Pertemuan 1

Mata Kuliah Pilihan Data Science (3 SKS)

Machine Learning and Data Science

Mustakim, S.T., M.Kom

Program Studi Sistem Informasi

https://mustakim.predatech.org | mustakim@uin-suska.ac.id | +6285275359942

Data Mining

Paradigma dan Mindset Tentang Data Mining

Menghitung → Matematika → Ngoding → Data Besar Sekali → Susah → Ribet

Tidak Serumit yang Dibayangkan

- Ada Data
- Ada Metode
- Ada Pengatahuan/ Rule

Tools dan Library

Paradigma dan Mindset Tentang Data Mining

Paradigma dan Mindset Tentang Data Mining

Pemula

- Cari **Data** → Diproses Pakai Tools → Gak Perlu Ngoding → Pengetahuan → Analisis
- Cari **Data** → Ngoding Dikit → Pakai Library → Pengetahuan → Analisis
- Cari **Data** → Ngoding Dikit + Tools → Pengetahuan → Analisis

Riset

- Baca Paper → Cari Kelemahan dan Kelebihan → Terapkan dan Perbaiki → Analisis
- Baca Paper → Cari Data Pembanding → Cari Metode Pembanding → Bandingkan → Analisis
- Baca Paper → Baca Paper → Baca Paper → Analisis

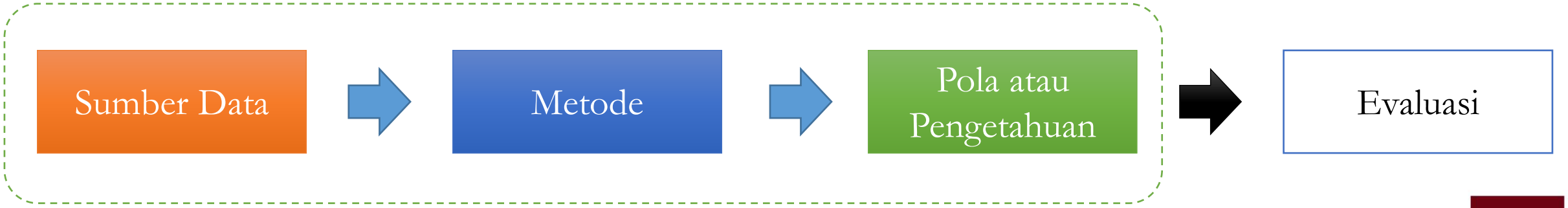
Data

Sekumpulan keterangan atau fakta mentah berupa simbol, angka, kata-kata, atau citra, yang didapatkan melalui proses pengamatan atau pencarian ke **sumber-sumber tertentu**.

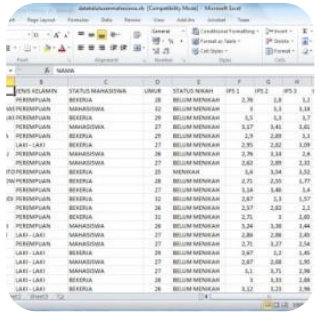
- Database
- Dokumen
- Web/ Sosial Media
- Log Sistem
- Sumber Lain

Data Mining

Disiplin ilmu yang mempelajari **metode** untuk **mengekstrak pengetahuan** atau **menemukan pola** dari suatu **data yang besar.**

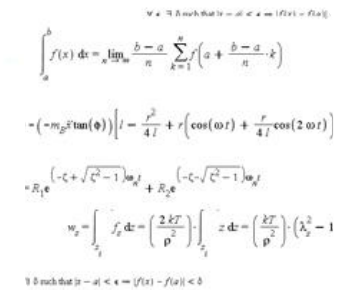


Alur Data Mining



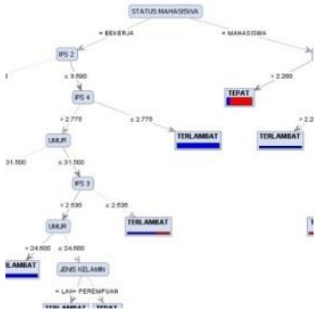
1. Himpunan Data

(Pemahaman dan Pengolahan Data)



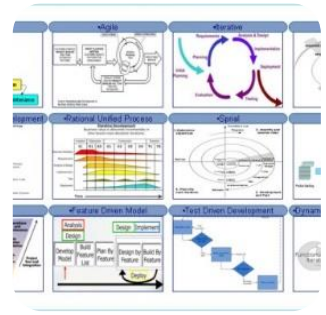
2. Metode Data Mining

(Pilih Metode Sesuai Karakter Data)



3. Pengetahuan

(Pola/ Model/ Rumus/ Tree/ Rule/ Cluster)



4. Evaluasi

(Akurasi, Validitas, RMSE, Lift Ratio,…)

Data dalam Machine Learning/ Data Mining/ Data Science

- Data Video = Computer Vision
- Data Gambar = Image Processing
- Data Terstruktur = **Data Mining**
- Data Text/ Non-Testruktur = Text Mining

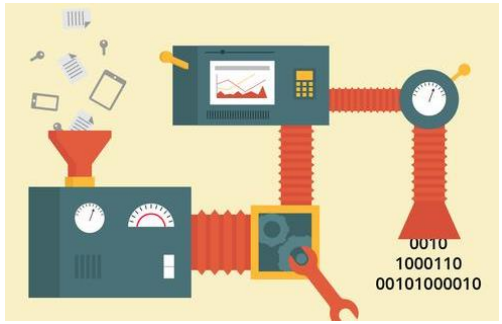
KLASIFIKASI TINGKAT KEMUNGKINAN SESEORANG TERKENA PENYAKIT DARAH TINGGI

No	Nama	Atribut						Kelas Data
		Jenis Kelamin	Umur	Tensi	Berat Badan	Golongan Darah	Konsumsi Obat	Kemungkinan Darah Tinggi
1	Ahmad Jaini	Laki-Laki	28	Rendah	55	O	Ya	Tidak
2	M. Zakiy Fauzi	Laki-Laki	30	Tinggi	48	A	Ya	Ya
3	M. Anang Ramadhan	Laki-Laki	19	Rendah	45	A	Tidak	Tidak
4	Celsa Bella	Perempuan	29	Sedang	50	O	Ya	Ya
5	Sasha Ervina Rahmadhani	Perempuan	20	Rendah	53	O	Ya	Tidak
6	Arianto Tarigan	Laki-Laki	32	Tinggi	52	A	Tidak	Ya
7	Trisdaningsih	Perempuan	21	Rendah	55	AB	Tidak	Tidak
8	Ervan Wahyudi	Laki-Laki	17	Tinggi	45	O	Tidak	Ya
9	Fachri Hadi	Laki-Laki	32	Rendah	60	O	Ya	Ya
10	Yulia Fitriani	Perempuan	17	Rendah	50	AB	Ya	Tidak
11	Assad Hidayat	Laki-Laki	18	Rendah	43	O	Ya	Tidak
12	Imaduddin Syukra	Laki-Laki	17	Sedang	74	B	Tidak	Ya
13	Dini Octari Rahmadia	Perempuan	17	Rendah	48	B	Tidak	Tidak
14	Rizki Handinata	Laki-Laki	34	Tinggi	45	A	Ya	Ya
15	M. Ridwan	Laki-Laki	20	Sedang	40	A	Ya	Ya
16	Nova Sestri Yeni	Perempuan	21	Rendah	50	AB	Tidak	Tidak
17	Taufiq Qurrohman	Laki-Laki	25	Rendah	46	O	Ya	Tidak
18	Yoga Rizola P	Laki-Laki	20	Sedang	50	O	Tidak	Tidak
19	Qumfa Anzir	Laki-Laki	21	Tinggi	50	B	Tidak	Ya
20	Hanafi MP	Laki-Laki	26	Tinggi	56	B	Ya	Ya

Proses Data Mining

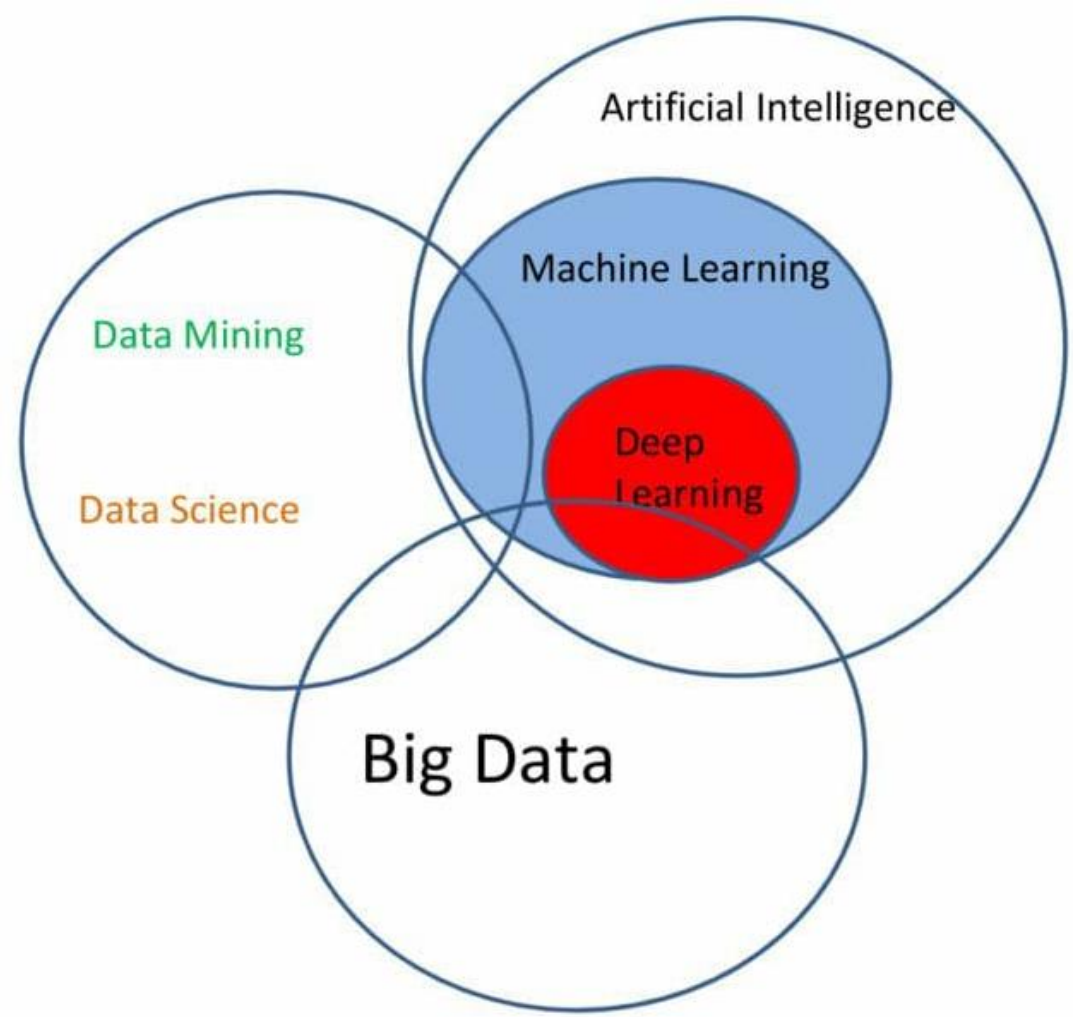
Input

Proses

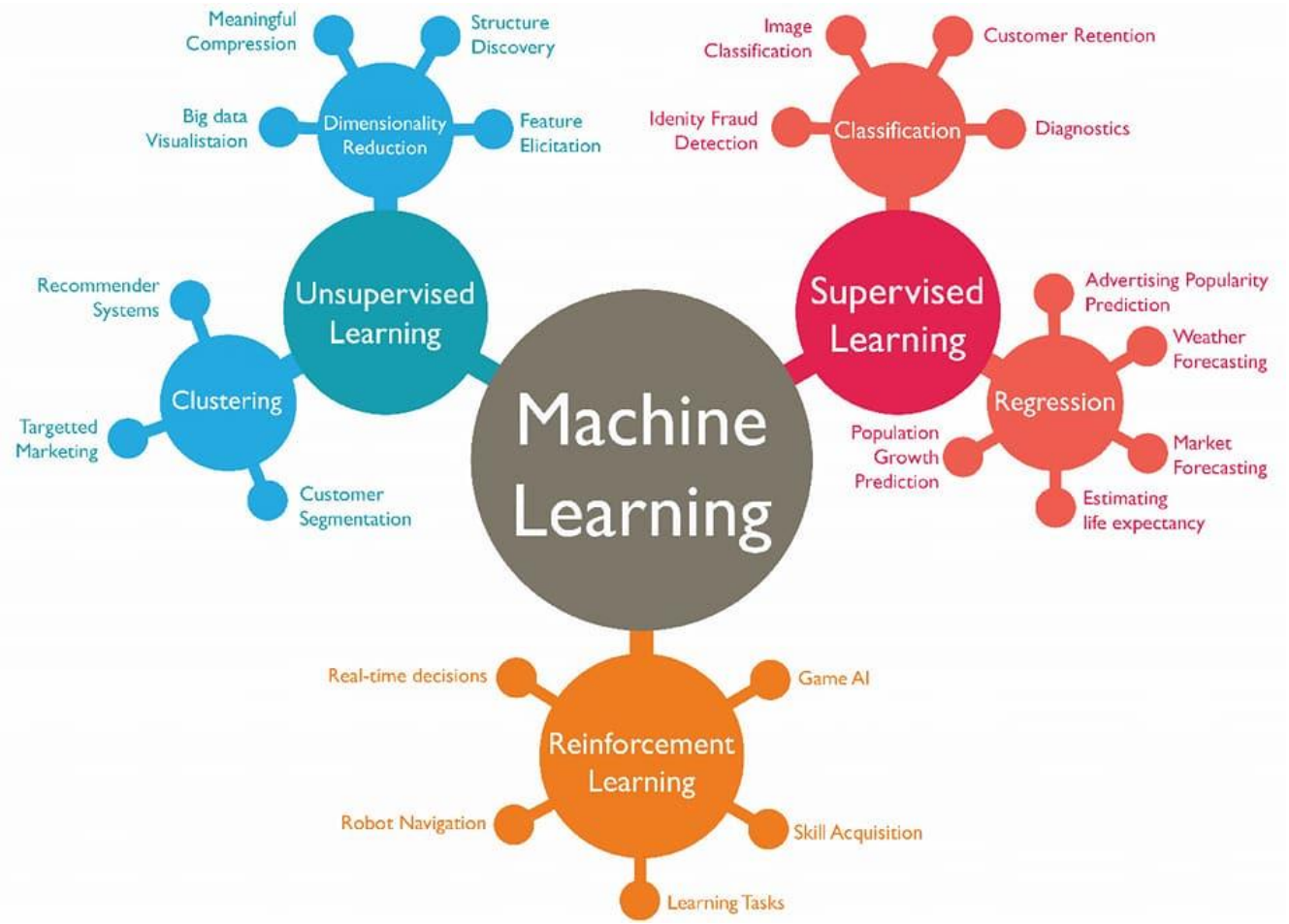


Output

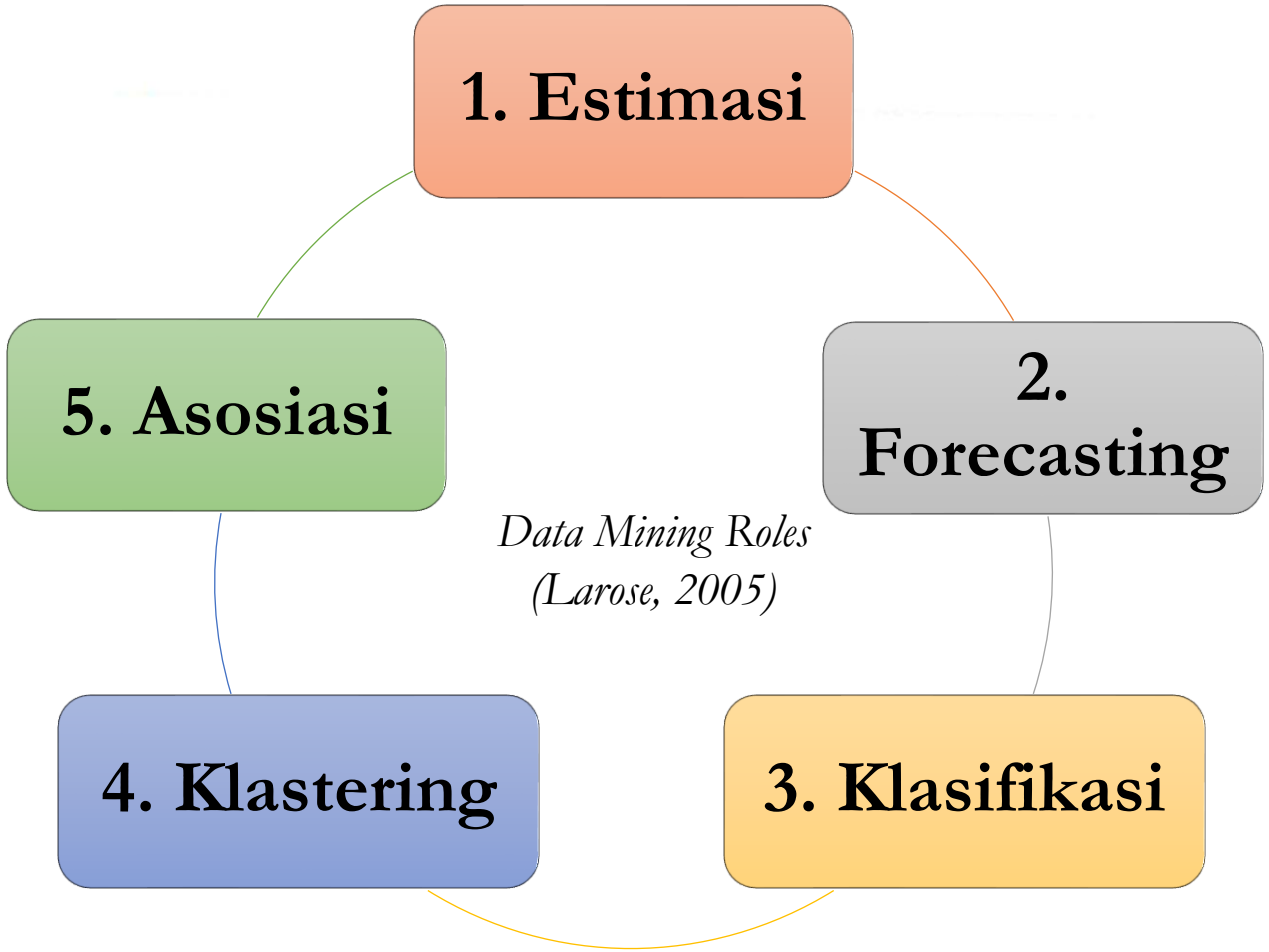
Bagan Irisan Keilmuan



Metode Learning dalam Data Mining



Peran Utama Data Mining



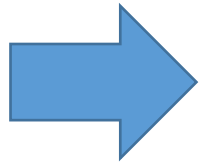
Data Mining Roles
(Larose, 2005)

Algoritma Data Mining

1. Estimation (Estimasi):
   - Linear Regression, Neural Network, Support Vector Machine, etc
2. Prediction/Forecasting (Prediksi/ Peramalan):
   - Linear Regression, Neural Network, Support Vector Machine, etc
3. Classification (Klasifikasi):
   - Naive Bayes, K-Nearest Neighbor, C4.5, ID3, CART, Linear Discriminant Analysis, Logistic Regression, etc
4. Clustering (Klastering):
   - K-Means, K-Medoids, Self-Organizing Map (SOM), Fuzzy C-Means, etc
5. Association (Asosiasi):
   - FP-Growth, A Priori, Coefficient of Correlation, Chi Square, etc

Evaluasi Data Mining

1. Estimation:
   - Error: Root Mean Square Error (RMSE), MSE, MAPE, etc
2. Prediction/Forecasting (Prediksi/Peramalan):
   - Error: Root Mean Square Error (RMSE) , MSE, MAPE, etc
3. Classification:
   - Confusion Matrix: Accuracy
   - ROC Curve: Area Under Curve (AUC)
4. Clustering:
   - Internal Evaluation: Davies–Bouldin index, Dunn index,
   - External Evaluation: Rand measure, F-measure, Jaccard index, Fowlkes–Mallows index, Confusion matrix
5. Association:
   - Lift Charts: Lift Ratio
   - Precision and Recall (F-measure)



Comparison

Optimization/ Improvement

Data Mining Course

1. Pengenalan dan Konsep Dasar Data Mining
2. Data dan Model Statistika
3. Model-model Statistika dalam Data Mining
4. *Knowledge Discovery in Database* (KDD)
5. *Supervised Learning* dan Contoh Terapan Algortima (KNN, NBC dan C4.5)
6. *Unsupervised Learning* dan Contoh Terapan Algortima (K-Means, K-Medoid, FCM dan DBSCAN)
7. *Association Rule* dan Terapannya pada Kasus Market Basked Analysis
8. *Estimation* dan *Forecasting*
9. *Spiliting Data* dengan *Cross Validation*, *Hold Out* dan *Clustering*
10. Evaluasi Model Data Mining
11. Terpan teknik *Classification* dan *Clustering* menggunakan RapidMiner
12. *Text Mining* dan *Sentiment Analysis*
13. Terapan Text Mining untuk *Clustering* dan *Classification*
14. *Improvement* dan *Comparison* Algoritma Data Mining

Data Science Course

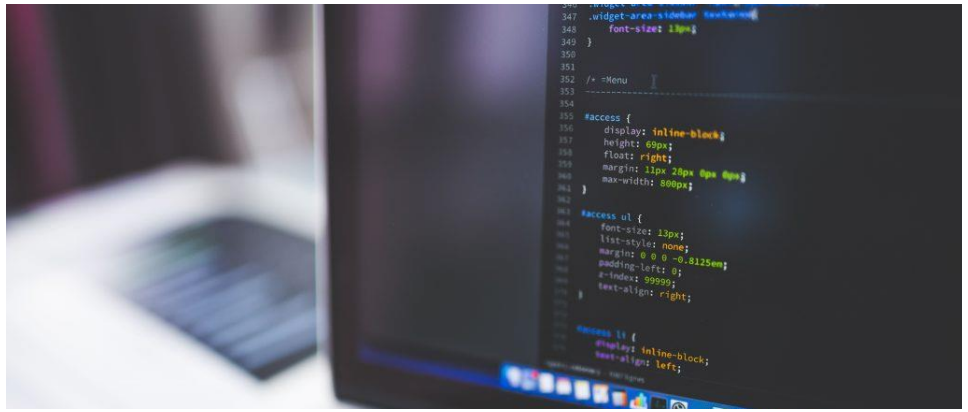
1. Introduction of Data Science
2. Mathematics and Statistics for Data Science
3. Implementation of Machine Learning, Data Mining and Data Science
4. Data Science Process and Data Collection Structures
5. Summarizing and Visualizing Data
6. Unstructured Data Pre-Processing and Text Analysis (Google Colaboratory)
7. Feature Engineering and Social Media Analysis
8. Intermediate Machine Learning and Applied Research
9. Algorithm of Dimensionality Reduction (PCA and LDA) with Scikit-Learn Library
10. Advanced: Supervised Learning and Unsupervised Learning (Google Colaboratory: Combine Method)
11. Advanced: Comparison and Improvement Algoritthm
12. Model of Evaluation and Optimization (Google Colaboratory)
13. Natural Language Processing (NLP)
14. Hadoop Big Data

Data Science

Data Science

- Data Science adalah ilmu pengetahuan interdisiplin tentang metode komputasi,
- Data: yang mencakup tiga fase yaitu Desain Data, Pengumpulan Data dan Analisis Data,
- Komponen Utama: Computer and IS, ML and DM, Math and Statistic, Unicorn dan Subject Matter Expertise (SME),
- SME: Penyelesaian permasalahan dalam kasus bisnis dalam fase Analisis.

Paradigma Data Science



- Data science merupakan ilmu yang menggabungkan sebuah kemahiran di bidang ilmu tertentu dengan keahlian pemrograman, matematika, dan statistik.
- Tujuannya adalah untuk mengekstrak sebuah pengetahuan atau informasi dari data.
- Orang yang mahir dalam bidang data science menggunakan algoritma machine learning atau pemelajaran mesin untuk mengolah teks, gambar, video, audio, dan lain-lain untuk menghasilkan sistem kecerdasan buatan.

Pilar Data Science

- **Bisnis**
  - Tujuan: untuk membantu perancangan strategi guna menyelesaikan masalah bisnis

- **Matematika dan statistika**
  - Data science sangat membutuhkan ilmu matematika, karena data harus diolah secara kuantitatif.
  - Statistik untuk data science adalah hal yang tak kalah penting.
  - Tidak hanya mengerti statistika klasik, seorang data scientist juga perlu memahami statistika Bayes

Pilar Data Science (1)

- **Teknologi**
  - Seorang data scientist perlu menguasai bahasa pemrograman seperti SQL, Phyton, R, SAS, Java, Scala, Julia, dan masih banyak lagi.
  - Seorang data scientist harus mampu berpikir layaknya algoritma dalam memecahkan permasalahan yang paling sulit sekalipun.

- **Obtain**
  - Data Collecting
  - Pemahaman terhadap jenis-jenis data
  - Pengambilan data dengan teknik dan pemahiran tertentu
  - Dapat menggunakan bahasa pemrograman atau beberapa metode lain dalam pengambilan data
  - Memahami ukuran data dan bagian-bagian dari data seperti atribut dan record

Proses Data Science (1)

- **Scrub**
  - Scrubing adalah proses pembersihan data
  - Pemahaman terhadap metode pembersihan data
  - Preprocessing data dan Langkah-langkahnya
  - Standarisasi data atau transformasi data
  - Pemisahan kategori-kategori berdasarkan kebutuhan

Proses Data Science (2)

- **Explore**
  - Pemeriksaan propertinya, karena tipe data yang berbeda memerlukan perlakuan yang berbeda pula
  - Statistik deskriptif harus dihitung untuk dapat mengekstrak fitur dan menguji variabel yang signifikan
  - Visualisasi data digunakan untuk mengidentifikasi pola dan tren signifikan dalam data yang sudah kamu dapatkan
  - Memperoleh gambaran yang lebih jelas dengan grafik agar pentingnya data dapat lebih dipahami.

Proses Data Science (3)

- **Model**
  - Membuat model data untuk mencapai tujuan yang diinginkan
  - Mengunakan regresi dan prediksi untuk memperkirakan nilai diwaktu mendatang serta melakukan klasifikasi dan pengelompokan grup nilai dari data

Proses Data Science (4)

- **Interpret**
  - Membuat output dari pemrosesan data, untuk dapat dipahami oleh orang awam
  - Presentasinya bertujuan untuk menjawab persoalan bisnis berdasarkan data yang diperoleh.
  - Pada tahap interpretasi data, kemampuan komunikasi yang baik juga sangat dibutuhkan untuk menyampaikan poin-poin pentingnya secara efektif pada semua orang yang berkepentingan

Peran dan Keunggulan Data Science

1. Membantu Manajemen dan Membuat Keputusan Lebih Baik
2. Membantu Analisis Prediksi
3. Membantu Mengidentifikasi Peluang Baru untuk Bisnis
4. Dapat Mengklasifikasi dan Personalisasi
5. Mengimplementasikan Ilmu Data Science di Masa Depan

Keunggulan Data Science untuk Kasus Bisnis

1. Akurat dalam Menganalisis
2. Hasil Analisis Relevan dengan Keinginan Pelanggan
3. Kemampuan Image Recognition
4. Penggunaan Fitur Chatbot

Mantafaat Data Science untuk Dongkrak Kasus Bisnis

1. Analisis Data Keuangan
2. Memprediksi Suatu Tren
3. Membantu Rekomendasi Produk
4. Filtering Data & Meminimalisir Aksi Penipuan
5. Customer Service Online Otomatis

Terima Kasih