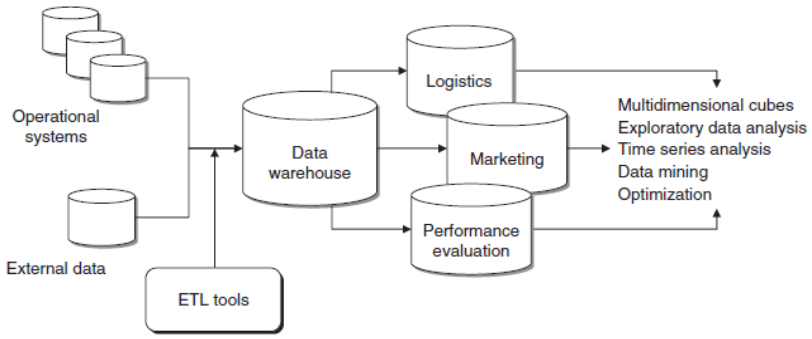
BAB II
LANDASAN TEORI

2.1 *Data Mining*

2.1.1 Pengertian *Data Mining*

Menurut Han dan Kamber(2006, p.9)*d*ata mining merupakan proses menggali informasi yang tersembunyi dari data berjumlah besar untuk pengambilan keputusan, prediksi ataupun pemecahan masalah. Dimana melibatkanintegrasiteknik dariberbagai disiplin ilmuseperti database danteknologi datawarehouse, statistik, *machine learning*, performa tinggi komputasi, pengenalan pola, jaringan saraf, visualisasi data, penemuan informasi, gambar danpemrosesan sinyal, dananalisis data spasialatau temporal.

Sedangkan Vercellis (2009, p.9-11), menggambarkan *data mining* sebagai bagian dari arsitektur sistem *business intelligence* (BI).



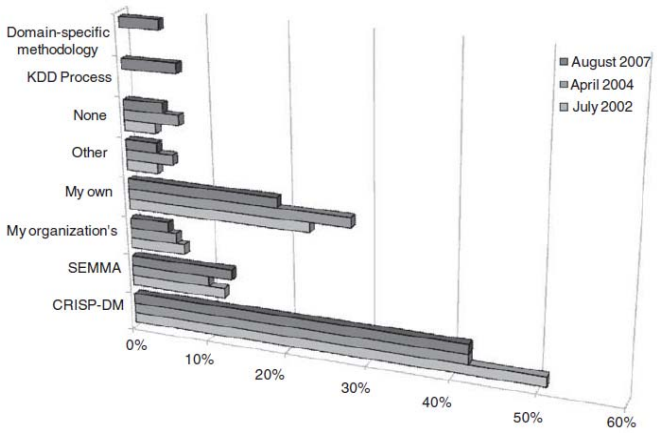
Gambar 2-1 Arsitektur system BI (Vercellis, 2009)

Dari gambar di atas, dapat dilihat karakteristik *data mining* yang tergantung pada prosedur untuk mengumpulkan data *history* yang berasal dari

berbagai sumber (*Operational systems* dan *External data*) kemudian memasukkannya ke dalam *database* (*data warehouse* atau *data marts*) melalui ETL *tools*. *Data mining* pada sistem BI merupakanmetodologi aktif dengan tujuan untuk ekstraksi informasi dan pengetahuan dari data yang ada di *data warehouse*/ *data marts*.*Data mining* termasuk model matematika untuk pengenalan pola, *machine learning*dan teknik *data mining*. *Data mining* tidak memerlukan hipotesis sebelumnya untuk diverifikasi tapi sebaliknya untuk memperluas pengetahuan pembuat keputusan.

2.1.2 *Cross-Industry Standard Process for Data Mining* (CRISP-DM)

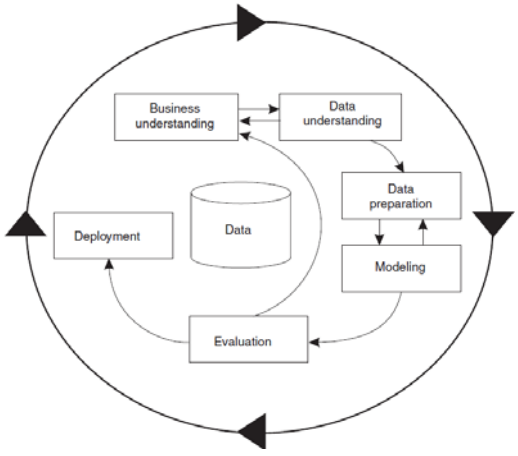
Dalam menanggapi isu-isu umum dan kebutuhan dalam proyek *data mining* dipertengahan 90-an, kelompok organisasi yang terlibat dalam *data mining* (Teradata, SPSS-ISL,Daimler-Chrysler dan Ohra) mengusulkan panduan referensi untuk mengembangkan proyek-proyek *data mining*, bernama CRISP-DM (*CrossIndustry Standard Process for Data Mining*).Mariscal, Marba dan Fernandez(2010) menyatakan CRISP-DM sebagai*defacto*standar untukpengembangan proyek*data mining* dan*knowledge discovery* karena paling banyak digunakan dalam pengembangan *data mining*. Hal tersebut dapat terlihat dari survei yang dilakukan terhadap penggunaan metodologi dalam proyek *data mining***:**



Gambar 2-2 Survei Penggunaan Metodologi Data Mining
(Mariscal, Marban, & Fernandez, 2010)

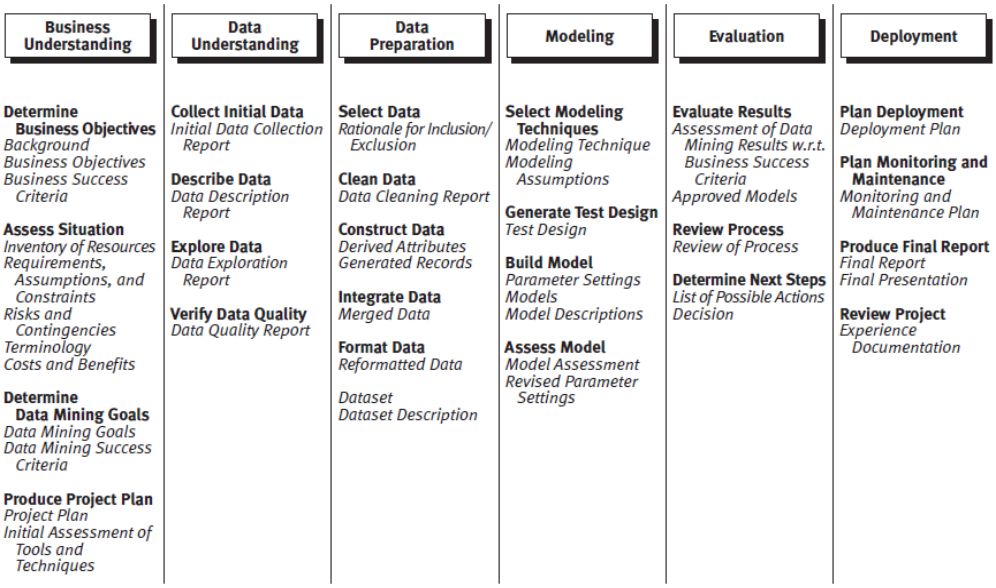
Hasil survei "Penggunaan Metodologi dalam Proyek *Data Mining"*, memperlihatkan pengguna CRISP-DM di tahun 2002 mencapai 51%, kemudian menurun menuju 41% di tahun 2004. Meskipun persentasi penggunaan CRISP-DM menurun 10%, jumlah pengguna metodologi ini masih terbilang lebih banyak daripada pengguna metodologi lain.

Model prosesCRISP-DM memberikan gambaran tentangsiklus hidup proyek *data mining*. Siklus hidupproyek*data mining*dalam CRISP-DM terdiri dari enam fase(Gambar 2-3).



Gambar 2-3 Fase CRISP-DM (Chapman, dkk,2000)

Pada CRISP-DM, urutanfasetidak kaku, dapat bergerakbolak-balik antarafase yang berbeda. Tanda panah pada gambar 2-3, menunjukkanfrekuensi ketergantunganantara fase.Berikut gambaran darigaris besarsetiap fasedisertai dengantugas-tugas(*bold*) dan output(*italic*).



Gambar 2-4 Tugas dan *output* dari model CRISP-DM
(Chapman, dkk,2000)

2.1.2.1 *Business Understanding*

2.1.2.1.1 *Determine Business Objectives*

Pada tahap awal, tim proyek harus benar-benar memahamidari perspektif bisnis, apa yang ingin dicapai. Adapun tahap ini bertujuan untuk memahami bidang masalah, menghasilkan solusiyang tepat,danmengungkapkanfaktor pentingyang berpengaruh pada hasilproyek.

Output dari tahap ini merupakan analisis dari hasil observasi, wawancaradan dokumen-dokumen yang dimiliki perusahaan.Berikut penjelasan dari *output* yang dihasilkan:

- ***Background***

Output ini berisi catatan informasi yang diketahui tentang situasi bisnis, organisasi pada awal proyek dan gambaran dasardarikonteksproyek. Konteks proyek yang dimaksud seperti proyek inibekerjadi bidangapa, identifikasi masalah, dan mengapa*data mining*muncul untukmemberikan solusi. Aktivitas pada point ini:

	Aktivitas
Organisasi	Memahami bagan organisasi, identifikasi divisi, departemen, dan kelompok proyek
	Identifikasi nama dan tanggung jawab manajer
	Mengidentifikasi orang-orang penting dalam bisnis dan peran mereka.
	Mengidentifikasi yang menjadi sponsor internal (sponsor keuangan dan user ahli/ domain utama). Sponsor proyek merupakan orang yang memberikan arahan atas proyek atau sebagai *client* dari proyek.
	Mengidentifikasi unit bisnis yang terkena proyek data mining (misalnya, pemasaran, penjualan, keuangan)
Identifikasi masalah	Mengidentifikasi bidang masalah (misalnya, pemasaran, layanan pelanggan, pengembangan bisnis, dan lain-lain)
	Periksastatusproyek(misalnya, apakahproyekdata miningakan dilakukan dalam unitbisnis? Atauapakahdata miningperlu dipromosikansebagai kunci teknologidalam bisnis?)

	Menjelaskanprasyaratproyek(misalnya, Apa motivasiproyek? Apakahbisnis sudah menggunakandata mining?)
	Mengidentifikasisasaranuntuk hasilproyek (misalnya, Apakah kitadiharapkan memberikanlaporan untukmanajemen puncak atausistem operasionalyang digunakan oleh*end user*?)
	Mengidentifikasikebutuhan dan harapan pengguna
Solusi sekarang	Jelaskansolusisaat ini yang digunakan untukmengatasi masalah
	Jelaskankeuntungan dan kerugiandari solusisaat ini dan tingkatyangditerima oleh pengguna

- ***Business objectives***

  Output ini menjelaskantujuan utamakonsumen, dari perspektif bisnis.Di samping tujuan utama bisnis,biasanya adapertanyaan lain yang berhubungan dengan bisnis. Misalnya, tujuanbisnis mungkinuntuk mempertahankan konsumensaat inidengan memprediksi kapan merekapindah kepesaing.Contohpertanyaan-pertanyaan bisnisyang berhubungan, "Apakah channelutama yang digunakan (misalnya, ATMdan Internet) mempengaruhi perginya konsumen?"atau"Apakah menurunkanbiayaATM secara signifikan mengurangi jumlahkonsumen yang pergi?". Aktivitas pada point ini terdiri dari:

  - Gambarkanmasalah yang harus dipecahkan.
  - Spesifikasikansemua pertanyaanbisnissecermat mungkin.

- Spesifikasikan kebutuhan bisnis lainnya(misalnya, tidakingin kehilanganpelanggan)
- Spesifikasikanmanfaat yang diharapkandalam istilah bisnis

- ***Business success criteria***

  Output ini menjelaskankriteria sukses atau kegunaan hasil proyek darisudut pandang bisnis. Penjelasan dilakukan dengan spesifikdan dapatdiukur secara obyektif, misalnya,penguranganpelanggan yang pergike tingkat tertentu, atau secara subyektif, seperti"memberikan wawasan yang berguna". Namun harus ditunjukkanpenilaian subjektif dibuat oleh siapa. Aktivitas dari point ini:

  - Spesifikasikankriteria keberhasilanusaha (misalnya, Meningkatkanrespondalam promosimailingsebesar 10% dan tingkatpendaftaransebesar 20%).
  - Mengidentifikasisiapa yang menilaikriteria keberhasilan.

2.1.2.1.2 *Assess situation*

Tugas ini mencakuppenemuan faktayang lebihrincitentang semuasumber daya, kendala, asumsi, dan faktor lain yangharus dipertimbangkandalam menentukantujuananalisisdata danrencana proyek. Adapun *output* dari tugas ini adalah:

- ***Inventory of resources***

  Output ini berisi daftarsumber daya yang tersediauntuk proyek,termasuk personil(*business experts, data experts, technical support, data mining experts*), data (*fixed extracts, access to live, warehoused, or operational data*),sumber daya komputasi(*hardware platforms*), dan

perangkatlunak(*data mining tools*, perangkat lunak lainyang relevan). Berikut aktivitas dari point ini:

	Aktivitas
Sumber daya hardware	Mengidentifikasiperangkat keras
	Menetapkanketersediaanperangkat kerasdasar untukproyek
	Periksa apakahjadwalpemeliharaanhardwarekonflikdengan ketersediaanperangkat kerasuntukproyek data mining
	Mengidentifikasiperangkat kerasyang tersediauntuk *tool*data miningyang akan digunakan(jika*tool* yang digunakan sudahdiketahui padatahap ini)
Sumber daya data dan pengetahuan	Mengidentifikasi sumber data
	Identifikasijenissumber data(sumber online, ahli, dokumentasi tertulis, dll)
	Mengidentifikasi sumber-sumberpengetahuan
	Identifikasijenissumberpengetahuan (sumber online, tenaga ahli, dokumentasi tertulis, dll)
	Periksaalat danteknik yang tersedia
	Jelaskanlatar belakang pengetahuanyang relevan(atau informal)
Sumber daya manusia	Identifikasisponsor proyek(jika berbeda dari sponsorinternaldalam bagian2.1.2.1.1.1)
	Mengidentifikasisistem administrator, database

	administrator, dan stafdukungan teknis untukpertanyaan lebih lanjut
	Mengidentifikasianalis pasar, ahlidata mining, dan ahli statistik, dan memeriksa ketersediaanmereka
	Periksaketersediaanpakar domainuntuktahap selanjutnya

- ***Requirements, assumptions, and constraints***

  Output yang dimaksud berupa:

  Daftar semua*requirements*proyek, termasukjadwalpenyelesaian, komprehensibilitasdan kualitashasil, keamanan, sertamasalah hukum seperti memastikan ijinuntuk penggunaandata.

  Daftarasumsi yang dibuatoleh proyek.Ini mungkinasumsi tentangdata yang dapatdiverifikasiselamadata mining, atauasumsinon-verifikasi tentangbisnis yang terkait denganproyek.

  Daftar batasanpada proyek.Daftar ini berisi batasan padaketersediaan sumber daya, termasukteknologisepertiukuran datasetyangdigunakanuntuk pemodelan.

  Berikut aktivitas dari point ini:

	Aktivitas
Requirements	Spesifikasikantarget
	Menangkap semuakeperluan padapenjadwalan
	Menangkapkeperluankomprehensibilitas, akurasi, menyebarkankemampuan, pemeliharaan, dan

	pengulangandari proyekdata miningdan model yang dihasilkan
	Menangkapkeperluankeamanan,batasan hukum, privasi,pelaporan, danjadwal proyek
Asumsi	Menjelaskansemua asumsi(termasuk yang implisit) dan menjadikannya eksplisit
	Daftarkanasumsipada kualitasdata (misalnya, akurasi, ketersediaan)
	Daftarkanasumsipada faktor eksternal(misalnya, isu-isu ekonomi, produk yang kompetitif, kemajuan teknis)
	Menjelaskanasumsiyang mengarahke salah satuperkiraan(misalnya, hargaalat tertentudiasumsikanlebih rendahdari $ 1.000)
	Daftar semuaasumsi tentangapakah perluuntuk memahami danmenggambarkan atau menjelaskanmodel(misalnya, bagaimana seharusnya menyampaikanmodeldan hasilkepadamanajemen senior/sponsor)
Batasan	Periksa batasanumum (misalnya, masalah hukum, anggaran, rentang waktu, dan sumber daya)
	Periksahak akseske sumber data(misalnya, pembatasanakses, password)
	Periksaaksesibilitasteknisdata (sistem operasi, sistem

	manajemendata,file ataudatabase format)
	Periksa apakahpengetahuan yang relevandapat diakses
	Periksa kendalaanggaran(biaya tetap, biaya pelaksanaan, dll)

- ***Risks and contingencies***

  Output ini berisi daftarresikoyang mungkin menundaataumenyebabkan proyekgagal.Termasuk rencana kontingensi, yaitu tindakan apa yang akandiambil jikaresikoterjadi. Aktivitas pada point ini:

	Aktivitas
Identifikasi resiko	Identifikasiresiko bisnis(misalnya, pesaing memiliki hasil yang lebih baik)
	Mengidentifikasi resikoorganisasi (misalnya, departemen yang meminta proyektidak memilikidana untukproyek)
	Identifikasiresiko keuangan(misalnya, pendanaan lebih lanjuttergantung padahasilawal data mining)
	Identifikasiresiko teknis
	Mengidentifikasi resiko yang bergantung pada data dan sumber data (misalnya, kualitas buruk dan cakupan)
Mengembangkan rencana kontigensi	Menentukankondisi di manaresikodapat terjadi
	Mengembangkan rencanakontigensi

- ***Terminology***

Output ini berisidaftar terminologiyang relevan denganproyek.Dua komponen yang mungkin ada pada point ini :

1. Glossarybisnis, yangmerupakan bagian daripemahaman bisnisuntukproyek
2. Glossarydata mining, diilustrasikandengan contohyang relevan dari pertanyaan bisnis

- ***Costs and benefits***

Point ini terdiri darianalisis biaya dan laba untuk proyek, yang membandingkanbiaya proyekdengan potensibagibisnisjika proyek berhasil.Analisis biaya dan laba menghasilkan ROI (*Return of Invesment*) yang didapatkan dari proyek. Berikut adalah rumus dari perhitungan ROI:

(total discounted benefits - total discounted costs)/discounted costs

*Discounted benefits = benefit*x*discount factors*

Discounted costs = cost x*discount factors*

Discount factors = $1/(1+r)^t$

*Discount factors*adalah pengali untuk setiap tahun berdasakan *discount rate* dan tahun, yang digunakan untuk perhitungan biaya dan laba setiap tahun.

r =*discount rate* (nilai yang digunakan untuk memotong *cash flow* di masa depan)

t = tahun ke-

2.1.2.1.3 Determine data mining goals

Tujuan*data mining*bagian daritujuan proyekdalamistilah-istilah teknis.Sebagai contoh,tujuan bisnis"Meningkatkan penjualan katalog

untukpelanggan yang sudah ada". Maka tujuan*data mining*"Memprediksi berapa banyak *widget*pelangganakan membeli, melihatpembelianmereka tiga tahun terakhir, informasi demografis(umur, gaji, kota, dll), dan hargaitem". *Output* yang dihasilkan dalam tugas ini:

- ***Data mining goals***

  Menjelaskan*output*proyekyang diharapkan untuk mencapaitujuan bisnis.

- ***Data mining success criteria***

  Menentukan kriteria suksesdalam halteknismisalnya, akurasiprediksi di tingkattertentu.Seperti halnya kriteria keberhasilanbisnis, mungkin kriteria suksesdigambarkandalam halsubyektif,sehingga orang ataupihak yang melakukanpenilaian harus diidentifikasi.

2.1.2.1.4 Produce project plan

Tugas ini menjelaskanrencana yang ditetapkanuntuk mencapai tujuan*data mining*.Rencana yang dibuat berisilangkah-langkahyang akan dilakukansampai akhirproyek, termasukpemilihanalat dan teknik.

- ***Project plan***

  Output ini berisi daftartahapanyang akan dieksekusidalam proyek, durasi, sumber daya yang dibutuhkan, *input*,*output*, dandependensi. Bagian dari*project plan*adalahuntuk menganalisisketergantungan antarawaktudan resiko.*Project plan*adalahdokumen yang dinamisdalam arti bahwapadaakhir setiap tahap,diperlukan *review*kemajuan, prestasidan*update*sesuaijadwal.

- ***Initial assessment of tools and techniques***

  Penilaian alat dan teknikdilakukan karena dapat mempengaruhiseluruh proyek.Aktivitas pada point ini:

- Membuat daftarkriteria untukseleksi alat dan teknik
- Pilihalatdanteknik yang potensial
- Evaluasikesesuaianteknik
- *Review*dan memprioritaskanteknik yang diterapkanmenurutevaluasidari alternatif

2.1.2.2*Data Understanding*

2.1.2.2.1 Collect initial data

Pada tugas ini dilakukan akses ke data yang tercantum dalam sumber daya proyek, mencakup *loaddata*.*Output*yang dihasilkan:

- *Initial Data Collection Report*

  Menjelaskansemuadatayang digunakandalam proyek, termasukkebutuhan*selection* untukdata yang lebih detail.Selain itu ditentukan atributyang lebih pentingdaripada atribut lain.Untuk itu, dilakukan observasi ke perusahaan untuk mengetahui prosedur pengambilan data dan ketersediaan data. Aktivitas pada point ini:

	Aktivitas
Data requirements planning	Mendaftarkan informasiyang diperlukan(misalnya, memberikanatribut, atauinformasi tambahantertentu)
	Periksa apakahsemua informasi yang dibutuhkantersedia
Selection criteria	Tentukankriteria *select*(misalnya, atributyang diperlukan untuktujuandata mining? apakah

	relevan? berapa banyakatributyang dapat ditanganidenganteknikyang dipilih?)
	Pilihtabel/fileberkepentingan
	Pilihdata dalamtabel/file
	Tentukan menggunakandata*history* berapa lama(misalnya, data18bulan yang tersedia,hanya 12bulanmungkin diperlukanuntuk latihan)

2.1.2.2.2 Describe Data

Tugas ini memeriksapropertidata yang diperolehdanmelaporkan hasilnya.*Output* yang dihasilkan:

- *Data Description Report*

  Menjelaskandata yang telahdiperoleh, formatdata, jumlahdata(misalnya, jumlah *record*dan*field* dalamsetiap tabel), identitas*field*, dan fitur lainnyayang ditemukan. Aktivitas pada point ini:

	Aktivitas
Analisisvolumedata	Identifikasidata danmetode*capture*
	Akses sumber data
	Gunakananalisis statistikjika diperlukan
	Laporantabel danhubungan mereka
	Periksavolume data,kompleksitas
	Catat datayang berisi teks bebas
Jenisdan nilai atribut	Periksaaksesibilitas danketersediaanatribut
	Periksajenisatribut(numerik, simbolik, dll)
	Periksarentangnilaiatribut
	Menganalisiskorelasiatribut
	Memahamiarti dari setiapatributdannilai atributdalam

	istilah bisnis
	Untuksetiap atribut, menghitungstatistik dasar(misalnya, menghitungdistribusi,rata-rata, max, min, standardeviasi, varians, modus, kemiringan, dll)
	Menganalisisstatistikdasar danmakna hasilnya yang berhubungan dengan bisnis
	Putuskan apakahatributrelevan untuktujuandata mining
	Menentukan apakahmaknaatributyang digunakankonsisten
	Wawancaradengan pakardomainuntuk mendapatkanpendapat mereka tentangrelevansi atribut
	Menentukan apakahperlu untukmenyeimbangkandata (berdasarkan teknik pemodelanyang akan digunakan)
Keys	Menganalisishubungan antar kunci
	Periksajumlahnilai atributkunci yang tumpang tindih ditabel lain
Tinjauasumsi/tujuan	Bila perlu, perbaharui daftarasumsi

2.1.2.2.3 Explore Data

Tugas inimenanganipertanyaan-pertanyaan*data mining*tentang*query*, visualisasi, danteknikpelaporan.Analisis inidapatsecara langsung menjawabtujuan*data mining*, kontribusi perbaikandeskripsidata dan kualitas laporan, dan menjadi masukan bagitransformasidan langkah-langkahpersiapan datalainnya yang diperlukansebelumanalisis lebih lanjutdapat terjadi.

Output yang dihasilkan:

- *Data Exploration Report*

Menjelaskanhasildaritugas ini, termasukhipotesisawal dandampaknya terhadapsisaproyek.Laporan inijuga dapat mencakupgrafik danplotyang menunjukkankarakteristikdata ataudatasubset yang menarikuntukpemeriksaan lebih lanjut. Aktivitas dari tahap ini:

	Aktivitas
Data exploration	Menganalisispropertidari atributsecara rinci(misalnya, statistik dasar, sub-populasi yang menarik)
	Mengidentifikasikarakteristiksub-populasi
Anggapan untuk analisis masa depan	Mempertimbangkan danmengevaluasi informasidan temuandalam laporandeskripsidata
	Bentukhipotesisdanmengidentifikasitindakan
	Bila memungkinkan,ubahhipotesismenjaditujuandata mining
	Memperjelastujuandata miningatau membuatnya lebih tepat. Misalnya, pencarian “blind”tidak selaluberguna,tapi pencarianlebihdiarahkanke tujuanbisnisadalah lebih baik.
	Lakukananalisis dasaruntuk memverifikasihipotesis

2.1.2.2.4 Verify data Quality

Pada tugas ini dilakukan pemeriksaankualitas data, menangani pertanyaan-pertanyaanseperti: Apakahdata lengkap? (tidak untuk semuakasus), Apakahadakesalahan? Jika ada kesalahan, berapa banyak kesalahan

tersebut?Apakah adanilai yang hilang dalamdata?Jika demikian, bagaimana hal tersebut direpresentasikan?. *Output* yang dihasilkan:

- *Data Quality Report*

  Berisi daftarhasilverifikasikualitas data, jika adamasalah kualitas, daftarkan jugasolusi yang dapat dilakukan.

	Aktivitas
Review key dan atribut	Periksaluascakupan yang digunakan (misalnya, apakah mungkinsemuanilai diwakili)
	Periksa*Key*
	Pastikan bahwamaknadariatribut dannilai-nilaiyang terkandungcocok satu sama lain
	Mengidentifikasiatributhilang dan*field* kosong
	Menetapkanmaknadari data yang hilang
	Periksaadakah atributdengan nilai yang berbedanamun memilikiarti yang sama(misalnya, rendah lemak, diet)
	Periksaejaan danformatnilai(misalnya, nilai yang samatapi kadang-kadangdimulai denganhuruf kecil, kadang-kadang denganhuruf kapital)
	Periksapenyimpangan, apakahpenyimpanganadalah"*noise*" ataumungkin menunjukkanfenomena menarik
	Periksaapakah nilai-nilai yang ada masuk akal, (misalnya, semua *field*harussamaatau mendekatinilai yang sama)

2.1.2.3*Data Preparation*

2.1.2.3.1 Select data

Pada tugas ini ditentukandata yang akandigunakan. Kriteriayang dipakai meliputirelevansi dengantujuan*data mining*, kualitas,dan kendalateknisseperti bataspada volumedata ataujenisdata.*Output* yang dihasilkan :

- *Rationale foe inclusion/exclusion*

  Berupa daftardata yang akandigunakan/dikeluarkandan alasan untukkeputusan ini.Aktivitas pada point ini:

  - Mengumpulkan datatambahan yang sesuai(dari berbagai sumber-internalmaupun eksternal)
  - Melakukansignifikansidankorelasites untukmemutuskan apakahsuatu field harus disertakan
  - Mempertimbangkan kembalikriteria *select data*(bagian 2.1.2.1.3.1) dengan melihat kualitas datadan dataeksplorasi(misalnya, mungkin inginmenyertakan*data set*yang lain)
  - Mempertimbangkan kembalikriteria*selectdata*(bagian 2.1.2.1.3.1) dengan melihat pemodelan(mungkin model menunjukkan dibutuhkannya *dataset*lain)
  - Pilih*data set* yang berbeda (misalnya, atribut yang berbeda, hanya data yangmemenuhikondisi tertentu)
  - Pertimbangkanpenggunaan tekniksampling (misalnya, sebuah solusi cepatmungkin memerlukan *splitting test* dandataset pelatihanataumengurangiukuran*dataset*, jika alattidak dapat menanganiukuran *dataset*)
  - Dasar pemikiran untukinklusi / eksklusi
  - Periksateknik yang tersediauntuk datasampel

2.1.2.3.2 Clean data

Tugas ini bertujuan meningkatkan kualitasdata ketingkat yang dibutuhkanolehteknikanalisis.Melibatkan pemilihan*subset*data yang bersih,penyisipan*default*yang cocok.*Output* yang dihasilkan:

- *Data Cleaning Report*

  Menjelaskankeputusan dan tindakanyang diambiluntuk mengatasimasalah kualitasdata yang dilaporkanpada tahap *verify data quality*. Aktivitas pada point ini:

  - Mempertimbangkan kembalibagaimana menangani*noise*.
  - Memperbaiki, menghapus, atau mengabaikan*noise.*
  - Putuskanbagaimana menanganinilai-nilaikhusus danmaknanya. Contohnilai-nilai khususyang bisatimbul melaluimengambilhasil surveidi manabeberapa pertanyaantidak ditanyakanatau tidakmenjawab. Hal ini dapatmengakibatkannilai99untuk datayang tidak diketahui. Sebagai contoh,99 untukstatus perkawinanatauafiliasi politik. nilai-nilai khususjugabisa munculketika datadipotong-misalnya, 00oranguntuk100tahunatausemua mobildengan100.000kmdi odometer.
  - Mempertimbangkan kembalikriteria seleksidata(Lihat Tugas2.1.2.1.3.1) dalam pembersihan data(misalnya, memasukkan/mengecualikan*dataset* lain).

2.1.2.3.3 Construct data

Tugas ini meliputioperasipersiapan konstruktifdatasepertimendapatkanatribut,menyelesaikanrekor baru, atau perubahan nilai-nilaiuntuk atributyang ada.*Output* yang dihasilkan:

- *Derived attributes*

  *Derived attributes*adalah atributbaruyangdibangun darisatu atau lebih atributyang adadirecord yang sama. Contoh: luas =panjang*lebar.*Derived attributes*dibangunkarena:

  - Latar belakang meyakinkan bahwa beberapafaktapentingdan harusdiwakilimeskipun saat initidak memilikiatributyang mewakili.
  - Hasil darifasepemodelanmenunjukkan bahwafakta-fakta tertentuyangtidak tercakup

  Aktivitas pada point ini terdiri dari:

  - Putuskan apakah setiap atribut harus dinormalisasi (misalnya, bila menggunakan algoritma *clustering* dengan usia dan pendapatan, dalam mata uang tertentu, pendapatan akan mendominasi)
  - Pertimbangkan untuk menambahkan informasi baru tentang pentingnya relevan atribut dengan menambahkan atribut baru (misalnya, bobot atribut, normalisasi bobot)
  - Bagaimana atribut hilang dibangun atau diperhitungkan?
  - Tambahkan atribut baru untuk data yang diakses.

- *Generated records*

  Generated records adalah record baru yang menambah pengetahuan baru.

2.1.2.3.4 Integrate data

Pada tugas ini dilakukan penggabungan informasi dari beberapa tabel atau sumber informasi lain untuk membuat *record* atau nilai-nilai baru. *Output* yang dihasilkan:

- *Merged data*:

Merged Table mengacu pada bergabungnya dua atau lebih tabel yang memiliki informasi yang berbeda dalam objek yang sama. Contoh: pada rangkaian ritel memilikisatu tabeldengan informasi tentangkarakteristik umummasing-masingtoko(misalnya, ruang lantai, jenismall), tabel lainberisi ringkasandata penjualan(misalnya, laba, perubahan persen penjualan daritahun sebelumnya), dan informasi lain tentangdemografidaerah sekitarnya.Masing-masingtabeltersebutberisi satu*record*untuk setiap toko. Tabel inidapat digabungkanketabel baru dengan*record* untuksetiap toko merupakan penggabungan*fields* dari tabelsumber.

- *Aggregation*

  Merupakan operasi di mana nilai-nilai baru dihitung dengan meringkas informasi dari beberapa *record* atau tabel. Aktivitas yang dilakukan:

  - Periksa apakah fasilitas integrasi mampu mengintegrasikan sumber masukan yangdiperlukan.
  - Mengintegrasikan sumber dan hasil penyimpanan
  - Mempertimbangkan kembali kriteria seleksi data (bagian 2.1.2.1.3.1) mengingat pengalaman integrasi data (misalnya, Anda mungkin ingin untuk memasukkan / mengeluarkan *dataset* yang lain)

2.1.2.3.5 Format Data

Transformasi format mengacu pada sintak modifikasi yang dibuat untuk tidak mengubah makna data, tapi mungkin diperlukan oleh alat pemodelan.*Output* yang dihasilkan:

- *Reformatted data*.

Beberapa alat memiliki persyaratan pada urutan atribut, seperti *field* pertama menjadi unik pengidentifikasi untuk setiap record atau *field* terakhir menjadi bidang hasil model adalah untuk memprediksi. Aktivitas pada point ini:

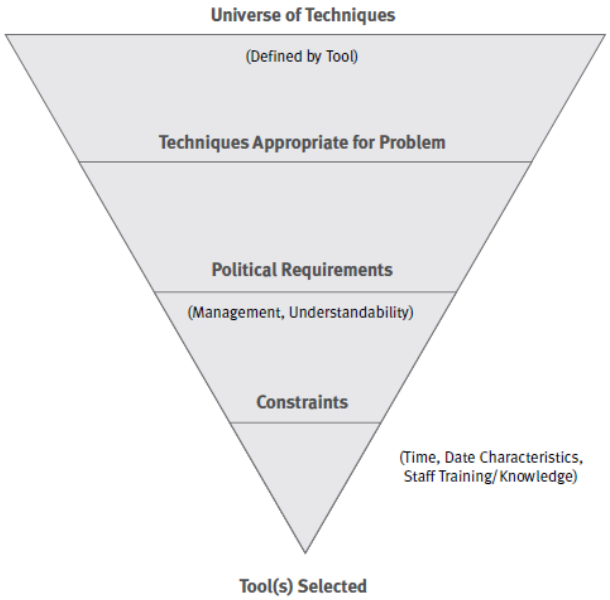
	Aktivitas
Rearranging attributes	Beberapa alat memiliki persyaratan pada urutan atribut, seperti *field* pertama menjadi unik pengidentifikasi untuk setiap record atau *field* terakhir menjadi bidang hasil model adalah untuk memprediksi.
Reordering records	Ini mungkin penting untuk mengubah urutan catatan dalam *dataset*. Mungkin alat pemodelan membutuhkan bahwa catatan diurutkan sesuai dengan nilai hasil atribut.
Reformatted within-value	Perubahan sintaksis yang dibuat untuk memenuhi persyaratan alat pemodelan tertentu
	Mempertimbangkan kembali kriteria seleksi data (bagian 2.1.2.1.3.1) berdasarkan pengalaman pembersihan data.

2.1.2.4 *Modelling*

2.1.2.4.1 Select modeling technique

Pada tahap *business understanding* telah dipilih alat dan teknik yang akan digunakan. Untuk tahap ini dipilih teknik pemodelan yang lebih spesifik, misalnya, *decision-tree*dengan5.0, atau *neural network generation*dengan *back*

propagation.Jika beberapateknik diterapkan, lakukan tugas inisecara terpisah untukmasing-masing teknik.



Gambar 2-5*Universe of Techniques*

Untukmasalah-masalah tertentu,hanyabeberapa teknikyang sesuai. "*Political Requirements*" dan *Constraints* dapat membatasipilihan teknik*data mining*.Mungkinhanyasalah satu alatatau teknikтersediauntuk memecahkanmasalah yang dihadapidanmungkin alat yang dipilihtidakbenar-benaryang terbaik, dari segiteknis.

Output yang dihasilkan pada tugas ini:

- *Modeling Technique*

  Catatan tentang teknik pemodelanyang digunakan. Tentukan teknik yang tepat, mengingatalat yang dipilih.

- *Modeling Assumptions*

Banyak teknikpemodelanmembuat asumsitertentutentang data, contohnyasemuaatributmemiliki distribusiseragam, tidak adanilai-nilaiyang hilangdiperbolehkan, atribut kelasharussimbolik, dll. Catatasumsiyangdibuat.

2.1.2.4.2 Generate test design

Sebelummembangun sebuah model, diperlukanprosedur ataumekanismeuntuk mengujikualitas danvaliditasmodel. Misalnya, dalamtugas*supervised**data mining*seperti klasifikasi, digunakantingkat kesalahansebagaiukuran kualitasuntuk modeldata mining. Oleh karena itu,kita biasanyamemisahkan*dataset*ke dalampelatihandan tes, membangunmodeldi set pelatihan, dan memperkirakankualitasdi settes terpisah.*Output* yang dihasilkan:

- *Testdesain*

  Jelaskanrencanayang ditujukan untukpelatihan, pengujian, dan evaluasimodel. Sebuah komponenutama dari rencanaadalah menentukanbagaimana membagidatasetyang tersediadalamdatasetpelatihan, pengujian, danvalidasi.

2.1.2.4.3 Build model

Jalankanalat pemodelanpada*dataset*untukmembuat satu atau lebihmodel.*Output* yang dihasilkan:

- *Parameter Settings*

  Denganalatpemodelan, sering adasejumlah besarparameter yang dapatdisesuaikan.Buatlah daftarparameterdan nilai-nilaiyang mereka pilih, bersama denganalasanuntuk pilihanpengaturan parameter.

- Model

  Ini adalah modelyang sebenarnyayang dihasilkan olehalat pemodelan, bukanlaporan.

- Model*descriptions*

  Jelaskanmodel yang dihasilkan. Laporkaninterpretasimodeldandokumenkesulitan yang ditemui denganmaknanya.

2.1.2.4.4 Assess model

*Engineer data mining*menginterpretasikan model menurut pengetahuan domain-nya,kriteria keberhasilan data mining, dan rancangan yang diinginkan. *Engineer data mining* menilai keberhasilanpenerapan pemodelan dan secara teknis teknik penemuan, menghubungkan analis bisnis dan ahli data mining dalam rangka untuk membahas hasil data mining dalam konteks bisnis. Harap dicatat bahwa tugas inihanya mempertimbangkan model, sedangkan tahap evaluasi memperhitungkan semua hasil lain yangdiproduksi dalam perjalanan proyek. *Engineer data mining* mencoba untuk menentukan peringkat model. Dia menilai model sesuai dengan kriteria evaluasi. Sebisa mungkin, ia juga memperhitungkan tujuan bisnis dan kriteriakesuksesan bisnis.Dalam kebanyakan proyek data mining, *Engineer data mining* menerapkan teknik tunggal lebih dari sekali, ataumenghasilkan hasil *data mining* dengan beberapa teknik yang berbeda. Dalam tugas ini, ia juga membandingkan semua hasilsesuai dengan kriteria evaluasi.*Output* yang dihasilkan:

- Penilaian Model

Merangkumhasildaritugas ini, daftarkualitasmodelyang dihasilkan(misalnya, dalam hal akurasi), dan peringkatkualitas merekadalam hubungan satu samalain.

- Revisi pengaturan parameter

  Berdasarkanpenilaian model, merevisipengaturanparameteruntukjangka berikutnyaada padatugas *BuildModel*.Iterasipembangunandan penilaian modelsampai Andapercaya bahwatelah menemukanmodel yang terbaik. Dokumentasikansemua revisidanpenilaian tersebut.

2.1.2.5 *Evaluation*

2.1.2.5.1 Evaluation Result

Langkah-langkahevaluasiditangani denganfaktor-faktor sepertiakurasi dangeneralitasmodel.langkah inimenilaisejauh manamodelmemenuhitujuan bisnisdan berusaha untukmenentukan apakah adabeberapa alasanmengapamodel ini kurang. Pilihan lainadalah untuk mengujimodel dengan testingdalamaplikasi nyata.

Hasil*data mining*melibatkanmodelyangselalu berhubungandengantujuan bisnisdansemua temuanlain yang tidakselalu berhubungandengantujuan bisnisasli, tetapimungkinjugamengungkaptantangan tambahan,informasi, ataupetunjuk untukarahmasa depan.*Output* yang dihasilkan:

- Penilaian hasil*data mining*berdasarkan kriteria keberhasilan bisnis.

  Meringkas hasil penilaian dalam hal kriteria keberhasilan usaha, termasuk pernyataan akhir mengenaiapakah proyek tersebut sudah memenuhi tujuan bisnis awal.

- Persetujuan model.

  Setelah menilai model berhubungan dengan kriteria keberhasilan bisnis, model yang dihasilkan yang memenuhi kriteria yang dipilih menjadi model yang disetujui.

2.1.2.5.2 Review process

Pada point ini, model yang dihasilkantampakmemuaskan danuntuk memenuhi kebutuhanbisnis. Tepat untuk melakukantinjauan lebihmenyeluruh tentangketerlibatan*data mining*untukmenentukan apakah adafaktorpentingatau tugasyang telah diabaikan.Ulasan inijugamencakupkualitasisu-jaminan,contoh:Apakah kitabenar membangunmodel? Apakah kitahanya menggunakanatribut yangdiijinkan untukdigunakan danyang tersedia untukanalisismasa depan?

Output yang dihasilkan adalah tinjauan ke bagian proses, ringkasanproses tinjauandankegiatan yang telahdilewatkan danharusdiulang.

2.1.2.5.3Determine next steps

Tergantung padahasil penilaiandantinjauanproses, tim proyekmemutuskanbagaimanamelanjutkannya. Tim memutuskanapakah akanmenyelesaikan proyekinidan beralih ke*deployment*, memulaiiterasilanjut, ataumengaturproyek*data mining*baru. Tugas inimeliputianalisissumber dayayang tersisa dananggaran, yang dapat mempengaruhikeputusan.*Output* yang dihasilkan:

- Daftar dari tindakan yang mungkin dilakukan

  Daftarpotensitindakan lebih lanjut, bersama denganalasan setiap pilihan.

- Keputusan

  Jelaskankeputusantentang bagaimana untukmelanjutkan, bersama dengandasar pemikiran.

2.1.2.6 *Deployment*

2.1.2.6.1 Plan deployment

Tugas inimembutuhkanhasilevaluasi danmenentukanstrategiuntuk *deploy*. Jika prosedurtelahtelah diidentifikasiuntuk menciptakanmodelyang relevan, prosedur inididokumentasikan di siniuntuk penyebarannanti.*Output* yang dihasilkan:

- *Deployment Plan*

  Merangkumstrategi penyebaran, termasuklangkah-langkahyang diperlukan dancara untuk melakukan itu.

2.1.2.6.2 Plan monitoring and maintenance

Pengawasan danpemeliharaanmerupakanisu-isu pentingjika hasil*data mining*menjadibagian daribisnis danlingkungannya.Persiapan yang cermatdaristrategi pemeliharaanmembantu untuk menghindariyang tidak perlu dalamjangka waktuyang lama untukpenggunaan yang salahdarihasil*data mining*.Untuk mengawasipenyebaranhasil*data mining*, dibutuhkan rencanaprosespengawasanyang rinci. Rencana inimemperhitungkanjenis spesifikdaripenyebaran.*Output* yang dihasilkan:

- *Monitoringandmaintenanceplan*

  Merangkumpengawasandan strategipemeliharaan, termasuklangkah-langkahyang diperlukan dancara untuk melakukan itu.

2.1.2.6.3 Produce final report

Padaakhir proyek, tim proyekmenulissebuahlaporan akhir. Tergantung padarencana penyebaran, laporanmungkin hanyaringkasanproyekdan pengalaman(jikabelumdidokumentasikan) atau mungkinpresentasiakhir dankomprehensif darihasil*data mining*.*Output* yang dihasilkan:

- *Final Report*

  Iniadalahlaporan tertulisakhir dariketerlibatan*data mining*. Ini mencakupsemuakirimansebelumnya,meringkasdan mengaturhasil.

- *Final Presentation*

  Menyampaikan hasilnyakepadapelanggan.

2.1.2.6.4 Review project

Menilaiapa yang benardanapa yang salah, apa yang telah dilakukandengan baik danapayangperlu ditingkatkan.*Output* yang dihasilkan:

- *Experience*Documentation

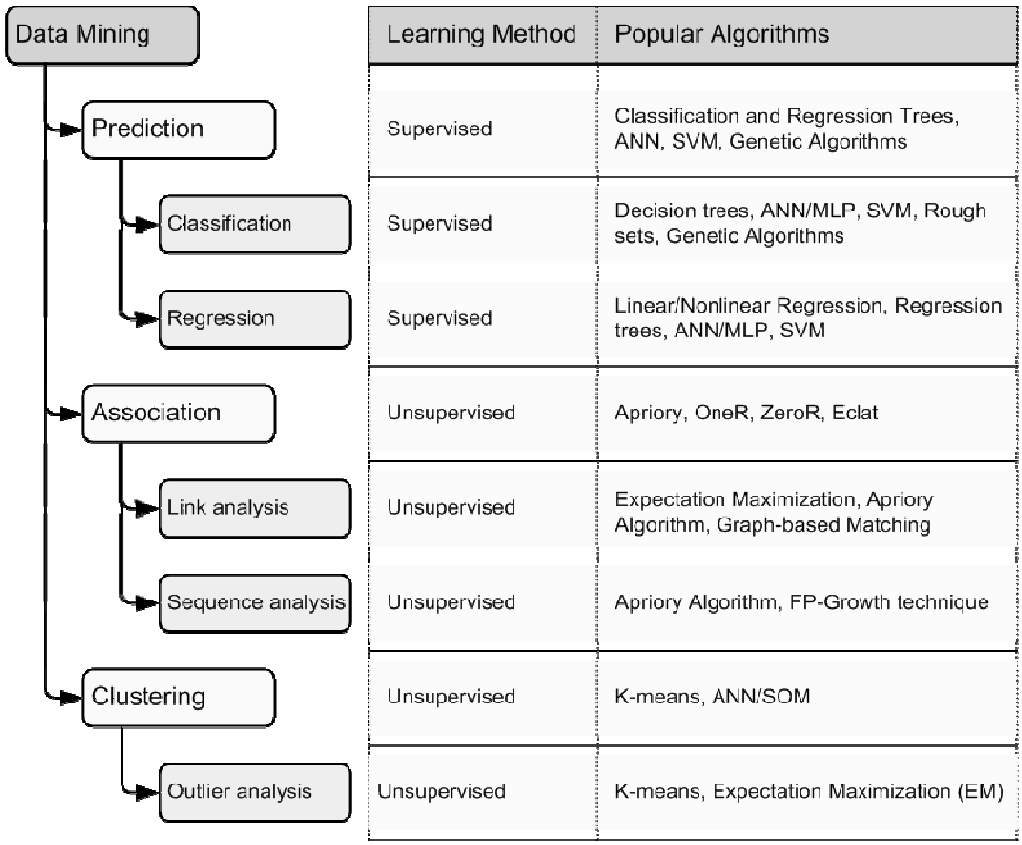
  Meringkaspengalaman pentingyang diperoleh selamaproyek.Sebagai contoh,pendekatanyang menyesatkan,ataupetunjuk untukmemilihyang terbaik sesuaiteknik*data mining*dalam situasi yang samabisa menjadi bagian daridokumentasi. Dalam proyek-proyekyang ideal, dokumentasi pengalamanjugamencakup setiaplaporanyang telahditulis oleh anggotaproyek individuselama faseproyek.

2.1.3 Tugas Data Mining

Turban, Sharda, Delen, danKing (2011, p.162) menjelaskan bahwa dengan menggunakandata yang adadan relevan, *data mining*membangunmodeluntuk

mengidentifikasipola-poladi antaraatributyang disajikan dalam*dataset*.Modeladalahrepresentasi matematis(hubungan linearsederhana dan/atau hubungannonlinearyang kompleks) yang mengidentifikasipola-poladi antaraatributdari objek.

Berdasarkan carauntuk mendapatkan pola dari data historis, algoritma pembelajaran dari metode *data mining*terbagi menjadi *supervised* dan *unsupervised*. Dengan algoritma *supervised learning*, data pelatihan mencakup atribut deskriptif (variabel independen) serta atribut kelas (variabel *output* atau variabelhasil).Sebaliknya, dengan*unsupervisedlearning*, datapelatihanhanya menyertakanatributdeskriptif



Learning Method	Popular Algorithms
Supervised	Classification and Regression Trees, ANN, SVM, Genetic Algorithms
Supervised	Decision trees, ANN/MLP, SVM, Rough sets, Genetic Algorithms
Supervised	Linear/Nonlinear Regression, Regression trees, ANN/MLP, SVM
Unsupervised	Apriory, OneR, ZeroR, Eclat
Unsupervised	Expectation Maximization, Apriory Algorithm, Graph-based Matching
Unsupervised	Apriory Algorithm, FP-Growth technique
Unsupervised	K-means, ANN/SOM
Unsupervised	K-means, Expectation Maximization (EM)

Gambar 2-6*Taxonomy* berdasarkan tugas *data mining*

(Turban, Sharda, Delen, dan King, 2011)

Secara umum,data miningberusaha untukmengidentifikasiempatjenis tugas:

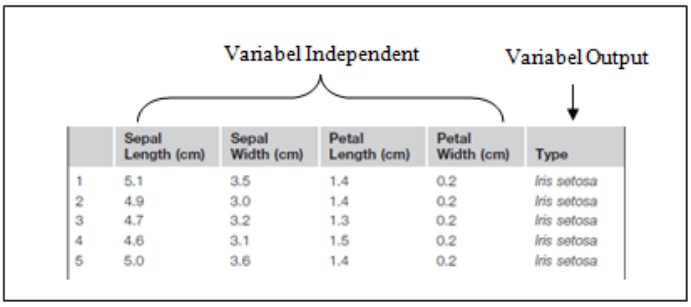
a. Asosiasimenemukanpengelompokanumumdimana terjadi pada saat bersamaan, seperti birdan popokakanbersama-samadalam analisis*market-basket*.
b. Prediksimemberitahuyang terjadi di masa depanpada peristiwatertentu, berdasarkanapa yang telah terjadidi masa lalu, seperti meramalkanpemenang*super bowl*atauperamalantemperaturhari tertentu.
c. Clustermengidentifikasipengelompokanberdasarkankarakteristikyang diketahui, sepertimenempatkanpelanggan dalamsegmen yang berbedaberdasarkandemografidanperilakupembelianmasa lalu.
d. Hubungansekuensialmenemukanwaktu-mengurutkan kejadian, seperti meramalkanbahwanasabah perbankanyang ada,yang sudah memilikirekening giroakanmembukarekening tabungandiikuti olehakuninvestasidalam setahun.

2.1.3.1 Klasifikasi

Turban, Sharda, Delen, dan King(2011, p.178) memaparkan klasifikasi sebagai*machine learning*yaitu pembelajaranuntuk menemukan pola daridata *history*(satu set informasi-ciri, variabel, fitur padakarakteristikitemyang sebelumnyadiberi label) dalam rangka untuk menempatkandata baru(tidakdiketahuilabelnya) ke dalam kelas.Misalnya,menggunakan klasifikasi awanuntuk memprediksi apakahcuacapada haritertentu akan"cerah", "hujan" atau"berawan". Selain itu tugasklasifikasidigunakan juga untukpersetujuan

kredit(risiko kredit misalnyabaik atau buruk), lokasi toko(misalnya baik, moderat, buruk), targetpemasaran (misalnya kemungkinanmenjadi pelanggan atau tidak).

Klasifikasimempelajarifungsiantarakarakteristik(variabel *independen/ input*) dankeanggotaan mereka(variabel *output*) melalui proses*supersived learning*, di manakedua jenisvariabel(*input dan output*) digunakan oleh algoritma. Berikut adalah contoh data yang digunakan pada klasifikasi:



	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)	Type
1	5.1	3.5	1.4	0.2	*Iris setosa*
2	4.9	3.0	1.4	0.2	*Iris setosa*
3	4.7	3.2	1.3	0.2	*Iris setosa*
4	4.6	3.1	1.5	0.2	*Iris setosa*
5	5.0	3.6	1.4	0.2	*Iris setosa*

Gambar 2-7 Contoh data klasifikasi

Dua langkah dalam metodologiklasifikasitipe prediksi:

- Model pembangunan/pelatihan

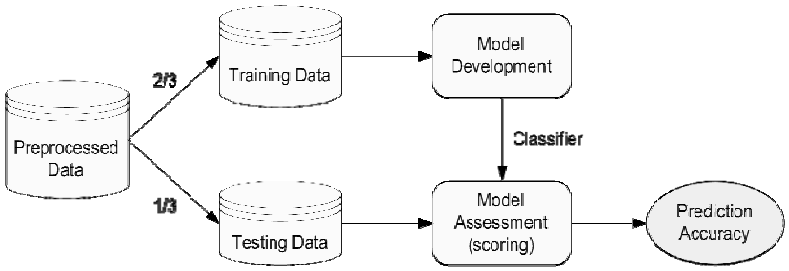
  Dalam model pembangunan/pelatihan dikumpulkandata yang akandigunakan terdiri dari variabel input dan labelkelas yang sebenarnya.Kemudian dilakukan pelatihan (*training*) menggunakan algoritma yang tersedia.

- Modelpengujian/penyebaran.

  Setelahdilatih, modeldiuji terhadapsampelterakhir untuk penilaianakurasidan digunakan untukmemprediksikelassebuah data baru(dimana labelkelastidak diketahui) dengan melihat nilai dari variabel *input* yang dimiliki.

2.1.3.1.1 Simple Split

Simple Split (atau *holdout*atau estimasisampel uji) melakukan partisidatamenjadi duasubseteksklusifdisebutsetpelatihan danset uji.Metode ini membagi 2/3 dari data sebagaisetpelatihan dan sisanya1/3 sebagai set uji. Set pelatihandigunakan oleh*modelbuilder*dan membangunklasifikasikemudiandiuji dengansampel uji. Aturan ini digunakan dengan asumsi bahwa data yang ada dalam dua subset memiliki jenis yang sama (misalnya, memiliki sifat yang benar-benar sama).



Gambar 2-8Metode *Simple Split*
(Turban, Sharda, Delen, dan King, 2011)

2.1.3.1.2 *Principal Component Analysis*(PCA)

Principal Component Analysis (PCA) melakukan pemetaan/transformasi set data dari dimensi lama ke dimensi baru (yang relative berdimensi lebih rendah) dengan memanfaatkan teknik dalam aljabar linier, tanpa memerlukan masukan parameter tertentu dalam memberikan keluaran hasil pemetaannya.

PCA memerlukan masukan data yang mempunyai sifat zero-mean pada setiap fiturnya. Sifat zero mean pada setiap fitur data bisa didapatkan dengan mengurangkan semua nilai dengan rata-ratanya. Set data X dengan dimensi MxN, dimana M adalah jumlah data dan N adalah jumlah fitur, akan tampak seperti berikut:

$$X=\begin{bmatrix} x_{11} & x_{12} & x_{1j} & x_{1N} \\ x_{21} & . & . & x_{2N} \\ x_{i1} & . & . & x_{iN} \\ x_{\mathrm{M1}} & x_{M2} & x_{Mj} & x_{MN} \end{bmatrix}$$

Untuk fitur ke-j, semua nilai pada kolom tersebut dikurang dengan rata-ratanya, diformulasikan dengan:

$$x_{ij} = x_{ij} - \bar{x}_j$$

i = 1,2,…..,M dan j adalah kolom ke-j

selanjutnya dilakukan perhitungan matriks kovarian dari matriks X, yaitu C_x. Formula yang digunakan adalah dot-product pada setiap fitur.

$$C_x = \frac{1}{M} X^T . X \quad \text{.....(3.1)}$$

N adalah jumlah fitur, sedangkan X^Tadalah matriks transpos dari X.

$$C_x = \frac{1}{M}\, x \begin{bmatrix} x_{11} & x_{21} & x_{i1} & x_{M1} \\ x_{12} & . & . & x_{M2} \\ x_{1j} & . & . & x_{Mi} \\ x_{1N} & x_{2N} & x_{iN} & x_{NM} \end{bmatrix} x \begin{bmatrix} x_{11} & x_{12} & x_{1j} & x_{1N} \\ x_{21} & . & . & x_{2N} \\ x_{i1} & . & . & x_{iN} \\ x_{\mathrm{M1}} & x_{M2} & x_{Mj} & x_{MN} \end{bmatrix}$$

$$= \frac{1}{M}\, x \begin{bmatrix} x_{11} & x_{21} & x_{1j} & x_{1N} \\ x_{21} & . & . & x_{2N} \\ x_{i1} & . & . & x_{iN} \\ x_{N1} & x_{N2} & x_{Nj} & x_{NN} \end{bmatrix}$$

Pada matriks C_x , elemen ke-ij adalah inner-product antara baris matriks X_T dengan kolom matriks X. Sifat-sifat yang dimiliki oleh matriks $C_x$adalah sebagai berikut:

- C_x adalah matriks simetris bujur sangkar berukuran NxN
- Bagian diagonal utama (dari kiri atas ke kanan bawah) adalah nilai varian masing-masing fitur sesuai dengan indeks kolomnya.
- Bagian selain diagonal utama adalah konvarian di antara pasangan dua fitur yang bersesuaian.

Jadi, matriks C_x menangkap kovarian diantara semua pasangan yang mungkin dari fitur data set data matriks X. Nilai kovarian merefleksikan noise dan redundansi pada fitur:

- Dalam diagonal utama, asumsinya adalah nilai yang tinggi berkorelasi dengan struktur yang penting.
- Dalam elemen selain diagonal utama, nilai jarak yang besar menandakan redundasi yang tinggi.

Jika Yadalah matriks set data hasil pemetaan dan C_y adalah matriks kovarian dari Y, yang diharapkan dalam PCA adalah:

- Semua elemen selain diagonal utama dalam C_y harus nol. Maka, C_y harus matriks diagonal. Dengan kata lain, Y adalah matriks terdekorelasi.
- Peletakan dimensi dalam Y dari kiri ke kanan diurutkan secara menurun.

Cara yang umum digunakan untuk mendapatkan C_y adalah dengan eigenvalue dan eigenvector. Eigenvalue dan eigenvektor dari matriks X berturut-turut adalah nilai skala λdan vector u yang memenuhi persamaan berikut:

$$Xu = \lambda u$$

Dengan mencari matriks ortonormal P dimana Y = PX dan $C_y = \frac{1}{M} YY^T$ adalah matriks diagonal, dan kolom dari P adalah komponen utama (principal components) dari X, persamaan $C_y$bisa dijabarkan sebagai berikut:

$$C_y = \frac{1}{\mathrm{M}} \mathrm{YY^T}$$

$$= \frac{1}{\mathrm{M}} (\mathrm{PX})(\mathrm{PX})^{\mathrm{T}}$$

$$= \frac{1}{\mathrm{M}} \mathrm{PXX^TP^T}$$

$$= \mathrm{P}\left(\frac{1}{\mathrm{M}}\mathrm{XX}^{\mathrm{T}}\right)\mathrm{P}^{\mathrm{T}}$$

Dengan mensubstitusikan persamaan 3.1, kita mendapatkan matriks $C_y$berdimensi NxN:

$$C_y = PC_xP^T$$

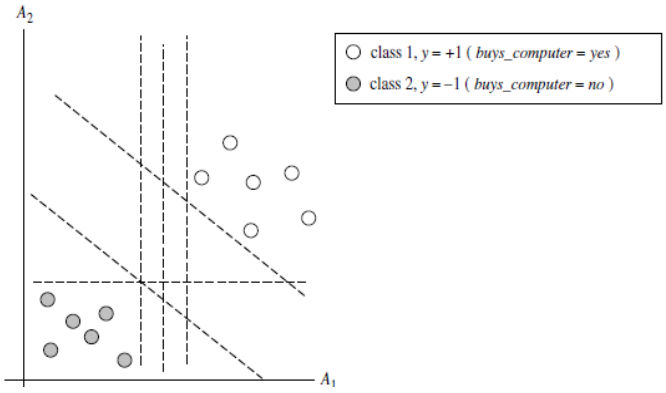
Setiap baris matriks P adalah eigen vector C_x

2.1.3.1.3 Algoritma Klasifikasi

2.1.3.1.3.1 Support Vector Machine(SVM)

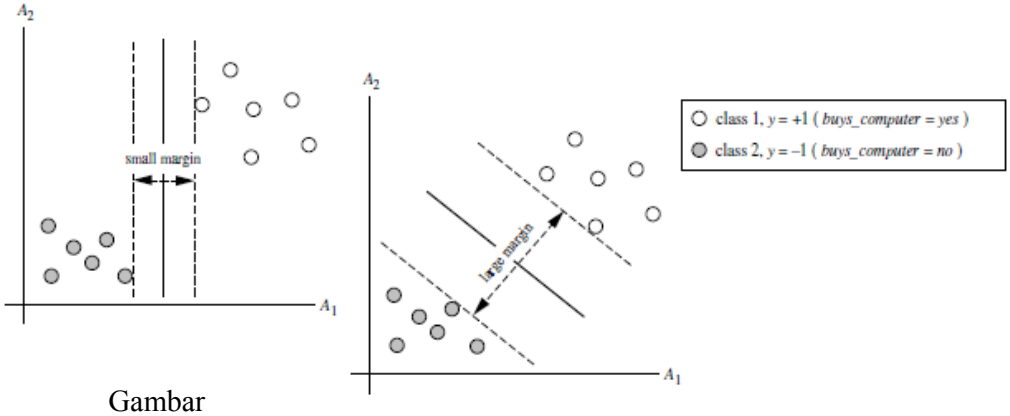
SVMadalahalgoritmaklasifikasibaik linier maupun nonlinear, pola data yang kompleksdan menghindari*overfitting*, situasidi manamodelmenghafalpolahanya yang relevan dengankasus-kasus analisis tertentu.Han dan Kamber (2006, p.337-343) memaparkan cara kerja algoritma SVMdengan pemetaannonlinieryaitu mengubahdatapelatihanasli kedimensi yang lebih tinggi.Dalamhal ini,dimensi baruakanmencarilinieroptimalyang memisahkan data ("*decision boundary*"memisahkantupeldarisatu kelasdenganyang lain). SVMmenemukan*hyperplane*menggunakan*support vektor*("esensial" pelatihantuple) dan margin.

Berikut adalah penjelasan SVM dalam dua kelasdimana kelas yang dipisahkan secaralinear.Kumpulan data D terdiri dari $(x_1, y_1), (x_2, y_2), \ldots, (x_{|d|}, y_{|d|})$, dimana X_i adalah himpunan tupelpelatihan dengan label kelas terkait, Y_i.Setiap Y_i dapat mengambil salah satu daridua nilai, baik +1 atau -1 (yaitu, $Y_i \epsilon \{+1, -1\}$), Untuk membantu dalam visualisasi,digunakan dua atribut masukan, A_1 dan A_2, seperti gambar berikut:



Gambar 2-9 Data *training*dua dimensi(Han & Kamber, 2006)

Dari grafik, kita melihatbahwa2-D datalinear, garis lurusdapat ditarikuntuk memisahkansemuatupelkelas+1darisemuatupelkelas-1. Sebuahpendekatanuntuk masalah menemukan*hyperplane*terbaik adalahdengan mencariMMH.



Gambar 2-10Kemungkinan duapemisah*hyperplane* danmarginyang terkait (Han & Kamber, 2006)

Kedua*hyperplane*dapatmengklasifikasikansemua datatupel yang diberikan. *Hyperplane*denganmarginyang lebih besarlebihakuratmengklasifikasikanmasa depandata tupeldaripada*hyperplane*denganmargin yang lebih kecil.

Sebuah*hyperplane*pemisahdapat ditulis sebagai:

$$WX+b=0$$

Wadalahvektor bobot, yaitu, $W=\{w_1, w_2, \ldots, w_n\}$, n adalahjumlahatribut,danbadalahskalar, sering disebut sebagaibias.Pada

tupelpelatihan2-D, misalnya, $X = (x_1, x_2)$, dimana$x_1$dan$x_2$adalah nilai-nilaiatributA_1danA_2, masing-masing. Jikakita berpikir tentangbsebagaibobot tambahan, w_0, kita dapatmenulis ulang*hyperplane*menjadi:

$$w_0 + w_1 x_1 + w_2 x_2 = 0$$

Dengan demikian, setiap titikyang terletak di atas*hyperplane*memenuhi:

$$w_0 + w_1 x_1 + w_2 x_2 > 0$$

Demikian pula, setiap titikyang terletak di bawah*hyperplane*:

$$w_0 + w_1 x_1 + w_2 x_2 < 0$$

Bobotdapatdisesuaikan sehingga*hyperplane*mendefinisikan"sisi" darimarginditulis sebagai:

$H_1 : w_0 + w_1 x_1 + w_2 x_2 \geq 1$ untuky$_i$=+1

dan

$H_2 : w_0 + w_1 x_1 + w_2 x_2 \leq -1$untuky$_i$=-1

Kedua persamaan di atas digabungkan, maka menghasilkan:

$$y_i (w_0 + w_1 x_1 + w_2 x_2) \geq 1, \forall i.$$

Setiaptuplepelatihanyang jatuh pada*hyperplane*H_1atauH_2(yaitu, "sisi" yang mendefinisikanmargin)yang memenuhipersamaan disebut*vektorSupport*. Artinya, merekasama-samadekat denganMMH. Untuk ukuranmarginmaksimal, jarak dari*hyperplane*untuk setiaptitik padaH_1adalah:

$$\frac{1}{||w||}$$

Dimana||W||adalah*Euclideannorm*dari W, yaitu$\sqrt{W.W}$ (jika W=$\{w_1, w_2, \ldots, w_n\}$ maka $\sqrt{W.W} = \sqrt{w_1^2 + w_2^2 + \cdots + w_3^2}$).

Dengan definisi, ini sama denganjarak darisetiap titik padaH_2ke*hyperplane*. Oleh karena itu, marginmaksimal adalah:

$$\frac{2}{||W||}$$

Setelah menemukan*support vektor*danMMH, kami memiliki *supportvektormachine*terlatih. MMHadalah batas kelaslinear, sehingga SVMyang sesuaidapat digunakan untukmengklasifikasikan datalinear.

Pendekatan yangdijelaskan untukSVMlineardapat diperluas untuk membuatSVMnonlinear.Ada dua langkah utama.Pada langkah pertama, mengubah input data asli ke dalam ruang dimensi yang lebih tinggimenggunakan pemetaan nonlinear. Beberapa pemetaan nonlinear umum dapat digunakan dalamlangkah ini. Setelah data telah berubah menjadiruang baru yang lebih tinggi, pencarian langkah kedua untuk *hyperplane* pemisahan linear dalam ruang baru.

Optimasi *quadratic* dapat diselesaikan dengan menggunakanperumusan linear SVM.MMH ditemukan di ruang yang barusesuai dengan *hypersurface* nonlinear dalam ruang aslinya.

Sebuahmasukan3DvectorX=(x_1, x_2, x_3) dipetakanke dalam ruang6D. Z, dengan menggunakanpemetaanφ_1(X) =x_1, φ_2(X) =x_2, φ_3(X) =x_3, φ_4(X) =$(x_1)^2$, φ_5(X) =x_1x_2, danφ_6(X) =x_1x_3. Sebuah*hyperplane*dalam ruangbaruadalah:

$$d(Z) = WZ+b$$

WdanZadalah vektor. Kami memecahkanW danbdankemudian mengganti kembalisehingga*hyperplane*keputusanlineardalam ruang(Z) yang barusesuai dengannonlinier- orde keduapolinomialdalam ruang3-D masukan asli:

$$d(Z) = w_1x_1 + w_2x_2 + w_3x_3 + w_4(x_1)^2 + w_5x_1x_2 + w_6x_1x_3 + b$$

$$= w_1z_1 + w_2z_2 + w_3z_3 + w_4z_4 + w_5z_5 + w_6z_6 + b$$

Tupelpelatihanhanya munculdalam bentuktitik,Φ(Xi).Φ(Xj), di manaΦ(X) hanyafungsipemetaannonlinearditerapkan untuk mengubahtupelpelatihan. Daripadamenghitungperkalian titikpadatupeldata yangberubah,ternyatasecara matematissetara denganmenerapkanfungsikernel, K(Xi,Xj), dengan datamasukan yang asli. Artinya,

$$K(Xi, Xj) = \Phi(Xi).\Phi(Xj)$$

Tigafungsikernelyang dapat diterimameliputi:

Polynominal kernel of degreeh: $K(X_i,X_j) = (X_i.X_j + 1)^h$

Gaussian radial basis function kernel: $K(X_i,X_j) = e^{-\|Xi-Xj\|/2\sigma^2}$

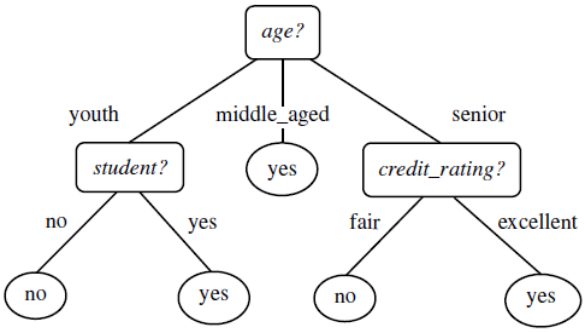
Sigmoid kernel:$K(X_i,X_j) = \tan h\, K(X_i \,.\, X_j - \delta)$.

Kelebihan pada SVM adalah solusi yang diberikan adalah global optima, tidak seperti Artificial Neural Network yang solusinya sering masuk pada wilayah local optima. Hal ini berarti SVM selalu memberikan model yang sama dan solusi dengan margin maksimal.

2.1.3.1.3.2 Decision Tree

*Decision tree*merupakan bagiandari"*machine learning* ", bagian penting darikecerdasan buatan(Lin, Shiue, Chen, & Cheng, 2009).Metode inimendekatiangka atausimbol-fungsi nilai sasaranyangkuat untukdata yang*noise* dan

mampu belajarekspresidisjungtif.Sedangkan Han dan Kamber (2006, p.337-343) menjelaskan *Decision tree*adalah flowchart- sepertistruktur pohon, di mana setiap simpul internal(simpultanpadaun) menunjukkanpengujian padaatribut, setiap cabangmerupakanhasiltes, dansetiap simpuldaun(atau simpulterminal) memilikilabelkelas.Simpulyang palingatas adalahsimpul akar.



Gambar 2-11 Struktur Decision Tree

Satu set sampel dalam partisi S, atribut X dipilihuntuk partisi set ke S1, S2,...,SL dan ini ditambahkan ke pohon keputusan sebagai anak-anak dari simpul untuk S. Simpul untuk S diberi label dengan uji X, dan partisi S1 , S2 , . . . , SL kemudianrekursif dipartisi . Berikut langkah dan perhitungan dalam algoritma *decision tree*:

- HitungInfo(S) untuk mengidentifikasikelaspadaset pelatihan S.

$$Info\ (S) = -\sum_{i=1}^{k}\{[\frac{freq(C_i,S)}{|S|}]log_2\ [\frac{freq(C_i,S)}{|S|}]\}$$

  Dimana|S|adalahjumlah kasuspada pelatihan set. $C_i$adalahkelas,i=1,2,...,k. kadalahjumlah kelasdanfreq(C_i, S) adalah nomor darikasus yang termasuk dalamC_i.

- Hitungnilaiinformasiyang diharapkan,$Info_x$(S) untuk fitur Xke partisiS.

$$Info_x(S) = -\sum_{i=1}^{L}\left[\left(\frac{|S_i|}{|S|}\right)info(S_i)\right],$$

  dimanaLadalah jumlahoutputuntuk fiturX, Siadalah bagian dariSyang sesuaidengan outputke idan|S|adalah jumlahkasusyang subsetS_i.

- Hitunginformation gainsetelahmempartisiberdasarkanfiturX.

- Hitungnilaiinformasipartisi,SplitInfo(X)diperoleh dariSyang dipartisi menjadisubsetL.

$$\frac{}{} \qquad \frac{}{}$$

- Hitungrasiogaindari Gain(X) terhadapSplitInfo(X)

GainRatio(X) = Gain(X) / SplitInfo(X)

TheGainRatio(X) mengimbangititik lemahdariGain(X), yang merupakan jumlahdariinformasi yang diberikan olehXdalam pelatihan set. Oleh karena itu,fiturdenganGainRatiotertinggi(X) diambil sebagai akarpohon keputusan.

2.1.3.1.4 Akurasi Metode Klasifikasi

Dalam klasifikasi, perhitungan akurasi menggunakan *confusion matrix*, disebut juga *classfication matrix*atau *contingency table*(Gambar 2-12).

		True Class: Positive	True Class: Negative
Predicted Class	Positive	True Positive Count (TP)	False Positive Count (FP)
Predicted Class	Negative	False Negative Count (FN)	True Negative Count (TN)

Gambar 2-12*Confusion Matrix*
(Turban, Sharda, Delen, dan King, 2011)

Angka sepanjang diagonal dari kiri atas ke kanan bawah mewakili keputusan yang benar, dan jumlah sisiluardiagonal mewakili kesalahan.Dari data *confusion matrix*, dapat dihitungakurasi umum untuk model klasifikasi.

Tabel 2-1Perhitungan akurasi untuk metode klasifikasi

Metric	Description
True Positive Rate = $\frac{TP}{TP+FN}$	The ratio of correctly classified positives divided by total positive count (i.e., hit rate or recall)
True Negative Rate = $\frac{TN}{TN+FP}$	The ratio of correctly classified negative divided by the total negative count (i.e., false alarm rate)
Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$	The ratio of correctly classified instances (positive and negative) divided by total number of instances

Perhitungan akurasimodelklasifikasi(atau *classifier*) dengan algoritmas*upervised learning*penting, karena dua alasan. Pertama, dapat digunakanuntuk memperkirakanakurasi prediksimasa depan, yangberartitingkat kepercayaan terhadap satu*outputclassifier*dalam sistemprediksi.Kedua, dapat digunakanuntukmemilih*classifier*(mengidentifikasi model klasifikasiterbaik).

2.2 *Data Warehouse*

Vercellis(2009, p.45) menjelaskan bahwa *data warehouse* merupakan repositoriutamauntuk menyediakan data dalammengembangkanarsitektur *business intelligence*dan *decision supportsystems*. Dimana *data warehouse* memiliki karakteristik: data statis, data saat inidan lampau, data agregasidankonsolidasi, dan berfungsi untuk analisis.

Sedangkan *data mart* berisi semuadata yang dibutuhkanoleh suatu departemen tertentu dalam perusahaanuntuk melakukananalisis *business intelligence*dan mengeksekusiaplikasi *decision support*sesuai fungsinya sendiri.

Dari pengertian di atas, *data mart*dapat dianggapsebagaifungsionalatau*data warehouse*departemen,denganukuran yang lebih kecildanjenis yang lebihspesifik daripada*data warehouse*.

Ada beberapa alternatif arsitektur *data warehouse* (Turban, Sharda, Delen, dan King,2011) yang diusulkan dalam merancang dan mengembangkan *data warehouse*, yaitu:

a. *Independent data mart*,arsitektur ini bisa dibilangyang paling sederhanadanalternatifarsitekturpaling mahal. *Data mart*yangdikembangkan memiliki tujuanberoperasi secara *independent*darisatu sama lain untukmelayani kebutuhanunit organisasimasing-masing. Karenakemandirian, memungkinkanuntuk memilikidefinisiyang tidak konsisten,dimensidan ukurandata yang berbeda, sehingga sulit untukmenganalisis dataseluruh*data mart*.

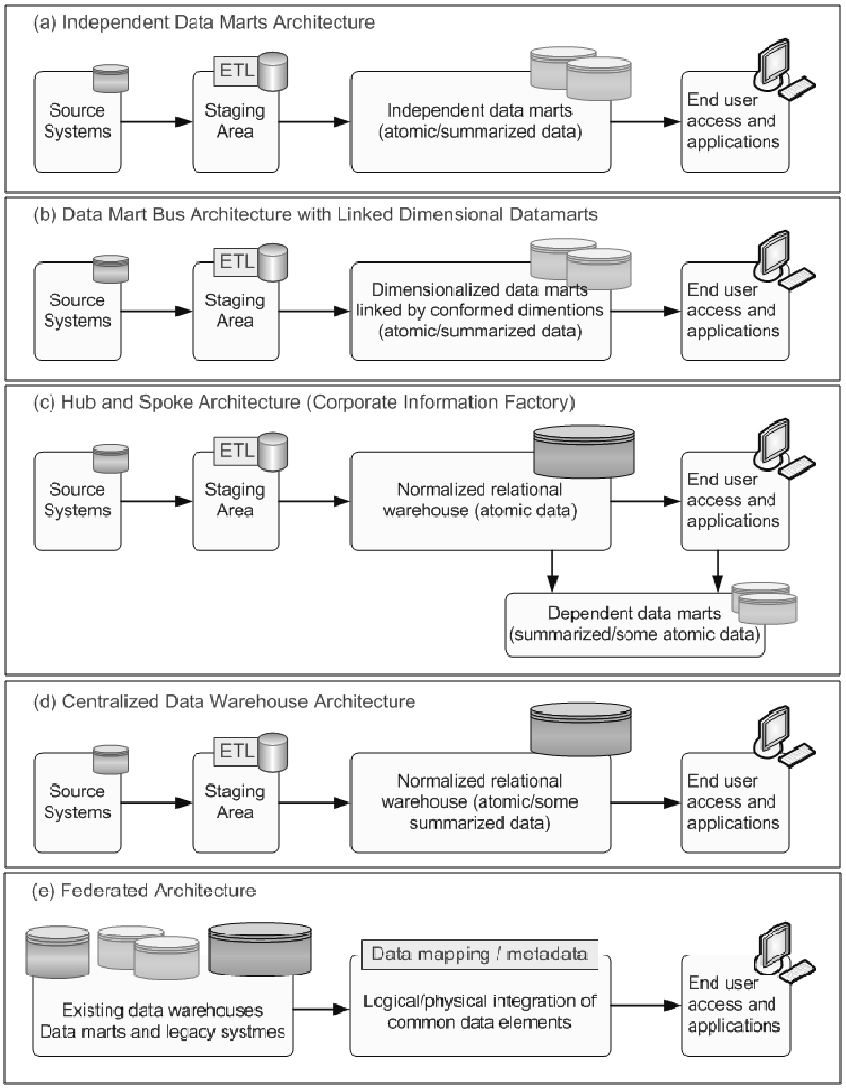
b. *Data mart bus architecture,* arsitektur ini merupakan alternatif untuk *data mart independent* dimana individu dihubungkan satu sama lain melalui beberapa jenis *middleware*. Karena data terkait antara individu*data mart*, ada kesempatan yang lebih baik menjaga konsistensi data seluruh perusahaan. Meskipun memungkinkan untuk permintaan data yang kompleks di seluruh *data mart*, kinerja untuk jenis analisis mungkin tidak pada tingkat yang memuaskan.

c. *Hub-and-spokearchitecture*, pada arsitektur iniperhatiandifokuskanpada pembangunanskalabilitasdankemampuan maintaininfrastrukturyang mencakup*data warehouse*terpusatdan beberapa*dependentdata mart*(masing-masing untuk unit organisasi). Arsitektur inimemungkinkan

untukkemudahan dankustomisasiantarmuka penggunadan laporan. Di sisi negatif, arsitektur inimemilikipandanganmenyeluruhperusahaan, dan dapat menyebabkanredundansi datadanlatencydata.

d. *Centralized data warehouse*,arsitektur *datawarehouse*terpusatmirip dengan*hub-and-spoke* architecturekecualitidak ada*data martdependent*, melainkan ada*datawarehouse*raksasayangmelayani untukkebutuhan semuaunit organisasi. Pendekatanterpusatmenyediakanpengguna denganakses ke semuadata dalam*data warehouse*. Selain itu, mengurangijumlah datatim teknisyang harusmentransferatau mengubah, sehinggamenyederhanakanpengelolaan datadan administrasi. Jikadirancang dan dilaksanakandengan baik, arsitektur inimemberikan pandangantepat waktu danmenyeluruhdariperusahaan untuksiapa, kapan, dandi mana punmereka beradadalam organisasi.

e. *Federated data warehouse*, menggunakansegala cara yang mungkinuntuk mengintegrasikansumber dayaanalitisdari berbagai sumberuntuk memenuhi perubahan kebutuhanatau kondisibisnis. Pendekatanfederasididukung oleh*vendormiddlewar*eyang mengusulkanpermintaandistribusidan kemampuanuntuk bergabung. Karena masalahkualitas kinerjadan data, kebanyakan ahli sepakat bahwapendekatanfederasibekerja dengan baikuntuk melengkapi*data warehouse*, tidakmenggantikannya.



Gambar 2-13 Arsitektur Data Warehouse
(Turban, Sharda, Delen, dan King, 2011)

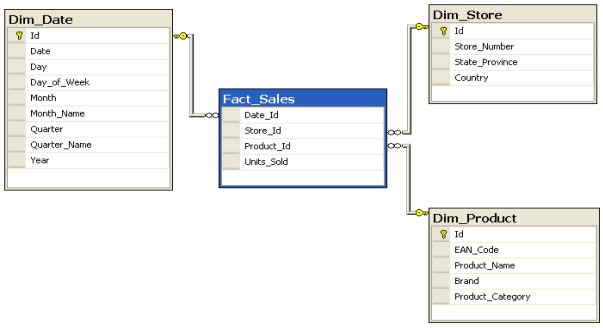
Beberapa perusahaanlebih memilih untukmerancang danmengembangkandengan carabertahapserangkaian *datamart*daripadasebuah c*entralized data warehouse*,dalam rangka untuk mengurangiwaktu pelaksanaan.

Step dalam pembuatan *data warehouse* yang paling utama adalah proses untuk ETL (*Extract Transform Load*).*Extract* yang dimaksud adalahproses untuk pengambilan data dari database sumber.*Transform*merupakan proses untuk

perubahan data yang telah diambil menyesuaikan dengan kebutuhan analisa. Dan *Load*adalah proses untuk menaikkan data hasil olahan ketarget.

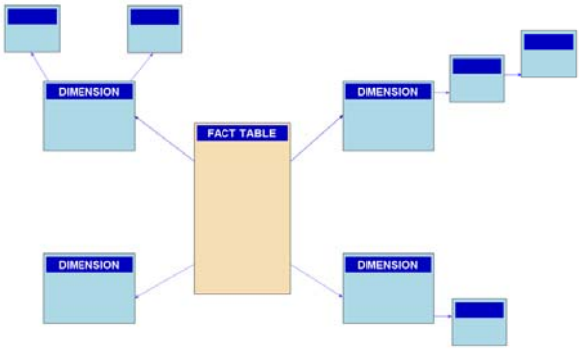
Dalam proses pembuatan *data warehouse*, diperlukan untuk mendesain schema dari database. Schema ini biasanya terdiri dari *fact* dan *measurements*, dimana *Fact* adalah sekumpulan data real yang sudah ada di sumber database, sedangkan *Measurement* adalah data olahan yang terdiri atas satu atau sekumpulan *fact* yang telah diolah menjadi suatu point tertentu.

Design *data warehouse* terdiri dari *Star Schema* dan *Snowflake Schema*. *Star schema* adalah skema yang paling banyak digunakan untuk desain *Data Warehouse*, dimana sekumpulan *Fact* dan *Measurements* dikelilingi oleh dimensi.



Gambar 2-14 Star Schema

Snowflake Schema adalah skema lanjutan dari *star scheme*, dimana dimensi yang ada *data warehouse* bisa berlapis dikarenakan data dimensi tersebut sudah dinormalisasi.



Gambar 2-15Snowflake Schema

2.3 Multi Finance dan Kredit

Perusahaan *Multifinance* merupakan perusahaan yang berperan serta dalam kegiatan yang berhubungan dengan produk-produk yang ada dalam pasar *multi finance*. Produk-produk yang ada yaitu sewa guna usaha atau *leasing*(salah satu cara perusahaan memperoleh asset atau kepemilikan tanpa harus melalui proses yang berkepanjangan), anjak piutang, *dan consumer financing*.

Anjak piutang merupakan salah satu instrument yang sering disebut dengan *factoring*, didefinisikan sebagai transaksi pembelian, penagihan serta pengurusan piutang atau tagihan jangka pendek klien (penjual) kepada perusahaan *factoring*. Kemudian akan ditagih oleh perusahaan anjak piutangkepada pembeli karena adanya pembayaran kepada klien oleh perusahaan *factoring*.

Consumer financing atau pembiayaan konsumen dimana dalam hal ini ada perusahaanyang bertindak sebagai penjamin dalam pemberian kredit kepada konsumen yang tidak menjadi satu kesatuan dalam perbankan melainkan perusahaan ini berdiri sendiri.Pembiayaan konsumen ini adalah suatupinjaman atau kredit yang diberikan oleh suatu perusahaan kepada debitur untuk pembelian barang atau jasa yang akan langsung digunakan oleh konsumen.

Kredit berasal dari bahasa italia yaitu kata *credere* yang memiliki arti percaya, dimana pihak yang mendapat kredit berarti orang tersebut dipercaya untuk mendapatkan pinjaman. Kasmir (2008 pp 103-105) menjelaskan unsur-unsur dalam kredit adalah:

a. Kepercayaan, yang berarti bahwa pemberi kredit yakin prestasi yang diberikannya baik dalam bentuk uang, barang, atau jasa, akan benar-benar diterimanya kembali dalam jangka waktu tertentu di masa yangakan datang.
b. Kesepakatan, di mana dituangkan dalam suatu perjanjian dan masing-masing pihak menandatangani hak dan kewajiban masing-masing.
c. Jangka Waktu, dimana mencakup masa pengembalian kredit yang telahdisepakati.
d. Resiko, faktor resiko kerugian dapat diakibatkan dua hal yaitu resiko kerugian yang diakibatkan nasabah sengaja tidak mau membayar kreditnya padahal mampu dan resiko kerugian yang diakibatkan karena nasabah tidak sengaja yaitu akibat terjadinya musibah seperti bencana alam. Penyebab tidak tertagih sebenarnya dikarenakan adanya suatu tenggang waktu pengembalian (jangka waktu). Semakin panjang jangka waktu suatu kredit semakin besar resikonya tidak tertagih, demikian pula sebaliknya.
e. Balas Jasa, dimana dalam bentuk bunga, biaya provisi, dan komisi serta biaya administrasi kredit ini merupakan keuntungan Lembaga Keuangan.

2.4 Analisis Kredit, Resiko Kredit dan Credit Scoring

Analisis kredit merupakan proses analisis atau menilai permohonan kredit yang diajukan dengan menggunakan pendekatan-pendekatan dan rasio-rasio keuangan sehingga dapat memberikan keyakinan kepada pihak penyedia kredit

bahwa proyek kredit yang diajukan cukup layak untuk disetujui. Adapun tujuan dari analisis kredit adalah mengurangi resiko kredit yang ada.

Purwanto (2011) menjelaskan resiko kredit sebagai resiko terjadinya kerugian akibat kegagalan pembayaran oleh debitur atau terjadinya kemerosotan kualitas kemampuan membayar pihak debitur. Resiko kredit menjadi kritis apabila nasabah dengan pemberian dana yang besar tidak dapat membayar pinjamannya sehingga menimbulkan kerugian dan kesulitan dalam likuiditas.

Sabato (2010) menulis bahwa sejak tahun 1960organisasi besartelahmemanfaatkan*credit scoring*dengan cepat danmenilai tingkatrisiko dari prospekpelamardankonsumen yang sudah adaterutamadalam bisniskredit konsumen.*Credit scoring* memprediksi kemungkinan bahwa calon konsumen akan sesuai dengan standar atau menunggak selama jangka waktu yang tetap dengan tujuan untuk menentukan pinjaman kredit. Sabato (2010) mendeskripsikan bahwa model ini memberikan peringkat risiko calon konsumen berdasarkan asumsi jika konsumen yang ada memiliki perilaku tertentu, ada kemungkinan calon konsumen dengan karakteristik serupa menunjukkan perilaku yang sama.