



YAYASAN PRIMA AGUS TEKNIK

APLIKASI AI DAN MACHINE LEARNING DALAM BISNIS

oleh:
Dr. Joseph Teguh Santoso, S.Kom, M.Kom

APLIKASI AI DAN MACHINE LEARNING DALAM BISNIS

oleh:
Dr. Joseph Teguh Santoso, S.Kom, M.Kom



YAYASAN PRIMA AGUS TEKNIK

PENERBIT :
YAYASAN PRIMA AGUS TEKNIK
Jl. Majapahit No. 605 Semarang
Telp. (024) 6723456. Fax. 024-6710144
Email : penerbit_ypat@stekom.ac.id

ISBN 978-623-8120-64-2 (PDF)



9 786238 120642

Aplikasi AI dan Machine Learning dalam Bisnis

Penulis :

Dr. Joseph Teguh Santoso, S.Kom., M.Kom

ISBN : 9 786238 120642

Editor :

Muhammad Sholikan, M.Kom

Penyunting :

Dr. Mars Caroline Wibowo. S.T., M.Mm.Tech

Desain Sampul dan Tata Letak :

Irdha Yuniyanto, S.Ds., M.Kom

Penebit :

Yayasan Prima Agus Teknik Bekerja sama dengan
Universitas Sains & Teknologi Komputer (Universitas STEKOM)

Redaksi :

Jl. Majapahit no 605 Semarang

Telp. (024) 6723456

Fax. 024-6710144

Email : penerbit_ypat@stekom.ac.id

Distributor Tunggal :

Universitas STEKOM

Jl. Majapahit no 605 Semarang

Telp. (024) 6723456

Fax. 024-6710144

Email : info@stekom.ac.id

Hak cipta dilindungi undang-undang

Dilarang memperbanyak karya tulis ini dalam bentuk dan dengan cara
apapun tanpa ijin tertulis dari penerbit

KATA PENGANTAR

Pemahaman tentang pembelajaran mesin penting karena memiliki dampak besar di banyak aspek kehidupan kita. Secara khusus, hal ini mendorong ledakan aplikasi “Kecerdasan Buatan” di banyak bidang seperti penerjemahan bahasa, robot otonom, dan diagnosis medis. Kecerdasan buatan dan pembelajaran mesin juga mempunyai dampak langsung pada banyak fungsi bisnis sehari-hari. Penerapan kecerdasan buatan (AI) dan pembelajaran *mesin (machine learning)* dalam bisnis telah menjadi tren yang semakin populer karena potensinya untuk meningkatkan efisiensi operasional, pengambilan keputusan yang lebih baik, dan inovasi.

Aplikasi AI dan machine learning dalam bisnis dapat membantu meningkatkan efisiensi, mempercepat inovasi, meningkatkan kepuasan pelanggan, dan mengoptimalkan pengambilan keputusan strategis. Namun, penggunaan teknologi ini juga harus mempertimbangkan etika, privasi, dan keamanan data. Penerapan AI dan machine learning dalam bisnis dapat membantu meningkatkan efisiensi, mempercepat inovasi, meningkatkan kepuasan pelanggan, dan mengoptimalkan pengambilan keputusan strategis.

Dalam buku ini terbagi menjadi 15 bab. Bab 1 buku ini akan memberikan pembaca pengenalan singkat tentang apa yang dimaksud dengan kecerdasan buatan. Bab 2 dalam buku ini membahas tentang apa yang dimaksud dengan kecerdasan buatan dan artificial intelligence (AI). Dalam bab 3 penulis akan menjelaskan tentang model prediktif dalam perhitungan kecerdasan buatan. Bab 4 dan 5 buku ini akan memberikan gambaran kepada penulis cara kerja pembelajaran mesin. Bab 6 dalam buku ini akan membahas tentang permodelan produktif untuk pengambilan keputusan. Bab 7 dalam buku ini akan memberikan gambaran tentang pohon keputusan dan kartu skor. Dua bab setelahnya (bab 8 dan 9) akan menjelaskan tentang jaringan syaraf tiruan yang diaplikasikan dalam pembelajaran tanpa pengawasan.

Bab 10 buku ini akan menjelaskan bagaimana membangun model prediktif dalam pembelajaran mesin. Melanjutkan di bab sebelumnya, bab 11 akan menerangkan dengan detail dalam memoperasikan pembelajaran mesin. Bab 12 akan membahas tentang hubungan big data dan pembelajaran mesin untuk menentukan keputusan. Buku ini juga tidak lupa membahas tentang etika dan hukum dalam penerapan AI dan machine learning dalam AI karena pentingnya merencanakan dan mengimplementasikan teknologi ini dengan bijak dan mempertimbangkan aspek keamanan dan privasi data akan dibahas di bab 13. Bab 14 buku ini akan membahas tentang pembelajaran mesin tercanggih. Bab terakhir buku ini, bab 15, membahas tentang salah satu aplikasi dan implementasi AI dan machine learning dalam bisnis berupa mobil selfdriving. Mobil self-driving atau kendaraan otonom adalah kendaraan yang mampu bergerak tanpa adanya pengemudi manusia dan dapat melakukan perjalanan, navigasi, dan mengambil keputusan sendiri berdasarkan pada perangkat keras (sensors) dan perangkat lunak (algoritma AI dan machine learning) yang tertanam di dalamnya. Buku ini juga dilengkapi dengan latihan soal dan lampiran-lampiran yang dapat membantu dalam pemahaman tentang lecerdasan buatan dan pembelajaran mesin. Akhir kata semoga buku ini berguna bagi para pembaca.

Semarang, Oktober 2023
Penulis

Dr. Dr. Joseph Teguh Santoso, S.Kom, M.Kom

DAFTAR ISI

Halaman judul	i
Kata pengantar	ii
Daftar isi	iii
BAB 1 PENGANTAR	1
BAB 2 APA ITU PEMBELAJARAN MESIN DAN KECERDASAN BUATAN (AI)	4
BAB 3 NILAI SKOR YANG DIHASILKAN OLEH MODEL PREDIKTIF.....	13
BAB 4 MENGAPA MENGGUNAKAN PEMBELAJARAN MESIN?	18
BAB 5 CARA KERJA PEMBELAJARAN MESIN	22
BAB 6 MODEL PREDIKTIF UNTUK MENGAMBIL KEPUTUSAN	30
BAB 7 KARTU SKOR DAN POHON KEPUTUSAN	33
BAB 8 JARINGAN SYARAF TIRUAN DAN PEMBELAJARAN MENDALAM	37
BAB 9 PEMBELAJARAN TANPA PENGAWASAN	45
BAB 10 MEMBANGUN MODEL PREDIKTIF	55
BAB 11 MENGOPERASIKAN PEMBELAJARAN MESIN	66
BAB 12 HUBUNGAN ANTARA BIG DATA DAN PEMBELAJARAN MESIN	74
BAB 13 ETIKA, HUKUM, DAN GDPR	78
BAB 14 PEMBELAJARAN MESIN TERCANGGIH	88
BAB 15 KAPAN SAYA BISA MEMBELI MOBIL SELFDRIVING?	95
Latihan soal	101
Lampiran A	103
Lampiran B	109
Daftar Pustaka	117

BAB 1

PENGANTAR

Apakah Anda memiliki ponsel pintar atau kartu kredit? Apakah Anda membeli barang dari supermarket atau bermain game komputer? Apakah Anda bekerja atau menggunakan layanan kesehatan? Jika jawaban atas pertanyaan-pertanyaan ini adalah ya, maka kecerdasan buatan dan pembelajaran mesin akan berdampak pada kehidupan Anda dalam satu atau lain cara. Hal ini karena mereka adalah alat utama yang digunakan organisasi untuk memanfaatkan data yang mereka miliki tentang Anda guna memutuskan bagaimana mereka akan berurusan dengan Anda. Mereka digunakan untuk memberikan informasi kepada organisasi tentang bagaimana Anda cenderung berperilaku dalam situasi yang berbeda, dan juga cara mereka memperlakukan Anda untuk memaksimalkan tujuan mereka (dan terkadang tujuan Anda).

Teknologi-teknologi ini kini digunakan di hampir setiap lapisan masyarakat untuk meningkatkan proses dan meningkatkan pengalaman sehari-hari masyarakat melalui mesin “kecerdasan buatan” dan antarmuka komputer. Echo dari Amazon, Siri dari Apple, dan Google Terjemahan hanyalah tiga produk perangkat lunak terkenal yang menunjukkan manfaat yang dapat diberikan oleh teknologi ini.

Saat ini, banyak produk dan layanan juga bersifat adaptif. Mereka menyesuaikan respons mereka terhadap perilaku masing-masing pengguna. Layanan streaming TV dan musik belajar mengidentifikasi konten yang Anda sukai, dan memberi Anda rekomendasi yang pasti akan Anda minati. Ubah jenis musik yang Anda dengarkan dan rekomendasinya juga akan berubah. Demikian pula, Anda dapat membeli sistem pemanas yang belajar mengantisipasi waktu terbaik untuk menyalakan pemanas sehingga Anda tidak perlu repot, sekaligus mengoptimalkan penggunaan energi untuk mengurangi tagihan Anda. Ini adalah contoh lebih lanjut dari tindakan kecerdasan buatan.

Teks ringkas ini memberikan gambaran umum manajerial (yaitu non-teknis dan tanpa rumus rumit) tentang kecerdasan buatan dan pembelajaran mesin, apa itu kecerdasan buatan dan bagaimana penggunaannya. Tidak ada pengetahuan sebelumnya yang diasumsikan. Dengan kata lain, jika Anda bisa membaca dan menulis serta mengerjakan aritmatika dasar (ada sedikit aritmatika, tapi tidak banyak) maka Anda seharusnya setuju dengan materi dalam buku ini.

Pertanyaan yang bagus untuk ditanyakan saat ini adalah: mengapa saya perlu mengetahui hal-hal ini? Salah satu alasannya bersifat pribadi. Sistem pengambilan keputusan yang cerdas, berdasarkan pembelajaran mesin, banyak digunakan oleh organisasi untuk memutuskan bagaimana memperlakukan Anda, teman, dan keluarga Anda. Mereka menggunakan alat-alat ini untuk memutuskan apakah Anda akan menerima tawaran yang bagus atau buruk, apakah Anda harus ditempatkan di barisan depan atau belakang, apakah Anda akan menjalani pemeriksaan pajak atau diperlakukan sebagai tersangka tindak pidana. kasus. Oleh karena itu, bukanlah ide yang buruk untuk mengetahui sesuatu tentang hal-hal ini

sehingga Anda dapat memahami mengapa suatu organisasi memperlakukan Anda dengan cara tertentu dan bukan cara lain.

Alasan lain untuk mempelajari kecerdasan buatan, dan salah satu yang menjadi fokus utama buku ini, adalah karena kecerdasan buatan kini menjadi alat bisnis yang mainstream. Belum lama berselang, kecerdasan buatan merupakan domain dari beberapa spesialis kutu buku yang bekerja terutama di bidang akademis, jasa keuangan, atau departemen pemasaran besar. Saat ini, apa pun bisnis yang Anda jalani, penerapan kecerdasan buatan, berdasarkan pembelajaran mesin, dapat ditemukan di seluruh aktivitas bisnis. Hal ini mencakup segala hal mulai dari pemeriksaan karyawan, menjawab pertanyaan pelanggan dan target pemasaran hingga robot di lini produksi, manajemen gudang, dan pengiriman pelanggan. Sebagai konsekuensinya, kecerdasan buatan mendukung atau menggantikan keahlian manusia di banyak bidang. Beberapa contohnya antara lain: mengganti penjamin emisi ketika menetapkan premi asuransi, membantu profesional HR memutuskan siapa yang akan dipekerjakan/dipecat, secara otomatis mengidentifikasi pelanggan saat mereka masuk ke toko Anda dan mendukung dokter ketika mendiagnosis penyakit seperti kanker dan penyakit jantung.

Kecerdasan buatan telah mencapai kemajuan pesat. Ini bukan sekedar iseng dan akan tetap ada. Organisasi-organisasi yang dapat menggunakannya untuk memecahkan masalah bisnis, meningkatkan efisiensi dan memotong biaya akan mendapatkan keuntungan dibandingkan pesaing mereka.

Ini tidak berarti bahwa Anda perlu mempelajari semua hal yang perlu diketahui oleh pakar teknis (ilmuwan data). Namun, memiliki pengetahuan tentang apa itu kecerdasan buatan dan pembelajaran mesin, dan mengetahui bagaimana keduanya dapat digunakan untuk membantu organisasi Anda memberikan produk dan layanan yang lebih baik, akan bermanfaat. Yang paling penting, karena agar alat-alat ini dapat digunakan secara efektif, mereka perlu fokus pada tujuan bisnis untuk mengatasi masalah spesifik yang dihadapi organisasi.

Jika di sisi lain Anda adalah seorang yang suka mengutip persamaan, ahli menyulap rumus, dan jenius matematika yang berpikir bahwa mereka mengetahui segala hal yang perlu diketahui tentang kecerdasan buatan, maka buku ini mungkin juga bermanfaat bagi Anda. Bahkan mungkin lebih banyak dari mereka yang tidak tahu apa-apa. Mengapa? Karena jika yang Anda pedulikan hanyalah aspek teoretis dari subjek tersebut, maka Anda menghadapi risiko nyata menemui jalan buntu ketika harus memberikan solusi yang berguna di ladang ranjau yaitu dunia nyata; dunia yang penuh dengan masalah sosial, etika dan politik. Hal ini, ditambah dengan semakin banyaknya undang-undang privasi dan perlindungan data, dapat menggagalkan solusi Anda, tidak peduli seberapa bagus solusi tersebut secara matematis.

Tanpa mempertimbangkan “masalah-masalah ringan” ini, kemungkinan terbaiknya adalah solusi yang Anda kembangkan tidak dapat diterapkan. Skenario terburuknya adalah Anda merancang sistem berbasis kecerdasan buatan yang akan membawa Anda ke pengadilan karena sistem tersebut mendiskriminasi secara tidak adil terhadap kelompok minoritas,

perempuan, atau kelompok masyarakat lainnya. Mungkin Anda bisa melewatkan beberapa bab sebelumnya, tapi Anda sebaiknya membaca bab berikutnya.

Untuk mendapatkan hasil maksimal dari kecerdasan buatan, data scientist perlu berinteraksi dengan pengguna bisnis untuk memahami masalah mereka. Ilmuwan data juga perlu memahami budaya organisasi dan pendekatannya terhadap adopsi ide, teknologi, dan praktik kerja baru. Permasalahan hukum dan peraturan di wilayah tempat klien mereka beroperasi juga memerlukan pertimbangan yang matang. Tidak peduli seberapa bagus solusi dalam hal perangkat keras dan perangkat lunak yang canggih, jika tidak selaras dengan tujuan bisnis dan proses operasional organisasi, maka itu semua hanya membuang-buang waktu dan uang.

Banyak pemasok solusi yang dapat menipu Anda dengan teknologi mewah dan terminologi terbaru mereka, yang seringkali hanya merupakan rebranding dari teknologi tahun lalu dengan sentuhan baru. Namun, pemasok yang memberi nilai tambah adalah mereka yang meluangkan waktu untuk memahami cara kerja Anda dan organisasi Anda. Mereka kemudian akan menentukan apakah dan bagaimana solusi mereka dapat digunakan untuk meningkatkan apa yang Anda lakukan dan menjelaskan hal ini kepada Anda dalam bahasa sederhana yang dapat Anda pahami tanpa perlu membuka Wikipedia.

Kecerdasan buatan yang sukses adalah hal dua arah. Ilmuwan data perlu mengetahui sesuatu tentang organisasi Anda dan apa yang dilakukannya, dan Anda perlu memahami sedikit tentang kecerdasan buatan dan pembelajaran mesin. Tanpa pemahaman bersama ini, kecil kemungkinan Anda atau organisasi Anda akan mampu mewujudkan manfaat penuh yang ditawarkan oleh kecerdasan buatan.

Baik. Jadi, apa yang akan Anda pelajari dari membaca buku ini? Topik utama yang akan kita bahas dalam bab-bab berikut adalah:

- Apa itu pembelajaran mesin dan kecerdasan buatan.
- Hal-hal yang digunakan organisasi untuk menggunakan kecerdasan buatan.
- Seperti apa model prediktifnya.
- Hubungan antara kecerdasan buatan, pembelajaran mesin, dan “Big Data.”
- Orang dan alat yang dibutuhkan untuk menerapkan kecerdasan buatan.
- Bagaimana menggunakan kecerdasan buatan untuk meningkatkan proses bisnis dan keuntungan.
- Masalah hukum dan etika yang perlu dipertimbangkan ketika mengembangkan solusi berbasis kecerdasan buatan yang akan digunakan untuk mengambil keputusan tentang manusia.
- Bagaimana bentuk pembelajaran mesin tingkat lanjut diterapkan untuk mendorong aplikasi kecerdasan buatan seperti pengenalan objek dan terjemahan bahasa.
- Keterbatasan pembelajaran mesin dan kecerdasan buatan saat ini.

BAB 2

APA ITU PEMBELAJARAN MESIN DAN KECERDASAN BUATAN (AI)?

Pembelajaran mesin adalah penggunaan prosedur matematika (algoritma) untuk menganalisis data. Tujuannya adalah untuk menemukan pola yang berguna (hubungan atau korelasi) antara item data yang berbeda. Setelah hubungan diidentifikasi, hubungan ini dapat digunakan untuk membuat kesimpulan tentang perilaku kasus-kasus baru ketika kasus-kasus tersebut muncul. Intinya, ini analog dengan cara orang belajar. Kita mengamati apa yang terjadi di sekitar kita dan menarik kesimpulan dari pengalaman kita tentang cara kerja dunia. Kita kemudian menerapkan apa yang telah kita pelajari untuk membantu kita menghadapi situasi baru yang kita hadapi. Semakin banyak kita mengalami dan belajar, semakin baik kemampuan kita dalam mengambil keputusan.

Salah satu penerapan pembelajaran mesin adalah pengenalan objek. Tujuannya adalah untuk mengembangkan sistem yang dapat mengidentifikasi objek sehari-hari dari gambar yang disajikan oleh sistem. Data yang digunakan untuk mengembangkan sistem pengenalan objek terdiri dari gambar-gambar objek yang berbeda seperti kursi, payung, mesin cuci dan lain sebagainya. Setiap gambar yang disajikan ke algoritme pembelajaran mesin diberi label untuk mengidentifikasi jenis objek yang ada di dalamnya. Untuk setiap jenis objek mungkin terdapat ratusan atau ribuan gambar berbeda, yang mewakili bentuk alternatif objek tersebut dari sudut pandang berbeda (Anda akan terkejut melihat banyaknya varian payung yang ada!)

Dengan menganalisis gambar yang berbeda, algoritme pembelajaran mesin mengenali bahwa objek tertentu dikaitkan dengan fitur (pola) tertentu. Kursi cenderung memiliki tonjolan (kaki) yang berasal dari alas (tempat duduk) yang datar dan seringkali berbentuk persegi. Mereka juga dibedakan dari tinja dengan adanya sandaran. Mesin cuci cenderung berbentuk kubus dengan kenop menyala dan hampir tidak pernah berwarna merah muda atau oranye (tolong beri tahu saya jika Anda pernah menemukan tempat di mana saya bisa membeli mesin cuci berwarna merah muda!) Begitu pula dengan payung yang panjang dan tipis (bila ditutup). Sering kali, namun tidak selalu, berwarna hitam dan sebagainya.

Salah satu penerapan pembelajaran mesin yang paling umum dan mungkin yang pertama adalah prediksi. Ini tentang menggunakan pembelajaran mesin untuk menentukan sesuatu yang saat ini tidak Anda ketahui, berdasarkan informasi yang Anda miliki saat ini. Pola yang ditemukan berkaitan dengan hubungan antara perilaku dan hasil. Seringkali hal ini berkaitan dengan perilaku seseorang di masa lalu dan apa yang kemudian mereka lakukan. Setelah mengidentifikasi hubungan-hubungan yang ada, maka dimungkinkan untuk membuat prediksi tentang perilaku seseorang di masa depan berdasarkan keadaannya saat ini. Jika Anda memberi saya contoh riwayat pembelian orang-orang sebelumnya, saya dapat memanfaatkan pembelajaran mesin untuk mengidentifikasi pola perilaku pembelian mereka. Saya kemudian dapat menggunakan pola-pola ini untuk memprediksi barang apa yang kemungkinan besar akan dibeli seseorang selanjutnya; yaitu pembelian di masa depan adalah hasil yang ingin saya

prediksi. Hal ini memungkinkan saya menargetkan mereka dengan penawaran promosi yang disesuaikan untuk produk spesifik tersebut.

Penggunaan pembelajaran mesin untuk prediksi terkadang disebut sebagai pemodelan prediktif atau Analisis Prediktif (PA). Faktanya, analitik prediktif adalah aplikasi pembelajaran mesin yang umum sehingga banyak orang (benar atau salah) sering menggunakan kedua istilah tersebut secara bergantian.

Memprediksi perilaku individu di masa depan adalah hal yang biasanya diasosiasikan orang dengan pembelajaran mesin, namun ada situasi dan masalah lain yang dapat diterapkan oleh pembelajaran mesin. Yang Anda butuhkan hanyalah peristiwa atau hal yang tidak diketahui yang ingin Anda tentukan (prediksi), dan ini bisa terjadi di masa lalu, sekarang, atau masa depan. Dokter memeriksa pasiennya, melakukan tes dan menanyakan gejalanya untuk mengumpulkan bukti (data). Mereka kemudian menggunakan data ini untuk menyimpulkan apa yang menurut mereka salah pada pasien. Mereka tidak membuat prediksi tentang kesehatan pasien di masa depan, namun mencoba mencari tahu apa yang salah dengan kesehatan pasien saat ini. Dokter dapat melakukan hal ini dengan tingkat akurasi yang tinggi karena mereka melakukan referensi silang antara informasi pasien yang mereka peroleh dengan apa yang telah mereka pelajari dari pelatihan dan praktik selama bertahun-tahun. Dengan kata lain, mereka mencari bagaimana gejala pasien berhubungan dengan pengetahuan mereka tentang penyakit yang diketahui. Pembelajaran mesin dapat diterapkan dengan cara yang sama. Mengingat banyaknya informasi mendetail tentang gejala berbagai penyakit, pembelajaran mesin dapat digunakan untuk memperkirakan kemungkinan seseorang menderita kondisi tertentu, berdasarkan gejala yang muncul.

Cara lain untuk memikirkan pembelajaran mesin/analisis prediktif adalah sebagai metode untuk mengurangi ketidakpastian. Ada banyak kemungkinan hasil yang dapat terjadi dalam situasi tertentu. Pembelajaran mesin tidak akan memberi tahu Anda dengan pasti hasil mana yang akan terjadi, namun dapat memberikan beberapa wawasan tentang kemungkinan, atau peluang, dari setiap hasil.

Anda mungkin tahu bahwa saat seseorang berbelanja bahan makanan, mereka sering kali membeli roti, anggur, dan ayam, namun dengan pembelajaran mesin, Anda dapat menentukan bahwa ada, misalnya, 80% kemungkinan produk berikutnya yang mereka beli adalah roti, 15% kemungkinan mereka membeli anggur dan peluang 5% mereka membeli ayam. Oleh karena itu, jika Anda ingin mendorong mereka untuk melakukan pembelian berikutnya di toko Anda, kemungkinan besar Anda akan memenangkan kebiasaan mereka dengan tawaran roti daripada tawaran anggur atau ayam.

Model prediktif (atau hanya model ke depan) adalah keluaran yang dihasilkan oleh proses pembelajaran mesin. Model tersebut menangkap hubungan (pola) yang telah ditemukan oleh proses analisis. Setelah model dibuat, model tersebut dapat digunakan untuk menghasilkan prediksi baru. Organisasi kemudian menggunakan prediksi model tersebut untuk memutuskan apa yang harus dilakukan atau bagaimana memperlakukan orang. Jadi, pembelajaran mesin adalah sebuah proses dan model prediktif adalah produk akhir dari proses tersebut.

Ada banyak jenis model prediktif, dan ada lusinan, bahkan ratusan, teknik dan algoritme pembelajaran mesin yang dapat digunakan untuk menghasilkan model. Namun, apa pun jenis model atau matematika yang digunakan untuk membuatnya, prediksi suatu model hampir selalu diwakili oleh angka – skor. Semakin tinggi skornya, semakin besar kemungkinan seseorang untuk berperilaku sesuai prediksi model, semakin rendah skornya, semakin kecil kemungkinannya untuk berperilaku seperti itu.

Pembelajaran mesin dapat diterapkan dalam segala situasi dan berbagai jenis masalah. Namun, penerapan pembelajaran mesin dalam bisnis yang paling umum, dan yang menjadi fokus utama buku ini, berkaitan dengan apa yang akan dilakukan orang atau bagaimana mereka akan berperilaku di masa depan, berdasarkan apa yang Anda ketahui tentang mereka saat ini.

Salah satu penerapan pembelajaran mesin yang sangat terkenal adalah penilaian kredit. Ketika seseorang menginginkan pinjaman, kartu kredit atau hipotek, pemberi pinjaman menanyakan pertanyaan individu tentang diri mereka sendiri dan gaya hidup mereka. Mereka kemudian menggabungkannya dengan informasi dari laporan kredit yang berisi rincian tentang riwayat pinjaman individu sebelumnya, yang disediakan oleh agen referensi kredit seperti Experian atau Equifax. Informasi tersebut kemudian dimasukkan ke dalam model prediktif untuk menghasilkan skor kredit.

Jika Anda tinggal di AS, Anda mungkin akrab dengan skor FICO dan/atau Vantage. Skor tinggi (>750) merupakan prediksi bahwa seseorang kemungkinan besar akan mengembalikan uang yang dipinjamnya; yaitu bahwa mereka layak mendapat kredit. Skor yang rendah (<500) menunjukkan seseorang sangat tidak layak mendapat kredit. Bank dan perusahaan pembiayaan di seluruh dunia menggunakan metode penilaian kredit yang serupa.

Penerapan umum lainnya dari pembelajaran mesin adalah pemasaran target. Dengan adanya informasi tentang usia seseorang, jenis kelamin, pendapatan, penelusuran web, riwayat pembelian, lokasi, dan sebagainya, departemen pemasaran dapat memperkirakan apakah orang tersebut tertarik pada produk tertentu atau tidak. Mereka kemudian menggunakan prediksi tersebut untuk memutuskan apakah akan menargetkan mereka dengan penawaran promosi atau tidak. Demikian pula, model prediktif juga dapat digunakan untuk menyimpulkan seberapa besar masyarakat bersedia membayar untuk produk seperti asuransi. Informasi ini kemudian digunakan untuk menyesuaikan strategi penetapan harga yang dipersonalisasi dengan keadaan individu masing-masing.

Contoh lain dari penerapan pembelajaran mesin adalah layanan kesehatan preventif. Sistem layanan kesehatan tradisional bersifat reaktif. Orang-orang mencari bantuan medis ketika mereka merasa sakit. Dokter kemudian melakukan yang terbaik untuk mengobati penyakit yang mereka derita – pengobatan yang bisa sangat mahal dan memakan waktu. Saat ini, sistem layanan kesehatan yang canggih semakin memfokuskan perhatiannya pada pencegahan dibandingkan pengobatan. Hal ini sangat mengurangi biaya dan meningkatkan hasil pasien. Pembelajaran mesin digunakan untuk menilai rekam medis seseorang dan memprediksi kemungkinan mereka terkena kondisi tertentu seperti penyakit jantung atau diabetes, seringkali bertahun-tahun sebelumnya. Individu yang berada di urutan teratas; yaitu

mereka yang diprediksi oleh model paling mungkin tertular penyakit, dihubungkan dengan maksud untuk memulai tindakan pencegahan. Misalnya saja melakukan perubahan gaya hidup atau mengonsumsi obat pencegahan.

Contoh terakhir penerapan pembelajaran mesin adalah menentukan jenis artikel berita (dan artikel lainnya) yang akan direkomendasikan kepada orang-orang. Penyedia media sosial menggunakan pembelajaran mesin untuk menganalisis artikel apa yang pernah Anda baca sebelumnya dan jenis topik yang Anda diskusikan dengan teman. Hal ini kemudian mendorong konten yang mereka promosikan kepada Anda.

Itu hanya beberapa cara penggunaan pembelajaran mesin. Saat ini, pembelajaran mesin mendukung sejumlah besar aplikasi. Faktanya, hampir semua aspek kehidupan melibatkan pengambilan keputusan dalam satu atau lain bentuk. Algoritme yang mencocokkan orang-orang di situs kencan, teknologi yang digunakan untuk mendeteksi penipuan kartu kredit, dan sistem untuk mengidentifikasi tersangka teroris, semuanya menggunakan model prediktif yang diperoleh dari pembelajaran mesin.

Hal ini membawa kita pada pertanyaan tentang apa yang dimaksud dengan Kecerdasan Buatan atau AI. Ada banyak definisi yang beragam tentang apa itu AI, dan seperti analisis prediktif, banyak orang menggunakan istilah AI dan pembelajaran mesin secara bergantian. Dalam lingkup penelitian AI secara keseluruhan, pembelajaran mesin adalah bidang studi utama, namun masih banyak bidang studi lainnya. Kecerdasan buatan yang sebenarnya lebih dari sekadar pengenalan pola dan prediksi. Beberapa ahli juga mempertanyakan apakah AI yang sebenarnya dapat dicapai hanya dengan mengikuti pendekatan “Brute force” yaitu mengembangkan algoritme yang semakin kompleks menggunakan perangkat keras komputer yang semakin canggih. Apakah ada elemen tambahan (yang belum diketahui) yang diperlukan untuk kecerdasan dan kesadaran diri seperti manusia yang tidak dapat direplikasi hanya melalui komputasi?

Jadi, di satu sisi, tidak tepat untuk mengatakan bahwa pembelajaran mesin dan kecerdasan buatan adalah hal yang sama. Namun, dalam praktiknya, hampir semua sistem AI yang digunakan saat ini sangat bergantung pada pembelajaran mesin. Oleh karena itu, untuk keperluan buku ini, definisi kerja sederhana tentang AI yang harus kita patuhi adalah:

Kecerdasan Buatan (AI) adalah replikasi kemampuan analitis dan/atau pengambilan keputusan manusia. Aplikasi AI yang baik adalah aplikasi yang dapat bekerja sama baik atau lebih baik dari rata-rata orang ketika menghadapi tugas sehari-hari. Misalnya, kemampuan mengidentifikasi orang-orang dari foto Facebook mereka, kemampuan menilai kelayakan kredit seseorang dengan lebih akurat dibandingkan penjamin emisi yang berpengalaman, kemampuan mengalahkan pemain Go dan catur terbaik, atau kemampuan lebih baik dalam mengenali tanda-tanda kanker melalui pemindaian medis daripada ahli radiologi ahli.

Pada satu tingkat, aplikasi AI bisa tampak ajaib bagi orang awam. Namun, seperti kebanyakan hal, begitu Anda masuk ke dalam kap mesin, mistiknya akan menguap. Dalam praktiknya, semua berita yang kita dengar tentang penerapan “AI” yang menakutkan sebenarnya hanyalah penerapan machine learning yang sangat canggih.

Kesalahan utama yang harus dihindari adalah berpikir bahwa aplikasi AI saat ini sama cerdasnya dengan kesadaran manusia. Tentu saja, mereka sangat kompleks, sangat pintar, dan kadang-kadang bisa terlihat sangat hidup, tetapi pada akhirnya, itu semua hanyalah matematika. Sebagian besar ahli sepakat bahwa kita masih membutuhkan waktu bertahun-tahun untuk dapat menciptakan sebuah mesin yang memiliki kesadaran diri seperti manusia, atau yang dapat dianggap sebagai manusia dari hari ke hari. Bukan berarti tidak ada chatbot yang bagus di luar sana!

Semua aplikasi AI yang digunakan saat ini adalah apa yang oleh industri disebut sebagai AI Sempit. Mereka sangat baik dalam berperilaku cerdas bila diterapkan pada satu bidang keahlian yang jelas. Namun, sistem ini sangat jauh dari General AI. General AI adalah sistem yang dapat belajar dan bertindak secara cerdas di berbagai lingkungan dan masalah dengan cara yang mirip dengan manusia. Aplikasi AI yang digunakan untuk mendeteksi penghindaran pajak misalnya, tidak berguna dalam mendeteksi tanda-tanda kanker dari scan medis. Namun, seseorang dapat belajar melakukan kedua tugas ini jika mereka diberi pelatihan yang sesuai. Demikian pula, sistem seperti Google Terjemahan sangat bagus dalam memahami kata-kata yang diucapkan tetapi tidak akan banyak berguna dalam menilai apakah seseorang di situs kencan mungkin cocok dengan Anda.

Komponen inti yang mendorong sebagian besar aplikasi AI/pembelajaran mesin adalah:

- **Masukan data.** Ini bisa berupa masukan sensorik dari kamera (mata), mikrofon (telinga) atau sumber lainnya. Ini juga mencakup data yang telah diproses sebelumnya seperti informasi yang diperoleh ketika seseorang mengisi formulir secara online, rincian tentang apa yang telah dibeli seseorang menggunakan kartu kreditnya, atau riwayat kredit seseorang yang diberikan oleh agen referensi kredit.
- **Pemrosesan (pra) data.** Input data mentah perlu diproses menjadi format standar “ramah komputer” sebelum siap digunakan.
- **Model prediktif.** Ini dihasilkan oleh proses pembelajaran mesin yang menggunakan pengalaman masa lalu; yaitu sejumlah besar data historis. Data yang telah diproses sebelumnya untuk kasus-kasus baru dimasukkan ke dalam model untuk menghasilkan prediksi baru di masa mendatang.
- **Aturan pengambilan keputusan (rule set).** Prediksi saja tidak ada gunanya. Anda harus memutuskan bagaimana menggunakannya. Aturan pengambilan keputusan digunakan bersama dengan masukan data dan skor dari model prediktif untuk memutuskan apa yang harus dilakukan. Terkadang aturan ini diturunkan secara otomatis oleh algoritme pembelajaran mesin, namun sering kali aturan tersebut menyertakan aturan tambahan yang ditentukan oleh pakar manusia/pengguna bisnis.
- **Respon/keluaran.** Tindakan perlu diambil berdasarkan keputusan yang telah dibuat. Jika keputusannya adalah seseorang layak mendapat kredit, maka kartu kredit perlu diterbitkan. Jika keputusannya adalah seseorang harus dipekerjakan, maka mereka perlu dikirim surat penawaran, diberikan kontrak untuk ditandatangani dan sebagainya.

Kombinasi dari masing-masing komponen inilah yang memberi kita “AI”. Apa yang membuat beberapa aplikasi AI tampak begitu pintar adalah kompleksitas algoritma yang mendukungnya, dikombinasikan dengan antarmuka pengguna yang apik untuk mengumpulkan data dan memberikan respons yang diperlukan dengan cara yang ramah manusia. Gabungkan komponen-komponen ini dengan mesin industri generasi terbaru, atau integrasikan ke dalam mobil dan kendaraan lain, dan kita akan mendapatkan robot yang dapat berinteraksi dengan lingkungannya dan berinteraksi dengan kita dengan cara yang sangat manusiawi.

Mari kita mulai dengan mempertimbangkan aplikasi pemasaran AI untuk perusahaan minuman. Aplikasi ini mengambil informasi yang dikumpulkan tentang individu dari jejaring sosial dan memasukkannya ke dalam model prediktif untuk menentukan seberapa besar kemungkinan mereka membeli wiski merek tertentu. Sistem kemudian menerapkan sejumlah aturan untuk memutuskan apakah seseorang harus dipasarkan. Aturan yang mungkin ada adalah:

1. Jika model prediktif memperkirakan peluang mereka membeli wiski lebih dari 90%, maka jangan lakukan apa pun. Mereka mungkin akan tetap membeli wiski itu.
2. Jika model prediktif memperkirakan bahwa peluang mereka membeli wiski adalah antara 1% dan 90%, kirimkan mereka kupon diskon Rp. 50.000 untuk mencoba menjadikan wiski tersebut sebagai penawaran yang lebih menarik; yaitu mempengaruhi perilaku pelanggan untuk meningkatkan peluang mereka membeli.
3. Jika model prediktif memperkirakan kemungkinan mereka membeli wiski kurang dari 1%, maka jangan lakukan apa pun. Mereka mungkin tidak akan membeli wiski apa pun yang Anda tawarkan. Oleh karena itu, tidak ada gunanya mencoba membujuk mereka.

Jadi, aturan-aturan ini akan diturunkan berdasarkan beberapa jenis analisis biaya-manfaat, di mana batas 1% dan 90% dianggap sebagai tingkat optimal untuk memicu aktivitas pemasaran. Namun, aturan bisnis lainnya juga akan berlaku, seperti:

1. **JANGAN PERNAH** menawarkan untuk menjual wiski kepada anak-anak, tidak peduli apa kecenderungan mereka untuk membelinya!
2. **JANGAN** mengirimkan penawaran kepada orang yang memiliki riwayat ketergantungan alkohol.

Kedua kelompok yang dimaksud dalam aturan ini akan berisi banyak orang yang ingin membeli wiski; yaitu >1% kemungkinannya, namun dari sudut pandang etika, menargetkan anak-anak atau orang-orang dengan masalah alkohol sulit untuk diperdebatkan. Dari sudut pandang yang murni berorientasi pada keuntungan, pemasaran kepada anak-anak kemungkinan besar merupakan tindakan ilegal dan menargetkan individu-individu tersebut dapat menghasilkan publisitas negatif dalam jumlah besar. Kedua aturan ini adalah contoh bagus mengapa keahlian manusia diperlukan untuk mendukung sistem berbasis pembelajaran mesin otomatis, terutama ketika sistem digunakan untuk membuat keputusan yang berisiko atau kontroversial mengenai manusia.

Salah satu contoh algoritma yang menjadi liar dan banyak dipublikasikan adalah kasus YouTube dan kebijakan penempatan iklannya. Pada tahun 2017, banyak organisasi besar

menarik iklan mereka dari YouTube. Hal ini karena YouTube ditemukan menempatkan beberapa iklannya di samping materi dari teroris dan sumber tidak menyenangkan lainnya. Algoritme pembelajaran mesin YouTube-lah yang menentukan iklan mana yang akan ditempatkan di tempat yang menyebabkan masalah. Akibatnya, YouTube harus melakukan peninjauan besar-besaran terhadap proses penempatan iklannya. Akibatnya, beberapa bulan kemudian pada tahun 2018, mereka memutuskan untuk kembali ke proses pemeriksaan manual. Setiap klip video harus ditinjau dan disetujui oleh orang sungguhan sebelum dimasukkan ke dalam layanan YouTube yang memasang pengiklan dengan konten populer.

Pertanyaan yang wajar untuk ditanyakan adalah mengapa YouTube membutuhkan waktu hampir satu tahun untuk menemukan solusinya? Alasannya tidak dapat dipastikan, namun jawaban yang sangat masuk akal adalah bahwa mereka menghabiskan banyak waktu untuk mencoba menyelesaikan masalah hanya dengan menggunakan pendekatan otomatis (machine learning/AI), sebelum menyadari bahwa mereka perlu memelihara manusia. elemen dalam proses penilaian.

Hal terakhir yang diperlukan dari sistem AI adalah mengeluarkan respons yang diperlukan. Dalam contoh wiski, ini adalah mengirimkan kode diskon atau kupon yang sesuai kepada individu yang menjadi sasaran. Artinya, untuk mencapai tindakan yang diperlukan, aplikasi AI harus terhubung dengan saluran pemasaran organisasi. Jika AI memberi tahu Anda bahwa respons yang tepat adalah mengirimkan SMS kepada jutaan pelanggan yang berisi penawaran, maka AI tersebut perlu terhubung ke sistem yang dapat mengirim SMS. Jika teks, kupon, pop-up internet, atau apa pun digunakan untuk mengkomunikasikan penawaran, maka masih diperlukan seseorang untuk merancang kata-kata dan format komunikasi dengan cara yang menarik bagi pelanggan dan sesuai dengan relevansi. undang-undang pemasaran.

Sebagian besar aplikasi AI sangat bergantung pada prediksi, didorong oleh pengenalan pola (yaitu pembelajaran mesin/analisis prediktif). Ambil contoh sistem pengenalan suara yang mendukung alat seperti Siri dari Apple dan Echo dari Amazon. Setelah mendengar suara, sistem pengenalan suara memproses suara tersebut menjadi sejumlah potongan suara standar. Ia kemudian akan mencoba mengidentifikasi kata apa yang menurutnya telah diucapkan. Hal ini didasarkan pada serangkaian probabilitas (skor) yang dihasilkan oleh model prediktif yang dibuat menggunakan pembelajaran mesin. Jika sistem menghitung bahwa ada 5% kemungkinan yang terdengar adalah "Halo", 20% adalah "Jell-O" dan 75% adalah "Mellow", maka sistem akan membuat tebakan bahwa kata yang diucapkan adalah "Mellow."

Sistem AI tingkat lanjut menggunakan kombinasi lapisan model prediktif yang kompleks untuk merumuskan pandangan tentang apa yang terjadi di sekitarnya, berdasarkan masukan yang diterima. Lapisan berikutnya dari sistem pengenalan suara akan mempertimbangkan frasa dan struktur kata. Elemen prediktif sistem tersebut kemudian digabungkan dengan aturan keputusan (logika keputusan) untuk menentukan tindakan apa yang harus diambil ketika kata atau frasa tertentu diucapkan.

Salah satu kesalahpahaman yang umum adalah bahwa pembelajaran mesin dan kecerdasan buatan adalah hal yang benar-benar baru dan baru memasuki arus utama dalam

beberapa tahun terakhir. Kenyataannya adalah penelitian terhadap aplikasi bisnis pembelajaran mesin pertama kali dilakukan pada awal tahun 1940an dan pembelajaran mesin telah digunakan dalam industri jasa keuangan sejak tahun 1950an (penilaian kredit). Teknik pembelajaran mesin kemudian diterapkan pada bidang-bidang seperti pemasaran langsung, pengumpulan pajak, dan penjaminan asuransi tidak lama setelah itu. Perkembangan terkini adalah penerapan alat jenis ini di rumah, di ponsel cerdas, dan perangkat lainnya.

Beberapa pakar pembelajaran mesin/AI mungkin membantah pandangan sejarah ini, dengan alasan bahwa hanya dengan kemajuan teknologi terkini pembelajaran mesin dapat berkembang dengan sendirinya. Namun, jika definisi pembelajaran mesin/AI mencakup penggunaan algoritme untuk mengidentifikasi pola perilaku yang mendorong sistem pengambilan keputusan otonom, dan sistem ini secara signifikan mengungguli pakar manusia, maka sistem penilaian kredit pada tahun 1950-an memenuhi definisi tersebut.

Yang berubah belakangan ini adalah tiga hal. Pertama adalah terminologinya. Pada tahun 1980-an/1990-an, apa yang sekarang disebut pembelajaran mesin/AI diberi nama “Data Mining” atau “statistik komputasi”. Teori yang mendasari pendekatan pembelajaran mesin yang paling umum digunakan saat ini juga dikembangkan sebelum atau selama tahun 1990an (walaupun banyak yang telah disempurnakan dan diperluas sejak saat itu).

Hal kedua yang berubah adalah perangkat keras dan perangkat lunak komputer yang mendukung pembelajaran mesin. Peralihan ke layanan online melalui ponsel pintar, tablet, dan internet berarti kini terdapat lebih banyak informasi yang tersedia tentang manusia dibandingkan sebelumnya, dan jumlah informasi tersebut terus bertambah dari tahun ke tahun. Kecepatan dan kemampuan penyimpanan komputer juga meningkat pesat. Hal ini memungkinkan pemrosesan data dalam jumlah besar dengan cepat dan hemat biaya. Demikian pula, kompleksitas algoritme pembelajaran mesin dan model prediktif yang dihasilkannya telah meningkat berkali-kali lipat hanya dalam beberapa tahun.

Sejalan dengan kemajuan dalam desain perangkat keras dan algoritme, kini terdapat banyak perangkat lunak yang tersedia untuk pembelajaran mesin. Sebagian besar gratis dan/atau open source (seperti bahasa pemrograman R dan Python) dan tersedia melalui PC, laptop, atau layanan cloud yang ditawarkan oleh raksasa IT seperti Microsoft dan Google.

Hal ketiga yang berubah adalah cara kita berinteraksi dengan model yang mendasari dan aturan keputusan. Kita tidak lagi harus mengetikkan instruksi yang sangat kaku pada keyboard atau harus menafsirkan halaman-halaman hasil matematika yang terperinci. Sebaliknya, sistem AI modern berinteraksi dengan kita dengan cara yang lebih mirip manusia.

Seringkali antarmuka antara manusia dan komputer didasarkan pada prinsip pembelajaran mesin, yang didorong oleh model prediktif. Jika kita kembali ke contoh penjualan wiski, saya dapat menggunakan pembelajaran mesin dalam desain chatbot untuk berbicara dengan orang-orang di media sosial tentang kebiasaan minum mereka. Data yang dikumpulkan oleh chatbot kemudian melengkapi data lain yang telah saya kumpulkan tentang orang-orang. Jika aplikasi wiski menentukan bahwa masyarakat harus diberikan penawaran diskon, maka penawaran ini dapat disampaikan kembali ke pelanggan melalui chatbot dalam

bentuk percakapan, dibandingkan dengan pendekatan tradisional yang mengirimkan mereka kupon atau kode diskon melalui teks atau email.

Prinsip yang sama berlaku untuk asisten pribadi digital seperti Amazon's Echo. Satu set algoritma digunakan untuk menerjemahkan apa yang Anda katakan ke dalam soundbites (pra-pemrosesan data). Soundbite yang telah diproses sebelumnya kemudian dimasukkan ke dalam rangkaian model prediktif lain yang mencoba memprediksi hasil apa yang Anda minta; seperti memutar lagu tertentu atau menanyakan cuaca. Amazon mengizinkan pengembang untuk menggunakan kemampuan pra-pemrosesan ini secara gratis – memungkinkan siapa saja untuk memasukkan pengenalan suara tingkat lanjut ke dalam program dan aplikasi mereka.

Artinya, biaya untuk memasuki dunia pembelajaran mesin dan AI telah berkurang secara signifikan. Beberapa tahun yang lalu, pembelajaran mesin hanya menjadi domain bisnis besar dan institusi akademis. Saat ini, siapa pun yang memiliki laptop dan memiliki waktu luang dapat terlibat dan mulai membuat model prediktif dengan relatif cepat dan hampir tanpa biaya, lalu mengembangkan antarmuka pengguna tingkat lanjut untuk mengimplementasikan model tersebut sebagai “aplikasi AI”.

Saat ini, bahkan ada situs kompetisi untuk pembelajaran mesin seperti Kaggle. Organisasi menyediakan datanya secara gratis, lalu tim amatir dan profesional bersaing untuk melihat siapa yang dapat membangun model dengan kekuatan prediksi terbesar menggunakan data yang telah disediakan. Pemenang membagikan pendekatan mereka dengan organisasi yang menyediakan data tersebut dan menerima hadiah sebagai imbalannya. Imbalan untuk memenangkan kompetisi Kaggle sangatlah signifikan. Hadiah uang tunai pada umumnya berjumlah beberapa ribu dolar, namun hadiah terbesar melebihi satu juta.

Saat membahas pembelajaran mesin, pemodelan prediktif, dan AI, Anda mungkin juga akan menemukan istilah ilmu data dan ilmuwan data. Ilmu data bukanlah ilmu seperti halnya fisika, kimia, atau biologi. Ini hanyalah deskripsi penerapan teknik pembelajaran mesin. Begitu pula dengan data scientist adalah seseorang yang dapat menerapkan pembelajaran mesin secara praktis. Ini lebih dari sekedar aspek matematika dari subjeknya. Seorang data scientist yang baik juga memiliki pemahaman tentang data dan sistem TI serta mampu memahami konteks bisnis di mana solusi mereka akan diterapkan.

BAB 3

NILAI SKOR YANG DIHASILKAN OLEH MODEL PREDIKTIF

Seperti disebutkan dalam bab sebelumnya, sebagian besar aplikasi pembelajaran mesin/AI yang digunakan dalam bisnis mengandalkan model prediktif. Model-model ini menghasilkan angka-angka (skor) yang menunjukkan apa yang mungkin terjadi, berdasarkan informasi yang disajikan dalam model tersebut.

Skor yang dihasilkan oleh model prediktif dapat terdiri dari dua jenis:

1. **Probabilitas (kemungkinan) skor.** Ini memprediksi kemungkinan terjadinya peristiwa tertentu. Misalnya, akankah seseorang melakukan sesuatu atau tidak? Nama teknis model prediktif jenis ini adalah model klasifikasi.
2. **Besaran (kuantitas) skor.** Ini memprediksi jumlah atau ukuran sesuatu. Misalnya, berapa banyak uang yang akan dibelanjakan seseorang di toko, atau berapa lama sebelum mereka melakukan sesuatu. Nama teknis model prediktif jenis ini adalah model regresi.

Model klasifikasi memprediksi kemungkinan terjadinya suatu peristiwa. Seringkali peristiwa-peristiwa ini bersifat biner; yaitu, skor tunggal dihasilkan, yang mewakili kemungkinan terjadinya peristiwa tersebut. Namun, model klasifikasi juga dapat digunakan untuk mengatasi masalah dengan banyak hasil (peristiwa).

Aplikasi bisnis umum dari model klasifikasi meliputi:

- **Diagnosa medis.** Skor tersebut merupakan ukuran kemungkinan seseorang mengidap atau terkena suatu penyakit tertentu.
- **Deteksi konten.** Model tersebut memprediksi apakah konten media berkaitan dengan subjek/topik tertentu. Model jenis ini digunakan oleh perusahaan media sosial untuk mengidentifikasi materi yang menyinggung atau ilegal yang diposting ke situs mereka. Materi dengan skor tinggi diblokir atau ditempatkan dalam antrian untuk ditinjau oleh penilai manusia.
- **Pengurangan pelanggan.** Model tersebut menghasilkan kemungkinan bahwa pelanggan akan membeli produk pesaing atau berhenti menggunakan suatu produk atau layanan. Misalnya, mengganti penyedia utilitas, membeli kartu kredit baru, atau membatalkan langganan majalah.
- **Deteksi penipuan.** Skor yang dihasilkan oleh model mewakili kemungkinan terjadinya transaksi penipuan. Misalnya, seseorang mencoba membeli sesuatu dengan kartu kredit curian.
- **Kompatibilitas kencan.** Skor model memberikan indikasi seberapa besar kemungkinan dua orang cocok berkencan.
- **Rekomendasi produk.** Skor tersebut mewakili kemungkinan seseorang akan membeli produk atau layanan tertentu. Pelanggan hanya dipasarkan jika kemungkinan mereka membeli produk atau jasa cukup tinggi.

- **Retensi staf.** Model tersebut memperkirakan kemungkinan seorang karyawan akan meninggalkan posisinya saat ini untuk mengambil posisi lain.
- **Kerusakan mesin.** Skor tersebut memprediksi apakah suatu mesin kemungkinan besar akan gagal dalam waktu dekat. Hal ini memungkinkan dilakukannya tindakan pencegahan seperti mengirim SMS kepada pengemudi bahwa mereka harus datang ke bengkel untuk memperbaiki mobilnya sebelum rusak.

Terkadang skor dari model klasifikasi mewakili probabilitas secara langsung. Seseorang yang mendapat skor 0,0 tentu tidak akan mengalami kejadian tersebut. Mereka yang mendapat skor 1,0 pasti akan melakukannya. Model prediktif tidak memberikan prediksi yang sempurna. Apa yang cenderung dilihat adalah penyebaran (distribusi) probabilitas antara 0,0 dan 1,0 untuk individu dalam populasi. Ketika model rekomendasi produk menghasilkan skor 0,70 untuk seseorang, ini berarti model tersebut memperkirakan bahwa terdapat 70% kemungkinan individu tersebut akan membeli produk dan 30% kemungkinan tidak akan membeli.

Dalam situasi lain, skor dari model klasifikasi diubah ke skala tertentu. Jika kita kembali ke dunia penilaian kredit, skor kredit cenderung berskala berkisar antara 400 hingga sekitar 800. Jika seseorang menerima skor 400 berarti kecil kemungkinannya dia akan membayar kembali kredit yang diberikan kepadanya (<5% kemungkinan pinjaman akan dilunasi). Skor 800 berarti mereka hampir pasti akan melunasi utangnya (>99,9% kemungkinan pinjaman akan dilunasi).

Kebanyakan model klasifikasi yang digunakan dalam bisnis menghasilkan skor tunggal. Skor tersebut memprediksi kejadian tipe ya/tidak yang sederhana, seperti yang dijelaskan dalam daftar di atas. Jika pembelajaran mesin diterapkan pada masalah dengan banyak hasil, maka skor terpisah akan dihasilkan untuk setiap kemungkinan hasil.

Secara ekstrim, beberapa model prediksi paling canggih yang digunakan saat ini menghasilkan ribuan skor. Setiap skor mewakili probabilitas dari satu kemungkinan hasil. Sistem pengenalan objek, yang dirancang untuk mengidentifikasi objek sehari-hari dalam gambar, adalah salah satu contohnya. Sistem menghasilkan skor (probabilitas) untuk setiap objek yang telah dilatih untuk diidentifikasi. Skor 1 memprediksi peluang 4% bahwa itu adalah gambar mesin cuci, Skor 2 memperkirakan peluang 6% bahwa itu adalah payung, Skor 3 memperkirakan peluang 80% bahwa itu adalah kursi dan seterusnya. Skor yang terkait dengan "Kursi" adalah yang tertinggi dan oleh karena itu sistem akan "Menebak" bahwa benda tersebut adalah kursi.

Untuk aplikasi kompleks seperti pengenalan objek dan ucapan, pengguna bisnis tidak akan melihat skor yang dihasilkan secara langsung. Sebaliknya, seperangkat aturan pengambilan keputusan akan diterapkan untuk menerjemahkan skor menjadi sesuatu yang lebih bermakna, dengan hasil yang disampaikan melalui antarmuka yang "ramah pelanggan".

Model regresi di sisi lain adalah tentang kuantitas; yaitu besarnya suatu peristiwa. Beberapa contoh penerapan model regresi dalam bisnis adalah:

- **Harapan hidup.** Skor tersebut memberikan prediksi berapa lama seseorang diperkirakan akan hidup. Model jenis ini sering digunakan untuk menetapkan premi asuransi jiwa dan anuitas pensiun.
- **Waktu perjalanan.** Mengingat informasi tentang kebiasaan mengemudi Anda, lokasi saat ini dan kondisi mengemudi, kapan Anda akan tiba di tujuan?
- **Kerugian kredit.** Ketika seseorang gagal membayar pinjamannya, berapa besar kemungkinan utangnya akan dihapuskan?
- **Membelanjakan.** Skor model tersebut memprediksi berapa banyak kemungkinan seseorang akan berbelanja di supermarket lokalnya dalam 12 bulan ke depan.
- **Interval pembelian.** Berapa lama setelah membeli suatu produk pembelian berulang dapat diharapkan? Aktivitas pemasaran kemudian dapat diatur waktunya agar terjadi sesaat sebelum pelanggan diperkirakan akan membeli produk tersebut lagi.
- **Durasi panggilan.** Berapa lama panggilan telepon diperkirakan akan berlangsung? Hal ini membantu perencanaan sumber daya di pusat panggilan.
- **Hunian.** Memprediksi jam berapa seseorang mungkin berada di rumah dan waktu terbaik untuk mengunjunginya.
- **Waktu merespon.** Berapa lama setelah tindakan diambil, respons terlihat? Setelah mengirim surat, email, atau SMS kepada seseorang, berapa lama waktu yang dibutuhkan untuk membalasnya?

Dalam beberapa kasus, skor dari model regresi merupakan perkiraan langsung dari apa yang ingin diprediksi. Model pembelian yang menghasilkan skor 5.981 berarti pelanggan diperkirakan akan membelanjakan Rp. 5.981.000 Dalam situasi lain, skor diskalakan untuk mewakili persentil atau untuk mengalokasikan nilai dari 1 hingga 10, A hingga E, atau serupa. Kelas A mewakili pelanggan terbaik (paling menguntungkan), Kelas E mewakili pelanggan yang paling tidak menguntungkan.

Baik model regresi maupun klasifikasi banyak digunakan, namun model klasifikasi cenderung menjadi yang paling populer, khususnya untuk aplikasi bisnis umum. Hal ini karena peristiwa yang sederhana akan/tidak akan mereka ketik biasanya paling mudah dipahami oleh organisasi. Ada juga cenderung lebih sedikit ambiguitas dalam mendefinisikan masalah. Hal ini mempermudah data scientist untuk menerjemahkan tujuan bisnis ke dalam representasi numerik yang sesuai sehingga proses pembelajaran mesin dapat diterapkan. Mendefinisikan apakah pelanggan membeli wiski atau tidak cukup mudah (klasifikasi). Namun, menentukan profitabilitas penjualan wiski individu jauh lebih kompleks (regresi).

Jenis model yang digunakan (regresi atau klasifikasi) sangat bergantung pada apa yang ingin dilakukan seseorang dengan skor model tersebut. Jika yang saya minati hanyalah mengidentifikasi pelanggan kartu kredit mana yang paling mungkin berpindah ke pesaing, maka yang saya perlukan adalah model klasifikasi untuk memprediksi kemungkinan pengurangan pelanggan. Strategi retensi (seperti menawarkan diskon atau hadiah gratis) kemudian diterapkan untuk mencoba mempertahankan pelanggan yang kemungkinan besar akan membelot.

Sebaliknya, jika perhatian utama saya adalah mengidentifikasi pelanggan yang menguntungkan dengan tujuan mendorong mereka membelanjakan lebih banyak, maka yang saya perlukan adalah model regresi. Model ini dirancang untuk memprediksi siapa yang akan menjadi pembelanja terbesar dan menargetkan mereka. Saya tidak perlu repot dengan basis pelanggan lainnya.

Namun, ada pilihan ketiga. Katakanlah yang benar-benar ingin saya lakukan adalah mempertahankan pelanggan yang menguntungkan, dan saya tidak peduli kehilangan pelanggan yang tidak menguntungkan yang tidak menghabiskan banyak uang untuk kartunya. Dalam situasi ini, pendekatan terbaik mungkin adalah dengan membangun dua model prediktif terpisah sebagai berikut:

1. Model atrisi untuk memprediksi pelanggan yang paling mungkin berpindah ke produk pesaing (Klasifikasi).
2. Model pendapatan untuk mengidentifikasi pelanggan mana yang paling banyak mengeluarkan uang (Regresi).

Kedua model tersebut kemudian digunakan bersama-sama untuk mendorong strategi retensi seperti yang ditunjukkan pada Gambar 3.1.

		Kemungkinan membelot ke produk pesaing (Skor dari model atrisi)		
		Rendah	Sedang	Tinggi
Perkiraan pembelajaran (Skor dari model pendapatan)	Rendah	Tidak melakukan apapun	Tidak melakukan apapun	Tidak melakukan apapun
	Sedang	Tidak melakukan apapun	Tidak melakukan apapun	Terapkan strategi retensi 1
	Tinggi	Tidak melakukan apapun	Terapkan strategi retensi 1	Terapkan strategi retensi 2

Gambar 3.1 Menggunakan dua model yang digabungkan

Gambar 3.1 menggunakan skor dari masing-masing model untuk mengelompokkan populasi berdasarkan pendapatan dan pengurangan. Dengan menggabungkan kedua model tersebut, serangkaian strategi retensi yang disesuaikan dapat diterapkan dan penghematan biaya yang signifikan dapat dicapai. Hal ini karena sumber daya tidak dihabiskan untuk mencoba mempertahankan pelanggan yang tidak akan melakukan pengurangan atau pelanggan yang memiliki kemungkinan besar akan pengurangan, namun tidak menghabiskan banyak uang untuk kartu mereka.

Sebagai contoh pada Gambar 3.1, tiga strategi diferensial diterapkan:

1. Jangan melakukan apa pun untuk pelanggan “terburuk”. Tidaklah efektif dari segi biaya untuk mencoba dan mempertahankannya.
2. Melakukan strategi retensi yang ringan (yaitu berbiaya rendah) bagi mereka yang berada di tingkat menengah (strategi 1).
3. Menghabiskan upaya maksimal pada kelompok pelanggan “terbaik” (strategi 2).

Tentu saja ada rumusan alternatif untuk masalah ini. Seseorang dapat membangun model prediktif tunggal yang hanya mencoba memprediksi pelanggan yang menghasilkan pendapatan tertinggi yang juga cenderung melakukan attrite, dan ini kemudian ditargetkan. Hal ini tentunya merupakan pendekatan yang masuk akal, namun hal ini akan mengakibatkan hilangnya rincian mengenai sifat pelanggan; yaitu model tunggal tidak akan memungkinkan Anda untuk mengeksplorasi trade-off antara pengurangan pendapatan dan pendapatan yang disediakan oleh pendekatan model kembar, seperti yang diilustrasikan pada Gambar 3.1.

Perhatikan bahwa model hanya memberikan prediksi tentang apa yang akan dilakukan pelanggan. Mereka tidak memberi tahu Anda apa yang harus dilakukan berdasarkan prediksi tersebut. Terkait dengan aktivitas apa yang harus dilakukan untuk strategi retensi 1 dan 2, hal ini merupakan pertimbangan tambahan yang perlu ditangani organisasi dalam keseluruhan cakupan proyek pembelajaran mesinnya. Hal ini penting karena kegagalan yang umum terjadi adalah organisasi membangun beberapa model prediktif yang sangat baik, dan kemudian tidak menggunakannya untuk apa pun. Membuat keputusan yang tepat dan kemudian menindaklanjuti keputusan tersebut sangatlah penting jika proyek pembelajaran mesin ingin berhasil.

BAB 4

MENGAPA MENGGUNAKAN PEMBELAJARAN MESIN?

Penerapan pembelajaran mesin, berdasarkan model prediktif, semakin banyak digunakan untuk menggantikan dan/atau melengkapi penilaian ahli dan pengambilan keputusan manual di berbagai bidang. Ini karena model prediktif cenderung:

- **Lebih akurat daripada pakar manusia.** Model prediktif terkadang memberikan prediksi yang salah seperti halnya manusia, namun banyak bukti yang menunjukkan bahwa rata-rata model tersebut tidak terlalu sering melakukan kesalahan. Di banyak bidang, model prediktif secara konsisten mengungguli pakar manusia sebesar 20-30%; yaitu mereka membuat kesalahan 20-30% lebih sedikit atau mengidentifikasi kasus-kasus 20-30% lebih baik (lebih penting/lebih menguntungkan).
- **Tidak memihak.** Berbeda dengan manusia, model prediktif tidak menampilkan bias prasangka yang disengaja terhadap seseorang karena jenis kelamin, ras, disabilitas, dll. Jangan salah paham, model prediktif memang menampilkan bias, namun jika dirancang dengan benar model tersebut hanya akan memberikan individu atau kelompok tertentu skor yang lebih tinggi atau lebih rendah dibandingkan populasi pada umumnya jika hal ini didasarkan pada bukti statistik yang kuat – hal ini tidak didasarkan pada prasangka atau stereotip yang tidak berdasar.
- **Cepat.** Sebagai bagian dari sistem pengambilan keputusan otomatis, model prediktif dapat digunakan untuk memprediksi perilaku jutaan individu dalam hitungan detik. Dalam kebanyakan kasus, akan memakan biaya dan waktu yang sangat mahal jika meminta orang membuat penilaian yang sama secara manual.
- **Murah.** Setelah dikembangkan, model prediktif seringkali lebih murah untuk diterapkan dibandingkan model manusia. Meskipun kita tidak boleh melupakan biaya di muka yang terkait dengan pengembangan awal, ditambah biaya pemeliharaan dan perizinan yang berkelanjutan.

Singkatnya, model prediktif dapat digunakan untuk membuat keputusan yang lebih baik, lebih murah dan lebih cepat dibandingkan dengan yang dibuat oleh manusia yang ahli. Artinya, di banyak bidang, model prediktif telah menggantikan manusia. Khususnya, peran kerah putih yang memerlukan staf yang sangat terlatih untuk menggunakan penilaian ahli mereka untuk memutuskan apa yang harus dilakukan.

Ini adalah fenomena yang dimulai di Amerika pada tahun 1950an dan 1960an di bidang jasa keuangan. Ini adalah masa ketika sebagian besar permohonan hipotek, kredit mobil, dan pinjaman pribadi dinilai oleh penjamin emisi yang secara individual meninjau setiap permohonan pinjaman. Mereka kemudian mengambil keputusan apakah orang tersebut layak mendapatkan kredit atau tidak, dan oleh karena itu, harus mendapatkan kredit yang telah mereka ajukan.

Dalam organisasi besar, seluruh gedung perkantoran ditempati oleh tim penjamin emisi yang menghabiskan hari-hari mereka untuk membuat keputusan pemberian pinjaman. Di cabang-cabang bank lokal, Manajer Bank adalah raja – seorang anggota masyarakat lokal yang dihormati dan memiliki keleluasaan mutlak dalam menentukan siapa yang boleh diberi pinjaman atau hipotek dan siapa yang tidak. Kadang-kadang mendapatkan pinjaman hanya tergantung pada apakah dia (dan selalu dia) sedang dalam suasana hati yang baik hari itu. Anda mungkin lebih mungkin mendapatkan pinjaman pada hari Jumat sore, ketika suasana hatinya sedang baik di akhir pekan, dibandingkan pada hari Senin pagi!

Pada akhir tahun 1980-an, sebagian besar peran ini sudah mubazir. Peran manajer bank telah direduksi menjadi lebih dari sekadar tenaga penjualan – yang menangani hubungan pelanggan namun tidak memiliki kemampuan mengambil keputusan. Sebagian besar keputusan pemberian pinjaman dibuat dengan model penilaian kredit yang merupakan inti dari sistem pengambilan keputusan otomatis yang berlokasi di kantor pusat. Dalam dunia pemberian kredit saat ini, hanya kasus-kasus yang tidak biasa atau berada di ambang batas yang cenderung ditinjau oleh manusia – sebagian besar keputusan pemberian pinjaman dibuat secara otomatis tanpa campur tangan manusia sama sekali.

Kehilangan pekerjaan merupakan kekhawatiran bagi mereka yang berada di industri yang terkena dampaknya, namun hal ini tidak berarti malapetaka dan kesuraman. Dalam banyak kasus, pembelajaran mesin memberikan nilai tambah pada peran, bukan menggantikannya. Di rumah sakit dan ruang operasi dokter, model prediktif melengkapi penilaian dokter sendiri, bukan menggantikannya. Dalam pemasaran, peningkatan penargetan, berdasarkan pembelajaran mesin, digunakan untuk mengurangi panggilan telepon yang mengganggu dan email sampah, sekaligus memaksimalkan kontak pelanggan yang bermanfaat; yaitu hanya telepon/sms/email kepada orang-orang yang kemungkinan besar tertarik dengan apa yang Anda jual. Di masa lalu, strategi kontak menyeluruh merupakan hal yang umum. Staf di pusat panggilan akan diberikan buku telepon dan disuruh untuk memeriksanya, menelepon sebanyak mungkin orang tanpa memandang kecenderungan mereka untuk membeli produk atau layanan yang ditawarkan. Departemen SDM juga memanfaatkan pembelajaran mesin untuk menyaring lamaran pekerjaan terlebih dahulu untuk menyingkirkan pelamar yang paling tidak mungkin cocok untuk peran tertentu. Hal ini memungkinkan staf HR untuk lebih memusatkan waktu mereka pada prospek yang paling menjanjikan.

Peluang lain yang dimungkinkan oleh pembelajaran mesin adalah memperkirakan jenis perilaku baru yang sebelumnya tidak dipertimbangkan atau tidak hemat biaya. Dalam kepolisian misalnya, model prediktif membantu petugas polisi mempersempit daftar tersangka dan menargetkan kejahatan dengan lebih baik, meningkatkan efisiensi dan memberi mereka lebih banyak waktu untuk fokus pada kejahatan yang paling serius. Model prediktif juga penting dalam layanan “pencocokan orang” seperti situs kencan dan pekerjaan. Tanpa model prediktif, hal ini akan menjadi masalah yang sangat berbeda – mencocokkan orang berdasarkan serangkaian kriteria yang sangat kasar dan/atau harus menggunakan orang

untuk melakukan pencocokan, menjadikan layanan tersebut jauh lebih mahal dan kurang efisien.

Terlepas dari semua argumen yang mendukung sistem pengambilan keputusan otomatis berdasarkan model prediktif, jangan salah dengan berpikir bahwa pembelajaran mesin itu sempurna. Model prediktif dapat dan memang sering melakukan kesalahan, namun rata-rata tidak sesering manusia. Demikian pula, jika ada masalah atau bias dalam pemberian data ke algoritma yang digunakan untuk membangun model prediktif, maka model yang dihasilkan akan cacat. Selain itu, jika orang yang merancang sistem pengambilan keputusan membuat asumsi yang salah tentang bagaimana model prediktif akan digunakan, model tersebut tidak akan berfungsi dengan baik.

Bayangkan Anda ingin menggunakan pembelajaran mesin untuk mengidentifikasi manajer senior yang bertanggung jawab mengelola tim di organisasi besar. Tujuannya, setelah mereka teridentifikasi, adalah untuk meminta mereka berkontribusi terhadap respons lintas industri terhadap undang-undang baru pemerintah yang bertujuan untuk meningkatkan penyediaan layanan pengasuhan anak. Model ini mungkin sangat prediktif, namun pasti akan bias terhadap laki-laki. Mengapa? Karena kita hidup dalam masyarakat yang masih menunjukkan bias gender di sebagian besar industri, dan masyarakat di mana perempuan masih dipandang sebagai sumber utama pengasuhan anak. Di satu sisi, model ini akan akurat selama model tersebut mengidentifikasi para pemimpin senior, namun hal ini akan mengakibatkan kurangnya keterwakilan perempuan yang mencerminkan bias masyarakat yang ada.

Salah satu situasi kehidupan nyata yang banyak dipublikasikan adalah pemberian pembebasan bersyarat kepada narapidana. Telah dilaporkan bahwa algoritme yang digunakan untuk mendukung keputusan pembebasan bersyarat di AS memiliki bias rasial yang sangat kuat. Jika Anda berkulit putih, kemungkinan besar Anda akan mendapatkan pembebasan bersyarat jauh lebih besar dibandingkan narapidana dari etnis lain. Mengapa? Karena model prediktif yang digunakan untuk menilai kasus pembebasan bersyarat dikembangkan dengan menggunakan informasi tentang keputusan historis yang bias. Itu semua kembali ke data yang digunakan dalam proses pembelajaran mesin. Jika data berisi informasi yang berkaitan dengan keputusan yang sebelumnya bias, maka informasi tersebut pasti akan diambil oleh algoritma pembelajaran mesin apa pun. Dengan kata lain, pembelajaran mesin tidak memiliki kesadaran sosial dan hati nurani. Jika ada pola dalam data yang secara etis tidak diinginkan atau ilegal, proses pembelajaran mesin tidak dapat mengetahuinya. Oleh karena itu, mereka tidak akan memperlakukan pola-pola yang tidak dapat diterima secara sosial ini dengan cara yang berbeda dari pola-pola lain yang mereka temukan.

Contoh yang lebih nyata dari bias yang dapat ditampilkan oleh sistem AI berbasis pembelajaran mesin adalah chatbot Tay, yang dikembangkan oleh Microsoft. Tay dimaksudkan untuk menghasilkan respons otomatis terhadap tweet remaja dan beradaptasi dengan percakapan baru yang muncul. Tay akhirnya ditarik dalam beberapa jam karena mulai men-tweet konten rasis dan tidak pantas lainnya. Alasan mereka melakukan hal ini adalah karena tweet yang digunakan untuk melatihnya berisi materi rasis dan materi lain yang tidak

dapat diterima. Tay mempelajari pola-pola ini dan memainkannya kembali dalam tanggapannya sendiri. Kemampuan Tay untuk merespons dengan menggunakan bahasa alami dapat dianggap hampir manusiawi, namun ketidakmampuannya untuk memoderasi responsnya untuk memenuhi norma-norma sosial dan etika yang diharapkan oleh masyarakat luas sangatlah tidak manusiawi.

Kekhawatiran lainnya adalah jika seorang data scientist memiliki prasangka tertentu, maka ada kemungkinan bagi mereka untuk secara sadar merancang prasangka tersebut ke dalam sistem, sehingga menumbangkan apa yang disampaikan oleh data tersebut. Namun hal ini juga dapat memberikan manfaat bagi masyarakat. Jika diyakini bahwa pengguna tidak menyukai keputusan yang dibuat, maka sistem yang dirancang dengan baik akan memungkinkan logika keputusan untuk dimodifikasi/ditimpa – seperti yang telah kita bahas terkait dengan contoh penjualan wiski di bab sebelumnya.

Artinya, perlu adanya pengawasan terhadap sistem pengambilan keputusan berbasis pembelajaran mesin, dan serangkaian pemeriksaan dan keseimbangan yang tepat untuk memastikan bahwa sistem bekerja dengan cara yang legal dan etis.

BAB 5

CARA KERJA PEMBELAJARAN MESIN

Sebagian besar aplikasi pembelajaran mesin didukung oleh model prediktif. Skor yang dihasilkan oleh model kemudian mendorong pengambilan keputusan dan tindakan selanjutnya yang diambil. Jadi, bagaimana model prediktif menghasilkan skor? Hal ini bergantung pada jenis model prediktif yang digunakan, dan terdapat beberapa opsi yang dapat dipilih. Namun, tiga jenis model paling populer, yang digunakan di sebagian besar aplikasi bisnis dunia nyata, adalah Kartu Skor (model linier), Pohon Keputusan (Pohon Klasifikasi dan Regresi atau CART) dan Jaringan Syaraf Tiruan (JST), yang paling umum digunakan. disebut sebagai Neural Networks (NNs).

Kartu skor dan pohon keputusan relatif mudah dipahami oleh orang non-teknis. Skor model dihitung dengan cara yang sederhana dan transparan. Sangat mudah untuk menentukan mengapa model ini membuat prediksi seperti itu. Jaringan saraf di sisi lain cenderung sedikit lebih kompleks dan bersifat “kotak hitam”. Artinya, mungkin sulit untuk memahami mengapa jaringan neural mencapai skor yang dimilikinya, dan oleh karena itu mengapa keputusan tertentu dibuat berdasarkan skor tersebut. Namun, jaringan neural menjadi pilihan yang semakin populer untuk banyak aplikasi pembelajaran mesin karena kecenderungannya untuk memberikan prediksi yang lebih akurat dibandingkan kartu skor atau pohon keputusan dalam banyak (tetapi tidak semua) situasi.

Bentuk jaringan saraf tingkat lanjut (jaringan dalam/pembelajaran mendalam) saat ini menjadi yang terdepan dalam pembelajaran mesin/penelitian AI. Akibatnya, jaringan saraf menjadi model pilihan untuk masalah AI yang kompleks seperti identifikasi objek, pengenalan suara, dan terjemahan bahasa.

Untuk mengilustrasikan ketiga jenis model prediktif ini, mari kita kembali ke dunia layanan kesehatan dan fokus pada satu kondisi spesifik: penyakit jantung. Penyakit jantung adalah salah satu penyebab utama kematian di seluruh dunia. Setiap tahunnya, sekitar 1 dari setiap 1.000 penduduk di negara-negara barat (Inggris, AS, dll.) meninggal karena penyakit jantung, dan sebagian besar penduduk hidup dengan penyakit jantung yang mungkin akan membunuh mereka suatu saat nanti. Jika Anda dapat mengidentifikasi orang-orang yang kemungkinan besar terkena penyakit jantung di masa depan dan mengambil tindakan pencegahan untuk mengurangi kemungkinan tersebut, hal ini akan mengurangi penderitaan manusia secara signifikan dan meningkatkan harapan hidup. Pengurangan biaya layanan kesehatan juga akan sangat besar.

Bayangkan pemerintah memiliki cita-cita utopis untuk melakukan skrining penyakit jantung pada setiap orang, dengan tujuan mengambil tindakan pencegahan untuk mengurangi kejadian penyakit ini di masa depan. Salah satu cara yang dapat mereka lakukan adalah dengan memulai kampanye kesehatan dimana setiap laki-laki, perempuan dan anak-anak diundang untuk mengunjungi dokter mereka. Mereka kemudian akan menjalani

pemeriksaan penunjang dan tes diagnostik serta diberikan nasihat tentang cara meminimalkan risiko terkena penyakit jantung di masa depan.

Secara teori, hal tersebut baik-baik saja, namun dalam praktiknya, biaya untuk melakukan tindakan seperti itu akan sangat besar dan tidak akan ada cukup dokter yang tersedia untuk menemui semua orang dalam waktu yang realistis. Pengobatan kondisi lain yang mengancam jiwa, seperti kanker dan diabetes, kemungkinan besar akan terhambat karena kurangnya sumber daya. Faktanya, dampak akhirnya mungkin adalah penurunan kesejahteraan masyarakat secara keseluruhan karena terlalu menekankan penyakit jantung dan mengorbankan kondisi lainnya.

Pendekatan menyeluruh untuk melakukan skrining terhadap penyakit jantung pada setiap orang jelas tidak realistis. Namun, tujuan yang lebih bisa dicapai adalah dengan mengidentifikasi, setidaknya setengah dari mereka yang akan terkena penyakit jantung dalam 5 tahun ke depan, namun menargetkan tidak lebih dari, katakanlah, 5% dari populasi secara keseluruhan; yaitu hanya 1 dari 20 orang yang diundang untuk menjalani serangkaian tes lengkap, namun mayoritas dari mereka yang kemungkinan terkena penyakit jantung adalah 5%. Oke – jadi setengah dari kasus penyakit jantung akan terlewatkan, namun penghematan biaya membuat tindakan ini dapat dibenarkan dan tidak mengancam sumber daya yang dialokasikan untuk pengobatan kondisi lainnya.

Untuk mencapai tujuan ini, kita memerlukan cara untuk mengidentifikasi kelompok yang paling berisiko. Untuk melakukan hal ini, kami akan menerapkan pendekatan pembelajaran mesin untuk membangun model prediktif yang memperkirakan kemungkinan seseorang terkena penyakit jantung dalam beberapa tahun ke depan (model klasifikasi). Tujuannya adalah menggunakan model tersebut untuk mengidentifikasi individu-individu yang paling berisiko. Orang-orang ini kemudian akan dihubungi untuk mengundang mereka melakukan pemeriksaan dan tinjauan gaya hidup.

Titik awal pembelajaran mesin adalah data. Data adalah bahan bakar yang mendukung proses analisis. Mencoba melakukan pembelajaran mesin tanpa data seperti mencoba membuat kue tanpa bahan apa pun atau mencuci mobil tanpa air.

Untuk membangun model penyakit jantung, tugas pertama adalah mengumpulkan beberapa data historis pasien. Untuk masalah khusus ini, kita akan melihat kembali dan memperoleh rekam medis dari sampel 500.000 orang (dipilih secara acak) yang tidak memiliki tanda-tanda penyakit jantung lima tahun lalu. Data ini mencakup, antara lain, hal-hal seperti: usia, jenis kelamin, Indeks Massa Tubuh (IMT), tekanan darah, riwayat kesehatan sebelumnya, seberapa sering mereka berolahraga, apakah mereka merokok, dan berapa banyak alkohol yang mereka konsumsi. Mungkin juga terdapat informasi non-medis lain yang tersedia seperti pendapatan masyarakat, status perkawinan, jumlah tanggungan, jenis rumah yang mereka tinggali, dan sebagainya, yang dikumpulkan dari sumber lain.

Cuplikan data pribadi dari lima tahun lalu ini kemudian dicocokkan dengan apa yang terjadi pada orang-orang tersebut selama jangka waktu perkiraan 5 tahun berikutnya; yaitu, ada catatan yang berisi orang-orang yang mengalami penyakit jantung pada suatu waktu

dalam 5 tahun ke depan, dan orang-orang yang tidak menderita penyakit jantung. Jadi, saat ini kami memiliki dua bagian data:

- **Data observasi.** Ini adalah informasi tentang orang-orang pada awal periode lima tahun; yaitu usia, tekanan darah, kebiasaan merokok dan sebagainya.
- **Data hasil.** Ini adalah catatan kesehatan mereka dalam jangka waktu perkiraan. Secara khusus, data ini mencatat apakah seseorang mengidap penyakit jantung atau tidak.

Kedua jenis data ini jika digabungkan akan membentuk sampel pengembangan yang akan digunakan untuk menghasilkan model. Mari kita asumsikan bahwa dari 500.000 kasus dalam sampel perkembangan, 30.000 (6%) kemudian menderita penyakit jantung dan 470.000 (94%) lainnya tidak. Perhatikan bahwa memiliki data tentang orang-orang yang tidak terkena penyakit jantung sama pentingnya dengan memiliki data tentang mereka yang mengidap penyakit jantung. Hal ini karena proses pembelajaran mesin bekerja dengan mengidentifikasi perbedaan antara kedua jenis peristiwa tersebut. Kesalahan umum dalam analisis data adalah hanya mencatat data tentang peristiwa yang diinginkan, dan melupakan “yang bukan peristiwa”. Tanpa informasi peristiwa dan non-peristiwa, sangat sulit untuk mencapai kesimpulan yang berarti menggunakan pembelajaran mesin (atau metode analisis data lainnya).

Tahap selanjutnya dari proses ini adalah saat matematika kompleks berperan. Berbagai algoritme diterapkan untuk menghasilkan model prediktif, berdasarkan apa yang dapat disimpulkan oleh algoritme dari data dalam sampel pengembangan. Secara khusus, algoritme berupaya menemukan hubungan yang berkorelasi dengan peristiwa atau bukan peristiwa, dan hubungan inilah yang ditangkap oleh model prediktif yang dihasilkan dari proses pembelajaran mesin.

Saat ini, ada banyak sekali paket komputer yang menerapkan matematika yang relevan untuk membuat model. Biasanya, Anda tidak perlu menjadi ilmuwan data terlatih untuk dapat menghasilkan model prediktif – perangkat lunak akan melakukan semuanya untuk Anda.

Oleh karena itu, memiliki pengetahuan tentang teknik matematika yang relevan sangatlah membantu. Hal ini karena memungkinkan Anda memahami diagnostik yang dihasilkan perangkat lunak dan mengidentifikasi apakah ada masalah dengan proses pembuatan model. Selain itu, seringkali ada beberapa parameter yang perlu diatur dalam perangkat lunak. Seorang data scientist yang berpengalaman akan mengetahui cara mengubah parameter ini untuk menghasilkan model yang sangat prediktif, dan juga memenuhi segala batasan bisnis yang diberlakukan oleh organisasi yang menugaskannya.

Seringkali proses pembuatan model bersifat berulang. Banyak model pengujian dibangun menggunakan algoritma yang berbeda sebelum model akhir diperoleh. Oleh karena itu, proses analitik memerlukan seseorang yang mengetahui apa yang mereka lakukan untuk memandu proses tersebut sehingga model keseluruhan terbaik dapat diperoleh.

Mari kita asumsikan bahwa data scientist memulai dengan mempertimbangkan model tipe kartu skor. Sebagai referensi, algoritma paling populer untuk menghasilkan model tipe kartu skor disebut Regresi Linier dan Regresi Logistik.

Setelah menggunakan perangkat lunak yang sesuai, untuk menerapkan algoritma yang relevan pada sampel pengembangan, perangkat lunak tersebut menghasilkan model prediktif yang ditunjukkan pada Gambar 5.1:

Skor awal (Konstan)	350		
Usia (Tahun)		Pendapatan kotor tahunan (Rp. 000)	
< 23	-57	< 22.000	11
23 - 32	-26	22.001 - 38.000	6
33 - 41	0	38.001 - 60.000	0
42 - 48	7	60.001 - 94.000	-3
49 - 57	15	94.001 - 144.000	-5
58 - 64	24	> 144.000	-6
65 - 71	31		
> 71	65	Perokok?	
		Ya	37
Berat badan (dalam Kg)		Tidak	0
< 19	2	Diabetes?	
19 - 25	0	Ya	21
27 - 29	8	Tidak	0
30 - 32	14		
> 32	29	Level kolesterol (mg/dl darah)	
Jenis Kelamin		Rendah (< 160 mg)	-2
Laki-laki	2	Normal (160 - 200 mg)	0
Perempuan	-4	High (201 -240 mg)	19
		Sangat Tinggi (> 240 mg)	32
Konsumsi Alkohol (Botol/minggu)		Tekanan darah	
0	4	Rendah (dibawah 90/60)	3
1 - 12	0	Normal (antara 90/60 - 140/90)	0
13 - 24	5	Tinggi (diatas 140/90)	36
25 - 48	10		

Gambar 5.1 Kartu skor untuk prediksi penyakit jantung

Dengan menggunakan kartu skor pada Gambar 5.1, skor seseorang dihitung sebagai berikut:

- Untuk memulai, setiap orang diberi “skor awal” 350.
 - Semua poin relevan yang berlaku kemudian ditambahkan atau dikurangi dari skor awal.
- Untuk wanita usia 45 tahun dengan ciri-ciri sebagai berikut:
- BMI 28.
 - Minum rata-rata 6 unit alkohol seminggu.
 - Lulusan universitas.
 - Penghasilan Rp. 50.000.000 setahun.
 - Merokok.
 - Bukan penderita diabetes.

- Kadar kolesterol normal.
- Tekanan darah rendah.

Skor awalnya adalah 350. Tujuh poin ditambahkan karena usianya, sehingga skornya menjadi 357. Delapan poin kemudian ditambahkan berdasarkan BMI-nya, empat poin dikurangi untuk jenis kelaminnya dan tidak ada poin yang ditambahkan (atau dikurangi) berdasarkan alkoholnya konsumsi dan sebagainya. Setelah menambah/mengurangi semua poin yang relevan, skor akhirnya adalah 404.

Skor individu dalam kartu skor merupakan representasi pola/hubungan yang ditemukan oleh algoritma pembelajaran mesin. Misalnya, peningkatan skor yang dialokasikan untuk orang lanjut usia mewakili hubungan antara peningkatan usia dan peningkatan risiko penyakit jantung, yang disimpulkan dari data (sampel perkembangan) yang digunakan untuk membangun model. Demikian pula, perokok mendapat poin lebih banyak dibandingkan bukan perokok, yang mewakili hubungan antara merokok dan penyakit jantung.

Melihat kartu skor pada Gambar 5.1, mudah untuk melihat mengapa kartu skor begitu menarik dan mengapa kartu skor digunakan secara luas di banyak sektor industri. Salah satu alasannya adalah Anda tidak memerlukan kemampuan teknis apa pun untuk melihat item data mana yang berkontribusi terhadap skor yang diterima seseorang. Juga mudah untuk mengukur kepentingan relatif setiap item data. Usia menyumbang lebih banyak poin pada skor akhir dibandingkan apa pun. Di sisi lain, gender hanya memberikan kontribusi kecil terhadap skor keseluruhan. Hal ini juga terjadi karena sebagian besar item data dalam model, seperti merokok, konsumsi alkohol, dan tekanan darah sesuai dengan apa yang diketahui secara umum dari penelitian medis; yaitu hubungan yang ditangkap oleh model terlihat masuk akal dan sesuai dengan apa yang telah diketahui mengenai faktor risiko penyakit jantung.

Fakta bahwa model tersebut sesuai dengan apa yang telah diketahui oleh para ahli berguna untuk membuat orang memercayai model tersebut dan merasa nyaman menggunakannya sebagai alat diagnostik dalam pekerjaan mereka sehari-hari. Jika model tersebut memberikan poin negatif untuk merokok dan diabetes, dan poin positif untuk tekanan darah normal dan usia muda, maka hal tersebut bertentangan dengan apa yang diketahui oleh setiap dokter. Akibatnya, orang tidak akan terlalu yakin bahwa model tersebut memprediksi kondisi dengan benar meskipun prediksi model tersebut benar.

Daya tarik terhadap akal sehat ini penting, terutama ketika model diterapkan di suatu wilayah untuk pertama kalinya. Orang sering kali curiga terhadap pendekatan baru seperti pembelajaran mesin dan mungkin secara naluriah tidak memercayai sesuatu yang menghilangkan elemen manusia dari proses pengambilan keputusan. Semakin banyak hal yang dapat dilakukan untuk meyakinkan masyarakat bahwa keputusan berdasarkan skor model adalah keputusan yang tepat, maka akan semakin mudah untuk mengatasi hambatan apa pun terhadap pembelajaran mesin yang dihadapi.

Satu hal dalam kartu skor yang mungkin sedikit tidak terduga adalah dampak pendapatan seseorang. Model tersebut mengatakan bahwa jika Anda berpenghasilan rendah maka Anda lebih mungkin terkena penyakit jantung. Jika Anda memikirkan hal ini, maka itu masuk akal. Jika seseorang berpenghasilan rendah, mereka mungkin tidak mampu membayar

keanggotaan gym (kurang berolahraga), dan mereka tidak akan menghabiskan banyak pendapatannya untuk makanan berkualitas baik. Namun, penting untuk tidak bingung antara korelasi dan sebab akibat. Model tersebut menangkap korelasi antara pendapatan dan penyakit jantung, namun tidak tepat untuk mengatakan bahwa memiliki pendapatan rendah selalu menyebabkan penyakit jantung. Mungkin saja, tetapi bisa juga hanya bertindak sebagai proksi untuk fitur-fitur lainnya.

Kemampuan untuk memilih item data baru yang bersifat prediktif, namun mungkin tidak dianggap penting oleh seorang pakar sebelumnya, adalah salah satu kekuatan besar pembelajaran mesin. Hal ini terutama berlaku ketika kita melangkah lebih jauh ke dunia “Big Data” di mana terdapat puluhan ribu item data yang tersedia, yang mana pun dapat ditampilkan dalam model prediktif – jauh lebih banyak daripada yang dapat dianalisis secara manual.

Pembelajaran mesin adalah cara otomatis untuk memilah semua item data tersebut untuk menemukan beberapa item yang berkorelasi dengan perilaku yang diprediksi. Oleh karena itu, penting untuk dicatat bahwa dalam sebagian besar kasus, di banyak area aplikasi, pembelajaran mesin cenderung memilih item data yang serupa dengan apa yang akan digunakan oleh pakar manusia saat mengambil keputusan. Hubungan prediktif yang baru, aneh, atau berlawanan dengan intuisi cenderung menjadi pengecualian daripada aturan.

Hal lain yang memberikan keunggulan pada model prediktif dibandingkan pengambil keputusan manusia adalah bahwa skor yang dialokasikan untuk setiap item data adalah optimal, berdasarkan data dalam sampel pengembangan; yaitu skor yang diberikan pada setiap item data dalam model adalah skor terbaik yang dapat dihasilkan, sehingga menghasilkan prediksi seakurat mungkin.

Pembelajaran mesin bukan satu-satunya cara untuk membuat model tipe kartu skor. Salah satu alternatifnya adalah dengan menerapkan pendapat ahli. Sekelompok ahli berkumpul dan secara kolektif memutuskan faktor mana yang penting dan poin apa yang harus diberikan pada faktor tersebut. Model prediktif yang dibuat dengan cara ini sering kali bekerja dengan sangat baik, tetapi cenderung lebih rendah dibandingkan model yang dibuat menggunakan algoritme untuk menentukan skor menggunakan sampel pengembangan yang sesuai.

Fitur menarik lainnya dari kartu skor pada Gambar 5.1 adalah kartu skor tersebut hanya berisi 9 item data – namun jumlah tersebut cukup untuk membuat prediksi yang cukup baik dan hal ini sering terjadi. Sangat sedikit model prediktif yang memerlukan lebih dari beberapa lusin item data untuk dapat menghasilkan prediksi yang sangat baik. Sekalipun terdapat puluhan ribu bit data yang tersedia, sebagian besar item data tersebut hanya memberikan sedikit atau tidak sama sekali terhadap keakuratan prediksi. Ini adalah fitur pembelajaran mesin yang sangat berguna karena ini berarti bahwa meskipun diperlukan sejumlah besar analisis data untuk membuat model, sumber daya yang dibutuhkan jauh lebih sedikit untuk menggunakan model tersebut. Untuk menerapkan model prediktif, Anda hanya memerlukan item data yang ditampilkan dalam model. Tidak diperlukan data lain.

Model scorecard pada Gambar 5.1 memprediksi kemungkinan seseorang terkena penyakit jantung dalam 5 tahun mendatang. Namun bagaimana skor individu yang dihasilkan oleh model tersebut diterjemahkan ke dalam probabilitas? Salah satu cara untuk membangun hubungan antara skor dan probabilitas adalah dengan menghasilkan tabel “distribusi skor” seperti yang ditunjukkan pada Gambar 5.2.

Untuk menghasilkan Gambar 5.2, sampel data yang benar-benar baru, sampel validasi, dikumpulkan untuk 500.000 orang lainnya (30.000 di antaranya kemudian menderita penyakit jantung selama 5 tahun berikutnya). Sampel validasi ini sepenuhnya independen dari sampel yang digunakan untuk membangun model. Oleh karena itu, hal ini memberikan gambaran yang representatif tentang bagaimana kinerja model ketika diterapkan pada orang baru, yang detailnya tidak digunakan untuk membuat model.

Kelompok	Kisaran skor jumlah		Jumlah populasi	% Populasi	Nomor dengan penyakit jantung setelahnya	% dengan penyakit jantung setelah 5 tahun
	dari	ke				
1	0	300	55,950	11.19%	40	0.07%
2	301	320	56,606	11.32%	68	0.12%
3	321	340	59,700	11.94%	129	0.22%
4	341	360	58,706	11.74%	216	0.37%
5	361	380	64,429	12.89%	403	0.63%
6	381	400	52,749	10.55%	575	1.09%
7	401	420	34,089	6.82%	600	1.76%
8	421	440	21,107	4.22%	632	2.99%
9	441	460	17,269	3.45%	878	5.09%
10	461	480	23,364	4.67%	2,020	8.65%
11	481	500	17,477	3.50%	2,553	14.61%
12	501	520	13,554	2.71%	3,366	24.84%
13	521	540	7,103	1.42%	3,463	48.76%
14	541	560	8,260	1.65%	6,587	79.74%
15	561	999	9,637	1.93%	8,469	87.88%
Total	Total		500,000		30,000	6.00%

Gambar 5.2 Tabel distribusi skor

Dua kolom paling kiri berisi rentang skor, menunjukkan rentang skor yang dilaporkan di setiap baris. Baris pertama berisi rincian setiap orang yang mendapat skor 300 atau kurang, baris kedua berisi rincian setiap orang yang mendapat skor antara 301 dan 320, dan seterusnya. Idealnya, akan ada baris terpisah untuk setiap skor individu, namun karena ada ratusan skor berbeda yang dapat dihasilkan oleh model, skor dalam contoh ini telah dikelompokkan ke dalam rentang 20 poin untuk memudahkan.

Kolom lain pada Gambar 5.2 memberikan informasi tentang individu di setiap rentang skor. Misalnya, 56.606 kasus mendapat skor antara 301 dan 320, mewakili 11,32% kasus dalam sampel validasi. Dari jumlah tersebut, 68 orang menderita penyakit jantung.

Kolom paling kanan menunjukkan persentase orang di setiap rentang skor yang mengidap penyakit jantung. Hal ini sebenarnya merupakan prediksi yang dibuat oleh model. Bagi mereka yang berada dalam kisaran skor 301 – 320, prediksi modelnya adalah 0,12%

Dengan kata lain, siapa pun yang skornya antara 301 dan 320 memiliki peluang 1 banding 833 terkena penyakit jantung ($1/0,0012$).

Untuk menilai risiko seseorang terkena penyakit jantung di kemudian hari, yang perlu dilakukan adalah:

- Hitung skor mereka menggunakan kartu skor pada Gambar 5.1.
- Temukan baris di mana skornya berada pada Gambar 5.2.
- Lihatlah ke kolom paling kanan pada Gambar 5.2 untuk mendapatkan prediksi.

Jika kita kembali ke kasus perempuan berusia 45 tahun yang mendapat skor 404, ia berada pada kisaran skor 401-420. Secara keseluruhan, proporsi orang dengan skor antara 401 dan 420 yang kemudian menderita penyakit jantung adalah 1,76%. yaitu model tersebut memperkirakan bahwa ia mempunyai peluang 1,76% (1 dari 57) terkena penyakit jantung dalam lima tahun ke depan.

Jika Anda memiliki informasinya, mengapa tidak mencoba menghitung skor Anda sendiri dan membuat prediksi risiko Anda sendiri terkena penyakit jantung?

Mengetahui arti skor itu penting, namun pertanyaan berikutnya adalah: Seberapa baik model kartu skor memprediksi penyakit jantung? Seberapa akuratnya? Salah satu cara untuk mengevaluasi model ini adalah dengan mempertimbangkan perbedaan tingkat penyakit jantung antara kasus dengan skor terendah dan tertinggi. Hanya 0,07% dari mereka yang mendapat skor 300 atau kurang akan menderita penyakit jantung dalam lima tahun ke depan. Bagi mereka yang mendapat skor 561 atau lebih, 87,88% mengembangkan kondisi tersebut. Dengan kata lain, orang dengan skor tertinggi memiliki kemungkinan 1.000 kali lebih besar terkena penyakit jantung dibandingkan orang dengan skor terendah – dan hal ini cukup bagus.

Cara lain untuk melakukan pendekatan ini adalah dengan membandingkan kinerja model dengan strategi pemilihan acak; yaitu jika Anda memilih orang secara acak untuk pemeriksaan, maka hanya 6% dari orang-orang tersebut yang akan terkena penyakit jantung, sedangkan kelompok dengan skor terbaik (561 - 999) memiliki kinerja sekitar 15 kali lebih baik dari ini ($87,88 / 6$).

BAB 6

MODEL PREDIKTIF UNTUK MENGAMBIL KEPUTUSAN

Pada titik ini kita:

- Memiliki model prediktif. Model tersebut menghasilkan skor yang menunjukkan seberapa besar kemungkinan seseorang terkena penyakit jantung.
- Pahami arti skornya.
- Ketahui kemungkinan seseorang terkena penyakit jantung berdasarkan.
- Rasakan seberapa baik kartu skor dalam mengidentifikasi penderita penyakit jantung dibandingkan kebijakan pemilihan acak.
- Mengetahui distribusi populasi pada rentang skor.

Semuanya baik-baik saja, tetapi tidak satu pun dari informasi ini yang memberi tahu kita apa yang harus dilakukan berdasarkan skor yang diterima seseorang. Jika model tersebut ingin memberikan nilai, maka keputusan perlu dibuat dan keputusan tersebut harus ditindaklanjuti. Kalau ada yang mendapat nilai katakanlah 517, apakah diajak periksa atau tidak?

Sebelum melangkah lebih jauh, mari kita mengingat kembali awal bab sebelumnya, yang memaparkan tujuan berikut:

- Identifikasi setidaknya separuh dari mereka yang akan terkena penyakit jantung dalam 5 tahun ke depan, lalu ajak mereka mengunjungi dokter untuk pemeriksaan lebih mendalam.

Tunduk pada batasan berikut:

- Tidak lebih dari 5% populasi dapat diundang untuk pemeriksaan; yaitu terdapat sumber daya yang cukup bagi 1 dari setiap 20 orang untuk memeriksakan diri ke dokter.

Kelompok	Kisaran skor jumlah		Jumlah populasi	% Populasi	Nomor dengan penyakit jantung setelahnya	% dengan penyakit jantung setelah 5 tahun
	dari	ke				
1	0	300	55,950	11.19%	40	0.07%
2	301	320	56,606	11.32%	68	0.12%
3	321	340	59,700	11.94%	129	0.22%
4	341	360	58,706	11.74%	216	0.37%
5	361	380	64,429	12.89%	403	0.63%
6	381	400	52,749	10.55%	575	1.09%
7	401	420	34,089	6.82%	600	1.76%
8	421	440	21,107	4.22%	632	2.99%
9	441	460	17,269	3.45%	878	5.09%
10	461	480	23,364	4.67%	2,020	8.65%
11	481	500	17,477	3.50%	2,553	14.61%
12	501	520	13,554	2.71%	3,366	24.84%
13	521	540	7,103	1.42%	3,463	48.76%
14	541	560	8,260	1.65%	6,587	79.74%
15	561	999	9,637	1.93%	8,469	87.88%
Total	Total		500,000		30,000	6.00%

Gambar 6.1 Skor distribusi penyakit jantung (2)

Untuk menjawab pertanyaan tersebut, kita perlu mengacu kembali pada distribusi skor yang diperkenalkan pada Gambar 6.1.

Faktor pembatas (constraint) dalam permasalahan ini adalah waktu dokter. Paling banyak 1 dari 20 orang (5%) bisa diundang untuk check up. Dengan menggunakan informasi dari kolom “% populasi” pada Gambar 6.1 untuk melakukan perhitungan sederhana, dapat ditentukan bahwa:

- 1,93% populasi mendapat skor 561 atau lebih.
- 3,58% populasi mendapat skor 541 atau lebih.
- 5,00% populasi mendapat skor 521 atau lebih.
- 7,71% populasi mendapat skor 501 atau lebih.
- Dan seterusnya...

Oleh karena itu, strategi cut-off (aturan keputusan) yang diterapkan adalah:

- Undang siapa pun yang mendapat skor 521 atau lebih untuk pemeriksaan dan tinjauan gaya hidup dengan dokter mereka.

Langkah terakhir adalah mengevaluasi dampak aturan keputusan. Seberapa baik proses pengambilan keputusan berdasarkan scorecard? Akankah model tersebut mampu mengidentifikasi setidaknya 50% penderita penyakit jantung dalam kelompok yang mendapat skor 521 atau lebih?

Melihat kolom paling kanan kedua pada Gambar 6.1, terdapat 18.519 ($3.463 + 6.587 + 8.469$) kasus dengan skor 521 atau lebih. Itu berarti 62% dari total 30.000 kasus penyakit jantung. Cukup bagus bukan? Dengan memilih hanya 5% dari populasi yang mendapat skor 521 atau lebih, 62% dari seluruh kasus penyakit jantung dapat diidentifikasi. Tujuannya, (untuk mengidentifikasi setidaknya 50%) telah terpenuhi.

Angka-angka ini didasarkan pada 500.000 orang dalam sampel validasi. Mengingat sampel independen yang diambil berjumlah besar, maka terdapat asumsi yang cukup aman bahwa hasilnya mewakili apa yang akan diamati jika model tersebut diterapkan pada seluruh populasi negara yang berjumlah puluhan juta jiwa.

Seperti yang telah saya sebutkan, model prediktif tidaklah sempurna. Beberapa dari mereka yang mendapat skor 521 atau lebih tidak akan terkena penyakit jantung. Dengan menggunakan kembali data dalam tabel distribusi skor, relatif mudah untuk mengetahui seberapa sering model tersebut melakukan dengan benar untuk aturan keputusan tertentu dan seberapa sering model tersebut melakukan kesalahan, sebagai berikut:

- Terdapat 25.000 kasus dengan skor 521 atau lebih.
- Dari jumlah tersebut, 18.519 menderita penyakit jantung.
- Oleh karena itu, tingkat keberhasilan keseluruhan adalah 74% ($100 * 18.519/25.000$).
- Artinya, meskipun 74% dari mereka yang diundang diperkirakan mengidap penyakit jantung, 26% sisanya yang diundang untuk pemeriksaan sebenarnya tidak memerlukannya.

Modelnya tidak sempurna. 26% kali melakukan kesalahan, namun kinerjanya jauh lebih baik daripada memilih orang secara acak untuk pemeriksaan. Dengan hanya 6% populasi yang

diperkirakan mengidap penyakit jantung dalam lima tahun ke depan, strategi undangan acak akan mengakibatkan 94% (100% – 6%) pemeriksaan kesehatan terbuang sia-sia bagi orang-orang yang tidak akan mengidap penyakit jantung. penyakit terlebih dahulu.

Tabel distribusi skor, seperti pada Gambar 6.1, mendasari semua pelaporan dan metrik kinerja yang mendukung evaluasi dan penggunaan model prediktif. Dalam aplikasi dunia nyata, kolom tersebut cenderung lebih terperinci (memiliki lebih banyak baris) dan berisi kolom tambahan yang berisi informasi seperti angka dan persentase kumulatif naik/turun, untuk membantu penghitungan.

Beberapa orang, terutama orang-orang tua seperti saya yang mulai melakukan pembelajaran mesin sebelum alat seperti Excel umum digunakan, lebih memilih bekerja langsung dari tabel distribusi skor seperti pada Gambar 6.1. Namun, saat ini, ada banyak alat visualisasi hebat yang dapat digunakan untuk menyajikan hasil tabel distribusi skor dengan cara yang lebih intuitif dan ramah pengguna. Alat terbaik juga bersifat interaktif. Mereka memungkinkan Anda bereksperimen dengan berbagai model, batasan, dan strategi cut-off untuk menemukan aturan keputusan terbaik untuk masalah khusus Anda.

Dua cara paling populer untuk memberikan representasi grafis dari tabel distribusi skor adalah diagram Peningkatan dan diagram Keuntungan. Hal ini dijelaskan dalam Lampiran A.

BAB 7

KARTU SKOR DAN POHON KEPUTUSAN

Pohon keputusan adalah jenis model prediktif populer lainnya. Seperti halnya kartu skor, mereka juga sangat mudah dipahami dan digunakan. Untuk mendemonstrasikan seperti apa pohon keputusan dan cara kerjanya, mari kita lanjutkan dengan masalah penyakit jantung. Jika Anda ingat, tujuannya adalah untuk menemukan cara untuk mengidentifikasi setidaknya setengah dari mereka yang akan terkena penyakit jantung, namun untuk mengundang tidak lebih dari 1 dari 20 (5%) populasi untuk memeriksakan diri ke dokter.

Untuk membangun model pohon keputusan, 500.000 kasus yang digunakan untuk membuat kartu skor digunakan lagi, namun kali ini jenis algoritma yang berbeda diterapkan pada sampel pengembangan. Semua algoritma pohon keputusan yang populer adalah variasi pada tema yang berulang kali membagi sampel pengembangan menjadi kelompok-kelompok yang lebih kecil dan lebih kecil. Berikut ini adalah penjelasan sederhana dari algoritma pohon keputusan:

Langkah 1. Pilih satu item data dalam sampel pengembangan (misalnya pendapatan).

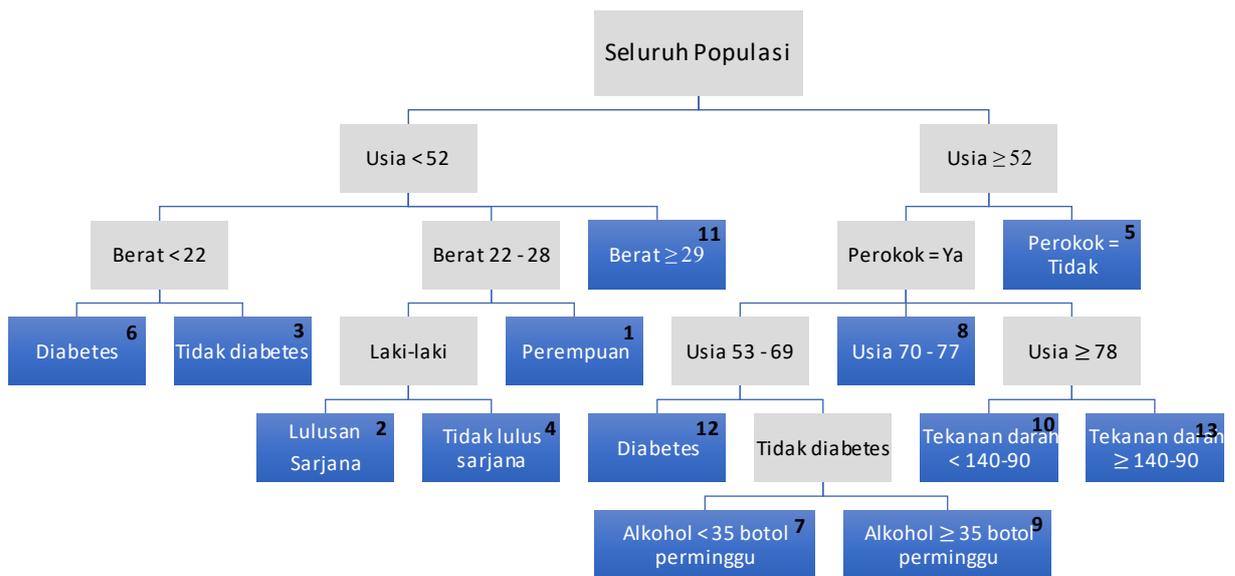
Langkah 2. Temukan nilai pendapatan yang menghasilkan pembagian data terbaik menjadi dua bagian. Yang saya maksud dengan “terbaik” adalah kelompok berpendapatan rendah dan kelompok berpendapatan tinggi sehingga kejadian penyakit jantung dimaksimalkan pada satu kelompok dan diminimalkan pada kelompok lain.

Langkah 3. Ulangi langkah 1 dan 2 untuk semua item data lainnya dalam sampel pengembangan (yaitu mengganti pendapatan dengan usia, BMI, perokok, dll.)

Langkah 4. Dari apa yang ditemukan pada langkah 2 dan 3, bagilah sampel pengembangan menjadi dua bagian menggunakan partisi terbaik yang ditemukan; yaitu membagi sampel perkembangan sedemikian rupa sehingga kejadian penyakit jantung dapat dimaksimalkan pada satu bagian dan diminimalkan pada bagian lainnya.

Langkah 5. Ulangi langkah 1-4 untuk masing-masing dua bagian yang dihasilkan sampel pengembangan. Lakukan terus hal ini sampai tidak ditemukan lagi perbedaan yang signifikan dari pemisahan selanjutnya.

Gambar 7.1 menunjukkan struktur model pohon keputusan yang diperoleh dari sampel perkembangan penyakit jantung, setelah semua kriteria pemisahan diterapkan.



Gambar 7.1 Model Pohon keputusan

Pada Gambar 7.1, algoritma dimulai dengan mengidentifikasi bahwa variabel yang paling signifikan adalah usia. Sampel perkembangan disegmentasi berdasarkan apakah seseorang berusia lebih dari atau kurang dari 52 tahun. Bagi mereka yang berusia kurang dari 52 tahun, algoritme kemudian menentukan bahwa pembagian paling signifikan berikutnya didasarkan pada BMI. Namun, bagi mereka yang berusia 52 tahun ke atas, variabel terpenting berikutnya adalah apakah seseorang tersebut seorang perokok, dan seterusnya.

Untuk menghitung skor seseorang dengan menggunakan pohon keputusan, seseorang memulainya dari puncak pohon pada simpul pertama (Keseluruhan populasi). Seseorang kemudian bergerak ke bawah pohon. Pada setiap titik percabangan dipilih simpul yang sesuai dengan karakteristik individu orang tersebut. Jika seseorang berusia kurang dari 52 tahun, sisi kiri pohon diikuti. Jika sudah berumur 52 tahun atau lebih maka diikuti cabang sebelah kanan. Proses ini diulangi pada setiap titik keputusan hingga tidak ada lagi keputusan yang harus diambil. Node terakhir di pohon (disebut simpul akhir atau simpul daun) tempat seseorang jatuh menentukan skornya.

Pada Gambar 7.1, terdapat 13 node akhir (diarsir dengan warna abu-abu dan diberi nomor) sehingga terdapat 13 skor yang dapat diterima seseorang. Mereka yang mendapat skor 1 memiliki peluang terendah terkena penyakit jantung, dan mereka yang mendapat skor 13 memiliki peluang tertinggi.

Dalam hal bagaimana skor yang dihasilkan oleh pohon keputusan digunakan, prosesnya hampir sama dengan cara seseorang menggunakan skor dari kartu skor. Rincian 500.000 orang dari sampel validasi dijalankan melalui pohon keputusan untuk memperoleh skor, dan kemudian dihasilkan tabel distribusi skor yang menunjukkan bagaimana skor didistribusikan.

Melalui proses yang sama seperti yang kami lakukan pada kartu skor, kelompok dengan skor tertinggi yaitu 5% dari populasi dicakup oleh mereka yang mendapat skor 12 atau 13. Dari jumlah tersebut, 16.910 kemudian menderita penyakit jantung, yang merupakan 55%

dari total jumlah penduduk yang menderita penyakit jantung. kasus. Oleh karena itu, strategi penargetan dalam mengundang masyarakat untuk melakukan pemeriksaan adalah dengan mengundang semua yang mendapat skor 12 atau 13.

Seperti kartu skor, strategi ini juga memenuhi tujuan yang dinyatakan, yaitu mengundang tidak lebih dari 5% dari total populasi untuk melakukan pemeriksaan, namun setidaknya setengah dari mereka yang akan menderita penyakit jantung berada dalam rentang 5% tersebut.

Node (Skor)	Jumlah Orang	% dari Populasi	Nomor dengan penyakit jantung	% dengan penyakit jantung setelah 5 tahun
1	98,760	19.75%	109	0.11%
2	104,324	20.86%	177	0.17%
3	32,176	6.44%	93	0.29%
4	78,287	15.66%	385	0.49%
5	19,998	4.00%	162	0.81%
6	28,675	5.74%	460	1.60%
7	48,748	9.75%	1,581	4.24%
8	22,801	4.56%	1,491	6.54%
9	18,884	3.78%	2,452	12.98%
10	15,034	3.01%	3,419	22.74%
11	7,313	1.46%	2,763	37.78%
12	18,901	3.78%	11,889	62.90%
13	6,099	1.22%	5,021	82.32%
Total	500,000		30,000	6.00%

Gambar 7.2 Distribusi skor untuk model pohon keputusan

Kami sekarang memiliki dua model. Pertanyaan yang kini ada di bibir banyak orang adalah: model mana yang terbaik? Apakah itu kartu skor atau pohon keputusan? Saat orang menanyakan pertanyaan ini, biasanya yang ingin mereka ketahui adalah: model mana yang menghasilkan prediksi paling akurat?

Dalam contoh ini, kedua model memberikan akurasi prediksi yang cukup untuk memenuhi tujuan yang ditetapkan. Namun, kartu skornya lebih baik. Alasan mengapa ini lebih baik adalah karena jika Anda menggunakan kartu skor untuk memutuskan 5% populasi mana yang akan diundang untuk melakukan pemeriksaan (mereka yang mendapat skor 521 atau lebih), Anda akan mengundang 62% dari mereka yang melanjutkan ke penyakit jantung. Sebaliknya, pohon keputusan hanya mengidentifikasi 55% dari mereka (yang mendapat skor 12 atau 13).

Pada titik ini ada baiknya melakukan beberapa observasi tentang persamaan dan perbedaan antara model pohon keputusan dan model kartu skor. Beberapa hal penting yang perlu diperhatikan adalah:

- Pohon keputusan dan kartu skor menggunakan item data yang serupa, namun tidak identik untuk membuat prediksi. Kartu skor mencakup pendapatan tahunan namun pohon keputusan tidak. Fitur “Lulusan” dalam model pohon keputusan tetapi tidak pada kartu skor.

- Kondisi cabang pada pohon keputusan tidak selalu selaras dengan rentang skor pada kartu skor. Dalam kartu skor, pembagian usia terjadi pada usia 49 dan 57 tahun, namun dalam pohon keputusan, pembagian usia utama adalah pada usia 52 tahun (dengan pembagian usia tambahan di bagian bawah pohon).
- Pohon keputusan menghasilkan 13 kemungkinan skor, sedangkan ada ratusan skor yang dapat dihasilkan oleh kartu skor; yaitu, kartu skor menghasilkan rentang skor yang jauh lebih terperinci dibandingkan pohon keputusan.

Fakta bahwa model kartu skor dan pohon keputusan memiliki perbedaan dalam data yang digunakan merupakan fitur algoritma yang digunakan untuk membangunnya. Ada ratusan algoritme pembelajaran mesin berbeda yang dapat digunakan untuk membuat berbagai jenis model prediktif. Setiap algoritme memiliki logikanya sendiri untuk memilih item data mana yang merupakan prediktor penting dan bobot yang harus diberikan pada prediktor tersebut dalam model. Masing-masing pendekatan mempunyai kekuatan dan kelemahannya masing-masing dan lebih baik dalam memprediksi jenis kasus tertentu dibandingkan jenis kasus lainnya.

Artinya dalam praktiknya adalah tidak ada model terbaik yang harus selalu digunakan untuk semua jenis masalah. Tidak ada satu jenis model prediksi yang dapat dikatakan lebih baik secara universal dibandingkan model lainnya. Kartu skor kali ini terbukti lebih baik (dalam hal akurasi prediksi) dibandingkan pohon keputusan, namun untuk jenis masalah yang berbeda, dengan menggunakan data yang berbeda dan dengan tujuan yang berbeda, pohon keputusan mungkin yang lebih unggul.

Artinya, jika Anda ingin memastikan bahwa Anda mendapatkan model terbaik untuk masalah khusus Anda (yang diukur dalam akurasi prediksi), Anda harus membuat model yang berbeda, membandingkannya satu sama lain, dan kemudian memutuskan model mana yang paling tepat. paling tepat.

Karena itu, dalam situasi praktis di dunia nyata, sebagian besar jenis model menghasilkan tingkat kinerja prediktif yang sangat mirip untuk banyak jenis masalah. Mungkin ada argumen teoretis mengenai mengapa satu jenis model lebih baik daripada yang lain, atau mengapa Algoritma A lebih unggul daripada Algoritma B, namun dalam praktiknya, argumen-argumen tersebut sangat banyak digunakan di banyak bidang penerapan. Oleh karena itu, ada banyak keleluasaan yang dapat diterapkan ketika memilih jenis model yang akan dipilih.

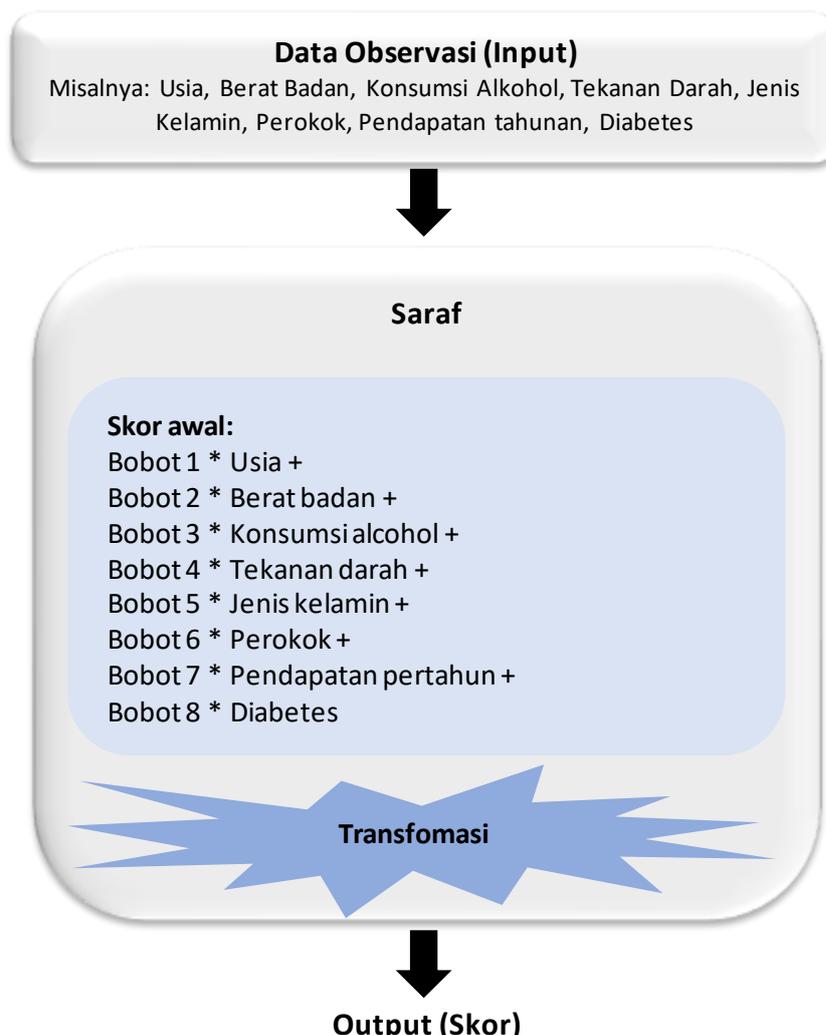
Ini adalah fitur hebat lainnya dari pembelajaran mesin. Hal ini berarti bahwa faktor-faktor lain, selain keakuratan prediksi, dapat ikut berperan dalam menentukan model mana yang terbaik untuk suatu permasalahan tertentu. Khususnya, dalam beberapa bidang permasalahan, memiliki model yang mudah dijelaskan dan selaras dengan pandangan yang masuk akal mengenai bagaimana keputusan harus diambil sangatlah penting. Hal ini terkadang jauh lebih penting daripada memiliki model yang secara teoritis lebih tepat atau sedikit lebih akurat, namun kurang dipahami oleh pengguna bisnis atau regulator industri.

BAB 8

JARINGAN SYARAF TIRUAN DAN PEMBELAJARAN MENDALAM

Otak kita mengandung milyaran neuron, masing-masing saling berhubungan untuk menghasilkan triliunan koneksi. Kombinasi neuron dan koneksi di antara keduanya adalah yang mendorong kemampuan kita untuk berpikir dan bertindak secara cerdas, memahami cara kerja dunia, dan membuat keputusan (baik secara sadar maupun tidak sadar) tentang cara kita berinteraksi dengan dunia.

Pada tahun 1950-an, ilmuwan komputer pertama kali mengemukakan konsep “Neuron Buatan”, namun baru pada pertengahan tahun 1980-an diusulkan penggunaan beberapa neuron buatan yang dihubungkan bersama (jaringan saraf). sebagai cara memecahkan masalah yang kompleks. Sejak saat itu, popularitas jaringan saraf terus berkembang dan bentuk-bentuk jaringan saraf yang lebih maju saat ini menjadi yang terdepan dalam penelitian kecerdasan buatan dan pembelajaran mesin.

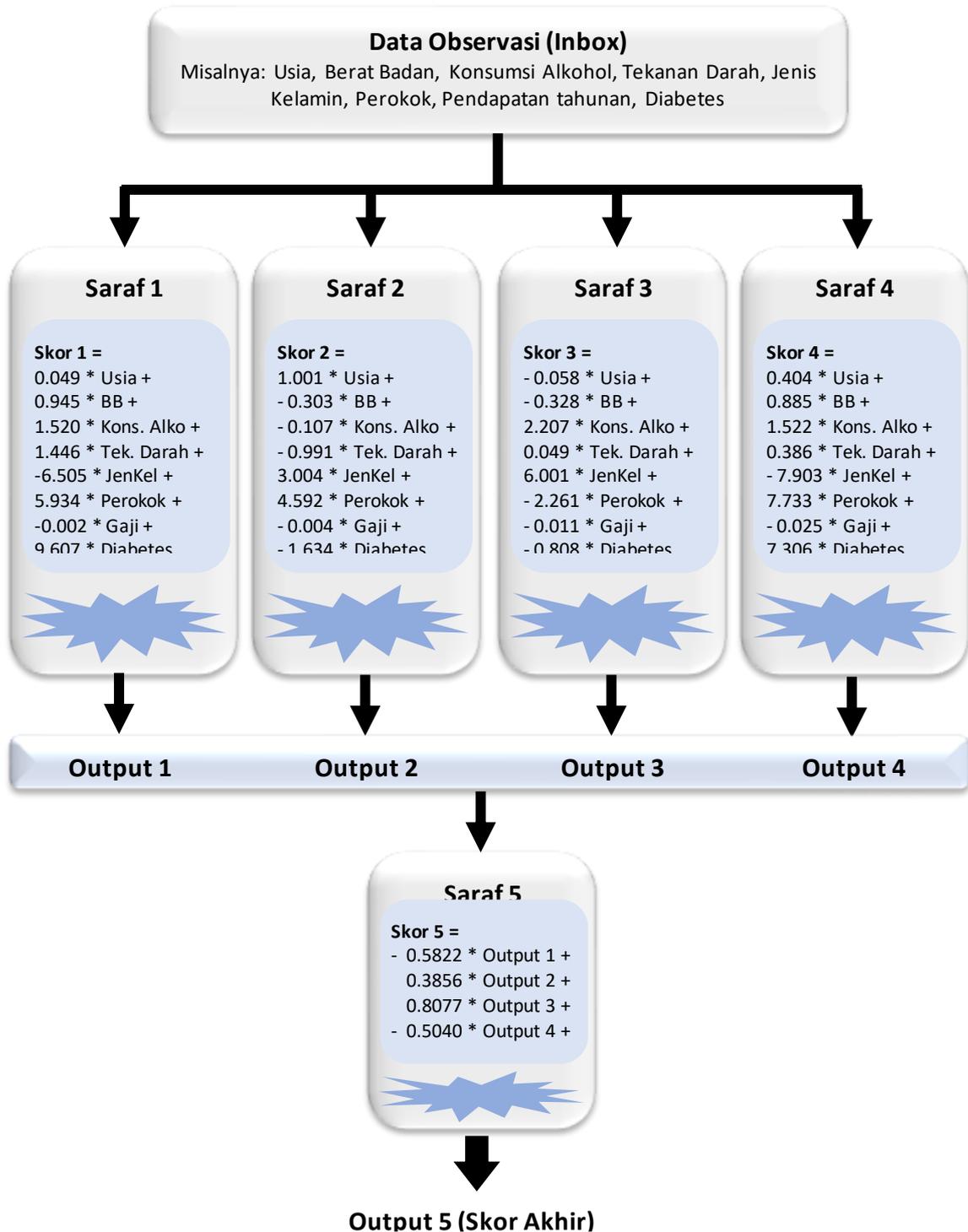


Gambar 8.1 Neuron Buatan

Hal pertama yang perlu dikatakan tentang jaringan saraf adalah bahwa meskipun kadang-kadang disebut-sebut sebagai sesuatu yang sangat kompleks “seperti otak”, jaringan saraf dapat dengan mudah dipahami jika seseorang bersedia meluangkan sedikit waktu dan upaya untuk mempelajarinya. Memang benar bahwa jaringan neural pada umumnya lebih kompleks daripada kartu skor dan pohon keputusan yang telah diperkenalkan pada bab-bab sebelumnya, namun prinsip dasarnya pada akhirnya tidak jauh berbeda – hanya ada lebih dari itu.

Untuk mengilustrasikan cara kerja jaringan saraf, kita akan kembali lagi ke masalah prediksi penyakit jantung yang telah kita gunakan sebelumnya. Untuk memulai, mari kita mulai dengan blok pembangun jaringan saraf – neuron. Neuron dalam istilah pembelajaran mesin bukanlah makhluk hidup yang dikembangkan oleh ilmuwan gila di laboratorium rahasia di suatu tempat, namun merupakan representasi sederhana dan sederhana tentang perilaku neuron alami, yang dibuat menggunakan persamaan yang diimplementasikan sebagai kode komputer. Gambar 8.1 memberikan ilustrasi cara kerja neuron buatan. Pengoperasian neuron pada Gambar 8.1 adalah sebagai berikut:

1. Data observasi, seperti Usia dan BMI, memberikan masukan ke neuron.
2. Setiap masukan dikalikan dengan bobot (bisa positif atau negatif). Untuk data non-numerik seperti jenis kelamin atau perokok, tanda 0/1 digunakan untuk mewakili setiap kondisi. Misalnya, 1 jika perempuan, 0 jika laki-laki.
3. Masukan yang dikalikan dengan bobotnya dijumlahkan untuk mendapatkan skor awal.
4. Setelah skor awal dihitung maka ditransformasikan. Seringkali hal ini untuk memaksa skor berada pada kisaran 0-1. Transformasi ini tidak mutlak penting namun dianggap sebagai “praktik yang baik.” Hal ini agar ketika neuron digabungkan dengan neuron lain untuk menghasilkan jaringan saraf, semua neuron menghasilkan nilai dalam rentang yang sama; yaitu antara 0 dan 1.
5. Versi transformasi dari skor awal adalah skor keluaran yang dihasilkan oleh neuron.



Gambar 8.2 Model jaringan syaraf tiruan

Jadi, neuron dalam konteks pembelajaran mesin bukanlah sesuatu yang misterius atau rumit, itu hanya beberapa rumus sederhana. Yang pertama mengalikan setiap variabel masukan dengan bobot dan menjumlahkan totalnya. Rumus kedua (disebut fungsi aktivasi[31] dalam ilmu data) kemudian memodifikasi (mengubah) nilai ini sehingga berada dalam rentang tetap, biasanya antara 0 dan 1. Sesederhana itu! Bagian cerdasnya adalah menentukan bobot yang seharusnya, yang akan segera kita bahas.

Dengan sendirinya, satu neuron sebenarnya tidak jauh berbeda dengan model tipe kartu skor. Satu-satunya perbedaan material adalah transformasi untuk memaksa skor berada pada kisaran yang tetap. Faktanya, dapat dibuktikan bahwa keluaran yang dihasilkan oleh neuron dan model tipe kartu skor adalah setara.

Untuk menghasilkan model jaringan saraf, sejumlah neuron dihubungkan bersama. Gambar 8.2 memberikan contoh bagaimana hal ini terjadi.

Pada Gambar 8.2, data masukan disuplai ke masing-masing dari empat neuron di lapisan pertama. Setiap neuron mempunyai bobot tersendiri. Bobot tersebut, jika digabungkan dengan data masukan, akan menghasilkan skor dasar (Skor1, Skor2, dll.). Skor ini kemudian ditransformasikan menjadi empat keluaran. Keluaran dari neuron lapisan pertama ini memberikan masukan ke satu neuron di lapisan kedua (Neuron 5). Neuron 5 kemudian mengalikan setiap keluaran dari 4 neuron pertama dengan bobot untuk menghasilkan Score5. Ini kemudian diubah untuk membuat skor akhir dari jaringan.

Fitur utama dari jaringan saraf pada Gambar 8.2 adalah:

- Masukan ke setiap neuron pada lapisan tertentu adalah sama (misalnya Usia, BMI, dll. untuk neuron lapisan pertama), namun bobot dalam setiap neuron berbeda. Oleh karena itu, setiap neuron menghasilkan skor yang berbeda.
- Keluaran (skor) dari kumpulan neuron pertama (lapisan pertama) memberikan masukan kepada kumpulan neuron kedua (lapisan kedua).
- Ada empat neuron di lapisan pertama dalam contoh ini. Dalam praktiknya, jumlah neuron optimal tidak dapat ditentukan sebelumnya. Sebaliknya, sejumlah model yang bersaing dibangun dengan jumlah neuron yang berbeda, dan model dengan kinerja terbaik akan dipilih. Aturan praktis yang berfungsi dengan baik untuk sebagian besar aplikasi bisnis adalah memiliki jumlah item data masukan antara setengah dan dua kali lipat (jadi antara 4 dan 16 dalam contoh ini).
- Ada banyak beban. Gambar 8.2 adalah jaringan yang sangat sederhana. Ia hanya memiliki 8 variabel masukan dan 5 neuron, tetapi terdapat 36 bobot. Aplikasi bisnis tipikal dengan, misalnya, 100 variabel masukan dan 75 neuron, akan memiliki 7.474 bobot.

Setelah jaringan saraf dibuat, jaringan tersebut dapat dinilai dan digunakan dengan cara yang sama seperti kartu skor dan pohon keputusan; yaitu skor dibuat untuk sampel validasi, yang kemudian digunakan untuk menghasilkan tabel distribusi skor.

Kumpulan bobot dalam jaringan mewakili pola yang telah diidentifikasi oleh algoritma pelatihan. Jadi bagaimana algoritma pelatihan menentukan bobot yang seharusnya? Ada banyak algoritme berbeda yang telah dibuat untuk mencari bobot dalam jaringan saraf, namun semuanya cenderung mengadopsi prinsip berikut:

1. Tetapkan setiap bobot nilai acak atau nol.
2. Hitung skor yang dihasilkan oleh jaringan untuk semua kasus dalam sampel pengembangan.
3. Menilai seberapa akurat skor akhir. Misalnya, dengan menilai properti lift dan gain model, sebagaimana dirinci dalam Lampiran A.
4. Sesuaikan bobot untuk meningkatkan akurasi prediksi model.

5. Ulangi langkah 1-4 hingga tidak ada peningkatan signifikan lebih lanjut pada performa model, atau jangka waktu tertentu telah berlalu.

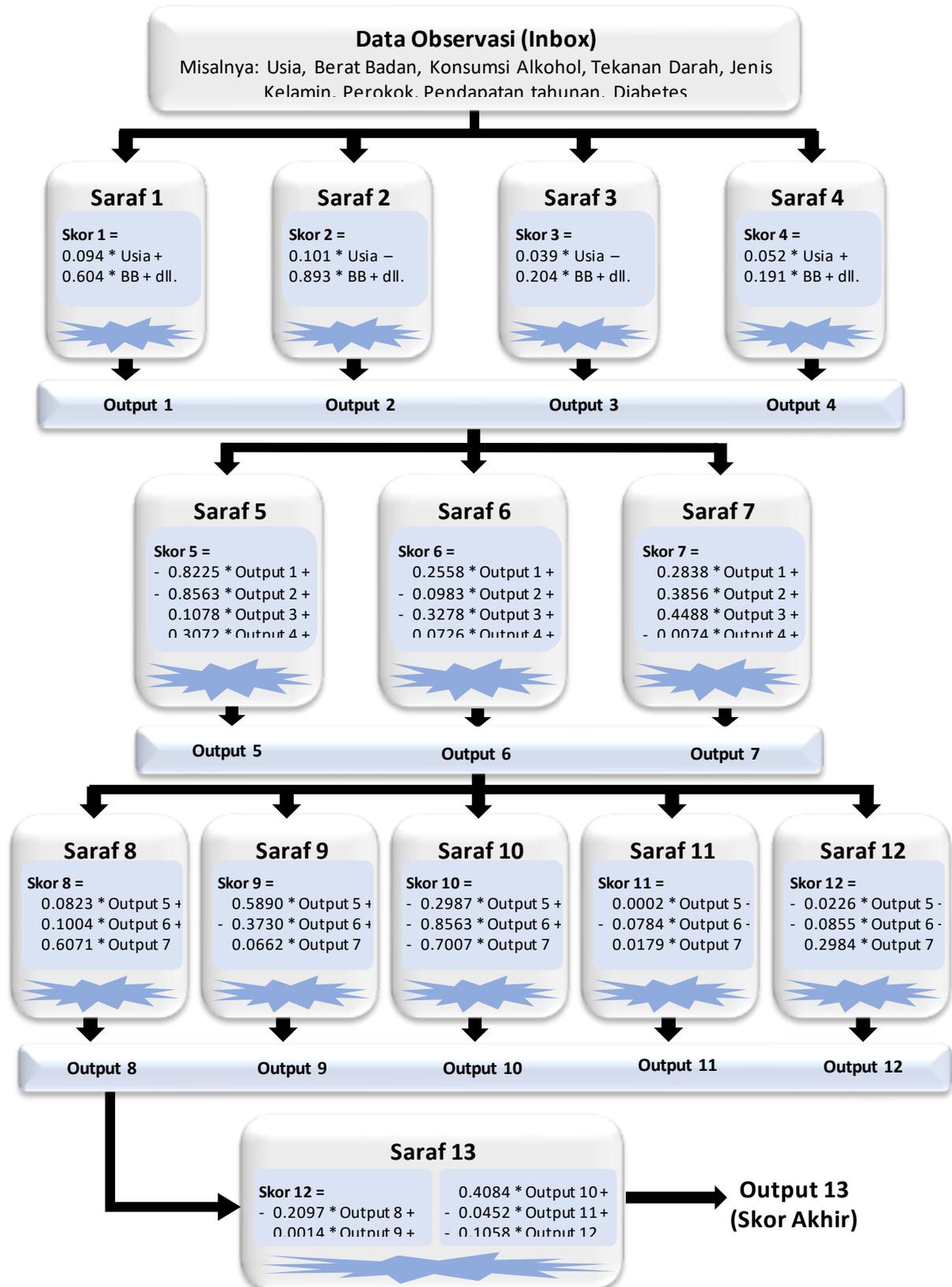
Bagian cerdasnya ada di Langkah 4, bagaimana seseorang menyesuaikan bobotnya. Pendekatan paling sederhana adalah dengan mencoba nilai yang berbeda secara acak dan melihat mana yang terbaik. Namun, hal ini sangat tidak efisien dan kemungkinan besar tidak akan menghasilkan solusi yang baik dalam waktu yang realistis. Semua algoritma pelatihan jaringan saraf praktis lebih pintar dari ini. Mereka mengadopsi berbagai strategi penyesuaian bobot yang berbeda berdasarkan perbedaan kinerja antara setiap iterasi algoritme. Mereka akan membuat penyesuaian besar pada bobot di awal proses ketika mudah untuk mendapatkan peningkatan besar dalam kinerja prediktif, dan kemudian secara bertahap membuat perubahan kecil seiring dengan peningkatan kinerja sehingga mendapatkan solusi optimal.

Jaringan saraf pada Gambar 8.2 memiliki satu neuron di lapisan kedua (Neuron 5) yang menghasilkan skor tunggal. Skor ini mewakili kemungkinan seseorang terkena serangan jantung dalam 5 tahun ke depan. Jika jaringan digunakan untuk suatu permasalahan dimana terdapat beberapa peristiwa yang diprediksi, maka akan ada neuron terpisah untuk mewakili setiap peristiwa. Untuk masalah seperti pengenalan objek, mungkin ada ratusan atau bahkan ribuan skor yang dihasilkan oleh neuron di lapisan terakhir. Setiap skor akan mewakili kemungkinan item tersebut menjadi satu objek tertentu. Objek yang dikaitkan dengan skor tertinggi kemudian diambil sebagai prediksi yang dihasilkan oleh sistem.

Alasan mengapa jaringan saraf sangat populer adalah kemampuannya untuk mendeteksi pola halus dalam data yang mungkin tidak dapat dideteksi oleh metode lain yang lebih sederhana, seperti kartu skor dan pohon keputusan. Artinya, mereka berpotensi menghasilkan prediksi yang lebih akurat. Hal ini dicapai dengan memiliki dua lapisan neuron, dan skor dari lapisan pertama memberikan masukan ke lapisan kedua.

Kelemahan utama jaringan saraf adalah skor yang dihasilkannya tidak intuitif. Ya, kita dapat melihat berbagai bobot dalam model dan memahami penghitungan skor keseluruhan, namun jika saya bertanya kepada Anda variabel input mana pada Gambar 8.2 yang memberikan kontribusi paling besar terhadap skor akhir, maka hal ini kurang jelas dibandingkan variabel input model kartu skor dan pohon keputusan penyakit jantung. Hal ini berpotensi menjadi masalah jika ada persyaratan bisnis atau hukum untuk menjelaskan bagaimana skor akhir model diperoleh.

Pembelajaran mendalam mewakili evolusi terbaru model tipe jaringan saraf. Jaringan saraf pada Gambar 8.2 memiliki dua lapisan neuron dan struktur ini sangat berhasil digunakan di banyak aplikasi jaringan saraf tradisional. Namun, tidak ada alasan mengapa tidak ada lebih banyak lapisan lagi. Perhatikan jaringan pada Gambar 8.3. Jaringan ini memiliki masukan awal yang sama dengan jaringan pada Gambar 8.2. Terdapat juga satu neuron keluaran (Neuron 13) yang memberikan skor model akhir. Namun jaringan pada Gambar 8.3 memiliki 4 lapisan neuron dibandingkan hanya 2 lapisan pada Gambar 8.2.



Gambar 8.3 Jaringan syaraf dalam

Secara teori, tidak ada alasan mengapa jaringan tidak dapat diperluas lebih jauh lagi. Mungkin ada 5, 6, 7, ..., 100+ lapisan jika diinginkan, dengan jumlah neuron berbeda di setiap

lapisan. Apa yang cenderung ditemukan adalah semakin banyak lapisan yang ditambahkan, maka kemampuan jaringan untuk mengidentifikasi pola yang kompleks dan/atau halus meningkat. Semakin banyak lapisan, semakin “dalam” jaringannya. Secara umum, apa pun yang memiliki lebih dari 2 atau 3 lapisan dapat diklasifikasikan sebagai jaringan “dalam”, namun tidak ada definisi tunggal yang diterima.

Selain memperluas jumlah neuron dan lapisan dalam jaringan, penelitian lain yang terkait dengan pembelajaran mendalam juga mempertimbangkan bagaimana neuron dalam jaringan terhubung. Dalam jaringan standar, seperti pada Gambar 8.2 dan 8.3, semua neuron di setiap lapisan terhubung ke semua neuron di lapisan berikutnya. Namun, konfigurasi lain juga dimungkinkan. Misalnya, tidak menghubungkan semua masukan ke semua neuron di lapisan pertama (jaringan saraf berbelit-belit). Ini bekerja sangat baik untuk jenis masalah tertentu seperti pengenalan gambar.

Varian lain pada jaringan saraf standar adalah membuat putaran umpan balik sedemikian rupa sehingga keluaran neuron di lapisan selanjutnya bertindak sebagai masukan ke lapisan sebelumnya (jaringan saraf berulang). Hal ini memungkinkan untuk menggabungkan latensi, untuk memberikan representasi waktu, yang tidak ditangkap oleh pendekatan pembelajaran mesin tradisional.

Artinya dalam praktiknya, jaringan saraf berulang berfungsi dengan baik ketika informasi terkandung dalam rangkaian peristiwa dalam sampel pengembangan yang digunakan untuk melatih jaringan. Jika Anda menggunakan pembelajaran mesin untuk menguraikan tulisan tangan satu huruf pada satu waktu, fakta bahwa huruf sebelumnya dalam sebuah kata diucapkan, huruf “Z” akan menjadi prediktor penting untuk huruf berikutnya, yang hampir pasti merupakan huruf vokal atau huruf. Huruf “S.” Jaringan saraf tradisional untuk pengenalan tulisan tangan tidak akan mempertimbangkan hubungan berurutan ini. Google Translate adalah salah satu contoh produk komersial yang menggunakan jaringan berulang.

Model jaringan saraf paling kompleks yang digunakan saat ini, seperti yang dikembangkan oleh anak perusahaan Google DeepMind untuk mengalahkan pemain Go terbaik di dunia, menggabungkan fitur-fitur ini dan memiliki jutaan neuron yang terhubung di lusinan lapisan.

Model pembelajaran mendalam yang kompleks, berdasarkan jaringan saraf, mendorong banyak penerapan kecerdasan buatan yang canggih, namun untuk banyak jenis masalah, hal ini berlebihan. Model yang sangat kompleks hanya diperlukan untuk permasalahan yang sangat kompleks. Ada pepatah lama: “Anda tidak memerlukan palu godam untuk memecahkan masalah.” Faktanya, jika Anda mencoba menerapkan pendekatan pembelajaran mesin yang sangat kompleks pada masalah yang sangat sederhana, terkadang hasil yang Anda dapatkan justru lebih buruk dibandingkan jika model yang sangat sederhana digunakan. Hal ini mungkin tampak berlawanan dengan intuisi, namun akan ada kecenderungan model untuk menemukan pola palsu yang sebenarnya tidak ada, sedangkan hal ini kecil kemungkinannya terjadi pada model sederhana.

Salah satu cara saya mencoba dan memikirkan masalah ini adalah dengan mempertimbangkan apakah segala sesuatunya dapat dianggap sebagai jenis prediksi/masalah AI yang “sederhana” atau “kompleks”:

- Masalah AI yang “sederhana”. Ini memiliki tujuan tunggal atau hal yang ingin ditentukan, yang mudah diukur. Untuk masalah klasifikasi, ini berarti bahwa hasilnya dapat dinyatakan sebagai “Ya” atau “Tidak” sederhana seperti apakah seseorang gagal membayar pinjamannya atau tidak? Untuk soal regresi, apakah hasilnya dapat dinyatakan sebagai bilangan tunggal? Misalnya berapa banyak uang yang dibelanjakan seseorang di toko terakhir kali mereka berbelanja.
- Masalah AI yang “kompleks”. Ini biasanya memiliki tujuan yang beragam dan sangat beragam. Contoh klasik dari masalah “kompleks” adalah pengenalan objek, terjemahan bahasa, dan permainan (misalnya catur atau Go). Hal ini juga mencakup masalah yang memerlukan beberapa pendekatan pembelajaran mesin untuk digunakan secara bersamaan, seperti pada kendaraan otonom dan asisten pribadi digital.

Hal ini tidak selalu sejelas ini, tetapi sebagai aturan praktis, ini bukanlah awal yang buruk. Sebagian besar aplikasi bisnis yang mungkin Anda temui (pemasaran target, penilaian kredit, siapa yang harus dipekerjakan/dipecat, pencocokan di situs kencan, dll.) termasuk dalam kategori “AI Sederhana”. Jaringan dengan beberapa ratus neuron yang tersebar di dua lapisan saja sudah lebih dari cukup dan mungkin akan menghasilkan kinerja yang serupa atau sedikit lebih baik daripada model jenis kartu skor yang dirancang dengan baik. Namun, Anda tidak pernah bisa yakin. Seperti aspek pembelajaran mesin lainnya, saran bagi siapa pun yang membuat model prediktif untuk Anda adalah mencoba sejumlah pendekatan berbeda dan kemudian melihat mana yang paling memenuhi kebutuhan Anda. Hal ini baik dalam hal akurasi prediksi mentah maupun persyaratan/kendala bisnis. Jika ada persyaratan bisnis atau hukum agar keluaran model dapat dijelaskan oleh orang awam, dapatkah Anda menjelaskan secara sederhana mengapa model tersebut menghasilkan skor yang dapat dijelaskan?

Seorang data scientist harus selalu menggunakan pendekatan sederhana seperti kartu skor sebagai tolok ukur awal, dan kemudian mencoba memperbaikinya menggunakan metode yang lebih kompleks/canggih.

BAB 9

PEMBELAJARAN TANPA PENGAWASAN

Teknik pembelajaran mesin yang telah kita bahas sejauh ini mengasumsikan bahwa terdapat sampel pengembangan yang berisi data observasi dan data hasil. Untuk masalah penyakit jantung, setiap data pasien dalam sampel perkembangan diberi label untuk mengidentifikasi apakah mereka kemudian mengidap penyakit jantung. Penerapan berbagai algoritma pembelajaran mesin dijelaskan untuk menemukan hubungan yang ada antara observasi dan data hasil. Hubungan ini kemudian ditangkap dalam bentuk model seperti kartu skor, pohon keputusan, atau jaringan saraf. Model yang dihasilkan kemudian dapat digunakan untuk memprediksi kemungkinan orang terkena penyakit jantung di masa depan, berdasarkan apa yang diketahui tentang mereka saat ini.

Pembelajaran mesin diterapkan pada data berlabel; dimana setiap kasus dalam sampel pengembangan memiliki data observasi dan hasil, disebut sebagai pembelajaran yang diawasi. Sebagian besar aplikasi AI/pembelajaran mesin yang akan Anda temui di dunia bisnis seperti pemasaran target, deteksi penipuan, dan pemeriksaan karyawan adalah contoh pembelajaran yang diawasi. Jika Anda memiliki data yang diberi label, maka pendekatan yang diawasi biasanya merupakan pendekatan yang tepat untuk diikuti dan akan memberikan hasil terbaik. Namun ada beberapa jenis masalah yang data hasilnya sangat terbatas atau tidak ada – yang ada hanyalah data observasi yang tidak berlabel. Oleh karena itu, meskipun buku ini terutama membahas pembelajaran yang diawasi, ada baiknya meluangkan waktu untuk mempertimbangkan apa yang dapat dilakukan dalam situasi seperti itu.

Ketika data hasil tidak tersedia, serangkaian teknik yang disebut pembelajaran tanpa pengawasan dapat diterapkan. Ciri utama pembelajaran tanpa pengawasan adalah bahwa pembelajaran ini terutama digunakan untuk penemuan pengetahuan; yaitu mendeteksi fitur-fitur menarik dalam data. Hal ini berbeda dengan pembelajaran terbimbing yang biasanya dikaitkan dengan memprediksi suatu hasil menggunakan data observasi yang cocok.

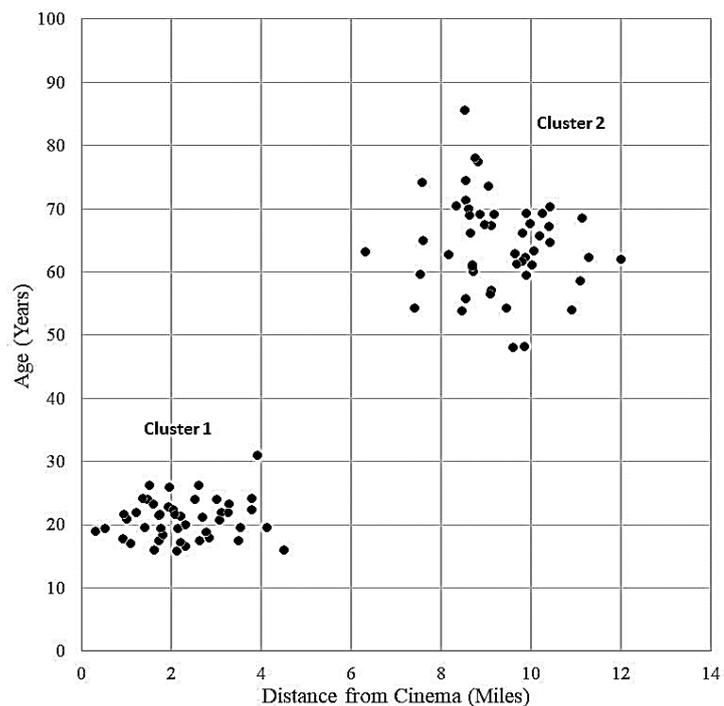
Artinya dalam praktiknya, pendekatan pembelajaran tanpa pengawasan sangat bagus dalam membantu mengidentifikasi dan mengelompokkan berbagai hal berdasarkan kesamaan di antara keduanya, namun pendekatan tersebut tidak memberi Anda prediksi yang memberi tahu Anda bagaimana seseorang atau sesuatu akan berperilaku. Oleh karena itu, tujuan dan keluaran dari pembelajaran yang diawasi dan tidak diawasi sangatlah berbeda. Dengan pembelajaran yang diawasi, permainan akhir adalah semacam model prediktif. Model tersebut kemudian dapat digunakan untuk memprediksi perilaku kasus-kasus baru ketika kasus-kasus tersebut muncul. Dengan pembelajaran tanpa pengawasan, keluarannya merupakan representasi struktur data yang merupakan hal yang berbeda sama sekali.

Jenis pembelajaran tanpa pengawasan yang paling umum digunakan saat ini adalah pengelompokan. Tujuan pengelompokan adalah untuk mengidentifikasi kesamaan dan/atau hubungan dalam data sehingga Anda dapat mengelompokkan (mengelompokkan) kasus serupa menjadi satu. Identy adalah karena kasus dalam kluster tertentu memiliki banyak

atribut yang sangat mirip, maka Anda dapat memperlakukan semua orang di klaster tersebut dengan cara yang sama.

Untuk mengilustrasikan tindakan pengelompokan, pertimbangkan Rebecca, manajer kompleks bioskop yang berlokasi di kota universitas. Sebagian besar pelanggannya menghadiri pertunjukan di malam hari dan akhir pekan. Hal ini mungkin karena mereka berada di tempat kerja atau sekolah sepanjang waktu. Namun, dia mengadakan beberapa pertunjukan awal pada pertengahan minggu, meskipun pertunjukan ini tidak terlalu banyak dihadiri. Apa yang Rebecca ingin lakukan adalah memahami lebih banyak tentang tipe orang yang menghadiri pertunjukan tengah minggu ini. Dia kemudian dapat menargetkan tawaran promosi kepada orang-orang serupa di komunitas lokal, yang mungkin tertarik untuk menonton film pada saat itu, dan dengan demikian meningkatkan jumlah penonton bioskop tersebut.

Ini adalah masalah yang tidak diawasi karena Rebecca tidak memiliki informasi tentang bagaimana masyarakat merespons aktivitas pemasaran sebelumnya. Tidak ada daftar orang-orang yang pernah menerima penawaran promosi di masa lalu dan tanda (label) terkait yang menunjukkan apakah mereka menerima tawaran tersebut atau tidak. Hal ini mengesampingkan pengembangan semua jenis model prediktif yang dibangun menggunakan pendekatan pembelajaran yang diawasi.



Gambar 9.1 Usia dan Jarak dari bioskop

Namun Rebecca tahu siapa saja di komunitas lokal yang baru-baru ini pergi ke bioskop. Pasalnya, saat ini kebanyakan orang mencatat kehadirannya melalui metode pembayaran yang digunakan atau informasi lokasi dari ponsel cerdasnya. Setelah Anda memilikinya, sejumlah informasi lain tentang orang-orang tersebut dapat diperoleh dari berbagai sumber.

Dua informasi yang menurutnya sangat relevan adalah usia penonton film dan seberapa jauh mereka tinggal dari bioskop. Gambar 9.1 menunjukkan plot kedua item ini untuk sampel acak pelanggan awal.

Dari Gambar 9.1 terlihat jelas dua pengelompokan (cluster). Masyarakat muda yang tinggal berdekatan dan masyarakat lanjut usia, sebagian besar sudah memasuki usia pensiun, dan tinggal agak jauh.

Terlihat dari Gambar 9.1 bahwa hanya sedikit pelanggan muda yang tinggal dekat dengan bioskop. Selain itu, rentang usia kelompok 1 juga cukup sempit; hampir seluruhnya adalah remaja akhir hingga pertengahan dua puluhan. Fitur lain yang menarik dari Gambar 10 adalah hampir tidak ada seorang pun berusia antara tiga puluh dan lima puluh tahun yang menghadiri pertunjukan *early bird*.

Rebecca tidak tahu mengapa cluster tersebut muncul seperti itu, namun dia berani menebak bahwa penonton bioskop yang lebih muda sebagian besar adalah mahasiswa yang memiliki jadwal mengajar yang berbeda. Mereka mungkin juga lebih bergantung pada transportasi umum, sehingga kurang nyaman melakukan perjalanan jauh untuk sampai ke bioskop. Demikian pula, orang lanjut usia mungkin memiliki lebih banyak fleksibilitas dalam kesehariannya dibandingkan orang kebanyakan, memiliki lebih banyak waktu luang, memiliki mobil, dan bersedia melakukan perjalanan lebih jauh. Mereka juga cenderung tinggal di rumah yang lebih besar di pinggiran kota, dibandingkan dengan mahasiswa yang cenderung tinggal di akomodasi universitas yang dekat dengan bioskop.

Meskipun sangat sederhana, informasi pada Gambar 9.1 dapat membantu Rebecca merumuskan strategi pemasaran yang disesuaikan dengan masing-masing dua kelompok pelanggan. Hal ini dapat terjadi dalam hal:

- Film apa yang menjadi sasaran individu di setiap kelompok.
- Penawaran perjalanan mencerminkan berbagai moda transportasi yang digunakan untuk pergi ke bioskop.
- Diskon tambahan yang mungkin menarik bagi setiap kelompok (mungkin soda/bir untuk anak muda dan kopi/teh/anggur untuk lansia).

Yang sama pentingnya, Rebecca mengetahui bahwa penonton bioskop *early bird* hanya berisi sedikit orang “usia paruh baya”. Oleh karena itu, ia tidak boleh menyia-nyaiakan anggaran pemasarannya untuk mencoba membujuk mereka agar datang ke bioskop.

Skenario sinema mengilustrasikan konsep clustering dengan cara yang sangat sederhana, hanya menggunakan dua item informasi. Prinsip yang sama dapat diterapkan ketika ada ratusan atau ribuan item data yang tersedia tentang manusia, namun tidak mudah untuk merepresentasikan cluster secara grafis ketika Anda memiliki lebih dari 2 atau 3 item data. Mencari tahu berapa banyak cluster yang ada dalam data dan apa saja fitur dari cluster tersebut, merupakan bagian mendasar dari proses clustering.

Ada banyak produk komersial berdasarkan clustering. Produk Mosaik Experian menetapkan 1,8 juta kode pos (kode pos) di Inggris ke salah satu dari 66 cluster. Experian memberi judul pada setiap klaster seperti “Elit kota”, “Pekerja berpenghasilan rendah”, dan “Kakek-nenek klasik” untuk mewakili tipe orang yang biasanya tinggal di wilayah tersebut.

Informasi geo-demografis tentang rata-rata usia, kekayaan, jumlah anak, penggunaan kredit, dan sebagainya disediakan untuk penduduk yang tinggal di kode pos setiap cluster. Organisasi kemudian menyesuaikan strategi keterlibatan pelanggan mereka agar sesuai dengan demografi orang-orang di setiap kode pos. Hal ini berhasil karena orang-orang yang tinggal dalam rumah tangga di lingkungan yang sama cenderung memiliki demografi yang sama. Jika Anda adalah seorang bankir bergaji tinggi yang tinggal di wilayah yang tergolong “elit pusat kota”, kemungkinan besar tetangga Anda juga merupakan profesional bergaji tinggi dibandingkan keluarga yang bertahan hidup dengan tunjangan atau upah minimum.

Banyak algoritma pengelompokan didasarkan pada prinsip meminimalkan jarak antar kasus dalam sampel pengembangan. Jarak dalam hal ini tidak selalu mengacu pada jarak fisik, namun seberapa berbeda dua kasus untuk item data tertentu. Jarak antara dua orang yang berusia 22 dan 24 tahun lebih kecil dibandingkan dengan dua orang yang berusia 18 dan 75 tahun. Jika kita berbicara tentang kebiasaan merokok, maka dua orang perokok mempunyai jarak nol sedangkan seorang perokok dan bukan perokok tidak. Jarak antara dua orang dengan tingkat gaji yang sama biasanya lebih kecil dari jarak antara dua orang dengan tingkat gaji yang berbeda dan seterusnya.

Sebelum algoritma pengelompokan diterapkan, setiap item data dalam sampel pengembangan diubah menjadi numerik standar. Sebelum algoritma pengelompokan diterapkan, setiap item data dalam sampel pengembangan diubah ke skala numerik standar. Misalnya, usia dan pendapatan dapat ditransformasikan sehingga nilai rata-rata masing-masingnya adalah nol, dan sebagian besar kasus berada di antara nilai -1 dan +1. Begitu pula dengan indikator 0/1 yang digunakan untuk mewakili hal-hal seperti apakah seseorang perokok atau bukan. Dengan cara ini, perbedaan untuk setiap item data memiliki bobot yang sama ketika algoritma pengelompokan diterapkan. Jika proses standarisasi ini tidak terjadi, maka pengaruh item data dengan skala besar seperti pendapatan dalam dolar, akan mengerdilkan item data yang nilainya jauh lebih kecil seperti jumlah tanggungan.

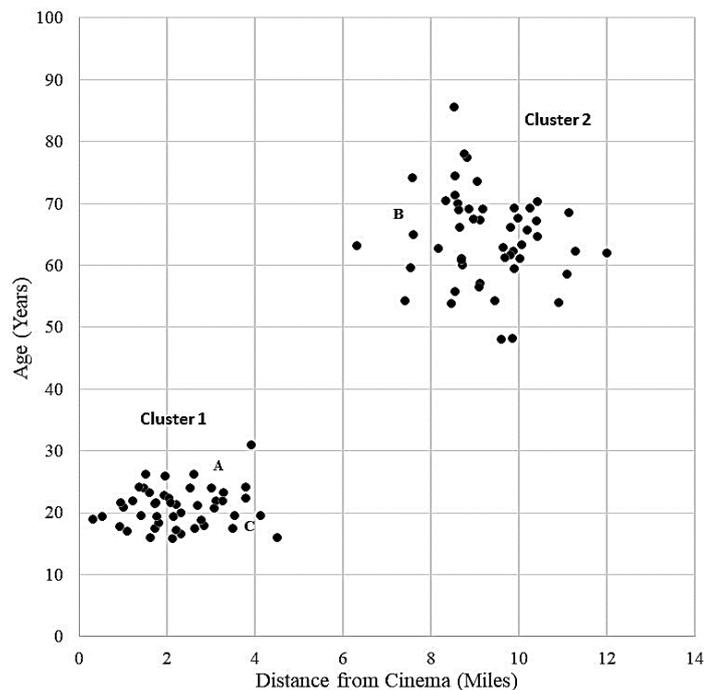
Ada banyak algoritma pengelompokan berbeda yang tersedia, tetapi salah satu yang paling populer disebut pengelompokan K-means. Ini adalah jenis pengelompokan yang dijelaskan Experian dalam desain Produk Mosaik mereka. Algoritme pengelompokan K-means berfungsi untuk menetapkan setiap kasus dalam sampel Anda ke salah satu dari K cluster yang berbeda, di mana K adalah nilai yang dipilih oleh pengguna. Algoritme melakukan hal ini dengan menugaskan kasus ke dalam cluster sehingga jumlah seluruh jarak individu dalam setiap cluster dapat diminimalkan.

Dengan clustering tidak ada model yang dihasilkan di akhir proses. Yang ada hanyalah pengidentifikasi yang menyatakan cluster K mana yang telah ditugaskan observasinya. Anda mungkin ditempatkan di klaster 4, namun saya sebaliknya, telah ditugaskan ke klaster 7. Ketika ada kasus baru yang perlu ditugaskan ke sebuah klaster, salah satu dari dua pendekatan dapat diadopsi.

1. Algoritma pengelompokan dijalankan kembali. Cluster tersebut ditugaskan kembali dengan memasukkan kasus-kasus baru.

2. Jarak antara masing-masing kasus baru dan pusat setiap cluster dihitung. Kasus ditugaskan ke cluster yang pusatnya juga paling dekat.

Opsi pertama adalah optimal dalam hal penetapan klaster karena klaster akan berkembang seiring waktu seiring dengan semakin banyaknya kasus yang disajikan. Namun, pengelompokan memerlukan banyak daya komputer. Jika kasus perlu ditetapkan dalam waktu (mendekati) waktu nyata, opsi 2 lebih praktis. Mari kita ilustrasikan hal ini dengan sebuah contoh, seperti yang ditunjukkan pada Gambar 9.2.



Gambar 9.2 Input kasus ke Cluster

Pada Gambar 9.2 huruf A, B dan C mewakili pelanggan baru yang akan ditugaskan ke salah satu dari dua cluster yang telah diidentifikasi. Setelah mereka ditugaskan ke sebuah cluster, Rebecca dapat menargetkan mereka dengan aktivitas pemasaran yang sesuai. Sangat mudah untuk melihat bahwa pelanggan A dan C lebih dekat ke cluster 1, sedangkan pelanggan B paling dekat dengan cluster 2. Jumlah jarak dalam cluster akan diminimalkan jika A dan C ditugaskan ke cluster 1 dan pelanggan B ke cluster 2.

Pembuatan profil pelanggan yang diterapkan pada pemasaran adalah salah satu penerapan pengelompokan yang paling dikenal, namun pendekatan berbasis pengelompokan juga berhasil diterapkan pada banyak masalah lainnya. Salah satu contohnya adalah pengelompokan dokumen. Dalam bidang-bidang seperti hukum dan penelitian akademis, ada keharusan untuk secara teratur menelusuri tumpukan literatur terbitan yang terus bertambah untuk menemukan informasi yang berkaitan dengan subjek atau jenis penelitian tertentu. Beberapa bidang studi mencakup jutaan makalah dan artikel. Basis data matematika zbMATH berisi sekitar 4 juta entri dengan ulasan atau abstrak dari 3.000 jurnal dan 180.000 buku. Di bidang kedokteran, lebih dari 800.000 makalah akademis diterbitkan setiap tahun.

Cara tradisional mencari dokumen mirip dengan menggunakan Google atau Bing – mengetikkan kata kunci ke dalam mesin pencari database, yang akan menemukan makalah yang cocok dengan kata kunci tersebut. Namun, istilah pencarian tidak selalu mengembalikan semua dokumen yang relevan dan banyak yang mengembalikan dokumen yang tidak relevan tetapi mengandung kata kunci dalam konteks yang berbeda. Beberapa dokumen, terutama dokumen lama, mungkin tidak memiliki kata kunci apa pun yang terkait dengannya atau menggunakan terminologi yang berbeda, yang berarti dokumen tersebut terlewatkan dalam pencarian.

Pengelompokan dokumen pada dasarnya tidak berbeda dengan contoh bioskop yang telah kita bahas sebelumnya. Alih-alih geodemografi; jumlah kata, struktur kalimat, dan fitur dokumen lainnya menyediakan item data yang digunakan dalam proses pengelompokan.

Pengelompokan dokumen juga dapat diterapkan pada tweet, umpan berita, postingan blog, dan media lain yang berubah dengan cepat secara real-time. Gambar dan video juga dapat dikategorikan dengan cara serupa. Organisasi berita menggunakan pendekatan ini untuk secara otomatis menandai postingan baru tentang topik tertentu saat postingan tersebut muncul. Hal ini agar mereka dapat segera memasukkannya ke dalam publikasi media mereka. Perusahaan media sosial dan pemerintah menggunakan metode serupa (serta model prediktif) untuk mendukung identifikasi konten ilegal atau tidak diinginkan.

Pendekatan pembelajaran yang diawasi bekerja dengan baik ketika terdapat sejumlah besar data observasi dan setiap observasi memiliki hasil yang jelas dan tidak ambigu. Jika tidak ada hasil yang tersedia, maka pembelajaran tanpa pengawasan terkadang terbukti bermanfaat; meskipun dalam konteks yang agak berbeda. Namun terdapat permasalahan ketika pada awalnya tidak ada data observasi atau data hasil, namun proses pembelajaran dapat menilai kinerjanya berdasarkan kasus per kasus seiring berjalannya waktu. Model ini disesuaikan setiap saat, berdasarkan beberapa ukuran keberhasilan atau imbalan, yang dihitung setiap kali suatu tugas dicoba. Jenis pembelajaran mesin ini disebut pembelajaran penguatan.

Contoh pembelajaran penguatan yang agak terlalu sederhana adalah melatih jaringan saraf untuk menemukan titik tertinggi di medan lokal menggunakan jumlah langkah sesedikit mungkin. Pada awalnya, yang diketahui hanyalah lokasi peta saat ini (bujur, lintang, dan ketinggian di atas permukaan laut). Jaringan dapat menghasilkan 4 kemungkinan skor, masing-masing mewakili gerakan maju, mundur, kiri, atau kanan. Keadaan saat ini; yaitu lokasi saat ini, memberikan masukan (data observasi). Keadaan yang dihasilkan, yaitu lokasi baru setelah perpindahan yang disarankan oleh jaringan dijalankan, adalah hasilnya.

Setelah setiap gerakan, keadaan baru dinilai. Apakah posisi kita saat ini lebih baik (lebih tinggi di atas permukaan laut) dibandingkan sebelumnya? Jika jawabannya ya, maka algoritme dianggap telah bekerja dengan baik - ukuran imbalan atau keberhasilannya tinggi. Jika tidak, maka algoritma tersebut dianggap berkinerja buruk – ukuran keberhasilannya rendah. Algoritme kemudian menyesuaikan bobot model berdasarkan tingkat keberhasilan yang dihasilkan dari tindakannya.

Versi yang lebih halus dari algoritma ini juga akan mempertimbangkan sejarah pergerakan masa lalu; yaitu dimana saja, sebagai bagian dari data masukan. Masuk akal juga untuk memasukkan pandangan jangka panjang tentang kesuksesan di masa depan, bukan hanya posisi satu langkah ke depan. Misalnya memberikan penalti terhadap keberhasilan jika lokasi yang sama dikunjungi lebih dari satu kali atau memberikan imbalan (peningkatan keberhasilan) karena memasuki area baru yang belum dijelajahi meskipun perpindahan awal ke area tersebut tidak menemukan titik yang lebih tinggi pada awalnya.

Ilmuwan data sering kali berbicara tentang tiga jenis pembelajaran mesin yang berbeda; yaitu pembelajaran yang diawasi, tidak diawasi dan penguatan. Namun, pembelajaran penguatan memiliki banyak kesamaan dengan pembelajaran yang diawasi. Seperti halnya pembelajaran yang diawasi, algoritme yang digunakan dalam pembelajaran penguatan menghasilkan model, biasanya didasarkan pada beberapa bentuk jaringan saraf. Demikian pula, model disesuaikan berdasarkan penilaian terhadap hasil yang dihasilkan dari serangkaian data masukan tertentu. Semakin banyak iterasi algoritme yang terjadi dan bobot dalam model disempurnakan, performa model akan meningkat.

Istilah “hadiah/sukses” dan “kegagalan/penalti” cenderung diterapkan pada pembelajaran penguatan, namun konsep ini sangat mirip dengan “akurasi” dan “kesalahan” yang digunakan untuk menggambarkan seberapa baik model prediktif, yang dikembangkan menggunakan pembelajaran yang diawasi, tampil. Kesalahan kecil antara nilai aktual dan prediksi dalam pembelajaran yang diawasi; yaitu prediksi yang sangat akurat, secara umum setara dengan keberhasilan tingkat tinggi dalam pembelajaran penguatan. Dengan cara yang sama, klasifikasi yang salah atau prediksi yang tidak akurat dalam pembelajaran yang diawasi dapat dianalogikan dengan rendahnya penghargaan/kegagalan dalam pembelajaran penguatan.

Contoh yang bagus tentang perbedaan antara pembelajaran yang diawasi dan pembelajaran penguatan adalah bagaimana pelatihan model dilakukan untuk membuat program bermain catur. Pendekatan yang diawasi akan mengambil ribuan gerakan permainan (atau urutan gerakan) dari permainan yang dimainkan sebelumnya sebagai data observasi, dengan data hasil yang diberi label memberikan indikasi apakah gerakan tersebut baik atau tidak. Skor yang dihasilkan oleh model digunakan untuk menunjukkan bidak mana yang harus dipindahkan dan ke mana. Algoritme kemudian menemukan bobot model yang menghasilkan rangkaian gerakan terbaik secara keseluruhan, diukur terhadap gerakan yang terdapat dalam sampel pengembangan.

Dengan pembelajaran penguatan, pada awalnya tidak ada data yang diberikan – tidak ada sama sekali! Skor model masih menunjukkan langkah mana yang harus dilakukan seperti pendekatan yang diawasi. Namun, pada awalnya hal ini akan lebih atau kurang acak, mengingat tidak ada data untuk melatih model tersebut. Setiap kali suatu langkah dilakukan, status dewan (keadaan baru) dievaluasi ulang. Algoritme kemudian menyesuaikan bobot dalam model berdasarkan seberapa sukses perpindahannya.

Mengevaluasi seberapa sukses suatu langkah dalam catur adalah hal yang rumit dan sering kali memperhitungkan kemungkinan keadaan di masa depan serta keadaan saat ini.

Namun, untuk keperluan contoh ini, kriteria keberhasilan yang sederhana adalah perbedaan nilai bidak masing-masing pemain yang tersisa di papan setelah gerakan dilakukan. Jika perpindahan Anda menghasilkan barang yang bernilai tinggi, tetapi tidak kehilangan barang tersebut, maka itu adalah kesuksesan besar. Bergerak dan kemudian salah satu bidak Anda diambil adalah sebuah kegagalan, dengan tingkat keberhasilan yang rendah. Keberhasilan atau kegagalan terbesar adalah kehilangan raja – skakmat. Dengan cara ini, dengan menilai setiap gerakan dan menyesuaikan bobot model, program belajar dengan sendirinya tanpa perlu diberikan informasi sebelumnya.

Keuntungan utama pembelajaran penguatan dibandingkan pembelajaran yang diawasi adalah tidak ada batasan pada rangkaian gerakan yang dapat dieksplorasi saat algoritme memodifikasi bobot dalam model. Dengan pembelajaran yang diawasi, Anda dibatasi pada contoh berlabel yang tersedia dalam sampel pengembangan. Untuk permainan seperti catur, bahkan sampel pengembangan yang sangat besar, yang berisi semua gerakan dan hasil dari jutaan permainan, hanya akan berisi sebagian kecil dari semua kemungkinan gerakan. Hal ini ditunjukkan dengan sangat efektif ketika tim AI DeepMind Google menggunakan dua algoritma penguatan untuk bermain melawan satu sama lain. Model yang dihasilkan tidak hanya mengungguli program catur terbaik yang ada pada saat itu, namun selama proses tersebut algoritma menemukan strategi permainan yang benar-benar baru, yang sebelumnya tidak diketahui oleh para grandmaster manusia.

Pembelajaran penguatan telah menghasilkan banyak kegembiraan karena proses pembelajarannya sangat mirip dengan cara makhluk hidup belajar. Jika saya mencoba belajar juggling maka saya tidak memiliki informasi apa pun untuk memulai. Saya mencoba, saya gagal, saya mencoba dan gagal lagi, saya mencoba dan saya berhasil melakukan juggling selama beberapa detik, dan sebelum Anda menyadarinya, saya melakukan juggling seperti seorang profesional! Dengan setiap upaya, otak saya secara tidak sadar belajar lebih banyak tentang lingkungannya dan menyempurnakan tindakannya berdasarkan setiap keberhasilan atau kegagalan; yaitu lamanya waktu saya mampu menjaga bola tetap di udara tanpa menjatuhkannya.

Pembelajaran penguatan memiliki potensi, namun memiliki kelemahan dan lebih cocok untuk beberapa tugas dibandingkan tugas lainnya. Salah satu masalahnya adalah algoritma pelatihan dibatasi oleh kecepatan proses coba-coba. Jika algoritme penguatan dilatih untuk membuat keputusan pinjaman hipotek, maka waktu antara tindakan yang diambil dan penilaian seberapa sukses tindakan tersebut dapat memakan waktu berbulan-bulan atau bertahun-tahun. Akibatnya, proses pelatihan akan memakan waktu terlalu lama untuk dapat digunakan secara praktis.

Memutuskan siapa yang akan meminjamkan uang juga merupakan masalah yang jauh lebih sederhana daripada catur dan biasanya terdapat banyak data pengembangan berlabel yang tersedia. Oleh karena itu, Anda akan mendapatkan hasil yang lebih baik dengan lebih cepat dengan pendekatan yang diawasi untuk masalah seperti penilaian risiko kredit. Masalah yang sama berlaku untuk semua jenis masalah dimana kesuksesan tidak dapat diukur dengan cepat. Jika kita kembali ke contoh catur, program bermain catur DeepMind perlu memainkan

68 juta permainan agar menjadi sebaik itu. Ia hanya bisa melakukan ini dengan bermain melawan komputer lain, sehingga permainan dapat dimainkan dalam sepersekian detik. Ia tidak akan bisa mencapai hasil yang sama dalam waktu yang wajar jika ia bermain melawan pesaing manusia di dunia nyata. Proses pelatihan akan menjadi terlalu lambat.

Kelemahan lain dari pembelajaran penguatan adalah biaya kegagalan selama proses pelatihan. Jika Anda memasukkan algoritme pembelajaran penguatan ke dalam kebijakan perekrutan organisasi Anda, maka pada awalnya Anda akan mendapatkan banyak karyawan yang buruk yang akan menyebabkan berbagai macam masalah. Dengan cara yang sama, Anda tidak ingin menghubungkan algoritma pembelajaran penguatan langsung ke kontrol pesawat penumpang.

Oke, jadi menemukan puncak bukit fiksi dan menjadi manusia super dalam catur itu baik-baik saja, tapi apa aplikasi bisnis pembelajaran penguatan di dunia nyata? Sejujurnya, jumlahnya saat ini tidak terlalu besar. Secara pribadi, saya tidak pernah terlibat dalam memberikan solusi yang menggunakannya dan saya tidak tahu siapa pun yang pernah melakukannya. Saya juga telah melakukan cukup banyak penelitian dan menemukan banyak makalah akademis yang sangat menarik dan artikel pers yang antusias, tetapi contoh dunia nyata jauh lebih sedikit. Kesimpulan saya adalah jumlah solusi berbasis pembelajaran penguatan yang digunakan saat ini sangat kecil dibandingkan dengan jumlah solusi pembelajaran yang diawasi. Jika saya harus menghitungnya, saya rasa lebih dari 95% aplikasi dunia nyata menggunakan pembelajaran yang diawasi dan sebagian besar dari 5% sisanya adalah pendekatan tanpa pengawasan berdasarkan pengelompokan. Namun, hal ini dapat berubah seiring berjalannya waktu seiring dengan semakin berkembangnya teknologi. Secara intuitif, jika Anda memiliki algoritma yang dapat membangun sesuatu yang bermain catur atau bermain lebih baik daripada orang lain, maka algoritma tersebut pasti memiliki potensi di bidang lain.

Salah satu bidang di mana pembelajaran penguatan tampaknya mempunyai dampak adalah peningkatan efisiensi sistem kontrol yang canggih dalam lingkungan yang tidak pasti atau kacau. Sistem kompleks seperti jaringan listrik, sistem pemanas, dan server farm, memiliki banyak kontrol yang disesuaikan untuk mengelola berbagai bagian sistem. Hubungan antara pengendalian dan kinerja sistem tidak selalu jelas. Ini bukan kasus sederhana dimana satu pengendalian hanya berdampak pada satu hal – semuanya saling berhubungan. Menyesuaikan tekanan air dalam sistem pendingin pembangkit listrik untuk meningkatkan kinerja turbin berdampak negatif pada efisiensi trafo. Sama seperti soal catur, di mana jumlah kemungkinan permainan yang dapat dimainkan hampir tak terhingga, dalam pembangkit listrik terdapat kombinasi pengaturan kontrol yang jumlahnya hampir tak terhingga. Pembelajaran penguatan menemukan cara yang lebih baik dalam mengatur kontrol melalui trial and error, untuk mengoptimalkan efisiensi sistem secara keseluruhan. Hal ini mencerminkan cara seorang insinyur berpengalaman menggunakan pengetahuan dan intuisi mereka, yang dipelajari selama bertahun-tahun dalam praktik, mengenai dampak luas sistem dari penyesuaian kontrol yang berbeda.

Bidang lain di mana pembelajaran penguatan menjanjikan adalah bidang robotika. Tujuannya adalah untuk melatih robot agar dapat melakukan tugas manual rumit yang sebelumnya hanya dapat dilakukan oleh orang yang terlatih. Perangkat robot diberi tugas untuk melakukan sesuatu seperti membalik burger, menyortir barang, atau menyusun rak, dan melalui uji coba, mereka berpotensi belajar melakukan hal ini dengan sangat efektif.

BAB 10

BAGAIMANA MEMBANGUN MODEL PREDIKTIF

Dalam bab ini kita akan melebur melalui lubang kunci ke sisi lain... Kita akan melintasi batas dari dunia sehari-hari yang waras dan masuk akal ke dalam dunia multi-dimensi yang terpelintir dari ilmuwan data; yaitu spesialis teknis dengan gelar matematika dan PhD. Dalam perjalanan kami, kami akan:

1. Jelaskan proses yang harus diikuti oleh data scientist saat menerapkan pembelajaran mesin untuk mengembangkan model prediktif, yang kemudian akan digunakan sebagai bagian dari sistem pengambilan keputusan otomatis.
2. Jelaskan proses ini sedemikian rupa sehingga orang normal dapat memahaminya. Hal ini untuk membantu Anda mengawasi apa yang dilakukan para data scientist. Hal ini pada gilirannya akan memastikan bahwa Anda mendapatkan apa yang Anda minta, bukan apa yang menurut data scientist dibutuhkan oleh organisasi Anda.

Seperti yang telah kita bahas di awal buku ini, proyek AI/pembelajaran mesin yang cenderung berhasil adalah proyek di mana pengguna bisnis dan ilmuwan data bekerja sama. Menurut pengalaman saya, ketika data scientist dibiarkan “melanjutkan saja” dan memberikan hasil sebagai sebuah kudeta di akhir proyek, yang cenderung dihasilkan adalah solusi yang kurang optimal. Dalam kasus terburuk, mereka mungkin gagal memberikan sesuatu yang berguna sama sekali.

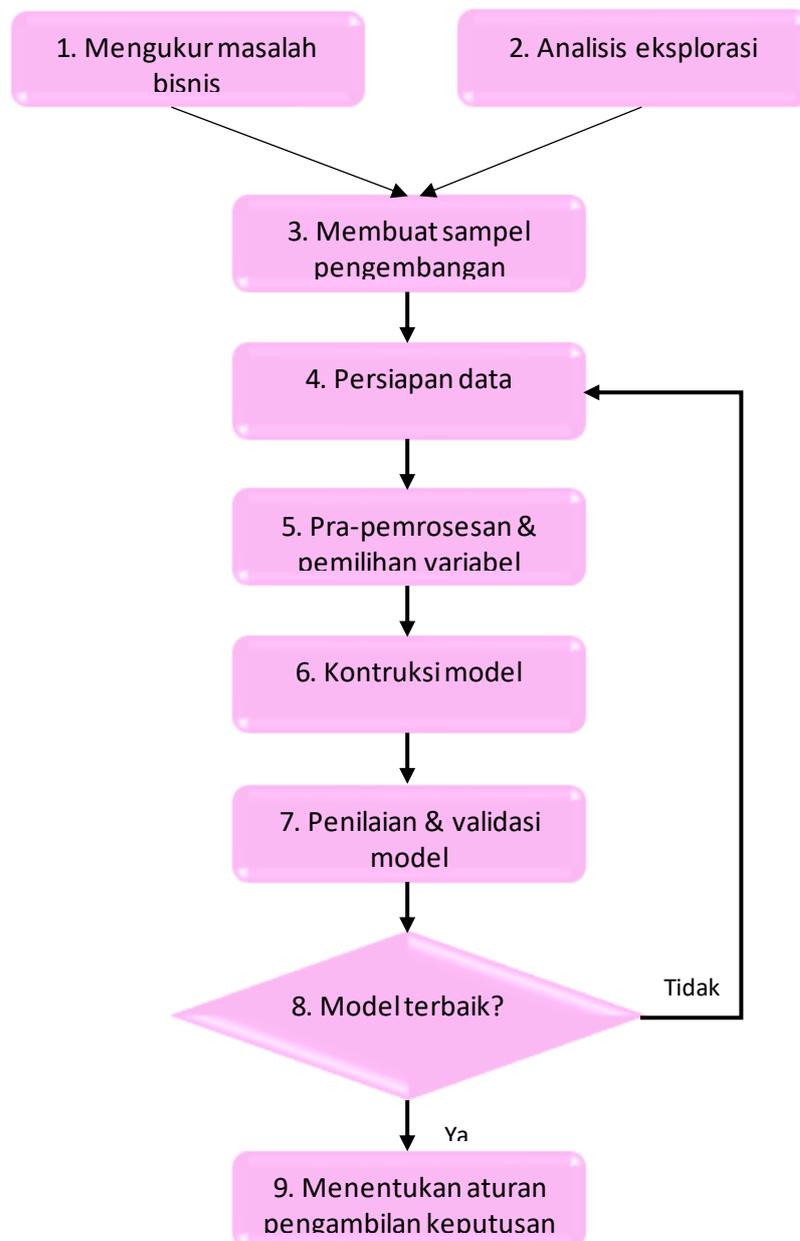
Sebagian besar ilmuwan data yang pernah bekerja dengan saya sangat menyadari perlunya melibatkan klien mereka dalam proses pembelajaran mesin dan telah bekerja keras untuk mendapatkan masukan bisnis untuk proyek mereka. Namun, sangat menyedihkan untuk mengatakan bahwa saya juga telah melihat lebih dari satu contoh upaya yang disengaja oleh para ilmuwan data untuk membingungkan orang-orang normal dengan pembicaraan teknologi atau menghindari berbicara dengan mereka secara bersamaan. Hal ini dimaksudkan agar mereka dapat mengerjakan semua hal penting yang berhubungan dengan matematika tanpa harus berurusan dengan para pebisnis jahat yang harus menjalankan organisasi sehari-hari. Saat mencoba menentukan apakah seorang data scientist melakukan tugasnya, a turan yang sangat sederhana yang harus diikuti adalah:

- Jika mereka tidak dapat menjelaskan kepada Anda dalam bahasa Inggris apa yang mereka lakukan dan alasannya, maka mereka telah gagal dalam tugasnya.

Ilmuwan data juga perlu memberikan pemahaman yang jelas kepada bisnis tentang dampak pekerjaan mereka di dunia nyata. Sebagai pengguna bisnis, saya tidak tertarik dengan kenyataan bahwa model prediktif 95% akurat atau 10% lebih diskriminatif dibandingkan proses yang akan digantikannya. Yang ingin saya ketahui adalah: berapa banyak uang yang dapat dihasilkan oleh model ini bagi saya, berapa banyak nyawa yang dapat diselamatkan, berapa banyak staf yang harus saya pekerjakan, atau apa pun yang akan berdampak pada tujuan bisnis saya. Ini adalah ilmuwan data yang buruk yang tidak dapat mengungkapkan

manfaat dari solusi mereka dalam kaitannya dengan manfaat yang diberikan kepada klien mereka.

Namun, ini bekerja dua arah. Tidaklah adil untuk menyerahkan segalanya pada data scientist. Bukanlah suatu permintaan yang tidak masuk akal bagi pengguna bisnis untuk berupaya mendapatkan pemahaman dasar tentang cara data scientist menjalankan proses pembelajaran mesin. Seiring berjalannya proyek, harus ada pertukaran informasi dua arah dan pertemuan rutin di mana temuan awal dan hasil, permasalahan dan isu dipresentasikan dan didiskusikan dengan cara yang tidak berfokus pada bisnis. Hal ini kemudian memandu bagaimana proyek ini dijalankan.



Gambar 10.1 Pemrosesan pembelajaran mesin

Dengan mengingat hal ini, saya akan menghindari mempelajari berbagai algoritma yang digunakan untuk pembelajaran mesin seperti yang digunakan untuk melatih jaringan saraf atau memutuskan bagaimana pembagian dalam pohon keputusan ditetapkan. Bagi Anda yang tertarik dengan detail teknisnya, algoritme paling umum yang digunakan untuk pembelajaran mesin semuanya terdokumentasi dengan cukup baik. Demikian pula, implementasi algoritme ini tersedia dalam paket statistik komersial yang dapat Anda beli atau tersedia secara gratis melalui perangkat lunak sumber terbuka (misalnya pustaka kode R dan Python). Anda tidak perlu mengetahui cara membuat kode algoritma pembelajaran mesin tertentu; Anda hanya perlu mengetahui cara memanggil fungsi atau prosedur yang relevan dalam paket perangkat lunak pilihan Anda.

Gambar 10.1 menguraikan langkah-langkah penting dalam proses pembelajaran mesin untuk mengembangkan model prediktif.

Hal pertama yang perlu dilakukan oleh data scientist adalah mengukur masalah bisnis (Langkah 1); yaitu untuk mewakili apa yang ingin dicapai oleh bisnis sebagai angka atau skor sederhana. Direktur penjualan toko ritel mungkin memulai percakapan dengan data scientist dengan pernyataan seperti:

“Kami menginginkan model prediktif untuk mengidentifikasi pelanggan terbaik kami.”

Jadi pertanyaan pertama yang diajukan adalah:

“Apa maksudmu dengan yang terbaik?”

Yang ditanggapi oleh direktur penjualan:

“Pelanggan yang akan memberikan kontribusi finansial terbesar bagi bisnis kami.”

Oke, sepertinya masuk akal, tapi apakah “kontribusi finansial” berhubungan dengan penjualan kotor atau laba bersih? Apakah kita berbicara tentang kontribusi per kunjungan toko, per klik, atau per satuan waktu (minggu/bulan/tahun)? Apakah kita melihat kontribusi individu atau rumah tangga ketika kita berbicara tentang “pelanggan”? dan daftarnya terus berlanjut. Seorang ilmuwan data perlu memberikan definisi yang tepat tentang apa yang diinginkan bisnis, dan kemudian mengungkapkannya dengan cara yang dapat diwakili oleh skor yang dihasilkan oleh model yang mereka bangun.

Dalam contoh ini, mungkin setelah berdiskusi dengan direktur pemasaran, disepakati bahwa model tersebut akan dirancang untuk memprediksi total pengeluaran kotor individu per kuartal tahun. Pembelanjaan kotor triwulanan dipilih karena sejalan dengan periode pelaporan organisasi dan total pendapatan (jumlah pengeluaran) merupakan indikator kinerja utama bagi direktur penjualan.

Belanja individu dibandingkan belanja rumah tangga dipilih karena disepakati bahwa meskipun belanja rumah tangga secara teori akan lebih baik, basis data pelanggan organisasi disusun pada tingkat individu, bukan rumah tangga. Tidak mungkin menerapkan model berdasarkan pengeluaran rumah tangga dengan infrastruktur TI organisasi yang ada saat ini. Bisakah TI ditingkatkan untuk memberikan gambaran rumah tangga? Tidak, tentu saja bukan

bagian dari proyek ini karena perusahaan belum mengalokasikan anggaran apa pun untuk perubahan TI tahun ini.

Dengan adanya informasi ini, ilmuwan data harus menentukan bahwa ini adalah masalah jenis regresi untuk memprediksi jumlah yang akan dibelanjakan setiap pelanggan selama jangka waktu perkiraan tiga bulan.

Sejalan dengan menentukan cara mengekspresikan tujuan bisnis secara kuantitatif, data scientist perlu melakukan analisis eksplorasi untuk memahami data organisasi dan aset TI, cara kerjanya, dan cara keduanya berinteraksi satu sama lain serta dengan bisnis yang lebih luas. Ini termasuk:

- Basis data apa yang dimiliki organisasi?
- Apa tujuan setiap database dan proses bisnis apa yang memanfaatkannya?
- Apa isi databasenya? yaitu item data apa dan format item data tersebut. Misalnya saja data numerik atau tekstual.
- Bagaimana setiap database diperbarui dan dipelihara? Sebagian besar organisasi memiliki gabungan database pelanggan “waktu nyata” yang diperbarui segera setelah ada perubahan, dan database batch yang lebih jarang diperbarui; misalnya semalaman atau di akhir bulan.
- Sistem pengambilan keputusan apa yang sudah ada? Akankah model tersebut diimplementasikan dalam sistem yang sudah ada atau akankah diperlukan fungsionalitas baru?
- Data apa yang saat ini digunakan organisasi untuk tujuan pengambilan keputusan? Meskipun data ini merupakan proses manual yang kuno, kemungkinan besar data ini akan sangat berharga untuk tujuan pembelajaran mesin karena organisasi telah mengidentifikasinya sebagai data yang berguna.

Di dunia yang sempurna, setiap organisasi akan memiliki dokumentasi komprehensif (metadata) yang menyediakan inventarisasi aset datanya. Hal ini berkaitan dengan jenis data yang dimilikinya, bagaimana data tersebut diformat, dari mana data tersebut berasal, dan bagaimana data tersebut digunakan. Namun, dalam praktiknya, informasi semacam ini sering kali tidak lengkap, tidak benar, atau ketinggalan zaman. Banyak informasi berguna yang tersimpan di kepala pengguna bisnis, yang menggunakannya secara rutin sebagai bagian dari pekerjaan sehari-hari mereka. Mungkin ada item data bernama “Demografi Pelanggan3” yang berisi huruf A, B,..., G. Apa maksudnya? Hal ini tidak didokumentasikan di mana pun, namun Claire di Bagian Operasional mengetahui bahwa hal tersebut sesuai dengan tingkat pendapatan yang mereka gunakan. “A” berarti pendapatan tahunan di bawah Rp. 20.000.000, “B” Rp. 20.000 – 35.000.000 dan seterusnya.

Banyak organisasi juga menyimpan data yang tidak lagi mereka pahami. Namun, mereka takut untuk menghapusnya jika hal tersebut merupakan hal penting yang menjadi sandaran proses penting bisnis. Potensi sumber data baru, baik internal maupun eksternal, juga harus diperiksa selama tahap analisis eksplorasi karena hal ini dapat meningkatkan kemampuan prediksi model apa pun yang dikembangkan.

Secara keseluruhan, diperlukan upaya besar untuk mengeksplorasi alam semesta dari mana data dapat dikumpulkan dan untuk memutuskan apa yang berpotensi digunakan dan apa yang tidak. Setelah seorang ilmuwan data memiliki pemahaman yang baik tentang data apa saja yang dimiliki suatu organisasi dan bagaimana strukturnya, mereka dapat melanjutkan ke Tahap 3 dan memikirkan tentang membuat sampel pengembangan yang akan digunakan dalam proses pembelajaran mesin.

Pendekatan yang sangat naif dalam membangun model prediktif adalah dengan memasukkan setiap informasi yang dimiliki organisasi ke dalam proses pembelajaran mesin. Dalam praktiknya, banyak data yang kedaluwarsa, tidak stabil, atau tidak sesuai. Oleh karena itu, hal tersebut harus dikeluarkan dari sampel pengembangan jika hal tersebut menurunkan kinerja proses pembelajaran mesin.

Beberapa alasan umum untuk mengecualikan data dari proses pembelajaran mesin adalah:

- **Data kedaluwarsa.** Jika data tersebut tidak lagi relevan dengan dunia saat ini maka jangan pertimbangkan. Data tentang cara orang menggunakan ponselnya pada masa sebelum ponsel pintar ada tidak akan terlalu berguna untuk memberi tahu Anda tentang penggunaan ponsel cerdas saat ini.
- **Tidak mewakili populasi sasaran.** Sampel pengembangan hanya boleh berisi kasus-kasus yang mencerminkan populasi dimana model tersebut akan diterapkan. Jika model pendapatan dikembangkan secara eksklusif untuk pelanggan online, Anda harus mengecualikan transaksi di dalam toko jika pola pembelian berbeda untuk kedua saluran ini. Demikian pula, jika Anda mengembangkan sistem pengenalan suara dari sampel suara, semua sampel harus terkait dengan ucapan. Jika Anda memiliki contoh anjing menggonggong atau kebisingan lalu lintas, maka hal tersebut perlu dihilangkan.
- **Stabilitas data.** Apakah data yang digunakan dalam proses pembelajaran mesin akan tersedia saat model diterapkan? Jika suatu item data tidak tersedia pada saat implementasi, maka item tersebut harus dikeluarkan.
- **Alasan hukum dan etika.** Apakah Anda ingin menggunakan data pribadi sensitif seperti ras, jenis kelamin, dan orientasi seksual? Jenis data ini bisa sangat bersifat prediktif, namun penggunaannya mungkin ilegal. Jika hal tersebut sah, maka penggunaan data tersebut dapat menimbulkan risiko reputasi jika diketahui bahwa Anda memperlakukan kelompok tertentu secara berbeda dibandingkan kelompok lainnya. Jika strategi diskon berdasarkan skor model mengakibatkan Anda hanya menawarkan diskon kepada pria heteroseksual – hal ini akan menjadi kontroversial.
- **Kasus deterministik.** Jika Anda memiliki pelanggan yang Anda ketahui dengan pasti bagaimana Anda akan memperlakukan mereka, maka Anda tidak memerlukan model prediktif untuk ini. Memasukkan mereka ke dalam sampel pembangunan hanya akan melemahkan kekuatan model tersebut, sebagaimana diukur pada populasi lainnya.
- **Ketidaktejelasan.** Prinsip dasar pembelajaran mesin adalah Anda harus sangat berhati-hati dalam membuat model menggunakan data yang tidak sepenuhnya Anda pahami. Hal ini terutama berlaku untuk model yang tunduk pada peraturan industri atau yang secara strategis penting bagi keberhasilan organisasi. Jika data scientist memberi Anda model

menggunakan data yang tidak dapat mereka jelaskan, maka ini adalah risiko yang harus Anda khawatirkan.

Jika suatu organisasi memiliki database yang sangat besar, hal ini mungkin menimbulkan masalah dalam hal sumber daya komputer yang diperlukan untuk menganalisis database tersebut sebagai bagian dari proses pembelajaran mesin. Bukan hal yang aneh jika hanya sampel; katakanlah 1:10 atau 1:100 kasus yang akan dipilih untuk dimasukkan dalam sampel pembangunan.

“Penganut Big Data Purists” mungkin berpendapat bahwa Anda harus menganalisis semua data yang tersedia karena kegagalan melakukan hal ini dapat mengakibatkan hilangnya pola halus, dan karenanya terciptalah model yang kurang optimal. Di satu sisi mereka benar. Jika Anda mengambil sampel daripada menggunakan seluruh populasi, model yang dihasilkan tidak akan seakurat itu. Namun, ini merupakan hasil yang semakin berkurang. Untuk hampir semua permasalahan bisnis praktis, manfaat yang diperoleh dari menangani lebih dari beberapa puluh ribu kasus sangatlah kecil. Jika Anda mengambil sampel dengan cara yang tepat, maka Anda akan mendapatkan 99% manfaat dari model yang dikembangkan menggunakan 100.000 kasus dibandingkan dengan model yang dikembangkan menggunakan 10 juta kasus. Dengan kata lain, hanya karena Anda memiliki data dalam jumlah besar tidak berarti Anda perlu berinvestasi pada perangkat keras/perangkat lunak khusus untuk dapat menerapkan pembelajaran mesin – Anda hanya perlu cerdas dalam melakukannya dan fokus pada data yang benar-benar penting.

Proses tahap keempat adalah persiapan data. Hal ini diperlukan karena data mentah mungkin memiliki format yang buruk, mengandung kesalahan, atau tidak memberikan representasi data terbaik; yaitu representasi yang berbeda akan memudahkan proses pembelajaran mesin untuk menemukan pola penting dalam data. Ada empat tugas utama yang terjadi selama persiapan data:

1. **Pembuatan data baru.** Usia lebih bersifat prediktif dibandingkan Tanggal Lahir untuk sebagian besar jenis masalah. Oleh karena itu, usia mungkin perlu dihitung jika hanya Tanggal Lahir yang tersedia. Jika Anda memiliki data tentang transaksi kartu kredit, sebaiknya buat variabel ringkasan yang mewakili rata-rata pembelanjaan per transaksi, jumlah transaksi bulan lalu, waktu sejak transaksi terakhir, dan seterusnya.
2. **Pembersihan data.** Data mungkin hilang atau salah. Tujuan utama dari persiapan data adalah untuk mengidentifikasi di mana jenis data ini berada, dan untuk menghapusnya (pengecualian) atau memformat ulang data tersebut; yaitu mengkode ulang semua data yang salah/hilang dengan nilai standar. Jika tanggal lahir yang digunakan untuk menghitung usia berada di masa depan atau mungkin jauh di masa lalu, Anda mungkin ingin menentukan nilai default katakanlah '-1' untuk usia, untuk mewakili kasus ini.
3. **Konsolidasi.** Merupakan praktik yang baik untuk merepresentasikan data serupa dengan cara yang serupa. Jika Anda memiliki item data bernama Tipe Properti, yang nilai “D”, “Det” dan “DH” semuanya memiliki arti yang sama; yaitu “Rumah Terpisah”, maka nilai-nilai ini harus dikonsolidasikan menjadi satu nilai yang digunakan untuk mewakili seluruh kasus Rumah Terpisah.

4. **Konversi ke numerik.** Algoritma pembelajaran mesin seperti data numerik. “Ya”/”Tidak”/”Mungkin” akan diubah menjadi tanda 0/1/2. Untuk teks atau ucapan yang kompleks seperti feed Twitter, maka salah satu pendekatannya adalah dengan menyertakan tanda untuk menunjukkan apakah kata kunci tertentu ditampilkan atau tidak, atau untuk menghasilkan hitungan berapa kali kata tertentu muncul.

Setelah sampel data yang sesuai telah disiapkan, maka inilah waktunya untuk pra-pemrosesan data dan pemilihan variabel awal (Tahap 5 pada Gambar 10.1). Dengan pra-pemrosesan data, seseorang memasukkan data ke dalam format yang paling sesuai dengan algoritma pembelajaran mesin yang dipilih.

Pra-pemrosesan biasanya melibatkan dua proses: standardisasi dan transformasi. Standardisasi memproses penskalaan data sehingga semuanya memiliki nilai yang serupa. Jika Anda memiliki dua item data seperti pendapatan kotor tahunan dalam dolar dan tinggi badan dalam meter, maka salah satu pendekatan standardisasi adalah dengan mengurangi rata-rata masing-masing dan kemudian membaginya dengan kisaran nilai (atau ukuran lainnya[50]). Dengan cara ini, keduanya akan berpusat pada nol dan memiliki rentang yang hampir sama.

Jika kita berbicara tentang penerapan pembelajaran mesin pada sesuatu seperti sistem pengenalan objek, maka prosedur standarisasi serupa diterapkan pada gambar yang membentuk sampel pengembangan. Dalam pengenalan objek, data pengembangan berasal dari kumpulan piksel di setiap gambar. Data tentang warna dan intensitas setiap piksel memberikan masukan untuk proses pembelajaran mesin. Gambar dalam sampel pengembangan mungkin diambil dalam kondisi pencahayaan berbeda. Oleh karena itu, intensitas piksel pada setiap gambar disesuaikan (distandarkan) dengan rata-rata intensitas piksel pada seluruh gambar.

Pemilihan variabel (reduksi variabel) adalah tentang memutuskan item data apa yang akan dibuang dan apa yang harus dipertahankan serta disajikan ke algoritma pembelajaran mesin. Alasan utama melakukan hal ini adalah meskipun komputer saat ini sangat canggih, seringkali terdapat terlalu banyak data untuk diproses dalam waktu yang realistis. Seringkali, terdapat ribuan atau bahkan jutaan item data individual yang telah dibuat selama persiapan data dan pra-pemrosesan data. Dalam kebanyakan kasus, hanya segelintir saja yang benar-benar terbukti penting dalam menghasilkan prediksi yang akurat.

Pemilihan variabel biasanya merupakan proses yang relatif cepat dan sederhana. Berbagai uji statistik dilakukan untuk menentukan seberapa baik korelasi setiap data observasi dengan hasil yang diprediksi. Item data hanya disimpan jika terbukti ada korelasinya. Biasanya, tidak lebih dari antara 1% dan 10% item data dipertahankan setelah proses ini. Ada juga metode seperti Analisis Komponen Prinsip (PCA) yang berupaya mengubah sekumpulan besar item data awal menjadi kumpulan item baru yang lebih kecil. Item data baru inilah yang kemudian digunakan untuk konstruksi model.

Langkah 6 dari proses ini adalah konstruksi model. Ini sering kali merupakan bagian paling menarik dari proses pembelajaran mesin bagi seorang data scientist. Hal inilah yang sering dibicarakan oleh para ilmuwan data, meskipun konstruksi model hanya mewakili beberapa persen dari keseluruhan proses end-to-end. Selama konstruksi model, data

pengembangan disajikan ke perangkat lunak pembelajaran mesin, algoritme yang relevan dijalankan, dan pada akhirnya model prediktif dihasilkan. Ini biasanya terlihat seperti kartu skor, pohon keputusan, dan jaringan saraf yang telah kita lihat sebelumnya di bab sebelumnya.

Perangkat lunak pembelajaran mesin modern sangat otomatis. Berbagai macam algoritma tersedia hanya dengan menekan satu tombol. Banyak juga yang menyediakan fitur yang berhubungan dengan persiapan data, pra-pemrosesan, dan pemilihan variabel. Akibatnya, beberapa orang berpendapat bahwa Anda tidak memerlukan gelar PhD di bidang statistik atau pembelajaran mesin untuk dapat mengembangkan model prediktif, dan di satu sisi hal itu benar. Namun, sedikit pengetahuan adalah hal yang berbahaya. Jika Anda tidak memiliki latar belakang teknis yang relevan maka Anda tidak boleh menyiapkan atau memproses data terlebih dahulu dengan tepat. Demikian pula, biasanya ada serangkaian opsi berbeda yang dapat dipilih untuk algoritme tertentu dan Anda perlu mengetahui bagaimana setiap opsi dapat memengaruhi solusi yang diberikan. Demikian pula, Anda mungkin tidak dapat menafsirkan keluaran dari perangkat lunak atau mengetahui tindakan apa yang harus Anda ambil berdasarkan apa yang disampaikan oleh keluaran tersebut kepada Anda.

Algoritma pembelajaran mesin hanyalah seperangkat aturan dan pernyataan logika yang diterapkan dengan cara tertentu untuk membuat model prediktif. Terkadang ada yang salah. Sangat mungkin perangkat lunak (dengan benar) melaporkan bahwa model yang dibuat sangat akurat padahal sebenarnya model tersebut tidak berguna.

Pertimbangkan model prediktif, yang dikembangkan menggunakan pembelajaran mesin, yang memprediksi apakah orang akan gagal membayar hipoteknya. Model ini akan digunakan untuk memutuskan pemohon hipotek mana yang akan menerima dana di muka dan pemohon hipotek mana yang harus ditolak kreditnya. Model ini sangat penting untuk berhasil karena miliaran dolar akan dipertaruhkan jika uang dipinjamkan kepada orang yang salah.

Perangkat lunak tersebut melaporkan bahwa model tersebut secara tepat memprediksi gagal bayar hipotek dengan akurasi 97%. Apakah ini model yang bagus? Jawaban yang naif (dan salah) adalah ya, tentu saja – hampir 100% dan Anda tidak bisa mendapatkan jawaban yang lebih baik dari itu. Selain itu, Anda mungkin mengharapkan model sampah akan melakukan tugasnya dengan benar sekitar 50% dari keseluruhan waktu; yaitu kinerjanya akan serupa dengan tebakan acak.

Jadi, modelnya jauh lebih bagus dari itu bukan? Mari kita mundur selangkah dan melihat contoh pengembangannya. Ini berisi informasi historis tentang permohonan hipotek dan kinerja pembayaran selanjutnya. Katakanlah 98% orang membayar hipotek mereka tepat waktu dan 2% gagal bayar. Saya dapat dengan mudah mengungguli model prediktif hanya dengan mengatakan bahwa saya memperkirakan semua hipotek akan dilunasi.

Saya hanya akan salah dalam 2% kasus yang default. Keakuratan prediksi saya adalah 98% Pasti ada yang salah karena aturan sederhana saya “terima semua orang” lebih akurat daripada aturan yang diperoleh menggunakan algoritme pembelajaran mesin canggih! Ini

hanyalah salah satu contoh hal-hal yang perlu dipertimbangkan ketika menilai seberapa baik model prediktif dan apakah model tersebut telah dibangun dengan benar.

Setelah model dibangun menggunakan sampel pengembangan, maka diperlukan upaya untuk menilai kinerja model (Langkah 7). Dari sudut pandang teknis, hal ini terutama melibatkan penilaian keakuratan prediksi model pada sampel data yang berbeda. Hal ini harus benar-benar independen dari data yang digunakan untuk membangun model sehingga dapat memberikan pandangan yang adil dan tidak memihak tentang bagaimana kinerja model ketika diterapkan secara operasional pada kasus-kasus baru. Khususnya, jika sampel pengembangan dibuat beberapa bulan sebelumnya, maka model tersebut harus dinilai menggunakan data yang lebih baru untuk memastikan bahwa model tersebut terus berfungsi seperti yang diharapkan. Seorang data scientist dapat melaporkan performa model yang diukur pada sampel pengembangan, namun mereka harus mendasarkan semua penilaian akhir performa model pada satu atau beberapa sampel validasi independen.

Aktivitas validasi lainnya mungkin lebih bersifat kualitatif, untuk memeriksa batasan hukum apa pun yang mungkin berlaku terhadap penggunaan jenis data tertentu dan untuk memastikan bahwa data yang digunakan dalam model akan tersedia saat model diterapkan; yaitu jika model akan diintegrasikan ke dalam salah satu sistem organisasi, maka data yang diperlukan untuk menghitung skor model harus tersedia dalam sistem tersebut.

Pembelajaran mesin adalah proses berulang. Seringkali, banyak model dibangun menggunakan varian algoritma yang berbeda dan/atau representasi data yang berbeda sebelum model akhir disepakati. Dengan lebih banyak waktu dan sumber daya, selalu mungkin untuk menemukan peningkatan kecil lainnya dalam akurasi prediksi. Namun, di dunia nyata, waktu dan sumber daya terbatas. Oleh karena itu, keputusan perlu diambil mengenai kapan harus menghentikan proses pembelajaran mesin, meninjau model yang telah dikembangkan, dan memutuskan mana yang akan diterima sebagai model akhir Anda.

Dari sudut pandang bisnis, penilaian model akhir harus dinyatakan dalam indikator kinerja utama dan manfaat yang diperoleh. Hal ini seharusnya sudah dipertimbangkan sejak awal proses. Pertimbangkan layanan saran otomatis yang telah dikembangkan untuk penyedia kartu kredit guna meningkatkan efisiensi dalam pusat kontak pelanggan mereka. Perusahaan yang mengembangkan solusi ini memutuskan untuk menyebutnya sebagai “Penasihat Robot Otonom” agar terdengar seperti produk tipe AI modern yang keren dan seksi.

Bagian pengambilan keputusan dari sistem telah dibangun menggunakan data dari panggilan layanan pelanggan yang direkam sebelumnya. Model prediktif (dalam hal ini jaringan saraf dalam), telah dikembangkan dari sampel pengembangan yang berisi beberapa ratus ribu percakapan telepon sebelumnya dengan pelanggan. Ini telah diklasifikasikan ke dalam jenis penyelidikan (peristiwa) berikut:

1. Menanyakan saldo di rekeningnya.
2. Menginginkan daftar transaksi terkini.
3. Melaporkan kartu kredit yang hilang atau dicuri.
4. Meminta untuk menutup akunnya.
5. Sesuatu yang lain.

Peristiwa 1 hingga 4 adalah jenis pertanyaan pelanggan yang paling umum, mencakup sekitar 60% panggilan pelanggan, dan saat ini ditangani oleh penasihat manusia. Mengingat bahwa empat peristiwa pertama memerlukan respons standar yang relatif sederhana, maka proses layanan pelanggan untuk jenis pertanyaan ini dapat diotomatisasi sepenuhnya.

Jika kueri berkaitan dengan hal lain (peristiwa 5) maka sistem akan merujuk panggilan tersebut ke manusia untuk ditangani. Demikian pula, jika di akhir respons otomatis terhadap peristiwa 1-4 pelanggan menunjukkan bahwa mereka tidak puas dengan respons yang diterima, maka mereka masih dapat berbicara dengan orang sungguhan jika mereka mau.

Saat pelanggan menelepon, ucapan mereka terlebih dahulu diproses menjadi sekumpulan item data yang memberikan masukan ke proses pembelajaran mesin. Item data ini sebagian besar berupa tanda atau penghitung yang mewakili kata, frasa, dan kalimat tertentu yang terkait dengan jenis pertanyaan pelanggan tertentu. Model prediktif yang dihasilkan dari proses pembelajaran mesin kemudian menghasilkan probabilitas mengenai kemungkinan bahwa pertanyaan pelanggan termasuk dalam masing-masing dari lima kategori tersebut. Setelah Penasihat Robot mengidentifikasi apa yang paling mungkin ditanyakan, jawaban yang sesuai dapat dirumuskan. Jika Robotic Advisor menentukan bahwa pelanggan kemungkinan besar menanyakan saldo mereka saat ini, maka informasi yang relevan diambil dari sistem manajemen akun perusahaan dan respons otomatis diberikan. "Saldo terutang Anda adalah Rp. xxx.xxx.xxx".

Saat menerapkan Robotic Advisor, apa yang ingin diketahui oleh manajer pusat panggilan adalah hal-hal seperti:

1. Jika kita mempertahankan tingkat layanan pelanggan yang sama, berapa banyak staf yang saya perlukan dan berapa banyak uang yang bisa saya hemat?
2. Jika kita mempekerjakan staf dalam jumlah yang sama seperti sekarang, seberapa besar peningkatan layanan? Staf harus memiliki lebih banyak waktu untuk pelanggan, mengingat Robot Advisor akan menangani sebagian besar panggilan yang diterima.
3. Seberapa sering pelanggan tidak mendapatkan jawaban yang benar atas pertanyaannya jika pertanyaan tersebut ditangani oleh Robot Advisor? Ini didasarkan pada model prediktif. Oleh karena itu, terkadang ada yang salah.

Akan ada juga trade-off antara hasil-hasil ini. Anda mungkin dapat menggunakan Penasihat untuk mengurangi jumlah staf, meningkatkan tingkat layanan, atau keduanya. Jika Anda ingin memastikan bahwa sistem hanya menjawab pertanyaan jika hampir pasti memberikan respons yang benar, maka Anda akan mendapatkan lebih banyak rujukan ke operator manusia dibandingkan jika Anda bersedia menerima tingkat kesalahan yang tinggi. Jika ambang batas ditetapkan hanya untuk merespons pelanggan jika probabilitas sistem memberikan jawaban yang benar, katakanlah, di atas 99%, Anda akan mendapatkan jauh lebih banyak rujukan ke penjamin emisi manusia dibandingkan jika Anda menetapkan ambang batas sebesar 80%. Masalahnya kemudian adalah salah satu layanan pelanggan. Jika Anda menetapkan ambang batas sebesar 80%, maka 20% pelanggan akan merasa tidak puas karena sistem akan memberi mereka jawaban yang salah dalam 1 dari setiap 5 kasus, dibandingkan dengan 1 dalam 100 kasus.

Untuk dapat menjawab pertanyaan seperti ini dan merumuskan aturan keputusan yang tepat (Tahap 9), data scientist perlu mengumpulkan informasi tentang sifat operasional organisasi. Hal ini agar mereka mampu mengukur manfaat bisnis dari pekerjaan mereka. Jika tidak, bagaimana mereka dapat membenarkan nilai dari apa yang telah mereka lakukan?

Jika tujuannya adalah menghemat biaya atau meningkatkan keuntungan (yang hampir selalu terjadi), maka dalam contoh ini, data scientist perlu mencari tahu berapa lama berbagai jenis panggilan biasanya berlangsung dan berapa biaya yang harus dikeluarkan perusahaan untuk setiap panggilan. Hal ini, dikombinasikan dengan angka-angka tentang proporsi panggilan yang akan diotomatisasi dan biaya pengembangan sistem, akan memungkinkan dilakukannya analisis biaya-manfaat dari proyek Robot Advisor.

Sejauh ini dalam bab ini kita telah fokus pada proses penerapan pembelajaran mesin. Namun berapa lama waktu yang dibutuhkan data scientist untuk memberikan model prediktif kepada Anda? Itu pertanyaan yang sangat berat. Banyak pemasok solusi akan memberi tahu Anda bahwa mereka dapat mengembangkan model untuk Anda hanya dalam beberapa menit. “Yang perlu Anda lakukan hanyalah mengunggah data Anda ke solusi AI berbasis cloud kami dan kami akan memberikan jawabannya dalam hitungan detik!” Hal ini memang benar, tetapi Anda perlu mempertimbangkan konteks penerapan pembelajaran mesin. Untuk mendapatkan hasil dalam hitungan menit, Anda perlu mengumpulkan semua data dengan baik di satu tempat. Anda tidak perlu terlalu memikirkan bagaimana prediksi model dihasilkan, bagaimana penyampaiannya, atau dalam format apa. Jika Anda memiliki beberapa database kompleks yang perlu disatukan, memiliki banyak data yang berantakan, dan model yang dihasilkan harus sesuai dengan berbagai undang-undang dan persyaratan bisnis, lain ceritanya. Diperlukan waktu berminggu-minggu, berbulan-bulan, atau bahkan bertahun-tahun sebelum solusi yang tepat dapat diperoleh.

BAB 11

MENGOPERASIKAN PEMBELAJARAN MESIN

Beberapa organisasi telah menggunakan pembelajaran mesin selama beberapa dekade. Hal ini terutama berlaku di industri jasa keuangan dan di departemen pemasaran organisasi besar yang berhubungan dengan konsumen seperti jaringan supermarket dan department store. Bagi tim Silicon Valley, yang mencakup perusahaan-perusahaan seperti Amazon, Google, dan Facebook, pembelajaran mesin adalah inti dari semua yang mereka lakukan. Organisasi-organisasi ini memiliki infrastruktur, sistem, dan proses TI yang mapan untuk menerapkan pembelajaran mesin dan menerapkan hasilnya. Mereka melakukannya hari demi hari – seperti biasa.

Bagi organisasi yang baru mengenal kecerdasan buatan dan pembelajaran mesin, membangun model prediktif untuk pertama kalinya dan kemudian mengoperasionalkannya sebagai bagian dari sistem pengambilan keputusan otomatis dapat menjadi tugas yang sulit. Terkadang, diperlukan waktu beberapa bulan sebelum data yang diperlukan siap dan proses pembelajaran mesin dapat dimulai. Namun, mengumpulkan data yang tepat dan kemudian menerapkan algoritme pembelajaran mesin yang relevan, seperti yang dibahas di bab sebelumnya, sering kali merupakan bagian termudah dari proyek pembelajaran mesin. Hal ini karena, pada akhirnya, model prediktif tidak lebih dari sekumpulan persamaan yang ditangkap dalam spreadsheet, dokumen Word, atau perangkat lunak lainnya. Setelah proses pembelajaran mesin selesai, hasilnya perlu dioperasionalkan agar dapat digunakan.

Tantangan utama, khususnya bagi organisasi yang belum pernah menggunakan pembelajaran mesin, adalah menerima penggunaan pengambilan keputusan otomatis. Mereka perlu:

1. Menyediakan infrastruktur yang memungkinkan model prediktif menjadi bagian penting dari kemampuan pengambilan keputusan operasional organisasi.
2. Merasa nyaman dengan keputusan yang dibuat berdasarkan prediksi model. Hal ini termasuk menetapkan prosedur tata kelola untuk memastikan hal tersebut keputusan berdasarkan model ditindaklanjuti sebagaimana dimaksud, mematuhi undang-undang dan bahwa keputusan ini tidak diabaikan atau dikesampingkan oleh karyawan kecuali dalam situasi tertentu yang telah disepakati sebelumnya.

Mungkin kesalahan terbesar yang dilakukan sebuah organisasi adalah berasumsi bahwa pembelajaran mesin yang sukses adalah: “Semua tentang model” padahal mereka seharusnya memikirkan berbagai hal dari perspektif: “Ini semua tentang bisnis.”

Pembelajaran mesin mungkin merupakan sesuatu yang dapat menambah nilai pada apa yang dilakukan suatu organisasi, namun mungkin juga tidak. Analogi yang berguna adalah membangun mesin mobil. Para insinyur dapat menghabiskan banyak waktu untuk membangun mesin yang sangat bertenaga dan efisien. Namun mesin tersebut tidak akan memberikan manfaat apapun kecuali seseorang telah memikirkan tentang desain bodi mobil, dudukan mesin dan lain sebagainya, di mana mesin tersebut akan ditempatkan. Tanpa sisa

mobil, mesin tidak ada gunanya. Hal yang sama berlaku untuk pembelajaran mesin. Kecuali ada proses bisnis untuk memasukkan model tersebut, model tersebut tidak akan banyak berguna.

Jika suatu organisasi ingin berhasil mengintegrasikan pembelajaran mesin ke dalam proses bisnisnya, ada lima bidang inti yang perlu dipertimbangkan, dengan urutan sebagai berikut:

1. **Masalah.** Untuk apa pembelajaran mesin akan digunakan? Perlu ada masalah yang jelas yang dapat dibantu oleh pembelajaran mesin.
2. **Budaya.** Bisakah bisnis dibujuk untuk menerima pengambilan keputusan otomatis tanpa melibatkan manusia yang ahli? Dari perspektif manajerial, adalah sumber daya; yaitu sumber daya manusia dan uang, tersedia untuk memungkinkan model prediktif diintegrasikan ke dalam proses bisnis yang relevan?
3. **Implementasi.** Sistem atau proses apa yang akan digunakan untuk menerapkan model tersebut? Bagaimana skor dihitung? Bagaimana keputusan yang dibuat berdasarkan skor tersebut akan ditindaklanjuti oleh fungsi bisnis yang relevan?
4. **Data Pembangunan.** Apakah tersedia cukup data untuk memungkinkan pembuatan model prediktif?
5. **Kemampuan analitis.** Apakah organisasi memiliki perangkat lunak dan keahlian yang diperlukan untuk menganalisis data, menerapkan pembelajaran mesin, dan membangun model prediktif berkualitas baik dan dapat digunakan?

Pada pandangan pertama, urutan daftar ini mungkin tampak berlawanan dengan intuisi. Mengapa, misalnya, implementasi model dicantumkan sebelum analisis? Tentunya seseorang membangun model terlebih dahulu dan kemudian berpikir untuk mengimplementasikannya? Pembelajaran mesin bermanfaat untuk banyak hal, namun solusi berbasis pembelajaran mesin otomatis tidak selalu sesuai dengan kebutuhan Anda. Demikian pula, hanya karena secara teknis memungkinkan untuk memprediksi sesuatu, bukan berarti Anda harus melakukannya atau prediksi tersebut terbukti berharga. Oleh karena itu, kita perlu memulai dengan spesifikasi masalah dan rencana bagaimana mengatasinya. Perlu dipikirkan bagaimana Anda akan menggunakan pembelajaran mesin dalam suatu organisasi sebelum Anda menerapkannya dan seterusnya.

Jika pembelajaran mesin ingin menjadi jawaban atas permasalahan tersebut, maka dua hal penting yang perlu ditetapkan sebelum melakukan hal lain adalah:

1. Peristiwa/nilai/hal apa yang tidak pasti yang ingin Anda prediksi? Ini harus menjadi sesuatu yang sangat jelas dan eksplisit. Nilai ini harus mudah diukur dan dapat direpresentasikan sebagai indikator sederhana untuk masalah klasifikasi atau sebagai nilai numerik untuk masalah regresi.
2. Keputusan dan tindakan apa yang akan diambil mengenai cara memperlakukan orang, berdasarkan prediksi yang dibuat oleh model?

Tidak ada gunanya mengembangkan model prediktif untuk memprediksi peristiwa yang pasti akan terjadi. Demikian pula, jika orang akan terus diperlakukan dengan cara yang sama, berapapun skor yang mereka terima, lalu apa gunanya model tersebut? Jika pelanggan lama

akan selalu ditawarkan tingkat diskon yang sama untuk pembelian berikutnya, maka tidak ada gunanya membangun model prediktif untuk mengidentifikasi pembeli berulang. Model pembelian tetap hanya akan berguna jika strategi insentif yang berbeda ditawarkan, berdasarkan skor yang dihasilkan oleh model tersebut. Misalnya saja, tidak menawarkan diskon kepada orang-orang yang diprediksi oleh model akan dibeli lagi (skor tinggi) dan menawarkan diskon besar kepada orang-orang yang berada di ujung skala yang lain (skor rendah), untuk mendorong mereka berbelanja lagi di toko tersebut.

Mengidentifikasi masalah yang tepat sangatlah penting. Salah satu keterbatasan pembelajaran mesin adalah sifatnya yang sangat spesifik. Misalnya dalam pemungutan pajak, apakah kita tertarik untuk mengidentifikasi pengembalian pajak yang salah (masalah klasifikasi – apakah jumlah yang dibayarkan sudah benar atau belum) atau memaksimalkan total penerimaan pajak (masalah regresi, jumlah yang dikumpulkan)? Banyak orang membuat kesalahan kecil dalam pengembalian pajaknya. Mengetahui hal ini memang membantu, tetapi tidak akan banyak meningkatkan jumlah pajak yang dipungut secara keseluruhan. Faktanya, banyak orang yang secara tidak sengaja membayar lebih dalam jumlah kecil. Jika model Anda mengidentifikasi kasus-kasus ini maka Anda mungkin harus membayar kembali sebagian pajak yang telah dibayarkan daripada memungut lebih banyak!

Otoritas pajak mempunyai sumber daya yang terbatas. Jika Anda ingin memaksimalkan pengambilan pajak maka Anda harus fokus hanya pada kasus-kasus terbesar, dimana kekurangan pembayaran sangat signifikan. Otoritas pajak tidak mempunyai tenaga untuk menindaklanjuti setiap kekurangan pembayaran. Oleh karena itu, permasalahannya perlu menangkap kebutuhan ini. Misalnya, mengabaikan perbedaan pajak kurang dari Rp. 1.000.000. Demikian pula, model yang dibuat untuk memprediksi penyakit jantung mungkin tidak terlalu baik dalam memprediksi demensia dan sebaliknya.

Pertimbangan kedua ketika membangun model prediktif adalah budaya organisasi, dan keinginan untuk melakukan perubahan dalam area bisnis yang akan terkena dampak model tersebut. Praktik kerja yang ada dan penolakan terhadap cara kerja baru sering kali menjadi hambatan dalam penerapan pembelajaran mesin. Hal ini terutama berlaku ketika staf diharapkan berperilaku berbeda berdasarkan apa yang diperintahkan model kepada mereka; yaitu pembelajaran mesin akan mengubah cara orang melakukan pekerjaannya, dan dalam beberapa kasus, hal ini akan mengakibatkan hilangnya pekerjaan dan/atau devaluasi keterampilan mereka.

Biasanya, para ahli di suatu bidang cenderung tidak mempercayai model prediktif sebaik mereka dalam memprediksi hasil. Artinya, seorang pakar sering kali mengesampingkan keputusan yang dibuat berdasarkan model prediktif jika mereka diberi kesempatan untuk melakukannya - yang pada dasarnya meniadakan manfaat apa pun yang dapat dihasilkan oleh model tersebut.

Apakah penggantian akan menjadi masalah atau tidak sering kali bergantung pada apakah model prediktif diterapkan secara aktif atau pasif:

- **Model Aktif.** Prediksi (skor) yang dihasilkan oleh suatu model ditindaklanjuti secara otomatis. Tidak diperlukan keterlibatan manusia untuk melakukan tindakan.

- **Model Pasif.** Skor model digunakan untuk menghasilkan kasus-kasus yang diteruskan ke masyarakat, yang kemudian memutuskan bagaimana menanganinya; yaitu tidak ada yang terjadi sampai seseorang melakukan sesuatu dengan kasus yang dihasilkan oleh proses analisis.

Situasi dimana model aktif digunakan adalah dalam penagihan utang. Model prediktif digunakan untuk memprioritaskan debitur berdasarkan debitur mana yang paling mungkin melunasi utangnya. Daftar yang diprioritaskan dimuat ke dalam dialer yang kemudian secara otomatis memanggil orang-orang dalam urutan daftar.

Ketika kontak dilakukan dengan debitur, panggilan tersebut dialihkan ke manusia penagih utang yang mencoba membujuk debitur untuk membayar kembali utangnya. Penagih utang sendiri tidak peduli dengan urutan pengajuan perkara kepada mereka. Mereka hanya memproses setiap kasus yang mereka terima. Akibatnya, setelah model diimplementasikan, tidak diperlukan masukan manusia lebih lanjut untuk memutuskan urutan daftar panggilan yang disimpan di dialer.

Sebaliknya, contoh model pasif adalah model yang dirancang untuk mengurutkan pelamar kerja berdasarkan kesesuaian mereka untuk peran tertentu. Model ini mungkin sangat baik dalam memilih prospek yang paling menjanjikan, namun wawancara manusia tetap diperlukan untuk bertemu dengan kandidat dan membuat keputusan akhir tentang tawaran pekerjaan. Jika wawancara tidak menyukai seorang kandidat karena alasan apa pun, maka mereka tidak akan menawarkan pekerjaan tersebut kepada kandidat tersebut, terlepas dari seberapa baik mereka sebenarnya dalam melakukannya.

Model pasif adalah yang paling bermasalah. Hal ini karena di banyak organisasi, mereka yang berada di garis depan mempunyai kemampuan untuk mengubah keputusan yang dibuat secara terpusat. Dalam contoh rekrutmen, kekuatan model semakin berkurang ketika wawancara memutuskan untuk mengesampingkan keputusan berbasis model dengan keputusan mereka sendiri. Selain itu, penilai manusia sering kali menunjukkan bias (disadari atau tidak) yang merugikan kandidat yang baik berdasarkan prasangka atau stereotip yang tidak berdasar, sehingga banyak kandidat yang baik ditolak lamaran kerjanya.

Contoh lain yang saya temui adalah dalam pemungutan pajak. Sebuah fungsi pusat bertanggung jawab untuk mengidentifikasi potensi penipuan pajak, dan kasus-kasus ini kemudian diserahkan kepada pemeriksa pajak lokal untuk melakukan penyelidikan rinci. Para pemeriksa pajak mempunyai wewenang untuk menolak kasus-kasus yang mereka rasa tidak sesuai atau yang kemungkinan besar tidak akan menghasilkan banyak pajak tambahan.

Ketika fungsi pusat menerapkan model prediktif untuk mengidentifikasi kasus-kasus yang kemungkinan besar merupakan penipuan, jumlah kasus yang ditolak oleh pemeriksa pajak meningkat secara dramatis. Hal ini bukan karena kasus yang dihasilkan oleh model tersebut bukanlah kasus penipuan. Hal ini karena model tersebut terlalu baik dalam mengidentifikasi aktivitas penipuan karena tanda-tanda yang sangat halus pada data.

Ketika sebuah kasus tiba di meja pemeriksa pajak, pemeriksa tersebut tidak dapat menentukan dengan tepat apa yang mencurigakan mengenai kasus tersebut. Mereka tidak tahu bagaimana cara memulai penyelidikan dan akibatnya mereka menolaknya. Demikian

pula, model tersebut sering kali memilih kasus-kasus yang sangat kompleks atau di mana akan sangat sulit untuk memungut pajak yang belum dibayar meskipun penipuan jelas-jelas telah terjadi. Kasus-kasus seperti ini juga cenderung ditolak.

Model aktif lebih mudah diimplementasikan dibandingkan model pasif. Hal ini karena masyarakat tidak memiliki kesempatan untuk mengesampingkan keputusan berdasarkan model. Misalnya, setelah keputusan dibuat untuk mengirimkan voucher diskon kepada pelanggan, proses pengiriman voucher, dengan catatan penutup, terjadi secara otomatis. Tidak ada seorang pun yang menebak-nebak sistem dan memberikan voucher kepada pelanggan yang mereka yakini tidak menginginkannya atau tidak akan menggunakannya. Inilah salah satu alasan mengapa pembelajaran mesin sukses dalam hal-hal seperti pemasaran sasaran, pemberian kredit, dan penetapan harga asuransi, namun kurang dimanfaatkan dengan baik di berbagai bidang seperti layanan kesehatan, sumber daya manusia, kepolisian, dan pendidikan.

Hal ketiga dalam daftar periksa, setelah definisi masalah dan budaya, adalah implementasi. Seperti yang telah dibahas sebelumnya, pembelajaran mesin adalah tentang membuat keputusan yang lebih cepat, lebih baik, dan lebih murah. Model prediktif dapat diterapkan secara manual dalam spreadsheet dan skor bahkan dapat dihitung menggunakan pena dan kertas, namun dalam banyak kasus, model prediktif tertanam dalam sistem pengambilan keputusan otomatis. Sistem ini menggunakan model untuk membuat prediksi tentang bagaimana orang akan berperilaku, dan kemudian mengambil keputusan tentang cara memperlakukan orang berdasarkan prediksi tersebut; misalnya keputusan untuk memberikan pinjaman jika skornya cukup tinggi.

Dengan mengotomatiskan proses pengambilan keputusan, strategi yang dirancang secara individual untuk menangani jutaan orang dapat diterapkan hanya dalam beberapa detik. Artinya dalam praktiknya adalah bahwa organisasi dapat memperbarui prediksi mereka tentang orang-orang dengan frekuensi yang sangat sering: bulanan, harian, atau bahkan secara real-time, setiap kali ada informasi baru tentang seseorang yang diperoleh.

Jauh lebih banyak model prediktif yang dikembangkan daripada yang pernah diterapkan secara operasional. Saya tidak dapat menghitung lagi jumlah organisasi yang pernah bekerja sama dengan saya, yang setiap orangnya sangat tertarik dengan proses pembuatan model, namun kemudian lupa memikirkan bagaimana mereka akan menggunakan model tersebut setelah model tersebut dibuat. Kecuali jika infrastruktur untuk menggunakan model prediktif sudah dibangun dengan baik, penerapan model prediktif seringkali memerlukan proses yang lebih lama dan memakan waktu dibandingkan membangun model terlebih dahulu. Paket perangkat lunak paling populer untuk pembelajaran mesin (R dan Python) tidak akan secara otomatis menerapkan model ke dalam sistem operasional untuk Anda. Implementasi model perlu dipertimbangkan, anggaran disetujui dan perencanaan dimulai jauh sebelum data scientist mulai mengumpulkan data dan membangun model[55]

Secara umum, ada dua cara organisasi menerapkan model. Pendekatan pertama adalah memperlakukan implementasi sebagai proyek TI. Seorang pemrogram dipekerjakan untuk “mengkodekan” model dalam sistem TI suatu organisasi sehingga skor dan aturan

pengambilan keputusan yang diperlukan dapat dihitung kapan dan di mana hal tersebut diperlukan. Ini adalah pendekatan yang umum, namun sungguh menakjubkan betapa seringnya model diberi kode yang salah karena kesalahan sederhana – seperti memberi seseorang 100 poin, bukannya 10, atau ditambah 15 poin, bukannya minus 15 poin dalam kartu skor. Ada juga masalah prioritas yang harus dihadapi. Bukan hal yang aneh jika implementasi model memakan waktu berbulan-bulan atau bahkan bertahun-tahun, karena ada tugas lain yang menjadi prioritas. Mengikuti implementasi model, mencoba membuat perubahan pada model atau aturan keputusan terkait juga merupakan sebuah mimpi buruk. Hal ini karena setiap perubahan harus kembali ke antrian pekerjaan IT untuk diprioritaskan.

Pendekatan kedua terhadap implementasi model adalah melalui perangkat lunak khusus yang menyediakan antarmuka antara perangkat lunak pembelajaran mesin yang digunakan untuk membangun model dan lingkungan operasional tempat model akan diterapkan. Ketika model sudah selesai, ilmuwan data cukup menekan tombol “unggah” di perangkat lunak, yang kemudian menerapkan model (dan aturan keputusan terkait) ke lingkungan operasional.

Manfaat utama dari pendekatan ini adalah, secara teori, model dapat diimplementasikan segera setelah model tersebut dibuat. Yang penting, ilmuwan data dan pengguna bisnislah yang mengontrol penerapan model dan aturan batas (keputusan) terkait. Sedikit atau tidak ada masukan yang diperlukan dari departemen TI. Hal ini menghindari penundaan panjang yang mungkin timbul jika implementasi model ditempatkan dalam antrian pekerjaan TI.

Pertimbangan kedua dari belakang sebelum proses analisis dapat dimulai adalah data. Data adalah unsur utama dalam proses analitik. Ini penting. Tanpa data, dan jenis data yang tepat, model yang berguna tidak dapat dibangun atau diterapkan. Oleh karena itu, sebelum memikirkan menjalankan algoritme pembelajaran mesin untuk membangun model prediktif, perlu mempertimbangkan data yang akan digunakan. Secara khusus:

- *Data sampel pengembangan.* Sebagaimana dibahas dalam Bab 5, diperlukan sampel data yang cukup besar. Sampel ini perlu mencakup data observasi historis dan data hasil yang lebih baru.
- *Ketersediaan operasional.* Data apa pun yang ditampilkan dalam model yang dihasilkan harus tersedia saat model tersebut akan diimplementasikan.
- *Stabilitas operasional.* Jika hubungan dalam data berubah, model yang didasarkan pada data tersebut akan kehilangan sebagian kemampuan prediktifnya. Oleh karena itu, diperlukan model baru.

Penting untuk diingat bahwa sampel pengembangan yang digunakan untuk membangun model prediktif bersifat historis, namun model prediktif yang dihasilkan akan diterapkan di masa depan. Dalam beberapa bulan atau tahun sejak data sampel pengembangan dicatat, mungkin terdapat perubahan signifikan terhadap cara organisasi mengumpulkan dan menyimpan data. Jika item data tertentu tidak tersedia di masa mendatang, maka item data tersebut tidak boleh digunakan untuk membangun model. Ini karena model tersebut tidak dapat dimasukkan ke dalam produksi. Data yang dibutuhkan oleh model tidak akan tersedia.

Bisa jadi dahulu laba bersih diperoleh dari pelanggan dan hal inilah yang tertuang dalam data sampel pengembangan. Namun, karena pendapatan bersih terkadang sulit diingat oleh pelanggan, departemen pemasaran memutuskan bahwa akan lebih ramah pelanggan jika beralih ke menangkap pendapatan kotor. Jika model prediktif dibangun menggunakan laba bersih, model tersebut tidak dapat diterapkan dengan benar karena laba bersih tidak lagi tersedia seiring dengan perubahan cara pengumpulan pendapatan.

Setelah masalah telah ditentukan, dan masalah budaya, implementasi, dan data telah diatasi, maka kita dapat mulai memikirkan proses analitis dalam membangun model. Di sinilah keahlian khusus seorang ilmuwan data berperan penting. Terkait aspek teknis dalam membangun model prediktif, data scientist perlu mempertimbangkan berbagai masalah, seperti yang telah dibahas di bab sebelumnya. Menghasilkan model yang paling prediktif jelas merupakan salah satu kekhawatiran, namun kompleksitas, penjelasan dan transparansi dari solusi yang dihasilkan juga perlu dipertimbangkan dengan mempertimbangkan persyaratan organisasi dan peraturan.

Jika kita mengambil contoh perbankan, model prediktif yang digunakan untuk menghasilkan perkiraan modal yang harus disimpan oleh bank sebagai cadangan akan sangat diawasi oleh regulator keuangan. Pembuat model harus menjelaskan secara lengkap proses analitis yang diterapkan untuk membangun model, menjelaskan dengan tepat bagaimana model menghasilkan prediksinya, dan kemudian menyoroti kelemahan material apa pun yang mungkin dimiliki model tersebut. Solusi “kotak hitam” yang hanya memberikan prediksi yang sangat akurat namun tidak dapat dijelaskan tidak akan bisa diterima. Demikian pula, di bidang-bidang seperti pengumpulan pajak, diagnosis medis, dan investigasi kriminal, para ahli sering kali memilih model yang dapat mereka pahami.

Namun, jika kita mengalihkan penekanan ke model yang digunakan dalam pemasaran target, yang digunakan untuk memilih jodoh di situs kencan, atau menjadi bagian dari sistem terjemahan bahasa, maka minat pengguna dalam memahami logika model yang mendasarinya akan jauh berkurang. Keakuratan prediksi adalah pertimbangan utama. Jika berhasil, maka berhasil, dan itulah yang terpenting.

Sebelum memilih jenis model yang akan dikembangkan dan algoritma apa yang akan diterapkan, seorang data scientist harus berkonsultasi dengan pihak bisnis. Mereka seharusnya mengajukan pertanyaan yang tepat untuk memastikan bahwa mereka memberikan model yang prediktif, sesuai dengan persyaratan peraturan dan mempertimbangkan masalah atau kendala lain yang disoroti oleh dunia usaha.

Setelah proses analisis selesai dan model prediktif tersedia, maka model tersebut dapat ditransfer ke area penerapannya. Dengan asumsi penerapan model dipertimbangkan pada titik yang tepat dalam proyek, maka proses ini akan relatif lancar. Hal ini karena perangkat lunak dan TI yang diperlukan seharusnya sudah diterapkan, dan staf operasional diberi pengarahan tentang perubahan yang akan terjadi ketika model tersebut sudah online.

Menerapkan proses pengambilan keputusan otomatis menggunakan model yang dikembangkan melalui pembelajaran mesin merupakan pencapaian yang luar biasa, terutama jika hal tersebut merupakan sesuatu yang baru bagi organisasi yang bersangkutan. Namun, itu

bukanlah akhir dari prosesnya. Setelah model diterapkan, proses pengambilan keputusan perlu ditinjau secara berkala untuk menilai seberapa baik model tersebut memprediksi dan untuk memastikan bahwa batasan dan aturan pengambilan keputusan yang tepat terus diterapkan.

Pemantauan diperlukan karena akurasi prediksi suatu model cenderung menurun seiring berjalannya waktu akibat perubahan hubungan mendasar antara data yang digunakan untuk membuat model dan hasil yang diprediksi. Pemantauan biasanya melibatkan pembuatan serangkaian laporan yang menunjukkan seberapa akurat prediksi yang dibuat oleh model terbukti. Ketika akurasi model mulai menurun, hal itu menunjukkan bahwa sudah waktunya untuk mengembangkan model baru.

Penurunan akurasi prediksi seiring bertambahnya usia model adalah alasan utama mengapa model perlu dikembangkan kembali, namun model prediktif juga dikembangkan kembali karena alasan lain. Terkadang ada perubahan hukum atau peraturan yang berarti item data tertentu dalam suatu model tidak dapat digunakan lagi. Demikian pula dengan tersedianya sumber data baru, yang menjanjikan bahwa model yang lebih akurat dengan kekuatan prediksi yang lebih besar dapat dibangun. Alasan lain untuk melakukan pemantauan model secara berkala adalah untuk menilai bagaimana keputusan berdasarkan model mempengaruhi berbagai kelompok dalam populasi. Jika departemen pemasaran sebuah bank ingin menargetkan orang-orang dengan profil geografis tertentu dengan jenis kartu kredit baru, maka penting bagi bank untuk memahami bagaimana model penilaian kreditnya memperlakukan orang-orang tersebut.

Hal terakhir yang ingin dilakukan bank adalah melakukan kampanye pemasaran besar-besaran untuk nasabah baru dan kemudian melihat semua pemohon kartu baru ditolak. Demikian pula, jika seorang model terlibat dalam pengambilan keputusan penting yang mengubah hidup seseorang, maka sebaiknya pantau model tersebut untuk memastikan bahwa bias apa pun yang ditampilkannya dapat diterima. Jika departemen SDM menggunakan model untuk menyaring pelamar kerja dan ternyata departemen tersebut menolak lebih dari rata-rata proporsi orang dari etnis minoritas atau perempuan, maka hal tersebut perlu diselidiki dan diambil tindakan perbaikan jika diperlukan.

BAB 12

HUBUNGAN ANTARA BIG DATA DAN PEMBELAJARAN MESIN

Sejak sekitar tahun 2010, “Big Data” telah menjadi istilah umum yang digunakan untuk menggambarkan semua data yang dihasilkan oleh orang-orang dari ponsel cerdas mereka, riwayat penjelajahan web, media sosial dan perilaku pembelian, serta informasi lain yang dimiliki organisasi tentang mereka.

Mengapa Big Data berbeda dengan jenis data lainnya? Di satu sisi tidak ada perbedaan; itu semua hanyalah angka nol dan satu pada akhirnya. Namun, istilah “Big Data” cenderung diterapkan pada kumpulan besar berbagai jenis data yang sering kali mudah berubah dan berubah, dan sulit untuk dianalisis menggunakan perangkat keras dan perangkat lunak komputer tradisional.

Big Data juga sering kali menggabungkan jenis data tertentu yang belum banyak digunakan untuk analisis pelanggan hingga saat ini. Secara khusus, Big Data meliputi:

- **Teks.** Apa yang orang tulis dan katakan dapat dianalisis untuk mengidentifikasi apa yang mereka bicarakan dan sentimen yang diungkapkan. Jika suatu produk dibicarakan dalam konteks positif atau negatif, hal ini mungkin dapat memprediksi apakah seseorang akan terus membeli produk tersebut.
- **Gambar-gambar.** Ini mencakup foto dan video, serta pencitraan medis. Salah satu penerapan pembelajaran mesin adalah dengan menggunakan fitur yang diidentifikasi dalam pemindaian dan sinar X untuk memprediksi kemungkinan seseorang mengidap penyakit tertentu.
- **(Sosial) data jaringan.** Ini adalah informasi tentang koneksi orang-orang dan siapa yang mereka kenal. Data jaringan mencakup jumlah dan jenis koneksi yang dimiliki orang-orang, serta data tentang individu yang terhubung. Jika semua teman Anda adalah pecinta fiksi ilmiah, itu mungkin merupakan indikasi bagus bahwa Anda juga termasuk salah satunya.
- **Geospasial.** Informasi tentang lokasi dan pergerakan orang, disediakan oleh ponsel pintar, sistem GPS mobil, dan perangkat seluler lainnya.
- **Biometrik.** Data tentang tekanan darah, detak jantung, dan sebagainya, dikumpulkan dari fit band, jam tangan pintar, dan sebagainya.
- **Produk (mesin) yang dihasilkan.** Perangkat sehari-hari mulai dari televisi hingga pembuat kopi dirancang untuk berbagi informasi satu sama lain dan melalui internet. Saat ini pemanas, ketel, mesin cuci, dan sebagainya, semuanya dapat dikontrol melalui ponsel cerdas Anda. Konsep “*Internet of Things*” (IoT) masih berkembang namun pada akhirnya dapat memberikan banyak data yang dapat digunakan untuk menyimpulkan perilaku masyarakat menggunakan pembelajaran mesin.

Di “masa lalu yang indah” di tahun 1990an, perangkat pintar belum ada. Saat itu hanya sedikit orang yang memiliki ponsel dan internet masih dalam tahap awal. Sangat sedikit data elektronik tentang manusia atau aktivitas mereka. Yang ada biasanya terbatas pada beberapa

geodemografi seperti alamat, usia, pendapatan, jenis kelamin dan sebagainya. Hal ini kemudian dapat dilengkapi dengan data yang diberikan dari perusahaan pemasaran langsung atau lembaga referensi kredit jika produk jasa keuangan terlibat (misalnya, status tunggakan pinjaman dan kartu kredit). Supermarket tidak tahu apa yang dibeli pelanggan setiap minggunya, perusahaan asuransi tidak tahu cara orang mengemudi, dan layanan kesehatan menyimpan sebagian besar catatan pasien mereka dalam bentuk kertas.

Kehidupan seorang data scientist pada saat itu cukup mudah karena semua data elektronik (yang sangat terbatas) ini biasanya disimpan dalam format baris dan kolom yang rapi dan rapi (seperti yang dapat ditemukan di spreadsheet). Datanya juga relatif statis. Biasanya hanya diperbarui sangat jarang dan secara berkala – biasanya pada akhir bulan atau tahun.

Di dunia sekarang ini, data diperbarui lebih sering, sering kali secara real time. Selain itu, lebih banyak lagi data tidak terstruktur “berbentuk bebas” seperti ucapan, email, tweet, blog, gambar, dan sebagainya.

Faktor lainnya adalah sebagian besar data ini sering kali dihasilkan secara independen dari organisasi yang ingin menggunakannya. Hal ini bermasalah karena berbagai alasan. Salah satunya adalah jika data diambil atau dihasilkan oleh organisasi itu sendiri, maka mereka dapat mengontrol bagaimana data tersebut diformat. Mereka juga dapat melakukan pemeriksaan dan pengendalian untuk memastikan bahwa data akurat dan lengkap. Namun, jika data dihasilkan oleh sumber eksternal maka tidak ada jaminan bahwa data tersebut benar.

Data yang bersumber dari luar sering kali “Berantakan”. Dibutuhkan banyak pekerjaan untuk merapkannya dan menjadikannya dalam format yang bisa digunakan. Selain itu, mungkin ada kekhawatiran mengenai stabilitas dan ketersediaan data tersebut, yang dapat menimbulkan risiko bisnis jika data tersebut menjadi bagian dari kemampuan analisis dan pengambilan keputusan inti suatu organisasi.

Artinya adalah arsitektur komputer tradisional (perangkat keras dan perangkat lunak) yang digunakan organisasi untuk hal-hal seperti memproses transaksi penjualan, memelihara catatan akun pelanggan dan pembuatan faktur, tidak cocok untuk menyimpan dan menganalisis semua jenis data baru dan berbeda yang ada sekarang. Akibatnya, selama beberapa tahun terakhir sejumlah solusi perangkat keras dan perangkat lunak baru dan menarik telah dikembangkan untuk menangani jenis data baru ini.

Secara khusus, sistem komputer Big Data modern sangat ahli dalam:

- *Menyimpan data dalam jumlah besar.* Basis data tradisional memiliki keterbatasan dalam jumlah data yang dapat disimpan dengan biaya yang wajar. Cara-cara baru dalam menyimpan data telah memungkinkan perluasan kapasitas penyimpanan murah yang hampir tak terbatas.
- *Pembersihan dan pemformatan data.* Data yang beragam dan berantakan perlu diubah ke dalam format standar sebelum dapat digunakan untuk pembelajaran mesin, pelaporan manajemen, atau tugas terkait data lainnya.

- *Memproses data dengan sangat cepat.* Big Data bukan hanya tentang ketersediaan lebih banyak data. Ini perlu diproses dan dianalisis dengan cepat agar dapat dimanfaatkan secara maksimal.

Masalah dengan sistem komputer tradisional bukanlah bahwa secara teoritis terdapat hambatan bagi mereka untuk melakukan pemrosesan yang diperlukan untuk memanfaatkan Big Data, namun dalam praktiknya mereka terlalu lambat, terlalu rumit dan terlalu mahal untuk melakukannya.

Paradigma pemrosesan dan penyimpanan data yang baru, seperti Hadoop/MapReduce, telah memungkinkan tugas-tugas yang membutuhkan waktu berminggu-minggu atau berbulan-bulan untuk diproses hanya dalam beberapa jam dan dengan biaya yang lebih murah dibandingkan pendekatan pemrosesan data yang lebih tradisional. Cara Hadoop melakukannya adalah dengan menyimpan kumpulan data besar-besaran di jaringan besar (cluster) komputer desktop murah, seperti yang mungkin Anda miliki di kantor atau di rumah.

MapReduce adalah salah satu contoh bahasa pemrograman yang berbagi tugas pemrosesan data melalui jaringan komputer, seperti cluster Hadoop. Setiap komputer di cluster ditugaskan untuk secara mandiri memproses data yang dimilikinya. Hasil dari masing-masing komputer kemudian disatukan kembali dan digabungkan untuk menghasilkan jawaban akhir.

Bayangkan sebuah perusahaan yang menggunakan Hadoop untuk menyimpan rincian miliaran transaksi pelanggannya di jaringan yang terdiri dari beberapa ratus komputer. Untuk meringkas informasi tentang transaksi, seperti menjumlahkan nilai keuangan dari semua transaksi yang terjadi pada tahun keuangan sebelumnya, MapReduce akan menugaskan setiap komputer untuk menjumlahkan rincian transaksi yang dimilikinya. Setiap komputer kemudian memasukkan hasilnya kembali ke satu komputer pengontrol, yang menjumlahkan hasil individual untuk mendapatkan angka akhir.

Secara teori, puluhan ribu PC dapat dihubungkan bersama untuk menghasilkan kemampuan komputasi besar yang sebanding dengan superkomputer terbesar yang pernah ada. Jika Anda membutuhkan lebih banyak kekuatan pemrosesan, yang perlu Anda lakukan hanyalah membeli beberapa PC desktop standar lagi dan menghubungkannya ke jaringan.

Hadoop tersedia sebagai seperangkat alat pemrograman gratis (sumber terbuka) dan digunakan oleh banyak perusahaan sebagai bagian dari solusi Big Data mereka. Banyak pemasok solusi juga telah menggunakan perangkat lunak inti Hadoop dan mengembangkan rangkaian antarmuka dan alat pendukung mereka sendiri untuk menyediakan penawaran produk yang disesuaikan. Contohnya termasuk Cloudera, Hitachi Vantara, IBM, Microsoft, SAP dan Teradata, dan masih banyak lagi.

Data (baik “Besar” atau “Kecil”) tidak memiliki nilai intrinsik. Sebuah kesalahan besar yang dilakukan suatu organisasi adalah berpikir bahwa jika mereka berinvestasi pada sistem penyimpanan massal seperti Hadoop dan mengumpulkan semua data tentang manusia, maka hal tersebut akan menambah banyak nilai bagi bisnis mereka. Namun, semua data itu harus diolah menjadi sesuatu yang berguna jika ingin bermanfaat. Pembelajaran mesin adalah alat

utama yang melakukan hal tersebut. Data perlu diproses untuk mengidentifikasi fitur dan hubungan signifikan dalam data, dan untuk menghasilkan model prediktif yang memberi tahu Anda sesuatu tentang perilaku masyarakat berdasarkan apa yang terjadi sebelumnya di masa lalu.

Cara yang baik untuk memahami hubungan antara Big Data dan pembelajaran mesin adalah bahwa data adalah bahan mentah yang mendukung proses pembelajaran mesin. Manfaat nyata bagi bisnis diperoleh dari model prediktif, pengelompokan pelanggan, atau keluaran analitis lainnya yang diberikan pada akhir proses, bukan data itu sendiri.

Oleh karena itu, kecerdasan buatan, pembelajaran mesin, dan Big Data sering kali dibicarakan secara bersamaan, tetapi ini bukanlah hubungan yang simetris. Anda memerlukan pembelajaran mesin untuk mendapatkan hasil terbaik dari Big Data, namun Anda tidak memerlukan Big Data untuk dapat menggunakan pembelajaran mesin secara efektif. Jika Anda hanya memiliki sedikit informasi tentang beberapa ratus orang, maka itu sudah cukup untuk mulai menerapkan pembelajaran mesin untuk melakukan hal-hal seperti membuat model dan membuat prediksi, menerapkan pengelompokan pada pelanggan, dan melakukan jenis analisis data lainnya.

Semakin banyak dan lebih baik data yang Anda miliki, semakin baik pula solusi pembelajaran mesin Anda, namun memiliki data sebesar gigabyte atau terabyte bukanlah prasyarat untuk pembelajaran mesin praktis.

BAB 13

ETIKA, HUKUM, DAN GDPR

Penggunaan pembelajaran mesin dan model prediktif menimbulkan beberapa pertanyaan etika yang menarik. Hal ini terutama berlaku ketika mereka digunakan untuk menciptakan sistem pengambilan keputusan otomatis yang memutuskan bagaimana masyarakat akan diperlakukan tanpa keterlibatan manusia dalam proses pengambilan keputusan.

Ketika sistem pengambilan keputusan otomatis menjadi semakin lazim di hampir setiap aspek kehidupan kita, maka apa yang kita katakan adalah bahwa kita tidak lagi memegang kendali. Mereka yang pernah memegang kekuasaan, entah itu manajer bank yang memutuskan siapa yang akan diberi pinjaman atau dokter yang memutuskan siapa yang harus dirawat, telah menyerahkan kendali kepada komputer.

Apakah ini sebuah masalah? Jika seseorang dapat mengatakan secara pasti bahwa hasilnya akan selalu lebih baik bagi individu, maka hal tersebut dapat diterima oleh banyak orang. Namun, penting untuk menyadari bahwa dalam banyak, dan mungkin sebagian besar, situasi, organisasi menggunakan pembelajaran mesin untuk keuntungan mereka sendiri. Mereka berusaha keras untuk mencapai tujuan mereka sendiri (yaitu memaksimalkan keuntungan) bukan tujuan Anda. Apakah hasil yang dihasilkan bermanfaat atau tidak bagi orang-orang yang menjadi subjek keputusan tersebut, bukanlah perhatian utama mereka.

Implikasinya adalah sebagai masyarakat kita harus merasa nyaman dengan cara model prediktif dikembangkan dan diterapkan dan hal ini sejalan dengan pemahaman kita tentang apa yang benar dan pantas. Pemeriksaan dan keseimbangan yang tepat perlu diterapkan untuk mencegah penyalahgunaan sistem pengambilan keputusan yang mengandalkan pembelajaran mesin. Salah satu masalahnya adalah apa yang dimaksud dengan “penggunaan etis” tidak selalu jelas. Ada berbagai pandangan dan opini mengenai apa yang etis atau tidak dalam hal data dan analisis. Oleh karena itu, pemeriksaan dan pengendalian apa yang diperlukan masih bisa diperdebatkan.

Salah satu argumennya adalah jika model prediktif menghasilkan prediksi yang lebih akurat dibandingkan manusia ahli dan penerapan model tersebut menghasilkan pengambilan keputusan yang lebih cepat dan efisien, maka penggunaan model tersebut jelas merupakan hal yang tepat untuk dilakukan. Namun, penting untuk disadari bahwa tujuan tidak selalu menghalalkan cara. Cara seseorang mengambil keputusan dan apa yang dianggap dapat diterima oleh masyarakat luas juga penting.

Mari kita mulai dengan memikirkan data yang digunakan untuk membuat prediksi tentang perilaku masyarakat. Menurut Anda, apakah boleh menggunakan informasi tentang jenis kelamin, usia, agama, status perkawinan, orientasi seksual, atau ras seseorang ketika memutuskan bagaimana seseorang akan diperlakukan? yaitu haruskah jenis data ini diizinkan untuk ditampilkan dalam model prediktif?

Agar adil, ini adalah pertanyaan jebakan. Jawabannya akan dipengaruhi oleh jenis keputusan yang diambil. Jika kita berbicara tentang diagnosis atau pengobatan suatu kondisi medis, maka kebanyakan orang mungkin tidak akan terlalu peduli dengan jenis data apa pun yang digunakan jika hal tersebut memberikan hasil yang lebih baik bagi individu. Fakta bahwa usia dan jenis kelamin, misalnya, merupakan faktor prediktif dalam model penyakit jantung (Bab 2) tidak menjadi masalah. Namun, jika kita berbicara tentang memutuskan siapa yang akan dipekerjakan atau dipecat berdasarkan hal-hal seperti gender atau status perkawinan, atau membebankan biaya lebih banyak kepada orang-orang atas produk dan layanan karena agama mereka, maka hal tersebut menjadi kurang jelas.

Dari sudut pandang statistik semata, hal-hal seperti status perkawinan, jenis kelamin, dan usia mungkin memang berkorelasi dengan hal-hal seperti seberapa baik kinerja seseorang dalam pekerjaannya (Dan siapa yang harus dipekerjakan atau dipecat). Namun, dari sudut pandang masyarakat, mengizinkan model prediktif untuk mendorong keputusan berdasarkan jenis data ini patut dipertanyakan secara etis dan ilegal di banyak wilayah hukum.

Solusi yang sangat sederhana dan sederhana untuk masalah seperti ini adalah dengan memastikan bahwa item data tertentu dicegah untuk ditampilkan dalam beberapa jenis model prediktif untuk memastikan kepatuhan hukum. Algoritme yang menghasilkan model prediktif diarahkan untuk mengecualikan data seperti jenis kelamin, ras, dan status perkawinan dari proses pembelajaran mesin. Tidak masalah, namun hanya karena model prediktif tidak menggunakan item data tertentu secara langsung, bukan berarti model tersebut tidak menampilkan bias yang tidak dapat diterima.

Contoh yang bagus dari hal ini adalah diskriminasi gender dalam asuransi. Sudah menjadi fakta umum bahwa perempuan memiliki risiko yang lebih rendah dibandingkan laki-laki untuk berbagai jenis asuransi. Namun, di semua negara anggota UE, menetapkan premi asuransi berdasarkan gender adalah tindakan ilegal. Oleh karena itu, perusahaan asuransi tidak menggunakan gender dalam model prediksi klaim yang mereka gunakan untuk menetapkan premi. Namun, ada variabel lain yang sering menjadi proksi gender. Pendapatan adalah salah satu item data tersebut. Mengapa? Karena ada bias gender dalam distribusi pendapatan. Perempuan rata-rata mendapat penghasilan lebih rendah dibandingkan laki-laki meskipun mereka melakukan pekerjaan yang sama dan diskriminasi upah berdasarkan gender adalah tindakan ilegal. Jika model prediktif menggunakan pendapatan sebagai variabel prediktor maka terdapat pengaruh tidak langsung yang berarti perempuan akan diperlakukan berbeda dengan laki-laki. Sebagai konsekuensinya, model prediktif mungkin harus dirancang menjadi kurang optimal dalam hal kemampuan prediktif. Hal ini untuk memastikan bahwa, jika hal-hal lain dianggap sama, hal ini menghasilkan prediksi yang sama untuk laki-laki dan perempuan dengan karakteristik yang sama dan karenanya, mematuhi hukum.

Model prediktif harus sesuai dengan persyaratan hukum dan peraturan yang relevan – seperti tidak menampilkan bias gender. Namun, itu bukanlah akhir dari cerita. Kesalahan yang sangat mudah dilakukan adalah berpikir bahwa jika hal tersebut legal maka hal tersebut juga harus etis. Tentu saja, ada beberapa tumpang tindih antara apa yang etis dan apa yang legal, namun keduanya bukanlah hal yang sama. Hukum yang baik sering kali berupaya

mendefinisikan perilaku yang dianggap tidak dapat diterima oleh masyarakat; yaitu tidak etis, namun yang terbaik, undang-undang merupakan generalisasi dari perilaku etis dan biasanya bersifat retrospektif.

Hukum juga cenderung mengatasi permasalahan yang terjadi di masa lalu. Mereka biasanya tidak mempertimbangkan situasi baru sebelum hal itu muncul. Artinya, selalu ada situasi yang tidak tercakup dalam undang-undang tertentu dan hampir selalu ada celah yang memungkinkan pihak yang tidak bertanggung jawab untuk mengelak dari hukum demi mendapatkan apa yang mereka inginkan. Inilah yang dimaksud orang-orang ketika mereka berbicara tentang isi undang-undang dan bukan tentang semangat undang-undang.

Oke - jadi sebuah organisasi dapat bertindak secara legal namun tidak etis, namun mengapa sebuah bisnis yang peduli dengan memaksimalkan keuntungan harus khawatir dengan hal tersebut? Salah satu alasannya adalah murni kepentingan pribadi. Ada banyak bukti bahwa jika seseorang berpikir untuk jangka panjang, penerapan kode etik perilaku akan memberikan keuntungan nyata.

Alasan lain untuk memikirkan masalah etika dalam konteks pembelajaran mesin, dan data yang digunakan model prediktif, adalah risiko kerusakan reputasi. Jika masyarakat memutuskan bahwa mereka tidak menyukai cara Anda beroperasi, menganggap Anda memperlakukan orang lain dengan cara yang curang, atau melakukan diskriminasi yang tidak adil terhadap kelompok tertentu, maka sentimen tersebut dapat sangat merendahkan nilai suatu merek. Berargumen bahwa model prediktif Anda valid secara statistik dan mematuhi hukum saja tidaklah cukup. Jika diketahui bahwa suatu model menempatkan veteran perang dan anak-anak di urutan paling belakang untuk mendapatkan perawatan medis (walaupun menjadi veteran perang atau anak-anak bukanlah variabel eksplisit dalam model tersebut) maka Anda akan tertantang mengenai hal tersebut – bahkan jika proses pengambilan keputusan yang menggunakan model prediktif menghasilkan hasil pasien yang optimal yang diukur pada populasi secara keseluruhan.

Cara lain untuk memikirkan hal ini adalah bahwa pertimbangan etis harus mendorong sejumlah kendala dalam sistem pengambilan keputusan, yang perlu dipertimbangkan secara tepat ketika sistem tersebut dirancang.

Apa yang harus dilakukan organisasi untuk memasukkan perspektif etika ke dalam penggunaan pembelajaran mesin dan pengambilan keputusan otomatis? Sayangnya, tidak ada jawaban yang jelas untuk pertanyaan ini. Salah satu masalahnya adalah sifat subjektif dari etika. Ini adalah hal yang sangat pribadi. Dua orang mungkin mempunyai pendapat yang berlawanan, namun sama-sama sah mengenai apa yang merupakan perilaku yang dapat diterima. Demikian pula, rezim legislatif yang berbeda mempunyai pendekatan terhadap data pribadi dan bagaimana organisasi dapat menggunakannya dengan cara yang sangat berbeda. Di AS, titik awal dalam hal data pribadi dan penggunaannya dalam sistem pengambilan keputusan otomatis sangat bersifat utilitarian. Data pribadi ada untuk dikumpulkan dan digunakan untuk memaksimalkan tujuan organisasi misalnya memaksimalkan keuntungan atau meminimalkan biaya. Jika terdapat masalah dalam penggunaan jenis data tertentu atau

terdapat bias yang tidak dapat diterima terhadap kelompok tertentu, maka undang-undang akan dibuat untuk mengatasi permasalahan tersebut.

Hal ini sangat berbeda dengan pendekatan yang diterapkan oleh negara-negara Uni Eropa (UE). Warga negara UE memiliki kendali atas siapa yang dapat menyimpan data mereka dan bagaimana data tersebut digunakan. Jika mereka tidak ingin datanya digunakan untuk tujuan tertentu, maka itu adalah keputusan mereka, meskipun hal tersebut memberikan hasil yang kurang optimal.

Dua perspektif yang berbeda ini adalah salah satu alasan mengapa perusahaan-perusahaan yang berbasis di AS seperti Google dan Facebook kesulitan menemukan titik temu dengan regulator mengenai bagaimana data pribadi dapat dikumpulkan dan digunakan di negara-negara UE.

Persoalan lainnya adalah bahwa pertimbangan etis sering kali merupakan masalah dan domain tertentu. Apa yang dapat diterima dalam satu situasi belum tentu dapat diterima dalam situasi lain. Seperti yang telah kita bahas, kebanyakan orang mungkin setuju jika informasi pribadi sensitif seperti agama dan orientasi seksual mereka digunakan untuk mendiagnosis suatu kondisi medis. Namun, menggunakan item data yang sama untuk memutuskan siapa yang bisa atau tidak bisa mendapatkan hipotek adalah proposisi yang jauh lebih dipertanyakan.

Mengingat kompleksitas permasalahan ini, merupakan praktik yang baik bagi pembuat sistem pengambilan keputusan otomatis untuk melakukan penilaian risiko etis sebagai bagian dari tahap desain proyek. Selain itu, banyak organisasi kini membentuk komite etika untuk berkontribusi pada proses tata kelola. Komite etika terlibat dalam poin-poin penting dalam proyek pembelajaran mesin dan akan mengajukan pertanyaan sulit yang perlu ditanyakan tentang bagaimana sistem akan memengaruhi manusia. Komite etika kemudian memberikan pendapatnya mengenai apakah mereka percaya bahwa proses tersebut sejalan dengan posisi etika organisasi mereka, dan jika tidak, akan merinci modifikasi apa yang diperlukan untuk mencapai hal tersebut.

Dalam menilai risiko etika yang terkait dengan sistem pengambilan keputusan otomatis, ada tiga aspek utama yang perlu dipertimbangkan:

1. *Penerima Manfaat*. Siapa yang akan mendapatkan keuntungan dari keputusan yang diambil? Semakin besar keuntungan yang diperoleh pengambil keputusan dibandingkan dengan individunya, semakin besar risiko etika yang ditimbulkan oleh keputusan tersebut. Majikan yang memutuskan siapa yang akan dipekerjakan adalah mengambil keputusan semata-mata untuk keuntungan mereka sendiri. Pekerjaan tidak ditawarkan berdasarkan manfaat yang diberikannya kepada karyawan.
2. *Kekekalan Data*. Ada beberapa karakteristik yang dimiliki seseorang sejak lahir. Ini tidak bisa diubah. Jika keputusan diambil berdasarkan faktor-faktor seperti usia dan asal etnis, maka hal ini jauh lebih kontroversial dibandingkan data yang dihasilkan dari pilihan gaya hidup masyarakat, yang lebih dinamis dan mudah berubah. Misalnya musik yang disukai orang atau apa yang mereka tonton di TV.

3. *Dampak (signifikansi)*. Apa pengaruh suatu keputusan? Keputusan hidup atau mati dalam pengobatan kanker jauh lebih penting daripada sesuatu yang sepele seperti memutuskan apakah akan mengirimkan voucher diskon 10% untuk pizza beku kepada seseorang atau tidak.

Dengan menggabungkan ketiga pertimbangan ini, maka ketika keputusan berdampak besar dibuat menggunakan data yang tidak dapat diubah, semata-mata untuk kepentingan pengambil keputusan, risiko etika terbesar akan muncul. Kehati-hatian harus diberikan untuk memastikan bahwa sistem pengambilan keputusan berlangsung adil, tidak menunjukkan bias yang tidak dapat diterima, dan oleh karena itu, tidak membuat pengambil keputusan dituduh melakukan kesalahan. Saya tidak mengatakan bahwa sebuah organisasi tidak dapat menggunakan data yang berdampak besar dan tidak dapat diubah untuk mencapai tujuan mereka, namun mereka harus berhati-hati dan siap merespons tantangan apa pun terkait cara mereka menggunakan data tersebut.

Jika sebuah organisasi mengidentifikasi bahwa mereka menggunakan model prediktif untuk mengambil keputusan etis yang berpotensi “berisiko tinggi”, lalu apa yang harus dilakukan? Jawabannya adalah dengan melakukan tindakan mitigasi sebagai berikut:

1. Cobalah untuk mengidentifikasi kelompok yang berisiko. Caranya adalah dengan membuat laporan distribusi skor terpisah untuk kelompok seperti perempuan, etnis minoritas, anak-anak dan sebagainya. Kemudian dapat dilihat apakah terdapat bias pada skor yang dihasilkan; yaitu kelompok mana yang cenderung mendapat skor lebih rendah dari rata-rata dan oleh karena itu akan mendapat perlakuan yang lebih buruk dibandingkan dengan populasi yang lebih luas.
2. Bagi kelompok-kelompok yang nilainya kurang baik dibandingkan rata-rata populasi, batasan dan peraturan yang mengesampingkan harus digunakan untuk memastikan bahwa mereka diperlakukan secara adil. Demikian pula, batasan yang berbeda mungkin perlu ditetapkan untuk kelompok berisiko.
3. Terus memantau situasi setelah sistem pengambilan keputusan mulai berjalan. Tinjau profil skor dan aturan pengambilan keputusan untuk kelompok-kelompok utama secara berkala. Batasan, batasan, dan penggantian harus disesuaikan sesuai kebutuhan.

Untuk mengilustrasikan pendekatan ini, mari kita pikirkan skenario ketenagakerjaan. Pertimbangkan model prediktif yang digunakan untuk memutuskan siapa yang akan dipekerjakan. Skor yang tinggi dari model menunjukkan bahwa orang tersebut kemungkinan besar akan berhasil dalam peran tersebut, sedangkan skor yang rendah akan kurang berhasil. Kekhawatirannya mungkin adalah bahwa laki-laki lanjut usia atau orang yang memiliki anak kecil akan diperlakukan tidak adil (yang berpotensi melanggar hukum UE jika model tersebut digunakan di wilayah tersebut). Hal ini tidak berarti bahwa tidak ada bukti bahwa laki-laki lanjut usia atau orang yang memiliki anak memiliki kinerja yang lebih buruk dalam peran mereka dan akan menjadi karyawan yang lebih buruk dibandingkan orang lain, namun bahwa memperlakukan kelompok-kelompok ini secara berbeda tidak dapat diterima secara sosial. Sistem yang secara etis dipertanyakan adalah sistem yang hanya memiliki satu model prediktif dan satu skor batas yang diterapkan untuk memutuskan siapa yang akan dipekerjakan.

Pendekatan yang lebih kuat adalah memulai dengan membuat distribusi skor terpisah untuk pria lanjut usia dan orang yang memiliki anak. Ini kemudian dibandingkan dengan distribusi skor untuk semua calon karyawan. Jika perlu, potongan terpisah kemudian diterapkan. Hal ini untuk memastikan bahwa proporsi penduduk yang ditawarkan pekerjaan pada kedua kelompok tersebut sama dengan jumlah penduduk lainnya.

Pendekatan serupa lainnya adalah dengan membangun model terpisah untuk setiap kelompok. Satu model dibuat untuk masyarakat umum, satu model untuk orang lanjut usia, dan satu lagi untuk orang yang memiliki anak kecil. Aturan batas terpisah kemudian diterapkan pada setiap model untuk memastikan terjadinya strategi pengambilan keputusan yang konsisten dan selaras.

Sejauh ini dalam bab ini kita telah berfokus pada pengembangan kerangka etika dalam menangani data pribadi dan bagaimana hal tersebut dapat diterapkan dalam lingkungan umum; yaitu apa yang telah kita diskusikan dapat diterapkan secara universal pada berbagai situasi di mana pun seseorang berada. Memikirkan isu-isu ini akan membantu Anda menghadapi individu dengan cara yang sesuai secara moral (artinya Anda dapat tidur nyenyak di malam hari). Namun, seperti disebutkan sebelumnya dalam bab ini, pemerintah dan otoritas pengatur di seluruh dunia menerapkan kontrol hukum dan peraturan yang lebih ketat terhadap cara organisasi mengumpulkan, menyimpan, dan menggunakan data pribadi. Oleh karena itu, terlepas dari seberapa etis suatu organisasi meyakinkannya, organisasi tersebut juga harus mematuhi semua undang-undang dan peraturan yang relevan di wilayah tempat organisasi tersebut beroperasi.

Undang-undang seputar penggunaan data pribadi dan penggunaannya dalam pengambilan keputusan otomatis telah ada di banyak negara selama beberapa dekade, namun cakupan dan penerapannya sangat berbeda. Beberapa negara memiliki peraturan yang sangat rinci terkait data pribadi, sementara negara lain tidak memiliki peraturan sama sekali.

Di AS, peraturan hanya ditujukan pada penggunaan data pribadi tertentu. Equal Credit Opportunity Act 1974 (ECOA) mewajibkan pemberi pinjaman untuk menjelaskan mengapa sistem penilaian kredit otomatis mereka menolak kredit orang. Undang-Undang Portabilitas dan Akuntabilitas Asuransi Kesehatan tahun 1996 (HIPAA) memberikan perlindungan seputar penggunaan catatan kesehatan masyarakat. Undang-Undang Perlindungan Privasi Daring Anak (COPPA) berkaitan dengan data tentang anak di bawah usia 13 tahun dan seterusnya. Di AS, tidak ada undang-undang yang secara khusus mengatur pengumpulan data pribadi atau bagaimana data tersebut digunakan dalam proses pengambilan keputusan otomatis.

Bagi 28 negara yang tergabung dalam Uni Eropa, yang mencakup 4 dari 10 negara dengan perekonomian terbesar di dunia dengan populasi gabungan lebih dari setengah miliar, situasinya sangat berbeda. Pendekatan yang jauh lebih komprehensif terhadap pengaturan data pribadi dan cara penggunaannya telah diterapkan selama bertahun-tahun. Hal yang dapat diperdebatkan adalah undang-undang perlindungan data paling signifikan yang pernah diberlakukan di mana pun di dunia dan mulai berlaku di semua negara UE pada tanggal 25 Mei 2018. Peraturan Perlindungan Data Umum (GDPR) memberikan tanggung jawab besar pada

organisasi mengenai cara mereka mengumpulkan, mengelola, dan memproses informasi pribadi seseorang.

GDPR berlaku untuk semua organisasi yang beroperasi di UE, meskipun basis operasinya berada di negara lain. Tidak ada pilihan bagi raksasa teknologi yang beroperasi di Silicon Valley jika mereka melakukan bisnis di negara anggota UE mana pun.

Ada yang mengatakan bahwa undang-undang itu tidak masuk akal, tetapi GDPR akan menendang Anda jika Anda mengabaikannya. Organisasi yang gagal mematuhi GDPR akan dikenakan denda hingga 4% dari omset global (bukan keuntungan) untuk setiap pelanggaran peraturan yang terjadi. Bahkan jika Anda tidak tinggal di UE, ada baiknya meluangkan sedikit waktu untuk memahaminya karena hal ini menyoroti arah perjalanan undang-undang perlindungan data di sejumlah negara dan yurisdiksi lain.

GDPR mengadopsi pendekatan berbasis hak. Landasan yang mendasarinya adalah kepemilikan pribadi atas data diri sendiri. Merupakan hak individu untuk memutuskan siapa yang memiliki akses ke data mereka dan bagaimana data tersebut digunakan. Data tentang saya adalah milik saya, itu bagian dari siapa dan apa saya. Anda tidak mempunyai hak untuk menyimpan atau menggunakan data saya kecuali saya memberi Anda izin untuk melakukannya. Menyimpan data saya atau membuat keputusan tentang saya tanpa izin adalah tindakan ilegal. Saya juga punya hak untuk berubah pikiran. Saya dapat memberi Anda akses ke data saya hari ini, tetapi jika saya tidak ingin Anda menggunakannya lagi, Anda harus berhenti melakukannya dan menghapus apa yang Anda ketahui tentang saya.

Peraturan perundang-undangan juga didasarkan pada prinsip. Artinya, hal ini tidak sepenuhnya bersifat preskriptif dalam hal apa yang harus Anda lakukan untuk mematuhi. Tujuannya adalah untuk mendorong kepatuhan terhadap semangat hukum, bukan sekadar isi undang-undang.

Faktor lain yang perlu diingat adalah bahwa fokus GDPR adalah pada individu. Tidak ada pertimbangan mengenai biaya yang dikeluarkan organisasi untuk mematuhi. Paling-paling, hanya ada ruang terbatas bagi sebuah bisnis untuk berpendapat bahwa hal tersebut tidak ekonomis atau mungkin menyebabkan mereka bangkrut kecuali mereka memproses data dengan cara tertentu.

GDPR mencakup berbagai masalah yang berkaitan dengan data pribadi dan berlaku untuk semua pengambilan keputusan dan pembuatan profil otomatis yang diterapkan pada individu yang masih hidup. Empat bidang yang sangat relevan dalam konteks pembelajaran mesin/AI adalah:

1. Persetujuan eksplisit harus diperoleh dari seseorang sebelum data pribadinya dapat dikumpulkan atau digunakan. Persetujuan tersebut tidak dapat disimpulkan atau menjadi “persetujuan secara default”.
2. Tidak diperbolehkan mengambil keputusan mengenai seseorang, menggunakan proses pengambilan keputusan otomatis yang mempunyai dampak hukum atau dampak signifikan serupa terhadap mereka kecuali mereka telah memberikan persetujuan bahwa pengambilan keputusan otomatis dapat digunakan untuk tujuan ini. Ini merupakan tambahan dari persetujuan untuk menyimpan data. Pengecualian utama terhadap hal ini

adalah ketika pemrosesan otomatis diperlukan untuk memungkinkan kontrak hukum diberlakukan.

3. Seseorang mempunyai hak untuk diberikan penjelasan tentang bagaimana keputusan otomatis diambil dan konsekuensi dari keputusan tersebut. Hal ini sejalan dengan persyaratan yang lebih umum dalam GDPR bagi organisasi untuk memberikan informasi yang diperlukan kepada individu guna memastikan terjadinya pemrosesan data mereka secara adil dan transparan.
4. Sistem pengambilan keputusan otomatis tidak boleh menampilkan diskriminasi yang tidak adil; yaitu memperlakukan orang secara berbeda berdasarkan ras, orientasi seksual, dan sebagainya.

Untuk mendapatkan persetujuan yang jelas, persetujuan yang tepat harus ada. Jika kita berbicara tentang pengambilan keputusan mengenai karyawan, klausul yang sesuai perlu dimasukkan dalam kontrak kerja mereka. Bagi konsumen, persetujuan perlu diberikan melalui orang yang mencentang kotak yang relevan atau sejenisnya.

Untuk butir 2, kata kuncinya adalah penting. Jika suatu keputusan akan mempunyai dampak nyata pada seseorang, maka orang tersebut berhak meminta agar keputusan tersebut ditinjau secara manual. Undang-undang sudah jelas bahwa tinjauan ini harus sepenuhnya independen dan dilakukan oleh seseorang yang diberi wewenang untuk membuat keputusan berbeda dari keputusan yang dibuat oleh sistem. Hal ini tidak bisa hanya menjadi kasus staf administrasi yang memberikan stempel pada keputusan awal.

Hal yang membuat butir ini sulit diterapkan dalam praktik adalah tidak adanya definisi yang diterima mengenai “signifikan”. Mungkin yang paling penting, apa yang mungkin tidak berarti bagi sebagian orang, mungkin sangat berarti bagi orang lain, meskipun hal tersebut hanya sebagian kecil dari orang-orang yang diproses melalui sistem. Tidak bisa dianggap remeh bahwa jenis pengambilan keputusan tertentu tidak berdampak signifikan terhadap masyarakat – pengambilan keputusan harus dilakukan berdasarkan kasus per kasus. Bayangkan sebuah jaringan supermarket yang memutuskan akan menerapkan strategi penetapan harga diferensial. Pelanggan yang diyakini tidak bisa atau tidak mau berbelanja di tempat lain akan dikenakan biaya 5% lebih banyak untuk bahan makanan mereka dibandingkan pelanggan yang lebih berubah-ubah dan berbelanja di tempat lain. Bagi sebagian besar pelanggan, hal ini hanya merupakan ketidaknyamanan kecil, namun bagi mereka yang berada di garis kemiskinan, hal ini mungkin memberikan dampak yang sangat nyata terhadap kualitas hidup mereka. Artinya secara praktis adalah bahwa akan lebih bijaksana untuk selalu memulai dari posisi bahwa semua keputusan berpotensi menjadi keputusan yang signifikan kecuali terdapat cukup bukti yang menyatakan sebaliknya.

Berkenaan dengan item 3 – hak untuk mendapatkan penjelasan tentang bagaimana suatu keputusan diambil – hal ini memerlukan pemahaman tentang bagaimana model prediktif digunakan. Misalnya, mengapa premi asuransi seseorang berjumlah tertentu atau mengapa pinjamannya ditolak. Minimal, sebuah organisasi perlu memberi tahu seseorang yang mengajukan pertanyaan seperti itu:

- Item data mana yang digunakan dalam proses pengambilan keputusan. Misalnya: “saat menilai permohonan Anda untuk kontrak telepon baru, kami menggunakan informasi tentang usia, pendapatan, dan status perkawinan Anda, serta informasi tentang riwayat kredit Anda untuk mengambil keputusan.”
- Sumber item data tersebut. Hal ini bisa dilakukan langsung dari individu melalui aplikasi atau formulir online, informasi yang tersedia untuk umum seperti catatan pemilih atau informasi yang telah mereka berikan izin untuk dibagikan kepada organisasi lain, seperti platform media sosial atau lembaga referensi kredit.
- Rincian item data mana yang berkontribusi terhadap keputusan dan dengan cara apa. Dalam memutuskan diterima atau tidaknya permohonan kontrak telepon seseorang, dapat dikatakan bahwa keputusan didasarkan pada nilai kredit. Skor tersebut menggunakan informasi tentang riwayat kredit masyarakat, dan jumlah serta terkini pembayaran yang terlewat berdampak negatif pada skor tersebut. Pemohon melewatkan dua pembayaran hipotek tahun lalu yang mengakibatkan nilai kredit rendah. Oleh karena itu, permohonan mereka untuk kontrak baru telah ditolak.

Dalam beberapa kasus, ada kemungkinan bahwa penjelasan teknis lengkap juga diperlukan untuk semua parameter model individual, meskipun pada saat penulisan, kondisi pasti yang memerlukan hal ini masih belum jelas. Kita mungkin berasumsi bahwa hal ini hanya diperlukan dalam beberapa kasus tertentu yang sampai ke pengadilan, dan bukan sesuatu yang diberikan setiap kali pelanggan mengajukan pertanyaan. Namun, dalam GDPR cukup jelas bahwa: “Kompleksitas bukanlah alasan untuk gagal memberikan informasi kepada subjek data.” Di Inggris, Kantor Komisaris Informasi membuat komentar berikut sehubungan dengan penjelasan model:

“Oleh karena itu, organisasi big data perlu berhati-hati sebelum mengandalkan keputusan pembelajaran mesin yang tidak dapat dirasionalisasikan dalam istilah yang dapat dimengerti manusia. Jika sebuah perusahaan asuransi tidak dapat menemukan seluk-beluk rumit yang menyebabkan sistem permohonan online mereka menolak beberapa orang namun menerima orang lain (betapapun masuk akal nya alasan mendasar tersebut), bagaimana mereka dapat menjelaskan hal ini kepada individu yang terkena dampaknya?”

Artinya, terdapat potensi hambatan dalam penggunaan pembelajaran mendalam dan model “kotak hitam” kompleks lainnya untuk mengambil keputusan tentang manusia, kecuali jika terdapat mekanisme penjelasan yang sesuai. Mekanisme tersebut dapat dikembangkan, namun hal ini memerlukan waktu dan biaya tambahan untuk memproduksinya.

Faktor lain yang perlu diingat adalah bahwa individu dapat berubah pikiran. Mereka mungkin telah memberikan izin untuk menggunakan datanya untuk tujuan tertentu pada suatu waktu, namun izin tersebut dapat dicabut di kemudian hari. Demikian pula, “hak untuk dilupakan” mengharuskan organisasi untuk menghapus data pribadi jika seseorang

memintanya. Oleh karena itu, proses untuk menangani perubahan persetujuan dan penghapusan data pribadi perlu dilakukan.

Terakhir, masalah diskriminasi. Semua sistem pengambilan keputusan (otomatis atau manual) bersifat diskriminatif – itulah sifat yang mereka lakukan. Pertanyaannya adalah apakah diskriminasi yang tidak adil terjadi. GDPR dengan jelas menyatakan bahwa hal-hal tertentu seperti ras, orientasi seksual, dan sebagainya, biasanya tidak boleh dimasukkan ke dalam sistem pengambilan keputusan otomatis. Namun, kita juga harus mampu mengidentifikasi dan mengoreksi bias tidak langsung seperti yang telah kita bahas sebelumnya dalam Bab ini.

Gabungkan semua hal ini, dan dalam praktiknya hal ini berarti bahwa pembelajaran mesin dan analisis tingkat lanjut “Wild West” sudah hampir berakhir jika menyangkut data pribadi. Praktik membiarkan data scientist membuat model prediktif yang “cepat dan kotor” untuk digunakan oleh fungsi bisnis Anda sudah tidak ada lagi – setidaknya di negara-negara UE.

Sebaliknya, organisasi yang berinvestasi pada Big Data dan proses pengambilan keputusan otomatis perlu:

- Menilai risiko dan masalah yang terkait dengan pembelajaran mesin dan pengambilan keputusan otomatis sejak awal. Ini bukan tambahan di akhir proyek.
- Jangan berharap untuk menghilangkan seluruh keahlian manusia. Harus ada fungsi operasional terlatih yang dapat memproses keputusan secara manual bila diperlukan.
- Memiliki proses tata kelola dan audit yang sesuai untuk mengelola dan memberikan pengawasan terhadap sistem pengambilan keputusan mereka. Hal ini mencakup pengembangan pendekatan untuk mengidentifikasi dan memitigasi risiko yang muncul, seperti yang telah dibahas sebelumnya.

Dari perspektif manajemen proyek, persyaratan tata kelola dan intervensi manual dapat berdampak signifikan terhadap biaya pengembangan dan pemeliharaan sistem pengambilan keputusan otomatis. Jika Anda mempertimbangkan untuk menggunakan otomatisasi dalam organisasi Anda, maka biaya tambahan ini perlu dimasukkan dalam kasus biaya-manfaat yang dilakukan sebelum proyek dimulai.

Dalam hal penggunaan operasional, jika Anda menggunakan model prediktif untuk melakukan hal-hal seperti memutuskan siapa yang akan Anda pecat, maka Anda perlu memiliki orang yang dapat meninjau data yang digunakan oleh model tersebut dan dapat mengesampingkan keputusan awal jika diperlukan. Di bidang seperti asuransi dan pemberian kredit, ini berarti Anda harus mempertahankan penjamin emisi ahli yang dapat menantang keputusan yang dibuat oleh sistem penjaminan emisi otomatis organisasi. Perlu juga ada proses untuk menangani pertanyaan pelanggan tentang pengambilan keputusan otomatis dan untuk memberikan informasi yang relevan kepada mereka sebagaimana diwajibkan oleh undang-undang.

BAB 14

PEMBELAJARAN MESIN TERCANGGIH

Pembelajaran mesin adalah subjek yang menarik dan terus berkembang yang didorong oleh perkembangan baru di tiga bidang utama:

1. *Model dan Algoritma*. Pendekatan pembelajaran mesin baru terus dikembangkan. Salah satu jalur penelitian adalah melihat model jenis baru. Upaya lain adalah untuk memperbaiki algoritme yang sudah ada yang menghasilkan kartu skor, pohon keputusan, jaringan saraf, dan sebagainya, untuk meningkatkan keakuratan prediksi jenis model ini.
2. *Data*. Model prediktif dan aplikasi AI lainnya hanya akan berfungsi dengan baik jika data yang digunakan untuk membangunnya. Data yang lebih banyak/lebih baik menghasilkan solusi yang lebih baik. Inilah salah satu alasan mengapa “Big Data” dan pembelajaran mesin sangat erat kaitannya.
3. *Sistem dan perangkat lunak*. Perbaikan pada sistem yang digunakan untuk mengembangkan dan menerapkan solusi berbasis pembelajaran mesin otomatis adalah hal yang penting karena semakin cepat solusi diterapkan, semakin cepat pula manfaatnya akan terwujud.

Mari kita mulai dengan membahas sisi model dan algoritma pembelajaran mesin. Kartu skor, pohon keputusan, dan jaringan saraf mungkin merupakan jenis model prediktif yang paling banyak digunakan saat ini. Namun, jika Anda berbicara dengan generasi muda di Silicon Valley, mereka mungkin akan tertawa dan kemudian memberi tahu Anda bahwa model semacam ini agak “vintage” – pohon keputusan sudah ada sejak tahun 1980-an! Ada begitu banyak jenis model prediktif yang lebih baik di luar sana saat ini.

Di satu sisi hal ini memang benar. Kartu skor, pohon keputusan, dan jaringan saraf tentu bukan hal baru. Rata-rata, jenis model prediktif yang lebih baru seperti jaringan saraf dalam, mesin dan ansambel vektor pendukung, juga menghasilkan prediksi yang lebih akurat, terutama di area yang memerlukan jenis pengambilan keputusan yang kompleks.

Kartu skor dan pohon keputusan pertama kali digunakan secara komersial pada tahun 1950an dan 1960an, pada zaman ketika komputer pada umumnya berukuran sebesar meja besar dan memiliki kemampuan komputasi kurang dari 1% dibandingkan ponsel pintar saat ini. Oleh karena itu, jenis model prediksi sederhana ini dapat dikembangkan dan diimplementasikan dengan relatif mudah. Kekuatan komputer tidak lagi menjadi masalah saat ini, namun model sederhana seperti kartu skor dan pohon keputusan tetap sangat populer karena alasan berikut:

- Sifatnya adalah “Kotak Putih”. Sangat mudah untuk memahami bagaimana skor dan prediksi tentang seseorang diperoleh. Demikian pula, mudah untuk melihat item data mana yang memberikan kontribusi paling signifikan terhadap skor, dan mana yang kurang penting.

- Mereka mudah untuk dikodekan. Perangkat lunak khusus tidak diperlukan untuk mengimplementasikannya. Jika sumber daya terbatas, maka Anda dapat mengimplementasikan kartu skor atau pohon keputusan sebagai proyek TI kecil tanpa perlu membeli perangkat keras/perangkat lunak tambahan, dan tanpa perlu mempekerjakan ilmuwan data yang sangat mahal.
- Mereka masih menghasilkan prediksi yang cukup bagus, jika bukan yang terbaik. Beberapa model prediktif ribuan kali lebih kompleks daripada kartu skor atau pohon keputusan sederhana. Namun, bahkan model prediktif paling canggih sekalipun sering kali hanya memberikan peningkatan hanya beberapa persen dibandingkan kartu skor sederhana atau pohon keputusan, dan terkadang tidak sama sekali.

Yang ingin saya perjelas adalah bahwa kemenangan besar bagi organisasi adalah dengan melakukan lompatan dalam penggunaan pengambilan keputusan otomatis, berdasarkan pembelajaran mesin. Manfaat tambahan dari penggunaan metode tercanggih yang tersedia seringkali lebih kecil. Hal ini terutama berlaku jika permasalahan bisnisnya sederhana dan dapat diungkapkan secara tepat dan ringkas, seperti menilai risiko gagal bayar suatu pinjaman atau seberapa besar kemungkinan seseorang merespons kampanye pemasaran. Jenis masalah yang lebih kompleks dan berbelit-belit seperti pengenalan wajah, terjemahan bahasa, dan mengemudi otonomlah yang paling mendapat manfaat dari bentuk pembelajaran mesin tingkat lanjut seperti jaringan saraf dalam.

Jika organisasi Anda saat ini tidak menggunakan pembelajaran mesin, mengembangkan beberapa model prediktif sederhana yang dapat diintegrasikan ke dalam infrastruktur pengambilan keputusan yang ada akan memberi Anda banyak manfaat. Hal-hal canggih yang canggih, yang seringkali membutuhkan perangkat keras dan/atau perangkat lunak khusus, akan memberikan manfaat yang lebih besar, namun tidak dalam jumlah yang besar. Oleh karena itu, jangan tunda lagi. Aturan 80/20 berlaku. Anda akan mendapatkan 80% manfaat untuk 20% usaha.

Jika Anda tidak dapat membuat model prediktif sederhana berfungsi, maka menggunakan pendekatan yang lebih kompleks atau membeli beberapa perangkat keras/perangkat lunak yang mahal kemungkinan besar tidak akan menyelesaikan masalah mendasar; yaitu kegagalan proyek pembelajaran mesin hampir selalu disebabkan oleh formulasi masalah yang salah, data yang mendasarinya, atau masalah organisasi. Masalahnya kemungkinan besar bukan disebabkan oleh jenis model prediktif yang telah dikembangkan.

Itulah argumen untuk menjaga keyakinan dengan model sederhana seperti kartu skor dan pohon keputusan. Namun, jika sebuah organisasi sudah menjadi pengguna aplikasi AI, yang dikembangkan menggunakan pembelajaran mesin, dan model prediktifnya bertanggung jawab atas pengambilan keputusan bernilai miliaran dolar setiap tahunnya, maka akan ada dorongan untuk mendapatkan keputusan terbaik (paling akurat). model yang mungkin – dan dengan alasan yang bagus. Untuk model yang mendukung pengambilan keputusan senilai miliaran dolar setiap tahunnya, peningkatan kinerja sebesar 0,1% saja setara dengan keuntungan sebesar Rp. 1 Milyar. Dalam skenario seperti ini, sangat dibenarkan untuk

mempekerjakan tim data scientist secara purna waktu untuk terus menantang dan menyempurnakan model yang diterapkan organisasi.

Bentuk model prediktif paling canggih yang digunakan saat ini adalah model ansambel, berdasarkan jaringan saraf yang kompleks (dalam). Dengan ansambel, alih-alih hanya memiliki satu kartu skor, pohon keputusan, atau model jaringan saraf, ratusan atau mungkin ribuan model berbeda dapat dibangun. Setiap model dikembangkan menggunakan sampel data yang berbeda dan/atau algoritma yang berbeda untuk menentukan parameter model. Oleh karena itu, setiap model membuat prediksi dengan cara yang sedikit berbeda. Skor yang dihasilkan oleh masing-masing model sering kali sama atau sangat mirip, namun terkadang tidak sesuai satu sama lain; yaitu beberapa model akan memberikan kasus-kasus tertentu skor yang sangat tinggi, sedangkan model-model lain akan memberikan kasus-kasus yang sama skor yang jauh lebih rendah dan sebaliknya.

Menggunakan model ansambel mirip seperti mengambil keputusan yang dibuat oleh sebuah komite yang terdiri dari para ahli, bukan oleh seorang ahli saja. Ada dua alasan mengapa pendekatan komite lebih baik daripada hanya memiliki satu pakar:

1. Jika salah satu pakar mempunyai pengetahuan khusus, sedangkan pakar lainnya tidak, maka hal ini dapat diikutsertakan dalam proses pengambilan keputusan.
2. Beberapa ahli terkadang mengambil keputusan yang buruk. Para ahli lainnya akan menggunakan pengetahuan kolektif mereka untuk mengesampingkan (mengeluarkan suara) dari kasus-kasus tersebut.

Sama seperti komite, beberapa model yang membentuk ansambel akan sangat baik dalam memprediksi hasil dari kasus-kasus tertentu. Jika salah satu model lemah di area tertentu (menghasilkan prediksi yang buruk) maka model tersebut akan dikesampingkan oleh model lainnya.

Setelah dibangun, cara kerja ansambel cukup mudah. Skor dari masing-masing model digunakan untuk mengambil keputusan. Keputusan akhir kemudian dibuat melalui pemungutan suara mayoritas sederhana. Jika kita kembali ke model kartu skor penyakit jantung yang telah dibahas sebelumnya, bayangkan bahwa alih-alih hanya menggunakan satu kartu skor, seribu kartu skor berbeda akan dibuat. Aturan keputusan awal adalah mengundang seseorang untuk pemeriksaan jika skornya 521 atau lebih. Dengan ansambel, jika setidaknya 501 model individu menghasilkan skor 521 atau lebih, maka keputusannya adalah mengundang.

Seberapa baik ansambel daripada model tunggal? Terkadang tidak ada! Namun, menurut pengalaman saya, bukan hal yang aneh jika sebuah ansambel memiliki hasil sekitar 5-10% lebih baik dibandingkan model tunggal. Jika perusahaan asuransi menemukan bahwa penggunaan pohon keputusan menghasilkan pengurangan klaim sebesar Rp. 40 Miliar dibandingkan proses manual sebelumnya untuk jumlah penjaminan yang sama, maka peralihan ke pendekatan ansambel diharapkan dapat memberikan manfaat tambahan sebesar Rp. 2-4 Miliar.

Jika yang Anda minati hanyalah keakuratan prediksi mentah, maka ansambel adalah pilihan yang tepat. Namun, jika penting bagi Anda untuk dapat menjelaskan bagaimana suatu

model sampai pada prediksi tertentu, Anda mungkin perlu berpikir dua kali sebelum mengambil jalur ansambel karena solusinya akan jauh lebih kompleks dan lebih sulit dipahami daripada solusi tunggal. pendekatan model.

Sekarang mari kita beralih ke pemikiran tentang data. Dari membaca literatur akademis tentang pembelajaran mesin, saya berani menebak bahwa 95% atau lebih di antaranya adalah tentang algoritma; yaitu diskusi yang sangat teknis tentang pendekatan matematika mutakhir yang dapat meningkatkan akurasi prediksi dari kumpulan data tertentu. Namun dalam praktiknya, dalam hal meningkatkan akurasi model prediktif, data adalah rajanya.

Jika diberi pilihan antara algoritme baru untuk membangun model prediktif dan memiliki data yang lebih banyak/lebih baik, maka data selalu menang. Dengan kata lain, model prediktif yang sangat sederhana yang dibuat menggunakan sejumlah besar data berkualitas tinggi hampir selalu mengungguli pendekatan lebih maju yang dibuat menggunakan sejumlah kecil data berkualitas rendah. Jika Anda benar-benar ingin memaksimalkan model prediktif Anda, maka meningkatkan kualitas data yang digunakan untuk membangunnya dan mencari sumber data baru dan lebih baik, harus menjadi prioritas utama Anda.

Pada masa awal Big Data, ketika biaya penyimpanan data turun drastis dalam waktu singkat, terdapat filosofi data “simpan dan analisis semuanya” di antara para pionir. Pesannya adalah setiap organisasi harus mengumpulkan dan menganalisis semua data yang mereka bisa. Saat itu, ada banyak pembicaraan tentang perlunya berinvestasi pada sistem penyimpanan massal seperti Hadoop. Hal ini memungkinkan organisasi menyimpan semua data yang dapat mereka peroleh agar dapat menghasilkan model prediktif terbaik dan karenanya, memperoleh keunggulan kompetitif. Namun, jumlah data yang dihasilkan terus meningkat dari tahun ke tahun dan tidak menunjukkan tanda-tanda melambat. Faktanya, volume data meningkat jauh lebih cepat dibandingkan penurunan biaya penyimpanan data.

Artinya, manfaat memiliki semua data yang tersedia sampai batas tertentu diimbangi dengan biaya penyimpanan dan analisis semua data tersebut. Seperti yang telah dibahas sebelumnya, hanya sebagian kecil dari seluruh data di luar sana yang benar-benar ditampilkan dalam model prediktif dan digunakan untuk membuat prediksi; yaitu setelah Anda mengetahui jenis data apa yang dapat memprediksi bagaimana orang akan berperilaku, maka Anda dapat membuang sebagian besar data lainnya karena Anda tidak memerlukannya. Melanjutkan pemeliharaan database besar yang berisi data “bernilai rendah” bukanlah penggunaan waktu dan sumber daya yang efisien.

Saat ini terdapat peralihan menuju penyimpanan dan agregasi data bersama – terutama jika menyangkut data yang bersumber dari luar dan data yang umum di seluruh organisasi. Jika seseorang mempunyai sepuluh aplikasi di ponselnya yang disediakan oleh sepuluh organisasi berbeda, maka sangat sia-sia jika masing-masing organisasi tersebut mengumpulkan data lokasi dan pergerakannya sendiri. Jauh lebih masuk akal jika satu organisasi mengelola data, dan kemudian menyediakan item data spesifik yang relevan bagi klien.

Jika Anda melihat perusahaan seperti Facebook, Google, Experian, Equifax dan sebagainya, maka inilah yang mereka lakukan. Mereka melakukan kerja keras mengumpulkan, memformat, menyiapkan dan merangkum data. Mereka kemudian mengemas bagian-bagian yang berguna dan menjualnya. Dengan cara ini, masing-masing organisasi hanya memperoleh data yang benar-benar berguna bagi mereka. Akibatnya, mereka tidak perlu membuang waktu dan sumber daya untuk mengumpulkan data dalam jumlah besar yang tidak diperlukan.

Pendorong ketiga perkembangan pembelajaran mesin adalah sistem dan perangkat lunak TI. Seiring dengan meningkatnya volume data pribadi, dan frekuensi perubahan data, waktu siklus antar pengembangan model telah berkurang di banyak industri.

Paradigma tradisional untuk mengembangkan dan menerapkan model prediktif adalah dengan memisahkan kedua bagian proses ini; yaitu kembangkan model Anda terlebih dahulu dan kemudian terapkan. Selama fase pengembangan, seorang data scientist menghabiskan waktu sehari-hari, berminggu-minggu, atau bahkan berbulan-bulan untuk mengumpulkan data dan melakukan analisis statistik yang diperlukan untuk membangun model. Ketika bagian proses tersebut selesai, ada latihan lebih lanjut untuk mengkodekan model dalam lingkungan produksi, menguji apakah model berfungsi, dan kemudian model tersebut dapat digunakan secara operasional.

Di banyak (dan mungkin sebagian besar) industri, pendekatan terhadap model prediktif ini masih diterapkan dan secara umum berfungsi dengan baik, salah satunya karena setelah suatu model dikembangkan, model tersebut harus lulus audit internal dan eksternal, dan kemudian harus melalui tinjauan peraturan sebelum dapat diterapkan. dimanfaatkan. Terlepas dari apa yang dinyatakan dalam undang-undang, memiliki model struktur tata kelola yang kuat adalah hal yang penting. Hal ini karena jika seluruh bisnis Anda bergantung pada keputusan yang tepat, dan Anda mengambil keputusan yang salah, maka dampaknya terhadap keuntungan akan sangat besar.

Agar model risiko yang digunakan di perbankan dan asuransi dapat menentukan berapa banyak modal yang perlu mereka simpan sebagai cadangan, diperlukan waktu satu tahun atau lebih antara dimulainya proyek pemodelan prediktif dan model yang diterapkan dalam bisnis. Setiap aspek model harus didokumentasikan sepenuhnya dan kemudian siklus diskusi, umpan balik, dan analisis lebih lanjut perlu dilakukan sebelum regulator menyetujui model tersebut layak untuk digunakan. Regulator perbankan bahkan tidak akan mempertimbangkan untuk meninjau model prediktif sampai model tersebut telah menjalani tinjauan lengkap oleh para ahli independen. Ini bisa memerlukan waktu yang sama atau lebih lama dari proses pembuatan model awal.

Namun di bidang lain, seperti pemasaran internet, segalanya bergerak lebih cepat. Data, dan hubungan dalam data tersebut, sering berubah, sebagian di antaranya terjadi secara real-time. Jika suatu organisasi ingin mempertahankan keunggulan kompetitifnya, maka organisasi tersebut memerlukan siklus pengembangan dan implementasi model yang jauh lebih cepat. Model dibangun kembali setiap hari atau lebih sering sebagai respons terhadap perubahan data yang konstan. Hal ini mengarah pada pengembangan sistem TI yang

mengintegrasikan data yang dimiliki organisasi, alat analisis yang digunakan untuk membuat model prediktif, dan sistem yang menerapkannya.

Sistem “dalam database” ini mengalirkan data ke alat pembelajaran mesin tanpa perlu mengekstrak data terlebih dahulu, sehingga secara drastis mengurangi waktu yang diperlukan untuk mengumpulkan sampel data, membuat model prediktif, dan kemudian menerapkan model tersebut secara operasional.

Setelah sistem dalam database dikonfigurasi, model dapat dikembangkan kembali dan diterapkan secara otomatis. Secara teori, model baru dan terkini dapat dibangun setiap kali ada data baru yang tersedia – sistem belajar dari setiap kasus baru yang ditangani. Model-model baru dikembangkan dan diterapkan setiap menitnya. Akibatnya, seorang data scientist menjadi tidak mungkin terlibat dalam detail setiap model yang dibangun.

Sebaliknya, peran data scientist adalah menjadi bagian dari tim yang merancang sistem yang lebih luas. Secara khusus, mereka memiliki tanggung jawab untuk memahami data yang dimasukkan ke dalam sistem dan bagaimana data tersebut dipetakan ke masalah bisnis yang memerlukan pembuatan model prediktif oleh sistem.

Setelah sistem berjalan, perangkat lunak menyediakan dasbor untuk data scientist yang melaporkan status sistem secara keseluruhan. Misalnya, seberapa baik performa model dalam sistem, bagaimana performa model berubah seiring waktu, bagaimana data yang digunakan untuk proses analisis berubah, dan seterusnya.

Para ilmuwan data sendiri hanya terlibat dalam detail ketika terjadi kesalahan, atau ketika beberapa fitur baru perlu dimasukkan ke dalam sistem. Jika terjadi penurunan kinerja model yang tidak terduga, data scientist perlu menyelidiki dan mencari tahu alasannya. Mereka kemudian akan melakukan tindakan perbaikan untuk memperbaiki masalah dan mengembalikan sistem ke kondisi pengoperasian optimal.

Pendekatan serupa dilakukan pada perangkat pembelajaran mandiri dan robot otonom. Setiap kali mereka melakukan suatu tugas, mereka mengumpulkan data tentang peristiwa itu. Model prediktif yang mendorong aktivitas mereka secara otomatis disempurnakan menggunakan data tambahan yang tersedia. Dengan cara ini, semakin banyak data yang tersedia, keakuratan model yang mendasarinya akan meningkat.

Fokus lebih lanjut dari beberapa perangkat lunak pembelajaran mesin yang lebih baru adalah alat perantara yang berupaya menyediakan antarmuka yang lebih baik antara pengguna bisnis non-teknis dan data serta algoritme dasar yang diperlukan untuk pembelajaran mesin.

Alat yang paling canggih ini mencoba menggantikan beberapa tugas yang biasanya dilakukan oleh data scientist. Mereka dapat menganalisis dan menyiapkan data dari berbagai sumber, menerapkan serangkaian algoritme, dan menyajikan hasilnya kembali kepada pengguna non-teknis dengan cara yang mudah dipahami tanpa rumus atau persamaan apa pun. Secara khusus, perangkat lunak ini berupaya menyajikan hasil dalam cara kontekstual yang masuk akal secara bisnis, dibandingkan memberikan perspektif statistik yang lebih formal yang biasa digunakan oleh para ilmuwan data.

Contoh utama dari pendekatan ini adalah yang dilakukan oleh IBM dengan Perangkat Lunak Watson Analytics-nya. Versi asli Watson terkenal mengalahkan beberapa pemain manusia dalam acara kuis pengetahuan umum Jeopardy di AS. Watson kini telah berevolusi menjadi produk komersial. Di balik layar, perangkat lunak ini menggunakan beberapa algoritma pembelajaran mesin yang sangat kompleks untuk mengekstrak informasi dari berbagai sumber data yang berbeda. Bagian depan perangkat lunak ini dirancang dengan mempertimbangkan manajer dan pengguna bisnis lainnya, bukan ilmuwan data.

Hasil akhirnya adalah ketika disajikan dengan data yang sesuai, wawasan dan pemahaman baru tentang perilaku pelanggan dapat disajikan kepada pengguna bisnis dalam hitungan jam, atau bahkan menit, tanpa perlu melibatkan pakar teknis dalam prosesnya.

BAB 15

KAPAN SAYA BISA MEMBELI MOBIL SELFDRIVING?

Saya yakin tidak akan lama lagi Anda dapat mengunjungi dealer setempat dan membelinya, dan beberapa orang akan berpendapat bahwa Anda sudah bisa melakukannya. Namun, jika yang Anda maksud adalah kendaraan yang sepenuhnya otonom, yang dapat disimpan oleh nenek Anda yang tunanetra di garasinya dan dikendarai ke mana pun dia mau, maka itu mungkin akan memakan waktu lebih lama daripada yang disarankan oleh semua berita media tentang mobil self-driving.

Bagi sebagian orang, mobil self-driving di abad ke-21 sama seperti pendaratan di bulan di abad ke-20. Ini adalah tantangan teknis yang sangat besar, yang membutuhkan investasi miliaran dolar selama satu dekade atau lebih. Kesuksesan akan menandai tonggak penting dalam kemampuan kita untuk meniru dan melampaui kemampuan manusia melalui cara-cara buatan. Sama seperti pendaratan di bulan, ada sejumlah tantangan yang harus diatasi dan banyak langkah yang harus ditempuh. Mampu menembakkan roket ke luar angkasa atau menempatkan satelit ke orbit memang membantu, namun hal itu tidak membuat Anda memenuhi syarat untuk mengirim manusia ke bulan. Itu adalah tantangan yang jauh lebih besar.

Baru-baru ini, saya sadar bahwa pertanyaan yang telah lama mengganggu saya adalah: “Di mana semua mobil self-driving?” Sepertinya tidak ada seorang pun yang menyebutkannya di dealer yang saya kunjungi baru-baru ini ketika mencari mobil baru.

Saya ingat pernah membaca bertahun-tahun yang lalu, pada tahun 2010 atau lebih awal, bahwa mobil self-driving sudah cukup banyak ada di sini. Salah satu kutipan yang terlintas dalam pikiran adalah: “Sungguh gila membayangkan kita akan terus mengendarai mobil seperti yang kita lakukan sekarang – bahwa dalam 10 hingga 20 tahun kita masih harus duduk di belakang kemudi...”. Jika Anda membaca media populer dan siaran pers yang dikeluarkan oleh produsen motor dan raksasa teknologi, maka Anda akan dimaafkan jika berpikir bahwa mobil self-driving sudah siap digunakan. Hanya kaum luddite di pemerintahan yang menghalangi hal-hal tersebut.

Saat membahas alasan mengapa nenek saya yang tunanetra mungkin tidak dapat membeli mobil yang sepenuhnya otonom pada tahun 2019/20, dan mungkin tidak selama satu dekade atau lebih setelah itu, sebuah pertanyaan yang sangat bagus untuk diajukan. Pertanyaan pada diri sendiri adalah: “Di mana semua kereta tanpa pengemudi (dan kapal dalam hal ini)?” Tentunya naik dan turun rel adalah tugas yang jauh lebih sederhana daripada mengemudi di jalan terbuka atau lalu lintas kota? Sebenarnya, ada beberapa sistem kereta tanpa pengemudi yang beroperasi, namun hanya sebagian kecil dari kereta api di dunia yang sepenuhnya otomatis. Yang dimaksud adalah transportasi lokal dan/atau sistem bawah tanah yang beroperasi di lingkungan kecil yang terkendali dengan baik, bukan kereta api antarkota yang bepergian ke seluruh negeri.

Mengapa sebagian besar, jika tidak semua, kereta api tidak sepenuhnya otomatis? Tidak ada jawaban tunggal atau sederhana untuk hal ini. Sebagian karena teknologi, sebagian lagi berkaitan dengan ekonomi. Serikat pekerja dan sentimen masyarakat memainkan peranannya sebagaimana halnya pemerintah dan perundang-undangan. Bisa dibayangkan masalah teknologi adalah bagian yang mudah. Hal-hal lainlah yang membuat hal-hal ini menjadi sulit.

Ujian sebenarnya dari mobil yang benar-benar otonom adalah ketika siapa pun dapat pergi dan membelinya tanpa batasan siapa mereka atau ke mana mereka dapat pergi. Mampu berkendara secara handsfree di jalan bebas hambatan sejauh beberapa mil atau memiliki taksi tanpa pengemudi yang beroperasi di pusat-pusat beberapa kota besar, tidaklah lulus ujian. Sebenarnya, ketika berbicara tentang mobil yang dapat mengemudi sendiri, jawabannya adalah teknologinya tidak cukup baik. Sudah sampai di sana, tapi belum sampai. Jika kita kembali ke analogi pendaratan di bulan, maka teknologi self-driving mungkin setara dengan tahap perkembangan orbit rendah bumi. Kemajuan yang baik telah dicapai namun perjalanan masih panjang.

Jangan salah paham, saya tidak mengatakan bahwa Anda tidak akan dapat membeli mobil yang sepenuhnya dapat dikendarai sendiri pada suatu saat dan menurut saya ini akan menjadi hal yang bagus jika Anda bisa, tetapi itu akan memakan waktu lama. Hype-nya lebih besar dari kenyataannya. Apa yang kita lihat dalam praktiknya adalah mobil-mobil dengan jumlah fitur otonom yang terus meningkat, mulai dari penghentian otomatis, parkir mandiri, dan bantuan jalur, dengan fitur-fitur yang lebih kompleks muncul setiap saat. Namun, untuk mencapai otonomi penuh bagi siapa pun, kapan pun, di mana pun transportasi tanpa pengemudi, ada banyak masalah rumit yang masih ditangani, dan yang saya maksud bukan dilema “troli belanja atau anak” yang sering dikutip. Melakukan putar balik secara ilegal ketika lalu lintas kembali normal (kita semua pernah melakukannya!), berhenti di lalu lintas yang melaju, menyingkir dari layanan darurat, menyalip traktor yang bergerak lambat di jalan satu jalur dan mengetahui ketika Anda harus berkendara di trotoar/trotoar. Ini hanyalah beberapa hal yang harus dilakukan oleh pengemudi manusia sehari-hari di seluruh dunia. Mereka tidak melakukan hal ini tanpa alasan. Terkadang, Anda harus menaiki trotoar atau menyalip ketika secara teknis ilegal, jika tidak, Anda hanya akan terjebak dengan lusinan pengemudi yang marah yang membunyikan klakson di belakang Anda. Isyarat tangan dari sesama pengemudi memberi tahu Anda bahwa Anda aman untuk keluar, meskipun kelihatannya tidak seperti itu. Dan bagi mereka yang mengatakan bahwa semua hal ini akan terselesaikan ketika kita menyingkirkan pengemudi manusia dari jalan – itu hanyalah negeri fantasi unicorn ungu. Tantangan untuk membuat orang Amerika melepaskan hak mengemudi mereka akan membuat pencabutan amandemen kedua tampak seperti piknik.

Terakhir, siapa bilang mobil self-driving adalah mobil self-driving? Tentunya harus ada kerangka resmi independen dan proses pengujian yang harus dilalui kendaraan tersebut sebelum pabrikan dapat mengatakan bahwa mobilnya dapat mengemudi sendiri? Jika saya harus lulus tes mengemudi, maka mobil tanpa pengemudi juga harus lulus. Banyak negara yang berupaya untuk melakukan hal-hal ini, namun hal ini masih dalam tahap awal.

Baik – cukup keluhan tentang mobil self-driving. Mengapa ini relevan dengan buku seperti ini? Fokus awal bab ini adalah: “Menghindari hype dan mengidentifikasi peluang.” Saya ingin menulis tentang cara-cara menarik pembelajaran mesin dan kecerdasan buatan mengubah masyarakat secara luas. Hal ini penting karena perubahan dalam cara masyarakat beroperasi menimbulkan ancaman dan peluang. Jalan perdagangan baru terbuka, dan cara-cara lama dalam melakukan sesuatu hanya tinggal kenangan akan kerja keras yang berulang-ulang serta proses yang lambat, manual, dan tidak efisien. Tidak seorang pun ingin apa yang terjadi pada Kodak dan Blockbuster terjadi pada mereka. Beradaptasi atau mati.

Namun, Anda harus mewaspadai risiko, hambatan, dan penghambat serta peluangnya. Ada alasan mengapa mereka menyebutnya sebagai teknologi terdepan. Di setiap kisah sukses teknologi baru, terdapat lusinan kegagalan, namun kita selalu memberikan lebih banyak perhatian pada keberhasilan dan melupakan kegagalan. Pepatah umum mengatakan: “Orang bijak adalah orang yang belajar dari kesalahan orang lain.” Namun, ketika saya mencari buku tentang “Kegagalan Bisnis” di Amazon, jumlah hasil yang dihasilkan kurang dari 5% dari jumlah yang dihasilkan saat mencari “Keberhasilan Bisnis.”

Perbedaan jumlah perhatian yang kita berikan terhadap kesuksesan dan kegagalan sering kali menghasilkan ekspektasi yang tidak realistis tentang bagaimana teknologi baru akan berdampak pada kita. Mobil self-driving adalah contoh utama dari efek ini. Mereka menunjukkan optimisme berlebihan Silicon Valley dan over-hyping terhadap teknologi baru yang kerap terjadi. Realitas virtual dan teknologi yang dapat dikenakan adalah beberapa contoh lain yang kini lebih maju. Awalnya sangat populer, tidak mengambil alih dunia, tetapi mereka memiliki beberapa aplikasi dunia nyata yang sangat berguna.

Faktanya, sebagian besar teknologi baru melewati jalur ini. Teknologi sering kali terbukti berguna dalam jangka panjang, namun jarang mencapai hasil maksimal seperti yang disarankan oleh para penginjil teknologi ketika teknologi pertama kali dibayangkan. Jika Anda ingin mendapatkan gambaran umum tentang efek over-hyping ini dalam praktiknya, maka saya sangat merekomendasikan Gartner Hype Cycle. Prinsip ini juga ditangkap oleh Hukum Amara:

“Kita cenderung melebih-lebihkan dampak suatu teknologi dalam jangka pendek dan meremehkan dampaknya dalam jangka panjang.”

Pada tahun 2012, Gartner melaporkan teknologi kendaraan otonom berusia 5-10 tahun dari kematangan. Lima tahun kemudian pada tahun 2017, angka ini direvisi naik menjadi lebih dari 10 tahun.

Contoh lain yang lebih umum dari hype seputar teknologi AI adalah perkiraan jumlah pekerjaan yang akan hilang akibat otomatisasi yang didorong oleh AI selama beberapa dekade mendatang. Jangan salah, AI adalah teknologi yang disruptif. Banyak orang akan kehilangan pekerjaan atau harus beradaptasi dengan cara kerja yang berbeda karenanya, namun hal ini tidak akan menggantikan semua yang kita lakukan – setidaknya dalam 10-20 tahun ke depan. Studi yang paling berpengaruh pada masanya pada tahun 2013, menyebutkan jumlah pekerjaan yang berisiko tinggi hilang karena otomatisasi sebesar 47%. Studi tersebut

memperkirakan bahwa pada tahun 2030-2035 hampir separuh pekerjaan di AS yang saat ini dilakukan oleh manusia akan dilakukan oleh mesin. Perkiraan serupa juga berlaku untuk negara-negara maju lainnya seperti Inggris dan Jerman. Seiring berjalannya waktu, perkiraan ini telah direvisi turun. Pada tahun 2017, PwC menyarankan batas atas sebesar 38%. Sebuah studi komprehensif yang dilakukan oleh OECD menyebutkan angka tersebut hanya sebesar 10% pada tahun 2018, dan angka tersebut belum memperhitungkan semua lapangan kerja tambahan yang akan diciptakan untuk mendukung teknologi baru ini. Dampaknya masih cukup besar, namun jauh lebih kecil dari apa yang awalnya dibicarakan. Jika kita membagi hilangnya pekerjaan tersebut selama periode 20 tahun, maka hal tersebut setara dengan hanya 1 dari 200 pekerjaan yang hilang setiap tahunnya – belum termasuk lapangan kerja baru yang akan tercipta.

Dengan mempertimbangkan semua ini, aturan yang biasa berlaku. Dari sudut pandang bisnis, AI tidak jauh berbeda dengan perkembangan teknologi lainnya. Anda perlu menilai dampaknya dan mempertimbangkan apakah, di mana, dan bagaimana teknologi berbasis AI akan berguna. Jangan mengikuti kawanan secara membabi buta. Hanya karena orang lain mengatakan mereka melakukannya, bukan berarti hal itu tepat untuk Anda. Tentu saja, Anda tidak boleh membenamkan kepala Anda di pasir. Anda harus proaktif dalam mencari ide-ide baru dan cara kerja yang berbeda, namun buatlah penilaian Anda sendiri mengenai nilai yang dapat ditambahkan oleh kecerdasan buatan dan pembelajaran mesin. Selain itu, jangan bergantung pada pihak ketiga, yang berkepentingan untuk menjual solusi kepada Anda, untuk mengukur manfaatnya bagi Anda.

Hampir setiap hari, saya menerima email dari perusahaan ilmu data yang menawarkan untuk merevolusi sistem TI saya, mengotomatiskan proses, dan meningkatkan kemampuan analisis data saya. Bagaimana cara mengetahui mana yang menawarkan sesuatu kepada saya dan mana yang hanya memanfaatkan teknologi ini?

Seperti yang telah kita bahas di bab sebelumnya, teknologi harus dapat diterapkan pada permasalahan dan proses bisnis nyata jika ingin memberikan nilai tambah. Solusi berbasis pembelajaran mesin hanya akan menambah nilai jika membantu Anda mencapai setidaknya satu hal berikut:

- **Mengurangi biaya.** Ini memfasilitasi melakukan sesuatu dengan lebih efisien, seperti pengendalian gudang dan inventaris yang lebih baik atau mengoptimalkan harga produk.
- **Peningkatan kualitas.** Solusi ini memberikan produk dan layanan yang lebih baik. Misalnya, pemenuhan pesanan pelanggan lebih cepat atau berkurangnya keluhan pelanggan.
- **Menciptakan peluang baru.** Pembelajaran mesin bertindak sebagai pendorong yang memungkinkan Anda melakukan sesuatu yang belum dapat Anda lakukan sebelumnya. Mungkin ini untuk memasuki pasar baru atau menyediakan jenis produk dan layanan yang berbeda.

Satu hal yang harus diwaspadai adalah apa yang saya sebut sebagai “oportunis data”. Saya telah bertemu dengan lebih dari satu perusahaan analitik yang menyarankan agar saya: “Pekerjakan beberapa ilmuwan data kami untuk melihat apa yang dapat kami temukan dari

data Anda.” Tidak ada penyelidikan mendalam mengenai apa yang saya atau organisasi saya lakukan, apa permasalahan kami, atau di mana letak ketidakefisienan kami. Hanya diasumsikan bahwa dengan membayar orang-orang ini untuk menjangkau data saya, dan mungkin menghasilkan beberapa model prediktif, mereka akan memberikan sesuatu yang berguna.

Saya ingat berbicara dengan seorang teman saya baru-baru ini, yang telah mempekerjakan beberapa ilmuwan data untuk melakukan hal ini untuk organisasinya (perusahaan jasa keuangan). Mereka mengatakan bahwa mereka dapat membantunya memprediksi pelanggan mana yang akan menguangkan investasi mereka. Jadi, ini adalah masalah tipe atrisi (attrition type problem) – mengidentifikasi pelanggan yang kemungkinan besar akan keluar dan kemudian mencoba membujuk mereka untuk tetap tinggal.

Teman saya memberi tahu saya, bahwa setelah bekerja selama berminggu-minggu, data scientist yang dia pekerjakan menghasilkan model prediktif yang pada dasarnya memperkirakan masa pensiun; yaitu orang-orang yang mendekati masa pensiun cenderung meninjau kembali investasi mereka sebagai persiapan untuk berhenti bekerja. Para ilmuwan data tampak sangat senang dengan apa yang mereka temukan. Namun, ini adalah sesuatu yang sudah diketahui oleh teman saya dan semua orang di industri ini. Latihan ini tidak memberikan nilai tambah tetapi memerlukan biaya puluhan ribu dolar untuk dilakukan. Saya dapat menjelaskan sejumlah situasi serupa lainnya, tetapi saya rasa Anda mengerti maksudnya.

Hal lain yang harus Anda lakukan adalah memikirkan apa yang terjadi di industri Anda dan pesaing Anda. Bicaralah dengan Anda jaringan kontak. Lihat apa yang dilakukan orang lain di ruang masalah Anda. Jika tidak ada orang lain yang menggunakan kecerdasan buatan/teknologi pembelajaran mesin untuk masalah semacam itu, maka Anda berada dalam ruang berisiko tinggi. Mungkin Anda bisa menjadi pelopor dan menjadi yang terdepan dengan menjadi orang pertama yang mendapatkan manfaatnya. Alternatifnya, mungkin Anda mengambil pertaruhan besar-besaran yang sepertinya tidak akan membuahkan hasil. Jika tidak ada orang lain yang mengadopsi teknologi untuk tujuan tersebut, maka itu adalah sebuah risiko.

Jika Anda memiliki budaya berisiko tinggi, dan mengambil risiko adalah bagian dari permainan, maka lakukanlah. Namun, jika seperti kebanyakan dari kita, Anda beroperasi di lingkungan yang lebih konservatif dan potensi risiko kegagalan semakin besar, maka pikirkan baik-baik sebelum terjun ke dalamnya. Anda perlu mengerjakan pekerjaan rumah Anda dan memastikan bahwa manfaat yang dijanjikan akan tercapai. disampaikan. Paling tidak, ketika masalah sudah mereda beberapa bulan kemudian, Anda harus mampu menunjukkan bahwa keputusan Anda rasional dan berdasarkan apa yang Anda ketahui saat itu.

Salah satu strategi yang dapat diikuti ketika berinteraksi dengan pemasok, khususnya ketika nilai suatu proposisi tidak jelas atau tidak pasti, adalah dengan menyarankan model pengembalian bersama dalam perjanjian kontrak. Serangkaian tujuan yang jelas telah disepakati di awal dan pemasok mengambil bagian dari keuntungan (atau kerugian) yang dihasilkan. Jika pemasok menjual kepada Anda semua manfaat dari solusi berbasis

pembelajaran mesin mereka, namun tidak mau menerima model kompensasi risiko/imbalan bersama, tanyakan pada diri Anda mengapa hal itu mungkin terjadi? Bahkan jika Anda tidak benar-benar ingin menempuh cara ini, respons yang Anda dapatkan dari pemasok akan sangat informatif.

Ingatlah juga bahwa hanya karena sesuatu dapat diotomatisasi bukan berarti Anda harus melakukannya atau Anda sebaiknya melakukannya. Silicon Valley memiliki reputasi dalam mengembangkan gadget inovatif dan kemudian mencoba menemukan pasar bagi gadget tersebut, dibandingkan memulai dengan masalah masyarakat dan kemudian mencoba mencari solusi. Jika kita kembali ke mobil self-driving – tahukah kita berapa banyak orang yang sebenarnya menginginkan mobil self-driving? Studi konsumen terkini menunjukkan bahwa saat ini jumlah konsumen hanya berjumlah sedikit. Hal ini mungkin berubah seiring berjalannya waktu, tetapi ini adalah sesuatu yang harus diwaspadai oleh produsen motor.

Secara keseluruhan, apa yang ingin saya sampaikan dalam bab ini adalah tetap berpikiran terbuka, namun tetap berkepala dingin. AI dan pembelajaran mesin menghadirkan peluang besar untuk melakukan berbagai hal dengan lebih baik, lebih efisien, dan lebih murah. Namun, sama seperti aspek bisnis lainnya – jangan langsung terjun sebelum Anda mengetahui peluang apa saja yang ada, Anda tahu apa yang ingin Anda lakukan dan alasannya, dan Anda telah melakukan perhitungan manfaat yang tepat untuk mendukung keputusan Anda.

LATIHAN SOAL

Dalam lampiran ini, kami akan fokus pada pertanyaan yang harus Anda jawab untuk memastikan bahwa proyek pembelajaran mesin Anda berhasil, dan memaksimalkan manfaat yang dapat diberikannya pada proses bisnis. Ketika saya mengatakan proses bisnis, saya pada dasarnya berbicara tentang mengganti atau melengkapi tugas-tugas pengambilan keputusan yang biasanya dilakukan karyawan.

Penekanannya adalah pada apa yang saya sebut sebagai masalah tipe “AI sederhana” di bab-bab sebelumnya. Di sinilah seseorang mungkin ingin menggunakan pendekatan pembelajaran mesin untuk melakukan hal-hal seperti memutuskan pelamar kerja mana yang akan diwawancarai, membuat diagnosis awal dari pemindaian medis, memutuskan bagaimana pelanggan harus diperlakukan, dan seterusnya. Saran di bagian ini mungkin kurang relevan dengan pengembangan sistem “AI yang kompleks” seperti pengoperasian kendaraan otonom atau pengembangan asisten digital pribadi. Namun, sebagian besar saran tetap berlaku untuk jenis aplikasi ini juga.

Berikut ini adalah daftar singkat pertanyaan yang harus Anda tanyakan saat mempertimbangkan penggunaan pembelajaran mesin dalam organisasi Anda. Kecuali Anda dapat menjawab semua pertanyaan ini dengan tingkat kepastian yang tinggi sebelum proyek meninggalkan fase persyaratan bisnis, maka ada kemungkinan besar proyek tersebut akan gagal. Akibatnya, Anda berisiko menghabiskan banyak waktu dan uang untuk proyek tersebut tanpa manfaat yang terlihat di akhir proyek.

1. Apa masalah bisnis Anda? Pembelajaran mesin tidak memiliki nilai intrinsik. Itu harus diterapkan pada masalah atau tujuan tertentu jika ingin bermanfaat. Jika Anda tidak dapat mengidentifikasi masalah bisnis yang perlu diatasi, berhentilah sekarang juga!
2. Metrik apa yang ingin Anda optimalkan? Ini harus menjadi sesuatu yang bisa diukur secara sederhana dan mana yang penting bagi Anda. Penghematan waktu, peningkatan keuntungan, pengurangan biaya, kepuasan pelanggan, dan nyawa yang diselamatkan adalah contoh umum. Inilah yang akan dioptimalkan oleh proses pembelajaran mesin.
3. Aturan dan tindakan pengambilan keputusan seperti apa yang akan Anda terapkan? Anda harus membuat keputusan berdasarkan keluaran dari proses pembelajaran mesin dan kemudian bertindak berdasarkan keputusan tersebut. Jika Anda tidak dapat memikirkan apa yang akan Anda lakukan dengan model prediktif, mengapa harus melakukannya?
4. Dari mana data untuk melakukan pembelajaran mesin berasal? Pembelajaran mesin memerlukan setidaknya beberapa ratus, dan idealnya ribuan, contoh kasus untuk dikerjakan. Untuk sebagian besar masalah pembelajaran mesin, metrik yang diinginkan (item 2 di atas) harus tersedia untuk kasus ini.
5. Bagaimana Anda mengoperasionalkan proses pembelajaran mesin? Sistem atau proses apa yang akan menerapkan model prediktif dan aturan keputusan? Ini akan memerlukan waktu dan uang. Siapa yang akan melakukannya dan siapa yang akan membayarnya? Model dan aturan pengambilan keputusan tidak akan dapat diterapkan dengan sendirinya.

6. Apa saja risiko etika dan hukum yang terkait dengan otomatisasi? Gunakan pedoman di Bab 13 untuk menilai hal ini berdasarkan dampak terhadap pelanggan, kekekalan data, dan penerima manfaat. Mitigasi apa yang akan Anda lakukan untuk mengelola risiko? Apakah Anda juga sudah mempertimbangkan undang-undang perlindungan data yang berlaku di wilayah Anda?
7. Apakah semua data yang diperlukan tersedia secara operasional? Minta ilmuwan data memberikan daftar periksa terhadap setiap item data yang ditampilkan dalam model mereka. Hal ini untuk memastikan bahwa data yang dibutuhkan model ada dalam platform implementasi. Jika ada item data yang tidak tersedia maka Anda tidak akan dapat menggunakan solusi yang telah dikembangkan.
8. Apakah solusi tersebut diterapkan secara aktif atau pasif, seperti dijelaskan pada Bab 11? Jika dalam bentuk pasif, kontrol apa yang Anda perlukan untuk memastikan bahwa penggantian hanya terjadi jika hal tersebut benar? Jika orang meninjau, dan berpotensi mengesampingkan, setiap keputusan yang dibuat oleh sistem, maka sistem tidak akan memberikan nilai tambah apa pun.
9. Bagaimana Anda menilai keberhasilan proyek? Manfaat teoritis yang disajikan pada akhir fase analitis proyek pembelajaran mesin hanya merupakan indikasi dari apa yang sebenarnya terjadi dalam kehidupan nyata. Proses pemantauan diperlukan untuk menilai sistem ketika sudah berjalan. Hal ini bertujuan agar Anda dapat mengukur manfaat aktual yang diperoleh dibandingkan dengan manfaat yang dijanjikan.
10. Berapa biaya pemeliharaan sistem yang berkelanjutan? Seiring berjalannya waktu, data baru, peraturan baru, perubahan kebutuhan bisnis, dan sebagainya, berarti bahwa sistem tersebut memerlukan modifikasi dan pada akhirnya mungkin perlu dikembangkan kembali dari awal. Seberapa sering hal ini diperkirakan akan terjadi dan berapa besar kemungkinan biayanya? Jangan lupa untuk memasukkan biaya untuk kepatuhan terhadap peraturan dan audit tahunan.

LAMPIRAN A

MENGEVALUASI MODEL PREDIKTIF

Setelah model prediktif dibangun, model tersebut perlu dinilai untuk menentukan seberapa baik model tersebut memprediksi perilaku. Terkadang terdapat satu skor batas yang menghasilkan tercapainya tujuan bisnis, yang dapat ditentukan dengan melihat tabel distribusi skor. Demikian pula, untuk hal-hal seperti sistem pengenalan objek yang setiap skor mewakili satu hasil, Anda cukup memilih hasil dengan skor tertinggi.

Namun dalam banyak situasi, tidak ada satu pun skor batas yang dapat dipilih. Sebaliknya, ada sejumlah kemungkinan pemutusan hubungan kerja (cut-off). Salah satu dari batasan ini akan mencapai tujuan bisnis awal, namun dengan kelebihan dan kekurangan yang berbeda. Demikian pula, praktik yang umum adalah membangun beberapa model prediksi yang berbeda. Kinerja model kemudian dibandingkan satu sama lain sehingga dipilih model terbaik.

Jika Anda mengingat kembali studi kasus penyakit jantung yang digunakan dalam buku ini, Anda akan ingat bahwa kita telah melihat model pohon keputusan dan model kartu skor. Kedua model tersebut memberikan solusi yang memenuhi tujuan bisnis yang dinyatakan;

yaitu mengidentifikasi setidaknya 50% orang yang akan mengidap penyakit jantung tetapi mengundang tidak lebih dari 5% dari seluruh populasi untuk melakukan pemeriksaan. Namun, model kartu skor secara keseluruhan memiliki kinerja yang lebih baik. Hal ini karena pohon keputusan mengidentifikasi 62% kasus dengan skor tertinggi 5%, sedangkan pohon keputusan hanya mengidentifikasi 55% kasus.

Anda juga dapat berargumentasi bahwa kedua model tersebut disampaikan secara berlebihan. Mereka mengidentifikasi lebih dari 50% kasus yang disyaratkan. Artinya, sebenarnya terdapat fleksibilitas dalam cara pemilihan batas waktu; yaitu terdapat lebih dari satu cut-off yang memenuhi tujuan. Dalam situasi ini, dua cara yang mungkin digunakan untuk menggunakan model tersebut adalah:

1. Maksimalkan hasil. Pilih batas yang menghasilkan tepat 5% populasi diundang untuk melakukan pemeriksaan; yaitu menggunakan semua sumber daya yang tersedia. Hal ini akan mengakibatkan lebih dari 50% kasus penyakit jantung teridentifikasi (ini adalah jawaban yang disajikan sebelumnya dalam studi kasus; yaitu skor batas 521 untuk kartu skor atau skor batas 12 untuk pohon keputusan).
2. Minimalkan sumber daya. Pilih batas yang menyebabkan tepat 50% kasus penyakit jantung diundang untuk pemeriksaan. Artinya, kurang dari 5% populasi yang akan diundang untuk melakukan pemeriksaan karena yang akan dipilih adalah nilai batas di atas 521.

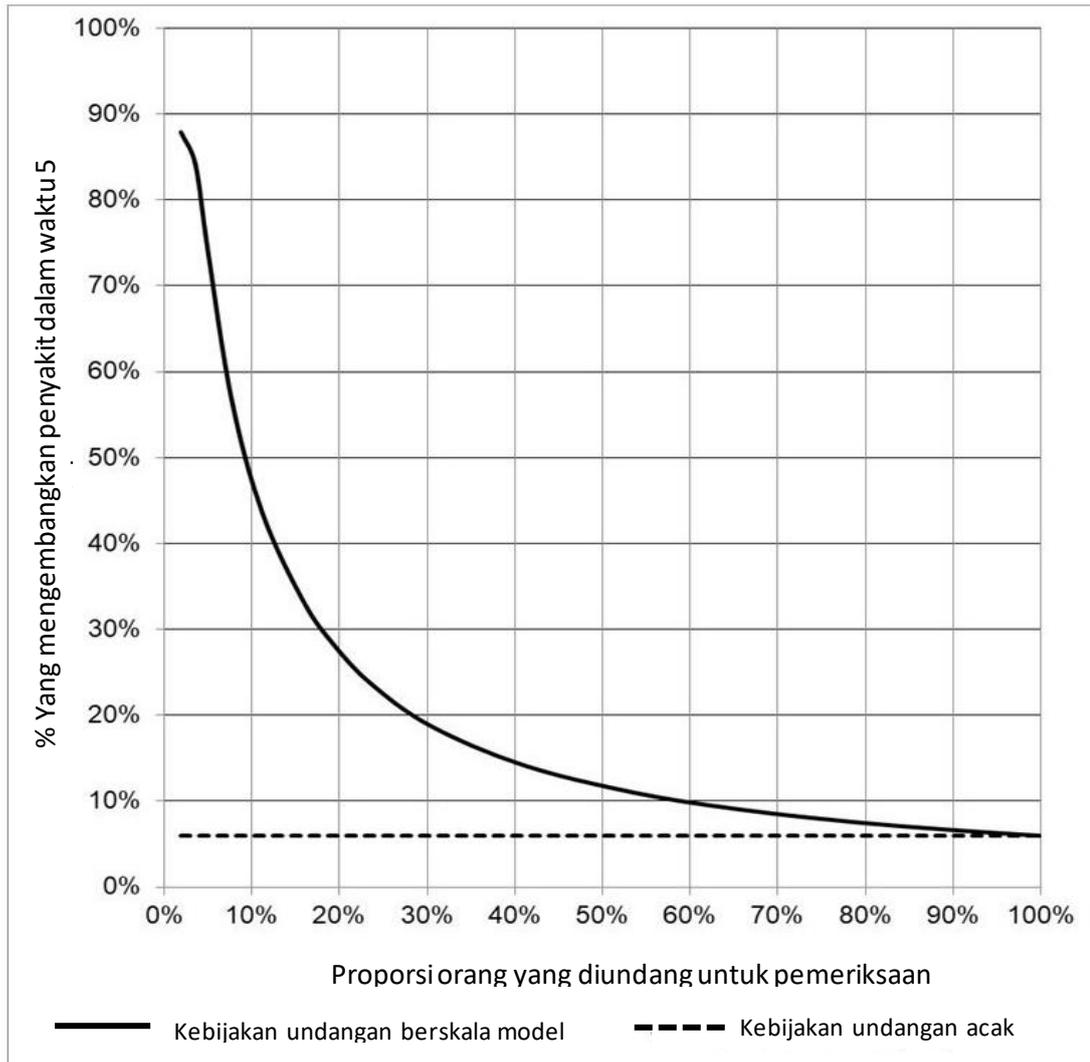
Pilihan mana yang harus dipilih bergantung pada hal yang paling penting bagi otoritas kesehatan. Jika ingin memaksimalkan identifikasi penyakit jantung maka pilihan 1 adalah yang terbaik. Jika ingin menggunakan sumber daya sesedikit mungkin maka opsi 2 lebih baik. Tentu saja, otoritas kesehatan juga dapat memilih batas skor yang berada di antara kedua ekstrem tersebut.

Secara umum, dalam situasi seperti ini ada dua pertimbangan ketika memilih batas waktu:

- *Kemurnian (Angkat)*. Jika dipilih cut-off yang tinggi, maka hanya sedikit kasus yang akan dipilih, namun dari kasus-kasus yang dipilih proporsi yang tinggi akan menampilkan karakteristik yang diminati.
- *Volume (Keuntungan)*. Jika batas yang lebih rendah dipilih, maka lebih banyak kasus yang akan dipilih dan secara keseluruhan akan lebih banyak kasus yang diminati dalam kelompok yang dipilih (kasus yang berada di atas batas). Namun, kemurnian kasus-kasus tertentu akan menurun; yaitu proporsi yang lebih rendah dari mereka yang berada di atas batas akan menampilkan karakteristik yang Anda minati.

Untuk membantu menentukan strategi cut-off dan trade-off antara kemurnian dan volume, dua alat populer yang digunakan untuk menilai dampak dari strategi cut-off yang berbeda adalah Lift Chart dan Gains Chart. Ini memberikan representasi visual yang mudah digunakan tentang kinerja model.

Dimulai dengan bagan peningkatan, ini memberikan gambaran kemurnian. Gambar A.1 memberikan contoh diagram lift untuk kartu skor penyakit jantung yang disajikan pada



Gambar A.1 Grafik peningkatan untuk kartu skor penyakit jantung

Pada Gambar A.1 sumbu X menunjukkan proporsi masyarakat yang diundang untuk pemeriksaan. Sumbu Y menunjukkan proporsi mereka yang diundang yang diperkirakan akan mengidap penyakit ini dalam waktu 5 tahun. Garis lurus putus-putus adalah tarif dasar. Angka dasar adalah persentase orang yang diperkirakan akan terkena penyakit jantung selama 5 tahun ke depan dalam populasi secara keseluruhan, yaitu 6%. Dengan kata lain, tarif dasar mewakili strategi pemilihan acak untuk mengundang orang melakukan pemeriksaan.

Garis lengkung padat menunjukkan apa yang akan terjadi jika proses penentuan prioritas dalam menentukan siapa yang akan diundang untuk pemeriksaan didasarkan pada skor dari model kartu skor. Misalnya, grafik menunjukkan bahwa jika 40% populasi dengan skor tertinggi diundang untuk melakukan pemeriksaan, maka sekitar 15% dari populasi tersebut akan menderita penyakit jantung di masa depan. Jika volume dikurangi dan katakanlah hanya 10% yang diundang untuk pemeriksaan, maka proporsi (kemurnian) akan meningkat menjadi sekitar 50%.

Semakin kecil proporsi sasaran intervensi medis, semakin tinggi proporsi penderita penyakit. Dengan kata lain, jika Anda hanya menargetkan sejumlah kecil orang dalam populasi

dengan skor tertinggi, maka kita dapat memperkirakan bahwa proporsi orang-orang terpilih yang mengidap penyakit jantung akan sangat tinggi. Namun jika kelompok yang diundang diperbesar, dengan memasukkan orang-orang yang skornya lebih rendah, maka secara keseluruhan proporsi yang diidentifikasi dengan benar (kemurnian) akan menurun.

Oleh karena itu, bagan pengangkatan adalah cara membandingkan trade-off antara volume dan kemurnian. Bagaimana diagram gaya angkat dibuat?

Kelompok	Rentang skor		Jumlah orang	% populasi	Penderita penyakit jantung setelah 5 tahun	% Penderita penyakit jantung setelah 5 tahun	Kumulatif menurun				Kemurnian	
	dari	Ke					Jumlah orang	% populasi	Penderita penyakit jantung setelah 5 tahun	% Penderita penyakit jantung setelah 5 tahun		Kemurnian
1	0	300	55,950	11.19%	40	0.07%	500,000	100.00%	30,000	6.00%	6.00%	
2	301	320	56,606	11.32%	68	0.12%	444,050	88.81%	29,960	6.75%	6.75%	
3	321	340	59,700	11.94%	129	0.22%	387,444	77.49%	29,892	7.72%	7.72%	
4	341	360	58,706	11.74%	216	0.37%	327,744	65.55%	29,763	9.08%	9.08%	
5	361	380	64,429	12.89%	403	0.63%	269,038	53.81%	29,457	10.98%	10.98%	
6	381	400	52,749	10.55%	575	1.09%	204,609	40.92%	29,144	14.24%	14.24%	
7	401	420	34,089	6.82%	600	1.76%	151,860	30.37%	28,569	18.81%	18.81%	
8	421	440	21,107	4.22%	632	2.99%	117,771	23.55%	27,969	23.75%	23.75%	
9	441	460	17,269	3.45%	878	5.09%	96,664	19.33%	27,337	28.28%	28.28%	
10	461	480	23,364	4.67%	2,020	8.65%	79,395	15.88%	26,459	33.33%	33.33%	
11	481	500	17,477	3.50%	2,553	14.61%	56,032	11.21%	24,439	43.62%	43.62%	
12	501	520	13,554	2.71%	3,366	24.84%	38,554	7.71%	21,885	56.77%	56.77%	
13	521	540	7,103	1.42%	3,463	48.76%	25,000	5.00%	18,519	74.08%	74.08%	
14	541	560	8,260	1.65%	6,587	79.74%	17,897	3.58%	15,056	84.12%	84.12%	
15	561	999	9,637	1.93%	8,469	87.88%	9,637	1.93%	8,469	87.88%	87.88%	
Total			500,000		30,000	6.00%						

Gambar A.2 versi tabel distribusi skor yang diperluas untuk kartu skor penyakit jantung

Empat kolom tambahan pertama di sebelah kanan Gambar A.2 menunjukkan angka kumulatif menurun; yaitu menunjukkan jumlah dan proporsi kasus yang mendapat skor sama atau di atas skor tertentu. Kemurnian / Lift (kolom paling kanan) dihitung dengan membagi jumlah kumulatif penderita penyakit jantung dengan total kumulatif. Misalnya, peningkatan pada skor 501 dihitung sebagai:

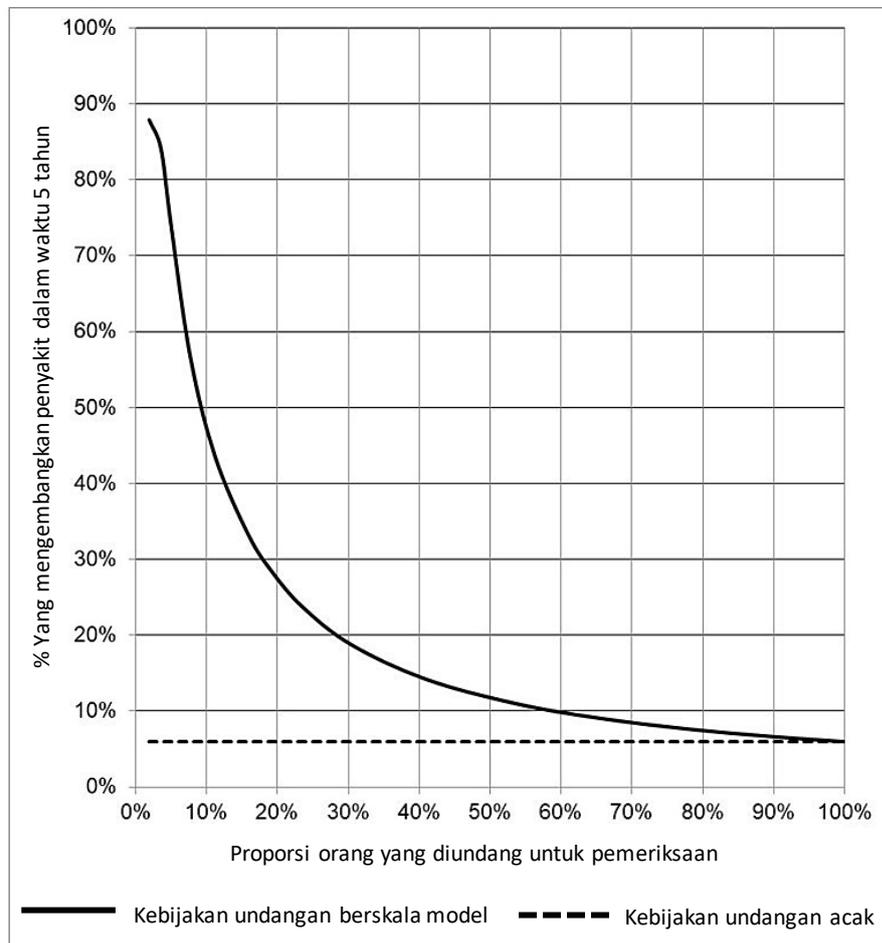
$$\text{Peningkatan} = 21.885 / 38.554 = 56,77\%$$

Artinya, 56,77% dari mereka yang mendapat skor di atas 501 diperkirakan mengidap penyakit jantung. Bagan peningkatan kemudian dibuat dengan memplot kolom % kumulatif populasi terhadap peningkatan.

Bagan peningkatan memberikan perspektif kemurnian populasi di seluruh rentang kemungkinan skor batas. Grafik keuntungan di sisi lain digunakan untuk memberikan tampilan berorientasi volume. Gambar A.3 menunjukkan grafik perolehan kartu skor penyakit jantung.

Grafik peningkatan dihasilkan dengan memplot persentase kumulatif total populasi yang berada pada atau di atas skor tersebut, dibandingkan dengan proporsi kumulatif kasus penyakit jantung yang berada pada atau di atas skor yang sama; yaitu kolom paling kanan kedua dan keempat pada Gambar A.2. Misalnya pada Gambar A.3 terlihat bahwa jika seseorang mengundang 10% penduduknya untuk memeriksakan diri, maka hal itu akan mengakibatkan sekitar 80% dari seluruh kasus penyakit jantung di Indonesia. populasi yang

dimasukkan dalam daftar undangan. Demikian pula, jika seseorang mengundang 30% populasi, maka sekitar 95% dari seluruh kasus akan masuk dalam daftar undangan.



Gambar A.3 Gambar peningkatan untuk kartu skor penyakit jantung

Bagan peningkatan dan bagan perolehan adalah alat yang sangat berguna yang dapat dibuat untuk semua jenis model klasifikasi guna membantu pengguna bisnis memutuskan bagaimana model harus digunakan. Grafik serupa juga dapat dihasilkan untuk model tipe regresi.

Ilmuwan data juga akan menggunakan metrik khusus untuk mengukur seberapa baik keakuratan prediksi suatu model secara keseluruhan. Mungkin tiga tindakan paling populer yang digunakan untuk melakukan hal ini adalah:

- AUC (area di bawah kurva). Ini adalah ukuran populer untuk mengevaluasi model klasifikasi. Nilai AUC berkisar antara 0,5 hingga 1. Nilai 0,5 menunjukkan bahwa suatu model tidak memiliki kemampuan prediktif; yaitu tidak berguna. Nilai 1,0 menunjukkan prediktor sempurna yang selalu memberikan hasil yang benar. Statistik GINI dan Somers'D yang sering dikutip merupakan ukuran yang serupa[88].
- Persentase yang diklasifikasikan dengan benar (PCC). Ini hanyalah proporsi kejadian yang diklasifikasikan dengan benar untuk batas skor tertentu. Jika suatu peristiwa mendapat skor di atas skor batas, maka model memprediksi dengan benar. Jika non-peristiwa terjadi

dan skornya di bawah batas, maka model tersebut juga telah melakukannya dengan benar. Untuk kartu skor penyakit jantung dengan batas 521, proporsi kasus yang diklasifikasikan dengan benar adalah 96,4%. Ketika orang berbicara tentang “akurasi model” atau model “akurat X%”, yang biasanya mereka maksud adalah PCC.

- (Disesuaikan) R-kuadrat. Ini adalah ukuran paling populer untuk mengevaluasi model regresi. Hal ini didasarkan pada perbedaan antara nilai aktual dan prediksi untuk setiap observasi dalam sampel pengembangan. Nilai R-kuadrat sebesar nol menunjukkan bahwa suatu model tidak memiliki kemampuan prediktif; yaitu tidak berguna. Nilai 1 menunjukkan prediktor sempurna yang selalu tepat sasaran.

Ukuran seperti Adjusted R-squared, AUC, dan PCC digunakan untuk membandingkan performa berbagai model dan untuk menilai bagaimana performa model berubah seiring waktu. Saat model prediktif baru dibuat, performa model lama dan model baru dibandingkan untuk melihat seberapa baik model baru tersebut. Demikian pula, jika dua atau lebih tipe model berbeda dikembangkan untuk suatu permasalahan tertentu, maka model dengan AUC, PCC, R-Squared tertinggi, dan seterusnya akan dianggap yang terbaik; yaitu yang paling prediktif.

Ada lusinan metrik lain yang telah dikembangkan orang untuk mengukur seberapa baik kinerja model, namun AUC dan PCC sejauh ini merupakan metrik paling populer untuk menilai kinerja model klasifikasi, dan Adjusted R-Squared adalah ukuran paling populer untuk menilai regresi. model.

Jika Anda ingin mempelajari lebih lanjut tentang hal ini dan langkah-langkah lain untuk menilai model prediktif, ada banyak sumber daya online yang tersedia, termasuk sejumlah besar entri Wikipedia.

LAMPIRAN B

ISTILAH POPULER DALAM PEMBELAJARAN MESIN DAN AI

Fungsi aktivasi. Persamaan (rumus matematika) yang digunakan untuk mengubah keluaran mentah suatu neuron menjadi keluaran akhir (skor). Fungsi aktivasi sering kali digunakan untuk memaksa keluaran neuron berada pada rentang yang tetap. Ini biasanya antara 0 dan 1, namun beberapa fungsi aktivasi menghasilkan keluaran dengan rentang yang berbeda. Fungsi aktivasi yang umum adalah fungsi logistik, fungsi tangen hiperbolik, dan fungsi identifikasi.

Algoritma. Algoritma adalah sekumpulan instruksi atau prosedur yang dijalankan secara berurutan untuk menyelesaikan suatu masalah tertentu. Dalam pembelajaran mesin, berbagai algoritma diterapkan untuk menemukan pola dalam data dan membuat model prediktif.

Kecerdasan Buatan (AI). Replikasi kemampuan analitis dan pengambilan keputusan biologis (terutama manusia). Misalnya, pengenalan objek dan ucapan, memprediksi perilaku masa depan masyarakat dan kemampuan bermain permainan seperti catur, poker, dan Go.

Data besar. Kumpulan data yang bervariasi dan dapat diubah dalam jumlah yang sangat besar, yang sulit diproses menggunakan PC/Server standar; yaitu biasanya berukuran terabyte atau lebih besar. Teknologi Big Data memanfaatkan arsitektur komputer canggih dan perangkat lunak khusus untuk memfasilitasi pemrosesan cepat kumpulan data ini. Pembelajaran mesin adalah salah satu alat utama yang digunakan untuk mengekstraksi nilai dari Big Data.

Model klasifikasi. Model prediktif yang memprediksi apakah suatu peristiwa akan terjadi atau tidak. Misalnya, kemungkinan dua orang akan cocok berkenan, kemungkinan seseorang akan membeli jenis mobil tertentu, atau kemungkinan seseorang terkena diabetes dalam 10 tahun mendatang. Skor yang dihasilkan oleh model klasifikasi merupakan perkiraan kemungkinan terjadinya suatu peristiwa. Jangan bingung dengan model regresi.

Clustering. Cluster adalah kelompok yang berisi orang-orang atau benda-benda yang mempunyai sifat serupa. Misalnya, orang-orang dengan usia dan pendapatan yang sama mungkin berada dalam satu kelompok, sedangkan orang-orang dengan peran pekerjaan dan ukuran keluarga yang sama mungkin berada di kelompok lain. Algoritme pengelompokan adalah salah satu bentuk pembelajaran tanpa pengawasan (lihat di bawah).

Jaringan (saraf) yang berbelit-belit / A convoluted (neural) network. Jenis jaringan saraf dalam yang tidak semua neuron pada suatu lapisan terhubung ke semua neuron pada lapisan berikutnya. Dengan bobot yang lebih sedikit, waktu (yang sangat besar) yang dibutuhkan untuk melatih jaringan menjadi jauh lebih sedikit. Aplikasi klasiknya adalah pengenalan objek. Bayangkan Anda memiliki gambar yang terdiri dari $256 * 256$ (65.536)

piksel. Jaringan saraf tradisional memerlukan 65.536 bobot untuk setiap neuron di lapisan pertama, satu untuk setiap piksel! Dengan mengelompokkan gambar menjadi 64 sub-wilayah yang masing-masing berukuran 32 *32 piksel dan hanya menghubungkan neuron dalam setiap sub-wilayah, jumlah bobot dikurangi menjadi hanya 1.024 per neuron. Hal ini berfungsi karena fitur terpenting dalam gambar biasanya berdekatan (yaitu dalam sub-wilayah yang sama atau bertetangga). Piksel di sisi berlawanan dari gambar kurang penting.

Korelasi. Variabel yang satu dikatakan berkorelasi dengan variabel yang lain jika perubahan pada variabel tersebut terjadi bersamaan dengan perubahan pada variabel yang lain. Penting untuk dipahami bahwa hal ini tidak berarti bahwa suatu hal disebabkan oleh hal lain (penyebab), walaupun mungkin saja demikian. Migrasi bangau berkorelasi dengan kelahiran (lebih banyak anak yang lahir di musim semi saat bangau bermigrasi), namun migrasi bangau tidak menyebabkan kelahiran!

Penambangan data / Data mining. Penambangan data adalah ilmu untuk menemukan informasi berguna dalam kumpulan data yang tidak dapat diidentifikasi dengan mudah oleh manusia hanya dengan memperhatikannya. Penambangan data menggunakan berbagai algoritma dan prosedur otomatis yang diambil dari berbagai disiplin ilmu yang mencakup statistik, komputasi, dan kecerdasan buatan/pembelajaran mesin.

Ilmu data / Data Science. Nama yang diberikan untuk keterampilan/seni yang mampu menggabungkan pengetahuan matematika (pembelajaran mesin) dengan data dan keterampilan TI secara pragmatis, untuk menghasilkan solusi berbasis pembelajaran mesin yang bernilai tambah dan praktis.

Ilmuwan data. Nama yang diberikan kepada seseorang yang melakukan ilmu data. Ilmuwan data yang baik fokus dalam memberikan solusi berguna yang dapat diterapkan di lingkungan dunia nyata. Mereka tidak terlalu terpaku pada teori.

Aturan pengambilan keputusan. Model prediktif menghasilkan skor. Aturan pengambilan keputusan kemudian digunakan untuk memutuskan bagaimana orang diperlakukan berdasarkan skor. Mereka yang mendapat skor di atas skor tertentu (skor batas) menerima satu perlakuan, dan mereka yang mendapat skor di bawah batas mendapat perlakuan lain. Misalnya, ketika menilai seseorang untuk suatu kondisi medis, hanya mereka yang mendapat skor di atas batas yang ditentukan; yaitu mereka yang memiliki risiko tertinggi terkena kondisi tersebut, ditawarkan pengobatan.

Pohon keputusan. Suatu jenis model prediktif yang dibuat menggunakan algoritma yang secara rekursif mengelompokkan suatu populasi ke dalam kelompok-kelompok yang lebih kecil dan lebih kecil. Juga dikenal sebagai Pohon Klasifikasi dan Regresi atau CART (Karena dapat digunakan untuk klasifikasi dan regresi.)

Pembelajaran mendalam. Model prediktif berdasarkan jaringan saraf kompleks (atau arsitektur terkait), yang berisi banyak neuron yang tersebar di banyak lapisan. Pembelajaran mendalam terbukti berhasil dalam masalah kompleks seperti pengenalan objek dan terjemahan bahasa.

- Sampel pengembangan.** Kumpulan data yang digunakan oleh proses pembelajaran mesin untuk membuat model prediktif. Sampel pengembangan harus berisi setidaknya beberapa ratus kasus, tetapi biasanya digunakan sampel yang lebih besar yang berisi ribuan atau jutaan kasus.
- Ansambel.** Model prediktif terdiri dari beberapa model tambahan (sedikitnya 3 atau terkadang ribuan). Semua model tambahan memprediksi hasil yang sama namun dibuat menggunakan metode berbeda dan/atau menggunakan data berbeda. Artinya model tidak selalu menghasilkan prediksi yang sama. Ansambel ini menggabungkan semua prediksi individu untuk menghasilkan satu prediksi akhir. Model ansambel sering kali (tetapi tidak selalu) secara signifikan mengungguli model individu terbaik.
- Umpan maju jaringan saraf.** Jaringan saraf yang semua koneksinya terjadi dalam satu arah, dari satu lapisan neuron ke lapisan berikutnya. Tidak ada koneksi ke belakang atau antar-lapisan. Sebagian besar jaringan saraf adalah jaringan feed forward.
- Cakrawala ramalan.** Banyak (tetapi tidak semua) model prediktif memperkirakan peristiwa atau kuantitas yang tidak diketahui di masa depan. Cakrawala perkiraan adalah kerangka waktu yang digunakan oleh suatu model untuk memprediksi. Model pemasaran biasanya memprediksi respons dalam jangka waktu perkiraan jam atau hari. Dalam dunia kedokteran, cakrawala perkiraan beberapa tahun digunakan ketika memprediksi probabilitas kelangsungan hidup untuk kondisi medis tertentu.
- Grafik keuntungan.** Alat visual yang umum digunakan untuk mendemonstrasikan manfaat suatu model. Sering digunakan bersamaan dengan diagram lift. Lihat Lampiran A untuk informasi lebih lanjut.
- Siklus Hype Gartner.** Siklus hype Gartner mengklasifikasikan teknologi baru ke dalam lima fase yang mencerminkan kecenderungan kita untuk memperkirakan nilainya secara berlebihan dan kemudian meremehkannya, sebelum mencapai tingkat di tengah-tengahnya.
- AI Umum.** Sistem buatan, yang bertindak dengan cara yang cerdas bagi manusia di berbagai bidang masalah, dan dapat beradaptasi dengan situasi baru seperti yang bisa dilakukan manusia, digambarkan memiliki AI umum. Tidak ada sistem AI yang digunakan saat ini yang dapat dikatakan memiliki AI umum. Lihat juga AI Sempit.
- Hadoop.** Solusi penyimpanan data yang memanfaatkan banyak PC desktop murah yang tersedia untuk menyimpan dan memproses data dalam jumlah besar dengan sangat cepat.
- IoT.** Ini menggambarkan perangkat cerdas sehari-hari seperti lemari es, mesin cuci, dan bahkan bola lampu yang dapat dihubungkan ke internet atau sistem lain. Hal ini memungkinkan data tentang penggunaannya dikumpulkan dan pengguna dapat mengontrol perangkat ini dari jarak jauh, misalnya, mengatur mesin cuci saat saya dalam perjalanan pulang melalui aplikasi ponsel pintar.
- Pengelompokan K-Means.** Pendekatan pengelompokan yang populer. Hal ini bertujuan untuk mengelompokkan observasi dalam suatu populasi ke dalam K kelompok serupa,

berdasarkan pada meminimalkan perbedaan observasi antar observasi pada setiap cluster.

Bagan angkat. Alat grafis yang banyak digunakan untuk menunjukkan manfaat praktis model prediktif. Sering digunakan bersamaan dengan grafik Keuntungan. Lihat Lampiran A untuk informasi lebih lanjut.

Model linier. Jenis model prediktif populer yang mudah dipahami dan digunakan. Skor dihitung dengan mengalikan nilai setiap karakteristik dengan bobot relevannya dan kemudian menjumlahkan seluruh hasilnya. Cara populer untuk merepresentasikan model linier adalah dalam bentuk kartu skor, seperti yang diperkenalkan pada Bab 5.

Regresi linier. Salah satu metode paling populer untuk membuat model linier dan kartu skor. Perkembangannya sudah ada sejak lebih dari 100 tahun yang lalu. Regresi linier membutuhkan daya komputasi yang jauh lebih kecil dibandingkan metode canggih seperti jaringan saraf. Oleh karena itu, ini dapat digunakan dengan sangat sukses sebagai alat pemilihan/pemodelan data awal, untuk mengurangi ukuran kumpulan data ketika metode seperti jaringan saraf (dalam) digunakan untuk model akhir.

Regresi logistik. Metode yang sangat sederhana dan populer digunakan untuk membuat model klasifikasi. Regresi logistik banyak digunakan untuk membuat model linier dan kartu skor, seperti kartu skor penyakit jantung yang diperkenalkan pada Bab 5.

Pembelajaran mesin. Pembelajaran mesin adalah proses menemukan fitur (pola dalam data). Algoritme pembelajaran mesin berasal dari penelitian kecerdasan buatan dan pengenalan pola. Algoritme yang digunakan untuk melatih jaringan saraf (dalam) dan mendukung mesin vektor adalah dua contoh pendekatan pembelajaran mesin.

Pengurangan Peta. Pendekatan pemrograman yang memungkinkan data yang disimpan di Hadoop (dan platform Big Data lainnya) diproses dengan sangat cepat. MapReduce bekerja dengan membagi tugas pemrosesan data menjadi banyak subtugas yang lebih kecil yang kemudian dapat diimplementasikan secara paralel di seluruh jaringan komputer yang membentuk jaringan Hadoop.

Model. Representasi matematis dari sistem atau situasi dunia nyata. Model ini digunakan untuk menentukan bagaimana sistem dunia nyata akan berperilaku dalam kondisi yang berbeda. Lihat juga model prediktif.

Pemantauan. Performa model prediktif cenderung menurun seiring waktu. Oleh karena itu, adalah bijaksana untuk memulai sistem pemantauan setelah penerapan model untuk mengukur kinerja model secara berkelanjutan. Model dikembangkan kembali ketika pemantauan menunjukkan bahwa telah terjadi penurunan kinerja model yang signifikan.

AI yang sempit. Aplikasi kecerdasan buatan yang sangat baik hanya dalam satu atau dua hal, namun tidak dapat diterapkan di luar permasalahan yang telah dirancang. Semua aplikasi AI yang digunakan saat ini dapat digambarkan sebagai sistem AI Sempit. Lihat juga AI Umum.

Jaringan syaraf. Jenis model prediktif populer yang diturunkan menggunakan pembelajaran mesin. Jaringan saraf sangat cocok untuk menangkap interaksi kompleks dan non-linier

dalam data dengan cara yang analog dengan pembelajaran manusia. Jaringan saraf “dalam” (Pembelajaran mendalam/Jaringan keyakinan mendalam) adalah jaringan saraf yang sangat besar dan kompleks, sering kali berisi ribuan atau jutaan neuron buatan, yang digunakan untuk tugas-tugas “AI” seperti pengenalan suara dan sistem navigasi pada mobil yang dapat mengemudi sendiri.

Saraf. Komponen kunci dari jaringan saraf, yang sering dianalogikan sebagai neuron biologis di otak manusia. Pada kenyataannya, neuron adalah model linier yang skornya kemudian mengalami transformasi (non-linier). Oleh karena itu, jaringan saraf dapat dianggap sebagai sekumpulan model linier dan transformasi non-linier yang saling berhubungan.

Kemungkinan. Suatu cara untuk mewakili kemungkinan suatu peristiwa terjadi. Peluang suatu kejadian sama dengan $(1/p) - 1$ dimana p adalah peluang kejadian tersebut. Demikian pula, probabilitasnya sama dengan $(1/Odds+1)$. Odds 1:1 sama dengan probabilitas 0,5, odds 2:1 probabilitas 0,33, 3:1 probabilitas 0,25 dan seterusnya.

Over-fitting. Kutukan bagi kehidupan seorang data scientist. Over-fitting terjadi ketika suatu algoritma bertindak terlalu jauh dalam mencari korelasi dalam data yang digunakan untuk mengembangkan model prediktif. Hasil akhirnya adalah model tersebut terlihat sangat prediktif ketika diukur terhadap sampel pengembangan, namun kinerjanya sangat buruk ketika digunakan untuk memprediksi hasil baru menggunakan data yang belum pernah disajikan ke model sebelumnya.

Ganti aturan. Terkadang, tindakan tertentu harus diambil terlepas dari prediksi yang dihasilkan oleh suatu model. Misalnya, model prediktif yang digunakan untuk menargetkan orang-orang yang mendapat tawaran bir mungkin memperkirakan bahwa beberapa anak kemungkinan besar akan menerima tawaran tersebut. Oleh karena itu, aturan penggantian diterapkan untuk mencegah penawaran dikirimkan kepada anak-anak, terlepas dari skor yang dihasilkan oleh model.

Analisis prediktif (PA). Penerapan teknik pembelajaran mesin untuk menghasilkan model prediktif. Beberapa orang berpendapat bahwa untuk semua tujuan praktis, pembelajaran mesin dan analisis prediktif kurang lebih sama, mengingat keduanya menggunakan jenis data yang sama sebagai masukan, menerapkan jenis algoritme yang sama, dan menghasilkan keluaran (skor) yang serupa.

Model prediktif. Model prediktif adalah keluaran utama yang dihasilkan oleh sebagian besar algoritme pembelajaran mesin. Model menangkap hubungan (korelasi) yang ditemukan oleh proses. Setelah model prediktif dibuat, model tersebut kemudian dapat diterapkan pada situasi baru untuk memprediksi peristiwa di masa depan, atau peristiwa yang tidak diketahui.

Analisis Komponen Prinsip (PCA). PCA adalah metode reduksi variabel yang populer. PCA menggunakan kumpulan variabel asli untuk membuat kumpulan variabel baru yang lebih kecil, yang kemudian digunakan dalam proses pembelajaran mesin. PCA bekerja berdasarkan premis bahwa banyak variabel berkorelasi satu sama lain sampai tingkat

tertentu; yaitu berisi beberapa informasi yang sama. Oleh karena itu, informasi ini dapat ditangkap secara lebih efisien dengan menggunakan variabel yang lebih sedikit.

Pembuatan profil. Pembuatan profil, sebagaimana didefinisikan oleh Peraturan Perlindungan Data Umum (GDPR), adalah: “segala bentuk pemrosesan otomatis data pribadi yang terdiri dari penggunaan data pribadi untuk mengevaluasi aspek pribadi tertentu yang berkaitan dengan seseorang, khususnya untuk menganalisis atau memprediksi aspek mengenai kinerja orang tersebut di tempat kerja, situasi ekonomi, kesehatan, preferensi pribadi, minat, keandalan, perilaku, lokasi atau pergerakan.”[89] Pembuatan profil mungkin atau mungkin bukan merupakan bagian dari proses pengambilan keputusan otomatis.

Python. Selain R, Python adalah salah satu paket perangkat lunak paling populer dan tersedia secara gratis yang digunakan untuk pembelajaran mesin.

R. Bahasa R adalah salah satu platform perangkat lunak paling populer untuk melakukan pembelajaran mesin. Perangkat lunak R gratis dan bersumber terbuka, dengan komunitas pengguna dapat mengembangkan fungsionalitas baru dan membaginya dengan pengguna lain. Bagi mereka yang ingin menjadi data scientist, disarankan untuk menguasai R (dan/atau Python).

Hutan acak. Hutan acak adalah metode ansambel yang didasarkan pada penggabungan keluaran sejumlah besar pohon keputusan. Setiap pohon keputusan dibuat dalam kondisi yang sedikit berbeda dan karenanya, menghasilkan skor yang berbeda. Hutan acak adalah salah satu metode ansambel yang paling sukses dan banyak diterapkan.

Jaringan saraf berulang. Jenis jaringan saraf yang menggunakan keluaran dari kasus sebelumnya dalam sampel pengembangan sebagai masukan untuk pelatihan berikutnya. Jenis jaringan ini sangat berguna jika terdapat pola berurutan dalam data; yaitu urutan kasus dalam sampel pengembangan itu penting. Salah satu contohnya adalah analisis video. Perubahan keadaan dari satu gambar ke gambar berikutnya (yaitu saat objek bergerak) dapat memberikan banyak informasi tentang sifat gambar saat ini.

Model regresi. Alat populer untuk memprediksi besarnya sesuatu. Misalnya, berapa banyak uang yang akan dibelanjakan seseorang atau berapa lama ia akan hidup. Hal ini berbeda dengan model klasifikasi yang memprediksi kemungkinan terjadinya suatu peristiwa.

Pembelajaran penguatan. Proses dimana bobot dalam model disempurnakan berdasarkan kasus per kasus, berdasarkan beberapa ukuran keberhasilan atau imbalan yang dihasilkan dari tindakan yang diambil. Setiap kali model menghasilkan suatu hasil, kualitas hasil tersebut dinilai. Algoritme pelatihan kemudian menyesuaikan bobot model bergantung pada seberapa baik kinerjanya. Pembelajaran penguatan dipandang lebih mirip dengan cara orang belajar dibandingkan jenis pembelajaran mesin lainnya; yaitu pembelajaran yang diawasi.

Pemodelan respon (pilihan). Istilah ini mengacu pada model pemasaran yang digunakan untuk memprediksi kemungkinan seseorang membeli produk atau layanan yang menjadi sasarannya. Model respons adalah jenis model klasifikasi.

- Skor.** Kebanyakan prediksi yang dihasilkan oleh model prediktif direpresentasikan dalam bentuk satu angka (skor). Untuk model klasifikasi, skor adalah representasi probabilitas suatu peristiwa terjadi, misalnya. seberapa besar kemungkinan seseorang menanggapi komunikasi pemasaran atau kemungkinan mereka gagal membayar pinjaman. Untuk model regresi, skor mewakili besarnya perilaku yang diprediksi, misalnya. berapa banyak yang mungkin dibelanjakan seseorang, atau berapa lama mereka bisa hidup.
- Kartu skor / catatan angka.** Kartu skor adalah cara menyajikan model linier yang mudah dipahami oleh non-ahli. Manfaat utama dari kartu skor adalah bersifat tambahan; yaitu skor model dihitung hanya dengan menjumlahkan poin yang berlaku. Tidak ada perkalian, pembagian atau aritmatika lain yang lebih kompleks.
- Distribusi skor.** Tabel atau grafik yang menunjukkan bagaimana skor dari model prediktif didistribusikan ke seluruh populasi yang diteliti. Bagan Peningkatan dan Perolehan merupakan cara menyajikan distribusi skor dalam bentuk grafik.
- Analisis sentimen.** Ini adalah teknik populer untuk menggali informasi tentang sikap masyarakat terhadap berbagai hal. Misalnya, apakah mereka mempunyai pengalaman positif atau negatif saat menggunakan produk atau layanan tertentu. Dalam pembelajaran mesin, analisis sentimen digunakan untuk mengekstrak informasi dari teks atau ucapan yang kemudian digunakan untuk membangun model prediktif atau memperoleh cluster.
- Pembelajaran yang diawasi.** Penerapan pembelajaran mesin di mana setiap kasus dalam sampel pengembangan memiliki hasil terkait yang ingin diprediksi. Kasus-kasus tersebut dikatakan "Berlabel." Contoh pembelajaran yang diawasi dalam pemasaran sasaran adalah ketika respons setiap pelanggan terhadap aktivitas pemasaran diketahui (mereka merespons atau tidak). Model yang dihasilkan oleh algoritma tersebut kemudian dioptimalkan untuk memprediksi apakah pelanggan akan merespons pemasaran atau tidak. Dalam praktiknya, sebagian besar pendekatan pembelajaran mesin adalah contoh pembelajaran yang diawasi.
- Mendukung mesin vektor.** Jenis model non-linier tingkat lanjut. Mesin vektor dukungan memiliki beberapa kesamaan dengan jaringan saraf.
- Aliran Tensor.** Pustaka sumber terbuka berisi algoritme pembelajaran mesin untuk digunakan dengan bahasa pemrograman Python. TensorFlow awalnya dikembangkan oleh Google. Fitur utama TensorFlow adalah memanfaatkan kemampuan pemrosesan kartu grafis kelas atas untuk secara signifikan meningkatkan kecepatan pengembangan model pembelajaran mesin yang kompleks (khususnya jaringan neural dalam).
- Algoritma pelatihan.** Istilah yang digunakan untuk menggambarkan algoritma pembelajaran mesin berulang yang digunakan untuk menentukan struktur model prediktif. Istilah ini paling banyak digunakan dalam kaitannya dengan model tipe jaringan saraf. Algoritme ini berulang beberapa kali, menyesuaikan bobot dalam jaringan untuk mengoptimalkan kinerja jaringan. Algoritme pelatihan berakhir setelah sejumlah iterasi tetap atau ketika tidak ada peningkatan signifikan lebih lanjut dalam performa model yang diperoleh.

Pembelajaran tanpa pengawasan. Penerapan pembelajaran mesin pada masalah yang sampel pengembangannya tidak berisi data hasil. Kasus dikatakan “Tidak Berlabel”. Algoritme tanpa pengawasan biasanya berupaya mengelompokkan kasus-kasus dengan karakteristik (fitur) yang serupa. Contoh pembelajaran tanpa pengawasan adalah organisasi yang ingin membuat kebijakan penempatan iklan untuk produk mewah yang mahal, namun tidak ada informasi tentang riwayat pembelian pelanggan. Pengelompokan diterapkan untuk mengelompokkan pelanggan serupa berdasarkan usia, pendapatan, jenis kelamin, dll. Strategi penempatan iklan kemudian ditargetkan pada individu dalam kelompok dengan pendapatan rata-rata tinggi, daripada kelompok dengan pendapatan lebih rendah.

Sampel validasi. Kumpulan data independen yang digunakan untuk mengevaluasi model prediktif setelah model tersebut dibuat. Sampel validasi harus benar-benar terpisah dari sampel pengembangan dan tidak boleh digunakan selama konstruksi model. Menggunakan satu (atau lebih) sampel validasi adalah hal yang penting karena pembelajaran mesin terkadang terlalu menyesuaikan model dengan sampel pengembangan. Artinya, jika Anda mengevaluasi model prediktif menggunakan sampel pengembangan, model tersebut mungkin tampak lebih prediktif daripada yang sebenarnya.

DAFTAR PUSTAKA

- Anggraini, K., & Sunarsih, D. (2022). Digital marketing. Pascal Books.
- Bagdare, S., & Jain, R. (2013). Measuring retail customer experience. *International Journal of Retail & Distribution Management*, 41(10), 790–804.
- Belayeh, W. G., Mussa, Y. O., & Gizaw, A. K. (2020). Approximate analytic solutions of twodimensional nonlinear Klein–Gordon equation by using the reduced differential transform method. *Mathematical Problems in Engineering*.
- Boujrad, M., & Iamlili, Y. el M. N. (2021). A New Artificial Intelligence-Based Strategy for Digital Marketing Reinforcement. In *Lecture Notes in Networks and Systems* (Vol. 183, pp. 689–699).
- Gao, F., & Zhang, L. (2020). Application of Artificial Intelligence and Big Data Technology in Digital Marketing. In *ACM International Conference Proceeding Series*.
- Guru, K., Raja, S., Sasiganth, J., Sharma, D. K., Tiwari, M., & Tiwari, T. (2023). The Future Impact of Technological Developments on Digital Marketing Through Artificial Intelligence. In *Smart Innovation, Systems and Technologies* (Vol. 290, pp. 217–225).
- Haenlein, M., Kaplan, A., Tan, C.-W., & Zhang, P. (2019). Artificial intelligence (AI) and management analytics. *Journal of Management Analytics*, 6(4), 341–343.
- Harden, A., & Thomas, J. (2005). Methodological issues in combining diverse study types in systematic reviews. *International Journal of Social Research Methodology: Theory and Practice*, 8(3), 257–271.
- Hassan, A. (2021). The usage of artificial intelligence in digital marketing: A review. *Studies in Computational Intelligence*, 954, 357–383.
- Hsu, S. F., & Liou, S. (2021). Artificial Intelligence Impact on Digital Content Marketing Research. In *2021 9th International Conference on Orange Technology, ICOT 2021*.
- Huang, M., & Rust, R. T. (2018). Artificial Intelligence in Service. *Journal of Service Research*.
- Karn, A. (2021). Artificial Intelligence in Computer Vision. *International Journal of Engineering Applied Sciences and Technology (IJEAST)*, 249–254.
- Lai, Z., & Yu, L. (2021). Research on Digital Marketing Communication Talent Cultivation in the Era of Artificial Intelligence. In *Journal of Physics: Conference Series* (Vol. 1757, Issue 1).
- Mahesh, B. (2018). Machine Learning Algorithms - A Review. *International Journal of Science and Research (IJSR)*, 9(1), 381–386.
- Ribeiro, T., & Reis, J. L. (2020). Artificial intelligence applied to digital marketing. In *Advances in Intelligent Systems and Computing: Vol. 1160 AISC* (pp. 158–169).

- Shahid, M. Z., & Li, G. (2019). Impact of Artificial Intelligence in Marketing: A Perspective of Marketing Professionals of Pakistan. *Global Journal of Management and Business Research: E-Marketing*, 19(2), 3–4.
- Siregar, H., dkk. (2020). Isu Proses Bisnis Berbasis Artificial Intelligence untuk Menyongsong Era Industri 4.0. *Jurnal Bisnis Strategi*, 29(2), 89-100.
- Sterne, J. (2017). *Artificial intelligence for marketing: practical applications*. John Wiley & Sons.
- Verma, S., Sharma, R., Deb, S., & Maitra, D. (2020). Artificial Intelligence in Marketing: Systematic Review and Future Research Direction. *International Journal of Information Management Data Insights*.

APLIKASI AI DAN MACHINE LEARNING DALAM BISNIS

oleh:
Dr. Joseph Teguh Santoso, S.Kom, M.Kom

BIODATA PENULIS



Dr. Joseph Teguh Santoso, S.Kom, M.Kom adalah Rektor dari Universitas Sains & Teknologi Komputer (Universitas STEKOM) Semarang yang memiliki banyak pengalaman praktis dalam bidang *e-commerce* sejak Tahun 2002. Beliau mempunyai 3 (tiga) toko *Official Online Store* di China untuk merek Sepeda Raleigh, dengan omzet tahunan pada Tahun 2019 mencapai lebih dari Rp. 35 Milyar rupiah dan terus meningkat. Dr. Joseph T.S memiliki lisensi tunggal sepeda merek “Raleigh” untuk penjualan *Online* di seluruh China. Di samping itu beliau juga memiliki pabrik sepeda dan sepeda listrik merek “Fengjiu”, yaitu Pabrik Sepeda Listrik yang masih tergolong kecil di China. Pengalaman beliau malang melintang di dunia *online store* di China seperti Alibaba, Tmall, Taobao, JD, Aliexpress sangat membantu mahasiswa untuk memiliki pengalaman teknis dan praktis untuk membuka toko *online* bersama beliau.



YAYASAN PRIMA AGUS TEKNIK

PENERBIT :
YAYASAN PRIMA AGUS TEKNIK
Jl. Majapahit No. 605 Semarang
Telp. (024) 6723456. Fax. 024-6710144
Email : penerbit_ypat@stekom.ac.id

ISBN 978-623-8120-64-2 (PDF)

