Teknik Utama Penyiapan Data untuk Data Scientist

dan Implementasinya di R

1. Peranan Penyiapan Data dalam Analitika
2. Pengenalan R
   1. Instalasi R dan Rstudio
   2. Instalasi Package
   3. Pengenalan data frame dan pengolahannya
3. Dasar-dasar Pemrograman R
4. Import Data
5. Mengenal Fungsi-Fungsi Dasar dplyr dan Operator Pipe di R
   1. Filter (memilih baris)
   2. Select (memilih kolom)
   3. Mutate (menambahkan data baru)
   4. Group\_By dan Summarize
   5. Pengenalan operator pipe
6. Bekerja dengan Satu Dataframe
   1. Filter
   2. Summary
   3. Sort
   4. Edit Data
   5. Reshaping (pivot)
7. Bekerja dengan Banyak Dataframe
   1. Konsep database relational
   2. Merge/Join

**Pendahuluan**

**1. Analitika dan Kegunaannya**

Data memiliki manfaat yang besar dalam pengembangan organisasi melalui pengambilan keputusan yang tepat dan terarah. Dalam bidang komersial, pemanfaatan data secara optimal mampu meningkatkan efisiensi dan kinerja yang berujung pada peningkatan keuntungan. Begitu berharganya data tersebut sehingga ada yang menyebut data sebagai “*new oil*” yang menggambarkan bahwa siapa yang menguasai data akan mampu menguasai perekonomian seperti halnya penguasaan terhadap minyak bumi pada beberapa dekade lalu.

Namun tentu saja, kepemilikan data tidak hanya berhenti pada proses menyimpan saja. Perlu upaya-upaya sistematis sehingga kita dapat memperoleh nilai tambah dari data yang kita miliki. Upaya ini dikenal sebagai analitika (*data analytics*).

Secara umum analitika dapat dipandang sebagai kegiatan mengubah data menjadi informasi yang berguna bagi pengambil keputusan untuk melakukan perencanaan dan penentuan tindakan yang efektif, baik untuk keperluan individu maupun organisasi. Ujung yang diharapkan dari proses analitika adalah keputusan perencanaan dan tindakan yang didasarkan pada data atau informasi dari proses mengolah data tersebut.

Berdasarkan klasifikasi tujuan dan jenis kegiatannya, proses analitika terbagi menjadi empat kategori yaitu: (1) *descriptive analytics*, (2) *diagnostic analytics*, (3) *predictive analytics*, dan (4) *prescriptive analytics*. Penjelasan singkat dari masing-masing kategori analitika tersebut dapat dipaparkan sebagai berikut.

Sesuai dengan namanya, *descriptive analytics* merupakan kegiatan untuk memaparkan kondisi yang ada berdasarkan data yang dimiliki. Pada kategori ini, upaya yang biasa dilakukan terhadap data adalah mengungkap dan memberikan informasi apa yang sedang atau telah terjadi. Misalnya dari data catatan penjualan sebuah supermarket kita bisa memperoleh informasi mengenai total penjualan per bulan selama setahun terakhir, atau tentang produk apa yang paling laku terjual pada sebulan terakhir, atau berapa rata-rata pengeluaran seorang pelanggan untuk satu kali kunjungan belanja. *Descriptive analytics* juga memuat informasi perbandingan dan pengenalan pola-pola umum pada data. Misalnya perbandingan kepadatan pengunjung supermarket pada hari kerja dan akhir pekan, atau pola kecenderungan perubahan naik-turun total pendapatan dari waktu ke waktu.

Umumnya pola-pola yang ditemukan pada saat melakukan descriptive analytics memerlukan jawaban mengenai bagaimana dan mengapa hal tersebut dapat terjadi. Jika ada penurunan angka penjualan supermarket, ingin diketahui apakah penurunan terjadi pada semua jenis komoditas dan apa penyebab terjadinya penurunan tersebut. Perusahaan pembiayaan yang memperoleh laporan adanya 2% nasabah yang macet ingin menggali lebih jauh bagaimana perbedaan karakteristik antara nababah yang bermasalah pembayarannya dan nasabah yang baik-baik saja. Hal-hal semacam ini dilakukan pada kelompok analitika yang lain yaitu *diagnostic analytics*. Pada bagian ini diharapkan ditemukan informasi lebih lanjut mengenai apa yang didapatkan dari proses pengenalan secara umum. Mengetahui informasi tambahan ini dapat bermanfaat dalam penentuan prioritas pengembangan selanjutnya.

Pengetahuan mengenai hal-hal yang terlibat dan mempengaruhi terjadinya suatu kejadian yang tergambar pada data selanjutnya dapat disusun dalam berbagai bentuk model hubungan. Model ini dapat menjadi representasi proses dengan input berbagai faktor yang telah diidentifikasi dan output berupa kejadian yang menjadi pusat perhatian. Algoritma memperoleh model ini ada yang sederhana dan mudah dipahami bentuknya, namun ada pula yang rumit dan perlu upaya ekstra untuk memahaminya. Salah satu yang utama dari penggunaan model ini adalah memprediksikan kejadian atau output dengan memasukkan informasi-informasi mengenai faktor-faktor yang menjadi input bagi model tersebut. Kategori analisis dalam memperoleh model yang demikian ini dikenal dengan *predictive analytics*. Pada proses memperoleh model untuk memprediksi tersebut terdapat beragam tahapan seperti proses penyiapan variabel prediktor, pemilihan variabel, perekayaan variabel, pembuatan model, hingga proses validasi. Tahapan tersebut umumnya dikerjakan sebagai upaya untuk memperoleh model dengan kualitas prediksi yang sangat baik.

Kategori yang keempat dalam proses analitika adalah *prescriptive analytics*, yang berfoku pada menemukan skenario tindakan terbaik berdasarkan data yang ada. Tentu saja skenario ini didasarkan pada tiga proses analitika yang lain. Berdasarkan hasil analisis secara deskriptif, diagnostik, dan prediktif, kita dapat saja dihadapkan pada beberapa pilihan solusi. Prescriptive analytics berupaya untuk mengevaluasi alternatif-alternatif tersebut dan mengidentifikasi pilihan mana yang paling baik untuk dikerjakan agar menghasilkan output yang optimal.

Guna memperjelas diskusi mengenai proses analitika, berikut ini akan dipaparkan ilustrasi dari beberapa bidang terapan.

Kita mulai dengan sebuah ilustrasi industri perkreditan baik yang dikejakan oleh perbankan maupun perusahaan pembiayaan, termasuk yang berbasis aplikasi online. Pada setiap bulan perusahaan secara rutin meng-*genarate* laporan mengenai kinerja produk kredit mereka, yang antara lain berisi informasi mengenai berapa besar *outstanding* kredit mereka yang menggambarkan seberapa besar skala usaha yang mereka miliki saat ini. Dengan membandingkan pada periode sebelumnya, perusahaan memiliki informasi apakah pada bulan tersebut ada kenaikan atau tidak. Selain nilai outstanding, juga didapatkan informasi mengenai berapa banyak kredit yang masuk dalam kategori macet atau menunggak pembayarannya. Angka-angka persentase yang menunggak dan trend-nya dari waktu ke waktu dapat disajikan pada laporan atau dashboard perusahaan. Sampai pada titik ini, itulah proses *descriptive analytics*.

Pihak manajemen terutama di bagian manajemen resiko umumnya akan tertarik untuk menelusuri siapa saja yang menunggak dan bagaimana karakteristiknya. Informasi mengenai karakteristik tersebut akan dibanding-bandingkan dengan dengan karakteristik nasabah yang pembayarannya lancar. Pada tahapan ini mungkin saja diperoleh informasi mengenai adanya kelompok nasabah yang resiko menunggaknya lebih besar dari yang lain. Misalnya saja, jika secara umum terdapat 2% nasabah yang macet, tapi ternyata pada kelompok nasabah dengan penghasilan 3 – 5 juta per bulan terdapat 5% yang macet atau dua setengah kali lebih banyak dari kondisi secara umum. Proses identifikasi seperti ini adalah bagian dari tahapan *diagnostic analytics*.

Hasil dari temuan di atas dapat menjadi alasan bahwa perusahaan memutuskan untuk menerapkan seleksi calon nasabah sebelum pembiayaan atau pinjaman diberikan. Variabel-variabel penting yang diperoleh dari proses diagnostik dapat dijadikan sebagai masukan bagi pengembangan model *approval scoring*, yaitu model yang yang digunakan untuk menentukan apakah aplikasi kredit seseorang akan disetujui atau ditolak. Model ini berbasis pada data sedemikian rupa sehingga orang-orang yang diprediksi akan memiliki resiko tinggi diberikan skor yang kecil sedangkan orang dengan resiko yang baik diberi skor tinggi. Teknik-teknik pemodelan prediktif akan diimplementasikan untuk memperoleh model dengan ketepatan yang tinggi. Tentu saja model akan diharapkan mampu membedakan beresiko dan yang tidak, sehingga yang disetujui aplikasinya adalah yang benar-benar akan mampu menunaikan kewajiban pembayarannya dengan baik dan model tidak salah identifikasi dengan menolak aplikasi dari orang-orang yang seperti itu.

Pada saat implementasi model prediktif untuk *approval scoring* tersebut, perusahaan akan dihadapkan pada pemilihan batas skor untuk memutuskan disetujui atau tidak. Memberikan batas yang terlalu besar akan membuat perusahaan hanya menerima sedikit *applicant* dan itu berarti skala bisnis akan mengecil. Sedangkan memberikan batas yang terlalu kecil akan mengakibatkan perusahaan memberikan kredit kepada orang-orang yang beresiko tinggi. *Prescrtiptive analytics* diterapkan untuk mendapatkan batas skor optimal dengan mempertimbangkan berbagai hal.

**2. Analitika: Art and Science**

Dalam melakan berbagai proses analitika, seorang *data analyst* atau *data scientist* akan dihadapkan pada banyak pilihan tahapan untuk dikerjakan. Karenanya untuk kasus yang sama untuk dipecahkan, dua orang *data scientist* bisa saja melakukan hal yang sangat berbeda namun menghasilkan keluaran yang sama-sama menarik untuk dipertimbangkan oleh pengambil keputusan.

Perbedaan proses dan hasil dari dua orang berbeda tersebut dapat disebabkan oleh berbagai hal, antara lain:

* Perbedaan cara pandang terhadap masalah yang dihadapi

Terhadap suatu permasalahan, seorang data scientist dapat saja memiliki pandangan yang berbeda dengan orang lain mengenai cara penyelesaiannya menggunakan analitika. Misalnya saja ketika dihadapkan pada persoalan upaya meningkatkan penjualan perusahaan, sangat memungkinkan untuk melihat dari berbagai sisi. Ada sebagian orang melihatnya sebagai permasalahan mencari strategi direct marketing yang tepat sehingga yang dilakukan adalah mengidentifikasi target pemasaran yang sesuai sehingga program pemasarannya dapat berjalan dengan efektif. Ada juga yang memandang sebagai permasalahan pengembangan produk yang harus lebih baik. Orang yang pertama akan melakukan analitika dengan fokus terhadap data pelanggan, sedangkan orang kedua bisa jadi akan fokus pada data produk dan transaksi. Karenanya proses analitika yang dikerjakan bisa sangatlah berbeda, dan selanjutnya akan menghasilkan rekomendasi tindakan yang berbeda pula.

* Perbedaan dalam penggunaan data

Perbedaan aktifitas yang dikerjakan dalam proses analitika juga dapat disebabkan karena adanya perbedaan data yang digunakan. Hal ini terjadi misalnya karena yang satu berpendapat harus menggunakan data 6 bulan terakhir, sedangkan orang lain menyarankan untuk menggunakan data lebih panjang yaitu setahun terakhir. Ada kemungkinan data-data yang lebih lama tersimpan pada database yang berbeda sehingga proses mengekstrak data untuk analisis akan memerlukan tahapan yang lebih lama dibandingkan menggunakan data yang lebih baru.

* Perbedaan dalam mendefinisikan variabel

Kemungkinan adanya perbedaan proses analitika juga sangat mungkin disebabkan oleh perbedaan dalam mendefinisikan variabel operasional yang digunakan. Ketika akan melakukan analisis terhadap pelanggan supermarket yang aktif, maka akan muncul perdebatan bagaimana mendefinisikan kata aktif. Jika didasarkan pada kapan terakhir menggunakan kartu anggota dalam melakukan pembelian, berapa batas waktu yang digunakan untuk mengkelaskan pelanggan aktif.

* Perbedaan dalam penangangan kualitas data

Proses analitika juga dapat berbeda dari satu orang ke orang yang lain karena perbedaan cara menangani masalah dengan kualitas data. Sangat mungkin pada data yang kita miliki ada nilai-nilai yang kosong (missing value). Bagaimana menanganinya? Ada yang berpendapat untuk menyisihkan saja amatan yang memuat nilai kosong tersebut. Namun ada pula yang yang mengusulkan untuk melakukan pengisian data kosong itu dengan cara imputasi sehingga tidak perlu menyisihkan amatan. Sebagian yang lain tetap menggunakan amatan dengan nilai kosong dan memandangnya sebagai nilai tersendiri untuk data hilang/kosong.

* Perbedaan dalam pemilihan algoritma analisis dan pemodelan

Kemungkinan lain yang jadi penyebab perbedaan proses analitika yang dilakukan oleh para data scientist dalam menghadapi permasalahan yang sama adalah penggunaan algoritma analisis dan pemodelan yang berbeda. Untuk tujuan yang sama, terdapat berbagai macam algoritma yang dapat dipilih oleh para pengguna. Misalnya saja pada pemodelan prediktif klasifikasi, seorang data scientist dapat menggunakan teknik regresi logistik, pohon klasifikasi sederhana, hingga penerapan teknik ensemble seperti boosting dan random forest untuk kasus yang sama. Seseorang bahkan dapat mengimplementasikan banyak algoritma pemodelan untuk satu kasus dan kemudian memilih mana yang memberikan hasil terbaik sesuai dengan kriteria yang dia jadikan acuan.

Dengan melihat berbagai hal di atas, banyang orang yang menilai bahwa analitika merupakan gabungan antara art dan science. Tidak hanya aspek kemampuan teknis programming dan pengetahuan statistika yang diperlukan, namun juga diperlukan pemahaman terhadap konteks permasalahan, pengetahuan proses bisnis yang berkaitan dengan data, hingga kepekaan dalam mengamati pola pada data.

**3. Peranan Penyiapan Data dalam Analitika**

Keberhasilan dari proses analitika atau analisis data secara keseluruhan sangat bergantung pada input yang kita gunakan. Input utama dari proses ini adalah data. Data yang berkualitas menjadi syarat perlu bagi diperolehnya hasil yang baik. Tidak hanya itu, data yang digunakan haruslah diperoleh dengan kerangka pemikiran analitika yang terstruktur sehingga interpretasi terhadap hasil analisis dapat dengan mudah dilakukan dan dikomunikasikan kepada pihak lain.

Tahapan penyiapan data merupakan tahapan awal dalam proses analitika dan dapat menentukan kualitas hasil akhir. Dalam tahapan ini kita diharapkan menghasilkan sebuah tabel data yang siap untuk analisis dan pemodelan. Tabel data tersebut hendaknya memuat berbagai variabel yang diperlukan untuk menghasilkan informasi yang dijadikan input analisis, memuat variabel-variabel utama dan tambahan yang penting dalam pemodelan, serta memiliki struktur yang sesuai dengan kebutuhan analisis dan pemodelan tersebut.

Banyak hal yang dapat disebutkan terkait dengan proses penyiapan data yang dapat menentukan keberhasilan proses analitika. Hal-hal tersebut antara lain adalah:

* Sumber data

Penggunaan sumber data yang tepat akan menentukan ketepatan informasi dihasilkan dan rekomendasi yang disarankan. Sebelum memulai tahapan penyiapan data, seorang data scientist harus memahami konten dari berbagai tabel yang ada pada database sumber data mereka. Pengetahuan mengenai variabel apa saja yang ada dalam data, bagaimana nilai-nilai data pada database diperoleh, dan kurun waktu berlakunya data tersebut merupakan hal-hal dasar yang harus dipahami oleh data scientist tersebut sejak awal. Diskusi dengan teman sejawat yang terlibat banyak dalam pengembangan database dan memahami proses bisnis yang terkait dengan data tersebut sangatlah diperlukan. Informasi yang serupa mungkin juga dapat diperoleh dari dokumentasi yang relevan dengan database tersebut. Pengetahuan ini penting karena dalam proses penyiapan data hampir selalu kita akan mengambil satu variabel dari suatu tabel data dan satu variabel lain dari tabel data yang berbeda.

* Definisi variabel

Pada saat menyiapkan data untuk analisis dan pemodelan, tentu saja perlu dipersiapkan keberadaan variabel-variabel yang diperlukan. Sebagian variabel, seperti jenis kelamin nasabah barangkali sudah langsung tersedia dan tinggal dipergunakan. Namun ada juga variabel lain yang harus digenerate atau diperoleh menggunakan formula-formula tertentu. Data scientist perlu mengetahui dengan baik definisi setiap variabel yang diperlukan dan bagaimana memperolehnya dari database.

* Struktur data

Dalam melakukan analisis, diperlukan struktur data yang spesifik tergantung pada kebutuhan analisis dan pemodelannya. Sewaktu menggunakan prosedur pemodelan data deret waktu (time series) dengan satuan waktu bulanan misalnya, maka kita memerlukan data waktu tersusun ke bawah dimana setiap baris pada tabel data merepresentasikan data pada bulan tertentu dengan urutan bulan yang benar. Namun ketika akan menggunakan data tersebut untuk melihat perilaku bulanan suatu individu pelanggan, maka informasi bulan akan berada pada posisi kolom. Data scientist perlu memahami struktur data dari database yang ada dan struktur data seperti apa yang diperlukan pada proses analitika. Dengan pemahaman yang baik mengenai hal ini, proses analitika dapat berjalan dengan lebih mudah dan lebih efisien waktunya.

* Format data

Pada tahapan penyiapan data, seorang data scientist juga harus memperhatikan format data yang perlu mereka hasilkan. Format data ini berhubungan dengan tipe data seperti numerik ataukah bukan, misalnya string (text) dan label kode. Proses analitika dapat saja terhambat karena data variabel tertentu yang seharusnya bertipe numerik, tapi disiapkan dalam format yang bertipe string pada tabel data.

**4. Kegiatan dalam Penyiapan Data**

Terdapat beragam aktifitas yang perlu dikerjakan dalam proses penyiapan data. Beberapa orang mengelompokkan kegiatan penyiapan data untuk proses analitika menjadi beberapa kelompok sebagai berikut:

* *Data Integration*, yaitu melakukan pengintegrasian data yang bersumber dari dua atau lebih dataset yang berbeda. Dalam prosesnya bisa saja diperlukan penggantian nama variabel agar seragam, penyeragamaan tipe data, dan lain sebagainya sehingga proses integrasi dapat dilakukan dengan baik. Butuh pemahaman mengenai konten dari setiap dataset yang akan digabung sehingga data scientist mengintegrasikan dataset yang memang diperlukan.
* *Data Cleaning*, yaitu melakukan pembenahan-pembenahan terhadap nilai data yang bermasalah. Yang dimaksudkan bermasalah antara lain adalah datanya kosong/hilang, adanya amatan-amatan yang mengalami duplikasi, serta adanya amatan dengan nilai yang anomali atau berbeda jauh dari yang lain atau bahkan terkesan tidak masuk akal.
* *Data Transformation*, yaitu menghasilkan variabel baru yang didapatkan nilainya dari variabel-variabel yang sudah ada pada database. Yang termasuk pada kegiatan ini antara lain adalah pembuatan variabel baru, mengagregasi data ke level yang lebih besar, melakukan penskalaan data, serta diskretisasi/binning yaitu mengkategorisasi data numerik menjadi kategorik atau diskret.
* *Data Reduction*, yaitu melakukan pereduksian data tanpa mengurangi informasi yang terkandung di dalam data secara signifikan. Aktifitas yang tergolong pada reduksi data antara lain adalah seleksi variabel, pereduksian dimensi, attribut sampling, dan sebagainya.

Pada buku ini, tidak semua aktifitas yang dinyatakan di atas akan dibahas detail. Yang menjadi fokus dari buku ini utamanya adalah *data integration* dan *data transformation*, sehingga yang akan dibahas detail beserta dengan hal-hal teknis pemrogramannya menggunakan R adalah aktifitas-aktifitas berikut:

* Filtering, yaitu memilih data yang berisi amatan-amatan sesuai dengan kriteria yang diperlukan untuk analisis dan pemodelan. Misalnya saja, pada saat melakukan pemodelan terhadap nasabah tabungan, kita barangkali hanya tertarik pada nasabah yang memiliki nilai saldo tertentu. Proses filtering dilakukan sehingga hanya nasabah yang sesuai kriteria saja yang dimasukkan pada data yang disiapkan.
* Generate variabel, yaitu membuat variabel-variabel baru dari data-data yang sudah ada baik merupakan fungsi matematis yang melibatkan satu variabel asal maupun yang melibatkan banyak variabel. Sebut saja misalnya kita memerlukan data besarnya transaksi dalam satu tahun yang dapat diperoleh dengan menjumlahkan data transaksi bulanan yang tersedia pada database.
* Agregasi, yang umumnya berupa ringkasan dari data pada level rendah ke level lebih tinggi. Misalnya saja dalam database tersedia data transaksi dimana baris merepresentasikan setiap produk yang dibeli oleh seorang pelanggan dan kita ingin memperoleh total pembelian setiap pelanggan. Agregasi dilakukan sehingga didapatkan data dimana setiap amatan merupakan total pembelian seorang pelanggan dalam periode waktu tertentu.
* Subsetting, yaitu melakukan seleksi variabel sehingga hanya variabel-variabel yang diperlukan saja yang akan dipersiapkan.
* Reshaping, yaitu mengubah struktur data yang semula ada di bagian baris dipindah menjadi di bagian kolom atau sebaliknya.
* Joining data, yaitu memperoleh satu gugus data baru dengan menggabungkan dari dua atau lebih dataset. Proses ini diperlukan karena satu variabel tersedia di suatu dataset sedangkan variabel yang lainnya adanya pada dataset yang berbeda.