


# Decision Tree



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



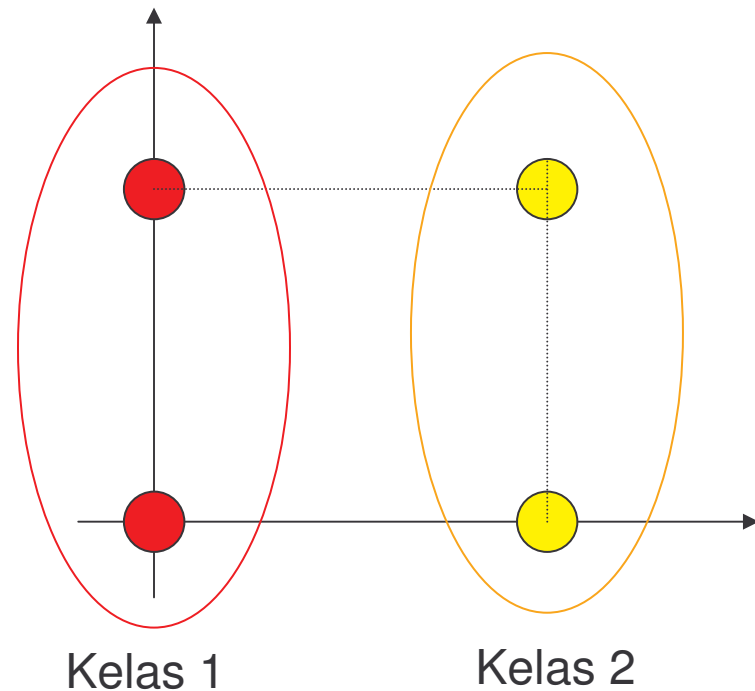
# Proses Klasifikasi Dalam Data Mining

- Klasifikasi adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Model itu sendiri bisa berupa aturan "jika-maka", berupa decision tree, formula matematis atau neural network. (Iko Pramudiono, Modul Pengantar Data Mining, [www.ilmukomputer.com](http://www.ilmukomputer.com))
- Proses klasifikasi biasanya dibagi menjadi dua fase : learning dan test. Pada fase learning, sebagian data yang telah diketahui kelas datanya diumpankan untuk membentuk model perkiraan. Kemudian pada fase test model yang sudah terbentuk diuji dengan sebagian data lainnya untuk mengetahui akurasi dari model tsb. Bila akurasinya mencukupi model ini dapat dipakai untuk prediksi kelas data yang belum diketahui.
- Klasifikasi dicirikan dengan data training mempunyai label, berdasarkan label ini proses klasifikasi memperoleh pola attribut dari suatu data.



# Proses Klasifikasi Dalam Data Mining

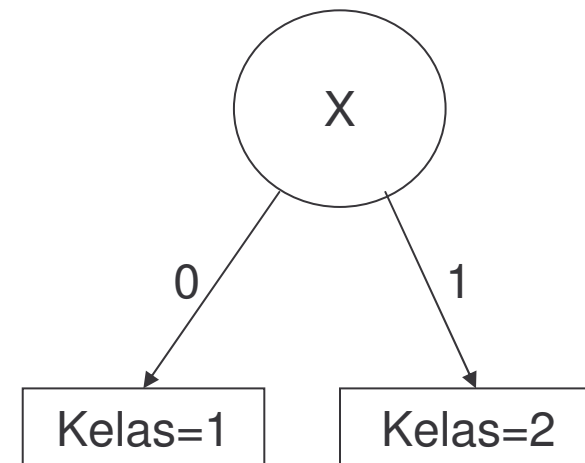
| X | Y | Kelas |
|---|---|-------|
| 0 | 0 | 1     |
| 0 | 1 | 1     |
| 1 | 0 | 2     |
| 1 | 1 | 2     |





# Proses Klasifikasi Dapat Menjadi Sebuah Tree

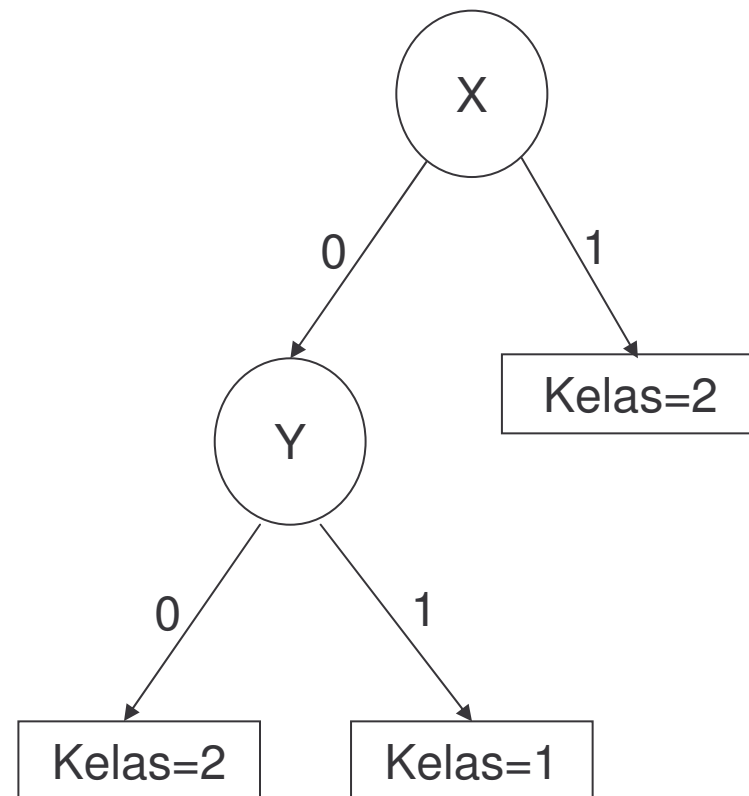
| X | Y | Kelas |
|---|---|-------|
| 0 | 0 | 1     |
| 0 | 1 | 1     |
| 1 | 0 | 2     |
| 1 | 1 | 2     |





# Proses Klasifikasi Dapat Menjadi Sebuah Tree

| X | Y | Z | Kelas |
|---|---|---|-------|
| 0 | 0 | 0 | 2     |
| 0 | 0 | 1 | 2     |
| 0 | 1 | 0 | 1     |
| 0 | 1 | 1 | 1     |
| 1 | 0 | 0 | 2     |
| 1 | 0 | 1 | 2     |
| 1 | 1 | 0 | 2     |
| 1 | 1 | 1 | 2     |





# Konsep Decision Tree

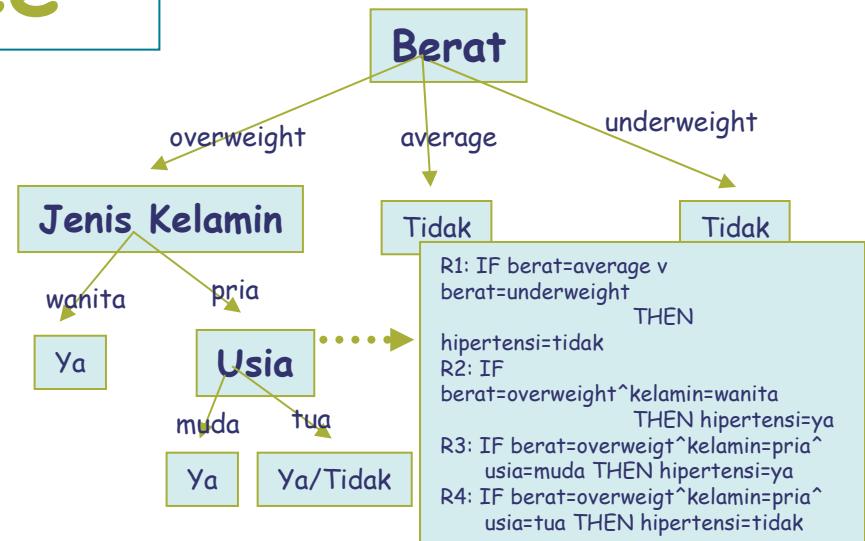
Mengubah data menjadi pohon keputusan (*decision tree*) dan aturan-aturan keputusan (*rule*)

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      |





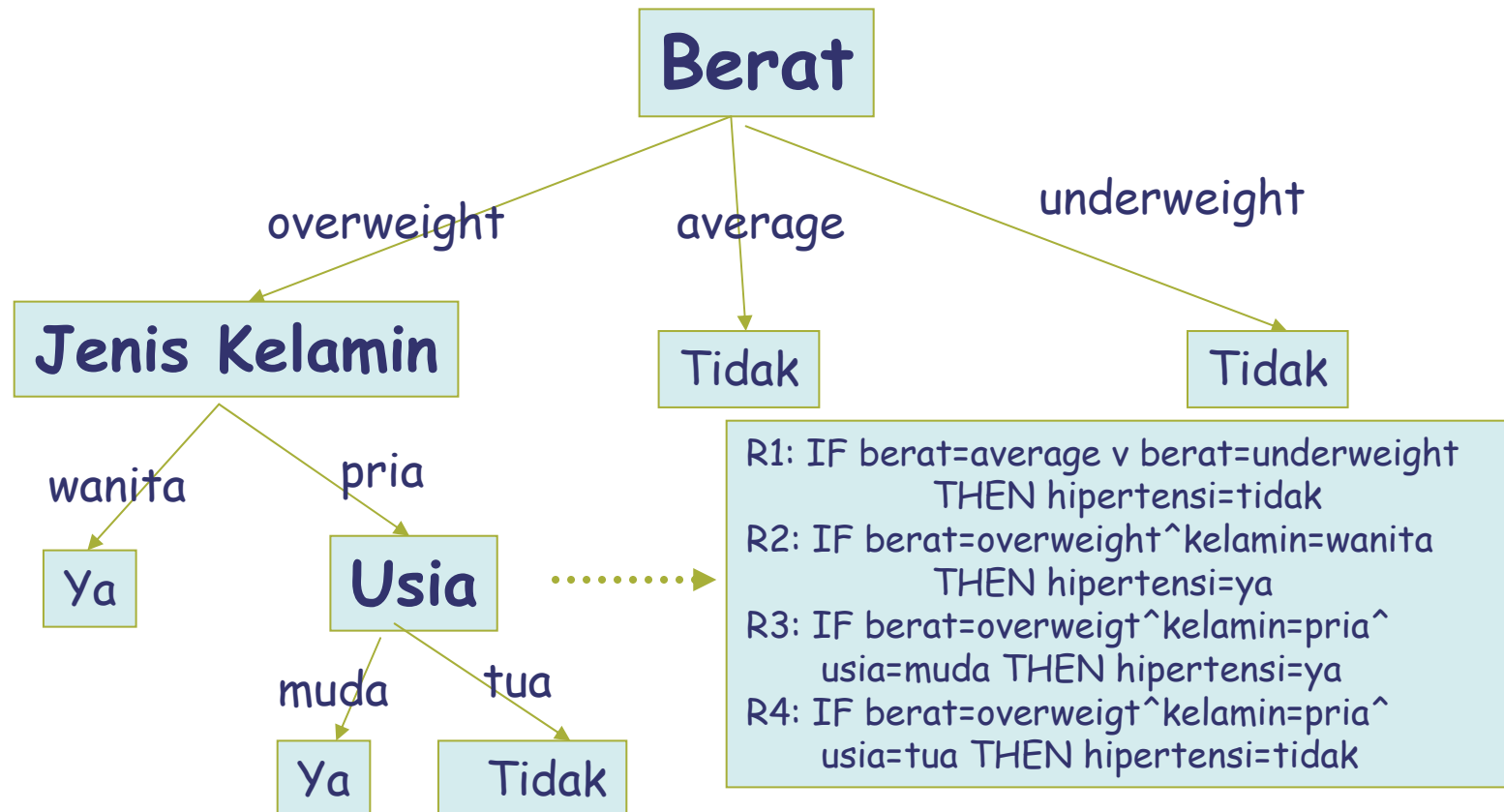
# Gambaran Pemakaian Decision Tree

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.

| # | Usia | Berat Badan | Kelamin | Hipertensi |
|---|------|-------------|---------|------------|
| 1 | muda | overweight  | pria    | ya         |
| 2 | muda | underweight | pria    | tidak      |
| 3 | muda | average     | wanita  | tidak      |
| 4 | tua  | overweight  | pria    | tidak      |
| 5 | tua  | overweight  | pria    | ya         |
| 6 | muda | underweight | pria    | tidak      |
| 7 | tua  | overweight  | wanita  | ya         |
| 8 | tua  | average     | pria    | tidak      |



# Gambaran Pemakaian Decision Tree







# 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

| 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

atribut

↓  
Target atribut



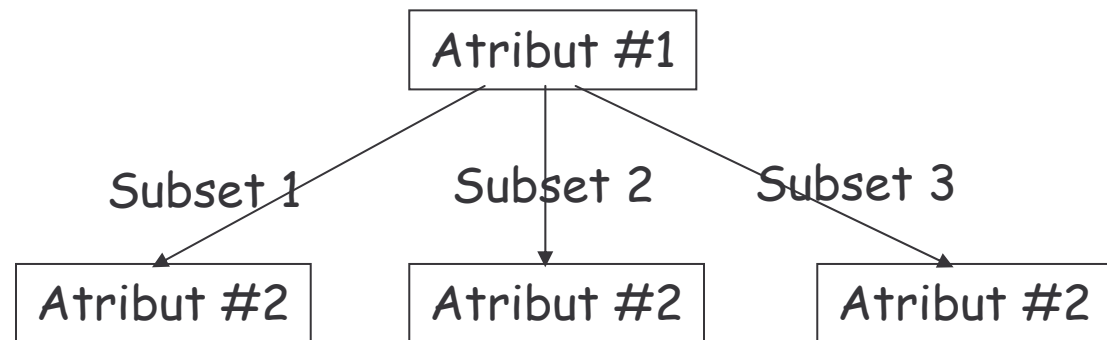
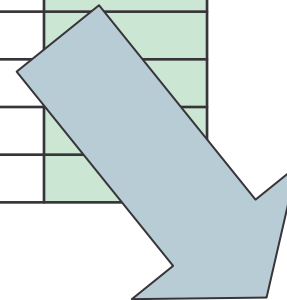
# Proses Dalam Decision Tree

- Mengubah bentuk data (tabel) menjadi model tree. Dalam Modul ini menggunakan algoritma ID3.
- Mengubah model tree menjadi rule
- Menyederhanakan Rule (Pruning)



# Proses Data Menjadi Tree

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

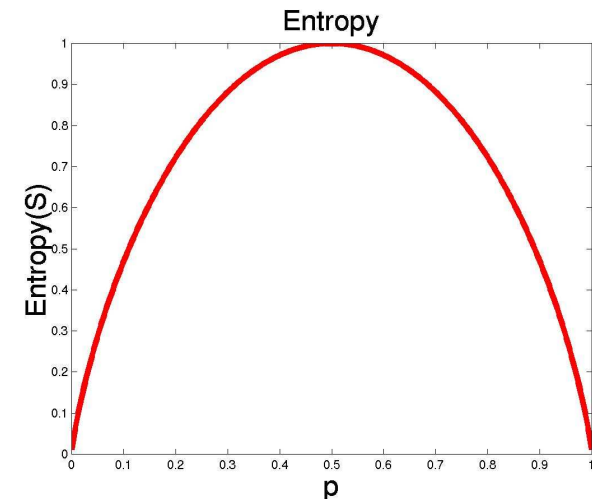




# Entropy

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

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



# Contoh Permasalahan Penentuan Hipertensi Menggunakan Decision Tree

Data diambil dengan 8 sample, dengan pemikiran bahwa yang mempengaruhi 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

| 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

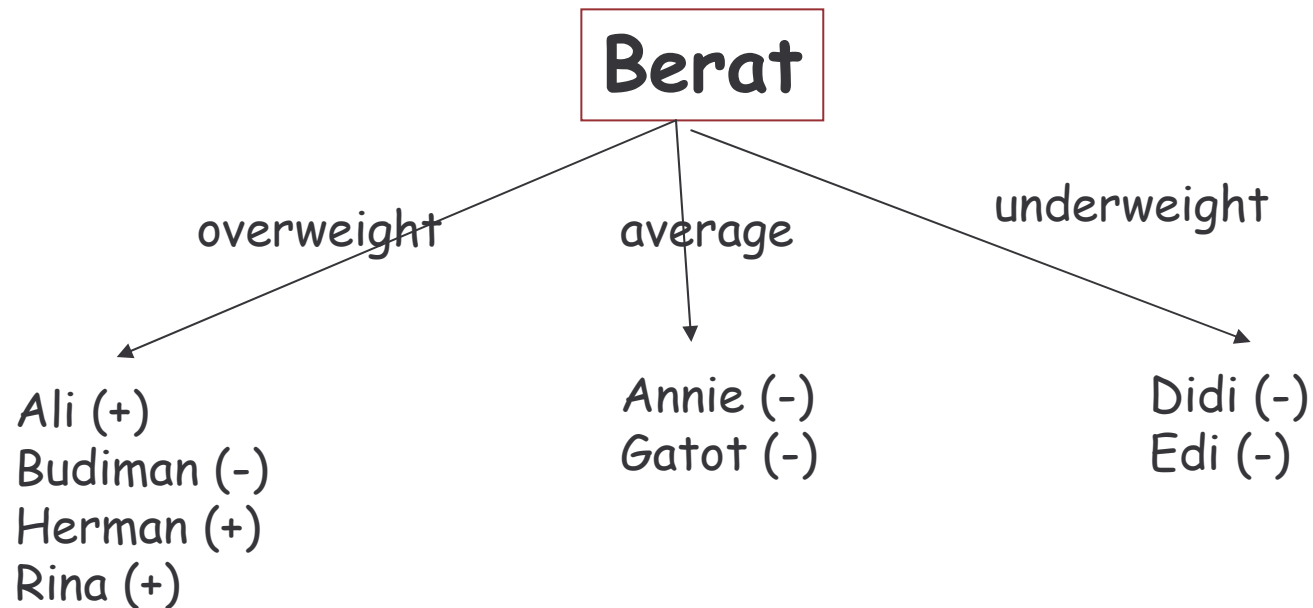
| Berat       | Hipertensi | Jumlah |
|-------------|------------|--------|
| overweight  | ya         | 3      |
| overweight  | tidak      | 1      |
| average     | ya         | 0      |
| average     | tidak      | 2      |
| underweight | ya         | 0      |
| underweight | tidak      | 2    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

Entropy = 0.41

Terpilih atribut BERAT BADAN  
sebagai node awal karena  
memiliki entropy terkecil



# Penyusunan Tree Awal



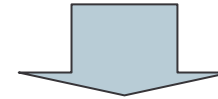
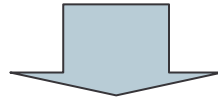
Leaf Node berikutnya dapat dipilih pada bagian yang mempunyai nilai + dan -, pada contoh di atas hanya berat=overweight yang mempunyai nilai + dan - maka semuanya pasti mempunya leaf node. Untuk menyusun leaf node lakukan satu-persatu.



# 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         |

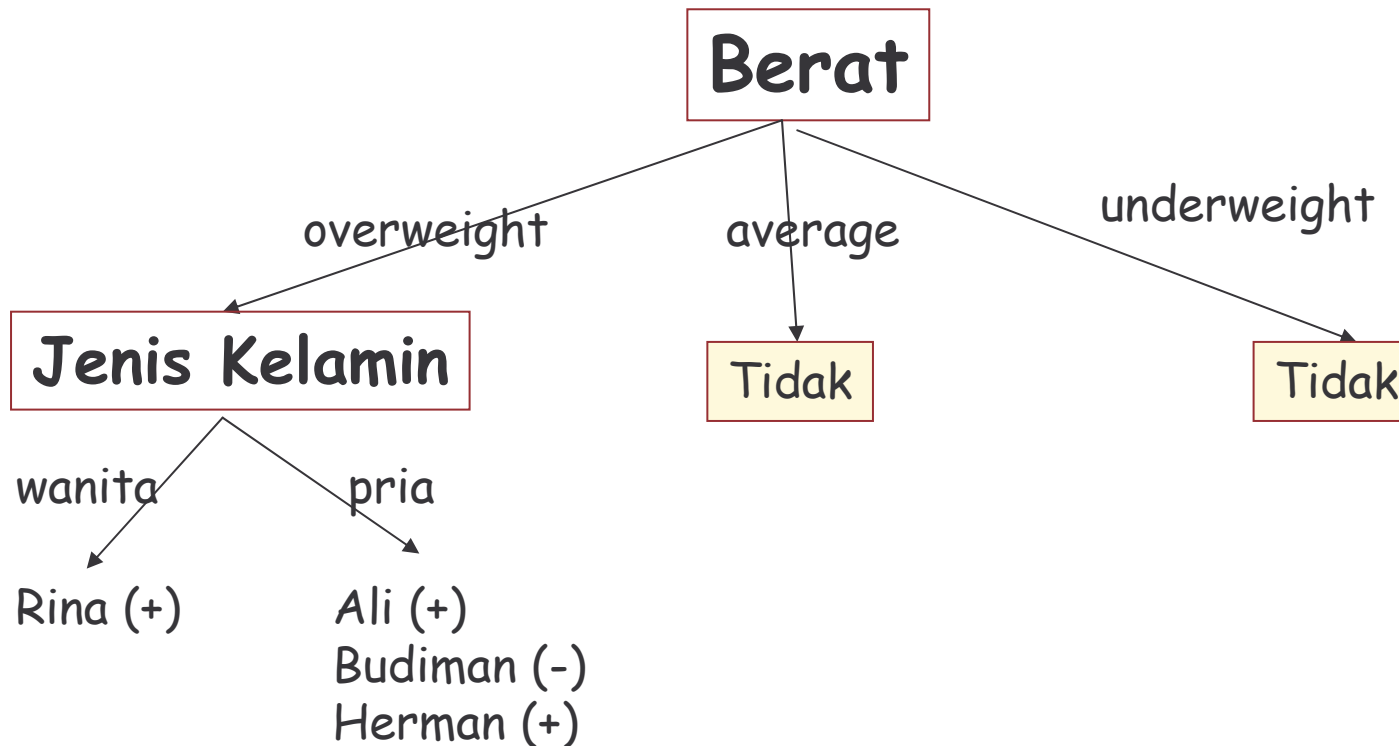


| Usia      | Hipertensi | Jumlah |
|-----------|------------|--------|
| muda      | ya         | 1      |
|           | tidak      | 0      |
| tua       | ya         | 2      |
|           | tidak      | 1      |
| Entropy = |            | 0,69   |

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



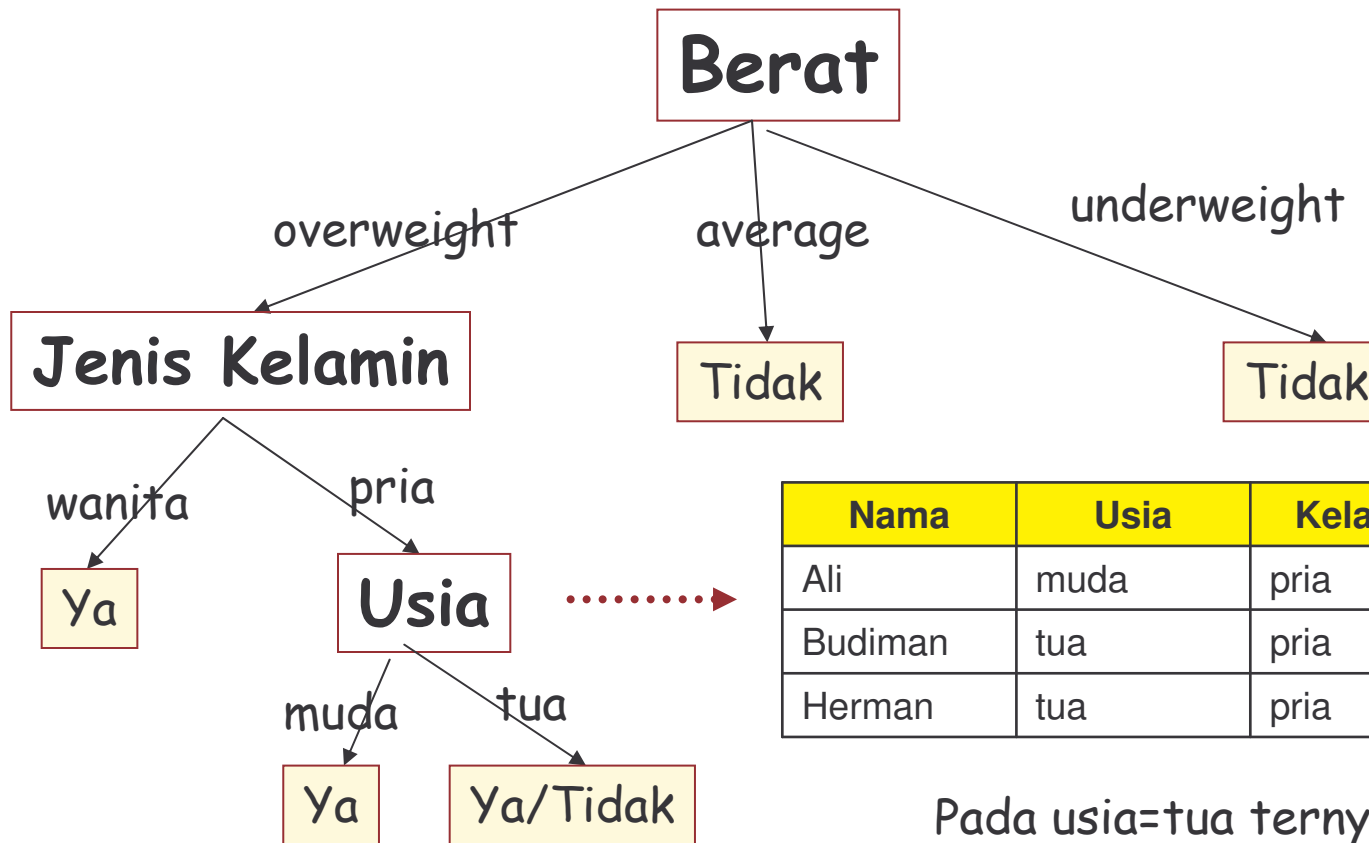
# Penyusunan Tree



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

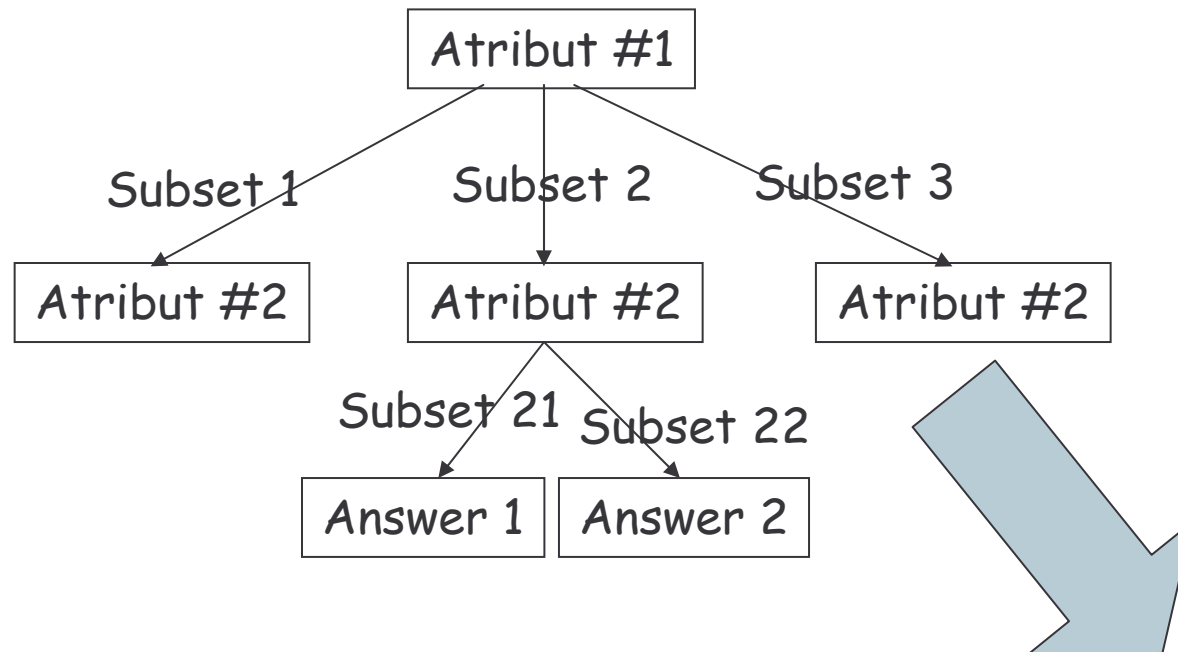


| Nama    | Usia | Kelamin | Hipertensi |
|---------|------|---------|------------|
| Ali     | muda | pria    | ya         |
| Budiman | tua  | pria    | tidak      |
| Herman  | tua  | pria    | ya         |

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



# Mengubah Tree Menjadi Rules

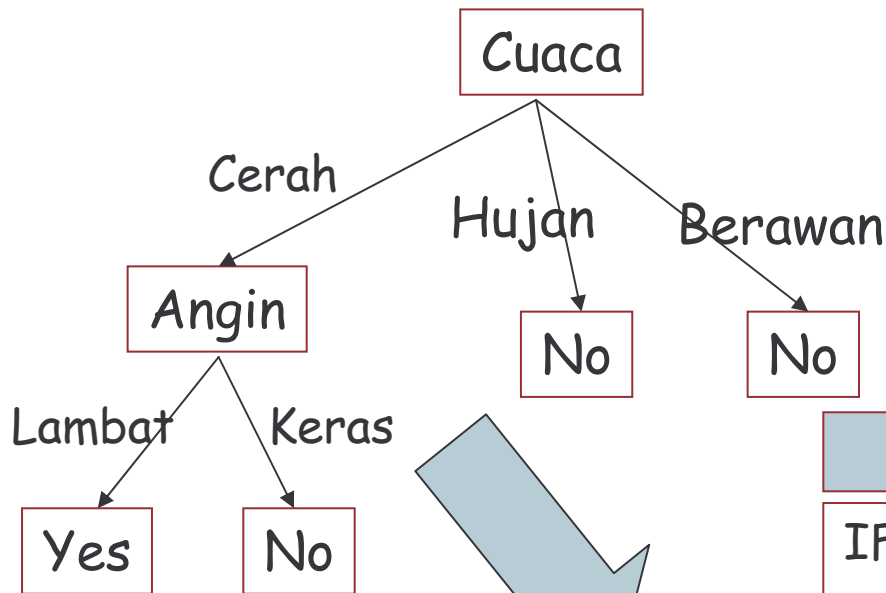


If atribut#1=subset2 ^ atribut#2=subset21  
then answer=answer1  
If atribut#1=subset2 ^ atribut#2=subset22  
then answer=answer2





# Conjunction & Disjunction



Disjunction  $\vee$

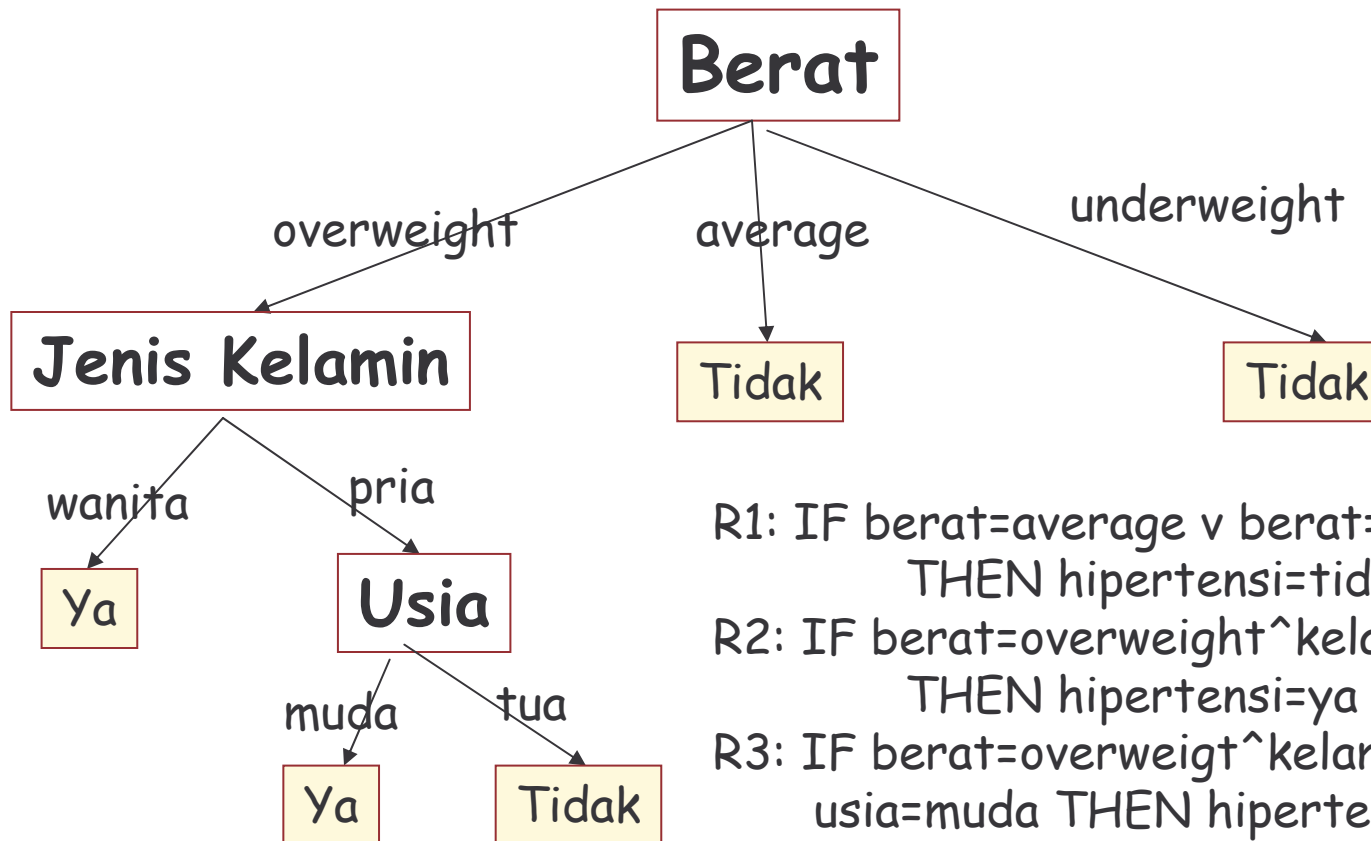
IF cuaca=hujan  $\vee$  cuaca=berawan THEN  
MainTenis=No

Conjunction  $\wedge$

IF cuaca=cerah  $\wedge$  angin=lambat THEN  
MainTenis=Yes  
IF cuaca=cerah  $\wedge$  angin=keras THEN  
MainTenis=No



# 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=overweight^kelamin=pria^  
usia=muda THEN hipertensi=ya
- R4: IF berat=overweight^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 )



# Data Uji Coba Decision Tree

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

1. Buatlah tree dan rule untuk mendeteksi adanya gangguan pada jaringan komputer menggunakan data di atas
2. Lakukan penyederhaan (Pruning)
3. Berapa persen besarnya error yang terjadi tanpa penyederhanaan (pruning) dan dengan penyederhanaan



# Data Uji Coba Decision Tree

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

1. Buatlah tree dan rule untuk mendeteksi penyakit jantung menggunakan data di atas
2. Lakukan Penyederhaan (Pruning)
3. Berapa persen besarnya error yang terjadi tanpa penyederhanaan (pruning) dan dengan penyederhanaan



# Data Uji Coba Decision Tree

| RED    | GREEN  | BLUE   | CONTENT    |
|--------|--------|--------|------------|
| TINGGI | RENDAH | TINGGI | BACKGROUND |
| TUA    | RENDAH | RENDAH | BACKGROUND |
| RENDAH | TINGGI | RENDAH | BACKGROUND |
| TINGGI | TUA    | TINGGI | OBYEK      |
| RENDAH | TUA    | TINGGI | BACKGROUND |
| RENDAH | TINGGI | TUA    | BACKGROUND |
| TINGGI | TUA    | RENDAH | OBYEK      |
| RENDAH | TUA    | RENDAH | BACKGROUND |
| TINGGI | TINGGI | TINGGI | OBYEK      |
| TUA    | TINGGI | RENDAH | BACKGROUND |
| TUA    | RENDAH | TUA    | OBYEK      |
| RENDAH | RENDAH | TINGGI | BACKGROUND |
| TUA    | RENDAH | RENDAH | BACKGROUND |
| RENDAH | TUA    | RENDAH | OBYEK      |
| RENDAH | RENDAH | TINGGI | OBYEK      |

1. Buatlah tree dan rule untuk mendeteksi apakah suatu warna itu obyek atau background
2. Lakukan penyederhaan (Pruning)
3. Berapa persen besarnya error yang terjadi tanpa penyederhanaan (pruning) dan dengan penyederhanaan