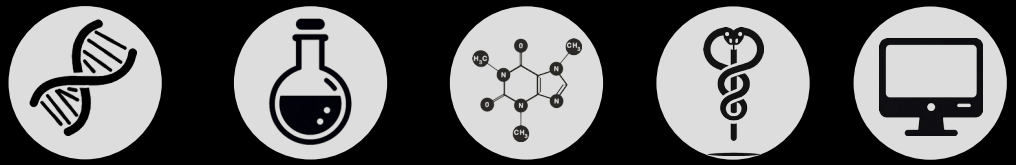
Grupo de Ciencia Computacional HIMFG



Introducción a Algoritmos de Machine Learning I

Random Forest

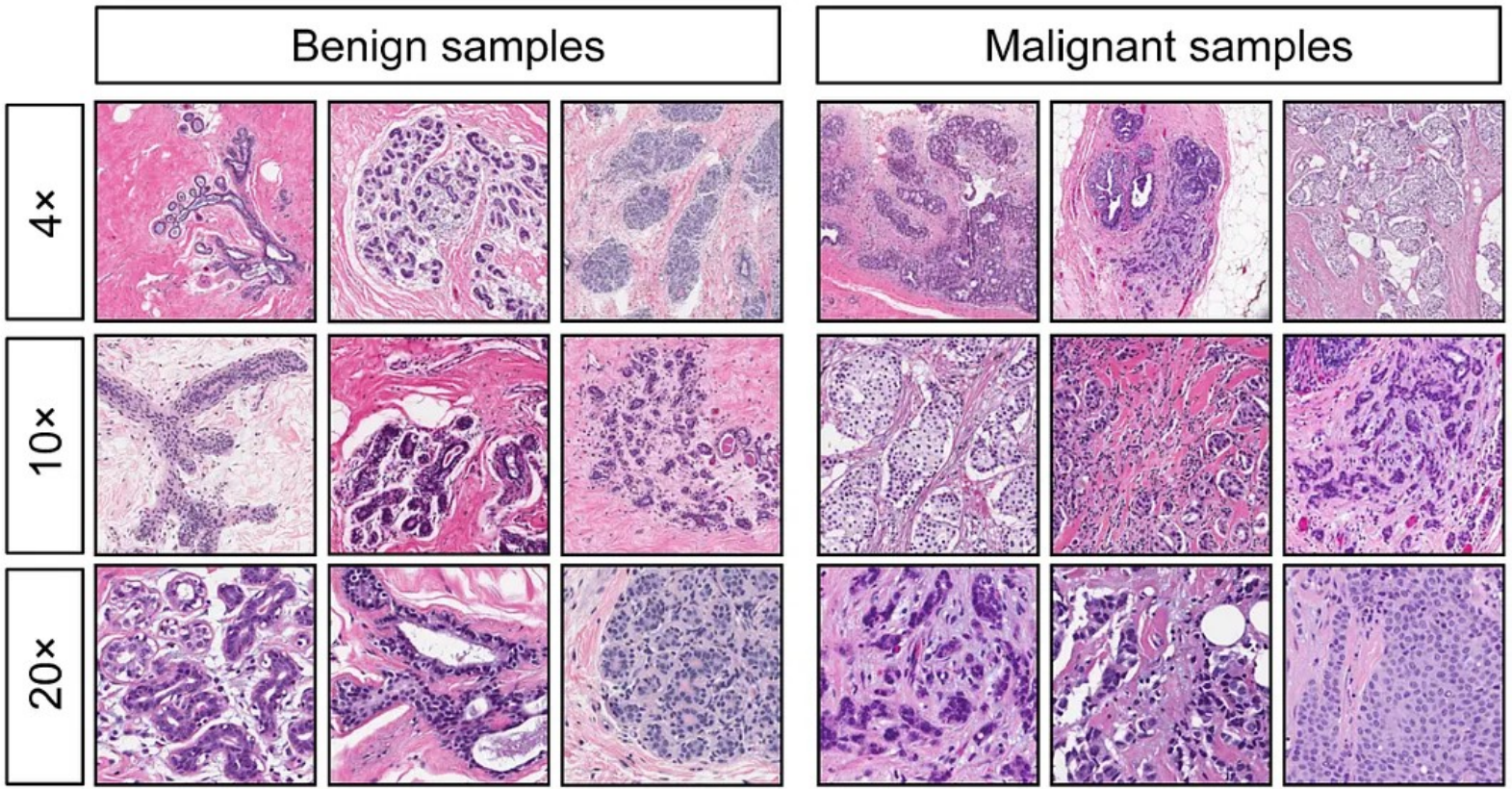
26 Febrero 2025 || Diego Prada Gracia @ Grupo de Ciencia Computacional HIMFG


creative
commons


BY NC SA

Un caso práctico

Wisconsin Breast Cancer Dataset



Mangasarian, O. L., Street, W. N., & Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. Operations research, 43(4), 570-577.

Wisconsin Breast Cancer Dataset

- Imágenes digitales de 569 Pacientes: 357 benignas, 212 malignas.

- Promedio, desviación estandard y valor máximo de 10 características en la imagen (30 atributos):

- radio
- perímetro
- suavidad
- concavidad
- dimensión fractal
- textura
- área
- compactación
- simetría
- puntos cóncavos

- 1 atributo “diagnóstico”: benigno - maligno

Dataset

Diagnóstico	***<Radio>***	***<Textura>***	***<Concavidad>***	***<Compactación>***
Maligno	*20.5*	*25.3*	*0.45*	*0.35*
Maligno	*18.7*	*20.1*	*0.40*	*0.32*
Benigno	*12.1*	*20.8*	*0.15*	*0.22*
Benigno	*14.3*	*19.6*	*0.18*	*0.24*
Maligno	*21.2*	*26.5*	*0.50*	*0.29*
Benigno	*13.0*	*18.5*	*0.12*	*0.27*
...				

Breast Cancer Wisconsin (Diagnostic)

Donated on 10/31/1995

Diagnostic Wisconsin Breast Cancer Database.

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Health and Medicine	Classification

Feature Type	# Instances	# Features
Real	569	30

Dataset Information

Additional Information

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. A few of the images can be found at http://www.cs.wisc.edu/~street/images/

Separating plane described above was obtained using Multisurface Method-Tree (MSM-T) [K. P. Bennett, "Decision Tree Construction Via Linear Programming." Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society, pp. 97-101, 1992], a classification method which uses linear programming to construct a decision tree. Relevant features were selected using an exhaustive search in the space of 1-4 features and 1-3 separating planes.

The actual linear program used to obtain the separating plane in the 3-dimensional space is that described in: [K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34].

DOWNLOAD (50.1 KB)

IMPORT IN PYTHON

CITE

37 citations

397303 views

Keywords

health cancer

Creators

William Wolberg

Olvi Mangasarian

Nick Street

W. Street

DOI

10.24432/C5DW2B

https://archive.ics.uci.edu/

Árbol de Decisión

Nuestra Experta Oncóloga

- Mayor **radio** → Maligno
- Mayor **textura** → Maligno
- Mayor **compactación** → Maligno

- Se tienen que cumplir las 3 condiciones

Dataset

Diagnóstico	***<Radio>***	***<Textura>***	***<Concavidad>***	***<Compactación>***
Maligno	*20.5*	*25.3*	*0.45*	*0.35*
Maligno	*18.7*	*20.1*	*0.40*	*0.32*
Benigno	*12.1*	*20.8*	*0.15*	*0.22*
Benigno	*14.3*	*19.6*	*0.18*	*0.24*
Maligno	*21.2*	*26.5*	*0.50*	*0.29*
Benigno	*13.0*	*18.5*	*0.12*	*0.27*
...				

Un árbol de decisión

¿Es el radio > 17?

Dataset

Diagnóstico	*<Radio>*	*<Textura>*	*<Concavidad>*	*<Compactación>*
?	***20.5***	*25.3*	*0.45*	*0.35*
?	***18.7***	*19.8*	*0.40*	*0.32*
?	*12.1*	*20.2*	*0.15*	*0.22*
?	*14.3*	*19.6*	*0.18*	*0.24*
?	***21.2***	*26.5*	*0.50*	*0.29*
?	*13.0*	*17.5*	*0.12*	*0.27*
...				

Dataset

Diagnóstico	*<Radio>*	*<Textura>*	*<Concavidad>*	*<Compactación>*
Maligno	*20.5*	*25.3*	*0.45*	*0.35*
Maligno	*18.7*	*19.8*	*0.40*	*0.32*
Benigno	*12.1*	*20.2*	*0.15*	*0.22*
Benigno	*14.3*	*19.6*	*0.18*	*0.24*
Maligno	*21.2*	*26.5*	*0.50*	*0.29*
Benigno	*13.0*	*17.5*	*0.12*	*0.27*
...	...	...	...	...

Un árbol de decisión

¿Es el radio > 17?

¿Es la textura > 20?

Dataset

Diagnóstico	<Radio>	<Textura>	<Concavidad>	<Compactación>
Maligno	20.5	25.3	0.45	0.35
Maligno	18.7	19.8	0.40	0.32
Benigno	12.1	20.2	0.15	0.22
Benigno	14.3	19.6	0.18	0.24
Maligno	21.2	26.5	0.50	0.29
Benigno	13.0	17.5	0.12	0.27
...	...	...	...	...

Dataset

Diagnóstico	*<Radio>*	*<Textura>*	*<Concavidad>*	*<Compactación>*
Maligno	***20.5***	***25.3***	*0.45*	*0.35*
Benigno	***18.7***	*19.8*	*0.40*	*0.32*
Benigno	*12.1*	***20.2***	*0.15*	*0.22*
Benigno	*14.3*	*19.6*	*0.18*	*0.24*
Maligno	***21.2***	***26.5***	*0.50*	*0.29*
Benigno	*13.0*	*17.5*	*0.12*	*0.27*
...	...	...	...	...

Un árbol de decisión

¿Es el radio > 17?

¿Es la textura > 20?

¿Es la compactación > 0.28?

Dataset

Diagnóstico	***<Radio>***	***<Textura>***	***<Concavidad>***	***<Compactación>***
Maligno	***20.5***	***25.3***	*0.45*	***0.35***
Benigno	***18.7***	*19.8*	*0.40*	***0.32***
Benigno	*12.1*	***20.2***	*0.15*	*0.22*
Benigno	*14.3*	*19.6*	*0.18*	*0.24*
Maligno	***21.2***	***26.5***	*0.50*	***0.29***
Benigno	*13.0*	*17.5*	*0.12*	*0.27*
...	...	...	...	...

¿Cómo caracterizo su acierto?

Dataset

Diagnóstico	***<Radio>***	***<Textura>***	***<Concavidad>***	***<Compactación>***
Maligno	*20.5*	*25.3*	*0.45*	*0.35*
Maligno	*18.7*	*20.1*	*0.40*	*0.32*
Benigno	*12.1*	*20.8*	*0.15*	*0.22*
Benigno	*14.3*	*19.6*	*0.18*	*0.24*
Maligno	*21.2*	*26.5*	*0.50*	*0.29*
Benigno	*13.0*	*18.5*	*0.12*	*0.27*
...				

Dataset

Diagnóstico	<Radio>	<Textura>	<Concavidad>	<Compactación>
Maligno	20.5	25.3	0.45	0.35
Benigno	18.7	19.8	0.40	0.32
Benigno	12.1	20.2	0.15	0.22
Benigno	14.3	19.6	0.18	0.24
Maligno	21.2	26.5	0.50	0.29
Benigno	13.0	17.5	0.12	0.27
...	...	...	...	...

Matriz de confusión

Diagnóstico

- ***Maligno***
- *Benigno*
- *Benigno*
- *Benigno*
- ***Maligno***
- *Benigno*

...

	Maligno real	***Benigno real***
Maligno predicción	*Verdaderos Positivos*	*Falsos Positivos*
Benigno predicción	*Falsos Negativos*	*Verdaderos Negativos*

Matriz de confusión

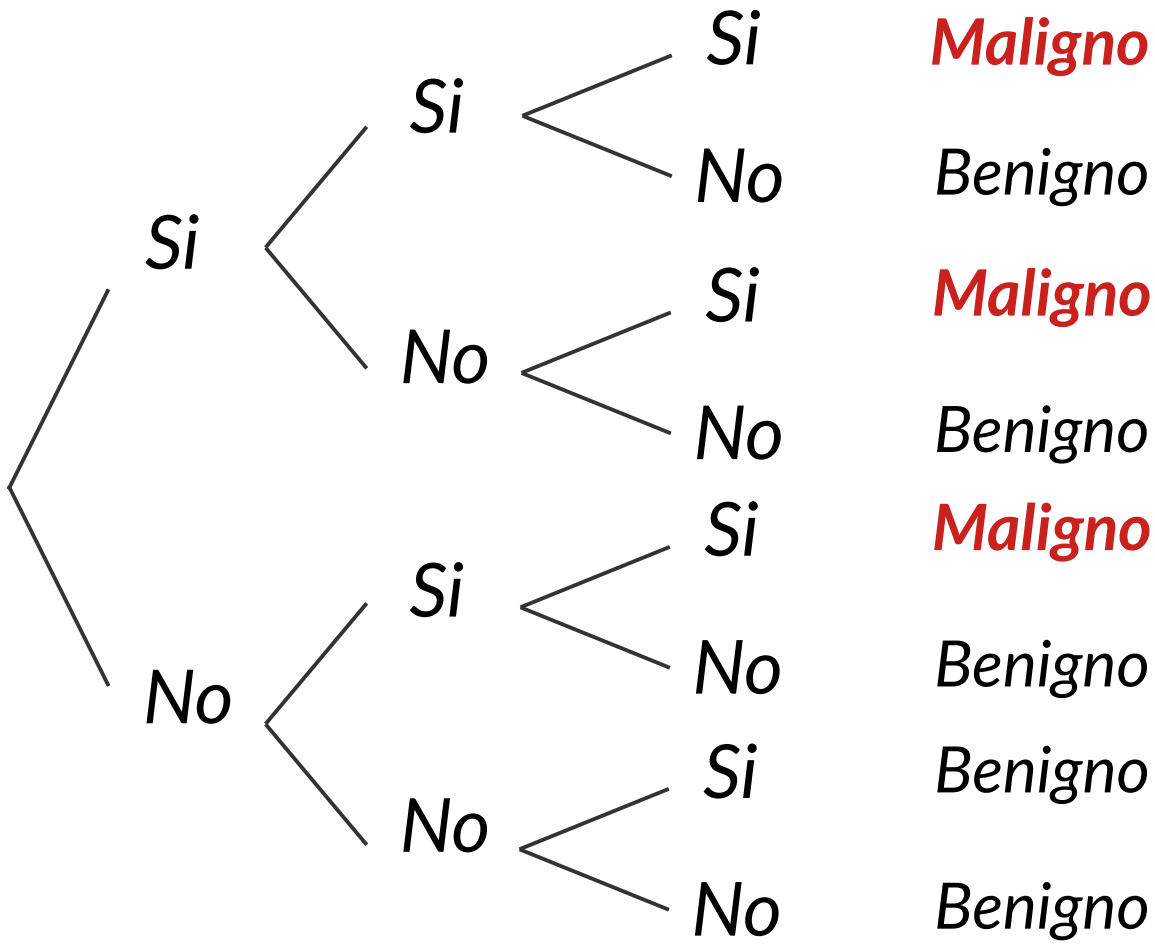
Diagnóstico
Maligno
Benigno
Benigno
Benigno
Maligno
Benigno
...

	Maligno real	Benigno real
Maligno predicción	2	0
Benigno predicción	1	3

Radio > 17 Textura > 20 Compact. > 0.28

Si	Si	Si	**Maligno**
		No	Benigno
	No	Si	Benigno
		No	Benigno
No	Si	Si	Benigno
		No	Benigno
	No	Si	Benigno
		No	Benigno

Compact. > 0.25 Radio > 18 Concavidad > 0.25



Dataset

Diagnóstico	<Radio>	<Textura>	<Concavidad>	<Compactación>
Maligno	20.5	25.3	0.45	0.35
Maligno	18.7	19.8	0.40	0.32
Benigno	12.1	20.2	0.15	0.22
Benigno	14.3	19.6	0.18	0.24
Maligno	21.2	26.5	0.50	0.29
Benigno	13.0	17.5	0.12	0.27
...	...	...	...	...

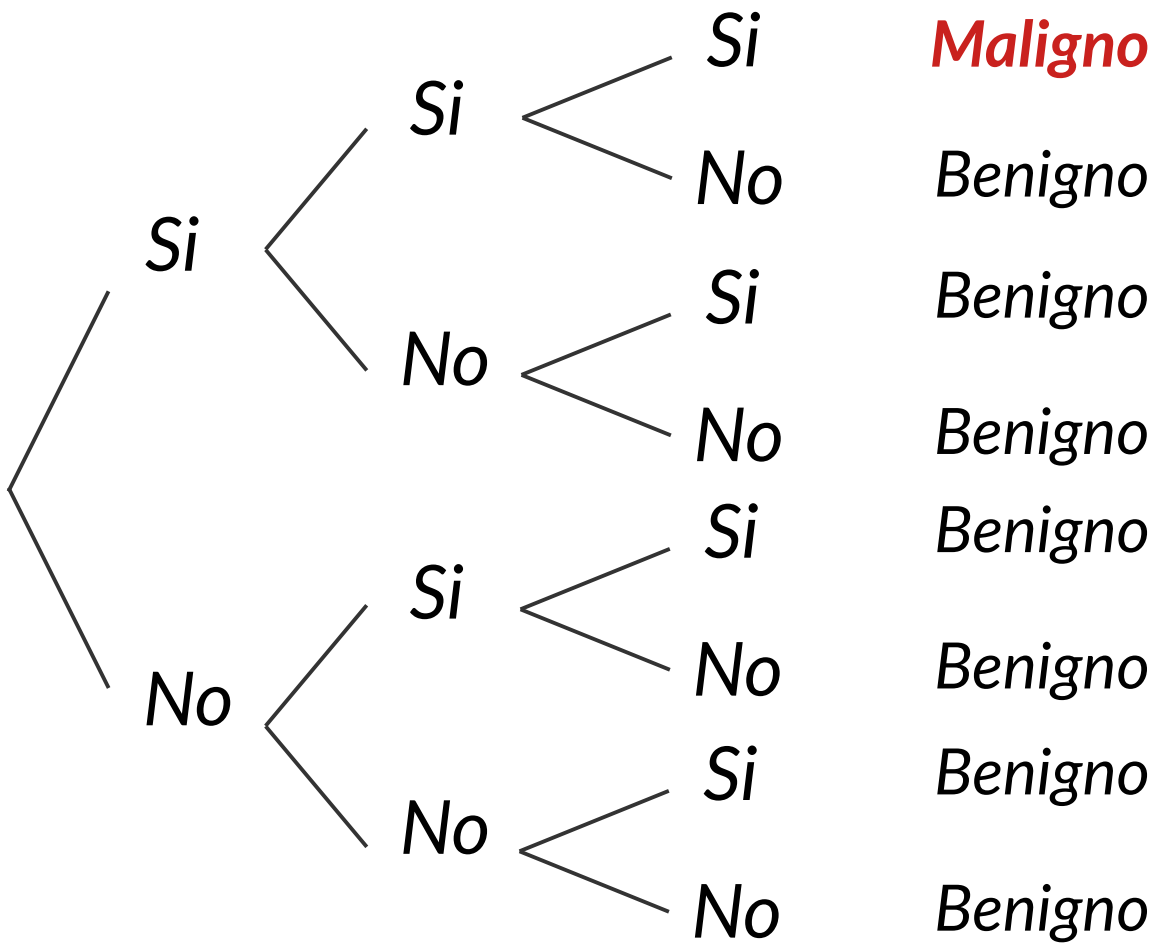
Matriz de confusión

Diagnóstico
Maligno
Maligno
Benigno
Benigno
Maligno
Benigno
...

	Maligno real	Benigno real
Maligno predicción	3	0
Benigno predicción	0	3

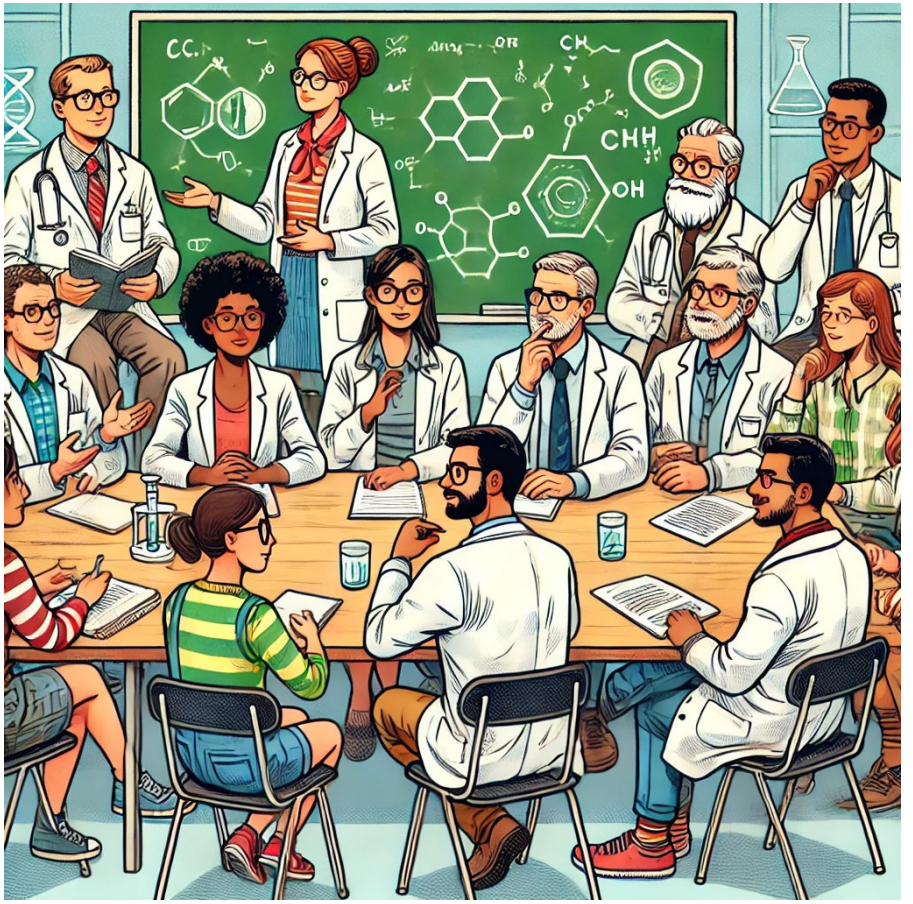
¿Cómo elegimos o entrenamos al experto más acertado?

Radio > 17 Textura > 20 Compact. > 0.28



¡Necesitamos un bosque!

Si consultamos a un sólo investigador, puede equivocarse.
Pero si consultamos a 100 investigadores, nuestra decisión es más confiable

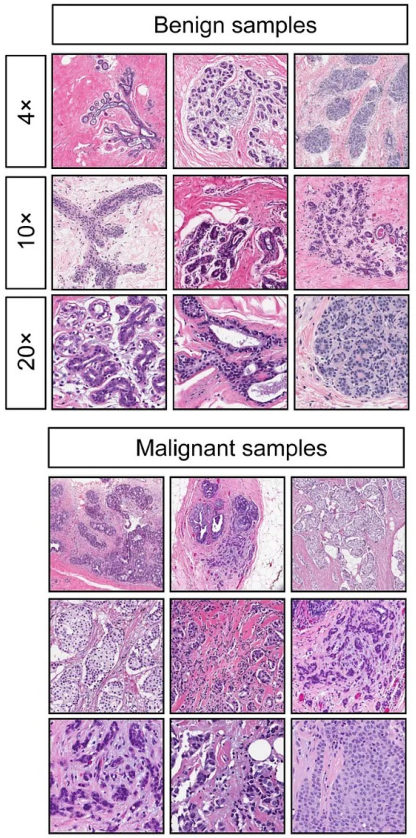


Random Forest

Algoritmo de aprendizaje automático

Entrenamiento
+
Evaluación

Wisconsin Breast Cancer Dataset



569 Pacientes
212 Malignas + 357 Benignas

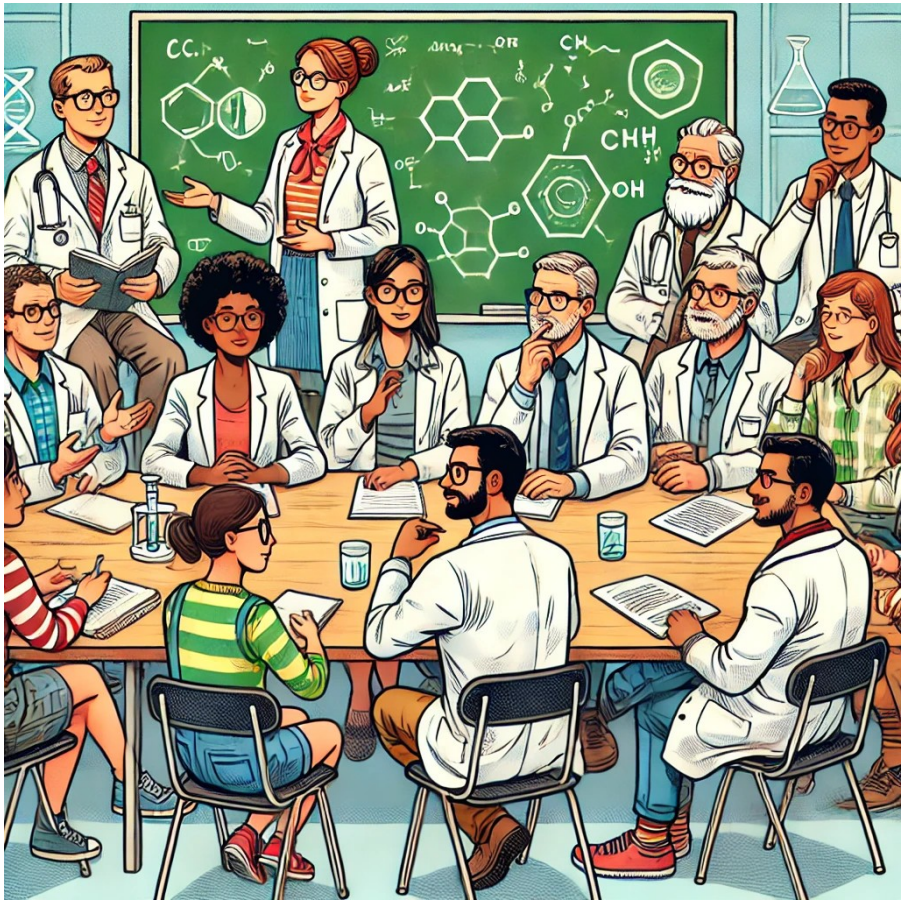
	Maligno real	*Benigno real*
Maligno predicción	*2*	*0*
Benigno predicción	*1*	*3*

75%
Entrenamiento

25%
Evaluación


¿Cómo entrenamos
el bosque?

Tomemos un conjunto de árboles plural e independiente
(muestro variado -> robusto)



Muestreo Bootstrap

100 árboles distintos

100 muestras distintas (muestra aleatoria con reemplazo)

+

7 características aleatorias distintas por árbol

Bootstrapping

“Pulling oneself up by one’s bootstraps”

Cada árbol se entrena independientemente

Evaluación de:

*- todas las **características** observadas.*
*- todos los posibles **puntos de corte.***

Corte que más minimiza la entropía

Diagnóstico	*<Radio>*
Maligno	*20.5*
Maligno	*18.7*
Benigno	*12.1*
Benigno	*14.3*
Maligno	*21.2*
Benigno	*13.0*
…	…

$$S = -\sum_{i=1}^{n} \left(p_i \cdot \log\left(p_i \right) \right)$$

Entropía final – Entropía inicial

Corte que más minimiza la entropía

Diagnóstico	*<Radio>*
Maligno	***20.5***
Maligno	***18.7***
Benigno	*12.1*
Benigno	***14.3***
Maligno	***21.2***
Benigno	*13.0*
…	…

Situación inicial

$p_M = 0.5 \qquad p_B = 0.5$

Situación final (corte en 14)

Si $p_M = 0.75 \qquad p_B = 0.25$

No $p_M = 0 \qquad p_B = 1$

Corte que más minimiza la entropía

Diagnóstico	*<Radio>*
Maligno	***20.5***
Maligno	***18.7***
Benigno	*12.1*
Benigno	*14.3*
Maligno	***21.2***
Benigno	*13.0*
…	…

Situación inicial

$p_M = 0.5$ $p_B = 0.5$

Situación final (corte en 16)

Si $p_M = 1$ $p_B = 0$

No $p_M = 0$ $p_B = 1$

Corte que más minimiza la entropía

Diagnóstico	*<Radio>*
Maligno	***20.5***
Maligno	*18.7*
Benigno	*12.1*
Benigno	*14.3*
Maligno	***21.2***
Benigno	*13.0*
…	…

Situación inicial

$p_M = 0.5$ $p_B = 0.5$

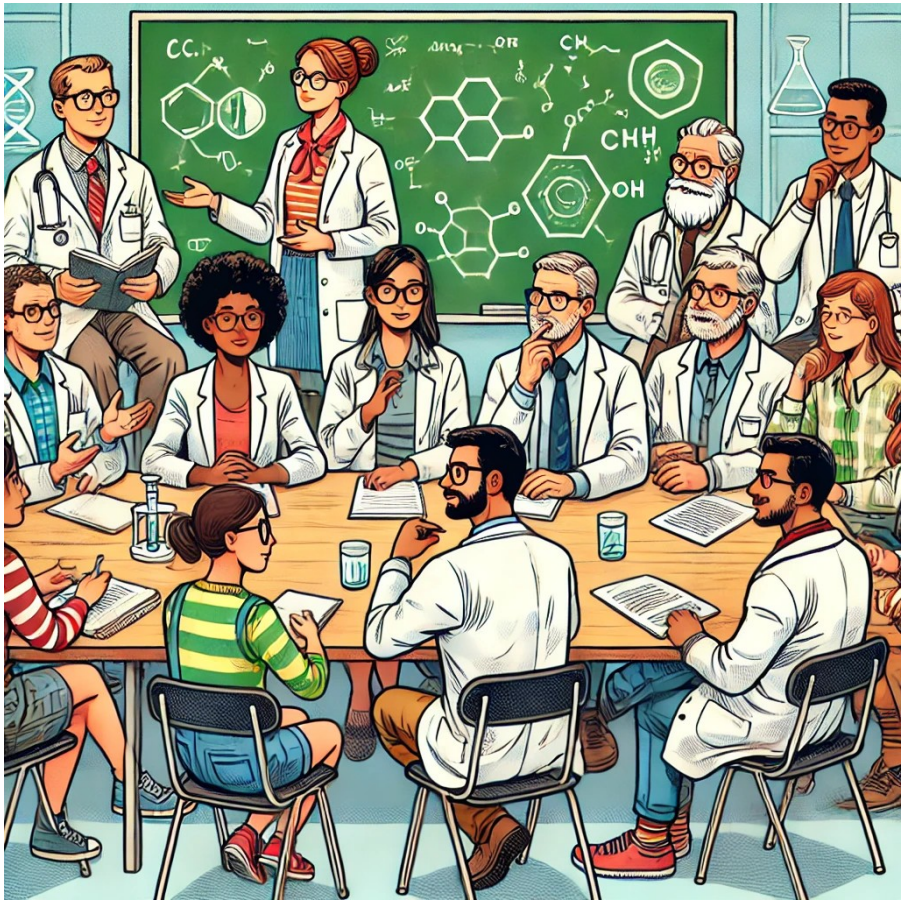
Situación final (corte en 19)

Si $p_M = 1$ $p_B = 0$

No $p_M = 0.25$ $p_B = 0.75$

¡Ya tenemos cada árbol entrenado!

La decisión será tomada ***por mayoría*** *(o expresando probabilidad de malignidad)*




¿Cómo evaluamos
el predictor?

Matriz de confusión

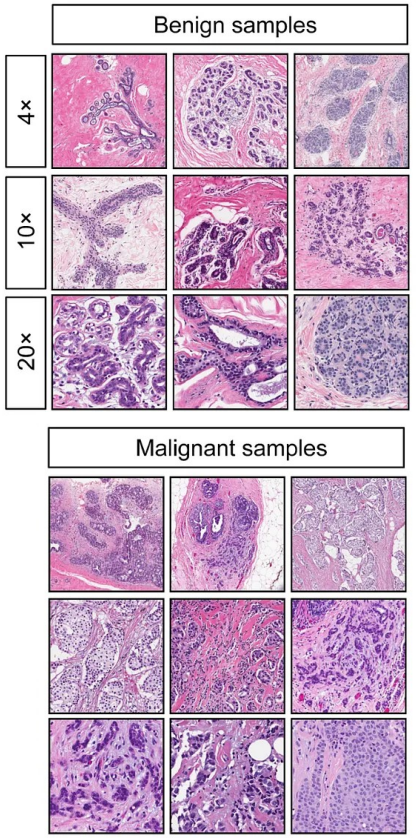
Diagnóstico

- ***Maligno***
- *Benigno*
- *Benigno*
- *Benigno*
- ***Maligno***
- *Benigno*

...

	Maligno real	***Benigno real***
Maligno predicción	*Verdaderos Positivos*	*Falsos Positivos*
Benigno predicción	*Falsos Negativos*	*Verdaderos Negativos*

Wisconsin Breast Cancer Dataset



569 Pacientes
212 Malignas + 357 Benignas

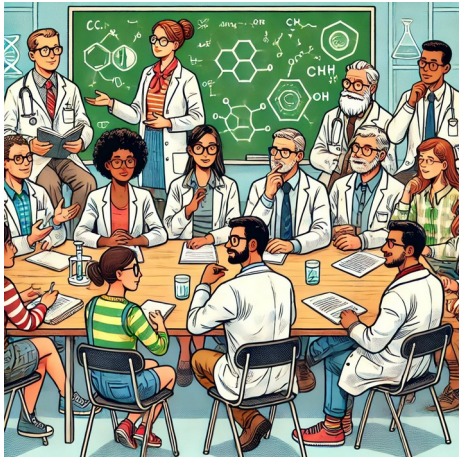
	Maligno real	*Benigno real*
Maligno predicción	*2*	*0*
Benigno predicción	*1*	*3*

75%
Entrenamiento

25%
Evaluación

Jugamos a predecir sabiendo el resultado

Diagnóstico	***<Radio>***	***<Textura>***	***<Concavidad>***	***<Compactación>***
Maligno	*20.5*	*25.3*	*0.45*	*0.35*



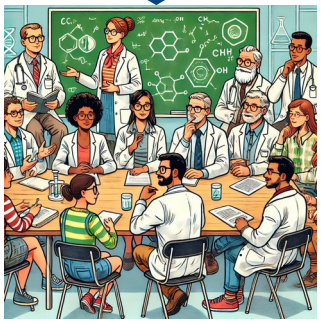
83%

17%

	Maligno real	***Benigno real***
Maligno predicción	*Verdaderos Positivos*	*Falsos Positivos*
Benigno predicción	*Falsos Negativos*	*Verdaderos Negativos*

Bootstrapping

Diagnóstico	***<Radio>***	***<Textura>***	***<Concavidad>***	***<Compactación>***
Maligno	*20.5*	*25.3*	*0.45*	*0.35*



83%

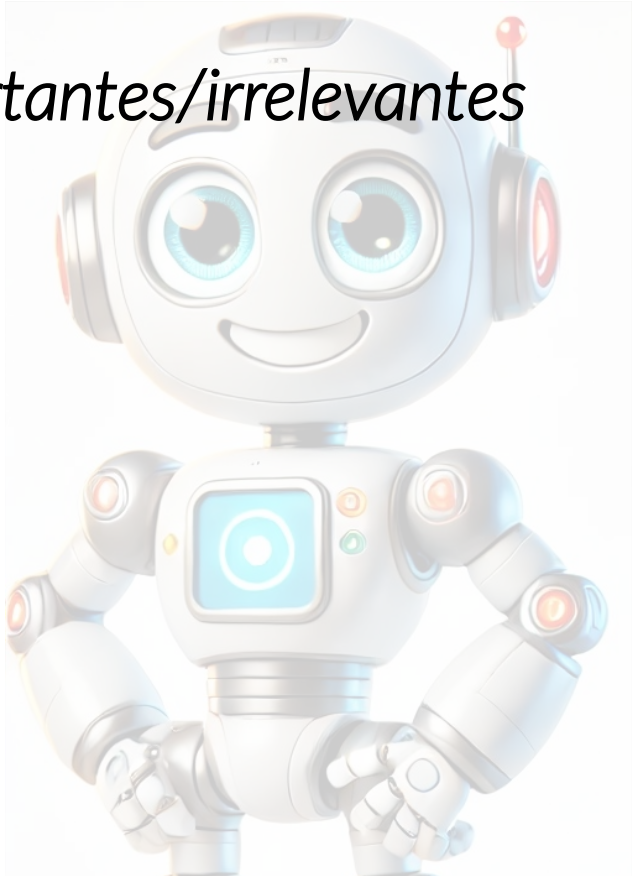
17%

	Maligno real	***Benigno real***
Maligno predicción	*Verdaderos Positivos*	*Falsos Positivos*
Benigno predicción	*Falsos Negativos*	*Verdaderos Negativos*

Ya tenemos nuestro clasificador/predictor entrenado y evaluado

Ventajas del Random Forest

- *Funciona con mezclas de datos importantes/irrelevantes*
- *Resistente al sobreajuste*
- *Funciona con datos ruidosos*
- *Funciona con datos faltantes*
- *No requiere normalización de datos*



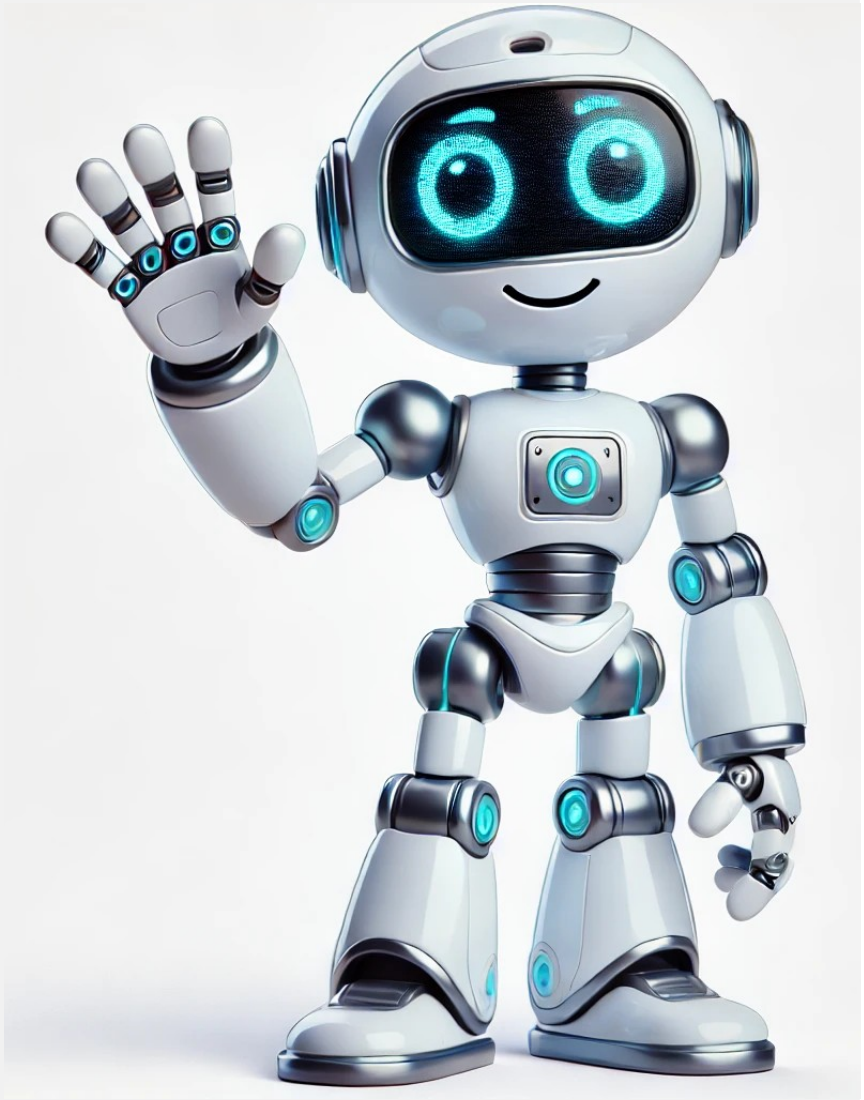
Desventajas del Random Forest

- *Lento en datasets grandes*
- *Require trabajo para ser interpretado, no tan sencillo como un árbol de decisión*
- *Puede ser problemático si el número de características irrelevantes es muy alto*

¿Podemos quitar características irrelevantes antes de entrenar?

- *Métodos de reducción de dimensionalidad:* ***PCA***

- *Métodos estadísticos: filtramos variables con baja correlación con las etiquetas (ANOVA,* ***mutual information****, test de chi-cuadrado,...)*



Últimos comentarios breves...

¿Por qué decimos que esto es IA?

- ***Inteligencia Artificial*** *es cualquier sistema capaz de* ***aprender de datos y tomar decisiones*** *sin ser programado explícitamente.*

- ***Machine Learning*** *es una rama de la IA que entrena modelos para* ***detectar patrones*** *en los datos.*

¿Qué es aprendizaje profundo o aprendizaje superficial?

-***Aprendizaje profundo***: *Redes neuronales con muchas capas, características complejas y alto número de parámetros. (Redes neuronales "complejas", Transformers, etc... Large Language Models)*

- ***Aprendizaje superficial***: *Modelos con pocas capas o reglas de decisión. (Random Forests... algoritmos de ML)*

¿Qué es aprendizaje supervisado, no supervisado y reforzado?

-***Aprendizaje supervisado***: *Se entrena con etiquetas. Por ejemplo,* ***nuestro Random Forest*** *con imágenes de tejidos marcados como malignos o benignos.*

- ***Aprendizaje no supervisado***: *Se entrena sin etiquetas. Por ejemplo algoritmos de clústering como K-means.*

- ***Aprendizaje no reforzado***: *Se entrena por ensayo-error con recompensa. Por ejemplo robots que aprenden a realizar tareas.*

¿Qué sirven los algoritmos de Machine Learning?

-***Agrupamiento***: *Algoritmos de aprendizaje no supervisado como algoritmos de clustering*

- ***Clasificador/Predictor***: *Distingue/Predice categorías como el Random Forest.*

- ***Regresión***: *Predice valores continuos como una serie temporal.*

¿Qué otros tipos de algoritmos de Machine Learning hay?

- *SVM*
- *LightGBM*
- *Regresión lineal*
- *...*
- *XGBoost*
- *KNN*
- *Redes neuronales*

¿Podemos saber qué características son relevantes para el clasificador?

- ***Random Forest*** *nos dice qué características son más importantes por construcción. Otros algoritmos no lo hacen…*

- ***SHAP (SHappely Additive Explanations)*** *mide el impacto de cada característica en la predicción final de cualquier modelo de Machine Learning.*


That's All Folks!

¿Te vas a atrever?

MUCHAS GRACIAS...