Naïve Bayes Classification

Instructor: Arindam Chaudhuri
Indian Institute of Quantitative Finance

Bayes rule

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$$

- H – hypothesis
- E – evidence related to the hypothesis H, i.e., the data to be used for validating (accepting/rejecting) the hypothesis H
- P(H) – probability of the hypothesis (prior probability)
- P(E) – probability of the evidence i.e., the state of the world described by the gathered data
- P(E|H) – (conditional) probability of evidence E given that the hypothesis H holds
- P(H|E) – (conditional) probability of the hypothesis H given the evidence E

Naive Bayes classifier

- Lets make an assumption that all attributes are mutually independent:

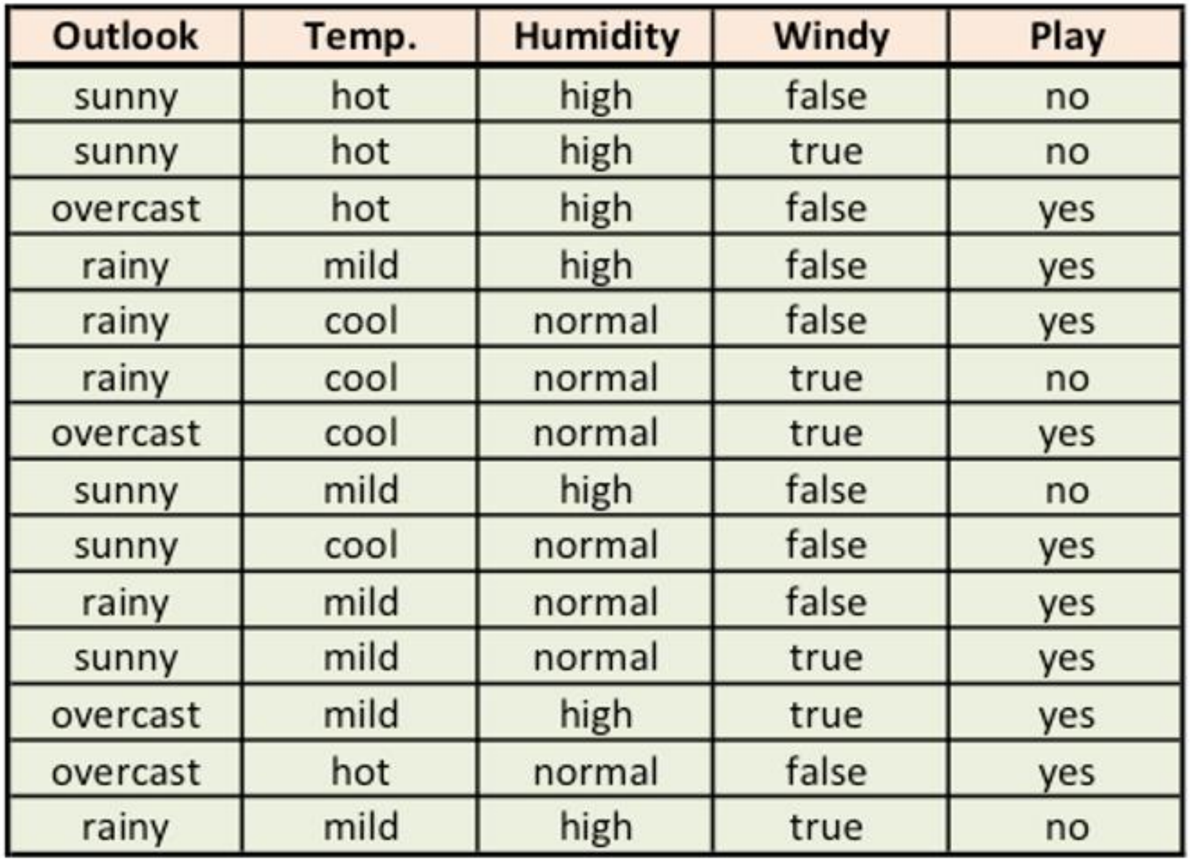
$$P(H|E) = \frac{P(E_1|H) * P(E_2|H) * \ldots * P(E_n|H) * P(H)}{P(E)}$$

Naive Bayes

- Makes two “naïve” assumptions over attributes:
    - all attributes are a priori equally important
    - all attributes are statistically independent (value of one attribute is not related to a value of another attribute)

- This assumptions mostly are not true, but in practice the algorithm gives good results

Example: Predicting whether a theater play will be performed



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

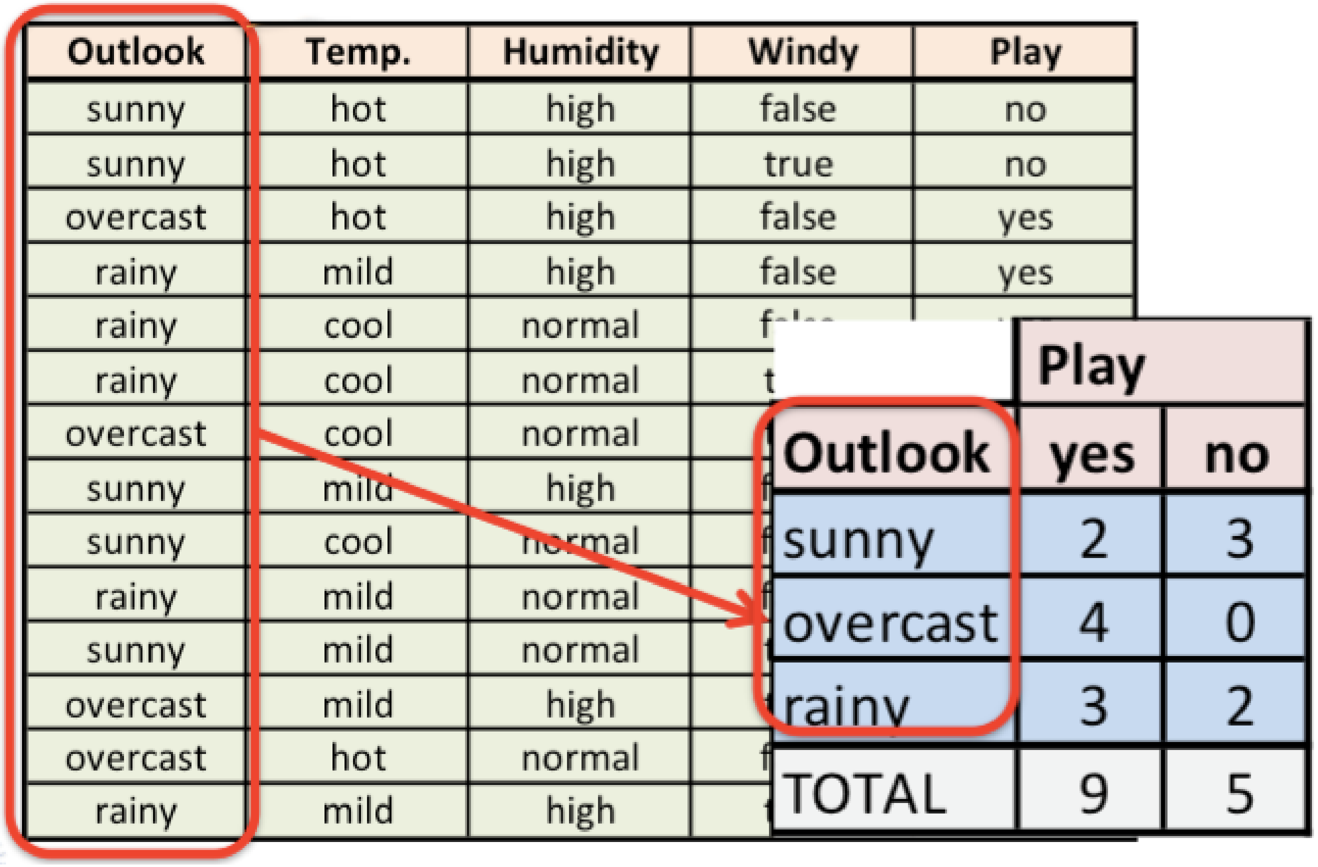
Sunny weather

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

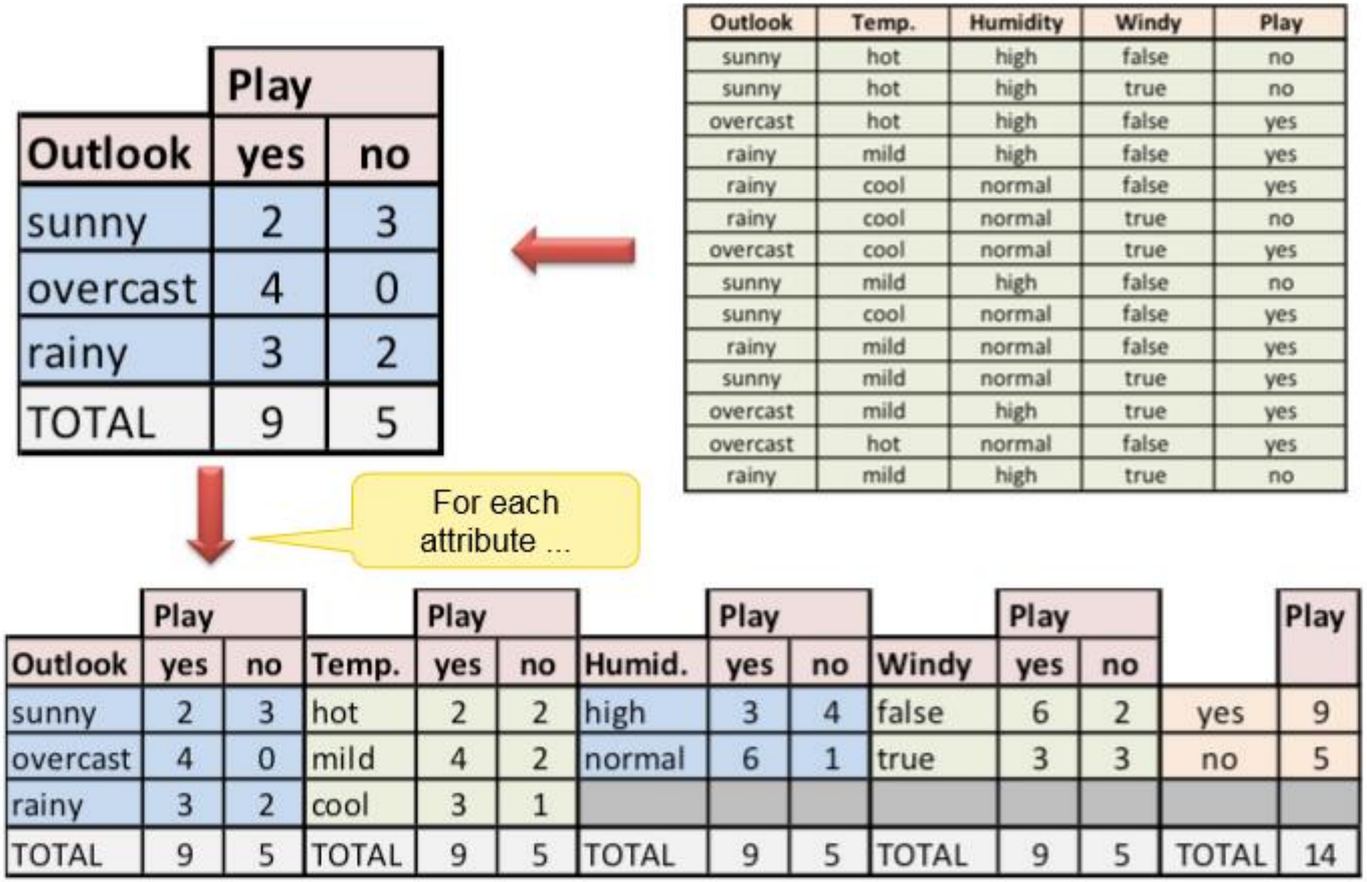
How well does outlook predict play?



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	[illegible]
rainy	cool	normal	t	[illegible]
overcast	cool	normal	[illegible]	[illegible]
sunny	mild	high	f	[illegible]
sunny	cool	normal	f	[illegible]
rainy	mild	normal	f	[illegible]
sunny	mild	normal	t	[illegible]
overcast	mild	high	t	[illegible]
overcast	hot	normal	f	[illegible]
rainy	mild	high	t	[illegible]

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

How well does outlook predict play?



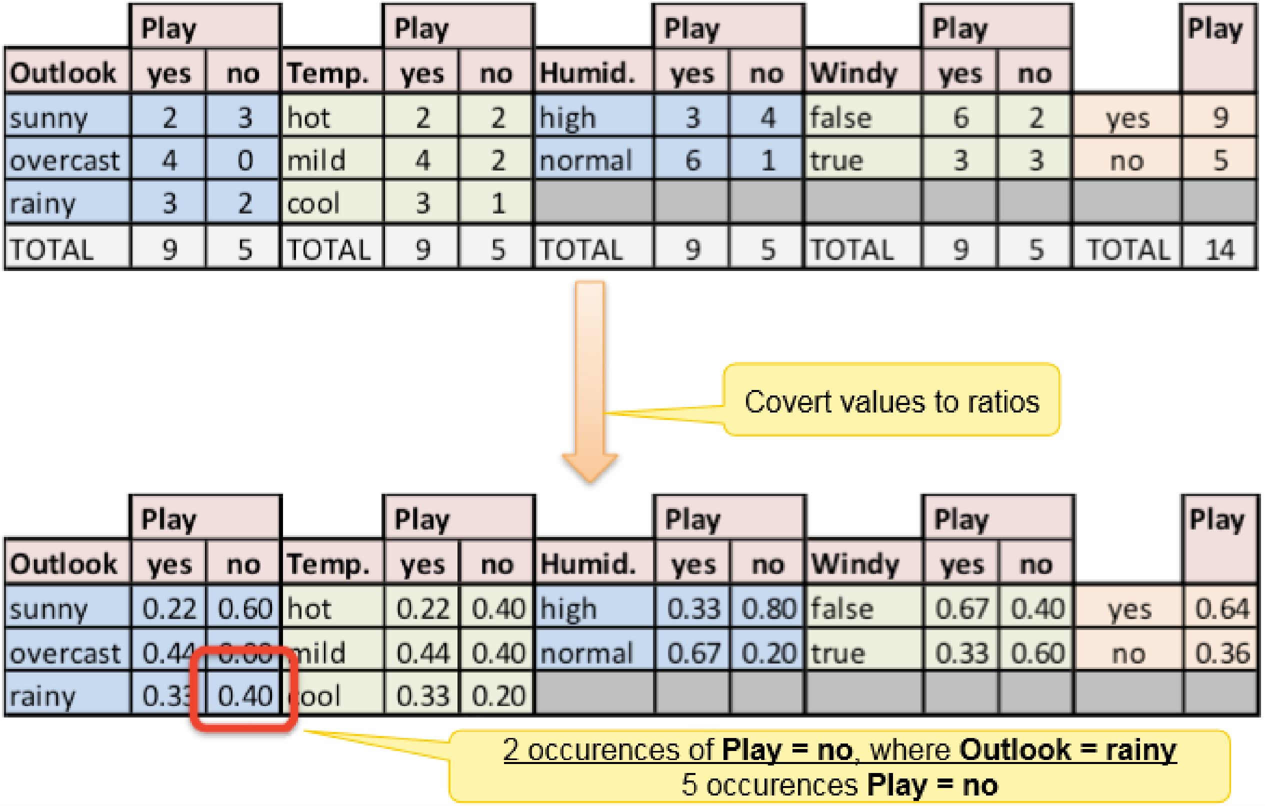
Outlook	Temp.	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny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

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

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 occurences **Play = no**

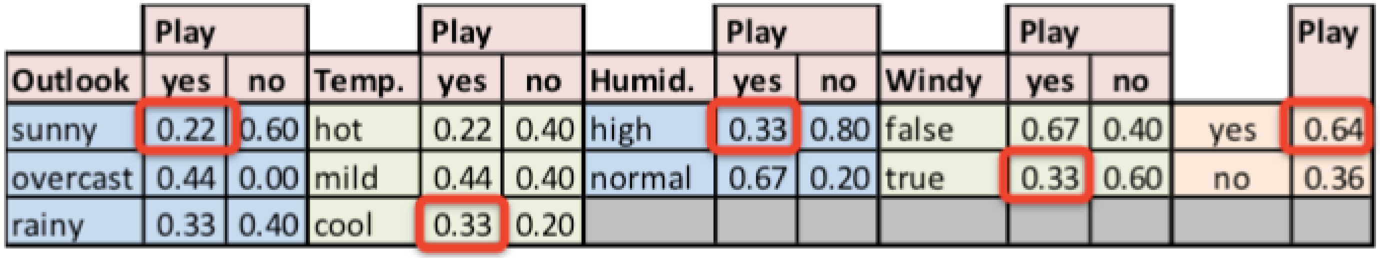
Likelihood of playing under these weather conditions

Calculate the likelihood that:

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

Likelihood of playing under these weather conditions

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



	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								

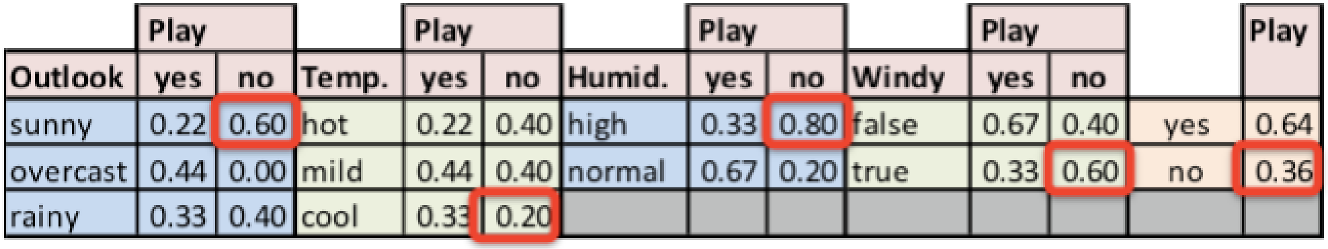
Likelihood of NOT playing under these weather conditions

Calculate the likelihood that:

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

Likelihood of **NOT** playing under these weather conditions

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



	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**: $$\frac{0.0053}{0.0053 + 0.0206} = \mathbf{20.5\%}$$

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

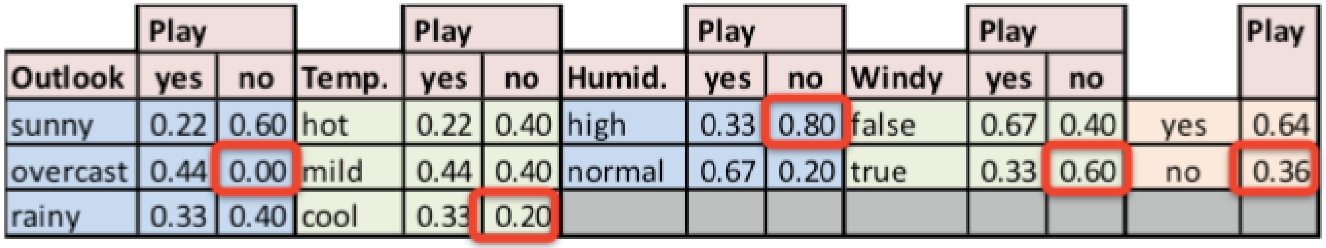
$$P(H|E) = \frac{P(E_1|H) * P(E_2|H) * \ldots * P(E_n|H) * P(H)}{P(E)}$$

Likelihood of NOT playing under these weather conditions

Calculate the likelihood that:

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

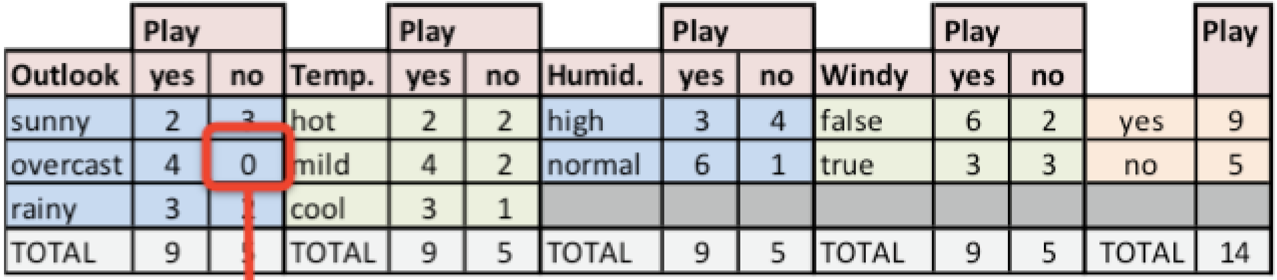
0.00 x 0.20 x 0.80 x 0.60 x 0.36 = **0.0000**



	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

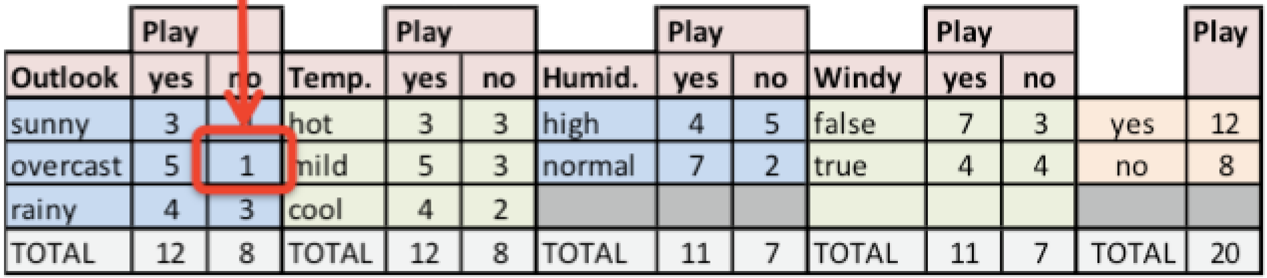
The original dataset



	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

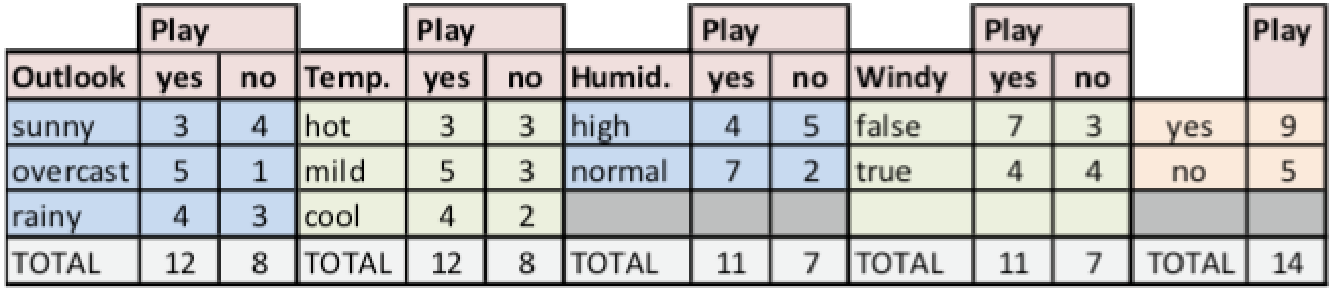
Laplace estimator:
Add 1 to each count

After the Laplace estimator



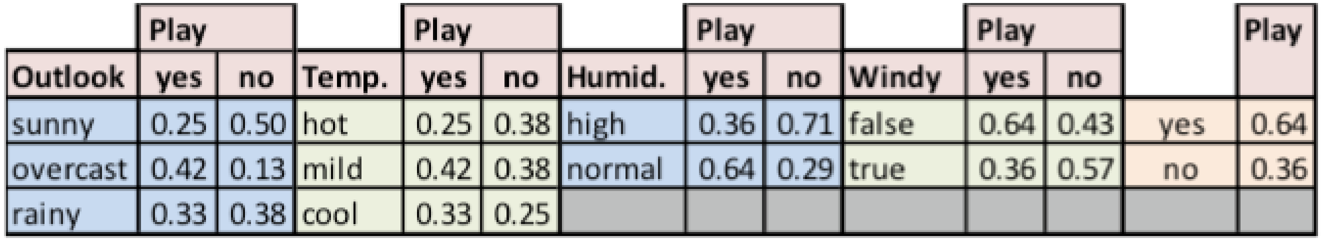
	Play			Play			Play			Play			Play
Outlook	yes	no	Temp.	yes	no	Humid.	yes	no	Windy	yes	no		
sunny	3	[illegible]	hot	3	3	high	4	5	false	7	3	yes	12
overcast	5	1	mild	5	3	normal	7	2	true	4	4	no	8
rainy	4	3	cool	4	2								
TOTAL	12	8	TOTAL	12	8	TOTAL	11	7	TOTAL	11	7	TOTAL	20

Laplace estimator



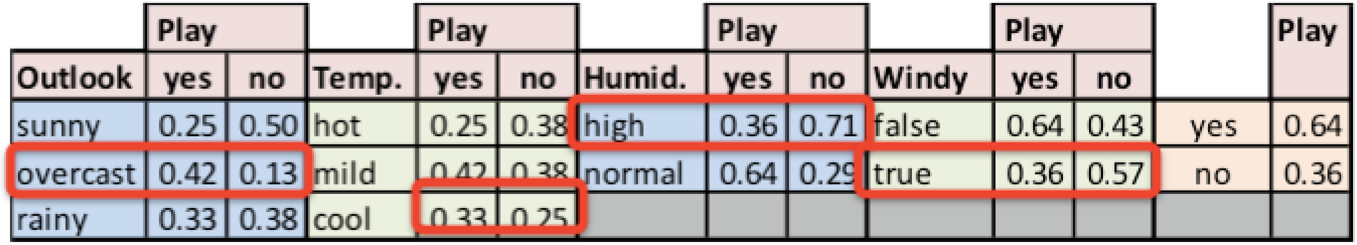
	Play			Play			Play			Play			Play
Outlook	yes	no	Temp.	yes	no	Humid.	yes	no	Windy	yes	no		
sunny	3	4	hot	3	3	high	4	5	false	7	3	yes	9
overcast	5	1	mild	5	3	normal	7	2	true	4	4	no	5
rainy	4	3	cool	4	2								
TOTAL	12	8	TOTAL	12	8	TOTAL	11	7	TOTAL	11	7	TOTAL	14

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**: $$\frac{0.0046}{0.0046 + 0.0118} = \mathbf{28\%}$$

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

Laplace estimator

Under these weather conditions:

Temperature = cool
Humidity = high
Windy = true

NOT using Laplace estimator:

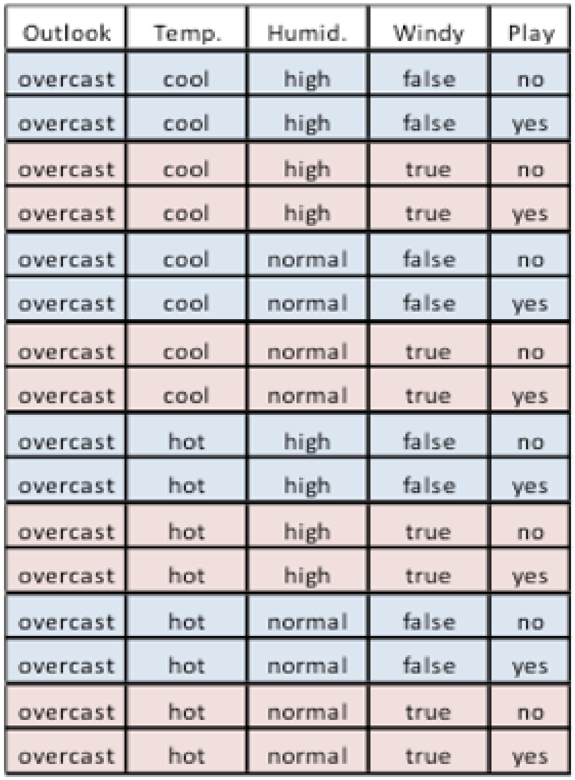
Play = no: 79.5%
Play = yes: 20.5%

Using Laplace estimator:

Play = no: 72.0%
Play = yes: 28.0%

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

Prediction rules



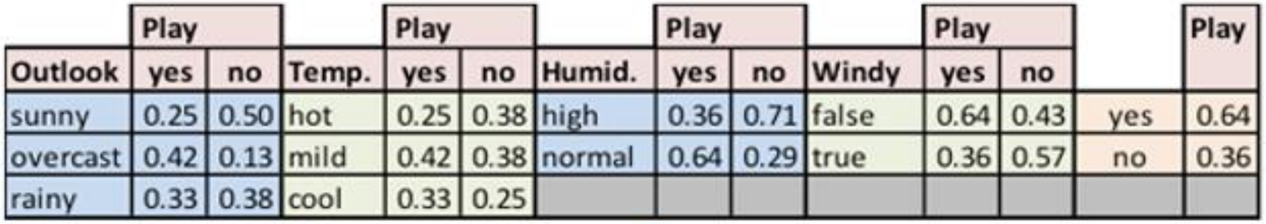
Outlook	Temp.	Humid.	Windy	Play
overcast	cool	high	false	no
overcast	cool	high	false	yes
overcast	cool	high	true	no
overcast	cool	high	true	yes
overcast	cool	normal	false	no
overcast	cool	normal	false	yes
overcast	cool	normal	true	no
overcast	cool	normal	true	yes
overcast	hot	high	false	no
overcast	hot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

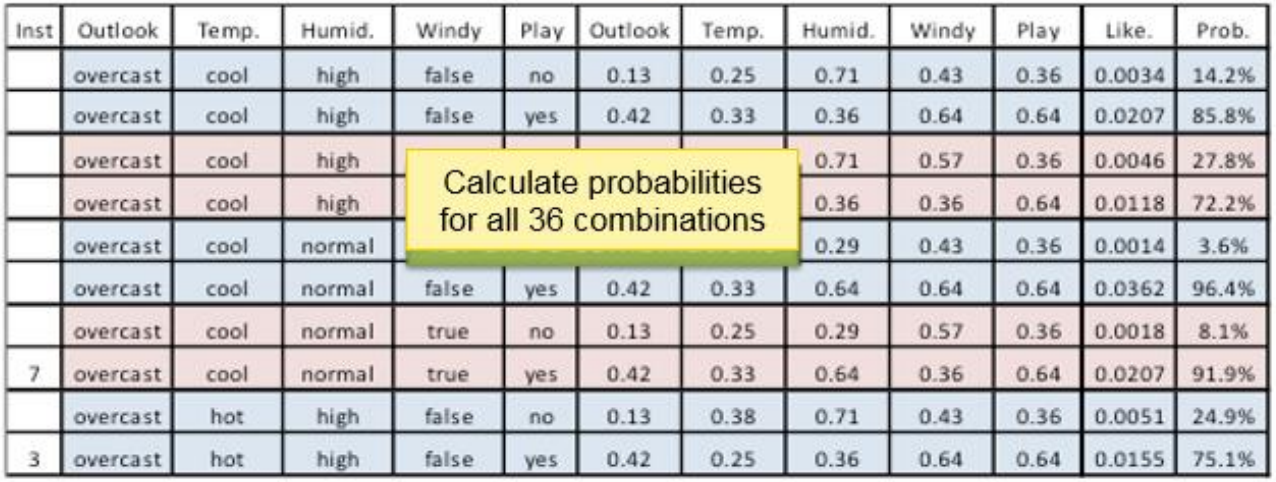
Calculate the rules for each pair.

Then throw out the rules with $p < 0.5$

Prediction rules



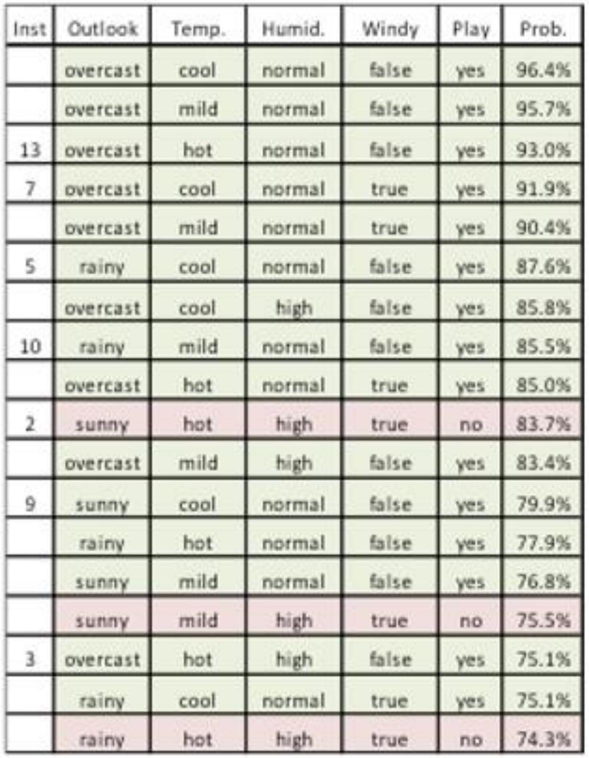
	Play			Play			Play			Play			Play
Outlook	**yes**	**no**	**Temp.**	**yes**	**no**	**Humid.**	**yes**	**no**	**Windy**	**yes**	**no**		
sunny	0.25	0.50	hot	0.25	0.38	high	0.36	0.71	false	0.64	0.43	yes	0.64
overcast	0.42	0.13	mild	0.42	0.38	normal	0.64	0.29	true	0.36	0.57	no	0.36
rainy	0.33	0.38	cool	0.33	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					0.71	0.57	0.36	0.0046	27.8%
	overcast	cool	high					0.36	0.36	0.64	0.0118	72.2%
	overcast	cool	normal					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%

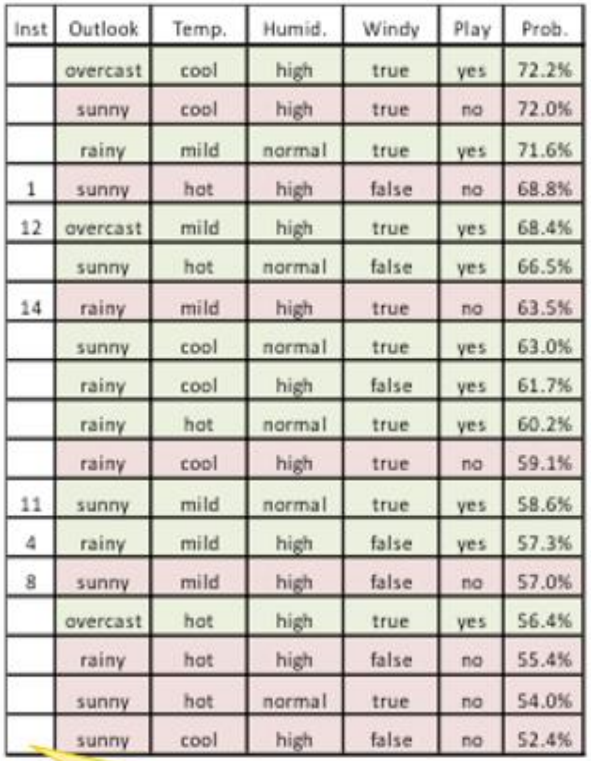
Calculate probabilities
for all 36 combinations

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%

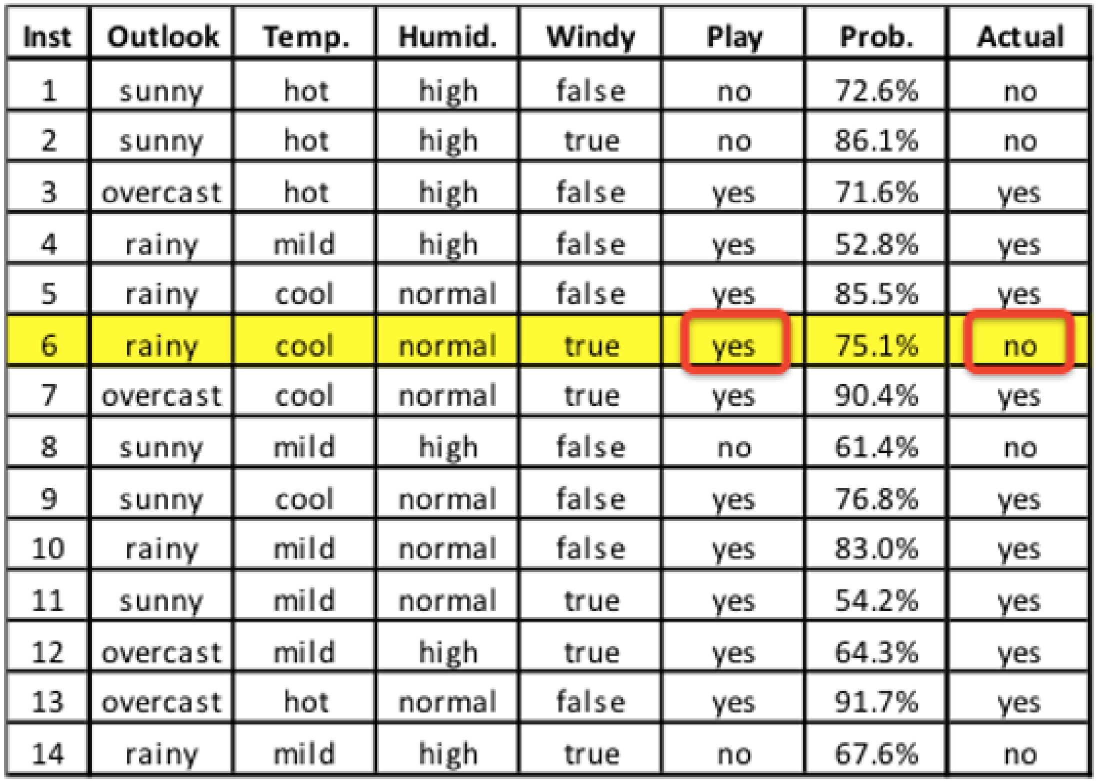
Rules predicting class for all combinations of attributes



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%

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