Maestría en Explotación de Datos y Descubrimiento del Conocimiento

Unidad I, Clase 3

Aprendizaje Automático

El problema del sobreajuste (overfitting)

Laura de la Fuente, Hernán Bocaccio
Ayudantes: Gastón Bujía, Diego Onna y Sofía Morena del Pozo

Dirección de e-mail de la materia:
datawillconfess@gmail.com

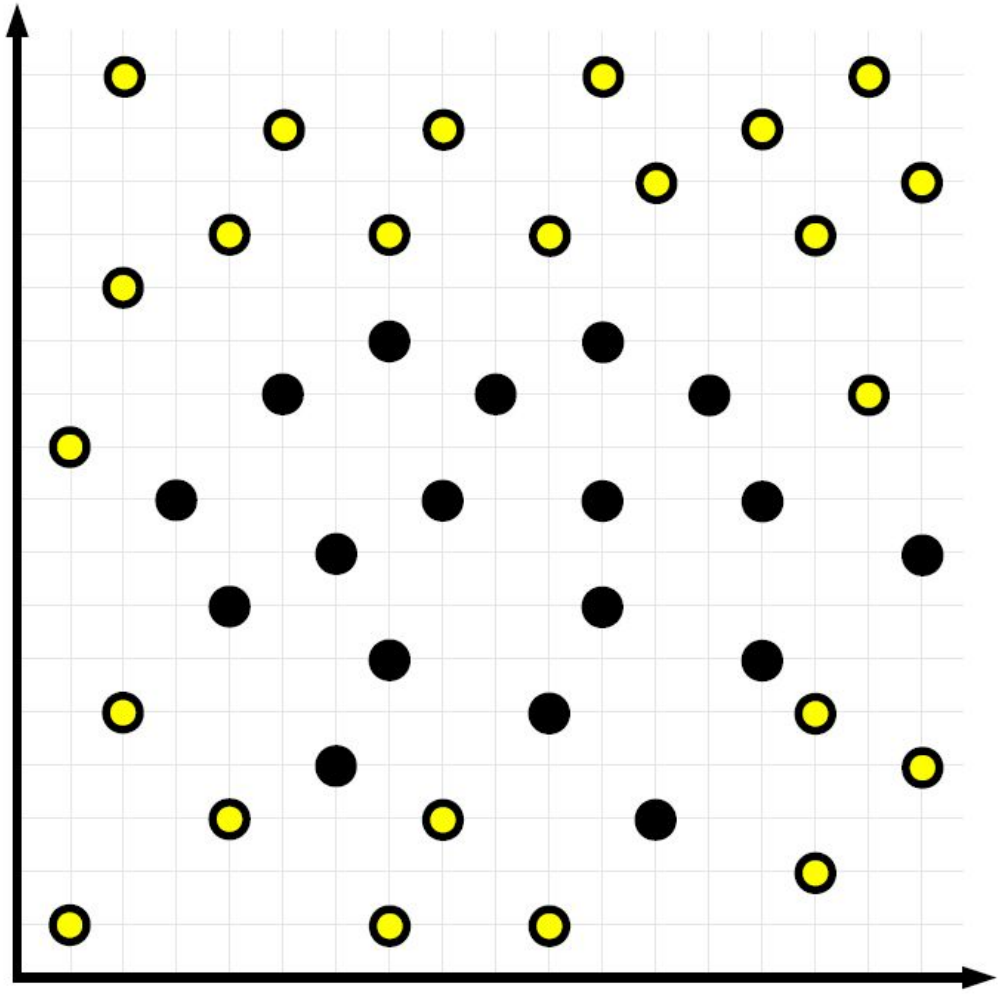
Itinerario de la clase de hoy

- Complejidad de modelos y generalización
- Separación de datos
- Sobre y sub ajuste
- Trade off sesgo varianza
- Parámetros e hiperparámetros
- Regularización
- Flujo general del trabajo en ciencia de datos
- Interpretabilidad y selección de modelos

Itinerario de la clase de hoy

- Complejidad de modelos y generalización
- Separación de datos
- Sobre y sub ajuste
- Trade off sesgo varianza
- Parámetros e hiperparámetros
- Regularización
- Flujo general del trabajo en ciencia de datos
- Interpretabilidad y selección de modelos

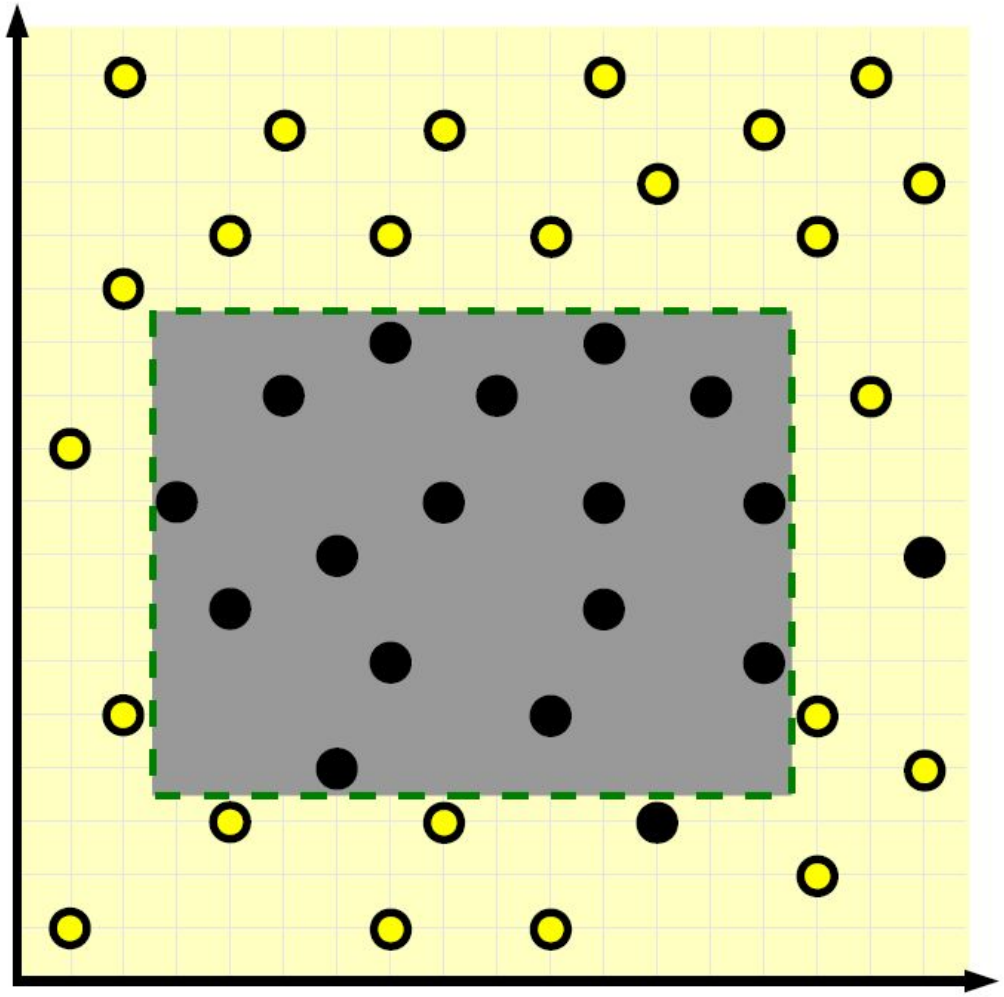
Complejidad de un modelo



Queremos aprender un concepto (función objetivo) desconocido.

Tenemos un conjunto de datos.

Complejidad de un modelo



Elegimos un modelo (hipótesis) rectangular.

Cada rectángulo tiene:

Base (b)

Altura (h)

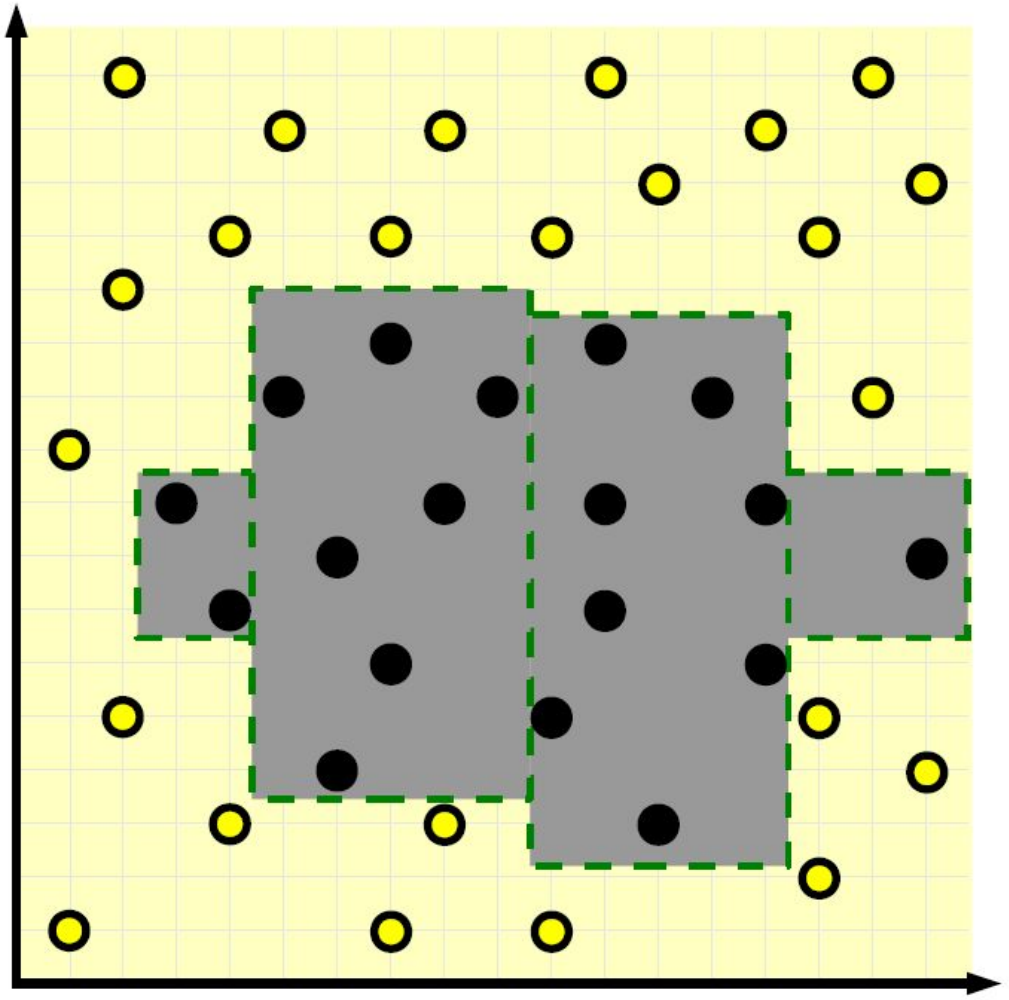
Posición (x, y)

Color interior (azul/rojo)

Espacio de hipótesis

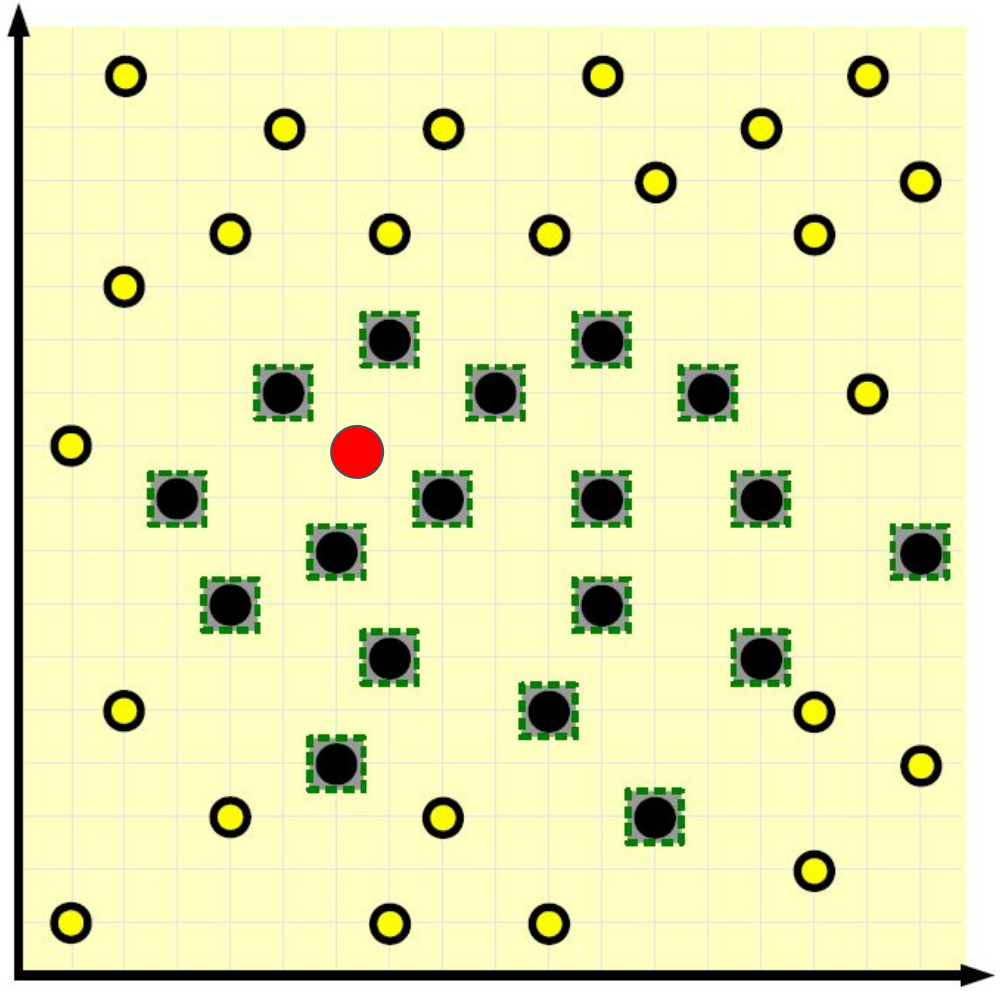
H = < b, h, x, y, colorint >

Complejidad de un modelo



Ahora, elegimos otro modelo (hipótesis) formado por varios rectángulos.

Complejidad de un modelo



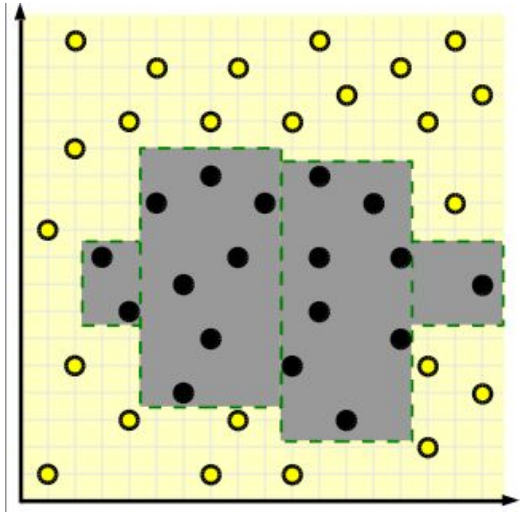
Elegimos otro con N rectángulos = N datos.

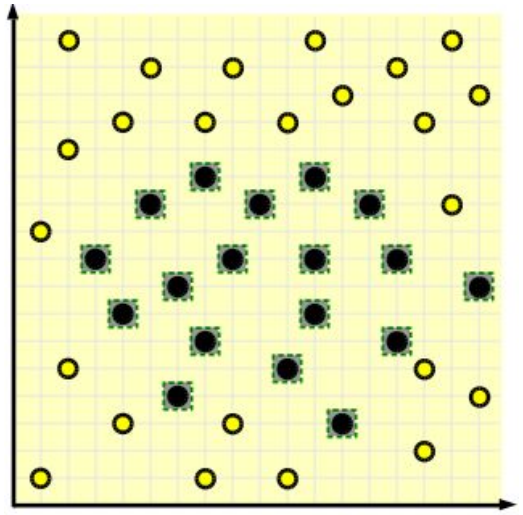
Para un dato nuevo, los distintos modelos darían distintas predicciones.

El punto de Schrodinger

Si el modelo es muy complejo, corremos el riesgo de sobreajuste (overfitting).

Complejidad de un modelo





Un modelo podría simplemente memorizar los casos y tener accuracy de 100% en los datos de entrenamiento.

Medir performance sobre los datos de entrenamiento tiende a sobre-estimar los resultados.

Generaliza a datos nuevos??

<u>Navaja de Ockham / Principio de parsimonia</u>

En igualdad de condiciones, elegir la explicación más simple.

Ante dos hipótesis que ajustan igualmente bien a los datos, la más simple es la más probablemente cierta.

Además, suele ser más fácil de interpretar y es esperable que **generalice** mejor.

Itinerario de la clase de hoy

- Complejidad de modelos y generalización
- Separación de datos
- Sobre y sub ajuste
- Trade off sesgo varianza
- Parámetros e hiperparámetros
- Regularización
- Flujo general del trabajo en ciencia de datos
- Interpretabilidad y selección de modelos

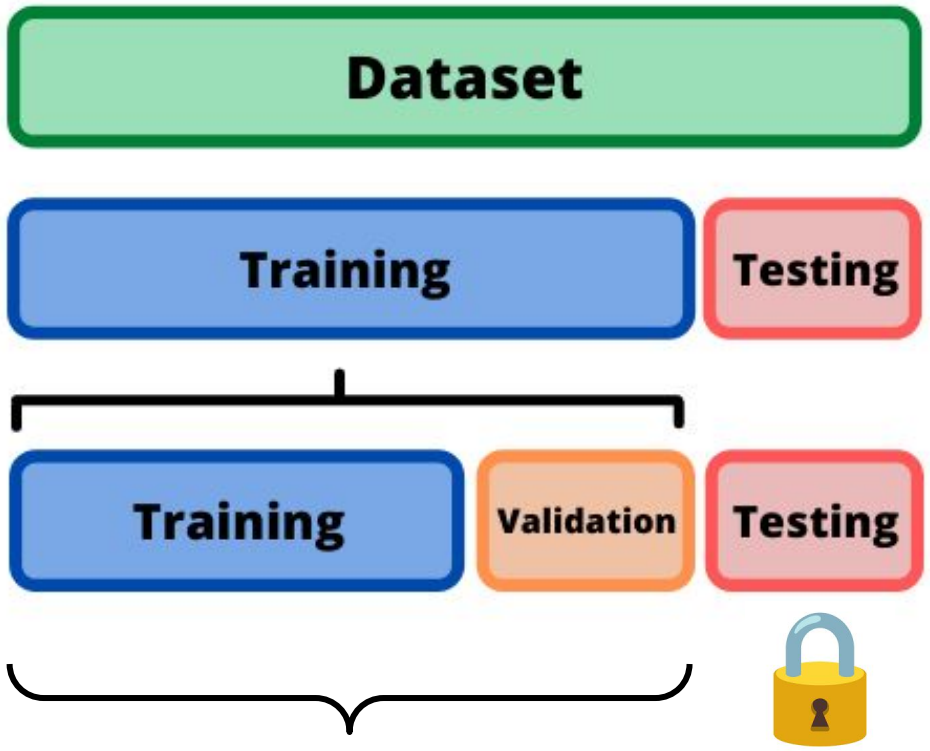
¿Dónde evaluamos un modelo?

Queremos **GENERALIZAR** y no MEMORIZAR

<u>Separación de datos</u>

Partir en Entrenamiento, Validación, y Testeo

Los datos se deben separar al azar (o pseudo), para evitar cualquier orden o estructura subyacente en los datos.



Problema de generalización

Para que un modelo tenga un buen desempeño debe funcionar bien en:

- Entrenamiento
- Validación
- **y Testeo!**

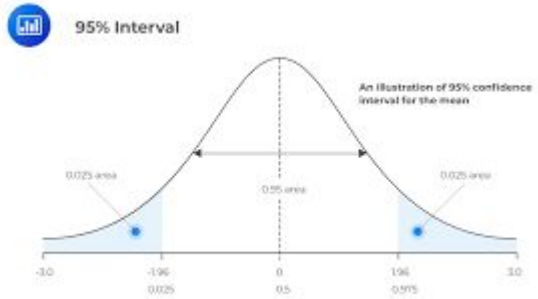
Partimos los datos para tener una mejor estimación del desempeño del modelo.

Es mi modelo un SUPER modelo?

No nos interesa un SUPER modelos, queremos estimar que tan bueno podría llegar a ser en términos estadísticos.

Para eso, podemos querer evaluar qué tan universales/robustos son mis modelos frente a particiones aleatorias de los datos.

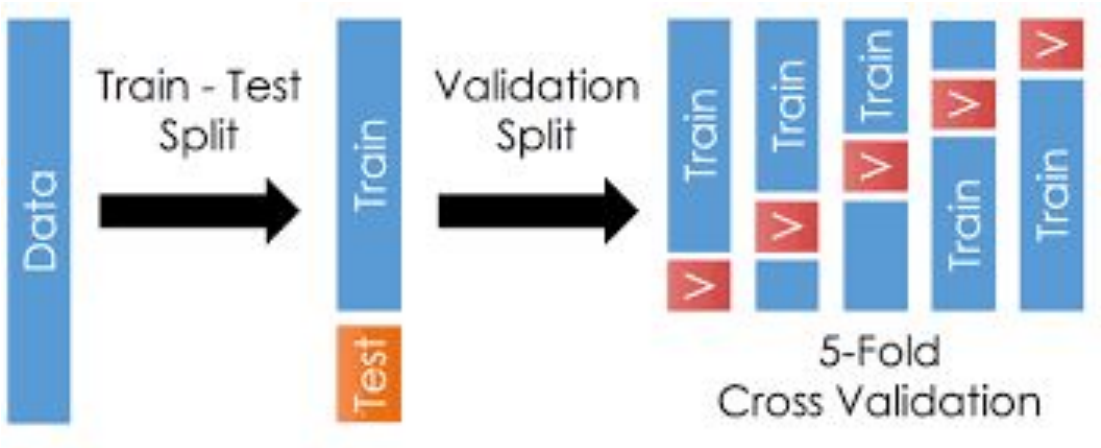
Técnicas de validación un poco más complejas.



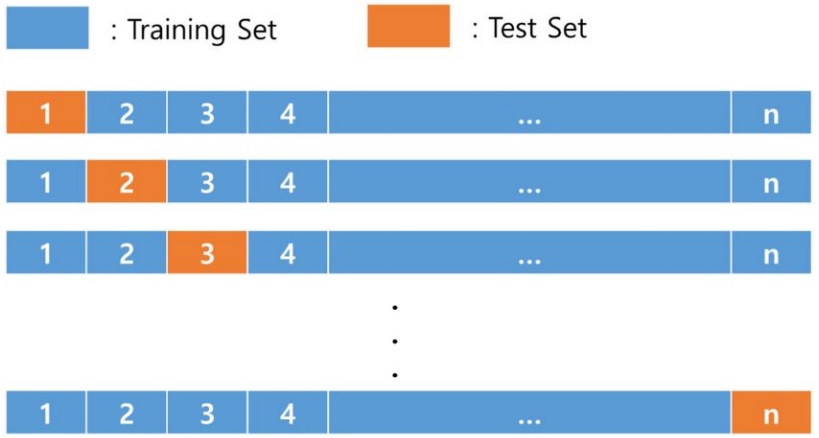
Validación cruzada

La validación cruzada o cross validation me permite tener una mejor estimación de un modelo evaluando estadísticamente su desempeño y garantizando que es independiente de la partición de datos.

Validación cruzada en k-grupos



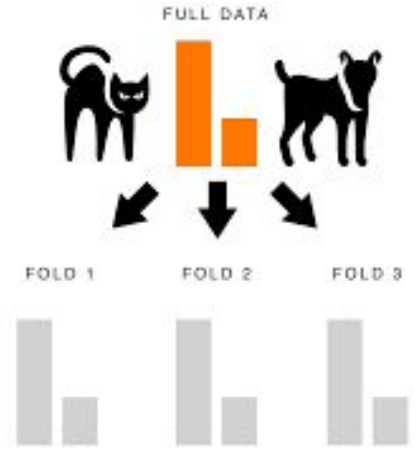
Leave one out

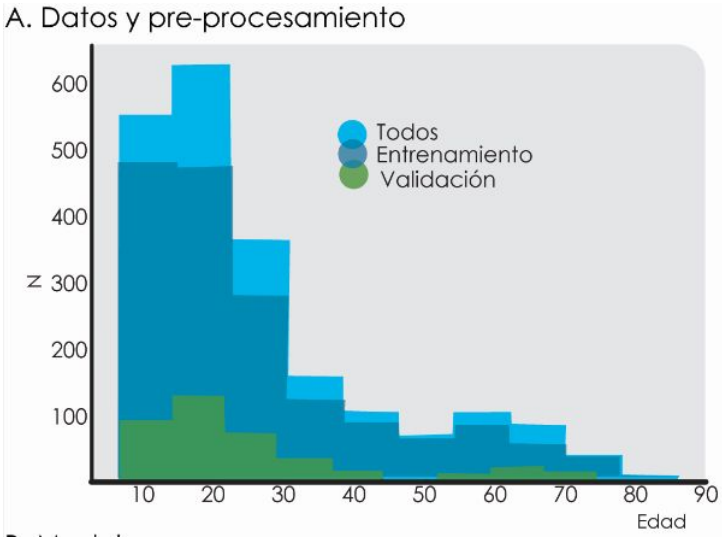


Los modelos igual comparten muchos datos de entrenamiento (cómo depende de la cantidad de k particiones?) por lo que no llegan a ser independientes del todo (diagnóstico no tan bueno de la varianza). Necesito también testeo independiente.

Datos desbalanceados

K-grupos estratificados





Reporte de desempeño

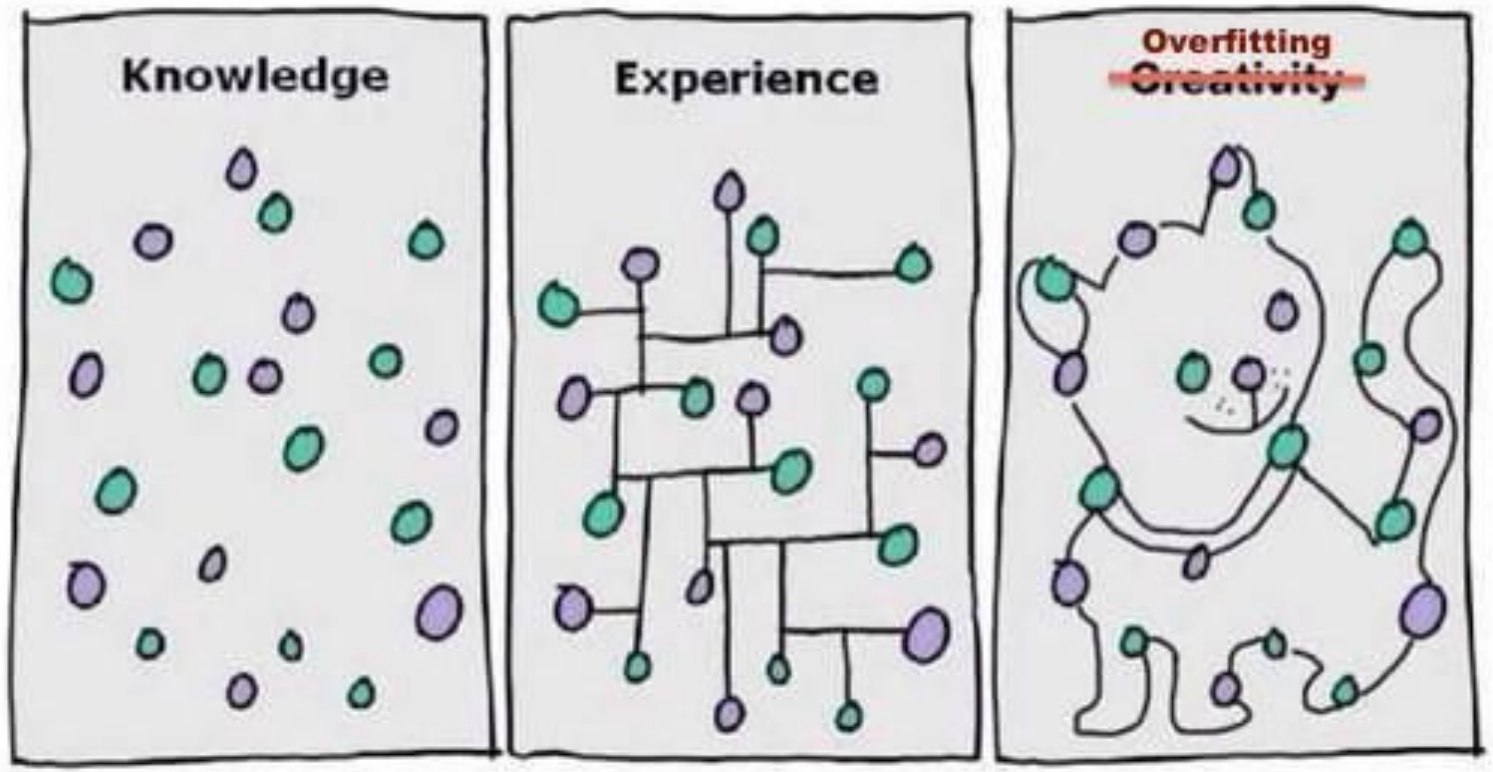
- Suele reportarse la medida de desempeño en los datos de testeo o validación
- En algunos casos luego se re-entrena en toda la base

Itinerario de la clase de hoy

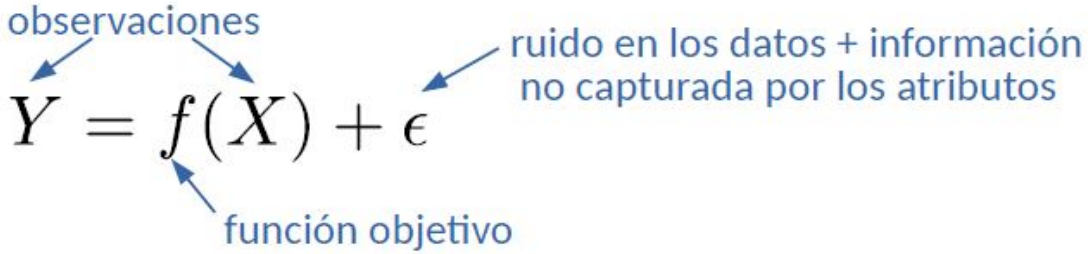
- Complejidad de modelos y generalización
- Separación de datos
- Sobre y sub ajuste
- Trade off sesgo varianza
- Parámetros e hiperparámetros
- Regularización
- Flujo general del trabajo en ciencia de datos
- Interpretabilidad y selección de modelos

Escenarios posibles

Recurrimos a nuestra medida de performance P y vemos cómo se comporta en diferentes sets. Así podremos diagnosticar si el modelo está sobreajustando.



Trade-off Sesgo / Varianza



Depende de la generación de los datos, irreducible, independiente del modelo

Para el caso de regresión

$$y = \omega_0 + \omega_1 \cdot x$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - t_i)^2$$

Experiencia = Muestras $\{x_i\} \quad i = 1 \dots N$

Valor de expectación $\mathbb{E}_f[x] \longrightarrow \bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$

Varianza $\mathrm{var}_f[x] \longrightarrow \bar{S} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{X})^2$

Error esperado del modelo sobre distintos sets de datos $\mathcal{D} = \{\mathcal{D}_1, \dots, \mathcal{D}_N\}$

$$\begin{aligned} \mathbb{E}_{\mathcal{D}}(MSE) &= \mathbb{E}_{\mathcal{D}}\{y - t\}^2 \\ &= \mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - h(\mathbf{x})\}^2\right] + \overbrace{\mathbb{E}_{\mathcal{D}}\left[\{h(\mathbf{x}) - t\}^2\right]}^{\sigma^2} \end{aligned}$$

h(x) = E[t|x]

Trade-off Sesgo / Varianza

Error esperado del modelo sobre distintos sets de datos $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_N\}$

$$\mathbb{E}_{\mathcal{D}}(MSE) = \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 + \mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right] + \sigma^2$$

h(x) = E[t|x]

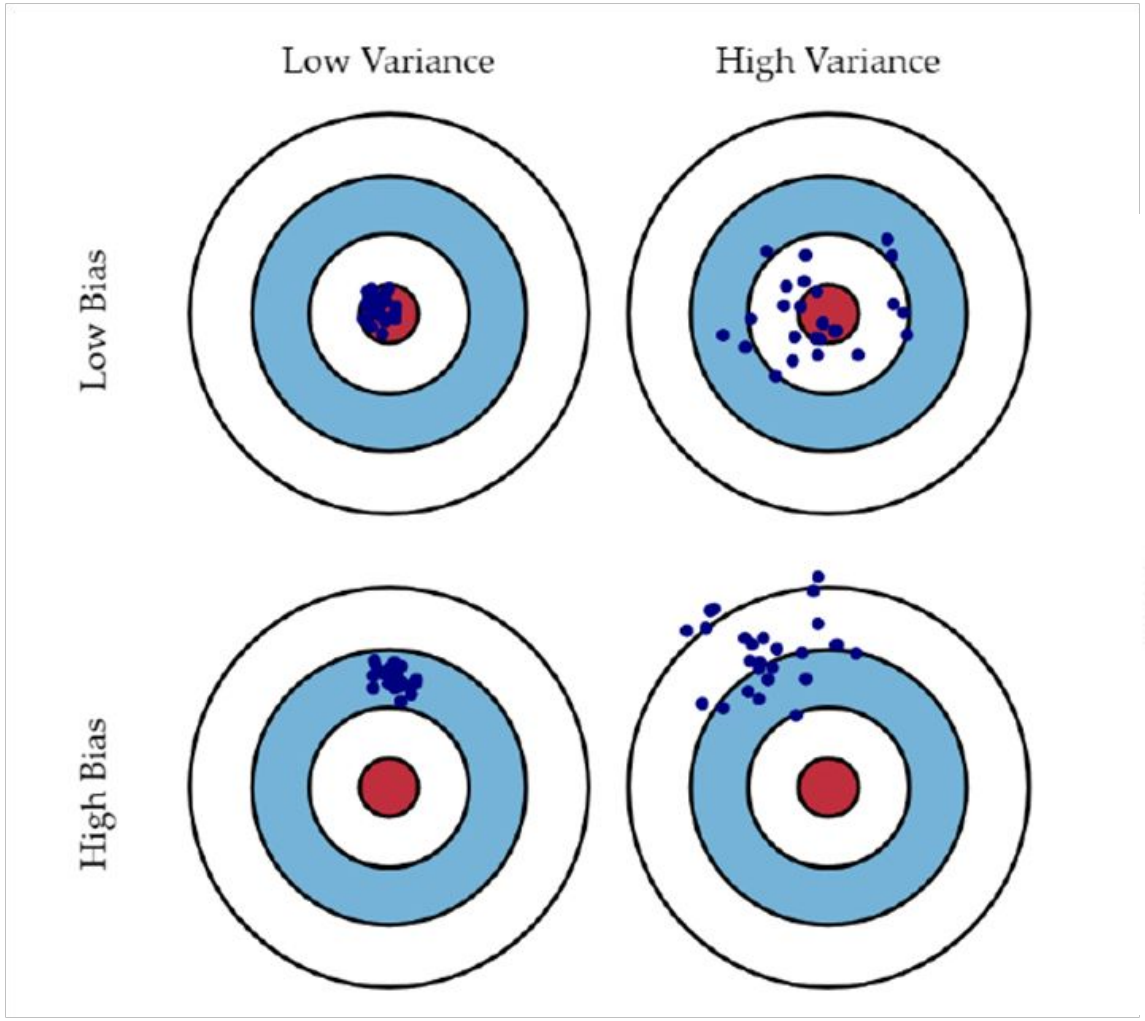
$$\boxed{\mathbb{E}_{\mathcal{D}}(MSE) = \text{Bias}^2 + \text{Variance} + \sigma^2}$$

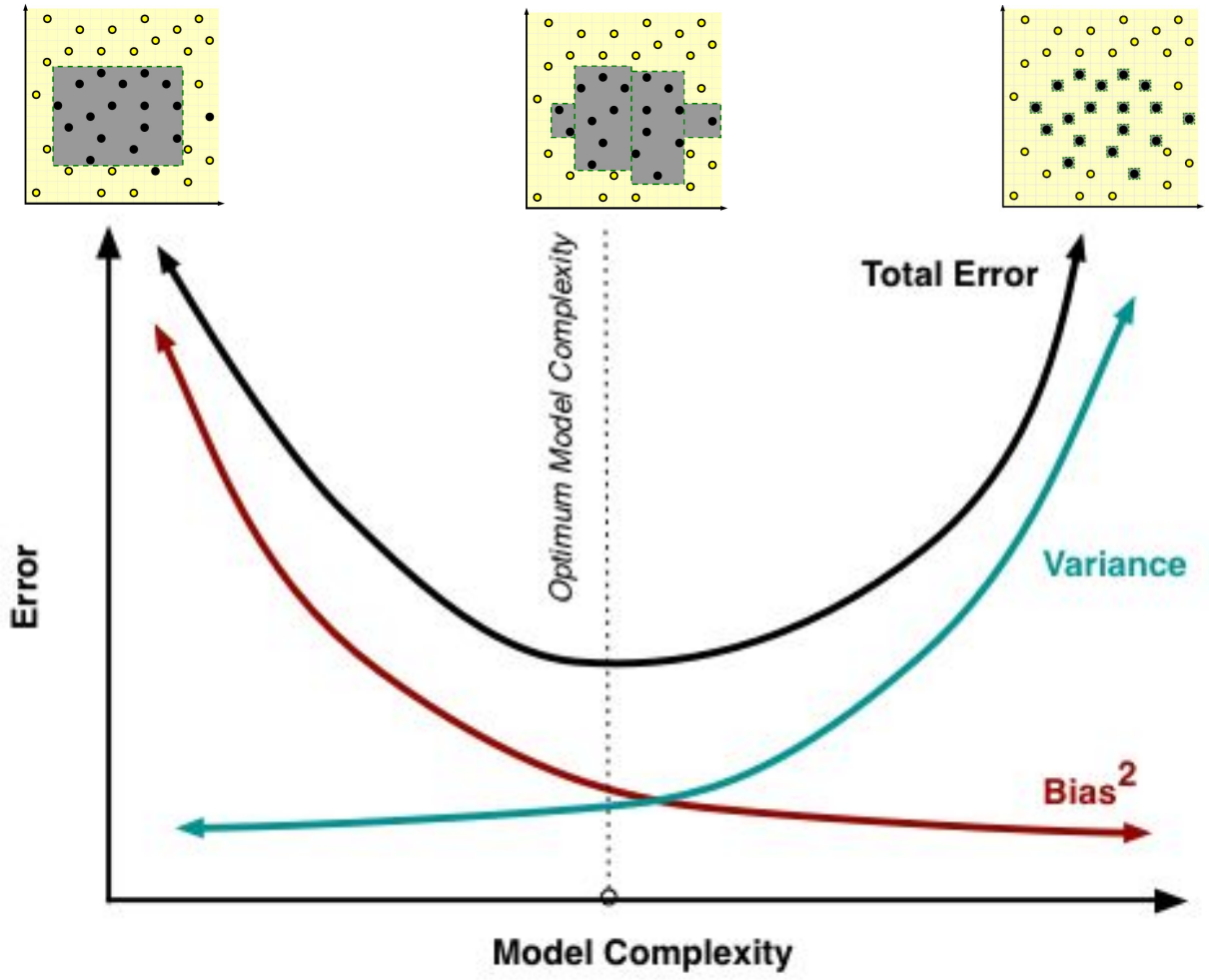
Error debido a sesgo (o bias): diferencia entre predicción del modelo (o promedio de predicciones) y valor correcto.

Error debido a varianza: variabilidad de la predicción de un modelo para distintos datos dados.

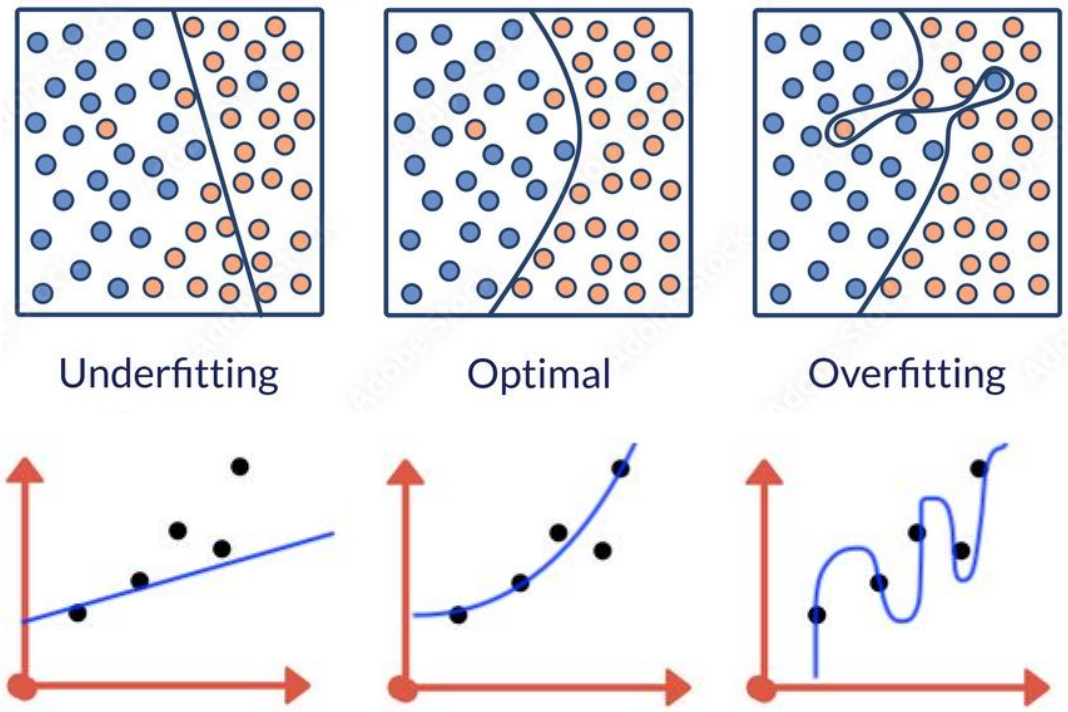
Disminuir el error: buscamos un método que tenga **bajo sesgo y baja varianza**.

Trade-off Sesgo / Varianza





Trade-off Sesgo / Varianza



Alto sesgo

Alta varianza

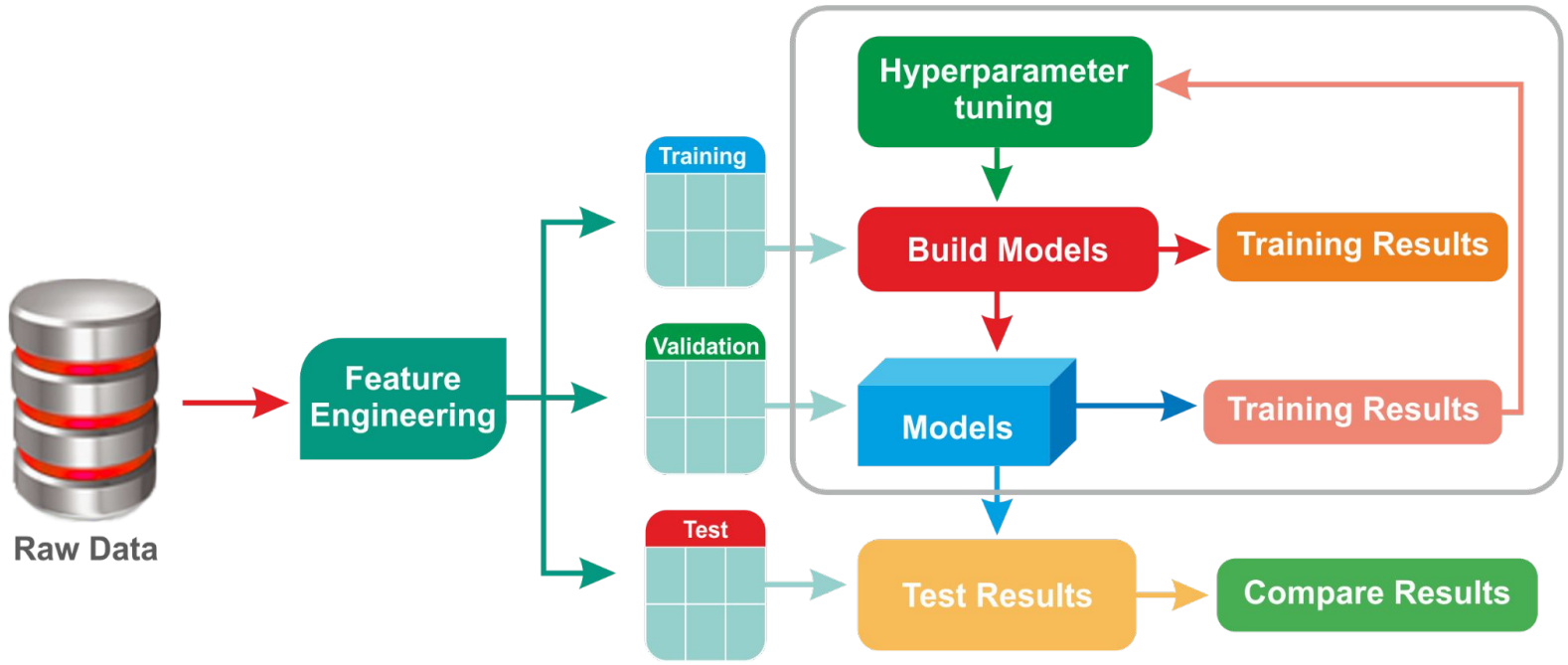
Itinerario de la clase de hoy

- Complejidad de modelos y generalización
- Separación de datos
- Sobre y sub ajuste
- Trade off sesgo varianza
- Parámetros e hiperparámetros
- Regularización
- Flujo general del trabajo en ciencia de datos
- Interpretabilidad y selección de modelos

Parámetros e hiperparámetros

Parámetros: Son quienes **aprenden** de los datos. Se **ajustan** en el entrenamiento (ej, w en RL).

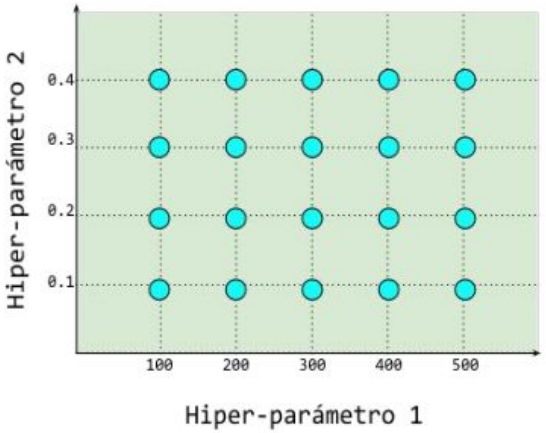
Hiperparámetros: Son quienes definen la **complejidad del modelo**. Se **seleccionan** (tuning) de las configuraciones de entrenamiento, según su impacto en el aprendizaje (ej, órden del polinomio en RL).

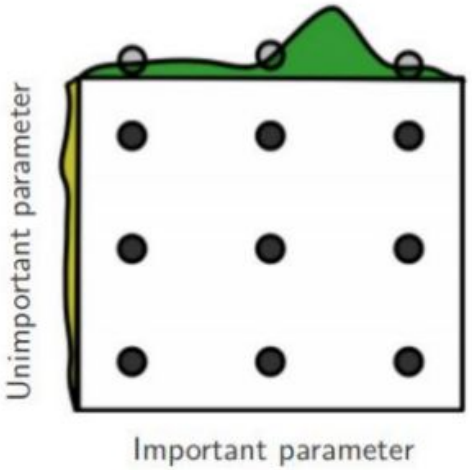


Selección de hiperparámetros

Grid Search:

Hacer una grilla

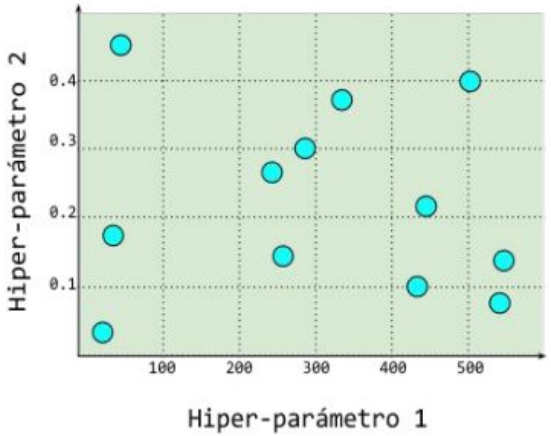


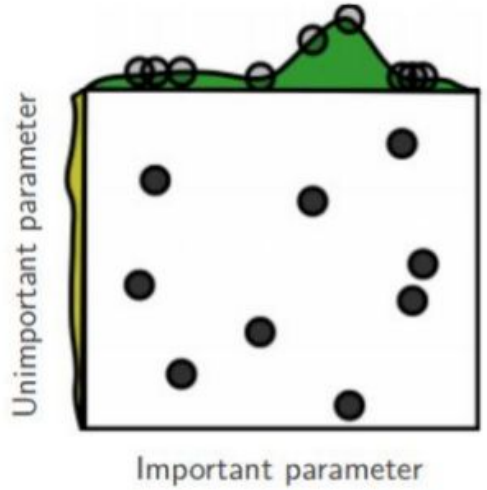


Evaluamos los modelos entrenados con distintos valores de hiperparámetros usando validación cruzada y seleccionamos el óptimo.

Random Search:

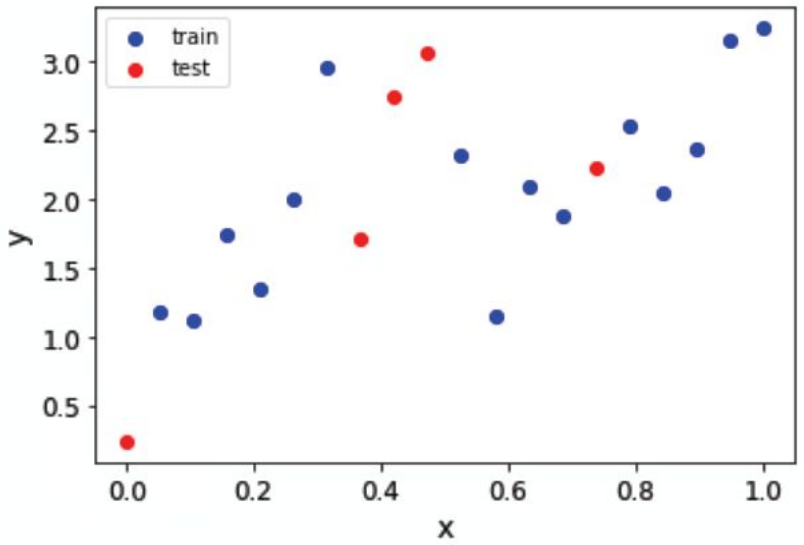
Explorar al azar

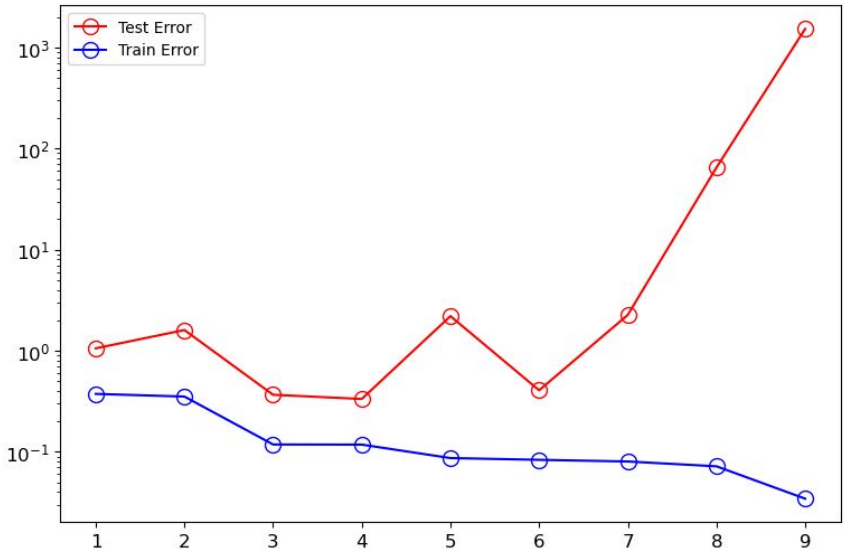


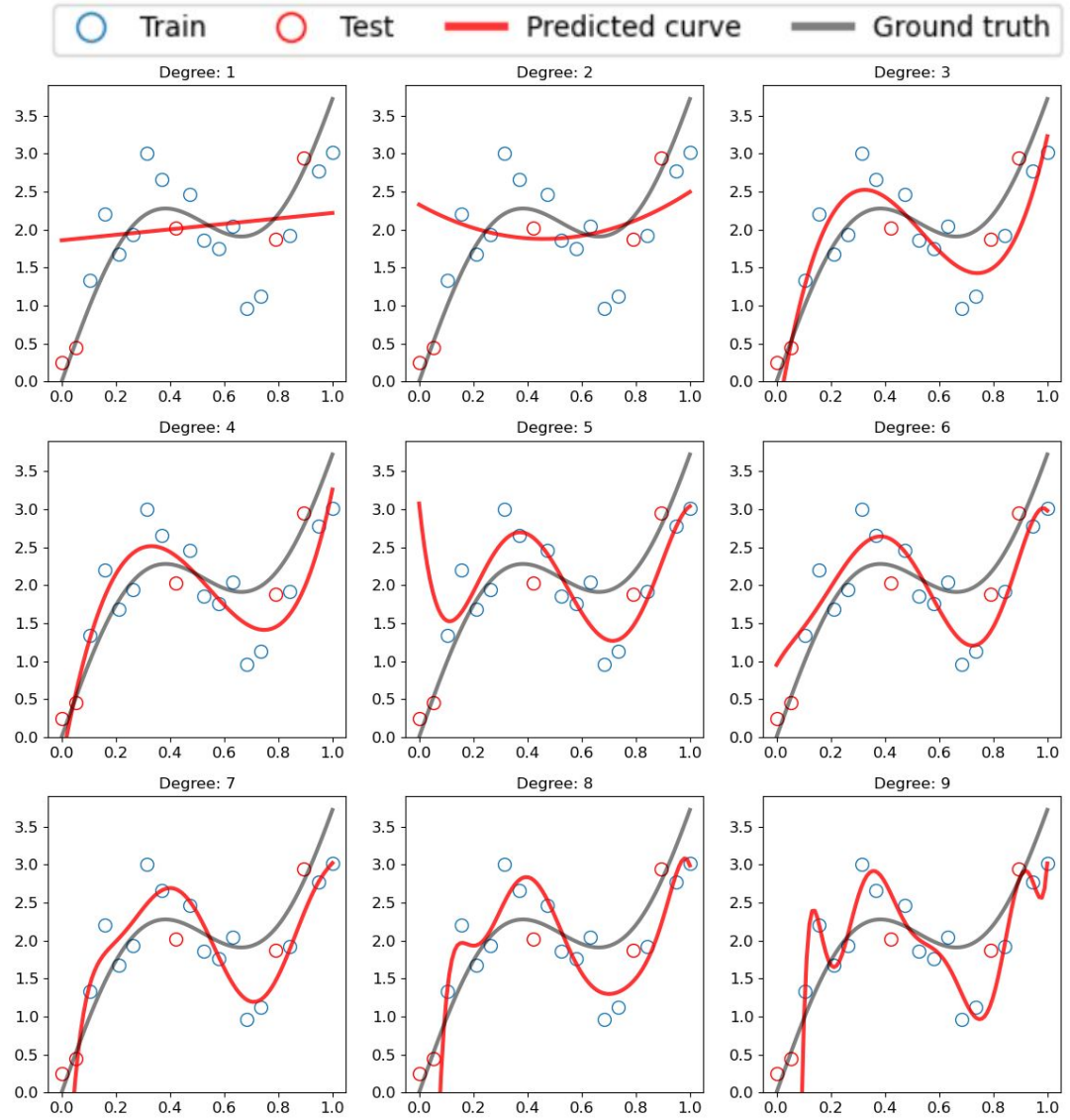


Con Random Search puedo encontrar valores de hiperparámetros que están más cerca del óptimo real.

Selección de hiperparámetros







Itinerario de la clase de hoy

- Complejidad de modelos y generalización
- Separación de datos
- Sobre y sub ajuste
- Trade off sesgo varianza
- Parámetros e hiperparámetros
- Regularización
- Flujo general del trabajo en ciencia de datos
- Interpretabilidad y selección de modelos

Regularización

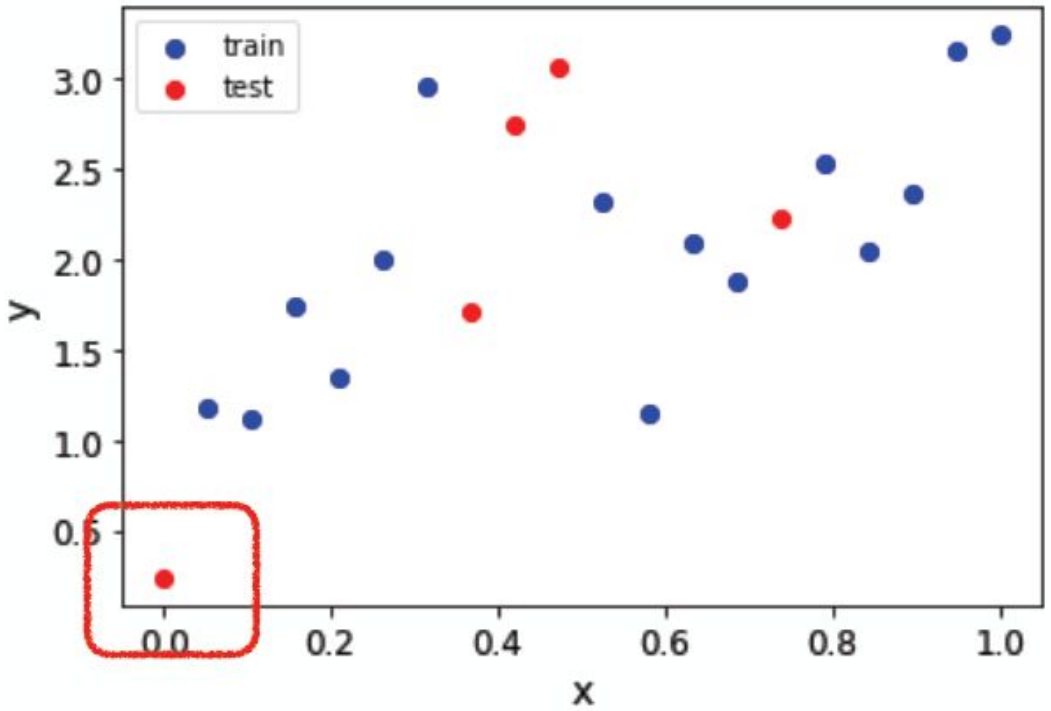
El error puede depender de los valores extremos (outliers).

Cuanto más complejo el modelo, más sensible a outliers.

Para evitar estos errores se suelen aplicar restricciones a los pesos.

Esta penalización se conoce como regularización, y su magnitud se regula con un hiperparámetro λ.

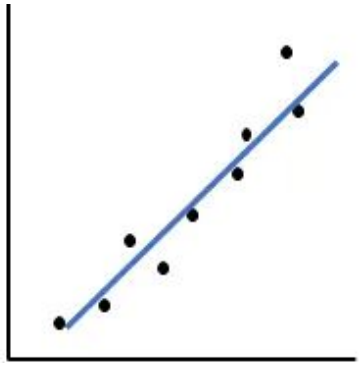
Consiste en agregar un término de penalización a la función costo.

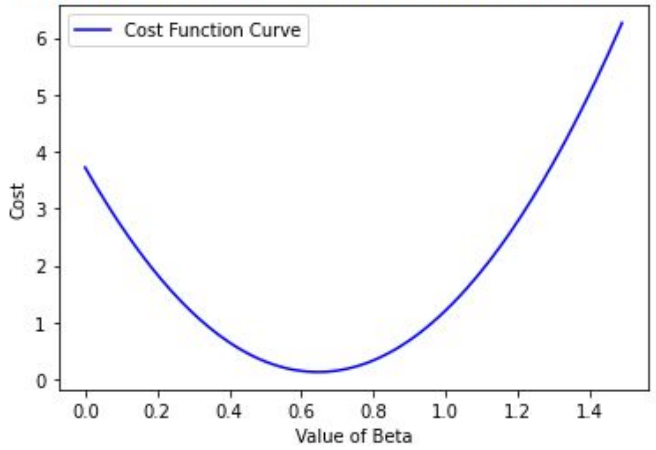


$$+\lambda \sum_{j=1}^{M} \left\| \hat{\beta}_j \right\|_p$$

Regularización

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$$





Error cuadrático medio

$$J(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2 + \lambda \sum_{j=1}^{M} \left\|\hat{\beta}_j\right\|_p$$

error

Función costo

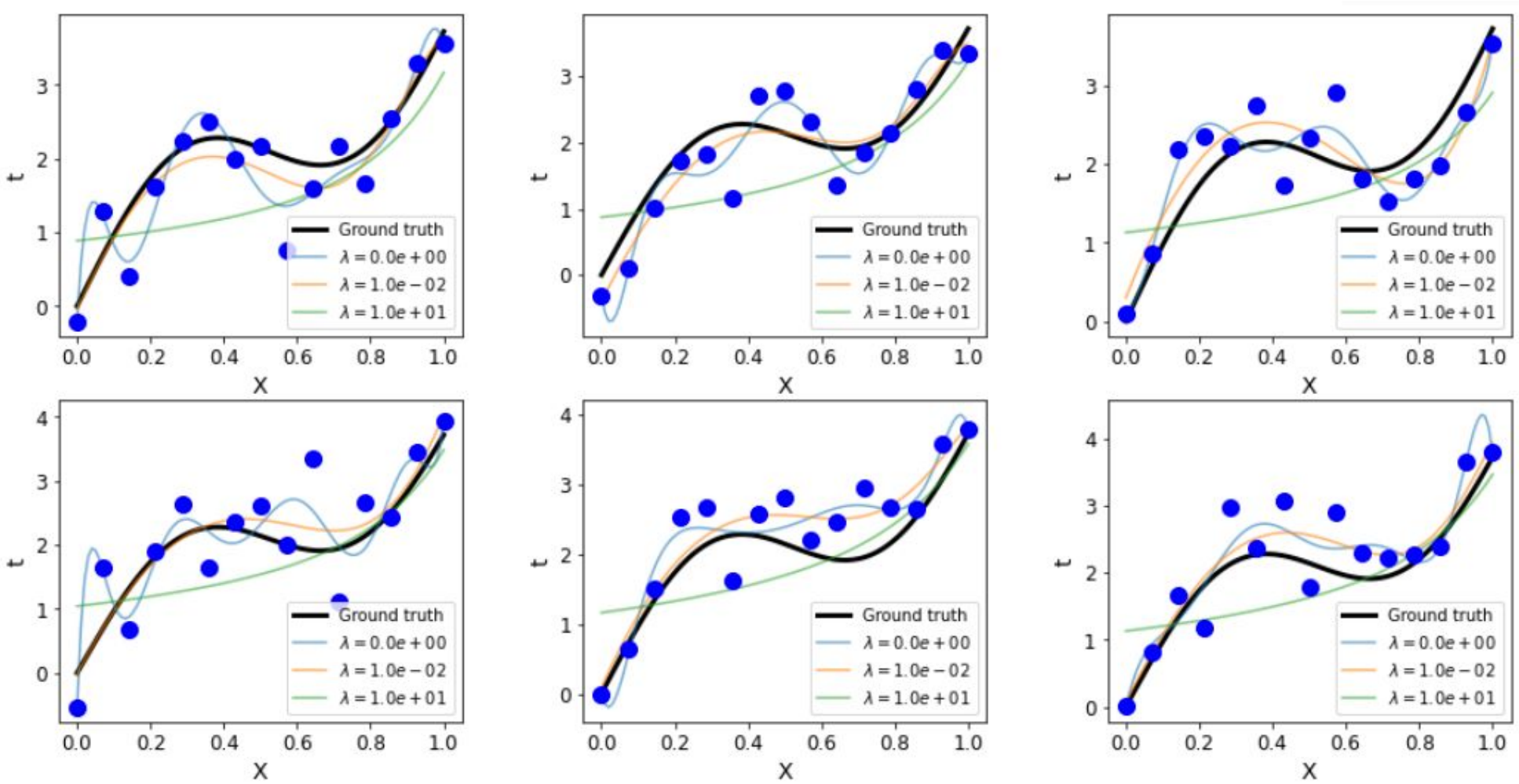
Penalización

Podemos agregar a la **función costo** un **término de penalización**. Si algún parámetro toma un valor muy alto, este término también será alto. Para minimizar la función costo voy a tener que penalizar ese parámetro y restringirlo para que no tome un valor alto.

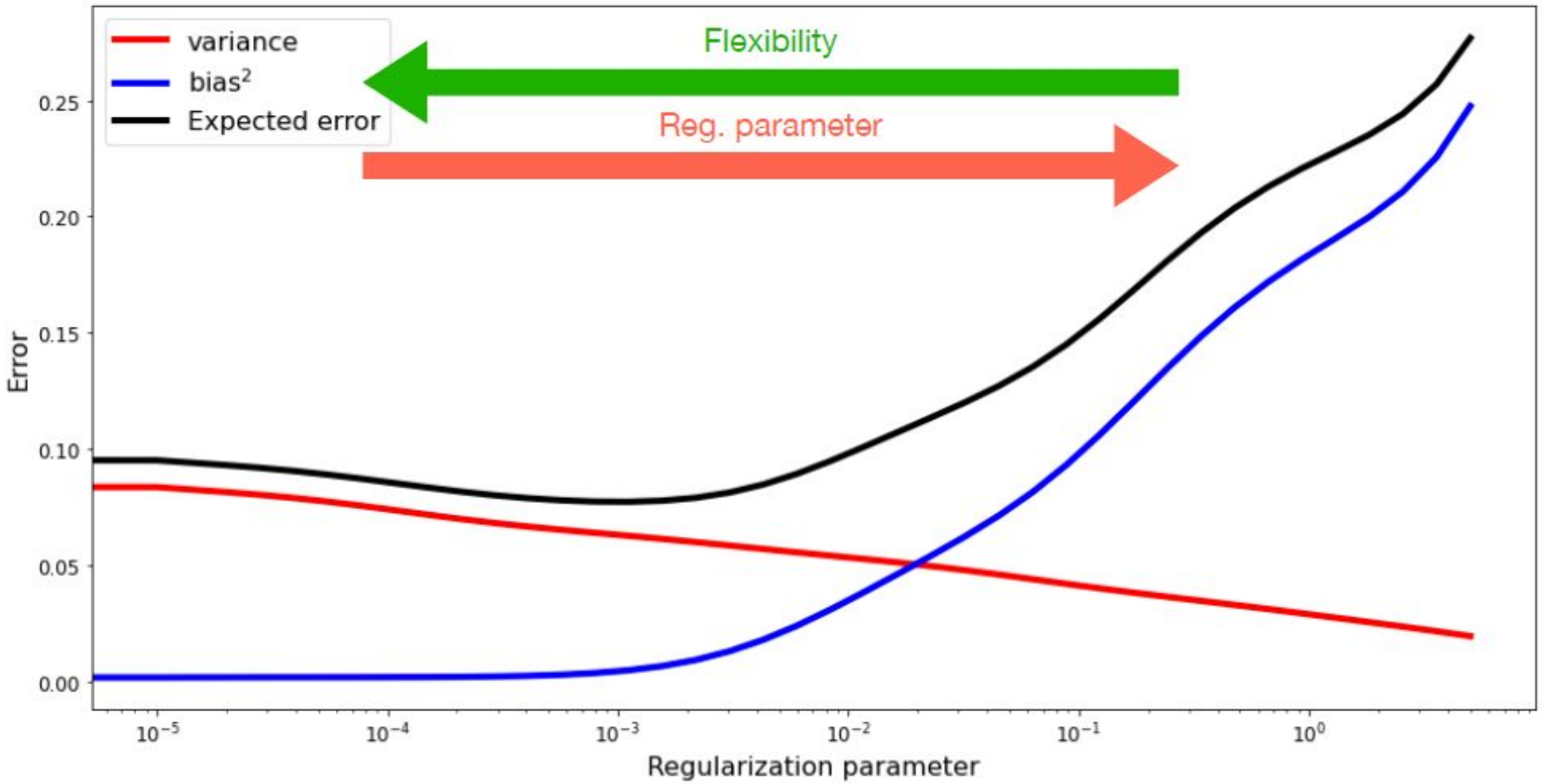
Estoy agregando un **hiperparámetro (lambda)** que regula la **magnitud de penalización**, asignándo más importancia a ciertos parámetros por sobre otros.

Permitirá regular la **flexibilidad del modelo**, haciendo que sea **menos sensible al ruido** y, por lo tanto, **reduciendo su varianza** (similar a reducir la complejidad del modelo).

Regularización y trade-off Sesgo / Varianza

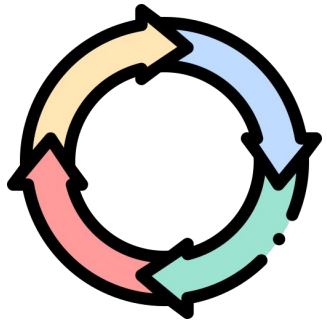


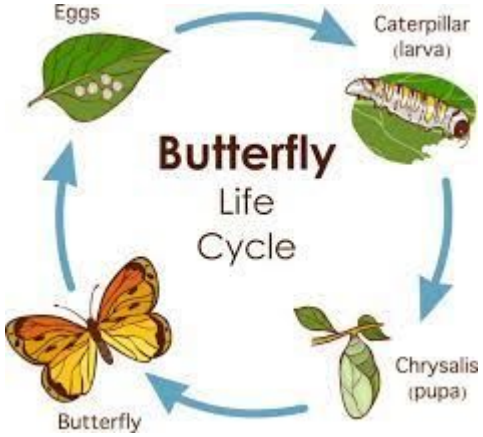
Regularización y trade-off Sesgo / Varianza



Ampliaremos…

Itinerario de la clase de hoy





- Complejidad de modelos y generalización
- Separación de datos
- Sobre y sub ajuste
- Trade off sesgo varianza
- Parámetros e hiperparámetros
- Regularización
- Flujo general del trabajo en ciencia de datos
- Interpretabilidad y selección de modelos

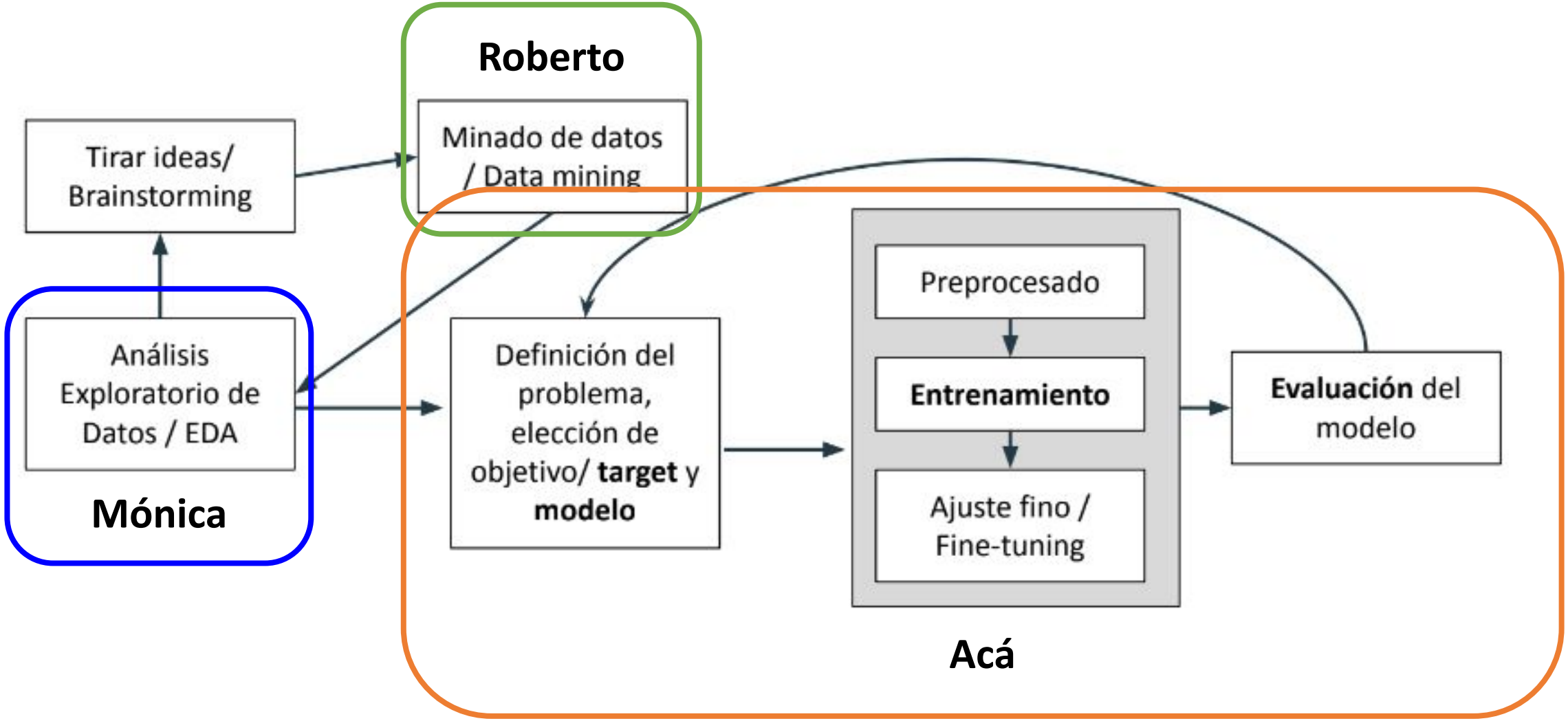
Esquema del aprendizaje automático

1. Definición de **Tarea objetivo T** a aprender.
2. Definición de medidas de **Performance P**.
3. Datos o **Experiencia E**.

Flujo de trabajo:

- Selección de técnica y algoritmo de aprendizaje automático.
- Selección de atributos o features.
- Elección de hiper-parámetros del modelo, ajuste.
- Evaluación del modelo sobre nuevos datos.

Flujo general de trabajo en Ciencias de Datos



Pre-procesado

Ingeniería de datos:

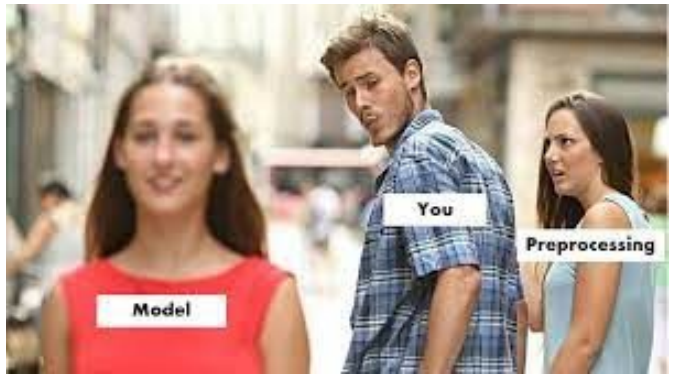
Adecuamos los datos de manera tal que los modelos funcionen mejor

Limpieza de datos:

Decidir que se hace con datos faltantes y anomalías (o outliers) estadísticas.

Preprocesado de datos:

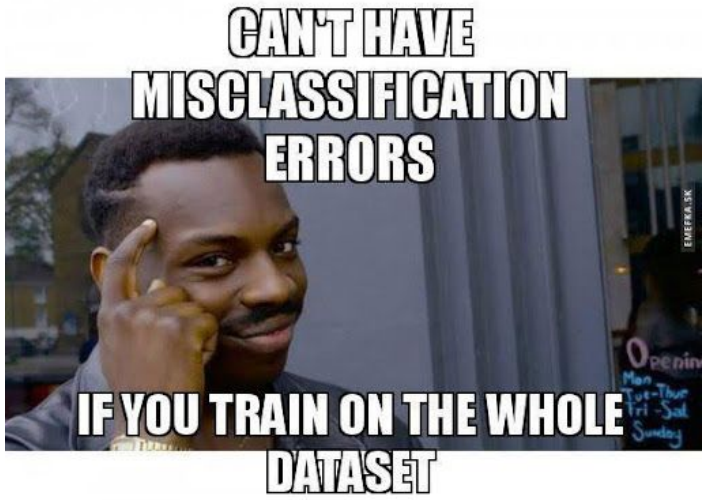
Estandarización, variables categóricas, etc.



Entrenamiento

- Comenzamos con modelos simples que sirvan como referencia.
- Exploración de distintos modelos.
- Utilizar los resultados iniciales / intermedios para ir mejorando la selección de modelo.
- Elegir los mejores modelos y ajustarlos a los datos.

Ojo con el sobreajuste!

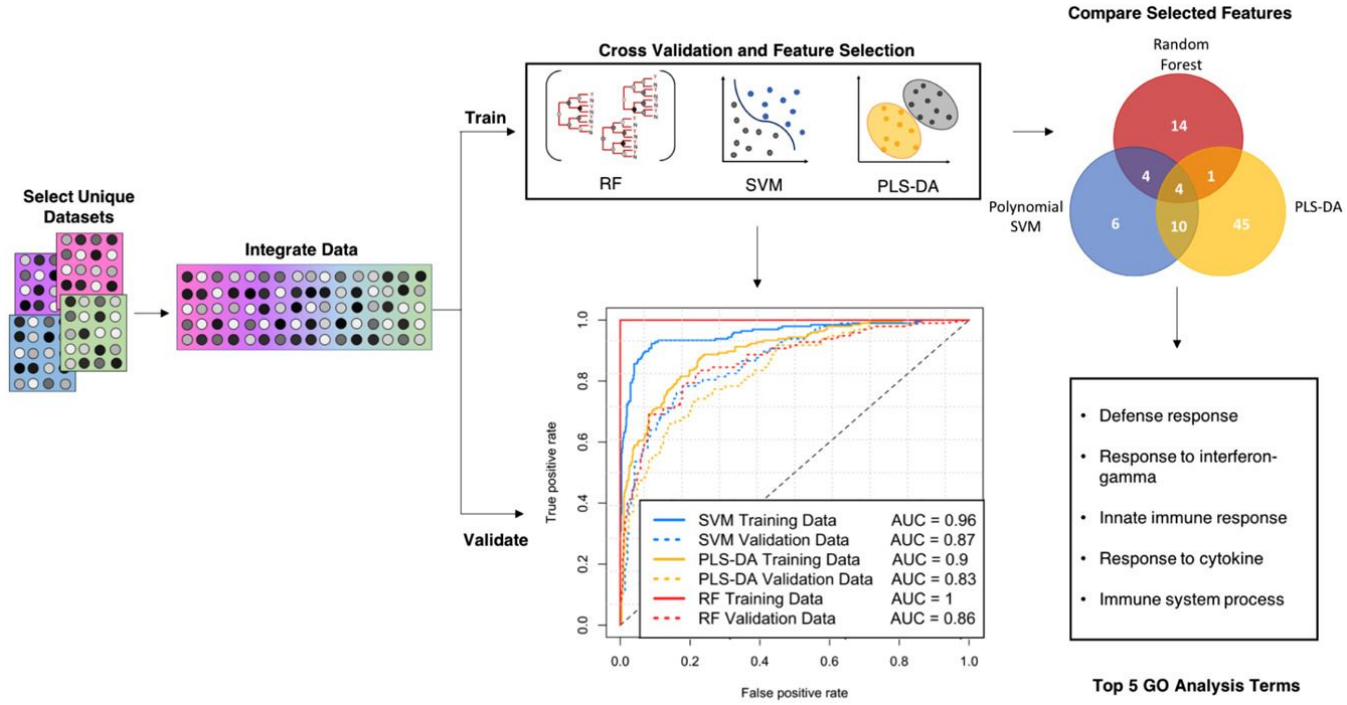


Selección y comparación de modelos

Australia's Solar City	Model	r	$RMSE$ (MJ·m^{-2}·day^{-1})	MAE (MJ·m^{-2}·day^{-1})	$RMSE_{ss}$
Blacktown	DBN_{10}	0.994	0.546	0.45	0.824
	DNN_{2Nadam}	0.99	0.706	0.503	0.773
	ANN	0.989	0.739	0.536	0.739
	DT	0.955	1.309	0.979	0.579
	RF	0.982	0.798	0.635	0.744
	GBM	0.988	0.664	0.568	0.787
	XGBR	0.985	0.727	0.589	0.766
Adelaide	DBN_{10}	0.997	0.503	0.426	0.863
	DNN_{2SGD}	0.996	0.636	0.546	0.826
	ANN	0.997	0.653	0.529	0.824
	DT	0.985	1.063	0.791	0.713
	RF	0.989	0.895	0.652	0.754
	GBM	0.988	0.906	0.72	0.758
	XGBR	0.992	0.737	0.577	0.801
Central Victoria	DBN_{10}	0.996	0.614	0.498	0.836
	DNN_{2SGD}	0.994	0.798	0.592	0.787
	ANN	0.984	1.276	0.995	0.682
	DT	0.961	1.696	1.217	0.553
	RF	0.984	1.094	0.854	0.714
	GBM	0.988	0.942	0.799	0.753
	XGBR	0.987	0.992	0.825	0.74
Townsville	DBN_{10}	0.974	0.773	0.627	0.718
	$DNN_{2RMSProp}$	0.967	0.868	0.646	0.682
	ANN	0.972	0.991	0.858	0.641
	DT	0.951	1.181	0.973	0.572
	RF	0.95	1.212	0.971	0.559
	GBM	0.947	1.254	1.006	0.539
	XGBR	0.953	1.205	0.959	0.56
Average of 5 Study Sites	DBN_{10}	0.990	0.609	0.500	0.810
	$DNN_{2RMSProp}$	0.987	0.752	0.572	0.767
	ANN	0.986	0.915	0.730	0.722
	DT	0.963	1.312	0.990	0.604
	RF	0.976	1.000	0.778	0.693
	GBM	0.978	0.942	0.773	0.709
	XGBR	0.979	0.915	0.738	0.717

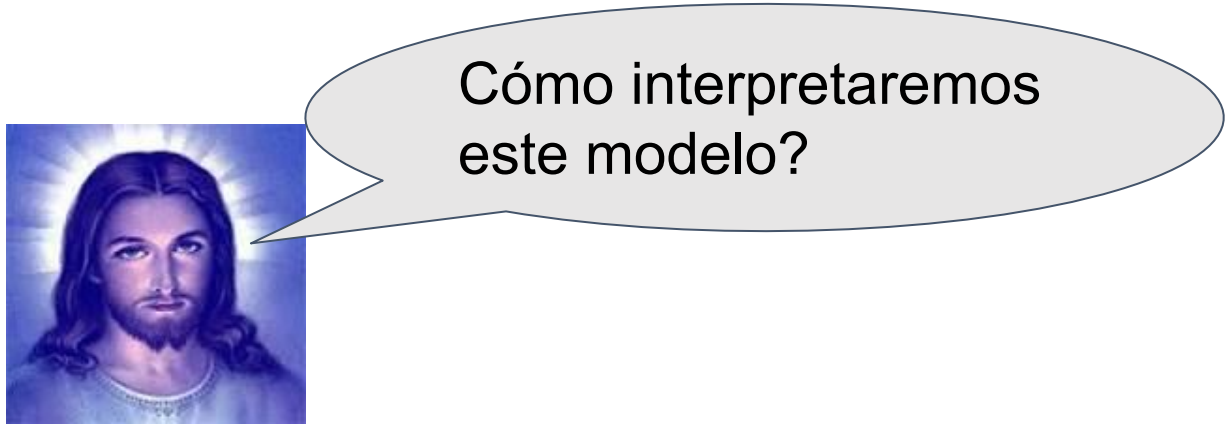
Usamos validación cruzada para comparar

- Tipos de modelos y algoritmos
- Atributos
- Hiperparámetros

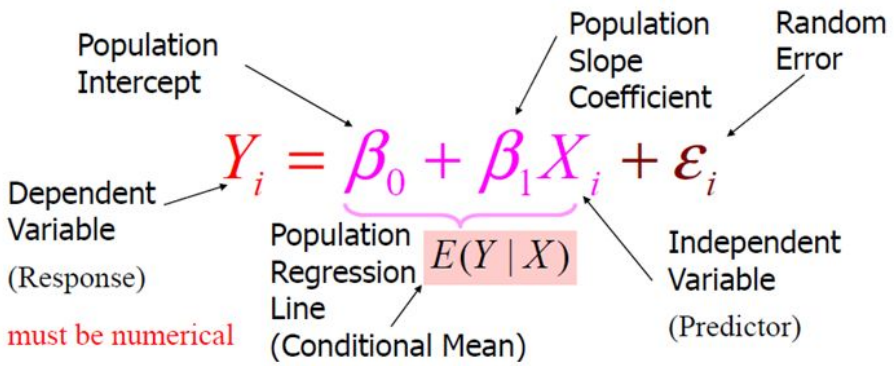


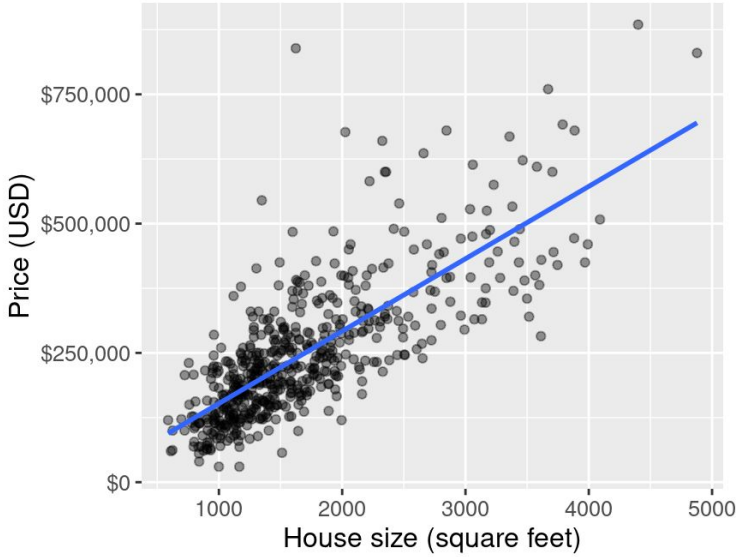
Evaluación del modelo

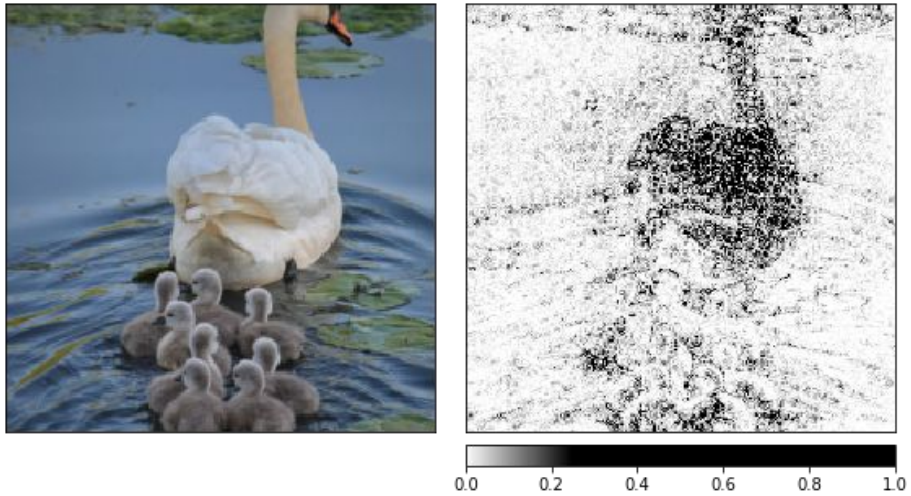
- Evaluamos el modelo en datos no explorados hasta ahora (conjunto de testeo)
- Obtenemos el error estimado en datos nuevos
- Estudiamos el modelo para ganar interpretabilidad (si es posible)
- Encontrar donde no funciona el modelo
- De ser necesario, volver a empezar

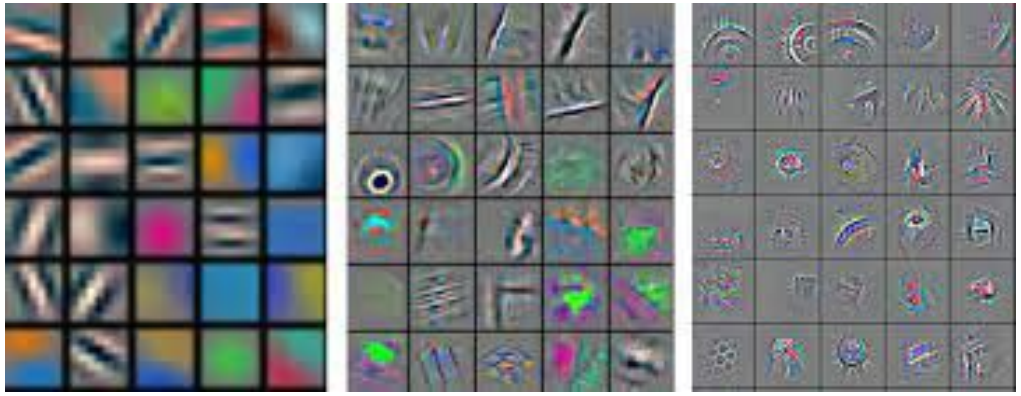


Interpretabilidad









Ampliaremos…

A los Colabs…!!