

BAB 2

LANDASAN TEORI

2.1 Teori –teori Dasar/Umum

2.1.1 Pengertian Data

Menurut (<http://id.wikipedia.org>), data merupakan bentuk jamak dari **datum**, berasal dari bahasa Latin yang berarti "sesuatu yang diberikan". Dalam penggunaan sehari-hari data berarti suatu pernyataan yang diterima secara apa adanya. Pernyataan ini adalah hasil pengukuran atau pengamatan suatu variabel yang bentuknya dapat berupa angka, kata-kata, atau citra.

Data adalah suatu gambaran yang belum jelas / abstrak bagi penggunanya. Selanjutnya, data yang ada harus diolah agar menjadi suatu informasi yang berguna bagi pengguna.

2.1.2 Pengertian Database

Menurut Connolly and Begg (2005, p 15), *database* adalah suatu kumpulan dari data yang berhubungan secara logikal dan deskripsi dari suatu data yang dirancang sebagai informasi yang dibutuhkan oleh sebuah organisasi. *Database* bersifat tunggal, memiliki tempat penyimpanan data yang besar di mana dapat digunakan secara bersama-sama oleh banyak departemen dan *user*.

Menurut (<http://id.wikipedia.org>), *database* juga diartikan kumpulan informasi yang disimpan di dalam komputer secara

sistematik sehingga dapat diperiksa menggunakan suatu program komputer untuk memperoleh informasi dari basis data tersebut. Secara sederhana *database* digambarkan sebagai potongan-potongan dari satu bagian yang utuh. Setiap potongan yang ada memiliki hubungan yang unik dengan potongan lainnya. Potongan-potongan ini dapat disusun dengan model yang berbeda-beda. Hingga membentuk suatu kesatuan bentuk yang memberikan fakta-fakta dari kumpulan potongan yang ada.

Menurut Date (2000, p10), *database* adalah sekumpulan data yang secara terus menerus dapat digunakan untuk sistem aplikasi suatu perusahaan. Sehingga secara keseluruhan *database* adalah sekumpulan data yang digunakan oleh banyak user dan terhubung secara *logical* untuk menghasilkan suatu sistem informasi yang dibutuhkan oleh perusahaan.

Perangkat lunak yang digunakan untuk mengelola dan memanggil kueri (*query*) basis data disebut sistem manajemen basis data (*database management system, DBMS*).

2.1.3 Definisi Database Management System (DBMS)

Connolly and Begg (2005, p 16), menjelaskan bahwa *DBMS* atau *Database Management System* adalah suatu *system software* yang memungkinkan pengguna untuk mendefinisikan, membuat, memelihara, dan mengatur akses ke *database*.

DBMS memungkinkan user untuk memahami *database*, dengan *Data Definition Language* (DDL). DDL memungkinkan pengguna untuk menspesifikasikan tipe data dan struktur dan konstrain data yang akan dimasukkan ke *database*. DBMS memungkinkan pengguna untuk melakukan proses *insert*, *update*, *delete*, dan *retrieve* data dari *database* menggunakan *Data Manipulating Language* (DML).

2.1.3.1 Kelebihan DBMS

DBMS juga memberikan kontrol akses ke *database*, seperti :

1. Sistem keamanan
2. Sistem integritas
3. Kontrol terhadap *Concurrency*
4. Sistem pemulihan data
5. Katalog yang dapat diakses pengguna

DBMS juga memberikan *database* berbagai kelebihan sebagai berikut :

- a. Kontrol terhadap pengulangan data (*redundancy*)
- b. Konsistensi data
- c. Lebih banyak informasi dari data yang sama
- d. Data dapat dipakai secara bersama-sama
- e. Memperbaiki integritas data
- f. Memperbaiki keamanan data

- g. Adanya standarisasi
- h. Skala ekonomi
- i. Mengatasi masalah kebutuhan
- j. Memperbaiki akses data dan respons data
- k. Meningkatkan produktivitas
- l. Memperbaiki pemeliharaan data melalui kemandirian data
- m. Meningkatkan *concurrency*
- n. Meningkatkan layanan *backup data* dan *recovery*

2.1.3.2 Kekurangan DBMS

Namun, DBMS juga mempunyai kelemahan sebagai berikut :

- a. Memiliki sistem kompleks yang mengakibatkan ukuran menjadi lebih besar
- b. DBMS memiliki harga yang bervariasi tergantung fungsi dan kebutuhannya
- c. Penambahan biaya untuk kebutuhan perangkat keras.
- d. Penambahan biaya untuk konversi
- e. DBMS dirancang untuk hal yang lebih umum, sehingga performancinya lebih rendah disbanding dengan aplikasi yang *file-based*
- f. Kegagalan DBMS dapat menyebabkan operasi tidak berjalan

2.1.4 *Data Warehouse*

Menurut Connolly and Begg (2005 , p 1151-1152), *data warehouse* adalah kumpulan data yang bersifat *subject oriented, integrated, time-variant, dan non volatile* yang mendukung proses pengambilan keputusan yang mendukung proses pengambilan keputusan bagi pihak manajemen.

Data mining menjadi mungkin karena adanya kuantitas data yang besar. Data-data ini haruslah disimpan dalam bentuk yang jelas dan terstruktur. *Data warehousing* adalah gudang fakta dan data-data yang berhubungan yang mudah diakses dan teratur untuk suatu keputusan bisnis yang lebih baik. *Data warehouse* menyediakan akses ke informasi tentang bisnis perusahaan, produk, dan konsumen.

2.1.4.1 *Manfaat Data Warehouse*

Menurut Connolly and Begg (2005, p 1152), Manfaat dari *data warehouse* itu sendiri juga banyak, seperti berikut ini :

- Kemampuan untuk mengakses data yang berskala perusahaan.
- Kemampuan memiliki data yang konsisten.
- Dapat melakukan analisis secara cepat.

- Dapat digunakan untuk mencari redundansi usaha di dalam perusahaan.
- Menemukan *gap* antara pengetahuan bisnis dan proses bisnis.
- Meminimalkan biaya administrasi.
- Meningkatkan kinerja pegawai perusahaan untuk dapat melakukan tugasnya dengan lebih efektif.

2.1.4.2 Karakteristik *Data Warehouse*

Menurut Connolly and Begg (2005 , p 1151), *data warehouse* memiliki karakteristik sebagai berikut :

- *Subject Oriented* : *data warehouse* berorientasi pada subjek, yaitu subjek-subjek utama yang berada dalam lingkungan bisnis dalam perusahaan.
- *Integrated* : *data warehouse* harus terintegrasi karena sumber-sumber *data warehouse* berasal dari lingkungan bisnis yang banyak dengan sistem aplikasi yang berbeda. Sumber data yang terintegrasi harus dibuat konsisten sebagai tampilan data kepada user.
- *Time Variant* : *data warehouse* hanya valid dan akurat pada poin-poin tertentu atau dalam interval waktu tertentu.

- *Non-Volatile* : data pada data warehouse tidak di-update dalam waktu yang sebenarnya (*real-time*) , tetapi data di-*refresh* dari sistem operasional. Data akan ditambahkan sebagai suplemen data yang ada , bukan digantikan. *Database data warehouse* akan selalu mengambil data yang baru dan secara berkala diintegrasikan dengan data yang sudah ada.

2.1.4.3 Model Data Warehouse

Menurut Connolly and Begg (2005, p 1183), setiap *data warehouse* memiliki table dengan *composite key* yang disebut dengan *fact table* dan kumpulan table-table kecil yang disebut dengan *dimension table*.

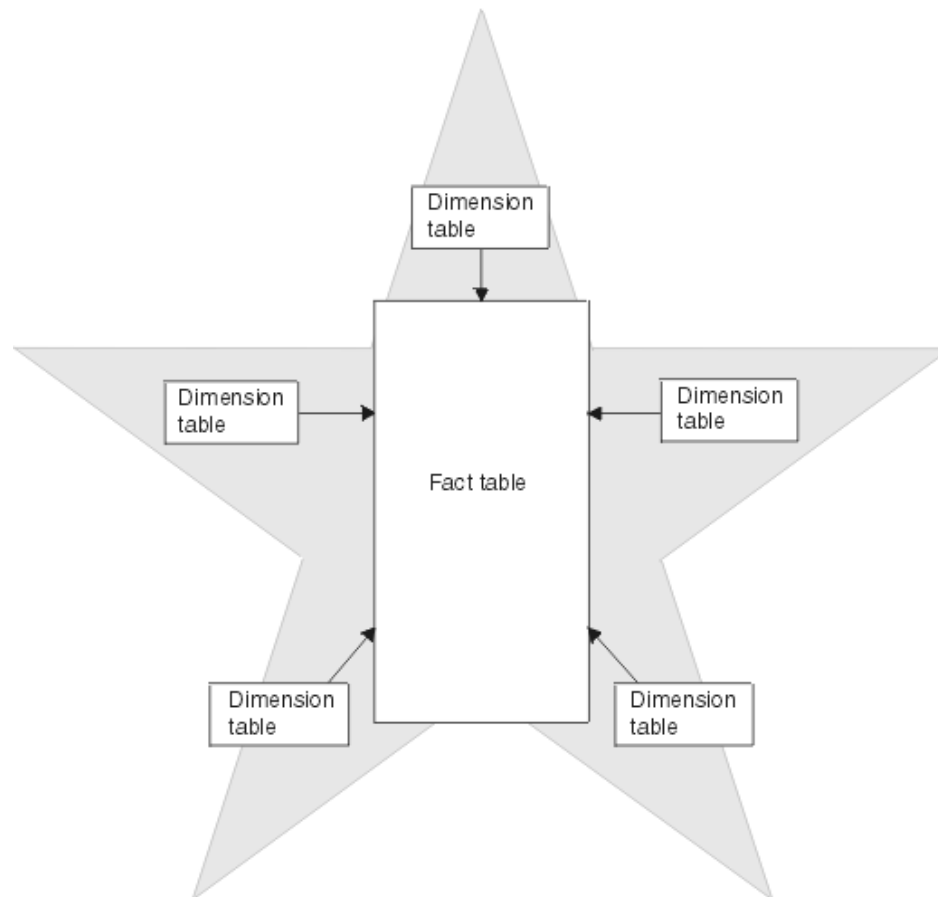
Menurut Connolly dan Begg (2005, p 1182), *Dimensionality modeling* adalah sebuah teknik desain logical yang memiliki tujuan menampilkan data dalam bentuk standar dan intuiti, yang memungkinkan akses ke performansi yang tinggi .

Beberapa contoh *dimensionality modeling* pada desain *data warehouse*, yaitu :

a. *Star Schema*

Star schema adalah struktur *logical* yang memiliki sebuah table fakta (*fact table*) yang berisi data faktual

yang diletakkan di tengah (pusat), dikelilingi oleh tabel-tabel dimensi yang berisi data referensi data yang dapat dinormalisasikan.

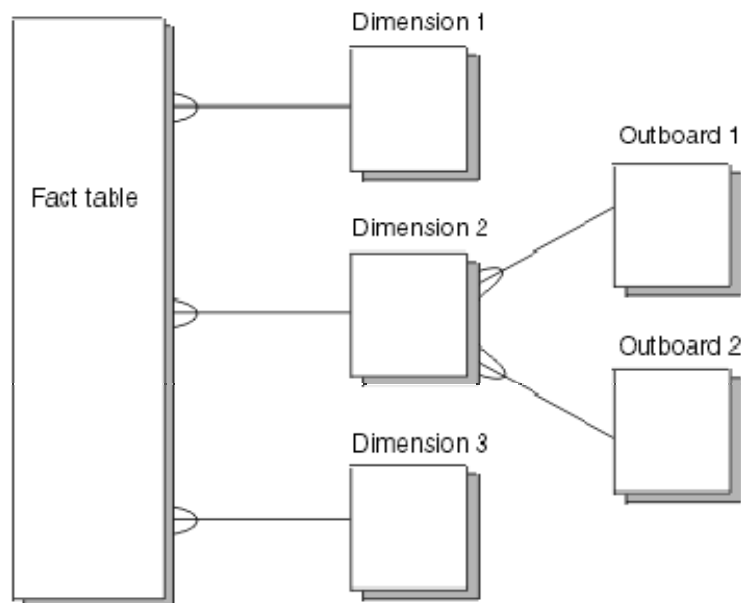


Sumber (<http://publib.boulder.ibm.com>)

Gambar 2.1 Model *Star Schema*

b. *Snowflake schema*

Snowflake schema adalah variasi dari *star schema*, namun table dimensi pada *schema* ini tidak mengandung denormalisasi yang memungkinkan sebuah dimensi untuk mempunyai dimensi yang lainnya. Suatu *schema* disebut *snowflake schema* jika satu atau lebih table dimensi tidak berhubungan langsung dengan *fact table* melainkan pada *dimension table*.



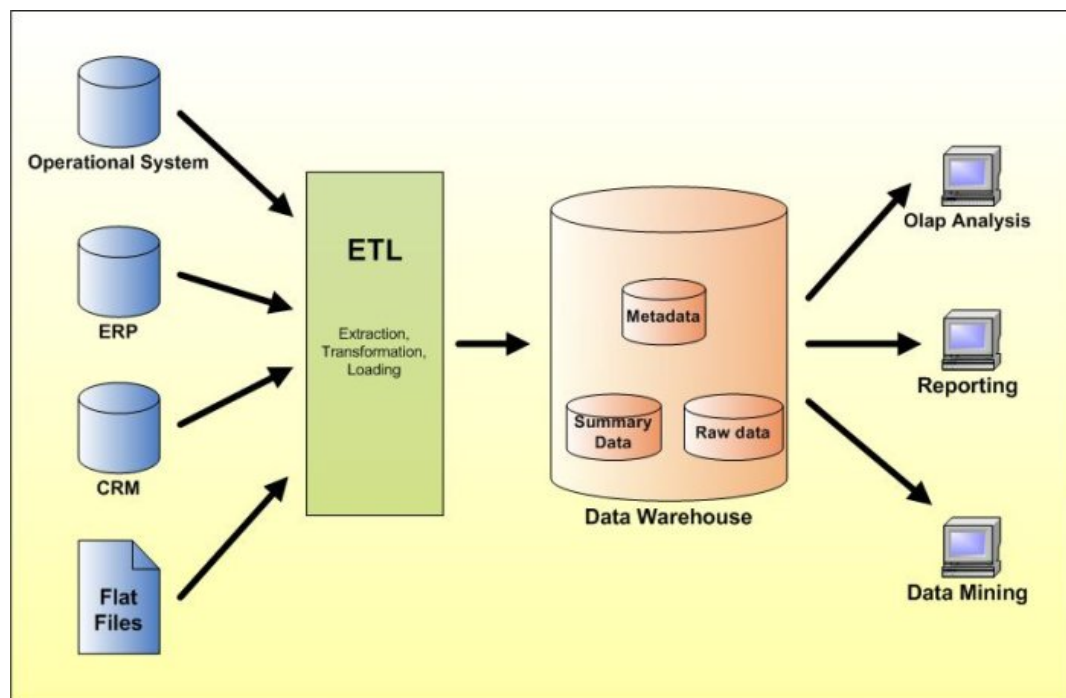
Sumber (<http://publib.boulder.ibm.com>)

Gambar 2.2 Model *Snow Flake Schema*

2.1.4.4 Arsitektur *Data Warehouse*

Menurut Connoly and Begg (2005, p 1056-p1161), komponen *data warehouse* yang terdapat pada arsitektur *data warehouse*, adalah sebagai berikut :

- *Operational data*
- *Operational Data Store (ODS)*
- *Load manager*
- *Warehouse Manager*
- *Query manager*
- *Detailed Data*
- *Lightly and highly summarized data*
- *Archive/ backup data*
- *Meta Data*



Sumber (<http://thepcweb.com/wp-content/uploads/2008/06/data-warehouse-architecture.jpg>)

Gambar 2.3 Bagan Arsitektur *Data Warehouse*

2.1.5 *Data Marts*

Menurut David Olson and Yong Shi (2007 , p 36), dalam membuat *data mining* , bentuk penyimpanan kelas menengah sudah dibutuhkan. *Data marts* kadang digunakan untuk meng-ekstrak item yang spesifik dari informasi untuk analisis pembuatan *data mining*.

Data marts biasa digunakan sebagai gudang dari data yang telah dikumpulkan yang akan memberikan informasi kepada beberapa *user*, menyediakan data yang diekstrak dari *data warehouse*, *data mart* juga memberikan fasilitas kepada *data-miners* untuk *men-transform* informasi menjadi suatu variabel yang baru (rasio,coded data,dll).

Data marts berbeda dengan data warehouse dalam hal kepemilikan. *Data warehouse* dimiliki oleh organisasi, dan sepenuhnya merupakan untuk perusahaan, sedangkan *Data marts* di pihak lain, dimiliki oleh suatu kelompok yang menggunakannya. Isi dari *Data marts* bergantung dari kebutuhan dari kelompok pengguna tersebut.

2.1.6 *OLAP (Online Analytic Processing)*

Menurut Connolly and Begg (2005 , p 1205), OLAP adalah sebuah perangkat yang menggambarkan teknologi gambaran

multidimensi sejumlah data untuk menyediakan akses yang cepat bagi strategi informasi dengan tujuan mempercepat analisis.

Menurut David Olson and Yong Shi (2007 ,p 37), OLAP adalah suatu pendekatan *multidimensional-spreadsheet* sebagai tempat penyimpanan data yang dirancang agar *user* dapat mengambil data dan menghasilkan report pada dimensi yang sesuai dengan bagian *user*.

2.1.7 *Data Mining*

2.1.7.1 **Pengertian Data Mining**

Menurut Connolly and Begg (2005 , p 1233), *data mining* adalah suatu proses ekstraksi atau penggalian data dan informasi dalam jumlah besar, yang belum diketahui sebelumnya, namun dapat dipahami dan berguna dari suatu *database* yang besar serta digunakan untuk membuat keputusan bisnis yang sangat penting.

Menurut David Olson and Yong Shi (2007 , p 5), *data mining* mengacu pada analisis dari data yang berukuran besar menggunakan sistem automata untuk mengetahui pola dan aturan tertentu.

Menurut Berson and Smith (2000, p 6), *data mining* , dengan mudah, otomatisasi pendeteksian dari pola yang berhubungan pada sebuah *database*. Bila dibayangkan ketika gunungan data telah berkumpul di dalam *datawarehouse*, maka *data mining* adalah sebuah alat untuk menjelajahi data-

data tersebut dan menemukan sebuah informasi berguna didalamnya.

2.1.7.2 Fungsi *Data Mining*

Data mining biasa digunakan sebagai sarana untuk mengetahui fakta-fakta yang ada dalam suatu database dengan cara mendeteksi dengan pola-pola yang ada. David Olson and Yong Shi (2007 , p 56), menjelaskan bahwa *data mining* memiliki beberapa fungsi seperti berikut :

1. **Klasifikasi** : menggunakan suatu data pembelajaran untuk mengidentifikasi kelas ataupun kluster. Contoh : Membuat suatu aplikasi tertentu yg dapat menjelaskan kategori resiko, atau resiko kredit bagi nasabah.
2. **Prediksi** : mengidentifikasi sebuah atribut kunci dari data untuk membangun suatu formula untuk prediksi ke depan, seperti model regresi.
3. **Asosiasi** : mengidentifikasi aturan yang menentukan hubungan dari setiap entity, seperti pada analisa pasar, atau gejala dengan penyakitnya.
4. **Deteksi** : menentukan anomali dan ketidakteraturan.
Cont: besarnya penemuan kasus penipuan.

2.1.7.3 *Cross industry Standard Process for Data Mining (CRISP – DM)*

Menurut David Olson and Yong Shi (2007 , p 20), saat ini CRISP-DM adalah merupakan standar metodologi *data mining* untuk industri. Metodologi ini membuat *data mining* yang besar dapat dilakukan dengan lebih cepat, lebih ekonomis, dan mudah untuk diatur. Bahkan, data mining yang berukuran kecil pun dapat memperoleh keuntungan dari CRISP-DM.

Berikut adalah enam fase yang disebut sebagai siklus :

1. Business understanding

Business understanding berkaitan dengan menentukan objektif atau tujuan dari bisnis yang dijalankan, melihat keadaan yang berjalan, menyusun suatu 'goal' dari *data mining* yang dibuat, dan yang terakhir adalah membuat suatu rencana kerja.

2. Data understanding

Ketika tujuan bisnis dan rencana kerja telah dibuat , maka perlu dilakukan *data understanding* berkaitan dengan kebutuhan dari data. Pada tahap ini termaksud didalamnya pengumpulan data awal, deskripsi data, eksplorasi data ,serta verifikasi dari kualitas data.

3. *Data preparation*

Saat sumber data telah terkumpul, tersedia untuk diidentifikasi, mereka harus dipilih, dibersihkan, dibentuk sesuai yang dikehendaki, lalu diformat. *Data*

cleansing dan transformasi data pada persiapan data berlangsung pada tahap ini. Eksplorasi data yang lebih dalam dapat dilakukan pada tahap ini, juga memberi kesempatan untuk melihat pola-pola yang berdasar pada bisnis.

4. *Modeling*

Program tools untuk *data mining* seperti visualisasi, dan analisis kluster, sangatlah berguna untuk analisis awal. Pada tahap modeling dilakukan Pemilihan dan penerapan berbagai variasi dari teknik *modelling* dan membuat standar parameter dari *modelling tool*. Jika data persiapan belum siap digunakan maka kita harus kembali ke tahap *data preparation*.

5. *Evaluation*

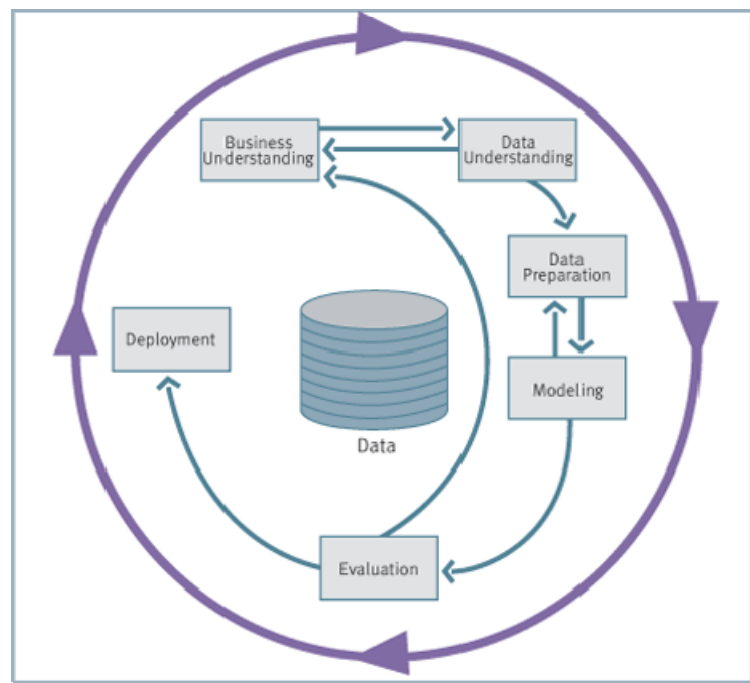
Hasil dari *modelling* harus kembali dievaluasi sesuai dengan konteks tujuan bisnis yang telah dibuat pada fase pertama. Tahap ini akan membawa kepada identifikasi dari kebutuhan lainnya (biasanya menuju kepada pengenalan pada pola). Pada akhir dari tahap ini, keputusan penggunaan hasil *data mining* telah ditentukan.

6. *Deployment*

Data mining dapat digunakan untuk memverifikasi hipotesis yang telah dibuat sebelumnya, ataupun

digunakan untuk menjelajah *knowledge* yang ada pada database. Pembuatan dari model bukanlah akhir dari proyek. Meskipun tujuan dari pemodelan adalah untuk meningkatkan *knowledge* dari data, *knowledge* data tersebut perlu dibangun dengan terorganisasi dan dibuat pada satu bentuk yang dapat digunakan oleh user. Bergantung pada kebutuhan, *deployment* dapat dilakukan semudah membuat laporan. Pada banyak kasus *deployment* dilakukan oleh user bukan dari data *analyst*.

Gambaran CRISP-DM *process* :



Sumber (<http://www.crisp-dm.org/Process/index.htm>)

Gambar 2.4 Siklus CRISP-DM

2.1.7.4 Data Mining Model

David Olson and Yong Shi (2007 , p 54) menjelaskan *data mining* memiliki banyak variasi model pengembangan sesuai dengan tujuannya masing-masing. Berbagai metode yang ada didasari oleh dua mata pelajaran besar yaitu statistik dan *artificial intelligence*. Tehnik yang didasari statistik biasanya merupakan *tool* yang memiliki diagnosa yang kuat, yang dapat digunakan untuk perkiraan parameter, testing hipotesis, dan lainnya. Tehnik yang berdasar pada *artificial intelligence* membutuhkan asumsi data yang lebih sedikit dan pada umumnya berjalan secara otomatis.

2.1.7.5 Model Data Mining Klasik

2.1.7.5.1 Statistik

Menurut Berson and Smith (2000, p 125-127) , berdasarkan pengertian yang seksama, statistik bukanlah data mining. Statistik telah digunakan lama sebelum bentuk data mining. Bagaimanapun, teknik statistik dikendalikan oleh data dan dapat digunakan untuk menemukan pola-pola yang terdapat pada data.

Statistik adalah cabang ilmu matematika yang mempelajari tentang sekumpulan data beserta deskripsinya. Semakin banyak dan semakin baik data

yang dikumpulkan maka pengertian yang diperoleh akan semakin baik.

2.1.7.5.2 *Nearest Neighbour*

Menurut Berson and Smith (2000, p 134-135), *Clustering* dan *nearest neighbour* adalah teknik prediksi data mining yang terbilang cukup tua. *Nearest neighbour* adalah teknik prediksi yang bentuknya hampir sama dengan *clustering*. Teknik ini digunakan untuk memprediksi nilai dalam suatu *record*, mencari *records* yang memiliki kesamaan nilai *predictor* di dalam basis data historis dan menggunakan nilai prediksi dari *record* yang ‘terdekat’ dengan *record* yang belum diklarifikasi.

Prediksi *Nearest neighbour* secara mudah diterangkan dengan pernyataan berikut :

Objek yang berdekatan akan memiliki prediksi yang hampir sama. Oleh karena itu, jika diketahui nilai prediksi dari salah satu objek yang ada, maka kita dapat memprediksi nilai objek yang terdekat.

2.1.7.5.3 *Clustering*

Menurut Berson and Smith (2000, p 139), *Clustering* adalah metode dimana record yang ada pada

database dikelompokkan bersama. Biasanya teknik ini dipakai untuk memberikan kepada user *high level view* terhadap apa yang ada dalam database. *Clustering* biasa dilakukan untuk membentuk segmentasi, dimana bagian pemasaran sangat memerlukannya.

Clustering membangun seluruh record yang terdapat didalamnya, yang memiliki nilai yang sama dari predictor tertentu yang dikelompokkan. Membentuk *cluster* yang homogen dengan nilai *predictor* yang sama sulit untuk dilakukan, jika terdapat banyak *predictor* atau *predictor* yang lain memiliki nilai yang berbeda.

Batasan dalam pembuatan *cluster* adalah jumlah *cluster* haruslah secara logis dibentuk. Logis atau tidak logis ditetapkan oleh pengguna. Namun sulit untuk mengukur *cluster* yang tidak dapat diterima dan banyak *cluster* dan *record* asli yang juga tidak diterima. Banyak algoritma *clustering* memperbolehkan pengguna untuk memilih jumlah kelompok yang ingin dibentuk dari database atau algoritma memperbolehkan pengguna membentuk sejumlah kelompok secara interaktif setelah *clustering* terbentuk.

Ketika *clustering* digunakan di dunia bisnis, *clustering* dapat terlihat lebih dinamis. Bahkan dapat berubah tiap minggu ataupun setiap bulannya, dan lebih banyak lagi keputusan yang terkait dengan *cluster* apakah yang akan dipilih, yang tentunya pemilihannya akan lebih sulit.

2.1.7.6 Model Data Mining Generasi Berikutnya

2.1.7.6.1 Decision Tree

Menurut Berson and Smith (2000, p 156), *Decision tree* adalah sebuah *predictive model* yang dapat dilihat seperti pohon. Masing-masing cabang dari tree mengklasifikasikan pertanyaan. Daun dari *tree* merupakan bagian dari set data dengan klasifikasinya. *Decision Tree* dalam kaitannya dalam dunia bisnis, membuat segmentasi bisnis dari dataset yang ada. Segmentasi ini dilakukan dengan tujuan untuk mendapatkan data dengan *high level view*, dengan tidak adanya ketentuan tertentu untuk membentuk segmentasi tersebut, kecuali keterkaitan antar *record* yang ada. Walaupun *decision tree* dan algoritmanya sangat kompleks, hasil dari algoritma ini menjadi mudah dimengerti dan sangat berguna bagi user yang menggunakannya. Algoritma untuk merancang

decision tree bervariasi, dua yang paling sering digunakan adalah CART (*Classification and Regression Trees*) dan CHAID (*Chi-square Automatic Interaction Detector*). CART adalah suatu algoritma prediksi dan eksplorasi data sedangkan CHAID adalah suatu algoritma yang menggunakan *chi-square* test untuk menentukan *predictor categorical* mana yang jauh lebih independen dari nilai yang diprediksi.

Hal yg menarik dari Decision Tree

- Decision Tree memiliki konsep yang mudah dimengerti.
- Decision Tree membagi data pada setiap leaf tanpa kehilangan data sedikitpun.
- Decision Tree cocok untuk digunakan dalam bidang bisnis ataupun pemasaran.

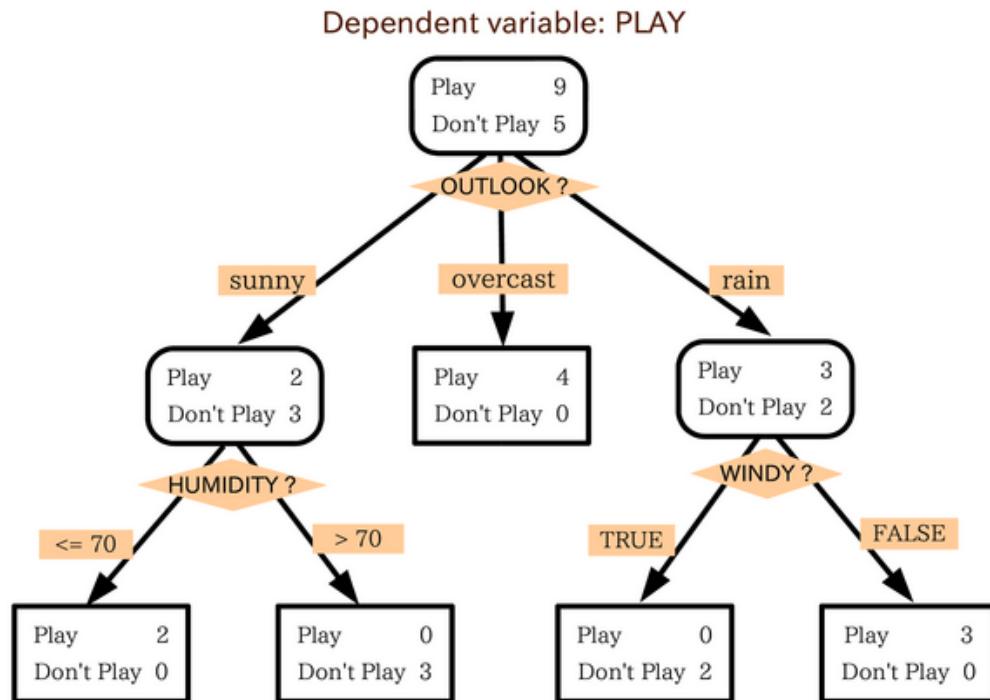
Decision Tree dan Rule induction

Decision tree sangatlah mirip dengan *rule induction*, kecuali dengan *rule* yang dibentuk pada *rule induction* tidak membuat partisi-partisi pada *database* yang bersifat saling eksklusif. Tidak ada *record* dari

database yang akan diklasifikasi lebih dari satu rule pada algoritma *decision tree*.

Decision tree membuat set rule yang paling efisien dan kemungkinan terkecil yang membuatnya menjadi *predictive model* yang baik. Jika terdapat *overlap* diantara dua prediktor maka yang terbaik dari keduanya yang akan diambil. Pada sistem *rule induction*, keduanya akan diambil dan pada faktanya, salah satunya akan menjadi lebih lemah atau kurang akurat.

Rule induction beroperasi 'bottom up' dan mengumpulkan semua pola yang menarik, lalu menggunakannya sebagai *predictive target*. Di pihak lain, *decision tree*, bekerja dari *prediction target* sebagai *root*, lalu terus sampai *leaf* terbawah.



Sumber (<http://gautam.lis.illinois.edu>)

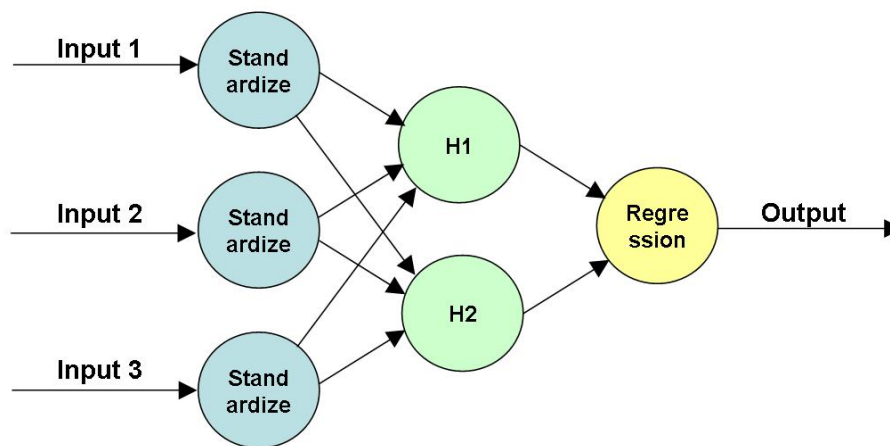
Gambar 2.5 Contoh *Decision Tree*

2.1.7.6.2 *Neural Network*

Menurut Berson and Smith (2000, p 166), *Neural Network* merupakan teknik *data mining* yang paling umum. Jika membicarakan *data mining* maka *neural network* dan *decision trees* merupakan yang paling populer sekarang ini. Teknik ini sangat menarik karena *neural network* menemukan pola dari data secara analogi, sama halnya seperti manusia.

Neural network memiliki kelebihan yaitu keakuratannya yang tinggi dan dapat digunakan pada berbagai jenis masalah. Disamping itu, juga terdapat

kelemahannya, model yang dibentuk oleh *neural network* sulit untuk dimengerti, bahkan bagi pakarnya, dan biasanya algoritmanya sensitif dengan format data yang digunakan, kebanyakan output yang diberikan dalam bentuk numerik, sehingga perlu didefinisikan lagi ke *value* yang aktual.



Sumber (<http://www.data-miners.com>)

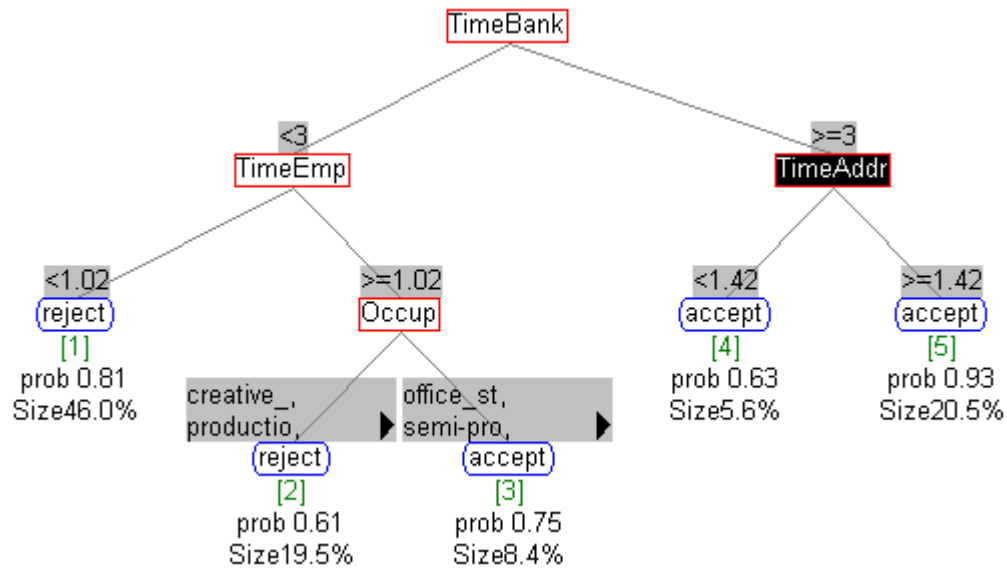
Gambar 2.6 Contoh *Neural Network*

2.1.7.6.3 Rule Induction

Menurut Berson and Smith (2000, p 183), *Rule induction* adalah salah satu bentuk utama dari *data mining*. Teknik ini juga merupakan bentuk *data mining* yang paling mendekati proses berpikir sebagian besar orang yang berpikir mengenai *data mining*. *Data*

mining menggali emas dalam database yang besar. Emas dalam hal ini merupakan sebuah rule yang menarik, yang memberikan informasi tentang sesuatu yang ada pada *database*, namun sebelumnya tidak diketahui atau tidak jelas dikemukakan.

Rule induction di dalam *database* dapat menjadi pekerjaan besar, dimana semua pola diambil dari data secara sistematis. Kemudian, akurasi dan kepentingan ditambahkan agar pengguna mengetahui seberapa kuat pola tersebut dan seberapa sering pola tersebut muncul kembali. Rules yang diambil dari database digali untuk disajikan kepada pengguna, berdasarkan atas persentase kebenarannya dan seberapa sering digunakan.



Sumber (<http://www.xpertrule.com>)

Gambar 2.7 Contoh *Rule Induction*

2.1.7.7 Penerapan Data Mining

Menurut Berson and Smith (2000, p71), Perusahaan-perusahaan biasa menggunakan *data mining* sebagai :

a. *Discover Knowledge*

Bertujuan menentukan dengan jelas relasi, pola, atau korelasi yang tersembunyi dari berbagai tempat penyimpanan di dalam database. Secara spesifik, data mining dapat digunakan untuk menunjukkan :

- Segmentasi : data kelompok pelanggan untuk penyesuaian pemasaran.

- Klasifikasi : Proses input data untuk kelas yang telah ditentukan, mencari dan memahami tren, klasifikasi dokumen teks.
- Asosiasi : Pencarian dari kesempatan pemasaran silang.
- Preferensi: Menentukan preferensi dari mayoritas pelanggan.

b. *Visualize Data*

Seorang analis harus dapat menemukan sebuah informasi yang besar yang disimpan di dalam *database*. Tujuannya untuk ‘mempermanusiakan’ data yang banyak dan menemukan cara yang terbaik untuk menampilkan data.

c. *Correct Data*

Pada saat proses menggabungkan *database* secara besar-besaran, banyak perusahaan menemukan data yang digabungkan tersebut tidak lengkap, dan terdiri dari informasi yang salah dan bertentangan. Dengan menggunakan teknik data mining dapat membantu untuk mengidentifikasi dan membetulkan kesalahan dengan cara yang konsisten. Jumlah dari aplikasi dalam kategori ini terbatas dikarenakan adanya kesulitan alami dari proses koreksi. Mengganti *missing values* atau memperbaiki nilai yang dianggap tidak valid membutuhkan pertimbangan yang sulit untuk disediakan secara otomatis.

2.2 Teori-Teori Khusus yang Berhubungan dengan Topik Yang Dibahas

2.2.1 *Business Data Mining*

Data mining telah sangat efektif pada berbagai bidang bisnis. Kuncinya adalah untuk menemukan informasi yang dapat ditindaklanjuti, atau informasi yang dapat dimanfaatkan pada jalur yang sesuai untuk meningkatkan keuntungan perusahaan. Berbagai jenis bisnis yang menjadi area *data mining* adalah bisnis retail, perbankan, asuransi, telekomunikasi, dan lain-lain.

2.2.1.1 *Data Mining Pemasaran*

Menurut Michael and Gordon (2004 , p 6-7), Faktanya, *data mining* ditemukan dengan algoritma yang berhubungan dengan sesuatu yang komersil. *Data miner* yang komersial mengambil teori-teori yang berasal dari statistik, ilmu komputer, dan juga bahasa mesin. Untuk membangun sebuah data mining yang ditujukan untuk membantu pemasaran, data mining haruslah disandingkan dengan strategi CRM yang cocok yang akhirnya memberikan output berupa action yang harus dilakukan oleh pihak user. Data mining merupakan tool, seperti halnya *tool* lainnya, *data mining* tidak cukup hanya dimengerti namun juga harus dipahami bagaimana data mining dapat bekerja bersama dengan CRM yang ada.

2.2.1.2 Data Mining Perbankan

Menurut David Olson and Yong Shi (2007 , p 9), Industri perbankan merupakan salah satu dari pengguna pertama dari data mining. Bank telah beralih ke teknologi untuk menemukan motif dari nasabah mereka, dan apa saja yang akan menjaga bisnis perbankan tetap berjalan (*Customer Relationship Management - CRM*). CRM menyangkut sebuah aplikasi untuk memonitor pelayanan terhadap user atau pelanggan, sebuah fungsi yang dapat ditingkatkan performanya dengan menggunakan *data mining*.

Data mining memberikan sebuah jalan bagi perbankan untuk mengenal pola dari nasabah. Ini sangat berguna pada bagian peminjaman, juga tentunya pada bagian pemasaran.

2.2.2 Data Mining Tool Yang Digunakan

Di dalam Oracle ada sebuah tool yang diperuntukkan bagi data miner, yaitu Oracle *Data Miner* (ODM). Oracle data miner merupakan *tool* yang dapat memining data dari *Oracle Database*.

Oracle *Data Miner* memiliki fungsi-fungsi dasar yang ada pada *data mining tools*, seperti :

- a. *Data Source*, disini ODM akan mengambil data yang akan dimining oleh *data miner* yang berasal dari database oracle ataupun dari *source* lain, yang memiliki tipe data teks (*.txt) atau *Excel binary Workbook* (*.csv) yang sebelumnya harus dikonversi ke bentuk *database* Oracle.
- b. *Activity Builder*, pada tahap ini ODM memberikan fasilitas pembentukan model yang terdiri dari 3 tahap, yaitu *build model*, *test model*, dan *apply model*.
 - *Build Model* : merupakan proses pembuatan model, seperti pemilihan algoritma data mining yang akan digunakan, pemilihan data source yang akan di *mining*. *Result* yang diberikan ODM adalah bentuk model data sebagai bentuk pola dari data yang akan dimining. Model disajikan dalam bentuk teks dan grafik.

Pemilihan algoritma *classification* yang diberikan Oracle *Data Mining*, terdiri dari :

1. *Decision Tree Algorithm*

2. *Naïve Bayes Algorithm*

3. *Adaptive Bayes Network Algorithm*

4. *Support Vector Machine (SVM) Algorithm*

- *Test Model* : pada tahap ini, model data yang sudah terbentuk akan dicek tingkat akurasi. Tujuan tahap ini agar data

miner dapat memperoleh model yang terbaik untuk diproses ke tahap selanjutnya.

- *Apply Model* : model yang sudah dibentuk lalu diaplikasikan kepada data sehingga didapat hasil sesuai dengan pencarian fakta data.