DATA SCIENCE APLICADO A NEGOCIOS

- **Contratar y retener empleados son tareas extremadamente complejas que requieren capital, tiempo y habilidades.**

- **“Los propietarios de pequeñas empresas dedican el 40% de su jornada laboral a tareas que no generan ingresos como la contratación”.**

- **“Las empresas gastan entre el 15% y el 20% del salario del empleado para contratar un nuevo candidato”.**

Source: https://toggl.com/blog/cost-of-hiring-an-employee

- **“Una empresa promedio pierde entre el 1% y el 2,5% de sus ingresos totales en el tiempo que lleva poner al día a un nuevo empleado”.**

- **Contratar a un nuevo empleado cuesta un promedio de $7645 (empresa de entre 0 y 500 empleados).**

- **En promedio, se necesitan 52 días para cubrir un puesto.**

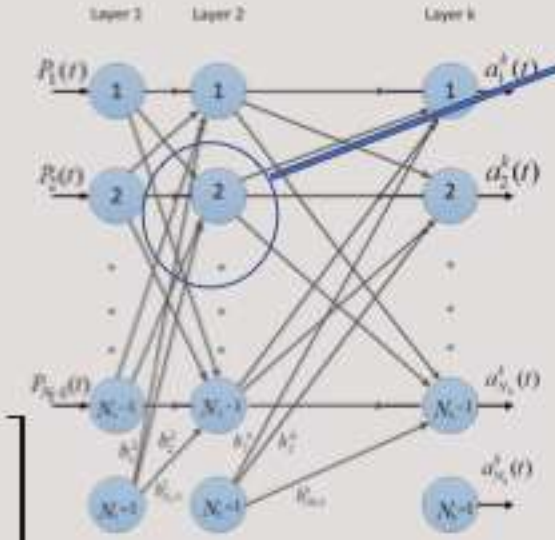
Source: https://toggl.com/blog/cost-of-hiring-an-employee

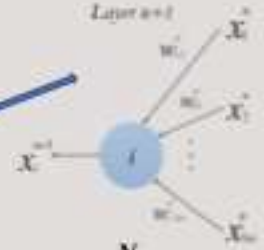
- Trabajas como científico de datos en una corporación multinacional.
- El equipo de recursos humanos recopiló una gran cantidad de datos sobre sus empleados y se acerca a ti para desarrollar un modelo que pudiera predecir qué empleados tienen más probabilidades de renunciar.
- El equipo nos ha proporcionado una gran cantidad de datos, aquí hay una muestra del conjunto de datos:
  - Participación laboral
  - Educación
  - Satisfacción laboral
  - Clasificación de Rendimiento
  - Satisfacción en las relaciones
  - Equilibrio trabajo-vida

Data Source: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

1. REDES NEURONALES ARTIFICIALES

$$P = \begin{bmatrix} P_1 \\ P_2 \\ \boxed{?} \\ P_{N_1} \end{bmatrix}$$





$$x_i^{n+1}(t) = \varphi\left(\sum_{j=1}^{N_n} w_{i,j}^n x_j^n(t)\right)$$

Node $(n+1, i)$ representation

$$\begin{bmatrix} W_{11} & W_{12} & \cdots & W_{1,\,N_1} \\ W_{21} & W_{22} & & W_{2,\,N_1} \\ \vdots & & \ddots & \vdots \\ W_{m-1,1} & W_{m-1,2} & \cdots & W_{m-1,\,N_1} \\ W_{m,1} & W_{m,2} & \cdots & W_{m,\,N_1} \end{bmatrix}$$

Función de Activación No Lineal Sigmoide

$$\varphi(w) = \frac{1}{1 + e^{-w}}$$

m: número de neuronas en la capa oculta

N_1: *número de entradas*

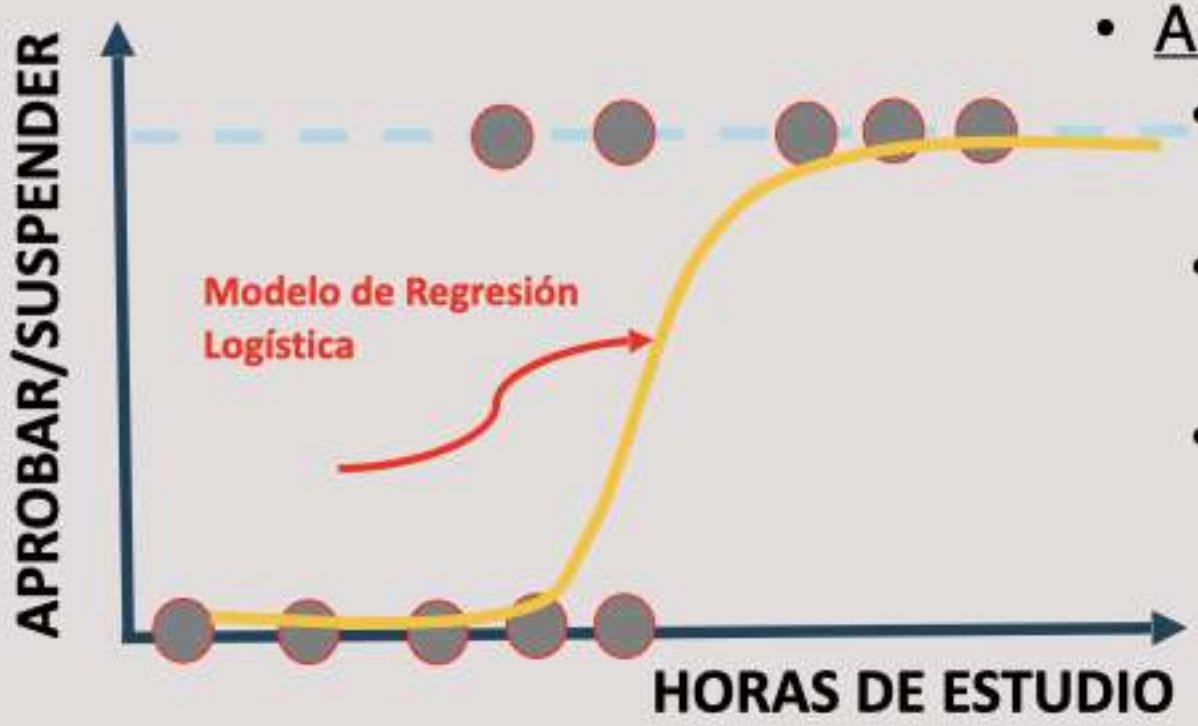
2. CLASIFICADOR DE REGRESIÓN LOGÍSTICA

- **La regresión lineal** se utiliza para predecir valores de salida en un espectro continuo.
  - Ejemplo: predecir los beneficios en función de la temperatura de aire exterior.
- **La regresión logística se utiliza para predicciones con salidas binarias** con dos posibles valores etiquetados con "0" o "1"
  - La salida de un modelo logístico puede ser de una de dos clases: aprobar/suspender, ganar/perder, sano/enfermo

Horas de Estudio	Aprobar/ Suspender
1	0
1.5	0
2	0
3	1
3.25	0
4	1
5	1
6	1

2. CLASIFICADOR DE REGRESIÓN LOGÍSTICA

- El algoritmo de regresión logística funciona implementando en primer lugar una ecuación lineal con predictores independientes para predecir un valor.
- Este valor luego se convierte en una probabilidad que podría oscilar entre los valores 0 y 1



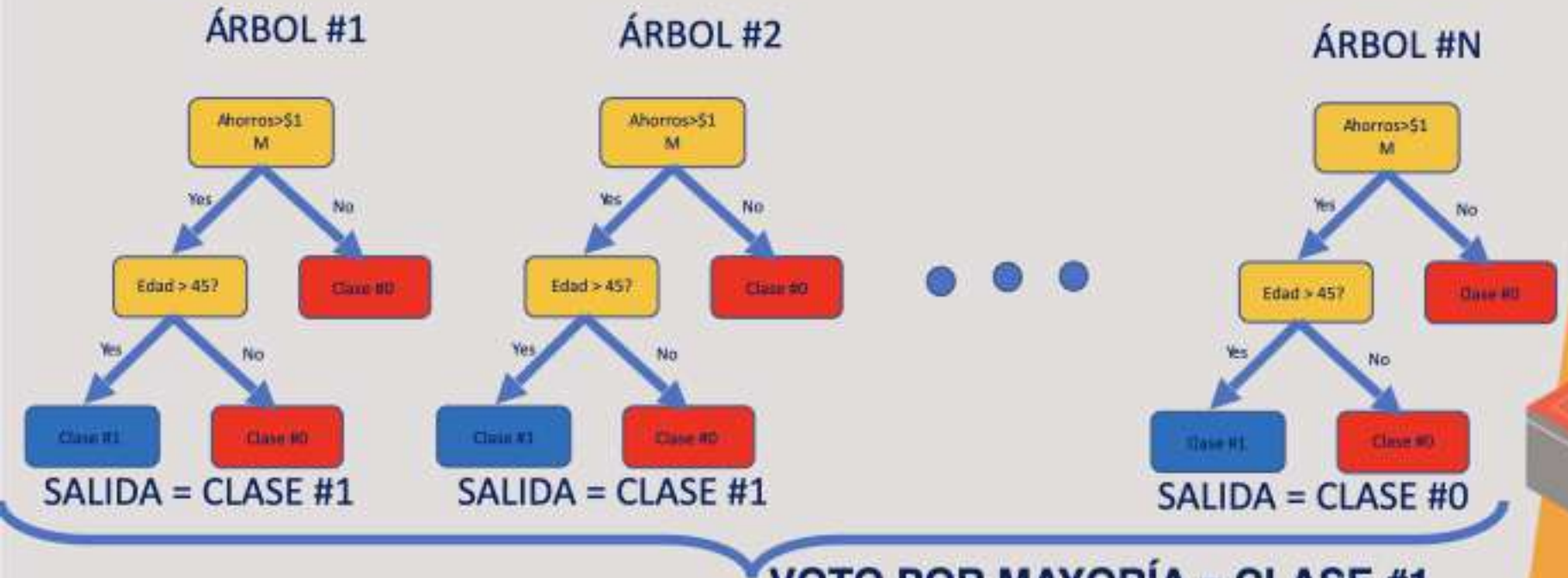
- Ecuación lineal:
  - $y = b_0 + b_1 * x$
- Aplicar la función sigmoide:
  - $P(x) = sigmoid\ (y)$
  - $P(x) = \dfrac{1}{1 + e^{-y}}$
  - $P(x) = \dfrac{1}{1 + e^{-(b_0 + b_1 * x)}}$

2. CLASIFICADOR DE REGRESIÓN LOGÍSTICA

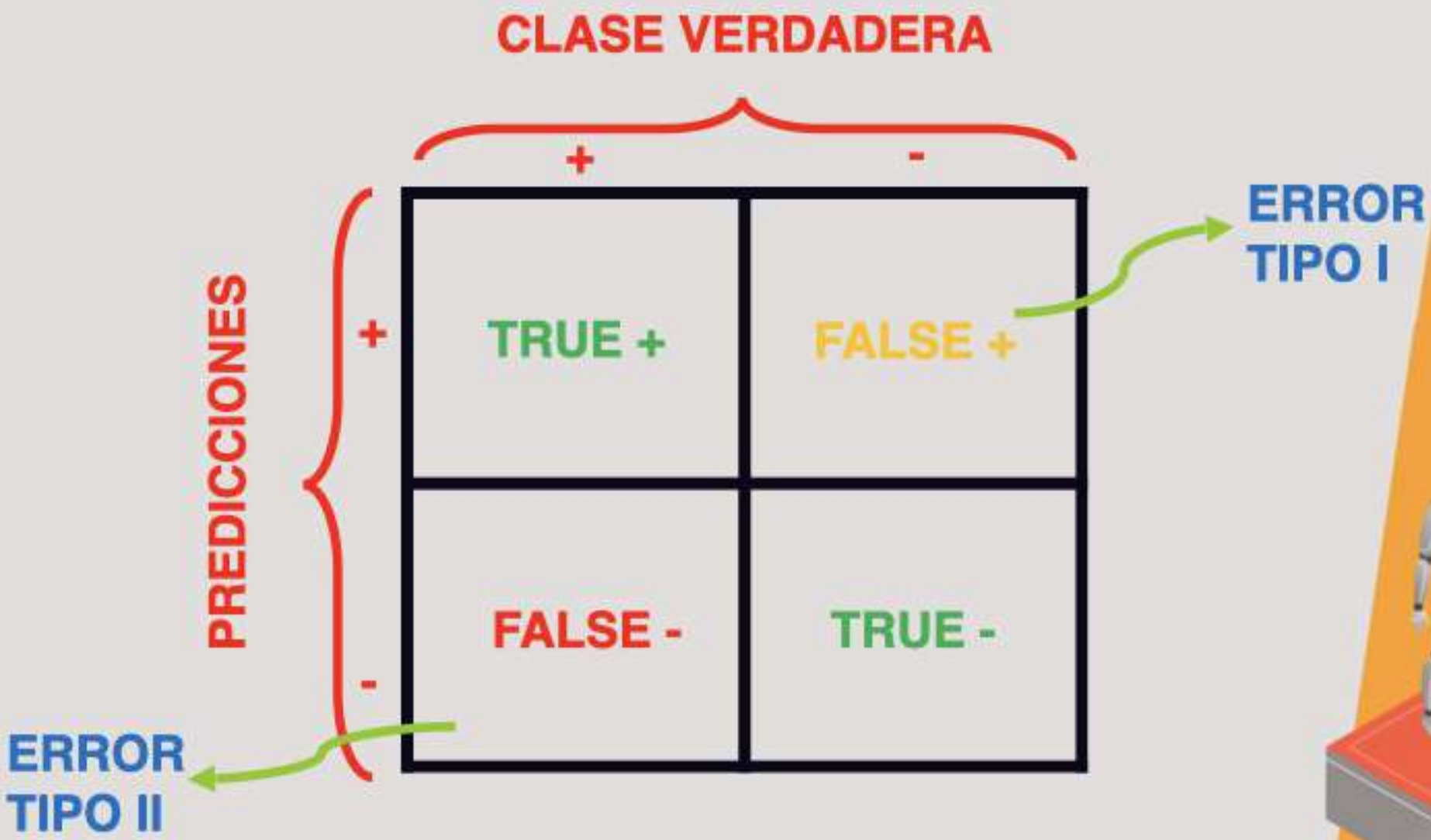
- **Luego necesitamos convertir la probabilidad a una de las dos categorías que son “0” o “1”.**

3. CLASIFICADOR DE BOSQUES ALEATORIOS

- Los árboles de decisión son técnicas de Machine Learning supervisadas donde los datos son divididos en función de cierta condición o parámetro.
- Por ejemplo, pongamos que queremos clasificar si un cliente se va a retirar en función de sus ahorros y su edad.
- El clasificador de bosque aleatorio es un tipo de algoritmo de ensamblado.
- Crea un conjunto de árboles de decisión a partir de un subconjunto seleccionado al azar del conjunto de entrenamiento.
- Luego combina el voto de diferentes árboles de decisión para decidir la clase final del objeto de prueba.



MATRIZ DE CONFUSIÓN



KPIS DEL MODELO DE CLASIFICACIÓN

- La matriz de confusión es una técnica utilizada para evaluar el resultado en un problema de clasificación

    - **Verdaderos positivos (TP):** casos donde el clasificador predice TRUE (predice enfermedad), y la clase verdadera es TRUE (el paciente tiene la enfermedad).
    - **Verdaderos negativos (TN):** casos donde el clasificador predice FALSE (predice no enfermedad), y la clase verdadera es FALSE (el paciente no tiene la enfermedad).
    - **Falsos positivos (FP) (Error de tipo I):** casos donde el clasificador predice TRUE, pero la clase correcta es FALSE (el paciente no tiene la enfermedad).
    - **Falsos negativos (FN) (Error de tipo II):** casos donde el clasificador predice FALSE (el paciente no tiene la enfermedad), pero si que la tiene.
    - **Acierto de la clasificación (Accuracy)** = (TP+TN) / (TP + TN + FP + FN)
    - **Precisión** = TP/Total TRUE Predictions = TP/ (TP+FP) (Cuando el modelo predice TRUE, ¿qué tan probable es que acierte?)
    - **Recall** = TP/ Actual TRUE = TP/ (TP+FN) (Cuando la clase verdadera es TRUE, ¿qué tan probable es que el modelo acierte?)

PRECISION VS. RECALL

HECHOS:
100 PACIENTES EN TOTAL
91 PACIENTES ESTÁN SANOS
9 PACIENTES TIENEN CANCER

CONDICIÓN VERDADERA

PREDICCIONES	+	-
+	TP = 1	FP = 1
-	FN = 8	TN = 90

- El acierto (accuracy) es generalmente engañosa y no es suficiente para evaluar el desempeño de un Clasificador..
- Recall es un KPI importante en situaciones donde:
  - El dataset está muy desbalanceado; es el caso donde hay muy pocos pacientes con cancer, comparados con los sanos.

- Accuracy = (TP+TN) / (TP + TN + FP + FN) = 91%
- Precision = TP/Total TRUE Predictions = TP/ (TP+FP) = ½=50%
- Recall = TP/ Actual TRUE = TP/ (TP+FN) = 1/9 = 11%

F1-SCORE

$$F1\ Score = \frac{2 * (PRECISION * RECALL)}{(PRECISION + RECALL)}$$

$$F1\ Score = \frac{2 * TP}{2 * TP + FP + FN}$$

- El F1 Score es una medida general de la precisión de un modelo que combina precisión y el recall.
- El F1 Score es la media armónica de precisión y el recall.
- ¿Cuál es la diferencia entre el F1 Score y la precisión?
- En conjuntos de datos desbalanceados, si tenemos una gran cantidad de verdaderos negativos (pacientes sanos), la precisión podría ser engañosa. Por lo tanto, el F1 Score podría ser un mejor KPI para usar, ya que proporciona un equilibrio entre el recall y la precisión en presencia de conjuntos de datos desbalanceados.