VNIVERSITAT DE VALÈNCIA

Machine Learning aplicado a Negocio

Otto F. Wagner
www.ottofwagner.com

Colegio de Economistas de Madrid
Colegiado Nº 26595

economistas Consejo General
REDI - Nº 098

SPAIN AI

AMKT Asociación de Marketing de España

OF.W

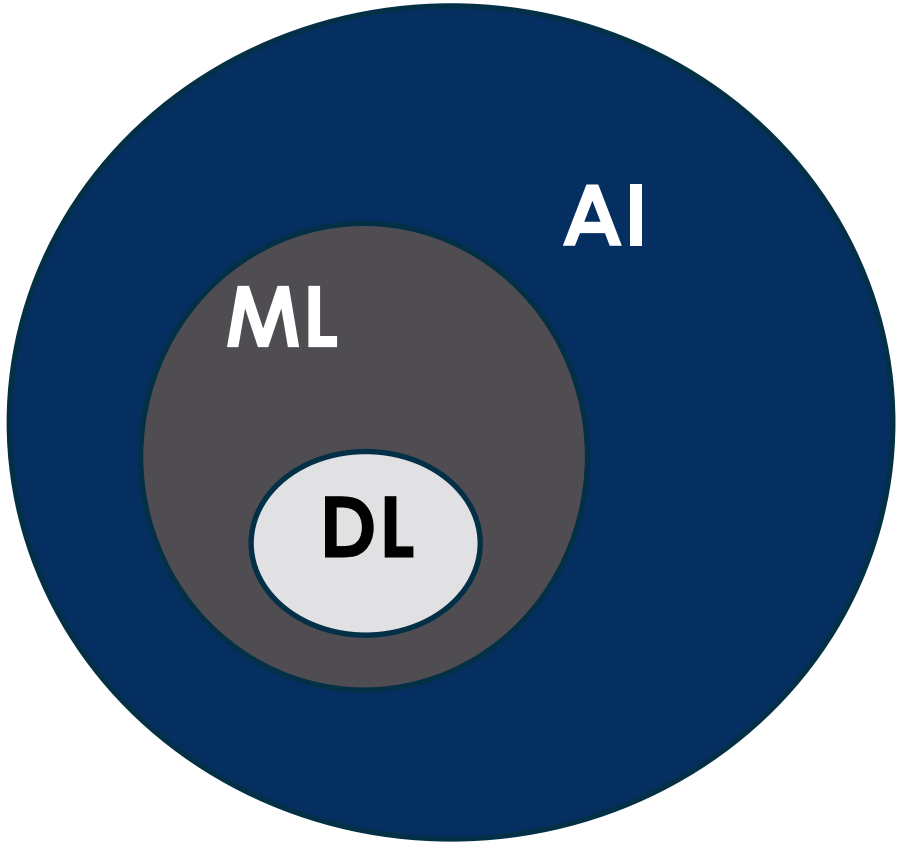
AGENDA

1. **Sesión 1**
    1. Introducción
    2. Modelos No Supervisados
2. **Sesión 2**
    1. Modelos Supervisados
    2. Series Temporales
3. **Sesión 3**
    1. Redes Neuronales
    2. Análisis del Sentimiento
    3. Más Allá del Deep Learning

MODELOS DE MACHINE LEARNING

INTRODUCCIÓN

DE LA AI AL DL



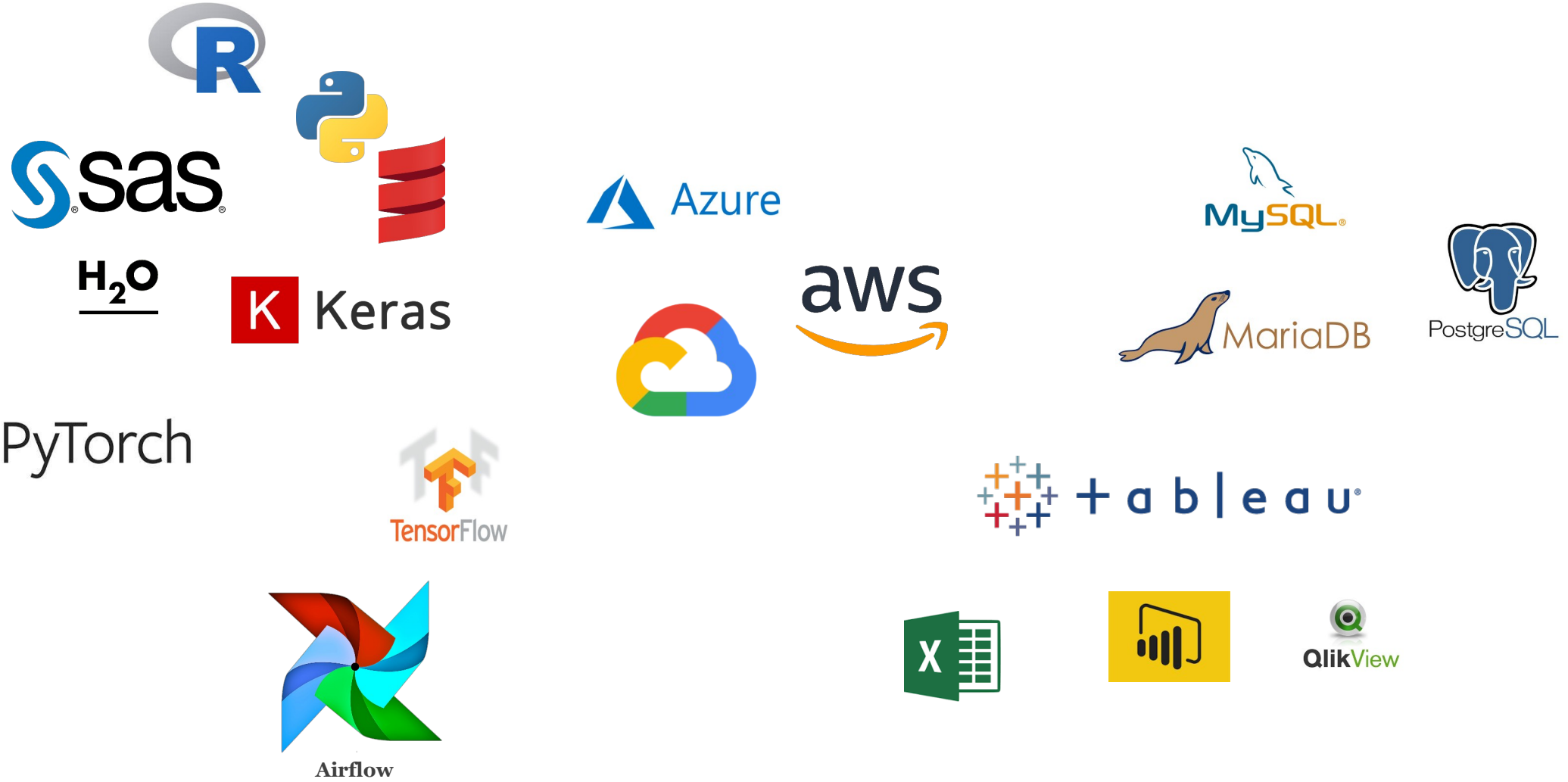
- Inteligencia Artificial (AI):
  - Posibilita que las máquina aprendan de la experiencia y realicen tareas como hacen los humanos
- Aprendizaje Automático (ML):
  - Permiten a las máquinas aprender. Generalizan comportamientos e inferencias para un conjunto más amplio (potencialmente infinito) de datos
- Aprendizaje Profundo (DL):
  - Usa las redes neuronales para obtener patrones, conocimiento abstracto y detección de estructuras complejas

SOFTWARE PARA EL APRENDIZAJE AUTOMÁTICO

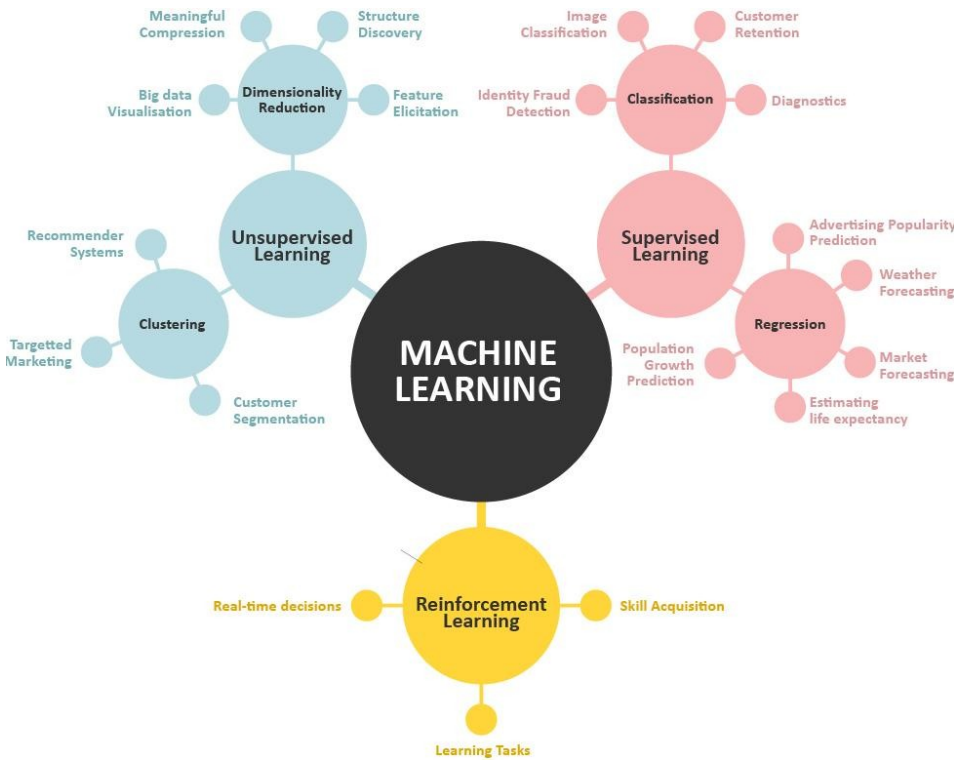
- Softwares estadísticos
  - Privado
    - SAS
    - SPSS
    - Stata
    - Eviews
  - Libre
    - R
    - Knime
    - Python (numpy, scipy)
    - PSPP
    - Orange (hecho en Python)
    - QGIS
    - Gephi
    - …

- Bases de Datos
  - SQL
    - MySQL
    - PostgreSQL
    - SQLite
  - Oracle SQL
  - Server SQL
  - …
  - Not Only SQL
    - MongoDB
    - Cassandra
    - Neo4j
    - …

SOFTWARE PARA EL DATA SCIENCE



http://dfkoz.com/ai-data-landscape/

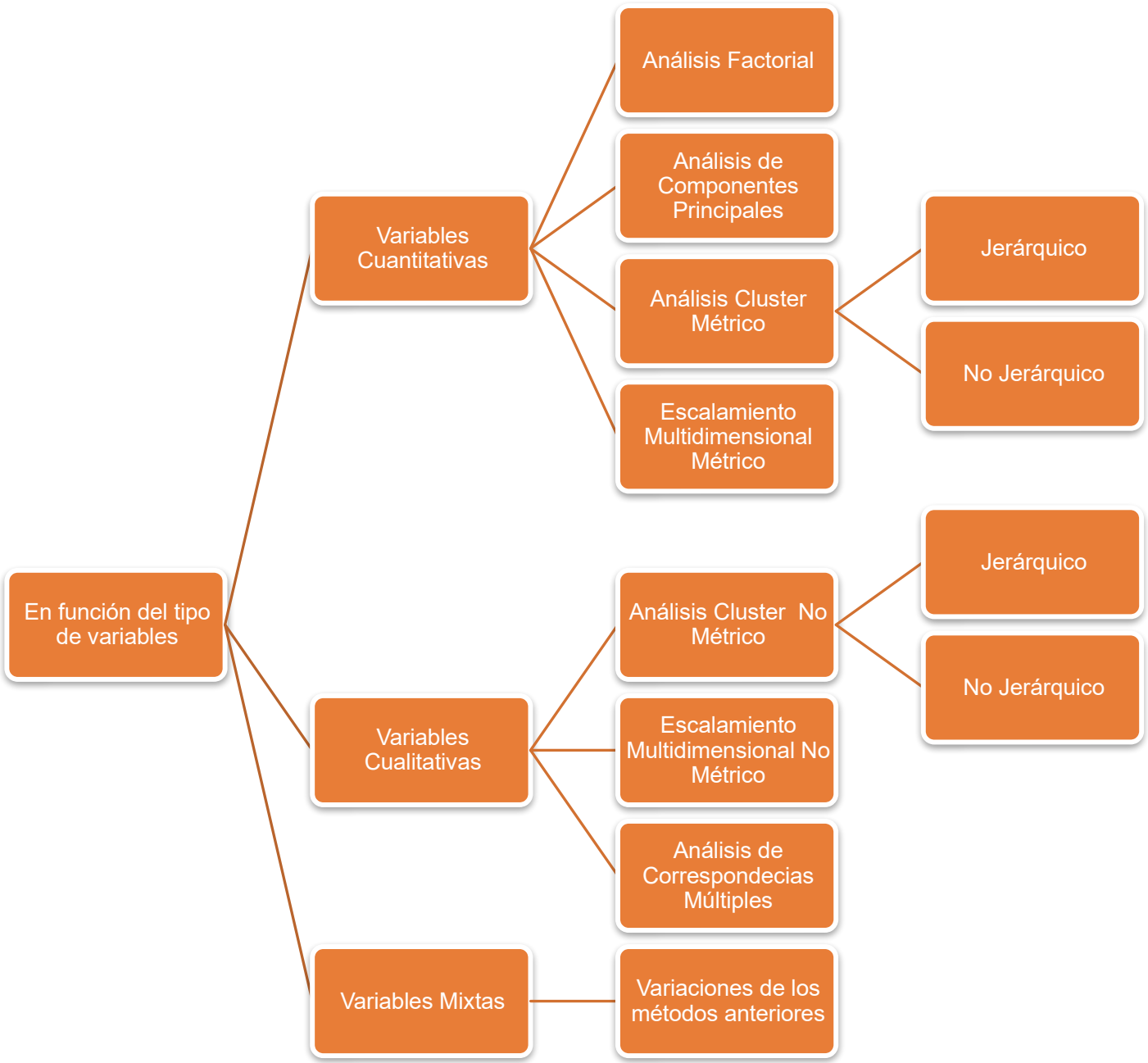


TIPOS DE APRENDIZAJES

- **Supervisados o de dependencia**
  - Son métodos de carácter explicativo
  - Se diferencia entre variables explicativas, independientes o predictivas y variables a explicar o dependientes
- **No supervisados o de interdependencia**
  - No hay distinción entre variables dependientes o independientes
  - Son métodos totalmente descriptivos: Estructura de los datos, síntesis de los datos y establecer clasificaciones.
- **Semi-supervisados**
  - Pseudolabelados
  - P. Ej: rentrenar: el modelo con predicciones con bajo error
- **Por refuerzo**
  - Determinar qué acciones debe escoger un agente de software en un entorno dado con el fin de maximizar alguna noción de "recompensa" o premio acumulado
  - Es un mecanismo de “prueba y error”

MODELOS NO SUPERVISADOS

MÉTODOS NO SUPERVISADOS



ANÁLISIS DE COMPONENTES PRINCIPALES

- Consiste en reducir nuestras variables $X_1,\ldots,X_n$ a un número inferior de variables (llamadas componentes principales) $Z_1,\ldots,Z_m$

$$Z_1=a_{1,1^*}X_1+\ldots+ a_{i,1^*}X_i+\ldots+a_{n,1^*}X_n$$

$$Z_2=a_{1,2^*}X_1+\ldots+ a_{i,2^*}X_i+\ldots+a_{n,2^*}X_n$$

$$\ldots$$

$$Z_1=a_{1,m^*}X_1+\ldots+ a_{i,m^*}X_i+\ldots+a_{n,m^*}X_n$$

Donde:

$Z_1,\ldots,Z_m$ son ortogonales

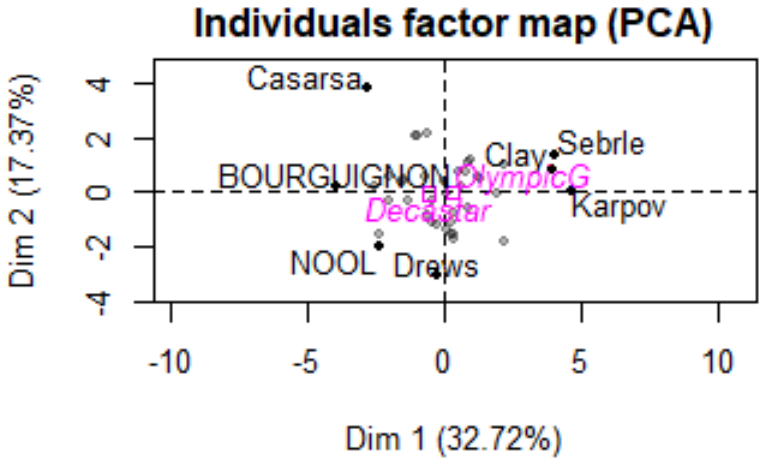
$m<=n$

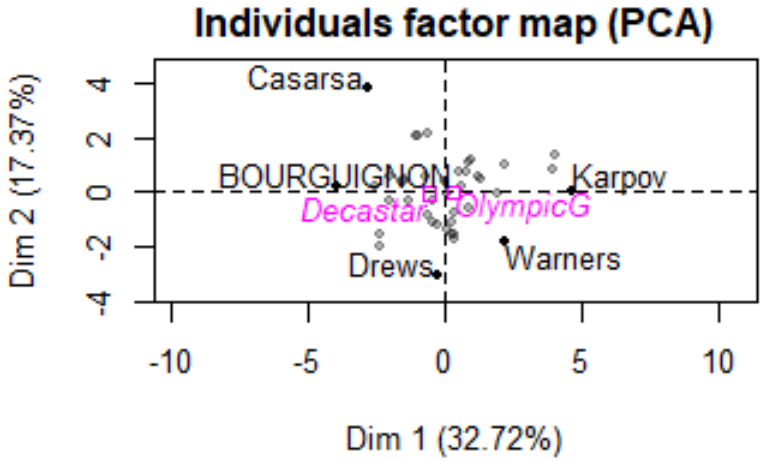
Nos ayuda a simplificar los problemas, tanto de métodos supervisados como de no supervisados.

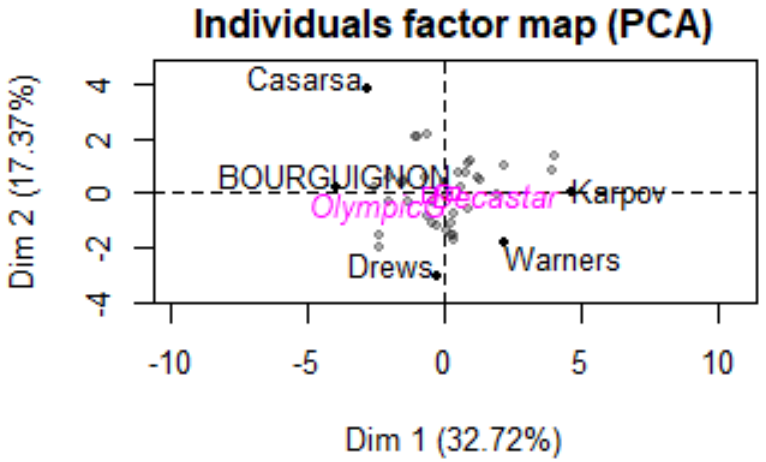
ANÁLISIS DE COMPONENTES PRINCIPALES

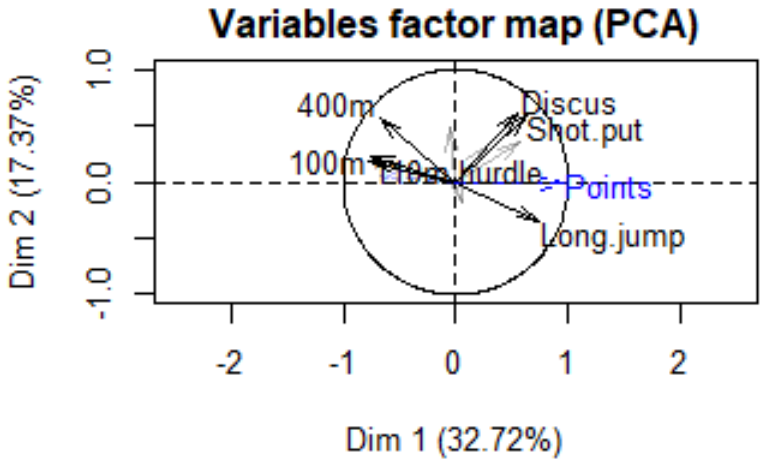
	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
100m	18.34376957	2.016090	2.42049891	0.13532858	13.336184
Long.jump	16.82246707	6.868559	2.36319121	0.98030118	0.196456
Shot.put	11.84353954	20.606785	0.03890276	3.43711486	1.804174
High.jump	9.99788710	7.063694	4.79362526	1.73967752	45.053306
400m	14.11622887	18.666374	1.23027094	0.08124195	1.122971
110m.hurdle	17.02011495	3.013382	0.61083225	8.00327927	3.943110
Discus	9.32848615	21.162245	0.13131711	6.38020830	1.604724
Pole.vault	0.07745541	1.872547	34.06090024	28.78266727	15.899147
Javeline	2.34696326	5.784369	10.80714169	48.00480246	13.596270
1500m	0.10308808	12.945954	43.54331962	2.45537861	3.443657

ANÁLISIS DE COMPONENTES PRINCIPALES





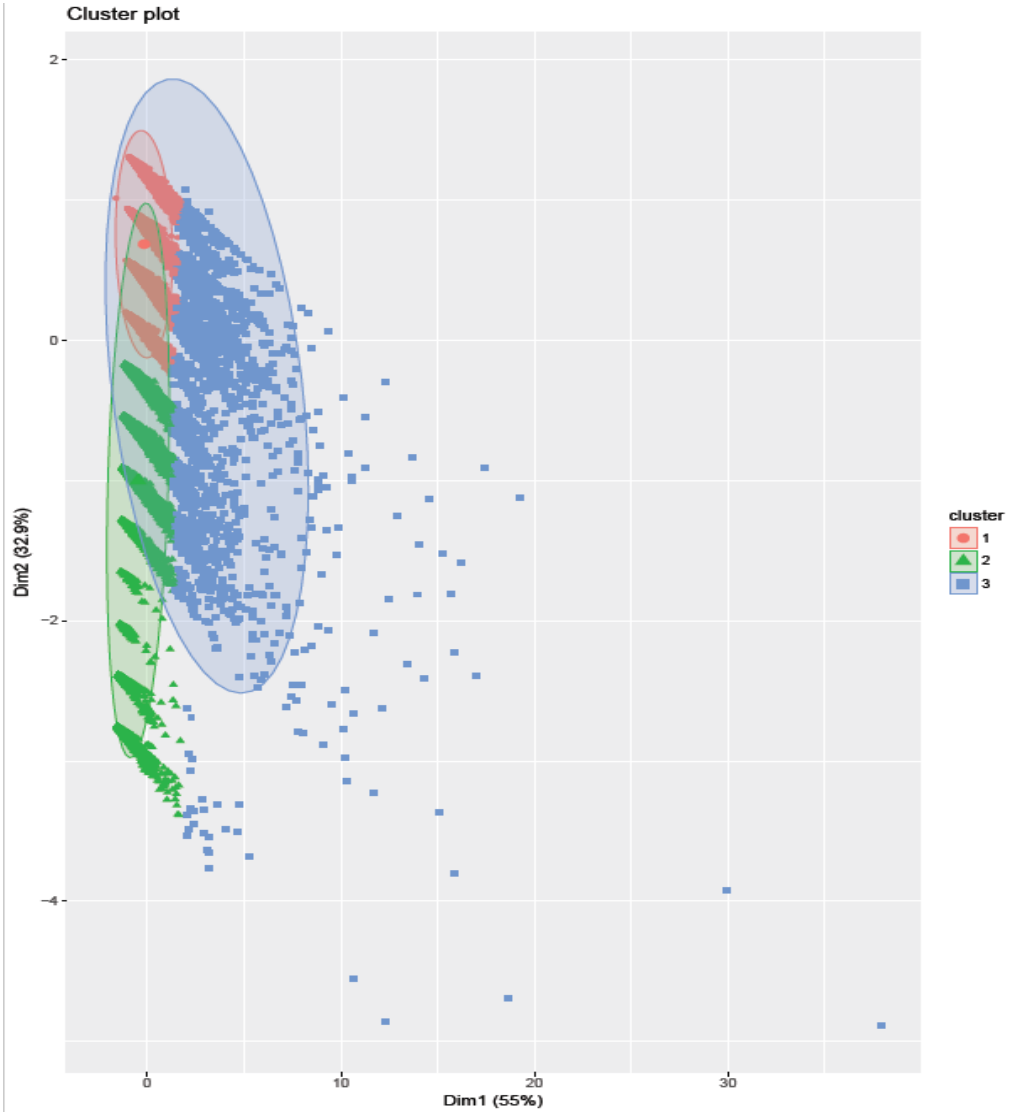




CLUSTERS: FUNDAMENTOS

- Dependiendo si el modelo tiene variables continuas, categóricas o ambas

- Métodos de clustering:
  - K-Medias
  - K-Modas
  - Jerárquico
  - …



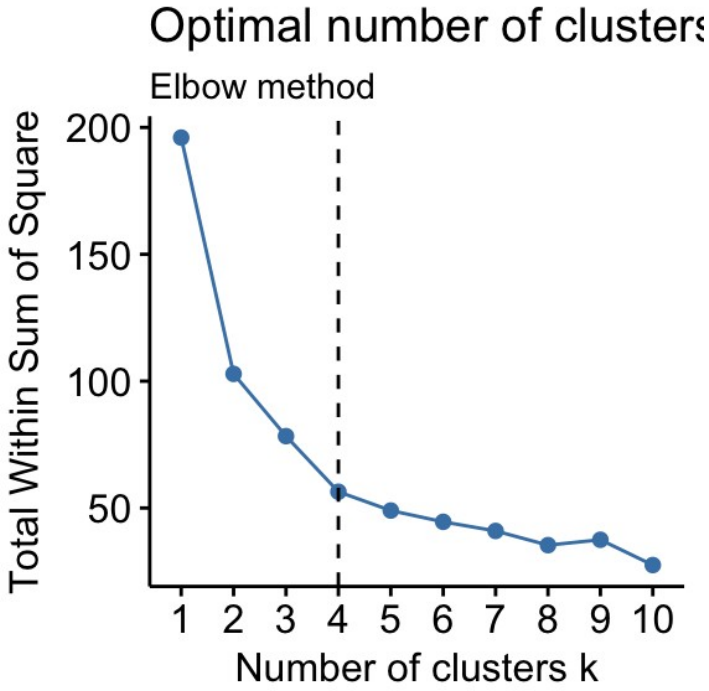
CLUSTERS: K-MEDIAS

- Es un procedimiento iterativo

- Consiste en la búsqueda de grupos (clusters) de los datos de tal forma que cada elemento esté asignado al grupo cuyo centro (centroide) sea el más cercano

- Hay que previamente especificar el número de grupos
  - Métodos gráficos: apoyados en técnicas de reducción de variables
  - Método analítico: "método del codo"
  - Criterio de negocio: "Quiero cinco segmentos..." 

- Específico para variables continuas:
  - Lo de la palabra "media" es por algo...



CLUSTERS: VARIABLES NO CONTINUAS

K-modas

- Similar al k-medias pero usando la moda como medida en vez de la media
- Especialmente útil para cuando las variables no son continuas

Cluster jerárquico

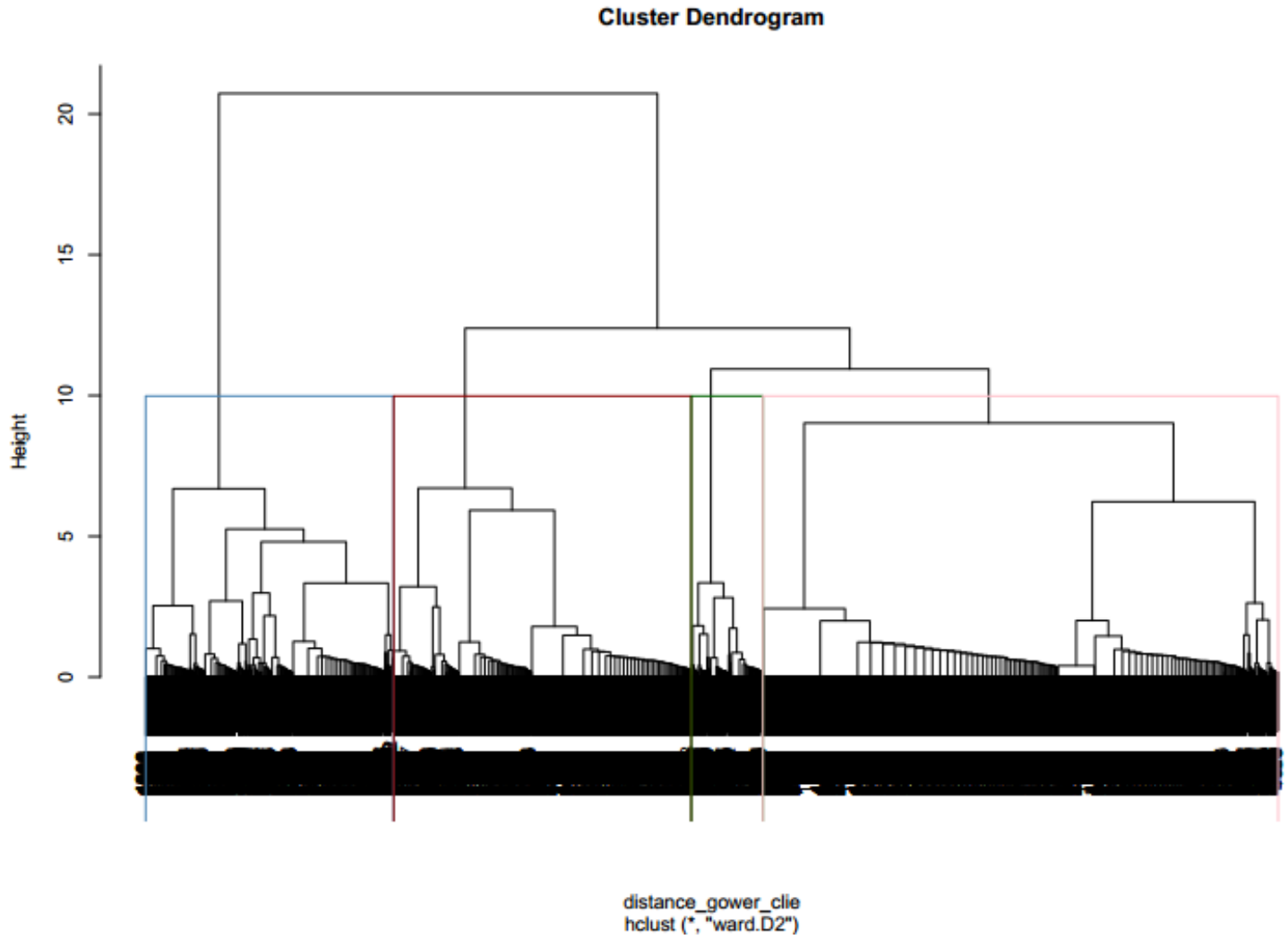
- Ascendente: partimos de todos los elementos por separado y vamos agrupando elementos comunes hasta tener los grupos definitivos
- Descendente: partimos de un único grupo el cual se va dividiendo sucesivamente
- Se debe calcular previamente una matriz de distancias, de similitudes o de disimilitudes (dependiendo del algoritmo y/o del tipo de datos)
- Suele tener un alto coste computacional
- También se puede utilizar como método auxiliar para obtener el número óptimo de clusters

"El truco de la distancia"

- Convertimos nuestro dataset en una matriz de distancia
- Se suele usar la distancia de Gower
- Finalmente Podemos usar cluster como: K-media, K-mediods,…

- https://towardsdatascience.com/clustering-on-mixed-type-data-8bbd0a2569c3

CLUSTERS: VARIABLES CONTINUAS Y/O DISCRETAS



CASO:

PCA y Cluster en R
y Python

MODELOS SUPERVISADOS

MÉTODOS SUPERVISADOS

Los tipos de problemas más habituales que se resuelven con técnicas de Machine Learning son:

- Regresión: En problemas de regresión los algoritmos aprendan a predecir el valor de una variable continua a partir de una o más variables explicativas.

- Clasificación: En los problemas de clasificación se busca que los algoritmos aprendan a predecir valores discretos a partir de una o más variables explicativas.

PREGUNTA

¿De qué tipo es la regresión logística?

MÉTODOS SUPERVISADOS

- El objetivo en los métodos supervisados es minimizar la función de coste o esfuerzo
- Para minimizar esta función se puede recurrir a diferentes técnicas que dependerán del modelo, los datos, la eficiencia requerida…
- Como técnicas de minimización podemos destacar:
  - Mínimos Cuadrados
  - Descenso por Gradiente
  - Máxima Verosimilitud
  - Newton-Raphson

MÉTODOS SUPERVISADOS

TÉCNICAS SUPERVISADAS "CLÁSICAS"		Variable Dependiente			
		Una		Muchas	
		Métrica	No Métrica	Métrica	No Métrica
Variable Independiente	Métrica	Regresión Polinómica Múltiple	Discriminante	Correlaciones Canónicas	Correlaciones Canónicas
		Árboles de Regresión	Logit	Ecuaciones Estructurales	Ecuaciones Estructurales
			Árboles de Decisión		
	No Métrica	ANOVA / ANCOVA*	Análisis Conjunto	MANOVA / MANCOVA*	Correlaciones Canónicas
		Regresión Polinómica Múltiple	Discriminante	Modelos Log-Lineales	Ecuaciones Estructurales
		Árboles de Regresión	Logit	Ecuaciones Estructurales	
			Árboles de Decisión		

* variables mixtas

MÉTODOS SUPERVISADOS

TÉCNICAS SUPERVISADAS "MACHINE LEARNING"	Redes Neuronales Artificiales
	Random Forests
	Máquinas Vector Soporte
	Algoritmos Bayesianos
	Algoritmos Genéticos
	Vecinos Cercanos
	Gradient Boosting Machines

ALGUNOS MODELOS DE ML SUPERVISADOS

Técnica	Principales Ventajas	Principales Inconvenientes
Máquinas de Vector Soporte	• Robustez incluso con muestras sesgadas • Única solución	• Baja velocidad de proceso y detección de patrones • Precisión media • Falta de transparencia de los resultados
Redes Bayesianas	• Alta velocidad de proceso y detección • Alta Precisión	• Excesiva necesidad de entrenamiento
Sistemas Basados en Lógica Difusa	• Buena precisión	• Alto consumo
Sistemas Expertos	• Facilidad de construcción y desarrollo del sistema • Alto grado de precisión • Facilidad de explicación • Buena combinación con otros modelos para la extracción de reglas	• Pobre en el manejo de datos incompletos o valores no esperados • Pobre en el proceso con tipos de datos diferentes • Aún estos sistemas no se aproximan a la flexibilidad del lenguaje humano

A Survey of Credit Card Fraud Detection Techniques. Sorournejad et al. (2016)

ALGUNOS MODELOS DE ML SUPERVISADOS

Técnica	Principales Ventajas	Principales Inconvenientes
Redes Neuronales Artificiales	• Aprenden del pasado • Extracción de reglas y predicción • Alta precisión • Alta velocidad de detección • Tiempo real • Adaptabilidad y mantenimiento	• Redes muy largas necesitan mucho tiempo de procesamiento • Poco "explicativas" • Sensible al formato de los datos • Dificultad de configuración
Sistemas Inmunes Artificiales	• Efecto memoria • Datos no balanceados	• Consumo computacional • Mucho entrenamiento
Algoritmos Genéticos	• Buen tratamiento del "ruido" • Fácilmente construibles y mantenibles • Mejoran la eficiencia de otras técnicas • Rápida detección	• De complejo entendimiento
Modelo Oculto de Markow	• Rápida detección	• Muchos recursos • Poca precisión • No escalable a grandes datos

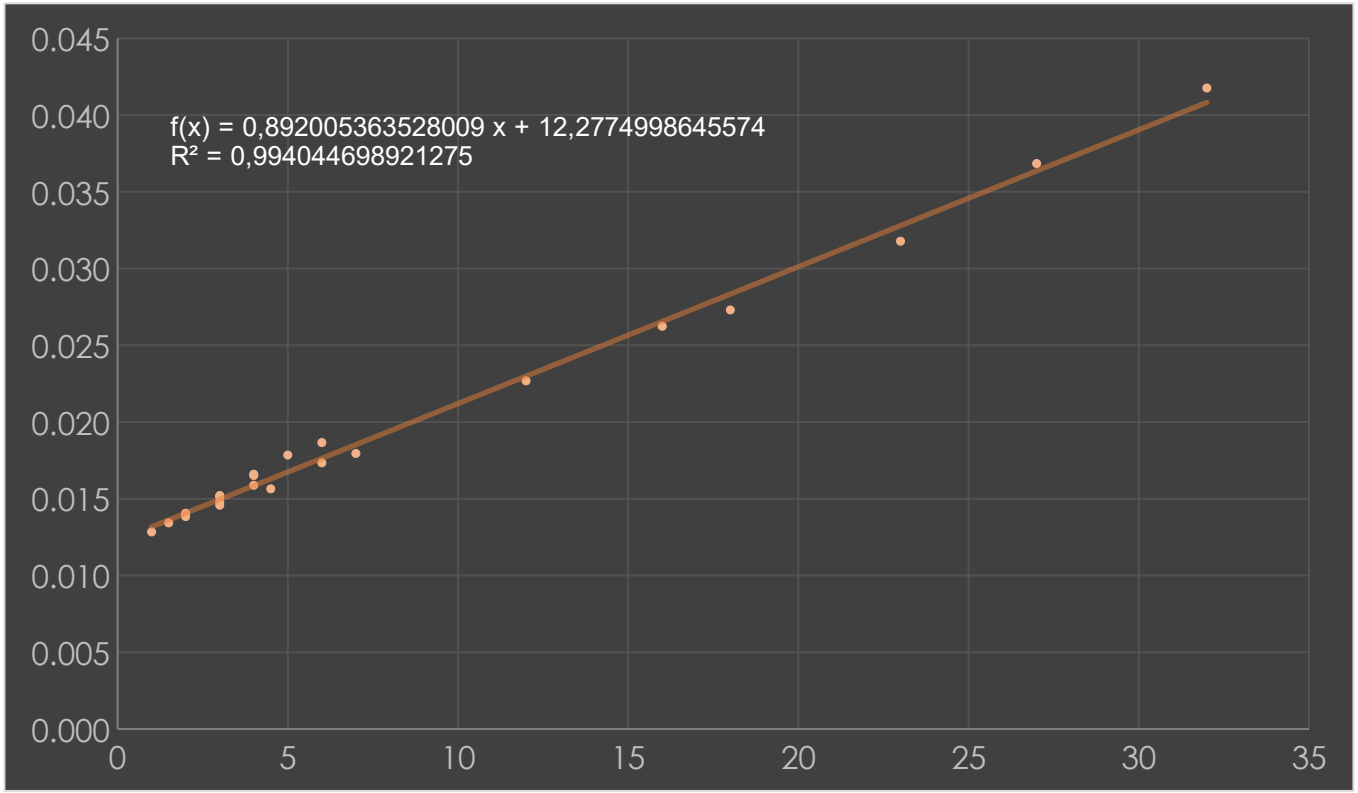
A Survey of Credit Card Fraud Detection Techniques. Sorournejad et al. (2016)

ALGUNOS MODELOS DE ML SUPERVISADOS

Técnica	Principales Ventajas	Principales Inconvenientes
Programación Lógica Inductiva	• Poderosa con datos de diferentes tipos • Modela relaciones complejas del lenguaje • Poderosa en el manejo de datos incompletos	• Baja precisión • Extremadamente sensible al ruido • Problemas con datos espurios
Razonamiento Basado en Casos	• Útil en dominios con muchos ejemplos • Efectivo, flexible y fácil de mantener • Se puede usar para modelos híbridos	• Puede sufrir por datos ruidosos y/o incompletos
Árboles de Decisión	• Alta Flexibilidad • Fácil de explicar e implementar	• Requiere de comprobaciones de cada condición una a una

A Survey of Credit Card Fraud Detection Techniques. Sorournejad et al. (2016)

REGRESIÓN: EL MODELO LINEAL



$$y = bx + n + \epsilon \longrightarrow \hat{y} = \hat{b}x + \hat{n}$$

REGRESIÓN: MÍNIMOS CUADRADOS ORDINARIOS (MCO)

Tenemos la siguiente función:

$y = f(x) = b_1x_1 + b_ix_i + \ldots + b_nx_n + n + \varepsilon$

Matricialmente sería:

$Y = XB + \varepsilon$

Pero como tenemos datos reales las cosas nos son perfectas...

$y' = b'_1x_1 + b'_ix_i + \ldots + b'_nx_n + n'$

$Y' = XB'$

Nuestro objetivo será estimar los b'_i

El método de los mínimos cuadrados ordinarios (MCO ó OLS) consiste en minimizar la distancia entre *y* (valores de la variable dependiente real) y *y'* (valores estimados de la variable dependiente)

Para poder usar este método se debe asumir los siguiente:

- Linealidad: la variable dependiente es una combinación lineal de las variables independientes
- Ausencia de multicolinealidad (entre variables independientes)
- Homocedasticidad de los errores
- No existencia de autocorrelación de los errores
- Normalidad de los residuos (en este caso existen discrepancias entre autores)

REGRESIÓN: EJEMPLO

OLS Regression Results

Dep. Variable:	prestige	R-squared:	0.828
Model:	OLS	Adj. R-squared:	0.820
Method:	Least Squares	F-statistic:	101.2
Date:	Wed, 17 Jul 2019	Prob (F-statistic):	8.65e-17
Time:	15:56:04	Log-Likelihood:	-178.98
No. Observations:	45	AIC:	364.0
Df Residuals:	42	BIC:	369.4
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-6.0647	4.272	-1.420	0.163	-14.686	2.556
income	0.5987	0.120	5.003	0.000	0.357	0.840
education	0.5458	0.098	5.555	0.000	0.348	0.744

Omnibus:	1.279	Durbin-Watson:	1.458
Prob(Omnibus):	0.528	Jarque-Bera (JB):	0.520
Skew:	0.155	Prob(JB):	0.771
Kurtosis:	3.426	Cond. No.	163.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

MODELOS DE REGRESIÓN: R^2

Coeficiente de Determinación

$$R^2 = \frac{\sum_{t=1}^{T} (\hat{Y}_t - \bar{Y})^2}{\sum_{t=1}^{T} (Y_t - \bar{Y})^2}$$

Coeficiente de Determinación Ajustado

$$\bar{R}^2 = 1 - \frac{N-1}{N-k-1}[1 - R^2]$$

Debe tomar valores entre 0 y 1, siendo 1 el "ajuste perfecto"

EVALUACIÓN DE MODELOS DE REGRESIÓN

MAE calculates the mean absolute error:

$$\frac{1}{n} \cdot \sum_{i=1}^{n} |ref_i - x_i|$$

MAPE calculates the mean absolute percentage error:

$$\frac{1}{n} \cdot \sum_{i=1}^{n} \left| \frac{ref_i - x_i}{ref_i} \right|$$

SMAPE calculates the symmetric mean absolute percentage error:

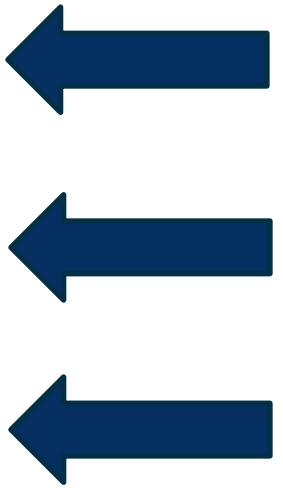
$$\frac{1}{n} \cdot \sum_{i=1}^{n} \frac{2 \cdot |ref_i - x_i|}{|ref_i| + |x_i|}$$

MSE calculates mean squared error:

$$\frac{1}{n} \cdot \sum_{i=1}^{n} (ref_i - x_i)^2$$

RMSE calculates the root mean squared error:

$$\sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (ref_i - x_i)^2}$$



SON ROBUSTOS

CONTROLES

	Name of Test	Test for	Null Hypothesis	Critical Value	P Value
1	Unit Root Test	Unit roots	Y has a unit Root	Varies	0.05 or less
2	a) R sq	Model Fit	N/A	85% or above	N/A
	b) Adjusted R sq	Model Fit	N/A	85% or above	N/A
3	Significance	Parameter Statistical Significance	beta = 0	1.96 or 1.6 or higher	0.05 or 0.10 or less
4	a) Correlation	Multicollinearity	N/A	Regressor Correlation < 0.75	N/A
	b) VIF	Multicollinearity	N/A	Centered VIF < 10	N/A
5	a) Breusch Godfrey	Residual serial Correlation	No Serial Correlation	Varies	Larger than 0.05 or 0.1
	b) D. W.	Residual serial Correlation	No Serial Correlation	2	
6	Jarque Bera	Residual Normality	Residuals are Normally Distributed		Larger than 0.05 or 0.1
7	White	Heteroscedasticity	Residuals are homoscedastic		Larger than 0.05 or 0.1
8	Ramsey	Specification Test	Correct Specification is Linear		Larger than 0.05 or 0.1
9	a) AIC	Model Selection	N/A	Min AIC	Min AIC
	b) BIC	Model Selection	N/A	Min BIC	Min BIC

MODELOS DE REGRESIÓN: Extensiones

- **Variables ficticias o *dummies*: *one hot encoding***
  - Cuando queremos pasar una variable categórica a continua
  - Si la variable categórica tiene n clases, crearemos n-1 variables

	España	Francia
España	1	0
Francia	0	1
Italia	0	0

MODELOS DE REGRESIÓN: Extensiones

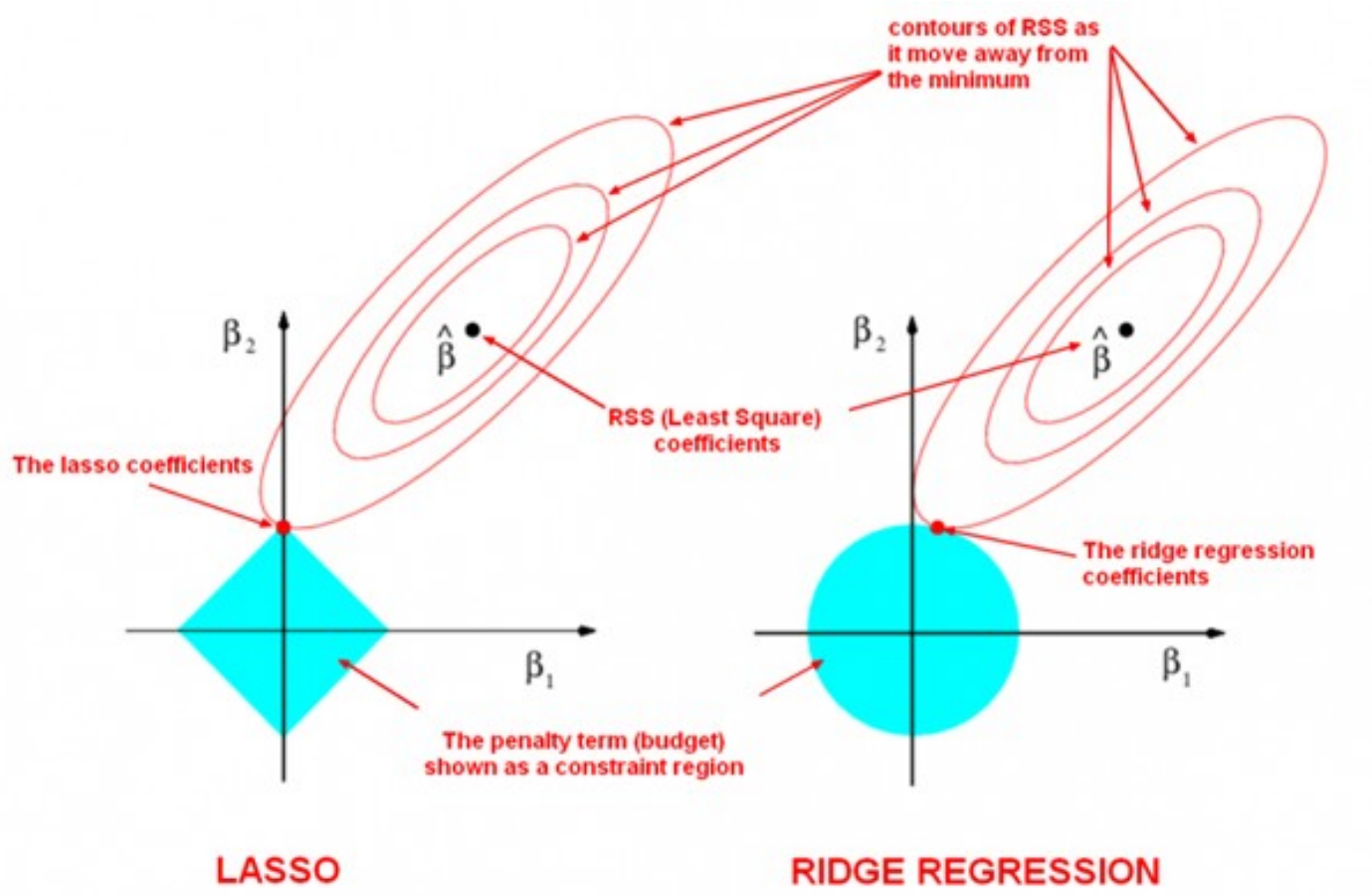
- **Regularización (Ridge y Lasso)**: se ajusta el modelo con todos los *p* predictores, y el valor de los coeficientes estimados es reducido (algunos pueden llegar a ser exactamente 0, con lo que son excluidos del ajuste). Esta reducción tiene el efecto de reducir la varianza.

- ***Subset selection***: se basa en identificar un subconjunto de los *p* predictores que pensamos están relacionados con la variable respuesta. Una vez seleccionados, se ajusta el modelo con dichos predictores mediante mínimos cuadrados.

- **Reducción dimensional**: este enfoque se basa en proyectar los *p* predictores en un subespacio M-dimensional, donde M < *p*, lo cual se consigue obteniendo M combinaciones lineales o proyecciones diferentes de las variables. Estas proyecciones son utilizadas como predictores para ajustar el modelo de regresión.

https://rstudio-pubs-static.s3.amazonaws.com/392216_5adc98e84ff246e3b6418f3426c8fe38.html

MODELOS DE REGRESIÓN: Extensiones



An Introduction of Statistic Learning with Applications in R (Gareth et al; 2013)

MODELOS DE REGRESIÓN:
¿Qué pasa si no cumplimos los supuestos MCO?

- **Métodos Robustos**
  - No son tan sensibles a los supuestos de MCO
  - Útiles cuando existe heterocedasticidad o valores atípicos
  - Estos métodos ponderan las observaciones dando un menor peso a los atípicos
  - Extensible a la estadística descriptiva
- **Optimización por Gradiente Descendente**
  - ¡Lo veremos más adelante!

CASO:

Modelos de Regresión en R y Python

ANOVA/ANCOVA

- Es un conjunto de modelos
- Basados en los métodos de regresión
- Nos ayuda a saber si distintas muestras de datos pertenecen o no a una misma población testando si hay diferencias significativas
- La variable dependiente debe ser continua o cuasi-continua
- Las variables independientes deben ser discretas
- Varios tipos:
  - ANOVA/ANCOVA un factor: una variable dependiente, una variable independiente
  - ANOVA/ANCOVA m factores: una variable dependiente, varias variables independientes
  - MANOVA/MANCOVA : varias variables dependientes, una o varias variables independientes

EVALUACIÓN DE MODELOS DE CLASIFICACIÓN

	Predicted: Yes	Predicted: No
Real: Yes	True Positive (TP)	False Positive (FP)
Real: No	False Negative (FN)	True Negative (TN)

$\text{TP Rate}_k = TP_k/(TP_k+FN_k)$

$\text{FP Rate}_k = FP_k/(FP_k+TN_k)$

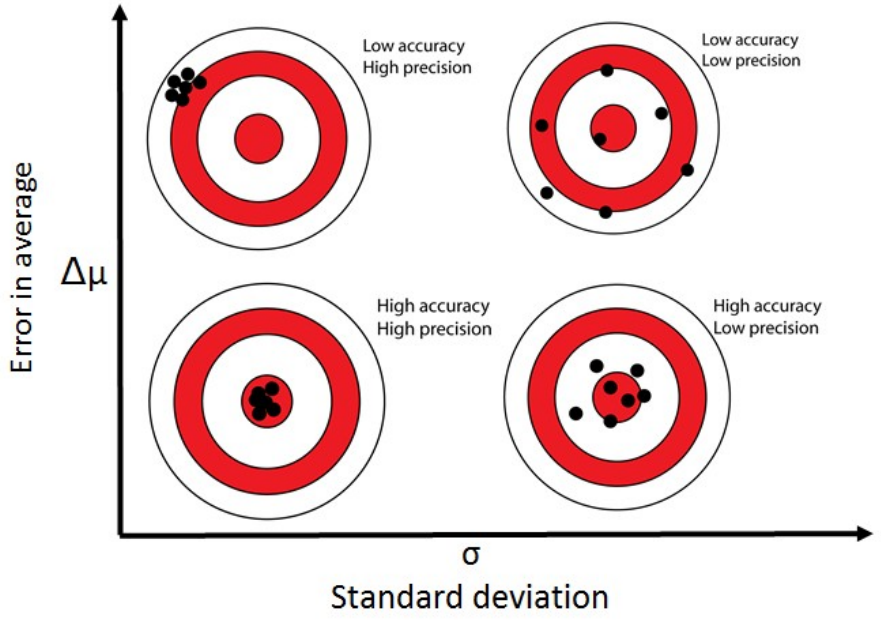
$\text{Accuracy}_k = (TN_k+TP_k)/(TN_k+TP_k+FN_k+FP_k)$

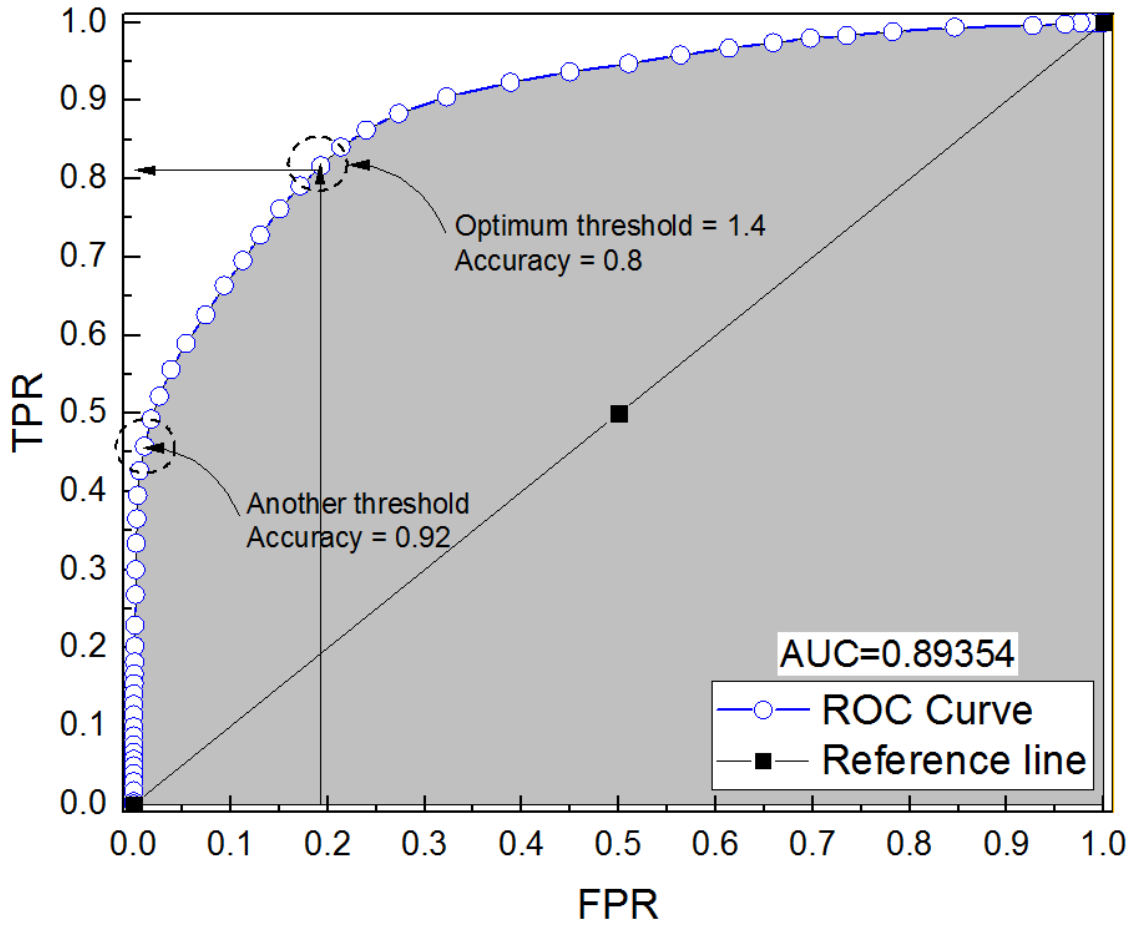
$\text{Precision}_k = TP_k/(TP_k+FP_k)$

$\text{Recall}_k = TPR_k = TP_k/(TP_k+FN_k)$

$\text{F1 Score}_k = 2*(\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$

→ **PROBLEMA DE OPTIMIZACIÓN**





CURVA ROC

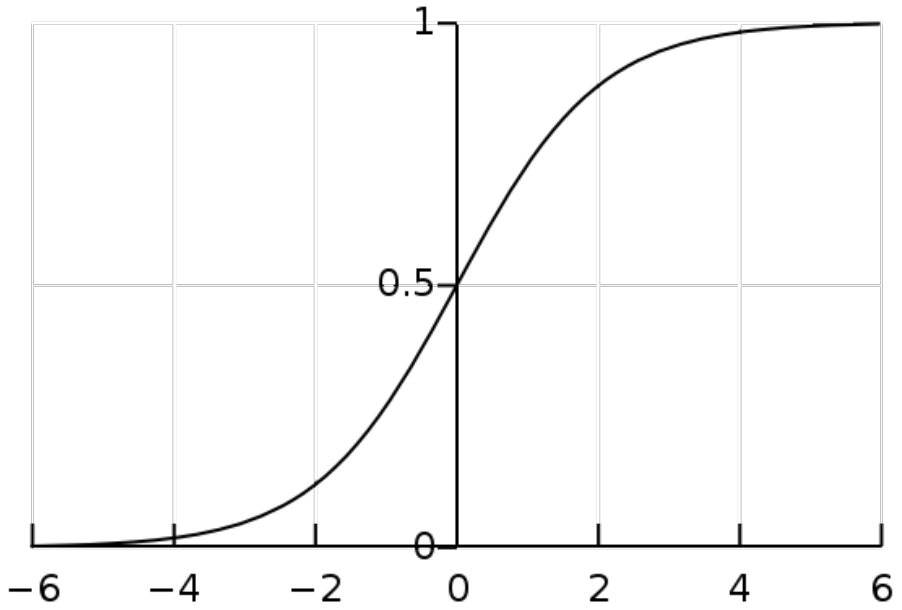
TPR = TP/(TP+FN)
FPR = (FP)/(TN+FP)

https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

https://www.bioestadistica.uma.es/analisis/roc1/

REGRESIÓN LOGÍSTICA

- Es un modelo discriminante binario, es decir nos ayuda a discriminar entre dos grupos.
- La variable dependiente deber ser binaria (dummy): toma valores entre 0 y 1.



$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}}$$

$$logit(p) = ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

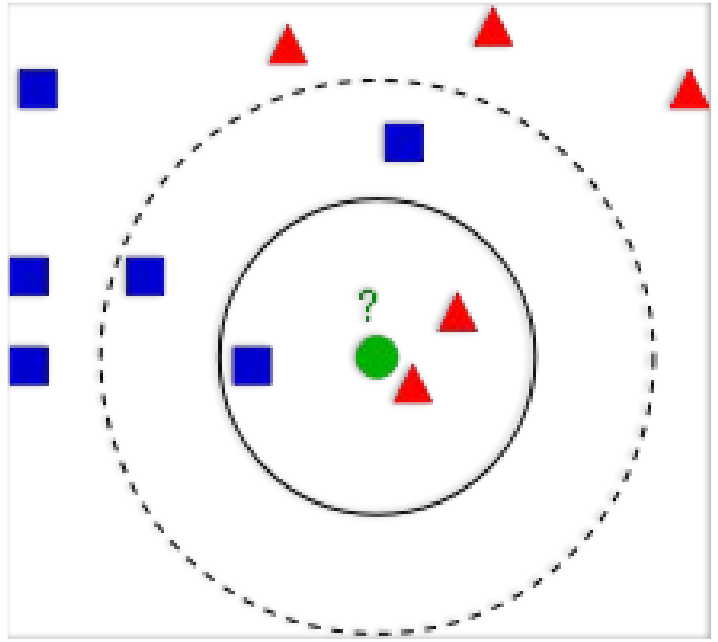
K VECINOS CERCANOS

- En el entrenamiento del algoritmo consiste en almacenar los vectores característicos y las etiquetas de las clases de los ejemplos de entrenamiento

- Para clasificar, se calcula la distancia entre los vectores almacenados y el nuevo vector, y se seleccionan los *k* ejemplos más cercanos. El nuevo ejemplo es clasificado con la clase que más se repite en los vectores seleccionados.

- Existen variantes que usan la distancia ponderada

- Es conveniente escalar los datos



ÁRBOLES DE DECISIÓN Y REGRESIÓN

- Uno de los grupos de métodos más utilizados para clasificar ya que es muy fácil de interpretar: reduce cada clase a reglas

- En combinación con el análisis de clúster es una de las herramientas más poderosas: Ayuda a validar los clusters

- Busca encontrar aquellos rasgos que diferencian los elementos de clases distintas y asemejan dentro de la misma clase

ÁRBOLES DE DECISIÓN Y REGRESIÓN

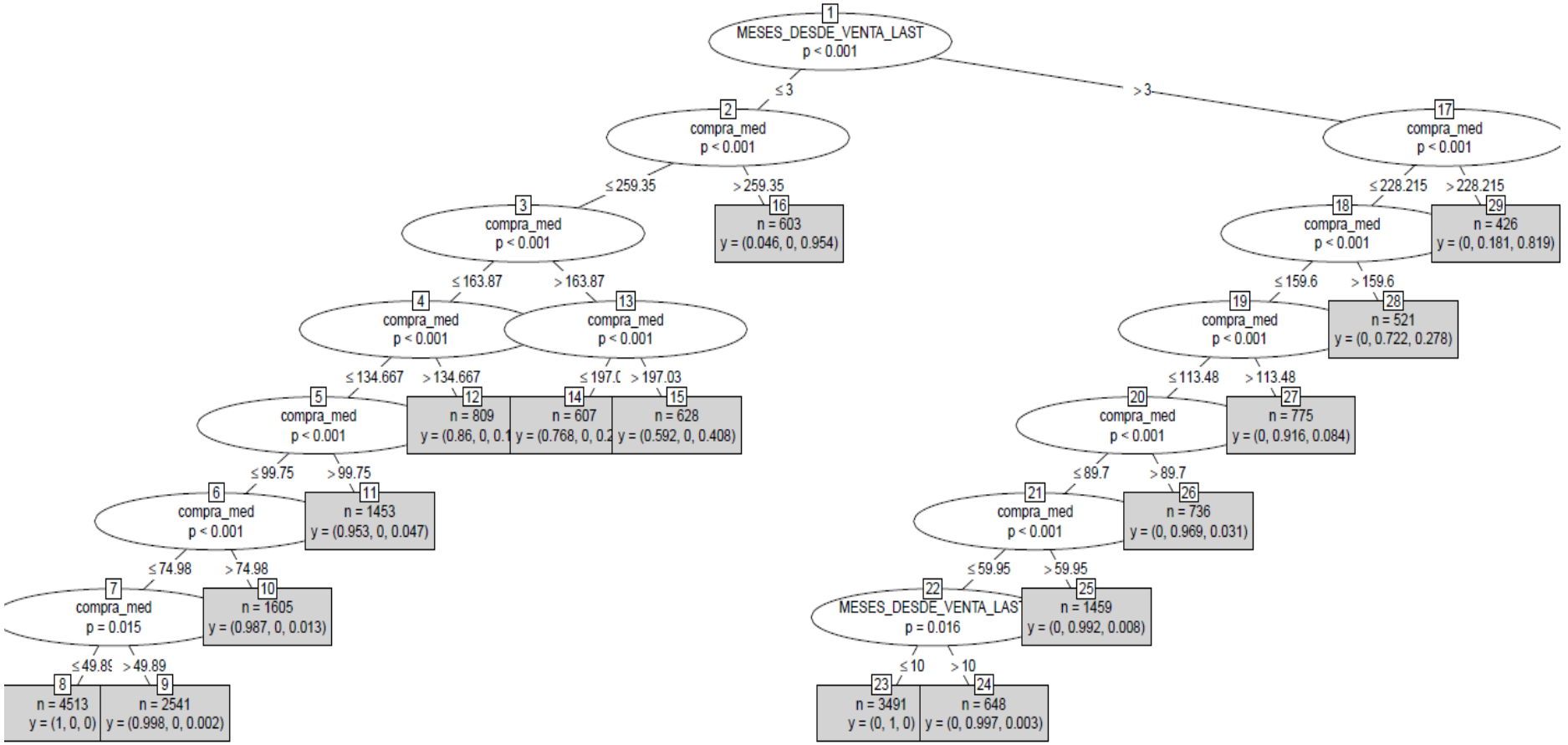
- Funcionamiento:
  - Al principio, todo el conjunto de entrenamiento se considera como raíz.
  - Los valores de las características son preferibles que sean categóricos.
  - Si los valores son continuos, entonces son discretizados antes de construir el modelo.
  - Las observaciones se distribuyen recursivamente sobre la base de valores de atributo.
  - El orden para colocar atributos como raíz o nodo interno del árbol se hace usando algún enfoque estadístico.

https://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/

ÁRBOLES DE DECISIÓN Y REGRESIÓN

- Diferentes estrategias y algoritmos en función del tipo de variables:
  - C5: https://rpubs.com/cyobero/C50
  - Árboles de Inferencia Condicional: https://rpubs.com/awanindra01/ctree
  - CHAID: https://www.r-bloggers.com/chaid-and-r-when-you-need-explanation-may-15-2018/
  - C&RT: http://www.sthda.com/english/articles/35-statistical-machine-learning-essentials/141-cart-model-decision-tree-essentials/
  - …
- Diferentes criterios de selección
  - Entropía: https://es.wikipedia.org/wiki/Entrop%C3%ADa_(informaci%C3%B3n)
  - Índice de Gini: https://es.wikipedia.org/wiki/Coeficiente_de_Gini
  - Test de significancia
  - Error de clasificación
  - Varianza

ÁRBOLES DE DECISIÓN Y REGRESIÓN

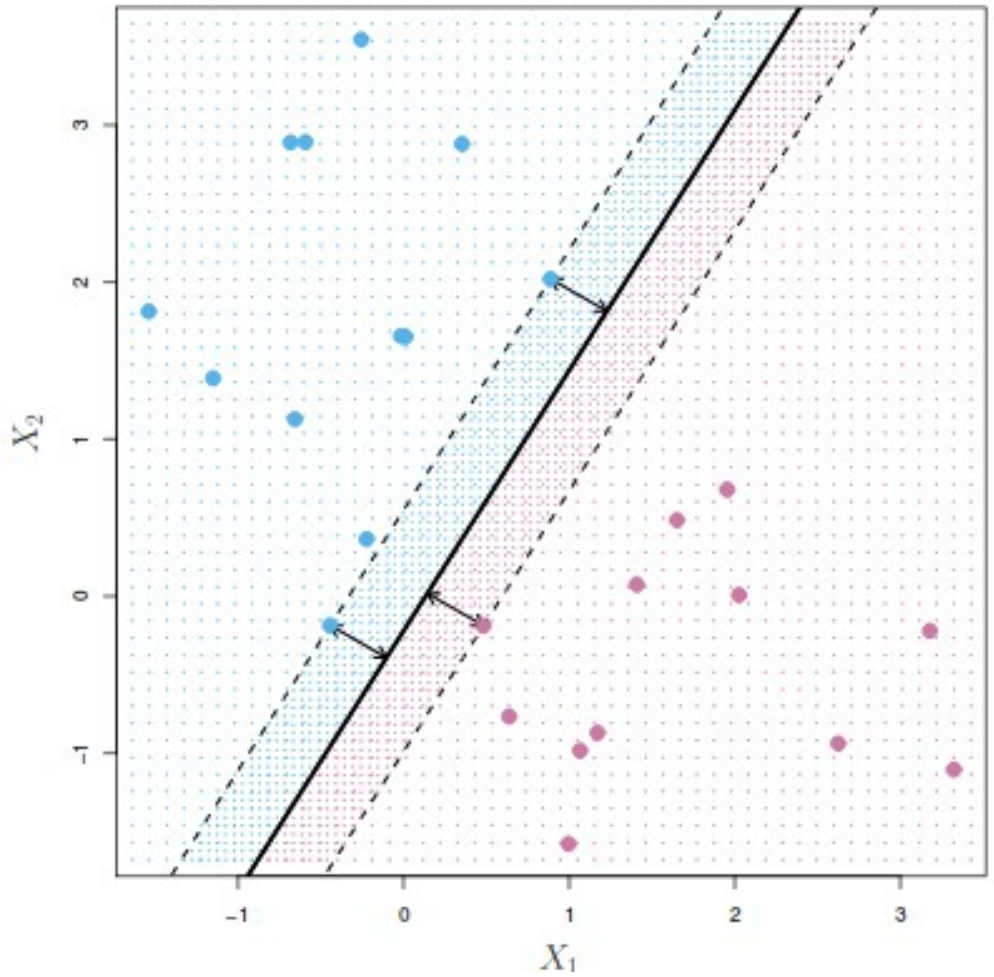


ÁRBOLES DE DECISIÓN

Grupo: Tipo de Cliente	Condiciones
VIP	• MESES_DESDE_VENTA_LAST <= 3 & compra_med > 400 • MESES_DESDE_VENTA_LAST > 3 & compra_med > 250
Buen Cliente	• MESES_DESDE_VENTA_LAST <= 3 & compra_med < 400
Estándar	• MESES_DESDE_VENTA_LAST > 3 & compra_med < 250

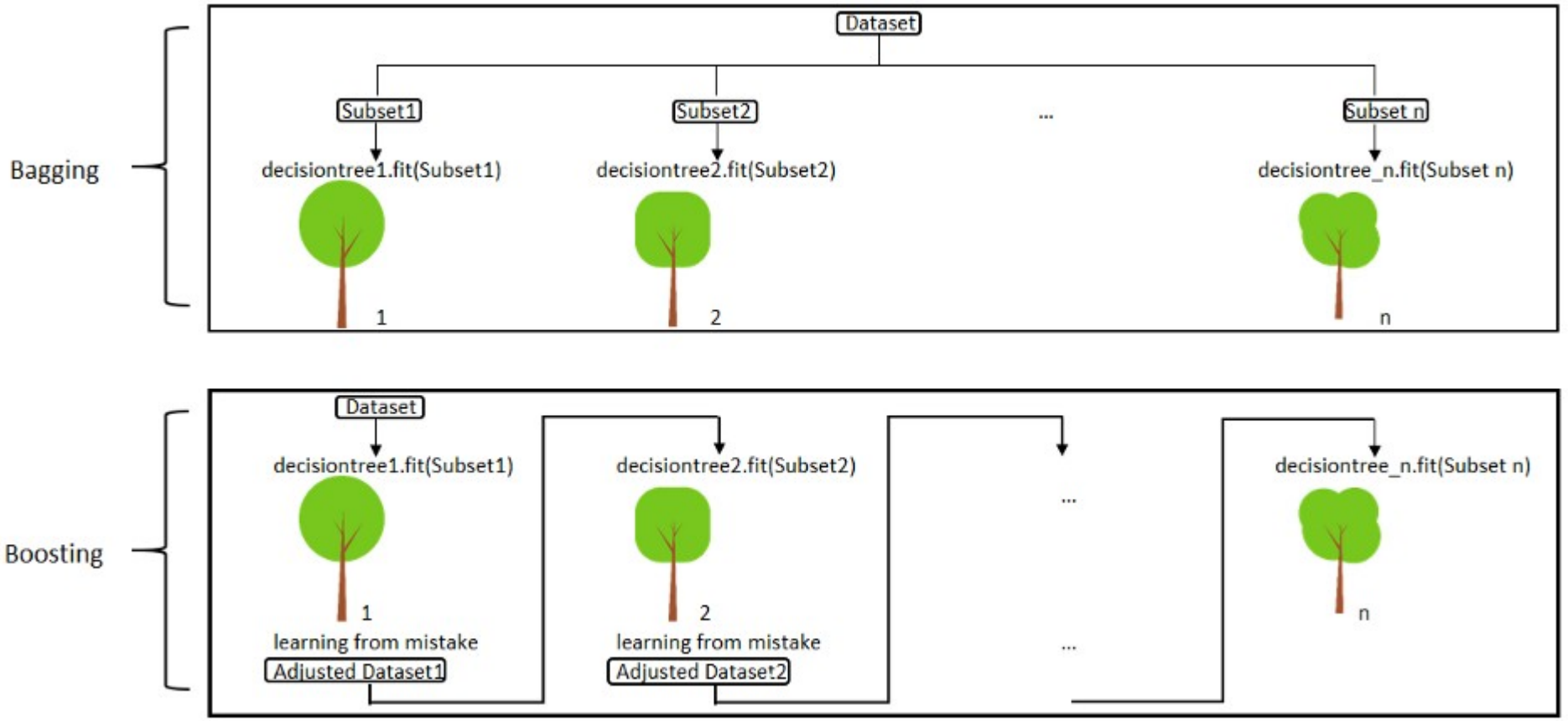
MÁQUINAS DE VECTOR DE SOPORTE

- Consiste en encontrar los hiperplanos que separan las distintas clases.
- Los hiperplanos no tienen porque ser euclídeos (truco del kernel).
- Es una de las técnicas más usadas de machine learning
- Útil si tienes pocas observaciones y muchas variables
- No tan fácil de interpretar como los árboles
- Se pueden usar una variante para regresión



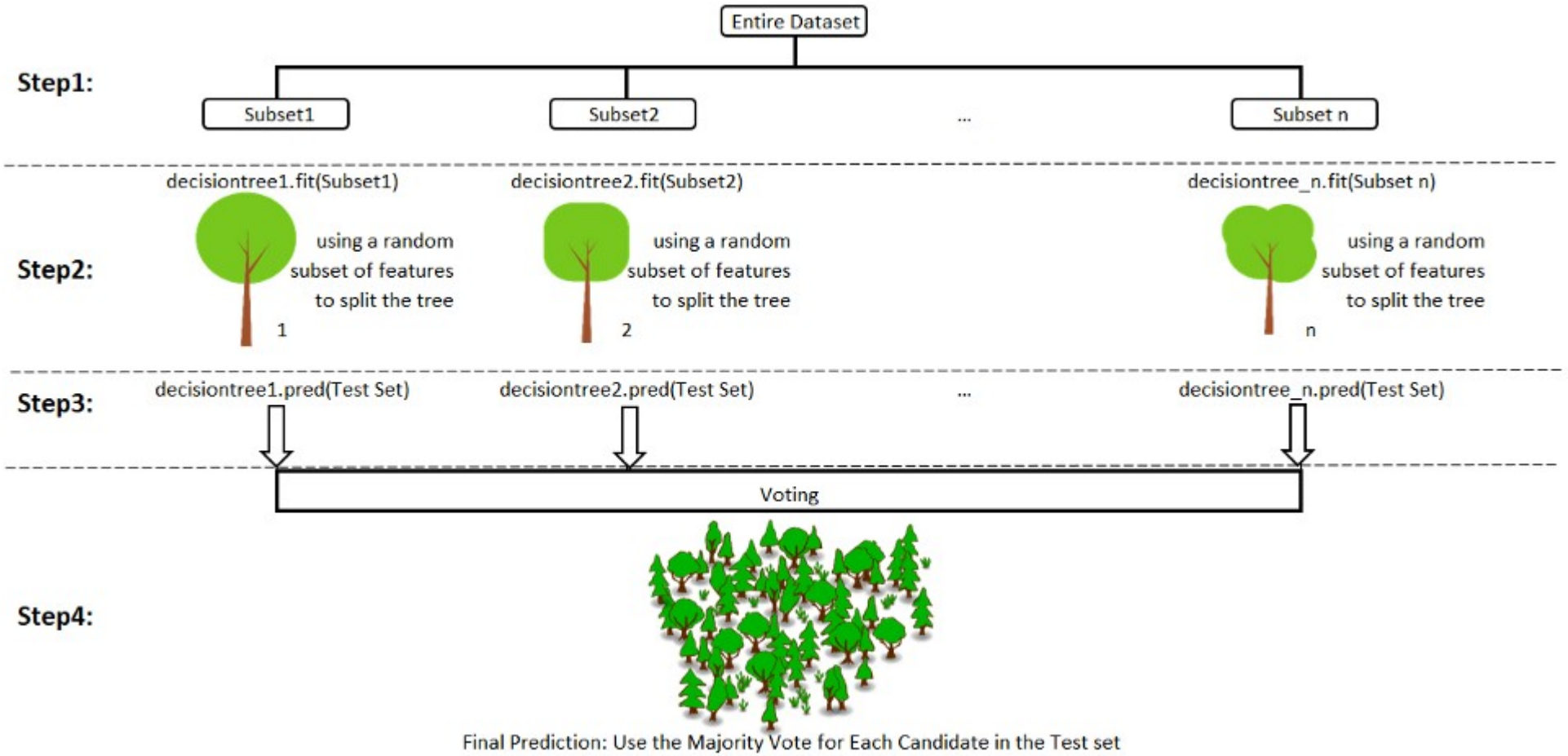
- https://rpubs.com/Joaquin_AR/267926
- http://ligdigonzalez.com/maquina-de-vectores-de-soporte-regresion-teoria/
- https://statinfer.com/204-6-8-svm-advantages-disadvantages-applications/

BAGGING Y BOOSTING

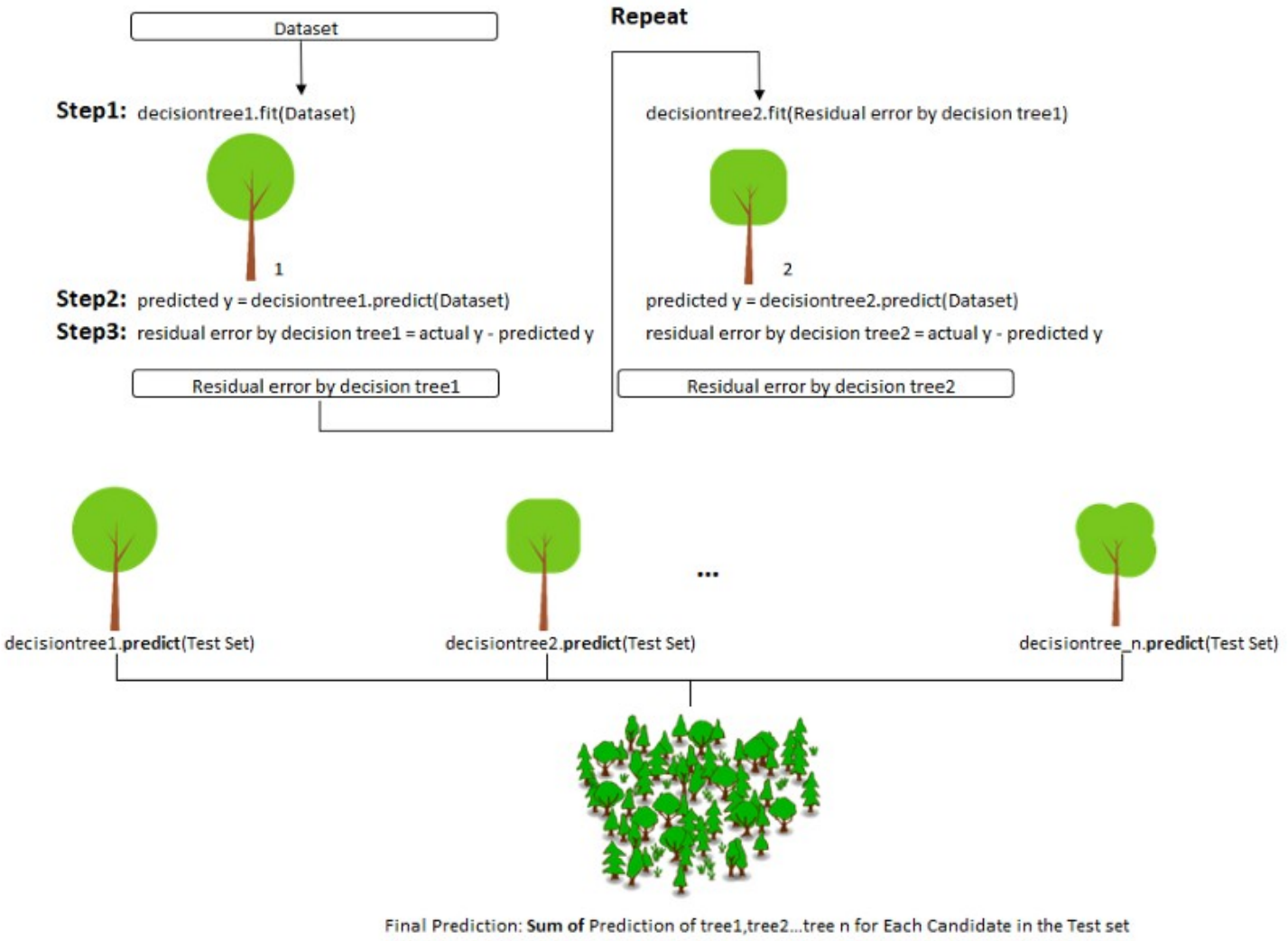


https://towardsdatascience.com/basic-ensemble-learning-random-forest-adaboost-gradient-boosting-step-by-step-explained-95d49d1e2725

RANDOM FOREST

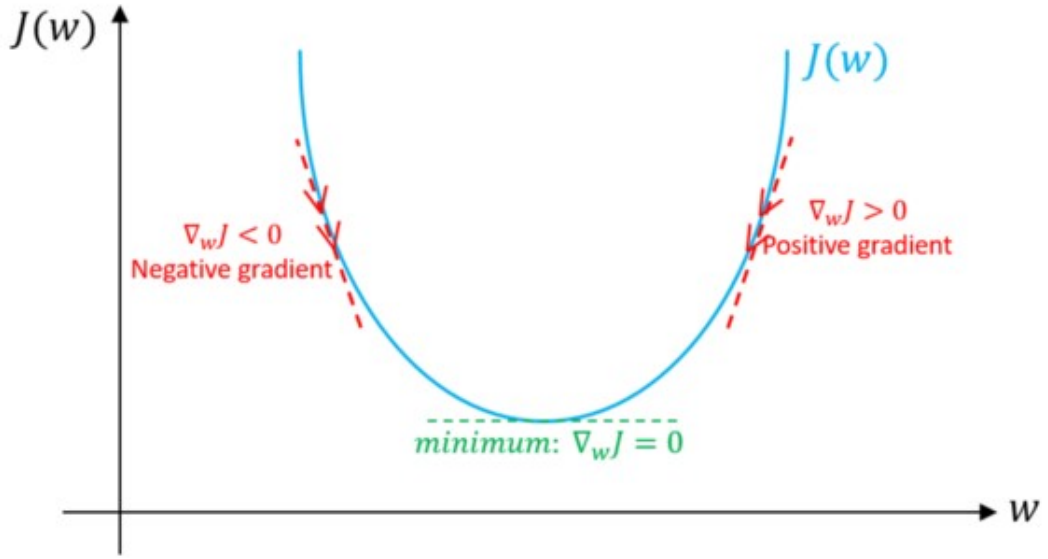


GRADIENT BOOSTING MACHINE



DESCENSO POR GRADIENTE

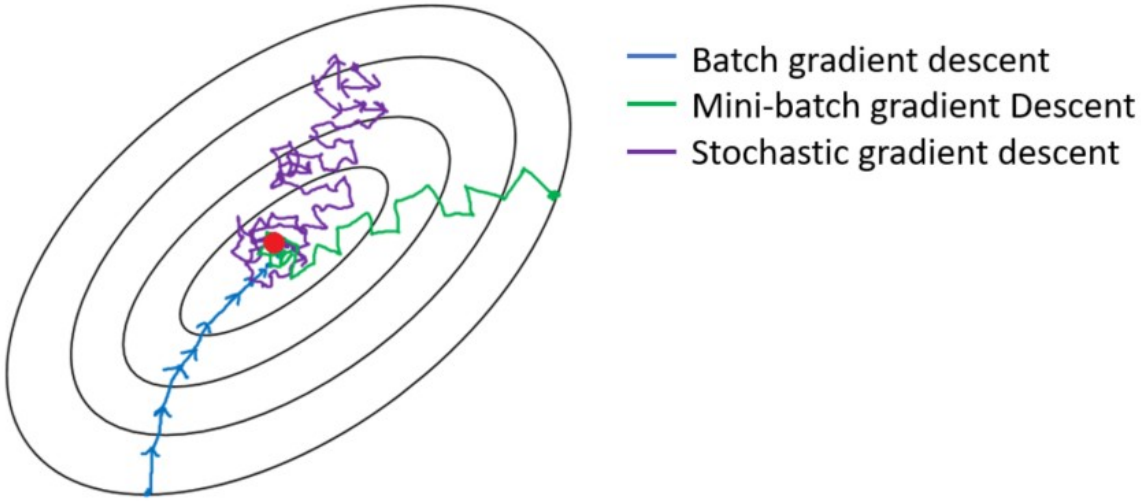
- Es una solución iterativa para minimizar la función de coste
- Si se amplía el conjunto de datos se refina el entrenamiento (ajuste fino) y no es necesario entrenar el modelo al completo (como ocurre con MCO)
- No es necesario que se cumplan algunos supuestos de MCO
- La idea parte del ir ajustando por mini-batches (submuestras de igual tamaño)
  - Existe un caso extremo en el que en cada iteración sólo se toma una observación: *descenso por gradiente estocástico*

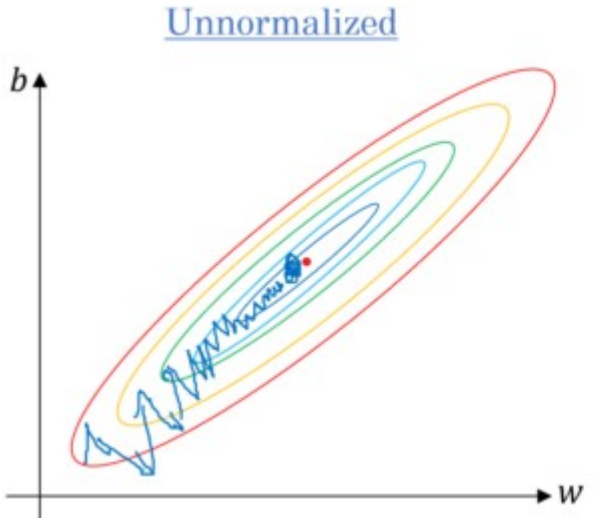


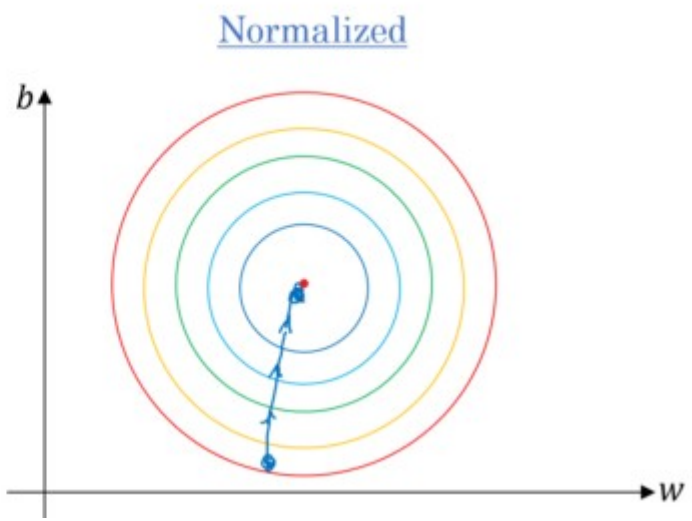
DESCENSO POR GRADIENTE

- Muy importante es que los datos estén normalizados
- Cuidado con mínimos locales o divergencia (explosión del gradiente)
- Hay que seleccionar adecuadamente la tasa de aprendizaje y las condiciones iniciales
- Este método es ideal para tratar con grandes datos

- https://en.wikipedia.org/wiki/Gradient_descent





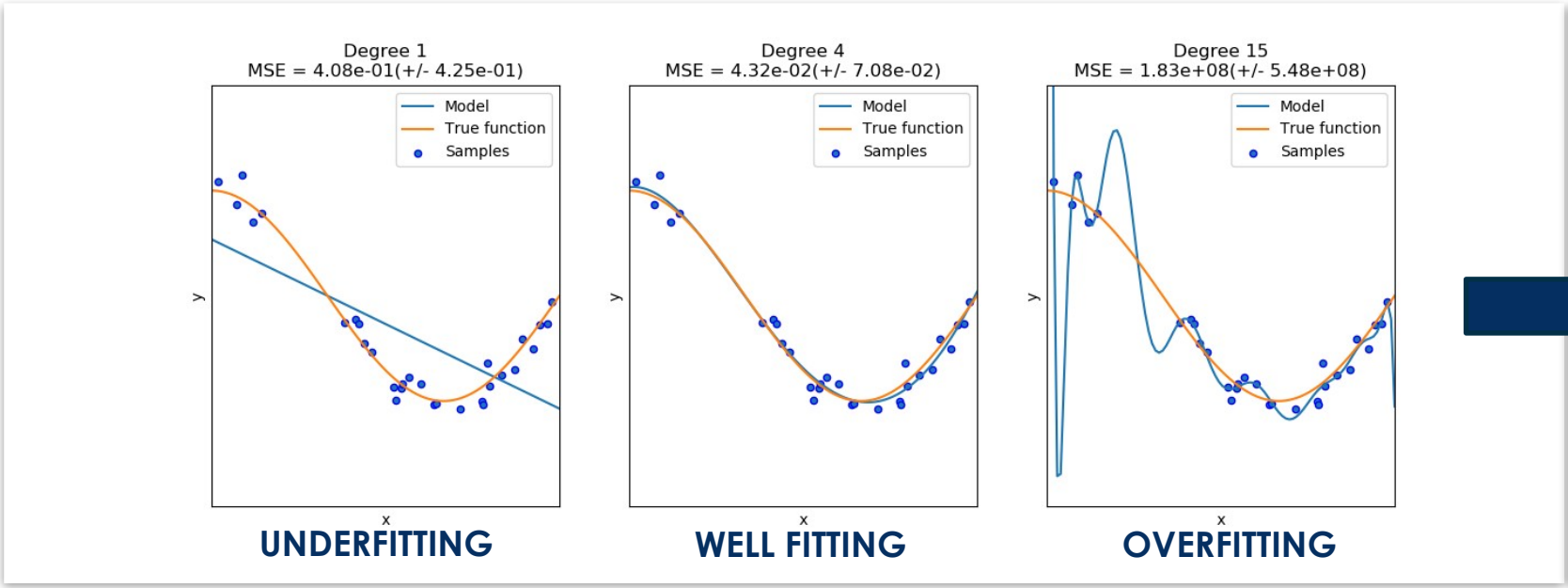


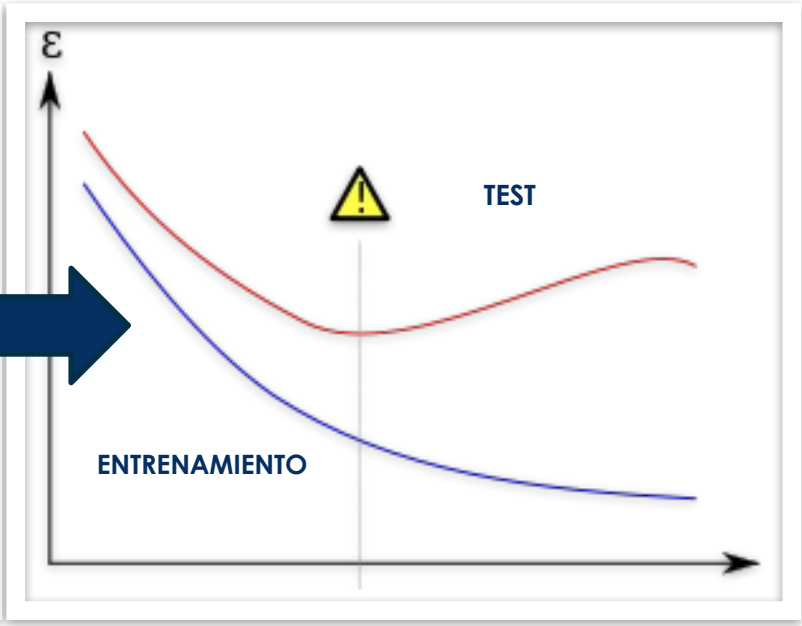
CASO:

Modelo de clasificación en Python

EL PROBLEMA DEL SOBREAJUSTE

- El modelo aprende "demasiado bien" el conjunto de entrenamiento y, por tanto, el modelo no es capaz de acertar en los datos de test.
- Obtenemos un rendimiento deficiente en el conjunto de datos de validación





CASO:

Ejemplo overfitting en R

OTRAS TÉCNICAS Y CONCEPTOS RELACIONADOS

Kernels

Muestreo: tamaño, generación de muestras…

Validación cruzada

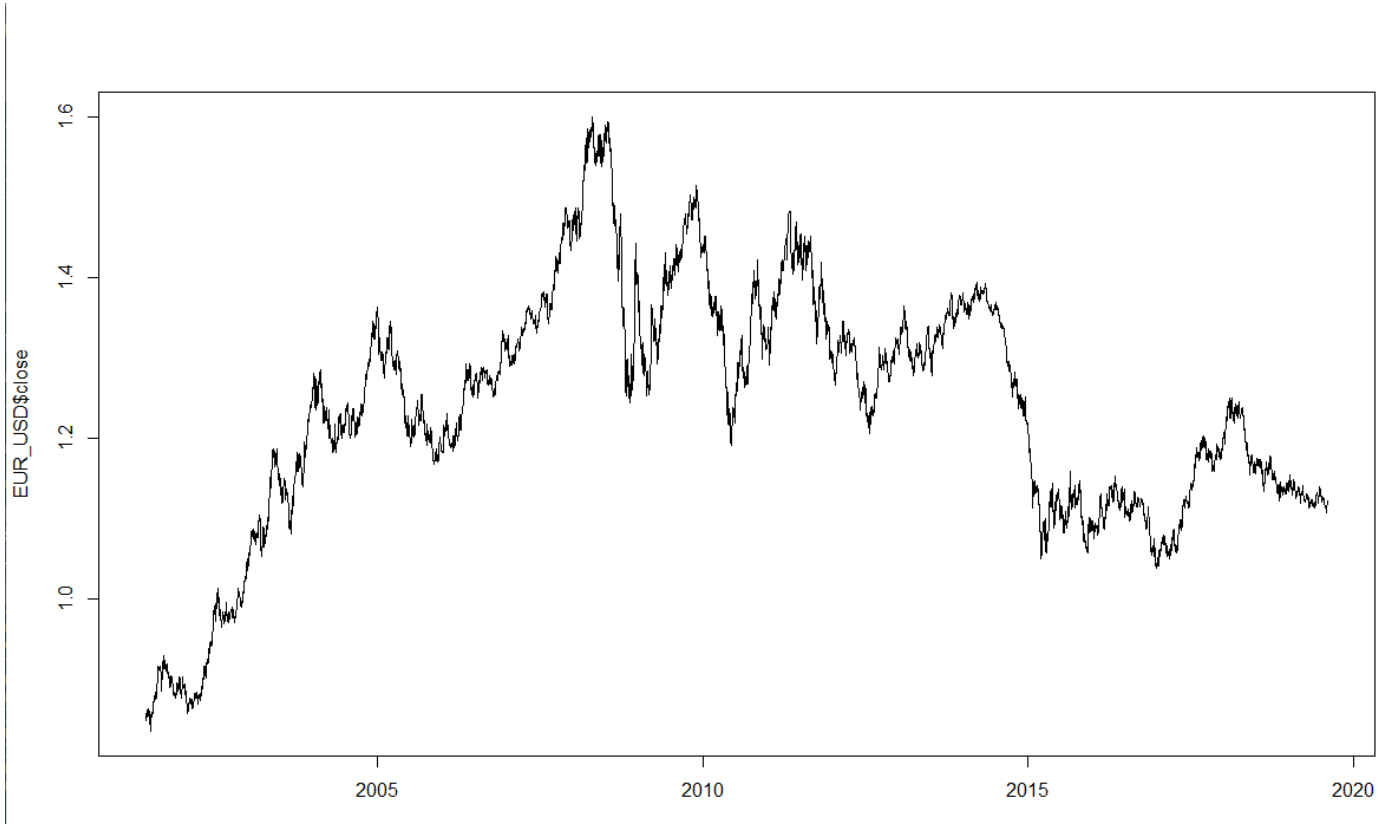
Series temporales

Estadística Espacial y Sistemas de Información Geográfica

SERIES TEMPORALES

CONCEPTOS

Una serie temporal es una estructura de datos en la que cada registro tiene asociado un identificador de tiempo (timestamp), ya sea diario, semanal, mensual, hora, etc.



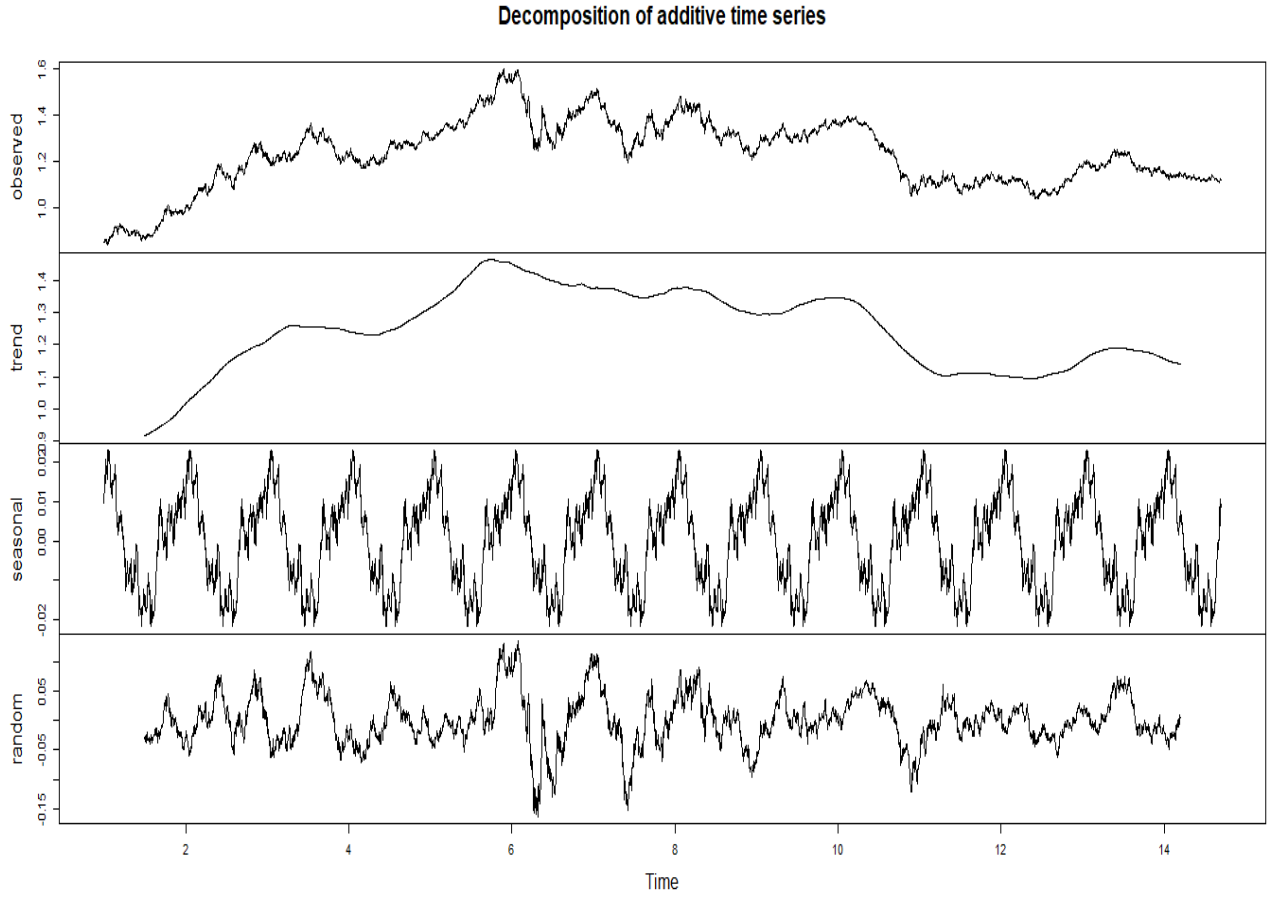
CONCEPTOS

Componentes:

Tendencia: Es el comportamiento creciente/decreciente de la serie a largo plazo

Estacional: Aparece cuando hay componentes de estacionalidad, es decir, parámetros que se repiten con determinada frecuencia

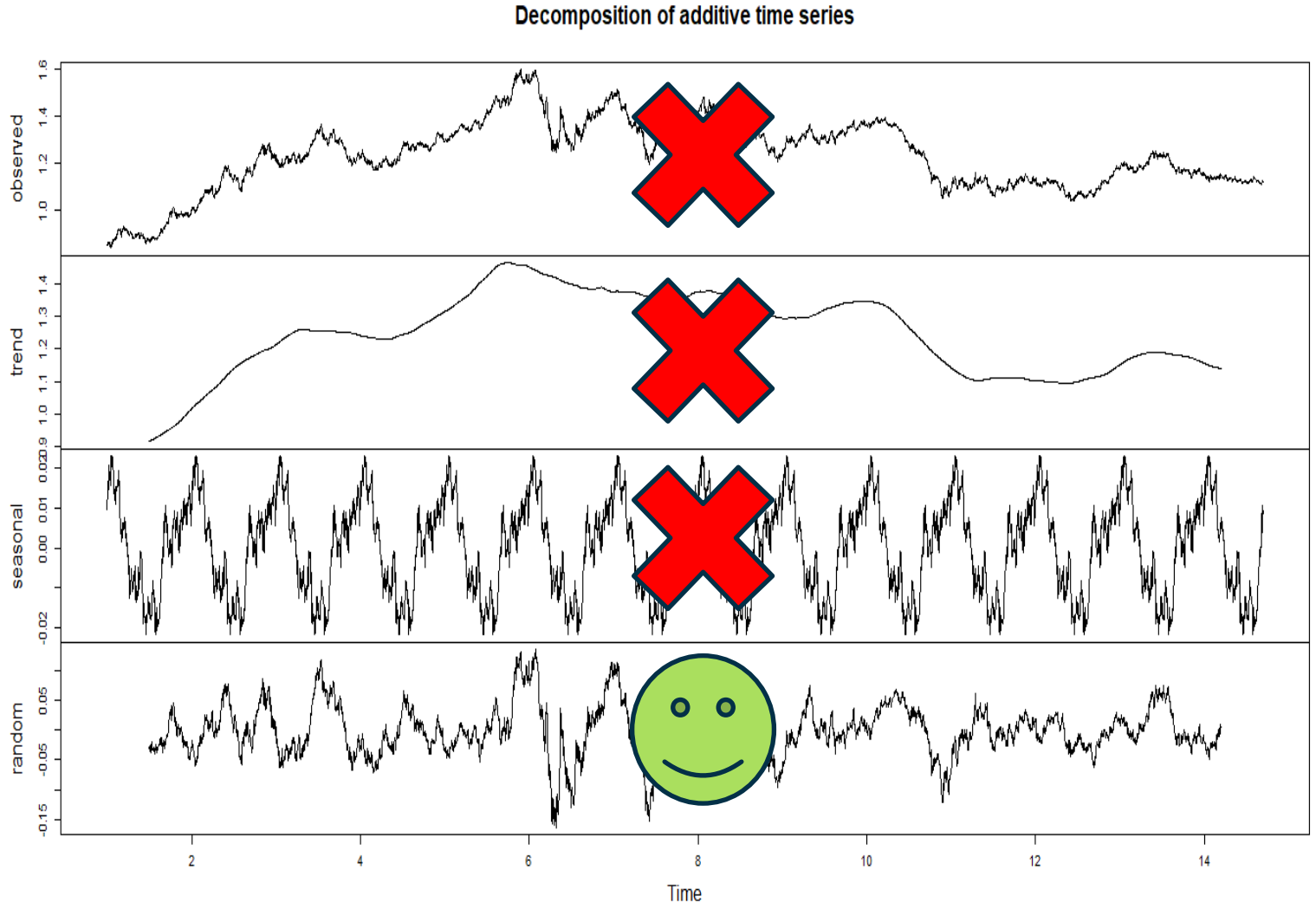
Ruido: Fluctuaciones que no ocurren siempre a la misma frecuencia.



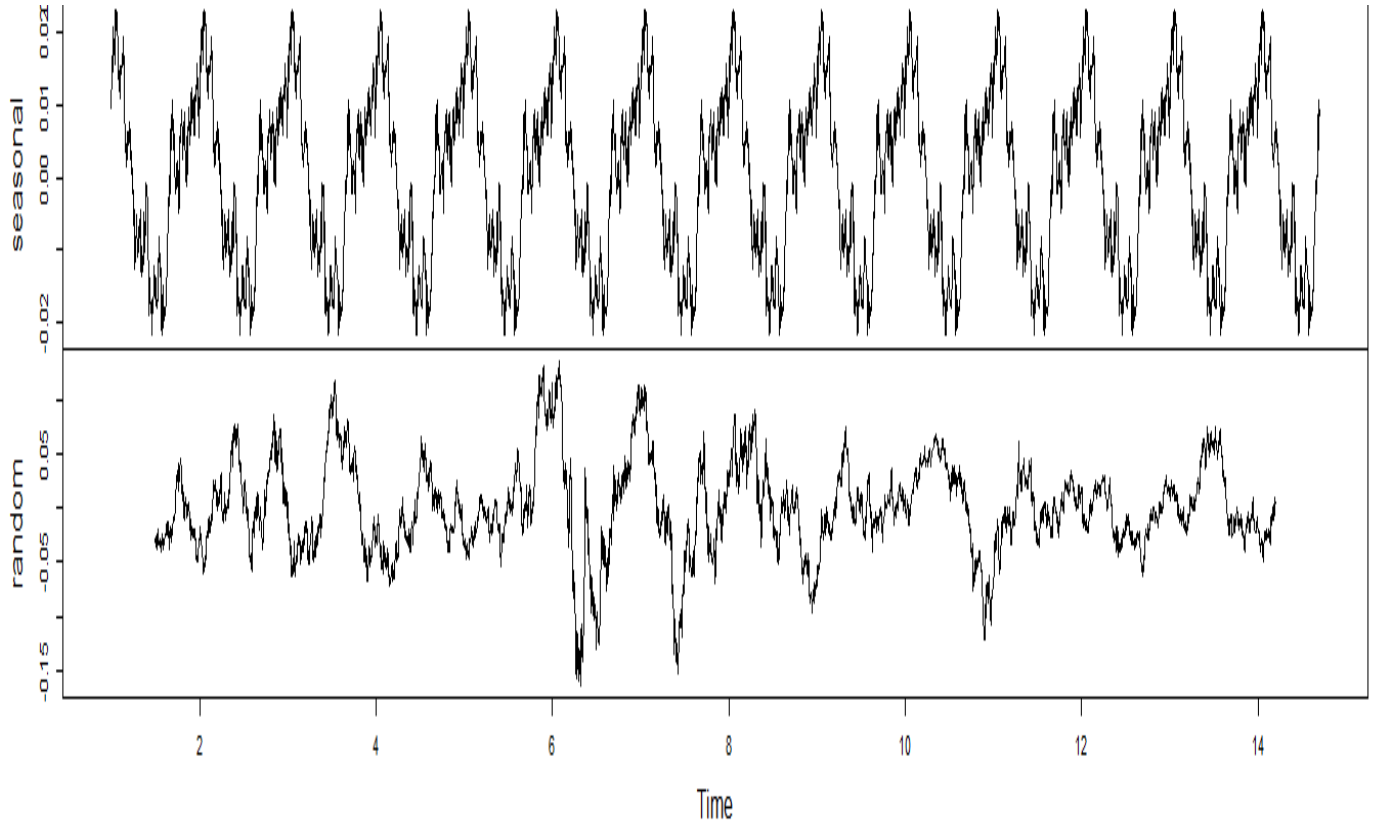
CONCEPTOS

Estacionaridad:

- La media es constante en el tiempo.
- La desviacion tpica es constante en el tiempo.
- La covariancia de yt con yt+k es constante en el tiempo.



MODELOS



- Variables dummies
- Análisis de Fourier
- Seasonal Auto Regresive Moving Average

- Auto Regresive Moving Average
- Otros…

EL MODELO CLÁSICO: ARIMA

$$Y_t = -(\Delta^d Y_t - Y_t) + \phi_0 + \sum_{i=1}^{p} \phi_i \Delta^d Y_{t-i} - \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t$$

DIFERENCIA | AUTOREGRESIVO | MEDIA MÓVIL

Se puede extender para incluir periodos estacionales

SARIMA

EL MODELO CLÁSICO: ARIMA

ARMA (p,q)

$$Y_t = \varepsilon_t + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

ARIMA (p, d, q)

$$Y_t = -(\Delta^d Y_t - Y_t) + \phi_0 + \sum_{i=1}^{p} \phi_i \Delta^d Y_{t-i} - \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon$$

Una diferencia se define como:

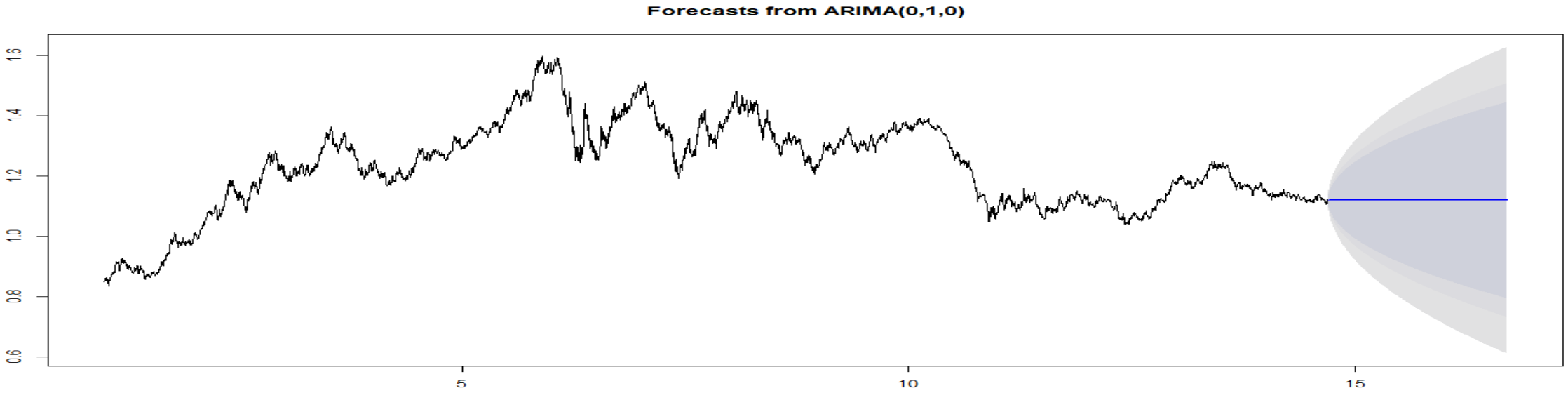
$$\Delta Y_t = Y_t - Y_{t-1}$$

Y una diferencia estacional:

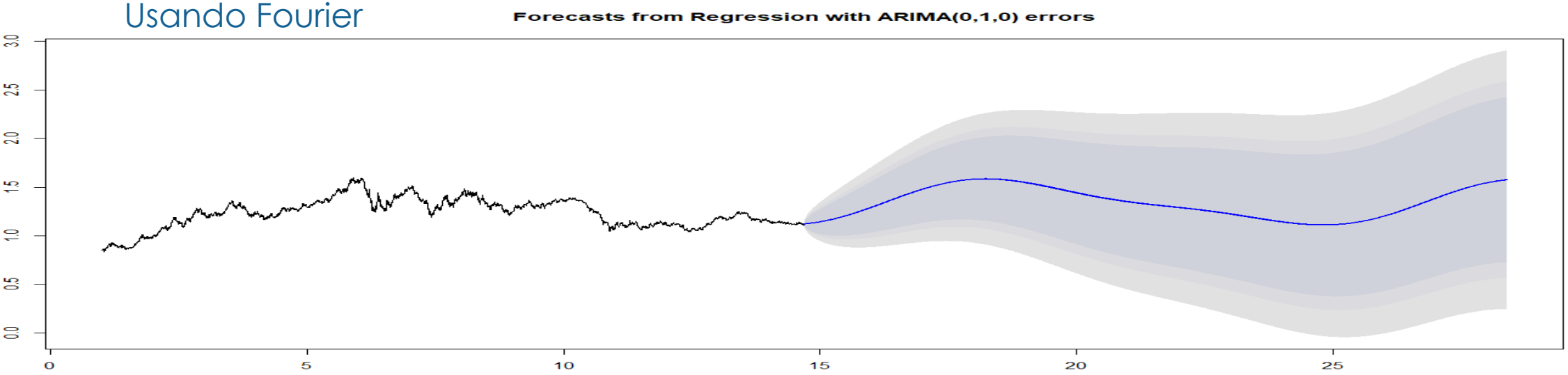
$$Y_t = Y_t - Y_{t-m} \quad \text{donde } m = \text{periodos}$$

Existe un modelo que también extiende los componentes auto-regresivas y media móvil a la estacionalidad SARIMA (p, d, q)x $(P, D, Q)_s$ donde P es el grado del modelo AR estacional, Q es el grado del modelo MA estacional, D es el grado del modelo estacional.

FORECAST



Usando Fourier



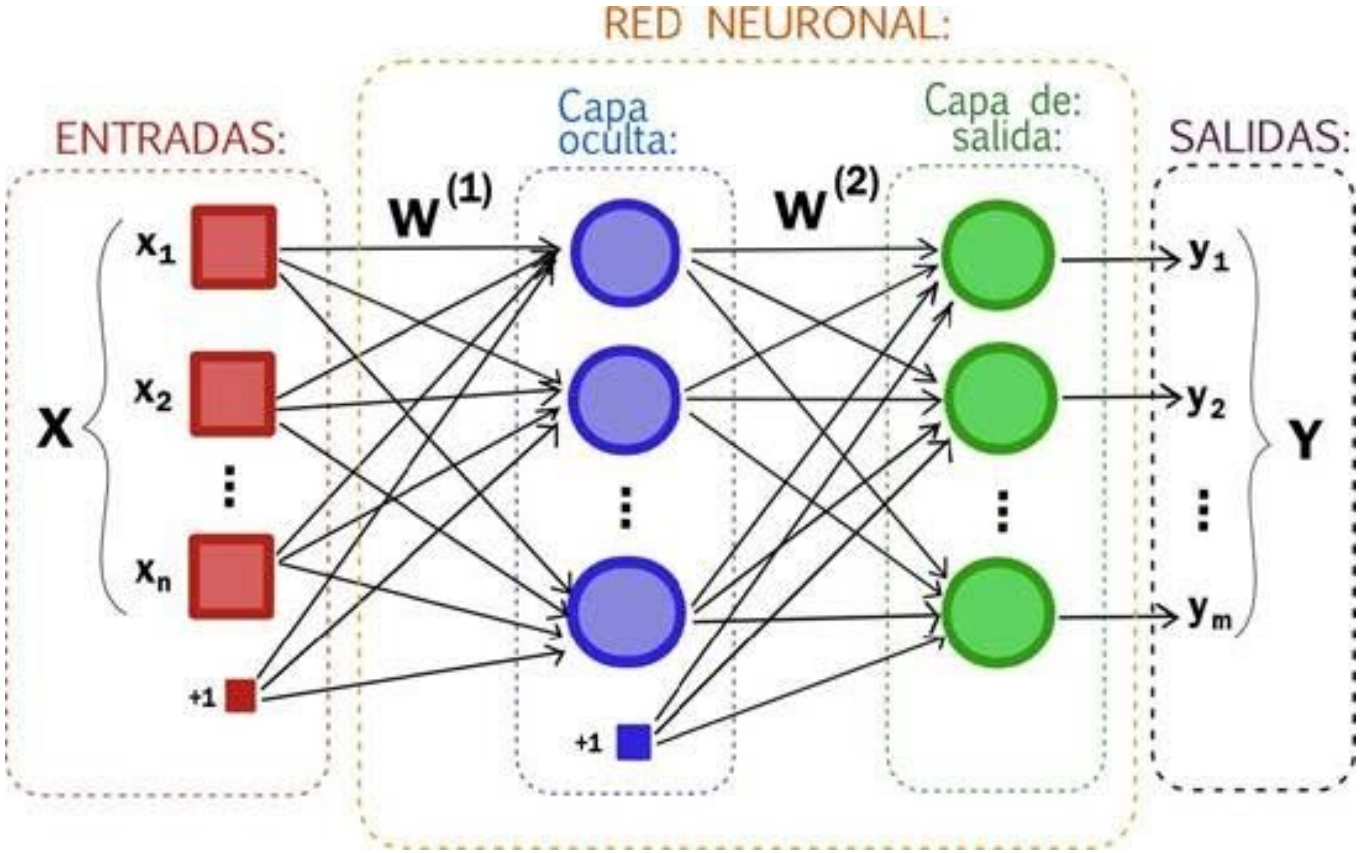
CASO:

Serie temporal
Sencilla

REDES NEURONALES

REDES NEURONALES ARTIFICIALES

- Una red neuronal es un modelo de inteligencia artificial que enseña a las máquinas a procesar datos de una forma inspirada en el sistema nervioso humano
- Nos permiten resolver problemas que no son linealmente separables gracias a las funciones de activación
- Son aproximadores universales, es decir, podemos construir una función que sea aproximación de cualquier función continua

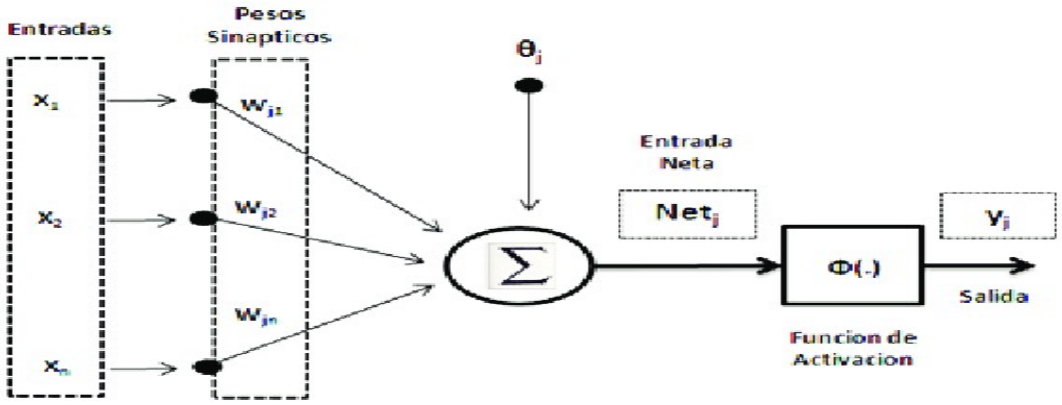


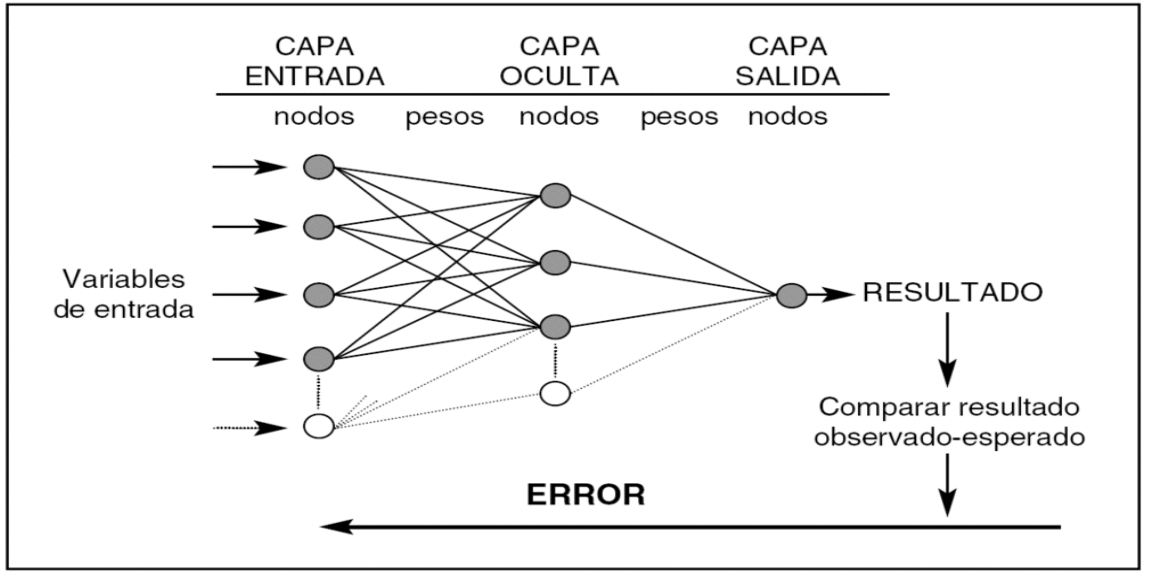
REDES NEURONALES ARTIFICIALES

- No deja de ser la propagación del valor calculado y retro-propagación del error a través de un grafo donde se ajustan iterativamente los pesos de cada conexión entre neuronas
- La información que pasa por cada neurona es a través de la función de activación (RELU, Lineal, TanH,…)
- El ajuste de pesos de cada capa en función del error y los pesos de la capa siguiente se basa en la regla de la cadena

https://www.researchgate.net/publication/307511858_Aplicacion_de_mapas_de_Kohonen_para_la_priorizacion_de_zonas_de_mercado_una_aproximacion_practica/figures?lo=1

http://www.medintensiva.org/es-redes-neuronales-artificiales-medicina-intensiva--articulo-13071859

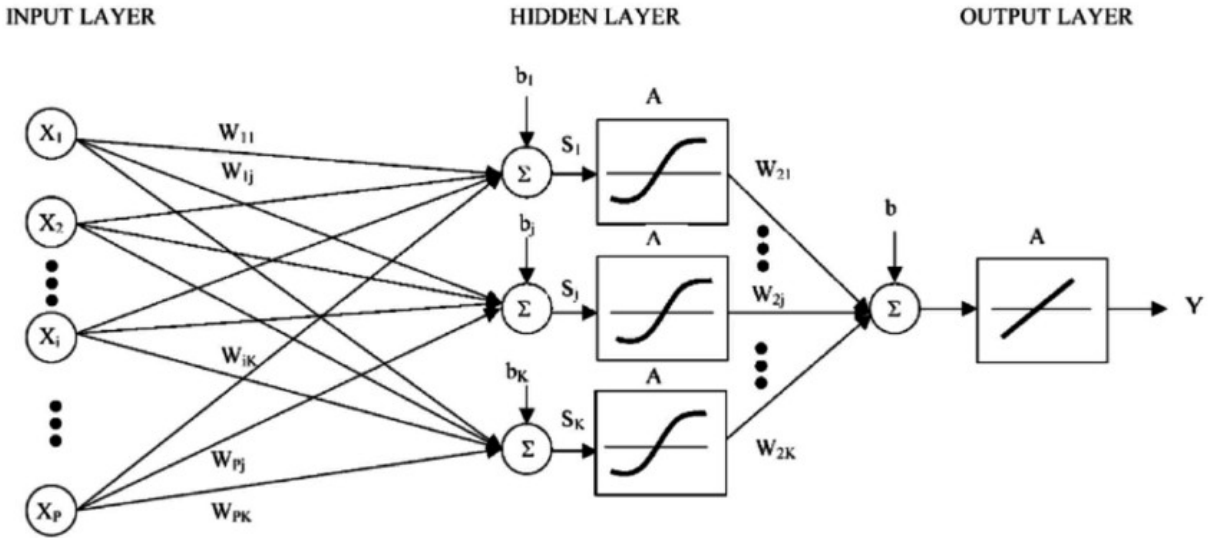




REDES NEURONALES ARTIFICIALES: FUNCIONES DE ACTIVACIÓN

- Identidad o lineal
- Tangente Hiperbólica
- Logística
- RELU
- Soft Max
- ….

- https://www.researchgate.net/publication/3978633_Artificial_neural_network_for_detecting_drowsiness_from_EEG_recordings/figures?lo=1&utm_source=google&utm_medium=organic

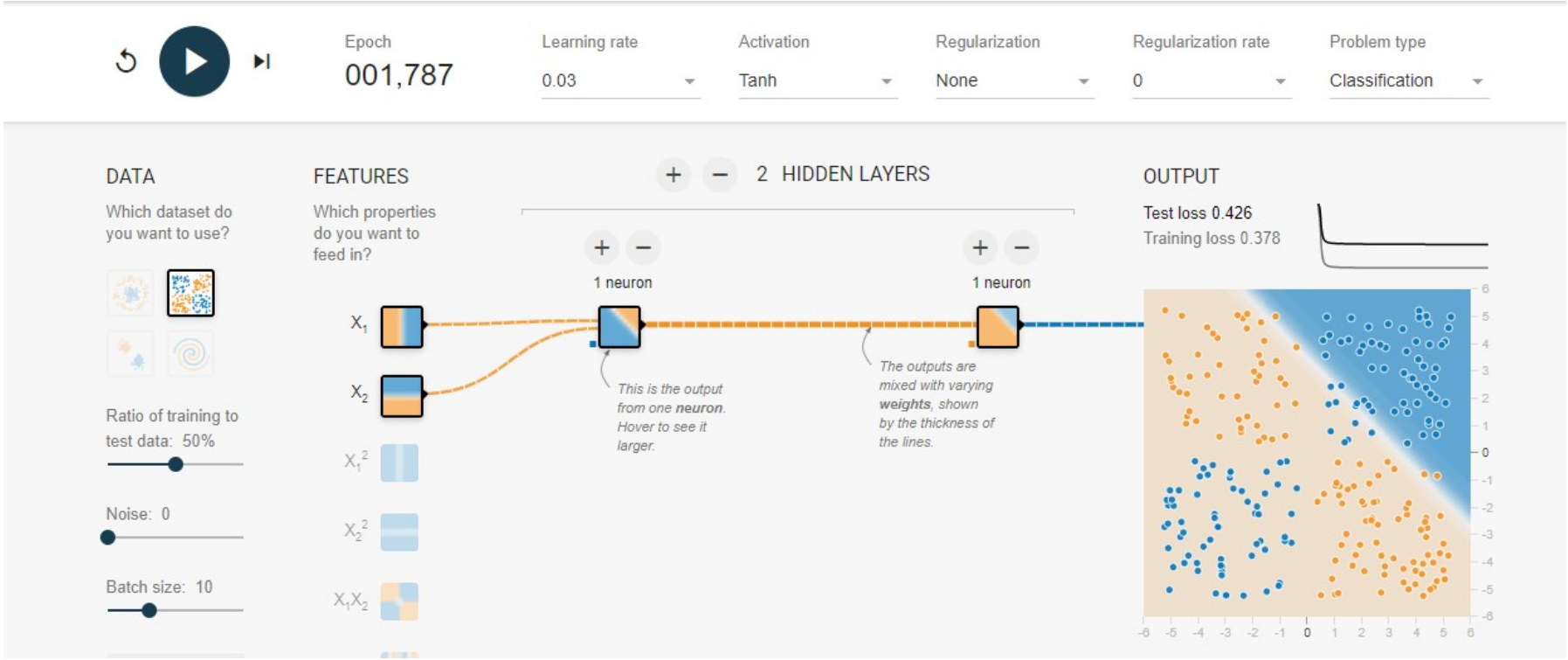


Activation Function	Equation	Example	1D Graph
Linear	$\phi(z) = z$	Adaline, linear regression	
Unit Step (Heaviside Function)	$\phi(z) = \begin{cases} 0 & z < 0 \\ 0.5 & z = 0 \\ 1 & z > 0 \end{cases}$	Perceptron variant	
Sign (signum)	$\phi(z) = \begin{cases} -1 & z < 0 \\ 0 & z = 0 \\ 1 & z > 0 \end{cases}$	Perceptron variant	
Piece-wise Linear	$\phi(z) = \begin{cases} 0 & z \leq -½ \\ z + ½ & -½ \leq z \leq ½ \\ 1 & z \geq ½ \end{cases}$	Support vector machine	
Logistic (sigmoid)	$\phi(z) = \frac{1}{1 + e^{-z}}$	Logistic regression, Multilayer NN	
Hyperbolic Tangent (tanh)	$\phi(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$	Multilayer NN, RNNs	
ReLU	$\phi(z) = \begin{cases} 0 & z < 0 \\ z & z > 0 \end{cases}$	Multilayer NN, CNNs	

REDES NEURONALES ARTIFICIALES

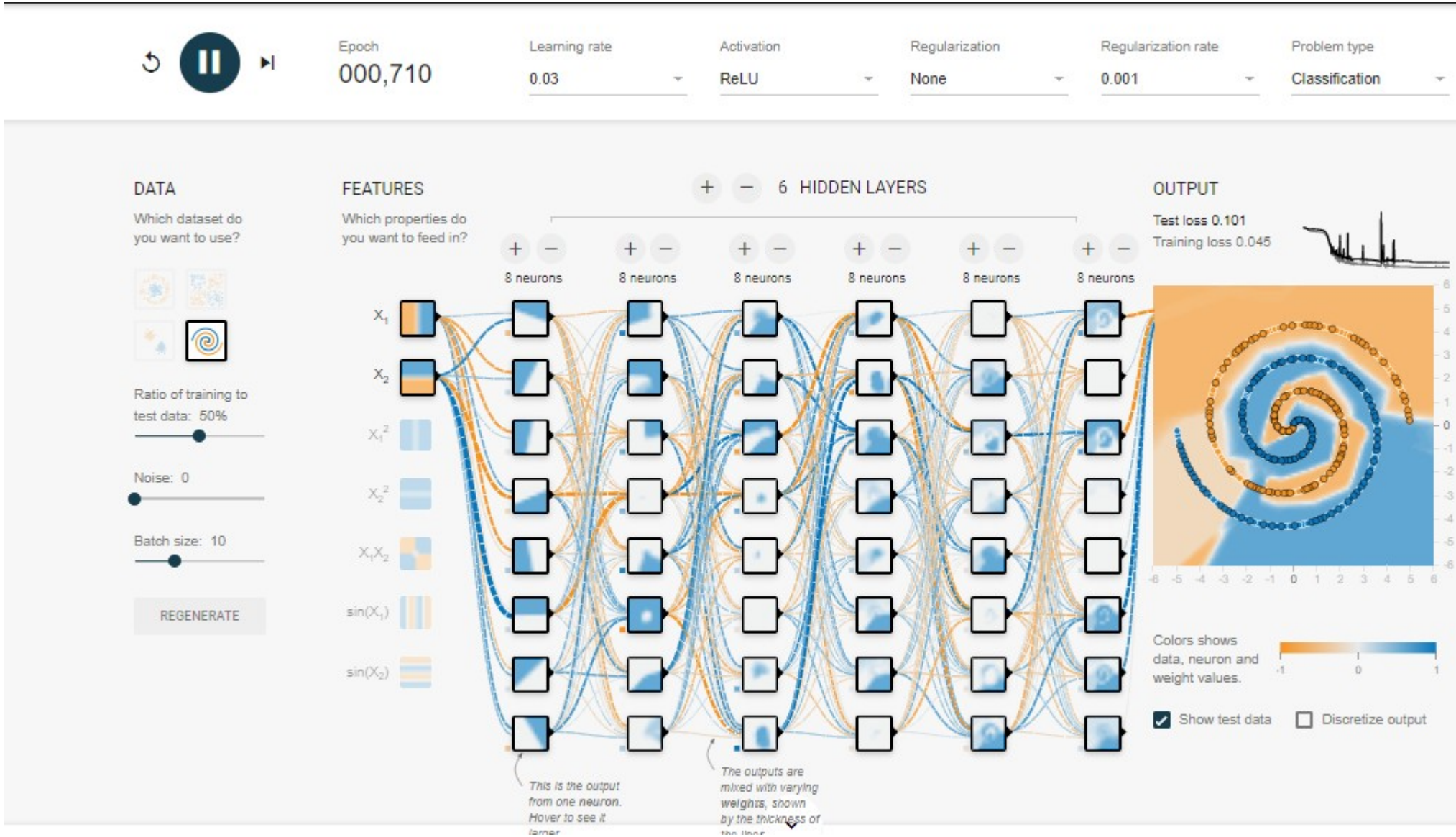
- Ideales para reconocimiento de datos no tabulares: sonidos, imágenes, textos…
- Funcionan muy bien con las series temporales
- También son muy útiles para modelos de clasificación con datos tabulares
- Cada vez hay arquitecturas más sofisticadas: transformers, attention, stable diffusion...

- Pero…
  - Sus hiperparámetros son muy difíciles de ajustar: arquitectura de la red, funciones de activación, regularizaciones, algoritmo y tasa de aprendizaje…
  - Generalmente con datos tabulares se consiguen resultados equivalentes o incluso superiores con modelos más sencillos como: GBM, RF…
  - Necesitan muchos datos para entrenar



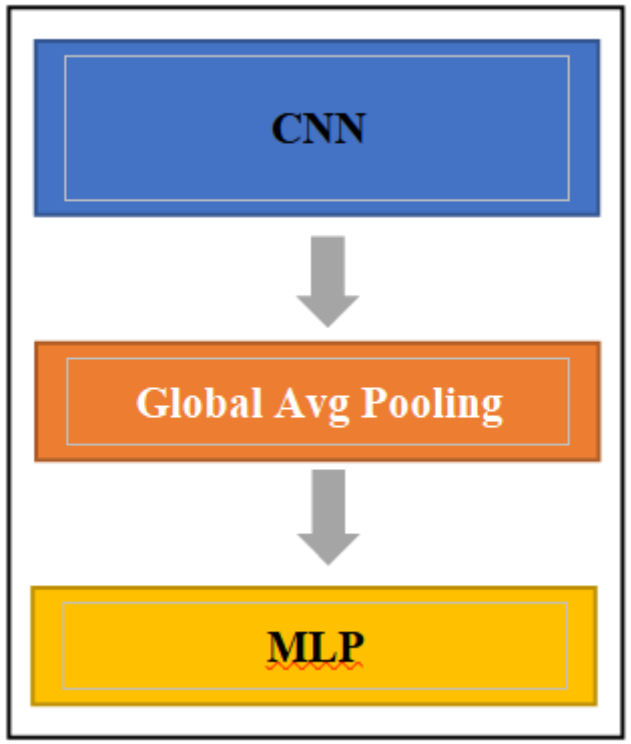
¿JUGAMOS UN POCO CON REDES NEURONALES?

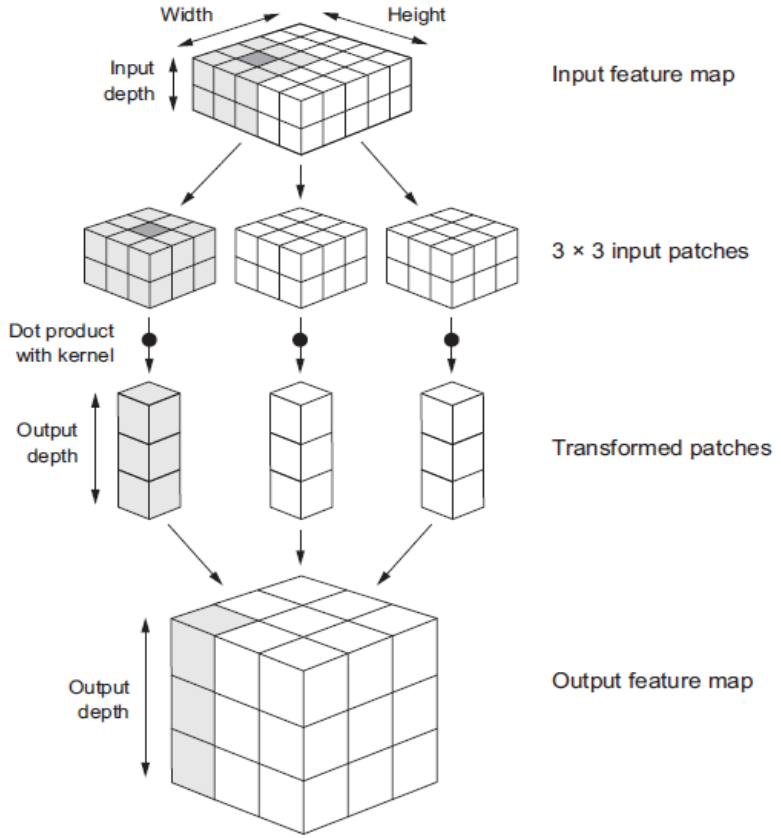
https://playground.tensorflow.org


Epoch
000,710
Learning rate
0.03
Activation
ReLU
Regularization
None
Regularization rate
0.001
Problem type
Classification
DATA
Which dataset do you want to use?
Ratio of training to test data: 50%
Noise: 0
Batch size: 10
REGENERATE
FEATURES
Which properties do you want to feed in?
6 HIDDEN LAYERS
8 neurons
8 neurons
8 neurons
8 neurons
8 neurons
8 neurons
X1
X2
X1²
X2²
X1X2
sin(X1)
sin(X2)
This is the output from one neuron. Hover to see it
The outputs are mixed with varying weights, shown by the thickness of
OUTPUT
Test loss 0.101
Training loss 0.045
Colors shows data, neuron and weight values.
Show test data
Discretize output

REDES NEURONALES CONVOLUCIONALES

En el análisis de imágenes, una convolución funciona moviendo ventanas de tamaño n × n × 3 (tensor de pesos) sobre la entrada en 3D (por ejemplo, los 3 canales RGB)





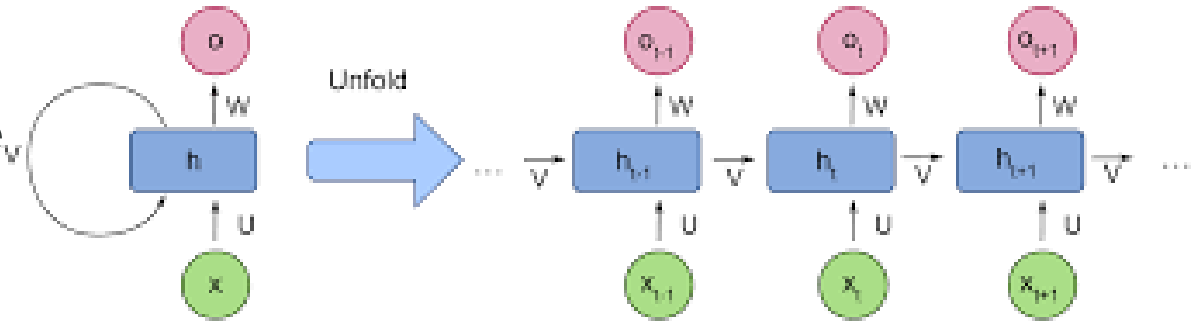
ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

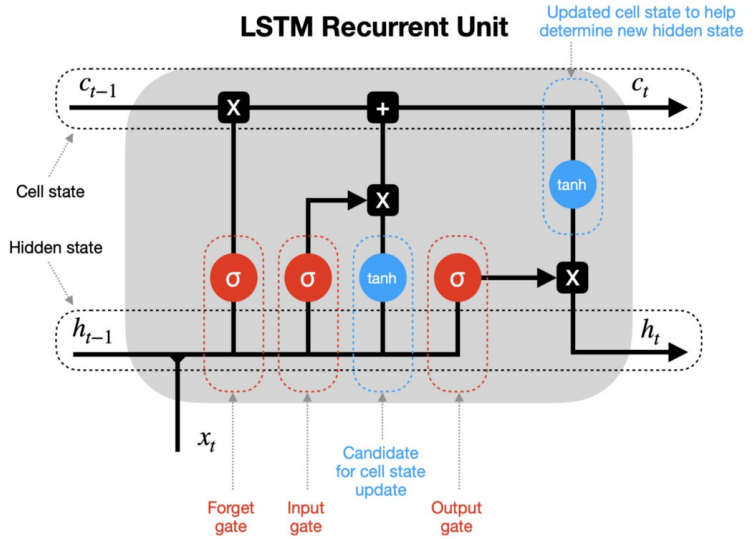
Network	A,A-LRN	B	C	D	E
Number of parameters	133	133	134	138	144

REDES NEURONALES RECURRENTES

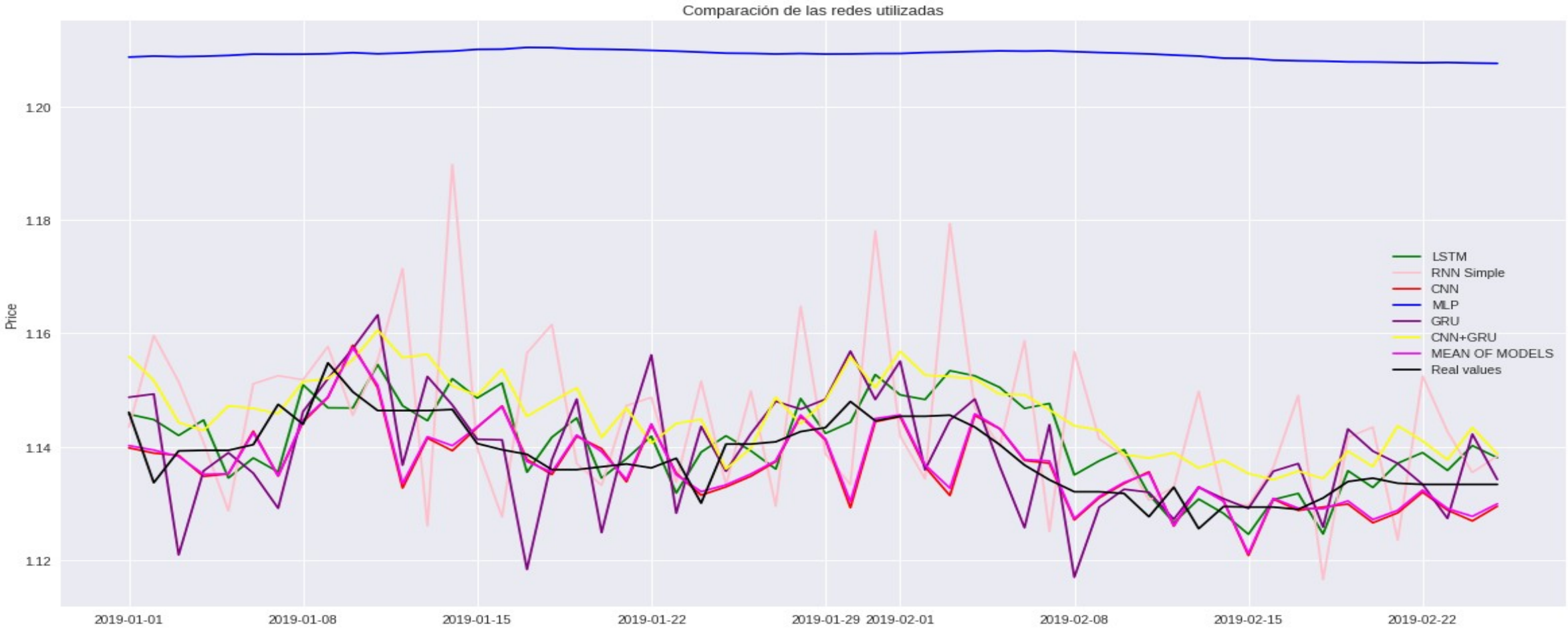
Es un tipo de red neuronal artificial que utiliza datos secuenciales o de series temporales.

Estos algoritmos de aprendizaje profundo se utilizan habitualmente para problemas ordinales o temporales, como la traducción de idiomas, el procesamiento del lenguaje natural (NLP), el reconocimiento del habla y el subtitulado de imágenes



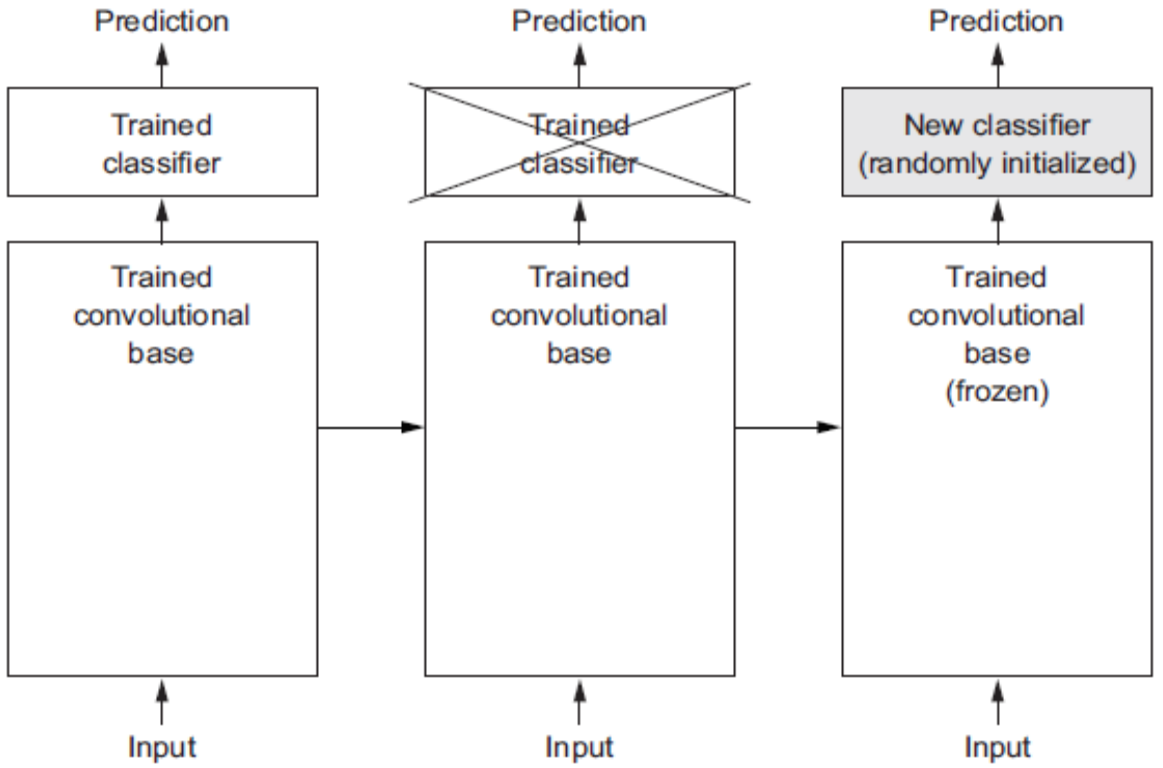


EL MODELO DE MACHINE LEARNING: REDES NEURONALES RECURRENTES



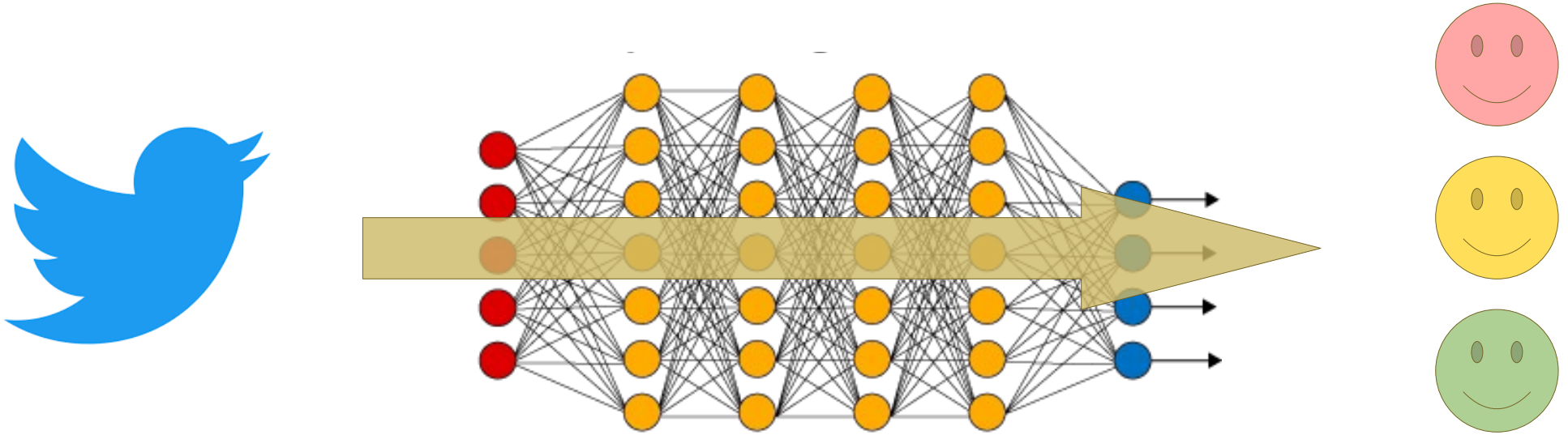
TRANSFER LEARNING

Se define como “reutilización de características extraídas para resolver un problema en la resolución de un problema diferente



ANÁLISIS DEL SENTIMIENTO

- Enfoque del procesamiento del lenguaje natural (NLP) que identifica el tono emocional detrás de un cuerpo de texto.
- Es una forma para determinar y categorizar las opiniones sobre un producto, marca, servicio o idea.
- Frecuentemente se resuelve usando redes neuronales recurrentes o arquitecturas basadas en éstas



CASO:

Ejemplos de
Redes Neuronales

MÁS ALLÁ DEL DEEP LEARNING

ALGORIMTOS BIOINSPIRADOS

- **Algoritmos Genéticos:** emulan la selección natural sobre un conjunto de individuos para buscar la mejor solución a un problema determinado. La "información genética" de cada individuo es una posible solución al problema; por analogía, hay un "gen" para cada variable o parámetro del problema sobre el que se desea ejecutar el proceso de optimización. Para emular la selección natural, se crea una población o conjunto de individuos y se le hace evolucionar de forma que los mejor adaptados, o sea, los que son mejor solución para el problema, se reproduzcan con mayor probabilidad y poco a poco vayan surgiendo individuos mejor adaptados al problema; en otras palabras, mejores soluciones.
- **Colonia de Hormigas:** algoritmo probabilístico que imita la habilidad de las hormigas para encontrar el camino más corto desde su hormiguero hasta una fuente de alimento. La forma en que las hormigas consiguen encontrar el camino más corto es la siguiente: en principio las hormigas vagabundean al azar alrededor de su hormiguero, y cuando encuentran alimento toman un poco y vuelven a su hormiguero dejando un rastro de feromonas. Si otras hormigas encuentran ese rastro, es probable que dejen de vagabundear al azar y lo sigan, ya que supuestamente conduce a una fuente de alimento. A su vez, las hormigas que vuelven con alimento dejan su propio rastro de feromona, reforzando así ese camino. Si hay varios caminos hacia una misma fuente de alimento el más corto acabará siendo el preferido por las hormigas, por la sencilla razón de que al ser más corto lo recorrerán más hormigas por unidad de tiempo, y por tanto la intensidad de su rastro de feromonas será mayor. Por otra parte, las feromonas se evaporan gradualmente, con lo que los caminos que no se utilizan van perdiendo atractivo. De esta forma, partiendo de una exploración aleatoria se consigue encontrar un camino óptimo o cercano al óptimo.
- **Sistemas Inmunes Artificiales:** algoritmo evolutivo que se basa en la dinámica del sistema inmune de los vertebrados para detectar y eliminar posibles amenazas para el organismo. Existen múltiples algoritmos.

  https://arxiv.org/ftp/arxiv/papers/1711/1711.07821.pdf

  Inteligencia artificial avanzada (Benitez, Escudero, Kanaan; 2013)