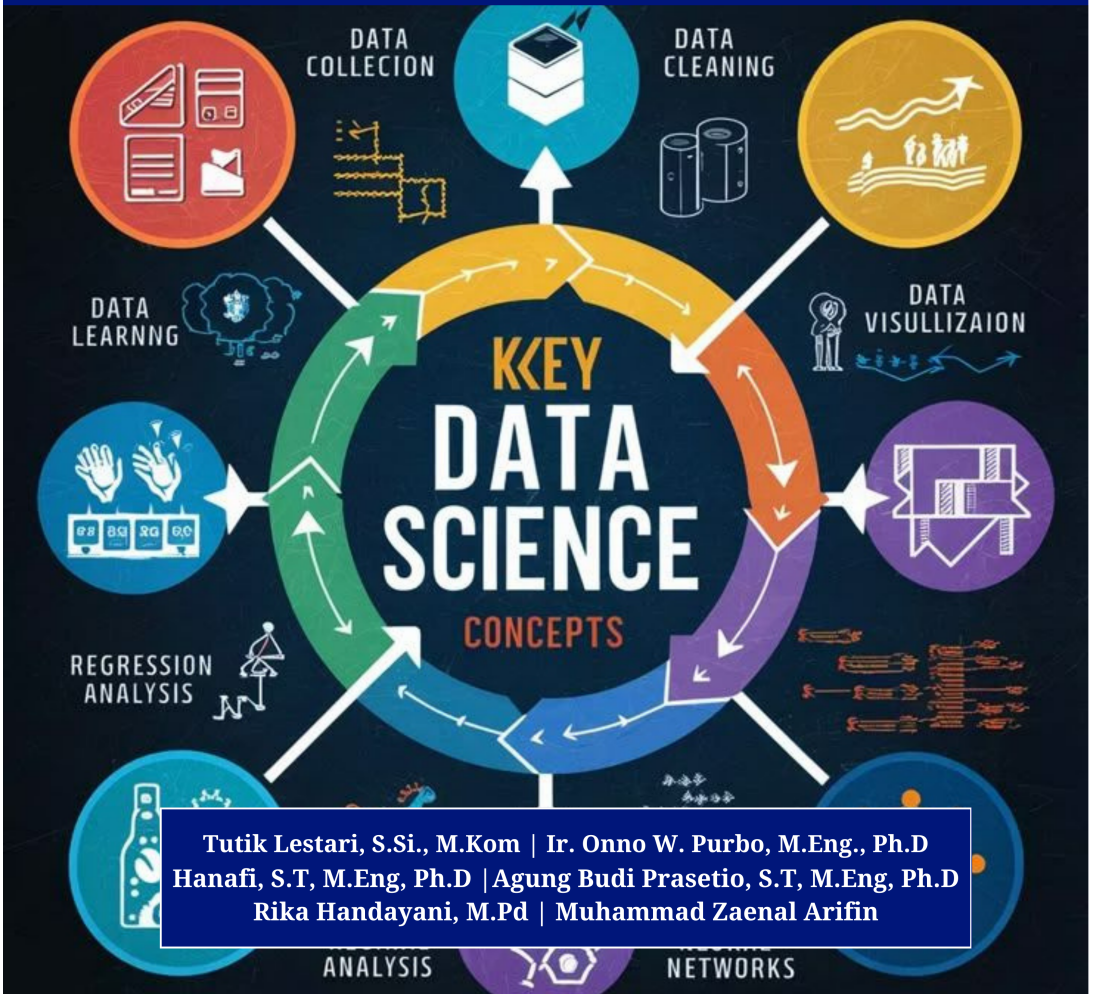
Rekomendasi Bagi Mahasiswa Modern

DATA SCIENCE

For Beginners

Where Data Meet Intelligence



Tutik Lestari, S.Si., M.Kom | Ir. Onno W. Purbo, M.Eng., Ph.D
Hanafi, S.T, M.Eng, Ph.D |Agung Budi Prasetio, S.T, M.Eng, Ph.D
Rika Handayani, M.Pd | Muhammad Zaenal Arifin

UNDANG-UNDANG REPUBLIK INDONESIA
NOMOR 28 TAHUN 2014 TENTANG HAK CIPTA

LINGKUP HAK CIPTA
Pasal 1

1. Hak Cipta adalah hak eksklusif pencipta yang timbul secara otomatis berdasarkan prinsip deklaratif setelah suatu ciptaan diwujudkan dalam bentuk nyata tanpa mengurangi pembatasan sesuai dengan ketentuan peraturan perundang-undangan.

KETENTUAN PIDANA
Pasal 113

1. Setiap Orang yang dengan tanpa hak melakukan pelanggaran hak ekonomi sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf i untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 1 (satu) tahun dan/atau pidana denda paling banyak Rp 100.000.000 (seratus juta rupiah).
2. Setiap Orang yang dengan tanpa hak dan/atau tanpa izin Pencipta atau pemegang Hak Cipta melakukan pelanggaran hak ekonomi Pencipta sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf c, huruf d, huruf f, dan/atau huruf h untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 3 (tiga) tahun dan/atau pidana denda paling banyak Rp500.000.000,00 (lima ratus juta rupiah).
3. Setiap Orang yang dengan tanpa hak dan/atau tanpa izin Pencipta atau pemegang Hak Cipta melakukan pelanggaran hak ekonomi Pencipta sebagaimana dimaksud dalam Pasal 9 ayat (l) huruf a, huruf b, huruf e, dan/atau huruf g untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 4 (empat) tahun dan/ atau pidana denda paling banyak Rp1.000.000.000,00 (satu miliar rupiah).
4. Setiap Orang yang memenuhi unsur sebagaimana dimaksud pada ayat (3) yang dilakukan dalam bentuk pembajakan, dipidana dengan pidana penjara paling lama 10 (sepuluh) tahun dan/atau pidana denda paling banyak Rp4.000.000.000,00 (empat miliar rupiah).

DATA SCIENCE FOR BEGINNERS

ISBN: 978-634-04-2976-3

Penulis:
Tutik Lestari, S.Si., M.Kom - Ir. Onno W. Purbo, M.Eng., Ph.D
Hanafi, S.T., M.Eng., Ph.D. - Agung Budi Prasetio, S.T, M.Eng, Ph.D
Rika Handayani, M.Pd - Muhammad Zaenal Arifin

Editor:
Tutik Lestari, S.Si., M.Kom
Falah Ibrahim, M.Comm

Layout
Falah Ibrahim, M.Comm

Desain Sampul:
Tutik Lestari, S.Si., M.Kom

Cetakan I, Juli 2025

Penerbit
UDN PRESS
No. Anggota IKAPI : 633/Anggota Luar Biasa/DKI/2024
Jln. Ciledug Raya No. 01. Ulujami Raya, Pesanggrahan
Jakarta Selatan, Provinsi DKI Jakarta
Website: www.press.darunnajah.ac.id
Instagram: @udn_press

KATA PENGANTAR

Puji syukur kami panjatkan ke hadirat Allah SWT, karena berkat rahmat, taufik, dan karunia-Nya, buku berjudul "**DATA SCIENCE FOR BEGINNERS**" ini dapat hadir di tengah para pembaca. Buku ini disusun dengan tujuan memberikan pengenalan yang komprehensif mengenai data science, sebuah bidang yang saat ini menjadi salah satu kompetensi paling dibutuhkan di era digital dan revolusi industri 4.0.

Data science tidak lagi hanya menjadi ranah para akademisi atau peneliti, melainkan juga telah menjadi keterampilan praktis yang dibutuhkan oleh berbagai sektor, mulai dari pendidikan, bisnis, industri, hingga pemerintahan. Melalui buku ini, penulis berusaha menghadirkan materi yang sistematis, ringkas, dan mudah dipahami, sehingga dapat menjadi pegangan awal bagi siapa saja yang tertarik mempelajari dasar-dasar data science.

Buku ini lahir dari kolaborasi berbagai penulis yang memiliki latar belakang dan keahlian di bidang ilmu komputer, teknik, serta pendidikan. Kehadiran karya ini diharapkan dapat membantu mahasiswa, dosen, praktisi, maupun masyarakat umum yang ingin memahami konsep dasar data science serta aplikasinya dalam kehidupan nyata.

Kami menyadari bahwa penyusunan buku ini masih jauh dari kesempurnaan. Oleh karena itu, kritik dan saran yang membangun sangat kami harapkan untuk penyempurnaan edisi-edisi berikutnya.

Akhir kata, kami mengucapkan terima kasih kepada semua pihak yang telah membantu, baik secara langsung maupun tidak langsung, dalam proses penyusunan hingga terbitnya buku ini. Semoga buku ini dapat memberikan manfaat dan menjadi salah satu kontribusi kecil dalam pengembangan ilmu pengetahuan, khususnya di bidang data science.

Jakarta, Agustus 2025

Tutik Lestari, S.Si, M.Kom

PRAKATA

Puji syukur kami panjatkan kehadirat Allah Subhanahu Wata'ala atas rahmat dan karunia-Nya yang tiada henti, sehingga Buku **DATA SCIENCE FOR BEGINNERS** ini dapat diselesaikan. Buku ini disusun sebagai panduan praktis bagi mahasiswa Program Studi Sains dan Teknologi, khususnya bagi mereka yang baru memulai perjalanan di dunia Data Science.

Buku ini dirancang dengan tujuan untuk memberikan pemahaman dasar mengenai Data Science dengan cara yang mudah dipahami, mengingat betapa pentingnya bidang ini di dunia teknologi dan ilmu pengetahuan saat ini. Buku ini akan membimbing pembaca melalui konsep-konsep dasar, teknik-teknik analisis data, serta penggunaan berbagai alat dan perangkat lunak yang sering digunakan dalam analisis data.

Kami mengucapkan terima kasih kepada seluruh pihak yang telah berkontribusi dalam penyusunan buku ini, mulai dari dosen, mahasiswa, hingga tim pengelola yang mendukung jalannya penyusunan. Terimakasih juga kami sampaikan kepada Universitas Darunnajah Jakarta yang telah memberikan kesempatan untuk menyusun buku ini sebagai referensi pembelajaran bagi mahasiswa.

Tak lupa, kami juga ingin mengucapkan terima kasih kepada para editor, penyunting, dan semua pihak yang membantu dalam proses penyelesaian buku ini. Semoga buku ini dapat bermanfaat sebagai sumber pengetahuan yang berguna, serta sebagai langkah awal dalam perjalanan para mahasiswa untuk lebih mendalami dunia Data Science. Kami menyadari bahwa buku ini masih memiliki banyak kekurangan, oleh karena itu, kami dengan senang hati menerima kritik dan saran yang membangun demi perbaikan buku ini di masa depan.

Jakarta, 2025

Dosen Universitas Darunnajah Jakarta
Tutik Lestari, S.Si, M.Kom

DAFTAR ISI

BAB 1
DATA SCIENCE

A. Tujuan Pembelajaran

Tujuan belajar Data Science adalah untuk memperoleh pemahaman dan keterampilan yang diperlukan dalam mengumpulkan, mengolah, menganalisis, dan menginterpretasi data guna membuat keputusan yang lebih baik dan berbasis bukti. Secara lebih rinci, tujuan belajar Data Science meliputi beberapa hal berikut:

1. **Menganalisis Data Secara Efektif**
   Mengembangkan kemampuan untuk memahami dan menganalisis data dalam jumlah besar (big data) menggunakan alat dan teknik statistik, matematis, serta algoritma komputasi.
2. **Menemukan Pola dan Insight dari Data**
   Kemampuan untuk menemukan pola, tren, atau insight dari data yang dapat membantu dalam pemecahan masalah atau perencanaan strategis.
3. **Meningkatkan Pengambilan Keputusan**
   Menggunakan hasil analisis data untuk mendukung keputusan yang lebih akurat, efisien, dan berbasis data, yang berdampak pada peningkatan kinerja dan produktivitas.
4. **Menguasai Teknologi dan Alat Data Science**
   Mempelajari berbagai alat dan perangkat lunak yang digunakan dalam Data Science, seperti Python, R, SQL, dan alat machine learning seperti TensorFlow, serta software visualisasi data seperti Tableau dan Power BI.
5. **Membangun Model Prediksi**
   Mengembangkan keterampilan dalam membangun model statistik dan machine learning untuk memprediksi berbagai variabel atau hasil di masa depan berdasarkan data yang ada.

6. **Memecahkan Masalah Dunia Nyata**
   Mengaplikasikan konsep Data Science untuk memecahkan masalah di berbagai bidang, seperti bisnis, kesehatan, ilmu pengetahuan, pendidikan, dan sektor lainnya, dengan solusi yang berbasis data.
7. **Pengolahan dan Manajemen Data**
   Mengembangkan keterampilan dalam mengelola data yang kompleks, termasuk membersihkan dan mempersiapkan data untuk analisis, serta menjaga kualitas dan keamanan data.
8. **Meningkatkan Kompetensi dalam Komunikasi Data**

   Mampu menyampaikan hasil analisis data dengan cara yang mudah dipahami oleh pemangku kepentingan, menggunakan visualisasi data yang tepat dan laporan yang jelas.

Dengan belajar Data Science, mahasiswa diharapkan dapat menguasai kompetensi-kompetensi tersebut, yang sangat dibutuhkan dalam dunia industri dan penelitian saat ini, yang semakin bergantung pada data untuk mengambil keputusan yang lebih baik dan terinformasi.

B. Pendahuluan

Data Science telah menjadi disiplin ilmu yang sangat penting dalam era digital saat ini, di mana volume data yang dihasilkan setiap hari terus meningkat secara eksponensial. Menurut Joel Grus dalam bukunya Data Science (Grus, Data Science from Scratch: First Principles with Python (2nd ed.)., 2020) memahami prinsip-prinsip dasar Data Science sangat penting untuk dapat mengolah dan menganalisis data secara efektif.

Grus menekankan pentingnya pemahaman mendalam tentang algoritma dan teknik dasar sebelum beralih ke alat dan pustaka yang lebih. Sebagai tambahan, buku Python Data Science Handbook oleh (VanderPlas, Python Data Science Handbook: Essential Tools for Working with Data. , 2021) memberikan panduan komprehensif mengenai penggunaan Python dalam Data

Science, mencakup berbagai pustaka penting seperti NumPy, Pandas, Matplotlib, dan scikit-learn. VanderPlas menyajikan contoh-contoh praktis yang memudahkan pembaca untuk memahami konsep-konsep yang diterapkan dalam analisis data. Pendekatan praktis ini sangat membantu bagi pemula yang ingin menguasai alat-alat dasar dalam Data Science.

Selain itu, buku Data Science for Beginners oleh (Park, 2020) menyajikan materi yang dirancang khusus untuk pemula, mencakup topik-topik seperti pemrograman Python, analisis data, dan pengenalan machine learning. Park menyusun materi secara bertahap dengan tutorial langkah demi langkah, memungkinkan pembaca untuk membangun pemahaman yang solid dari dasar hingga konsep yang lebih kompleks.

Mengintegrasikan konsep-konsep dari ketiga buku tersebut, pembaca dapat memperoleh pemahaman yang menyeluruh mengenai Data Science, mulai dari dasar hingga aplikasi praktis menggunakan Python. Pendekatan yang sistematis dan berbasis praktik ini diharapkan dapat memfasilitasi pemula dalam memulai perjalanan mereka di dunia Data Science.

C. Apa sih Data Science itu ?

Data Science, apa itu ? Jenis makhluk apa ya, hehe untuk lebih detailnya yuk kita pelajari bareng-bareng Data Science lebih dalam lagi.

Data Science adalah bidang interdisiplin yang menggunakan metode, proses, algoritma dan sistem ilmiah untuk mengekstraksi pengetahuan dan insights dari data dalam berbagai bentuk, baik terstruktur maupun tidak terstruktur, mirip dengan data mining.

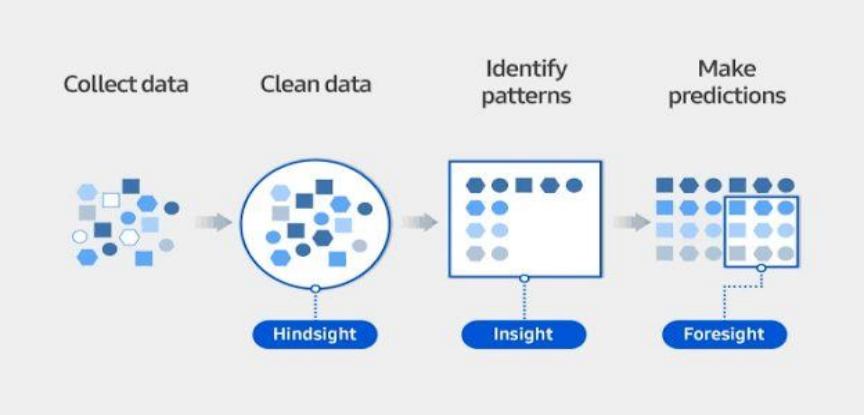
Data science adalah **"konsep untuk menyatukan statistik, analisis data, pembelajaran mesin dan metode terkait"** untuk **"memahami dan menganalisis fenomena aktual"** dengan data. Ini menggunakan teknik dan teori yang diambil dari banyak bidang dalam konteks matematika, statistik, information science, dan ilmu komputer.

BAB 2
MENGULIK DATA SCIENCE LEBIH JAUH

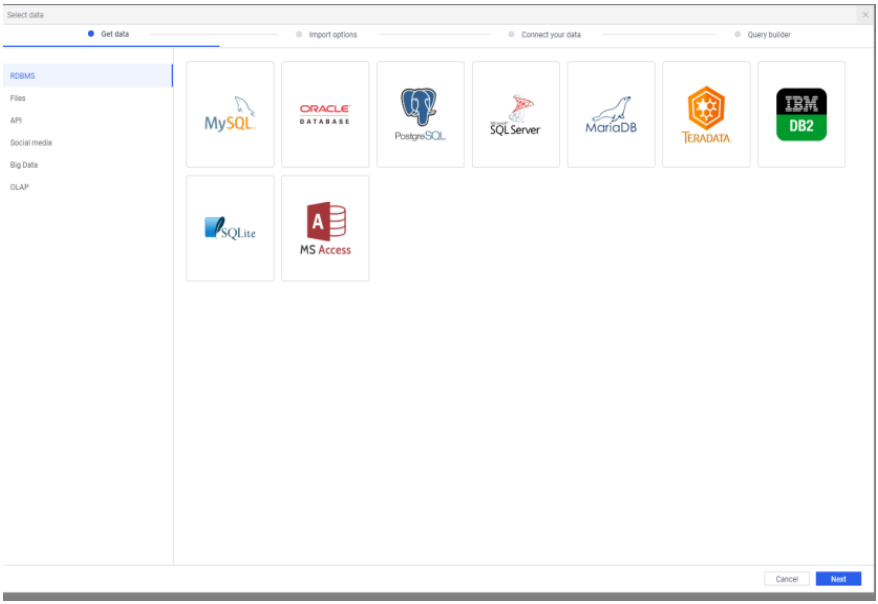
A. Tujuan Pembelajaran

Data science atau ilmu data telah menjadi salah satu bidang yang berkembang pesat dalam beberapa tahun terakhir. Dengan kemajuan teknologi informasi, akses terhadap data yang sangat besar dan beragam, serta kemampuan untuk menganalisisnya dengan menggunakan berbagai teknik statistik dan komputasi, data science kini menjadi kunci penting dalam pengambilan keputusan di berbagai sektor. Bidang ini tidak hanya terbatas pada analisis data numerik, tetapi juga mencakup pemrosesan data tidak terstruktur, seperti teks dan gambar, untuk menghasilkan wawasan yang bernilai. Artikel ini akan membahas lebih lanjut mengenai konsep dasar, metode, serta aplikasi nyata dari data science dalam kehidupan sehari-hari, dan mengapa keahlian dalam data science semakin dibutuhkan di dunia modern.

B. Data Science: Data Skill



Gambar 1. Penjabaran manfaat Data Science ke semua Scope



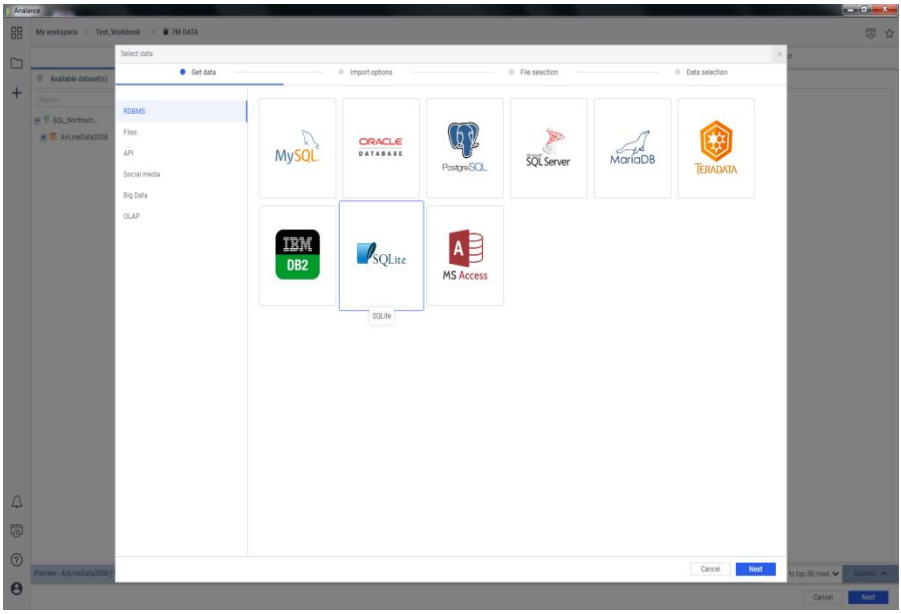
Gambar 2. Data Science Tools

Analance memiliki lebih dari 20 konektor data. Menghubungkan ke berbagai sumber data untuk menyatukan semua data ke dalam satu platform dapat dilakukan dalam hitungan menit.

C. Mengapa Persiapan Data Diperlukan?

Setelah Anda mengumpulkan bahan untuk adonan kue Anda, langkah selanjutnya adalah mencari tahu resep apa yang akan Anda ikuti untuk membuat kue itu dipanggang.

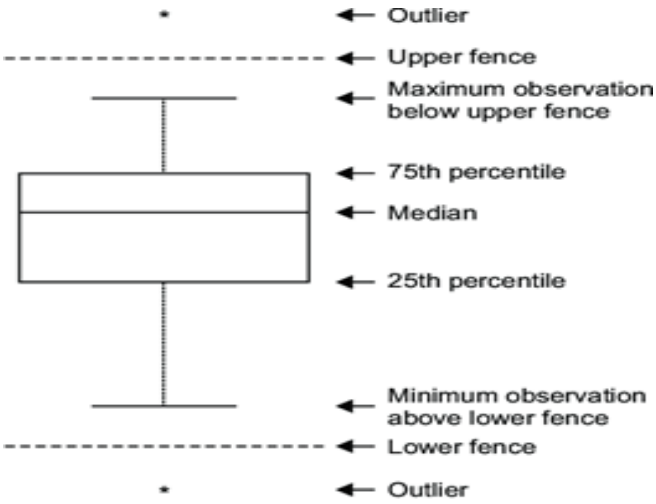
Ada beberapa hal yang harus Anda pertimbangkan jika Anda ingin memanggang Anda berjalan dengan lancar: Apakah adonan dalam bentuk yang tepat untuk Anda gunakan? Apakah Anda membeli tepung yang tepat? Apakah tanggal pembuatan dan kedaluwarsa masih dalam batas yang dapat diterima? Apakah ada cukup dalam satu kotak untuk jumlah cookie yang perlu Anda buat?



Gambar 3. Data Science Tools for ML

Alat perantara yang kuat, SQLite, di Analance dapat membantu pengguna dengan persiapan data. Setelah terhubung ke sumber data, semua informasi yang berasal darinya disimpan dalam kumpulan database SQLite yang kuat dan aman untuk digunakan dalam analisis.

D. Data Cleaning Techniques



Gambar 4. Data Cleaning Tehniques

Jadi sekarang Anda sudah mendapatkan adonan kue dan tampaknya dalam bentuk yang benar — bagaimana selanjutnya? Nah, Anda perlu memastikan bahwa itu hanya adonan dan tidak ada pemalsuan. Anda ingin menggigit chocolate chip, bukan gumpalan tepung yang rapat, bukan?

Dalam statistik dan analitik, langkah ini disebut pembersihan data, sebuah proses yang cenderung diabaikan oleh banyak analis data pemula. Metode pembersihan data melibatkan penyisihan anomali dan nilai yang tidak perlu untuk memastikan bahwa Anda bekerja dengan kumpulan data yang tidak tercemar.

Anomali datang dalam berbagai bentuk tetapi kebanyakan dari mereka termasuk dalam keluarga pencilan, nilai yang berada di luar distribusi normal data. Ada dua jenis: pencilan atas dan pencilan bawah.

Jadi, bagaimana tepatnya Anda menentukan nilai mana yang merupakan pencilan dalam kumpulan data Anda? Pertama, Anda

perlu mengidentifikasi rentang antar kuartil (IQR). Ini adalah perbedaan antara kuartal pertama dan kuartal terakhir dari kumpulan data. Sekarang tambahkan dan kurangi 1,5 kali deviasi standar (SD) dari nilainya untuk mendapatkan pagar yang ditunjukkan pada gambar di atas.

Sekarang Anda akan memiliki jangkauan. Apa pun yang berada di luar rentang adalah pencilan. Lebih khusus lagi, apa pun yang berada di bawah IQR - 1,5 * (SD) adalah pencilan yang lebih rendah, dan apa pun yang berada di atas IQR + 1,5 * (SD) adalah pencilan atas.

E. Identify Pattern

Pola adalah segalanya di dunia digital ini. Suatu pola dapat dilihat secara fisik atau dapat diamati secara matematis dengan menerapkan algoritma.
Contoh: Warna pada pakaian, pola bicara, dll. Dalam ilmu komputer, sebuah pola direpresentasikan dengan menggunakan nilai fitur vektor.

Apa itu Pengenalan Pola?

Pengenalan pola adalah proses mengenali pola dengan menggunakan algoritma pembelajaran mesin. Pengenalan pola dapat didefinisikan sebagai klasifikasi data berdasarkan pengetahuan yang telah diperoleh atau informasi statistik yang diambil dari pola dan / atau representasi mereka. Salah satu aspek penting dari pengenalan pola adalah potensi aplikasinya.
Contoh: Pengenalan ucapan, identifikasi pembicara, pengenalan dokumen multimedia (MDR), diagnosis medis otomatis.
Dalam aplikasi pengenalan pola yang khas, data mentah diproses dan diubah menjadi bentuk yang dapat digunakan oleh mesin. Pengenalan pola melibatkan klasifikasi dan kelompok pola.

Dalam klasifikasi, label kelas yang sesuai diberikan ke pola berdasarkan abstraksi yang dihasilkan menggunakan sekumpulan pola pelatihan atau pengetahuan domain. Klasifikasi digunakan dalam pembelajaran terbimbing.

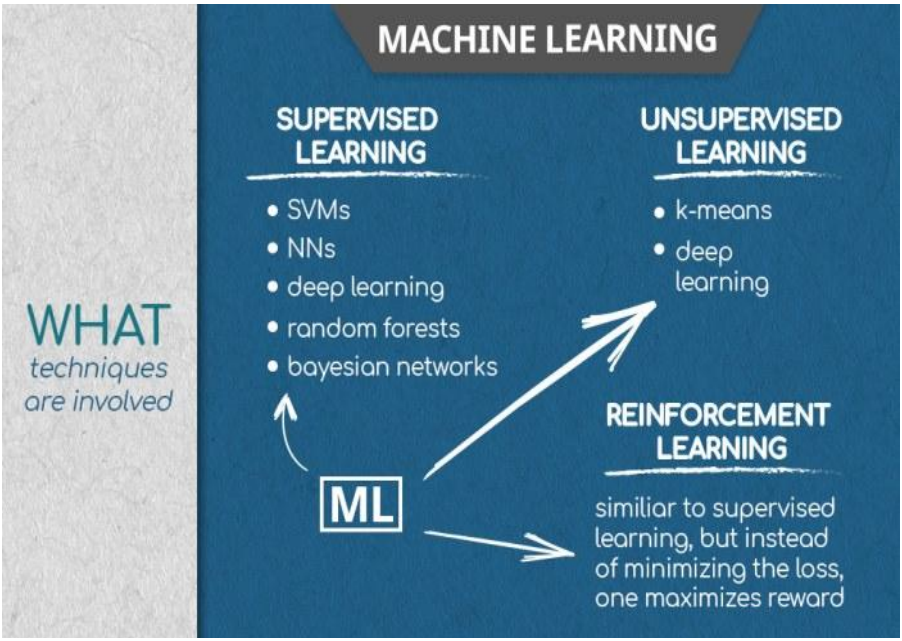


Figure 5. Objektivitas Machine Learning

F. Supervised Learning

Pembelajaran yang diawasi bersandar pada penggunaan data berlabel. Mesin mendapatkan data yang terkait dengan jawaban yang benar; jika kinerja mesin tidak mendapatkan jawaban yang benar, algoritme pengoptimalan menyesuaikan proses komputasi, dan komputer melakukan uji coba lagi. Ingatlah bahwa, biasanya, mesin melakukan ini pada 1000 titik data sekaligus.

Mendukung mesin vektor, jaringan neural, pembelajaran mendalam, model hutan acak, dan jaringan Bayesian adalah contoh pembelajaran yang diawasi.

G. Unsupervised Learning

Ketika datanya terlalu besar, atau data scientist berada di bawah tekanan yang terlalu besar bagi sumber daya untuk melabeli data, atau mereka sama sekali tidak tahu apa labelnya, ilmu data

terpaksa menggunakan pembelajaran tanpa pengawasan. Ini terdiri dari memberikan data tanpa label pada mesin dan memintanya untuk mengekstrak wawasan darinya. Hal ini sering mengakibatkan data dibagi dengan cara tertentu sesuai dengan propertinya. Dengan kata lain, itu dikelompokkan.

Pembelajaran tanpa pengawasan sangat efektif untuk menemukan pola dalam data, terutama hal-hal yang akan terlewatkan oleh manusia yang menggunakan teknik analisis tradisional.

Ilmu data sering kali menggunakan pembelajaran yang diawasi dan tidak diawasi bersama-sama, dengan pembelajaran tanpa pengawasan yang memberi label pada data, dan pembelajaran yang diawasi menemukan model terbaik yang sesuai dengan data. Salah satu contohnya adalah pembelajaran semi-supervisi.

Pembelajaran penguatan

Ini adalah jenis pembelajaran mesin yang berfokus pada kinerja (berjalan, melihat, membaca), bukan keakuratan. Setiap kali mesin berkinerja lebih baik dari sebelumnya, ia menerima hadiah, tetapi jika bekerja secara sub-optimal, algoritme pengoptimalan tidak menyesuaikan penghitungan. Pikirkan perintah belajar anak anjing. Jika mengikuti perintah, ia mendapat hadiah; jika tidak mengikuti perintah, hadiah tidak akan datang. Karena camilannya enak, anjing akan berangsur-angsur membaik dengan mengikuti perintah. Artinya, alih-alih meminimalkan kesalahan, pembelajaran penguatan memaksimalkan hadiah.

H. Contoh-contoh Machine Learning dalam Ilmu Data & Bisnis

Deteksi penipuan

Dengan pembelajaran mesin, pembelajaran yang diawasi secara khusus, bank dapat mengambil data masa lalu, memberi label transaksi sebagai sah, atau curang, dan melatih model untuk mendeteksi aktivitas penipuan. Ketika model ini mendeteksi kemungkinan pencurian sekecil apa pun, mereka menandai transaksi, dan mencegah penipuan secara real time.

BAB 3
Implementasi Data Science Tanpa Programming

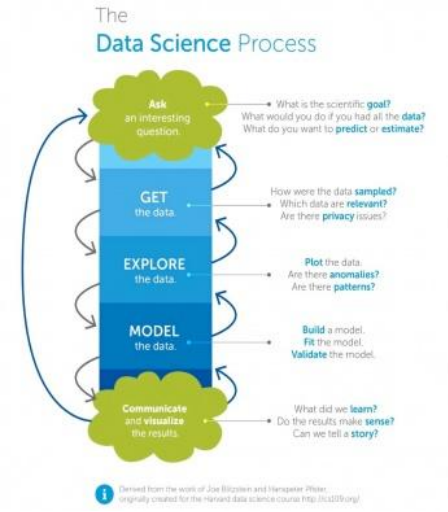
A. Tujuan Pembelajaran

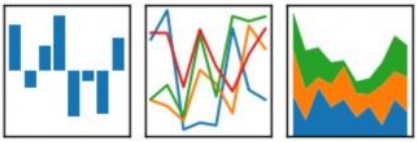
Tujuan utama dari pembelajaran ini adalah untuk memberikan pemahaman kepada peserta mengenai konsep dasar data science dan bagaimana cara mengimplementasikannya tanpa memerlukan keterampilan pemrograman yang mendalam. Dalam era digital saat ini, akses terhadap alat-alat analisis data yang user-friendly memungkinkan siapa saja, tanpa latar belakang teknis, untuk memanfaatkan data untuk pengambilan keputusan yang lebih baik.

Melalui pembelajaran ini, peserta akan mempelajari berbagai alat dan platform yang dapat digunakan untuk analisis data secara intuitif, seperti software visualisasi data, analisis statistik, dan machine learning tanpa kode. Selain itu, peserta juga akan memahami bagaimana melakukan eksplorasi data, pembersihan data, dan penerapan model analitik sederhana tanpa menulis satu baris kode pun.

Di akhir pembelajaran, peserta diharapkan dapat mengaplikasikan konsep-konsep dasar data science dalam pekerjaan mereka sehari-hari dan mampu menghasilkan wawasan dari data yang ada, meskipun tidak memiliki pengalaman dalam pemrograman. Pembelajaran ini bertujuan untuk membuka akses kepada lebih banyak orang untuk terlibat dalam dunia data science, meningkatkan kemampuan analitik mereka, dan mengoptimalkan penggunaan data dalam berbagai konteks.

B. Machine Learning Mindmap





persyaratan untuk skenario ilmu data
Setelah Anda mengetahui apa yang ingin Anda lakukan dengan data Anda, Anda perlu menentukan persyaratan tambahan untuk solusi Anda.

Tentukan pilihan dan kemungkinan trade-off untuk persyaratan berikut:
a) Ketepatan
b) Waktu pelatihan
c) Linearitas
d) Jumlah parameter
e) Jumlah fitur

Ketepatan
Akurasi dalam pembelajaran mesin mengukur keefektifan model sebagai proporsi hasil sebenarnya terhadap kasus total. Dalam desainer Machine Learning, modul Evaluate Model menghitung sekumpulan metrik evaluasi standar industri. Anda dapat menggunakan modul ini untuk mengukur keakuratan model terlatih.

Tidak selalu perlu mendapatkan jawaban seakurat mungkin. Kadang-kadang perkiraan sudah cukup, tergantung untuk apa Anda ingin menggunakannya. Jika itu masalahnya, Anda mungkin dapat memotong waktu pemrosesan Anda secara dramatis dengan tetap menggunakan metode yang lebih mendekati. Metode perkiraan juga secara alami cenderung menghindari overfitting.

Ada tiga cara untuk menggunakan modul Evaluate Model:

1. Hasilkan skor atas data pelatihan Anda untuk mengevaluasi model
2. Hasilkan skor pada model, tetapi bandingkan skor tersebut dengan skor pada set pengujian yang dipesan
3. Bandingkan skor untuk dua model yang berbeda tetapi terkait, menggunakan kumpulan data yang sama

Untuk daftar lengkap metrik dan pendekatan yang dapat Anda gunakan untuk mengevaluasi keakuratan model machine learning, lihat modul Mengevaluasi Model.

Waktu pelatihan
Dalam supervised learning, pelatihan berarti menggunakan data historis untuk membuat model pembelajaran mesin yang meminimalkan kesalahan. Jumlah menit atau jam yang diperlukan untuk melatih model sangat bervariasi di antara algoritma. Waktu pelatihan sering kali terkait erat dengan akurasi; yang satu biasanya menemani yang lain.

Selain itu, beberapa algoritma lebih sensitif terhadap jumlah titik data daripada algoritma lainnya. Anda mungkin memilih algoritma tertentu karena Anda memiliki batasan waktu, terutama saat kumpulan datanya besar.

Di desainer Machine Learning, membuat dan menggunakan model machine learning biasanya merupakan proses tiga langkah:

1. **Konfigurasikan model,** dengan memilih jenis algoritme tertentu, lalu tentukan parameter atau hyperparameternya.

2. **Berikan set data yang diberi label dan memiliki data yang kompatibel dengan algoritme**. Hubungkan data dan model ke modul Model Latih.

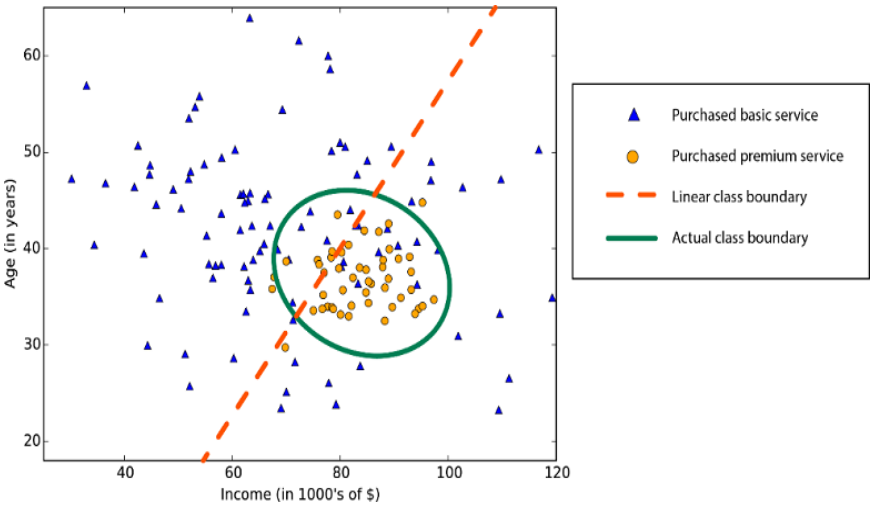
3. Setelah pelatihan selesai, **gunakan model terlatih dengan salah satu modul penilaian** untuk membuat prediksi pada data baru.

Linearitas
Linearitas dalam statistik dan pembelajaran mesin berarti ada hubungan linier antara variabel dan konstanta dalam kumpulan data Anda. Misalnya, algoritma klasifikasi linier mengasumsikan bahwa kelas dapat dipisahkan oleh garis lurus (atau analog berdimensi lebih tinggi).

Banyak algoritma pembelajaran mesin menggunakan linieritas. Di desainer Pembelajaran Mesin Azure, mereka mencakup:

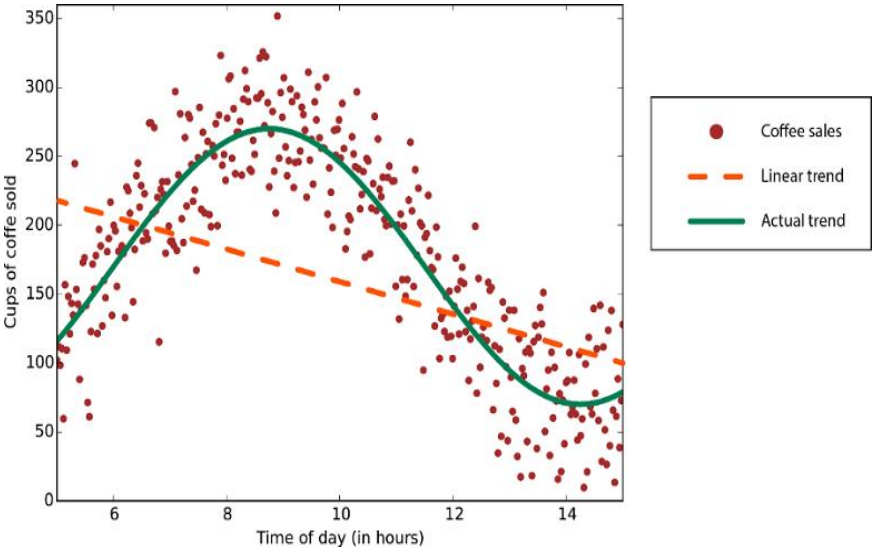
1. Regresi logistik multikelas
2. Regresi logistik dua kelas
3. Mendukung mesin vector

Algoritma regresi linier mengasumsikan bahwa tren data mengikuti garis lurus. Asumsi ini tidak buruk untuk beberapa masalah, tetapi untuk masalah lain ini mengurangi akurasi. Terlepas dari kekurangannya, algoritma linier populer sebagai strategi pertama. Mereka cenderung sederhana secara algoritme dan cepat dilatih.



Gambar 6. Algorithms Machine Learning Linier

Batas kelas nonlinier: Mengandalkan algoritme klasifikasi linier akan menghasilkan akurasi yang rendah.



Gambar 7. Algorithms Machine Learning Non Linier

Data dengan tren nonlinier: Menggunakan metode regresi linier akan menghasilkan kesalahan yang jauh lebih besar daripada yang diperlukan.

Jumlah parameter

Parameter adalah tombol yang dapat diputar oleh data scientist saat menyiapkan algoritme. Mereka adalah angka yang memengaruhi perilaku algoritme, seperti toleransi kesalahan atau jumlah iterasi, atau opsi di antara varian cara algoritme berperilaku. Waktu pelatihan dan keakuratan algoritme terkadang sensitif untuk mendapatkan pengaturan yang tepat. Biasanya, algoritme dengan jumlah parameter yang besar membutuhkan trial and error paling banyak untuk menemukan kombinasi yang baik.

Selain itu, ada modul Tune Model Hyperparameters di desainer Machine Learning: Tujuan modul ini adalah menentukan hyperparameter yang optimal untuk model machine learning. Modul ini membangun dan menguji beberapa model dengan menggunakan kombinasi pengaturan yang berbeda. Ini membandingkan metrik di semua model untuk mendapatkan kombinasi pengaturan.

Meskipun ini adalah cara yang bagus untuk memastikan Anda telah memperluas ruang parameter, waktu yang dibutuhkan untuk melatih model meningkat secara eksponensial dengan jumlah parameter. Keuntungannya adalah memiliki banyak parameter biasanya menunjukkan bahwa algoritme memiliki fleksibilitas yang lebih besar. Seringkali dapat mencapai akurasi yang sangat baik, asalkan Anda dapat menemukan kombinasi pengaturan parameter yang tepat.

Jumlah fitur

Dalam pembelajaran mesin, fitur adalah variabel yang dapat diukur dari fenomena yang Anda coba analisis. Untuk jenis data tertentu, jumlah fiturnya bisa sangat besar dibandingkan dengan jumlah titik data. Ini sering terjadi pada genetika atau data tekstual.

Sejumlah besar fitur dapat menghambat beberapa algoritme pembelajaran, membuat waktu pelatihan menjadi sangat lama. Mesin vektor pendukung sangat cocok untuk skenario dengan banyak fitur. Untuk alasan ini, mereka telah digunakan di banyak aplikasi mulai dari pencarian informasi hingga klasifikasi teks dan gambar. Mesin vektor pendukung dapat digunakan untuk tugas klasifikasi dan regresi.

Pemilihan fitur mengacu pada proses penerapan uji statistik ke masukan, dengan memberikan keluaran yang ditentukan. Tujuannya adalah untuk menentukan kolom mana yang lebih memprediksi keluaran. Modul Pemilihan Fitur Berbasis Filter di desainer Machine Learning menyediakan beberapa algoritme pemilihan fitur untuk dipilih. Modul tersebut mencakup metode korelasi seperti korelasi Pearson dan nilai chi-kuadrat.

Anda juga dapat menggunakan modul Permutasi Pentingnya Fitur untuk menghitung sekumpulan skor kepentingan fitur untuk kumpulan data Anda. Anda kemudian dapat memanfaatkan skor ini untuk membantu Anda menentukan fitur terbaik untuk digunakan dalam model.

mencakup tugas-tugas seperti **»» Menerjemahkan permintaan bisnis yang ambigu ke dalam masalah atau peluang yang konkret dan terdefinisi dengan baik ««** Mendalam ke dalam konteks permintaan untuk lebih memahami seperti apa solusi potensial itu, termasuk data mana yang akan dibutuhkan »» Menguraikan (jika mungkin) prioritas bisnis strategis yang ditetapkan oleh perusahaan yang mungkin memengaruhi pekerjaan ilmu data Sekarang saya telah membuat menjelaskan pentingnya menangkap dan memahami permintaan bisnis dan pelingkupan awal data yang diperlukan, saya ingin beralih ke mendeskripsikan aspek dari proses pengambilan data itu sendiri. Ini adalah antarmuka utama ke sumber data yang perlu Anda manfaatkan dan mencakup area seperti **»» Mengelola kepemilikan data dan mengamankan hak hukum untuk pengambilan dan penggunaan data ««** Penanganan informasi pribadi dan mengamankan privasi data melalui berbagai teknik **anonimisasi »» Menggunakan perangkat keras dan perangkat lunak untuk memperoleh data melalui upload batch atau streaming data secara real-time Framing Data Science Strategy »» Menentukan seberapa sering data perlu diperoleh, karena frekuensi biasanya bervariasi antara tipe dan kategori data ««** Mandat bahwa preprocessing data terjadi di titik pengumpulan, atau bahkan sebelum pengumpulan (di tepi perangkat IoT, misalnya). Ini termasuk pemrosesan dasar, seperti pembersihan dan penggabungan data, tetapi juga dapat mencakup aktivitas yang lebih canggih, seperti menganonimkan data untuk menghapus informasi sensitif.

Anonimisasi mengacu pada penghapusan informasi sensitif seperti nama seseorang, nomor telepon, alamat, dan sebagainya dari kumpulan data.) Dalam kebanyakan kasus, data harus dianonimkan sebelum ditransfer dari sumber data. Biasanya prosedur juga tersedia untuk memvalidasi kumpulan data dalam hal kelengkapan. Jika datanya tidak lengkap, pengumpulan mungkin perlu diulang beberapa kali untuk mencapai cakupan data yang diinginkan. Melakukan jenis validasi ini sejak awal berdampak positif pada kecepatan proses dan biaya. »» Mengelola proses transfer data ke titik penyimpanan yang dibutuhkan (lokal dan / atau global). Sebagai bagian dari transfer data, Anda mungkin harus mengubah data - menggabungkannya untuk membuatnya lebih kecil, misalnya. Anda

mungkin perlu melakukan ini jika Anda menghadapi batasan kapasitas bandwidth dari tautan transfer yang Anda gunakan. Menjaga Aktivitas pemeliharaan data mencakup penyimpanan dan pemeliharaan data. Perhatikan bahwa data biasanya diproses dalam banyak langkah berbeda sepanjang siklus hidupnya. Kebutuhan untuk melindungi integritas data selama siklus hidup elemen data sangat penting selama aktivitas pemrosesan data. Sangat mudah untuk tidak sengaja merusak kumpulan data melalui kesalahan manusia saat memproses data secara manual, menyebabkan kumpulan data menjadi tidak berguna untuk analisis di langkah berikutnya. Cara terbaik untuk melindungi integritas data adalah dengan mengotomatiskan sebanyak mungkin langkah aktivitas manajemen data yang mengarah ke titik analisis data. Menjaga kepercayaan bisnis pada fondasi data sangat penting agar pengguna bisnis dapat mempercayai dan memanfaatkan wawasan yang diperoleh.

Dalam hal pemeliharaan data, dua aspek penting adalah »» Penyimpanan data: Anggaplah ini sebagai segala sesuatu yang terkait dengan apa yang terjadi di data lake. Aktivitas penyimpanan data termasuk mengelola periode retensi yang berbeda untuk berbagai jenis data, serta membuat katalog data dengan benar untuk memastikan bahwa data mudah diakses dan digunakan.

Mengoptimalkan Investasi Ilmu Data Anda »» Persiapan data: Dalam konteks pemeliharaan data, persiapan data mencakup tugas-tugas pemrosesan dasar seperti pembersihan data tingkat kedua, pementasan data, dan agregasi data, yang semuanya biasanya melibatkan penerapan filter secara langsung saat data disimpan. Anda tentu tidak ingin memasukkan data dengan kualitas buruk ke dalam data lake Anda. Periode retensi data dapat berbeda untuk tipe data yang sama, bergantung pada tingkat agregasi. Misalnya, data mentah mungkin menarik untuk disimpan hanya dalam waktu singkat karena biasanya volumenya sangat besar sehingga mahal untuk disimpan. Di sisi lain, data yang digabungkan seringkali berukuran lebih kecil dan lebih murah serta lebih mudah disimpan dan oleh karena itu dapat disimpan untuk waktu yang lebih lama, bergantung pada kasus penggunaan yang ditargetkan. Proses Pemrosesan data adalah lapisan pemrosesan data utama yang difokuskan pada persiapan data untuk analisis, dan mengacu pada

penggunaan metodologi rekayasa data yang lebih canggih, seperti »» Klasifikasi data: Ini mengacu pada proses pengorganisasian data ke dalam kategori agar lebih efektif dan penggunaan yang efisien, termasuk aktivitas seperti pelabelan dan penandaan data. Sistem klasifikasi data yang terencana dengan baik membuat data penting mudah ditemukan dan diambil. Ini juga bisa menjadi sangat penting untuk bidang-bidang seperti hukum dan kepatuhan. »» Pemodelan data: Ini membantu representasi visual dari data dan menegakkan aturan bisnis yang mapan terkait data. Anda juga akan membangun model data untuk menerapkan kebijakan tentang bagaimana Anda harus menghubungkan tipe data yang berbeda secara konsisten.

Model data juga memastikan konsistensi dalam konvensi penamaan, nilai default, semantik, dan prosedur keamanan, sehingga memastikan kualitas data. »» Peringkasan data: Di sini tujuan Anda adalah menggunakan berbagai cara untuk meringkas data, seperti menggunakan teknik pengelompokan yang berbeda. »» Data mining: Ini adalah proses menganalisis kumpulan data besar untuk mengidentifikasi pola atau penyimpangan serta untuk membangun hubungan agar masalah dapat dipecahkan melalui analisis data selanjutnya. Penambangan data adalah sejenis analisis data, berfokus pada pemahaman data yang ditingkatkan, juga disebut sebagai literasi data. Membangun literasi data dalam tim ilmu data adalah komponen kunci dari kesuksesan ilmu data. Dengan literasi data yang rendah, dan tanpa benar-benar memahami data yang Anda persiapkan, analisis, dan dapatkan wawasannya, Anda berisiko tinggi gagal dalam hal investasi ilmu data. BAB 1 Strategi Ilmu Data Framing 13Analisis Analisis data adalah tahap di mana data menjadi hidup dan Anda akhirnya dapat memperoleh wawasan dari penerapan teknik analisis yang berbeda. Wawasan dapat difokuskan untuk memahami dan menjelaskan apa yang telah terjadi, yang berarti analisisnya bersifat deskriptif dan lebih reaktif. Ini juga kasus dengan analisis waktu nyata: Analisis masih reaktif bahkan ketika terjadi di sini-dan-sekarang. Kemudian ada metode analisis data yang bertujuan untuk menjelaskan tidak hanya mengapa sesuatu terjadi tetapi juga apa yang terjadi. Jenis analisis data ini biasanya disebut sebagai analisis diagnostik. Metode deskriptif dan diagnostik biasanya dikelompokkan ke dalam area pelaporan, atau business intelligence (BI).

C. Memilih Konsep Data-Driven Organization

Menyortir Konsep Organisasi Berdasarkan Data Data adalah warna hitam baru! Atau minyak baru! Atau emas baru! Apa pun yang Anda bandingkan dengan data, itu mungkin benar dari perspektif nilai konseptual. Sebagai masyarakat, kita sekarang memasuki era baru data dan mesin cerdas. Dan ini bukanlah tren yang lewat atau sesuatu yang dapat atau harus Anda hindari. Sebaliknya, Anda harus menerimanya dan bertanya pada diri sendiri apakah Anda cukup memahami tentang hal itu untuk memanfaatkannya dalam bisnis Anda. Bersikaplah terbuka dan ingin tahu! Berani bertanya pada diri sendiri apakah Anda benar-benar memahami apa artinya menjadi data-driven. Konsep dari data-driven adalah landasan yang perlu Anda pahami untuk melaksanakan pekerjaan strategis dalam ilmu data dengan benar, dan dibahas di beberapa bagian buku ini.

Dalam bab ini, saya mencoba memberi Anda gambaran besar tentang bagaimana berpikir dan bernalar seputar gagasan didorong oleh data. Jika Anda mulai dengan menempatkan perubahan yang sedang terjadi di masyarakat ke dalam konteks yang lebih luas, merupakan pemahaman umum bahwa kita manusia sekarang mengalami revolusi industri keempat, didorong oleh akses ke data dan teknologi canggih. Ini juga disebut sebagai revolusi digital. Tapi waspadalah! Mendigitalkan atau mendigitalkan bisnis Anda tidak sama dengan bisnis berdasarkan data. Digitalisasi adalah konsep yang banyak digunakan yang pada dasarnya mengacu pada transisi dari analog ke digital, seperti konversi data ke format digital.

Sehubungan dengan itu, digitalisasi mengacu pada pembuatan informasi digital yang berfungsi dalam bisnis Anda. Konsep digitalisasi bisnis terkadang dicampur dengan data-driven. Namun, penting untuk diingat bahwa mendigitalkan data bukan hanya hal yang baik untuk dilakukan - ini adalah dasar untuk memungkinkan perusahaan berbasis data.

Tanpa digitalisasi, Anda tidak bisa menjadi didorong oleh data. Mendekati berdasarkan data Dalam organisasi berdasarkan data, titik awalnya adalah data. Itu benar-benar dasar dari segalanya. Tapi apa maksud yang sebenarnya? Nah, menjadi data-driven berarti Anda harus siap menangani data dengan serius. Dan apa artinya itu?

Nah, dalam praktiknya, itu berarti data adalah titik awal dan Anda menggunakan data untuk menganalisis dan memahami jenis bisnis apa yang harus Anda lakukan. Anda harus menanggapi hasil analisis dengan cukup serius agar siap mengubah model bisnis Anda. Anda harus siap untuk mempercayai dan menggunakan data untuk memajukan bisnis Anda. Ini harus menjadi perhatian utama Anda di perusahaan. Anda harus menjadi "**terobsesi dengan data".**

Sebelum saya menjelaskan apa artinya terobsesi dengan data, pertimbangkan cara Anda melakukan berbagai hal saat ini di perusahaan Anda. Apakah ini agak berdasarkan data? Atau mungkin tidak sama sekali? Di manakah titik awal dalam area bisnis yang berbeda? Gambar 1-5 menunjukkan model (dengan contoh) untuk membandingkan pendekatan yang lebih tradisional dengan pendekatan berbasis data yang terkait dengan pendekatan aspek bisnis yang berbeda.

Mengoptimalkan Investasi Ilmu Data

Anda Perbedaan antara bisnis tradisional dan bisnis berbasis data. Membandingkan pendekatan dalam bisnis tradisional versus organisasi berbasis data adalah bermanfaat. Banyak pemimpin perusahaan yang benar-benar berpikir bahwa perusahaan mereka berdasarkan data hanya karena mereka mengumpulkan dan menganalisis data. Namun, yang terpenting adalah bagaimana data mendorong (atau tidak mendorong) prioritas, keputusan, dan eksekusi bisnis yang memberi tahu Anda bagaimana sebenarnya bisnis Anda berdasarkan data.

D. Memilah Konsep Machine Learning

Memilah-milah Konsep Pembelajaran Mesin Orang-orang sering meminta saya untuk menjelaskan perbedaan antara analitik tingkat lanjut dan pembelajaran mesin dan untuk mengatakan kapan sebaiknya menggunakan satu pendekatan atau yang lain. Saya selalu memulai dengan mendefinisikan pembelajaran mesin. Pembelajaran mesin (ML) adalah Mengoptimalkan Investasi Ilmu Data Anda studi ilmiah tentang algoritme dan model statistik yang digunakan sistem komputer untuk meningkatkan kinerjanya secara progresif pada tugas tertentu. Algoritme pembelajaran mesin membangun model matematika berdasarkan data sampel, yang dikenal sebagai data

pelatihan, untuk membuat prediksi atau keputusan tanpa diprogram secara eksplisit untuk melakukan tugas tersebut. Jadi, berikut ini cara analitik dan ML canggih memiliki beberapa karakteristik yang sama: **»» Teknik analitik dan pembelajaran mesin tingkat lanjut digunakan untuk membangun dan menjalankan model matematika dan statistik tingkat lanjut serta membangun model yang dioptimalkan yang dapat digunakan untuk memprediksi peristiwa sebelum terjadi . ««** Kedua metode menggunakan data untuk mengembangkan model, dan keduanya memerlukan kebijakan model yang ditentukan. »» Otomatisasi dapat digunakan untuk menjalankan model analitik dan model pembelajaran mesin setelah dimasukkan ke dalam produksi. Bagaimana dengan perbedaan antara analitik tingkat lanjut dan pembelajaran mesin? »» Ada perbedaan dalam hal siapa aktornya saat membuat model Anda. Dalam model analitik tingkat lanjut, aktornya adalah manusia; dalam model pembelajaran mesin, aktor (jelas) adalah mesin. »» Ada juga perbedaan dalam format model.

Model Analytics dikembangkan dan diterapkan dengan desain yang ditentukan manusia, sedangkan model ML bersifat dinamis dan mengubah desain dan pendekatan saat dilatih oleh data, sehingga mengoptimalkan desain selama proses tersebut. Model pembelajaran mesin juga dapat diterapkan sebagai model dinamis, yang berarti model tersebut terus melatih, mempelajari, dan mengoptimalkan desain saat dihadapkan pada data kehidupan nyata dan konteks langsungnya. »» Perbedaan lain antara model analitik dan model pembelajaran mesin berkaitan dengan perbedaan dalam cara model diuji menggunakan data (untuk analitik) dan dilatih menggunakan data (untuk pembelajaran mesin).

Dalam analitik, data digunakan untuk menguji bahwa hasil yang ditentukan tercapai seperti yang diharapkan, sedangkan dalam pembelajaran mesin, data digunakan untuk melatih model untuk mengoptimalkan desainnya tergantung pada sifat datanya. »» Terakhir, teknik dan alat yang digunakan untuk mengembangkan model analitik tingkat lanjut dan model ML berbeda. Teknik pemodelan pembelajaran mesin jauh lebih maju dan dibangun di atas prinsip-prinsip lain yang berkaitan dengan bagaimana mesin akan belajar untuk mengoptimalkan kinerja model.

Bagaimana model yang berbeda dapat dikembangkan, diuji, atau dilatih dan kemudian diterapkan. Seperti yang Anda lihat, model analitik selalu dikembangkan dan diuji secara statis, di mana aktor manusia memutuskan metode statistik mana yang akan digunakan dan cara menguji model menggunakan kumpulan data sampel yang ditentukan untuk mencapai kinerja model yang optimal. Dan, terlepas dari berapa banyak data (atau data mana) yang Anda dorong melalui model analitis, itu tetap sama sampai aktor manusia memutuskan untuk memperbaiki atau mengembangkan model.

Dalam pengembangan ML, aktor manusia juga menentukan teknik atau metode yang akan digunakan. Metode pelatihan dalam ML berbeda-beda bergantung pada teknik yang digunakan - Anda dapat menggunakan pembelajaran yang diawasi, misalnya, atau pembelajaran tanpa pengawasan, pembelajaran semi-supervisi, pembelajaran penguatan, atau bahkan pembelajaran mendalam, yang merupakan metode yang lebih kompleks. Bahkan mungkin untuk menggabungkan dua metode, seperti menggabungkan pembelajaran penguatan dengan pembelajaran mendalam hingga apa yang disebut sebagai pembelajaran penguatan mendalam. Alih-alih pendekatan statis yang digunakan dalam pengujian model tradisional, dengan model ML Anda terlebih dahulu melatih model menggunakan kumpulan data pelatihan yang dipilih yang harus mewakili lingkungan target tempat Anda ingin menerapkan model ML.

Selama pelatihan, kinerja model diujicobakan untuk memantau kemajuan pembelajaran sekaligus mengukur keakuratan model. Dalam cakupan metode ML yang dipilih, Anda kemudian membiarkan algoritme (aktor mesin) melatih dirinya sendiri pada kumpulan data pelatihan untuk mencapai target yang telah ditetapkan. Mesin tersebut kemudian terus melatih model ML untuk berkembang dan menemukan performa model yang paling optimal selama Anda membiarkannya. Akan tiba waktunya ketika akurasi model tidak dapat ditingkatkan dengan menggunakan set pelatihan. Pada tahap tersebut, Anda harus mengevaluasi apakah akurasi model cukup baik untuk penerapan.

Mengoptimalkan Investasi Ilmu Data Anda

Jika Anda memutuskan bahwa tingkat pelatihan yang memadai telah dicapai oleh aktor mesin, Anda perlu memutuskan cara menerapkan model di lingkungan target, - dengan kata lain, menerapkan ke produksi. Anda memiliki dua opsi saat ini. Anda dapat memutuskan bahwa model tersebut cukup terlatih untuk mencapai tujuannya dan Anda dapat menerapkannya sebagai model statis - artinya model tidak akan lagi mempelajari dan mengoptimalkan kinerja berdasarkan data, terlepas dari perubahan apa yang terjadi di lingkungan target. Atau, Anda dapat memutuskan untuk menerapkan model ML ke dalam produksi sebagai model dinamis, yang berarti model tersebut akan terus berkembang dan mengoptimalkan kinerjanya yang didorong oleh data dan perilaku yang mengisi model di lingkungan produksi.

Ini terkadang juga disebut sebagai pelatihan online. Jadi, kapan Anda harus mencari jenis model dan pendekatan penerapan apa? Ya, itu tergantung banyak faktor. Sebagai aturan panduan, Anda tidak boleh menggunakan ML jika Anda dapat menyelesaikan pekerjaan menggunakan pendekatan analitik. Mengapa? Untuk alasan yang sama Anda tidak menggunakan palu godam untuk menancapkan paku. Anda mungkin berhasil, tetapi Anda dapat dengan mudah menghancurkan paku dan melukai diri sendiri, menyebabkan hilangnya waktu dan uang. Ketika datang ke penyebaran statis atau dinamis, itu tergantung pada model bisnis dan apakah lingkungan target statis (perubahan jarang terjadi dan biasanya kecil) atau dinamis (perubahan sering terjadi dan dalam skala besar). Jika Anda mengembangkan algoritme untuk membuat rekomendasi online berdasarkan perilaku pengguna sebelumnya, misalnya, perlu untuk menerapkan model ML dinamis; jika tidak, Anda tidak dapat memenuhi tujuan Anda.

Sebaliknya, jika tujuan model ML adalah membiarkan mesin menemukan cara optimal untuk mengotomatiskan sekumpulan tugas kompleks yang Anda harapkan tetap sama dari waktu ke waktu, sebaiknya terapkan model ML sebagai statis model di lingkungan targetnya. Ketahuilah bahwa menerapkan model ML di lingkungan langsung membutuhkan lebih banyak resource dari Anda. Pelatihan machine learning itu rumit dan membutuhkan banyak kapasitas pemrosesan serta lebih banyak pemantauan model ML. Anda perlu memastikan bahwa model ML terus bekerja seperti

yang diharapkan dan tidak menurunkan atau menyimpang dari tujuannya sebagai bagian dari pelatihan langsungnya. Aspek lain yang perlu dipertimbangkan adalah kebutuhan untuk memastikan bahwa model tersebut dapat berinteraksi dengan model ML dinamis lainnya di lingkungan target tanpa mengganggu tujuan satu sama lain atau bertindak sedemikian rupa sehingga membuat model membatalkan satu sama lain. (Apa yang Anda lakukan di sini sering disebut sebagai memastikan interoperabilitas model.)

E. Definisi dan Scope

Mendefinisikan dan Mencakup Strategi Ilmu Data Untuk memahami bagian konstituen dari strategi ilmu data serta signifikansi strategi saat ini dan masa depan, ada baiknya untuk melihat beberapa komponen utama pada tingkat yang tinggi. Saya kemudian membahas masing-masing bagian yang berbeda ini secara rinci di seluruh buku ini. Namun sebelumnya saya perlu membuat klarifikasi singkat tentang perbedaan antara strategi ilmu data dan strategi data. Pada tingkat tinggi, strategi ilmu data mengacu pada strategi yang Anda tetapkan terkait dengan seluruh investasi ilmu data di perusahaan kami. Ini mencakup bidang-bidang seperti tujuan ilmu data secara keseluruhan dan pilihan strategis, strategi regulasi, kebutuhan data, kompetensi dan kumpulan keterampilan, arsitektur data, serta bagaimana mengukur hasilnya.

Strategi data di sisi lain, merupakan bagian dari strategi ilmu data, dan difokuskan pada garis besar arah strategis yang terkait langsung dengan data. Ini termasuk area seperti ruang lingkup data, persetujuan data, pertimbangan hukum, peraturan dan etika, frekuensi pengumpulan penyimpanan, periode retensi penyimpanan data, proses dan prinsip manajemen data, dan terakhir, namun tidak kalah pentingnya; tata kelola data. Kedua strategi tersebut diperlukan agar berhasil dengan investasi ilmu data Anda dan harus saling melengkapi agar berhasil.

Tujuan Jika saya bertanya tentang tujuan strategi ilmu data, saya bertanya apakah ada tujuan perusahaan yang ditetapkan dan disepakati dengan jelas untuk setiap investasi yang dilakukan dalam ilmu data. Apakah tujuan dirumuskan sedemikian rupa sehingga

dapat berujung pada pengembangan model statistik (model penjelasan, model kausal, dan sebagainya) dan dasbor.

F. Memulai Rencana Data Driven Transformation

Memulai Saat organisasi mengumpulkan lebih banyak data dan membangun model yang lebih akurat, manajer perubahan akan dapat dengan percaya diri menggunakannya untuk menentukan strategi yang memungkinkan organisasi mencapai tujuan mereka. Mereka akan dapat menjawab pertanyaan penting, seperti ini: 62 BAGIAN 1 Mengoptimalkan Investasi Ilmu Data Anda »» Siapa pemangku kepentingan yang terlibat? Jenis pendekatan perubahan apa yang berhasil dengan kelompok yang memiliki karakteristik ini? »» Risiko apa yang terkait dengan program yang menggunakan fitur-fitur ini? »» Apa saja teknik yang mempercepat penyampaian manfaat bisnis, dan berapa biaya relatifnya?

Apa sebab-akibat dari jenis investasi tertentu? Semua pertanyaan ini dapat dijawab dengan data dan akan mendukung rencana transformasi berbasis data. Mengembangkan jenis metrik ini tidaklah cepat atau mudah. Mereka bukanlah instalasi satu kali, melainkan komitmen beberapa tahun untuk menangkap data, membuat model, dan menyempurnakan dasbor. Membuat kumpulan data yang stabil dan andal membutuhkan waktu.

Kualitas data menjadi masalah di mana-mana, begitu pula kebutuhan akan bahasa data umum yang memungkinkan organisasi mengetahui bahwa mereka mengukur apa yang ingin mereka ukur. Ini telah menjadi masalah untuk analitik data di bidang lain; tidak ada alasan untuk berpikir bahwa manajemen perubahan akan berbeda. Meskipun membutuhkan waktu, pada akhirnya Anda akan dapat menutup lingkaran penyebab dan membuat prediksi yang andal tentang bagaimana tindakan atau inisiatif dalam program perubahan akan memengaruhi metrik tertentu.

Ini akan memindahkan investasi dalam perubahan dari tindakan keyakinan menjadi keputusan berdasarkan data. Manajemen perubahan akan beralih dari disiplin berbasis proyek yang berjuang untuk membenarkan investasi yang memadai ke yang memberi nasihat tentang hasil bisnis dan cara menyampaikannya. Ini akan menyebabkan penurunan satu metrik yang terkenal di seluruh

program perubahan - tingkat kegagalan. Dan, sebagai bagian dari memperkenalkan manajemen perubahan yang digerakkan oleh data, pada akhirnya harus mungkin untuk memecahkan teka-teki besar mengapa begitu banyak upaya transformasi gagal.

DAFTAR PUSTAKA

Grus, J. (2020). *Data Science from Scratch: First Principles with Python (2nd ed.).* O'Reilly Media.

Grus, J. (2020). *Data Science from Scratch: First Principles with Python (2nd ed.).* 2020: O'Reilly Media.

VanderPlas, J. (2021). *Python Data Science Handbook: Essential Tools for Working with Data.* O'Reilly Media.

VanderPlas, J. (2021). *Python Data Science Handbook: Essential Tools for Working with Data.* . O'Reilly Media.

Grus, J. (2020). *Data Science from Scratch: First Principles with Python (2nd ed.).* O'Reilly Media.

Park, A. (2020). *Data Science for Beginners: 4 Books in 1: Python Programming, Data Analysis, Machine Learning.* . Independently published.

GLOSARIUM

- **Algorithm (Algoritma)**: Prosedur atau instruksi langkah-demi-langkah yang digunakan untuk memecahkan masalah atau mencapai tujuan tertentu dalam pemrograman dan analisis data.
- **Big Data**: Kumpulan data yang sangat besar dan kompleks yang tidak dapat dikelola atau diproses menggunakan metode atau perangkat lunak tradisional.
- **Clustering**: Teknik dalam unsupervised learning untuk mengelompokkan data ke dalam kelompok atau klaster berdasarkan kesamaan atau kedekatannya.
- **Cross-validation**: Teknik untuk mengevaluasi kinerja model dengan membagi data menjadi beberapa bagian dan melatih model pada bagian yang berbeda untuk mengurangi bias.
- **Data Cleaning**: Proses pembersihan data dengan mengidentifikasi dan mengoreksi kesalahan atau inkonsistensi dalam data sebelum analisis dilakukan.
- **Data Mining**: Proses menemukan pola atau informasi yang berguna dari kumpulan data besar menggunakan teknik statistik dan algoritma.
- **Data Preprocessing**: Tahap persiapan data sebelum dianalisis, yang meliputi pembersihan, transformasi, dan pengurangan dimensi.
- **Data Visualization**: Teknik untuk menyajikan data dalam bentuk grafis atau visual untuk mempermudah pemahaman dan analisis.
- **Deep Learning**: Subset dari machine learning yang menggunakan jaringan saraf yang lebih kompleks untuk menganalisis data yang sangat besar dan tidak terstruktur.
- **Feature Engineering**: Proses menciptakan fitur baru dari data mentah yang dapat meningkatkan kinerja model prediksi.
- **Hypothesis Testing**: Proses statistik untuk menguji asumsi atau dugaan tentang data dengan menggunakan pengujian statistik.
- **Machine Learning (Pembelajaran Mesin)**: Metode analisis data yang memungkinkan sistem untuk belajar dan membuat prediksi atau keputusan tanpa pemrograman eksplisit.

- **Model Prediktif**: Model yang digunakan untuk memprediksi hasil atau kejadian masa depan berdasarkan data yang ada.
- **Neural Networks**: Struktur komputasi yang meniru cara kerja otak manusia, digunakan dalam deep learning untuk tugas-tugas kompleks seperti pengenalan gambar dan suara.
- **Overfitting**: Kondisi di mana model terlalu cocok dengan data pelatihan sehingga kehilangan kemampuan untuk menggeneralisasi data baru.
- **Regression**: Teknik analisis data yang digunakan untuk memprediksi nilai kontinu (misalnya, harga, suhu) berdasarkan variabel input.
- **Supervised Learning**: Jenis pembelajaran mesin di mana model dilatih menggunakan data berlabel untuk memprediksi hasil pada data baru.
- **Time Series**: Kumpulan data yang terorganisir dalam urutan waktu, digunakan untuk analisis prediksi di mana urutan waktu memiliki peran penting, seperti pada ramalan cuaca.
- **Underfitting**: Kondisi di mana model tidak cukup kompleks untuk menangkap pola dalam data, sehingga menghasilkan performa buruk baik pada data pelatihan maupun data baru.
- **Unsupervised Learning**: Teknik pembelajaran mesin di mana model dilatih dengan data tanpa label untuk menemukan pola atau struktur dalam data.

INDEKS

HASIL SCANNING SIMILARITY

Berisi tentang hasil scanning plagiat dengan batas toleransi 20%.

KOMENTAR REVIEWER

Ir. Onno Widodo Purbo, M.Eng., Ph.D.

Buku *Data Science untuk Pemula* adalah panduan yang luar biasa untuk siapa saja yang ingin memulai perjalanan mereka dalam dunia data science. Buku ini menyajikan konsep-konsep yang kompleks dengan cara yang sangat mudah dipahami, cocok untuk pembaca yang tidak memiliki latar belakang teknis atau pengalaman sebelumnya dalam analisis data.

Penulis berhasil mengemas teori dan praktik data science dengan cara yang terstruktur, dimulai dari dasar-dasar yang fundamental, seperti algoritma, big data, dan machine learning, hingga teknik yang lebih canggih seperti deep learning dan neural networks. Penjelasan yang jelas dan contoh praktis yang diberikan memberikan pemahaman yang mendalam, serta membantu pembaca merasa lebih percaya diri dalam menerapkan ilmu yang didapatkan. Salah satu kelebihan buku ini adalah penekanan pada pentingnya data preprocessing dan visualisasi data. Penulis dengan cermat menjelaskan pentingnya pembersihan dan transformasi data sebelum analisis dilakukan, yang merupakan salah satu aspek yang sering diabaikan oleh pemula. Dengan begitu, pembaca tidak hanya belajar tentang teori, tetapi juga cara praktis untuk mengelola dan menganalisis data secara efektif.

Buku ini juga memperkenalkan berbagai tools yang relevan dalam data science, seperti Python, R, dan berbagai paket analisis data lainnya. Setiap bab dilengkapi dengan latihan dan contoh soal yang membantu memperkuat pemahaman pembaca, yang membuat buku ini tidak hanya informatif tetapi juga interaktif. Buku ini sangat layak untuk dibaca oleh siapa saja yang tertarik dengan dunia data science,

baik itu mahasiswa, profesional, atau bahkan mereka yang baru ingin memulai karier di bidang ini. Buku ini memberikan fondasi yang kuat dan dapat dijadikan referensi yang berguna untuk pengembangan keterampilan lebih lanjut di dunia data science. Secara keseluruhan, *Data Science untuk Pemula* adalah sumber daya yang sangat berharga dan patut dimiliki bagi siapa saja yang ingin mempelajari dasar-dasar analisis data secara menyeluruh dan aplikatif.

BIOGRAFI PENULIS

Tutik Lestari, S.Si., M.Kom., adalah seorang akademisi dan peneliti di bidang Teknologi Informasi dan Komunikasi, khususnya dalam topik E-commerce, Data Science, dan Kecerdasan Buatan. Saat ini, beliau menjabat sebagai dosen di Universitas Darunnajah Jakarta dan aktif dalam berbagai penelitian serta publikasi ilmiah. Beliau juga seorang ibu dari 2 anak laki-laki dan berfokus pada perkembangan anak-anaknya untuk mendapat Pendidikan yang lebih baik.

Latar Belakang Pendidikan dan Karier

Tutik Lestari memperoleh gelar Sarjana Sains (S.Si) di Universitas Mercu Buana Jakarta dan Magister Komputer (M.Kom) dari universitas terkemuka di Indonesia, Universitas Budi Luhur Jakarta. Beliau memiliki pengalaman mengajar dan membimbing mahasiswa di bidang Teknik Informatika, serta berperan dalam pengembangan kurikulum dan penelitian di institusinya.

Bidang Keahlian dan Penelitian

Sebagai peneliti, fokus utama Tutik Lestari meliputi:

- **E-commerce dan Ekonomi Digital**: Menganalisis peran e-commerce dalam mendukung ekonomi digital Indonesia.
- **Data Science dan Big Data Analytics**: Mengembangkan metode analisis data untuk berbagai aplikasi, termasuk analisis sentimen dan sistem informasi berbasis web.
- **Kecerdasan Buatan (Artificial Intelligence)**: Menerapkan teknik AI dalam berbagai bidang, seperti pendidikan dan kesehatan.

Karya Ilmiah dan Publikasi

Tutik Lestari telah berkontribusi dalam berbagai jurnal ilmiah dan prosiding konferensi, baik nasional maupun internasional. Beberapa publikasi terbarunya antara lain:

- **"Peran e-Commerce dalam Mendukung Ekonomi Digital Indonesia"**: Ditulis bersama Sofian Lusa dan Onno W. Purbo, diterbitkan oleh Penerbit Andi pada tahun 2024.Google Scholar
- **"Sentiment analysis for IMDb movie review using support vector machine (SVM) method"**: Bersama DDN Cahyo dan lainnya, dipublikasikan dalam Inform: Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi pada tahun 2023.Google Scholar
- **"Transformation of Pesantren Education in the Digital Era: AI Innovation and Adaptation for Technology-Based Learning"**: Ditulis bersama A. Rahmayana, diterbitkan dalam The Electronic Integrated Computer Algorithm Journal pada tahun 2025.Google Scholar

Publikasi-publikasi tersebut mencerminkan dedikasi beliau dalam mengembangkan ilmu pengetahuan dan teknologi, serta kontribusinya terhadap perkembangan pendidikan dan teknologi informasi di Indonesia.

Keterlibatan dalam Pengabdian Masyarakat

Selain kegiatan akademik, Tutik Lestari juga aktif dalam program pengabdian masyarakat. Beliau terlibat dalam berbagai kegiatan, seperti Pelatihan Pengembangan Kompetensi Artificial Intelligence dalam Pembelajaran di Kelas untuk santri dan guru, serta pengembangan sistem informasi untuk mendukung kegiatan sosial dan pendidikan.

Kontak dan Jejak Digital

Untuk informasi lebih lanjut mengenai karya dan publikasi beliau, dapat mengunjungi profil Google Scholar: Tutik Lestari, S.Si., M.Kom. Di sana, tersedia daftar lengkap publikasi, kutipan, dan kolaborasi penelitian yang telah dilakukan.

Tutik Lestari, S.Si., M.Kom., merupakan contoh akademisi yang tidak hanya berkontribusi dalam dunia pendidikan, tetapi juga aktif dalam penelitian dan pengabdian masyarakat, dengan fokus pada pengembangan teknologi informasi untuk kemajuan bangsa.