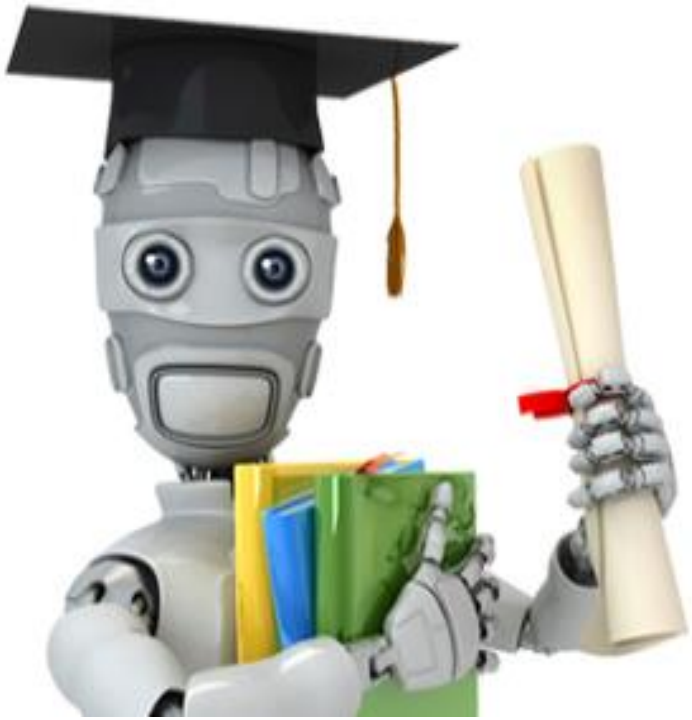
APLICACIÓN E IMPLEMENTACIÓN DE ALGORITMOS DE MACHINE LEARNING EN LA INDUSTRIA

André Omar Chávez Panduro
Correo : andrecp38@gmail.com / 09140205@unmsm.edu.pe
LinkedIn : www.linkedin.com/in/andré-chávez-a90078b9.

Agenda

- ✓ Introducción al Machine Learning.
- ✓ Supervised Learning.
- ✓ Unsupervised Learning.
- ✓ Reinforcement Learning.
- ✓ Deep Learning.
- ✓ Desarrollo de Algoritmos de Machine Learning.
- ✓ Implementación de Algoritmos de Machine Learning.
- ✓ Recursos.

Introducción a los Algoritmos de Machine Learning



¿Qué es el Machine Learning (ML)?

- Rama de la inteligencia artificial que pretende que una máquina sea capaz de mejorar su actuación a la hora de resolver un problema mediante la adquisición de experiencia en una tarea determinada.

- **Multitud de algoritmos** con finalidades específicas.

- Ramas de Machine Learning:

  - ✓ Supervised Learning
  - ✓ Unsupervised Learning
  - ✓ Reinforcement Learning
  - ✓ Deep Learning

Definiciones bàsicas

- **Conjunto de Datos (Data Set):** El total del conjunto de datos sobre los que queremos desarrollar un algoritmo de Machine Learning con el fin de obtener un modelo que lo represente lo mejor posible. Contendrá variables independientes y dependientes.

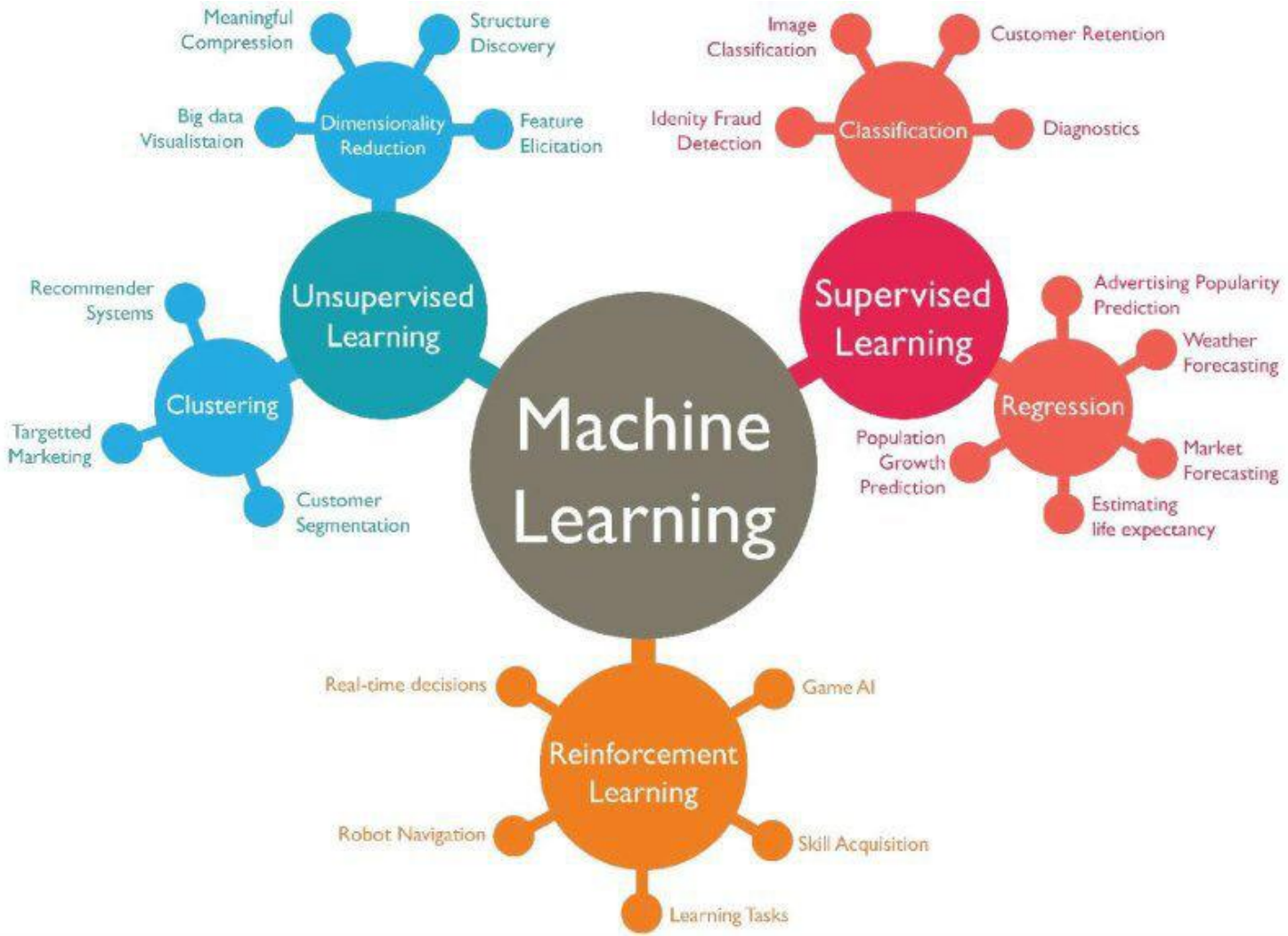
- **Variables Independientes (Features), (VI):** Aquellas columnas del Data Set que serán usadas por el algoritmo para generar un modelo que prediga lo mejor posible las variables dependientes.

- **Variables dependientes (Labels,Target), (VD):** Columna del data set que responde a una correlación de VI y que debe ser predicha por el futuro modelo

- **Conjunto de Datos de Entrenamiento (Training Set):** Subconjunto del Data Set que será utilizado para entrenar el modelo que se pretende generar.

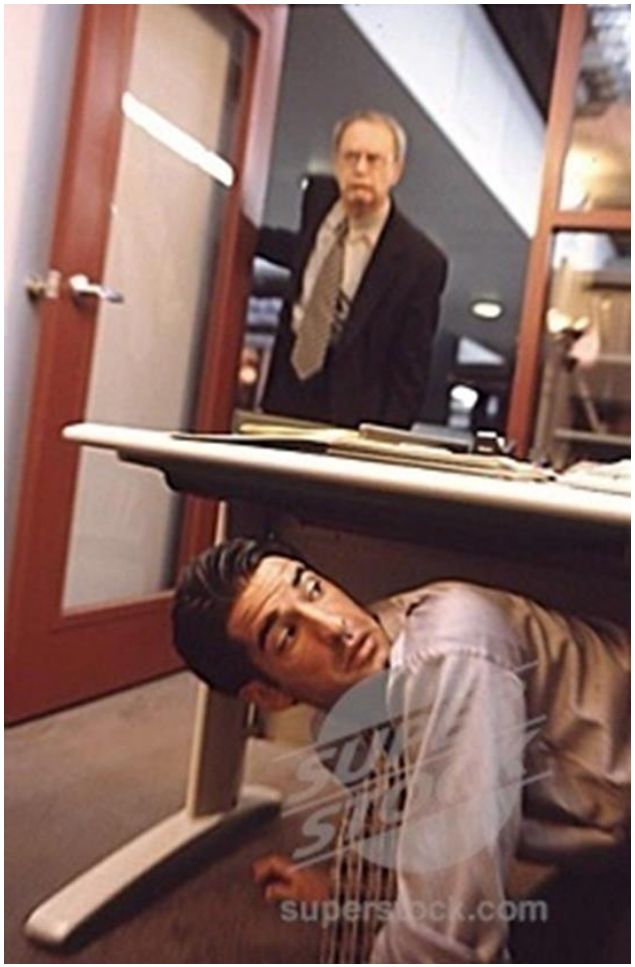
- **Conjunto de Datos de Test (Test Set):** Subconjunto del data set que se le pasará al modelo una vez haya sido entrenado para comprobar, mediante el uso de diferentes métricas, sus indicadores más importantes de calidad.

Introducción



Supervised Learning (Modelos Supervisados)

- Se tiene una **variable objetivo** (Variable de Salida).
- Variables que ayudan a predecir a la variable de salida (Variables de entrada).
- Existe una dependencia de las variables de entrada con las variables de salida.

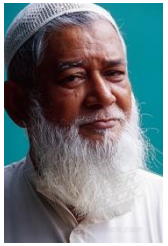


Supervised Learning

- Género.

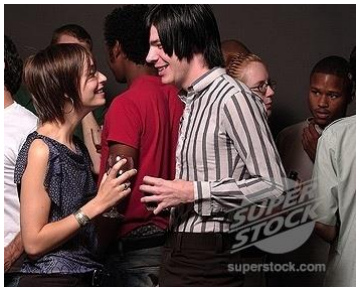

- Rangos de Edad.





- Ingresos.


- Estado Civil.

- Si Compra
- No Compra

Unsupervised Learning (Modelos No Supervisados)

- No hay una variable objetivo (Variable de Salida).
- No hay variables que ayudan a predecir a la variable de salida.

- Todas las variables tienen la misma importancia.
- Se busca la interdependencia de las variables.

Modelos no Supervisados





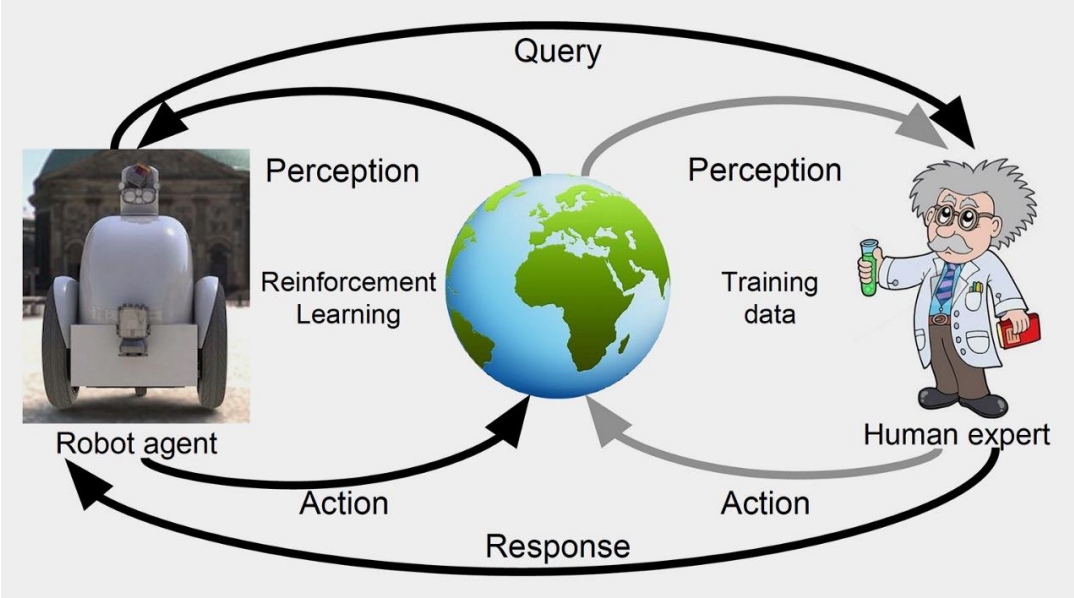






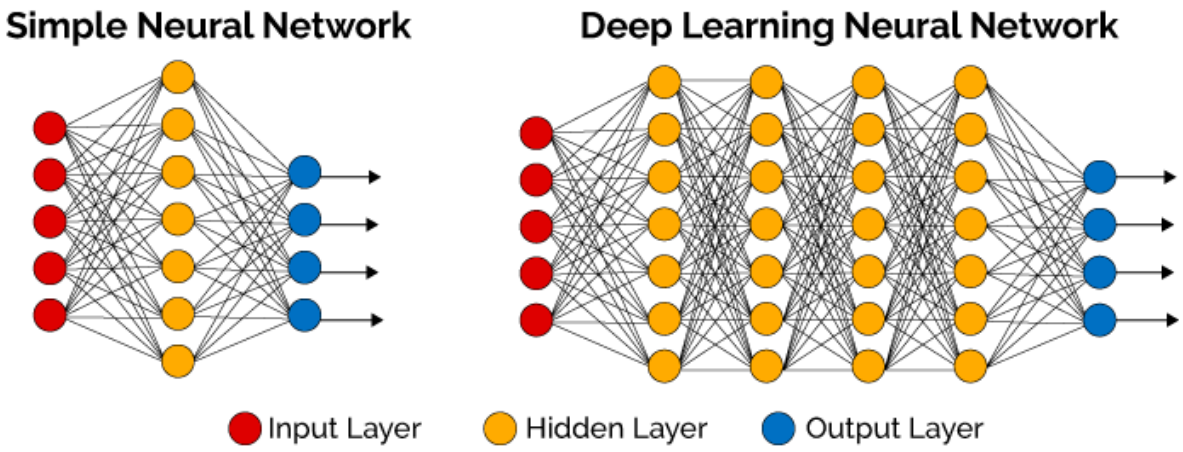

Reinforcement Learning (Aprendizaje por refuerzo)

- El algoritmo de aprendizaje recibe un tipo de valoración acerca de la idoneidad de la respuesta dada.
- Cuando la decisión es correcta es muy parecido al aprendizaje supervisado, sin embargo difiere mucho cuando la decisión es incorrecta.

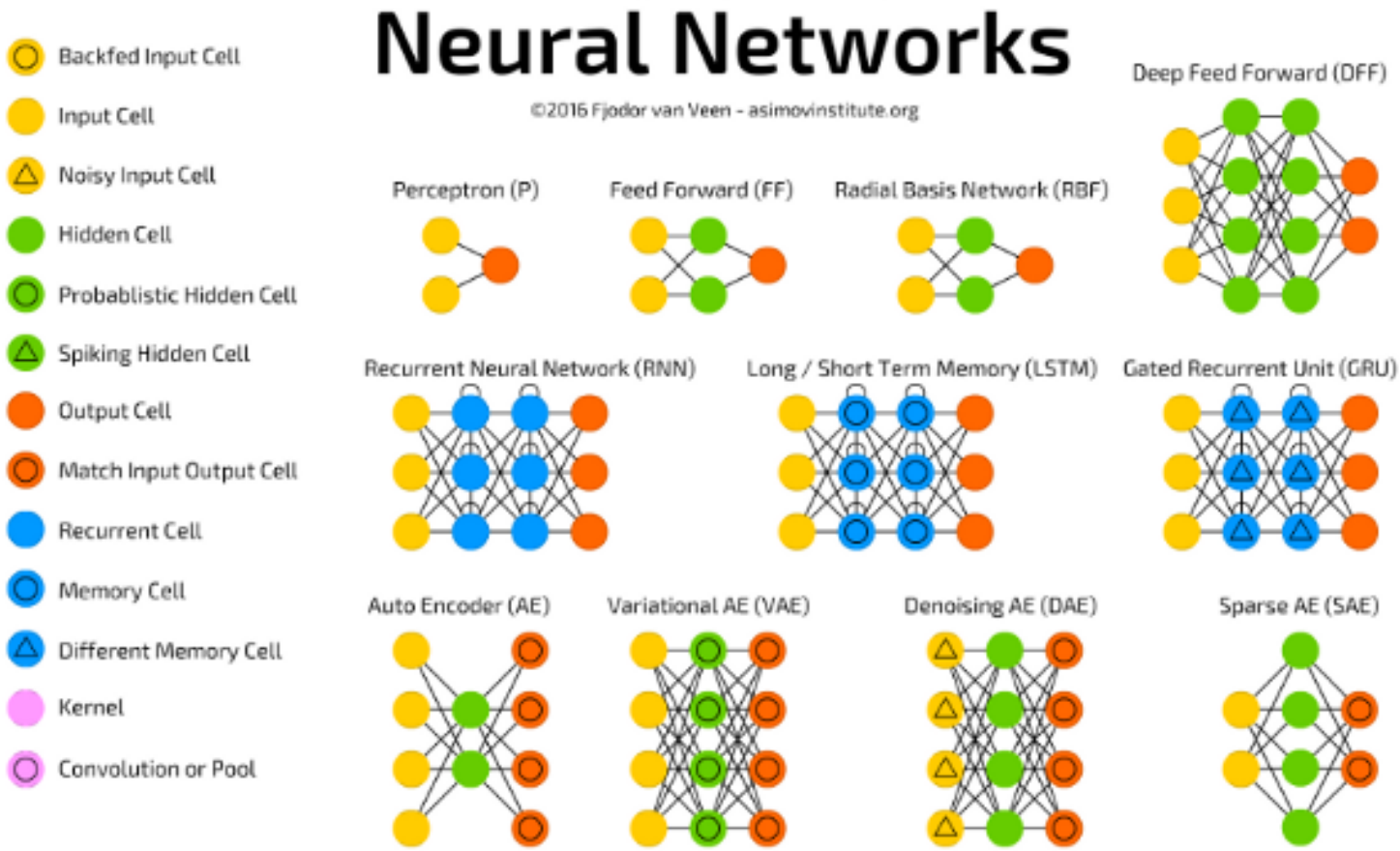


Deep Learning (Aprendizaje Profundo)

- Es un conjunto de algoritmos de Machine Learning que intenta modelar abstracciones de alto nivel usando arquitecturas compuestas como redes neuronales profundas, redes neuronales convolucionales y redes de creencia profunda para resolver problemas como visión del computador, reconocimiento automático del habla, reconocimiento del audio y música, etc.

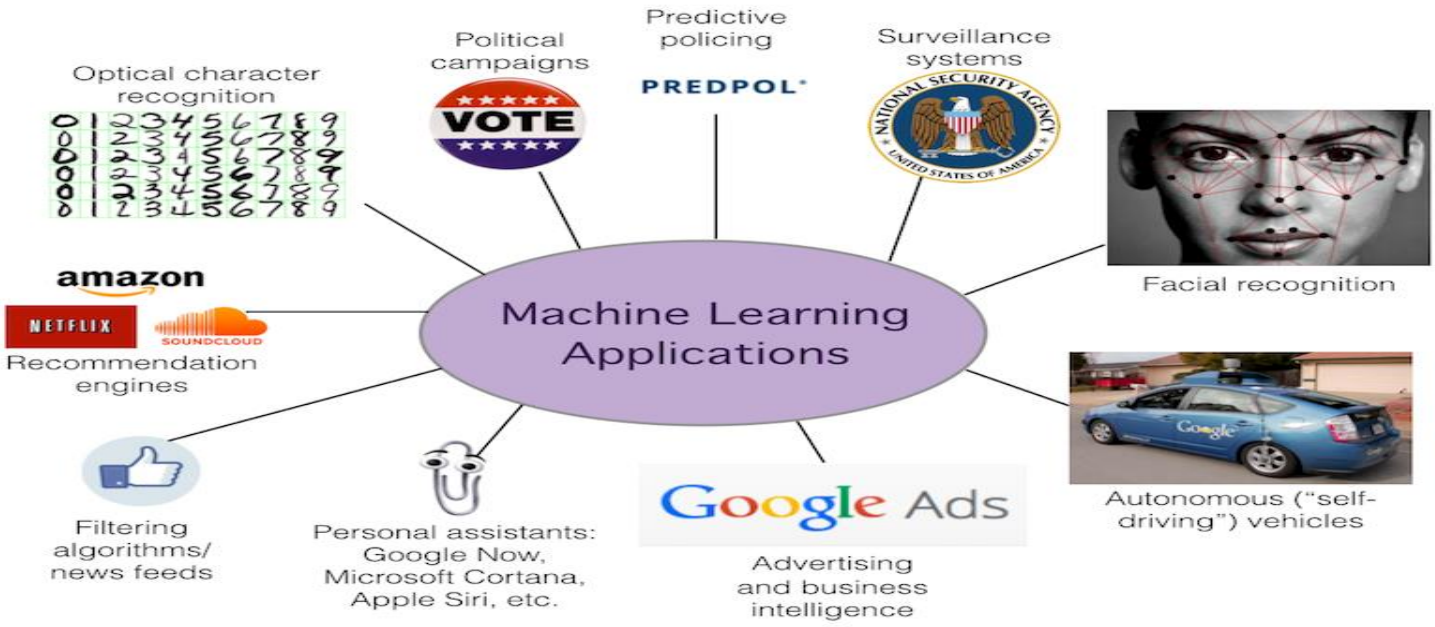


Deep Learning (Aprendizaje Profundo)

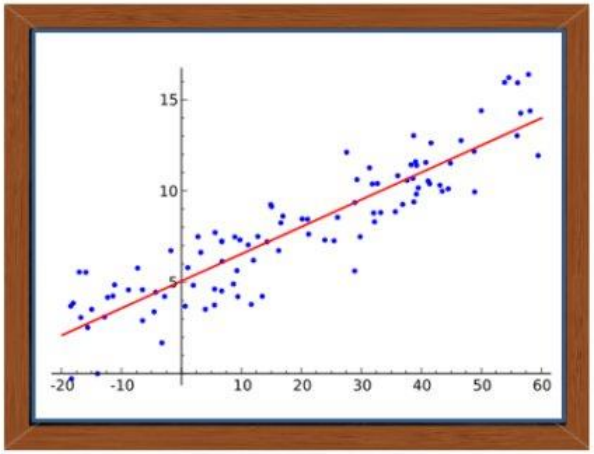


Aplicaciones Machine Learning

- Sistemas de Recomendación.
- Detección de Spam.
- Natural Language Processing (NLP).
- Photo OCR (Optical Character Recognition).
- Visión Artifical.
- Diagnósticos médicos.
- Conducción Autónoma.
- AMD Ryzen ANN.



Algoritmos Machine Learning: Regresión



Introducción

- Determinar la ecuación de regresión sirve para:
  - Describir de manera concisa la relación entre variables.
  - **Predecir los valores de una variable en función de la otra.**
- Veremos EXCLUSIVAMENTE relaciones lineales.
- La regresión lineal simple estudia la relación entre sólo dos variables (el caso de relación más sencillo posible).

ALGORITMOS MACHINE LEARNING: REGRESIÓN

Regresión lineal Simple y múltiple

- Para tratar este tipo de problemas se requiere expandir el análisis de regresión:

Regresión Lineal Simple

Regresión Lineal Múltiple

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots\ldots + \beta_p x_p + \varepsilon$$

ALGORITMOS MACHINE LEARNING: REGRESIÓN

Interpretación del modelo de regresión lineal

$$\hat{Y} = 600 + 300X$$

- Supongamos que tenemos la ecuación de regresión, donde X es el número de años de experiencia profesional, e Y es el sueldo mensual.

  - ✓ Interpreta a y b.
  - ✓ Una persona con 3 años de experiencia laboral, ¿qué sueldo mensual tendrá? Interpreta el resultado.
  - ✓ Si una persona con 3 años de experiencia laboral tiene un sueldo mensual de 1700 €, ¿cuál será su error asociado?

ALGORITMOS MACHINE LEARNING: REGRESIÓN

Interpretación del modelo de regresión lineal

$$\hat{Y} = 600 + 300X$$

- ✓ b=300 → Cambio en Y por cada unidad de cambio en X. Por cada año de experiencia laboral, el sueldo mensual aumenta 300 €.

- ✓ a=600 → Valor medio de Y cuando X=0. Sueldo medio de aquellas personas sin experiencia laboral.

- ✓ Una persona con 3 años de experiencia laboral, ¿qué sueldo mensual tendrá? Interpreta el resultado.

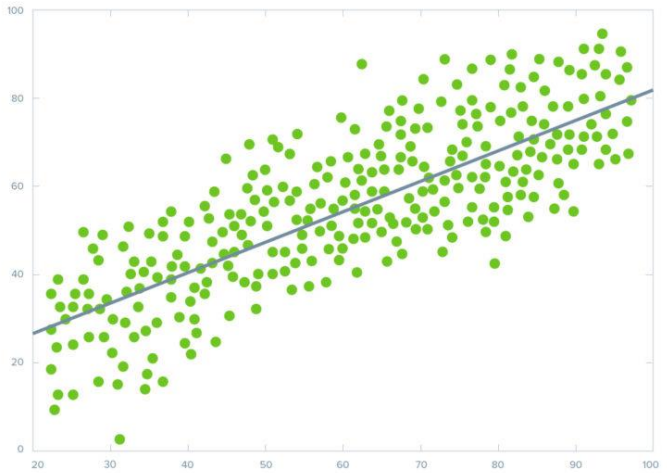
$$X = 3 \Rightarrow Y = 600 + 300 * 3 = 1500$$

$$\hat{Y} = 1500$$

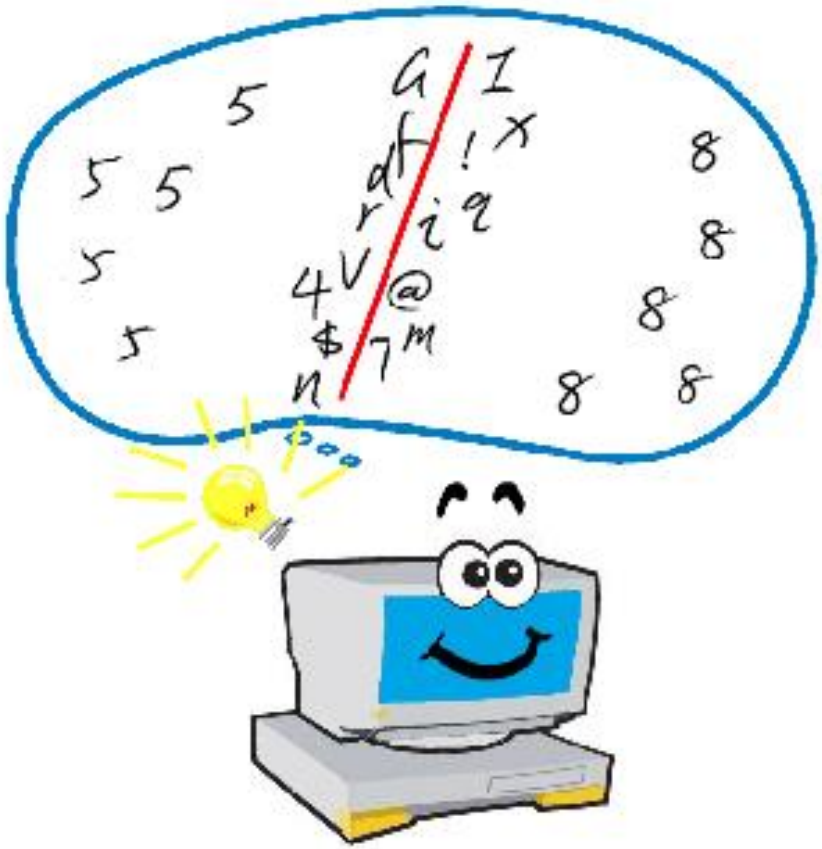
- ✓ → Valor promedio previsto para todos los sujetos que han obtenido en la variable X un valor de X_i. Las personas con 3 años de experiencia tienen un sueldo promedio de 1500 €.

Algoritmos de regresión

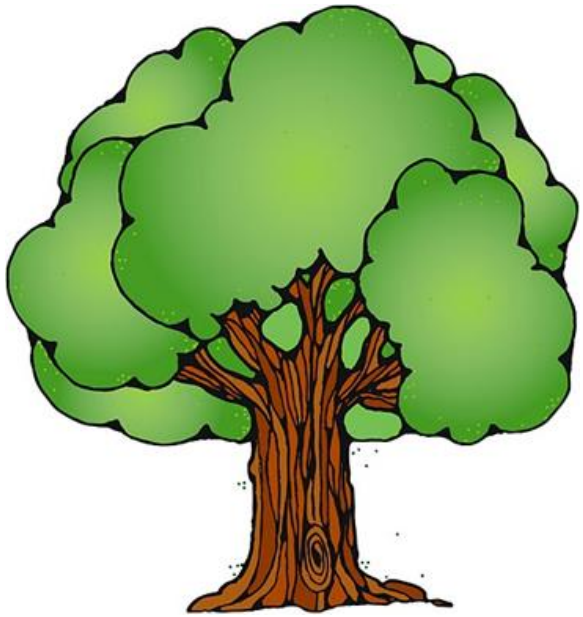
- ✓ **Simple Linear Regression.**
- ✓ **Multiple Linear Regression.**
- ✓ Polynomial Regression.
- ✓ Support Vector Regression (SVR).
- ✓ Decision Tree Regression.
- ✓ Random Forest Regression.
- ✓ XGBoost, LightGBM Regression.

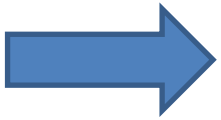


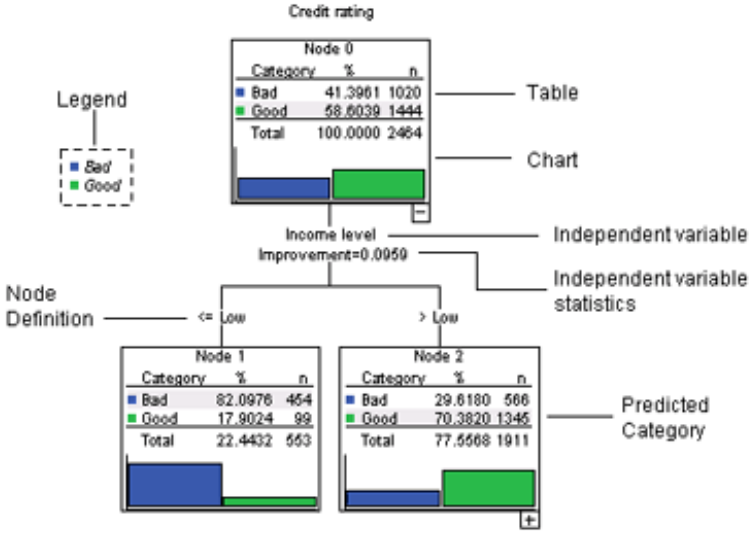
Algoritmos Machine Learning: Clasificaciòn



Árboles de Clasificación



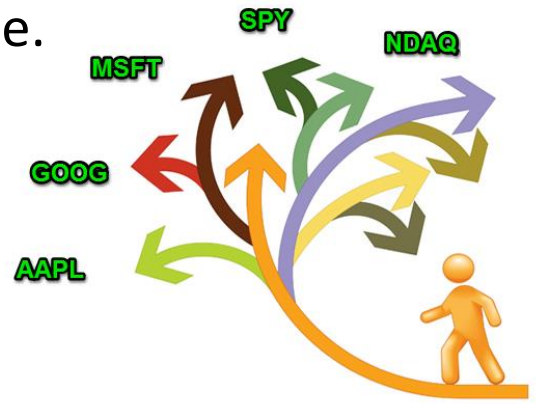




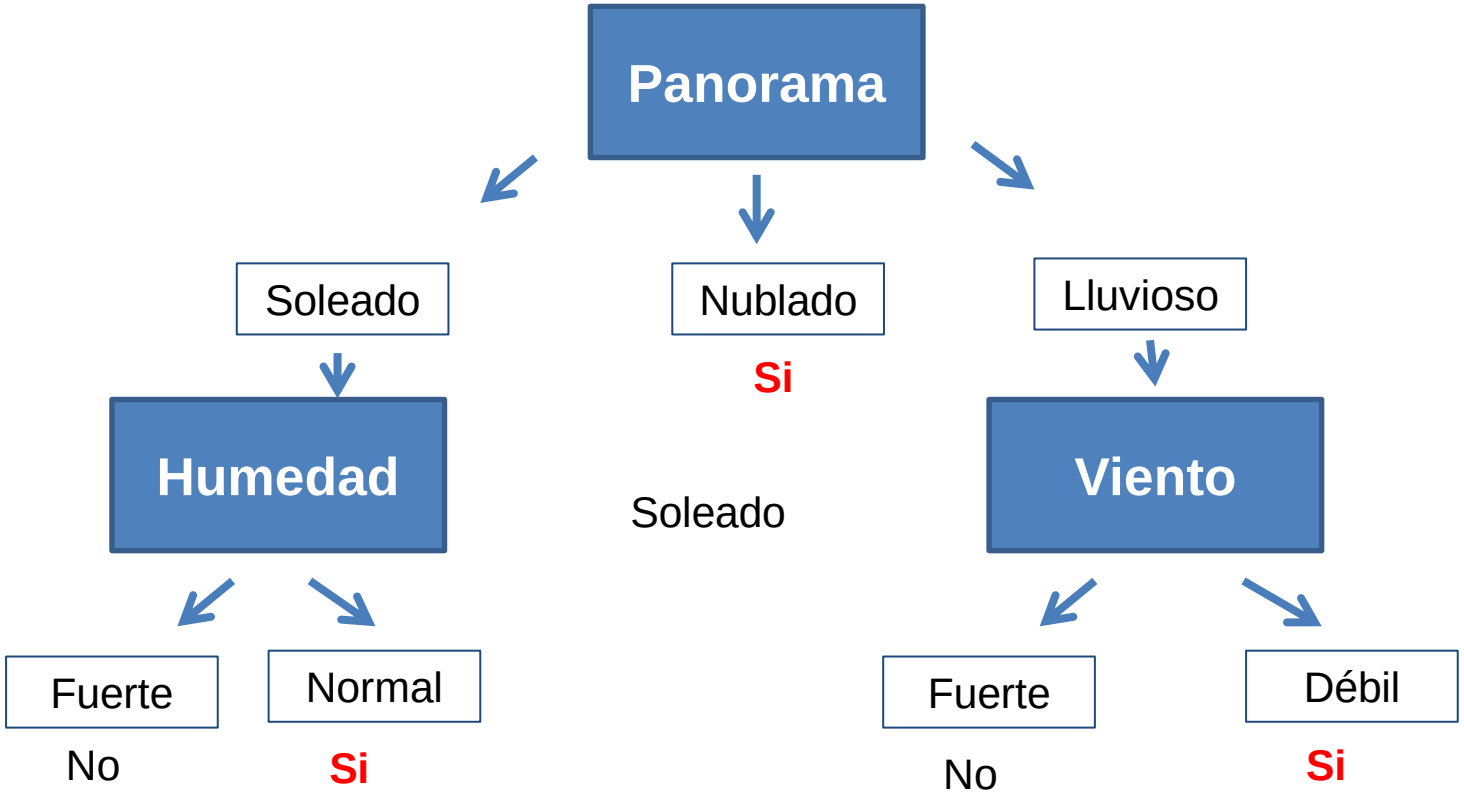
ALGORITMOS MACHINE LEARNING: CLASIFICACIÒN

Árboles de Clasificación

- Entrada:
- Objetos caracterizables mediante propiedades.
- Salida:
  - En árboles de decisión: una decisión (sí o no).
  - En árboles de clasificación: una clase.
- Conjunto de reglas.

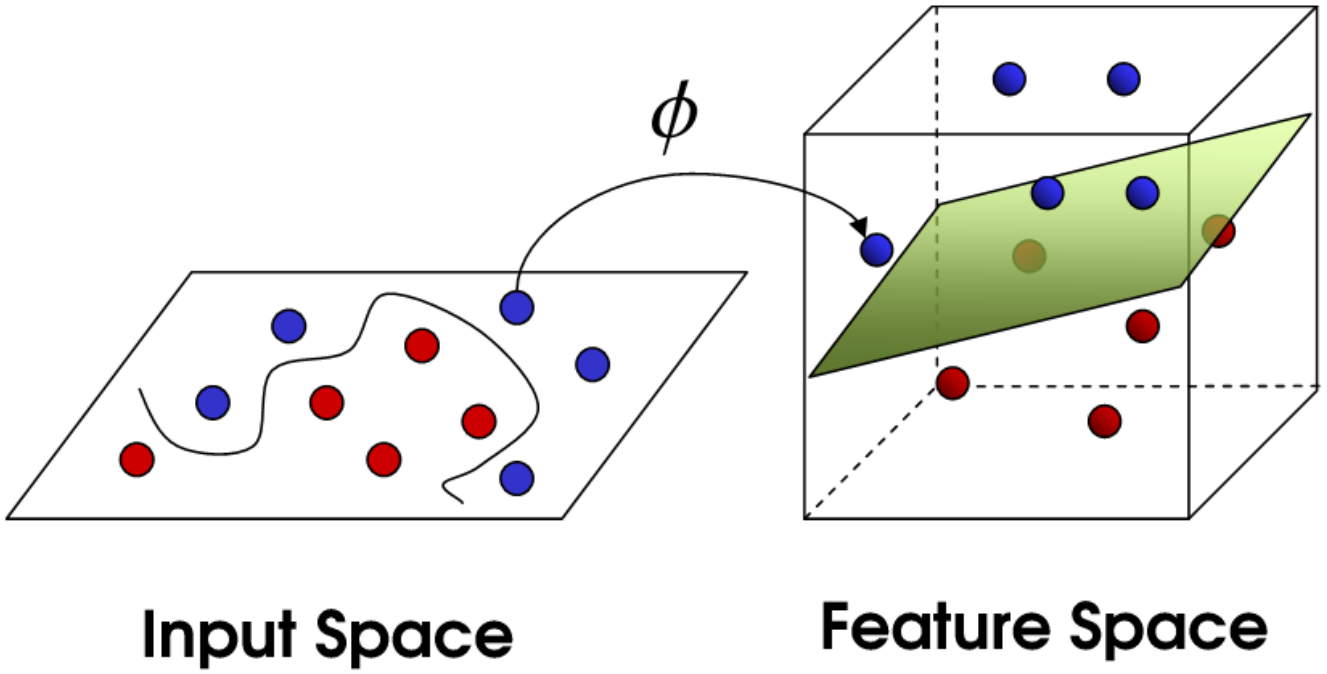


Juego Tennis ?



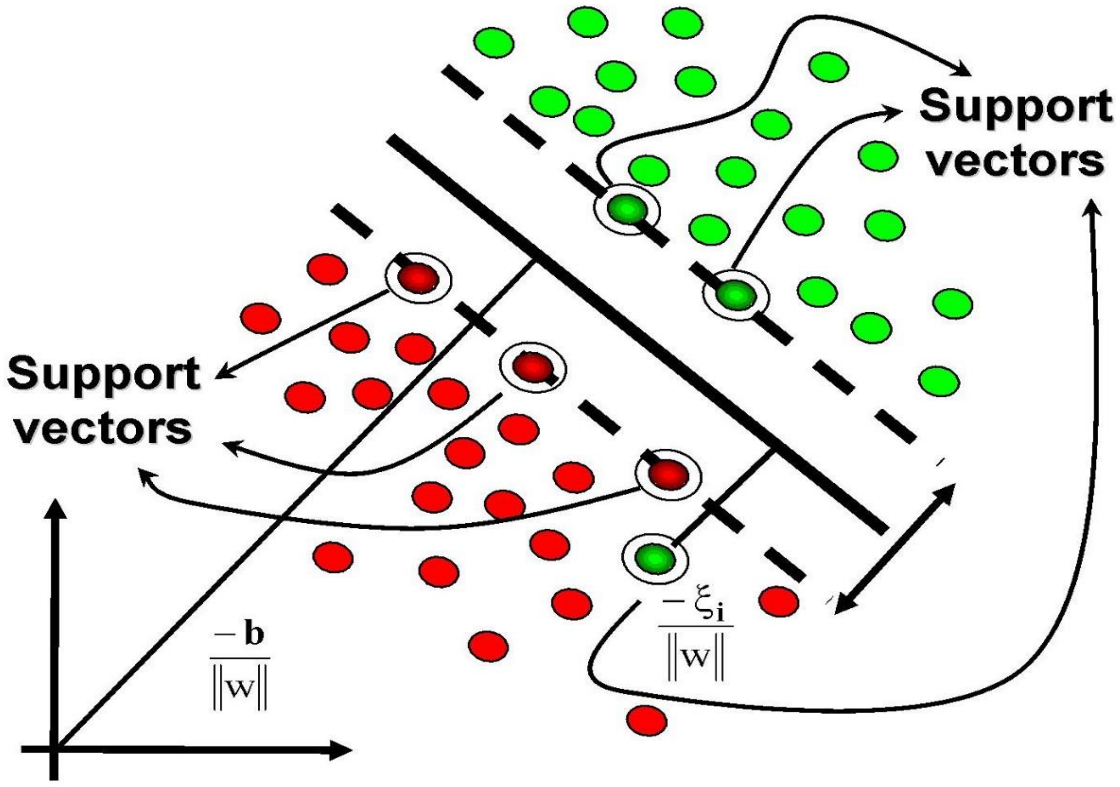
ALGORITMOS MACHINE LEARNING: CLASIFICACIÒN

SVM



ALGORITMOS MACHINE LEARNING: CLASIFICACIÒN

SVM

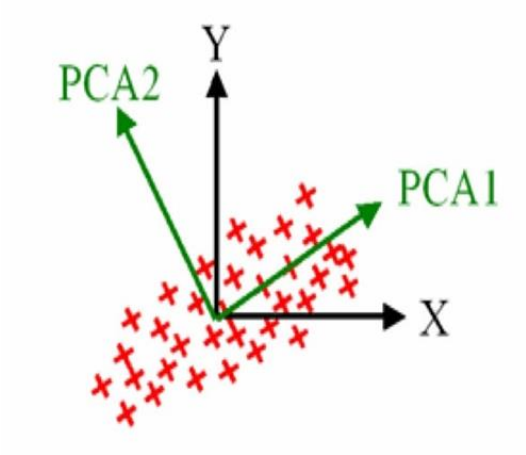


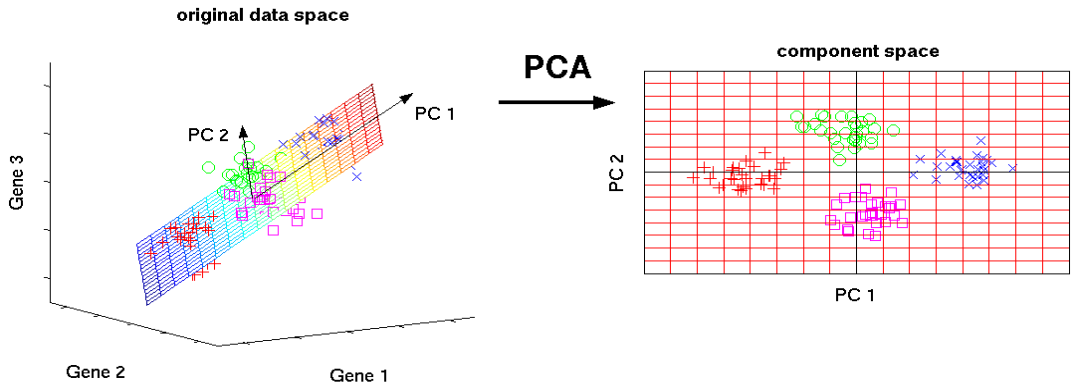
Algoritmos de clasificación

- ✓ Logistic Regression
- ✓ **Support Vector Machine (SVM)**
- ✓ Naïve Bayes
- ✓ **Decision Tree Classification**
- ✓ Random Forest Classification
- ✓ LGBM , CatBoost y XGBoost.

Algoritmos machine learning: Reducción de Dimensión

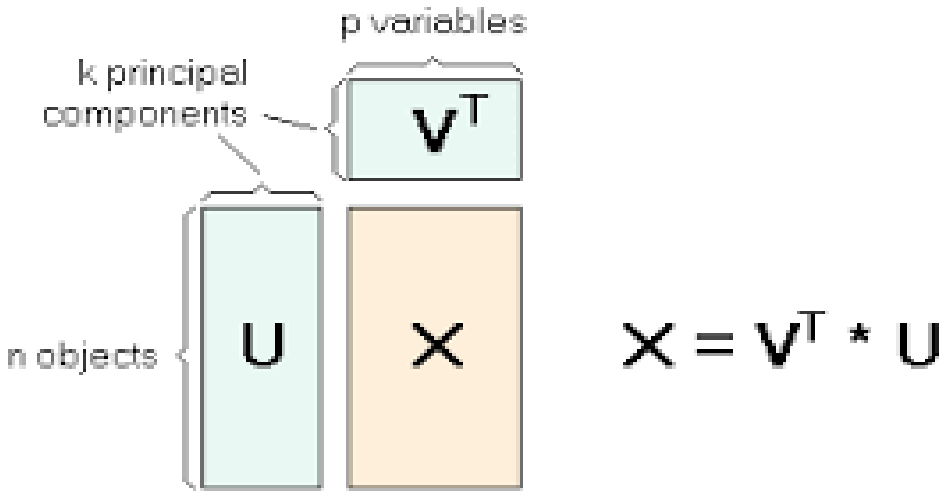
- **Karl Pearson**





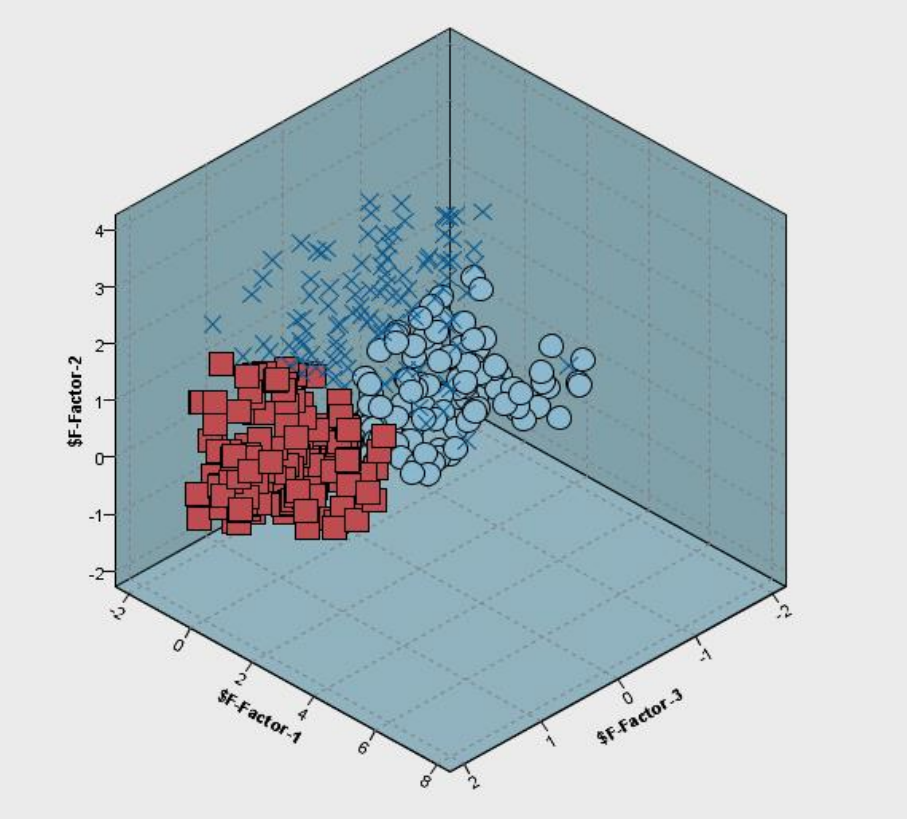
ALGORITMOS MACHINE LEARNING: REDUCCIÓN DE DIMENSIÓN

- **Objetivo:** Dada una matriz de datos de dimensiones nxp que representa los valores de p variables en n individuos, investigar si es posible representar los individuos mediante k variables (k<p) con poca (o ninguna si es posible) pérdida de información.

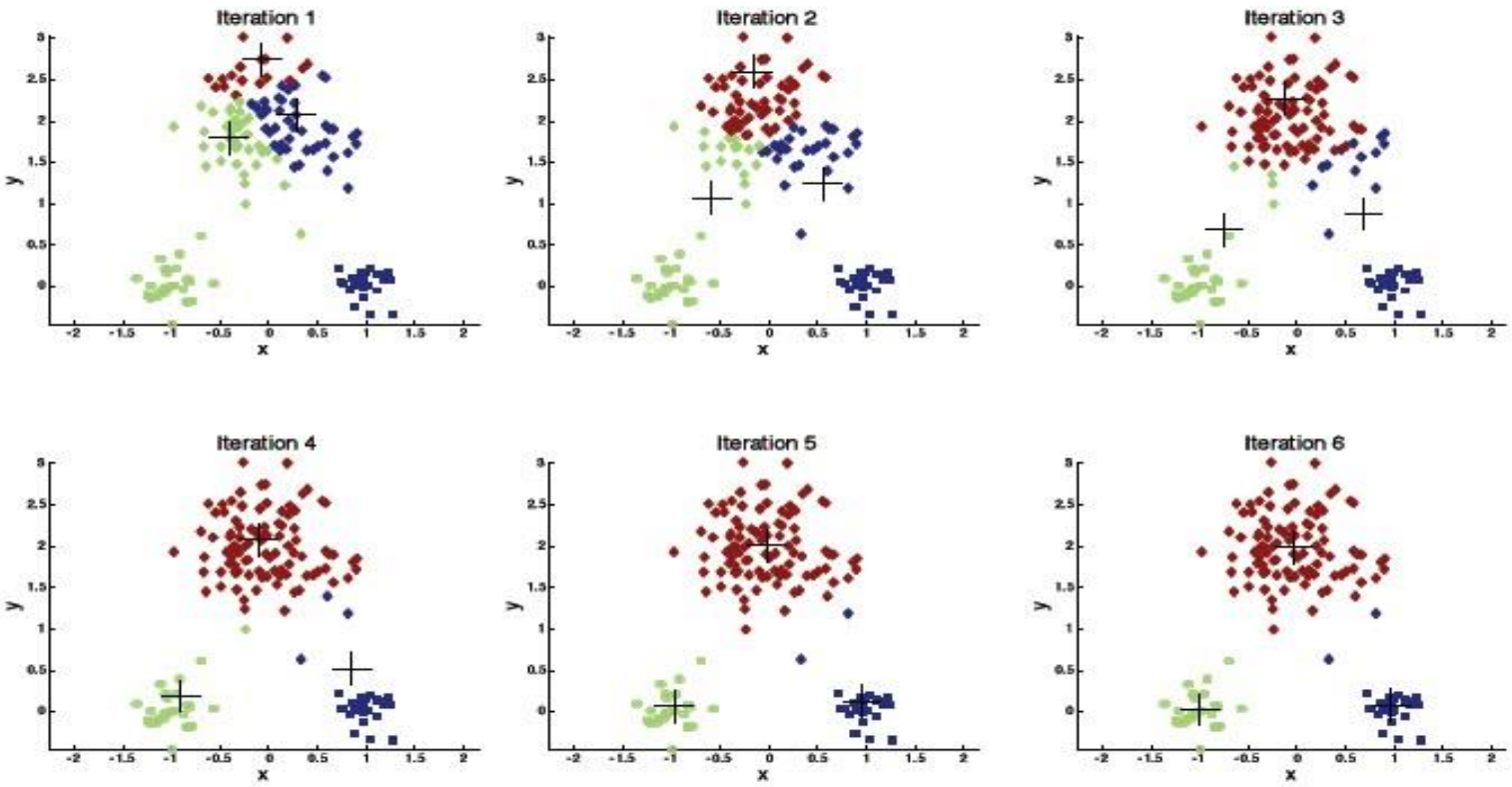


Algoritmos machine learning: Clustering

ALGORITMO DE K - MEANS



Método de K - Means



Teorema: Igualdad de Fisher

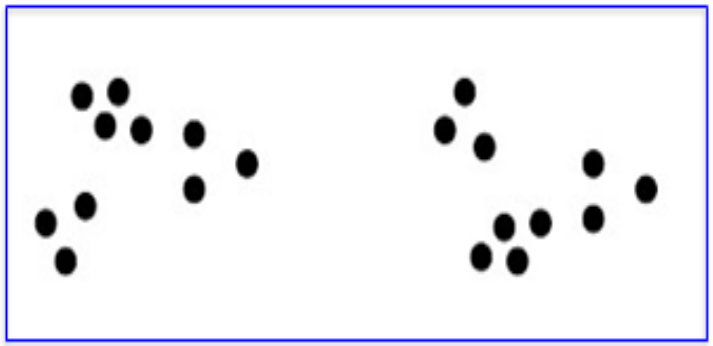
Inercia total = Inercia Inter - clases

+

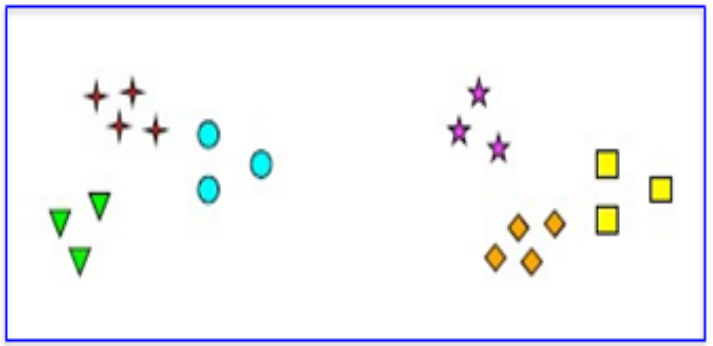
Inercia Intra-clases

$$I = B(P) + W(P)$$

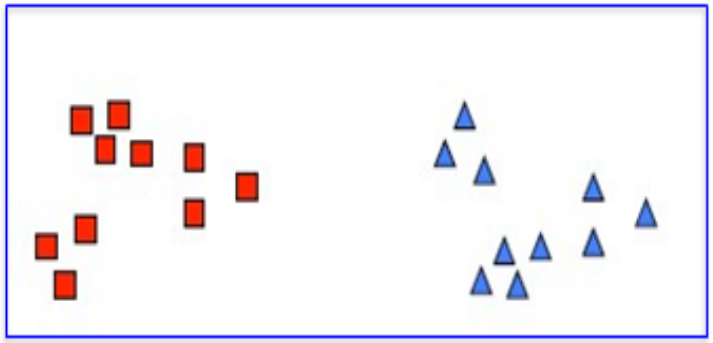
¿ Cuántos clústeres?



Datos originales



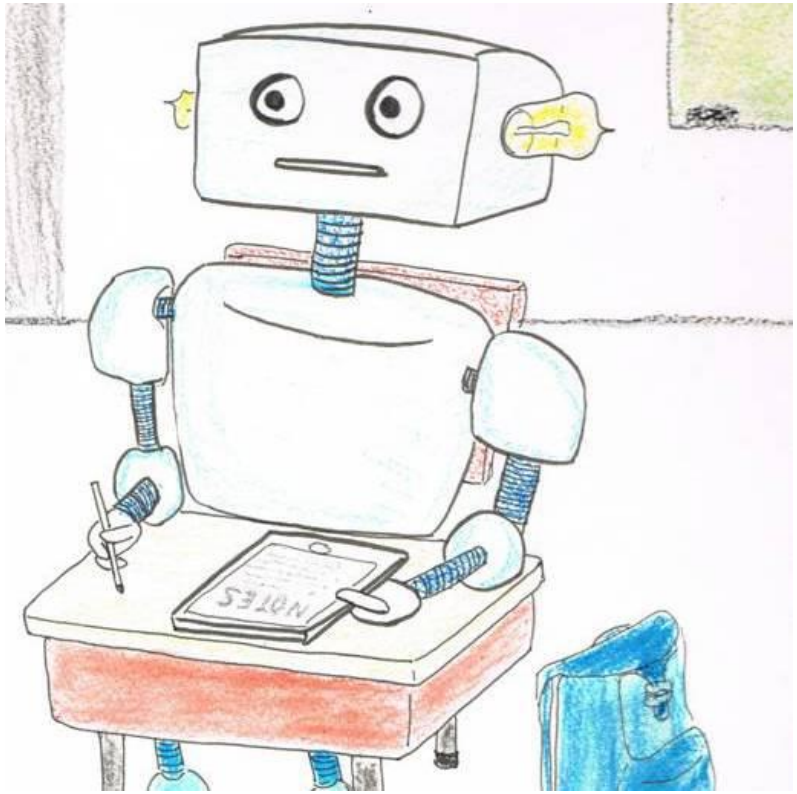
6 clústeres



2 clústeres

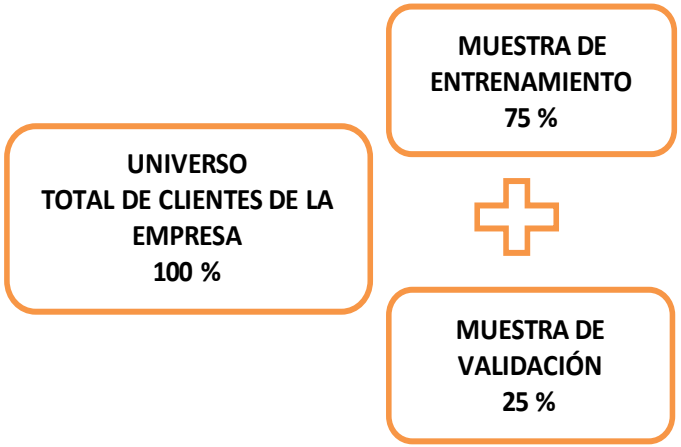
4 clústeres

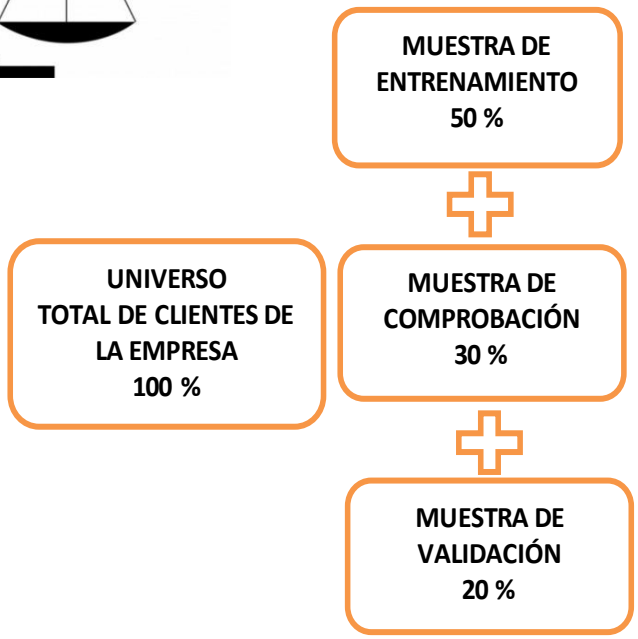
Evaluando un Algoritmo de Machine Learning



Evaluando un Algoritmo de Machine Learning

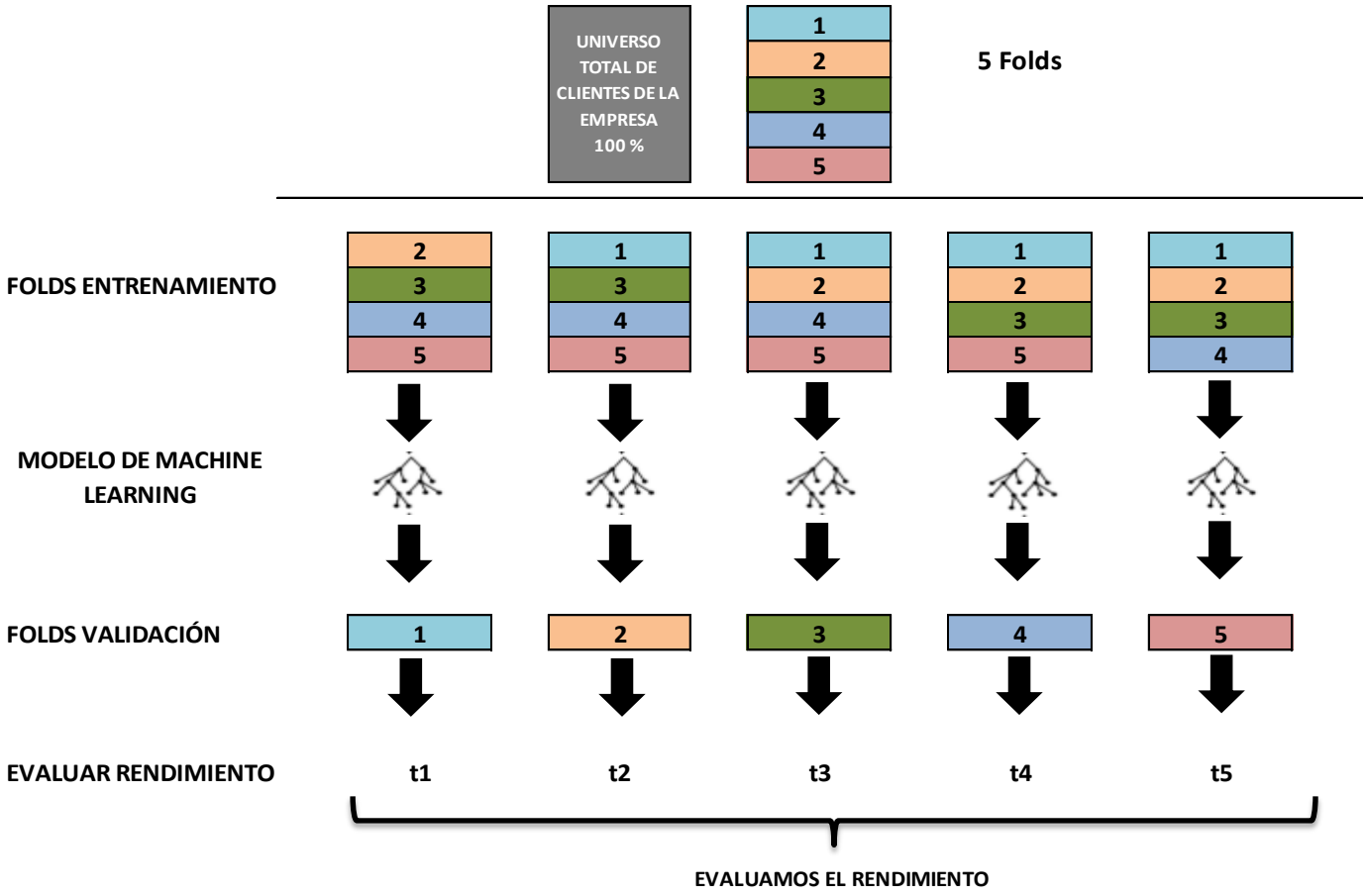
Muestra de entrenamiento y validación





Evaluando un Algoritmo de Machine Learning

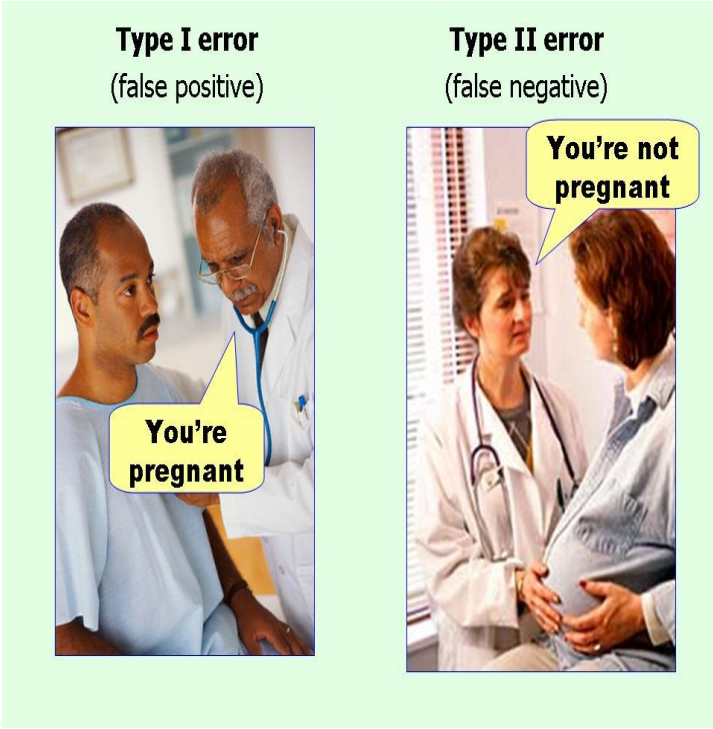
VALIDACIÓN CRUZADA



Evaluando un Algoritmo de Machine Learning

MATRIZ DE CONFUSIÓN Y MATRIZ DE COSTOS

MATRIZ DE CONFUSIÒN		PREDICCIÒN	
		NO MOROSOS	MOROSOS
REALIDAD	NO MOROSOS	DECISIÓN CORRECTA VN	**FP**
	MOROSOS	**FN**	DECISIÓN CORRECTA VP



PRECISIÓN = (VN + VP) / (VN + VP + FP + FN)

SENSIBILIDAD = VP / (VP + FN)

ESPECIFICIDAD = VN / (VN + FP)

F-SCORE = 2 *((VP/ VP + FP) * (VP/ VP + FN)) / ((VP/ VP + FP) + (VP/ VP + FN))

Aplicación de Machine Learning

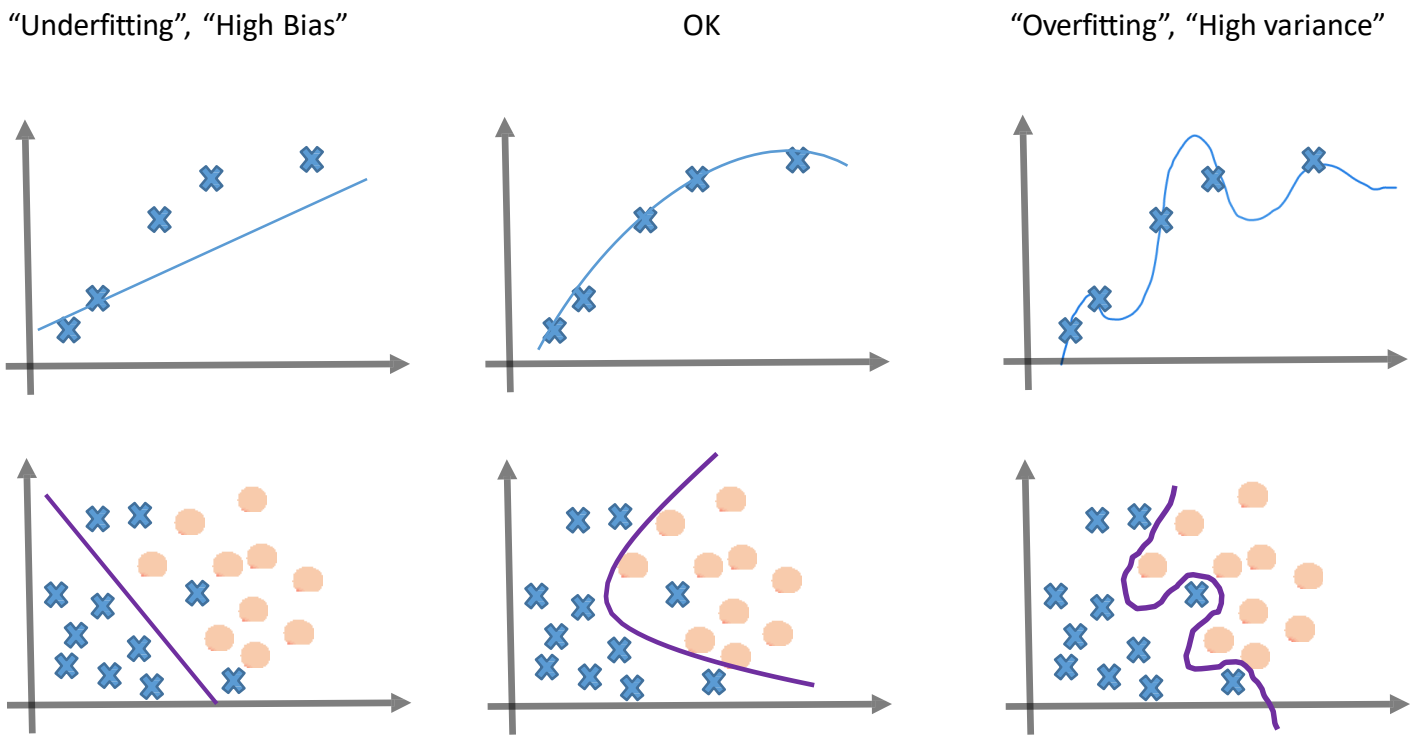
Caso práctico: Clasificación del cáncer

- Casos en los que el nº de ejemplos negativos es mucho mayor que el de ejemplos positivos
- Ejemplo:
  - Modelo regresión logística $y = 1$ $cáncer$; $y = 0$ $no\ cáncer$
  - Se tiene un 1 % de error en el set de test (99 % de diagnósticos correctos)
  - Sólo el 0,5 % de los pacientes tiene cáncer

Exactitud vs. Precisión (Accuracy vs. Precision)

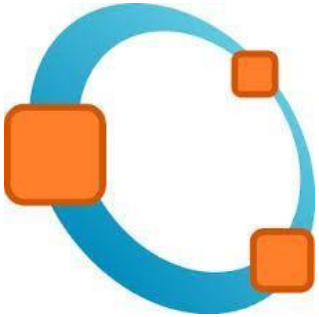
El problema del sobreajuste y Sobregeneralización

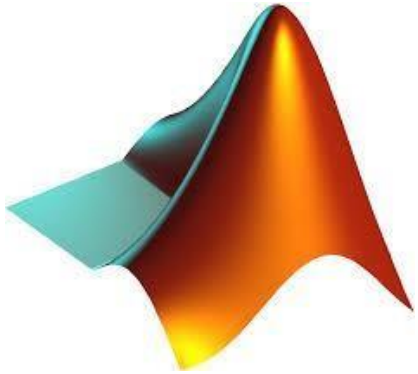
- La sobregeneralización (Underfitting) se da cuando un modelo es demasiado simple y no se ajusta a los datos de entrenamiento.
- El sobreajuste (Overfitting) aparece cuando un modelo es muy complejo y se ajusta demasiado bien a los ejemplos de entrenamiento pero mal a los de test (no generaliza)



Panorama tecnológico : Softwares Machine Learning

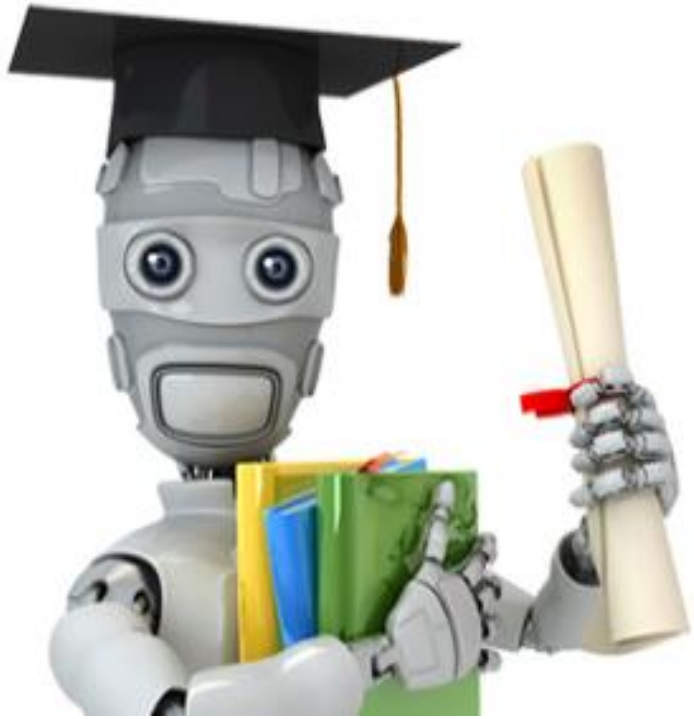
SPSS Modeler



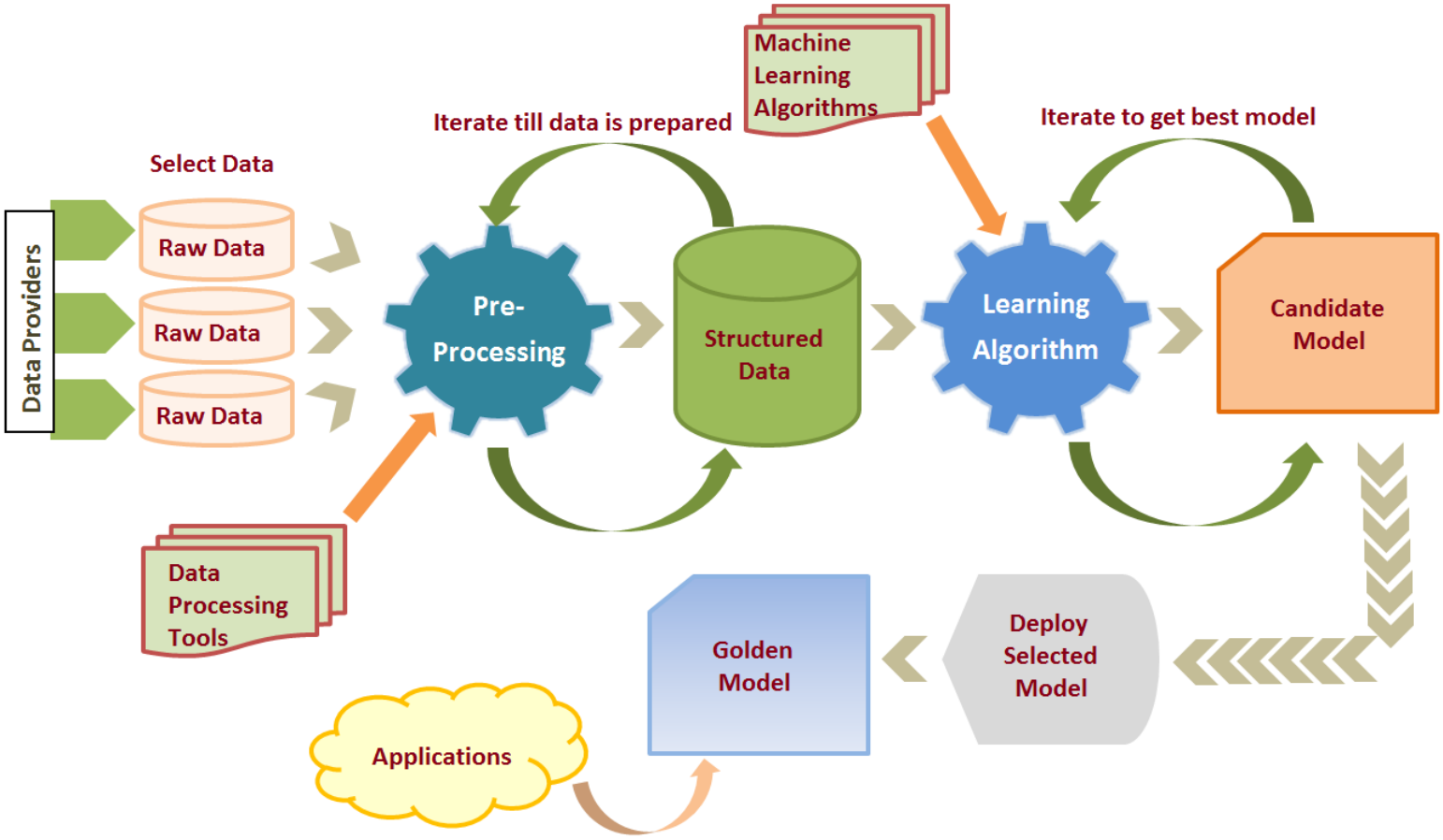


Desarrollo de Algoritmos de Machine Learning

70 %



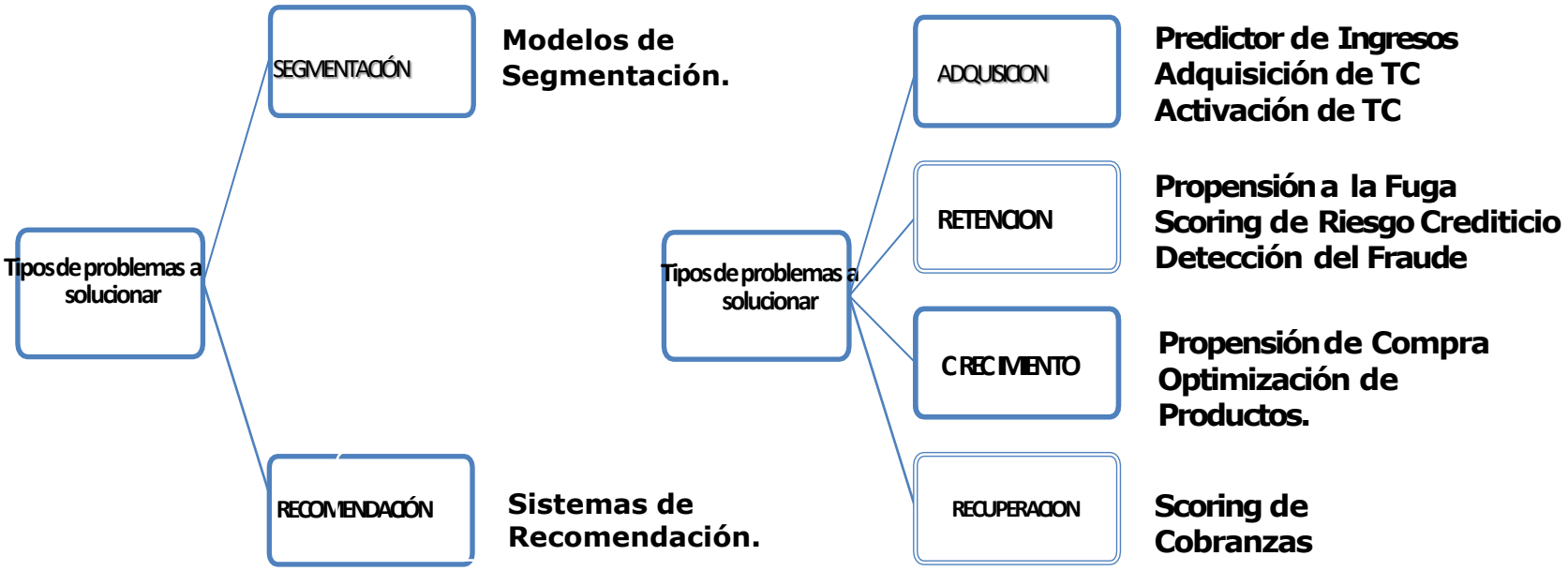
Machine learning en los negocios



ENTENDIMIENTO DEL PROBLEMA

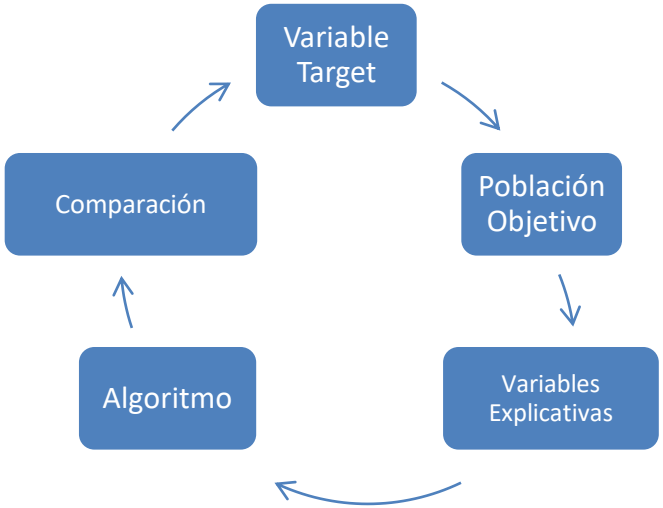
PROPÓSITO DEL ANÁLISIS

Descubrir eventos o resultados futuros en base al conocimiento previo de los datos, utilizando para ello métodos estadísticos, matemáticos, computacionales y de base de datos, así como de la aplicación de los algoritmos de machine learning. En cualquier negocio el éxito depende de:

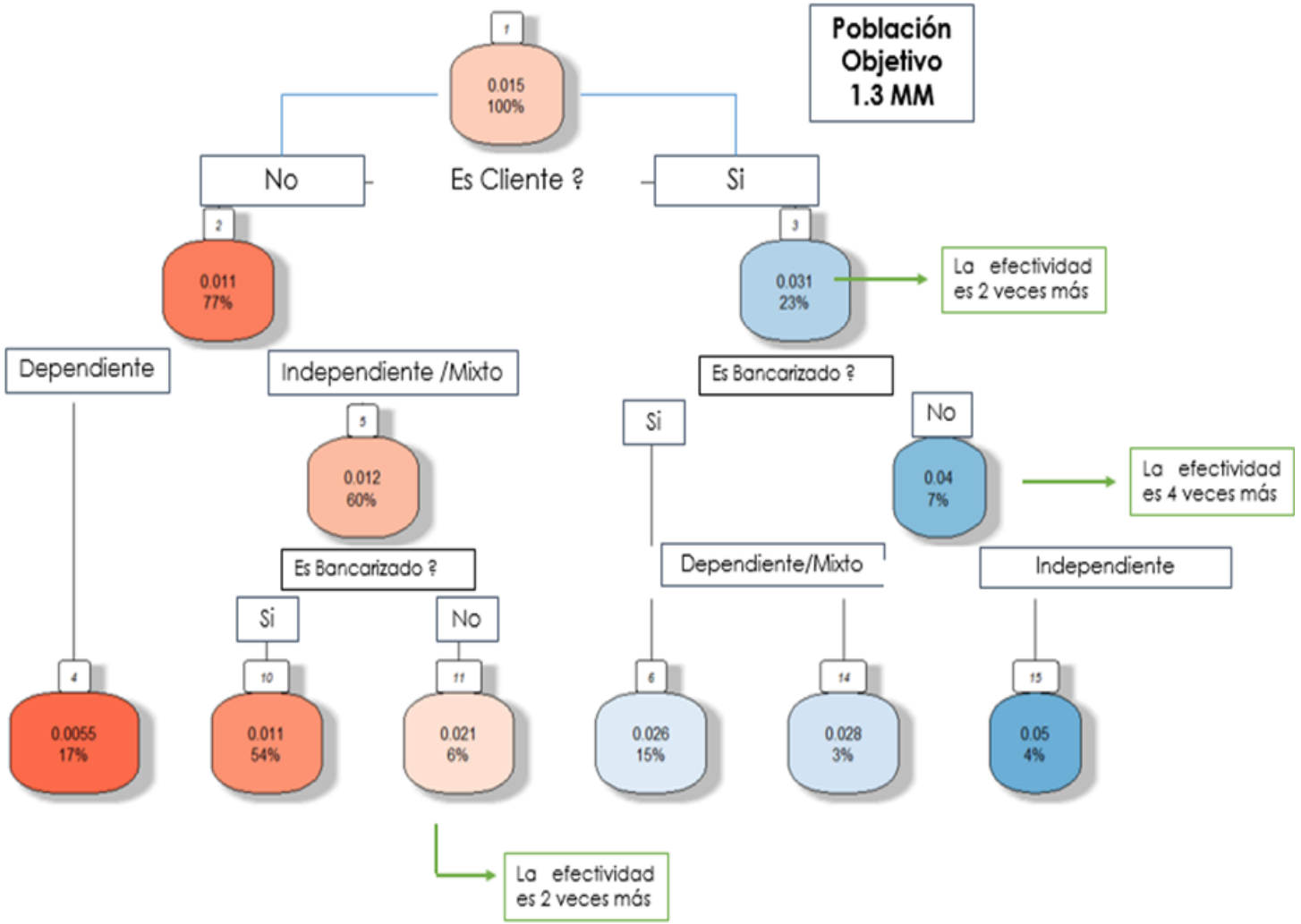


ENTENDIMIENTO DEL PROBLEMA Y POSIBLE SOLUCIÓN

1.¿Que problema quiero solucionar?	Variable objetivo o de respuesta (Y)	Ejem: Estimar ingresos de personas no bancarizados
2.¿Con qué población analizo el problema?	Población objetivo	Ejem: Dependientes e Independientes
3.¿Qué indicadores pueden explicar el problema?	Covariables (Xs)	Ejem: NSE (Reniec), Tipo de automovil (Sunarp)
4. ¿Qué técnica estadistica o informática se ajusta al análisis?	Métrica o algoritmo	Ejem: Arboles de decisión



SEGMENTACIÓN DE LA PO

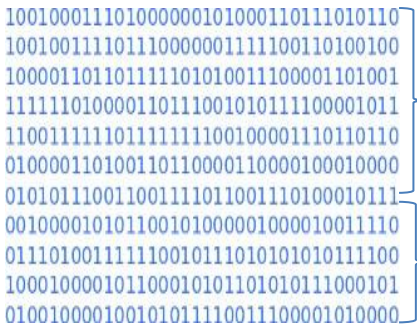


Criterios de Inclusión y Exclusión de la Información

1. Data Original

Criterios de exclusión

2. Data de Estudio

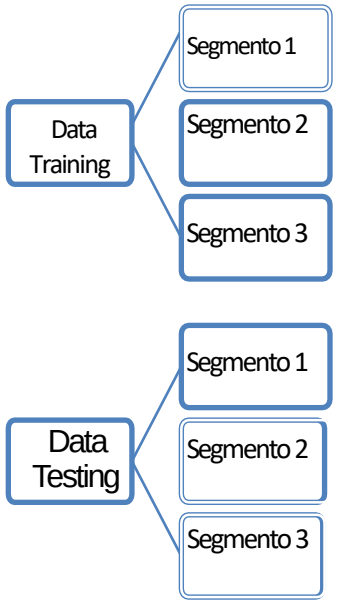


Training

Testing

Información pulida y lista para el proceso de modelación

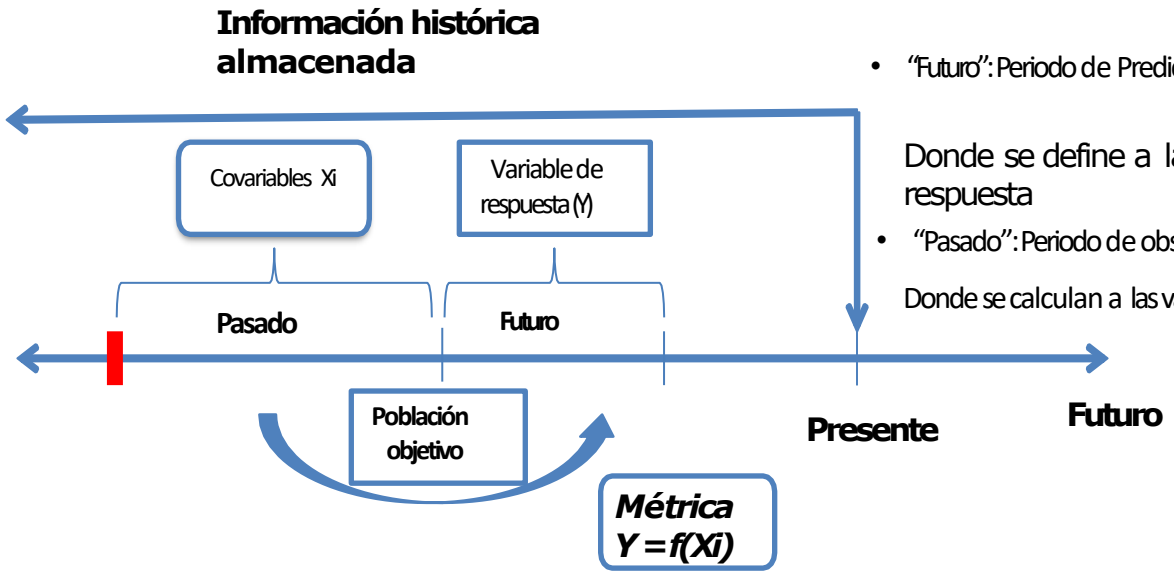
3. Segmentación



- ✓ Tienen poca información histórica
- ✓ Son riesgosos para el negocio
- ✓ Están fuera de las políticas o de las estrategias del negocio
- ✓ Son datos erróneos
- ✓ Son casos especiales que no se volverán a recopilar
- ✓ Ciclos económicos
- ✓ Temas regulatorios / legales

Definición de la Variable Target y horizonte Temporal

Enfoque temporal:



- "Futuro": Periodo de Predicción o Performance

Donde se define a la variable de respuesta

- "Pasado": Periodo de observación

Donde se calculan a las variables explicativas

Enfoque matricial:

ID	Segment_Target	Var_Target	Var_X1	Var_X2	Var_X3	Var_X4	Var_X5	Var_X6
1	Segment 1	1	-0.243257655	216	952.4800	1	4	3
2	Segment 2	1	1.696358794	191	633.4949	0	7	4
3	Segment 3	1	0.561226988	192	637.5107	0	6	3
4	Segment 1	1	-1.673888687	205	927.2513	0	8	3
5	Segment 2	0	-0.315746538	200	988.0877	0	2	3
6	Segment 3	0	0.402197729	201	927.5218	1	6	2
7	Segment 1	1	0.668736379	202	582.0028	0	6	2
8	Segment 2	1	1.489475004	197	701.1748	0	6	2
9	Segment 3	0	0.308647509	201	526.3747	0	8	4
10	Segment 1	1	0.090616380	189	989.2571	0	7	4
11	Segment 2	1	0.081223506	200	789.0298	0	8	2
12	Segment 3	1	-0.443663814	207	937.3809	0	2	2
13	Segment 1	1	-1.416088194	220	819.6118	0	9	1
14	Segment 2	1	-0.316298576	187	995.7736	1	2	5

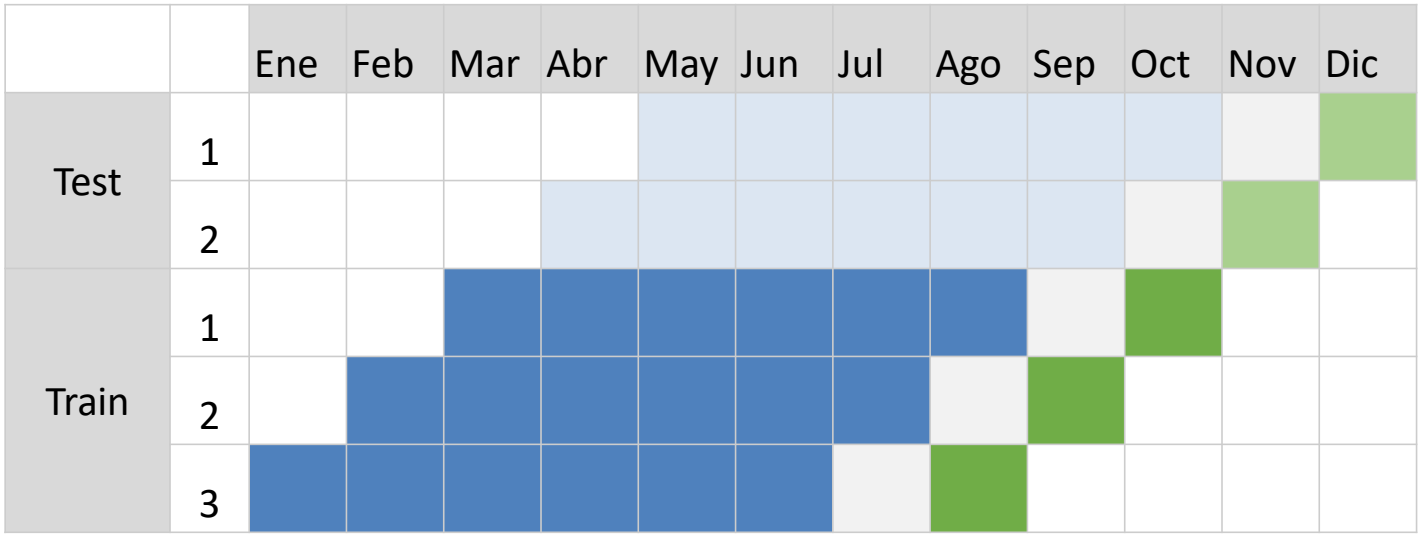
Población objetivo

Variable de respuesta (Y)

Covariables Xi

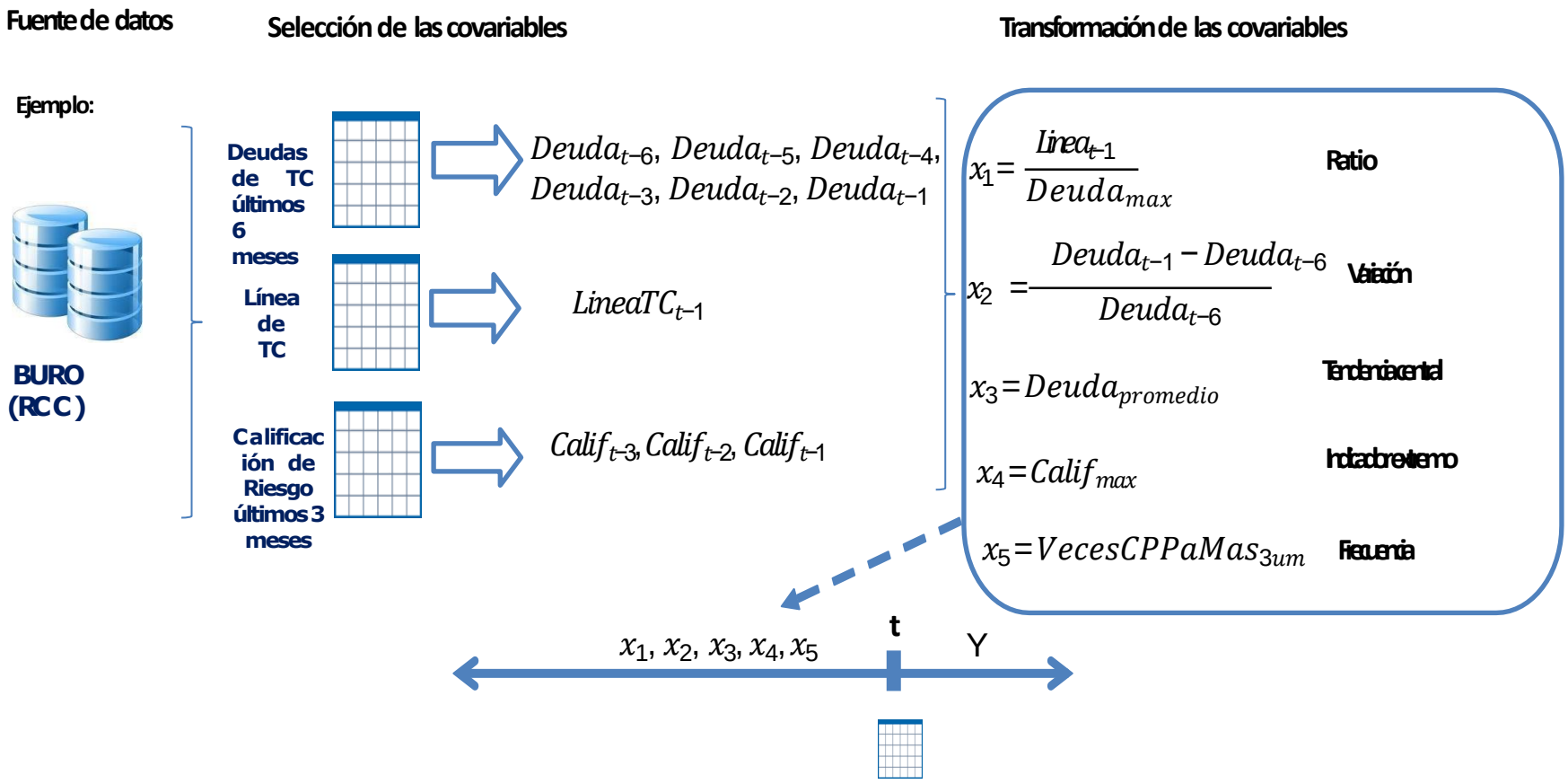
Métrica
Var_Target= f(Var_X1, Var_X2, Var_X3, Var_X4, Var_X5, Var_X6)

Periodicidad , Cosechas y Matrices de Transición



Definición y Creación de Drivers

Las variables a seleccionar para la solución del problema propuesto deben tener **sentido para el negocio**. En otras palabras al seleccionarlas se espera que estén correlacionadas con la variable de respuesta del modelo. La transformación tiene como propósito optimizar el aporte de las Xi en el modelo.



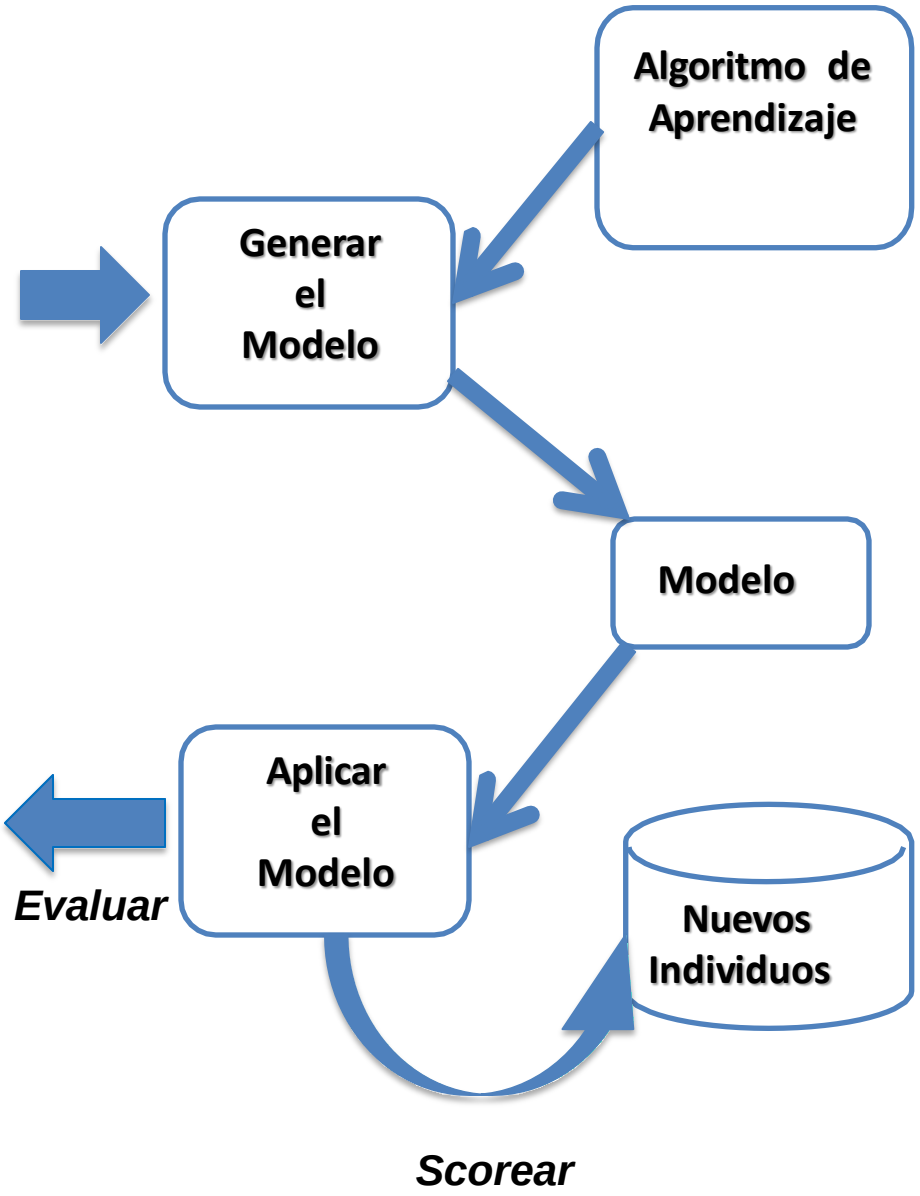
Modelo general de Implementación de Modelos

ID	REEMBOLSO	ESTADO CIVIL	INGRESOS ANUALES	FRAUDE
1	SI	SOLTERO	S/ 1,000	NO
2	SI	CASADO	S/ 5,000	NO
3	NO	CASADO	S/ 3,500	SI
4	SI	VIUDO	S/ 4,500	NO
5	NO	SOLTERO	S/ 2,000	NO
6	NO	SOLTERO	S/ 1,500	SI

Tabla de Aprendizaje

ID	REEMBOLSO	ESTADO CIVIL	INGRESOS ANUALES	FRAUDE
7	SI	SOLTERO	S/ 4,000	NO
8	SI	CASADO	S/ 5,500	NO
9	NO	CASADO	S/ 6,500	SI

Tabla de Testing



METODOLOGÍA DE DESARROLLO DE ALGORITMOS PREDICTIVOS

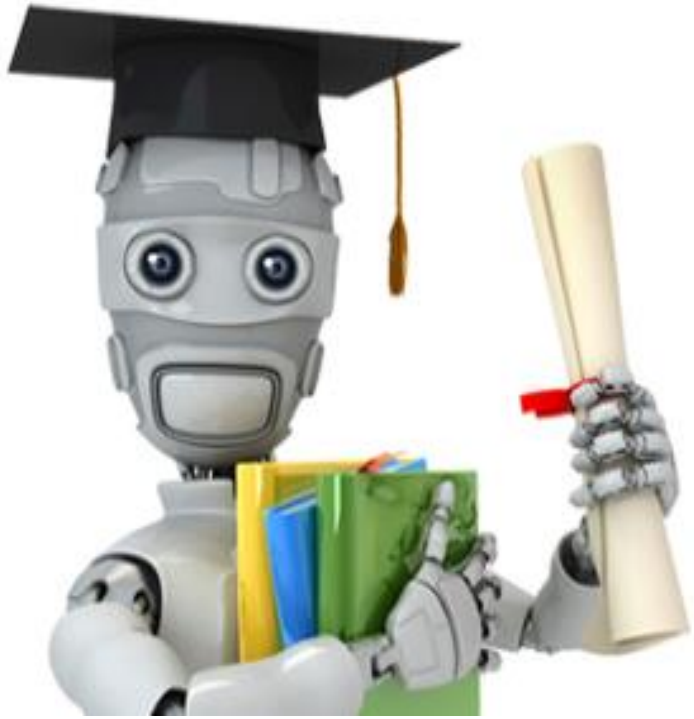
- Definición y diseño del modelamiento a realizar.
- Análisis exploratorio de datos:

  - ✓ Detección de Outliers.
  - ✓ Imputación de valores perdidos.
  - ✓ Transformaciones.
  - ✓ Recodificaciones.

- Balanceo de datos.
- Selección de variables. (Met. Estadísticas vs ML).
- Modelamiento.
- Validación .
- Implementación.

Implementación de Algoritmos de Machine Learning

30 %

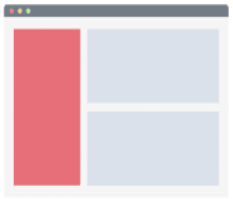


Implementación de Modelos : Generación de Grupos de Ejecución

Probabilidad	N° Clientes	N° Sucesos VD	% Sucesos VD/ N° Clientes
0,9	10 000	100	36%
0,8	10 000	60	22%
0,7	10 000	40	14%
0,6	10 000	33	12%
0,5	10 000	20	7%
0,4	10 000	10	4%
0,3	10 000	5	2%
0,2	10 000	5	2%
0,1	10 000	3	1%
0	10 000	3	1%
Total	100 000	279	100%

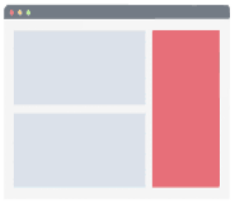
GRUPOS DE EJECUCIÓN	N° CLIENTES	% SUCESOS ACUMULADOS	EFECTIVIDAD	LIFT
RECOMENDADO	**30 000**	**200**	**0,67%**	**2,39**
MEDIO	30 000	63	**0,21%**	0,75
BAJO	40 000	16	**0,04%**	0,14
TOTAL	**100 000**	**279**	**0,28%**	

Implementación de Modelos : A/B Testing



= 16% de Nivel de Morosidad en la Cartera

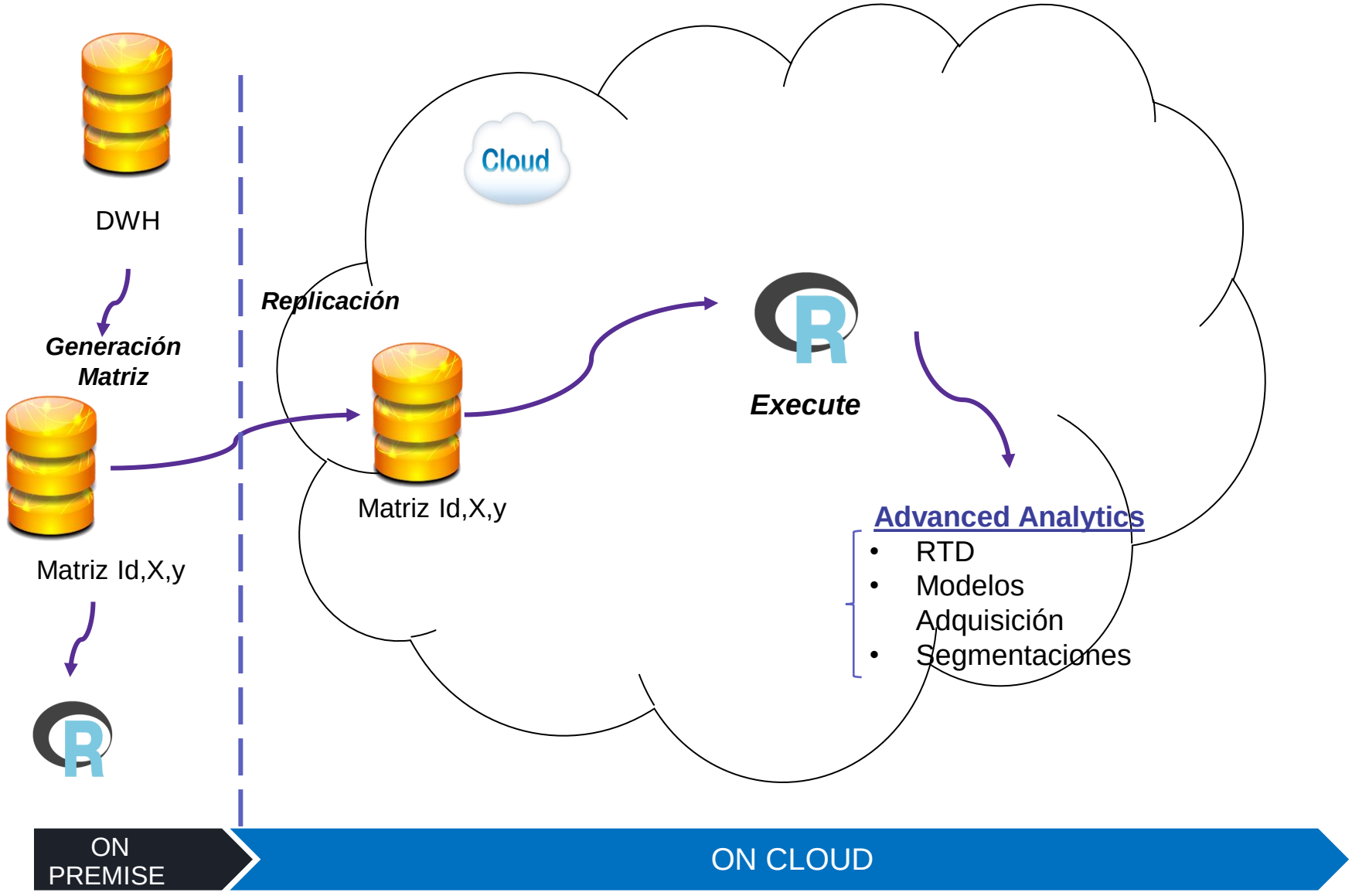
Clientes con Gestión del Modelo de Riesgo



= 22 % de Nivel de Morosidad en la Cartera

Clientes sin Gestión del Modelo de Riesgo

Implementación de Modelos : Infraestructura



¡Gracias!

André Omar Chávez Panduro
UNMSM
MSc in Data Science Candidate
Promotion "Erwin Kraenau Espinal"
Universidad Ricard Palma