Minggu ke-6

Decision Tree

Ali Ridho Barakbah, Entin Martiana

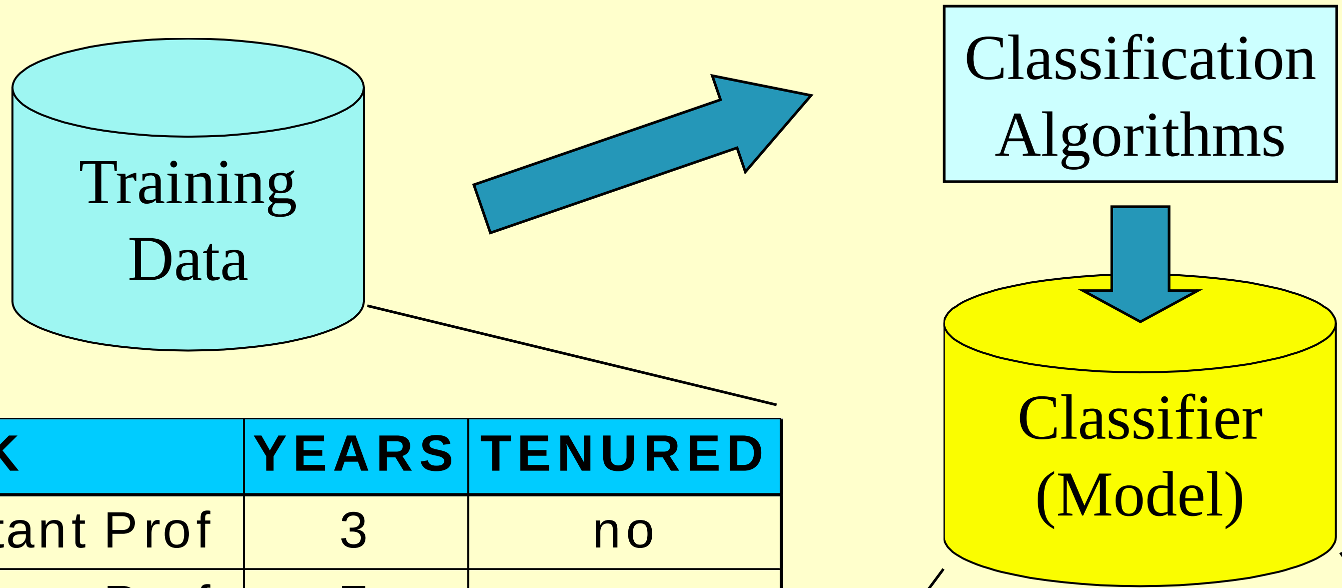
Knowledge Engineering Research Group
Department of Information and Computer Engineering
Politeknik Elektronika Negeri Surabaya

What is a Decision Tree?

- An *inductive learning task*
  - Use particular facts to make more generalized conclusions

- A predictive model based on a branching series of Boolean tests
  - These smaller Boolean tests are less complex than a one-stage classifier

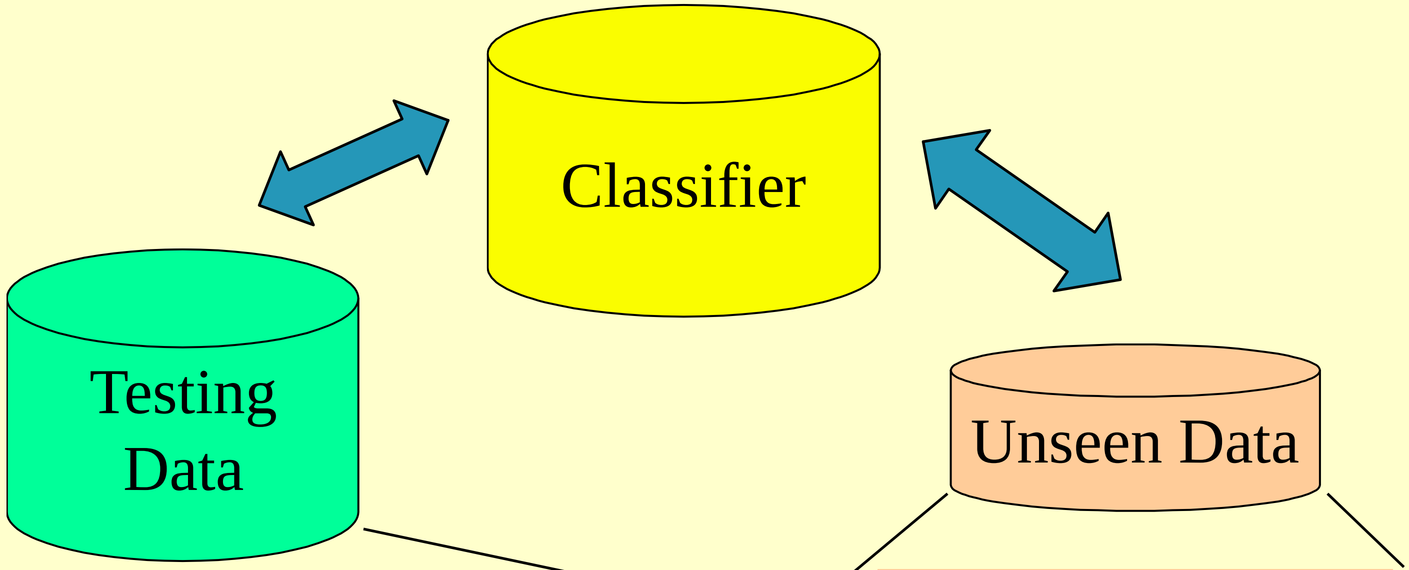
Process (1): Model Construction



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

IF rank = 'professor'
OR years > 6
THEN tenured = 'yes'

Process (2): Using the Model in Prediction



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

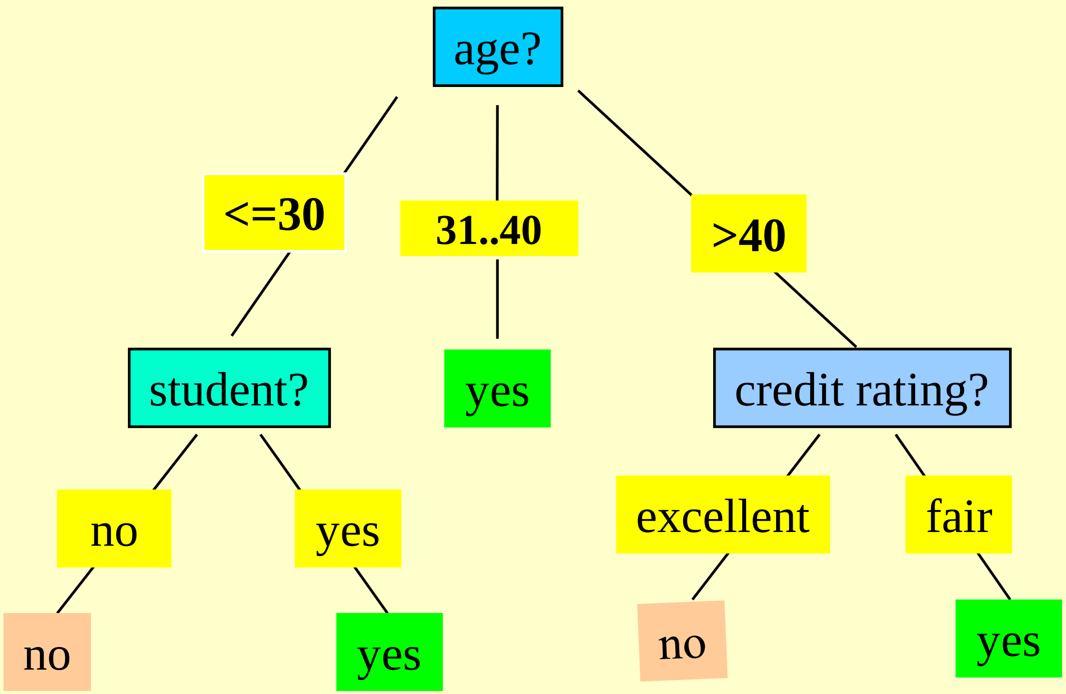
(Jeff, Professor, 4)

Tenured?

Yes

Decision Tree Induction: An Example

- Training data set: Buys_computer
- The data set follows an example of Quinlan's ID3 (Playing Tennis)
- Resulting tree:

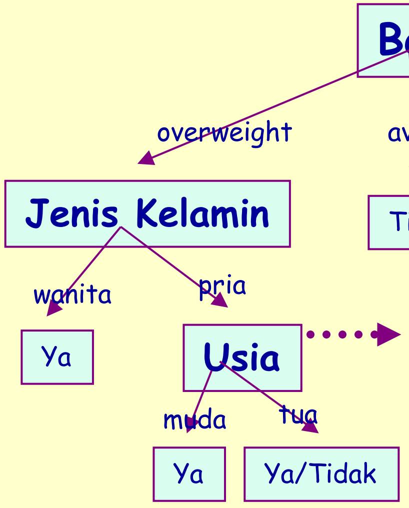


age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent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

Concept of Decision Tree

Data → **Decision Tree** → **Rule**

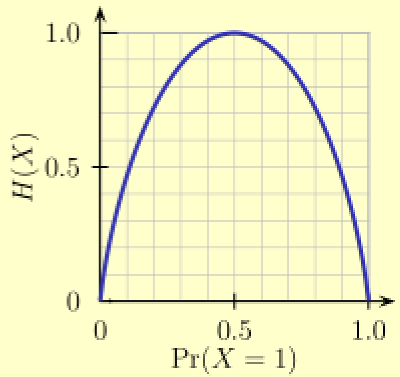
Nama	Usia	Berat	Kelamin	Hipertensi
Ali	muda	overweight	pria	ya
Edi	muda	underweight	pria	tidak
Annie	muda	average	wanita	tidak
Budiman	tua	overweight	pria	tidak
Herman	tua	overweight	pria	ya
Didi	muda	underweight	pria	tidak
Rina	tua	overweight	wanita	ya
Gatot	tua	average	pria	tidak



R1: IF berat=average v berat=underweight THEN hipertensi=tidak
R2: IF berat=overweight^kelamin=wanita THEN hipertensi=ya
R3: IF berat=overweigt^kelamin=pria^ usia=muda THEN hipertensi=ya
R4: IF berat=overweigt^kelamin=pria^ usia=tua THEN hipertensi=tidak

Brief Review of Entropy

- Entropy (Information Theory)
  - A measure of uncertainty associated with a random variable
  - Calculation: For a discrete random variable Y taking m distinct values $\{y_1, \ldots, y_m\}$,
    - $H(Y) = -\sum_{i=1}^{m} p_i \log(p_i)$, where $p_i = P(Y = y_i)$
  - Interpretation:
    - Higher entropy => higher uncertainty
    - Lower entropy => lower uncertainty
- Conditional Entropy
  - $H(Y|X) = \sum_x p(x) H(Y|X = x)$



m = 2

Example: Training Data

Nama	Usia	Berat	Kelamin	Hipertensi
Ali	muda	overweight	pria	ya
Edi	muda	underweight	pria	tidak
Annie	muda	average	wanita	tidak
Budiman	tua	overweight	pria	tidak
Herman	tua	overweight	pria	ya
Didi	muda	underweight	pria	tidak
Rina	tua	overweight	wanita	ya
Gatot	tua	average	pria	tidak

Entropy untuk Usia

Usia	Hipertensi	Jumlah
muda	Ya (+)	1
muda	Tidak (-)	3
tua	ya	2
tua	tidak	2

Usia = muda

$$q_1 = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} = 0.81$$

Usia = tua

$$q_2 = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$$

Entropy untuk Usia

$$\boxed{E = \frac{4}{8}q_1 + \frac{4}{8}q_2 = \frac{4}{8}(0.81) + \frac{4}{8}(1) = 0.91}$$

Memilih Node Awal

Usia	Hipertensi	Jumlah
muda	ya	1
muda	tidak	3
tua	ya	2
tua	tidak	2

Entropy = 0.91

Berat	Hipertensi	Jumlah
overweight	ya	3
overweight	tidak	1
average	ya	0
average	tidak	2
underweight	ya	0
underweight	tidak	2

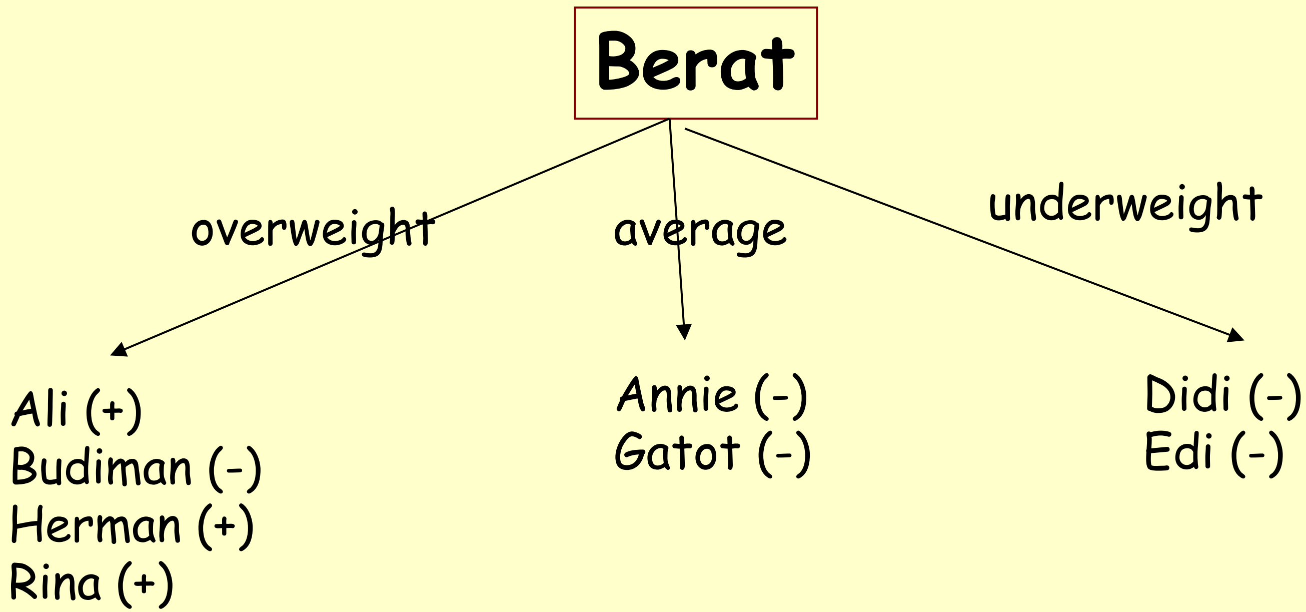
Entropy = 0.41

Kelamin	Hipertensi	Jumlah
pria	ya	2
pria	tidak	4
wanita	ya	1
wanita	tidak	1

Entropy = 0.94

Terpilih atribut BERAT BADAN sebagai node awal karena memiliki entropy terkecil

Penyusunan Tree Awal



Penentuan Leaf Node Untuk
Berat=Overweight

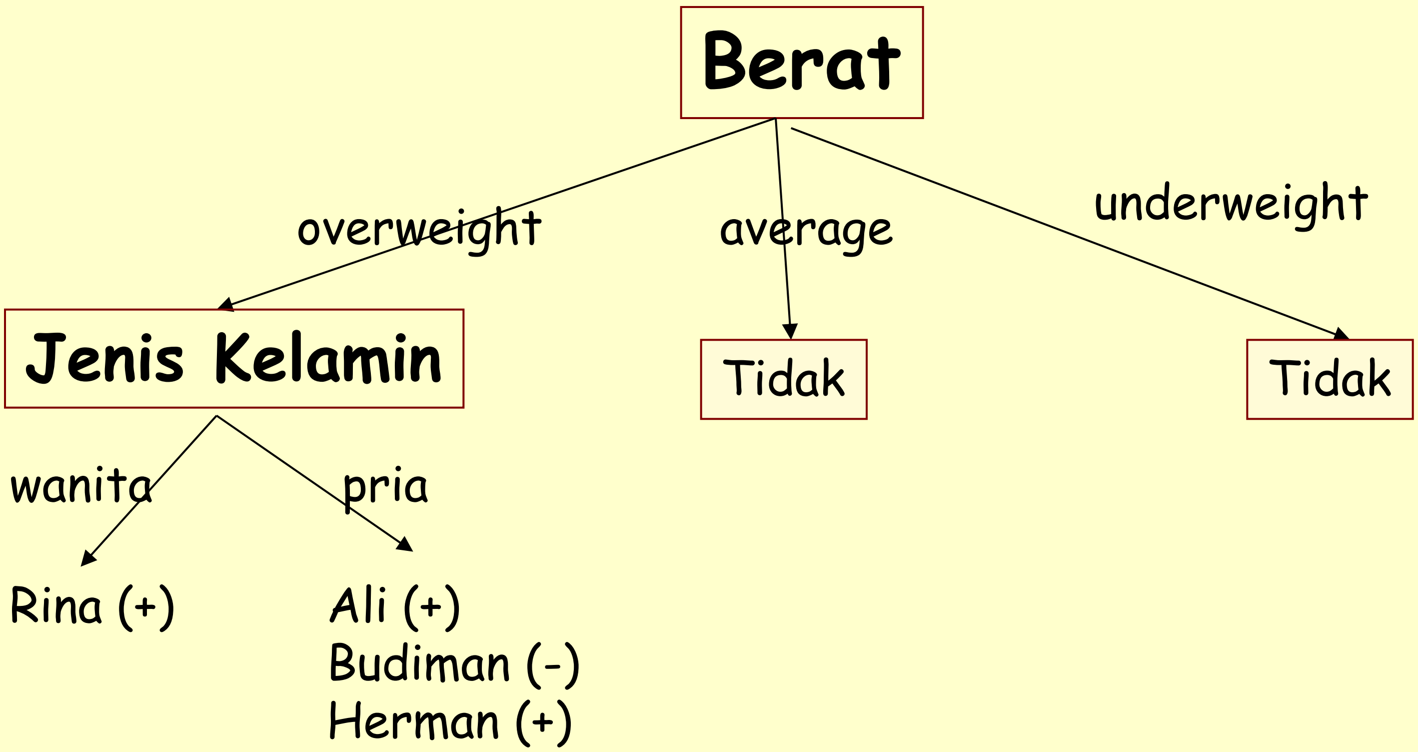
Data Training untuk berat=overweight

Nama	Usia	Kelamin	Hipertensi
Ali	muda	pria	ya
Budiman	tua	pria	tidak
Herman	tua	pria	ya
Rina	tua	wanita	ya

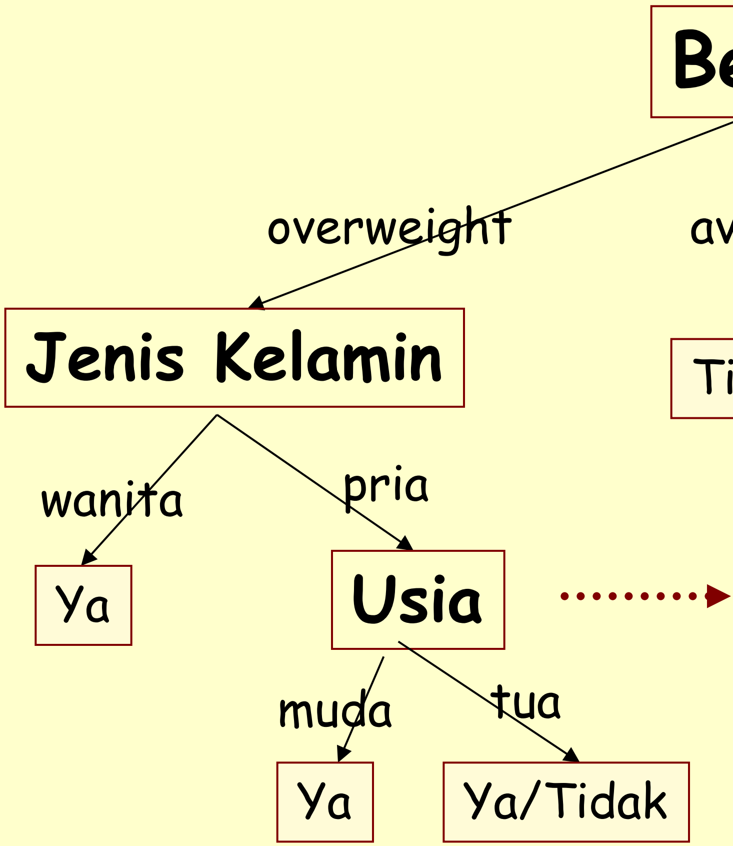
Usia	Hipertensi	Jumlah
muda	ya	1
	tidak	0
tua	ya	2
	tidak	1
	Entropy =	0,69

Kelamin	Hipertensi	Jumlah
pria	ya	2
	tidak	1
wanita	ya	1
	tidak	0
	Entropy =	0,69

Penyusunan Tree

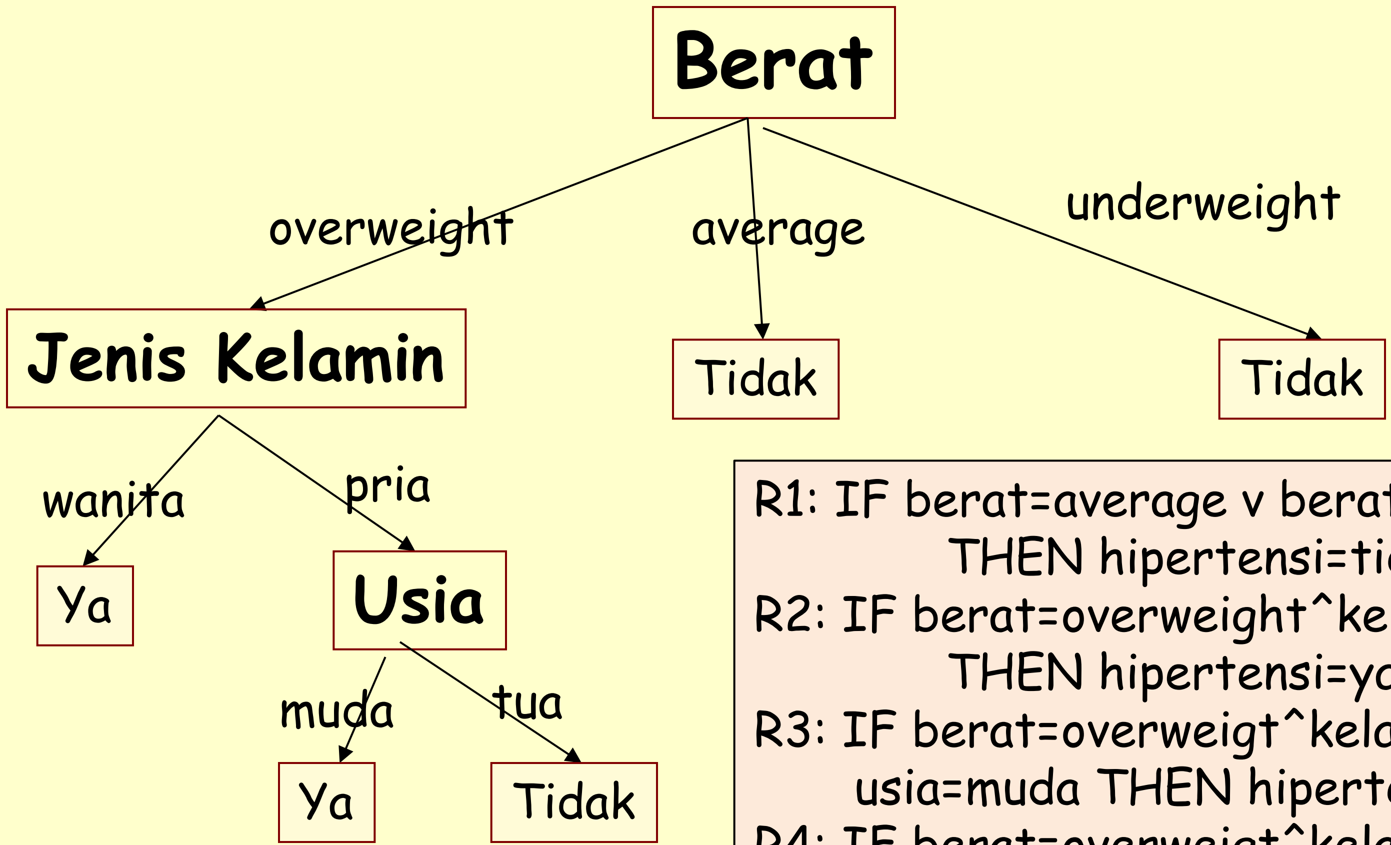


Hasil Tree



Nama	Usia	Kelamin	Hipertensi
Ali	muda	pria	ya
Budiman	tua	pria	tidak
Herman	tua	pria	ya

Mengubah Tree Menjadi Rule



R1: IF berat=average v berat=underweight
THEN hipertensi=tidak

R2: IF berat=overweight^kelamin=wanita
THEN hipertensi=ya

R3: IF berat=overweigt^kelamin=pria^
usia=muda THEN hipertensi=ya

R4: IF berat=overweigt^kelamin=pria^
usia=tua THEN hipertensi=tidak

Konversi Numerical Attribute ke Categorical Attibute (dengan Gini Index)

- If a data set D contains examples from n classes, gini index, $gini(D)$ is defined as

$$gini(D) = 1 - \sum_{j=1}^{n} p_j^2$$

where p_j is the relative frequency of class j in D

- If a data set D is split on A into two subsets D_1 and D_2, the *gini* index $gini(D)$ is defined as

$$gini_A(D) = \frac{|D_1|}{|D|} gini(D_1) + \frac{|D_2|}{|D|} gini(D_2)$$

- Reduction in Impurity:

$$\Delta gini(A) = gini(D) - gini_A(D)$$

- The attribute provides the smallest $gini_{split}(D)$ (or the largest reduction in impurity) is chosen to split the node (*need to enumerate all the possible splitting points for each attribute*)

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

Computation of Gini Index

- Ex. D has 9 tuples in buys_computer = "yes" and 5 in "no"

$$gini(D) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.459$$

- Suppose the attribute income partitions D into 10 in D_1: {low, medium} and 4 in D_2

$$gini_{income \in \{low,medium\}}(D) = \left(\frac{10}{14}\right) Gini(D_1) + \left(\frac{4}{14}\right) Gini(D_2)$$

$$= \frac{10}{14}\left(1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2\right) + \frac{4}{14}\left(1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right)$$

$$= 0.443$$

$$= Gini_{income \in \{high\}}(D).$$

$Gini_{\{low,high\}}$ is 0.458; $Gini_{\{medium,high\}}$ is 0.450. Thus, split on the {low,medium} (and {high}) since it has the lowest Gini index

- All attributes are assumed continuous-valued
- May need other tools, e.g., clustering, to get the possible split values
- Can be modified for categorical attributes

Contoh

$$Gini(Dataset) = 1 - \left(\frac{3}{8}\right)^2 - \left(\frac{5}{8}\right)^2 = 0.46875$$

#	Usia	Berat Badan	Jenis Kelamin	Hipertensi
1	22	overweight	pria	ya
2	27	underweight	pria	tidak
3	31	average	wanita	tidak
4	46	overweight	pria	tidak
5	59	overweight	pria	ya
6	23	underweight	pria	tidak
7	48	overweight	wanita	ya
8	43	average	pria	tidak

Usia 22 23 27 31 43 46 48 59

$Split_1$ 22 | 23 27 31 43 46 48 59

A B

$$Gini_{Split1} = \frac{1}{8} Gini_A + \frac{7}{8} Gini_B$$

$$= \frac{1}{8}[1 - (\frac{1}{1})^2 - (\frac{0}{1})^2] + \frac{7}{8}[1 - (\frac{2}{7})^2 - (\frac{5}{7})^2] = 0.4$$

$$\Delta Gini(Usia) = Gini(Usia) - Gini_{Split1}$$
$$= 0.46875 - 0.4$$
$$= 0.06875$$

Attribute Selection Measure dengan Gini Index

Split$_1$ 22 | 23 27 31 43 46 48 59 → 0.4

A B

Split$_2$ 22 23 | 27 31 43 46 48 59 → ?

A B

Split$_3$ 22 23 27 | 31 43 46 48 59 → ?

A B

⋮

Split$_n$ 22 23 27 31 43 46 48 | 59 → ?

A B

Pilih yang terkecil