EDEM
Escuela de Empresarios

EDEM Escuela de Empresarios

Machine Learning 0 - Intro

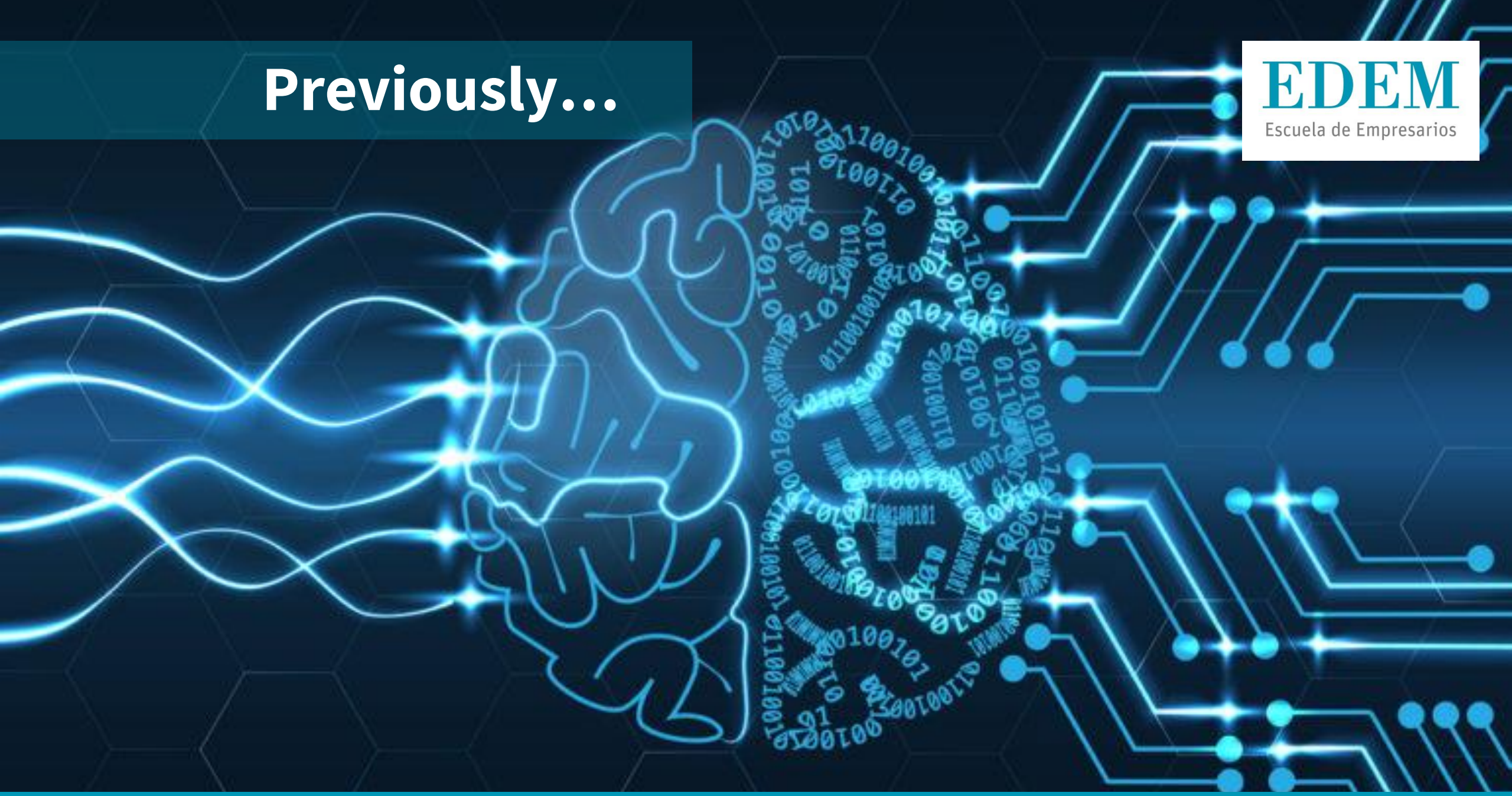
horus·ml

Jesús Prada Alonso - HORUS ML

Curso Máster en Data Analytics - Edición 4

Fecha 09/03/2023

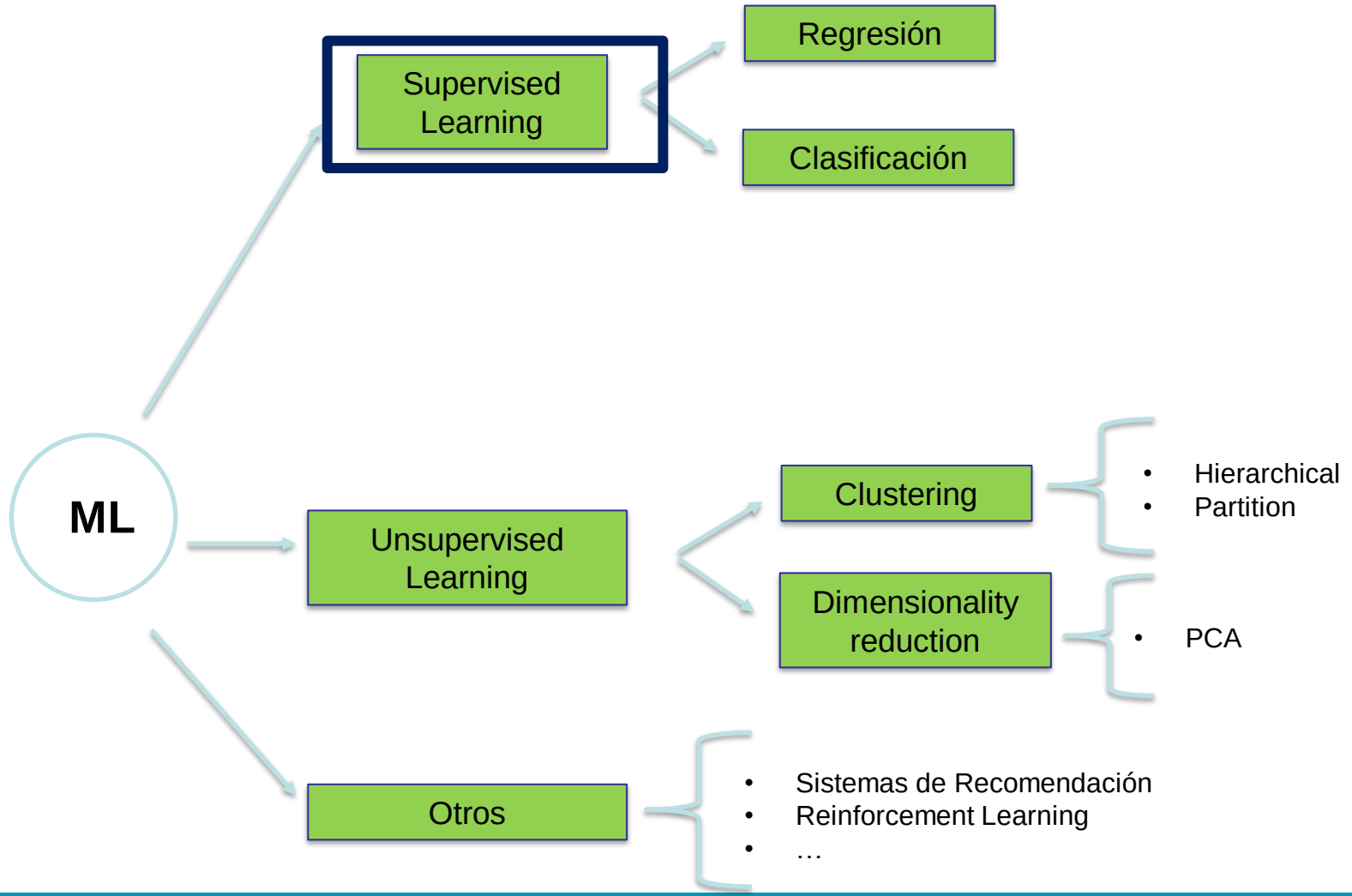
Previously...



EDEM
Escuela de Empresarios

ML SUMMARY

EDEM
Escuela de Empresarios



ML SUPERVISADO

EDEM
Escuela de Empresarios

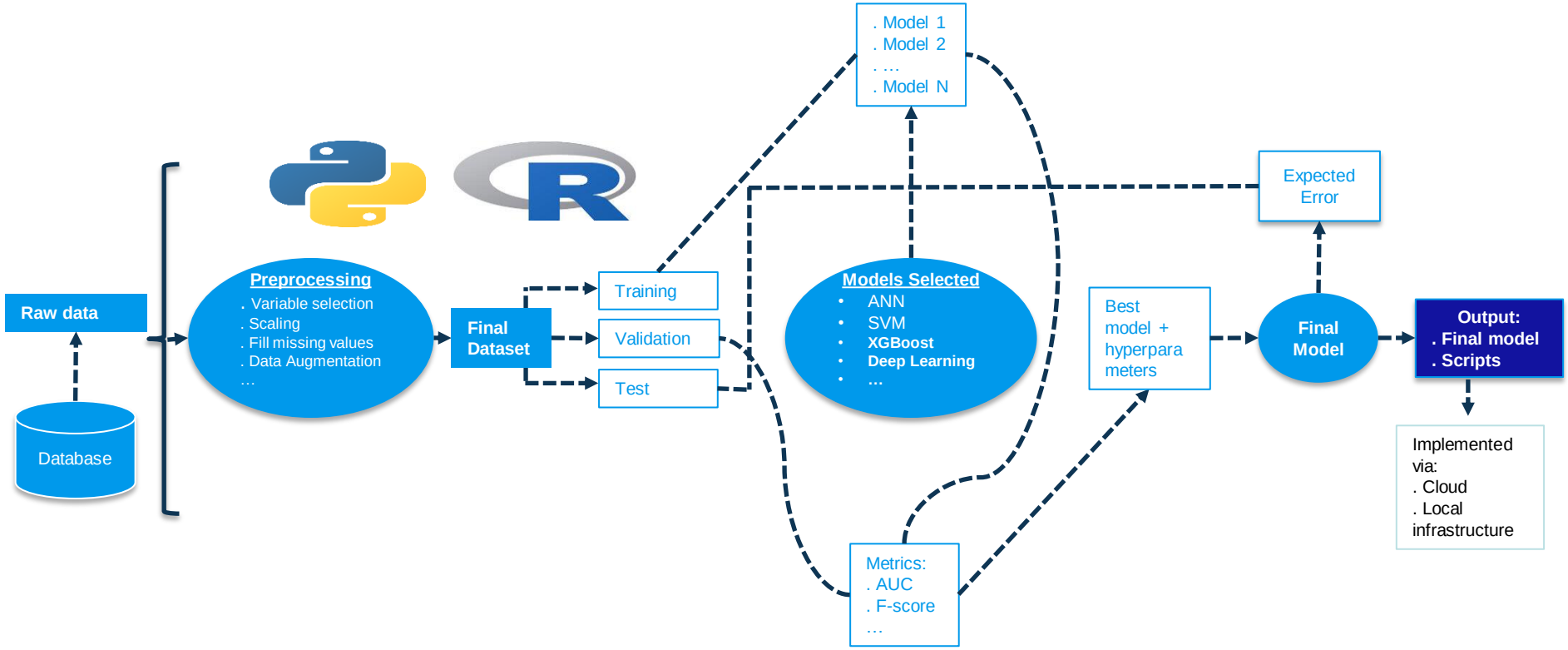
CONCEPTOS ML

EDEM
Escuela de Empresarios

- Conjuntos Train/Validación/Test. Cross-validation.
- Métricas.
- Metaparametrización.
- Trade off bias/variance. Overfitting/Underfitting.

ESQUEMA ML

EDEM
Escuela de Empresarios



EDEM
Escuela de Empresarios

Conjuntos Train/Validación/Test

Tipos de conjuntos

- **Muestra de Entrenamiento (TRAINING):** Datos de los que los modelos extraen patrones. Son los únicos para los que el modelo “ve” el target o etiqueta a predecir.
- **Muestra de Validación (VALIDATION):** Se emplea para seleccionar el mejor de los modelos entrenados cuando realizamos el ajuste de parámetros o metamodelización.
- **Muestra de Prueba (TEST):** Proporciona el error real esperado con el modelo seleccionado.

Consideraciones de los conjuntos de train, validación y test

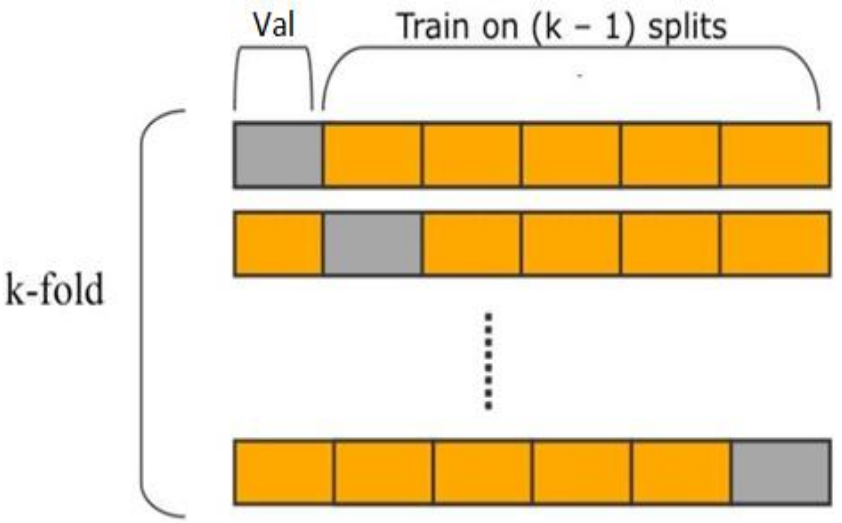
- Que sean lo **suficientemente grandes** como para generar resultados significativos desde el **punto de vista estadístico**.
- Que sean **representativos** de todo el conjunto de datos. Es decir, no elegir un conjunto de prueba con características diferentes (**sesgo**) al del conjunto de entrenamiento.
- **No existe** una solución óptima (**golden rule**) para elegir el porcentaje del total de datos asignado a cada conjunto, ya que depende del problema. Pero un estándar típico es 70/15/15.

EDEM
Escuela de Empresarios

Cross-validation

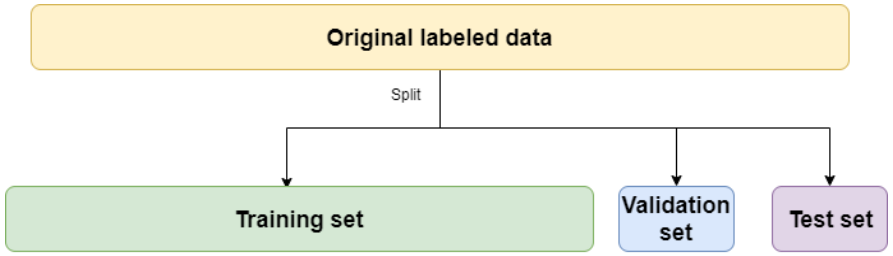
Es un método alternativo a la división en train/val/test para realizar la optimización de hiperparámetros. Permite no tener que crear un conjunto de validación, sustituyendo su funcionalidad por la siguiente metodología:

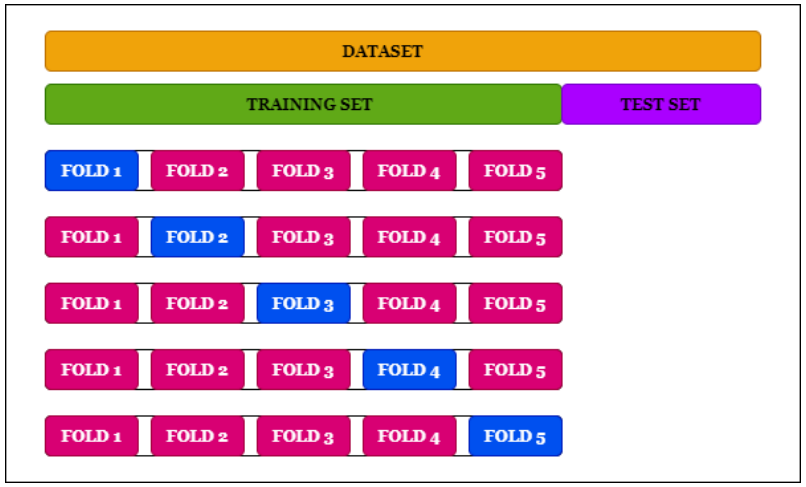
- Se hace una separación del datatest en k subconjuntos del mismo tamaño.
- Se realiza el k -1 conjuntos para entrenar y 1 para validación.
- Se repite el procedimiento k veces rotando el conjunto de validación.
- Se evalúa con la métrica seleccionada.
- Se promedian los errores de validación para obtener el error de cross-validation.



EDEM
Escuela de Empresarios

Validación fija VS. Cross-validation





✓ Ventajas CV	✘ Desventajas CV
Permite aprovechar más volumen del dataset para su uso como train en el entrenamiento de los modelos	Puede tener efectos negativos cuando existe una dimensión temporal en el problema
Implica tener que estimar un porcentaje óptimo para 2 conjuntos en lugar de 3	Es más costoso computacionalmente

EDEM
Escuela de Empresarios

Métricas

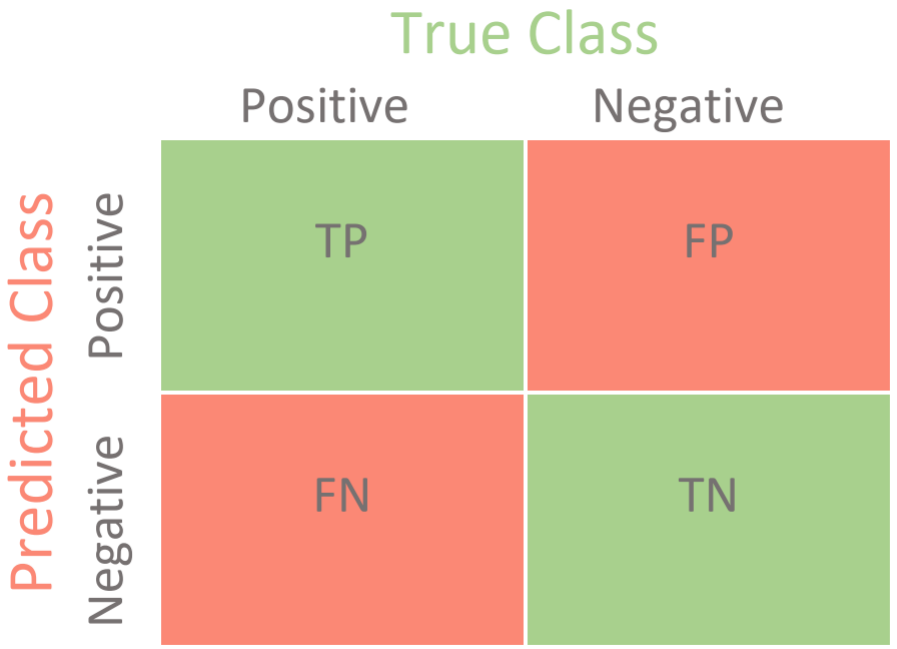
- Para comparar el rendimiento obtenido por cada combinación de tipo de modelos y conjunto de hiperparámetros necesitaremos de un valor numérico que nos informe de su bondad predictiva.
- Este valor numérico vendrá dado por la métrica elegida.
- La elección de esta métrica dependerá del tipo de problema, de los datos y del objetivo a solucionar.
- Dicha elección tiene mucho impacto en el modelo final elegido, por lo que tiene una gran importancia.
- Ejemplo: MAE

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |e_t|, \text{ where } e_t = original_t - predict_t$$

MÉTRICAS (I)

EDEM
Escuela de Empresarios

Clasificación



MÉTRICAS (II)

EDEM
Escuela de Empresarios

Clasificación

- **Accuracy:** Porcentaje de aciertos del modelo.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

- **Sensibilidad, o recall, VPR:** ratio de verdaderos positivos.

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

- **Especifidad, VNR:** ratio de verdaderos negativos.

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

- **Precisión**: probabilidad de que, dada una predicción positiva, la realidad sea positiva también.

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

- **F1-score**: f1-score es una medida que mezcla la precision y el recall. Mide si nuestro modelo tiene falsos positivos y falsos negativos a la vez.

$$F1 = 2 * \frac{precision*sensibilidad}{precision+sensibilidad}$$

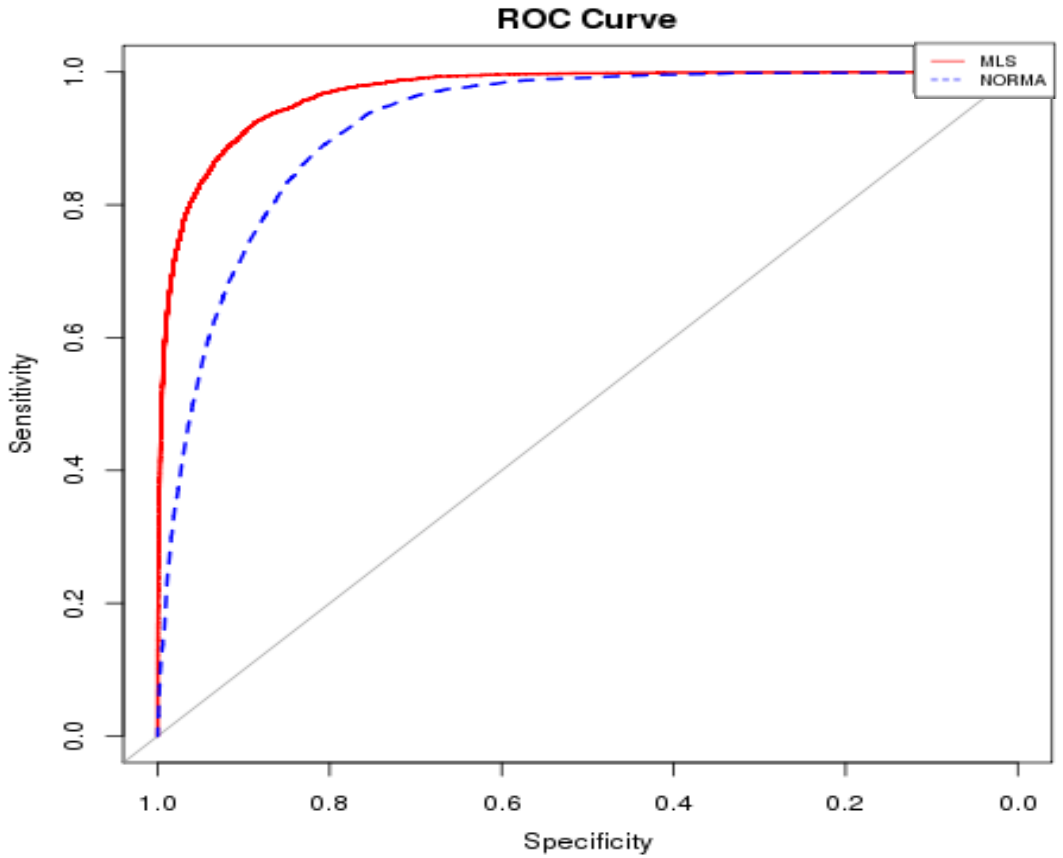
MÉTRICAS (III)

EDEM
Escuela de Empresarios

Clasificación. AUC

Sensibilidad, recall, VPR: ratio de verdaderos positivos.

$$\text{Sensibilidad} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$



Especifidad, VNR: ratio de verdaderos negativos.

$$\text{Especifidad} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

MÉTRICAS (IV)

EDEM
Escuela de Empresarios

Regresión

- **MAE o Error absoluto medio:** es la media de la diferencia absoluta entre los puntos de datos reales y el valor de predicción**.**

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

- **MSE o Error cuadrático medio:** es la media de la diferencia entre los puntos reales de datos y el valor de predicción al cuadrado. Penaliza más las diferencias mayores o extremas.

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **RMSE:** Raíz cuadrada del MSE. Proporciona mayor intuición que el MSE.

$$RMSE = \sqrt{MSE}$$

- **MAPE o Error absoluto porcentual medio:** Permite medir error relativos a la magnitud del valor real.

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{|y_i|}$$

EDEM
Escuela de Empresarios

Metaparametrización (I)

- Los modelos ML suelen incluir un conjunto de **hiperparámetros** que nos permiten controlar su comportamiento.
- De su correcta elección dependerá la bondad del modelo entrenado.
- Los hiperparámetros dependen del perfil de los datos que estamos analizando (**problem-dependent**), por lo que no es sencillo establecer un procedimiento estándar para su obtención.

EDEM
Escuela de Empresarios

Metaparametrización (II). Grid search

1. Elegimos una familia de modelos.
2. Elegimos unos hiperparámetros a optimizar, les llamaremos par1 y par2.
3. Para cada hiperparámetro, elegimos una serie de valores a probar.

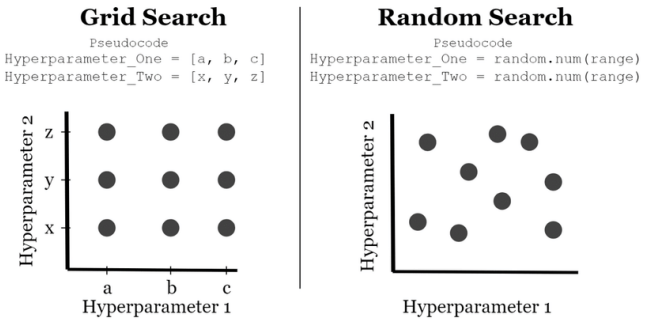
par1/par2	10	100	1000
0.1	0.3	0.22	0.25
0.01	0.15	0.14	0.14
0.001	0.35	0.05	0.11

4. Entrenamos nuestro modelo sobre el conjunto de train con los diferentes hiperparámetros haciendo todas las combinaciones posibles.
5. Hacemos la predicción de los diferentes modelos sobre el conjunto de validación y calculamos el error con la métrica seleccionada.
6. Escogemos el que mejor métrica obtenga y lo aplicamos sobre el conjunto de test para ver el error final esperado de nuestro modelo.

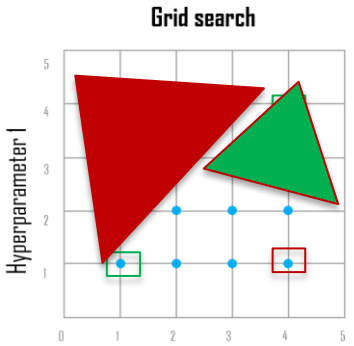
EDEM
Escuela de Empresarios

Variantes Grid search

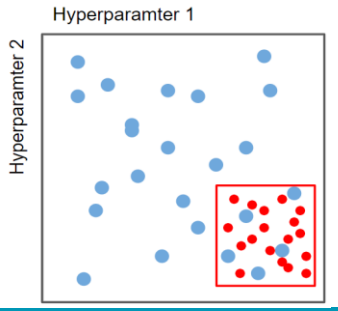
Random Grid Search



Heuristic Grid Search



Zoom-in Grid Search

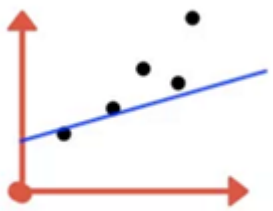


EDEM
Escuela de Empresarios

Trade off bias/variance

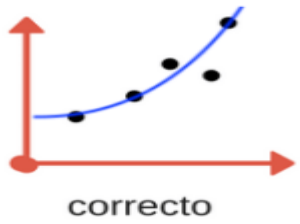
Bias:

- El sesgo es la diferencia entre la predicción promedio de nuestro modelo y el valor correcto que estamos tratando de predecir.
- El modelo con alto sesgo presta muy poca atención a los datos de entrenamiento y simplifica en exceso el modelo.
- Un modelo muy sesgado siempre da un error alto en los datos de train. **Underfitting**.



Variance:

- Es la variabilidad de las predicciones cuando se introducen datos que difieren entre sí.
- El modelo con alta variación se ajusta mucho a los datos de entrenamiento y no generaliza bien con datos que no ha visto antes.
- Dichos modelos funcionan muy bien con los datos de entrenamiento pero tienen altos índices de error en los datos de prueba. **Overfitting**.



EDEM
Escuela de Empresarios

Overfitting/underfitting (I)

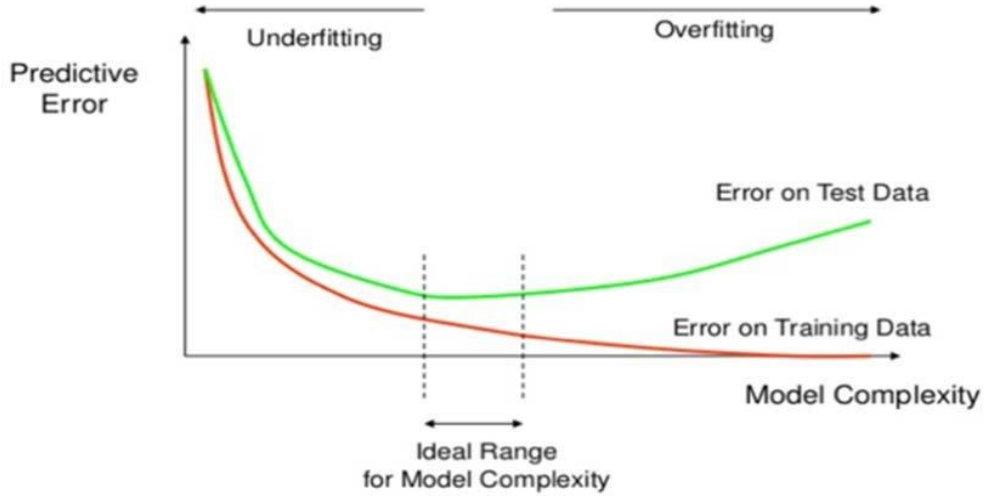
Las principales causas al obtener malos resultados en ML son el overfitting o el underfitting de los datos.

- Si nuestros datos de entrenamiento son muy pocos o nuestro modelo es demasiado sencillo no será capaz de aprender a resolver el problema → **underfitting** (High bias - Low variance).

- Cuando el algoritmo sólo se ajusta a aprender los casos particulares que le enseñamos y es incapaz de reconocer nuevos datos de entrada → **overfittting** (Low bias - High variance)

EDEM
Escuela de Empresarios

Overfitting/underfitting (II)



¿Cómo detectar el overfitting?

Si el modelo entrenado tiene en el conjunto de validación un error mucho mayor que en el conjunto de train, esto sugiere la posibilidad de un problema de overfitting.

¿Cómo detectar el underfitting?

Cuándo el error de train parece demasiado elevado, podemos tener sospechas de overfitting.

También si en el conjunto de validación sólo se acierta un tipo de clase o el único resultado que se obtiene es siempre el mismo valor, o valores similares.

EDEM
Escuela de Empresarios

Overfitting/underfitting (III)

Prevenir el overfitting

- **Cantidad mínima de muestras** tanto para entrenar el modelo como para validarlo.
- **Clases variadas y equilibradas** en cantidad: es importante que los datos de entrenamiento estén balanceados.
- **Conjunto de validación.** Subdividir el conjunto de datos y mantener una porción del mismo «oculto» al modelo.
- Parameter Tunning o **Ajuste de Parámetros**: deberemos experimentar con distintas configuraciones hasta encontrar el equilibrio.
- A veces conviene eliminar o reducir la cantidad de características que utilizaremos para entrenar el modelo, por ejemplo cuando se tiene una cantidad excesiva de dimensiones (features), con muchas variantes distintas, sin suficientes muestras. Una herramienta útil para hacerlo es PCA.

Prevenir el underfitting

- Entrenar un **modelo más complejo.**
- **Aumentar el número de variables** en el dataset**.**

CLASIFICACIÓN

PROBABILIDADES CALIBRADAS

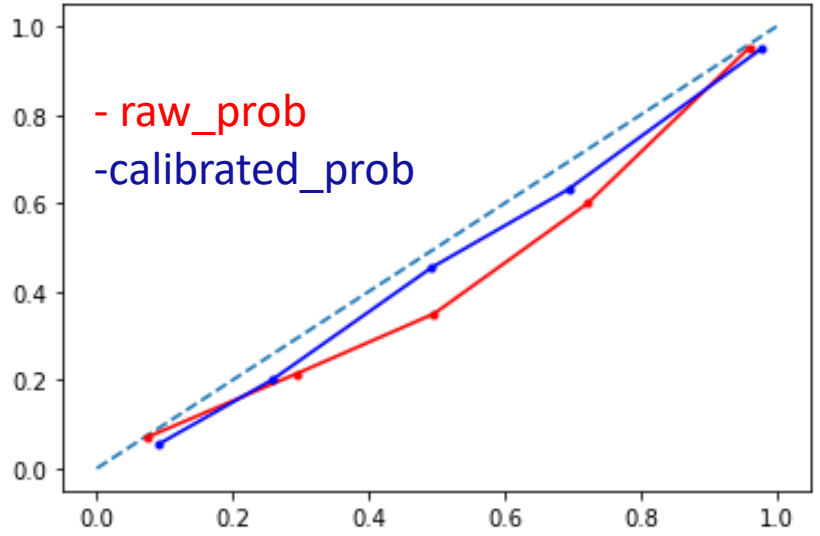
EDEM
Escuela de Empresarios

- Los modelos de **clasificación dan una puntuación** numérica**, pero no todos** son **una probabilidad** verdadera **en un sentido de frecuencia.**

- En estos casos, este valor numérico se puede utilizar para clasificar las predicciones y, por lo tanto, la puntuación AUC no se verá afectada, pero no permite utilizar esta puntuación numérica directamente como probabilidad.

- Estas puntuaciones numéricas se denominan **probabilidades no calibradas**.

- Puede **transformarse en una probabilidad real**, o algo cercano, mediante un proceso llamado **calibración**.



MULTICLASIFICACIÓN

EDEM
Escuela de Empresarios

- La **clasificación** se puede aplicar a **más de dos clases (k > 2). Dos modificaciones** necesarias**:**

- **Modelo:**

    - La regresión logística tiene una formulación especial para este caso**.**

    - **Para otros modelos,** un enfoque común es el **enfoque de uno contra todos,** donde se entrenan k modelos binarios, uno para cada clase.

    - Cada modelo dará la probabilidad de pertenecer a esa clase. **QUIZ!**

- **Métricas:**

    - Las **métricas** de clasificación que vimos se basan en valores de matriz de confusión: TP, TN, FP, FN.

    - Todos estos conceptos son **válidos sólo para la clasificación binaria.**

    - El enfoque estándar es calcular una **matriz de confusión para cada clase y** resultados métricos **promedio** en todas las clases.

EJEMPLO

EDEM
Escuela de Empresarios

https://ml-playground.com/#

RESUMEN

EDEM
Escuela de Empresarios

RESUMEN (I)

EDEM
Escuela de Empresarios

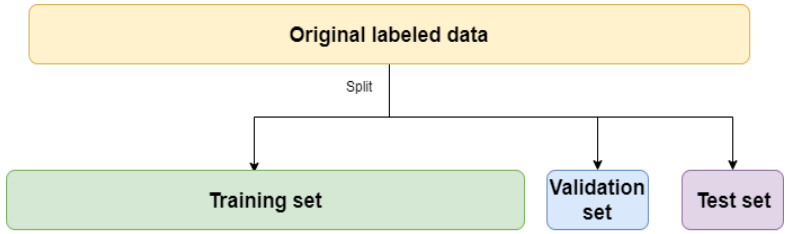
Supervisado VS no Supervisado

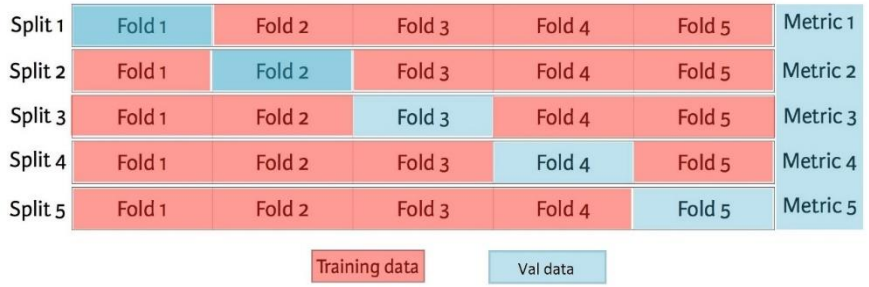
	Supervisado	No supervisado
Etiquetas	SI	NO
Objetivo	Dar predicciones a futuro sobre el conjunto de test	Encontrar patrones en los datos o reducir dimensiones
Modelos	Regresión lineal, árboles, SVM, Redes Neuronales	Clustering, PCA
Ejemplo	Predecir si una transacción es fraudulenta	Encontrar clientes con perfiles similares

Clasificación VS Regresión

	Clasificación	Regresión
Etiquetas	Categóricas.	Numéricas.
Ejemplo	Una imagen es un gato (1) o no (0). 	Precio de alquiler de una casa 
Métrica	AUC	MSE

Cross Validation y validación fija





- TRAINING : Datos de los que los modelos extraen patrones.
- VALIDATION : Se emplea para seleccionar el mejor de los modelos entrenados en metamodelización.
- TEST : Proporciona el error real esperado con el modelo seleccionado.

RESUMEN (II)

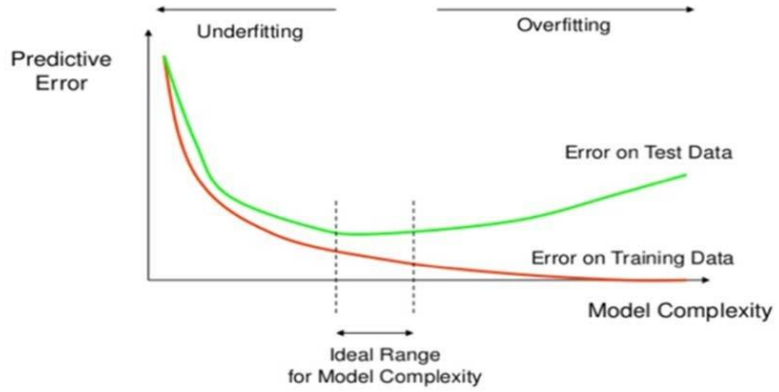
EDEM
Escuela de Empresarios

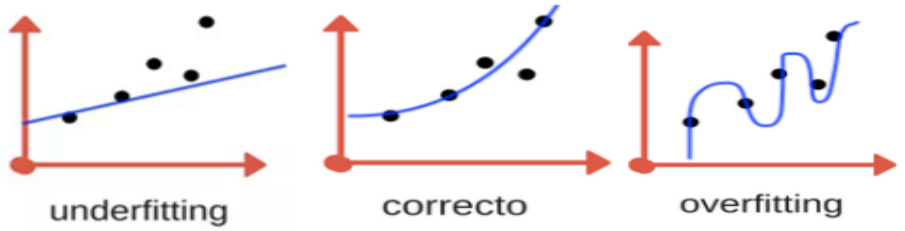
Grid Search

- Los modelos ML suelen incluir un conjunto de **hiperparámetros** que nos permiten controlar su comportamiento.
- De su correcta elección dependerá la bondad del modelo entrenado.

par1/par2	10	100	1000
0.1	0.3	0.22	0.25
0.01	0.15	0.14	0.14
0.001	0.35	0.05	0.11

Overfitting y Underfitting





¿Cómo detectar el overfitting?

Validación tiene un error mucho mayor que en train

¿Cómo detectar el underfitting?

El error de train parece demasiado elevado o da la misma respuesta siempre.

EDEM
Escuela de Empresarios

EDEM Escuela de Empresarios

Jesús Prada Alonso
jesus.prada@horusml.com