Rakamin Academy

M Mirza Fahmi

4 years experience as Data Specialist

Muhammad Mirza Fahmi

A self-taught data scientist starting from an startups and enterprise company handling wide range of analytics including (but not limited) IT Security, Fraud, and Telecommunication. Currently focus on building data ecosystem at Digital Payment.

Education Background

2007-2013
Bachelor Degree
Physics

2014-2017
Master Degree
Computational Sciences

Rakamin Academy

Machine Learning Preparation

Feature Extraction

Outline Pembelajaran

Rakamin Academy

Feature Extraction

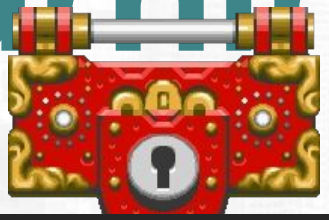
- Apa itu Machine Learning?
- Review: Machine Learning dan Data Science
- Tipologi Machine Learning
- Workflow Machine Learning
- Kutukan Dimensi
- Mengolah Feature

Apa itu Machine Learning (ML)?

Belajar mengajari mesin cara belajar

Rakamin Academy

Machine Learning



Berikan 3 tebakan untuk lanjut ke slide berikutnya!

Mari mulai dengan yang lebih sederhana.

Apa itu ~~Machine~~ Learning (~~M~~L)?

*dalam konteks Machine Learning

Kasus #1
(Easy)

Data Rekap Kelulusan Murid SMP Negeri 666 Paku (Fiktif)

Presensi	Nilai Ujian Akhir	Donasi Orangtua	Lulus
90%	60	1 Milyar	YA
70%	70	0	YA
90%	60	0.5 Milyar	TIDAK
50%	100	1 Milyar	YA
100%	60	0	TIDAK
20%	10	5 Milyar	YA
80%	80	1 Milyar	YA

90%	0	10 Milyar	?

Pertanyaan #1: Berdasarkan data yang kalian lihat di tabel, apakah murid dengan presensi 90%, nilai ujian akhir 0, dan donasi orangtua 10 Milyar akan lulus?

Apakah syarat kelulusan SMP Negeri 666 Paku?

Data Rekap Kelulusan Murid SMP Negeri 666 Paku (Fiktif)

Presensi	Nilai Ujian Akhir	Donasi Orangtua	Lulus
90%	60	1 Milyar	YA
70%	70	0	YA
90%	60	0.5 Milyar	TIDAK
50%	100	1 Milyar	YA
100%	60	0	TIDAK
20%	10	5 Milyar	YA
80%	80	1 Milyar	YA
90%	**0**	**10 Milyar**	**YA**

Syarat Kelulusan (salah satu terpenuhi → lulus):

- **Presensi > 60% & Nilai Ujian > 60**
  Atau
- **Donasi > 0.5 Milyar**

Kasus #2 (Medium)

Data Nilai Ujian Bahasa Qalbu SMP Negeri 666 Paku (Fiktif)

Jenis Kelamin	Presensi	Uang Jajan/Hari	Nilai Ujian
Laki-laki	60%	100 Ribu	70
Laki-laki	80%	50 ribu	85
Perempuan	100%	80 ribu	92
Laki-laki	80%	0	90
Perempuan	40%	0	70
Perempuan	60%	50 ribu	75
Perempuan	100%	100 ribu	90

Perempuan	100%	0	?

Pertanyaan #2:
Berdasarkan data di atas, berapa nilai ujian murid berjenis kelamin Perempuan, presensi 100%, dan tanpa uang jajan?

Apakah formula menghitung nilai ujian SMP negeri 666?

Data Nilai Ujian Bahasa Qalbu SMP Negeri 666 Paku (Fiktif)

Jenis Kelamin	Presensi	Uang Jajan/Hari	Nilai Ujian
Laki-laki	60%	100 Ribu	70
Laki-laki	80%	50 ribu	85
Perempuan	100%	80 ribu	92
Laki-laki	80%	0	90
Perempuan	40%	0	70
Perempuan	60%	50 ribu	75
Perempuan	100%	100 ribu	90

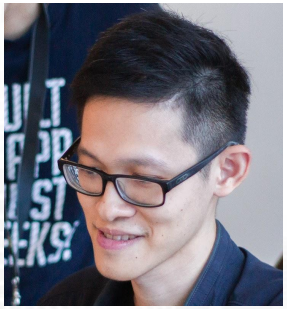
Perempuan	100%	0	100

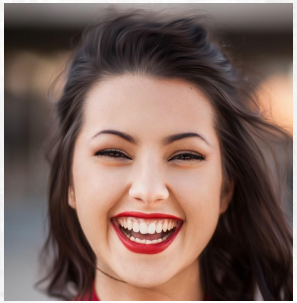
Rumus nilai ujian:

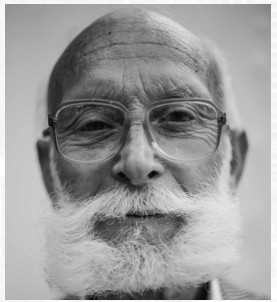
- **50 + (Presensi Kelas / 2) - (Uang Jajan / 10 ribu)**

Kasus #3
(Hard)





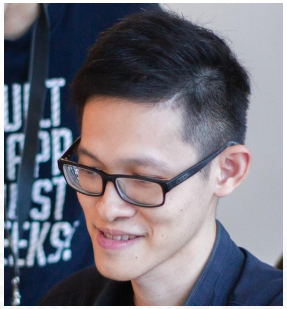



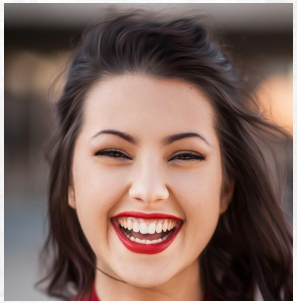


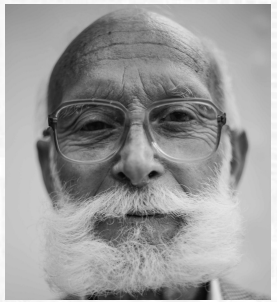
Pertanyaan #3: Di antara foto-foto di atas manakah yang bukan foto murid SMP Negeri 666 Paku?

Kenapa ini (hard)? Bayangkan kalian **tidak** tahu apa-apa sebelumnya laiknya bayi yang baru lahir.







Kenapa yang kanan bawah? Karena satu-satunya yang tua? Kenapa kita bisa tahu bahwa dia tua? Pixel-pixel mana yang menunjukkan bahwa dia tua?

Learning

- Dari data/sekumpulan observasi yang dimiliki
- Pelajari pola/aturan yang menggambarkan bagaimana data/observasi tsb. dihasilkan

+ Gunakan pola yang ditemukan untuk membuat keputusan mengenai data/observasi baru

Machine Learning

Sebuah proses:

- Dari data/sekumpulan observasi yang dimiliki
- Secara otomatis mempelajari pola/aturan yang ada dengan bermacam algoritma

\+ **Gunakan pola yang ditemukan untuk membuat keputusan mengenai data/observasi baru**

Machine Learning

2 pertanyaan:

- Apakah kita dapat menangkap pola-pola/aturan yang tersimpan di dalam data?

\+ Apakah pola-pola yang kita tangkap dapat diterapkan untuk data yang baru?

Belajar ML = belajar:

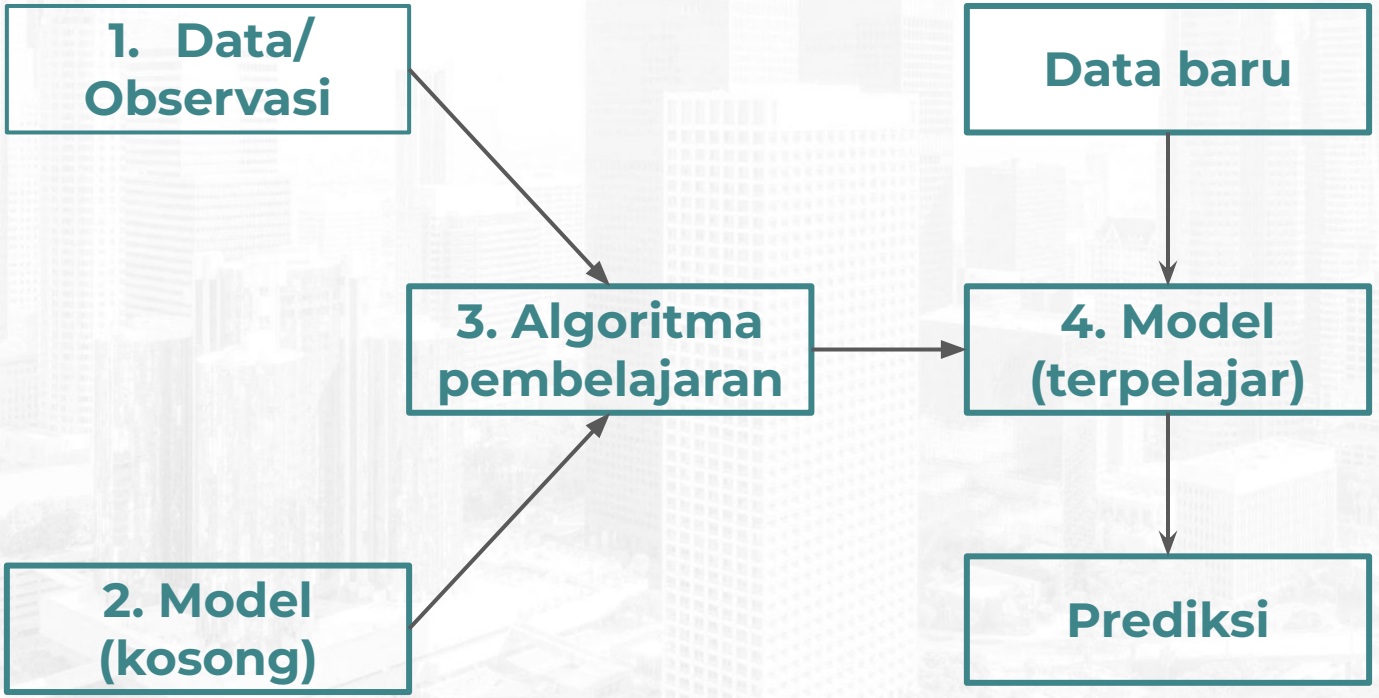
1. Bagaimana **cara mempersiapkan data** agar dapat digunakan dalam proses pembelajaran?

2. **Apa saja model** yang dapat diajari untuk menemukan pola-pola/aturan dalam data?

3. Bagaimana **cara mengajari model** tersebut dengan data yang kita punya?

4. Bagaimana cara memastikan **model** yang sudah kita ajari dapat **mengaplikasikan ilmunya** pada data-data baru?

Diagram Konsep ML

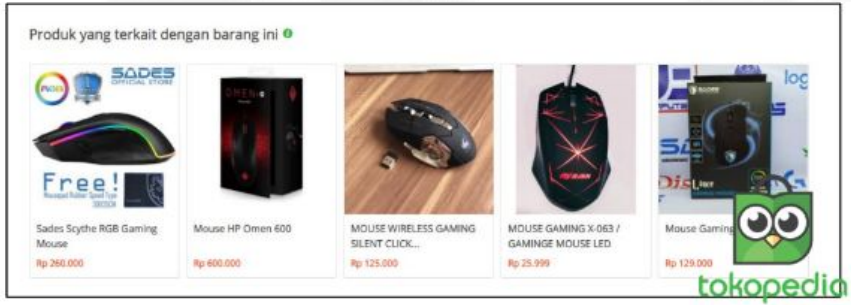


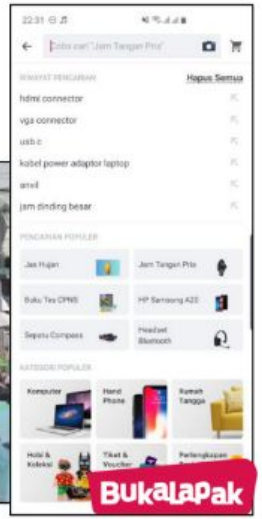
Rakamin Academy

Perhatian:

- Mesin belajar HANYA dari data yang dimiliki
- Mesin TIDAK PUNYA pengetahuan dasar
- Pada dasarnya mesin hanya mengerti ANGKA

Machine Learning di Dunia Nyata





Outline Pembelajaran

Rakamin Academy

Feature Extraction

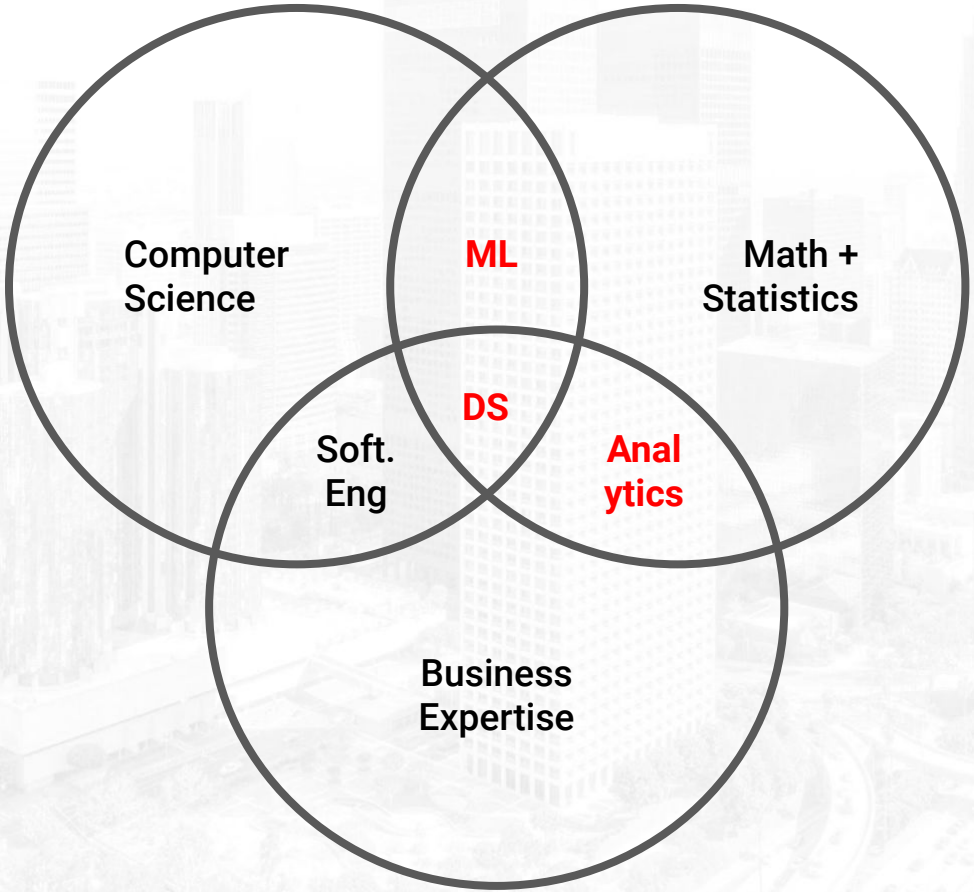
- [x] Apa itu Machine Learning?
- [ ] Review: Machine Learning dan Data Science
- [ ] Tipologi Machine Learning
- [ ] Workflow Machine Learning
- [ ] Kutukan Dimensi
- [ ] Mengolah Feature

Rakamin Academy

ML dan Data Science

Posisi ML dalam disiplin Data Science

Data Science = Programming + Statistics + Business Domain



*dari sesi 1!

...sebelumnya, di Rakamin...

1. SQL

SQL (Structured Query Language) adalah bahasa standar untuk mengakses DBMS dan mengolah data dalam database.

2. Statistics

3. Python

Python is an interpreted , high-level, general-purpose programming language.

Mudahnya...
Python adalah bahasa pemrograman yang mudah dipahami oleh manusia dan dapat digunakan untuk berbagai tujuan, mulai dari analisis data, membuat website, aplikasi dll

Bagaimana semuanya terhubung?

Ketiga alat tadi digunakan untuk
“melakukan data science”

Data Science

Data Mining/Analytics: Inspiration

- Masuk: Data
- Proses: Data extraction, manipulation, visualization
- Keluar: Insight tentang korelasi -> hipotesis

Statistical Inference: Rigor

- Masuk: Hipotesis
- Proses: Hypothesis testing, A/B testing
- Keluar: Validasi hipotesis, keputusan

Machine Learning: Scale

- Masuk: Data (optional: insight)
- Proses: Data Prep -> Model Train -> Evaluation -> Test
- Keluar: Resep (model) pengambilan keputusan

“Marketplace A: Penjualan kosmetik kucing di semester ini naik.”

Data Mining/ Analytics: Inspiration

- In: Data penjualan + data pembeli
- Out: “Ada korelasi positif antara pembelian mi instan dan kosmetik kucing.”

Statistical Inference: Rigor

- In: (hipotesis) “Rekomendasi kosmetik kucing setelah pembelian Mi Instan akan meningkatkan penjualan.”
- Out: “Kita yakin 99.XX% bahwa rekomendasi kosmetik kucing pada pembeli mi instan meningkatkan total penjualan.”

Machine Learning: Scale

- In: Data (optional: insight, domain knowledge)
- Out: Rumus kemungkinan seorang pembeli akan membeli kosmetik kucing 1 minggu kedepan.

Outline Pembelajaran

Rakamin Academy

Feature Extraction

- [x] Apa itu Machine Learning?
- [x] Review: Machine Learning dan Data Science
- [ ] Tipologi Machine Learning
- [ ] Workflow Machine Learning
- [ ] Kutukan Dimensi
- [ ] Mengolah Feature

Jenis-jenis Masalah ML

spoiler: ada banyak

Jenis - jenis Machine Learning

Supervised Learning

- Tersedia data dengan target
- Tujuan: memprediksi data baru dengan benar
- Jenis: Klasifikasi dan regresi

Unsupervised Learning

- Tersedia data tanpa target
- Tujuan: menyingkap pola hubungan yang tersembunyi
- Clustering, reduksi dimensi

Reinforcement Learning

- Trial and error learning pada suatu lingkungan dengan aturan spesifik
- Tujuan: Melatih ‘agen’ dalam suatu ‘task’ untuk memaksimalkan ‘reward’

* diluar scope Bootcamp

Supervised Learning

- Tersedia data dengan target
- Klasifikasi dan regresi

Data = Fitur + Target

No	Mahasiswa	Nilai Quiz	Nilai UTS	Nilai UAS	Nilai Akhir	Lulus
1	Budi	55	70	80	68.33	Tidak
2	Ahmad	65	75	80	73.33	Ya
3	Sandi	70	70	75	71.67	Ya
4	Robert	90	65	50	68.33	Tidak
5	Bagja	80	70	65	71.67	Ya

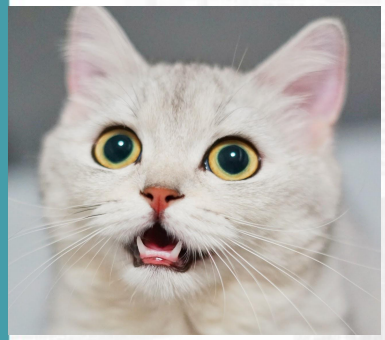
Fitur **Target**

- **Fitur** : Data yang kita diduga berpengaruh terhadap target
- **Target** : Hal yang ingin kita prediksi menggunakan fitur-fitur baru

Note :
ada yang menyebut fitur dan target dengan istilah lain

- Istilah lain untuk fitur : predictor, variabel, dll
- Istilah lain untuk target : label, kelas, dll

Data = Fitur + Target (2)



Fitur

Kucing

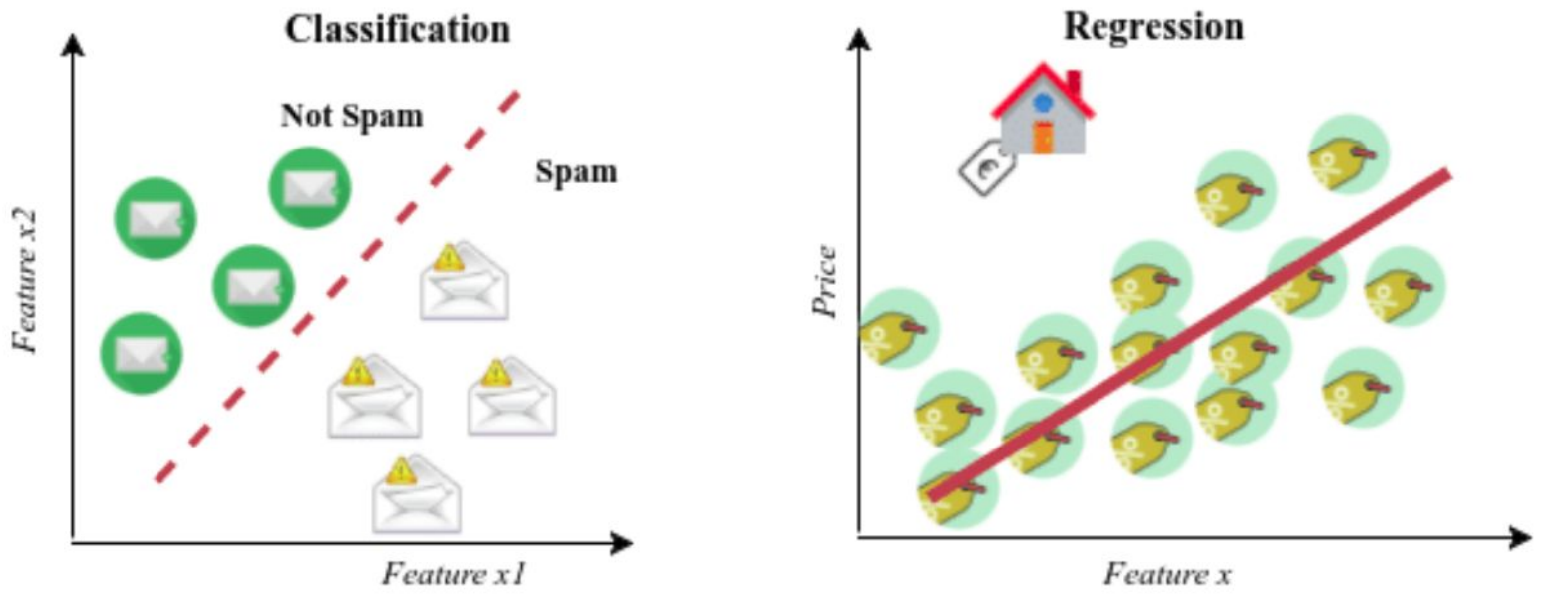
Kucing

Kucing

Bukan Kucing

Target

2 Tipe Supervised Learning



Klasifikasi

Rakamin Academy

Input (Fitur)

Presensi	Nilai Ujian Akhir	Donasi Orangtua
90%	60	1 Milyar
70%	70	0
90%	60	0.5 Milyar
50%	100	1 Milyar
100%	60	0
20%	10	5 Milyar
80%	80	1 Milyar

Kategori

Output (Target)

Lulus
YA
YA
TIDAK
YA
TIDAK
YA
YA

Regresi

Rakamin Academy

Input (Fitur)

Jenis Kelamin	Presensi	Uang Jajan/Hari
Laki-laki	60%	100 Ribu
Laki-laki	80%	50 ribu
Perempuan	100%	80 ribu
Laki-laki	80%	0
Perempuan	40%	0
Perempuan	60%	50 ribu
Perempuan	100%	100 ribu

Angka

Output (Target)

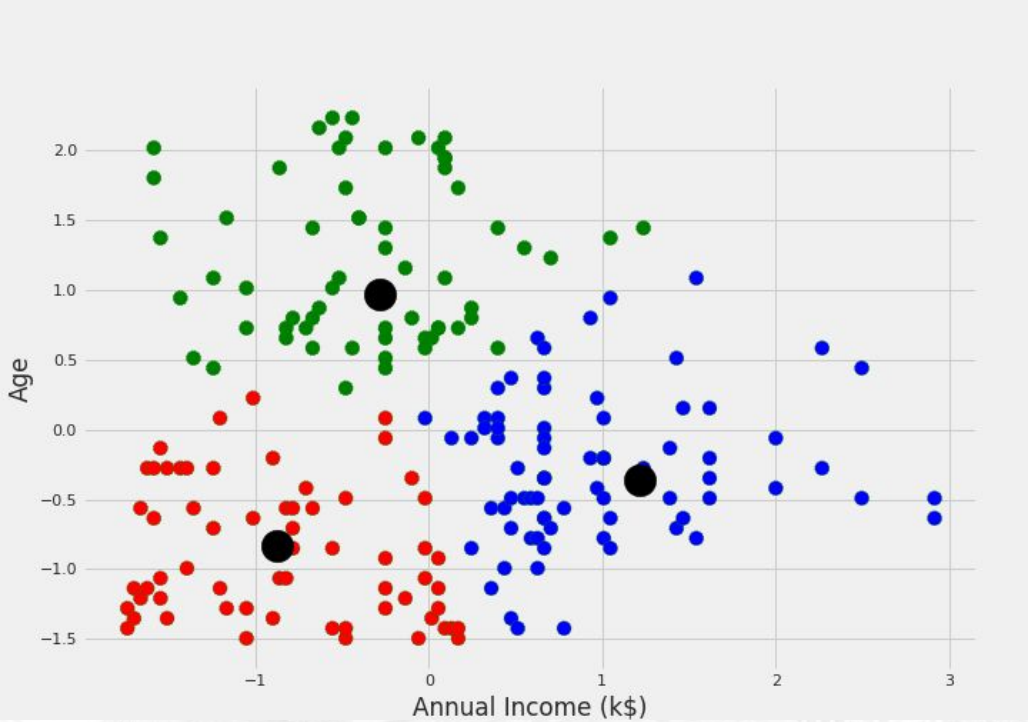
Nilai Ujian
70
85
92
90
70
75
90

Rakamin Academy

Unsupervised Learning

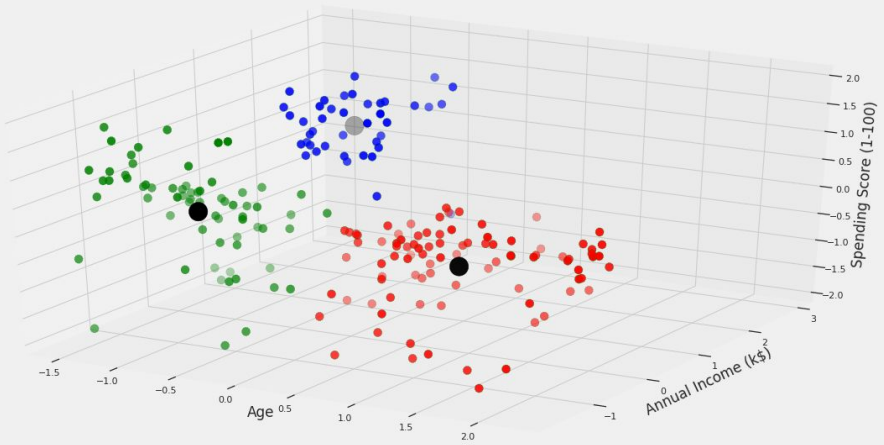
- Tersedia data tanpa target
- Clustering, reduksi dimensi

Clustering 2D



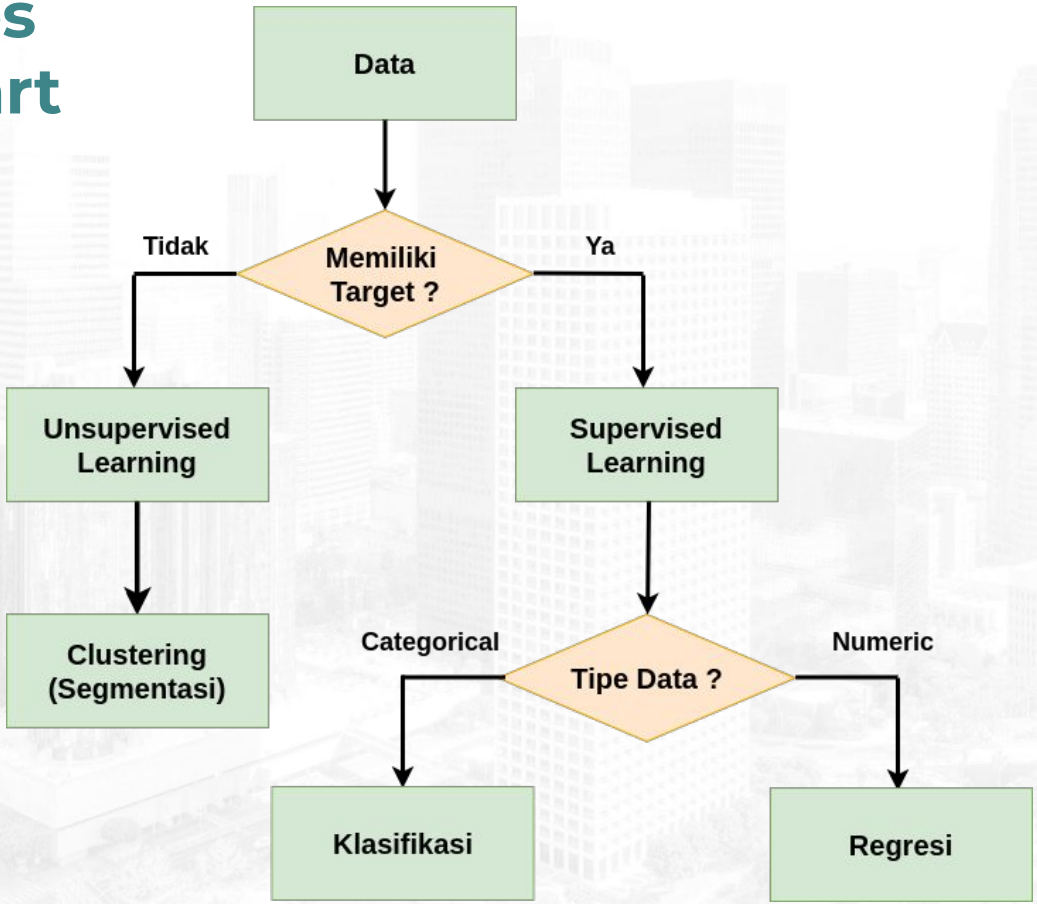
Segmentasi pelanggan berdasarkan data umur & pendapatan per tahun

Clustering 3D



Segmentasi pelanggan berdasarkan data umur, pendapatan per tahun, dan spending score

ML Types Flowchart



Reinforcement Learning

Rakamin Academy

- Trial and error learning pada suatu lingkungan dengan aturan spesifik
- Tujuan: melatih ‘agen’ dalam suatu ‘task’ untuk memaksimalkan ‘reward’

AI Go, Dota 2, etc.

Reinforcement Learning

Outline Pembelajaran

Rakamin Academy

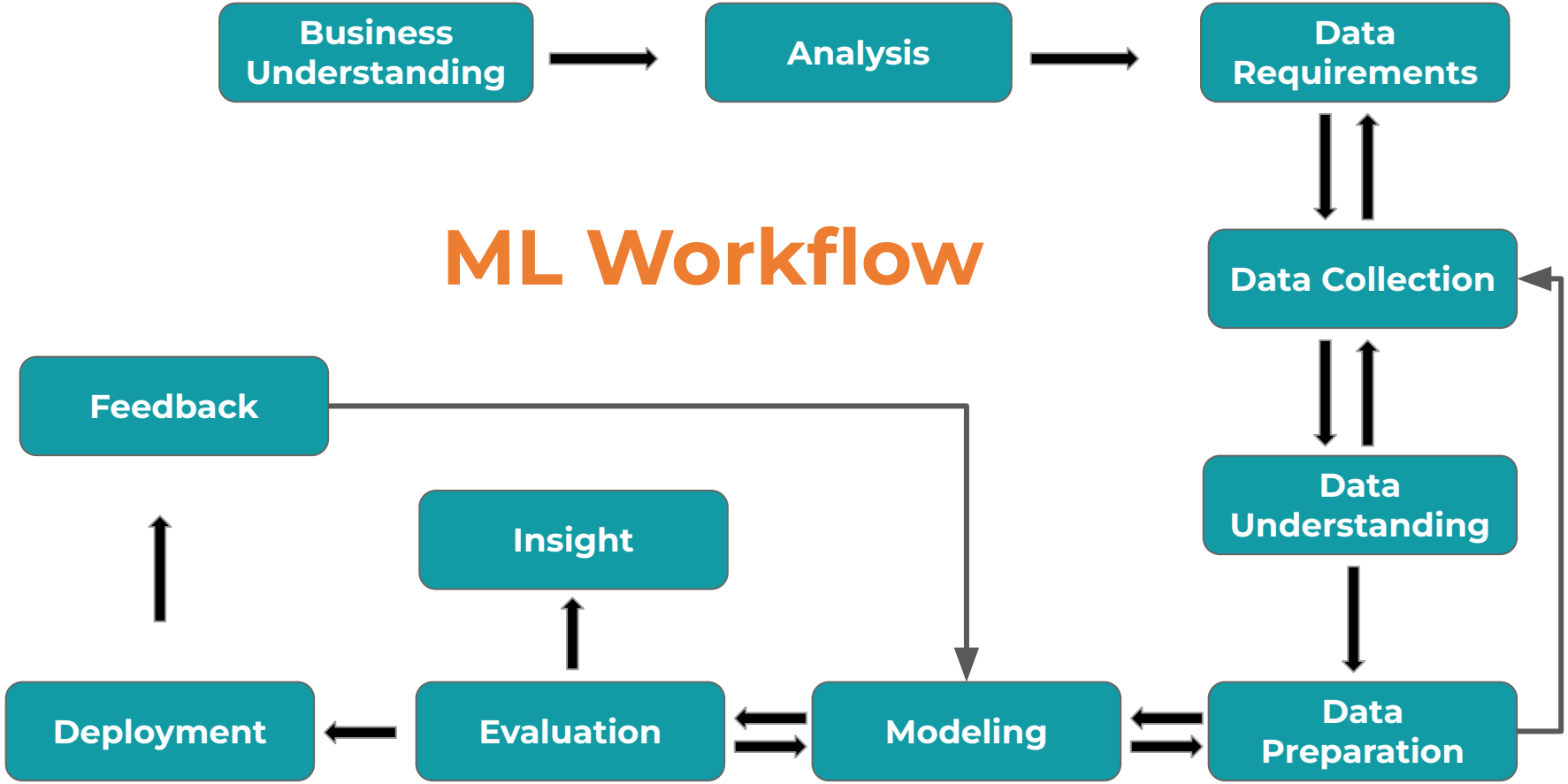
Feature Extraction

- [x] Apa itu Machine Learning?
- [x] Review: Machine Learning dan Data Science
- [x] Tipologi Machine Learning
- [ ] Workflow Machine Learning
- [ ] Kutukan Dimensi
- [ ] Mengolah Feature

Rakamin Academy

Workflow Machine Learning

Step-by-step ML


Business Understanding
Analysis
Data Requirements
ML Workflow
Data Collection
Feedback
Data Understanding
Insight
Deployment
Evaluation
Modeling
Data Preparation

Tahap	Masuk	Proses	Keluar
Data collection	-	• Survey/Labelling • ETL	• Data mentah
Data understanding	• Data mentah	• Exploratory Data Analysis	• Data mentah • Insight
Data preparation	• Data mentah • Insight	• Pre-processing • Feature processing	• Data *training* • Data *test/validation*
Modelling	• Data *training*	• Model training • Hyperparameter tuning	• ML Model
Evaluation	• Data *test/ validation*	• Validation	• Performance measure
Insight	• Data *training & test*	• Model analysis	• feature importance, coef, dll
Deployment	• Data baru	• Prediction	• Prediksi

- **Data mentah:** semua kandidat *feature*, format apapun, kotor, ***feature + label***
- **Data training:** *feature* terpilih, format sesuai algoritma, ***feature + label***
- **Data test/validation:** *feature* terpilih, format sesuai algoritma, ***feature + label***
- **Data baru:** *feature* terpilih, format sesuai algoritma, ***feature* tanpa *label***


ITERATIVE PROCESS

Outline Pembelajaran

Rakamin Academy

Feature Extraction

- [x] Apa itu Machine Learning?
- [x] Review: Machine Learning dan Data Science
- [x] Tipologi Machine Learning
- [x] Workflow Machine Learning
- [ ] Kutukan Dimensi
- [ ] Mengolah Feature

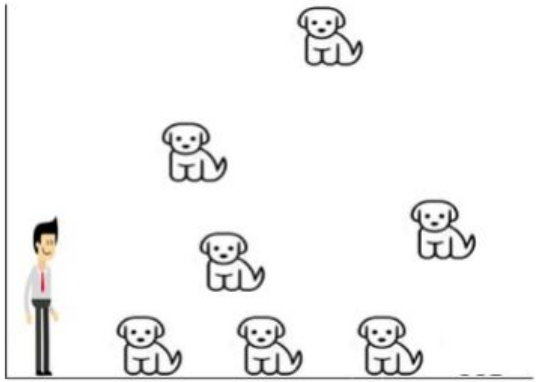
Kutukan Dimensi

(Eng: *Curse of Dimensionality*)

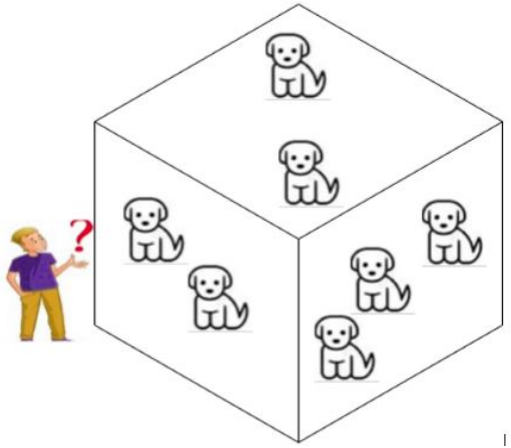
Curse of Dimensionality?

Seiring dengan bertambahnya dimensi ruang fitur, jumlah konfigurasi (kemungkinan kombinasi nilai fitur) yang ada bisa bertambah secara eksponensial dan jumlah konfigurasi yang dicakup pada pengamatan berkurang

1 Dimensi



2 Dimensi



3 Dimensi

Kenapa tidak masukkan saja semua fitur yang ada?

Semakin banyak fitur = semakin banyak informasi = bisa jadi semakin banyak masalah juga

Apa itu "Data Lengkap"?
Semua kemungkinan kombinasi fitur-fitur yang ada

Mata dadu	Cuaca
1	Hujan
2	Hujan
3	Tidak hujan
4	TIdak hujan
5	Tidak hujan
6	Hujan

- **Misal, kita berandai bahwa cuaca ditentukan oleh nilai mata dadu yang dilempar oleh seseorang**
- **Jumlah kemungkinan kombinasi kasus/fitur ada 6**
  - **i.e. dengan mengetahui ke-6 kasus ini, kita dapat menebak dengan pasti kondisi cuaca pada apapun data baru yang muncul (data lengkap)**

Muka uang logam	Mata dadu	Cuaca
Angka	1	Hujan
Angka	2	Hujan
Angka	3	Tidak hujan
Angka	4	TIdak hujan
Angka	5	Tidak hujan
Angka	6	Hujan
Garuda	1	Tidak hujan
Garuda	2	Hujan
Garuda	3	Tidak hujan
Garuda	4	Hujan
Garuda	5	Tidak hujan
Garuda	6	Hujan

Jika ternyata cuaca ditentukan oleh mata dadu & muka uang logam, maka kita perlu 6x2=12 kasus yang berbeda agar data kita lengkap.

Ilustrasi Curse of Dimensionality: Kombinasi fitur yang mungkin tumbuh eksponensial

	Nilai Ujian Akhir (0-100) Kelipatan 10	Donasi Orangtua (0-10 M) Kelipatan 0.5M	Lulus
	60	1 Milyar	YA
	70	0	YA
	60	0.5 Milyar	TIDAK
	100	1 Milyar	YA
	60	0	TIDAK
	10	5 Milyar	YA
	80	1 Milyar	YA

- **Kombinasi fitur: 11* 21 = 231**
- **Hanya perlu 231 baris data untuk mengetahui dengan pasti pola untuk menentukan kelulusan**

Ilustrasi Curse of Dimensionality: Kombinasi fitur yang mungkin tumbuh eksponensial

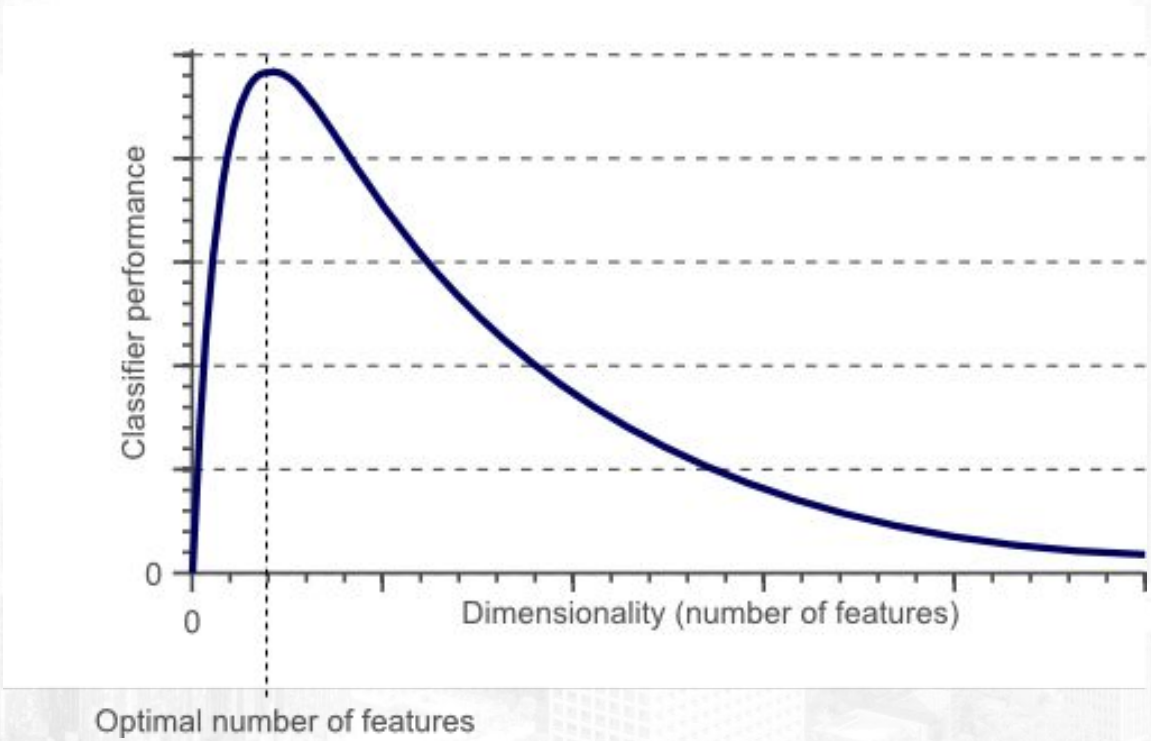
Presensi (0-100%) Kelipatan 10%	Nilai Ujian Akhir (0-100) Kelipatan 10	Donasi Orangtua (0-10 M) Kelipatan 0.5M	Lulus
90%	60	1 Milyar	YA
70%	70	0	YA
90%	60	0.5 Milyar	TIDAK
50%	100	1 Milyar	YA
100%	60	0	TIDAK
20%	10	5 Milyar	YA
80%	80	1 Milyar	YA

- **Kombinasi fitur: 11 * 11 * 21 = 2541**
- **Perlu 2541 baris data**

Bagaimana dengan 4 fitur? 5? 10? 20?

Beberapa hal penting untuk diingat:

- Semakin 'lengkap' data kita (+baris) -> semakin bagus *performance model*
- Semakin banyak fitur -> semakin banyak data yang diperlukan agar 'lengkap'
- Karena data kita biasanya terbatas (konstan), **ada titik dimana menambahkan fitur malah mengurangi *performance***



Pesan utama

- **Menambah lebih banyak fitur pada model machine learning tak selalu meningkatkan performa model**

Outline Pembelajaran

Rakamin Academy

Feature Extraction

- [x] Apa itu Machine Learning?
- [x] Review: Machine Learning dan Data Science
- [x] Tipologi Machine Learning
- [x] Workflow Machine Learning
- [x] Kutukan Dimensi
- [ ] Mengolah Feature

Rakamin Academy

Mengolah Feature

Ekstraksi Fitur

Rakamin Academy

Feature?

Selection

- Membuang feature-feature yang tidak relevan
- Membuang feature-feature yang redundant
- **Mengurangi fitur**

Extraction/Engineering

- Membuat fitur baru dari fitur-fitur yang ada
- Mengambil saripati dari fitur-fitur yang ada
- **Menambah fitur**

Transformation

- Mengubah fitur ke dalam bentuk yang lebih mudah dipakai oleh algoritma/model

Dimensionality Reduction

- Mereduksi dimensi fitur ke dalam dimensi yang lebih rendah

Feature Selection

Presensi	Nilai Ujian	Donasi	# Skors	Prestasi	Beasiswa	...	...	Lulus
90%	60	1 Milyar	4 Hari	...	...	...	...	YA
70%	70	0	2 Hari	...	...	...	...	YA
90%	60	0.5 Milyar	0	...	...	...	...	TIDAK
50%	100	1 Milyar	99 Hari	...	...	...	...	YA
100%	60	0	20 Hari	...	...	...	...	TIDAK
20%	10	5 Milyar	0	...	...	...	...	YA
80%	80	1 Milyar	0	...	...	...	...	YA
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
20%	10	5 Milyar	0	...	...	...	...	YA
80%	80	1 Milyar	0	...	...	...	...	YA

Feature Selection

		Donasi						Lulus
		1 Milyar						YA
		0						YA
		0.5 Milyar						TIDAK
		1 Milyar						YA
		0						TIDAK
		5 Milyar						YA
		1 Milyar						YA
		...						...
		...						...
		...						...
		5 Milyar						YA
		1 Milyar						YA

Feature Extraction/Engineering

Nilai Ujian 1	Nilai Ujian 2	Nilai Ujian 3	Nilai Ujian 4	...	...	...	Nilai Ujian 100	Lulus
50	60	20	60	...	...	...	0	YA
90	70	100	70	...	...	...	0	YA
60	60	60	60	...	...	...	0	TIDAK
100	20	10	100	...	...	...	0	YA
60	100	80	60	...	...	...	0	TIDAK
10	60	10	100	...	...	...	0	YA
80	80	80	60	...	...	...	0	YA
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
10	50	10	40	...	...	...	0	YA
60	30	10	80	...	...	...	0	YA

Feature Extraction/Engineering: Agregasi

Rata-rata Nilai	Nilai Ujian Maksimum	Nilai Ujian Minimum	Lulus
50	90	0	YA
70	80	0	YA
60	60	0	TIDAK
10	20	0	YA
60	100	0	TIDAK
70	90	0	YA
40	70	0	YA
...	...	...	...
...	...	...	...
...	...	...	...
50	60	0	YA
60	90	0	YA

Feature Extraction/Engineering: Rescaling

Nilai Ujian	Donasi	Umur	Uang Jajan
50	1M	13	200rb
90	3M	12	300rb
60	5M	15	100rb
100	2M	14	500rb
60	1M	11	250rb
10	0	20	1 jt
...	...	...	...
...	...	...	...
...	...	...	...
10	2M	10	400rb

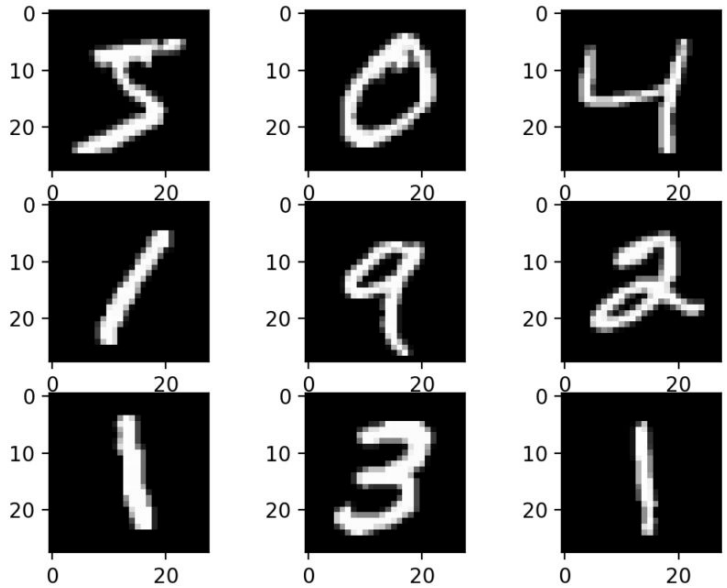
Nilai Ujian	Donasi	Umur	Uang Jajan
0.5	0.1	0.55	0.2
0.9	0.3	0.4	0.3
0.6	0.5	0.6	0.1
1	0.2	0.575	0.5
0.6	0.1	0.35	0.25
0.1	0	1	1
...	...	...	...
...	...	...	...
...	...	...	...
0.1	0.2	0.3	0.4

Feature Extraction/Engineering: Datetime extraction

Datetime
2021-10-01 09:00:00
2021-10-01 10:00:00
2021-10-01 11:00:00
2021-10-01 12:00:00
2021-10-02 09:00:00
2021-10-02 10:00:00
...
...
...
2021-11-17 12:00:00

Month	Date	Day Name	Is Weekend	Hour
10	1	Friday	No	9
10	1	Friday	No	10
10	1	Friday	No	11
10	1	Friday	No	12
10	2	Saturday	Yes	9
10	2	Saturday	Yes	10
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
11	17	Wednesday	No	12

Feature Extraction/Engineering: Image



# Pixel Putih	# Pixel Hitam	Label
...	210	1
...	455	5
...	334	3
...	300	0
...	400	2
...	260	7
...	444	9
...	...	...
...	...	...
...	...	...
...	423	6
...	588	8

Outline Pembelajaran

Rakamin Academy

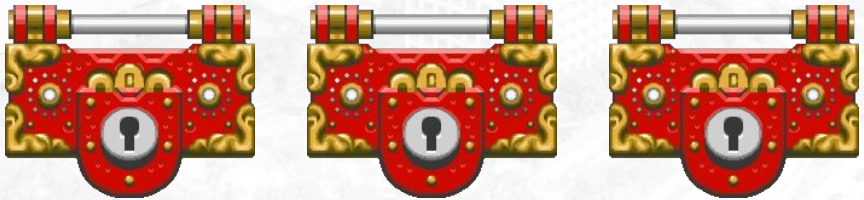
Feature Extraction

- Apa itu Machine Learning?
- Review: Machine Learning dan Data Science
- Tipologi Machine Learning
- Workflow Machine Learning
- Kutukan Dimensi
- Mengolah Feature

Challenge #1!
Fitur apa saja yang bisa diturunkan dari NIK?

Sampel NIK

3 5 1 5 1 3 1 1 0 4 6 2 0 0 0 8



Challenge #1!
Fitur apa saja yang bisa diturunkan dari NIK?

Challenge #2!
Dari feature-feature di bawah, sebutkan 3 contoh pengolahannya!

Aviation Accidents and Incidents (NTSB, FAA, WAAS) | Kaggle

Feature Name	Description	Feature Name	Description
Event Id	Unique NTSB ID for event	Airport Code	Code for nearest airport
Investigation Type	Type of investigation (accident/incident)	Airport Name	Full name of nearest airport
Accident Number	Unique ID for accident	Injury Severity	Severity of injury in event
Event Date	Date of event	Aircraft Damage	Severity of damage to involved aircraft
Location	Location where event occurred	Aircraft Category	Category of aircraft involved
Country	Country where even occurred	Reg. Number	Registration number aircraft involved
Latitude	Latitude where event occurred	Make	Make of aircraft involved
Longitude	Longitude where event occurred	etc.	

Rakamin Academy

Bagaimana kita tahu fitur mana yang penting dan yang mana yang perlu dibuang?
Bagaimana cara mentransformasi data yang baik?

Tunggu sesi-sesi selanjutnya!


Rakamin
Academy
Thank You