

Analítica Empresarial para la Toma de Decisiones Efectivas. ¿Cómo usar Big Data?

Rolando de la Cruz

Director Académico Magíster en Data Science

rolando.delacruz@uai.cl

Agenda

- ✓ Machine Learning - Aprendizaje Automático
- ✓ Evaluación del clasificador
- ✓ Software

Aprendizaje Automático

Una máquina *aprende* una particular tarea **T** considerando las experiencias de tipo **E** respecto de una medida de performance **P**, si la máquina efectivamente mejora su performance **P** en la tarea **T** a partir de la experiencia **E**

Tom Mitchell (1997)

Aprendizaje Automático

- ✓ El Aprendizaje Automático estudia cómo construir programas que mejoren automáticamente con la experiencia.

Aprendizaje Automático

Es una derivación natural desde la intersección entre la Estadística y la Ciencia de la Computación.

- ✓ **Inferencia Estadística**
 - modelar la población usando la muestra (datos)

- ✓ **Ciencia de la Computación**
 - algoritmos que realicen la tarea eficientemente

Aprendizaje Automático: Terminología

✓ Instancia, ejemplo o registro

Una *instancia* es cada uno de los datos de los que se disponen para hacer un análisis. Si se quiere predecir el comportamiento de los clientes de un servicio de telefonía, cada instancia correspondería a un abonado. Cada instancia, a su vez, está compuesta de características que la describen, como la antigüedad del cliente en la compañía, el gasto diario en llamadas, etc. En una hoja de cálculo, las instancias serían las filas.

Aprendizaje Automático: Terminología

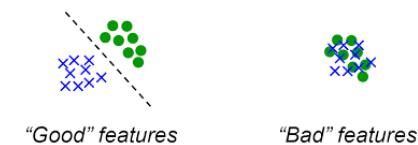
✓ Clase

Es el atributo o factor que queremos predecir, el objetivo de la predicción, como puede ser la probabilidad de reingreso de un paciente tras una intervención quirúrgica. En una hoja de cálculo, sería una de las columnas.

Aprendizaje Automático: Terminología

✓ Atributo o característica

Son los atributos (*features*) que describen cada una de las instancias del conjunto de datos. En el caso de una cartera de clientes, estaríamos hablando del número de compras de cada cliente, antigüedad, si es seguidor en redes sociales, si se ha dado de alta en la newsletter, qué productos ha comprado, etc. En una hoja de cálculo, serían las columnas.



Aprendizaje Automático: Tipos

- ✓ Aprendizaje **supervisado**
 - los datos de entrenamiento ya están clasificados
- ✓ Aprendizaje **semi supervisado**
 - una parte de los datos de entrenamiento están clasificados y la otra no
- ✓ Aprendizaje **no supervisado**
 - los datos de entrenamiento no están clasificados
- ✓ Aprendizaje **reforzado**
 - aprendizaje gradual en base a premios y castigos

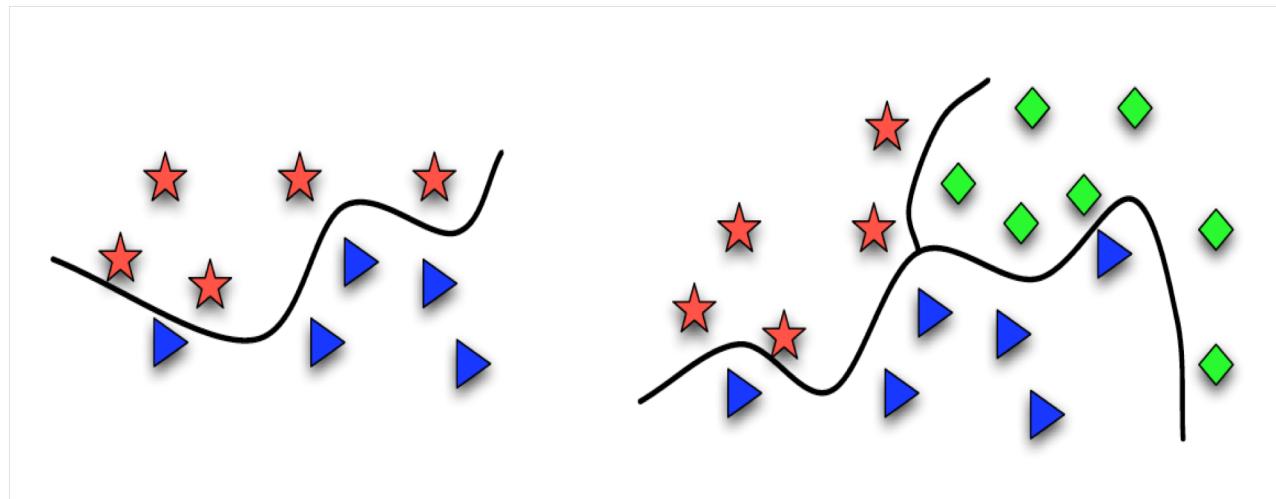
Aprendizaje Automático: Tipos

- ✓ Aprendizaje **supervisado**
 - Es la técnica más usada del ML.

credit scoring. Una entidad financiera clasifica como **bajo** o **alto** el riesgo de dar un crédito a un cliente en función de su información financiera pasada (créditos, edad, impagos, ingresos, profesión, etc). En este caso, ML genera un algoritmo utilizando los datos pasados, el cual es capaz de decidir si acepta o no una nueva petición de un cliente

Aprendizaje Automático: Tipos

Aprendizaje **supervisado**



Clasificación **binaria**

Clasificación **multiclas**

Aprendizaje Automático: Tipos

Ejemplos de aprendizaje supervisado:

- diagnóstico médico, acciones preventivas
- solicitudes de tarjetas de crédito
- cálculo de la cuota de un seguro (scoring)
- detección de fraude en transacciones electrónicas
- identificación de spam
- recomendación: artículos, películas, libros, música, productos
- inversiones financieras
- reconocimiento de lenguaje escrito y hablado
- etiquetado de imágenes

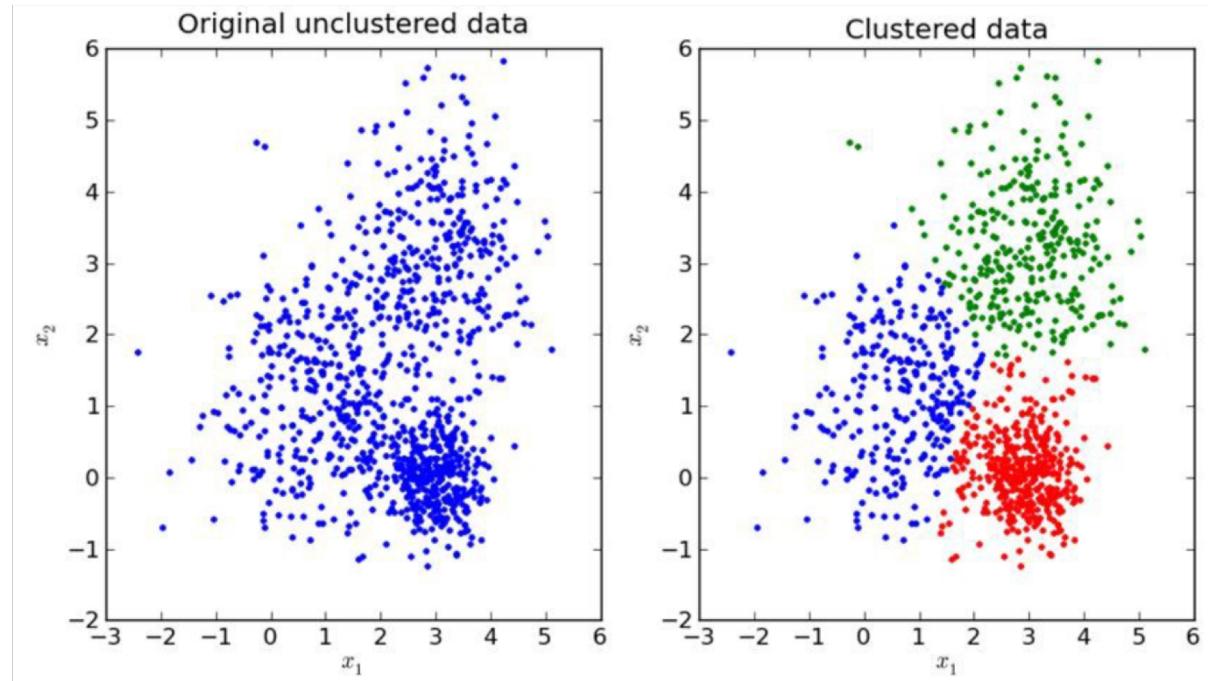
Aprendizaje Automático: Tipos

- ✓ Aprendizaje **no supervisado**

segmentación. Se suele utilizar el aprendizaje no supervisado para crear y descubrir patrones no conocidos en el comportamiento de los clientes de una web, app o comercio

Aprendizaje Automático: Tipos

- ✓ Aprendizaje **no supervisado**



Aprendizaje Automático: Tipos

Ejemplos de aprendizaje no supervisado:

- segmentación de clientes para campañas dirigidas
- agrupar pacientes por estadíos de la enfermedad
- segmentar contribuyentes del IVA para detectar grupos de mayor propensión a la evasión
- segmentar individuos en distintos tipos de personalidad
- segmentar curvas de luz
- etc.

APRENDIZAJE SUPERVISADO

Aprendizaje Supervisado

- ✓ Objetivo central:
 - Predicción, no causalidad

El foco está puesto en poder realizar predicciones de interés bajo condiciones complejas, no en estudiar el mecanismo causal que rige bajo esas condiciones.

Aprendizaje Supervisado

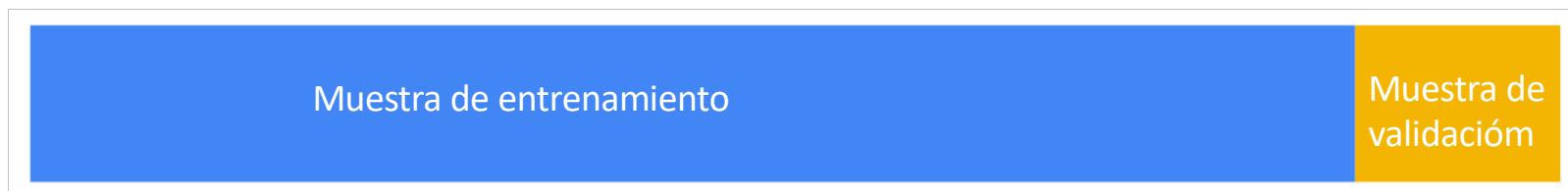
Esquema general:

- ✓ Datos
 - Separar datos de entrenamiento (desarrollo) y validación.
 - Definir instancias, clases y atributos.
- ✓ Experimentación
 - Selección de atributos.
 - Medidas de performance.
 - Validación cruzada.
- ✓ Validación de los modelos

Aprendizaje Supervisado

- ✓ Materia prima: Datos
 - **Muestra de entrenamiento:** Un subconjunto para entrenar un modelo.
 - **Muestra de validación:** Un subconjunto para probar el modelo entrenado.

Dividir en forma aleatoria el único conjunto de datos de la siguiente manera:



Aprendizaje Supervisado

- ▶ **Datos.** Muestra de entrenamiento

$$(X_1, Y_1), \dots, (X_n, Y_n) \in \Xi \times \Upsilon$$

- Usualmente $\Xi = \mathbb{R}^p$
- $\Upsilon = \mathbb{R}$ en este caso hablamos de regresión.
- $\Upsilon = \{0,1\}$ o $\Upsilon = \{-1,1\}$ en este caso hablamos de clasificación binaria.
- $\Upsilon = \{1,2,\dots,m\}$ en este caso es clasificación multiclas.

Aprendizaje Supervisado

➤ Regresión:

- Y_i puede ser el precio de una vivienda, X_i características de la casa y del entorno: el número de dormitorios, el tamaño de la casa, tasa de delincuencia del barrio, etc.

➤ Clasificación:

- Los valores Y_i son ‘etiquetas’ que indican al grupo al que pertenece X_i . Los valores de Y_i podrían ser las ‘etiquetas’ que podemos asignar a un cliente bancario {pagador, no pagador}, y las X_i son características observables sobre los clientes: edad, profesión, etc.

Aprendizaje Supervisado

Problema: **clasificar** una nueva instancia X del cual no se conoce la etiqueta Y

- ▶ Dado $X \in \Xi$
- ▶ Queremos saber a qué grupo $Y \in \{1, \dots, m\}$ pertenece
- ▶ Para ello utilizamos un *claseificador*

Claseificador:

$$g: \Xi \rightarrow \Upsilon$$

es decir, una función que a X le asigne una etiqueta Y

Aprendizaje Supervisado

***g* puede ser cualquier función:**

- ▶ Regresión logística (solo con $m = 2$)
- ▶ Análisis discriminante lineal de Fisher
- ▶ Análisis discriminante cuadrático
- ▶ k -vecinos más cercanos
- ▶ Árboles de clasificación
- ▶ Máquinas de soporte vectorial
- ▶ Bosques aleatorios
- ▶ Red neuronal artificial
- ▶ Red neuronal artificial profunda (*deep learning*), etc.

Evaluación del Clasificador

Una vez encontrado el mejor clasificador debemos:

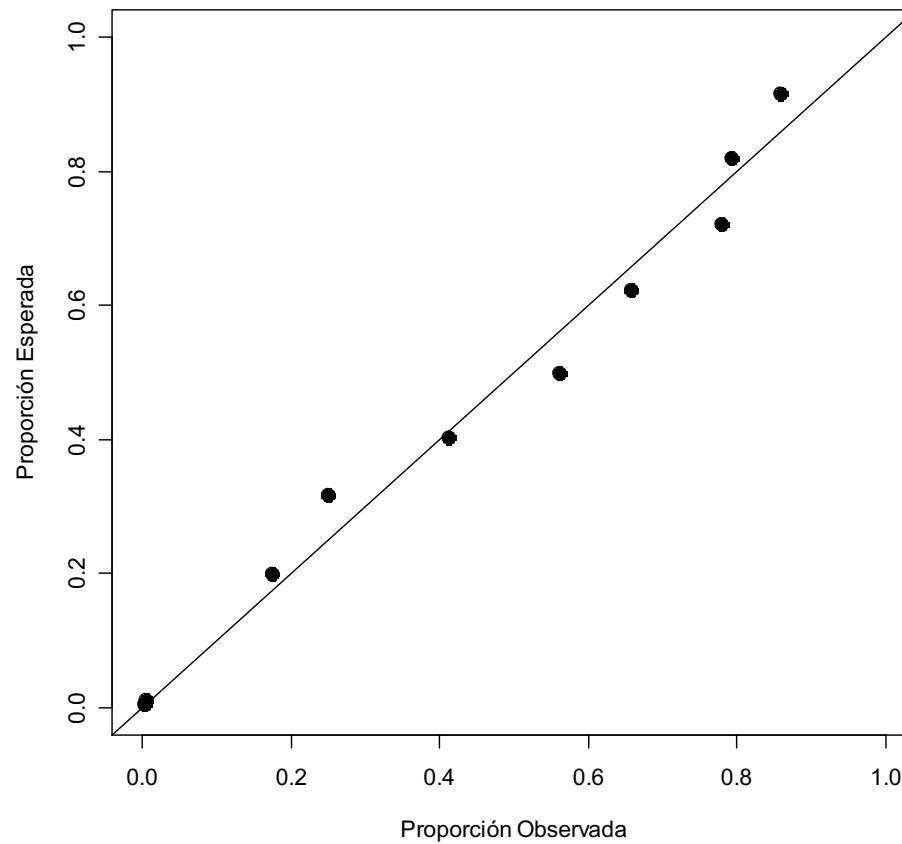
- ▶ ver si predice bien la clase (Y) en una nueva instancia (X_{new})
 - ▶ validez (“accuracy”)
 - ▶ generalizabilidad (“generalizability”)

Evaluación del Clasificador

- ▶ **Validez** es el grado en que las predicciones coinciden con las observaciones:
calibración y discriminación
 - ▶ *Calibración* compara el número predicho de eventos con el número observado en grupos de instancias (caso binario: test de Hosmer-Lemeshow)
 - ▶ *Discriminación* evalúa el grado en que el clasificador distingue entre instancias en las que ocurre el evento y las que no (caso binario: área bajo la curva ROC, K-S, Gini)

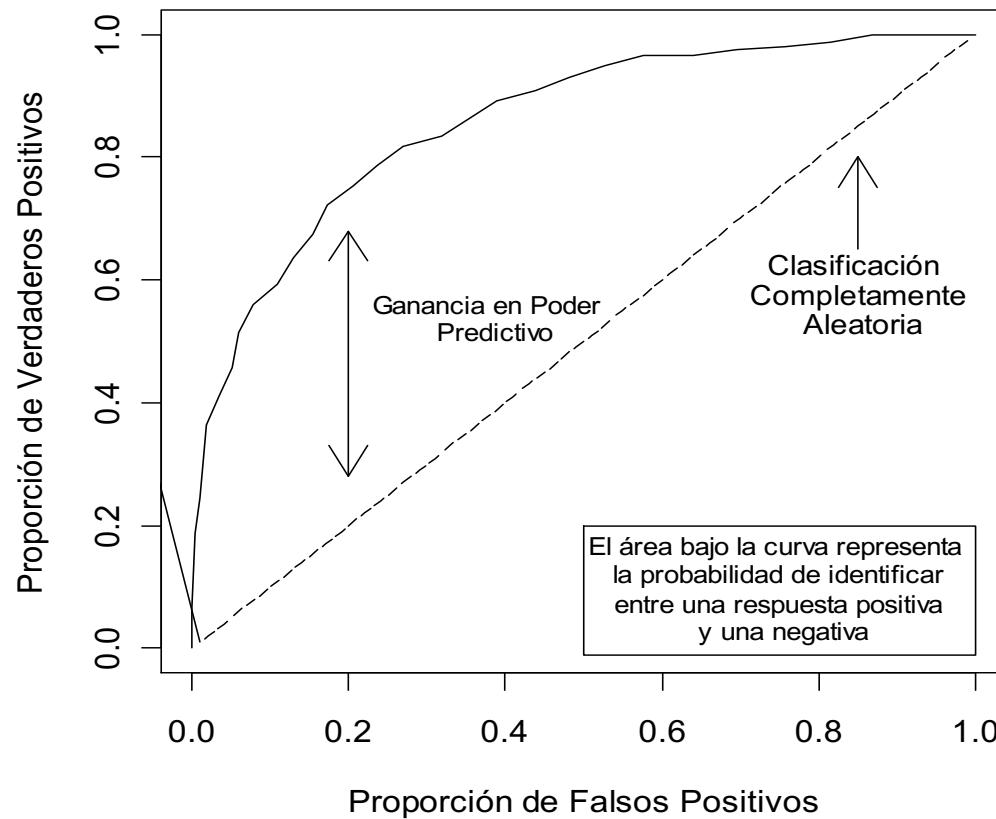
Evaluación del Clasificador

Hosmer - Lemeshow



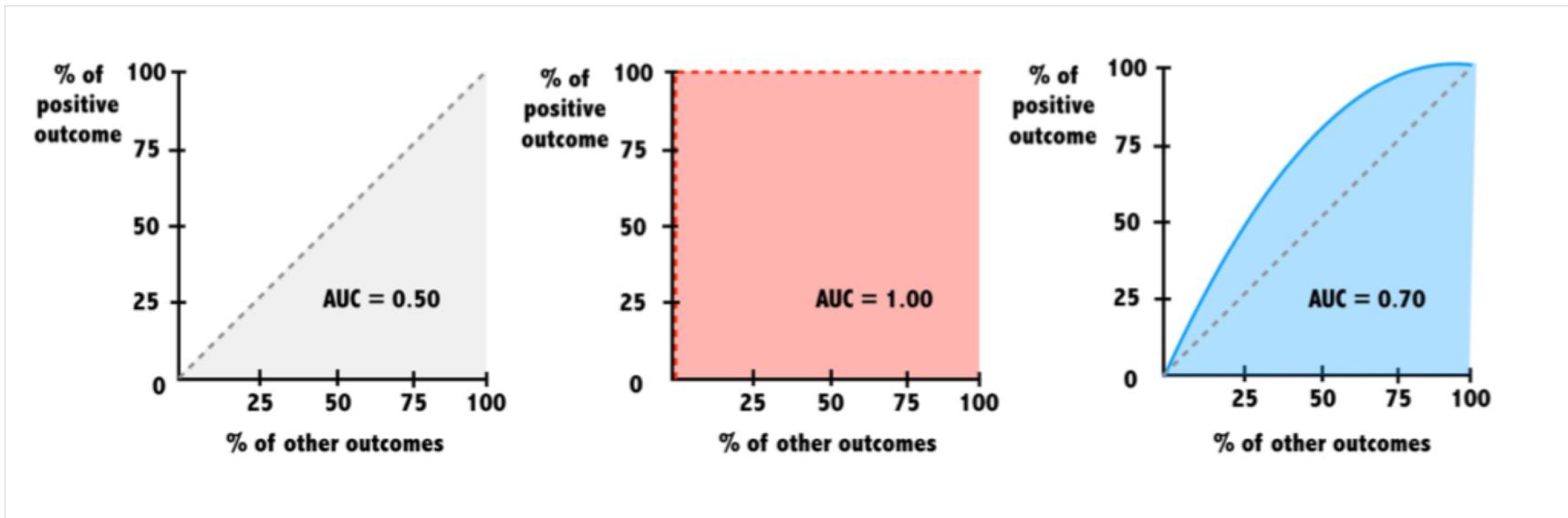
Evaluación del Clasificador

Curva ROC



Evaluación del Clasificador

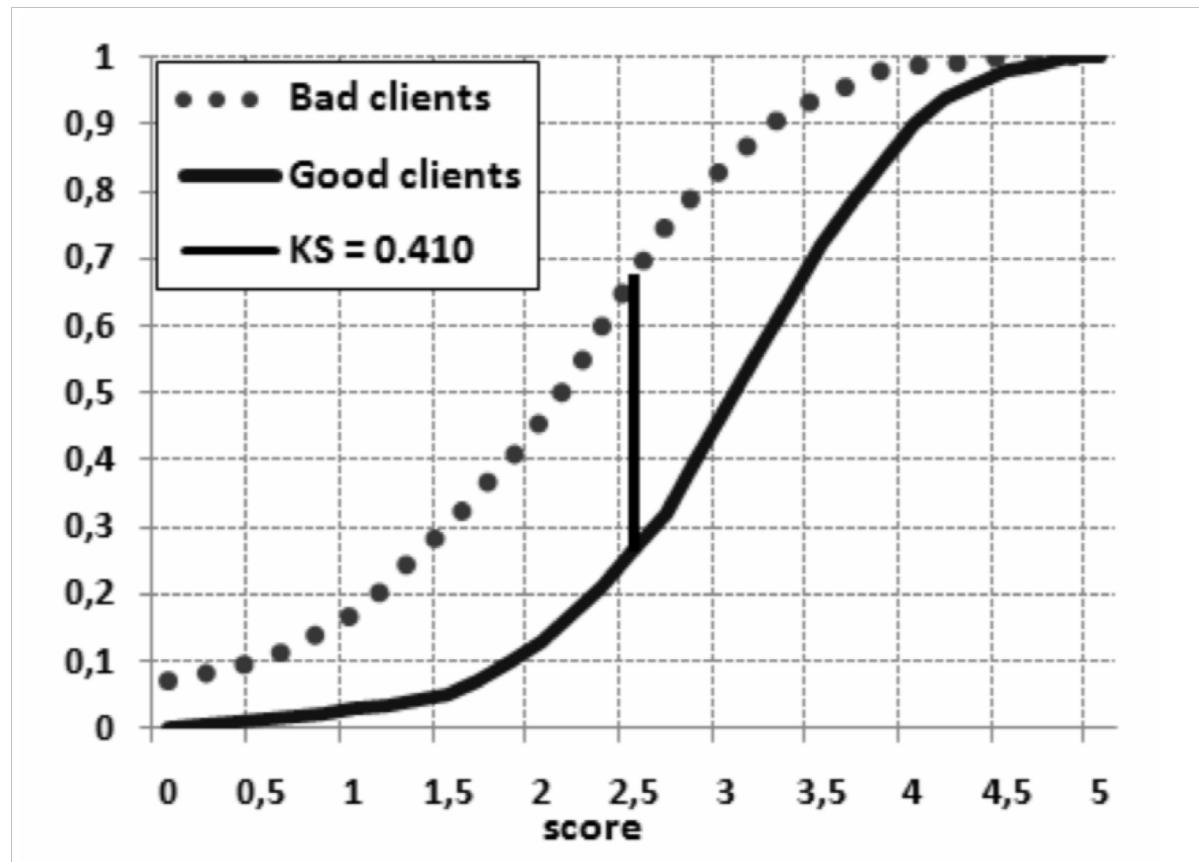
AUC: Área bajo la curva ROC



Evaluación del Clasificador

Estadístico K-S

Kolmogorov-Smirnov



Evaluación del Clasificador

- ▶ **Generalizabilidad** es la capacidad del clasificador de realizar predicciones válidas en instancias diferentes de aquellas en las que se ha generado: *reproducibilidad* y *transportabilidad*
 - ▶ *Reproducibilidad* capacidad del clasificador de realizar predicciones válidas en instancias no incluidas en la muestra con la que se ha generado, pero procedente de la misma población.
 - ▶ *Transportabilidad* capacidad de realizar predicciones válidas en instancias procedentes de una población distinta pero relacionada.

Evaluación del Clasificador

Matriz de Confusión

Es una herramienta que permite la visualización del desempeño de un clasificador que se emplea en aprendizaje supervisado.

Evaluación del Clasificador

		<u>True class</u>				
		<u>p</u>	<u>n</u>			
<u>Hypothesized class</u>	<u>Y</u>	True Positives	False Positives	fp rate = $\frac{FP}{N}$	tp rate = $\frac{TP}{P}$	
	<u>N</u>	False Negatives	True Negatives	precision = $\frac{TP}{TP+FP}$	recall = $\frac{TP}{P}$	
Column totals:		P	N	accuracy = $\frac{TP+TN}{P+N}$		

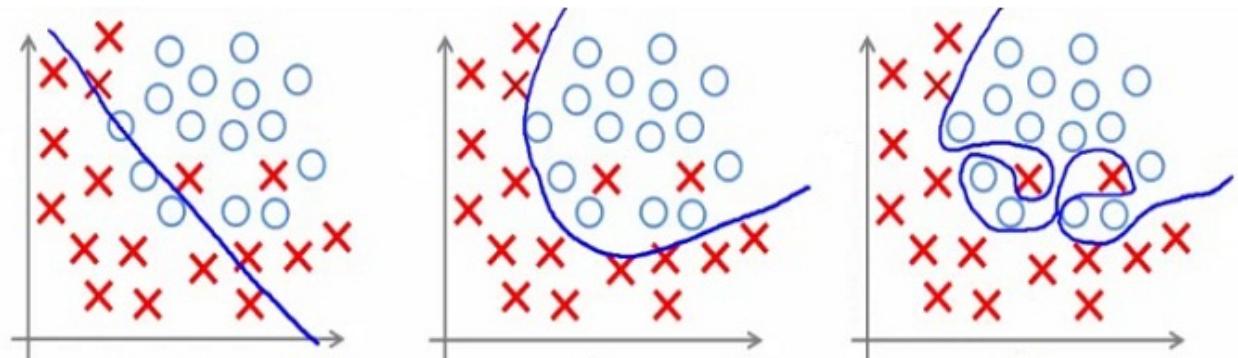
F-measure = $\frac{2}{1/\text{precision}+1/\text{recall}}$

Fig. 1. Confusion matrix and common performance metrics calculated from it.

Fuente: T. Fawcett. (2006). An introduction to ROC analysis. Pattern Recognition Letters 27, 861–874.

Aprendizaje Supervisado

Equilibrio en el proceso de aprendizaje supervisado



Under-fitting

(too simple to
explain the
variance)

Appropriate-fitting

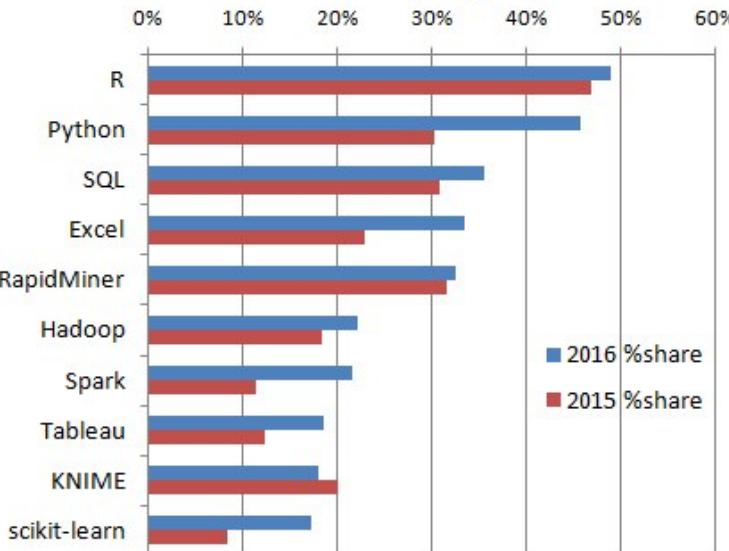
Over-fitting

(forcefitting -- too
good to be true)

Software

Según KDnuggets, R y Python son los lenguajes de programación más preferidos para realizar analítica de datos:

KDnuggets Analytics/Data Science 2016 Software Poll, top 10 tools



**KDnuggets 2017 Data Science Software Poll:
Top Tools Associations**



Analítica Empresarial para la Toma de Decisiones Efectivas. ¿Cómo usar Big Data?

Rolando de la Cruz

Director Académico Magíster en Data Science

rolando.delacruz@uai.cl