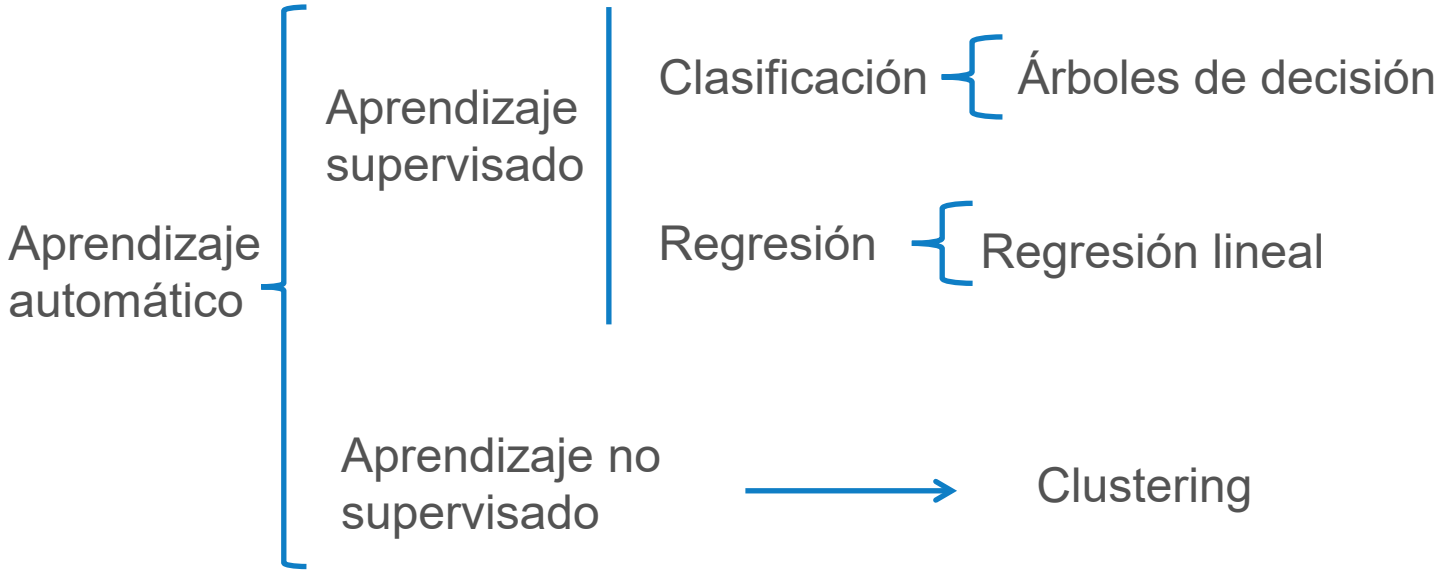
Inteligencia Artificial e Ingeniería del Conocimiento

Elena Verdú Pérez

Aprendizaje Supervisado (Continuación)

¿Cómo estudiar este tema?

SEMANAS	TEMAS	ACTIVIDADES (15.0 PUNTOS)	CLASES EN DIRECTO
Semana7 25-abr-2022 - 29-abr-2022	**Tema 6. Aprendizaje supervisado** 6.1. ¿Cómo estudiar este tema? 6.2. Introducción al Aprendizaje Automático 6.3. Clasificación		Clase del tema 6
Semana8 02-may-2022 - 06-may-2022	**Tema 6. Aprendizaje supervisado (continuación)** 6.4. Regresión 6.5. Validación de resultados	Actividad: Laboratorio Aprendizaje Supervisado con Weka (5.0 puntos) Fecha de entrega: 23/05/2022 Test - Tema 06 (0.1 puntos) Fecha de entrega: 03/07/2022	Clase del tema 6 y presentación del laboratorio (2h x 2 turnos)



¿Cómo estudiar este tema?

- Ideas clave
- Lectura de **páginas 71-75, 77-81 y 83-85** del libro: Gironés, J., Casas, J., Minguillón, J. y Caihuelas, R. (2017). *Minería de datos: modelos y algoritmos.* Barcelona: Editorial UOC.

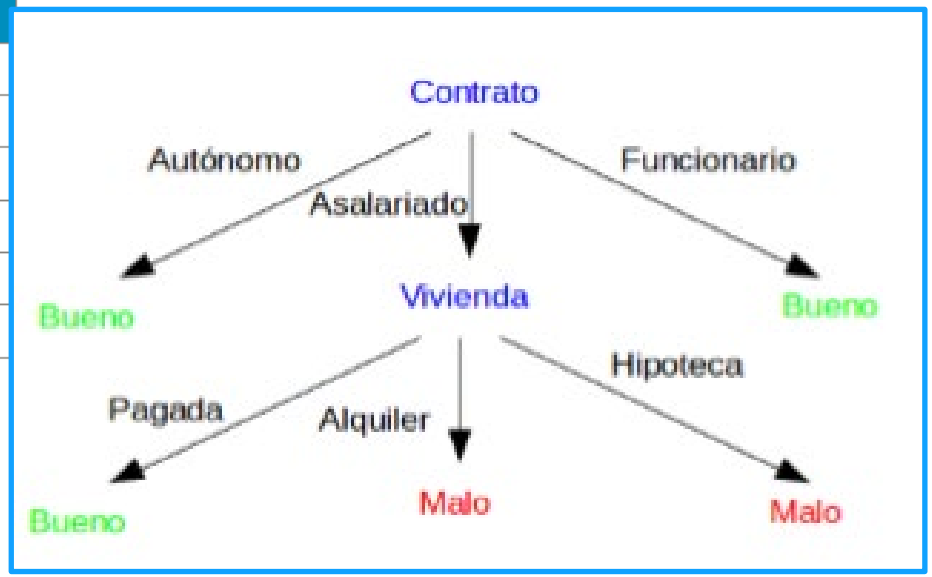
» TEMA 6. APRENDIZAJE SUPERVISADO

IDEAS CLAVE	LO + RECOMENDADO	+ INFORMACIÓN
¿Cómo estudiar este tema?	**Lecciones magistrales**	**A fondo**
Introducción al Aprendizaje Automático	TV Aprendizaje supervisado	Aprendizaje Automático en Open Course Ware
Clasificación	**No dejes de leer...**	Six Novel Machine Learning Applications
Regresión	Machine Learning y Data Mining	**Bibliografía**
Validación de resultados	**No dejes de ver...**	**Recursos externos**
	Aprendizaje Automático en Coursera TV	WEKA

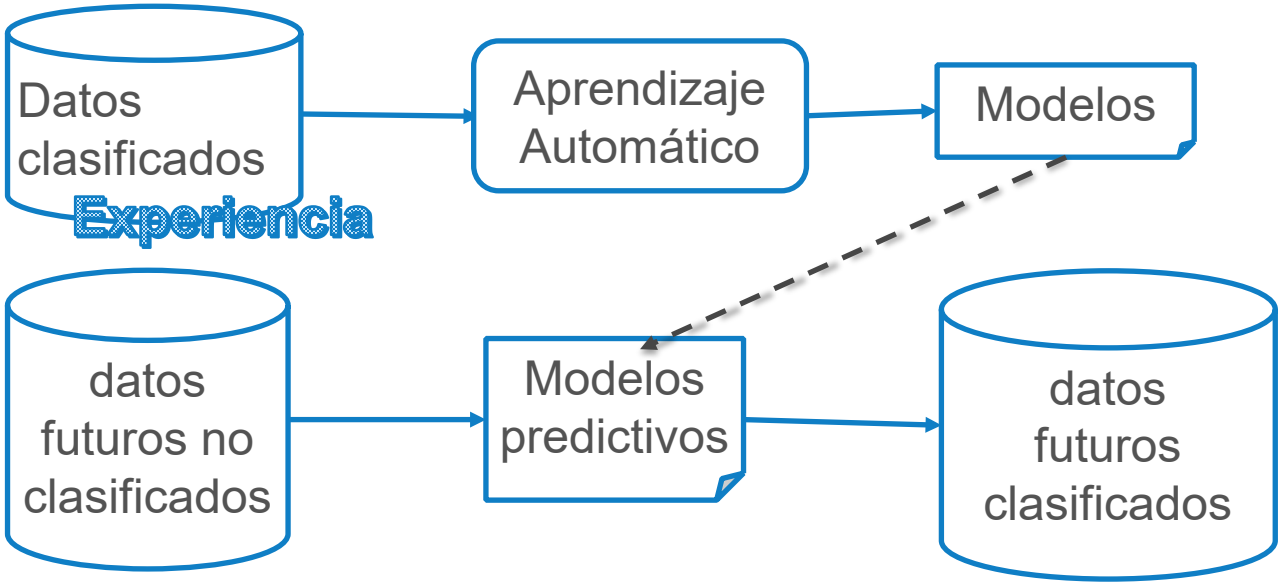
Validación

- El modelo generado “describe” los datos de los clientes ya clasificados.
- Ante un cliente que se presente en el futuro ¿será el modelo generado capaz de clasificarlo correctamente?

Vivienda	Hijos	Tarjeta	Contrato	Tipo
Hipoteca	0	Débito	Funcionario	Bueno
Hipoteca	0	Crédito	Asalariado	Malo
Hipoteca	2	Débito	Autónomo	Bueno
Pagada	2	Débito	Asalariado	Bueno
Hipoteca	1	Débito	Asalariado	Malo
Alquiler	2	Débito	Asalariado	Malo



Validación



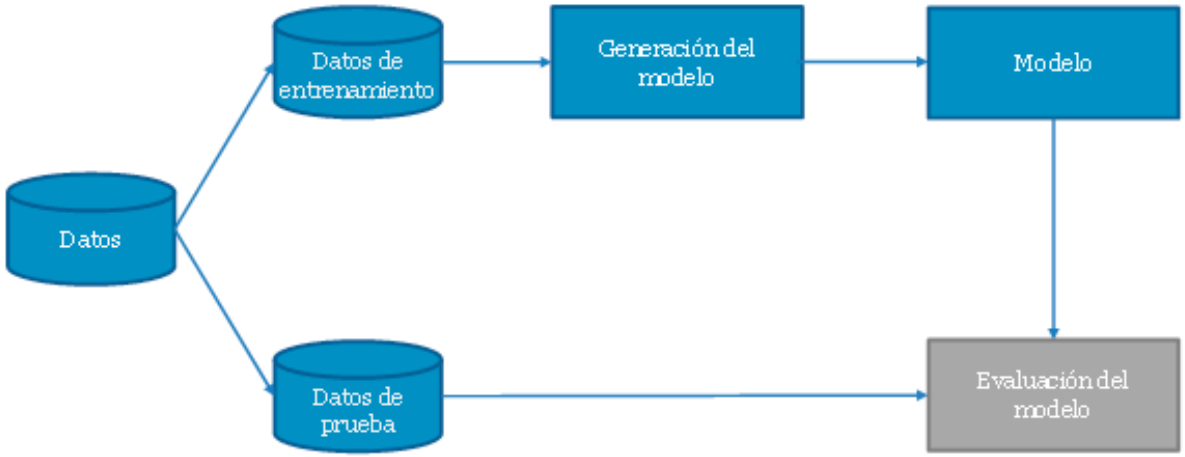
¿podemos estar seguros de que esas futuras instancias estarán correctamente clasificadas?

Validación

- La validación del modelo permite medir su capacidad de predicción de la clase de nuevas instancias que le lleguen en un futuro.

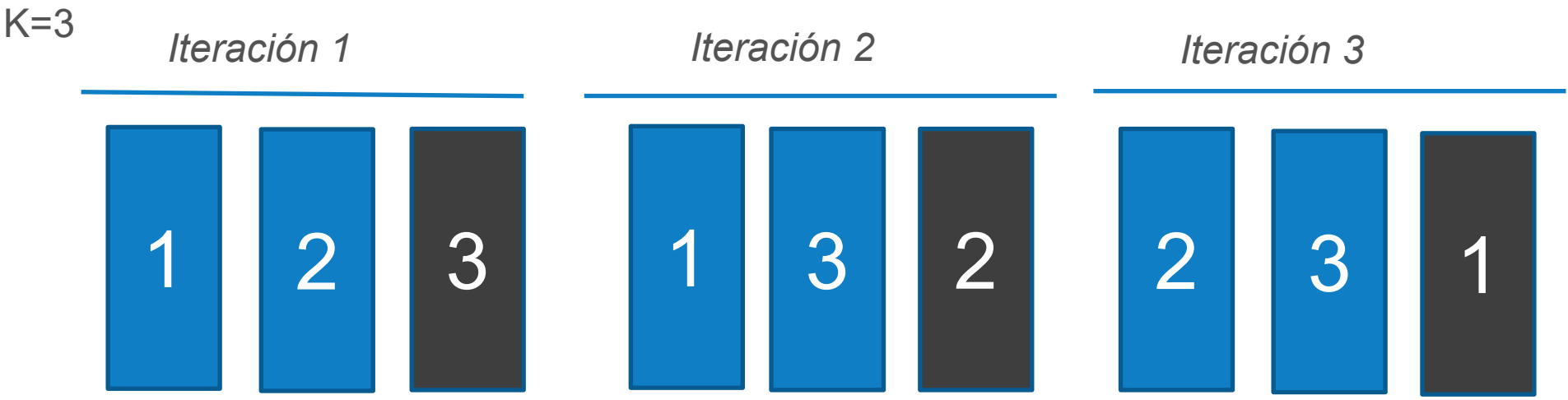
- Podemos utilizar la tasa de error del clasificador al predecir la clase de un conjunto de datos de prueba.

Tasa de error = número de errores / número total de instancias de prueba



Validación. Validación cruzada

- Validación cruzada de *k-iteraciones* *(K-fold cross validation)*
  - Los datos se dividen en k particiones disjuntas de igual tamaño
  - Se realizan k iteraciones, en cada una:
    - Generación del modelo a partir de k-1 participaciones
    - Validación mediante la partición no utilizada
      - en la generación del modelo de esta iteración
      - en la validación de iteraciones previas
  - Se promedian los resultados de evaluación de las k iteraciones



Validación. Matriz de confusión

- Cada elemento de la matriz → número de ejemplos de prueba cuya clase real es la indicada en la cabecera de la fila y la clase estimada es la indicada en la cabecera de la columna.

- Tasa de Éxito es la suma de los valores en la diagonal dividido por el total de instancias.

predicción →

```
a  b  c   <-- classified as
5  0  0 |  a = soft
0  3  1 |  b = hard
1  2 12 |  c = none
```

↑ Clase real

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

                 TP Rate   FP Rate
                   1         0.053
                   0.75      0.1
                   0.8       0.111
Weighted Avg.      0.833     0.097
```

Validación. Tasas TP/FP

- 2 clases
- **TP** y **TN** → clasificaciones correctas

		Clase Predicha	
		sí	No
Clase Real	Sí	Verdadero Positivo (TP)	Falso Negativo (FN)
	no	Falso Positivo (FP)	Verdadero Negativo (TN)

- **FP** → Una instancia es incorrectamente clasificada como “Sí” o positivo cuando es negativa
- **FN** → Una instancia es incorrectamente clasificada como “No” o negativa cuando es positiva

Validación. Tasas TP/FP

- Tasa de verdaderos positivos (TP)

$$\frac{TP}{TP + FN}$$

- Tasa de falsos positivos (FP)

$$\frac{FP}{FP + TN}$$

```
=== Confusion Matrix ===

 a b   <-- classified as
 7 2 | a = yes
 1 3 | b = no
```

¿Tasa de verdaderos positivos?

$$\frac{7}{7 + 2}$$

Validación. Tasas TP/FP

- Tasa de verdaderos positivos (TP)

$$\frac{TP}{TP + FN}$$

- Tasa de falsos positivos (FP)

$$\frac{FP}{FP + TN}$$

```
=== Confusion Matrix ===

 a b   <-- classified as
 7 2 | a = yes
 1 3 | b = no
```

¿Tasa de falsos positivos?

$$\frac{1}{1 + 3}$$

Validación. Precisión

- Precisión

$$\frac{TP}{TP + FP}$$

```
=== Confusion Matrix ===

 a b   <-- classified as
 7 2 | a = yes
 1 3 | b = no
```

¿Precisión?

$$\frac{7}{7 + 1}$$

Validación. Tasa de éxito

- Tasa de éxito general

$$t_{exito} = \frac{TP + TN}{TP + TN + FP + FN}$$

- Tasa de error

$$t_{error} = 1 - t_{exito}$$

```
=== Confusion Matrix ===

 a b   <-- classified as
 7 2 | a = yes
 1 3 | b = no
```

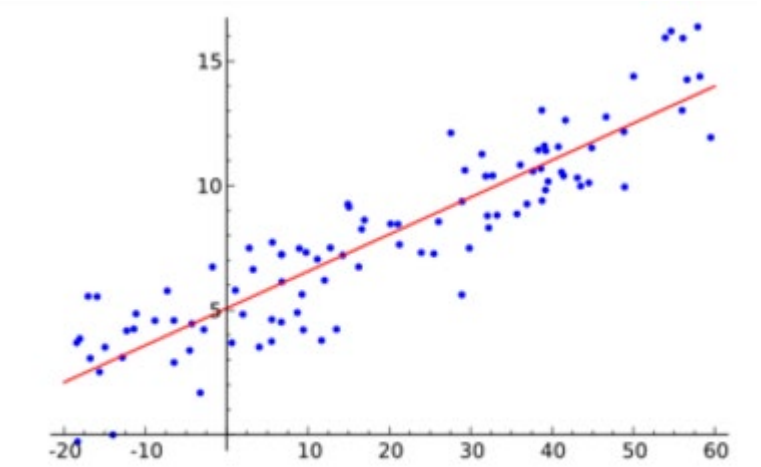
¿Tasa de éxito?

$$\frac{7 + 3}{7 + 2 + 1 + 3}$$

Regresión

La variable de salida es continua → Modelo de regresión

La **regresión lineal** consiste en **encontrar una función lineal** lo más cercana posible a la función real del modelo.



Fuente: wikipedia

Regresión

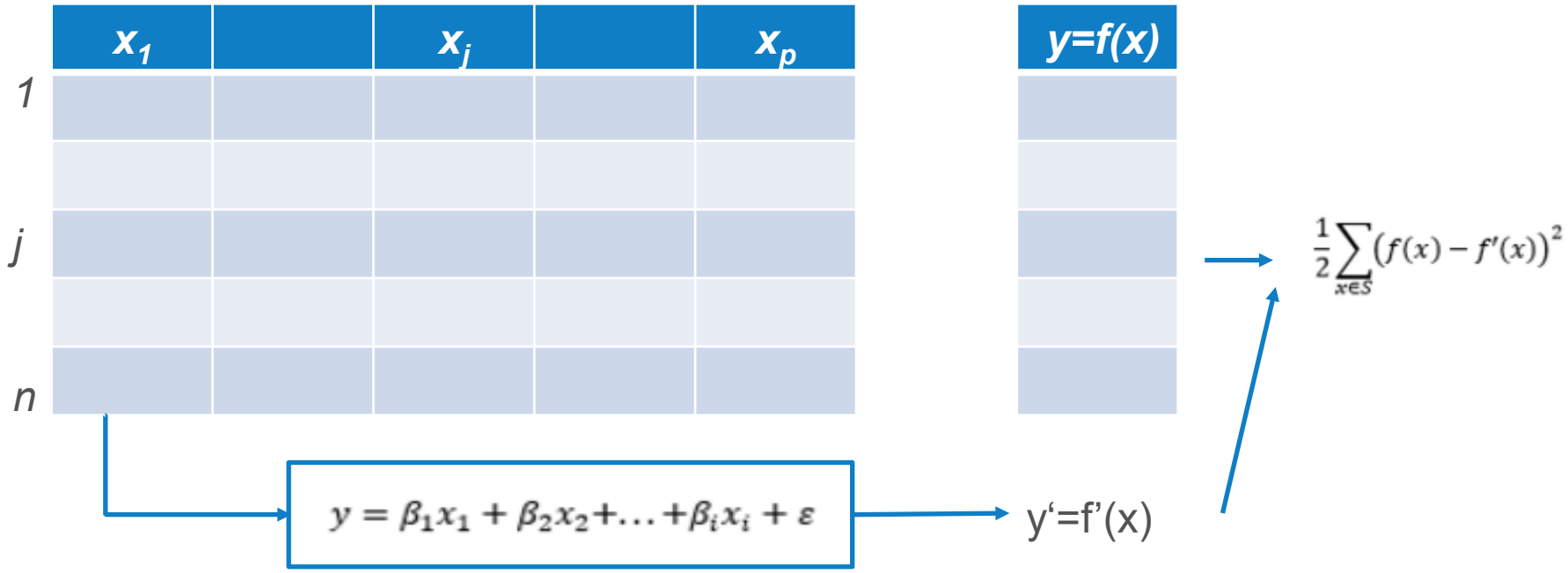
Las funciones lineales son una combinación lineal de sus parámetros más un valor constante (a menudo denominado término de error o ruido):

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + \varepsilon$$

Aprendizaje automático → estimación de la función de regresión lineal → aprender los parámetros β_1, β_2, β_i… y ε que mejor ajustan el modelo lineal al comportamiento real de los ejemplos

Regresión

- Conjunto de ejemplos S (ej: *n instancias con p atributos o parámetros continuos y la salida continua*)



Método de minimización del error entre la función aproximada f'(x) y los valores reales de los ejemplos f(x) → minimización de la suma del error cuadrático sobre el conjunto de entrenamiento total.

Regresión

El método de mínimos cuadrados consiste en una búsqueda en el espacio de parámetros β_1, β_2, β_i… y ε tal que se minimice la suma del error cuadrático sobre los ejemplos.

Algoritmo- Método Descenso de gradiente

1. **Se inicializan los parámetros a un valor inicial aleatorio pequeño**
2. **Para cada parámetro β_i se define un diferencial $\Delta\beta_i$ inicializado a 0**
3. **Dado un ratio de aprendizaje η definido por el usuario, por cada ejemplo x modificar $\Delta\beta_i$ de la siguiente manera: $\Delta\beta_i = \Delta\beta_i - \eta(f'(x) - f(x))x_i$**
4. **Cada parámetro β_i se modifica de la siguiente manera: $\beta_i = \beta_i + \Delta\beta_i$**

Validación en predicción numérica

El **error absoluto medio** (*mean absolute error-MAE*) es un promedio de los errores de clasificación de cada una de las instancias. Si tenemos n instancias con unos valores predichos $p_1, p_2, p_3 \ldots p_n$, y unos valores reales $x_1, x_2, x_3, \ldots x_n$, el error absoluto medio se calcula según la siguiente expresión:

$$MAE = \frac{|p_1 - x_1| + |p_2 - x_2| + \cdots + |p_n - x_n|}{n}$$

Si tenemos n instancias con unos valores predichos $p_1, p_2, p_3 \ldots p_n$, y unos valores reales $x_1, x_2, x_3, \ldots x_n$, la **raíz del error cuadrático medio** se calcula según la siguiente expresión:

$$RMSE = \sqrt{\frac{(p_1 - x_1)^2 + (p_2 - x_2)^2 + \cdots + (p_n - x_n)^2}{n}}$$

¿Dudas?

¡Muchas gracias por vuestra atención!

¡Feliz y provechosa semana!

www.unir.net