Machine Learning

Decission Tree

Entin Martiana
Knowledge Engineering Research Group
Soft Computing Laboratory
Department of Information and Computer Engineering
Politeknik Elektronika Negeri Surabaya

Politeknik Elektronika Negeri Surabaya
Departemen Teknik Informatika dan Komputer

Konten

- Konsep Decission Tree
- Proses dalam Decission Tree
- Definisi Entropy
- Pembentukan Node

Tujuan Instruksi Umum

Mahasiswa mampu menyelesaikan masalah – masalah menggunakan metode mesin pembelajaran yang tepat berdasarkan supervised, unsupervised dan reinforcement learning, baik secara individu maupun berkelompok/kerjasama tim.

Tujuan Instruksi Khusus

- Memahami klasifikasi menggunakan Decission Tree
- Mampu menerapkan Decission Tree

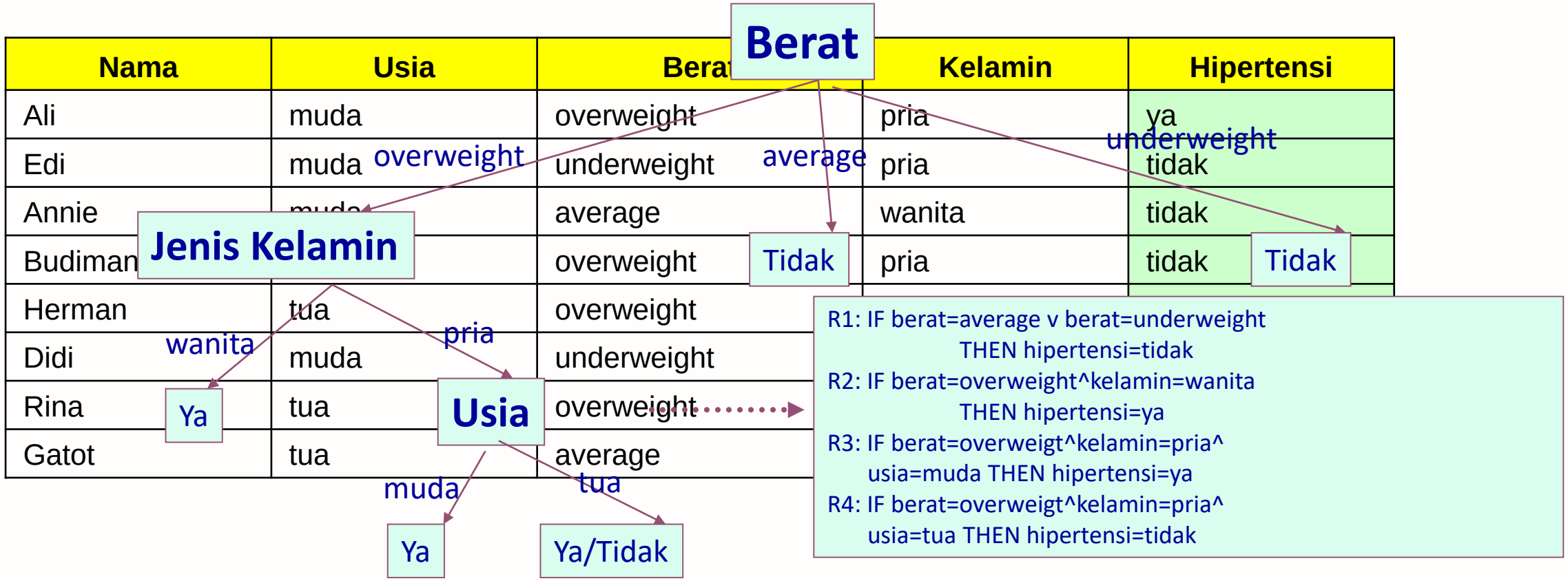
Konsep Decision Tree

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

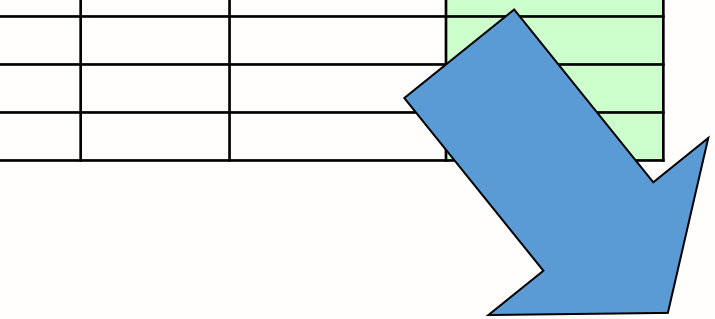
Target atribut

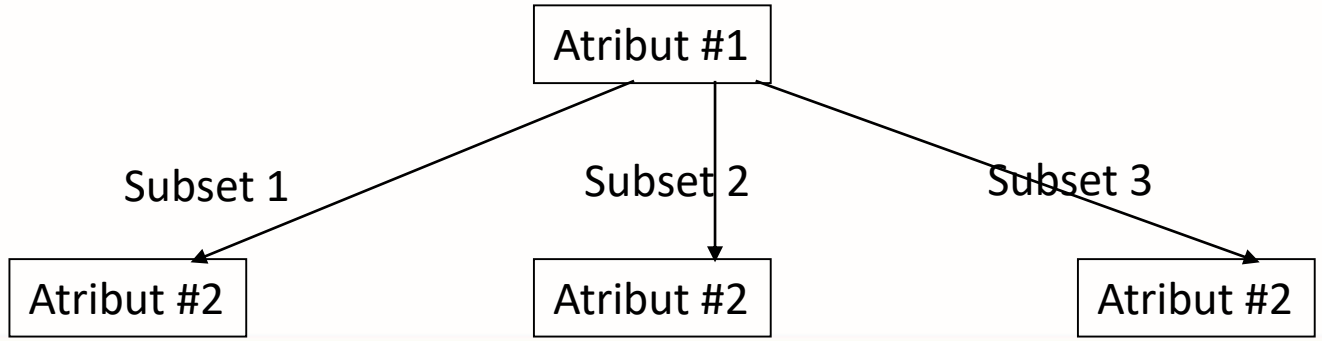
Proses Dalam Decision Tree

- 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

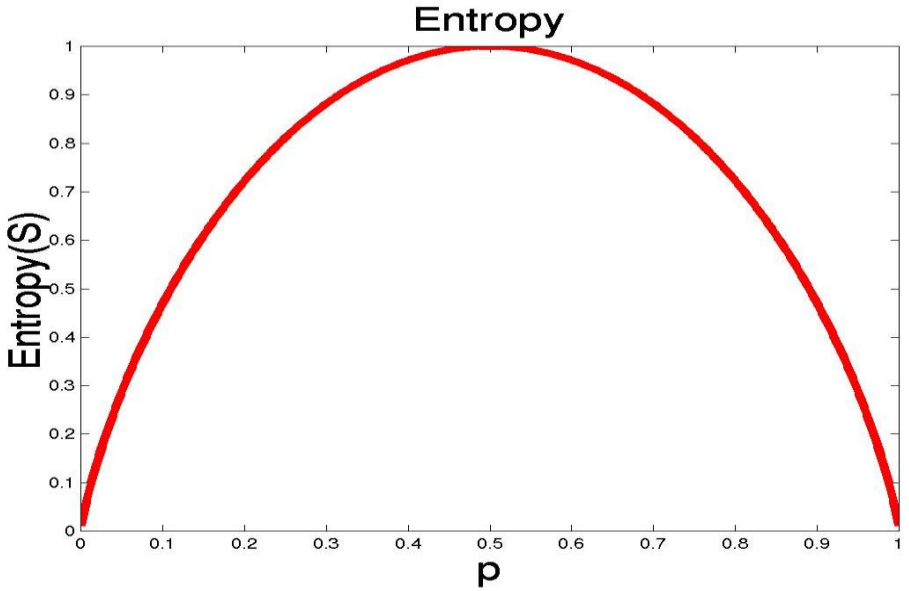




Entropy

- 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:

$$\text{Entropy}(S) = -p_+ \log_2 p_+ - p_- \log_2 p_-$$

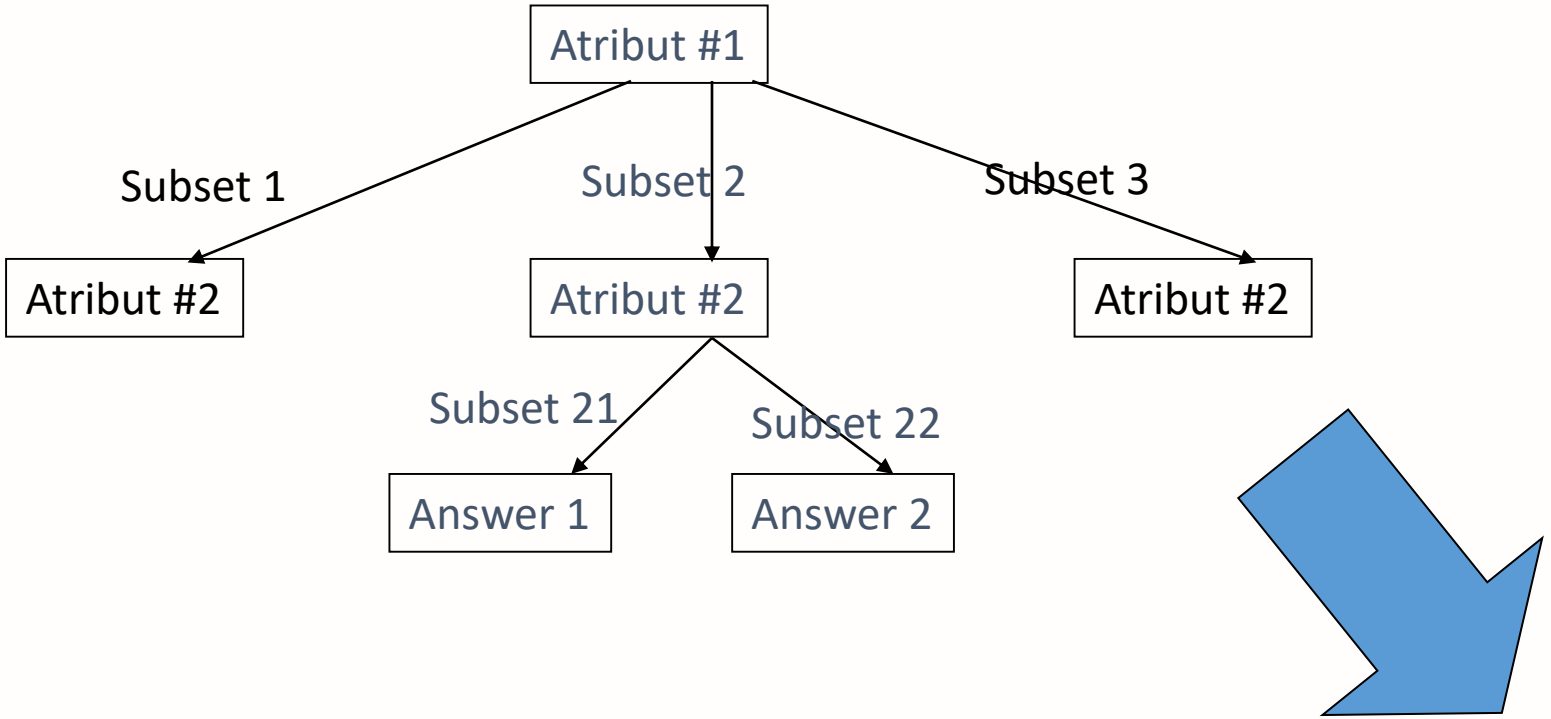


Definisi Entropy

- 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 $\mathbf{-\log_2 p}$ bits untuk messages yang mempunyai probabilitas p.
- Sehingga jumlah bit yang diperkiraank untuk mengekstraksi S ke dalam kelas adalah:

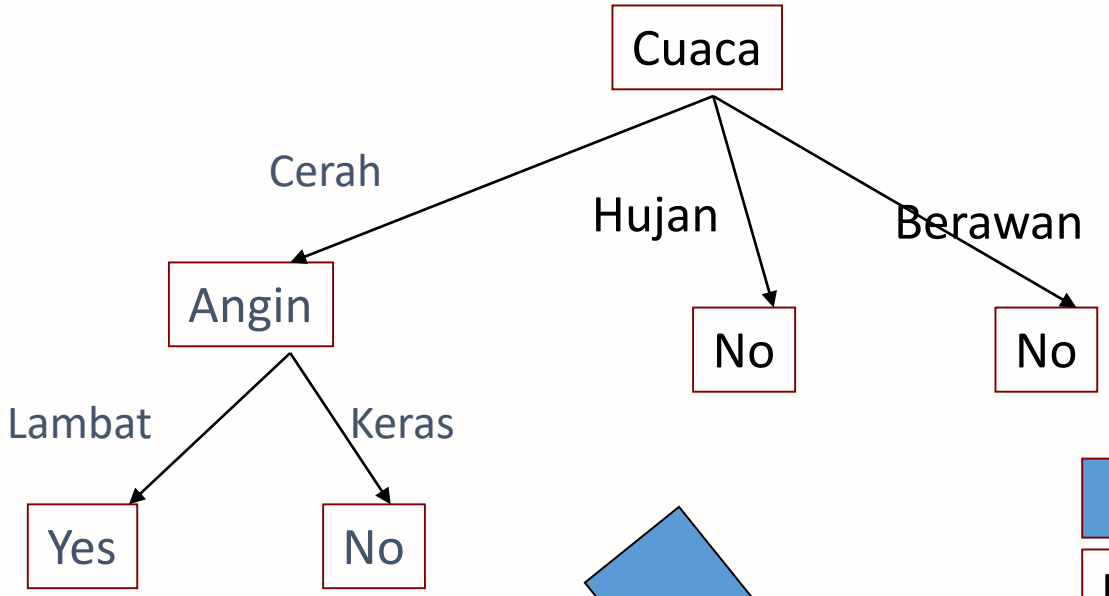
$$\mathbf{-p_+ \log_2 p_+ - p_- \log_2 p_-}$$

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 Mentah

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

Decision Tree??

Pembentukan Node

- Hitung Entropy

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

- Hitung Information Gain

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

- Atribut dengan Infromation Gain tertinggi dijadikan Node

Entropy Awal

- Jumlah instance = 8
- Jumlah instance positif = 3
- Jumlah instance negatif = 5

$$Entropy(Hipertensi) = -P_{instance\_positif} \log_2 P_{instance\_positif} - P_{instance\_negatif} \log_2 P_{instance\_negatif}$$

$$= -\left(\left(\frac{3}{8}\right) \times \log_2 \left(\frac{3}{8}\right)\right) - \left(\left(\frac{5}{8}\right) \times \log_2 \left(\frac{5}{8}\right)\right)$$

$$= -(0.375 \times \log_2 0.375) - (0.625 \times \log_2 0.625)$$

$$= -(0.375 \times -1.415) - (0.625 \times -0.678)$$

$$= 0{,}531 + 0{,}424$$

$$= 0{,}955$$

Entropy Usia

- Jumlah instance = 8
- Instance Usia
  - Muda
    - Instance positif = 1
    - Instance negatif = 3
  - Tua
    - Instance positif = 2
    - Instance negatif = 2
- Entropy Usia

$$Entropy(Muda) = -P_{instance\_positif} \log_2 P_{instance\_positif} - P_{instance\_negatif} \log_2 P_{instance\_negatif}$$

$$Entropy(Tua) = -P_{instance\_positif} \log_2 P_{instance\_positif} - P_{instance\_negatif} \log_2 P_{instance\_negatif}$$

  - Entropy(muda) = 0.906
  - Entropy(tua) = 1

Gain Usia

$$Gain(S, Usia) = Entropy(S) - \sum_{v \in Muda, Tua} \frac{|S_v|}{S} Entropy(S_v)$$

$$= Entropy(S) - \frac{S_{Muda}}{S} Entropy(S_{Muda}) - \frac{S_{Tua}}{S} Entropy(S_{Tua})$$

$$= (0.955) - \frac{4}{8}(0.906) - \frac{4}{8}(1)$$

$$= 0.955 - 0.453 - 0.5$$

$$= 0.002$$

Entropy Berat

- Jumlah instance = 8
- Intance Berat
  - Overweight
    - Instance positif = 3
    - Instance negatif = 1
  - Average
    - Instance positif = 0
    - Instance negatif = 2
  - Underweight
    - Instance positif = 0
    - Instance negatif = 2

$$Entropy(Overweight) = -P_{instance\_positif} \log_2 P_{instance\_positif} - P_{instance\_negatif} \log_2 P_{instance\_negatif}$$

$$Entropy(Average) = -P_{instance\_positif} \log_2 P_{instance\_positif} - P_{instance\_negatif} \log_2 P_{instance\_negatif}$$

$$Entropy(Underweight) = -P_{instance\_positif} \log_2 P_{instance\_positif} - P_{instance\_negatif} \log_2 P_{instance\_negatif}$$

- Entropy(Overweight)=0.906
- Entropy(Average)=0.5
- Entropy(Underweight)=0.5

Gain Berat

$$Gain(S, Berat) = Entropy(S) - \sum_{v \in Overwight, Average, Underweigh} \frac{|S_v|}{S} Entropy(S_v)$$

$$= Entropy(S) - \frac{S_{Overweight}}{S} Entropy(S_{Overweight}) - \frac{S_{Average}}{S} Entropy(S_{average}) - \frac{S_{Underweigh}}{S} Entropy(S_{Underweigh})$$

$$= (0.955) - \frac{4}{8}(0.906) - \frac{2}{8}(0.5) - \frac{2}{8}(0.5)$$

$$= 0.955 - 0.453 - 0.125 - 0.125$$

$$= 0{,}252$$

Entropy Jenis Kelamin

- Jumlah instance = 8
- Intance Jenis Kelamin
  - Pria
    - Instance positif = 2
    - Instance negatif = 4
  - Wanita
    - Instance positif = 1
    - Instance negatif = 1

$$Entropy(\mathrm{Pr}\,ia) = -P_{instance\_positif} \log_2 P_{instance\_positif} - P_{instance\_negatif} \log_2 P_{instance\_negatif}$$

$$Entropy(Wanita) = -P_{instance\_positif} \log_2 P_{instance\_positif} - P_{instance\_negatif} \log_2 P_{instance\_negatif}$$

- Entropy(Pria)=1
- Entropy(Wanita)=0.75

Gain Jenis Kelamin

$$Gain(S, JenisKelamin) = Entropy(S) - \sum_{v \in Pria, Wanita} \frac{|S_v|}{S} Entropy(S_v)$$

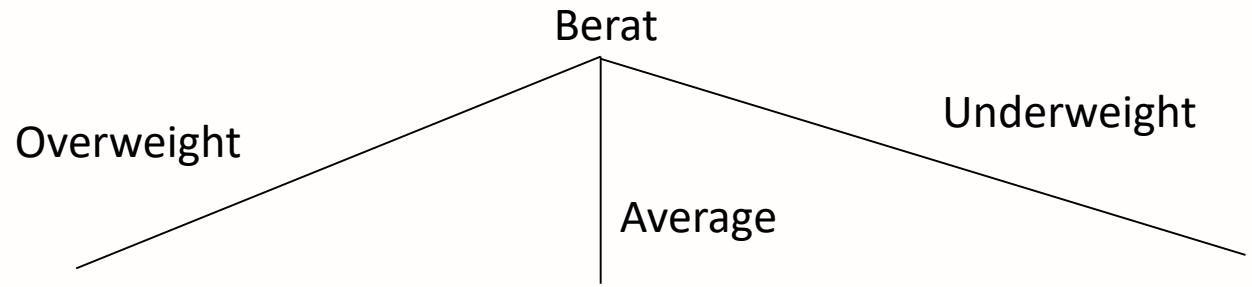
$$= Entropy(S) - \frac{S_{Pria}}{S} Entropy(S_{Pria}) - \frac{S_{Wanita}}{S} Entropy(S_{Wanita})$$

$$= (0.955) - \frac{6}{8}(1) - \frac{2}{8}(0.75)$$

$$= 0.955 - 0.75 - 0.188$$

$$= 0{,}017$$

- Atribut yang dipilih adalah atribut berat karena nilai Information Gainnya paling tinggi

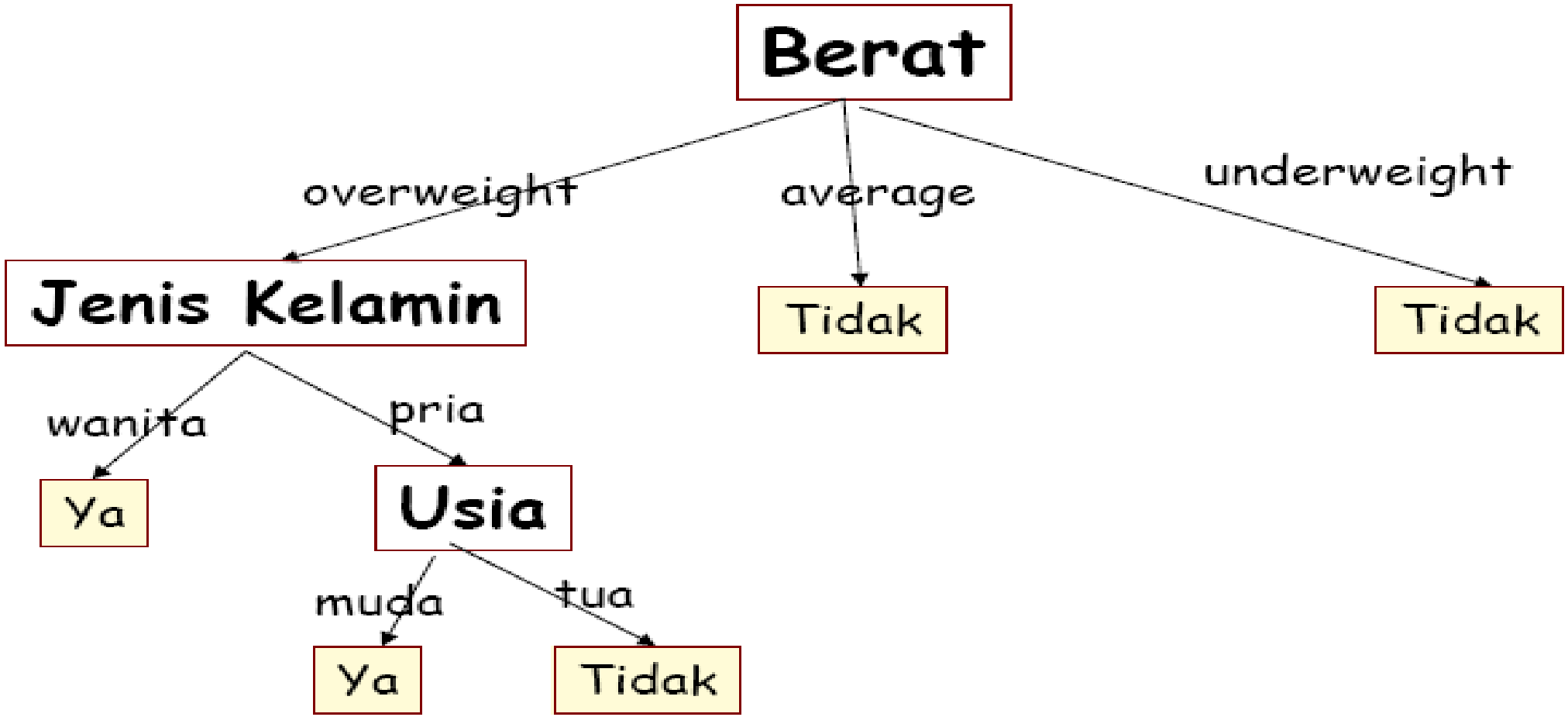


- Jumlah Instance untuk Overweight = 4
- Jumlah Instance untuk Average = 2
- Jumlah Instance untuk Underweight = 2
- Hitung Gain paling tinggi untuk dijadikan cabang berikutnya

Node untuk cabang Overweight

- Jumlah instance = 4
- Instance (Berat = Overwight) & Usia =
  - Muda
    - Instance positif = 1
    - Instance negatif = 0
  - Tua
    - Instance positif = 2
    - Instance negatif = 1
- Instance (Berat = Overwight) & Jenis Kelamin =
  - Pria
    - Instance positif = 2
    - Instance negatif = 1
  - Wanita
    - Instance positif = 1
    - Instance negatif = 0

Decision Tree yang dihasilkan



Latihan Soal

WAKTU	PAKET	FREKWEKSI	PRIORITAS	GANGGUAN
PENDEK	BESAR	SEDANG	RENDAH	GANGGUAN
PENDEK	KECIL	RENDAH	TINGGI	GANGGUAN
PANJANG	BESAR	SEDANG	TINGGI	NORMAL
PANJANG	KECIL	TINGGI	RENDAH	NORMAL
PENDEK	BESAR	TINGGI	TINGGI	GANGGUAN
PANJANG	KECIL	RENDAH	TINGGI	GANGGUAN
PANJANG	KECIL	TINGGI	RENDAH	GANGGUAN
PANJANG	KECIL	SEDANG	RENDAH	NORMAL
PANJANG	BESAR	TINGGI	TINGGI	NORMAL
PANJANG	KECIL	SEDANG	RENDAH	GANGGUAN
PENDEK	BESAR	SEDANG	TINGGI	NORMAL
PANJANG	BESAR	RENDAH	TINGGI	NORMAL

1. Buatlah tree dan rule untuk mendeteksi adanya gangguan pada jaringan komputer menggunakan data di atas
2. Berapa persen besarnya error yang terjadi jika data tersebut dimasukkan pada rule yang didapatkan?

Latihan Soal

USIA	KELAMIN	MEROKOK	OLAHRAGA	JANTUNG
TUA	PRIA	TIDAK	YA	TIDAK
TUA	PRIA	YA	YA	TIDAK
MUDA	PRIA	YA	TIDAK	TIDAK
TUA	PRIA	TIDAK	TIDAK	TIDAK
MUDA	WANITA	TIDAK	TIDAK	YA
MUDA	PRIA	TIDAK	YA	YA
MUDA	PRIA	TIDAK	YA	TIDAK
TUA	WANITA	TIDAK	TIDAK	YA
MUDA	PRIA	YA	TIDAK	TIDAK
TUA	PRIA	YA	TIDAK	TIDAK
MUDA	PRIA	YA	YA	YA
TUA	PRIA	YA	TIDAK	TIDAK
MUDA	PRIA	TIDAK	TIDAK	TIDAK
TUA	PRIA	TIDAK	YA	TIDAK
MUDA	PRIA	YA	TIDAK	TIDAK

1. Buatlah tree dan rule untuk mendeteksi penyakit jantung menggunakan data di atas
2. Berapa persen besarnya error yang terjadi jika data tersebut dimasukkan pada rule yang didapatkan?

Referensi

- Decission Tree, *https://danangjunaedi.files.wordpress.com/2011/02/decision-tree.ppt*
- Decission Tree, **Achmad Basuki, Iwan Syarif,** Politeknik Elektronika Negeri Surabaya, 2003.
- Machine Learning, Tom Mitchell, McGraw-Hill. 2008.

bridge to the future

http://www.eepis-its.edu