurut Eko Prasetyo, (Prasetyo, 2012), jika dilacak akar keilmuannya, *data mining* mempunyai empat akar bidang ilmu, sebagai berikut :

1. Statistik

   Bidang ini merupakan akar paling tua, tanpa ada statistik maka *data mining* mungkin tidak ada. Dengan menggunakan statistik klasik ternyata data yang diolah dapat diringkas dalam apa yang umun dikenal sebagai *Exploratory Data Analysis* (EDA). EDA berguna untuk mengidentifikasi hubungan sistematis antar variabel/fitur ketika tidak ada cukup informasi alami yang dibawanya. Teknik EDA klasik yang digunakan dalam *data mining* diantaranya :

a. Metode komputasional : statistik deskriptif (distribusi, parameter statistik klasik (mean, median, rata-rata, varian, dan sebagainya), korelasi, tabel frekuensi, teknik eksplorasi multivariate (analisis cluster, analisis faktor, analisis komponen utama dan klasifikasi, analisis kanonik, analisis diskriminan, *classification tree*, analisis korespondensi), model linear/nonlinear lanjutan (regresi linear/nonlinear, *time series/forecasting*, dan sebagainya).
b. Visualisasi data : mengarah pada representasi informasi dalam bentuk visual dan dapat dipandang sebagai satu yang paling berguna. Teknik visualisasi yang paling umum dikenal adalah histogram semua jenis (kolom, silinder, kerucut, piramida, lingkaran, batang, dan sebagainya), kotak, scatter, kontur, matriks, ikon, dan sebagainya.

2. Kecerdasan buatan atau artificial intelligence (AI)

   Bidang ilmu ini berbeda dengan statistik. Teorinya dibangun berdasarkan teknik heuristik sehingga AI berkontribusi terhadap teknik pengolahan informasi berdasarkan pada model penalaran manusia. Salah satu cabang dari AI, yaitu mesin atau *machine learning*.

3. Pengenalan pola

   Sebenarnya *data mining* juga menjadi turunan bidang pengenalan pola, tetapi hanya mengola data dari basis data. Data yang diambil dari basis data untuk diolah bukan dalam bentuk relasi, melainkan dalam bentuk normal pertama sehingga set data dibentuk menjadi bentuk normal pertama. Kan tetapi, *data mining* mempunyai ciri khas yaitu pencarian pola asosiasi dan pola skuensial.

4. Sistem basis data.

   Akar bidang ilmu ini menyediakan informasi berupa data yang akan ‘digali’ menggunakan metode – metode yang disebutkan sebelumnya.

   Kebutuhan ‘penggalian’ informasi dalam data dapat dilihat pada kasus dunia nyata, diantaranya sebagai berikut :

a. Ekonomi :jumlah data yang sangat besat yang dikumpulkan dari berbagai bidang seperti data web, *e-commerce, supermarket*, transaksi keuangan dan perbankan, dan sebagainya yang siap dianalisis dengan tujuan untuk mendapatkan keputusan yang optimal.
b. Pelayanan kesehatan : ada banyak basis data berbeda dalam bidang pelayanan kesehatan (medis dan farmasi), yang dianalisis secara parsial, khususnya dengan cara medis sendiri, padahal sebenarnya dalam data tersembunyi banyak informasi yang belum dibuka secara tepat.
c. Riset pengetahuan : ada basis data besar yang dikumpulkan bertahun – tahun dalam bermacam – macam bidang (astronomi, meteorologi, biologi, linguistik, dan sebagainya) yang tidak dapat dieksplorasi menggunakan cara tradisional.

2.2 Proses *data mining*

Secara sistematis, ada tiga langkah utama dalam *data mining* :

1. Eksplorasi pemrosesan awal

   Eksplorasi data terdiri dari ‘pembersihan data’ normalisasi data, transformasi data, penanganan data yang salah, reduksi dimensi, pemilihan subset fitur, dan sebagainya.
2. Membangun model dan melakukan validasi terhadapnya

   Maksudnya melakukan analisis berbagai model dan memilih model kinerja prediksi yang terbaik. Dalam langkah ini, digunakan metode-metode seperti, klasifikasi, regresi, analisis cluster, deteksi anomali, analisis asosiasi, analisis pola skuensial, dan sebagainya.
3. Penerapan

   Penerapan berarti menerapkan model pada data yang baru untuk menghasilkan perkiraan/prediksi masalah yang diinvestigasi.

2.3 Pengelompokan *Data mining*

Data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu (Kusrini dan Emha Taufiq Luthfi, 2009):

1. Deskripsi

   Terkadang peneliti dan analis secara sederhana ingin mencoba mencari data untuk menggambarkan pola dan kecenderungan yang terdapat dalam data. Sebagai contoh, petugas pengumpulan suara mungkin tidak dapat menentukan keterangan atau fakta bahwa siapa yang tidak cukup profesional akan sedikit didukung dalam pemilihan presiden.

2. Estimasi

   Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih kearah numerik dari pada kearah kategori. Model dibangun menggunakan *record* lengkap yang menyediakan nilai dari variabel target sebagai prediksi. Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel predikasi. Sebagai contoh akan dilakukan estimasi tekanan darah sistolik pada pasien rumah sakit berdasarkan umur pasien, jenis kelamin, indeks berat badan, dan level sodium darah. Hubungan antara tekanan darah sistolik 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 prediksi bisnis dan penelitian adalah:

   a. Prediksi harga beras dalam tiga bulan yang akan datang.
   b. Prediksi persentasi kenaikan kecelakaan lalu lintas tahun depan jika batas bawah kecepatan dinaikkan.

   Beberapa metode dan teknik yang digunakan dalam klasifikasi dan estimasi dapat pula digunakan (untuk keadaan yang tepat) untuk prediksi.

4. Klasifikasi

   Klasifikasi adalah fungsi pembelajaran yang memetakan (mengklasifikasi) sebuah unsur (*item*) data ke dalam salah satu dari beberapa kelas yang sudah didefinisikan. Contoh lain klasifikasi dalam bisnis dan penelitian adalah :

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

5. Pengklusteran *(Clustering)*

   Pengelompokan (*clustering*) merupakan tugas deskripsi yang banyak digunakan dalam mengidentifikasi sebuah himpunan terbatas pada kategori atau *cluster* untuk mendeskripsikan data yang ditelaah. Kategori-kategori ini dapat bersifat eksklusif dan ekshaustif mutual, atau mengandung representasi yang lebih kaya seperti kategori yang hirarkis atau saling menumpu (*overlapping*). Contoh pengklusteran dalam bisnis dan penelitian adalah:

   a. Mendapatkan kelompok-kelompok konsumen untuk target pemasaran dari satu 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 *attribut* yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut 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. Menentukan barang dalam supermarket yang dibeli secara bersamaan dan yang tidak pernah dibeli secara bersamaan.

2.4 Clustering

Clustering atau klasterisasi adalah metode pengelompokan data. Menurut (Tan, 2006) *clustering* adalah sebuah proses untuk mengelompokkan data ke dalam beberapa *cluster* atau kelompok sehingga data dalam satu *cluster* memiliki tingkat kemiripan yang maksimum dan data antar *cluster* memiliki kemiripan yang minimum. *Clustering* merupakan proses partisi satu set objek data ke dalam himpunan bagian yang disebut dengan *cluster*. Objek yang di dalam *cluster* memiliki kemiripan karakteristik antar satu sama lainnya dan berbeda dengan *cluster* yang lain. Partisi tidak dilakukan secara manual melainkan dengan suatu algoritma *clustering*.

Salah satu cara mengelompokkan data yang efektif adalah dengan menggunakan teknik *data mining clustering*, salah satu metode dari *clustering* adalah *K-means*, *K-means* merupakan salah satu metode dari *clustering* nonhirarki yang berusaha mempartisi data yang ada ke dalam bentuk dua atau lebih kelompok. Metode ini mempartisi data ke dalam kelompok sehingga data berkarakteristik berbeda dikelompokkan ke dalam kelompok yang lain. Adapun tujuan pengelompokan data ini adalah untuk meminimalkan variasi dalam suatu kelompok dan memaksimalkan variasi antar kelompok.

2.5 Algoritma *K-means*

Pengelompokan *K-means* merupakan metode analisis kelompok yang mengarah pada pemartisian N objek pengamatan kedalam K kelompok *(cluster)* di mana setiap objek pengamatan dimiliki oleh sebuah kelompok dengan *mean* (rata-rata) terdekat. (Prasetyo, 2012)

K-means merupakan salah satu metode pengelompokan data nonhirarki yang berusaha mempartisi data yang ada ke dalam bentuk dua atau lebih kelompok. Metode ini mempartisi data ke dalam kelompok sehingga data berkarakteristik berbeda dikelompokkan ke dalam kelompok yang lain. Adapun tujuan pengelompokan data ini adalah untuk meminimalkan variasi dalam suatu kelompok dan memaksimalkan variasi antar kelompok. Adapun bebrapa kelebihan pada algoritma *K-means*, yaitu :

1. Mudah untuk diimplementasikan dan dijalankan.
2. Waktu yang dibutuhkan untuk menjalankan pembelajaran ini relatif cepat.
3. Mudah untuk diadaptasi.
4. Umun digunakan.

Algoritma *K-means* :

1. Menentukan nilai K sebagai jumlah *cluster* yang ingin dibentuk. Jumlah *cluster* yang akan dibentuk akan ditentukan sendiri oleh pengguna *system*.
2. Membangkitkan K *centroid* (titik pusat cluster) secara random. Dalam menentukan n buah pusat *cluster* awal dilakukan pembangkitan secara random yang mempresentasikan urutan data input. Pusat awal *cluster* di dapatkan dari data sendiri bukan dengan menentukan titik baru, yaitu dengan merandom pusat awal dari data.
3. Menghitung jarak antara data dengan pusat *cluster* digunakan rumus *Euclidean distance*.
    1. Ambil nilai data dan nilai titik pusat *cluster*.
    2. Hitung jarak data ke pusat *cluster*.

$$D(x_2, x_1) = \left||x_2 - x_1|\right| = \sqrt{\sum_{j=1}^{p} |x_{2j} - x_{1j}|^2} \text{.....(2.1)}$$

Keterangan :

$D(x_2, x_1)$ = jarak antara data dengan pusat *cluster*

x_{2j} = nilai data dua ke-j

x_{1j} = nilai data pertama ke-*j*

4. Cari jarak terdekat dan masukkan X kedalam *cluster* sesuai dengan centroid tersebut. Jarak hasil perhitungan akan dilakukan perbandingan dan dipilih jarak terdekat (terkecil) Antara data dan pusat *cluster*. Jarak ini menunjukkan bahwa data tersebut berada dalam satu kelompok dengan pusat *cluster* terdekat.

   Algoritma pengelompokan data :

   1. Ambil nilai jarak tiap pusat *cluster* dengan data.
   2. Cari nilai jarak yang terkecil.
   3. Kelompokkan data dengan pusat *cluster* yang memiliki jarak terkecil.
5. Menentukan posisi *centroid* baru dengan cara mengitung rata-rata dari data-data yang terpilih pada *centroid* yang sama. Untuk mendapatkan pusat *cluster* baru, bisa dihitung dengan rata-rata nilai anggota *cluster* yang baru.

   Algoritma penentuan pusat *cluster* baru :

   1. Cari jumlah anggota tiap *cluster*.
   2. Hitung pusat baru dengan rumus :

      $$v_{ij} = \frac{1}{N}\sum_{k=0}^{N_i} X_{kj} \quad \text{.....(2.2)}$$

      Dimana :

      v_{ij} = *centroid*/rata-rata cluster ke-*i* untuk variabel ke-*j*

      N_i = jumlah data yang menjadi anggota *cluster* ke-*i*

      i, k = indeks dari *cluster*

      j = indeks dari variabel

      X_{kj} = nilai data ke-*k* yang ada didalam *cluster* tersebut untuk variabel ke-*j*
6. Lakukan langkah 3 – 5 hingga posisi anggota *cluster* baru dengan anggota *cluster* lama tidak berubah. Pusat *cluster* yang baru digunakan untuk melakukan perhitungan iterasi selanjutnya, jika hasil yang didapatkan belum konvergen, dan data akan berhenti jika hasilnya yang dicapai sudah konvergen (pusat *cluster* baru sama dengan pusat *cluster*

lama) atau apabila ada perubahan nilai *centroid* diatas nilai ambang atau nilai pada fungsi objektif yang telah ditentukan. Dimana nilai ambang (*threshold*) adalah 0.0000 < 1.

2.6 Evaluasi *Cluster Davies Bouldin Index*

Evaluasi *cluster* yang akan digunakan dalam sistem ini adalah evaluasi validitas internal, yakni dalam evaluasi hasil *cluster* tanpa menggunakan informasi dari luar/eksternal. Evaluasi validitas *cluster* akan membandingkan hasil *cluster* dengan nilai K=3 dengan metode validitas *Davies Bouldin Index* yang diperkenalkan oleh David L. Davies dan Donald W. Bouldin pada tahun 1972. Dimana nilai DBI yang terkecil maka *cluster* tersebut yang paling bagus / valid.

Pendekatan perhitungan validitas Davies Bouldin Index ini untuk memaksimalkan jarak *inter-cluster* di antara *Cluster Ci* dan *C j* atau *Sum-of-square-between-cluster* (SSB) dan pada waktu yang sama mencoba untuk meminimalkan jarak antara titik dalam sebuah *cluster* atau *Sum-of-square-within-cluster* (SSW). Dimana rumus SSW dan SSB sebagai berikut :

$$SSB_{i,j} = d(c_i, c_j) \quad \text{..... (2.3)}$$

Keterangan :

$SSB_{i,j}$ = maksimal jarak diantara *Cluster Ci* dan *Cj*

$d(c_i, c_j)$ = jarak antara *cluster* ke-*i* dengan *cluster* ke-*j*

$$SSW_i = \frac{1}{m_i} \sum_{j=1}^{m_i} d(x_j, c_i) \quad \text{..... (2.4)}$$

Keterangan :

SSW_i = minimal jarak antara titik *cluster* ke-*i*

$d(x_j, c_i)$ = jarak antara data ke-*j* dengan *centroid cluster* ke-*i*

m = jumlah yang berada dalam *cluster* ke-*i*

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \quad \text{..... (2.5)}$$

Keterangan :

$R_{i,j}$ = rasio nilai perbandingan antara *cluster* ke-*i* dan cluster ke-*j*

SSW_I = minimal jarak antara titik *cluster* ke-*i*

SSW_J = minimal jarak antara titik *cluste*r ke-*j*

$SSB_{I,J}$ = maksimal jarak diantara *Cluster Ci* dan *Cj*

$$DBI = \frac{1}{K}\sum_{i=1}^{K} \max_{i \neq j}\left(R_{i,j}\right) \text{..... (2.6)}$$

Keterangan :

DBI = nilai skalar evaluasi DBI

$R_{i,j}$ = rasio nilai perbandingan antara *cluster* ke-*i* dan cluster ke-*j*

K = jumlah *cluster* yang digunakan

2.7 Prestasi akademik

Akademik secara Bahasa berasal dari kata akademi yang berarti lembaga pendidikan tinggi setingkat universitas, institut, atau sekolah tinggi. Akademik adalah kata sifat yang menunjukkan sesuatu yang bersifat ilmiah dan berhubungan dengan ilmu pengetahuan. Sesuatu yang berdasarkan teori – teori yang telah diuji kebenarannya dan bersifat objektif. (Sobur, 2006)

Pengertian akademik adalah kemampuan yang dapat diukur secara pasti karena ilmu pengetahuan itu sendiri. Akademik berkaitan dengan kegiatan formal yang diadakan sebuah institusi atau lembaga tertentu dengan syarat tertentu. Kemampuan akademik seseorang sering diidentikkan dengan kecerdasan otak kiri, karena berhubungan dengan logika. Prestasi akademik adalah kemampuan, kecakapan, dan prestasi yang didapatkan seseorang dimana kemampuan tersebut dapat bertambah dari waktu ke waktu karena adanya proses belajar dan bukan disebabkan karena proses pertumbuhan.

Faktor – faktor yang mempengaruhi prestasi belajar dalam teori kognitif sosial menurut Bandura dibangun dari dua faktor utama, yaitu :

1. Faktor perilaku (internal) yaitu *self regulated learning* (SRL) meruapakan variabel laten eksogenus terhadap prestasi belajar peserta didik, sekaligus berfungsi sebagai variabel laten endogenus terhadap sikap orang tua terhadap anak dan sikap guru terhadap peserta didik.
2. Faktor lingkungan (eksternal) yaitu sikap orang tua terhadap anak dan guru terhadap peserta didik.

2.8 Penelitian sebelumnya

Beberapa artikel yang digunakan sebagai referensi pembelajaran, di dapatkan beberapa contoh kasus yang hampir sama dengan permasalahan yang dihadapi, berikut artikel yang digunakan sebagai bahan wacanan Antara lain :

1. Penelitian yang dilakukan oleh Arga Yuavy Hertanto (2014) adalah "Sitem Pengelompokan Jurusan Siswa SMA NU 2 Gresik pada media IST (*Intelligent Structure Test*) Menggunakan Metode *K-means*". Pengelompokan dilakukan dengan menggunakan variabel melengkapi kalimat, mencari kata yang berbeda, mencari hubungan kata, mencari kata yang mencakup dua pengertian, hitungan sederhana, deret angka, menyusun bentuk, bangun ruang kubus, dan mengingat kata. Hasil uji sistem dengan evaluasi klaster internal Davies-Bouldin Index yang paling bagus pada cluster 3 yakni 1.4804, dengan jumlah prosentase 90,20 % dapat dikatakan sebagai hasil cluster yang baik karena mendekati kondisi riil pengelompokan jurusan, sehingga hasil clustering tersebut dapat diterapkan. (Hertanto, 2017)
2. Penelitian yang dilakukan oleh Banatus Sa'adah (2013) adalah "Pengelompokan Potensi Akademik Siswa RA Tarbiyatul Aulad dengan Metode *K-means*". Penulis menjelaskan masalah yang terjadi dalam penelitian ini adalah perkembangan sekolah dasar yang melakukan seleksi bagi siswa TK maupun RA, hal ini dilakukan untuk mengetahui kemampuan dan pengetahuan siswa tersebut sehingga sekolah dasar dapat mengelompokkan kemampuan siswa yang berbeda – beda. Metode yang

digunakan adalah metode *K-means*. Hasil dari penelitian ini menghasilkan bahwasanya metode *K-means* dapat digunakan untuk mengelompokkan potensi akademik siswa RA Tarbiyatul Aulad. (Sa'adah, 2013)

3. Penelitian yang dilakukan oleh Mochamad Jainul Arifin (2017) adalah "Aplikasi Pengelompokan Santri SHQ Menggunakan Metode *K-means*". Pengelompokan dilakukan dengan parameter uji kemampuan siswa yaitu usia, kemampuan baca, jumlah hafalan, kelancaran menghafal. Hasil uji sistem dengan evaluasi klaster internal Davies-Bouldin Index yang paling bagus pada cluster 3 yakni 0.4476, dari hasil pengujian sistem yang telah dilakukan didapatkan nilai DBI dan jumlah anggota masing-masing cluster yang berbeda pada setiap pengujian. Hal ini dipengaruhi oleh masukan data titik pusat awal atau centroid awal yang dapat mempengaruhi perhitungan dengan metode *K-means* secara keseluruhan. (Arifin, 2017)