Análisis de modelos de machine learning. Regresión lineal y regresión logística.

Pablo Vargas Ibarra

Trabajo de fin de grado

Grado en Ingeniería Matemática

Tutor: Javier Yañez

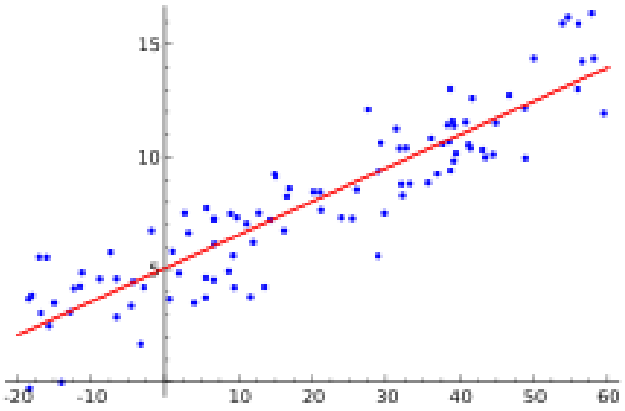
Motivación

- ***Modelización***
- *Probabilidad / Estadística: F.coste y regularización.*
- *Investigación Operativa: Descenso del gradiente.*
- *Machine Learning: Uso de la librería de Python scikit-learn.*
- *Programación: Repositorio en GitHub.*

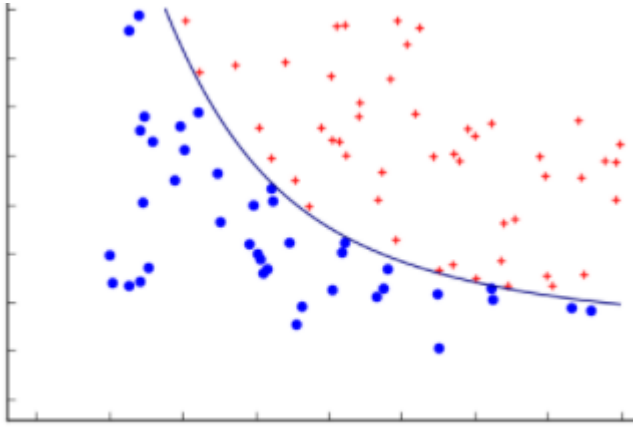
Aprendizaje supervisado

$$Y = f(X)$$

- Regresión



- Clasificación

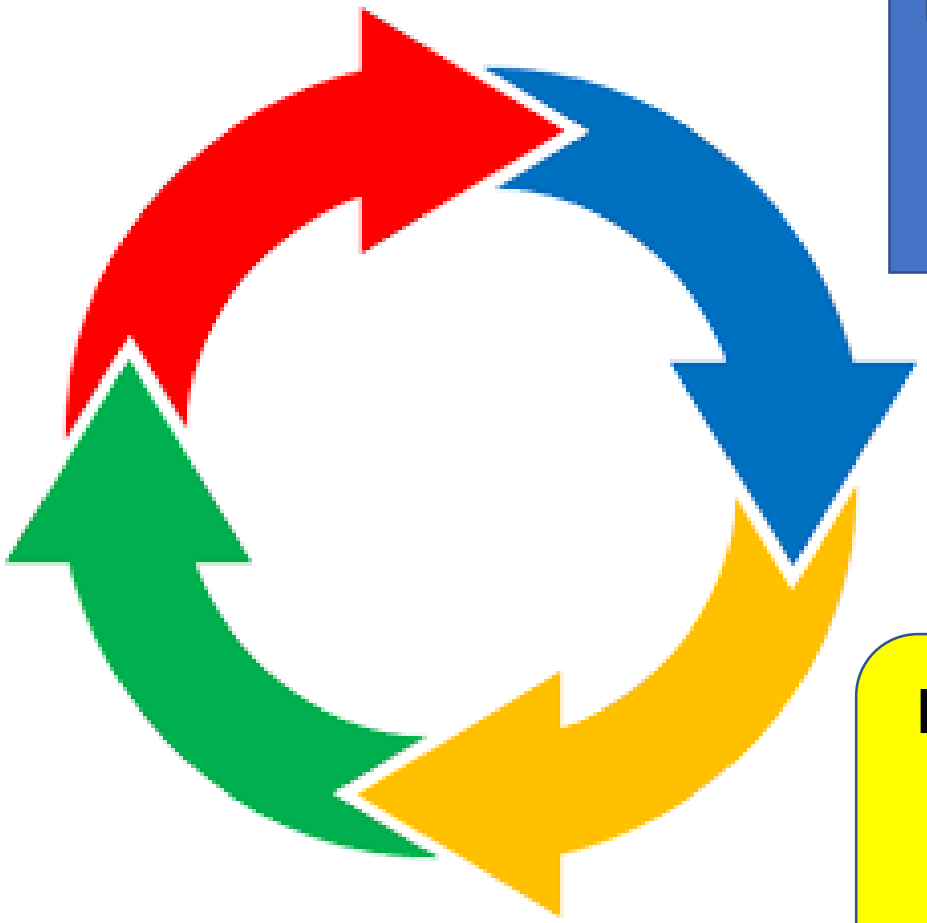


Evaluación

Métricas de éxito

Modelización

$h_\theta(X)$



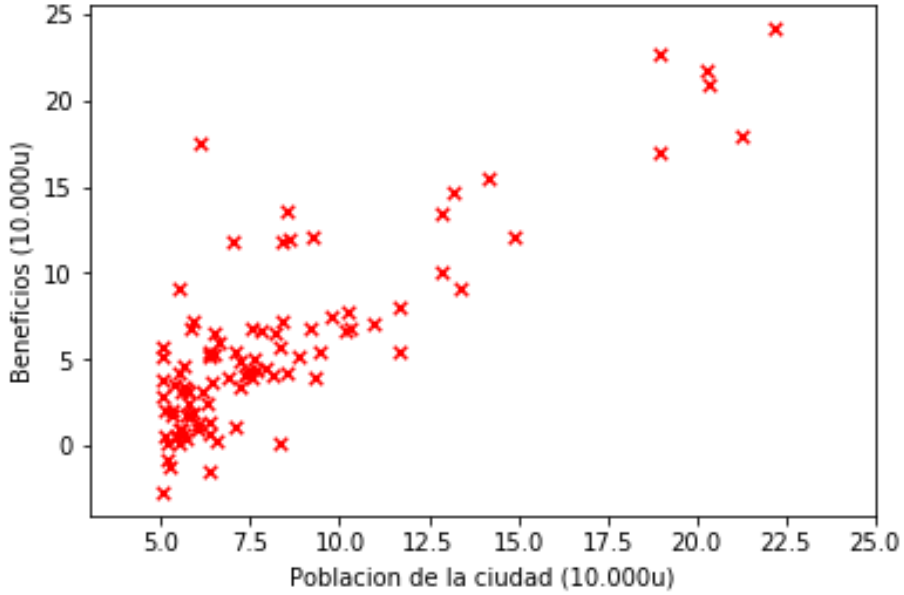
Optimización

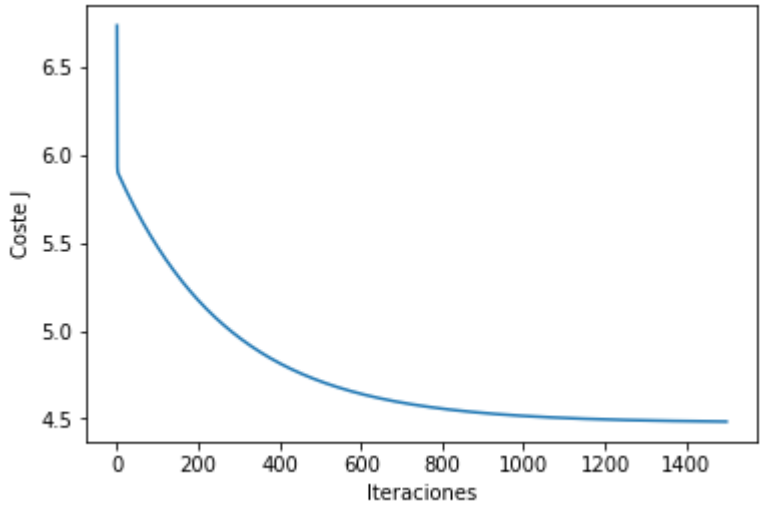
Gradient descent

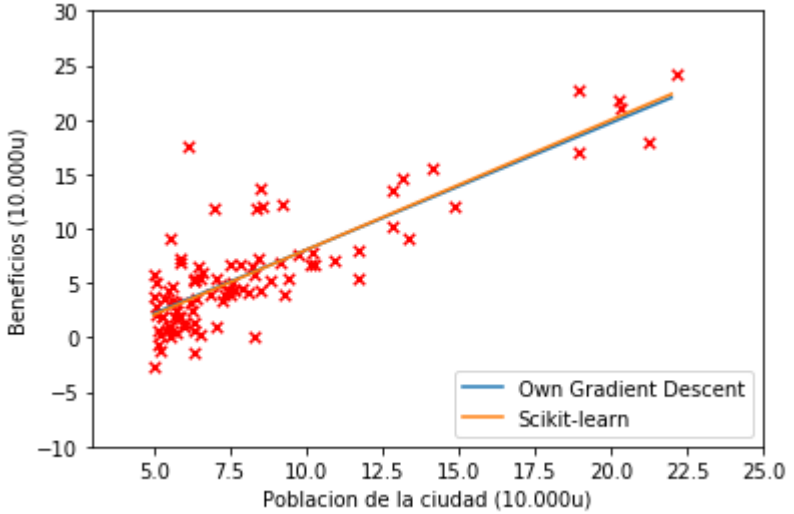
Función de coste

MSE
LogLoss
Regularización

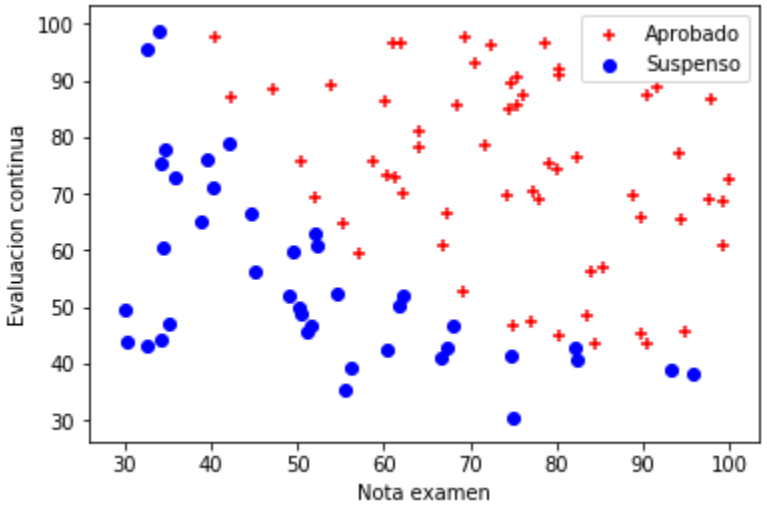
Regresión Lineal



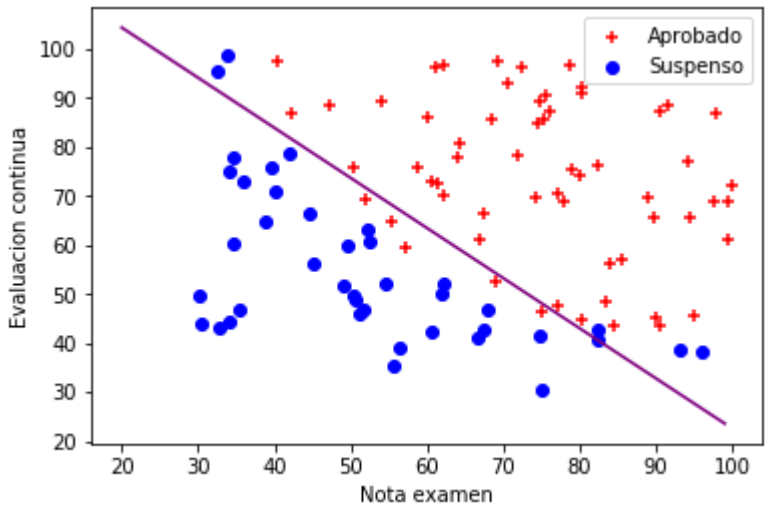




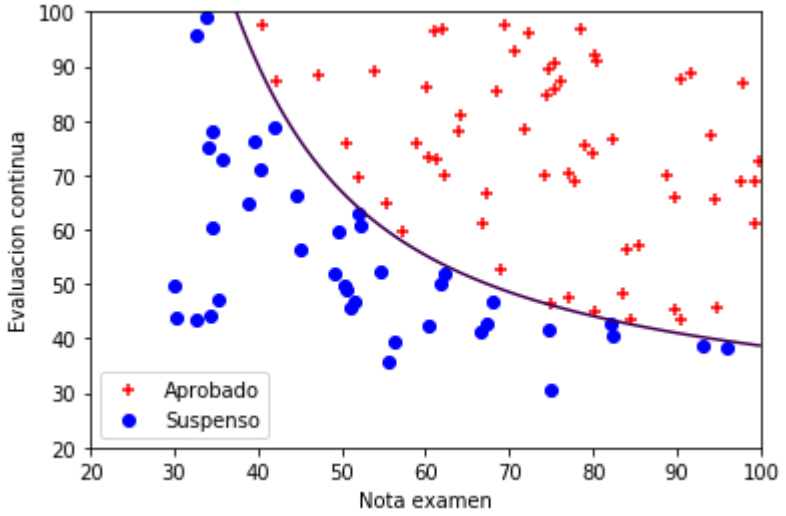
Regresión Logística



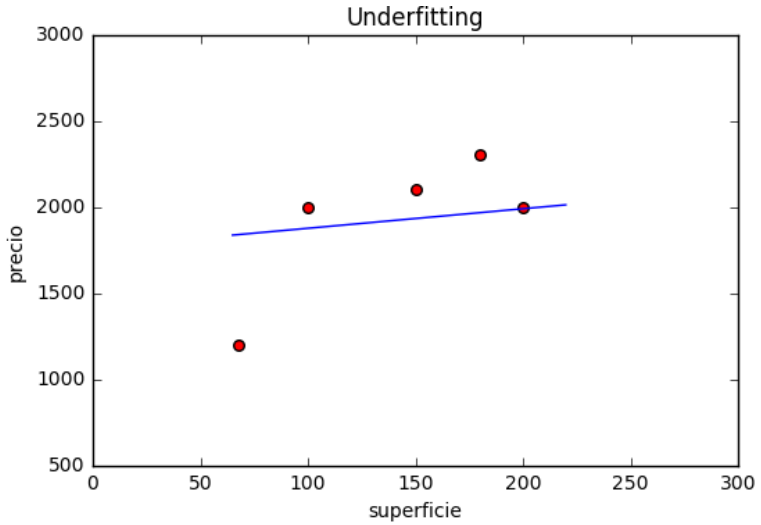
Clasificador lineal

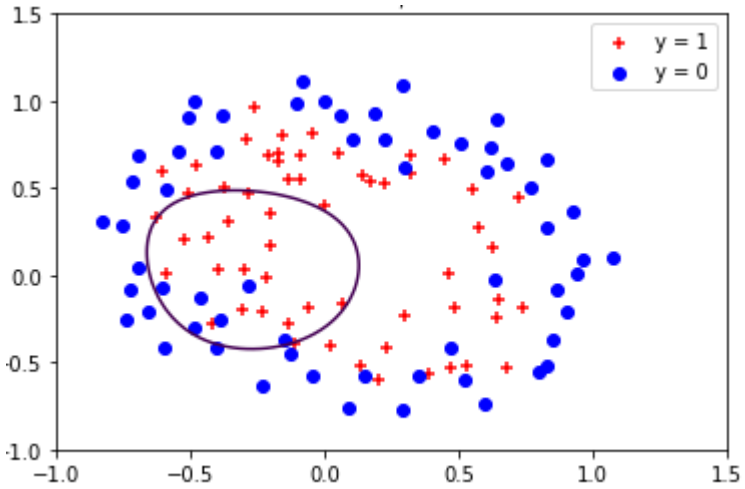


Clasificador no lineal

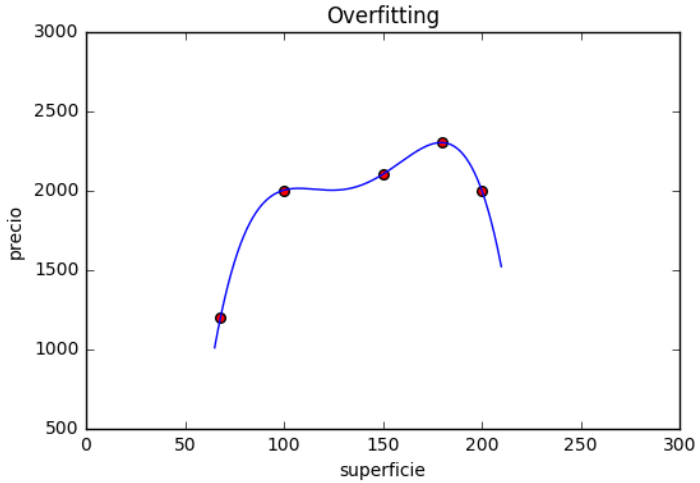


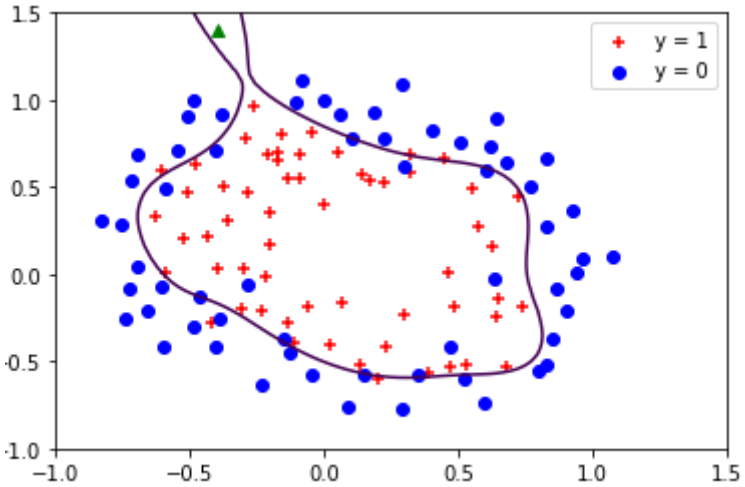
Alto Sesgo





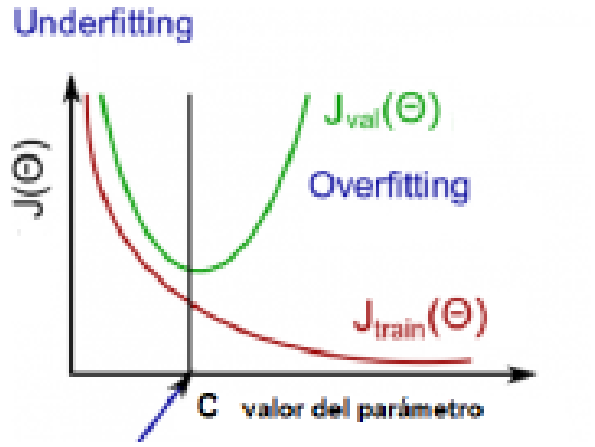
Alta Varianza





Regularización

$$R(\theta) = \frac{1}{2C}\sum_{j=1}^{n}\theta_j^2 \; ; \; C > 0$$



Underfitting

(-) Regularización

- ⬆ C

(+) $h_\theta(X)$

- ⬆ Degree
- ⬆ # Variables

Overfitting

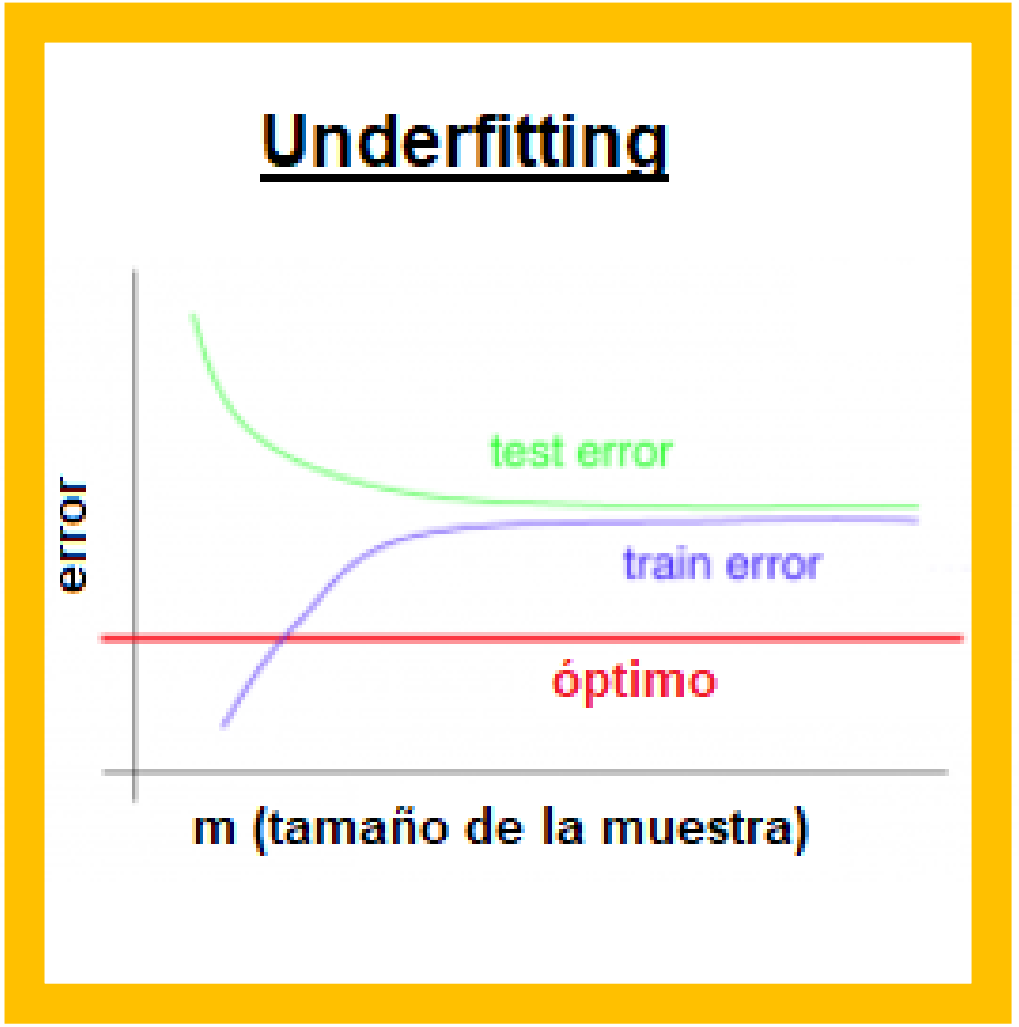
(+) Regularización

- ⬇ C

(-) $h_\theta(X)$

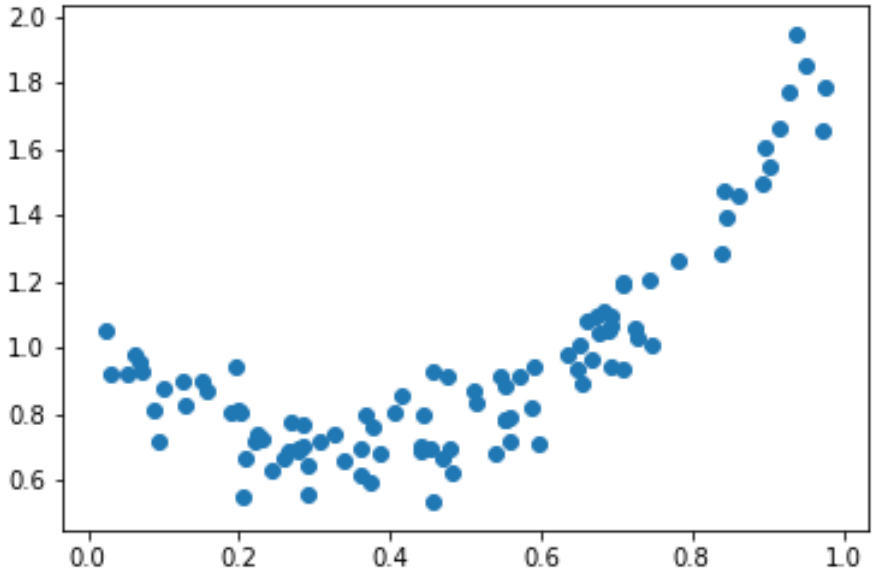
- ⬇ Degree
- ⬇ # Variables

Curvas de aprendizaje ¿# Datos?



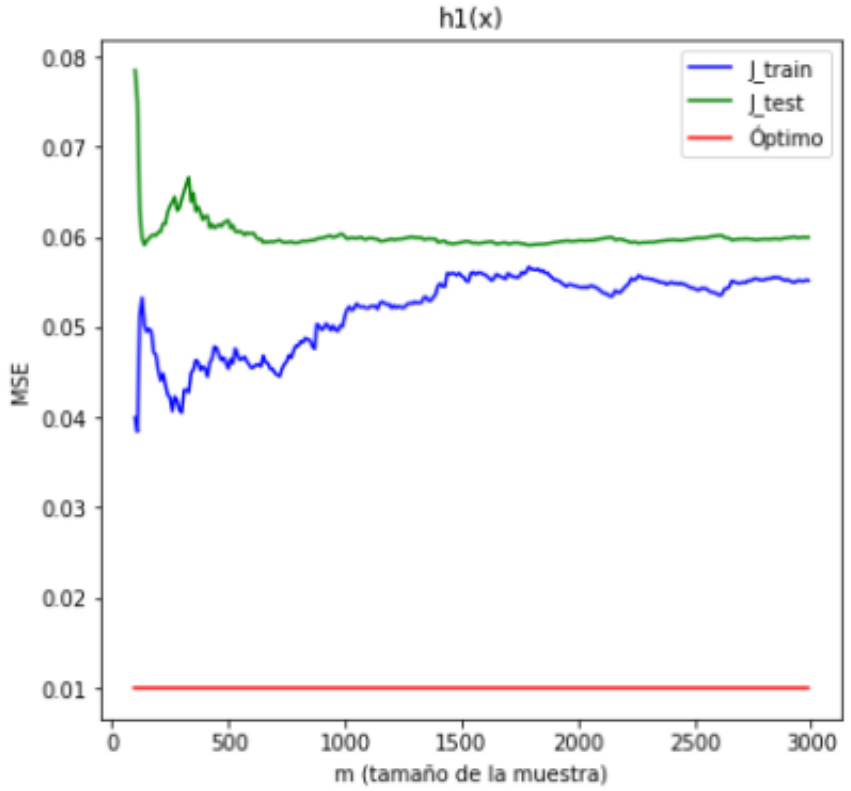
Simulation

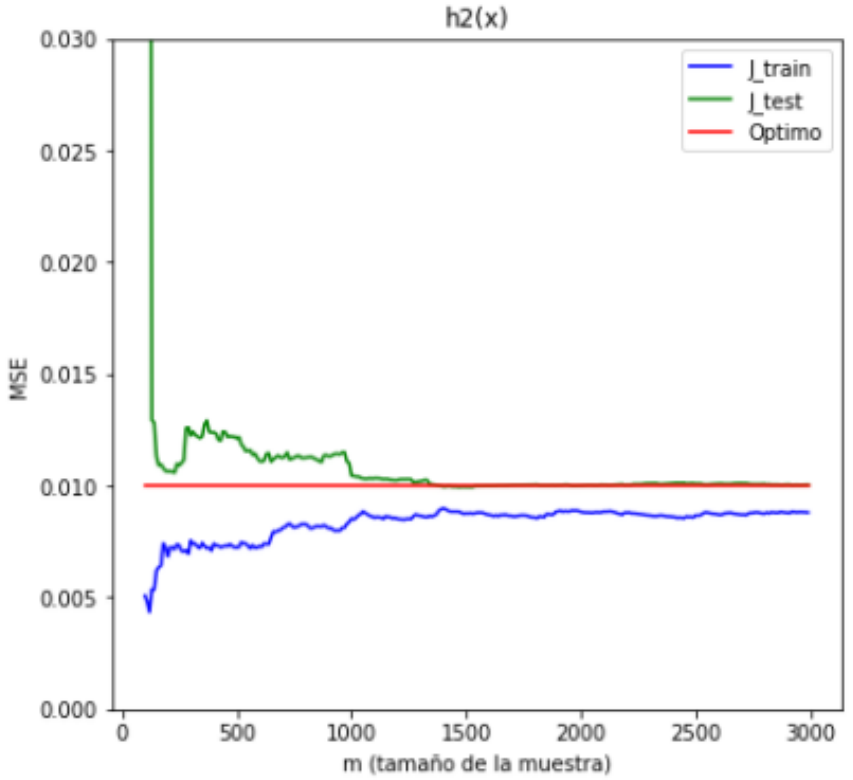
$$y = 1 - 2x + 3x^2 + \epsilon\,;\ \epsilon \sim \mathcal{N}(0,\ \sigma^2 = 0.01) \quad x \in (0, 1)$$



$$h_1(x) = \theta_0 + \theta_1 x$$

$$h_2(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4 + \theta_4 x^5 + \theta_6 x^6 + \theta_7 x^7 + \theta_8 x^8$$





Metodología Propuesta

1. Entrenamiento/Validación/Test
2. Estandarizar (según muestra de entrenamiento)
3. Elegir distintos C y h.
4. Calcular los parámetros para cada C y h.
5. Elegir el major modelo en base a una métrica de éxito evaluada en el conjunto de validación.
6. Evaluar la capacidad de generalización en el conjunto de test.
7. Entrenar el modelo final con todos los datos. (h* y C*)

Breast Cancer Prediction

	C	Train Accuracy	Validation Accuracy	Test Accuracy	Test Recall	Test Precision
0	0.00001	0.935484	0.912281	0.947368	0.914894	0.955556
1	0.00010	0.938416	0.912281	0.947368	0.914894	0.955556
2	0.00100	0.944282	0.938596	0.947368	0.914894	0.955556
3	0.01000	0.976540	0.964912	0.938596	0.893617	0.954545
4	0.20000	0.988270	0.982456	0.947368	0.893617	0.976744