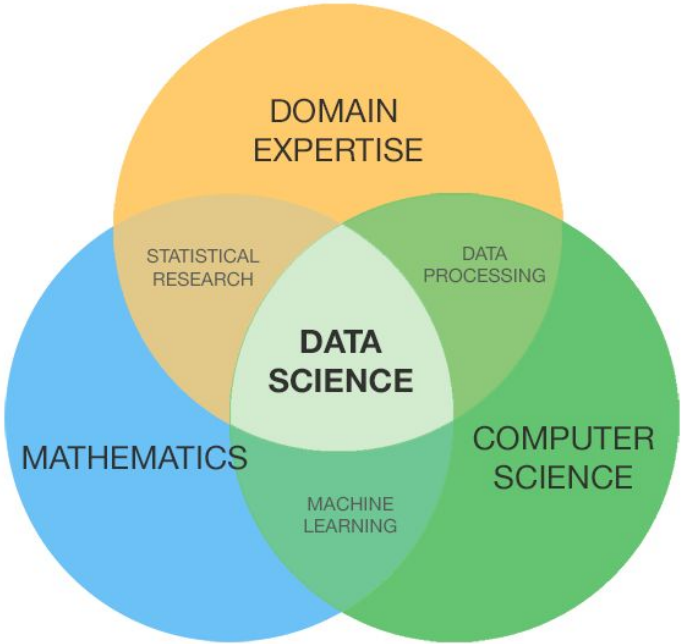
Aprendizaje Automático

Fabián Villena

Ciencia de datos

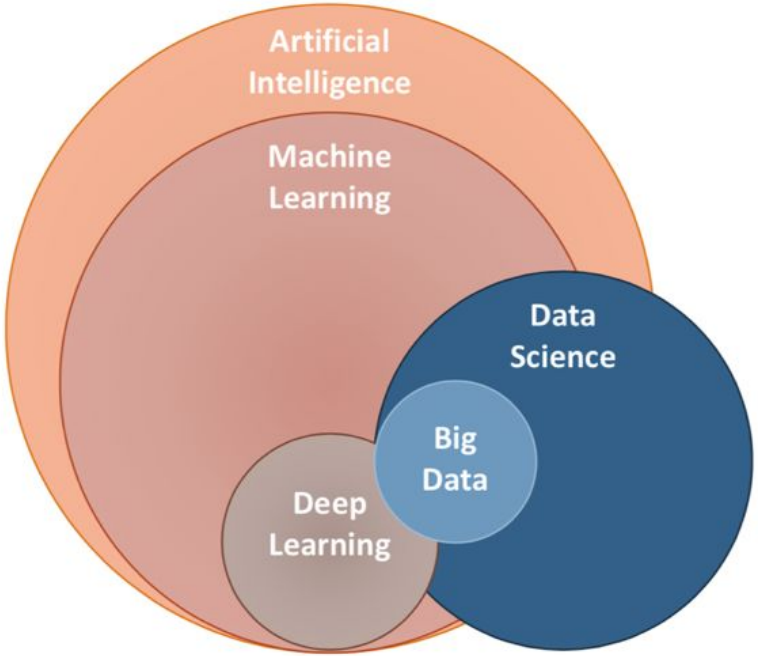
La ciencia de datos es un concepto que contempla el uso de métodos estadísticos, aprendizaje de máquinas y el **conocimiento específico de un área** para entender fenómenos desde los datos.



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

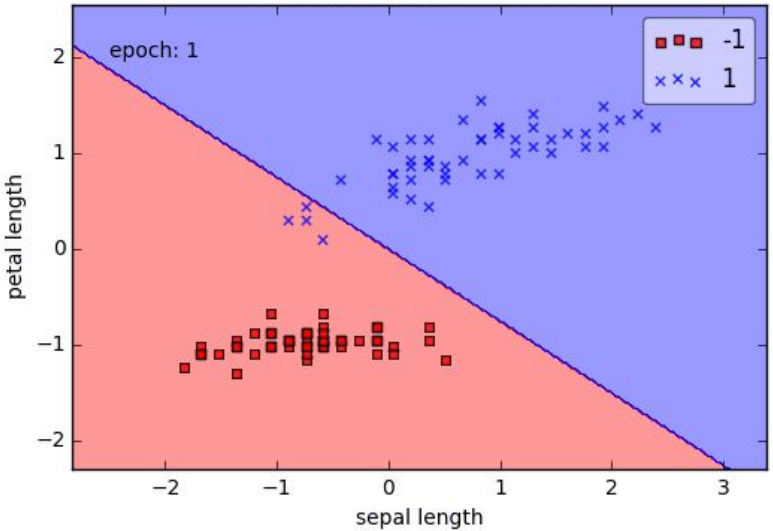
Inteligencia Artificial

- Es la clasificación más general de un conjunto de metodologías para la **generación de modelos**.
- Estos modelos pueden ser utilizados para la toma de decisiones.
- El método más básico es la generación de una serie de reglas para modelar un fenómeno.



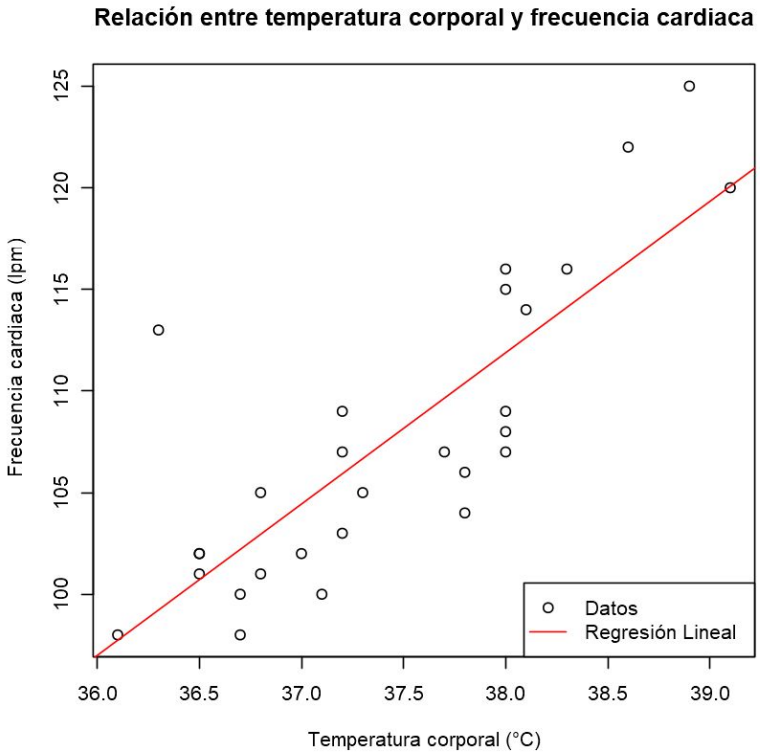
Aprendizaje Automático

- El aprendizaje automático es el estudio de algoritmos que automáticamente **mejoran su rendimiento a través de la experiencia**.
- Estos algoritmos construyen modelos basados en datos de muestra con la intención de realizar predicciones sin ser explícitamente programados para hacerlo.



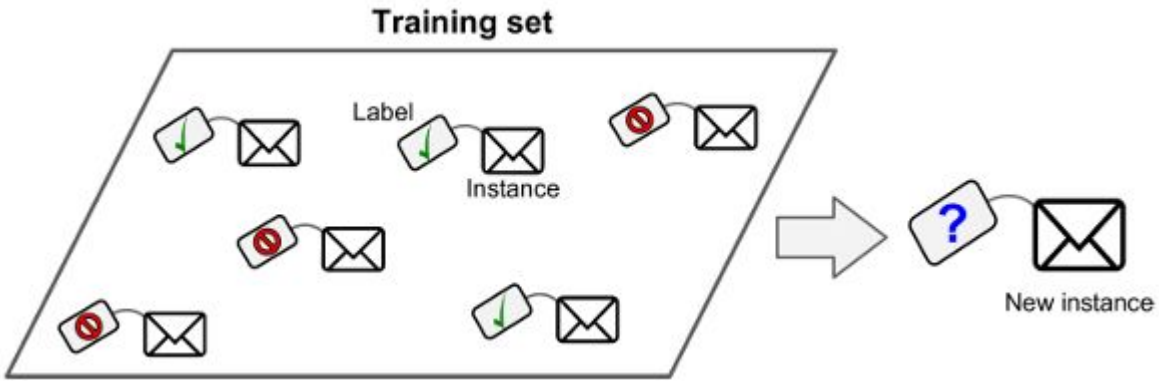
Aprendizaje supervisado

Este tipo de aprendizaje tiene la tarea de **aprender una función** que estime una salida dada una serie de características. Se infiere una función desde **datos de entrenamiento previamente etiquetados**.



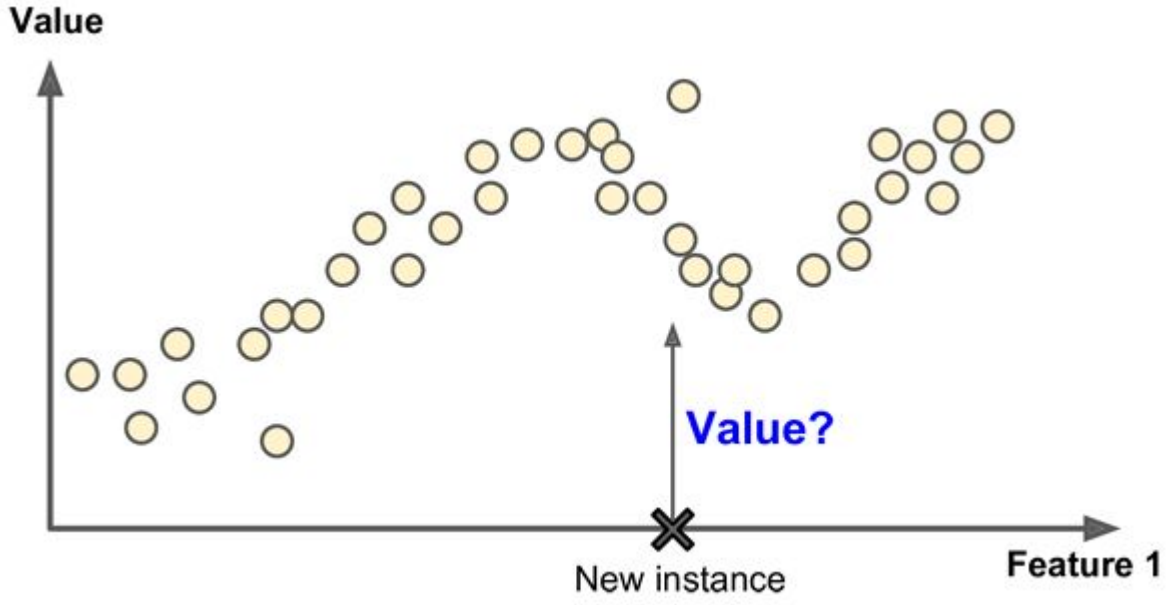
Clasificación

Esta tarea es una de las más comunes junto a la regresión. En esta tarea **buscamos predecir la clase** a la cual pertenece un objeto.



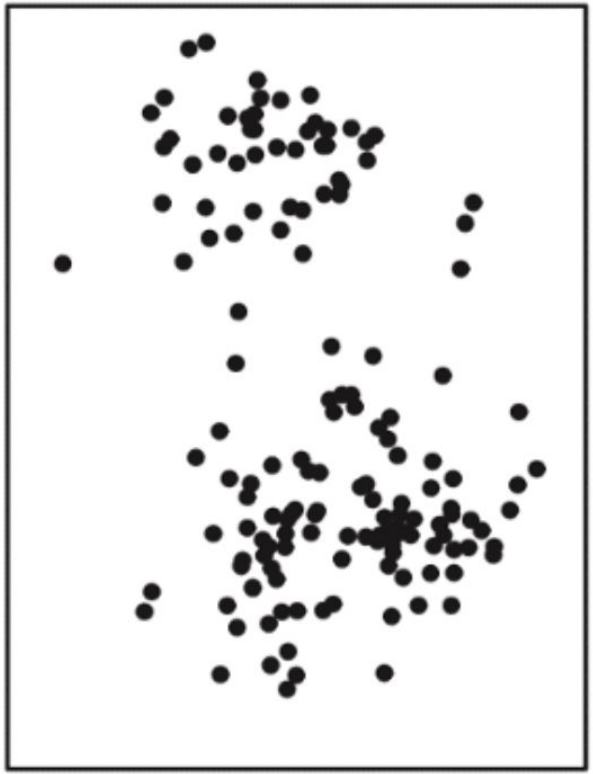
Regresión

Esta tarea es una de las más comunes junto a la clasificación. En esta tarea **buscamos predecir un valor numérico** asociado al objeto.



Aprendizaje no supervisado

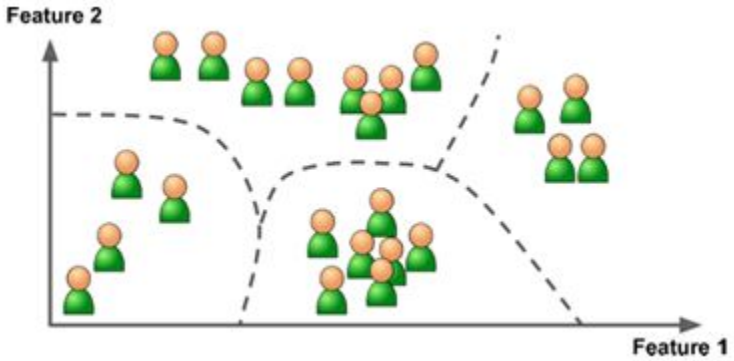
En este tipo de aprendizaje de máquinas se busca **detectar patrones en conjuntos de datos**, sin tener etiquetas de datos previas. Principalmente se utiliza en etapas exploratorias de investigación.



Clustering

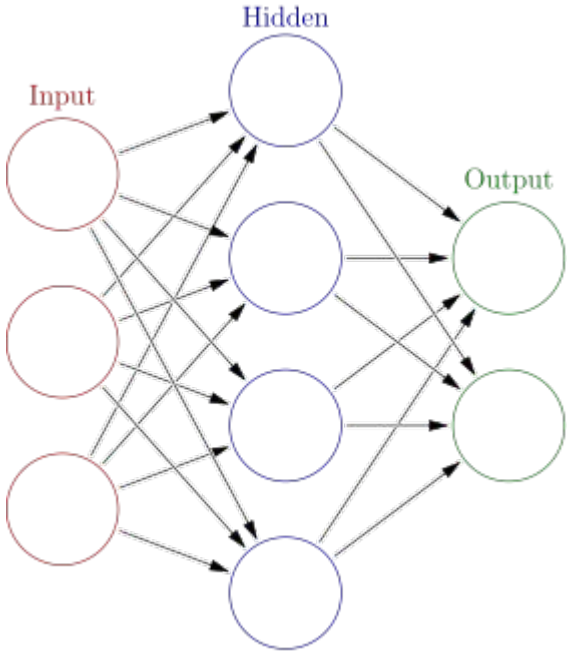
Es la tarea que busca grupos de objetos similares dentro de un conjunto de datos.

En ningún momento se le comunica al algoritmo a qué grupo pertenece cada objeto (porque no lo sabemos), sino que detecta agrupaciones sin supervisión.



Deep Learning

- Estos métodos están dentro del aprendizaje automático.
- Se realiza el ajuste del modelo a través de sus unidades mínimas llamadas **neuronas**, las cuales están conectadas entre sí.
- Estas conexiones están determinadas por pesos que se ajustan para optimizar el rendimiento.



Datos

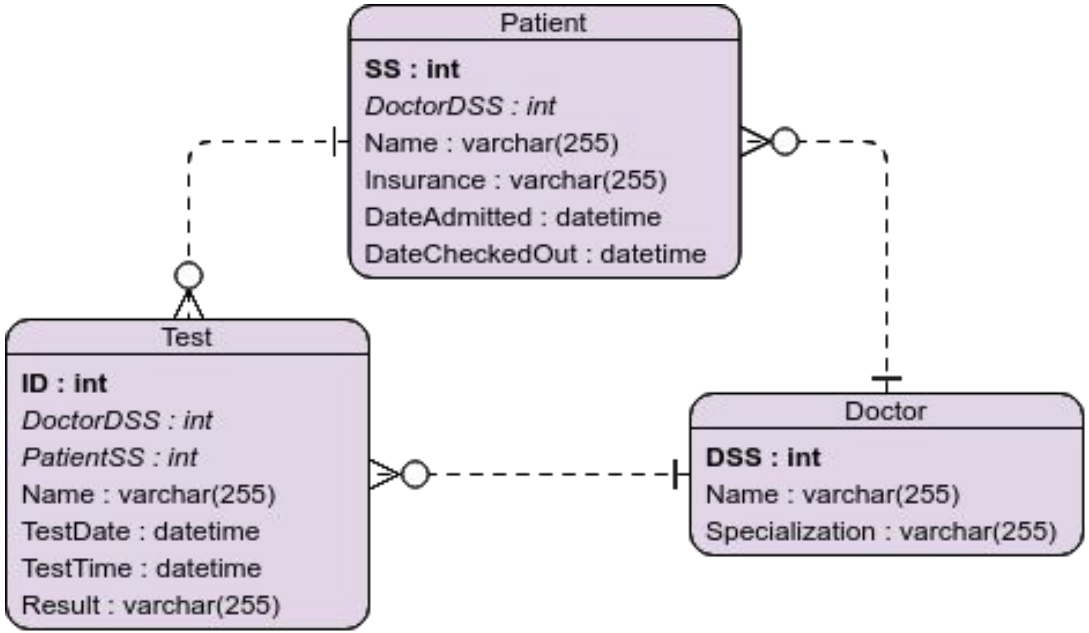
- Los datos son características de objetos coleccionadas a través de la observación.
- **No son información**, esta es producto de un análisis.
- Un conjunto de datos típicamente se operacionaliza como un conjunto de instancias de datos de una misma clase. Estas instancias cuentan con atributos y valores.

Paciente:
    id: <número>
    Nombre: <texto>
    Fecha de Nacimiento: <fecha>
    Salario: <número>
    COPD: <número>

id	**Nombre**	**...**	**COPD**
1	Juan	...	10
2	Pedro	...	8
...			

Datos estructurados

Con datos estructurados, nos referimos a que bajo el conjunto de datos **existe un modelo abstracto que estandariza** (modelo de datos) los valores de cada uno de los atributos y cómo se relacionan entre sí.



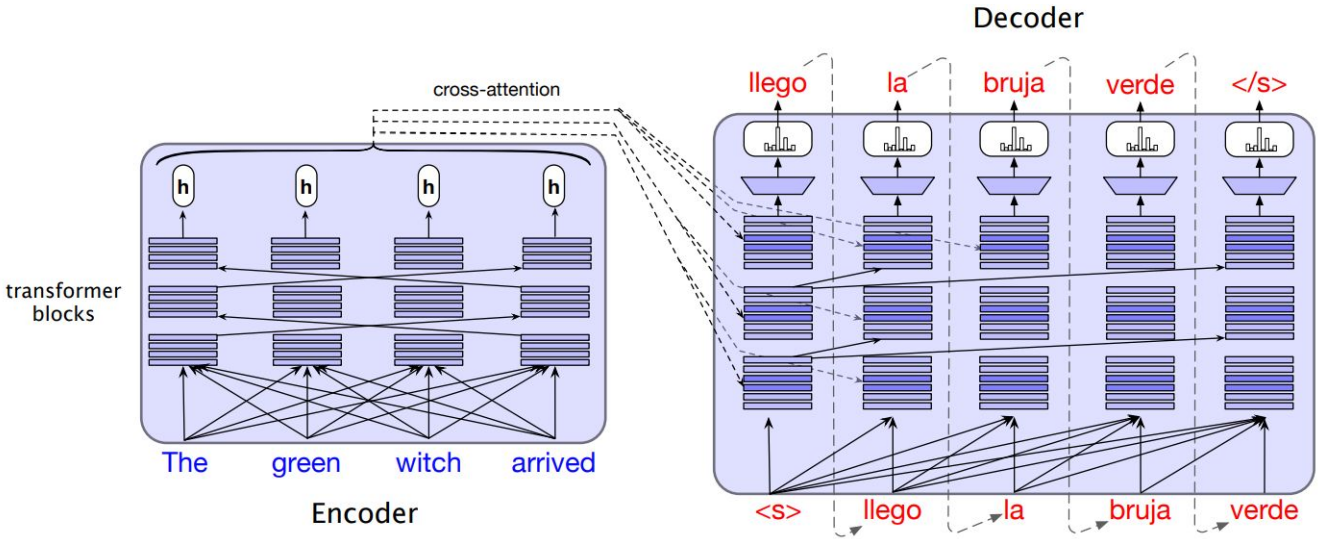
Datos no estructurados

Los datos no estructurados son aquellos que **no cuentan con un modelo de datos predefinido** o no están organizados de una manera predefinida. Esta característica genera irregularidades o ambigüedades que dificultan la extracción de información.

PCTE CON CUADROS DE PERICORONITIS RECURRENTE EN ZONA PZA 3.8 SEMIERUPCIONADA, SE RUEGA EVALUACION PARA EVENTUAL CIRUGIA DE EXODONCIA PZA 3.8 Y POSIBLEMENTE PZA 4.8

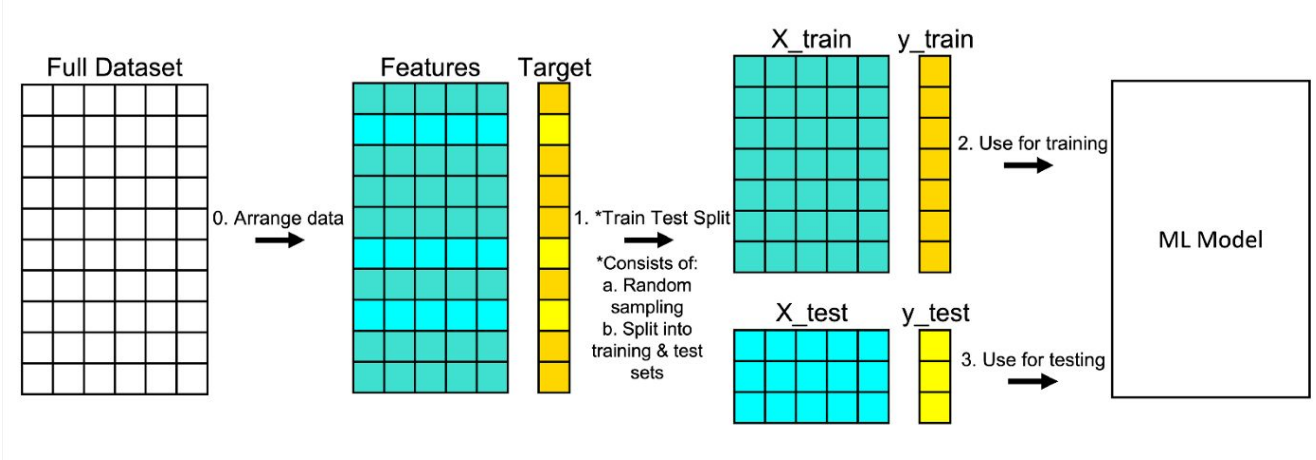
Para datos no estructurados usamos Deep Learning

En general, los modelos basados en Deep Learning se comportan mejor que modelos clásicos en tareas que utilizan datos no estructurados, tales como texto, audio, imágenes, grafos, etc.



Prueba y validación

Para saber qué tan bien generaliza un modelo hay que dividir el conjunto de datos en dos subconjuntos: Un subconjunto de entrenamiento, al cual ajustaré mi modelo y un subconjunto de prueba, con el cual se evaluará la generalización.



Métricas de rendimiento

Para poder entender el rendimiento de un modelo y poder valorar si está funcionando de manera correcta, debemos calcular un valor numérico. Cada métrica de rendimiento va a tener un significado específico y debemos interpretarlas de manera correcta.

	precision	recall	f1-score	support
0	0.77	0.86	0.81	37584
1	0.84	0.75	0.79	37577
accuracy			0.80	75161
macro avg	0.81	0.80	0.80	75161
weighted avg	0.81	0.80	0.80	75161

Matriz de confusión

Una manera de evaluar el rendimiento de un clasificador es observar la matriz de confusión. La idea general es contar la cantidad de veces que los ejemplos de la clase A son clasificados como clase B. Este cálculo normalmente se debe realizar en el subconjunto de prueba.

		Predicho	
		Negativo	Positivo
Real	Negativo	Verdaderos negativos (FN)	Falsos positivos (FN)
	Positivo	Falsos negativos (FN)	Verdaderos positivos (FN)

Accuracy

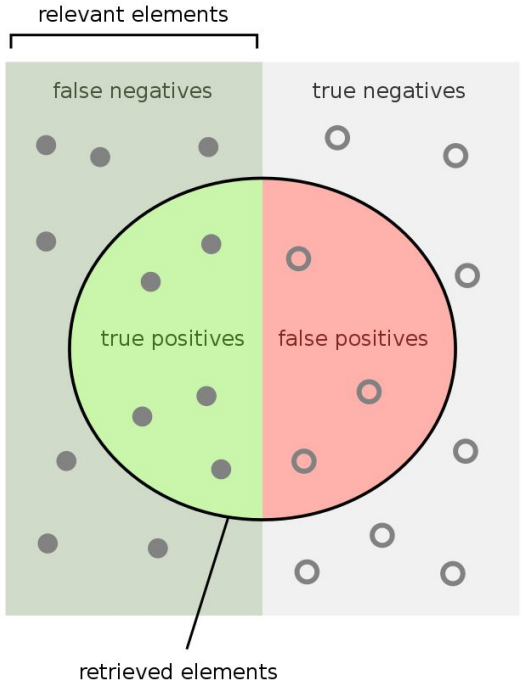
La *accuracy* es una de las métricas que podemos utilizar para evaluar el rendimiento de un modelo. La accuracy es la fracción de predicciones que nuestro modelo realizó correctamente. El problema es que en problemas desbalanceados, esta métrica puede llevarnos a falsas conclusiones.

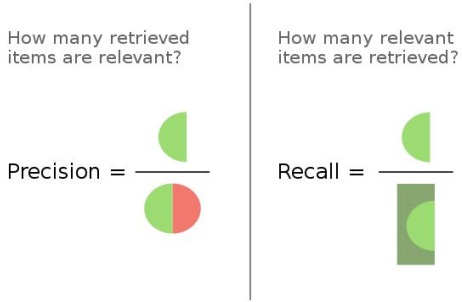
$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision y *recall*

- La *precision* busca identificar la proporción de predicciones positivas que realmente eran positivas.
- El *recall* busca identificar la proporción de elementos realmente positivos que se predijeron como positivos.

Para evaluar el rendimiento de nuestro modelo debemos evaluar ambas métricas. Típicamente existe un compromiso entre ambas métricas que debemos manejar con cuidado.





F-score

El F-score es una métrica resumen que nos comunica el resultado de la *precision* y el *recall* en un sólo valor. La versión más utilizada es el F_1-score, que pondera con el mismo peso la *precision* y el *recall*.

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Métricas de rendimiento para regresión

Las métricas más utilizadas para evaluar el rendimiento de un regresor son el *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE) y el *Root Mean Squared Error* (RMSE).

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} \left| y_i - \hat{y}_i \right|}{n}$$

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

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \left| (y_i - \hat{y}_i)^2 \right|}{n}}$$