Pengenalan Data mining

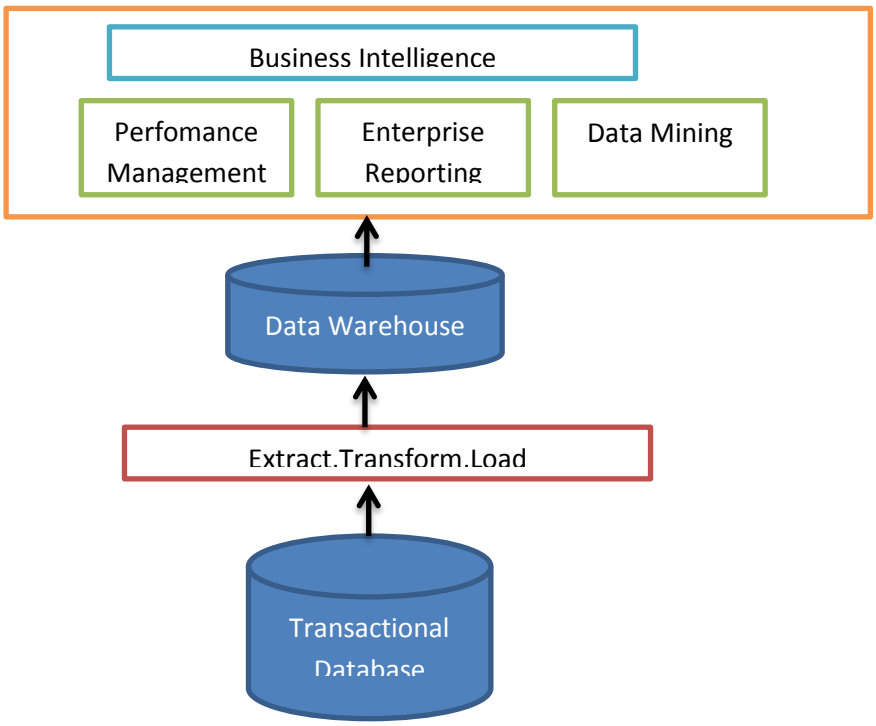
Data Mining merupakan suatu proses otomatis/semi otomatis dalam penemuan pola di dalam data. Pola yang ditemukan harus memiliki manfaat (Witten, 2011).

Tan, P. et al.(2006) mendefinisikan Data Mining sebagai proses untuk mendapatkan informasi yang berguna dari gudang basis data 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 juga disebut sebagai Knowledge Discovery.

Teknik yang digunakan di dalam data mining salah satunya adalah dengan menelusuri data yang ada untuk membangun sebuah model, kemudian menggunakan model tersebut agar dapat mengenali pola data lainnya yang tidak terdapat di dalam basisdata yang tersimpan. Teknik yang lain adalah untuk melakukan prediksi, berikutnya teknik pengelompokkan data dengan tujuan mengetahui pola universal data-data yang ada. Teknik deteksi anomali untuk mengetahui data yang berbeda dari data lainnya yang normal.

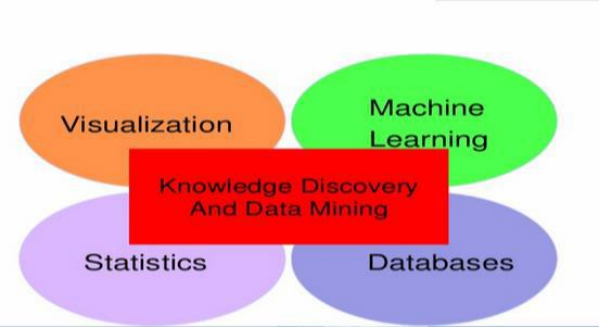
Data mining merupakan bidang yang menggunakan data yang dihasilkan dari data warehouse, bersama dengan bidang yang menangani masalah pelaporan data dan manajemen data. Sementara data warehouse sendiri memiliki tugas menarik data dari basis data mentah dan melakukan pengolahan sehingga data dapat digunakan untuk keperluan pelaporan,manajemen data datamining.



Gambar posisi datamining dalam *business intelligence*

Posisi data mining dalam berbagai bidang ilmu.

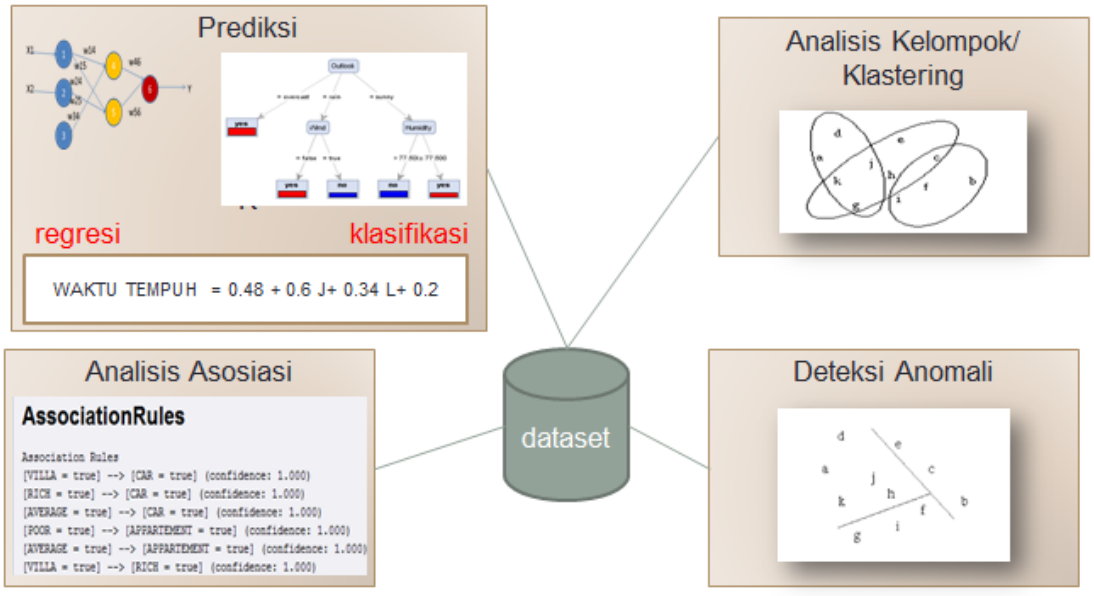
Para ahli berusaha menentukan posisi bidang data mining diantara bidang-bidang lainnya. Hal ini dikarenakan ada kesamaan antara sebagian bahasan dalam data mining dengan bahasan di bidang lain. Kesamaan data mining dengan bidang statistik adalah tentang hal sampling,estimasi, dan pengujian hipotesis. Kesamaan dengan *artificial intellegent*,*pattern recognition*,dan *machine learning* adalah pada algoritma pencarian,teknik pemodelan, dan teori pembelajaran.



Bidang lain yang mempengaruhi data mining adalah teknologi basis data, yang mendukung penyediaan penyimpanan yang efisien,pengindexan, dan pemrosesan query. Teknik komputasi paralel digunakan untuk memberikan kinerja yang tinggi untuk dataset yang berukuran besar, sedangkan komputasi terdistribus digunakan untuk menangani masalah ketika data tidak dapat disimpan pada satu tempat.

Pekerjaan dalam data mining

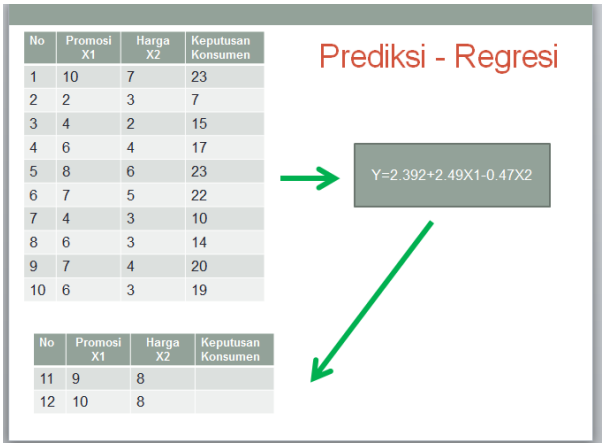
Pekerjaan datamining dapat dibagi menjadi empat kelompok, yaitu model prediksi,analisis kelompok, analisis asosiasi, dan deteksi anomali.



Kategori pekerjaan dalam data mining:

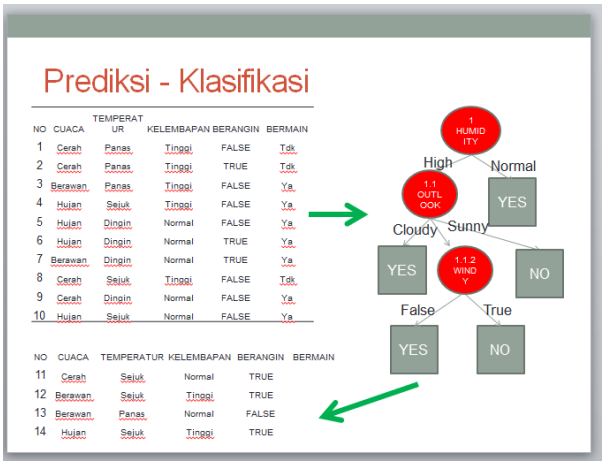
1. Model prediksi
   Model prediksi berkaitan dengan pembuatan model yang dapat melakukan pemetaan dari setiap himpunan variabel ke setiap targetnya, kemudian menggunakan model

tersebut untuk memberikan nilai target pada himpunan baru. Ada dua jenis model prediksi yaitu regresi dan klasifikasi.



No	Promosi X1	Harga X2	Keputusan Konsumen
1	10	7	23
2	2	3	7
3	4	2	15
4	6	4	17
5	8	6	23
6	7	5	22
7	4	3	10
8	6	3	14
9	7	4	20
10	6	3	19

No	Promosi X1	Harga X2	Keputusan Konsumen
11	9	8	
12	10	8	



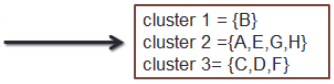
NO	CUACA	TEMPERATUR	KELEMBAPAN	BERANGIN	BERMAIN
1	Cerah	Panas	Tinggi	FALSE	Tdk
2	Cerah	Panas	Tinggi	TRUE	Tdk
3	Berawan	Panas	Tinggi	FALSE	Ya
4	Hujan	Sejuk	Tinggi	FALSE	Ya
5	Hujan	Dingin	Normal	FALSE	Ya
6	Hujan	Dingin	Normal	TRUE	Ya
7	Berawan	Dingin	Normal	TRUE	Ya
8	Cerah	Sejuk	Tinggi	FALSE	Tdk
9	Cerah	Dingin	Normal	FALSE	Ya
10	Hujan	Sejuk	Normal	FALSE	Ya

NO	CUACA	TEMPERATUR	KELEMBAPAN	BERANGIN	BERMAIN
11	Cerah	Sejuk	Normal	TRUE	
12	Berawan	Sejuk	Tinggi	TRUE	
13	Berawan	Panas	Normal	FALSE	
14	Hujan	Sejuk	Tinggi	TRUE	

2. Analisis kelompok

Analisis kelompok melakukan pengelompokan data-data ke dalam sejumlah kelompok(cluster) berdasarkan kesamaan karakteristik masing-masing data pada kelompok-kelompok yang ada. Data yang masuk dalam batas kesamaan dengan kelompoknya akan bergabung dalam kelompok tersebut, dan akan terpisah dengan kelompok yang berbeda jika diluar batas kesamaan kelompok tersebut.

Nasabah	Jumlah Rumah	Jumlah Mobil
A	1	3
B	3	3
C	4	3
D	5	3
E	1	2
F	4	2
G	1	1
H	2	1



3. Analisis asosiasi

Menemukan pola yang menggambarkan kekuatan hubungan fitur/variabel dalam data. Pola yang ditemukan biasanya merepresentasikan bentuk aturan implikasi(subset fitur).

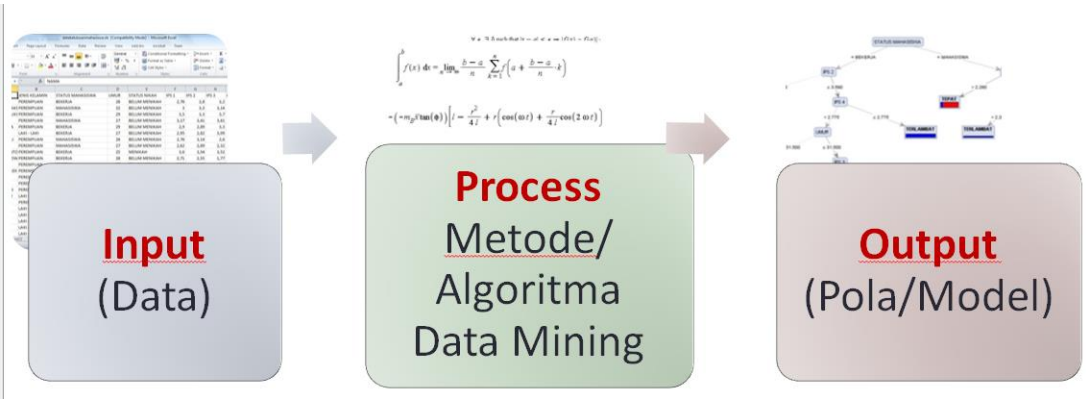
No.TX	Item
1	susu,Teh,Gula
2	Teh,Gula,Roti
3	Teh,Gula
4	Susu,Roti
5	Susu,Gula,Roti
6	teh,Gula
7	Gula,Kopi,Susu
8	Gula,Kopi,Susu
9	Susu,Roti,Kopi
10	Gula,Teh,Kopi

Aturan	Confidence	
IF Teh THEN Gula	5/5	100%
IF Kopi THEN Gula	3/4	75%
IF Kopi THEN Susu	3/4	75%
IF Roti THEN Susu	3/4	75%

4. Deteksi anomali
   Pekerjaan deteksi anomali berkaitan dengan pengamatan sebuah data dari sejumlah data yang secara signifikan mempunyai karakteristik berbeda dengan sisa data lainnya (outlier).
   Contoh penggunaan:
   Deteksi anomali diterapkan pada sistem jaringan untuk mengetahui pola data yang memasuki jaringan sehingga penyusupan bisa ditemukan jika pola kerja data yang datang berbeda.

Proses Utama Datamining

Input-proses-output

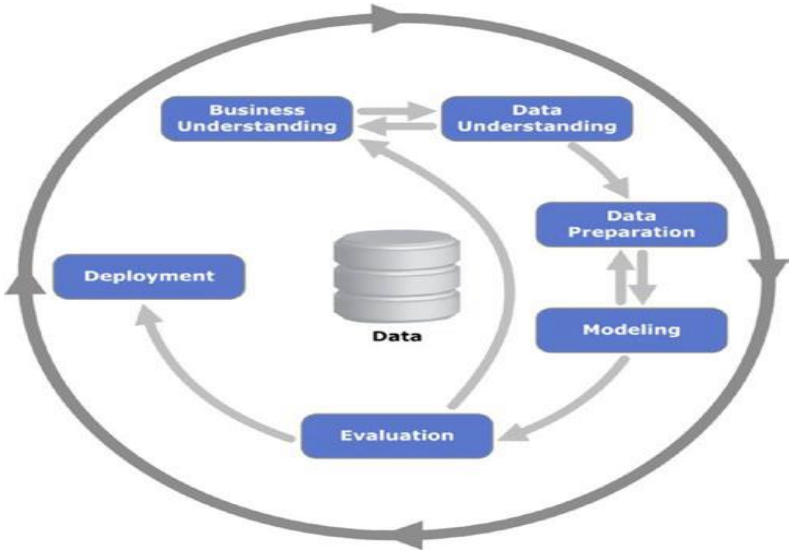


Contoh Penerapan Data Mining:

- Ecommerce:
  Pembelian buku lewat Amazon tentang data mining, akan disarankan pula buku-buku lain yang seharusnya dibeli, karena Amazon dapat melakukan clustering terhadap data buku-buku-buku yang dibeli
- Medis:
  Memprediksi apakah seorang pasien yang diopname akan mendapatkan serangan jantung berikutnya berdasarkan catatan kesehatan sebelumnya dan pola makan
- Prakiraan Cuaca:
  Memprediksi apakah akan terjadi tornado berdasarkan informasi dari sebuah radar tentang data kondisi angin dan data kondisi atmosfir yang lain.

Cross-Industry Standard Process for Data Mining

Umumnya dikenal dengan singkatan CRISP-DM. Itu didirikan oleh Program Strategis Eropa pada penelitian dalam inisiatif Teknologi Informasi dengan tujuan untuk membuat metodologi yang tidak memihak yang tidak tergantung domain. Ini adalah upaya untuk mengkonsolidasikan praktik terbaik proses datamining diikuti oleh para ahli untuk mengatasi masalah penambangan data. framework tersebut disusun pada tahun 1996 dan pertama kali diterbitkan pada tahun 1999 dan dilaporkan sebagai metodologi terkemuka untuk proyek-proyek penambangan data / analisis prediktif dalam jajak pendapat yang dilakukan pada tahun 2002, 2004, dan 2007. Ada rencana antara 2006 dan 2008 untuk memperbarui CRISP-DM tetapi itu pembaruan tidak terjadi, dan hari ini situs web asli CRISP-DM.org tidak lagi aktif.



Tahapan yang dilakukan pada framework CRISP-DM:

- Business Understanding
  Tahap ini fokus pada proses memahami tujuan proyek secara keseluruhan dan harapan dari perspektif bisnis. Tujuan ini dikonversi menjadi definisi masalah data mining /machine learning dan rencana kegiatan seputar kebutuhan terhadap data, masukan pemilik bisnis, dan bagaimana mengukur evaluasi kinerja dari hasil yang sudah dirancang.
- Data Understanding
  Dalam fase ini, data awal dikumpulkan sesuai dengan persyaratan pada fase sebelumnya. Kegiatan dilakukan untuk memahami relevansi data dengan tujuan yang proyek, masalah kualitas data, dan pandangan thd yang tujuannya adalah untuk menghasilkan hipotesis yang sesuai. Hasil dari fase ini akan disajikan secara iteratif kepada owner bisnis sehingga pemahaman bisnis dan tujuan proyek menjadi lebih jelas.
- Data Preparation:
  Fase ini merupakan fase proses membersihkan data sehingga siap digunakan untuk fase pembuatan model. Pembersihan data mencakup mengisi kesenjangan data yang diketahui dari langkah sebelumnya, perlakuan terhadap *missing value*, mengidentifikasi fitur-fitur penting, menerapkan transformasi, dan membuat fitur-fitur baru yang relevan yang bisa diterapkan. Keakuratan model akan sangat bergantung

pada kualitas data yang dimasukkan ke dalam algoritma yang mempelajari pola data tersebut.

- Modeling:
  Merupakan tahap pembuatan model dimana beberapa algoritma machine learning digunakan untuk memecahkan masalah yang diberikan. Jadi berbagai algoritma machine learning yang tepat diterapkan pada dataset yang bersih, dan parameternya disesuaikan dengan nilai optimal yang memungkinkan. masing-masing model yang diterapkan dicatat Kinerjanya.
- Evaluation
  Pada tahap ini pembandingan kinerja dilakukan di antara semua model yang digunakan yang memiliki akurasi tinggi. Model akan diujikan pada data yang tidak digunakan sebagai bagian dari pelatihan untuk mengevaluasi konsistensi dari kinerjanya. Hasil akan bandingkan dengan kebutuhan bisnis yang diidentifikasi dalam fase 1. Pakar dari permasalahan bisnis tersebut dilibatkan untuk memastikan bahwa hasil model akurat dan dapat digunakan sesuai dengan tujuan proyek.
- Deployment
  Fokus utama dalam fase ini adalah penggunan model yang telah dihasilkan.Jadi model akhir yang telah disetujui oleh pakar/ahli akan diimplementasikan,dan pengguna dari model akan dilatih untuk menggunakannya dalam mengambil keputusan bisnis yang sesuai dengan ayng ditetapkan dalam fase pemahaman bisnis.