

**PELATIHAN ARTIFICIAL INTELLIGENCE & BIG DATA (DATA SCIENCE)  
PROGRAM UPSKILLING DAN RESKILLING GURU KEJURUAN SMK**



**Made Agus Andi Gunawan**  
**Student Ambassador BISA AI**

# Decision Tree

---

Aplikasi Decision Tree pada  
proses pengambilan keputusan

LAMPUNG, 13 OKTOBER 2021

# Outline

1. Pengenalan Decision Tree
2. Pengenalan komponen Decision Tree: root, node, leaf
3. Pengenalan Gini Impurity
4. Pengenalan Information Gain
5. Membangun Decision Tree
6. Training model Decision Tree Classifier
7. Visualisasi model Decision Tree
8. Evaluasi model Decision Tree



# Decision Tree

salah satu model decision tree  
ditemukan oleh

J. Ross Quinlan



pada bukunya



1986



# Decision Tree

*Decision tree* atau pohon keputusan adalah salah satu algoritma supervised learning yang dapat dipakai untuk masalah klasifikasi dan regresi. Decision tree merupakan algoritma yang powerful alias mampu dipakai dalam masalah yang kompleks. Decision tree juga merupakan komponen pembangun utama algoritma Random Forest, yang merupakan salah satu algoritma paling powerful saat ini.

Decision tree memprediksi sebuah kelas (klasifikasi) atau nilai (regresi) berdasarkan aturan-aturan yang dibentuk setelah mempelajari data.

# **komponen pohon terdiri dari**



**akar**

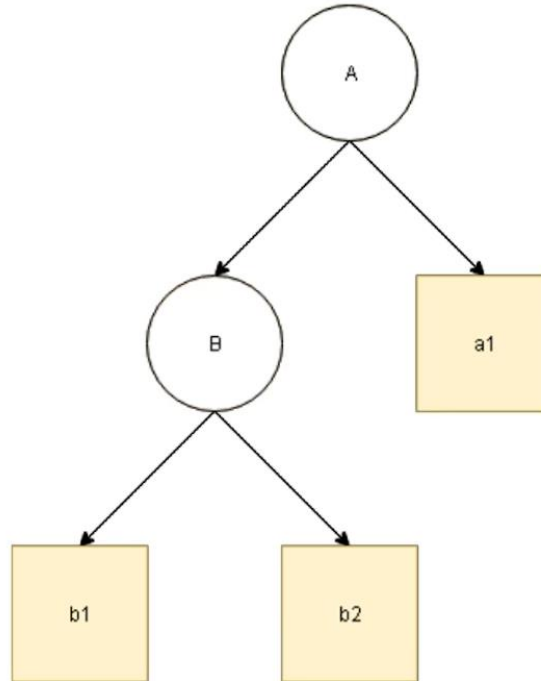


**batang**



**daun**

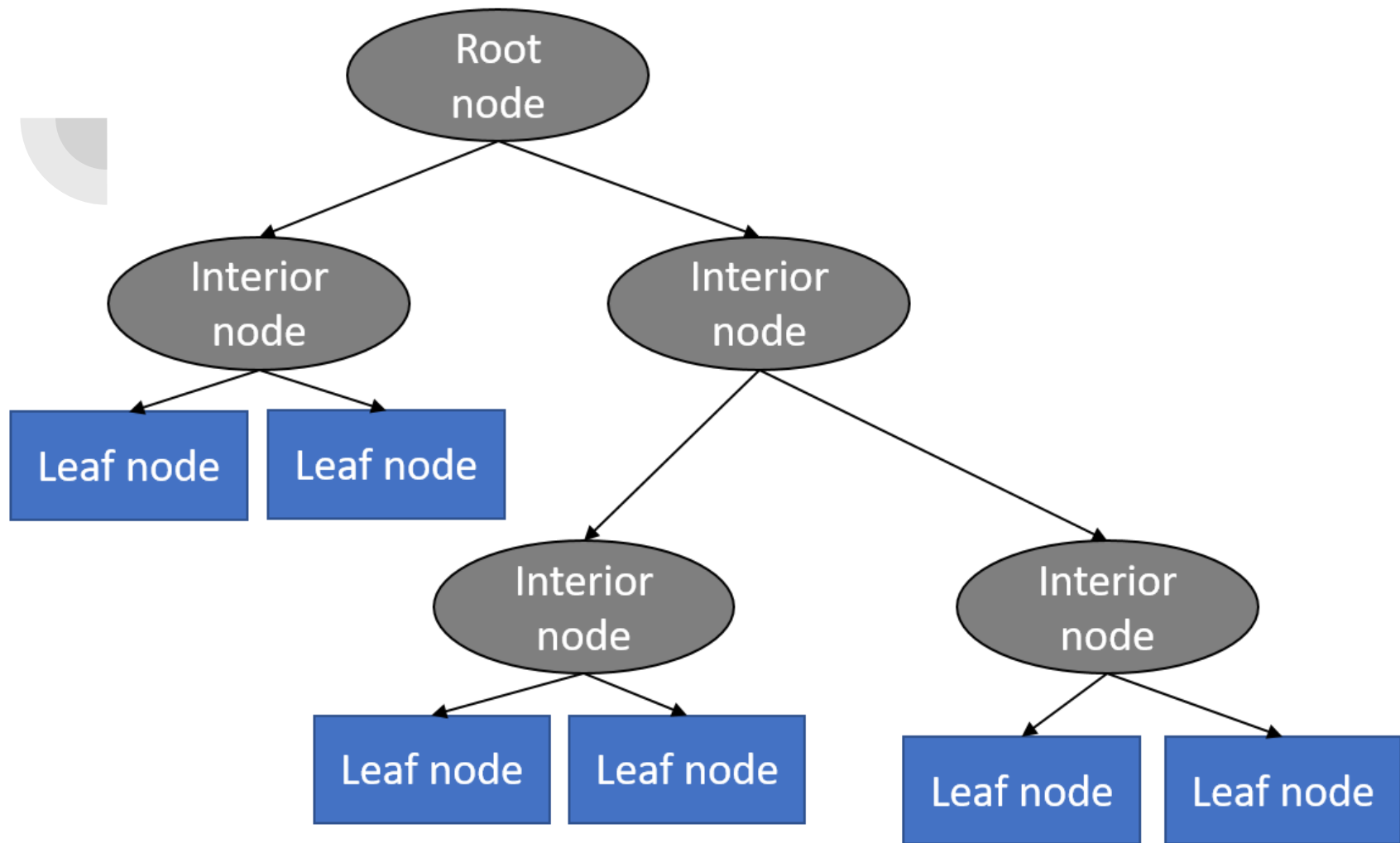
# Komponen Decision Tree



**Root Node**  
terletak paling atas  
dari suatu pohon.

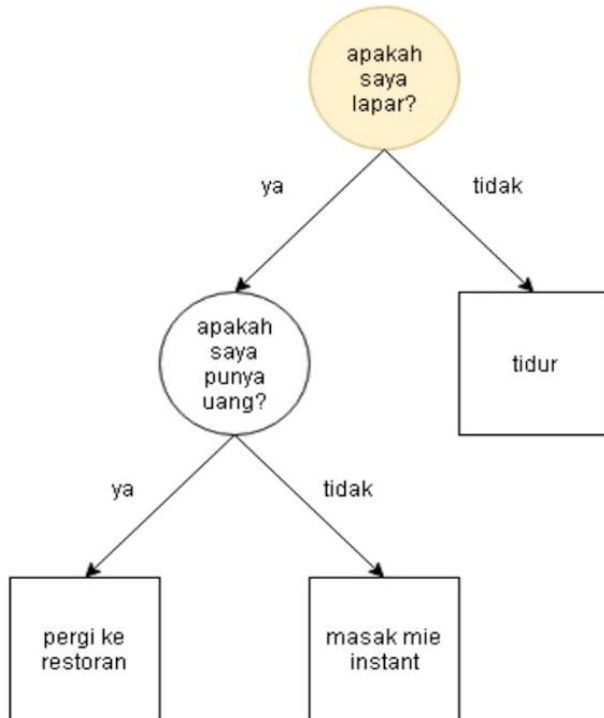
**Internal Node**  
node percabangan,  
memiliki 1 input dan  
minimal 2 output

**Leaf Node**  
node akhir,  
memiliki 1 input dan  
tidak memiliki output





# Studi Kasus Sederhana



- Laper mau makan tapi bingung harus ngapain?

Langkah pertama yaitu membuat root node yang berisi 'apakah saya lapar?'

Solusinya ngapain biar gak laper?

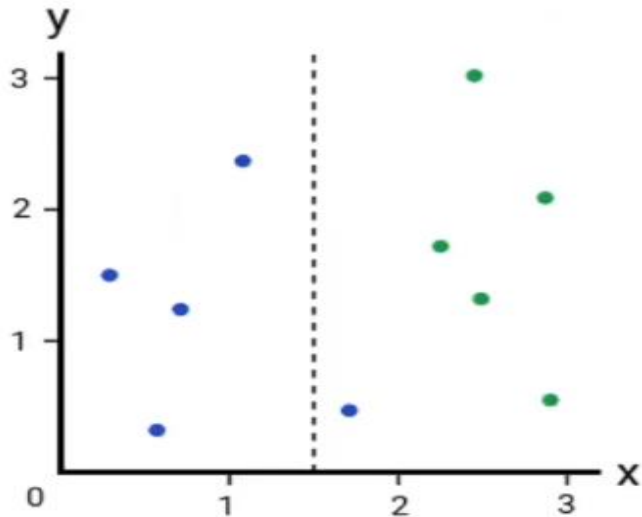


# Keunggulan vs Kekurangan

1. Dapat digunakan untuk regresi atau klasifikasi.
2. Dapat ditampilkan secara grafis. Sangat bisa ditafsirkan.
3. Dapat ditentukan sebagai serangkaian aturan, dan lebih mendekati pengambilan keputusan manusia daripada model lainnya.
4. Prediksinya cepat.
5. Fitur tidak perlu penskalaan.
6. Secara otomatis mempelajari interaksi fitur.
7. Cenderung mengabaikan fitur yang tidak relevan.

1. Kinerja (umumnya) tidak kompetitif dengan metode pembelajaran terawasi terbaik.
2. Diperlukan penyetelan
3. varian tinggi
4. Pemisahan biner rekursif membuat keputusan "optimal secara lokal" yang mungkin tidak menghasilkan pohon yang optimal secara global.
5. Tidak berfungsi dengan baik dengan kumpulan data yang tidak seimbang.

# Gini Impurity



Ruas Kiri:

$$\begin{aligned} G &= 1 - \sum_i^n P_i^2 \\ &= 1 - P(\text{biru})^2 \\ &= 1 - \left(\frac{4}{4}\right)^2 = 0 \end{aligned}$$

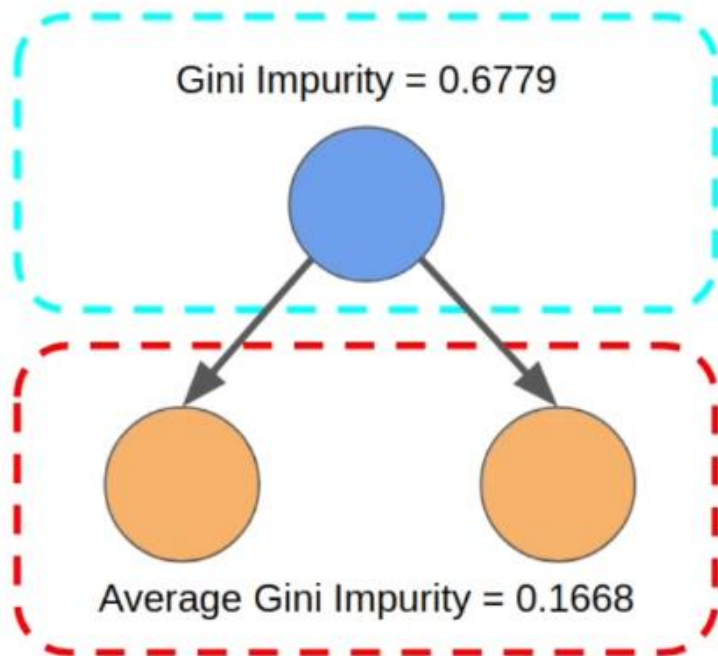
Ruas Kanan:

$$\begin{aligned} G &= 1 - \sum_i^n P_i^2 \\ &= 1 - (P(\text{biru})^2 + P(\text{hijau})^2) \\ &= 1 - \left(\left(\frac{1}{6}\right)^2 + \left(\frac{5}{6}\right)^2\right) = 0.278 \end{aligned}$$

Average Gini Impurity:

$$\begin{aligned} G &= \frac{4}{4+6} \times 0 + \frac{6}{4+6} \times 0.278 \\ &= 0.1668 \end{aligned}$$

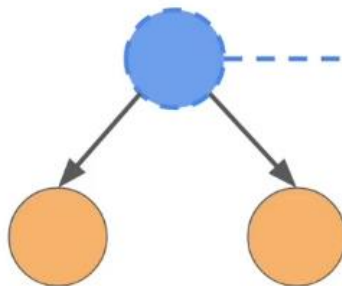
# Information Gain



$$\text{Information Gain} = 0.6779 - 0.1668 = 0.51$$

# Membangun Decision Tree

	Color	Diameter	Label
0	Green	3	Apple
1	Yellow	3	Apple
2	Red	1	Grape
3	Red	1	Grape
4	Yellow	3	Lemon




Kemungkinan pertanyaan untuk splitting:

1. Color == Green?
2. Color == Yellow?
3. Color == Red?
4. Diameter <=1
5. Diameter <=3

Highest Information Gain

$$\begin{aligned} G &= 1 - (P(apple)^2 + P(grape)^2 + P(lemon)^2) \\ &= 1 - ((\frac{2}{5})^2 + (\frac{2}{5})^2 + (\frac{1}{5})^2) \\ &= 0.63 \end{aligned}$$



Outlook	Temperature	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cold	Normal	False	Yes
Sunny	Cold	Normal	True	No
Overcast	Cold	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cold	Normal	False	Yes
Rainy	Mild	Normal	False	Yes

## Attributes

## Classes

Outlook	Temperature	Humidity	Windy	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Sunny	Mild	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No



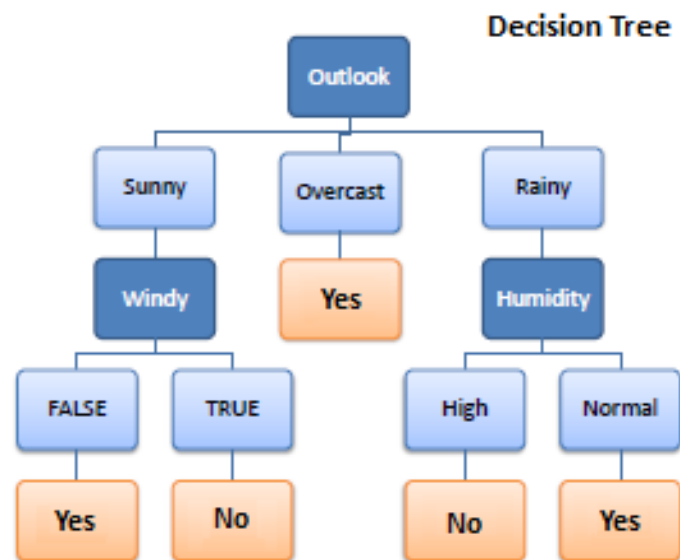
# Decision Tree

Misalnya kita memiliki data seperti di atas. Data berisi informasi mengenai kondisi cuaca pada hari tertentu dan apakah cocok untuk bermain golf di kondisi cuaca tersebut.

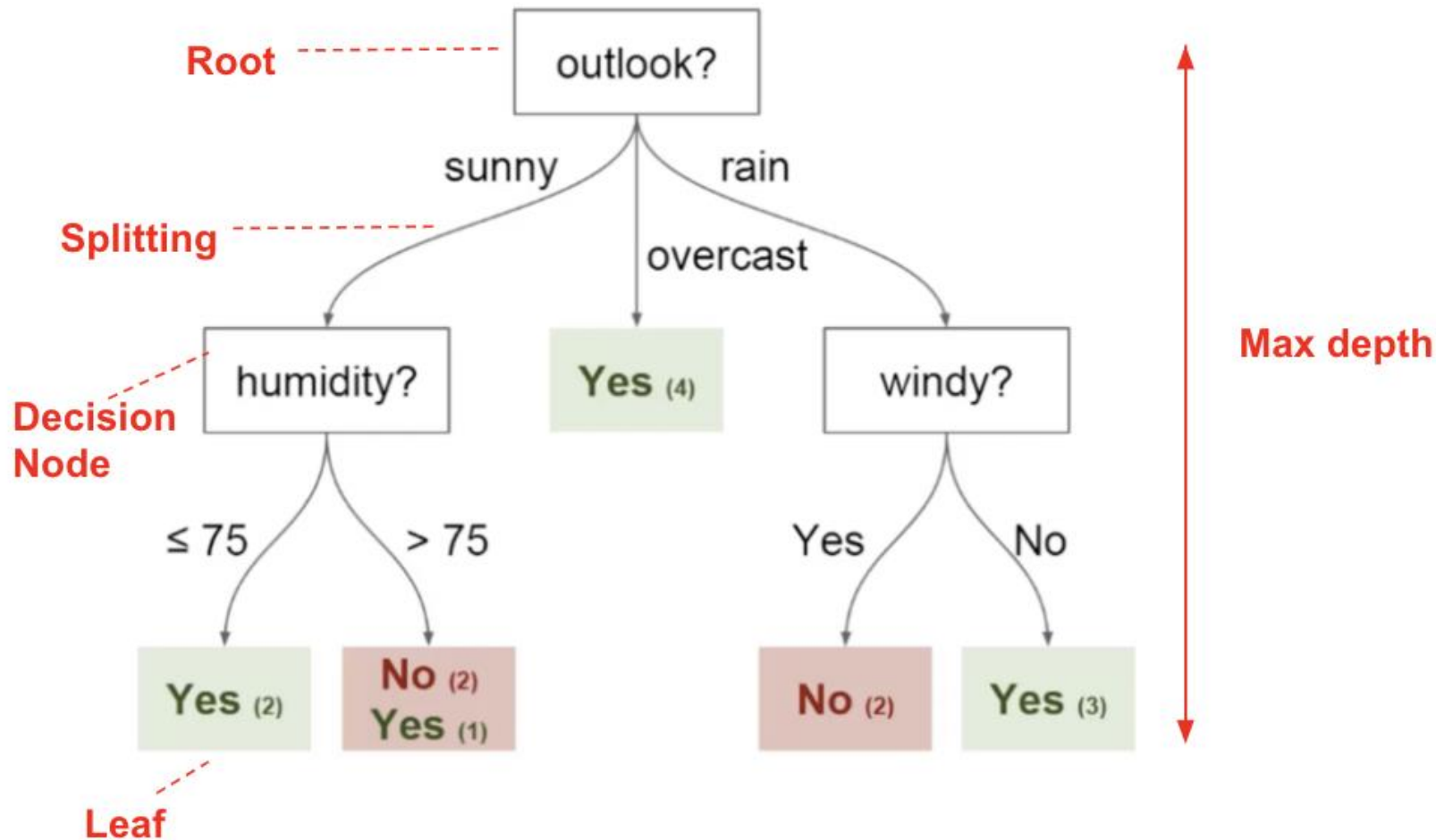
Sebuah pohon keputusan dapat dibuat dari data sebelumnya. Perhatikan contoh pohon keputusan di bawah. Pohon ini menggunakan beberapa atribut, diantaranya adalah kondisi langit dan kecepatan angin untuk menentukan bermain golf atau tidak.



Predictors				Target
Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

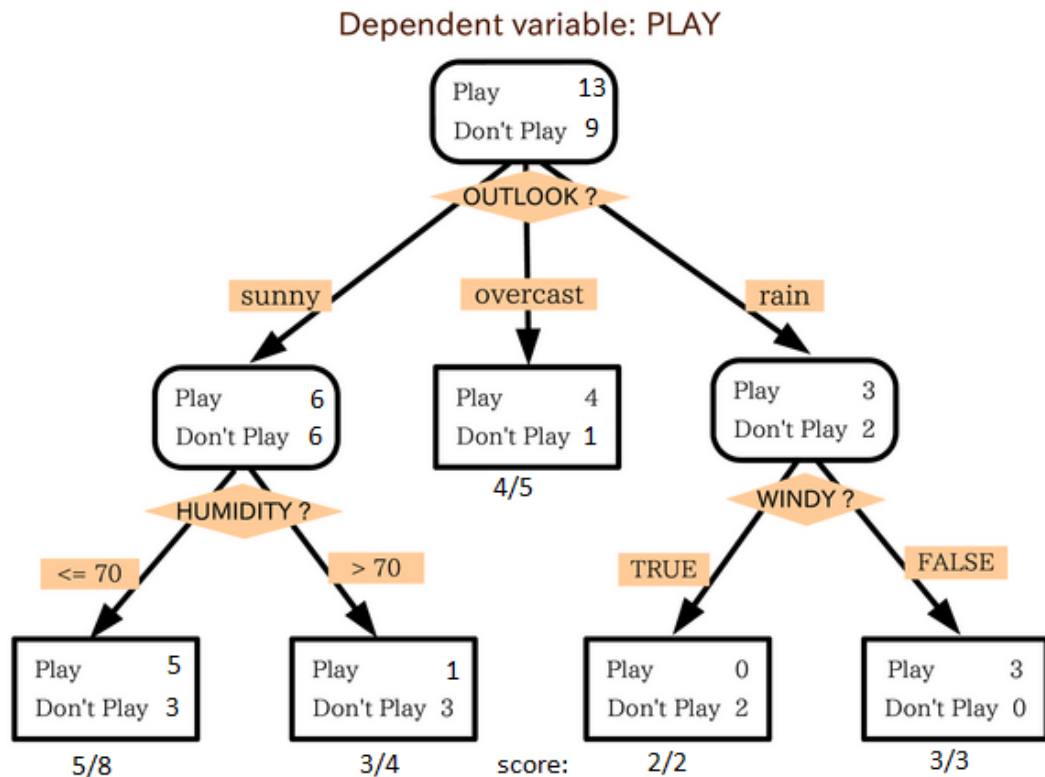


# Decision Tree Diagram





# Decision Tree





# Decision Tree Algorithm

- a. [ID3](#) (Iterative Dichotomiser 3)
- b. [C4.5](#) (successor of ID3)
- c. [CART](#) (Classification And Regression Tree)<sup>[6]</sup>
- d. [Chi-square automatic interaction detection](#) (CHAID). Performs multi-level splits when computing classification trees.<sup>[14]</sup>
- e. [MARS](#): extends decision trees to handle numerical data better.
- f. Conditional Inference Trees. Statistics-based approach that uses non-parametric tests as splitting criteria, corrected for multiple testing to avoid overfitting. This approach results in unbiased predictor selection and does not require pruning.<sup>[15][16]</sup>



<b>Training Algorithm</b>	<b>CART</b>  (Classification and Regression Trees)
<b>Target(s)</b>	Classification and Regression
<b>Metric</b>	Gini Index
<b>Cost function</b> (Based on what to split?)	Select its splits to achieve the subsets that minimize Gini Impurity

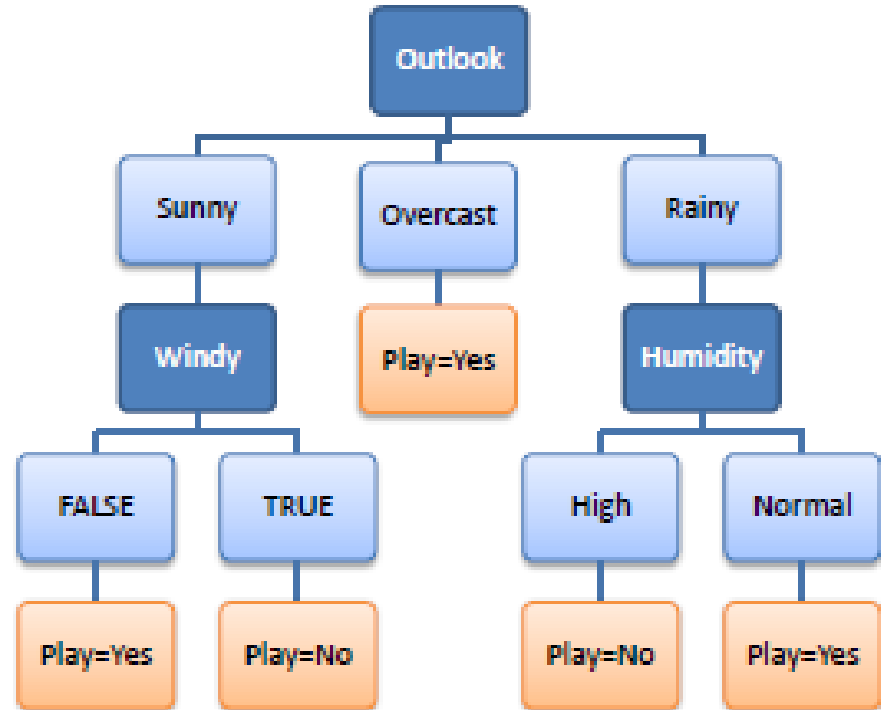
$R_1$ : IF (Outlook=Sunny) AND  
(Windy=FALSE) THEN Play=Yes

$R_2$ : IF (Outlook=Sunny) AND  
(Windy=TRUE) THEN Play=No

$R_3$ : IF (Outlook=Overcast) THEN  
Play=Yes

$R_4$ : IF (Outlook=Rainy) AND  
(Humidity=High) THEN Play=No

$R_5$ : IF (Outlook=Rain) AND  
(Humidity=Normal) THEN  
Play=Yes





## Asset

[https://drive.google.com/drive/folders/168bLU4P9kNAtIA\\_PVTLPSADz1POoll3o?usp=sharing](https://drive.google.com/drive/folders/168bLU4P9kNAtIA_PVTLPSADz1POoll3o?usp=sharing)



# Referensi

<https://medium.com/analytics-vidhya/decision-tree-algorithm-explained-bd6b7b22eab9>

[https://www.saedsayad.com/decision\\_tree.htm](https://www.saedsayad.com/decision_tree.htm)

<https://dinhanhthi.com/decision-tree-classifier/>

<https://www.kdnuggets.com/2020/02/decision-tree-intuition.html>

<https://rapidminer.com/>



# Terima Kasih

See You Next Time