CLASSIFICATION – NAIVE BAYES

NIKOLA MILIKIĆ
nikola.milikic@fon.bg.ac.rs

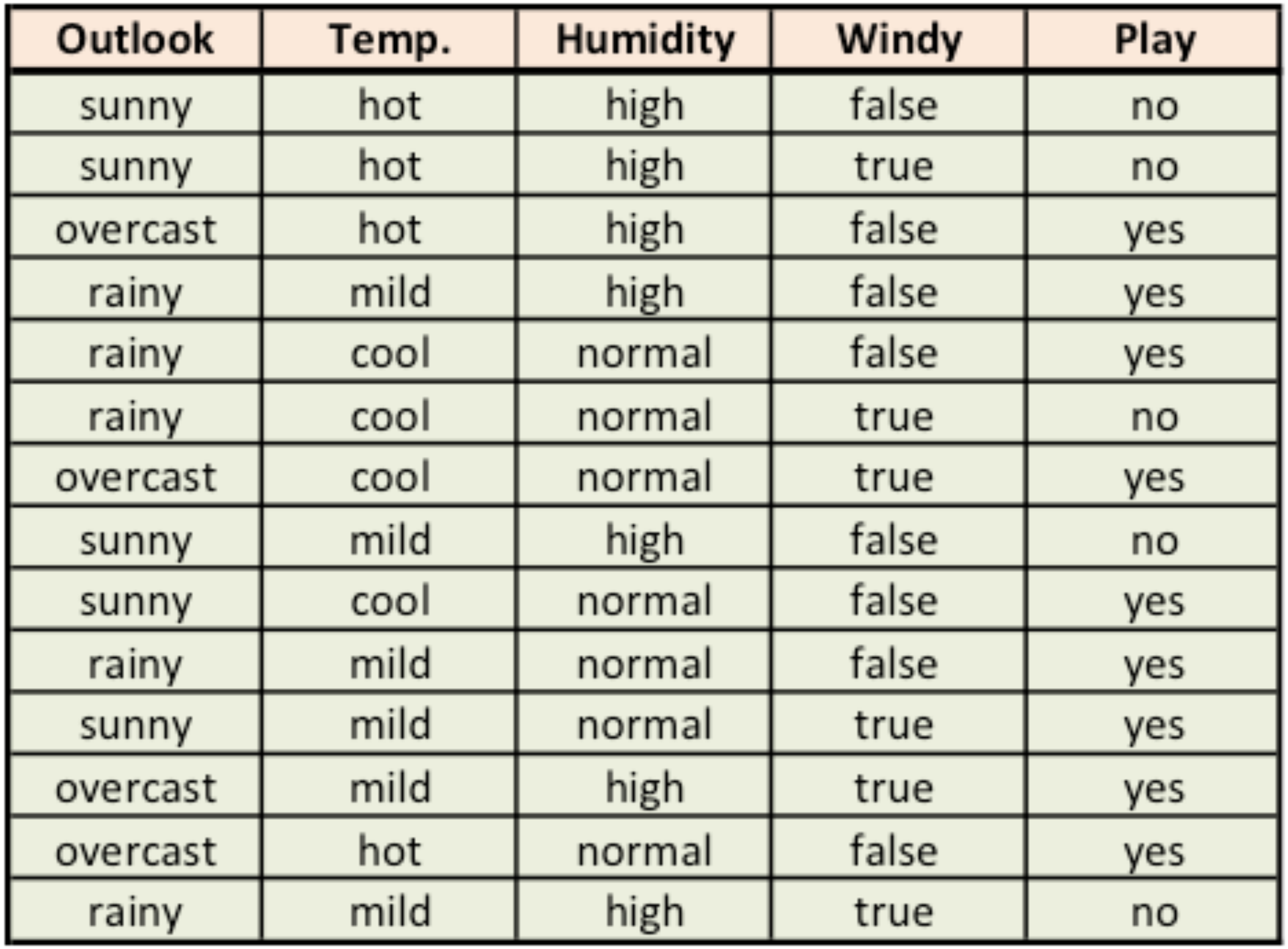
UROŠ KRČADINAC
uros@krcadinac.com

WHAT IS CLASSIFICATION?

- **A supervised learning task of determining the class of an instance; it is assumed that:**
  - feature values for the given instance are known
  - the set of possible classes is known and given
- **Classes are given as nominal values; for instance:**
  - classification of email messages: spam, not-spam
  - classification of news articles: politics, sport, culture i sl.

Example 1

ToPlayOtNotToPlay.arff dataset

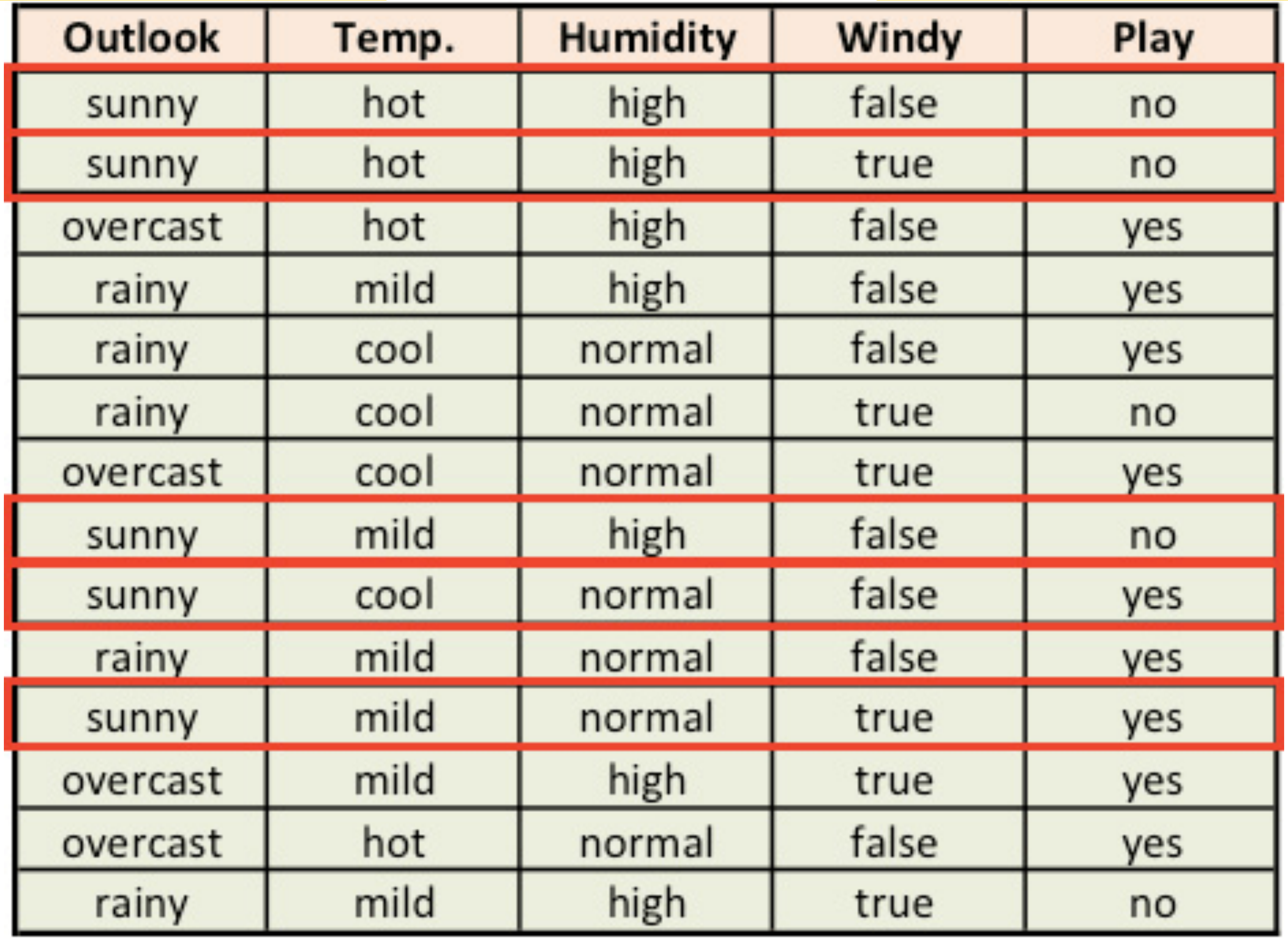


Outlook	Temp.	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Sunny weather

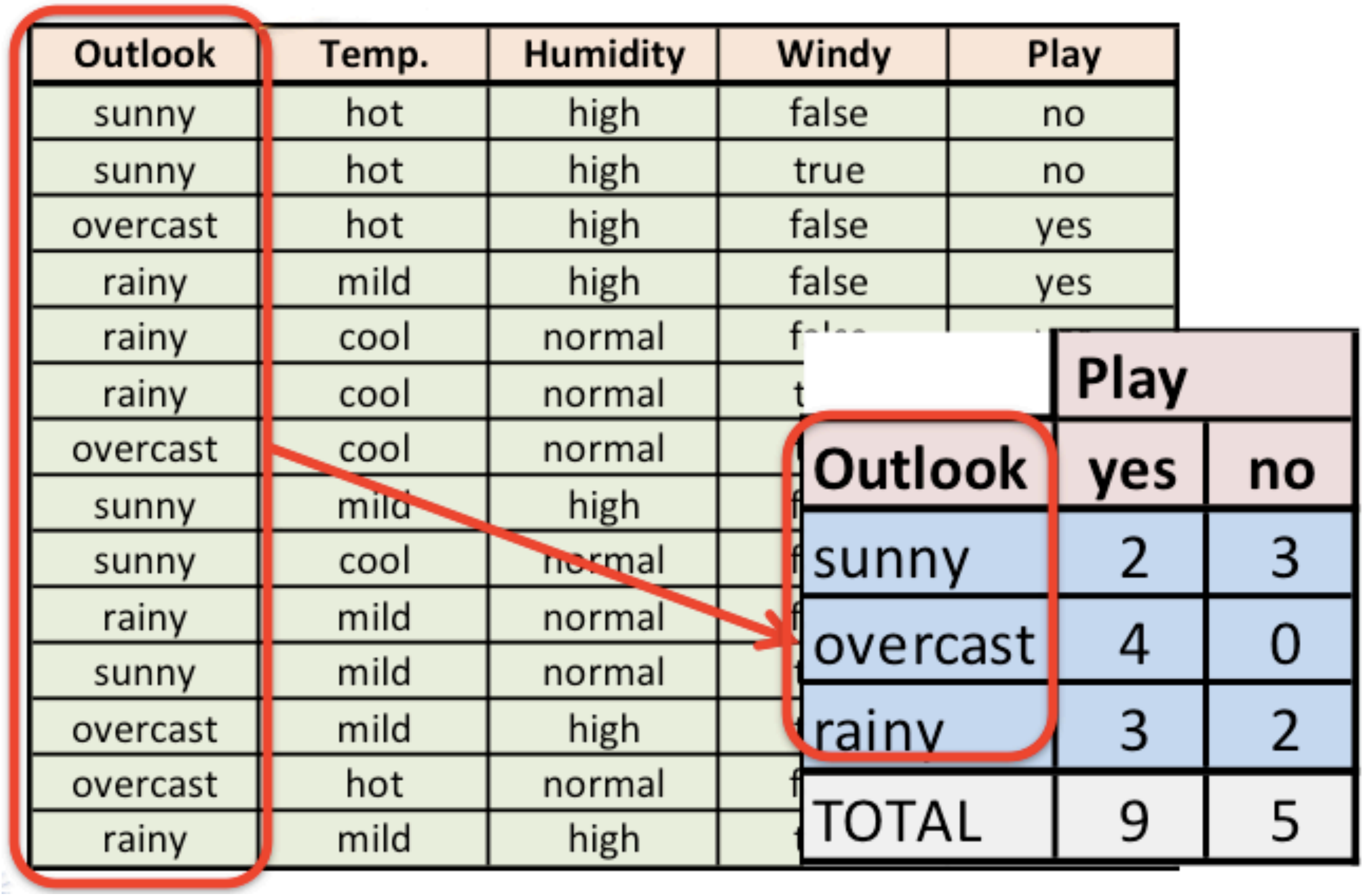
Suppose you know that it is sunny outside

Then 60% chance that Play = no



Outlook	Temp.	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

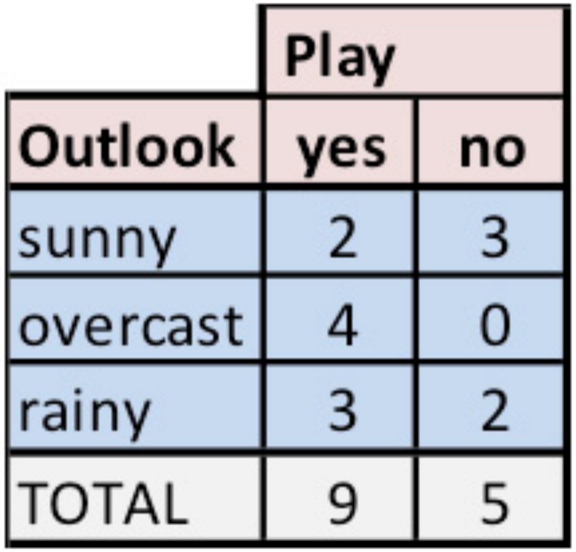
How well does outlook predict play?



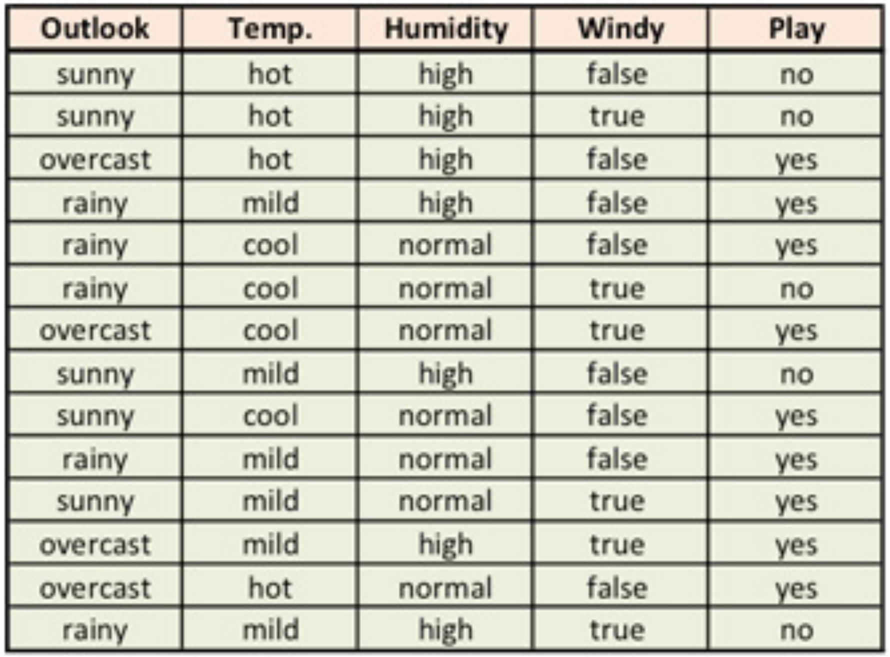
Outlook	Temp.	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	f	
rainy	cool	normal	t	
overcast	cool	normal		
sunny	mild	high	f	
sunny	cool	normal	f	
rainy	mild	normal	f	
sunny	mild	normal		
overcast	mild	high		
overcast	hot	normal	f	
rainy	mild	high		

	Play	
Outlook	yes	no
sunny	2	3
overcast	4	0
rainy	3	2
TOTAL	9	5

How well does outlook predict play?

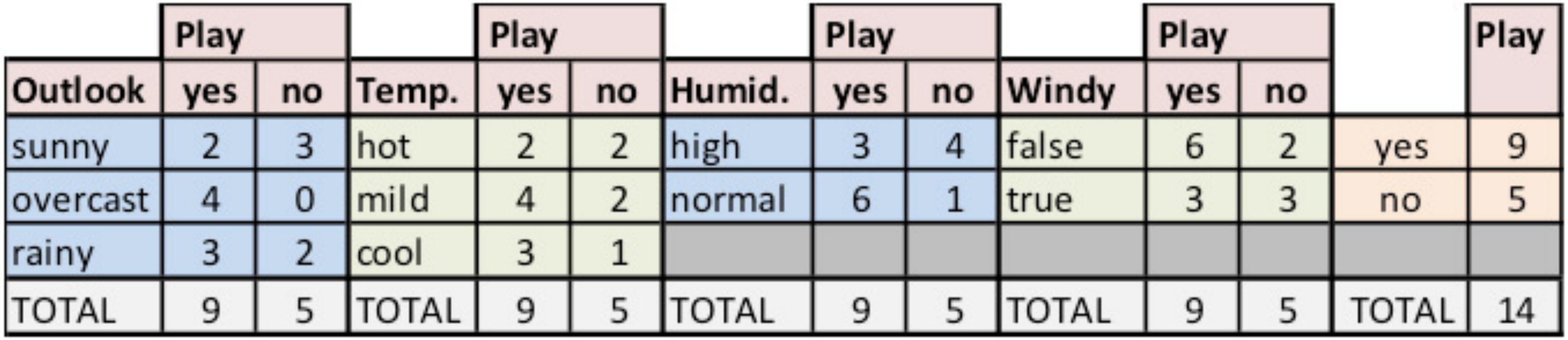


	Play	
Outlook	yes	no
sunny	2	3
overcast	4	0
rainy	3	2
TOTAL	9	5



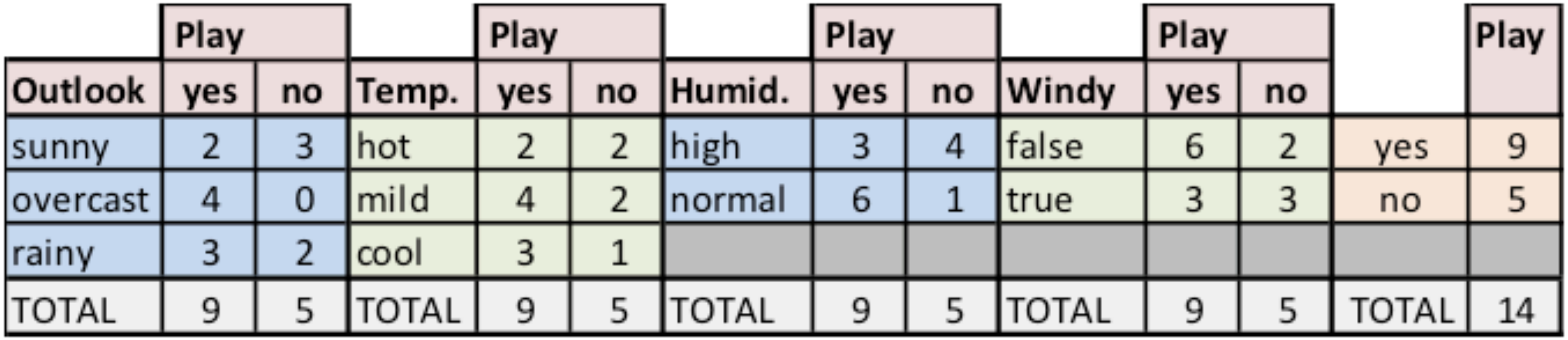
Outlook	Temp.	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

For each attribute…



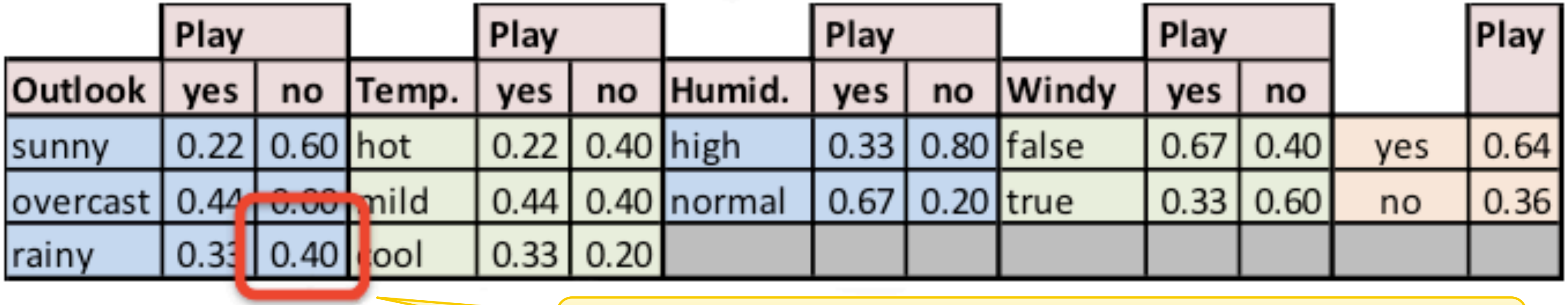
	Play			Play			Play			Play			Play
Outlook	yes	no	Temp.	yes	no	Humid.	yes	no	Windy	yes	no		
sunny	2	3	hot	2	2	high	3	4	false	6	2	yes	9
overcast	4	0	mild	4	2	normal	6	1	true	3	3	no	5
rainy	3	2	cool	3	1								
TOTAL	9	5	TOTAL	9	5	TOTAL	9	5	TOTAL	9	5	TOTAL	14

Values to ratios



	Play			Play			Play			Play			Play
Outlook	yes	no	Temp.	yes	no	Humid.	yes	no	Windy	yes	no		
sunny	2	3	hot	2	2	high	3	4	false	6	2	yes	9
overcast	4	0	mild	4	2	normal	6	1	true	3	3	no	5
rainy	3	2	cool	3	1								
TOTAL	9	5	TOTAL	9	5	TOTAL	9	5	TOTAL	9	5	TOTAL	14

Covert values to ratios



	Play			Play			Play			Play			Play
Outlook	yes	no	Temp.	yes	no	Humid.	yes	no	Windy	yes	no		
sunny	0.22	0.60	hot	0.22	0.40	high	0.33	0.80	false	0.67	0.40	yes	0.64
overcast	0.44	0.00	mild	0.44	0.40	normal	0.67	0.20	true	0.33	0.60	no	0.36
rainy	0.33	0.40	cool	0.33	0.20								

2 occurences of **Play = no**, where **Outlook = rainy**
5 occurrences of **Play = no**

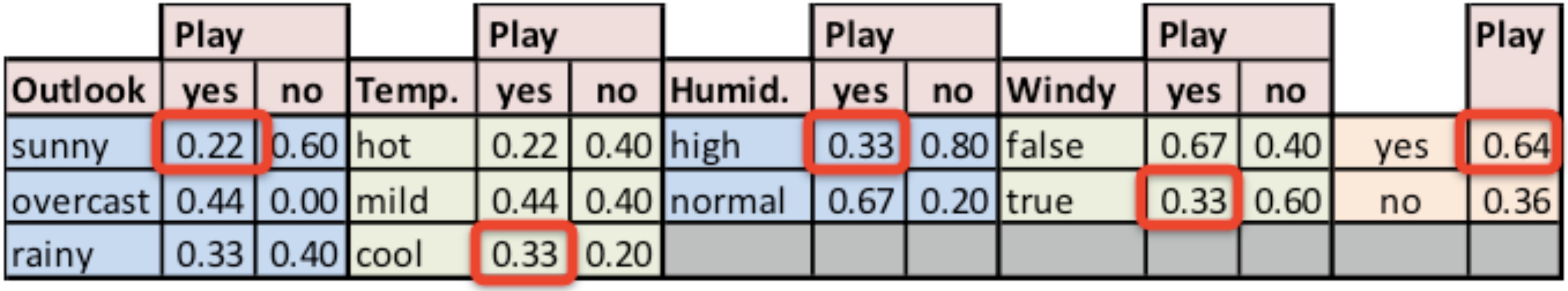
Likelihood of playing under these weather conditions

Calculate the likelihood that:

Outlook = sunny (0.22)
Temperature = cool (0.33)
Humidity = high (0.33)
Windy = true (0.33)
Play = yes (0.64)

$$0.22 \times 0.33 \times 0.33 \times 0.33 \times 0.64 = \mathbf{0.0053}$$

Likelihood of playing under these weather conditions



	Play			Play			Play			Play			Play
Outlook	yes	no	Temp.	yes	no	Humid.	yes	no	Windy	yes	no		
sunny	0.22	0.60	hot	0.22	0.40	high	0.33	0.80	false	0.67	0.40	yes	0.64
overcast	0.44	0.00	mild	0.44	0.40	normal	0.67	0.20	true	0.33	0.60	no	0.36
rainy	0.33	0.40	cool	0.33	0.20								

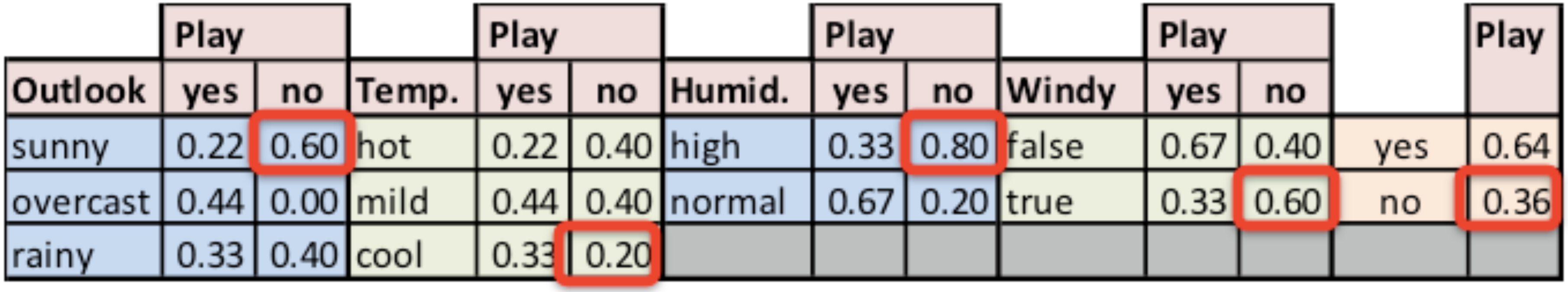
Likelihood of NOT playing under these weather conditions

Calculate the likelihood that:

Outlook = sunny (0.60)
Temperature = cool (0.20)
Humidity = high (0.80)
Windy = true (0.60)
Play = no (0.36)

0.60 x 0.20 x 0.80 x 0.60 x 0.36 = **0.0206**

Likelihood of NOT playing under these weather conditions



	Play			Play			Play			Play			Play
Outlook	yes	no	Temp.	yes	no	Humid.	yes	no	Windy	yes	no		
sunny	0.22	0.60	hot	0.22	0.40	high	0.33	0.80	false	0.67	0.40	yes	0.64
overcast	0.44	0.00	mild	0.44	0.40	normal	0.67	0.20	true	0.33	0.60	no	0.36
rainy	0.33	0.40	cool	0.33	0.20								

The Bayes Theorem

Given these weather conditions:

Outlook = sunny
Temperature = cool
Humidity = high
Windy = true

Probability of **Play = yes**: $\dfrac{0.0053}{0.0053 + 0.0206}$ = **20.5%**

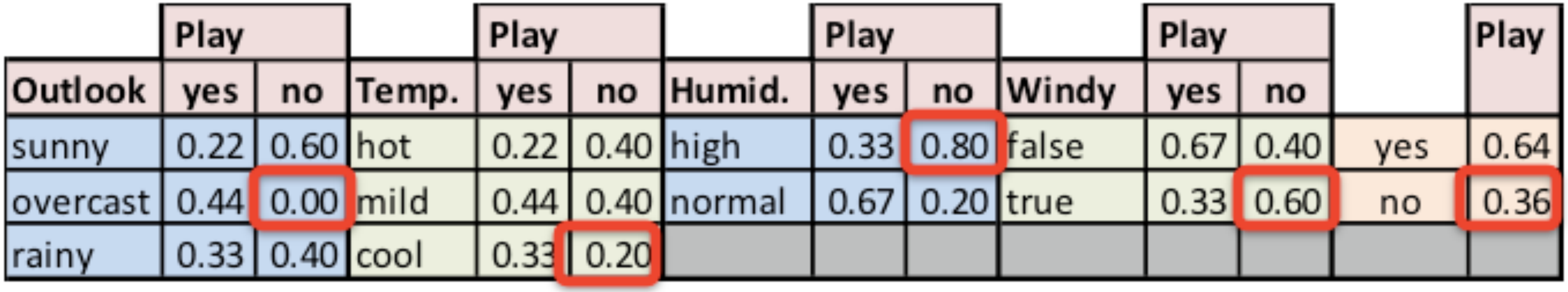
Probability of **Play = no**: $\dfrac{0.0206}{0.0053 + 0.0206}$ = **79.5%**

Likelihood of NOT playing under these weather conditions

Calculate the likelihood that:

Outlook = overcast (0.00)
Temperature = cool (0.20)
Humidity = high (0.80)
Windy = true (0.60)
Play = no (0.36)

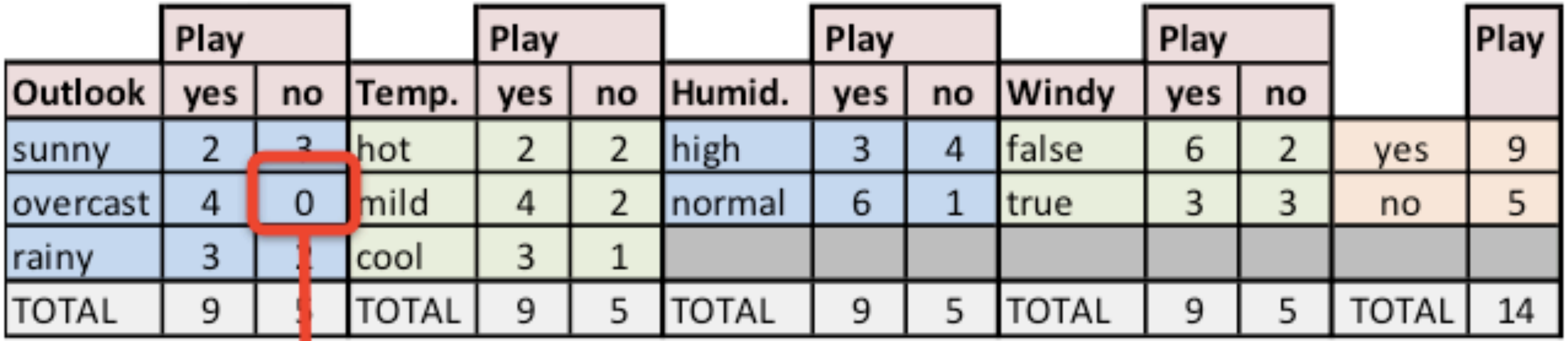
0.00 x 0.20 x 0.80 x 0.60 x 0.36 = **0.0000**



	Play			Play			Play			Play			Play
Outlook	yes	no	Temp.	yes	no	Humid.	yes	no	Windy	yes	no		
sunny	0.22	0.60	hot	0.22	0.40	high	0.33	0.80	false	0.67	0.40	yes	0.64
overcast	0.44	0.00	mild	0.44	0.40	normal	0.67	0.20	true	0.33	0.60	no	0.36
rainy	0.33	0.40	cool	0.33	0.20								

Laplace estimator

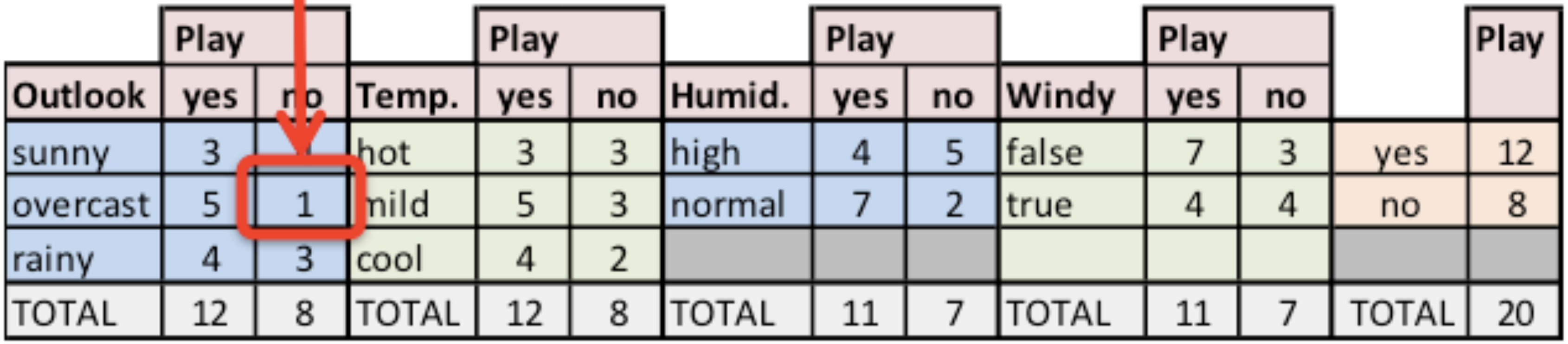
The original dataset



	Play			Play			Play			Play			Play
Outlook	yes	no	Temp.	yes	no	Humid.	yes	no	Windy	yes	no		
sunny	2	3	hot	2	2	high	3	4	false	6	2	yes	9
overcast	4	0	mild	4	2	normal	6	1	true	3	3	no	5
rainy	3	2	cool	3	1								
TOTAL	9	5	TOTAL	9	5	TOTAL	9	5	TOTAL	9	5	TOTAL	14

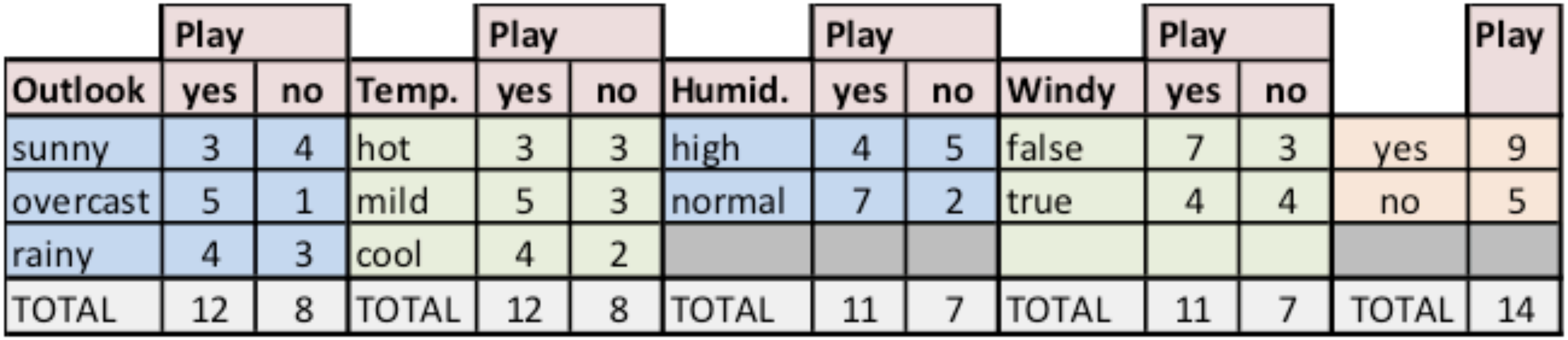
Laplace estimator:
Add 1 to each count

After the Laplace estimator



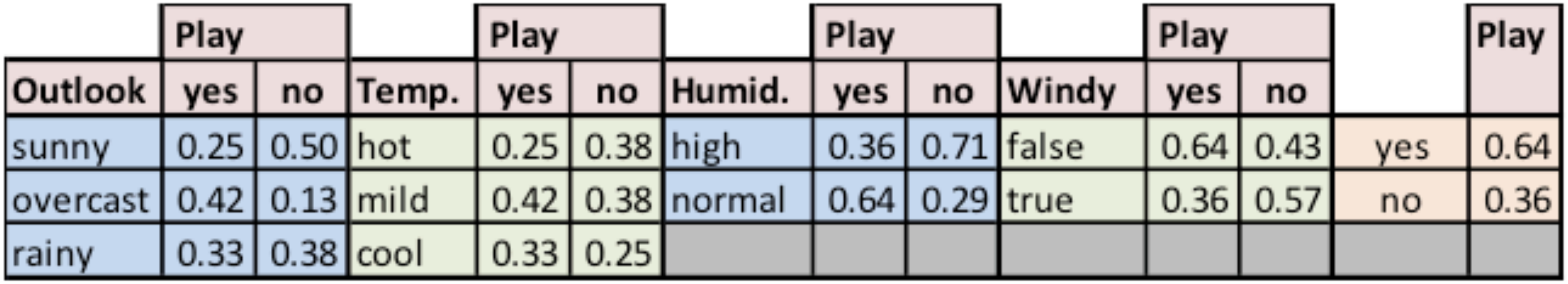
	Play			Play			Play			Play			Play
Outlook	yes	no	Temp.	yes	no	Humid.	yes	no	Windy	yes	no		
sunny	3	[illegible]	hot	3	3	high	4	5	false	7	3	yes	12
overcast	5	1	mild	5	3	normal	7	2	true	4	4	no	8
rainy	4	3	cool	4	2								
TOTAL	12	8	TOTAL	12	8	TOTAL	11	7	TOTAL	11	7	TOTAL	20

Laplace estimator



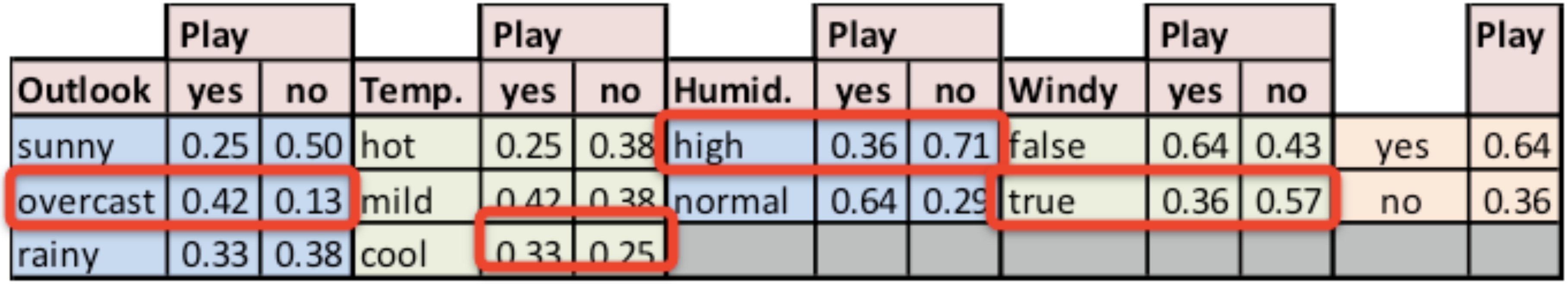
	Play			Play			Play			Play			Play
Outlook	yes	no	Temp.	yes	no	Humid.	yes	no	Windy	yes	no		
sunny	3	4	hot	3	3	high	4	5	false	7	3	yes	9
overcast	5	1	mild	5	3	normal	7	2	true	4	4	no	5
rainy	4	3	cool	4	2								
TOTAL	12	8	TOTAL	12	8	TOTAL	11	7	TOTAL	11	7	TOTAL	14

Convert incremented counts to ratios after implementing the Laplace estimator



	Play			Play			Play			Play			Play
Outlook	yes	no	Temp.	yes	no	Humid.	yes	no	Windy	yes	no		
sunny	0.25	0.50	hot	0.25	0.38	high	0.36	0.71	false	0.64	0.43	yes	0.64
overcast	0.42	0.13	mild	0.42	0.38	normal	0.64	0.29	true	0.36	0.57	no	0.36
rainy	0.33	0.38	cool	0.33	0.25								

Laplace estimator



	Play			Play			Play			Play			Play
Outlook	yes	no	Temp.	yes	no	Humid.	yes	no	Windy	yes	no		
sunny	0.25	0.50	hot	0.25	0.38	high	0.36	0.71	false	0.64	0.43	yes	0.64
overcast	0.42	0.13	mild	0.42	0.38	normal	0.64	0.29	true	0.36	0.57	no	0.36
rainy	0.33	0.38	cool	0.33	0.25								

Outlook = ovecast, Temperature = cool, Humidity = high, Windy = true

Play = no: 0.13 x 0.25 x 0.71 x 0.57 x 0.36 = 0.046
Play = yes: 0.42 x 0.33 x 0.36 x 0.36 x 0.64 = 0.0118

Probability of **Play = no**: $\dfrac{0.0046}{0.0046 + 0.0118}$ = **28%**

Probability of **Play = yes**: $\dfrac{0.0118}{0.0046 + 0.0118}$ = **72%**

Laplace estimator

Under these weather conditions:

Outlook = sunny
Temperature = cool
Humidity = high
Windy = true

NOT using Laplace estimator:

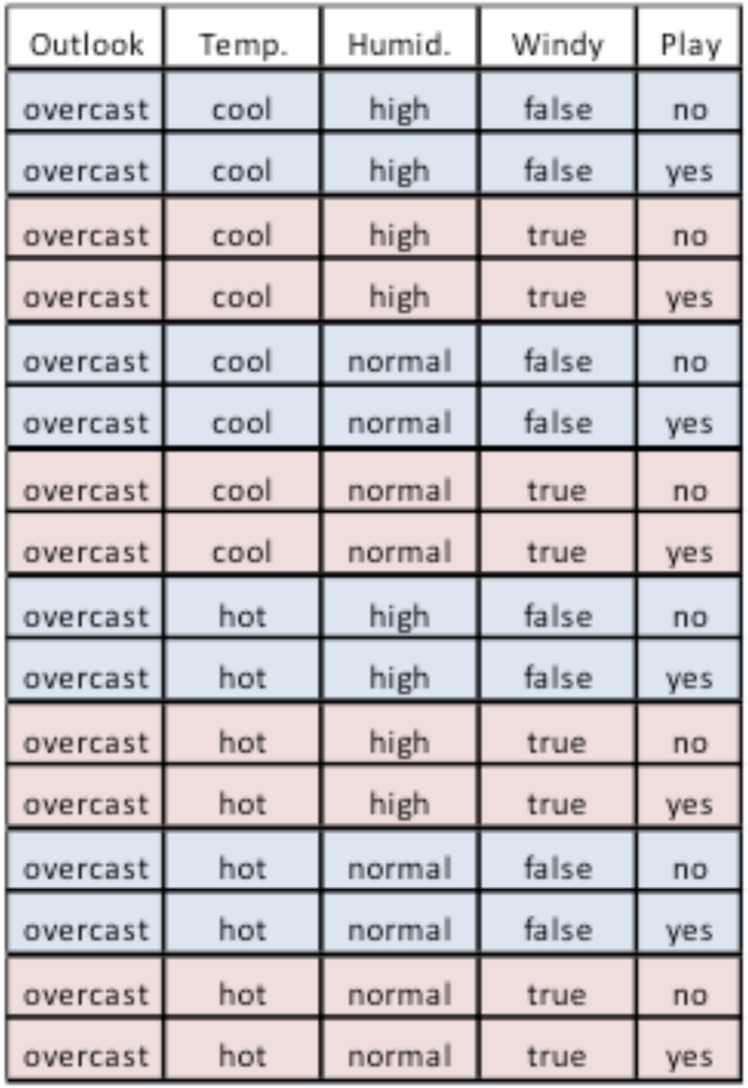
Play = no: 79.5%
Play = yes: 20.5%

Using Laplace estimator:

$\boxed{\text{Play = no: 72.0\%}}$
Play = yes: 28.0%

The effect of **Laplace estimator** has little effect as sample size grows.

Prediction rules



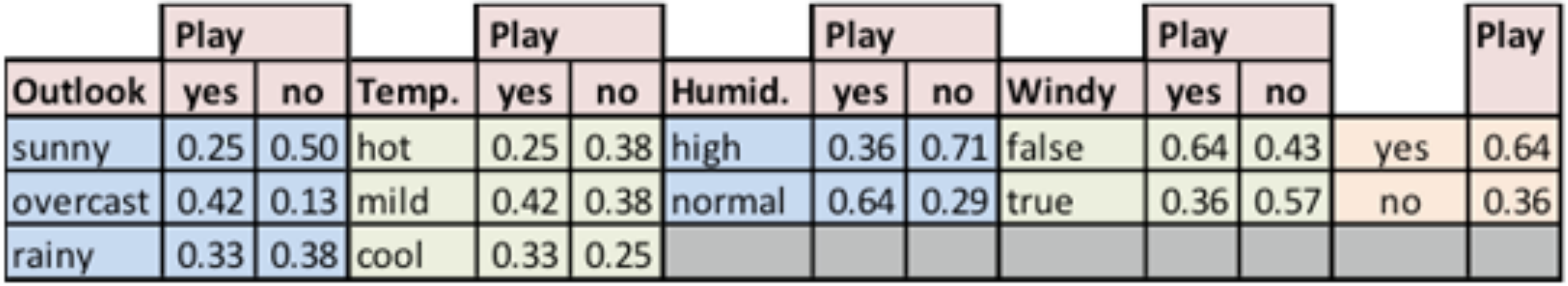
Outlook	Temp.	Humid.	Windy	Play
overcast	cool	high	false	no
overcast	cool	high	false	yes
overcast	cool	high	true	no
overcast	cool	high	true	yes
overcast	cool	normal	false	no
overcast	cool	normal	false	yes
overcast	cool	normal	true	no
overcast	cool	normal	true	yes
overcast	hot	high	false	no
overcast	hot	high	false	yes
overcast	hot	high	true	no
overcast	hot	high	true	yes
overcast	hot	normal	false	no
overcast	hot	normal	false	yes
overcast	hot	normal	true	no
overcast	hot	normal	true	yes

Repeat previous calculation for all other combinations of weather conditions.

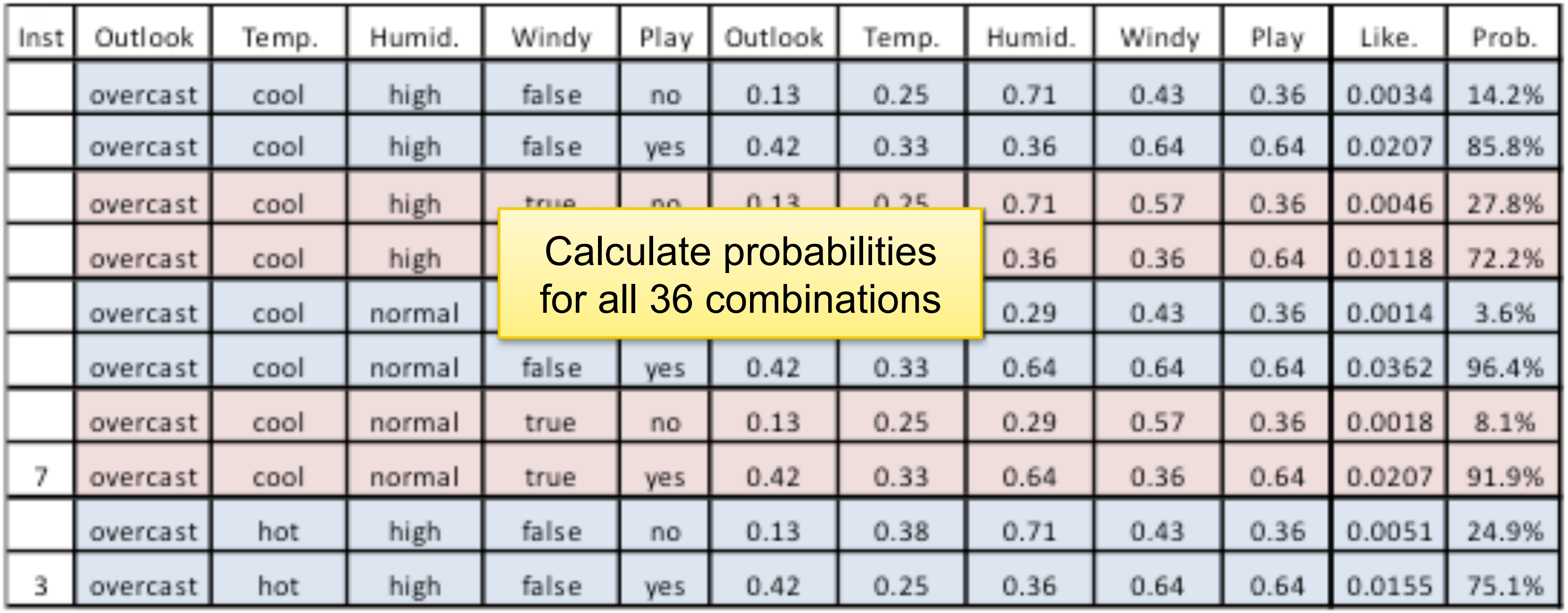
Calculate the rules for each pair.

Then throw out the rules with $p < 0.5$

Prediction rules



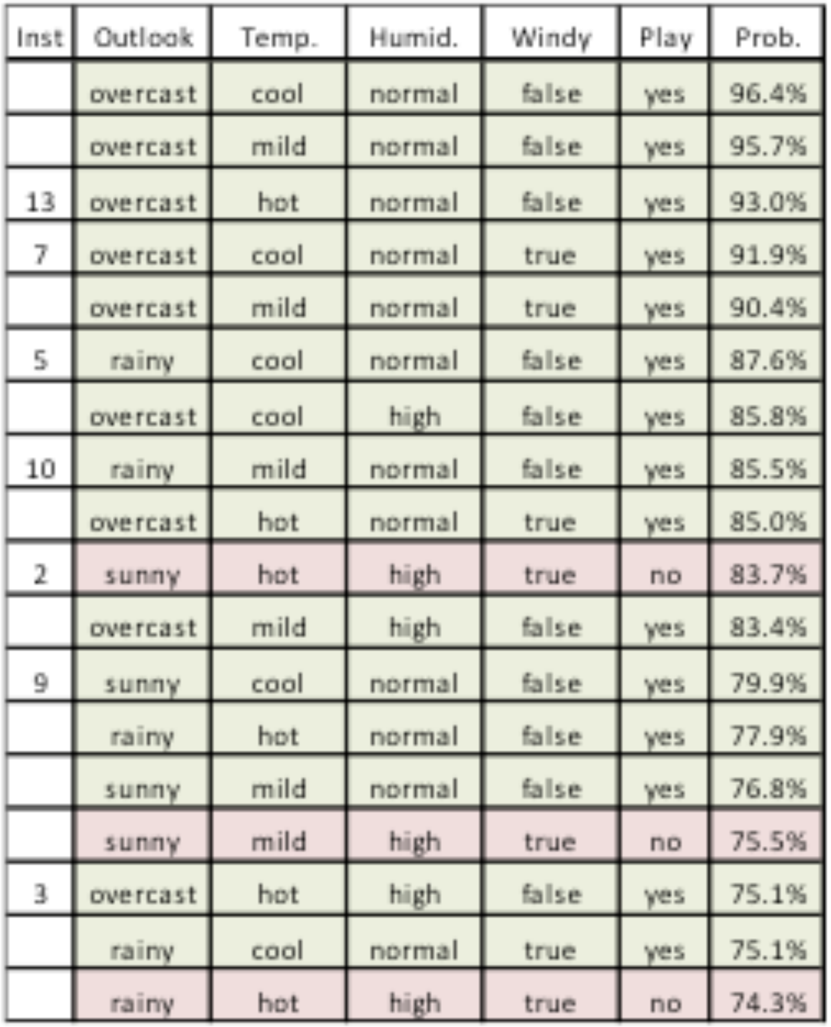
	Play			Play			Play			Play			Play
Outlook	yes	no	Temp.	yes	no	Humid.	yes	no	Windy	yes	no		
sunny	0.25	0.50	hot	0.25	0.38	high	0.36	0.71	false	0.64	0.43	yes	0.64
overcast	0.42	0.13	mild	0.42	0.38	normal	0.64	0.29	true	0.36	0.57	no	0.36
rainy	0.33	0.38	cool	0.33	0.25								



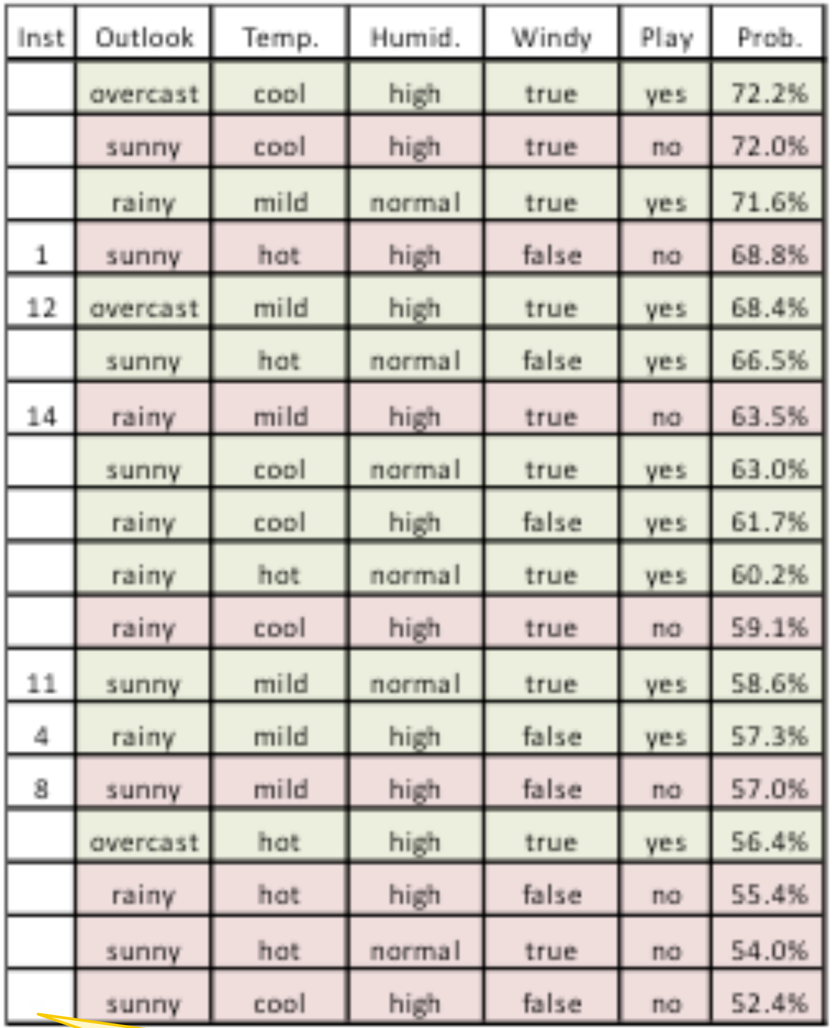
Inst	Outlook	Temp.	Humid.	Windy	Play	Outlook	Temp.	Humid.	Windy	Play	Like.	Prob.
	overcast	cool	high	false	no	0.13	0.25	0.71	0.43	0.36	0.0034	14.2%
	overcast	cool	high	false	yes	0.42	0.33	0.36	0.64	0.64	0.0207	85.8%
	overcast	cool	high	true	no	0.13	0.25	0.71	0.57	0.36	0.0046	27.8%
	overcast	cool	high	[illegible]	[illegible]	[illegible]	[illegible]	0.36	0.36	0.64	0.0118	72.2%
	overcast	cool	normal	[illegible]	[illegible]	[illegible]	[illegible]	0.29	0.43	0.36	0.0014	3.6%
	overcast	cool	normal	false	yes	0.42	0.33	0.64	0.64	0.64	0.0362	96.4%
	overcast	cool	normal	true	no	0.13	0.25	0.29	0.57	0.36	0.0018	8.1%
7	overcast	cool	normal	true	yes	0.42	0.33	0.64	0.36	0.64	0.0207	91.9%
	overcast	hot	high	false	no	0.13	0.38	0.71	0.43	0.36	0.0051	24.9%
3	overcast	hot	high	false	yes	0.42	0.25	0.36	0.64	0.64	0.0155	75.1%

Calculate probabilities for all 36 combinations

Prediction rules



Inst	Outlook	Temp.	Humid.	Windy	Play	Prob.
	overcast	cool	normal	false	yes	96.4%
	overcast	mild	normal	false	yes	95.7%
13	overcast	hot	normal	false	yes	93.0%
7	overcast	cool	normal	true	yes	91.9%
	overcast	mild	normal	true	yes	90.4%
5	rainy	cool	normal	false	yes	87.6%
	overcast	cool	high	false	yes	85.8%
10	rainy	mild	normal	false	yes	85.5%
	overcast	hot	normal	true	yes	85.0%
2	sunny	hot	high	true	no	83.7%
	overcast	mild	high	false	yes	83.4%
9	sunny	cool	normal	false	yes	79.9%
	rainy	hot	normal	false	yes	77.9%
	sunny	mild	normal	false	yes	76.8%
	sunny	mild	high	true	no	75.5%
3	overcast	hot	high	false	yes	75.1%
	rainy	cool	normal	true	yes	75.1%
	rainy	hot	high	true	no	74.3%

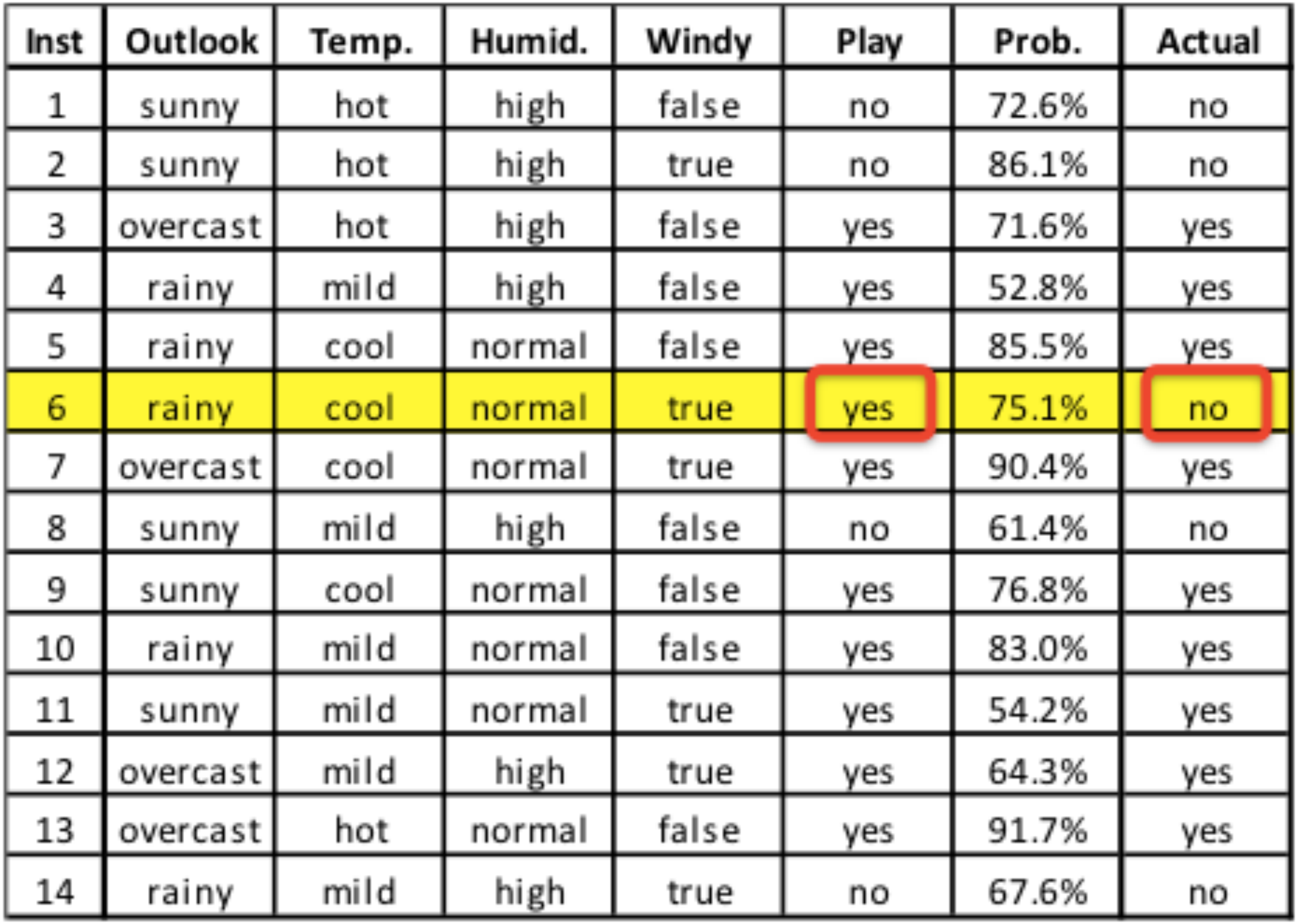


Inst	Outlook	Temp.	Humid.	Windy	Play	Prob.
	overcast	cool	high	true	yes	72.2%
	sunny	cool	high	true	no	72.0%
	rainy	mild	normal	true	yes	71.6%
1	sunny	hot	high	false	no	68.8%
12	overcast	mild	high	true	yes	68.4%
	sunny	hot	normal	false	yes	66.5%
14	rainy	mild	high	true	no	63.5%
	sunny	cool	normal	true	yes	63.0%
	rainy	cool	high	false	yes	61.7%
	rainy	hot	normal	true	yes	60.2%
	rainy	cool	high	true	no	59.1%
11	sunny	mild	normal	true	yes	58.6%
4	rainy	mild	high	false	yes	57.3%
8	sunny	mild	high	false	no	57.0%
	overcast	hot	high	true	yes	56.4%
	rainy	hot	high	false	no	55.4%
	sunny	hot	normal	true	no	54.0%
	sunny	cool	high	false	no	52.4%

Rules predicting class for all combinations of attributes

The instance 6 is missing

Comparing the prediction with the original data



Inst	Outlook	Temp.	Humid.	Windy	Play	Prob.	Actual
1	sunny	hot	high	false	no	72.6%	no
2	sunny	hot	high	true	no	86.1%	no
3	overcast	hot	high	false	yes	71.6%	yes
4	rainy	mild	high	false	yes	52.8%	yes
5	rainy	cool	normal	false	yes	85.5%	yes
6	rainy	cool	normal	true	yes	75.1%	no
7	overcast	cool	normal	true	yes	90.4%	yes
8	sunny	mild	high	false	no	61.4%	no
9	sunny	cool	normal	false	yes	76.8%	yes
10	rainy	mild	normal	false	yes	83.0%	yes
11	sunny	mild	normal	true	yes	54.2%	yes
12	overcast	mild	high	true	yes	64.3%	yes
13	overcast	hot	normal	false	yes	91.7%	yes
14	rainy	mild	high	true	no	67.6%	no

Weka

- Waikato Environment for Knowledge Analysis
- Java Software for data mining
- Set of algorithms for machine learning and data mining
- Developed at the University of Waikato, New Zealand
- Open-source
- Website: http://www.cs.waikato.ac.nz/ml/weka

Datasets we use

- We use datasets from the Technology Forge:

http://www.technologyforge.net/Datasets

ARFF file

- Attribut-Relation File Format – ARFF
- Text file

Attributes could be:
- Numerical
- Nominal

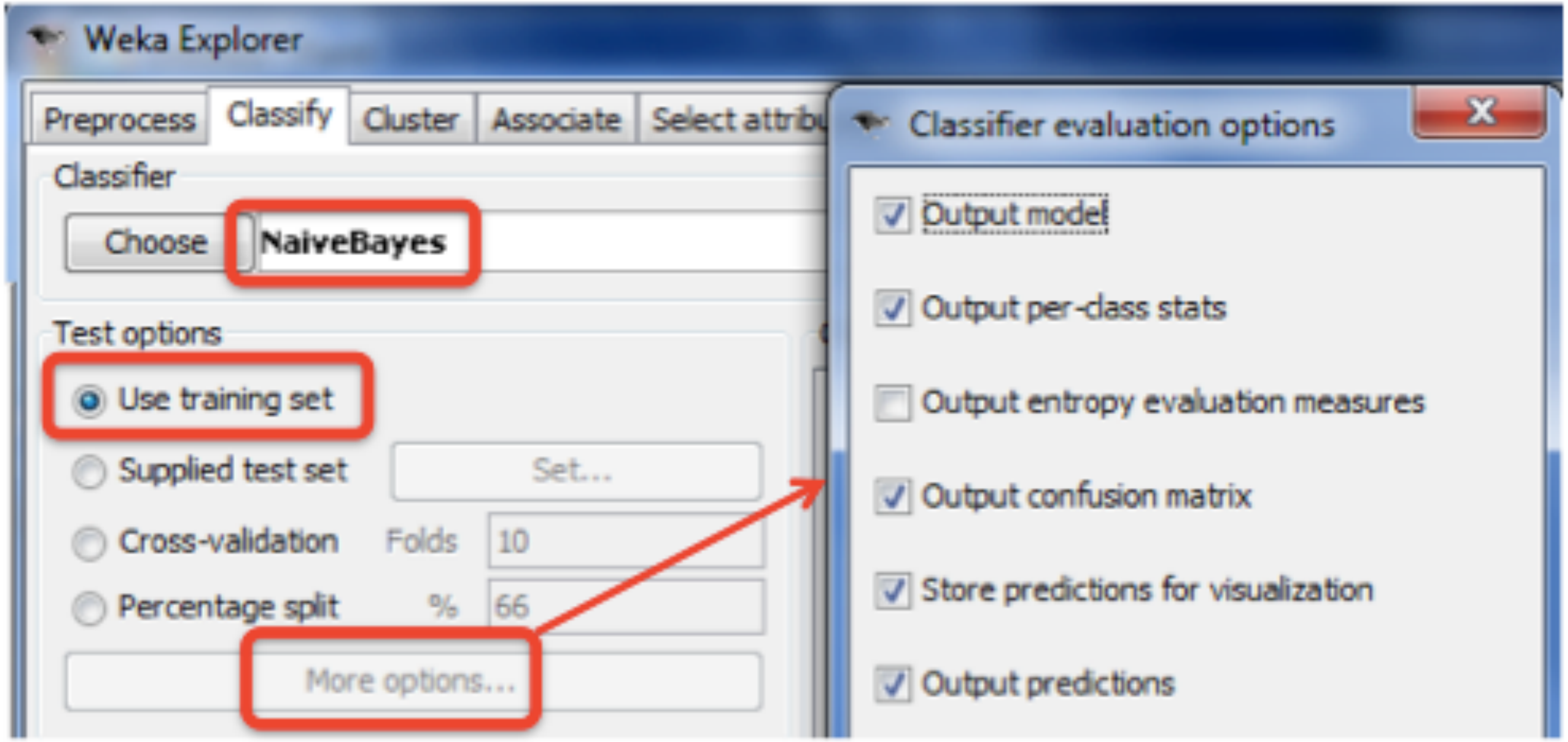
```
@relation TPONTPNom

@attribute Outlook {sunny, overcast, rainy}
@attribute Temp. {hot, mild, cool}
@attribute Humidity {high, normal}
@attribute Windy {'false', 'true'}
@attribute Play {no, yes}

@data
sunny, hot, high, 'false', no
sunny, hot, high, 'true', no
overcast, hot, high, 'false', yes
...
```

Classification in Weka

ToPlayOtNotToPlay.arff dataset



Classifier

NaiveBayes

Test options

● Use training set

○ Supplied test set

○ Cross-validation Folds 10

○ Percentage split % 66

☑ Output model

☑ Output per-class stats

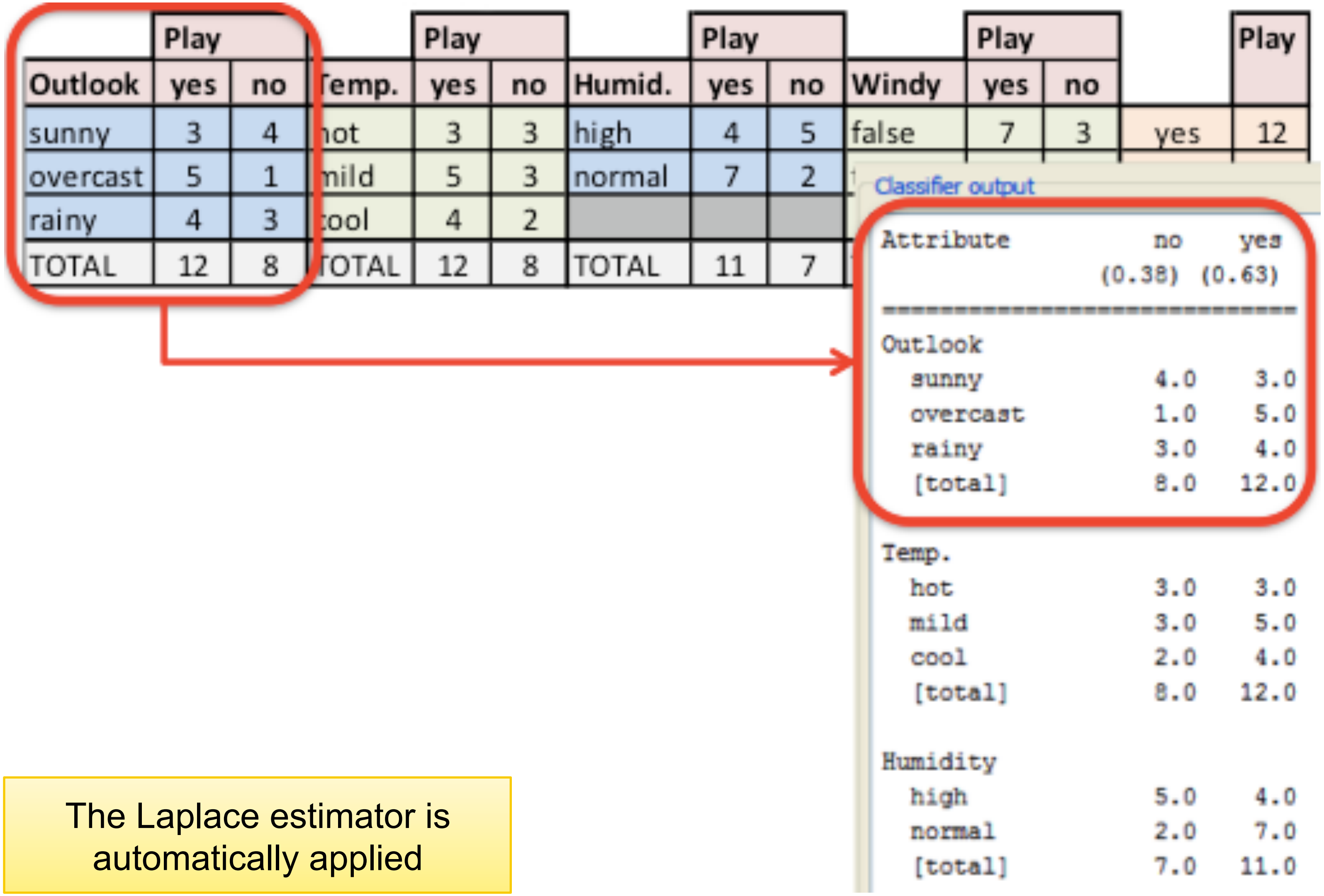
□ Output entropy evaluation measures

☑ Output confusion matrix

☑ Store predictions for visualization

☑ Output predictions

Classification results



	Play	
Outlook	yes	no
sunny	3	4
overcast	5	1
rainy	4	3
TOTAL	12	8

	Play	
Temp.	yes	no
hot	3	3
mild	5	3
cool	4	2
TOTAL	12	8

	Play	
Humid.	yes	no
high	4	5
normal	7	2
TOTAL	11	7

	Play	
Windy	yes	no
false	7	3

	Play
yes	12

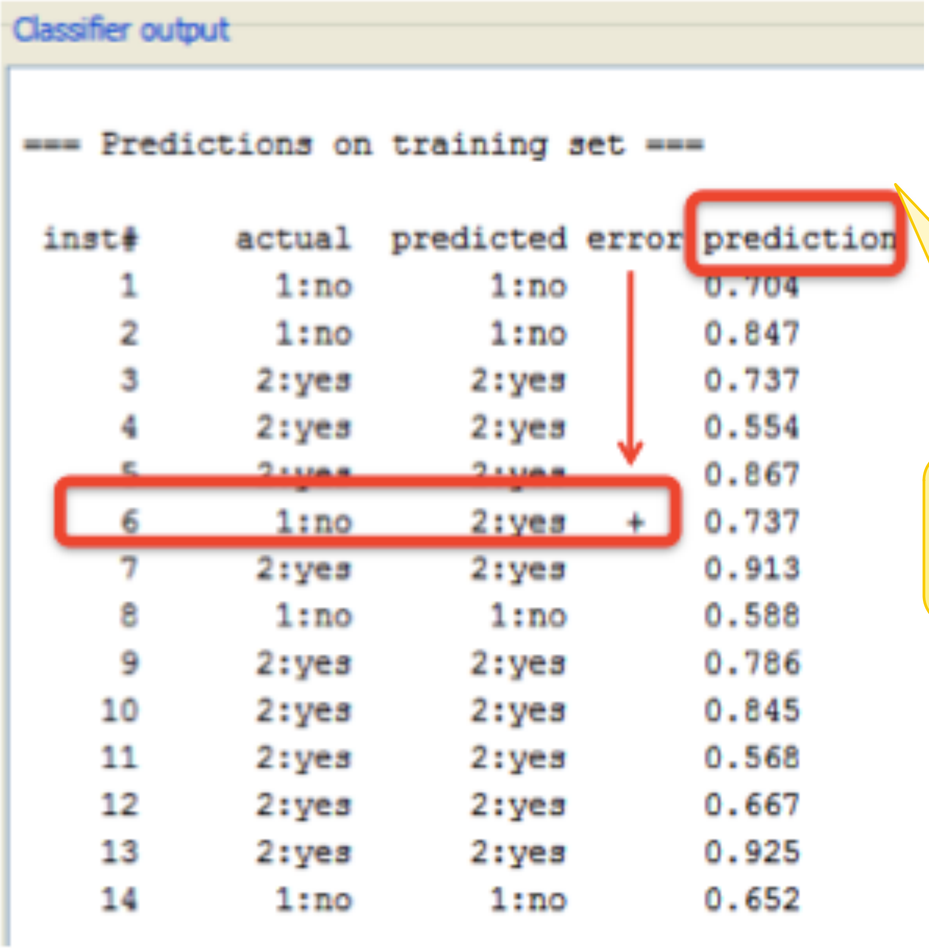
```
Attribute                  no      yes
                       (0.38)   (0.63)
======================================
Outlook
  sunny                   4.0      3.0
  overcast                1.0      5.0
  rainy                   3.0      4.0
  [total]                 8.0     12.0

Temp.
  hot                     3.0      3.0
  mild                    3.0      5.0
  cool                    2.0      4.0
  [total]                 8.0     12.0

Humidity
  high                    5.0      4.0
  normal                  2.0      7.0
  [total]                 7.0     11.0
```

The Laplace estimator is automatically applied

Classification results



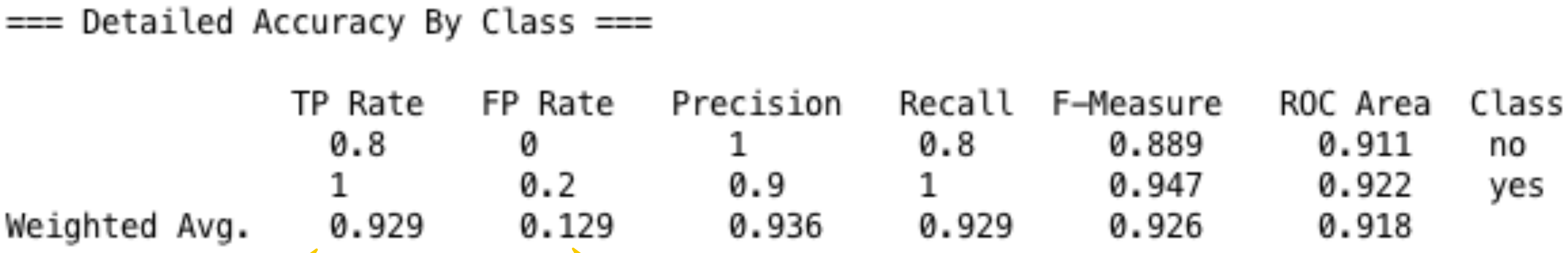
```
=== Predictions on training set ===

 inst#     actual  predicted error prediction
     1       1:no       1:no       0.704
     2       1:no       1:no       0.847
     3      2:yes      2:yes       0.737
     4      2:yes      2:yes       0.554
     5      2:yes      2:yes       0.867
     6       1:no      2:yes   +   0.737
     7      2:yes      2:yes       0.913
     8       1:no       1:no       0.588
     9      2:yes      2:yes       0.786
    10      2:yes      2:yes       0.845
    11      2:yes      2:yes       0.568
    12      2:yes      2:yes       0.667
    13      2:yes      2:yes       0.925
    14       1:no       1:no       0.652
```

Instance 6 is marked as a wrong identified instance

Probability of each instance in the dataset

Precision, Recall, and F-Measure



```
=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.8       0          1         0.8      0.889      0.911    no
                   1         0.2        0.9       1        0.947      0.922    yes
Weighted Avg.      0.929     0.129      0.936     0.929    0.926      0.918
```

True Positives Rate

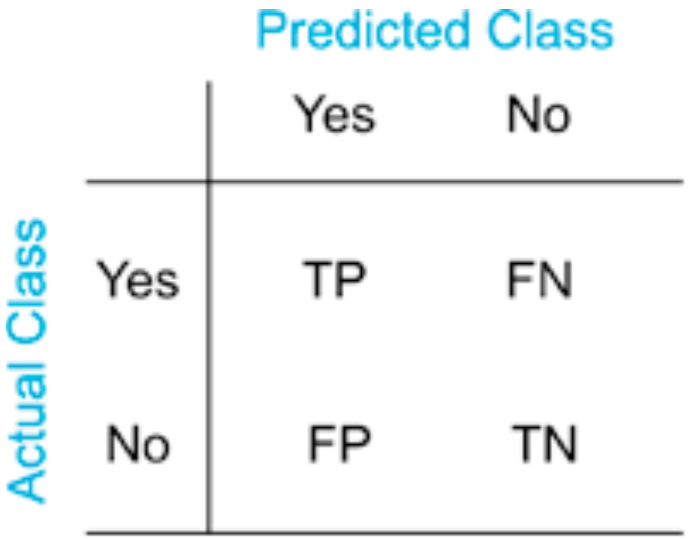
False Positives Rate

Precision = $\frac{TP}{(TP + FP)}$

Recall = $\frac{TP}{(TP + NP)}$

F measure = $\frac{2 * Precision * Recall}{Precision + Recall}$

Confusion Matrix



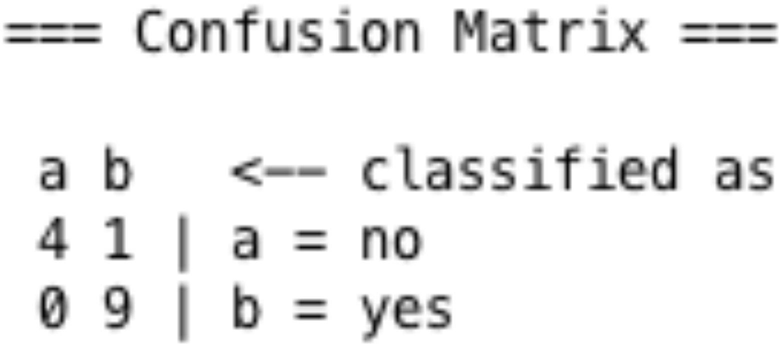
		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

TP = True Positive

FP = False Positive

TN = True Negative

FN = False Negative



```
=== Confusion Matrix ===

 a b   <-- classified as
 4 1 | a = no
 0 9 | b = yes
```

Example 2 – Eatable Mushrooms dataset

- Eatable Mushrooms dataset based on “National Audubon Society Field Guide to North American Mushrooms”
- Hypothetical samples with descriptions corresponding to 23 species of mushrooms
- There are 8124 instances with 22 nominal attributes which describe mushroom characteristics; one of which is whether a mushroom is eatable or not
- Our goal is to predict whether a mushroom is eatable or not

Thank you!

Weka Tutorials and Assignments @ The Technology Forge

- Link: http://www.technologyforge.net/WekaTutorials/

(Anonymous) survey for your comments and suggestions

http://goo.gl/cqdp3I

ANY QUESTIONS?

NIKOLA MILIKIĆ
nikola.milikic@fon.bg.ac.rs

UROŠ KRČADINAC
uros@krcadinac.com