

An abstract network diagram consisting of numerous teal-colored nodes connected by thin teal lines. The nodes are scattered across the left side of the slide, with some forming small clusters and others standing alone. The lines vary in length and orientation, creating a complex web-like structure.

Técnicas básicas de Machine Learning

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Técnicas básicas de Machine Learning



Entrenamiento

Ajustar un modelo a partir de un conjunto de datos que permita realizar predicciones (regresión y clasificación) o extraer patrones (clustering).

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Métodos paramétricos y no paramétricos

Recordemos que el objetivo es estimar una función f que exprese la relación entre la salida y el vector de atributos. Esta relación funcional se puede obtener usando dos tipos de métodos estadísticos:

- Métodos paramétricos
- Métodos no paramétricos

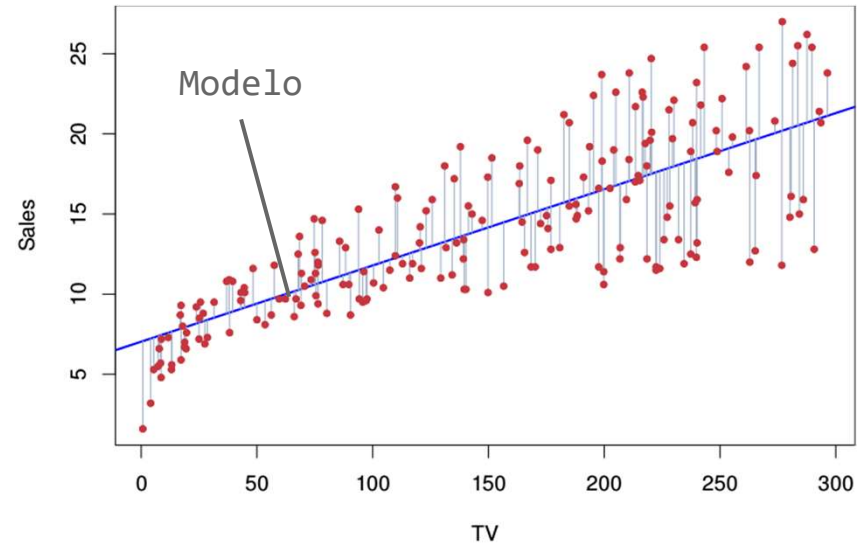
$$y \approx f(\mathbf{X})$$

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Métodos paramétricos y no paramétricos

Métodos Paramétricos

Asumen a priori que la función f tiene una forma determinada. Luego ajustan ciertos parámetros para acercar la función estimada a los datos de entrenamiento.



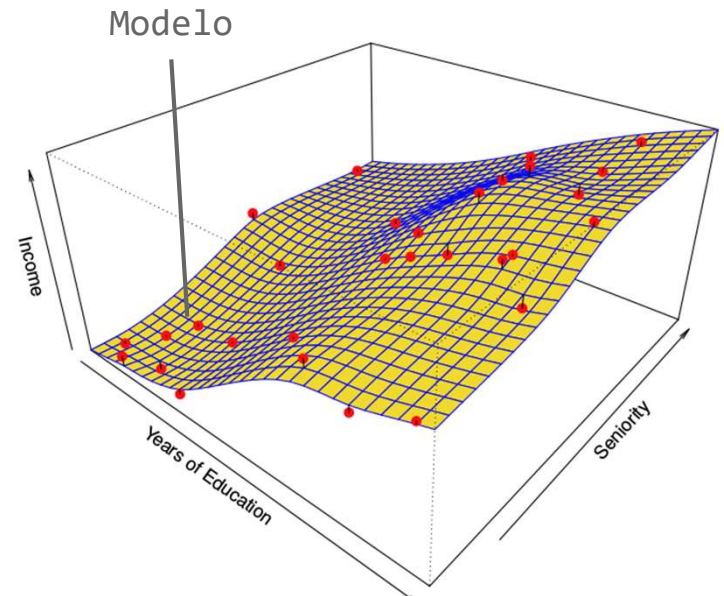
En este caso el método es paramétrico porque se asume que la función tiene una forma lineal, con lo cual el problema se reduce a estimar el valor de los parámetros que ajustan a los datos de entrenamiento a la hipótesis.

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Métodos paramétricos y no paramétricos

Métodos No Paramétricos

No asumen a priori la forma de la función f , sino que buscan ajustar la forma de la función lo mejor posible a los datos de entrenamiento.



En este caso, el método no asume que la función tenga una forma determinada, sino que busca ajustar la función a los puntos de los datos de entrenamiento.

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Algoritmos y técnicas básicas

Supervisado

Regresión

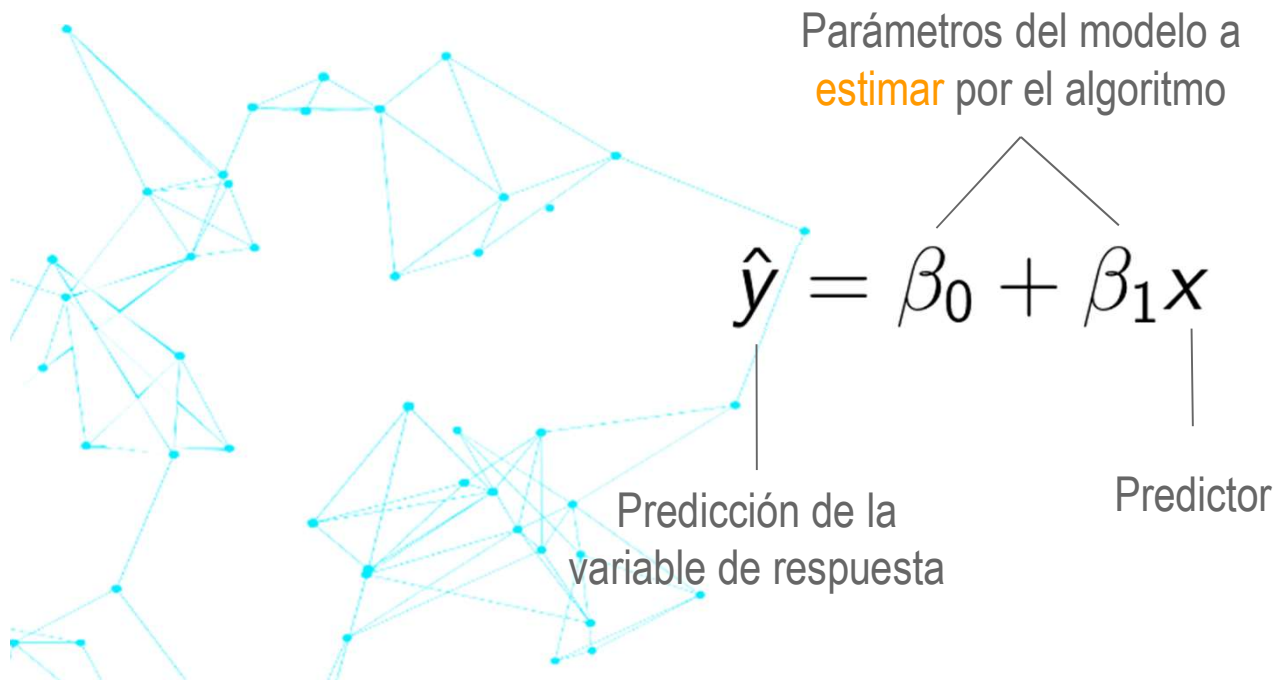
Paramétrico

Lineal

Regresión lineal

Es un modelo estadístico que permite predecir el valor de una variable **cuantitativa** (numérica), como una función lineal de las variables de entrada o predictores.

El modelo...



INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Algoritmos y técnicas básicas

Supervisado

Regresión

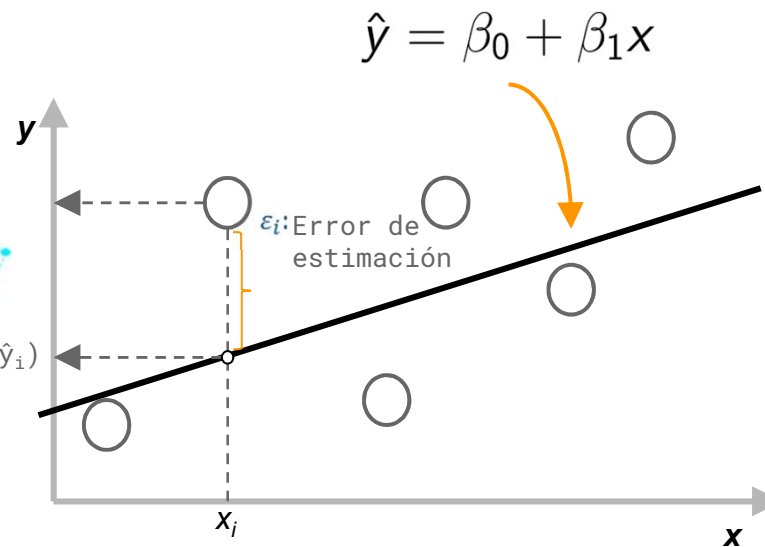
Paramétrico

Lineal

Regresión lineal

Es un modelo estadístico que permite predecir el valor de una variable **cuantitativa** (numérica), como una función lineal de las variables de entrada o predictores.

... de manera gráfica



INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Algoritmos y técnicas básicas

Supervisado

Regresión

Paramétrico

Lineal

Regresión lineal

Para estimar los coeficientes utilizamos los datos del dataset de entrenamiento. Representemos los datos de entrenamiento como n observaciones de x y y :

... el algoritmo

Mínimos cuadrados

A partir de los datos se calculan los parámetros del modelo que minimizan la suma de errores cuadráticos (RSS)

n muestras de
entrenamiento

x	y
x_1	y_1
x_2	y_2
\vdots	\vdots

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Algoritmos y técnicas básicas

Supervisado

Clasificación

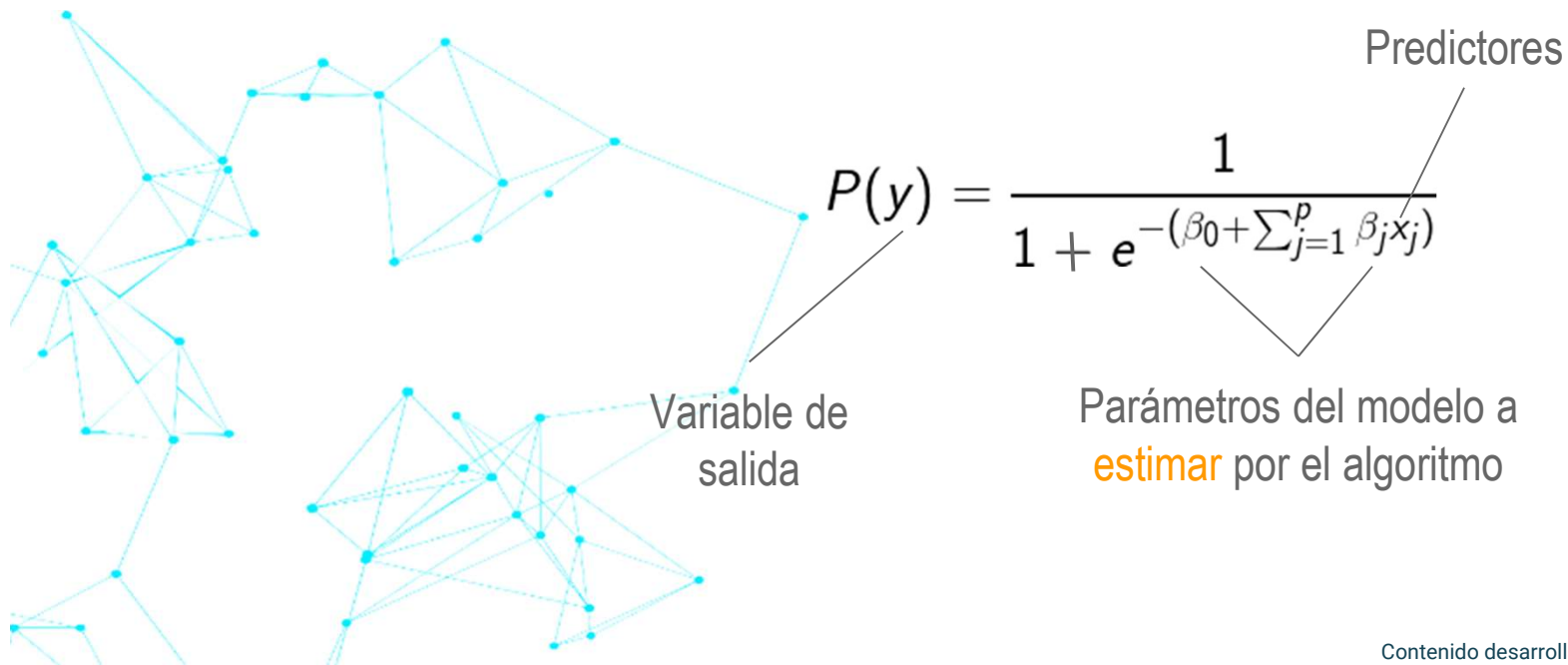
Paramétrico

Lineal

Regresión logística

Es un modelo de regresión generalizado utilizado como método de clasificación binaria, puesto que en lugar de valores numéricos, éste permite estimar la probabilidad de que ocurra (o no) un evento como función de otras variables.

El modelo...



INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Algoritmos y técnicas básicas

Supervisado

Clasificación

Paramétrico

Lineal

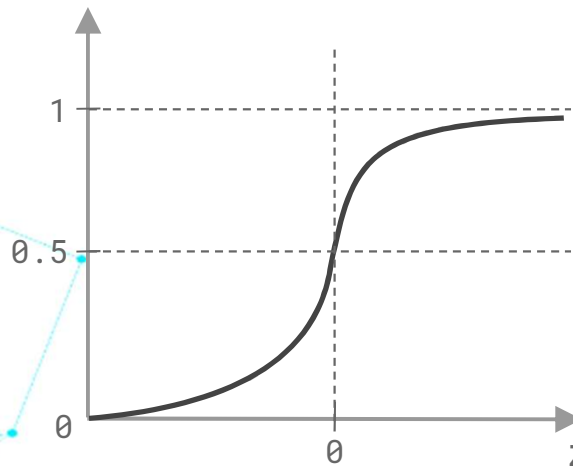
Regresión logística

En lugar de utilizar la función de regresión lineal, vamos a utilizar la función sigmoide o logística.

... de manera gráfica

$$g(z) = \frac{1}{1 + e^{-z}}$$

$P(y) = g(z)$



si $z < 0$, entonces
 $g(z) < 0.5$ por lo que
 $y = 0$

si $z \geq 0$, entonces
 $g(z) \geq 0.5$ por lo que
 $y = 1$

Podemos utilizar esta función para establecer un umbral de decisión, p. ej. en 0.5, que nos permita fijar el valor de y a 0 o 1.

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Algoritmos y técnicas básicas

Supervisado

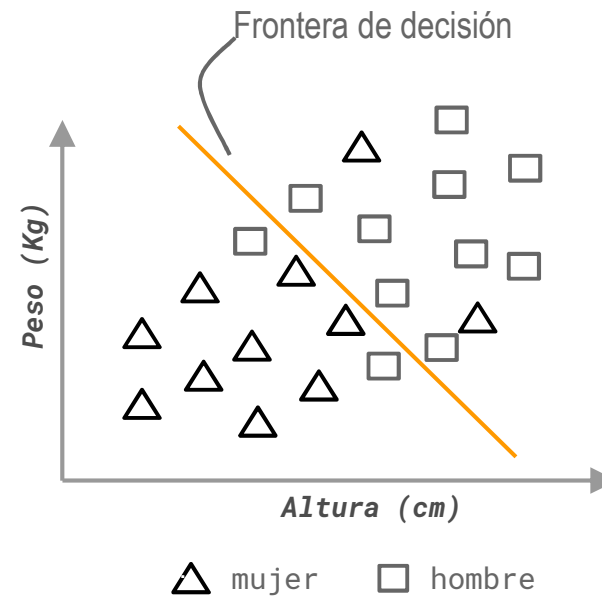
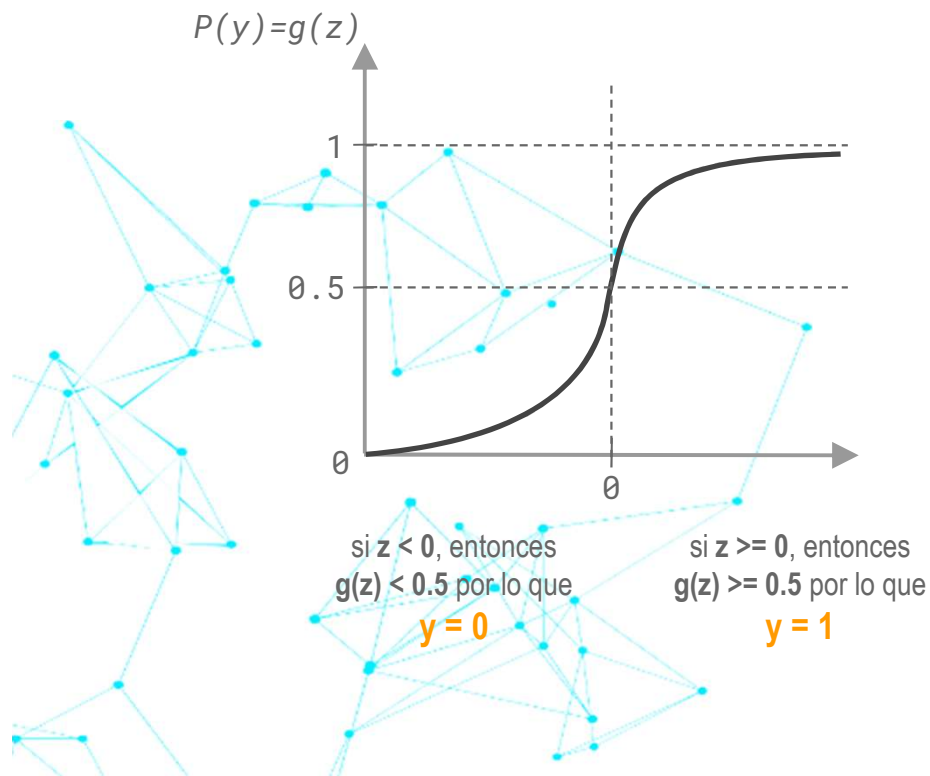
Clasificación

Paramétrico

Lineal

Regresión logística

La función logística nos permite establecer un umbral de decisión que separe las dos clases:



INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Algoritmos y técnicas básicas

Supervisado

Clasificación

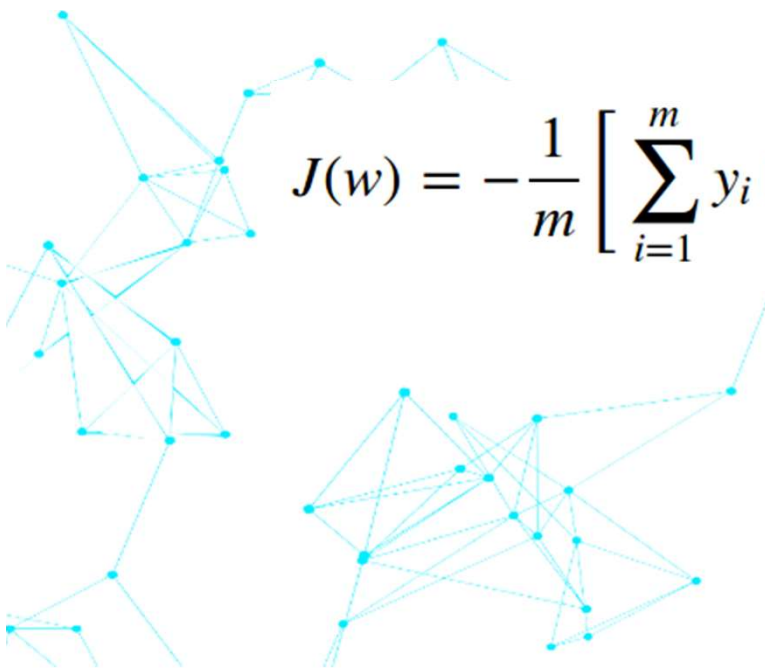
Paramétrico

Lineal

Regresión logística

Al igual que la regresión lineal el algoritmo estima el valor óptimo de los coeficientes de la función logística minimizando una función de coste.

... el algoritmo


$$J(w) = -\frac{1}{m} \left[\sum_{i=1}^m y_i \log(h_w(x_i)) + (1 - y_i) \log(1 - h_w(x_i)) \right]$$

Función de coste

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Algoritmos y técnicas básicas

Supervisado

Regresión

Clasificación

No Paramétrico

No Lineal

Métodos basados en árboles

- Los métodos basados en árboles consisten en segmentar el espacio de predictores en varias regiones.
- Dentro de cada región, se utiliza la media o la moda de las observaciones de entrenamiento en esa región para hacer la predicción.
- Se dice que son **métodos basados en árboles** porque las reglas que se utilizan para dividir el espacio de predictores pueden ser representadas en forma de diagrama de árbol.
- El método más sencillo es el árbol de decisión básico. Luego existen otros métodos como bagging, random forest y boosting que están basados en el árbol básico pero que mejoran la precisión de este modelo.

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Algoritmos y técnicas básicas

Supervisado

Regresión

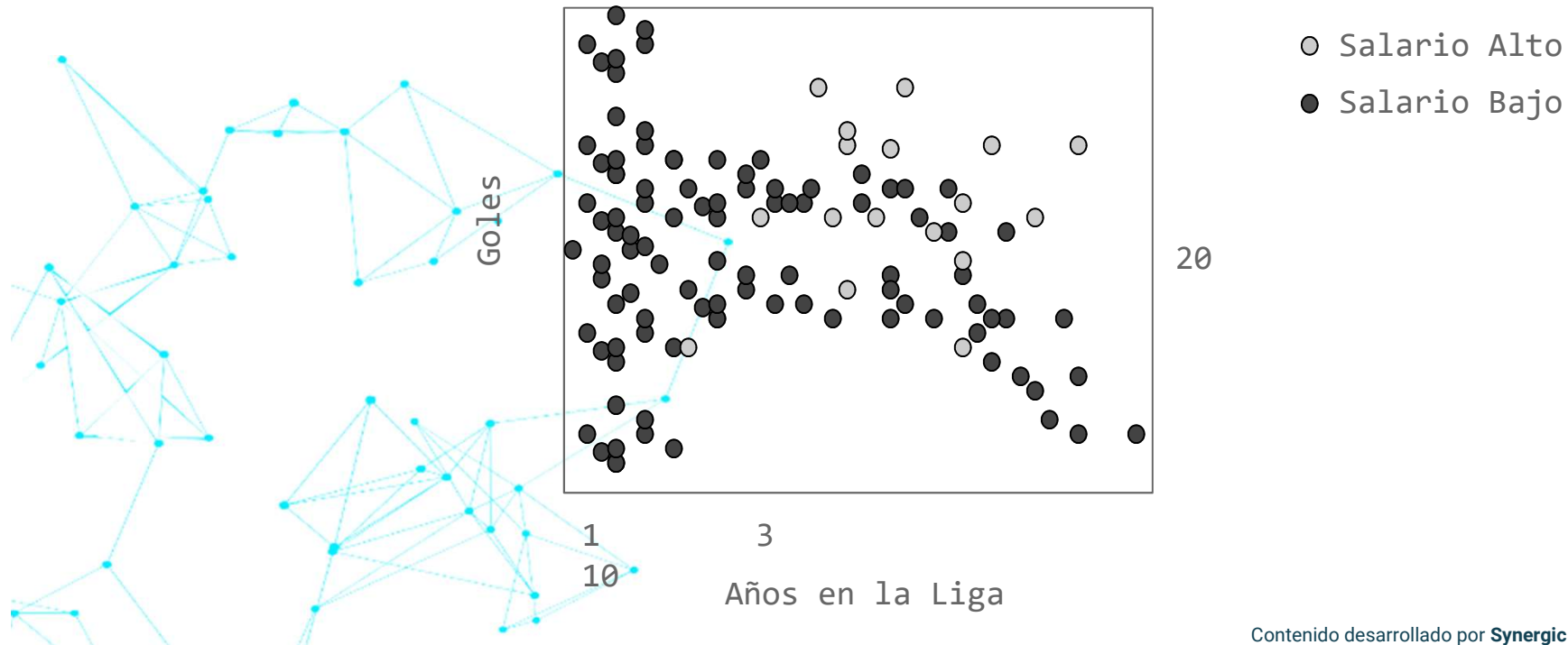
Clasificación

No Paramétrico

No Lineal

Métodos basados en árboles

Para entender cómo funciona un árbol de decisión observemos el siguiente ejemplo. Se trata de predecir el salario de un jugador de futbol por la cantidad de años que lleva jugando en la liga y por la cantidad de goles que marcó en las últimas temporadas:



INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Algoritmos y técnicas básicas

Supervisado

Regresión

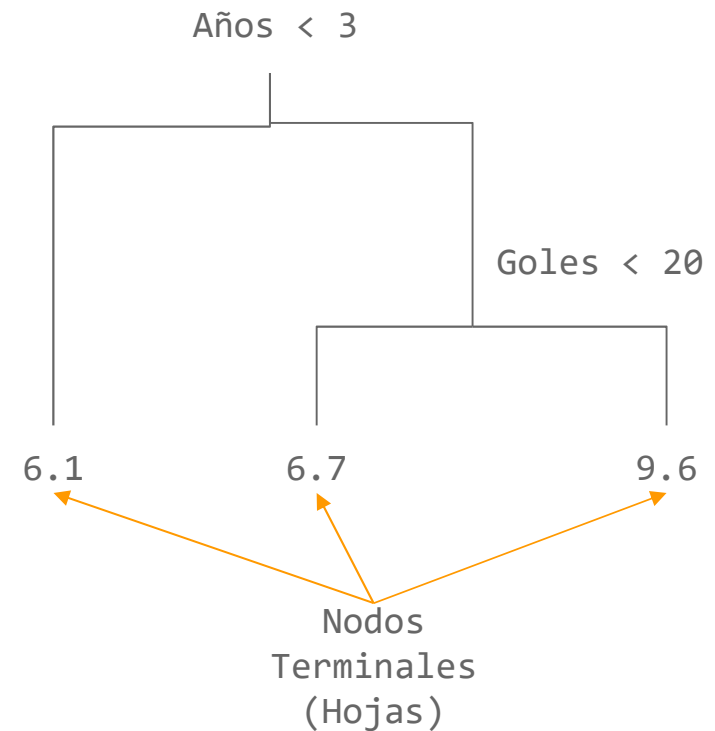
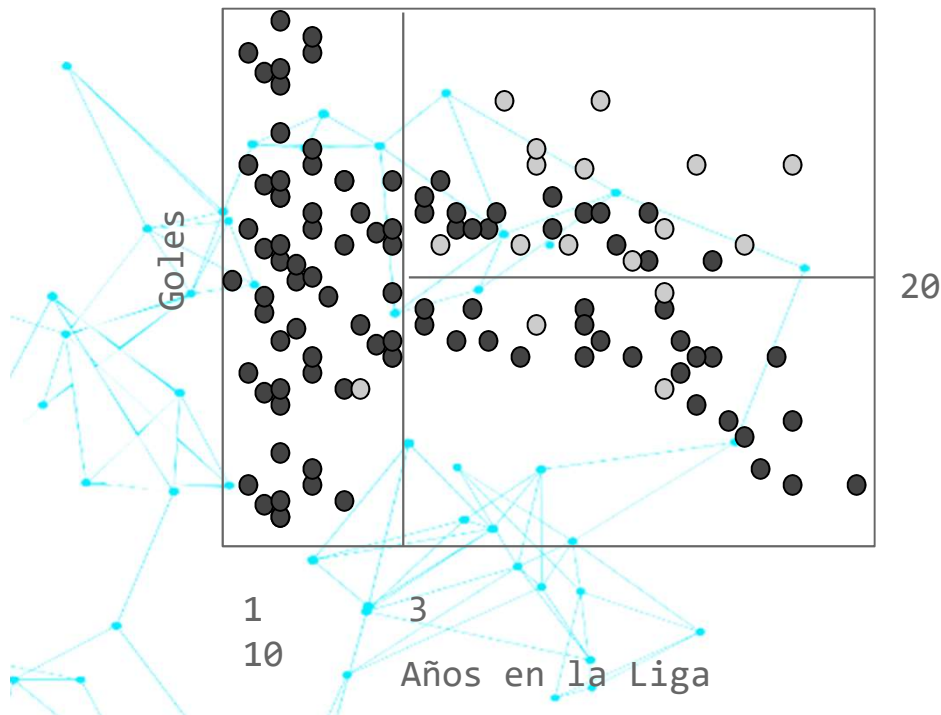
Clasificación

No Paramétrico

No Lineal

Métodos basados en árboles

Utilizando un árbol de decisión obtenemos un conjunto de reglas que nos permite predecir el salario del futbolista.



Contenido desarrollado por Synergic Partners

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Algoritmos y técnicas básicas

Supervisado

Regresión

Clasificación

No Paramétrico

No Lineal

Métodos basados en árboles

- Los métodos basados en árboles se pueden utilizar para resolver problemas de **regresión** y **clasificación**.
- El proceso de construcción de un árbol de clasificación es muy similar al proceso de construcción de un árbol de regresión.
- La principal diferencia es la medida que utilizamos para determinar las particiones de las regiones del árbol. En Regresión utilizamos el **RSS**. En clasificación utilizamos medidas estadísticas de pureza como el **índice de Gini** y la **Cross-Entropy**.
- En el árbol de **regresión** predecimos cada observación como la **media** de todas las observaciones pertenecientes a la misma región del árbol. En **clasificación** predecimos la observación con la clase **más común** de la misma región del árbol.

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Algoritmos y técnicas básicas

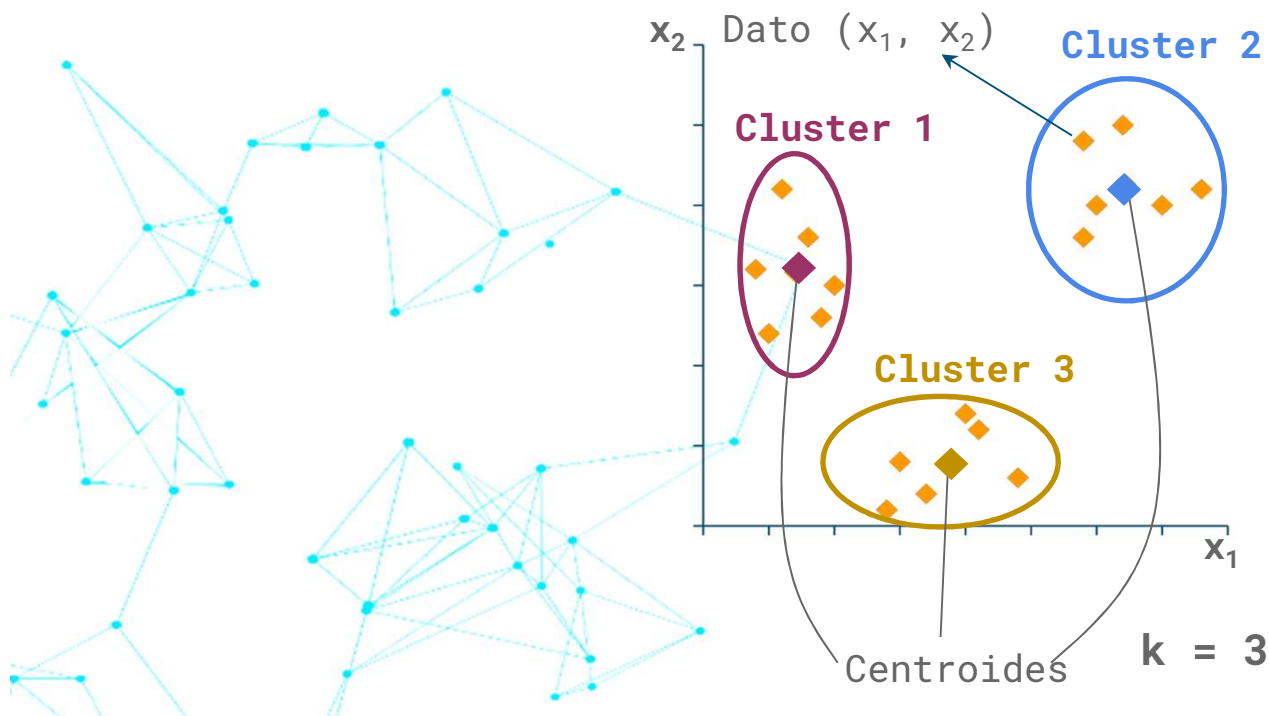
No Supervisado

Clustering

No Paramétrico

K-means

Es un método que permite crear **clusters** de datos numéricos. La entrada del sistema son las mediciones numéricas de interés y la salida son los centroides de los **clusters** resultantes y la asignación de cada dato a un **cluster** determinado.



INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Algoritmos y técnicas básicas

No Supervisado

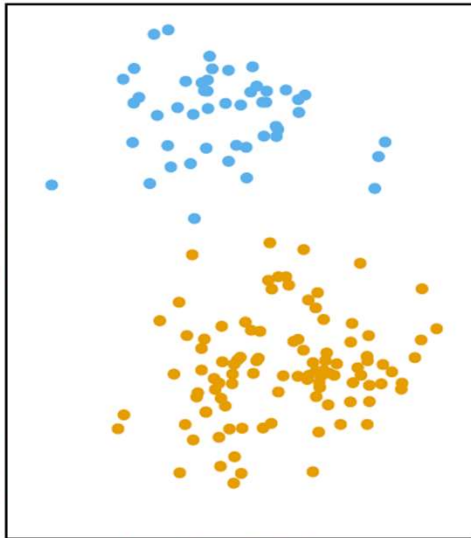
Clustering

No Paramétrico

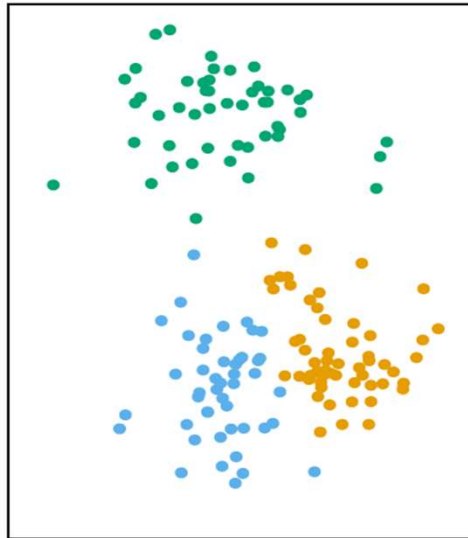
K-means

- El objetivo es agrupar las observaciones de un dataset en un número **K** de clusters.
- El número **K** es dado a priori al algoritmo (hiperparámetro).

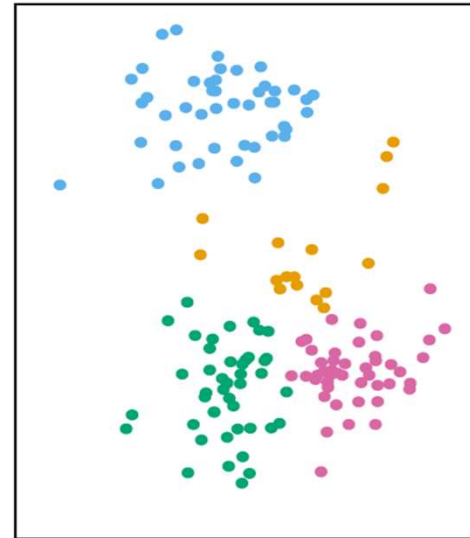
K=2



K=3



K=4



INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Algoritmos y técnicas básicas

No Supervisado

Clustering

No Paramétrico

K-means

En términos generales el algoritmo de K-means se puede resumir en los siguientes pasos:

1. De manera aleatoria asignar un número de 1 a K a cada observación. Esto será la asignación inicial a los cluster de cada observación.
2. Iterar sobre los siguientes pasos hasta que las asignaciones a los cluster deje de cambiar:
 - a. Para cada cluster, calcular el centroide. El centroide será un vector compuesto por la media de los p predictores de las observaciones del mismo cluster.
 - b. Reasigne cada observación al cluster cuyo centroide esté más cercano a la observación.

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Algoritmos y técnicas básicas

Supervisado

No Supervisado

Regresión

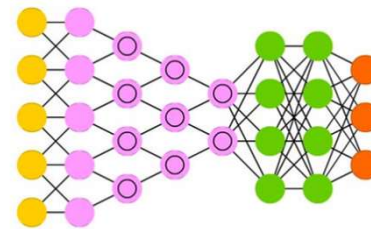
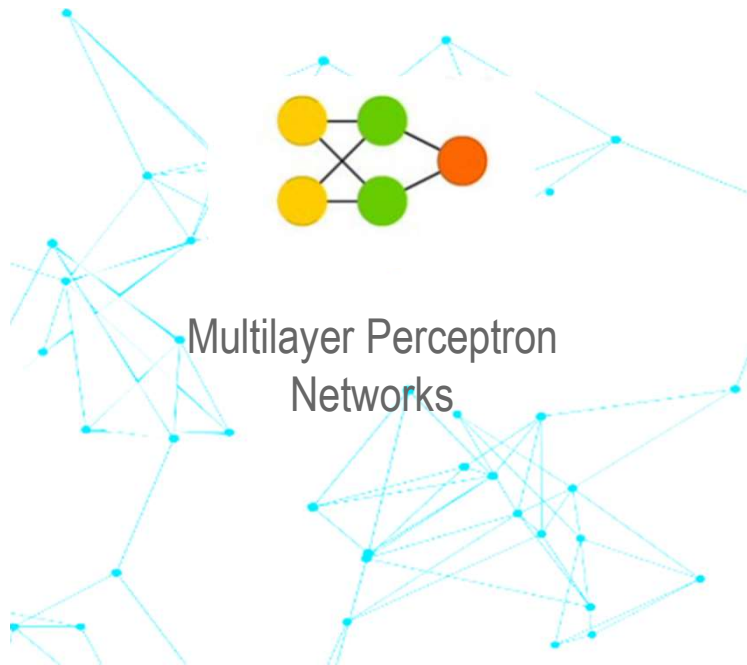
Clasificación

Paramétrico

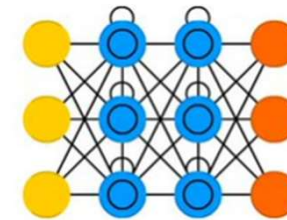
No Lineal

Deep Learning

Son métodos que llevan a cabo el proceso de machine learning usando redes neuronales artificiales compuesta por muchas capas organizadas de forma jerárquica. Algunas técnicas populares son:



Convolutional Neural Networks



Long Short-Term Memory Recurrent Neural Networks

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Algoritmos y técnicas básicas

Supervisado

No Supervisado

Regresión

Clasificación

Paramétrico

No Lineal

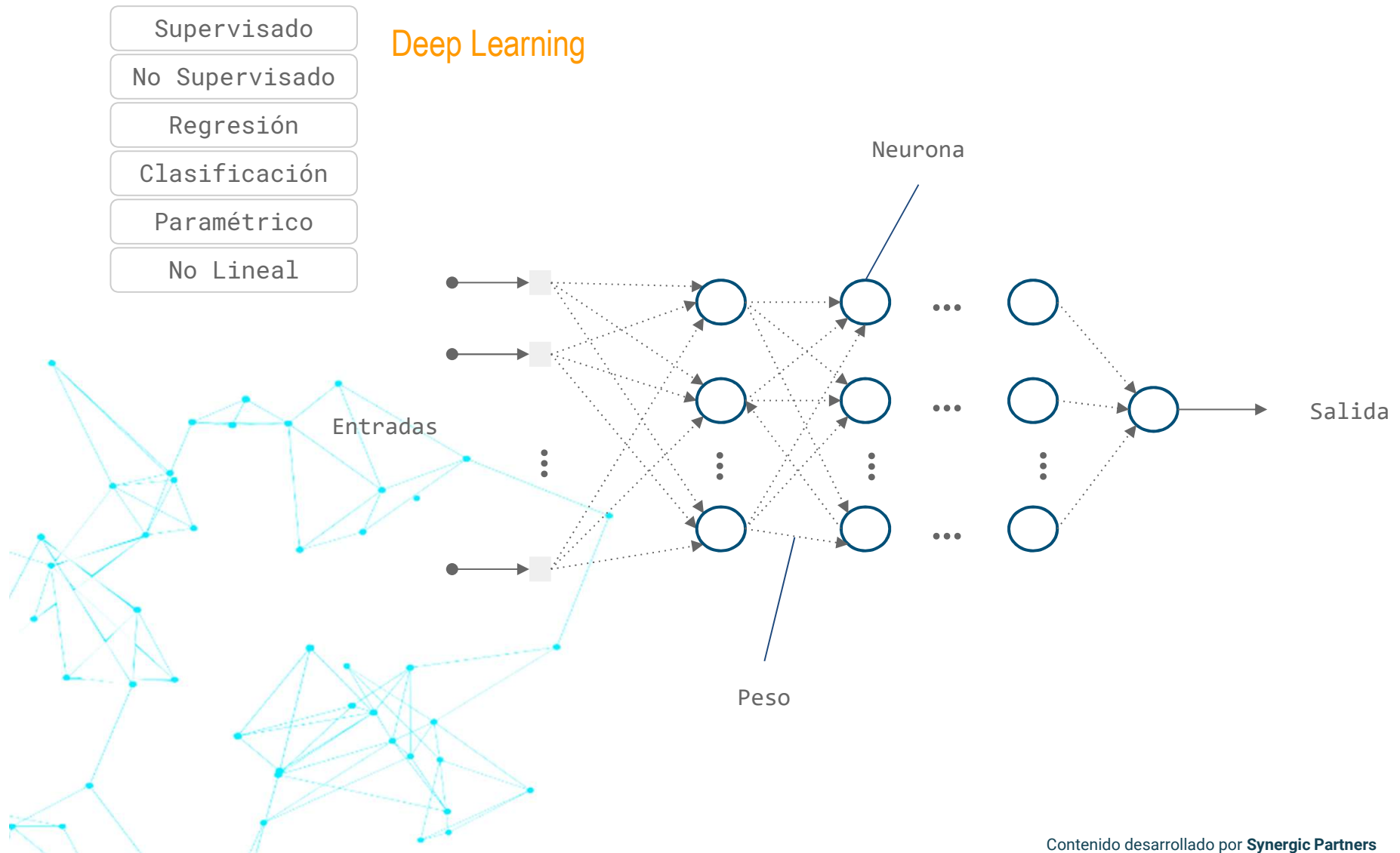
Deep Learning

- Una red neuronal artificial con una arquitectura multicapa consiste en un grafo finito.
- La red está organizada en una capa de **entrada**, una o más capas **ocultas**, y una capa de **salida**.
- Cada capa estando compuesta por un conjunto de neuronas o unidades de cómputo.
- La entrada es procesada y transmitida de una capa a la siguiente hasta que se calcula el resultado final y es presentado a la salida.
- El aprendizaje consiste en ajustar los pesos de las conexiones entre capas.



INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Algoritmos y técnicas básicas



INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Algoritmos y técnicas básicas

Supervisado

No Supervisado

Regresión

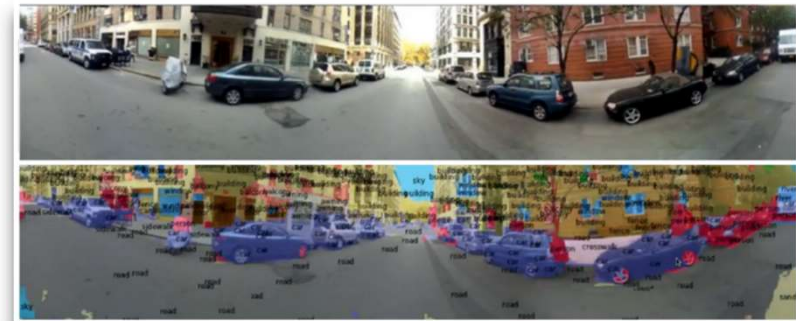
Clasificación

Paramétrico

No Lineal

Deep Learning

“Classical” applications:
object classification, detection and segmentation.



Speech translation



INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Algoritmos y técnicas básicas

	Supervisado		No supervisado
Técnicas principales	Clasificación	Regresión	Clustering
Regresión		X	
Regresión Logística	X		
k-NN	X	X	
Naive Bayes	X		
Árboles de Decisión	X	X	
Random Forest	X	X	
Redes Neuronales	X	X	
Support Vector Machine	X	X	
K-means			X
Clustering jerárquico			X

An abstract network diagram consisting of numerous teal-colored nodes connected by thin teal lines. The nodes are scattered across the left and center of the slide, forming a complex web of connections. Some nodes are isolated, while others are part of larger, interconnected clusters.

Evaluación y selección de modelos

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Evaluación y selección del modelo



Evaluación & Selección

Evaluar un modelo consiste en estimar su desempeño al predecir la salida con el fin de:

1. Estimar la capacidad de generalización de nuestro modelo sobre datos futuros aún no observados.
2. Mejorar el desempeño predictivo ajustando los hiperparámetros del algoritmo de aprendizaje.
3. Comparar diferentes algoritmos sobre el mismo problema y conjunto de datos de entrenamiento.

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Problemas de ajuste del modelo

Sesgo (*Bias*)

Es el error que se produce por asumir una forma de la función que aproxima a la salida que no se adapta bien a los datos de entrenamientos.

Varianza (*Variance*)

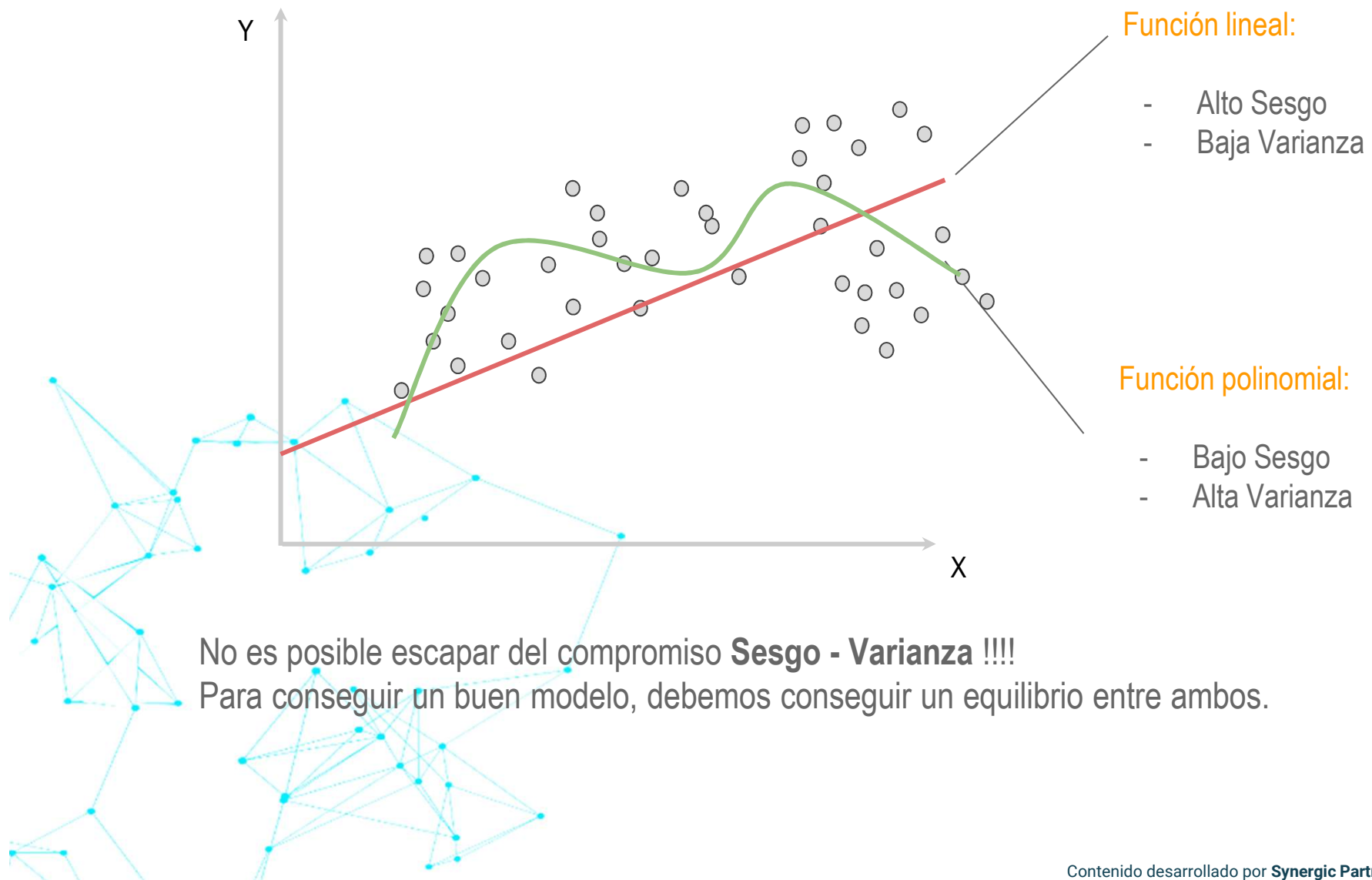
Es el error que se produce por la **sensibilidad** del modelo a pequeñas variaciones en los datos de entrenamiento. Por ejemplo, un modelo con alta variabilidad puede intentar aprender a seguir el ruido en lugar de los datos reales.



Al escoger un método estadístico para estimar la función f , vamos a tener que asumir un compromiso entre dos opciones, **sesgo** y **varianza**:

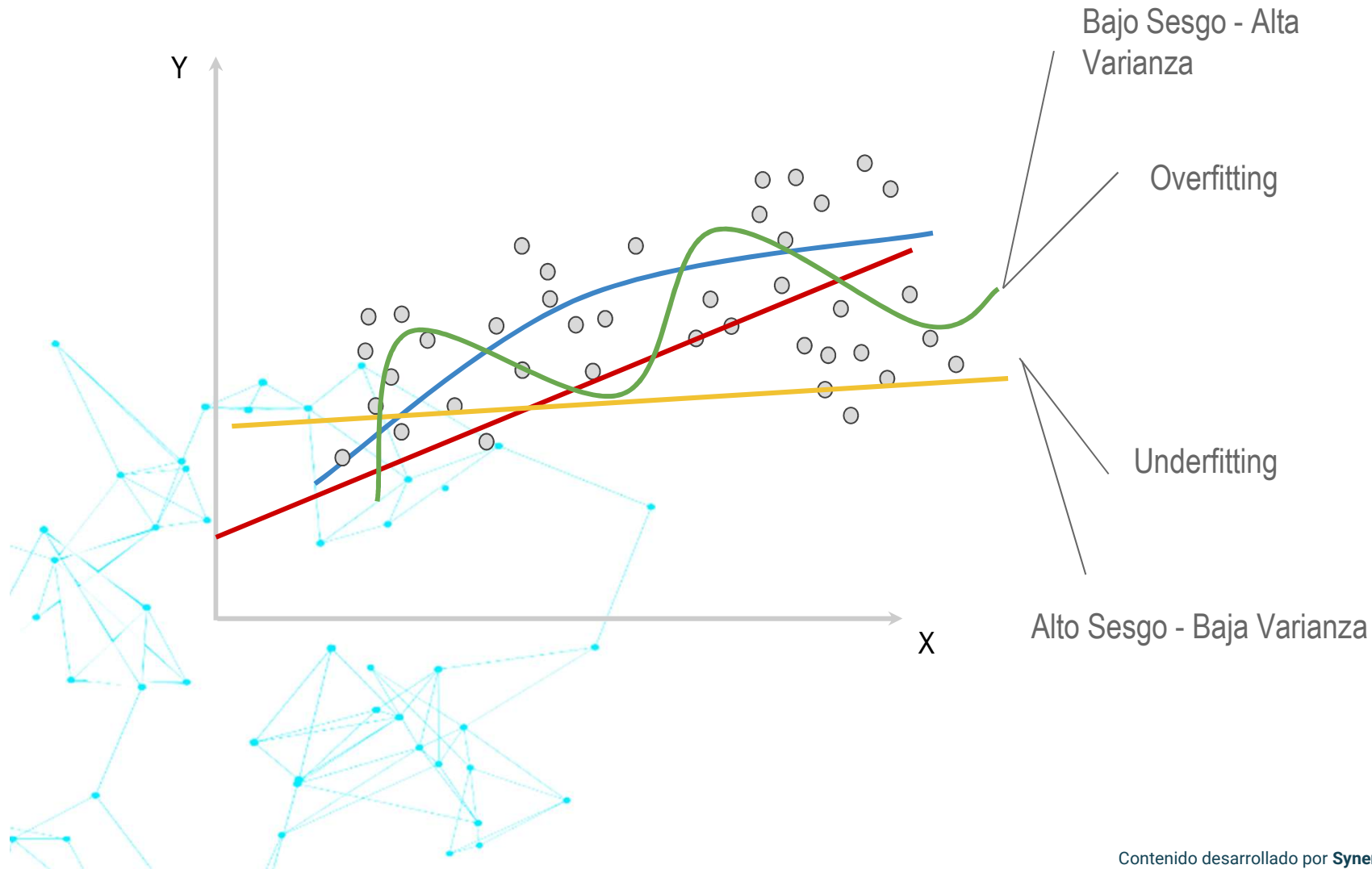
INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Problemas de ajuste del modelo



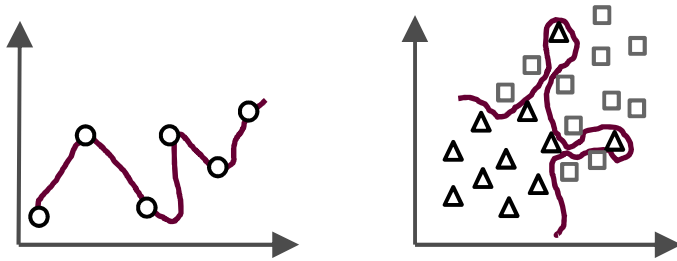
INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Problemas de ajuste del modelo



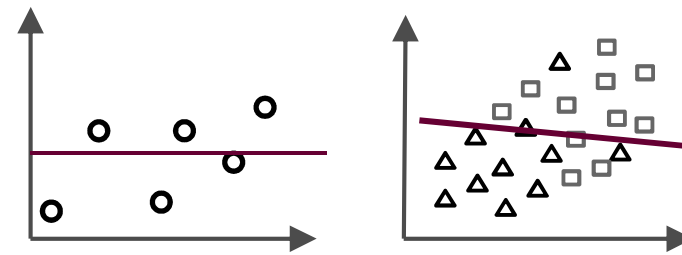
INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Problemas de ajuste del modelo



Overfitting

Se detecta cuando el modelo tiene un alto desempeño de predicción sobre los **datos** de entrenamiento, pero un pobre rendimiento ante nuevos datos (test data).



Underfitting

Ocurre cuando el algoritmo de aprendizaje no es capaz de capturar la relación funcional entre los atributos y el predictor. Generalmente es el resultado del uso de modelos muy sencillos.

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Hiperparámetros

- Definen conceptos de alto nivel como la complejidad y capacidad de aprendizaje del algoritmo.
- Parámetros que el algoritmo no es capaz de aprender por sí mismo a partir de los datos.
- Se ajustan fijando varios valores, entrenando el modelo y seleccionando aquellos valores que maximicen el desempeño del algoritmo.

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Hiperparámetros. Ejemplos

- Número de hojas o profundidad de un árbol de decisión
- Número de clusters en K-means
- Número de capas ocultas de una red neuronal

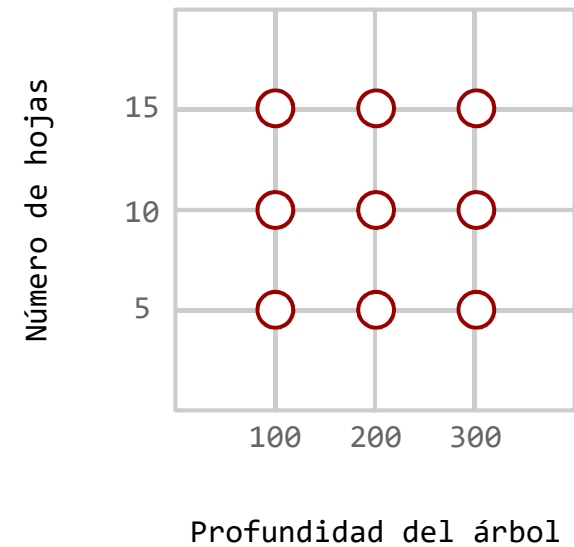


INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Ajuste de hiperparámetros

Grid Search

Consiste en realizar una búsqueda exhaustiva del espacio de hiperparámetros fijando manualmente un subconjunto de valores de los mismos.

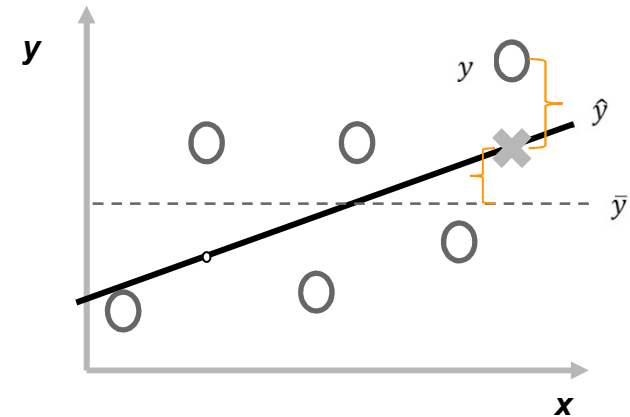


INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Métricas de evaluación - Aprendizaje supervisado

Regresión

Coeficiente de Determinación R^2 , es un estadístico que mide la bondad del ajuste como la proporción de variación de los resultados que puede explicarse por el modelo.



$$R^2 = \frac{SSR}{SSR + SSE}$$

$$SSR = \sum_i (\hat{y}_i - \bar{y})$$

Desviación respecto a la media explicada por el modelo

$$SSE = \sum_i (y_i - \hat{y}_i)$$

Desviación respecto a la media no explicada por el modelo

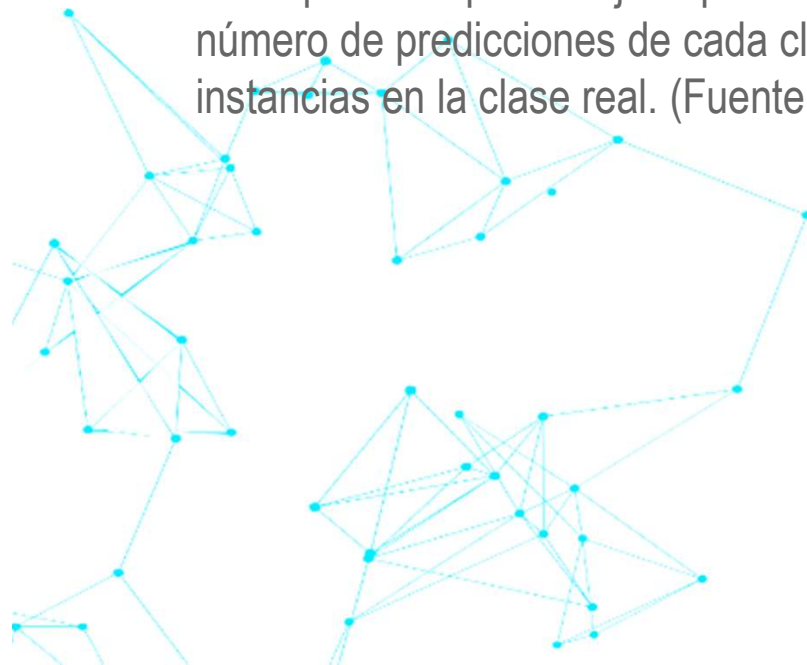
INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Métricas de evaluación - Aprendizaje supervisado

Clasificación

Matrices de confusión o Tabla de contingencia

Es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. (Fuente: Wikipedia)



INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Métricas de evaluación - Aprendizaje supervisado

Clasificación

Predicción del Modelo

Valor Verdadero

Verdaderos Positivos

TP

nº elementos positivos clasificados como positivos.

Falsos Negativos

FN

nº elementos positivos clasificados como negativos.

Falsos Positivos

FP

nº elementos negativos clasificados como positivos.

Verdaderos Negativos

TN

nº elementos negativos clasificados como negativos.

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Métricas de evaluación - Aprendizaje supervisado

Clasificación

Métricas de evaluación

Accuracy: Frecuencia de predicciones correctas.

$$\frac{TP + TN}{P + N}$$

Recall: Proporción de valores positivos predichos correctamente.

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

Precisión: Valor predictivo positivo sobre toda la muestra.

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

F-score: Combina precisión y sensibilidad en una misma métrica como la media armónica de ambas.

$$\frac{2 * Precision * Recall}{Precision + Recall}$$

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Métricas de evaluación - Aprendizaje supervisado

Clasificación

	Predicted <small>TRUE</small>	Predicted <small>FALSE</small>
Total	2951	1484
Actually <small>TRUE</small>	1806	283
Actually <small>FALSE</small>	1145	1201

$$\text{Accuracy} = \frac{TP + TN}{P + N}$$

$$\text{Precisión} = \frac{TP}{TP + FP}$$

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

$$\text{F-score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

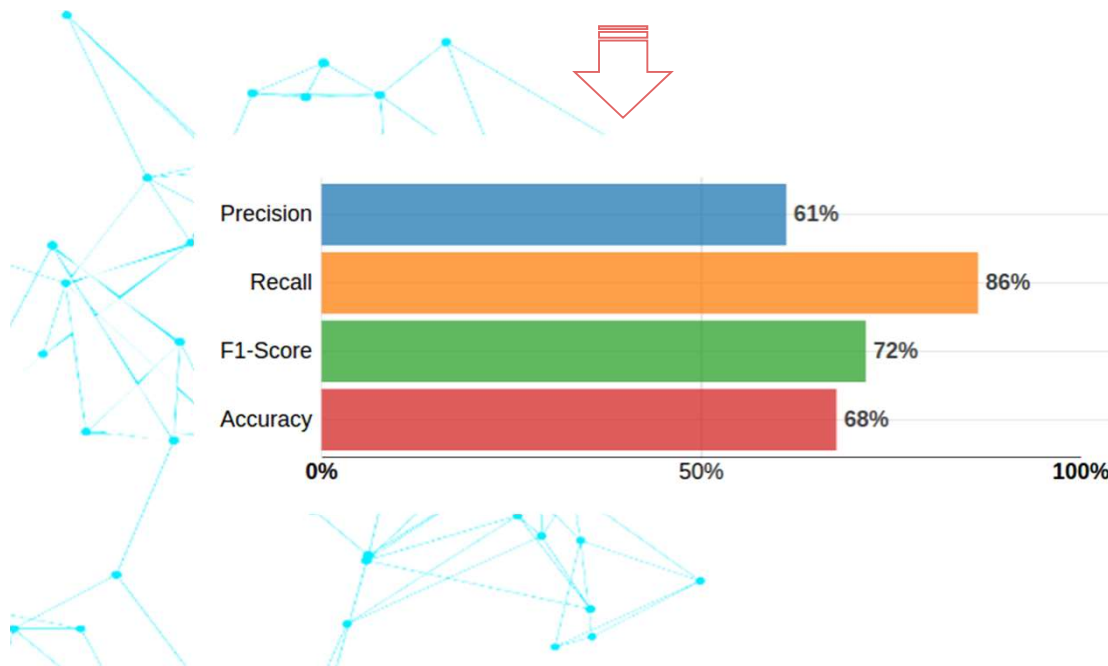


INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Métricas de evaluación - Aprendizaje supervisado

Clasificación

	Predicted <small>TRUE</small>	Predicted <small>FALSE</small>
Total	2951	1484
Actually <small>TRUE</small>	1806	283
Actually <small>FALSE</small>	1145	1201



INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Métricas de evaluación - Aprendizaje supervisado

Clasificación

Supongamos que disponemos de la siguiente población:

Casos Positivos (1)	0001
Casos Negativos (0)	9999

Y que nuestro modelo predice siempre, sea cual sea el dato de entrada, un valor 0

$$TP = 0, TN = 9999, FP = 0, FN = 1$$

Accuracy

$$\frac{TP + TN}{P + N}$$

Accuracy

0.9999

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

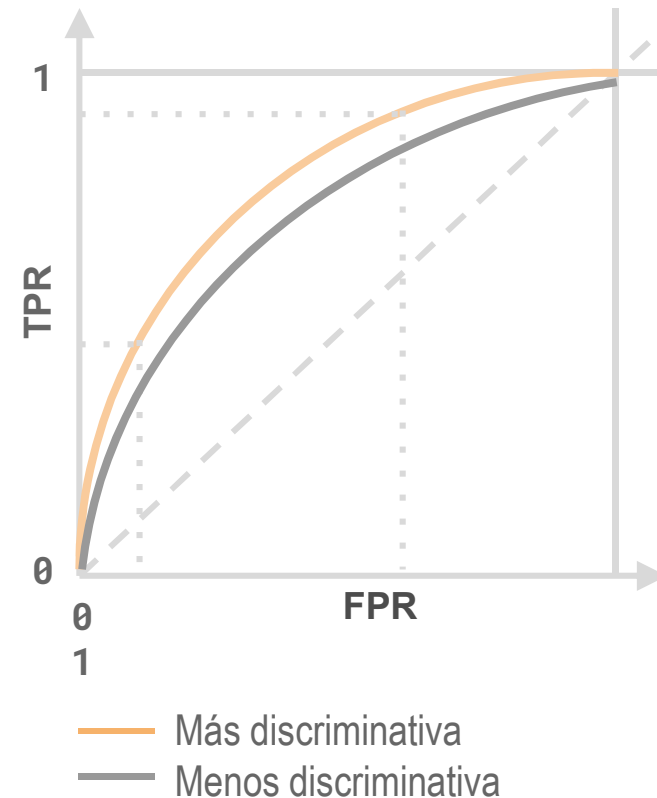
Métricas de evaluación - Aprendizaje supervisado

Clasificación

Curva ROC

Es una representación de la tasa de verdaderos positivos frente a la tasa de falsos positivos, según se va variando el umbral.

1. Cuanta mayor sea el área debajo de la curva (**Area Under Curve, AUC**), mejor es el algoritmo (idealmente 1)
2. Ayuda a **calibrar el umbral** en el punto de trabajo que se requiera.



INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

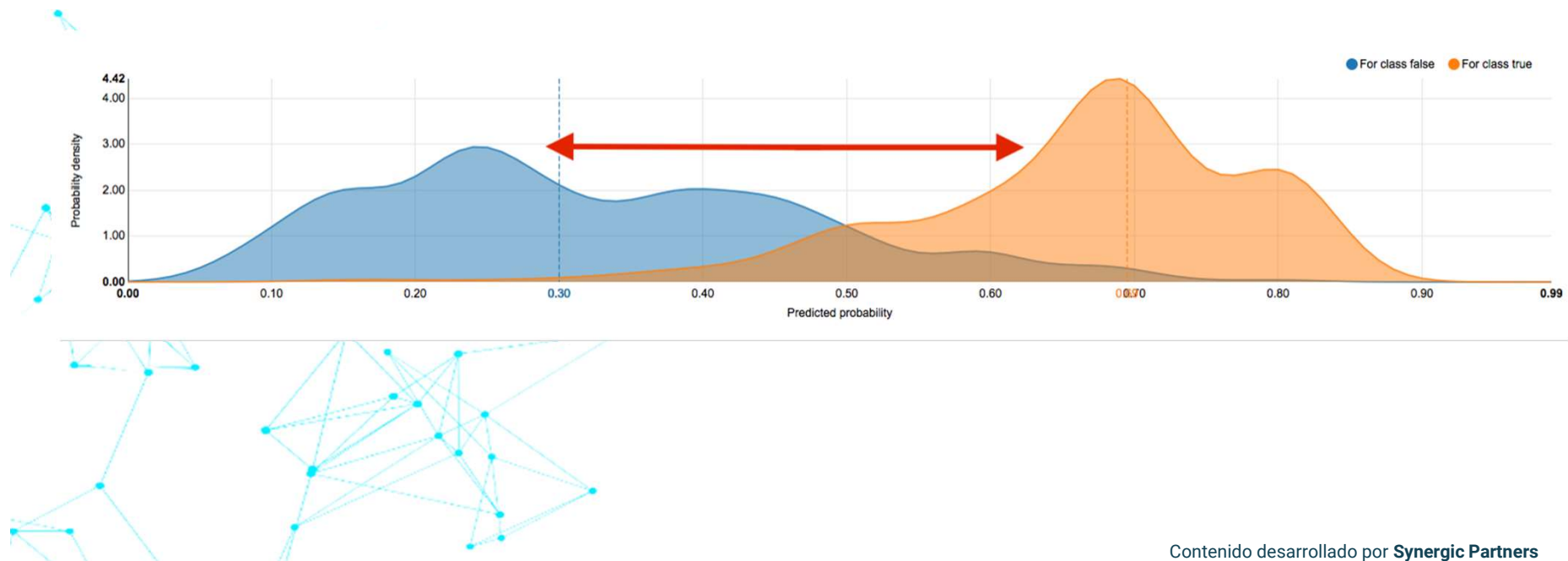
Métricas de evaluación - Aprendizaje supervisado

Clasificación

Curvas densidad

Las curvas de densidad representan la capacidad del modelo de identificar y separar las clases correctamente.

Muestra cómo se reparten las clases en el conjunto de validación de acuerdo a su probabilidad de pertenecer dicha clase según el modelo. Idealmente las clases no se deben sobreponer.



INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

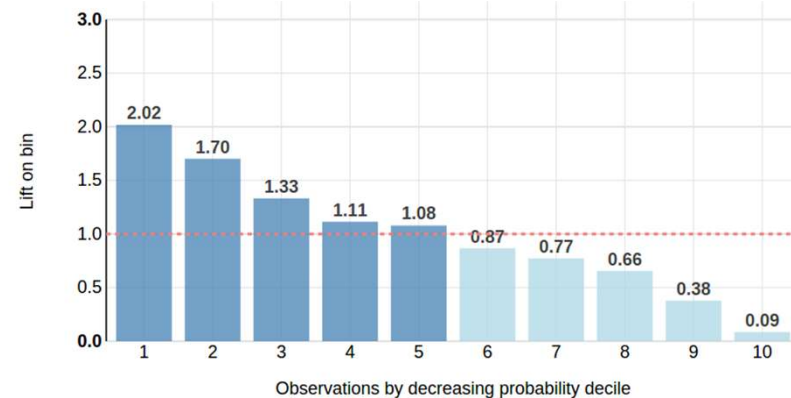
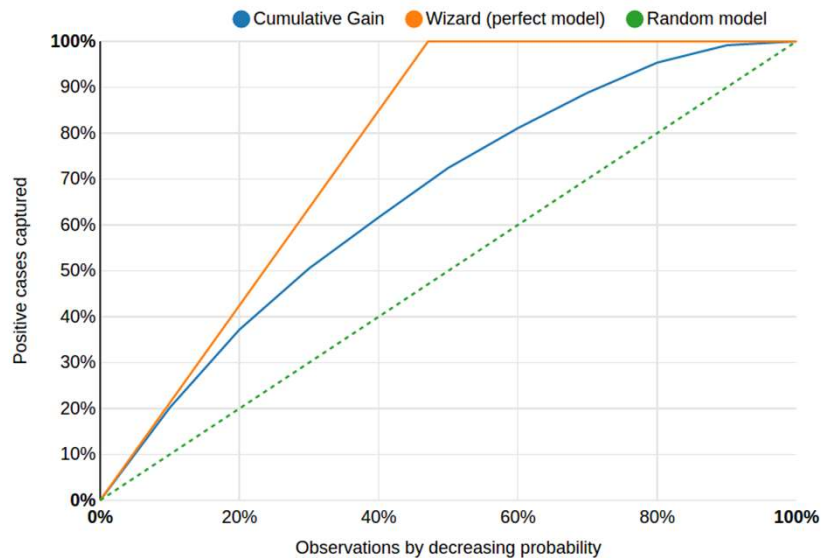
Métricas de evaluación - Aprendizaje supervisado

Clasificación

Curva Lift

El **lift** es la proporción entre los resultados obtenidos con el modelo y los resultados obtenidos con un modelo aleatorio.

La gráfica de la derecha ordena las observaciones por decil en orden decreciente de probabilidad predicha mostrando su valor de lift correspondiente.



INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Métricas de evaluación - Aprendizaje no supervisado

Clustering

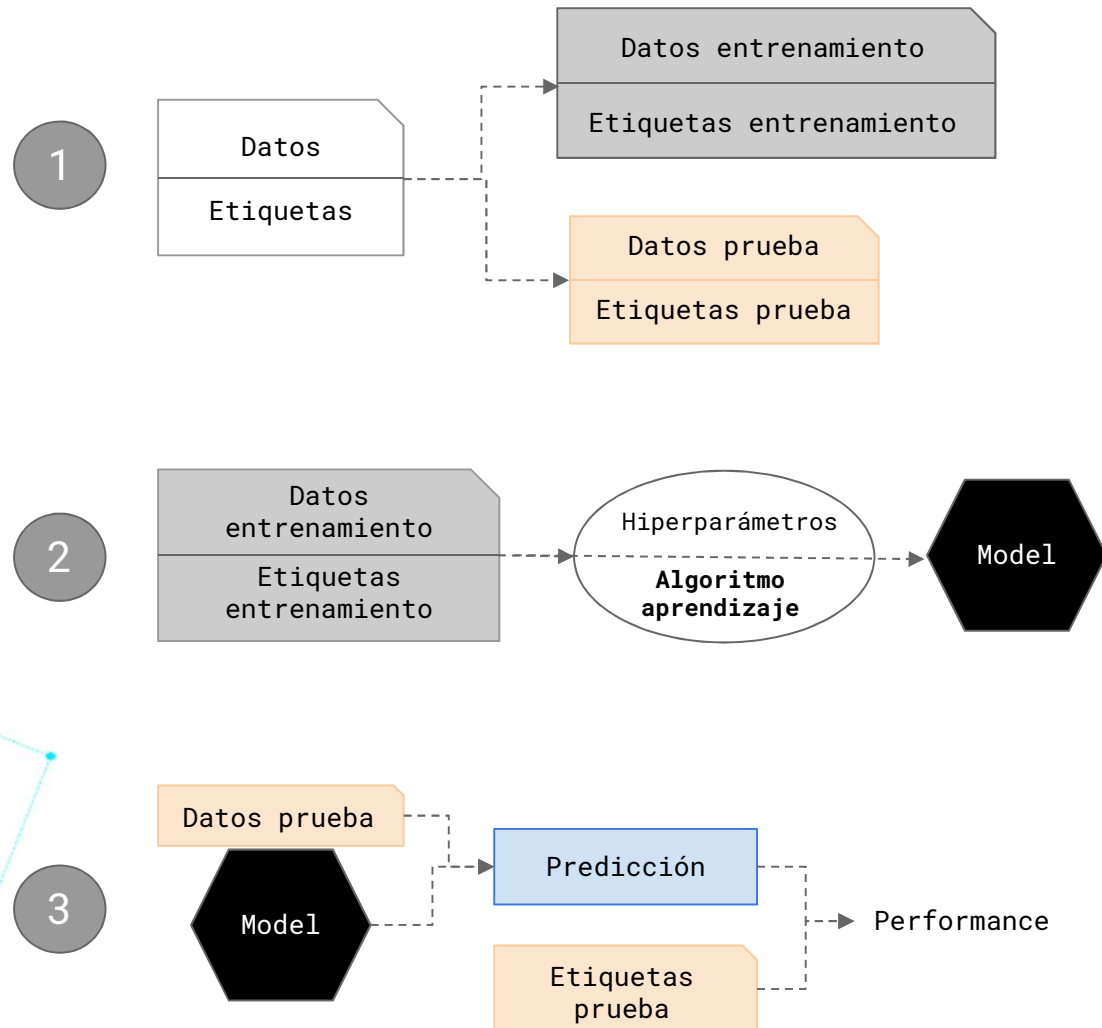
Modelos de clustering

- La forma de evaluar este modelo es heurística y los puntos que se suelen comprobar son los siguientes:
 - Si hay algún 'cluster' con muy pocos datos significa que el número de 'clusters' es demasiado alto, es necesario disminuir k .
 - Si hay 'centroides' que están demasiado cerca entre sí, quiere decir que el número de 'clusters' es demasiado alto, es necesario disminuir k .
 - Se pueden realizar representaciones en dos dimensiones de pares de características de las que se componen los datos, para ver si hay una clara agrupación gráfica de los 'clusters'. Esto sólo es útil cuando el número de características es bajo.
 - Si no se aprecia agrupación según características de las que se esperaba que existiera agrupación, significa que hay pocos centroides.
- La evaluación de los modelos de clustering no puede hacerse de forma exhaustiva ya que al ser una técnica no supervisada, no existen datos etiquetados.

INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Métodos de evaluación. Método hold-out

Método Hold-out. Los datos son divididos en **datos de entrenamiento** (training data) y **datos de prueba**. El algoritmo de aprendizaje ajusta el modelo usando los datos de entrenamiento, y se evalúa usando los datos de prueba, es decir, con datos desconocidos por el algoritmo durante la fase de entrenamiento.



INTRODUCCIÓN A DATA SCIENCE Y MACHINE LEARNING

Métodos de evaluación. Validación cruzada

Validación cruzada. Es una técnica que permite evaluar el valor predictivo de un algoritmo de aprendizaje automático, y evitar problemas de *overfitting* cuando no disponemos del conjunto explícito de datos de prueba.



Telefónica
FUNDACIÓN

Conecta Empleo

