Intro Machine Learning

Juantomás García

¿Que es AI?

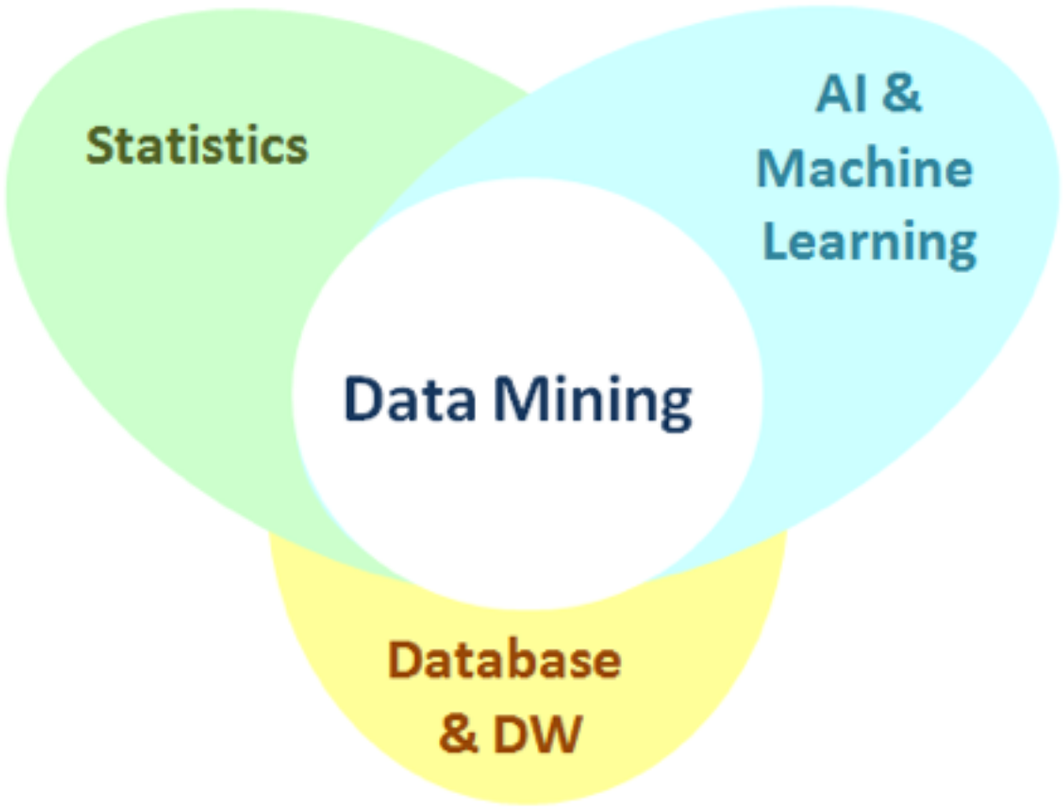
"AI is the study of making computers do things, at which, at the moment, people are better"

-- Elaine Rich (author of Artificial Intelligence book)

Data Mining

- No es fácil definirlo
- Machine Learning es una parte
- Conviene conocer todo el entorno
- Data Mining intenta explicar el pasado y predecir el futuro.

Data Mining



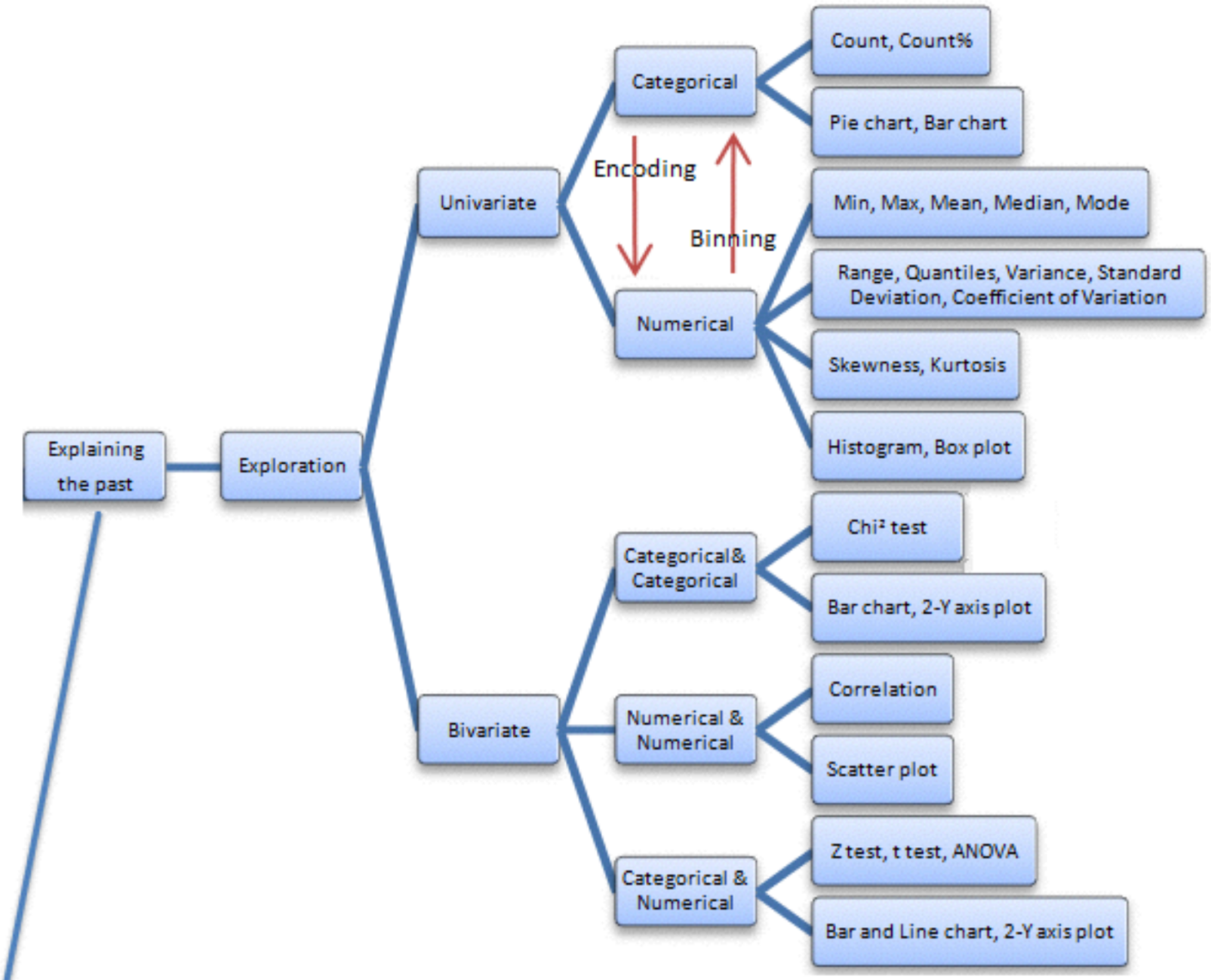
- Estadísticas
- AI & Machine Learning
- Database & Datawarehouse

Data Mining

- Podemos dividirlo en dos subcategorias:
  - Explicando el Pasado
  - Prediciendo el Futuro
  - Incluso una tercera Prescribiendolo ;-)

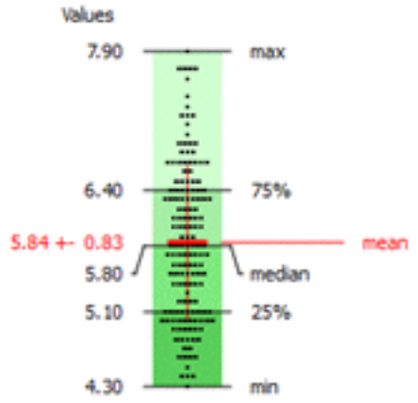
Data Mining

Explicando el Pasado

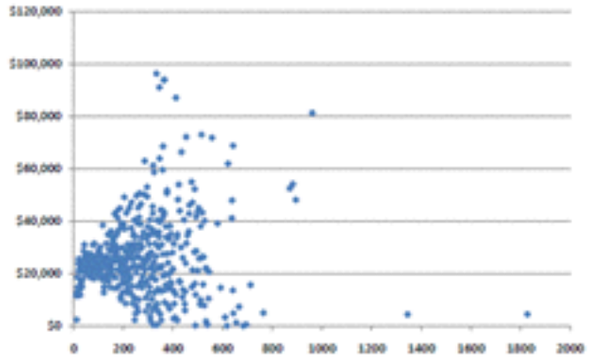


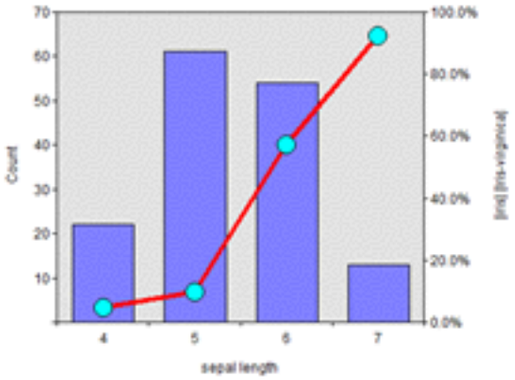
Explorando los datos

- 1 Variable



- 2 Variables





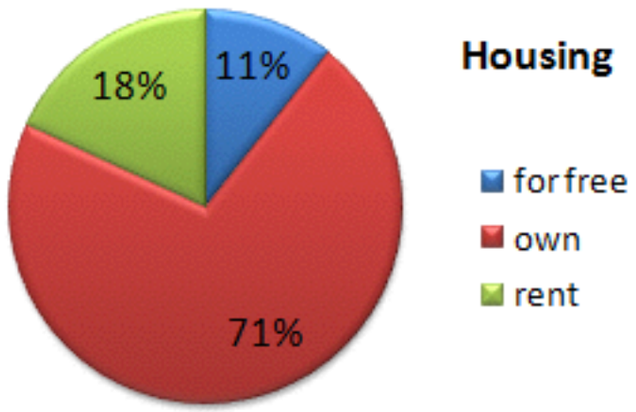
1 Variable

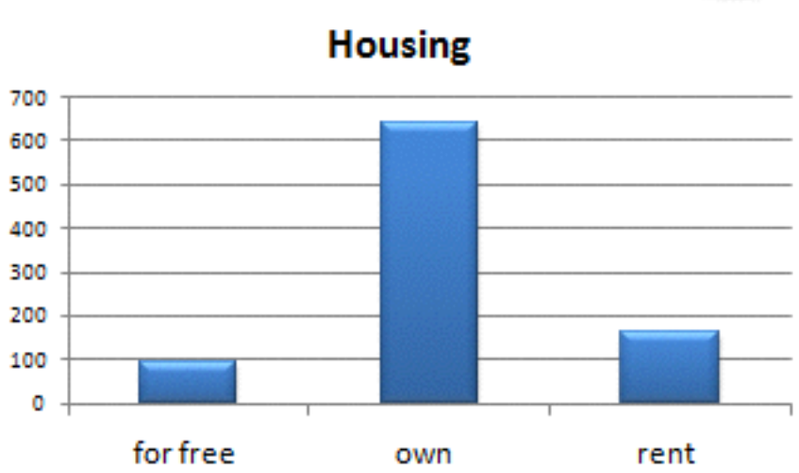
- Exploramos las variables (o atributos) de una en una.
  - Categorias (nominales, ordenadas)
  - Numéricas

Categorias

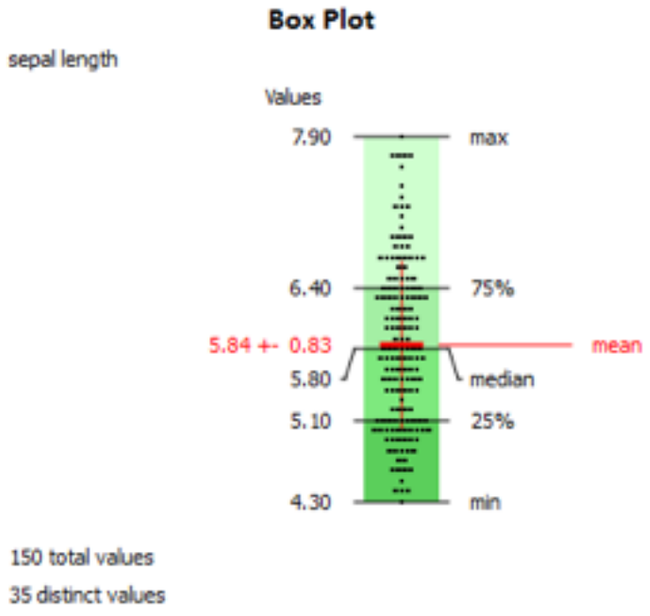
Frequency Table

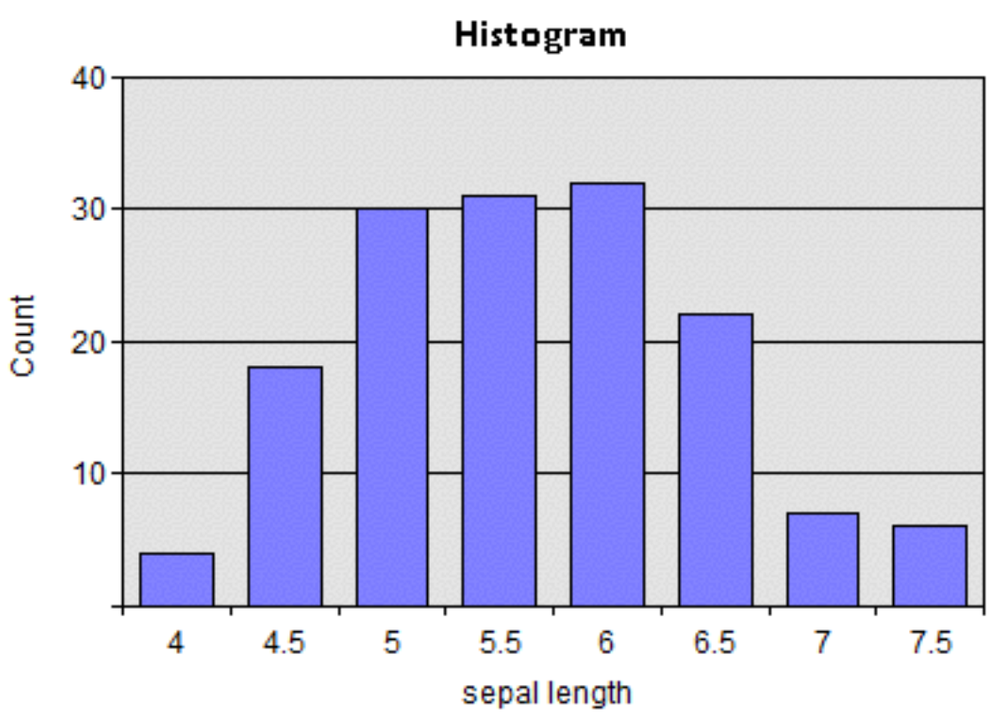
Housing	Count	Count%
for free	96	10.67%
own	641	71.22%
rent	163	18.11%





Numéricas



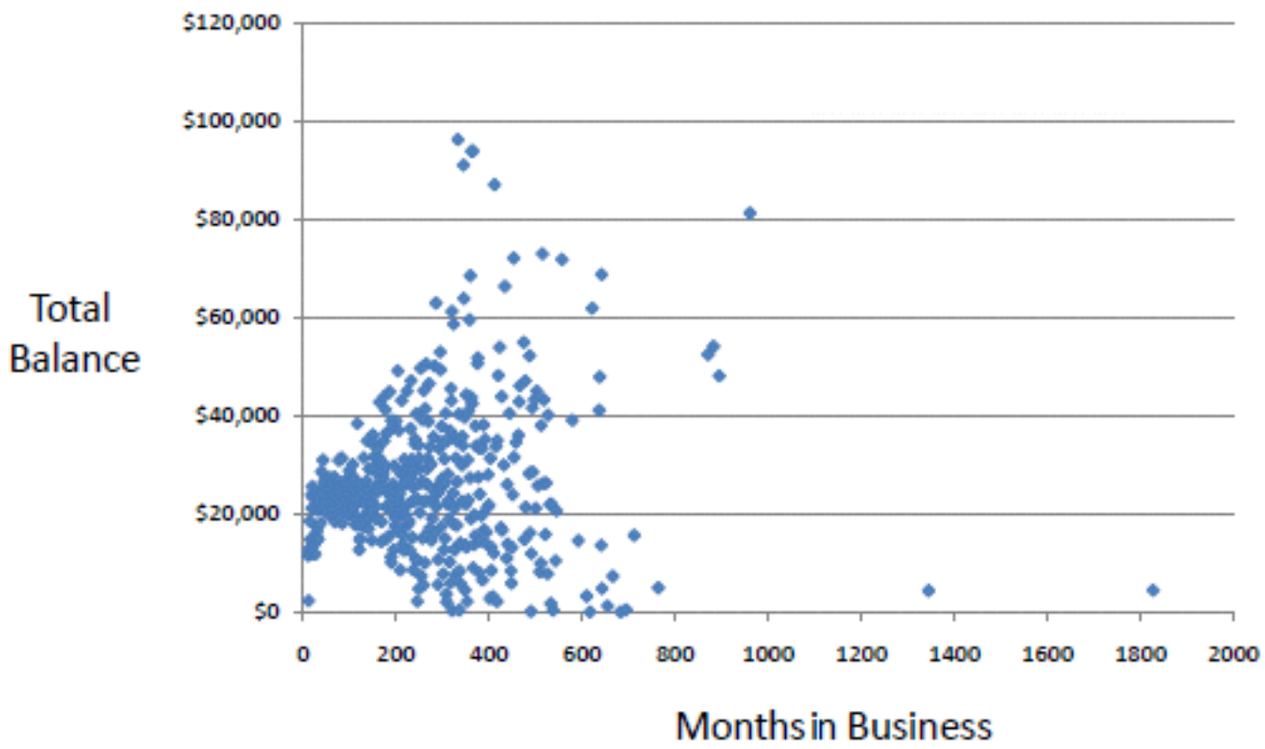


sepal length	
Count	150
Minimum	4.3
Maximum	7.9
Mean	5.84
Median	5.8
Mode	5
Quartile 1	5.1
Range	3.6
Variance	0.69
Standard Deviation	0.83
Coefficient of Variation	14.2%
Skewness	0.31
Kurtosis	-0.55

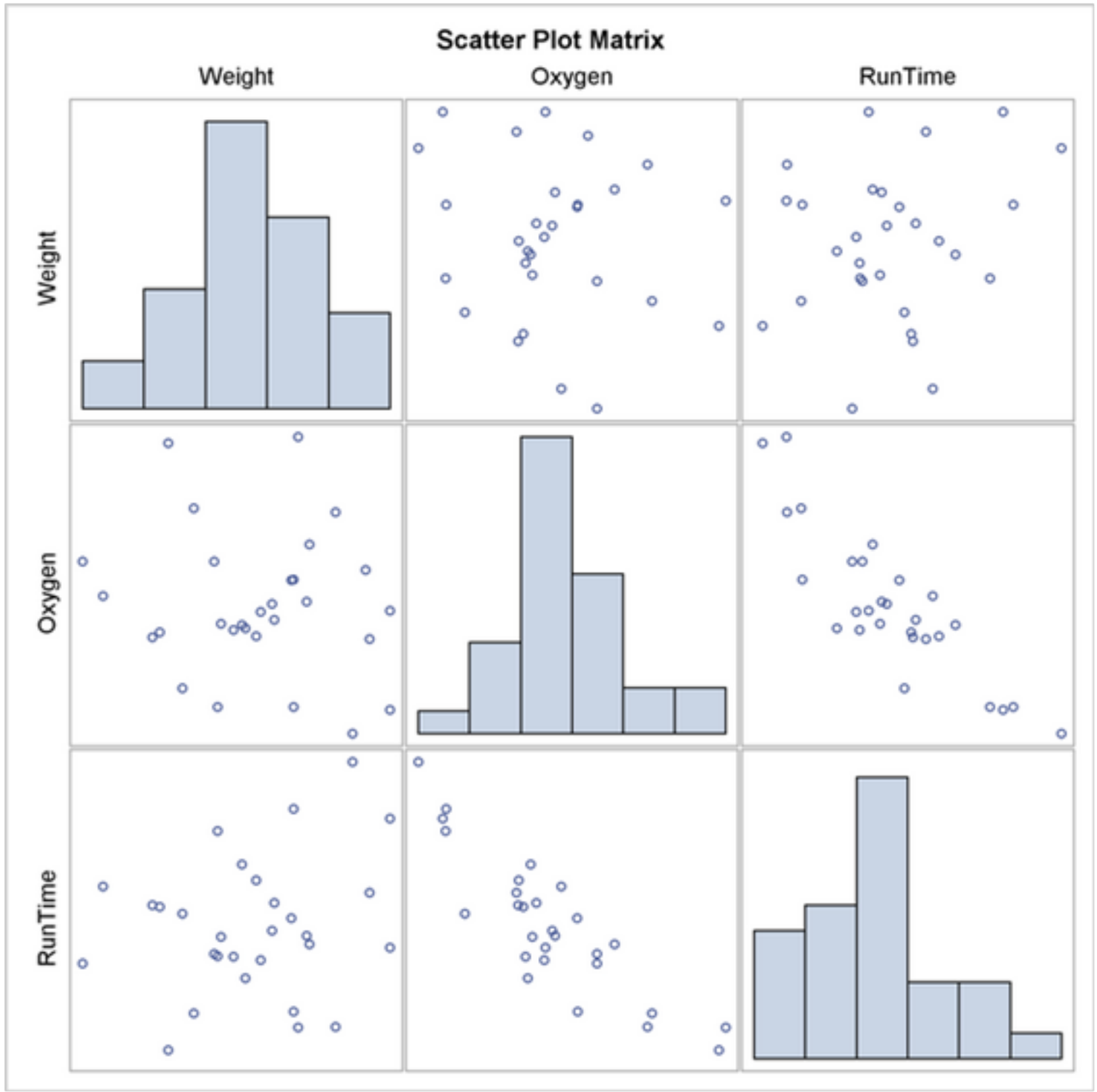
2 Variables

- Númericas vs Númericas
- Categorías vs Categorías
- Números vs Categorías

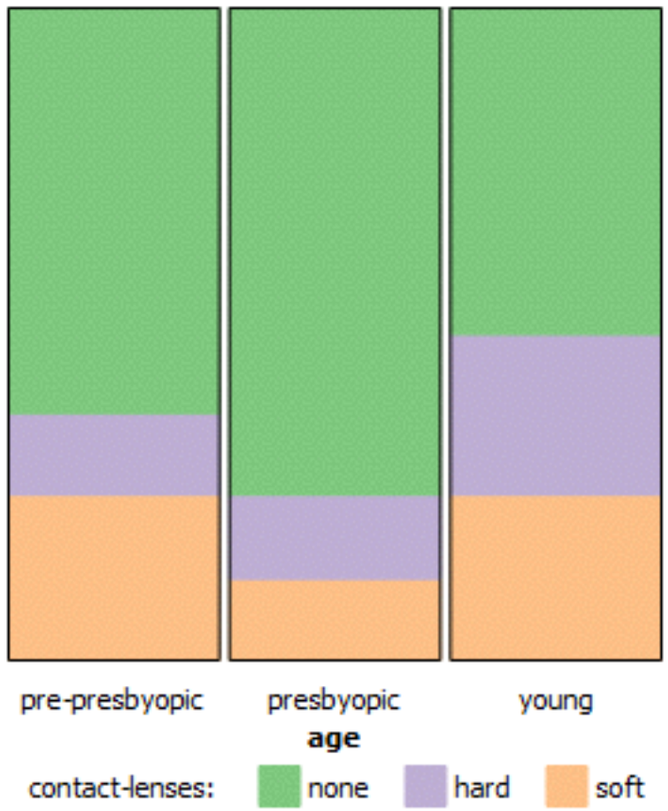
Númericas vs Númericas



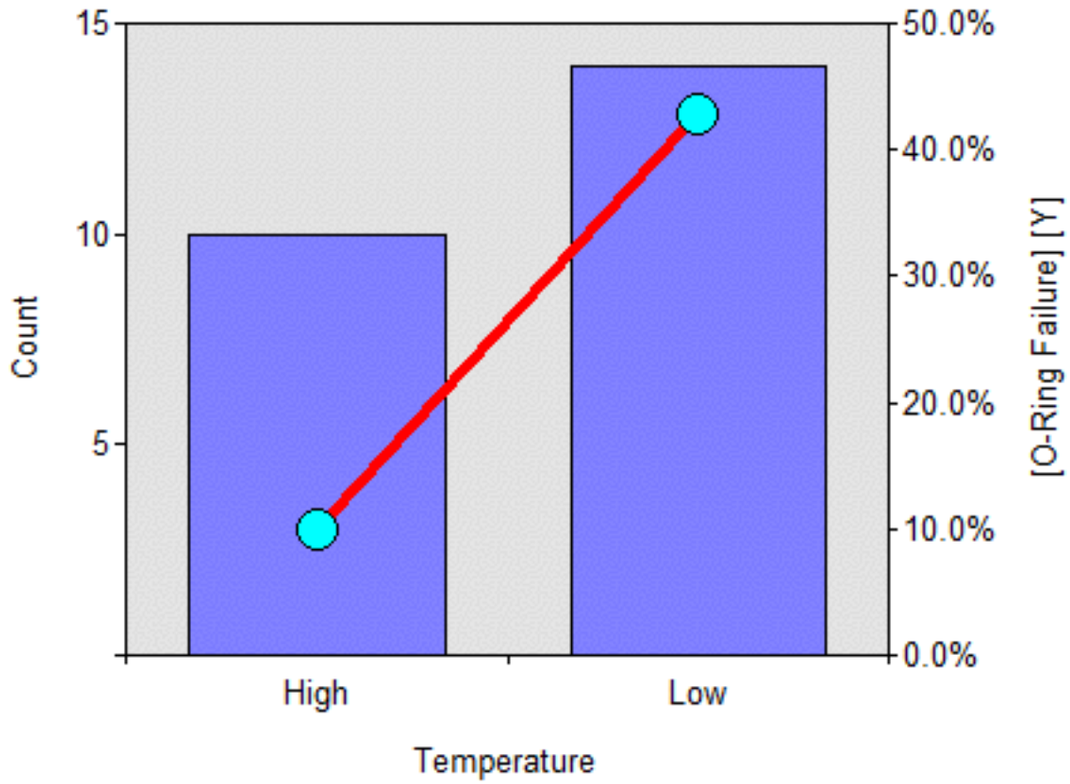
Númericas vs Númericas



C vs C: Stacked Plot



C vs C: Combination Chat



C vs C: Chi Square

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}$$

$$e_{ij} = \frac{n_{i.}n_{.j}}{n}$$

$$df = (r-1)(c-1)$$

Tchouproff Contingency Coefficient

$$\rho_\tau = \sqrt{\frac{\chi^2}{n\sqrt{(c-1)(r-1)}}}$$

e=(44 * 52)/95=24.1

		Hair		
		Light	Dark	
Eye	Black	**32** **(24.1)**	**12** **(19.9)**	44
	Green/Blue	**14** **(19.7)**	**22** **(16.3)**	36
	Others	**6** **(8.2)**	**9** **(6.8)**	15
		52	43	95

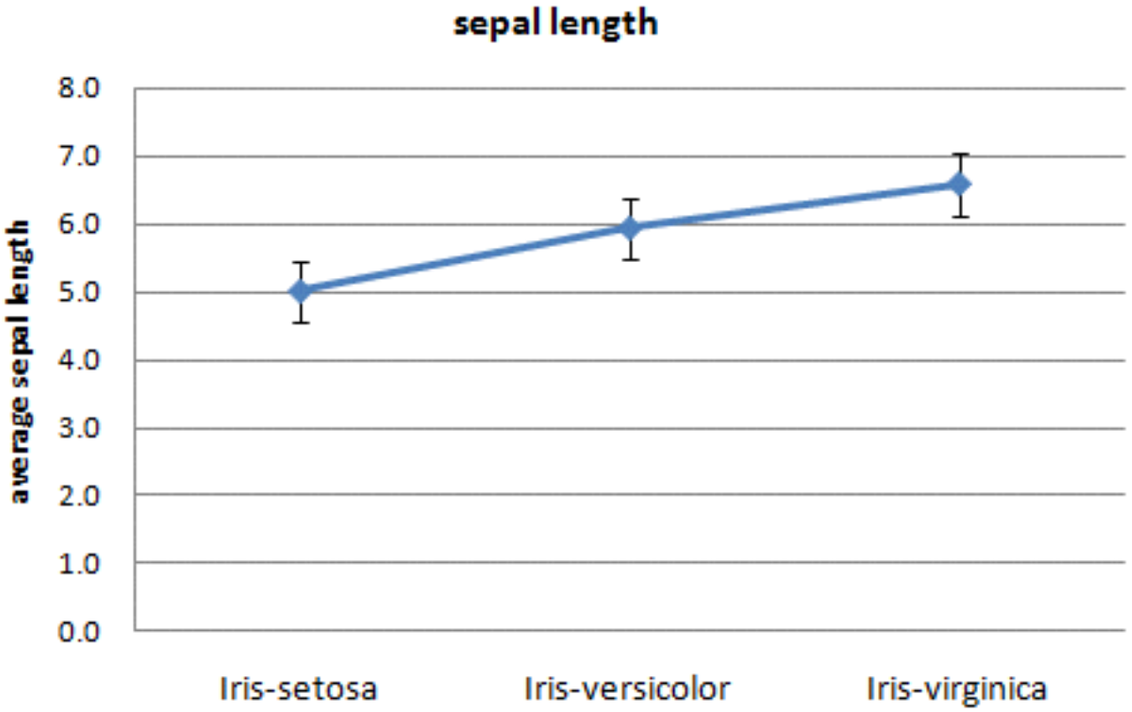
$$\chi^2 = 10.67$$

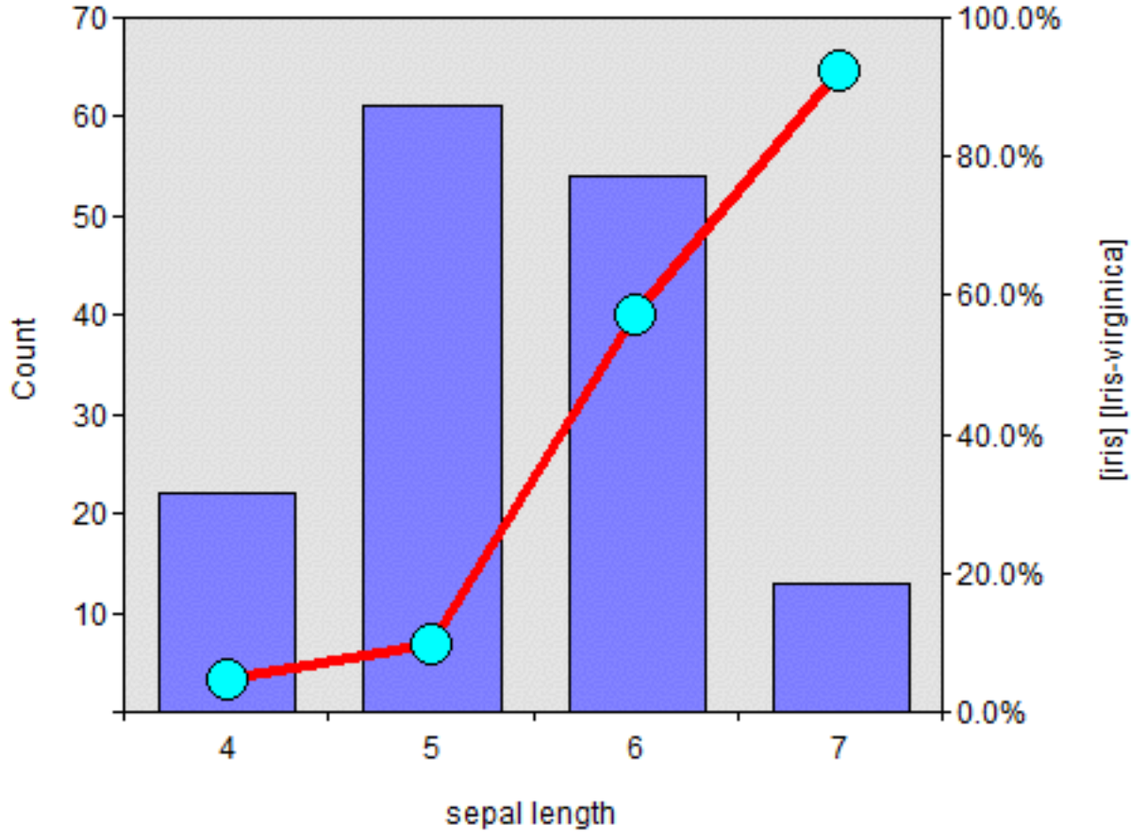
$$df = (r-1)(c-1) = (3-1)(2-1) = 2$$

$$p = 0.005$$

$$\rho_\tau = \sqrt{\frac{10.67}{95\sqrt{(3-1)(2-1)}}} = 0.28$$

Categorías vs Numéricas





Categorías vs Numéricas

- Y además no podían faltar:
- El Z-test y t-test
- Análisis de Varianza (ANOVA)

Prediciendo el Futuro

Prediciendo el Futuro

- Generamos Modelos
  - Clasificación: resultado categoria
  - Regresión: resultado numérico
  - Clustering
  - Reglas Asociativas

Prediciendo el Futuro

- La Clasificaciones las podemos dividir en grupos, basados:
  - Tabla de Frecuencias (ZeroR, OneR, Naive Bayesian, Decision Tree
  - Matriz de Covariancia (A.Linear Discriminante, Regresión Logistica

Prediciendo el Futuro

- La Clasificaciones las podemos dividir en grupos, basados en:
  - Funciones de similitud: KNN (K vecinos más cercanos)
  - Otros: ANN (Redes Neuronales Artificiales), SVM (Support Vector Machine)

Zero R

- La más simple.
- De hecho ignora las variables

Predictors				Target
Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

OneR

- Se generan los modelos para cada variable
- Se escoge el que menos error tiene

Which one is the best predictor ?

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Frequency Tables

★		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

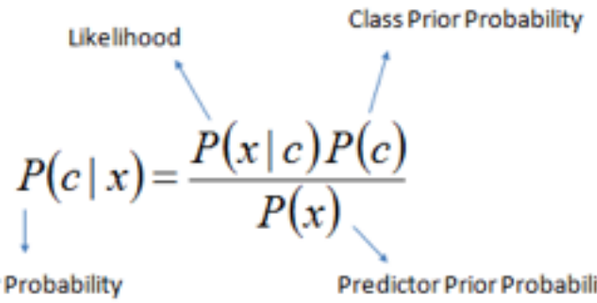
		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3

Naive Bayesian

- Basado en el teorema de Bayes
- Mucho más sencillo de construir



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Frequency Table		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

$P(x \mid c) = P(Sunny \mid Yes) = 3/9 = 0.33$

Likelihood Table		Play Golf		
		Yes	No	
Outlook	Sunny	3/9	2/5	5/14
	Overcast	4/9	0/5	4/14
	Rainy	2/9	3/5	5/14
		9/14	5/14	

$P(x) = P(Sunny) = 5/14 = 0.36$

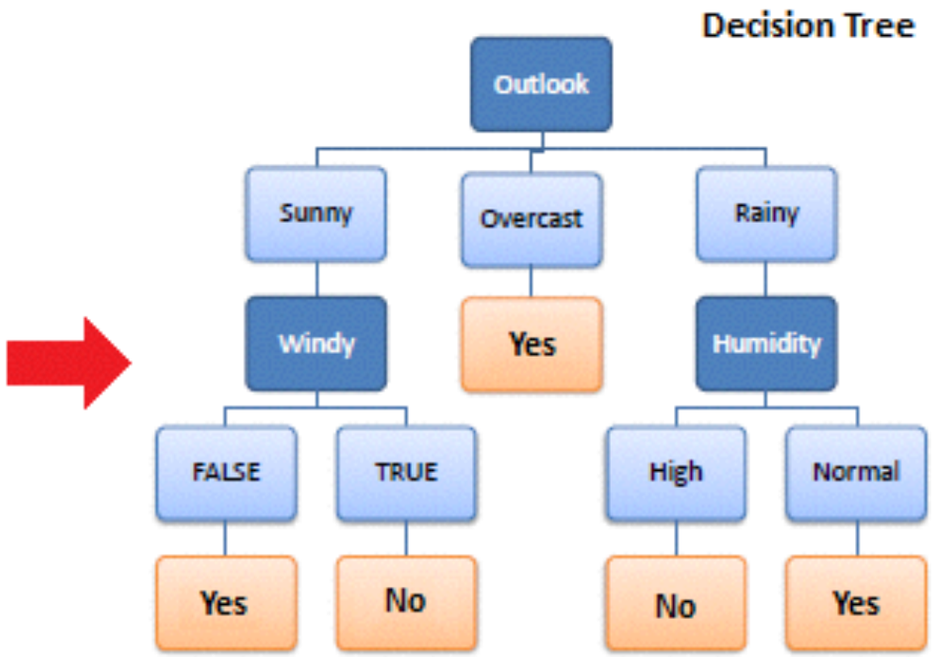
$P(c) = P(Yes) = 9/14 = 0.64$

Posterior Probability: $P(c \mid x) = P(Yes \mid Sunny) = 0.33 \times 0.64 \div 0.36 = 0.60$

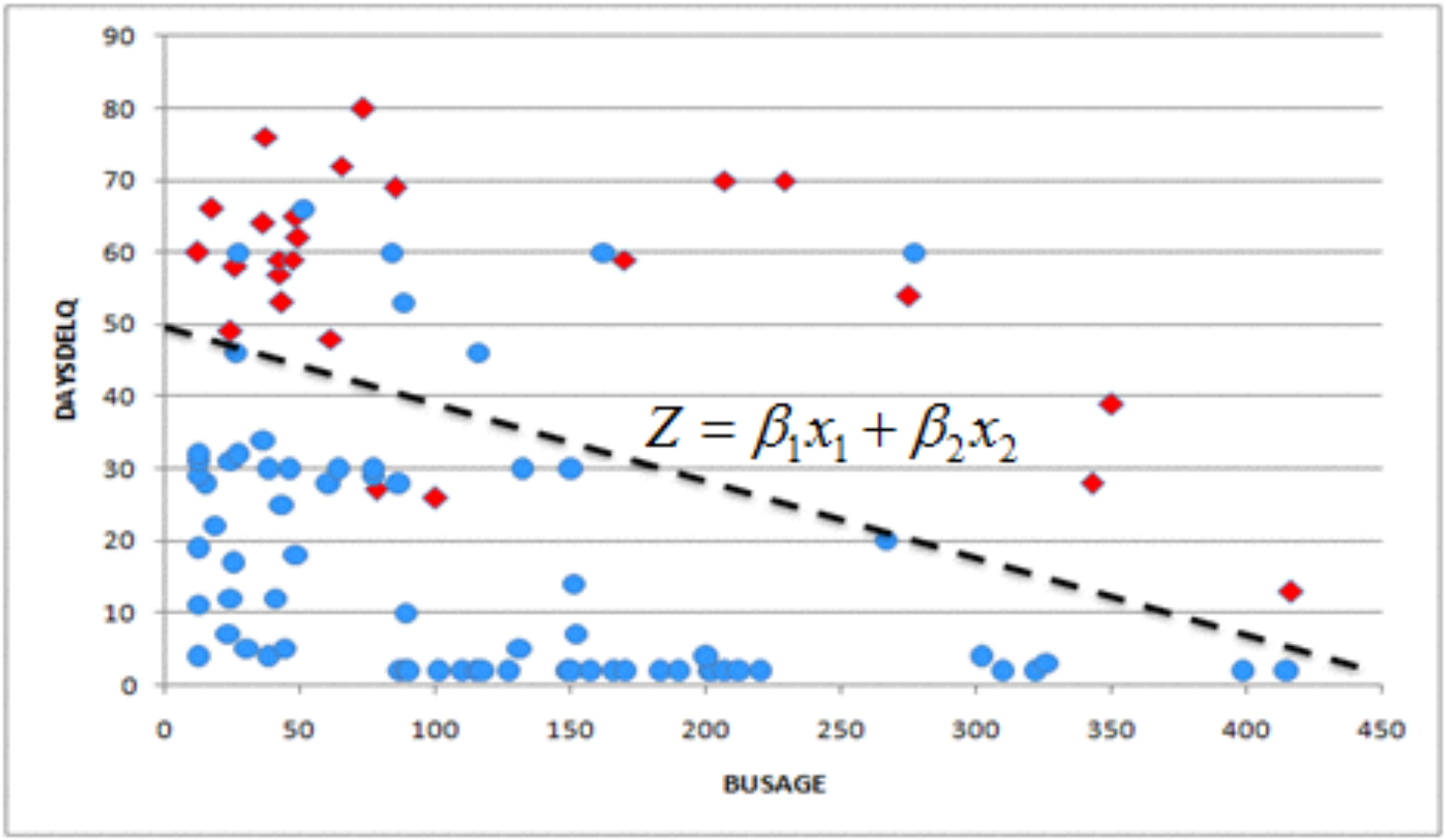
Decision Tree

- El árbol de decisiones se monta basado en la entropía

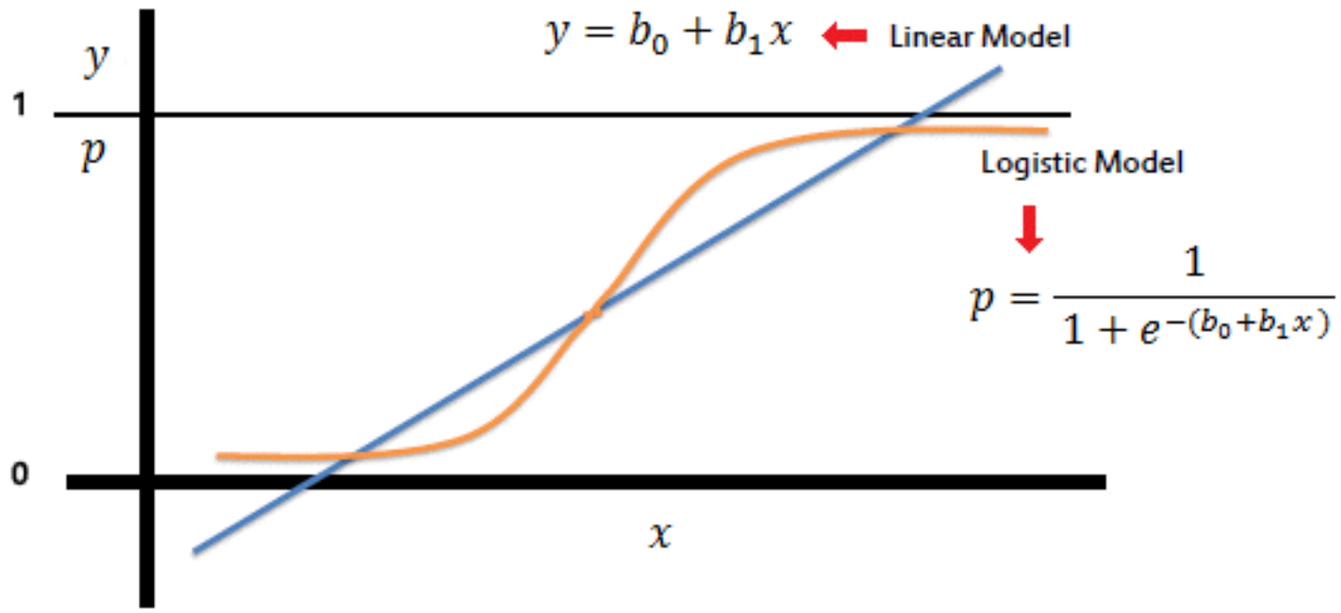
Predictors				Target
Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No



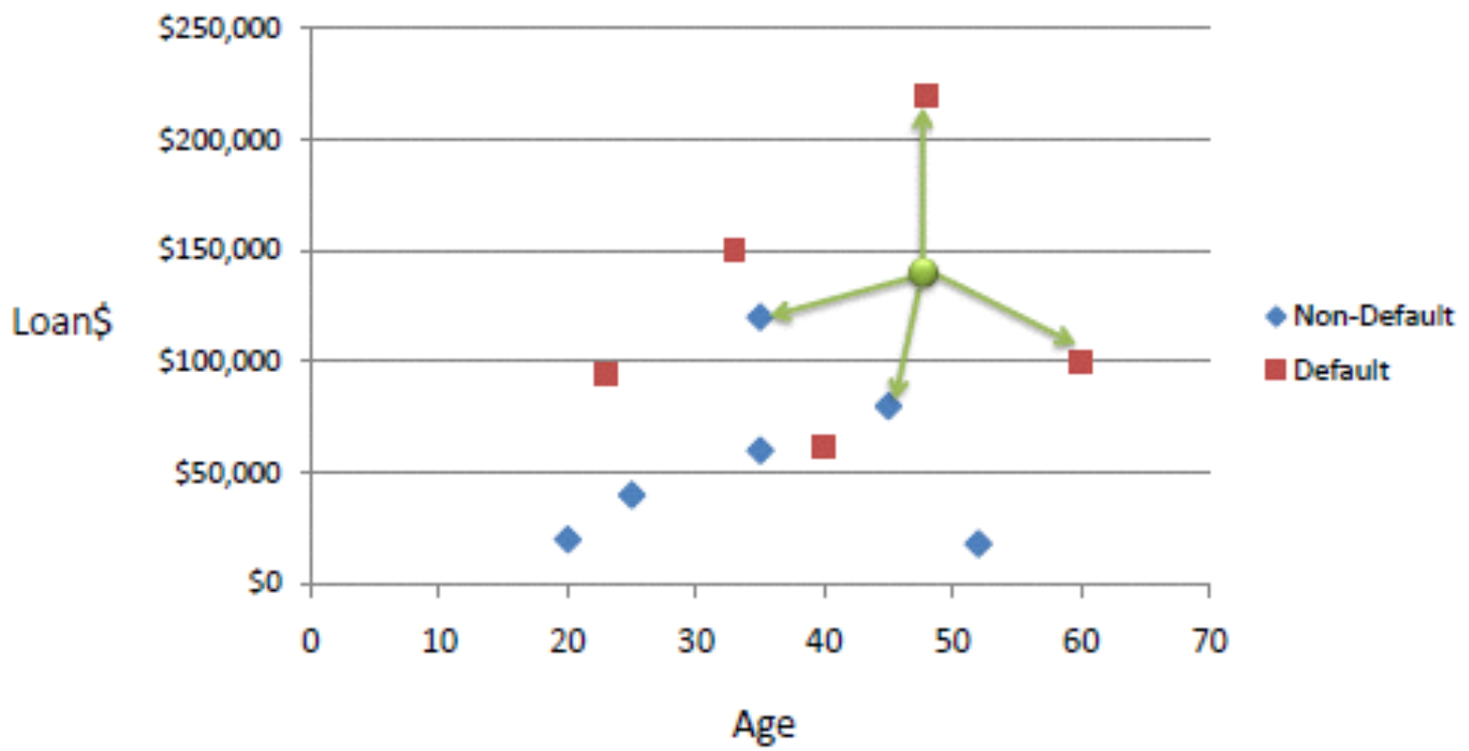
Analisis Discrim. Linear



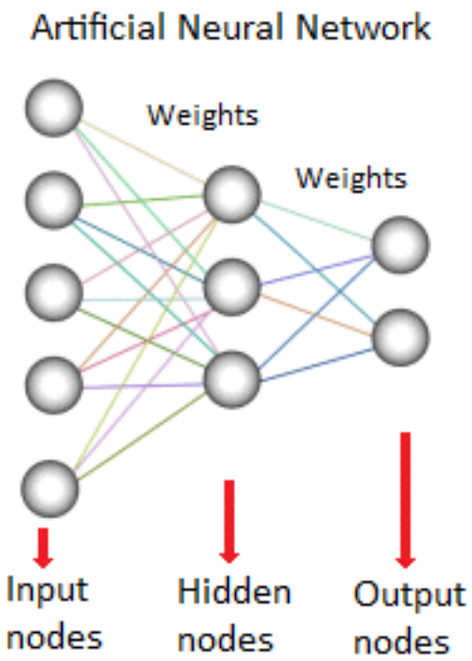
Regresión Logistica

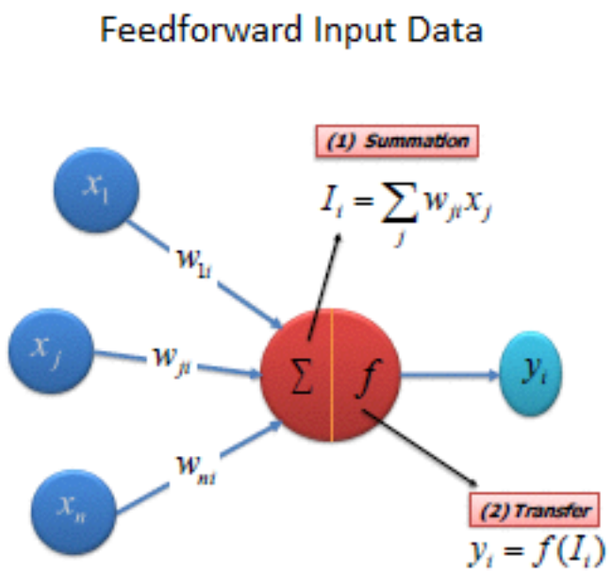


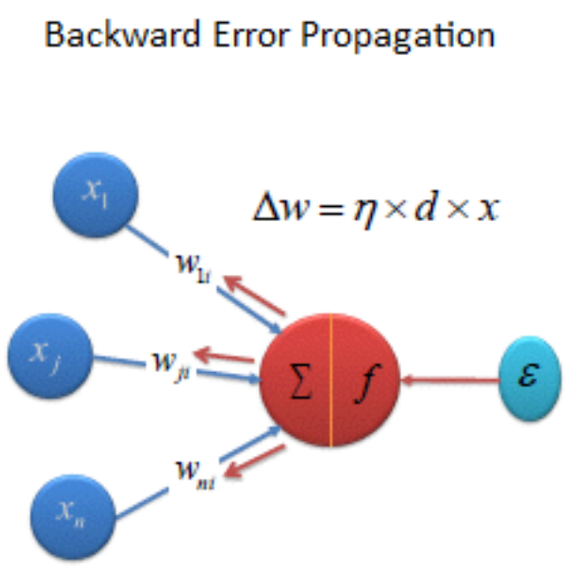
KNN K Nearest Neighbors



Artificial Neural Networks



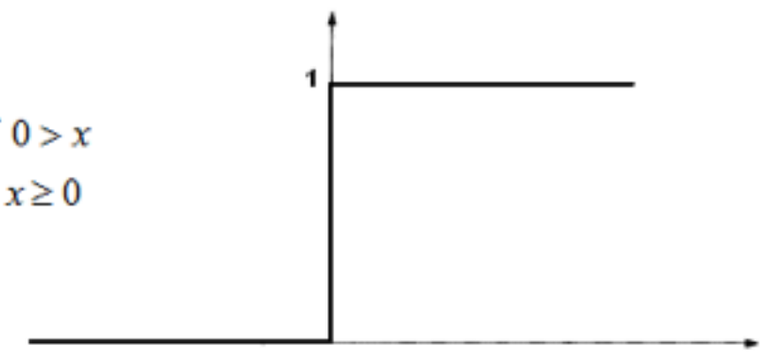




Artificial Neural Networks

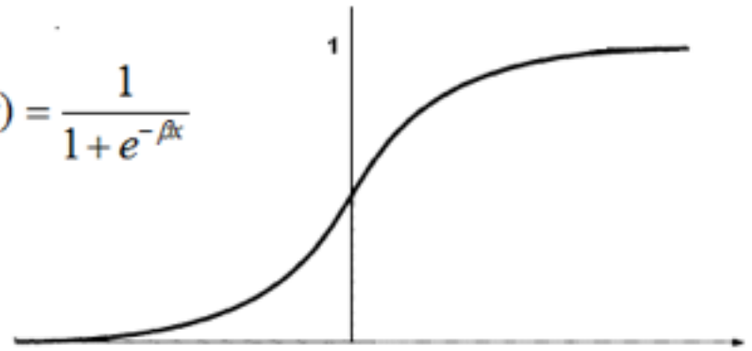
Unit step (threshold)

$$f(x) = \begin{cases} 0 \text{ if } 0 > x \\ 1 \text{ if } x \geq 0 \end{cases}$$



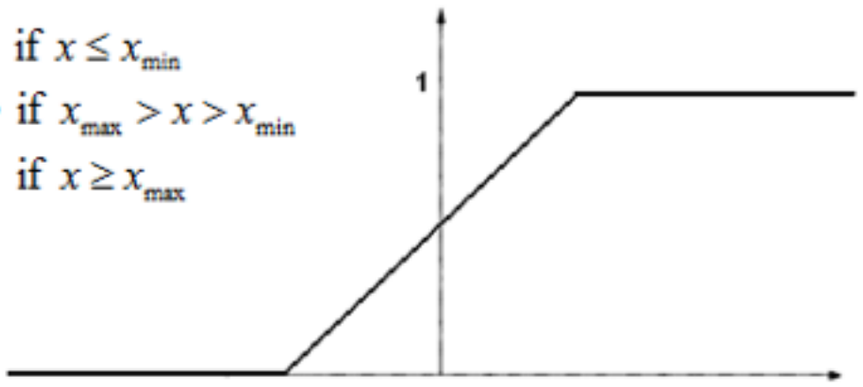
Sigmoid

$$f(x) = \frac{1}{1 + e^{-\beta x}}$$



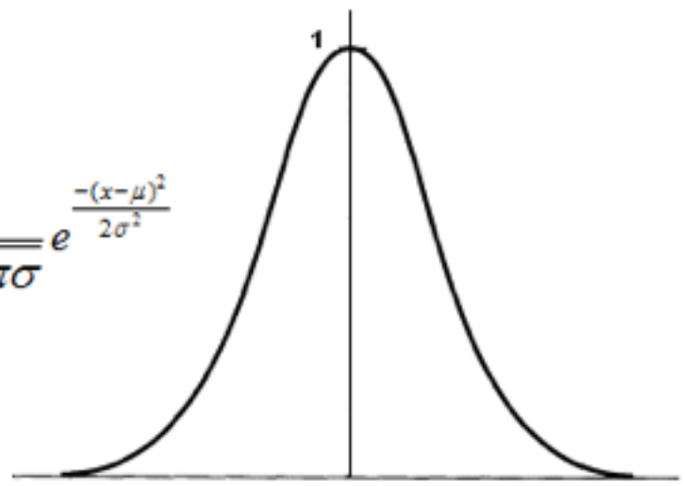
Piecewise Linear

$$f(x) = \begin{cases} 0 & \text{if } x \leq x_{\min} \\ mx + b & \text{if } x_{\max} > x > x_{\min} \\ 1 & \text{if } x \geq x_{\max} \end{cases}$$

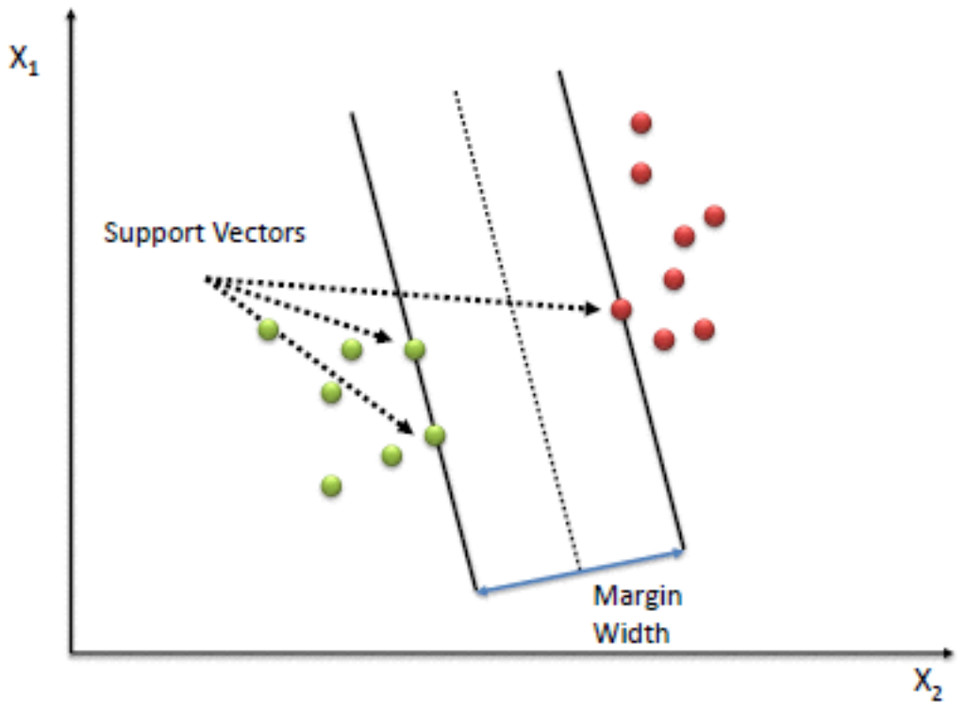


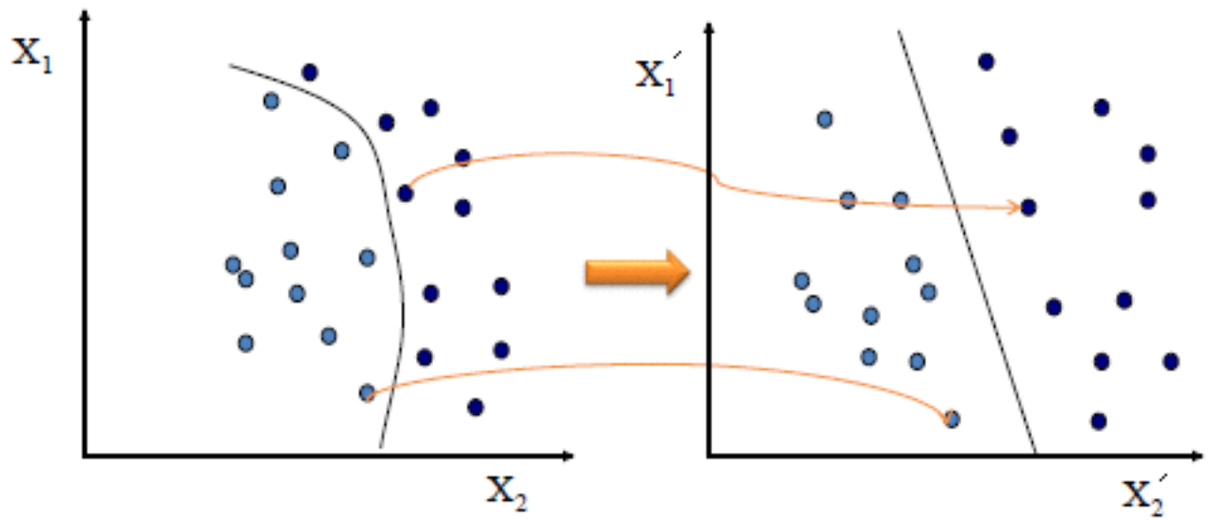
Gaussian

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$



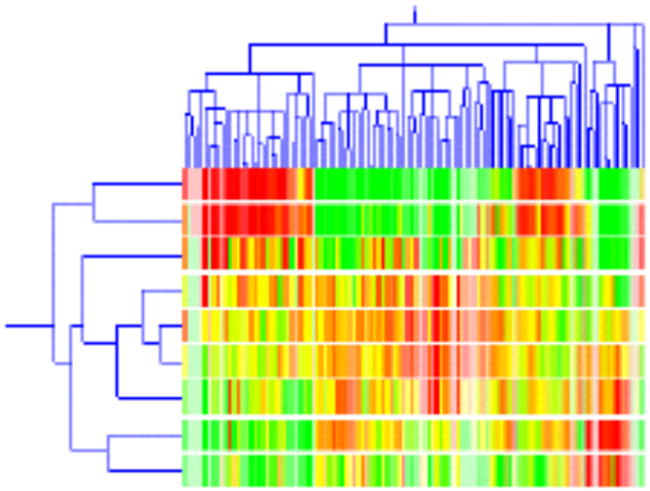
Support Vector Machine



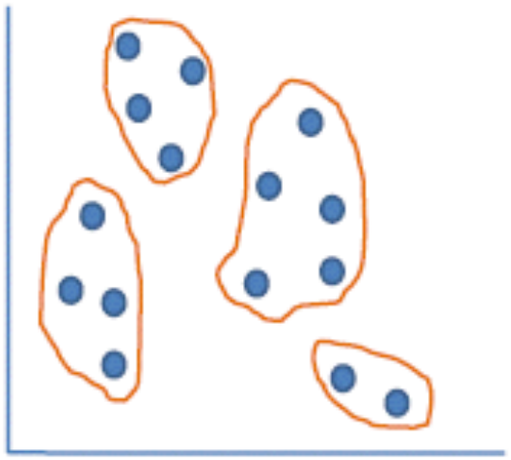


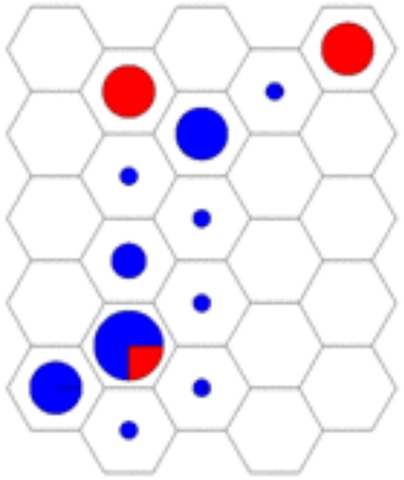
Clustering

Hierarchical Clustering

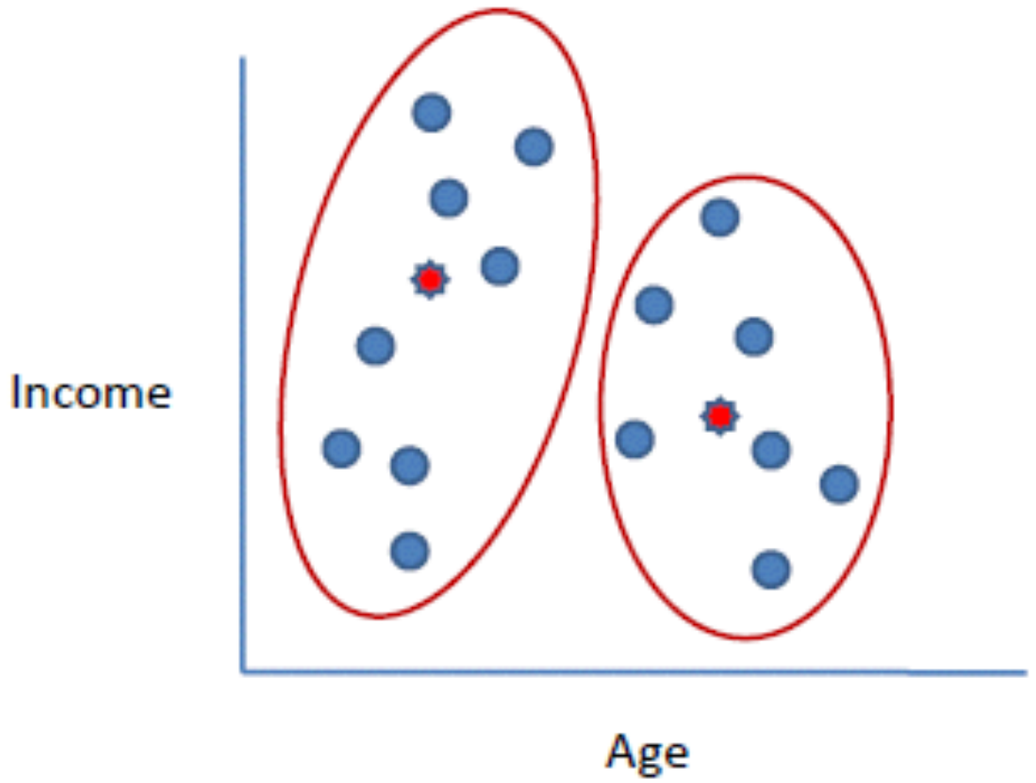


Partitive Clustering

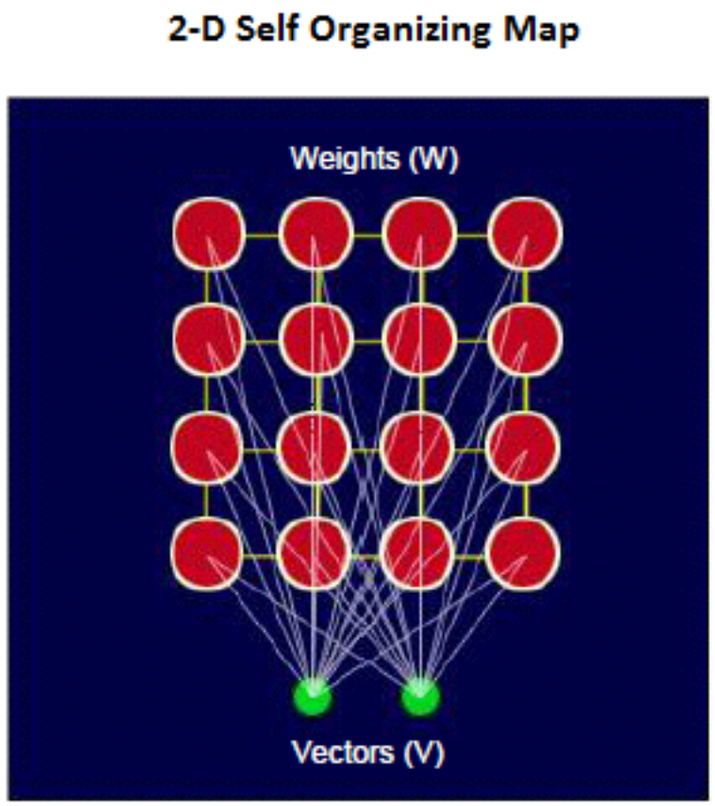




K-Means



SOM Mapas Autoorganiza.



Every node is connected to the input the same way, and no nodes are connected to each other.

Algunos Casos de Éxito

- Hay muchos ;-)
- Un par de mis Favoritos
  - TARGET
  - Envios

n

- Curso Machine Learning de Andrew Ng en coursera
- Smart Data: Foreman, John
- kaggle.com

- Apuntarse al meetup del machine learning spain
- La siguiente charla

Muchas Gracias

Juantomás García

juantomas@aspgems.com

@juantomas