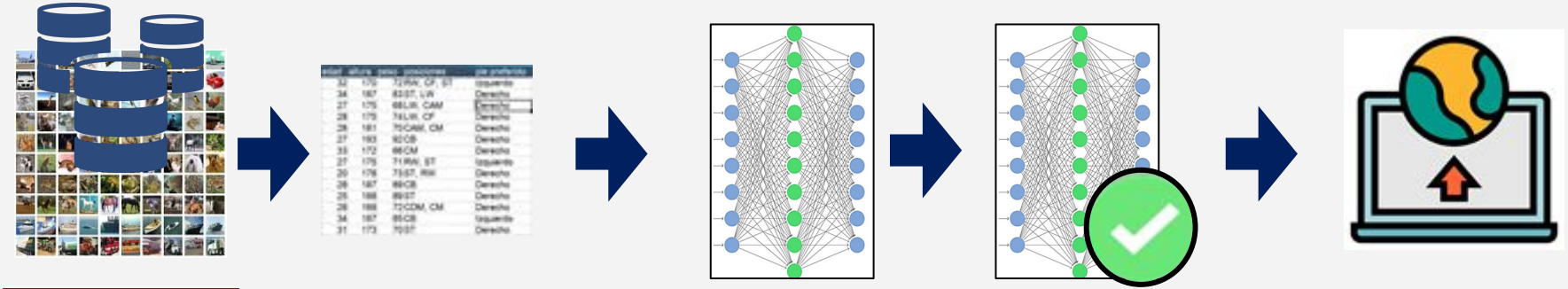
Evaluación de modelos

CRISP-DM

Pipeline de procesamiento en Machine Learning



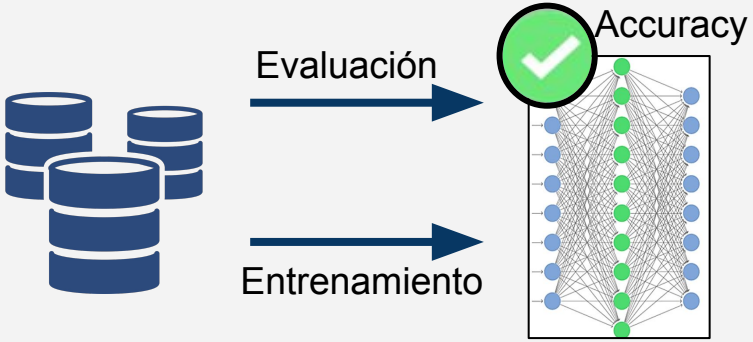
Comprensión de los Datos | Preprocesamiento | Modelado | Evaluación | Despliegue

- **Evaluación**
  - ¿Qué tan bueno es mi modelo?
  - Con datos nuevos(Sin etiquetar, No “vistos”)

Evaluación de Modelos de Clasificación

Hasta ahora:

- Accuracy con datos de Entrenamiento



Analogía:

Enseñar a multiplicar

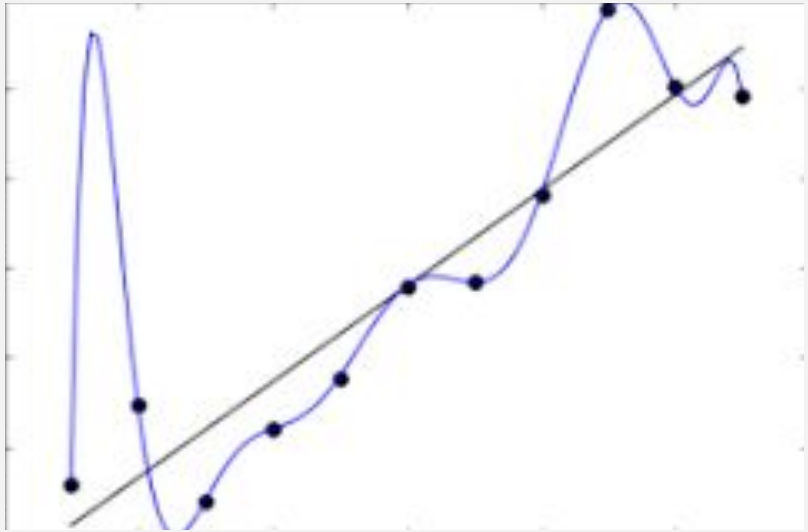
- Práctica: 10 ejercicios
- Examen: los mismos 10 ejercicios
  - ¿aprenden el dominio? o ¿memorizan los resultados?

142 × 309 1,278 42,600 43,878	883 × 470 61,810 353,200 415,010	556 × 288 4,448 44,480 111,200 160,128	491 × 622 982 9,820 294,600 305,402	321 × 508 2,568 160,500 163,068
554 × 628 4,432 11,080 332,400 347,912	668 × 664 2,672 40,080 400,800 443,552	469 × 909 4,221 422,100 426,321	887 × 138 7,096 26,610 88,700 122,406	908 × 219 8,172 9,080 181,600 198,852

Evaluación de Modelos de Clasificación

Qué sería aprender los datos “de memoria” en Redes neuronales?

- Podríamos tener un polinomio que pase exactamente por todos los patrones de entrenamiento.
- Spoiler: *overfitting*.



Evaluación de Modelos de Clasificación

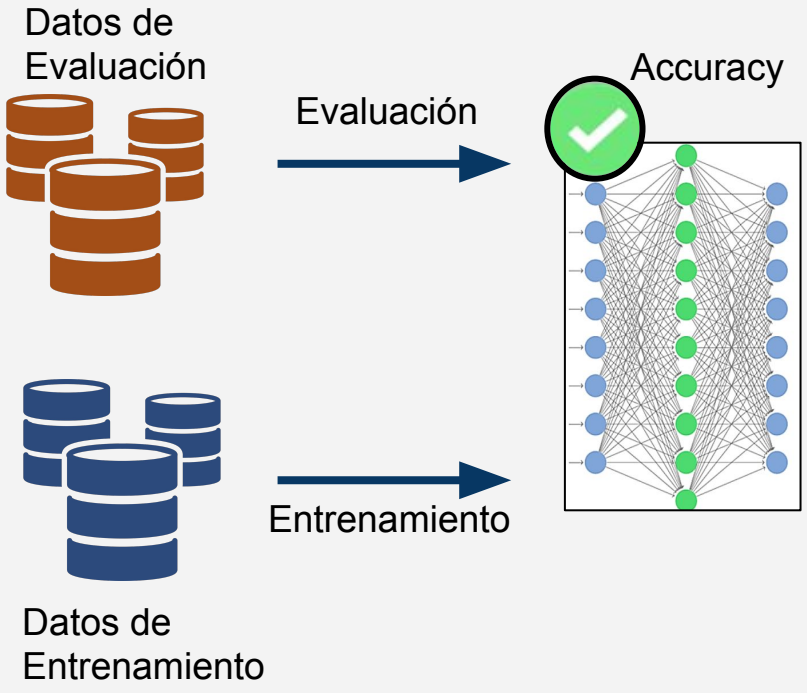
Lo que haremos:

- Accuracy con datos de Evaluación

Analogía:

Enseñar a multiplicar

- Práctica: 10 ejercicios
- Examen: otros 10 ejercicios
  - ¿aprenden el dominio? o ¿memorizan los resultados?



Evaluación de Modelos de Clasificación

Datos de evaluación

- Sinónimos: Test Set, Testing set, datos de prueba.
- Deberían tener el mismo formato que los datos de entrenamiento.
- Generalmente se obtienen de la misma base de datos.
  - Dividir en dos conjuntos
  - Generalmente 80%- 20%

Evaluación de Modelos de Clasificación

Accuracy en ambos conjuntos

- Acc en Train set: ¿Fue entrenado correctamente el modelo?
- Acc en Test set: ¿Cómo se comporta el modelo con datos nuevos?

Cómo generar conjunto de evaluación

División fija

- 12 ejemplos
  - 3 Quemados - 9 Ninguno
- Divido ⅔ y ⅓
  - Train: ⅔ = 66% = 8 ej
  - Test: ⅓ = 33% = 4 ej
- Problemas?
  - Test sin clase Quemado
  - Ordenado por Pelo

Nombre	Pelo	Estatura	Peso	Protector	Resultado
Sara	Rubio	Promedio	Ligero	No	Quemado
Diana	Rubio	Alta	Promedio	Si	Quemado
Alexis	Rubio	Baja	Promedio	Si	Ninguno
Ana	Rubio	Baja	Promedio	No	Ninguno
Emilia	Rubio	Promedio	Pesado	No	Quemado
Pedro	Rubio	Alta	Pesado	No	Ninguno
Jacinta	Castaño	Alta	Promedio	Si	Ninguno
María	Castaño	Baja	Promedio	No	Ninguno
Juan	Castaño	Promedio	Pesado	No	Ninguno
Catalina	Castaño	Baja	Ligero	Si	Ninguno
Felipe	Pelirrojo	Promedio	Pesado	No	Ninguno
Roberta	Pelirrojo	Alta	Ligero	Si	Ninguno

Train

Nombre	Pelo	Estatura	Peso	Protector	Resultado
Sara	Rubio	Promedio	Ligero	No	Quemado
Diana	Rubio	Alta	Promedio	Si	Quemado
Alexis	Rubio	Baja	Promedio	Si	Ninguno
Ana	Rubio	Baja	Promedio	No	Ninguno
Emilia	Rubio	Promedio	Pesado	No	Quemado
Pedro	Rubio	Alta	Pesado	No	Ninguno
Jacinta	Castaño	Alta	Promedio	Si	Ninguno
María	Castaño	Baja	Promedio	No	Ninguno

Test

Nombre	Pelo	Estatura	Peso	Protector	Resultado
Juan	Castaño	Promedio	Pesado	No	Ninguno
Catalina	Castaño	Baja	Ligero	Si	Ninguno
Felipe	Pelirrojo	Promedio	Pesado	No	Ninguno
Roberta	Pelirrojo	Alta	Ligero	Si	Ninguno

Cómo generar conjunto de evaluación

División Estratificada

- 12 ejemplos
  - 3 Quemados - 9 Ninguno
- Divido ⅔ y ⅓
  - Train: ⅔ = 66% = 8 ej
  - Test: ⅓ = 33% = 4 ej
- Estratificada
  - Misma proporción de clases
  - Sigue ordenado

Nombre	Pelo	Estatura	Peso	Protector	Resultado
Sara	Rubio	Promedio	Ligero	No	Quemado
Diana	Rubio	Alta	Promedio	Si	Quemado
Alexis	Rubio	Baja	Promedio	Si	Ninguno
Ana	Rubio	Baja	Promedio	No	Ninguno
Emilia	Rubio	Promedio	Pesado	No	Quemado
Pedro	Rubio	Alta	Pesado	No	Ninguno
Jacinta	Castaño	Alta	Promedio	Si	Ninguno
María	Castaño	Baja	Promedio	No	Ninguno
Juan	Castaño	Promedio	Pesado	No	Ninguno
Catalina	Castaño	Baja	Ligero	Si	Ninguno
Felipe	Pelirrojo	Promedio	Pesado	No	Ninguno
Roberta	Pelirrojo	Alta	Ligero	Si	Ninguno

Train

Nombre	Pelo	Estatura	Peso	Protector	Resultado
Sara	Rubio	Promedio	Ligero	No	Quemado
Diana	Rubio	Alta	Promedio	Si	Quemado
Alexis	Rubio	Baja	Promedio	Si	Ninguno
Ana	Rubio	Baja	Promedio	No	Ninguno
Pedro	Rubio	Alta	Pesado	No	Ninguno
Jacinta	Castaño	Alta	Promedio	Si	Ninguno
María	Castaño	Baja	Promedio	No	Ninguno
Juan	Castaño	Promedio	Pesado	No	Ninguno

Test

Nombre	Pelo	Estatura	Peso	Protector	Resultado
Emilia	Rubio	Promedio	Pesado	No	Quemado
Catalina	Castaño	Baja	Ligero	Si	Ninguno
Felipe	Pelirrojo	Promedio	Pesado	No	Ninguno
Roberta	Pelirrojo	Alta	Ligero	Si	Ninguno

Cómo generar conjunto de evaluación

División Estratificada Aleatoria

- 12 ejemplos
  - 3 Quemados - 9 Ninguno
- Divido ⅔ y ⅓
  - Train: ⅔ = 66% = 8 ej
  - Test: ⅓ = 33% = 4 ej
- Estratificada y aleatoria
  - Misma proporción de clases
  - Orden Aleatorio

Nombre	Pelo	Estatura	Peso	Protector	Resultado
Sara	Rubio	Promedio	Ligero	No	Quemado
Diana	Rubio	Alta	Promedio	Si	Quemado
Alexis	Rubio	Baja	Promedio	Si	Ninguno
Ana	Rubio	Baja	Promedio	No	Ninguno
Emilia	Rubio	Promedio	Pesado	No	Quemado
Pedro	Castaño	Alta	Pesado	No	Ninguno
Jacinta	Castaño	Alta	Promedio	Si	Ninguno
María	Castaño	Baja	Promedio	No	Ninguno
Juan	Castaño	Promedio	Pesado	No	Ninguno
Catalina	Castaño	Baja	Ligero	Si	Ninguno
Felipe	Pelirrojo	Promedio	Pesado	No	Ninguno
Roberta	Pelirrojo	Alta	Ligero	Si	Ninguno

Train

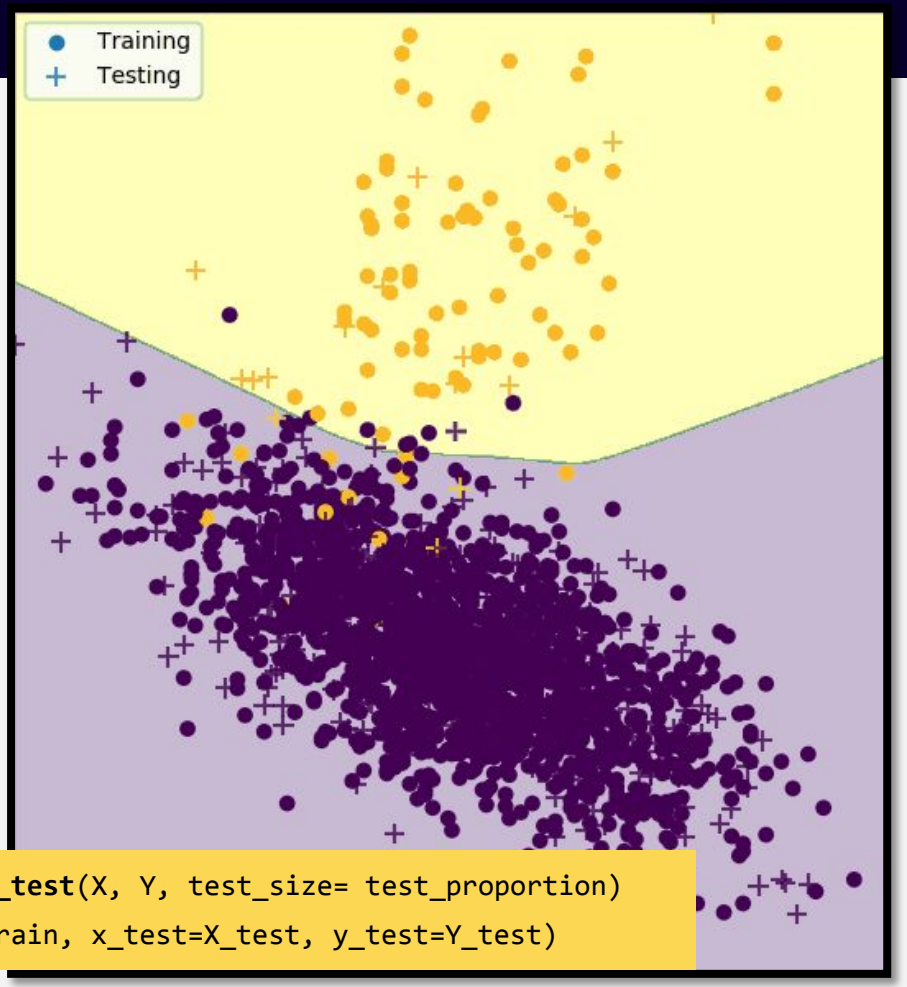
Nombre	Pelo	Estatura	Peso	Protector	Resultado
Diana	Rubio	Alta	Promedio	Si	Quemado
Alexis	Rubio	Baja	Promedio	Si	Ninguno
Emilia	Rubio	Promedio	Pesado	No	Quemado
Jacinta	Castaño	Alta	Promedio	Si	Ninguno
María	Castaño	Baja	Promedio	No	Ninguno
Juan	Castaño	Promedio	Pesado	No	Ninguno
Catalina	Castaño	Baja	Ligero	Si	Ninguno
Roberta	Pelirrojo	Alta	Ligero	Si	Ninguno

Test

Nombre	Pelo	Estatura	Peso	Protector	Resultado
Sara	Rubio	Promedio	Ligero	No	Quemado
Ana	Rubio	Baja	Promedio	No	Ninguno
Pedro	Castaño	Alta	Pesado	No	Ninguno
Felipe	Pelirrojo	Promedio	Pesado	No	Ninguno

Training / Testing

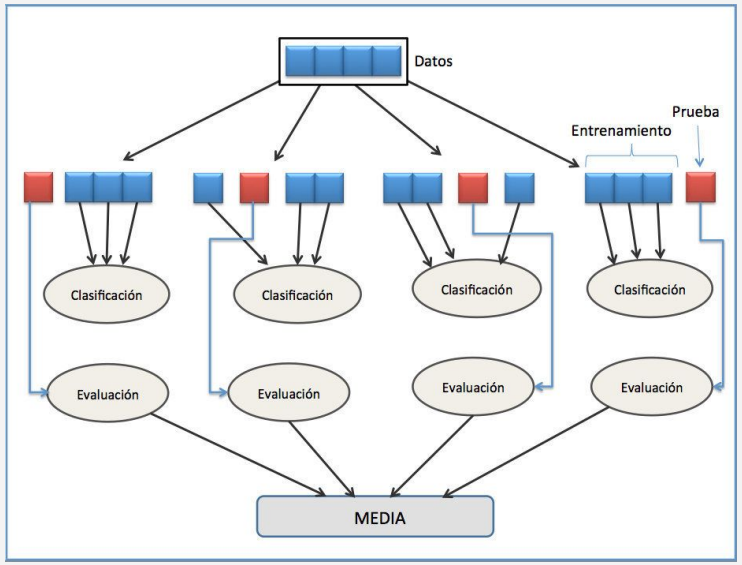
Ejemplo dataset 2D
estratificado aleatoriamente




```
X_train, X_test, Y_train, Y_test = AAPutils.dividir_train_test(X, Y, test_size= test_proportion)
AAPutils.plot_frontera_de_decision_2D(model, X_train, Y_train, x_test=X_test, y_test=Y_test)
```

Validación cruzada

- Una sola evaluación no es suficiente para garantizar que mi modelo funciona correctamente
  - ¿Qué ocurre si los datos de “testing” son muy “fáciles” de clasificar/predecir?
- Una forma de solucionar esto es con la validación cruzada.
  - Hacer K ejecuciones de forma estratificada partiendo en k lugares distintos el conjunto de datos.
  - Muchas veces se reemplaza esto por K ejecuciones aleatorias independientes.



Interpretación Accuracy en Train/Test set

Situaciones con Accuracy:

- Train **alto** - Test **alto**
  - Buen modelo
  - sirve para nuevos datos

Situación ideal. El modelo Generaliza

- Train **alto** - Test **bajo**
  - Modelo “bien” entrenado
  - No sirve para nuevos datos.

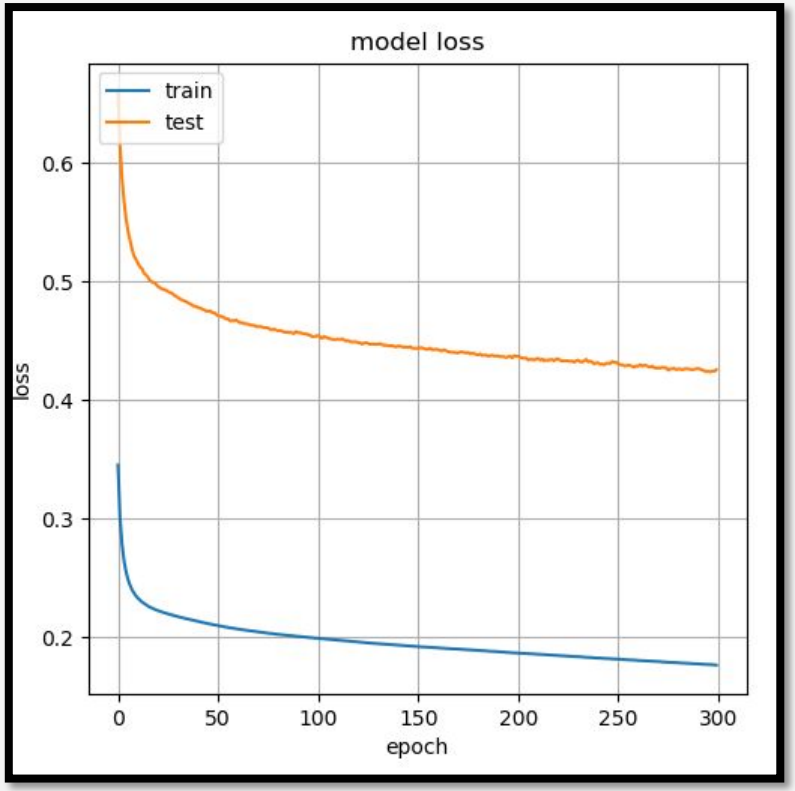
Síntomas de sobreajuste

- Train **bajo** - Test no importa
  - Modelo mal entrenado

Síntomas de sub-ajuste

Curvas de entrenamiento

- Las curvas de entrenamiento nos permiten ver cómo se comporta el modelo (error) en cada iteración.
- No siempre la última iteración será el mejor modelo.
- La diferencia entre la curva de training y de testing nos permite visualizar el comportamiento del modelo con nuevos datos.

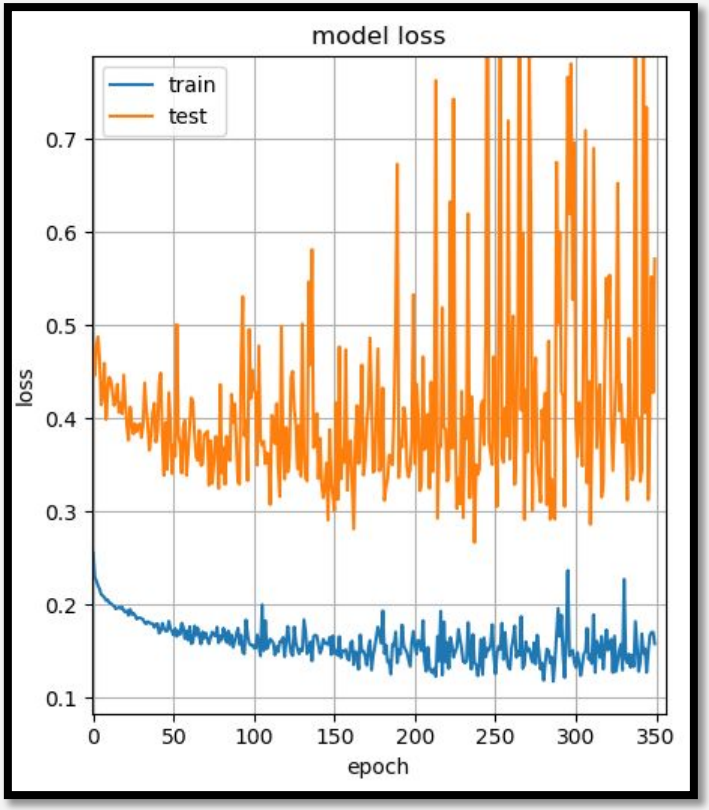


Curvas de entrenamiento

También nos permiten visualizar errores en la configuración de los hiperparámetros. Por ejemplo ¿Qué ocurre aquí?

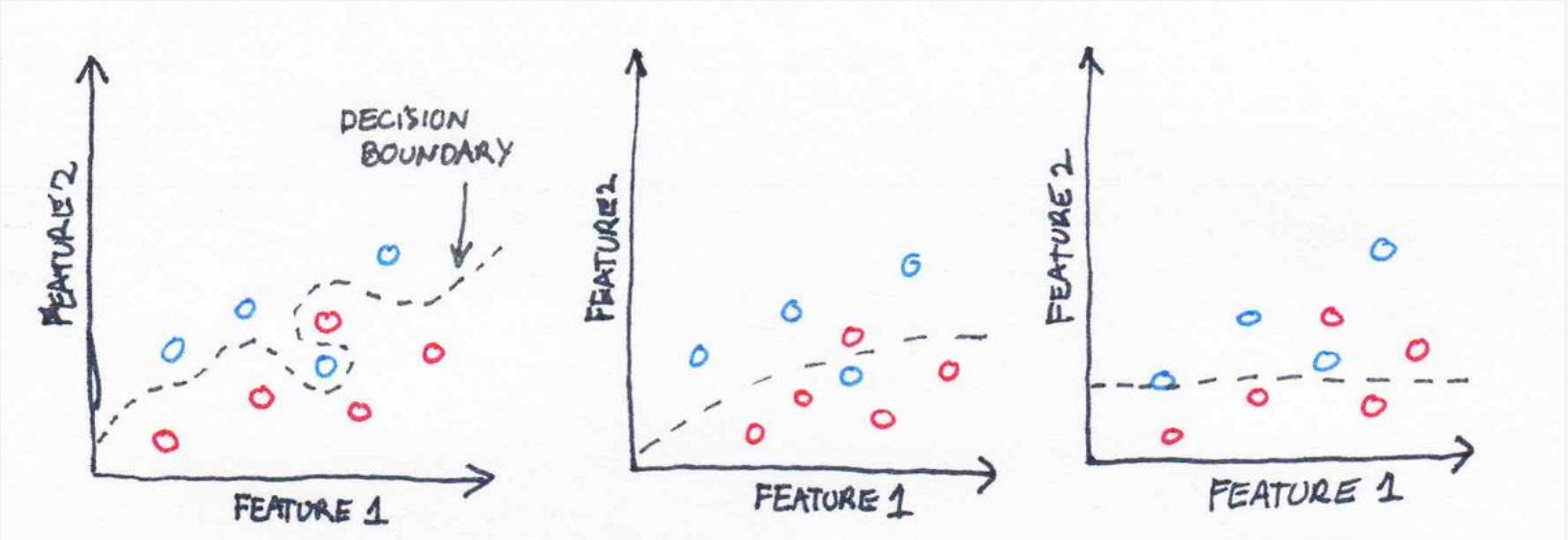
Alfa demasiado grande!

```
history= model.fit(X_train, Y_train,
                   epochs= EPOCAS,
                   batch_size= LOTE,
                   validation_data=(X_test, Y_test))
AAPutils.plot_training_curves(history)
```



Sobreajuste (overfitting)

El *Overfitting* ocurre cuando el modelo es muy complejo e intenta adecuarse perfectamente a los datos de entrenamiento.



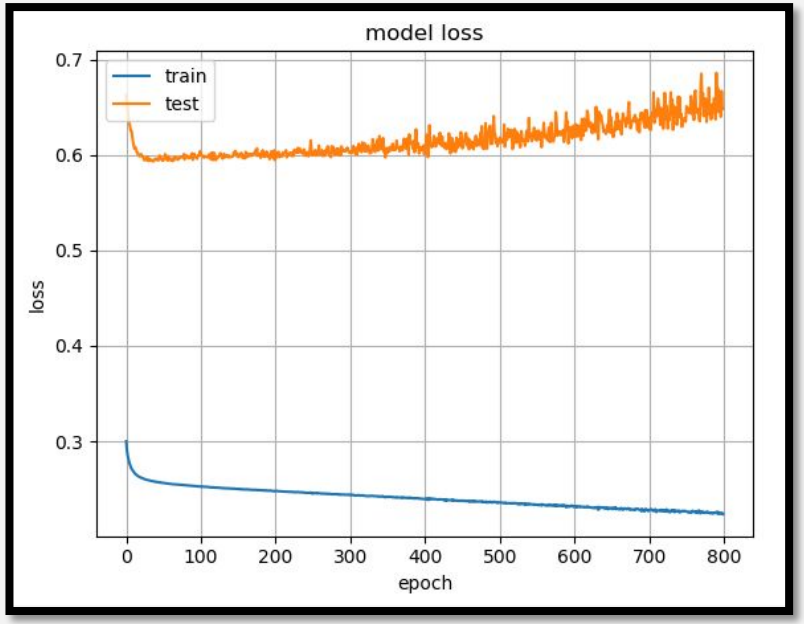
Modelo con Overfitting.

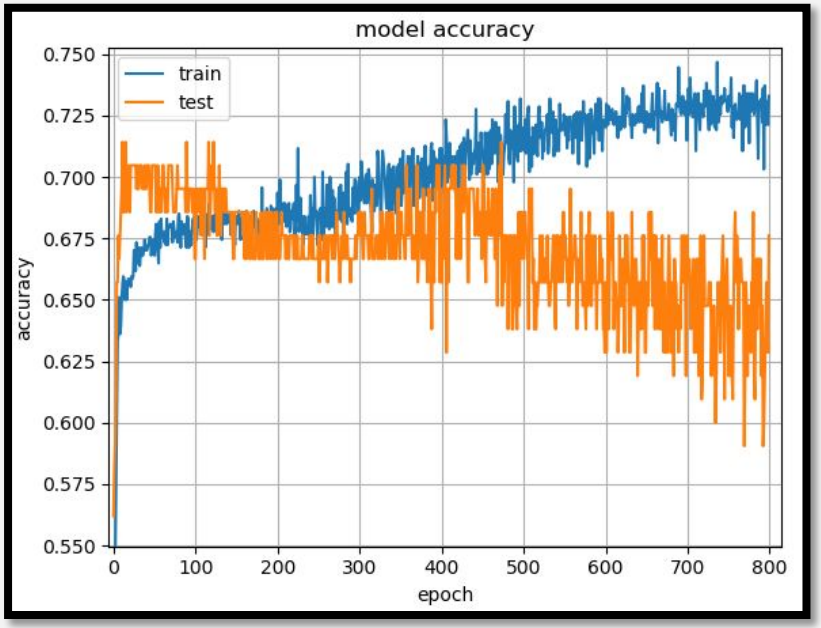
Modelo “correcto”.

Modelo con UnderFitting.

Overfitting

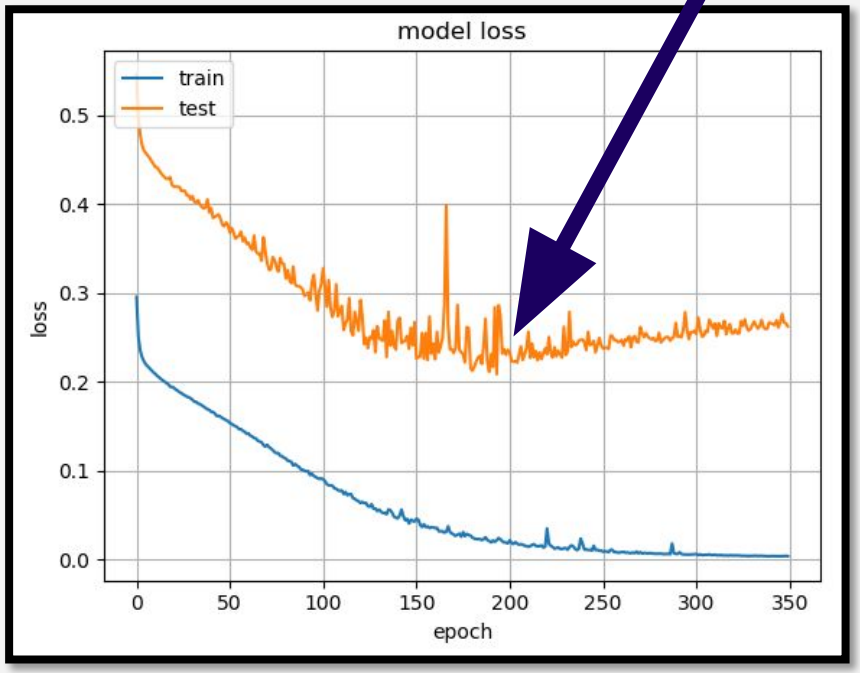
Podemos verificar el Overfitting si el error en Training es bajo y el de testing es alto. Si observamos las curvas de aprendizaje podemos detectar el momento en que comienza el sobreajuste del modelo.



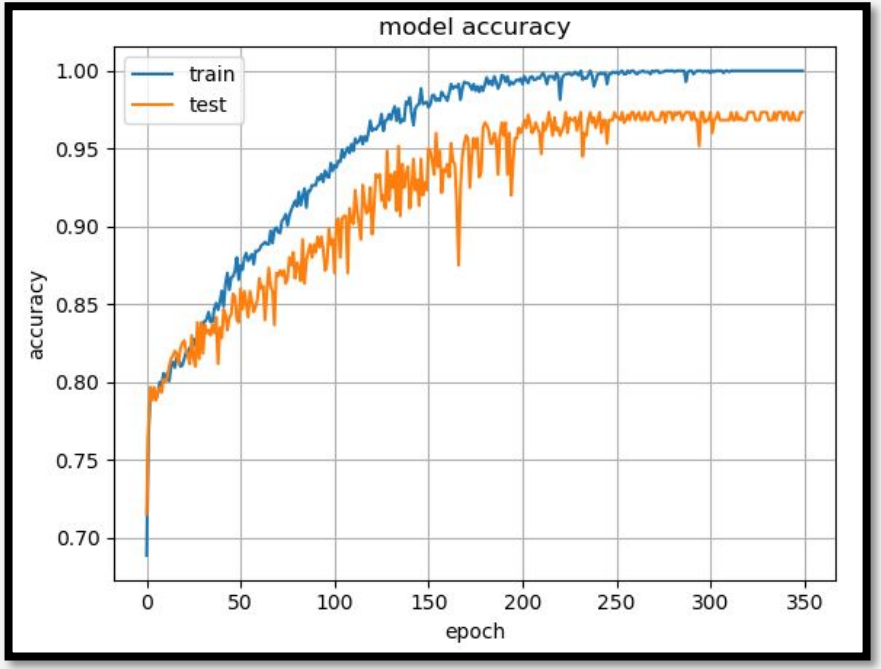


Overfitting

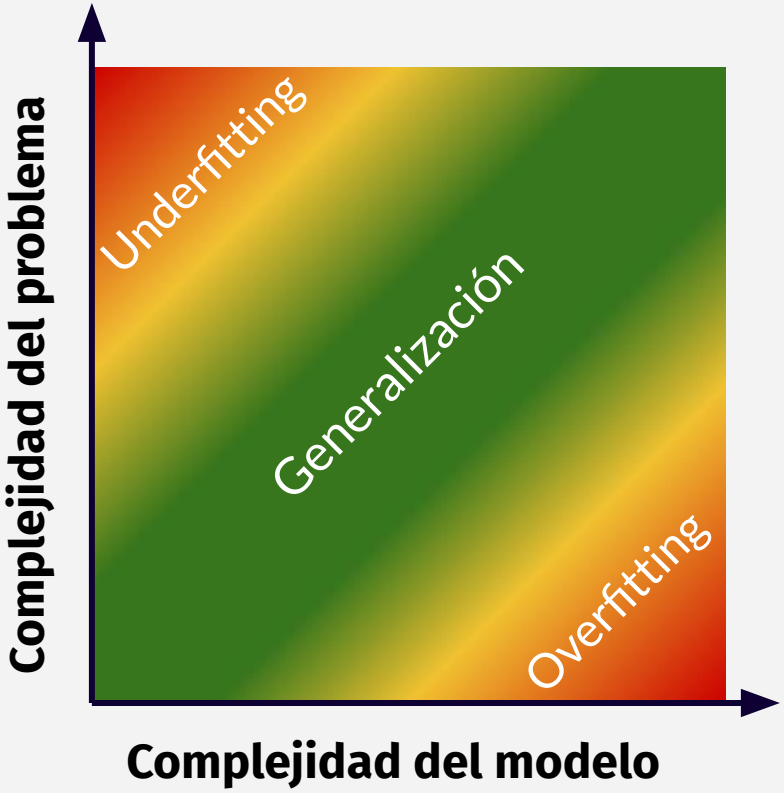
Momento en que el modelo comenzó a Sobreentrenar



En este caso, el accuracy no se vio demasiado afectado.

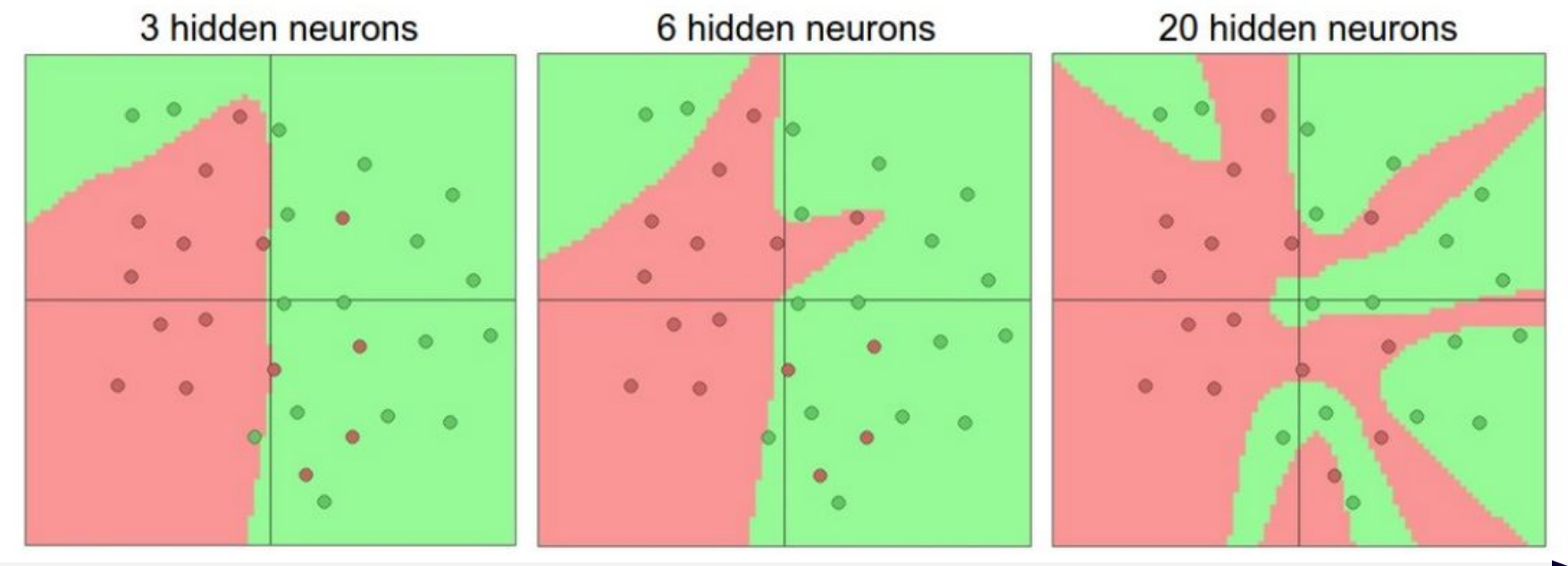


Complejidad del modelo vs problema



- Generalización
  - Capacidad de un modelo para predecir datos no vistos (sin etiquetar).
  - Se mide con Testing set
- Overfitting
  - Predice bien train
  - No generaliza en test
- Underfitting
  - Predice mal train

Complejidad del modelo vs problema



Hiperparámetros

Los hiperparámetros son los que guían el algoritmo de optimización de nuestro modelo.
No existe una configuración ideal, sino que depende del modelo y del dataset específico.

Cantidad de capas y cant. de neuronas por capa.

- Fx activación
- Learning rate (alfa)
- Batch size
- # Epochs
- Regularización
- ...

Regularización

La regularización es un modo de restringir la complejidad de un modelo.

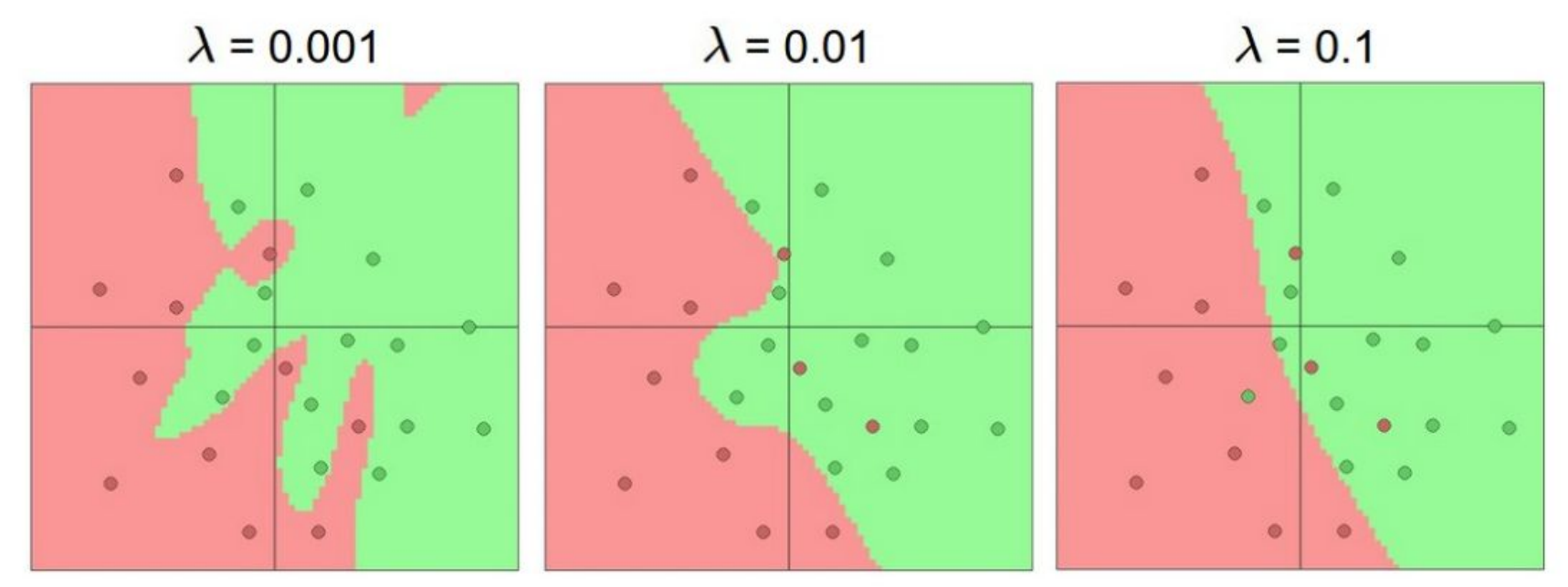
Es útil cuando hay mucha correlación entre features, para filtrar ruido en los datos y también para evitar el Overfitting.

Aprendizaje Loss + Penalización

L2 Regularization:

$$\boldsymbol{Loss} = \underbrace{\frac{1}{n}\sum_{i}^{n} E_i}_{\text{Error plano}} + \underbrace{\lambda \sum_{j}^{m} w_j^2}_{\text{Penalización}}$$

Regularización



Preprocesamiento

¿Cuándo preprocesar los datos?

Ej: **TRAIN - petal size: min =0.1, max= 2**
TEST - petal size: min =0.2, max= 1.5

- **Solo preprocesar datos de entrenamiento**
  - Distintos formatos y rangos

TRAIN - petal size: min =0, max= 1
TEST - petal size: min =0.2, max= 1.5

- **Preprocesar Train y Test por separado**
  - Problemas al tener nuevas distribuciones

TRAIN - petal size: min =0, max= 1
TEST - petal size: min =0, max= 1
(en train 1= 2 y en test 1= 1.5)

- **Preprocesar antes de dividir**
  - Contamina train Set (hacemos trampa)

TRAIN - petal size: min =0, max= 1
TEST - petal size: min =0, max= 1
Ok, pero se entrenó con info de Test set

Preprocesamiento

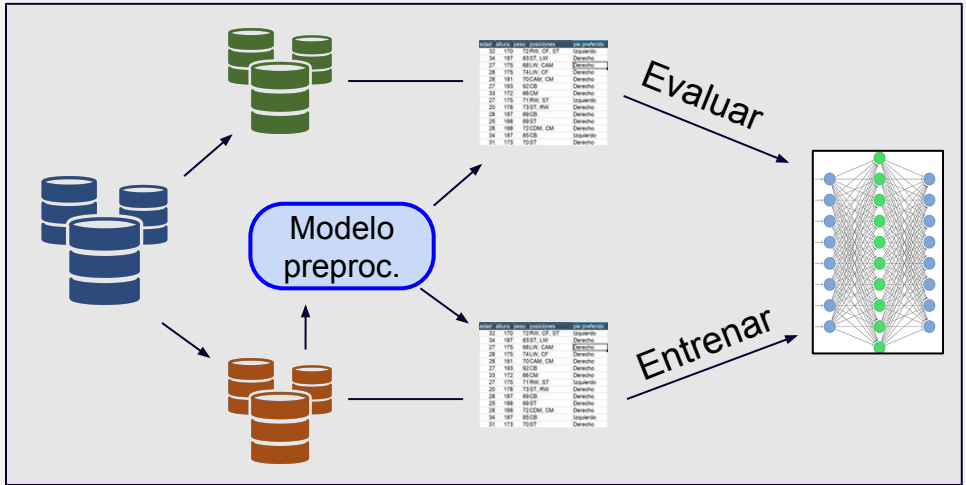
¿Cuándo preprocesar los datos?

- **Modelo de preprocesamiento**
  - Se genera un modelo de datos con Train set
  - Se utiliza el mismo modelo con Test Set (Ej. media y desviación).
  - Otra opción: El modelo de clasificación también preprocesa (Ej. ConvNets)

Ej: **TRAIN - petal size: min =0.1, max= 2**
TEST - petal size: min =0.2, max= 1.5

TRAIN -
Modelo Preproc.= min=0.1, max=2
petal size: min =0, max= 1

Aplicar mismo modelo de preproc.
TEST - petal size: min =0, max= 1



Análisis y visualización del entrenamiento

- **Regresión con webcam**

  https://editor.p5js.org/AndreasRef/sketches/B1g7ds0wm

- **Clasificación o Regresión simple en datasets 2D:**

  https://playground.tensorflow.org/

  https://cs.stanford.edu/people/karpathy/convnetjs/demo/classify2d.html

- **Red entrenada para clasificar imágenes de números:**

  http://scs.ryerson.ca/~aharley/vis/fc/