BAB II

LANDASAN TEORI

2.1. Data, Informasi dan *Knowledge*

2.1.1. Data

Data merupakan hal yang memegang peranan penting sebagai penghubung antara user dengan mesin. Data bisa berupa angka, karakter, simbol, gambar, tanda-tanda, isyarat, tulisan, suara, ataupun bunyi yang merepresentasikan keadaan sebenarnya. Sebagai penguat, definisi data menurut Connolly & Begg (2010) adalah komponen yang paling penting dalam *Data mining Management System (DBMS),* berasal dari sudut pandang *end-user*. Sedangkan menurut Turban (2010), data adalah komponen deskripsi dasar dari hal, kejadian, kegiatan dan transaksi yang direkam, diklarifikasikan dan disimpan tetapi tidak diorganisasikan untuk menyampaikan makna tertentu.

2.1.2. Informasi

Informasi adalah data yang penting yang memberikan pengetahuan yang berguna. Informasi biasanya merupakan hasil dari suatu pengolahan data seperti hasil analisa, hasil gabungan ataupun hasil penyimpulan. Suatu informasi dinilai berguna atau tidak tergantung pada beberapa aspek, yaitu:

a. Tujuan si penerima

   Apabila informasi itu tujuannya untuk memberikan bantuan maka informasi itu harus membantu si penerima dalam usahanya untuk mendapatkannya.

b. Ketelitian penyampaian dan pengolahan data

Penyampaian dan mengolah data, inti, dan pentingnya info harus dipertahankan.

c. Waktu

Informasi yang disajikan harus sesuai dengan perkembangan informasi itu sendiri.

d. Ruang dan tempat

Informasi yang didapat harus tersedia dalam ruangan atau tempat yang tepat agar penggunaannya lebih terarah bagi si pemakai.

e. Bentuk

Dalam hubungannya bentuk informasi harus disadari oleh penggunaannya secara efektif, hubungan-hubungan yang diperlukan, kecenderungan-kecenderungan dan bidang-bidang yang memerlukan perhatian manajemen serta menekankan informasi tersebut ke situasi-situasi yang ada hubungannya.

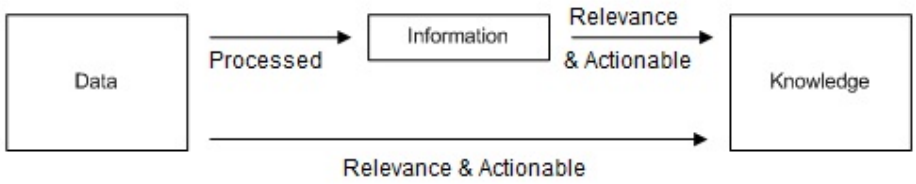
f. Semantik

Agar informasi efektif informasi harus ada hubungannya antara kata-kata dan arti yang cukup jelas dan menghindari kemungkinan salah tafsir.

2.1.3. *Knowledge*

Knowledge merupakan keseluruhan bagan dari data informasi yang orang bawa untuk digunakan pada kegunaan praktis dalam tindakan, supaya dapat membawa tugas kita dan menciptakan informasi baru.

Menurut Turban (2010), terdapat hubungan antara data, informasi dan *knowledge* yang digambarkan pada bagan berikut:



Gambar 2.1. Data, Informasi dan Knowledge

2.2. Data Warehouse

2.2.1. Pengertian *Data Warehouse*

Data Warehouse menurut (Han, Kamber, & Pei, 2011) adalah suatu penyimpanan data yang memiliki beberapa karakteristik yaitu berorientasi objek, terintegrasi, mempuya variant waktu dan menyimpan data dalam bentuk non-volatile sebagai pendukung dalam proses pengambilan keputusan.

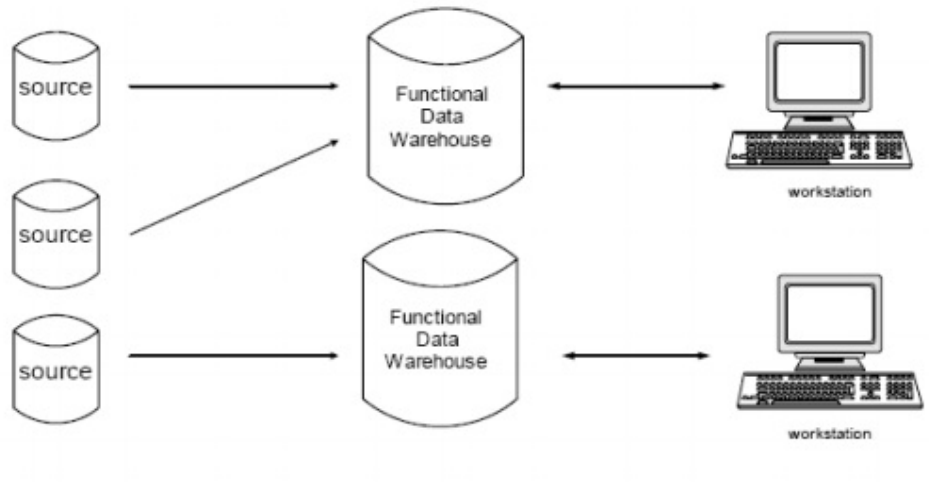
Pembangunan *data warehouse* terdiri dari pembersihan data, penggabungan data dan transformasi data. Data warehouse ini dapat menyatukan data ke dalam bentuk multi dimesi dan dapat dilihat sebagai praproses yang dapat digunakan dalam proses data mining. Data warehouse juga mendukung OLAP (Online Analytical Processing), beberapa metode data mining dapat diintegrasikan dengan operasi OLAP untuk meningkatkan proses mining yang interaktif. Karena itulah data warehouse ini menjadi platform yang penting bagi data analisis dan OLAP untuk mengefektifkan proses data mining.

Terdapat tiga jenis sistem *data warehouse*, yaitu:

1. Data Warehouse Fungsional

   Data Warehouse fungsional ini dibuat lebih dari satu dan dikelompokkan berdasarkan fungsi-fungsi yang ada. Dengan

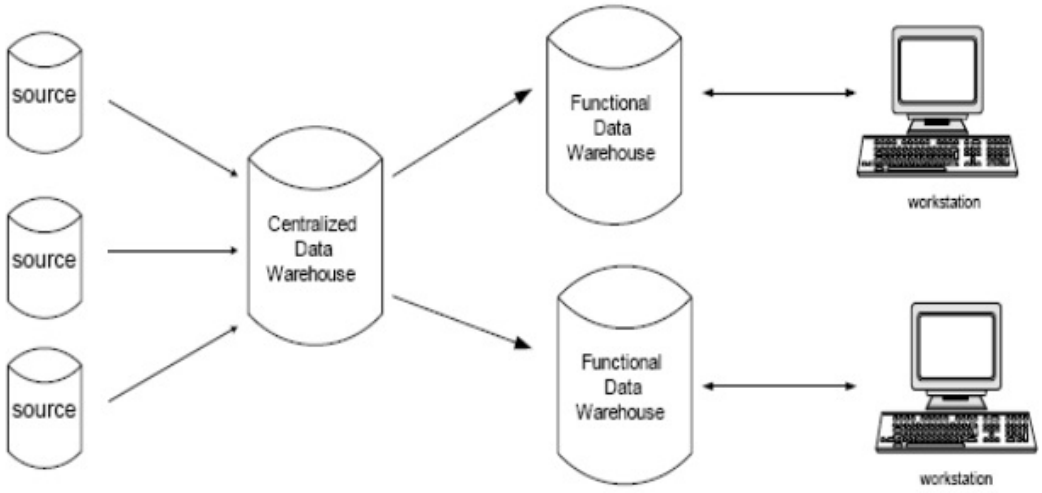
menggunakan data warehouse fungsional ini sistem lebih mudah dibangun dan biaya yang dikeluarkan relatif murah tetapi lebih beresiko untuk kehilangan konsistensi data dan terbatasnya kemampuan dalam pengumpulan data bagi pengguna.



Gambar 2.2. *Data warehouse* Fungsional

2. Data Warehouse Terpusat

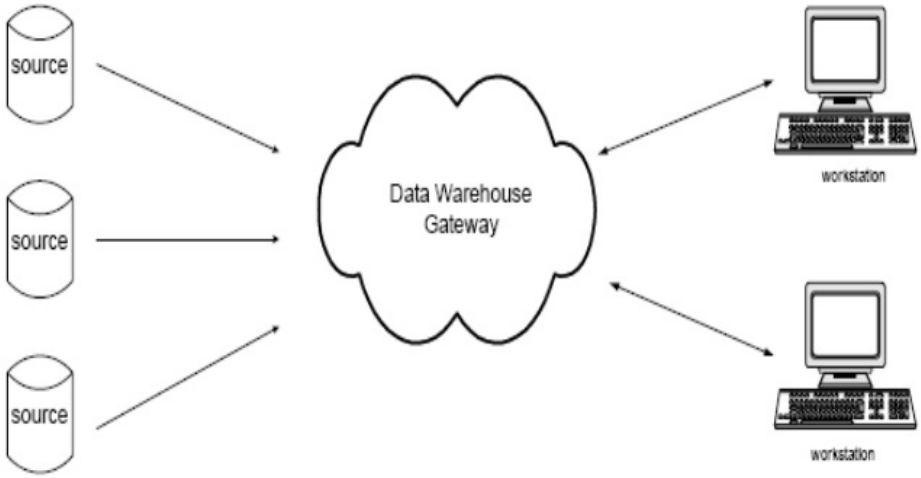
Data warehouse terpusat ini biasanya digunakan oleh perusahaan yang memiliki jaringan eksternal. Data warehouse terpusat ini terlihat seperti bentuk data warehouse fungsional, tetapi sumber sumber data dikumpulkan dalam satu tempat terpusat lalu data tersebut disebar ke fungsinya masing-masing sesuai dengan kebutuhan. Dengan menggunakan data warehouse terpusat ini data yang dihasilkan memiliki konsistensi yang tinggi tetapi untuk membangun data warehouse terpusat ini membutuhkan biaya yang tinggi dan waktu yang lama.



Gambar 2.3. *Data warehouse* Terpusat

3. Data Warehouse Terdistribusi

Data warehouse terdistribusi ini menggunakan gateway yang berfungsi sebagai jembatan penghubung antara data warehouse dengan workstation yang menggunakan sistem yang beraneka ragam. Dengan data warehouse yang terdistribusi ini memungkinkan perusahaan dapat mengakses sumber data yang berasda di luar lokasi perusahaan. Dengan menggunakan data warehouse terdistribusi ini data yang ada tetap konsisten karena data yang digunakan sebelumnya mengalami proses sinkronisasi terlebih dahulu, tetapi data warehouse terdistribusi ini lebih kompleks dalam pembuatannya karena sistem operasi dikelola secara terpisah selain itu biaya yang dikeluarkan lebih besar daripada data warehouse fungsional maupun terpusat.



Gambar 2.4. *Data warehouse* Terdistribusi

2.2.2. ETL (Extract, Tranform, Load)

Proses ETL *(Extract, Tranform, Load)* merupakan proses yang harus dilakukan dalam pembangunan data warehouse (Kimball & Ross, 2002).:

- *Extract*

  Ekstraksi data adalah proses saat data diambil dari berbagai sumber data.

- *Transform*

  Transformasi data adalah proses saat data hasil ekstraksi difilter dan diubah sesuai dengan format yang telah ditentukan.

- *Load*

  Pemuatan data merupakan proses terakhir dari ETL dimana terjadi proses pemuatan data yang telah ditransformasi ke dalam data warehouse.

ETL dapat membaca data dari suatu *data store*, merubah bentuk data, dan menyimpan ke *data store* yang lain. *Data store* yang dibaca ETL disebut *data source*, sedangkan *data store* yang disimpan ETL disebut target. Proses pengubahan data digunakan agar data sesuai dengan format dan kriteria, atau sebagai validasi data

dari *source system*. Proses ETL tidak hanya menyimpan data ke *data warehouse*, tetapi juga digunakan untuk berbagai proses pemindahan data. Kebanyakan ETL mempunya mekanisme untuk membersihkan data dari *source system* sebelum disimpan ke *warehouse*.

2.3. Business Intelligence

2.3.1. Pengertian *Business Intelligence*

Business Intelligence menurut Carlo Vercellis (Vercellis, 2009) adalah kumpulan model matematika dan metodologi analisa yang secara sistematik menghasilkan data untuk menghasilkan suatu informasi dan pengetahuan yang berguna untuk mendukung proses pengambilan keputusan yang kompleks.

2.3.2. Arsitektur *Business Intelligence*

Arsitektur *business intelligence* memiliki enam komponen utama, yaitu:

- *Data source*

  Pada tahap pertama perlu dilakukan pengumpulan dan pengintegrasian data dari berbagai macam sumber yang berbeda beda.

- *Data warehouse and datamarts*

  Pada tahap berikutnya perlu melakukan proses ETL (Extract, Transform and Load) dengan menggunakan tools untuk melakukan extraction dan transformation pada data yang berasal dari berbagai sumber tersebut lalu disimpan kedalan database yang berguna untuk mendukung analisis business intelligence, database inilah yang dikenal dengan sebutan datawarehouse dan datamarts.

- *Data exploration*

  Pada tahap ini , tools-tools yang berguna untuk keperluan analisis business intelligence pasif digunakan. Metodologi ini bersifat pasif karena para pengambil keputusan harus mengambil keputusan berdasarkan hipotesa mereka sendiri kemudian menggunakan tools analisis untuk menemukan jawaban dan mencocokkannya dengan hipotesa awal.

- *Data mining*

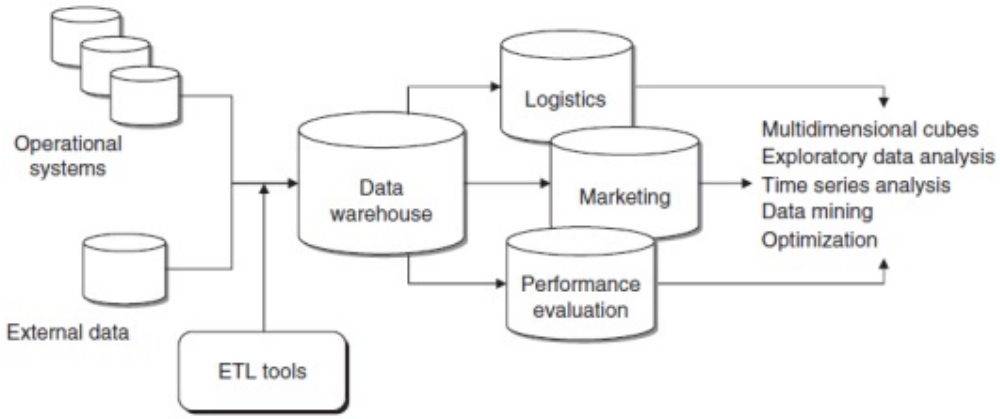
  Pada tahap ini terdiri dari sejumlah metodologi *business intelligence* yang bersifat aktif yang tujuannya adalah untuk mengekstrak informasi dan pengetahuan dari data yang ada. Metodologi ini berisi sejumlah model matematika untuk pengenalan pola, pembelajaran mesin dan teknik data mining. Tidak seperti tools yang digunakan pada tahap sebelumnya, model dari business intelligence yang bersifat aktif ini tidak mengharuskan para pengambilan keputusan untuk mengeluarkan hipotesa apapun.

- *Optimization*

  Pada tahap ini, solusi terbaik harus dipilih dari sekian alternative yang ada yang biasanya sangat banyak dan beragam.

- *Decisions*

  Pada tahap terakhir ini yang menjadi permasalahan utama adalah bagaimana menentukan keputusan akhir yang akan diambil . Walaupun metodologi *business intelligence* berhasil diterapkan, pilihan untuk mengambil sebuah keputusan ada pada para pengambil keputusan. Pertimbangan untuk mengambil keputusan ini biasanya diambil juga dari informasi yang tidak terstruktur seta tidak formal dan memodifikasi rekomendasi serta kesimpulan yang dicapai melalui penggunaan model matematika.



Gambar 2.5. Arsitektur Business Intelligence (Vercellis,2009)

2.4. Data mining

2.4.1. Pengertian *Data mining*

Data mining merupakan suatu kemampuan pencarian data yang canggih dengan menggunakan algoritma statistic untuk menemukan pola dan korelasi yang tersembunyi dalam data. *Data mining* sering juga dikenal dengan sebutan *Knowledge Discovery from Data* (KDD). Tanpa adanya verifikasi, *data mining* tidak dapat digunakan untuk menemukan pola maupun knowledge. *Data mining* dapat

digunakan untuk membantu analisis bisnis dengan mengahsilkan hipotesis, tetapi tidak bisa digunakan untuk memvalidasi hipotesis (Rygielski, Wang, & Yen, 2002).

Menurut Varcellis (2009), *data mining* adalah aktivitas yang menggambarkan sebuah proses analisis yang terjadi secara iterative pada *data mining* yang besar dengan tujuan untuk mengekstrak informasi dan *knowledge* yang akurat dan berpotensial berguna untuk *knowledge workers* yang berhubungan dengan pengambilan keputusan dan pemecahan masalah. Sedangkan definisi *data mining* berdasarkan Han, Kamber, & Pei (2011) adalah proses untuk meneukan *interesting knowledge* dari sejumlah besar data yang disimpan dalam *data mining, data warehouse* atau media penyimpanan yang lainnya. Berikut merupakan karakteristik dari *data mining*:

a. *Data mining* berhubungan dengan penemuan sesuatu yang tersembunyi dengan pola data tertentu yang tidak diketahui sebelumnya.
b. *Data mining* biasanya menggunakan data yang sangat besar. *Data mining* menggunakan data yang besar agar hasilnya dapat lebih dipercaya.
c. *Data mining* berguna untuk membuat keputusan yang strategis, terutama dalam menetukan strategi.

Tahapan-tahapan pada *data mining* (Han, Kamber, & Pei, 2011):

a. Data cleaning

Data cleaning atau pembersihan data adalah suatu proses menghilangkan *noise* dan data yang tidak relevan atau data yang tidak konsisten.

b. Data integration

Data integration atau integrasi data adalah penggabungan data dari berbagai *data mining* ke dalam suatu *data mining* yang baru.

c. Data selection

Data selection atau seleksi data adalah proses penyeleksian data- data pada suatu *data mining* yang akan dianalisis, karena data-data yang ada di dalam *data mining* seringkali tidak semuanya dipakai untuk proses analisis.

d. Data transformation

Data transformation atau transformasi data adalah proses pengubahan atau penggambungan data ke dalam format yang sesuai untuk diproses dalam *data mining*.

e. Application of techniques data mining

Application of techniques data mining atau aplikasi teknik *data mining* merupakan suatu proses utama di mana terdapat suatu metode yang diterapkan untuk menemnukan *knowledge* yang berharga dan tersembunyi dari suatu data.

f. Pattern evaluation

Pattern evaluation atau evaluasi pola adalah proses untuk mengidentifikasi pola-pola yang menarik untuk direpresentasikan ke dalam *knowledge* base yang d*item*ukan. Pada tahap ini, hasil dari teknik *data mining* berupa pola-pola yang unik dan hasil prediksi dievaluasi untuk mengetahui apakah hipotesa yang dilakukan sudah tercapai atau belum.

g. Knowledge presentation

Knowledge presentation atau presentasi pengetahuan merupakan visualisasi dan penyajian *knowledge* mengenai teknik yang digunakan untuk memperoleh pengetahuan yang diperoleh oleh pengguna. Tahap

terakhir dari proses *data mining* adalah bagaimana memformulasikan keputusan atau aksi dari hasil analisa yang didapat.

2.4.2. Teknik-Teknik *Data mining*

Pada umumnya teknik *data mining* terbagi menjadi dua kategori, yaitu:

- **Deskriptif**

  Teknik deskriptif ini digunakan untuk mengenali *pattern (cluster, korelasi, trajectory, trend dan anomaly)* yang merupakan *summary* dari relasi-relasi di dalam data.

- **Prediktif**

  Teknik prediktif ini digunakan untuk memprediksi nilai dari atribut tertentu berdasarkan nilai dari atribut yang lain. Atribut yang akan diprediksi umumnya dikenal sebagai target atau *dependent variable* dan atribut yang digunakan untuk membuat prediksi disebut dengan independent variable

Terdapat empat metode utama pada data mining, yaitu:

a. Cluster Analysis

b. Predictive Modelling

c. Association Analysis

d. Anomaly Detection

2.4.2.1. Cluster Analysis

Cluster analysis ini digunakan untuk menemukan kelompok yang memiliki hubungan yang berdekatan pada suatu observasi yang dilakukan.

2.4.2.2. Predictive Modelling

Predictive modeling digunakan untuk menemukan variable tujuan dengan membangun suatu model. Pada metode ini terdapat dua macam model yaitu, regresi yang digunakan untuk variable tujuan yang kontinyu dan klasifikasi yang digunakan untuk variable tujuan yang diskrit.

Contoh dari penggunaan *predictive modeling* adalah memprediksikan harga barang di masa mendatang yang merupakan metode regresi karena harga barang merupakan atribut nilai kontinyu, sedangkan contoh lainnya adalah pelanggan yang melakukan pembelian pada *online shopping* termasuk dalam metode klarifikasi karena variable tujuannya adalah nilai biner.

Dengan menggunakan *predictive modeling* ini diharapkan bisa dapat digunakan untuk mempelajari sebuah model yang akan meminimalkan tingkat kesalahan dan error antara prediksi dengan nilai yang sebenarnya pada suatu variable tujuan. *Predictive modeling* juga dapat digunakan unutk memprediksi gangguan dan kesalahan ataupun pelanggan yang berpotensial.

2.4.2.3. Association Analysis

Association analysis atau seringkali disebut juga dengan *association rule* adalah salah satu teknik utama pada *data mining* dan merupakan bentuk yang paling umum dipakai dalam menemukan *pattern* dari suatu kumpulan data. *Association rules* merupakan salah satu metode yang umumnya digunakan untuk mencari hubungan suatu *item*. Proses untuk menemukan hubungan antar *item* ini memerlukan pembacaan data transaksi secara berulang-ulang dalam jumlah data transaksi yang besar untuk menemukan *pattern-pattern* hubungan yang berbeda-beda. Oleh karena

itu, memerlukan biaya yang besar dan waktu yang lama, sehingga diperlukanlah algoritma yang efisien untuk menemukan hubungan tersebut.

Dalam menetukan suatu *association rules*, terdapat parameter yang menyatakan bahwa suatu informasi atau *knowledge* dianggap menarik (*interestingness measure*). Parameter ini didapatkan dari hasil pengolahan data dengan perhitungan tertentu. Parameter yang yang digunakan adalah *support* (nilai penunjang) yaitu proporsi dari transaksi pada suatu *data mining* yang mengandung *antecedent*dan *consequent*, sedangkan *confidence* adalah ukuran ketepatan suatu rule yaitu presentase transaksi dalam *data mining* yang mengandung *antecedent*dan *consequent*.

Association rules didefinisikan sebagai suatu proses untuk menemukan semua aturan minimum untuk *support (minimum support)* dan syarat minimum untuk *confidence (minimum confidence).*

Support* = Jumlah transaksi yang mengandung *antecedent*dan *consequent

Jumlah tranksaksi

Confidence = Jumlah transaksi dengan *item* dalam *antecedent*dan *consequent*

Jumlah tranksaksi dengan *item* dalam antecedent

2.4.2.4. Anomaly Detection

Anomaly detection adalah suatu metode yang digunakan untuk mengidentifikasikan hal-hal yang secara signifikan berbeda dengan data-data yang lain. Tujuan dari metode ini adalah untuk menemukan adanya suatu anomalis serta untuk menghindari kesalahan dengan melabelkan obyek biasa sebagai suatu

anomalis. *Anomaly detection* ini biasanya digunakan untuk mendeteksi penyakit yang tidak biasa, gangguan ekosistem seperti pendeteksian lubang pada lapisan ozon, segala bentuk penipuan, gangguan jaringan pada komputer dan sebagainya.

2.4.3. Market Basket Analysis

Market basket analysis merupakan suatu metodologi untuk melakukan analisis pola belanja konsumen dengan menemukan asosiasi antara beberapa item yang berbeda yang diletakkan konsumen ke dalam *shopping basket* pada suatu transaksi tertentu. Tujuan dari *market basket analysis* ini adalah uutk mengetahui produk apa saja yang ada kemungkinan untuk dibeli secara bersamaan. Analisis data transaksi bisa menghasilkan pola pembelian yang sering terjadi. Hasil yang didapatkan bisa dijadikan sebagai informasi yang berguna bagi penjual dalam pengambilan keputusan dan pengembangan strategi dengan melihat produk-produk mana saja yang sering dibeli secara bersamaan oleh konsumen. Pengembangan strategi penjualan yang bisa dilakukan misalnya dengan cara meletakkan secara berdekatan produk-produk yang sering dibeli bersamaan di suatu toko. *Market basket analysis* ini telah diterapkan oleh banyak toko baik retail maupun grosir (Olson & Delen, 2008).

2.4.3.1. Apriori Algorithm

Algoritma apriori adalah algoritma untuk menemukan pola frekuensi tinggi, Pola frekuensi tinggi adalah adalah pola-pola *item* di dalam suatu *data mining* yang memiliki frekuensi atau *support* di atas ambang batas tertentu yang disebut dengan minimum *support*. Pola frekuensi tinggi ini digunakan untuk menyusun *association rule* dan juga beberapa teknik *data mining* lainnya.

Association rule terdiri dari 2 sub persoalan yaitu menemukan semua kombinasi dari *item*, disebut dengan *frequent item*set yang memiliki *support* lebih besar daripada minimum *support*, dan menggunakan *frequent item*set untuk menjalankan aturan yang ditetapkan.

Algoritma Apriori ini berguna untuk menemukan *frequent item*sets yang diajalankan pada sekumpulan data. Pada iterasi ke-k maka akan d*item*ukan semua *item*set yang memiliki k *item*. Tiap iterasi terdiri dari dua tahap, yaitu:

a. Menggunakan *frequent* (k-1) *item*set untuk membangun kandidat *frequent* k-*item*set.
b. Menggunakan scan *data mining* dan pencocokan *pattern* untuk mengumpulkan hitungan pada kandidat *item*set.

Berikut faktor-faktor yang dapat mempengaruhi tingkat kompleksitas pada algoritma apriori, yaitu:

a. Pemilihan minimum *support*
   - Dengan menurunkan batas minimum *support* dapat menyebabkan semakin banyaknya *frequent item*set yang didapatkan.
   - Hal ini juga menyebabkan peningkatan jumlah dari kandidat dan panjang maksimum dari *frequent item*set.
b. Dimensi atau jumlah *item* pada dataset atau parameter
   - Lebih banyak ruang yang dibutuhkan untuk menyimpan hitungan *support* untuk setiap *item*.
   - Jika jumlah pada *frequent item* juga meningkat, baik komputasi dan *input output cost* juga akan meningkat.

c. Besarnya ukuran *data mining*

- Karena algoritma apriori membuat *multiple pass, run time* dari algoritma juga akan meningkat dengan jumlah dari transaksi.

d. Rata-rata panjang transaksi

- Lebar transaksi akan meningkatkan kepadatan *dataset* atau *parameter*.
- Hal ini akan meningkatkan panjang maksimum dari *frequent item*set dan garis lintang pada *hash tree* (jumlah dari subset di dalam sebuah transaksi meningkatkan lebarnya).

2.4.3.2. FP Growth Algorithm

FP Growth algorithm digunakan untuk melakukan pencarian *frequent itemset* tanpa harus melalui *candidate generation*. *FP Growth algorithm* menggunakan struktur data *FP Tree* sehingga cara kerja yang dilakukan oleh algoritma ini adalah dengan melakukan *scan database* yang dilakukan hanya dua kali saja, lalu kemudian data ditampilkan dalam bentuk *FP Tree*. Kemudian setelah *FP Tree* terbentuk, digunakanlah pendekatan *devide* dan *conquer* untuk mendapatkan *frequent itemset*.

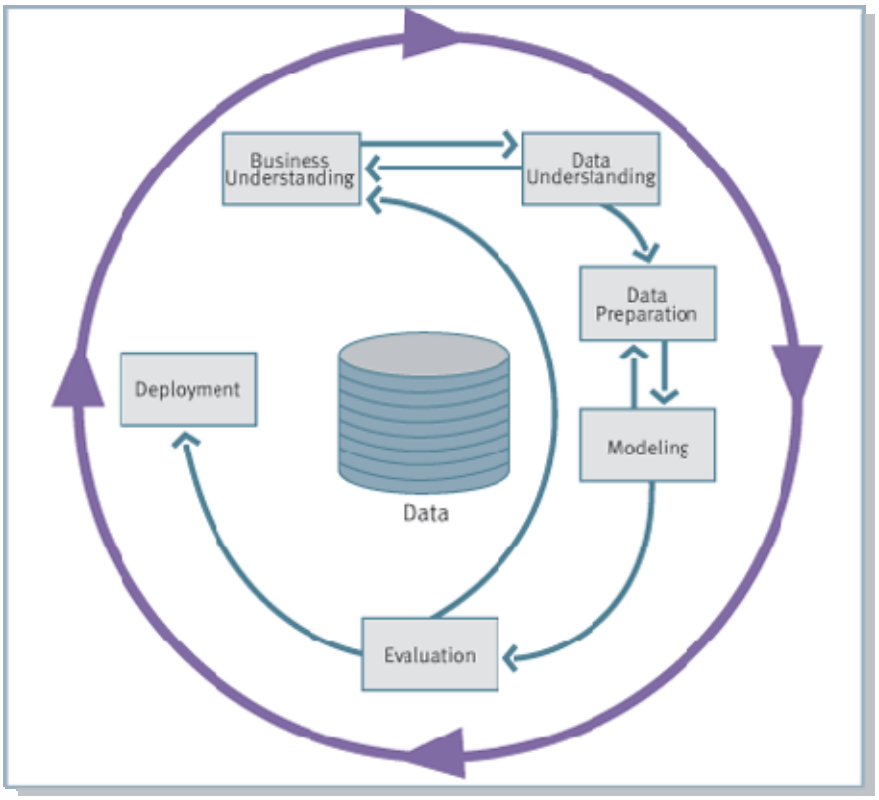
2.4.3.3. Vertical Format Algorithm

Vertical format algorithm merupakan sebuah algoritma baru yang digunakan untuk mencari frequent itemset dengan format vertikan. Algoritma ini hanya melakukan satu kali scanning database untuk mendapatkan frequent 1-itemset dan tidak memerlukan scan database lagi untuk tahap selanjutnya.

Keuntungan dari menggunakan *vertical format algorithm* ini adalah dapat menentukan itemset non-frequent sebelum generate itemset kandidat sehingga dapat menghemat waktu. Dengan menggunakan *vertical format algorithm* ini tidak perlu

lagi melakukan scan database untuk mencari tingkat support dari (k+1) itemset karena setiap id transaksi pada k-itemset membawa informasi yang lengkap yang dapat digunakan untuk mengkalkulasi tingkat support.

2.4.4. Metodologi *Data Mining*



Gambar 2.6. Metodologi CRISP-DM

Pada tahun 1996, sebuah konsorsium dari vendor dan user yang terdiri dari NCR System Engineering Copenhagen (Denmark), Daimler – Benz AG (Germany), SPSS/Integral Solutions Ltd. (England) dan OHRA Verzekeringen en Bank Groep BV (The Netherlands) membangun atau merumuskan sebuah spesifikasi sebagai standardisasi metodologi data mining dengan nama Cross Industry Standard Process for Data Mining (CRISP-DM) (Chapman, et al., 2000). Tahapan tersebut terdiri dari:

1. Business Understanding

Tujuan dari tahap ini adalah untuk mengerti tujuan proyek dan kebutuhan dasar dari perspektif bisnis dan kemudian mengkonversikan pengetahuan menjadi definisi masalah data mining dan kemudian membuat perencaan awal.

a. Determine Business Objection

Pada tahap ini tim proyek harus benar-benar memahami tujuan dari diadakannya proyek ini dan apa saja yang dibutuhkan untuk proyek ini dilihat dari perspektif bisnis agar dapat mengetahui apa yang benar-benar ingin dicapai. Untuk mengetahui informasi tersebut dapat dilakukan beberapa teknik seperti observasi langsung ke perusahaan, wawancara dengan pihak eksekutif, kuesioner, dll sehingga tim proyek dapat memahami masalah dan menghasilkan jawaban yang tepat dari pertanyaan yang benar.

- **Background**

  Tahapan background ini merupakan bagian pertama dari determine business objection yaitu melakukan pencarian informasi tentang situasi organisasi bisnis pada awal proyek dan gambaran dasar dari konteks proyek, di bidang apa proyek ini bekerja, identifikasi masalah, dan mengapa data mining muncul untuk memberikan solusi. Contohnya adalah menjelaskan latar belakang organisasi perusahaan serta struktur organisasi yang ada pada perusahaan tersebut.

- **Business Objectives**

  Tahapan berikutnya adalah menjelaskan kegiatan utama dan tujuan dari perspektif bisnis lalu menspesifikasikan apa saja yang dibutukan dalam menjalani proses bisnis serta manfaat apa saja yang diharapkan dalam melakukan kegiatan tersebut. Untuk Business Objectives ini bisa digambarkan dengan DFD, Flowchart, dsb untuk mengetahui proses kegiatan yang lebih rinci.

- **Business Succes Criteria**

  Setelah menjelaskan business objectives, tahapan berikutnya adalah menjelaskan kriteria sukses atau kegunaan hasil proyek dari sudut pandang bisnis. ini mungkin cukup spesifik dan dapat diukur secara obyektif, misalnya, pengurangan pelanggan yang pergi ke tingkat tertentu, atau mungkin umum dan subyektif, seperti "memberikan wawasan yang berguna dalam hubungan." Dalam kasus terakhir, harus ditunjukkan yang membuat penilaian subjektif. Aktivitas dari point ini adalah menyespesifikasikan kriteria keberhasilan usaha (misalnya, Meningkatkan tingkat respons dalam promosi mailing sebesar 10 persen dan tingkat pendaftaran sebesar 20 persen) setelah itu mengidentifikasi siapa yang menilai kriteria keberhasilan.

b. Asses Situation

Pada tahap ini dilakukan kegiatan penemuan fakta lebih rinci tentang sumber daya, kendala, asumsi, dan faktor lain yang harus

dipertimbangkan dalam menentukan tujuan analisis data dan rencana proyek.

- **Inventory of Resources**

  Daftar sumber daya yang tersedia untuk 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 perangkat lunak (data mining tools, perangkat lunak lain yang relevan).

- **Requirement, Assumptions and Constrains**

  Daftar semua requirements proyek, termasuk jadwal penyelesaian, komprehensibilitas dan kualitas hasil, dan keamanan, serta masalah hukum. Bagian dari output yang dihasilkan yaitu memastikan ijin untuk penggunaan data.

  Daftar asumsi yang dibuat oleh proyek. Ini mungkin asumsi tentang data yang dapat diverifikasi selama data mining, tetapi juga dapat mencakup asumsi non-diverifikasi tentang bisnis yang terkait dengan proyek. Hal ini sangat penting untuk mendaftarkan keduanya jika mempengaruhi keabsahan hasil.

  Daftar batasan pada proyek. Mungkin batasan pada ketersediaan sumber daya, tetapi juga dapat termasuk teknologi seperti ukuran dataset yang digunakan untuk pemodelan.

- **Risk and Contingencies**

  Daftar risiko atau kejadian yang mungkin menunda atau menyebabkan kegagalan proyek. Daftar rencana untuk kontingensi

yang sesuai dan tindakan apa yang akan diambil jika risiko ini atau peristiwa terjadi.

- **Terminology**

  Menyusun daftar istilah terminologi yang relevan dengan proyek. Ini mungkin meliputi dua komponen:

  - Sebuah glossary terminologi bisnis yang relevan, yang merupakan bagian dari pemahaman bisnis untuk proyek. Membangun glossary ini berguna sebagai "pengetahuan elisitasi" dan latihan pendidikan.
  - Sebuah glossary data mining terminologi, diilustrasikan dengan contoh-contoh yang relevan dengan pertanyaan masalah bisnis

- **Costs and Benefits**

  Buatlah sebuah analisis costs and benefits untuk proyek tersebut, yang membandingkan biaya proyek dengan potensi manfaat bagi bisnis jika berhasil. Perbandingan harus sespesifik mungkin. Sebagai contoh, menggunakan langkah-langkah moneter dalam situasi komersial.

c. Determine Data Mining Goals

Menjelaskan tujuan dari kegiatan bisnis. Tujuan dengan adanya data mining ini didapatkan berdasarkan dari tujuan bisnis yang dijelaskan dengan menggunakan istilah teknis. Sebagai contoh, tujuan bisnis adalah "meningkatkan penjualan katalog untuk pelanggan yang sudah ada". Tujuan data mining nya adalah "memprediksi berapa banyak

widget pelanggan akan membeli, mengingat pembelian mereka tiga tahun terakhir, informasi demografis (umur, gaji, kota, dll), dan harga item. "

- **Data Mining Goals**

  Menjelaskan output proyek yang diharapkan memungkinkan untuk pencapaian tujuan bisnis.

- **Data Mining Success Criteria**

  Menentukan kriteria sukses untuk hasil proyek dalam hal teknis. Misalnya, akurasi prediksi atau kecenderungan untuk membeli barang dengan tingkatan tertentu.

d. Produce Project Plan

Menjelaskan rencana yang ditetapkan untuk mencapai tujuan data mining sehingga dapat mencapai tujuan bisnis. Rencana harus terdiri dari langkah-langkah yang akan dilakukan dari awal sampai akhir proyek, termasuk pemilihan tools dan teknik yang dilakukan dalam proses pemgembangan data mining.

- **Project Plan**

  Berisi tahapan-tahapan yang akan dilakukan dalam proyek, waktu yang dibutuhkan, sumber daya yang dibutuhkan, input dan output.

  Rencana proyek berisi detail rencana untuk setiap tahap yang dilakukan. Pada tahapan ini harus ditentukan strategi evaluasi apa yang akan digunakan untuk tahap evaluasi. Rencana proyek berupa dokumen yang dinamis dimana pada akhir dari setiap tahapan dilakukan review untuk kemajuan serta update sesuai

dengan rencana proyek. Update review poin secara spesifik merupakan bagian dari rencana proyek.

- **Initial Assessment of Tools and Techniques**

  Pada akhir dari tahap pertama ini dilakukan proses penilaian awal terhadap toolsdan teknik yang digunakan. Penting untuk menilai tools dan teknik di tahap awal karena dapat mempengaruhi seluruh proyek. Kegiatan pada tahapan ini meliputi:

  - Membuat daftar yang berisi criteria pemilihan untuk tools dan teknik yang digunakan.
  - Memilih tool dan teknik yang paling berpotensial untuk digunakan dalam pengembangan data mining ini.
  - Mengevaluasi kelayakan dari teknik yang digunakan
  - Mereview dan memprioritaskan teknik yang digunakan berdasarkan hasil evaluasi dari pilihan solusi yang ada.

2. Data Understanding

Dimulai dengan mengumpulkan data kemudian mengenal data serta mengerti data dengan baik, dengan tujuan:

- Untuk mengidentifikasikan masalah kualitas data
- Untuk menemukan arti dari data
- Untuk membuat hipotesis dari data mengenai informasi yang tersembunyi

a. Collect Initial Data

Tahapan ini merupakan proses untuk mendapatkan data yang digunakan dalam pengembangan data mining. Jika data didapatkan berasal dari beberapa sumber data perlu diintegrasikan.

- ***Initial Data Collection***

  Menjelaskan dataset yang digunakan, diambil dari sumber mana serta metode yang digunakan untuk mendapatkan data tersebut (copy, create, transform).

b. Describe Data

- ***Data Description Report***

  Menjelaskan data yang digunakan, termasuk tipe data, panjang data serta identitas data. Tahapan ini perlu dilakukan untuk mengetahui apakah data yang digunakan sudah sesuai dengan kebutuhan.

c. Explore Data

Pada tahapan ini dilakukan proses pengeskploran data agar dapat mencapai tujuan data mining. Pengeksploran data meliputi proses querying, visualization dan reporting. Tahapan ini berkontribusi untuk memperbaiki deskripsi data dan laporan kualitas dari hasil data mining serta berguna untuk proses transformasi serta persiapan data untuk analisis berikutnya.

- ***Data Exploration Report***

  Menjelaskan hasil dari tahapan explore data ini, termasuk penemuan atau hipotesis awal serta dampakna terhadap kelanjutan proyek.

d. Verify Data Quality

Pada tahapan ini dilakukan kegiatan memeriksa kualitas data, untuk menjawab pertanyaan-pertanyaan seperti: Apakah data yang tersedia sudah lengkap? Apakah data yang digunakan itu sudah benar? Dll.

- ***Data Quality Report***

  Membuat daftar hasil verifikasi kualitas data,Jika terdapat masalah pada kualitas, maka buatlah solusi yang dapat mengatasi masalah tersebut. Solusi untuk masalah kualitas data umumnya sangat bergantung pada data dan pengetahuan bisnis.

3. Data Preparation

Mencakup semua kegiatan yang dibutuhkan untuk membangun dataset akhir dari data-data mentah. Dataset adalah data yang akan digunakan untuk dimasukkan kedalam modeling tool. Kegiatan ini meliputi: memilih table, case/record dan atribut untuk kemudian ditransformasi dan dibersihkan lalu setelah itu data dapat digunakan untuk modeling tool.

a. Select Data

Tentukan data yang akan digunakan untuk analisis. Kriteria meliputi relevansi dengan tujuan dari data mining, kualitas, dan kendala teknis seperti batas pada volume data atau tipe data. Seleksi data meliputi pemilihan atribut (kolom) serta pemilihan record (baris) dalam tabel.

- ***Rationale for Inclusion/Exclusion***

  Daftar data yang akan dimasukkan / dikeluarkan**.**

b. Clean Data

Data cleaning atau pembersihan data adalah suatu proses menghilangkan *noise* dan data yang tidak relavan atau data yang tidak konsisten.

- ***Data Cleaning Report***

  Menjelaskan hasil dari pembersihan data sehingga menghasilkan data yang siap untuk dimasukkan kedalam mining tools.

c. Construct Data

Tahapan ini mencakup operasi persiapan data konstruktif seperti produksi untuk derived attributers atau mengubah nilai untuk atribut yang ada.

- ***Derived Attributes Generated Records***

  Atribut yang diturunkan adalah atribut baru yang dibangun dari satu atau lebih atribut yang ada di record yang sama. Contoh: luas = panjang * lebar.

d. Integrate Data

Ini adalah metode dimana informasi dikombinasikan dari beberapa tabel untuk membuat nilai baru.

- ***Merged Data***

  Penggabungan tabel mengacu pada dua atau lebih tabel yang memiliki informasi yang berbeda tentang hal yang sama. Data gabungan juga mencakup agregasi . Agregasi mengacu pada operasi di mana nilai-nilai baru dihitung dengan meringkas informasi dari beberapa tabel . Misalnya, mengkonversi table pembelian di mana ada satu record untuk setiap pembelian ke tabel baru di mana ada satu record untuk setiap pelanggan , dengan field-field seperti jumlah pembelian , jumlah pembelian rata-rata , diskon, dll.

e. Formal Data

Transformasi format mengacu pada modifikasi sintaksis untuk data yang tidak berubah artinya, tetapi mungkin diperlukan oleh modeling tool.

- ***Reformatted Data***

  Beberapa tools memiliki persyaratan untuk urutan atribut, seperti field pertama harus unik untuk setiap record atau field terakhir menjadi outcome field untuk prediksi. Ini mungkin penting untuk mengubah urutan catatan dalam dataset. Mungkin alat pemodelan membutuhkan bahwa catatan akan diurutkan sesuai dengan nilai hasil atribut. Umumnya, catatan dataset yang awalnya memerintahkan dalam beberapa cara, tetapi algoritma pemodelan kebutuhan mereka berada dalam cukup acak. Misalnya, ketika menggunakan neural network, biasanya cara terbaik untuk catatan akan disajikan secara acak, meskipun beberapa tools menangani hal ini secara otomatis tanpa campur tangan user.

  Selain itu, ada perubahan sintaksis murni dibuat untuk memenuhi persyaratan modeling tools tertentu. Contoh: menghapus koma dari dalam bidang teks dalam file data dipisahkan oleh koma, pemangkasan semua nilai maksimum 32 karakter.

f. Dataset

Ini adalah dataset yang dihasilkan oleh tahap persiapan data, yang akan digunakan untuk pemodelan atau pekerjaan analisis utama proyek.

- ***Dataset Description***

  Menjelaskan dataset yang akan digunakan untuk pemodelan dan analisis pekerjaan utama proyek.

4. Modelling

Memilih dan menerapkan berbagai variasi dari teknik modeling dan membuat standar parameter dari modeling tool sampai nilai optimalnya. Ada beberapa teknik yang dapat digunakan dalam kasus data mining yang sama. Beberapa teknik mempunyai requirement yang berbeda terhadap data yang akan digunakan. Jika data yang akan digunakan belum siap maka kita harus kembali ke tahap data preparation.

a. Select Modelling Techniques

Sebagai langkah pertama dalam pemodelan, pilih teknik pemodelan yang yang akan digunakan. Meskipun Anda mungkin telah memilih tools selama fase business understanding, tahapan ini mengacu pada spesifik teknik pemodelan, misalnya, pembuatan decision-tree dengan 5,0. Jika menggunakan beberapa teknik untuk data mining yang sama maka tahapan ini dilakukan secara terpisah untuk masing-masing teknik.

- ***Modelling Techniques***

  Mendokumentasikan teknik pemodelan yang akan digunakan.

- ***Modelling Assumptions***

  Beberapa teknik pemodelan memiliki asumsi tertentu untuk data yang digunakan. misalnya, bahwa semua atribut memiliki distribusi seragam, tidak boleh ada nilai yang hilang, atribut kelas harus berupa simbol, dll.

b. Generate Test Design

Sebelum kita benar-benar membangun sebuah model, kita perlu untuk menghasilkan suatu prosedur atau mekanisme untuk menguji kualitas dan validitas model. Misalnya, jika menggunakan teknik data mining klasifikasi, umumnya menggunakan tingkat kesalahan sebagai ukuran kualitas untuk model data mining.

- ***Test Design***

  Menjelaskan rencana yang ditujukan untuk pelatihan, pengujian, dan evaluasi model. Sebuah komponen utama dari rencana adalah menentukan bagaimana membagi dataset yang tersedia dalam dataset pelatihan, pengujian, dan validasi.

c. Build Model

Menjalankan modeling tool pada saat menyiapkan dataset siap untuk menghasilkan satu atau lebih model.

- ***Parameter Settings***

  Dengan beberapa modeling tool, terdapat banyak parameter yang dapat disesuaikan. Buatlah daftar parameter dan nilai-nilai yang telah pilih sebagai pertimbangan untuk pengaturan parameter.

- ***Models***

  Ini adalah model yang sebenarnya dihasilkan oleh modeling tool, bukan berupa laporan.

- ***Model Descriptions***

  Menjelaskan model yang dihasilkan. Laporan penafsiran model dan mendokumentasikan kesulitan yang ditemui serta pemecahannya.

d. Assess Model

Data mining menafsirkan model berdasarkan pengetahuan domain-nya , data mining success criteria, dan rancangan yang diinginkan . Pada tahapan ini hanya melakukan mempertimbangkan model. Dalam tahap ini terdapat proses untuk menilai model sesuai dengan evaluasi kriteria . Sebisa mungkin , juga memperhitungkan tujuan bisnis dan kriteria kesuksesan bisnis . Dalam banyak proyek data mining bisa saja menggunakan beberapa teknik atau menghasilkan hasil data mining dengan

beberapa teknik yang berbeda . Dalam tugas ini , ia juga membandingkan semua hasil menurut kriteria evaluasi.

- ***Model Asessment***

  Merangkum hasil dari tahapan ini membuat, daftar kualitas model yang dihasilkan (misalnya, dalam hal akurasi), dan peringkat kualitas mereka dalam hubungan satu sama lain.

- ***Revised Parameter Setting***

  Berdasarkan asses model, buatlah dokumentasi yang berisi semua revisi dan assessment.

5. Evaluation

Evaluasikan model secara menyeluruh dan lihat kembali langkah-langkah membuat model yang telah dikerjakan untuk meyakinkan apakah model telah mencapai tujuan dari bisnis. Tentukan juga beberapa factor bisnis penting yang tidak ditangani oleh model. Pada akhir tahap ini, keputusan penggunaan hasil data mining telah ditentukan.

a. Evaluate Result

Langkah evaluasi sebelumnya berurusan dengan faktor-faktor seperti akurasi dari model. langkah ini menilai sejauh mana model memenuhi tujuan bisnis. Selain itu evaluasi dilakukan untuk menguji model, lalu evaluasi juga menilai hasil data mining lain yang dihasilkan. Hasil Data mining melibatkan model yang selalu berhubungan dengan tujuan bisnis dan

semua temuan lain yang tidak selalu berhubungan dengan tujuan bisnis.

- ***Assesment of DataMining Result w.r.t. Business Success Criteria***

  Merangkum hasil penilaian dalam hal kriteria keberhasilan usaha, termasuk pernyataan akhir tentang apakah proyek sudah memenuhi tujuan bisnis awal.

- ***Aproved Models***

  Setelah menilai model sesuai berdasarkan kriteria keberhasilan bisnis. Maka, model yang memenuhi kriteria yang dipilih menjadi model yang disetujui.

b. Review Process

Pada tahap ini, model yang dihasilkan harus dapat memenuhi kebutuhan bisnis. Dan perlu dilakukan review secara lebih menyeluruh untuk keterlibatan data mining dalam rangka untuk menentukan apakah terdapat faktor-faktor penting atau tahapan diabaikan. Ulasan ini juga mencakup kualitas isu-jaminan contoh: Apakah kita benar membangun model? Apakah hanya menggunakan atribut yang diijinkan untuk digunakan dan yang tersedia untuk analisis masa depan?

- ***Review of Process***

  Merangkum proses review dan menandakan proses mana yang belum dilakukan atau proses mana yang perlu dilakukan ulang.

c. Determine Next Steps

Berdasarkan pada hasil penilaian dan review proses, proses berikutnya adalah memutuskan bagaimana kelanjutan dari pengembangan proyek apakah akan menyelesaikan proyek ini dan beralih ke penyebaran, memulai lanjut iterasi, atau mendirikan proyek-proyek data mining baru. Tugas ini meliputi analisis sumber daya yang tersisa dan anggaran, yang dapat mempengaruhi keputusan.

- ***List of Possible Actions Decision***

  Membuat daftar tindakan lebih lanjutnya beserta alasan untuk setiap pilihan tindakan.

6. Deployment

Pada tahapan deployment ini dilakukan proses untuk mengorganisir dan mempresentasikan hasil dari data mining. Deployment dapat saja semudah membuat report otomatis, ataupun sesusah mengimplementasikan proses data mining secara berulang.

a. Plan Deployment

Proses ini membutuhkan hasil evaluasi dan menentukan strategi untuk penyebaran hasil data mining untuk digunakan oleh pihak eksekutif. ***Deployment Plan***

Merangkum strategi penggunaan hasil data mining, termasuk langkah-langkah yang diperlukan dan cara untuk melakukan hal tersebut.

b. Plan Monitoring and Maintenance

Perencaan untuk kegiatan monitoring dan maintenance merupakan hal penting dalam proses akhir pengembangan data mining ini untuk mencegah hal-hal yang tidak diinginkan terjadi serta memantau hasil dari data mining apakah sudah sesuai dengan kebutuhan yang diperlukan.

- ***Monitoring and Maintenance Plan***

  Pada tahap ini terdapat rangkuman monitoring dan strategi untuk melakukan proses pemeliharaan dan tahapan-tahapan untuk melakukannya.

c. Produce Final Report

Pada akhir proyek perlu dilakukan proses penulisan laporan akhir. Laporan berupa ringkasan proyek dan hasil data mining.

- ***Final Report***

  Final report adalah laporan tertulis akhir dari pengembangan data mining ini.

- ***Final Presentation***

  Final presentationadalah pertemuan dimana penulis mempresentasikan hasil dari pengembangan data mining ini.

d. Review Project

Dalam kegiatan review project ini dilakukan proses untuk menilai apakah proyek ini sudah dilakukan dengan baik dan apa saja yang perlu ditingkatkan.

- ***Experience Document***

  Meringkas pengalaman penting yang diperoleh selama proyek dan membuat laporan yang berisi dokumentasi dan hasil proyek.