Clasificación binaria, multiclase y métricas. Curva ROC

Juan José Silva Torres

Etapas en el proceso de modelización

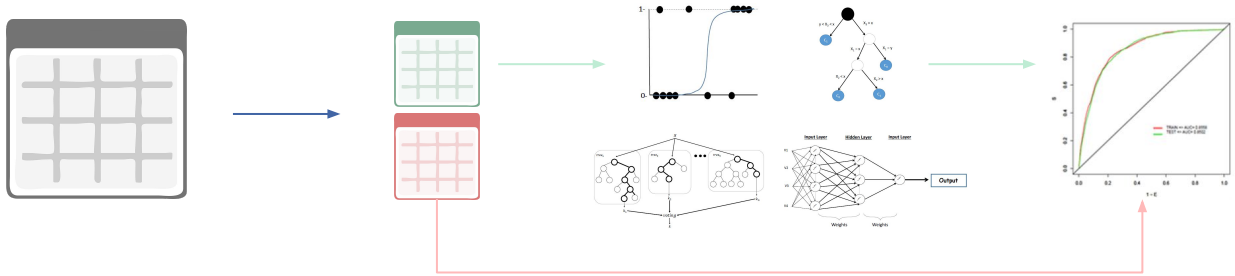
Visión global

Análisis de conjunto de datos

Muestreo

Selección del algoritmo

Validación



Comprender la naturaleza del target a predecir y de las variables explicativas de las que se dispone.

A partir del conjunto de datos, se define uno de entrenamiento y otro de test.

Se selecciona uno o varios algoritmos de modelización y se ajusta el modelo.

Se evalúa la bondad y precisión del modelo.

Visión global

Análisis de conjunto de datos

Comprender la naturaleza del target a predecir y de las variables explicativas de las que se dispone.

Muestreo

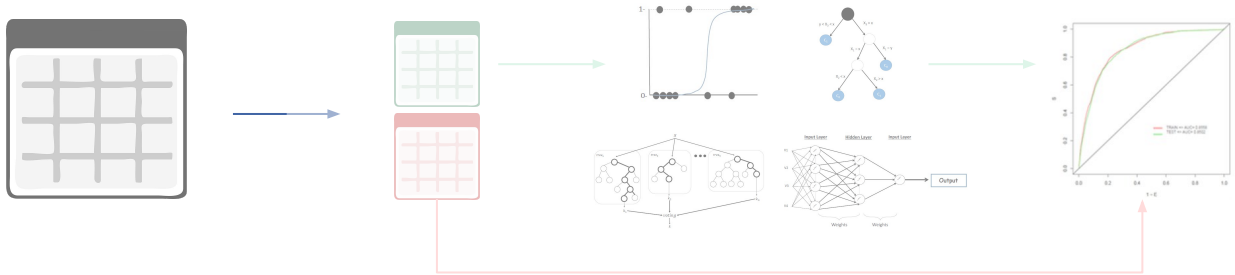
A partir del conjunto de datos, se define uno de entrenamiento y otro de test.

Selección del algoritmo

Se selecciona uno o varios algoritmos de modelización y se ajusta el modelo.

Validación

Se evalúa la bondad y precisión del modelo.



Análisis del conjunto de datos

La validación inicial de los datos es la fase más importante de cualquier proceso de modelización.

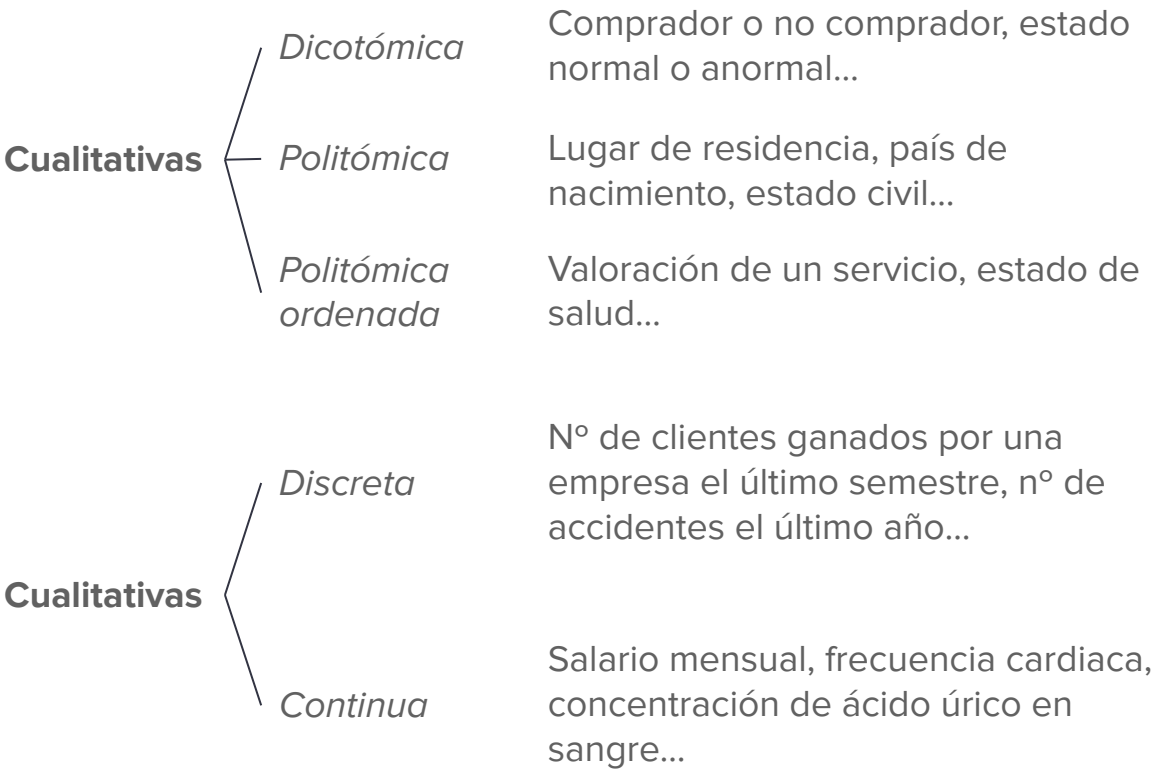
El análisis depende de **la naturaleza de las variables.**

Las variables pueden clasificarse en:

a) Cualitativas. Dicotómicas y politómicas.

b) Cuantitativas. Discretas y continuas.

Algunos ejemplos



Cómo realizar el análisis descriptivo

Cualitativas y cuantitativas discretas con pocos valores:

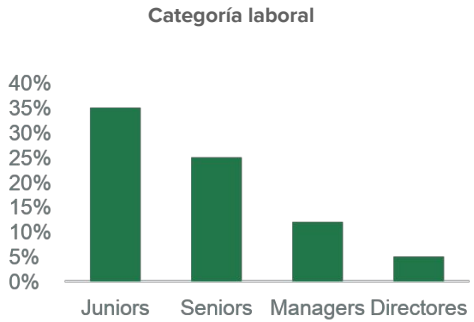
- Distribución de frecuencias.
- Detección de valores faltantes (*missing*) o fuera de rango (*outliers*).

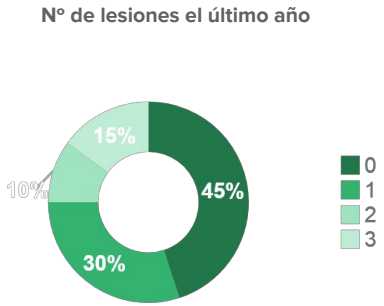
Cuantitativas continuas:

- Medidas de centralización, localización y dispersión.
- Histogramas y gráficos de cajas y patillas (*box-plot*).
- Detección de valores faltante (*missing*) o fuera de rango (*outliers*).

Distribución de frecuencias

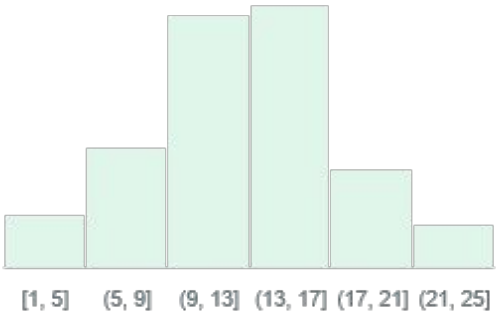
Se cuenta el número de observaciones que cae en cada una de las categorías de la variable. Se puede representar tanto el valor en términos absolutos como porcentuales.



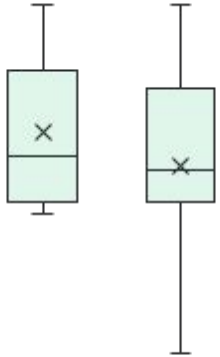


Histograma y *box plot*

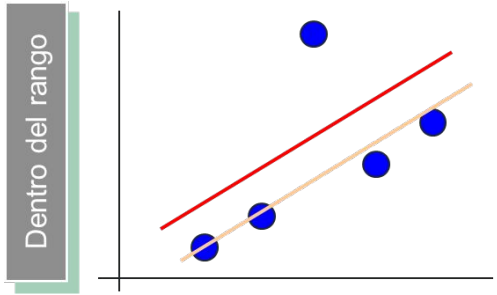
El **histograma** permite evaluar de manera visual la distribución de una variable continua. Para ello se crean categorías de la variable continua y se cuenta cuantas observaciones cae en cada una de ellas. Posteriormente se representa el gráfico como una distribución de frecuencias.



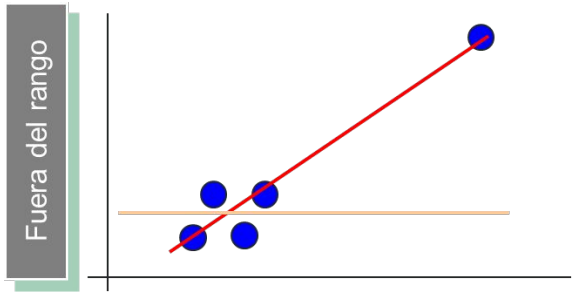
El gráfico de cajas y patillas (**box plot**) sintetiza las medidas de centralización y localización de una variable continua.



Cuidado con los *outliers*



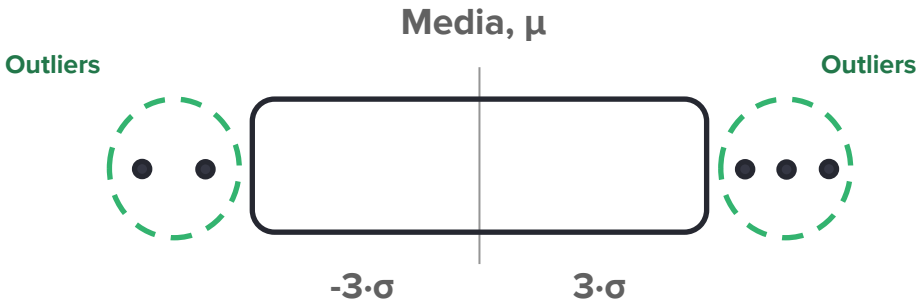
- Cambian el nivel pero no la pendiente.
- Estos errores conviene explicarlos, incluyendo alguna nueva variable.
- No debe quitarse el dato del estudio.



- Cambian la pendiente.
- Difíciles de explicar.
- Lo mejor es quitar el dato del estudio.

Definición de *outliers*

A partir de media (μ) y desviación típica (σ),

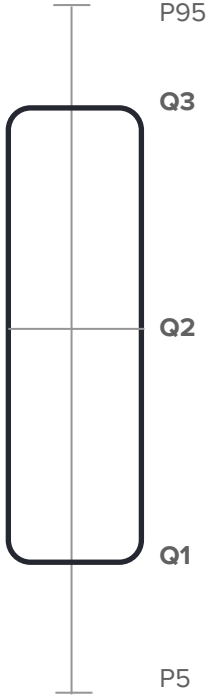


¿Cómo imputar estos valores?

A los valores qué están fuera del rango **[μ - 3·σ, μ + 3·σ]** se les asigna un valor mínimo y máximo.

Definición de *outliers*

A partir de la mediana y del rango intercuartílico (IQR),



La **mediana (Q2)** es el percentil 50 de los datos, es decir, el valor de la variable que deja a su izquierda el 50% de las observaciones.

El **rango intercuartílico** (IQR))es la diferencia entre el percentil 75 y el 25,
IQR = Q3 – Q1

Imputación de *outliers*

De cara a reducir el impacto de *outliers* en el conjunto de datos, se suele seguir algunas de las siguientes estrategias:

- *Anclar* los valores de las variables continuas a partir de cierto percentil suficientemente extremo (p99, p95, p1, p5...).

- *Categorizar la variable continua* pasando a tratarla como una variable de clase (discreta) en la que se estima un parámetro para cada clase.

- La *eliminación* de los registros asociados a valores altos de las variables se antoja una estrategia peligrosa sobre todo si el número de variables explicativas es muy alto.

Qué es un valor *missing*

Un valor *missing* es aquel que no está informado en la BBDD.

Motivos:

- No se dispone de la información.
- El valor no informado tiene algún significado.
- Pérdida del dato por un error humano.

¿Cómo detectarlos?

Deben detectarse en la fase descriptiva con las técnicas vistas anteriormente.

Imputación de *missings*

¿Qué hacer con ellos?

- *Si la variable es de cualitativa o cuantitativa discreta*: Imputar el valor más frecuente en variables de clase.
- *Si la variable es continua*: Imputar alguna medida de centralización (media, mediana o moda) o Imputar de acuerdo a la distribución de los datos para no sesgar el resultado.

Existen algunas algoritmos de modelización que tratan a los valores *missing* como un valor más.

Debe tenerse en cuenta que, en ocasiones, dicho valor puede ser interpretable.

Creación de nuevas variables

A partir de las variables existentes se crean nuevas:

- Transformaciones funcional variables continuas:

$$edad2 = edad^2$$

$$nep\_edad = \ln(edad)$$

$$\text{sqrt}\_edad = \sqrt{edad}$$

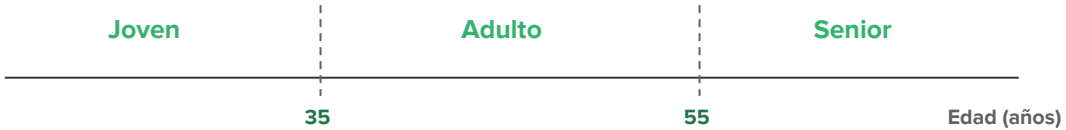
- Normalización de variables continuas (z – score):

$$Z = \frac{x - \mu}{\sigma}$$

Creación de nuevas variables

- Agrupación de categorías en variables cualitativas o cuantitativas discretas.

- Categorización de variables continuas



Visión global

Análisis de conjunto de datos

Comprender la naturaleza del target a predecir y de las variables explicativas de las que se dispone.

Muestreo

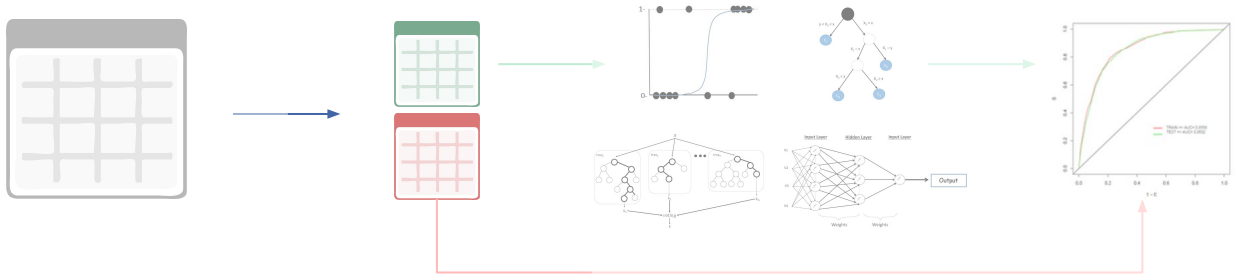
A partir del conjunto de datos, se define uno de entrenamiento y otro de test.

Selección del algoritmo

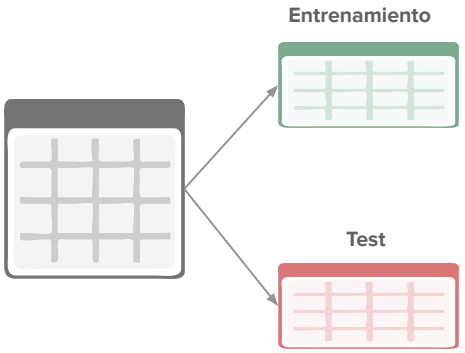
Se selecciona uno o varios algoritmos de modelización y se ajusta el modelo.

Validación

Se evalúa la bondad y precisión del modelo.



Conjunto de entrenamiento y test



Conjunto de datos a partir del cual va a aprender el modelo.

Este conjunto de datos no participa en el proceso de entrenamiento. Sirve para evaluar la precisión y bondad del modelo.

50 - 50. En caso de disponer de observaciones suficientes, normalmente el 50% de los datos disponibles se destinan al conjunto de entrenamiento y el otro 50% al de test.

Otros repartos comunes suelen ser: **60 – 40** o **70 – 30**.

Es común que se consideren varios a la hora de modelizar.

¿Cómo realizar el reparto?

Muestreo Probabilístico es una técnica en la que se seleccionan a los individuos de manera aleatoria.

Muestreo No probabilístico es una técnica de muestreo en la cual el investigador selecciona muestras basadas en un juicio subjetivo en lugar de hacer la selección al azar.

Tipos de muestreo probabilísticos

Aleatorio simple. Cada observación tiene la misma probabilidad de ser incluida en la muestra de estudio. Las observaciones se escogen al azar, sin utilizar ningún otro criterio.

Estratificado. Se divide la población en estratos y luego se selecciona aleatoriamente unos individuos de cada estrato para formar toda la muestra del estudio. Así que habrá como mínimo un integrante de cada estrato en la muestra.

Sistemático. Se selecciona la primera observación de manera aleatoria, y el resto de observaciones de la muestra se seleccionan utilizando un intervalo fijo.

Qué hacer cuando el evento es raro

Puede ocurrir que el evento a predecir sea raro o tenga poca frecuencia. Buenos ejemplos son enfermedades raras, conversiones en una página web u operaciones fraudulentas.

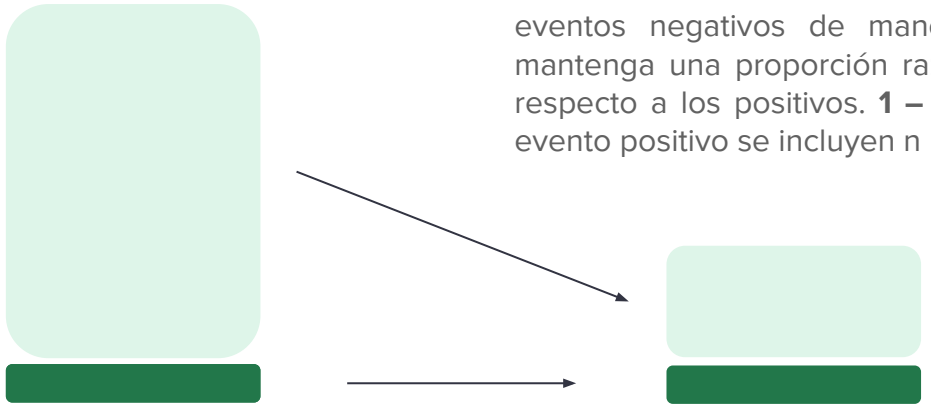
Se plantean dos problemáticas:

- El número de eventos negativos es muy superior al de positivos. El modelo no aprende a estimar los eventos positivos.

- No es posible crear un conjunto de test para la validación del modelo.

Qué hacer cuando el evento es raro

- *El número de eventos negativos es muy superior al de positivos.*

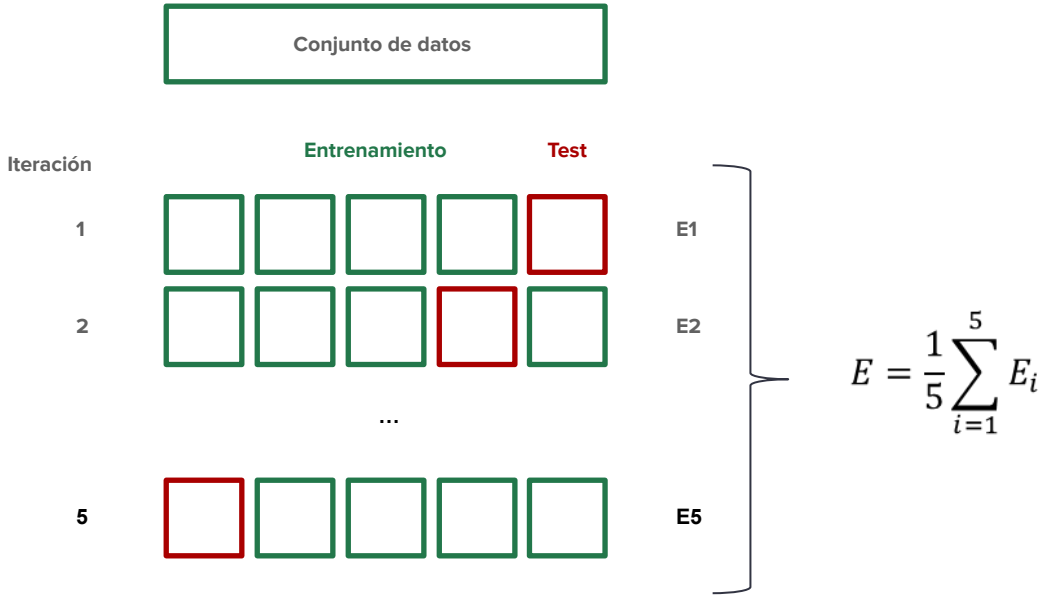


Se seleccionan una muestra aleatoria de eventos negativos de manera que se mantenga una proporción razonable con respecto a los positivos. **1 – n** (por cada evento positivo se incluyen n negativos).

Se seleccionan todos los eventos positivos

Qué hacer cuando el evento es raro

- *No es posible crear un conjunto de test para la validación del modelo □* ***Validación cruzada***



$$E = \frac{1}{5}\sum_{i=1}^{5} E_i$$

Visión global

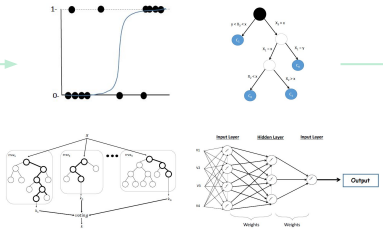
Análisis de conjunto de datos

Comprender la naturaleza del target a predecir y de las variables explicativas de las que se dispone.

Muestreo

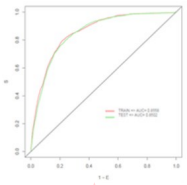
A partir del conjunto de datos, se define uno de entrenamiento y otro de test.

Selección del algoritmo



Se selecciona uno o varios algoritmos de modelización y se ajusta el modelo.

Validación



Se evalúa la bondad y precisión del modelo.

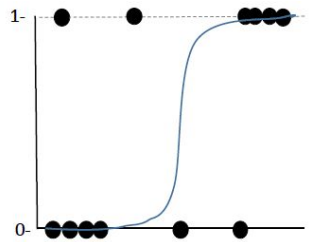
Selección del algoritmo

¿Qué factores hay que tener en cuenta?

- Naturaleza del target
- Necesidad de explicar el modelo
- Precisión del modelo
- Coste computacional
- Rapidez en la ejecución
- Necesidad de trabajar valores *outliers* y *missing*

Principales algoritmos

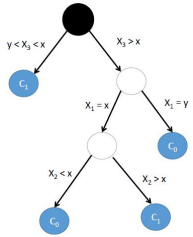
Regresión logística



- Es posible interpretar el peso y la importancia de las variables.
- Bajo coste computacional.

- Sensible a *outliers* y no acepta valores missing.
- Menor precisión y bondad de ajuste.

Árbol de decisión

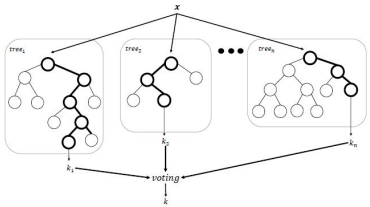


- Es posible interpretar y sencillo de explicar.
- Bajo coste computacional.

- Cuidado con el sobreajuste.
- Menor precisión y bondad de ajuste.

Principales algoritmos

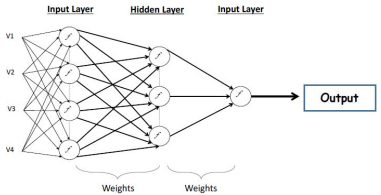
Random forest



- Mayor precisión.

- No es interpretable.
- Cuidado con el sobreajuste.
- Alto coste computacional.

Red neuronal

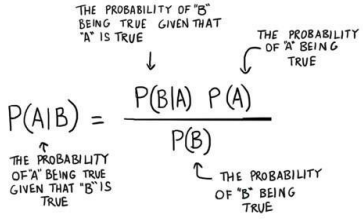


- Mayor precisión.

- No es interpretable.
- Cuidado con el sobreajuste.
- Alto coste computacional.

Principales algoritmos

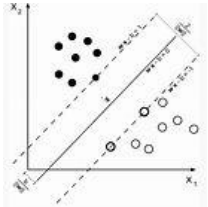
Naive Bayes



- Simple de entender e implementar.
- Muy rápido en la ejecución.

- Si no se cumplen ciertas hipótesis el ajuste suele ser pobre.

Support Vector Machine



- Mayor precisión.

- No es interpretable.
- Cuidado con el sobreajuste.
- Alto coste computacional.

Visión global

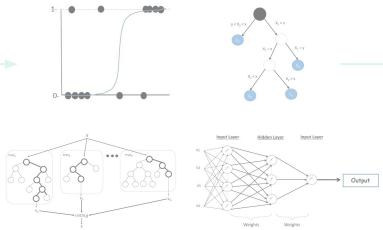
Análisis de conjunto de datos

Comprender la naturaleza del target a predecir y de las variables explicativas de las que se dispone.

Muestreo

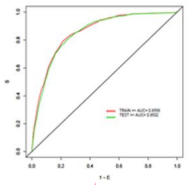
A partir del conjunto de datos, se define uno de entrenamiento y otro de test.

Selección del algoritmo



Se selecciona uno o varios algoritmos de modelización y se ajusta el modelo.

Validación



Se evalúa la bondad y precisión del modelo.

Validación del algoritmo

Regresión:

R^2 o coeficiente de determinación. R^2 ajustado.

Error Estándar de la regresión.

Clasificación:

Matriz de confusión.

Curva ROC y AUC (área bajo la curva).

R^2 o coeficiente de determinación.

Métrica de **validación** utilizada en algoritmos de **regresión**.

Se interpreta como la **variabilidad explicada** por el modelo.

Cuanto más **próximo a 1 mejor** será el ajuste del modelo.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \varepsilon_i^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} = 1 - \frac{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

Error estándar de la regresión

Métrica de **validación** utilizada en algoritmos de **regresión**.

Indica cuanto se equivoca el modelo, por término medio, al estimar el valor del target Y en función de la estimación del modelo a partir de las variables explicativas.

Cuanto más pequeño sea el valor menor será el error cometido en la estimación.

$$S_{x,y} = \sqrt{\frac{\sum_{i=1}^{n} \varepsilon_i^2}{n-2}}$$

Matriz de confusión

Métrica de **validación** utilizada en algoritmos de **clasificación**.

		Real	
		Enfermo	Sano
Estimado	Enfermo	**153**	**7**
	Sano	**8**	**268**

Sensibilidad = *% de enfermos que acierta el modelo* = 153 / 161 = 95%

Especificidad = *% de sanos que acierta el modelo* = 268 / 275 = 97%

Precisión = *% de aciertos* = (153 + 268) / (153 + 7 + 8 + 268) = 97%

Curva ROC y Área Bajo la Curva

Métrica de **validación** utilizada en algoritmos de **clasificación binaria**.

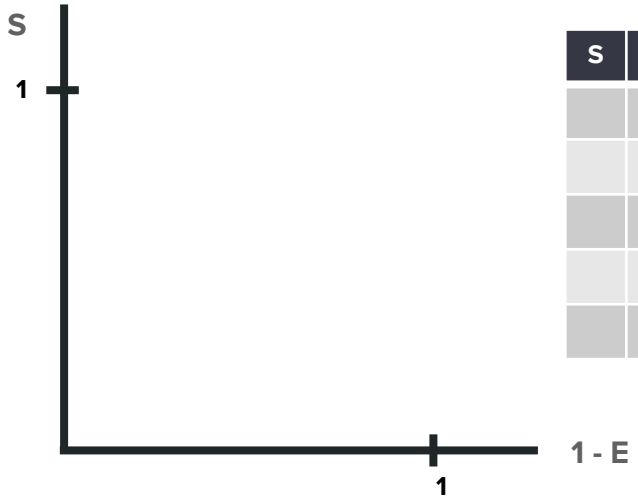
ROC son las siglas de **R**eceiver **O**perating **C**haracteristic

Curva ROC y Área Bajo la Curva

P	Nivel de Ácido	Cáncer	Estimado
0,24	40	0	
0,32	60	0	
0,42	80	0	
0,52	100	1	
0,62	120	0	
0,71	140	1	
0,78	160	1	
0,84	180	1	
0,89	200	0	
0,92	220	1	

Probabilidad Umbral:

		Real	
		1	0
Predicho por el modelo	1		
	0		



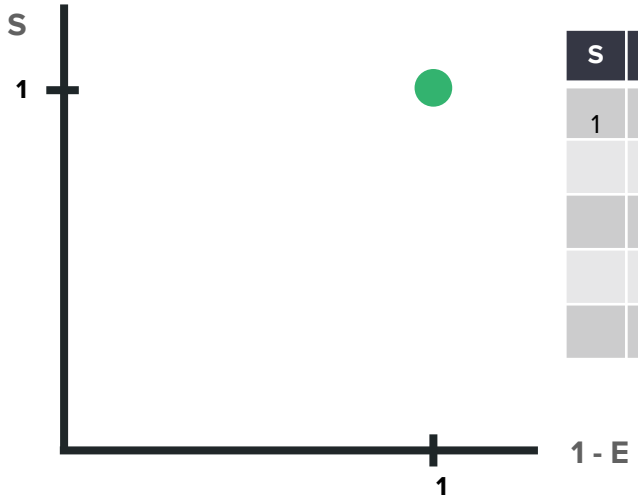
S	1 - E

Curva ROC y Área Bajo la Curva

P	Nivel de Ácido	Cáncer	Estimado
0,24	40	0	1
0,32	60	0	1
0,42	80	0	1
0,52	100	1	1
0,62	120	0	1
0,71	140	1	1
0,78	160	1	1
0,84	180	1	1
0,89	200	0	1
0,92	220	1	1

Probabilidad Umbral: 0

		Real	
		1	**0**
Predicho por el modelo	**1**	5	5
	0	0	0



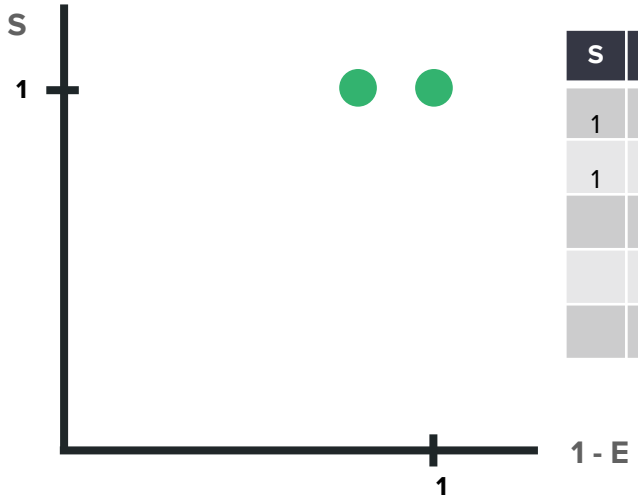
S	1 - E
1	1

Curva ROC y Área Bajo la Curva

P	Nivel de Ácido	Cáncer	Estimado
0,24	40	0	0
0,32	60	0	1
0,42	80	0	1
0,52	100	1	1
0,62	120	0	1
0,71	140	1	1
0,78	160	1	1
0,84	180	1	1
0,89	200	0	1
0,92	220	1	1

Probabilidad Umbral: 0,3

		Real	
		1	**0**
Predicho por el modelo	**1**	5	4
	0	0	1



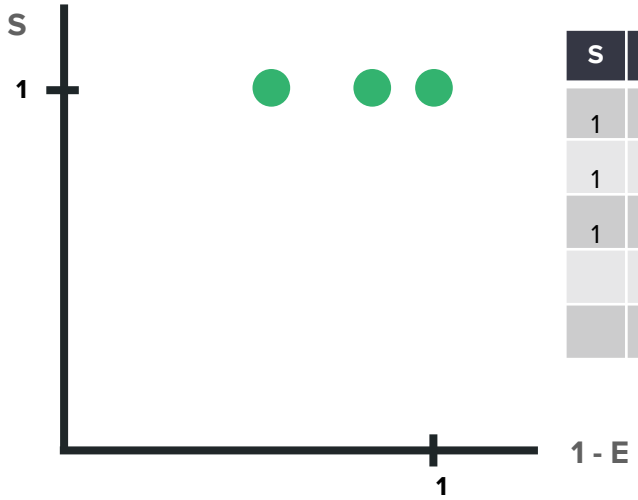
S	1 - E
1	1
1	0,8

Curva ROC y Área Bajo la Curva

P	Nivel de Ácido	Cáncer	Estimado
0,24	40	0	0
0,32	60	0	0
0,42	80	0	0
0,52	100	1	1
0,62	120	0	1
0,71	140	1	1
0,78	160	1	1
0,84	180	1	1
0,89	200	0	1
0,92	220	1	1

Probabilidad Umbral: 0,5

		Real	
		1	**0**
Predicho por el modelo	**1**	5	2
	0	0	3



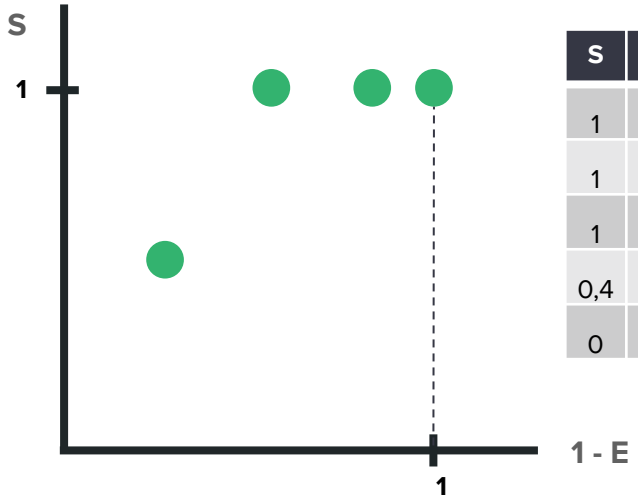
S	1 - E
1	1
1	0,8
1	0,4

Curva ROC y Área Bajo la Curva

P	Nivel de Ácido	Cáncer	Estimado
0,24	40	0	0
0,32	60	0	0
0,42	80	0	0
0,52	100	1	0
0,62	120	0	0
0,71	140	1	0
0,78	160	1	0
0,84	180	1	1
0,89	200	0	1
0,92	220	1	1

Probabilidad Umbral: 0,8

		Real	
		1	**0**
Predicho por el modelo	**1**	2	1
	0	3	4



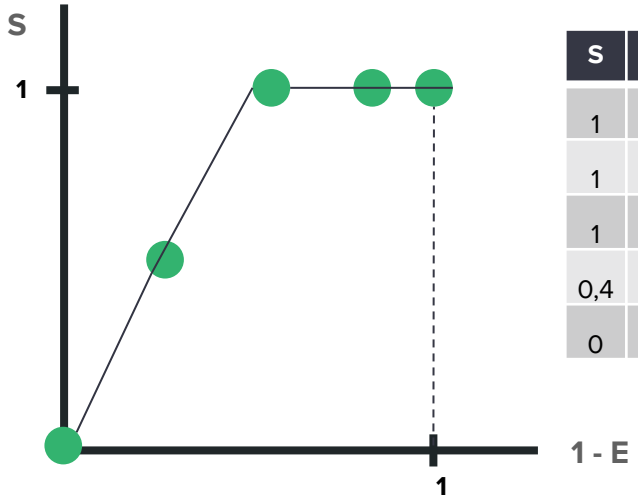
S	1 - E
1	1
1	0,8
1	0,4
0,4	0,2
0	0

Curva ROC y Área Bajo la Curva

P	Nivel de Ácido	Cáncer	Estimado
0,24	40	0	0
0,32	60	0	0
0,42	80	0	0
0,52	100	1	0
0,62	120	0	0
0,71	140	1	0
0,78	160	1	0
0,84	180	1	0
0,89	200	0	0
0,92	220	1	0

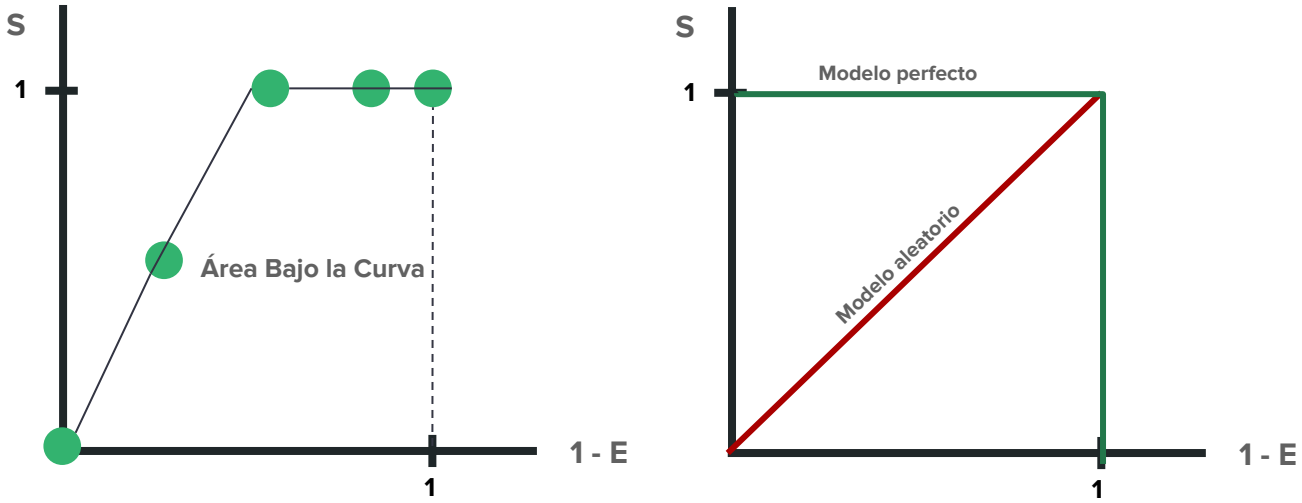
Probabilidad Umbral: 1

		Real	
		1	0
Predicho por el modelo	1	0	0
	0	5	5



S	1 - E
1	1
1	0,8
1	0,4
0,4	0,2
0	0

Área Bajo la Curva



Cuanto más **próximo a 1 mejor** será el ajuste del modelo.

Un **modelo aleatorio** tiene un área bajo la curva igual a **0,5**.

Esta métrica permite **comparar** varios modelos entre sí.

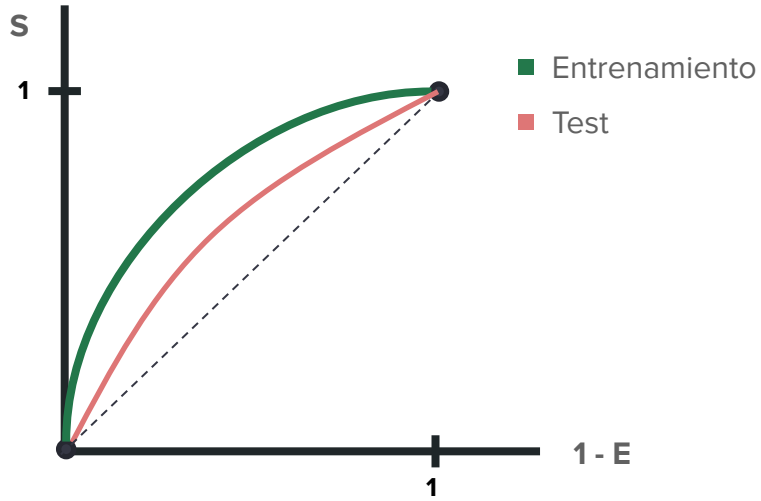
¿Sobreajuste?

Cuando el modelo aprende en exceso del conjunto de entrenamiento y ofrece una precisión pobre en el de test se dice que el modelo está **sobreajustado**.

Esto implica que el modelo obtenido **no es generalizable** y por lo tanto no va a predecir correctamente en conjuntos de datos que no participan en el proceso de entrenamiento.

Cómo detectarlo

A partir de la curva ROC y el área bajo la curva.



Cómo resolverlo

- Modificando los hiperparámetros de un algoritmo.

- Probando con otro algoritmo de modelización supervisada.

- Reducir la variabilidad de las variables mediante categorizaciones de variables continuas o agrupación de categorías.

- Realizando un muestreo distinto para obtener los conjuntos de entrenamiento y test.

Conclusiones

El proceso de modelización es mitad ciencia, mitad arte. La habilidad del *data scientist* para validar e interpretar los datos resulta clave para ajustar un modelo robusto.

Lo que diferencia a un *data scientist* senior de un junior es la capacidad crítica para evaluar si los datos disponible para ajustar un modelo son correctos.

¡Gracias!

Contacto:

juanjo.silva.t@gmail.com