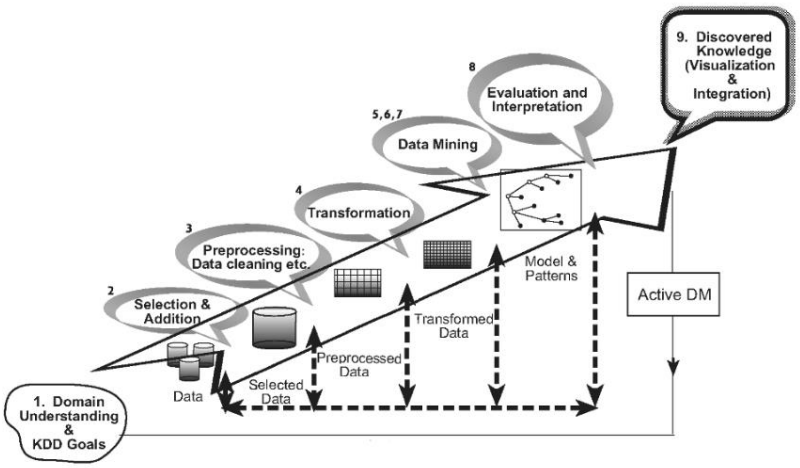
BAB 2
LANDASAN TEORI

2.1 Data Mining

Data mining atau yang dikenal dengan Knowledge Discovery in Databases (KDD) merupakan proses untuk menemukan pola yang menarik atau pengetahuan dari sejumlah sumber data seperti database, teks, gambar, Web, dll. Pola atau pengetahuan yang didapat haruslah valid, berpotensi berguna, dan dapat dimengerti [7].



Gambar 2.1 Alur Proses KDD

Berikut adalah tahapan dari KDD :

1. *Domain Understanding & KDD Goals*, yaitu mengembangkan pemahaman tentang domain aplikasi. Ini adalah langkah persiapan awal. Mempersiapkan keadaan untuk memahami apa yang harus dilakukan dengan banyak pertimbangan (tentang transformasi, algoritma, representasi, dll.). Orang yang bertanggung jawab atas proyek KDD perlu memahami dan menentukan tujuan pengguna akhir dan lingkungan di mana proses penemuan pengetahuan akan berlangsung (termasuk pengetahuan sebelumnya yang relevan). Saat proses KDD berlangsung, mungkin ada revisi dan penyetelan langkah ini.

2. *Selection & Addition*, setelah menentukan tujuan, data yang akan digunakan untuk penemuan pengetahuan harus ditentukan. Ini termasuk mencari tahu data apa yang tersedia, memperoleh data tambahan yang diperlukan, dan kemudian mengintegrasikan semua data untuk penemuan pengetahuan ke dalam satu kumpulan data, termasuk atribut yang akan dipertimbangkan untuk proses. Proses ini sangat penting karena Data Mining belajar dan menemukan dari data yang tersedia. Ini adalah dasar bukti untuk membangun model. Jika beberapa atribut penting hilang, maka seluruh studi mungkin gagal. Dari keberhasilan proses, ada baiknya mempertimbangkan sebanyak mungkin atribut pada tahap ini. Di sisi lain, untuk mengumpulkan, mengatur, dan mengoperasikan repositori data yang kompleks itu mahal, dan ada pertukaran dengan peluang untuk memahami fenomena dengan baik. *Tradeoff* ini merupakan aspek di mana aspek interaktif dan iteratif dari KDD berlangsung. Ini dimulai dengan kumpulan data terbaik yang tersedia dan kemudian berkembang dan mengamati efeknya dalam hal penemuan dan pemodelan pengetahuan.
3. *Preprocessing*, pada tahap ini, keandalan data ditingkatkan. Ini termasuk pembersihan data, seperti menangani nilai yang hilang dan menghilangkan noise atau outlier.
4. *Transformation*, pada tahap ini, pembuatan data yang lebih baik untuk penambangan data disiapkan dan dikembangkan. Metode di sini termasuk pengurangan dimensi (seperti pemilihan dan ekstraksi fitur, dan pengambilan sampel rekaman), dan transformasi atribut (seperti diskritisasi atribut numerik dan transformasi fungsional). Langkah ini sering kali penting untuk keberhasilan seluruh proyek KDD, tetapi biasanya sangat spesifik untuk proyek. Misalnya, dalam pemeriksaan medis, hasil bagi dari atribut mungkin sering menjadi faktor yang paling penting, dan tidak masing-masing dengan sendirinya. Dalam pemasaran, kita mungkin perlu mempertimbangkan efek di luar kendali kita serta upaya dan masalah temporal (seperti mempelajari efek akumulasi iklan). Namun, bahkan jika kita tidak menggunakan transformasi yang tepat di

awal, kita dapat memperoleh efek mengejutkan yang mengisyaratkan kepada kami tentang transformasi yang diperlukan (dalam iterasi berikutnya). Dengan demikian proses KDD mencerminkan dirinya sendiri dan mengarah pada pemahaman tentang transformasi yang dibutuhkan (seperti pengetahuan singkat dari seorang ahli di bidang tertentu mengenai indikator utama).

5. *Data Mining - Choosing the appropriate Data Mining task.* Memilih tugas Data Mining yang sesuai. Memutuskan jenis Data Mining yang akan digunakan, misalnya, klasifikasi, regresi, atau pengelompokan. Ini sebagian besar tergantung pada tujuan KDD, dan juga pada langkah-langkah sebelumnya. Ada dua tujuan utama dalam Data Mining: prediksi dan deskripsi. *Prediction* sering disebut sebagai Supervised Data Mining, sedangkan Deskriptif Data Mining mencakup aspek *unsupervised* dan visualisasi dari Data Mining. Sebagian besar teknik data mining didasarkan pada pembelajaran induktif, di mana model dibangun secara eksplisit atau implisit dengan menggeneralisasi dari sejumlah contoh pelatihan yang memadai. Asumsi yang mendasari pendekatan induktif adalah bahwa model yang dilatih dapat diterapkan untuk kasus-kasus masa depan. Strategi ini juga memperhitungkan tingkat meta-learning untuk kumpulan data tertentu yang tersedia.
6. *Data Mining - Choosing the Data Mining algorithm.* Tahap ini termasuk memilih metode spesifik yang akan digunakan untuk pola pencarian (termasuk beberapa penginduksi). Untuk setiap strategi meta-learning ada beberapa kemungkinan bagaimana hal itu dapat dicapai. Meta-learning berfokus untuk menjelaskan apa yang menyebabkan suatu algoritma Data Mining berhasil atau tidak dalam suatu masalah tertentu. Dengan demikian, pendekatan ini mencoba untuk memahami kondisi di mana algoritma Data Mining paling tepat. Setiap algoritma memiliki parameter dan taktik pembelajaran.
7. *Data Mining -Employing the Data Mining algorithm.* Dalam langkah ini kita mungkin perlu menggunakan algoritma beberapa kali sampai hasil

yang memuaskan diperoleh, misalnya dengan menyetel parameter kontrol algoritma, seperti jumlah minimum *instance* dalam satu daun pohon keputusan.

8. *Evaluation.* Dalam tahap ini kita mengevaluasi dan menafsirkan pola yang ditambang (aturan, keandalan, dll.), sehubungan dengan tujuan yang ditentukan pada langkah pertama. Di sini kami mempertimbangkan langkah-langkah prapemrosesan sehubungan dengan efeknya pada hasil algoritme Data Mining (misalnya, menambahkan fitur di Langkah 4, dan mengulanginya dari sana). Langkah ini berfokus pada pemahaman dan kegunaan model yang diinduksi. Dalam langkah ini, pengetahuan yang ditemukan juga didokumentasikan untuk penggunaan lebih lanjut.
9. *Visualization & Integration.* Pengetahuan menjadi aktif dalam arti bahwa kita dapat membuat perubahan pada sistem dan mengukur efeknya. Sebenarnya keberhasilan langkah ini menentukan efektivitas seluruh proses KDD. Ada banyak tantangan dalam langkah ini, seperti kehilangan "kondisi laboratorium" tempat beroperasi. Misalnya, pengetahuan ditemukan dari *snapshot* statis tertentu (biasanya sampel) dari data, tetapi sekarang data menjadi dinamis. Struktur data dapat berubah (atribut tertentu menjadi tidak tersedia), dan domain data dapat dimodifikasi (seperti, atribut mungkin memiliki nilai yang tidak diasumsikan sebelumnya).

2.1.1 Analisis Prediktif

Analisis Prediktif pada data mining mengacu pada tugas penambangan data dengan melakukan induksi pada data untuk membuat prediksi. Analisis prediktif tergolong pada Supervised Learning yang berarti penambangan data dilakukan pada data yang telah memiliki label [8].

2.1.1.1 Classification

Classification merujuk pada proses menemukan model atau suatu fungsi yang dapat dengan baik menggambarkan perbedaan antara kelas pada data atau konsep. Model atau fungsi diturunkan berdasarkan hasil analisis dengan penerapan

algoritma pada kumpulan data pelatihan atau data yang diketahui label kelasnya. Nantinya model atau fungsi akan digunakan untuk memprediksi label kelas pada data uji atau data yang tidak diketahui label kelasnya [8].

2.2 Web Data Mining

Web Data Mining bertujuan untuk menemukan informasi atau pengetahuan yang berguna dari struktur *hyperlink* Web, konten halaman, dan data penggunaan. Web Data Mining merupakan pengaplikasian dari data mining pada data website yang memiliki sifat semi-terstruktur atau tidak terstruktur [4].

2.2.1 Web Content Mining

Web Content Mining merupakan bagian dari web data mining yang mengekstrak dan menambang informasi yang berguna atau pengetahuan dari konten website. Web Content Mining dapat digunakan untuk mengklasifikasikan website secara otomatis berdasarkan topiknya [4].

2.2.2 Web Crawling

Web Crawling merujuk kepada kegiatan untuk memahami bagaimana cara memperoleh data teks dari berbagai macam website. Pada saat melakukan Web crawling kita dapat menggunakan *web crawler* yang secara otomatis melintasi struktur *hyperlink* Web dan mengunduh setiap halaman lalu ditautkan ke penyimpanan lokal [4].

2.2.3 Meta-tag

Meta-tag adalah *non-displaying*, atau tag HTML tersembunyi yang dapat memberikan pemilik situs dan penulis dengan tingkat kontrol atas bagaimana halaman Web di indeks. Meta-tag harus selalu ditempatkan di kepala dokumen HTML, tepat setelah elemen <TITLE>, di antara tag <HEAD> yang sebenarnya, dan sebelum tag <BODY>. Namun, meta-tag tidak muncul di tampilan halaman Web, tetapi dapat dibaca dan digunakan oleh mesin pencari atau oleh *br* pengguna. Meta-tag dapat digunakan untuk mengidentifikasi properti dokumen (misalnya

penulis, tanggal kedaluwarsa, kata kunci, dll.) dan memberikan nilai pada properti tersebut [9].

2.2.3.1 Server-side rendering

Server-side rendering mengacu pada website yang memiliki mekanisme rendering yang dilakukan di server. Pada server-side rendering konten website akan terlihat secara langsung yang berdampak pada lebih mudah di indeks oleh mesin pencari. Konten website yang terlihat secara langsung membuat informasi yang terkandung di dalam website dapat diperoleh dengan baik.

2.3 Text Mining

Text mining merupakan proses dimana pengguna berinteraksi dengan kumpulan dokumen dari waktu ke waktu menggunakan seperangkat alat analisis. Text mining berusaha mengekstraksi informasi yang berguna dari sumber data melalui identifikasi dan eksplorasi pola yang menarik. Sumber data yang dimaksud adalah data tekstual yang tidak terstruktur [10].

2.4 Text Preprocessing

Text Preprocessing dilakukan pada dokumen atau data tekstual sebelum dilakukannya ekstraksi pengetahuan [4]. Text Preprocessing diperlukan agar data tekstual bersih dari data yang tidak diperlukan. Prosesnya mencakup menghilangkan *stopword* atau data yang berulang kali muncul tetapi tidak mewakili makna pada dokumen, *Stemming* atau mengubah kata menjadi bentuk dasarnya, *Case folding* atau menyeragamkan bentuk huruf pada teks, menghapus *tags* pada HTML

2.5 Support Vector Machine

Algoritma Support Vector Machine (SVM) bekerja dengan membuat suatu sistem persamaan linear yang nantinya akan memisahkan dua kelas. Algoritma ini memiliki landasan teoritis yang kuat dan melakukan klasifikasi dua kelas dengan lebih akurat terutama pada data yang berdimensi tinggi seperti data teks [4].