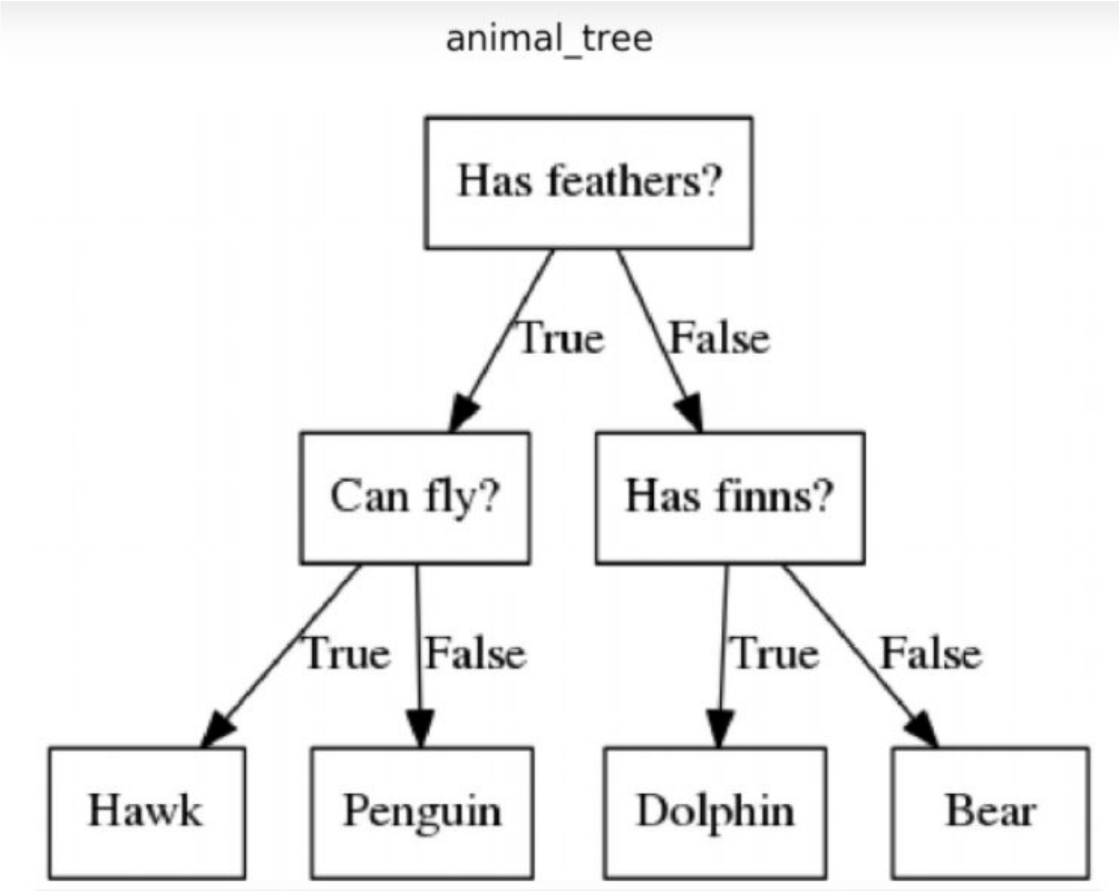
Decision Tree

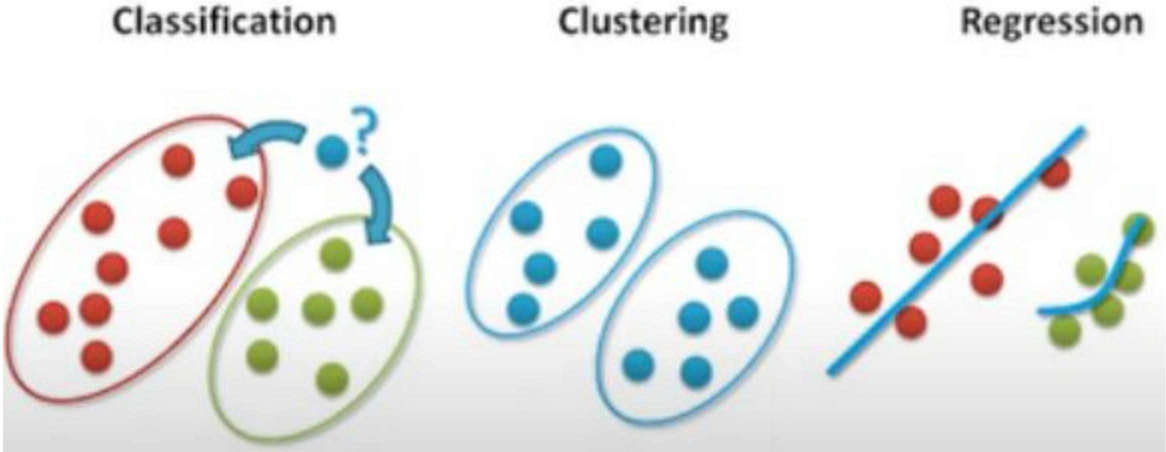
CART - Classification And Regression Tree

Example of Decision Tree

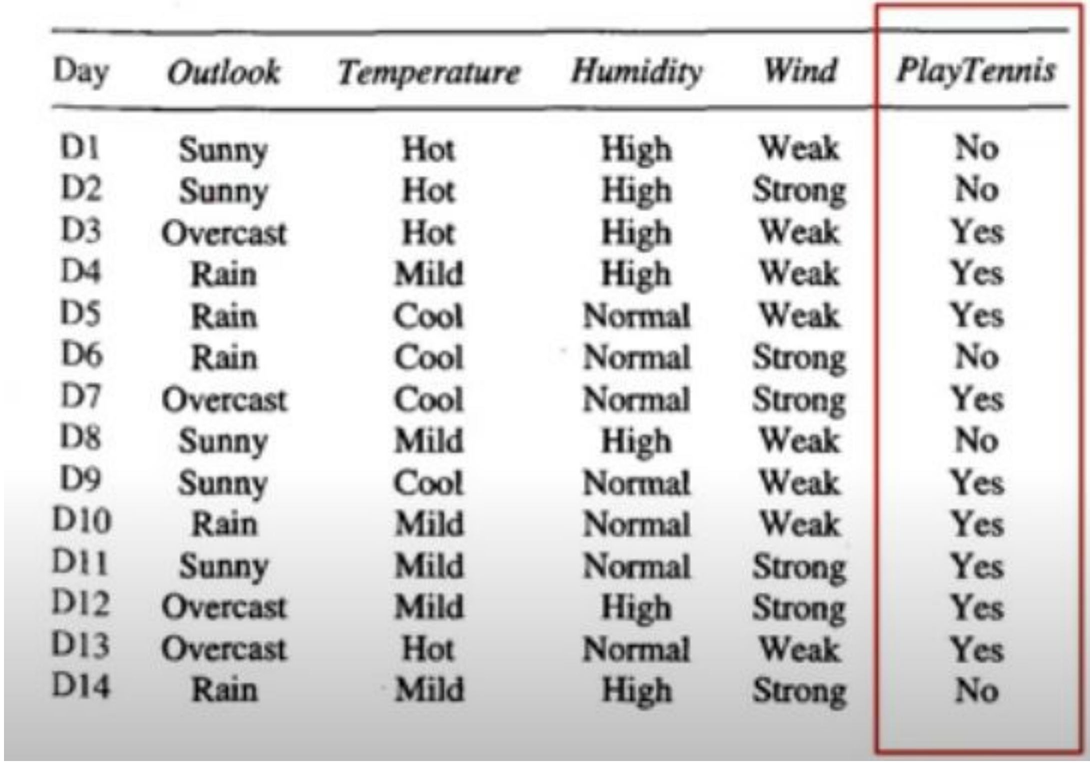


A **decision tree** is one of the supervised **machine learning algorithms**.

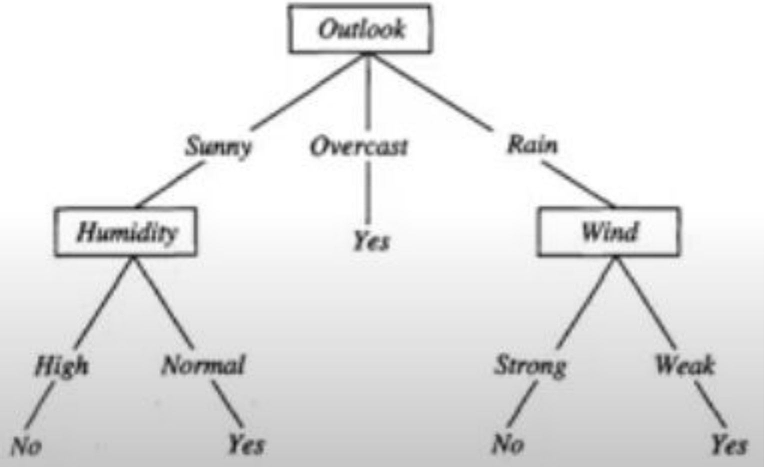
A decision tree follows a set of if-else conditions to visualize the data and classify it according to the conditions.


Classification
?
Clustering
Regression

Decision Tree- Representation



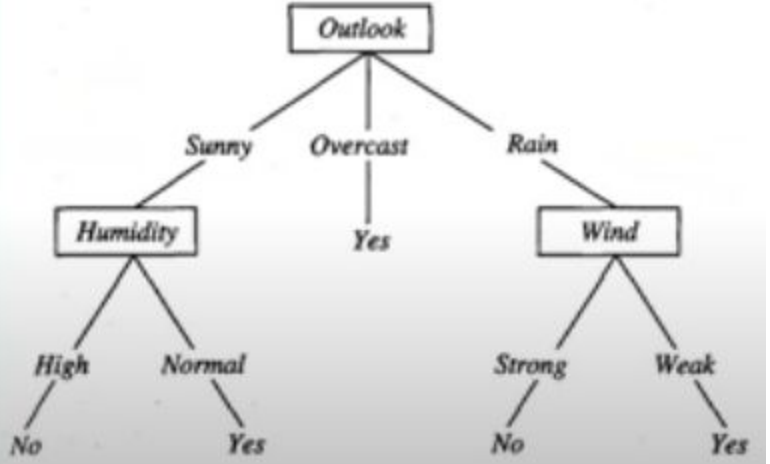
Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

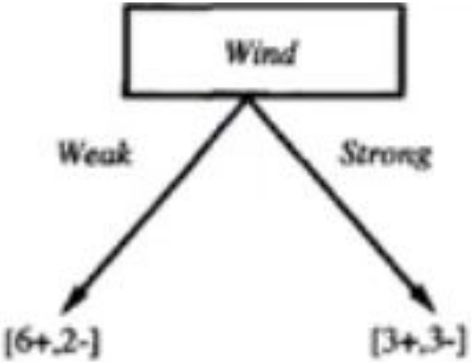


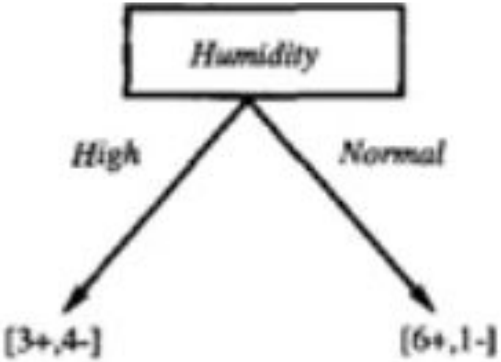
Decision Tree- Representation

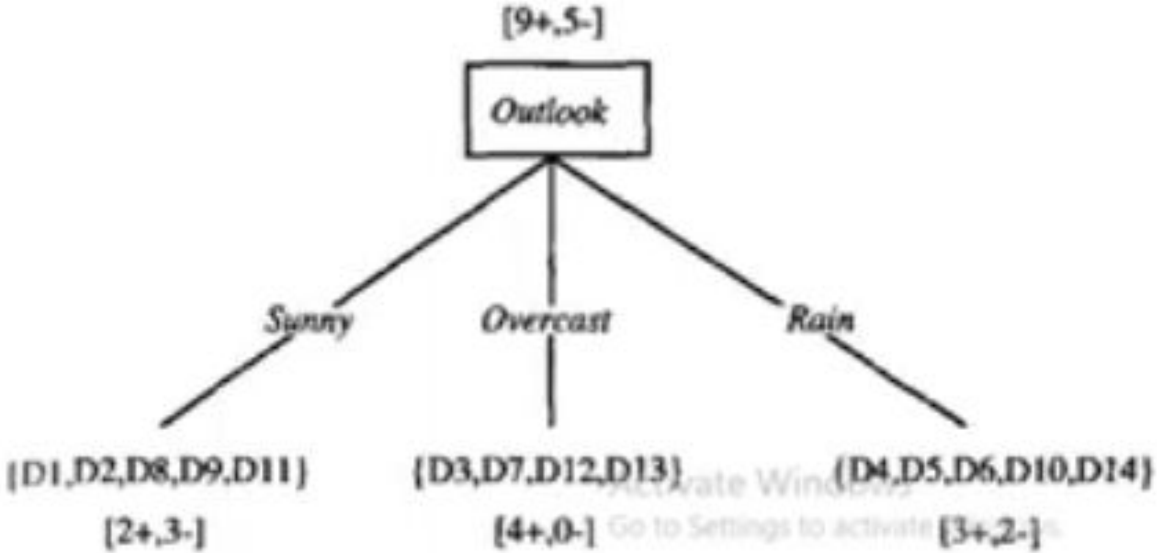
Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Consider instance-
(*Outlook*= Sunny, *Temperature*= Hot, *Humidity*= High, *Wind*= Strong)
Prediction- *PlayTennis*= No




Wind
Weak
Strong
[6+,2-]
[3+,3-]


Humidity
High
Normal
[3+,4-]
[6+,1-]


[9+,5-]
Outlook
Sunny
Overcast
Rain
{D1,D2,D8,D9,D11}
[2+,3-]
{D3,D7,D12,D13}
[4+,0-]
{D4,D5,D6,D10,D14}
[3+,2-]

The Gini impurity index

Measuring the diversity of a dataset

Which set is more diverse?

Gini = 0.42

■■■■■■■●●●

■	■	Same
■	●	**Different**
●	■	**Different**
■	■	Same
●	●	Same
■	●	**Different**
●	●	Same
■	■	Same
●	■	**Different**
■	■	Same

Different:
4 out of 10

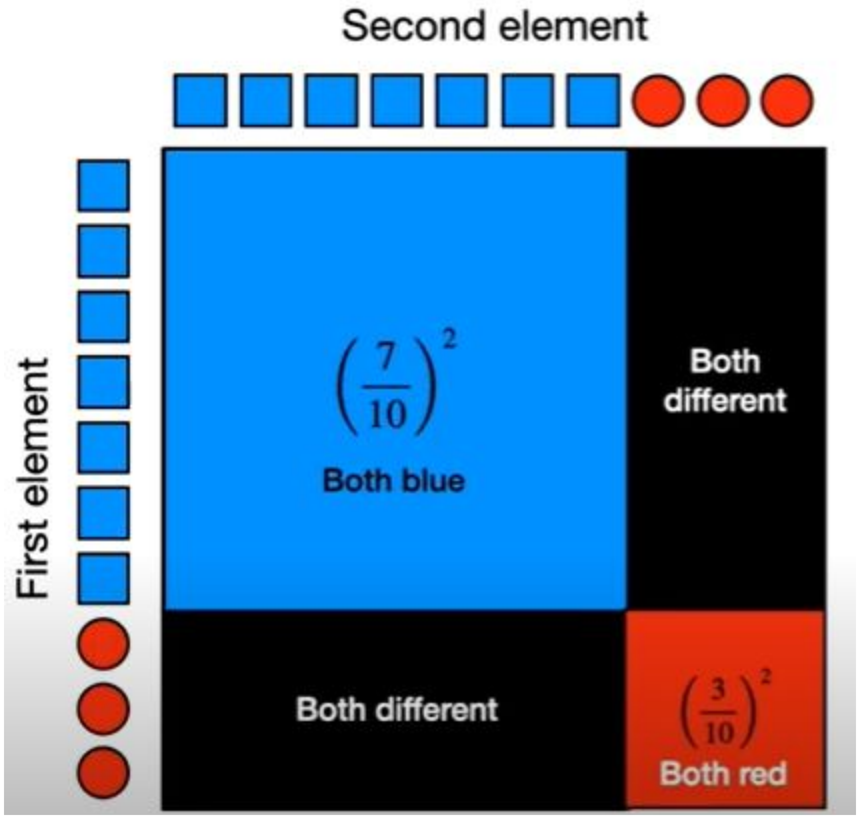
Gini = 0.7

More diverse

■■■■●●●▲▲☆

●	▲	**Different**
■	■	Same
▲	■	**Different**
☆	●	**Different**
■	▲	**Different**
■	■	Same
●	●	Same
▲	●	**Different**
■	☆	**Different**
●	■	**Different**

Different:
7 out of 10


Second element
First element
$\left(\frac{7}{10}\right)^2$
Both blue
Both different
Both different
$\left(\frac{3}{10}\right)^2$
Both red

CART- Gini Index

L. Breiman, J. Friedman, R. Olshen and C. Stone in 1984 proposed an algorithm to build a binary decision tree also called CART decision tree.

in CART, for each node only two children are created.

CART uses Gini index as a measure to select the best attribute to be splitted, It is also known as Gini Index of Diversity and is denote as γ.

Gini Index

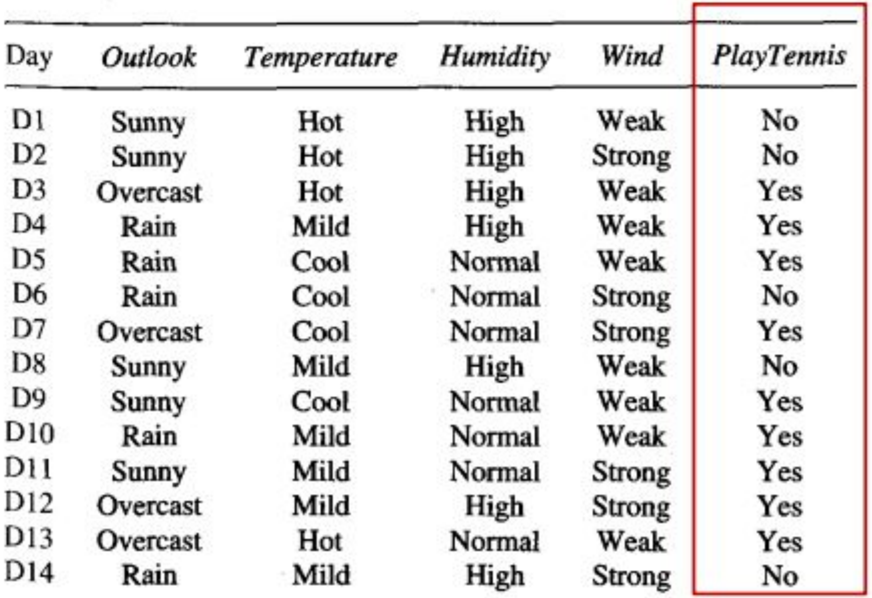
$$\boldsymbol{G(D) = 1 - \sum_{i=1}^{k} p_i^2}$$

Gini index

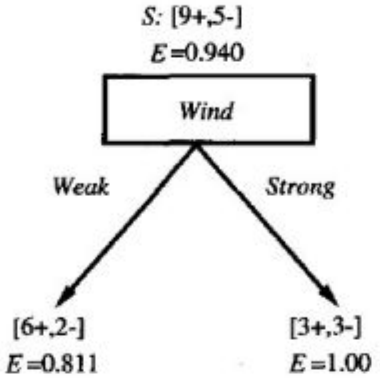
The measure of the degree of probability of a particular variable being wrongly classified when it is randomly chosen is called the Gini index or Gini impurity. The data is equally distributed based on the Gini index.

$$\text{Gini} = 1 - \sum_{i=1}^{n} (p_i)^2$$

Example-



Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



Gini Index $G(D) = 1 - \sum_{i=1}^{k} p_i^2$

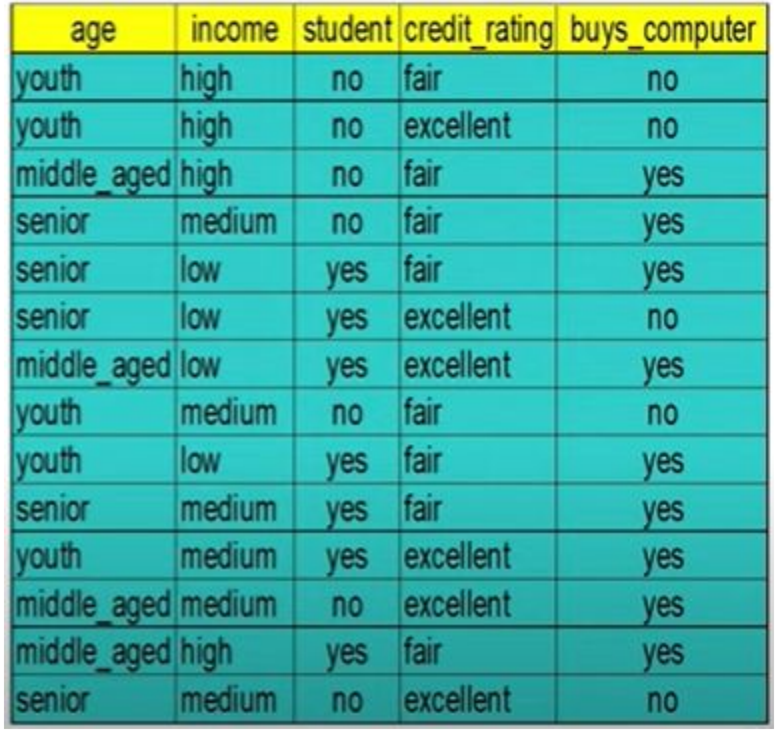
$$= 1 - \left[\left(\frac{9}{14}\right)^2 + \left(\frac{5}{14}\right)^2\right]$$

$$= 0.4592$$

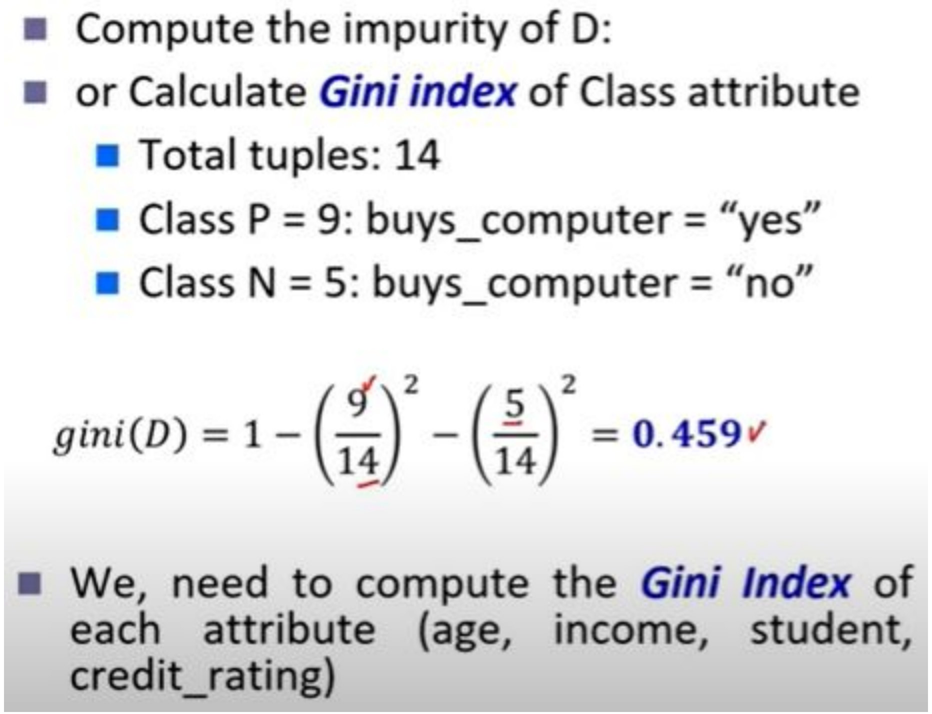
$$\boxed{Gini(D) = 1 - \sum_{i=1}^{m} p_i^2,}$$

$$gini_{A}(D) = \frac{|D_1|}{|D|} gini(D_1) + \frac{|D_2|}{|D|} gini(D_2)$$

$$\Delta gini(A) = gini(D) - gini_{A}(D)$$



age	income	student	credit_rating	buys_computer
youth	high	no	fair	no
youth	high	no	excellent	no
middle_aged	high	no	fair	yes
senior	medium	no	fair	yes
senior	low	yes	fair	yes
senior	low	yes	excellent	no
middle_aged	low	yes	excellent	yes
youth	medium	no	fair	no
youth	low	yes	fair	yes
senior	medium	yes	fair	yes
youth	medium	yes	excellent	yes
middle_aged	medium	no	excellent	yes
middle_aged	high	yes	fair	yes
senior	medium	no	excellent	no



- Compute the impurity of D:
- or Calculate ***Gini index*** of Class attribute
  - Total tuples: 14
  - Class P = 9: buys_computer = “yes”
  - Class N = 5: buys_computer = “no”

$$gini(D) = 1 - \left(\frac{9}{\underline{14}}\right)^2 - \left(\frac{\underline{5}}{14}\right)^2 = \mathbf{0.459} \checkmark$$

- We, need to compute the ***Gini Index*** of each attribute (age, income, student, credit_rating)

- Lets now consider: **credit_rating**
  - It is a binary attribute

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2,$$

$$gini_A(D) = \frac{|D_1|}{|D|} gini(D_1) + \frac{|D_2|}{|D|} gini(D_2)$$

$$\Delta gini(A) = gini(D) - gini_A(D)$$

$$\boldsymbol{gini}_{credit-rating}(D) = \left(\frac{D_1}{14}\right) gini(D_1) + \left(\frac{D_2}{14}\right) gini(D_2)$$

$$= \frac{8}{14}\left(1 - \left(\frac{6}{8}\right)^2 - \left(\frac{2}{8}\right)^2\right) + \frac{6}{14}\left(1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2\right) = \mathbf{0.4285}$$

		Class		
		yes	no	
credit_rating	fair	6	2	8
	excellent	3	3	6
				14

age	income	student	credit_rating	buys_computer
youth	high	no	fair	no
youth	high	no	excellent	no
middle_aged	high	no	fair	yes
senior	medium	no	fair	yes
senior	low	yes	fair	yes
senior	low	yes	excellent	no
middle_aged	low	yes	excellent	yes
youth	medium	no	fair	no
youth	low	yes	fair	yes
senior	medium	yes	fair	yes
youth	medium	yes	excellent	yes
middle_aged	medium	no	excellent	yes
middle_aged	high	yes	fair	yes
senior	medium	no	excellent	no

Gini Index [CART] - Example

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2,$$

$$gini_A(D) = \frac{|D_1|}{|D|} gini(D_1) + \frac{|D_2|}{|D|} gini(D_2)$$

$$\Delta gini(A) = gini(D) - gini_A(D)$$

- Lets now consider: **student**
  - It is a binary attribute

		Class		
		yes	no	
student	yes	6	1	7
	no	3	4	7
				14

$$\boldsymbol{gini_{student}}(D) = \left(\frac{D_1}{14}\right) gini(D_1) + \left(\frac{D_2}{14}\right) gini(D_2)$$

$$= \frac{7}{14}\left(1 - \left(\frac{6}{7}\right)^2 - \left(\frac{1}{7}\right)^2\right) + \frac{7}{14}\left(1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2\right) = \mathbf{0.3673}$$

age	income	student	credit_rating	buys_computer
youth	high	no	fair	no
youth	high	no	excellent	no
middle_aged	high	no	fair	yes
senior	medium	no	fair	yes
senior	low	yes	fair	yes
senior	low	yes	excellent	no
middle_aged	low	yes	excellent	yes
youth	medium	no	fair	no
youth	low	yes	fair	yes
senior	medium	yes	fair	yes
youth	medium	yes	excellent	yes
middle_aged	medium	no	excellent	yes
middle_aged	high	yes	fair	yes
senior	medium	no	excellent	no

Gini Index [CART] - Example

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2,$$

$$gini_A(D) = \frac{|D_1|}{|D|} gini(D_1) + \frac{|D_2|}{|D|} gini(D_2)$$

$$\Delta gini(A) = gini(D) - gini_A(D)$$

- Lets now consider: **age:** {youth, middle_aged, senior}
  - Now consider each possible splitting subsets

{{youth, middle_aged}, {youth, senior}, {middle_aged, senior}, {youth},{middle_aged}, {senior}}

$$gini_{age \in \{youth, middle\_aged\}}(D) = \left(\frac{D_1}{14}\right) gini(D_1) + \left(\frac{D_2}{14}\right) gini(D_2)$$

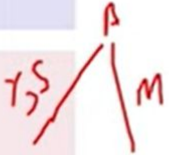
$$= \frac{9}{14}\left(1 - \left(\frac{6}{9}\right)^2 - \left(\frac{3}{9}\right)^2\right) + \frac{5}{14}\left(1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2\right) = \mathbf{0.4571}$$

$$= gini_{age \in \{senior\}}(D)$$

$$gini_{age \in \{youth, senior\}}(D) = \left(\frac{D_1}{14}\right) gini(D_1) + \left(\frac{D_2}{14}\right) gini(D_2)$$

$$= \frac{10}{14}\left(1 - \left(\frac{5}{10}\right)^2 - \left(\frac{5}{10}\right)^2\right) + \frac{4}{14}\left(1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2\right) = \mathbf{0.3571}$$

$$= gini_{age \in \{middle\_aged\}}(D)$$



$$gini_{age \in \{middle\_aged, senior\}}(D) = \left(\frac{D_1}{14}\right) gini(D_1) + \left(\frac{D_2}{14}\right) gini(D_2)$$

$$= \frac{9}{14}\left(1 - \left(\frac{7}{9}\right)^2 - \left(\frac{2}{9}\right)^2\right) + \frac{5}{14}\left(1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2\right) = \mathbf{0.3936} = gini_{age \in \{youth\}}(D)$$

age	income	student	credit_rating	buys_computer
youth	high	no	fair	no
youth	high	no	excellent	no
middle_aged	high	no	fair	yes
senior	medium	no	fair	yes
senior	low	yes	fair	yes
senior	low	yes	excellent	no
middle_aged	low	yes	excellent	yes
youth	medium	no	fair	no
youth	low	yes	fair	yes
senior	medium	yes	fair	yes
youth	medium	yes	excellent	yes
middle_aged	medium	no	excellent	yes
middle_aged	high	yes	fair	yes
senior	medium	no	excellent	no

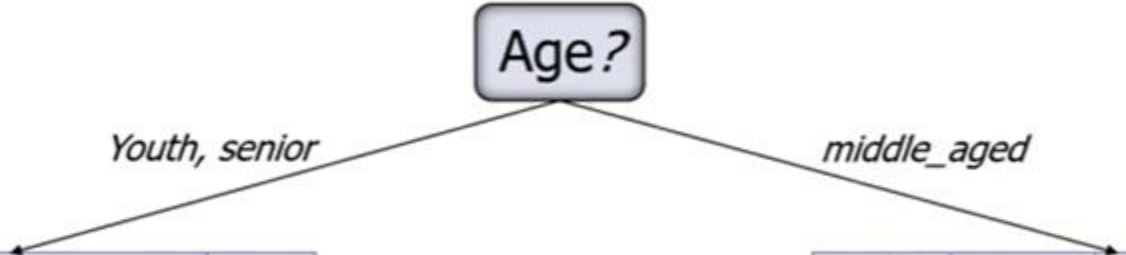
		Class		
		yes	no	
age	youth	2	3	5
	middle_aged	4	0	4
	senior	3	2	5
				14

Gini Index [CART] - Example

- Best binary split for age is {youth, senior} or {middle_aged} with minimum Gini index.
- And best binary split for income is {medium, high} or {low} with minimum Gini index.

Attribute	Split	Gini index	Reduction in impurity $\Delta gini = gini(D) - gini_A(D)$
age	{youth, senior} or {middle_aged}	0.3571	0.459 - 0.3571 = 0.1019
income	{medium, high} or {low}	0.4428	0.459 - 0.4428 = 0.0162
student	Binary	0.3673	0.459 - 0.3673 = 0.0917
credit_rating	Binary	0.4285	0.459 - 0.4285 = 0.0305

Gini Index [CART] - Example



Age?

Youth, senior

middle_aged

income	student	credit_rating	class
high	no	Fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	Yes
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

income	student	credit_rating	class
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes