Pohon Regresi

Pengantar Data Sains 13

Supervised Learning

Regresi

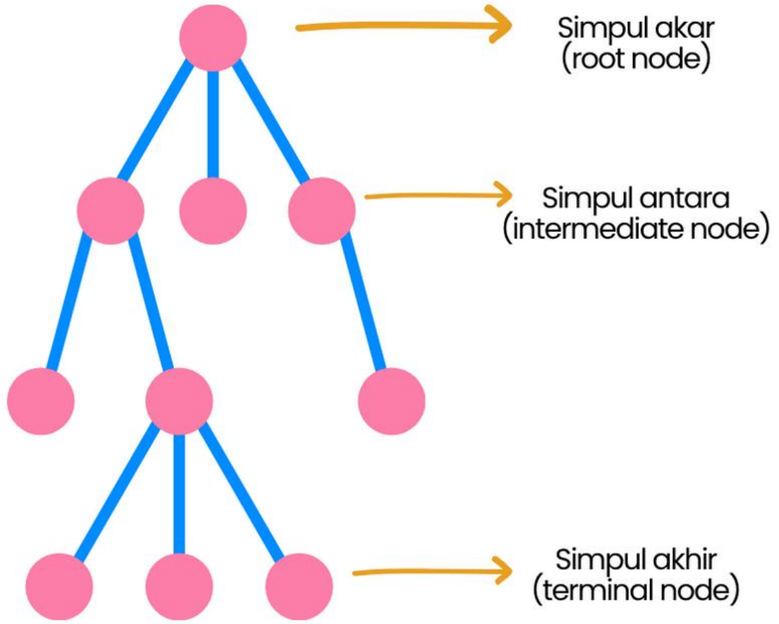
Variabel respon numerik
contoh:
- Simple linier regression
- Multiple regression
- SVM regression
- **Decision tree regression**
- Random forest regression

Klasifikasi

Variabel respon kategorik
contoh:
- Logistik regression
- Classification tree
- SVM
- Random forest
- Neural network
- Boosting

Pohon Regresi

Model bekerja dengan cara **membagi data ke dalam kelompok kelompok lebih kecil** berdasarkan nilai fitur (variabel input), hingga setiap kelompok menghasilkan **prediksi berupa rata-rata nilai target**.



Impurity (penentuan split): Variance/MSE, dan MAE

Contoh:

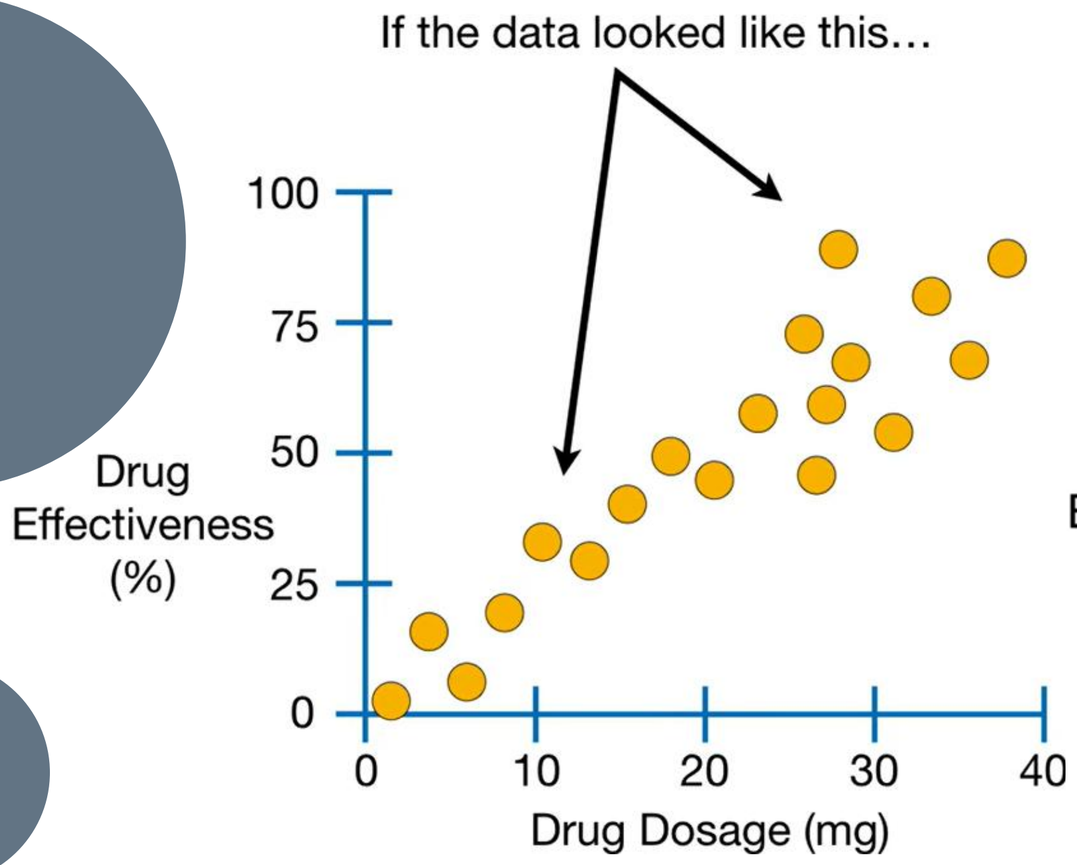
- Prediksi harga rumah
- Prediksi efektivitas obat
- Prediksi hasil panen

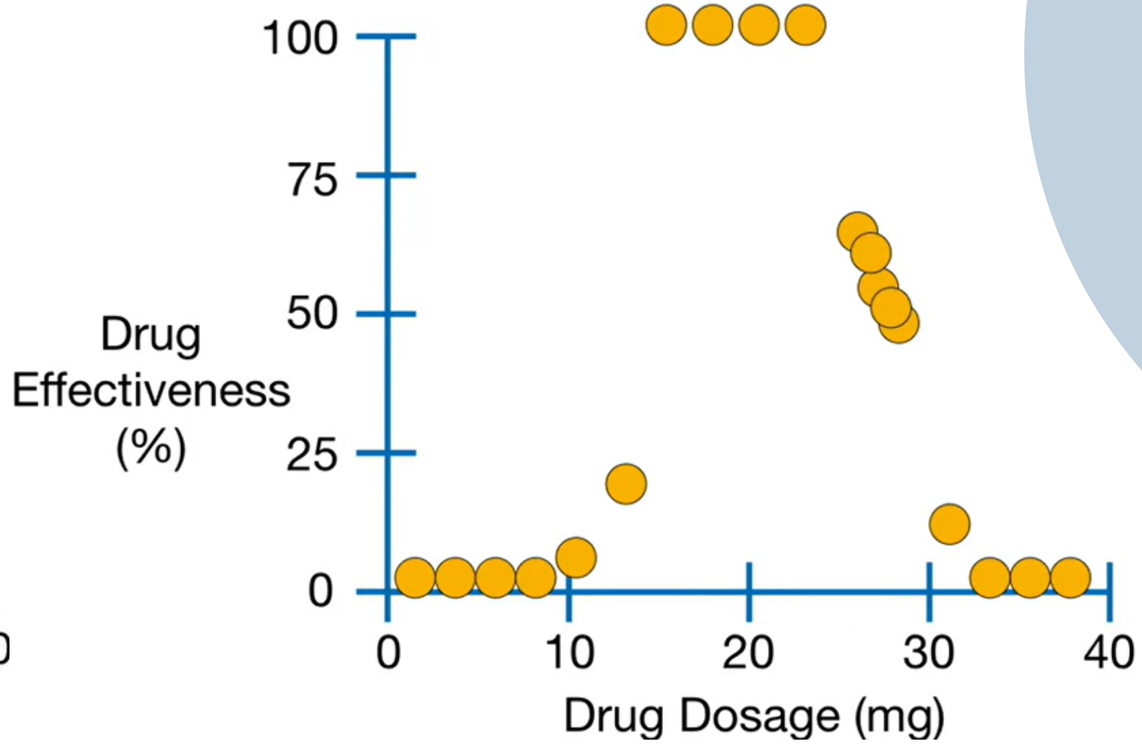
Pohon Regresi

Bentuk dan komponennya

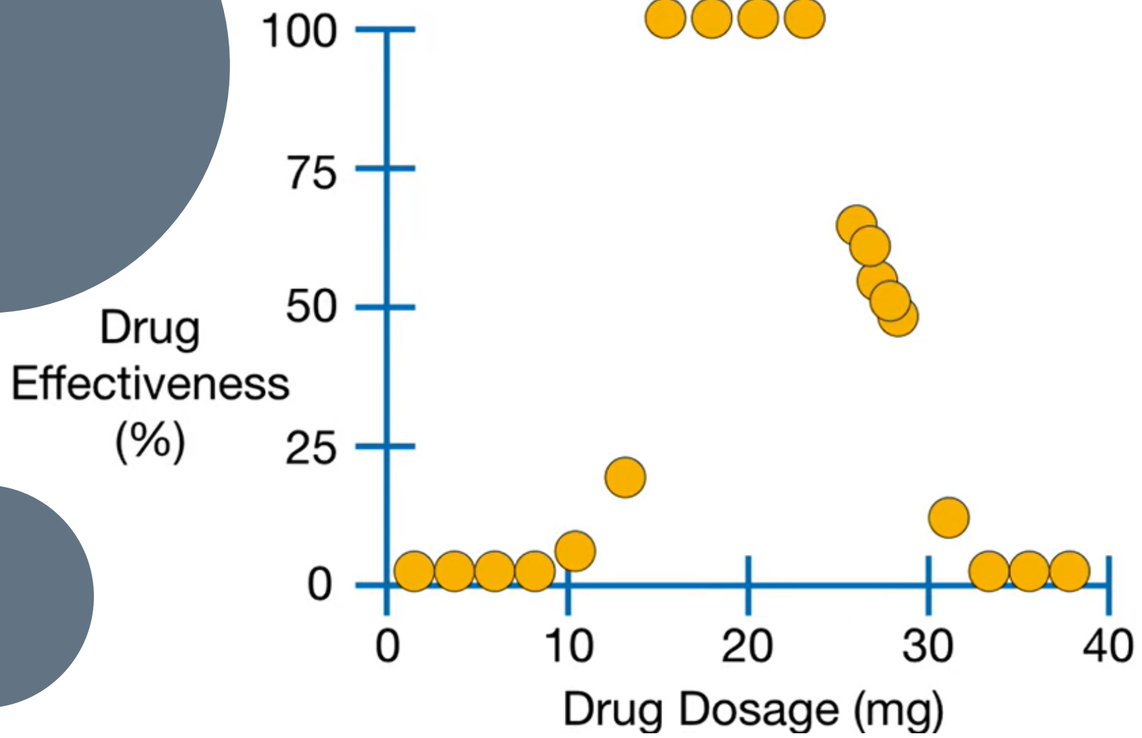
- setiap simpul mewakili suatu dataset atau sub-dataset
- suatu cabang merepresentasikan aturan partisi/penyekatan
- Data dari dua simpul jika digabungkan isinya sama dengan simpul diatasnya

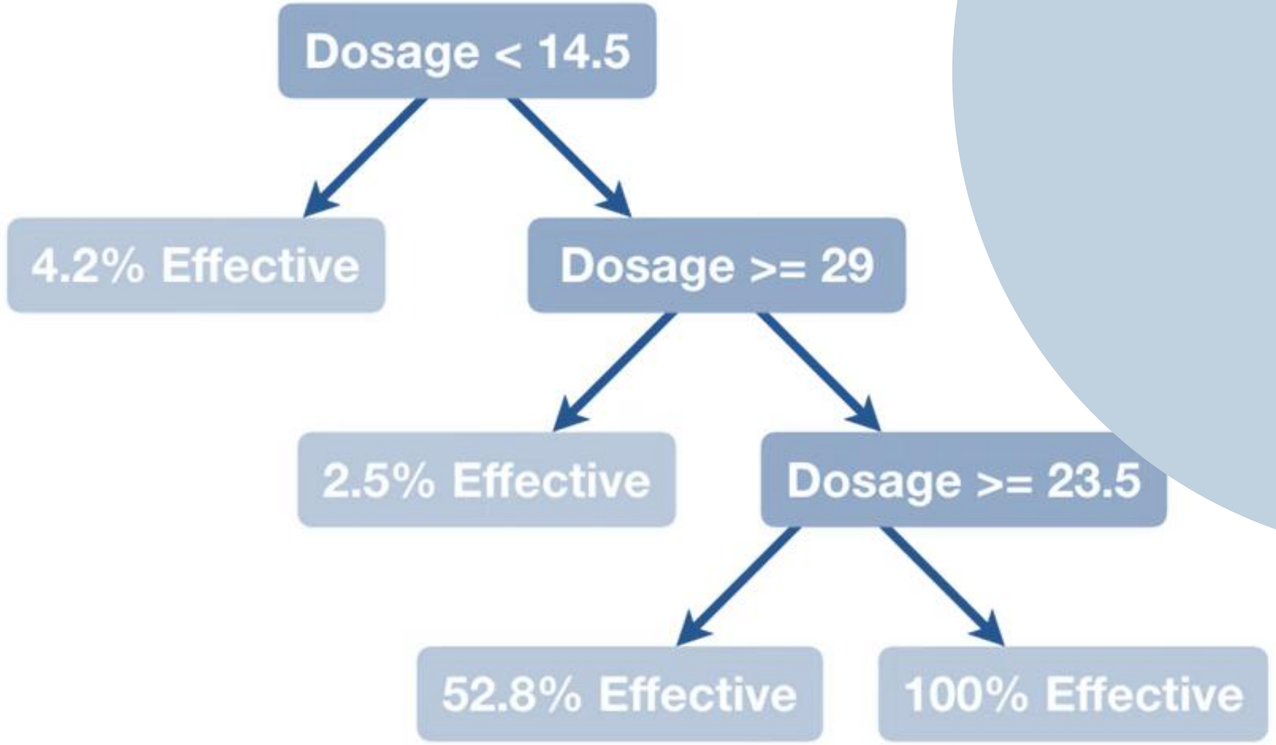
Regression


If the data looked like this…
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)

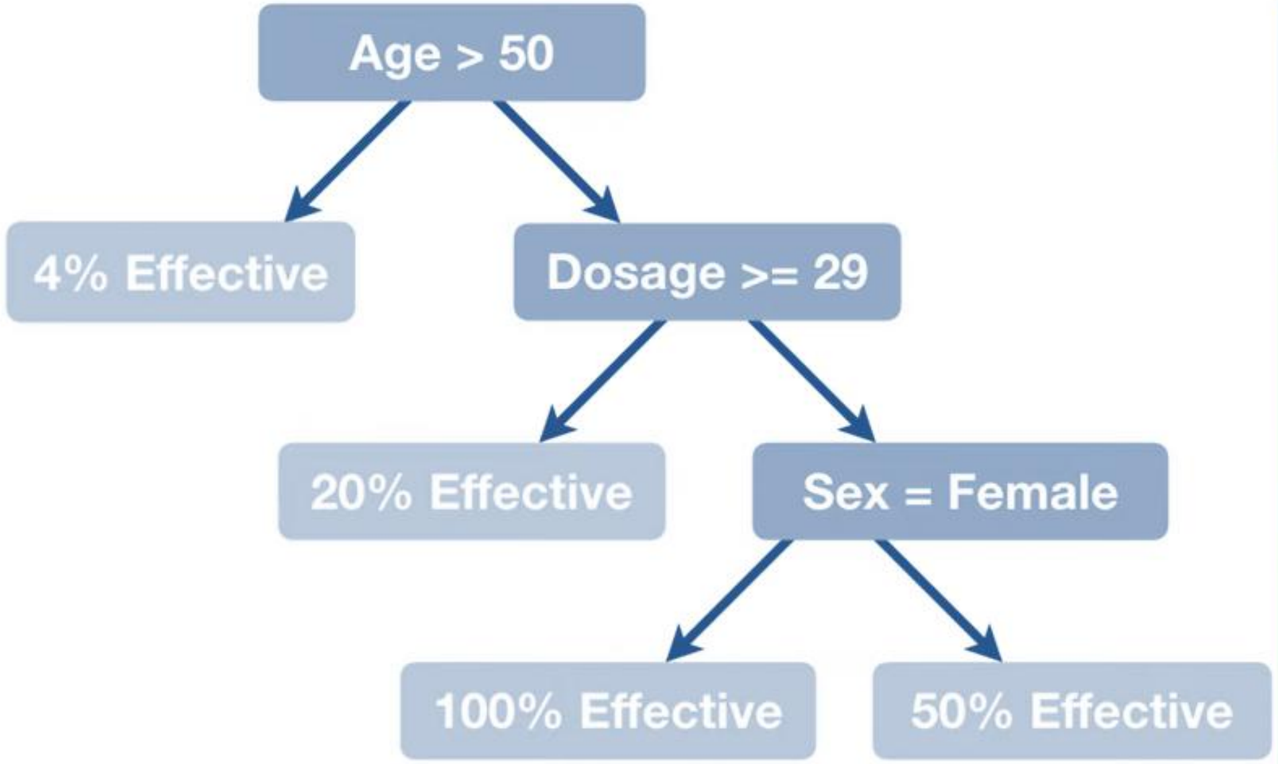

Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)

Pohon Regresi 1 Prediktor





Pohon Regresi >1 Prediktor



Dosage	Age	Sex	Etc.	Drug Effect.
10	25	Female	…	98
20	73	Male	…	0
35	54	Female	…	100
5	12	Male	…	44
etc…	etc…	etc…	etc…	etc…

Tahapan Proses Partisi/Penyekatan

Using MSE

1. cari batas **partisi terbaik** untuk setiap variabel prediktor
2. bandingkan partisi terbaik dari semua variabel prediktor, **pilih yang paling baik**
3. **lakukan penyekatakan** berdasarkan variabel yang dihasilkan pada langkah ke-2
4. Lakukan proses 1,2,3 untuk setiap simpul, sampai tercapai kriteria penghentian algoritma

Algoritma berhenti jika

- ✓ semua amatan pada simpul hanya terdiri atas **satu kelas variabel respon** saja
- ✓ semua amatan pada simpul memiliki **nilai variabel prediktor yang sama**
- ✓ simpul hanya berisi **amatan yang sedikit** -> minsplit
- ✓ pohon sudah **terlalu besar** -> maxdepth

Alur Kerja Pohon Regresi

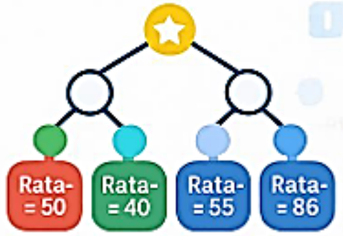
1 **Model melihat seluruh data**

2 **Model mencoba memecah data** berdasarkan fitur tetentu (misalnya usia, luas tanah, pendapatan)

3 **Menghitung error** (biasanya MSE) untuk setiap kemungkinan pemisahan

	$MSE_1 = 100$	60 ✓
	$MSE_2 = 60$	80 ✓
	$MSE_2 = 20$	50 ✓

4 **Memilih split terbaik,** yaitu error yang paling kecil



5 **Memilih split terbaik,** yaitu error yang paling kecil

Proses diulang hingga tidak bisa dibagi lagi atau mencapai batas kedalaman pohon.

6 **Setiap "daun" pada** pohon berisi prediksi nilai, yaitu rata-rata target dari datam kelompok tersebut

MSE dan Standar Deviasi

$$MSE = \frac{1}{n}\sum(y_i - \bar{y})^2$$

Karena Variance = MSE, maka

$$\boldsymbol{SD = \sqrt{Variance} = \sqrt{MSE}}$$

Menurunkan SD setara dengan menurunkan MSE karena akar tidak mengubah urutan sehingga split terbaik tetap sama

$$RSD = SD_{global} - wieghted\ SD$$
$$RSD = SD_{global} - \left(\frac{n_l}{n}SD_L + \frac{n_r}{n}SD_R\right)$$

$$red\ MSE = MSE_{global} - weighted\ MSE$$
$$red\ MSE = MSE_{global} - \left(\frac{n_l}{n}MSE_L + \frac{n_r}{n}MSE_R\right)$$

Reduction Standard Deviation (RSD) setara dengan Reduction MSE

Proses pembentukan Pohon Regresi

Day	Outlook	Temp	Humidity	Wind	Golf Players
D1	Sunny	Hot	High	Weak	25
D2	Sunny	Hot	High	Strong	30
D3	Overcast	Hot	High	Weak	46
D4	Rain	Mild	High	Weak	45
D5	Rain	Cool	Normal	Weak	52
D6	Rain	Cool	Normal	Strong	23
D7	Overcast	Cool	Normal	Strong	43
D8	Sunny	Mild	High	Weak	35
D9	Sunny	Cool	Normal	Weak	38
D10	Rain	Mild	Normal	Weak	46
D11	Sunny	Mild	Normal	Strong	48
D12	Overcast	Mild	High	Strong	52
D13	Overcast	Hot	Normal	Weak	44
D14	Rain	Mild	High	Strong	30

1. Hitung standar deviasi global (respon)
   SD golf player = 9.32
2. Hitung standar deviasi golf player untuk setiap variabel prediktor
   Outlook → {overcast, rain , sunny}

$$SD_{overcast} = 3.49$$
$$SD_{rain} = 10.87$$
$$SD_{sunny} = 7.78$$

3. Hitung RSD

Outlook	Stdev of Golf Players	Instances
Overcast	3.49	4
Rain	10.87	5
Sunny	7.78	5

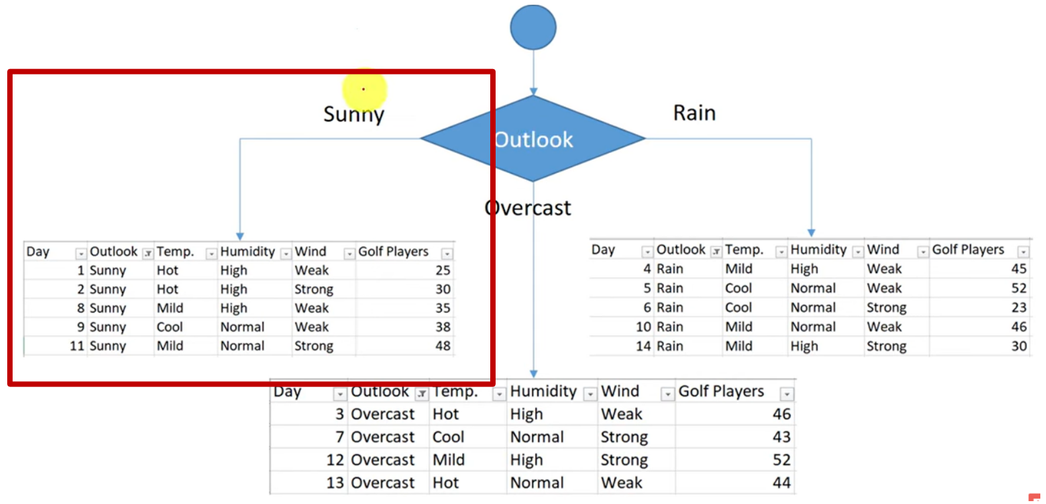
- Weighted SD outlook = $\left(\frac{4}{14}\right) 3.49 + \left(\frac{5}{14}\right) 10.87 + \left(\frac{5}{14}\right) 7.78 = 7.66$
- RSD outlook = 9.32 – 7.66 = 1.66

Proses pembentukan Pohon Regresi

4. Hitung Semua RSD variabel prediktor

Prediktor	RSD
Outlook	1.66
Temp	0.47
Humidity	0.27
Wind	0.29

→ Simpul Akar



Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
11	Sunny	Mild	Normal	Strong	48

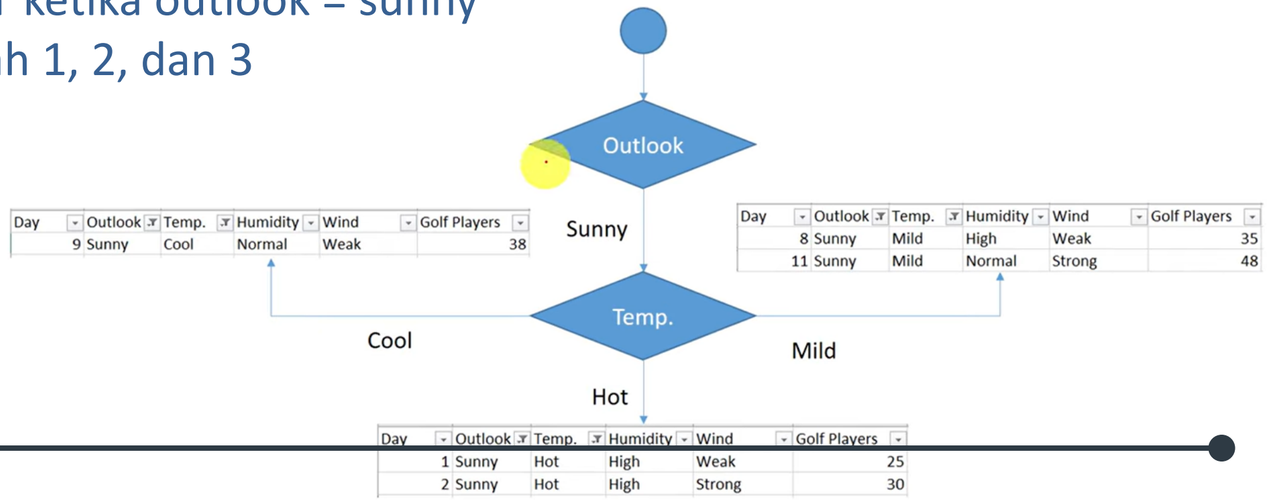
Day	Outlook	Temp.	Humidity	Wind	Golf Players
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
10	Rain	Mild	Normal	Weak	46
14	Rain	Mild	High	Strong	30

Day	Outlook	Temp.	Humidity	Wind	Golf Players
3	Overcast	Hot	High	Weak	46
7	Overcast	Cool	Normal	Strong	43
12	Overcast	Mild	High	Strong	52
13	Overcast	Hot	Normal	Weak	44

5. Hitung RSD untuk setiap variabel prediktor ketika outlook = sunny menggunakan cara yang sama pada langkah 1, 2, dan 3

Prediktor	RSD
Temp	4.18
Humidity	3.33
Wind	0.85

→ Next Simpul



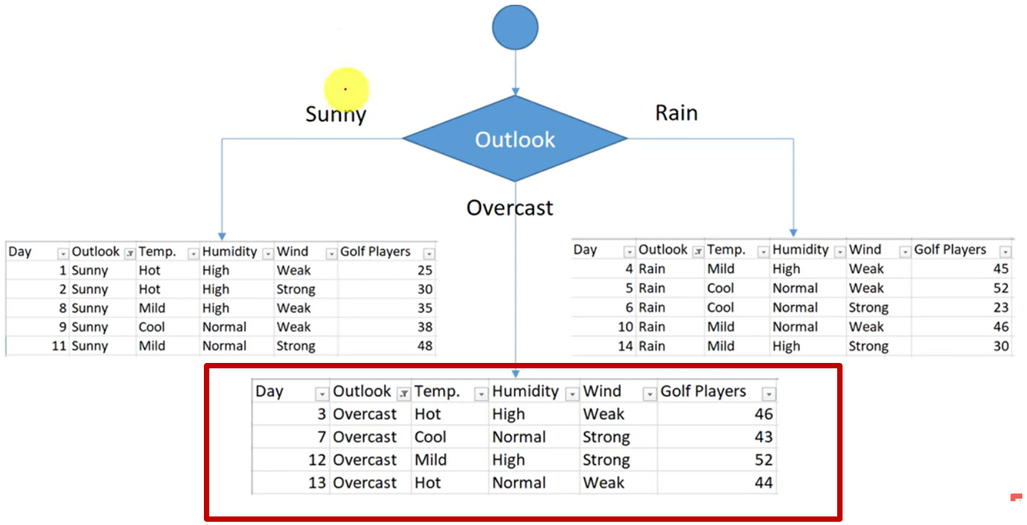
Day	Outlook	Temp.	Humidity	Wind	Golf Players
9	Sunny	Cool	Normal	Weak	38

Day	Outlook	Temp.	Humidity	Wind	Golf Players
8	Sunny	Mild	High	Weak	35
11	Sunny	Mild	Normal	Strong	48

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30

Proses pembentukan Pohon Regresi

6. Ulangi langkah 5, untuk menghitung SD ketika outlook = overcast dan rain

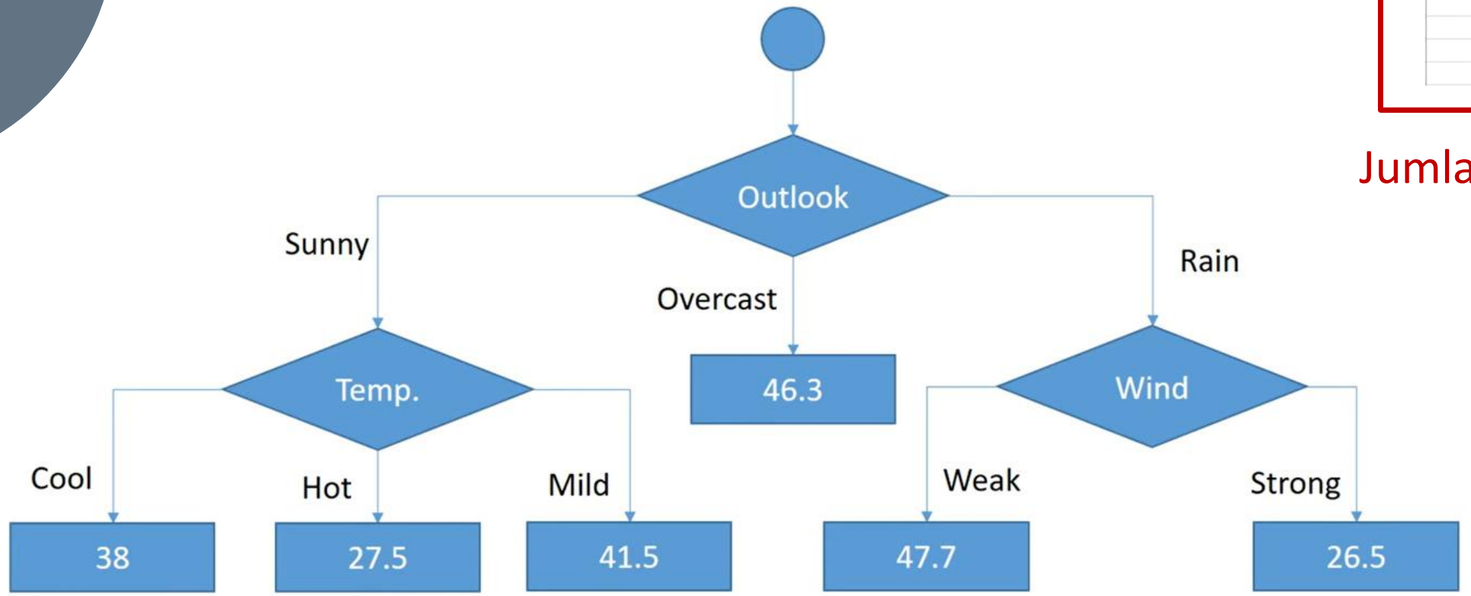


Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
11	Sunny	Mild	Normal	Strong	48

Day	Outlook	Temp.	Humidity	Wind	Golf Players
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
10	Rain	Mild	Normal	Weak	46
14	Rain	Mild	High	Strong	30

Day	Outlook	Temp.	Humidity	Wind	Golf Players
3	Overcast	Hot	High	Weak	46
7	Overcast	Cool	Normal	Strong	43
12	Overcast	Mild	High	Strong	52
13	Overcast	Hot	Normal	Weak	44

Jumlah data < 5 → pemangkasan



7. Hitung rata-rata dari golf player untuk setiap simpul akhir

Evaluasi Model

1. Mean Squared Error (MSE)
2. Root Mean Squared Error (RMSE)
3. Mean Absolute Error (MAE)
4. Mean Absolute Percentage Error (MAPE)
5. R-squared

Python

```
from sklearn.tree import DecisionTreeRegressor
model = DecisionTreeRegressor(max_depth=3)
model.fit(X_train, y_train)
pred = model.predict(X_test)
```

Hyperparameter Pohon Regresi

- max_depth: kedalaman pohon
- min_samples_split: minimum sampel untuk split
- min_samples_leaf: minimum sampel di leaf
- max_features: fitur yang dipertimbangkan saat
- split

TUGAS

Lakukan pemodelan pohon regresi menggunakan data asuransi berikut

- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
- children: Number of children covered by health insurance / Number of dependents
- smoker: Smoking
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance

Thank You