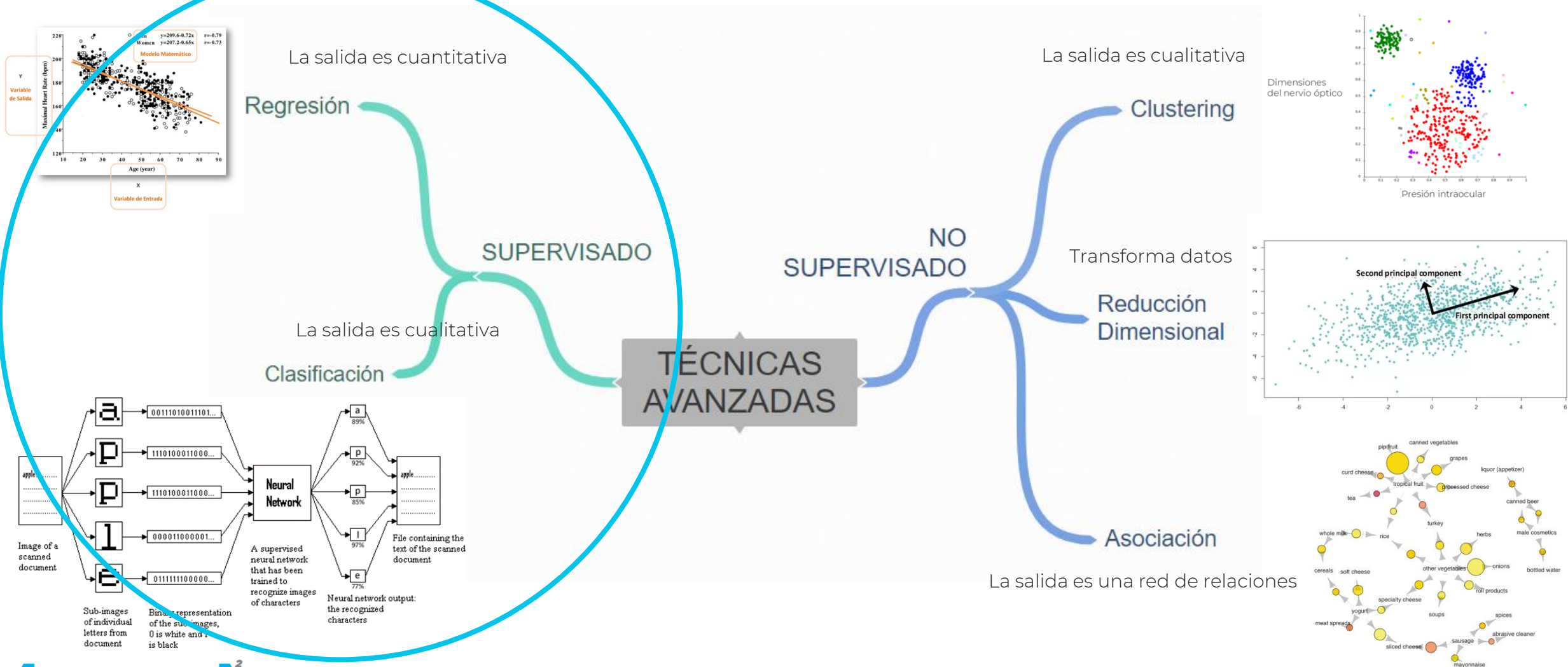


Las claves de los modelos predictivos (ML-supervisado)

Te muestro las claves para llevar a cabo grandes algoritmos
y modelos de predicción

El mapa de aplicaciones prácticas de ciencia de datos



¿Cuáles son las claves?

- El resampling o la partición de los datos en training data y validation data
- Las métricas para evaluar un modelo
- Herramientas gráficas de comparación de modelos (con el uso de las métricas)

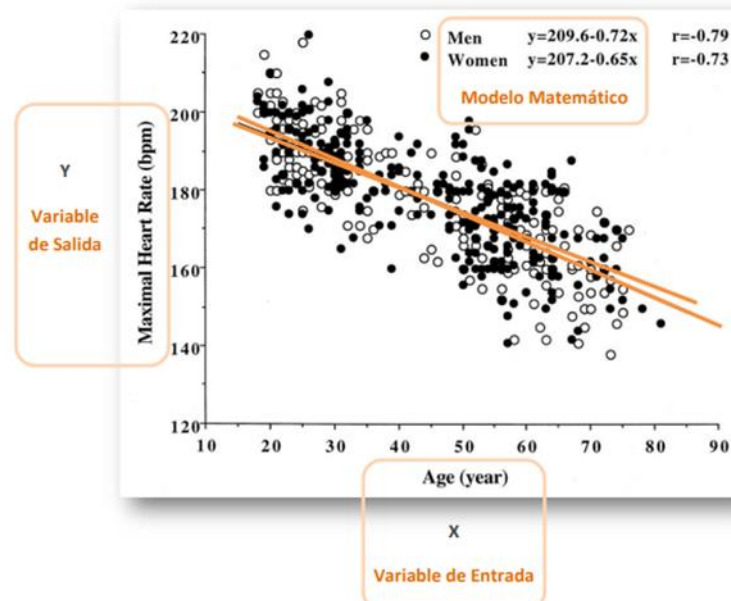
¿Cómo validar un modelo?

Cómo dividir los datos de forma inteligente

Existen 4 maneras de dividir los datos

- Partir por una proporción 70/30
 - 70% - Training Data
 - 30% - Validation Data
- Bootstrapping – para generar muestras aleatorias similares a los datos
- K-fold – “k” grupos obtenidos de los datos y con “L” repeticiones
- LOOCV

Aplicando ciencia de datos – SUPERVISADO



SUPERVISADO

EL PROGRAMA NECESITA
INFORMACIÓN
DE LAS ENTRADAS Y LAS SALIDAS

Paso 1 – Partir los datos 70 - 30

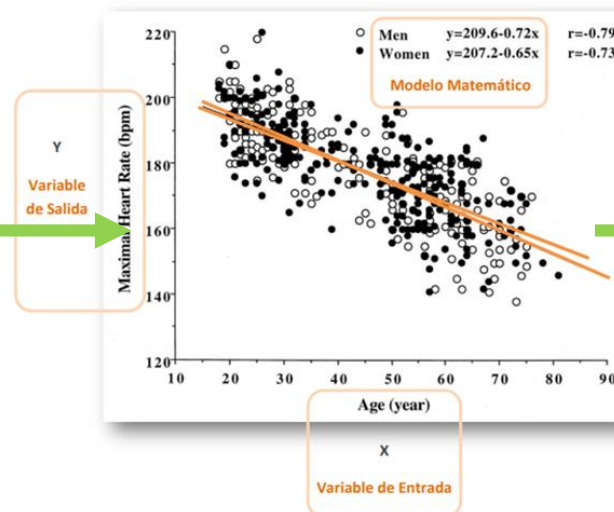
ETAPA DE TRAINING TRAINING DATA

Cálculo del modelo en función de los datos que tenemos



VARIABLES
DE
ENTRADA

70 %



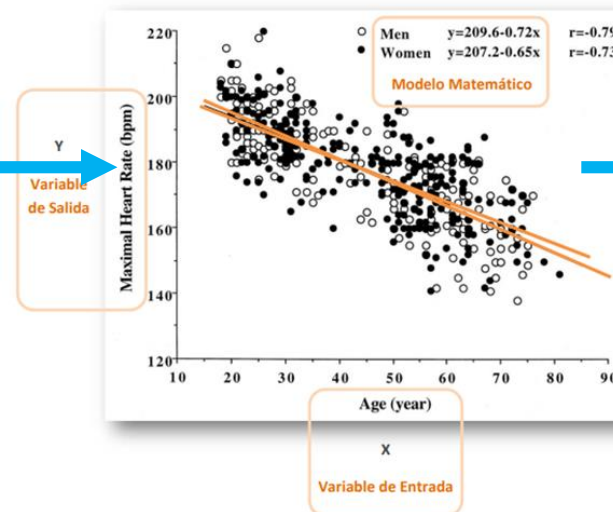
VARIABLE
DE
SALIDA

ETAPA DE PREDICCIÓN VALIDATION DATA

Predicción de la salida con nuevos datos de entrada

VARIABLES
DE
ENTRADA

30%



PREDICCIÓN
VARIABLE DE
SALIDA

Paso 2 – Aplicar un método para generar subgrupos

ETAPA DE TRAINING
TRAINING DATA

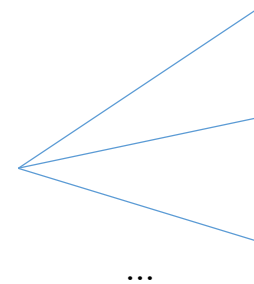
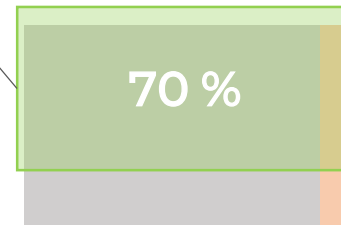
Cálculo del modelo en función de los datos que tenemos



RESAMPLING DATA
GENERANDO GRUPOS DE
TRAINING DATA

Esto nos ayuda a mejorar el error del modelo y el overfitting

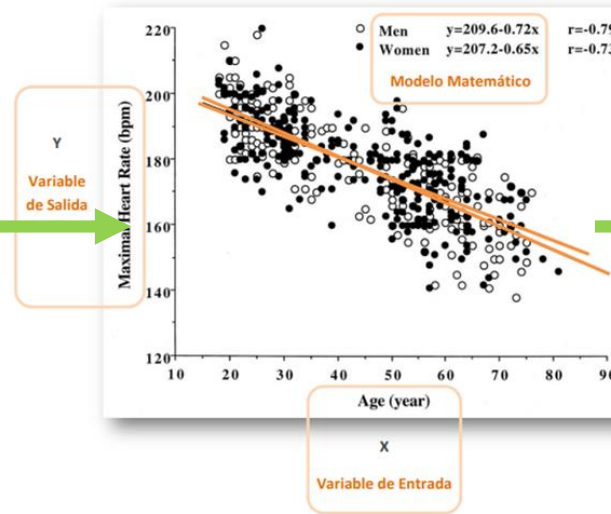
VARIABLES
DE
ENTRADA



...



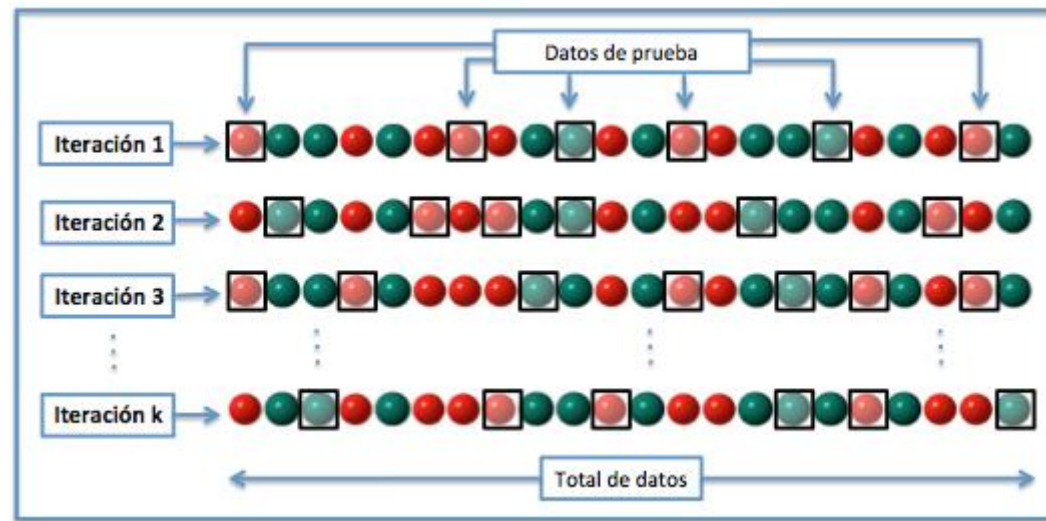
...



VARIABLE
DE
SALIDA

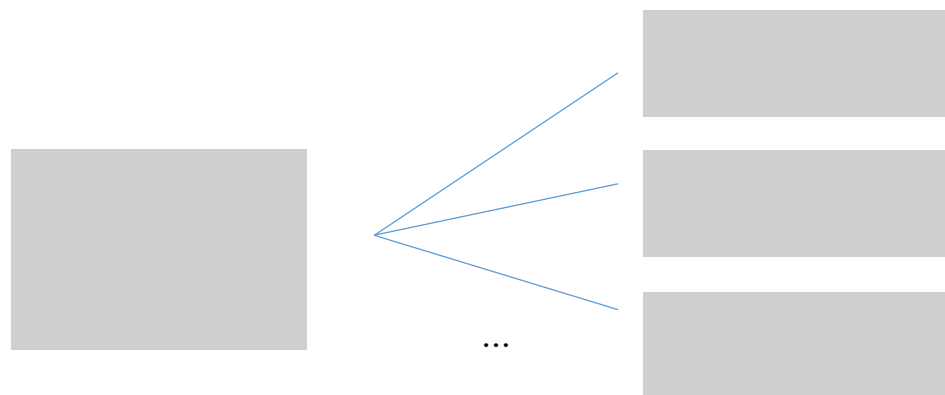
- Crear 10 grupos de 3 repeticiones – **k-fold**
- Crear 100 nuevos grupos generando la muestra a partir de la distribución de training data- **bootstrapping**
- Quitar una variable de entrada cada vez - **LOOCV**

Bootstrapping



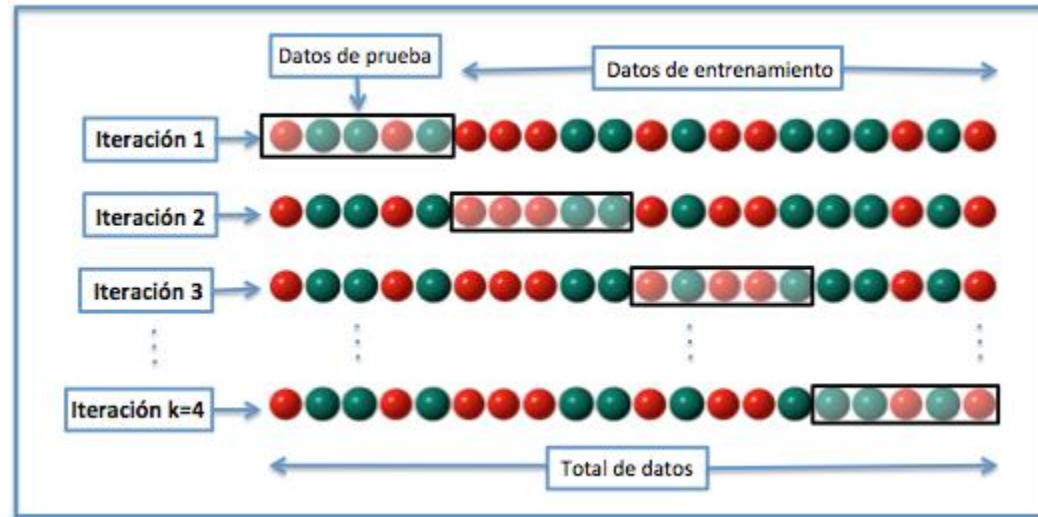
RESAMPLING DATA
GENERANDO GRUPOS DE
TRAINING DATA

Esto nos ayuda a mejorar el
error del modelo y el
overfitting



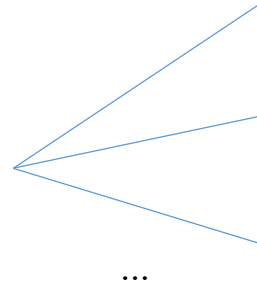
- Crear 10 grupos de 3 repeticiones – **k-fold**
- Crear 100 nuevos grupos generando la muestra a partir de la distribución de training data- **bootstrapping**
- Quitar una variable de entrada cada vez - **LOOCV**

K iteraciones



RESAMPLING DATA
GENERANDO GRUPOS DE
TRAINING DATA

Esto nos ayuda a mejorar el
error del modelo y el
overfitting



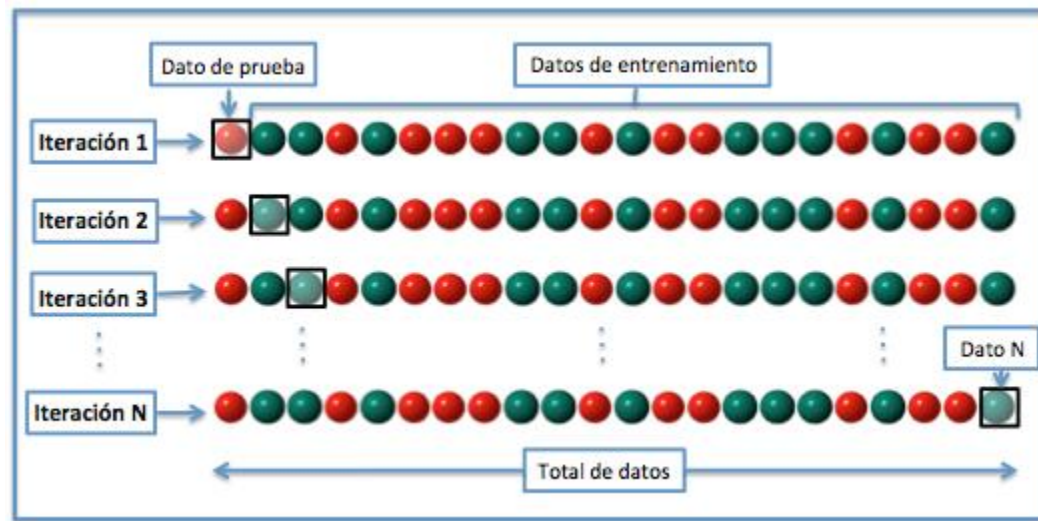
...



...

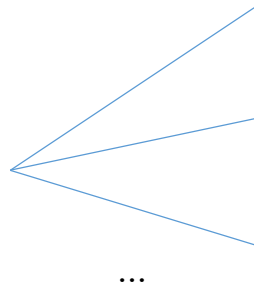
- Crear 10 grupos de 3 repeticiones – **k-fold**
- Crear 100 nuevos grupos generando la muestra a partir de la distribución de training data- **bootstrapping**
- Quitar una variable de entrada cada vez - **LOOCV**

LOOCV - Leave-one-out cross-validation



RESAMPLING DATA
GENERANDO GRUPOS DE
TRAINING DATA

Esto nos ayuda a mejorar el
error del modelo y el
overfitting



...

...

- Crear 10 grupos de 3 repeticiones – **k-fold**
- Crear 100 nuevos grupos generando la muestra a partir de la distribución de training data- **bootstrapping**
- Quitar una variable de entrada cada vez - **LOOCV**

Take away

El resumen de la lección

Lo más importante de la lección

- Para mejorar el error y evitar el overfitting utilizamos 2 pasos:
 - Paso 1 – dividir los datos en dos grupos (70/30)
 - Paso 2 – generar resampling de los datos para mejorar el error con
 - Creando nuevas muestras – bootstrapping
 - Creando pequeños grupos con repeticiones – 10 grupos y 5 repeticiones se utiliza mucho
 - LOOCV – quitando variables una a una

Tú turno

Un pequeño ejercicio para empezar

Tú turno

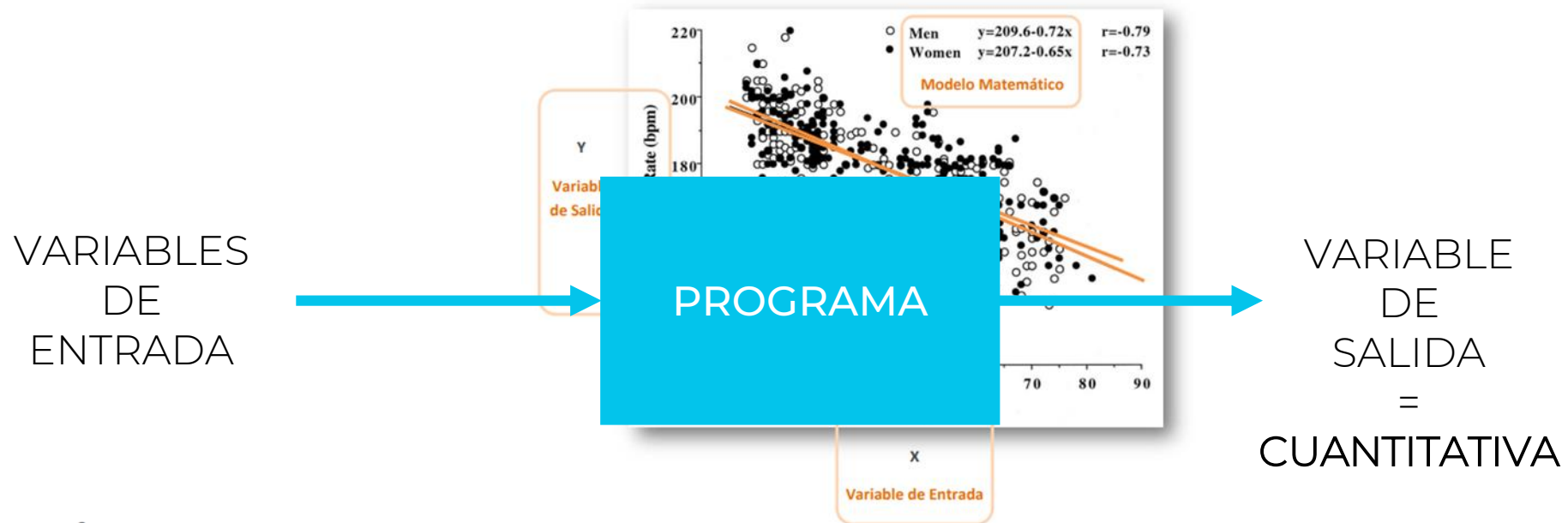
- Descarga la hoja de trabajo
- Utiliza estos métodos para calcular un mismo modelo y compara el accuracy que es una métrica para saber lo buena que es la predicción del modelo
- ¡A por ello!

Métricas para comparar modelos

Conoce las métricas para saber lo bueno que es un modelo

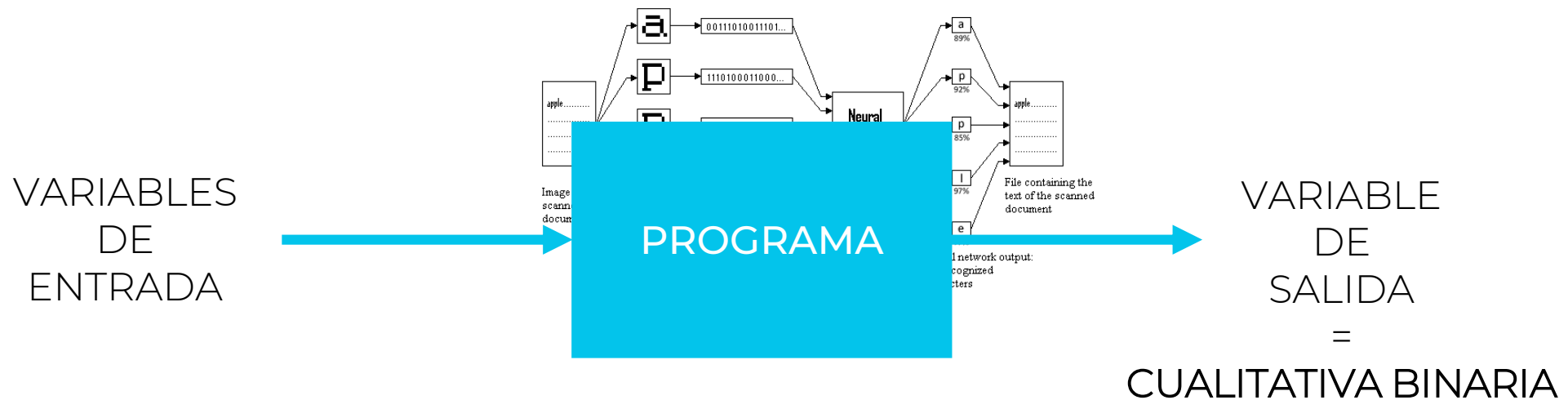
Métricas de Regresión

- RMS – root mean square – error cuadrático medio
$$\text{RMSD} = \sqrt{\frac{\sum_{t=1}^T (\hat{y}_t - y_t)^2}{T}}.$$
- R-squared – R cuadrado – en tanto por 1, es de alguna forma el RMSD normalizado



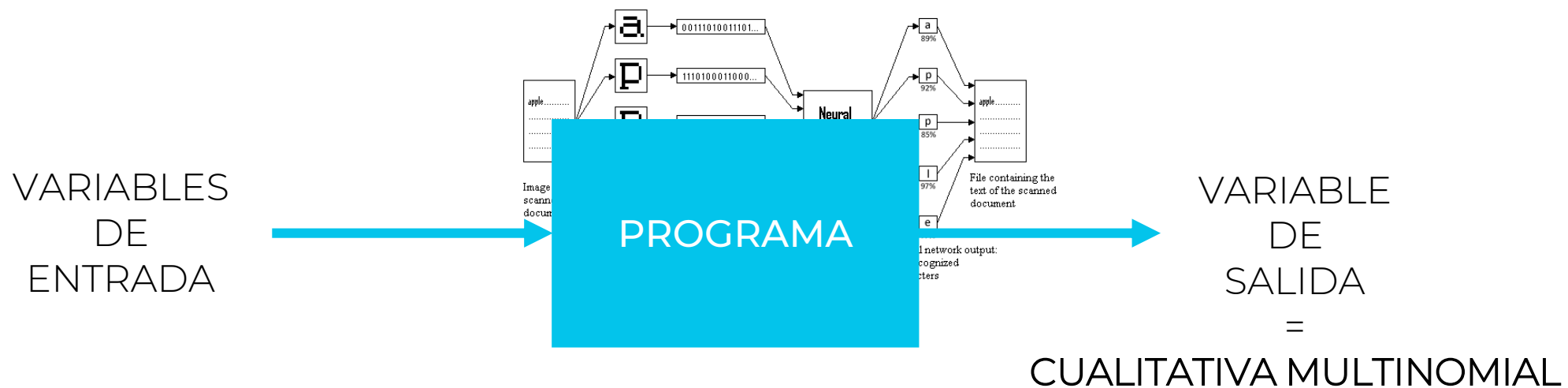
Métricas de Clasificación Binaria

- Kappa – para problemas con las clases no balanceadas
- Accuracy - % de clases bien clasificadas del total (problemas con clases balanceadas)
- AUROC – Área debajo la curva ROC, de 0 a 1



Métricas de Clasificación Multiclase

- Kappa – para problemas con las clases no balanceadas
- Accuracy - % de clases bien clasificadas del total (problemas con clases balanceadas)
- Logarithmic Loss – Utilizada para clasificadores multiclase



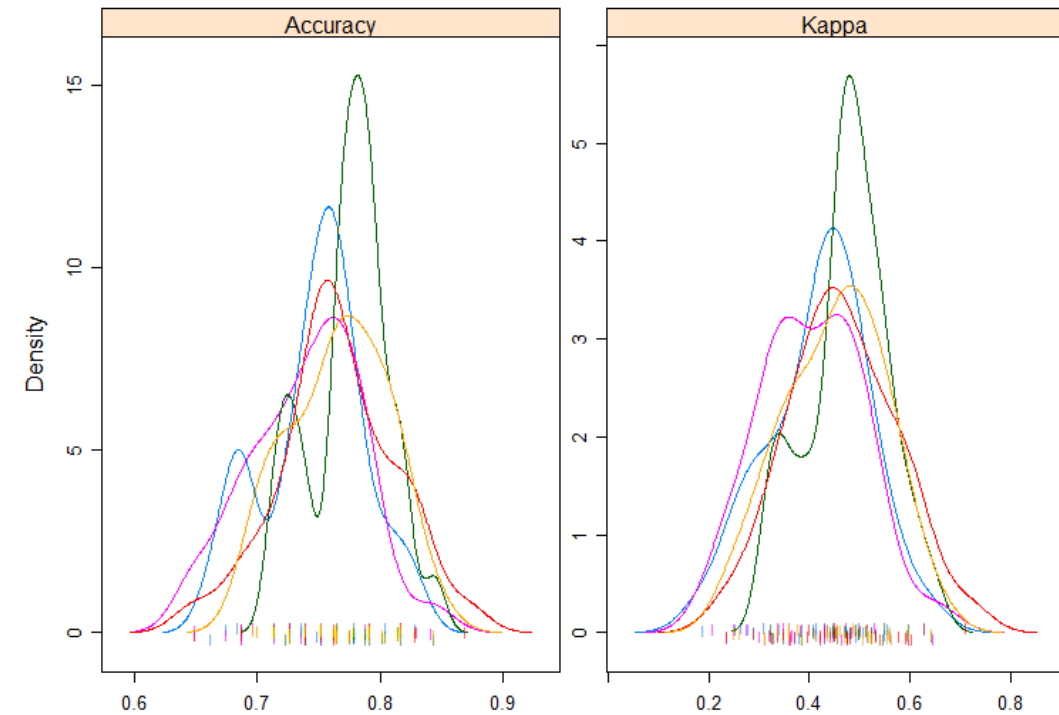
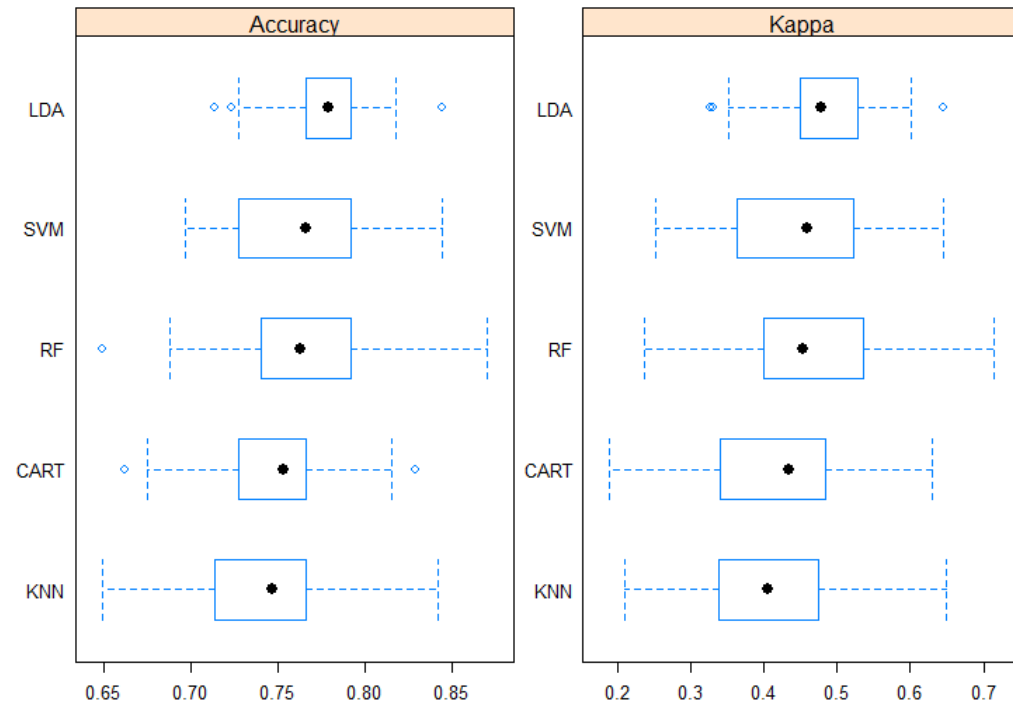
Gráficos para comparar modelos con la ayuda de las métricas

Gráficos interesantes para comparar modelos

Ejemplos para comparar modelos

Accuracy							
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
CART	0.6623377	0.7272727	0.7532468	0.7465596	0.7662338	0.8289474	0
LDA	0.7142857	0.7662338	0.7792208	0.7747608	0.7922078	0.8441558	0
SVM	0.6973684	0.7305195	0.7662338	0.7665243	0.7922078	0.8441558	0
KNN	0.6493506	0.7142857	0.7467532	0.7400832	0.7662338	0.8421053	0
RF	0.6493506	0.7402597	0.7631579	0.7643825	0.7922078	0.8701299	0

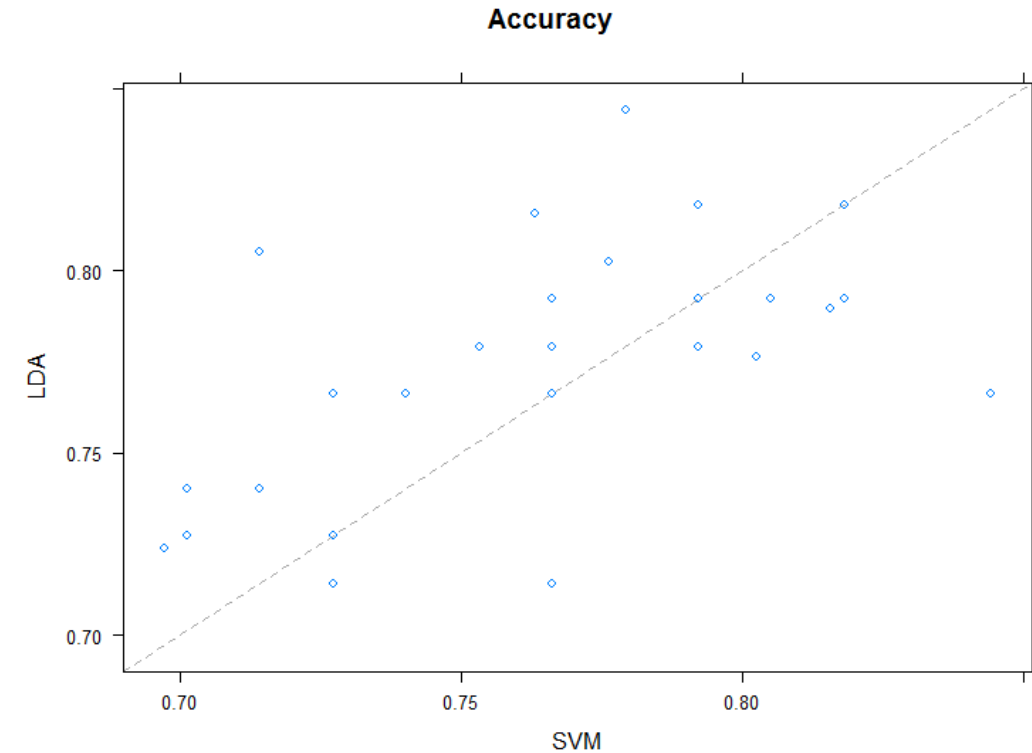
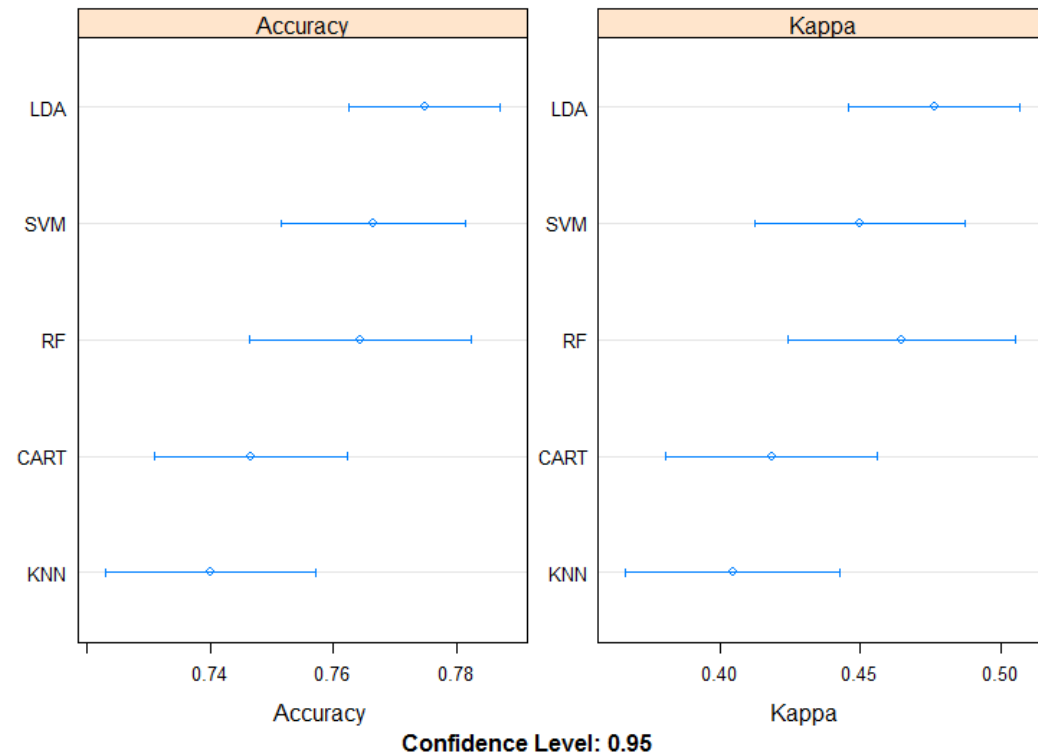
Kappa							
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
CART	0.1894737	0.3495477	0.4340426	0.4184446	0.4827744	0.6302395	0
LDA	0.3267091	0.4491256	0.4784008	0.4764965	0.5276074	0.6457055	0
SVM	0.2517123	0.3670435	0.4590164	0.4500126	0.5211405	0.6457055	0
KNN	0.2098062	0.3388961	0.4058531	0.4047018	0.4733369	0.6492308	0
RF	0.2365039	0.3997116	0.4535834	0.4649871	0.5337481	0.7148148	0



Ejemplos para comparar modelos

Accuracy							
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
CART	0.6623377	0.7272727	0.7532468	0.7465596	0.7662338	0.8289474	0
LDA	0.7142857	0.7662338	0.7792208	0.7747608	0.7922078	0.8441558	0
SVM	0.6973684	0.7305195	0.7662338	0.7665243	0.7922078	0.8441558	0
KNN	0.6493506	0.7142857	0.7467532	0.7400832	0.7662338	0.8421053	0
RF	0.6493506	0.7402597	0.7631579	0.7643825	0.7922078	0.8701299	0

Kappa							
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
CART	0.1894737	0.3495477	0.4340426	0.4184446	0.4827744	0.6302395	0
LDA	0.3267091	0.4491256	0.4784008	0.4764965	0.5276074	0.6457055	0
SVM	0.2517123	0.3670435	0.4590164	0.4500126	0.5211405	0.6457055	0
KNN	0.2098062	0.3388961	0.4058531	0.4047018	0.4733369	0.6492308	0
RF	0.2365039	0.3997116	0.4535834	0.4649871	0.5337481	0.7148148	0



Take away

El resumen de la lección

Lo más importante de la lección

- Las métricas para regresión son:
 - RMS
 - R cuadrado
- Las métricas para la clasificación son:
 - Kappa
 - Accuracy
 - Logarithmic loss (muticlase)
 - AUROC (binaria)

Tú turno

Un pequeño ejercicio para empezar

Tú turno

- Descarga la hoja de trabajo que te he preparado
- Utiliza los scripts para calcular diferentes parámetros y compararlos con la ayuda de los gráficos que has visto
- ¡A por ello!