Klasifikasi (Decision Tree Algorithm & Naïve Bayes)

Pertemuan 7

Learning Objective

- Mahasiswa mampu menjelaskan konsep klasifikasi
- Mahasiswa mampu menjelaskan algoritma Decision Tree (ID3)
- Mahasiswa mampu menjelaskan algoritma Decision Tree (C4.5)
- Mahasiswa mampu menjelaskan algoritma Decision Tree (CART)
- Mahasiswa mampu menjelaskan konsep klasifikasi dengan Naïve Bayes
- Mahasiswa mampu menyelesaikan kasus klasifikasi dengan Naïve Bayes

Course Materials

Decision Tree (ID3)	Decision Tree (C4.5)
Decision Tree (CART)	Naïve Bayes

Konsep Decision Tree

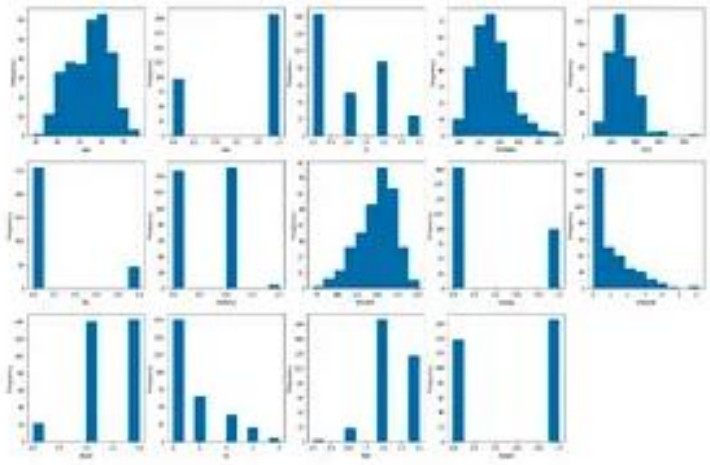
Membahas pengantar algoritma decision tree

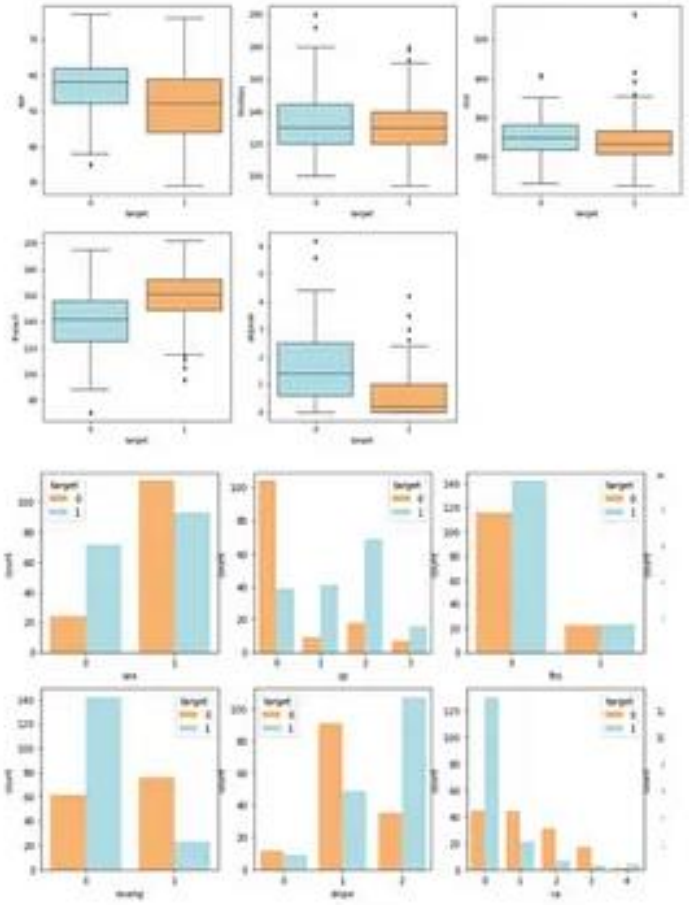
Classification Task

Eksplorasi & Visualisasi Data

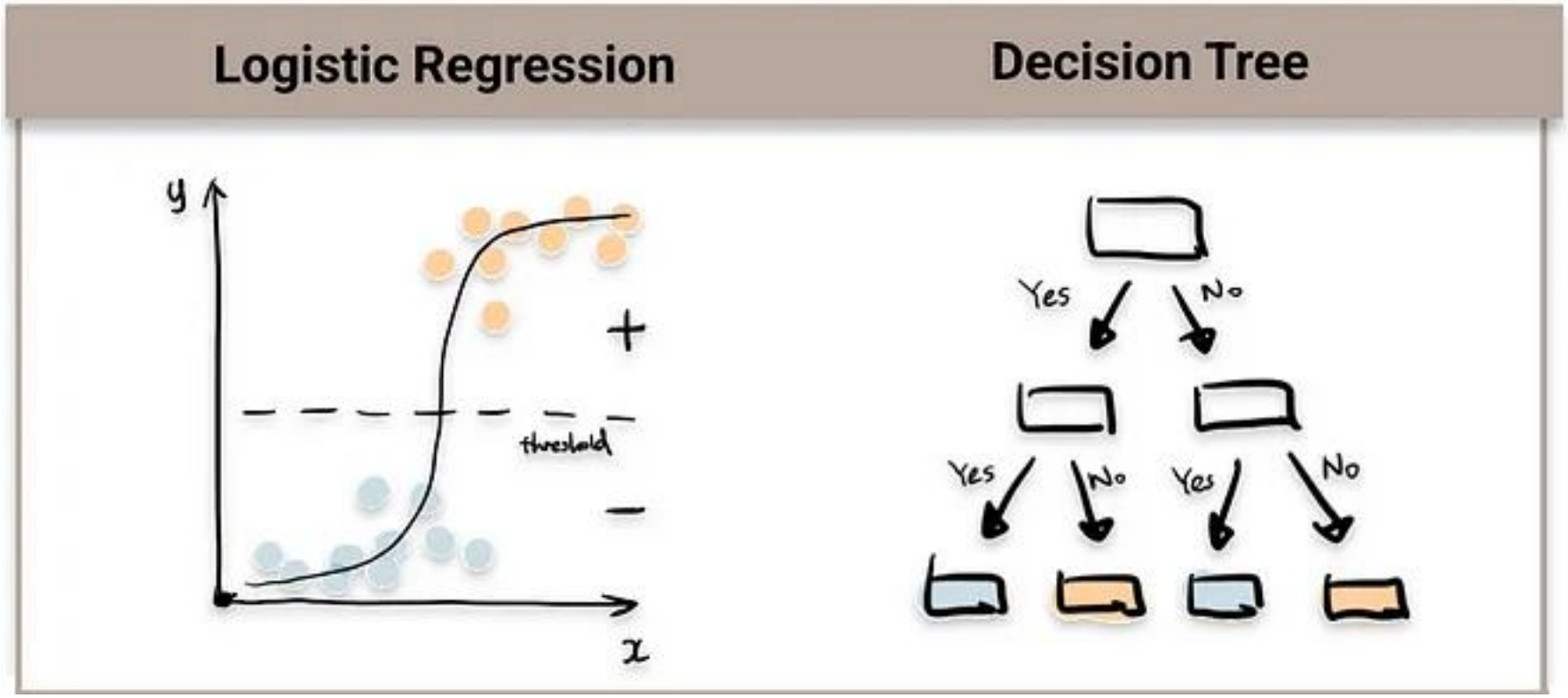
Exploratory Data Analysis (EDA)

1) **Histogram:** df.plot(kind = 'hist')
2) **Box Plot:** sns.boxplot()
3) **Grouped Bar Chart:** sns.countplot()

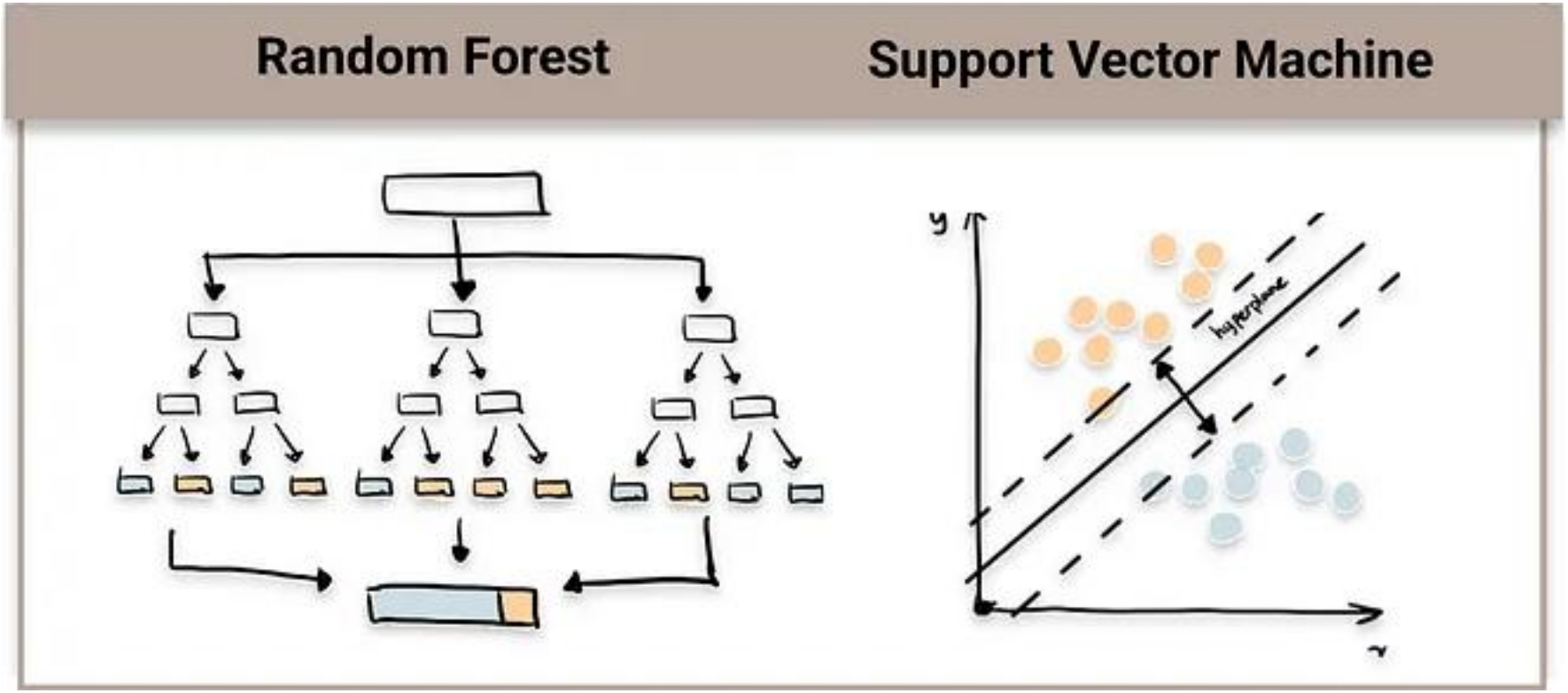




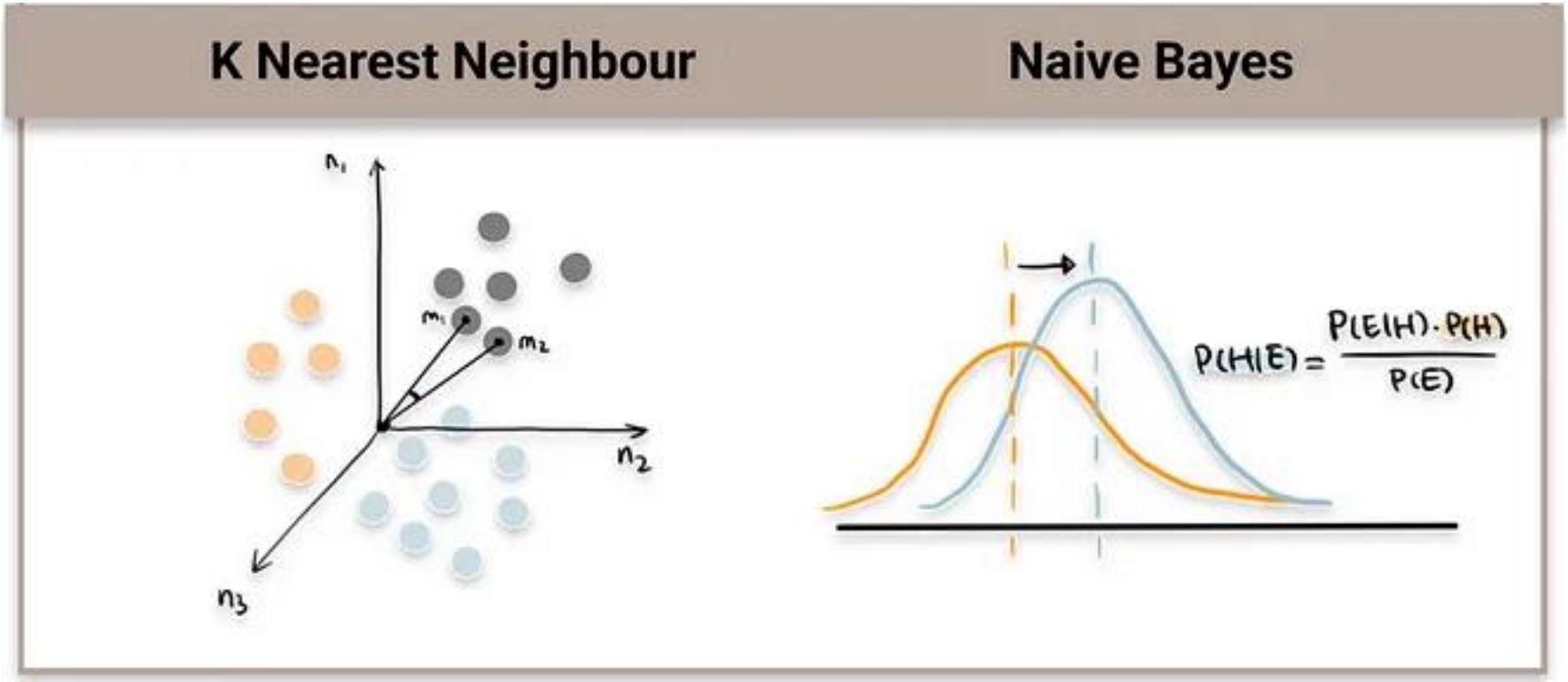
(1.1) Algoritma Klasifikasi



(1.1) Algoritma Klasifikasi



(1.1) Algoritma Klasifikasi

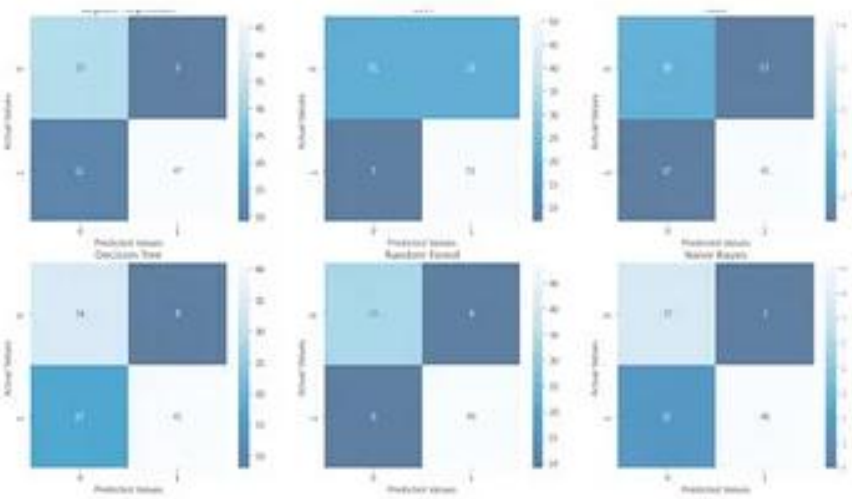


(1.2) Evaluasi Algoritma Klasifikasi

Model Evaluation

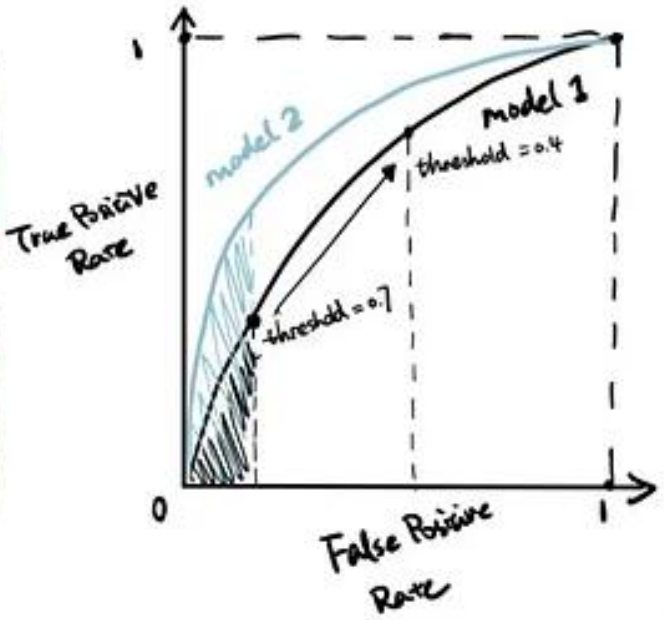
Confusion Matrix

confusion_matrix(y_test, y_pred)

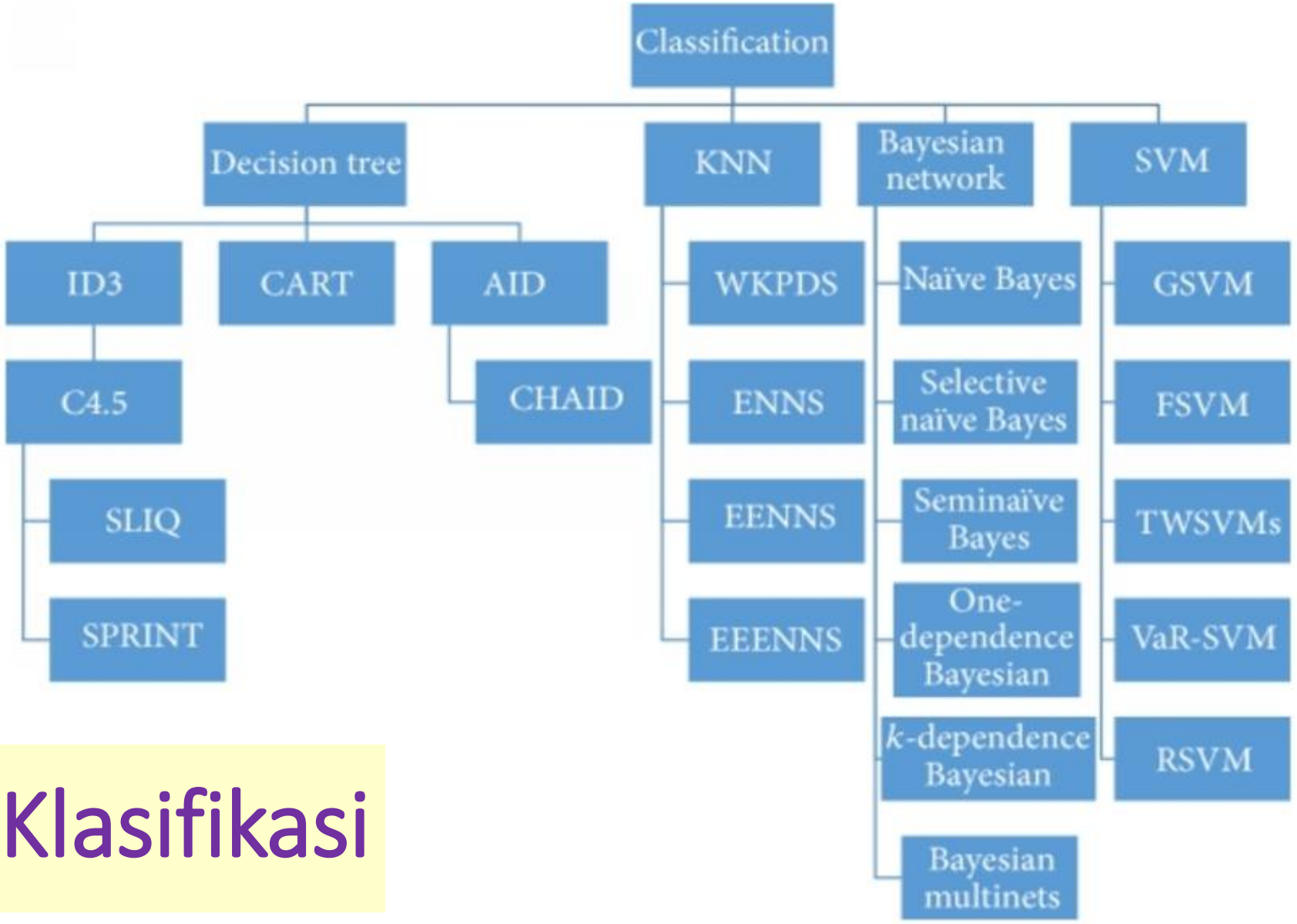


ROC & AUC

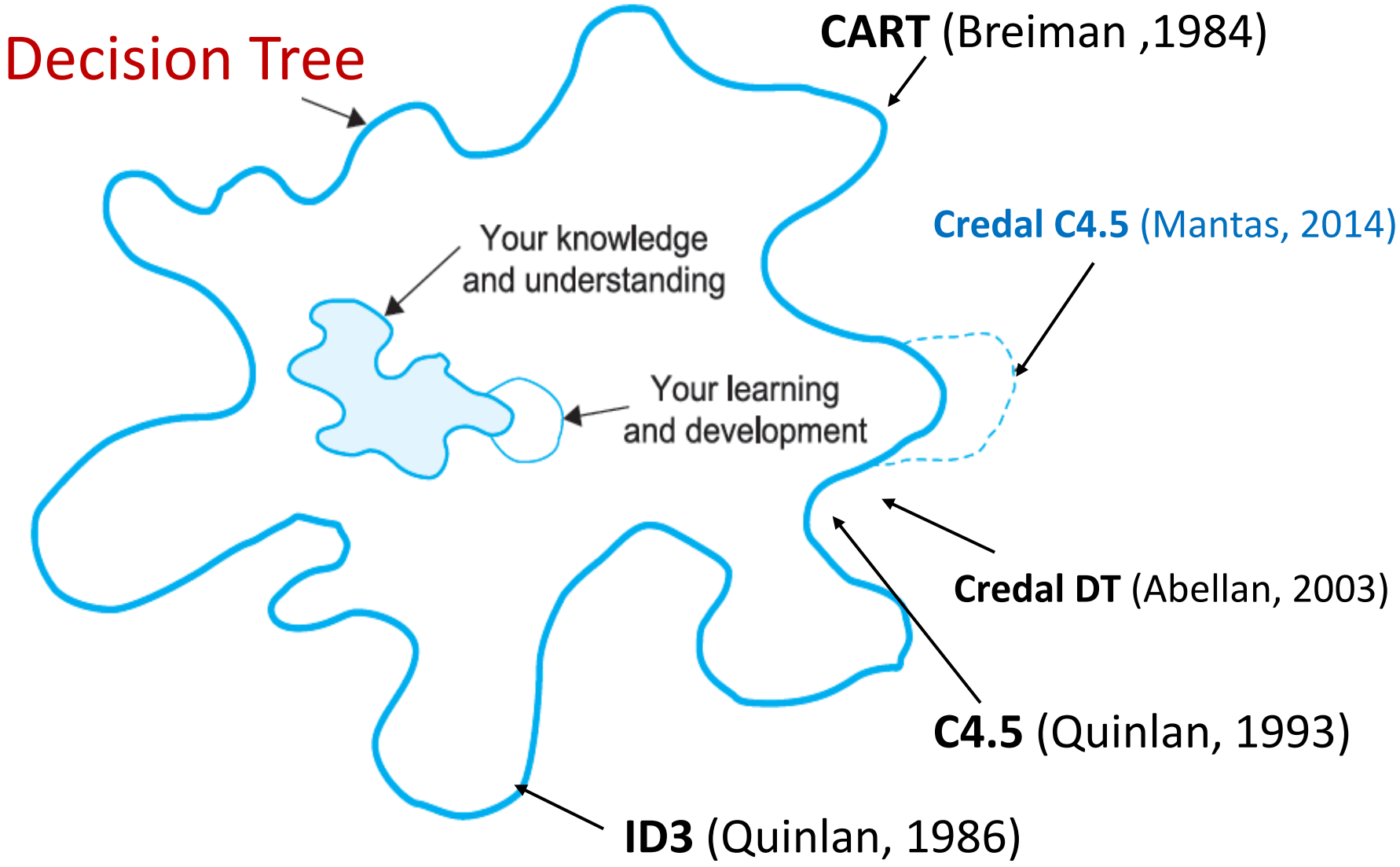
metrics.auc(fpr, tpr)



Adaptive Credal C4.5

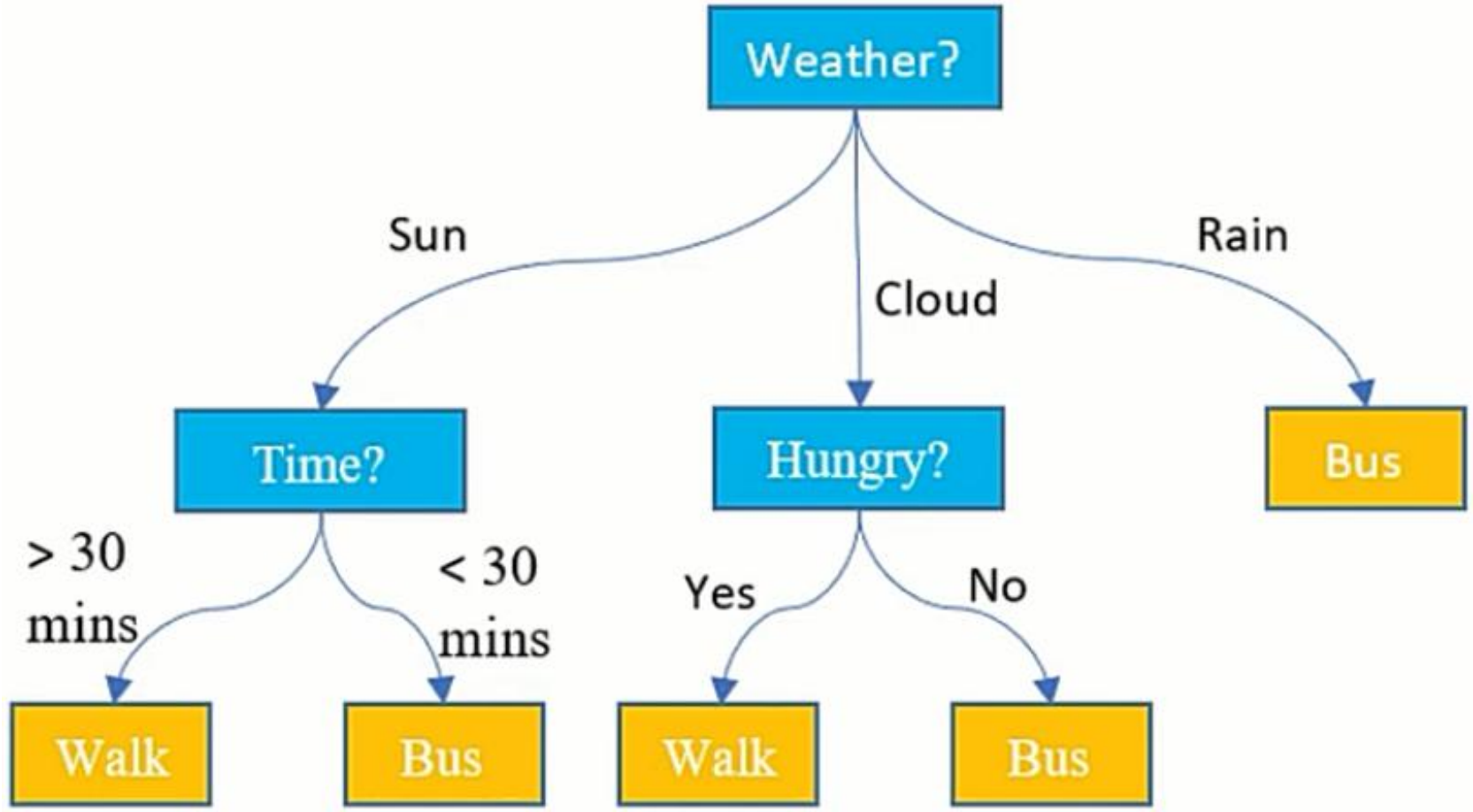


Algoritma Klasifikasi


Decision Tree
CART (Breiman ,1984)
Credal C4.5 (Mantas, 2014)
Your knowledge and understanding
Your learning and development
Credal DT (Abellan, 2003)
C4.5 (Quinlan, 1993)
ID3 (Quinlan, 1986)

Decision Tree

"*Decision tree* adalah **model prediksi** menggunakan **struktur pohon** atau **struktur hiraraki**"



Kebutuhan Klasifikasi dengan Decision Tree

1

Weather/cuaca	Time/Waktu	Hungry/Lapar	Pilihan Kendaraan
Rain	>30	Yes	Bus
Cluod	>30	No	Walk
Rain	<30	Yes	Bus
Rain	>30	No	Bus
Sun	<30	No	Bus
Cloud	<30	Yes	Bus
.....	....	.....	.....

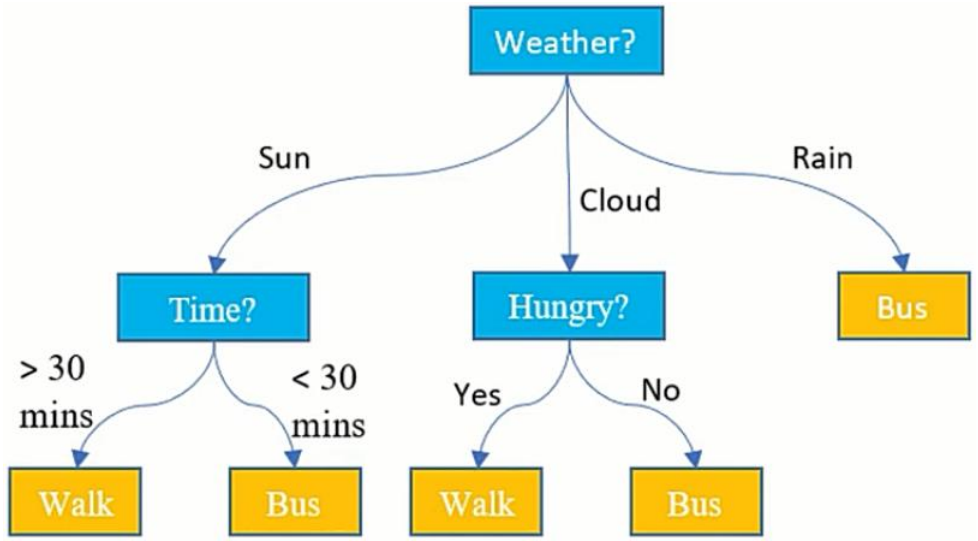
Struktur dataset harus memilik output/target atribut

Kebutuhan Klasifikasi dengan Decision Tree

2

Proses/Parameter: *splitting* (melakukan percabangan) dengan ***information gain***, ***gain ratio***, dan ***gini index***

3

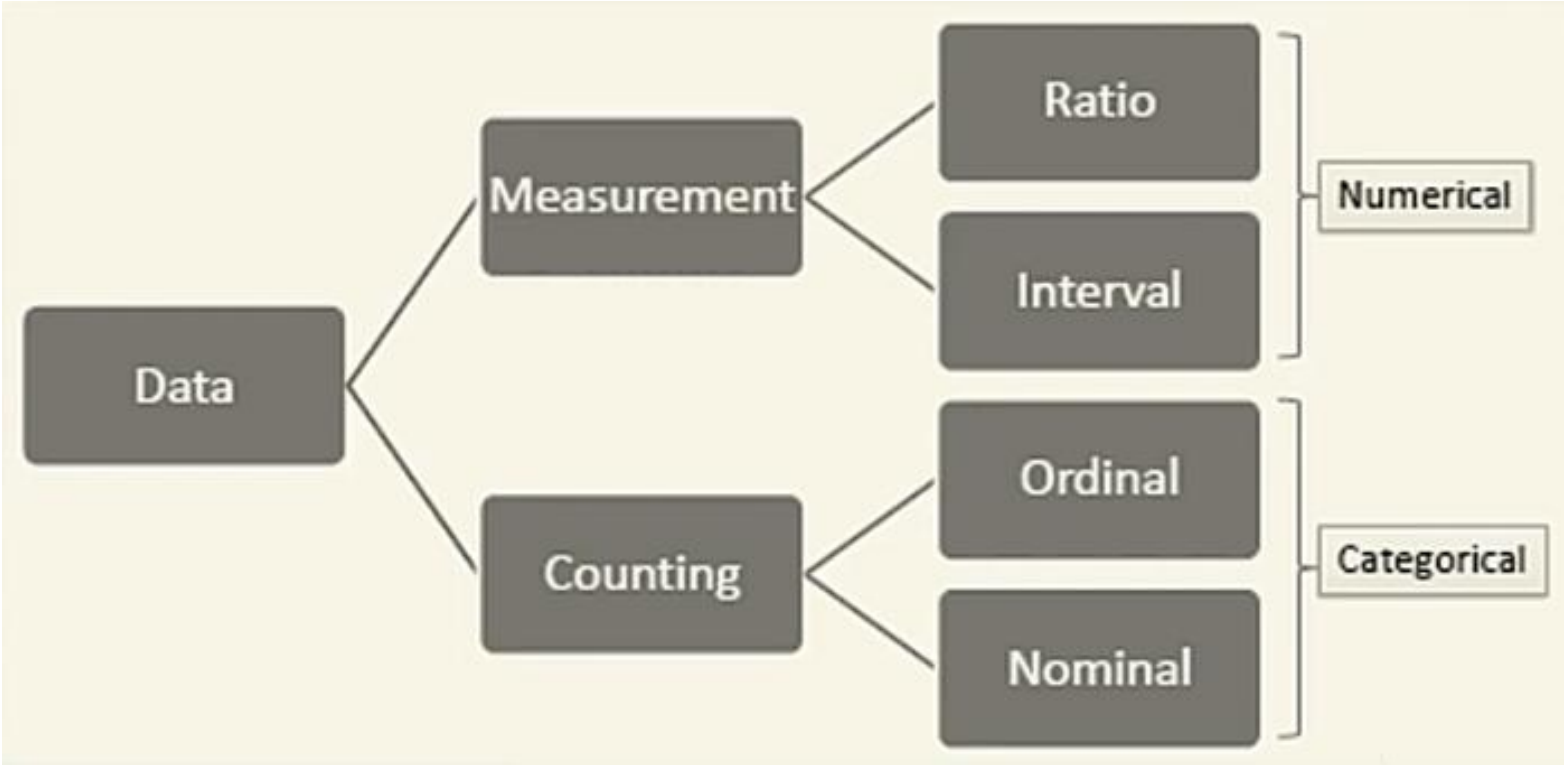


Output: **pohon keputusan** yang mengilustrasikan sebuah rule/aturan

Tujuan Model Decision Tree

Jika data berupa **model/pohon sudah dibuat**, maka ketika **ada data baru** model tersebut dapat digunakan untuk melakukan **prediksi** berdasarkan kaidah yang ada pada model tersebut

Tipe Data pada Decision Tree



Decision Tree menggunakan tipe data **Categorical**

Gambaran Umum Decision Tree

- ***Tree*** dibangun dalam **struktur hirarki**
- **Dimulai** dari mencari variabel ***root***
- **Atribut** bernilai **kategori** (jika nilainya **kontinyu**, maka **didiskritkan dahulu**)
- **Atribut uji** dipilih berdasarkan ukuran heurstik atau statistik (***information gain***, ***gain ratio***, ***gini index***)

Decision Tree
(ID3 algorithm)

Membahas prinsip kerja dari algoritma ID3 pada kasus diskrit

Iterative Dichotomiser 3 (ID3)

- **Dasar Algoritma** (Greedy algorithm)
  1. ***Tree*** dibangun dalam **struktur hirarki**
  2. **Dimulai** dari mencari variabel ***root***
  3. **Atribut** bernilai **kategori** (jika nilainya **kontinyu**, maka **didiskritkan dahulu**)
  4. Sampel dipartisi secara **rekursif** berdasarkan atribut yang dipilih
  5. **Atribut uji** dipilih berdasarkan ukuran heurstik atau statistik (***information gain***, ***gain ratio***, ***gini index***)

- Kondisi **penghentian partisi**
  - Semua sampel untuk node yang diberikan milik kelas yang sama
  - Tidak ada atribut yang tersisa untuk partisi lebih lanjut – **voting majority** digunakan untuk mengklasifikasikan daun
  - Tidak ada sampel yang tersisa

Konsep Entropy

- Entropy mengukur ketidakpastian suatu variabel acak.

Kepastian Angka
Dadu yang muncul?

Kepastian Gambar
atau Angka?

Masalahnya jika dadu yang dilempar memiliki ketidakpastian yang lebih tinggi dari uang logam yang dilempar, berapa besar?

(2.1) Rumus Entropy

- **Entropy** menggunakan **konsep probabilitas** dalam menentukan besar entropy suatu kejadian.
- Misal probabilitas uang yang normal adalah ½ untuk gambar dan ½ untuk angka,
- sementara untuk dadu tiap angka memiliki peluang yang sama yaitu 1/6

$$Entropy(S) = \sum_{i=1}^{n} -pi * log_2\ pi$$

Keterangan:

- S : Himpunan kasus
- A : Fitur atau atribut
- N : jumlah partisi S
- pi : proporsi dari S_i terhadap S

(2.2) Entropy pada Decision Tree

- **Entropi** berfungsi sebagai pengontrol pohon keputusan untuk memutuskan di mana data akan dibagi.
- **Entropi** digunakan untuk menghitung homogenitas sampel.
- Interpretasi **nilai entropy**:
  - Semakin **tinggi** → ketidakpastian semakin tinggi (semakin tidak homogen)
  - Semakin **rendah** → ketidakpastian semakin rendah (semakin homogen)

Information Gain

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i)$$

S	=	Himpunan Kasus
A	=	Atribut
n	=	Jumlah Partisi Atribut *A*
| Si |	=	Jumlah Kasus pada partisi ke-*i*
| S |	=	Jumlah Kasus dalam *S*

Langkah ID3

1. Siapkan **data training**
2. Pilih **atribut** sebagai **akar**
3. Hitung **entropi** $$Entropy(S) = \sum_{i=1}^{n} - pi * \log_2 pi$$
4. **Seleksi atribut** $$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i)$$
5. Buat **cabang** untuk tiap-tiap nilai ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Contoh Kasus

No	OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Pilih akar sebagai atribut

- Pemilihan atribut akar, didasarkan pada nilai Gain tertinggi dari atribut-atribut yang ada.
- Nilai Gain, harus ditentukan terlebih dahulu nilai Entropy-nya.

NODE			Jml Kasus (S)	Tidak (S_1)	Ya (S_2)	Entropy	Gain
1	TOTAL						
	OUTLOOK						
		CLOUDY					
		RAINY					
		SUNNY					
	TEMPERATURE						
		COOL					
		HOT					
		MILD					
	HUMIDITY						
		HIGH					
		NORMAL					
	WINDY						
		FALSE					
		TRUE					

Penghitungan Entropy Total

Langkah pertama menghitung **Entropi Total**

	Play		
	Yes	No	Total
Total Kasus	10	4	14

$$Entropy(Total) = (-\frac{4}{14} * \log_2(\frac{4}{14})) + (-\frac{10}{14} * \log_2(\frac{10}{14}))$$

$$Entropy(Total) = 0.863120569$$

Penghitungan Entropy Akar

No	OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

- Entropy (Outlook)

$$Entropy(Cloudy) = (-\frac{0}{4} * \log_2(\frac{0}{4})) + (-\frac{4}{4} * \log_2(\frac{4}{4})) = 0.000000000$$

$$Entropy(Rainy) = (-\frac{1}{5} * \log_2(\frac{1}{5})) + (-\frac{4}{5} * \log_2(\frac{4}{5})) = 0.721928095$$

$$Entropy(Sunny) = (-\frac{3}{5} * \log_2(\frac{3}{5})) + (-\frac{2}{5} * \log_2(\frac{2}{5})) = 0.970950594$$

- Entropy (Temperature)

$$Entropy(Cool) = (-\frac{0}{4} * \log_2(\frac{0}{4})) + (-\frac{4}{4} * \log_2(\frac{4}{4})) = 0.000000000$$

$$Entropy(Hot) = (-\frac{2}{4} * \log_2(\frac{2}{4})) + (-\frac{2}{4} * \log_2(\frac{2}{4})) = 1.000000000$$

$$Entropy(Mild) = (-\frac{2}{6} * \log_2(\frac{2}{6})) + (-\frac{4}{6} * \log_2(\frac{4}{6})) = 0.918295834$$

$$\frac{log_{10}\,8}{log_{10}\,2}$$

- Entropy (Humidity)

$$Entropy(High) = (-\frac{4}{7} * \log_2(\frac{4}{7})) + (-\frac{3}{7} * \log_2(\frac{3}{7})) = 0.985228136$$

$$Entropy(Normal) = (-\frac{0}{7} * \log_2(\frac{0}{7})) + (-\frac{7}{7} * \log_2(\frac{7}{7})) = 0.000000000$$

- Entropy (Windy)

$$Entropy(False) = (-\frac{2}{8} * \log_2(\frac{2}{8})) + (-\frac{6}{8} * \log_2(\frac{6}{8})) = 0.811278124$$

$$Entropy(True) = (-\frac{4}{6} * \log_2(\frac{4}{6})) + (-\frac{2}{6} * \log_2(\frac{2}{6})) = 0.918295834$$

Hasil Penghitungan Entropy Akar

NODE	ATRIBUT		JML KASUS (S)	YA (Si)	TIDAK (Si)	ENTROPY	GAIN
1	**TOTAL**		14	10	4	0,86312	
	OUTLOOK						
		CLOUDY	4	4	0	0	
		RAINY	5	4	1	0,72193	
		SUNNY	5	2	3	0,97095	
	TEMPERATURE						
		COOL	4	0	4	0	
		HOT	4	2	2	1	
		MILD	6	2	4	0,91830	
	HUMADITY						
		HIGH	7	4	3	0,98523	
		NORMAL	7	7	0	0	
	WINDY						
		FALSE	8	2	6	0,81128	
		TRUE	6	4	2	0,91830	

Penghitungan Gain Akar

$$Gain\,(Total, Outlook) = Entropy(Total) - \sum_{i=1}^{n} \frac{|\,Outlook_i\,|}{|\,Total\,|} * Entropy\,(Outlook_i)$$

$$Gain\,(Total, Outlook) = 0.863120569 - \left(\left(\frac{4}{14} * 0.000000000\right) + \left(\frac{5}{14} * 0.721928095\right) + \left(\frac{5}{14} * 0.970950594\right)\right)$$

$$Gain\,(Total, Outlook) = 0.258521037$$

$$Gain(Total, Temperature) = Entropy(Total) - \sum_{i=1}^{n} \frac{|\,Temperature_i\,|}{|\,Total\,|} * Entropy\,(Temperature_i)$$

$$Gain(Total, Temperature) = 0.863120569 - \left(\left(\frac{4}{14} * 0.000000000\right) + \left(\frac{4}{14} * 1.000000000\right) + \left(\frac{6}{14} * 0.918295834\right)\right)$$

$$Gain(Total, Temperature) = 0.183850925$$

$$Gain(Total, Humidity) = Entropy(Total) - \sum_{i=1}^{n} \frac{|\,Humidity_i\,|}{|\,Total\,|} * Entropy\,(Humidity_i)$$

$$Gain(Total, Humidity) = 0.863120569 - \left(\left(\frac{7}{14} * 0.985228136\right) + \left(\frac{7}{14} * 0.000000000\right)\right)$$

$$Gain(Total, Humidity) = 0.370506501$$

$$Gain(Total, Windy) = Entropy(Total) - \sum_{i=1}^{n} \frac{|\,Windy_i\,|}{|\,Total\,|} * Entropy\,(Windy_i)$$

$$Gain(Total, Windy) = 0.863120569 - \left(\left(\frac{8}{14} * 0.811278124\right) + \left(\frac{6}{14} * 0.918295834\right)\right)$$

$$Gain(Total, Windy) = 0.005977711$$

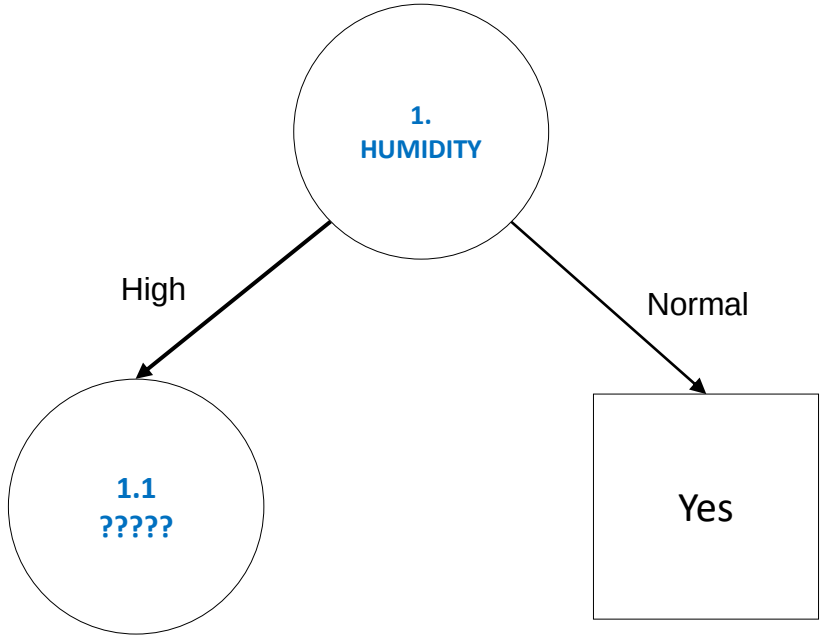
Hasil Penghitungan Gain Akar

NODE	ATRIBUT		JML KASUS (S)	YA (Si)	TIDAK (Si)	ENTROPY	GAIN
1	**TOTAL**		14	10	4	0,86312	
	OUTLOOK						0,25852
		CLOUDY	4	4	0	0	
		RAINY	5	4	1	0,72193	
		SUNNY	5	2	3	0,97095	
	TEMPERATURE						0,18385
		COOL	4	0	4	0	
		HOT	4	2	2	1	
		MILD	6	2	4	0,91830	
	HUMIDITY						0,37051
		HIGH	7	4	3	0,98523	
		NORMAL	7	7	0	0	
	WINDY						0,00598
		FALSE	8	2	6	0,81128	
		TRUE	6	4	2	0,91830	

Gain Tertinggi Sebagai Akar

- Dari hasil pada **Node 1**, dapat diketahui bahwa atribut dengan Gain tertinggi adalah HUMIDITY yaitu sebesar 0.37051
  - Dengan demikian HUMIDITY dapat menjadi node akar

- Ada 2 nilai atribut dari HUMIDITY yaitu HIGH dan NORMAL. Dari kedua nilai atribut tersebut, nilai atribut NORMAL sudah mengklasifikasikan kasus menjadi 1 yaitu keputusan-nya Yes, sehingga tidak perlu dilakukan perhitungan lebih lanjut
  - Tetapi untuk nilai atribut HIGH masih perlu dilakukan perhitungan lagi



2. Buat cabang untuk tiap-tiap nilai

- Untuk memudahkan, dataset di filter dengan mengambil data yang memiliki kelembaban **HUMADITY = HIGH** untuk membuat tabel Node 1.1

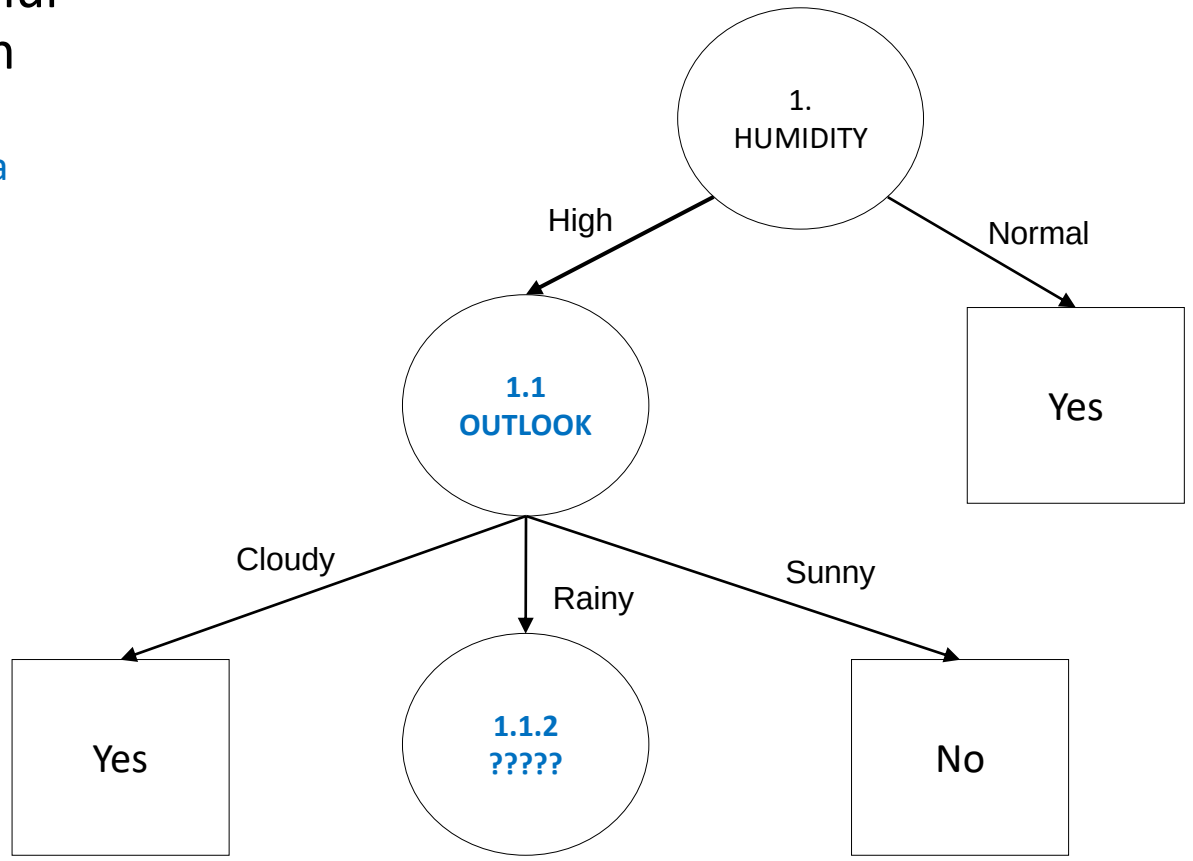
OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
Sunny	Hot	High	FALSE	No
Sunny	Hot	High	TRUE	No
Cloudy	Hot	High	FALSE	Yes
Rainy	Mild	High	FALSE	Yes
Sunny	Mild	High	FALSE	No
Cloudy	Mild	High	TRUE	Yes
Rainy	Mild	High	TRUE	No

Perhitungan Entropi Dan Gain Cabang

NODE	ATRIBUT		JML KASUS (S)	YA (Si)	TIDAK (Si)	ENTROPY	GAIN
1.1	**HUMADITY**		7	3	4	0,98523	
	OUTLOOK						0,69951
		CLOUDY	2	2	0	0	
		RAINY	2	1	1	1	
		SUNNY	3	0	3	0	
	TEMPERATURE						0,02024
		COOL	0	0	0	0	
		HOT	3	1	2	0,91830	
		MILD	4	2	2	1	
	WINDY						0,02024
		FALSE	4	2	2	1	
		TRUE	3	1	2	0,91830	

Gain Tertinggi Sebagai Node 1.1

- Dari hasil pada Tabel Node 1.1, dapat diketahui bahwa atribut dengan Gain tertinggi adalah OUTLOOK yaitu sebesar 0.69951
  - Dengan demikian OUTLOOK dapat menjadi node kedua
- Artibut CLOUDY = YES dan SUNNY= NO sudah mengklasifikasikan kasus menjadi 1 keputusan, sehingga tidak perlu dilakukan perhitungan lebih lanjut
  - Tetapi untuk nilai atribut RAINY masih perlu dilakukan perhitungan lagi



Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yg sama

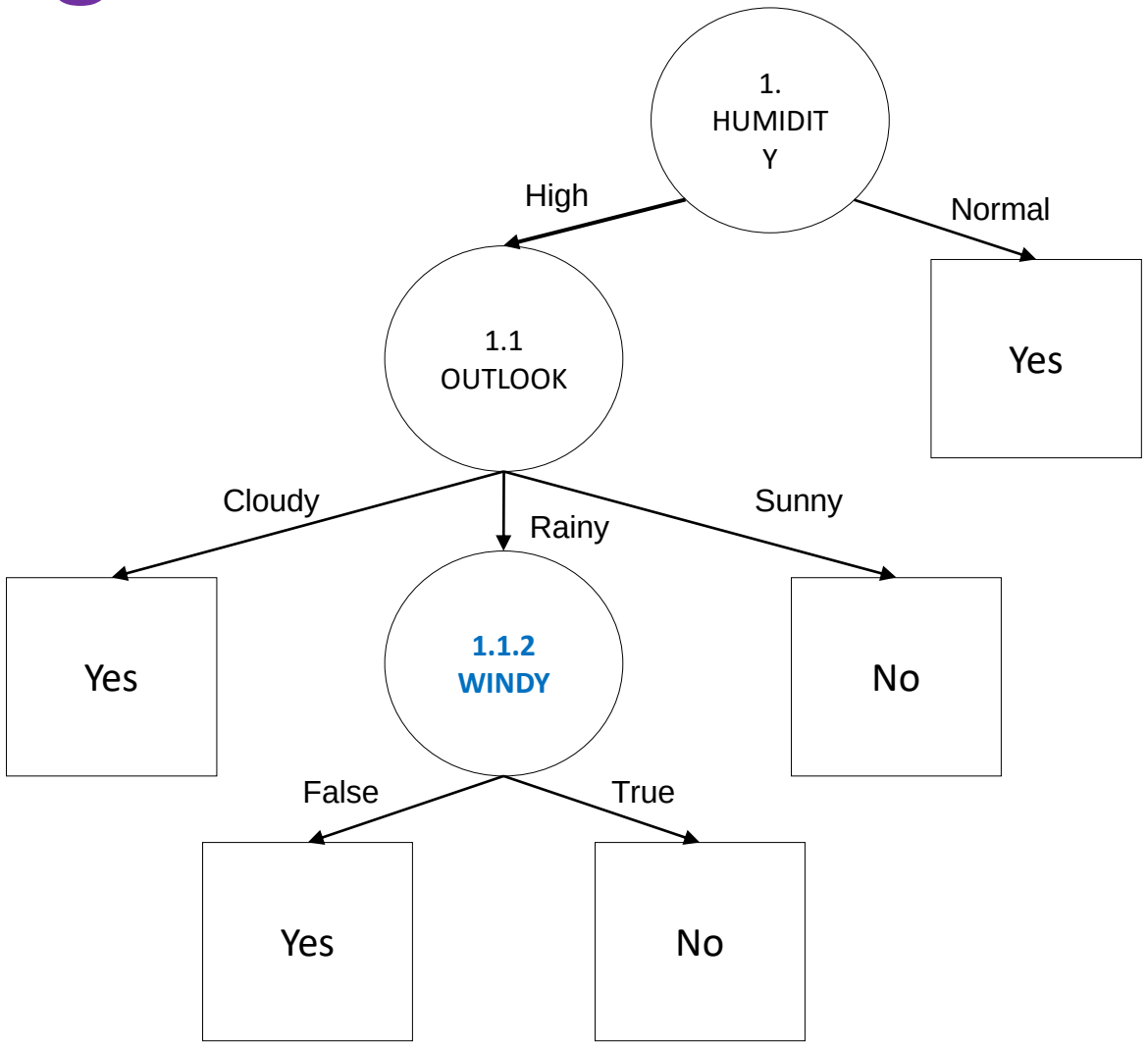
OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
Rainy	Mild	High	FALSE	Yes
Rainy	Mild	High	TRUE	No

NODE	ATRIBUT		JML KASUS (S)	YA (Si)	TIDAK (Si)	ENTROPY	GAIN
1.2	**HUMADITY HIGH & OUTLOOK RAINY**		2	1	1	1	
	TEMPERATURE						0
		COOL	0	0	0	0	
		HOT	0	0	0	0	
		MILD	2	1	1	1	
	WINDY						1
		FALSE	1	1	0	0	
		TRUE	1	0	1	0	

Gain Tertinggi Sebagai Node 1.1.2

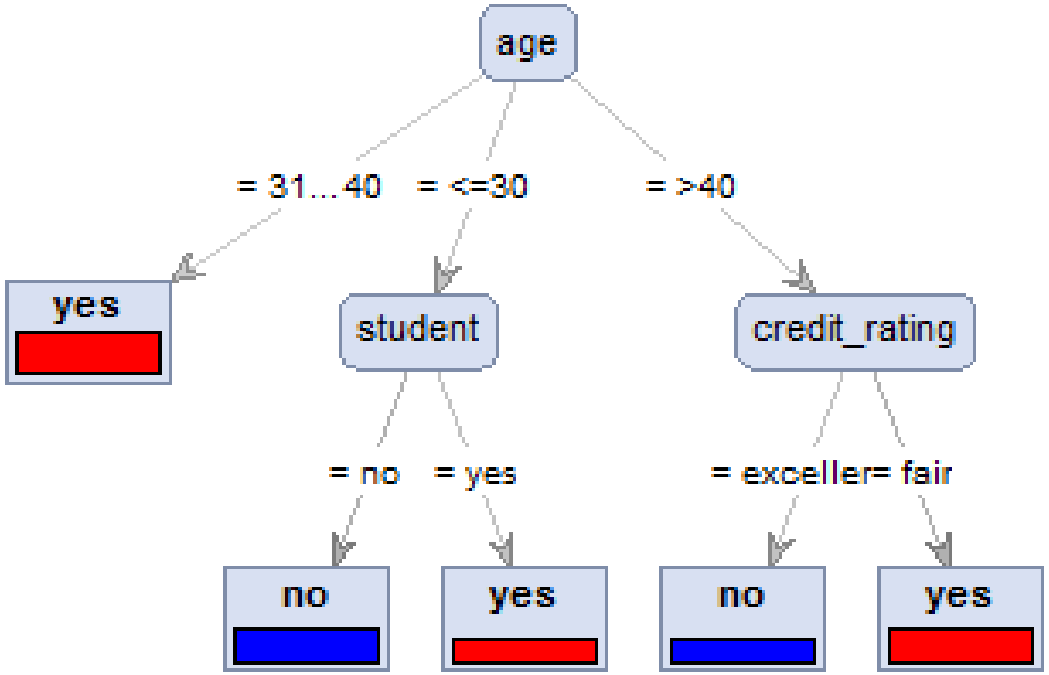
- Dari tabel, Gain Tertinggi adalah WINDY dan menjadi node cabang dari atribut RAINY

- Karena semua kasus sudah masuk dalam kelas
  - Jadi, pohon keputusan pada Gambar merupakan pohon keputusan terakhir yang terbentuk



Latihan

- Training data set:
  Buys_computer



age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Decision Tree
(C4.5 algorithm)

Membahas prinsip kerja dari algoritma C4.5 pada kasus diskrit

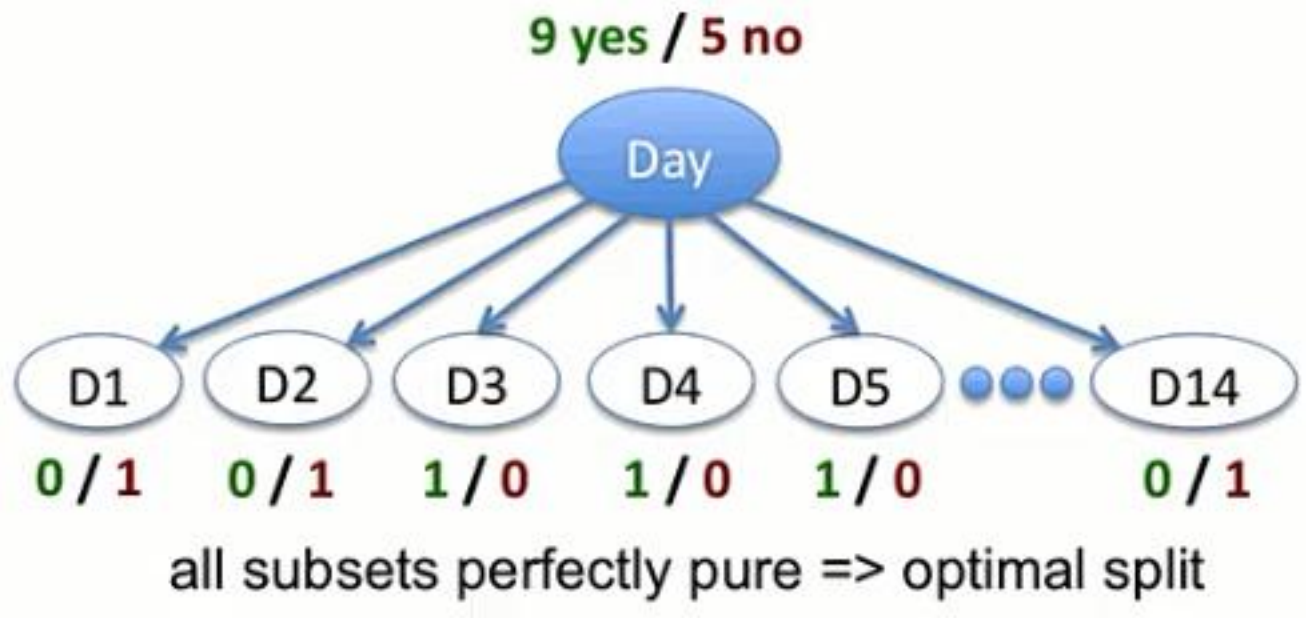
C4.5

- Decision Tree model **C4.5** merupakan pengembangan dari ID3.
- **C4.5** menggunakan **Gain Ratio** sebagai fungsi seleksi atribut untuk membagi dataset, sedangkan **ID3** menggunakan **Information Gain**.
- **Information Gain** cenderung memilih atribut dengan lebih banyak kategori karena cenderung memiliki entropi yang lebih rendah, sehingga **menyebabkan *overfitting*** pada data *training*.
- **Gain Ratio** mengurangi masalah ini dengan menghukum atribut karena memiliki banyak kategori.

Information Gain Problems

- **Biased** towards attributes with many values
- Won't work for new data:

  D15 Rain High Weak



Rumus Umum Gain Ratio

$$SplitEntropy(S,A) = -\sum_{i=1}^{n} \frac{|S_i|}{|S|} \times log_2\left(\frac{|S_i|}{|S|}\right)$$

Keterangan:

- *S* = Himpunan Kasus
- *A* = Atribut
- *n* = Jumlah Partisi Atribut *A*
- | Si | = Jumlah Kasus pada partisi ke-*i*
- | S | = Jumlah Kasus dalam *S*

$$GainRatio(A) = \frac{Gain(A)}{SplitEntropy(S,A)}$$

Atribut dengan nilai gain tertinggi dipilih sebagai atribut akar

(3.1) Contoh Penggunaan Gain Ratio

Langkah pertama menghitung **Entropi Total**

	Play		
	Yes	No	Total
Total Kasus	10	4	14

$$Entropy(Total) = (-\frac{4}{14} * \log_2(\frac{4}{14})) + (-\frac{10}{14} * \log_2(\frac{10}{14}))$$

$$Entropy(Total) = 0.863120569$$

Langkah kedua menghitung **Entropi Akar** setiap atribut

Contoh entropi akar pada atribut "Outlook"

$$Entropy(Cloudy) = (-\frac{0}{4} * \log_2(\frac{0}{4})) + (-\frac{4}{4} * \log_2(\frac{4}{4})) = 0.000000000$$

$$Entropy(Rainy) = (-\frac{1}{5} * \log_2(\frac{1}{5})) + (-\frac{4}{5} * \log_2(\frac{4}{5})) = 0.721928095$$

$$Entropy(Sunny) = (-\frac{3}{5} * \log_2(\frac{3}{5})) + (-\frac{2}{5} * \log_2(\frac{2}{5})) = 0.970950594$$

(lanjutan) Contoh Penggunaan Gain Ratio

Langkah ketiga menghitung **Gain Akar** setiap atribut

Contoh Gain akar pada atribut "Outlook"

$$Gain\,(Total, Outlook) = Entropy(Total) - \sum_{i=1}^{n} \frac{|\,Outlook_i\,|}{|\,Total\,|} * Entropy\,(Outlook_i)$$

$$Gain\,(Total, Outlook) = 0.863120569 - \left(\left(\frac{4}{14} * 0.000000000\right) + \left(\frac{5}{14} * 0.721928095\right) + \left(\frac{5}{14} * 0.970950594\right)\right)$$

$$Gain\,(Total, Outlook) = 0.258521037$$

Langkah keempat menghitung **Gain Ratio** setiap atribut

Attribute	Value	Play
		Total
Outlook	Sunny	5
	Cloudy	4
	Rainy	5
		14

$$SplitEntropy(Outlook) = -\frac{5}{14} \times log_2 \frac{5}{14} - \frac{4}{14} \times log_2 \frac{4}{14} - \frac{5}{14} \times log_2 \frac{5}{14}$$

$$SplitEntropy(Outlook) = 1.577$$

$$GainRatio(Outlook) = \frac{Gain(Outlook)}{SplitEntropy(outlook)} = \frac{0.258521037}{1.577} = 0.1639$$

(lanjutan) Contoh Penggunaan Gain Ratio

Langkah kelima menentukan **atribut akar**

Atribut dengan nilai gain ratio tertinggi dipilih sebagai atribut akar

Ulangi langkah awal untuk menentukan atribut cabang

Decision Tree
(CART)

Membahas prinsip kerja dari algoritma CART pada kasus diskrit

Classification and Regression Tree (CART)

- **CART** terdiri dari dua metode yaitu metode regresi dan pohon klasifikasi.
- Jika variabel bertipe **kategorik** maka **CART** menghasilkan **pohon klasifikasi** (***decision trees***).
- Jika variabel bertipe **kontinu** (**numerik**) maka **CART** menghasilkan **pohon regresi** (***regression trees***).
- Proses klasifikasi menggunakan **CART** mirip dengan **ID3**.
- Jika pada **ID3** menggunakan **entropi** dan **information gain**, **CART** menggunakan **Gini Index** atau **Gini Impurity**

Gini Index

- **Gini Index** merupakan metrik untuk mengukur seberapa sering elemen yang dipilih secara acak salah diidentifikasi.
- Nilai Gini Index berkisar dari **0** hingga **1**
  - Nilai **0** mewakili persamaan sempurna (***homogenitas*** atau ***pure***)
  - Nilai **1** mewakili ketidaksetaraan sempurna (***heterogenitas*** atau *impurity*).
  - Atribut dengan **Gini Index yang lebih rendah** diutamakan
- **Gini Index** digunakan untuk mengevaluasi kualitas pemisahan (*split criteria*) dengan mengukur perbedaan antara ketidakmurnian simpul induk dan ketidakmurnian bobot dari simpul anak pada **Decision Tree**.

Rumus Umum Gini Index

$$Gini(D) = 1 - \sum_{j=1}^{n} P_j^2$$

Keterangan:

- *Gini(D)* = Gini Index
- *D* = Dataset
- *n* = Jumlah kelas
- Pj = Frekuensi relative kelas J di D

(3.1) Contoh Penggunaan Gini Index

Langkah pertama menghitung **Gini Index Total**

	Play		
	Yes	No	Total
Total Kasus	10	4	14

$$Gini(D) = 1 - \left[\left(\frac{10}{14}\right)^2 + \left(\frac{4}{14}\right)^2\right] = 0.4081$$

(lanjutan) Contoh Penggunaan Gini Index

Langkah selanjutnya menghitung **Gini Gain** tiap atribut.
Contoh **Gini Gain** pada atribut "**Outlook**"

Attribute	Value	Play		
		Yes	No	Total
Outlook	Sunny	2	3	5
	Cloudy	4	0	4
	Rainy	4	1	5
				14

$$Gini(D, Outlook) = \left(\frac{5}{14}\right) gini(sunny) + \left(\frac{4}{14}\right) gini(cloudy) + \left(\frac{5}{14}\right) gini(rainy)$$

$$Gini(D, Outlook) = \left(\frac{5}{14}\right)\left(1 - \left[\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right]\right) + \left(\frac{4}{14}\right)\left(1 - \left[\left(\frac{4}{4}\right)^2 + \left(\frac{0}{4}\right)^2\right]\right) + \left(\frac{5}{14}\right)\left(1 - \left[\left(\frac{4}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right]\right)$$

$$Gini(D, Outlook) = (0.375)(1 - [0.16 + 0.36]) + (0.285)(1 - [1 + 0]) + (0.375)(1 - [0.64 + 0.04])$$

$$Gini(D, Outlook) = (0.375)(0.48) + (0.285)(0) + (0.375)(0.32) \quad Gini(D, Outlook) = 0.3$$

- Ketika semua nilai **Gini Gain** diperoleh, maka **pilih** **Gini Gain** **terendah** sebagai **Atribut akar**
- **Ulangi lagi langkah awal** untuk menentukan **atribut cabang**

Naïve Bayes

Membahas konsep algoritma Naïve Bayes untuk kasus klasifikasi

Naïve Bayes

- **Naïve Bayes** adalah sebuah metode klasifikasi menggunakan **metode probabilitas dan statistika** (penemu Thomas Bayes).
- Algoritma ini memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya (**Teorema Bayes**).
- **Ciri utama** Naive Bayes Classifier adalah asumsi yang sangat kuat (naif) akan independensi dari masing-masing kondisi/kejadian.

Kelebihan vs Kekurangan Naïve Bayes

Kelebihan:

- Hanya membutuhkan jumlah data latih (training data) yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses klasifikasi
- Mudah dibuat
- Hasil bagus

Kekurangan:

- Asumsi independensi antar atribut membuat akurasi berkurang (karena biasanya ada keterkaitan)

Algoritma Naïve Bayes

Tahapan:

1. Menghitung jumlah kelas/label
2. Menghitung jumlah kasus per kelas
3. Mengalikan semua variabel kelas
4. Membandingkan hasil per kelas

(3.1) Rumus Naïve Bayes

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)}$$

- P(H|E): Probabilitas akhir bersyarat (conditional probability) suatu hipotesis H terjadi jika diberikan bukti (evidence) E terjadi.
- P(E|H):Probabilitas sebuah bukti E akan mempengaruhi hipotesis H.
- P(H) :Probabilitas awal hipotesis H terjadi tanpa memandang bukti apapun.
- P(E) :Probabilitas awal bukti E terjadi tanpa memandang hipotesis/bukti yang lain.

Contoh Kasus
Naïve Bayes

Membahas penyelesaian kasus klasifiaksi dengan Naïve Bayes

Contoh Kasus

Jika ada data baru (x) berikut:

- Outlook = Sunny
- Temperature = Cool
- Humidity = High
- Wind = Strong

Apakah akan bermain tennis atau tidak?

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

(4.1) Menghitung Probabilitas Jumlah Kelas

- P(PlayTennis = Yes) = 9/14 = 0.64
- P(PlayTennis = No) = 5/14 = 0.36

(4.2) Menghitung Probabilitas Jumlah Kasus Per Kelas

- P(Outl=Sunny | PlayTennis = Yes) = 2/9 = 0.22
- P(Outl=Sunny | PlayTennis = No) = 3/5 = 0.6

- P(Temp=Cool | PlayTennis = Yes) = 3/9 = 0.33
- P(Temp=Cool | PlayTennis = No) = 1/5 = 0.2

- P(Wind=strong | PlayTennis = Yes) = 3/9 = 0.33
- P(Wind=strong | PlayTennis = No) = 3/5 = 0.6

- P(Hum=High | PlayTennis = Yes) = 3/9 = 0.33
- P(Hum=High | PlayTennis = No) = 4/5 = 0.8

(4.3) Mengalikan Semua Variabel Kelas

- data X : Outl=Sunny, Temp=Cool, Hum=High, Wind=strong

- P(X|Yes) = P(Yes) x P(Sunny|Yes) x P(Cool|Yes) x P(High|Yes) x P(Strong|Yes)
  =0.64 x 0.22 x 0.33 x 0.33 x 0.33 = 0.00505

- P(X|No) = P(No) x P(Sunny|No) x P(Cool|No) x P(High|No) x P(Strong|No)
  =0.36 x 0.6 x 0.2 x 0.8 x 0.6 = 0.02073

- Kesimpulan : Karena P(X|No) > P(X|Yes)

- Maka prediksi / klasifikasi untuk data X adalah “No”

(4.4) Membandingkan Hasil Per Kelas

- Kesimpulan : Karena P(X|No) > P(X|Yes)
- Maka prediksi / klasifikasi untuk data X adalah “No”

Latihan

Usia	Tekanan Darah	Jml Bayi	Riwayat Persalinan	Riwayat Abortus	Nutrisi	Penyakit Lain	Masalah Saat Hamil	Usia Kelahiran
Lebih	Tinggi	1	Riwayat Normal	Tidak	Normal	Tidak Ada	PEB	*Postdate*
Kurang	Normal	1	Riwayat Normal	Tidak	Normal	Tidak Ada	Tidak Ada	Normal
Lebih	Normal	1	Riwayat Prematur	Ya	Normal	Anemia	Tidak Ada	*Premature*
Cukup	Tinggi	1	Anak Pertama	Tidak	Normal	Anemia	PER	*Postdate*
Cukup	Normal	1	Riwayat Prematur	Tidak	Normal	Tidak Ada	Tidak Ada	Normal
Cukup	Tinggi	1	Anak Pertama	Tidak	Normal	Hipertensi	PEB	*Premature*
Lebih	Normal	1	Riwayat Normal	Tidak	Normal	Tidak Ada	Tidak Ada	Normal
Lebih	Tinggi	1	Riwayat Prematur	Ya	Normal	Asma	PER	*Premature*
Lebih	Normal	1	Anak Pertama	Tidak	Kurang	Asma	Tidak Ada	*Premature*
Cukup	Normal	2	Riwayat Normal	Tidak	Normal	Tidak Ada	Tidak Ada	*Premature*

Latihan: Data baru yang akan diprediksi

Usia	Tekanan Darah	Jml Bayi	Riwayat Persalinan	Riwayat Abortus	Nutrisi	Penyakit Lain	Masalah Saat Hamil	Usia Kelahiran
Cukup	Tinggi	1	Anak Pertama	Tidak	Normal	Hipertensi	PEB	?