NicoElNino/iStock

Lima Konsep Statistika yang Data Scientist Perlu Tahu

Data science ditopang oleh tiga bidang ilmu termasuk statistika. Oleh karena itu, seorang *data scientist* perlu memiliki pemahaman mengenai konsep statistika.

PENULIS

BAGUS SARTONO
Dosen di Departemen Statistika IPB dan Wakil Ketua FORSTAT)

YUNANTO PUTRANTO
Data Scientist di DataLabs Analytics

BELAKANGAN INI, jargon "kalau bicara pakai data" makin sering terdengar di media. Ini indikasi bahwa kepedulian terhadap data makin tumbuh, seiring dengan makin menggeliatnya industri digital serta industri-industri lain yang santer memanfaatkan data. Indonesia sendiri menargetkan untuk menjadi kekuatan ekonomi digital terbesar ASEAN pada tahun 2020, dengan proyeksi nilai transaksi *e-commerce* mencapai 130 juta dolar AS. Kebutuhan akan orang bertalenta untuk peran di bidang teknologi data, otomatis ikut meningkat.

Salah satu lembaga edukasi dan *hiring solution* di bidang teknologi data di Indonesia menyatakan bahwa ketersediaan orang berbakat yang siap terjun ke dunia ini masih terbatas, karena *data science* meliputi banyak bidang keilmuan. Orang bersangkutan dituntut untuk tidak hanya menguasai teknologinya, tetapi juga memahami bisnisnya. Seperti ditunjukkan pada **Gambar 1**, *data science* ditopang oleh tiga bidang ilmu, yaitu teknologi infomasi atau ilmu komputer, matematika dan statistika, serta pengetahuan dari aspek bisnis atau bidang terapannya. Memiliki dan mampu mengombinasikan keahlian pada tiga bidang tersebut akan membawa seseorang menjadi *data scientist* yang mumpuni.

Statistika, yang merupakan ilmu tentang data, tidak dapat dilepaskan dari dunia *data science*. Seorang *data scientist* perlu memiliki pemahaman konsep statistika yang lima di antaranya diulas berikut ini.

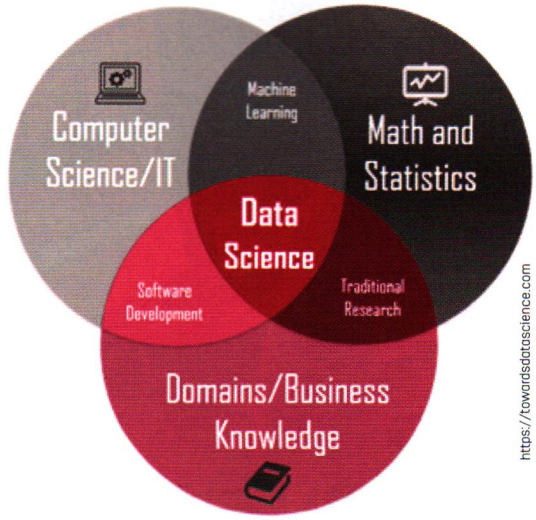
1 Metode Sampling

Dalam banyak situasi, data yang dihadapi adalah sampel. Memperoleh data dari seluruh objek studi atau populasi hampir mustahil dan memerlukan usaha yang besar, baik dari segi waktu, biaya, maupun tenaga. Agar karakteristik data sampel yang diambil tetap dapat menggambarkan kondisi populasi secara keseluruhan, diperlukan metode sampling yang tepat agar penarikan kesimpulan menjadi sah.

Metode sampling dibagi menjadi dua kelompok besar yaitu *random sampling* dan *nonrandom sampling*. Teknik *random sampling* memungkinkan kita mengidentifikasi peluang terpilihnya masing-masing individu, sehingga proses generalisasi dapat dikerjakan. Kalaupun tidak terlibat dalam pengumpulan data, setidaknya *data scientist* harus mengetahui proses samplingnya.

2 Skala Pengukuran Data

Agar dapat memproses data dengan tepat, *data scientist* perlu mengetahui skala pengukuran variabel yang ada seperti nominal, ordinal, interval, dan rasio. Variabel berskala nominal hanya melakukan pengkategorian atau penggolongan seperti jenis kelamin atau pekerjaan. Sementara pada skala ordinal, antara kategori sudah terdapat urutan dengan makna tertentu. Yang termasuk pada skala ini antara lain adalah tingkat pendidikan, tingkat kepuasan (mulai dari tidak puas, kurang puas, sampai sangat puas), dan tingkat risiko (rendah, sedang, tinggi). Variabel dengan kedua skala tersebut sering dikenal sebagai variabel kategorik. Dua skala yang lain, interval dan rasio, adalah untuk variabel numerik.



Gambar 1.
Komponen utama penopang bidang *data science*.

Pengetahuan tentang skala ini penting karena berhubungan dengan perbedaan cara menangani data sampai metode analisis dan pemodelan. Ketidakpahaman dalam hal skala pengukuran dapat berakibat pada kekeliruan dalam proses analitik.

3 Statistika Deskriptif

Data yang besar akan lebih mudah dipahami jika dirangkum dalam bentuk yang lebih sederhana, baik dalam bentuk angka atau penyajian dengan bentuk grafik yang sesuai. Perangkuman dan penyajian data inilah yang disebut sebagai statistika deskriptif. Selain mendapatkan gambaran tentang data, statistika deskriptif ini juga seringkali mengungkap informasi atau *insight* dari dalam data yang tidak terlihat sebelumnya. Statistika deskriptif itu di antaranya mencakup ukuran pemusatan data seperti nilai rata-rata, median, modus, serta ukuran persebaran data seperti standar deviasi dan varian.

Data scientist perlu saksama memperhatikan ukuran-ukuran statistika deskriptif agar bisa mendapatkan gambaran data dengan benar. Sebagai contoh, hanya mengungkapkan nilai rata-rata laba dari seratus gerai *reseller* bisa berujung pada keputusan bisnis yang tidak tepat. Bisa jadi, ada sejumlah gerai yang memiliki ni laba yang amat tinggi sehingga mendongkrak nilai rata-rata laba keseluruhan. Padahal kenyataanya, kebanyakan gerai labanya masih tidak besar.

4 Distribusi dan Uji Hipotesis

Data sampel sering juga digunakan sebagai bahan mengonfirmasi suatu hipotesis. *Data scientist* perlu memahami bahwa statistik-statistik yang mereka hasilkan dari data samp hanya satu dari tak terhingga kemungkinan. Statistik-statistik tersebut memiliki distribusi peluang. Pada saat melakukan pengujian hipotesis terdapat peluang membuat kesalahan dalam mengambil kesimpulan. Dalam uji hipotesis ini *data scientist* perlu familiar dengan istilah-istilah seperti hipotesis n hipotesis alternatif, taraf nyata, nilai dan daerah kritis, statistik uji dan kesalahan tipe I dan II, *p-value*, serta kuasa uji.

5 Konsep Bayesian

Teorema Bayes dikenalkan kurang lebih 250 tahun lalu sebagai pengembangan dari pengetahuan mengenai peluang bersyarat. Prinsip Bayesian bany digunakan dalam melakukan pendugaan dan pengujian hipotesis, dengan pemahaman bahwa distribusi posterior dari suatu parameter populasi dapat berbeda, tergantung pada data sampel yang dimiliki. Pada analisis klasifikasi, *Bayesian classifier* yang mengandalkan perhitungan peluang bersyarat berdasarkan data *training*, mam memberikan prediksi-prediksi peluang kejadian dengan sangat baik. IK