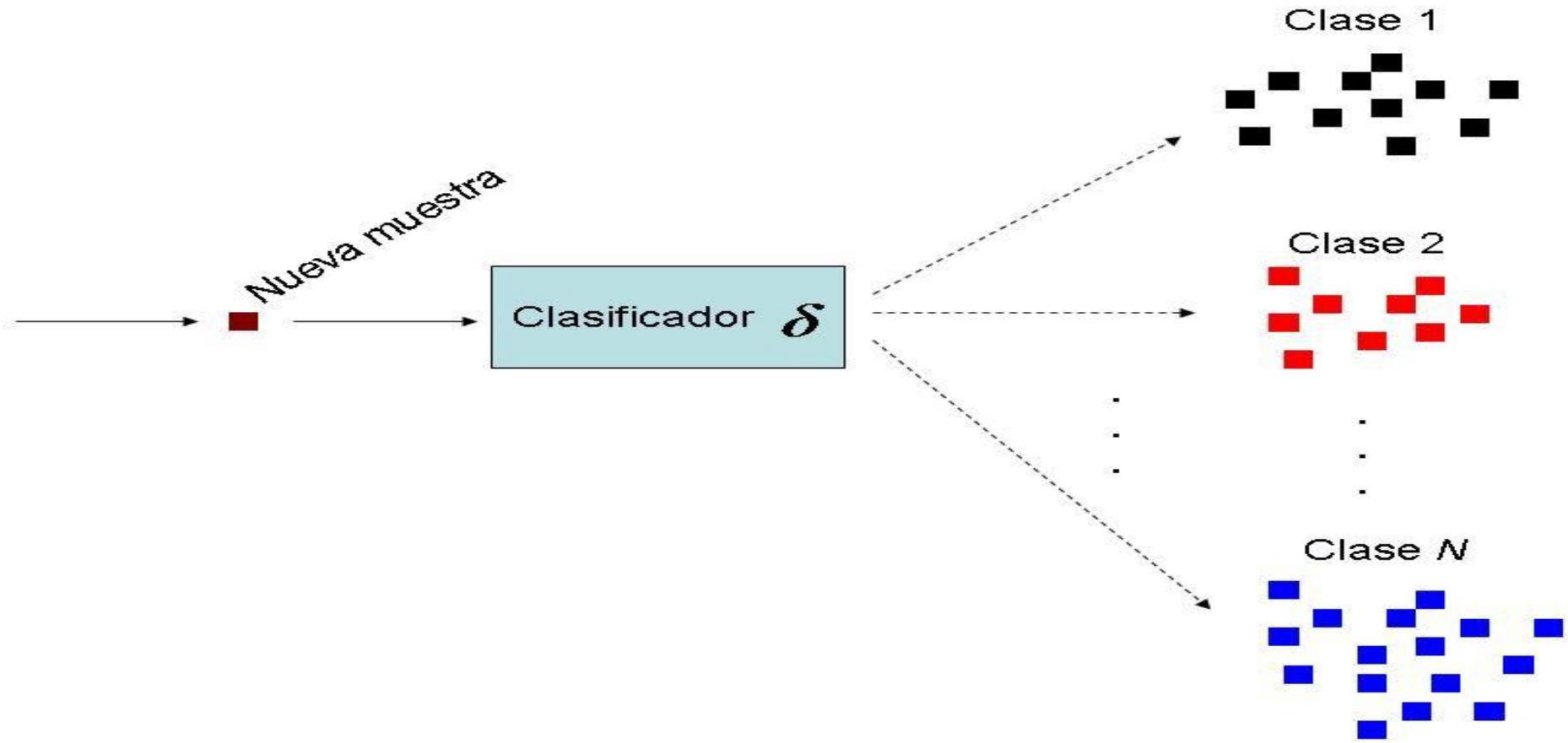
Clasificación

Clasificación

- Regresión logística
- Matriz de confusión
- Análisis de Componentes principales ACP
- Diagrama ROC
- Arboles de Decisión para Clasificación
- Vecino mas cercano

¿En dónde aplicar clasificación?



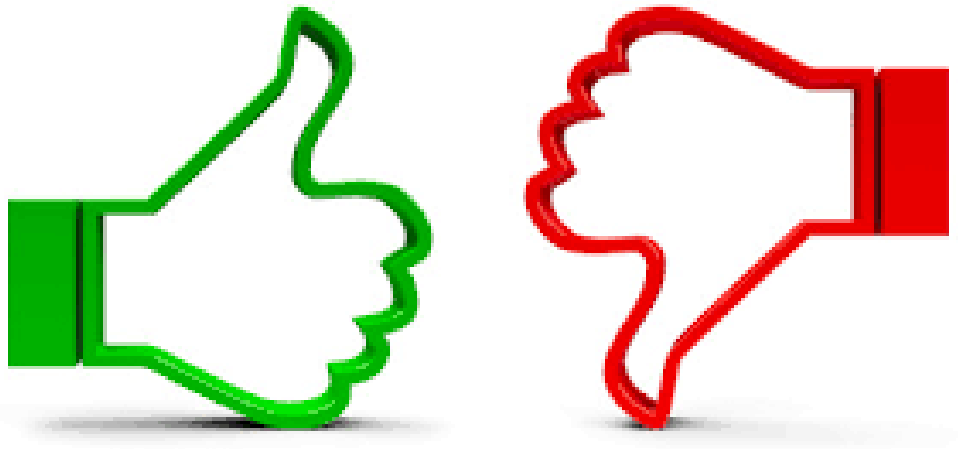
¿En dónde aplicar clasificación?

- ¿comprará el cliente este producto? [sí, no]
- ¿tipo de tumor? [maligno, benigno]
- ¿subirá el índice bursátil? IBEX mañana [sí, no]
- ¿es este comportamiento una anomalía? [sí, no]
- ¿nos devolverá este cliente un crédito? [sí, no]
- ¿qué deporte estás haciendo? tal y como lo detectan los relojes inteligentes [caminar, correr, bicicleta, nadar]
- ¿obtendrá una historia un número alto de visitas en un agregador de noticias? [sí, no]

Regresión Logísitica gml()
Caso probabilidad de fraude por impago (*default*)

la probabilidad de fraude por impago (*default*) en función del balance de la cuenta bancaria (*balance*).

https://rpubs.com/rpizarro/584634



Regresión Logísitica gml()
Convertir probabilidad en clasificación

- Una de las principales aplicaciones de un modelo de regresión logística es clasificar la variable cualitativa en función de valor que tome el predictor.

- Para conseguir esta clasificación, es necesario establecer un *threshold* (límite) de probabilidad a partir de la cual se considera que la variable pertenece a uno de los niveles.
- Por ejemplo, se puede asignar una observación al grupo 1

$$1 \text{ si } \hat{p}(Y = 1|X) > 0.5$$

- Asignar al grupo 0 en caso contrario

Regresión Logísitica gml()
Ejemplo de Cuadro de honor y matemáticas

- *Un estudio quiere establecer un modelo que permita calcular la probabilidad de obtener una matrícula de honor al final del bachillerato en función de la nota que se ha obtenido en matemáticas.*

- *La variable matrícula está codificada como 0 si no se tiene matrícula y 1 si se tiene.*

https://rpubs.com/rpizarro/584508

Regresión Logísitica gml()

Ejemplo de Cuadro de honor y matemátcas

https://rpubs.com/rpizarro/584508

El modelo es capaz de clasificar correctamente $\frac{140+22}{140+22+27+11} = 0.81(81\%)$ de las observaciones cuando se emplean los datos de entrenamiento.

Regresión Logísitica gml()
Ejemplo de Cuadro de honor y matemátcas

https://rpubs.com/rpizarro/584508

```
> matriz_confusion
                predicciones
observaciones   0   1
            0 140  11
            1  27  22
```

Matriz de Confusión

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Matriz de Confusión

En el campo de la inteligencia artificial una **matriz de confusión** es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado.

Matrices de confusión. La matriz de confusión es una herramienta fundamental a la hora de evaluar el desempeño de un algoritmo de clasificación, ya que dará una mejor idea de cómo se está clasificando dicho algoritmo, a partir de un conteo de los aciertos y errores de cada una de las clases en la clasificación. Así se puede comprobar si el algoritmo está clasificando mal las clases y en qué medida.

Caso matriz de confusión de Notas Estudiantes

https://rpubs.com/rpizarro/584650

	Predecido		
Actual	Low	Medium	High
Low	86.34	6.31	7.36
Medium	7.77	84.28	7.96
High	6.59	7.16	86.25

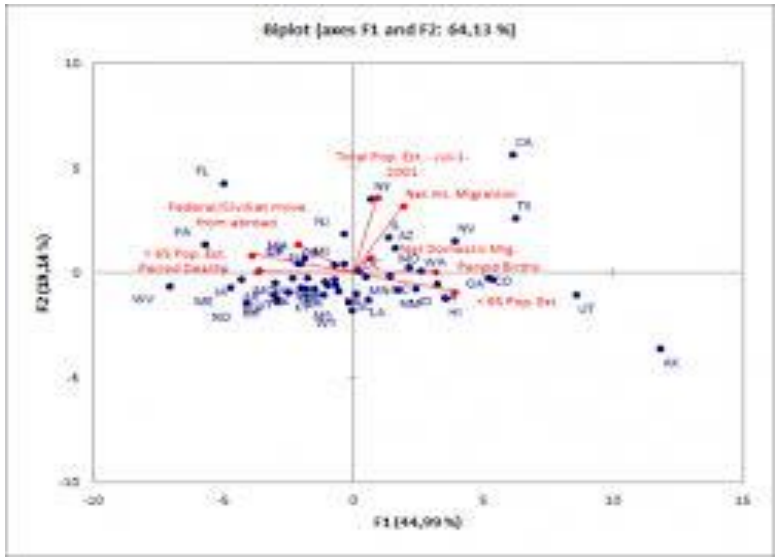
Análisis de Componentes principales ACP

En estadística, el **análisis de componentes principales** (en español ACP, en inglés, PCA) es una técnica utilizada para describir un conjunto de datos en términos de nuevas variables ("**componentes**") no correlacionadas. ... El ACP se emplea sobre todo en **análisis** exploratorio de datos y para construir modelos predictivos.

Para estudiar las relaciones que se presentan entre p variables correlacionadas (que miden información común) se puede transformar el conjunto original de variables en otro conjunto de nuevas variables incorreladas entre sí (que no tenga repetición o redundancia en la información) llamado conjunto de componentes principales.
Las nuevas variables son combinaciones lineales de las anteriores y se van construyendo según el orden de importancia en cuanto a la variabilidad total que recogen de la muestra

Análisis de Componentes principales ACP

[Caso USA Arrest
https://rpubs.com/rpizarro/584627](https://rpubs.com/rpizarro/584627)



Digrama ROC

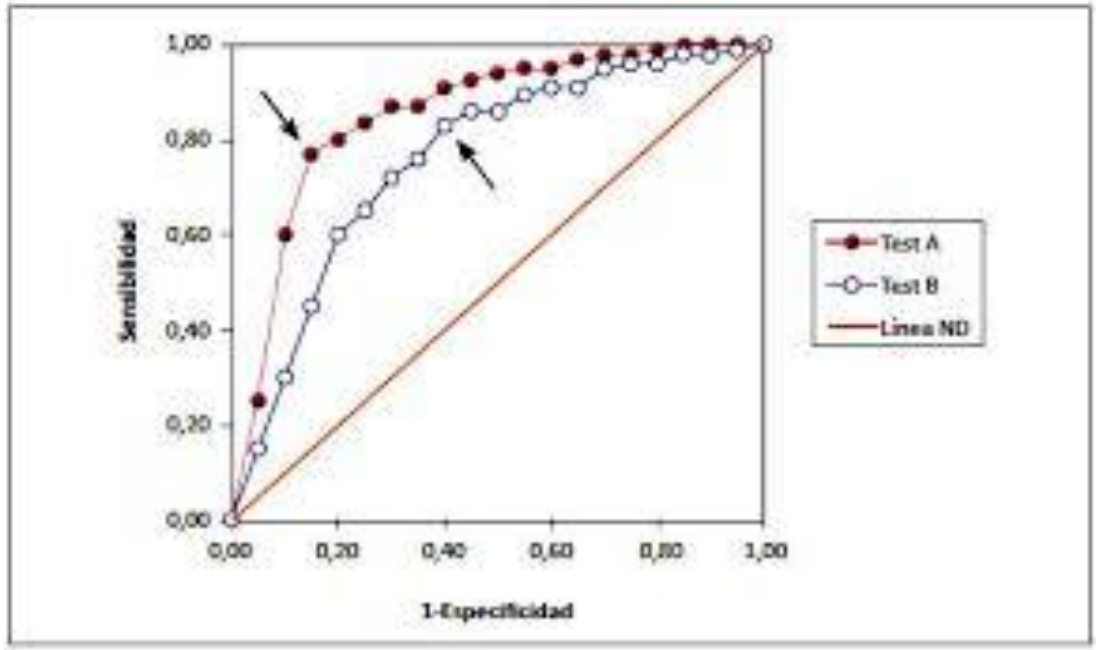
Una curva ROC (acrónimo de Receiver Operating Characteristic, o Característica Operativa del Receptor) es una representación gráfica de la sensibilidad frente a la especificidad para un sistema clasificador binario según se varía el umbral de discriminación.

Otra interpretación de este gráfico es la representación de la razón o ratio de verdaderos positivos (VPR = Razón de Verdaderos Positivos) frente a la razón o ratio de falsos positivos (FPR = Razón de Falsos Positivos) también según se varía el umbral de discriminación.

Como algoritmo de clasificación podría ser útil para determinar la certezaa de que un caso tenga **éxito o fracaso**

Diagrama ROC

https://rpubs.com/rpizarro/584672



Arboles de decisión para clasificación

- **CART**: **Classification And Regression Trees**. Esta es una técnica de aprendizaje supervisado que se puede utilizar para clasificación

- Se tiene una variable objetivo (dependiente) y nuestra meta es obtener una **función** que permita predecir, a partir de variables predictoras (independientes), el valor de la variable objetivo para casos desconocidos.

- Lo que hace este algoritmo es encontrar la variable independiente que mejor separa nuestros datos en grupos, que corresponden con las categorías de la variable objetivo. Esta mejor separación es expresada con una **regla**. A cada **regla** corresponde un **nodo**.

Arboles de decisión para clasificación

https://rpubs.com/rpizarro/584684