



Regression Tree

disusun oleh:
Bagus Sartono
bagusco@apps.ipb.ac.id
0852-1523-1823

Prodi Statistika dan Sains Data
Fakultas Matematika dan Ilmu Pengetahuan Alam
Institut Pertanian Bogor

2020



IPB University
— Bogor Indonesia —

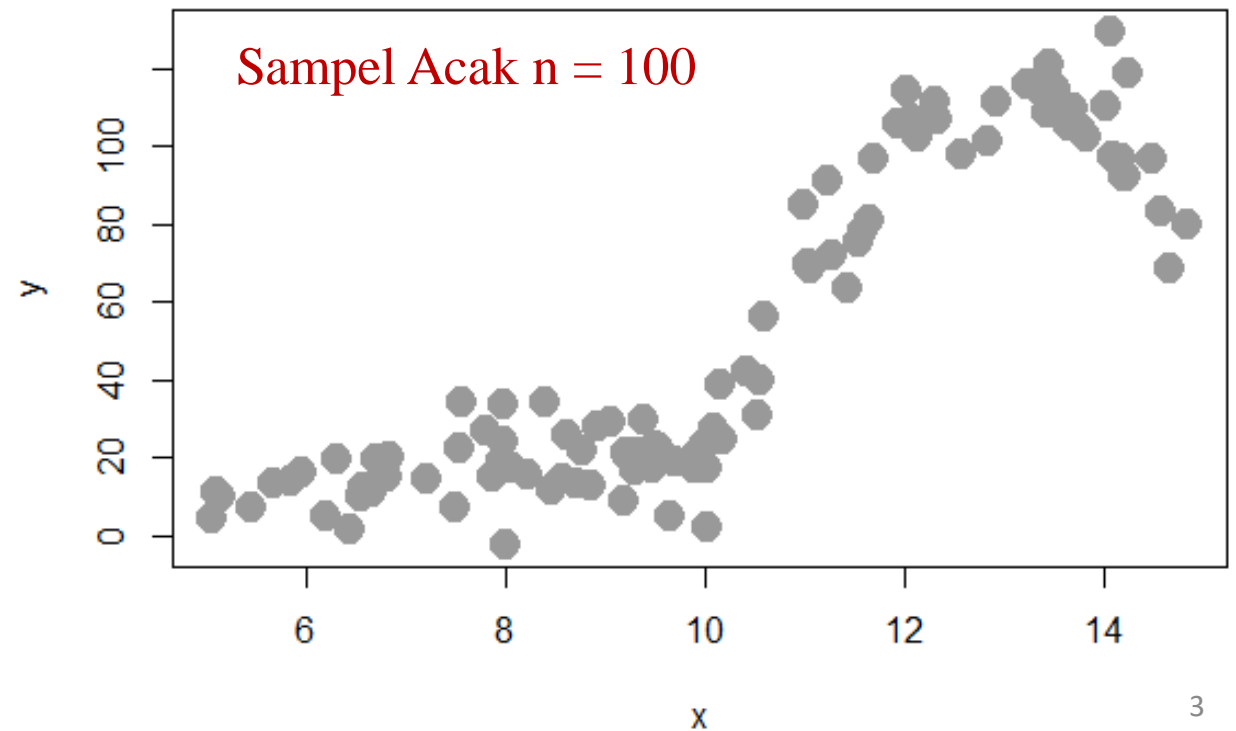
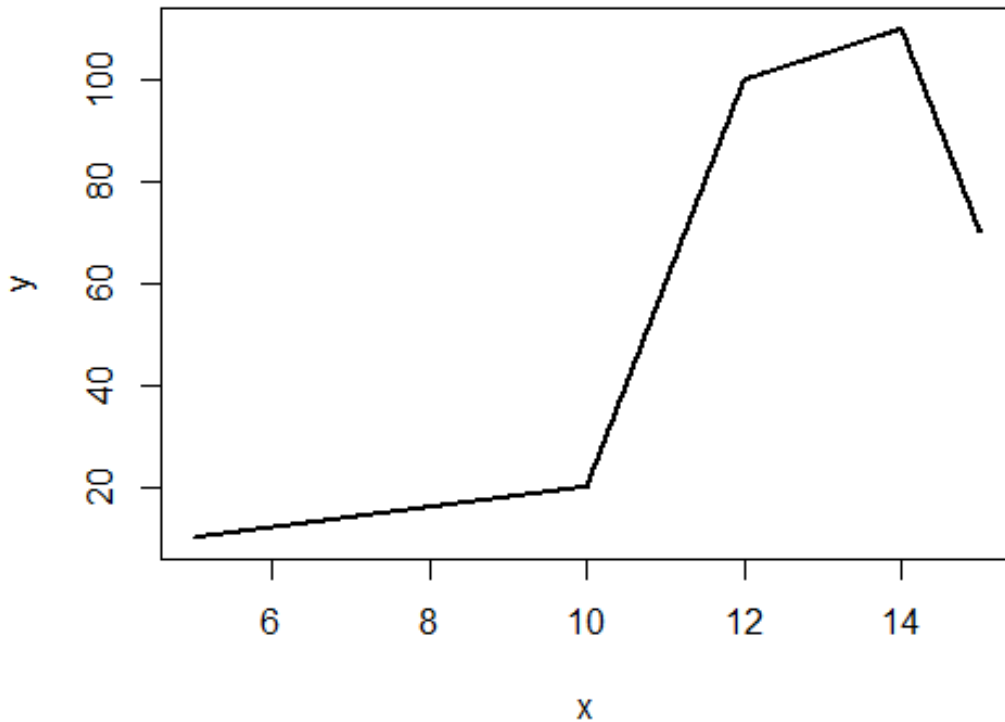


Outline

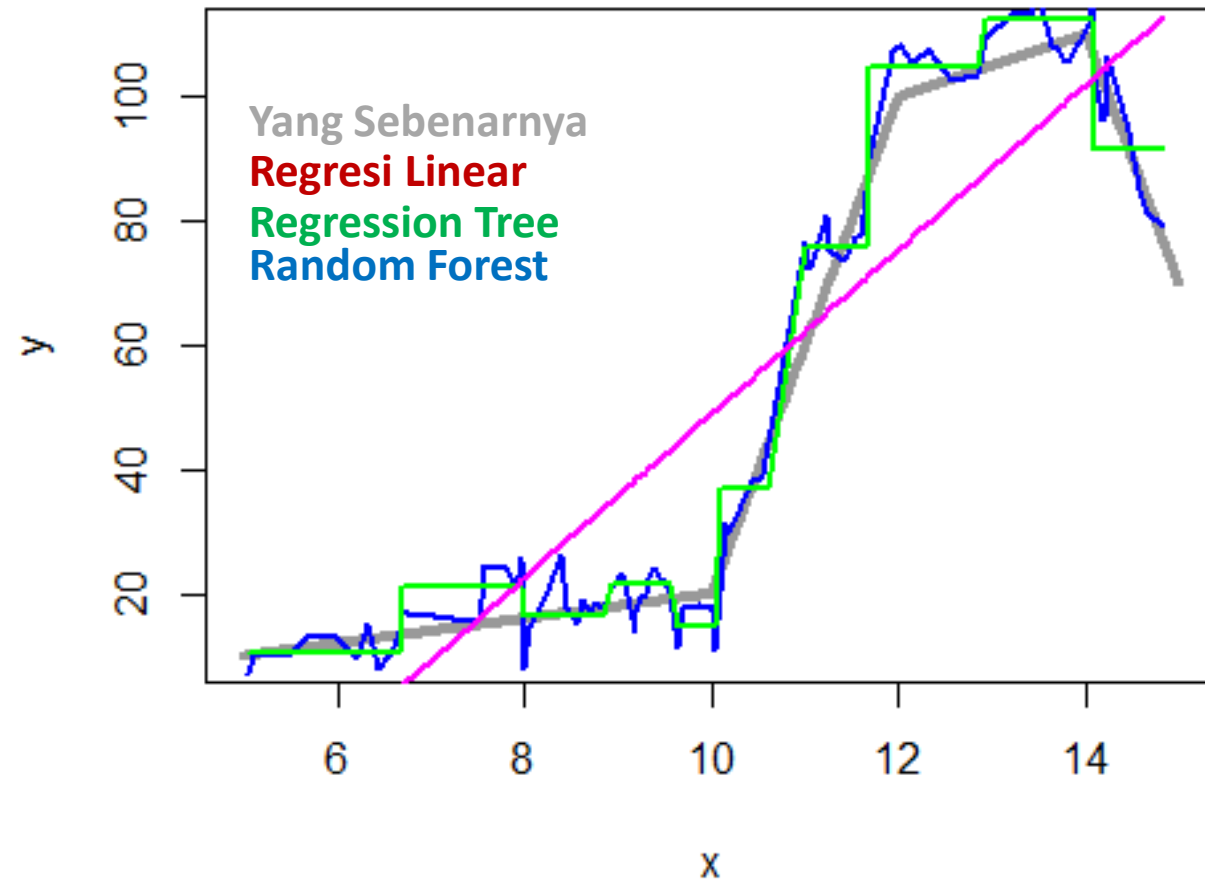
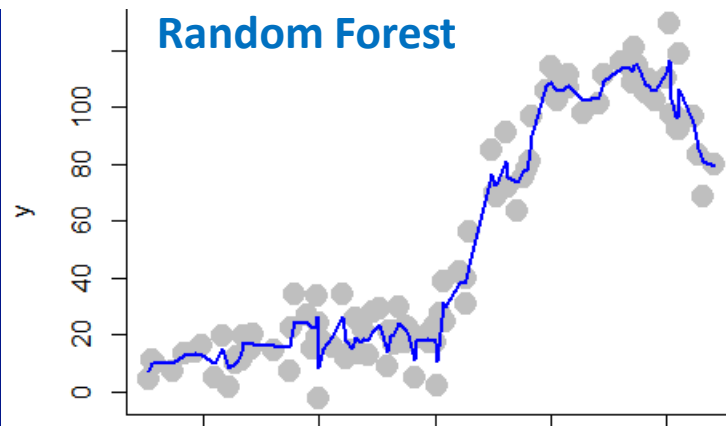
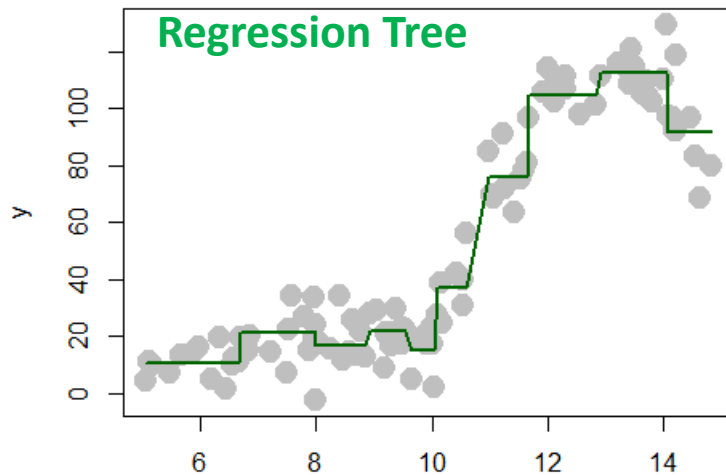
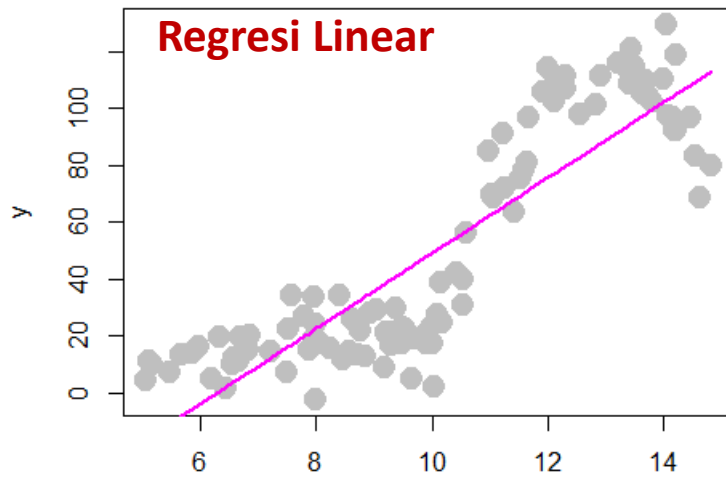
- Pendahuluan
- Bentuk dan Kegunaan Model Regression Tree
- Bagaimana Memperoleh Modelnya?
- Implementasi-nya di R
- Pengenalan Random Forest

$$y = \begin{cases} 2x & \text{untuk } 5 < x \leq 10 \\ -380 + 40x & \text{untuk } 10 < x \leq 12 \\ 40 + 5x & \text{untuk } 12 < x \leq 14 \\ 670 - 40x & \text{untuk } 14 < x < 15 \end{cases}$$

+ Error ~ Normal (0, 10)



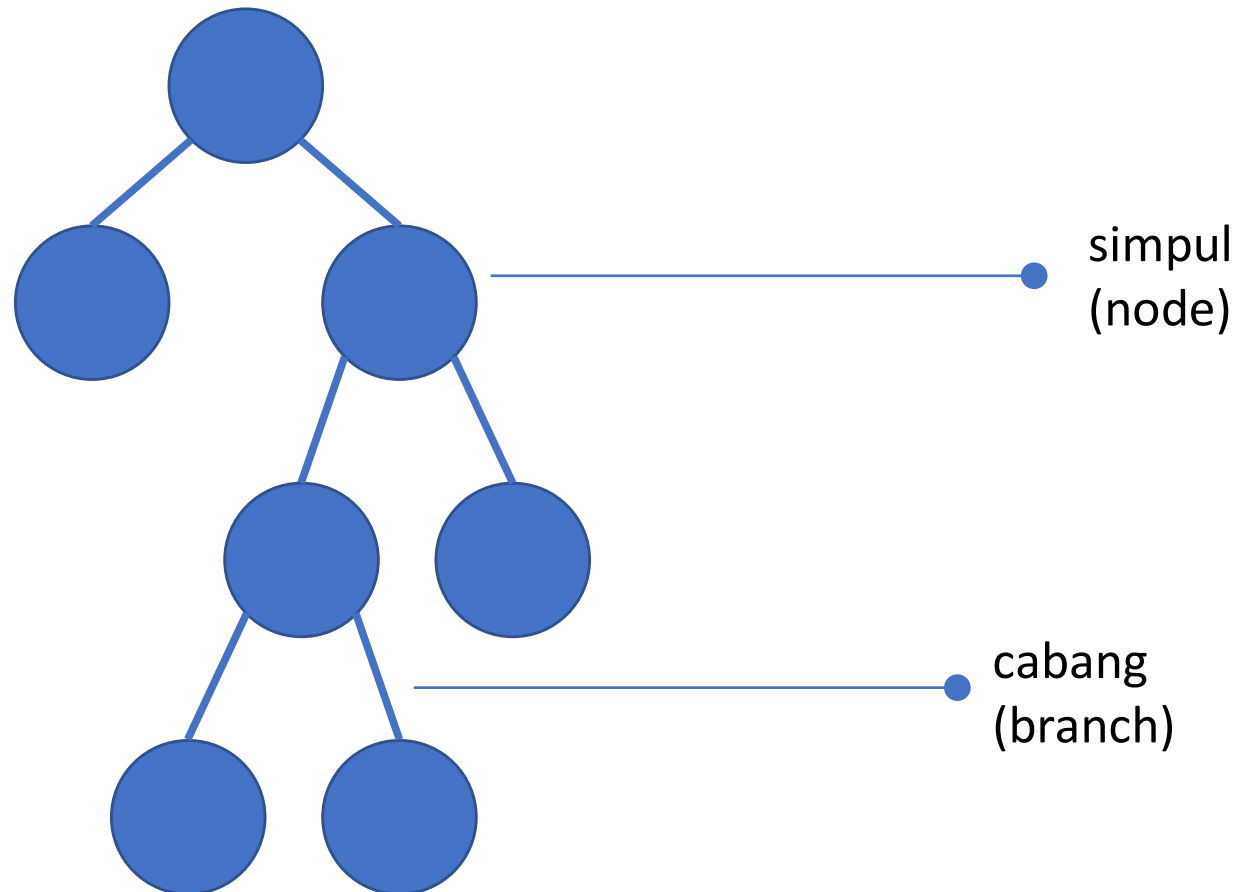
Regresi Linear vs Pohon Regresi



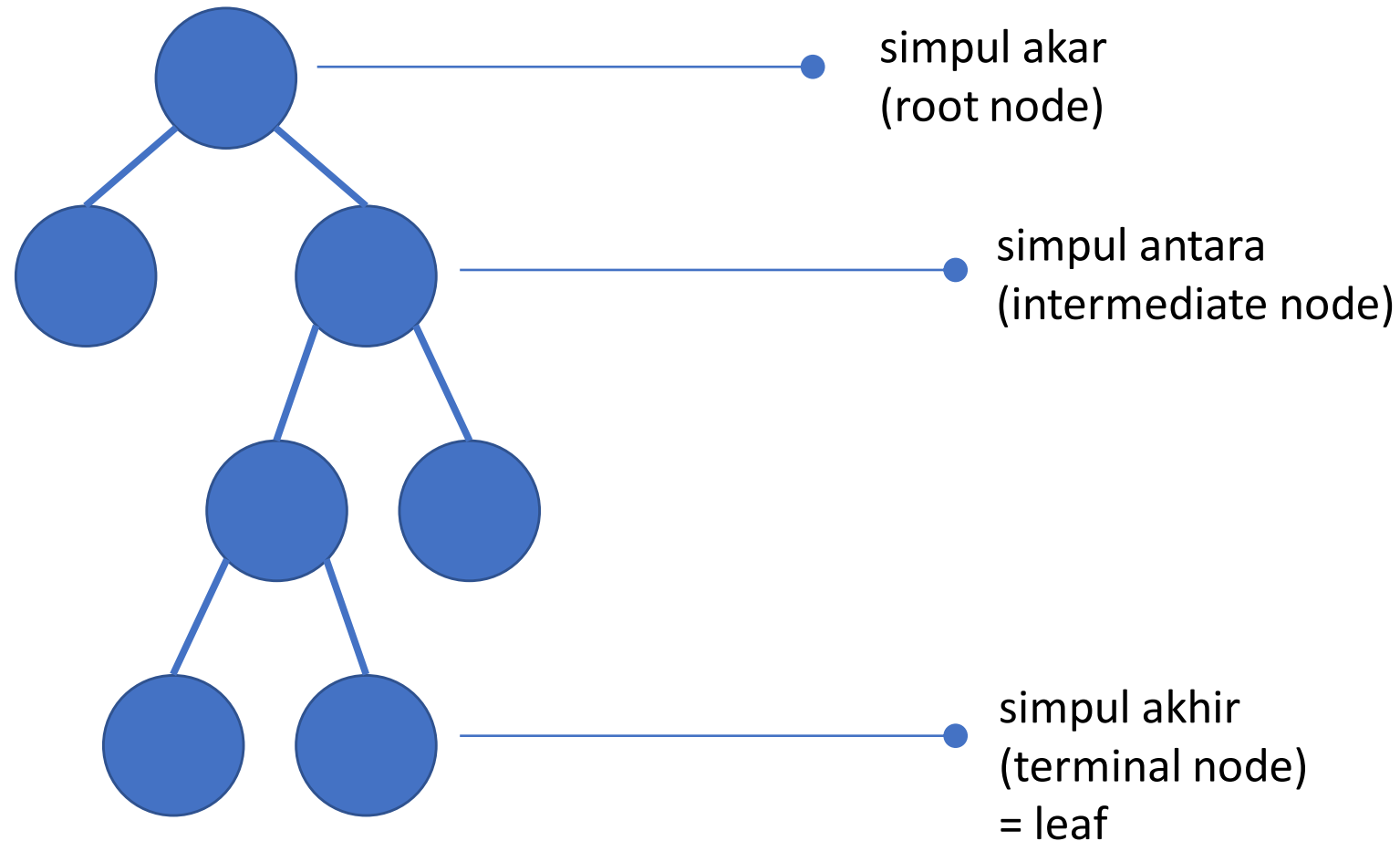
Bentuk dan Kegunaan



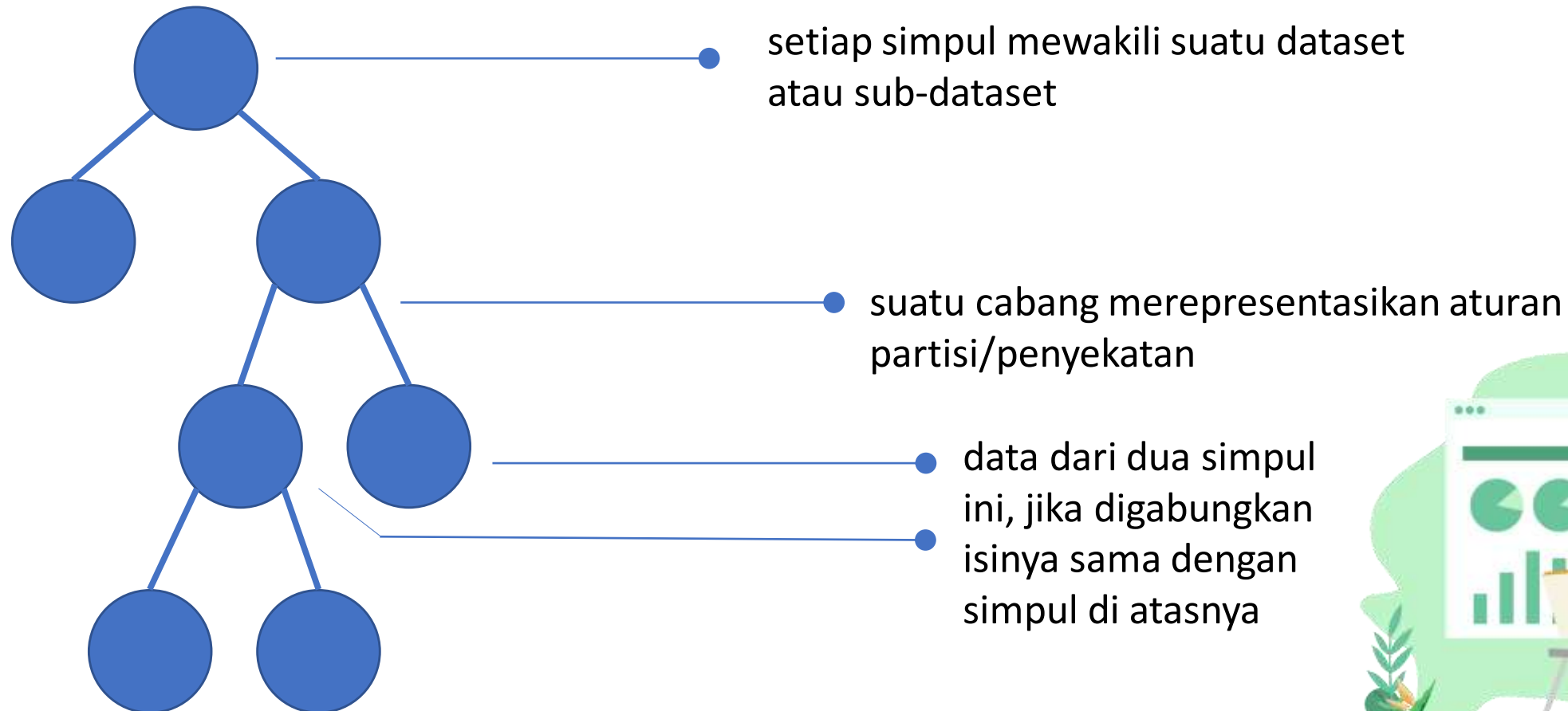
Bentuk dan Komponennya



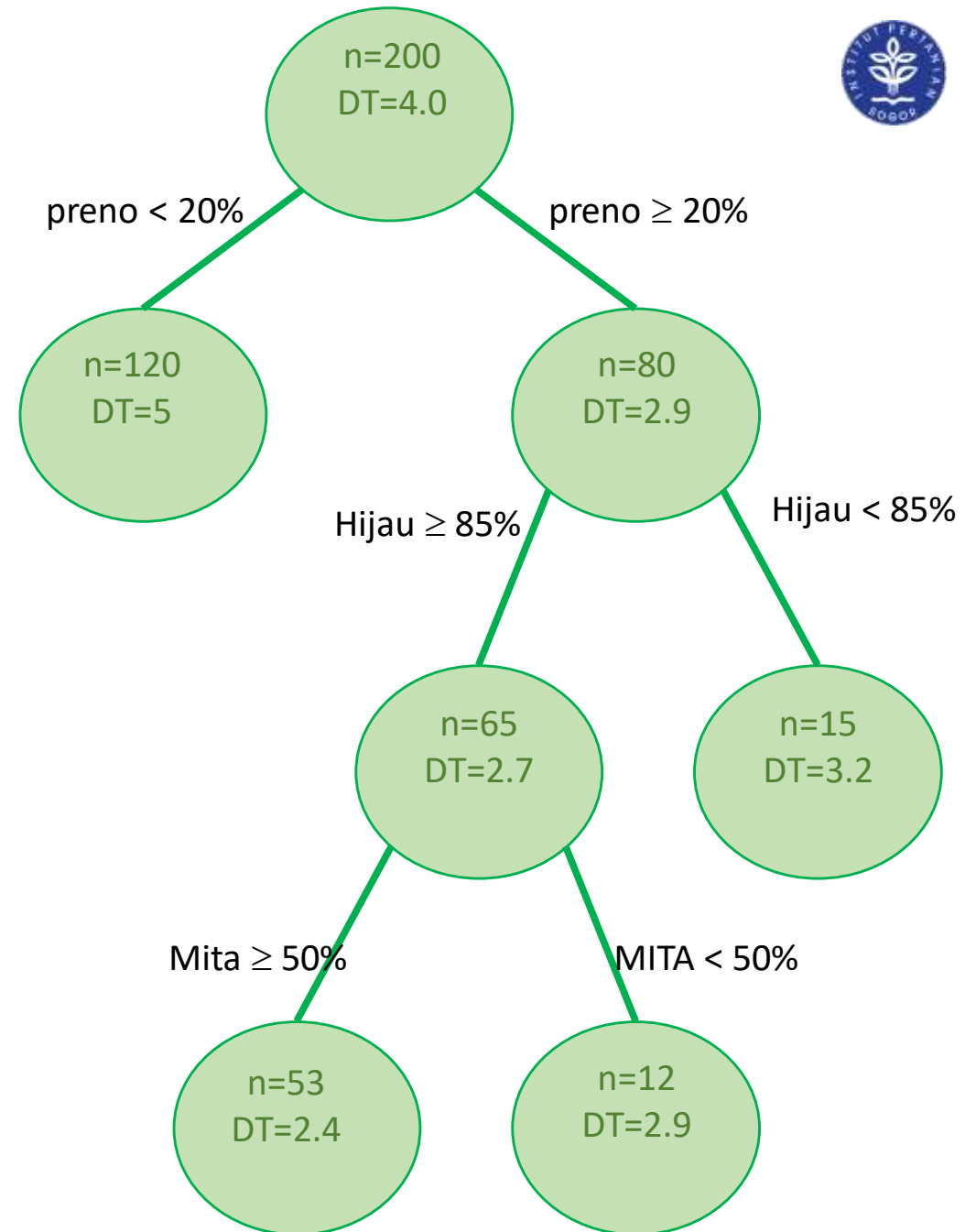
Bentuk dan Komponennya



Bentuk dan Komponennya



Apa gunanya?

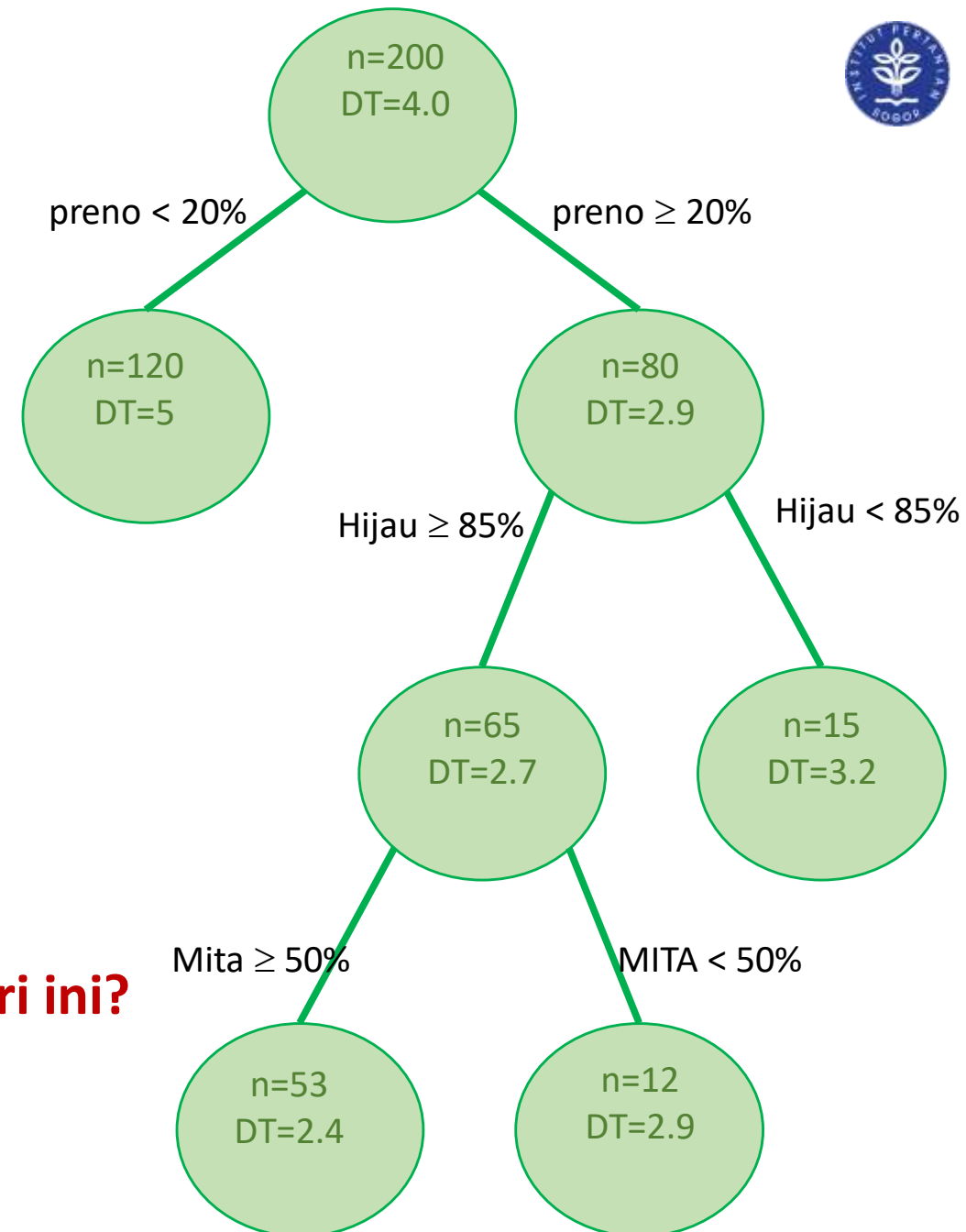


Kegunaan #1: Memprediksi Nilai dari Individu Baru



Preno = 34%
Jalur Hijau = 87%
MITA = 40%

**Berapa rata-rata
Dwelling Time hari ini?**



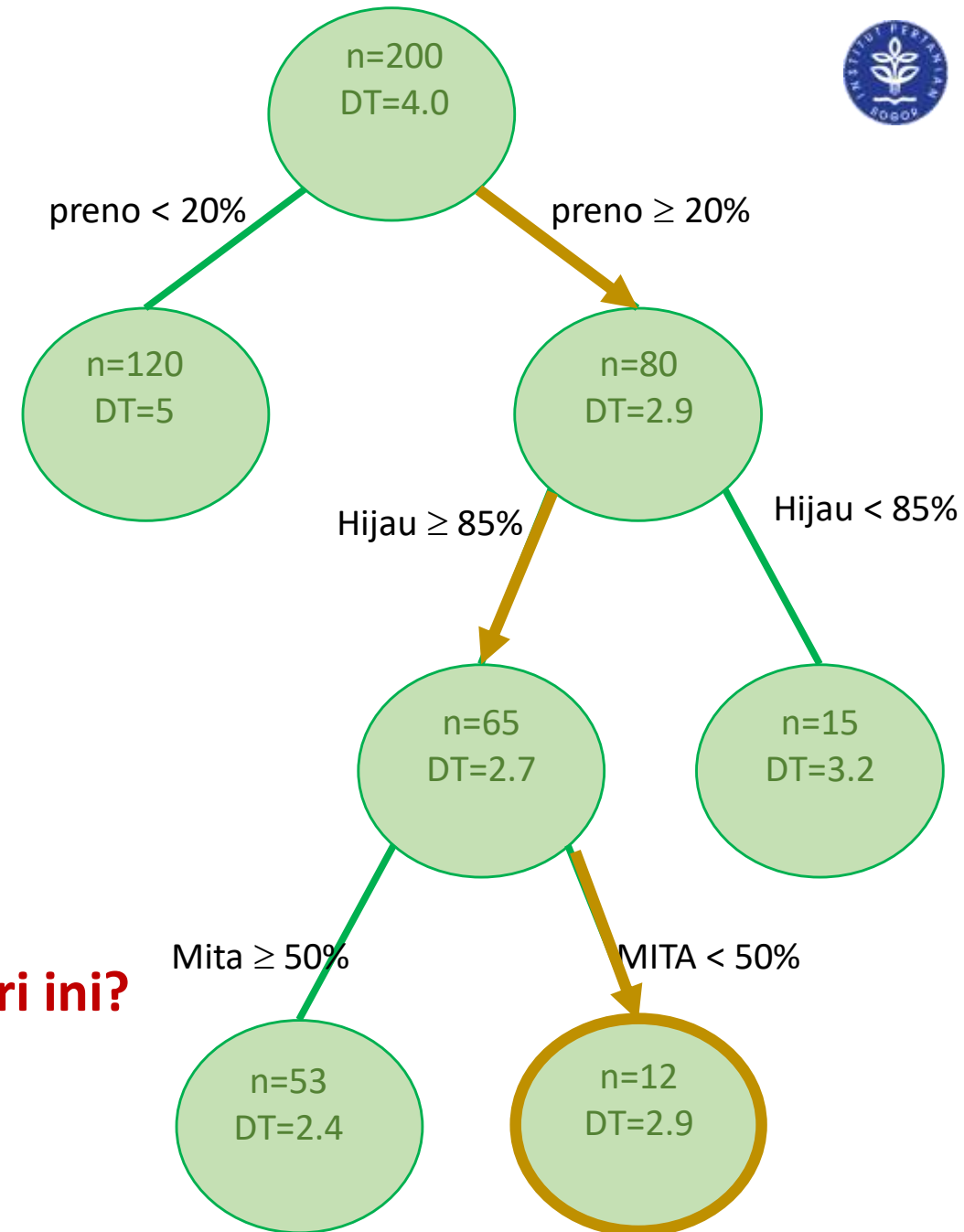
Kegunaan #1: Memprediksi Nilai dari Individu Baru



Preno = 34%
Jalur Hijau = 87%
MITA = 40%

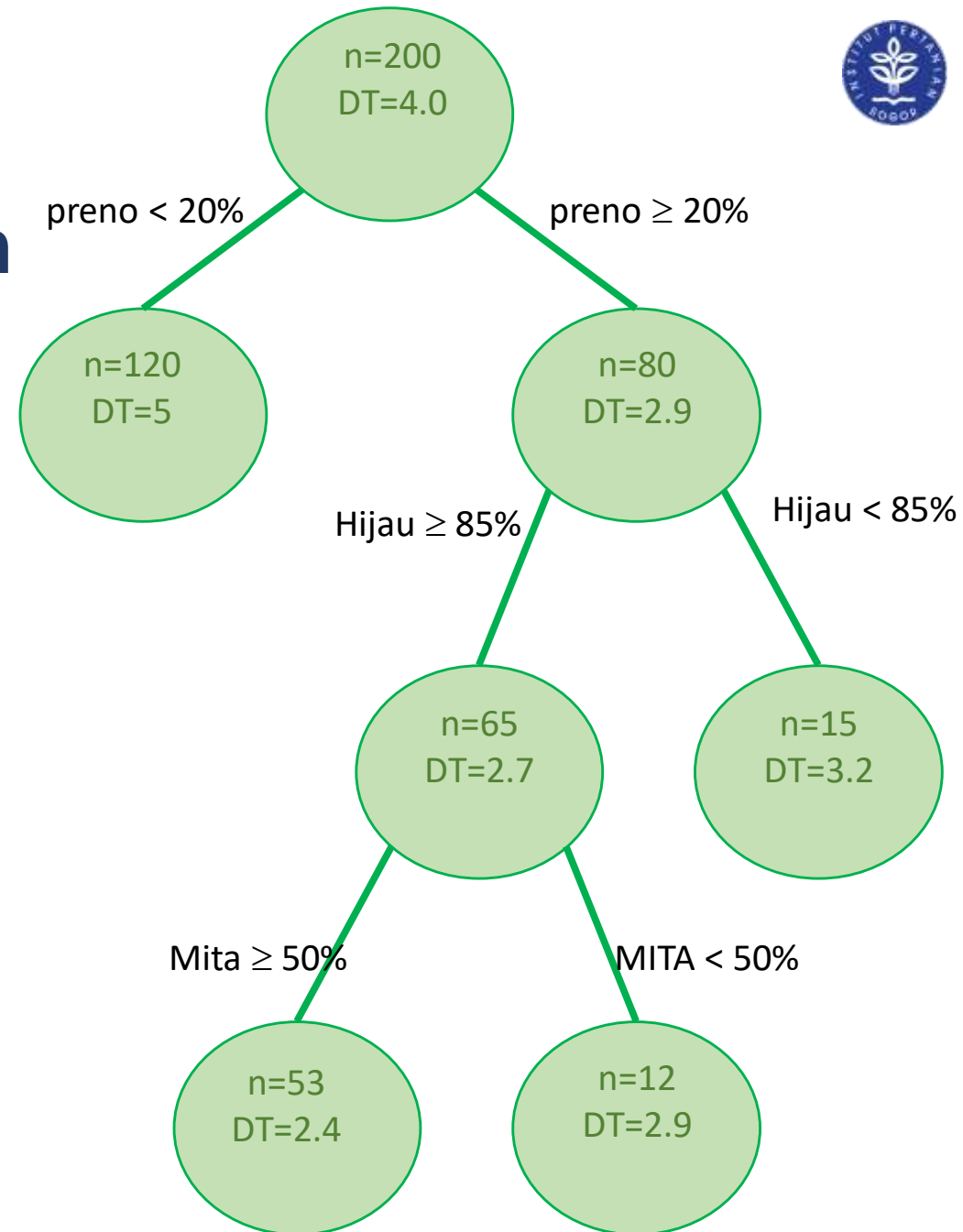
**Berapa rata-rata
Dwelling Time hari ini?**

DT = 2.9 hari



Kegunaan #2: Mengidentifikasi karakteristik dari segmen tertentu

... misal yang DT-nya singkat itu seperti apa?



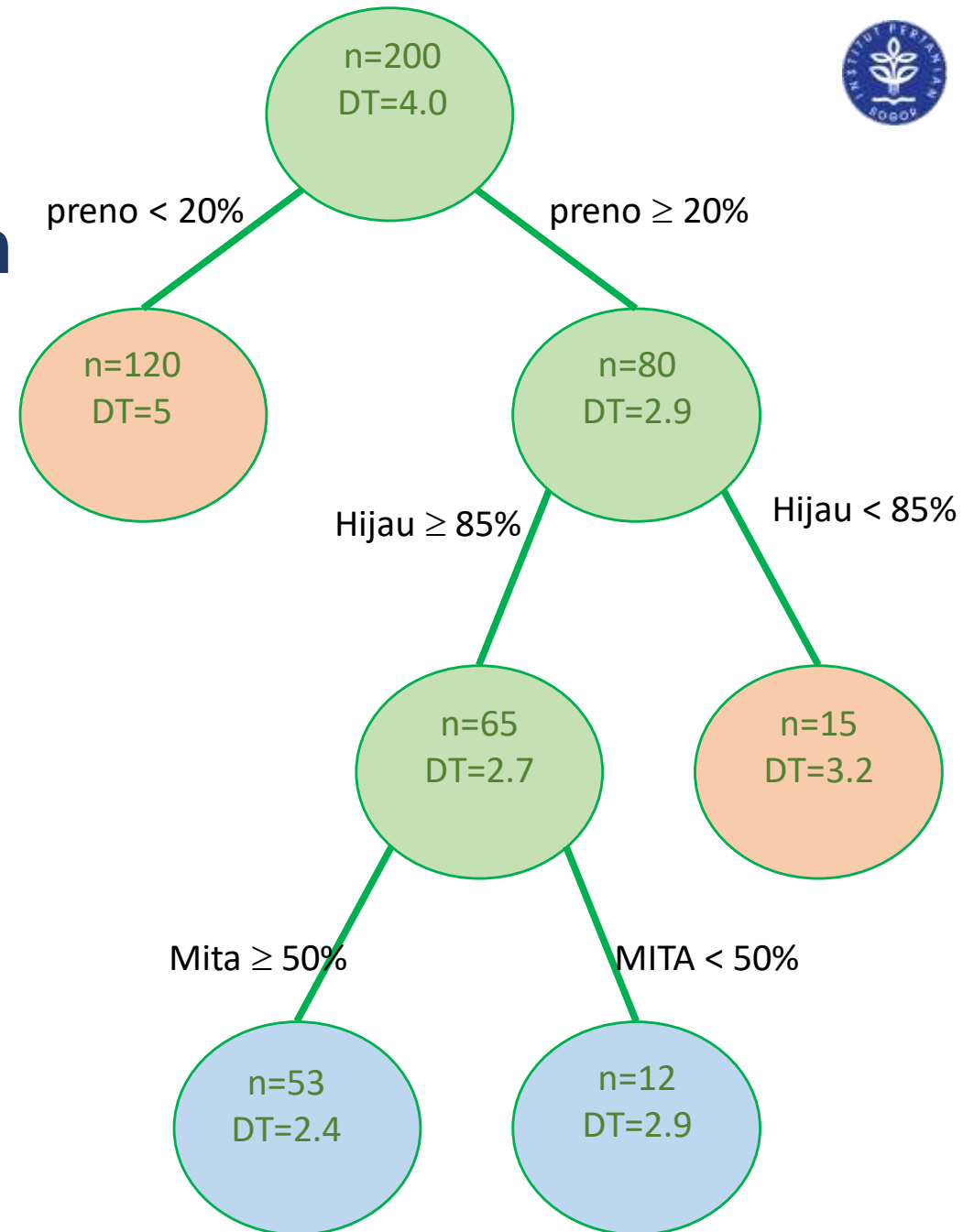
Kegunaan #2: Mengidentifikasi karakteristik dari segmen tertentu

... misal yang DT-nya singkat itu seperti apa?

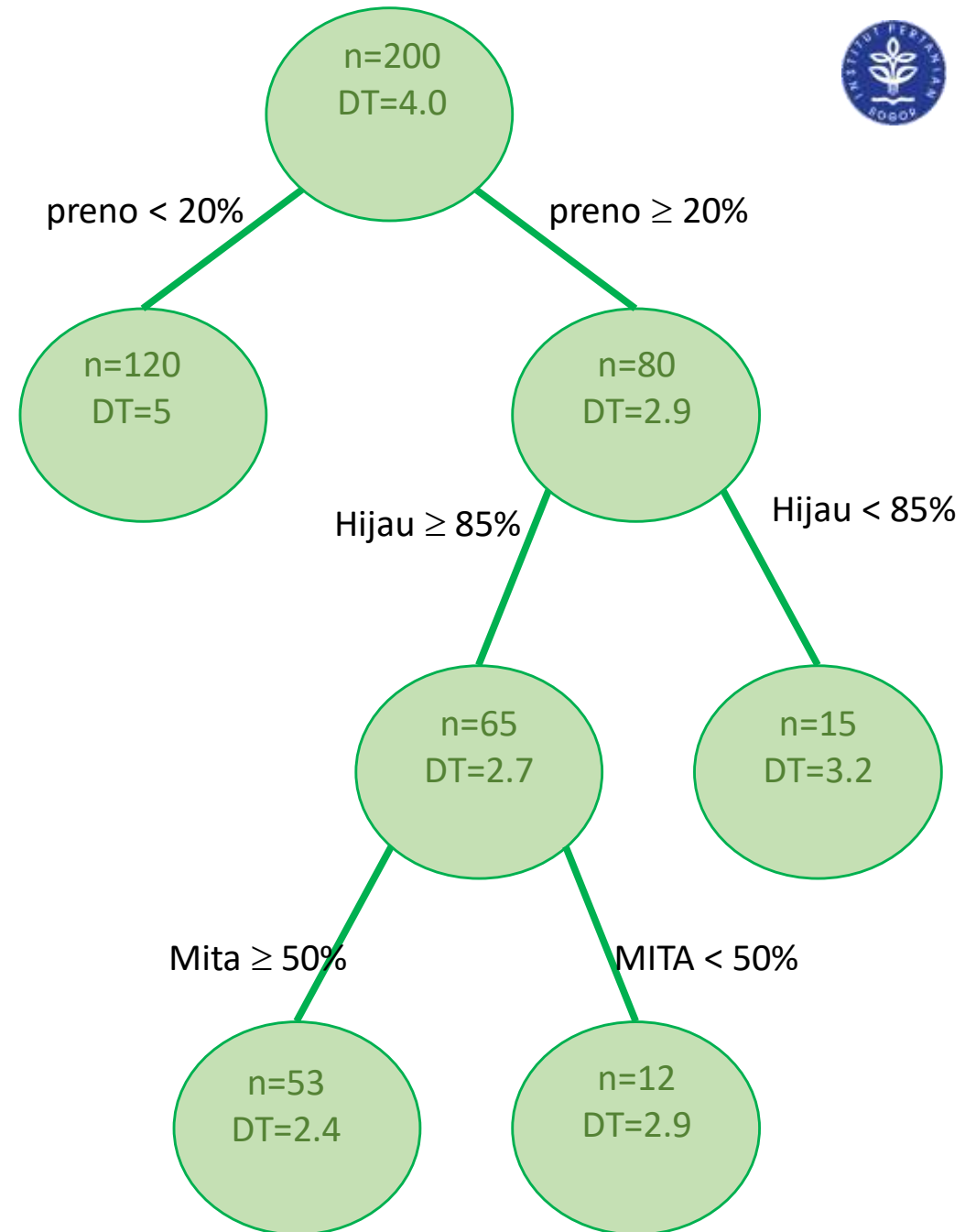
... ditandai dengan simpul biru

...
preno lebih dari 20%
jalur hijau lebih dari 85%

dan MITA lebih dari 50%



Bagaimana memperoleh pohon regresi?




Tahapan Umum Proses Partisi/Penyekatan (splitting)

1. Cari batas partisi/sekatan terbaik untuk masing-masing variabel prediktor

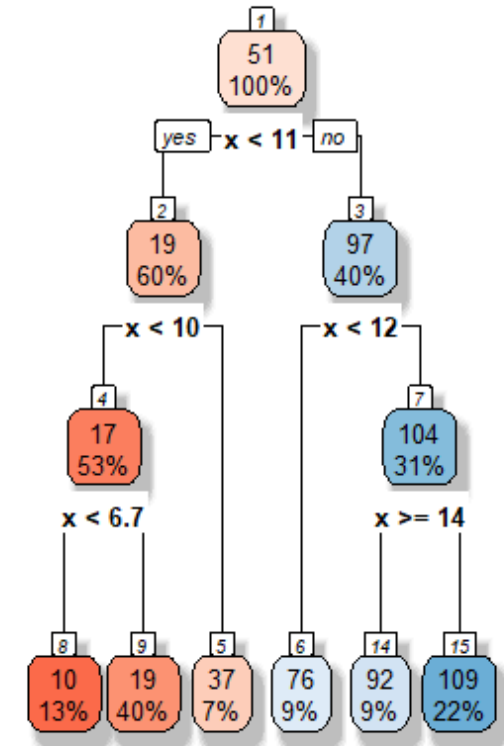
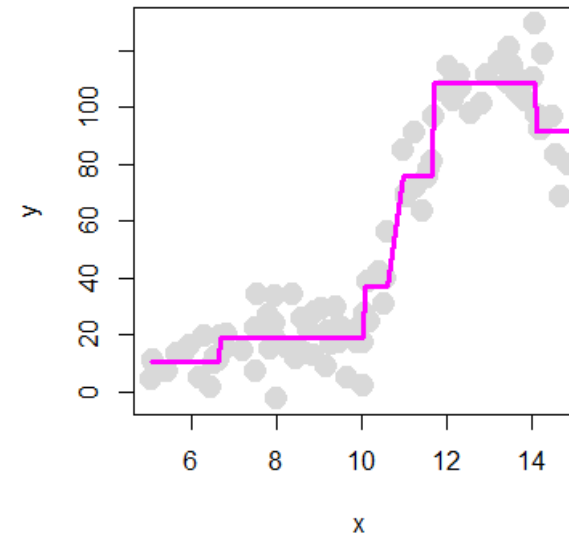
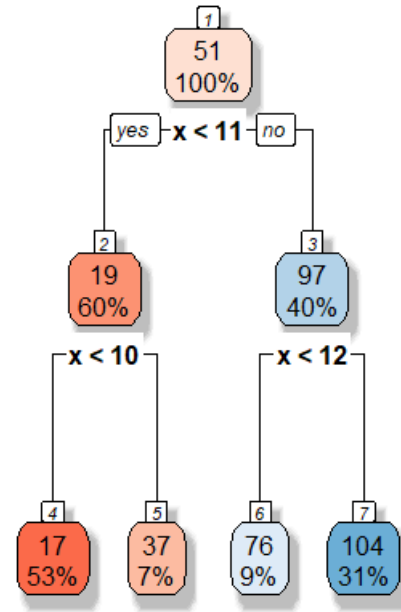
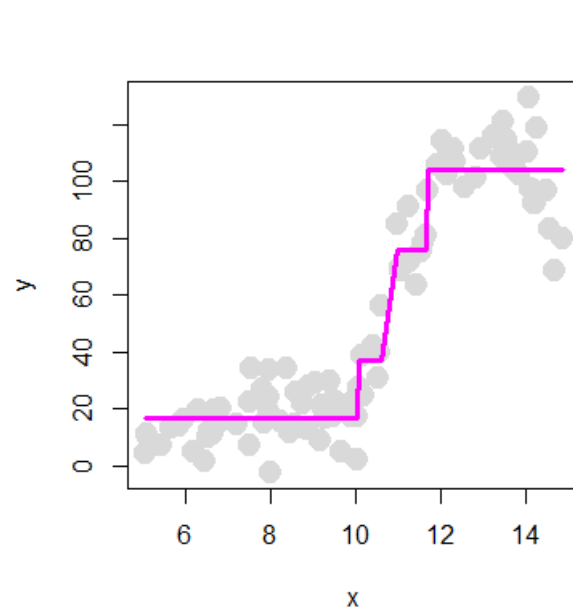
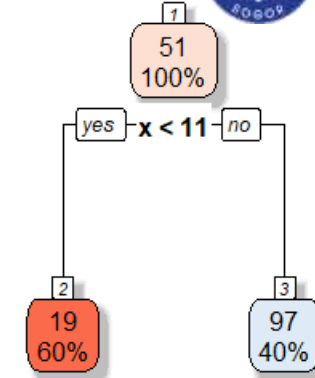
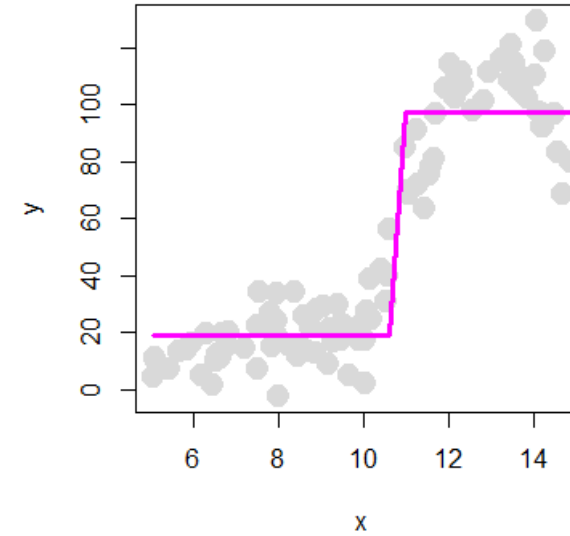
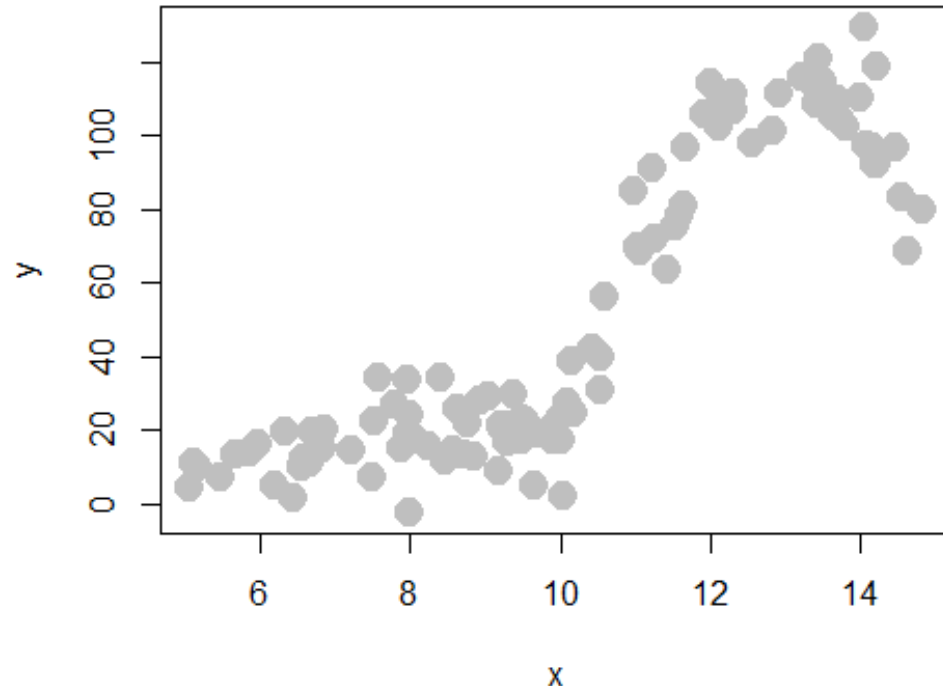
2. Bandingkan partisi terbaik dari semua variabel prediktor... pilih yang paling baik

3. Lakukan penyekatan berdasarkan variabel yang dihasilkan pada langkah ke-2

4. Lakukan 1-2-3 untuk setiap simpul, sampai tercapai kriteria penghentian algoritma



**Penurunan MSE
terbesar**



Details on tree growing

- Let j_r denote the **index of the split variable at the step r** , i.e. the coordinate X_{j_r} will be used for the r th split.
- Let t_r be the **split point to be used at the step r** .
- How to determine j_r and t_r efficiently?
- Let $R_{1,r-1}, \dots, R_{m_{r-1},r-1}$ be the sets obtained for the $r - 1$ partition (leaves at the step $r - 1$).
- Fix j so that a potential split variable is x_j .
- A potential split point t of $R_{k,r-1}$ must be one of x_j 's for those x 's that are in $R_{k,r-1}$.
- For such x_j 's let consider the split $R_1(x_j), R_2(x_j)$ of $R_{k,r-1}$.
- Choose x_j 's such that the reduction of the mean square error is the largest, i.e. choose x_j 's maximizing

$$N_{k,r-1} Q_k(T_{r-1}) - \sum_{x_i \in R_1(x_j)} (y_i - \hat{c}_{1,x_j})^2 - \sum_{x_i \in R_2(x_j)} (y_i - \hat{c}_{2,x_j})^2,$$

$N_{k,r-1}$ – the number of x 's in $R_{k,r-1}$, \hat{c}_{1,x_j} and \hat{c}_{2,x_j} – averages of the responses over the splits $R_1(x_j), R_2(x_j)$, respectively.

Aturan penghentian algoritma partisi/splitting

Algoritma splitting akan berhenti jika tercapai salah satu dari kriteria berikut:

1. Simpul hanya berisi amatan yang **sedikit**... pada fungsi `rpart()` ditentukan menggunakan opsi **minsplit** dan **minbucket**
2. Pohon sudah **terlalu besar**... pada fungsi `rpart()` ditentukan menggunakan fungsi **maxdepth**

Menilai Keباikan Hasil Prediksi

- MAPE (mean absolute percentage error)

$$MAPE = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i} \times 100\%$$

- Gunakan data lain untuk melakukan penilaian!



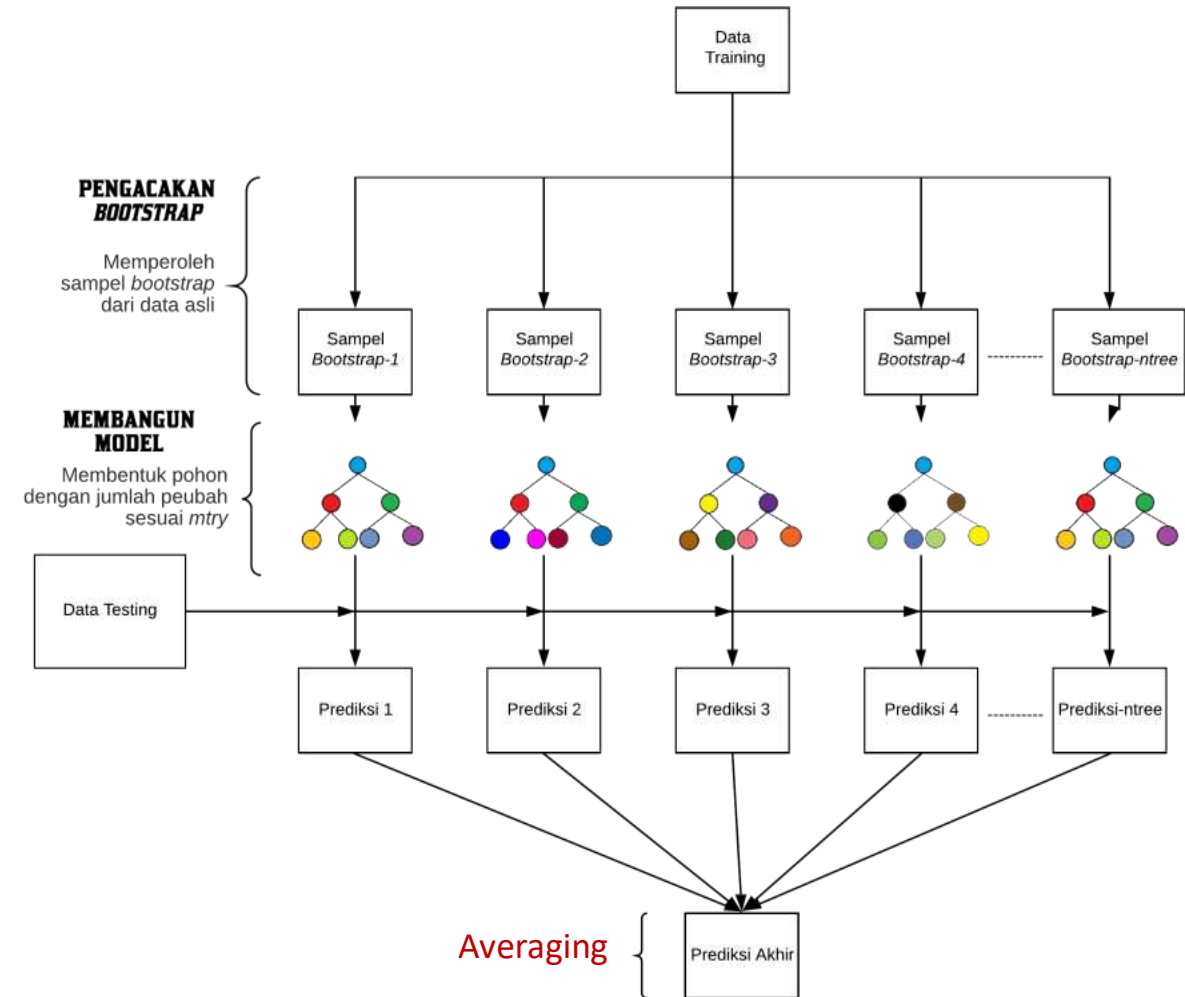
Kita lihat implementasinya di R

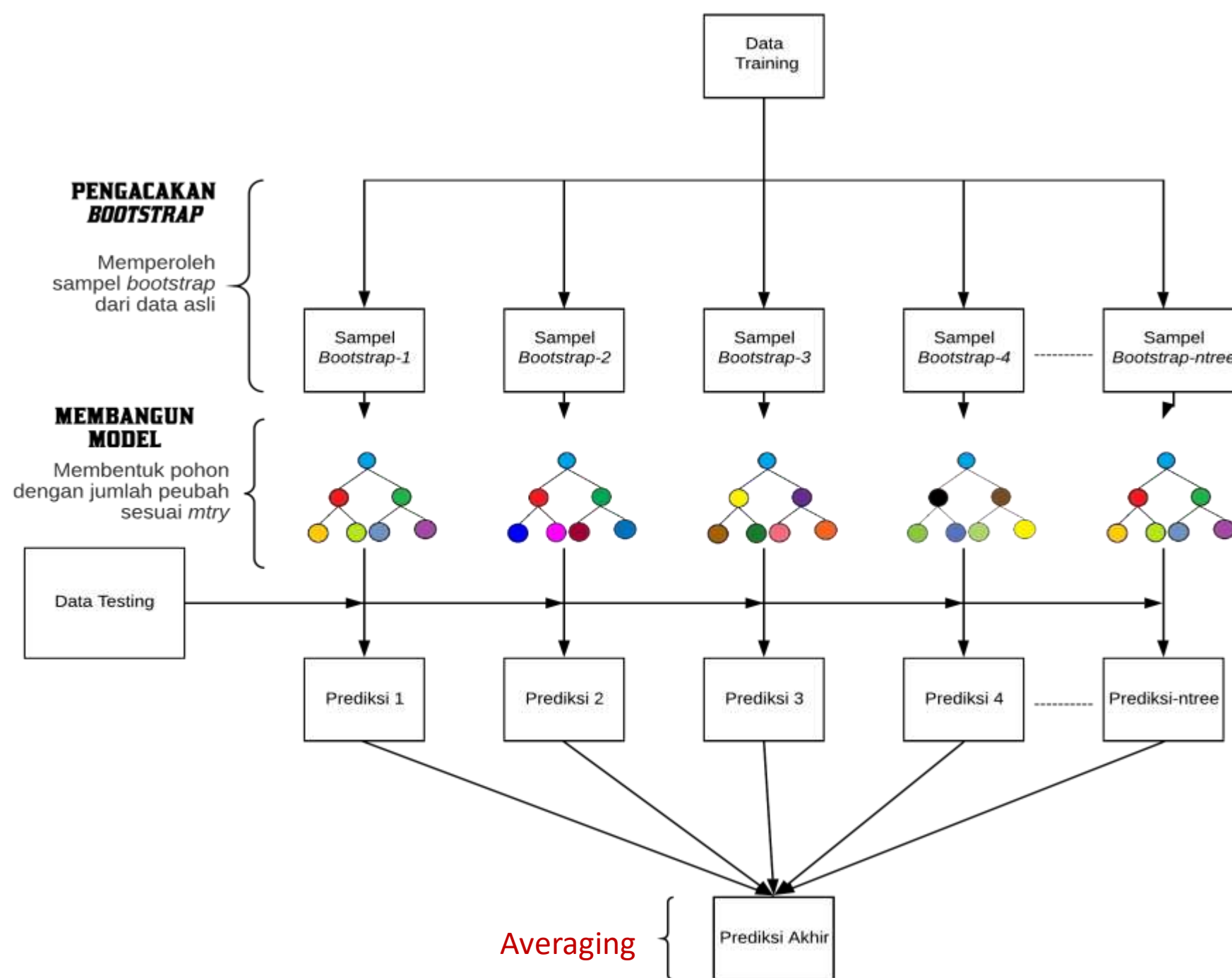
- <https://rpubs.com/bagusco/regressiontree>

Random Forest mampu
meningkatkan kualitas prediksi
pohon tunggal

Random Forest

- pengembangan dari metode *bagging* (Bootstrap and Aggregating)
- Awalnya, terdapat sebuah gugus data training.
- Lakukan resampling bootstrap
- jalankan algoritma pohon klasifikasi dan diperoleh sebuah pohon (variabel pemisah diambil yang terbaik dari sampel acak variabel)
- Proses ini diulang sebanyak k kali untuk menghasilkan k buah pohon yang berbeda
- Lakukan prediksi menggunakan k pohon dan prediksi akhir diperoleh dengan cara rata-rata (averaging)





Algoritma

For $b = 1$ to B :

- (a) Draw a bootstrap sample Z^* of size N from the training data.
- (b) Grow a random-forest tree to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.
 - i. Select m variables at random from the p variables.
 - ii. Pick the best variable/split-point among the m .
 - iii. Split the node into two daughter nodes.

Output the ensemble of trees.

To make a prediction at a new point x we do:

For regression: average the results

For classification: majority vote



Kita lihat implementasinya di R

- <https://rpubs.com/bagusco/regressiontree>

Terima Kasih





IPB University
— Bogor Indonesia —

Inspiring Innovation with Integrity
in Agriculture, Ocean and Biosciences for a Sustainable World