

Classification Schema

Classification Schema (0)

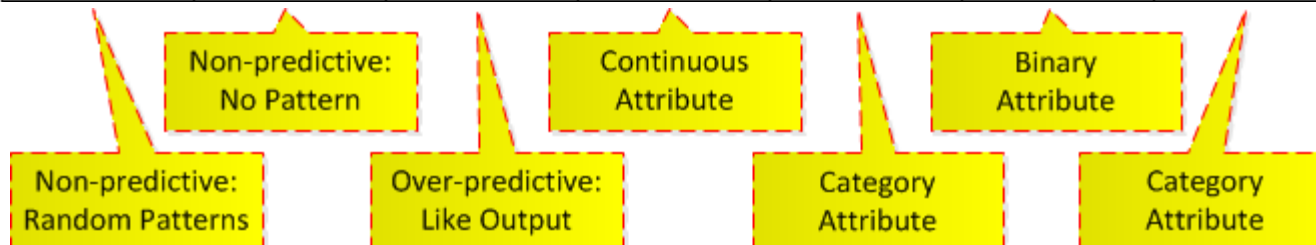
- Rectangular Modeling Dataset
 - Schema
 - Input columns
 - Output column (target column, outcome)
 - Classification: Category Column
 - Partition For Training and Test Data
 - Incremental Data
- Algorithm
 - Classification
 - Logistic Regression
 - Neural Network
 - Decision Tree
 - Naïve Bayes

Classification Schema (1)

Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Attribute 6	Attribute 7
330-272-449	Seaborg	Good	0.123	red	1	Yes
330-272-450	Seaborg	Bad	0.987	green	1	No
330-272-451	Seaborg	Yes	0.245	blue	0	Yes
720-273-500	Seaborg	Yes	0.254	blue	1	Yes
720-273-501	Seaborg	Bad	0.244	blue	0	No
720-273-502	Seaborg		0.415	green	0	Maybe
110-272-461	Seaborg	Yes	0.925	red	1	Yes
110-272-462	Seaborg	Yes	0.376	green	0	Yes
220-273-700	Seaborg	Bad	0.615	green	1	No
220-274-701	Seaborg		0.321	blue	0	Maybe
220-275-703	Seaborg	Bad	0.098	green	0	No
220-275-704	Seaborg	Bad	0.765	red	1	No

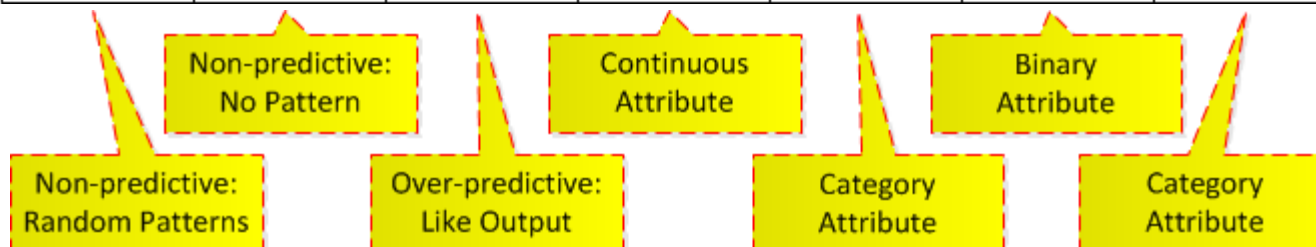
Classification Schema (2)

Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Attribute 6	Attribute 7
330-272-449	Seaborg	Good	0.123	red	1	Yes
330-272-450	Seaborg	Bad	0.987	green	1	No
330-272-451	Seaborg	Yes	0.245	blue	0	Yes
720-273-500	Seaborg	Yes	0.254	blue	1	Yes
720-273-501	Seaborg	Bad	0.244	blue	0	No
720-273-502	Seaborg		0.415	green	0	Maybe
110-272-461	Seaborg	Yes	0.925	red	1	Yes
110-272-462	Seaborg	Yes	0.376	green	0	Yes
220-273-700	Seaborg	Bad	0.615	green	1	No
220-274-701	Seaborg		0.321	blue	0	Maybe
220-275-703	Seaborg	Bad	0.098	green	0	No
220-275-704	Seaborg	Bad	0.765	red	1	No



Classification Schema (3)

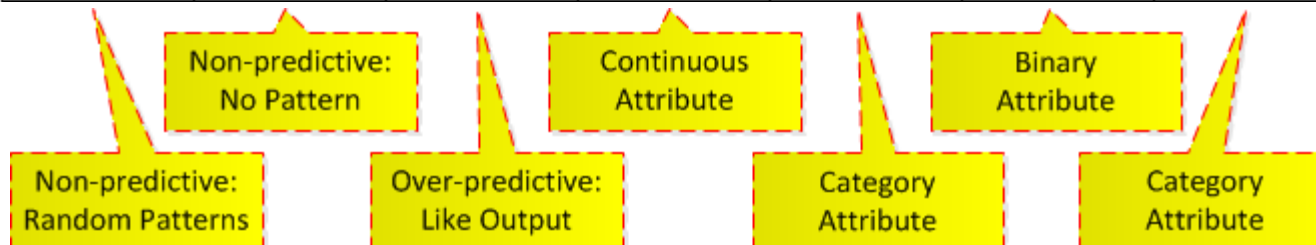
Key Column	Constant	Proxy Column	Input Column 1	Input Column 2	Input Column 3	Outcome Column
330-272-449	Seaborg	Good	0.123	red	1	Yes
330-272-450	Seaborg	Bad	0.987	green	1	No
330-272-451	Seaborg	Yes	0.245	blue	0	Yes
720-273-500	Seaborg	Yes	0.254	blue	1	Yes
720-273-501	Seaborg	Bad	0.244	blue	0	No
720-273-502	Seaborg		0.415	green	0	Maybe
110-272-461	Seaborg	Yes	0.925	red	1	Yes
110-272-462	Seaborg	Yes	0.376	green	0	Yes
220-273-700	Seaborg	Bad	0.615	green	1	No
220-274-701	Seaborg		0.321	blue	0	Maybe
220-275-703	Seaborg	Bad	0.098	green	0	No
220-275-704	Seaborg	Bad	0.765	red	1	No



Classification Schema (4)

Outcome ~ Column 1 + Column 2 + Column 3

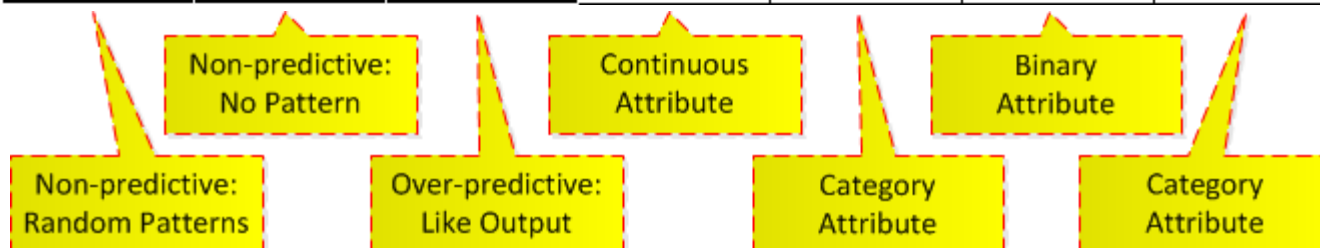
Key Column	Constant	Proxy Column	Input Column 1	Input Column 2	Input Column 3	Outcome Column
330-272-449	Seaborg	Good	0.123	red	1	Yes
330-272-450	Seaborg	Bad	0.987	green	1	No
330-272-451	Seaborg	Yes	0.245	blue	0	Yes
720-273-500	Seaborg	Yes	0.254	blue	1	Yes
720-273-501	Seaborg	Bad	0.244	blue	0	No
720-273-502	Seaborg		0.415	green	0	Maybe
110-272-461	Seaborg	Yes	0.925	red	1	Yes
110-272-462	Seaborg	Yes	0.376	green	0	Yes
220-273-700	Seaborg	Bad	0.615	green	1	No
220-274-701	Seaborg		0.321	blue	0	Maybe
220-275-703	Seaborg	Bad	0.098	green	0	No
220-275-704	Seaborg	Bad	0.765	red	1	No



Classification Schema (5)

Outcome ~ Column 1 + Column 2 + Column 3

			Input Column 1	Input Column 2	Input Column 3	Outcome Column
			0.123	red	1	Yes
			0.987	green	1	No
			0.245	blue	0	Yes
			0.254	blue	1	Yes
			0.244	blue	0	No
			0.415	green	0	Maybe
			0.925	red	1	Yes
			0.376	green	0	Yes
			0.615	green	1	No
			0.321	blue	0	Maybe
			0.098	green	0	No
			0.765	red	1	No



Classification Schema (6)

Outcome ~ Column 1 + Column 2 + Column 3

Input Column 1	Input Column 2	Input Column 3	Outcome Column
0.123	red	1	Yes
0.987	green	1	No
0.245	blue	0	Yes
0.254	blue	1	Yes
0.244	blue	0	No
0.415	green	0	Maybe
0.925	red	1	Yes
0.376	green	0	Yes
0.615	green	1	No
0.321	blue	0	Maybe
0.098	green	0	No
0.765	red	1	No

Continuous
Attribute

Binary
Attribute

Category
Attribute

Category
Attribute

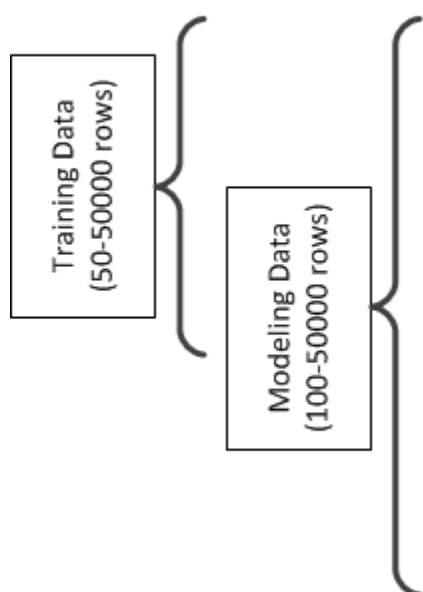
Classification Schema (7)

Outcome ~ Column 1 + Column 2 + Column 3

Modeling Data (100-50000 rows)	Input Column 1	Input Column 2	Input Column 3	Outcome Column
	0.123	red	1	Yes
	0.987	green	1	No
	0.245	blue	0	Yes
	0.254	blue	1	Yes
	0.244	blue	0	No
	0.415	green	0	Maybe
	0.925	red	1	Yes
	0.376	green	0	Yes
	0.615	green	1	No
	0.321	blue	0	Maybe
	0.098	green	0	No
	0.765	red	1	No

Classification Schema (8)

Outcome ~ Column 1 + Column 2 + Column 3

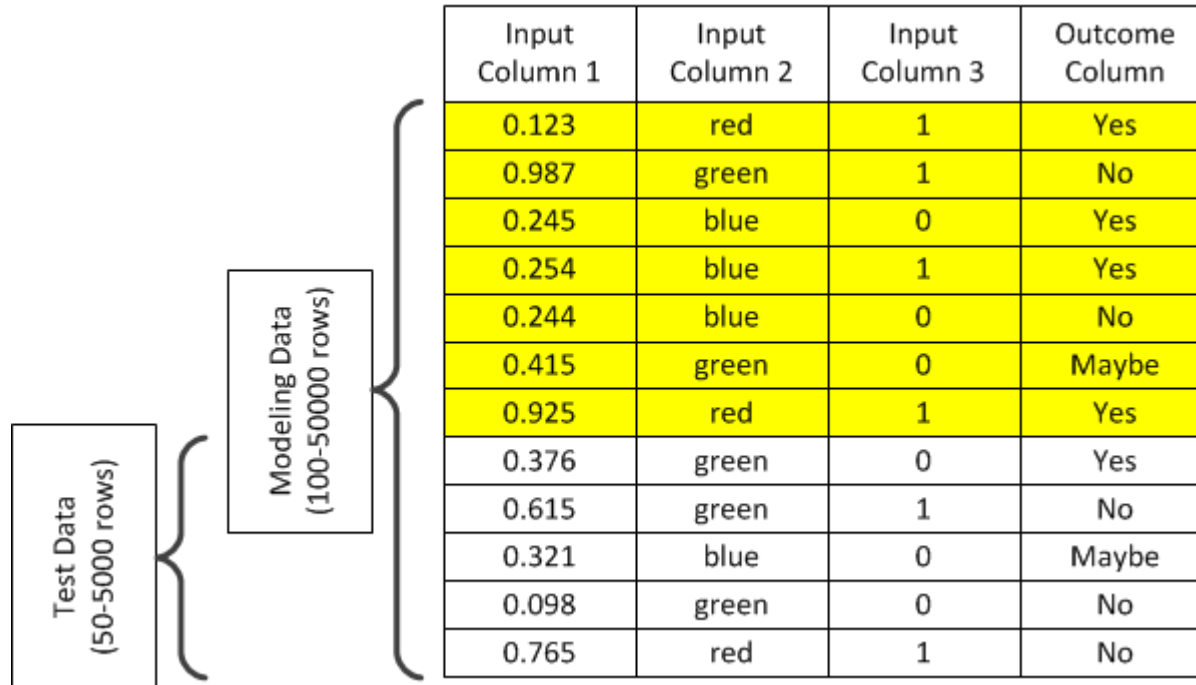


The diagram illustrates the data partitioning process. A large bracket on the left groups the data into two categories: 'Training Data (50-50000 rows)' and 'Modeling Data (100-50000 rows)'. The 'Modeling Data' is further detailed in a table with four columns: 'Input Column 1', 'Input Column 2', 'Input Column 3', and 'Outcome Column'. The table contains 15 rows of data, with the first 7 rows highlighted in yellow.

Input Column 1	Input Column 2	Input Column 3	Outcome Column
0.123	red	1	Yes
0.987	green	1	No
0.245	blue	0	Yes
0.254	blue	1	Yes
0.244	blue	0	No
0.415	green	0	Maybe
0.925	red	1	Yes
0.376	green	0	Yes
0.615	green	1	No
0.321	blue	0	Maybe
0.098	green	0	No
0.765	red	1	No

Classification Schema (9)

Outcome ~ Column 1 + Column 2 + Column 3



Input Column 1	Input Column 2	Input Column 3	Outcome Column
0.123	red	1	Yes
0.987	green	1	No
0.245	blue	0	Yes
0.254	blue	1	Yes
0.244	blue	0	No
0.415	green	0	Maybe
0.925	red	1	Yes
0.376	green	0	Yes
0.615	green	1	No
0.321	blue	0	Maybe
0.098	green	0	No
0.765	red	1	No

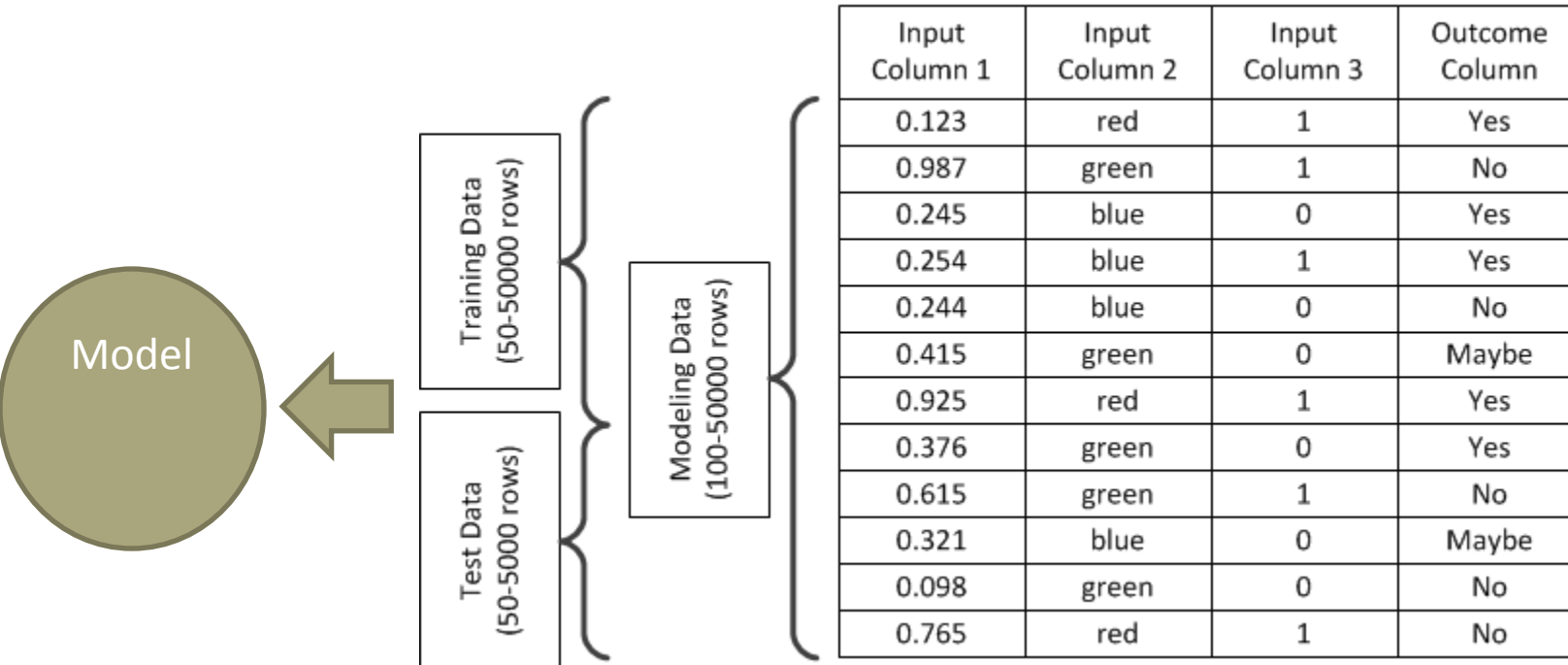
Classification Schema (10)

Outcome ~ Column 1 + Column 2 + Column 3

<div>Training Data (50-50000 rows)</div> <div>Test Data (50-5000 rows)</div>		<div>Modeling Data (100-50000 rows)</div>	Input Column 1	Input Column 2	Input Column 3	Outcome Column
			0.123	red	1	Yes
			0.987	green	1	No
			0.245	blue	0	Yes
			0.254	blue	1	Yes
			0.244	blue	0	No
			0.415	green	0	Maybe
			0.925	red	1	Yes
			0.376	green	0	Yes
			0.615	green	1	No
			0.321	blue	0	Maybe
			0.098	green	0	No
			0.765	red	1	No

Classification Schema (11)

Outcome ~ Column 1 + Column 2 + Column 3



Classification Schema (12)

Outcome ~ Column 1 + Column 2 + Column 3

Model

Input Column 1	Input Column 2	Input Column 3	Outcome Column
0.123	red	1	Yes
0.987	green	1	No
0.245	blue	0	Yes
0.254	blue	1	Yes
0.244	blue	0	No
0.415	green	0	Maybe
0.925	red	1	Yes
0.376	green	0	Yes
0.615	green	1	No
0.321	blue	0	Maybe
0.098	green	0	No
0.765	red	1	No
0.234	green	1	
0.567	blue	0	
0.890	green	1	
0.314	red	1	
0.310	blue	1	
0.284	blue	1	

Classification Schema (13)

Outcome ~ Column 1 + Column 2 + Column 3

Model

		Input Column 1	Input Column 2	Input Column 3	Outcome Column
Training Data (50-50000 rows)	Modeling Data (100-50000 rows)	0.123	red	1	Yes
		0.987	green	1	No
		0.245	blue	0	Yes
		0.254	blue	1	Yes
		0.244	blue	0	No
0.415		green	0	Maybe	
0.925		red	1	Yes	
0.376		green	0	Yes	
0.615		green	1	No	
0.321		blue	0	Maybe	
0.098		green	0	No	
0.765		red	1	No	
Incremental Data (1 to ? rows)		0.234	green	1	
		0.567	blue	0	
		0.890	green	1	
	0.314	red	1		
	0.310	blue	1		
	0.284	blue	1		

Classification Schema (14)

Outcome ~ Column 1 + Column 2 + Column 3

Model

Incremental Data (1 to ? rows)	Input Column 1	Input Column 2	Input Column 3	Outcome Column
	0.234	green	1	
	0.567	blue	0	
	0.890	green	1	
	0.314	red	1	
	0.310	blue	1	
	0.284	blue	1	

Classification Schema (15)

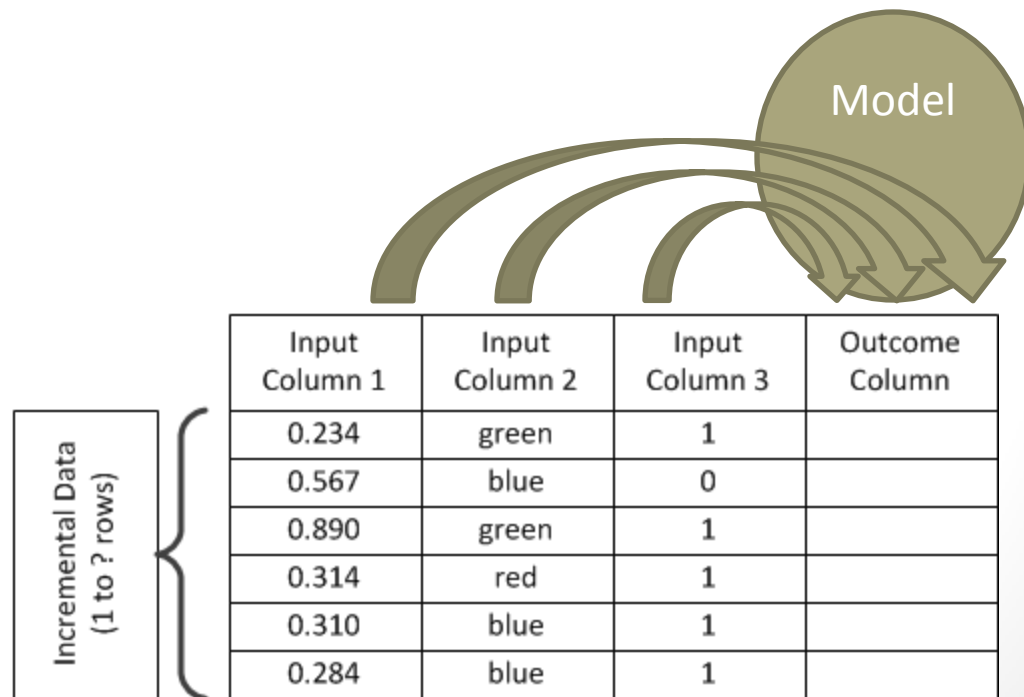
Outcome ~ Column 1 + Column 2 + Column 3

Model

Incremental Data (1 to ? rows)	Input Column 1	Input Column 2	Input Column 3	Outcome Column
	0.234	green	1	
	0.567	blue	0	
	0.890	green	1	
	0.314	red	1	
	0.310	blue	1	
	0.284	blue	1	

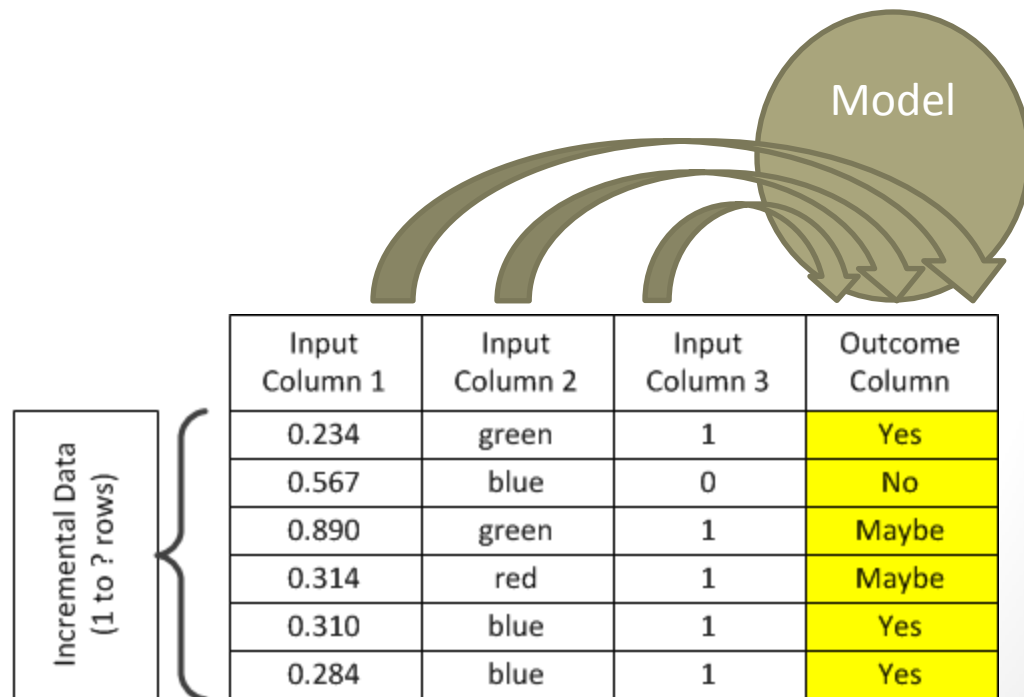
Classification Schema (16)

Outcome ~ Column 1 + Column 2 + Column 3



Classification Schema (17)

Outcome ~ Column 1 + Column 2 + Column 3



Classification Schema (18)

Outcome ~ Column 1 + Column 2 + Column 3

Model

Incremental Data (1 to ? rows)	Input Column 1	Input Column 2	Input Column 3	Outcome Column
	0.234	green	1	Yes
	0.567	blue	0	No
	0.890	green	1	Maybe
	0.314	red	1	Maybe
	0.310	blue	1	Yes
	0.284	blue	1	Yes

Classification Schema (19)

- Attributes
 - All the columns are attributes
- Input Column
 - Input columns are columns that can help predict the outcome. Input columns can be of type binary, ordinal, or category.
- Target Outcome
 - The term "Target Outcome" is redundant. The outcome is the target and vice versa. The target or outcome is the output of a predict function. Providing target or outcome values during modeling makes the process supervised. Creating a model using a outcome is called supervised learning.
- Proxy Column
 - A proxy column is a column that predicts too well. It is too good to be true. Something from the target leaked. This is also called target leakage. The leaked information is "not fair" to use in a prediction. Values for that attribute will not be available when you want to predict the target outcome.
- Key Column
 - In principle, a key column should not affect the model's prediction. The relationship between a key and any other attribute should be random. In practice, the algorithm will find a pattern in the key column and train on this pattern. This pattern is likely to be fortuitous, that means: random. The pattern will not hold for test data or when the model is applied. As a consequence, the key column will affect the model in a bad way.
- Constant Column
 - A constant column should have no affect on the model's predictions. The constant column may increase computation time and cause other problems. It is standard practice to remove all constant columns prior to modeling.

Classification Schema