بِسْمِ اللهِ الرَّحْمٰنِ الرَّحِيمِ ;اَلْحَمْدُ للهِ رَبِّ الْعالَمِينَ
اَللّٰهُمَّ صَلِّ عَلٰى سَيِّدِنَا مُحَمَّدٍ وَعَلٰى اٰلِ سَيِّدِنَا مُحَمَّدٍ

Tulisan ini diterjemahkan dan disesuaikan dari buku *Data Mining For Dummies*, karya Meta S Brown (2014).

CRISP-DM: PENGANTAR

Data mining tidak memiliki aturan resmi. kita memiliki fleksibilitas luar biasa untuk menentukan dan menyempurnakan metode kerja kita sendiri. Namun, kita akan mendapatkan manfaat jika memahami dan mengikuti pendekatan yang berhasil bagi orang lain. *Cross-Industry Standard Process for Data Mining* (CRISP-DM) adalah kerangka proses yang dominan untuk data mining. Ini adalah standar terbuka; siapa pun dapat menggunakannya. Bab ini menjelaskan setiap tahapan proses.

Standar Siapakah CRSIP-DM?

Model proses CRISP-DM adalah pendekatan langkah demi langkah dalam data mining yang dibuat oleh *data miner* untuk *data miner*. Peserta dari lebih dari 200 organisasi (terutama berbagai kelompok bisnis yang berkepentingan untuk menggunakan data mining secara internal atau mempromosikan penggunaan data mining secara luas) memberikan masukan untuk mengembangkan kerangka kerja, yang menguraikan tugas-tugas utama data mining dalam istilah bisnis dan membuat pengguna bebas menentukan pilihannya sendiri mengenai pendekatan matematika dan komputasi tertentu, serta masalah teknis lainnya.

Penjelasan proses CRISP-DM dalam bab ini sangat mirip dengan versi asli yang diterbitkan. Namun, terdapat perbedaan, seperti perubahan terminologi atau diagram, yang dimaksudkan untuk membuat informasi lebih jelas bagi pemula data mining. Selain itu, penjelasan dalam buku ini lebih singkat dan ringan gayanya. Jika kita ingin membaca dokumen asli yang belum diencerkan (semuanya berjumlah 76 halaman dalam cetakan kecil), kita bisa mendapatkannya secara online (gratis) di:

ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf

Pendekatan Proses Secara Bertahap

Model proses CRISP-DM memiliki enam fase utama, yaitu:

1. **Business understanding**
   Dapatkan pemahaman yang jelas tentang masalah yang ingin kita selesaikan, bagaimana dampaknya terhadap organisasi kita, dan tujuan kita untuk mengatasinya.
2. **Data understanding**
   Tinjau data yang kita miliki, dokumentasikan, dan identifikasi masalah pengelolaan data dan kualitas data.
3. **Data preparation**
   Siapkan data kita untuk digunakan untuk pemodelan.
4. **Modeling**
   Gunakan teknik matematika untuk mengidentifikasi pola dalam data kita.
5. **Evaluation**
   Tinjau pola yang kita temukan dan nilai potensinya untuk penggunaan bisnis.
6. **Deployment**
   Manfaatkan penemuan kita dalam bisnis sehari-hari.

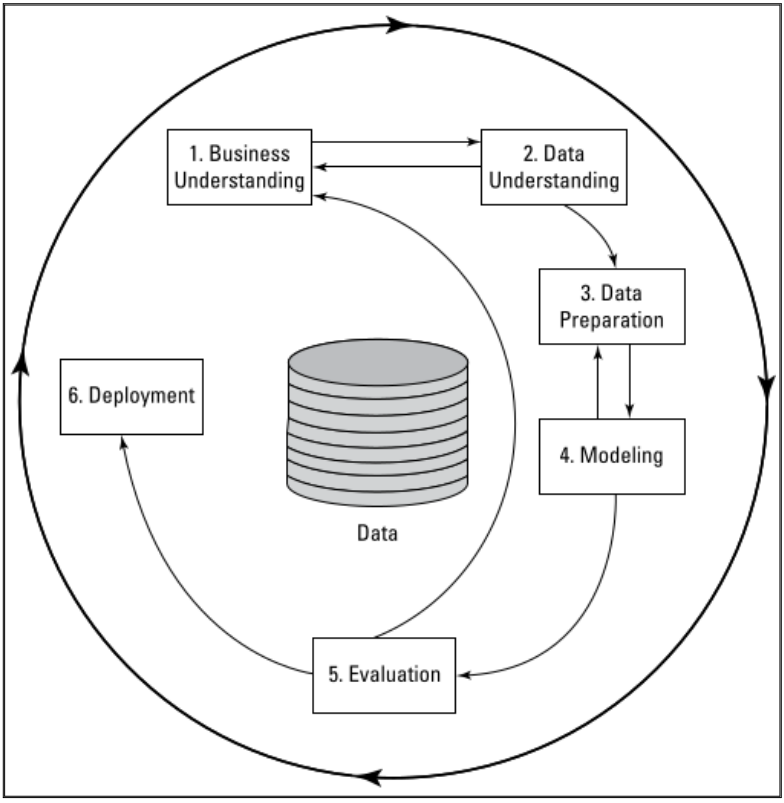
Masing-masing fase ini melibatkan beberapa pekerjaan utama, dan setiap pekerjaan memerlukan beberapa hasil, terutama laporan yang merangkum pekerjaan yang dilakukan dan informasi yang dipelajari dalam fase proses data mining tersebut. Namun, CRISP-DM tidak menentukan template untuk hasil kerja ini. kita harus merencanakan dan membuatnya agar sesuai dengan kebutuhan spesifik dan gaya tempat kerja kita.

Perlu untuk Diingat

CRISP-DM mendefinisikan proses data mining terutama dari sudut pandang bisnis. Ini memberi tahu kita banyak hal tentang apa yang perlu kita lakukan, tetapi tidak menjelaskan semua detail teknisnya.

Melewati Siklus Melalui Fase dan Proyek

Data mining bukanlah sesuatu yang kita lakukan sekali dan kemudian lupakan. Ini adalah siklus aktivitas yang berkelanjutan. Dalam proyek apa pun, kita mungkin hanya mengatasi elemen kecil dari masalah besar dan penting, namun kita akan kembali menemui masalah itu lagi dan lagi dengan proyek baru. Karena pekerjaan kita juga dapat diterapkan pada proyek baru, kita akan sering meninjau kembali proyek sebelumnya, untuk melihat apakah model yang kita kembangkan di masa lalu masih efektif dan untuk mencari peluang untuk meningkatkan apa yang telah kita lakukan. Mendaur ulang pekerjaan kita dengan cara ini meminimalkan upaya dan membantu kita menghindari kebingungan.



Gambar 1. Model proses CRISP-DM

Model proses CRISP-DM (bukan model matematika, namun seperangkat pedoman untuk pekerjaan data mining) adalah sebuah siklus yang sering diwakili oleh diagram seperti yang ditunjukkan pada Gambar 1. Setiap proyek dimulai dengan pemahaman bisnis dan langkah-langkah yang melalui lima fase proses. Dalam siklus tersebut, kita akan menemukan siklus yang lebih kecil, sehingga kita dapat melakukan beberapa langkah bolak-balik saat kita berupaya memahami bisnis dan datanya, atau untuk menyiapkan data dan membangun model. Siklus ini berulang seiring evaluasi proyek dan pengalaman kita selama penerapan menambah pemahaman kita tentang bisnis dan menginspirasi proyek baru.

بِسْمِ اللهِ الرَّحْمٰنِ الرَّحِيمِ ;اَلْحَمْدُ للهِ رَبِّ الْعالَمِينَ
اَللّٰهُمَّ صَلِّ عَلٰى سَيِّدِنَا مُحَمَّدٍ وَعَلٰى اٰلِ سَيِّدِنَا مُحَمَّدٍ

Tulisan ini diterjemahkan dan disesuaikan dari buku *Data Mining For Dummies*, karya Meta S Brown (2014).

CRISP-DM: *Business Understanding* (Pemahaman Bisnis)

Data mining bukanlah pengganti untuk *business understanding* (pemahaman bisnis). Pengetahuan bisnis kita memiliki nilai lebih daripada alat (*tool*) *data mining* atau analisis data apa pun. Alat tidak berarti apa-apa jika sendirian; mereka hanya menambah kecepatan dan kekuatan untuk membantu proses berpikir kita sendiri. Jika kita tidak tahu apa-apa tentang domain masalah, kita harus bekerja sama dengan seseorang yang memiliki pengetahuan itu.

Business understanding adalah kegiatan mendapatkan pemahaman yang jelas tentang masalah yang ingin kita selesaikan, bagaimana hal itu memengaruhi organisasi kita, dan tujuan kita untuk mengatasinya.

Pekerjaan-Pekerjaan di *Business Understanding*

Pada tahap pertama proyek *data mining*, sebelum kita menggunakan pendekatan data atau alat, kita harus menentukan apa yang ingin kita capai dan menentukan alasan ingin mencapai tujuan tersebut.

Pekerjaan-pekerjaan di *business understanding* mencakup empat pekerjaan (kegiatan utama, yang masing-masing dapat melibatkan beberapa bagian yang lebih kecil), yaitu:

1. Mengidentifikasi tujuan bisnis
2. Menilai situasi
3. Menentukan tujuan penambangan/analisis data
4. Membuat rencana proyek

Pekerjaan 1: Mengidentifikasi tujuan bisnis

Hal pertama yang harus kita lakukan dalam proyek apa pun adalah mencari tahu dengan tepat apa yang ingin kita capai! Itu tentu saja ini tidak

semudah kedengarannya. Banyak *data miner* telah menghabiskan waktu untuk analisis data, tetapi hanya untuk menemukan bahwa manajemen/ perusahaan/ organisasi mereka tidak terlalu tertarik dengan masalah yang sedang mereka selidiki. kita harus mulai dengan pemahaman yang jelas tentang hal-hal berikut ini:

- Masalah yang ingin ditangani oleh manajemen
- Tujuan bisnis
- Kendala (batasan tentang apa yang boleh dilakukan, jenis solusi yang dapat digunakan, kapan pekerjaan harus diselesaikan, dan sebagainya)
- Dampak (bagaimana masalah dan kemungkinan solusi cocok dengan bisnis)

Contoh kasus: E-Retail - Mengidentifikasi Tujuan Bisnis

Karena semakin banyak perusahaan melakukan transisi untuk menjual melalui Web, pengecer elektronik (e-retailer) pada produk komputer/elektronik yang telah ada menghadapi persaingan yang semakin ketat dari situs yang lebih baru. Dihadapkan pada kenyataan bahwa toko Web tumbuh lebih cepat daripada migrasi pelanggan ke Web, perusahaan harus menemukan cara untuk tetap menguntungkan meskipun biaya akuisisi pelanggan meningkat. Salah satu solusi yang diusulkan adalah memupuk hubungan pelanggan yang ada untuk memaksimalkan nilai dari masing-masing pelanggan perusahaan saat ini.

Dengan demikian, sebuah penelitian ditugaskan dengan tujuan sebagai berikut:

- Tingkatkan *cross-sales* dengan membuat rekomendasi yang lebih baik
- Tingkatkan loyalitas pelanggan dengan layanan yang lebih personal

Untuk sementara, penelitian akan dinilai berhasil jika:

- Peningkatan *cross-sale* sebesar 10%
- Pelanggan menghabiskan lebih banyak waktu dan melihat lebih banyak halaman di situs per kunjungan
- Studi selesai tepat waktu dan sesuai anggaran

Hasil kerja untuk pekerjaan ini mencakup tiga laporan (biasanya laporan singkat yang berfokus hanya pada poin utama):

- **Latar Belakang**
  Laporan ini untuk menjelaskan situasi bisnis yang mendorong sebuah proyek. Laporan ini, seperti kebanyakan yang telah dilakukan, hanya berjumlah beberapa paragraf saja. Berikut ini contoh laporan latar belakang:

  > Klien kami, komisi perencanaan regional, berupaya mempengaruhi penggunaan properti untuk meningkatkan kualitas hidup penduduk lokal. Komisi perencanaan memiliki hak yang luas yang memungkinkan untuk mempertimbangkan isu-isu luas termasuk pekerjaan, rekreasi, lingkungan, dan banyak aspek lain dari kehidupan masyarakat; namun peran komisi ini murni sebagai penasehat. Komisi ini memiliki kebebasan yang besar untuk memilih masalah apa yang dipelajari, melakukan penelitian, dan membuat rekomendasi kebijakan kepada pembuat undang-undang dan staf lokal, tetapi tidak memiliki kekuatan independen untuk menetapkan peraturan atau mempengaruhi pemilik properti.
  >
  > Anggota komisi (dan anggota lainnya di pemerintah daerah dan organisasi sipil) percaya kesempatan terbaik untuk mempengaruhi penggunaan properti terjadi ketika properti berpindah tangan. Ini menyiratkan bahwa upaya perencanaan pemerintah daerah dapat mencapai dampak terbesar dengan berfokus pada properti yang akan berpindah kepemilikan. Tetapi masalah yang dihadapi dalam hal ini adalah: waktu terbaik untuk bertindak adalah sebelum properti berpindah tangan, tetapi pemerintah daerah tidak memiliki informasi yang dapat diandalkan

tentang properti mana yang kemungkinan besar akan dialihkan. (Daftar *real estate* komersial mungkin berguna, tetapi tidak mencakup semua transfer properti, dan waktu terbaik untuk bertindak sebelum properti terdaftar).

Penelitian sebelumnya telah mengidentifikasi sejumlah faktor yang diyakini yang menunjukkan perubahan kepemilikan yang akan datang; ini termasuk kepemilikan nonlokal, beberapa pelanggaran kode bangunan, dan penyitaan. Meskipun komisaris memiliki alasan kuat untuk percaya bahwa faktor-faktor ini memengaruhi kemungkinan properti berpindah tangan, efeknya belum dihitung.

- **Tujuan bisnis**
  Laporan ini untuk menentukan apa yang ingin dicapai organisasi kita dengan proyek tersebut. Ini biasanya merupakan tujuan yang lebih luas daripada yang dapat kita capai sebagai data miner secara mandiri. Misalnya, tujuan bisnis mungkin untuk meningkatkan penjualan sebesar 10 persen dari tahun ke tahun dari kampanye iklan liburan.

- **Kriteria Keberhasilan Bisnis**
  Laporan ini untuk menentukan bagaimana hasil akan diukur. Cobalah untuk mendapatkan kriteria keberhasilan kuantitatif yang jelas. Jika kita harus menggunakan kriteria subyektif (petunjuk: istilah seperti mendapatkan wawasan atau menangani kriteria subyektif yang tersirat), setidaknya dapatkan kesepakatan tentang siapa yang akan menilai apakah kriteria tersebut telah dipenuhi atau tidak.

Pekerjaan 2: Menilai situasi

Pada fase inilah kita mendapatkan detail lebih lanjut tentang masalah yang terkait dengan tujuan bisnis kita. Sekarang kita akan masuk

lebih dalam ke pencarian fakta, membangun penjelasan yang lebih lengkap tentang masalah yang diuraikan dalam tugas tujuan bisnis.

Contoh kasus: E-Retail – Menilai Situasi

Ini adalah upaya pertama e-retailer elektronik di Web *mining*, dan perusahaan telah memutuskan untuk berkonsultasi dengan spesialis *data mining* untuk membantu memulai. Salah satu tugas pertama yang dihadapi konsultan adalah menilai sumber daya perusahaan untuk *data mining*.

Personil

Jelas bahwa ada keahlian internal dalam mengelola log server dan database produk dan pembelian, tetapi sedikit pengalaman dalam *data warehouse* dan pembersihan data untuk analisis. Dengan demikian, spesialis *database* juga dapat dikonsultasikan. Karena perusahaan berharap hasil penelitian akan menjadi bagian dari proses Web mining yang berkelanjutan, manajemen juga harus mempertimbangkan apakah posisi apa pun yang dibuat selama upaya saat ini akan menjadi posisi permanen.

Data

Karena ini adalah perusahaan yang mapan, ada banyak log Web dan data pembelian untuk diambil. Untuk studi awal ini, perusahaan akan membatasi analisis pada pelanggan yang telah "terdaftar" di situs tersebut. Jika berhasil, program dapat diperluas.

Risiko

Selain pengeluaran uang untuk konsultan dan waktu yang dihabiskan oleh karyawan untuk penelitian, tidak ada banyak risiko langsung dalam usaha ini. Namun, waktu selalu penting, jadi proyek awal ini dijadwalkan untuk satu kuartal keuangan.

Selain itu, tidak ada banyak arus kas tambahan saat ini, jadi sangat penting agar studi dilakukan di bawah anggaran. Jika salah satu dari tujuan ini berada dalam bahaya, manajer bisnis telah menyarankan agar ruang lingkup proyek dikurangi.

Hasil kerja untuk tugas ini mencakup lima laporan mendalam:

- **Inventarisasi Sumber Daya**
  Laporan ini merupakan daftar semua sumber daya yang tersedia untuk proyek. Ini mungkin termasuk orang (bukan hanya *data miner*), tetapi juga mereka yang memiliki pengetahuan ahli tentang masalah bisnis, manajer data, dukungan teknis, dan lainnya), data, perangkat keras, dan perangkat lunak.

- **Kebutuhan**
  Kebutuhan akan mencakup jadwal penyelesaian, kewajiban hukum dan keamanan, dan kebutuhan untuk pekerjaan selesai yang dapat diterima. Inilah poin untuk memverifikasi bahwa kita akan memiliki akses ke data yang sesuai!

- **Risiko dan Kontingensi**
  Identifikasi penyebab yang dapat menunda penyelesaian proyek, dan siapkan rencana kontingensi untuk masing-masingnya. Misalnya, jika pemadaman Internet di kantor kita dapat menimbulkan masalah, kemungkinan kita harus bekerja di kantor lain hingga pemadaman berakhir.

- **Terminologi**
  Buat daftar istilah bisnis dan istilah *data mining* yang relevan dengan proyek kita dan tuliskan dalam glosarium dengan definisi (dan mungkin contoh), sehingga setiap orang yang terlibat dalam proyek dapat memiliki pemahaman yang sama tentang istilah-istilah tersebut.

- **Biaya dan Manfaat**
  Persiapkan analisis biaya-manfaat untuk proyek tersebut. Cobalah untuk menyatakan semua biaya dan manfaat dalam dolar (rupiah, euro, pound, yen, dan sebagainya). Jika manfaatnya tidak melebihi biaya secara signifikan, hentikan dan pertimbangkan kembali analisis ini dan proyek kita.

TIP
Pengambil keputusan sering kali merasa lebih nyaman mengalokasikan sumber daya untuk proyek yang mengurangi biaya daripada proyek yang bertujuan untuk meningkatkan pendapatan, jadi selalu cari potensi penghematan biaya, dan nyatakan peluang penghematan terlebih dahulu dalam laporan biaya dan manfaat kita.

Pekerjaan 3: Menentukan Tujuan *Data Mining*

Dalam mencapai tujuan bisnis seringkali membutuhkan aksi dari banyak orang, bukan hanya *data miner*. Jadi sekarang, kita harus menentukan bagian kecil kita dalam gambaran yang lebih besar. Jika tujuan bisnis adalah untuk mengurangi pengurangan pelanggan, misalnya, tujuan data mining kita mungkin untuk mengidentifikasi tingkat pengurangan untuk beberapa segmen atau kelompok pelanggan, dan mengembangkan model untuk memprediksi pelanggan mana yang memiliki risiko terbesar.

Hasil kerja untuk tugas ini mencakup dua laporan:

- **Tujuan data mining**
  Tentukan hasil data mining, seperti model, laporan, presentasi, dan kumpulan data yang diproses.

- **Kriteria keberhasilan data mining**
  Tetapkan kriteria teknis *data mining* yang diperlukan untuk mendukung kriteria keberhasilan bisnis. Cobalah untuk mendefinisikannya dalam istilah kuantitatif (seperti akurasi model atau peningkatan prediktif dibandingkan dengan metode yang ada). Jika kriteria harus kualitatif, identifikasikan orang yang membuat penilaian.

Pekerjaan 4: Membuat Rencana Proyek

Sekarang kita menentukan setiap langkah yang ingin kita lakukan, sebagai *data miner*, buatlah langkah hingga proyek selesai dan hasilnya dipresentasikan dan direview.

Hasil kerja untuk tugas ini meliputi dua laporan:

- **Rencana proyek**
  Uraikan rencana tindakan langkah demi langkah kita untuk proyek tersebut. Perluas garis besar dengan jadwal penyelesaian setiap langkah, sumber daya yang diperlukan, input (seperti data atau pertemuan dengan pakar materi pelajaran), dan output (seperti data yang dibersihkan, model, atau laporan) untuk setiap langkah, dan dependensi (langkah-langkah yang tidak dapat dimulai hingga langkah ini selesai). Nyatakan secara eksplisit bahwa langkah-langkah tertentu harus diulangi (misalnya, pemodelan dan evaluasi biasanya memerlukan beberapa pengulangan bolak-balik).

- **Penilaian awal terhadap alat dan teknik**
  Identifikasi kemampuan yang diperlukan untuk memenuhi tujuan *data mining* atau analisis data kita dan nilai alat, dan sumber daya yang kita miliki. Jika ada sesuatu yang tidak terpenuhi, kita harus mengatasi masalah itu sejak awal proses.

بِسْمِ اللهِ الرَّحْمٰنِ الرَّحِيمِ ;اَلْحَمْدُ للهِ رَبِّ الْعالَمِينَ
اَللّٰهُمَّ صَلِّ عَلٰى سَيِّدِنَا مُحَمَّدٍ وَعَلٰى اٰلِ سَيِّدِنَا مُحَمَّدٍ

Tulisan ini diterjemahkan dan disesuaikan dari buku *Data Mining For Dummies*, karya Meta S Brown (2014).

CRSIP-DM: *Data Understanding* (Pemahaman Data)

Fase kedua proyek data mining dilakukan setelah kita menentukan tujuan dan membuat rencana. Pada fase sebelumnya, kita telah mendapatkan data dan memverifikasi bahwa data tersebut sesuai dengan kebutuhan kita. Pada fase tersebut kita mungkin mengidentifikasi masalah yang menyebabkan kita harus kembali mempelajari pemahaman bisnis dan merevisi rencana kita. Kita bahkan mungkin menemukan kekurangan dalam pemahaman bisnis kita, alasan lain untuk memikirkan kembali tujuan dan rencana.

Pekerjaan-Pekerjaan di Fase Data Understanding

Fase *data understanding* mencakup empat pekerjaan, yaitu:

- Mengumpulkan data
- Mendeskripsikan data
- Mengeksplorasi data
- Memverifikasi kualitas data

Pekerjaan 1: Mengumpulkan data

Di fase *business understanding* kita baru saja menetapkan tujuan dan menentukan rencana data mining. Setiap langkah rencana bergantung pada apakah kita memiliki data yang tepat atau tidak. Lebih baik untuk memastikan bahwa kita benar-benar memiliki data tersebut!

Hanya ada satu hasil untuk tugas ini: laporan pengumpulan data awal. Dalam laporan kita, kita perlu memverifikasi bahwa kita telah memperoleh data atau setidaknya memperoleh akses ke data, menguji

proses akses data, dan memverifikasi bahwa data tersebut ada. Kita juga perlu me-*load* data ke *tool* apa pun yang akan kita gunakan untuk data mining guna memverifikasi bahwa alat tersebut kompatibel dengan data. Kita mungkin saja melakukan banyak pekerjaan untuk mengumpulkan data yang kita butuhkan sebelum kita dapat menulis laporan ini. Untuk itu, pertama kita akan membuat rencana kita, sebagai berikut:

- **Buat *outline* persyaratan data**
  Buat daftar jenis data yang diperlukan untuk mencapai tujuan data mining. Perluas daftar dengan detail seperti rentang waktu dan format data yang diperlukan.

- **Verifikasi ketersediaan data**
  Konfirmasikan bahwa data yang diperlukan ada, dan kita dapat menggunakannya. Jika beberapa data yang kita inginkan tidak tersedia, putuskan bagaimana kita akan mengatasi masalah tersebut. Pertimbangkan alternatif seperti:
  - Mengganti dengan sumber data alternatif
  - Mempersempit ruang lingkup proyek
  - Mengumpulkan data baru

- **Menentukan kriteria seleksi**
  Identifikasi sumber data tertentu (database, file, dokumen, dan sebagainya.) yang akan kita gunakan. Dalam sumber tersebut, tentukan tabel, *field*, dan rentang kasus yang relevan dengan proyek ini.

Setelah kita melewati langkah-langkah ini, kita harus benar-benar mendapatkan datanya. Pada tahap ini, impor data ke platform *data mining* yang akan kita gunakan untuk proyek guna mengonfirmasi bahwa hal itu mungkin dilakukan dan kita memahami prosesnya. Selama uji coba ini, kita mungkin menemukan batasan perangkat lunak (atau perangkat keras) yang tidak kita antisipasi, seperti

- Batasan jumlah kasus atau *field*, atau jumlah memori yang dapat kita gunakan
- Ketidakmampuan untuk membaca format data dari sumber kita

- Kesulitan menangani ketidaksempurnaan dalam data (misalnya, kita mungkin menemukan produk yang tidak dapat mengimpor atau menganalisis kumpulan data yang tidak lengkap)

Terakhir, rangkum proses pengumpulan dalam sebuah laporan. Laporan tersebut harus menggambarkan kebutuhan kita, dan menjelaskan secara rinci data apa yang telah kita kumpulkan dan dari sumber apa. Di sini kita mengonfirmasi bahwa kita benar-benar memperoleh data dan data tersebut kompatibel dengan platform *data mining* kita. Jika kita mengalami kesulitan, kita akan menjelaskan apa itu kesulitannya dan bagaimana kita mengatasinya (menggunakan sumber alternatif, merevisi rencana, mengubah format).

Hasil untuk tugas ini hanyalah laporan sederhana, tetapi pekerjaan yang perlu kita lakukan sebelum kita dapat menulis laporan tersebut tidak akan sederhana! Akses data dapat menjadi salah satu bagian yang paling menantang dan membuat frustrasi dari proses *data mining*, penuh dengan tantangan teknis dan bisnis.

Pekerjaan 2: Mendeskripsikan data

Sekarang setelah kita telah memiliki data, siapkan gambaran umum tentang data yang kita miliki. Hasil tugas ini adalah laporan deskripsi data. Di dalamnya, kita mendeskripsikan sumber dan format data, jumlah kasus, jumlah dan deskripsi *field*, dan informasi umum lainnya yang mungkin penting. kita juga membuat evaluasi singkat tentang kesesuaian data untuk tujuan *data mining* kita. Misalnya, verifikasikan bahwa data menyertakan *field* yang kita harapkan dan perlu ada serta kasus yang memadai untuk analisis.

Pekerjaan 3: Mengeksplorasi data

Dalam pekerjaan ini, kita memeriksa data lebih dalam. Untuk setiap variabel, kita akan mencari rentang nilai dan distribusinya. Kita akan menggunakan manipulasi data sederhana dan teknik statistik dasar untuk pemeriksaan lebih lanjut ke dalam data. Eksplorasi data adalah untuk mendukung beberapa tujuan, yaitu:

- Mengenali data
- Menemukan tanda-tanda masalah kualitas data

- Mengatur tahapan langkah-langkah persiapan data

Hasil untuk pekerjaan ini adalah laporan eksplorasi data. Laporan adalah tempat untuk mendokumentasikan setiap hipotesis atau temuan awal yang telah kita kembangkan selama eksplorasi data. Laporan ini harus mencakup deskripsi data yang lebih rinci daripada laporan deskripsi data, termasuk distribusi, ringkasan, dan tanda-tanda masalah kualitas data.

Pekerjaan 4: Memverifikasi kualitas data

Kita memiliki data dan kita telah memeriksanya, dan sekarang kita harus menentukan apakah itu cukup baik untuk mendukung tujuan kita. Kita akan sering memiliki beberapa masalah kualitas yang harus diatasi namun masih dapat bergerak maju, namun terkadang juga kualitas data yang sangat buruk sehingga tidak dapat mendukung rencana kita dan kita harus mencari alternatif. Beberapa masalah data terburuk adalah:

- Data yang kita butuhkan tidak ada. (Apakah tidak pernah ada, atau dibuang? Bisakah data ini dikumpulkan dan disimpan untuk digunakan di masa mendatang?)
- Datanya ada, tetapi kita tidak dapat memilikinya. (Dapatkah pembatasan ini diatasi?)
- Kita menemukan masalah kualitas data yang parah (banyak nilai yang hilang atau salah yang tidak dapat diperbaiki).

Hasil untuk pekerjaan ini adalah laporan kualitas data. Laporan ini adalah ringkasan data yang kita miliki, masalah kecil dan besar tentang kualitas yang kita temukan, dan solusi yang mungkin untuk masalah kualitas atau alternatif (seperti menggunakan sumber data alternatif). Jika kita menghadapi masalah kualitas data yang sangat serius dan tidak dapat mengidentifikasi solusi yang memadai, kita mungkin harus merekomendasikan untuk mempertimbangkan kembali tujuan atau rencana.

بِسْمِ اللهِ الرَّحْمٰنِ الرَّحِيمِ ;اَلْحَمْدُ للهِ رَبِّ الْعالَمِينَ
اَللّٰهُمَّ صَلِّ عَلٰى سَيِّدِنَا مُحَمَّدٍ وَعَلٰى اٰلِ سَيِّدِنَا مُحَمَّدٍ

Tulisan ini diterjemahkan dan disesuaikan dari buku *Data Mining For Dummies*, karya Meta S Brown (2014).

CRISP-DM: DATA PREPARATION (Persiapan Data)

Data miner menghabiskan sebagian besar waktunya pada fase ketiga proses data mining, yaitu persiapan data. Sebagian besar data yang digunakan untuk data mining pada awalnya dikumpulkan dan disimpan untuk tujuan lain dan memerlukan beberapa penyempurnaan sebelum siap digunakan untuk pemodelan.

Tahap persiapan data mencakup lima pekerjaan, yaitu:

- Memilih data
- Membersihkan data
- Membangun data
- Mengintegrasikan data
- Memformat data

Panduan langkah demi langkah CRISP-DM tidak secara eksplisit menyebutkan dataset sebagai hasil untuk setiap pekerjaan persiapan data, namun dataset tersebut sudah ada dengan lebih baik dan diarsipkan serta didokumentasikan dengan baik. Dataset tidak akan berhubungan satu per satu dengan pekerjaan, namun informasi tentang data yang digunakan harus disertakan dalam setiap laporan yang dapat disampaikan.

Pekerjaan 1: Memilih data

Sekarang kita akan memutuskan bagian mana dari data yang kita miliki yang sebenarnya akan digunakan untuk data mining.

Hasil dari pekerjaan ini adalah alasan inklusi dan eksklusi. Di dalamnya, kita akan menjelaskan data apa yang akan dan tidak akan digunakan untuk pekerjaan data mining lebih lanjut. kita akan menjelaskan alasan untuk menyertakan atau mengecualikan setiap bagian data yang kita miliki, berdasarkan relevansi dengan sasaran, kualitas data, dan masalah teknis, seperti batasan jumlah field atau baris yang dapat ditangani alat kita, atau kesesuaian format data dengan kebutuhan kita.

Pekerjaan 2: Membersihkan data

Data yang kita pilih untuk digunakan kemungkinan besar tidak benar-benar bersih (bebas kesalahan). Kita akan membuat perubahan, mungkin melacak sumber untuk melakukan koreksi data tertentu, mengecualikan beberapa kasus atau sel individual (item data), atau mengganti beberapa item data dengan nilai default atau penggantian yang dipilih dengan teknik pemodelan yang lebih canggih. kita dapat memilih untuk hanya menggunakan sebagian data untuk semua atau sebagian pekerjaan data mining kita.

Hasil dari pekerjaan ini adalah laporan pembersihan data, yang mendokumentasikan, dengan sangat rinci, setiap keputusan dan tindakan yang digunakan untuk membersihkan data kita. Laporan ini harus mencakup dan mengacu pada setiap masalah kualitas data yang diidentifikasi dalam pekerjaan verifikasi kualitas data pada tahap proses pemahaman data. Laporan kita juga harus mengatasi potensi dampak terhadap hasil pilihan yang kita buat selama pembersihan data.

Pekerjaan 3: Membangun data

Kita mungkin perlu mendapatkan beberapa kolom baru (misalnya, menggunakan tanggal pengiriman dan tanggal pelanggan melakukan pemesanan untuk menghitung berapa lama pelanggan menunggu untuk menerima pesanan), menggabungkan data, atau membuat bentuk data baru.

Hasil pekerjaan ini mencakup dua laporan:

- Atribut turunan

  Laporan yang menjelaskan bidang (kolom) baru yang telah kita buat, bagaimana kita melakukannya, dan alasannya.

- Catatan yang dihasilkan
  Laporan yang menjelaskan kasus (baris) baru yang telah kita buat, bagaimana kita melakukannya, dan alasannya.

Informasi
Meskipun pekerjaan menggabungkan data dan memformat data dicantumkan terakhir dalam fase proses ini, pekerjaan tersebut tidak selalu berada di urutan terakhir, dan mungkin tidak muncul hanya sekali. kita mungkin harus melakukan penggabungan atau pemformatan ulang di awal tahap persiapan data.

Pekerjaan 4: Mengintegrasikan data

Data kita sekarang mungkin berada dalam beberapa dataset yang berbeda. kita harus menggabungkan beberapa atau semua dataset yang berbeda tersebut untuk bersiap menghadapi fase pemodelan. Hasil pekerjaan ini adalah data yang digabungkan. (Dan tidak ada salahnya untuk mendokumentasikan bagaimana penggabungan tersebut dilakukan).

Pekerjaan 5: Memformat data

Data sering kali datang kepada kita dalam format selain format yang paling nyaman untuk pemodelan. (Perubahan format biasanya didorong oleh desain alat kita). Jadi konversikan format tersebut sekarang.

Hasil pekerjaan ini adalah data kita yang telah diformat ulang. (Dan sedikit laporan yang menjelaskan perubahan yang kita buat akan menjadi hal yang cerdas untuk disertakan). Kita harus mengakhiri fase persiapan data dari proses data mining dengan dataset yang siap untuk dimodelkan dan laporan menyeluruh yang menjelaskan dataset tersebut.

بِسْمِ اللهِ الرَّحْمٰنِ الرَّحِيمِ ;اَلْحَمْدُ للهِ رَبِّ الْعالَمِينَ
اَللّٰهُمَّ صَلِّ عَلٰى سَيِّدِنَا مُحَمَّدٍ وَعَلٰى اٰلِ سَيِّدِنَا مُحَمَّدٍ

Tulisan ini diterjemahkan dan disesuaikan dari buku *Data Mining For Dummies*, karya Meta S Brown (2014).

CRISP-DM MODELING
(Pemodelan)

Pengantar

Ini adalah bagian dari proses yang paling disukai oleh sebagian besar data miner. Data kita sudah dalam kondisi yang baik, dan sekarang kita dapat mencari pola yang berguna dalam data kita.

Pekerjaan-Pekerjaan di Fase Modeling

Fase pemodelan mencakup empat pekerjaan, yaitu:

- Memilih teknik pemodelan
- Merancang pengujian
- Membangun model
- Menilai model

Pekerjaan 1: Memilih teknik pemodelan

Dunia data mining yang menakjubkan menawarkan banyak sekali teknik pemodelan, namun tidak semuanya sesuai dengan kebutuhan kita. Persempit daftar berdasarkan jenis variabel yang terlibat, pilihan teknik yang tersedia di tools yang kita miliki, dan pertimbangan yang penting bagi kita. (Misalnya, banyak organisasi menyukai metode dengan keluaran yang mudah diinterpretasikan, sehingga pohon keputusan atau regresi logistik mungkin dapat diterima, namun jaringan saraf mungkin tidak akan diterima.)

Hasil pekerjaan ini mencakup dua laporan:

- Teknik pemodelan: Tentukan teknik yang akan kita gunakan.

- Asumsi pemodelan: Banyak teknik pemodelan yang didasarkan pada asumsi tertentu. Misalnya, tipe model mungkin dimaksudkan untuk digunakan dengan data yang memiliki tipe distribusi tertentu. Dokumentasikan asumsi-asumsi ini dalam laporan ini.

Para ahli statistik mempunyai pengetahuan luas, ketat, dan “cerewet” mengenai asumsi. Hal ini belum tentu berlaku bagi data miner, dan ini bukan merupakan persyaratan untuk menjadi seorang data miner. Jika kita memiliki pengetahuan statistik yang mendalam dan memahami asumsi di balik model yang kita pilih, kita bisa bersikap ketat dan “cerewet” dalam menentukan asumsi. Namun banyak data miner, terutama data miner pemula, tidak terlalu mempermasalahkan asumsi. Alternatifnya adalah menguji (melakukan banyak sekali pengujian) model kita.

Pekerjaan 2: Merancang pengujian
Pengujian dalam pekerjaan ini adalah pengujian yang akan kita gunakan untuk menentukan seberapa baik model kita bekerja. Ini mungkin sesederhana, membagi data kita menjadi beberapa kelompok kasus untuk pelatihan model dan kelompok lain untuk pengujian model. Data pelatihan digunakan untuk menyesuaikan bentuk matematika dengan model data, dan data pengujian digunakan selama proses pelatihan model untuk menghindari overfitting (membuat model yang sempurna untuk satu kumpulan data, namun tidak untuk kumpulan data lainnya). kita juga dapat menggunakan data holdout, data yang tidak digunakan selama proses pelatihan model, untuk pengujian tambahan.

Hasil tugas ini adalah desain pengujian kita. Hal ini tidak perlu rumit, namun kita setidaknya harus berhati-hati agar data pelatihan dan pengujian kita serupa dan kita menghindari bias apa pun pada data.

Pekerjaan 3: Membangun model
Pemodelan adalah apa yang dibayangkan banyak orang sebagai keseluruhan pekerjaan data mining, padahal itu hanya satu dari lusinan tugas! Meskipun demikian, pemodelan untuk mencapai tujuan bisnis tertentu adalah inti dari profesi data mining.

Hasil pekerjaan ini mencakup tiga item:

- Pengaturan parameter: Saat membuat model, sebagian besar alat memberi kita opsi untuk menyesuaikan berbagai pengaturan, dan pengaturan ini berdampak pada struktur model akhir. Dokumentasikan pengaturan ini dalam laporan.
- Deskripsi model: Jelaskan model kita. Nyatakan jenis model (seperti regresi linier atau jaringan saraf tiruan) dan variabel yang digunakan. Jelaskan bagaimana model diinterpretasikan. Dokumentasikan setiap kesulitan yang dihadapi dalam proses pemodelan.
- Model: Hasil ini adalah model itu sendiri. Beberapa tipe model dapat dengan mudah didefinisikan dengan persamaan sederhana; yang lainnya terlalu rumit dan harus dikirimkan dalam format yang lebih canggih.

Pekerjaan 4: Menilai model

Sekarang kita akan meninjau model yang telah kita buat, dari sudut pandang teknis dan juga dari sudut pandang bisnis (seringkali dengan masukan dari pakar bisnis di tim proyek kita).

Hasil pekerjaan ini mencakup dua laporan:

- Penilaian model: Meringkas informasi yang dikembangkan dalam tinjauan model kita. Jika kita telah membuat beberapa model, kita dapat memberi peringkat pada model tersebut berdasarkan penilaian kita terhadap nilainya untuk aplikasi tertentu.
- Merevisi pengaturan parameter: kita dapat memilih untuk menyempurnakan pengaturan yang digunakan untuk membuat model dan melakukan putaran pemodelan lainnya dan mencoba meningkatkan hasil kita.

Tips

Data mining, seperti bawang, torte Dobos, atau batuan sedimen, memiliki banyak lapisan. Saat kita baru memulai data mining, kita bisa memulai dengan membiarkan pengaturan parameter pada nilai defaultnya (bahkan, kita mungkin tidak menyadari adanya opsi kecuali kita berusaha mencarinya). Saat kita merasa nyaman dengan karir data mining baru kita,

masuk akal bagi kita untuk mencari tahu tentang parameter model dan mengetahui bagaimana kita dapat menggunakannya. Pilihan kita akan sangat bervariasi tergantung jenis model dan alat spesifik yang kita gunakan.

بِسْمِ اللهِ الرَّحْمٰنِ الرَّحِيمِ ;اَلْحَمْدُ لِلّٰهِ رَبِّ الْعالَمِينَ
اَللّٰهُمَّ صَلِّ عَلٰى سَيِّدِنَا مُحَمَّدٍ وَعَلٰى اٰلِ سَيِّدِنَا مُحَمَّدٍ

Tulisan ini diterjemahkan dan disesuaikan dari buku *Data Mining For Dummies*, karya Meta S Brown (2014).

CRISP-DM: EVALUATION
(Evaluasi)

Kita telah mengeksplorasi data dan menemukan polanya, dan sekarang kita harus bertanya: Apakah hasilnya bagus? kita akan mengevaluasi tidak hanya model yang kita buat tetapi juga proses yang kita gunakan untuk membuatnya, dan potensi penggunaan praktisnya.

Fase pemahaman data mencakup tiga pekerjaan, yaitu:

- Mengevaluasi hasil
- Meninjau proses
- Menentukan langkah selanjutnya

Pekerjaan 1: Mengevaluasi hasil

Pada tahap ini, kita akan menilai nilai model kita untuk memenuhi tujuan bisnis yang memulai proses data mining. Kita akan mencari alasan mengapa model tersebut tidak memuaskan untuk penggunaan bisnis. Jika memungkinkan, kita akan menguji model tersebut dalam aplikasi praktis, Untuk menentukan apakah model tersebut berfungsi dengan baik di tempat kerja seperti yang terjadi pada pengujian kita.

Hasil pekerjaan ini mencakup dua item:

- Penilaian hasil (untuk tujuan bisnis)
  Rangkum hasil sehubungan dengan kriteria keberhasilan bisnis yang kita tetapkan pada fase pemahaman bisnis. Nyatakan secara eksplisit apakah kita telah mencapai tujuan bisnis yang ditentukan di awal proyek.

- Model yang disetujui
  Model ini mencakup model apa pun yang memenuhi kriteria keberhasilan bisnis.

Pekerjaan 2: Meninjau proses

Sekarang setelah kita mengeksplorasi data dan mengembangkan model, luangkan waktu untuk meninjau proses kita. Ini adalah kesempatan untuk menemukan masalah yang mungkin kita abaikan dan mungkin menarik perhatian kita pada kelemahan dalam pekerjaan yang telah kita lakukan sementara kita masih punya waktu untuk memperbaiki masalah sebelum penerapan. Juga pertimbangkan cara-cara yang dapat kita lakukan untuk meningkatkan proses kita untuk proyek-proyek masa depan.

Hasil pekerjaan ini adalah peninjauan laporan proses. Di dalamnya, kita harus menguraikan proses peninjauan dan temuan kita serta menyoroti segala kekhawatiran yang ada memerlukan perhatian segera, misalnya langkah-langkah yang terabaikan atau perlu ditinjau kembali.

Pekerjaan 3: Menentukan langkah selanjutnya

Fase evaluasi diakhiri dengan rekomendasi kita untuk langkah selanjutnya. Model tersebut mungkin siap untuk diterapkan, atau kita mungkin menilai akan lebih baik jika mengulangi beberapa langkah dan mencoba memperbaikinya. Temuan kita mungkin menginspirasi proyek data mining yang baru.

Hasil pekerjaan ini mencakup dua item:

- Daftar tindakan yang mungkin dilakukan
  Uraikan setiap alternatif tindakan, beserta alasan terkuat yang mendukung dan menentang tindakan tersebut.
- Keputusan
  Menyatakan keputusan akhir atas setiap tindakan yang mungkin dilakukan, beserta alasan di balik keputusan tersebut.

بِسْمِ اللهِ الرَّحْمٰنِ الرَّحِيمِ ;اَلْحَمْدُ لِلّٰهِ رَبِّ الْعالَمِينَ
اَللّٰهُمَّ صَلِّ عَلٰى سَيِّدِنَا مُحَمَّدٍ وَعَلٰى اٰلِ سَيِّدِنَا مُحَمَّدٍ

Tulisan ini diterjemahkan dan disesuaikan dari buku *Data Mining For Dummies*, karya Meta S Brown (2014).

CRISP-DM: DEPLOYMENT
PENERAPAN

Deployment adalah saat data mining membuahkan hasil. Tidak peduli seberapa cemerlang penemuan kita, atau seberapa sempurna model kita sesuai dengan data, jika kita tidak benar-benar menggunakan hal-hal tersebut untuk meningkatkan cara kita berbisnis.

Fase deployment mencakup empat pekerjaan. Ini, yaitu

- Merencanakan deployment (metode kita untuk mengintegrasikan penemuan data mining ke dalam penggunaan)
- Perencanaan pemantauan dan pemeliharaan
- Melaporkan hasil akhir
- Meninjau hasil akhir

Pekerjaan 1: Merencanakan deployment

Ketika model kita siap digunakan, kita memerlukan strategi untuk menerapkannya dalam bisnis kita. Hasil pekerjaan ini adalah rencana deployment. Ini adalah ringkasan strategi deployment kita, langkah-langkah yang diperlukan, dan petunjuk untuk melakukan langkah-langkah tersebut.

Pekerjaan 2: Merencanakan pemantauan dan pemeliharaan

Pekerjaan data mining adalah sebuah siklus, jadi harap untuk tetap terlibat secara aktif dengan model kita saat model tersebut diintegrasikan ke dalam penggunaan sehari-hari. Hasil dari pekerjaan ini adalah rencana pemantauan dan pemeliharaan. Ini adalah ringkasan strategi kita untuk peninjauan berkelanjutan terhadap performa model. Kita harus memastikan

bahwa model tersebut digunakan dengan benar secara berkelanjutan, dan setiap penurunan performa model akan terdeteksi.

Pekerjaan 3: Melaporkan hasil akhir

Hasil pekerjaan ini mencakup dua item:

- Laporan akhir
  Laporan akhir merangkum keseluruhan proyek dengan mengumpulkan semua laporan yang dibuat hingga saat ini, dan menambahkan gambaran umum yang merangkum keseluruhan proyek dan hasil-hasilnya.
- Presentasi akhir
  Ringkasan laporan akhir disajikan dalam pertemuan dengan manajemen. Ini juga merupakan kesempatan untuk menjawab pertanyaan terbuka apa pun.

Pekerjaan 4: Meninjau proyek

Terakhir, tim pengumpulan data bertemu untuk mendiskusikan apa yang berhasil dan apa yang tidak, apa yang sebaiknya dilakukan lagi, dan apa yang harus dihindari! Langkah ini juga mempunyai hasil, meskipun hanya untuk penggunaan tim data mining, bukan manajer (atau klien). Ini adalah laporan dokumentasi pengalaman. Di sinilah kita harus menguraikan metode kerja apa pun yang berhasil dengan baik, sehingga metode tersebut didokumentasikan untuk digunakan lagi di masa mendatang, dan perbaikan apa pun yang mungkin dilakukan pada proses kita. Ini juga merupakan tempat untuk mendokumentasikan masalah dan pengalaman buruk, dengan rekomendasi kita untuk menghindari masalah serupa di masa depan.

Data mining adalah aktivitas tim. Jadi, jika proses ini tampaknya mencakup banyak langkah, sadarilah bahwa mungkin bukan tanggung jawab pribadi kita untuk melakukan semuanya, dan meminta bantuan orang lain saat kita membutuhkannya adalah hal yang wajar. Pada awal proyek, kita membuat daftar orang-orang yang menjadi sumber daya untuk proyek data mining. Itu adalah direktori kecil berisi bantuan kita!