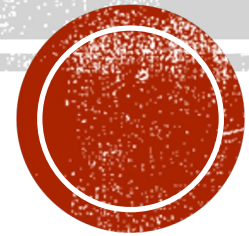


PROYECTO APLICADO



“You can’t manage what you don’t measure”,
Tom De Marco, 1982

“We’re drowning in data but starving for knowledge”,
John Naisbitt, 1982

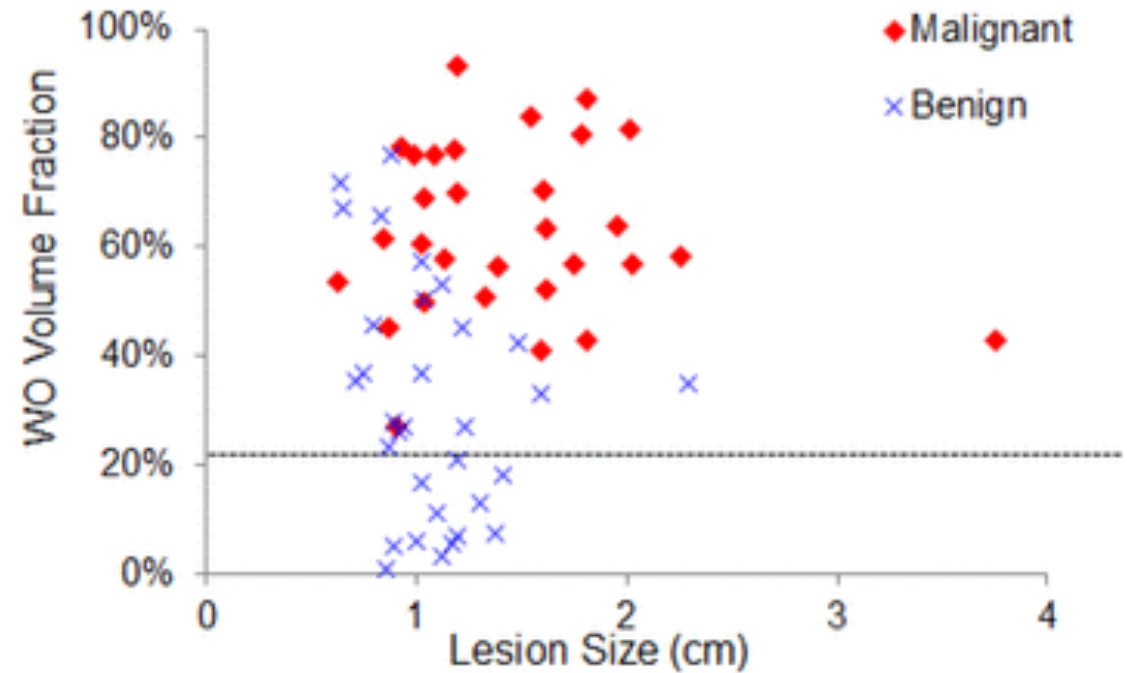
AGENDA

- Tareas de clasificación
- Métricas de clasificación
- Modelos de clasificación
- Proyecto de clasificación



TAREAS DE CLASIFICACIÓN

- Encontrar modelos que describan clases para futuras predicciones:
 - KNN
 - Árboles de decisión
 - Regresión logística
 - Redes neuronales
 - ...
- Valores **discretos** de la variable objetivo (categóricos)
- Incluye modelos que no solo clasifican sino que estiman las **probabilidades** de cada clase



http://www.jacmp.org/index.php/jacmp/article/view/5187/html_374

MÉTRICAS DE CLASIFICACIÓN

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$



MÉTRICAS DE CLASIFICACIÓN

Cohen Kappa Statistic

Measure The Performance of Classification Models

Kappa
Statistic


$$k = \frac{2 * (TP * TN - FN * FP)}{(TP + FP) * (FP + TN) + (TP + FN) * (FN + TN)}$$

Assess the level of agreement between an actual and predicted

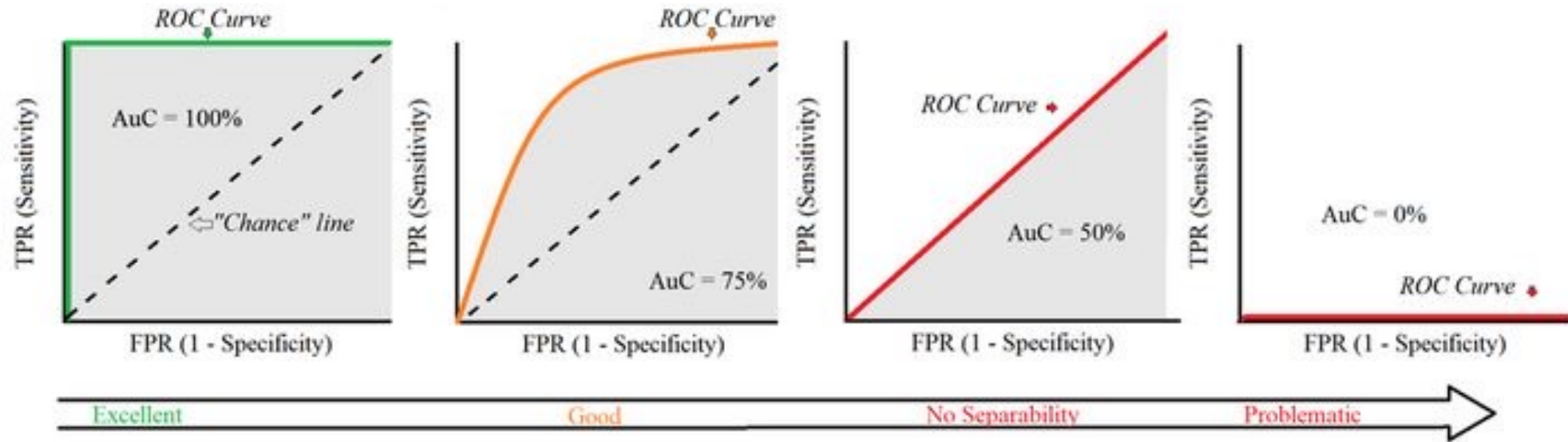
Predicted (rater 2) \ Actual (rater 1)	YES	NO	
	YES	NO	
YES	45 (TP)	15 (FN)	60
NO	25 (FP)	15 (TN)	40
	70	30	

Kappa Score Interpretation

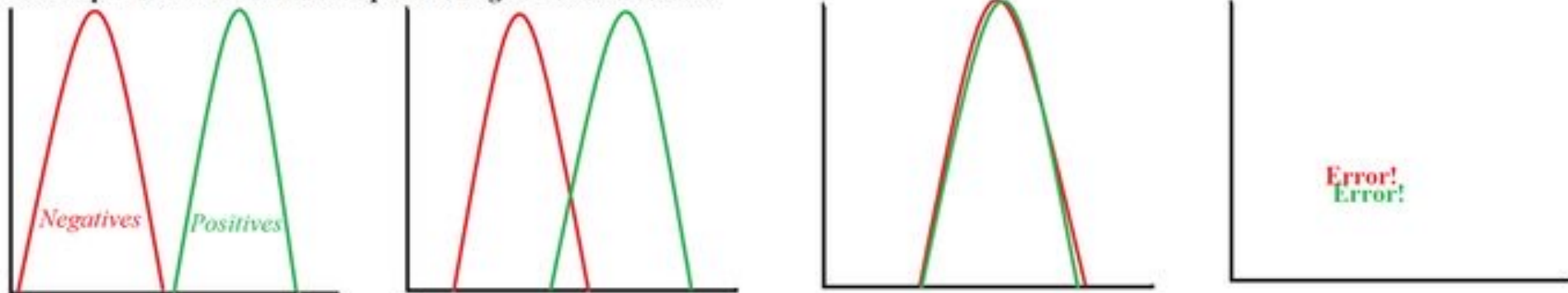
Kappa	Agreement
<0	Less than chance agreement
0.01-0.20	Slight agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-0.99	Almost perfect agreement


$$k = \frac{2 * (45 * 15 - 15 * 25)}{(45 + 25) * (25 + 15) + (45 + 15) * (15 + 15)} = 0.13 \text{ (13\%)}$$

MÉTRICAS DE CLASIFICACIÓN



Overlap = How well the model separates Negatives and Positives



MODELOS DE CLASIFICACIÓN

- **Regresión Logística**
- **KNN**
- **Naïve Bayes**
- **Random Forest**
- **Boosting**



REGRESIÓN LOGÍSTICA

- Parte de la idea de la regresión lineal, cuyo resultado es modificado para poder obtener una salida binaria: sólo permite distinguir entre 2 clases

- El modelo pasa de:

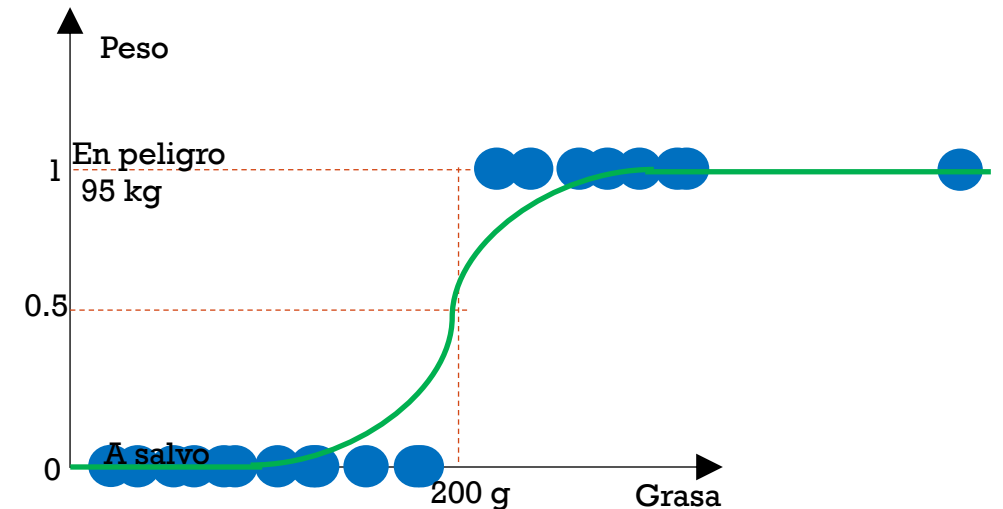
$$h_{\theta}(X) = \theta_0 + \theta_1 x_1 + \dots + \theta_n x_n$$

$$h_{\theta}(X) = f(z) = \sigma(\theta_0 + \theta_1 x_1 + \dots + \theta_n x_n),$$

con $\max(f(z))=1$ y $\min(f(z))=0$

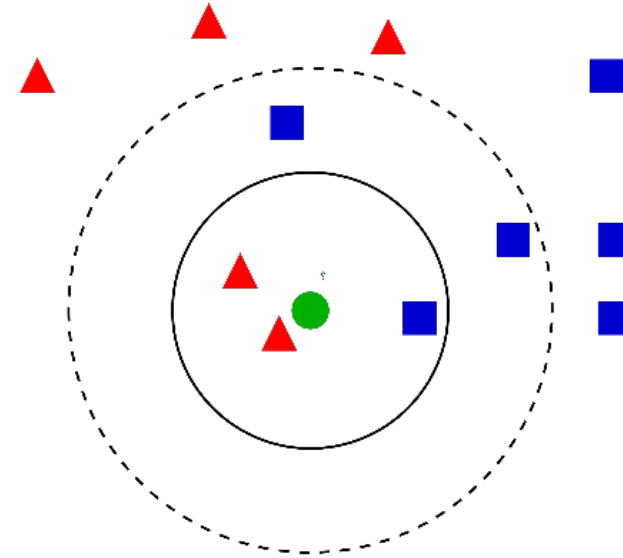
- $\sigma(z)$ es la función **sigmoide** o **logística**
- Se pueden interpretar los valores de $\sigma(z)$ como **probabilidades** de que una instancia con atributos X pertenezca a la clase $Y=1$:
 $P(Y = 1|x_1, \dots, x_n) = p_1(X) = \sigma(\theta_0 + \theta_1 x_1 + \dots + \theta_n x_n)$

$$p_1(X) = \frac{1}{1 + e^{-\theta^T X}}$$



KNN

- Algoritmo de aprendizaje supervisado para clasificación y regresión.
- Asigna la clase o valor agregado de las instancias conocidas que se encuentran mas cerca de la instancia a predecir.



NAÏVE BAYES

- **Los clasificadores bayesianos** asignan cada observación a la clase j más probable, dados los valores observados de sus variables predictivas:

$$\operatorname{argmax}_j p(Y = y_j | X = x_{\text{observados}})$$

- Si se conocen las distribuciones de probabilidad, el clasificador resultante da la frontera de separación óptima en términos de error
- No siempre se tienen las probabilidades condicionales necesarias.
- **Naïve Bayes** es un algoritmo basado en el Teorema de Bayes

La regla de clasificación es:

$$\operatorname{argmax}_j p(y_j) \prod_{i=1}^n p(x_i | y_j)$$

Sólo se debe especificar :

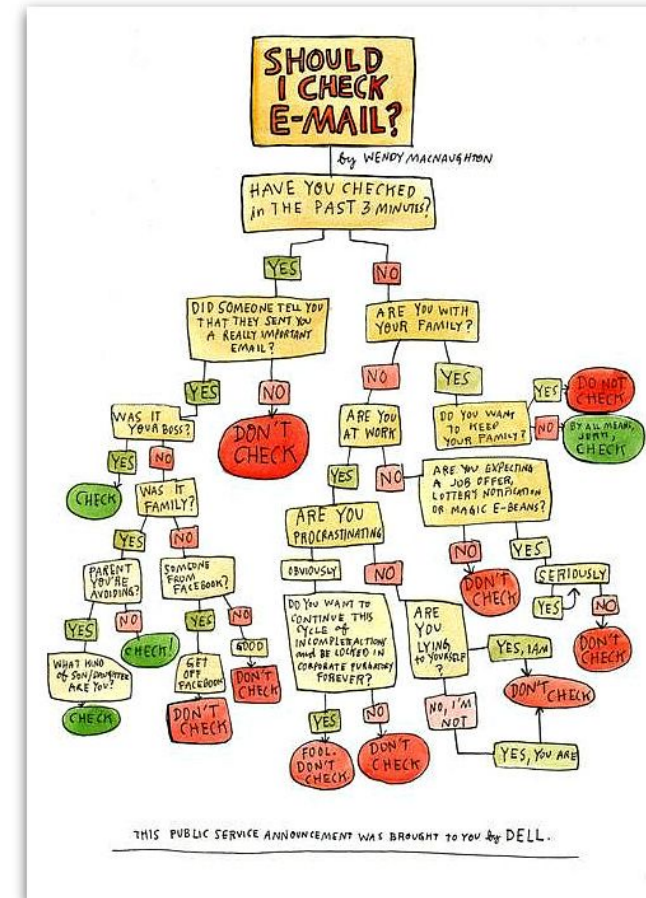
- Las probabilidades a priori de cada clase
- Las distribuciones de probabilidad de las variables predictivas para cada clase (condicionadas a la clase)

Esta información se constituye en los **parámetros** del modelo, y en el caso de variables categóricas se obtienen a partir de tablas de frecuencias (conteos)



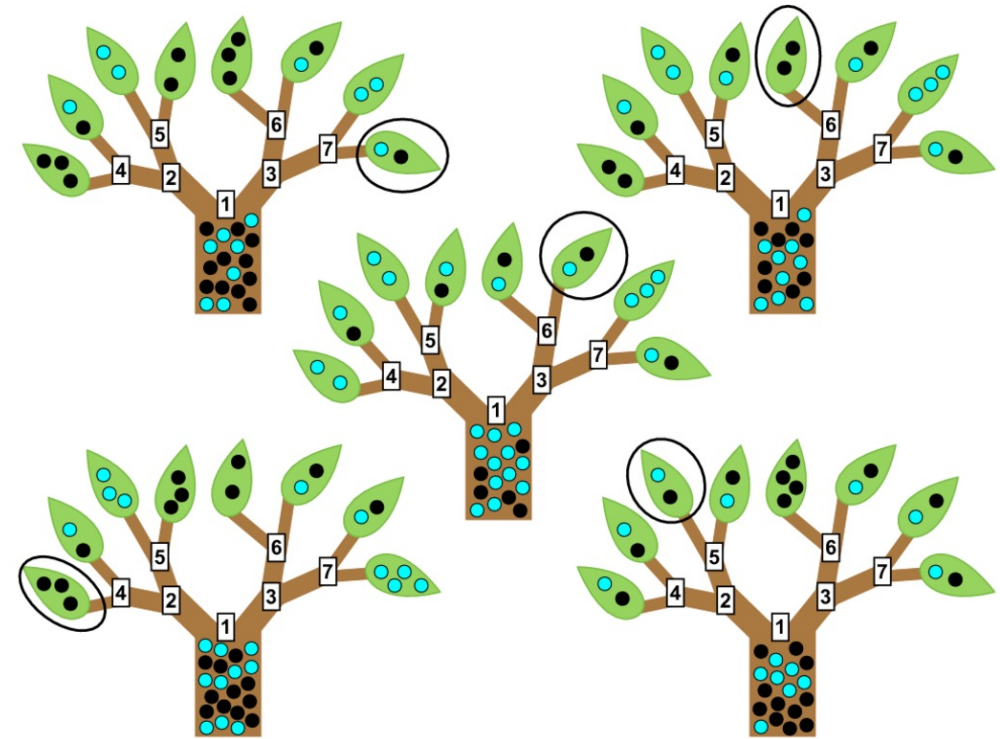
ÁRBOLES DE DECISIÓN

- Algoritmo de aprendizaje supervisado para clasificación y regresión.
- Posee una estructura de árbol, donde sus nodos internos representan las características, las ramas las decisiones y cada hoja el resultado.



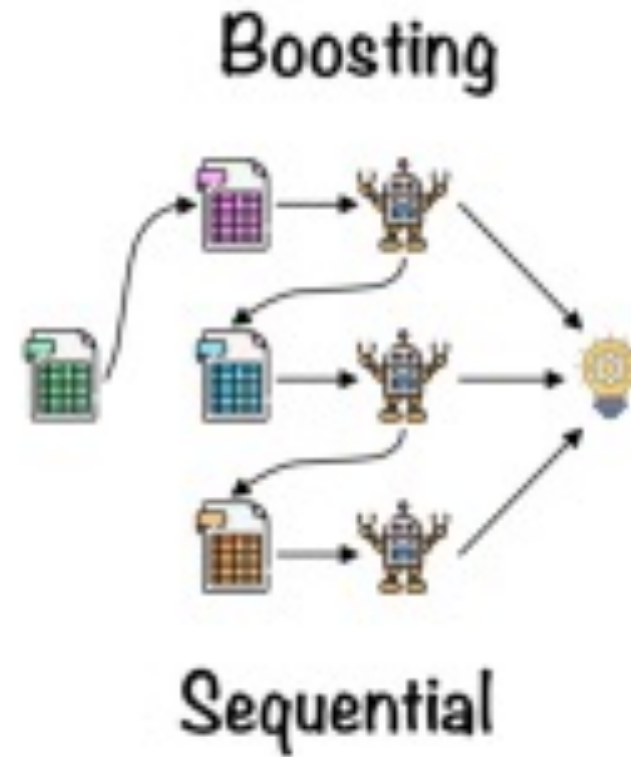
RANDOM FOREST

- Algoritmo de aprendizaje supervisado para clasificación y regresión.
- Basado en el concepto de ensambles, el cual es un proceso de combinación de múltiples regresores o clasificadores para resolver un problema complejo y mejorar el rendimiento del modelo.



BOOSTING

- Algoritmo de aprendizaje supervisado para clasificación y regresión.
- Mejora la exactitud y rendimiento de los modelos de aprendizaje automático convirtiendo múltiples modelos débiles en uno mucho más potente.



Proyecto de clasificación

- Análisis exploratorio de datos
 - Limpieza de datos
 - Valores faltantes y tipos de datos
 - Identificación (explícitos e implícitos)
 - Imputación
 - Valores atípicos
 - Visualización (boxplots e histogramas)
 - Limitación inferior y superior (winsorizing)
- Exploración de datos
 - Análisis univariado
 - Variables continuas
 - Variables categóricas
 - Análisis bivariado
 - Independientes continuas vs variable objetivo
 - Variables independientes continuas entre sí
 - Variables independientes categóricas entre sí
 - Variables categóricas con respecto a la variable objetivo
- Ingeniería de características
 - Creación de nuevas características
 - Dummificación
 - Eliminación de características
 - Variables independientes continuas altamente correlacionadas entre sí
 - Variables independientes categóricas con baja correlación con la objetivo
 - PCA

Proyecto de clasificación

- Protocolos de evaluación
- Métricas de evaluación
 - Accuracy
 - Kappa
 - Precision
 - Recall
 - ROC AUC
- Implementación de modelos
 - Baseline
 - Persistencia
 - Uso de pickle para guardar modelos y métricas
 - Regresión logística
 - KNN
 - Naïve Bayes
 - Árbol de clasificación
 - Random forest
 - Gradient Boosting
 - XGBoosting
- Comparación de modelos



Ejercicio

Replicar los mismos pasos realizados con el conjunto de datos de riesgo automotor, pero en este caso con respecto al conjunto de datos de cardiología.

Con este dataset se busca predecir el riesgo cardíaco.

