



# Classification Tree CART (2)

---

Kuliah 4 - STA1382 Teknik  
Pembelajaran Mesin

Septian Rahardiantoro



# Outline

- Classification Tree CART
  1. Pengenalan Konsep Entropy dan Information Gain
  2. Pengenalan Algoritma Dasar Pohon Klasifikasi
  3. Menilai Kemampuan Prediksi Pohon Klasifikasi

# Classification Tree CART

- Merupakan pohon keputusan yang diaplikasikan pada kasus regresi dengan peubah  $Y$  berskala kategorik
- Seringkali disebut dengan istilah:
  - Classification Tree
  - Decision Tree
  - Recursive Partition
  - Iterative Dichotomiser

## Kegunaan

- Mengidentifikasi peubah apa yang dapat dijadikan sebagai pembeda antar kelompok
- Memprediksi keanggotaan kelompok suatu individu berdasarkan karakteristiknya
- Terapannya antara lain:
  - Marketing: Mengidentifikasi prospective customer (cross-sell, up-sell, new acquisition)
  - Risk: Credit scoring, menentukan apakah calon penerima kredit akan mampu bayar atau tidak
  - Customer Relationship: churn analysis, menentukan customer yang berpotensi akan meninggalkan jasa/produk
  - Health: menentukan tingkat resiko penyakit
  - dll
- Metode yang setara kegunaannya: Regresi Logistik, k-Nearest Neighbor, Discriminant Analysis, Support Vector Machine, Bayesian Classifier, dll

# 1. Pengenalan Entropy dan Information Gain

## Entropy

- Andaikan sebuah gugus data D berisi individu-individu dengan dua kelas yaitu kelas YES dan NO, dengan proporsi yang YES sebesar  $p$ , dan tentu saja  $(1 - p)$  lainnya tergolong kelas NO.

- Entropi dari gugus data tersebut adalah

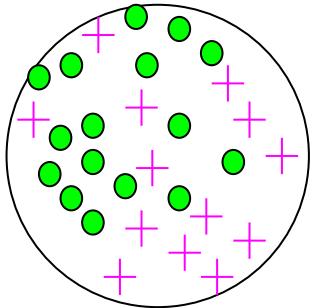
$$E(D) = -p \log_2(p) - (1-p) \log_2(1-p) \quad \checkmark$$

$$\begin{array}{l} \text{Yes} \\ p = 1 \\ \text{No} \\ p = 0 \end{array} \quad 1-p = 0 \quad 1-p = 1$$

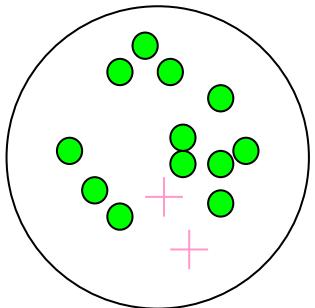
$$E(D) = 0 - 1(\log_2(1)) = 0$$

- Gugus data yang seluruh amatannya dari kelas YES akan memiliki  $E(D) = 0$
- Gugus data yang seluruh amatannya dari kelas NO juga akan memiliki  $E(D) = 0$
- Entropi ini adalah ukuran keheterogenan data (impurity)

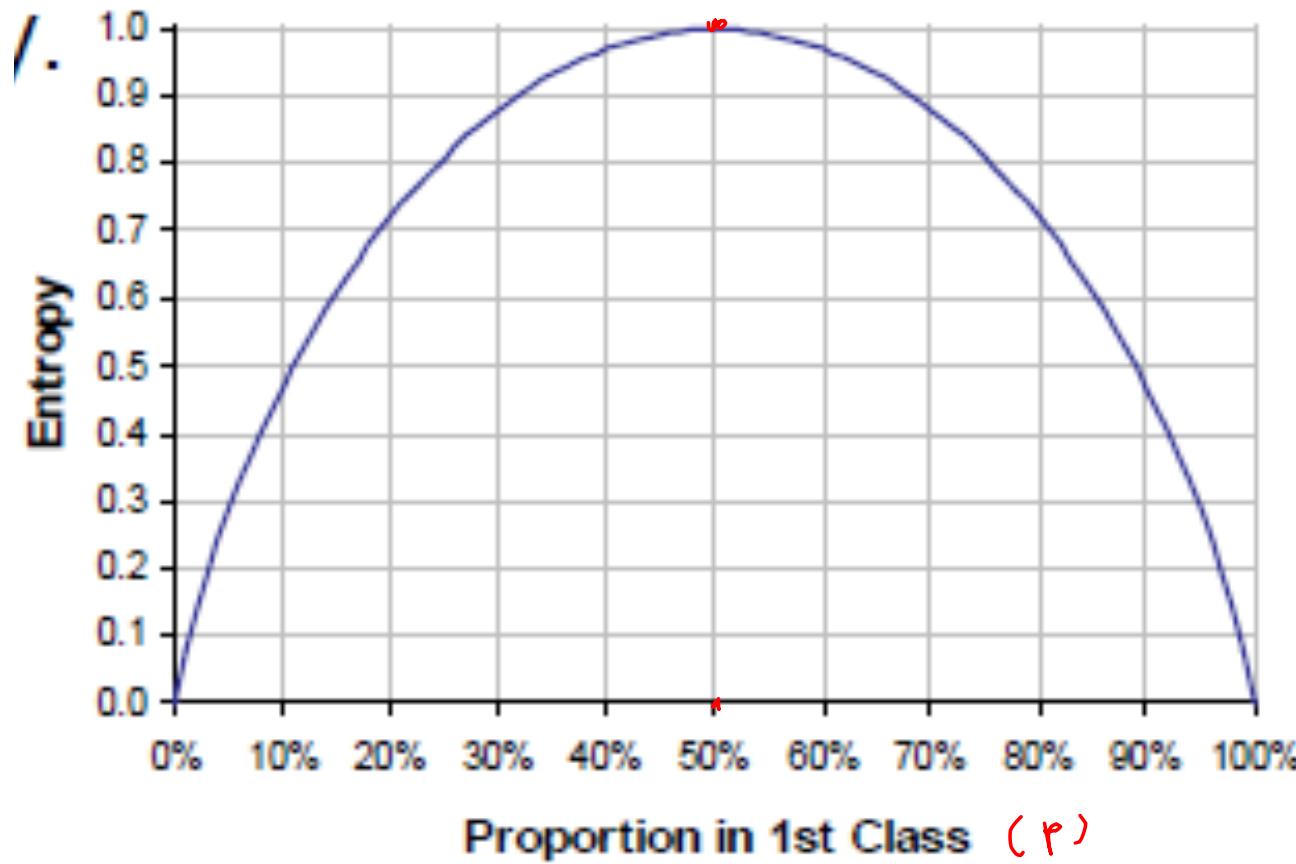
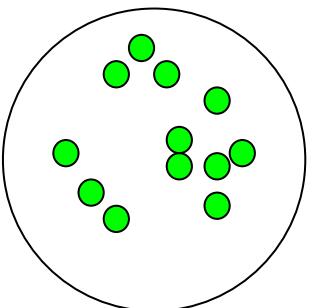
Very impure group



Less impure



Minimum impurity



Entropi

$$E(D) = -p \log_2(p) - (1-p) \log_2(1-p)$$

## Ilustrasi

Tertarik Beli (Y)

Tidak	Tertarik	Total
750	334	1084
0.6919	0.3081	

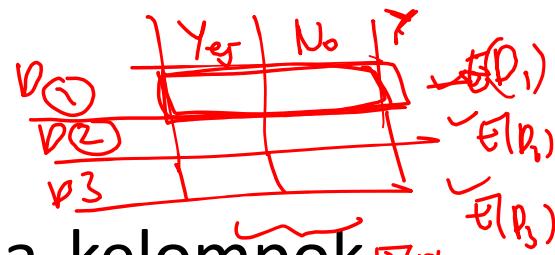
Entropi

$$E(D) = -p \log_2(p) - (1-p) \log_2(1-p)$$

$$\begin{aligned} E(D) &= -0.3081 (\log_2(0.3081)) \\ &\quad - (0.6919) (\log_2(0.6919)) \\ &= 0.4910 \end{aligned}$$

$$\begin{aligned} E(D) &= -p \log_2(p) - (1-p) \log_2(1-p) \\ &= -0.3081 \log_2(0.3081) - 0.6919 \log_2(0.6919) \\ &= 0.4910 \end{aligned}$$

## Information Gain



- Andaikan sebuah gugus data  $D$  dibagi menjadi beberapa kelompok, misalnya  $D_1, D_2, \dots, D_k$  berdasarkan variabel prediktor  $V$

- Dari setiap  $D_i$  bisa dihitung entropinya, yaitu  $E(D_i)$

- Information Gain adalah

$$IG(D, V) = E(D) - \sum_{i=1}^k \frac{|D_i|}{|D|} E(D_i)$$

- Pemisahan yang menghasilkan kelompok-kelompok yang homogen → memiliki information gain yang semakin besar

# Entropy & Information Gain

Frequency  
Percent  
Row Pct  
Col Pct

Table of Jenis_Kelamin by Tertarik_Beli			
Jenis_Kelamin(Jenis Kelamin)	Tertarik_Beli(Tertarik Beli)		
	tidak	tertarik	Total
perempuan $D_1$	561	27	588 ✓
	51.75	2.49	54.24
	95.41	4.59	
	74.80	8.08	
laki-laki $D_2$	189	307	496
	17.44	28.32	45.76
	38.10	61.90	
	25.20	91.92	
Total		750	1084
		69.19	100.00

X

Pada tabel kontingensi di samping, berapa nilai IG?

$$E(D) = 0,8910$$

$$E(D_1) = -0,0159 \log_2(0,0159) - 0,9541 \log_2(0,9541) \\ = A$$

$$E(D_2) = -0,619 \log_2(0,619) - 0,381 \log_2(0,381) \\ = B$$

$$IG = 0,8910 - \left( \frac{588}{1084} \cdot A + \frac{496}{1084} \cdot B \right)$$

$$E(D) = -p \log_2(p) - (1-p) \log_2(1-p)$$

$$IG(D, V) = E(D) - \sum_{i=1}^k \frac{|D_i|}{|D|} E(D_i)$$

# Entropy & Information Gain

Frequency  
Percent  
Row Pct  
Col Pct

Table of Jenis_Kelamin by Tertarik_Beli			
Jenis_Kelamin(Jenis Kelamin)	Tertarik_Beli(Tertarik Beli)		
	tidak	tertarik	Total
perempuan	561	27	588
	51.75	2.49	54.24
	95.41	4.59	
	74.80	8.08	
laki-laki	189	307	496
	17.44	28.32	45.76
	38.10	61.90	
	25.20	91.92	
Total	750	334	1084
	69.19	30.81	100.00

$$IG(D,V) = E(D) - \sum_{i=1}^k \frac{|D_i|}{|D|} E(D_i)$$

$$\begin{aligned} E(\text{TOTAL}) &= -p \log_2(p) - (1-p) \log_2(1-p) \\ &= -0.3081 \log_2(0.3081) - 0.6919 \log_2(0.6919) \\ &= 0.8910 \end{aligned}$$

$$\begin{aligned} E(\text{Perempuan}) &= -p \log_2(p) - (1-p) \log_2(1-p) \\ &= -0.0459 \log_2(0.0459) - 0.9541 \log_2(0.9541) \\ &= 0.2687 \end{aligned}$$

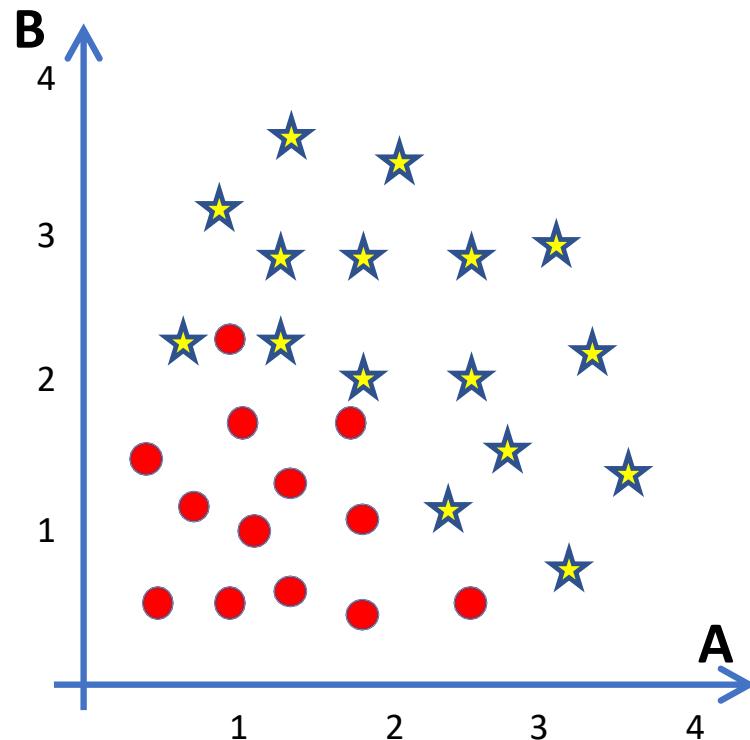
$$\begin{aligned} E(\text{Laki-Laki}) &= -p \log_2(p) - (1-p) \log_2(1-p) \\ &= -0.6190 \log_2(0.6190) - 0.3810 \log_2(0.3810) \\ &= 0.9587 \end{aligned}$$

Information Gain dari pemisahan berdasarkan Jenis Kelamin

$$\begin{aligned} IG &= 0.8910 - (588/1084 * 0.2687 + 496/1084 * 0.9587) \\ &= 0.8910 - 0.5845 \\ &= 0.3065 \end{aligned}$$

## 2. Pengenalan Algoritma Dasar Pohon Klasifikasi

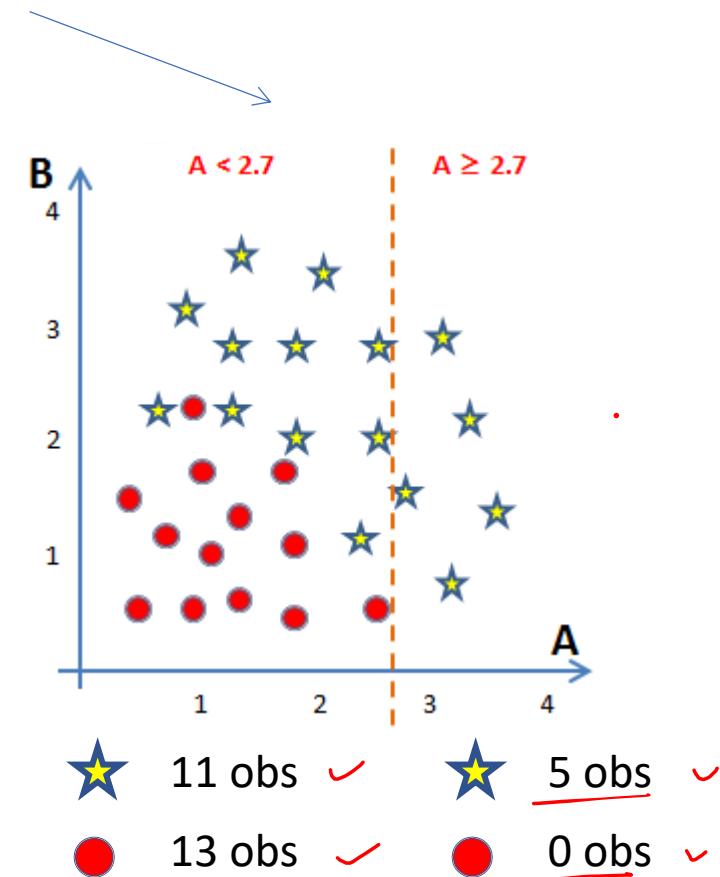
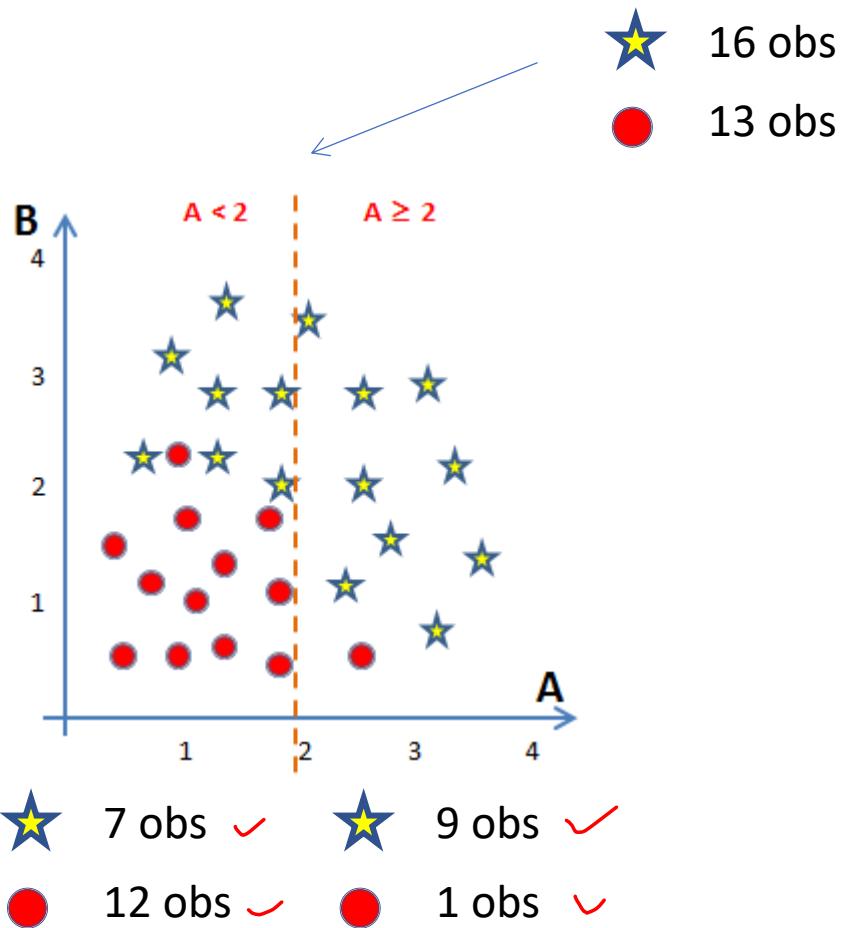
### Ide Dasar



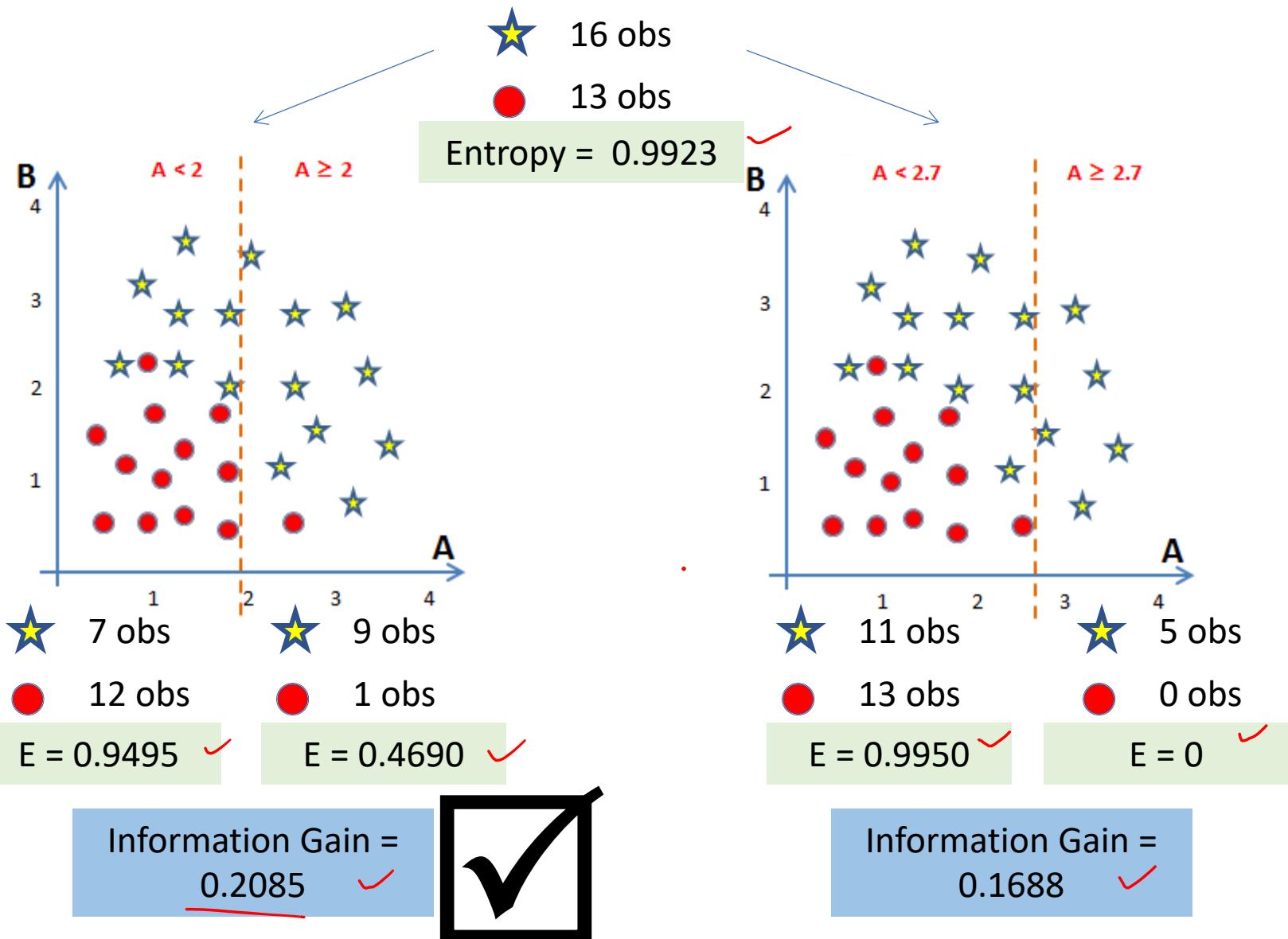
Mencari pemisah terbaik antara individu dengan individu

Pemisahan dilakukan untuk masing-masing variabel, bukan kombinasinya.

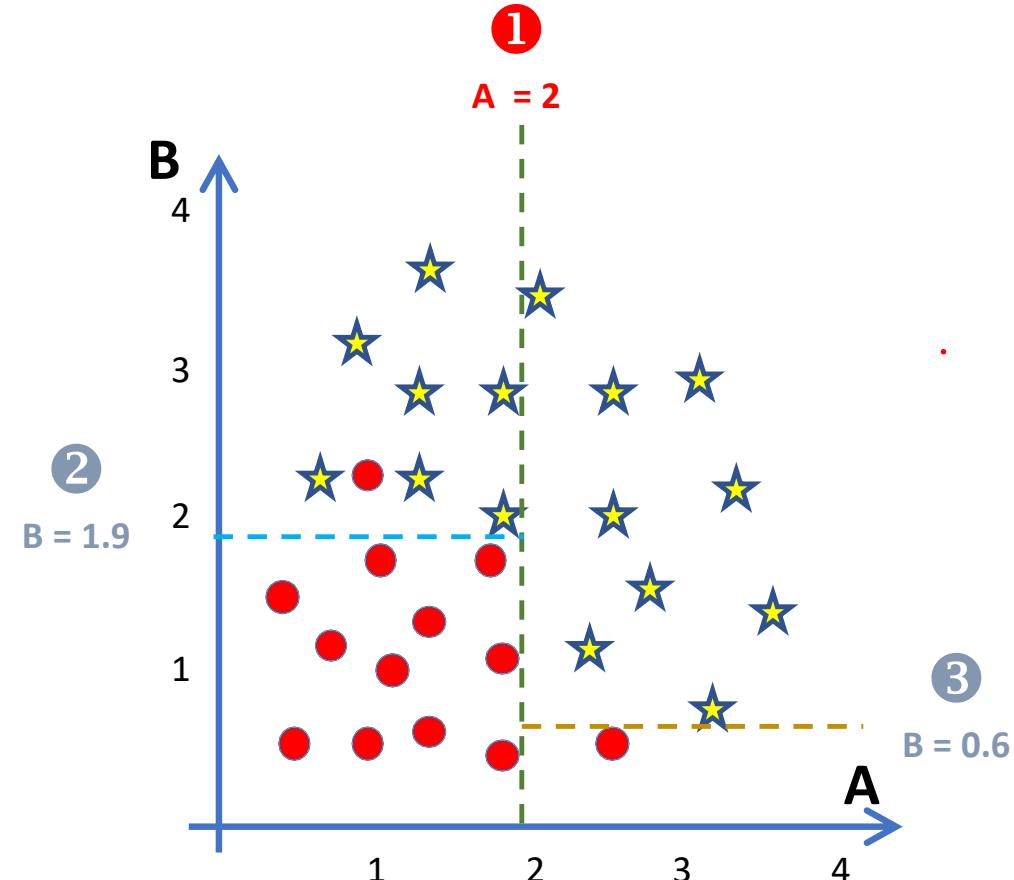
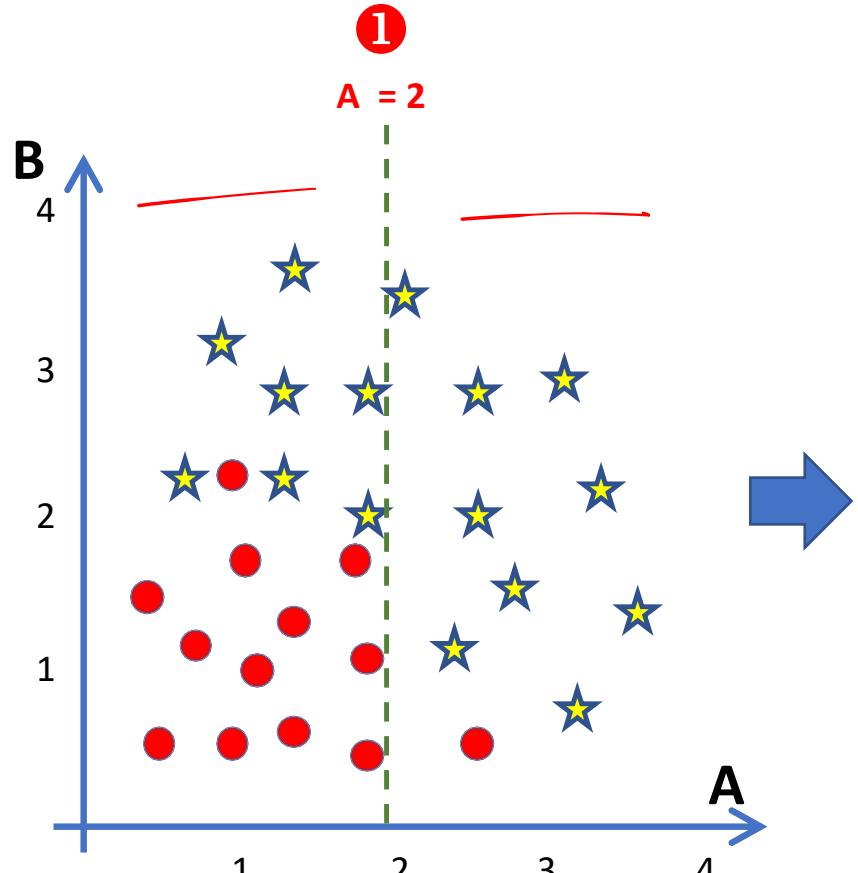
Pemisah yang dicari adalah yang menyebabkan data hasil pemisahannya bersifat homogen kelasnya.



Mana yang lebih baik?

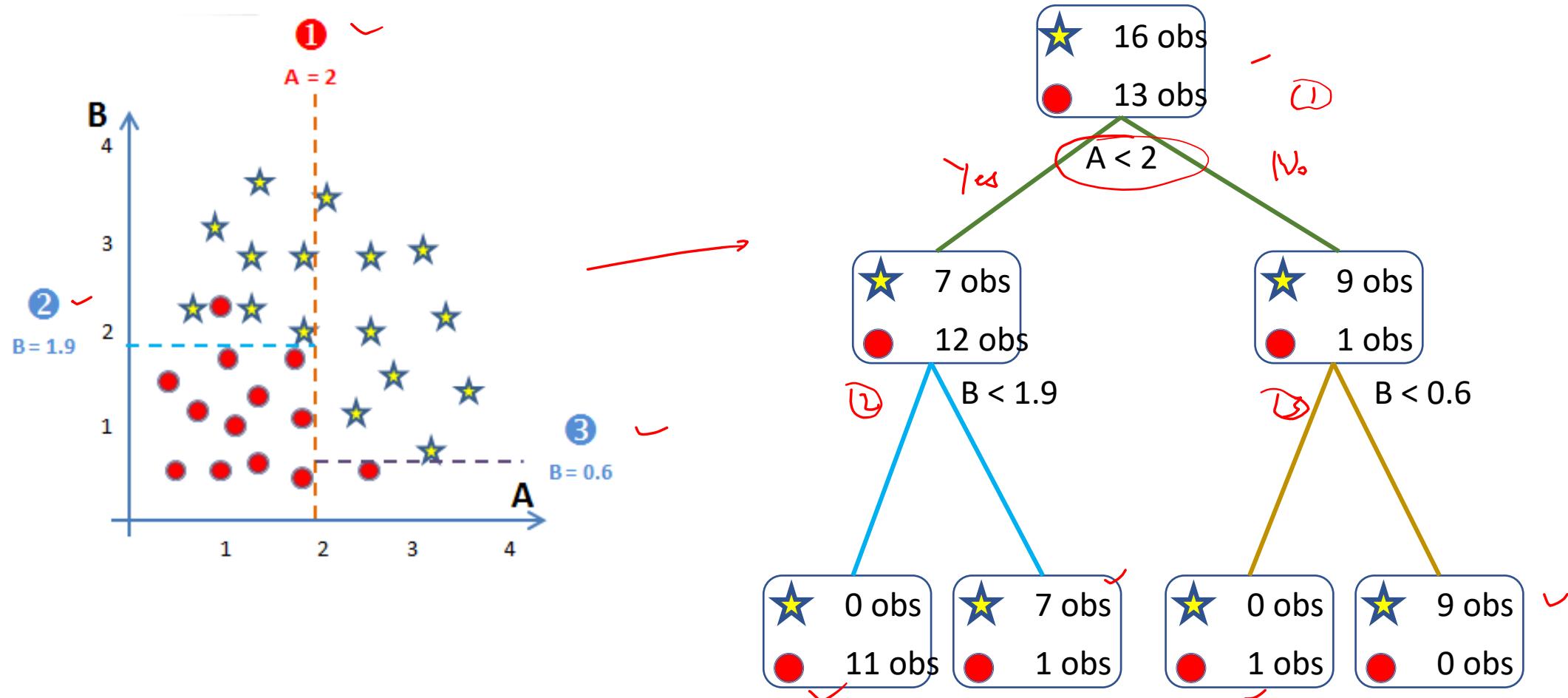


# Ide Dasar: Tahapan

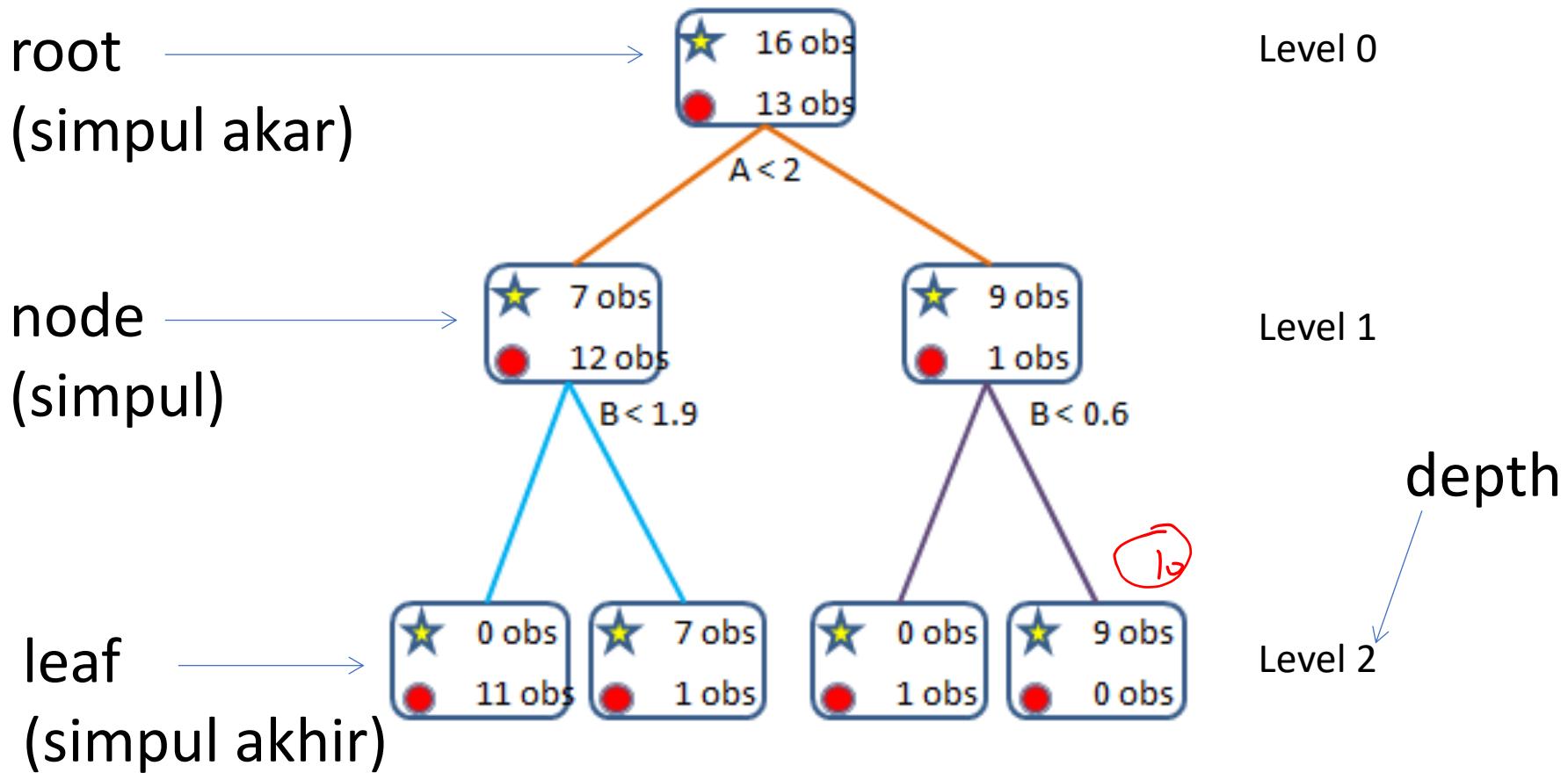


Lanjutkan mencari pemisahan untuk masing-masing kelompok....

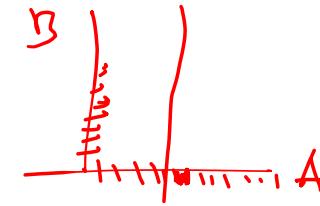
# Representasi Hasil Pemisahan



# Beberapa istilah



# Algoritma Dasar Pohon Klasifikasi



- Tahap 1:  
Mencari pemisahan/penyekatan (splitting) terbaik di setiap variabel
- Tahap 2:  
Menentukan variabel terbaik untuk penyekatan
- Tahap 3:  
Melakukan penyekatan berdasarkan hasil dari Tahap 2, dan memeriksa apakah sudah waktunya menghentikan proses

Lakukan tiga tahapan di atas untuk setiap simpul dan hasil sekatannya

- Proses pemisahan akan berhenti dengan kriteria:
  1. Simpul berisi amatan yang berasal dari satu kelas variabel respon ✓
  2. Simpul berisi amatan yang seluruh variabel prediktornya identik ✓
  3. Simpul berisi amatan yang kurang dari ukuran simpul minimal yang ditentukan di awal ✓
  4. Kedalaman pohon sudah mencapai kedalaman maksimal ✓

# Ilustrasi Sederhana

- Gunakan “datatree01.csv”
- Variabel respon:
  - Tertarik\_Beli
- Variabel Prediktor:
  - Jenis\_Kelamin
  - Single
  - Perokok
  - Tinggal
  - Usia
  - Budget

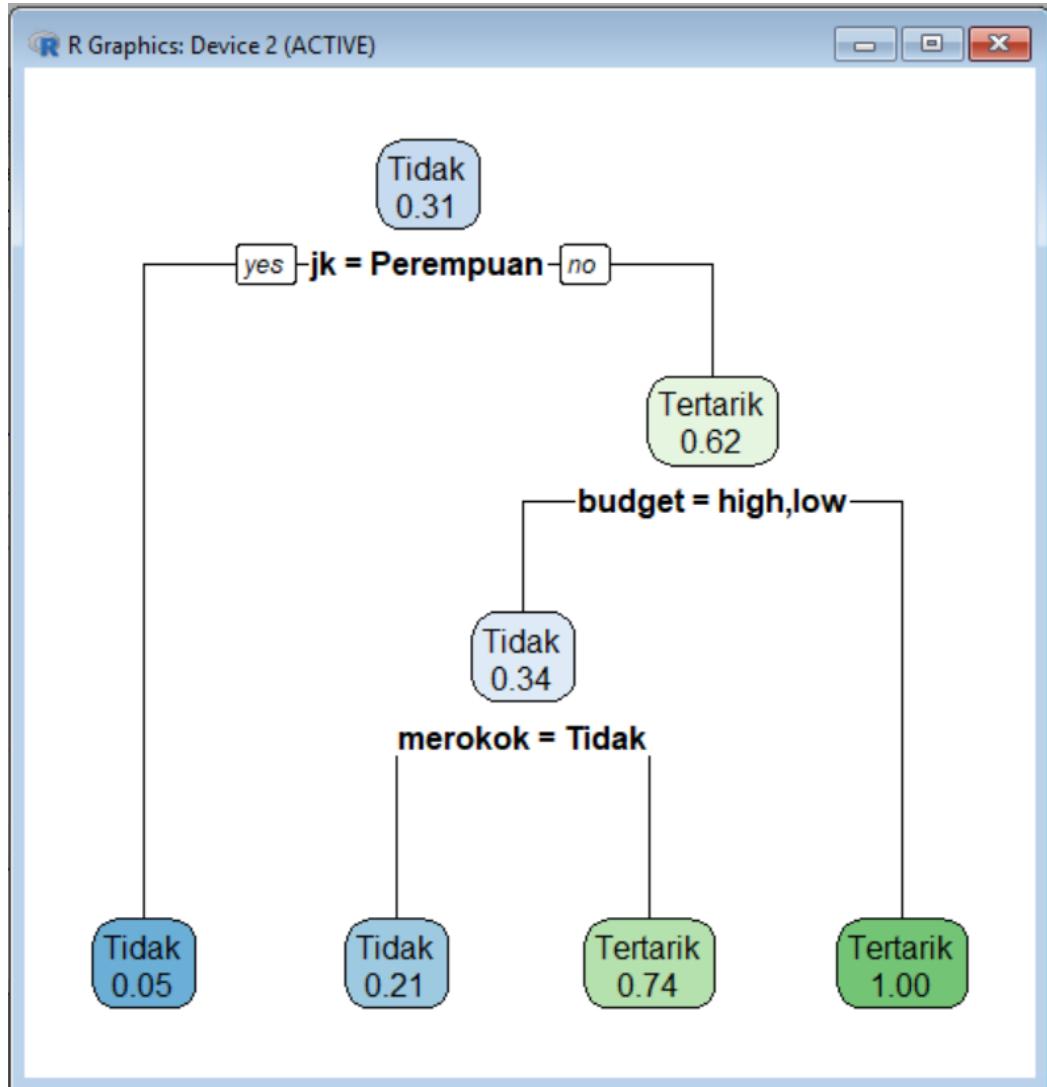
```
##Classification Tree

propensity <- read.csv("D:/datatree01.csv", sep=";", header=TRUE)
tertarik <- factor(propensity$Tertarik.Beli., levels = 0:1, labels = c("Tidak", "Tertarik"))
jk <- factor(propensity$Jenis.Kelamin,    levels = 0:1, labels = c("Perempuan", "Laki-Laki"))
kota <- factor(propensity$Tinggal.di.Kota, levels = 0:1, labels = c("Tidak", "Ya"))
single <- factor(propensity$Single, levels = 0:1, labels = c("Menikah", "Single"))
merokok <- factor(propensity$Perokok, levels = 0:1, labels = c("Tidak", "Ya"))
budget <- propensity$Budget
usia <- propensity$usia

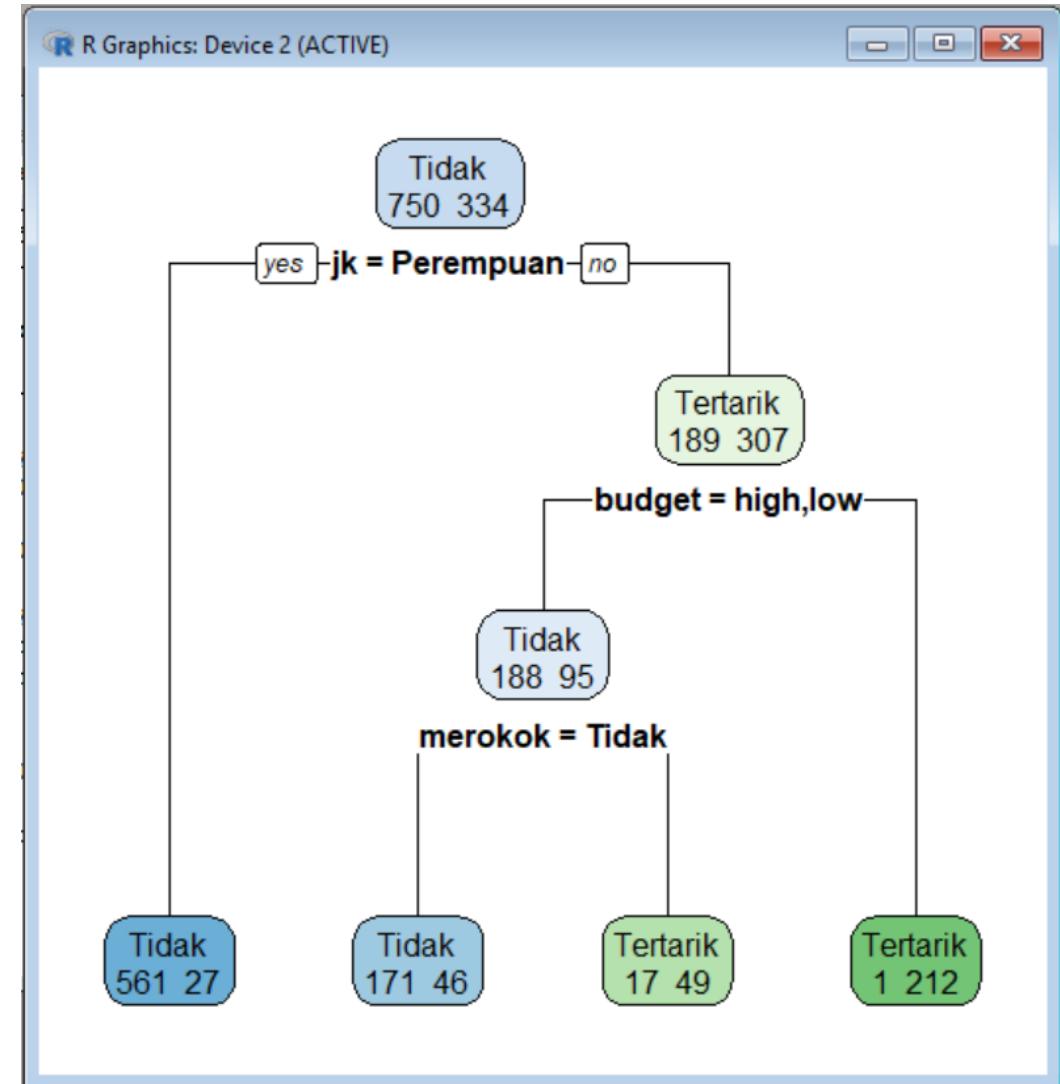
library(rpart)

model.01 <- rpart(tertarik ~  jk + kota + single + usia + merokok + budget,
                  method="class", control = rpart.control(minsplit = 100))
model.01
```

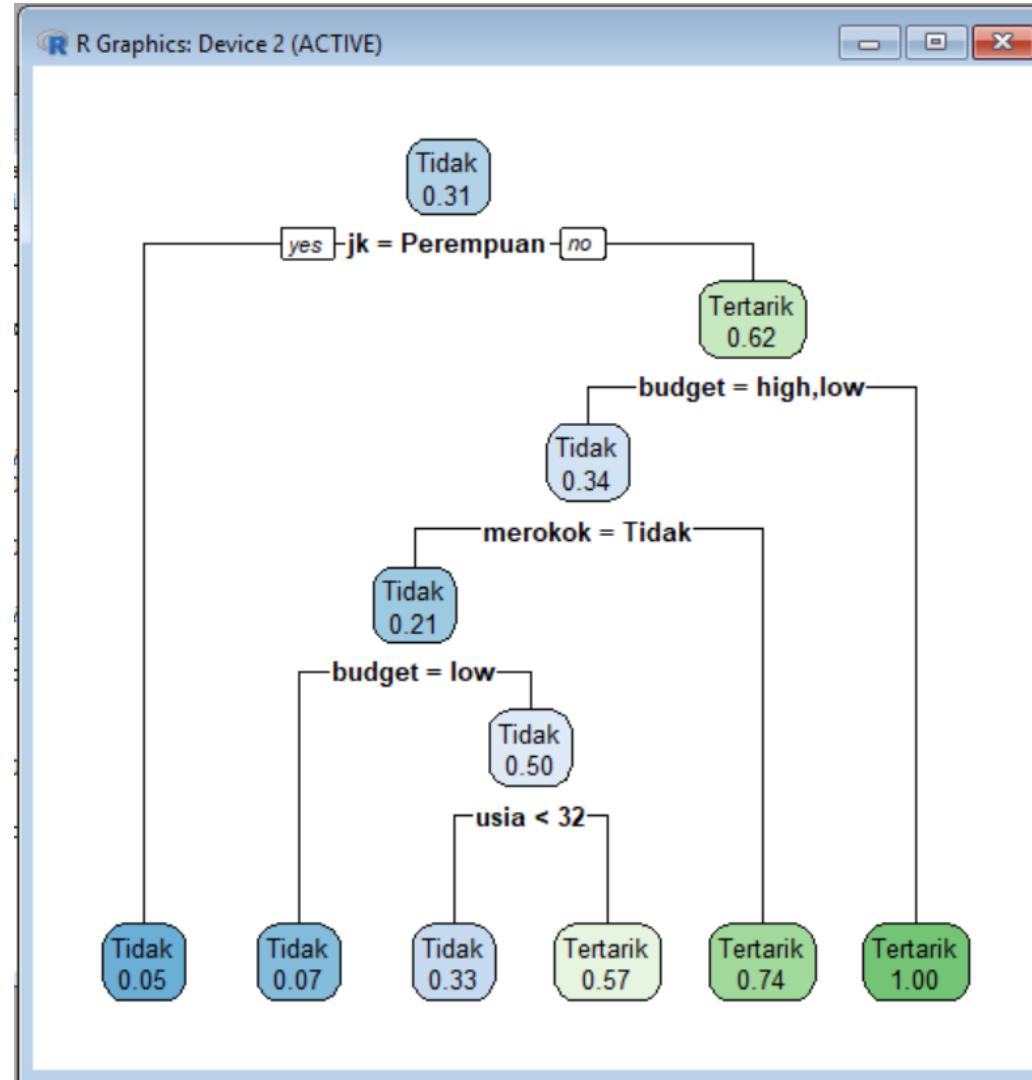
```
library(rpart.plot)
rpart.plot(model.01, extra=6)
```



```
rpart.plot(model.01, extra=1)
```



```
model.02 <- rpart(tertarik ~ jk + kota + single + usia + merokok + budget,  
                  method="class", control = rpart.control(minsplit = 50))  
rpart.plot(model.02, extra=6)
```



### 3. Menilai Kemampuan Prediksi Pohon Klasifikasi

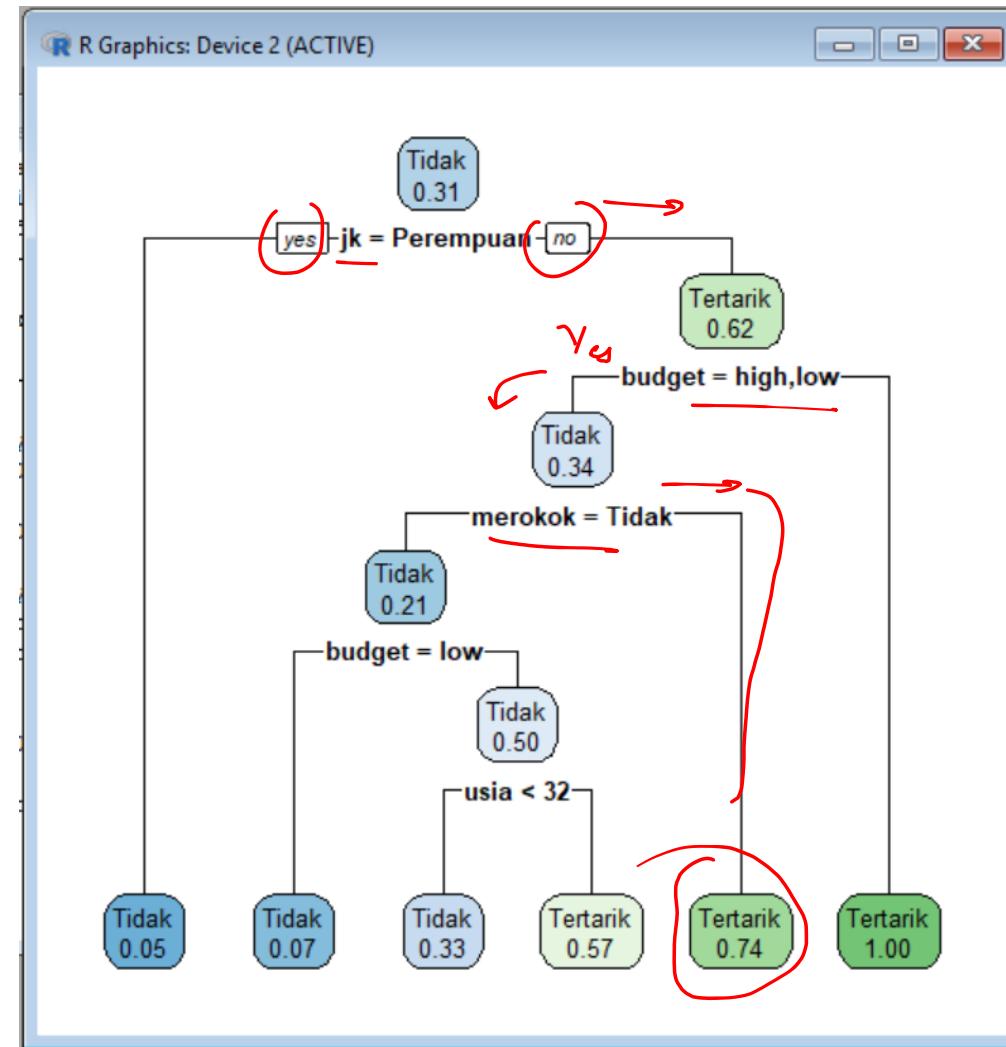
#### Prediksi Peubah Respon

- Untuk setiap individu yang diketahui nilai-nilai variabel prediktor yang muncul pada pohon klasifikasi, kita dapat melakukan prediksi kelas variabel respon. Misalnya jika diketahui usia, jenis kelamin, apakah merokok, dan klasifikasi budget dari seseorang, maka kita dapat memprediksi apakah orang tersebut akan tertarik atau tidak.
- Bagaimana caranya? Gunakan alur pencabangan yang ada pada pohon klasifikasi sampai berhenti di simpul akhir. Berdasarkan simpul akhir itulah kita prediksi dia masuk ke kategori apa.

## Prediksi Peubah Respon

Misal

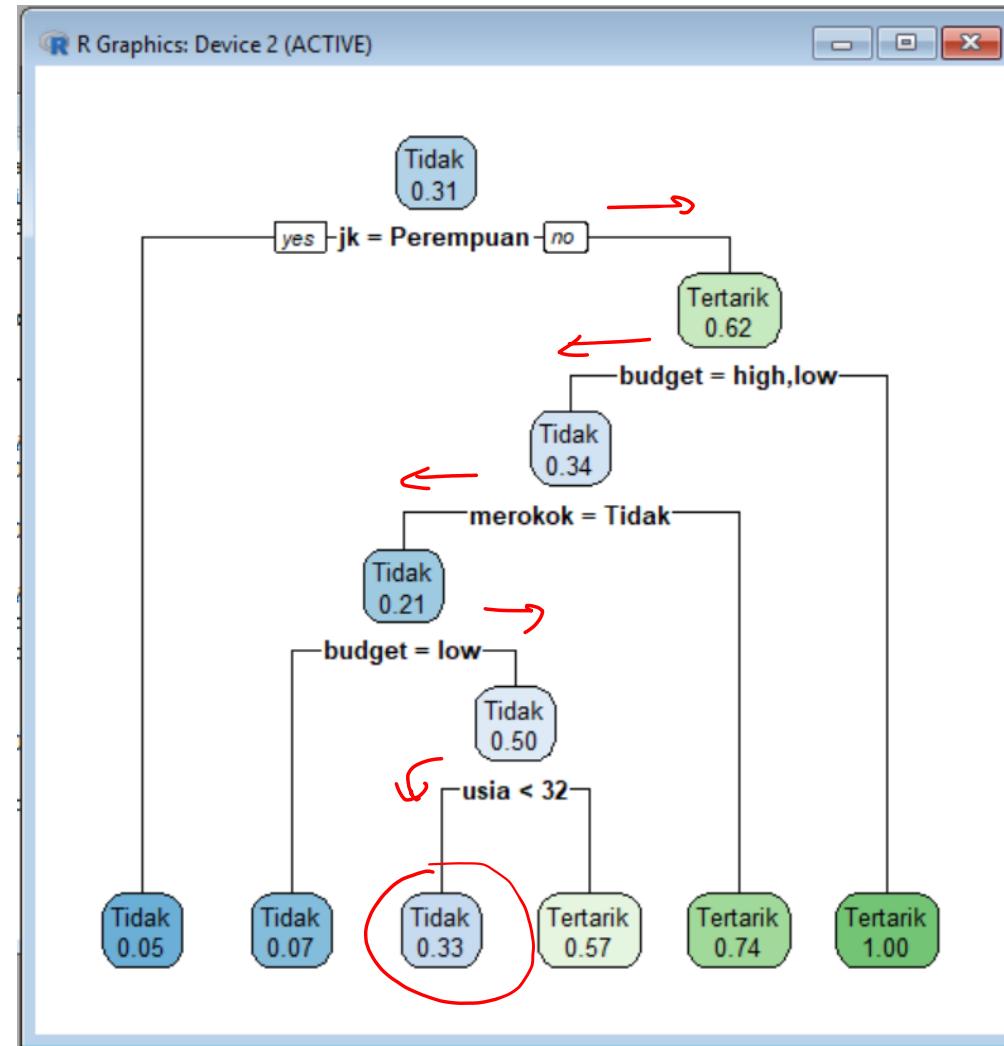
- Jenis Kelamin = Laki-Laki ✓
- Budget = Low ✓
- Merokok = Ya ✓
- Usia 25 tahun
- Tinggal di Kota
- Single
- Probability TERTARIK = 0.74 ✓



## Prediksi Peubah Respon

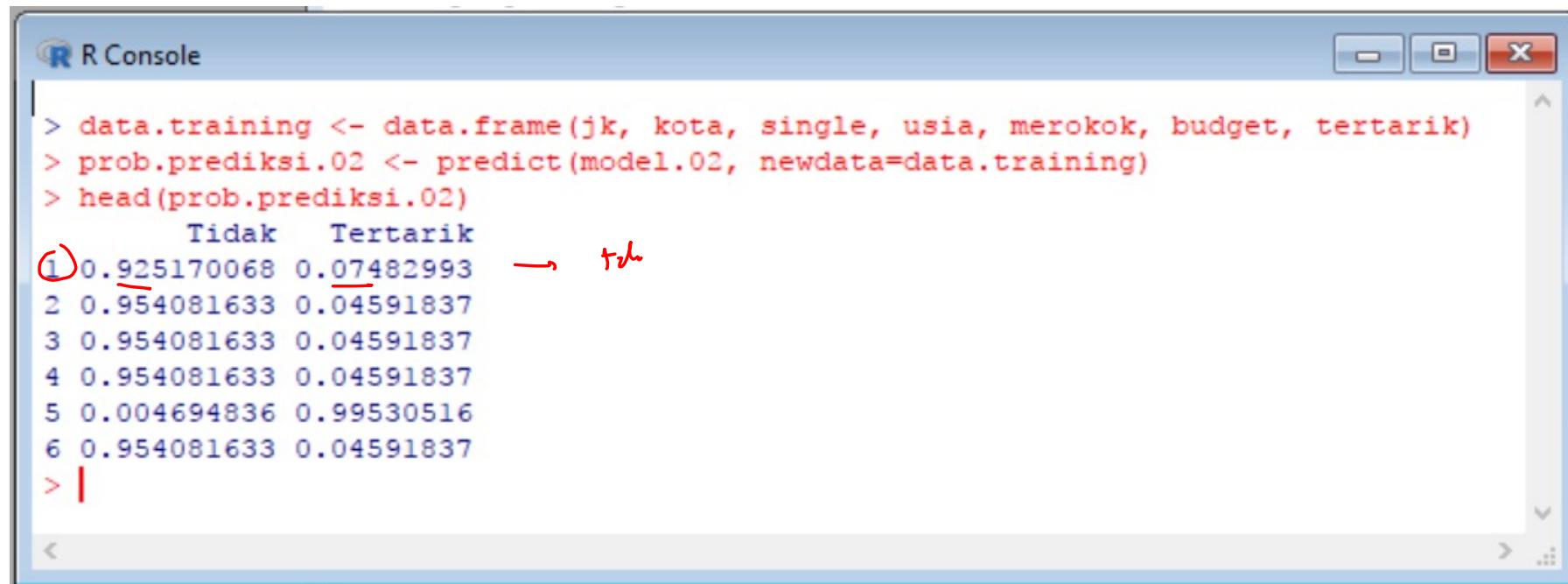
Misal

- Jenis Kelamin = Laki-Laki
  - Budget high
  - Tidak Merokok
  - Usia 25 tahun ↘
  - Tinggal di Kota
  - Single
- 
- Probability TERTARIK = 0.33



## Prediksi Peubah Respon

```
data.training <- data.frame(jk, kota, single, usia, merokok, budget, tertarik)
prob.prediksi.02 <- predict(model.02, newdata=data.training)
head(prob.prediksi.02)
```



The screenshot shows the R Console window with the following output:

```
> data.training <- data.frame(jk, kota, single, usia, merokok, budget, tertarik)
> prob.prediksi.02 <- predict(model.02, newdata=data.training)
> head(prob.prediksi.02)
      Tidak Tertarik
1 0.925170068 0.07482993 → tidak
2 0.954081633 0.04591837
3 0.954081633 0.04591837
4 0.954081633 0.04591837
5 0.004694836 0.99530516
6 0.954081633 0.04591837
> |
```

The first row of the output is annotated with a red circle around the value 0.925170068, followed by an arrow pointing to the word "tidak".

## Prediksi Peubah Respon

Andaikan digunakan batasan 0.5 untuk mengelompokkan ketertarikan, sehingga kalau

Prob(Tertarik) > 0.5 → Tertarik ✓

Prob(Tertarik) ≤ 0.5 → tidak ✓

Maka kita akan dapatkan

	Tidak	Tertarik	Prediksi
1	0.925170068	0.07482993	→ Tidak ✓
2	0.954081633	0.04591837	→ Tidak ✓
3	0.954081633	0.04591837	→ Tidak ✓
4	0.954081633	0.04591837	→ Tidak ✓
5	0.004694836	0.99530516	→ Tertarik ✓
6	0.954081633	0.04591837	→ Tidak ✓
7	0.004694836	0.99530516	→ Tertarik ✓
8	0.925170068	0.07482993	→ Tidak ✓
9	0.954081633	0.04591837	→ Tidak ✓
10	0.428571429	0.57142857	→ Tertarik ✓

Perbandingan antara respon yang sebenarnya dengan dugaan

	Tertarik_beli	dugaan	prediksi
1	Tidak	Tidak	Tidak
2	Tidak	Tidak	Tidak
3	Tidak	Tidak	Tidak
4	Tidak	Tidak	Tidak
5	Tertarik	Tertarik	Tertarik
6	Tidak	Tidak	Tidak
7	Tertarik	Tertarik	Tertarik
8	Tidak	Tidak	Tidak
9	Tidak	Tidak	Tidak
10	Tidak	✓	Tertarik → salah prediksi ✓

# Classification Table

		Predicted Class	
		0	1
Actual Class	0	True Negative	False Positive
	1	False Negative	True Positive
Predicted Negative		Predicted Positive	

Actual Negative

Actual Positive

Proporsinya harus tinggi

## R Console

```
> confusionMatrix(prediksi.02, tertarik)
```

Confusion Matrix and Statistics

Reference

Prediction Tidak Tertarik

Tidak	711	45
Tertarik	39	289

Accuracy : 0.9225

95% CI : (0.905, 0.9377)

No Information Rate : 0.6919

P-Value [Acc > NIR] : <2e-16

Kappa : 0.8173

McNemar's Test P-Value : 0.5854

Sensitivity : 0.9480

Specificity : 0.8653

Pos Pred Value : 0.9405

Neg Pred Value : 0.8811

Prevalence : 0.6919

Detection Rate : 0.6559

Detection Prevalence : 0.6974

Balanced Accuracy : 0.9066

'Positive' Class : Tidak

prediksi.02 <- factor(ifelse(prob.prediksi.02[,2] > 0.5, 1, 0),  
levels = 0:1, labels = c("Tidak", "Tertarik"))

library(caret) *prediksi.02* *tertarik*  
confusionMatrix(prediksi.02, tertarik)

## Goodness of Classification Tree

Andaikan batas peluangnya diganti dari 0.5 menjadi 0.6

```
R Console
Confusion Matrix and Statistics

      Reference
Prediction Tidak Tertarik
 Tidak      732      73
 Tertarik    18     261

      Accuracy : 0.9161
      95% CI  : (0.8979, 0.9319)
No Information Rate : 0.6919
P-Value [Acc > NIR] : < 2.2e-16

      Kappa : 0.7937
McNemar's Test P-Value : 1.507e-08

      Sensitivity : 0.9760
      Specificity : 0.7814
Pos Pred Value : 0.9093
Neg Pred Value : 0.9355
      Prevalence : 0.6919
Detection Rate : 0.6753
Detection Prevalence : 0.7426
Balanced Accuracy : 0.8787

'Positive' Class : Tidak
```

```
prediksi.02 <- factor(ifelse(prob.prediksi.02[,2] > 0.6, 1, 0),
levels = 0:1, labels = c("Tidak", "Tertarik"))
```

```
library(caret)
confusionMatrix(prediksi.02, tertarik)
```

## Goodness of Classification Tree

Andaikan batas peluangnya diganti dari **0.5** menjadi **0.3**

```
R Console
Confusion Matrix and Statistics

      Reference
Prediction Tidak Tertarik
 Tidak      697      38
Tertarik     53     296

      Accuracy : 0.9161
      95% CI  : (0.8979, 0.9319)
No Information Rate : 0.6919
P-Value [Acc > NIR] : <2e-16

      Kappa : 0.8055
McNemar's Test P-Value : 0.1422

      Sensitivity : 0.9293
      Specificity : 0.8862
      Pos Pred Value : 0.9483
      Neg Pred Value : 0.8481
      Prevalence : 0.6919
      Detection Rate : 0.6430
      Detection Prevalence : 0.6780
      Balanced Accuracy : 0.9078

'Positive' Class : Tidak
```

```
prediksi.02 <- factor(ifelse(prob.prediksi.02[,2] > 0.3, 1, 0),
levels = 0:1, labels = c("Tidak", "Tertarik"))
```

```
library(caret)
confusionMatrix(prediksi.02, tertarik)
```

## Goodness of Classification Tree

0, 1

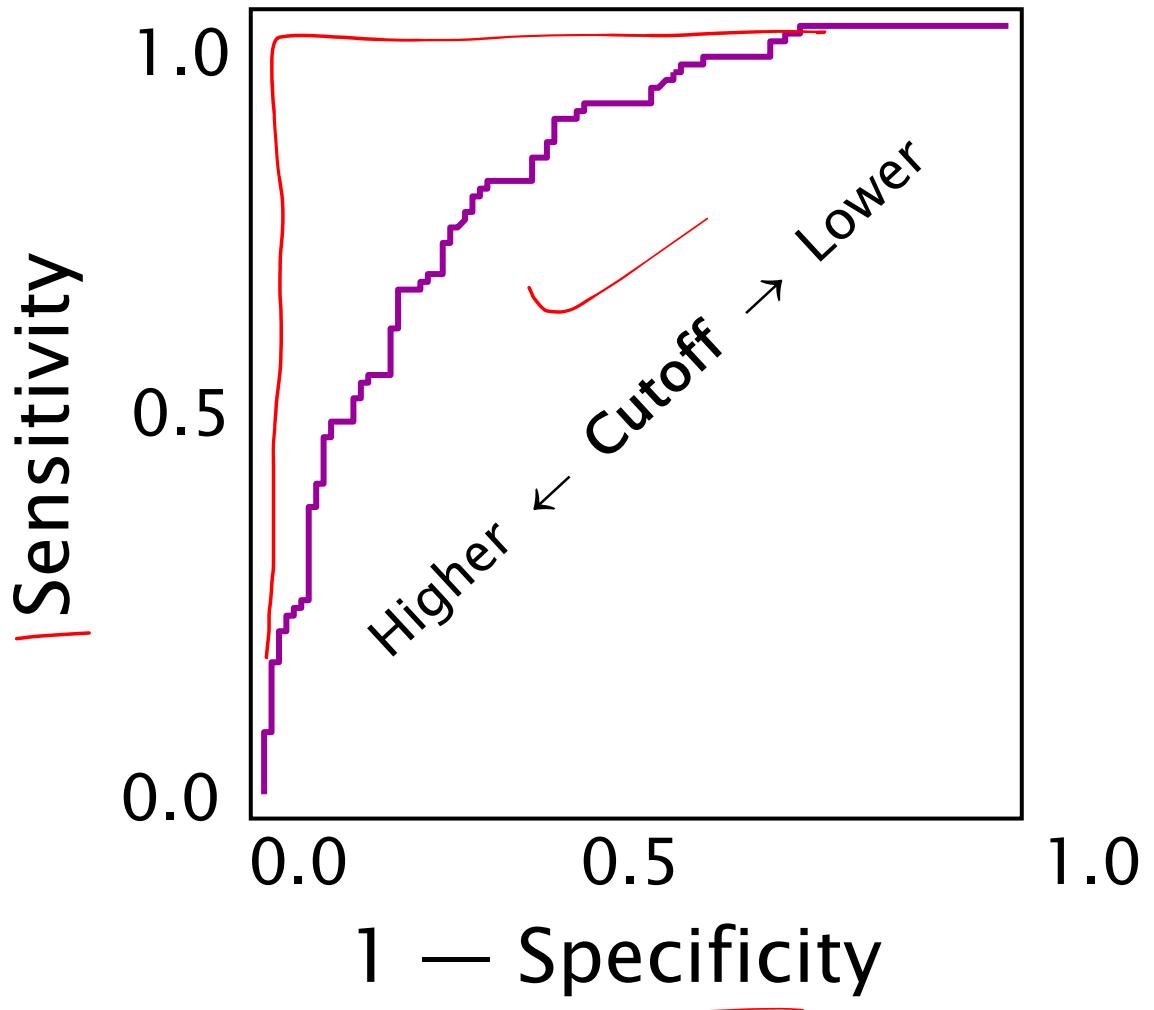
✓

✓

Cut-Off	Accuracy	Sensitivity	Specificity
0.3	91.61%	92.93%	88.62%
<u>0.5</u>	92.25%	94.80%	86.53%
<u>0.6</u>	91.61%	97.60%	78.14%
:			

0, 1

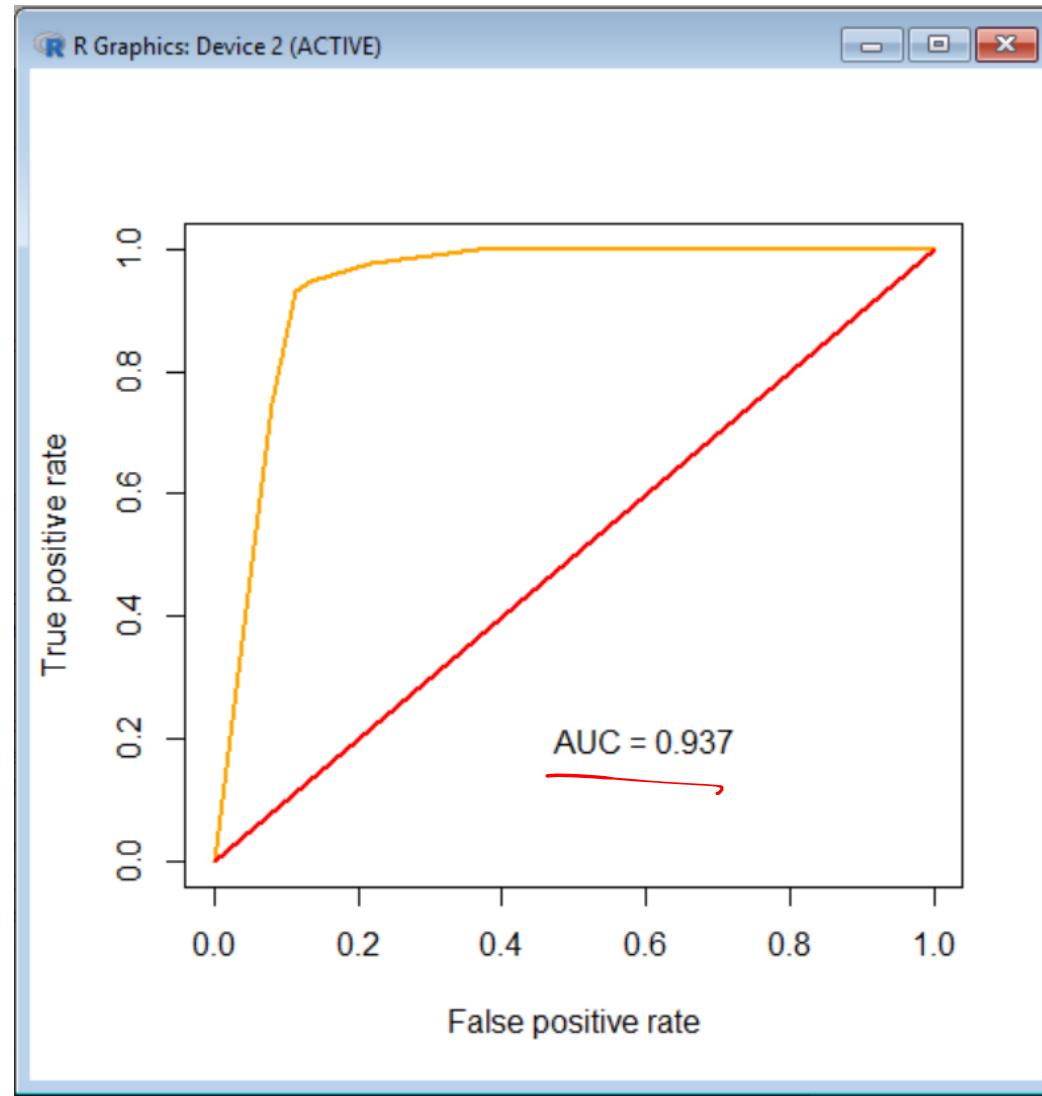
## ROC Curve



- Kurva ROC yang ideal adalah garis vertikal dari asal ke sensitivitas 1, dan kemudian garis horizontal sepanjang sensitivitas 1 untuk semua tingkat ( $1 - \text{spesifisitas}$ )
- Kurva ROC biasanya digunakan untuk membandingkan model pesaing
- Ukuran numerik seberapa dekat kurva ROC cocok dengan kurva ideal dihitung dengan membandingkan area di bawah kurva dengan 1. Ternyata luasnya sama dengan statistik c.

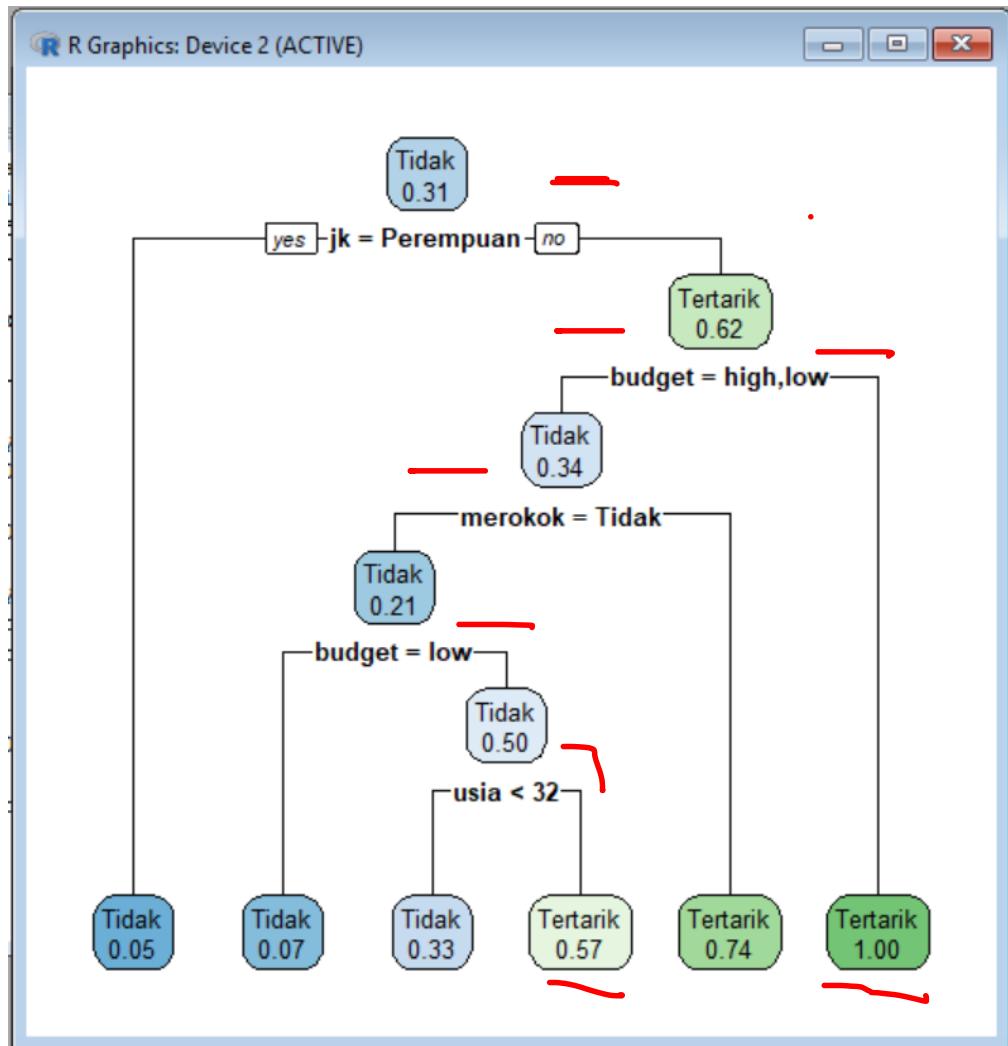
## ROC Curve

```
library(ROCR)  
  
pred <-  
prediction(prob.prediksi.02[,1],  
tertarik)  
  
roc <- performance(pred,  
measure="tpr", x.measure="fpr")  
  
auc <- performance(pred, 'auc')  
AUC <- auc@y.values[[1]]  
  
plot(roc, col="orange", lwd=2)  
lines(x=c(0, 1), y=c(0, 1),  
col="red", lwd=2)  
text(0.6, 0.2, paste0("AUC = ",  
round(AUC,3)) )
```



# Latihan (Isian singkat)

1. Tujuan penggunaan classification tree adalah ...
2. Information Gain yang semakin besar mengindikasikan kelompok yang dihasilkan bersifat ...
3. Nilai entropi suatu gugus data akan bernilai maksimum diperoleh ketika nilai  $p = \dots$
4. Nilai entropi suatu gugus data akan bernilai minimum diperoleh ketika ...
5. Yang dimaksud dengan istilah minimum impurity adalah ...



Diketahui classification tree di samping.

1. Berapa kedalaman (depth) pohon di samping?

Tentukan Probability TERTARIK dengan kondisi:

Budget = Med

Merokok = Ya

Usia 25 tahun

Tinggal di Kota

Jenis Kelamin = Laki-Laki  
Single

Budget = High

Merokok = Tidak

Usia 45 tahun

Tinggal di Kota

Jenis Kelamin = Laki-Laki  
Single

Terima kasih 😊