① First we have to find the best split for the data using entropy based information gain

The dataset S consist of 11 instances. 6 of them are positive and 5 negative. This leads us to:

$$I(S) = -p_{pos}\log_2(p_{pos}) - p_{neg}\log_2(p_{neg})$$
$$= -\frac{6}{11}\log_2\left(\frac{6}{11}\right) - \frac{5}{11}\log_2\left(\frac{5}{11}\right) \approx 0{,}994$$

For each of the attributes we get the following gain:

<u>Outlook</u>: Sunny $[2^+, 2^-]$
Overcast $[1^+, 2^-]$
Rain $[3^+, 1^-]$

Gain: $I(S) - \frac{4}{11}I(S_{sunny}) - \frac{3}{11}I(S_{overcast}) - \frac{4}{11}I(S_{rain})$
$= 0{,}0849$

<u>Wind</u>: Weak $[3^+, 2^-]$
Strong $[3^+, 3^-]$

Gain: $I(S) - \frac{5}{11}I(S_{weak}) - \frac{6}{11}I(S_{strong}) \approx 0{,}0072$

→ <u>Temperature</u>: 7° $[1^+,0^-]$ 8° $[1^+,0^-]$ 10° $[1^+,0^-]$ 11° $[1^+,0^-]$
sort examples 20° $[1^+,0^-]$ | 21° $[0^+,1^-]$ 25° $[1^+,1^-]$ | 31 $[0^+,1^-]$
32° $[0^+,1^-]$ 34° $[0^+,1^-]$

Gain: $I(S) - \frac{1}{11}I(S_{7°}) - \frac{1}{11}I(S_{8°}) - \frac{1}{11}I(S_{10°}) - \frac{1}{11}I(S_{11°})$
$- I(S_{20°}) - \frac{1}{11}I(S_{21°}) - \frac{2}{11}I(S_{25°}) - \frac{1}{11}I(S_{31°})$
$\frac{1}{11}I(S_{32°}) - \frac{1}{11}I(S_{34°}) \approx 0{,}812$

// everything 0, except $I(S_{25°}) = 1$

The temperature gives the best information gain. So this category will be our first node. Find leafs next. The two red lines above mark adjacent examples with different classes, values already sorted, see above (should be sorted now, already done by accident)

Mean values for the two class-changes:

a) $\frac{20+21}{2}$ b) $\frac{23+31}{2}$

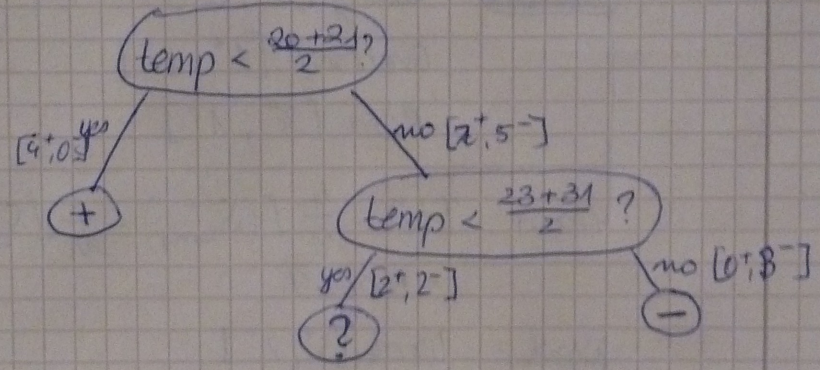
Calculate the information gain for both mean tests, namely temp $< \frac{20+21}{2}$ and temp $\frac{23+31}{2}$.

$$I(S_{temp<\frac{20+21}{2}}) = -\frac{5}{11}\log\left(\frac{5}{11}\right) - \frac{6}{11}\log\left(\frac{6}{11}\right) \simeq 0{,}994$$

$$I(S_{temp<\frac{23+31}{2}}) = -\frac{8}{11}\log\left(\frac{8}{11}\right) - \frac{3}{11}\log\left(\frac{3}{11}\right) \simeq 0{,}8454$$

test a) gives the higher information gain.

$\Rightarrow$



The children hot (temp > 27°C) and cool (temp < $\frac{20+21}{2}$) are classified perfectly, no further operations here.

For the descendant with temperatures between "hot" and "cool" another split has to be done.

There are four instances with temperature belonging to this group. The instances are:

	Outlook	Wind	Enjoy sport
S_{sunny}^{+}	Rain	weak	+
	Sunny	strong	+
	Overcast	strong	=
	Rain	strong	-

Let's call this Set S_{sunny}^{+}.

↪

Now we must calculate the gain for these attributes.

$I(S^+) = 1$ since there are equally distributed (+) and (−) instances (2 of each)

Outlook: Rain $[1^+, 1^-]$

Sunny $[1^+, 0^-]$

Overcast $[0^+, 1^-]$

Gain: $I(S^*) - \frac{2}{4} \underbrace{I(S^*_{Rain})}_{1} - \frac{1}{4} \underbrace{I(S^*_{Sunny})}_{0} - \frac{1}{4} \underbrace{I(S^*_{Overcast})}_{0}$

$= 1 - \frac{1}{2} = \frac{1}{2} = 0{,}5$

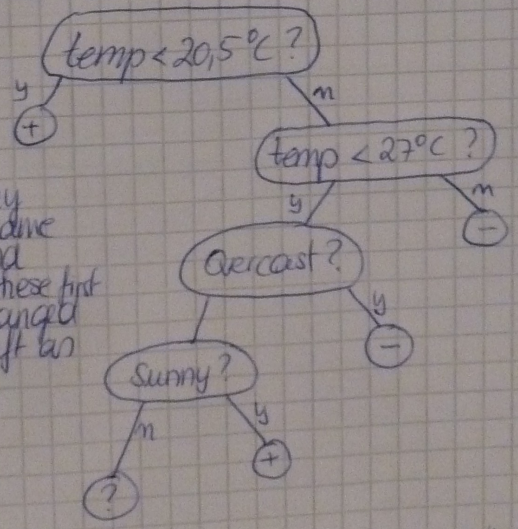
Wind: Weak $[1^+, 0^-]$

Strong $[1^+, 2^-]$

Gain: $I(S^*) - \frac{1}{4} \underbrace{I(S^*_{Weak})}_{0} - \frac{3}{4} I(S^*_{Strong}) = 1 - \frac{3}{4} \cdot 0{,}9185$

$\approx 0{,}3113$

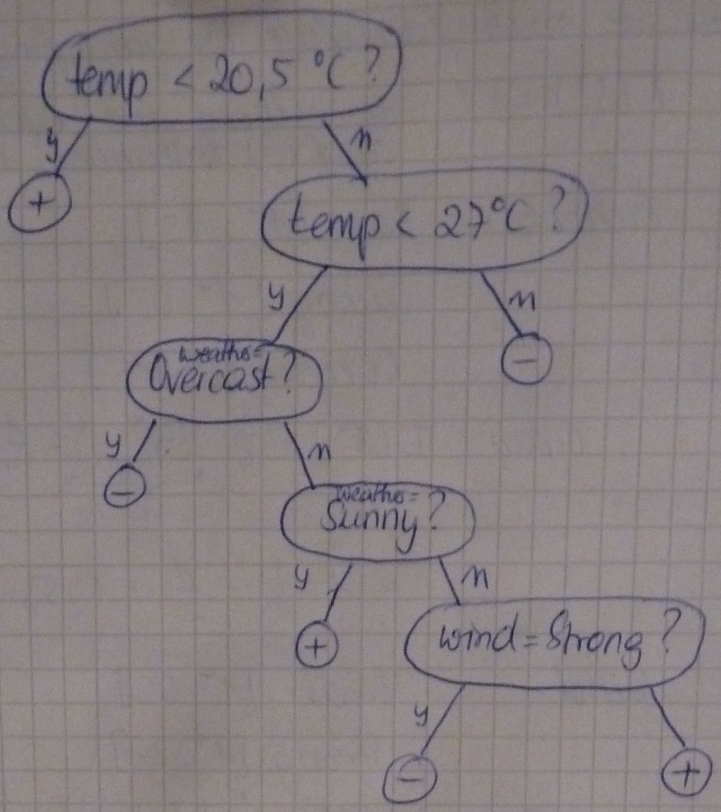
Outlook gives the largest gain and is the next attribute for the subtree



Overcast and Sunny obviously give the same information gain and Rainy is left, so these first values can be exchanged and Rainy is left on last test node.

The descendants "Overcast" and "Sunny" are perfectly ~~spot~~ classified. Thus leaves the attribute Wind. This classifies the rest perfectly hence the following tree:

Result:


temp < 20,5 °C ?
y
n
+
temp < 27°C ?
y
n
−
weather = Overcast ?
y
n
−
weather = Sunny ?
y
n
+
wind = Strong ?
y
−
+