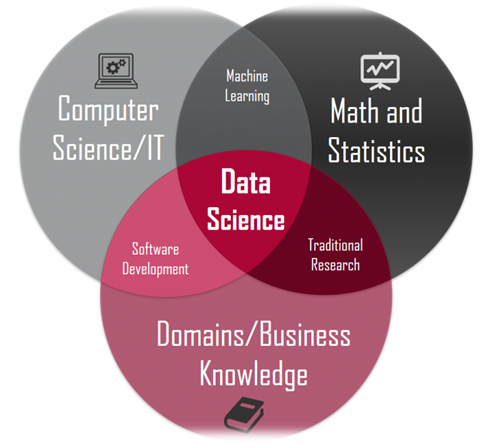
|  |  |
| --- | --- |
|  |  |
|  | **MODUL 2 PERKULIAHAN** |
|  |  |
|  | **Proyek Data Science** |
|  |  |
|  | **Pengelompokkan algoritma data science dan Jenis algoritma data science** |
|  |  |

# Pembahasan

**Pertemuan 2**

**Pengelompokkan algoritma data science dan Jenis algoritma data science**

Data science adalah ilmu yang menggabungkan matematika, statisika dengan ilmu komputer dengan tujuan analisa data (data analysis) dari suatu himpunan data baik skala kecil (sampel) maupun besar (populasi) dengan mengaplikasikan algoritma tertentu untuk tujuan menggali data (data mining) dan mendapatkan pola data serta dapat melakukan prediksi data (prediction) dengan cukup akurat yang dapat membantu dalam pengambilan keputusan dan dapat digunakan untuk membuat sistem yang cerdas (AI) yang dapat terus belajar dengan sendirinya (machine learning).

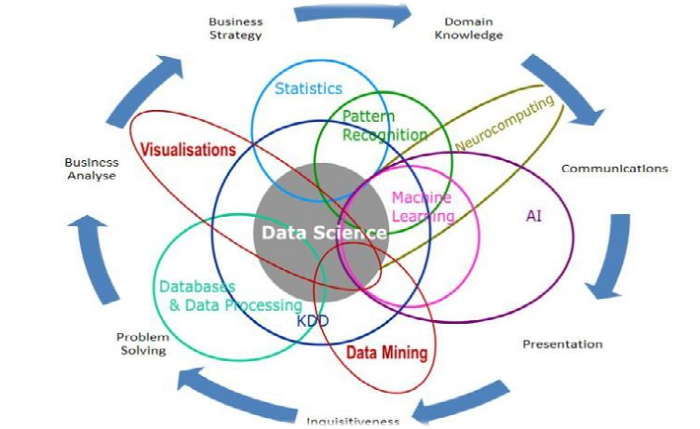


Gambar 1 Kemampuan yang dibutuhakan data science

Sumber : https://medium.com/@jrendz/pengenalan-data-science-b49a52eeef9c

1. Kemampuan Matematika dan Statistika
2. Kemampuan Pemrograman (R, Python, dan lainnya)
3. Kemampuan Database dan Query (SQL dan lainnya) dan pengolahan data
4. Kemampuan analisa data dan visualisasi data
5. Kemampuan pemahaman masalah terkait bisnis atau suatu bidang lainnya

Apa beda Data Science, Data Mining & Machine Learning

Data Science melibatkan proses berikut :

1. Data Mining adalah proses pengambilan informasi dari pola data dari himpunan data yang sebelumnya tidak diketahui, kadang disebut juga Data Discovery.
2. Data Mining fokus pada mengekstrak pola menggunakan metode statistik untuk dianalisa dan dapat juga melakukan prediksi.
3. Machine learning adalah bidang yang merupakan bagian dari Artificial Intelligence (AI) yang digunakan agar sistem komputer secara otomatis dapat belajar dengan sendirinya tanpa diberi instruksi pemrograman dan dapat meningkatkan prediksi yang akurat dan pengunaannya biasanya sifatnya realtime.

Jadi Data Mining dan Machine Learning merupakan bagian dari Data Science

Tujuan Data Science

1. Menggali data dan memberikan informasi seakurat mungkin yang digunakan untuk :
2. Deskripsi yaitu menampilkan pola data untuk dianalisa dan penemuan masalah.
3. Prediksi yaitu melakukan prediksi berupa nilai, probabilitas maupun data dan kemudian merekomendasikan hasilnya untuk digunakan sebagai alat bantu pengambil keputusan maupun secara langsung digunakan secara otomatis oleh sistem.



Jenis Pembelajaran Data Science

Supervised Learning (Prediksi)

Untuk membentuk sistem yang cerdas, sistem harus diberikan pelatihan terlebih dahulu (training) dengan data fakta (labelled training), sistem akan belajar dan membentuk pola data yang ada baru kemudian digunakan untuk melakukan prediksi, proses belajarnya seperti anak murid yang diajarkan oleh guru.

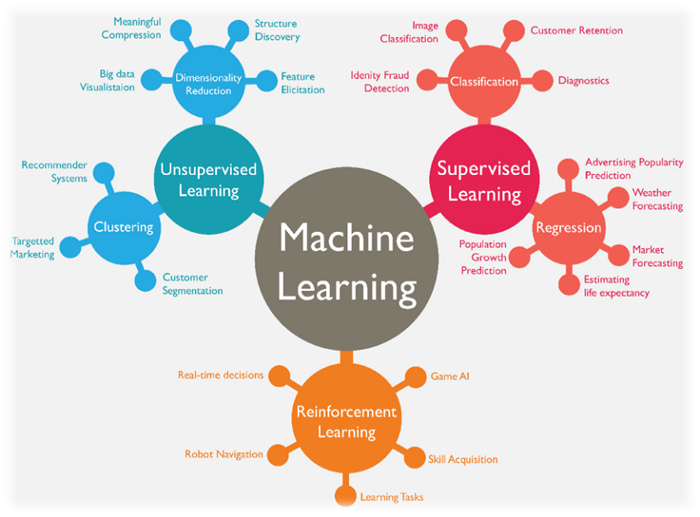
Unsupervised Learning (Deskripsi)

Sistem dapat mengandalkan data yang belum dilatih sebelumnya (unlabelled training) dan dapat membentuk pola data yang sifatnya deskriptif, bukan untuk prediksi.

Reinforced dan Deep Learning (Prediksi)

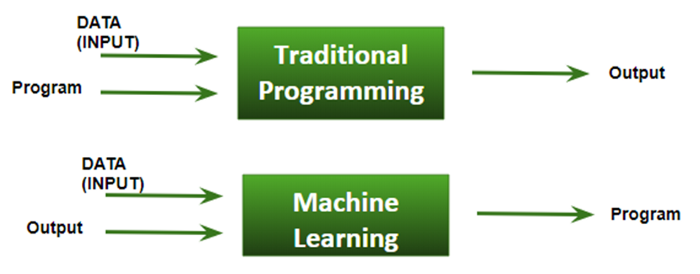
Sistem belajar dari feedback lingkungan dengan teknik learning yang iteratif (berulang-ulang) dan adaptif (menyesuaikan) seperti cara manusia belajar dengan sendirinya, algoritma digunakan untuk memaksa sistem belajar menemukan nilai optimal dengan coba-coba (trial and error). Teknik pembelajaran menggunakan algoritma Neural Network berlapis yang sangat mirip dengan cara kerja otak manusia dimana neuron-neuron satu sama lain membentuk jaringan neuron yang sangat rumit.

Biasanya digunakan untuk apa



Digunakan untuk banyak jenis seperti : Prediksi populasi, cuaca, kondisi pasar, iklan, pendeteksi penipuan, klasifikasi gambar, pola kebiasaan pelanggan, jenis atau karakteristik pelanggan, marketing yang lebih terarah, rekomendasi produk, AI Game, pendeteksi anomali dan masih banyak kegunaan lainnya.

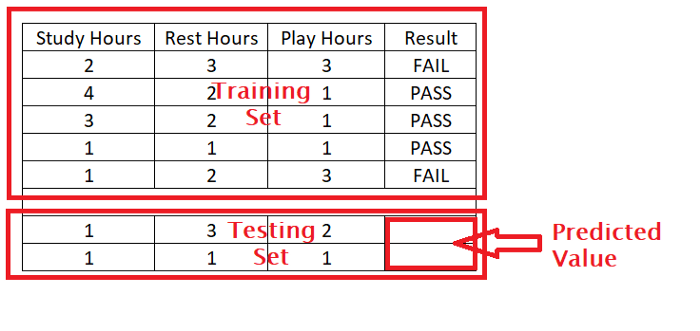
Bagaimana Machine Learning bekerja dan apa bedanya dengan pemrograman tradisional



Pemrograman tradisional membutuhkan Data sebagai input dan program yang sudah dibuat untuk menghasilkan output.

Machine Learning hanya membutuhkan Data sebagai input dan contoh output yang diinginkan, dan sistem akan mencari dengan sendirinya program perhitungannya.

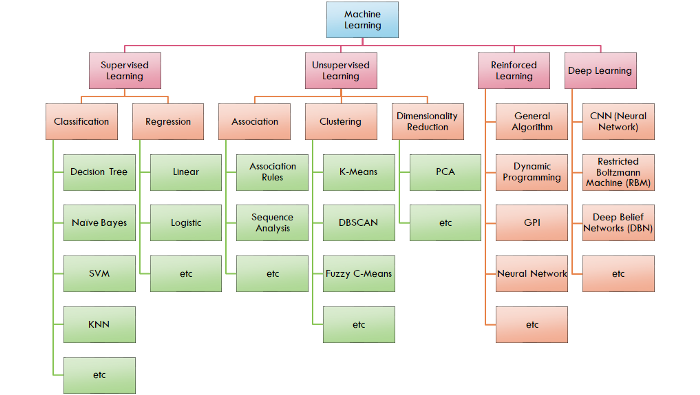
Contoh:



Agar mesin dapat belajar, dibutuhkan 2 jenis data yaitu

1. Data Training berisi data fakta yang ada
2. Data Tes berisi data yang ingin diprediksi

Contoh Algoritma Machine Learning



Algoritma Machine Learning biasanya terbagi menjadi beberapa kelompok :

1. Klasifikasi (Classification)

Digunakan untuk melakukan klasifikasi suatu data termasuk dalam suatu kategori tertentu.

1. Regresi (Regression)

Digunakan untuk melakukan proyeksi nilai berdasarkan pola data yang sudah ada.

1. Asosiasi (Association)

Digunakan untuk melihat keterkaitan dan hubungan antar data.

1. Pengelompokan (Clustering)

Digunakan untuk melihat pola pengelompokan data berdasarkan ciri-ciri yang mirip.

1. Pengurangan Dimensi (Dimensionality Reduction)

Digunakan untuk mengurangi jumlah dimensi yang kompleks menjadi sederhana agar dapat memudahkan dilihat secara visual server digunakan untuk perhitungan yang lebih sederhana.

1. Reinforced dan Deep Learning

Digunakan untuk membuat sistem berpikir dan memiliki kecerdasan seperti manusia, misalnya bisa membedakan gambar (Computer Vision), bisa membedakan kalimat dan kata (NLP), dan kecerdasan buatan lainnya misalnya seperti ChatBot.

Ranking 10 algoritma teratas

1. C 4.5

Pohon Keputusan (Decision Tree) merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan aturan. Aturan dapat dengan mudah dipahami dengan bahasa alami. Aturan ini juga dapat diekspresikan dalam bentuk bahasa basis data seperti SQL untuk mencari record pada kategori tertentu. Pohon keputusan juga berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan sebuah variabel target. Karena pohon keputusan memadukan antara eksplorasi data dan pemodelan, pohon keputusan ini sangat bagus sebagai langkah awal dalam proses pemodelan bahkan ketika dijadikan sebagai model akhir dari beberapa teknik lain(J R Quinlan, 1993).

2. K-Means

K-means merupakan salah satu metode clustering non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih cluster. Metode ini mempartisi data ke dalam cluster sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu cluster yang sama dan data yang mempunyai karateristik yang berbeda di kelompokan ke dalam cluster yang lain.

3. Support Vector Machines

SVM adalah metode machine learning yang bekerja atas prinsip Structural Risk Minimization (SRM) dengan tujuan menemukan hyperplane terbaik yang memisahkan dua buah class pada input space. SVM merupakan sistem pembelajaran yang menggunakan ruang hipotesis berupa fungsi-fungsi linier dalam sebuah ruang fitur (feature space) berdimensi tinggi, dilatih dengan algoritma pembelajaran yang didasarkan pada teori optimasi dengan mengimplementasikan learning bias yang berasal dari teori pembelajaran statistik.

4. Apriori

Algoritma apriori adalah sebuah algoritma pencarian pola yang sangat populer dalam teknik penambangan data (datamining). Algoritma ini ditujukan untuk mencari kombinasi item-set yang mempunyai suatu nilai keseringan tertentu sesuai kriteria atau filter yang diinginkan. Hasil dari algoritma ini dapat digunakan untuk membantu dalam pengambilan keputusan pihak manajemen.

5. Expectation Maximisation Algorithm

Expectation Maximisation Algorithm (EM Algorithm) adalah algoritma yang sering digunakan untuk menemukan nilai estimasi Maximum Likelihood (ML) dari parameter dalam sebuah model probabilistic, dimana model juga tergantung pada latent variabel yang belum diketahui. Dalam algoritma ini, ada dua hal yang dilakukan secara bergantian yaitu E step yang menghitung nilai ekspektasi dari likelihood termasuk laten variabel seolah-olah seperti mereka ada, dan M step menghitung nilai estimasi ML dari parameter dengan memaksimalkan nilai ekspektasi dari likelihood yang ditemukan pada E step.

6. PageRank

PageRank adalah sebuah algoritma yang telah dipatenkan yang berfungsi menentukan situs web mana yang lebih penting/populer. PageRank merupakan salah satu fitur utama mesin pencari Google dan diciptakan oleh pendirinya, Larry Page dan Sergey Brin yang merupakan mahasiswa Ph.D. Universitas Stanford.

7. k-Nearest Neighbors

K-Nearest Neighbor (KNN) adalah suatu metode yang menggunakan algoritmasupervised dimana hasil dari query instance yang baru diklasifikan berdasarkan mayoritas dari kategori pada KNN. Tujuan dari algoritma ini adalah mengklasifikasikan obyek baru bedasarkan atribut dan training sample. Algoritma metode KNN sangatlah sederhana, bekerja berdasarkan jarak terpendek dariquery instance ke training sample untuk menentukan KNN-nya. Training samplediproyeksikan ke ruang berdimensi banyak, dimana masing-masing dimensi merepresentasikan fitur dari data. Ruang ini dibagi menjadi bagian-bagian berdasarkan klasifikasi training sample.

8. Naıve Bayes

Naïve Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai teorema Bayes. Teorema tersebut dikombinasikan dengan ”naive” dimana diasumsikan kondisi antar atribut saling bebas [1]. Pada sebuah dataset, setiap baris/dokumen I diasumsikan sebagai vector dari nilai-nilai atribut <x1,x2,…,x3> dimana tiap nilai-nilai menjadi peninjauan atribut Xi (iЄ[1,n])).

9. Classiﬁcation and Regression Trees

CART (Classification And Regression Trees) yaitu metode pohon regresi dan pohon klasifikasi. Jika variabel dependen yang dimiliki bertipe kategorik maka CART menghasilkan pohon klasifikasi (classification trees), sedangkan jika variabel dependen yang dimiliki bertipe kontinu atau numerik maka CART menghasilkan pohon regresi (regression trees).

10. Adaboost

Model standard dari algoritma adaboost terdiri dari dua bagian, yaitu bagian offline training dan bagian online recognizing. Bagian offline training adalah bagian proses pelatihan data yang tidak bekerja secara realtime. Bagian ini meliputi penginputan sampel gambar positif dan sampel gambar negatif, preprocessing, pelatihan data oleh algoritma adaboost sampai membangun detektor. Setelah detektor terbentuk kita bisa melakukan pendeteksian secara realtime/online recognizing terhadap data pengujian. Sebelum melakukan pendeteksian dengan algoritma adaboost, terlebih dahulu data pengujian sudah harus mengalami preprocessing.

Quiz Pertemuan 2

Berikan 1 contoh penerapan algoritma tebaik dengan studi kasus yang ada saat ini. (Tidak boleh sama dengan kawan yang lain) tuliskan perhitungan algoritma.

# Daftar Pustaka

[1] “Data Scientist: The Sexiest Job of The 21st Century” Diakses 01 April 2019 dari https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

[2] “50 Best Jobs in America 2019” Diakses 01 April 2019 dari https://www.glassdoor.com/List/Best-Jobs-in-America-LST\_KQ0,20.htm

[3] “Tech in Asia Jobs” Diakses 02 Desember 2018 dari https://www.techinasia.com/jobs

[4] “Kalibrr: Where Jobs Find You” Diakses 02 Desember 2018 dari https://www.kalibrr.com/

[5] “Data Scientist Earning More Than CAs, Engineers” Diakses 01 April 2019 dari https://timesofindia.indiatimes.com/india/Data-scientists-earning-more-than-CAs-engineers/articleshow/52171064.cms

[6] “5 Facts About Software Engineers, Like Which One Gets Paid The Most” Diakses 01 Apil 2019 dari https://learning.linkedin.com/blog/tech-tips/the-american-city-that-pays-software-engineers-the-most–and-oth

[7] “Top 10 Reasons Why You Should Learn Data Analytics” Diakses 01 April 2019 dari https://bigdata-madesimple.com/10-reasons-why-you-should-learn-data-analytics/

[8] “5 Reasons Why Everybody Should Learn Data Analytics” Diakses 01 April 2019 dari https://www.sas.com/en\_au/insights/articles/analytics/5-reasons-why-everybody-should-learn-data-analytics.html

[9] “Permintaan Tenaga Data Scientist Melonjak, Jadikan Profesi Ini Kian Menjanjikan” Diakses 10 Desember 2018 dari https://id.techinasia.com/talk/profesi-data-scientist-menjanjikan

[10] “Go-Jek Buka Kantor Data Science di Singapura, Apa Alasannya?” Diakses 01 April 2019 dari https://www.liputan6.com/tekno/read/2998722/go-jek-buka-kantor-data-science-di-singapura-apa-alasannya

[11] Introduction to Data Science. A Python Approach to Concepts,Techniques and Applications. Laura Igual, Santi Segui. Tahun

2017.

[12] Cathy O'Neil and Rachel Schutt. Doing Data Science, Straight Talk from The Frontline. O'Reilly. 2014.

[13] Jure Leskovek, Anand Rajaraman and Jerey Ullman. Mining of Massive Datasets. v2.1, Cambridge University Press. 2014. (free online)

[14] Mohammed J. Zaki andWagner Miera Jr. Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press. 2014.

[15] https://anaktik.com/data-science/

[16] https://anaktik.com/skill-data-scientist/