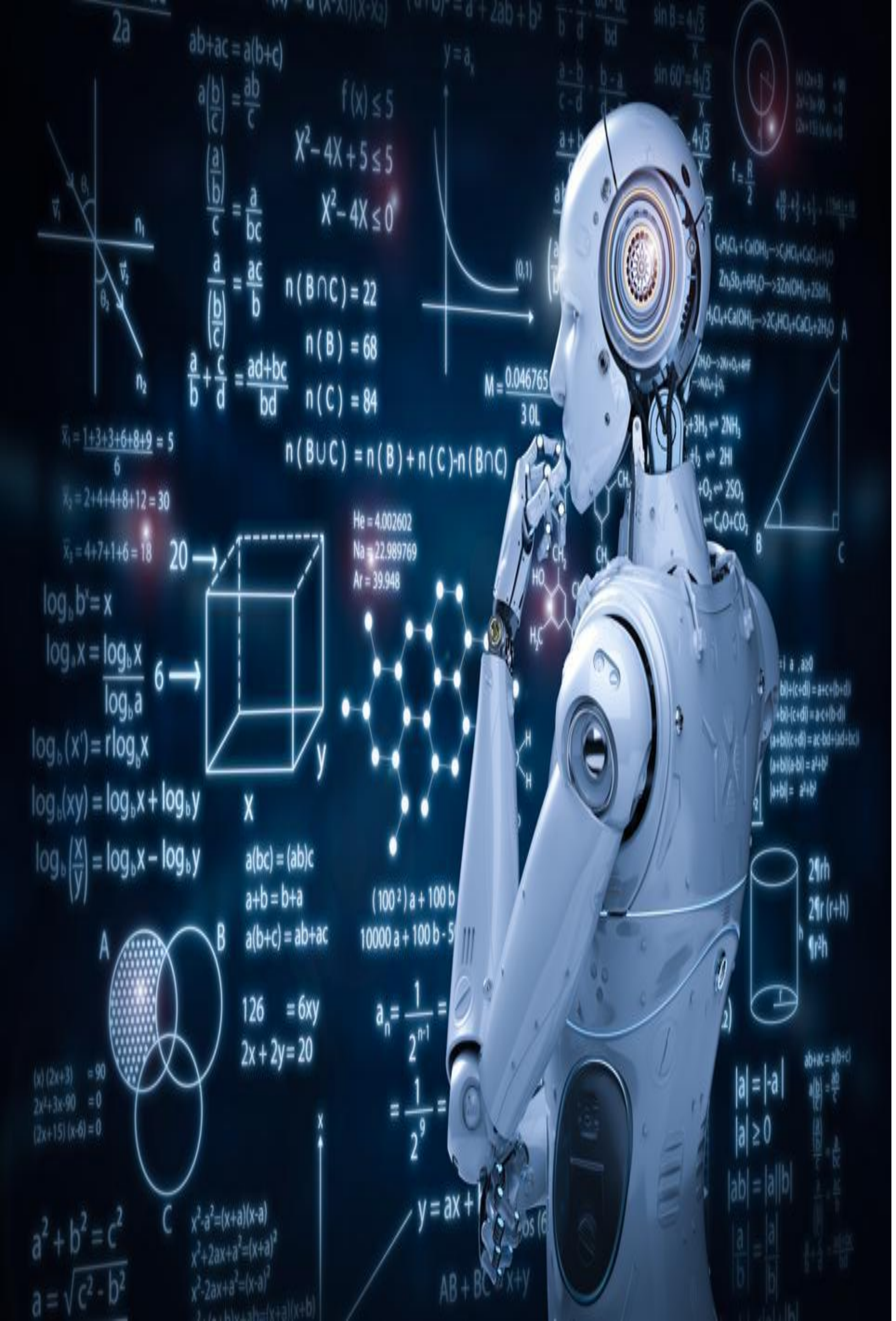


Practical Machine Learning

Day 8: Mar22 DBDA

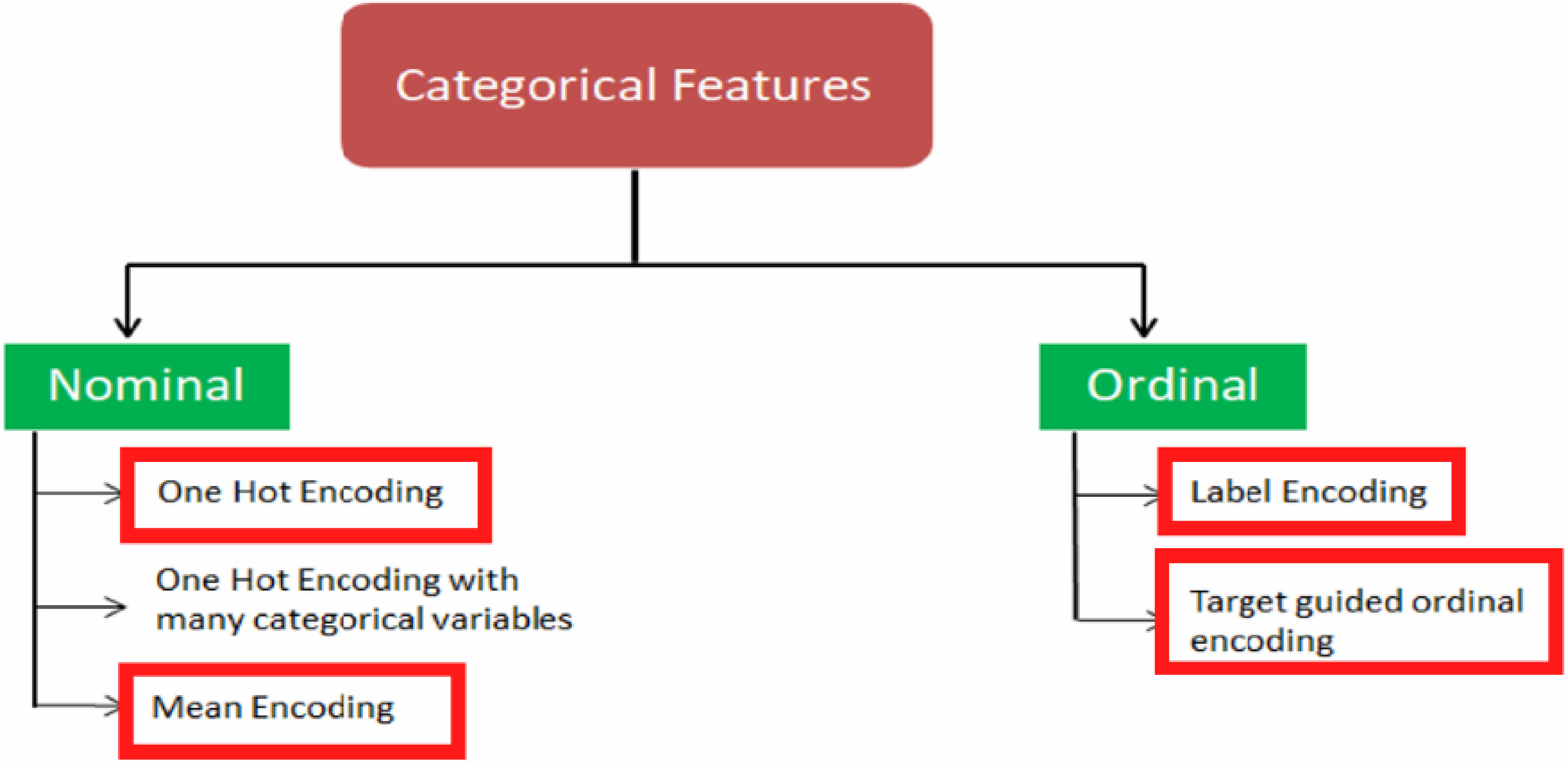
Kiran Waghmare

Agenda

- Classification
- Measures for classification
- KNN

Problem Statement

- **Titanic dataset**
- **Explore:** How does each feature relate to whether a person survives/alives?
- Do the EDA in more detail than usual and explain the results!
  - Splitting: 80-20, stratify: y, random_state = 0
- **Preprocessing:**
  - * Drop decks
  - * Fill in the missing value using a simple imputer
  - * One hot encoding: sex, alone
  - * Ordinal encoding: class
  - * Binary encoding: embark town
- **Model selection:**
  - * Evaluation metrics used: F1_score
  - Logistic Regression



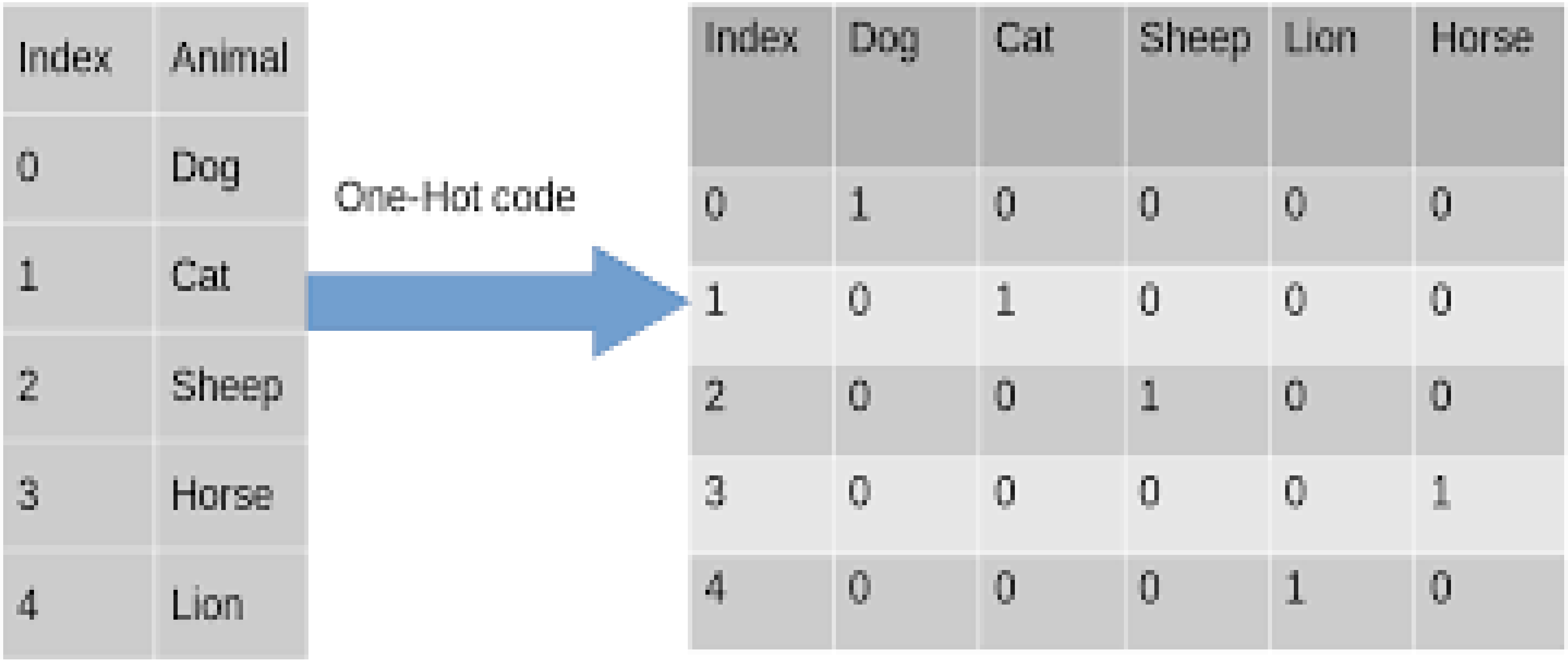
Categorical Features

Nominal

- One Hot Encoding
- One Hot Encoding with many categorical variables
- Mean Encoding

Ordinal

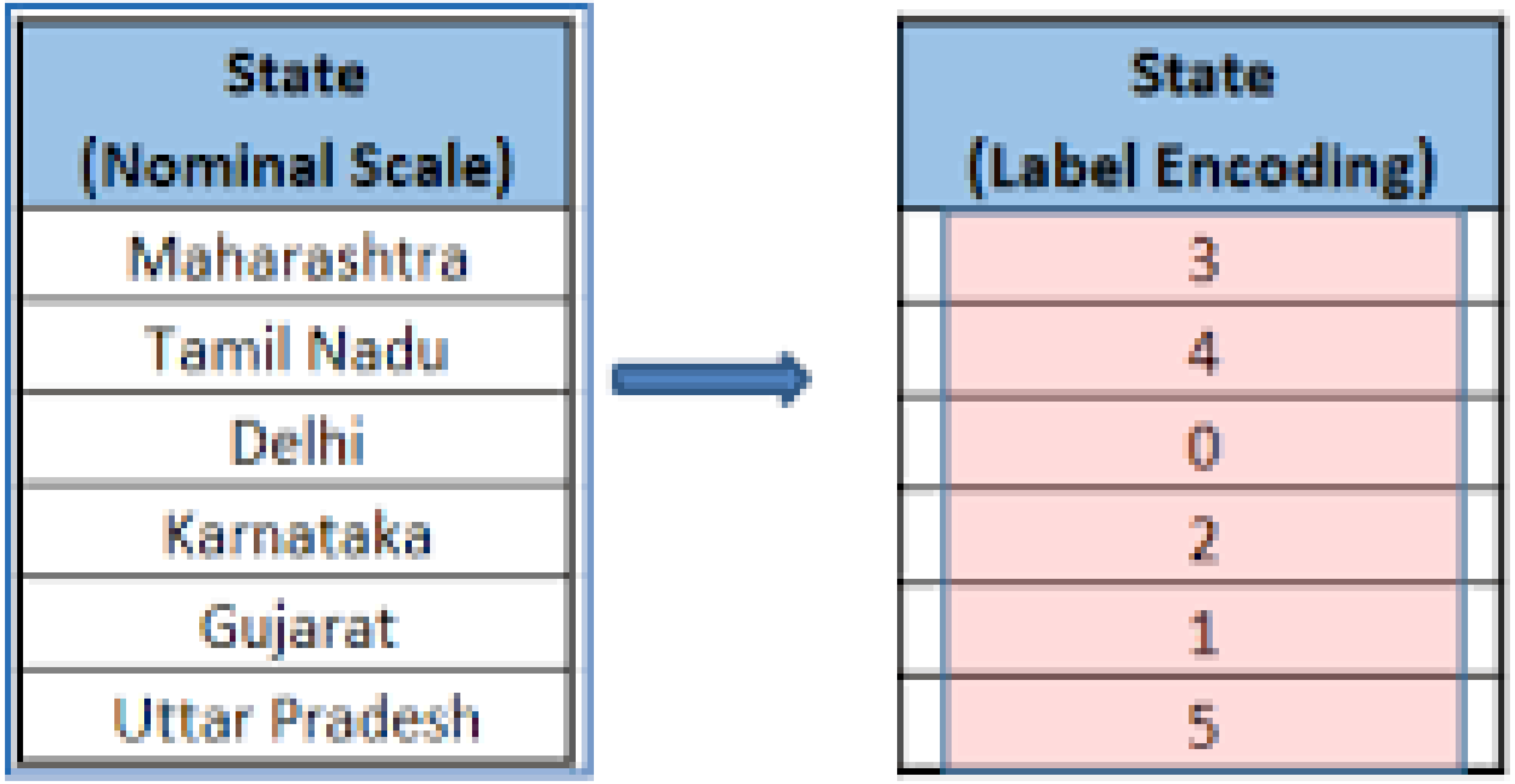
- Label Encoding
- Target guided ordinal encoding



Index	Animal
0	Dog
1	Cat
2	Sheep
3	Horse
4	Lion

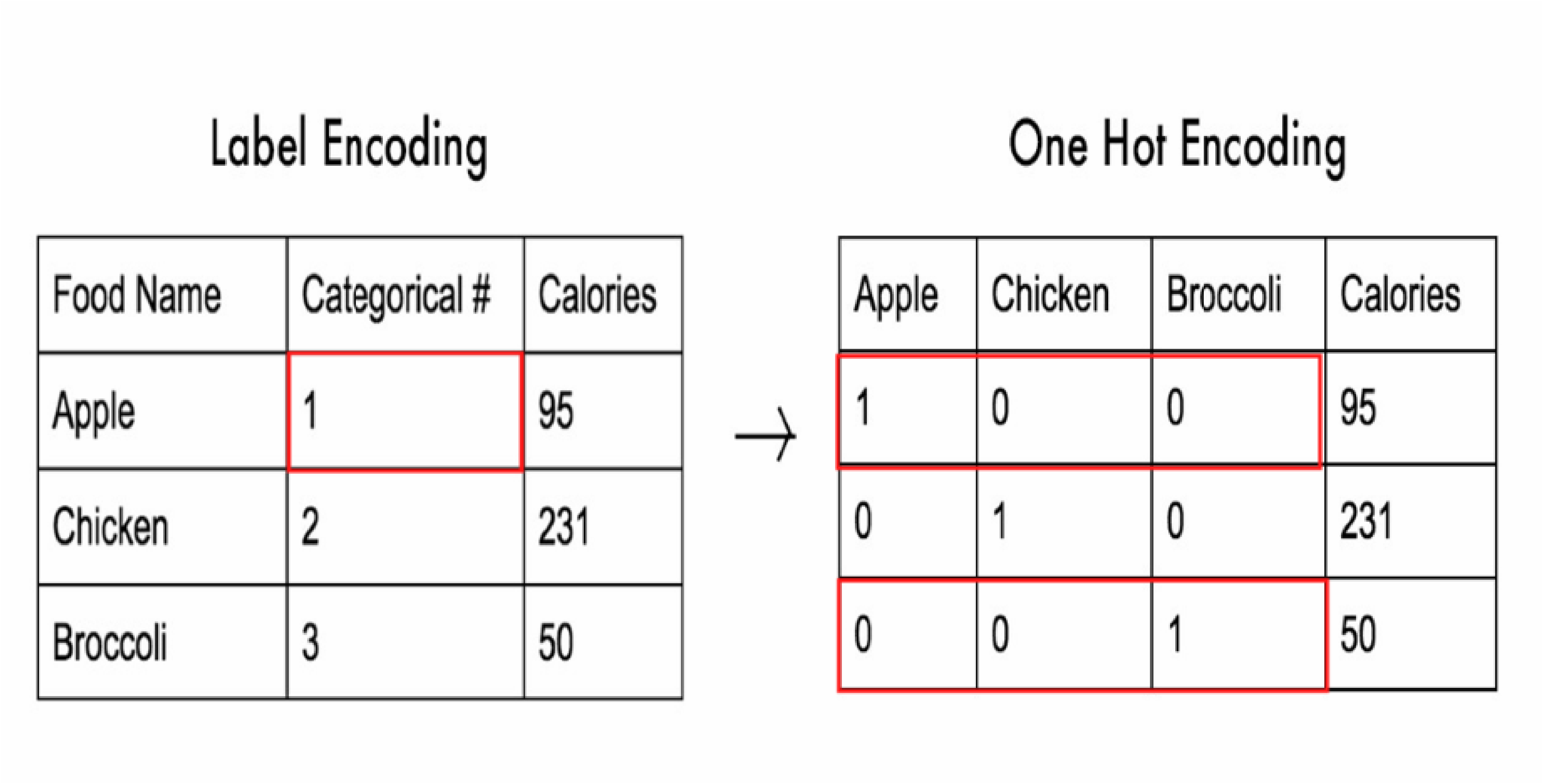
One-Hot code

Index	Dog	Cat	Sheep	Lion	Horse
0	1	0	0	0	0
1	0	1	0	0	0
2	0	0	1	0	0
3	0	0	0	0	1
4	0	0	0	1	0



State (Nominal Scale)
Maharashtra
Tamil Nadu
Delhi
Karnataka
Gujarat
Uttar Pradesh

State (Label Encoding)
3
4
0
2
1
5



Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

Target Mean Encoding

Height
Short
Tall
Short
Medium

Target
100
50
70
60

Height	Target Mean
Short	(100+70)/2 =85
Medium	60
Tall	50

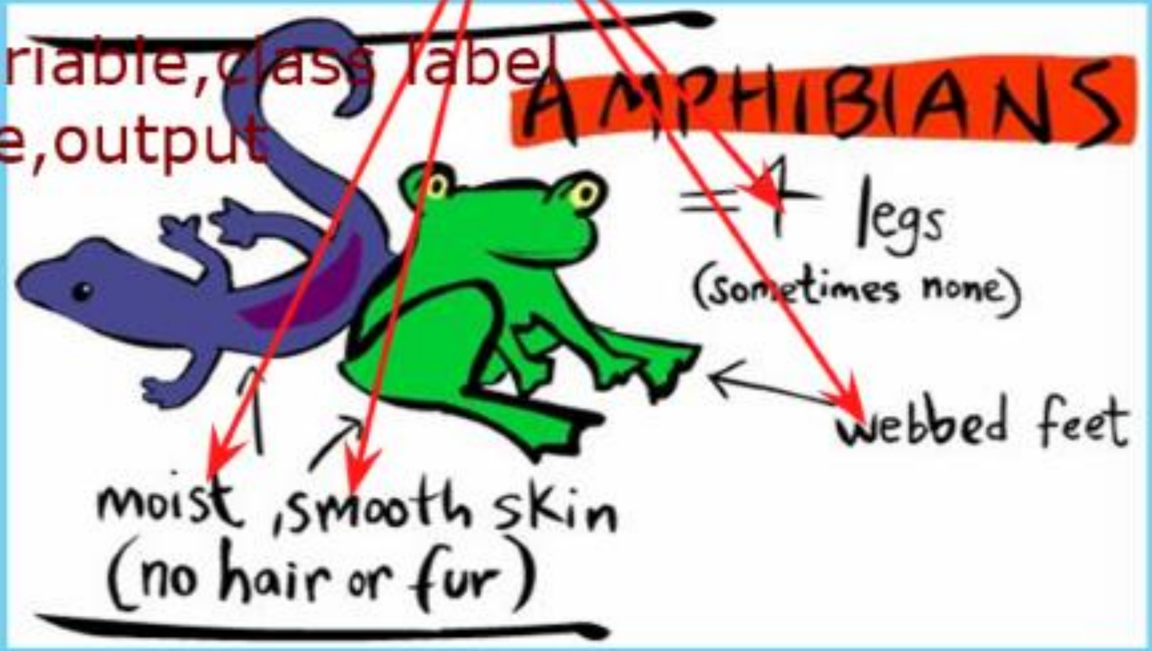
Height
Short
Tall
Short
Medium

Height
85
50
85
60

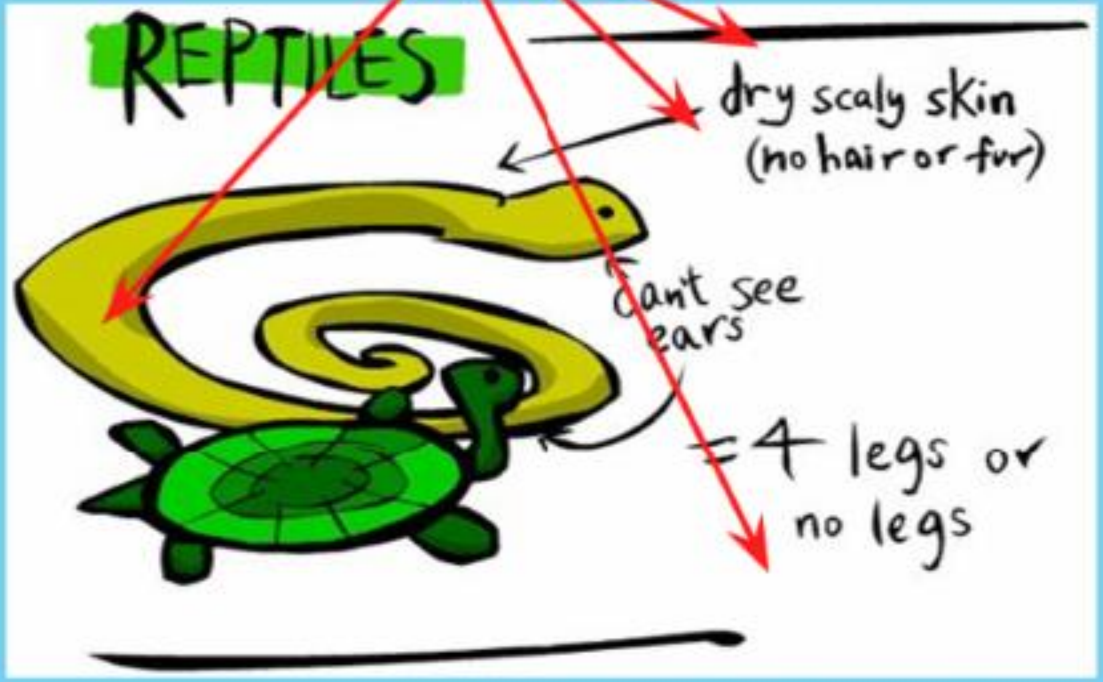
Classification:

Training dataset:collection of records
-tuple(x,y)
-x:Independent,attribute,predictor,variable,class label
-y:Dependent ,class,response,variable,output

Amphibians

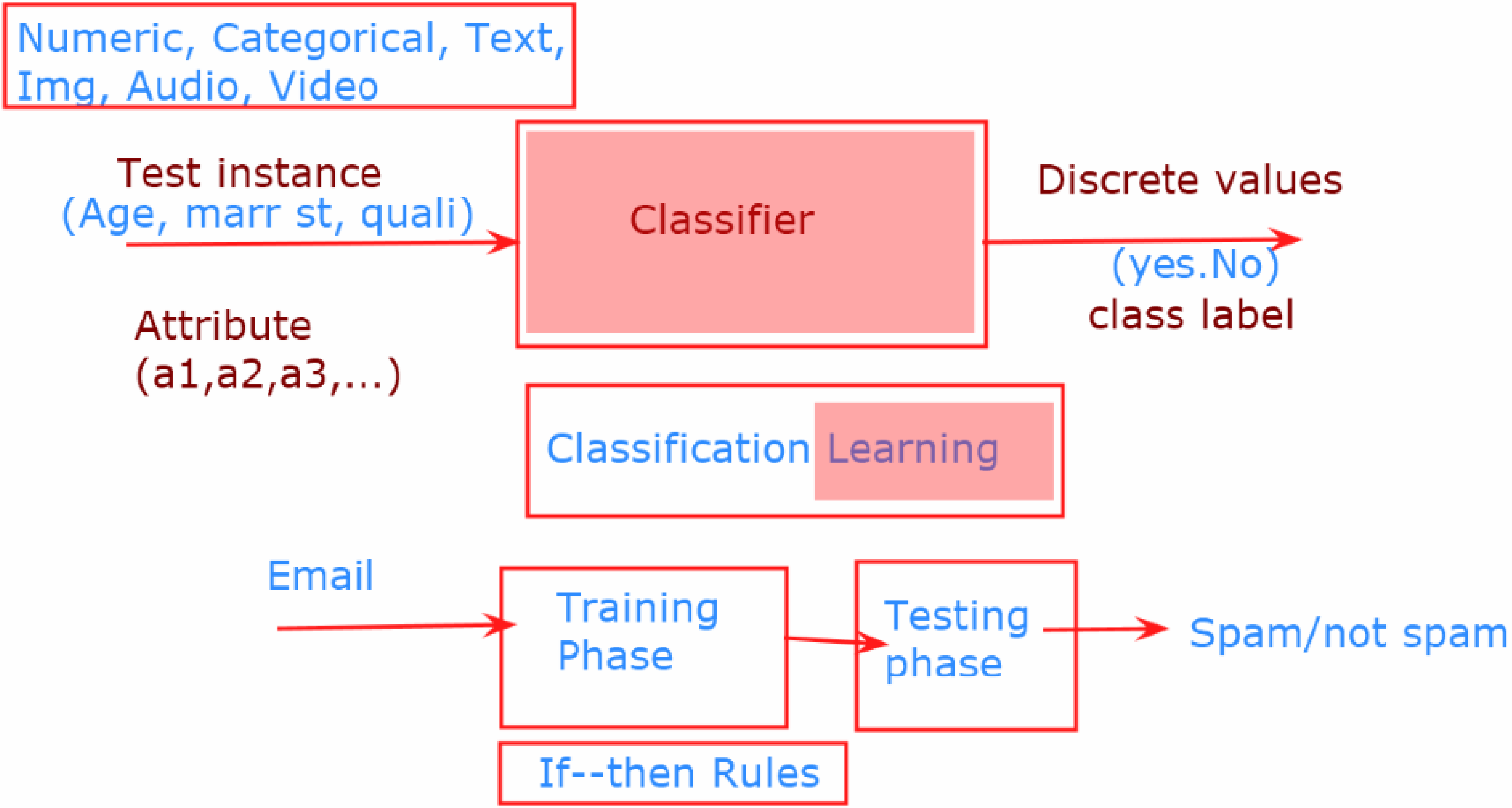


Reptiles



Task:

-Learning of a model
-Mapping of x, y attributes



Numeric, Categorical, Text, Img, Audio, Video

Test instance (Age, marr st, quali)

Classifier

Discrete values

(yes.No)

class label

Attribute (a1,a2,a3,...)

Classification Learning

Email

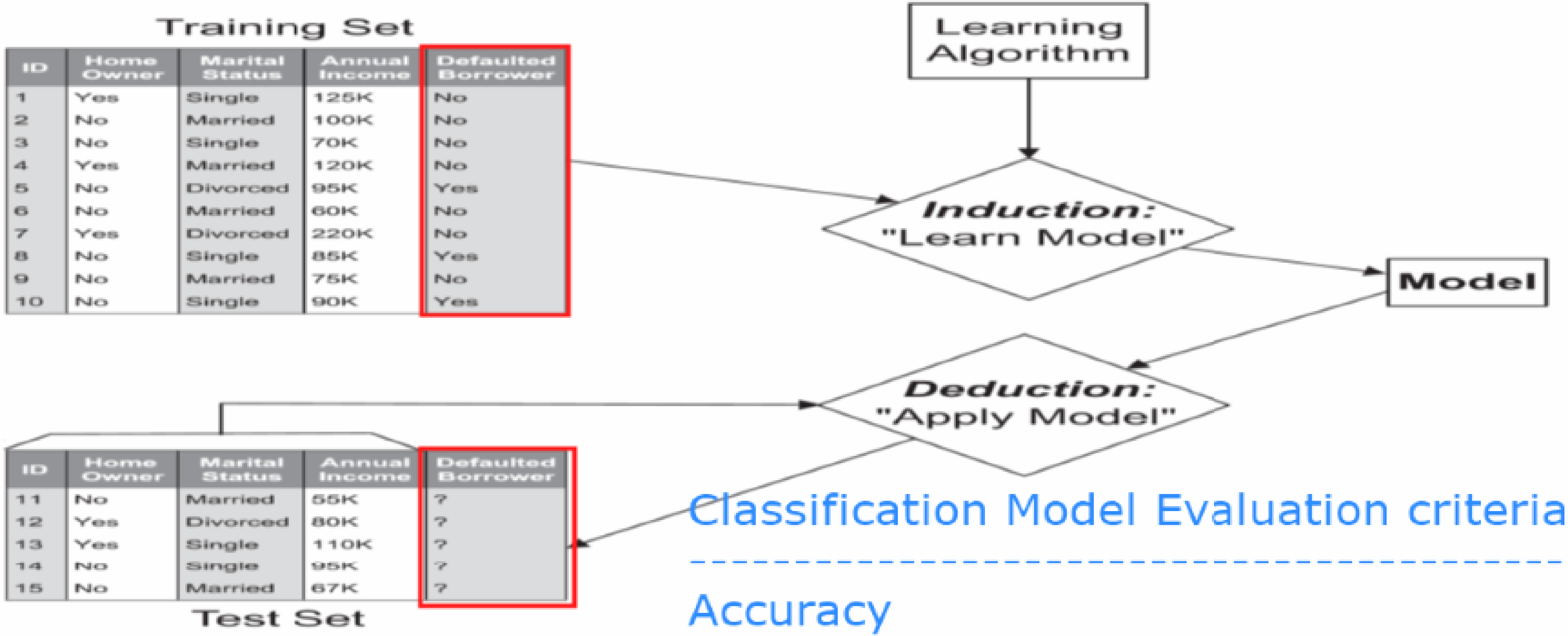
Training Phase

Testing phase

Spam/not spam

If--then Rules

General Approach for Building Classification Model



Training Set

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Learning Algorithm

Induction: "Learn Model"

Model

Deduction: "Apply Model"

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
11	No	Married	55K	?
12	Yes	Divorced	80K	?
13	Yes	Single	110K	?
14	No	Single	95K	?
15	No	Married	67K	?

Test Set

Classification Model Evaluation criteria

Accuracy

Confusion matrix

ROC curse

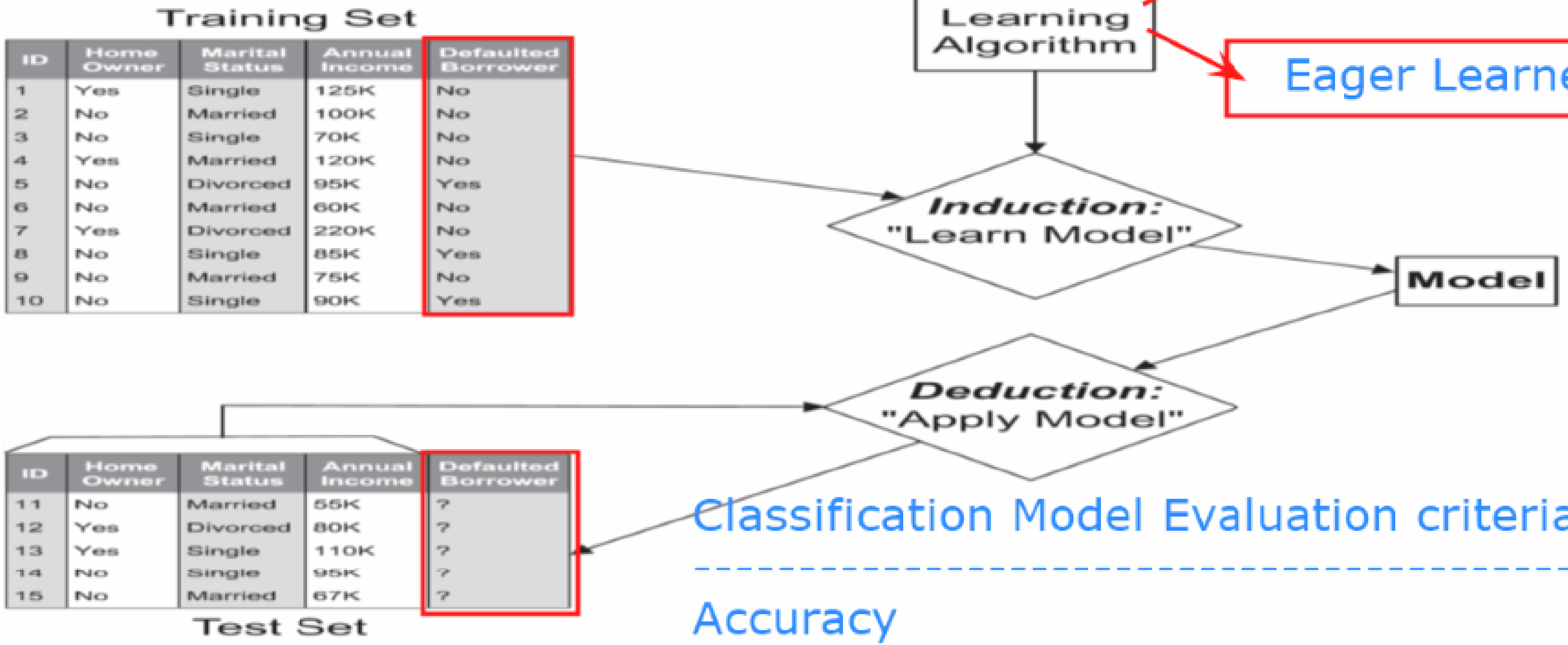
Cost-sensitive

General Approach for Building Classification Model

Association,prob,bayes,hyperplanes

Lazy Learners

Eager Learners



Training Set

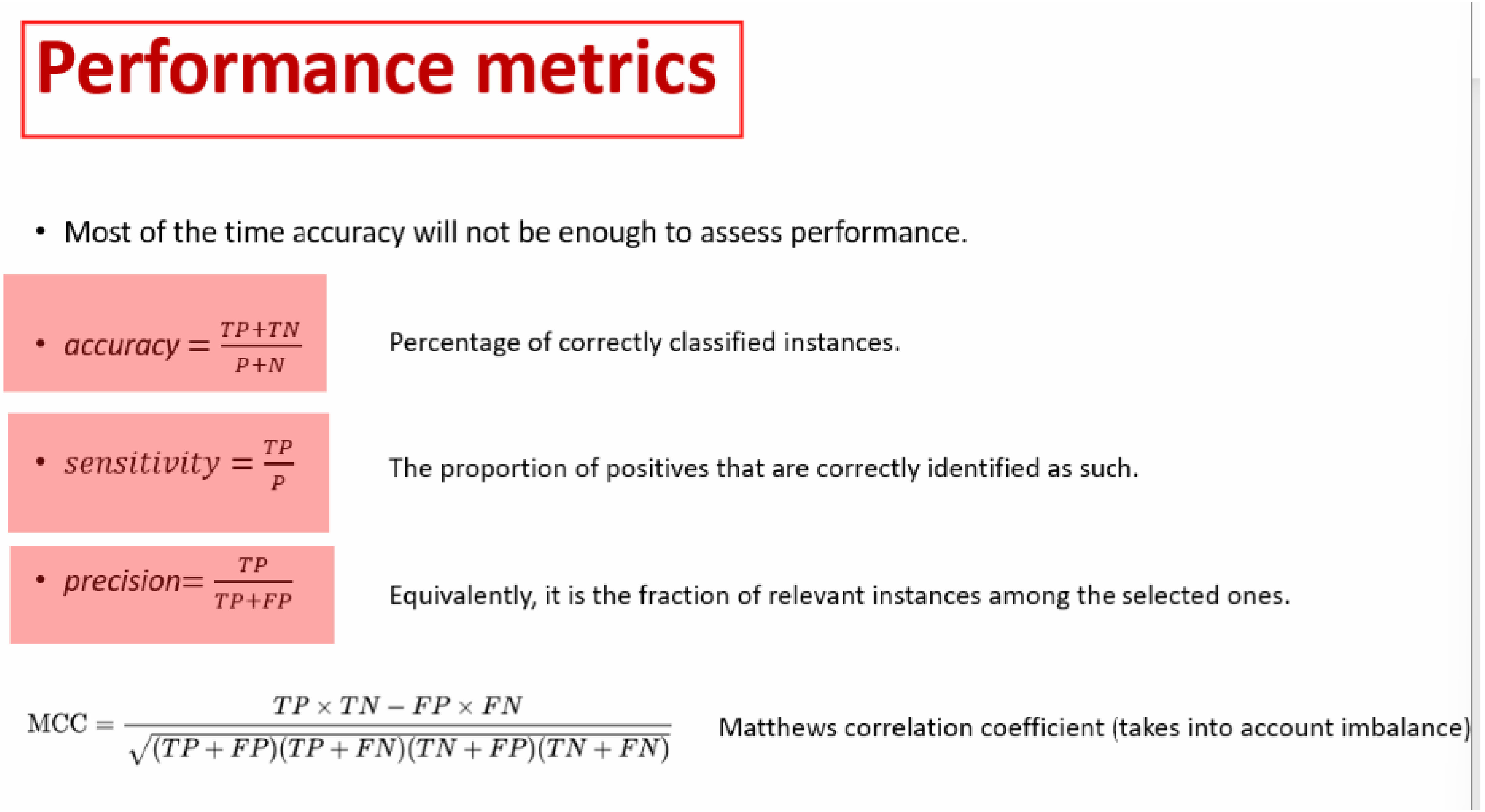
ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
11	No	Married	55K	?
12	Yes	Divorced	80K	?
13	Yes	Single	110K	?
14	No	Single	95K	?
15	No	Married	67K	?

Test Set

Classification Model Evaluation criteria

Accuracy
Confusion matrix
ROC curse
Cost-sensitive



Performance metrics

- Most of the time accuracy will not be enough to assess performance.

- *accuracy* $= \frac{TP+TN}{P+N}$

Percentage of correctly classified instances.

- *sensitivity* $= \frac{TP}{P}$

The proportion of positives that are correctly identified as such.

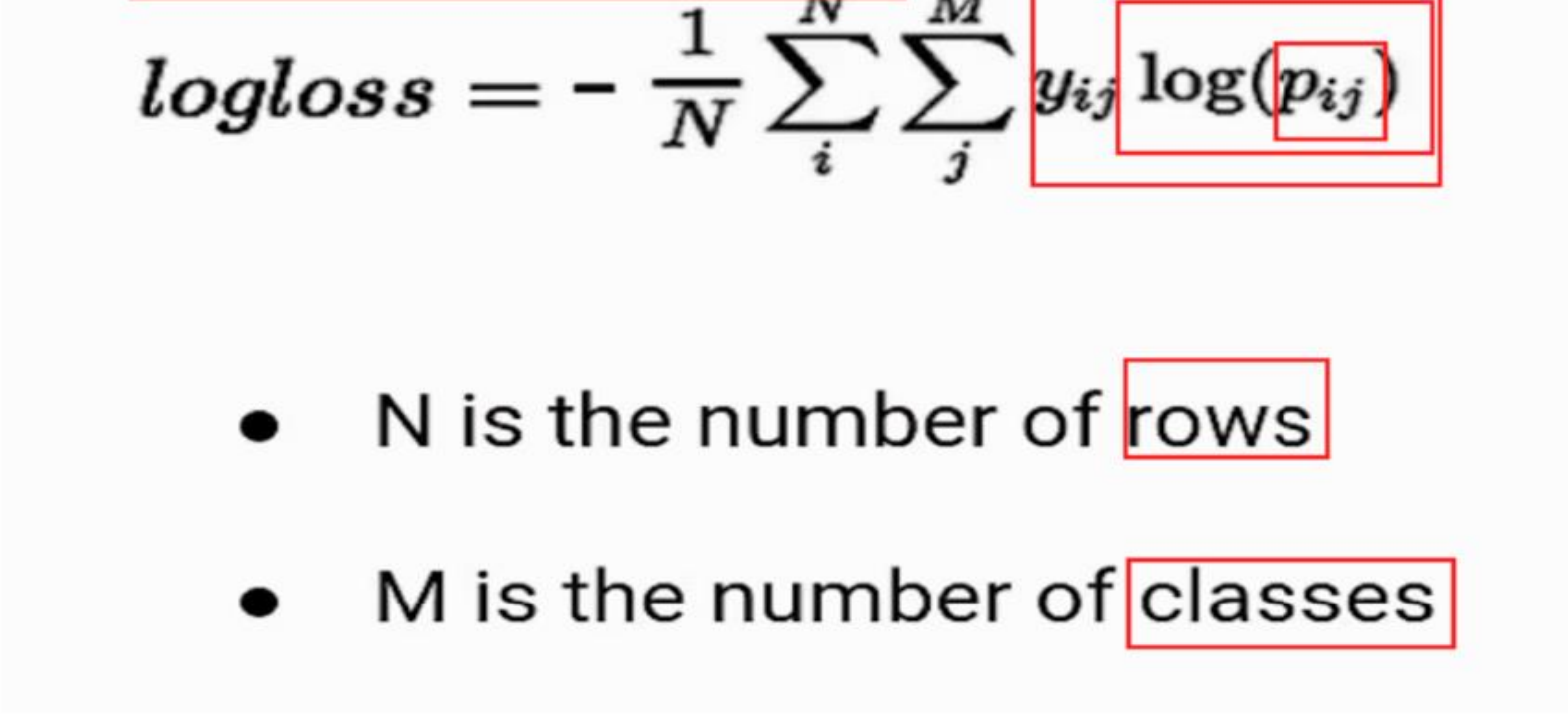
- *precision*$= \frac{TP}{TP+FP}$

Equivalently, it is the fraction of relevant instances among the selected ones.

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

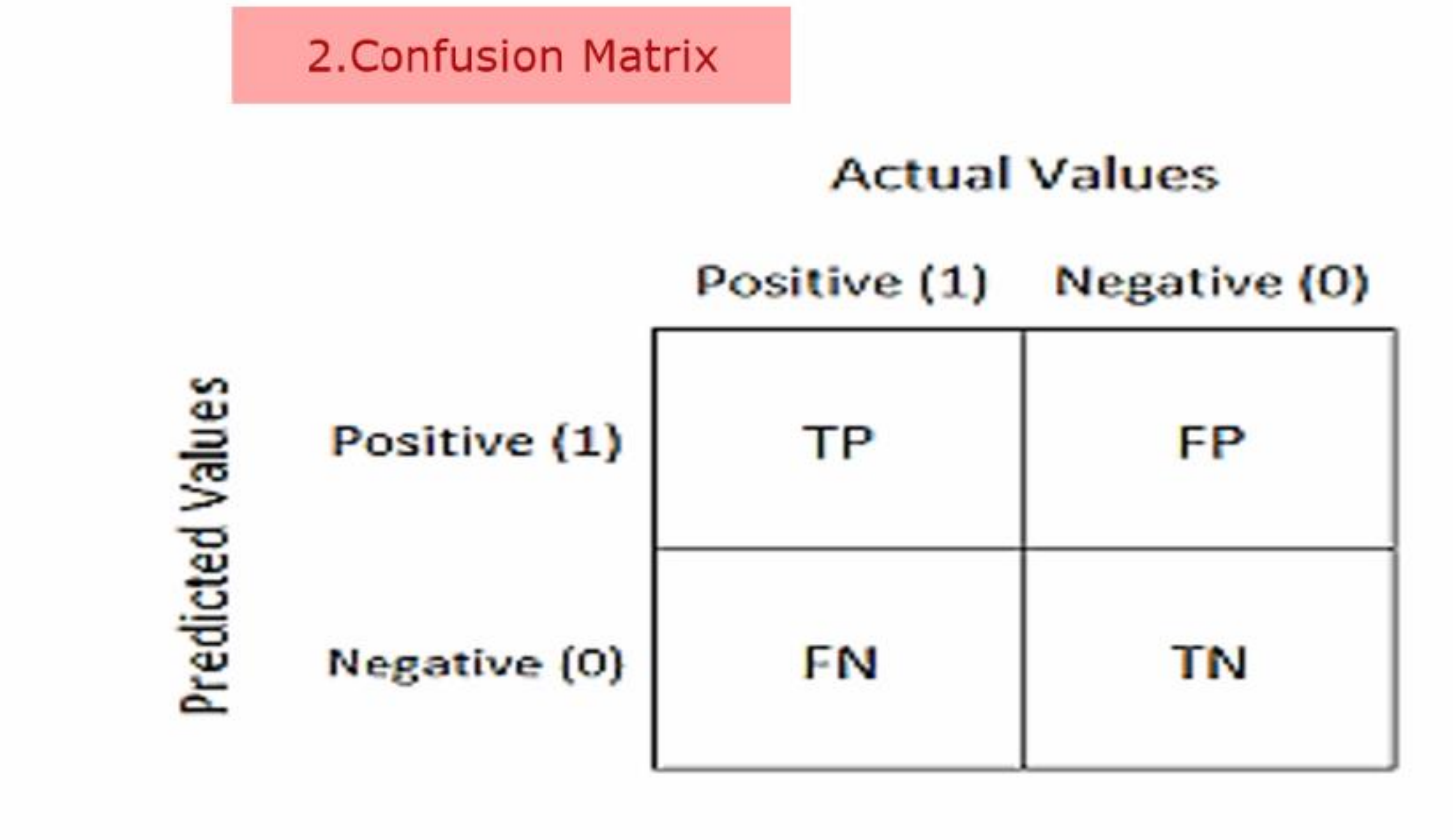
Matthews correlation coefficient (takes into account imbalance)

1.Probablistic way: Entropy



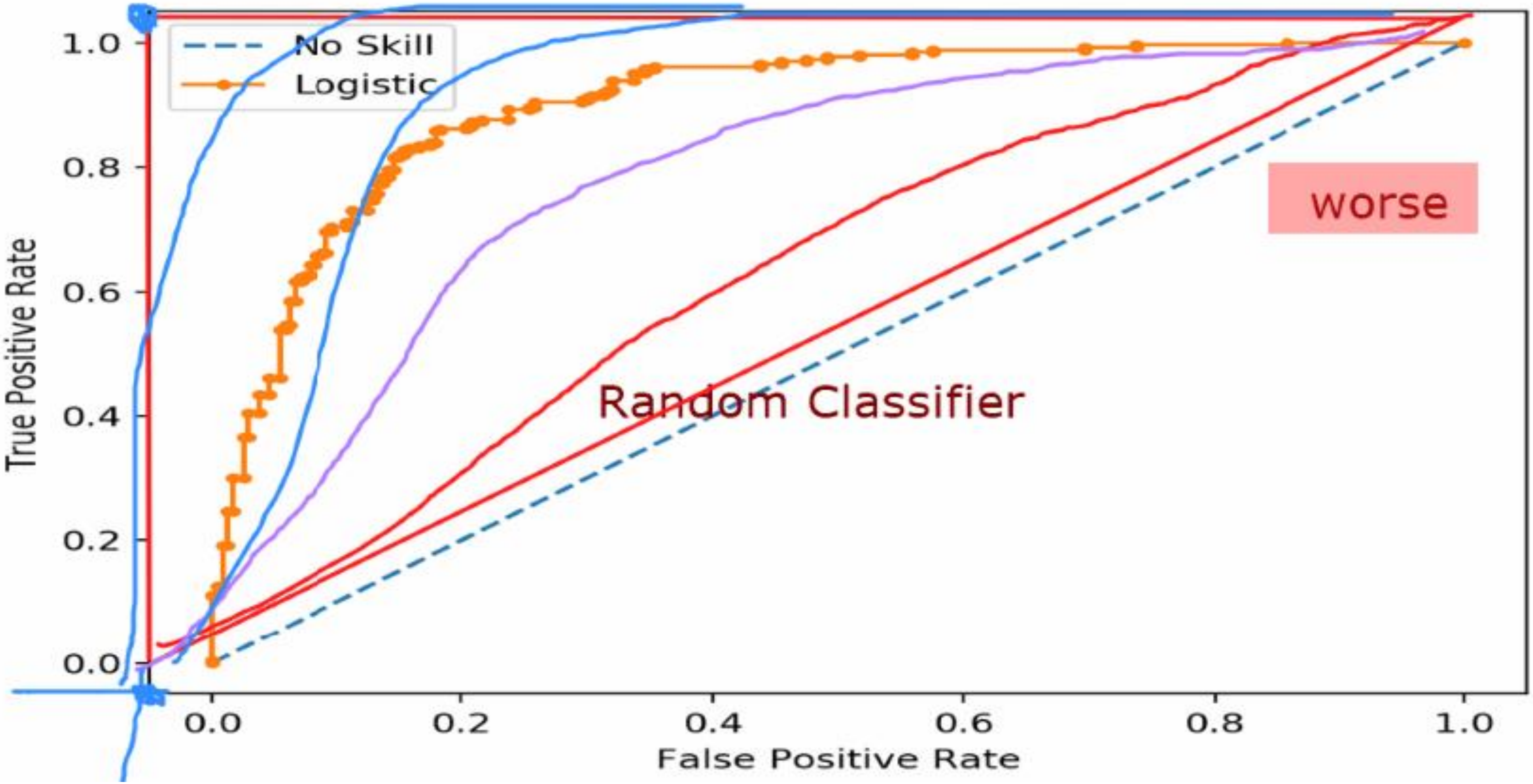
$$logloss = -\frac{1}{N}\sum_{i}^{N}\sum_{j}^{M}\boxed{y_{ij}\boxed{\log(\boxed{p_{ij}})}}$$

- N is the number of rows
- M is the number of classes



2.Confusion Matrix

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

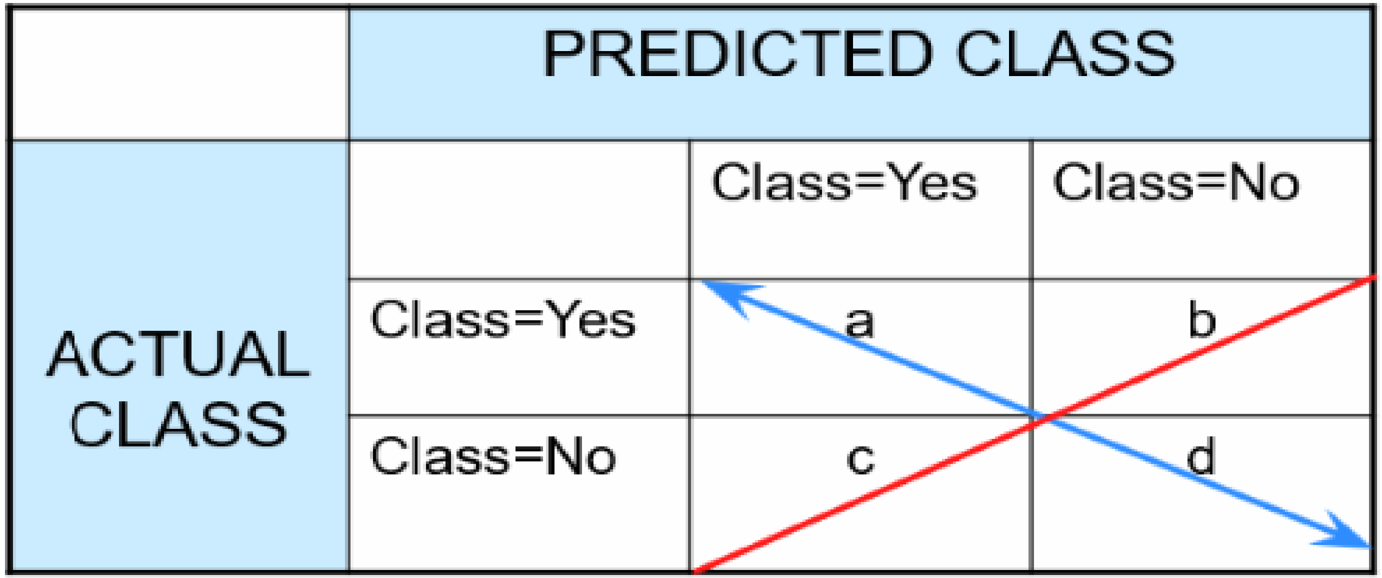

No Skill
Logistic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
worse
Random Classifier

Confusion Matrix

Error Matrix

- Confusion Matrix:

y_pred



	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a	b
	Class=No	c	d

y_test

Error

Accuracy

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

Accuracy

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

- Most widely-used metric:

Alternative Measures

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a	b
	Class=No	c	d

$$\text{Precision (p)} = \frac{a}{a+c}$$

$$\text{Recall (r)} = \frac{a}{a+b}$$

$$\text{F-measure (F)} = \frac{2rp}{r+p} = \frac{2a}{2a+b+c}$$