Data Mining and Knowledge Discovery

Knowledge Discovery and Knowledge Management in e-Science

Petra Kralj Novak

Petra.Kralj.Novak@ijs.si

Practice, 2008/11/12

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

ROC space exercise

Simple mushroom dataset

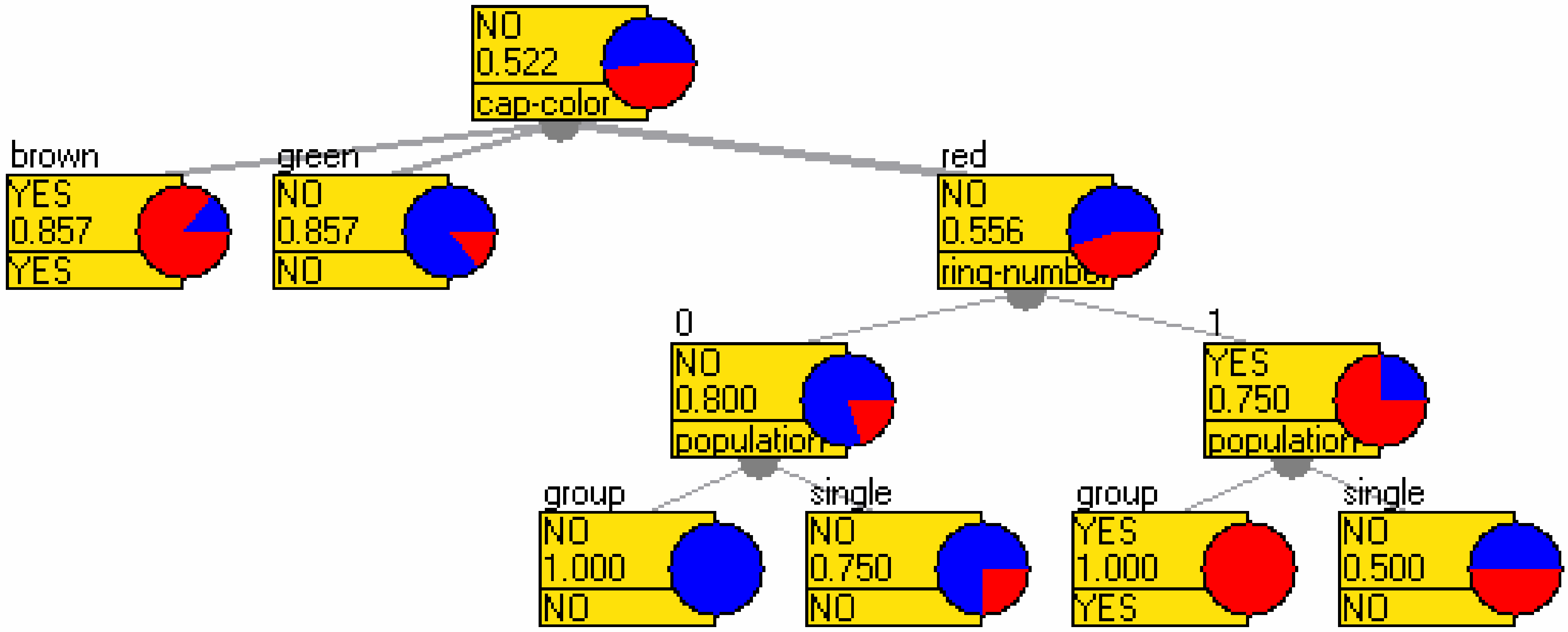
Train set

cap-color	ring-number	population	EDIBLE
red	1	single	YES
green	1	group	NO
brown	1	single	YES
brown	1	single	YES
brown	1	single	YES
brown	1	single	YES
red	1	single	NO
red	0	group	NO
green	0	group	NO
green	0	single	NO
green	0	single	NO
red	1	group	YES
red	1	group	YES
brown	1	group	YES
brown	0	single	YES
brown	0	single	NO
green	0	group	NO
green	0	group	NO
red	0	single	NO
red	0	single	YES
red	0	single	NO
green	0	group	YES
red	0	single	NO

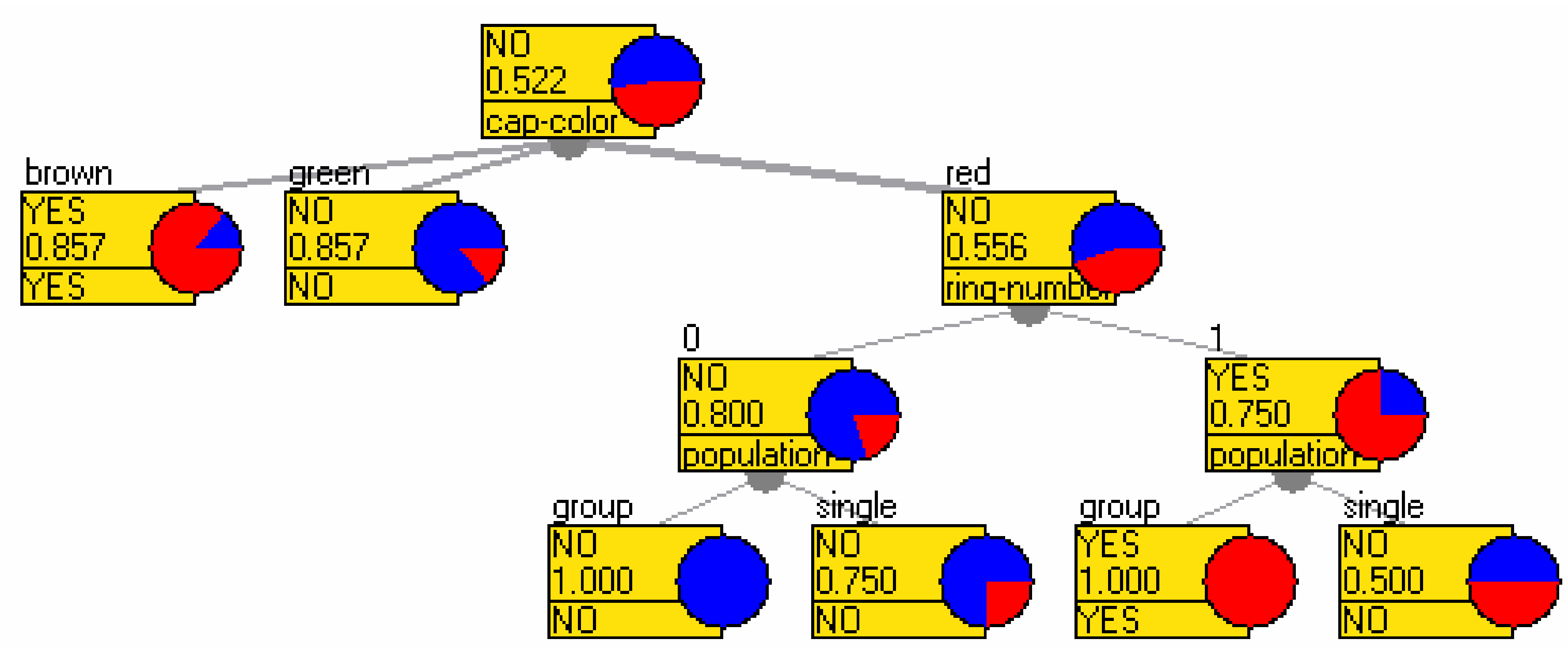
Test set

cap-color	ring-number	population	EDIBLE
brown	1	single	NO
green	0	group	NO
red	1	single	YES
red	0	group	NO
red	1	group	YES

Decision tree induced on the train set



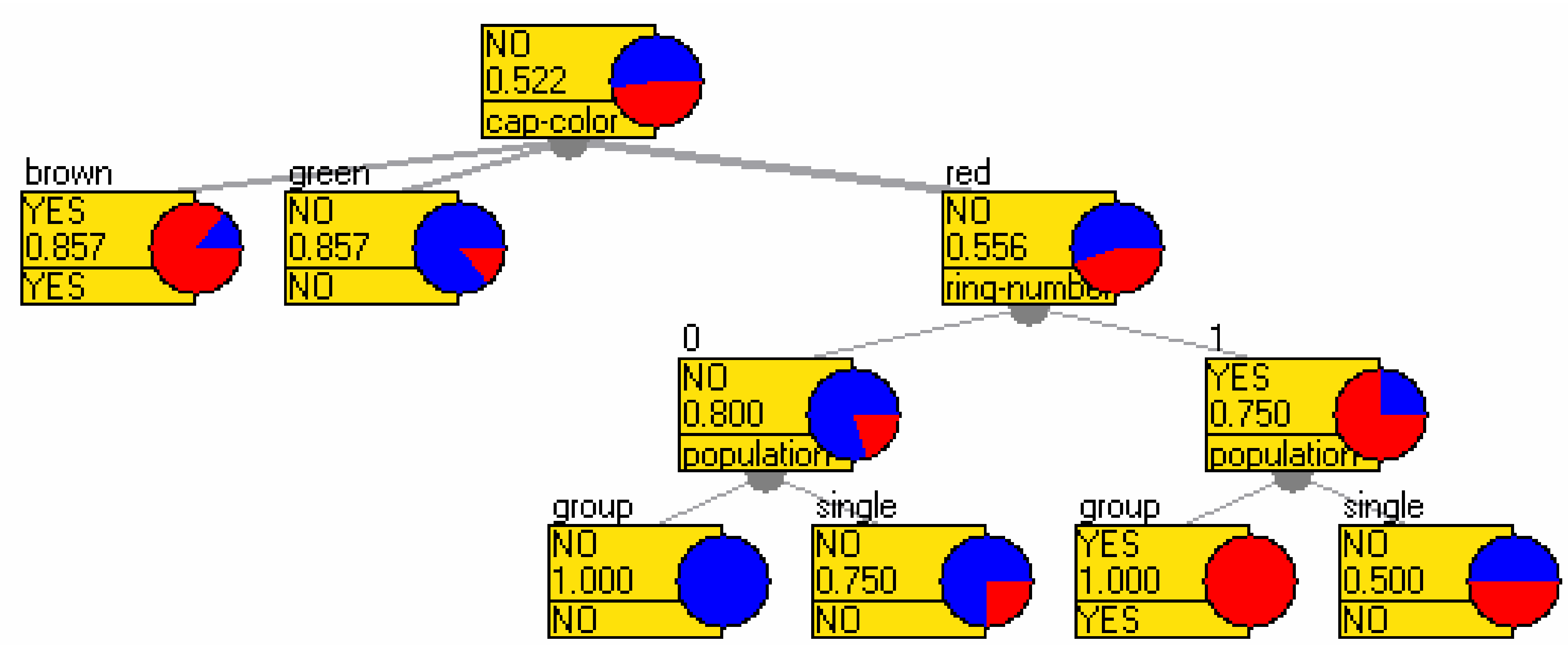
Confusion matrix



cap-color	ring-number	population	EDIBLE	DT1
brown	1	single	NO	
green	0	group	NO	
red	1	single	YES	
red	0	group	NO	
red	1	group	YES	

	Predicted YES	Predicted NO
Actual YES		
Actual NO		

Confusion matrix



cap-color	ring-number	population	EDIBLE	DT1
brown	1	single	NO	YES
green	0	group	NO	NO
red	1	single	YES	NO
red	0	group	NO	NO
red	1	group	YES	YES

	Predicted YES	Predicted NO
Actual YES	1	1
Actual NO	1	2

ROC space

	Predicted YES	Predicted NO
Actual YES	1	1
Actual NO	1	2

- True positive rate =

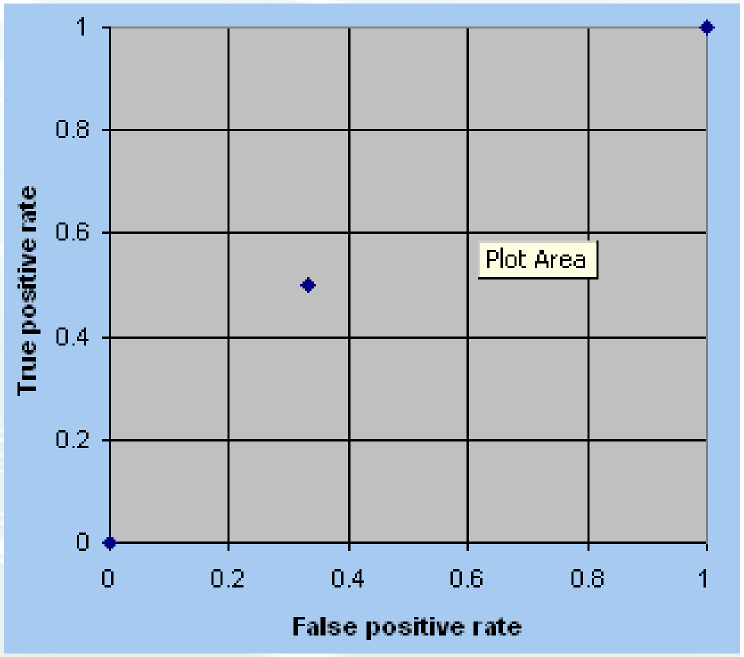
  = # true positives / # all positives =

  = TPr = 1/2

- False positive rate =

  = # false positives / # all negatives =

  = FPr = 1/3

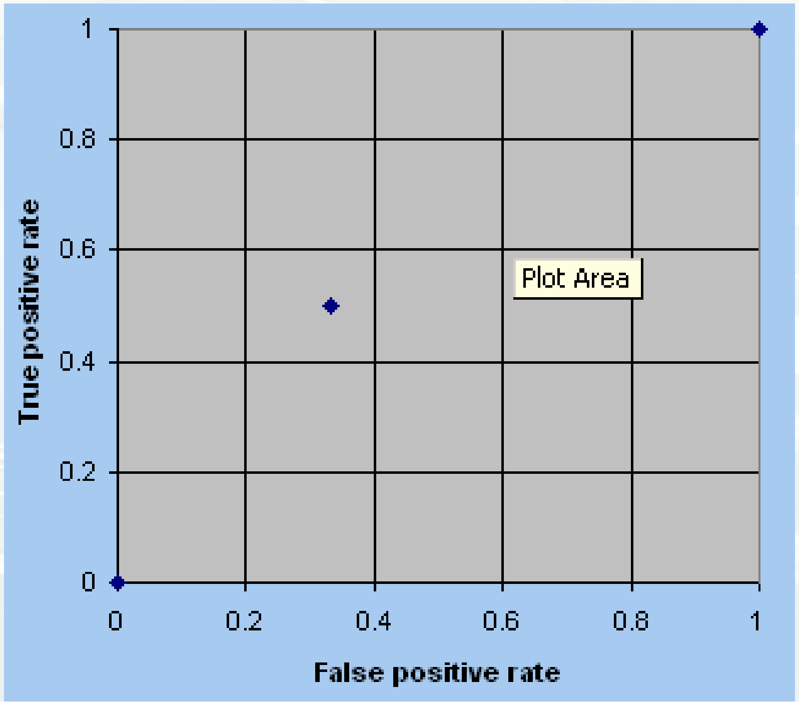


ROC space 2

- Classifier “always YES”

	Predicted YES	Predicted NO
Actual YES	2	0
Actual NO	3	0

- TPr = 1
- FPr = 1

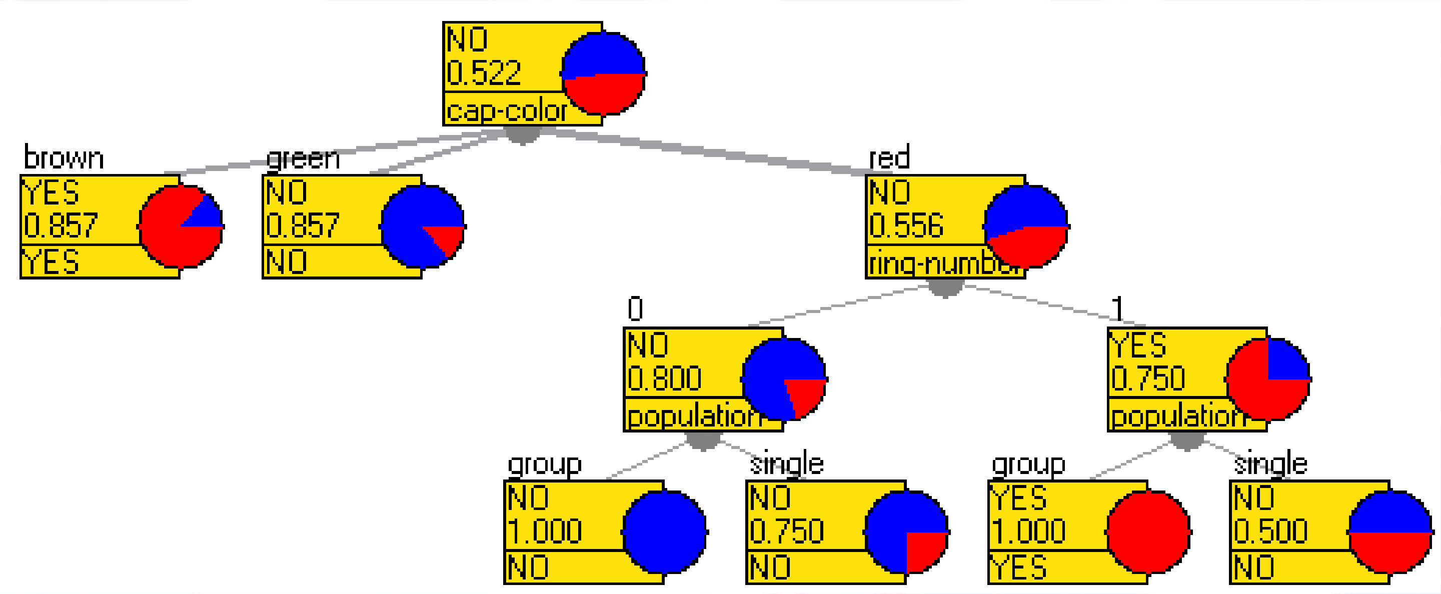


- Classifier “always NO”

	Predicted YES	Predicted NO
Actual YES	0	2
Actual NO	0	3

- TPr = 0
- FPr = 0

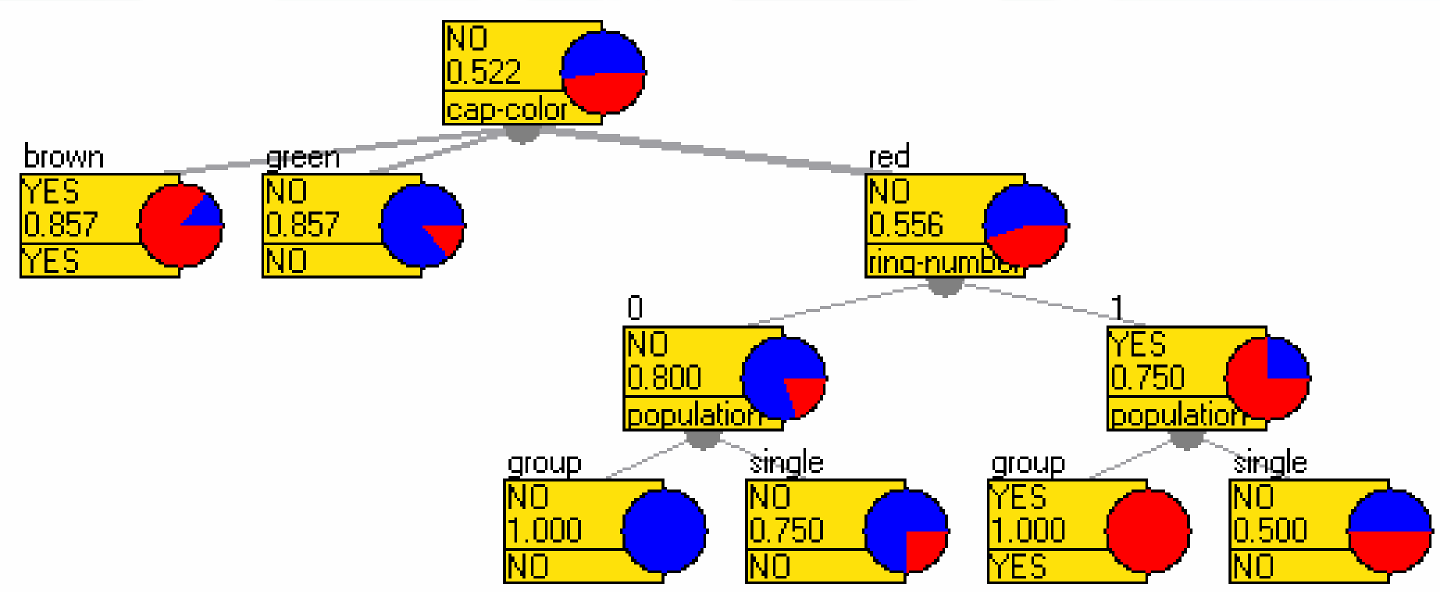
Confusion matrix 2:
A mushroom is edible if the model is at least 90% sure of this



cap-color	ring-number	population	EDIBLE	DT2
brown	1	single	NO	
green	0	group	NO	
red	1	single	YES	
red	0	group	NO	
red	1	group	YES	

	Predicted YES	Predicted NO
Actual YES		
Actual NO		

Confusion matrix 2:
A mushroom is edible if the model is at least 90% sure of this



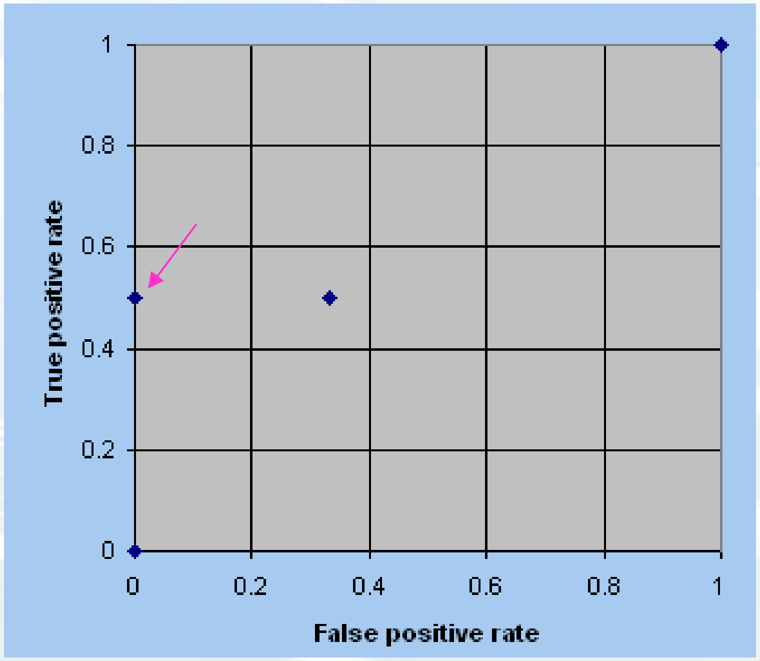
cap-color	ring-number	population	EDIBLE	DT2
brown	1	single	NO	NO
green	0	group	NO	NO
red	1	single	YES	NO
red	0	group	NO	NO
red	1	group	YES	YES

	Predicted YES	Predicted NO
Actual YES	1	1
Actual NO	0	3

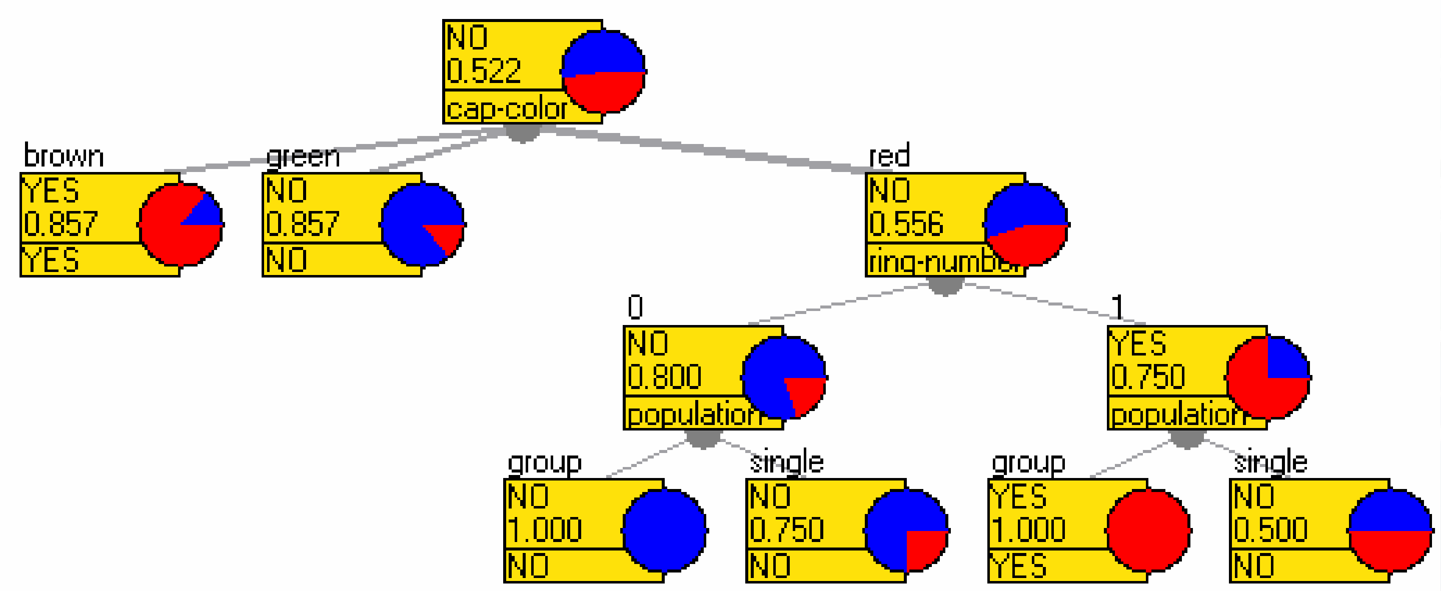
ROC space

	Predicted YES	Predicted NO
Actual YES	1	1
Actual NO	0	3

- True positive rate TPr = 1/2
- False positive rate FPr = 0



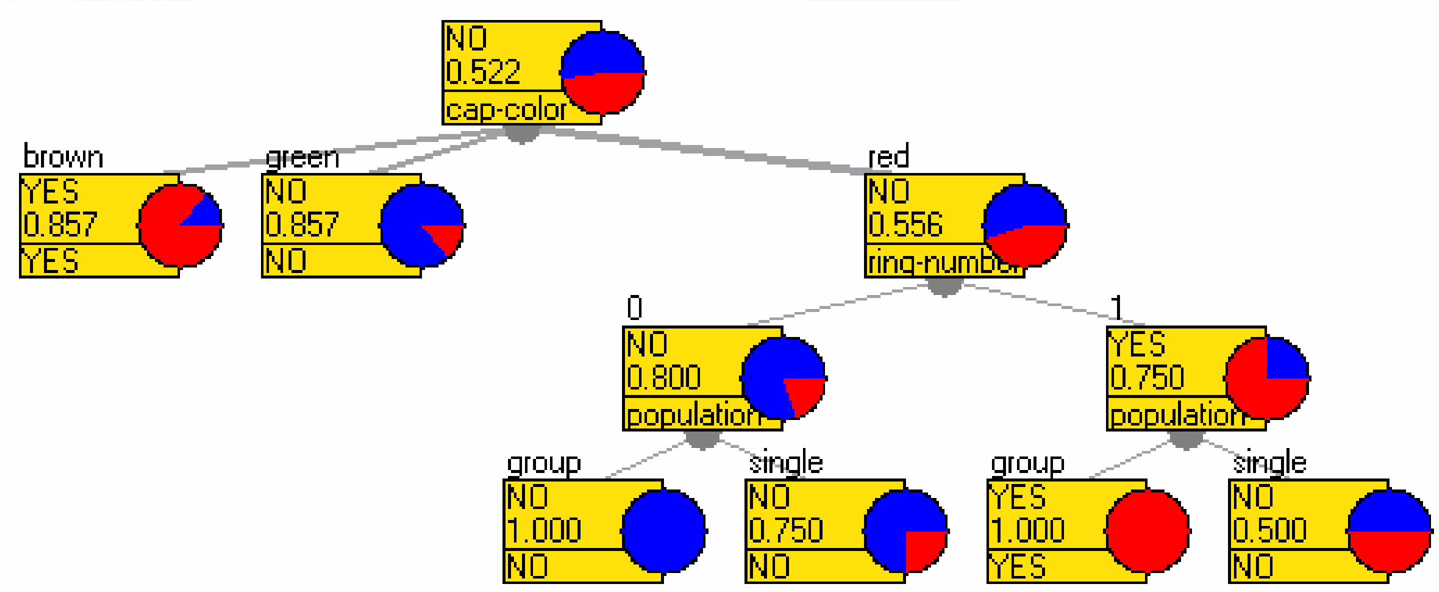
Confusion matrix 3: A mushroom is edible if the model is at least 20% sure of this



cap-color	ring-number	population	EDIBLE	DT3
brown	1	single	NO	
green	0	group	NO	
red	1	single	YES	
red	0	group	NO	
red	1	group	YES	

	Predicted YES	Predicted NO
Actual YES		
Actual NO		

Confusion matrix 3:
A mushroom is edible if the model is at least 20% sure of this



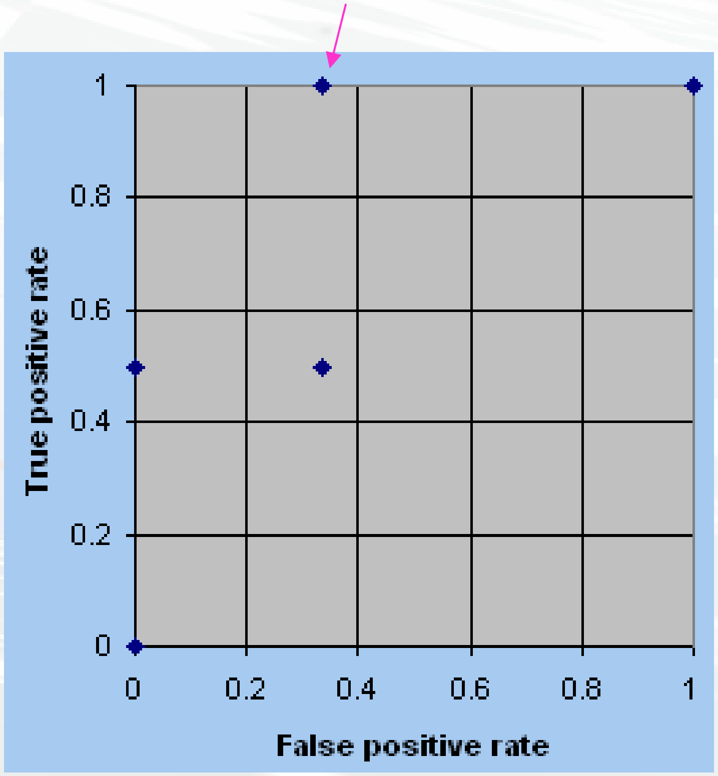
cap-color	ring-number	population	EDIBLE	DT3 (20%)
brown	1	single	NO	YES
green	0	group	NO	NO
red	1	single	YES	YES
red	0	group	NO	NO
red	1	group	YES	YES

	Predicted YES	Predicted NO
Actual YES	2	0
Actual NO	1	2

ROC space

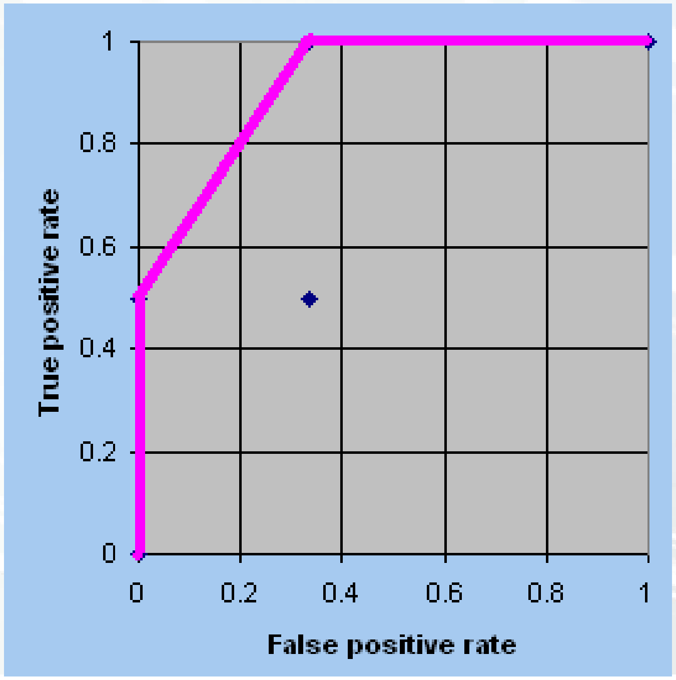
	Predicted YES	Predicted NO
Actual YES	2	0
Actual NO	1	2

- True positive rate TPr = 1
- False positive rate FPr = 1/3



ROC convex hull

cap-color	ring-number	population	EDIBLE	DT1 (50%)	DT2 (90%)	DT3 (20%)	YES	NO
brown	1	single	NO	YES	NO	YES	YES	NO
green	0	group	NO	NO	NO	NO	YES	NO
red	1	single	YES	NO	NO	YES	YES	NO
red	0	group	NO	NO	NO	NO	YES	NO
red	1	group	YES	YES	YES	YES	YES	NO

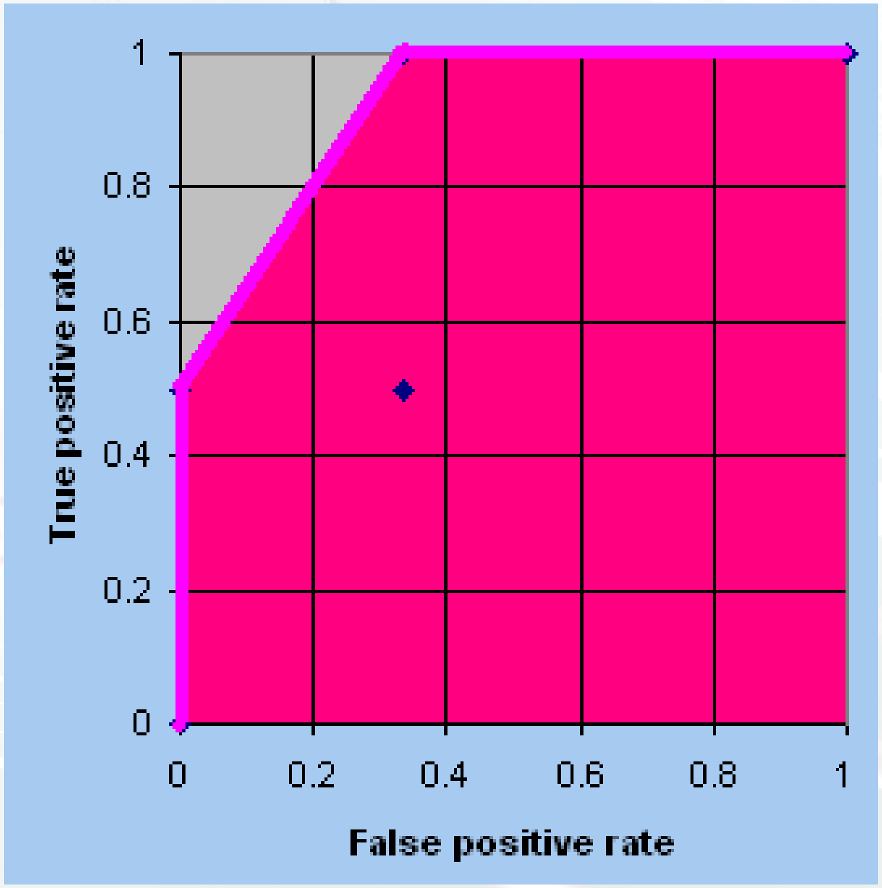


AUC – Area Under Curve

AUC =

= (0.5+1)/2*1/3+2/3

= 0.917


True positive rate
False positive rate