BINUS UNIVERSITY ONLINE LEARNING

People Innovation Excellence

LECTURE NOTES

ISYS8036 - Business Intelligent and Analytics

Topic 2

Business Problem and Data Science Solutions

LEARNING OUTCOMES

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

1. Tujuan umum dari mata kuliah
2. Literatur Utama yang digunakan
3. Software pendukung
4. Decision support system
5. Pengertian Business Intelligence, Business Analytics, , Data science and data mining

OUTLINE MATERI :

1. Masalah Bisnis dan Solusi Data Sains
2. Pendekatan "Supervised" Versus "Unsupervised"
3. Data Analytical Thinking
4. Data Mining dan Hasilnya
5. Proses Data Mining
6. Kesimpulan

ISI MATERI

MASALAH BISNIS DAN SOLUSI DATA SAINS

Setiap masalah pengambilan keputusan bisnis berbasis data adalah unik, memiliki kombinasi tujuan, keinginan, batasan, dan kepribadiannya sendiri. Seperti banyak teknik, ada beberapa masalah umum yang mendasari masalah bisnis. Bekerja sama dengan pemangku kepentingan bisnis, ilmuwan data menguraikan masalah bisnis menjadi lebih spesifik. Solusi untuk subtugas kemudian dapat disusun untuk memecahkan keseluruhan masalah. Beberapa subtugas ini unik untuk masalah bisnis tertentu, namun ada beberapa tugas penambangan data yang umum. Misalnya, masalah churn telekomunikasi kita unik untuk MegaTelCo: ada beberapa masalah spesifik yang berbeda dari masalah churn perusahaan telekomunikasi lainnya. Namun, subtugas yang kemungkinan akan menjadi bagian dari solusi untuk masalah churn adalah memperkirakan dari data historis kemungkinan pelanggan yang menghentikan kontraknya tidak lama setelah masa berlakunya habis. Begitu data MegaTelCo yang khas telah diramu menjadi format tertentu (dijelaskan di sesi berikutnya), estimasi probabilitas ini sesuai dengan format tugas penambangan data yang sangat umum.

Keterampilan kritis dalam ilmu data adalah kemampuan untuk menguraikan masalah analisis data menjadi beberapa bagian sehingga masing-masing sesuai dengan tugas yang diketahui solusinya. Mengenali masalah yang sudah dikenal dan solusi mereka menghindari pemborosan waktu dan sumber daya. Hal ini juga memungkinkan orang untuk memusatkan perhatian pada bagian-bagian yang lebih menarik dari proses yang memerlukan bagian keterlibatan manusia yang belum otomatis, sehingga kreativitas dan kecerdasan manusia harus ikut dilibatkan.

Meskipun sejumlah besar algoritma data mining telah dikembangkan selama bertahun-tahun, hanya ada sedikit konsep fundamental yang dipakai oleh algoritma algoritma ini. Akan didefinisikan konsep konsep ini dengan jelas. Dalam beberapa sesi selanjutnya akan diuraikan dua konsep yang pertama yaitu klasifikasi dan regresi untuk menggambarkan beberapa konsep dasar.

Istilah "individu/ instans" akan mengacu pada entitas yang memiliki data, seperti pelanggan atau konsumen, atau mungkin entitas tak bernyawa seperti bisnis. Dalam banyak proyek analisis bisnis, ingin ditemukan "korelasi" antara variabel tertentu yang menggambarkan individu dan variabel lainnya. Misalnya, dalam data historis, mungkin akan dapat diketahui pelanggan mana yang meninggalkan perusahaan setelah kontrak mereka berakhir. Selain itu mungkin ingin

diketahui variabel lain yang berkorelasi dengan pelanggan yang tidak beralih dalm waktu dekat. Menemukan korelasi semacam itu adalah contoh paling dasar dari tugas klasifikasi dan regresi.

Klasifikasi dan pendugaan probabilitas kelas mencoba memprediksi, untuk setiap individu dalam suatu populasi, termasuk dalam kelas yang mana. Biasanya kelas saling eksklusif. Contoh pertanyaan klasifikasi adalah: "Di antara semua pelanggan MegaTelCo, pelanggan manakah yang cenderung menanggapi tawaran yang diberikan?" Dalam contoh ini, dua kelas dapat dipisahkan sebagai yang akan merespons dan tidak akan merespons. Untuk tugas klasifikasi, prosedur data mining menghasilkan sebuah model yang, jika diberikan individu baru, model menentukan kelas mana individu itu berada. Tugas yang terkait erat adalah penilaian probabilitas atau penilaian kelas. Model penilaian yang diterapkan pada individu menghasilkan, bukan prediksi kelas, melainkan skor yang mengatakan probabilitas (atau beberapa kuantifikasi kemungkinan lainnya) bahwa individu tersebut termasuk dalam suatu kelas. Dalam skenario respons pelanggan, model dapat mengevaluasi setiap pelanggan dan menghasilkan skor seberapa besar kemungkinan masing-masing untuk merespons penawaran tersebut.

Regresi (estimasi nilai) mencoba untuk memperkirakan atau memprediksi, untuk setiap individu, nilai numerik dari suatu variabel untuk individu tersebut. Contoh pertanyaan regresi adalah: "Berapa banyakkah seorang pelanggan akan menggunakan layanan ini?" Properti (variabel) yang diprediksi di sini adalah penggunaan layanan, dan suatu model dapat dihasilkan dengan melihat individu lain yang serupa dalam populasi dan historis penggunaan layanan dari individu tersebut. Prosedur regresi menghasilkan sebuah model yang, jika diberikan seorang individu, model memperkirakan nilai variabel khusus yang spesifik untuk individu tersebut. Regresi berhubungan dengan klasifikasi, namun keduanya agak berbeda. Secara informal, klasifikasi memprediksi apakah sesuatu akan terjadi, sedangkan regresi memprediksi berapa banyak sesuatu akan terjadi.

Pencocokan kesamaan berupaya mengidentifikasi individu yang sama berdasarkan data yang diketahui tentang mereka. Pencocokan kesamaan dapat digunakan secara langsung untuk menemukan entitas serupa. Misalnya, IBM tertarik untuk menemukan perusahaan yang serupa dengan pelanggan bisnis terbaik mereka, untuk memusatkan tenaga penjualan mereka pada peluang terbaik. Mereka menggunakan pencocokan kesamaan berdasarkan data "firmographic" yang menggambarkan karakteristik perusahaan. Pencocokan kesamaan adalah dasar untuk salah satu metode yang paling populer untuk membuat rekomendasi produk (menemukan orang-orang yang serupa dengan Anda dalam hal produk yang mereka sukai atau yang telah mereka beli). Ukuran kesamaan menjadi dasar bagi beberapa solusi tertentu dalam data mining seperti klasifikasi, regresi, dan clustering (pengelompokan).

Clustering mencoba mengelompokkan individu dalam suatu populasi berdasarkan kesamaan mereka, namun tidak didorong oleh tujuan tertentu. Contoh pertanyaan clustering adalah: "Apakah pelanggan kita membentuk kelompok atau segmen alami?" Clustering berguna dalam

eksplorasi domain awal untuk melihat kelompok alami mana yang ada karena kelompok-kelompok ini yang pada gilirannya dapat menyarankan tugas atau pendekatan penambangan data lainnya. Clustering juga digunakan sebagai masukan untuk proses pengambilan keputusan yang

berfokus pada pertanyaan seperti: Produk apa yang harus kita tawarkan atau kembangkan? Bagaimana seharusnya tim layanan pelanggan (atau tim penjualan) terstruktur?

Co-occurrence grouping (juga dikenal sebagai frequent item set mining, association rule discovery, dan market-basket analysis) mencoba menemukan asosiasi antar entitas berdasarkan transaksi yang melibatkan mereka. Contoh pertanyaan co-occurrence adalah: barang apa yang biasa dibeli bersama? Sementara pengelompokan terlihat pada kesamaan antara objek berdasarkan atribut objek, pengelompokan bersama menunjukkan kesamaan objek berdasarkan kemunculannya bersama dalam transaksi. Misalnya, menganalisis catatan pembelian dari supermarket dapat mengungkap bahwa daging dibeli bersama dengan saus tertentu melampaui yang sering diduga. Memutuskan bagaimana bertindak atas penemuan ini mungkin memerlukan beberapa kreativitas, namun bisa menyarankan promosi khusus, tampilan produk, atau penawaran kombinasi. Kemunculan produk dalam pembelian adalah jenis pengelompokan umum yang dikenal dengan market basket analysis. Beberapa sistem rekomendasi juga melakukan jenis pengelompokan dengan menemukan, misalnya, pasangan buku yang sering dibeli oleh orang yang sama ("orang-orang yang membeli X juga membeli Y"). Hasil pengelompokan co-occurrence adalah deskripsi item yang terjadi bersamaan. Deskripsi ini biasanya mencakup statistik tentang frekuensi kejadian bersama dan perkiraan probabilitasnya.

Profiling (juga dikenal sebagai deskripsi perilaku) mencoba untuk mengkarakterisasi perilaku khas individu, kelompok, atau populasi. Contoh pertanyaan profiling adalah: "Apa jenis ponsel yang disukai segmen pelanggan ini?" Perilaku mungkin tidak memiliki deskripsi sederhana; Profiling pengguna telepon mungkin memerlukan deskripsi kompleks tentang rata-rata *airtime* malam dan akhir pekan, penggunaan sambungan internasional, biaya *roaming*, dan sebagainya. Perilaku dapat digambarkan secara umum di seluruh populasi, atau sampai ke tingkat kelompok kecil atau bahkan individu.

Profiling sering digunakan untuk menetapkan norma perilaku untuk aplikasi pendeteksian anomali seperti deteksi kecurangan dan pemantauan intrusi ke suatu sistem komputer (seperti seseorang yang masuk ke akun iTunes Anda). Misalnya, jika kita mengetahui jenis pembelian yang sering dilakukan seseorang dengan kartu kredit, kita mungkin dapat menentukan apakah tagihan baru pada kartu tersebut sesuai dengan profil pembelinya. Kita bisa menggunakan tingkat ketidakcocokan sebagai nilai kecurigaan dan mengeluarkan alarm jika ketidak cocokan sangat mencolok.

Prediksi tautan (link prediction) mencoba memprediksi hubungan antara item data, biasanya dengan menyarankan existensi tautan(link), dan mungkin juga memperkirakan bobot tautan. Prediksi link umum terjadi pada sistem jejaring sosial: "Karena Anda dan Ali memiliki 10 teman

yang sama, mungkin Anda ingin menjadi teman Ali" Prediksi tautan juga dapat memperkirakan kekuatan tautan. Misalnya, untuk merekomendasikan film kepada pelanggan, seseorang dapat menggunakan graf antara pelanggan dan film yang mereka tonton atau beri nilai. Dalam graf, dicari link yang tidak ada antara pelanggan dan film, tapi yang seharusnya ada dan cukup kuat. Tautan ini membentuk dasar rekomendasi.

Pengurangan data (data reduction) mencoba menangani sekumpulan data besar dan menggantinya dengan kumpulan data dengan jumlah lebih kecil yang berisi informasi penting. Dataset yang lebih kecil mungkin lebih mudah ditangani atau diproses. Selain itu, dataset yang lebih kecil mungkin lebih baik dalam mengungkapkan informasi tersembunyi. Misalnya, kumpulan data yang besar pada preferensi menonton film konsumen dapat dikurangi ke kumpulan data yang jauh lebih kecil yang mengungkapkan preferensi rasa konsumen laten dalam data tampilan (misalnya, preferensi *genre* pemirsa). Pengurangan data biasanya melibatkan hilangnya informasi. Namun yang penting adalah *trade off* untuk meningkatkan pemahaman terhadap perilaku konsumer.

Pemodelan kausal (causal modelling) mencoba untuk membantu kita memahami kejadian atau tindakan apa yang mempengaruhi kejadian lain. Misalnya, perhatikan bahwa pemodelan prediktif untuk menargetkan iklan kepada konsumen, dan kita amati bahwa memang konsumen yang ditargetkan membeli pada tingkat yang lebih tinggi setelah menjadi sasaran. Apakah ini karena iklan tersebut mempengaruhi konsumen untuk membeli? Atau apakah model prediktif hanya melakukan pekerjaan dengan baik untuk mengidentifikasi konsumen yang pasti akan membeli? Teknik pemodelan kausal termasuk melibatkan investasi besar dalam data, seperti eksperimen terkontrol secara acak (misalnya, apa yang disebut "tes A / B"), serta metode yang canggih untuk menarik kesimpulan kausal dari data pengamatan. Metode eksperimental dan observasional untuk pemodelan kausal umumnya dapat dipandang sebagai analisis "kontrafaktual": mereka berusaha memahami apa perbedaan antara situasi yang tidak dapat terjadi baik di mana peristiwa "perlakuan" (misalnya, menunjukkan iklan kepada individu tertentu) terjadi, dan tidak akan terjadi.

Dalam semua kasus, ilmuwan data yang hati-hati harus selalu menyertakan kesimpulan kausal asumsi pasti yang harus dilakukan agar kesimpulan kausal dapat ditahan (selalu ada asumsi seperti itu yang selalu diajukan). Ketika melakukan pemodelan kausal, bisnis perlu menimbang trade off investasi yang meningkat untuk mengurangi asumsi yang dibuat, versus memutuskan bahwa kesimpulan tersebut cukup baik mengingat asumsi tersebut. Bahkan dalam eksperimen acak terkontrol yang paling hati-hati, asumsi dibuat yang dapat membuat kesimpulan kausal tidak valid. Penemuan "efek plasebo" dalam kedokteran menggambarkan situasi yang tidak terkenal dimana asumsi diabaikan dalam eksperimen acak yang dirancang dengan cermat.

Dalam kuliah ini, akan disajikan kumpulan prinsip data sains yang paling mendasar, prinsip-prinsip yang bersama-sama mendasari semua jenis tugas ini. Prinsip-prinsip yang dibahas terutama dengan menggunakan klasifikasi, regresi, kesamaan pencocokan, dan pengelompokan,

dan akan membahas hal-hal lain saat memberikan ilustrasi penting dari prinsip-prinsip dasar (menjelang akhir kuliah).

Kembali ke jenis tugas yang sesuai dengan masalah prediksi churn. Seringkali, para praktisi merumuskan prediksi churn sebagai masalah dalam menemukan segmen pelanggan yang cenderung beralih ke provider lain (churn). Masalah segmentasi ini terdengar seperti masalah klasifikasi, atau mungkin clustering, atau bahkan regresi. Untuk menentukan rumusan terbaik, pertama perlu diperkenalkan beberapa perbedaan penting.

Metode “Supervised” Versus “Unsupervised”

Dua pertanyaan serupa yang mungkin muncul tentang pelanggan. Yang pertama adalah: "Apakah pelanggan kita secara alami jatuh ke dalam kelompok yang berbeda?" Di sini tidak ada tujuan atau target khusus yang telah ditentukan untuk pengelompokan tersebut. Bila tidak ada target seperti itu, masalah data mining disebut tidak diawasi (unsupervised). Bandingkan dengan pertanyaan yang sedikit berbeda: "Bisakah kita menemukan kelompok pelanggan yang memiliki kemungkinan sangat tinggi untuk beralih ke provider yang lain segera setelah kontrak mereka berakhir?" Di sini ada target spesifik yang ditetapkan: akankah pelanggan beralih? Dalam kasus ini, segmentasi dilakukan dengan alasan tertentu: melakukan tindakan berdasarkan kemungkinan churn. Ini disebut masalah data mining yang diawasi (supervised).

Istilah supervised dan unsupervised diwarisi dari bidang pembelajaran mesin. Secara metaforis, seorang guru "mengawasi" pelajar dengan memberikan informasi target melalui sejumlah contoh. Belajar tanpa pengawasan mungkin melibatkan serangkaian contoh yang sama namun tidak memasukkan informasi target. Pembelajar tidak diberi informasi tentang tujuan pembelajaran, namun dibiarkan untuk membuat kesimpulan sendiri berdasarkan kesamaan antar sampel.

Perbedaan antara pertanyaan-pertanyaan ini tidaklah kelihatan namun penting. Jika target tertentu diberikan, masalahnya tergolong diawasi. Masalah yang diawasi memerlukan teknik yang berbeda dari masalah yang tidak diawasi, dan hasilnya seringkali jauh lebih bermanfaat. Teknik yang diawasi diberikan tujuan khusus untuk pengelompokan, memprediksi target. Clustering, sebuah prosedur tanpa pengawasan, menghasilkan pengelompokan berdasarkan persamaan, namun tidak ada jaminan bahwa kesamaan ini bermakna atau akan berguna untuk tujuan tertentu.

Secara teknis, ada kondisi yang harus dipenuhi untuk data mining yang diawasi: yaitu harus ada data target. Tidaklah cukup bahwa informasi target tersedia secara prinsip; Target harus ada secara explisit dalam data. Misalnya, mungkin berguna untuk mengetahui apakah seorang pelanggan akan bertahan setidaknya enam bulan, namun jika dalam data historis informasi retensi ini hilang atau tidak lengkap (jika, katakanlah, data hanya disimpan selama dua bulan)

nilai target akan hilang. Memperoleh data target sering merupakan investasi kunci dalam data sains. Nilai untuk variabel target suatu individu sering disebut label individu. Seringkali (tidak selalu) seseorang harus mengeluarkan biaya untuk secara aktif memberi label data.

Klasifikasi, regresi, dan pemodelan kausal umumnya diselesaikan dengan metode yang diawasi. Pencocokan kesamaan, link prediction, dan reduksi data seringkali bisa menggunakan pendekatan ini. *Clustering, co-occurrence grouping,* dan *profiling* umumnya tidak diawasi. Prinsip dasar data mining yang akan dibahas mendasari semua jenis teknik ini.

Dua subkelas dalam klasifikasi dan regresi yang diawasi, dibedakan berdasarkan jenis targetnya. Regresi melibatkan target numerik sementara klasifikasi melibatkan target kategoris (sering biner). Perhatikan pertanyaan serupa berikut yang mungkin muncul dari data mining yang diawasi:

"Apakah pelanggan A akan menggunakan layanan S1 jika diberikan insentive I ?" Ini adalah masalah klasifikasi karena melibatkan target biner.

"Apakah pelanggan ini akan membeli layanan S1, S2 atau tidak sama sekali jika diberikan insentive I ?" Masalah ini masih tergolong klasifikasi meski melibatkan target lebih dari dua.

"Seberapa banyakkah seorang pelanggan akan menggunakan layanan ini ?" Ini adalah masalah regresi karena target bernilai numerik. Variabel target adalah jumlah pemakaian (aktual atau prediksi) per pelanggan.

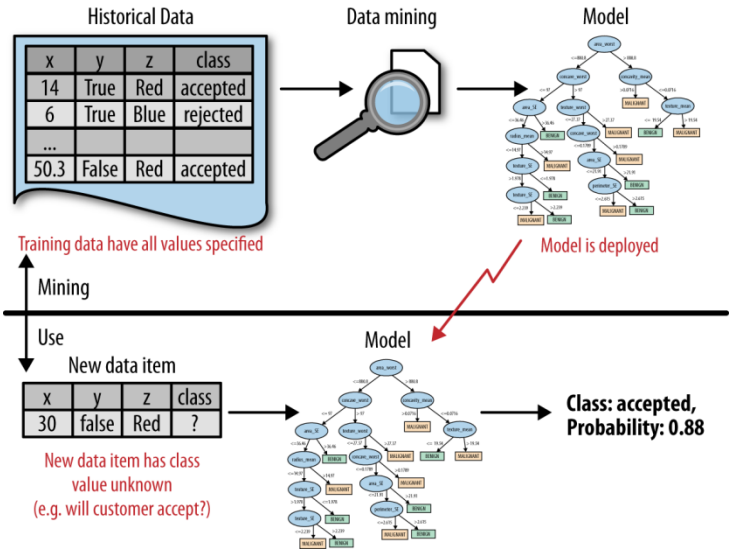
Terdapat perbedaan tipis di antara pertanyaan-pertanyaan di atas yang harus dicermati. Dalam suatu aplikasi bisnis kita sering menginginkan prediksi numerik atas target kategoris. Dalam contoh churn, prediksi dasar ya / tidak seorang pelanggan cenderung tetap, mungkin tidak mencukupi; lebih dari itu dikehendaki informasi mengenai berapa probabilitas bahwa pelanggan tersebut akan melanjutkan. Ini masih dianggap pemodelan klasifikasi bukan regresi karena target dasarnya bersifat kategoris. Inilah yang disebut dengan "estimasi probabilitas kelas".

Bagian penting dalam tahap awal proses data mining adalah (i) memutuskan apakah pendekatan yang digunakan diawasi atau tidak diawasi, dan (ii) jika diawasi, perlu definisi yang tepat terhadap variabel target. Variabel ini harus mempunyai kuantitas tertentu yang akan menjadi fokus dari proses data mining.

Data mining dan hasilnya

Ada perbedaan penting lainnya yang berkaitan dengan proses data mining: (1) Data mining untuk menemukan pola dan membangun model, dan (2) menggunakan hasil data mining.

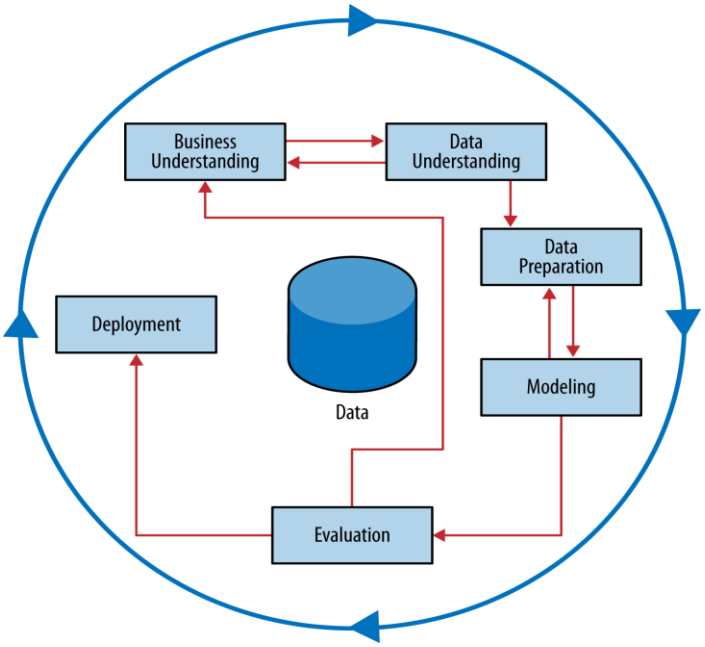
Dalam contoh churn,asumsikan bahwa data mining telah menciptakan model estimasi probabilitas kelas M. M menggunakan karakteristik sebagai input dan menghasilkan skor atau perkiraan probabilitas dari atrisi. Ini adalah penggunaan hasil data mining.



Gambar 2-1. Proses Data Mining versus penggunaan hasil Data Mining. Gambar atas memperlihatkan penambangan data historis untuk menghasilkan model. Data historis memiliki nilai target ("kelas") yang ditentukan. Gambar bawah menunjukkan hasil dari penambangan data yang digunakan, di mana model diterapkan ke data baru yang belum diketahui kelasnya. Model memprediksi nilai kelas dan probabilitas.

Proses Data Mining

Data mining adalah seni. Ini melibatkan penerapan sejumlah besar sains dan teknologi, tetapi aplikasi yang tepat perlu melibatkan seni. Tetapi seperti banyak produk seni yang matang, ada proses yang perlu dipahami dengan baik. Sebuah kodifikasi yang berguna dari proses data mining diberikan oleh Proses Standar Industri Cross untuk Penambangan Data (CRISP-DM; Shearer, 2000), yang diilustrasikan pada Gambar 2-2.



Gambar 2-2. Proses penambangan data CRISP.

Diagram proses ini membuat eksplisit fakta bahwa iterasi adalah sebuah prinsip dasar. Dalam sekali iterasi tanpa memecahkan masalah, secara umum, bukan merupakan kegagalan. Berikut akan dijelaskan secara singkat proses utam dalam diagram. Pada pertemuan sesi berikut akan dijelaskan secara detil setiap tahapan dalam proses tersebut.

KESIMPULAN

Terdapat banyak bukti yang meyakinkan bahwa pengambilan keputusan berbasis data dan teknologi big data secara substansial memperbaiki kinerja bisnis. Ilmu data (data sains) mendukung pengambilan keputusan berbasis data dan terkadang melakukan pengambilan keputusan secara otomatis dan bergantung pada teknologi penyimpanan dan rekayasa big data yang digunakan, namun prinsipnya terpisah. Prinsip data sains yang dibahas dalam kuliah ini juga berbeda, dan saling melengkapi dengan Teknik lain seperti pengujian hipotesis statistik dan query database. Pada sesi sesi berikut akan dijelaskan beberapa perbedaan ini secara lebih rinci.

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.