Decision Tree

Achmad Basuki, Iwan Syarif
Politeknik Elektronika Negeri Surabaya
PENS-ITS 2003

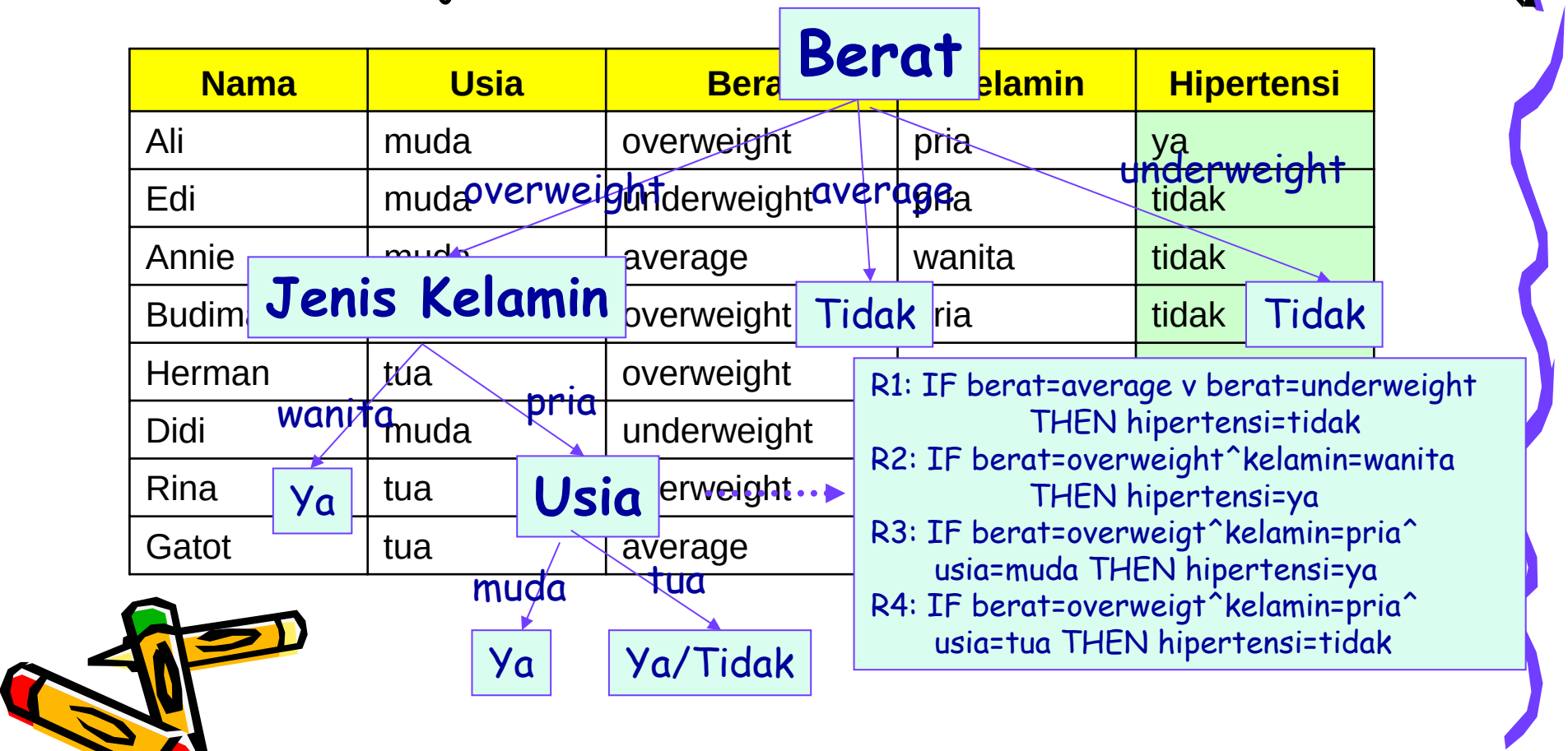
Konsep Decision Tree

Mengubah data menjadi pohon keputusan (*decision tree*) dan aturan-aturan keputusan (*rule*)

Gambaran Pemakaian Decision Tree

Membuat aturan (rule) yang dapat digunakan untuk menentukan apakah seseorang mempunyai potensi untuk menderita hipertensi atau tidak berdasarkan data usia, berat badan dan jenis kelamin.

Nama	Usia	Bera	elamin	Hipertensi
Ali	muda	overweight	pria	ya
Edi	muda	underweight	pria	tidak
Annie	mud	average	wanita	tidak
Budim		overweight	ria	tidak
Herman	tua	overweight		
Didi	muda	underweight		
Rina	tua	erweight		
Gatot	tua	average		



Beberapa contoh pemakaian Decision Tree

- Diagnosa penyakit tertentu, seperti hipertensi, kanker, stroke dan lain-lain
- Pemilihan produk seperti rumah, kendaraan, komputer dan lain-lain
- Pemilihan pegawai teladan sesuai dengan kriteria tertentu
- Deteksi gangguan pada komputer atau jaringan komputer seperti Deteksi Entrusi, deteksi virus (trojan dan varians)
- Masih banyak lainnya.

Konsep Data Dalam Decision Tree

- Data dinyatakan dalam bentuk tabel dengan atribut dan record.
- **Atribut** menyatakan suatu parameter yang dibuat sebagai kriteria dalam pembentukan tree. Misalkan untuk menentukan main tenis, kriteria yang diperhatikan adalah cuaca, angin dan temperatur. Salah satu atribut merupakan atribut yang menyatakan data solusi per-item data yang disebut dengan **target atribut**.
- Atribut memiliki nilai-nilai yang dinamakan dengan **instance**. Misalkan atribut cuaca mempunyai instance berupa cerah, berawan dan hujan.

Konsep Data Dalam Decision Tree *(Cont...)*

Nama	Cuaca	Angin	Temperatur	Main
Ali	cerah	keras	panas	tidak
Budi	cerah	lambat	panas	ya
Heri	berawan	keras	sedang	tidak
Irma	hujan	keras	dingin	tidak
Diman	cerah	lambat	dingin	ya

Sample

attribut

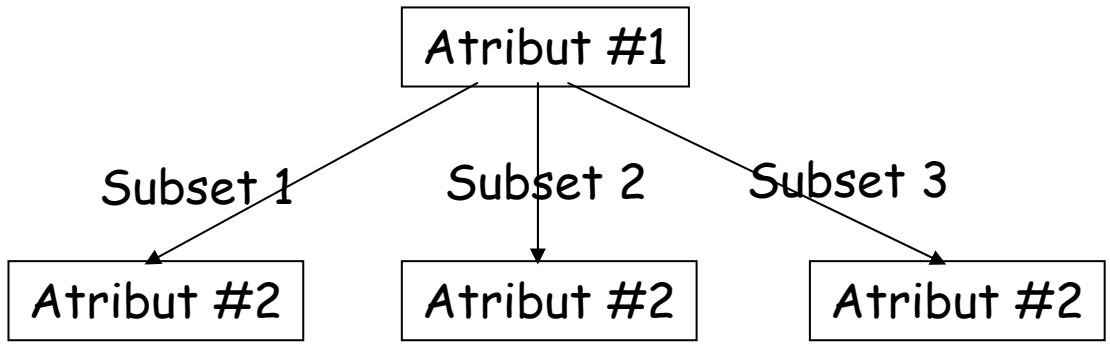
Target atribut

Proses Dalam Decision Tree

- Mengubah bentuk data (tabel) menjadi model tree.
- Mengubah model tree menjadi rule
- Menyederhanakan Rule (Pruning)

Proses Data Menjadi Tree

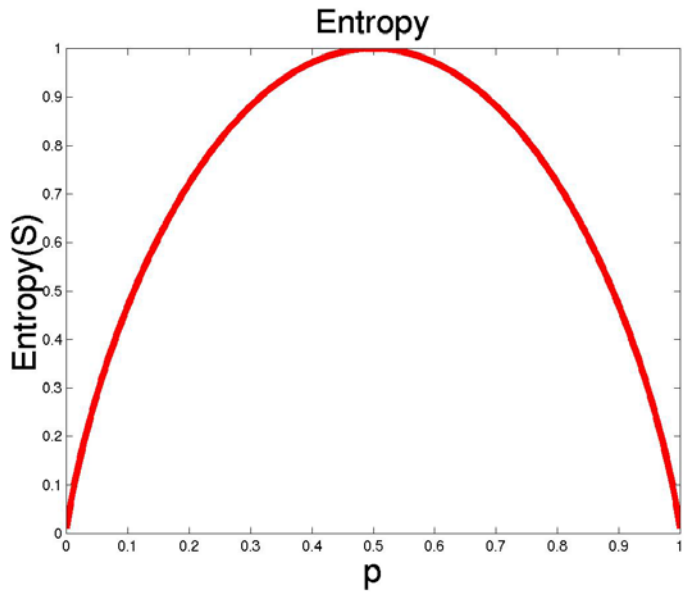
Indentity Atribut	Atribut 1	Atribut 2	Atribut 3	.....	Atribut n	Target Atribut



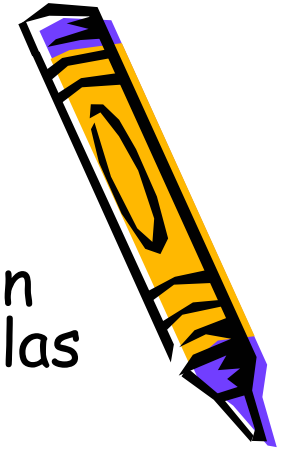
Entropy

- S adalah ruang (data) sample yang digunakan untuk training.
- P+ adalah jumlah yang bersolusi positif (mendukung) pada data sample untuk kriteria tertentu.
- P+ adalah jumlah yang bersolusi negatif (tidak mendukung) pada data sample untuk kriteria tertentu.
- Besarnya Entropy pada ruang sample S didefinisikan dengan:

$$Entropy(S) = -p_+ \log_2 p_+ - p_- \log_2 p_-$$



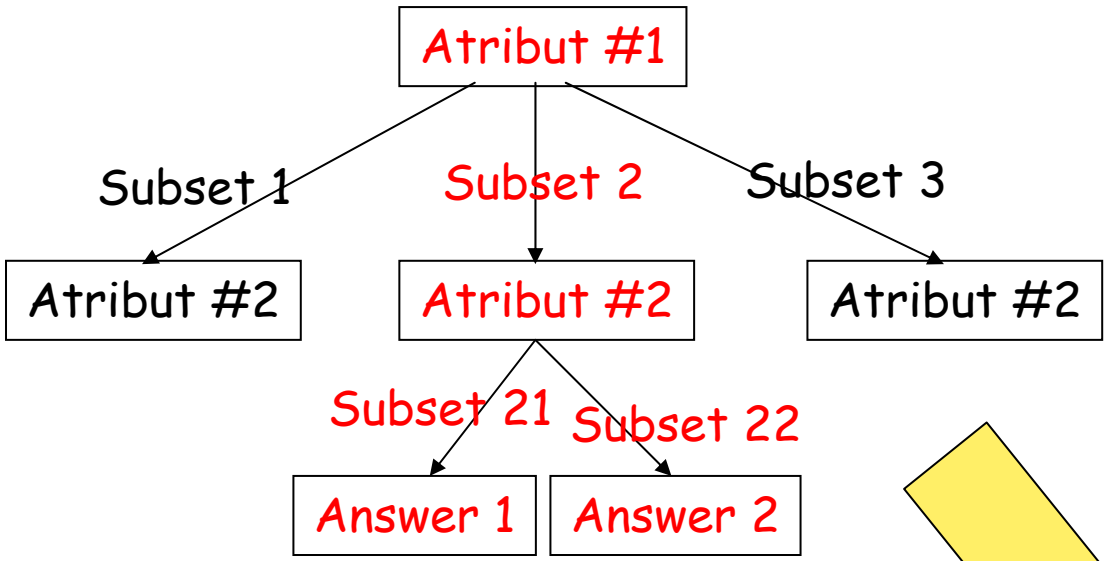
Definisi Entropy



- Entropy(S) adalah jumlah bit yang diperkirakan dibutuhkan untuk dapat mengekstrak suatu kelas (+ atau -) dari sejumlah data acak pada ruang sample S.
- Entropy bisa dikatakan sebagai kebutuhan bit untuk menyatakan suatu kelas. Semakin kecil nilai Entropy maka semakin baik untuk digunakan dalam mengekstraksi suatu kelas.
- Panjang kode untuk menyatakan informasi secara optimal adalah $-\log_2 p$ bits untuk messages yang mempunyai probabilitas p.
- Sehingga jumlah bit yang diperkiraank untuk mengekstraksi S ke dalam kelas adalah:

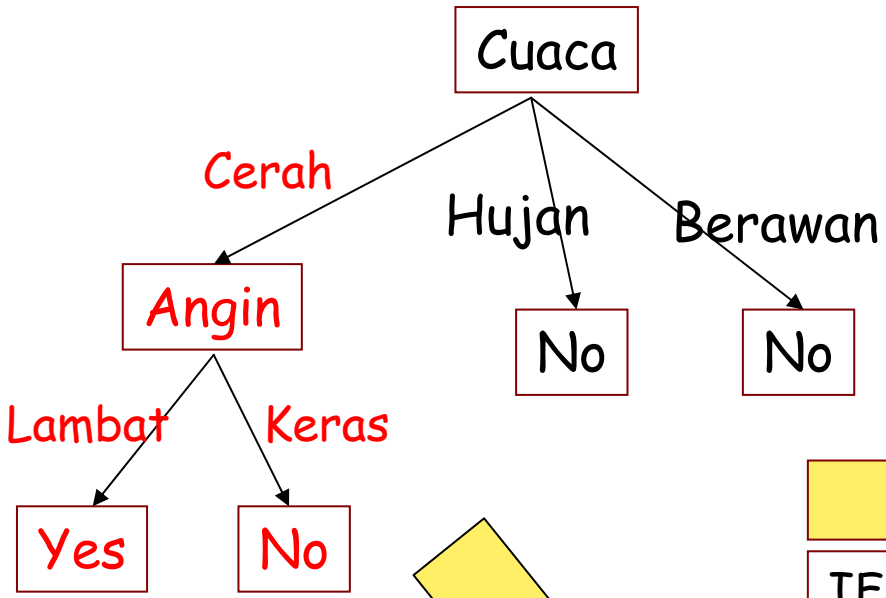
$$-p_+ \log_2 p_+ - p_- \log_2 p_-$$

Mengubah Tree Menjadi Rules



If atribut#1=subset2 ^ atribut#2=subset21
then answer=answer1
If atribut#1=subset2 ^ atribut#2=subset22
then answer=answer2

Conjunction & Disjunction



Disjunction v

IF cuaca=hujan v cuaca=berawan THEN
MainTenis=No

Conjunction ^

IF cuaca=cerah ^ angin=lambat THEN
MainTenis=Yes
IF cuaca=cerah ^ angin=keras THEN
MainTenis=No

Contoh Permasalahan Penentuan Seseorang Menderita Hipertensi Menggunakan Decision Tree

Data diambil dengan 8 sample, dengan pemikiran bahwa yang memperngaruhi seseorang menderita hipertensi atau tidak adalah usia, berat badan, dan jenis kelamin.

Usia mempunyai instance:
muda dan tua
Berat badan mempunyai instance:
underweight, average dan overweight
Jenis kelamin mempunyai instance:
pria dan wanita

Data Sample yang Digunakan Untuk Menentukan Hipertensi

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

Langkah Mengubah Data Menjadi Tree

- Menentukan Node Terpilih
- Menyusun Tree

Menentukan Node Terpilih

- Untuk menentukan node terpilih, gunakan nilai Entropy dari setiap kriteria dengan data sample yang ditentukan.
- Node terpilih adalah kriteria dengan Entropy yang paling kecil.

Memilih Node Awal

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:

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

Memilih Node Awal (cont)

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

Entropy = 0.91

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

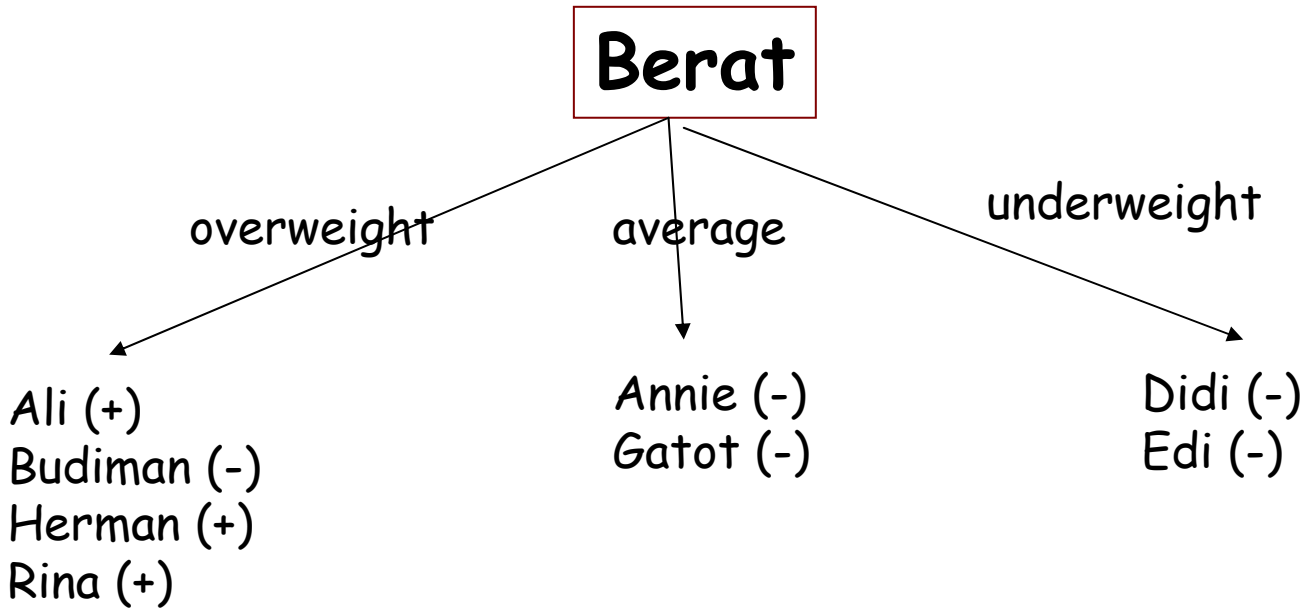
Entropy = 0.94

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

Entropy = 0.41

Terpilih atribut BERAT BADAN sebagai node awal karena memiliki entropy terkecil

Penyusunan Tree Awal



Leaf Node berikutnya dapat dipilih pada bagian yang mempunyai nilai + dan -, pada contoh di atas hanya berat=overweight yang mempunyai nilai + dan – maka semuanya pasti mempunya leaf node. Untuk menyusun leaf node lakukan satu-persatu.

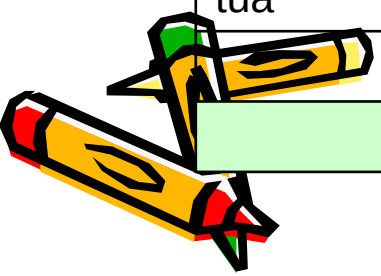
Penentuan Leaf Node Untuk Berat=Overweight

Data Training untuk berat=overweight

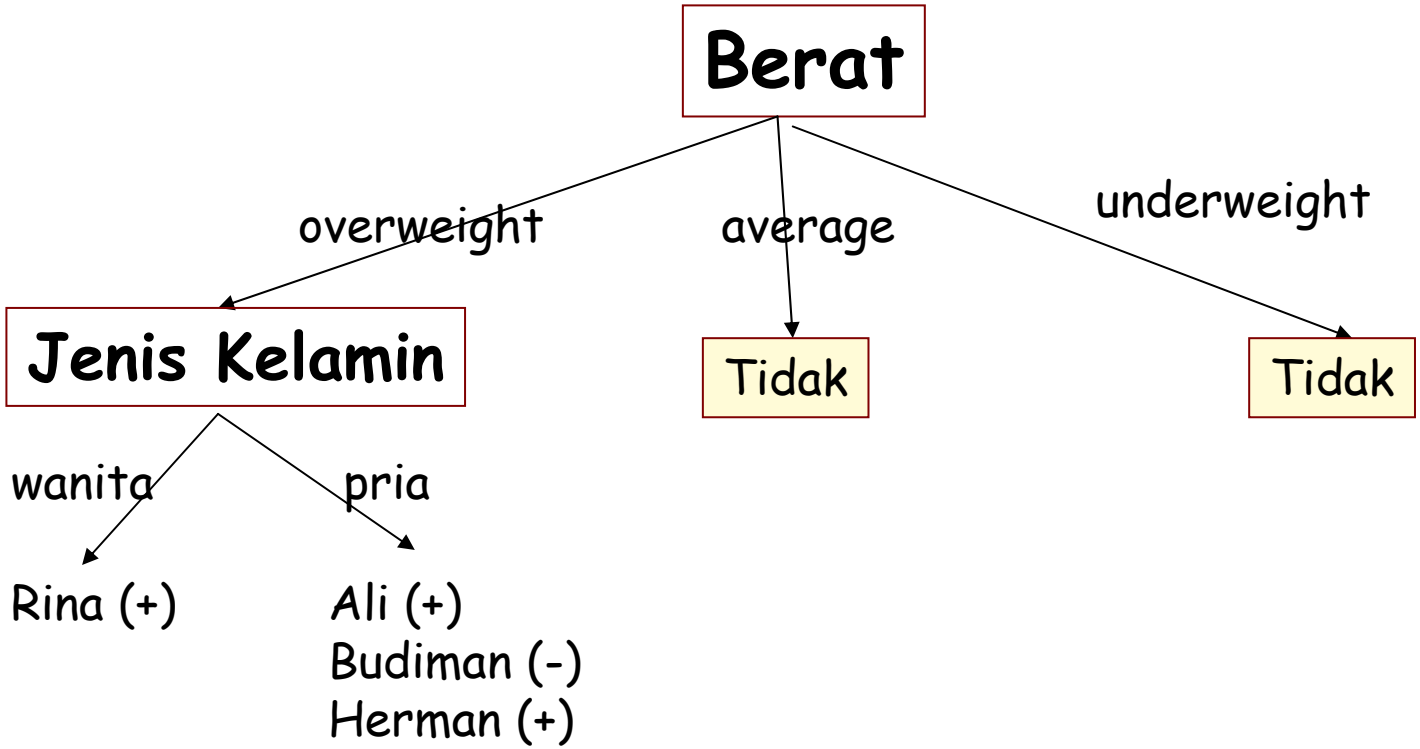
Nama	Usia	Kelamin	Hipertensi
Ali	muda	pria	ya
Budiman	tua	pria	tidak
Herman	tua	pria	ya
Rina	tua	wanita	ya

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

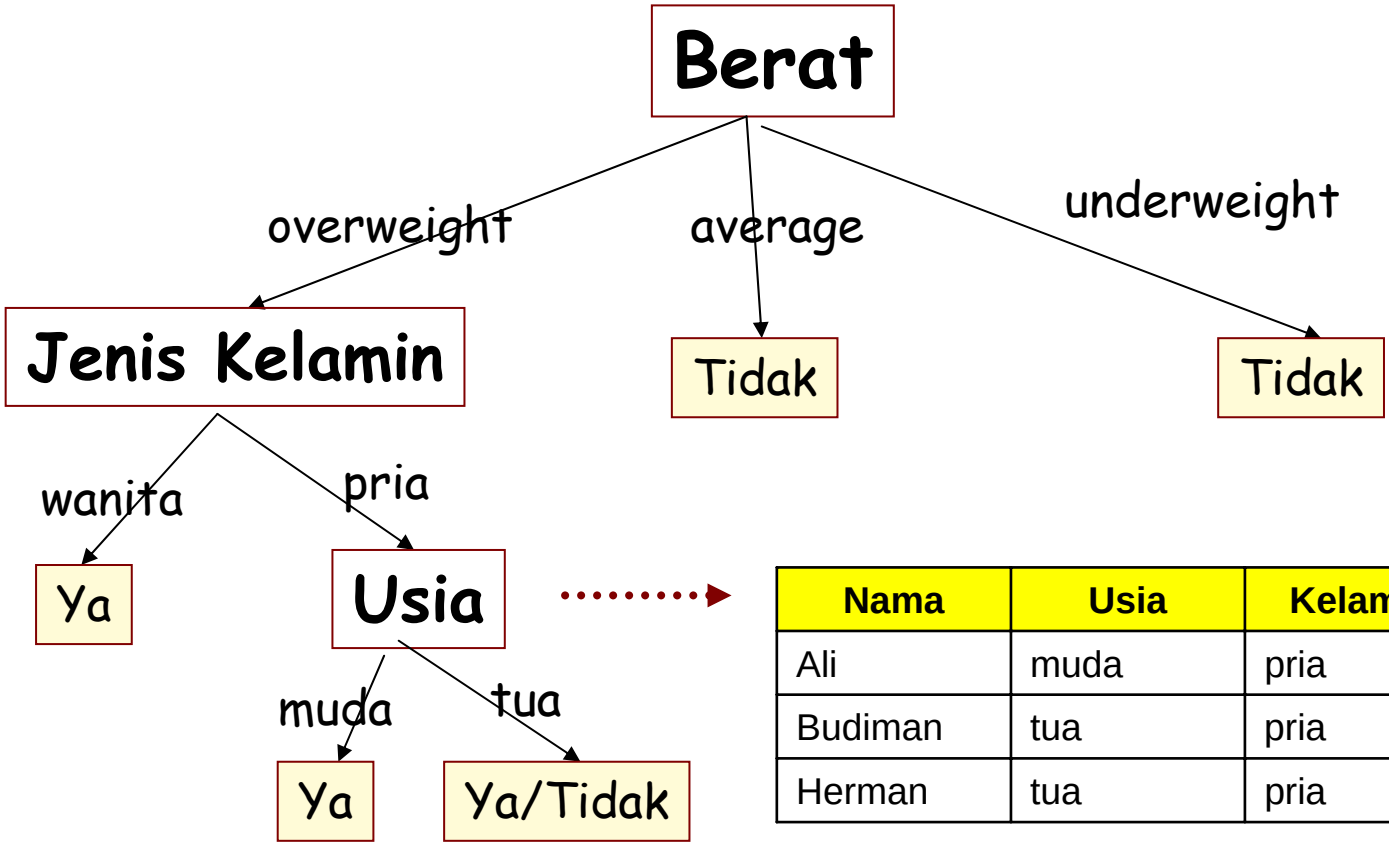


Penyusunan Tree (cont)



Leaf Node Usia dan Jenis Kelamin memiliki Entropy yang sama, sehingga tidak ada cara lain selain menggunakan pengetahuan pakar atau percaya saja pada hasil acak.

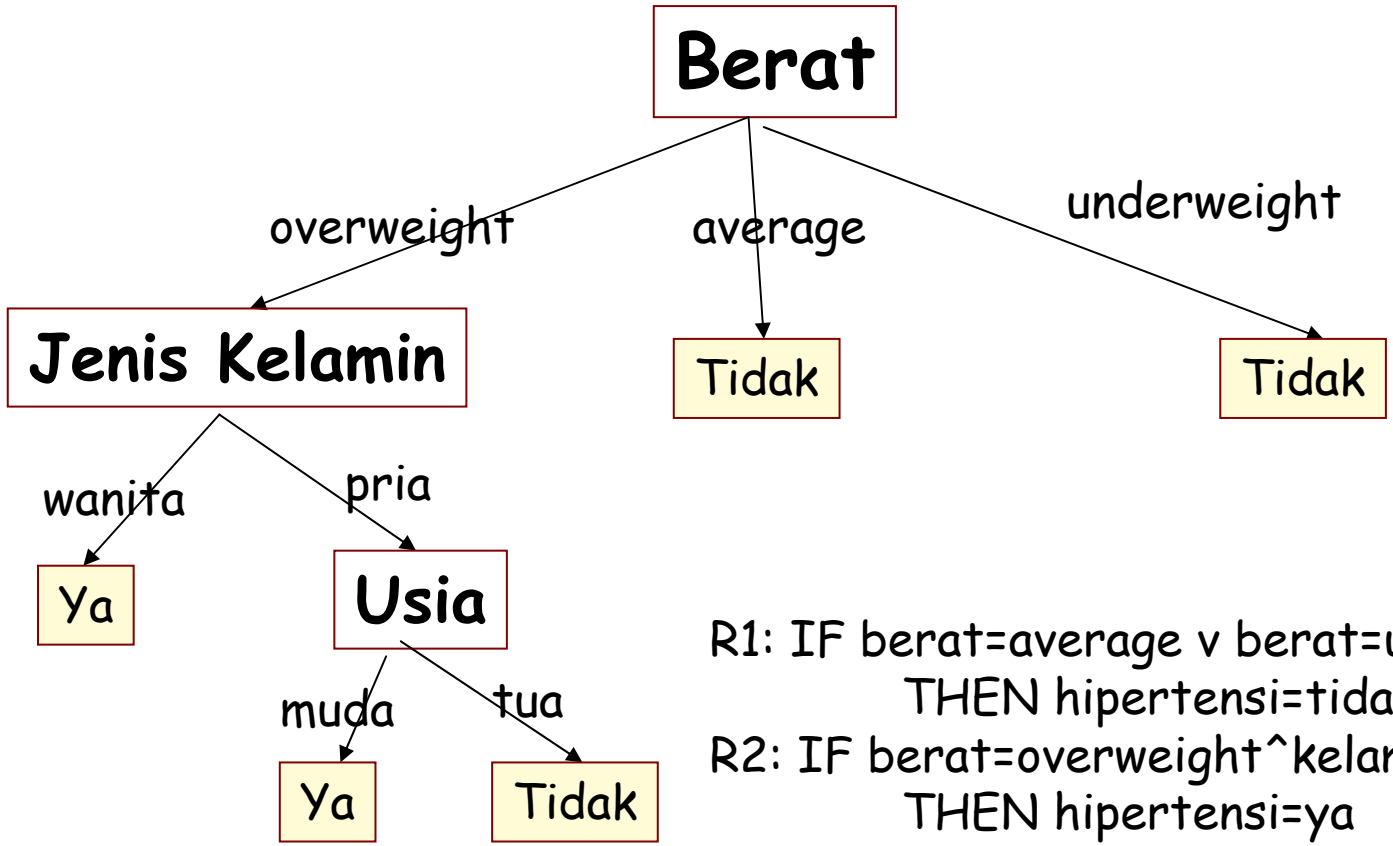
Hasil Tree



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

Pada usia=tua ternyata ada 1 data menyatakan ya dan 1 data menyatakan tidak, keadaan ini perlu dicermati. Pilihan hanya dapat ditentukan dengan campur tangan seoranng pakar.

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

Hasil Prediksi Pada Data Training

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

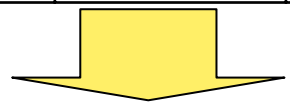
Kesalahan (e) = 12.5 %
(1 dari 8 data)

Menyederhanakan Dan Menguji Rule

- Membuat table distribusi terpadu dengan menyatakan semua nilai kejadian pada setiap rule.
- Menghitung tingkat independensi antara kriteria pada suatu rule, yaitu antara atribut dan target atribut.
- Mengeliminasi kriteria yang tidak perlu, yaitu yang tingkat independensinya tinggi.

Distribusi Terpadu

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



	overweight	average	underweight	Marginal
hipertensi	3	0	0	3
tidak	1	2	2	5
Marginal	4	2	2	8

	overweight	average	underweight	Marginal
hipertensi	12	0	0	12
tidak	4	8	8	20
Marginal	16	8	8	32

Uji Independensi Dengan Distribusi Chi-Square

O_{ij}

	overweight	average	underweight	Marginal
hipertensi	12	0	0	12
tidak	4	8	8	20
Marginal	16	8	8	32

Derajat Kebebasan adalah (jumlah baris-1)(jumlah kolom-1) = (2-1)(3-1)
dan nilai tingkat kepercayaan α=0.05
Nilai χ^2_a yang didapat dari tabel distribusi chi-square adalah 6.27

e_{ij}

	overweight	average	underweight	Marginal
hipertensi	6	3	3	12
tidak	10	5	5	20
Marginal	16	8	8	32

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$$

$$= \frac{(12-6)^2}{6}+\frac{(0-3)^2}{3}+\frac{(0-3)^2}{3}+\frac{(4-10)^2}{10}+\frac{(8-5)^2}{5}+\frac{(8-5)^2}{5}$$

$$= 19.2$$

Karena nilai $\chi^2 > \chi^2_a$ maka kriteria berat ini dependent, sehingga tidak bisa dihilangkan

Hasil Uji Independensi untuk Semua Kriteria

- Kriteria Berat mempunyai nilai chi-square 19.2, ini lebih besar dari nilai acuan chi-square dengan derajat kebebasan 2 yaitu 6.27. Jadi kriteria dependent dan tidak bisa dieliminasi.
- Kriteria Usia mempunyai nilai chi-square 2.13, ini lebih besar dari nilai acuan chi-square dengan derajat kebebasan 1 yaitu 3.89. Jadi kriteria independent dan bisa dieliminasi.
- Kriteria Jenis Kelamin mempunyai nilai chi-square 0.71, ini lebih besar dari nilai acuan chi-square dengan derajat kebebasan 1 yaitu 3.89. Jadi kriteria independent dan bisa dieliminasi.

Rule Hasil Penyederhanaan:

R1: IF berat=average
v berat=underweight
THEN hipertensi=tidak
R2: IF berat=overweight
THEN hipertensi=ya
R3: IF berat=overweigt
THEN hipertensi=ya
R4: IF berat=overweigt
THEN hipertensi=tidak

Rule Hasil Penyederhanaan:

R1: IF berat=average
v berat=underweight
THEN hipertensi=tidak
R2: IF berat=overweight
THEN hipertensi=ya

Hasil Prediksi Pada Data Training

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

Kesalahan (e) = 12.5 %
(1 dari 8 data)

Saran

- Sebaiknya pelajari lagi Statistik untuk bisa benar-benar mendukung penguasaan ilmu-ilmu Data Mining, dan Decision Tree pada khususnya
- Lebih banyak mencoba dengan berbagai macam model data dan kasus
- Belajar dan belajar terus, karena ilmu tidak akan ada habisnya