BINUS UNIVERSITY ONLINE LEARNING

People
Innovation
Excellence

LECTURE NOTES

ISYS8036 - Business Intelligent and Analytics

Topik 3

Proses Data Mining

LEARNING OUTCOMES

Setelah mempelajari materi ini peserta kuliah diharapkan mampu mengidentifikasikan dan memahami:

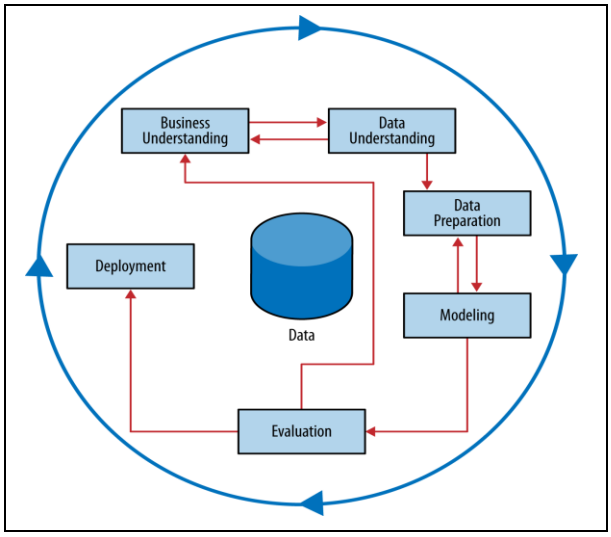
1. Fundamental concepts: A set of canonical data mining tasks;
2. The data mining process;
3. Supervised versus unsupervised data mining.

OUTLINE MATERI :

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Kesimpulan

ISI MATERI

Proses data mining adalah seni yang melibatkan penerapan sejumlah besar sains dan teknologi, Seperti banyak produk kesenian, ada proses yang dipahami dengan baik yang menempatkan struktur pada masalah, memungkinkan konsistensi yang wajar, pengulangan, dan obyektivitas. Sebuah kodifikasi atau kerangka kerja yang sangat popular dalam proses data mining adalah the Cross Industry Standard Process for Data Mining (CRISP-DM; Shearer, 2000), yang digambarkan dalam Gambar 2.1.



Gambar 2.1. Data mining proses menurut CRISP-DM (Cross Industry Standard Process for Data Mining).

Diagram ini menampilkan secara eksplisit fakta bahwa iterasi adalah sebuah keharusan. Melewati seluruh proses sekali tanpa memecahkan masalah secara umum, bukan dianggap sebagai kegagalan. Seringkali seluruh proses hanya terkait dengan eksplorasi data, dan setelahnya para anggota tim data sains dapat memahami lebih banyak. Iterasi selanjutnya bisa berjalan lebih bermakna karena didukung pemahaman yang lebih detil terhadap data. Bagian berikut akan didiskusikan langkah-langkah ini secara detail.

Business Understanding

Pada tahap paling awal, sangatlah penting untuk memahami masalah yang akan dipecahkan. Hal ini mungkin terlihat jelas, tetapi masalah bisnis sangat jarang berbentuk sebagai masalah data mining secara eksplisit. Memformulasikan kembali masalah bisnis dan merancang solusi yang

dituju adalah proses yang sering kali dilakukan secara berulang. Diagram yang ditunjukkan pada Gambar 2-1 memperlihatkan hal ini sebagai siklus dalam siklus, bukan sebagai proses linear sederhana. Formulasi awal mungkin tidak lengkap atau optimal sehingga beberapa iterasi mungkin diperlukan untuk formulasi solusi yang dapat diterima.

Tahap Pemahaman Bisnis merupakan bagian dari seni di mana kreativitas dan analis memainkan peran besar. Dalam konteks ini, data sains memiliki beberapa hal yang akan dijelaskan, tetapi sesungguhnya kunci sukses terletak pada perumusan masalah yang kreatif oleh beberapa analis mengenai bagaimana memformulasikan masalah bisnis sebagai satu atau lebih masalah data sains. Pengetahuan tingkat tinggi tentang prinsip prinsip fundamental data mining akan sangat membantu para analis bisnis yang kreatif merumuskan formulasi baru.

Pada Sesi berikut akan disajikan seperangkat tool/ teknik untuk memecahkan masalah penambangan data tertentu. Biasanya, pada tahap awal dipikirkan rencana yang memanfaatkan kemampuan solusi dari tool ini.

Pada tahap pertama ini, tim desain harus berpikir dengan hati-hati tentang skenario penggunaan. Ini merupakan salah satu konsep ilmu data yang paling penting, yang akan dibahas pada sesi sesi selanjutnya. Pertayaan pertanyaan seperti Apa sebenarnya yang ingin kita lakukan? Bagaimana tepatnya kita melakukannya? Bagian mana dari skenario penggunaan ini merupakan model penambangan data yang mungkin? Dalam membahas ini secara lebih rinci, kita akan mulai dengan pandangan yang disederhanakan dari skenario penggunaan, tetapi ketika kita maju kita akan memutar balik dan menyadari bahwa sering kali skenario penggunaan harus disesuaikan untuk lebih mencerminkan kebutuhan bisnis yang sebenarnya. Kita akan menyajikan teknik konseptual untuk membantu pemikiran, misalnya membingkai masalah bisnis dalam hal nilai yang diharapkan dapat memungkinkan kita untuk secara sistematis menguraikannya menjadi tugas data mining yang riil.

Data Understanding

Apabila pemecahkan masalah bisnis adalah tujuannya, maka data merupakan bahan baku dimana solusi akan dibangun. Sangatlah penting untuk memahami kemudahan dan keterbatasan data karena jarang terjadi terdapat kecocokan sama persis dengan masalah yang dihadapi. Data historis sering dikumpulkan untuk tujuan yang tidak terkait dengan masalah bisnis yang menjadi

fokus, atau tanpa tujuan yang jelas. Database pelanggan, data transaksi, dan data respons pemasaran berisi informasi yang berbeda, mencakup populasi yang berbeda namun mungkin beririsan, dan mungkin memiliki tingkat keandalan yang berbeda.

Merupakan sesuatu yang umum bahwa biaya dalam memperoleh data bervariasi. Beberapa data praktis tersedia secara gratis sementara yang lain membutuhkan upaya untuk mendapatkannya. Beberapa data mungkin dapat dibeli, namun dapat pula terjadi bahwa data yang dibutuhkan tidak tersedia dan membutuhkan pupaya ekstra untuk mengkoleksi. Bagian penting dari fase pemahaman data adalah memperkirakan biaya dan manfaat dari setiap sumber data dan memutuskan apakah investasi lebih lanjut layak. Bahkan setelah semua data set diperoleh, menyusunnya mungkin memerlukan upaya tambahan. Misalnya, catatan pelanggan dan pengenal produk dikenal bervariasi dan banyak error. Membersihkan dan mencocokkan catatan pelanggan untuk memastikan hanya satu catatan per pelanggan itu sendiri merupakan masalah analitik yang rumit (Hernandez & Stolfo, 1995; Elmagarmid, Ipeirotis, & Verykios, 2007).

Seiring semakin berkembangnya pemahaman terhadap data, jalur solusi dapat berubah arah sebagai respons, dan upaya tim bahkan mungkin bercabang. Sebagai contoh masalah deteksi penipuan (fraud detection). Data mining telah digunakan secara ekstensif untuk mendeteksi penipuan, dan banyak masalah deteksi penipuan melibatkan tugas penambangan data yang diawasi secara klasik. Perhatikan pekerjaan mengidentifikasi penipuan kartu kredit. Tagihan muncul di setiap akun pelanggan, sehingga biaya penipuan biasanya diketahui, bila bukan oleh perusahaan, namun akhirnya oleh pelanggan sendiri saat membaca aktivitas akun. Dapat diasumsikan bahwa hampir semua penipuan dapat diidentifikasi dan diberi label, karena pelanggan yang sah dan orang yang melakukan penipuan adalah orang yang berbeda dan memiliki niat yang berlawanan. Dengan demikian transaksi kartu kredit memiliki label yang dapat dipercaya (penipuan dan sah) yang dapat berfungsi sebagai target untuk teknik yang diawasi (supervised).

Mari kita tinjau masalah mengidentifikasi penipuan Medicare (Ausransi Kesehatan). Ini adalah masalah besar di Amerika Serikat dengan biaya miliaran dolar setiap tahun. Meskipun tampak seperti masalah deteksi penipuan konvensional, saat mempertimbangkan kaitan masalah bisnis dengan data, perlu disadri bahwa masalahnya sangat berbeda. Para pelaku penipuan, yang tidak lain adalah penyedia layanan medis yang mengajukan klaim palsu, dan kadang-kadang para pasien, semuanya merupakan pihak pihak yang sah untuk bertransaksi. Penipu adalah pengguna yang sah. Akibatnya, data penagihan Medicare tidak memiliki variabel target yang dapat diandalkan yang menunjukkan penipuan, dan karenanya pendekatan pembelajaran yang diawasi yang dapat digunakan untuk mendeteksi penipuan kartu kredit tidak dapat digunakan dalam

kasus ini. Masalah seperti itu biasanya membutuhkan pendekatan yang tidak diawasi seperti pembuatan profil, pengelompokan, deteksi anomali, dan pengelompokan co-occurrence.

Fakta bahwa keduanya adalah masalah deteksi penipuan adalah penyamaan yang semu yang sebenarnya menyesatkan. Dalam pemahaman data kita perlu menggali di bawah permukaan untuk mengungkap struktur masalah bisnis dan data yang tersedia, dan kemudian mencocokkannya dengan satu atau lebih tugas penambangan data yang mungkin kita miliki untuk menerapkan ilmu dan teknologi yang secara substansial tepat. Adalah hal yang tidak biasa, suatu masalah bisnis mengandung beberapa tugas penambangan data, dari jenis yang berbeda.

Data Preparation

Contoh umum persiapan data adalah mengkonversi data ke format tabel, menghapus atau menyimpulkan nilai yang hilang, dan mengonversi data ke jenis yang berbeda. Beberapa teknik penambangan data dirancang untuk data simbolis dan kategoris, sementara yang lain hanya menangani nilai numerik. Selain itu, nilai-nilai numerik harus sering dinormalisasi atau diskalakan sehingga dapat dibandingkan. Teknik standar dan aturan praktis tersedia untuk melakukan konversi seperti itu.

Secara umum, dalam kuliah ini tidak akan difokuskan pada teknik persiapan data. Namun akan didefinisikan format data dasar dalam Sesi-Sesi berikutnya, dan rincian persiapan data akan diberikan ketika menjelaskan beberapa prinsip dasar ilmu data atau diperlukan untuk menyajikan contoh konkret.

Sebagian ilmuwan data mungkin menghabiskan banyak waktu di awal proses mendefinisikan variabel yang digunakan dalam proses. Ini adalah salah satu poin utama di mana kreativitas manusia, akal sehat, dan pengetahuan bisnis ikut bermain. Seringkali kualitas solusi penambangan data terletak pada seberapa baik para analis menyusun masalah dan menyusun variabel.

Salah satu yang menjadi perhatian umum dan penting selama persiapan data adalah untuk berhati-hati terhadap apa yang disebut "leaks" (Kaufman et al. 2012). Leaks adalah situasi di mana variabel yang dikumpulkan dalam data historis memberikan informasi tentang variabel target; namun ketika diterapkan pada kondisi sebenarnya variable ini tidak tersedia ketika keputusan harus dibuat. Sebagai contoh, ketika memprediksi apakah pada titik waktu tertentu pengunjung situs web akan mengakhiri sesinya atau melanjutkan menjelajah ke halaman lain, variabel "jumlah total halaman web yang dikunjungi dalam sesi" bersifat prediktif. Namun,

jumlah total halaman web yang dikunjungi dalam sesi tidak akan diketahui sampai setelah sesi berakhir (Kohavi et al., 2000); pada titik mana orang akan tahu nilai untuk variabel target! Sebagai contoh ilustratif lainnya, pandang pekerjaan memprediksi apakah seorang pelanggan berperilaku “boros" dalam belanja; mengetahui kategori barang yang dibeli (atau lebih buruk, jumlah pajak yang dibayar) sangat prediktif, tetapi tidak diketahui pada waktu pengambilan keputusan (Kohavi & Parekh, 2003). Leaks harus dipertimbangkan dengan hati-hati selama persiapan data, karena persiapan data biasanya dilakukan setelah fakta; dari data historis.

Modeling

Pemodelan merupakan topik yang akan dibahas pada sesi sesi selanjutnya. Namun perlu dipahami bahwa output dari pemodelan adalah model atau pola yang menangkap keteraturan dalam data.

Tahap pemodelan adalah fase utama di mana teknik data mining diterapkan pada data. Penting untuk memiliki pemahaman tentang ide-ide dasar penambangan data, termasuk jenis teknik dan algoritma yang ada, karena ini adalah bagian dari seni di mana sebagian besar ilmu pengetahuan dan teknologi diracik.

Evaluation

Tujuan tahap ini adalah untuk menilai hasil penambangan data secara ketat dan untuk mendapatkan keyakinan bahwa kesimpulan yang dihasilkan valid dan dapat diandalkan sebelum melanjutkan. Jika kita mengamati secara detil pada set data apa pun, kita akan menemukan pola, tetapi pola ini mungkin tidak dapat digeneralisasi. Kita ingin memiliki keyakinan bahwa model dan pola yang ditemukan dari data adalah keteraturan yang benar dan bukan hanya berasal dari sampel yang anomali. Sangatlah mungkin untuk memanfaatkan hasil segera setelah penambangan data, tetapi ini tidak disarankan. Biasanya cukup mudah, murah, cepat, dan lebih aman untuk menguji model dalam “laboratory setting”

Sesuatu yang sama pentingnya, adalah bahwa tahap ini juga berfungsi untuk membantu memastikan bahwa model memenuhi tujuan bisnis. Ingat bahwa tujuan utama dari ilmu data untuk bisnis adalah untuk mendukung pengambilan keputusan, dan bahwa kita memulai proses dengan berfokus pada masalah bisnis yang ingin kita pecahkan. Biasanya solusi penambangan data hanyalah bagian dari solusi yang lebih besar, dan karenanya perlu dievaluasi dalam kerangka itu. Lebih lanjut, bahkan jika model melewati tes evaluasi ketat di “di lab,” mungkin ada pertimbangan eksternal yang membuatnya tidak praktis. Misalnya, suatu kekurangan yang umum dengan solusi deteksi (seperti deteksi penipuan, deteksi spam, dan pengawasan intrusi)

adalah bahwa solusi tersebut menghasilkan terlalu banyak alarm palsu. Sebuah model mungkin sangat akurat (> 99%) menurut standar laboratorium, tetapi evaluasi dalam konteks bisnis yang sebenarnya dapat mengungkapkan bahwa masih menghasilkan terlalu banyak alarm palsu yang tidak layak digunakan dari sisi ekonomis. (Berapa biaya untuk menyediakan staf untuk menangani semua alarm palsu? Berapa biaya ketidakpuasan pelanggan?)

Mengevaluasi hasil penambangan data mencakup penilaian kuantitatif dan kualitatif. Berbagai pemangku kepentingan memiliki kepentingan dalam pengambilan keputusan bisnis yang akan dicapai atau didukung oleh model yang dihasilkan. Dalam banyak kasus, para pemangku kepentingan ini perlu "menandatangani" penerapan model, dan untuk melakukannya harus puas dengan kualitas keputusan model. Ini bervariasi dari aplikasi ke aplikasi, tetapi sering kali para pemangku kepentingan mencari untuk melihat apakah model akan memberikan hasil yang lebih baik daripada bahaya yang didatangkan, dan terutama bahwa model tersebut tidak mungkin membuat kesalahan besar. Untuk memfasilitasi penilaian kualitatif semacam itu, ilmuwan data harus berpikir tentang komprehensibilitas model kepada para pemangku kepentingan (tidak hanya untuk para ilmuwan data). Dan jika model itu sendiri tidak dapat dipahami (misalnya, mungkin modelnya adalah rumus matematika yang sangat rumit), adalah tugas para ilmuwan data untuk menyajikan perilaku model agar lebih dapat dipahami.

Akhirnya, evaluasi yang komprehensif menjadi sangat penting karena mendapatkan informasi rinci tentang kinerja model yang telah terlanjur diterapkan mungkin sulit atau tidak mungkin. Seringkali hanya ada akses terbatas ke lingkungan penyebaran sehingga membuat evaluasi komprehensif "dalam produksi" itu sulit. Sistem yang digunakan biasanya berisi banyak "bagian yang bergerak," dan menilai kontribusi dari satu bagian itu sulit. Perusahaan dengan tim data sains yang canggih perlus secara bijaksana membangun lingkungan uji coba yang mencerminkan data produksi sedekat mungkin, untuk mendapatkan evaluasi paling realistis sebelum mengambil risiko dalam deployment.

Meskipun demikian, dalam beberapa kasus kita mungkin ingin memperluas proses evaluasi ke tahap pengembangan, misalnya dengan memfasilitasi system agar dapat melakukan eksperimen acak. Dalam contoh churn, jika kita telah memutuskan dari uji laboratorium bahwa data model yang ditambang akan memberikan pengurangan churn yang lebih baik, kita mungkin ingin melanjutkan ke evaluasi "in vivo", di mana sistem secara acak menerapkan model tersebut kepada beberapa pelanggan sambil mempertahankan pelanggan lain sebagai kelompok control. Eksperimen semacam itu harus dirancang dengan hati-hati Kita mungkin juga ingin menerapkan sistem yang diterapkan untuk evaluasi guna memastikan bahwa dunia tidak berubah hingga merugikan pengambilan keputusan model. Misalnya, perilaku dapat berubah, dalam beberapa kasus, seperti penipuan atau spam, dalam tanggapan langsung terhadap penyebaran model. Selain itu, output dari model sangat bergantung pada data input; data input dapat berubah dalam format dan substansi, sering tanpa peringatan dari tim data sains. Raeder dkk. (2012) menyajikan

diskusi terperinci tentang desain sistem untuk membantu menangani masalah-masalah terkait evaluasi.

Deployment

Dalam menerapkan hasil penambangan data, dan teknik penambangan data itu sendiri, tujuannya adalah untuk mewujudkan ROI. Kasus deployment yang paling jelas melibatkan penerapan model prediktif dalam beberapa sistem informasi atau proses bisnis. Dalam contoh churn, model untuk memprediksi kemungkinan churn dapat diintegrasikan dengan proses bisnis untuk manajemen churn, misalnya, dengan mengirimkan penawaran khusus kepada pelanggan yang diprediksi menjadi sangat berisiko. Model deteksi penipuan baru dapat dibangun ke dalam sistem informasi manajemen tenaga kerja, untuk memantau akun dan membuat "kasus" bagi analis penipuan untuk memeriksa.

Terjadi kecenderungan dimana semakin banyak, teknik penambangan data itu sendiri diterapkan. Misalnya, untuk menargetkan iklan daring, sistem yang diterapkan secara otomatis membuat (dan menguji) model dalam produksi saat kampanye iklan baru ditampilkan. Dua alasan utama untuk menerapkan sistem penambangan data itu sendiri daripada model yang dihasilkan oleh sistem penambangan data adalah (i) dunia dapat berubah lebih cepat daripada tim sains data dapat beradaptasi, seperti dengan penipuan dan deteksi intrusi, dan (ii) bisnis memiliki terlalu banyak tugas pemodelan untuk tim sains data mereka untuk secara manual mengkurasi masing-masing model secara individual. Dalam kasus ini, mungkin yang terbaik adalah menerapkan fase penambangan data ke dalam produksi. Dengan demikian, sangat penting untuk instrumen proses untuk mengingatkan tim sains data dari setiap tampak anomali dan untuk menyediakan operasi (Raeder et al., 2012).

Tahap penerapan bersifat kurang “teknis.” Dalam kasus yang terkenal, penambangan data menemukan seperangkat aturan yang dapat membantu mendiagnosis dan memperbaiki kesalahan umum dalam pencetakan industri dengan cepat. Penerapan dapat berhasil hanya dengan merekam selembar kertas yang berisi aturan ke sisi printer (Evans & Fisher, 2002). Penerapan juga bisa lebih halus, seperti perubahan pada prosedur akuisisi data, atau perubahan pada strategi, pemasaran, atau operasi yang dihasilkan dari wawasan yang diperoleh dari penambangan data.

Menerapkan suatu model ke dalam sistem produksi biasanya mengharuskan model tersebut disesuikan dengan lingkungan produksi, biasanya untuk kecepatan yang lebih besar atau kompatibilitas dengan sistem yang ada. Ini mungkin menimbulkan biaya besar dan investasi. Dalam banyak kasus, tim data sains bertanggung jawab untuk memproduksi prototipe kerangka kerja, bersama dengan evaluasinya. Ini diteruskan ke tim pengembangan.

Terlepas dari apakah proses deployment berhasil, proses sering kembali ke fase Pemahaman Bisnis. Proses data mining menghasilkan banyak pemahaman terhadap masalah bisnis dan kesulitan solusinya. Iterasi kedua dapat menghasilkan solusi yang lebih baik. Hanya pengalaman berpikir tentang bisnis, data, dan tujuan kinerja sering mengarah pada ide-ide baru untuk meningkatkan kinerja bisnis, dan bahkan lini bisnis baru atau usaha baru.

Perhatikan bahwa tidak perlu menunggu kegagalan dalam tahap deployment untuk memulai satu siklus baru. Tahap Evaluasi dapat mengungkapkan bahwa hasil tidak cukup bagus untuk diterapkan, dan kita perlu menyesuaikan definisi masalah atau mendapatkan data yang berbeda. Ini diwakili oleh tautan "pintas" dari Evaluasi kembali ke Pemahaman Bisnis dalam diagram proses. Dalam praktiknya, harus ada jalan pintas kembali dari setiap tahap ke tahap sebelumnya karena proses tersebut selalu mempertahankan beberapa aspek eksplorasi, dan proyek harus cukup fleksibel untuk meninjau kembali langkah-langkah sebelumnya berdasarkan penemuan dalam suatu tahap.

Mengelola Tim Data Sains

Biasanya sangat menggoda, tetapi merupakan kesalahan besar, untuk melihat proses penambangan data sebagai siklus pengembangan perangkat lunak. Memang, proyek-proyek penambangan data sering diperlakukan dan dikelola sebagai proyek rekayasa, yang dapat dimengerti ketika mereka diprakarsai oleh departemen perangkat lunak, dengan data yang dihasilkan oleh sistem perangkat lunak dan hasil analisis dikembalikan ke dalamnya. Manajer biasanya akrab dengan teknologi perangkat lunak dan merasa nyaman mengelola proyek perangkat lunak. Pentahapan dapat saja disepakati dan keberhasilan biasanya tidak membingngkan. Manajer perangkat lunak mungkin melihat siklus penambangan data CRISP dan menganggapnya mirip dengan siklus pengembangan perangkat lunak secara umum, sehingga bagi mereka mengelola proyek data analisis dapat dilakukan dengan cara yang sama.

Ini merupakan kesalahan karena penambangan data adalah eksplorasi yang lebih mirip dengan tugas meneliti daripada rekayasa. Siklus CRISP didasarkan pada eksplorasi; melakukan iterasi pada pendekatan dan strategi daripada desain perangkat lunak. Merancang solusi penambangan data secara langsung untuk deployment dapat menjadi komitmen prematur yang mahal. Sebaliknya, proyek analitik harus bersiap untuk berinvestasi dalam informasi untuk mengurangi ketidakpastian dalam berbagai cara. Investasi kecil dapat dilakukan melalui studi pilot dan prototipe. Para ilmuwan data harus meninjau literatur untuk melihat apa lagi yang telah dilakukan dan bagaimana cara kerjanya. Pada skala yang lebih besar, tim dapat berinvestasi secara substansial dalam membangun lab uji eksperimental untuk memungkinkan eksperimen

yang lebih luas. Jika Anda seorang manajer perangkat lunak, ini akan terlihat lebih seperti penelitian dan eksplorasi daripada biasanya, dan mungkin lebih dari yang Anda pikirkan.

Kesimpulan

Proses penambangan data adalah seni. Seperti banyak kesenian yang lain, ada proses yang terdefinisi dengan baik yang dapat membantu meningkatkan kemungkinan sukses. Proses ini adalah alat konseptual yang krusial untuk berpikir tentang proyek sains data. Kita merujuk kembali ke proses penambangan data berulang kali di sepanjang kuliah ini, menunjukkan bagaimana ketepatan masing-masing konsep dasar. Pada gilirannya, memahami dasar-dasar ilmu data secara substansial meningkatkan kemungkinan berhasil.

Berbagai bidang studi yang terkait dengan ilmu data telah mengembangkan satu set jenis tugas kanonik, seperti klasifikasi, regresi, dan pengelompokan. Setiap jenis tugas melayani tujuan yang berbeda dan memiliki serangkaian teknik solusi terkait. Seorang ilmuwan data biasanya menyerang proyek baru dengan menguraikannya sedemikian rupa sehingga satu atau lebih dari tugas kanonik ini terungkap, memilih teknik solusi untuk masing-masing, kemudian menyusun solusi. Melakukan hal ini secara ahli mungkin membutuhkan banyak pengalaman dan keterampilan. Proyek penambangan data yang sukses melibatkan kompromi cerdas antara apa yang dapat dilakukan oleh data (yaitu, apa yang dapat mereka prediksi, dan seberapa baik) dan sasaran proyek. Karena alasan ini, penting untuk diingat bagaimana hasil penambangan data akan digunakan, dan gunakan ini untuk menginformasikan proses penambangan data itu sendiri.

Data mining berbeda dari, dan saling melengkapi dengan teknologi pendukung lainnya seperti pengujian hipotesis statistik dan query database. Meskipun batas antara penambangan data dan teknik terkait tidak selalu tajam, penting untuk mengetahui tentang kemampuan dan kekuatan teknik lain untuk mengetahui kapan mereka harus digunakan.

Bagi seorang manajer bisnis, proses penambangan data berguna sebagai kerangka kerja untuk menganalisis proyek atau proposal penambangan data. Proses ini menyediakan pengorganisasian yang sistematis, termasuk serangkaian pertanyaan yang dapat ditanyakan tentang proyek atau proyek yang diusulkan untuk membantu memahami apakah proyek tersebut dipahami dengan baik atau secara fundamental cacat.

DAFTAR PUSTAKA

1. Foster Provost & Tom Fawcett (2013) Data Science for Business: What you need to know about data mining and data analytic thinking, O'Reilly, ISBN: 978-1-449-36132-7.

2. Sharda, R., Delen, D., Turban, E., (2018). Business intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson.