Info gain for attribute Age - step by step calculation

0,6429	0,6374	0,4098		
0,3571	1,4854	0,5305	**Info(D)**	
		0,9403		
0,4000	1,3219	0,5288		
0,6000	0,7370	0,4422		
		0,9710	0,3571	0,3468
1,0000	0,0000	0,0000		
0,0000	0,0000	0,0000		
		0,0000	0,2857	0,0000
0,4000	1,3219	0,5288		
0,6000	0,7370	0,4422		
		0,9710	0,3571	0,3468
				0,6935
				0,2467

Attribute Selection: Information Gain

- Class P: buys_computer = "yes"
- Class N: buys_computer = "no"

$$Info(D) = I(9,5) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) = 0.940$$

$$Info_{age}(D) = \frac{5}{14}I(2,3) + \frac{4}{14}I(4,0) + \frac{5}{14}I(3,2) = 0.694$$

age	p_i	n_i	$I(p_i, n_i)$
<=30	2	3	0.971
31…40	4	0	0
>40	3	2	0.971

$\frac{5}{14}I(2,3)$ means "age <=30" has 5 out of 14 samples, with 2 yes'es and 3 no's. Hence

$$Gain(age) = Info(D) - Info_{age}(D) = 0.246$$

Similarly,

$$Gain(income) = 0.029$$
$$Gain(student) = 0.151$$
$$Gain(credit\_rating) = 0.048$$

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

14

Not: Normalde 2 ya da herhangi bir tabanda 0 sayısının logaritması tanımsızdır. Ama, tanımsızlık ya da hatayı önlemek adına, veri ve metin madenciliğinde info. gain ve Entropy hesabında 0 sayısının herhangi bir tabanda logaritması 0 olarak varsayılmaktadır.