CALVIN
INSTITUTE OF TECHNOLOGY
God's People for God's Glory

Decision Tree

Hendrik Santoso Sugiarto

IBDA2032 – *Artificial Intelligence*

Capaian Pembelajaran

- Decision Tree
- Alternatif Algoritma
- Decision Tree Regressor

Decision Tree

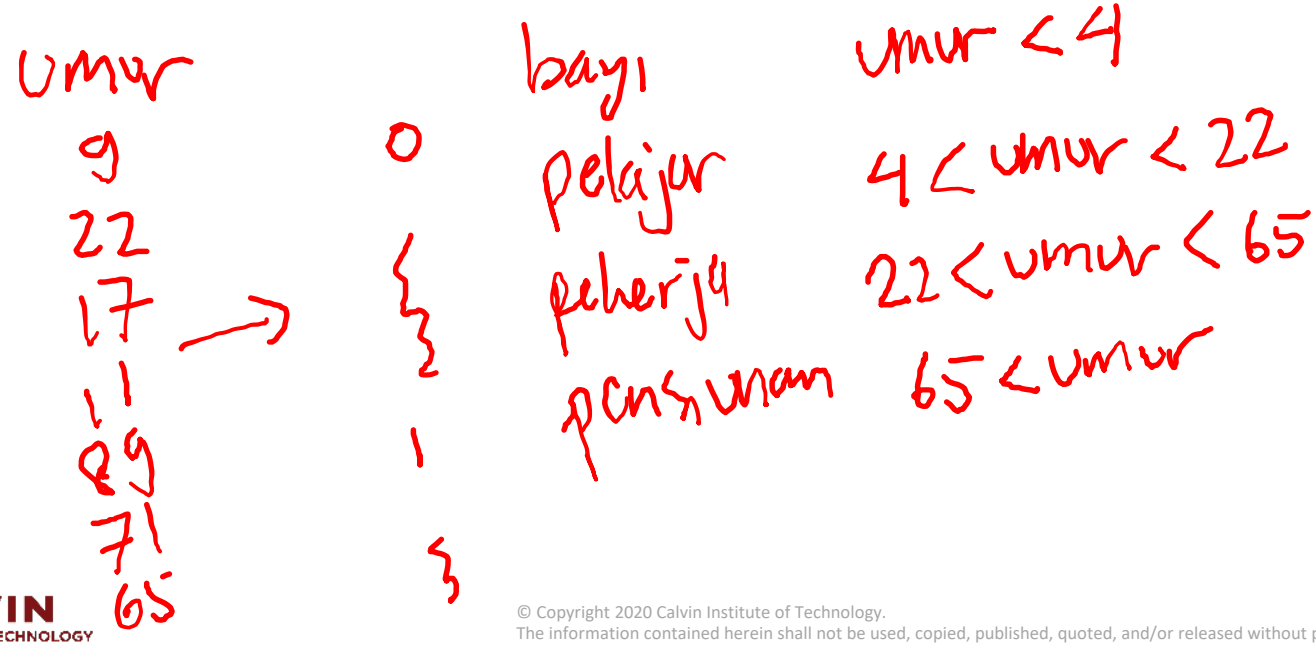
Decision Tree

- Paradigma machine learning yang menggunakan perspektif logika
- Dapat digunakan untuk regresi, klasifikasi, unsupervised
- Mudah diinterpretasi
- Dapat dijadikan probabilistik

CALVIN INSTITUTE OF TECHNOLOGY

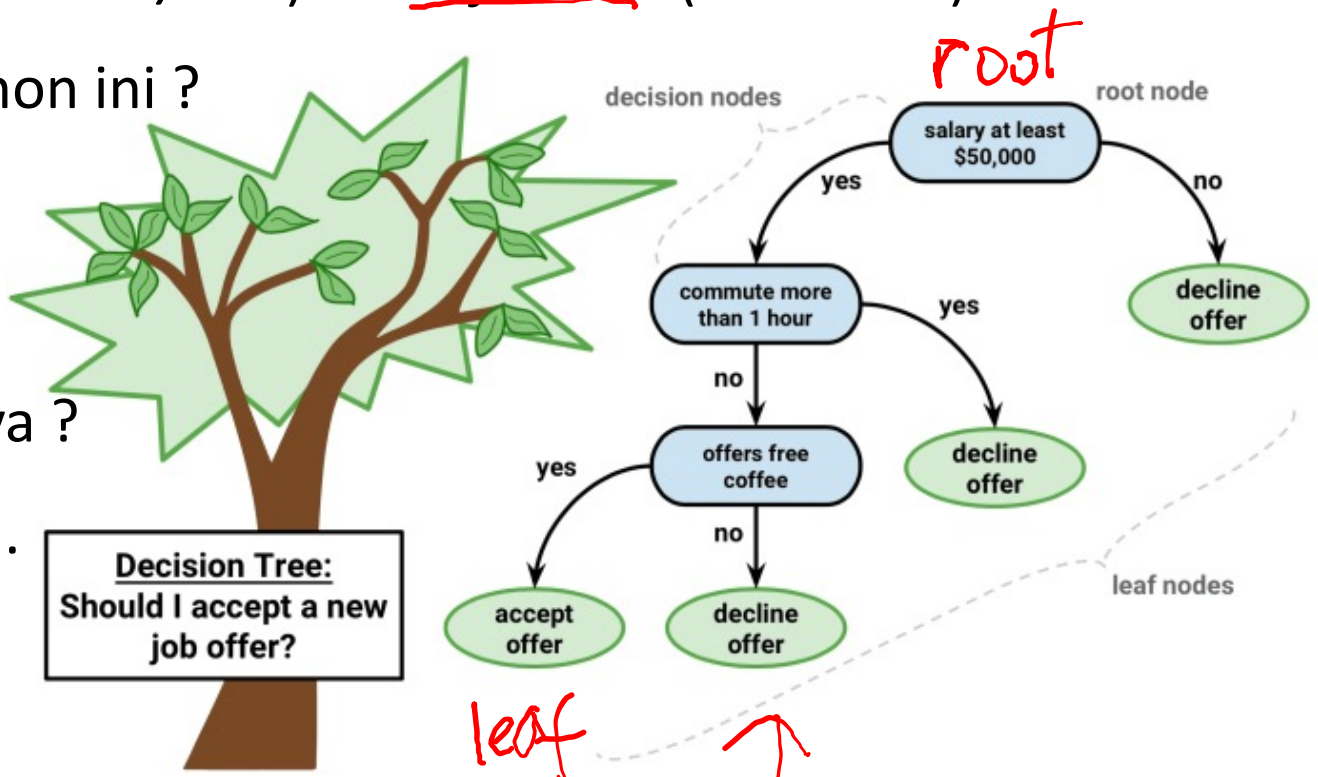
Decision Tree

- Algoritma non-parametrik
- Menggunakan sejumlah aturan berbentuk hirarki pertanyaan if/else
- Tujuan: sesedikit mungkin aturan if/else untuk menghasilkan prediksi yang akurat
- Jika depth decision tree besar→cenderung overfitting
- Tidak memerlukan feature scaling→invarian terhadap berbagai jenis ukuran



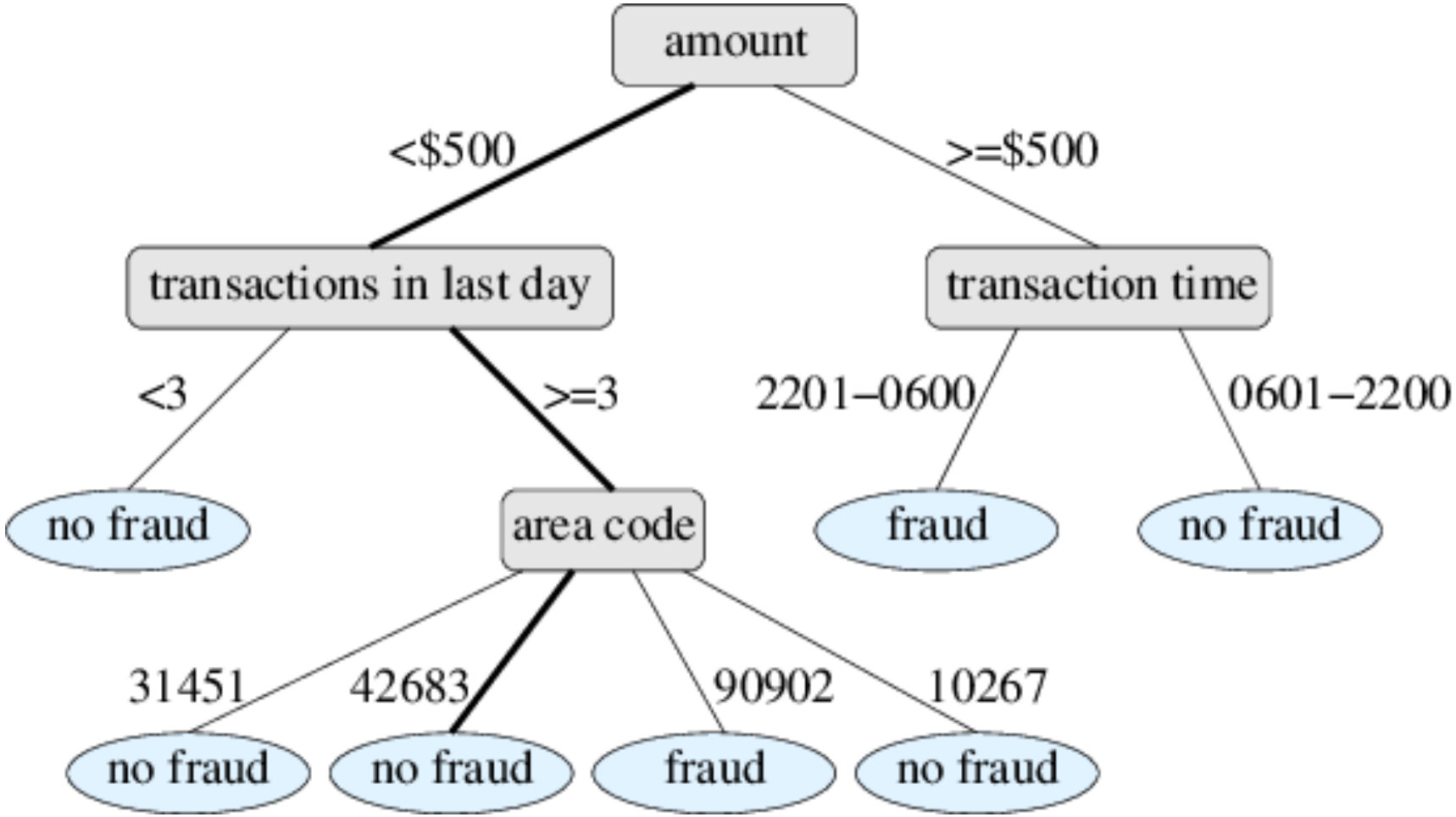
Pohon Keputusan

- Mulai dengan pertanyaan, kita membentuk pohon dengan *nodes* (*features*): *root node (starting feature)* → *branch* (*decision/rule*) → *leaf nodes* (*outcomes*)
- Apa saja *features* yang ada di contoh pohon ini ?
- Bagaimana membentuk *decision tree* ?
- Bagaimana menentukan *root node* ?
- Bagaimana menentukan *nodes* berikutnya ?
- Berbagai algoritma: CART, CHAID, ID3, dll.



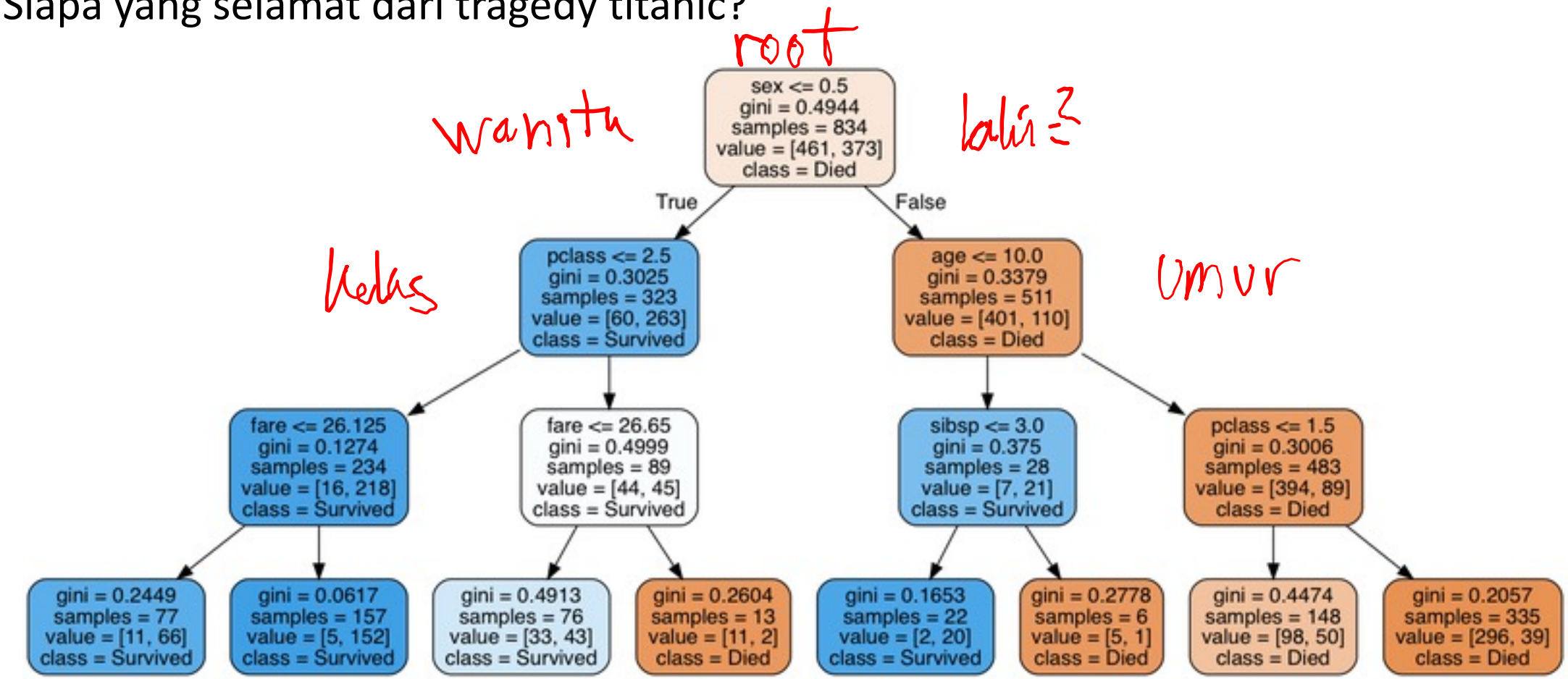
	salary	ct	fc	lubc
A	500K	2h	0	No
B	60K	0.5h	1	yes
C	70K	0.1h	1	yes
D	40K	0.3h	0	No

Contoh Pohon Keputusan untuk menentukan Fraud

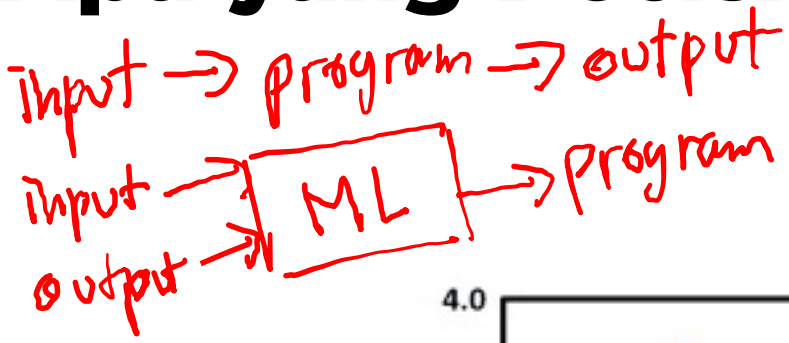


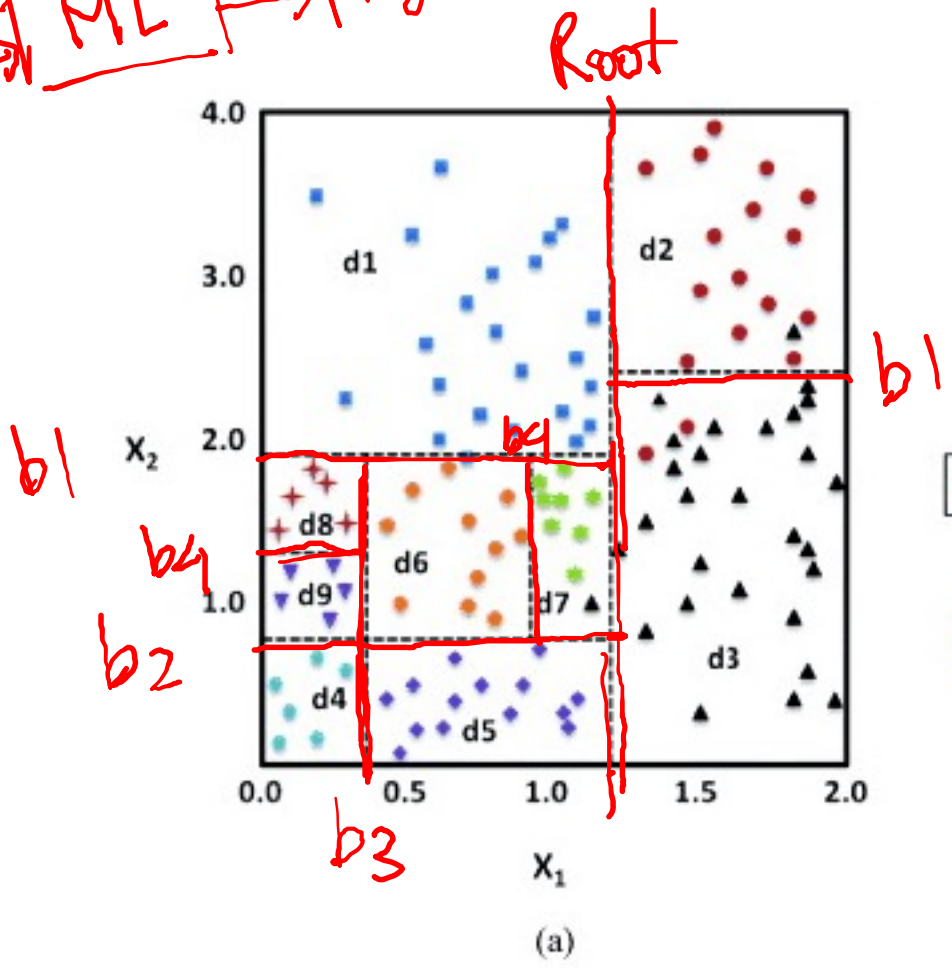
Contoh Hasil Pembelajaran Pohon Keputusan

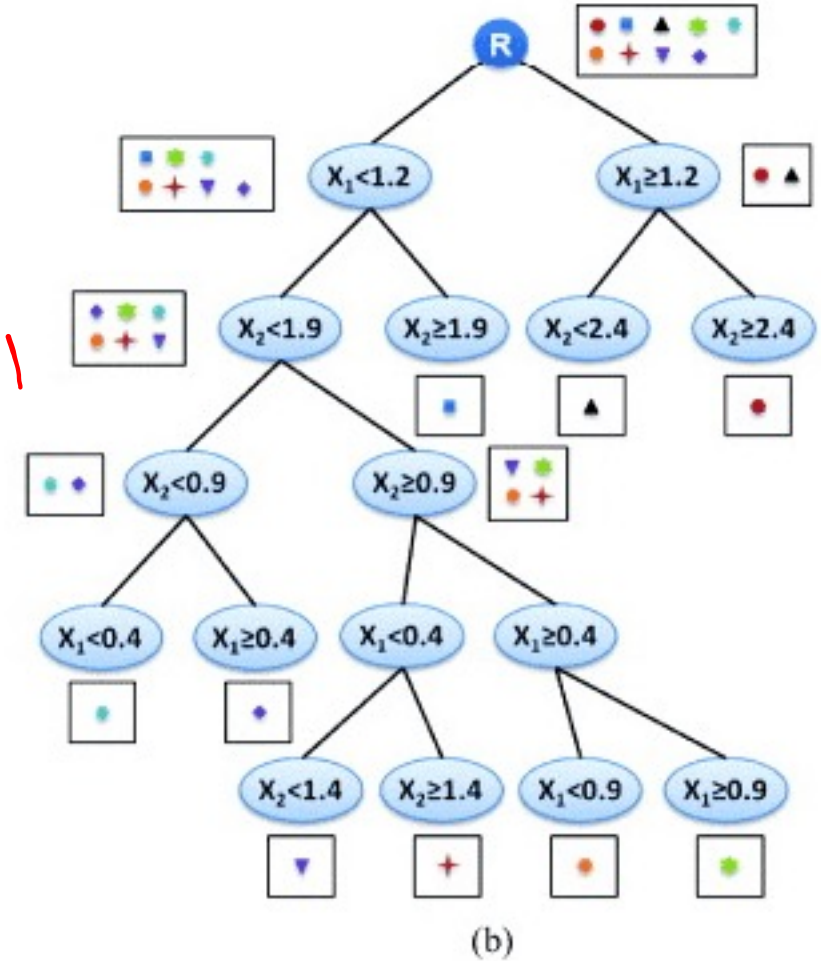
- Siapa yang selamat dari tragedy titanic?



Apa yang Decision Tree lakukan terhadap Data







root

branch 1

branch 2

branch 3

branch 4

Alternatif Algortima

Konsep ketidakmurnian (*impurity*)

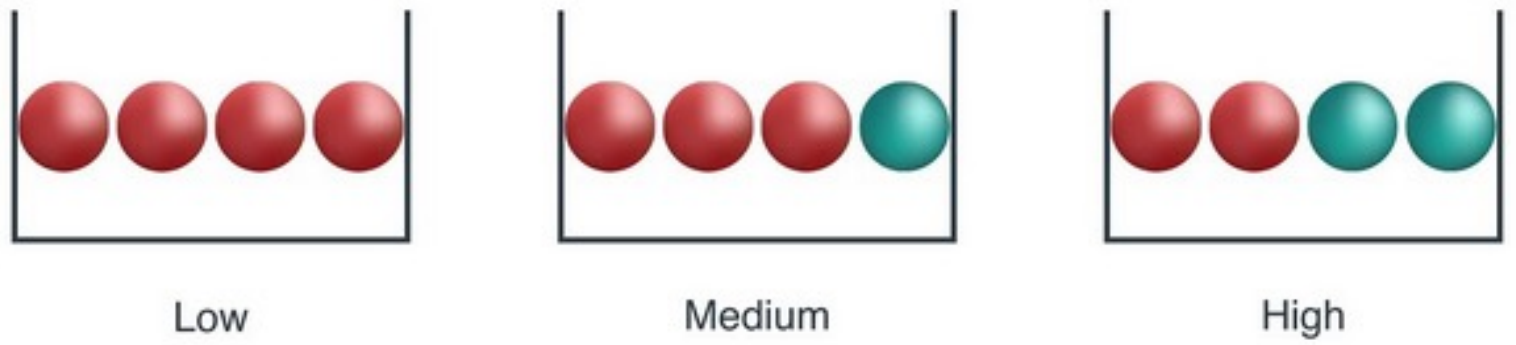
- Ilustrasi: 3 buah wadah masing-masih terdiri dari 4 bola

  Kasus 1: 1 bola diambil acak dari wadah 1. Warna apa? Merah (100%)

  Kasus 2: 1 bola diambil acak dari wadah 2. Warna apa? Merah (75%), hijau (25%)

  Kasus 3: 1 bola diambil acak dari wadah 3. Warna apa? Merah (50%), hijau (50%)

- Wadah 1 → murni (*pure node*) sedangkan Wadah 3 → most *impure*



Uji Pemahaman

- Manakah diantara ketiga kasus di atas yang paling banyak elemen “surprise”? wadah 3
- Manakah diantara ketiga kasus di atas yang paling sedikit elemen “surprise”? wadah 1

Pengukuran *impurity* – entropi

$$\int dS = \int \frac{dQ}{T}$$

- Apa itu entropi?
- Definisi termodinamika: tendensi suatu zat atau sistem menuju ketidakberaturan (*molecular disorder*)
- Definisi teori informasi (*information theory*) dari Wikipedia:

The entropy of a random variable is the average level of "information", "surprise", or "uncertainty" inherent in the variable's possible outcomes

- Contoh: pelemparan koin mata uang → kemungkinan hanya *Head or tail*, maka:

1. Jika koin tidak bias → probabilitas *head* $p_H = 1/2$, *tail* $p_T = 1 - p_H = 1/2$ (sama)
2. Jika koin itu bias → p_H dan p_T tidak sama (misal $p_H = 3/4$, $p_T = 1/4$)
3. Kasus ekstrim: $p_H = 1$, $p_T = 0$ → pasti *head*

Pengukuran *impurity* – entropi

- Apa itu entropi?

  Derajat "keterkejutan" nilai luaran dari suatu variabel random, atau
  Seberapa banyak informasi yang diperlukan untuk dapat secara akurat
  mendeskripsikan sampel data

- Semakin *impure* suatu dataset, maka informasi yang diperlukan semakin banyak (derajat keterkejutan lebih tinggi) untuk dapat menggolongkan setiap sampel dengan akurat

Pengukuran *impurity* – indeks Gini

- Derajat / ukuran ketidakseragaman (*inequality*) dalam sampel → antara 0 dan 1
- Nilai 0 → sampel homogen → semuanya dari 1 kelas
  Nilai 1 → sampel tidak seragam seluruhnya
- *Sum of squares of probabilities* dari setiap kelas (n = jumlah kelas, n = 2 : biner)

$$Indeks\ Gini(K) = \sum_{i=1}^{n} p_{i,K}\left(1 - p_{i,K}\right) = 1 - \sum_{i=1}^{n} p_{i,K}^2$$

  $p_{i,K}$ adalah probabilitas bahwa kategori K memiliki kelas i
- *Impurity* → ukuran ketidakseragaman

Pengukuran *impurity* – Chi-square

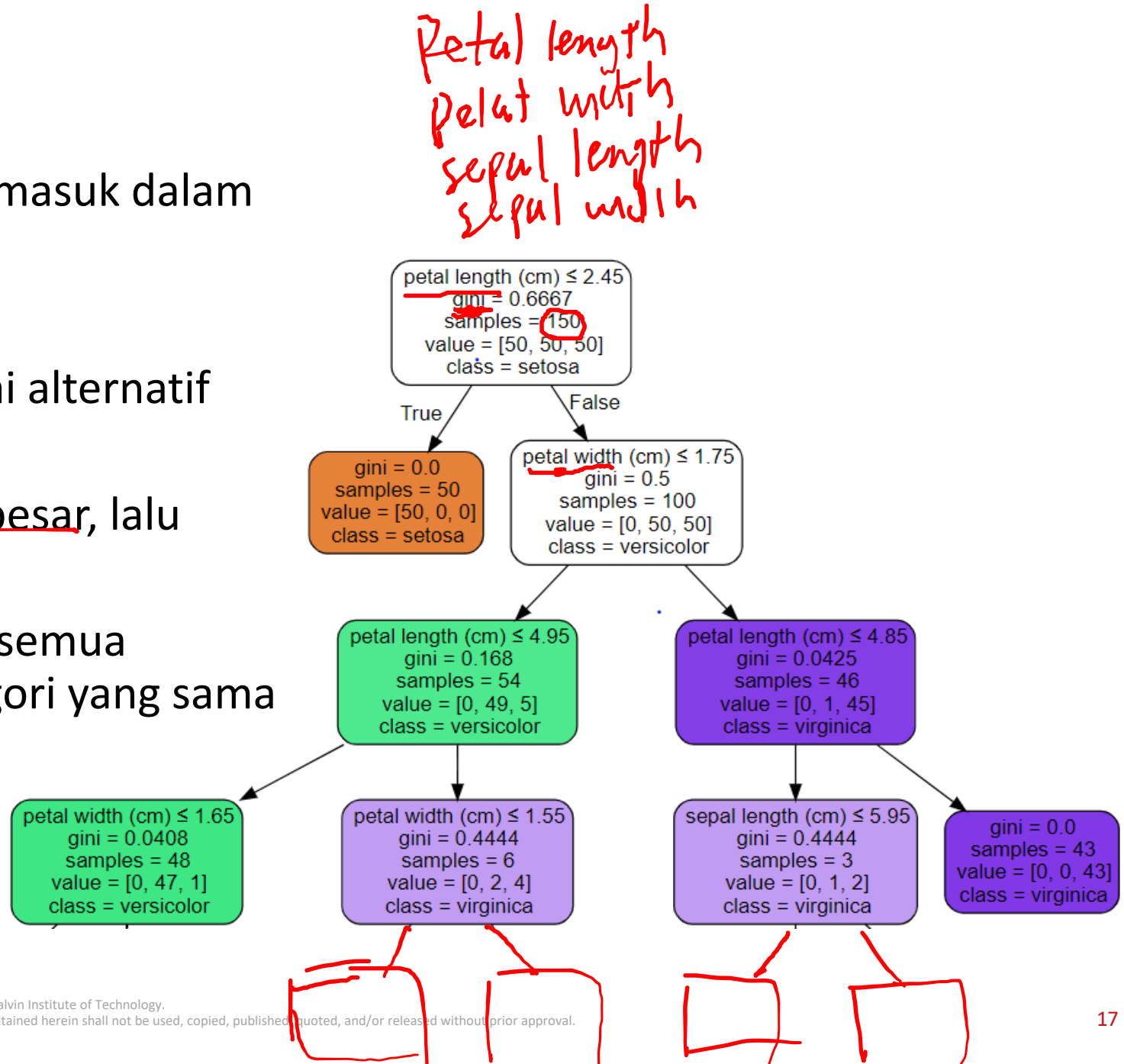
- Chi-squared (dibaca “kai squared”) merupakan contoh uji hipotesis untuk menguji independensi antara variabel kategorial.
- Chi-squared:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

	Yes	No	Total	Expected	Chi-square Yes	Chi-square No
Weak	5	2	7	3.5	0.802	0.802
Strong	3	3	6	3	0.000	0.000

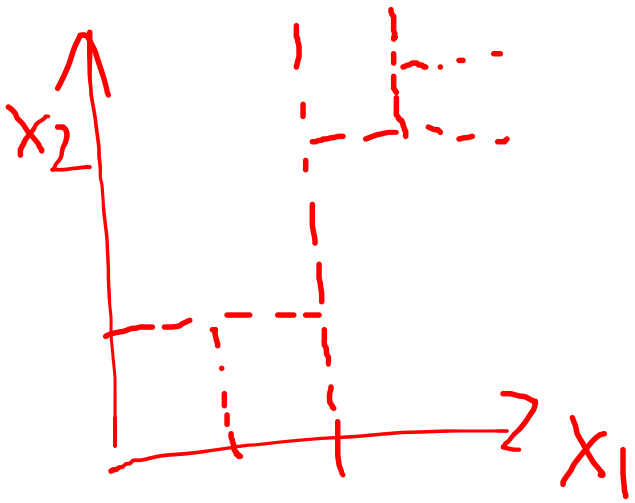
Decision Tree

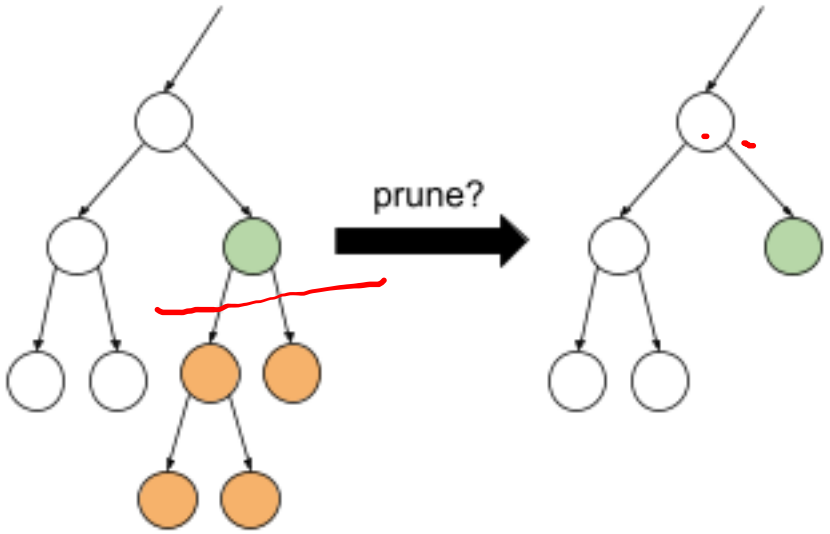
- Pada awalnya seluruh data termasuk dalam root node
- Pilih impurity measure
- Hitung impurity untuk berbagai alternatif fitur
- Pilih fitur dengan impurity terbesar, lalu lakukan splitting
- Ulangi proses splitting sampai semua anggota masuk ke dalam kategori yang sama

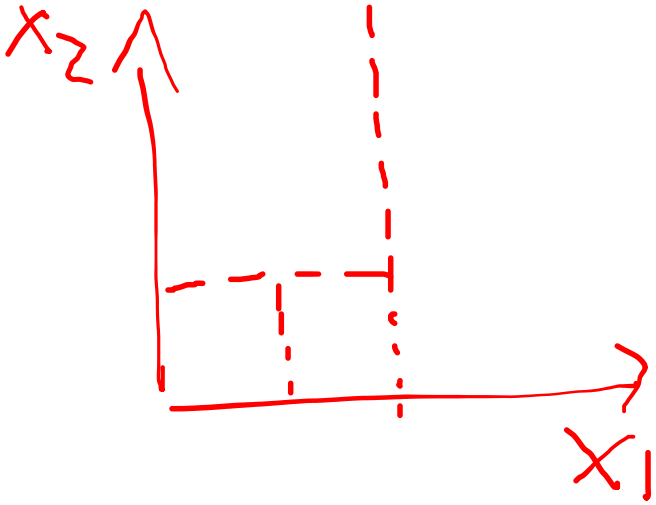


Pruning

- https://developers.google.com/machine-learning/decision-forests/overfitting-and-pruning
- Pruning adalah usaha untuk mengurangi overfitting dengan memangkas sebagian batang dan daun
- Bentuk pohon setelah pruning akan menjadi lebih sederhana







CALVIN INSTITUTE OF TECHNOLOGY

CHAID – Chi-square automatic interaction detection

- Digunakan untuk menghasilkan pohon klasifikasi dan pohon regresi
- Klasifikasi multikelas: menggunakan chi-square untuk mengevaluasi pembagian dalam pemilihan urutan fitur
- Chi-square adalah statistical measure untuk menemukan perbedaan antara child dan parent nodes

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

CALVIN INSTITUTE OF TECHNOLOGY

CART – *Classification And Regression Tree*

- Digunakan untuk menghasilkan pohon klasifikasi dan pohon regresi
- Klasifikasi biner: menggunakan indeks Gini sebagai *cost function* untuk mengevaluasi pembagian (*split*) dalam pemilihan urutan fitur
- Regresi: menggunakan *least squares* sebagai *cost function*

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

CALVIN INSTITUTE OF TECHNOLOGY

ID3 – Iterative Dichotomizer

- Singkatan dari iterative dichotomizer
- Menggunakan entropi dan *information gain* sebagai metrik

$$Entropy(S) = -\sum P(I) \times \log_2(P(I))$$

$$Information\,Gain(S, A) = Entropy(S) - \sum P(S|A) \times Entropy(S|A)$$

CALVIN INSTITUTE OF TECHNOLOGY

C4.5 & C5

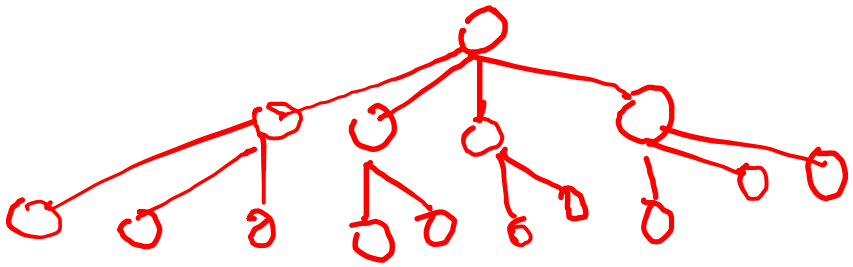
- C4.5 merupakan pengembangan dari ID3
- Jauh lebih cepat dari ID3 dan melibatkan pre-pruning
- Tidak dapat digunakan untuk boosting dan tidak dapat menerima missing value

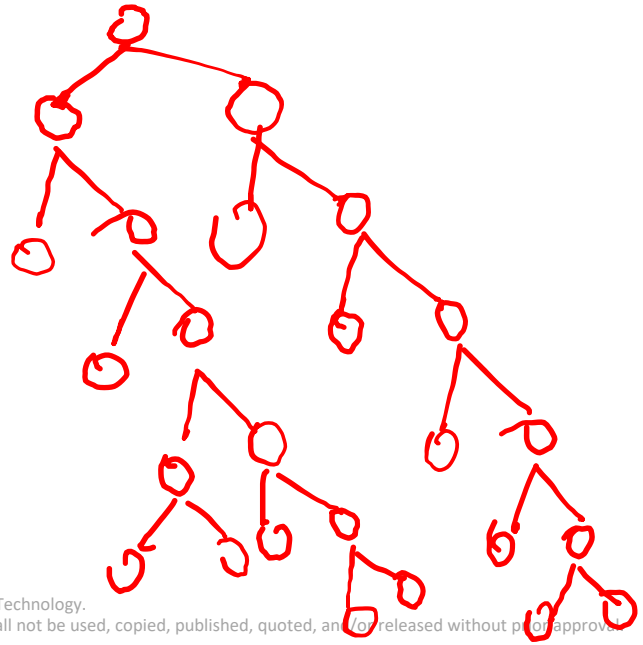
- C5 merupakan pengembangan dari C4.5
- Dapat digunakan untuk boosting dan dapat menerima missing value

CALVIN INSTITUTE OF TECHNOLOGY

Perbandingan

	CHAID	CART	ID3	C4.5 / C5
Aturan segmentasi	Chi-square	Gini	Entropy	Entropy
Aturan trimming	Otomatis	Estimasi error	Estimasi error	Estimasi error
Split	Multiple	Biner	Multiple	Multiple
Variabel	Kategori	Numerik/kategori	Kategori	Numerik/kategori
Bentuk	Melebar	Memanjang	Balance	Balance



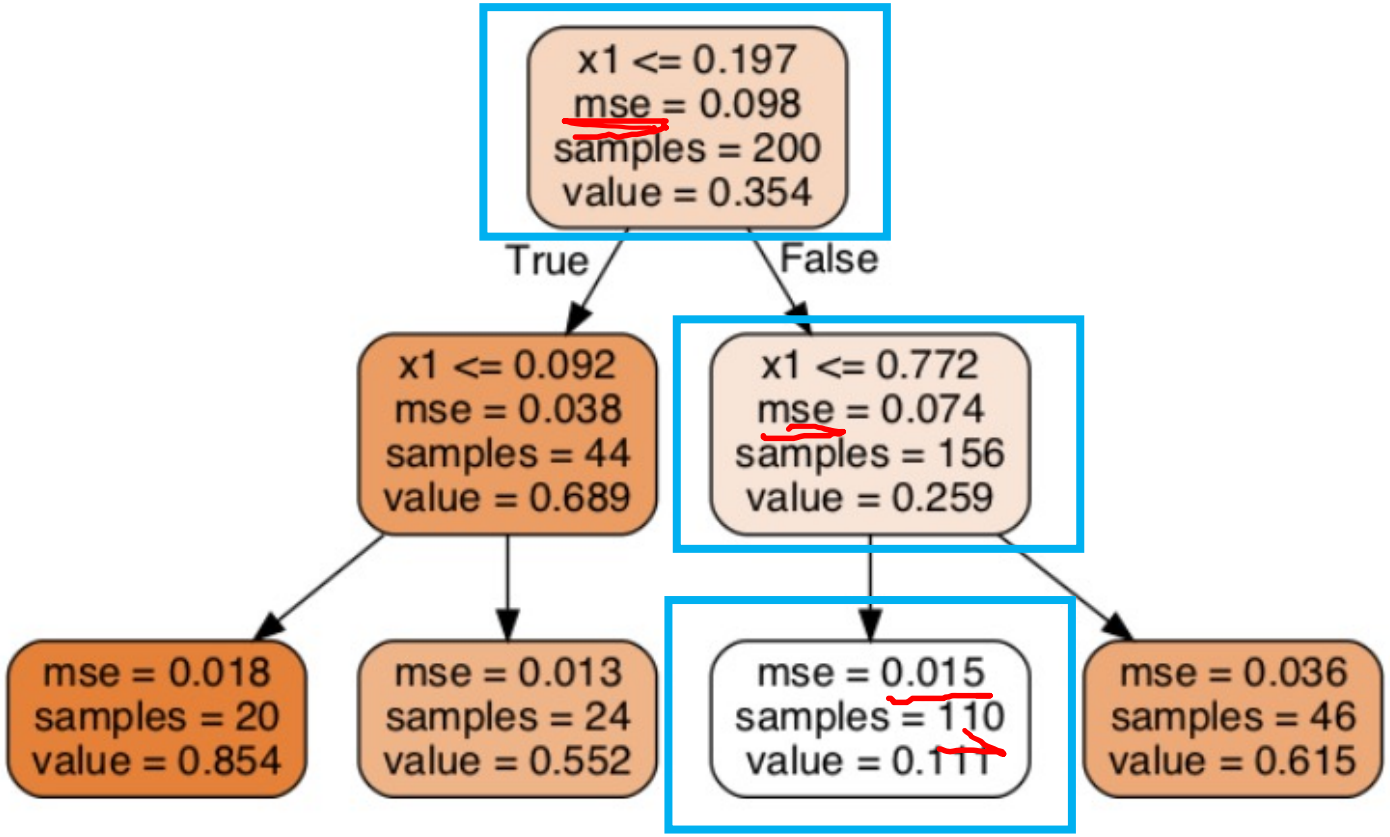


CALVIN INSTITUTE OF TECHNOLOGY

Decision Tree Regressor

Regresi dengan *decision tree*

- Luaran adalah nilai prediksi
- Model ini memiliki *depth = 2*
- Misalnya mau prediksi *output* $\hat{y}$ dari *input* x1 = 0.6
- Dari *root node* turun ke kanan
- Lalu turun ke *leaf node* dengan MSE = 0.015 dan *value* $\hat{y}$ = 0.111
- Nilai dan MSE di atas adalah hasil rerata 110 sampel
- Bagaimana menentukan *best split?*
  Goal: max. variance reduction



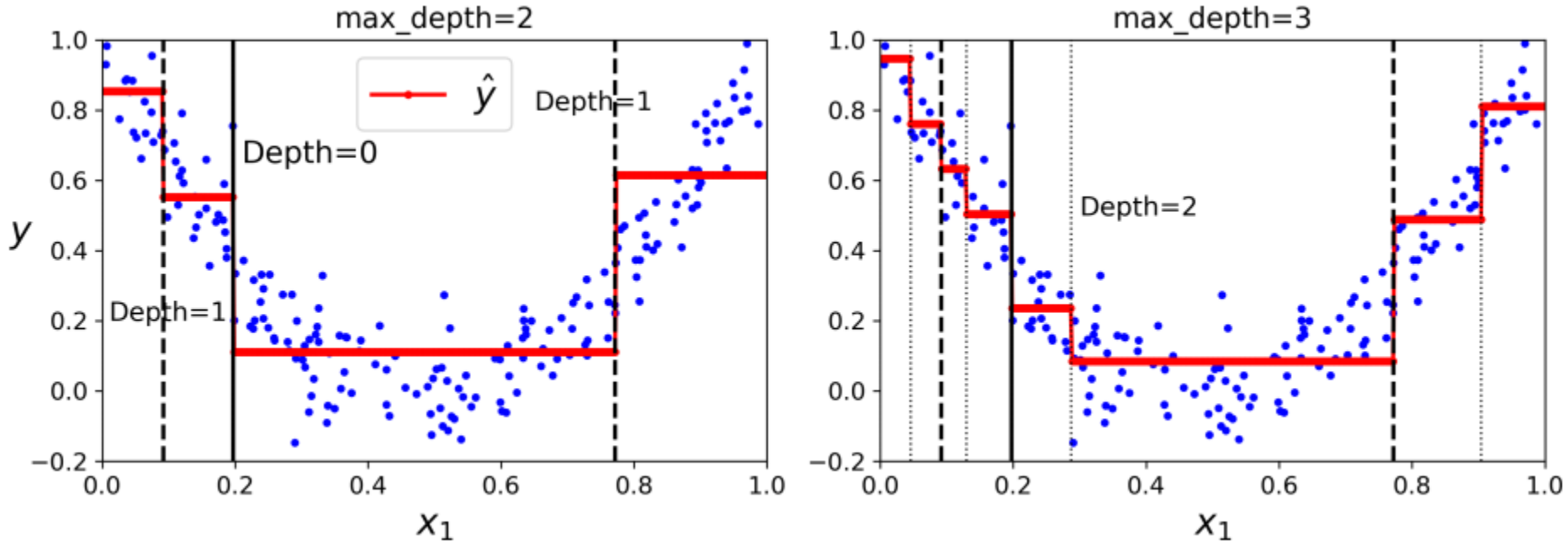
$$Var\ Reduction = Var(parent) - \sum w_i\ Var(child_i)$$

$$w_i = \frac{\#sample\ in\ child\ i}{\#sample\ in\ parent}$$

CALVIN INSTITUTE OF TECHNOLOGY

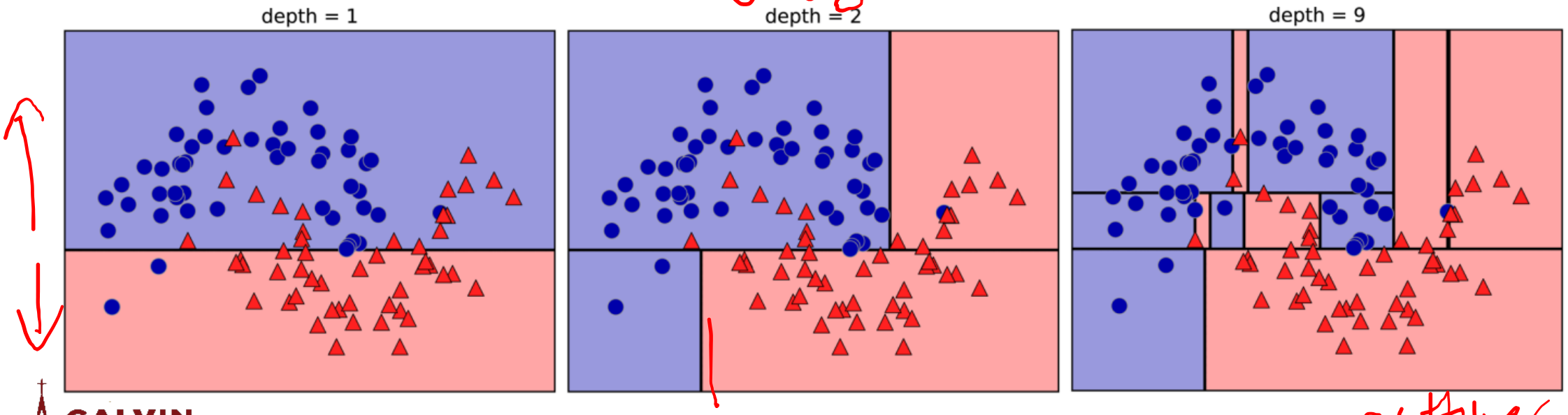
Regresi dengan *decision tree*

- Gambar di kiri: prediksi dengan *depth = 2*, sedangkan kanan : *depth = 3*
- Nilai prediksi di setiap wilayah (*region*) = rerata dari sampel latihan di wilayah itu
- Algoritma *decision tree* membagi wilayah sehingga hampir semua *training instances* berada dekat pada nilai prediksi



Kompleksitas *decision tree*

- *Pure leaf* : daun dimana semua *instances* ada di satu kategori atau *target value*
- Melatih pohon sampai semua daun *pure* → model sangat kompleks → overfit
- Regularisasi *pre-pruning (early stopping)* dengan *hyperparameters* seperti *decrease max. depth*, *max. no. of leaf nodes*, *max. no. of features evaluated for splitting*, *increase min. no. of samples in leaf*, *min. no. of samples in node to split*
- Regularisasi *post-pruning* (memangkas *nodes* tanpa menaikkan *error* secara signifikan)*: reduced error pruning, cost complexity pruning* (https://en.wikipedia.org/wiki/Decision_tree_pruning)



overfitting

Tuhan Memberkati