

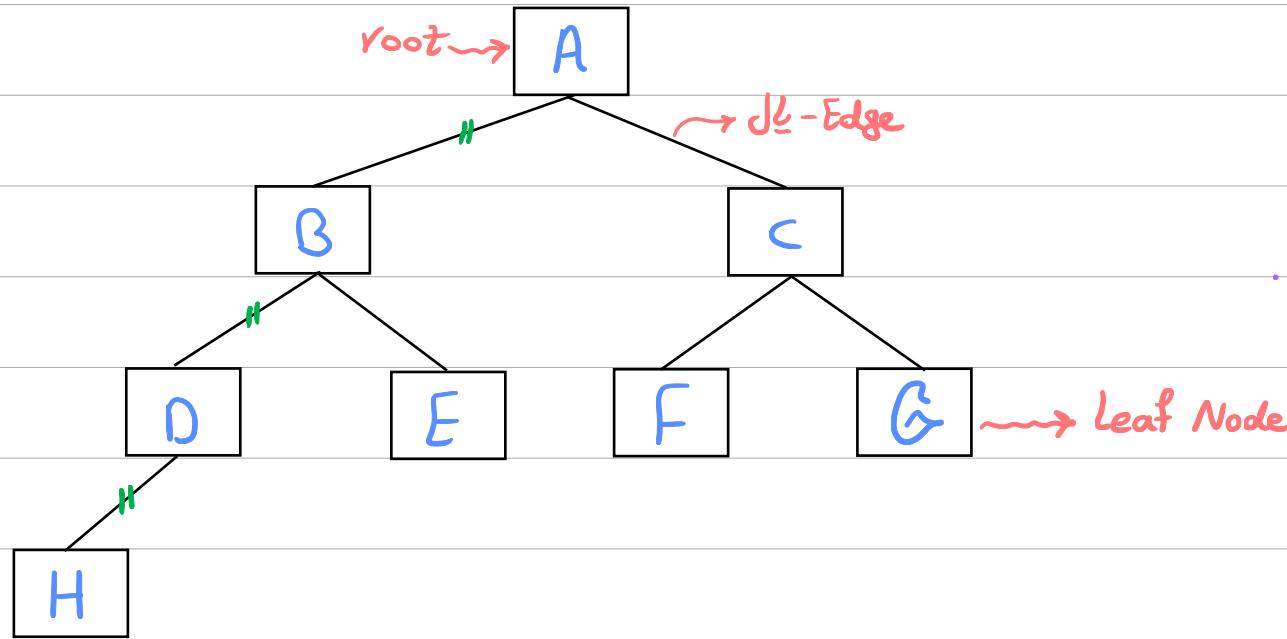
Session 9

Decision Tree



درخت تصمیم — (DT) Decision Tree

DT یکی از قدیمی‌ترین روش‌های طبقه‌بندی است.



ساختار داده درخت:

عمر درخت → تعداد تمام بالها از ریشه تا عیتی ترین برگ.

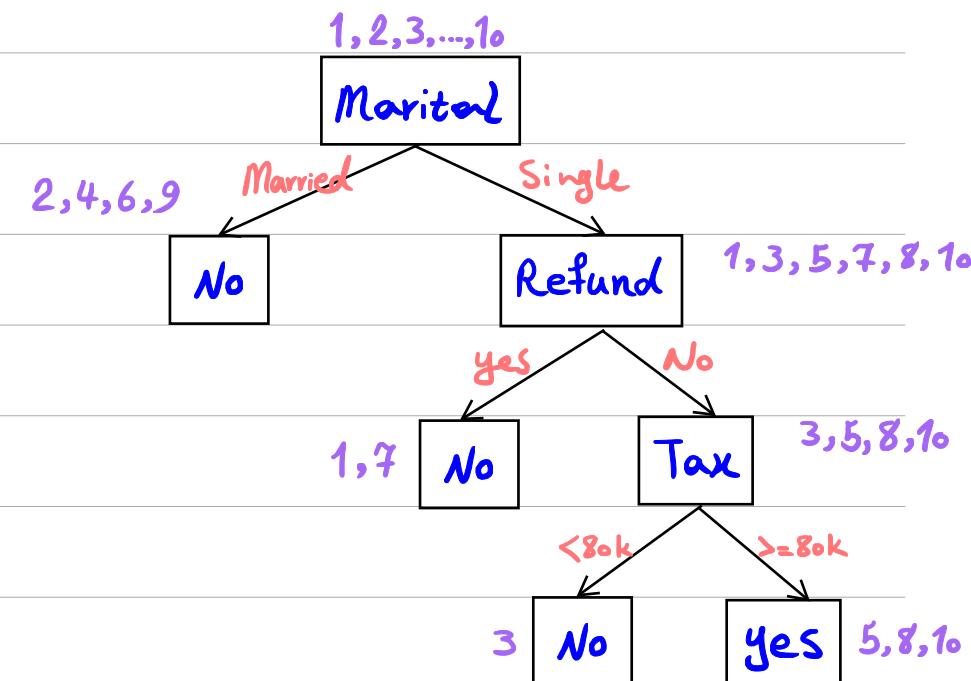
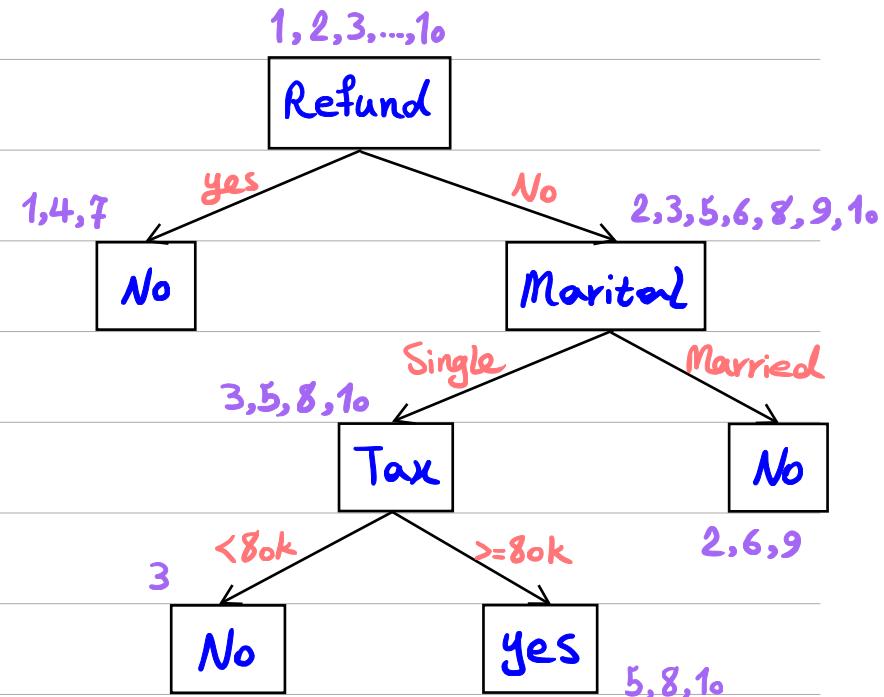
3 = عمر درخت

گره داخلی ← ویژگی

گره برگ ← class label

بازپرداخت وام قبلی
با محدوده مالیات برآورده شدن

	Refund	Marital Status	Taxable Income	Class
1	Yes	Single	125k	No
2	No	Married	100k	No
3	No	Single	70k	No
4	Yes	Married	120k	No
5	No	Single	95k	Yes
6	No	Married	60k	No
7	Yes	Single	220k	No
8	No	Single	85k	Yes
9	No	Married	75k	No
10	No	Single	90k	Yes



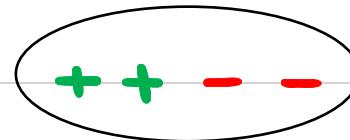
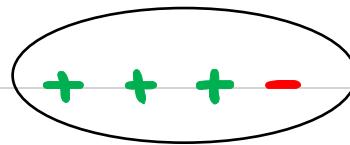
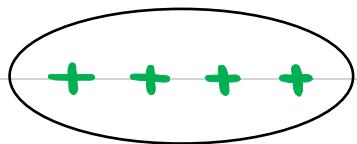
χ^2 test : No, Married, 80k \rightarrow Class=? No

؟: کدام درخت بکسر است؟ از کدام دیگری شروع کنم؟
برای پاسخ به این سوال نیاز به اطلاعات بیشتری داریم.

الگوریتم‌های DT اغلب بر پایه‌ی جستجوی متریانه کار می‌کنند.

$$\text{Entropy}(S) = \sum_{i=1}^c -P_i \log_2 P_i$$

Measure of Node Impurity ← معیار ناخالعی یک که در Entropy



$$\text{Entropy} = 0$$

$$\text{Entropy} = 0.811$$

$$\text{Entropy} = 1$$

$$P_+ = \frac{4}{4} = 1, \quad P_- = \frac{0}{4} = 0$$

$$P_+ = \frac{3}{4}, \quad P_- = \frac{1}{4}$$

$$P_+ = \frac{2}{4}, \quad P_- = \frac{2}{4}$$

$$\text{Entropy}([4+, 0-]) =$$

$$-\frac{4}{4} \log_2 \frac{4}{4} - \frac{0}{4} \log_2 \frac{0}{4} = 0$$

$$\text{Entropy}([3+, 1-]) =$$

$$-\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} = 0.811$$

$$\text{Entropy}([2+, 2-]) =$$

$$-\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4} = 1$$

$$0 < \text{Entropy} \leq 1$$

* هرچقدر Entropy کمتر خلوص بیشتر (خالع تر)

	Wind	Temp	OutLook	Humidity	Play?
1	Weak	Hot	Sunny	High	Play
2	Strong	Hot	Sunny	High	Play
3	Weak	Hot	Rain	High	Stay
4	Weak	Mid	Overcast	High	Play
5	Strong	Cold	Rain	Normal	Stay
6	Weak	Cold	Overcast	Normal	Play
7	Strong	Cold	Rain	Normal	Stay
8	Weak	Mid	Sunny	Normal	Play
9	Weak	Cold	Sunny	Normal	Play
10	Strong	Mid	Overcast	Normal	Play
11	Weak	Mid	Sunny	High	Stay
12	Strong	Mid	Rain	High	Stay
13	Weak	Hot	Overcast	Normal	Play
14	Weak	Cold	Rain	High	Play

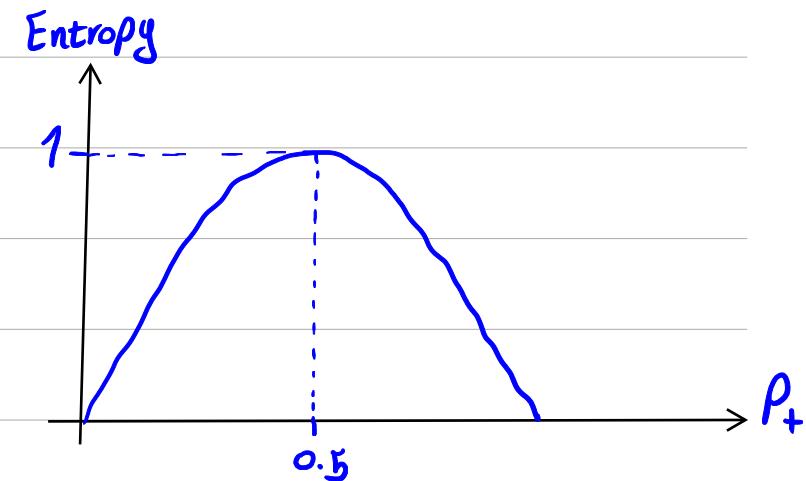
Entropy (S) =

$$-\frac{9}{14} \log_2 \left(\frac{9}{14}\right) - \frac{5}{14} \log_2 \left(\frac{5}{14}\right) =$$

0.940

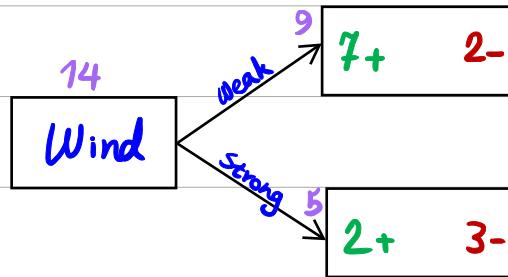
$$P_+ = \frac{9}{14}$$

$$P_- = \frac{5}{14}$$



Gain ← سمجھ

$$\text{Gain}(S, A) = \text{Ent}(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} \text{Ent}(S_v)$$



$$\text{Ent}(\text{weak}) = -\frac{7}{9} \log_2 \frac{7}{9} - \frac{2}{9} \log_2 \frac{2}{9} = 0.764$$

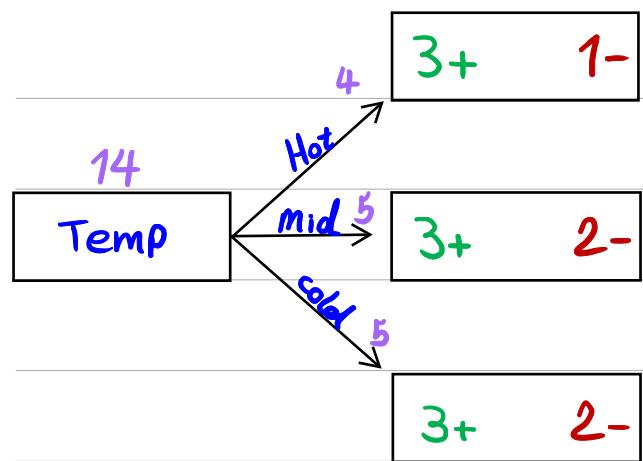
$$\text{Ent}(\text{Strong}) = -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} = 0.97$$

$$E(D) = -\frac{9}{14} \log_2 \left(\frac{9}{14}\right) - \frac{5}{14} \log_2 \left(\frac{5}{14}\right) = 0.94$$

$$\text{Gain}(D, \text{Wind}) = E(D) - \sum_{v \in \text{value(Wind)}} \frac{|D_v|}{|D|} \text{Ent}(D_v) =$$

$$0.94 - \left(\frac{9}{14} \text{Ent}(\text{weak}) + \frac{5}{14} \text{Ent}(\text{strong}) \right) =$$

$$0.94 - (0.491 + 0.346) = 0.94 - 0.837 = 0.102$$



$$Ent(Hot) = -\frac{3}{4} \log_2 \left(\frac{3}{4}\right) - \frac{1}{4} \log_2 \left(\frac{1}{4}\right) = 0.811$$

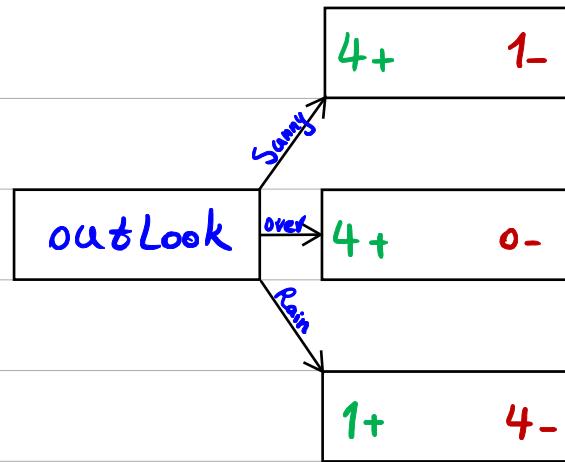
$$Ent(Mid) = -\frac{3}{5} \log_2 \left(\frac{3}{5}\right) - \left(\frac{2}{5}\right) \log_2 \left(\frac{2}{5}\right) = 0.97 = Ent(Cold)$$

Gain(D, Temp) = Ent(D) - $\sum_{v \in \text{Values(Temp)}} \frac{|D_v|}{|D|} Ent(D_v) =$

$$0.94 - \left(\frac{4}{14} Ent(Hot) + \frac{5}{14} Ent(Mid) + \frac{5}{14} Ent(Cold) \right) =$$

$$0.94 - (0.231 + 0.346 + 0.346) = 0.94 - 0.932 =$$

$$0.008$$

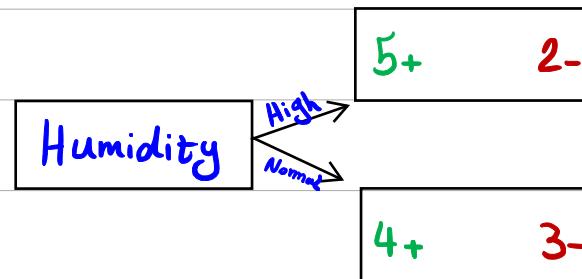


$$Ent(\text{Sunny}) = -\frac{4}{5} \log_2 \left(\frac{4}{5}\right) - \frac{1}{5} \log_2 \left(\frac{1}{5}\right) = 0.722$$

$$Ent(\text{overcast}) = -\frac{4}{4} \log_2 \frac{4}{4} - \frac{0}{4} \log_2 \frac{0}{4} = 0$$

$$Ent(\text{Rain}) = -\frac{1}{5} \log_2 \left(\frac{1}{5}\right) - \frac{4}{5} \log_2 \left(\frac{4}{5}\right) = 0.722$$

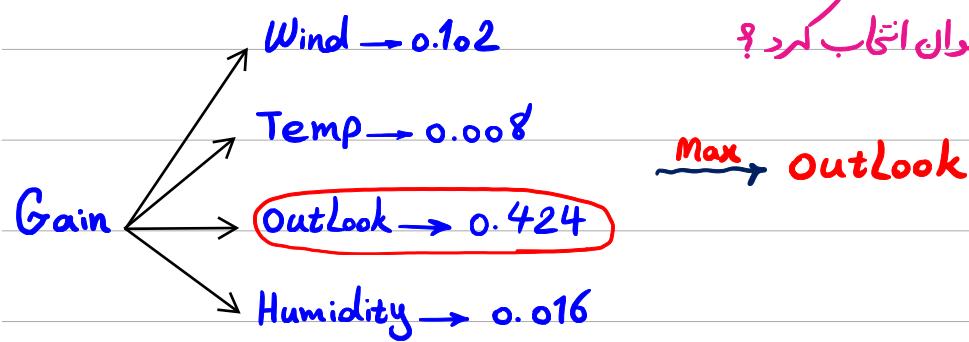
$$\text{Gain}(O, \text{outLook}) = Ent(D) - \left(\frac{5}{14} Ent(S) + \frac{4}{14} Ent(O) + \frac{5}{14} Ent(R) \right) = 0.94 - (0.258 + 0 + 0.258) = 0.94 - 0.516 = 0.424$$



$$Ent(\text{High}) = -\frac{5}{7} \log_2 \frac{5}{7} - \frac{2}{7} \log_2 \frac{2}{7} = 0.863$$

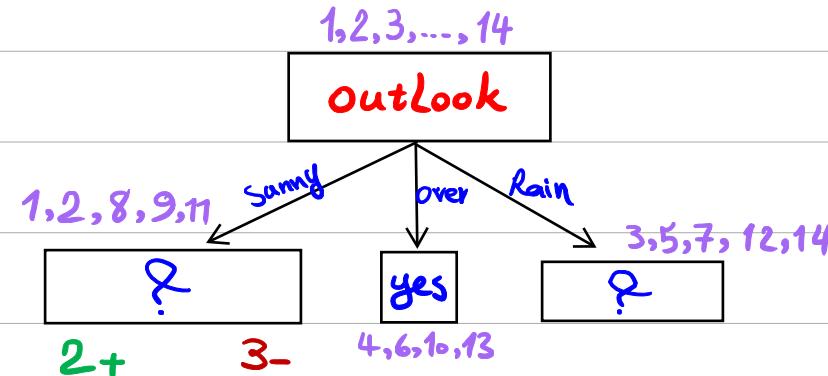
$$Ent(\text{Normal}) = -\frac{4}{7} \log_2 \frac{4}{7} - \frac{3}{7} \log_2 \frac{3}{7} = 0.985$$

$$\text{Gain}(O, \text{Humidity}) = Ent(D) - \left(\frac{7}{14} Ent(H) + \frac{7}{14} Ent(N) \right) = 0.94 - (0.431 + 0.493) = 0.94 - 0.924 = 0.016$$



ماکسیمم Gain را برای انتخاب کرد و از feature pl. 1 را برای دستگاه آوردید. feature pl. 2 را برای دستگاه آوردید.

ماکسیمم Gain بیشترین خود را داشت.



$$D = \{1, 2, 8, 9, 11\}$$

$$\text{Gain}(D, \text{Humidity}) = \text{Ent}(D) - \sum \frac{|D_s|}{|D|} \text{Ent}(D_s) = 0.97 - \frac{3}{5} \times \text{Ent}(H) - \frac{2}{5} \times \text{Ent}(N) = 0.97 - 0 - 0 = 0.97$$

$$\text{Gain}(D, \text{Temp}) = \text{Ent}(D) - \frac{2}{5} \text{Ent}(H) - \frac{2}{5} \text{Ent}(M) - \frac{1}{5} \text{Ent}(C) = 0.97 - 0 - 0.4 - 0 = 0.57$$

$$\text{Gain}(D, \text{wind}) = \text{Ent}(D) - \frac{0.97}{5} \text{Ent}(S) - \frac{3}{5} \text{Ent}(W) = 0.019$$

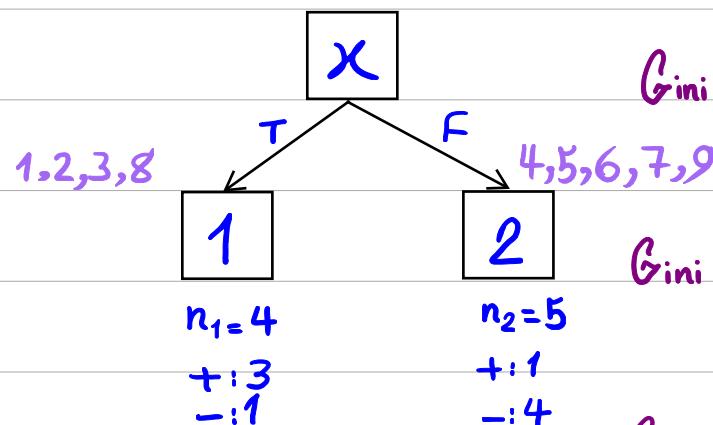
{ max → Humidity

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

$$GiniSplit(A) = \sum_{i=1}^k \frac{n_i}{n} Gini(i)$$

: Gini ضریب

	x	y	Class
1	T	T	+
2	T	T	+
3	T	F	-
4	F	F	+
5	F	T	-
6	F	T	-
7	F	F	-
8	T	F	+
9	F	T	-

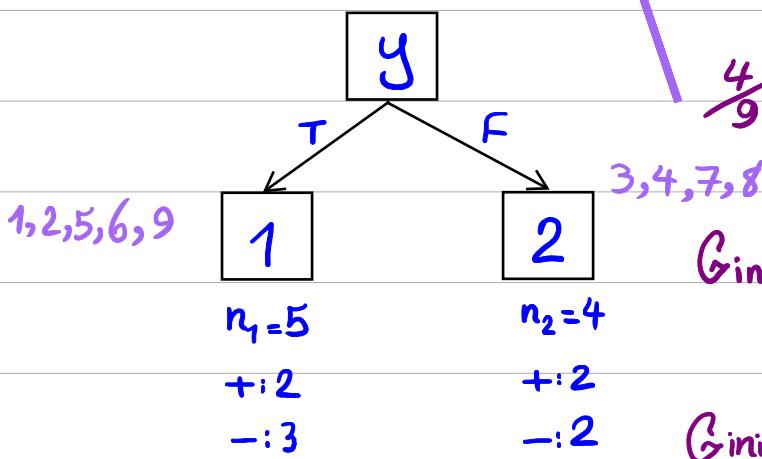


$$Gini(1) = 1 - \left(\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2 \right) = 1 - \left(\frac{9}{16} + \frac{1}{16} \right) = \frac{6}{16}$$

$$Gini(2) = 1 - \left(\left(\frac{1}{5}\right)^2 + \left(\frac{4}{5}\right)^2 \right) = 1 - \left(\frac{1}{25} + \frac{16}{25} \right) = \frac{8}{25}$$

$$GiniSplit(x) = \frac{n_1}{n} Gini(1) + \frac{n_2}{n} Gini(2) =$$

$$\frac{4}{9} \times \frac{6}{16} + \frac{5}{9} \times \frac{8}{25} = 0.166 + 0.177 = \underline{\underline{0.343}}$$



$$Gini(1) = 1 - \left(\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2 \right) = 1 - \left(\frac{9}{25} + \frac{4}{25} \right) = \frac{12}{25}$$

$$Gini(2) = 1 - \left(\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2 \right) = 1 - \left(\frac{4}{16} + \frac{4}{16} \right) = \frac{8}{16}$$

$x \leftarrow \min$

$$GiniSplit(y) = \frac{n_1}{n} Gini(1) + \frac{n_2}{n} Gini(2) =$$

مقدار ضریب Gini

$$\frac{5}{9} \times \frac{12}{25} + \frac{4}{9} \times \frac{8}{16} = 0.222 + 0.266 = 0.488$$

ج: اگر دیش کی مال چند مدلی باشد، از داده انتساب استفاده کنم یا چند انتساب؟

Car Type	class
Family	C1
Sport	C1
Luxury	C1
Sport	C1
Family	C2
Family	C2
Family	C2
Sport	C2
Luxury	C2

✓

class	a	Car Type	c
	Family	Sport	Luxury
C1	1	2	1
C2	4	1	1

5 3 2

$$\text{GiniSplit} = \frac{5}{10} \times \text{Gini}(a) + \frac{3}{10} \times \text{Gini}(b) + \frac{2}{10} \times \text{Gini}(c)$$

$$= 0.392$$

$$\text{Gini}(a) = 1 - \left(\left(\frac{1}{3}\right)^2 + \left(\frac{4}{3}\right)^2 \right) = 0.32$$

$$\text{Gini}(b) = 1 - \left(\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2 \right) = 0.44$$

$$\text{Gini}(c) = 1 - \left(\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 \right) = 0.5$$

class	Car Type	
	Sport, Luxury	Family
C1	a 3	b 1
C2	2	4

5 5

$$\text{GiniSplit} = \frac{5}{10} \times \text{Gini}(a) + \frac{5}{10} \times \text{Gini}(b)$$

$$= 0.4$$

$$\text{Gini}(a) = 1 - \left(\left(\frac{3}{3}\right)^2 + \left(\frac{2}{3}\right)^2 \right) = 0.48$$

$$\text{Gini}(b) = 1 - \left(\left(\frac{1}{5}\right)^2 + \left(\frac{4}{5}\right)^2 \right) = 0.32$$

class	Car Type		
	Sport	Luxury	Family
C1	2 a	2 b	
C2	1	5	

$$\text{GiniSplit} = \frac{3}{10} \times \text{Gini}(a) + \frac{7}{10} \times \text{Gini}(a) = 0.417$$

$$\text{Gini}(a) = 1 - \left(\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2 \right) = 0.44$$

$$\text{Gini}(b) = 1 - \left(\left(\frac{2}{7}\right)^2 + \left(\frac{5}{7}\right)^2 \right) = 0.408$$

کمترین Gini برای چند انتساب است

3

7

0 < Gini < 0.5

نوبتی حکم باش، نوبت شمات ::

RID	Age	Income	Student	Credit rating	Class: buys computer
1	Youth	High	No	Fair	No
2	Youth	High	No	Excellent	No
3	middle_aged	High	No	Fair	Yes
4	Senior	Medium	No	Fair	Yes
5	Senior	Low	Yes	Fair	Yes
6	Senior	Low	Yes	Excellent	No
7	middle_aged	Low	Yes	Excellent	Yes
8	Youth	Medium	No	Fair	No
9	Youth	Low	Yes	Fair	Yes
10	Senior	Medium	Yes	Fair	Yes
11	Youth	Medium	Yes	Excellent	Yes
12	middle_aged	Medium	No	Excellent	Yes
13	middle_aged	High	Yes	Fair	Yes
14	Senior	Medium	No	Excellent	No

کدام دسته کی را برای ریشه دخواست انتخاب کنیم؟
از میان این دسته کنی Gain کیا است؟

Gain (age) = ?

Gain (income) = ?

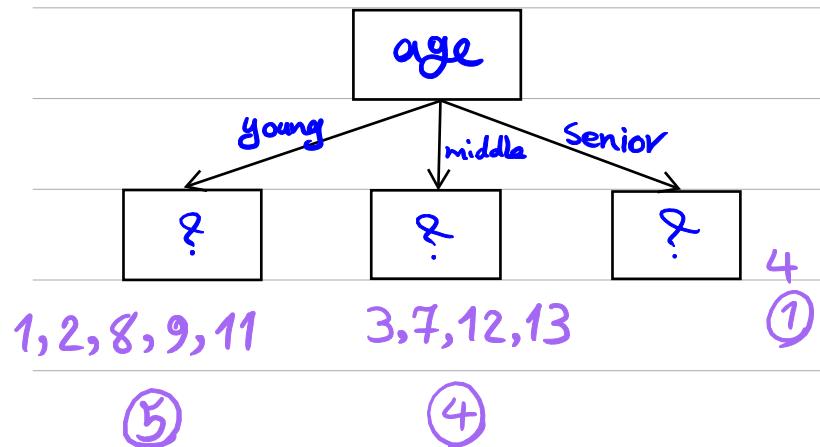
Gain (student) = ?

Gain (credit rating) = ?

↶ + -
9 5

Gain(age) =

$$\text{Entropy}(S) = \sum_{i=1}^c -P_i \log_2 P_i$$



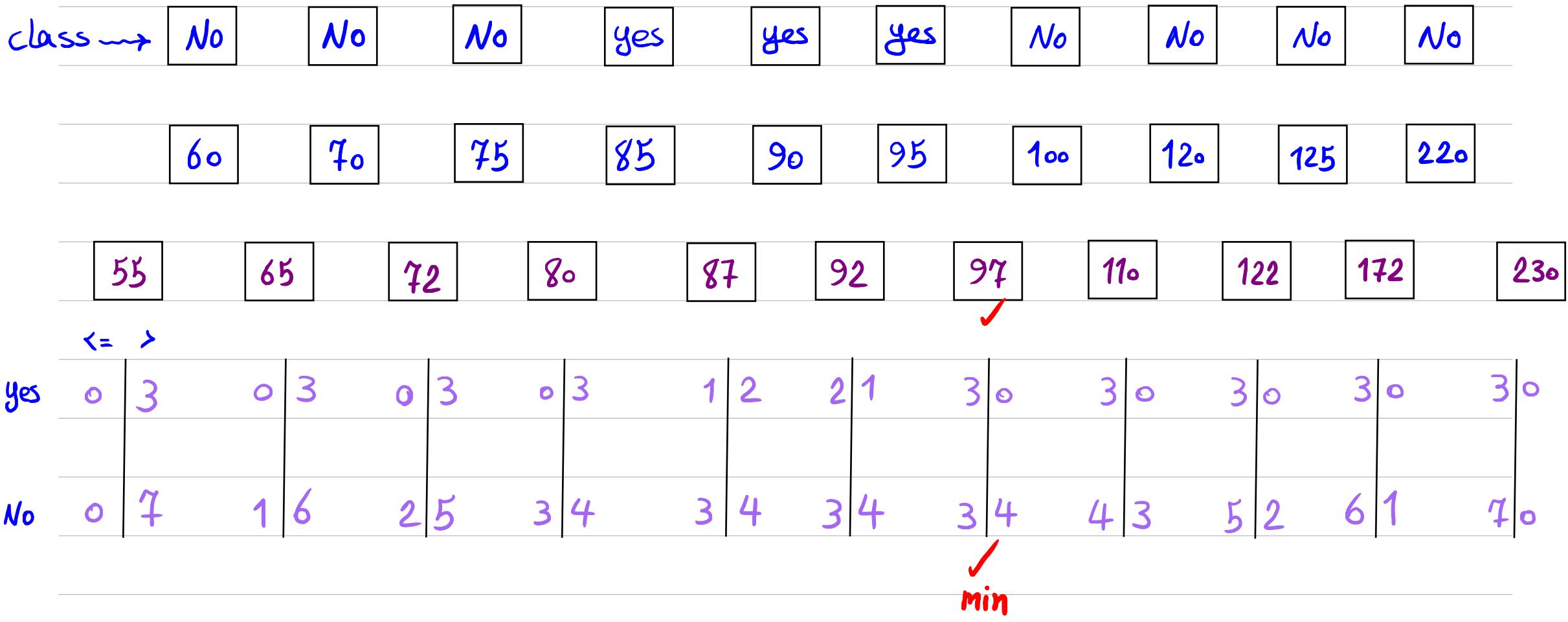
$$\text{Ent}(S) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.94$$

$$\text{Gain}(S, A) = \text{Ent}(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} \text{Ent}(S_v)$$

$$0.94 - \frac{5}{14} \times \left(-\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} \right) + \frac{4}{14} \times \left(-\frac{4}{4} \log_2 \frac{4}{4} \right) + \frac{5}{14} \times \left(-\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} \right) = 0.694$$

سایات برداشت
بازپرداخت وام قبلی
وضعیت تأهل

	Refund	Marital Status	Taxable Income	Class
1	Yes	Single	125k	No
2	No	Married	100k	No
3	No	Single	70k	No
4	Yes	Married	120k	No
5	No	Single	95k	Yes
6	No	Married	60k	No
7	Yes	Single	220k	No
8	No	Single	85k	Yes
9	No	Married	75k	No
10	No	Single	90k	Yes

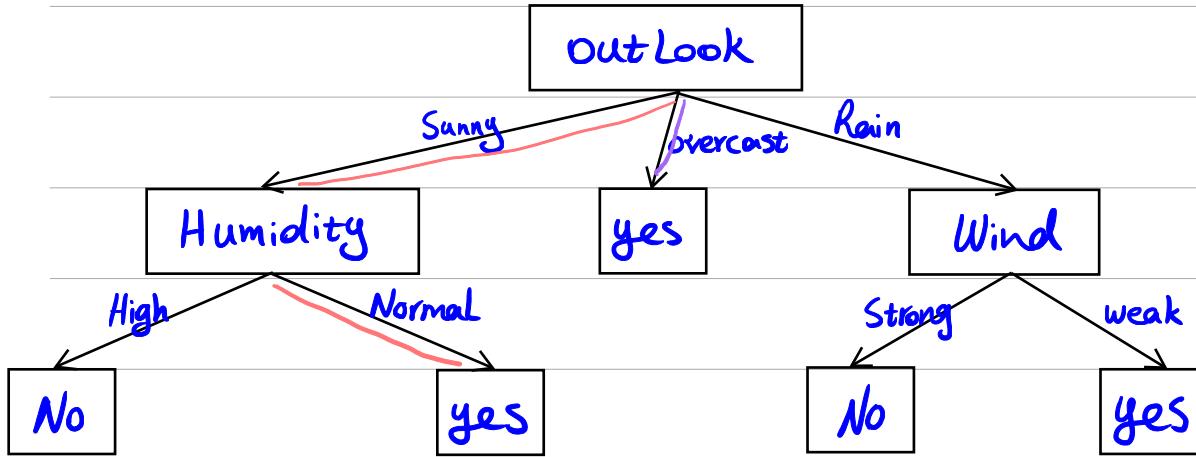


$$\text{GiniSplit}(97) = \frac{6}{10} \left(1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2\right) + \frac{4}{10} \left(1 - \left(\frac{0}{4}\right)^2 - \left(\frac{4}{4}\right)^2\right) = 0.3$$

$\leftarrow 97$ $\rightarrow 97$

اما مانند که احتمال حدس نا زنی، دقت DT برای داده های پیوسته پایین است.

قوانین در DT :



بھر جیسا ریٹھے ہے بگرڈ یا گوئند کی قانون.

If ($\text{outlook} = \text{Sunny} \wedge \text{Humidity} = \text{Normal}$) Then $\text{Play} = \text{yes}$

If ($\text{outlook} = \text{overcast}$) Then $\text{Play} = \text{yes}$

؟: چھ طور سنجھ شویں کی مکانوں درست اسات یا ہم تراز بقیہ است؟

بازپرداخت وام قبلی
و ضمیمه تأثیر
سایت برداشت
کاتواند وام پلیردج

	Refund	Marital Status	Taxable Income	Class
1	Yes	Single	125k	No
2	No	Married	100k	No
3	No	Single	70k	No
4	Yes	Married	120k	No
5	No	Single	95k	Yes
6	No	Married	60k	No
7	Yes	Single	220k	No
8	No	Single	85k	Yes
9	No	Married	75k	No
10	No	Single	90k	Yes

$$\text{Coverage}(R) = \frac{n_{\text{covers}}}{101}$$

$$\text{accuracy}(R) = \frac{n_{\text{covers}}}{n_{\text{covers}}}$$

If (Status=Single) Then class=No

چند نفر از کل جمعیت
 $\frac{6}{10} \rightarrow 60\%$

$\text{accuracy} = \frac{3}{6} = 50\%$

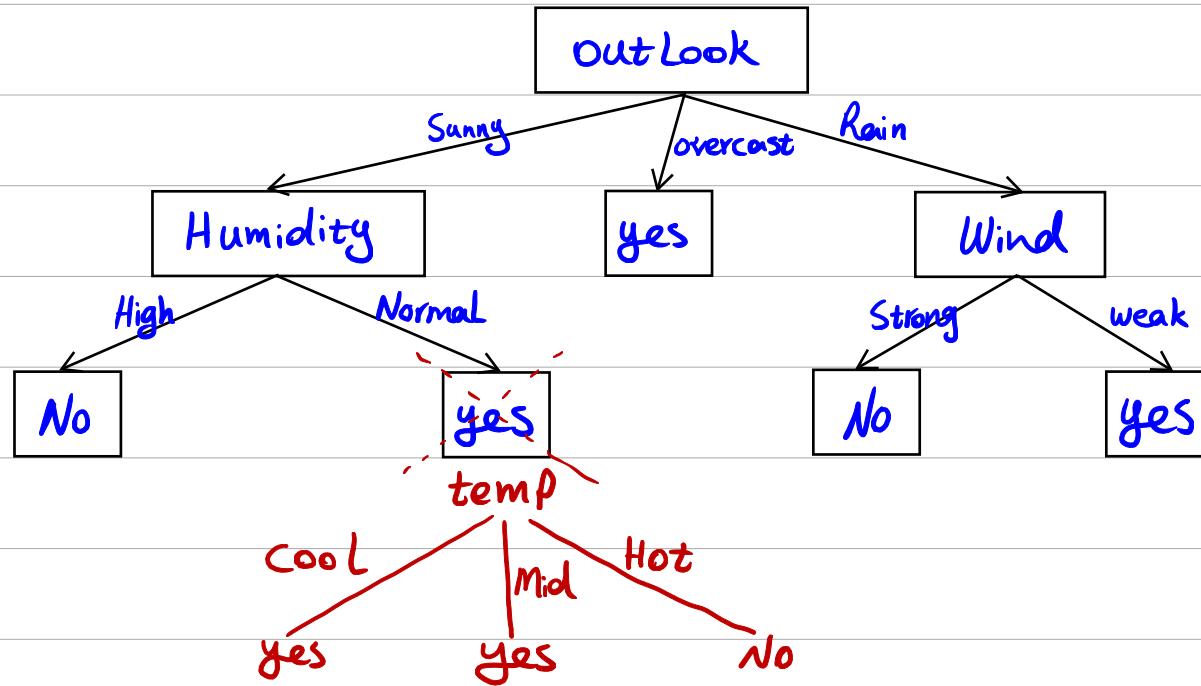
نیز شتر بایش accuracy , coverage همچو

قانون ۷ بازش تراست.

در برابر noise چه طور رفتاری کند؟ DT ؟

outLook	Temp	Humidity	Wind	Play=?
Sunny	Hot	Normal	Strong	No

نویز →



الرعداد اشغال هادرخت به صورت غیرقابل قبولی انزواش می‌باشد، دچار overfitting خواهد شد.

چه طوری توانیم شکل overfitting را بر طرف کنیم؟

Pruning یا هرس کردن:

: پیش هرس (Pre-Pruning)

درخت با داده های Validation رهگذام آموزش نست شده الگوریتم برآورزش رخ داده باشد، عملیات Split درخت متوقف می شود

: پس هرس (Post-Pruning)

1. ساخت درخت با توجه به داده های آموزشی، اجازه رشد دادن به رفت تا جایی که روی داده های آموزشی به خوبی $f(x)$ شود و overfitting رخ دهد.

2. تبدیل درخت به چرخه قوانین هم ارز با ایجاد یک قانون برای هر سیر از ریشه تا برگ.

3. هرس هر قانون یعنی حذف پیش شرط های آن به شرطی که نتیجه بحبود یابد.

4. مرتب کردن قوانین هرس شده به ترتیب دقت ثان و در نظر گرفتن آنها در دسته بندی مخونه های بعدی.

