PENGANTAR DATA SCIENCE

Mawaddah Harahap, M.Kom
Amir Mahmud Husein, M.Kom

Pengantar Data Science

PENULIS
Mawaddah Harahapp, M.Kom
Amir Mahmud Husein, M.Kom

EDITOR
Agung Prabowo, M.Kom

PENERBIT
UNPRI PRESS
ANGGOTA IKAPI

ISBN: **978-623-8299-27-0**

UNIVERSITAS
PRIMA INDONESIA
UNPRI PRESS

Pengantar Data Science

Penulis : *Mawaddah Harahap, M.Kom*
Amir Mahmud Husein, M.Kom
Editor : *Agung Prabowo, M.Kom*

Desain Isi : *Mawaddah Harahap, M.Kom*
Desain Cover : *Felix Thedora*

PENERBIT :

UNPRI PRESS
(ANGGOTA IKAPI)

Alamat Redaksi
Kampus 2
Jl. Sampul No. 4 Medan

KATA PENGANTAR

Puji syukur kami panjatkan kehadirat Tuhan Yang Maha Esa, berkat limpahan rahmat dan karunia-Nya, sehingga kami dapat menyusun Pengantar Data Science Tahun 2023 diUniversitas Prima Indonesia (UNPRI) Medan. Terima kasih kepadan Kementrian Pendidikan dan Kebudayaan, Riset, dan Teknologi Republik Indonesia(Kemendikbud) Republik Indonesia dalam rangka Kanpus Merdeka Indonesia Jaya.

Buku ini hadir sebagai panduan komprehensif bagi para pembaca yang ingin memahami konsep, teknik, dan aplikasi Data Science. Ditulis dengan tujuan untuk menjembatani kesenjangan antara teori dan praktik, buku ini dirancang agar dapat diakses oleh berbagai kalangan—mulai dari pelajar, akademisi, hingga profesional yang ingin memperdalam pengetahuan mereka dalam bidang ini.

Kami menyadari bahwa Data Science adalah bidang yang terus berkembang dengan cepat. Oleh karena itu, buku ini juga disusun dengan pendekatan yang dinamis, mencakup teori dasar sekaligus menerapkan teknik-teknik modern yang relevan dengan tantangan dunia nyata. Selain itu, kami berupaya menyajikan materi ini dengan cara yang mudah dipahami, lengkap dengan contoh-contoh kasus nyata yang akan membantu pembaca menghubungkan konsep teoretis dengan aplikasi praktis.

Kami berharap buku ini tidak hanya memberikan wawasan mendalam tentang Data Science, tetapi juga menginspirasi pembaca untuk terus belajar dan berkembang di bidang yang menantang ini. Semoga buku ini menjadi sumber pengetahuan yang bermanfaat dan dapat memperkaya pemahaman tentang dunia Data Science.

Medan, 10 September 2023

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB I

PENGANTAR DATA SCIENCE

A. PENDAHULUAN

Memahami dan menjelaskan pengantar data science mulai dari penerapan data science, defenisi data science dan data scientist, bagaiaman cara menggunakan aplikasi data science, alasan kenapa data science itu penting, manfaat dari data science dalam Dunia Usaha Dunia Industri(DUDI). beberapa siklus(metodologi) dalam data science serta mampu menyelesaikan masalah dengan menggunakan proses data science.

B. INTI

1. Capaian Pembelajaran
    a. Mampu memahami dan menjelaskan pengantar data science
    b. Mampu memahami dan menjelaskan peneraparan data science dalam berbagai bidang(hal).
    c. Mampu memahami dan menggunakan aplikasi(sotware) yang dapat digunakan pada data science.
    d. Mampu memahami dan menjelaskan defenisi data science dan data scientist.
    e. Mampu mengetahui alasan kenapa data science itu penting.
    f. Mampu memahami manfaat dari data science dalam Dunia Usaha Dunia Industri(DUDI).
    g. Mampu memahami dan menjelaskan beberapa siklus(metodologi) dalam data science.
    h. Mampu memahami dan menyelesaikan masalah dengan menggunakan proses data science.

2. Materi
    a. Peneraparan data science

b. Aplikasi data science
c. Defenisi data science
d. alasan data science penting
e. Manfaat data science
f. Siklus data science
g. Proses Data Science

3. Uraian Materi / Bahan Ajar / Kajian

a. Peneraparan data science

Data Science dapat diterapkan pada Dunia Usaha dan Dunia Industri (DUDI). Contoh dari penerapan ilmu data science dibidang atau dunia perindustrian terdiri dari :

1. Industri E-Commerce

Beberapa online shop telah menerapkan harga berdasarkan profil penghasilan konsumen. Saat ini orang-orang lebih banyak berbelanja secara digital atau biasa kita kenal dengan marketplace. Hampir semua orang menggunakan marketplace, penjual online secara otomatis akan menyesuaikan etalasenya berdasarkan profil 2data pembeli. Dengan mengubah tata letak halaman dan menyesuaikan produk jualannya secara otomatis dan real-time. Beberapa online shop juga menyesuaikan harga berdasarkan profil penghasilan konsumen, atau disebut dengan harga yang dipersonalisasi. Teknologi ini merupakan salah satu aplikasi penerapan ilmu Data Science.

2. Industri finance atau keuangan

Membantu analisis prediksi secara real-time. Mendeteksi penipuan(Fraud Detection) adalah bagian terpenting dari segala industri keuangan. Data Science dan AI adalah kedua ilmu yang sering digunakan disini. Bahkan kerusakan dan gangguan kecil akan dapat menyebabkan kerugian finansial. Analisis prediksi real-time membantu dalam peningkatan deteksi penipuan dan juga keamanan cyber. Dengan bantuan Data Science, perusahaan dapat menyediakan layanan

keuangan yang lebih efektif. Teknologi ini membantu untuk mengidentifikasi potensi transaksi penipuan yang dilakukan di setiap aktivitas. Dan ini juga akan membantu untuk memblokir sesi atau akun jika terdeteksi ada aktivitas keuangan yang tidak biasa.

3. Industri Travel

   Setiap kelompok pelanggan kemudian dapat ditawari produk dengan harga berbeda. Penetapan dynamic pricing sangat berguna dalam industri travel. Penetapan harga dinamis ini digunakan oleh perusahaan dengan menggunakan data untuk mensegmentasikan pelanggan atau konsumen secara akurat. Setiap kelompok pelanggan kemudian dapat ditawari produk dengan harga berbeda. Penawaran ini didasari pada informasi yang dihasilkan oleh Data Science dan berbagai faktor lainnya. Salah satu contoh perusahaan yang menerapkannya adalah Airbnb. Airbnb menggunakan ilmu Data Science dan algoritma dynamic pricing yang berfokus pada harga. Algoritma ini memperhitungkan berbagai macam kategori. Seperti lead time, review properti dan fasilitas yang disediakan. Algoritma ini juga dapat digunakan oleh pemilik properti untuk menentukan biaya per malam secara otomatis.

4. Industri Kesehatan

   Menerapkan model ML/DL untuk menemukan parameter optimal untuk tugas-tugas. seperti klasifikasi kondisi kesehatan pasien.

5. Industri Hiburan : Menerapkan Sistem Rekomendasi untuk menyusun daftar lagu berdasarkan genre musik atau band yang Anda ikuti saat ini.

b. Aplikasi data science

   Berikut ini adalah 8 top aplikasi pada data science:

   1. Deteksi Anomali

Menerapkan Analisis Statistik untuk menemukan anomali dalam kumpulan data. Mendeteksi perilaku pembelanjaan yang curang dalam data transaksi, secara real time untuk deteksi penipuan dan penggunaan lainnya

2. Pengenalan Pola

   Mengidentifikasi pola dalam kumpulan data (Amazon, Walmart, Buka Lapak, Tokopedia). Pengenalan pola membantu Retail dan perusahaan e-Commerce melihat tren dalam perilaku pembelian pelanggan, Membuat penawaran produk yang relevan kepada pelanggan. Strategi pemasaran yang lebih efektif.

3. Pemodelan Prediktif

   Meningkatkan kemampuan pengambilan keputusan dengan menciptakan model yang lebih baik dalam memprediksi perilaku pelanggan, risiko keuangan, tren pasar, dan banyak lagi. Aplikasi peramalan berbasis data yang lebih mampu merespons perilaku pelanggan yang berkembang

4. Mesin Rekomendasi dan personalisasi

   Rekomendasi produk dan layanan disesuaikan dengan kebutuhan Pelanggan. Menawarkan produk yang tepat di waktu yang tepat. Membangun profil terperinci dari pelanggan individu

5. Klasifikasi dan kategorisasi

   Memilah-milah volume data yang besar dan mengkategorikan atau mengklasifikasikannya berdasarkan karakteristik yang dipelajari. Data tidak terstruktur, dokumen, gambar, video, file audio, teks dan informasi. Klasifikasi transportasu jenis kendaraan, mengkategorikan daerah rawan kecalakaan berdasarkan data insiden. Mengkategorikan wilayah kemacetan di waktu dan kondisi tertentu.

6. Analisis sentimen dan perilaku

   Mengidentifikasi pola pembelian dan pengguna, mengetahui pendapat (Sikap) Pelanggan tentang produk dan layanan. Perusahaan Travel dan perhotelan telah mengadopsi analisis sentimen ini untuk mengidentifikasi pelanggan yang

memiliki pengalaman positif atau negatif sehingga dapat merespons dengan cepat untuk meningkat layanan.

7. Sistem percakapan :

   Melatih sistem ini pada sejumlah besar teks sehingga dapat memperoleh pola percakapan dari data, misalnya, untuk mencari informasi, membantu memproses transaksi, dan memberikan layanan kepada pelanggan

8. Sistem Otonom :

   Mobil tanpa pengemudi

c. Defenisi data science

Data science merupakan sebuah cabang ilmu yang mengggabungkan ilmu inferensi data, penggabungan algoritmik dan penggunaan teknologi untuk memecahkan masalah analitik yang kompleks. Urban Institute, Data science merupakan keterampilan yang membutuhkan ilmu komputer, pemrograman, teknologi, dan statistic Chikio Hayashi dari Institut Statistika Matematika Sakuragaoka, Ilmu pengetahuan interdisiplin tentang metode komputasi untuk mendapatkan wawasan berharga yang dapat ditindaklanjuti dari kumpulan data yang mencakup tiga fase yaitu desain data, mengumpulkan data, dan analisis data.

Data Science terus berkembang sebagai salah satu jalur karir yang paling menjanjikan dan diminati oleh para profesional yang terampil. Saat ini, para profesional data yang sukses memahami bahwa mereka harus melampaui keterampilan tradisional dalam menganalisis sejumlah besar data, penambangan data, dan keterampilan pemrograman. Untuk mengungkap kecerdasan yang berguna bagi organisasi mereka, data scientist harus menguasai spektrum penuh dari siklus hidup ilmu data dan memiliki tingkat fleksibilitas dan pemahaman untuk memaksimalkan pengembalian pada setiap fase proses.

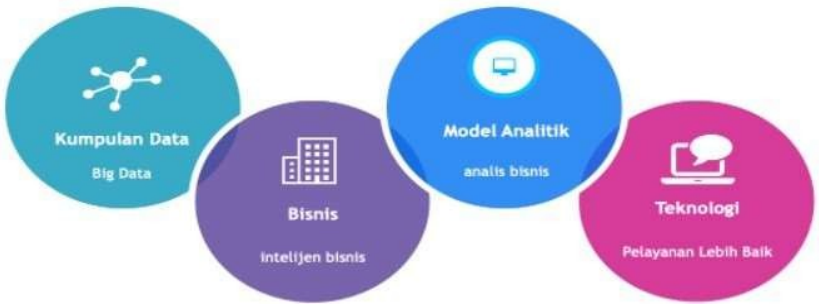
Kemampuan untuk mengambil data, untuk dapat memahaminya, memprosesnya, mengekstraksi nilai darinya, memvisualisasikannya, mengomunikasikannya itu akan menjadi keterampilan yang sangat penting dalam beberapa dekade mendatang. Data Scientist yang efektif mampu mengidentifikasi pertanyaan yang relevan, mengumpulkan data dari banyak sumber data yang berbeda, mengatur informasi, menerjemahkan hasil menjadi solusi, dan mengomunikasikan temuan mereka dengan cara yang secara positif memengaruhi keputusan bisnis. Keterampilan ini diperlukan di hampir semua industri, menyebabkan ilmuwan data yang terampil menjadi semakin berharga bagi perusahaan.

Berikut ini 10 pertanyaan penting data scientist:

1. Apa Permintaan Bisnis?
2. Apa yang dibutuhkan Bisnis?.
3. Siapa saja pemangku kepentingan dan apa kebutuhan individu mereka?
4. Apakah pemangku kepentingan memiliki harapan yang jelas?
5. Apa solusi paling sederhana yang menambah nilai bagi para pemangku kepentingan?
6. Berapa nilai proyek ini? Bagaimana itu akan diukur?
7. Mengapa melakukan proyek ini?
8. Apa saja risikonya?
9. Orang dan sumber daya apa yang dibutuhkan?
10. Apa pertanyaan lain yang harus dijawab

d. Alasan data science penting

Mengapa Data Scientist penting?



Gambar 1. Alasan data science itu penting

Data science adalah ilmu yang penting, karena bisnis kecil maupun besar sangat bergantung pada data. Jika perusahaan tidak mampu mengolah data, bisnis tidak akan memiliki pedoman untuk strategi operasi yang efektif dan efisien untuk mendapat keuntungan. Saat ini, perusahaan-perusahaan di seluruh dunia semakin menyadari pentingnya data science, kecerdasan buatan, dan machine learning. Jika sebuah bisnis ingin berkompetisi dan tetap relevan, ia harus mampu mengimplementasi data science. Hal Varian, seorang ahli ekonom Google dan dosen Ilmu Komputer, Bisnis, dan Ekonomi UC Berkeley mengatakan bahwa kemampuan mengambil, memahami, memroses, dan menyaring nilai dari suatu data serta memvisualisasikannya adalah keahlian yang semakin penting di dekade yang akan datang. Tentu saja, orang yang memiliki pemahaman data science yang baik akan menjadi berharga dan banyak dicari. Berikut ini adalah tantangan yang ditemukan dalam data science:

1. Menghilangkan bias dalam kumpulan data dan aplikasi analitik.
2. Mengelola penerapan model analitik, mengukur nilai bisnis, dan mempertahankan model.
3. Menemukan data yang tepat untuk dianalisis.
4. Mengkomunikasikan ide-ide yang kompleks dan membuat keputusan organisasi yang didorong oleh data.

Kondisi saat ini yang terjadi, sehingga data science diperlukan adalah sebagai berikut:

1. Kumpulan data yang terus meningkat (Big Data)
2. Kebutuhan akan respon cepat terhadap perubahan lingkungan bisnis

3. Industri saat ini membutuhkan berbagai aplikasi yang memberikan wawasan dan nilai bisnis yang lebih baik

Data Scientist bertanggung jawab dalam mengumpulkan, menganalisa, dan mengolah data untuk menghasilkan informasi yang berguna bagi perusahaan(organisasi) dalam membuat keputusan. Pekerjaan yang dilakukan Data Scientist adalah : Defenisi masalah, pengumpulan data, pembersihan data, explorasi data, analisis data, permodelan data, hasil dan wawasan dimasa depan.

Kemampuan yang dimiliki seorang data scientist adalah sebagai berikut:

1. Pemrograman (Python, R, dan lainnya)
2. Matematikan dan Statistik.
3. Database dan Query
4. Data Mining
5. Machine Learning/Deep Learning

Contoh dari jabatan pekerjaan dibidang data science adalah: Data Scientist, Data Expert, Ahli Data, Pengolah Data dan Analis Data

Tugas yang dilakukan Data Scientist terdiri dari:

1. Memilih fitur, membangun dan mengoptimalkan pengklasifikasi menggunakan teknik belajar mesin
2. Data mining menggunakan metode state-of-the-art
3. Memperluas data perusahaan dengan sumber informasi pihak ketiga bila diperlukan
4. Meningkatkan prosedur pengumpulan data untuk memasukkan informasi yang relevan untuk membangun sistem analitik
5. Pengolahan, pembersihan, dan verifikasi integritas data yang digunakan untuk analisis
6. Melakukan analisis ad-hoc dan menyajikan hasilnya secara jelas
7. Membuat sistem deteksi anomali otomatis dan pelacakan konstan kinerjanya

Berikut ini adalah pengetahuan yang dimiliki Data Scientist :

1. Terbukti pengalaman sebagai Data Scientist atau Data Analyst
2. Pengalaman di data mining
3. Pengertian penelitian mesin belajar dan operasi
4. Pengetahuan tentang R, SQL dan Python; Keakraban dengan Scala, Java atau C ++ adalah aset
5. Pengalaman menggunakan alat intelijen bisnis (misalnya Tableau) dan kerangka data (misalnya Hadoop)
6. Kecerdasan analitis dan ketajaman bisnis
7. Kemampuan matematika yang kuat (misalnya statistik, aljabar)
8. Kemampuan memecahkan masalah
9. Memiliki kemampuan komunikasi dan presentasi yang baik
10. BSc / BA di bidang Ilmu Komputer, Teknik atau bidang yang relevan; Gelar sarjana di bidang Ilmu Data atau bidang kuantitatif lainnya lebih diutamakan.

Yang dilakukan data science dan keterampilan yang diperlukan, terdapat pada gambar berikut ini.

Gambar 2. Yang dilakukan data science dan keterampilan yang diperlukan

e. Manfaat data science

Berikut ini adalah manfaat Data Science:

1. Organisasi memiliki metodologi, alat, dan teknologi untuk mendapatkan informasi berharga dari jumlah data yang sangat bervariasi yang terus meningkat.
2. Organisasi memiliki keunggulan kompetitif atas persaingan bisnis; layanan yang lebih baik kepada pelanggan, dan kemampuan untuk merespon secara lebih efektif terhadap lingkungan bisnis yang berubah dengan cepat yang menuntut adaptasi terus-menerus.
3. Industri saat ini membutuhkan berbagai aplikasi yang memberikan wawasan dan nilai bisnis yang lebih baik di perusahaan

Tujuan utama data science adalah untuk mengolah dan mengekstrak data supaya mendapatkan data yang benar. Kebutuhan untuk posisi Data Scientist pun telah tumbuh pesat hingga mencapai angka lebih dari 650% sejak tahun 2012. Namun, belum banyak orang yang mengeksploitasi kesempatan kerja di bidang ini. Maka dari itu, bisa dikatakan peluang untuk menggeluti bidang data science sangatlah menjanjikan.

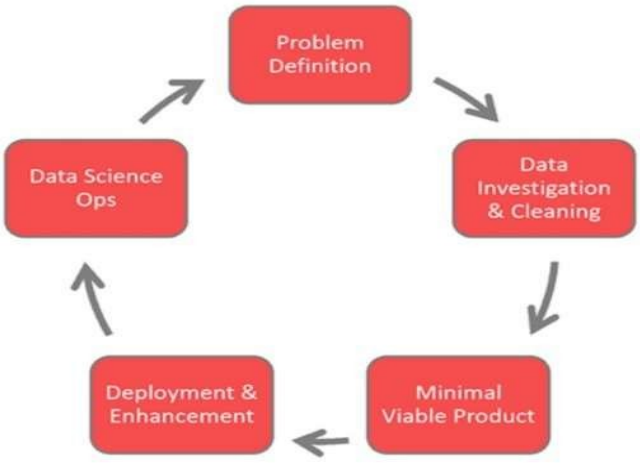
Perusahaan perlu menggunakan data untuk menjalankan dan mengembangkan bisnis mereka sehari-hari. Tujuan utama data science adalah untuk membantu perusahaan membuat keputusan yang lebih cepat dan lebih baik. Hal ini akan membawa perusahaan ke puncak pasar sesuai bidang perusahaan tersebut.

Jumlah perusahaan yang siap menggunakan data yang besar pun semakin meningkat. Bahkan studi menunjukkan bahwa, 40% dari non-pengguna berharap bisa mengadopsi dan mengolah data yang besar di tahun-tahun mendatang. Terlebih lagi, kamu dapat menerapkan pembelajaran pada set data yang lebih kecil, seperti data dari media sosial pada suatu perusahaan.

Ini memberikan lebih banyak peluang dan meningkatkan permintaan untuk seorang Data Scientist. Saat ini, banyak perusahaan sedang membangun tim data science untuk menjadikannya bagian integral dari kesuksesan mereka. Sebagian besar Data Scientist dalam industri telah berkembang dalam statistik, matematika, dan ilmu komputer.

Pengalaman mereka adalah cakrawala luas yang juga melingkupi ranah visualisasi data, data mining, dan manajemen informasi. Selain itu, sudah umum bagi mereka untuk memiliki pengalaman sebelumnya dalam desain infrastruktur, cloud computing, dan data warehousing.

f. Siklus data science



Gambar 3. Siklus Data Science

Berikut ini adalah Siklus Data Science:

1. Defenisi Masalah.: Identifikasi jenis masalah yang dipecahkan
2. Investigasi dan Pembersihan Data:
    a. Apakah data tersedia secara internal?
    b. Dokumentasikan kualitas data, Bersihkan data, Visualisasi Data (EDA).
    c. Menyajikan temuan awal kepada pemangku kepentingan dan meminta umpan balik
3. Modeling Data yang layak
    a. Apakah model bekerja di lingkungan yang terkendali?
    b. Bagaimana kinerja model pada data saat ini?
4. Penerapan dan Penyempurnaan

a. Perluas model ke kasus penggunaan serupa (yaitu fase "Definisi Masalah" baru)

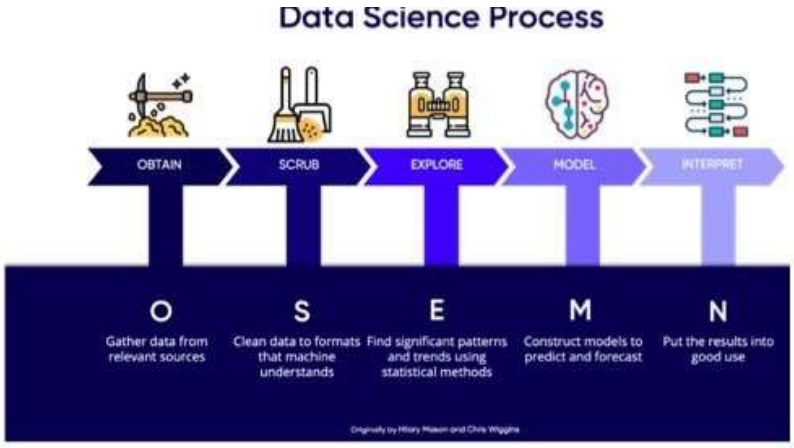
b. Menambah dan membersihkan kumpulan data (yaitu fase “Investigasi dan Pembersihan Data” baru)

c. Mencoba teknik pemodelan baru (yaitu mengembangkan “Modeling Data") berikutnya

5. Implementasi (Operasi): Mengukur keakuratan model dengan menggunakan metrik tertentu untuk melihat seberapa baik kinerjanya.

g. Proses Data Science

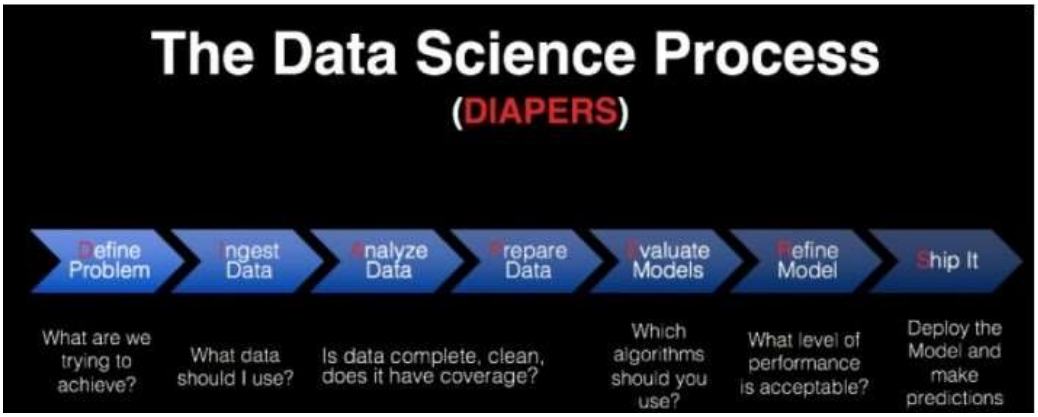
Berikut ini adalah proses data science dengan pendekatan OSEMN Framework



Gambar 4. Data Science Process (OSEMN Framework)

Keterangan :

1. Obtain/Data Collection : mengumpulkan data dari sumber yang relevan)
2. Scrub/Data Preparation : membersihkan data menjadi format yang dimengerti machine
3. Explore : menemukan pola menggunakan statistical methods
4. Model : membangun model untuk predict dan forecast
5. nterpret : menginterpretasi hasil analisis menjadi executable



Gambar 5. Data Science Project Life Cycle

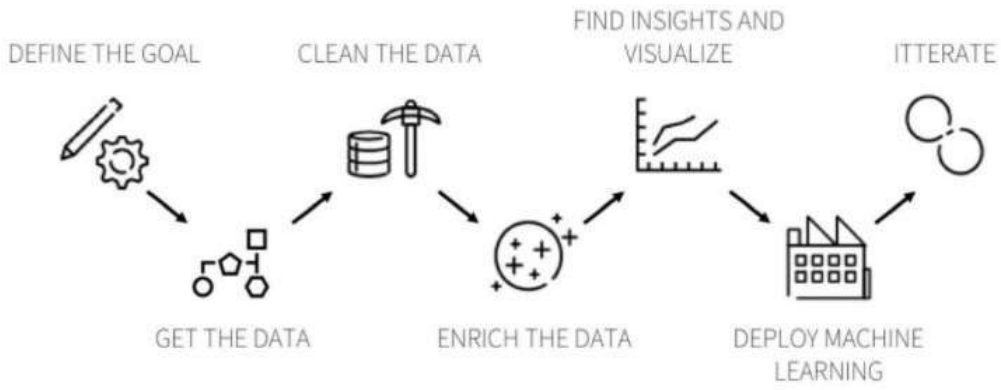
Berikut ini adalah Data Science Project Life Cycle

1. Define Problem Statetment: Tiga pertimbangan terpenting dalam real estat adalah lokasi, lokasi, lokasi. Dalam proyek data science, terdiri dari: Use Case, Use Case, Use Case. Pernyataan masalah yang kohesif dan kasus penggunaan yang memadai merupakan dasar bagi keberhasilan suatu proyek. Apa yang Anda coba capai; apa solusi saat ini; apa kesenjangan? Dilempar ke dalam asumsi untuk ukuran yang baik, dan Anda mungkin siap untuk berbaris maju di sisa perjalanan ini.
2. Ingest Data: Memahami sumber data, jenis, dan mekanisme penyerapan adalah langkah penting berikutnya. Data dapat terstruktur, tidak terstruktur atau semi terstruktur; mekanisme dan desain konsumsi bervariasi sesuai. Memahami frekuensi dan kecepatan (batch atau real-time) adalah pertimbangan lain. Langkah ini adalah dimana akan merancang solusi penyerapan data. Bergantung pada kebutuhan,mungkin perlu mempertimbangkan Big Data pipaline dan Data Lake Architcture.
3. Analyze Data: Pemahaman yang mendetail tentang data akan diperlukan untuk mendapatkan wawasan paling banyak dari eksperimen ilmu data kami. Ini dapat dilakukan dengan dua cara:
    a. Keahlian domain - Seorang ahli materi pelajaran (UKM) akan sangat penting untuk ini. Misalnya, UKM yang memahami domain data

perawatan kesehatan menambah nilai luar biasa pada proyek ilmu data perawatan kesehatan. Namun, ada contoh menarik baru-baru ini di mana orang-orang tanpa banyak keahlian domain telah menciptakan hasil yang sukses. Berikut adalah contoh lulusan perguruan tinggi tanpa pengetahuan domain apa pun yang membuat aplikasi untuk mengidentifikasi sel kanker.

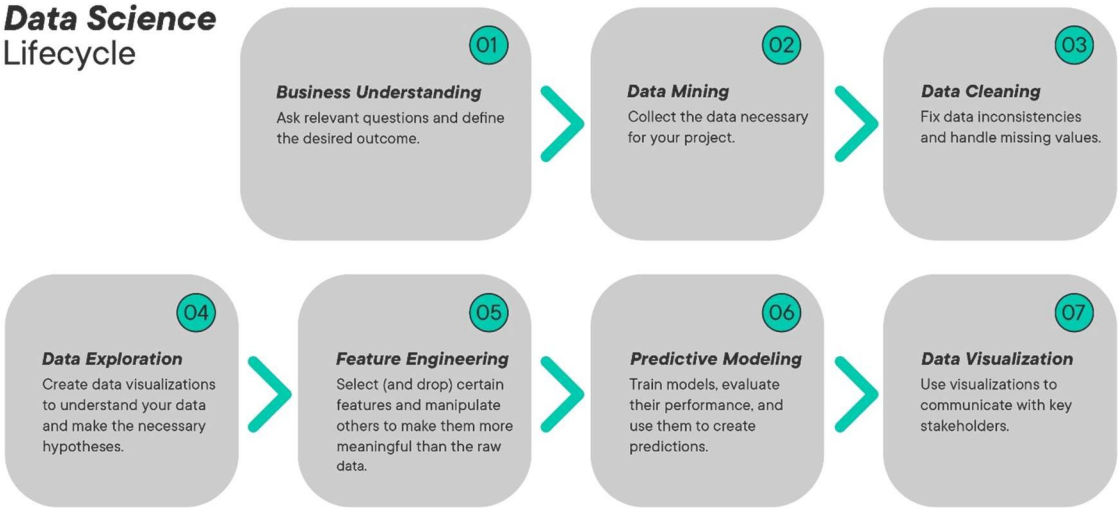
b. Eksplorasi Data - Ketersediaan UKM dilengkapi dengan eksplorasi data menggunakan visualisasi (misalnya Matplotlib) dan alat analisis dan profil lainnya.

4. Prepare Data For Learning: Eksplorasi data sering kali mengidentifikasi kebutuhan untuk membersihkan, menstandarisasi, dan menganonimkan data (ini adalah hasil yang SANGAT umum di dunia nyata). Kami mungkin perlu menghapus ID dan kode yang tidak berguna, menghapus informasi pengenal pribadi agar sesuai dengan standar dan undang-undang, dan membersihkan nilai data (menstandarkan ejaan, dll.) Hal ini dapat menyebabkan Seleksi Fitur dan/atau Rekayasa Fitur (yaitu membuat fitur baru dari yang sudah ada).
5. Evaluate Model: menggunakanbanyak algoritma yang berbeda dan mengevaluasi model mana yang memberikan skor yang lebih baik (bias dan keseimbangan presisi, dll.) Berdasarkan analisis yang dilakukan, beberapa finalis adalah dipilih untuk tahap selanjutnya.
6. Refine model: Untuk mendapatkan model yang paling optimal, teknik lanjutan seperti penyetelan Hyperparameter, validasi silang, dll dapat diterapkan. Kadang-kadang kompromi dibuat pada akurasi untuk memilih model yang lebih sederhana.
7. Ship It: model final yang dioptimalkan kedalam produksi, termasuk persyaratan non-fungsional seperti logging, keamanan, dan kinerja. Dan tentu saja, karena tidak ada yang permanen, pada titik tertentu akan tiba saatnya untuk melatih kembali model dan memulai proses dari awal.



Gambar 6. Step of Data Project

Ada lebih banyak hal dalam ilmu data daripada hanya memilih, menerapkan, dan menyetel algoritme Pembelajaran Mesin. Proyek ilmu data sering kali mencakup tahapan berikut:



Gambar 7. Proyek Data Science

Dibagian ini, akan melalui setiap tahapan ini dan melihat apa yang terlibat.

1. Pemahaman Bisnis / Pengetahuan Domain

   Sebelum mencoba memecahkan masalah terkait data, penting bahwa Ilmuwan/Analis Data memiliki pemahaman yang jelas tentang domain masalah dan jenis pertanyaan yang perlu dijawab oleh analisis mereka. Beberapa pertanyaan yang mungkin ditanyakan oleh Data Scientist meliputi:

a. Berapa banyak atau berapa banyak? Misalnya Mengidentifikasi jumlah pelanggan baru yang kemungkinan akan bergabung dengan perusahaan Anda pada kuartal berikutnya. (Analisis regresi)
b. Kategori yang mana? Misalnya Menetapkan dokumen ke kategori tertentu untuk sistem manajemen dokumen. (Analisis klasifikasi)
c. Grup yang mana? Misalnya Membuat sejumlah grup (segmen) pelanggan Anda berdasarkan nilai uang mereka. (Kekelompokan)
d. Apakah ini aneh? Misalnya, mendeteksi aktivitas mencurigakan pelanggan oleh perusahaan kartu kredit untuk mengidentifikasi potensi penipuan. (Deteksi anomali)
e. Item mana yang lebih disukai pengguna? Misalnya Merekomendasikan produk baru (seperti film, buku, atau musik) kepada pelanggan yang sudah ada (Sistem rekomendasi)

Gambar 8. Industry domain knowledge

2. Penambangan Data/Data Mining

Setelah mengidentifikasi tujuan analisis Anda dan menyetujui pertanyaan analitis yang perlu dijawab, langkah selanjutnya adalah mengidentifikasi dan mengumpulkan data yang diperlukan.

Data mining adalah proses mengidentifikasi dan mengumpulkan data yang menarik dari berbagai sumber - database, file teks, API, Internet, dan bahkan dokumen tercetak. Beberapa pertanyaan yang mungkin Anda tanyakan pada diri sendiri pada tahap ini adalah:

a. Data apa yang saya perlukan untuk menjawab pertanyaan analitis saya?
b. Di mana saya dapat menemukan data ini?
c. Bagaimana saya bisa mendapatkan data dari sumber data?
d. Bagaimana cara mengambil sampel dari data ini?
e. Apakah ada masalah privasi/hukum yang harus saya pertimbangkan sebelum menggunakan data ini?

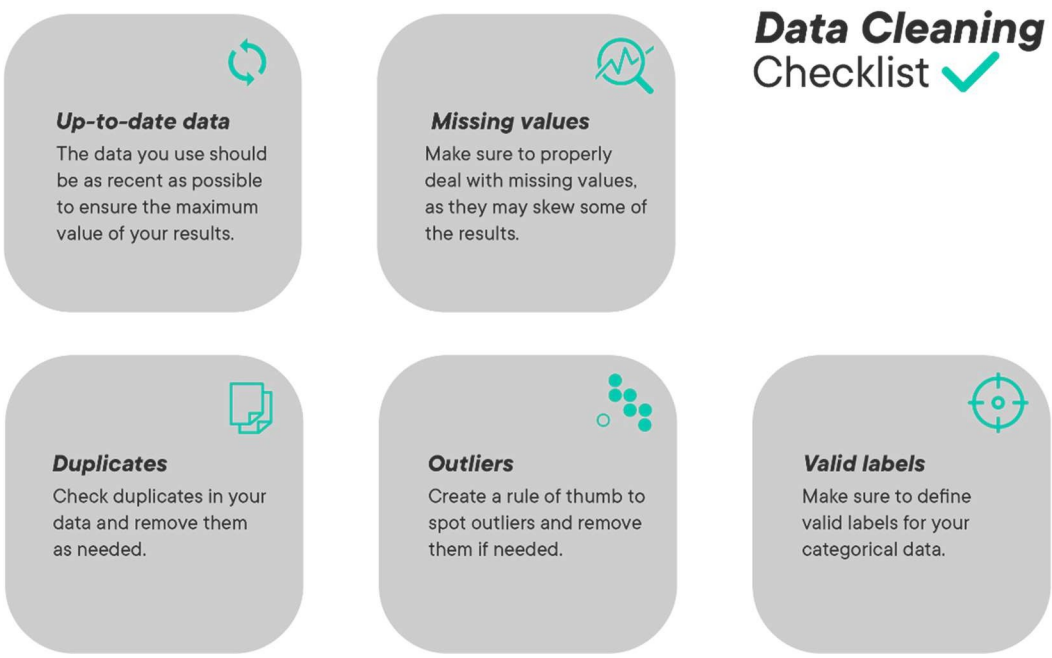
Gambar 9. Sumber Data

3. Pembersihan Data/Data Cleaning

Pembersihan data biasanya merupakan tahap proses Ilmu Data yang paling memakan waktu. Tahap ini mungkin memakan waktu hingga 50-80% dari waktu Data Scientist karena ada banyak kemungkinan masalah yang membuat data "kotor" dan tidak cocok untuk dianalisis. Beberapa masalah yang mungkin Anda lihat dalam data adalah:

a. Inkonsistensi dalam data
b. Data teks salah eja
c. Pencilan
d. Data tidak seimbang
e. Data tidak valid/kedaluwarsa
f. Data hilang

Tahap ini membutuhkan pengembangan strategi yang cermat tentang bagaimana menangani masalah ini. Strategi tersebut dapat bervariasi secara substansial antara analisis yang berbeda tergantung pada sifat masalah yang dipecahkan.

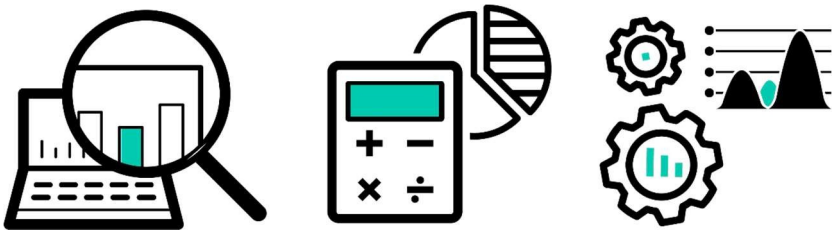


Gambar 10. Tahapan Cleaning Data

4. Eksplorasi Data

   Eksplorasi data atau Exploration Data Analysis (EDA) membantu menyoroti pola dan hubungan dalam data. Analisis eksplorasi dapat melibatkan kegiatan berikut:

   a. Menghitung statistik deskriptif dasar seperti mean, median, dan modus
   b. Membuat berbagai plot termasuk histogram, plot sebar, dan kurva distribusi untuk mengidentifikasi tren dalam data
   c. Visualisasi interaktif lainnya untuk fokus pada segmen data tertentu



Gambar 11. Jenis EDA

5. Rekayasa Fitur

   Sebuah "Fitur" adalah atribut terukur dari fenomena yang diamati - jumlah kamar tidur di rumah atau berat kendaraan. Berdasarkan sifat pertanyaan analitis yang diajukan pada langkah pertama, Ilmuwan Data mungkin harus merekayasa fitur tambahan yang tidak ditemukan dalam kumpulan data asli. Rekayasa fitur adalah proses menggunakan pengetahuan ahli untuk mengubah data mentah menjadi fitur bermakna yang secara langsung mengatasi masalah yang Anda coba pecahkan. Misalnya, mengambil berat dan tinggi badan untuk menghitung Indeks Massa Tubuh untuk individu dalam dataset. Tahap ini secara substansial akan mempengaruhi keakuratan model prediksi yang Anda buat di tahap berikutnya.

Gambar 12. Rekayasa Fitur

6. Pemodelan Prediktif

   Pemodelan adalah tahap di mana Anda menggunakan pendekatan matematis dan/atau statistik untuk menjawab pertanyaan analitis Anda. Pemodelan Prediktif mengacu pada proses menggunakan metode statistik probabilistik untuk mencoba memprediksi hasil dari suatu peristiwa. Misalnya, berdasarkan data karyawan, organisasi dapat mengembangkan model prediktif untuk mengidentifikasi tingkat pengurangan karyawan untuk mengembangkan strategi retensi yang lebih baik.

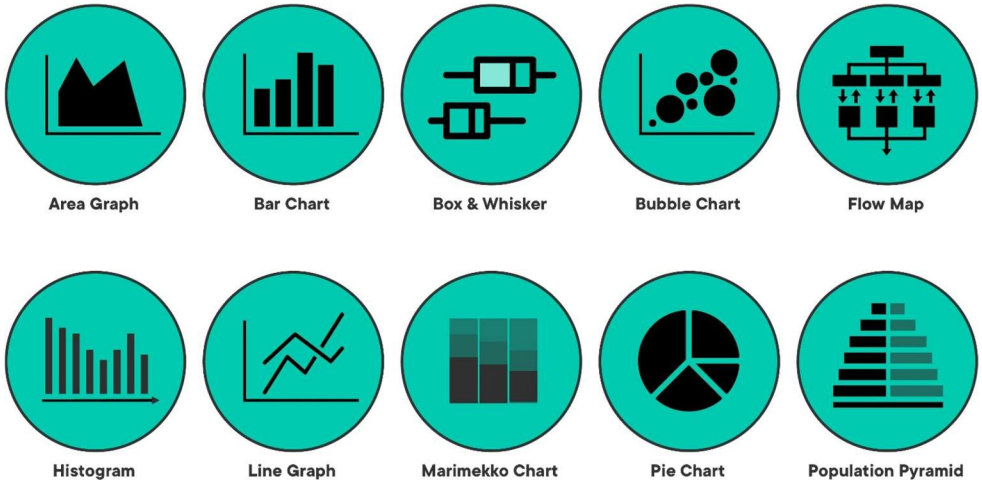
   Memilih model yang "benar" seringkali merupakan keputusan yang menantang karena tidak pernah ada satu jawaban yang benar. Memilih model melibatkan keseimbangan akurasi dan biaya komputasi dari proses analisis. Misalnya, beberapa pendekatan terbaru dalam pemodelan prediktif seperti pembelajaran mendalam telah terbukti menawarkan akurasi hasil yang jauh lebih baik, tetapi dengan biaya komputasi yang sangat tinggi.

Gambar 13. Pemodelan Prediktif

7. Visualisasi data

   Setelah memperoleh hasil yang diperlukan dari model statistik, visualisasi biasanya digunakan untuk meringkas dan menyajikan temuan dari proses analisis dalam bentuk yang mudah dimengerti oleh pengambil keputusan non-teknis.

Visualisasi data dapat dianggap sebagai evolusi teknik komunikasi visual karena berkaitan dengan representasi visual data. Ada berbagai macam teknik visualisasi data yang berbeda, dari grafik batang, grafik garis dan plot pencar hingga diagram aluvial dan visualisasi spatio-temporal, yang masing-masing akan bekerja lebih baik untuk menyajikan jenis informasi tertentu.



Gambar 14. Visualisasi Data

4. Forum Diskusi

a. Berdasarkan aplikasi(software) yang digunakan dalam data science, pelajari dan fahami peran dan fungsi pada setiap aplikasi tersebut. Dengan cara anda membentuk tim diskusi, dimana masing-masing anggota tim kelompok anda memilih masing-masing 1 aplikasi(software) lalu membahas dan menjelaskan dari setiap aplikasi(software) yang yang telah difahami oleh setiap anggota.

b. Carilah kelebihan, kekurangan, fungsi dari masing-masing aplikasi yang dibahas sehingga dapat dijadikan rujukan dalam memilih aplikasi yang tepat berdasarkan masalah yang akan diselesaikan.

C. PENUTUP

1. Rangkuman

Memasuki era Industri 4.0, istilah Data Science, Artificial Intelligence, dan Machine Learning, mungkin sudah familiar untuk kita. Hal ini diiringi dengan semakin banyaknya perusahaan yang mulai menyadari pentingnya menerapkan Data Science dan menggunakan Big Data untuk bisnis mereka. Profesi talenta digital di bidang tersebut pun semakin diminati di berbagai kalangan saat

ini.Metodologi data science adalah langkah- langkah digunakan dalam proyek data science agar dapat menghasilkan hasil yang optimal yang dapat menjawab pertanyaan dari suatu masalah yang ingin diselesaikan. Tahapan dalam metodologi data science merupakan proses berulang, yang mana jika dalam suatu tahapan dirasa masih belum sesuai, bisa kembali ke tahap sebelumnya.

2. Tes Formatif

Sebagai evaluasi pemahaman anda terhadap materi, jawablah pertanyaan berikut ini.

1. Sebutkan beberapa kemampuan data scientist yang efektif
2. Berikut ini yang merupakan aplikasi pada bidang data sceince, adalah
   a. Deteksi anomali
   b. Pengenalan Pola
   c. Analisis sentimen dan perilaku
   d. Sistem Percakapan
3. Berikut ini adalah beberapa kondisi yang terjadi pada saat ini sehingga diperlukannya data science adalah
   a. Kumpulan data yang terus meningkat (Big Data)
   b. Kebutuhan akan respon cepat terhadap perubahan lingkungan bisnis
   c. Industri saat ini membutuhkan berbagai aplikasi yang memberikan wawasan
   d. Industri saat ini membutuhkan berbagai aplikasi yang memilki nilai bisnis yang lebih baik
4. Berikut ini adalah manfaat Data Science:
   a. Mendapatkan informasi yang berharga
   b. Organisasi memiliki keunggulan kompetitif

c. Kemampuan organisasi untuk merespon secara lebih efektif terhadap lingkungan bisnis yang berubah dengan cepat yang menuntut adaptasi terus-menerus.

d. Memperoleh berbagai aplikasi yang memberikan wawasan dan nilai bisnis yang lebih baik di perusahaan

5. Berikut ini yang merupakan bidang penerapan Data Science, adalah
   a. Industri E-commerce
   b. Industri Finance atau keuangan
   c. Industri Travel
   d. Industri Kesehatan
6. Dinamyc Pricing merupakan salah satu penerapan Data Sceince yang digunakan oleh perusahaan dengan menggunakan data untuk mensegmentasikan pelanggan atau konsumen secara akurat. Hal tersebut meruJpakan penerapan data sceince dalam bidang apa?
   a. Industri E-commerce
   b. Industri Finance atau keuangan
   c. Industri Travel
   d. Industri kesehatan
7. Berikut ini yang bukan merupakan pertanyaan penting tentang Data Science adalah
   a. Dimana data science diterapkan?
   b. Apa permintaan bisnis?
   c. Apa yang dibutuhkan bisnis?
   d. Siapa saja pemangku kepentingan dan apa kebutuhan individu mereka?
8. Berikut ini yang merupakan tahapan proses data sceince yang benar adalah
   1) Define Problem Statetment
   2) Analyze Data
   3) Ingest Data
   4) Evaluate Model

5) Refine Model
6) Prepare Data for Learning
7) Ship It

a. 1-3-2-6-4-5-7
b. 1-2-3-4-5-6-7
c. 1-3-2-4-6-5-7
d. 1-3-2-6-5-4-7

9. Berikut ini Siklus Data Science yang benar adalah
   1) Defenisi Masalah.
   2) Modeling Data yang layak
   3) Investigasi dan Pembersihan Data
   4) Implementasi(Operasi)
   5) Penerapan dan Penyempurnaan

   a. 1-2-3-4-5
   b. 2-1-3-4-5
   c. 3-2-1-5-4
   d. 1-3-2-5-4

10. Berikut ini yang merupakan tantangan pada Data Science adalah:
    a. Menghilangkan bias dalam kumpulan data dan aplikasi analitik.
    b. Mengelola penerapan model analitik, mengukur nilai bisnis, dan mempertahankan model.
    c. Menemukan data yang tepat untuk dianalisis.
    d. Mengkomunikasikan ide-ide yang kompleks dan membuat keputusan organisasi yang didorong oleh data.
11. Sebuah cabang ilmu yang mengggabungkan ilmu inferensi data, penggabungan algoritmik dan penggunaan teknologi untuk memecahkan masalah analitik yang kompleks, hal tersebut disebut dengan?
12. Saat ini, perusahaan-perusahaan di seluruh dunia semakin menyadari pentingnya data science, kecerdasan buatan, dan machine learning.

a. True
b. False

Daftar Pustaka

(IBM Cognitive Class-1, 2020) IBM Cognitive Class, Introduction to Data Science, https://cognitiveclass.ai/courses/data-science-101

Data Science for Everyone, https://learn.datacamp.com/courses/data-science-for-everyone

Data Science From Scratch first principal with Python, published by O'Reilly Media, Inc.,1005 gravenstein Highway North Sebatopol, CA

E-Modul Data Science, IlmudataPy.com

Pengantar Data Science dan Aplikasinya bagi Pemula Oleh Program Data Science, Informatika UNPAR, 2020

BAB II

BUSINESS UNDERSTANDING AND DATA COLLECTION

A. PENDAHULUAN

Mampu memahami bisnis(masalah), menjelaskan tentang latar belakang terjadinya masalah atau kendala pada bisnis tersebut dan mengusulkan solusi dengan menguraikan metodologi. Persyaratan data yang diperlukan dalam mendukung penyelesaian masalah, bagaiaman Teknik pengumpulan data dan mampu memahami dan menyelesaikan contoh studi kasus dengan pendekatan secara analitik.

B. INTI

1. Capaian Pembelajaran
    a. Mampu memahami dan menjelaskan Business Understanding and Data Collection

b. Mampu memahami dan menjelaskan pengantar bisnis

c. Mampu memahami dan menjelaskan persyaratan data yang dibutuhkan dalam menyelesaikan masalah

d. Mampu memahami bagaimana metode pengumpulan data

e. Mampu memahami dan menggunakan proses data science dalam menyelesaikan masalah

f. Mampu memahamu dan menjelaskan apa itu Business Undrestanding

g. Mampu memahami dan menyusun laporan keluar proyek untuk pelanggan(konsumen)

h. Mampu memahami contoh studi kasus Metode Pengumpulan Data

i. Mampu memahami studi kasus dengan pendekatan analitik

2. Materi

a. Pengantar bisnis

b. Persyaratan data

c. Metode pengumpulan data

d. Proses data science

e. Business Undrestanding

f. Laporan keluar proyek untuk pelanggan

g. Studi kasus Metode Pengumpulan Data

h. Pendekatan Analitik (Studi Kasus)

3. Uraian Materi

a. Pengantar bisnis

Sebuah bisnis didefinisikan sebagai sebuah organisasi atau badan usaha yang terlibat dalam kegiatan komersial, industri, atau profesional. Bisnis dapat berupa entitas nirlaba atau mereka dapat menjadi organisasi nirlaba yang beroperasi untuk memenuhi misi amal atau lebih jauh lagi untuk tujuan sosial.

Istilah "bisnis" juga mengacu pada upaya dan kegiatan terorganisir individu untuk memproduksi dan menjual barang dan jasa untuk keuntungan. Bisnis berkisar dalam skala dari kepemilikan tunggal hingga perusahaan internasional. Beberapa jalur teori terlibat dengan pemahaman administrasi bisnis termasuk perilaku organisasi, teori organisasi, dan manajemen strategis.

Umumnya, sebuah bisnis dimulai dengan konsep bisnis (ide) dan nama. Bergantung pada sifat bisnisnya, riset pasar yang ekstensif mungkin diperlukan untuk menentukan apakah mengubah ide menjadi bisnis itu layak dan apakah bisnis tersebut dapat memberikan nilai kepada konsumen. Nama bisnis dapat menjadi salah satu aset paling berharga dari sebuah perusahaan; pertimbangan yang cermat harus diberikan ketika memilihnya. Bisnis yang beroperasi dengan nama fiktif harus terdaftar di negara bagian.

Bisnis paling sering terbentuk setelah pengembangan rencana bisnis, yang merupakan dokumen formal yang merinci tujuan dan sasaran bisnis, dan strateginya tentang bagaimana hal itu akan mencapai tujuan dan sasaran. Rencana bisnis hampir penting ketika meminjam modal untuk memulai operasi.

Penting juga untuk menentukan struktur hukum bisnis. Tergantung pada jenis bisnisnya, mungkin perlu mendapatkan izin, mematuhi persyaratan pendaftaran, dan mendapatkan lisensi untuk beroperasi secara legal. Di banyak negara, korporasi dianggap badan hukum, yang berarti bahwa bisnis dapat memiliki properti, mengambil hutang, dan dituntut di pengadilan.

b. Persyaratan Data

Data yang salah, apabila digunakan sebagai dasar bagi pembuatan keputusan, akan menghasilkan keputusan yang salah. Persyaratan data yang baik, antara lain, objektif, representatif (mewakili kelompok data asalnya), memiliki kesalahan baku yang kecil, tepat waktu up to date) dan relevan.

1. Objektif

Data yang objektif berarti bahwa data harus sesuai dengan keadaan yang sebenarnya (as it is). Misalnya, produksi yang turun dilaporkan naik ini tidak objektif; harga satu satuan barang Rp500, dilaporkan Rp750, walaupun ada kuitansi, tetap tidak objektif.

2. Representatif (mewakili)

   Data harus mewakili objek yang diamati. Misalnya, jika laporan produksi padi dari suatu daerah hanya didasarkan atas hasil sawah-sawah yang subur saja, ini jelas tidak mewakili; laporan harga yang hanya didasarkan atas pasarpasar yang murah saja juga tidak mewakili; laporan konsumsi susu hanya dari golongan orang kaya saja juga tidak mewakili.

3. Kesalahan sampling (sampling error) kecil

   Suatu perkiraan (estimte) dikatakan baik (mempunyai tingkat ketelitian yang tinggi) apabila kesalahan samplingnya kecil.

Ketiga syarat tersebut di atas sering disebut syarat data yang dapat diandalkan (reliable). Sedangkan kedua syarat berikut lebih menunjukkan manfaat atau kegunaannya, yaitu:

1. Tepat waktu. Apabila data akan dipergunakan untuk melakukan pengendalian atau evaluasi, maka syarat tepat waktu ini penting sekali agar sempat dilakukan penyesuaian atau koreksi seperlunya kalau ada kesalahan atau penyimpangan yang terjadi di dalam implementasi suatu perencanaan.
2. Relevan. Data yang dikumpulkan harus ada hubungannya dengan masalah yang akan dipecahkan. Misalnya, pemerintah mengetahui faktor-faktor yang menyebabkan kemerosotan produksi padi selama beberapa tahun terakhir.

Data dapat dikelompokkan, antara lain: menurut sifat, sumber, cara memperoleh, dan waktu pengumpulan

Data menurut sifatnya dibedakan antara data kualitatif dan data kuantitatif. Data kualitatif adalah data yang tidak berbentuk angka (nonnumen's). Misalnya, produksi daging sapi meningkat, harga daging ayam mahal, penyaluran pupuk berjalan lancar, dan sebagainya. Data kuantitatif adalah data yang dinyatakan dalam bentuk angka.

Misalnya, produksi padi meningkat 10 persen, harga daging sapi per kilogram rata-rata adalah Rp 15.000, sebanyak 99 persen pupuk telah disalurkan, penduduk Indonesia pada tahun 1990 adalah 200 juta, dan sebagainya.

Metode kuantitatif, berurusan dengan sesuatu yang dapat dihitung. Lainnya bersifat kualitatif, artinya mereka mempertimbangkan faktor selain nilai numerik. Secara umum, kuesioner, survei, dan dokumen serta catatan bersifat kuantitatif, sedangkan wawancara, kelompok fokus, observasi, dan sejarah lisan bersifat kualitatif. Bisa juga terjadi persilangan antara kedua metode tersebut.

c. Metode Pengumpulan Data

Saat ini bisnis dan organisasi terhubung dengan klien, pelanggan, pengguna, karyawan, vendor, dan terkadang bahkan pesaing mereka. Data dapat menceritakan sebuah kisah tentang salah satu dari hubungan ini, dan dengan informasi ini, organisasi dapat meningkatkan hampir semua aspek operasi mereka.

Meskipun data bisa berharga, terlalu banyak informasi yang berat, dan data yang salah tidak berguna. Metode pengumpulan data yang tepat dapat berarti perbedaan antara wawasan yang berguna dan penyesatan yang membuang-buang waktu. Untungnya, organisasi memiliki beberapa alat untuk pengumpulan data primer. Metodenya berkisar dari tradisional dan sederhana, seperti wawancara tatap muka, hingga cara yang lebih canggih untuk mengumpulkan dan menganalisis data.

Berikut adalah enam metode pengumpulan data:

1. Wawancara
2. Kuesioner dan survei
3. Pengamatan
4. Dokumen dan catatan
5. Grup fokus
6. Sejarah lisan
7. Wawancara

Jika Anda bertanya kepada seseorang yang sama sekali tidak mengetahui analisis data tentang cara terbaik untuk mengumpulkan informasi dari orang-orang, jawaban yang paling umum mungkin adalah wawancara. Hampir semua orang dapat mengajukan daftar pertanyaan, tetapi kunci wawancara yang efisien adalah mengetahui apa yang harus ditanyakan. Efisiensi dalam wawancara sangat penting karena, dari semua metode pengumpulan data primer, wawancara langsung bisa menjadi yang paling mahal.

Ada beberapa cara untuk membatasi biaya wawancara, seperti melakukan wawancara melalui telepon atau melalui antarmuka obrolan web. Tapi terkadang wawancara langsung bisa sepadan dengan biayanya, karena pewawancara dapat menyesuaikan pertanyaan lanjutan berdasarkan tanggapan dalam pertukaran waktu nyata.

Wawancara juga memungkinkan untuk pertanyaan terbuka. Dibandingkan dengan metode pengumpulan data primer lainnya, seperti survei, wawancara lebih dapat disesuaikan dan responsif.

1. Pengamatan

   Observasi melibatkan pengumpulan informasi tanpa mengajukan pertanyaan. Metode ini lebih subjektif, karena mengharuskan peneliti, atau pengamat, untuk menambahkan penilaian mereka ke data. Namun dalam beberapa keadaan, risiko bias minimal.

   Misalnya, jika sebuah penelitian melibatkan jumlah orang di sebuah restoran pada waktu tertentu, kecuali jika pengamat menghitung dengan tidak benar, data tersebut harus cukup andal. Variabel yang mengharuskan pengamat untuk membuat perbedaan, seperti berapa banyak milenial mengunjungi restoran dalam periode tertentu, dapat menimbulkan potensi masalah.

   Pada umumnya observasi dapat menentukan dinamika suatu keadaan, yang pada umumnya tidak dapat diukur melalui teknik pengumpulan data lainnya. Observasi juga dapat dikombinasikan dengan informasi tambahan, seperti video.

2. Dokumen dan catatan

Terkadang Anda dapat mengumpulkan sejumlah data besar tanpa menanyakan apapun kepada siapapun. Penelitian berbasis dokumen dan catatan menggunakan data yang ada untuk penelitian. Catatan kehadiran, notulen rapat, dan catatan keuangan hanyalah beberapa contoh dari jenis penelitian ini.

Menggunakan dokumen dan catatan dapat menjadi efisien dan murah karena Anda sebagian besar menggunakan penelitian yang telah selesai. Namun, karena peneliti kurang memiliki kendali atas hasil, dokumen dan catatan dapat menjadi sumber data yang tidak lengkap.

3. Grup focus

Gabungan dari wawancara, survei, dan observasi, kelompok fokus adalah metode pengumpulan data yang melibatkan beberapa individu yang memiliki kesamaan. Tujuan dari grup fokus adalah untuk menambahkan elemen kolektif ke pengumpulan data individu. Studi kelompok fokus dapat meminta peserta untuk melihat presentasi, kemudian mendiskusikan konten sebelum menjawab pertanyaan bergaya survei atau wawancara.Kelompok fokus sering menggunakan pertanyaan terbuka seperti, "Bagaimana perasaan Anda tentang presentasi?" atau “Apa yang paling Anda sukai dari produk ini?” Moderator grup fokus dapat meminta grup untuk memikirkan kembali pengalaman bersama, daripada meneruskan ke masa depan.

Pertanyaan terbuka mendasarkan penelitian dalam keadaan pikiran tertentu, menghilangkan gangguan eksternal.

4. Sejarah lisan

Sekilas, sejarah lisan mungkin terdengar seperti wawancara. Metode pengumpulan data melibatkan mengajukan pertanyaan. Namun sejarah lisan lebih tepat diartikan sebagai pencatatan, pelestarian, dan interpretasi informasi sejarah berdasarkan pendapat dan pengalaman pribadi orang-orang yang terlibat dalam peristiwa tersebut. Tidak seperti wawancara dan survei, sejarah lisan terkait dengan satu fenomena. Sebagai contoh, seorang peneliti mungkin tertarik

untuk mempelajari pengaruh banjir pada suatu komunitas. Sejarah lisan dapat menjelaskan dengan tepat apa yang terjadi. Ini adalah pendekatan holistik untuk evaluasi yang menggunakan berbagai teknik.

Seperti dalam wawancara, peneliti dapat menjadi variabel pengganggu. Variabel pengganggu adalah variabel tambahan yang tidak diinginkan yang dapat mengubah hasil Anda dengan memperkenalkan bias dan menyarankan korelasi di mana tidak ada.

Contoh klasiknya adalah korelasi antara tingkat pembunuhan dan penjualan es krim. Kedua tokoh itu, pada satu waktu atau lainnya, bangkit bersama. Kesimpulan yang tidak ilmiah mungkin bahwa semakin banyak orang membeli es krim, semakin tinggi terjadinya pembunuhan.Namun, ada kemungkinan ketiga bahwa variabel tambahan mempengaruhi kedua kejadian ini. Dalam kasus es krim dan pembunuhan, variabel lainnya adalah cuaca. Cuaca yang lebih hangat adalah variabel pengganggu baik untuk tingkat pembunuhan dan penjualan es krim.

5. Kuesioner dan survei

Kuesioner dan survei dapat digunakan untuk mengajukan pertanyaan yang memiliki jawaban tertutup. Data yang dikumpulkan dari kuesioner dan survei dapat dianalisis dengan berbagai cara. Anda dapat menetapkan nilai numerik ke data untuk mempercepat analisis. Ini dapat berguna jika Anda mengumpulkan data dalam jumlah besar dari populasi yang besar.

Agar bermakna, survei dan kuesioner perlu direncanakan dengan cermat. Tidak seperti wawancara, di mana seorang peneliti dapat bereaksi terhadap arah jawaban responden, kuesioner yang dirancang dengan buruk tidak akan membawa penelitian ke mana-mana dengan cepat. Meskipun survei seringkali lebih murah daripada wawancara, survei tidak akan bernilai jika tidak ditangani dengan benar.

Survei dapat dilakukan sebagai wawancara, tetapi dalam banyak kasus, masuk akal untuk melakukan survei menggunakan formulir.

Formulir online adalah cara modern dan efektif untuk melakukan survei. Tidak seperti survei tertulis yang bersifat statis, pertanyaan yang disajikan dalam formulir online dapat berubah sesuai dengan bagaimana seseorang merespons berkat fitur formulir logika kondisional. Misalnya, jika Anda menggunakan JotForm untuk membuat formulir, ketika seseorang menjawab tidak untuk pertanyaan tentang alergi, mereka tidak perlu menggulir melewati semua pertanyaan lanjutan terkait tentang alergi tertentu. Sebaliknya, mereka akan langsung menjawab pertanyaan tentang topik yang berbeda.

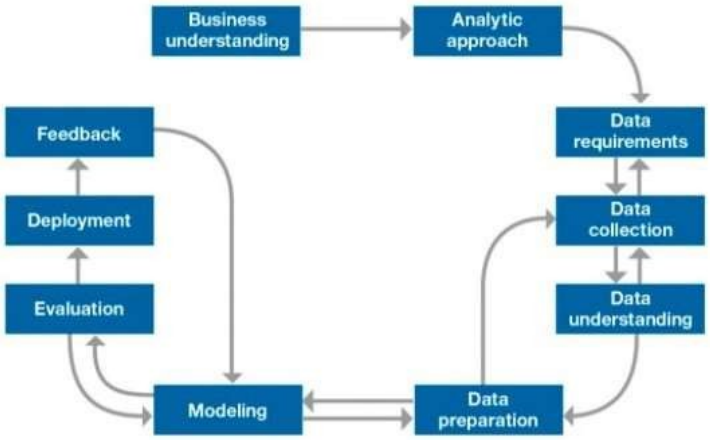
Pembuatan formulir modern juga menekankan pengumpulan data seluler, sehingga formulir dapat dengan mudah dilihat dan diisi di perangkat seluler.

Satu kekhawatiran ketika mengumpulkan data secara elektronik di UE adalah Peraturan Perlindungan Data Umum (GDPR) Uni Eropa. Peraturan yang baru diberlakukan ini memberikan perlindungan privasi kepada penduduk dan warga negara Uni Eropa dan dapat mengakibatkan denda yang mahal untuk ketidakpatuhan. Jika Anda ingin mempelajari lebih lanjut tentang cara memastikan formulir Anda mematuhi GDPR, JotForm memiliki semua informasi yang Anda butuhkan.

Metode pengumpulan data kualitatif vs kuantitatif

Metode kuantitatif, berurusan dengan sesuatu yang dapat dihitung. Lainnya bersifat kualitatif, artinya mereka mempertimbangkan faktor selain nilai numerik. Secara umum, kuesioner, survei, dan dokumen serta catatan bersifat kuantitatif, sedangkan wawancara, kelompok fokus, observasi, dan sejarah lisan bersifat kualitatif. Bisa juga terjadi persilangan antara kedua metode tersebut.

d. Proses Data Science



Gambar 15. Proses Data Science

Masing-masing tahapan dalam metodologi data science di atas dimaksudkan untuk menjawab 10 pertanyaan dasar seperti dibawah ini:

1. What is the problem that you are trying to solve?
2. How can you use data to answer the question?
3. What data do you need to answer the question?
4. Where is the data coming from (identify all sources) and how will you get it?
5. Is the data that you collected representative of the problem to be solved?
6. What additional work is required to manipulate and work with the data?
7. In what way can the data be visualized to get the answer that is required?
8. Does the model used really answer the initial question or does it need to be adjusted?
9. Can you put the model into practice?
10. Can you get constructive feedback into answering the question?

Adapun proses data science yang wajib dipahami sebagai berikut:

1. Business Understanding : dimulai dari menentukan masalah, tujuan, dan solusi dari perspektif bisnis
2. Analytic Approach: melakukan pendekatan dari masalah tadi, dengan teknik descriptive, predictive, atau perspective.

3. Data requirements: mendefinisikan data yang dibutuhkan
4. Data Collection: mengumpulkan data yang relevan dengan masalah, pengetahuan mengenai tools-tools data science diperlukan selain python tapi juga ada SQL dan Hadoop untuk proses pengambilan data.
5. Data Understanding: dengan proses visualisasi data dapat membantu untuk memahami kualitas data.
6. Data Preparation: kemampuan eksplorasi data sangat diperlukan seperti melakukan cleaning, transforming, selection dan normalize
7. Modelling: memilih dan membuat model yang paling optimal dengan algoritma machine learning, bisa klasifikasi, regresi, clustering, atau rekomendasi.
8. Evaluation: menilai tingkat keoptimalan dari proses modelling dengan matrix atau confusion matrix.
9. Deployment: merupakan tahapan implementasi dari rangkaian proses diatas
10. Feedback : Umpan balik tentang kinerja model

e. Business Understanding

Hal-hal yang dilkuakan dalam Business Understanding adalah menentukan masalah, tujuan, dan solusi dari perspektif bisnis. Kemudian mancari latar belakang pada bisnis tersebut dengan menentukan siapa kliennya, di domain bisnis apa klien itu, dan masalah bisnis apa yang akan coba atasi?. Langkah selanjutnya adalah menentukan cakupan yang terdapat dalam bisnis tersebut, diantaraya : mencari solusi data science apa yang coba dibangun?, Apa yang akan dilakukan?, Bagaimana hal tersebut akan dikonsumsi oleh pelanggan?

Kemudian memiliih personil yang terlibat pada bisnis tersebut. Siapa saja yang terlibat dalam proyek ini:

1. Microsoft : Pimpinan proyek, PM, Data Scientist, Manajer Akuntansi
2. Klien: administrator data, Kontak bisnis

3. Metrik: Apa tujuan kualitatif? (misalnya mengurangi churn pengguna). Apa yang dimaksud dengan metrik terukur (mis., kurangi fraksi pengguna yang tidak aktif selama 4 minggu). Hitung peningkatan apa dalam nilai metrik yang berguna untuk skenario pelanggan (mis., kurangi fraksi pengguna dengan tidak aktif selama 4 minggu sebesar 20%). Berapa nilai dasar (saat ini) dari metrik? (misalnya fraksi pengguna saat ini dengan tidak aktif selama 4 minggu = 60%). Bagaimana kita akan mengukur metrik? (misalnya pengujian A/B pada subset tertentu untuk periode tertentu; atau perbandingan kinerja setelah implementasi dengan baseline)
4. Rencana : Fase (tonggak sejarah), garis waktu, deskripsi singkat tentang apa yang akan kita lakukan di setiap fase.
5. Arsitektur :
    a. Data

       Data apa yang kita harapkan? Data mentah di sumber data pelanggan (misalnya file lokal, SQL, Hadoop lokal, dll.)
    b. Pergerakan data dari lokal ke Azure menggunakan ADF atau alat pergerakan data lainnya (Azcopy, EventHub, dll.) untuk memindahkan baik : semua datanya, setelah beberapa pra-agregasi di tempat, Data sampel cukup untuk pemodelan
    c. Alat dan penyimpanan data/sumber daya analitik apa yang akan digunakan dalam solusi misalnya: ASA untuk agregasi aliran; HDI/Hive/R/Python untuk konstruksi fitur, agregasi, dan pengambilan sampel, AzureML untuk pemodelan dan operasionalisasi layanan web
    d. Bagaimana skor atau layanan web yang dioperasionalkan (RRS dan/atau BES) digunakan dalam alur kerja bisnis pelanggan? Jika berlaku, tuliskan kode semu untuk API panggilan layanan web: Bagaimana pelanggan akan menggunakan hasil model untuk membuat keputusan?. Pipa pergerakan data dalam produksi. Buat diagram slide 1 yang menunjukkan aliran data dan arsitektur keputusan ujung ke ujung. Jika ada perubahan

substansial dalam alur kerja bisnis pelanggan, buatlah diagram sebelum/sesudah yang menunjukkan aliran data tersebut.

6. Komunikasi : Bagaimana kita akan tetap berhubungan? Rapat mingguan? Siapa orang yang dapat dihubungi di kedua sisi?

f. Laporan Keluar Proyek untuk Pelanggan

Petunjuk: Template untuk kriteria keluar untuk proyek data science. Ini adalah dokumen ringkas yang mencakup ikhtisar keseluruhan proyek, termasuk rincian setiap tahap dan pembelajaran. Jika bagian tidak berlaku (misalnya proyek tidak menyertakan model ML), cukup tandai bagian itu sebagai "Tidak berlaku". Panjang yang disarankan antara 5-20 halaman. Kode sebagian besar harus berada dalam repositori kode (bukan dalam dokumen ini).

Pelanggan: <Masukkan Nama Pelanggan>

Anggota Tim: <Masukkan nama anggota tim. Harap masukkan juga nama pihak terkait, seperti ketua tim, tim Akun, pemangku kepentingan Bisnis, dll.>

Gambaran <Ringkasan eksekutif dari seluruh solusi, tinjauan singkat non-teknis>

Domain Bisnis <Industri, domain bisnis pelanggan>

Masalah Bisnis <Masalah bisnis dan kasus penggunaan yang tepat, mengapa itu penting>

Pengolahan data <Skema kumpulan data asli, bagaimana data diproses, skema data input akhir untuk model>

Pemodelan, Validasi <Teknik pemodelan yang digunakan, hasil validasi, detail cara validasi dilakukan>

Arsitektur Solusi <Arsitektur solusi, jelaskan dengan jelas apakah ini benar-benar diterapkan atau arsitektur yang diusulkan. Sertakan diagram dan detail yang relevan untuk mereproduksi arsitektur serupa. Sertakan detail mengapa arsitektur ini dipilih versus arsitektur lain yang dipertimbangkan, jika relevan>

Manfaat

Manfaat Perusahaan (khusus internal. Periksa kembali apakah Anda ingin membagikan ini dengan pelanggan Anda) <Apa yang diperoleh perusahaan kami dari pertunangan ini? ROI, pendapatan, dll>

Manfaat Pelanggan <Apa manfaatnya (ROI, tabungan, peningkatan produktivitas, dll) bagi pelanggan? Jika hanya POC, berapa perkiraan ROI? Jika metrik yang tepat tidak tersedia, mengapa hal itu berdampak pada pelanggan?>

Pembelajaran

Eksekusi projek <Pelajaran seputar proses keterlibatan pelanggan>

Data Science / Rekayasa <Pembelajaran terkait data science/engineering, tips/trik, dll>

Domain <Pembelajaran seputar domain bisnis, >

Produk <Pembelajaran seputar produk dan layanan yang digunakan dalam solusi >

Apa yang unik dari proyek ini, tantangan khusus <Masalah atau pengaturan khusus, hal-hal unik, tantangan khusus yang harus diatasi selama keterlibatan dan bagaimana hal itu dicapai>

Tautan <Tautan ke studi kasus yang dipublikasikan, dll.; Tautan ke repositori git tempat semua kode berada>

Langkah selanjutnya <Langkah selanjutnya. Ini harus mencakup tonggak untuk tindak lanjut dan siapa yang 'memiliki' tindakan ini. Misal Post-Proof of Concept check-in status pada 1/12/2016 oleh X, rapat check-in bulanan oleh Y, dll.>

Lampiran <Materi lain yang tampaknya relevan – coba pertahankan non-lampiran hingga <20 halaman tetapi detail lebih lanjut dapat dimasukkan dalam lampiran jika diperlukan>

g. Studi Kasus Metode Pengumpulan Data

1. Topik mana yang Anda pilih untuk menerapkan metodologi data science?

ANS: Saya memilih kartu kredit dalam skenario dimana kita harus menggunakan informasi perusahaan kartu kredit untuk mengidentifikasi peluang bisnis. Selanjutnya, Anda akan berperan sebagai klien dan data scientist.

Dengan menggunakan topik yang Anda pilih, selesaikan tahap Pemahaman Bisnis dengan memunculkan masalah yang ingin Anda pecahkan dan menyusunnya dalam bentuk pertanyaan yang akan Anda gunakan datanya untuk dijawab.

Anda diminta untuk:

Jelaskan masalahnya, terkait dengan topik yang Anda pilih. Ungkapkan masalah sebagai pertanyaan yang harus dijawab menggunakan data. Misalnya, dengan menggunakan kasus penggunaan resep makanan yang dibahas di laboratorium, pertanyaan yang kami definisikan adalah, "Dapatkah kami secara otomatis menentukan masakan hidangan tertentu berdasarkan bahan-bahannya?".

ANS: Deskripsi masalah:

Menemukan peluang bisnis merupakan tantangan besar bagi pengusaha, kecenderungan alaminya adalah memulai bisnis terkait dengan yang sudah sukses, namun mungkin ada beberapa peluang yang terlewatkan untuk produk pelengkap.

Mengevaluasi situasi pasar dan mengidentifikasi potensi pertumbuhan berbagai jenis bisnis dari lapangan bisa sangat sulit dan dengan akurasi rendah, namun kartu kredit memiliki data konsumsi per bisnis yang dapat diklasifikasikan dan diurutkan berdasarkan beberapa kriteria, mereka juga memiliki informasi tentang setiap bisnis, produk mereka, lokasi, dll.

Tentu saja datanya tidak akan lengkap karena beberapa orang membayar tunai, tetapi selama kita mencari tren, kesenjangannya mungkin tidak relevan.

Pertanyaan:

Berdasarkan data konsumsi kartu kredit bisnis di sektor geografis tertentu, apakah mungkin untuk mengidentifikasi peluang bisnis?

Jelaskan secara singkat bagaimana Anda akan menyelesaikan setiap tahap berikut untuk masalah yang Anda jelaskan di tahap Pemahaman Bisnis, sehingga Anda akhirnya dapat menjawab pertanyaan yang Anda buat. (5 tanda): Pendekatan

Analitik, Persyaratan Data, Pengumpulan Data, Pemahaman dan Persiapan Pemodelan, dan Evaluasi Anda selalu dapat merujuk ke lab sebagai referensi dengan menjelaskan bagaimana Anda akan menyelesaikan setiap tahap untuk masalah Anda.

Pendekatan Analitik Selama kita mencari bisnis yang entah bagaimana bisa terkait, maka perlu menggunakan model deskriptif. Ini juga berguna untuk mendefinisikan cluster dalam area tertentu.

2. Persyaratan Data

Selalu ingat tujuan kami adalah untuk mengidentifikasi peluang bisnis terkait, kami tidak perlu mengumpulkan nama bisnis tetapi lokasi (alamat), jenis bisnis (makanan, masakan, olahraga, dll). Data historis konsumsi di setiap bisnis. Anda perlu mengumpulkan informasi dari beberapa sumber setidaknya yang paling umum digunakan atau diterima di sebagian besar bisnis. Juga akan diperlukan daftar produk per jenis industri dengan informasi penjualan sehingga bisnis yang dipilih sebelumnya dapat diperluas dalam produk terkait. Kartu kredit menggunakan informasi mereka untuk memberikan layanan penargetan ulang ke perusahaan lain, informasi rinci bisnis mungkin rahasia, namun mengkonsolidasikan informasi tanpa nama dan pendaftaran fiskal mungkin tersedia untuk tujuan komersial. Penting untuk menghubungi departemen B2B dan meminta informasi yang dijelaskan di atas, setidaknya selama 3 tahun terakhir sehingga kami dapat mengidentifikasi musim, dan tren pasar. Daftar produk dapat ditemukan di organisasi statistik nasional resmi mana pun. Di Ekuador adalah Bank Sentral yang mengumpulkan dan mempublikasikan data per produk yang diklasifikasikan oleh kode CIIU. Informasi ini seharusnya terlambat satu tahun, tetapi selama ada cukup data dari beberapa tahun, itu mungkin untuk mendapatkan pendekatan -to date- data. Mungkin ada dua kemungkinan skenario di mana ada peluang bisnis.

3. Pengumpulan data

Informasi dari kartu kredit harus dalam bentuk terstruktur tetapi akan diperlukan untuk mendapatkan konsumsi mingguan dan bulanan, selama informasi penjualan

produk dari Bank Sentral juga memiliki data lain yang dilampirkan pada setiap produk, maka informasi yang tidak diinginkan itu perlu dihilangkan. Data produk bisa dalam bentuk yang tidak terstruktur sehingga perlu untuk mendefinisikan bidang, tergantung pada jumlah baris yang dapat digunakan Python atau program lain di notebook Jupiter.

4. Pemahaman dan Persiapan Data

Beberapa data mungkin tidak lengkap sehingga kita harus mencari tahu apakah itu perlu atau jika kita bisa melewatkannya. Beberapa informasi lain dapat menghadirkan produk yang sama dengan nama yang berbeda sehingga kami ingin memperbaikinya. Akan ada di database produk banyak informasi produk yang berbeda seperti bensin yang kita tidak akan tertarik melihat daftar produk dan jumlah catatan akan membantu kita untuk berkonsentrasi pada mereka yang ingin kita analisis. Maka kita harus menghapus data yang tidak perlu.

5. Pemodelan dan Evaluasi

Selama kami mengerjakan model deskriptif, kami tidak memerlukan satu set pelatihan, namun mendapatkan angka yang layak dari konsumsi produk Terkini yang dihitung diperlukan karena kami bergantung pada informasi ini untuk mengambil keputusan yang tepat. 'berasumsi bahwa orang yang mengkonsumsi beberapa produk di area tertentu cenderung mengkonsumsi beberapa produk terkait, jadi proses pemodelan harus membawa kita untuk memperkirakan penjualan produk terkait yang disarankan berbeda di area yang ditentukan. Beberapa pertimbangan pasar lainnya mungkin muncul saat melakukan analisis sehingga prosesnya dapat didefinisikan ulang dan disesuaikan untuk mendapatkan informasi terbaik untuk membuat keputusan.

h. Pendekatan Analitik (Studi Kasus)

Pemahaman Bisnis & Pembingkaian Masalah

Cara cepat memahami konteks bisnis

Informasi Latar Belakang Tugas

PowerCo adalah utilitas gas dan listrik utama yang memasok ke pelanggan korporat, UKM (Usaha Kecil & Menengah), dan perumahan. Liberalisasi kekuatan pasar energi di Eropa telah menyebabkan churn pelanggan yang signifikan, terutama di segmen UKM. Mereka telah bermitra dengan BCG untuk membantu mendiagnosis sumber pelanggan UKM yang berputar.

Salah satu hipotesis yang dipertimbangkan adalah bahwa churn didorong oleh sensitivitas harga pelanggan dan memungkinkan untuk memprediksi pelanggan yang cenderung churn menggunakan model prediktif. Klien juga ingin mencoba strategi diskon, dengan kepala divisi UKM menyarankan bahwa menawarkan pelanggan dengan kecenderungan tinggi untuk menghasilkan diskon 20% mungkin efektif.

Lead Data Scientist (LDS) mengadakan rapat tim awal untuk membahas berbagai hipotesis, termasuk churn karena sensitivitas harga. Setelah berdiskusi dengan tim Anda, Anda telah diminta untuk lebih mendalami hipotesis bahwa churn didorong oleh sensitivitas harga pelanggan.

Tujuan

Tugas pertama adalah memahami apa yang terjadi pada klien dan memikirkan bagaimana Anda akan mendekati masalah seperti ini untuk menguji hipotesis spesifik ini.

Merumuskan hipotesis sebagai masalah ilmu data dan menjabarkan langkah-langkah utama yang diperlukan untuk menguji hipotesis ini. Komunikasikan pemikiran dan temuan Anda dalam email ke OSZA Anda, dengan fokus pada data potensial yang Anda perlukan dari klien dan model analitis yang akan Anda gunakan untuk menguji hipotesis semacam itu.

4. Forum Diskusi

   a. Dengan tim anda, lakukanlah pemahaman terhadap bisnis yang anda temukan disekitar anda, pilihlah metode pengumpulan data yang bagaiman yang dapat dilakukan berdasarkan situasi dan kondisi yang terjadi pada saat ini.

b. Berdasarkan proses data science pada gambar 7, jawablah 10 pertanyaan yang berhubungan dengan pemahaman bisnis yang kelompok(tim) anda temukan.

C. PENUTUP

1. Rangkuman

   Langkah selanjutnya adalah memahami machine learning. Machine learning sendiri merupakan sub bidang kecerdasan buatan yang berkaitan dengan pemodelan algoritma agar sistem dapat belajar sendiri berdasarkan data mentah. Seperti yang sudah disinggung sebelumnya, pada tahap modelling kita harus memilih algoritma machine learning yang paling optimal.

2. Tes Formatif

   Sebagai evaluasi pemahaman materi pada Bab 2, jawalah pertanyaan berikut ini.

   1. Mengidentifikasi peluang bisnis terkait, hal apa yang perlu dikumpulkan selain megumpulkan nama bisnis
   2. Hal-hal yang dilakukan dalam busniness undertanding adalah;
      a. menentukan masalah
      b. menentukan tujuan
      c. mencari solusi dari perspektif bisnis
      d. menentukan pengaruh masalah terkait
   3. Berikut ini yang merupakan metode pengumpulan data adalah
      a. Wawancara
      b. Kuesioner dan survei
      c. Observasi
      d. Focus Group Discussion (FGD)
   4. Data yang salah, apabila digunakan sebagai dasar bagi pembuatan keputusan, akan menghasilkan keputusan yang salah. Persyaratan data yang baik, antara lain:
      a. Objektif
      b. representatif (mewakili kelompok data asalnya)

c. memiliki kesalahan baku yang kecil
d. tepat waktu up to date) dan relevan.

5. Langkah awal yang dilakukan pada proses Data Science adalah:
   a. Business understanding
   b. data preparation
   c. modelling
   d. analytic approach
6. Cara modern dan efektif untuk melakukan survei adalah
7. Sebuah organisasi atau badan usaha yang terlibat dalam kegiatan komersial, industri, atau profesional. Hal tersebut disebut dengan apa?
8. Selama mengerjakan model deskriptif, tidak memerlukan satu set pelatihan, namun mendapatkan angka yang layak dari konsumsi produk Terkini yang dihitung diperlukan karena bergantung pada informasi untuk mengambil keputusan yang tepat.
   a. True
   b. False
9. Pendekatan Analitik Selama kita mencari bisnis yang entah bagaimana bisa terkait, maka perlu menggunakan model deskriptif
   a. True
   b. False
10. Bisnis paling sering terbentuk setelah pengembangan rencana bisnis, yang merupakan dokumen formal yang merinci tujuan dan sasaran bisnis, dan strateginya tentang bagaimana hal itu akan mencapai tujuan dan sasaran.
    a. True
    b. False

Daftar Pustaka

Data Science for Business https://learn.datacamp.com/courses/data-science-for-business

Data Science for Everyone, https://learn.datacamp.com/courses/data-science-for-everyone

Data Science From Scratch first principal with Python, published by O'Reilly Media, Inc.,1005 gravenstein Highway North Sebatopol, CA

E-Modul Data Science, IlmudataPy.com

Pengantar Data Sceince dan Aplikasinya bagi Pemula Oleh Program Data Science, Informatika UNPAR, 2020

BAB III

STATISTIK FOR DATA SCIENCE

A. PENDAHULUAN

Mampu memahami menerapakan fungsi statistik pada data science dalam melakukan pengolahan data berdasarkan dataset yang dikumpulkan dalam mencari nilai yang paling optimal sebagai ketentuan dalam pemilihan solusi berdasarkan masalah yang difahami.

B. INTI

1. Mampu memahami dan menjelaskan Statistik for data science
    - **a.** Mampu memahami dan menjelaskan apa itu perbedaan statistik, statistika dan data science
    - **b.** Mampu memahami dan menjelaskan tipe-tipe data statistik
    - **c.** Mampu memahami dan menjelaskan populasi dan sample data pada statistika
    - **d.** Mampu memahami dan menjelaskan apa itu statistik deskriptif
    - **e.** Mampu memahami dan menjelaskan bagaimana visualisasi data dan macam-macam grafik dalam statistika
    - **f.** Mampu memahami dan menjelaskan apa itu korelasi dan regresi

2. Materi
    - **a.** Perbedaan statistik, statistika dan data science
    - **b.** Tipe-tipe data statistik
    - **c.** Populasi dan sample data pada statistika
    - **d.** Statistik deskriptif
    - **e.** Visualisasi data dan macam-macam grafik dalam statistika
    - **f.** Korelasi dan regresi

3. Uraian Materi

a. Perbedaan statistik, statistika dan data science

Statistik, adalah Kumpulan data yang bisa memberikan gambaran tentang keadaan sampel. Statistika, adalah Ilmu yang mempelajari tentang statistik. Bagaimana mengumpulkan, mengolah, menganalisis, menyajikan hingga menarik atau menghasilkan kesimpulan atas data yang besar. Statistika pada umumnya mempelajari masa lalu atau pengalaman logika pengambilan keputusan di masa depan. Contoh penerapan statistika adalah seperti prakiraan cuaca, google predictive search dan sebagainya.

Lalu apa hubungannya dengan Data Science? Data Science adalah pembelajaran tingkat lanjut setelah statistik dan statistika.

Ada tiga ilmu dasar Data Science

1. Statistik
2. Programming
3. Keahlian di bidang tertentu

Secara umum alur dari cara kerja Data Science ada 5 tahapan yaitu:

1. Data Collection
2. Data Exploration
3. Model Training dan Testing
4. Visualisasi dan Interpretasi
5. Model Implementasi

Dan Ilmu Statistika berperan pada tiga poin di atas yaitu tiga poin di atas yaitu Data Exploration, Model Training dan Testing serta Visualisasi dan Interpretasi

b. Tipe-tipe Data Statistik

Pada umumnya data dikelompokkan berdasarkan :

1. Jenis
2. Cara Memperoleh data, dan
3. Cara pengukurannya

Tipe data berdasarkan jenisnya dibedakan menjadi 2 kategori, yaitu Data Kategorikal dan Data Numerikal.

1. Data Kategorikal, adalah data yang berupa kelompok dengan skala yang digunakan yaitu skala nominal dan sekala ordinal. Skala Nominal, adalah kelompok data yang kategorinya tidak dapat diurutkan (misal, data pekerjaan dan data golongan darah). sedangkan Skala Ordinal adalah kelompok data yang kategorinya bisa diurutkan (misal, Tingkat pendidikan dan jenis stadium penyakit).
2. Data Numerikal, adalah data yang didapatkan dari hasil pengukuran dengan menggunakan skala interval dan skala rasio . Skala Interval, adalah skala yang berdasarkan satuan unit yang sama dan konstan. sedangkan Skala Rasio, adalah data dengan interval 0 adalah mutlak (contoh, Berat dalam Kg dan Kwintal memilki titik nol yang sama atau mutlak).

Tipe data berdasarkan cara perolehannya:

1. Data primer adalah data yang diperoleh secara langsung dari obyak yang diteliti baik secara individu maupun kelompok/organisasi.
2. Data sekunder adalah data yang diperoleh secara tidak langsung untuk mendapatkan informasi/keterangan dari obyek yang diteliti.
3. Data tersier yaitu data yang diperoleh secara tidak langsung dari obyek yang diteliti biasanya data tersebut diperoleh dari pihak ketiga baik dari individu maupun kelompok yang sengaja mengungkapkan fakta dari pihak kedua.

Tipe data berdasarkan cara pengukurannya

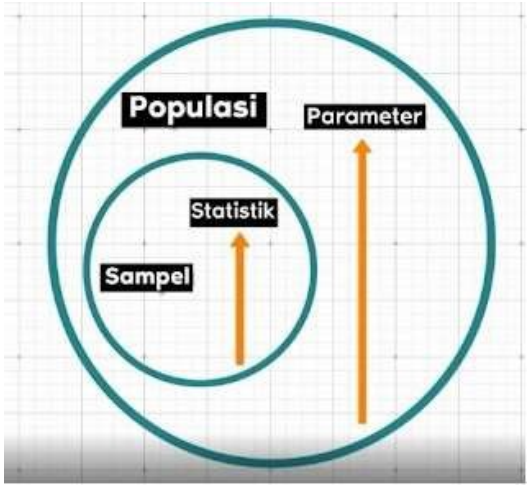
1. Data kualitatif adalah data yang dinyatakan dalam bentuk bukan angka (sifat).

2. Data kuantitatif adalah data yang dinyatakan dalam bantuk angka yang diasumsikan sebagai informasi dalam bentuk pernyataan “bilangan” yang didasarkan pada hasil perhitungan

c. Populasi dan Sampel Pada Statistika

Populasi, adalah Himpunan keseluruahn objek yang sedang diamati. Nilai yang dihitung dari populasi dan memberi deskripsi pada populasi disebut Parameter. Sampel, adalah Himpunan bagian dari populasi yang bertindak sebagai perwakilan dari populasi. Nilai yang dihitung dari sampel disebut statistik.

Berikut adalah contoh dari hubungan populasi dan sampel dalam statistika



Gambar 16. Hubungan populasi dan sampel dalam statistika

Penjelasan Gambar 8 adalah ;

Dalam Contoh, kita memiliki Populasi Mahasiswa dalam suatu perguruan tinggi, dan memiliki Parameter nilai rata-rata tinggi badan setiap mahasiswa dikampus tersebut. sementara kita juga memiliki Sampel Mahasiswa pada salah satu program studi di kampus tersebut, selain itu kita juga memiliki nilai rata-rata tinggi badan setiap mahasiswa di salah satu Program studi tersebut.

Dalam ilmu Statistika untuk memilah atau menentukan nilai-nilai Populasi dan Sampel kita juga perlu melakukan yang namanya Sampling . Kenapa sampling itu diperlukan

1. Karena Ukuran Populasi

2. Masalah Biaya (Sampling lebih sedikit memakan biaya, apalagi jika ukuran populasinya besar)

3. Masalah Waktu (Sampling akan memberikan data dan hasil lebih cepat)

4. Jika terdapat percobaan yang merusak

Teknik Sampling dalam ilmu statistika

1. Simple Random Sampling

   Penarikan sampel acak sederhana (simple random sampling) merupakan pengambilan sampel dari populasi secara acak tanpa memperhatikan strata yang ada dalam populasi dan setiap anggota populasi memiliki kesempatan yang sama untuk dijadikan sampel. Dua cara sampel acak sederhana yaitu:

   a. Sistem kocokan yaitu sistem sampel acak sederhana dengan cara sama sistem arisan.

   b. Menggunakan tabel acak yaitu memilih sampel dengan menggunakan suatu tabel. Dalam penggunaannya ditentukan dahulu titik awal (starting point).

   Keuntungan

   a. Memerlukan pengetahuan minimum dari populasi

   b. Bebas dari kesubjektifan dan bebas dari kesalahan personal

   c. Memberikan data yang sesuai untuk tujuan kita

   d. Observasi dari suatu sampel dapat digunakan untuk tujuan inferensial

   Kekurangan

a. Keterwakilan dari sampel tidak dapat dipastikan pada metode ini

b. Metode ini tidak menggunakan pengetahuan tentang populasi

c. Keakuratan inferensial dari penemuan tergantung pada ukuran sampel

2. Stratified Random Sampling

   Penarikan sampel acak terstruktur (stratified random sampling) dilakukan dengan membagi anggota populasi dalam beberapa sub kelompok yang disebut strata, lalu suatu sampel dipilih dari masing-masing stratum. Contoh kasus berdasarkan usia

dari 1000 populasi pemilih pada pemilu akan diambil 100 orang sebagai sample berdasarkan usia pemilih secara proporsional.

Penyelesaian :

Tabel 1. Usia pemilu berdasarkan usia

Usia Pemilih	**Populasi**	**Jumlah/ ditentukan**	**Terpilih**
15 – 25	100	10 %	10
26 – 35	200	10 %	20
36 – 45	500	10 %	50
45 >	200	10 %	20
	1000		100

Keuntungan

a. Perbaikan dari metode yang sebelumnya

b. Ini adalah metode objektif sampling

c. Observasi dapan digunakan untuk tujuan inferensial

Kekurangan

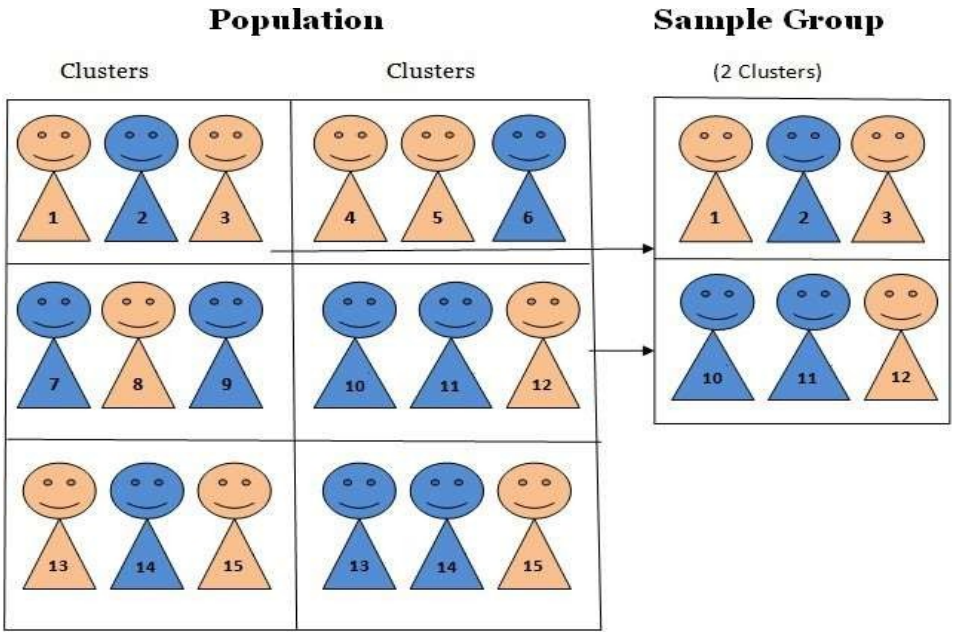
a. Kekurangan berat dari metode ini adalah sulit bagi peneliti untuk memutuskan kriteria relevan untuk stratifikasi

b. Hanya satu kriteria yang dapat digunakan untuk stratifikasi, tapi itu secara umum tampak lebih dari satu kriteria relevan untuk stratifikasi

c. Banyak memakan biaya dan waktu

3. Clustering Sampling

Penarikan sampel cluster (cluster sampling) dilakukan dengan membagi anggota populasi dalam beberapa kelompok, lalu suatu sampel dipilih dari masing-masing area kelompok.



Gambar 17. Population Sample Group

Contoh cluster random sampling : seorang peneliti ingin membuat survei mengenai performa akademis siswa SMA di Medan. Berikut ini langkah-langkah yang dapat diambil:

a. Peneliti membagi populasi siswa SMA Medan menjadi beberapa kelompok atau cluster berdasarkan kecamatan;

b. Kemudian peneliti memilih beberapa cluster sesuai dengan penelitian yang sedang dilakukan melalui pemilihan sampel acak yang sistematis;

c. Lalu dari beberapa cluster kecamatan yang telah dipilih secara acak, peneliti dapat memilih untuk memasukkan semua siswa SMA sebagai subjek, atau memilih beberapa subjek dari tiap cluster kecamatan melalui random sampling yang sistematis.

Keuntungan

a. Merupakan perwakilan yang baik dari populasi

b. Metode yang mudah

c. Metode yang ekonomis

d. Praktis dan sangat berlaku pada penelitian

e. Hasil penelitian dapat digunakan untuk tujuan inferensial

Kekurangan

a. Cluster sampling tidak bebas dari kesalahan

b. Tidak menyeluruh, jadi semua teknik diatas merupakan probabilitas sampling

4. Systematic Random Sampling

Penarikan sampel sistematis (systematic sampling) yaitu penarikan sampel apabila setiap unsur atau anggota dalam populasi disusun dengan cara tertentusecara alfabetis, dari besar kecil atau sebaliknya-kemudian dipilih titik awal secara acak lalu setiap anggota ke K dari populasi dipilih sebagai sampel

Contoh Systematic Sampling

Untuk memahami pengertian systematic random sampling, mari kita cerna langkah-langkah berikut ini.

a. Berikan nomor pada setiap penyusun populasi. Misalnya ada 50 orang dalam populasi. Beri label angka 1 sampai 50 pada anggota populasi;

b. Tentukan ukuran sampel yang Anda perlukan. Misalnya Anda memerlukan 10 orang;

c. Bagi populasi berdasarkan ukuran sampel Anda. Pada contoh systematic random sampling ini berarti 50 dibagi ukuran sampel 10. 50/10 = 5. Maka angka 5 ini menjadi sampling interval pada penelitian Anda. Di bawah ini dapat Anda lihat sampel yang dipilih mengikuti sampling interval 5, maka pilihlah anggota setiap kelipatan 5.

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50

Keuntungan

a. Merupakan metode yang sederhana untuk memilih sampel

b. Meminimalisir biaya (saat ke lapangan)

c. Menggunakan statistik inferensial

d. Komperhensif (menyeluruh) dan mewakili populasi

e. Observasi sampel dapat digunakan untuk menggeneralisasikan dan menarik kesimpulan

Kekurangan

a. Tidak bebas dari kesalahan karena subjektifitas mengarah pada cara yang berbeda dari daftar sistematis oleh individu yang berbeda. Pengetahuan tentang populasi sangat penting.

b. Informasi dari setiap individu sangat penting

c. Metode ini tidak dapat menjamin keterwakilan

d. Adanya resiko dalam menggambarkan kesimpulan dari penelitian sampel

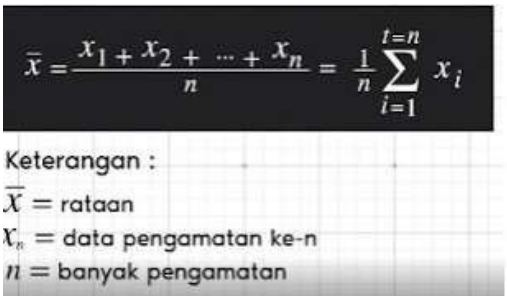
d. Pengertian Statistik Deskriptif

Statistika Deskriptif. adalah Analisis untuk menyajikan dan meringkas data sehingga menjadi mudah dipahami. Ada 2 Jenis Stistika Deskripstif yaitu Ringkasan Numerik dan Visualisasi Data.

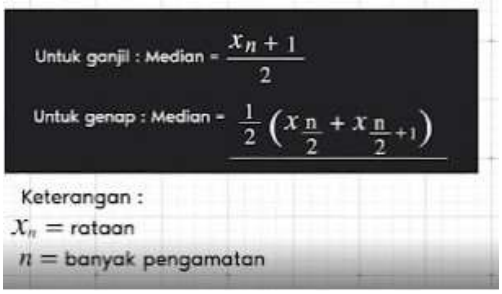
1. Ringkasan Numerik, adalah cara meringkas data menjadi ukuran numerik Misalnya sering kita temui dalam ilmu matematika yaitu Mean dan Median. Secara garis besar ringkasan numerik terbagi menjadi 2 bagian yaitu pemusatan data dan penyebaran data.

    a. Pemusatan Data, adalah Ringkasan numerik yang menggambarkan letak di mana data berpusat. Rumus yang sering digunakan adalah Mean, Median dan Modus.

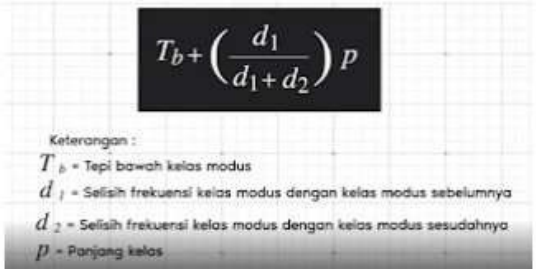
        1. Mean, adalah Jumlah semua data dibagi banyaknya data atau bisa disebut dengan rata-rata. rumusnya sebagai berikut :



2. Median, adalah nilai yang berada ditengah sekumpulan data yang telah diurutkan, dari data terkecil hingga terbesar. rumusnya sebagai berikut :



3. Modus, adalah nilai yang paling sering muncul dalam kumpulan data. suatu data mungkin memilki modus lebih dari satu. rumunya sebagai berikut :



b. Penyebaran Data, adalah ringkasan numerik yang menunjukkan seberapa jauh data menyebar dari rata-rata. Rumus yang biasa digunakan adalah Variansi dan Standar Deviasi

1. Variansi (Ragam), adalah Nilai yang menggambarkan variabilitas data terhadap rata-ratanya. rumusnya sebagai berikut:

$$\sigma^2 = \left[\sum (x_i - \bar{x})^2\right] / n$$

Keterangan :
σ^2 = Variasi
$\sum$ = Jumlah data
x_i = Data ke i
i = 1,2,....,n
$\bar{x}$ = Rata rata keseluruhan data
n = Banyak data

2. Standar Deviasi. adalah cara untuk mengetahui sebaran data. DImanfaatkan untuk mengetahui sampel data yang diambil mewakili populasi atau tidak. Rumusnya sebagai berikut :

$$\sigma = \sqrt{\left[\sum (x_i - \bar{x})^2\right] / n}$$

Keterangan :
σ = Variasi
$\sum$ = Jumlah data
x_i = Data ke i
i = 1,2,....,n
$\bar{x}$ = Rata rata keseluruhan data
n = Banyak data

e. Visualisasi Data dan Macam-macam Grafik dalam Statstik

Visualisasi data dalam ilmu statistika sangat diperlukan untuk menjadikan data agar lebih mudah dibaca, dimengerti, dan yang paling penting dipresentasikan. Salah satu cara mem visualisasikan data biasa kita temui yaitu dengan Grafik dan Tabel.

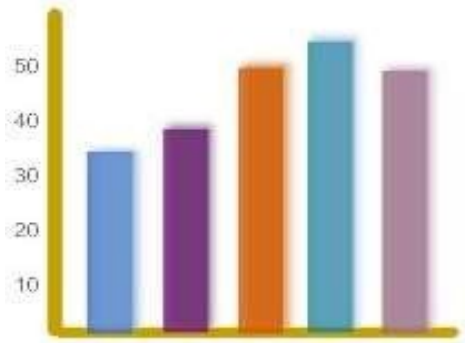
Grafik, adalah alat untuk menyajikan informasi agar terlihat dinamis, informatif, dan menarik. sedangkan table adalah penyajian data yang disesuaikan dengan tujuan dan kompleksitas data.

Macam-macam Grafik dalam Statistika yaitu :

1. Grafik Perbandingan

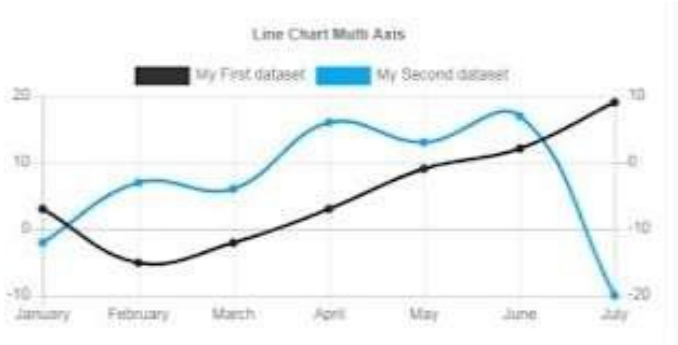
   Grafik Perbandingan terdiri dari dua kategori yaitu Bar Chart dan Line Chart

   a. Bar Chart (Diagram Batang) terdiri dari sedikit kategori, bentuknya mudah dipahami dan sifatnya fleksibel atau menyajikan banyak hal.



Gambar 18. Contoh Bar Chart(Diagram batang)

b. Line Chart (Diagram Garis), digunakan untuk non-cyclical data atau tidak mengandung unsur siklus dan digunakan untuk menyajikan data yang berkesinambungan atau data yang bersifat kontinu. contoh

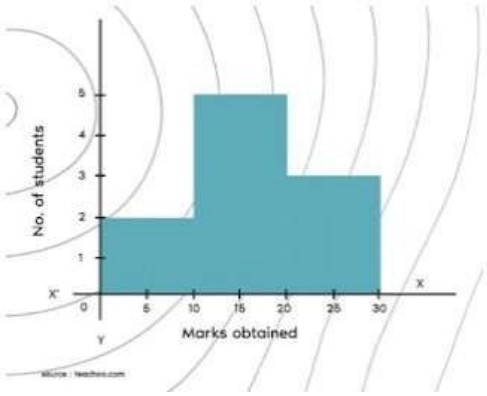


Gambar 19. Line Chart (Diagram Garis)

2. Grafik Distribusi

Grafik distribusi terdiri dari Bar Histogram dan Line Histogram.

a. Bar Histogram terdiri dari atas satu variabel dan memiliki sedikit data poin yang berfungsi untuk menunjukkan frekuensi.



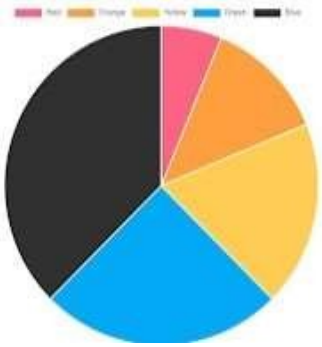
Gambar 20. Bar Histogram

b. Line Histogran terdiri atas Satu variabel, memiliki banyak data poin dan memudahkan untuk melihat apakah data berdistribusi normal atau tidak.

3. Grafik Komposisi

Grafik Komposisi dilihat dari fungsinya terdiri dari grafik statis dan grafik yang berganti setiap waktu. Grafik Statis terdiri dari dua macam yaitu pie chart dan waterfall chart.

a. Pie Chart berbentuk lingkaran yang tidak memakan banyak tempat. Karateristiknya adalah bisa memperlihatkan perbandingan ukuran data besar sektornya secara langsung dan tidak memperlihatkan ukuran atau frekuensi.



Gambar 21. Pie Chart

b. Waterfall Chart atau biasa dikenal dengan mario chart merupakan grafik yang menggambarkan bagaimana nilai awal dipengaruhi secara positif

dan negatif oleh berbagai faktor. Ini sangat berfungsi dibidang akuntansi.

Gambar 22. Waterfall Chart

Grafik yang berganti setiap waktu mempunyai 4 jenis grafik yaitu

1. Stacked Bar Chart merupakan grafik yang memiliki sedikit periode dan bisa membandingkan relative differences dan absolute differences. seperti menggambarkan banyak populasi global berdasarkan negaranya yang dibedakan menjadi 3 warna dataset .

Gambar 23. Stacked Bar Chart

2. Stacked 100% Bar Chart merupakan grafik yang sama dengan Stacked Bar Chart bedanya tanpa absolute differences.
3. Stacked Area Bar Chart, grafik yang memiliki banyak periode dan bisa membandingkan relative differences dan absolute differences. Karateristinya

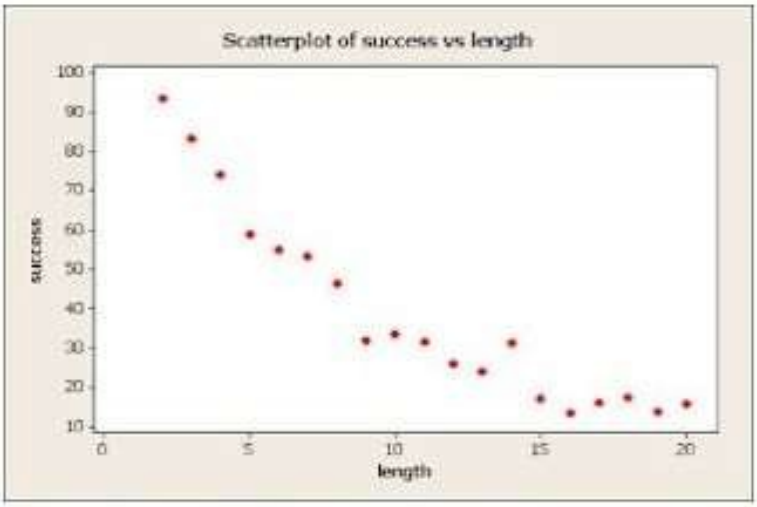
adalah setiap garis mewakilkan kategori yang berbeda, data muda dibandingakan dan biasanya area di bawah tiap garis diberi warna yang berbeda. contoh kasar sebagai berikut :

Gambar 24. Stacked Area Bar Chart

4. Grafik Hubungan

Terdiri dari dua grafik yaitu Scatter plot & Scatter plot bubble size

a. Scatter Plot, merupakan grafik dua variabel yang biasa digunakan untuk membandingakan pasangan nilai antar variabel saling terkait atau tidak. Memiliki Karateristik Jika kedua variabel saling berhubungan jika titik-tiknya bisa membentuk garis lurus dan ini biasa digunakan dalam Analisis Regresi.



Gambar 25. Scatter Plot

b. Scatter plot bubble size hampir mirip dengan yang diatas, hanya saja memeiliki karateristik menampilkan pengaruh variabel melalui volume dan warna bubble. Dan umumnya terdiri dari 3 variabel atau lebih

f. Korelasi dan Regresi

1. Korelasi

Secara sederhana, korelasi dapat diartikan sebagai hubungan. Namun ketika dikembangkan lebih jauh, korelasi tidak hanya dapat dipahami sebatas pengertian tersebut. Korelasi merupakan salah satu teknik analisis dalam statistik yang digunakan untuk mencari hubungan antara dua variabel yang bersifat kuantitatif. Hubungan dua variabel tersebut dapat terjadi karena adanya hubungan sebab akibat atau dapat pula terjadi karena kebetulan saja.

Pola/ Bentuk Hubungan antara dua variabel :

a. Korelasi linier positif (+)

Perubahan salah satu Nilai Variabel diikuti perubahan Nilai Variabel yang lainnya secara teratur dengan arah yang sama. Jika Nilai Variabel X mengalami kenaikan, maka Variabel Y akan ikut naik. Jika Nilai Variabel X mengalami penurunan, maka Variabel Y akan ikut turun.

Apabila Nilai Koefisien Korelasi mendekati +1 (positif Satu) berarti pasangan data Variabel X dan Variabel Y memiliki Korelasi Linear Positif yang kuat/Erat.

b. Korelasi linier negatif (-)

Perubahan salah satu Nilai Variabel diikuti perubahan Nilai Variabel yang lainnya secara teratur dengan arah yang berlawanan. Jika Nilai Variabel X mengalami kenaikan, maka Variabel Y akan turun. Jika Nilai Variabel X mengalami penurunan, maka Nilai Variabel Y akan naik.

Apabila Nilai Koefisien Korelasi mendekati -1 (Negatif Satu) maka hal ini menunjukan pasangan data Variabel X dan Variabel Y memiliki Korelasi Linear Negatif yang kuat/erat.

c. Tidak Berkorelasi (0)

Kenaikan Nilai Variabel yang satunya kadang-kadang diikut dengan penurunan Variabel lainnya atau kadang-kadang diikuti dengan kenaikan Variable yang lainnya. Arah hubungannya tidak teratur, kadang-kadang searah, kadang-kadang berlawanan.

Apabila Nilai Koefisien Korelasi mendekati 0 (Nol) berarti pasangan data Variabel X dan Variabel Y memiliki korelasi yang sangat lemah atau berkemungkinan tidak berkorelasi.

Korelasi Sederhana merupakan suatu Teknik Statistik yang dipergunakan untuk mengukur kekuatan hubungan 2 Variabel dan juga untuk dapat mengetahui bentuk hubungan antara 2 Variabel tersebut dengan hasil yang sifatnya kuantitatif. Kekuatan hubungan antara 2 variabel yang dimaksud disini adalah apakah hubungan tersebut ERAT, LEMAH, ataupun TIDAK ERAT sedangkan bentuk hubungannya adalah apakah bentuk korelasinya Linear Positif ataupun Linear Negatif.

Kekuatan Hubungan antara 2 Variabel biasanya disebut dengan Koefisien Korelasi dan dilambangkan dengan symbol "r". Nilai Koefisian r akan selalu berada di antara -1 sampai +1.

Rumus Pearson Product Moment

Koefisien Korelasi Sederhana disebut juga dengan Koefisien Korelasi Pearson karena rumus perhitungan Koefisien korelasi sederhana ini dikemukakan oleh Karl Pearson yaitu seorang ahli Matematika yang berasal dari Inggris.

Rumus yang dipergunakan untuk menghitung Koefisien Korelasi Sederhana adalah sebagai berikut :

(Rumus ini disebut juga dengan Pearson Product Moment)

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{\{n\Sigma x^2 - (\Sigma x)^2\}\{n\Sigma y2 - (\Sigma y)2\}}}.$$

Dimana :

n = Banyaknya Pasangan data X dan Y

Σx = Total Jumlah dari Variabel X

Σy = Total Jumlah dari Variabel Y

Σx^2= Kuadrat dari Total Jumlah Variabel X

Σy^2= Kuadrat dari Total Jumlah Variabel Y

Σxy= Hasil Perkalian dari Total Jumlah Variabel X dan Variabel Y

2. Regresi Linier

Analisis Regresi Sederhana adalah sebuah metode pendekatan untuk pemodelan hubungan antara satu variabel dependen dan satu variabel independen. Dalam model regresi, variabel independen menerangkan variabel dependennya. Dalam analisis regresi sederhana, hubungan antara variabel bersifat linier, dimana perubahan pada variabel X akan diikuti oleh perubahan pada variabel Y secara tetap.

Regresi Linier merupakan suatu alat ukur yang dapat digunakan untuk mengetahui adanya korelasi antara beberapa variabel. Dalam Regresi Linier ada beberapa hal yang harus dipahami diantaranya Variabel Terikat, Variabel Bebas, Konstanta dan Koefisien Regresi. Kalau ditinjau keakurasiannya dalam pemecahan sebuah kasus, regresi memiliki tingkat akurasi yang lebih baik di dalam konsep analisis sebuah hubungan antara 1(satu) variabel dengan variabel lainnya .

Untuk fungsi regresi terdapat beberapa rumus yang dapat dijadikan untuk pembentukan rumus regresi baru yaitu:

Y = a+bX

Keterangan:

Y = variabel terikat

X = variabel bebas

a = konstanta (intersep)

b = koefisien Regresi (slop)

Untuk mencari nilai a (konstanta) dan b(koefisien Regresi) maka ada beberapa rumus yang dapat digunakan yaitu:

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y}{n(\Sigma x^2) - (\Sigma x)^2}$$

Dimana:

n = Banyaknya pasangan data X dn Y

ΣX =Total jumlah dari variabel X

ΣY = Total jumlah dari variabel Y

ΣX^2 = Total jumlah kuadrat variabel X

ΣY^2 = Total jumlah kuadrat variabel Y

ΣXY = Total jumlah Perkalian variabel X dan Y

4. Forum Diskusi

a. Berdasarkan data yang tersedia sebelumnya, dengan menggunakan software Ms. Excel lakukan perhitngan untuk mencari nilai unsur yang terdapada korelasi dan regresi

b. Hitunglah nilai unsur-unsur yang terdapat pada pemusatan data

C. PENUTUP

1. Rangkuman

   Pada Pertemuan Topik 6 ini, kita telah mempelajari mengenai perbedaan statistik, statistika dan data Science. Tipe - tipe data statistik, populasi dan sampel statistik, statistik deskriptif, visualisasi data serta Korelasi dan Regresi Linier

   Setelah mempelajari ini diharapkan mahasiswa dapat melihat hubungan antara statistik dengan data Science

2. Tes Formatif

Sebagai evaluasi pemahaman materi pada Bab 3, jawablah pertanyaan berikut ini:

1. Keuntungan ke lima pada cluster sampling adalah....
   a. Hasil penelitian dapat digunakan untuk tujuan inferensial
   b. Merupakan perwakilan yang baik dari populasi
   c. Metode yang mudah
   d. Praktis dan sangat berlaku pada penelitian
2. Berapakah macam dalam grafik dalam statistika?
   a. 4
   b. 3
   c. 5
   d. 2
3. Tipe Data berdasarkan cara perolehannya terdiri atas .....
   a. 3
   b. 2
   c. 4
   d. 5
4. Nilai yang berada ditengah sekumpulan data yang telah diurutkan, dari data terkecil hingga terbesar disebut dengan...
5. Ada 2 Jenis Stistika Deskripstif yaitu .....dan Visualisasi Data
6. Tipe Data berdasarkan jenisnya dibedakan menjadi 2 kategori, yaitu Data Kategorikal dan Data....
7. Mean, adalah nilai yang paling sering muncul dalam kumpulan data.
   a. True
   b. False
8. Analisis Regresi Sederhana adalah sebuah metode pendekatan untuk pemodelan hubungan antara satu variabel dependen dan satu variabel independen.
   a. True
   b. False

9. Nilai yang dihitung dari populasi dan memberi deskripsi pada populasi disebut statistik. Sedangkan Nilai yang dihitung dari sampel disebut parameter.
   a. True
   b. False
10. Skala Interval, adalah skala yang berdasarkan satuan unit yang sama dan konstan
   a. True
   b. False

Daftar Pustaka

Data Science for Business https://learn.datacamp.com/courses/data-science-for-business

Data Science From Scratch first principal with Python, published by O'Reilly Media, Inc.,1005 gravenstein Highway North Sebatopol, CA

Data Science with R, https://learn.datacamp.com/career-tracks/data-scientist-with-r

E-Modul Data Science, IlmudataPy.com

Pengantar Data Science dan Aplikasinya bagi Pemula Oleh Program Data Science, Informatika UNPAR, 2020

BAB IV

DATA VISUALIZATION

A. PENDAHULUAN

Mampu memahami dan mengatur serta menampilkan data dengan cara yang menarik dan mudah dicerna. Merepresentasikan data dalam konteks visual, seperti bagan atau peta, untuk membantu siapapun yang melihatnya lebih memahami pentingnya data tersebut dan mengatur agar data sesuai dengan yang diinginkan dengan cara melakukan scrubuung and cleaning data

B. INTI

1. Mampu memahami dan menjelaskan teori visualisai data, data cleaning dan exploratory data analysis.
    - **a.** Mampu memahami dan menjelaskan visualisasi data dan pentingnya visualisasi data
    - **b.** Mampu menguraikan jenis-jenis visualisasi data
    - **c.** Mampu mengetahui waktu menggunakan jenis visualisasi data
    - **d.** Mampu memahami dan menjelaskan visualisasi data untuk tim internal dan stakeholder
    - **e.** Mampu memahami contoh visualisasi data pemasaran pada tim internal
    - **f.** Mampu memvisualisasikan data untuk pemangku kepentingan eksternal dan audiens publik
    - **g.** Mampu memahami dan menjelaskan alat dan sumber daya visualisasi data
    - **h.** Mampu memahami dan menjelaskan apa itu data visualisasi tools
    - **i.** Mampu memahami dan menjelaskan anjuran dan larangan dalam visualisasi data
    - **j.** Mampu memahami dan menjelaskan tentang scrub and cleaning data

2. Materi

- **a.** Apa itu visualisasi data dan mengapa itu penting
- **b.** Jenis-jenis visualisasi data
- **c.** Kapan harus menggunakan berbagai jenis visualisasi data
- **d.** Visualisasi data untuk tim internal dan stakeholder
- **e.** Contoh visualisasi data pemasaran pada tim internal
- **f.** Visualisasi data untuk pemangku kepentingan ekstrnal dan audiensi public
- **g.** Alat dan sumber daya visualisasi data
- **h.** Anjuran dan larangan dalam visualisasi data
- **i.** Scrub and cleaning data

3. Uraian Materi

a. Apa itu visualisasi data dan mengapa itu penting

Data anda hanya sebaik kemampuan anda untuk memahami dan mengomunikasikannya. Anda harus dapat memahami dan secara efektif menceritakan kisah di balik angka-angka. Data yang benar dengan visual yang salah tidak berguna dan bahkan dapat menyesatkan audiens Anda.

Ketika Anda dapat (dengan benar) memvisualisasikan data Anda menggunakan grafik dan bagan, Anda akan lebih mungkin untuk mengungkap pola, korelasi, dan outlier, mengomunikasikan wawasan kepada atasan Anda, tim Anda, perusahaan, atau audiens media sosial, dan membuat data didukung keputusan.

Pentingnya Data sebagai Marketer

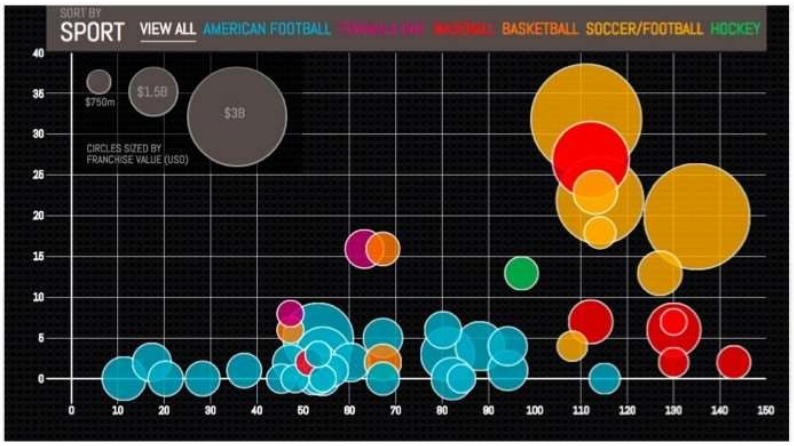
Data adalah informasi yang digunakan untuk dianalisis dan "digunakan untuk membantu pengambilan keputusan," menurut Kamus Cambridge. Dalam bisnis, mengumpulkan data setiap hari, dari lokasi pengunjung situs web, pengikut media sosial, hingga penjualan berdasarkan wilayah. Dengan mengungkap informasi yang tepat dan memasangkannya dengan data yang tepat organisasi dan analisis, dapat

mengambil tindakan dan membantu tim Anda dalam pengambilan keputusan yang didukung data.

Visualisasi data mengacu pada menampilkan data, angka, dan statistik melalui gambar dan bagan. Kapan menampilkan data secara visual, dapat dengan lebih mudah menemukan pola yang bermakna dari serangkaian angka yang tidak dapat diuraikan, dan menarik kesimpulan yang mengarah pada keputusan yang tepat.

Visualisasi data mengacu pada menampilkan data, angka, dan statistik melalui gambar dan bagan. Saat menampilkan data secara visual, Anda dapat dengan lebih mudah menemukan pola yang bermakna dari serangkaian angka yang tidak dapat diuraikan, dan menarik kesimpulan yang mengarah pada keputusan yang tepat.

Berikut adalah contoh sederhana menggunakan visualisasi data. Bagan dibawah ini menunjukkan waralaba olahraga berdasarkan nilai, yang kemudian diplot terhadap kejuaraan dan tahun-tahun yang bersaing.



Gambar 26. Contoh sederhana menggunakan visualisasi data

Setiap lingkaran menunjukkan nilai waralaba dan warna menunjukkan olahraga. Anda mungkin mengharapkan sepak bola Amerika menjadi lingkaran kuning besar itu, tetapi kenyataannya, sepak bola adalah olahraga paling berharga diseluruh dunia. Visualisasi data dapat menunjukkan tren dalam skala yang lebih besar dan mencakup beberapa variabel, mengungkapkan temuan menarik!

Visualisasi data sangat membantu saat menyajikan data kepada orang lain. Dapat lebih mudah mengidentifikasi tren, menjawab pertanyaan, membuktikan teori, membuatnya

untuk bisnis—menunjukkan merek. Saat memvisualisasikan data, dapat lebih mudah melihat angka dalam konteks dan memahami bagaimana mereka berhubungan satu sama lain.

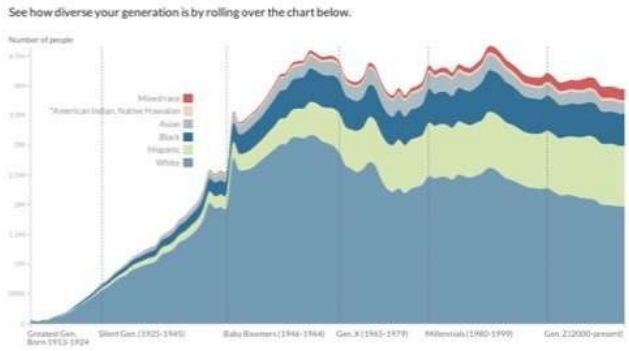
Ketika memproses informasi, kita terhubung untuk memahami gambar lebih cepat dan lebih mudah daripada teks. Menurut penelitian dari American Management Association, ketika dipasangkan dengan visual, informasi 70% lebih mudah diingat daripada ketika ditampilkan sebagai teks saja. Pemasar dapat menggunakan visualisasi data untuk menampilkan hasil kampanye atau menganalisis hasil survei pelanggan dengan lebih baik, sehingga memengaruhi strategi pemasaran dan penjualan atau keputusan anggaran Anda.

Mempersenjatai diri dengan data, dapat mengesankan pemangku kepentingan utama dengan membuktikan nilai kampanye Anda melalui pelaporan ROI, menunjukkan bagaimana Anda telah mengurangi biaya per akuisisi dengan strategi SEO organik Anda, atau memperjuangkan lebih banyak anggaran untuk tim sosial Anda setelah menunjukkan pertumbuhan audiens mereka dan mencapai.

b. Jenis-jenis visualisasi data

Anda dapat menggunakan visualisasi data untuk menampilkan sebagian data, peringkat, korelasi, distribusi geografis, penyimpangan, garis waktu, dan skala. Anda dapat membandingkan pertumbuhan perusahaan Anda dari waktu ke waktu dengan merek atau pesaing inspirasi Anda. Data dapat mewakili hasil survei, atau kata kunci umum di tiket layanan pelanggan, yang menunjukkan di mana tim produk Anda mungkin perlu melakukan perbaikan.

Perhatikan kumpulan data berikut ini.

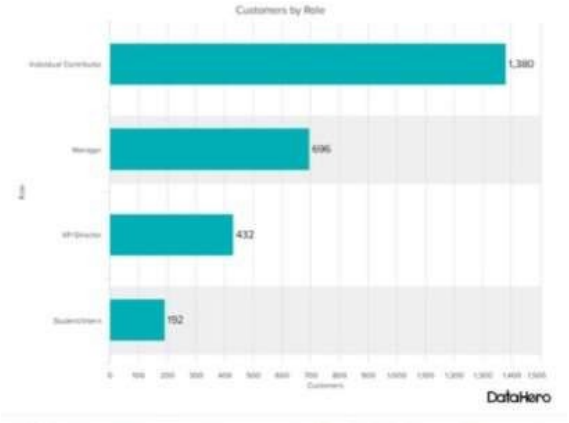


Gambar 27. Kumpulan Data

Ditampilkan sebagai spreadsheet angka sederhana, dampak gambar ini tidak akan terlihat. Saat melihat data ini sebagai bagan area yang ditumpuk satu sama lain, dari waktu ke waktu, dan menggunakan warna yang berbeda, penonton dapat melihat dengan jelas bagaimana susunan masyarakat saat ini telah berubah dari apa yang tampak seperti 100 tahun yang lalu.

Jenis visualisasi data yang umum termasuk bagan, grafik, peta, dan tabel. Anda akan sering melihat banyak visualisasi data yang disertakan dalam infografis, dan data dapat dianimasikan atau dikodekan untuk dimanipulasi dan mengungkapkan data secara perlahan atau seiring waktu.

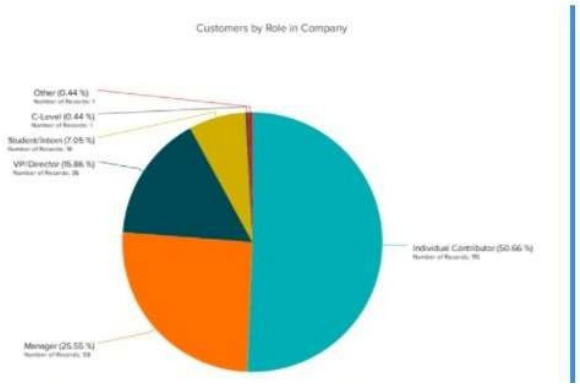
1. Grafik Batang



Gambar 28. Grafik batang

Grafik batang sangat serbaguna, paling baik digunakan untuk menunjukkan perubahan dari waktu ke waktu, membandingkan kategori yang berbeda, atau membandingkan bagian dari keseluruhan.

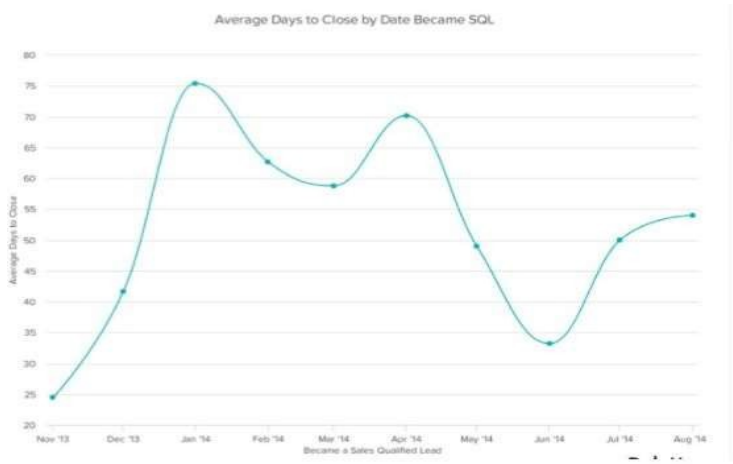
2. Diagram lingkaran



Gambar 29. Diagram Lingkaran

Diagram lingkaran paling baik digunakan untuk membuat perbandingan sebagian-ke-seluruh dengan diskrit atau data terus menerus. Mereka paling berdampak dengan kumpulan data kecil.

3. Grafik Garis



Gambar 30. Diagram Garis

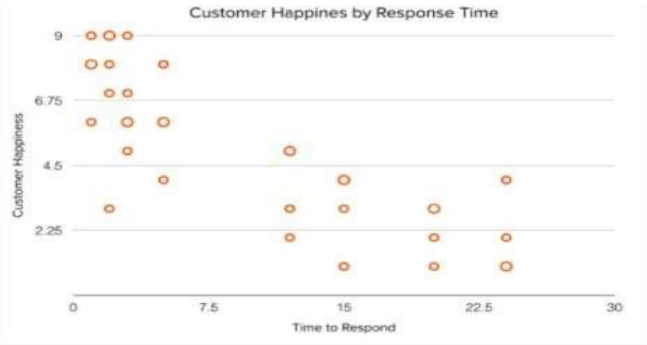
Grafik garis digunakan untuk menunjukkan hubungan deret waktu dengan kontinu data. Mereka membantu menunjukkan tren, akselerasi, deselerasi, dan volatilitas

4. Grafik Area

Gambar 31. Grafik area

Bagan area juga menggambarkan hubungan deret waktu, tetapi mereka berbeda dari diagram garis karena dapat mewakili volume.

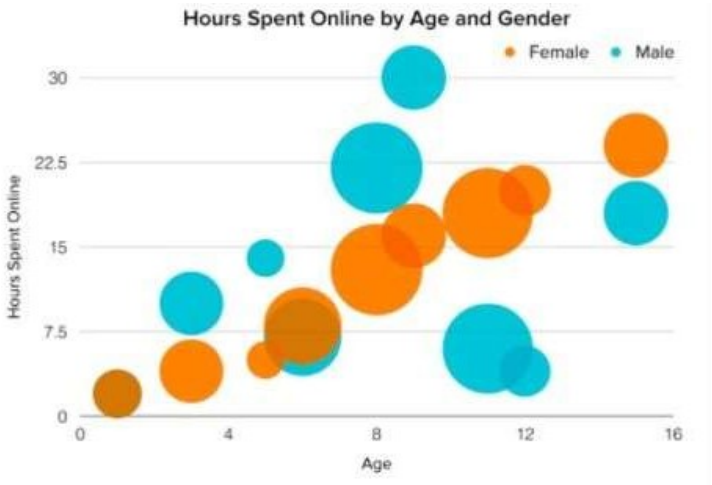
5. Scatter Plots



Gambar 32. Scatter Plots

Plot pencar menunjukkan hubungan antar item berdasarkan dua set variabel. Mereka paling baik digunakan untuk menunjukkan korelasi dalam sejumlah besar data.

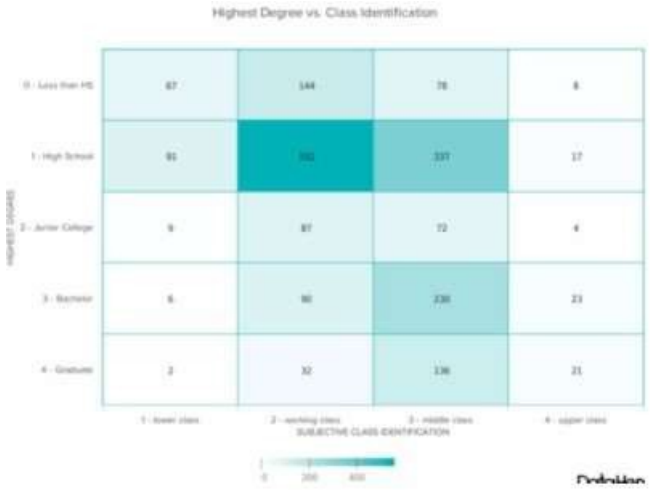
6. Bubble Chart



Gambar 33. Bubble Chart

Bagan gelembung bagus untuk menampilkan perbandingan nominal atau hubungan peringkat. Mereka menampilkan tiga variabel data dan merupakan kombinasi dari scatterplot dan diagram area proporsional. Ini sangat membantu dalam membandingkan beberapa titik data pada grafik yang sama.

7. Heat Maps



Gambar 34. Heat Maps

Heat Maps menampilkan data kategoris, menggunakan intensitas warna untuk mewakili nilai wilayah geografis atau tabel data.

8. Geographical Maps

Gambar 35. Geographical Maps

Peta yang memplot data pada peta geografis, menunjukkan distribusi data lintas wilayah.

9. Tabel

Gambar 36. Tabel

Tabel adalah cara sederhana untuk membandingkan penjualan menurut perwakilan penjualan dengan cepat, seperti tabel dibawah ini, atau lihat dari mana penawaran teratas atau tampilan blog Anda berasal.

10. Waterfall Chart

Gambar 37. Waterfall Chart

Waterfall chart menunjukkan nilai dari waktu ke waktu dan kinerja berbagai kelompok dari waktu ke waktu.

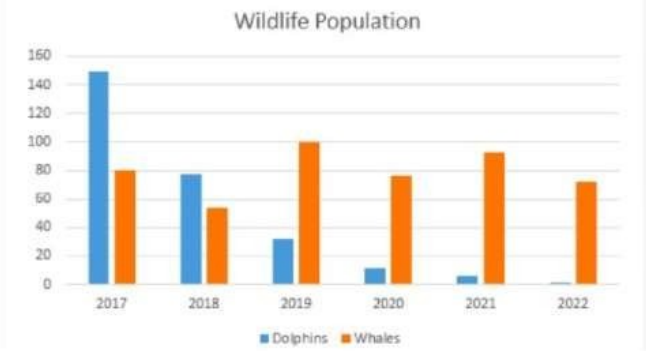
11. Pictograms

Gambar 38. Pictograms

Pictogram adalah cara sederhana untuk menampilkan statistik menggunakan ikon.

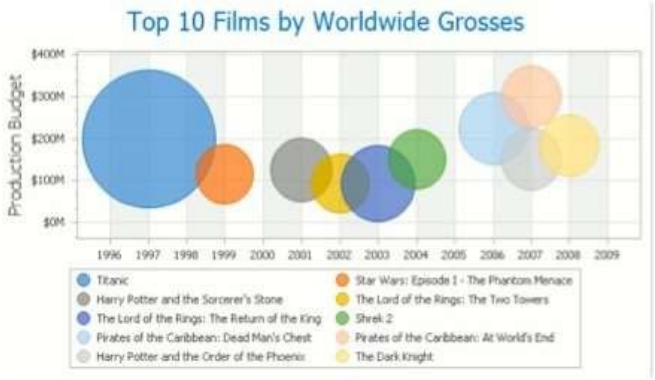
c. Kapan harus menggunakan berbagai jenis visualisasi data

Saat mempertimbangkan jenis bagan, grafik, atau peta yang akan digunakan saat menampilkan data Anda, pertama-tama, pikirkan tentang apa yang ingin Anda tunjukkan, pelajari, atau buktikan. Saat membandingkan dua variabel dengan kumpulan data yang sama, bagan batang kolom, seperti dibawah ini, memungkinkan anda untuk membandingkan keduanya secara bersamaan dan dengan mudah melihat kesenjangan.



Gambar 39. Perbandingan jenis bagan atau grafik

Untuk menunjukkan bagaimana pengelompokan secara keseluruhan, seperti pengunjung di beranda Anda berdasarkan lokasi, atau sumber tampilan blog, coba peta atau bagan area.



Gambar 40. Contoh pengelompokan pengunjung secara keseluruhan

Bagan gelembung dapat membantu Anda melihat dengan lebih baik hubungan antara banyak hal dan dapat membantu membandingkan topik gambaran yang lebih besar seperti anggaran.

Gambar 41. Contoh scatterplot

Untuk lebih memahami distribusi data Anda, dan mengidentifikasi tren, outlier, dan pola, gunakan scatterplot atau grafik garis. Banyak jenis bagan dan grafik dapat digunakan untuk berbagai jenis visualisasi data. Cobalah sepuluh kiat ini untuk menyempurnakan data Anda, seperti membalik sumbu x dan y, menggunakan jenis bagan yang berbeda, atau menghilangkan titik data yang tidak terkait untuk menceritakan kisah dengan lebih baik.

d. Visualisasi data untuk tim internal dan stakeholder

Visualisasi data adalah alat penting untuk rapat internal atau stakeholer. Bergantung pada tim dan sasaran Anda, gunakan data saat menyajikan hal-hal seperti pertumbuhan pelanggan dari waktu ke waktu (mungkin dari bulan ke bulan), atau tarif buka email berdasarkan baris subjek.

Tidak peduli peran Anda dalam tim pemasaran perusahaan Anda, Anda akan menggunakan data untuk melakukan pekerjaan Anda dan menyajikan informasi tentang proyek yang sedang Anda kerjakan. Hampir semua keputusan pemasaran didukung oleh data. Data mencakup informasi kualitatif dan kuantitatif dan mencakup segala sesuatu yang dapat diukur, dikumpulkan, dan dianalisis.

Misalnya, jika Anda bertanggung jawab atas pemasaran konten, Anda akan memiliki sasaran atau KPI (indikator kinerja utama). Pada tim pemasaran konten, tujuan ini

kemungkinan akan berpusat di sekitar jumlah posting blog yang ditulis, peringkat halaman, tampilan halaman organik, pembagian halaman, waktu yang dihabiskan di halaman, jumlah posting yang dibaca atau digulir, dan prospek yang dihasilkan dari posting Anda. Semua ini akan memiliki laporannya sendiri, dan Anda akan memperbaruinya secara teratur untuk memeriksa kemajuan Anda.

Jika Anda menggunakan alat seperti HubSpot untuk blog dan pelaporan Anda, Anda akan dapat dengan mudah mengakses laporan bawaan seperti lalu lintas untuk blog Anda berdasarkan sumber.

Gambar 42. HubSpot untuk blog dan pelaporan

Memiliki pelaporan yang akurat dan sederhana akan membantu Anda berdua berhasil dalam pekerjaan Anda dengan mengidentifikasi topik mana yang beresonansi dengan audiens mana, tetapi mereka akan membantu Anda membuktikan nilai Anda kepada atasan dan pemangku kepentingan internal lainnya, seperti bos atasan Anda ketika tiba saatnya untuk meminta lebih banyak anggaran atau sumber daya untuk mengembangkan program konten Anda.

e. Contoh visualisasi data pemasaran pada tim internal

1. Visualisasi Data Konten Marketing

Gambar 43. Tampilan blog menurut sumber

2. SEO Data Visualization

Gambar 44. Lalu lintas organic berdasarkan sumber

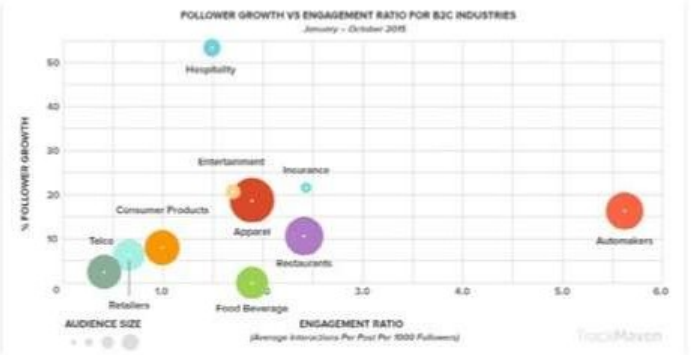
3. Paid Data Visualization

Gambar 45. Biaya rata-rata per klik untuk kampanye iklan menurut kata kunci

4. Email and customer data visualization

Gambar 46. Jumlah rata-rata email yang dibuka oleh pelanggan setiap bulan

5. Social Media Data Visualization



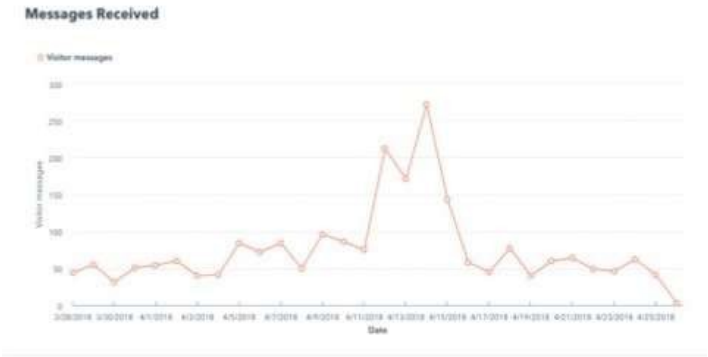
Gambar 47. Pengikut dari waktu ke waktu berdasarkan jaringan dalam Bubble Chart

f. Visualisasi data untuk pemangku kepentingan ekstrnal dan audiensi public

Visualisasi data akan memainkan peran penting dalam laporan tahunan perusahaan anda, infografis, studi kasus, laporan benchmark, dan banyak lagi. Dengan data dan presentasi yang tepat, laporan benchmark perusahaan Anda dapat menjadi sumber daya utama bagi pemasar lain di industri Anda, sehingga meningkatkan otoritas dan peringkat perusahaan Anda untuk laporan tersebut.

Misalnya, katakanlah anda bekerja di tim pemberdayaan penjualan perusahaan Anda. anda mungkin membuat studi kasus dan testimonial dari klien anda yang paling sukses.

Visualisasi data dapat membantu anda menunjukkan dampak bagaimana produk atau layanan Anda memengaruhi metrik klien Anda, seperti studi kasus Rock and Roll Hall of Fame HubSpot di bawah ini.

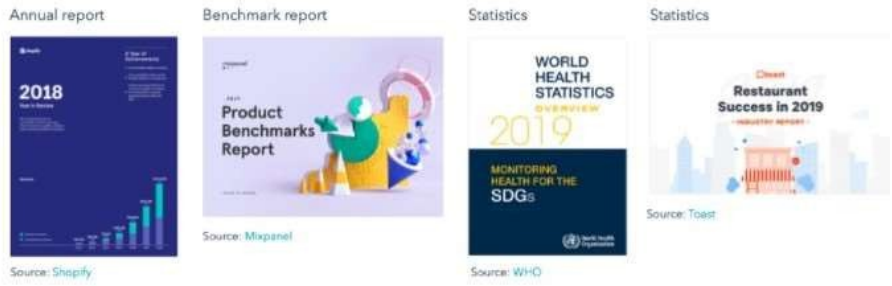


Gambar 48. Studi kasus Rock and Roll Hall of Fame HubSpot

Visualisasi bagan garis ini menunjukkan lonjakan pesan yang diterima, yang menyebabkan peningkatan 81% dalam ukuran audiens mereka.

Menggunakan data membantu membawa pesan ke rumah dan dapat membantu perwakilan penjualan Anda mendapatkan kesepakatan, menggambarkan kemanjuran produk Anda, dan memamerkan merek Anda di media sosial. Visualisasi data dapat digunakan dalam PDF yang dapat dilihat publik, studi kasus, halaman situs web, dan laporan. Pastikan untuk menyinkronkan dengan tim merek Anda untuk memastikan bahwa desain apa pun cocok dengan panduan gaya yang disetujui merek anda.

Berikut ini adalah contoh visualisasi data eksternal:



Gambar 49. Contoh visualisasi data eksternal

Anda juga dapat menggunakan visualisasi data untuk alasan-alasan ini:

1. Studi kasus
2. Social media images
3. PDFs
4. Follower Data
5. User Data

g. Alat dan sumber daya visualisasi data

Untuk memulai visualisasi data, Anda memerlukan beberapa data. Banyak CMS dan alat blogging memiliki beberapa pelaporan bawaan, atau Anda mungkin gunakan alat seperti Google Analytics untuk mengumpulkan data situs web. Setelah Anda mengumpulkan data Anda dengan mengekspornya dari sejumlah alat (termasuk platform media sosial, platform iklan berbayar, data pelanggan, hasil survei, dan banyak lagi), Anda dapat menggunakan Microsoft Excel atau Google Spreadsheet untuk membiasakan diri Anda mengatur data .

Anda perlu mempelajari beberapa fungsi dan rumus utama untuk membawa data Anda ke tempat di mana Anda dapat mulai mengidentifikasi tren dan statistik. Berikut adalah panduan lengkap untuk Excel yang akan memandu Anda melalui filter, tabel pivot, vlookup, dan fungsi lain yang cukup universal di semua versi Excel dan Google Spreadsheet. Dengan menggunakan tabel pivot, Anda dapat melakukan referensi silang data dengan cepat dan mudah, dan menghemat banyak waktu untuk melakukan pekerjaan manual.

h. Data visualization tools

Setelah Anda menghapus semua data yang tidak relevan, memfilter apa yang Anda butuhkan, dan membuat beberapa tabel awal, tabel pivot, dan lembar, Anda siap untuk mulai membuat dan menyempurnakan visualisasi anda. Pelajari dasar-dasar desain

grafis dan buat bagan dan grafik yang paling menarik secara visual untuk konten data yang dapat dilihat publik.

Ada banyak alat dan perangkat lunak visualisasi data yang berbeda di luar sana yang dapat Anda gunakan pada berbagai tingkat keahlian untuk menganalisis dan menyajikan data. Microsoft Excel dan Google Spreadsheet menawarkan visualisasi dasar dan memungkinkan Anda membuat tabel pivot, bagan, dan grafik dengan mudah, serta melihat bagaimana data ditampilkan.

Setelah Anda merasa nyaman dengan kemampuan untuk memanipulasi data dan menelusuri tren, bagan, dan grafik, saatnya untuk mendapatkan desain yang mewah dan membuat bagan yang layak untuk dibagikan. Mulailah dengan kursus Dasar-Dasar Desain Grafis HubSpot, di mana Anda akan mempelajari teori warna, alat desain grafis, dan menyusun gambar dalam identitas merek dan panduan gaya. Setelah itu, Anda dapat menggunakan alat seperti Canva untuk membuat bagan bermerek dan gambar sosial tambahan.

Dalam hal pelaporan internal, Anda dapat menggunakan opsi pelaporan dan analitik internal CMS Anda, seperti dasbor analitik HubSpot, atau alat eksternal seperti Google Analytics atau Hotjar, perangkat lunak pemetaan panas.

1. Microsft Excel

   Ini adalah alat yang paling umum untuk visualisasi data. Excel memungkinkan Anda mengurutkan dan menganalisis data, lalu membuat visualisasi data menggunakan panduan bagan. Untuk mempelajari lebih lanjut tentang menggunakan Excel, lihat sumber Excel.

2. Statista

   Statista adalah database studi, statistik, dan prakiraan yang menyediakan data pasar untuk presentasi.

Gambar 50. Statista

3. Google Trends

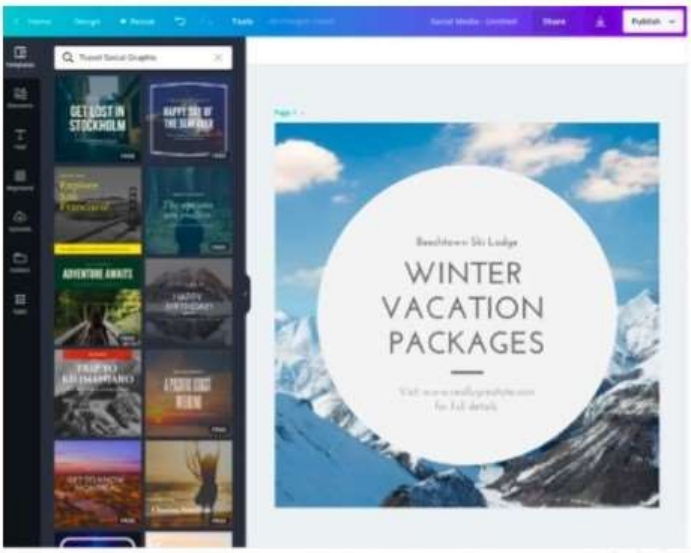
Google Trends adalah alat gratis dari Google yang menunjukkan riwayat pencarian topik dan kata kunci tertentu dari waktu ke waktu. Anda dapat mengevaluasi popularitas istilah tertentu, membandingkannya dengan variasi kata kunci lainnya, menganalisis bagaimana popularitasnya bervariasi dari waktu ke waktu dan di berbagai wilayah/bahasa, dan menampilkan kata kunci terkait, yang dapat membantu dalam mendapatkan saran kata kunci baru.

Gambar 51. Google Trends

4. Canva

Canva adalah alat gratis yang sangat mudah digunakan dengan fitur bagan baru untuk membuat bagan pai sederhana, grafik batang, dan lainnya dalam

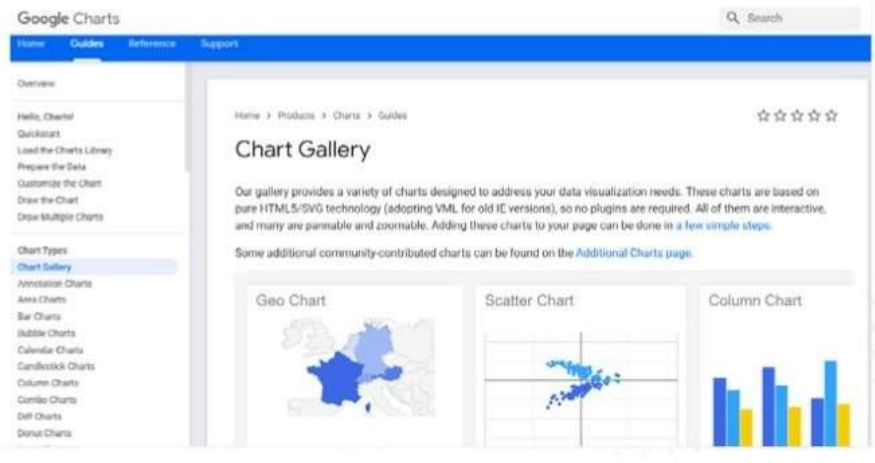
berbagai ukuran template. Ini adalah pilihan yang bagus untuk membuat presentasi, infografis, dan gambar sosial.



Gambar 52. Canva

5. Google Charts and Google Sheets

Google Charts memungkinkan Anda menyematkan grafik dan bagan ke halaman web. Ini adalah alat API yang memungkinkan Anda membuat grafik dan bagan khusus dengan opsi untuk menganimasikannya untuk visualisasi yang lebih dinamis. Google Spreadsheet adalah pilihan yang baik untuk membuat bagan dan grafik sederhana seperti yang Anda lakukan di Excel.



HubSpot Analytics (or your CMS's analytics tool)

Gambar 53. Google Charts and Google Sheets

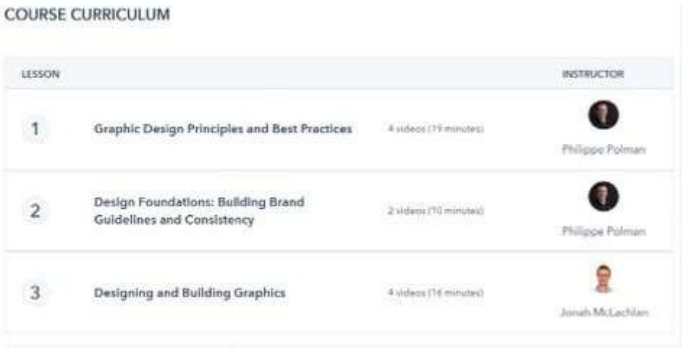
6. HubSpot Analytics (or your CMS's analytics tool)

   Penuhi laporan khusus atau gunakan laporan HubSpot sendiri, lalu perbarui bagan sesuai keinginan Anda. Telusuri HubSpot

Gambar 54. HubSpot Analytics (or your CMS's analytics tool) 1

7. HubSpot Academy's Course on Graphic Design Essentials

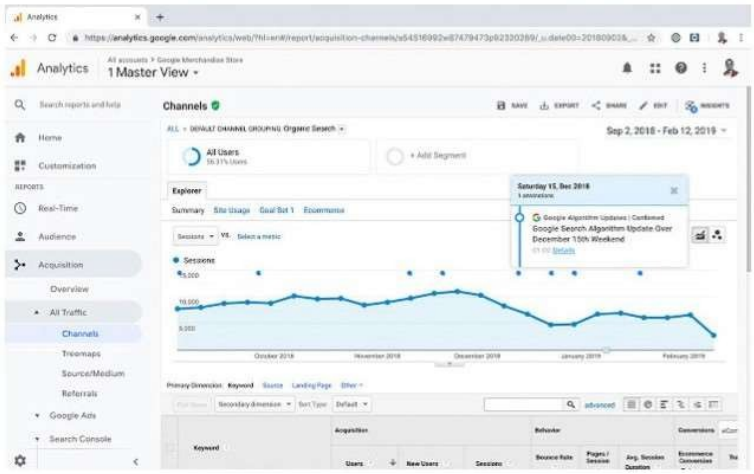
   Pelajari dasar-dasar desain grafis dan buat bagan dan grafik yang paling menarik secara visual untuk konten data yang dapat dilihat publik.



Gambar 55. HubSpot Analytics (or your CMS's analytics tool) 2

8. Google Analytic

   Menganalisis pengunjung situs web Anda dan menyediakan dasbor langsung dengan data visualisasi, bagan, dan grafik perilaku dan tindakan pengunjung Anda. Telusuri Google Chrome.



Gambar 56. Google Analytic

i. Anjuran dan larangan dalam visualisasi data

Terdapat 6 Anjuran dan larangan dalam Vusualisasi Data

1. Anjuran

a. Gunakan warna untuk membantu meningkatkan tren.

b. Sertakan statistik dalam format yang dapat dibaca bagi mereka yang memiliki kesulitan penglihatan dan dengan angka yang tepat.

c. Gunakan data dari lima tahun terakhir.

d. Bandingkan data Anda dengan sebelumnya tolak ukur.

e. Sajikan data Anda dalam urutan terorganisir yang membantu pembaca melihat tren.

f. Gunakan untuk keterbacaan, sertakan penanda data dan persentase jika berlaku.

2. Larangan

a. Membanjiri audiens dengan warna-warna cerah yang berbenturan.

b. Lupa untuk memasukkan angka yang tepat dan hanya menyertakan grafik tanpa persentase.

c. Menggunakan data yang lebih tua dari lima tahun.

d. Kehilangan data Anda segera setelah Anda mempresentasikannya.

e. Sajikan data Anda secara tidak berurutan, memaksa audiens menggali kesimpulan sendiri.
f. Pilih daya tarik estetika atau visual daripada kepraktisan dan keterbacaan.

j. Scrub and Cleaning Data

Langkah "Scrub"

Selama proses bekerja dengan data, Anda akan selalu mencapai titik di mana Anda telah mengumpulkan semua data yang Anda perlukan (langkah "Dapatkan" kami), tetapi data tersebut belum dalam format yang dapat Anda gunakan untuk pemodelan. Semua pekerjaan yang akan Anda lakukan di lab berikutnya adalah mendapatkan kumpulan data dalam format yang dapat dengan mudah Anda jelajahi dan buat modelnya.

Subsampling untuk Mengurangi Ukuran

Saat membangun model untuk tujuan prediktif, lebih banyak data selalu lebih baik saat melatih model akhir. Namun, selama proses pengembangan saat bekerja dengan kumpulan data besar, biasanya bekerja hanya dengan subsampel kumpulan data. Membangun model adalah proses berulang -- sering kali, Anda menyesuaikan model, menyelidiki hasilnya, lalu melatih model lagi dengan beberapa penyesuaian kecil berdasarkan apa yang Anda perhatikan. Karena ini adalah proses berulang, Anda ingin menghindari runtime yang lama, dan iterasi secepat mungkin. Saat Anda puas dengan model yang Anda buat di subsampel data, Anda akan menyesuaikan model di seluruh kumpulan data.

Di lab berikutnya, Anda akan bekerja dengan subsampel set data untuk meningkatkan kecepatan iterasi dalam menyempurnakan model Anda.

Berurusan Dengan Tipe Data

Salah satu masalah paling umum yang harus Anda tangani selama langkah penggosokan data adalah kolom yang dikodekan sebagai tipe data yang salah. Misalnya, masalah pemformatan umum yang mungkin Anda temui adalah data numerik yang salah dikodekan sebagai data string. Ini membuat operasi numerik pada data tidak mungkin

dilakukan tanpa memformat ulang data terlebih dahulu. Operasi sederhana seperti 2 + 2akan kembali 22jika data numerik secara tidak sengaja diformat sebagai string ('2'+'2'='22'). Demikian pula, data kategorikal sering dikodekan sebagai nilai integer. Jika Anda tidak mengkonseptualisasikan dengan benar bagaimana data direpresentasikan, Anda akan gagal merumuskan model dan wawasan yang bermakna. Langkah pertama untuk mengungkap dan menyelidiki masalah tersebut adalah dengan menggunakan .info()metode yang tersedia untuk semua Pandas DataFrames. Ini akan memberi tahu jenis data yang ada di setiap kolom, serta jumlah nilai yang ada di dalam kolom itu (yang juga dapat membantu kami mengidentifikasi kolom yang berisi data yang hilang)! Berikut ini contoh tanggapan:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 97839 entries, 0 to 97838
Data columns (total 16 columns):
Store           97839 non-null object
Dept            97839 non-null object
Date            97839 non-null object
Weekly_Sales    97839 non-null float64
IsHoliday       97839 non-null bool
Type            97839 non-null object
Size            97839 non-null int64
Temperature     97839 non-null float64
Fuel_Price      97839 non-null float64
MarkDown1       35013 non-null float64
MarkDown2       27232 non-null float64
MarkDown3       32513 non-null float64
MarkDown4       34485 non-null float64
MarkDown5       35013 non-null float64
CPI             97839 non-null float64
Unemployment    97839 non-null float64
dtypes: bool(1), float64(10), int64(1), object(4)
memory usage: 12.0+ MB
```

Dari sini, langkah selanjutnya yang baik adalah melihat contoh dari setiap kolom yang dikodekan sebagai string (ingat, Pandas mengacu pada kolom string sebagai object) dan konfirmasikan bahwa data ini seharusnya dikodekan sebagai string. Salah satu metode untuk melakukannya adalah dengan melihat pratinjau versi terpotong dari output dari .value_counts(). Misalnya, Anda dapat mempratinjau 5 entri paling sering dari setiap kolom dengan loop sederhana seperti ini:

```
for col in df.columns:
    try:
        print(col, df[col].value_counts()[:5])
    except:
        print(col, df[col].value_counts())
        # If there aren't 5+ unique values for a column the first print statement
        # will throw an error for an invalid idx slice
    print('\n') # Break up the output between columns
```

Biasanya juga merupakan ide yang baik untuk memeriksa kolom bilangan bulat untuk memastikan bahwa data yang dikandungnya dimaksudkan untuk mewakili data numerik yang sebenarnya, dan bukan hanya data kategorikal yang dikodekan sebagai bilangan bulat. Anda juga dapat menemukan nilai nol yang dikodekan secara keras sebagai string seperti "?", "999999"atau nilai asing lainnya tergantung pada kumpulan data dan siapa yang membuatnya.

Data Numerik Dikodekan sebagai String

Jika Anda telah mengidentifikasi data numerik yang dikodekan sebagai string, biasanya masalah ini cukup mudah untuk dipecahkan. Seringkali sesederhana casting data string ke tipe numerik:

```
df['numeric_string_col'] = df['numeric_string_col'].astype('float')
```

Sayangnya, tidak selalu sesederhana itu. Misalnya, jika ada sel tunggal yang berisi huruf atau karakter non-numerik seperti koma atau simbol moneter ($) pernyataan di atas akan gagal. Dalam kasus seperti itu, fungsi pembersihan yang lebih kompleks harus dibuat secara manual. Ini bisa melibatkan pengupasan simbol asing seperti ',/$/%', atau cukup casting string yang tidak dapat dikonversi sebagai null. Ingatlah bahwa ketika NumPy melihat beberapa tipe data dalam array, defaultnya adalah mentransmisikan semuanya sebagai string. Jika Anda mencoba mentransmisikan kolom dari string ke tipe data numerik dan mendapatkan kesalahan, pertimbangkan untuk memeriksa nilai unik di kolom itu -- kemungkinan Anda memiliki satu huruf yang bersembunyi di suatu tempat yang perlu dihapus!

Data Kategoris Dikodekan sebagai Bilangan Bulat

Itu juga umum untuk melihat data kategorikal dikodekan sebagai bilangan bulat. Mengingat bahwa langkah besar dalam proses pembersihan data adalah mengonversi semua kolom kategorikal menjadi ekuivalen numerik, ini mungkin tidak tampak seperti masalah pada pandangan pertama. Namun, membiarkan data kategorikal dikodekan sebagai bilangan bulat dapat memiliki efek negatif dengan memasukkan informasi buruk ke dalam model kami. Ini karena penyandian bilangan bulat secara keliru menambahkan hubungan matematis antara kategori yang berbeda -- model kami mungkin salah mengira bahwa kategori yang diwakili oleh bilangan bulat 4dua kali lebih banyak dari kategori 2, dan seterusnya.

Cara terbaik untuk mengatasi masalah ini adalah dengan melemparkan seluruh kolom ke tipe data string, yang akan lebih mewakili sifat kategoris kolom. Karena ini kategorikal, kita dapat menanganinya dengan benar saat kita mengkodekan data kategorikal dalam prosesnya nanti.

Contoh berikut menunjukkan sintaks yang diperlukan untuk mengonversi kolom dari satu tipe data ke tipe data lainnya:

```
# Cast to a numeric type
df['some_column'] = df.['some_column'].astype('float32')

# Cast back to a string type
df['some_column'] = df.['some_column'].astype('str')
```

Setelah selesai, biasanya variabel kategoris ini diteruskan ke fungsi lain seperti pd.get_dummies()untuk mengubah fitur ini menjadi representasi yang lebih cocok untuk algoritme pembelajaran mesin. Mungkin perlu untuk menjatuhkan dummy pertama untuk menghindari jebakan variabel dummy.

Mendeteksi dan Menangani Nilai Null

Pemeriksaan pembersihan data penting lainnya adalah memeriksa nilai yang hilang atau nol. Ingat dari lab sebelumnya bahwa Pandas menunjukkan nilai yang hilang sebagai NaN.

Memeriksa NaNs

Anda dapat dengan mudah memeriksa berapa banyak nilai yang hilang yang terkandung dalam setiap kolom dengan meminta Pandas membuat tabel kebenaran di mana sel-sel yang berisi NaNditandai sebagai Truedan yang lainnya ditandai sebagai False.

```
# Create a truth table for missing values
df.isna()
```

Sejak False=0dan True=1dalam pemrograman, Anda kemudian dapat sum()tabel kebenaran ini untuk mendapatkan jumlah kolom demi kolom dari jumlah nilai yang hilang dalam kumpulan data.

```
# Check how many missing values in each column
df.isna().sum()
```

Seperti disebutkan di atas, ingatlah bahwa kumpulan data Anda mungkin juga berisi nilai nol yang dilambangkan dengan nilai placeholder. Sebagian besar kumpulan data yang melakukan ini akan menyebutkan hal ini dalam kamus data kumpulan data. Namun, Anda mungkin juga melihat ini dilambangkan dengan nilai ekstrem yang tidak masuk akal (misalnya berat badan seseorang disetel ke sesuatu seperti 0 atau 10.000). Melakukan pemeriksaan manual cepat terhadap nilai teratas untuk setiap fitur seringkali merupakan satu-satunya cara untuk mendeteksi anomali tersebut.

Berurusan Dengan Nilai Null

Ada beberapa opsi untuk menangani nilai nol. Anda selalu dapat menghapus baris observasi dengan nilai yang hilang atau menghapus fitur dengan sparity berlebihan yang disebabkan oleh nilai null. Yang mengatakan, melakukan hal itu membuang informasi yang berpotensi berharga. Mungkin ada alasan penting mengapa informasi tersebut hilang. Meskipun demikian, banyak algoritme pembelajaran mesin tidak akan mentolerir nilai nol dan karena itu Anda harus memasukkan nilai atau menghapus data. Beberapa opsi yang Anda miliki untuk memasukkan data meliputi:

Data Numerik

1. Mengganti Nulls dengan median kolom
2. Binning data dan konversi kolom ke format kategoris (Klasifikasi Kasar)

Kategori data

1. Membuat Null menghargai kategori mereka sendiri
2. Mengganti nilai nol dengan kategori yang paling umum

Sebagai data science, Anda harus memutuskan metode imputasi mana yang sesuai untuk kumpulan data dan tujuan bisnis Anda.

Memeriksa Multikolinearitas

Sebelum melanjutkan ke pemodelan, Anda juga ingin memeriksa bahwa data tidak memiliki multikolinearitas atau korelasi/kovarians yang tinggi antar kolom prediktor. Cara termudah untuk melakukan ini untuk membangun dan menafsirkan peta panas korelasi dengan seabornpaket.

4. Forum Diskusi
   a. Berdasarkan tim anda, buatlah grafik 3 dimensi dimana dataset dapat diambil pada website https://covid19.go.id/ dengan menyajikan data dengan table perkembangan covid di Indonesia pada 1 tahun terakhir yaitu pada tahun januari 2020 sampai dengan desember 2020.
   b. Lalu lakukan Analisa berdasarkan table tersebut dan tampilkan visualisasi data yang menarik agar dapat mudah difahani penerima informasi

C. PENUTUP

1. Rangkuman

Visualisasi data bisa tampak sulit dan berlebihan, tetapi dengan alat dan sumber daya yang disediakan dalam panduan ini, Anda akan siap untuk mulai membuat visualisasi

data yang menarik, presentasi persuasif, dan menemukan tren dalam data yang sudah Anda miliki.

Data dapat memaksa orang untuk mengubah hasil kesehatan global atau memengaruhi persepsi. Ini juga dapat membantu tim pemasaran untuk membuat kampanye yang didukung data dan hanya berinvestasi pada peluang yang dapat dilacak dan dapat membuktikan ROI mereka sendiri. Dengan visualisasi data, Anda dapat membagikan apa yang telah Anda temukan dan menampilkan penelitian yang telah Anda lakukan dengan tampilan yang menarik dan cara yang efektif. Dengan data di pihak Anda, Anda akan membantu menciptakan peningkatan eksponensial dalam kinerja tim pemasaran, lalu pamerkan dengan kemampuan baru Anda untuk membuat bagan yang luar biasa.

2. Tes Formatif

Sebagai evaluasi pemahaman materi pada bab 4. Jawablah pertanyaan berikut ini.

1. Uraikan manfaat dari visualisasi Data?

2. Berikut ini yang merupakan Contoh visualisasi data pemasaran pada tim internal adalah:

- **a.** Visualisasi Data Konten Marketing
- **b.** Paid Data Visualization
- **c.** Email and customer data visualization
- **d.** Social Media Data Visualization

3. Berikut ini yang merupakan jenis visualisasi data adalah:

- **a.** Grafik batang
- **b.** Grafik lingkaran
- **c.** Grafik garis
- **d.** Grafik Area

4. Berikut ini adalah grafik jenis apa?

- **a.** Grafik batang
- **b.** Grafik lingkaran
- **c.** Grafik garis
- **d.** Grafik Area

5. Berikut ini yang merupakan larangan dalam visualisasi data adalah:
   a. Gunakan warna untuk membantu meningkatkan tren.
   b. Sertakan statistik dalam format yang dapat dibaca bagi mereka yang memiliki kesulitan penglihatan dan dengan angka yang tepat.
   c. Sajikan data Anda dalam urutan terorganisir yang membantu pembaca melihat tren
   d. Menggunakan data yang lebih tua dari lima tahun.
6. Peta yang memplot data pada peta geografis, menunjukkan distribusi data lintas wilayah, disebut dengan apa?
   a. Geographical Maps
   b. Tabel
   c. Waterfall Chart
   d. Pictograms
7. Berikut ini yang merupakan tools data visualization adalah
   a. MS. Excel
   b. Statista
   c. Canva
   d. Google Charts and Google Sheets
8. Alat penting untuk rapat internal atau stakeholer adalah
9. Informasi yang digunakan untuk dianalisis dan "digunakan untuk membantu pengambilan keputusan," menurut Kamus Cambridge. Hal tersebut adalah defenisi dari apa?
10. Banyak CMS dan alat blogging memiliki beberapa pelaporan bawaan, atau Anda mungkin gunakan alat seperti Google Analytics untuk mengumpulkan data situs web
    a. True
    b. False
11. Memiliki pelaporan yang akurat dan sederhana akan membantu Anda berdua berhasil dalam pekerjaan Anda dengan mengidentifikasi topik mana yang beresonansi dengan audiens mana, tetapi mereka akan membantu Anda

membuktikan nilai Anda kepada atasan dan pemangku kepentingan internal lainnya, seperti bos atasan Anda ketika tiba saatnya untuk meminta lebih banyak anggaran atau sumber daya untuk mengembangkan program konten Anda

a. True

b. False

12. Plot pencar menunjukkan hubungan antar item berdasarkan dua set variabel. Mereka paling baik digunakan untuk menunjukkan korelasi dalam sejumlah besar data.

a. True

b. False

13. Saat menampilkan data secara visual, Anda dapat dengan lebih mudah menemukan pola yang bermakna dari serangkaian angka yang tidak dapat diuraikan, dan menarik kesimpulan yang mengarah pada keputusan yang tepat.

a. True

b. False

Daftar Pustaka

An Introduction to Data Visualization: How to Create Compelling Charts & Graphs, https://offers.hubspot.com/hubfs/DataVisualization_Ebook.pdf?hubs_post-cta=author&_ga=2.253808911.605801064.1631361271-1844418442.1630874161&hubs_offer=offers.hubspot.com%2Fdata-visualization-guide&hubs_post=blog.hubspot.com%2Fmarketing%2Fdata-visualization-guide&submissionGuid=267317a5-73a5-4921-ae52-f2f651a3fc2f&hubs_signup-url=offers.hubspot.com/thank-you/data-visualization-guide&hubs_signup-cta=typ-download

E-Modul Data Science, IlmudataPy.com

BAB V

MODELING TO EVALUATION

A. PENDAHULUAN

Mampu memahami dan menjelaskan bagian dari metodologi data sciene pada tahap modeling ke evaluation, yaitu dengan cara apa data dapat divisualisasikan untuk mendapatkan jawaban yang diperlukan. Dan apakah model yang digunakan benar-benar dapat menjawab pertanyaan(kebutuhan) atau apakah perlu penyesuaian terhadap kebutuhan data. Dengan cara memilih model yang tepat dan cara mengevaluasi model tersebut atau model ini siap untuk diterapkan atau tidak.

B. INTI

1. Mampu memahami dan menguraikan proses dari tahap modeling ke tahap evaluation
   - **a.** Mampu memahami dan menjelaskan apa itu modeling pada proses data science
   - **b.** Mamapu memahami dan menjelaskan modeling pada studi kasus
   - **c.** Mampu memahami dan menjelaskan apa itu evaluation pada proses data science
   - **d.** Mampu memahami dan menganalisa evaluation pada studi kasus
   - **e.** Mampu memahami dan menjelaskan tahap modelling ke evaluation pada proses data science
   - **f.** Mampu memahami dan menjelaskan kurva ROC dan UAC pada klasifikasi

2. Materi
   - **a.** Modeling
   - **b.** Studi Kasus Modeling
   - **c.** Evaluation
   - **d.** Studi Kasus Evaluation

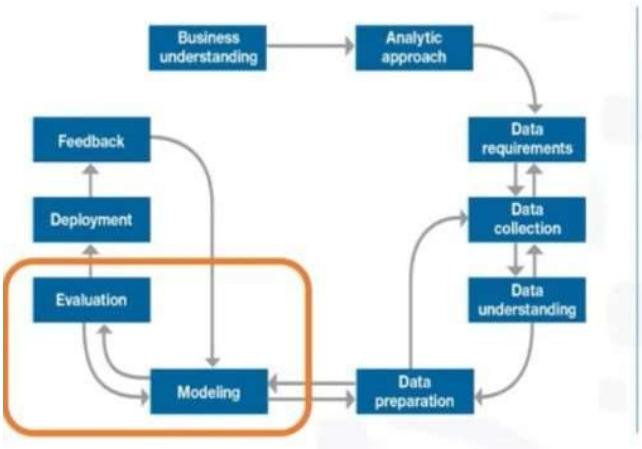
e. Model Evaluatoin

f. Matrik Evaluation

g. kurva ROC dan UAC

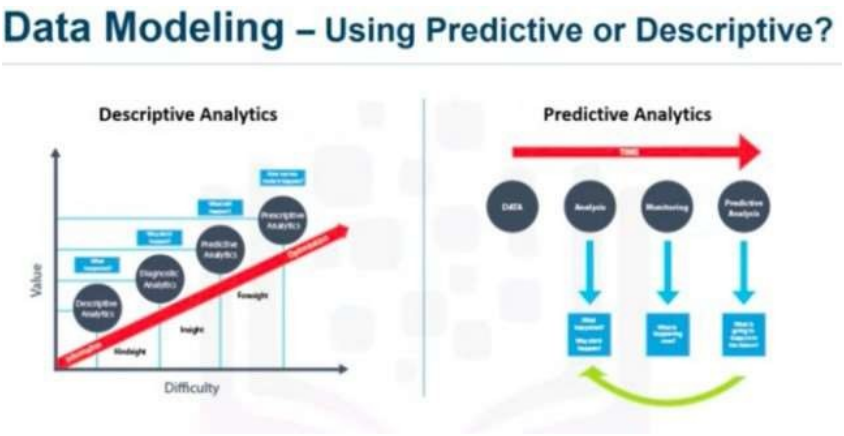
3. Uraian Materi

a. Modeling



Gambar 57. Tahap modeling ke evaluation

Pemodelan adalah salah satu fase dari metodologi data science dimana data scientist dapat memiliki kesempatan untuk mereview hasil dan menentukan hasil tersebut layak atau tidak untuk diterapkan pada sebuah kasus tertentu. Pemodelan data adalah proses menghasilkan diagram deskriptif hubungan antara berbagai jenis informasi yang akan disimpan dalam database. Salah satu tujuan pemodelan data adalah untuk menciptakan metode penyimpanan informasi yang paling efisien sambil tetap menyediakan akses dan pelaporan yang lengkap. Pemodelan data berfokus pada pengembangan model deskriptif atau prediktif.



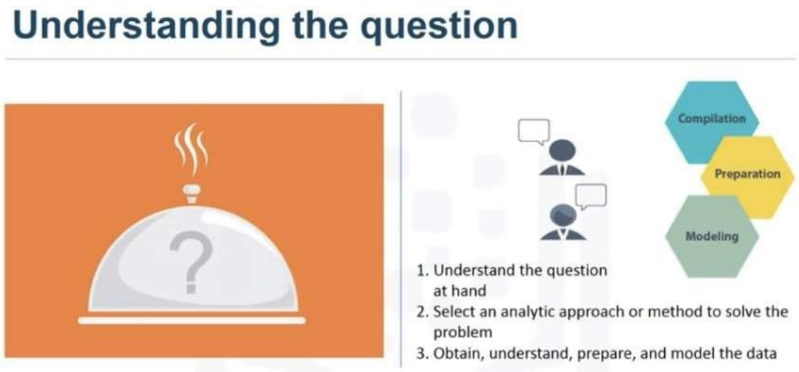
Gambar 58. Data Modeling menggunakan prediktif atau deskriptif

Contoh model deskriptif adalah : jika seseorang melakukannya, mereka mungkin lebih menyukainya.

Sebuah model prediksi merupakan upaya untuk memberikan berhasil atau tidak berhasil, atau berhenti atau melanjutkan. Model-model ini didasarkan pada pendekatan analitik yang dipelajari baik secara statistik atau machine learning. Data Scientist akan menggunakan set training untuk pemodelan prediktif.

Satu training set adalah satu set data historis dimana hasilnya sudah diketahui. Training set berfungsi sebagai indikator untuk menentukan apakah model perlu dikalibrasi .

Pada titik ini, data scientist akan menggunakan beberapa algoritma untuk memastikan bahwa variabel yang terlibat benar-benar dibutuhkan.



Gambar 59. Memahami pertanyaan(kebutuhan)

Keberhasilan dari pengumpulan data, penyusunan dan modeling tergantung pada pemahaman tentang masalah tersebut dan pendekatan analitis yang tepat. Data yang

mendukung jawaban atas pertanyaan dan kualitas bahan di dapur menjadi dasar dari hasil tersebut. Setiap langkah membutuhkan peningkatan, penyesuaian, dan penyesuaian yang konstan untuk memastikan kekuatan hasilnya.

Dalam metodologi data science deskriptif John Rollins, kerangka kerja dirancang untuk tiga hal:

1. Pertama, pahami pertanyaan yang menjadi perhatian anda.
2. Kedua, memilih pendekatan atau metode analitis untuk memecahkan masalah.
3. Ketiga, memperoleh, memahami, menyiapkan dan memodelkan data.

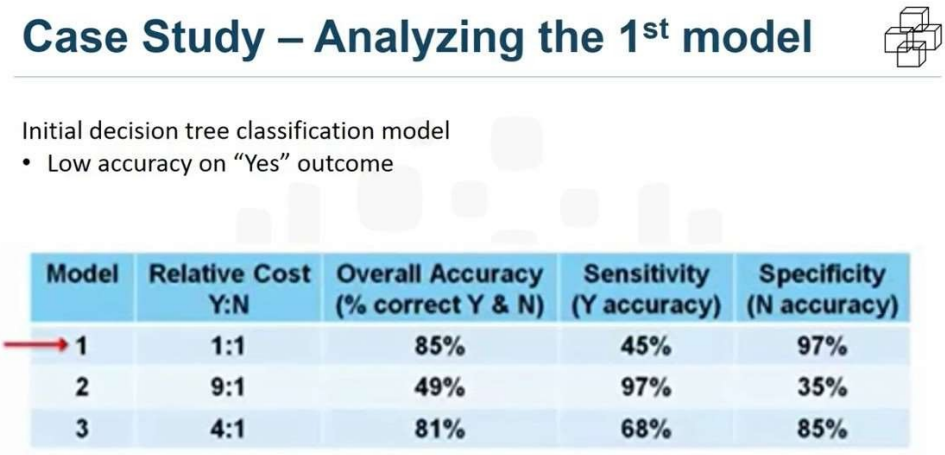
Tujuan utamanya adalah untuk membawa data scientist ke titik dimana dimungkinkan untuk membuat model data untuk menjawab pertanyaan.

Dalam menguji pertanyaan yang sudah terjawab. Misalkan pada sebuah restoran, saat makan malam disajikan dan tamu yang lapar duduk di meja, pertanyaan kuncinya adalah: apakah saya sudah cukup siap untuk makan? Kami berharap bahwa pada tahap metodologi ini, evaluasi model, penyebaran dan siklus umpan balik dari model akan memastikan bahwa responsnya relevan dan mendekati hasil .

Relevansi ini penting untuk seluruh bidang data science, karena ini adalah bidang yang relatif baru dan sangat menarik dengan kemungkinan yang ditawarkannya.

Dalam hal ini akan membahas salah satu dari banyak aspek konstruksi model, dengan ini mengoptimalkan parameter untuk memperbaiki model.

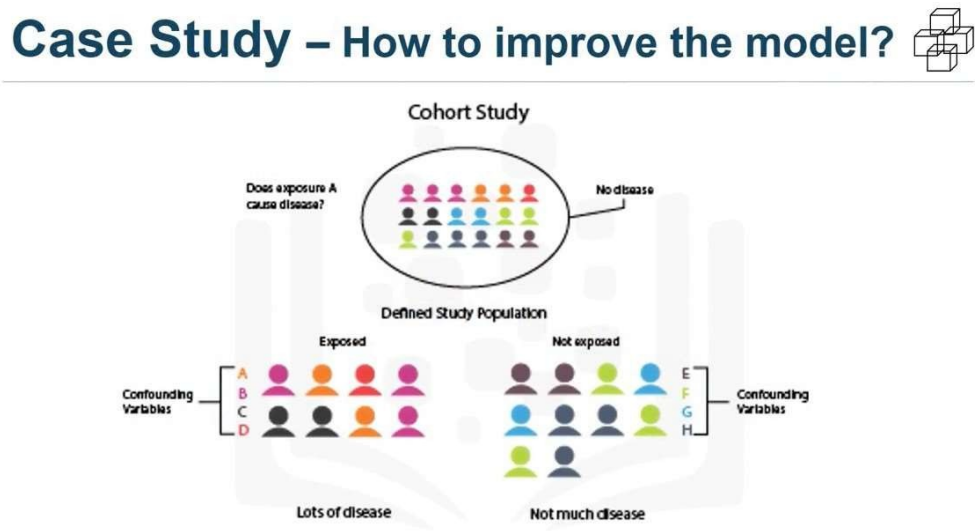
b. Contoh Studi Kasus Modeling



Gambar 60. Studi kasus-analisa model pertama

Dengan satu set data pelatihan yang disiapkan, dimungkinkan untuk membangun model klasifikasi pertama dari decission tree untuk penerimaan kembali kongestif untuk gagal jantung. Kami mencari pasien dengan risiko tinggi masuk kembali. Hasil yang menarik bagi kami adalah penerimaan kembali kongestif untuk gagal jantung yang setara dengan "ya". Pada model pertama ini, akurasi klasifikasi hasil secara keseluruhan adalah 85% dan bukan 85%. Kedengarannya bagus, tetapi hanya mewakili 45% dari "ya". Penerimaan kembali yang sebenarnya diberi peringkat dengan benar, yang berarti bahwa modelnya tidak terlalu akurat.

Pertanyaannya adalah: bagaimana meningkatkan akurasi model untuk memprediksi hasil itu sendiri?. Untuk klasifikasi decision tree, parameter terbaik untuk menyesuaikan adalah biaya relatif dari hasil ya dan tidak diklasifikasikan salah.



Gambar 61. Bagaimana memperbaiki sebuah model

Pikirkan seperti ini: Ketika non-penerimaan kembali yang sebenarnya salah diklasifikasikan dan tindakan diambil untuk mengurangi risiko pasien ini, biaya kesalahan ini adalah intervensi yang sia-sia.

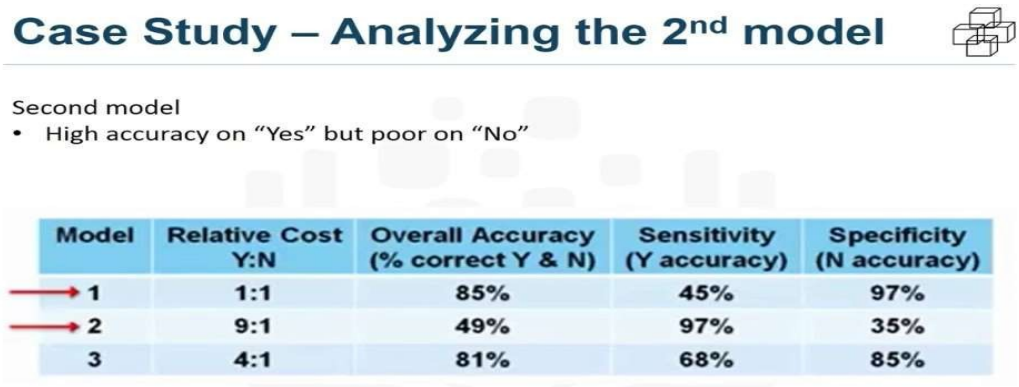
		True condition			
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$	F_1 score = $\frac{1}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$

Gambar 62. Prediction Condition

Seorang ahli statistik menyebut ini sebagai kesalahan Tipe I atau positif palsu. Tetapi ketika penerimaan kembali yang sebenarnya salah diklasifikasikan dan tidak ada tindakan yang diambil untuk mengurangi risiko ini , biaya kesalahan tersebut adalah penerimaan kembali dan semua biaya terkait, serta trauma pada pasien.

Ini adalah kesalahan Tipe II atau negatif palsu . Kemudian kita dapat melihat bahwa biaya dari dua jenis kesalahan klasifikasi yang salah bisa sangat berbeda. Untuk alasan ini, masuk akal untuk menyesuaikan bobot relatif dari klasifikasi hasil yang salah ya dan tidak.

Standarnya adalah antara 1 dan 1, tetapi algoritme pohon keputusan memungkinkan Anda menetapkan nilai yang lebih tinggi u ntuk diri sendiri.



Gambar 63. Studi kasus-analisa model kedua

Untuk model kedua , biaya relatif ditetapkan pada 9/1 . Laporan ini sangat tinggi, tetapi memberikan lebih banyak informasi tentang perilaku model. Kali ini, model 97% bekerja dengan baik , tetapi dengan biaya yang sangat rendah , dengan akurasi umum hanya 49%. Jelas, ini bukan model yang baik.

Masalah dengan hasil ini adalah sejumlah besar positif palsu , menyarankan tidak perlu dan intervensi mahal untuk pasien yang tidak pernah kembali mengakui. Oleh karena itu, data scientist harus mencoba lagi untuk mendapatkan keseimbangan yang lebih baik antara data ya dan tidak.

Case Study – Analyzing the 3rd model

Third model
- Better balance on "Yes" and "No" accuracy

Model	Relative Cost Y:N	Overall Accuracy (% correct Y & N)	Sensitivity (Y accuracy)	Specificity (N accuracy)
1	1:1	85%	45%	97%
2	9:1	49%	97%	35%
3	4:1	81%	68%	85%

Gambar 64. Studi kasus-analisa model ketiga

Untuk model ketiga, biaya relatif ditetapkan ke rasio 4:1 yang lebih masuk akal . Kali ini, 68% diperoleh ya , tetapi ahli statistik menyebutnya sensitivitas, dan akurasi 85% untuk tidak, disebut spesifisitas, dengan akurasi keseluruhan 81%.

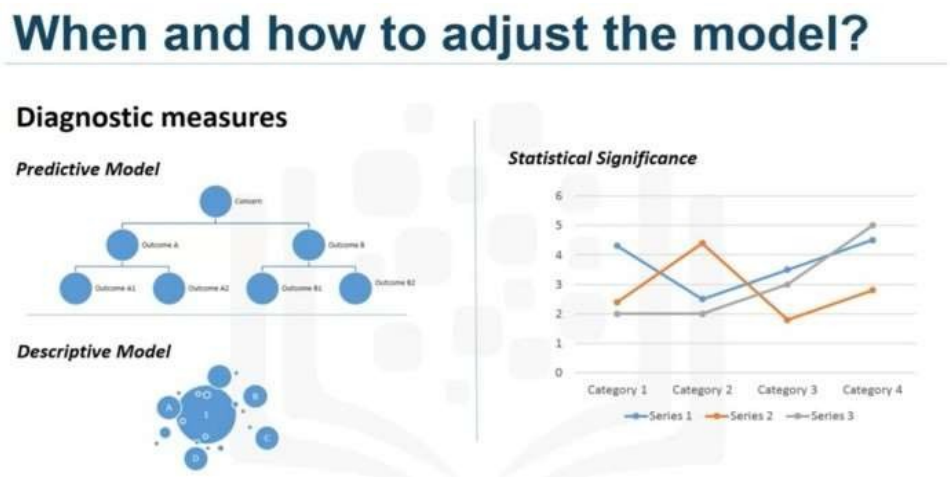
Ini adalah keseimbangan terbaik yang dapat dicapai dengan rangkaian latihan yang relatif terbatas dengan menyesuaikan biaya relatif dari parameter hasil ya dan tidak yang salah klasifikasi. Tentu saja, pemodelan membutuhkan lebih banyak pekerjaan, termasuk iterasi dalam fase persiapan data, untuk mendefinisikan kembali beberapa variabel lain agar lebih mewakili informasi yang mendasarinya dan dengan demikian meningkatkan model.

c. Evaluation

Evaluasi model berjalan seiring dengan pembuatan model. Langkah-langkah pemodelan dan evaluasi dilakukan secara iteratif. Evaluasi model dilakukan selama pengembangan model dan sebelum penerapan. Evaluasi mengevaluasi kualitas model, tetapi juga memberikan kesempatan untuk menentukan apakah memenuhi persyaratan awal.

Evaluasi menjawab pertanyaan: "Apakah model yang digunakan benar-benar menjawab pertanyaan awal atau harus disesuaikan?"

Evaluasi model dapat memiliki dua fase utama.



Gambar 65. Bila dan bagaimana menyesuaikan sebuah model

Fase pertama adalah fase pengukuran diagnostik, yang memastikan bahwa model berfungsi sebagaimana dimaksud. Jika modelnya prediktif , pohon keputusan dapat digunakan untuk menilai apakah respons yang diberikan oleh model sesuai dengan desain aslinya. Ini memungkinkan area untuk ditampilkan di mana penyesuaian diperlukan. Jika model adalah model deskriptif yang mengevaluasi hubungan, satu set tes dengan hasil yang dikenal dapat diterapkan dan model halus yang diperlukan.

Tahap evaluasi kedua yang dapat digunakan adalah uji signifikansi statistik. Jenis evaluasi ini dapat diterapkan pada model untuk memastikan bahwa data model diproses dan diinterpretasikan dengan benar. Ini untuk menghindari asumsi kedua yang tidak perlu ketika jawabannya terungkap.

d. Studi Kasus Evaluation

Mari kembali ke studi kasus untuk menerapkan komponen Evaluasi dalam metodologi data science.

Case Study – Misclassification costs

Misclassification cost tuning
- Tune the relative misclassification costs
- Balance true-positive rate and false-positive rate for best model

Model	Relative Cost Y:N	True Positive Rate (Sensitivity)	Specificity (accuracy on N)	False Positive Rate (1 – Specificity)
1	1:1	0.45	0.97	0.03
2	1.5:1	0.60	0.92	0.08
3	4:1	0.68	0.85	0.15
4	9:1	0.97	0.35	0.65

Gambar 66. Studi kasus-klasifikasi biaya kesalahan

Mari kita cari cara untuk menemukan model yang optimal melalui pengukuran diagnostik berdasarkan konfigurasi salah satu parameter konstruksi model. Kita akan memeriksa lebih dekat bagaimana biaya relatif dari kesalahan klasifikasi hasil positif dan negatif dapat disesuaikan. Seperti yang ditunjukkan pada tabel ini, empat model dibangun dengan empat biaya kesalahan klasifikasi relatif yang berbeda.

Case Study – Relative costs

Misclassification cost tuning
- Tune the relative misclassification costs
- Balance true-positive rate and false-positive rate for best model

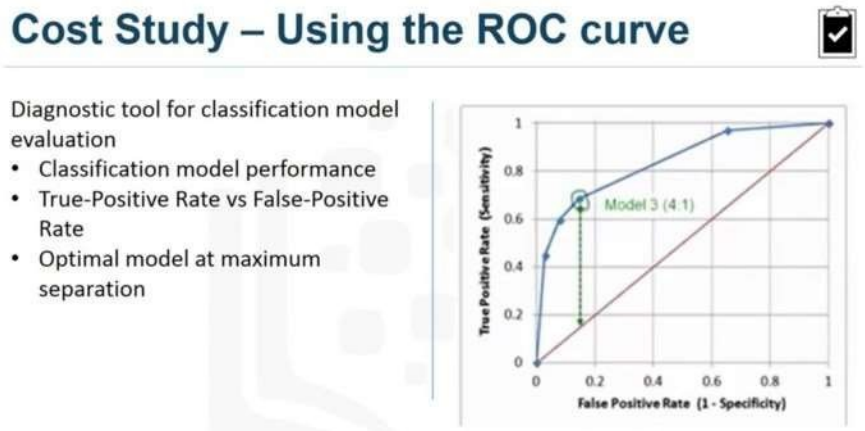
Model	Relative Cost Y:N	True Positive Rate (Sensitivity)	Specificity (accuracy on N)	False Positive Rate (1 – Specificity)
1	1:1	0.45	0.97	0.03
2	1.5:1	0.60	0.92	0.08
3	4:1	0.68	0.85	0.15
4	9:1	0.97	0.35	0.65

Gambar 67. Studi kasus-biaya relative

Pada gambar 67, setiap nilai parameter konstruksi model ini meningkatkan tingkat positif sejati, atau sensitivitas, akurasi dalam prediksi ya, sehingga merugikan akurasi yang lebih rendah dalam prediksi no yaitu, peningkatan tingkat positif palsu.

Pertanyaannya adalah, model mana yang terbaik berdasarkan pengaturan parameter ini? Untuk alasan anggaran, intervensi pengurangan risiko tidak dapat diterapkan pada sebagian besar pasien dengan gagal jantung, banyak diantaranya tidak akan

diterima kembali. Di sisi lain, intervensi tidak akan seefektif yang seharusnya untuk meningkatkan perawatan pasien, karena jumlah pasien gagal jantung beresiko tinggi tidak cukup.



Gambar 68. Studi kasus-menggunakan kurva ROC

Jadi bagaimana kita menentukan model mana yang optimal? Seperti yang Anda lihat pada gambar di atas, model optimal adalah model yang memberikan pemisahan maksimum antara kurva ROC biru dan baseline merah.

Kita bisa melihat bahwa model ketiga 3, dengan biaya relatif kesalahan klasifikasi dari 4-1, adalah yang terbaik dari 4 model. Dan jika ditanya, ROC mewakili kurva operasi karakteristik penerima, yang pertama kali dikembangkan selama Perang Dunia II untuk mendeteksi pesawat musuh di radar.

Sejak itu, itu juga telah digunakan di banyak daerah lain. Saat ini, ini biasa digunakan dalam machine learning dan penambangan data. Kurva ROC adalah alat diagnostik yang berguna untuk menentukan model klasifikasi yang optimal. Kurva ini mengkuantifikasi kinerja model klasifikasi biner, mendeklasifikasi hasil ya dan tidak ketika kriteria diskriminasi diubah.

Dalam hal ini, kriterianya adalah biaya relatif dari kesalahan klasifikasi. Dengan memplot tingkat positif benar terhadap tingkat positif palsu untuk nilai yang berbeda dari biaya relatif kesalahan klasifikasi, kurva ROC memfasilitasi pemilihan model yang optimal.

e. Model Evaluation

Mengapa Evaluasi Model Penting? Data scientist membuat sebuah model. Seringkali, terdengar Data scientist mendiskusikan bagaimana mereka bertanggung jawab untuk membangun model sebagai produk atau membuat banyak model yang saling membangun yang memengaruhi strategi bisnis. Aspek pengembangan model machine learning yang mendasar dan menantang adalah mengevaluasi kinerjanya. Tidak seperti model statistik yang mengasumsikan bahwa distribusi data akan tetap sama, distribusi data dalam model pembelajaran mesin dapat bergeser dari waktu ke waktu. Mengevaluasi model dan mendeteksi penyimpangan distribusi memungkinkan orang untuk mengidentifikasi kapan pelatihan ulang model machine learning diperlukan. Dalam laporan "Evaluating Machine Learning Models" dianjurkan untuk mempertimbangkan evaluasi model diawal proyek apapun karena akan membantu menjawab pertanyaan seperti "bagaimana saya bisa mengukur keberhasilan proyek ini?" dan hindari "bekerja pada proyek yang dirumuskan dengan buruk dimana pengukuran yang baik tidak jelas atau tidak layak"

Ada beberapa tahap dalam mengembangkan model machine learning dan itu berarti ada banyak tempat dimana seseorang perlu mengevaluasi model. Untuk mempertimbangkan evaluasi model selama tahap pembuatan prototipe, atau ketika "kami mencoba model yang berbeda untuk menemukan yang terbaik (pemilihan model)". Metrik evaluasi terkait dengan tugas machine learning" dan bahwa "ada metrik yang berbeda untuk tugas". Beberapa metrik evaluasi yang disusun dalam membuat laporan terdiri klasifikasi, regresi, dan peringkat untuk pembelajaran yang diawasi.

f. Matrik Evaluation

1. Klasfikasi

Mengenai klasifikasi, bahwa diantara metrik yang paling populer untuk mengukur kinerja klasifikasi termasuk akurasi, matriks kebingungan, log-loss, dan AUC (Area Under Curve). Sementara akurasi "mengukur seberapa sering pengklasifikasi

membuat prediksi yang benar" karena "rasio antara jumlah prediksi yang benar dan jumlah prediksi (jumlah data pada pengujian), matriks kebingungan" menunjukkan perincian yang lebih rinci tentang klasifikasi yang benar dan salah untuk setiap kelas.“ Menggunakan matriks konfusi berguna ketika ingin memahami perbedaan antara kelas, terutama ketika “biaya kesalahan klasifikasi mungkin berbeda untuk dua kelas, atau salah satu kelas mungkin berbeda. memiliki lebih banyak data uji dari satu kelas daripada yang lain. Misalnya, konsekuensi dari membuat positif palsu atau negatif palsu dalam diagnosis kanker berbeda.

Adapun kerugian log (kerugian logaritmik), bahwa “jika output mentah dari pengklasifikasi adalah probabilitas numerik alih-alih label kelas 0 atau 1, maka kerugian log dapat digunakan. Probabilitas dapat dipahami sebagai ukuran kepercayaan karena "adalah pengukuran akurasi "lunak" yang menggabungkan gagasan kepercayaan probabilistik ini." Adapun AUC, sebagai “salah satu cara untuk meringkas kurva ROC menjadi satu angka, sehingga dapat dibandingkan dengan mudah dan otomatis.” Kurva ROC adalah kurva utuh dan “memberikan detail bernuansa pengklasifikasi”.

2. Rangking

Salah satu metrik peringkat utama, presisi-recall, juga populer untuk tugas klasifikasi”. Meskipun ini adalah dua metrik, mereka biasanya digunakan bersama. Secara matematis, presisi dan daya ingat dapat didefinisikan sebagai berikut:

precision = #jawaban benar yang menyenangkan/# total item yang dikembalikan oleh ranker

recall = # jawaban benar yang menyenangkan/ # total item yang relevan."

Dalam implementasi yang mendasarinya, pengklasifikasi dapat menetapkan skor numerik untuk setiap item kategori label kelas, dan pemeringkat dapat dengan mudah memesan item dengan skor mentah. Rekomendasi berpotensi menjadi contoh lain, masalah peringkat atau model regresi. Rekomendasi dapat bertindak baik sebagai pemeringkat atau prediktor skor. Dalam kasus pertama, outputnya adalah daftar peringkat item untuk setiap pengguna. Dalam kasus prediksi skor,

pemberi rekomendasi perlu mengembalikan skor yang diprediksi untuk setiap pasangan item pengguna—ini adalah contoh model regresi".

3. Regresi

Dengan regresi, tugas regresi, model belajar memprediksi skor numerik." Seperti disebutkan sebelumnya, rekomendasi yang dipersonalisasi adalah ketika "mencoba memprediksi peringkat pengguna untuk suatu item". Salah satu "metrik yang paling umum digunakan untuk tugas regresi adalah RMSE (root-mean-square-error" yang juga dikenal sebagai RMSD (root-mean-square-deviation). RSME umum digunakan, ada beberapa tantangan. RSME sangat "sensitif terhadap outlier besar. Jika regressor berkinerja sangat buruk pada satu titik data, kesalahan rata-rata bisa sangat besar" atau bahwa "rata-rata tidak kuat (untuk outlier besar) Akan selalu ada "pencilan" dengan data nyata dan "model mungkin tidak akan berkinerja sangat baik pada mereka. Jadi penting untuk melihat penaksir kinerja yang kuat yang tidak terpengaruh oleh outlier besar." bahwa melihat persentase absolut median berguna karena "memberi kita ukuran relatif dari kesalahan tipikal."

4. Mekanisme Evaluasi offline

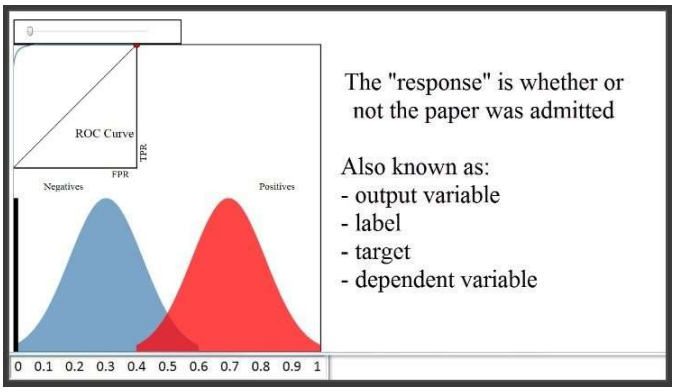
Model harus dievaluasi pada kumpulan data yang secara statistik independen dari yang dilatihnya. Mengapa? Karena kinerjanya pada set pelatihan merupakan perkiraan yang terlalu optimis dari kinerja sebenarnya pada data baru. Proses pelatihan model sudah disesuaikan dengan data pelatihan. Evaluasi yang lebih adil akan mengukur kinerja model pada data yang belum terlihat. Dalam istilah statistik, ini memberikan perkiraan kesalahan generalisasi, yang mengukur seberapa baik model digeneralisasi ke data baru."

Bahwa peneliti dapat menggunakan validasi terus-menerus sebagai cara untuk menghasilkan data baru. Validasi tahan, "dengan asumsi bahwa semua titik data adalah i.i.d. (terdistribusi secara independen dan identik), secara acak memegang sebagian data untuk validasi. Melatih model pada bagian data yang lebih besar dan mengevaluasi metrik validasi pada set penangguhan yang lebih kecil." Teknik

resampling seperti bootstrap atau validasi silang juga dapat digunakan saat membutuhkan mekanisme yang menghasilkan kumpulan data tambahan. Bootstrapping “menghasilkan beberapa kumpulan data dengan mengambil sampel dari satu kumpulan data asli. Setiap kumpulan data "baru" dapat digunakan untuk memperkirakan jumlah minat. Karena ada beberapa kumpulan data dan oleh karena itu banyak perkiraan, seseorang juga dapat menghitung hal-hal seperti varians atau interval kepercayaan untuk perkiraan tersebut." Validasi silang, berguna ketika kumpulan data pelatihan sangat kecil sehingga seseorang tidak mampu melakukannya, menyimpan sebagian data hanya untuk tujuan validasi.” Meskipun ada banyak varian validasi silang, salah satu yang paling umum digunakan adalah validasi silang k-fold yang membagi set data pelatihan menjadi k-folds….setiap k folds secara bergiliran menjadi set validasi hold-out; sebuah model dilatih pada sisa lipatan k-1 dan diukur pada lipatan yang ditahan. Performa keseluruhan diambil sebagai rata-rata performa pada semua k fold. Ulangi prosedur ini untuk semua pengaturan hyperparameter yang perlu dievaluasi, lalu pilih hyperparameter yang menghasilkan rata-rata k-fold tertinggi.”

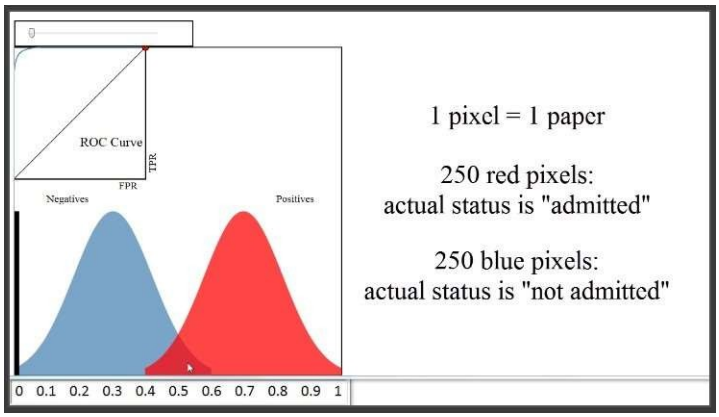
g. Kurva ROC dan Under Area Curve(UAC)

Kurva ROC adalah cara yang paling umum digunakan untuk memvisualisasikan kinerja pengklasifikasi biner, dan AUC adalah cara terbaik untuk meringkas kinerjanya dalam satu angka. Misalnya, membuat pengklasifikasi untuk memprediksi apakah paper akan diterima di jurnal, berdasarkan berbagai faktor. Fitur-fiturnya mungkin panjang makalah, jumlah penulis, jumlah makalah yang sebelumnya telah dikirimkan oleh penulis ke jurnal, dan lain-lain. Tanggapannya adalah apakah paper itu diterima atau tidak.



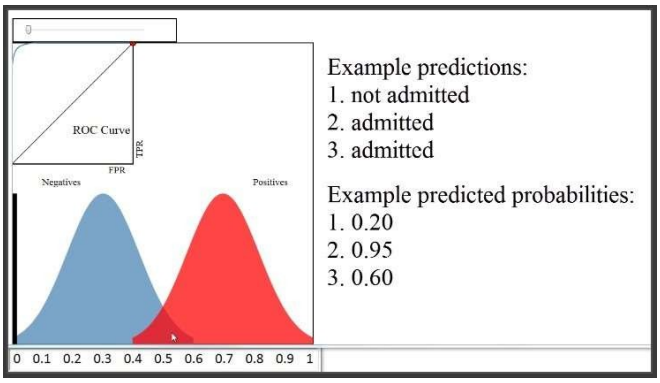
Gambar 69. Diagram: tanggapan paper diterima atau tidak

Perhatikan distribusi biru dan merah. Setiap piksel biru dan merah mewakili paper yang ingin diprediksi status penerimaannya. Ini adalah set validasi ("hold-out"), jadi dapat diketahui status penerimaan sebenarnya dari setiap makalah. 250 piksel merah adalah makalah yang benar-benar diterima, dan 250 piksel biru adalah makalah yang tidak diterima.



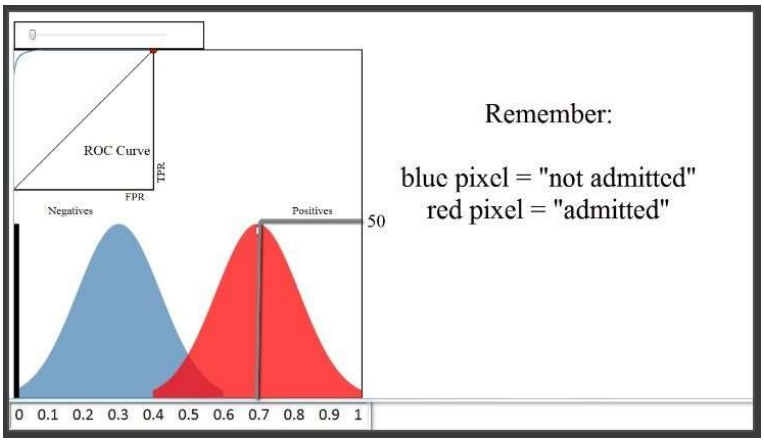
Gambar 70. Diagram: visualisasi probabilitas prediksi

Karena ini adalah set validasi, dapat dinilai seberapa baik kinerja model dengan membandingkan prediksi model dengan status penerimaan sebenarnya dari 500 paper tersebut. Asumsi bahwa menggunakan metode klasifikasi seperti regresi logistik yang tidak hanya dapat membuat prediksi untuk setiap paper, tetapi juga dapat menampilkan prediksi probabilitas penerimaan untuk setiap paper. Distribusi biru dan merah ini adalah salah satu cara untuk memvisualisasikan bagaimana probabilitas yang diprediksi tersebut dibandingkan dengan status sebenarnya.



Gambar 71. Diagram: contoh probabilitas prediksi

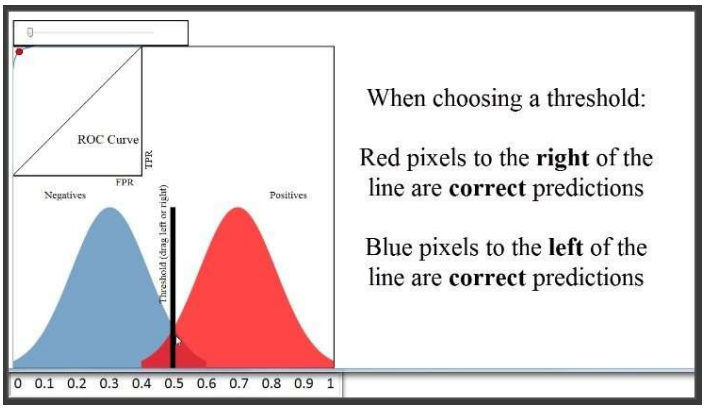
Sumbu x mewakili probabilitas yang diprediksi, dan sumbu y mewakili hitungan pengamatan, seperti histogram. Perkirakan bahwa ketinggian pada 0,1 adalah 10 piksel. Plot ini memberitahu bahwa ada 10 makalah yang Anda prediksi probabilitas masuknya 0,1, dan status sebenarnya untuk semua 10 makalah adalah negatif (artinya tidak diterima). Ada sekitar 50 makalah yang diprediksikan kemungkinan masuknya 0,3, dan tidak satu pun dari 50 itu yang diterima. Ada sekitar 20 makalah yang diprediksi probabilitasnya 0,5, dan setengahnya diterima dan separuh lainnya tidak. Ada 50 makalah yang probabilitasnya Anda perkirakan 0,7, dan semuanya diterima. Dan seterusnya.



Gambar 72. Diagram: classifier-nilai ambang

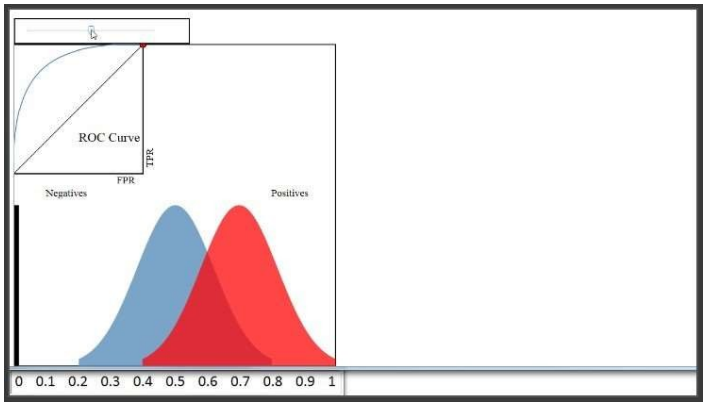
Bahwa classifier bekerja dengan cukup baik, karena melakukan pekerjaan yang baik untuk memisahkan kelas. Untuk benar-benar membuat prediksi kelas, dapat menetapkan "ambang" Anda pada 0,5, dan mengklasifikasikan semua di atas 0,5

sebagai diterima dan semua di bawah 0,5 sebagai tidak diterima, yang akan dilakukan sebagian besar metode klasifikasi secara default. Dengan ambang itu, tingkat akurasi akan berasda diatas 90%, yang mungkin sangat bagus.



Gambar 73. Diagram: classifier-nilai ambang yang diubah

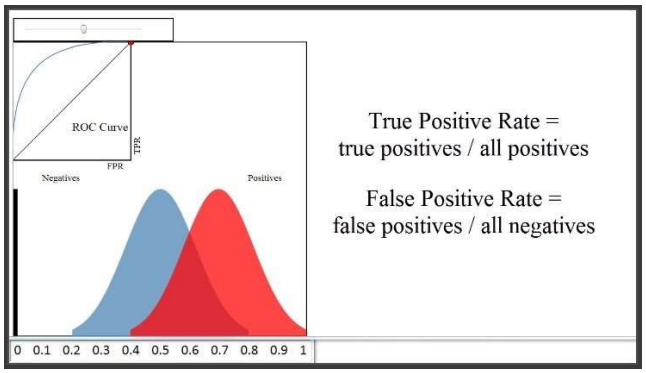
Bila classifier tidak melakukannya dengan baik dan pindahkan ke distribusi biru. Dapat terlihat bahwa ada lebih banyak tumpang tindih disini, dan dimana pun menetapkan ambang batas, akurasi klasifikasi akan jauh lebih rendah daripada sebelumnya.



Gambar 74. Diagram: classifier- tingkat Positif Benar dan Palsu

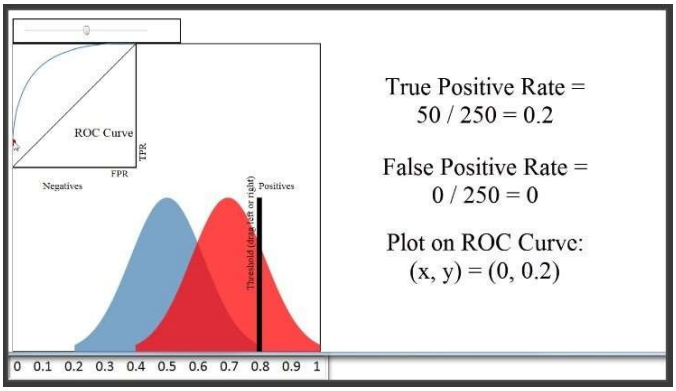
Kurva ROC adalah plot Tingkat Positif Benar (pada sumbu y) versus Tingkat Positif Palsu (pada sumbu x) untuk setiap ambang klasifikasi yang mungkin. Sebagai pengingat, True Positive Rate menjawab pertanyaan, "Bila klasifikasi sebenarnya positif (artinya diakui), seberapa sering pengklasifikasi memprediksi positif?" Tingkat Positif Palsu menjawab pertanyaan, "Bila klasifikasi sebenarnya negatif (artinya tidak

diterima), seberapa sering pengklasifikasi salah memprediksi positif?" Baik Tingkat Positif Benar dan Tingkat Positif Palsu berkisar dari 0 hingga 1.



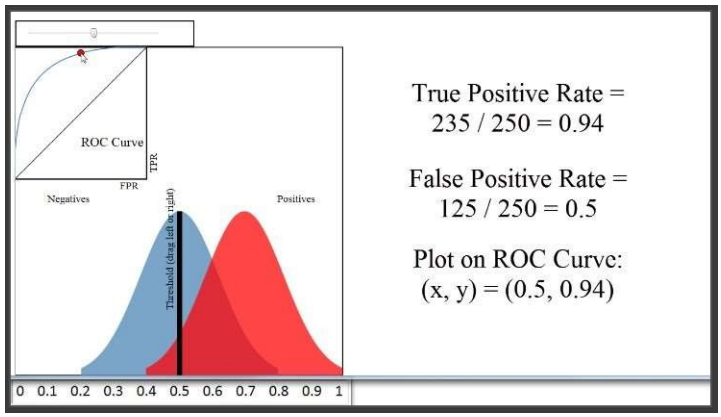
Gambar 75. Diagram: classifier- penentuan nilai ambang batas

Contoh ambang batas untuk mengklasifikasikan paper diterima. Ambang batas 0,8 akan mengklasifikasikan 50 makalah diterima, dan 450 makalah tidak diterima. True Positive Rate adalah piksel merah di sebelah kanan garis dibagi semua piksel merah, atau 50 dibagi 250, yaitu 0,2. Tingkat Positif Palsu adalah piksel biru disebelah kanan garis dibagi dengan semua piksel biru, atau 0 dibagi 250, yaitu 0. Dengan demikian, kita akan memplot titik pada 0 pada sumbu x, dan 0,2 pada sumbu sumbu y yang benar



Gambar 76. Diagram: classifier- plot titik pada sumbu x dan y

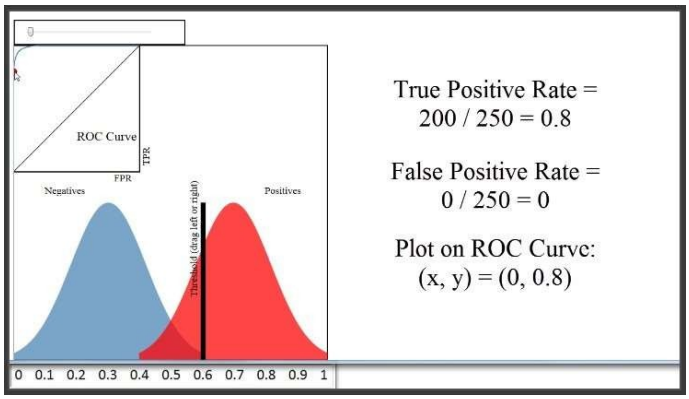
Menetapkan ambang batas yang berbeda dari 0,5. Akan mengklasifikasikan 360 makalah sebagai diterima, dan 140 makalah tidak diterima. Tingkat Positif Benar akan menjadi 235 dibagi 250, atau 0,94. Tingkat Positif Palsu akan menjadi 125 dibagi 250, atau 0,5. Jadi, kita akan memplot sebuah titik di 0,5 pada sumbu x, dan 0,94 pada sumbu y, yang ada disini.



Gambar 77. Diagram: classifier- plot on ROC curve

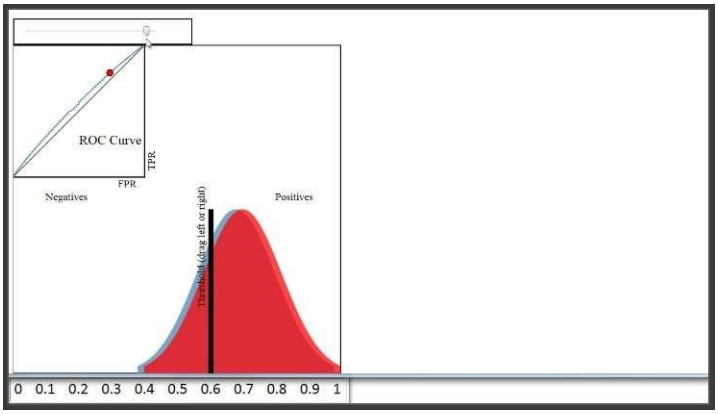
Plot dua poin, tetapi untuk menghasilkan seluruh kurva ROC, yang harus dilakukan adalah memplot True Positive Rate versus False Positive Rate untuk semua kemungkinan ambang klasifikasi yang berkisar dari 0 hingga 1. Itu adalah keuntungan besar dari menggunakan kurva ROC untuk mengevaluasi pengklasifikasi alih-alih metrik yang lebih sederhana seperti tingkat kesalahan klasifikasi, di mana kurva ROC memvisualisasikan semua ambang klasifikasi yang mungkin, sedangkan tingkat kesalahan klasifikasi hanya mewakili tingkat kesalahan untuk satu ambang batas. Perhatikan bahwa tidak dapat benar-benar melihat ambang batas yang digunakan untuk menghasilkan kurva ROC dimana pun pada kurva itu sendiri.

Pindahkan distribusi biru kembali ke tempat sebelumnya. Karena pengklasifikasi melakukan pekerjaan yang sangat baik dalam memisahkan biru dan merah, tetapkan ambang 0,6, memiliki Tingkat Positif Benar 0,8, dan masih memiliki Tingkat Positif Palsu 0.



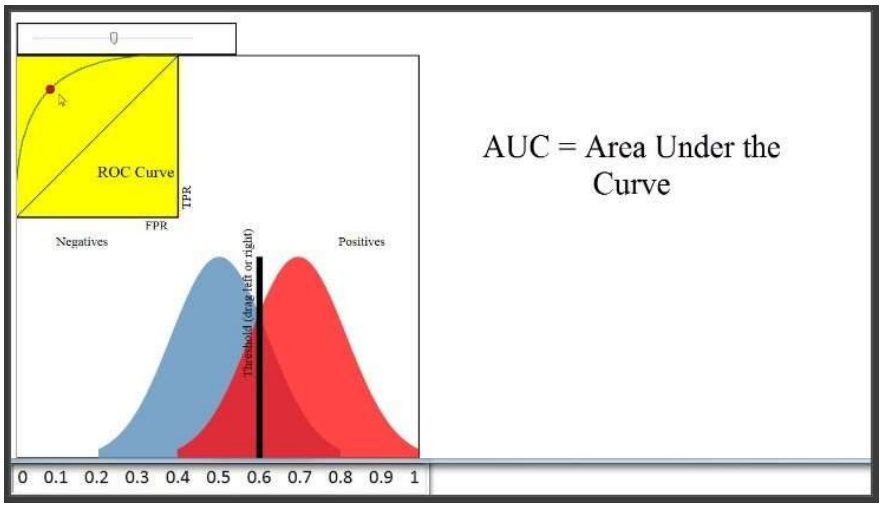
Gambar 78. Diagram: classifier- plot baik dan buruk

Oleh karena itu, pengklasifikasi yang melakukan pekerjaan yang sangat baik memisahkan kelas akan memiliki kurva ROC yang memeluk sudut kiri atas plot. Sebaliknya, pengklasifikasi yang melakukan pekerjaan yang sangat buruk dalam memisahkan kelas akan memiliki kurva ROC yang dekat dengan garis diagonal hitam ini. Baris itu pada dasarnya mewakili pengklasifikasi yang tidak lebih baik dari tebakan acak.



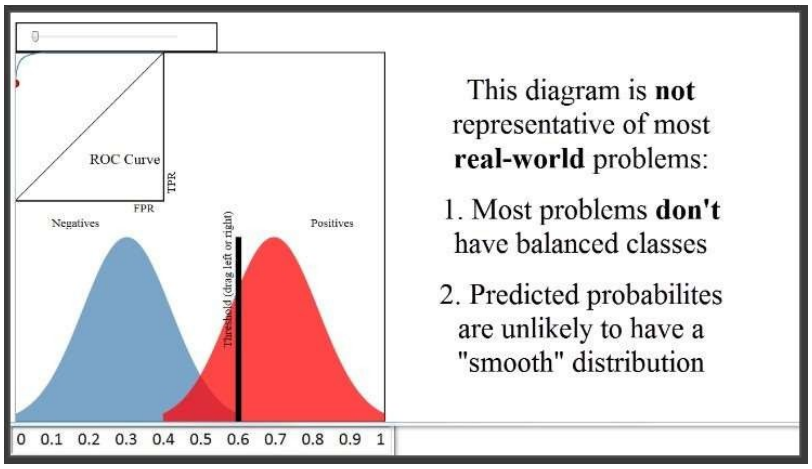
Gambar 79. Diagram: classifier- pemisahan class yang buruk

Secara alami, ingin menggunakan kurva ROC untuk mengukur kinerja pengklasifikasi, dan memberikan skor yang lebih tinggi untuk klasifikasi. Itulah tujuan dari AUC, yang merupakan singkatan dari Area Under the Curve. AUC secara harfiah hanyalah persentase dari kotak ini yang berada di bawah kurva ini. Pengklasifikasi ini memiliki AUC sekitar 0,8, pengklasifikasi yang sangat buruk memiliki AUC sekitar 0,5, dan pengklasifikasi ini memiliki AUC mendekati 1.



Gambar 80. Diagram: classifier- UAC

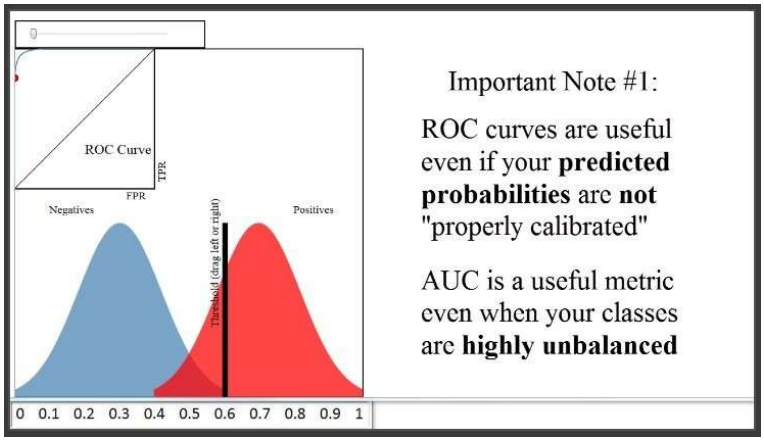
Diagram ini menunjukkan kasus dimana kelas seimbang sempurna, itulah sebabnya ukuran distribusi biru dan merah identik. Dalam sebagian besar masalah dunia nyata, ini tidak terjadi. Misalnya, jika hanya 10% makalah yang diterima, distribusi biru akan sembilan kali lebih besar dari distribusi merah. Namun, itu tidak mengubah bagaimana kurva ROC dihasilkan. Diagram ini menunjukkan kasus dimana probabilitas prediksi memiliki bentuk yang sangat halus, mirip dengan distribusi normal. Itu hanya untuk tujuan demonstrasi. Probabilitas yang dihasilkan oleh pengklasifikasi tidak selalu mengikuti bentuk tertentu.



Gambar 81. Diagram: classifier- tidak representative

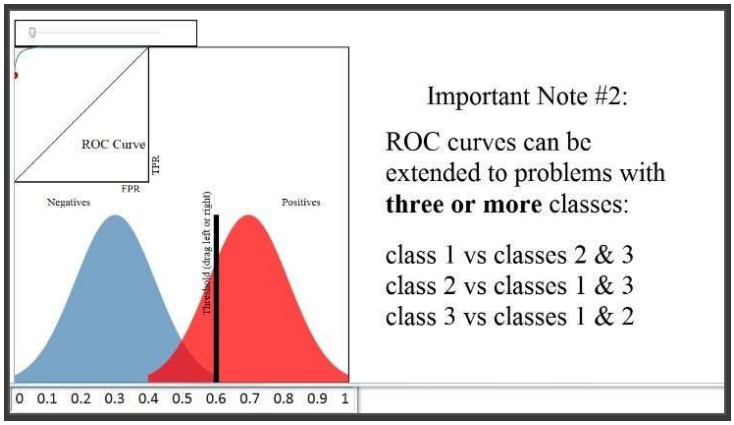
Catatan

Kurva ROC dan AUC tidak sensitif terhadap apakah probabilitas prediksi dikalibrasi dengan benar untuk benar-benar mewakili probabilitas keanggotaan kelas. Dengan kata lain, kurva ROC dan AUC akan identik bahkan jika probabilitas prediksi berkisar antara 0,9 hingga 1, bukan 0 hingga 1, selama urutan pengamatan berdasarkan probabilitas yang diprediksi tetap sama. Yang diperhatikan oleh metrik AUC adalah seberapa baik pengklasifikasi memisahkan dua kelas, dan dengan demikian dikatakan hanya sensitif terhadap urutan peringkat. Dapat dianggap AUC mewakili probabilitas bahwa pengklasifikasi akan memberi peringkat pengamatan positif yang dipilih secara acak lebih tinggi daripada pengamatan negatif yang dipilih secara acak, dan dengan demikian ini adalah metrik yang berguna bahkan untuk kumpulan data dengan kelas yang sangat tidak seimbang.



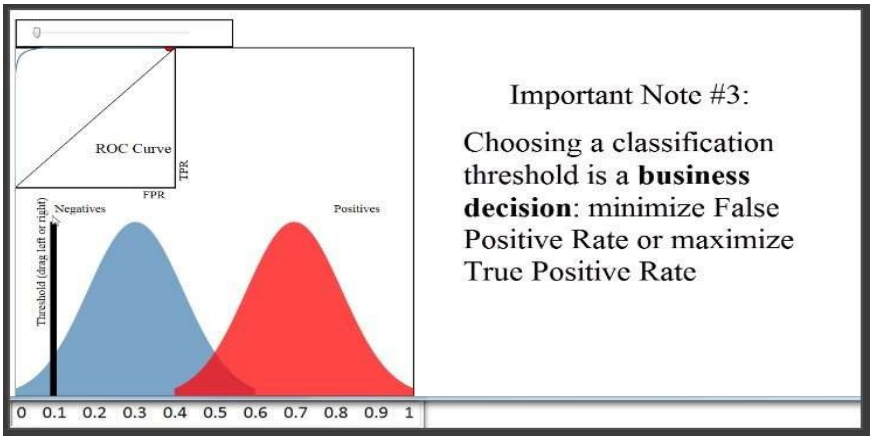
Gambar 82. Diagram: ROC-not properly calibrated

Bahwa kurva ROC dapat diperluas ke masalah klasifikasi dengan tiga atau lebih kelas menggunakan apa yang disebut pendekatan "satu lawan semua". Itu berarti jika memiliki tiga kelas, akan membuat tiga kurva ROC. Pada kurva pertama, memilih kelas pertama sebagai kelas positif, dan mengelompokkan dua kelas lainnya menjadi kelas negatif. Pada kurva kedua, akan memilih kelas kedua sebagai kelas positif, dan mengelompokkan dua kelas lainnya menjadi kelas negatif. Dan seterusnya.



Gambar 83. Diagram: ROC-3 class atau lebih

Menetapkan ambang klasifikasi untuk memprediksi data di luar sampel. Itu sebenarnya lebih merupakan keputusan bisnis, karena harus memutuskan apakah lebih suka meminimalkan Tingkat Positif Palsu atau memaksimalkan Tingkat Positif Sejati. Dalam contoh ini, tidak jelas apa yang harus lakukan. Tapi katakanlah pengklasifikasi digunakan untuk memprediksi apakah transaksi kartu kredit tertentu mungkin penipuan dan dengan demikian harus ditinjau oleh pemegang kartu kredit. Keputusan bisnis mungkin untuk menetapkan ambang batas yang sangat rendah. Itu akan menghasilkan banyak kesalahan positif, tetapi itu mungkin dianggap dapat diterima karena akan memaksimalkan tingkat positif yang sebenarnya dan dengan demikian meminimalkan jumlah kasus di mana contoh nyata penipuan tidak ditandai untuk ditinjau.



Gambar 84. Diagram: ROC-business decision

Selalu harus memilih ambang klasifikasi, tetapi kurva ROC akan membantu memahami secara visual dampak dari pilihan tersebut.

4. Forum Diskusi

   a. Mengapa Evaluasi Model Penting?

   b. Bersama dengan tim kelompok anda, pilihlah 1 dataset yang ada pada materi-materi sebelumnya, berdasarkan Analisa kebutuhan masalah yang diselesaikan, tenting nilai ROC dan UAC untuk memberikan keputusan pada solusi yang ditawarkan.

   c. Berdasarkan dataset yang digunakan tersebut, lakukanlah tahap modelling dan evaluation

C. PENUTUP

1. Rangkuman

   Karena data scientist menghabiskan begitu banyak waktu untuk membuat model, mempertimbangkan metrik evaluasi sejak dini dapat membantu data scientist mempercepat pekerjaan dan menyiapkan proyek untuk sukses. Namun, mengevaluasi model machine learning adalah tantangan yang diketahui. Memperoleh pemahaman mendalam tentang kurva ROC dan AUC bermanfaat bagi data scientist, praktisi machine learning, peneliti medis dan lain sebagainya.

2. Tes Formatif

   Sebagai evaluasi pemahaman materi pada Bab 5, jawablah pertanyaan berikut ini.

   1. Apakah yang dimaksud dengan modelling, dan pada proses data science modelling terdapat pada proses ke berapa?
   2. Apakah tujuan dari modelling, dan pemodelan tersebut berfikus pada model apa saja, serta apakah perbedaan dari tiap-tiap model tersebut?
   3. Dalam metodologi data science deskriptif John Rollins, kerangka kerja dirancang untuk tiga hal, terdiri dari :

a. Pertama, pahami pertanyaan yang menjadi perhatian anda.

b. Kedua, memilih pendekatan atau metode analitis untuk memecahkan masalah.

c. Ketiga, memperoleh, memahami, menyiapkan dan memodelkan data.

d. Kedua, memilih metode analitis untuk memecahkan masalah.

4. Satu training set adalah dua set data historis dimana hasilnya sudah diketahui

   a. True

   b. False

5. Evaluasi mengevaluasi kualitas model, tetapi juga memberikan kesempatan untuk menentukan apakah memenuhi persyaratan awal.

   a. True

   b. False

6. Kurva yang digunakan sebagai alat diagnostik yang berguna untuk menentukan model klasifikasi yang optimal, disebur dengan apa?

7. Beberapa metrik evaluasi yang disusun dalam membuat laporan terdiri dari :

   a. Klasifikasi

   b. Regresi

   c. peringkat untuk supervised learning

   d. peringkat untuk unsupervised learning

8. Aspek pengembangan model machine learning yang mendasar dan menantang adalah?

9. Cara yang paling umum digunakan untuk memvisualisasikan kinerja pengklasifikasi biner, dan AUC adalah cara terbaik untuk meringkas kinerjanya dalam satu angka, hal tersebut menggunakan kurva yang disebut dengan kurva apa?

10. Menggunakan kurva ROC untuk mengukur kinerja pengklasifikasi, dan memberikan skor yang lebih tinggi untuk klasifikasi, hal tersebut merupakan tujuan dari?

Daftar Pustaka

Model Evaluatioan, by Domino On June 14, 2018, http://blog.dominodatalab.com/model-evaluation

Part-4 Data Science Methodology From Modelling to Evaluation,by Ashish Patel, ML Research Lab, Medium, https://medium.com/ml-research-lab/part-4-data-science-methodology-from-modelling-to-evaluation-3fb3c0cdf805

ROC curves and Area Under the Curve explained, https://www.dataschool.io/roc-curves-and-auc-explained/

BAB VI

MARKETING ANALYTIC

A. PENDAHULUAN

Memahami dan menjelaskan marketing analytic, customer analytic dan recommendation system.

B. INTI

1. Capaian Pembelajaran

- **a.** Mampu memahami dan menjelaskan Marketing Analityc pada Data Science
- **b.** Mampu memahami dan menjelaskan apa itu marketing analityc
- **c.** Mampu memahami dan menjelaskan tantangan yang ditemukan dalam menganalisa data
- **d.** Mampu memahami dan menjelaskan tujuan menggunakan software pada Marketing Analityc, kemampuan menggunakan software tersebut, keterampilan yang diperlukan manajer analisis pemasaran
- **e.** Mampu memahami dan menjelaskan tentang defenisi Customer Analytic, pertanyaan yang dapat dijawab, cara kerjanya dan bagaimana cara menerapkannya
- **f.** Mampu memahami dan menjelaskan tentang recomendation system dan metode yang digunakan.

2. Materi

- **a.** Marketing Analityc
- **b.** Tantangan Analisa Data
- **c.** Software Marketing Analityc
- **d.** Customer Analytic

e. Recommended System

3. Uraian Materi

a. Marketing Analityc

Marketing Anaityc(analisis pemasaran) adalah studi tentang data yang dikumpulkan melalui kampanye pemasaran untuk membedakan pola antara hal-hal seperti bagaimana kampanye berkontribusi pada konversi, perilaku konsumen, preferensi regional, preferensi kreatif, dan banyak lagi. Tujuan Marketing Anaityc sebagai praktik adalah menggunakan pola dan temuan ini untuk mengoptimalkan kampanye masa depan berdasarkan apa yang berhasil.

Analisis pemasaran menguntungkan pemasar dan konsumen. Analisis ini memungkinkan pemasar untuk mencapai ROI yang lebih tinggi pada investasi pemasaran dengan memahami apa yang berhasil dalam mendorong konversi, kesadaran merek, atau keduanya. Analytics juga memastikan bahwa konsumen melihat lebih banyak iklan yang ditargetkan dan dipersonalisasi yang sesuai dengan kebutuhan dan minat khusus mereka, daripada komunikasi massal yang cenderung mengganggu. Data pemasaran dapat dianalisis menggunakan berbagai metode dan model tergantung pada (Key Performance Indicators) KPIs yang diukur. Misalnya, analisis kesadaran merek bergantung pada data dan model yang berbeda dari analisis konversi. Beberapa model dan metode analitik populer meliputi:

1. Media Mix Model (MMM): Model atribusi yang melihat data agregat dalam jangka waktu yang lama.
2. Multi-Touch Attribution (MTA): Model atribusi yang menyediakan data tingkat orang dari seluruh perjalanan pembeli.
3. Unified Marketing Measurement (UMM): Suatu bentuk pengukuran yang mengintegrasikan berbagai model atribusi termasuk MMM dan MTA ke dalam metrik keterlibatan yang komprehensif.

Dalam pemasaran modern, analitik yang akurat lebih penting dari sebelumnya. Konsumen menjadi sangat selektif dalam memilih media bermerek yang mereka gunakan dan media yang mereka abaikan. Jika mereka ingin menarik perhatian pembeli yang ideal, mereka harus mengandalkan analitik untuk membuat iklan pribadi yang ditargetkan berdasarkan minat individu, bukan asosiasi demografis yang lebih luas. Ini akan memungkinkan tim pemasaran untuk menayangkan iklan yang tepat, pada waktu yang tepat, disaluran yang tepat untuk menggerakkan konsumen ke saluran penjualan.

Bagaimana Organisasi Menggunakan Analisis Pemasaran? Data analitik pemasaran dapat membantu bisnis membuat keputusan tentang berbagai hal termasuk pembaruan produk, pencitraan merek, dan lainnya. Penting untuk mengambil data dari berbagai sumber (online dan offline) untuk mencegah tampilan terfragmentasi. Dengan menggunakan data ini, dapat memperoleh wawasan tentang hal-hal berikut:

1. Intelijen Produk

   Kecerdasan produk melibatkan menyelam jauh ke dalam produk merek serta bagaimana produk tersebut menumpuk di pasar. Biasanya dilakukan dengan berbicara kepada konsumen, polling audiens target atau melibatkan mereka dengan survei, organisasi dapat lebih memahami pembeda dan keunggulan kompetitif produk mereka. Dari sana, tim dapat lebih menyelaraskan produk dengan minat dan masalah unik konsumen yang membantu mendorong konversi.

2. Tren dan Preferensi Pelanggan

   Analytics dapat memberi tahu banyak tentang konsumen Anda. Pesan/materi iklan apa yang sesuai dengan mereka? Produk mana yang mereka beli dan mana yang pernah mereka teliti sebelumnya? Iklan mana yang menghasilkan konversi dan mana yang diabaikan?

3. Tren Pengembangan Produk

Analytics juga dapat menawarkan wawasan tentang jenis fitur produk yang diinginkan konsumen. Tim pemasaran dapat meneruskan informasi ini ke pengembangan produk untuk iterasi di masa mendatang.

4. Dukungan Pelanggan

   Analytics juga membantu mengungkap area perjalanan pembeli yang dapat disederhanakan atau ditingkatkan. Di mana klien Anda berjuang? Apakah ada cara untuk menyederhanakan produk Anda atau membuat proses check-out lebih mudah?

5. Pesan dan Media

   Analisis data dapat menentukan di mana pemasar memilih untuk menampilkan pesan untuk konsumen tertentu. Ini menjadi sangat penting karena banyaknya saluran. Selain saluran pemasaran tradisional seperti media cetak, televisi, dan siaran, pemasar juga harus mengetahui saluran digital dan jaringan media sosial mana yang disukai konsumen. Analytics menjawab pertanyaan kunci berikut: Media apa yang harus Anda beli? Mana yang paling banyak mendorong penjualan? Pesan apa yang beresonansi dengan audiens Anda?

6. Kompetisi

   Bagaimana upaya pemasaran Anda dibandingkan dengan pesaing Anda? Bagaimana Anda bisa menutup celah itu jika memang ada? Apakah ada peluang yang dimanfaatkan pesaing Anda yang mungkin Anda lewatkan?

7. Prediksi Hasil Masa Depan

   Jika Anda memiliki pemahaman menyeluruh tentang mengapa kampanye berhasil, Anda akan dapat menerapkan pengetahuan itu ke kampanye mendatang untuk meningkatkan ROI.

b. Tantangan Analisis Data

Sementara analisis pemasaran sangat penting untuk kampanye yang sukses, proses analisis menimbulkan tantangan utama karena banyaknya jumlah data yang sekarang dapat dicapai oleh pemasar. Ini berarti bahwa pemasar harus menentukan cara terbaik

untuk mengatur data kedalam format yang dapat dicerna untuk mendapatkan wawasan yang dapat ditindaklanjuti.

Beberapa tantangan analitik pemasaran terbesar yang dihadapi saat ini adalah:

1. Kuantitas Data: Data besar muncul selama era digital, memungkinkan tim pemasaran mencatat setiap klik, tayangan, dan tampilan konsumen. Namun jumlah data ini tidak relevan jika tidak dapat disusun dan dianalisis untuk wawasan yang memungkinkan pengoptimalan dalam kampanye. Ini membuat pemasar bergulat dengan cara terbaik mengatur data untuk mengevaluasi artinya. Faktanya, penelitian menunjukkan bahwa ilmuwan data berpengalaman menghabiskan sebagian besar waktu mereka untuk berdebat dan memformat data, daripada menganalisisnya
2. Kualitas Data: Tidak hanya ada masalah dalam hal banyaknya informasi yang harus disaring oleh organisasi, tetapi data ini sering dianggap tidak dapat diandalkan. Menurut Forrester, 21 persen anggaran media responden terbuang percuma karena kualitas data yang buruk. Ini berarti satu dolar dari setiap 5 dolar tidak digunakan secara efektif. Selama setahun, dolar ini dapat bertambah, menghasilkan $1,2 juta dolar dan $16,5 juta dolar dari anggaran yang terbuang untuk perusahaan skala menengah dan tingkat perusahaan. Organisasi membutuhkan proses untuk menjaga kualitas data, sehingga karyawan dapat memanfaatkan informasi yang akurat untuk membuat keputusan yang tepat.
3. Kurangnya Data Sciantist: Bahkan jika perusahaan memiliki akses ke data yang tepat, banyak yang tidak memiliki akses ke orang yang tepat. Faktanya, menurut survei oleh The CMO, hanya 1,9% perusahaan yang percaya bahwa mereka memiliki orang yang tepat untuk sepenuhnya memanfaatkan analitik pemasaran.
4. Memilih Model Atribusi: Menentukan model yang memberikan wawasan yang tepat bisa jadi rumit. Misalnya, pemodelan campuran media dan atribusi multi-sentuh menawarkan wawasan yang sama sekali berbeda – data gabungan yang

berfokus pada kampanye dan data konsumen tingkat orang. Model yang dipilih pemasar akan menentukan jenis wawasan yang mereka terima. Analisis keterlibatan di begitu banyak saluran dapat menimbulkan kebingungan saat tiba waktunya untuk memilih model yang tepat.

5. Korelasi Data: Dalam nada yang sama, karena pemasar mengumpulkan data dari begitu banyak sumber yang berbeda, mereka harus menemukan cara untuk menormalkannya agar dapat dibandingkan. Sangat menantang untuk membandingkan keterlibatan online dan offline, karena biasanya diukur dengan model atribusi yang berbeda. Di sinilah pengukuran pemasaran terpadu dan platform analisis pemasaran menunjukkan nilai sebenarnya, mengatur data dari sumber yang berbeda.

c. Software Marketing Analityc

Software analitik pemasaran mengatasi tantangan dengan mengumpulkan, mengatur, dan menghubungkan data berharga dengan cepat, memungkinkan pemasar melakukan pengoptimalan kampanye secara real-time. Platform pemasaran modern sangat berharga untuk kecepatan penyimpanan dan pemrosesan data dalam jumlah besar. Salah satu kelemahan utama dari memiliki akses ke begitu banyak data adalah bahwa pemasar tidak mungkin menguraikan semuanya dalam waktu untuk membuat pengoptimalan waktu nyata. Di situlah kekuatan pemrosesan platform analitik canggih berperan, memungkinkan pemasar menyesuaikan materi iklan atau penempatan iklan sesuai kebutuhan sebelum kampanye berakhir, sehingga meningkatkan potensi ROI. Selain itu, banyak platform sekarang memanfaatkan pengukuran pemasaran terpadu, untuk menormalkan dan menggabungkan data pemasaran dari berbagai saluran dan kampanye, menyederhanakan analisis. Platform analitik canggih lebih dari sekadar mengukur keterlibatan konsumen untuk menawarkan wawasan tentang ekuitas merek dan bagaimana segmen audiens tertentu bereaksi terhadap elemen kreatif. Ini membantu pemasar menentukan ROI pembangunan merek dengan lebih baik, serta cara mempersonalisasi pengalaman bermerek lebih lanjut.

Kemampuan Marketing Analytic Software

Saat menerapkan solusi pengukuran pemasaran, pertimbangkan fitur dan kemampuan utama berikut dari perangkat lunak analisis pemasaran:

1. Analisis dan Wawasan real-time
2. Kemampuan Pengukuran Merek
3. Granular, Data Tingkat Orang
4. Kemampuan untuk Menghubungkan Metrik Atribusi Online dan Offline
5. Wawasan Pelanggan dan Pasar yang Dikontekstualisasikan
6. Rekomendasi Rencana Media Tahunan

Keterampilan yang Dibutuhkan Manajer Analisis Pemasaran

Saat tim pemasaran berusaha melakukan analisis kualitas yang mengarah pada kampanye yang lebih menarik dan menguntungkan, mereka harus fokus pada mempekerjakan manajer analitik yang dapat:

1. Melakukan Analisis Kualitas: Pertama dan yang paling jelas, manajer analitik harus memiliki pengalaman mengevaluasi kumpulan data besar untuk membedakan wawasan termasuk pola pembelian dan tren keterlibatan dalam audiens target.
2. Buat Rekomendasi Pengoptimalan: Setelah wawasan data diperoleh, kemampuan untuk menghasilkan rekomendasi untuk meningkatkan kampanye yang berkinerja buruk berdasarkan tren sangat penting. Misalnya, data mungkin menunjukkan bahwa satu konsumen terlibat dengan konten bermerek hanya di malam hari, menginformasikan perubahan strategi untuk menayangkan iklan di perjalanan pulang konsumen, bukan di perjalanan pagi.

3. Memahami Tren Konsumen dan MarTech: Manajer Analytics juga harus tetap mengikuti tren konsumen dan MarTech. Memahami tuntutan konsumen untuk pengalaman omnichannel yang mulus dan bagaimana pembeli terlibat dengan augmented reality dan virtual reality tentu akan berperan dalam menentukan langkah selanjutnya untuk peluang pengoptimalan.
4. Bekerja dengan Alat Analytics: Selanjutnya, manajer analitik harus siap dan nyaman dengan berbagai alat otomatisasi dan platform analitik, karena peran penting alat ini dalam mengurangi waktu dari keterlibatan konsumen hingga wawasan konsumen.
5. Berkolaborasi dengan Pemangku Kepentingan: Terakhir, anggota tim analitik harus dapat menggunakan data yang mereka gunakan untuk menceritakan kisah yang menarik kepada pemangku kepentingan, dan menunjukkan cara departemen lain, seperti penjualan atau pengembangan produk, dapat menggunakan temuan ini untuk mendorong keterlibatan dan konversi.

Cara memulai proses analisis pemasaran. Jika ingin meningkatkan kemampuan analitik, berikut adalah empat langkah yang harus diambil diawal program Anda:

1. Pahami Apa yang Ingin Anda Ukur

   Ada banyak aspek kampanye pemasaran yang dapat diukur: tingkat konversi, prospek yang diperoleh, dan pengenalan merek, untuk beberapa nama. Pahami masalah yang Anda coba pecahkan atau wawasan yang coba Anda kumpulkan saat mulai menganalisis data Anda.
2. Tetapkan Tolok Ukur

   Seperti apa kampanye yang berhasil? Ini akan menentukan jenis data dan metrik yang dikumpulkan pemasar. Misalnya, jika tujuannya adalah untuk meningkatkan kesadaran merek – tolok ukur keberhasilan mungkin berupa peningkatan persentase loyalitas merek yang ditunjukkan di panel pelanggan, bukan klik atau tayangan online.
3. Nilai Kemampuan Anda Saat Ini

Apa yang perusahaan Anda lakukan hari ini? Apa titik lemah Anda? Baik menilai hasil kampanye offline atau mengidentifikasi media yang paling mungkin berkonversi, memahami titik-titik lemah ini dapat membantu Anda memperkuat program Anda.

4. Terapkan Alat Analisis Pemasaran

   Alat analisis pemasaran akan semakin penting karena konsumen menjadi lebih selektif dan kumpulan data tumbuh. Platform lanjutan, seperti Platform Pengukuran dan Pengoptimalan Pemasaran kami menggunakan pengukuran pemasaran terpadu untuk membantu pemasar mengidentifikasi pesan yang beresonansi dan jenis media yang berkonversi. Ini memberikan gambaran menyeluruh tentang kampanye mana yang berhasil dan mana yang berkinerja buruk secara real time.

d. Customer Analityc

Customer Analityc adalah proses yang digunakan perusahaan untuk menangkap dan menganalisis data pelanggan untuk membuat keputusan yang lebih baik. Analisis pelanggan sering kali datang dalam bentuk perangkat lunak yang melengkapi perusahaan dengan wawasan tentang perilaku pengguna mereka. Wawasan ini memperkuat upaya dan studi penjualan, pemasaran, dan pengembangan produk bisnis menunjukkan bahwa perusahaan yang menggunakan analitik pelanggan lebih menguntungkan.

Mengapa perusahaan menggunakan analisis pelanggan?

Analisis pelanggan membantu bisnis memecahkan masalah besar menjadi jawaban yang dapat dikelola. Ketika perusahaan perlu melihat bagaimana pelanggan mereka berperilaku, baik sebagai individu atau secara keseluruhan, analitik pelanggan menerjemahkan tindakan mereka sehingga lebih mudah dipahami. Ini membantu perusahaan membuat keputusan yang lebih baik tentang penetapan harga, promosi, dan manajemen. Keputusan yang lebih baik sangat bagus untuk bisnis. Menurut McKinsey, “Perusahaan yang menggunakan analisis pelanggan secara komprehensif

melaporkan melampaui pesaing mereka dalam hal keuntungan hampir dua kali lebih sering daripada perusahaan yang tidak." Platform analisis konsumen yang baik dapat membantu peningkatan tim Anda:

1. Mobile adoption
2. Retensi pelanggan
3. User engagement/Keterlibatan pengguna
4. In-app purchases

Saat ini, perusahaan lebih bergantung pada analitik daripada sebelumnya karena membantu mereka mengimbangi konsumen yang semakin canggih. Menurut sebuah studi IBM, 85% pelanggan sekarang mengharapkan pengalaman yang mulus dan menginginkan respons yang lebih cepat, mengandalkan dukungan optichannel, dan memiliki lebih sedikit perhatian. Analytics membantu perusahaan mengukur dan meningkatkan produk mereka untuk melayani pelanggan modern.

Bagaimana cara kerja analisis pelanggan?

Analisis pelanggan memberi perusahaan visibilitas penuh tentang bagaimana pelanggan menggunakan produk mereka. Analytics sangat berguna bagi perusahaan dengan penawaran teknologi karena mereka dapat mengumpulkan data langkah demi langkah tentang bagaimana pelanggan, pengguna, atau pelanggan mengalir melalui situs atau aplikasi mereka. Pada tingkat makro, ini memperlihatkan tren utama seperti bagaimana pengguna menemukan produk mereka, fitur mana yang paling mereka sukai, di mana mereka menemukan nilai, dan apa yang menyebabkan mereka pergi.

Pada tingkat mikro, analisis pelanggan memungkinkan perusahaan memahami siapa penggunanya sebagai individu. Mereka dapat mengelompokkan pengguna berdasarkan demografi, minat, dan perilaku serta melihat perjalanan unik mereka. Pengetahuan ini membantu bisnis melayani setiap persona pelanggan dengan lebih baik. Elavon, misalnya, adalah aplikasi pembayaran seluler yang mendapati penggunanya mengeluh bahwa mereka tidak dapat mengunduh aplikasi tersebut. Dengan menggunakan perangkat lunak analitik pelanggan Mixpanel, mereka dapat secara instan

mengidentifikasi semua pengguna yang mencoba mengunduh aplikasi pada sistem OS yang tidak kompatibel, menjangkau, dan menyarankan perbaikan. Pelanggan Mixpanel lainnya, STARZ PLAY, mampu mengelompokkan pelanggan mereka berdasarkan perilaku untuk mendeteksi penipuan dan menguranginya dengan faktor 1.000. Tim mana yang membutuhkan analisis pelanggan?

1. **Pemasaran** dapat menciptakan segmen dan pemirsa yang serupa.
2. **Penjualan** dapat menilai prospek, prospek, dan pengguna.
3. **Produk** dapat mengukur fitur, penggunaan, dan perjalanan pelanggan.

Pertanyaan apa yang dapat dijawab oleh analisis pelanggan?

Platform analitik pelanggan dapat menjawab pertanyaan tentang apa pun yang dapat diukur atau dilacak. Yang mengatakan, jawaban yang mereka berikan hanya sebaik pertanyaan yang diajukan kepada mereka. Pertanyaan terbuka, subjektif, atau relatif seperti, "Apakah produk kami yang terbaik?" sulit untuk dijawab dengan jelas. Pertanyaan bagus bersifat konkret, mudah dibuktikan atau dibantah, dan terkait kembali dengan tujuan bisnis inti seperti akuisisi, pendapatan, retensi, dan keterlibatan. Contoh pertanyaan umum tentang analisis pelanggan:

1. Akuisisi
   a. Saluran mana yang mendorong pelanggan baru paling banyak?
   b. Apa perjalanan pelanggan yang paling umum dari kesadaran ke advokasi?
2. Pendapatan
   a. Apa saluran pendapatan kami yang paling menguntungkan?
   b. Pengguna mana yang paling menguntungkan?
3. Penyimpanan
   a. Di mana kita kehilangan pelanggan dan mengapa?
   b. Perilaku apa yang berkorelasi dengan tingkat retensi yang tinggi?
4. Keterikatan
   a. Fitur apa yang sesuai dengan pelanggan mana?

b. Apa pengalaman optimal bagi pengguna?

Untuk menerima jawaban yang akurat dari platform analisis pelanggan mereka, bisnis menangkap dan menganalisis data mereka. Untuk ini, implementasi yang tepat adalah kuncinya.

Bagaimana cara menerapkan analisis pelanggan?

Untuk menerapkan analisis pelanggan dan memperoleh wawasan yang bermanfaat, perusahaan harus melakukan dua hal:

a. Tangkap, simpan, dan atur data mereka.

b. Analisis dan buat keputusan dengan data tersebut.

Untuk menangkap dan memanfaatkan data, perusahaan harus mengumpulkan banyak data. Mereka dapat menjalankan survei, melakukan riset pengguna, membeli data pihak ketiga, dan, jika mereka menawarkan solusi teknologi, secara pasif mengumpulkan data penggunaan melalui situs atau aplikasi mereka. Dalam kasus situs web, sebagian besar platform analisis pelanggan secara pasif mengumpulkan semua data pengunjung. Dengan aplikasi, perusahaan mungkin perlu menentukan aktivitas atau "peristiwa" tempat data dikumpulkan, seperti login, logout, dan tindakan pengguna. Untuk menyimpan data, sangat berguna untuk memiliki repositori pusat yang menyatukan semua sumber data Anda ke dalam satu tampilan tunggal pelanggan Anda. Ini adalah fitur utama dari sebagian besar platform analisis pelanggan.

Perusahaan yang tidak memelihara data mereka dengan benar akan kesulitan mendapatkan jawaban langsung dari analitik pelanggan mereka. Jika data basi, tidak terstruktur, atau tidak lengkap, itu dapat menyebabkan hasil yang menyesatkan. Ini sering terjadi ketika ada kekurangan proses, banyak pemilik selama bertahun-tahun, atau ketika perusahaan telah membuat platform analitik pelanggan mereka sendiri. Untuk menganalisis dan menginterpretasikan hasil, ada baiknya memiliki manajer produk khusus, desainer UX/UI, atau ilmuwan data. Mereka akan dapat membingkai pertanyaan analitik pelanggan dalam hal tujuan bisnis inti dan mengarahkan perusahaan menuju hasil yang lebih baik. Perusahaan yang berinvestasi dalam analisis

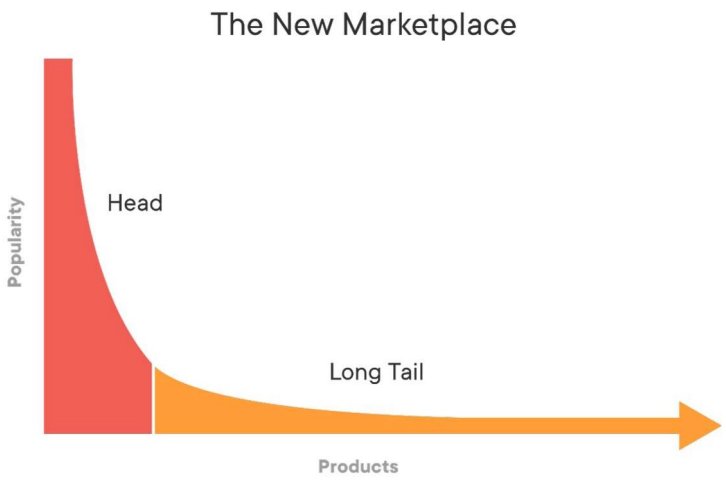
pelanggan akan mendapatkan keuntungan luar biasa atas pesaing mereka karena mereka dapat lebih mudah memahami dan melayani pelanggan mereka.

e. Recommended System/Sistem Rekomendasi

Tujuan dari sistem rekomendasi adalah untuk mengekspos orang ke item yang mereka sukai. Dalam istilah yang tepat, sistem rekomendasi memprediksi preferensi masa depan dari satu set item untuk pengguna, dan merekomendasikan item teratas dari set ini. Di dunia sekarang ini, karena internet dan jangkauan globalnya, orang memiliki lebih banyak pilihan untuk dipilih daripada sebelumnya.

Pertimbangkan untuk membeli album dari toko musik tradisional yang pilihannya selalu terbatas dan sebagian besar bergantung pada ukuran dan jenis toko. Ada batasan fisik berapa banyak lagu, album, dan artis yang bisa ditawarkan. Namun, produk online seperti Spotify memiliki plafon yang jauh lebih tinggi dalam hal ruang penyimpanan. Dengan metode pemilihan produk baru ini, sistem rekomendasi adalah cara yang populer bagi pengguna untuk memilah jutaan lagu untuk menemukan lagu yang disesuaikan secara tepat untuk mereka. Sistem rekomendasi memberikan dampak langsung pada profitabilitas dan kepuasan pelanggan untuk sebagian besar bisnis saat ini. Dengan pilihan yang hampir tak terbatas yang dimiliki konsumen untuk produk online, mereka memerlukan beberapa panduan!

Ide ini dapat diwakili oleh sebuah konsep yang disebut "Long Tail", yang merupakan kumpulan distribusi statistik yang memiliki "ekor" distribusi yang sangat panjang, yang mewakili banyak kejadian frekuensi rendah. Dalam konteks produk konsumen, ada beberapa produk yang akan dibeli semua orang: bola lampu, tisu toilet, roti, dll. Ada juga barang yang jauh lebih tidak jelas: mainan khusus, peralatan olahraga, film. Sistem rekomendasi dibuat untuk membantu konsumen memanfaatkan ekor panjang ini untuk membantu mereka memilih dari banyaknya pilihan yang tersedia bagi mereka melalui internet.



Gambar 85. Konsep Long Tail

Berikut definisi formal sistem rekomendasi dari penulis Bo Xiao dan Izak Benbasat, 2017 :

Sistem Rekomendasi adalah agen perangkat lunak yang memperoleh minat dan preferensi konsumen individu […] dan membuat rekomendasi yang sesuai. Mereka memiliki potensi untuk mendukung dan meningkatkan kualitas keputusan yang dibuat konsumen saat mencari dan memilih produk secara online.

Aplikasi Sistem Rekomendasi

Mari kita pahami apa yang dapat dilakukan semua sistem rekomendasi untuk bisnis:

1. Bantuan dalam menyarankan pedagang/barang yang mungkin diminati pelanggan setelah membeli produk di pasar
2. Perkirakan untung & rugi banyak item yang bersaing dan buat rekomendasi kepada pelanggan (misalnya jual beli saham)
3. Berdasarkan pengalaman pelanggan, rekomendasikan penawaran yang berpusat pada pelanggan atau yang berpusat pada produk
4. Tingkatkan keterlibatan pelanggan dengan memberikan penawaran yang bisa sangat menarik bagi pelanggan

Pendekatan Sistem Rekomendasi

Ada dua jenis utama sistem rekomendasi: tidak dipersonalisasi dan dipersonalisasi. Di sebagian besar bagian ini, kami akan fokus pada sistem rekomendasi yang dipersonalisasi karena di situlah ilmuwan data dapat memberikan nilai paling banyak kepada perusahaan, tetapi untuk memulai, mari kita selidiki beberapa sistem yang tidak dipersonalisasi karena mereka dapat menjadi produktif dengan caranya sendiri.

1. Rekomendasi yang Tidak Dipersonalisasi

   Sistem rekomendasi yang tidak dipersonalisasi telah terjadi jauh sebelum pembelajaran mesin ada di basis pengetahuan publik. Contoh rekomendasi yang tidak dipersonalisasi adalah di YouTube saat merekomendasikan video yang paling banyak dilihat. Ini adalah video yang paling banyak ditonton orang. Untuk sebagian besar, rekomendasi ini tidak terlalu buruk. Lagi pula, ada alasan mengapa sesuatu menjadi populer. Pendekatan ini, bagaimanapun, tidak akan membantu lebih banyak video niche mendapatkan eksposur. Ini juga tidak akan sangat bermanfaat bagi mereka yang memiliki selera yang sangat khusus. Tentu saja, ada kalanya pendekatan sederhana seperti ini mungkin yang terbaik. Contoh rekomendasi popularitas sederhana yang bekerja dengan baik adalah dengan berita.

Gambar 86. Contoh rekomendasi popularitas sederhana

Karena rekomendasi yang tidak dipersonalisasi didasarkan pada seluruh kelompok pengguna, item apa pun yang paling populer pada waktu tertentu akan direkomendasikan kepada Anda, bahkan jika itu adalah sesuatu yang sama sekali tidak Anda minati. Ada begitu banyak item yang terlalu kabur untuk menjadi item "paling populer" yang mungkin membuat hari seseorang. Untuk membuat rekomendasi yang lebih terinformasi, sistem rekomendasi yang dipersonalisasi menggunakan data besar untuk memastikan bahwa pengguna mendapatkan item yang disesuaikan dengan minat pribadi di sana, tidak peduli seberapa niche mereka.

2. Rekomendasi yang Dipersonalisasi

Masalah umum dari sistem rekomendasi yang dipersonalisasi dapat diringkas sebagai:

Mengingat : Profil pengguna "aktif" dan mungkin beberapa konteks situasional, yaitu pengguna menelusuri produk atau melakukan pembelian, dll.

Diperlukan : Membuat satu set item, dan skor untuk setiap item yang direkomendasikan dalam set itu

Profil :

Profil pengguna dapat berisi pembelian sebelumnya, peringkat dalam bentuk implisit atau eksplisit, demografi, dan skor minat untuk fitur item

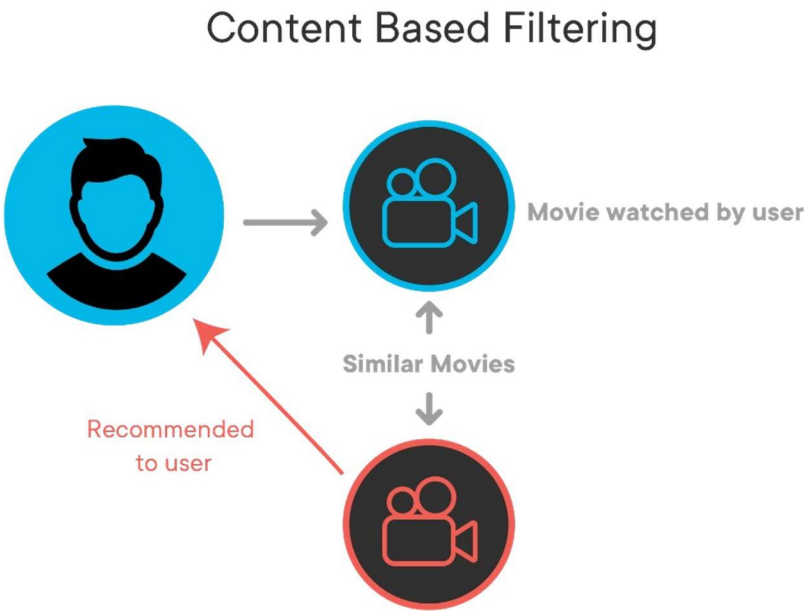
Ada dua cara untuk mengumpulkan data tersebut. Metode pertama adalah meminta peringkat eksplisit dari pengguna, biasanya pada skala peringkat konkret (seperti memberi peringkat film dari satu hingga lima bintang). Yang kedua adalah mengumpulkan data secara implisit karena pengguna berada dalam domain sistem - yaitu, untuk mencatat tindakan pengguna di situs.

Masalah : Kami ingin mempelajari fungsi yang memprediksi skor relevansi untuk item tertentu (biasanya tidak terlihat) berdasarkan profil dan konteks pengguna

pengguna. Dalam sistem rekomendasi yang dipersonalisasi, ada banyak kemungkinan algoritme yang berbeda. Kita akan membahas yang penting sekarang. Masing-masing teknik ini menggunakan metrik kesamaan yang berbeda untuk menentukan seberapa "mirip" item satu sama lain. Metrik kesamaan yang paling umum adalah jarak Euclidean , kesamaan kosinus , korelasi Pearson dan indeks Jaccard (berguna dengan data biner) . Masing-masing metrik jarak ini memiliki kelebihan dan kekurangan tergantung pada jenis peringkat yang Anda gunakan dan karakteristik data Anda.

Rekomendasi Berbasis Konten

Ide Utama : Jika Anda menyukai suatu item, Anda juga akan menyukai item yang "serupa".



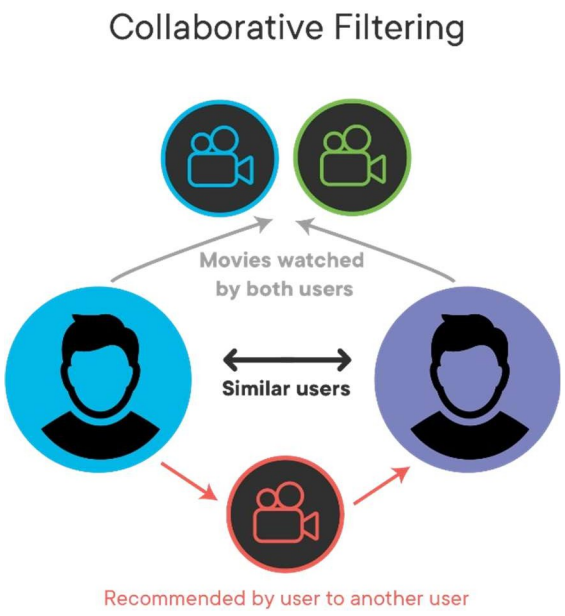
Gambar 87. Content based filtering

Sistem ini didasarkan pada karakteristik item itu sendiri. Jika Anda pernah melihat iklan spanduk yang mengatakan "coba item lain seperti ini", kemungkinan besar itu adalah sistem pemberi rekomendasi berbasis konten. Keuntungan dari sistem pemberi rekomendasi berbasis konten adalah sistem pemberi rekomendasi yang memberi pengguna sedikit lebih banyak informasi tentang mengapa mereka melihat

rekomendasi ini. Jika mereka berada di halaman buku yang sangat mereka sukai, mereka akan senang melihat buku lain yang mirip dengannya. Jika mereka diberitahu bahwa buku ini mirip dengan buku favorit mereka, kemungkinan besar mereka akan mendapatkan buku itu. Kelemahan dari sistem pemberi rekomendasi berbasis konten adalah bahwa mereka sering membutuhkan penandaan manual atau semi-manual untuk setiap produk. Versi sistem pemberi rekomendasi berbasis konten yang lebih canggih memungkinkan pengembangan rata-rata semua item yang disukai pengguna. Hal ini memungkinkan pendekatan yang lebih bernuansa untuk memasukkan lebih dari satu item saat menghitung item mana yang paling mirip.

Sistem Penyaringan Kolaboratif

Ide Utama : Jika pengguna A menyukai item 5, 6, 7, dan 8 dan pengguna B menyukai item 5, 6, dan 7, maka kemungkinan besar pengguna B juga akan menyukai item 8.



Gambar 88. Collaborative Filtering

Sistem penyaringan kolaboratif menggunakan kumpulan peringkat pengguna item untuk membuat rekomendasi. Masalah dengan pemfilteran kolaboratif adalah Anda memiliki apa yang disebut "masalah awal yang dingin". Ide di baliknya adalah, bagaimana merekomendasikan sesuatu berdasarkan aktivitas pengguna jika Anda

tidak memiliki aktivitas pengguna untuk memulai! Hal ini dapat diatasi melalui berbagai teknik. Hal terpenting yang harus disadari adalah bahwa tidak ada satu pun teknik sistem rekomendasi terbaik. Pada akhirnya, yang paling penting adalah sistem apa yang benar-benar membuat orang mendapatkan rekomendasi yang akan mereka tindak lanjuti. Mungkin secara keseluruhan, merekomendasikan item yang paling populer adalah cara yang paling hemat biaya untuk memperkenalkan produk baru kepada pengguna.

Gagasan utama dibalik pemfilteran kolaboratif adalah bahwa pengguna serupa memiliki minat yang sama dan pengguna cenderung menyukai item yang serupa satu sama lain.

Meskipun ini mungkin tidak sepenuhnya benar pada setiap kesempatan, jika kita memiliki kumpulan data yang cukup besar, jika ada pola, pola itu akan mulai muncul. Asumsikan ada beberapa pengguna yang telah membeli item tertentu, kita dapat menggunakan matriks dengan ukuran $\text{num_users} * \text{num_items}$ untuk menunjukkan perilaku pengguna sebelumnya. Setiap sel dalam matriks mewakili pendapat terkait yang dipegang pengguna. Matriks seperti ini disebut Matriks Utilitas. Misalnya, $M_{i, j}$ menunjukkan bagaimana pengguna u menyukai item i. Terkadang peringkat individu ini ditulis sebagai r_{ui} untuk peringkat untuk pengguna tertentu dan item tertentu. Dengan menggunakan tabel di bawah ini sebagai titik referensi, jika kita mengganti subskrip variabel u dan i dengan nilai sebenarnya, maka akan terlihat seperti $r_{\text{Mike},\text{Little Mermaid}} = 3$.

Tabel 2. Matriks Utilitas

	Cerita mainan	**Cinderella**	**Putri duyung kecil**	**Raja singa**
Matt		2		5
Pengetahuan	2		4	
mike		5	3	2

	Cerita mainan	**Cinderella**	**Putri duyung kecil**	**Raja singa**
hutan	5		1	
Taylor	1	5		2

Tugas sistem rekomendasi adalah memberikan peringkat untuk pengguna di tempat yang saat ini kosong. Seperti yang dapat Anda bayangkan, sebagian besar waktu, nilai-nilai ini sebagian besar akan kosong. Untuk pengguna 1, sistem rekomendasi kami akan mencoba dan memprediksi pengguna 1 yang akan menilai Toy Story dan Little Mermaid dan kemudian merekomendasikan produk manapun yang menurut prediksi model kami akan mereka nilai tertinggi. Matriks utilitas diatas adalah apa yang dikenal sebagai peringkat eksplisit. Setiap orang telah menilai film yang telah mereka lihat. Seringkali, kita harus menyimpulkan beberapa arti dari data dan menggunakan penilaian kita sendiri untuk menentukan bagaimana menggunakannya untuk sistem rekomendasi. Asumsikan bahwa alih-alih peringkat, kami hanya tahu apakah pengguna membeli film dari situs web streaming atau tidak. Mari kita lihat seperti apa tabel ini:

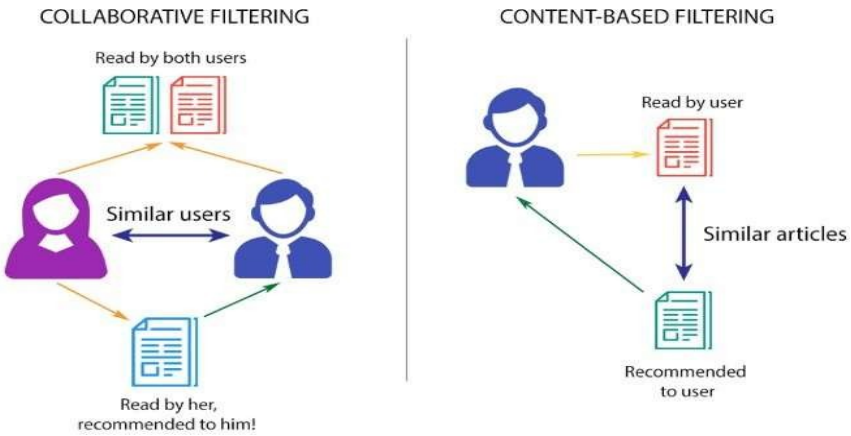
Tabel 3. Asumsi rekomendasi peringkat

	Cerita mainan	**Cinderella**	**Putri duyung kecil**	**Raja singa**
Matt		1		1
Pengetahuan	1		1	
mike		1	1	1
hutan	1		1	
Taylor	1	1		1

Ini adalah peringkat implisit karena kami berasumsi bahwa karena seseorang telah membeli sesuatu, mereka ingin membeli barang lain yang serupa. Tentu saja, ini belum tentu benar, tetapi lebih baik daripada tidak sama sekali!

Recommendation system adalah sebuah sistem yang mengacu pada memprediksi sejumlah item atau data untuk pengguna dimasa mendatang, kemudian dijadikan rekomendasi item paling teratas. Salah satu alasan mengapa perlu digunakannya recommendation system karena pengguna memiliki banyak pilihan untuk digunakan karena prevalensi internet.

Meskipun jumlah informasi yang tersedia meningkat, masalah baru muncul karena para pengguna kesulitan memilih item yang ingin mereka lihat. Di sinilah recommendation system masuk.



Gambar 89. Collaborative filtering and content-based filtering

Recommendation System terbagi atas 3 metode :

1. Content Based Recommendation

   Content Based Recommendation memanfaatkan informasi beberapa item / data untuk direkomendasikan kepada pengguna sebagai referensi yang terkait dengan informasi yang digunakan sebelumnya. Tujuan dari content based recommendation agar dapat memprediksi persamaan dari sejumlah informasi yang didapat dari pengguna.

   Contoh, seorang pengguna sedang menonton video di Youtube. Konten yang dilihat oleh pengguna, yaitu tentang sepak bola. Youtube secara sistem akan

merekomendasikan si pengguna untuk melihat video lain yang berhubungan dengan konten sepak bola.

Dalam pembuatannya, content based filtering menggunakan konsep perhitungan vector, TF-IDF, dan Cosine Similarity yang intinya dikonversikan dari data / text menjadi berbentuk vektor.

$$Cosine\ Similarity : Sim(u_i, u_k) = \frac{r_i \cdot r_k}{|r_i||r_k|} = \frac{\sum_{j=1}^{m} r_{ij} r_{kj}}{\sqrt{\sum_{j=1}^{m} r_{ij}^2 \sum_{j=1}^{m} r_{kj}^2}}$$

2. Collaborative Filtering

Collaborative Filtering memanfaatkan transaksi suatu produk / item yang didasarkan kepada perilaku / kebiasaan si pengguna. Tujuannya agar pengguna yang sama dan item yang serupa dapat disukai oleh pengguna sebagai rekomendasi pilihan. Collaborative Filtering umumnya terbagi atas 2 metode :

a. User-Based Collaborative Filtering (UB-CF)

User-Based Collaborative Filtering (UB-CF) merekomendasikan item suatu produk dengan menemukan pengguna yang mirip dengan pengguna yang lainnya. Contoh, kita ingin merekomendasikan film kepada teman kita. Kita bisa berasumsi bahwa orang yang sama akan memiliki selera yang sama. Misalkan saya dan si B telah menonton film yang sama, dan kami menilai semuanya hampir identik. Tapi si B belum melihat film ‘Harry Potter : Bagian II’ dan saya tahu. Jika saya suka film itu, itu masuk akal untuk berpikir bahwa dia juga.

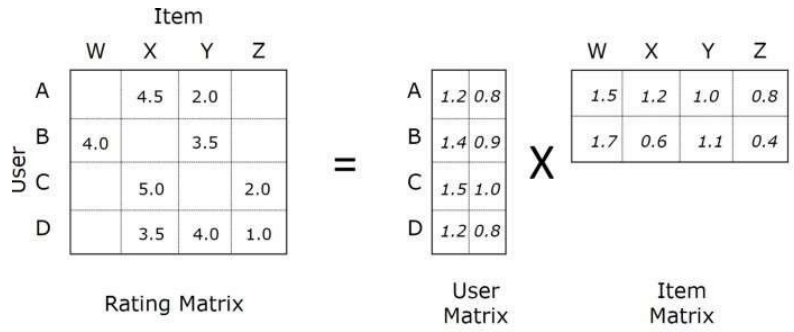
UB-CF mempunyai 2 metode perhitungan, yaitu Cosine similarity dan Pearson Correlation.

$$Pearson\ Correlation : Sim(u_i, u_k) = \frac{\sum_j (r_{ij} - r_i)(r_{kj} - r_k)}{\sqrt{\sum_j (r_{ij} - r_i)^2 \sum_j (r_{kj} - r_k)^2}}$$

Kedua metode tersebut biasa digunakan. Perbedaannya adalah bahwa Pearson Correlation tidak berubah untuk menambahkan konstanta pada semua elemen. Pada pembuatannya, algoritma yang umum digunakan K-Nearest Neighbor (KNN). Dan algoritma KNN bisa dilihat di Scikit-learn.

b. Item-Based Collaborative Filtering (IB-CF) merekomendasikan kesamaan antara item yang menargetkan / berinteraksi dengan pengguna dan item lainnya. Contoh, seorang pengguna menyukai tim sepak bola "Bayern Munich". Sistem secara otomatis akan merekomendasikan suatu item yang berhubungan dengan "Bayern Munich", seperti jersey, bola, poster, dan lain sebagainya yang berhubungan.

   Pada pembuatannya, IB-CF menggunakan metode perhitungan yang sama dengan UB-CF. Hanya saja, berbeda dalam menargetkan suatu rekomendasi yang dibuat. Selain peggunaan algoritma KNN, collaborative filtering juga bisa menggunakan algoritma Matrix Facrotization (MF) dengan mendekomposisi interaksi matriks item pengguna menjadi produk dari dua matriks dua dimensi yang lebih rendah. Satu matriks dapat dilihat sebagai matriks pengguna di mana baris mewakili pengguna dan kolom adalah latent factor. Matriks lain adalah matriks item di mana baris merupakan latent factor dan kolom mewakili item.



Gambar 90. Algoritma Matrix Factorization

Untuk menentukan latent factor, algoritma MF bisa menggunakan singular value decomposition (SVD). Pada python, MF bisa digunakan dengan scikit-learn yang sudah tersedia library nya.

3. Content—Collaborative Recommendation System (Hybrid Recommendation System)

   Hybrid Recommendation System adalah gabungan antara metode content based filtering dengan collaborative filtering untuk mendapatkan keuntungan dari keunggulan komplementernya.

   Salah satu penelitian dengan judul Hybrid Recommender Systems: A Systematic Literature Review oleh Erion Çano, Maurizio Morisio, membahas tentang Sistem rekomendasi hibrid menggabungkan dua atau lebih strategi rekomendasi dengan cara yang berbeda untuk mendapatkan keuntungan dari keunggulan komplementernya.

4. Forum Diskusi
   a. Bersama dengan tim diskusi anda, lakukanlah analisis pemasaran berdasarkan studi kasus(bisnis) yang anda fahami.
   b. Jawablah pertanyaan dapat dijawab oleh analisis pelanggan? (Pertanyaan terdapat pada naskah Bab 6)

C. PENUTUP

1. Rangkuman

Recommendation system dapat menjadi salah satu teknik dimana dapat menentukan rekomendasi kepada pengguna lewat kebiasaan atau interaksi dengan pengguna lainnya maupun dengan beberapa item yang terkait dengan item yang digunakan oleh pengguna. Selain penggunaan library scikit-learn, pembuatan recommendation system bisa juga menggunakan library lain, seperti : Fast.ai, Surprise, LightFM, Implicit, PySpark : mllib.recommendation

2. Tes Formatif

Sebagai evaluasi pemahaman materi pada Bab 6, jawablah pertanyaan berikut ini.

1. Jelaskan yang dimaksud dengan Marketing Analityc

2. Berikut ini yang merupakan beberapa model dan metode analitik yang populer adalah :

a. Media Mix Model (MMM

b. Multi-Touch Attribution (MTA)

c. Unified Marketing Measurement (UMM)

d. Multi Media Touch Model(MMTM)

3. Dengan menggunakan data dari berbagai sumber (online dan offline) untuk mencegah tampilan terfragmentasi, berikut ini yang merupakan beberapa hal wawasan yang dapat diperoleh adalah :

a. Intelijen Produk

b. Tren dan Preferensi Pelanggan

c. Tren Pengembangan Produk

d. Dukungan Pelanggan

4. Berikut ini yang merupakan tantangan analitik pemasaran terbesar yang dihadapi saat ini adalah :

a. Kuantitas Data

b. Kualitas Data

c. Kurangnya Data Sciantist

d. Memilih Model Atribusi

5. Dengan cara apa saja yang dilakukan software analitik pemasaran mengatasi tantangan?

a. mengumpulkan data

b. mengatur data yang diperlukan

c. menghubungkan data berharga dengan cepat

d. memungkinkan pemasar melakukan pengoptimalan kampanye secara real-time

6. Saat menerapkan solusi pengukuran pemasaran, yang menjadi pertimbangkan fitur dan kemampuan utama dari perangkat lunak analisis pemasaran adalah

a. Analisis dan Wawasan real-time

b. Kemampuan Pengukuran Merek

c. Granular, Data Tingkat Orang

d. Kemampuan untuk Menghubungkan Metrik Atribusi Online dan Offline

7. Saat tim pemasaran berusaha melakukan analisis kualitas yang mengarah pada kampanye yang lebih menarik dan menguntungkan, mereka harus fokus pada mempekerjakan manajer analitik yang dapat:

a. Melakukan Analisis Kualitas

b. Buat Rekomendasi Pengoptimalan

c. Memahami Tren Konsumen dan MarTech

d. Bekerja dengan Alat Analytics

8. Jika ingin meningkatkan kemampuan analitik, berikut ini yang merupakan empat langkah yang harus diambil diawal program adalah :

a. Pahami Apa yang Ingin Anda Ukur

b. Tetapkan Tolok Ukur

c. Nilai Kemampuan Anda Saat Ini

d. Terapkan Alat Analisis Pemasaran

9. "Perusahaan yang menggunakan analisis pelanggan secara komprehensif melaporkan melampaui pesaing mereka dalam hal keuntungan hampir dua kali

lebih sering daripada perusahaan yang tidak." Platform analisis konsumen yang baik dapat membantu peningkatan tim Anda adalah :

a. Mobile adoption

b. Retensi pelanggan

c. User engagement/Keterlibatan pengguna

d. In-app purchases

10. Pelanggan Mixpanel lainnya, STARZ PLAY, mampu mengelompokkan pelanggan mereka berdasarkan perilaku untuk mendeteksi penipuan dan menguranginya dengan faktor 1.000. Tim mana yang membutuhkan analisis pelanggan?

a. Pemasaran

b. Penjualan

c. Produk

d. Konsumen

11. Pertanyaan bagus bersifat konkret, mudah dibuktikan atau dibantah, dan terkait kembali dengan tujuan bisnis inti seperti akuisisi, pendapatan, retensi, dan keterlibatan. Contoh pertanyaan umum tentang analisis pelanggan dalam hal akuisi adalah:

a. Saluran mana yang mendorong pelanggan baru paling banyak?

b. Apa perjalanan pelanggan yang paling umum dari kesadaran ke advokasi?

c. Apa saluran pendapatan kami yang paling menguntungkan?

d. Pengguna mana yang paling menguntungkan?

12. Sebuah sistem yang mengacu pada memprediksi sejumlah item atau data untuk pengguna dimasa mendatang, kemudian dijadikan rekomendasi item paling teratas. Hal tersebut disebut dengan apa?

13. Berikut ini yang merupakan metode Recommendation System, yaitu:

a. Content Based Recommendation

b. Collaborative Filtering

c. Content—Collaborative Recommendation System (Hybrid Recommendation System)

d. User-Based Collaborative Filtering (UB-CF)

Daftar Pustaka

Recommendation System Dengan Python: Definisi (Part1), https://medium.com/data-folks-indonesia/recommendation-system-dengan-python-definisi-part-1-71154dc3f700

What is customer analytics?, The Signal by Mixpanel, https://mixpanel.com/blog/what-is-customer-analytics/

What is Marketing Analytics? Definition, Tips, and Tools, https://www.marketingevolution.com/marketing-essentials/marketing-analytics

BAB VII

PREDICTIVE ANALYTICS FOR BUSINESS

A. PENDAHULUAN

Mampu memahami dan menjelaskan bagaimana memprediksi analisa hasil dimasa depan berdasarkan data historis dan teknik analitik seperti pemodelan statistik dan machine learning.

B. INTI

1. Capaian Pembelajaran
    - **a.** Memahami dan menjelaskan bagaimana menganalisa prediksi pada sebuah bisnis dan perbedaan Data Collection dengan Analysis Data
    - **b.** Mampu memahami dan menjelaskan apa itu predictive analytics dan bagaimana di tempat kerja
    - **c.** Mampu memahami dan menjelaskan keuntungan predictive analytic, contoh predictive analytics, tools yang digunakan pada analisis prediksi
    - **d.** Mampu memahami dan menjelaskan predictive analytic models, predictive modelling techniques, predictive analytics algorithms, Analisis prediktif dalam perawatan kesehatan, Bagaimana seharusnya sebuah organisasi memulai dengan analitik prediktif
    - **e.** Mampu memahami dan menjelaskan Predictive Analytics
    - **f.** Mampu memecahkan masalah dengan Analisis Prediktif untuk Bisnis, bagaiamana melihat data dalam konteks untuk analisis prediktif untuk bisnis,
    - **g.** Mampu memahami dan menjelaskan perbedaan data collection dan analysis data
    - **h.** Mampu memahami dan menyelesaikan masalah yang dapat diselesaikan dengan data science
    - **i.** Mampu memahami dan menjelaskan data science dan Teknik machine learning
2. Materi

a. Predictive Analytics

b. Keuntungan predictive analytic

c. Predictive Analytic Models

d. Pentingnya Predictive Analytic

e. Memecahkan Masalah dengan Analisis Prediktif untuk Bisnis

f. Perbedaan Data Collection dan Analysis Data

g. Masalah yang dapat diselesaikan dengan data science

h. Data science dan Teknik machine learning

3. Uraian Materi

a. Predictive Analytics

Predictive Analytics(Analisis Prediksi) adalah kategori analitik data yang bertujuan membuat prediksi tentang hasil masa depan berdasarkan data historis dan teknik analitik seperti pemodelan statistik dan machine learning. Ilmu Predictive Analytics dapat menghasilkan wawasan masa depan dengan tingkat presisi yang signifikan. Dengan bantuan alat dan model analitik prediktif yang canggih, organisasi mana pun kini dapat menggunakan data masa lalu dan saat ini untuk memperkirakan tren dan perilaku dalam hitungan milidetik, hari, atau tahun ke depan dengan andal.

Analisis prediktif telah mendapatkan dukungan dari berbagai organisasi, dengan pasar global diproyeksikan mencapai sekitar $10,95 miliar pada tahun 2022, tumbuh pada tingkat pertumbuhan tahunan gabungan (CAGR) sekitar 21 persen antara tahun 2016 dan 2022, menurut laporan tahun 2017 yang dikeluarkan oleh Riset Pasar Zion.

Analisis prediktif di tempat kerja

Analitik prediktif memanfaatkan kekuatannya dari berbagai metode dan teknologi, termasuk big data, data mining, pemodelan statistik, machine learning, dan berbagai macam proses matematika. Organisasi menggunakan analitik prediktif untuk menyaring data terkini dan historis untuk mendeteksi tren dan memperkirakan

peristiwa dan kondisi yang akan terjadi pada waktu tertentu, berdasarkan parameter yang disediakan.

Dengan analitik prediktif, organisasi dapat menemukan dan memanfaatkan pola yang terkandung dalam data untuk mendeteksi risiko dan peluang. Model dapat dirancang, misalnya, untuk menemukan hubungan antara berbagai faktor perilaku. Model tersebut memungkinkan penilaian baik janji atau risiko yang disajikan oleh serangkaian kondisi tertentu, memandu pengambilan keputusan berdasarkan informasi di berbagai kategori rantai pasokan dan peristiwa pengadaan.

b. Keuntungan Predictive Analityc

Analitik prediktif membuat melihat ke masa depan lebih akurat dan andal daripada alat sebelumnya. Dengan demikian dapat membantu pengadopsi menemukan cara untuk menyimpan dan mendapatkan uang. Pengecer sering menggunakan model prediktif untuk memperkirakan kebutuhan inventaris, mengelola jadwal pengiriman, dan mengonfigurasi tata letak toko untuk memaksimalkan penjualan. Maskapai penerbangan sering menggunakan analitik prediktif untuk menetapkan harga tiket yang mencerminkan tren perjalanan masa lalu. Hotel, restoran, dan pelaku industri perhotelan lainnya dapat menggunakan teknologi untuk memperkirakan jumlah tamu pada malam tertentu untuk memaksimalkan hunian dan pendapatan.

Dengan mengoptimalkan kampanye pemasaran dengan analitik prediktif, organisasi juga dapat menghasilkan tanggapan atau pembelian pelanggan baru, serta mempromosikan peluang penjualan silang. Model prediktif dapat membantu bisnis menarik, mempertahankan, dan memelihara pelanggan mereka yang paling berharga.

Analitik prediktif juga dapat digunakan untuk mendeteksi dan menghentikan berbagai jenis perilaku kriminal sebelum kerusakan serius terjadi. Dengan menggunakan analitik prediktif untuk mempelajari perilaku dan tindakan pengguna, organisasi dapat mendeteksi aktivitas yang tidak biasa, mulai dari penipuan kartu kredit hingga mata-mata perusahaan hingga cyber crime.

Contoh Predictive Analytic

Organisasi saat ini menggunakan analitik prediktif dalam berbagai cara yang hampir tak ada habisnya. Teknologi ini membantu pengadopsi di berbagai bidang seperti keuangan, perawatan kesehatan, ritel, perhotelan, farmasi, otomotif, kedirgantaraan, dan manufaktur.

Berikut ini adalah beberapa contoh bagaimana organisasi memanfaatkan analitik prediktif:

1. Dirgantara: Memprediksi dampak operasi pemeliharaan khusus pada keandalan pesawat, penggunaan bahan bakar, ketersediaan, dan waktu kerja.
2. Otomotif: Memasukkan catatan kekokohan dan kegagalan komponen ke dalam rencana manufaktur kendaraan yang akan datang. Pelajari perilaku pengemudi untuk mengembangkan teknologi bantuan pengemudi yang lebih baik dan, pada akhirnya, kendaraan otonom.
3. Energi: Perkiraan harga jangka panjang dan rasio permintaan. Menentukan dampak peristiwa cuaca, kegagalan peralatan, peraturan dan variabel lain pada biaya layanan.
4. Layanan keuangan: Mengembangkan model risiko kredit. Prakiraan tren pasar keuangan. Memprediksi dampak kebijakan, undang-undang, dan peraturan baru terhadap bisnis dan pasar.
5. Manufaktur: Memprediksi lokasi dan tingkat kegagalan mesin. Mengoptimalkan pengiriman bahan baku berdasarkan proyeksi permintaan di masa mendatang.
6. Penegakan hukum: Gunakan data tren kejahatan untuk menentukan lingkungan yang mungkin memerlukan perlindungan tambahan pada waktu-waktu tertentu dalam setahun.
7. Ritel: Ikuti pelanggan online secara real-time untuk menentukan apakah memberikan informasi produk tambahan atau insentif akan meningkatkan kemungkinan penyelesaian transaksi.

Predictive Analytic Tools

Alat analitik prediktif memberi user wawasan yang mendalam dan real-time kedalam rangkaian aktivitas bisnis yang hampir tak ada habisnya. Alat dapat digunakan untuk memprediksi berbagai jenis perilaku dan pola, seperti bagaimana mengalokasikan sumber daya pada waktu tertentu, kapan harus mengisi kembali stok atau saat terbaik untuk meluncurkan kampanye pemasaran, mendasarkan prediksi pada analisis data yang dikumpulkan selama periode waktu tertentu. .

Hampir semua pengadopsi analitik prediktif menggunakan alat yang disediakan oleh satu atau lebih pengembang eksternal. Banyak alat tersebut disesuaikan untuk memenuhi kebutuhan perusahaan dan departemen tertentu. Perangkat lunak analitik prediktif dan penyedia layanan utama meliputi: Aksioma, IBM, Pembuat Informasi, Microsoft, GETAH, Institut SAS, Perangkat Lunak Tableau, Teradata, Perangkat Lunak TIBCO.

c. Predictive analytics models

Model adalah dasar dari analitik prediktif — template yang memungkinkan pengguna mengubah data masa lalu dan saat ini menjadi wawasan yang dapat ditindaklanjuti, menciptakan hasil jangka panjang yang positif. Beberapa tipe tipikal model prediktif meliputi:

1. Model Nilai Seumur Hidup Pelanggan: Tentukan pelanggan yang kemungkinan besar akan berinvestasi lebih banyak dalam produk dan layanan.
2. Model Segmentasi Pelanggan: Mengelompokkan pelanggan berdasarkan karakteristik dan perilaku pembelian yang serupa
3. Model Pemeliharaan Prediktif: Memperkirakan kemungkinan kerusakan peralatan penting.
4. Model Jaminan Kualitas: Temukan dan cegah cacat untuk menghindari kekecewaan dan biaya tambahan saat menyediakan produk atau layanan kepada pelanggan.

Predivtive Modeling Techniques

Pengguna model memiliki akses ke berbagai teknik pemodelan prediktif yang hampir tak ada habisnya. Banyak metode yang unik untuk produk dan layanan tertentu, tetapi inti dari teknik umum, seperti pohon keputusan, regresi — dan bahkan jaringan saraf — kini didukung secara luas di berbagai platform analitik prediktif.

Pohon keputusan, salah satu teknik yang paling populer, bergantung pada skema, diagram berbentuk pohon yang digunakan untuk menentukan tindakan atau untuk menunjukkan probabilitas statistik. Metode percabangan juga dapat menunjukkan setiap hasil yang mungkin dari keputusan tertentu dan bagaimana satu pilihan dapat mengarah ke pilihan berikutnya. Teknik regresi sering digunakan dalam perbankan, investasi dan model berorientasi keuangan lainnya. Regresi membantu pengguna memperkirakan nilai aset dan memahami hubungan antar variabel, seperti komoditas dan harga saham. Di ujung tombak teknik analitik prediktif adalah jaringan saraf — algoritme yang dirancang untuk mengidentifikasi hubungan mendasar dalam kumpulan data dengan meniru cara pikiran manusia berfungsi.

Predictive analytics algorithms

Pengadopsi analitik prediktif memiliki akses mudah ke berbagai algoritma statistik, penambangan data, dan pembelajaran mesin yang dirancang untuk digunakan dalam model analisis prediktif. Algoritma umumnya dirancang untuk memecahkan masalah bisnis tertentu atau serangkaian masalah, meningkatkan algoritma yang ada atau menyediakan beberapa jenis kemampuan unik.

Algoritme pengelompokan, misalnya, sangat cocok untuk segmentasi pelanggan, deteksi komunitas, dan tugas terkait sosial lainnya. Untuk meningkatkan retensi pelanggan, atau untuk mengembangkan sistem rekomendasi, algoritma klasifikasi biasanya digunakan. Algoritma regresi biasanya dipilih untuk membuat sistem penilaian kredit atau untuk memprediksi hasil dari banyak peristiwa yang didorong oleh waktu.

Analisis prediktif dalam perawatan kesehatan

Organisasi perawatan kesehatan telah menjadi beberapa pengadopsi analitik prediktif yang paling antusias karena alasan yang sangat sederhana: Teknologi ini membantu mereka menghemat uang. Organisasi perawatan kesehatan menggunakan analitik prediktif dalam beberapa cara berbeda, termasuk mengalokasikan sumber daya fasilitas secara cerdas berdasarkan tren masa lalu, mengoptimalkan jadwal staf, mengidentifikasi pasien yang berisiko untuk rawat inap kembali jangka pendek yang mahal dan menambahkan kecerdasan untuk akuisisi dan manajemen farmasi dan pasokan.

Laporan Society of Actuaries tahun 2017 tentang tren industri perawatan kesehatan dalam analitik prediktif, menemukan bahwa lebih dari setengah eksekutif perawatan kesehatan (57 persen) di organisasi yang sudah menggunakan analitik prediktif percaya bahwa teknologi akan memungkinkan mereka untuk menghemat 15 persen atau lebih dari total anggaran mereka selama lima tahun ke depan. Tambahan 26 persen diprediksi penghematan 25 persen atau lebih.

Studi ini juga mengungkapkan bahwa sebagian besar eksekutif layanan kesehatan (89 persen) berasal dari organisasi yang sekarang menggunakan analitik prediktif atau berencana untuk melakukannya dalam lima tahun ke depan. 93 persen eksekutif layanan kesehatan yang mengesankan menyatakan bahwa analitik prediktif penting untuk masa depan bisnis mereka.

Bagaimana seharusnya sebuah organisasi memulai dengan analitik prediktif?

Meskipun memulai analitik prediktif bukanlah hal yang mudah, ini adalah tugas yang dapat ditangani oleh hampir semua bisnis selama seseorang tetap berkomitmen pada pendekatan tersebut dan bersedia menginvestasikan waktu dan dana yang diperlukan untuk menjalankan proyek. Dimulai dengan proyek percontohan skala terbatas di area bisnis penting adalah cara terbaik untuk membatasi biaya awal sambil meminimalkan waktu sebelum imbalan finansial mulai bergulir. Setelah model diterapkan, umumnya memerlukan sedikit pemeliharaan saat berlanjut. untuk menggiling wawasan yang dapat ditindaklanjuti selama bertahun-tahun.

d. Pentingnya Predictive Analytic

Organisasi beralih ke analitik prediktif untuk membantu memecahkan masalah sulit dan mengungkap peluang baru. Penggunaan umum meliputi:

1. Mendeteksi penipuan. Menggabungkan beberapa metode analitik dapat meningkatkan deteksi pola dan mencegah perilaku kriminal. Ketika keamanan siber menjadi perhatian yang berkembang, analitik perilaku berkinerja tinggi memeriksa semua tindakan di jaringan secara real time untuk menemukan kelainan yang mungkin mengindikasikan penipuan, kerentanan zero-day, dan ancaman persisten tingkat lanjut.
2. Mengoptimalkan kampanye pemasaran. Analisis prediktif digunakan untuk menentukan tanggapan atau pembelian pelanggan, serta mempromosikan peluang penjualan silang. Model prediktif membantu bisnis menarik, mempertahankan, dan menumbuhkan pelanggan mereka yang paling menguntungkan.
3. Meningkatkan operasi. Banyak perusahaan menggunakan model prediktif untuk memperkirakan persediaan dan mengelola sumber daya. Maskapai menggunakan analitik prediktif untuk menetapkan harga tiket. Hotel mencoba memprediksi jumlah tamu untuk setiap malam tertentu untuk memaksimalkan hunian dan meningkatkan pendapatan. Analisis prediktif memungkinkan organisasi berfungsi lebih efisien.
4. Mengurangi risiko. Skor kredit digunakan untuk menilai kemungkinan default pembeli untuk pembelian dan merupakan contoh analitik prediktif yang terkenal. Skor kredit adalah angka yang dihasilkan oleh model prediktif yang menggabungkan semua data yang relevan dengan kelayakan kredit seseorang. Penggunaan terkait risiko lainnya termasuk klaim dan penagihan asuransi.

e. Memecahkan Masalah dengan Analisis Prediktif untuk Bisnis

Analisis prediktif untuk bisnis berlaku untuk berbagai masalah perusahaan yang dihadapi saat ini, dan semakin banyak orang mulai mengenali nilainya. Perusahaan

menggunakan analitik prediktif untuk bisnis untuk menjawab pertanyaan bisnis nyata seperti "Segmen pelanggan potensial mana yang paling merespons pesan kami" dan "Mengapa saya kehilangan pelanggan, dan bagaimana cara menghentikan mereka untuk pergi?".

Meskipun penggunaan analitik prediktif untuk bisnis memiliki begitu banyak nilai bagi perusahaan, masalah umum dengan menerapkannya adalah bahwa pengetahuan tentang bagaimana melakukan analitik prediktif untuk bisnis tidak tersedia. Banyak orang kesulitan ketika mencoba memahami praktik analisis yang baik, memilih model prediksi yang tepat untuk situasi tertentu, dan memahami statistik yang mendasarinya.

Melihat Data dalam Konteks untuk Analisis Prediktif untuk Bisnis

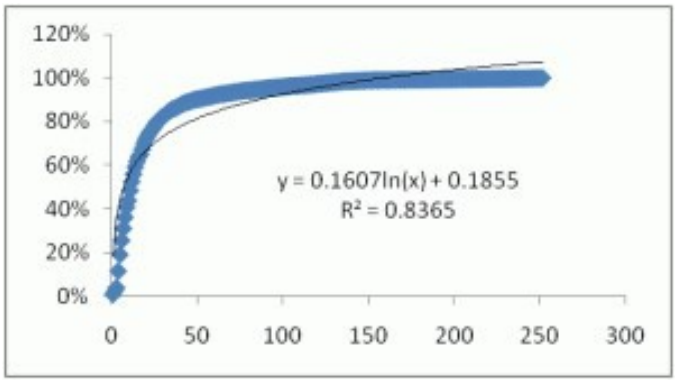
Untuk memulai dengan analitik prediktif untuk bisnis, memahami apa yang dikatakan data kepada Anda dalam konteks situasi bisnis yang sedang dianalisis sangatlah penting. Ini akan membantu Anda menghindari membuat kesimpulan yang salah dan menjaga analisis Anda sesuai dengan pertanyaan bisnis yang dijawab. Cara terbaik untuk mempelajari dasar ini adalah dengan melihatnya dalam tindakan, jadi kita akan mengambil contoh.

Dalam analitik prediktif untuk model bisnis ini, diketahui berapa banyak panggilan yang diharapkan masuk ke pusat panggilan setelah menjalankan kampanye. Pertama, mengambil beberapa data historis yang menunjukkan persentase total panggilan yang masuk sesuai dengan jumlah hari setelah memulai kampanye email, yang ditunjukkan dibawah ini.

Days After Mailing	% Total Incoming Calls
1	1%
2	3%
3	4%
4	12%
5	19%
6	26%
7	31%
8	36%
9	41%

Gambar 91. Persentasi total panggilan masuk

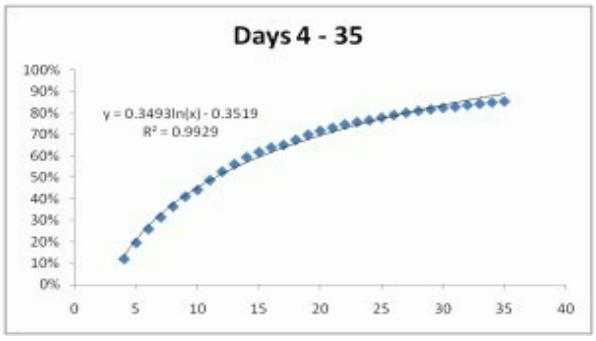
Setelah membuat plot pencar dari data, dibentuk garis regresi logaritmik sebagai model, yang ditunjukkan pada gambar 91.



Gambar 92. Garis regresi logaritmik

Meskipun R2 memberi tahu kita bahwa kecocokannya baik, model mungkin bukan cara terbaik untuk menjelaskan data ini ketika konteks dan tujuan analisis ini dipertimbangkan. Kami ingin analitik prediktif kami untuk model bisnis dapat memprediksi berapa persentase total panggilan yang akan masuk dari kampanye pengiriman surat sehingga kami dapat menjadi staf pusat panggilan. Jika saya menggunakan saluran di atas sebagai model, saya akan memprediksi nilai rendah untuk panggilan masuk antara sekitar hari 20 dan 100, dan nilai tinggi setelahnya. Karena kesalahan ini, kami tidak akan mengatur staf call center dengan benar.

Untuk membuat analitik prediktif yang lebih baik untuk model bisnis, akan dipertimbangkan bahwa fakta dalam konteks ini, tidak perlu menyesuaikan model tren ke seluruh kumpulan data. Pertimbangkan model berikut, yang dapat digunakan untuk memprediksi persentase total panggilan yang masuk antara hari ke 4 dan 35 setelah kampanye pengiriman surat:



Gambar 93. Persentase total panggilan yang masuk 4 dan 35

Anda akan melihat bahwa analitik prediktif untuk model tren bisnis ini tidak mengandung kesalahan tinggi dan rendah yang sama seperti model sebelumnya. Selanjutnya, setelah melakukan beberapa perhitungan pada data dalam spreadsheet, kita tahu bahwa apa pun sebelum hari ke-4 hanya menghasilkan 8% dari semua panggilan, dan apa pun setelah hari ke-35 hanya menghasilkan 15% dari semua panggilan. Dengan model periode waktu pertumbuhan terbesar untuk persentase panggilan, sambil meringkas persentase yang tersisa di kedua sisi. Ini akan memberikan jumlah informasi yang tepat yang diperlukan untuk staf pusat panggilan, sambil meminimalkan kesalahan saat mencoba menyesuaikan model tren tunggal dengan data.

f. Perbedaan Data Collection dan Analysis Data

Data adalah kebutuhan saat ini, dan pengumpulan serta analisisnya adalah dasar dari kesuksesan bisnis dan penelitian apa pun sekarang. Napoleon Bonaparte lebih dari 200 tahun yang lalu mengatakan tentang pentingnya data atau informasi dengan kutipan terkenalnya “Perang adalah 90% informasi”. . Oleh karena itu pengumpulan data dan analisis data akan menjadi kunci sukses di banyak bidang.

Data Collection adalah proses mengumpulkan dan mengukur data pada variabel yang ditargetkan melalui sistem yang ditetapkan secara menyeluruh untuk mengevaluasi hasil dengan menjawab pertanyaan yang relevan. **Analisis Data** adalah proses yang melibatkan data yang dibentuk untuk diperiksa untuk interpretasi untuk menemukan informasi yang relevan, mengusulkan kesimpulan, dan membantu dalam pengambilan keputusan masalah penelitian.

Data Collection adalah pengumpulan informasi dari berbagai sumber, dan analisis data adalah memprosesnya untuk mendapatkan wawasan yang berguna darinya. Perbedaan diantara keduanya selain dari fungsi utama adalah dalam cara melakukan kegiatan yang saling terkait. Untuk data yang dikumpulkan dari berbagai sumber dan metode, diperlukan metode dan alat analisis data khusus untuk memproses dan mendapatkan wawasan darinya.

Beberapa sumber pengumpulan data yang berbeda:

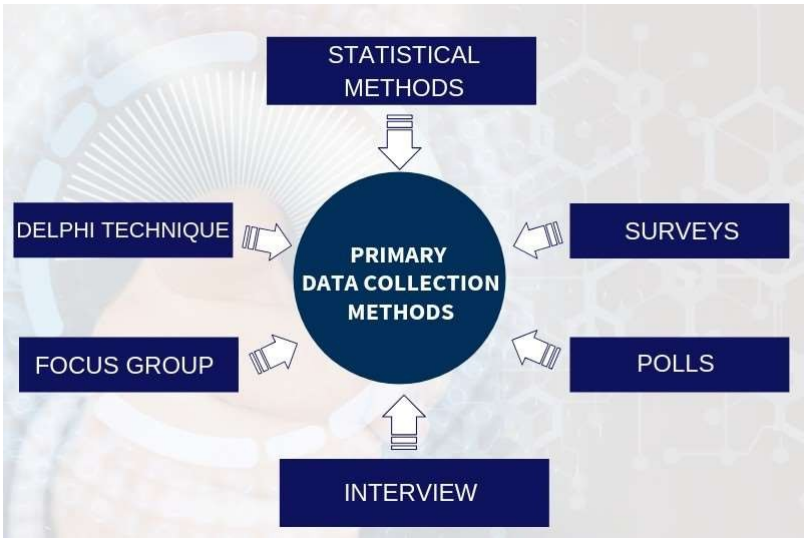
1. Mengumpulkan data baru dari internet dan sumber lainnya
2. Menggunakan data yang dikumpulkan dan disimpan sebelumnya
3. Menggunakan kembali data orang lain
4. Data pembelian

Metode pengumpulan data bergantung pada hal-hal berikut:

1. Masalah penelitian yang diteliti
2. Desain penelitian
3. Informasi yang dikumpulkan tentang variabel

Ada dua metode pengumpulan data:

1. Metode pengumpulan data primer



Gambar 94. Metode pengumpulan data primer

Data yang dikumpulkan pertama kali oleh peneliti, yang asli adalah data primer. Ini dikumpulkan untuk penelitian masalah yang unik dimanapun dan siapa pun tidak melakukan pemeriksaan terkait lainnya. Hasil penelitian tersebut justru menggunakan data primer yang dikumpulkan oleh peneliti. Tapi itu membutuhkan waktu dan biaya

2. Metode pengumpulan data sekunder

Gambar 95. Metode pengumpulan data sekunder

merupakan data yang disediakan oleh orang lain untuk pekerjaan penelitian. Dan pasti sudah melewati analisis statistik.

Perbedaan antara data primer dan sekunder adalah bahwa hasil sekunder tidak akan begitu otentik seperti primer dan secara komparatif lebih murah dan tersedia daripada data primer.

Berbagai jenis manajemen data sesuai dengan metode pengumpulannya:

Sesuai dengan metode pengumpulan, pengelolaan data akan memerlukan metode pengelolaan yang berbeda, yang meliputi:

1. Quantitative information:

Gambar 96. Quantitative Data Collection Tools

Merupakan informasi numerik yang diperoleh dari metode penelitian seperti survei populasi dan prosedur eksperimental berulang. Selama perekaman mereka, penting untuk memasukkan informasi rinci seperti tanggal, tempat pengumpulan, unit pengukuran, dan metodenya.

2. Qualitattive Information

Gambar 97. Qualitative Data Collection Tools

Informasi non numerik yang dikumpulkan dalam bentuk video dan audio adalah informasi kualitatif. Itu bisa ditranskripsikan dalam bentuk tertulis nanti.

Mengerjakan data Anda:

Data yang dikumpulkan peneliti seringkali dalam bentuk besar, yang perlu diringkas sebelum mendapatkan kesimpulan darinya. Ini mungkin dalam bentuk spreadsheet data numerik, transkrip atau wawancara atau deskripsi.

Menyajikan data Anda:

Saat menulis paper, penting untuk menggambarkan data dalam bentuk tabel, gambar, untuk menunjukkan poin secara mendalam.

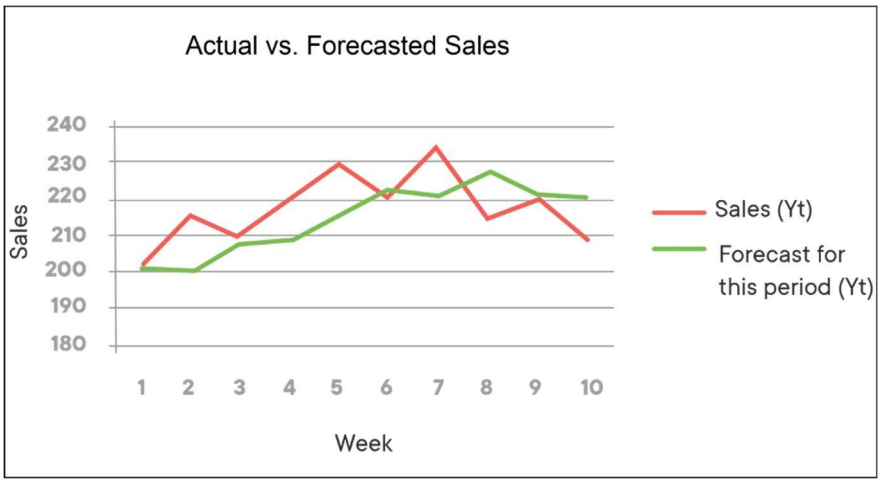
g. Masalah yang dapat diselesaikan dengan data science

Selamat atas keputusannya untuk menjadi ilmuwan data! Sebelum kita menggali detail alat dan teknik yang perlu Anda pelajari, penting untuk meluangkan sedikit waktu untuk memahami apa yang dapat Anda lakukan setelah lulus. Berikut adalah daftar beberapa jenis masalah bisnis umum yang diharapkan dapat dipecahkan oleh para ilmuwan data.

1. Regresi: Berapa banyak atau berapa banyak?

   Analisis regresi digunakan untuk memprediksi nilai berkelanjutan - seperti jumlah staf yang Anda perlukan untuk shift sibuk atau kemungkinan harga jual rumah.

   Contoh: Prakiraan Penjualan atau Pasar



Gambar 98. Penjualan dan ramalan penjualan

Analisis tren tradisional hanya melihat bagaimana satu entitas bisnis berubah sehubungan dengan yang lain. Analisis regresi dapat memberikan wawasan tentang bagaimana suatu hasil akan berubah ketika beberapa variabel lain dimodifikasi.

2. Klasifikasi: Kategori yang mana?

   Klasifikasi digunakan untuk memprediksi ke dalam kategori mana sesuatu akan masuk. Jika Anda mencoba mencari tahu apakah klien cenderung gagal membayar pinjaman (yaitu, gagal bayar atau tidak gagal bayar) atau produk mana yang mungkin disukai pelanggan, Anda menghadapi masalah klasifikasi.

   Contoh: Peringkat Kredit

Tabel 4. Peringkat Kredit

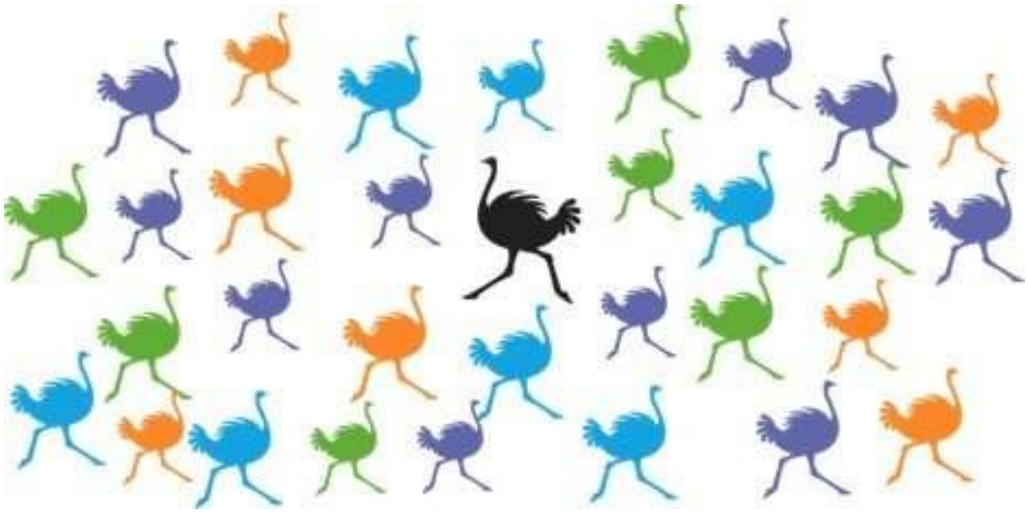
Moody's Investors Service	FitchRatings	S&P Global Ratings	Rating Grade Description
Aaa	AAA	AAA	Highest credit quality, lowest level of credit risk
Aa1 Aa2 Aa3	AA+ AA AA-	AA+ AA AA-	Very high credit quality with very low credit risk
A1 A2 A3	A+ A A-	A+ A A-	High credit quality with low credit risk
Baa1 Baa2 Baa3	BBB+ BBB BBB-	BBB+ BBB BBB-	Good credit quality with moderate credit risk
Ba1 Ba2 Ba3	BB+ BB BB-	BB+ BB BB-	Speculative with substantial credit risk
B1 B2 B3	B+ B B-	B+ B B-	Highly speculative with high credit risk
Caa1 Caa2 Caa3	CCC+ CCC CCC-	CCC+ CCC CCC-	Substantial credit risk with default as a real possibility
Ca	CC	CC	Very high levels of credit risk with default either occurring or about to occur
C	C	C	Default or default-like process has begun
	SD	RD	Selective Default (SD): Issuers have defaulted on one or more specific issues but are expected to meet their other payment obligations. Restricted Default (RD): Issuers have missed one or more payments but are not under supervision for reorganization or liquidation.
	D	D	Default: Issuers are unlikely to pay their obligations and have likely entered into bankruptcy filings, administration, receivership, liquidation or other formal winding-up procedures.

Perusahaan kartu kredit menerima ratusan ribu aplikasi untuk kartu kredit baru setiap minggu. Aplikasi ini berisi informasi rinci tentang atribut sosial, ekonomi, dan pribadi pelamar. Analisis klasifikasi dapat memungkinkan perusahaan untuk mengkategorikan aplikasi ini berdasarkan kualitas kredit mereka.

3. Deteksi anomali: Apakah ini aneh?

Deteksi anomali adalah teknik ilmu data umum yang digunakan untuk menemukan pola yang tidak biasa yang tidak sesuai dengan perilaku yang diharapkan. Ini memiliki aplikasi diberbagai industri mulai dari deteksi intrusi (mengidentifikasi pola aneh dalam lalu lintas jaringan yang dapat menandakan peretasan) hingga deteksi penipuan dalam transaksi kartu kredit hingga deteksi kesalahan di lingkungan operasi.

Contoh: Mengidentifikasi Penipuan



Gambar 99. Mengidentifikasi penipuan

Pendekatan ini berfokus pada pencarian outlier dalam data yang tampak memiliki pola yang tidak biasa. Ini berfungsi sebagai indikasi pertama adanya aktivitas penipuan. Pendekatan semacam itu juga sering diterapkan oleh jejaring sosial besar seperti Facebook, Twitter, dll.

4. Sistem pemberi rekomendasi: Item mana yang lebih disukai pengguna?

   Sistem rekomendasi adalah salah satu aplikasi ilmu data yang paling populer saat ini. Mereka digunakan untuk memprediksi preferensi pengguna terhadap suatu produk/layanan. Hampir setiap perusahaan teknologi besar (Amazon, Netflix, Google, Facebook) telah menerapkannya dalam beberapa bentuk. Anda mungkin telah memperhatikan frasa seperti "Jika Anda menyukai produk ini, Anda mungkin juga menyukai ...", "Pengguna yang membeli item ini juga membeli ...", dan "Berdasarkan preferensi Anda, kami merekomendasikan produk berikut untuk Anda .. .". Anda mengerti, ini semua adalah sistem rekomendasi yang sedang beraksi.

   Sistem pemberi rekomendasi dapat membantu bisnis mempertahankan pelanggan dengan memberikan saran yang disesuaikan dengan kebutuhan mereka. Mereka dapat membantu meningkatkan penjualan dan menciptakan loyalitas merek melalui personalisasi yang relevan. Ketika pelanggan merasa seolah-olah mereka dipahami

oleh merek Anda, mereka cenderung tetap setia dan terus membeli melalui situs Anda. Menurut sebuah studi baru-baru ini oleh McKinsey, hingga 75% dari apa yang ditonton konsumen di Netflix berasal dari sistem rekomendasi perusahaan. Raksasa ritel Amazon memuji sistem pemberi rekomendasi dengan 35% dari pendapatan mereka. Best Buy memutuskan untuk fokus pada penjualan online mereka, dan pada kuartal kedua 2016 mereka melaporkan peningkatan 23,7%, sebagian berkat sistem rekomendasi mereka.

h. Data science dan Teknik machine learning

Keterampilan Data Science dan Pasar Kerja

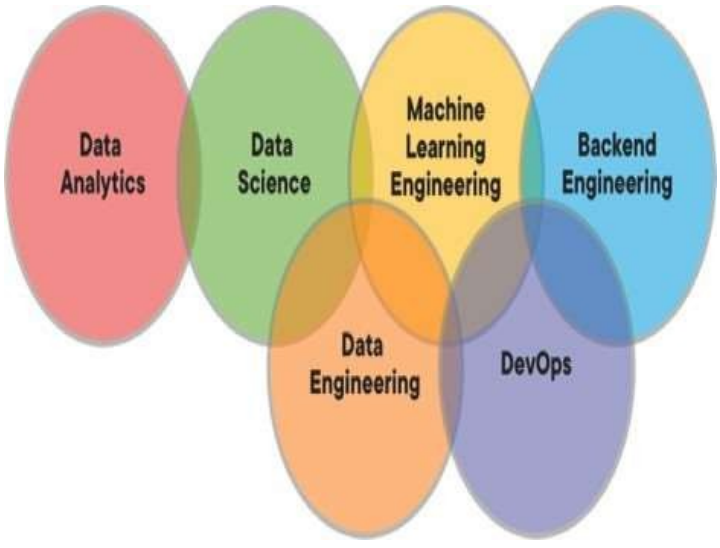
Saat Anda mulai melihat lowongan pekerjaan, Anda mungkin memperhatikan bahwa pekerjaan "Data Scientist" bisa berarti banyak hal yang berbeda. Di beberapa perusahaan, ini berarti seorang analis data atau DBA yang berfokus pada database atau saluran data. Pada orang lain, itu berarti seseorang dengan pola pikir ilmiah yang terampil menjalankan tes A/B. Namun yang lain mungkin merupakan peran machine learning yang sangat terspesialisasi di bidang-bidang seperti NLP, computer vision, atau deep learning -- dan peran ini dapat dipecah lebih lanjut menjadi spesialisasi yang berfokus pada penelitian atau implementasi. Sebagai Junior Data Scientist yang memasuki dunia kerja, kemungkinan besar Anda akan mendapatkan peran generalis, menghabiskan beberapa tahun pertama karir Anda mengerjakan berbagai tugas yang berfokus pada semua area ini setidaknya sedikit. Spesialisasi terjadi kemudian dalam karir Anda. Keluar dari gerbang,

Selama studi Anda, Anda telah mengambil banyak keterampilan berbeda di banyak area domain berbeda yang memungkinkan Anda berkontribusi pada proyek ilmu data di lingkungan profesional. Namun, ketika datang ke pasar kerja, tidak semua keterampilan Ilmu Data diciptakan sama. Sementara ilmuwan data atau perekrut yang berbeda dapat memberi peringkat keterampilan ini secara berbeda berdasarkan pengalaman atau kebutuhan mereka sendiri, satu hal yang paling disepakati adalah bahwa kemampuan untuk menghasilkan model sangat berharga dan sangat langka jika menyangkut Ilmuwan Data Junior. Ini memberikan peluang besar bagi Anda -- jika

Anda dapat menjadi mahir dalam memproduksi model yang telah Anda buat (dan menunjukkan kemahiran ini dalam portofolio proyek Anda!), Anda menjadi kandidat yang jauh lebih menarik untuk peran apapun.

Memproduksi Model sebagai Keterampilan Karir

Di perusahaan besar seperti Google dan Facebook, Ilmuwan Data biasanya menjalankan eksperimen dan melatih model sampai mereka menemukan solusi yang berhasil. Setelah mereka melatih model yang divalidasi, mereka biasanya kemudian menyerahkan produksi model kepada Insinyur Pembelajaran Mesin. Sedangkan Data Scientist menciptakan prototipe dasar, tugas Machine Learning Engineer adalah menempatkan model itu ke dalam sistem produksi dengan cara yang berkinerja dan dapat dipelihara. Sementara Ilmuwan Data di perusahaan besar fokus pada "gambaran besar" dengan menemukan solusi untuk masalah bisnis, Insinyur Pembelajaran Mesin fokus pada detail, menerapkan solusi yang dibuat oleh para ilmuwan data dengan cara terbaik. Ilmuwan Data lebih fokus pada analitik dan statistik, sedangkan Insinyur Pembelajaran Mesin akan memiliki komando yang lebih kuat dalam rekayasa backend, struktur dan algoritme data, dan rekayasa perangkat lunak secara keseluruhan. Diagram berikut menjabarkan hubungan antara peran teknis yang berbeda dengan baik:



Gambar 100. hubungan antara peran teknis yang berbeda dengan baik

Seperti yang dapat Anda lihat dari tumpang tindih antara Data Scientist dan Machine Learning Engineer , masih ada tumpang tindih yang signifikan antara keduanya -- seorang Data Scientist harus dapat menghasilkan model pembelajaran mesin dasar, seperti halnya seorang Machine Learning Engineer harus dapat melatih model, memvalidasi hasil, dan menangani overfitting.

Menjadi Ilmuwan Data 'Scrappy'

Banyak ilmuwan data junior memiliki setidaknya satu atau dua area di mana mereka memiliki lubang yang signifikan dalam pengetahuan mereka. Ada banyak jalan menuju ilmu data, dan banyak ilmuwan data junior berasal dari latar belakang yang tumpang tindih dengan bagian tertentu dari ilmu data. Misalnya, tidak jarang profesional atau ahli statistik bioinformatika mengubah citra diri mereka sebagai ilmuwan data untuk mengambil keuntungan dari gaji yang lebih tinggi di bidang ini. Meskipun latar belakang mereka mungkin memberi mereka pemahaman yang sangat kuat tentang analitik, eksperimen ilmiah, atau pemahaman tentang cara kerja model pembelajaran mesin, para profesional ini sering memiliki sedikit pengetahuan tentang teknik, dan dengan demikian dapat membuat dan melatih model, tetapi tidak dapat menerapkannya. produksi sehingga perusahaan benar-benar dapat menggunakannya. Demikian pula, banyak ilmuwan data junior di pasar kerja tidak lebih dari gelar sarjana dalam ilmu komputer dan pemahaman yang lewat tentang pembelajaran mesin -- dalam hal ini, membuat model itu mudah, tetapi mereka mungkin kurang dalam pemahaman tentang model diri. Ini bukan masalah yang tidak dapat diatasi -- perusahaan besar hampir selalu memiliki peran di mana keterampilan kandidat dapat berguna, dan mereka dapat berinvestasi dalam melatih karyawan dan melatih mereka di area di mana mereka agak lemah.

Dengan perusahaan kecil dan menengah, ini adalah masalah yang jauh lebih besar. Ilmuwan Data di organisasi yang lebih kecil diharapkan sedikit lebih mandiri, dan kemungkinan harus "mengenakan lebih banyak topi". Untuk peran ilmu data di sebuah startup, sudah menjadi harapan umum bagi para ilmuwan data mereka untuk

menangani setiap bagian dari proyek ilmu data. Ini berarti memulai dengan mewawancarai pemangku kepentingan utama dan mengidentifikasi masalah yang harus dipecahkan, diikuti dengan membuat prototipe solusi dengan cepat sampai Anda melatih/menyetel/memvalidasi model yang memenuhi standar Anda, diikuti dengan benar-benar menempatkan model itu dalam produksi! Ini berarti sangat penting untuk menjadi 'scrappy', dan mampu menangani apa pun yang dilemparkan kepada Anda sebagai ilmuwan data. Perusahaan kecil sering tidak t memiliki dana atau infrastruktur untuk tim Teknik Pembelajaran Mesin terpisah untuk menangani detail implementasi. Dalam hal ini, kemampuan untuk memproduksi model pembelajaran mesin adalah keterampilan paling praktis dan berguna yang dapat Anda miliki di kotak peralatan Ilmu Data Anda. Untuk semua kecuali perusahaan terbesar, tidak masalah seberapa hebat Anda dalam melatih model -- sampai Anda memasukkan model itu ke dalam produksi sehingga seluruh organisasi dapat benar-benarmenggunakannya, mungkin juga tidak ada!

TL;DR di sini cukup sederhana:

1. Banyak ilmuwan data tidak tahu cara memasukkan model pembelajaran mesin ke dalam produksi.
2. Menempatkan model ke dalam produksi adalah keterampilan wajib bagi para ilmuwan data di sebagian besar perusahaan kecil hingga menengah.
3. Mampu memproduksi model akan membuat Anda menjadi kandidat yang jauh lebih menarik bagi pemberi kerja, dan memberi Anda keunggulan kompetitif!

Data Science and Cloud Computing

Kemampuan untuk memproduksi machine learning model adalah keterampilan yang berharga -- jadi bagaimana kami melakukannya? Jawaban ini telah berubah dalam beberapa tahun terakhir berkat platform komputasi awan seperti Amazon Web Services . Satu dekade yang lalu, memproduksi model pembelajaran mesin berarti membangun server web Anda sendiri dengan sesuatu seperti Flask atau Django dan menghosting di suatu tempat, sama seperti yang Anda lakukan dengan aplikasi web apa pun. Namun, penciptaan platform komputasi awan mengubah banyak hal, dan ilmu data

tidak terkecuali. Sekarang, kita bahkan tidak perlu khawatir tentang hal-hal seperti kode server -- sebagai gantinya, kita dapat menggunakan layanan yang sudah ada sebelumnya dari AWS yang dibuat khusus untuk menyederhanakan proses produksi solusi pembelajaran mesin!

Untuk sisa bagian ini, kita akan menggali lebih dalam ke semua alat luar biasa yang disediakan AWS, dan mempelajari bagaimana kami dapat menggunakannya untuk menjadikan Anda ilmuwan data yang lebih efektif!

4. Forum Diskusi

Berdasarkan tim diskusi anda, jawablah pertanyaan berikut ini.

a. Untuk apa Anda akan menggunakan model prediksi?

b. Apakah perlu untuk menyesuaikan model ke seluruh kumpulan data?

c. Seberapa tepat Anda harus dengan prediksi?

d. Apa bagian terpenting dari kumpulan data untuk dimodelkan?

C. PENUTUP

1. Rangkuman

Pengumpulan data adalah proses mengumpulkan informasi yang diperlukan dengan metode primer atau sekunder dan mengelolanya dalam format yang tepat. Analisis data adalah langkah pengumpulan data berikutnya dan di sini data yang dikelola diperiksa dan diperiksa ulang kualitasnya. Ini juga menemukan informasi yang hilang dan informasi yang tidak relevan serta cara untuk memanfaatkan dan memvalidasinya.

2. Tes Formatif

Sebagai evaluasi pemahaman materi pada Bab 7, jawablah pertanyaan berikut ini.

1. Apa yang dimaksud dengan Predictive Analytics(Analisis Prediksi)

2. Berikut ini yang termasuk analitik prediktif memanfaatkan kekuatannya dari berbagai metode dan teknologi adalah:

a. Big data

b. Data Mining

c. Pemodelan statistik

d. Machine Learning

3. Dengan menggunakan analitik prediktif untuk mempelajari perilaku dan tindakan pengguna, organisasi dapat mendeteksi aktivitas yang biasa, mulai dari penipuan kartu kredit hingga mata-mata perusahaan hingga cyber crime.

a. True

b. False

4. Berikut ini adalah beberapa contoh bagaimana organisasi memanfaatkan analitik prediktif:

a. Dirgantara

b. Otomotif

c. Energi

d. Layanan Keuangan

5. Perangkat lunak analitik prediktif dan penyedia layanan utama meliputi:

a. Aksioma

b. IBM

c. GETAH

d. Microsft

6. Beberapa tipe tipikal model prediktif meliputi:

a. Model nilai seumur hidup pelanggan

b. Model Segmentasi Pelanggan

c. Model Pemeliharaan Prediktif

d. Model Jaminan Kualitas

7. Organisasi beralih ke analitik prediktif untuk membantu memecahkan masalah sulit dan mengungkap peluang baru. Penggunaan umum meliputi:

a. Mendeteksi penipuan

b. Mengoptimalkan kampanye pemasaran

c. Meningkatkan operasi

d. Mengurangi risiko.

8. Apa yang dimaksud dengan data collection dan analisi data?

9. Berikut ini yang merupakan beberapa sumber pengumpulan data yang berbeda adalah:

a. Mengumpulkan data baru dari internet dan sumber lainnya

b. Informasi yang dikumpulkan tentang variable

c. Menggunakan data yang dikumpulkan dan disimpan sebelumnya

d. Menggunakan kembali data orang lain

10. Sebutkan dan jelaskan 2 metode pengumpulan data(data collection)

Daftar Pustaka

Predictive Analytics, What is predictive analytics? Transforming data into future insights, By John Edwards, https://www.cio.com/article/3273114/what-is-predictive-analytics-transforming-data-into-future-insights.html

Predictive Analytics, What it is and why it matters, https://www.sas.com/en_us/insights/analytics/predictive-analytics.html

Solve Problems with Predictive Analytics for Business, Data Crunch, https://datacrunchcorp.com/predictive-analytics-for-business/What is the difference between data collection and data analysis?, PhD assistance, https://www.phdassistance.com/blog/what-is-the-difference-between-data-collection-and-data- analysis/

Data Science telah menjadi salah satu disiplin ilmu yang paling penting di era digital ini. Dengan perkembangan teknologi dan meningkatnya volume data, kemampuan untuk mengolah, menganalisis, dan menafsirkan data menjadi semakin krusial. Buku "Pengantar Data Science" hadir sebagai panduan komprehensif untuk memahami dasar-dasar ilmu Data Science. Buku ini dirancang bagi siapa saja yang ingin memulai perjalanan mereka di dunia data, baik itu mahasiswa, profesional, atau siapa pun yang tertarik pada analisis data. Melalui pendekatan yang sistematis, buku ini membahas konsep-konsep fundamental, metode analisis, serta pengenalan terhadap Python. Dilengkapi dengan studi kasus nyata, buku ini tidak hanya menjelaskan teori, tetapi juga menunjukkan bagaimana Data Science dapat diterapkan dalam berbagai konteks untuk menghasilkan wawasan yang berharga. Mulailah perjalanan Anda menuju masa depan yang berbasis data dengan memahami ilmu yang akan membentuk dunia kita di masa mendatang.

ISBN 978-623-8299-27-0 (PDF)