Clase 16 Modelamiento predictivo: Regresión lineal

Diplomado en Análisis de Datos y Modelamiento Predictivo con Aprendizaje Automático para la Acuicultura

Dra. María Angélica Rueda Calderón

Pontificia Universidad Católica de Valparaíso

17 June 2023

PLAN DE LA CLASE

1.- Introducción

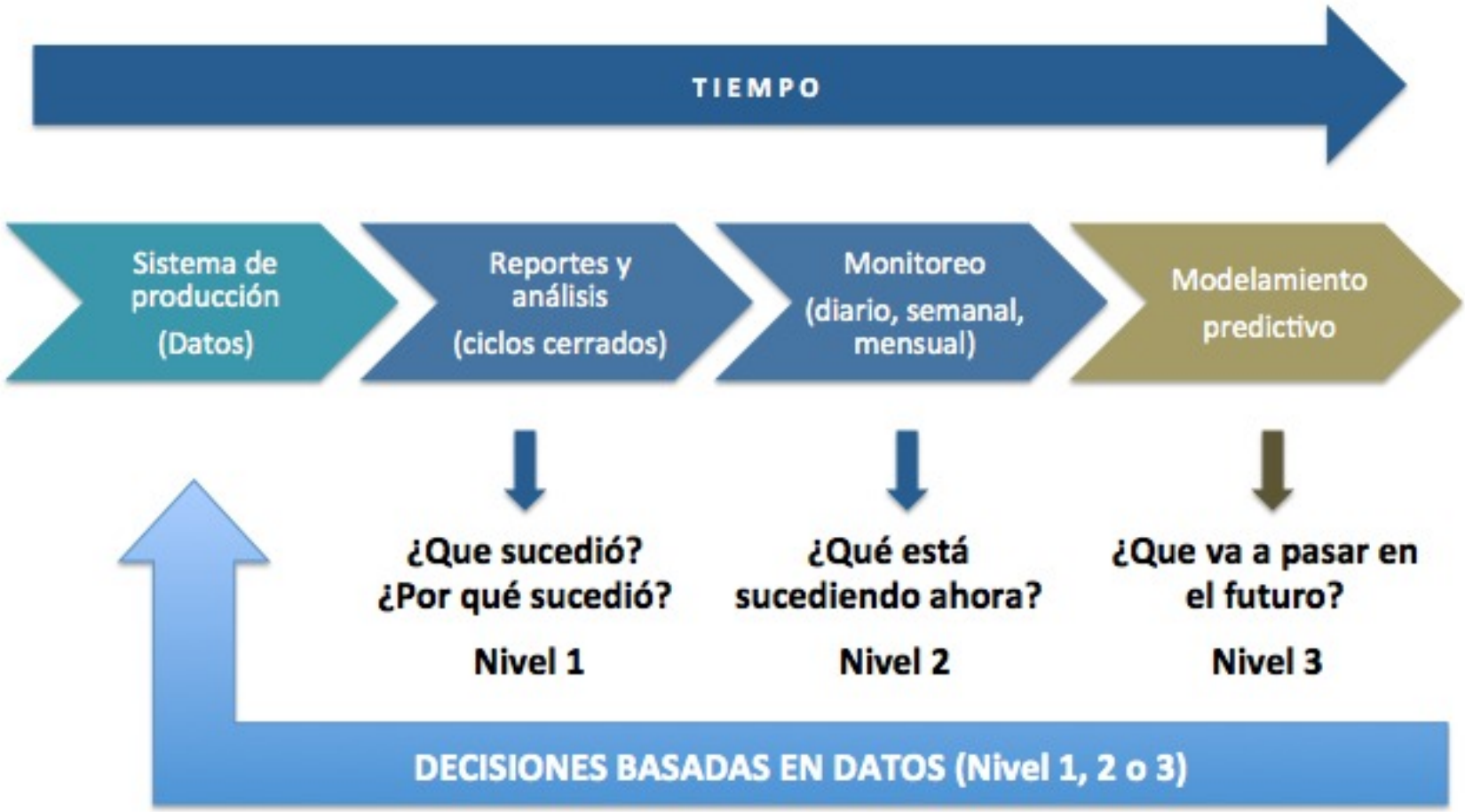
- ¿Qué es el modelamiento predictivo?
- Tipos de modelos predictivos.
- Etapas del modelamiento predictivo.
- Partición del conjunto de datos:entrenamiento y testeo.
- Métodos de validación.
- ¿Cómo medir la precisión/desempeño del modelo?

2.- Práctica con R y Rstudio cloud.

- Realizar modelamiento predictivo para regresión lineal.
- Realizar gráficas avanzadas con ggplot2.

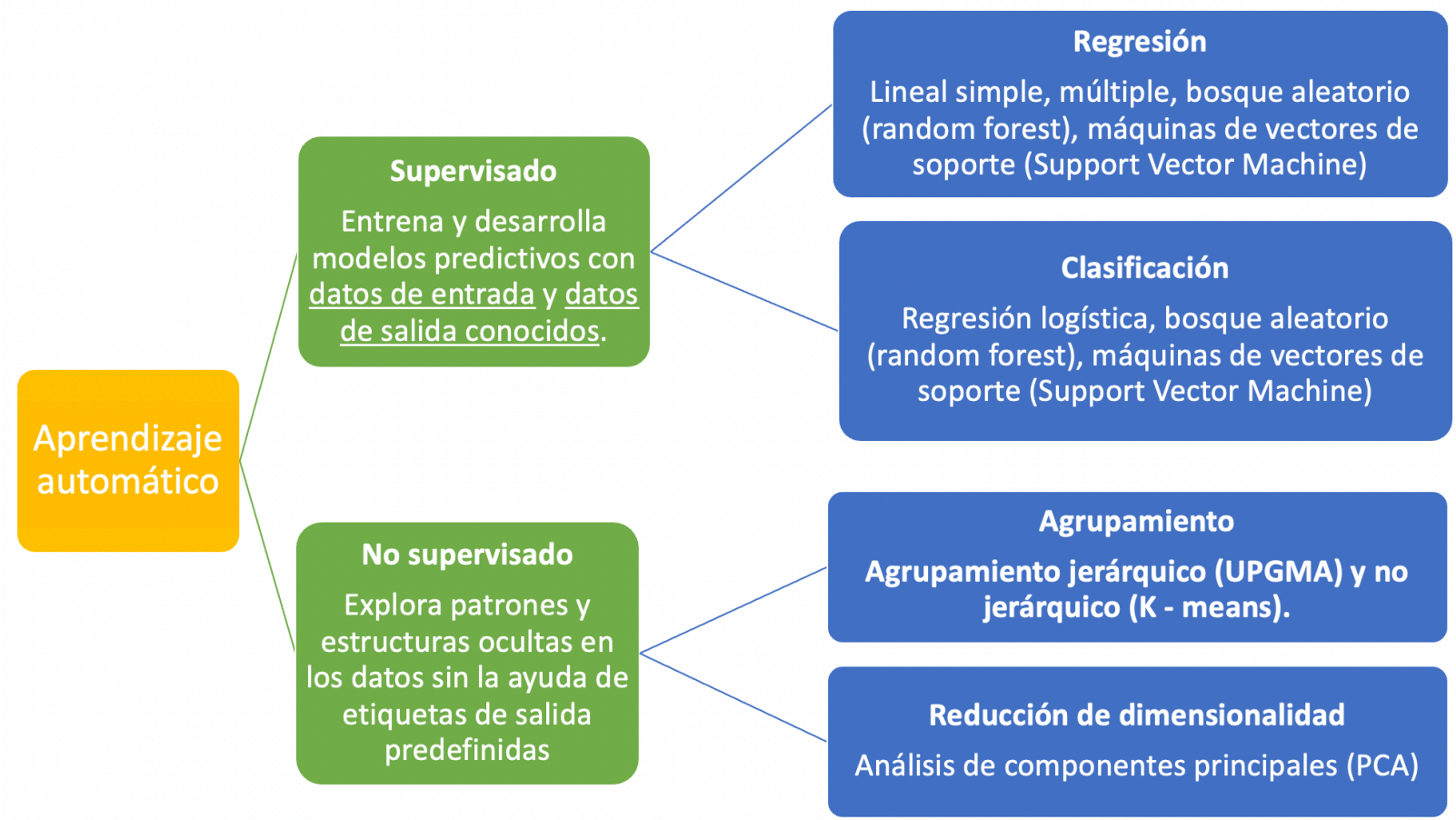
¿QUÉ ES EL MODELAMIENTO PREDICTIVO?

El modelamiento predictivo es una técnica utilizada en el campo del análisis de datos y la inteligencia artificial para predecir eventos futuros o hacer estimaciones basadas en datos históricos.

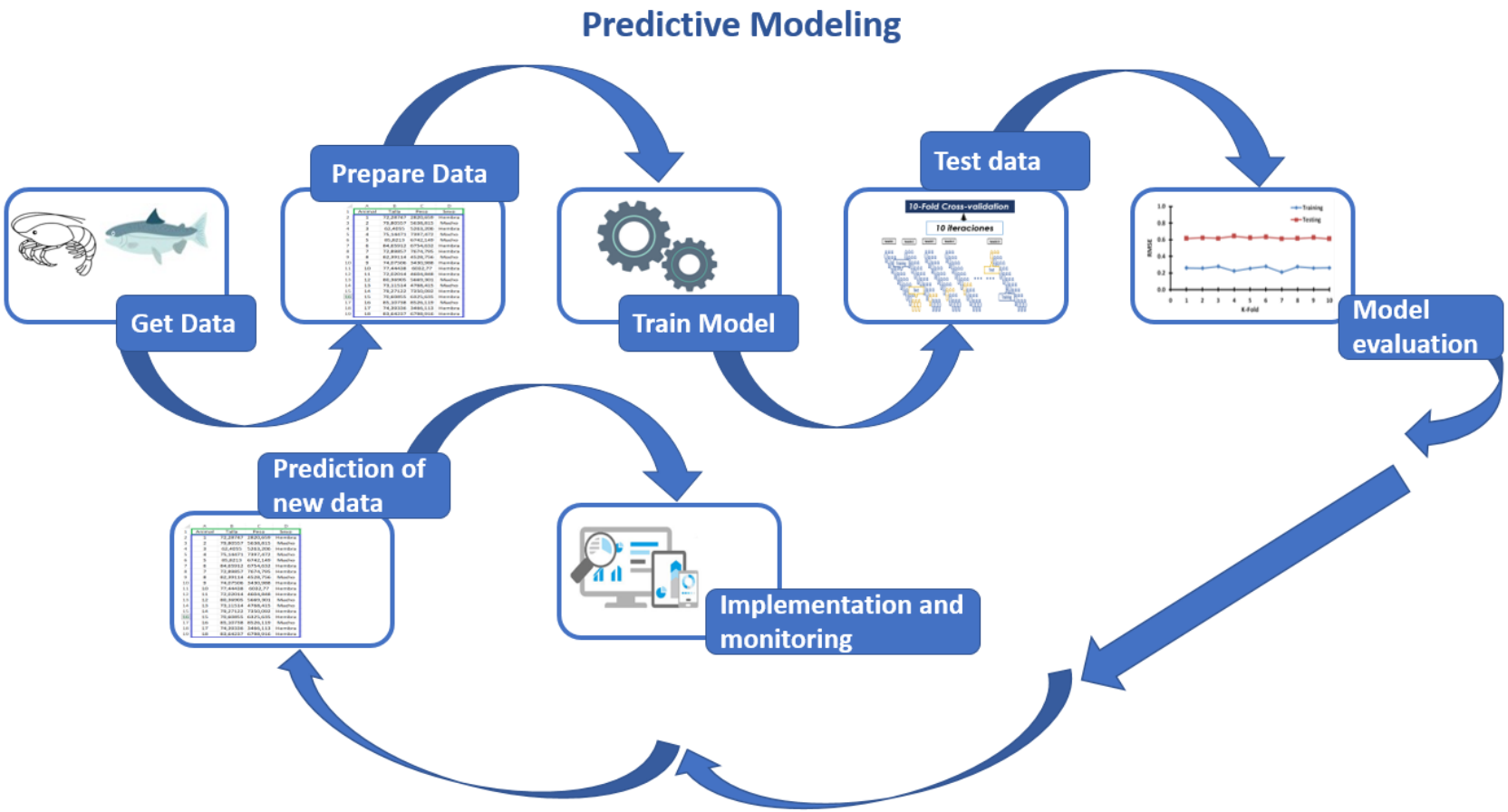


MACHINE LEARNING Y MODELOS PREDICTIVOS

Existen diferentes modelos predictivos, los cuales se clasifican bajo el marco del aprendizaje automático, machine learning en ingés, como algoritmos o metodos supervizados y no supervizados.



ETAPAS DEL MODELAMIENTO PREDICTIVO



PARTICIÓN DEL CONJUNTO DE DATOS

El **training dataset** y el **testing dataset** son dos conjuntos de datos utilizados en el aprendizaje automático y la minería de datos para entrenar y evaluar modelos predictivos.

Conjunto de entrenamiento	**Conjunto de prueba/testeo**
- Conjunto de datos utilizado para entrenar un modelo predictivo. - Contiene variables predictoras y variable a predecir. - El modelo utiliza este conjunto de datos para aprender patrones y relaciones entre las variables predictoras y variable repuesta.	- Conjunto de datos utilizado para evaluar el desempeño y la capacidad de generalización de un modelo entrenado. - No se utiliza para entrenar el modelo. - Los resultados obtenidos en el testing dataset ayudan a estimar el desempeño del modelo en situaciones del mundo real.

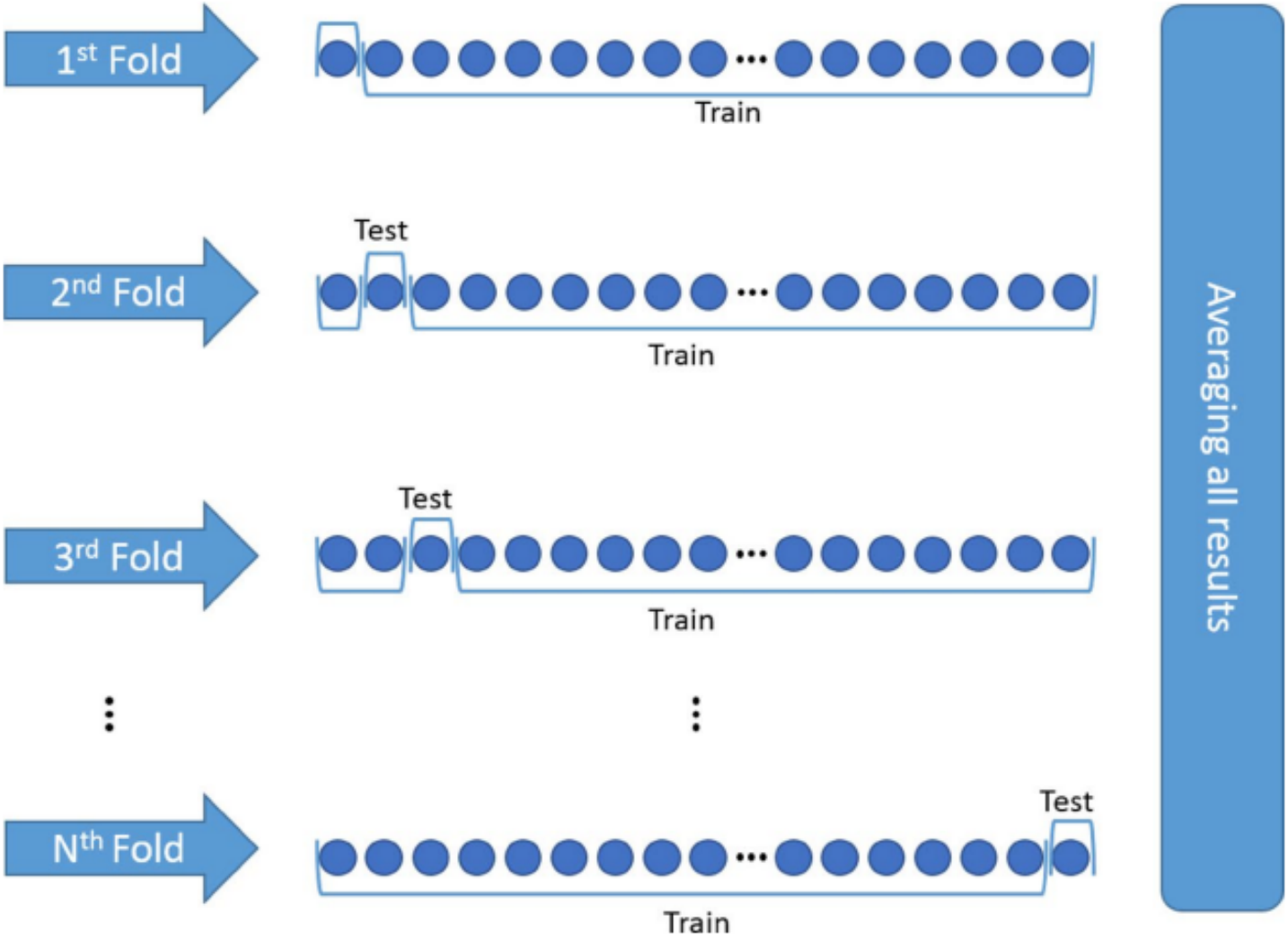
¿PORQUÉ VALIDAR UN MODELO PREDICTIVO?

- Saber qué tan bien funciona: La validación nos dice qué tan preciso es nuestro modelo en la predicción de nuevos datos.
- Evitar errores de sobreajuste: Validar nos ayuda a evitar el error común de ajustar demasiado nuestro modelo a los datos de entrenamiento, lo cual puede llevar a predicciones poco confiables.
- Comparar modelos: Podemos validar varios modelos y comparar su desempeño, para elegir el mejor en función de las métricas de evaluación.
- Optimizar hiperparámetros: La validación nos permite ajustar los hiperparámetros del modelo para mejorar su rendimiento y precisión.

ALGUNOS TIPOS DE VALIDACIÓN

- **Validación cruzada Leave-One-Out:** Proceso de dejar una observación fuera del conjunto de entrenamiento y utilizarla como conjunto de prueba, mientras que el resto de las observaciones se utilizan para entrenar el modelo.
- **Validación cruzada K-fold:** Consiste en dividir el conjunto de datos en K grupos o "pliegues" de aproximadamente igual tamaño. Se selecciona uno de los pliegues como conjunto de prueba y los K-1 pliegues restantes como conjunto de entrenamiento.

VALIDACIÓN CRUZADA LEAVE-ONE-OUT



VALIDACIÓN CRUZADA K-FOLD

Dividir el conjunto de datos en *k*-grupos de igual o similar tamaño

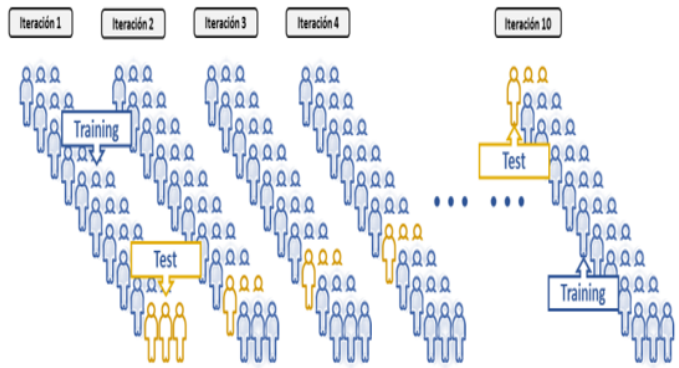
K-Fold Cross-validation

K-1 subconjuntos para entrenar el modelo (Training dataset)

Subconjunto restante para probar/testear el modelo (Testing dataset)

10-Fold Cross-validation

10 iteraciones



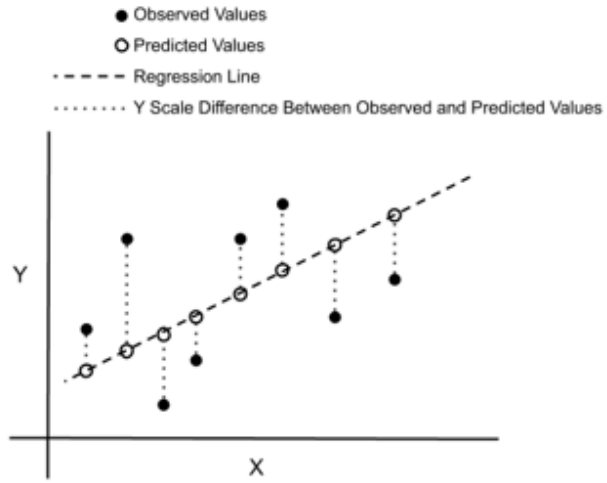
¿CÓMO MEDIR EL DESEMPEÑO DEL MODELO?

Métricas de evaluación

Rasgos cuantitativos continuos

- GF3
- COLOR BELLY

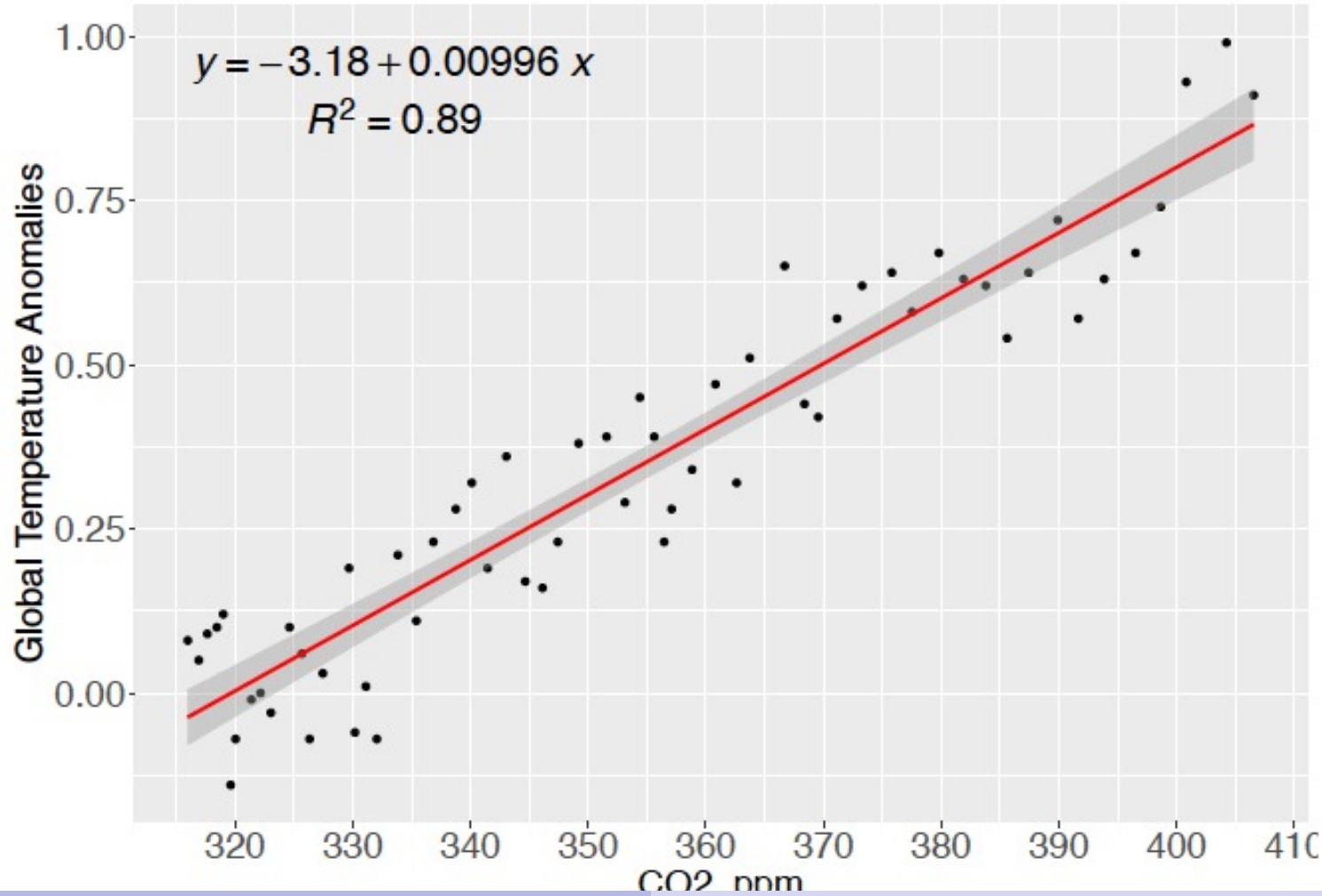
$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$



Mejor es el modelo si tiene el valor más bajo

ESTUDIO DE CASO: REGRESIÓN LINEAL SIMPLE

Queremos predecir las anomalías de la temperatura global en función del CO2.



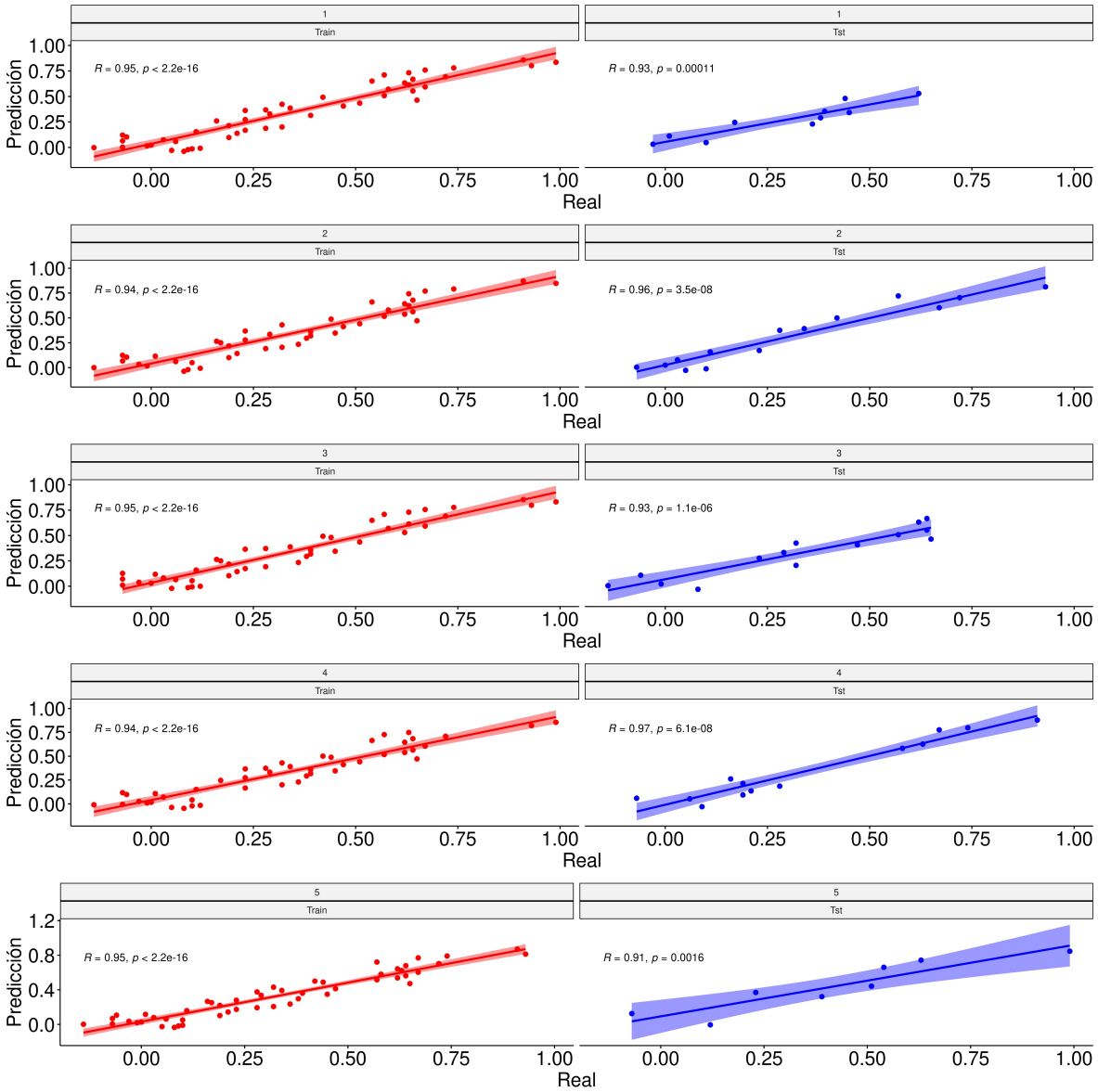
VALIDACIÓN CRUZADA K-FOLD (K=5)

	True	Predicted	Fold	Data
1	0.08	-0.04	1	Train
2	0.05	-0.03	1	Train
3	0.09	-0.02	1	Train
4	0.10	-0.01	1	Train
5	0.12	-0.01	1	Train
110	-0.03	0.03	1	Tst
210	0.10	0.05	1	Tst
310	0.01	0.11	1	Tst
410	0.36	0.23	1	Tst
51	0.17	0.25	1	Tst

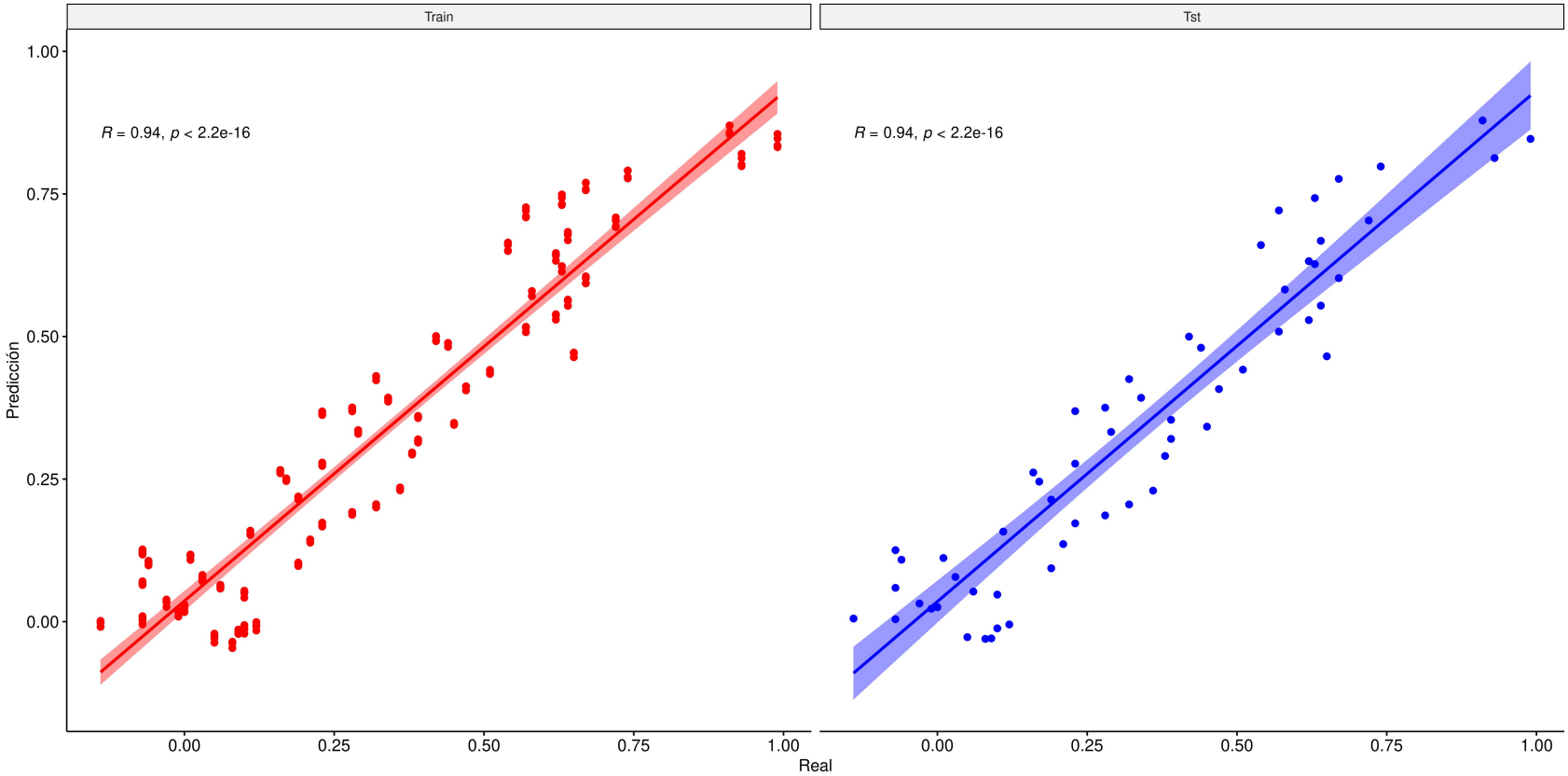
VALIDACIÓN CRUZADA K-FOLD (K=5)

Fold	Data	R	R2	MSE	RMSE	n
1	Train	0.95	0.90	0.01	0.09	49
1	Tst	0.93	0.86	0.01	0.08	10
2	Train	0.94	0.88	0.01	0.10	45
2	Tst	0.96	0.93	0.01	0.08	14
3	Train	0.95	0.90	0.01	0.09	45
3	Tst	0.93	0.87	0.01	0.10	14
4	Train	0.94	0.88	0.01	0.10	46
4	Tst	0.97	0.94	0.01	0.08	13
5	Train	0.95	0.90	0.01	0.09	51
5	Tst	0.91	0.83	0.02	0.13	8

VALIDACIÓN 5-FOLD

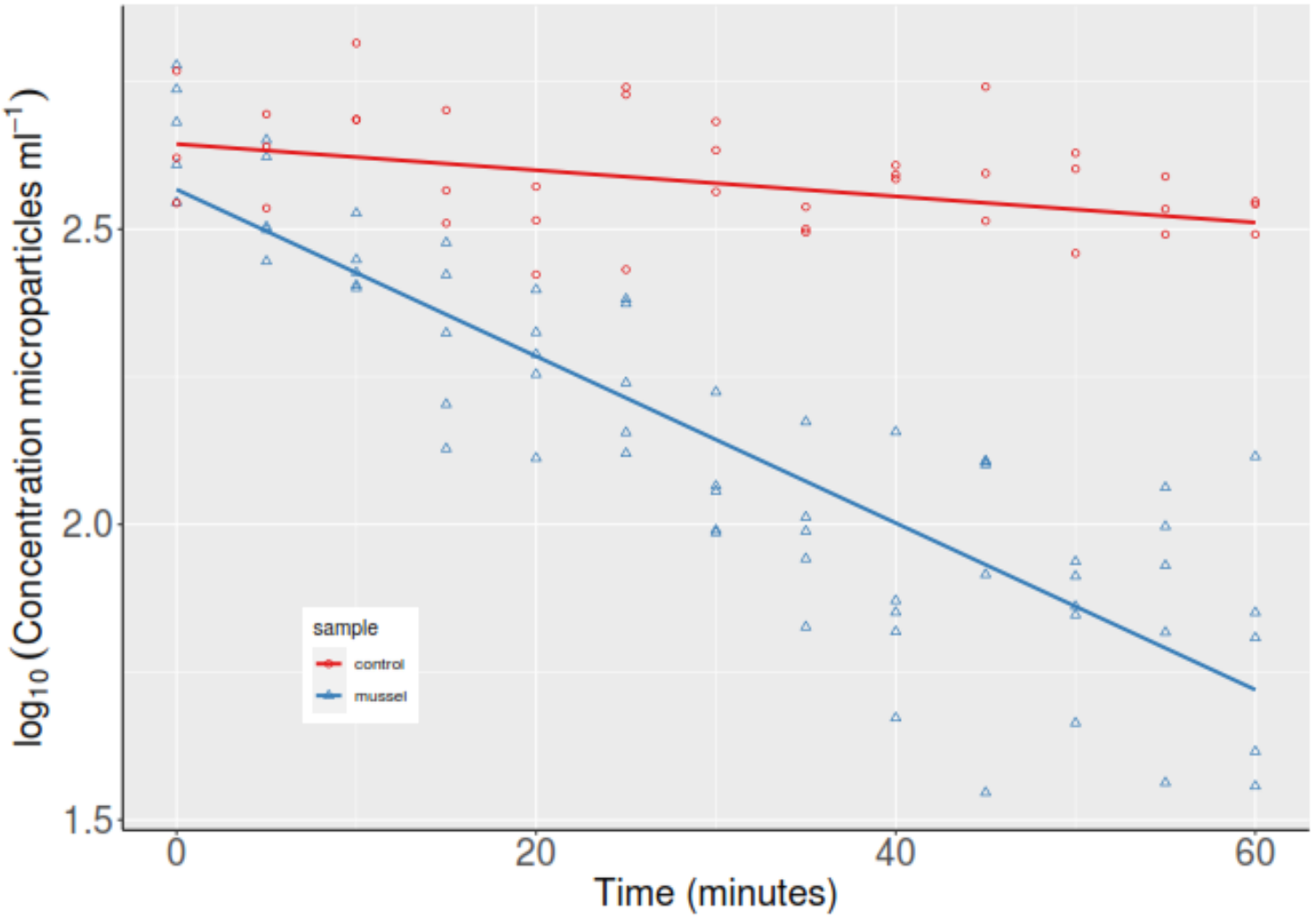


VALIDACIÓN 5-FOLD




```
## # A tibble: 2 x 3
##   Data  MSE_PROMEDIO RMSE_PROMEDIO
##   <chr>        <dbl>          <dbl>
## 1 Train      0.00855         0.0924
## 2 Tst        0.00925         0.0944
```

ESTUDIO DE CASO: REGRESIÓN LINEAL MÚLTIPLE

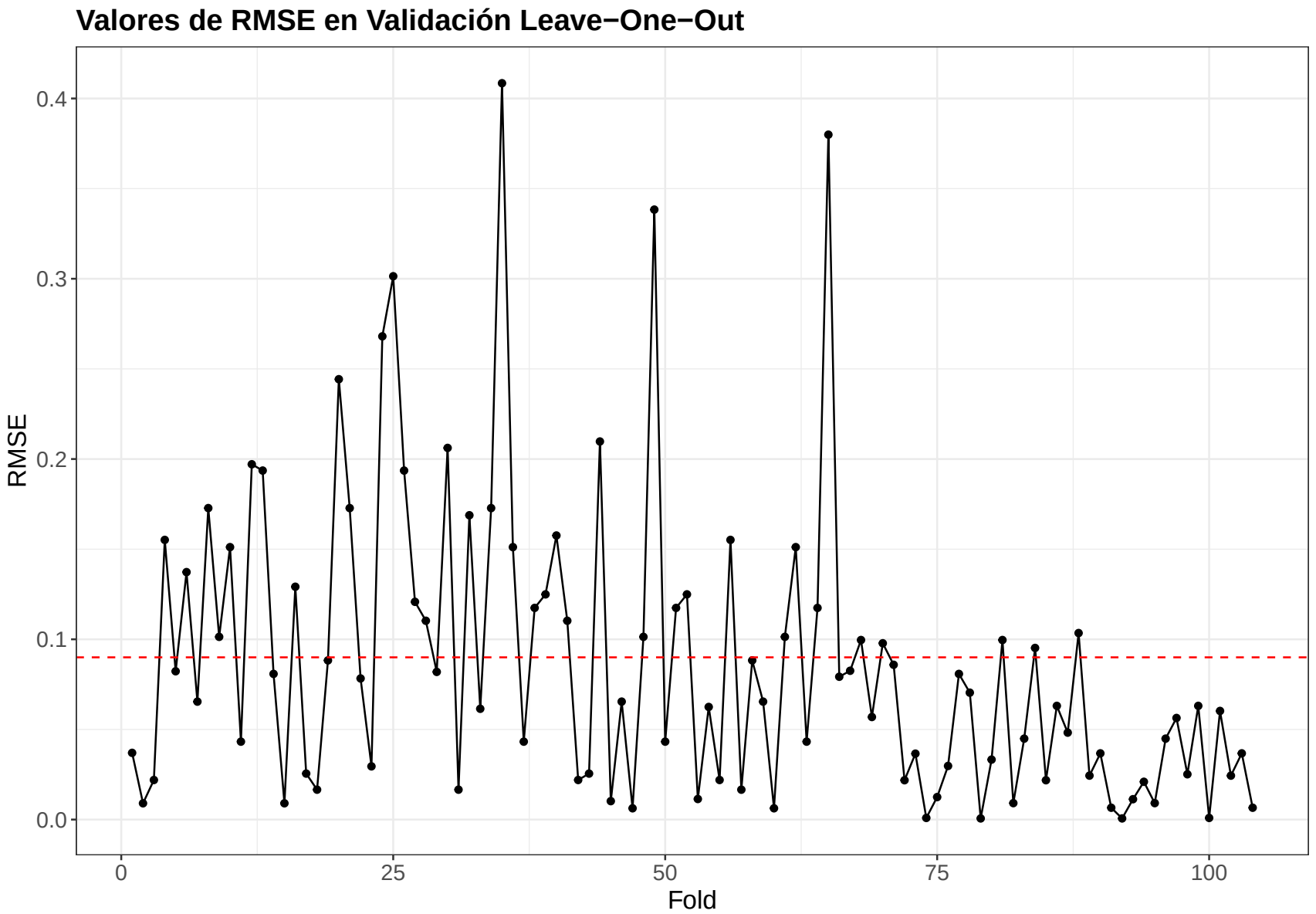


VALIDACIÓN LEAVE-ONE-OUT

	Fold	RMSE
1	1	0.04
2	2	0.01
3	3	0.02
4	4	0.16
5	5	0.08
100	100	0.00
101	101	0.06
102	102	0.02
103	103	0.04
104	104	0.01

```
## [1] 0.09
```

VALIDACIÓN LEAVE-ONE-OUT



PRÁCTICA ANÁLISIS DE DATOS

- El trabajo práctico se realiza en Rstudio.cloud.
- Guía de clase 16 en formato html.

RESUMEN DE LA CLASE

- Modelamiento predictivo Supervisado y No supervisado.
- Partición del conjunto de datos: entrenamiento y testeo.
- Tipos de validación (leave-one-out, K-fold).
- Métricas de evaluación (MSE, RMSE).