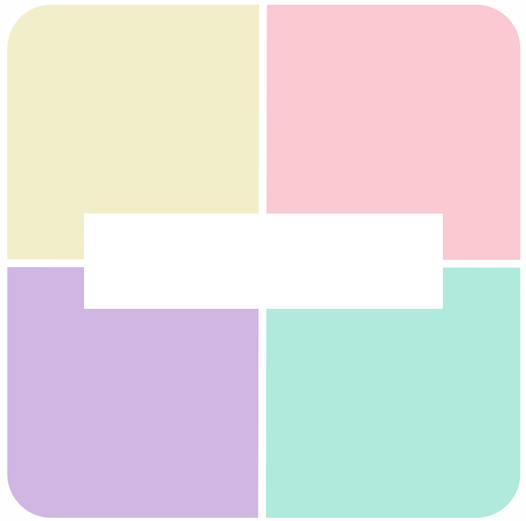
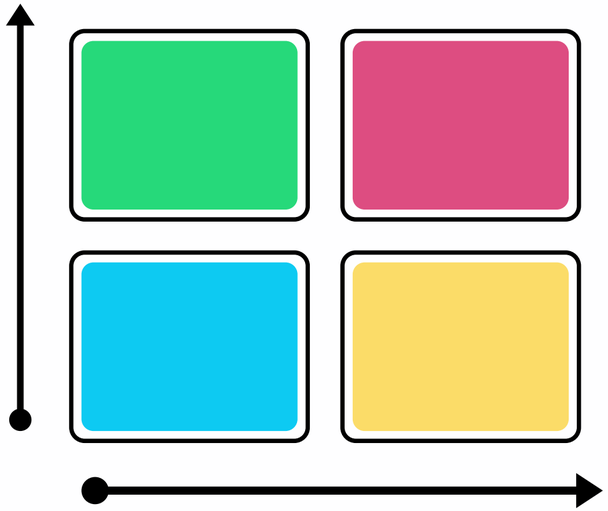


Matriz de Confusión



Supongamos que tenemos estos datos médicos:

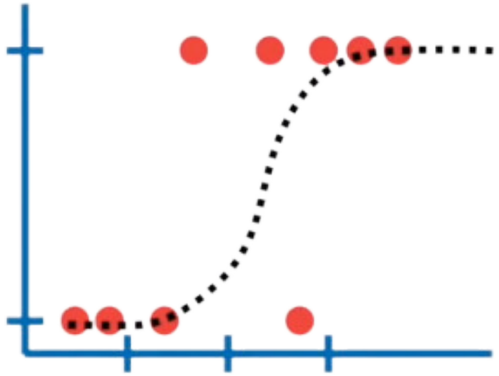
Dolor de Pecho	Buena Circulación Sanguínea	Arterias Bloqueadas	Peso	Enfermedad Cardíaca
No	No	No	57	No
Si	Si	Si	82	Si
Si	Si	No	95	No
...	...	...	...	...

Tenemos algunas medidas médicas como:

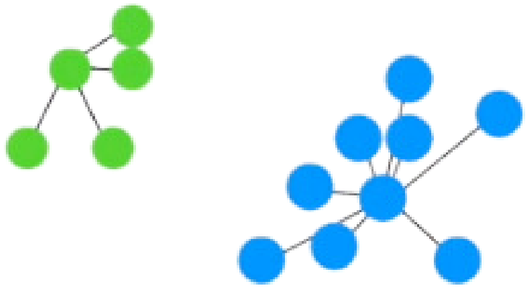
- Dolor de Pecho
- Buena Circulación Sanguínea
- Arterias Bloqueadas
- Peso

Con estos datos clínicos queremos aplicar un método de aprendizaje automático para predecir si alguien **desarrollará o no una enfermedad cardiaca.**

Para esto podríamos usar:



Regresión Logística



KNN

Random Forest

O algún otro método de los que existen (hay muchos algoritmos de clasificación).

¿Cómo podríamos decidir cuál de los algoritmos es el que funciona mejor con nuestros datos?

Primero dividimos nuestros datos en data de entrenamiento **(Training)** y data de prueba **(Testing)**

Dolor de Pecho	Buena Circulación Sanguínea	Arteries Bloqueadas	Peso	Enfermedad Cardíaca
No	No	No	57	No
...	...	...	...	...

Datos de Entrenamiento

Dolor de Pecho	Buena Circulación Sanguínea	Arteries Bloqueadas	Peso	Enfermedad Cardíaca
Yes	Yes	No	95	No
...	...	...	...	...

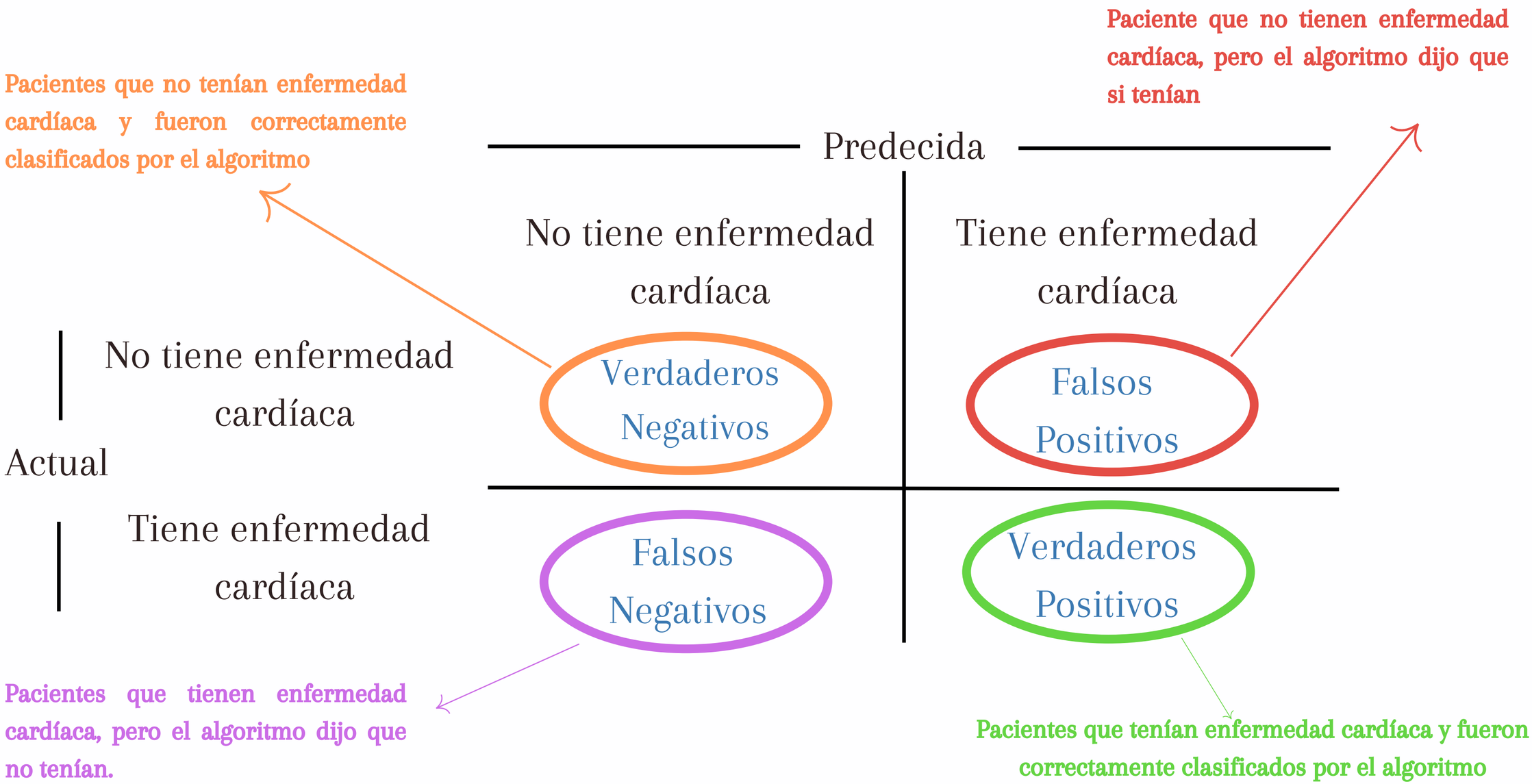
Datos de Prueba

Luego entrenamos todos los métodos que elegimos con la data de entrenamiento y testeamos cada método con la data de prueba.

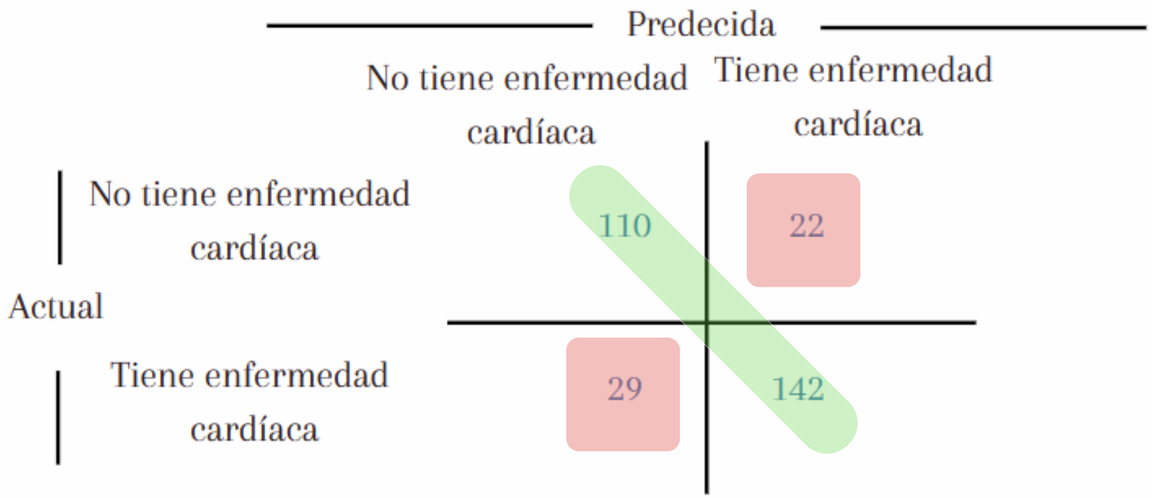
Ahora necesitamos resumir como cada método rindió al momento de pasarle la data de prueba:
Una forma de hacer esto es creando una **matriz de confusión** para cada método.
En general una matriz de confusión se ve así:

		Predecida	
		Clase Negativa	Clase Positiva
Actual	Clase Negativa	Verdaderos Negativos	Falsos Positivos
	Clase Positiva	Falsos Negativos	Verdaderos Positivos

En este caso la clase positiva serían las personas que tienen enfermedad cardíaca y la negativa las que no.


Pacientes que no tenían enfermedad cardíaca y fueron correctamente clasificados por el algoritmo
Paciente que no tienen enfermedad cardíaca, pero el algoritmo dijo que si tenían
Predecida
No tiene enfermedad cardíaca
Tiene enfermedad cardíaca
No tiene enfermedad cardíaca
Actual
Tiene enfermedad cardíaca
Verdaderos Negativos
Falsos Positivos
Falsos Negativos
Verdaderos Positivos
Pacientes que tienen enfermedad cardíaca, pero el algoritmo dijo que no tenían.
Pacientes que tenían enfermedad cardíaca y fueron correctamente clasificados por el algoritmo

Por ejemplo, cuando aplicamos el algoritmo Random Forest a la data de prueba, obtenemos la siguiente matriz de confusión:



- Los números a lo largo de la diagonal nos dicen cuántas veces se clasificaron correctamente las observaciones
- Los números que no están en la diagonal son observaciones clasificadas incorrectamente.

Podemos ver claramente que:

- Hubieron 142 Verdaderos Positivos, pacientes con enfermedad cardíaca que fueron correctamente clasificados
- Hubieron 110 Verdaderos Negativos, pacientes sin enfermedad cardíaca que fueron correctamente clasificados.
- El algoritmo clasifico erróneamente a 29 pacientes que tenían enfermedad cardíaca, diciendo que no tenían enfermedad cardíaca. (Falsos Negativos)
- El algoritmo clasificó erróneamente a 22 pacientes que no tenían enfermedad cardíaca, diciendo que si tenían enfermedad cardíaca. (Falsos Positivos)

Supongamos que otro algoritmo que usamos fue el KNN, entonces podemos comparar la matriz de confusión de Random Forest, con la matriz de confusión que obtuvimos al aplicar KNN.

Random Forest

Actual \ Predecida	No tiene enfermedad cardíaca	Tiene enfermedad cardíaca
No tiene enfermedad cardíaca	110	22
Tiene enfermedad cardíaca	29	142

KNN

Actual \ Predecida	No tiene enfermedad cardíaca	Tiene enfermedad cardíaca
No tiene enfermedad cardíaca	79	53
Tiene enfermedad cardíaca	64	107

Podemos ver que:

- **KNN tuvo un peor rendimiento que el Random Forest al predecir los pacientes con enfermedad cardíaca (107 vs 142)**
- **KNN también fue peor al predecir pacientes sin enfermedad cardíaca. (79 vs 110)**

Por lo que si tuvieramos que elegir entre KNN y Random Forest, elegiríamos Random Forest.

Supongamos que decidimos aplicar otro algoritmo para seguir comparando, en este caso aplicamos una Regresión Logística a nuestros datos de prueba, por lo cual obetenemos otra matriz de confusión.

Regresión Logística

		Predecida	
		No tiene enfermedad cardíaca	Tiene enfermedad cardíaca
Actual	No tiene enfermedad cardíaca	112	20
	Tiene enfermedad cardíaca	32	139

Random Forest

		Predecida	
		No tiene enfermedad cardíaca	Tiene enfermedad cardíaca
Actual	No tiene enfermedad cardíaca	110	22
	Tiene enfermedad cardíaca	29	142

En este caso vemos que ambas matrices de confusión son muy similares, lo que hace díficil el poder elegir que algoritmo es mejor para nuestros datos. Es aquí donde entran otras métricas más sofisticadas como la sensibilidad, especificidad, ROC y AUC que nos ayudarán en estos casos.

Ahora que hemos resuelto la matriz de confusión básica, veamos una más complicada.

Ahora tenemos este conjunto de datos, la pregunta es: ¿Basado en lo que las personas piensan sobre estas películas podemos usar un método de aprendizaje automático para predecir su película favorita?

Supongamos que en este caso las únicas opciones son:

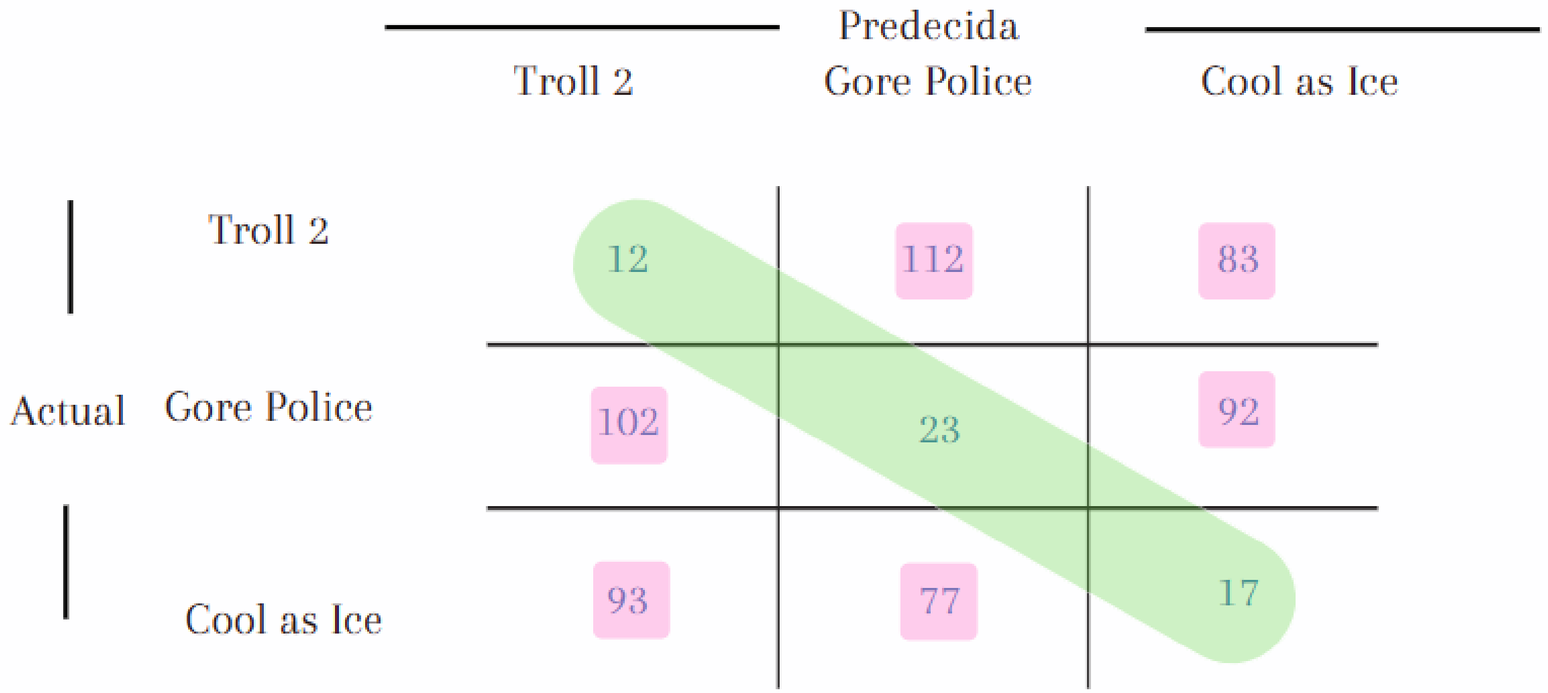
- Troll 2
- Gore Police
- Cool As Ice

Jurasic Park III	Run for your Wife	Out Kold	Howard the Duck	Película Favorita
Me gusto	No me gusto	Me gusto	Me gusto	Troll 2
No me gusto	No me gusto	Me gusto	No me gusto	Gore Police
No me gusto	Me gusto	Me gusto	Me gusto	Cool As Ice
...	...	...	...	...

Entonces la matriz de confusión tendrá **3 filas x 3 columnas**.

- **Al igual que antes en la diagonal es donde el algoritmo de aprendizaje automático hizo las cosas correctas.**
- **Y todo lo que se encuentra fuera de la diagonal es donde el algoritmo se equivocó.**

En este caso el algoritmo de aprendizaje automático no lo hizó tan bien ... pero ¿podemos culparlo? ... Todas estas películas son **terribles**.



Por último, el tamaño de una matriz de confusión dependerá del número de cosas que queremos predecir.
En el primer ejemplo solo estabamos tratando de predecir dos cosas:

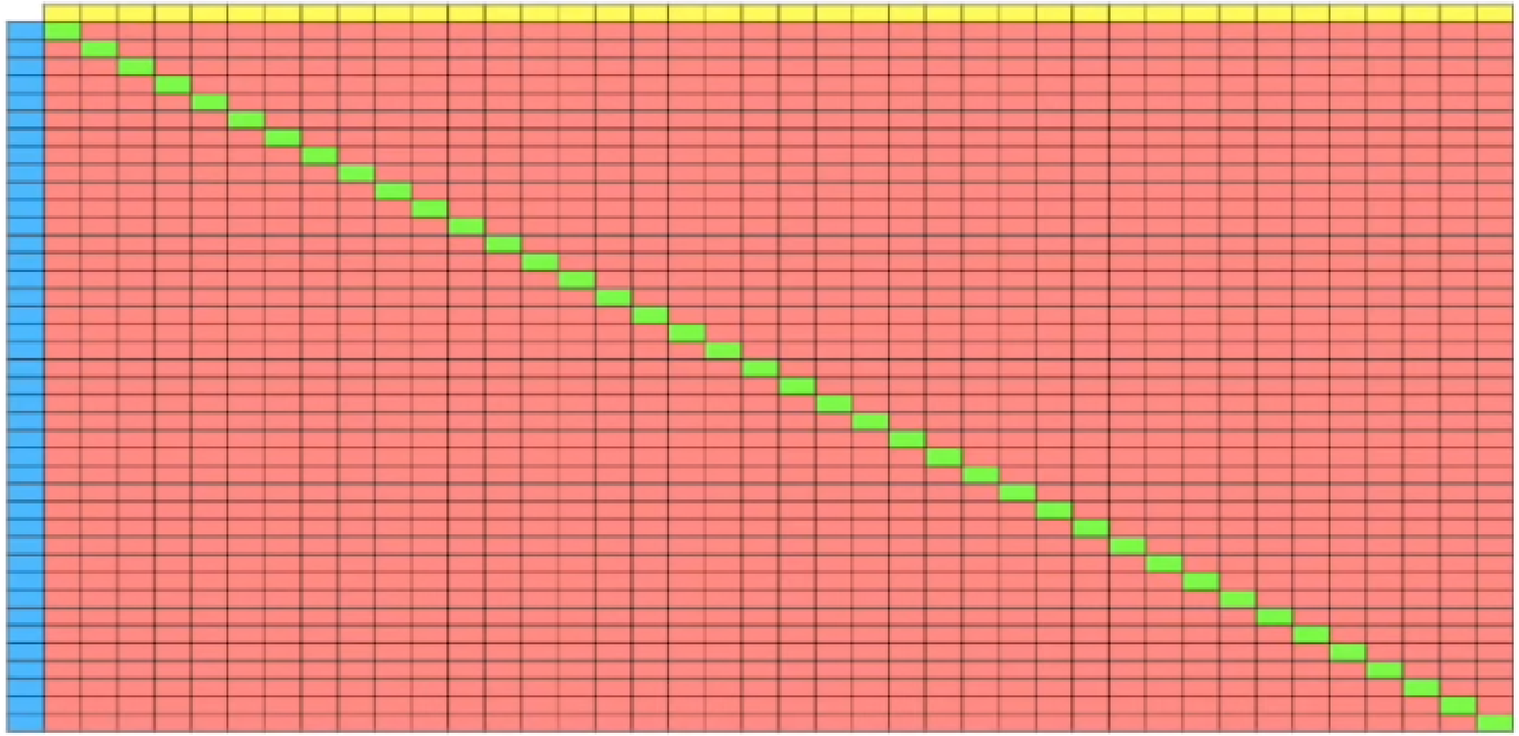
- La persona tiene una enfermedad cardíaca
- La persona no tiene una enfermedad cardíaca

Y esto nos dió una matriz de **2 filas x 2 columnas**.

En el segundo ejemplo teníamos 3 películas para elegir como la favorita de una persona, por lo que la matriz de confusión resulto en una de **3 filas x 3 columnas**.

Entonces ... Si hubieramos tenido 4 opciones para elegir, la matriz habría sido **4 filas x 4 columnas**

Y si tuvieramos 40 opciones para elegir, tendríamos una matriz de confusión de **40 filas x 40 columnas**



En resúmen, una **matriz de confusión**, nos dice que es lo que tu algoritmo de aprendizaje automático hizo correcto y que hizo incorrecto