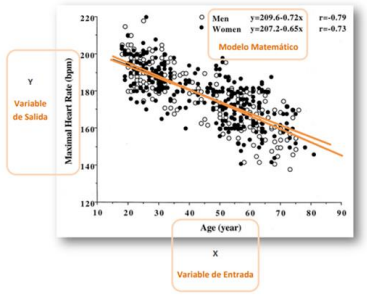
Algoritmos y técnicas de clasificación

Conoce las técnicas más importantes de clasificación

El mapa de aplicaciones prácticas de ciencia de datos



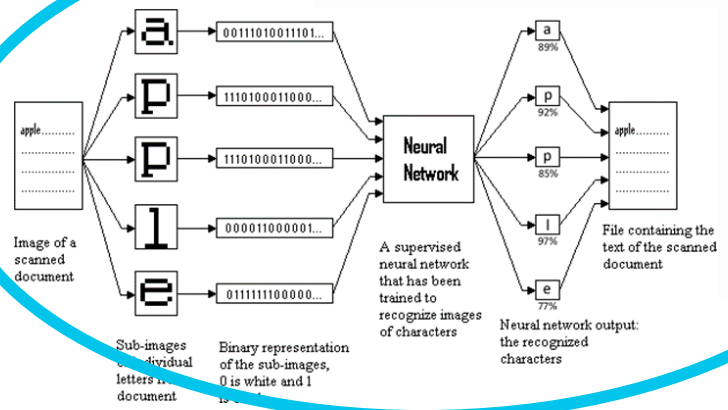
La salida es cuantitativa

Regresión

SUPERVISADO

La salida es cualitativa

Clasificación

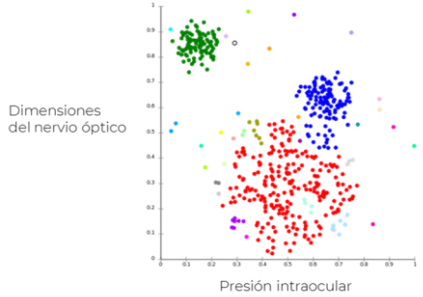


TÉCNICAS AVANZADAS

NO SUPERVISADO

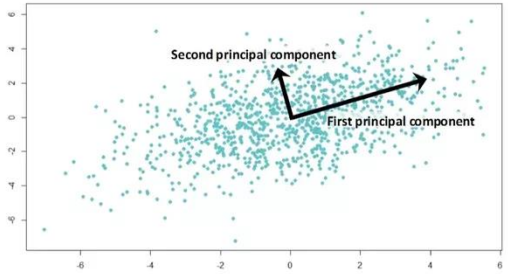
La salida es cualitativa

Clustering



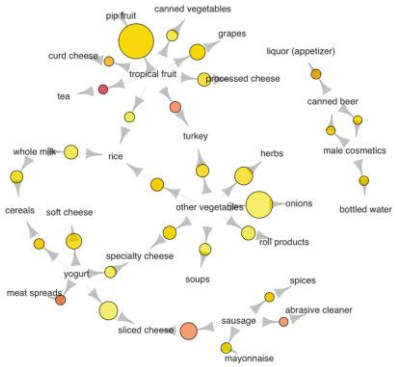
Transforma datos

Reducción Dimensional



Asociación

La salida es una red de relaciones

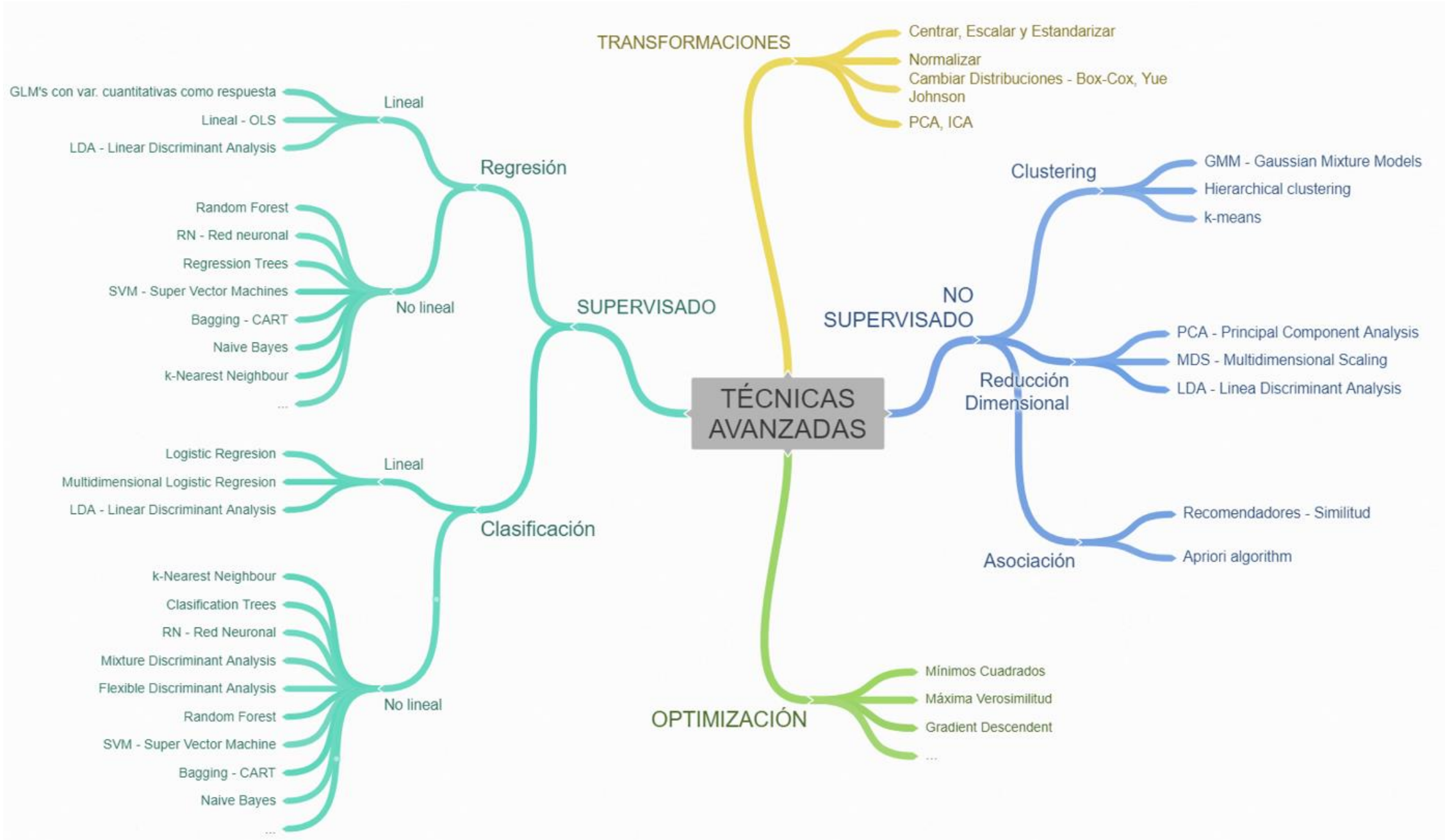


Conceptosclaros

Las técnicas de clasificación

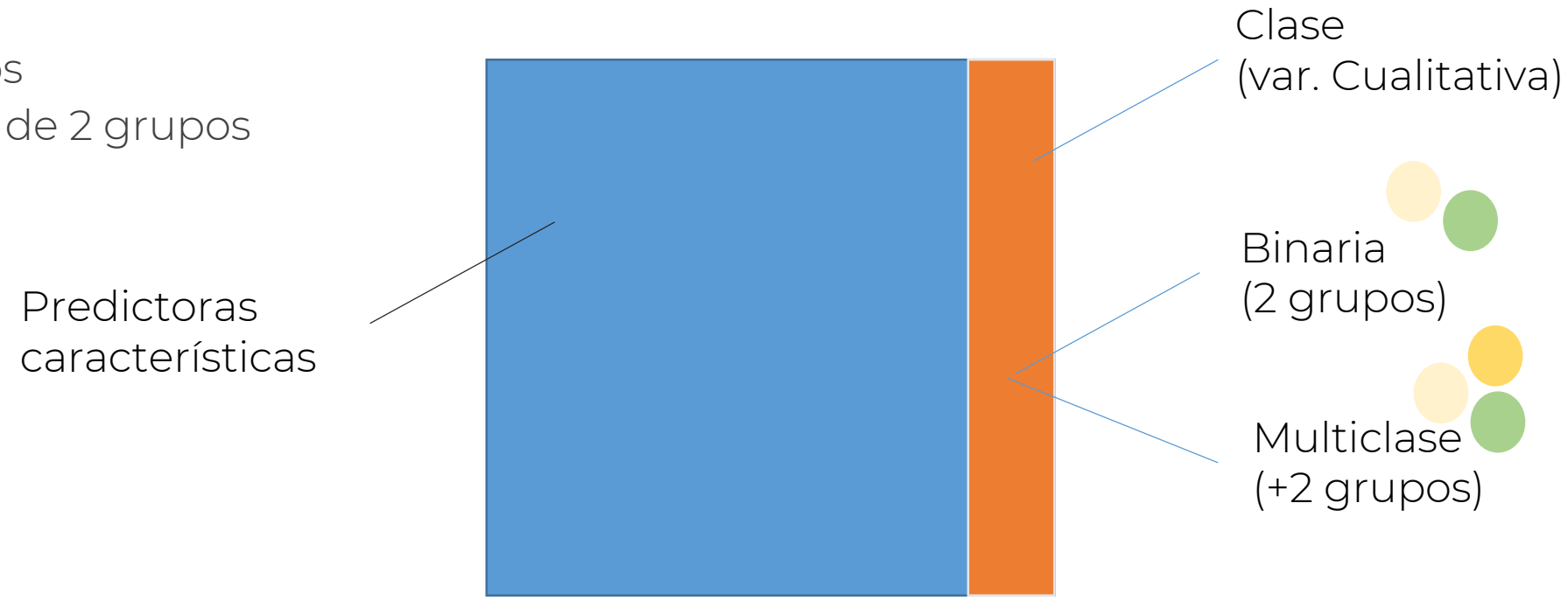
- Lineales
  - Regresión logística binaria
  - Regresión logística multinomial
  - LDA – lineal discriminant analysis

- NO lineales
  - CART – Árboles de clasificación y regresión
  - Random forest – Árboles aleatorios (generalización de CART)
  - Naive Bayes
  - KNN – k vecinos más cercanos
  - SVM – super vectors machine
  - NN – redes neuronales
  - Algoritmos de bagging y boosting


TÉCNICAS AVANZADAS
SUPERVISADO
Regresión
Lineal
GLM's con var. cuantitativas como respuesta
Lineal - OLS
LDA - Linear Discriminant Analysis
No lineal
Random Forest
RN - Red neuronal
Regression Trees
SVM - Super Vector Machines
Bagging - CART
Naive Bayes
k-Nearest Neighbour
...
Clasificación
Lineal
Logistic Regresion
Multidimensional Logistic Regresion
LDA - Linear Discriminant Analysis
No lineal
k-Nearest Neighbour
Clasification Trees
RN - Red Neuronal
Mixture Discriminant Analysis
Flexible Discriminant Analysis
Random Forest
SVM - Super Vector Machine
Bagging - CART
Naive Bayes
...
TRANSFORMACIONES
Centrar, Escalar y Estandarizar
Normalizar
Cambiar Distribuciones - Box-Cox, Yue Johnson
PCA, ICA
NO SUPERVISADO
Clustering
GMM - Gaussian Mixture Models
Hierarchical clustering
k-means
Reducción Dimensional
PCA - Principal Component Analysis
MDS - Multidimensional Scaling
LDA - Linea Discriminant Analysis
Asociación
Recomendadores - Similitud
Apriori algorithm
OPTIMIZACIÓN
Mínimos Cuadrados
Máxima Verosimilitud
Gradient Descendent
...

Cómo ordenar las variables en un algoritmo de clasificación

- Las predictoras – o features (normalmente cuantitativas, pueden ser cualitativas también)
- La clase
  - Binaria – 2 grupos
  - Multiclase – más de 2 grupos

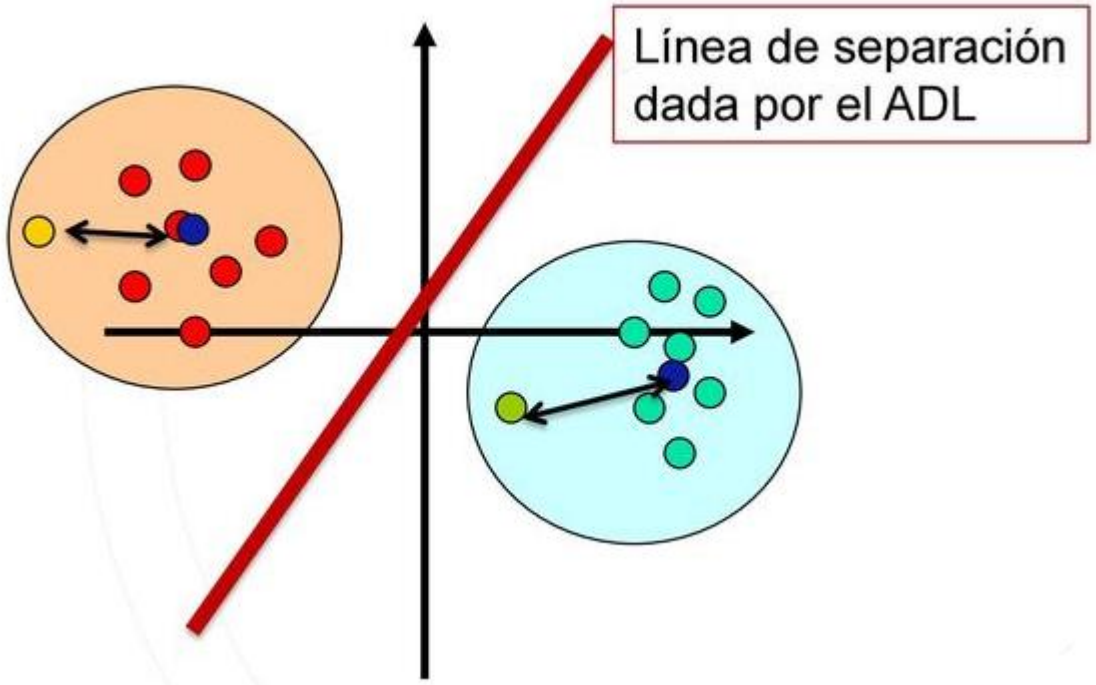


LDA - Linear Discriminant Analysis

Conoce el análisis de discriminante lineal

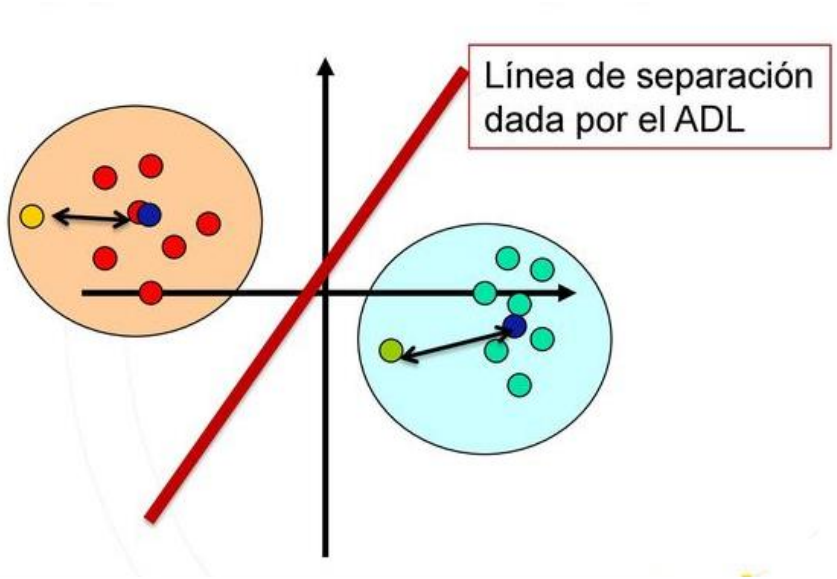
LDA – Análisis de Discriminante Lineal

- Trata de calcular rectas que separen los diferentes grupos.
- Se utiliza como clasificador multicase ya que la regresión logística se utiliza para clasificaciones binaria
- Las predictoras deben ser cuantitativas



LDA – Análisis de Discriminante Lineal

- Calcula una parámetro llamada discriminante para cada punto y para cada clase
- Es de alguna forma el valor que indica lo cercano que está del grupo. Cuanto más grande el discriminante más cerca está del grupo
- Identificará el grupo con el valor más grande



Medida del discriminante

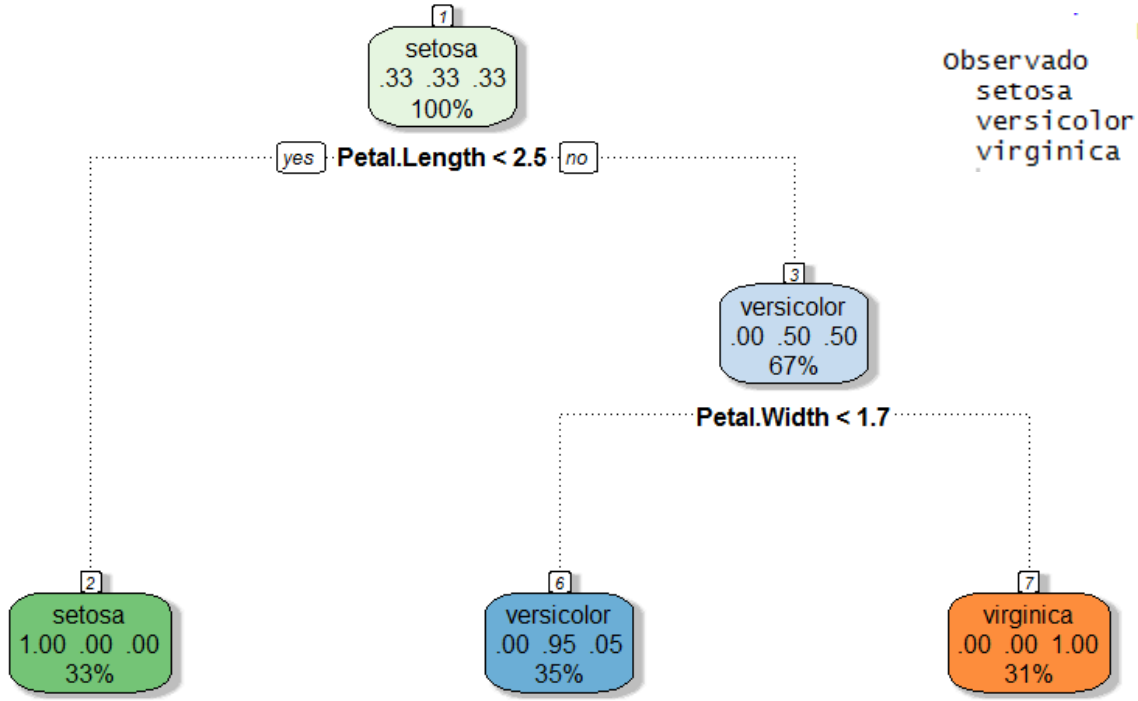
$$D_k(x) = x \times \frac{mean_k}{sigma^2} - \frac{{mean_k}^2}{2 \times sigma^2} + ln(P(k))$$

CART – Árboles de decisión

Una técnica muy interesante. Los árboles de clasificación

Árboles de clasificación o de regresión

- Se utilizan para tomar decisiones o para el uso de técnicas de clasificación
- El cálculo se basa en algoritmo de probar cortes y ver qué combinación es la ganadora



	Esperado		
Observado	setosa	versicolor	virginica
setosa	15	0	0
versicolor	0	13	2
virginica	0	2	13

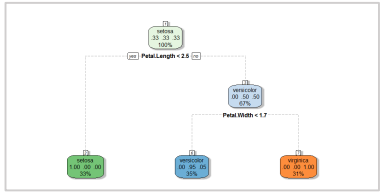
Random Forest

Creando un bosque aleatorio con muchos árboles juntos.
Una herramienta extremadamente potente

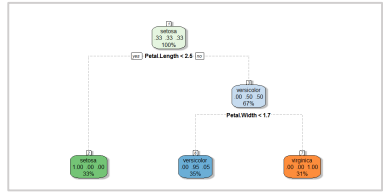
Árboles de clasificación o de regresión

- Es una técnica más potente que los CART porque evita el overfitting y mejora el error. De hecho es utilizar muchos árboles para crear un bosque
- Es muy costosa computacionalmente. Si tienes muchos datos es difícil trabajarlo
- Es un conglomerado de algoritmos CART, se calculan muchos árboles y se promedian los resultados

Árbol 1

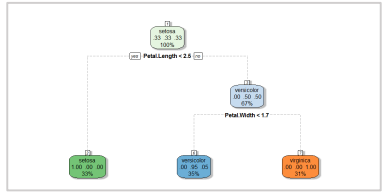


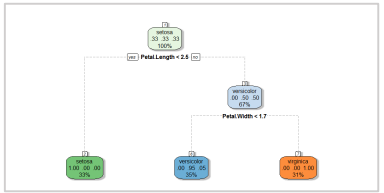
Árbol 2



...

Árbol m





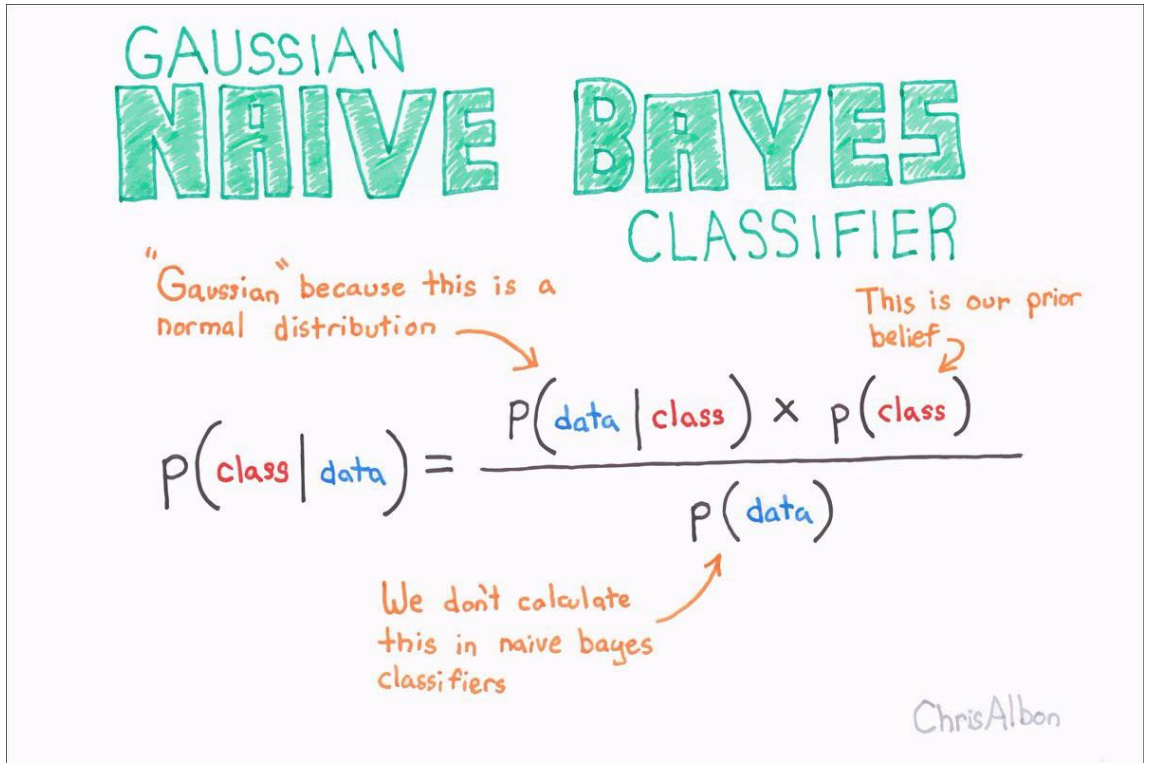
Árbol Promedio

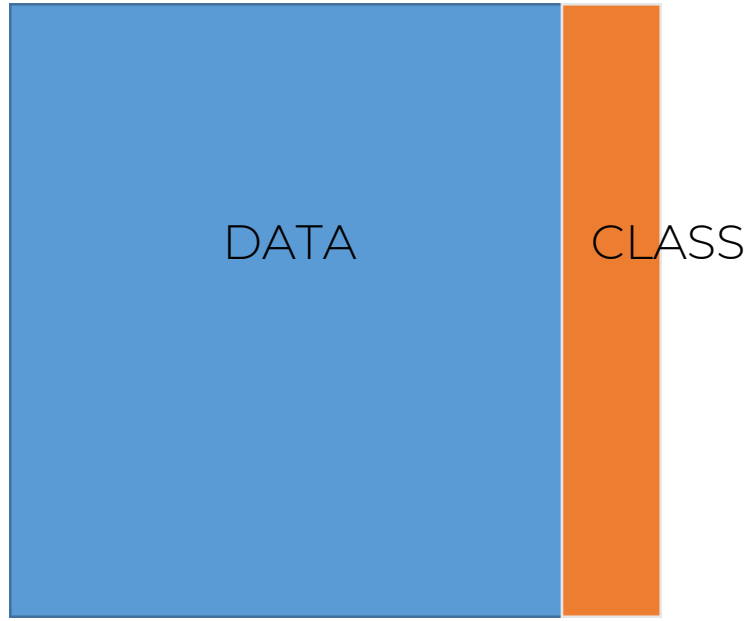
Naive Bayes

El teorema de Bayes para clasificar

Teorema de Bayes para utilizarlo como modelo clasificador

- Utiliza una simplificación del teorema de Bayes para decidir dónde se encuentran los puntos
- Se puede utilizar tanto para datos cuantitativos como cualitativos





Un ejemplo simple para entenderlo

Clasificación de sexo [editar]

Problema: Clasificar una persona en hombre o mujer basándonos en las características de sus medidas: peso, altura y número de pie.

Entrenamiento [editar]

Entrenamiento previo.

sexo	altura (pies)	peso (lbs)	talla del pie (inches)
hombre	6	180	12
hombre	5.92 (5'11")	190	11
hombre	5.58 (5'7")	170	12
hombre	5.92 (5'11")	165	10
mujer	5	100	6
mujer	5.5 (5'6")	150	8
mujer	5.42 (5'5")	130	7
mujer	5.75 (5'9")	150	9

Haciendo una distribución Gaussiana extraemos los datos y obtenemos la media y la varianza de cada característica.

sexo	media (altura)	varianza (altura)	media (peso)	varianza (peso)	media (foot size)	varianza (foot size)
hombre	5.855	0.035033	176.25	122.92000	11.25	0.91667
mujer	5.4175	0.097225	132.5	558.33000	7.5	1.66670

En este caso nos encontramos en una distribución equiprobable, es decir que tienen la misma probabilidad. P(hombre)=0.5 y P(mujer)=0.5.

Testing [editar]

Ahora recibimos unos datos para ser clasificado como hombre o mujer

sex	altura (pies)	peso (lbs)	número de pie(inches)
muestra	6	130	8

Ahora nos interesa saber la probabilidad a posteriori de los dos casos, según es hombre o mujer.

hombre

$$posteriori(hombre) = \frac{P(hombre)\, p(altura|hombre)\, p(peso|hombre)\, p(numerodepie|hombre)}{Evidencia}$$

mujer

$$posteriori(mujer) = \frac{P(mujer)\, p(altura|mujer)\, p(peso|mujer)\, p(numerodepie|mujer)}{Evidencia}$$

Un ejemplo simple para entenderlo

En este caso nos encontramos en una distribución equiprobable, es decir que tienen la misma probabilidad. P(hombre)=0.5 y P(mujer)=0.5.

$$P(\text{hombre}) = 0.5$$

$$p(\text{altura}|\text{hombre}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(6-\mu)^2}{2\sigma^2}\right) \approx 1.5789,$$

donde $\mu = 5.855$ y $\sigma^2 = 0.035033$ son los parámetros de la distribución normal que han sido determinados previamente en el entrenamiento .

$$p(\text{peso}|\text{hombre}) = 5.9881e-06$$

$$p(\text{NumeroDePie}|\text{hombre}) = 1.3112e-3$$

$$\text{posteriori (hombre)} = \text{el producto} = 6.1984e-09$$

$$P(\text{mujer}) = 0.5$$

$$p(\text{altura}|\text{mujer}) = 2.2346e-1$$

$$p(\text{peso}|\text{mujer}) = 1.6789e-2$$

$$p(\text{numero de pie}|\text{mujer}) = 2.8669e-1$$

$$\text{posteriori (mujer)} = \text{el producto} = 5.3778e-04$$

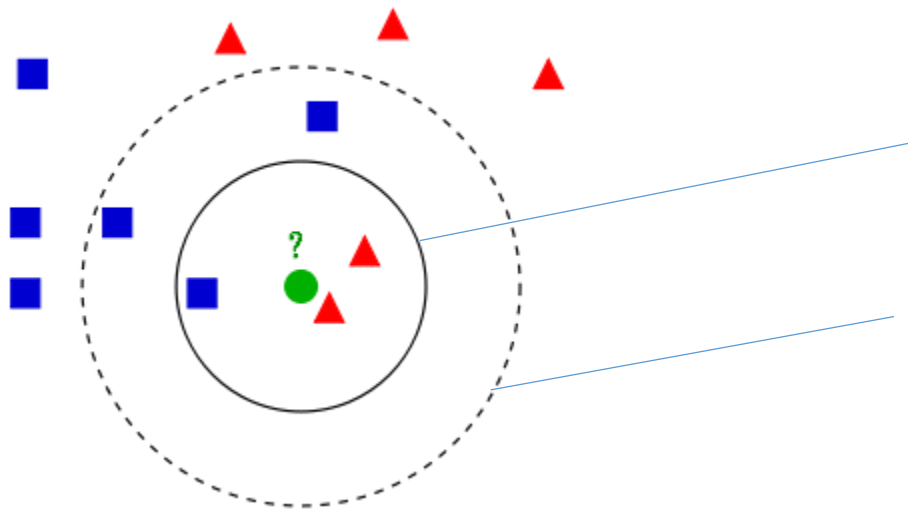
En este caso el numerador a posteriori más grande es el de la mujer, por eso determinamos que los datos son de mujer.

KNN – Los k vecinos más cercanos

Creando k vecindarios en los datos con los puntos más cercanos

K vecinos más cercanos

- Tenemos que calcular k. Es el número de vecinos en el vecindario
- Escoger k – el número de vecinos es clave para el clasificador



Vecindario 1 - ¿Qué clase será?

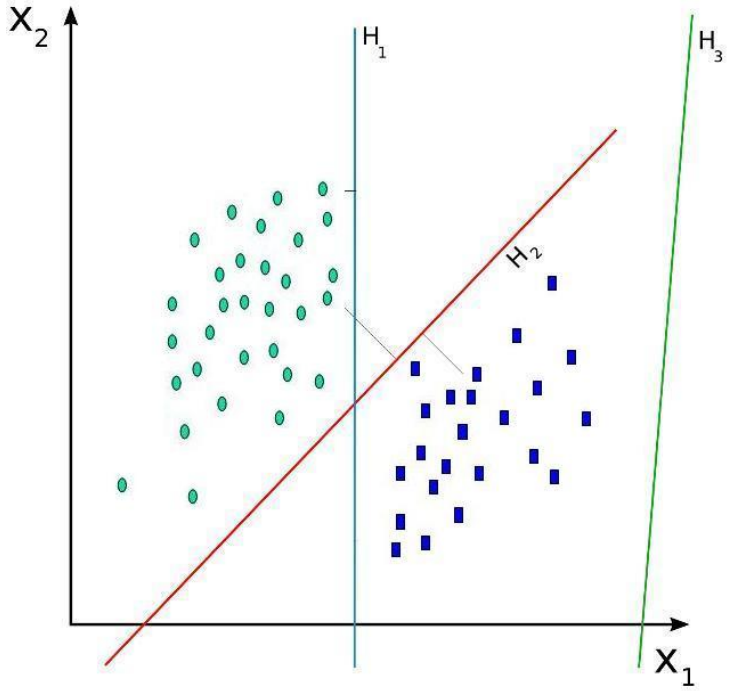
Vecindario 2 - ¿Qué clase será?

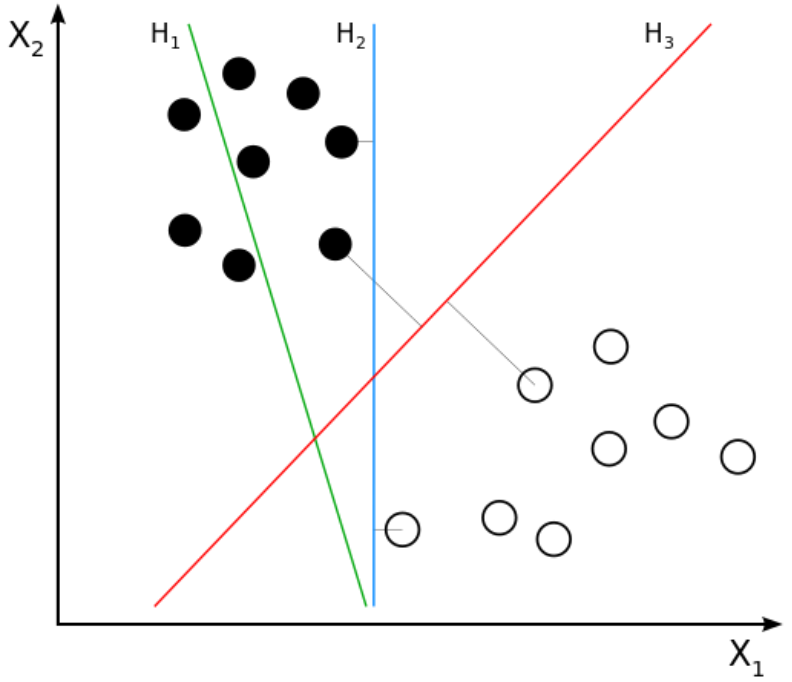
SVM – Super Vectors Machine

Una máquina de vectores. Una herramienta poderosa

La intuición del super vector machine - SVM

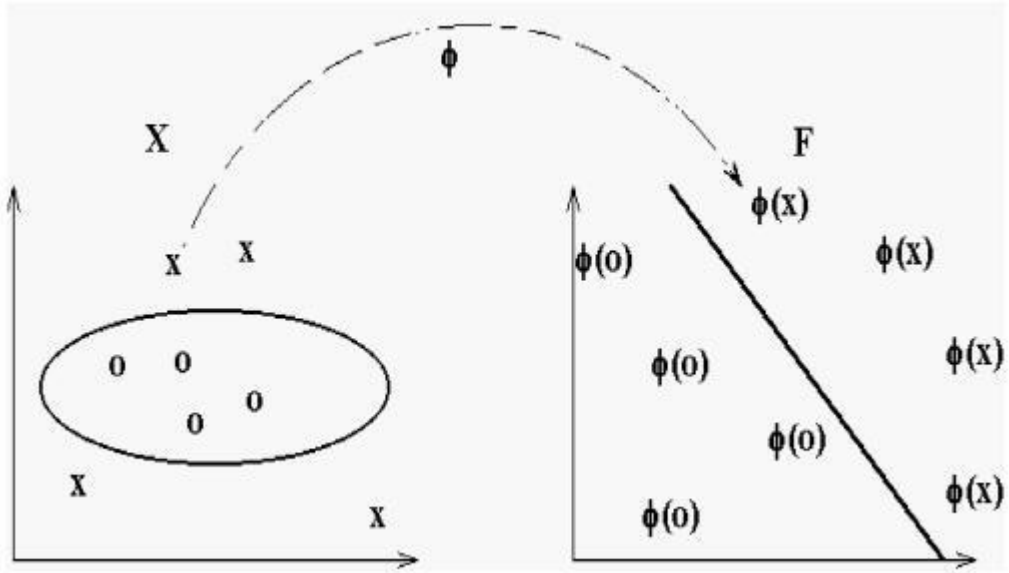
- Trata de calcular vectores (direcciones) de rectas que mejor separen a los grupos
- La recta (hiperplano) con un margen más grande a los puntos de clases será la ganadora





La intuición del super vector machine - SVM

- El enfoque lineal del algoritmo es limitado. Estamos trabajando con planos y líneas rectas.
- La solución es transformar los puntos con una transformada. Una función llamada Kernel. A partir de aquí aparecen muchos tipos de SVM

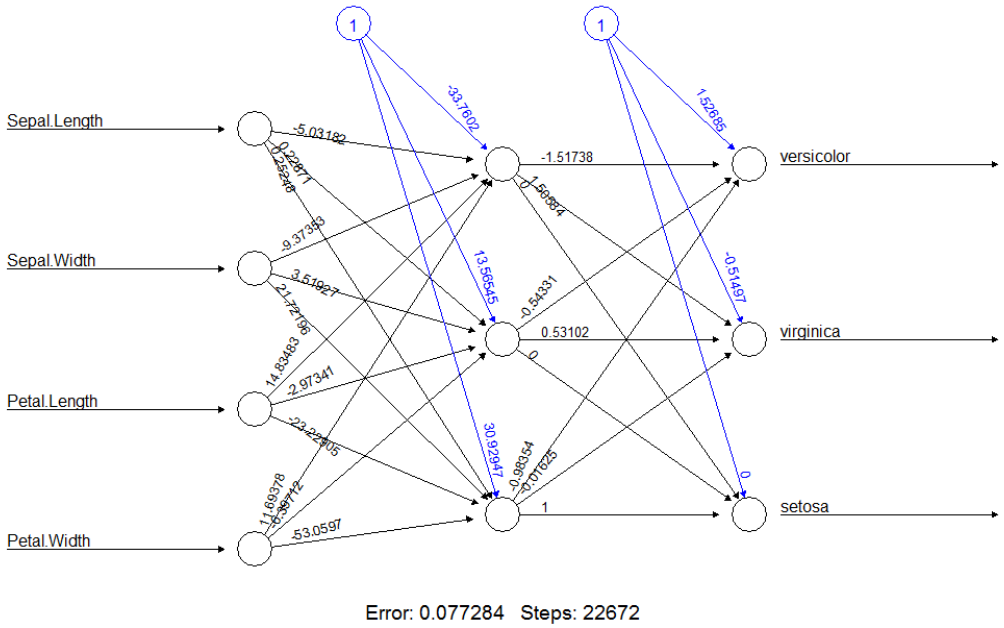


NN – Redes neuronales

Una maravilla de la matemática o la caja negra por excelencia

Las famosas redes neuronales

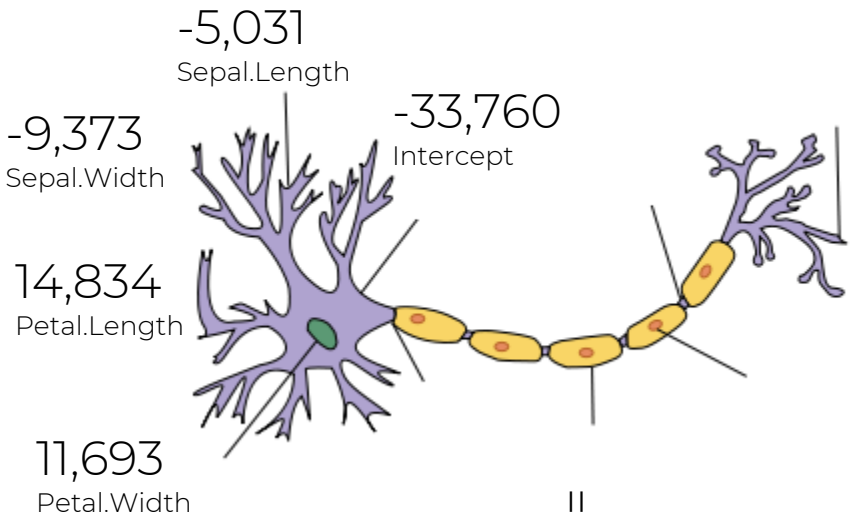
- La caja negra de los algoritmos – Un arma de doble filo
- Las redes neuronales se basan en la idea de neuronas
- Las neuronas están formadas por axones



Probabilidad de ser Versicolor

Probabilidad de ser Virginica

Probabilidad de ser Setosa

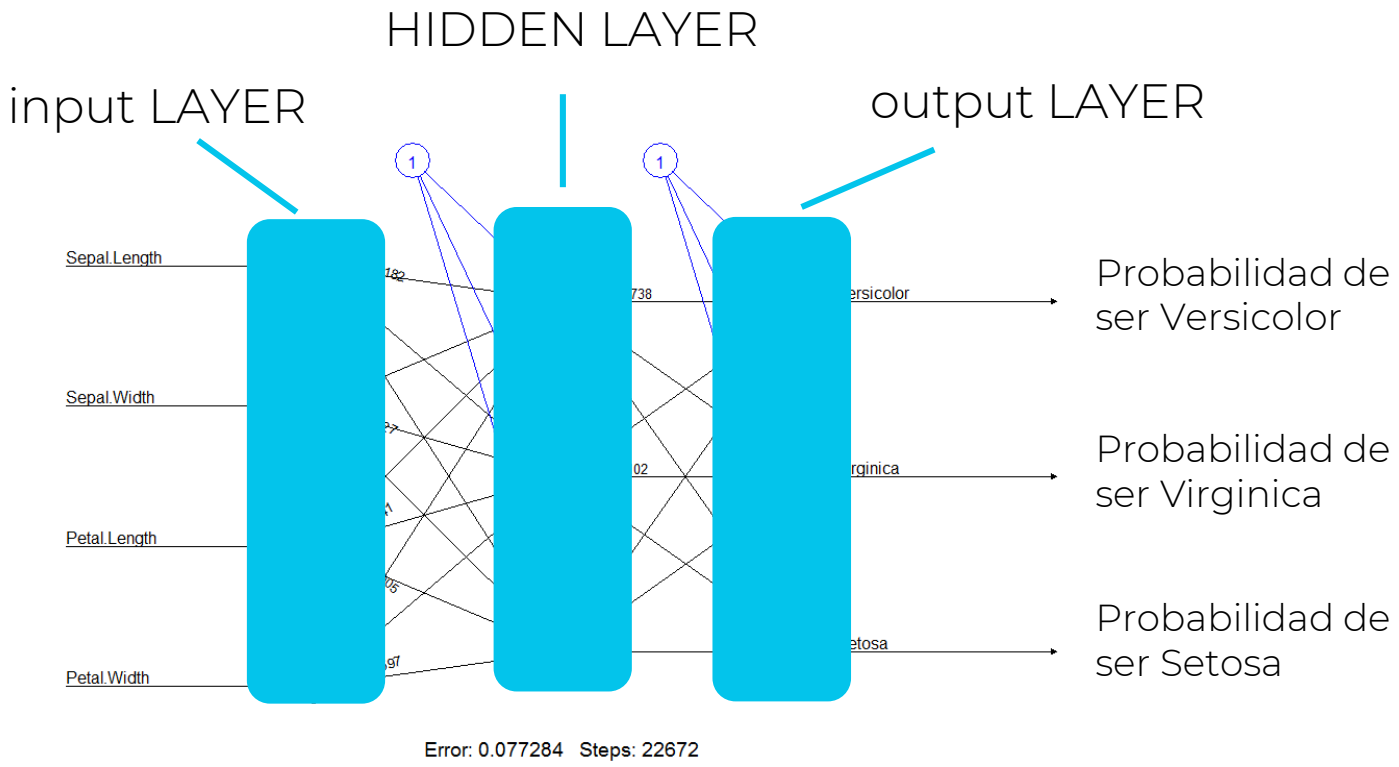


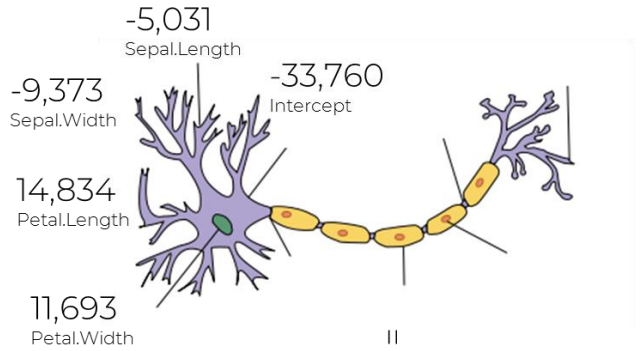
Es una combinación lineal

$$-33{,}760 - 5{,}031 \cdot \text{Sepal.Length} - 9{,}373 \cdot \text{Sepal.Width} + 14{,}834 \cdot \text{Petal.Length} + 11{,}693 \cdot \text{Petal.Width}$$

Las famosas redes neuronales

- La estructura neuronal puede ser tan complicada como se quiera
- Y los axones pueden tener funciones no lineales



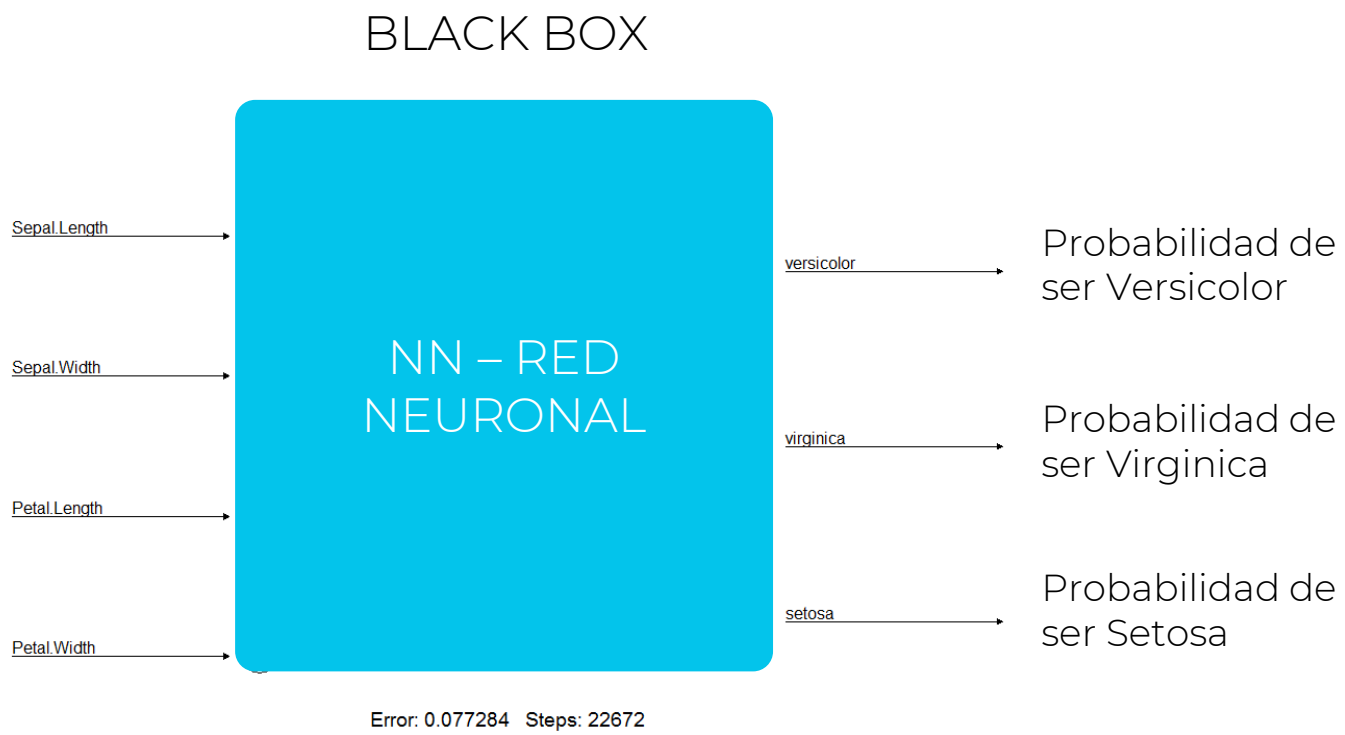


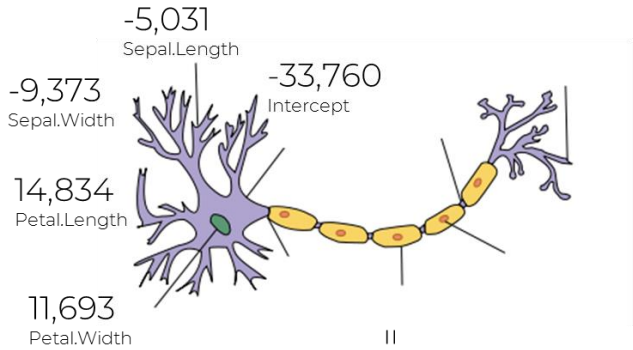
Es una combinación lineal

$$-33{,}760 - 5{,}031 \cdot \text{Sepal.Length} - 9{,}373 \cdot \text{Sepal.Width} + 14{,}834 \cdot \text{Petal.Length} + 11{,}693 \cdot \text{Petal.Width}$$

Las famosas redes neuronales

- La estructura neuronal puede ser tan complicada como se quiera
- Y los axones pueden tener funciones no lineales





Es una combinación lineal

$$-33{,}760 - 5{,}031 \cdot \text{Sepal.Length} - 9{,}373 \cdot \text{Sepal.Width} + 14{,}834 \cdot \text{Petal.Length} + 11{,}693 \cdot \text{Petal.Width}$$

Algoritmos de ensamblaje

Multiplicando el poder de los algoritmos no lineales.
Bagging and boosting

Ensamblando algoritmos - Bagging

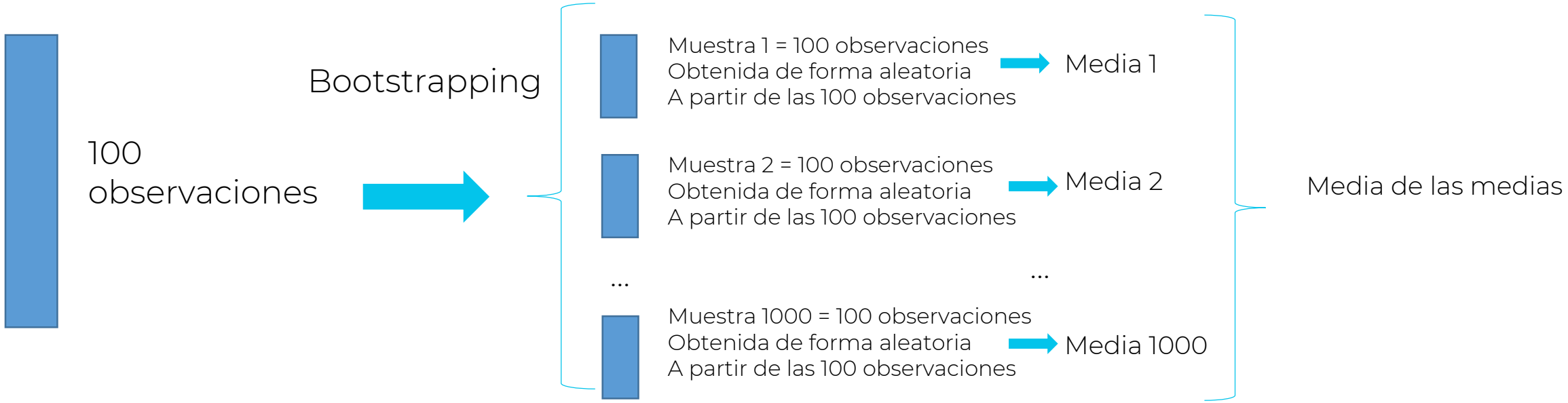
- Los random forest o bosques aleatorios es un algoritmo de ensamblaje
- Se basa en el bootstrapping
- Es una técnica muy buena para poder calcular una medida, una característica o un modelo a base de repeticiones

100
observaciones

Puedes calcular el promedio como sabes
Sumando las 100 observaciones y dividiendo por 100

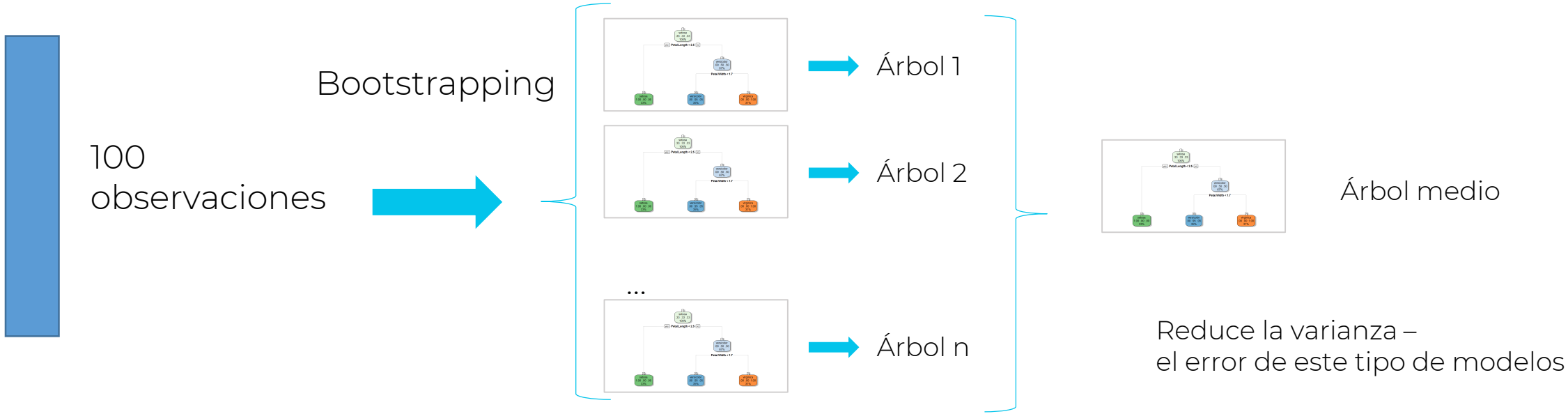
Ensamblando algoritmos - Bagging

- Los random forest o bosques aleatorios es un algoritmo de ensamblaje
- Se basa en el bootstrapping
- Es una técnica muy buena para poder calcular una medida, una característica o un modelo a base de repeticiones



Ensamblando algoritmos - Bagging

- Los random forest o bosques aleatorios es un algoritmo de ensamblaje
- Se basa en el bootstrapping
- Es una técnica muy buena para poder calcular una medida, una característica o un modelo a base de repeticiones



Reduce la varianza –
el error de este tipo de modelos

Ensamblando algoritmos - Boosting

- Se trata de mejorar el modelo de forma iterativa
- Se calcula un modelo. Se ajusta teniendo en cuenta dónde a fallado
- Y así iterativamente hasta conseguir un modelo más óptimo
- Se utiliza para la clasificación binaria en general – Adaboost es el agoritmo por excelencia

Take away

El resumen de la lección

Lo más importante de la lección

- Lo algoritmos de clasificación pueden ser:
  - Lineales
  - NO lineales

- Los algoritmos no lineales son más potentes (menos error) pero tienden al overfitting y a la varianza como el caso de los random forest
- Los algoritmos lineales tienden al tener un error grande pero menor overfitting. Son mucho más fáciles de interpretar que los no lineales

Tú turno

Un pequeño ejercicio para empezar

Tú turno

- Utiliza las hojas de trabajo para practicar con los modelos que acabas de ver en esta lección
- Te ayudará a consolidar los conceptos

- ¡Ya casi estás! Estás a punto de terminar con los algoritmos ☺