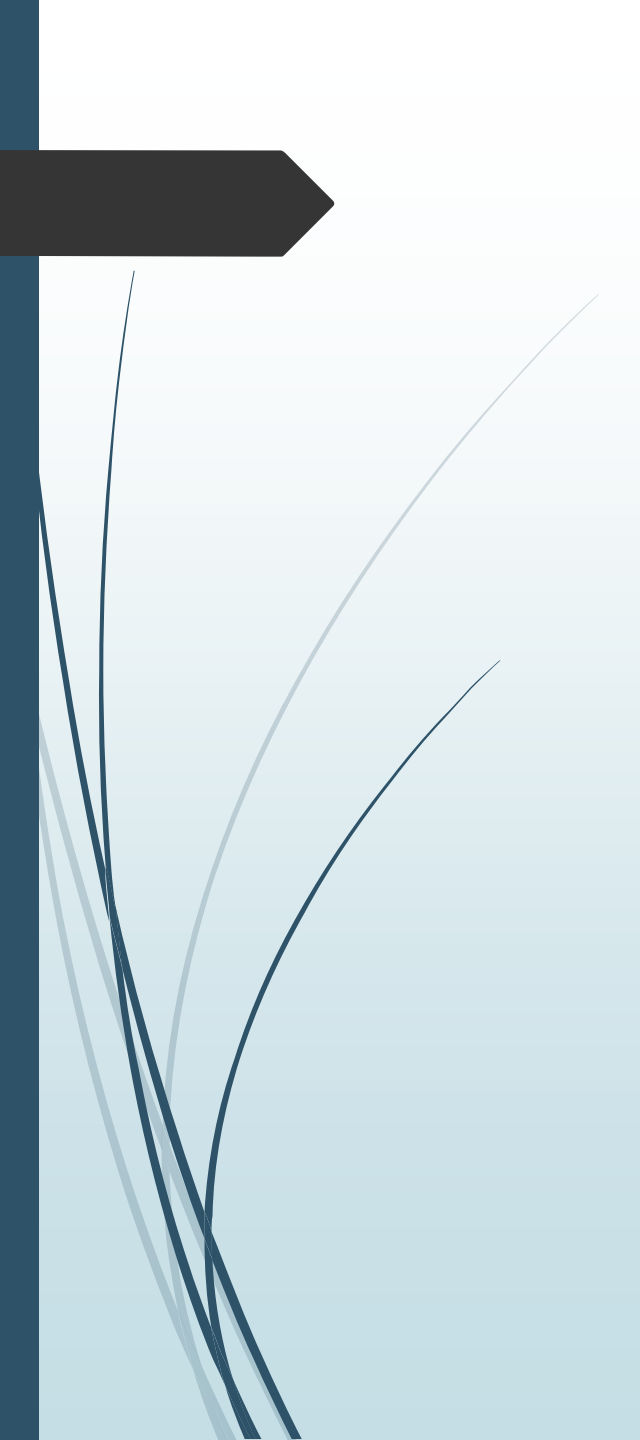


C4.5 Algorithm in Decision Tree



Atefeh Khosravani



بِه نام خدا

الگوریتم C4.5

- این الگوریتم یکی از تعمیم های الگوریتم ID3 است که از معیار نسبت بهره (Gain Ratio) استفاده می کند.
- الگوریتم هنگامی متوقف می شود که تعداد نمونه ها کمتر از مقدار مشخص شده ای باشد.
- از تکنیک پس هرس استفاده می کند و داده های عددی را نیز می پذیرد.

از نقاط ضعف الگوریتم ID3 که در C4.5 رفع شده است می‌توان به موارد زیر اشاره کرد:

- الگوریتم C4.5 می‌تواند مقادیر گسسته یا پیوسته را در ویژگی‌ها درک کند.
- الگوریتم C4.5 قادر است با وجود مقادیر گمشته نیز درخت تصمیم (decision tree) خود را بسازد، در حالی که الگوریتمی مانند ID3 و بسیاری دیگر از الگوریتم‌های طبقه‌بندی نمی‌توانند با وجود مقادیر گمشته، مدل خود را بسازند.
- الگوریتم C4.5 این قابلیت را دارد که وزن‌های مختلف و غیر یکسانی را به برخی از ویژگی‌ها بدهد.

از نقاط ضعف الگوریتم ID3 که در C4.5 رفع شده است می‌توان به موارد زیر اشاره کرد:

- چهارمین موردی که باعث بهینه شدن الگوریتم C4.5 نسبت به ID3 می‌شود، عملیات هرس کردن جهت جلوگیری از بیش برآزش (Overfitting) می‌باشد. الگوریتم‌هایی مانند ID3 به خاطر اینکه سعی دارند تا حد امکان شاخه و برگ داشته باشند (تا به نتیجه مورد نظر برسند) با احتمال بالاتری دارای پیچیدگی در ساخت مدل و این پیچیدگی در بسیاری از موارد الگوریتم را دچار بیش برآزش و خطای بالا می‌کند.
- اما با عملیات هرس کردن درخت که در الگوریتم C4.5 انجام می‌شود، می‌توان مدل را به یک نقطه بهینه رساند که زیاد پیچیده نباشد (و البته زیاد هم ساده نباشد) و بیش برآزش یا کم برآزش (Underfitting) رخ ندهد.

مثال:

می خواهیم یک جدول تصمیم برای مجموعه داده زیر که همان مثال قبلی بازی تنیس می باشد ایجاد کنیم. تنها تفاوت در این است که ستون های دما و رطوبت به جای مقادیر اسمی دارای مقادیر پیوسته هستند.



Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	85	85	Weak	No
2	Sunny	80	90	Strong	No
3	Overcast	83	78	Weak	Yes
4	Rain	70	96	Weak	Yes
5	Rain	68	80	Weak	Yes
6	Rain	65	70	Strong	No
7	Overcast	64	65	Strong	Yes
8	Sunny	72	95	Weak	No
9	Sunny	69	70	Weak	Yes
10	Rain	75	80	Weak	Yes
11	Sunny	75	70	Strong	Yes
12	Overcast	72	90	Strong	Yes
13	Overcast	81	75	Weak	Yes
14	Rain	71	80	Strong	No

روش:

- همان روند ID3 را به کار میگیریم که ابتدا آنтроپی کلی را محاسبه میکنیم.
- در الگوریتم ID3، بهره (Gain) را برای هر ویژگی محاسبه میکردیم اما در اینجا، به جای سود، نسبت بهره (Gain Ratio) را محاسبه میکنیم.

$$\text{Entropy(Decision)} = \sum - p(I) \cdot \log_2 p(I) = - p(\text{Yes}) \cdot \log_2 p(\text{Yes}) - p(\text{No}) \cdot \log_2 p(\text{No}) = - (9/14) \cdot \log_2(9/14) - (5/14) \cdot \log_2(5/14) = 0.940$$

$$\text{GainRatio}(A) = \text{Gain}(A) / \text{SplitInfo}(A)$$

$$\text{SplitInfo}(A) = - \sum |D_j|/|D| \times \log_2 |D_j|/|D|$$

بررسی ویژگی باد (Wind)

➡ ابتدا ویژگی باد را که یک ویژگی اسمی است، بررسی میکنیم. مقادیر احتمالی آن ضعیف (weak) و قوی (strong) هستند.

$$\text{Gain}(\text{Decision}, \text{Wind}) = \text{Entropy}(\text{Decision}) - \sum (p(\text{Decision}|\text{Wind}) . \text{Entropy}(\text{Decision}|\text{Wind}))$$

$$\text{Gain}(\text{Decision}, \text{Wind}) = \text{Entropy}(\text{Decision}) - [p(\text{Decision}|\text{Wind}=\text{Weak}) . \text{Entropy}(\text{Decision}|\text{Wind}=\text{Weak})] + [p(\text{Decision}|\text{Wind}=\text{Strong}) . \text{Entropy}(\text{Decision}|\text{Wind}=\text{Strong})]$$

8 مورد باد ضعیف (weak) وجود دارد. 2 تای آنها خیر، 6 تای آنها بله نتیجه گیری شده است.

$$\text{Entropy}(\text{Decision}|\text{Wind}=\text{Weak}) = - p(\text{No}) . \log_2 p(\text{No}) - p(\text{Yes}) . \log_2 p(\text{Yes}) = - (2/8) . \log_2 (2/8) - (6/8) . \log_2 (6/8) = 0.811$$

$$\text{Entropy}(\text{Decision}|\text{Wind}=\text{Strong}) = - (3/6) . \log_2 (3/6) - (3/6) . \log_2 (3/6) = 1$$

$$\text{Gain}(\text{Decision}, \text{Wind}) = 0.940 - (8/14).(0.811) - (6/14).(1) = 0.940 - 0.463 - 0.428 = 0.049$$

8 تصمیم برای باد ضعیف و 6 تصمیم برای باد قوی وجود دارد.

$$\text{SplitInfo}(\text{Decision}, \text{Wind}) = -(8/14).\log_2(8/14) - (6/14).\log_2(6/14) = 0.461 + 0.524 = 0.985$$

$$\text{GainRatio}(\text{Decision}, \text{Wind}) = \text{Gain}(\text{Decision}, \text{Wind}) / \text{SplitInfo}(\text{Decision}, \text{Wind}) = 0.049 / 0.985 = 0.049$$

بررسی ویژگی Outlook

➤ Outlook نیز یک ویژگی اسمی است. مقادیر احتمالی آن آفتابی (sunny)، ابری (overcast) و بارانی (rain) است.

$$\text{Gain}(\text{Decision}, \text{Outlook}) = \text{Entropy}(\text{Decision}) - \sum (p(\text{Decision}|\text{Outlook}) \cdot \text{Entropy}(\text{Decision}|\text{Outlook})) =$$

$$\begin{aligned} \text{Gain}(\text{Decision}, \text{Outlook}) &= \text{Entropy}(\text{Decision}) - p(\text{Decision}|\text{Outlook}=\text{Sunny}) \cdot \\ &\text{Entropy}(\text{Decision}|\text{Outlook}=\text{Sunny}) - p(\text{Decision}|\text{Outlook}=\text{Overcast}) \cdot \\ &\text{Entropy}(\text{Decision}|\text{Outlook}=\text{Overcast}) - p(\text{Decision}|\text{Outlook}=\text{Rain}) \cdot \\ &\text{Entropy}(\text{Decision}|\text{Outlook}=\text{Rain}) \end{aligned}$$

5 مورد آفتابی وجود دارد. 3 تای آنها خیر، 2 تای آنها بله نتیجه گیری شده است.

$$\begin{aligned} \text{Entropy}(\text{Decision}|\text{Outlook}=\text{Sunny}) &= -p(\text{No}) \cdot \log_2 p(\text{No}) - p(\text{Yes}) \cdot \log_2 p(\text{Yes}) = -(3/5) \cdot \log_2(3/5) \\ &- (2/5) \cdot \log_2(2/5) = 0.441 + 0.528 = 0.970 \end{aligned}$$

$$\begin{aligned} \text{Entropy}(\text{Decision}|\text{Outlook}=\text{Overcast}) &= -p(\text{No}) \cdot \log_2 p(\text{No}) - p(\text{Yes}) \cdot \log_2 p(\text{Yes}) = - \\ &(0/4) \cdot \log_2(0/4) - (4/4) \cdot \log_2(4/4) = 0 \end{aligned}$$

$$\begin{aligned} \text{Entropy}(\text{Decision}|\text{Outlook}=\text{Rain}) &= -p(\text{No}) \cdot \log_2 p(\text{No}) - p(\text{Yes}) \cdot \log_2 p(\text{Yes}) = -(2/5) \cdot \log_2(2/5) - \\ &(3/5) \cdot \log_2(3/5) = 0.528 + 0.441 = 0.970 \end{aligned}$$

$$\text{Gain}(\text{Decision}, \text{Outlook}) = 0.940 - (5/14) \cdot (0.970) - (4/14) \cdot (0) - (5/14) \cdot (0.970) - (5/14) \cdot (0.970) = 0.246$$

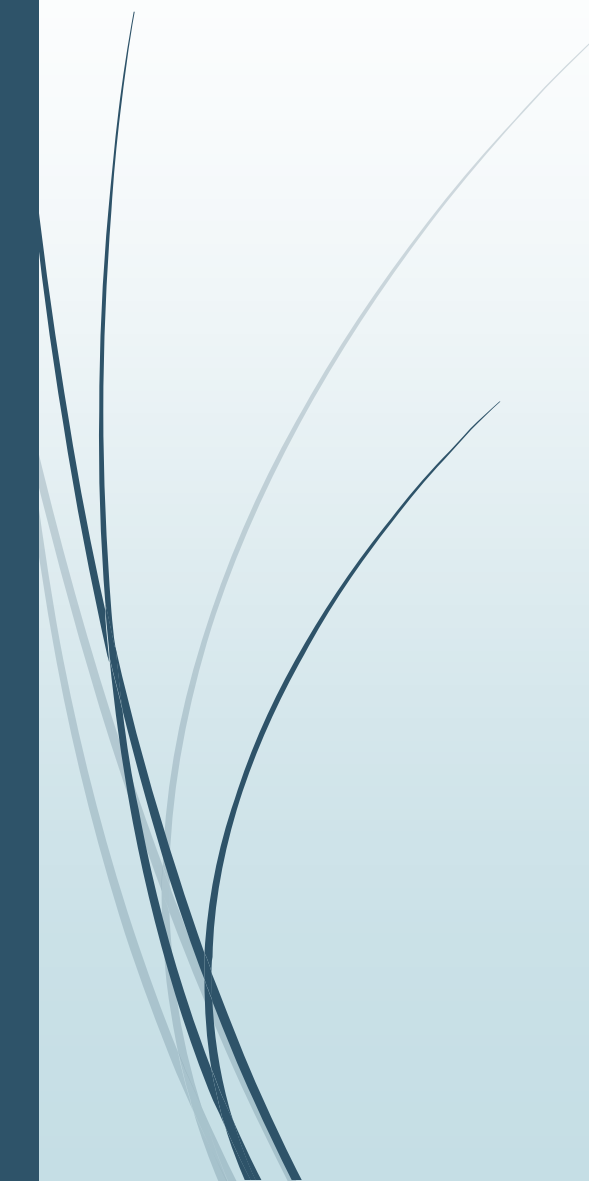
5 مورد آفتابی، 4 مورد ابری و 5 مورد باران وجود دارد.

$$\text{SplitInfo}(\text{Decision}, \text{Outlook}) = -(5/14) \cdot \log_2(5/14) - (4/14) \cdot \log_2(4/14) - (5/14) \cdot \log_2(5/14) = 1.577$$

$$\begin{aligned} \text{GainRatio}(\text{Decision}, \text{Outlook}) &= \text{Gain}(\text{Decision}, \text{Outlook}) / \text{SplitInfo}(\text{Decision}, \text{Outlook}) = \\ &0.246 / 1.577 = 0.155 \end{aligned}$$

بررسی ویژگی رطوبت (Humidity)

- به عنوان یک استثنا، رطوبت یک ویژگی پیوسته است. باید مقادیر پیوسته را به اسمی تبدیل کنیم. C4.5 انجام تقسیم باینری را بر اساس مقدار آستانه (threshold) پیشنهاد می کند. آستانه باید مقداری باشد که حداکثر بهره را برای آن ویژگی ارائه دهد.
- برای بررسی این ویژگی ابتدا باید مقادیر رطوبت را از کوچکترین به بزرگترین مرتب کنیم.




Day	Humidity	Decision
7	65	Yes
6	70	No
9	70	Yes
11	70	Yes
13	75	Yes
3	78	Yes
5	80	Yes
10	80	Yes
14	80	No
1	85	No
2	90	No
12	90	Yes
8	95	No
4	96	Yes

بررسی ویژگی رطوبت (Humidity)

■ اکنون، باید همه مقادیر رطوبت را تکرار کنیم و مجموعه داده‌ها را به دو قسمت به عنوان نمونه‌های کمتر یا مساوی با مقدار فعلی و نمونه‌های بزرگتر از مقدار فعلی جدا کنیم. نسبت بهره یا بهره را برای هر مرحله محاسبه می‌کنیم. مقداری که بهره را به حداکثر می‌رساند آستانه خواهد بود.

■ 65 را به عنوان آستانه رطوبت بررسی می‌کنیم.


$$\text{Entropy}(\text{Decision}|\text{Humidity} \leq 65) = -p(\text{No}) \cdot \log_2 p(\text{No}) - p(\text{Yes}) \cdot \log_2 p(\text{Yes}) = -(0/1) \cdot \log_2(0/1) - (1/1) \cdot \log_2(1/1) = 0$$

$$\text{Entropy}(\text{Decision}|\text{Humidity} > 65) = -(5/13) \cdot \log_2(5/13) - (8/13) \cdot \log_2(8/13) = 0.530 + 0.431 = 0.961$$

$$\text{Gain}(\text{Decision}, \text{Humidity} <> 65) = 0.940 - (1/14) \cdot 0 - (13/14) \cdot (0.961) = 0.048$$

* عبارت فوق به این اشاره دارد که چه شاخه درخت تصمیم برای کمتر یا مساوی 65 و بزرگتر از 65 خواهد بود. به این اشاره نمی شود که رطوبت برابر با 65 نیست!

$$\text{SplitInfo}(\text{Decision}, \text{Humidity} <> 65) = -(1/14) \cdot \log_2(1/14) - (13/14) \cdot \log_2(13/14) = 0.371$$

$$\text{GainRatio}(\text{Decision}, \text{Humidity} <> 65) = 0.126$$

70 را به عنوان آستانه رطوبت بررسی می کنیم.

$$\text{Entropy}(\text{Decision}|\text{Humidity} \leq 70) = - (1/4) \cdot \log_2(1/4) - (3/4) \cdot \log_2(3/4) = 0.811$$

$$\text{Entropy}(\text{Decision}|\text{Humidity} > 70) = - (4/10) \cdot \log_2(4/10) - (6/10) \cdot \log_2(6/10) = 0.970$$

$$\begin{aligned} \text{Gain}(\text{Decision}, \text{Humidity} <> 70) &= 0.940 - (4/14) \cdot (0.811) - (10/14) \cdot (0.970) = 0.940 - 0.231 - 0.692 \\ &= 0.014 \end{aligned}$$

$$\text{SplitInfo}(\text{Decision}, \text{Humidity} <> 70) = - (4/14) \cdot \log_2(4/14) - (10/14) \cdot \log_2(10/14) = 0.863$$

$$\text{GainRatio}(\text{Decision}, \text{Humidity} <> 70) = 0.016$$

► سایر مقادیر را نیز بررسی می کنیم.

$$\text{Entropy}(\text{Decision}|\text{Humidity} \leq 75) = - (1/5) \cdot \log_2(1/5) - (4/5) \cdot \log_2(4/5) = 0.721$$

$$\text{Entropy}(\text{Decision}|\text{Humidity} > 75) = - (4/9) \cdot \log_2(4/9) - (5/9) \cdot \log_2(5/9) = 0.991$$

$$\text{Gain}(\text{Decision}, \text{Humidity} <> 75) = 0.940 - (5/14) \cdot (0.721) - (9/14) \cdot (0.991) = 0.940 - 0.2575 - 0.637 = 0.045$$

$$\text{SplitInfo}(\text{Decision}, \text{Humidity} <> 75) = -(5/14) \cdot \log_2(4/14) - (9/14) \cdot \log_2(10/14) = 0.940$$

$$\text{GainRatio}(\text{Decision}, \text{Humidity} <> 75) = 0.047$$

$$\text{Gain}(\text{Decision}, \text{Humidity} <> 78) = 0.090, \text{GainRatio}(\text{Decision}, \text{Humidity} <> 78) = 0.090$$

$$\textbf{Gain}(\text{Decision}, \text{Humidity} <> 80) = \textbf{0.101}, \textbf{GainRatio}(\text{Decision}, \text{Humidity} <> 80) = \textbf{0.107}$$

$$\text{Gain}(\text{Decision}, \text{Humidity} <> 85) = 0.024, \text{GainRatio}(\text{Decision}, \text{Humidity} <> 85) = 0.027$$

$$\text{Gain}(\text{Decision}, \text{Humidity} <> 90) = 0.010, \text{GainRatio}(\text{Decision}, \text{Humidity} <> 90) = 0.016$$

$$\text{Gain}(\text{Decision}, \text{Humidity} <> 95) = 0.048, \text{GainRatio}(\text{Decision}, \text{Humidity} <> 95) = 0.128$$

➤ در اینجا مقدار 96 را به عنوان آستانه نادیده می گیریم زیرا رطوبت نمی تواند بیشتر از این مقدار باشد.

➤ همانطور که مشاهده شد، بهره زمانی به حداکثر می رسد که آستانه برابر با 80 برای رطوبت باشد. این بدان معنی است که برای ایجاد شاخه ای در درخت خود باید سایر ویژگی های اسمی و مقایسه رطوبت با 80 را مقایسه کنیم.

➤ ویژگی دما نیز پیوسته است. هنگامی که تقسیم دودویی را به دما برای تمام نقاط تقسیم ممکن اعمال می کنیم، قانون تصمیم گیری زیر برای هر دو نسبت بهره و بهره حداکثر می شود.

$$\text{Gain(Decision, Temperature } \leq 83) = 0.113, \text{ GainRatio(Decision, Temperature } \leq 83) = 0.305$$

- در تصویر نسبت های سود و بهره محاسبه شده را به صورت خلاصه میبینید. اگر از سنجش بهره استفاده کنیم، outlook گره ریشه خواهد بود زیرا بالاترین مقدار بهره را دارد. از طرف دیگر، اگر از متریک نسبت بهره استفاده کنیم، دما گره ریشه خواهد بود زیرا بالاترین مقدار نسبت بهره را دارد. که در اینجا ترجیح داده شده از gain مشابه ID3 استفاده شود.
- پس از آن، مراحل مشابه را مانند ID3 اعمال می کنیم و درخت تصمیم زیر را ایجاد می کنیم. Outlook در گره ریشه قرار می گیرد. اکنون، ما باید به دنبال تصمیم گیری برای حالات مختلف Outlook باشیم.

Attribute	Gain	GainRatio
Wind	0.049	0.049
Outlook	0.246	0.155
Humidity <> 80	0.101	0.107
Temperature <> 83	0.113	0.305

Outlook = Sunny

- ▶ رطوبت را برای بیشتر از 80، و کمتر یا مساوی 80 تقسیم کرده‌ایم. اگر رطوبت بیشتر از 80 باشد، وقتی چشم‌انداز آفتابی و تصمیم‌گیری منفی است. به طور مشابه، اگر رطوبت کمتر یا مساوی 80 برای چشم‌انداز آفتابی باشد، تصمیم مثبت خواهد بود.

Day	Outlook	Temp.	Hum. > 80	Wind	Decision
1	Sunny	85	Yes	Weak	No
2	Sunny	80	Yes	Strong	No
8	Sunny	72	Yes	Weak	No
9	Sunny	69	No	Weak	Yes
11	Sunny	75	No	Strong	Yes

Outlook = Overcast

► اگر Outlook ابری باشد، بدون توجه به دما، رطوبت یا باد، تصمیم همیشه مثبت خواهد بود.

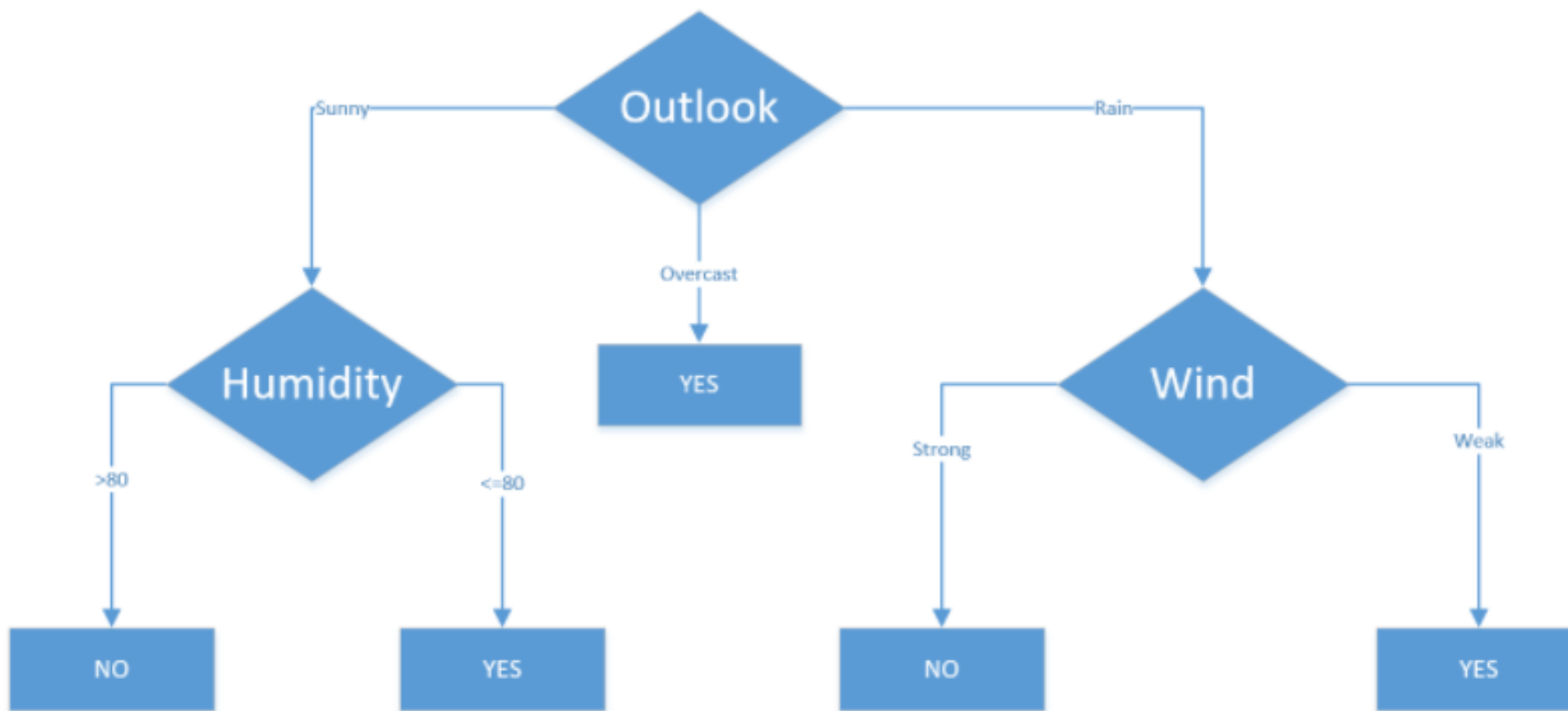
Day	Outlook	Temp.	Hum. > 80	Wind	Decision
3	Overcast	83	No	Weak	Yes
7	Overcast	64	No	Strong	Yes
12	Overcast	72	Yes	Strong	Yes
13	Overcast	81	No	Weak	Yes

Outlook = Rain

همانطور که مشاهده میشود، زمانی که باد ضعیف باشد، تصمیم بله خواهد بود، و اگر باد قوی باشد، خیر.

Day	Outlook	Temp.	Hum. > 80	Wind	Decision
4	Rain	70	Yes	Weak	Yes
5	Rain	68	No	Weak	Yes
6	Rain	65	No	Strong	No
10	Rain	75	No	Weak	Yes
14	Rain	71	No	Strong	No

➡ شکل نهایی جدول تصمیم گیری در زیر نشان داده شده است.



Decision tree generated by C4.5