



BSG Institute

conocimiento para crecer

PROGRAMA CERTIFIED BIG DATA APPLICATIONS DEVELOPER

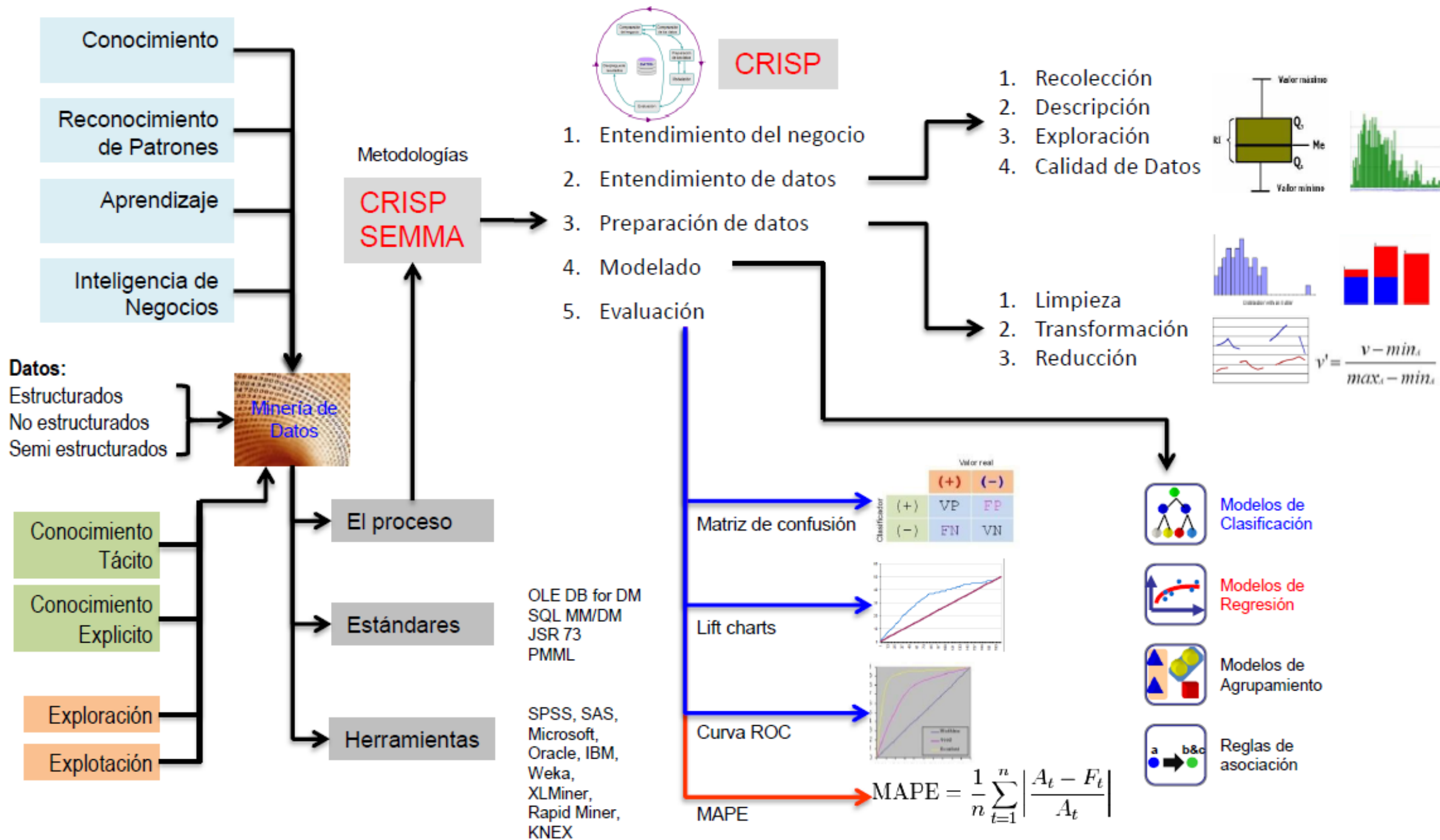


Desarrollo de Aplicaciones Big Data con Python

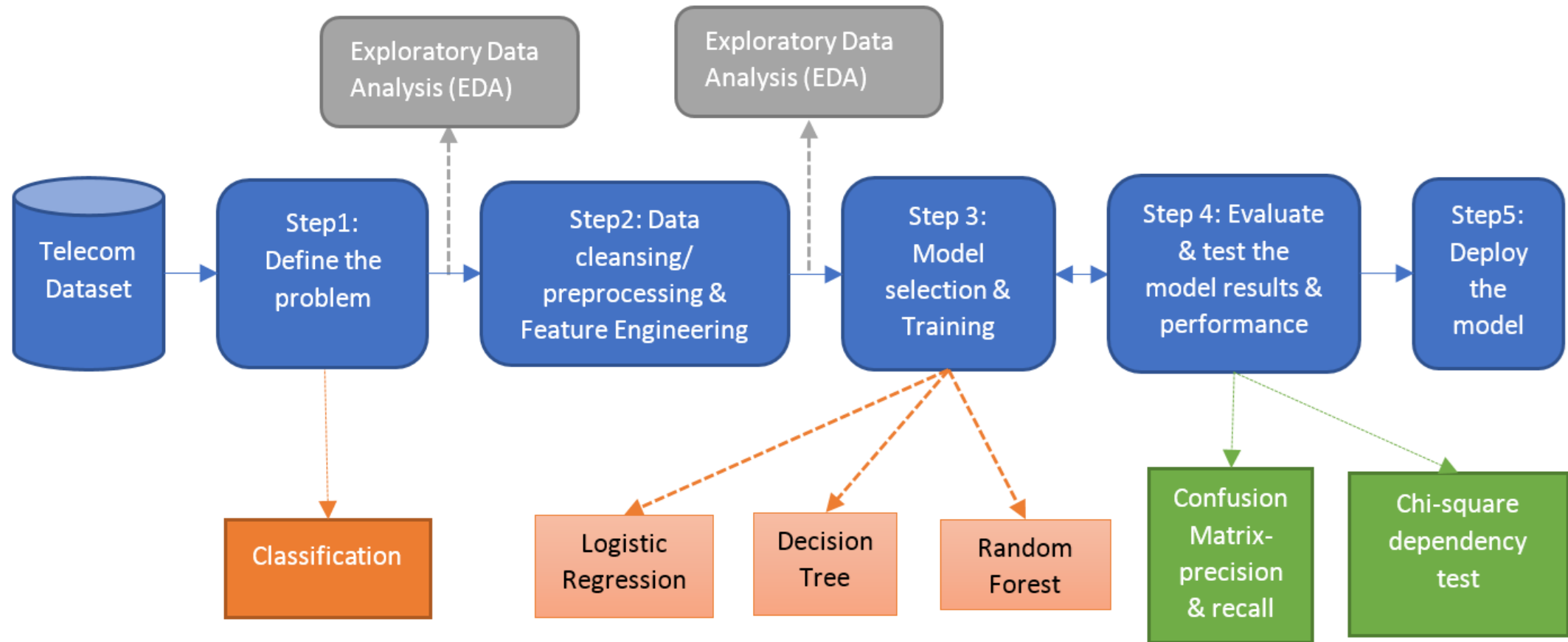
COMO FUNCIONAN LOS MODELOS ?

Prof. Daniel Alfredo Chávez Gallo

dacg160381@Hotmail.com



Modelos



Modelos



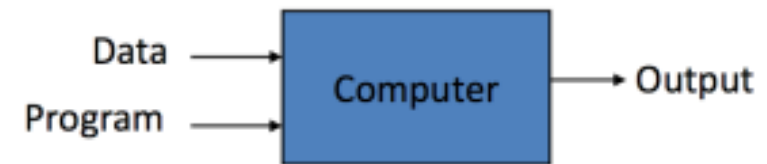
Modelos y algoritmos

- Un **modelo** es, en general, una **función o una estructura de datos**
 - Es una **representación específica hecha a partir de los datos**
 - Árbol de decisión, Red Neuronal, Conjunto de coeficientes de una recta de regresión...
- Un **algoritmo (o método)** es una secuencia de pasos con un fin.
 - Un algoritmo de aprendizaje se encarga de que el modelo aprenda de los datos (se ajuste a los datos)
 - C4.5, Regresión lineal por mínimos cuadrados, “Probar coeficientes hasta que esto funcione”...

Modelos y métodos

- **Modelo (entrenado):** resultado de la aplicación de un método sobre unos datos específicos
- El **método (de aprendizaje)** se encargará de dar valores a las variables que forman el modelo
 - Árbol de decisión
 - Red Neuronal
 - Conjunto de coeficientes

Traditional Programming

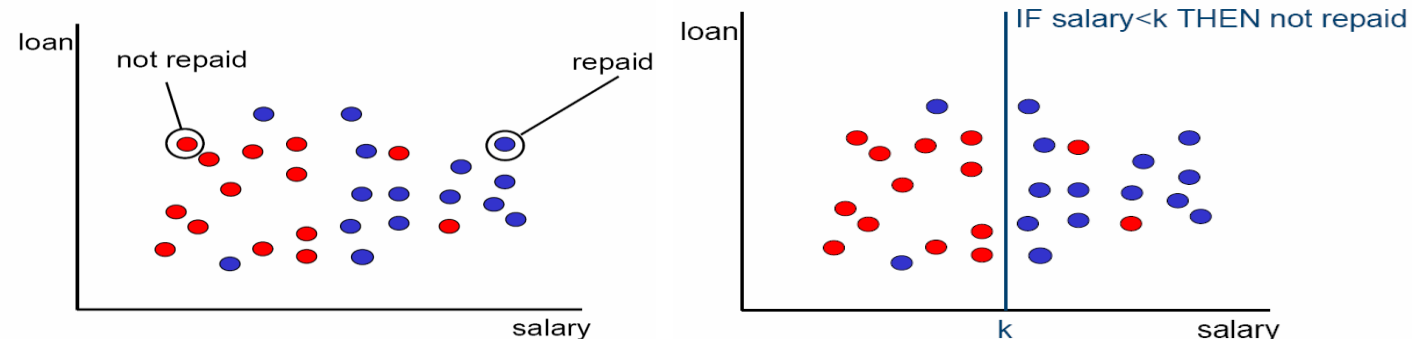


Machine Learning

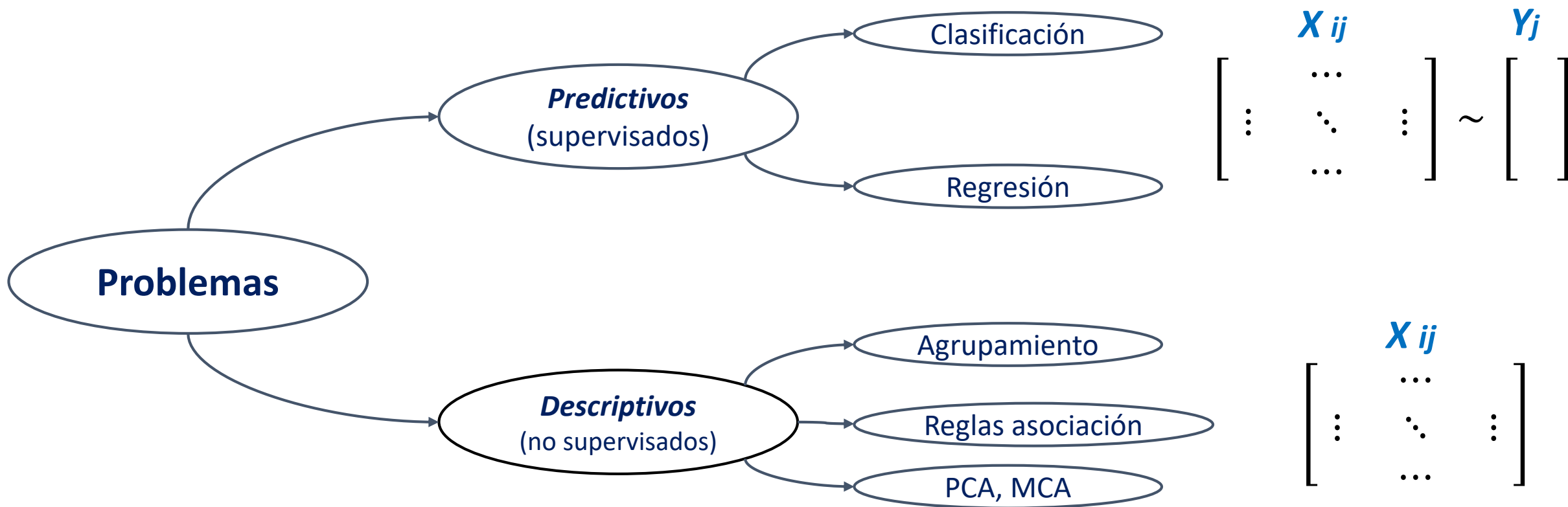


Modelos

- **Modelo = Método(Datos)**
- **Riesgo de un crédito**
 - Modelo \rightarrow Una regla: *"If salary < k THEN not repaid"*
 - Método \rightarrow *"Probar 1000 valores aleatorios de k y asignarlo al que mejor resultado dé"*
 - Modelo (entrenado) \rightarrow *"If salary < k THEN not repaid"*
 - Siendo k el valor obtenido por el método



Modelos



Modelos supervisados

$$\begin{bmatrix} \square & \cdots & \square \\ \vdots & \ddots & \vdots \\ \square & \cdots & \square \end{bmatrix} \sim \begin{bmatrix} \square \\ \square \\ \square \end{bmatrix}$$

X_{ij} Y_j

- Se desea crear un modelo que relacione las **variables** y las **respuestas** con el objetivo de predecir las respuestas de futuras observaciones.
- Variable respuesta **Y** (Dependent variable, objective, response, target, class)
- Variables predictoras llamadas **X** (inputs, regressors, covariates, features, independent variables).
- Tenemos datos de entrenamiento (training data) que son observaciones (ejemplos, instancias) de estas medidas.

Modelos supervisados

$$\begin{bmatrix} \square & \dots & \square \\ \vdots & \ddots & \vdots \\ \square & \dots & \square \end{bmatrix} \sim \begin{bmatrix} \square \\ \square \\ \square \end{bmatrix}$$

X_{ij} Y_j

- **Clasificación:** Cuando la variable a predecir es una categoría.
 - Binaria: {Sí, No}, {Azul, Rojo}, {Fuga, No Fuga}...
 - Múltiple: {Comprará Producto1, Producto2...}...
 - Ordenada: {Riesgo Bajo, Medio, Alto}...
- **Regresión:** Cuando la variable a predecir es una cantidad
 - Precio, cantidad, tiempo,...

Modelos no supervisados

$$\begin{matrix} & X_{ij} \\ \begin{bmatrix} \square & \dots & \square \\ \vdots & \ddots & \vdots \\ \square & \dots & \square \end{bmatrix} \end{matrix}$$

- El aprendizaje no supervisado **busca patrones en los datos**.
- **Se utiliza** cuando se desconoce la estructura de los datos.
 - Por ejemplo, cuando se desconoce cuántos grupos de usuarios similares existen.
- La clave, consisten en buscar buenas *variables* capaces de distinguir entre las diferentes instancias.

Modelos no supervisados

$$\begin{bmatrix} \boxed{} & \dots & \boxed{} \\ \vdots & \ddots & \vdots \\ \boxed{} & \dots & \boxed{} \end{bmatrix}$$

X_{ij}

- **Agrupamiento - Clustering:** Buscan encontrar grupos dentro de los datos de elementos similares
 - ✓ Clientes con hábitos de compra similares
 - ✓ Productos vendidos en fechas similares
- **Asociación:** Buscan reglas que describen la mayor parte posible de los datos de los que se disponen
 - ✓ Productos que se compran juntos
- **Reducción de Dimensionalidad - Análisis de Correlación:** Explicar la misma información contenida en los datos originales con un menor número de variables buscadas de manera óptima.

Tipos de modelos

- **Paramétricos**

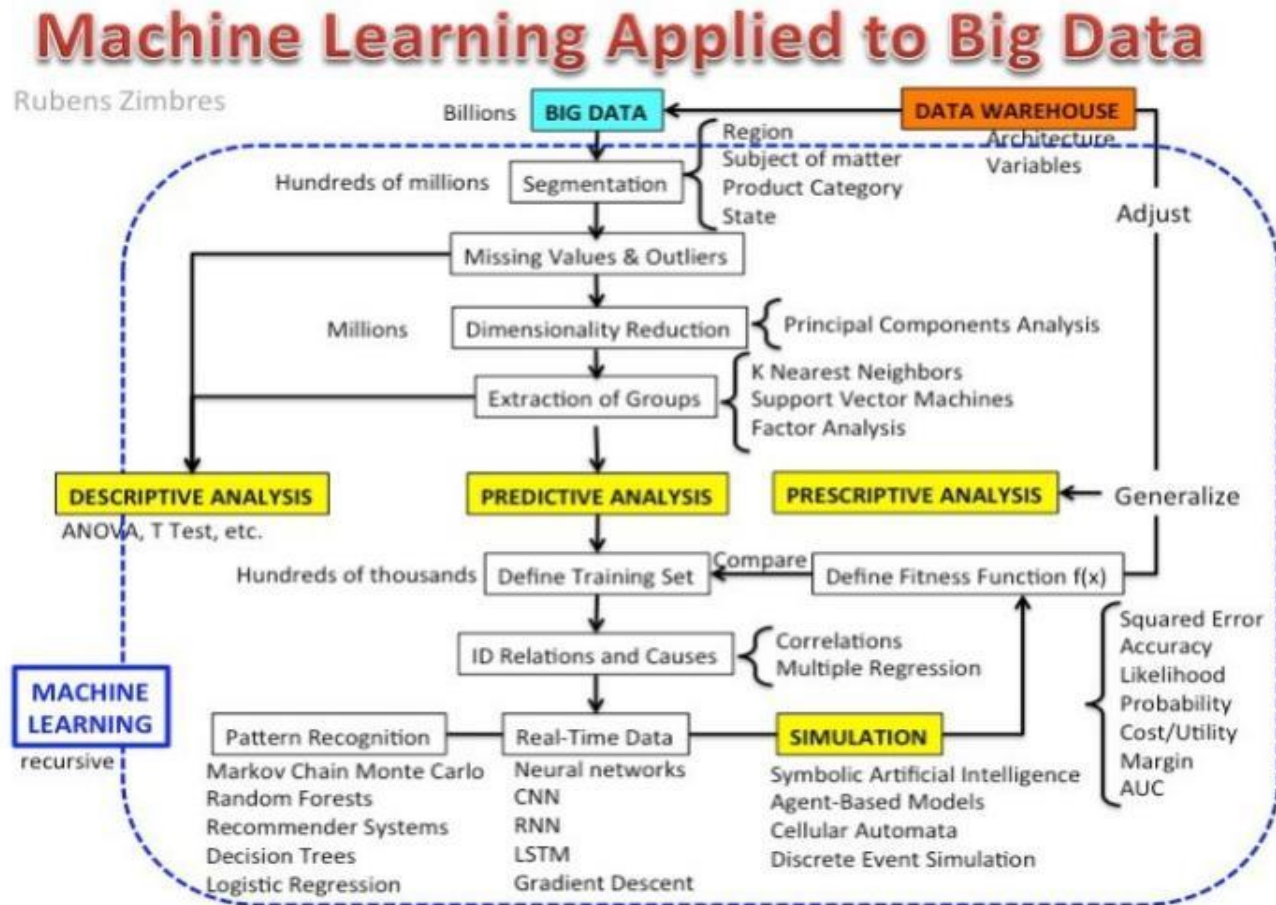
- Hacen asunciones acerca de la “forma” de los datos
 - Asumen una serie de parámetros a modificar en función de los datos de entrada
- El número de parámetros es independiente de la cantidad de datos
 - Regresión Lineal, Perceptrón,...

- **No paramétricos**

- Tienen la “libertad” de aprender sea cual sea la forma y cantidad de datos mediante el aprendizaje
 - Árboles de decisión,...

- ¿Ventajas? ¿Inconvenientes? ¿Similaridades? ¿Diferencias?

Modelos de machine learning



Modelizar

- **Modelizar el problema de Machine Learning**

- Ejemplos

- Dado el perfil de un cliente y su actividad pasada, ¿en qué productos financieros estaría más interesado?
 - Dados los resultados de un test clínico, ¿sufre el paciente de cáncer?
 - Dada una imagen de una resonancia magnética, ¿hay un tumor?
 - Vista la actividad pasada de una tarjeta de crédito, ¿es esta operación fraudulenta?
 - Si la pregunta es sobre la predicción de una cantidad, generalmente real, estamos en un problema de regresión.
 - Dada la descripción de un piso, ¿cuál es el valor de un piso?
 - Con mi historial de notas en la universidad, ¿qué nota sacaré en el próximo examen?
 - Con mi historial de uso de aplicaciones en mi móvil, ¿durante cuánto tiempo voy a utilizar la última aplicación que me he descargado?

Modelizar

- **Modelizar el problema de Machine Learning**

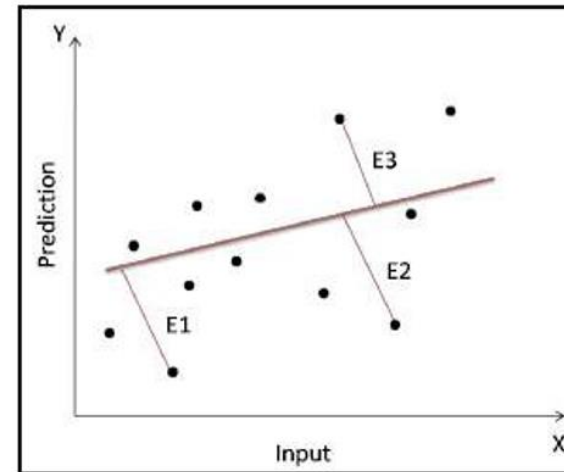
- Muchos de estos problemas se pueden resolver tanto por problemas de clasificación como de regresión
- Muchos algoritmos de clasificación funcionan con umbrales
- Importante cómo formulamos la pregunta porque la solución será distinta
- Utilizar siempre la solución más sencilla
 - Ejercicio: ¿qué tipo de problema es la predicción del tiempo?

Modelos, como funciona el aprendizaje?

- El procedimiento de dividir los datos en training y test se conoce como **hold-out**
 - El conjunto de **training** se utiliza para crear el modelo que es evaluado con el conjunto de **test**. Normalmente se utiliza un **80% - 20%**. Para asegurarse de que no existen diferencias sistemáticas se obtienen las instancias de forma aleatoria en cada uno de los 2 grupos.
 - Para que este método sea riguroso no se pueden utilizar instancias del conjunto de test para crear el clasificador. Es común caer en el error de utilizar el conjunto de test para elegir el mejor clasificador. Para ello se suele utilizar otro conjunto: **validation set**.

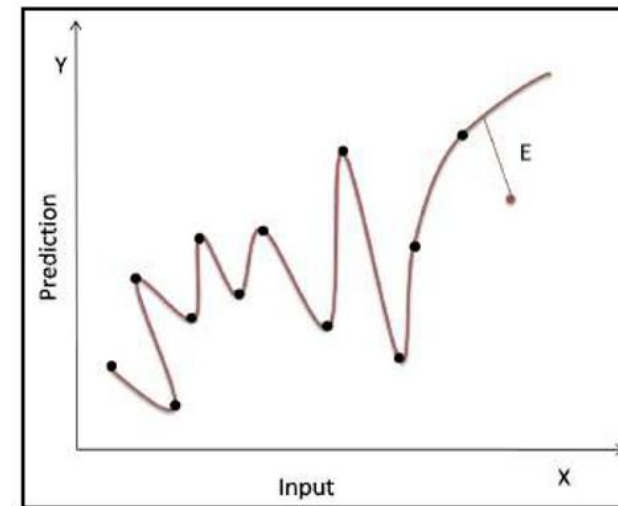
Sesgo

- La diferencia entre el valor que se predice y el valor actual.
- Un error de sesgo grande querrá decir que se tiene un modelo que no está rindiendo al nivel que se esperaba.
- El modelo está omitiendo tendencias importantes.
- A medida que aumenta el número de muestras de entrenamiento, los errores tienden al mismo valor



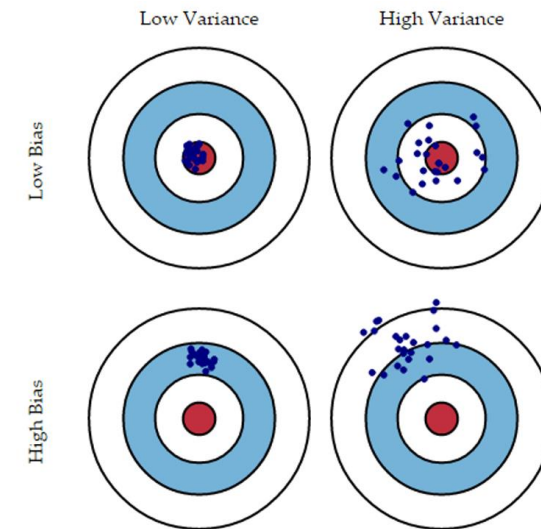
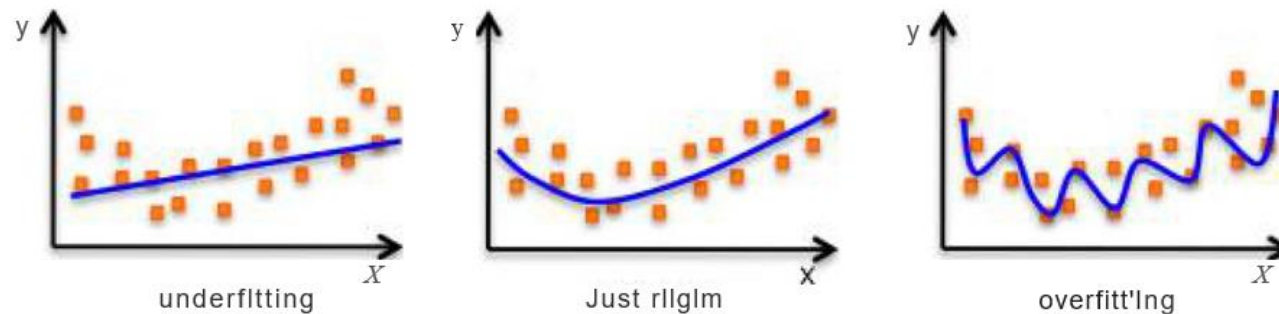
Varianza

- La diferencia entre las predicciones hechas sobre la misma observación.
- Un modelo con alta variación tendrá sobre ajuste sobre la población de entrenamiento y tendrá un rendimiento malo en cualquier observación más allá del entrenamiento.
- Es importante controlar la varianza desde el análisis de los predictores, en la etapa de la preparación de los datos.



Overfitting

- A medida que aumenta la complejidad se reduce el error de entrenamiento, pero a partir de un cierto nivel de complejidad aumenta el error en la muestra de validación. **Consejo:** Una vez elegido el modelo siempre podemos reentrenarlo con la muestra de training+validation.
- ***“El modelo se ha aprendido bien los datos de entrenamiento, pero no generaliza bien”***



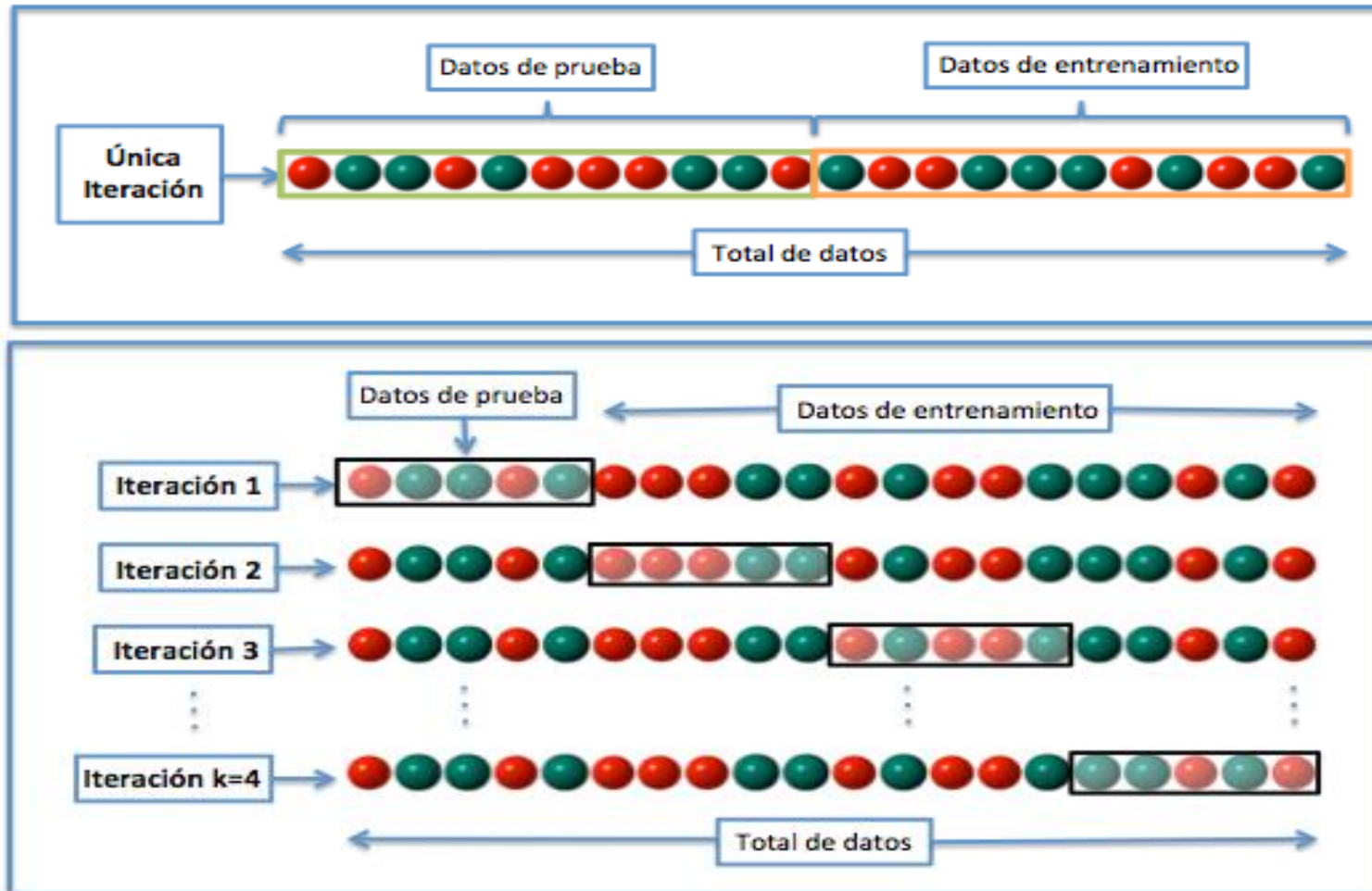
Modelos, que hacer cuando?

- Nuestro algoritmo tiene gran **sesgo**:
 - Meter más variables (sobre todo si podemos generar nuevas variables discriminantes)
 - Usar un modelo más sofisticado
- Si tiene gran **varianza**:
 - Usar menos variables
 - Usar un modelo más simple
 - Usar más datos de entrenamiento
 - Técnicas de remuestreo

Modelos, como funciona el aprendizaje?

- La repetición de la **técnica de Hold-Out es la base para el Cross-Validation**
 - Se trata del estándar de la industria para estimar el rendimiento de los modelos.
 - K-Fold divide de forma aleatoria los datos en K particiones separadas llamadas folds.
 - Aunque K puede ser cualquier número, lo habitual es utilizar $K=5$ ó $k=10$. ¿Por que 10? Porque la evidencia empírica indica que hay poco beneficio en utilizar más de 10 folds.
 - Para cada uno de los 10 folds (que comprenden un 10% del total de los datos) se crea un modelo en los restantes 9 folds (90% de los datos) y se evalúa en ese 10%. El proceso se realiza 10 veces y la media del rendimiento es reportada.

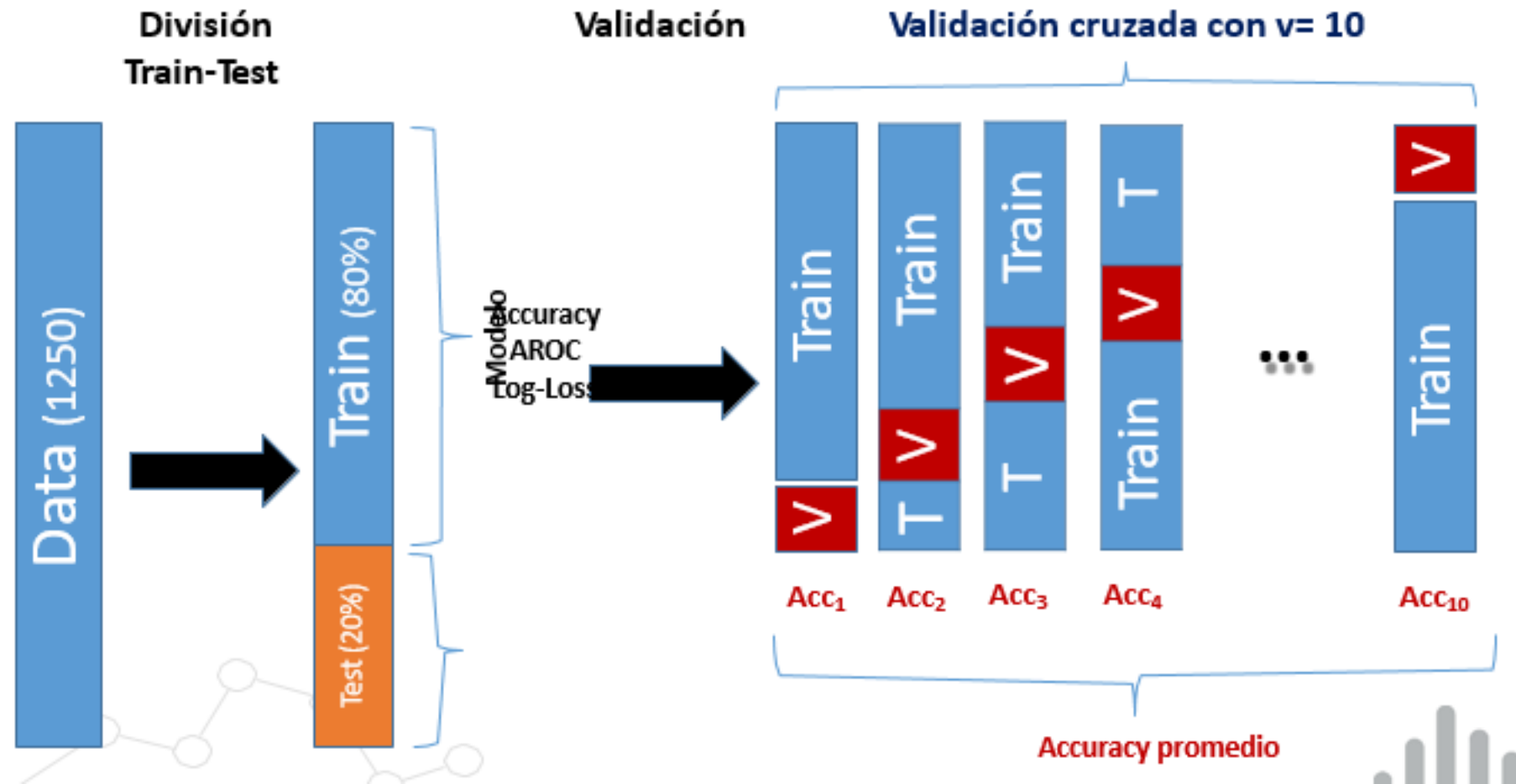
Modelos, como funciona el aprendizaje?



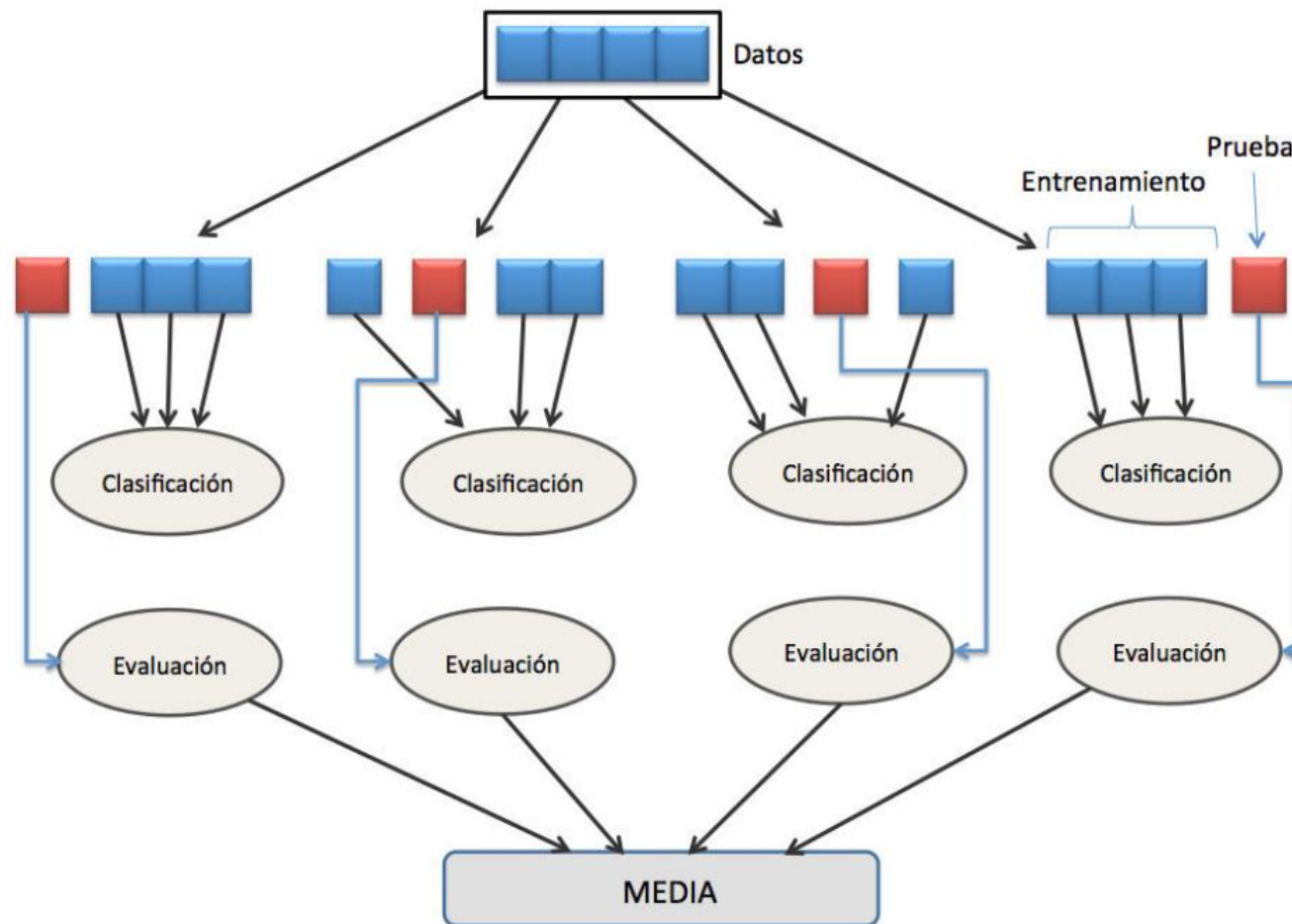
División sin
K - fold

**Validación
cruzada de
K iteraciones**

Modelos, como funciona el aprendizaje?



Modelos, como funciona el aprendizaje?



Matriz de confusión

$$error = \frac{FP + FN}{total}$$

$$exito = \frac{VP + VN}{total}$$

$$sensibilidad = \frac{VP}{VP + FN}$$

$$especificidad = \frac{VN}{VN + FP}$$

Caso mal clasificados

Caso bien clasificados

		Valor real	
		(+)	(-)
Clasificador	(+)	VP	FP
	(-)	FN	VN

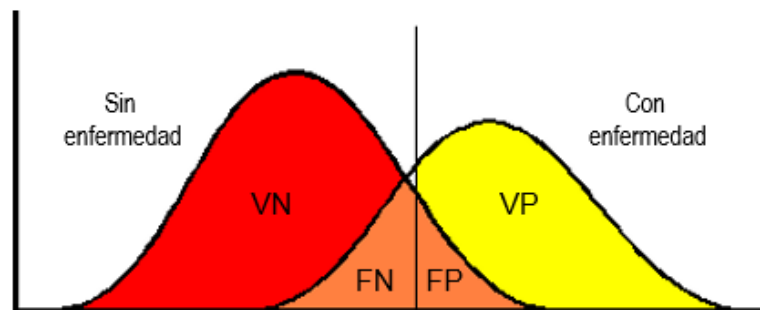
Probabilidad de clasificar correctamente a un individuo con el valor de interés (+)

Probabilidad de clasificar correctamente a un individuo sin el valor de interés (-)

FP = error de tipo I (α)
Muy costoso

FN = error de tipo II (β)
 β <5%, 20%>

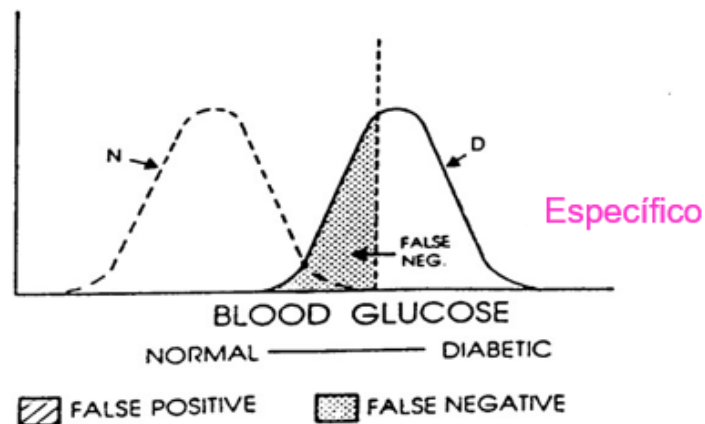
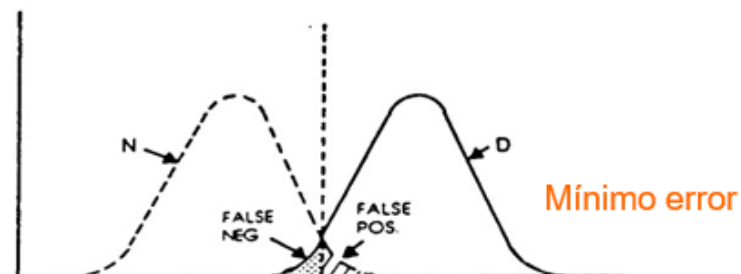
Matriz de confusión



$$\text{especificidad} = \frac{VN}{VN + FP}$$

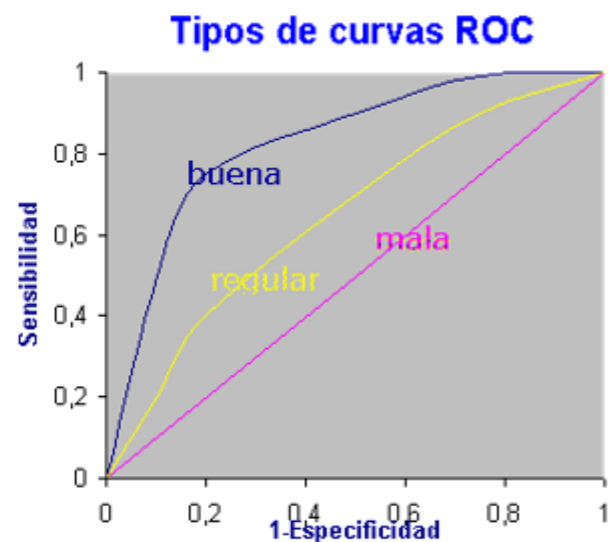
$$\text{sensibilidad} = \frac{VP}{VP + FN}$$

	(+)	(-)
(+)	VP	FP
(-)	FN	VN

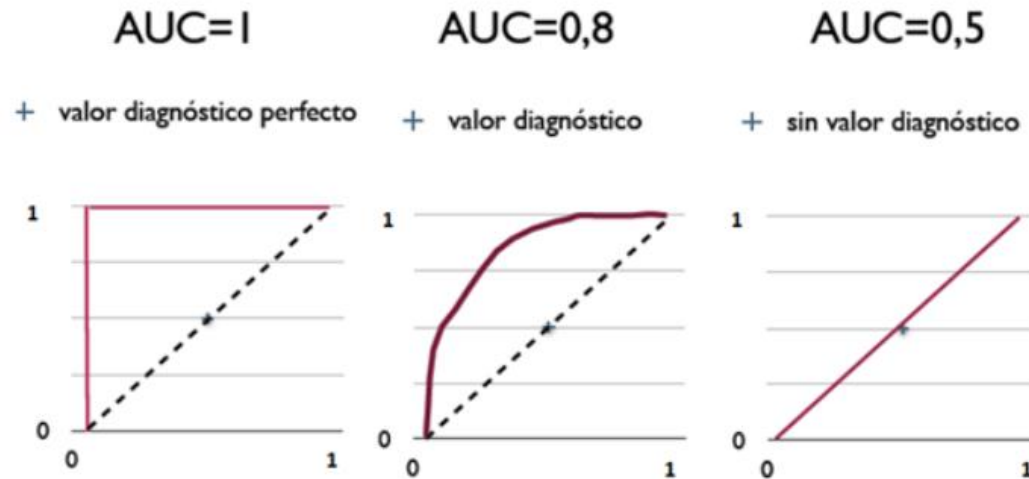


Matriz de confusión

- El desempeño de un modelo se mide en términos de precisión diagnóstica (Sensibilidad + Especificidad)
- La Curva ROC proporcionan un índice de la capacidad de un modelo para discriminar entre estados alternativos de la clase.
- Es útil para comparar modelos y seleccionar umbrales de decisión (puntos de corte entre (+) y (-))

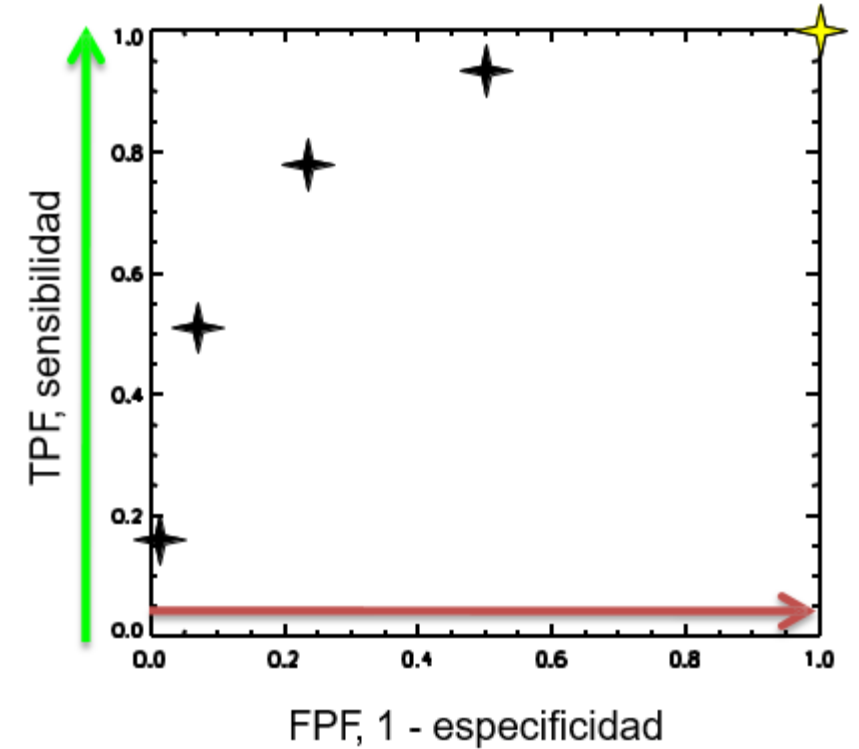


Coeficiente de Gini



[0.5]: ES COMO LANZAR UNA MONEDA.

[0.5, 0.6):	Test malo.
[0.6, 0.75):	Test regular.
[0.75, 0.9):	Test bueno.
[0.9, 0.97):	Test muy bueno.
[0.97, 1):	Test excelente



$$\text{Gini} = 2 * \text{AUC(ROC)} - 1$$

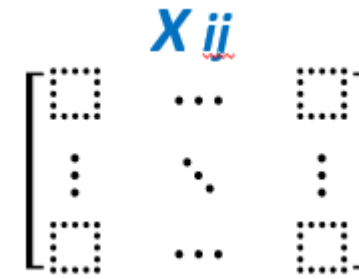
Desarrollo de Aplicaciones Big Data con Python

Algoritmos de machine learning

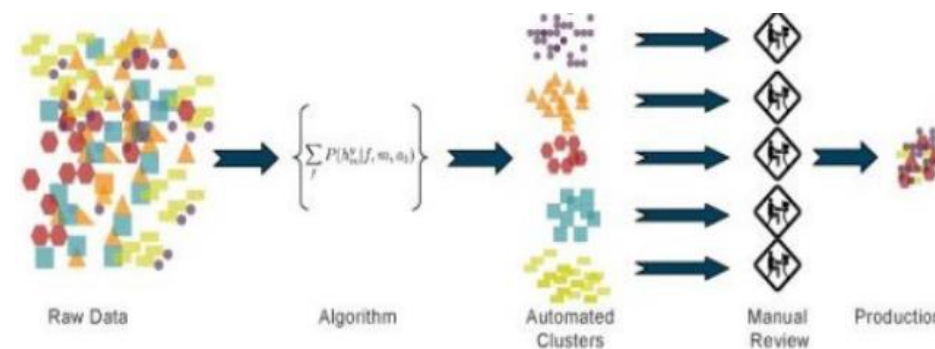
Prof. Daniel Alfredo Chávez Gallo

dacg160381@Hotmail.com

Modelos no supervisados

$$X_{ij}$$


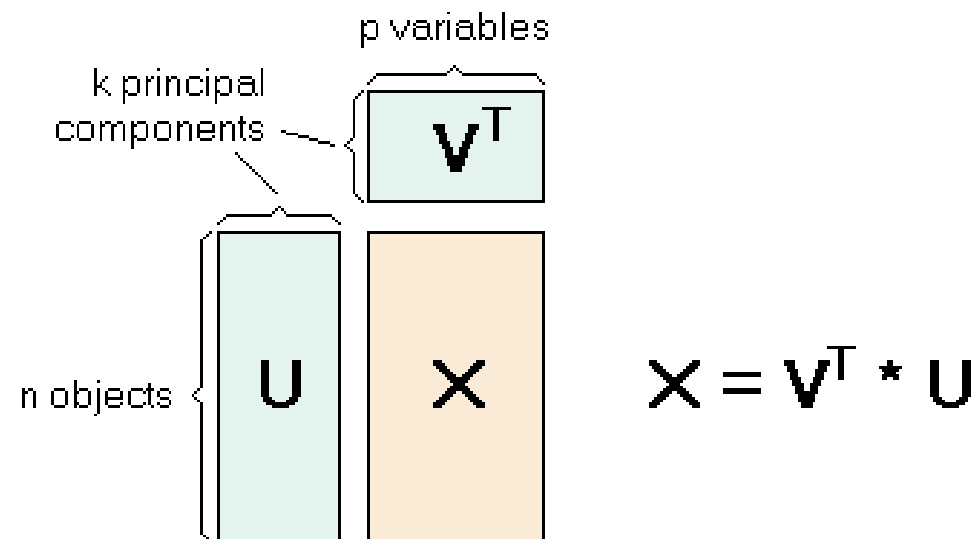
- El aprendizaje no supervisado **busca patrones en los datos**.
- **Se utiliza** cuando se desconoce la estructura de los datos.
 - Por ejemplo, cuando se desconoce cuántos grupos de usuarios similares existen.
- La clave, consisten en buscar buenas *variables* capaces de distinguir entre las diferentes instancias.



MNS – Reducción de dimensionalidad

ACP

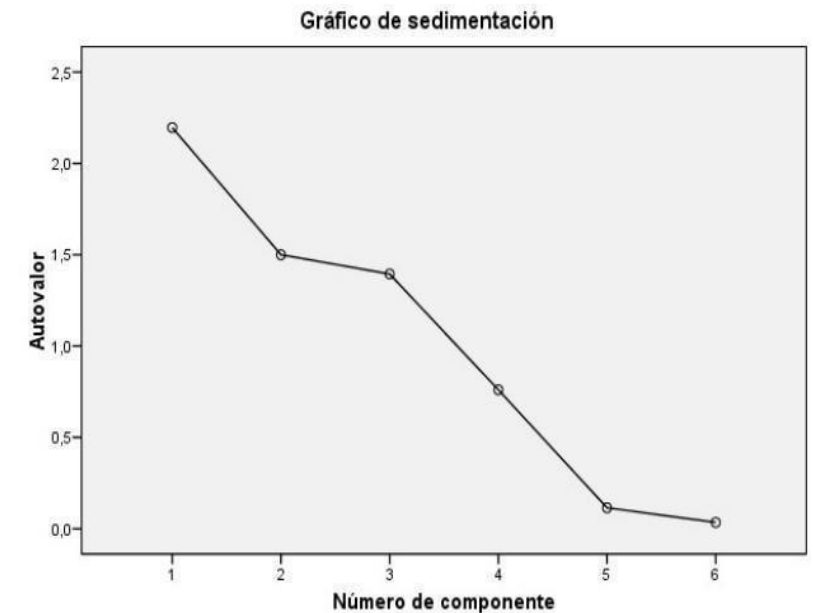
- Técnicas de reducción de dimensionalidad en la cual se reemplaza la matriz de datos por una matriz de componentes, los cuales resumen la misma información que la matriz de datos original.



MNS – Reducción de dimensionalidad

ACP

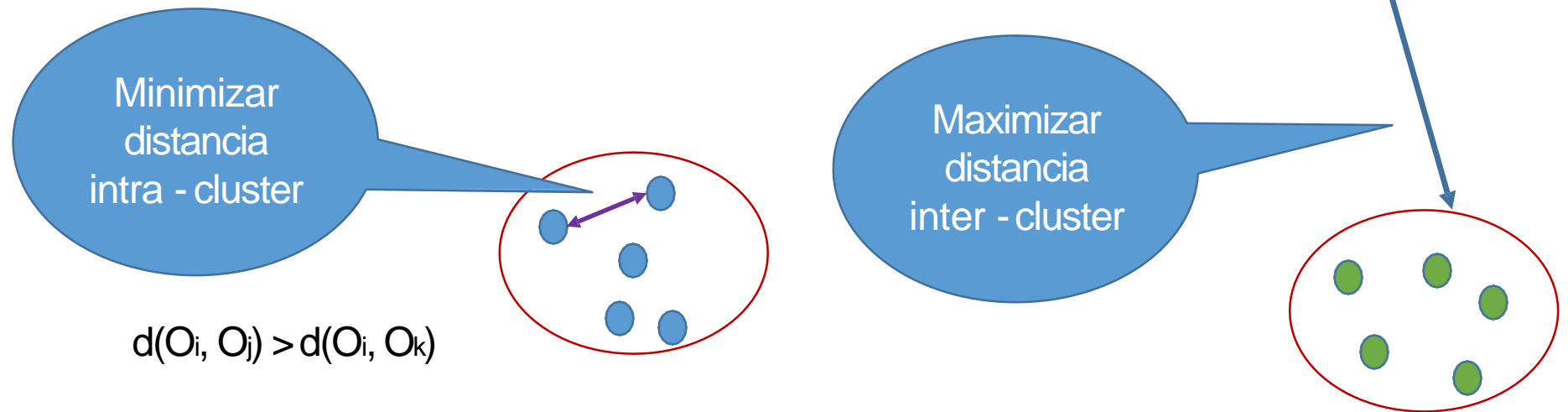
- Se escogen los k primeros componentes, de manera que la varianza acumulada sea mayor al 75%.
- Otro criterio de elección es a través del gráfico de sedimentación (Scree Plot).
- También pueden escogerse los componentes con autovalor mayor que 1.



MNS - Agrupamiento

Los resultados obtenidos dependerán de:

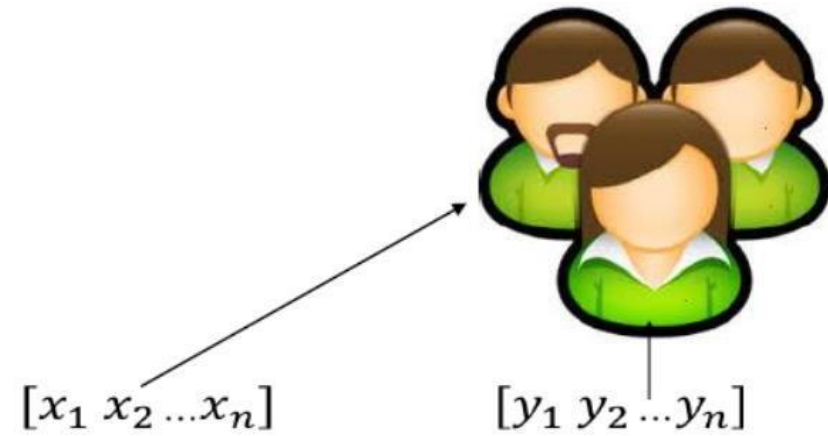
- El algoritmo de agrupamiento seleccionado.
- El conjunto de datos disponible
- La medida de similitud utilizada para comparar objetos.



MNS - Agrupamiento

Noción de similitud

Dada una representación vectorial de dos clientes \mathbf{x} y \mathbf{y} , podemos determinar el grado de similitud entre ellos a través del uso de una **métrica**.



$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^n |x_i - y_i|^2}$$

MNS - Agrupamiento

- **¿Cuántos grupos?**

Grupos o clústers no definidos a priori. Diferencia con los métodos supervisados

- **¿Cómo buscarlos?**

Los objetos dentro de un cluster sean similares o cercanos entre sí en algún sentido (gran similaridad intra-clase) y diferentes o alejados a los objetos de otro cluster (baja similaridad inter-clase)

MNS - Agrupamiento

Para medir la distancia entre las instancias de datos, es necesario que todos los atributos estén en la misma escala

- **Normalización:** escala los valores numéricos en el Rango [0,1]

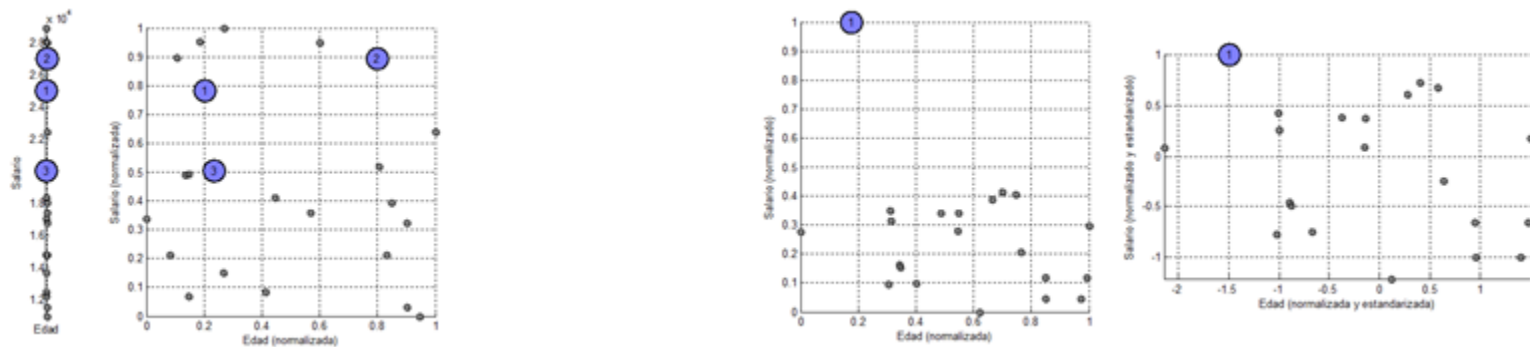
$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- **Estandarización:** hace que la distribución de los datos sea normal

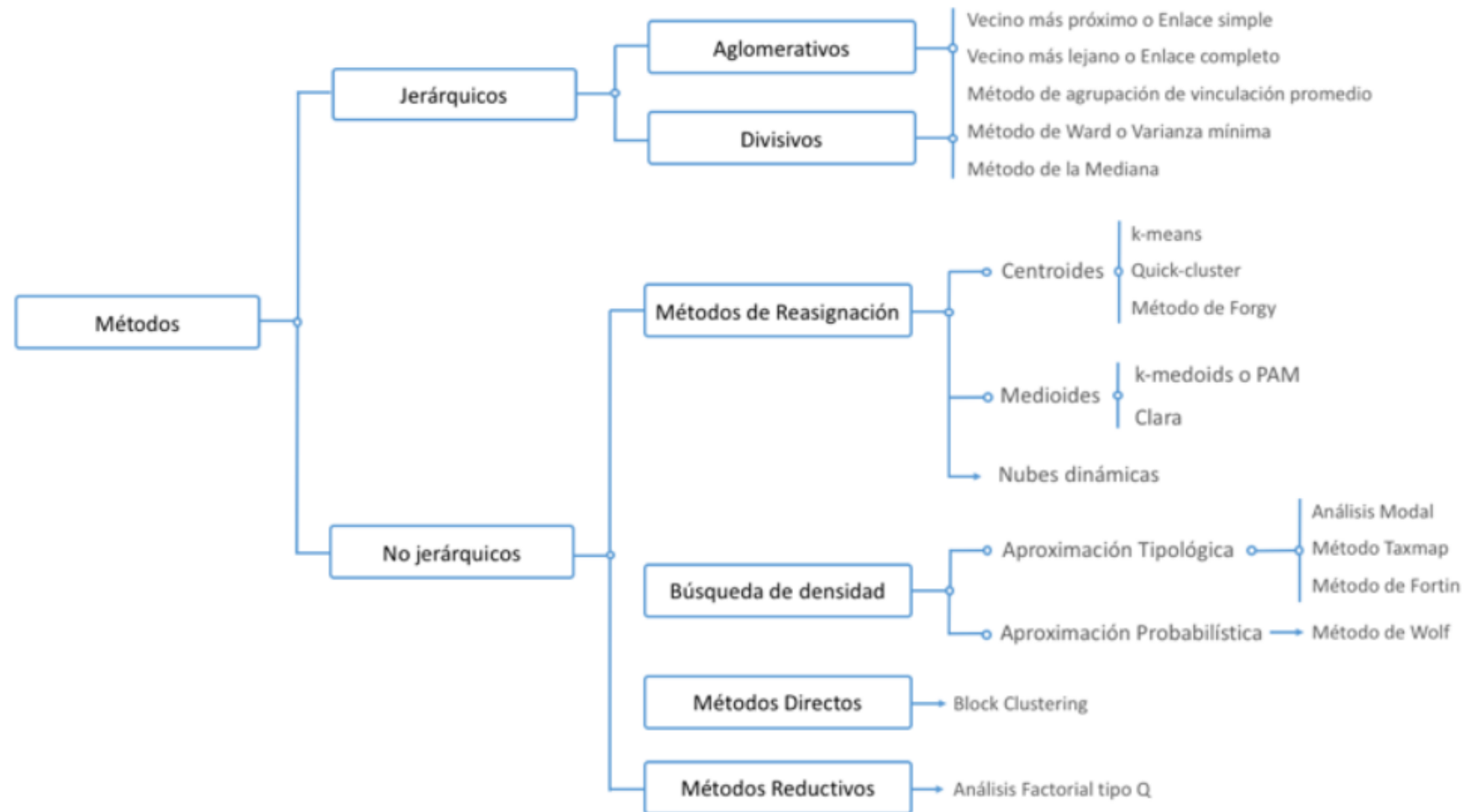
$$x_{new} = \frac{x - \mu}{\sigma}$$

MNS - Agrupamiento

- Tanto normalizar como estandarizar tienen sus ventajas e inconvenientes:
 - Normalizar nos asegura que los valores estarán entre 0 y 1, para todos los atributos
 - (Estandarizar, no)
 - Estandarizar mitiga el efecto negativo de los outliers, hace que los datos estén “mejor distribuidos”, lo que facilita la tarea de minería
 - (Normalizar, no)

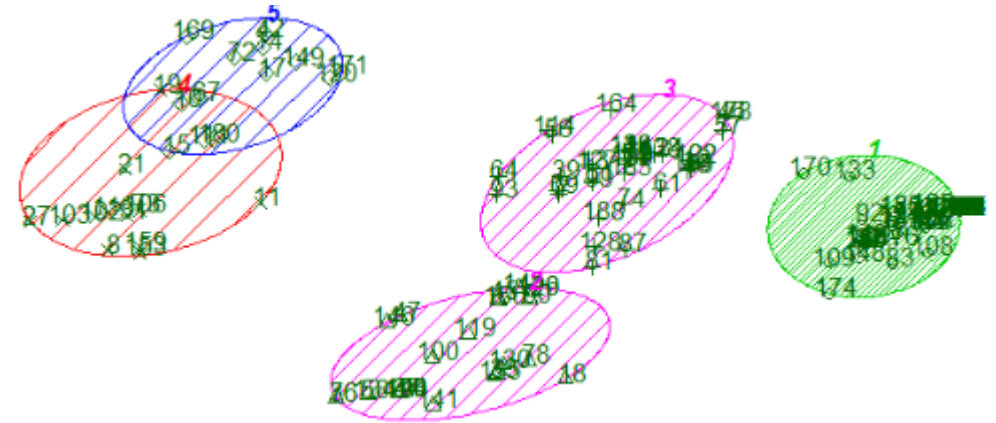


MNS - Clústers



MNS - Algoritmo K Means

- Probablemente el más utilizado y conocido
- Asigna cada observación a uno de los k clusters
- K es un número definido a priori
- Minimizar las distancias intra cluster y maximizar las inter clase



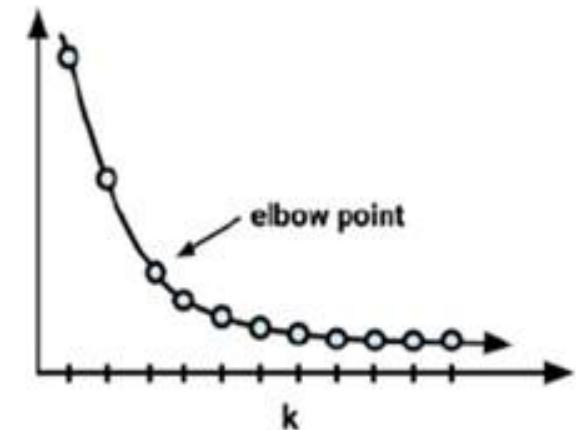
MNS - Algoritmo K-Means

- ¿Cómo funciona el algoritmo?
 1. Elegir el valor de K (número de clusters).
 2. Asignar cada objeto al grupo más cercano (por ejemplo distancia euclídea)
 3. Re-estimar los centros de los k clusters, asumiendo que las asignaciones a los grupos están bien.
 4. Repetir el paso 3 hasta que no haya más cambios
- Se puede cambiar el punto 2, empezando con k centroides iniciales
- La mayor parte de las reasignaciones ocurren en la primera iteración del algoritmo

MNS - Algoritmo K-Means

Elegir el número de clústers

- Conocimiento a priori: por ejemplo, si clasificamos películas, $k = \text{nº de géneros}$
- Dirigidos por el negocio: por ejemplo, el departamento de Marketing sólo tiene recursos para hacer 3 campañas distintas de marketing
- Sin nada de lo anterior: $k = \text{raíz}(n/2)$



MNS - Algoritmo K-Means

➤ **Ventajas**

- Principios no estadísticos
- Muy flexible
- Funciona bien en casos de la vida real
- Rápido: no hay calcular las distancias entre todas y cada una de las observaciones

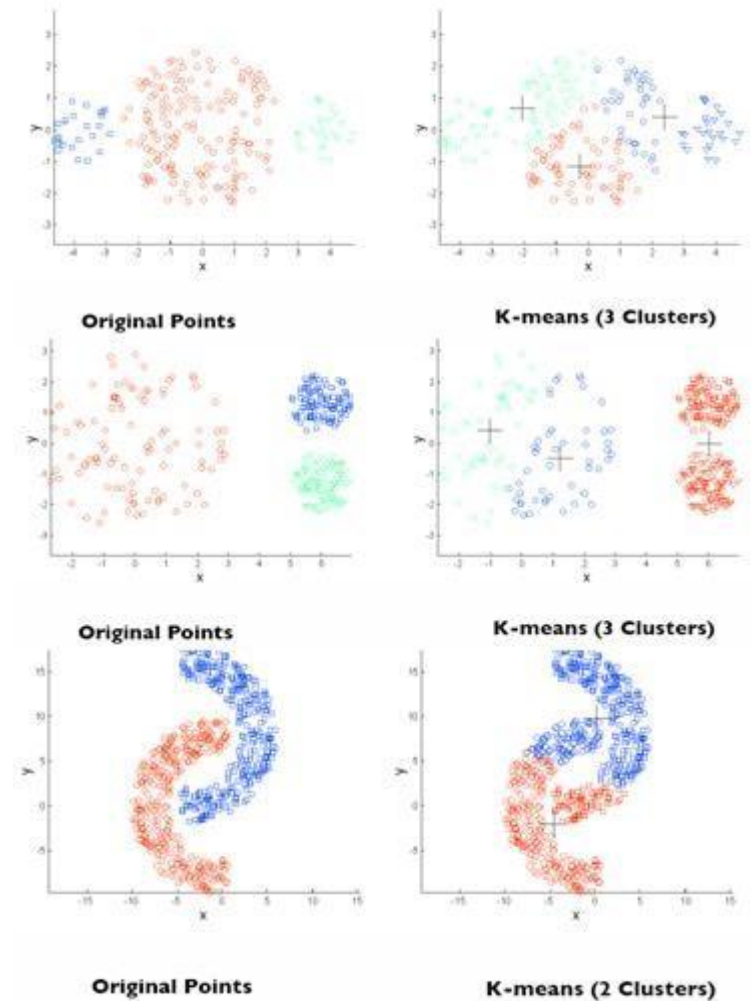
➤ **Desventajas**

- No muy sofisticado
- No está garantizado encontrar en número de clusters óptimo
- Sensible a outliers que pueden formar clusters propios
- La solución final depende del punto de partida

MNS - Algoritmo K-Means

Limitaciones

- Principalmente, su desempeño se ve mermado cuando los clusters tienen
 - Diferentes tamaños
 - Diferentes densidades
 - Formas no globulares
- (Al igual que casi todos) También presenta problemas cuando los datos contienen outliers
- Una solución puede ser hacer un número superior de clusters, y luego “unir las partes”



Modelos supervisados

$$\begin{matrix} & X_{ij} & & Y_i \\ \begin{bmatrix} \square & \dots & \square \\ \vdots & \ddots & \vdots \\ \square & \dots & \square \end{bmatrix} & \sim & \begin{bmatrix} \square \\ \square \\ \square \end{bmatrix} \end{matrix}$$

- **Clasificación:** Cuando la variable a predecir es una categoría.
 - Binaria: {Sí, No}, {Azul, Rojo}, {Fuga, No Fuga}...
 - Múltiple: {Comprará Producto1, Producto2...}...
 - Ordenada: {Riesgo Bajo, Medio, Alto}...
- **Regresión:** Cuando la variable a predecir es una cantidad
 - Precio, cantidad, tiempo,...

MS - Regresión

- La regresión intenta aproximar el comportamiento de una variable **Y** en función **f(x)**, variables **x**.

$$Y = f(x) + \text{error}$$

- El objetivo de la regresión es minimizar el **error** entre la función aproximada y el comportamiento de la variable **Y**

$$\text{error} = Y - f(x)$$

- Donde **x**, son variables independientes

MS - Regresión supuestos

- **Independencia**

Entre los residuales (e_i), mediante Durbin Watson: Si $DW = 2$ los e_i son completamente independientes. Entre 1.5 y 2.5 se considera que existe independencia, $DW < 2$ autocorrelación negativa

- **Homocedasticidad**

Igualdad de varianzas de e_i y los pronósticos. El supuesto implica que la variación de los e_i sea uniforme en todo el rango de valores de los pronósticos (gráfico sin pautas de asociación).

- **Normalidad**

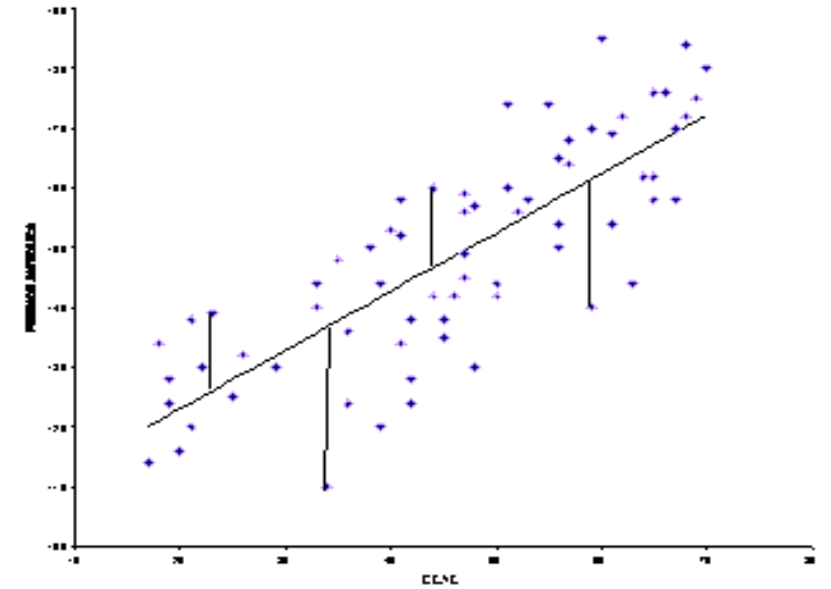
Los residuales tiene distribución normal ?, por la prueba de Shapiro, si el Pvalue > 0.05 entonces se afirma que poseen esta distribución.

- **No Colinealidad**

No correlación entre variables independientes.

MS - Regresión lineal

- Modelo matemático para predecir el efecto de una variable **x** sobre otra **y**
- Se acepta que la relación entre ellas tiene un comportamiento lineal.
- Es **simple** si existe solo una variable predictora y es **múltiple** si existen varias variables predictoras.



$$y = bx + e$$

MS - Regresión lineal - dummy

Conversión a Dummy

- Toma dos valores 1 y 0
 - 1, si toma la modalidad
 - 0, si no toma la modalidad.
- Una variable categórica con **q** estados se reemplaza por **q-1** variables dummy.

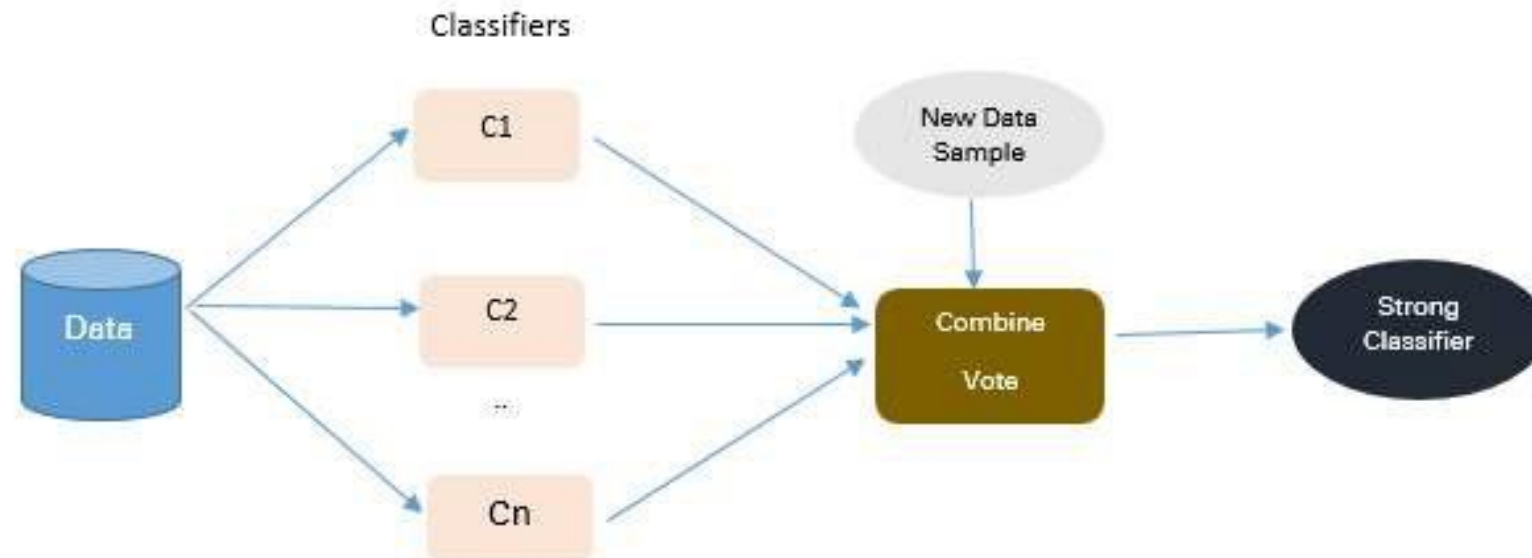
“Numerización”

- Se enumera cada estado con un valor secuencial.
- Se genera solo una variables adicional.

Estado Civil	Numerizado
S	1
C	2
V	3
D	4

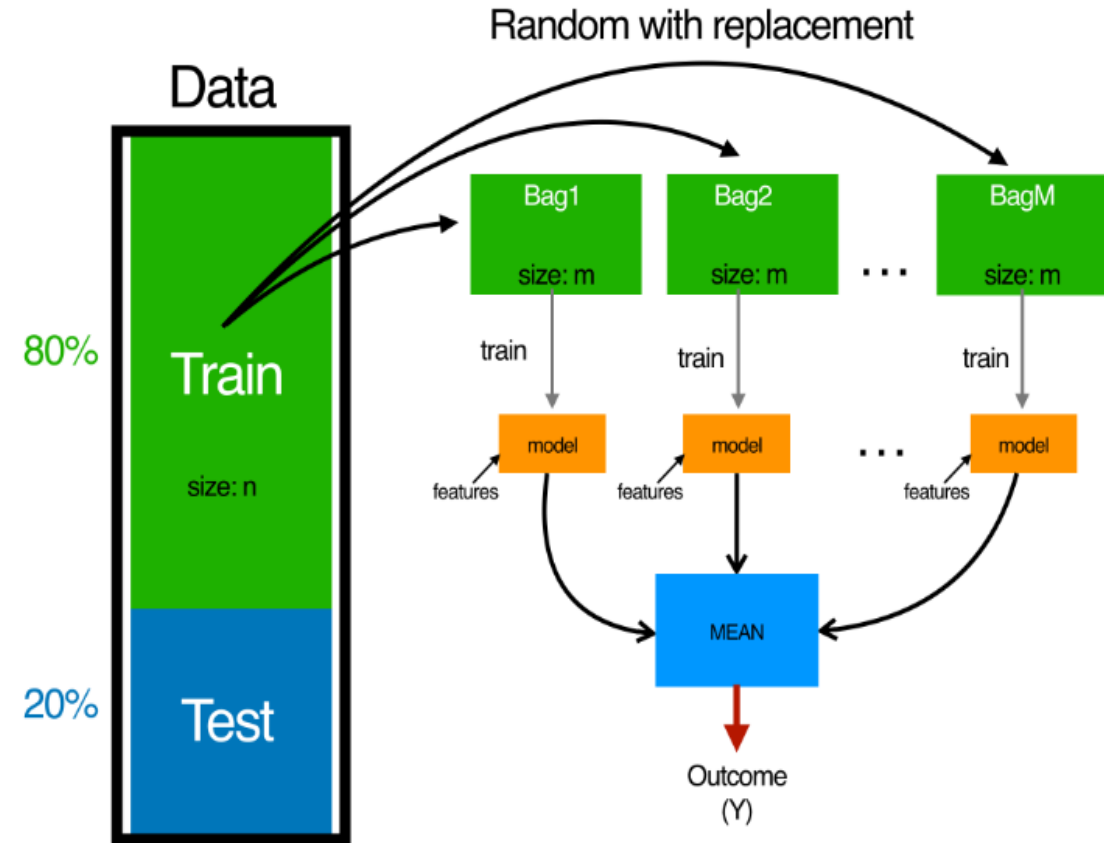
MS - Métodos de ensamble

- Los métodos de ensamble son técnicas para combinar varios algoritmos de aprendizaje débiles con la finalidad de construir un algoritmo de aprendizaje más fuerte.
- Los métodos de ensamble combinan múltiples hipótesis para formar una mejor hipótesis.



MS - Métodos de ensamble

- Evaluar la predicción de un ensamble, requiere mayor computación que un modelo simple.
- Algunos de los principales métodos de ensamble son:
 - Bagging
 - Boosting
 - Stacking



MS - Métodos de ensamble

Learners débiles y fuertes

1. Strong (PAC) Learner

- Utilizamos datos etiquetados para entrenamiento
- Producimos un clasificador que es preciso arbitrariamente
- Objetivo del Machine Learning

2. Weak (PAC) Learner

- Utilizamos datos etiquetados para entrenamiento
- Producimos un clasificador que es más preciso que el arbitrario

- Given some training data

$$\mathcal{D}_{\text{train}} = \{\mathbf{x}_n, y_n\}_{n=1, \dots, N_{\text{train}}}$$

- Inductive learning

$$\mathcal{L}: \mathcal{D}_{\text{train}} \rightarrow h(\cdot), \text{ where } h(\cdot): \mathcal{X} \rightarrow \mathcal{Y}$$

- Ensemble learning

$$\mathcal{L}_1: \mathcal{D}_{\text{train}} \rightarrow h_1(\cdot)$$

$$\mathcal{L}_2: \mathcal{D}_{\text{train}} \rightarrow h_2(\cdot)$$

...

$$\mathcal{L}_T: \mathcal{D}_{\text{train}} \rightarrow h_T(\cdot) \Rightarrow \text{Ensemble: } \{h_1(\cdot), h_2(\cdot), \dots, h_T(\cdot)\}$$

MS - Métodos de ensamble

Que son?

El arte de combinar un conjunto diverso de modelos predictivos.

- ¡¡ Importante !!
 - Diversidad
 - Variedad, desemejanza, diferencia.
 - Abundancia, gran cantidad de varias cosas distintas.
 - Diferencia
 - Cualidad o accidente por el cual algo se distingue de otra cosa.
 - Variedad entre cosas de una misma especie.
 - Controversia, disensión u oposición de dos o más personas entre sí.

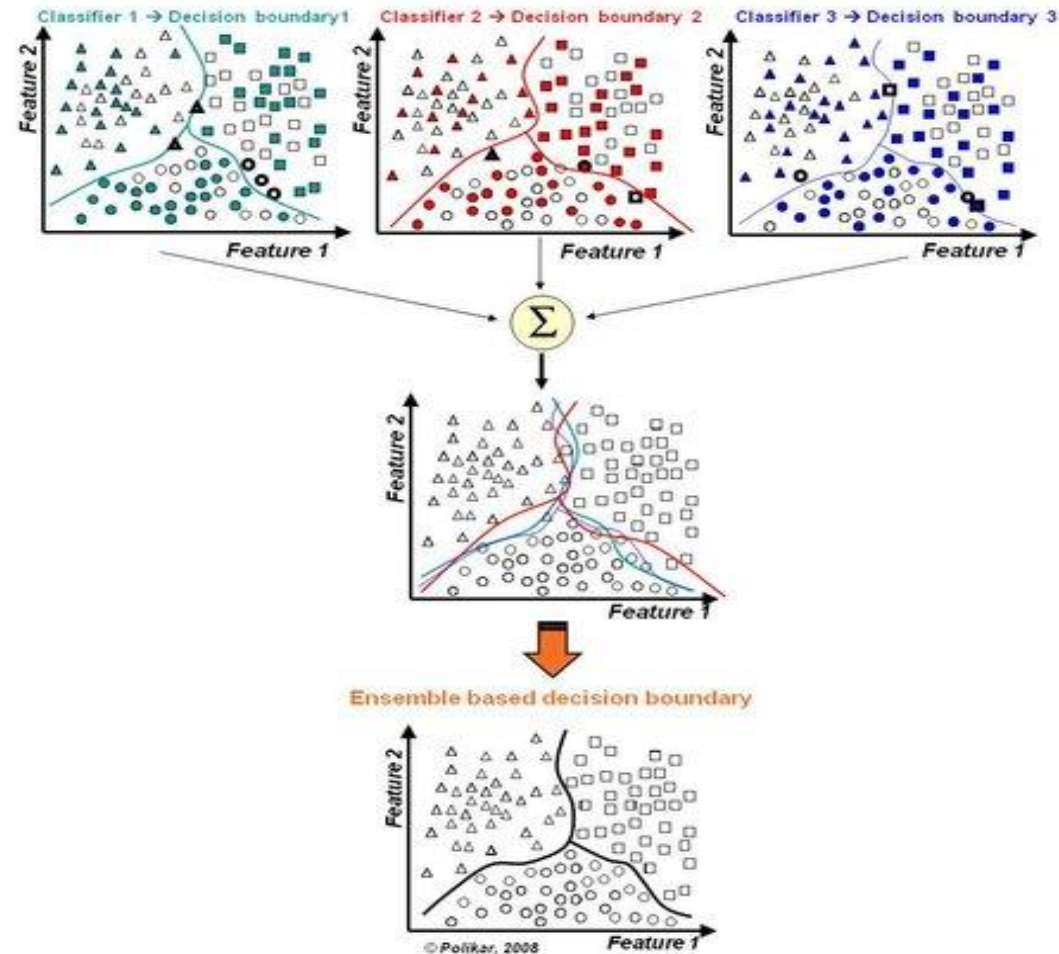
MS - Métodos de ensamble

Que son?

- Son técnicas que crean múltiples modelos que se combinan posteriormente para obtener mejores resultados
- Son modelos que crean mejores resultados que cada uno de los modelos por sí mismos
 - Hablamos de algoritmos como los modelos de regresión logística, árboles, de decisión, etc.
 - Cuando estos modelos son usados como entradas a los Ensemble Methods, los denominamos “modelos de base” (base models)

MS - Métodos de ensamble

Reducción del error de clasificación y/o selección del modelo



MS - Métodos de ensamble

Ventajas	Desventajas
Mejora la precisión	Difícil interpretación de los resultados
Funciona en la gran mayoría de problemas y situaciones	Así, se dificulta la obtención de conclusiones de negocio en ocasiones
Modelos ganadores en hackathones y concursos	Lleva mucho más tiempo que otros modelos → difícil integración en sectores y negocios con decisiones en tiempo real
Aportan robustez y estabilidad	La selección de modelos a ensamblar es un arte que suele costar
Son buenos modelizando relaciones lineales y no lineales	

MS - Métodos de ensamble

Muchos o pocos datos

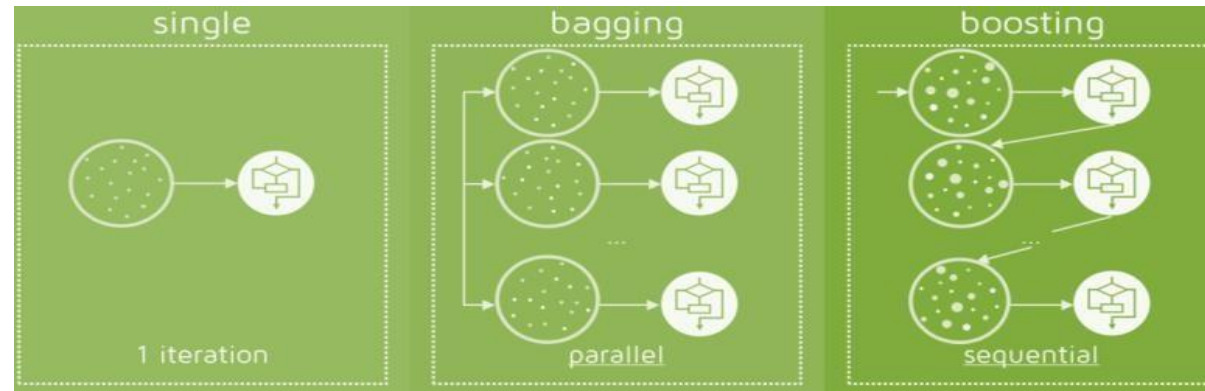
Los Ensemble Methods funcionan bien tanto con grandes como con pequeños volúmenes de datos

Cuando son muchos datos, pueden ser estratégicamente separados para cada que cada parte entrene un modelo diferente

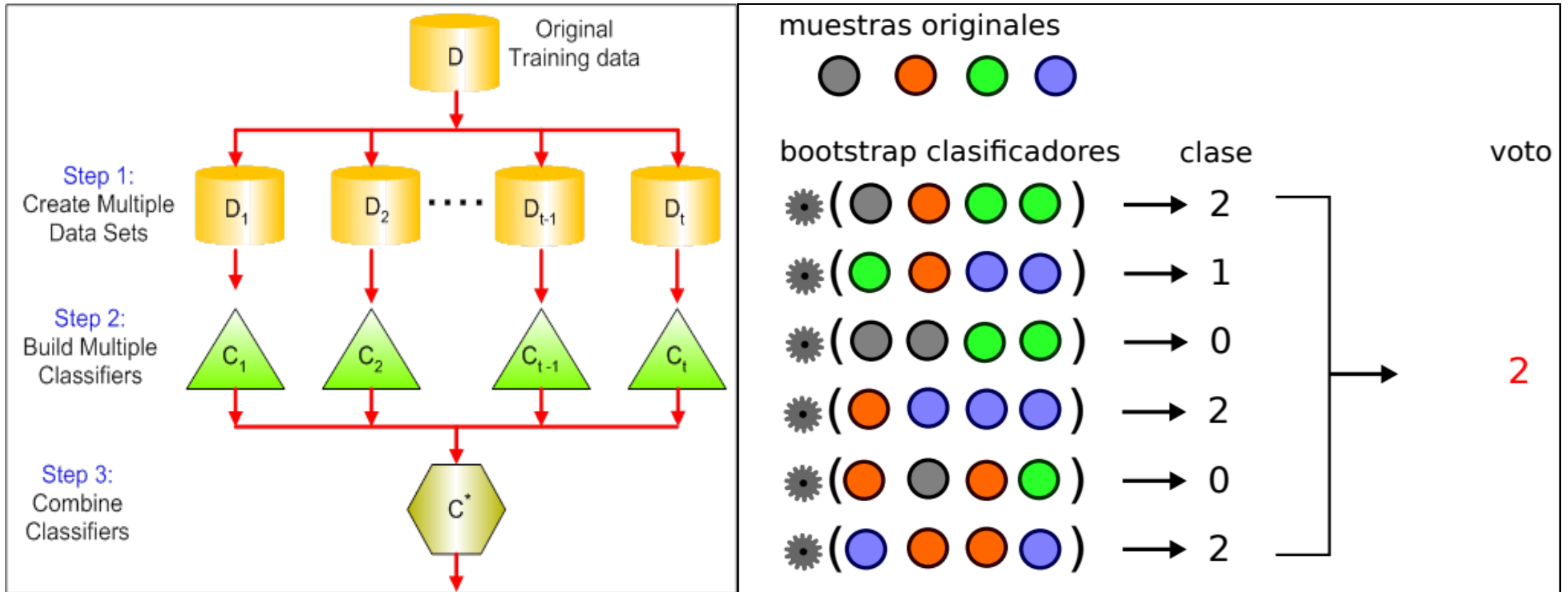
Cuando hay pocos datos, la técnica del bootstrapping puede ser empleada para hacer diferentes muestras “bootstrap” (siendo cada una de ellas una muestra aleatoria e independiente de la distribución de los datos originales)

MS - Bagging

- Una forma de mejorar un modelo predictivo es usando la técnica creada por Leo Breiman que denominó Bagging (o Bootstrap Aggregating).
- Esta técnica consiste en crear diferentes modelos usando muestras aleatorias con reemplazo y luego combinar o ensamblar los resultados.



MS - Bagging



MS - Bagging

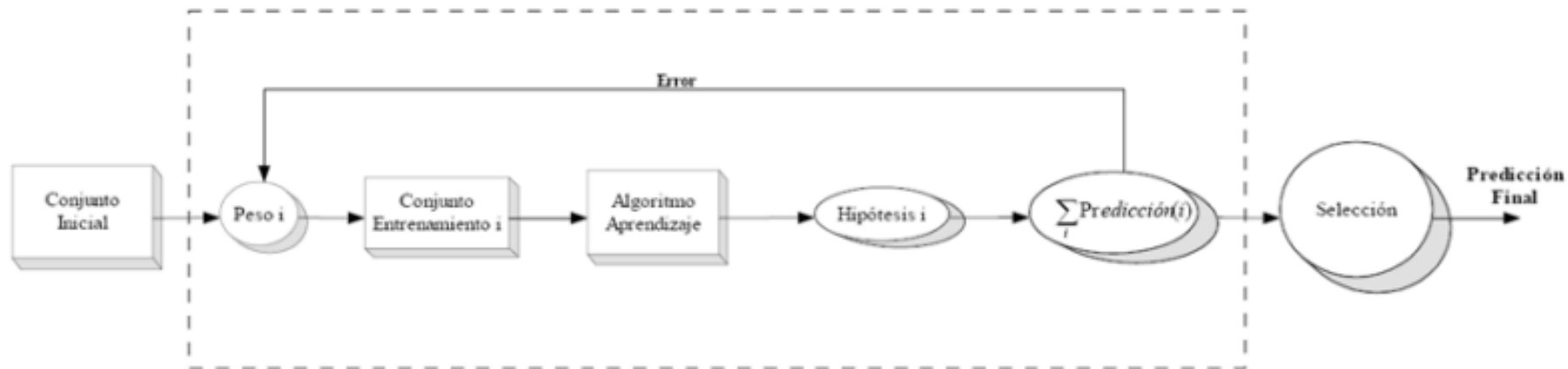
- De esta manera, podemos tener múltiples muestras “bootstrapped” de los mismos datos
- Algunas observaciones pueden aparecer varias veces
- Algunas observaciones pueden no aparecer
- Ahora, podemos hacer “crecer árboles” de estas muestras, y luego utilizar o bien el “voto mayoritario” o la “media” para hacer la predicción final
- Esto se hace para reducir la variación

MS - Bagging

- Cada sub-muestra, se puede generar de manera independiente a la anterior.
- La generación y el entrenamiento se pueden hacer en paralelo.
- También hay algoritmos que implementan esta estrategia de bagging.
- Random Forest, por ejemplo, utilizar una selección aleatoria de características, y su modelo de algoritmo base es un árbol de decisión.
- Se emplean en modelos con muy poco sesgo pero mucha varianza, agregando muchos de estos modelos se consigue reducir la varianza sin apenas inflar el sesgo.

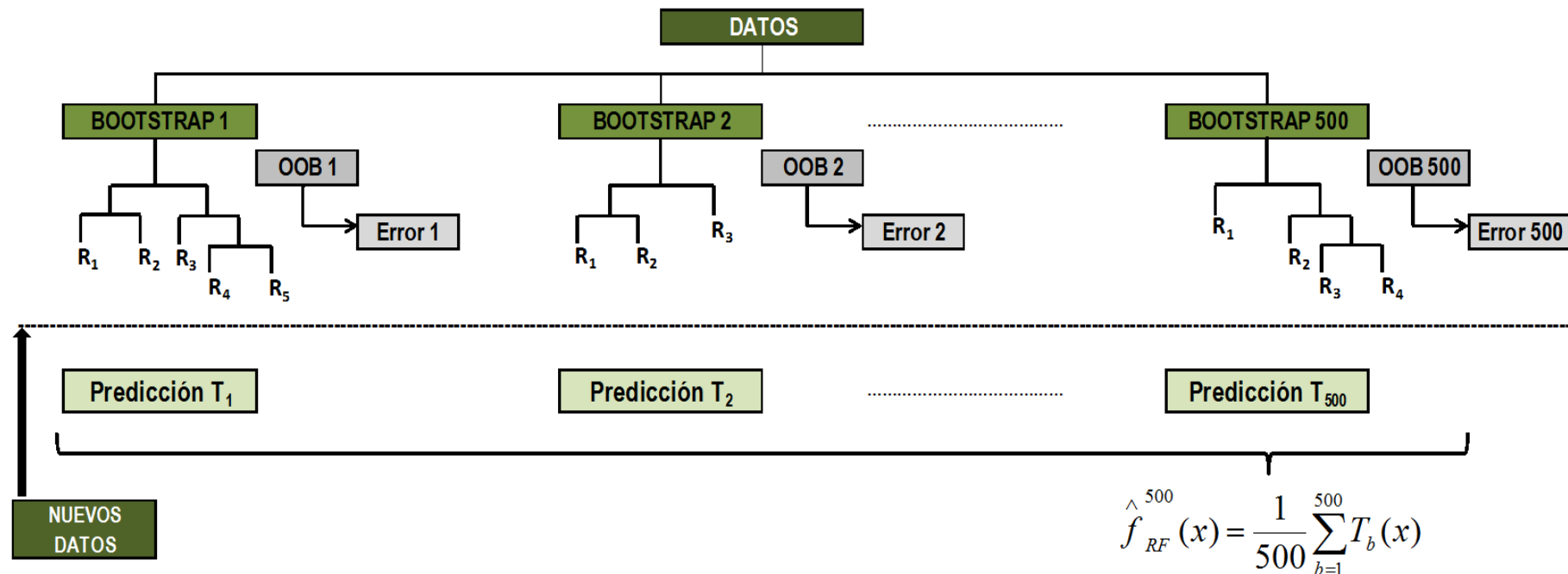
MS - Bagging

Se entrena una serie de clasificadores débiles, así que en cada paso mejoramos el clasificador anterior, terminando en un clasificador fuerte.



MS - Random Forest

- Es un algoritmo predictivo que usa la técnica de Bagging para combinar diferentes arboles, donde cada árbol es construido con observaciones y variables aleatorias.



MS - Random Forest

- Desarrollado por Leo Breiman (Berkeley) en 2001 , tiene su base en los arboles CART.
- La agregación de modelos de árbol se realiza por Bootstrap aggregation (Bagging)
- Comercializado por Salford Systems en la herramienta RandomForestSTM.
- Implementado por Andy Liaw y Matthew Wiener en la librería randomForest del entorno R de programación

MS - Random Forest

- Se define la tasa de error out of bag (OOB) de una observación X_i como el error obtenido al ser clasificada por los árboles del bosque contruidos sin su intervención.
- La estimación OOB del error es el promedio de todos los OOBis para todas las observaciones del conjunto de datos.
- Es mejor estimador que el error aparente. Parecida a la estimación por validación cruzada.
- La medida se puede extrapolar al problema de regresión describiéndola en términos del ECM.

MS - Grid search de algoritmos

- Es utilizado para aumentar el poder predictivo de un modelo
- Para hacer que el modelo sea mas rápido o puntual.
- Ventajas
- Manejo personalizado de los parámetros de entrada para los algoritmos como rpart, C50, random forest, Xgboost, entre otros.
- Desventajas
- Se requiere de un buen performance de maquina, según la exactitud o presición

MS - Grid search de algoritmos

Tomando de ejemplo el Random Forest:

Aumentar el poder predictivo

n_estimadores: Número de árboles necesarios.

Un numero mayor de árboles aumenta el rendimiento y hace que las predicciones sean más estables, pero también ralentiza el cálculo

max_features: Número máximo de variables.

Considerados para dividir un nodo

min_sample_leaf: número mínimo de hojas.

Requieren para dividir un nodo interno

MS - Grid search de algoritmos

Tomando de ejemplo el Random Forest:

n_jobs: # de procesadores a utilizar.

Un numero mayor de árboles aumenta el rendimiento y hace que las predicciones sean más estables, pero también ralentiza el cálculo

Random_state: salida del modelo sea replicable.

Considerados para dividir un nodo

oob_score: método de validación cruzada.

Aprox. es un tercio de los datos no se utiliza para entrenar el modelo y se puede usar para evaluar su desempeño.

Aumentar la velocidad

Muchas Gracias!



BSG Institute

conocimiento para crecer

Av. José Pardo 650, Miraflores - Lima
Urb. León XIII Calle 2 N° 107, Cayma - Arequipa
Carrera 45 #108-27 torre 1 oficina 1008 - Bogotá
Av. Marcelo Terceros Bánzer 304 (3er anillo externo) - Santa Cruz