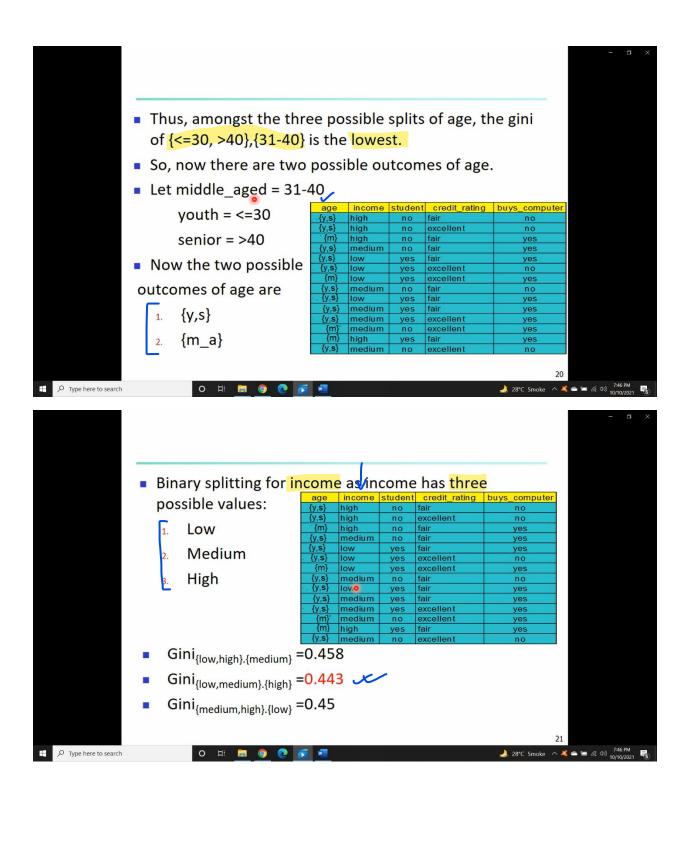
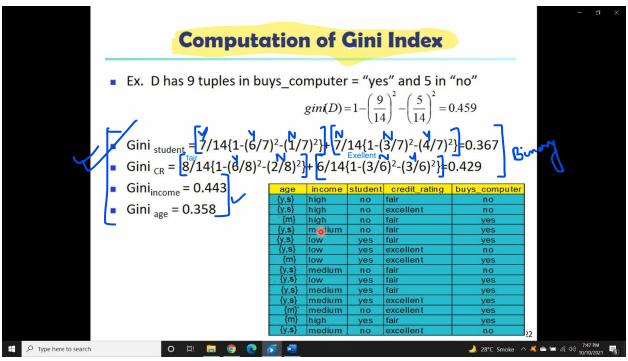


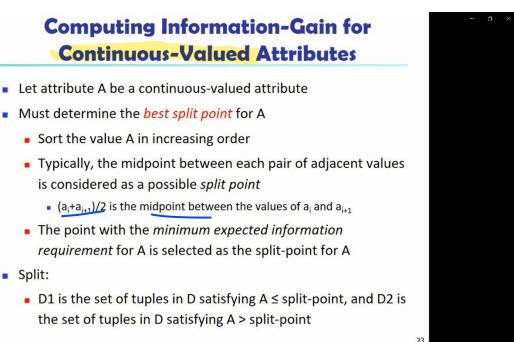
 $Gini_{\{<=30, >40\}, \{31-40\}} = 10/14\{1-(5/10)^2-(5/10)^2\}+4/14\{1-(4/4)^2-0\}$

 $Gini_{\{31-40,>40\},\{<=30\}} = 9/14\{1-(7/9)^2-(2/9)^2\} + 5/14\{1-(2/5)^2-(3/5)^2\}$

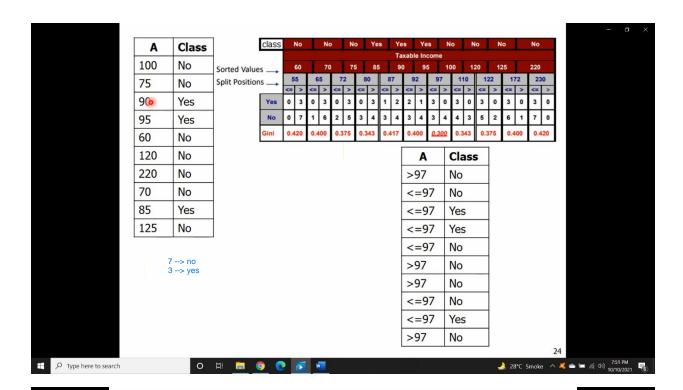
=0.39







Type here to search



Gain Ratio for Attribute Selection (C4.5)

- ✓ Information gain measure is biased towards attributes with a large number of values
- C4.5 (a successor of ID3) uses gain ratio to overcome the problem (normalization to information gain)

$$SplitInfo_A(D) = -\sum_{j=1}^{\nu} \frac{|D_j|}{|D|} \times \log_2(\frac{|D_j|}{|D|})$$

- GainRatio(A) = Gain(A)/SplitInfo(A)
- **EX.** $SplitInfo_{income}(D) = -\frac{4}{14} \times \log_2(\frac{4}{14}) \frac{6}{14} \times \log_2(\frac{6}{14}) \frac{4}{14} \times \log_2(\frac{4}{14}) = 1.557$
 - gain_ratio(income) = 0.029/1.557 = 0.019
- The attribute with the maximum gain ratio is selected as the splitting attribute

