

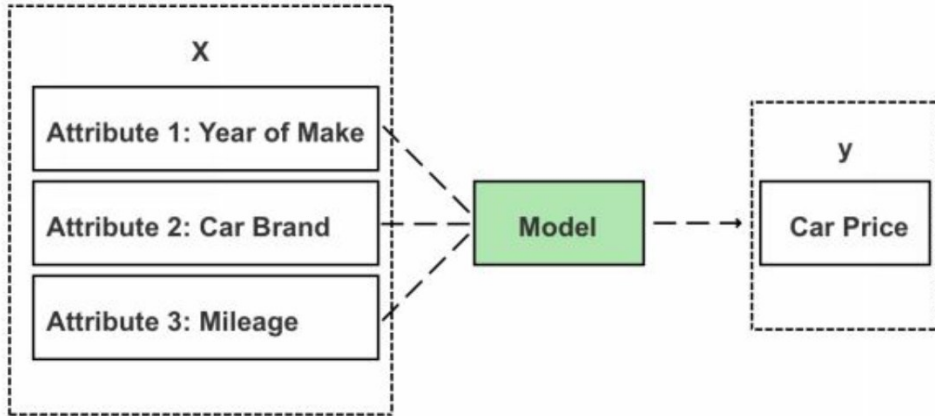


SUPERVISED LEARNING - CLASSIFICATION

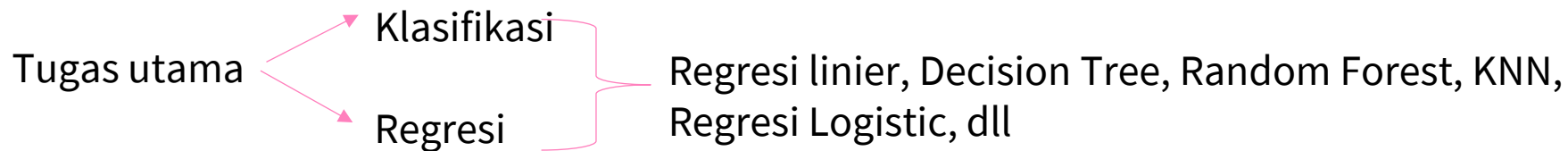
TEAM TEACHING



METODE SUPERVISED LEARNING



- Supervised Learning adalah pembelajaran **terarah/ terawasi**.
- Tujuan utama dari metode **Supervised Learning** adalah untuk mempelajari hubungan antara sampel data input dan output yang sesuai setelah melakukan beberapa training data sampel





PENDAHULUAN

- Klasifikasi adalah proses menemukan atau menggali model atau fungsi yang membantu dalam memisahkan data menjadi beberapa kelas kategorikal yaitu nilai diskrit.
- Ini adalah tugas mendekati fungsi pemetaan (f) dari variabel input (X) ke variabel output diskrit (Y).
- Variabel keluaran sering disebut label atau kategori.
- Fungsi pemetaan memprediksi kelas atau kategori untuk pengamatan yang diberikan.





PENDAHULUAN

Klasifikasi

- Memprediksi kelas suatu item
- Membuat model berdasarkan data pelatihan dan digunakan untuk mengklasifikasi data.

Prediksi

- Memprediksi nilai yang belum diketahui





PENDAHULUAN

Aplikasi

- Persetujuan kredit
- Diagnosis penyakit
- Target marketing
- Fraud detection





JENIS KLASIFIKASI

- Klasifikasi biner
- Klasifikasi multi-kelas





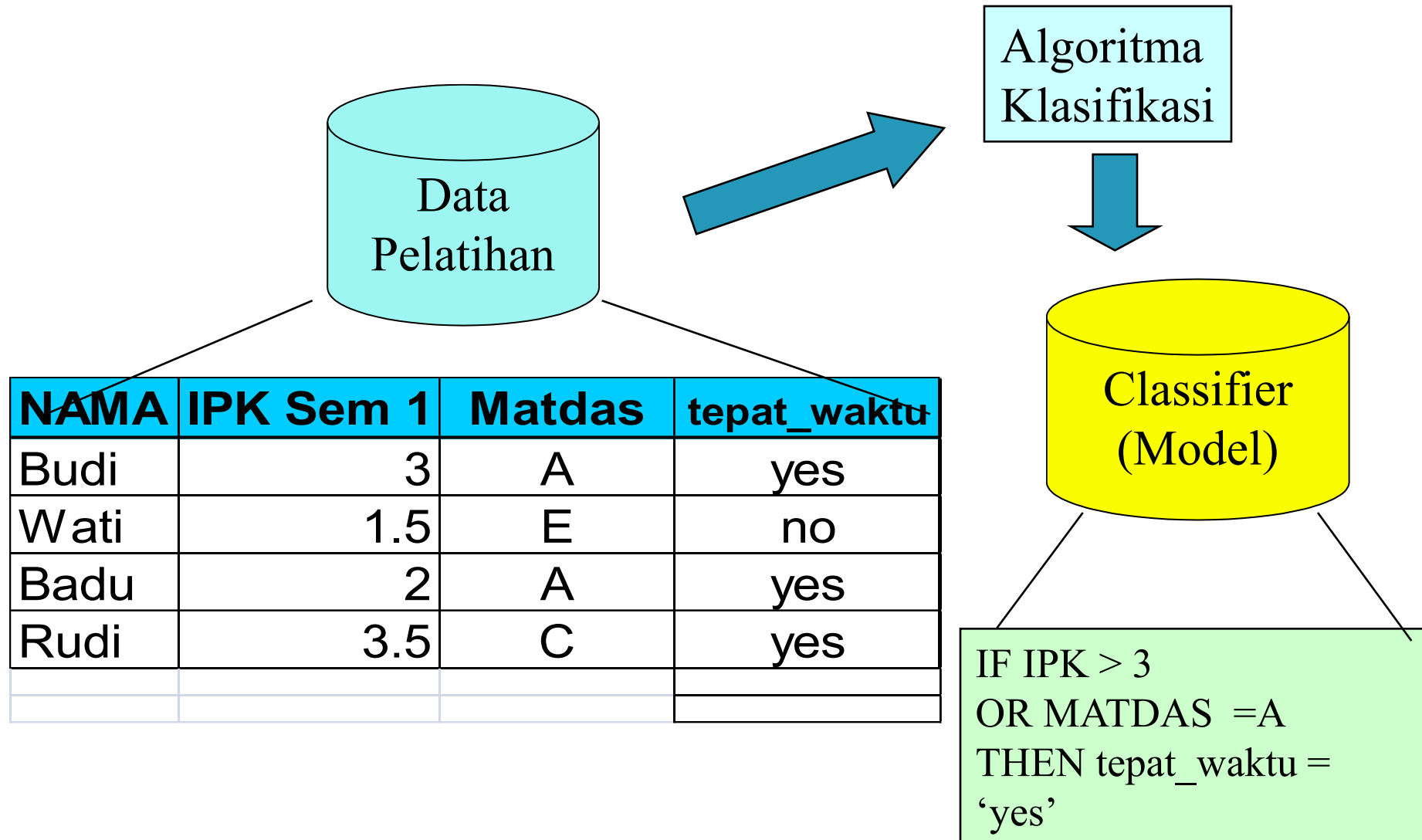
CONTOH KASUS

- Input: data mahasiswa
- Output: dua kelas (lulus_tepat_waktu dan lulus_terlambat)

Bagaimana kalau diberikan data input mahasiswa, sistem secara otomatis menentukan mhs tersebut akan lulus tepat waktu atau terlambat?

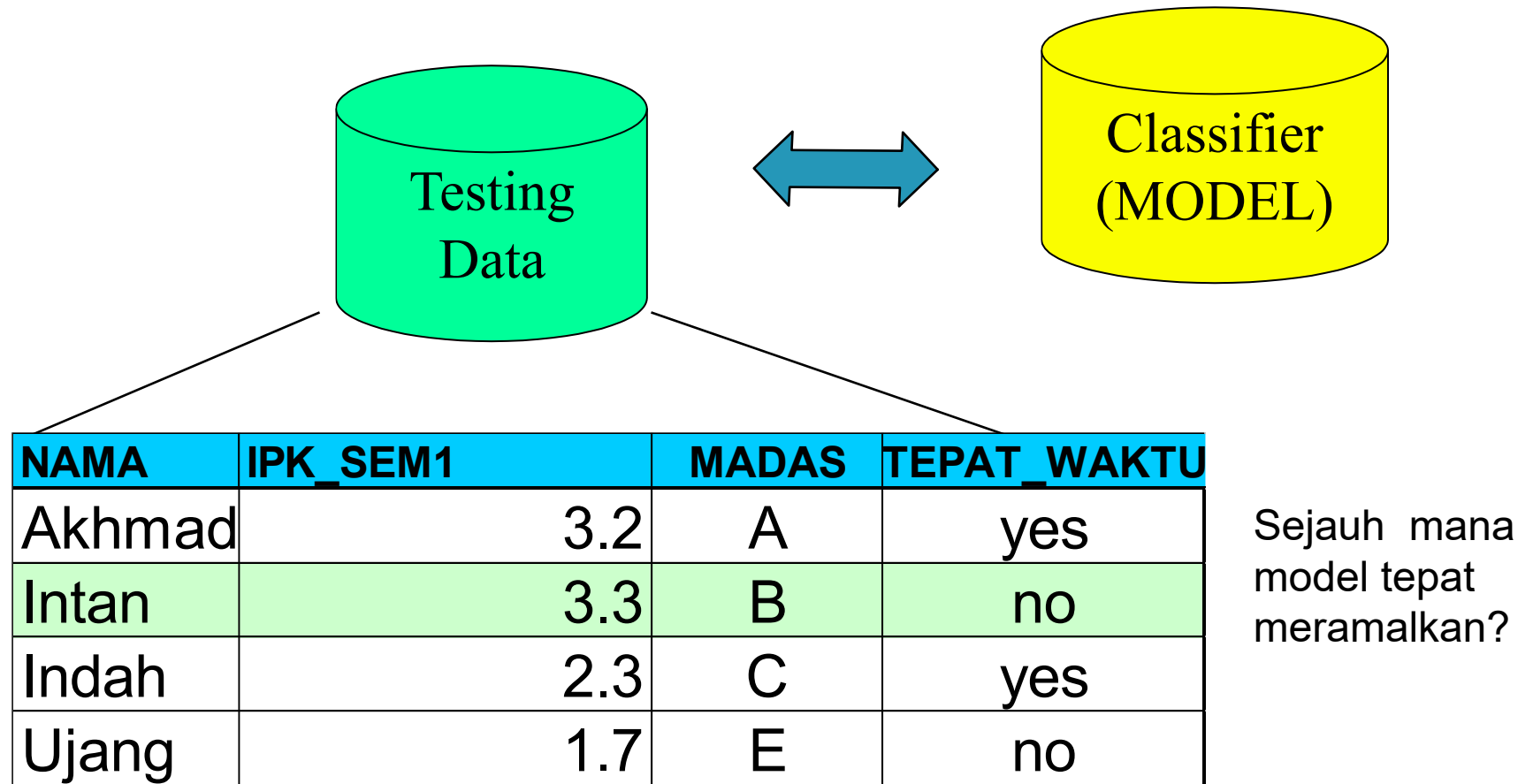


PEMBUATAN MODEL

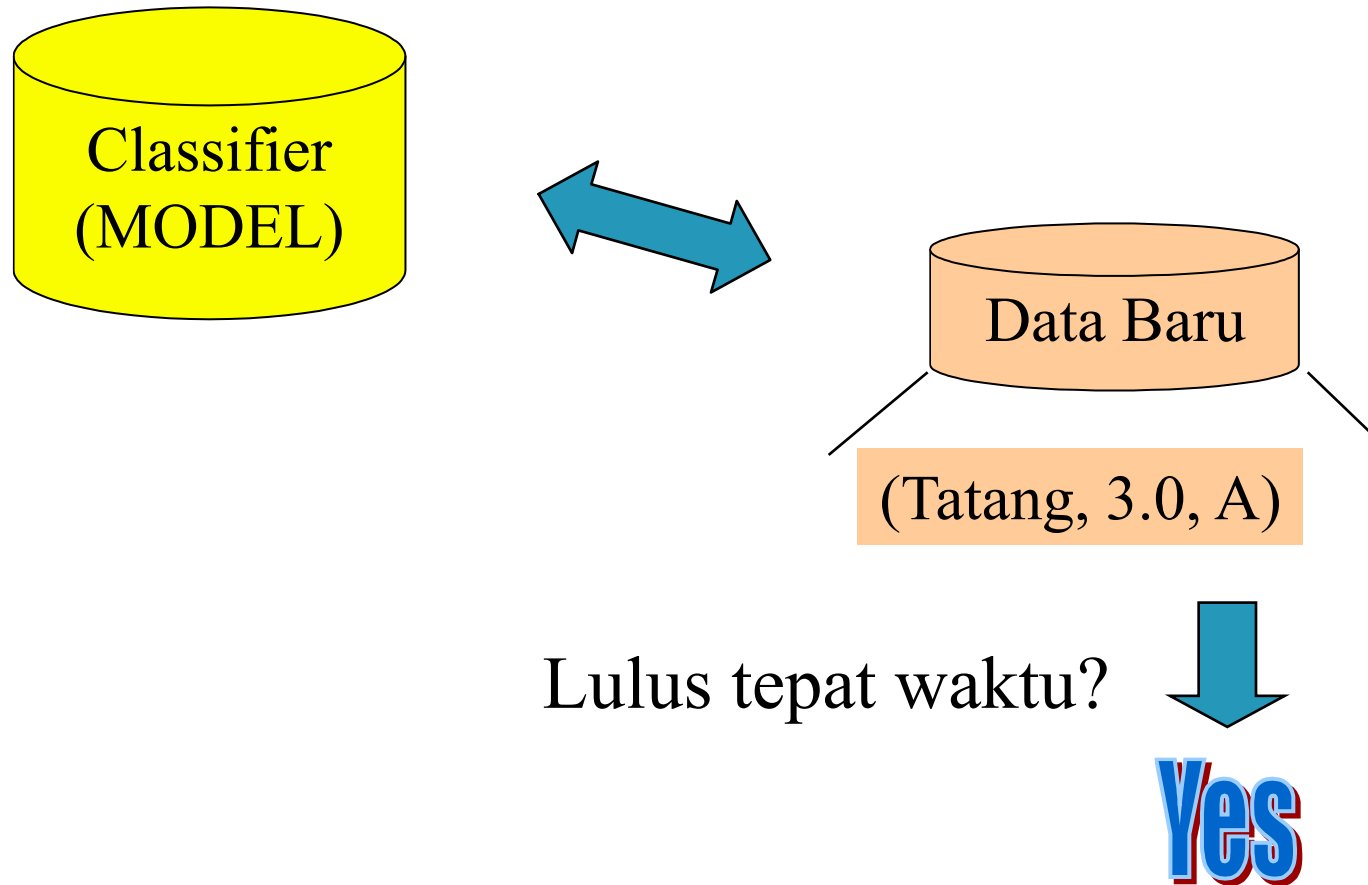




PROSES TESTING MODEL



Proses Klasifikasi





PROSES KLASIFIKASI

- Proses pembuatan model
 - Data latih → Model Klasifikasi
- Proses testing model
 - Data testing → Apakah model sudah benar?
- Proses klasifikasi
 - Data yang tidak diketahui kelasnya → kelas data





SEBELUM KLASIFIKASI

- Data cleaning
 - Preprocess data untuk mengurangi noise dan missing value
- Relevance analysis (feature selection)
 - Memilih atribut yang penting
 - Membuang atribut yang tidak terkait atau duplikasi.
- Data transformation
 - Generalize and/or normalize data





EVALUASI METODE KLASIFIKASI

- Akurasi
 - classifier accuracy: memprediksi label kelas
 - predictor accuracy: memprediksi nilai atribut
- Kecepatan
 - Waktu untuk membuat model (training time)
 - Waktu untuk menggunakan model (classification/prediction time)





EVALUASI METODE KLASIFIKASI

- Robustness: menanggapi noise dan missing value.
- Scalability: efisien untuk proses dengan DBMS
- Interpretability
 - Model mudah dimengerti





CLASSIFICATION TECHNIQUES

- Decision Tree based Methods
- Rule-based Methods
- Naïve Bayes and Bayesian Belief Networks
- Neural Networks
- Support Vector Machines
- and more...





DECISION TREE

- Diciptakan oleh Ross Quinlan
- ID3, C4.5, C5.0
- Model direpresentasikan dalam bentuk tree





DECISION TREE: CONTOH INPUT (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





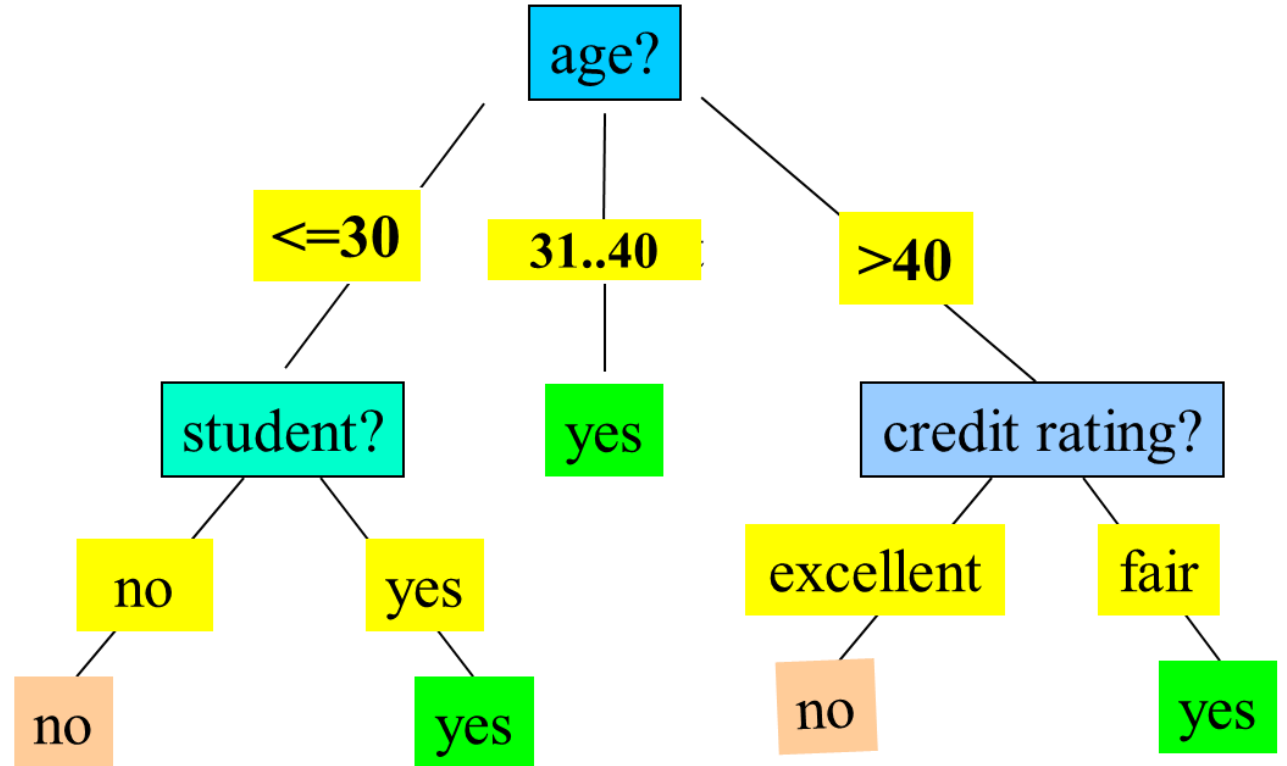
MASALAH

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



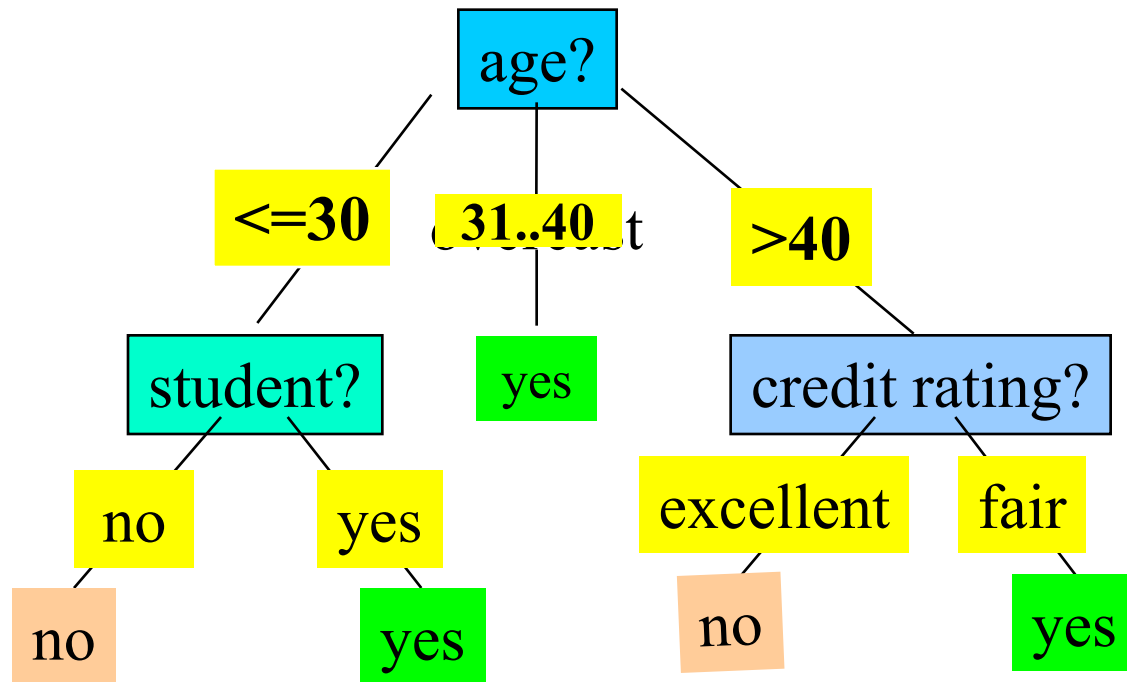
MODEL: DECISION TREE

Dari data latih, model ini
dibangkitkan secara
otomatis...





TREE DAPAT DIREPRESENTASIKAN SEBAGAI RULE



((age<=30) and
(student))

OR

(age=31..40)

OR

(age>40) and
(credit_rating=fair)

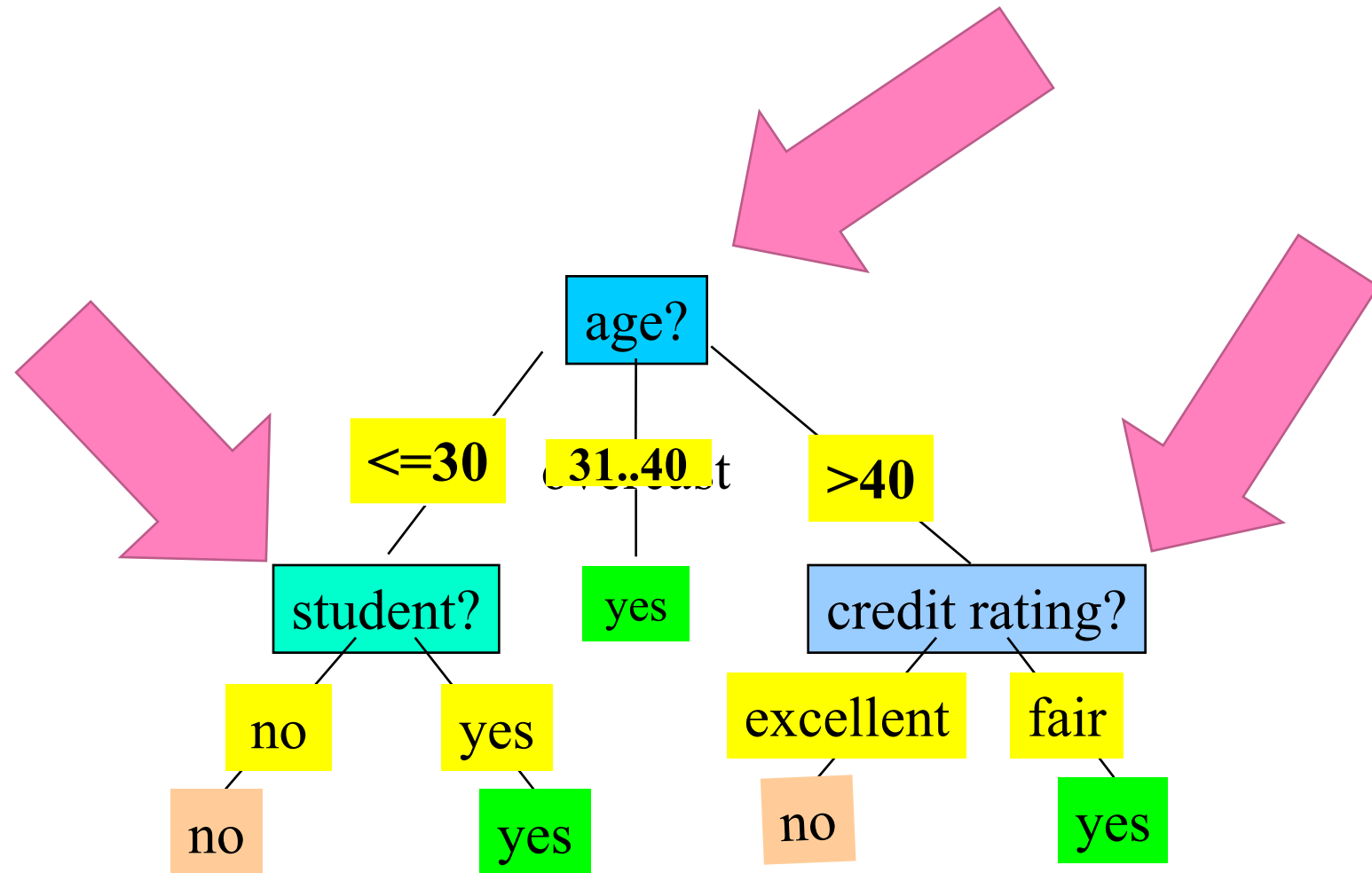
THEN

BELI_PC=YES





BAGAIMANA CARA PEMILIHAN URUTAN ATRIBUT?





CARA PEMILIHAN ATRIBUT

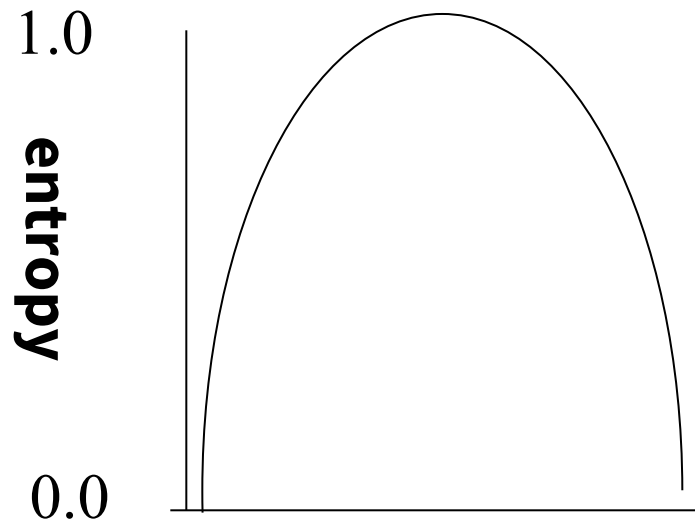
- Entropy: Ukuran kemurnian, semakin murni, semakin homogen, semakin rendah nilainya.
- Information Gain: pengurangan entropy disebabkan oleh partisi berdasarkan suatu atribut.
- Semakin besar info gain = atribut itu semakin membuat homogen = semakin bagus
- Idanya → pilih atribut dengan info gain yg paling besar





ENTROPY UNTUK DUA KELAS: + DAN -

$$\text{Entropy}(S) \equiv -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$$



- Entropy([9+,5-] ((9 positif, 5 neg))
= $-(9/14) \log_2(9/14) - (5/14) \log_2(5/14)$
= 0.940
- Entropy([9+,5-]) = 0.940
- Entropy([7+,7-]) = 1
- Entropy([14+,0]) = 0
- Entropy([0+,14-]) = 0





ENTROPHY UNTUK KELAS > 2

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

Info (D) = Entrophy (D) (istilah dibuku J. HAN)





INFORMATION GAIN

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

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

- Gain(A) seberapa besar entropy berkurang akibat atribut A.
- Makin besar makin bagus.



CONTOH PEMILIHAN ATRIBUT

- Class P: buys_computer = “yes”
- Class N: buys_computer = “no”

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

age	p_i	n_i	$I(p_i, n_i)$
≤ 30	2	3	0.971
31...40	4	0	0
> 40	3	2	0.971

$$Info_{age}(D) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2) = 0.694$$

$\frac{5}{14} I(2,3) \Rightarrow$ berarti ada 5 dari 14

“age ≤ 30 ” dgn 2 yes dan 3 no.

$$\begin{aligned} \text{Gain (Age)} &= \text{Info}(D) - \text{Info age}(D) \\ &= 0.940 - 0.694 = 0.246 \end{aligned}$$

$$\text{Gain (Income)} = 0.029$$

$$\text{Gain (Student)} = 0.151$$

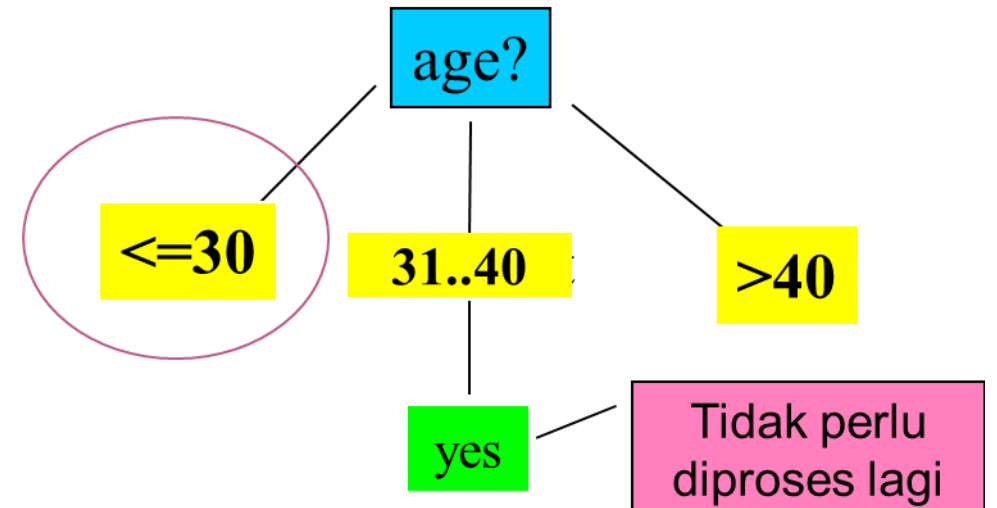
$$\text{Gain (Credit_Rating)} = 0.048$$





PEMILIHAN ATRIBUT

- Gain (Age) = 0.246 $\leftarrow$ yang terbesar, dipilih
- Setelah AGE, atribut apa selanjutnya?
- Diproses untuk setiap cabang selama masih ada > 1 kelas
- Selanjutnya... proses data yang ≤ 30





PEMILIHAN ATRIBUT

- Gain(age) tidak perlu dihitung lagi, hitung gain(student), gain(credit_rating), gain(income)

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

- hitung gain(student)

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

$$Gain(student) = Info(D) - Info_{student}(D) = 0.97 - 0 = 0.97$$

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





PEMILIHAN ATRIBUT

- hitung gain(credit_rating)

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

$$\begin{aligned} \text{Gain}(\text{credit_rating}) &= \text{Info}(D) - \text{Info}_{\text{credit_rating}}(D) \\ &= 0.97 - 0.95 = 0.02 \end{aligned}$$

- hitung gain(income)

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

$$\begin{aligned} \text{Gain}(\text{income}) &= \text{Info}(D) - \text{Info}_{\text{income}}(D) \\ &= 0.97 - 0.4 = 0.57 \end{aligned}$$

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



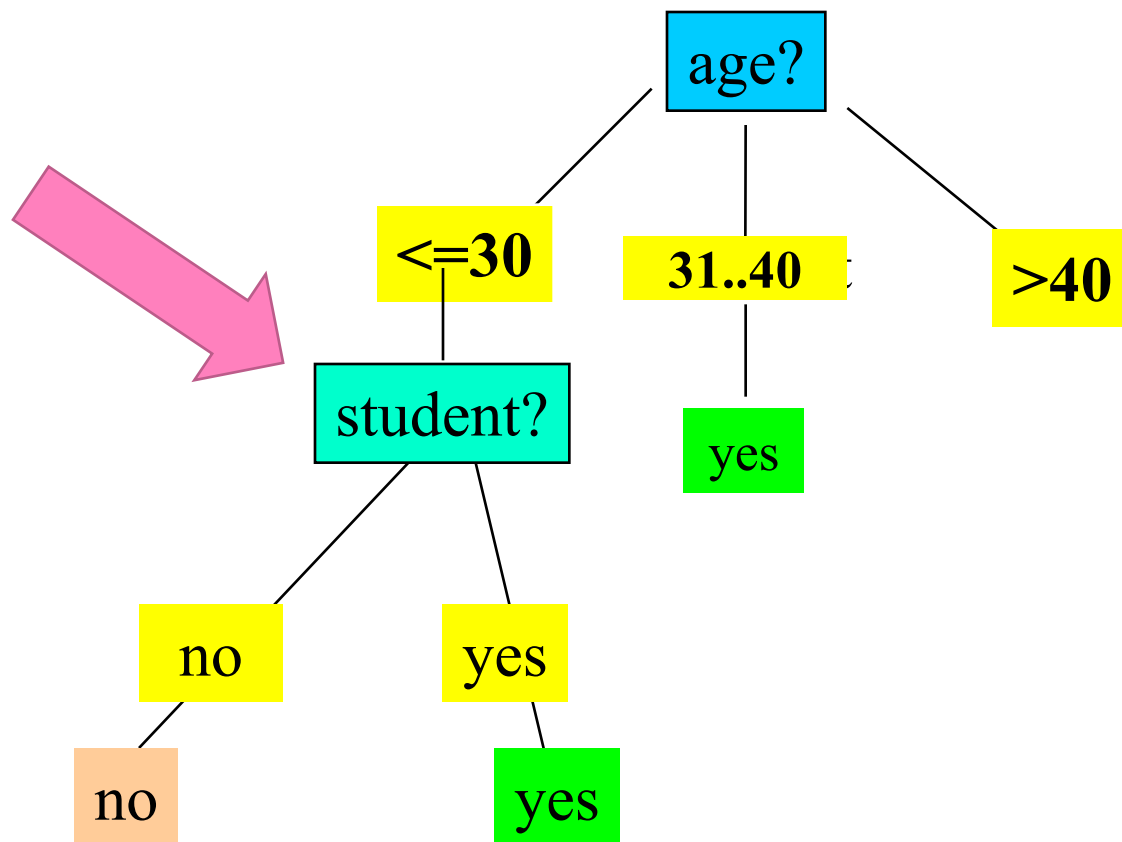


PEMILIHAN ATRIBUT

- Bandingkan semua gain, ambil yang paling besar
- **Gain (studet) = 0.97**
- Gain (credit_rating) = 0.02
- Gain (income) = 0.57



PEMILIHAN ATRIBUT





MENGAPA DECISION TREE?

- Mudah diimplementasikan
- Hipotesis yang dihasilkan mudah dipahami
- Efisien





DECISION TREE COCOK UNTUK MASALAH?

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





Masalah DT

Overfitting: terlalu mengikuti training data

- Terlalu banyak cabang, merefleksikan anomali akibat noise atau outlier.
- Akurasi rendah untuk data baru





Masalah DT

Dua pendekatan untuk menghindari overfitting

- Prepruning: Hentikan pembuatan tree di awal. Tidak mensplit node jika goodness measure dibawah threshold.
 - Sulit untuk menentukan threshold
- Postpruning: Buang cabang setelah tree jadi
 - Menggunakan data yang berbeda dengan training untuk menentukan pruned tree yang terbaik.





TERIMA KASIH

