



# Machine Learning

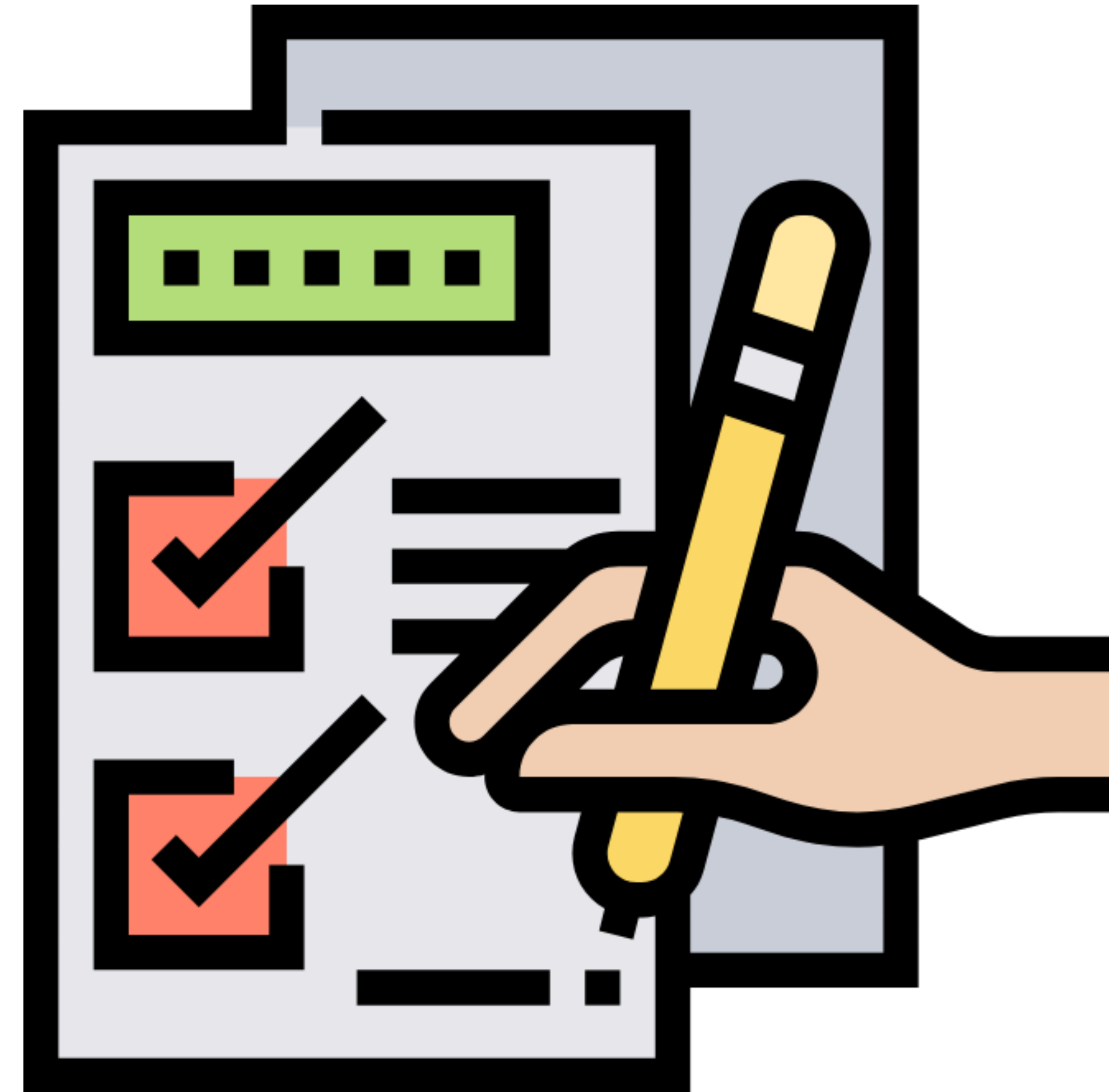
Unidad # 2 - Aprendizaje Supervisado  
CC57 – 2019-1

Profesor  
Andrés Melgar



# Competencias a adquirir en la sesión

- Al finalizar la sesión el alumno comprenderá los conceptos de **overfitting** y **underfitting**.
- Al finalizar la sesión el alumno analizará la **matriz de confusión** resultantes de los modelos algorítmicos.
- Al finalizar la sesión el alumno comprenderá los conceptos de **verdaderos positivos, verdaderos negativos, falsos positivos** y **falsos negativos**.
- Al finalizar la sesión el alumno **analizará** modelos algorítmicos usando la **precisión, recall** y **número F**.





# Revisión de la sesión anterior

- ¿En qué se **fundamenta** el algoritmo **Naïve Bayes**?
- ¿Qué similitudes encuentra en el algoritmo **OneR** y el algoritmo **Naïve Bayes**?
- ¿Cómo gestiona el algoritmo **Naïve Bayes** los **atributos numéricos**?





# Naïve Bayes

## Texto guía

Witten, Ian H., Frank, Eibe, and Hall, Mark A.. 2011. *Data Mining : Practical Machine Learning Tools and Techniques with Java Implementations*. San Francisco: Elsevier Science & Technology.

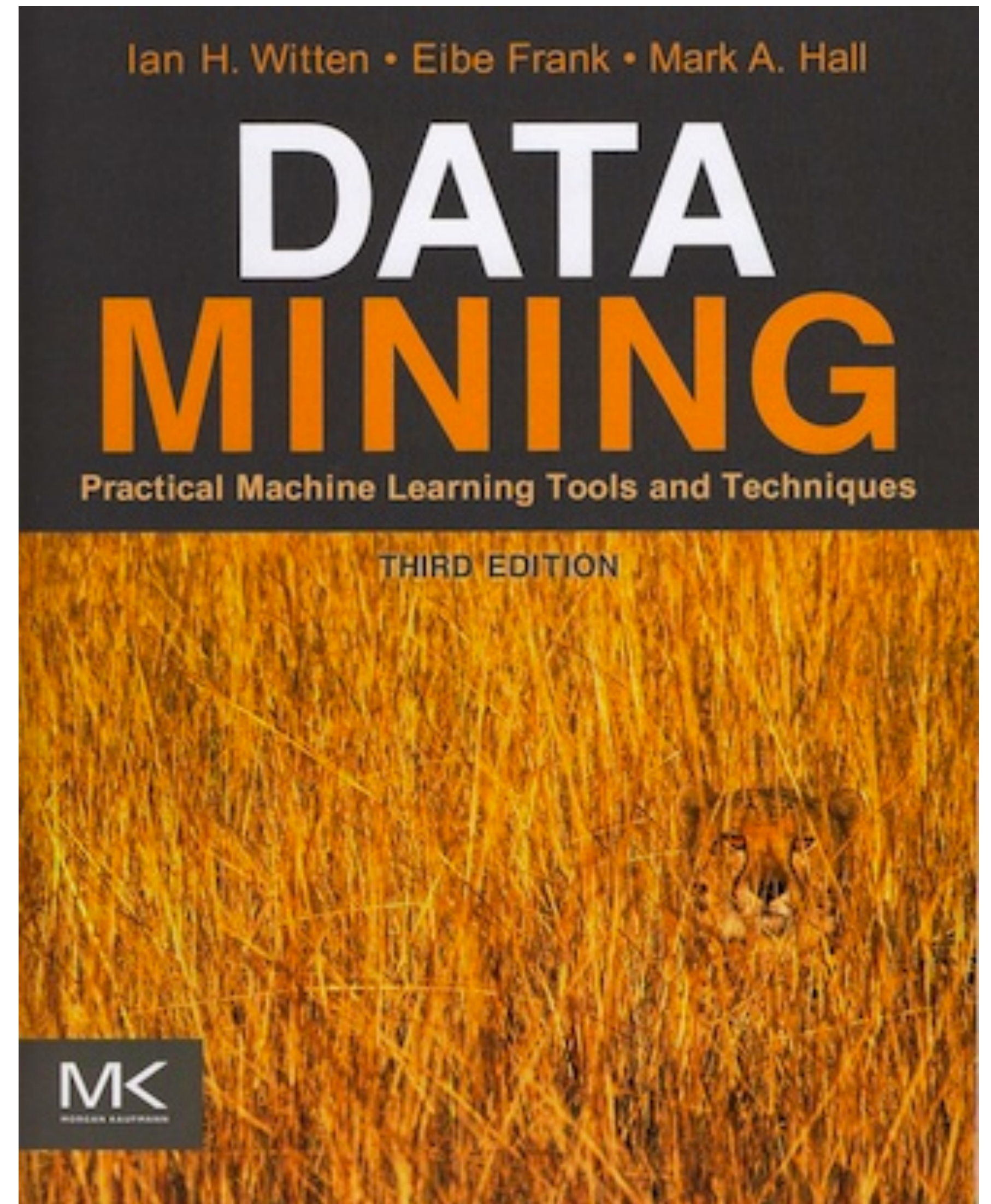
	CHAPTER
Credibility: Evaluating What's Been Learned	5

---

### 5.7 COUNTING THE COST

---

5.7 COUNTING THE COST

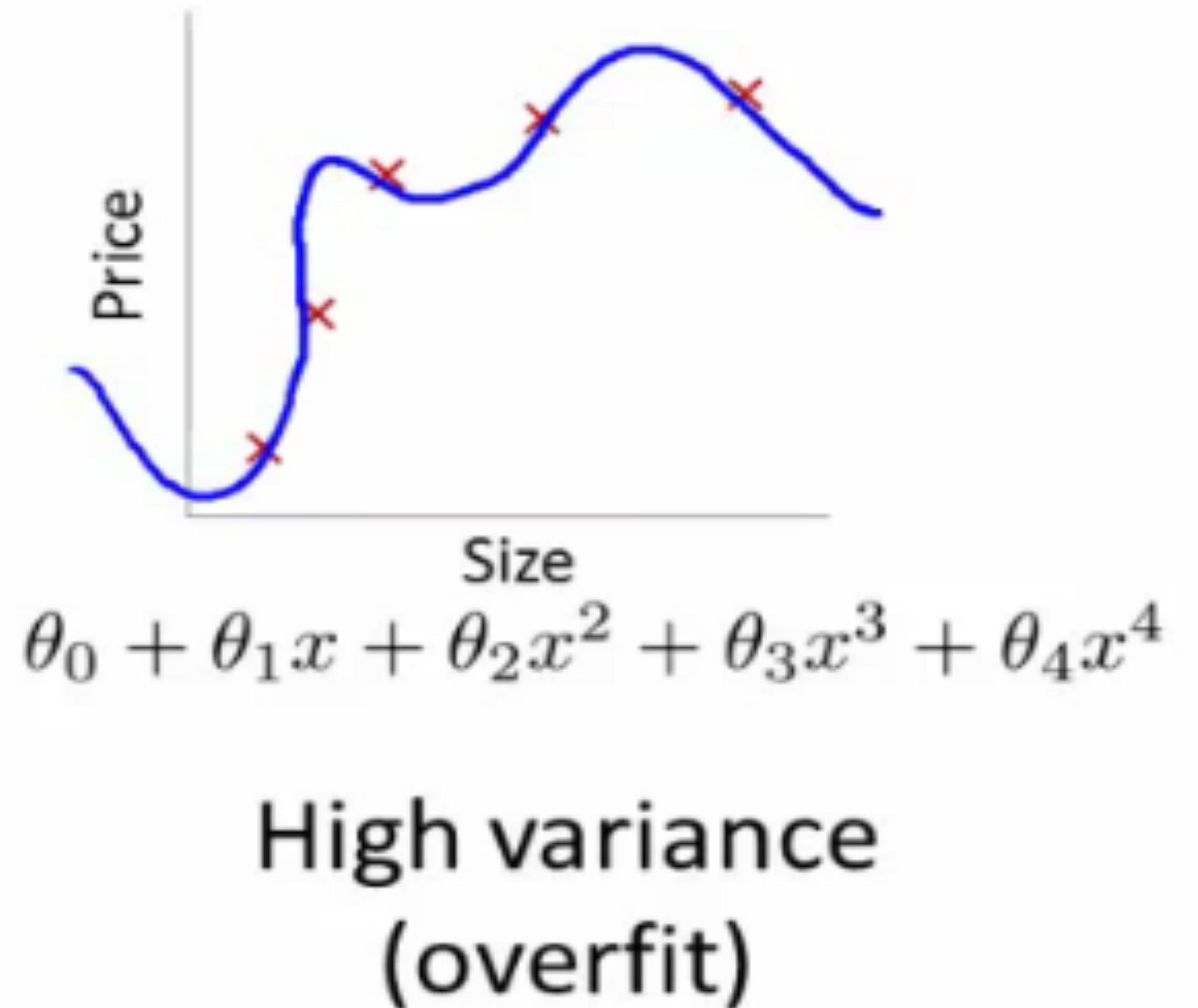






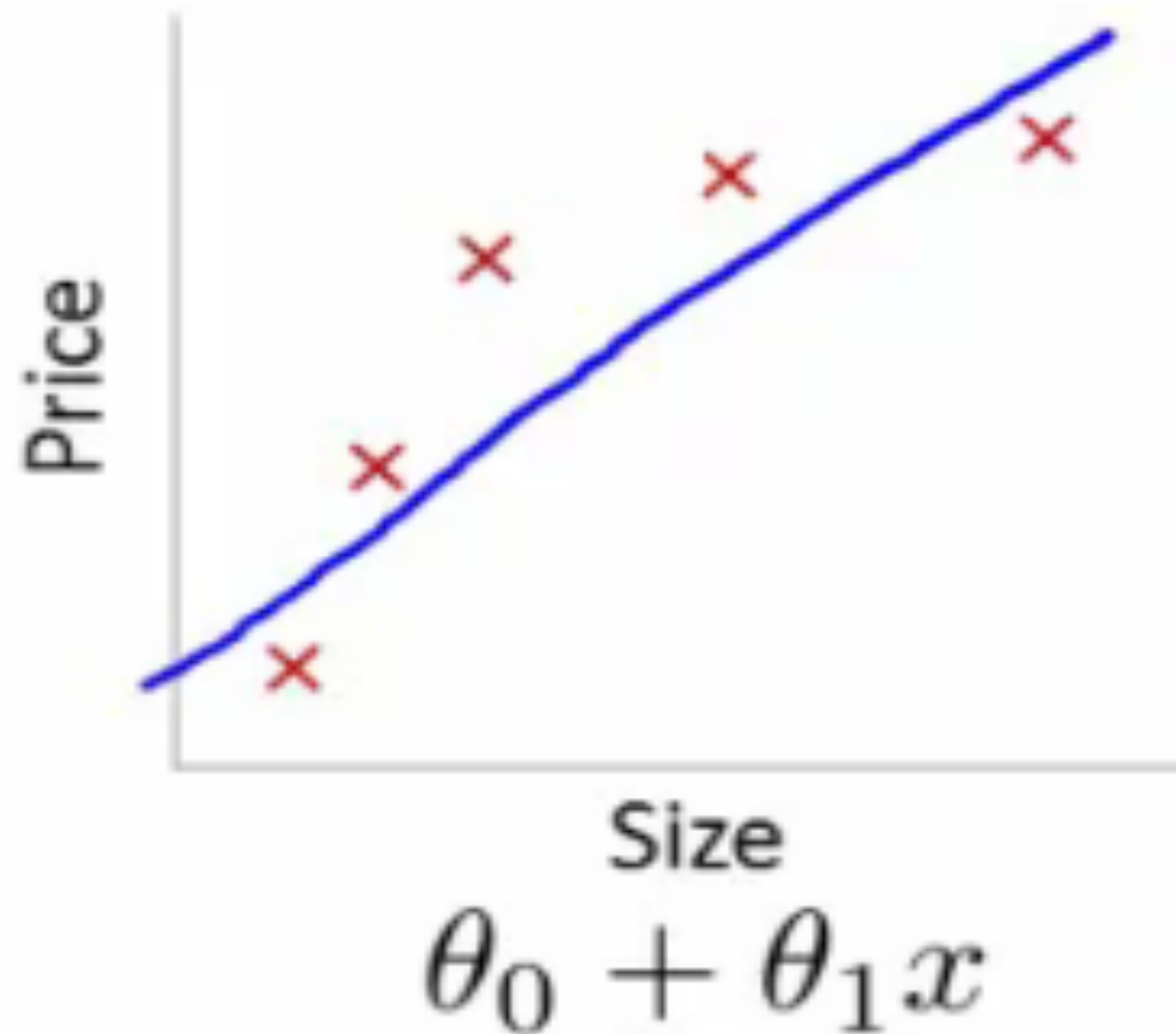
# Overfitting

- El modelo **capta el ruido** de los datos.
- Se da cuando existe **poco sesgo** y **alta varianza**.
- Es resultado de un **modelo extremadamente complicado**.





# Underfitting



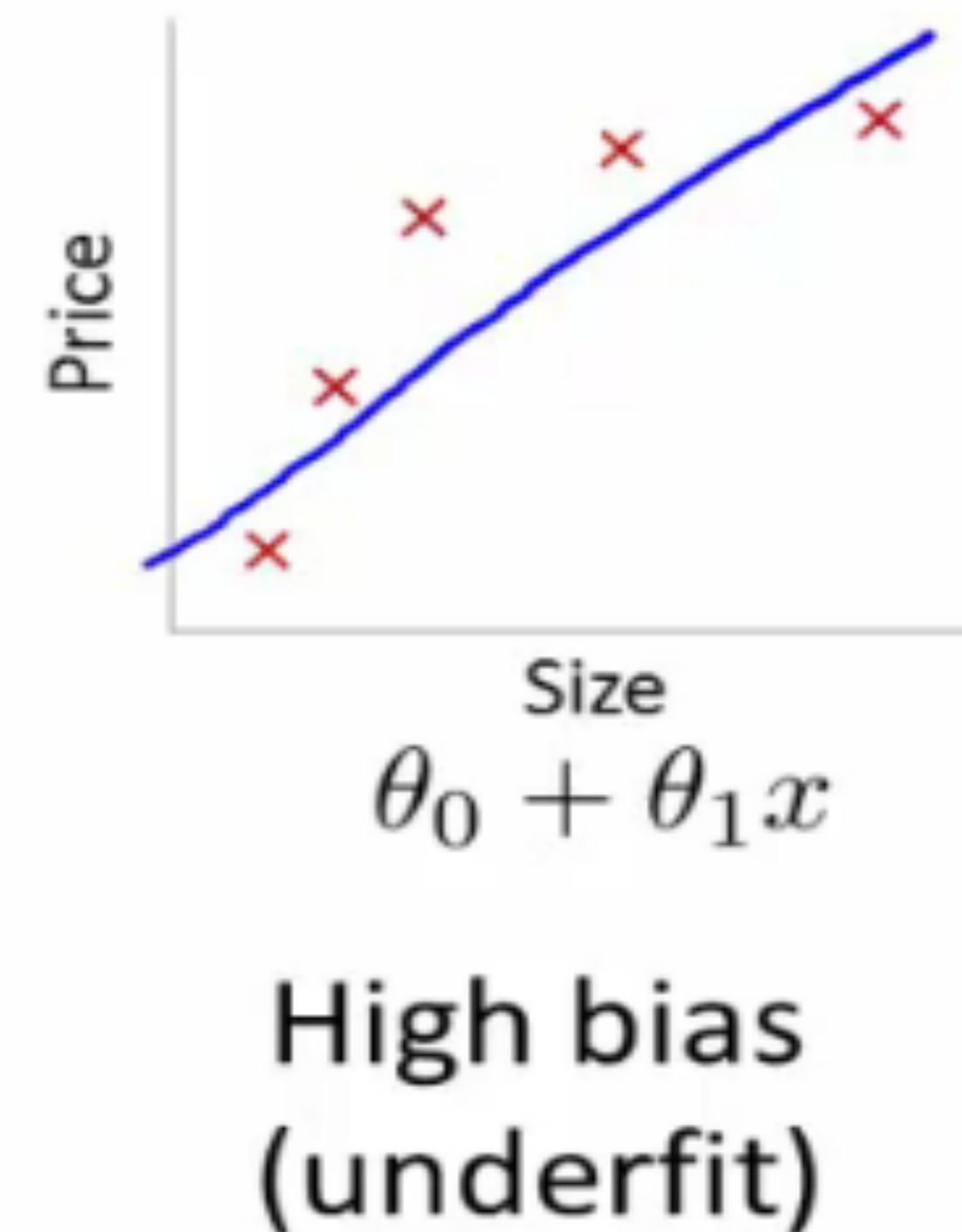
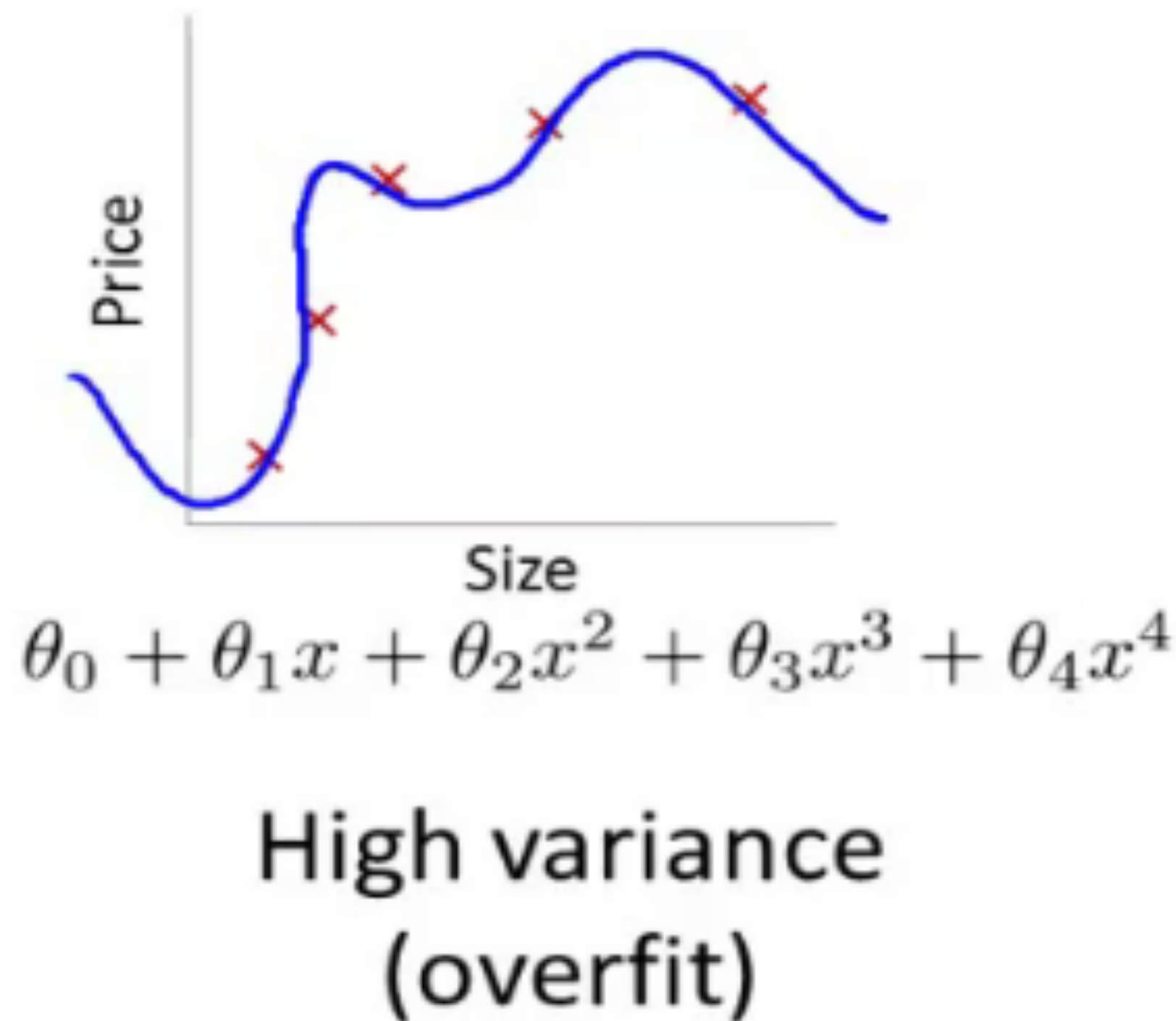
High bias  
(underfit)

- El modelo **no logra captar el ruido** de los datos.
- Se da cuando existe **alto sesgo** y **varianza pequeña**.
- Es resultado de un **modelo extremadamente simple**.



# Overfitting vs Underfitting

- El **overfitting** representa el **sobreaprendizaje** de un modelo.
- El **underfitting** representa el **subaprendizaje** de un modelo.





# Matriz de confusión

- Contiene información acerca de la **clase actual** y la **clase predicha** por un clasificador.
- Las **filas** representan la clase actual y las **columnas** la clase predicha.
- Se busca que la diagonal principal tenga los valores más altos.

	Clase predicha	
	yes	no
Clase real	a	b
	c	d

$$AC = \frac{a + d}{a + b + c + d}$$





# Matriz de Confusión

## Actividad en Weka

- Ejecutar el algoritmo **Naïve Bayes** usando el conjunto de datos **weather.nominal.arff** y analice la matriz de confusión el modelo algoritmo resultante. Calcule la exactitud del modelo usando la matriz de confusión.
- Ejecutar el algoritmo **J48** usando el conjunto de datos **weather.nominal.arff** y analice la matriz de confusión el modelo algoritmo resultante. Calcule la exactitud del modelo usando la matriz de confusión.



# Matriz de Confusión

## Actividad en Weka

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

**Classifier**

Choose J48 -C 0.25 -M 2

**Test options**

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds 10

☐ Percentage split % 66

More options...

(Nom) play

Start Stop

**Result list (right-click for options)**

18:44:03 - bayes.NaiveBayes

18:44:10 - trees.J48

**Classifier output**

=== Classifier model (full training set) ===

J48 pruned tree

-----

outlook = sunny  
| humidity = high: no (3.0)  
| humidity = normal: yes (2.0)  
outlook = overcast: yes (4.0)  
outlook = rainy  
| windy = TRUE: no (2.0)  
| windy = FALSE: yes (3.0)

Number of Leaves : 5

Size of the tree : 8

Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===

=== Summary ===

=== Confusion Matrix ===

a	b		<-- classified as
5	4		a = yes
3	2		b = no



# Matriz de Confusión

## Actividad en RapidMiner

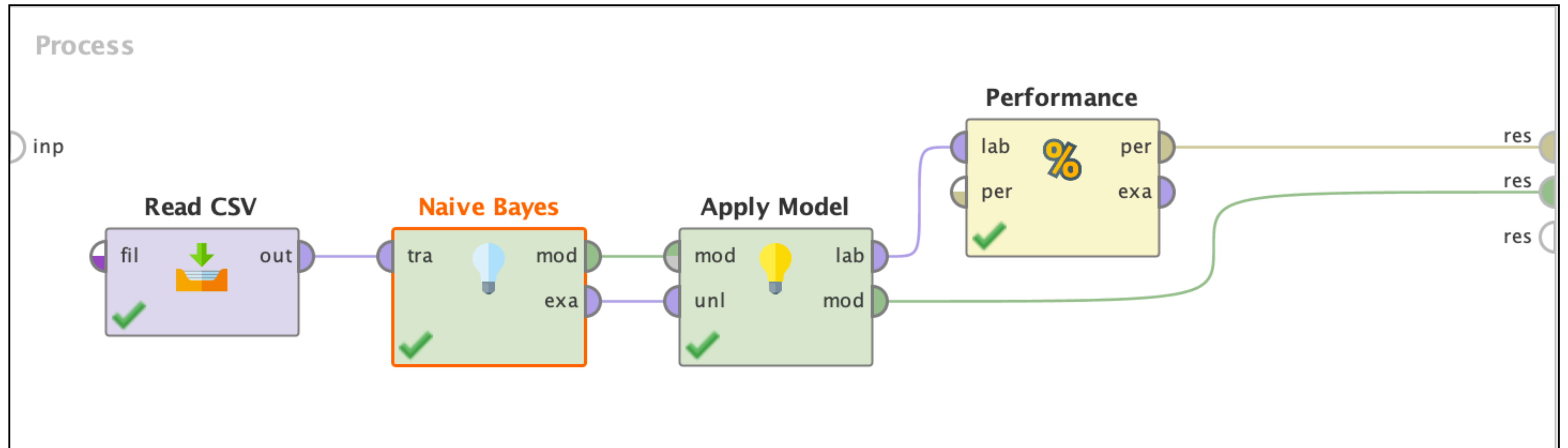
- Ejecutar el algoritmo **Naïve Bayes** usando el conjunto de datos **weather.nominal.csv** y analice la matriz de confusión el modelo algoritmo resultante. Calcule la exactitud del modelo usando la matriz de confusión.
- Ejecutar el algoritmo **Decision Tree** usando el conjunto de datos **weather.nominal.csv** y analice la matriz de confusión el modelo algoritmo resultante. Calcule la exactitud del modelo usando la matriz de confusión.





# Matriz de Confusión

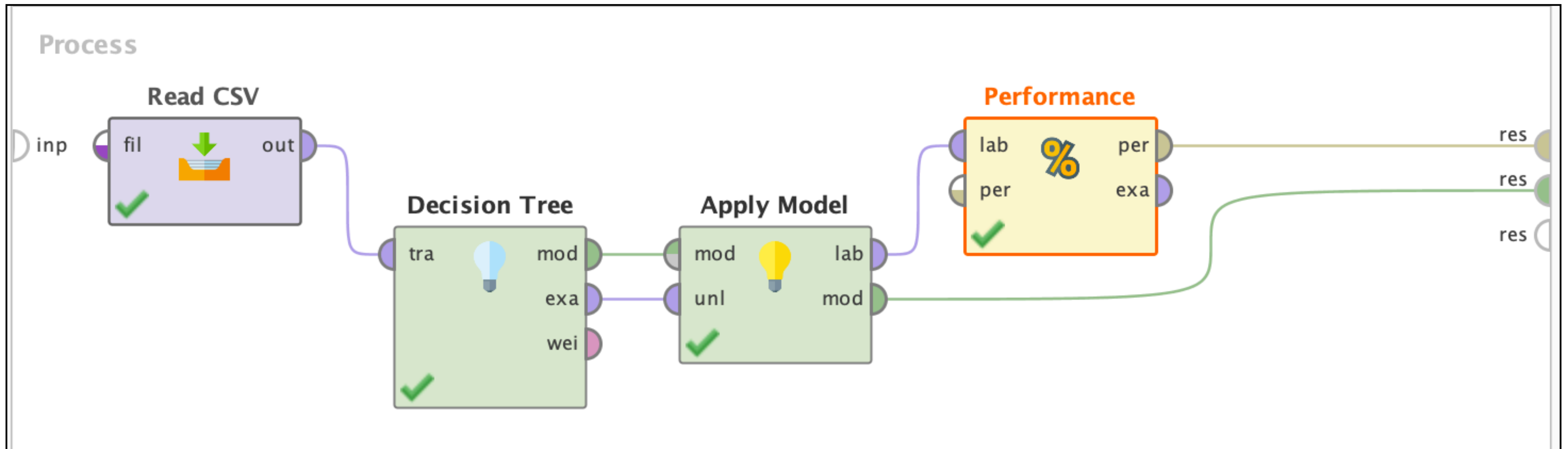
## Actividad en RapidMiner





# Matriz de Confusión

## Actividad en RapidMiner





# Matriz de Confusión

## Actividad en Python

- Ejecutar el algoritmo **Naïve Bayes** usando el conjunto de datos **iris.data** y analice la matriz de confusión el modelo algoritmo resultante. Calcule la exactitud del modelo usando la matriz de confusión.





# Matriz de Confusión

## Actividad en Python

*# Cargamos el conjunto de datos*

**import** pandas

archivo="iris.data"

columnas=['longitud-sépalo', 'ancho-sépalo', 'longitud-pétalo', 'ancho-pétalo', 'clase']

conjunto\_de\_datos = pandas.read\_csv(archivo, names=columnas)

X = conjunto\_de\_datos.iloc[:,0:4].values

y = conjunto\_de\_datos.iloc[:,4].values

**print**(X)

**print**(y)

```
[ [ 5.1  3.5  1.4  0.2 ]  
  [ 4.9  3.   1.4  0.2 ]  
  [ 4.7  3.2  1.3  0.2 ]  
  [ 4.6  3.1  1.5  0.2 ]  
  [ 5.   3.6  1.4  0.2 ]  
  [ 5.4  3.9  1.7  0.4 ]  
  [ 4.6  3.4  1.4  0.3 ]
```

```
[ 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa'  
  'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa'  
  'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa'  
  'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa'
```



# Matriz de Confusión

## Actividad en Python

```
from sklearn.naive_bayes import GaussianNB  
from sklearn.metrics import accuracy_score  
from sklearn.metrics import confusion_matrix
```

```
gnb = GaussianNB()  
modelo = gnb.fit(X, y)  
y_predecido = modelo.predict(X)
```

0.96

```
print(accuracy_score(y, y_predecido))  
print(confusion_matrix(y, y_predecido))
```

```
[ [ 50   0   0 ]  
  [  0  47   3 ]  
  [  0   3  47 ] ]
```



# Precisión y Recall

- La **precisión** y la **exhaustividad (recall)** son dos métricas empleada en la **medida del rendimiento** de los sistemas de búsqueda, recuperación de información y reconocimiento de patrones.
- En este contexto se denomina:
  - **Precisión** como a la fracción de instancias recuperadas que son relevantes.
  - **Recall** es la fracción de instancias relevantes que han sido recuperadas.
- Tanto la precisión como la exhaustividad son entendidas como **medidas de la relevancia**.






# Conceptos

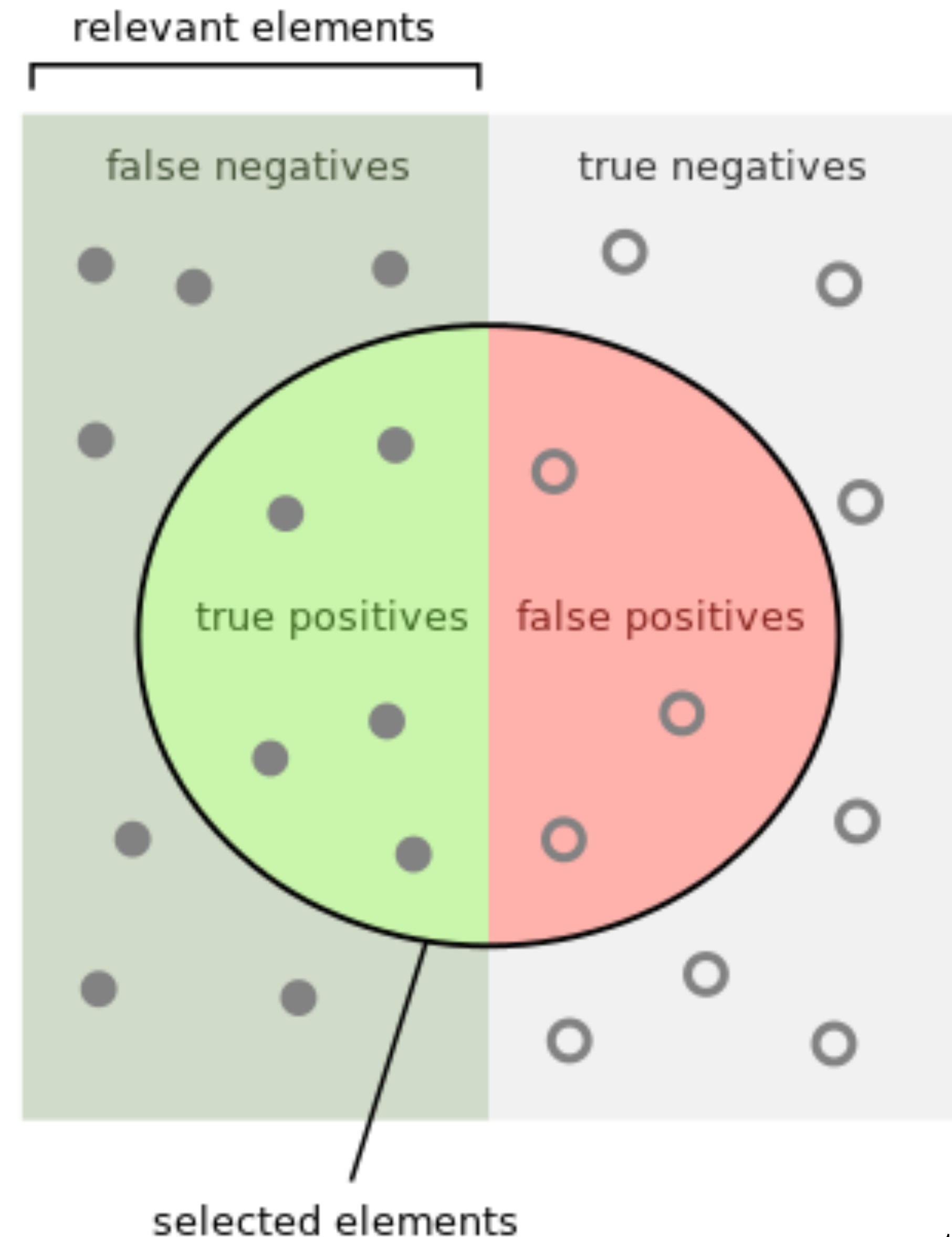
- **True Positives (TP):** son instancias pertenecientes a la clase que se clasifican correctamente en dicha clase.
- **True Negatives (TN):** son instancias no pertenecientes a la clase y que no se clasifican como dicha clase.
- **False Positives (FP):** son instancias no pertenecientes a la clase pero que se clasifican como dicha clase.
- **False Negatives (FN):** son instancias pertenecientes a la clase pero que no se clasifican como dicha clase.



# Precisión

- Propuesto por Gerald Salton en 1983.
- **Fracción de instancias recuperadas que son relevantes.**
- Si el resultado es 1, todos los documentos son relevantes.

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$


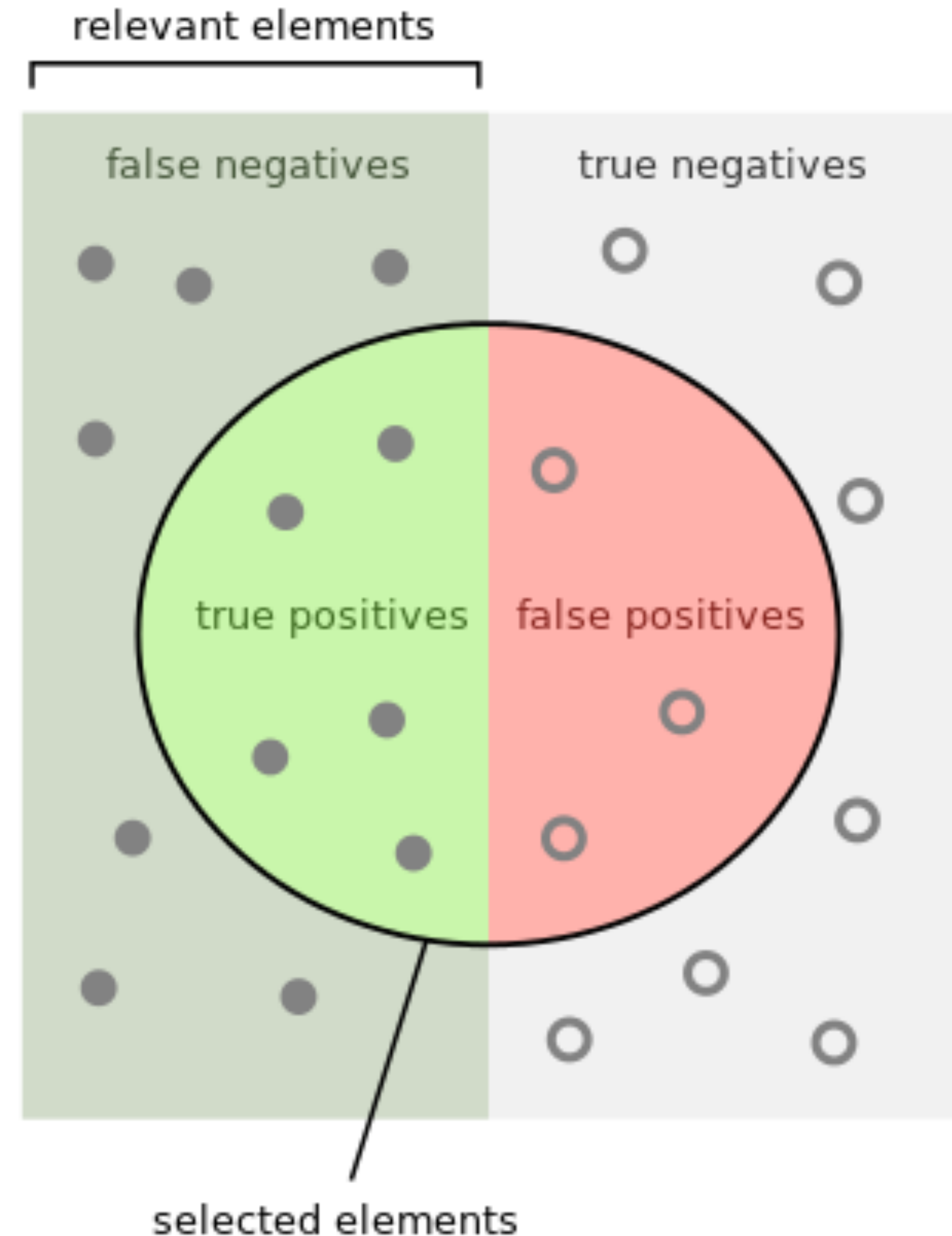




# Recall

- Propuesto a mediados del siglo XX.
- **fracción de instancias relevantes que han sido recuperadas.**
- Cuantos elementos relevantes son seleccionados.
- Si el valor es 1, se encontraron todos los documentos relevantes.

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$







# F-Measure

- Medida de precisión que tiene un *test*.
- Obtener valor único ponderado de precisión y recall.

$$F_{\beta} = \frac{Precision \times Recall}{(\beta^2 \times Precision) + Recall}$$

- Si  $\beta$  es igual a uno, se está dando la misma importancia a Precisión que al Recall.
- Si  $\beta$  es mayor que uno le damos más importancia al Recall.
- Si  $\beta$  es menor que uno se le da más importancia a la Precisión.



# Precisión, Recall y F1

## Actividad en Weka

- Ejecutar el algoritmo **One Rule** usando el conjunto de datos **weather.nominal.arff** y analice la precisión, recall y número F1.



# Precisión, Recall y F1

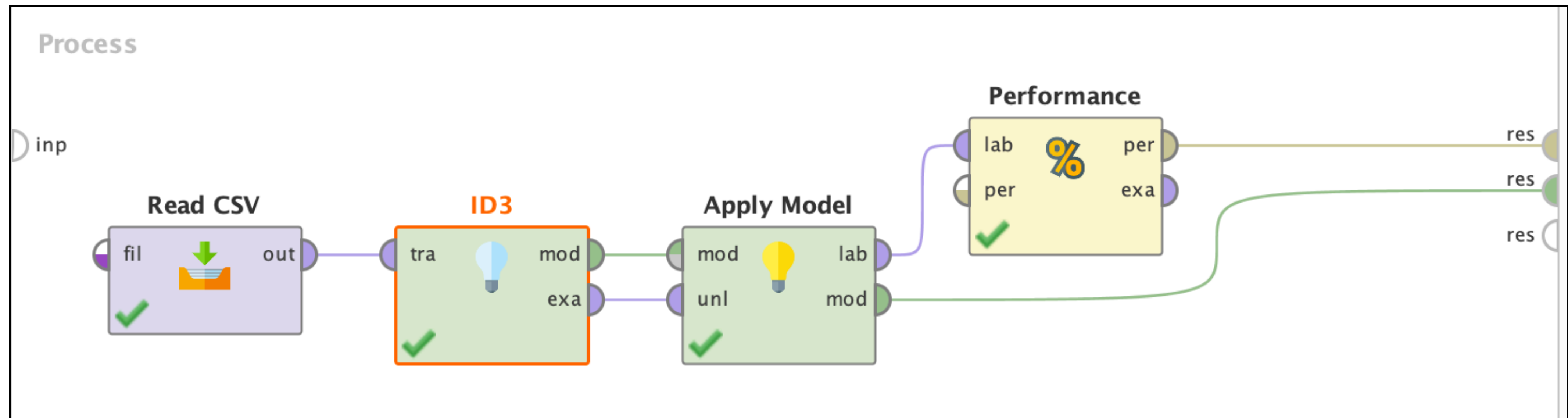
## Actividad en RapidMiner

- Ejecutar el algoritmo **IDE3** usando el conjunto de datos **weather.nominal.csv** y analice la precisión, recall y número F1.



# Precisión, Recall y F1

## Actividad en RapidMiner







# Precisión, Recall y F1

## Actividad en Python

*# Cargamos el conjunto de datos*

**import** pandas

archivo="iris.data"

columnas=['longitud-sépalo', 'ancho-sépalo', 'longitud-pétalo', 'ancho-pétalo', 'clase']

conjunto\_de\_datos = pandas.read\_csv(archivo, names=columnas)

X = conjunto\_de\_datos.iloc[:,0:4].values

y = conjunto\_de\_datos.iloc[:,4].values

**print**(X)

**print**(y)

```
[ [5.1 3.5 1.4 0.2]
  [4.9 3.  1.4 0.2]
  [4.7 3.2 1.3 0.2]
  [4.6 3.1 1.5 0.2]
  [5.  3.6 1.4 0.2]
  [5.4 3.9 1.7 0.4]
  [4.6 3.4 1.4 0.3]
```

```
['Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa'
 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa'
 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa'
 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa' 'Iris-setosa']
```



# Precisión, Recall y F1

## Actividad en Python

```
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
```

```
gnb = GaussianNB()
modelo = gnb.fit(X, y)
y_predecido = modelo.predict(X)
```

0.96

```
print(accuracy_score(y, y_predecido))
print(classification_report(y, y_predecido))
```

	precision	recall	f1-score	support
Iris-setosa	1.00	1.00	1.00	50
Iris-versicolor	0.94	0.94	0.94	50
Iris-virginica	0.94	0.94	0.94	50
micro avg	0.96	0.96	0.96	150
macro avg	0.96	0.96	0.96	150
weighted avg	0.96	0.96	0.96	150



# Competencias a adquirir en la sesión

- Al finalizar la sesión el alumno comprenderá los conceptos de **overfitting** y **underfitting**.
- Al finalizar la sesión el alumno analizará la **matriz de confusión** resultantes de los modelos algorítmicos.
- Al finalizar la sesión el alumno comprenderá los conceptos de **verdaderos positivos, verdaderos negativos, falsos positivos** y **falsos negativos**.
- Al finalizar la sesión el alumno **analizará** modelos algorítmicos usando la **precisión, recall** y **número F**.

