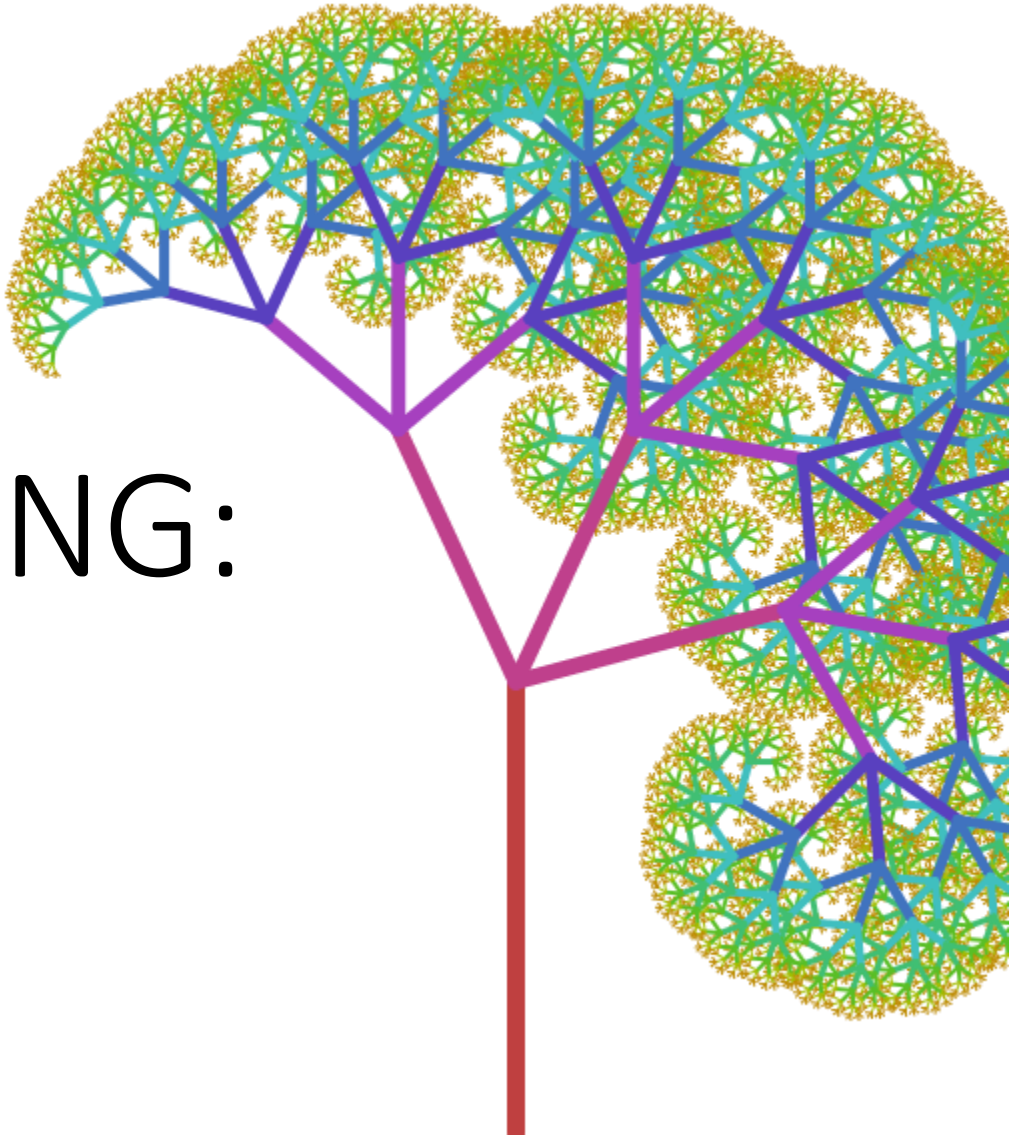
SUPERVISED LEARNING:

NAÏVE BAYES



Tujuan Pembelajaran

- Mahasiswa mampu memahami konsep klasifikasi menggunakan metode Naïve Bayes
- Mahasiswa mampu mengimplementasikan metode Naïve Bayes untuk menyelesaikan permasalahan klasifikasi

Topik Pembahasan

- Dasar Teorema Bayes
- Klasifikasi Naïve Bayes
- Contoh Implementasi Naïve Bayes
- Contoh Penelitian Naïve Bayes
- Kesimpulan

Dasar Teorema Bayes

- Diketahui X merupakan sample data.
  - Dalam bayes X disebut “evidence” atau fakta
  - Label class tidak diketahui
  - Umumnya X merupakan record data yang disusun dari n atribut

- H merupakan suatu hypothesis bahwa X termasuk dari kelas C

- Classification adalah untuk menentukan P(H|X), peluang hipotesis dari data sample X
  - Dengan kata lain, dicari peluang bahwa record X termasuk kelas C, dengan diketahui atribut yang menjelaskan X.
  - Atau, peluang keluarnya hasil H jika diketahui nilai X tertentu.

Dasar Teorema Bayes

- P(H) (prior probability), peluang awal
  - Misal: X akan membeli komputer, tanpa memperhatikan umur, penghasilan, …
- P(X) (prior probabilitiy) peluang bahwa data sampel X diamati tanpa memperhatikan nilai yang lain
- P(X|H) (posteriori probability), peluang diamatinya data sampel X dengan mempertimbangkan H
  - Misal: Jika X akan membeli komputer, peluang X adalah berumur 31..40, medium income

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Dasar Teorema Bayes

- Jika diberikan data training X, posteriori probability dari suatu hypothesis H, P(H|X), mengikuti teorema Bayes

$$P(H|\mathbf{X}) = \frac{P(\mathbf{X}|H)P(H)}{P(\mathbf{X})}$$

- Dengan kata lain, dapat ditulis sebagai berikut:

  posteriori = likelihood x prior/evidence

- Memperkirakan X termasuk dalam kelas Ci jika peluang P(Ci|X) merupakan tertinggi diantara semua P(C*k*|X) untuk semua klas *k*
- Permasalahan nyata: diperlukan pengetahuan awal dari banyak peluang, hal ini dapat merupakan biaya komputasi yang mencolok

Dasar Teorema BayesLearning

- Misal terdapat beberapa alternatif hipotesa h -> h ϵ H.
- Bayes Learning:
  - Memaksimalkan hipotesis yang paling mungkin h, maksimum apriori (MAP)

$$h_{MAP} = \arg\max P(h \mid x)$$
$$= \arg\max \frac{P(x \mid h)P(h)}{P(x)}$$
$$= \arg\max P(x \mid h)P(h)$$

Klasifikasi Naïve Bayes

- Simple Naïve Bayesian Classifier (NBC) merupakan salah satu metode pengklasifikasi berpeluang sederhana yang berdasarkan pada penerapan Teorema Bayes dengan asumsi antar variabel penjelas saling bebas (independen).
- Algoritma ini memanfaatkan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi probabilitas di masa depan berdasarkan pengalaman di masa sebelumnya.

Contoh Kasus 1

- Misalnya terdapat ingin diketahui apakah suatu objek masuk dalam ketegori dipilih untuk perumahan atau tidak dengan algoritma Naive Bayes Classifier. Untuk menetapkan suatu daerah akan dipilih sebagai lokasi untuk mendirikan perumahan, telah dihimpun 10 data.
- Ada 4 atribut yang digunakan, yaitu:
  - harga tanah per meter persegi (C1),
  - jarak daerah tersebut dari pusat kota (C2),
  - ada atau tidaknya angkutan umum di daerah tersebut (C3), dan
  - keputusan untuk memilih daerah tersebut sebagai lokasi perumahan (C4).

Contoh Kasus 1: Dataset

Atribut / Parameter

Label / Kelas

Aturan ke-	Harga tanah (C1)	Jarak dari pusat kota (C2)	Ada angkutan umum (C3)	Dipilih untuk perumahan (C4)
1	Murah	Dekat	Tidak	Ya
2	Sedang	Dekat	Tidak	Ya
3	Mahal	Dekat	Tidak	Ya
4	Mahal	Jauh	Tidak	Tidak
5	Mahal	Sedang	Tidak	Tidak
6	Sedang	Jauh	Ada	Tidak
7	Murah	Jauh	Ada	Tidak
8	Murah	Sedang	Tidak	Ya
9	Mahal	Jauh	Ada	Tidak
10	Sedang	Sedang	Ada	Ya

Probabilitas Atribut C1

- Probabilitas kemunculan setiap nilai untuk atribut Harga Tanah (C1)

Harga tanah	Jumlah kejadian "Dipilih"		Probabilitas	
	Ya	Tidak	Ya	Tidak
Murah	2	1	2/5	1/5
Sedang	2	1	2/5	1/5
Mahal	1	3	1/5	3/5
Jumlah	*5*	*5*	*1*	*1*

Probabilitas Atribut C2

- Probabilitas kemunculan setiap nilai untuk atribut Jarak dari Pusat Kota (C2)

Jarak Dari Pusat Kota	Jumlah kejadian "Dipilih"		Probabilitas	
	Ya	Tidak	Ya	Tidak
Dekat	3	0	3/5	0
Sedang	2	1	2/5	1/5
Jauh	0	4	0	4/5
Jumlah	*5*	*5*	*1*	*1*

Probabilitas Atribut C3

- Probabilitas kemunculan setiap nilai untuk atribut Ada Angkutan Umum (C3)

Angkutan Umum	Jumlah kejadian "Dipilih"		Probabilitas	
	Ya	Tidak	Ya	Tidak
Ada	1	3	1/5	3/5
Tidak	4	2	4/5	2/5
Jumlah	*5*	*5*	*1*	*1*

Probabilitas Atribut C4

- Probabilitas kemunculan setiap nilai untuk atribut Dipilih untuk perumahan (C4)

Dipilih Untuk Perumahan	Jumlah kejadian "Dipilih"		Probabilitas	
	Ya	Tidak	Ya	Tidak
Jumlah	*5*	*5*	*1/2*	*1/2*

5/10 *5/10*

Klasifikasi Data Uji

- Berdasarkan data tersebut, apabila diketahui suatu daerah dengan harga tanah (C1)= **MAHAL**, jarak dari pusat kota (C2)= **SEDANG**, dan angkutan umum (C3) = **ADA**, maka apakah daerah tersebut dapat dipilih untuk perumahan atau tidak?

Klasifikasi Data Uji

YA = P(Ya|Tanah=MAHAL) . P(Ya|Jarak=SEDANG) .
P(Ya|Angkutan=ADA) . P(Ya)
= 1/5 x 2/5 x 1/5 x 5/10 = 10/1250 = 0,008

TIDAK=P(Tidak| Tanah=MAHAL) . P(Tidak|Jarak=SEDANG) .
P(Tidak|Angkutan=ADA) . P(Tidak)
= 3/5 x 1/5 x 3/5 x 5/10 = 45/1250 = 0,036

Klasifikasi Data Uji

- Nilai probabilitas dapat dihitung dengan melakukan normalisasi terhadap likelihood tersebut sehingga jumlah nilai yang diperoleh = 1

Probabilitas Ya = $\frac{0{,}008}{0{,}008 + 0{,}036} = 0{,}182.$

Probabilitas Tidak = $\frac{0{,}036}{0{,}008 + 0{,}036} = 0{,}818.$

Klasifikasi : TIDAK

Karena probabilitas **TIDAK (0,818)** lebih besar daripada **YA (0,182),** maka tanah tersebut TIDAK cocok dijadikan perumahan

Contoh Kasus 2

- Untuk jenis data harga tanah dan jarak pusat kota yang kontinue, misalnya :

Aturan ke-	Harga tanah (C1)	Jarak dari pusat kota (C2)	Ada angkutan umum (C3)	Dipilih untuk perumahan (C4)
1	100	2	Tidak	Ya
2	200	1	Tidak	Ya
3	500	3	Tidak	Ya
4	600	20	Tidak	Tidak
5	550	8	Tidak	Tidak
6	250	25	Ada	Tidak
7	75	15	Ada	Tidak
8	80	10	Tidak	Ya
9	700	18	Ada	Tidak
10	180	8	Ada	Ya

Contoh Kasus 2

- Namun jika atribut ke-i bersifat **kontinu**, maka P(xi|C) diestimasi dengan fungsi **densitas Gauss**.
- Distribusi normal adalah distribusi dari variabel acak kontinu. Kadang-kadang distribusi normal disebut juga dengan distribusi Gauss. Distribusi ini merupakan distribusi yang paling penting dan paling banyak digunakan di bidang statistika.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Probabilitas Atribut C1

- Probabilitas kemunculan setiap nilai untuk atribut Harga Tanah (C1)

	Ya	Tidak
1	100	600
2	200	550
3	500	250
4	80	75
5	180	700
Mean (μ) Deviasi standar (σ)	212 168,8787	435 261,9637

Probabilitas Atribut C2

- Probabilitas kemunculan setiap nilai untuk atribut Jarak dari Pusat Kota (C2)

	Ya	Tidak
1	2	20
2	1	8
3	3	25
4	10	15
5	8	18
Mean (μ) Deviasi standar (σ)	4,8 3,9623	17,2 6,3008

Probabilitas Atribut C3

- Probabilitas kemunculan setiap nilai untuk atribut Angkutan Umum (C3)

Angkutan Umum	Jumlah kejadian "Dipilih"		Probabilitas	
	Ya	Tidak	Ya	Tidak
Ada	1	3	1/5	3/5
Tidak	4	2	4/5	2/5
Jumlah	*5*	*5*	*1*	*1*

Probabilitas Atribut C4

- Probabilitas kemunculan setiap nilai untuk atribut Dipilih untuk Perumahan (C4)

Dipilih Untuk Perumahan	Jumlah kejadian "Dipilih"		Probabilitas	
	Ya	Tidak	Ya	Tidak
Jumlah	*5*	*5*	*1/2*	*1/2*

5/10 *5/10*

Klasifikasi

- Apabila diberikan C1 = 300, C2 = 17, C3 = Tidak, maka:

$$f(C1 = 300|ya) = \frac{1}{\sqrt{2\pi(168.8787)}} e \frac{-(300-212)^2}{2(168.8787)^2} = 0.02869$$

$$f(C1 = 300|tidak) = \frac{1}{\sqrt{2\pi(261.9637)}} e \frac{-(300-435)^2}{2(261.9637)^2} = 0.02307$$

$$f(C2 = 17|ya) = \frac{1}{\sqrt{2\pi(3.9623)}} e \frac{-(17-4.8)^2}{2(3.9623)^2} = 0.01874$$

$$f(C2 = 17|tidak) = \frac{1}{\sqrt{2\pi(6.3008)}} e \frac{-(17-17.2)^2}{2(6.3008)^2} = 0.15893$$

$(\text{C3}=\text{Tidak}|\ \text{Ya}) = 4/5$ dan $(\text{C3} = \text{Tidak}|\text{Tidak} = 2/5)$

$(\text{C4} = \text{Ya} = 5/10$ dan $\text{C4} = \text{Tidak} = 5/10$

Ingat
$\pi = 3.14$

Rumus Excel (C1 dan C2)

- C1(300|Ya) =1/SQRT(2*3.14*168.8787)*EXP(-(((300-212)^2/(2*168.8787)^2)))
- C1(300|Tidak) = 1/SQRT(2*3.14*261.9637)*EXP(-(((300-435)^2/(2*261.9637)^2)))

- C2 (17|Ya) = 1/SQRT(2*3.14*3.9623)*EXP(-(((17-4.8)^2/(2*3.9623)^2)))
- C2 (17|Tidak) = 1/SQRT(2*3.14*6.3008)*EXP(-(((17-17.2)^2/(2*6.3008)^2)))

Klasifikasi

$$\text{Likelihood Ya} = (0.02869)\ \text{x}\ (0.01874)\ \text{x}\ 4/5\ \text{x}\ 5/10 \qquad = 0.000215$$

$$\text{Likelihood Tidak} = (0.02307)\ \text{x}\ (0.15893)\ \text{x}\ 2/5\ \text{x}\ 5/10 \quad = 0.000733$$

- Nilai probabilitas dapat dihitung dengan melakukan normalisasi terhadap likelihood tersebut sehingga jumlah nilai yang diperoleh = 1

$$\text{Probabilitas Ya} = \frac{0.000215}{0.000215 + 0.000733} = 0.226752$$

$$\text{Probabilitas Tidak} = \frac{0.000733}{0.000215 + 0.000733} = 0.773248$$

Klasifikasi Tidak

Karena probabilitas **TIDAK (0,7732)** lebih besar daripada **YA (0,2268),** maka tanah tersebut **TIDAK** cocok dijadikan perumahan

Penelitian Naïve Bayes

- Ahmad Zainul Mafakhir, Achmad Solichin. 2020. Penerapan Metode Naïve Bayes Classifier Untuk Penjurusan Siswa Pada Madrasah Aliyah Al-Falah Jakarta. Fountain Informatics Journals (FIJ), Vol. 5, No. 1, Hal. 21-26, 2020. ISSN 2541-4313 (Print). DOI: http://dx.doi.org/10.21111/fij.v5i1.4007
- Achmad Solichin. 2019. Comparison of Decision Tree, Naïve Bayes and K-Nearest Neighbors for Predicting Thesis Graduation. Proceeding of The 6th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI 2019). https://doi.org/10.23919/EECSI48112.2019.8977081
- Supardi Salmu, Achmad Solichin. Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu Menggunakan Naive Bayes: Studi Kasus UIN Syarif Hidayatullah Jakarta. Prosiding Seminar Nasional Multidisiplin Ilmu (SENMI) 2017. Universitas Budi Luhur, Jakarta, 22 April 2017. ISSN: 2087-0930, hal 701-709.
- Imam Riadi, Rusydi Umar, Fadhilah Dhinur Aini. 2019. Analisis Perbandingan Detection Traffic Anomaly Dengan Metode Naive Bayes Dan Support Vector Machine (Svm). ILKOM Jurnal Ilmiah, 11(1), 17-24. doi:https://doi.org/10.33096/ilkom.v11i1.361.17-24
- Safitri Juanita. 2019. Analisis Sentimen Persepsi Masyarakat Terhadap Pemilu 2019 Pada Media Sosial Twitter Menggunakan Naive Bayes. Jurnal Media Informatika Budidarma. DOI: http://dx.doi.org/10.30865/mib.v4i3.2140

Kesimpulan

- A statistical classifier:
  - menyelesaikan prediksi probabilitas, sebagai contoh memprediksi peluang keanggotaan suatu class

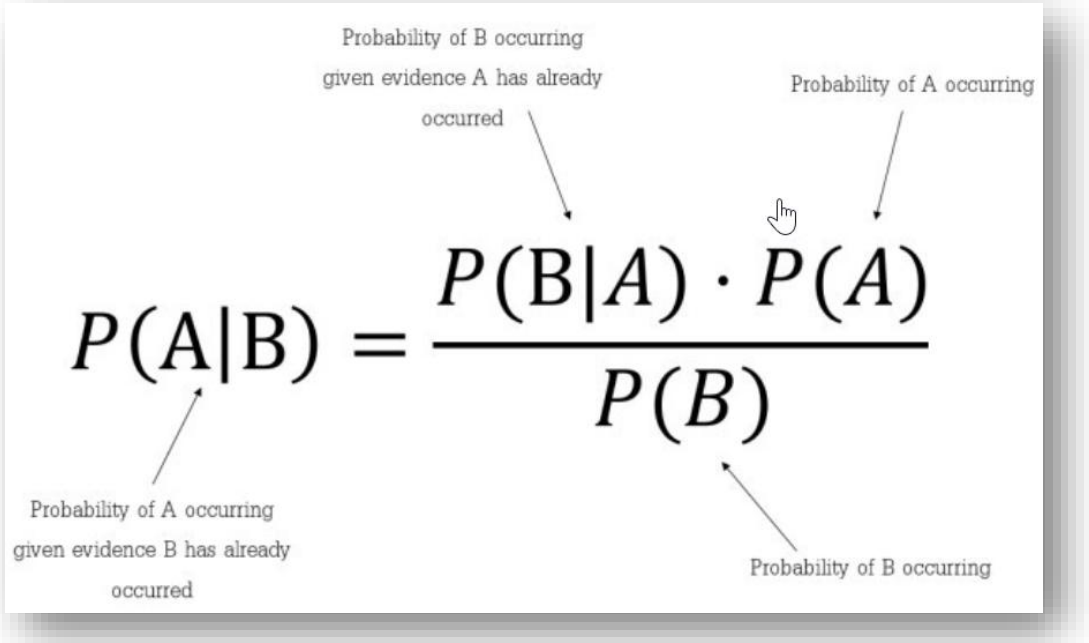
- Foundation:
  - Teorema Bayes

- Performance:
  - pengklasifikasi Bayesian sederhana, memiliki kinerja yang dapat dibandingkan pengklasifikasi *decision tree* dan *neural network*

Kesimpulan

Naïve Bayes Classifier

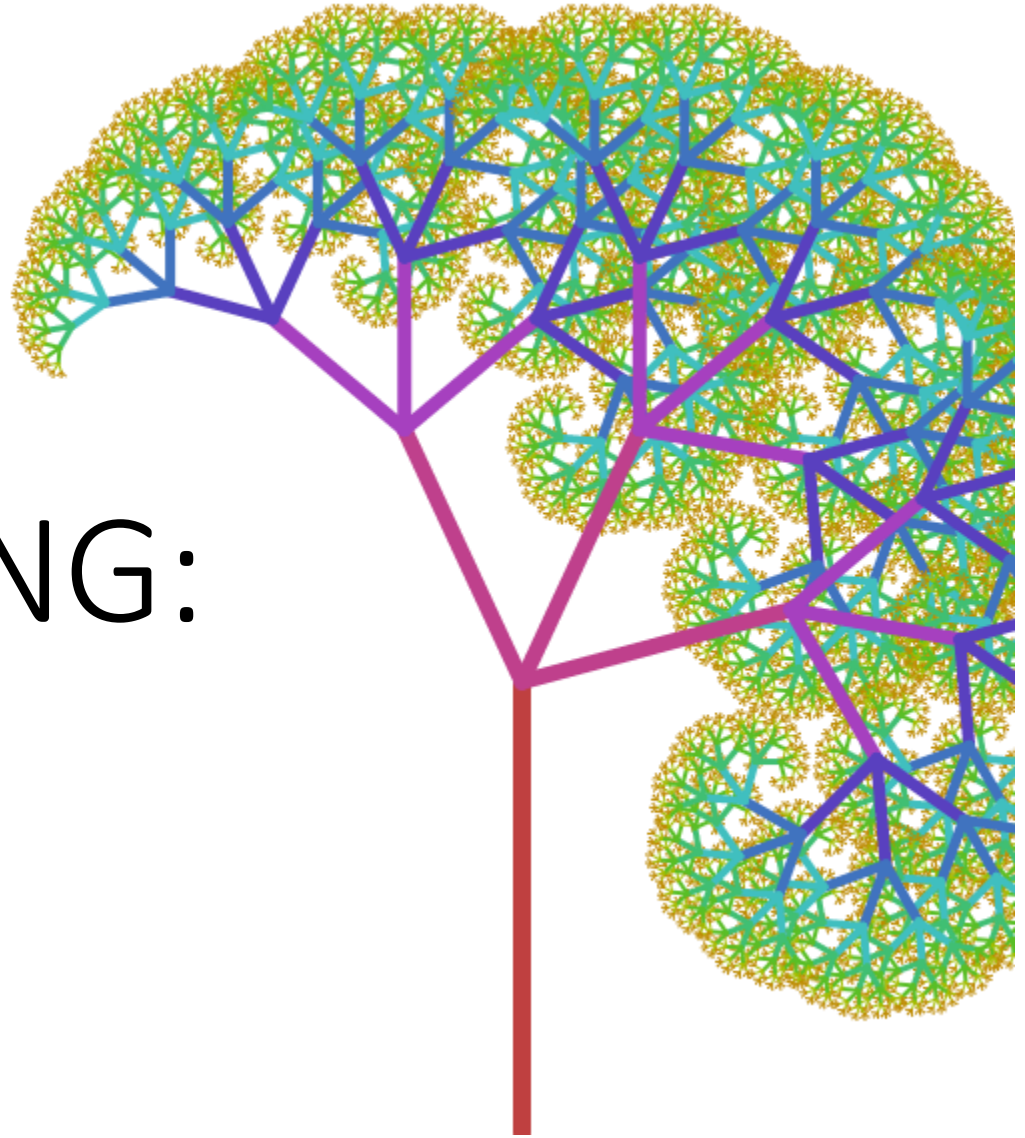
- **Statistical** classifier
- Berdasarkan teorema **Bayesian** (peluang keanggotaan setiap kelas)
- **Sederhana**, tanpa modeling
- **Kinerja baik**, untuk jumlah data terbatas maupun banyak.



Penjelasan selengkapnya bisa simak di video:
https://youtu.be/CJZN51tudS0

SUPERVISED LEARNING:

DECISION TREE



Tugas Naïve Bayes

TUGAS / LATIHAN

Mata Kuliah : **Penambangan Data**

Pokok Bahasan : **Naïve Bayes Classifier**

Petunjuk: untuk dataset di bawah ini, terapkan metode **Naïve Bayes** untuk memprediksikan kelas (class) dari data uji (data yang baru).

Studi Kasus 1. Penentuan Lokasi SPBU

Jalan	Lebar Jalan	Volume Kendaraan	Jumlah Pesaing	Jumlah Pemukiman	Lokasi Strategis
Jl. Pahlawan	Jalan 2 mobil	Sedang	1	<3000	Tidak
Jl. Berduri	Jalan tol	Ramai	>1	3000-5000	Ya
Jl. Insiden	Jalan 4 mobil	Ramai	0	<3000	Ya
Jl. Deklarasi	Jalan 4 mobil	Sepi	1	<3000	Tidak
Jl. Bangkok	Jalan 2 mobil	Ramai	0	>5000	Ya
Jl. Harapan	Jalan tol	Sepi	>1	<3000	Tidak
Jl. Marzuki	Jalan 4 mobil	Sedang	0	3000-5000	Ya
Jl. Denpasar	Jalan 2 mobil	Sepi	0	<3000	Tidak
Jl. H.Soleh	Jalan 4 mobil	Ramai	0	<3000	Ya
Jl. M.Said	Jalan tol	Sepi	1	>5000	Ya

Jalan	Lebar Jalan	Volume Kendaraan	Jumlah Pesaing	Jumlah Pemukiman	Lokasi Strategis
Jl. Merah Putih	Jalan tol	Sepi	>1	<3000	????

Tugas Naïve Bayes

Studi Kasus 2. Playing Tennis

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

A new day				
outlook	temperature	humidity	windy	play
sunny	cool	high	true	?

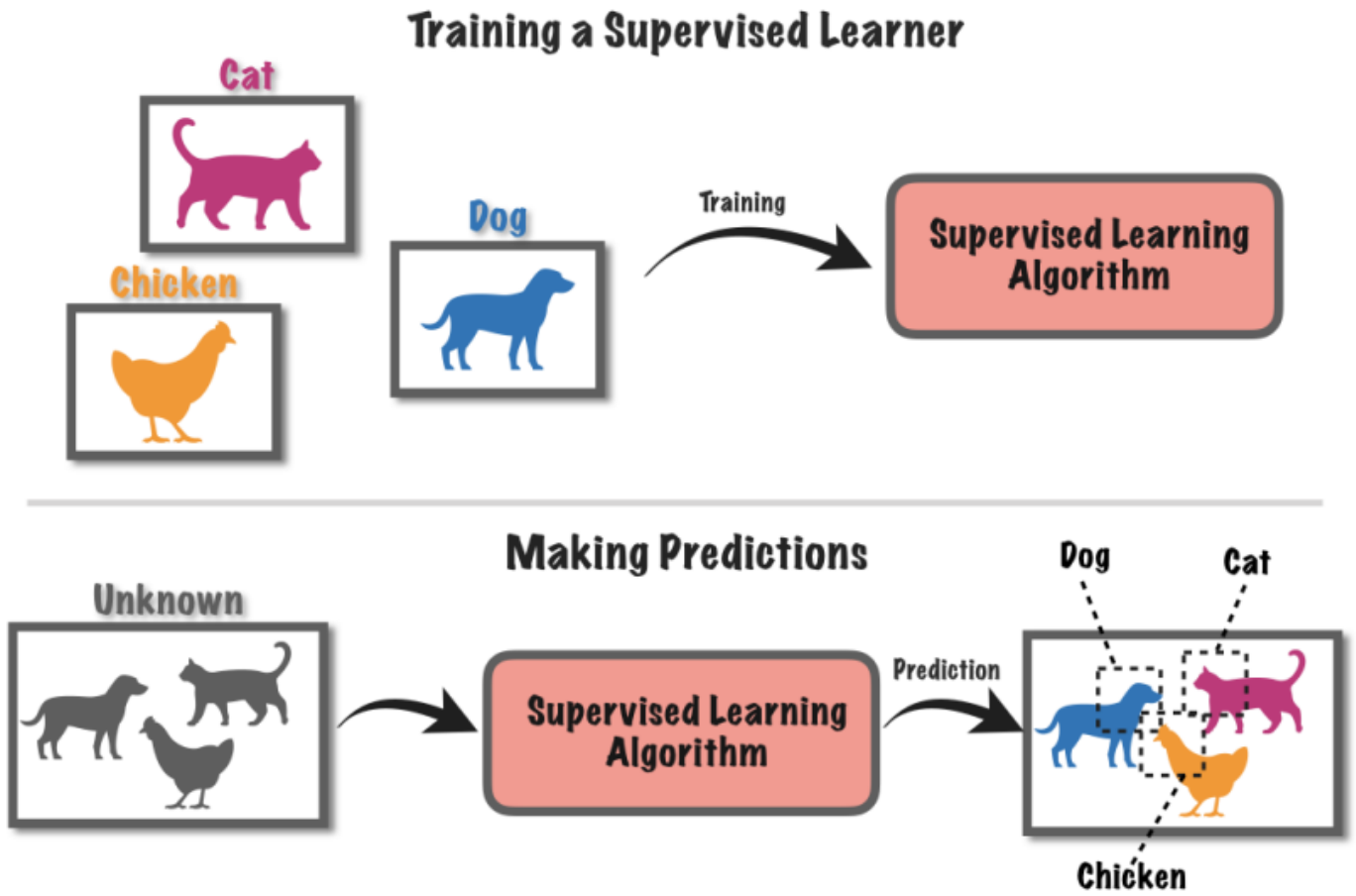
Tujuan Pembelajaran

- Mahasiswa dapat memahami perbedaan konsep pembelajaran tersupervisi (supervised learning) dan pembelajaran tidak tersupervisi (unsupervised learning).
- Mahasiswa dapat memahami konsep pembelajaran (learning) tersupervisi, khususnya menggunakan metode Decision Tree (DT)

Topik Pembahasan

- ❑ Pengantar
- ❑ Supervised vs Unsupervised Learning
- ❑ Decision Tree
- ❑ Pembentukan Tree dalam Decision Tree
- ❑ Latihan

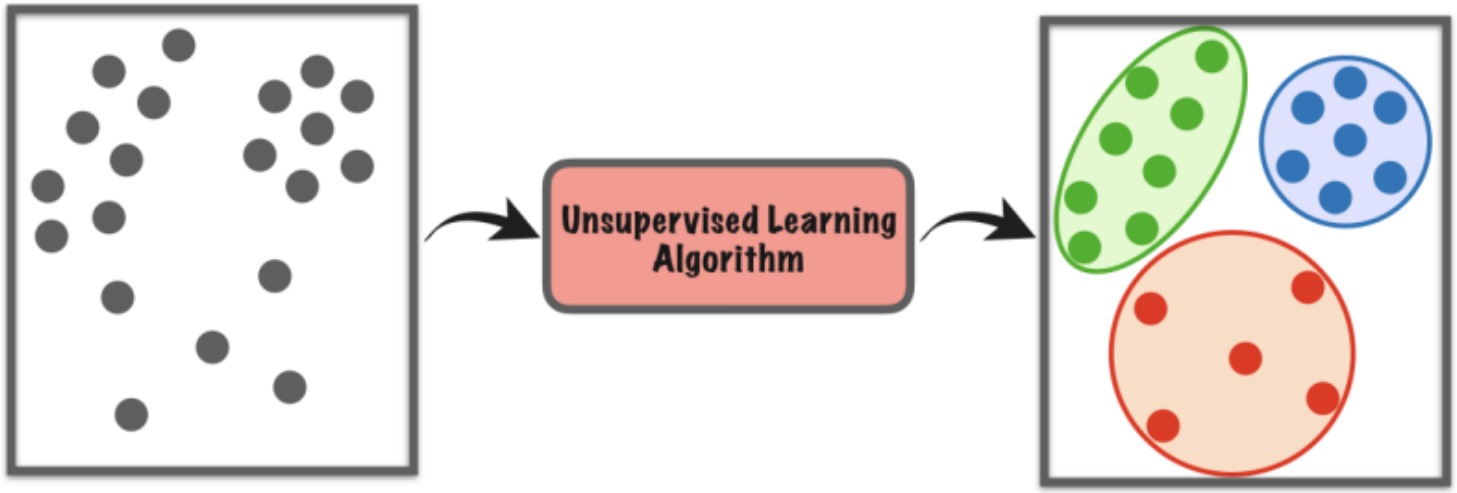
Supervised vs Unsupervised Learning



SUPERVISED LEARNING

Kita "**melatih model**", lalu dengan pengetahuan tersebut, si model dapat memprediksikan data yang baru atau belum diketahui

Supervised vs Unsupervised Learning

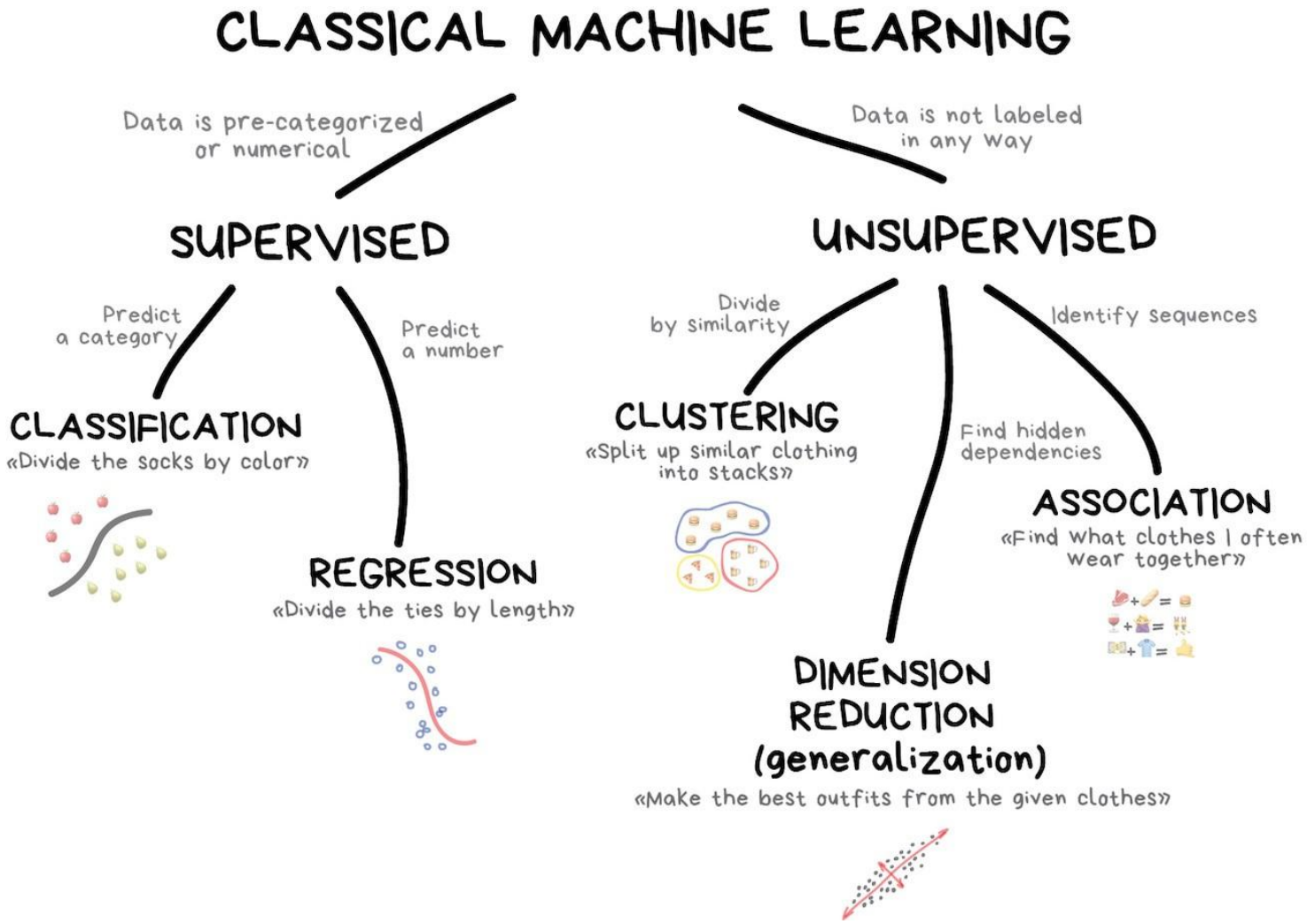


UNSUPERVISED LEARNING

Model / algoritma berusaha untuk menemukan pengetahuan (pola, informasi, dll) dari sekumpulan data

Supervised vs Unsupervised Learning

- Supervised
  - Classification
  - Regression
- Unsupervised
  - Clustering
  - Association
  - Dimension Reduction

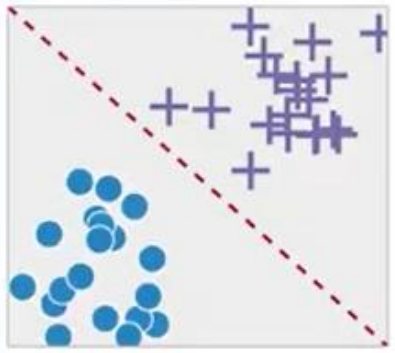


Classification vs Regression

- **Klasifikasi** berusaha memprediksi label atau kelas yang bersifat diskret / kategorikal

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

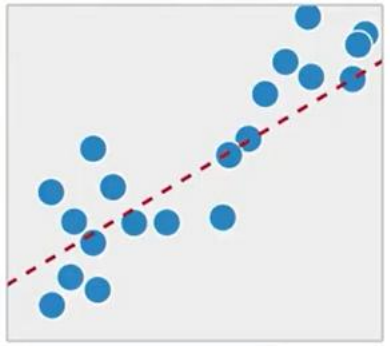
Categorical Values



- **Regresi** berusaha memprediksi label atau kelas yang bersifat kontinu / numerik

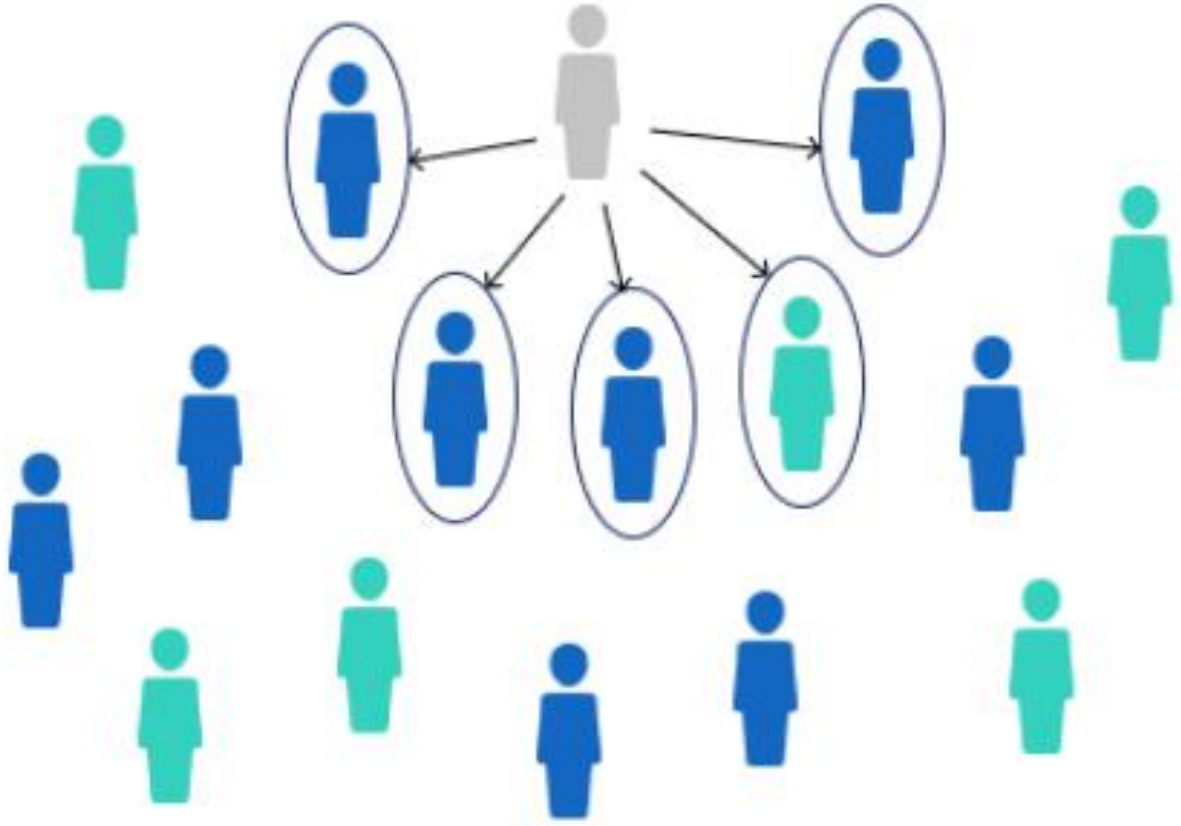
	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Continuous Values

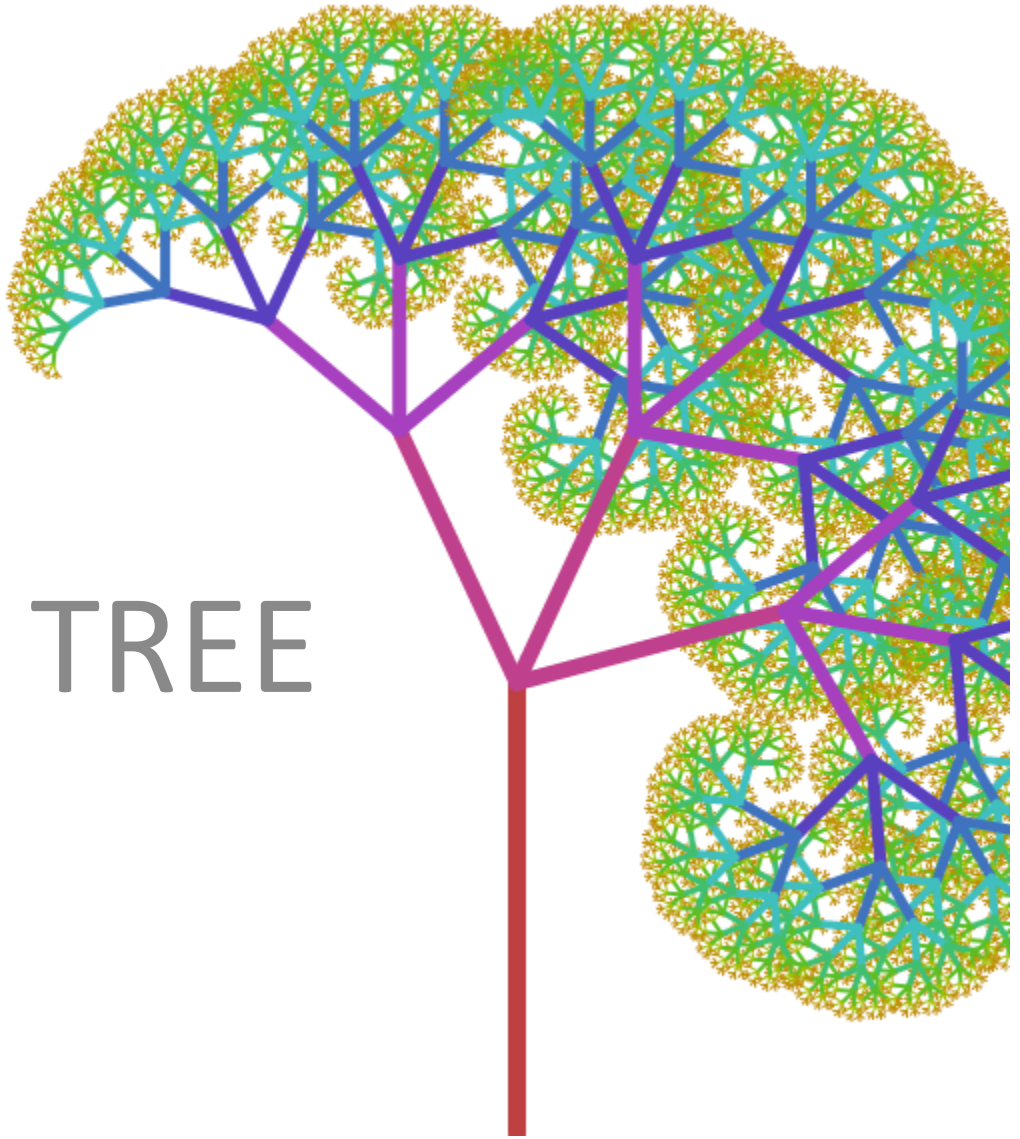


Algoritma Klasifikasi

- **Naïve Bayes.**
- **Decision Tree.**
- Logistic Regression.
- Stochastic Gradient Descent.
- K-Nearest Neighbours.
- Random Forest.
- Support Vector Machine.



ALGORITMA DECISION TREE



Membentuk Decision Tree Dari Data Latih

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Perlu dipahami terlebih dahulu mengenai:

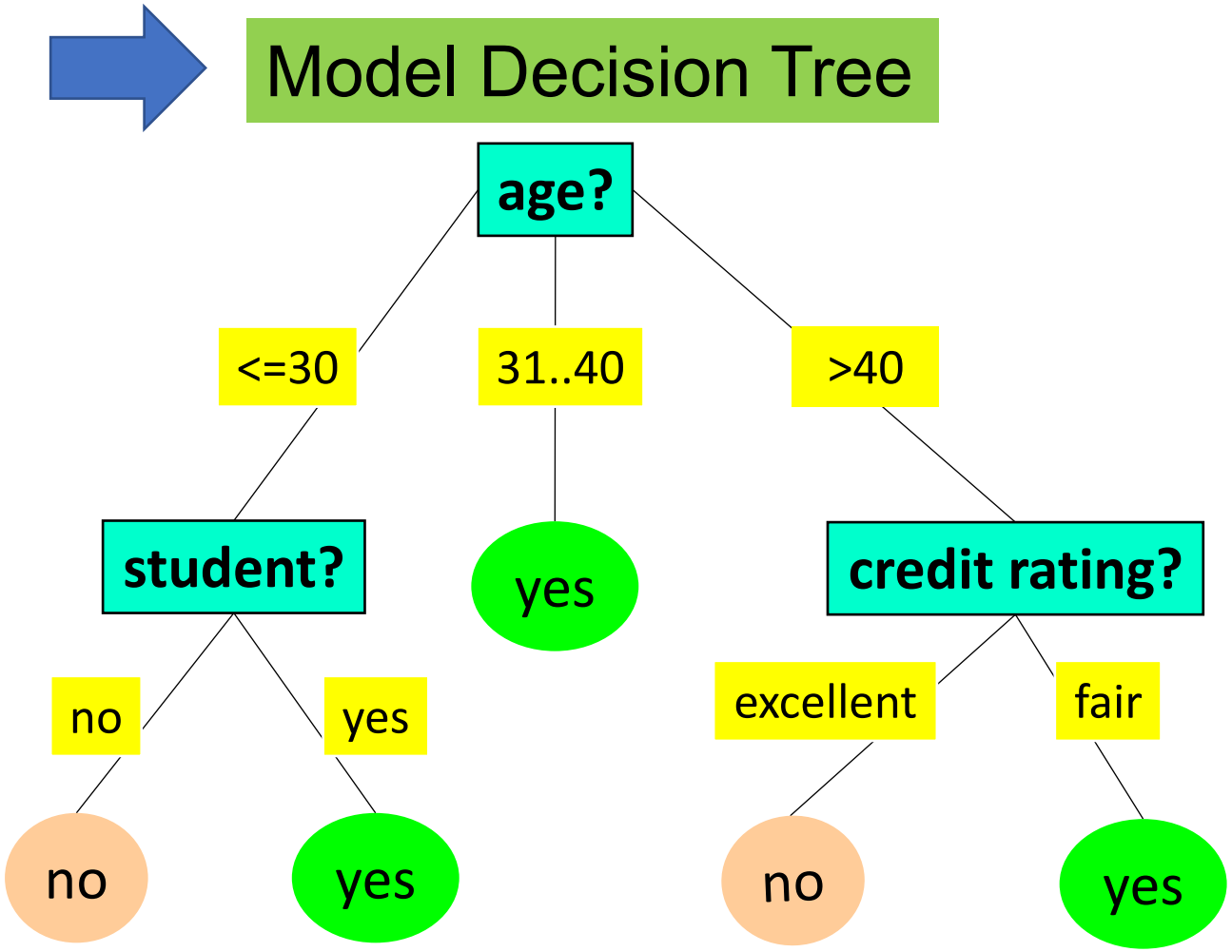
- Dataset
- Atribut
- Label / class
- Tipe data

Bagaimana dari data latih tersebut dapat diperoleh **model** yang bisa mengklasifikasikan secara otomatis?

Membangun Decision Tree Dari Data Latih

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

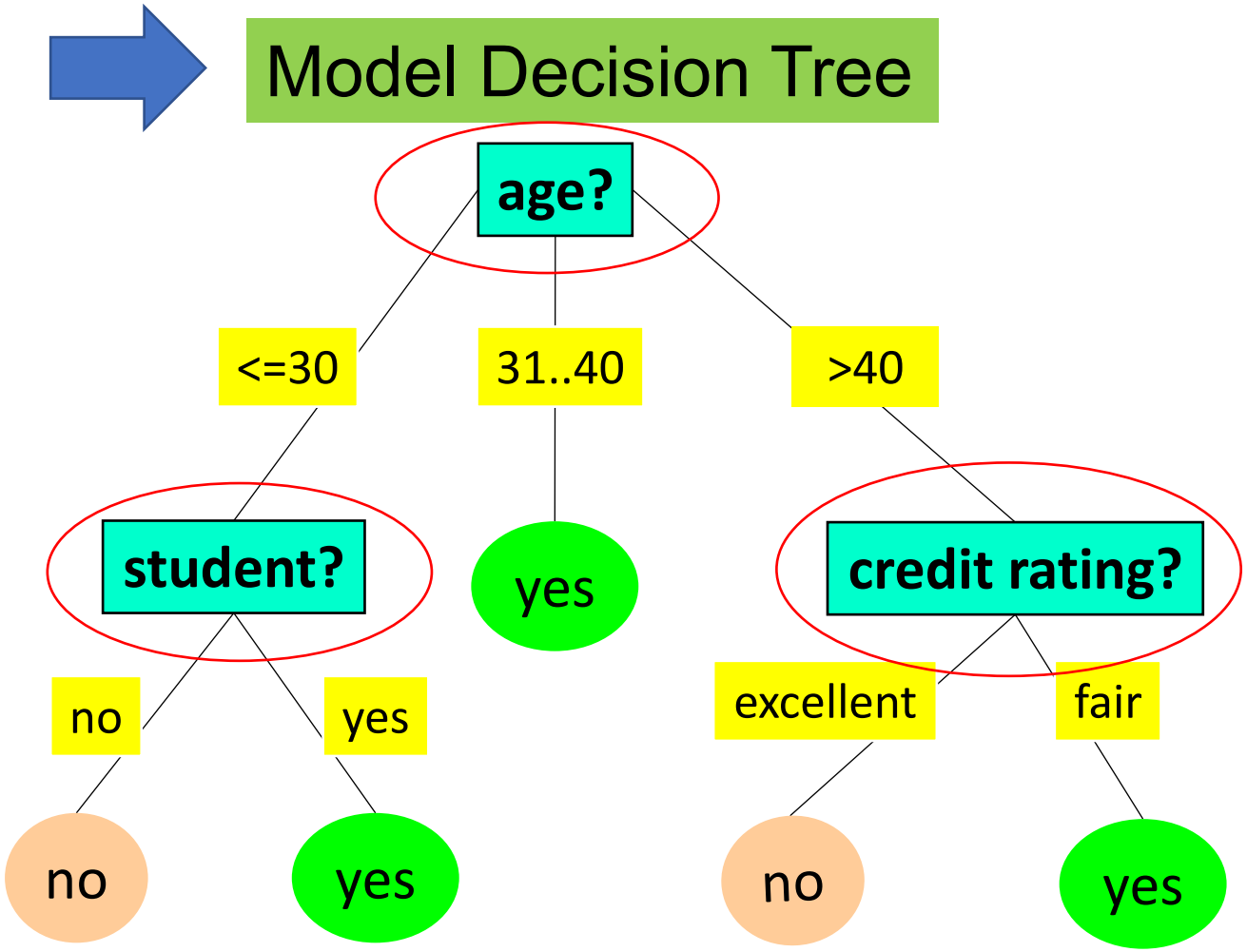
Rule:
IF ((age<=30) and (student)) OR (age=31..40) OR
((age>40) and (credit_rating=fair))
THEN
BELI_PC=YES



Membangun Decision Tree Dari Data Latih

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

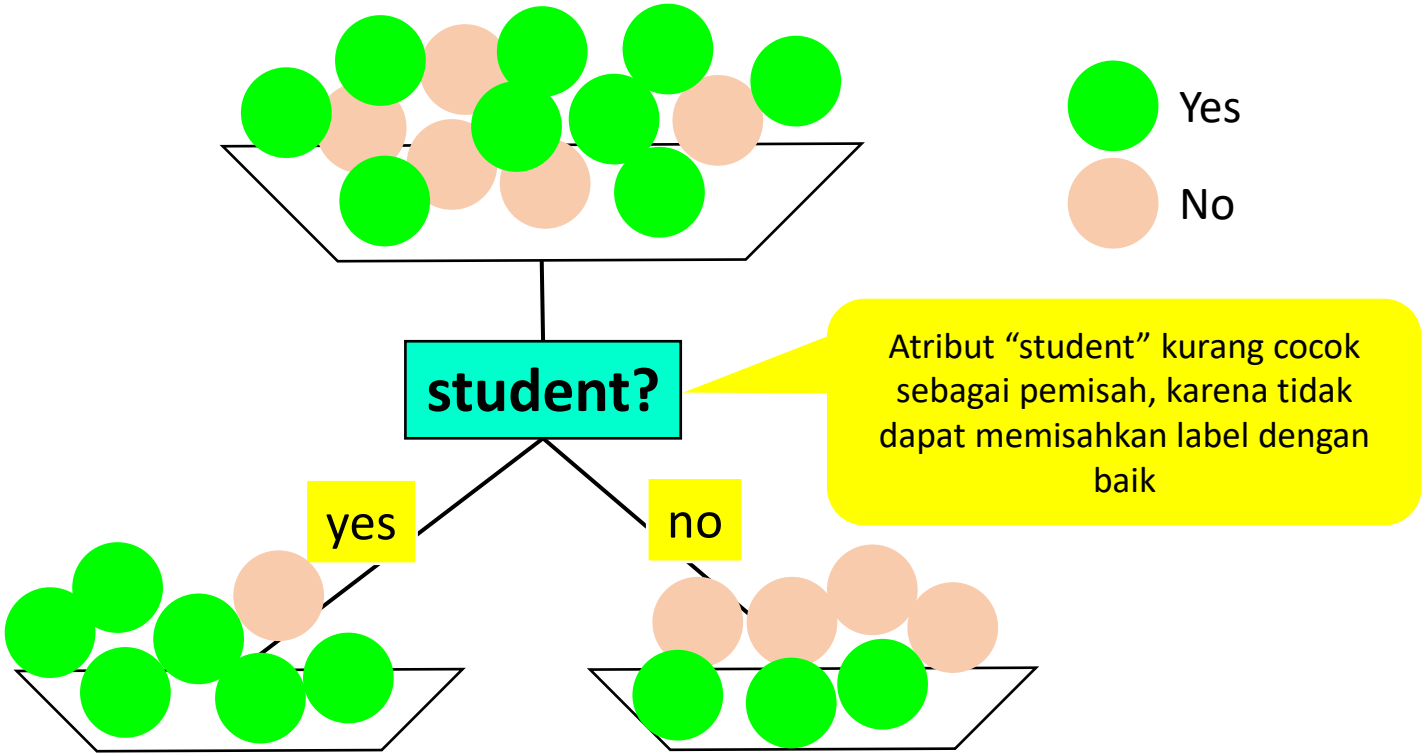
Bagaimana memilih atribut mana yang menjadi root? Disajikan terlebih dahulu, dst



Memilih Atribut Terbaik

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

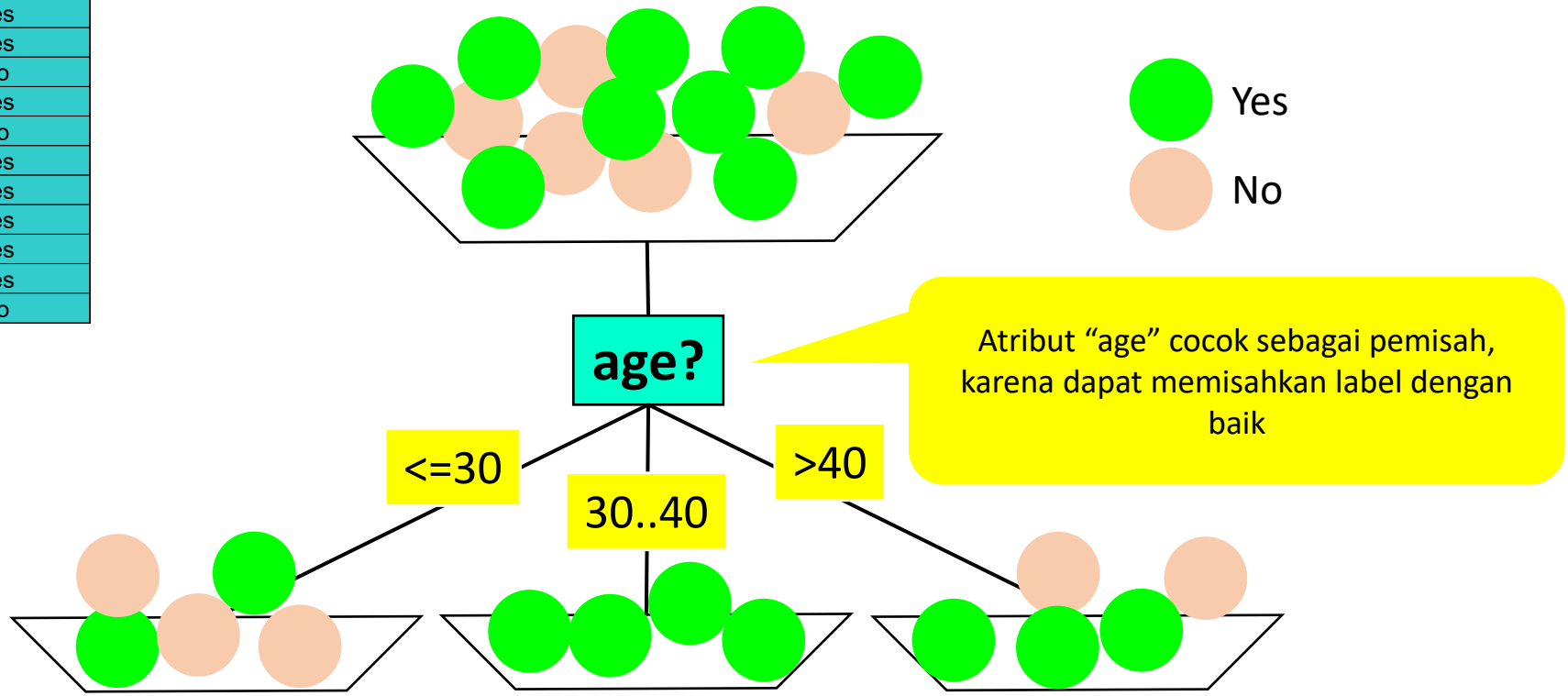
Kita coba atribut “student”



Memilih Atribut Terbaik

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

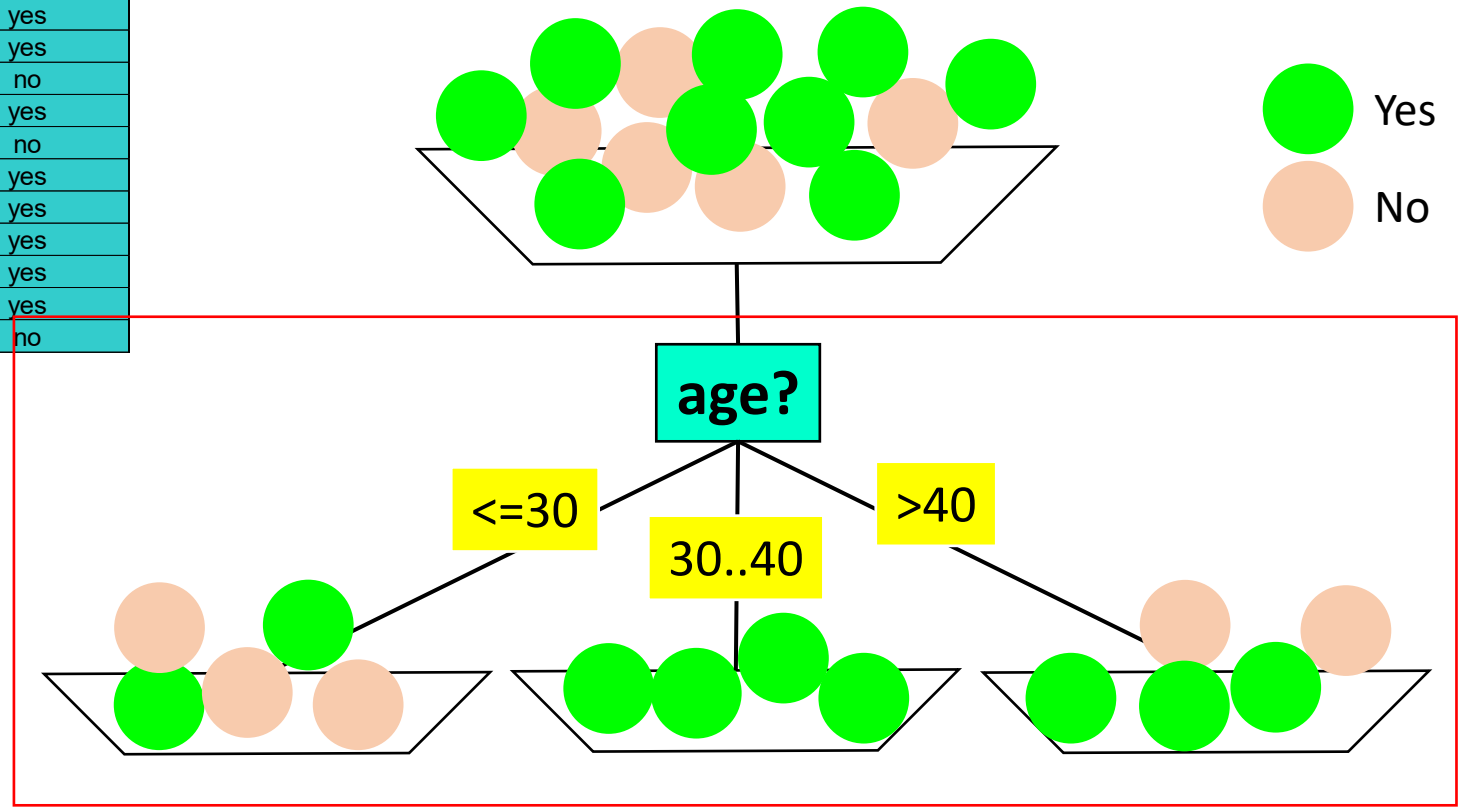
Kita coba atribut yang lain: “age”



Memilih Atribut Terbaik

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Kita coba atribut yang lain: “age”

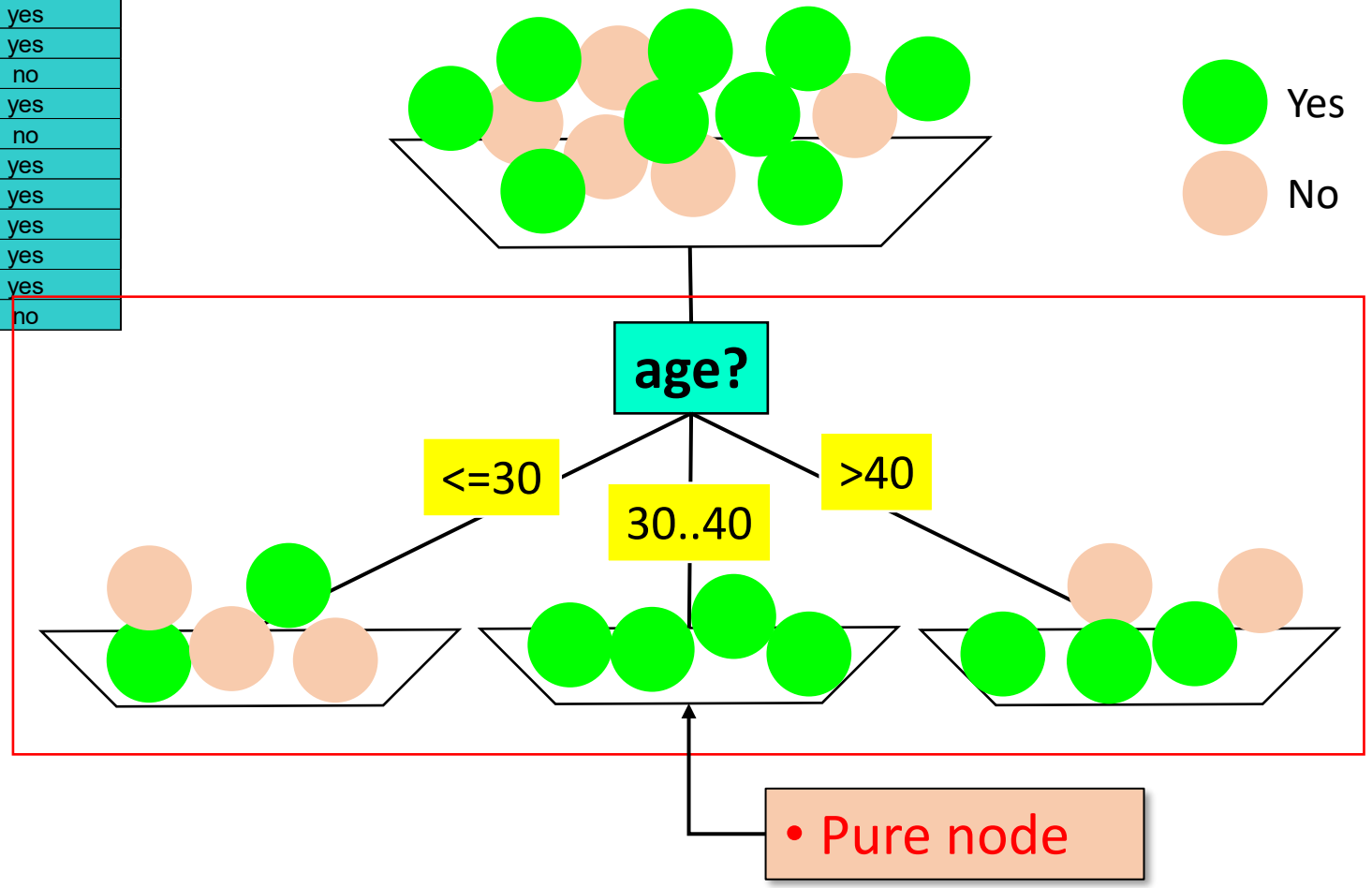


- More **Predictiveness**
- Less **Impurity**
- Lower **Entropy**

Memilih Atribut Terbaik

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Kita coba atribut yang lain: “age”



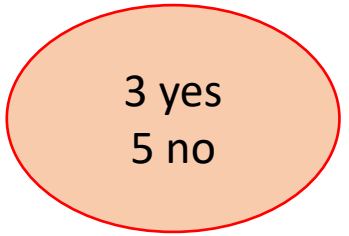
Entropy

- Ukuran kemurnian, keberagaman, randomness atau uncertainty
- Nilai entropy semakin kecil, distribusi semakin homogen, node semakin murni (pure)

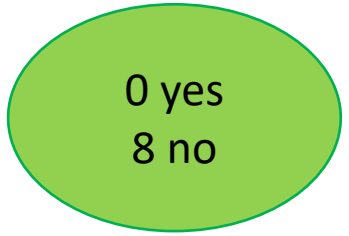
$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

*Info(D) = Entropy (D)

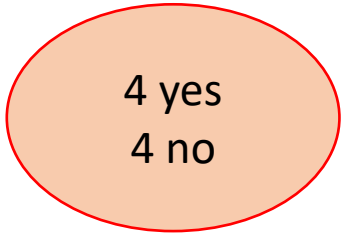
Entropy rendah



Entropy tinggi



Entropy = 0



Entropy = 1

Perhitungan Entropy

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

1 Hitung entropy dari class / label "**buys_computer**" (entropy data)

- buys_computer = "yes" → **9 data**
- buys_computer = "no" → **5 data**
- Total → **14 data**

$$Info(D) = I(9,5) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) = 0.940$$

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j)$$

Perhitungan Entropy

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

2 Hitung entropy dari setiap atribut

Atribut "age"

age	"yes"	"no"	I(yes,no)
<=30	2	3	0,971
31..40	4	0	0
>40	3	2	0,971

$$Info(2,3) = I(2,3) = -\frac{2}{5}\log_2(\frac{2}{3}) - \frac{3}{5}\log_2(\frac{3}{5}) = 0{,}971$$

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j)$$

$$Info_{age}(D) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2) = 0{,}694$$

Perhitungan Entropy

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

3 Hitung entropy dari setiap atribut

Atribut "income"

income	"yes"	"no"	I(yes,no)
low	3	1	0,811
medium	4	2	0,918
high	2	2	1

$$Info_{income}(D) = \frac{4}{14} I(3,1) + \frac{6}{14} I(4,2) + \frac{4}{14} I(2,2) = 0{,}911$$

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j)$$

Perhitungan Entropy

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

4 Hitung entropy dari setiap atribut

Atribut "student"

student	"yes"	"no"	I(yes,no)
yes	6	1	0,592
no	3	4	0,985

$$Info_{student}(D) = \frac{7}{14} I(6,1) + \frac{7}{14} I(3,4) = 0,788$$

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j)$$

Perhitungan Entropy

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

5 Hitung entropy dari setiap atribut

Atribut "credit_rating"

credit_rating	"yes"	"no"	I(yes,no)
fair	6	2	0,811
excellent	3	3	1

$$Info_{credit\_rating}(D) = \frac{8}{14} I(6,2) + \frac{6}{14} I(3,3) = 0{,}892$$

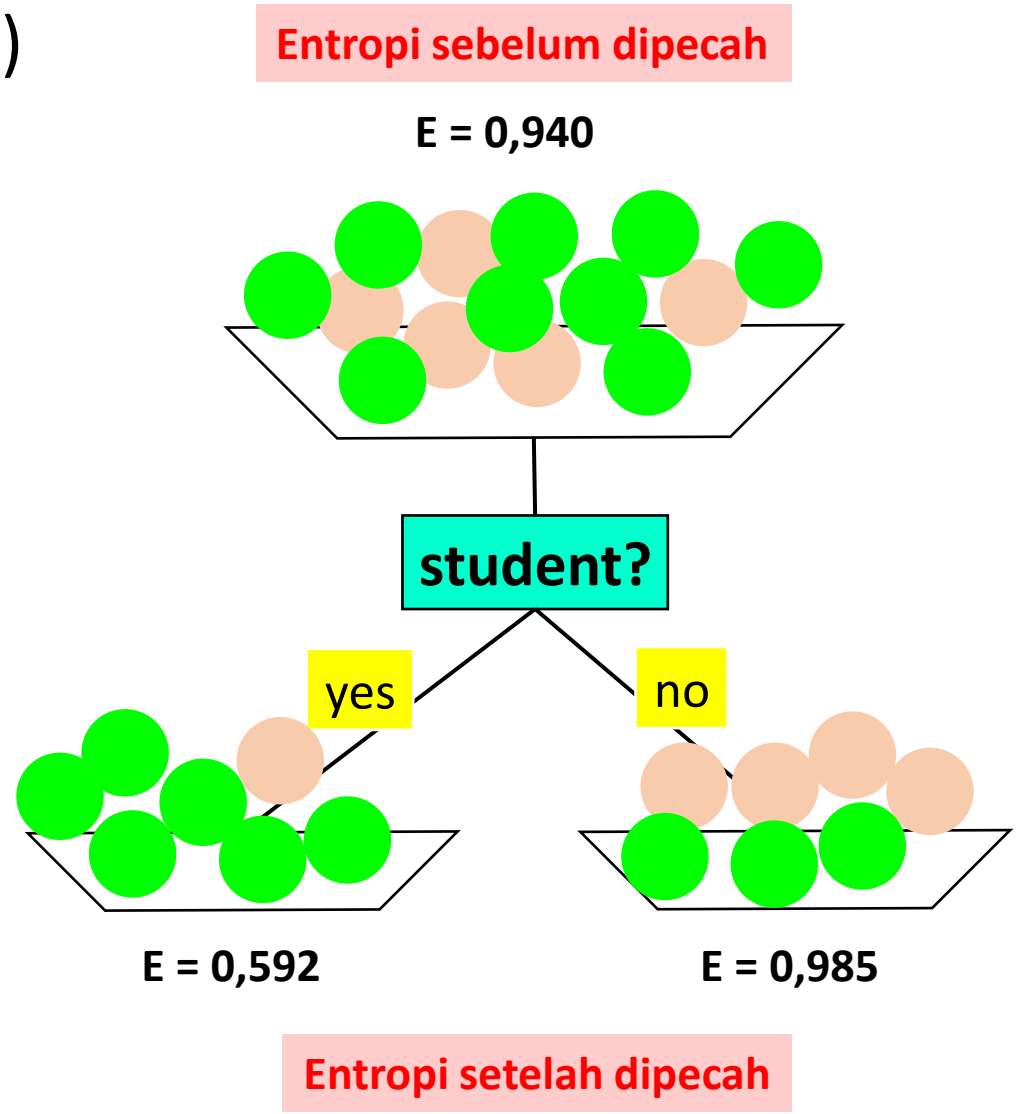
$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j)$$

Information Gain

- Information gain (sering disebut “gain” saja) merupakan informasi / nilai peningkatan derajat kepastian dari suatu atribut setelah dipecah (split)
- Gain(A) menggambarkan seberapa besar entropy berkurang akibat atribut A. Semakin besar, semakin bagus.

$$Gain(A) = Info(D) - Info_A(D)$$



Hitung Information Gain Setiap Atribut

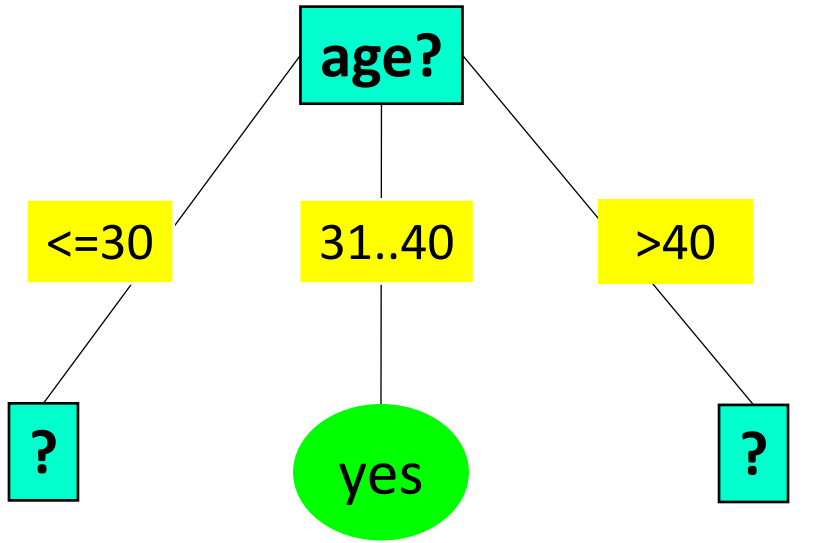
$$Gain(A) = Info(D) - Info_A(D)$$

Atribut	Entropi Atribut
buys_computer	0,904
age	0,694
income	0,911
student	0,788
credit_rating	0,892

- Gain (age) = 0,940 – 0,694 = **0,246**
- Gain (income) = 0,940 – 0,911 = **0,029**
- Gain (student) = 0,940 – 0,788 = **0,152**
- Gain (credit_rating) = 0,940 – 0,892 = **0,048**

Atribut "**age**" memiliki nilai information gain yang tertinggi, sehingga dipilih sebagai node awal (root)

Membentuk Decision Tree



Tidak perlu diproses, karena hanya 1 kemungkinan kelas

- Setelah atribut “age”, atribut apa selanjutnya?
- Diproses untuk setiap cabang selama masih ada > 1 kelas

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Membentuk Decision Tree

Selanjutnya... proses data age<=30

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
<=30	medium	yes	excellent	yes

$$Info(D) = I(2,3) = -\frac{2}{5}\log_2(\frac{2}{5}) - \frac{3}{5}\log_2(\frac{3}{5}) = 0.97$$

Hitung Gain atribut:

- Gain(Age): tidak perlu dihitung lagi
- Gain(income)
- Gain(student)
- Gain(credit_rating)

$$Info_{income}(D) = \frac{2}{5}I(0,2) + \frac{2}{5}I(1,1) + \frac{1}{5}I(1,0) = 0.4$$

Gain (income) = Info(D) – Info$_{income}$(D) = 0.97 – 0.4 = 0.57

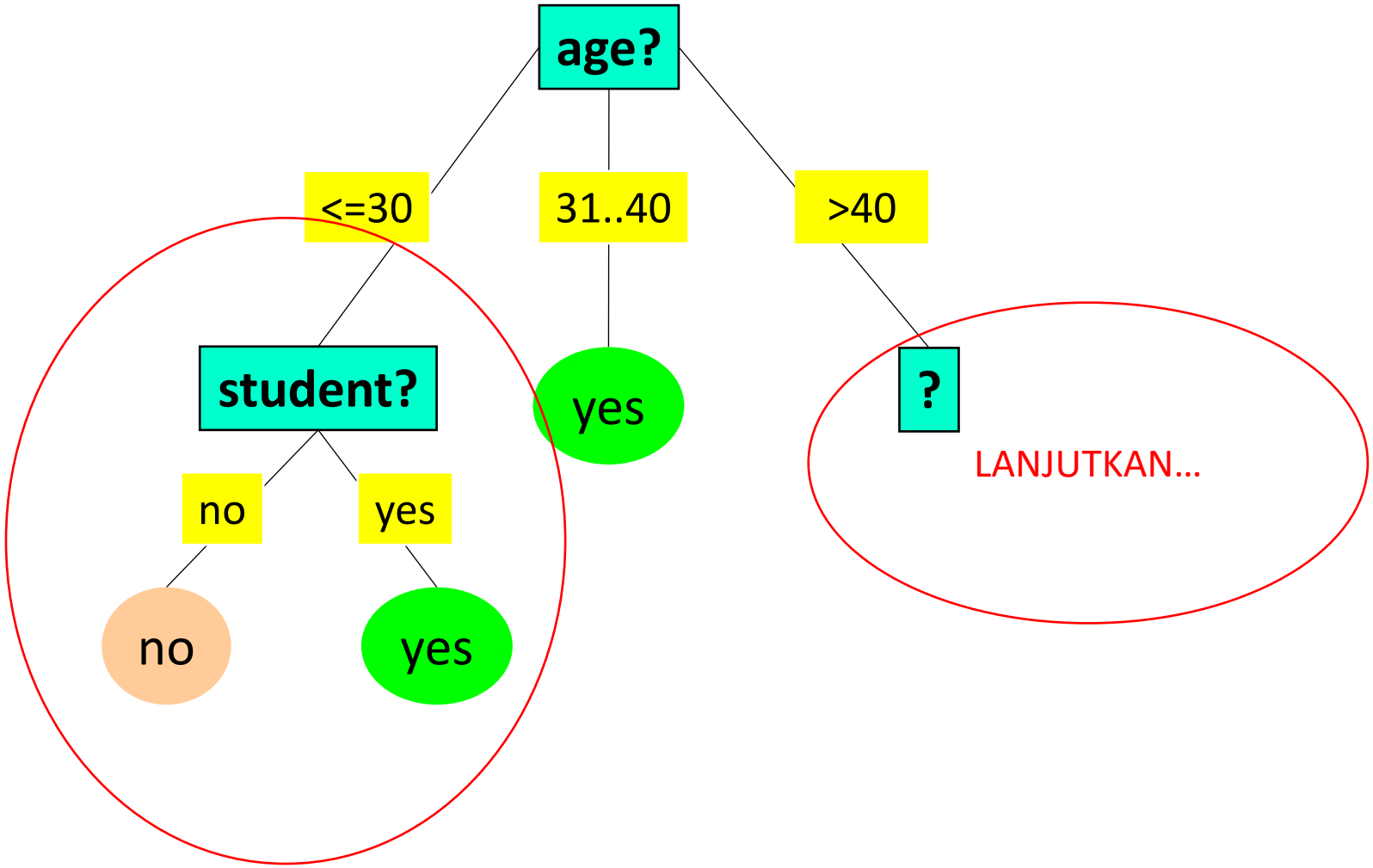
$$Info_{student}(D) = \frac{3}{5}I(0,3) + \frac{2}{5}I(2,0) = 0$$

Gain (student) = Info(D) – Info$_{student}$(D) = 0.97 – 0 = **0.97**

$$Info_{credit\_rating}(D) = \frac{3}{5}I(1,2) + \frac{2}{5}I(1,1) = 0.95$$

Gain (credit_rating) = Info(D) – Info$_{credit_rating}$(D) = 0.97 – 0.95 = 0.02

Membentuk Decision Tree



Cara lain dalam membentuk Decision Tree

- Menggunakan Gini Index (GI) atau Impurity
- Langkah:
  - Menghitung Gini Index (GI) untuk setiap atribut
  - Menentukan root berdasarkan nilai GI. Root adalah atribut dengan nilai GI terkecil.
  - Ulangi langkah 1 dan 2 untuk level berikutnya pada tree hingga nilai GI = 0.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Kapan Menggunakan Decision Tree?

- Data dalam bentuk atribut-nilai. Kondisi ideal adalah jika isi nilai jumlahnya sedikit. Misalnya: “panas”, “sedang”, “dingin”.
- Output diskrit.
- Training data dapat tidak lengkap

Kelebihan dan Kekurangan Decision Tree

- Kelebihan:
  - Mudah diimplementasikan
  - Hipotesis yang dihasilkan mudah dipahami
  - Efisien
- Kekurangan:
  - Overfitting: terlalu mengikuti training data
    - Terlalu banyak cabang, merefleksikan anomali akibat noise atau outlier.
    - Akurasi rendah untuk data baru
  - Proses pembangunan pohon keputusan pada data numerik menjadi lebih rumit dan memungkinkan terdapat informasi yang hilang
  - Pohon keputusan dapat tumbuh menjadi sangat kompleks pada data yang rumit.

Pengembangan Decision Tree

- Mengatasi overfitting
  - **Pre-pruning:** Hentikan pembuatan tree di awal. Tidak mensplit node jika goodness measure dibawah threshold.
  - **Post-pruning:** Buang cabang setelah tree jadi
- Pengembangan Metode:
  - C4.5, C5.0
  - Conditional Decision Tree
  - Gradient-boosted Trees
  - Random Forest

Studi Kasus 3 (Tugas Decision Tree) :

No	Kelas	Kulit Buah	Warna	Ukuran	Bau
1	Aman	Kasar	Coklat	Besar	keras
2	Aman	Kasar	Hijau	Besar	keras
3	Berbahaya	Halus	Merah	Besar	Lunak
4	Aman	Kasar	Hijau	Besar	Lunak
5	Aman	Kasar	Merah	Kecil	Keras
6	Aman	Halus	Merah	Kecil	Keras
7	Aman	Halus	Coklat	Kecil	Keras
8	Berbahaya	Kasar	Hijau	Kecil	Lunak
9	Berbahaya	Halus	Hijau	Kecil	Keras
10	Aman	Kasar	Merah	Besar	Keras
11	Aman	Halus	Coklat	Besar	Lunak
12	Berbahaya	Halus	Hijau	Kecil	Keras
13	Aman	Kasar	Merah	Kecil	Lunak
14	Berbahaya	Halus	Merah	Besar	Keras
15	Aman	Halus	Merah	Kecil	Keras
16	Berbahaya	Kasar	Hijau	Kecil	Keras

Penjelasan Tugas:

Terdapat 3 soal (2 studi kasus NB dan 1 DT)

Kerjakan dengan tulis tangan, foto, kemudian masukkan ke dlm file word atau pdf.
Uplot ke lms, perhatikan batas akhir tgl submit, jangan terlambat.

Selamat Bertugas

REFERENSI

- Russel, S., & Norvig, P. (2003). Artificial Intelligence A Modern Approach . New Jersey : Pearson Education, Inc.
- Slide Jiawei Han http://www.cs.uiuc.edu/~hanj/bk2/
- Course “Machine Learning with Python” dari CognitiveClass.Ai
- Slide Materi Pelatihan Digital Talent Academy – Kominfo.
- https://towardsdatascience.com/supervised-vs-unsupervised-learning-in-2-minutes-72dad148f242