MACHINE LEARNING

Tema de la clase: Introducción a las hojas de cálculo

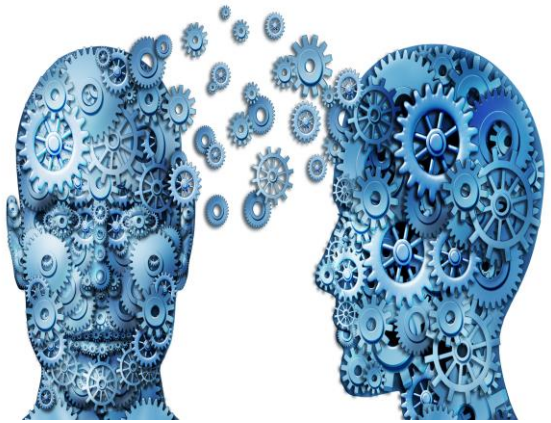
- ¿Qué es machine learning?
- Inteligencia artificial
- Aplicaciones

¿Qué es machine learning?

Machine Learning

Arthur Samuel (1959). Campo de estudio que da a las computadoras la capacidad de aprender sin ser explícitamente programadas.

Tom Mitchell (1998). Problema de aprendizaje bien planteado: Se dice que un programa de computadora aprende de la experiencia E con respecto a alguna tarea T y alguna medida de rendimiento P, si su desempeño en T, medido por P, mejora con la experiencia E.



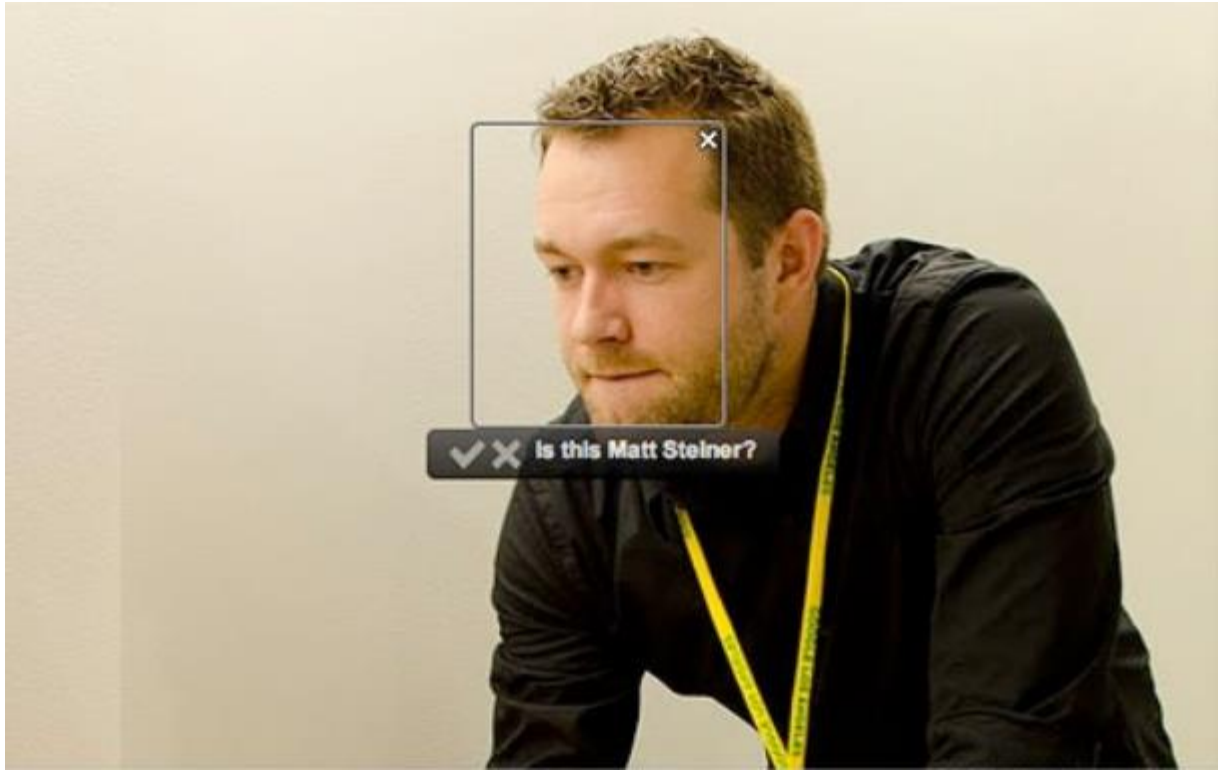
Machine learning

Ciertas tareas son extremadamente difíciles de programar a mano:

Filtrado de spam

Machine learning

Ciertas tareas son extremadamente difíciles de programar a mano:



Reconocimiento facial

Machine learning

Máquina traductora

Google Translate

This text has been automatically translated from Arabic:

Moscow stressed tone against Iran on its nuclear program. He called Russian Foreign Minister Tehran to take concrete steps to restore confidence with the international community, to cooperate fully with the IAEA. Conversely Tehran expressed its willingness

Translate text

شددت موسكو لهجتها ضد إيران بشأن برنامجها النووي. ودعا وزير الخارجية الروسي طهران إلى اتخاذ خطوات ملموسة لاستعادة الثقة مع المجتمع الدولي والتعاون الكامل مع الوكالة الذرية. بالمقابل أبدت طهران استعدادها لاستئناف السماح بعمليات التفتيش المفاجئة بشرط إسقاط مجلس الأمن ملفها النووي.

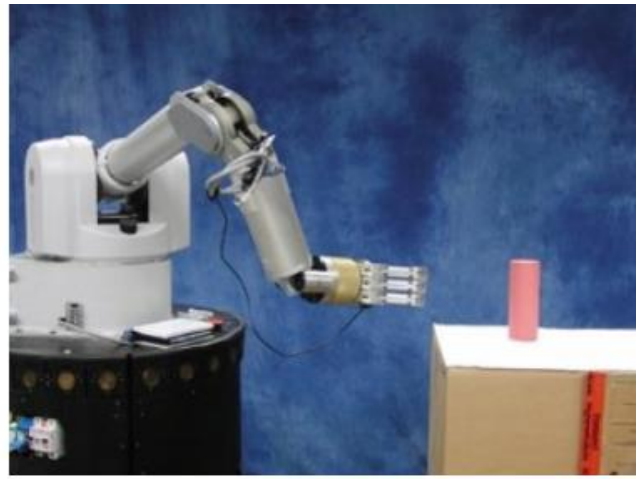
from Arabic to English BETA | Translate

Reconocimiento de voz

Machine learning

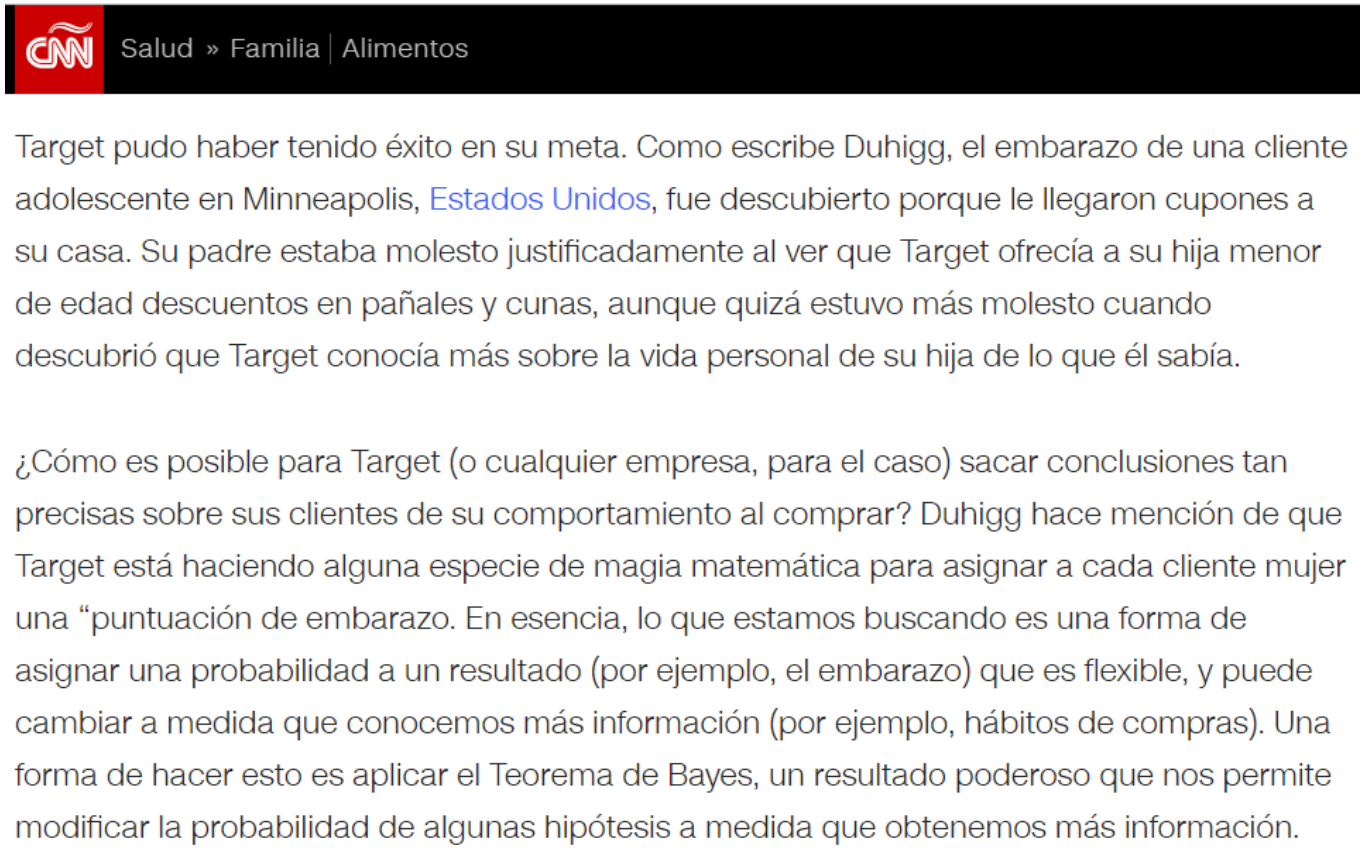
Minería de datos

Movimiento robot



Machine Learning

¿Cómo se entera una tienda antes que tus padres de que estás embarazada?



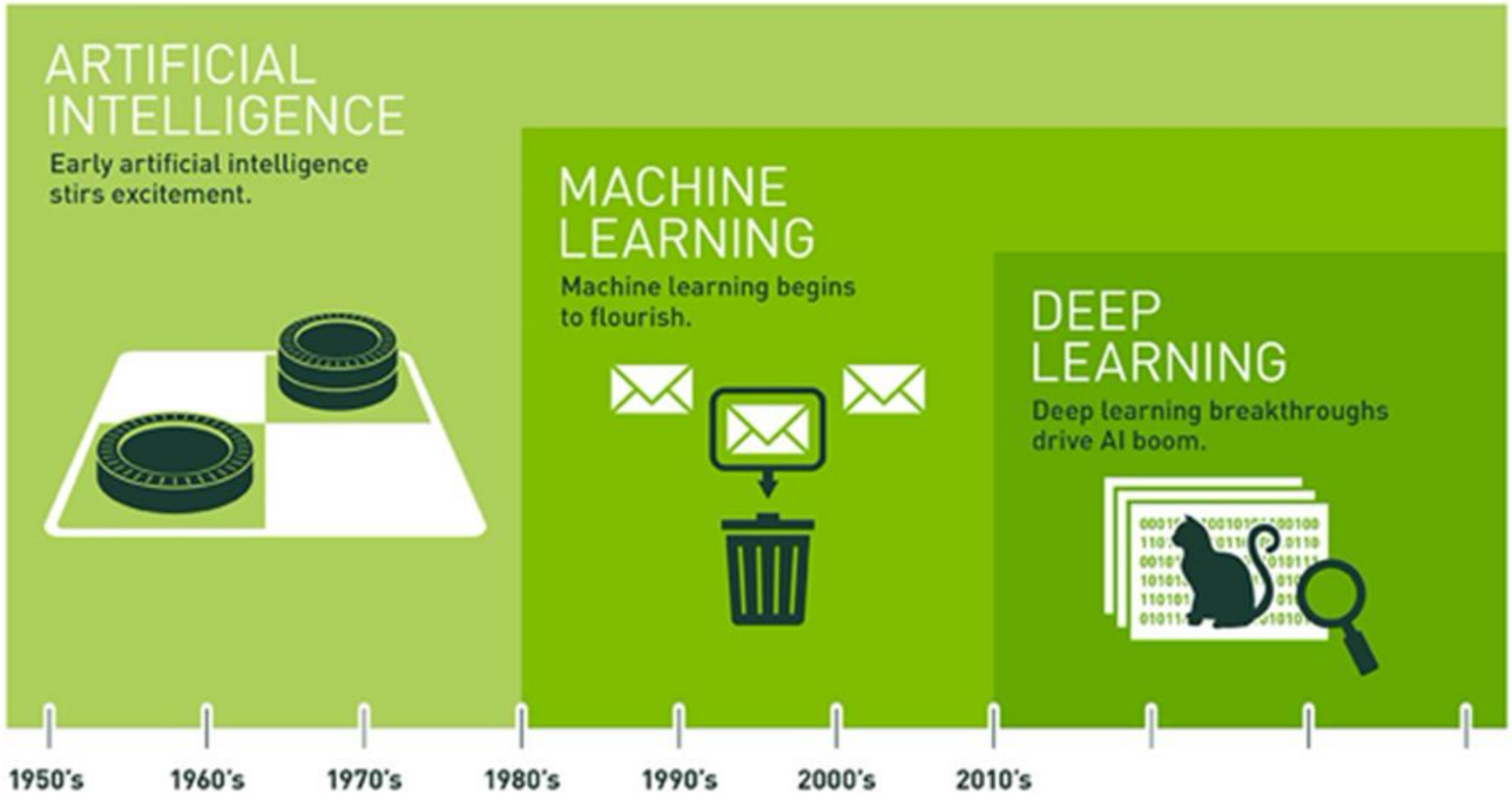
CNN Salud » Familia | Alimentos

Target pudo haber tenido éxito en su meta. Como escribe Duhigg, el embarazo de una cliente adolescente en Minneapolis, Estados Unidos, fue descubierto porque le llegaron cupones a su casa. Su padre estaba molesto justificadamente al ver que Target ofrecía a su hija menor de edad descuentos en pañales y cunas, aunque quizá estuvo más molesto cuando descubrió que Target conocía más sobre la vida personal de su hija de lo que él sabía.

¿Cómo es posible para Target (o cualquier empresa, para el caso) sacar conclusiones tan precisas sobre sus clientes de su comportamiento al comprar? Duhigg hace mención de que Target está haciendo alguna especie de magia matemática para asignar a cada cliente mujer una "puntuación de embarazo. En esencia, lo que estamos buscando es una forma de asignar una probabilidad a un resultado (por ejemplo, el embarazo) que es flexible, y puede cambiar a medida que conocemos más información (por ejemplo, hábitos de compras). Una forma de hacer esto es aplicar el Teorema de Bayes, un resultado poderoso que nos permite modificar la probabilidad de algunas hipótesis a medida que obtenemos más información.

https://cnnespanol.cnn.com/2012/04/23/como-se-entera-una-tienda-antes-que-tus-padres-de-que-estas-embarazada/

Inteligencia Artificial, Machine Learning y Deep Learning



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Machine learning

Idea general:

- Recopilar datos para nuestro problema.
- Use estos datos para aprender cómo resolver la tarea.

Ventajas clave:

- Puede resolver robustamente tareas complejas
- Confianza en los datos del mundo real en lugar de pura intuición.
- Puede adaptarse a nuevas situaciones (recopilar más datos)

Machine learning

- Machine learning se define como un proceso automatizado que extrae patrones de los datos. Para construir los modelos utilizados en aplicaciones de análisis de datos predictivos o descriptivos.

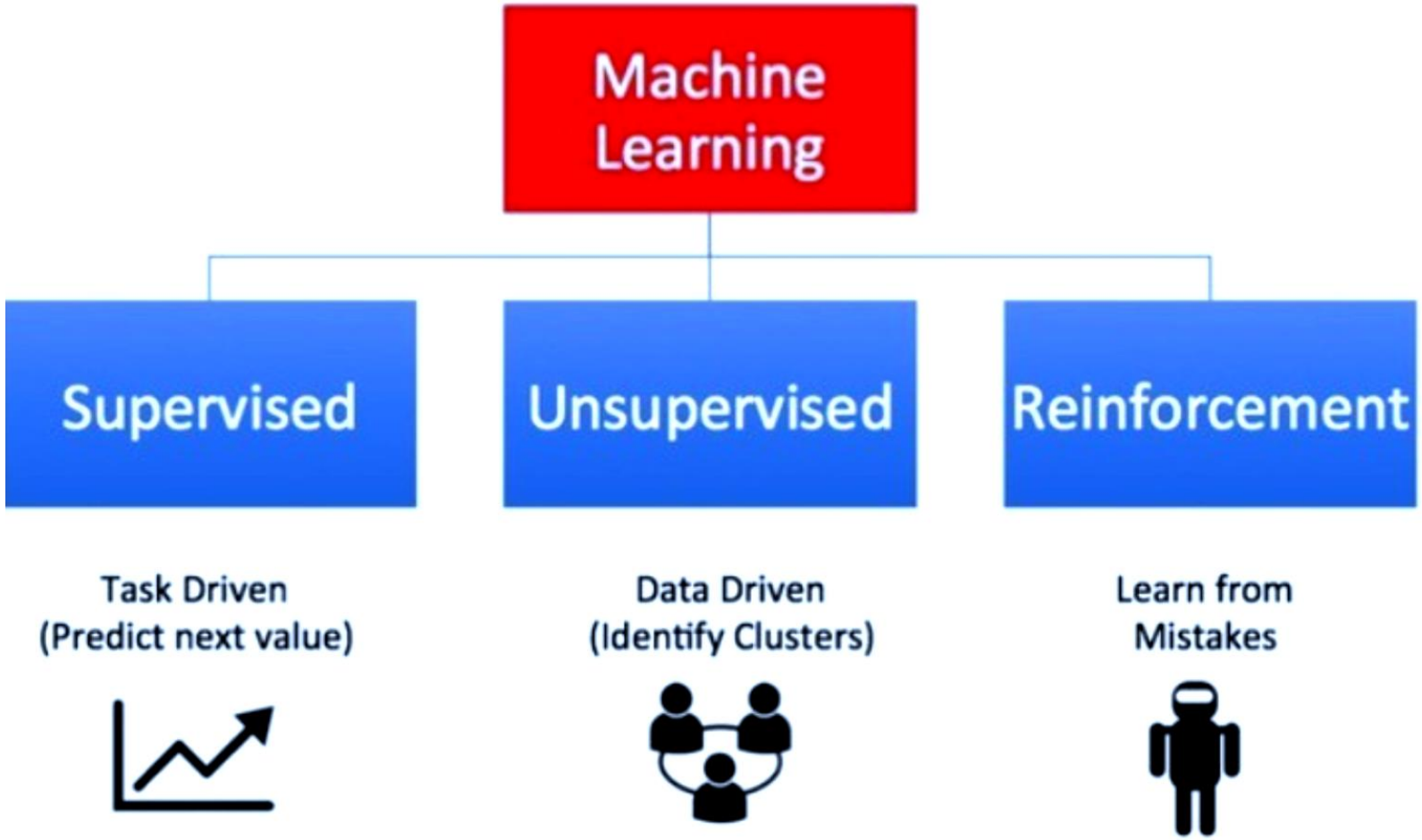
Machine learning

¿cómo aprende este mapeo (patrones)?

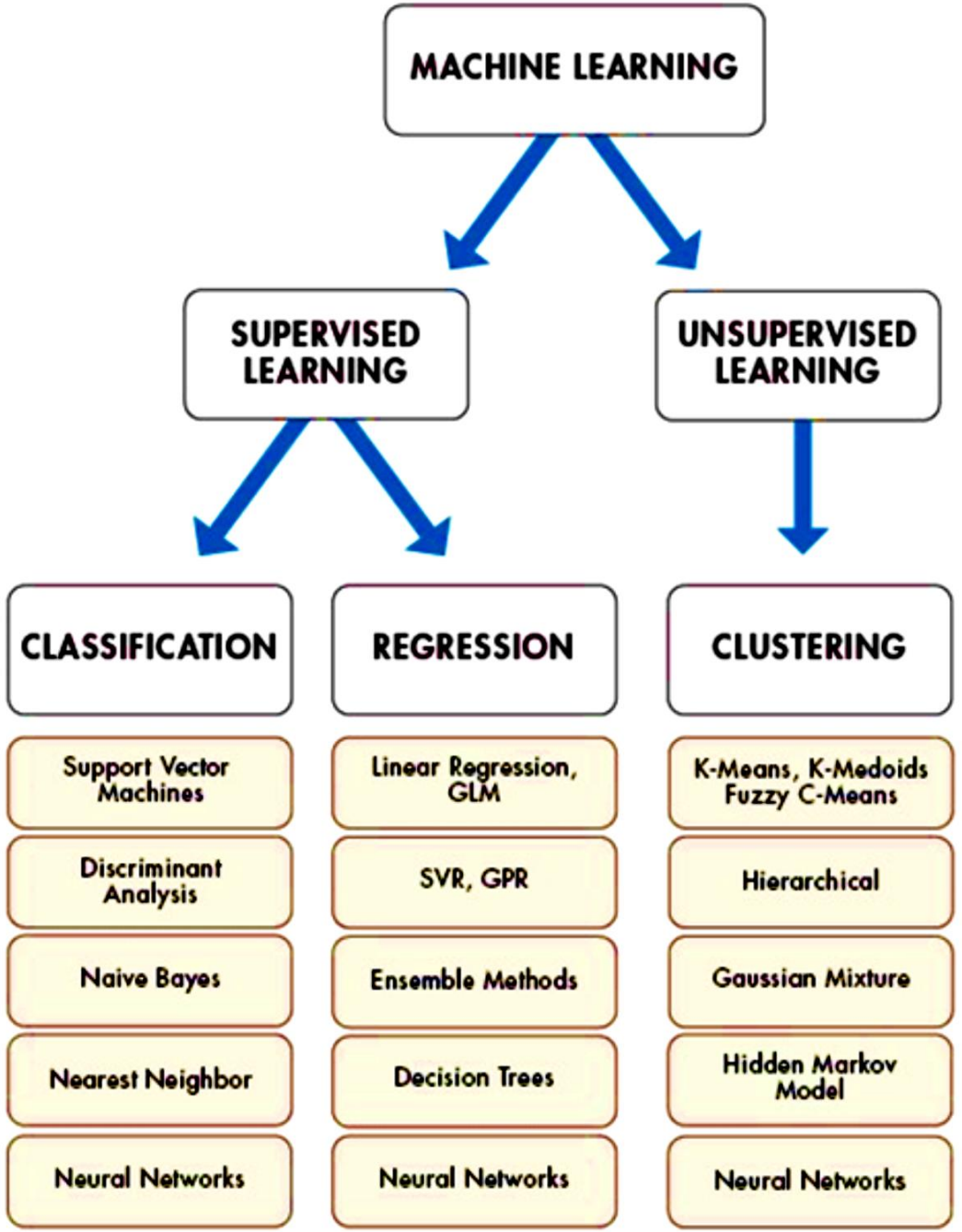
El método de aprendizaje depende del tipo de datos que tengamos a nuestra disposición.

- Podemos tener ejemplos de datos donde tenemos tanto las entradas como las salidas: (i, o)
- Para algunos datos, solo tenemos las entradas i
- A veces no tenemos acceso directo a la salida «correcta», pero podemos obtener alguna medida de la calidad de una salida **o**, después de la entrada **i**

Machine learning



Machine learning

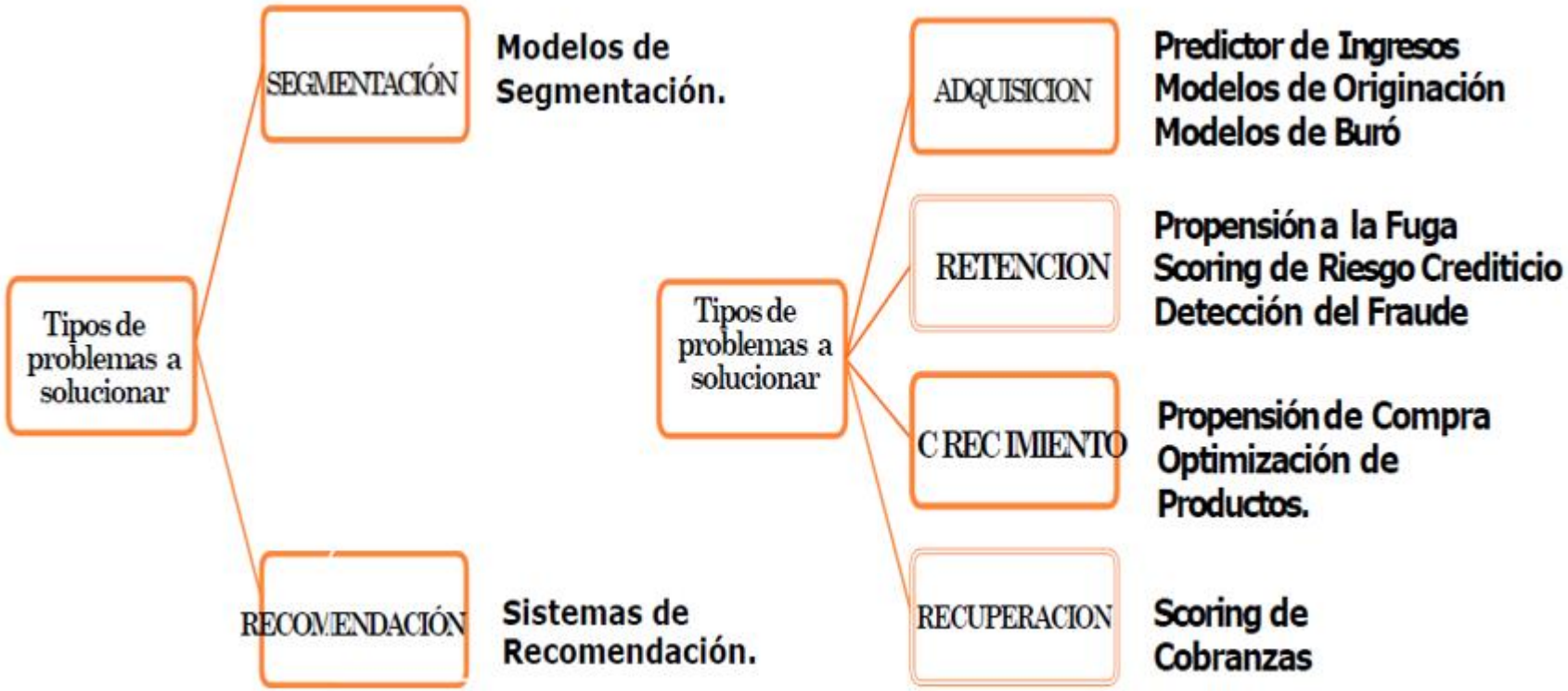


Proceso ML

1. ENTENDIMIENTO DEL NEGOCIO

ENTENDIMIENTO DEL NEGOCIO

1. PROPÓSITO DEL ANÁLISIS



Datos

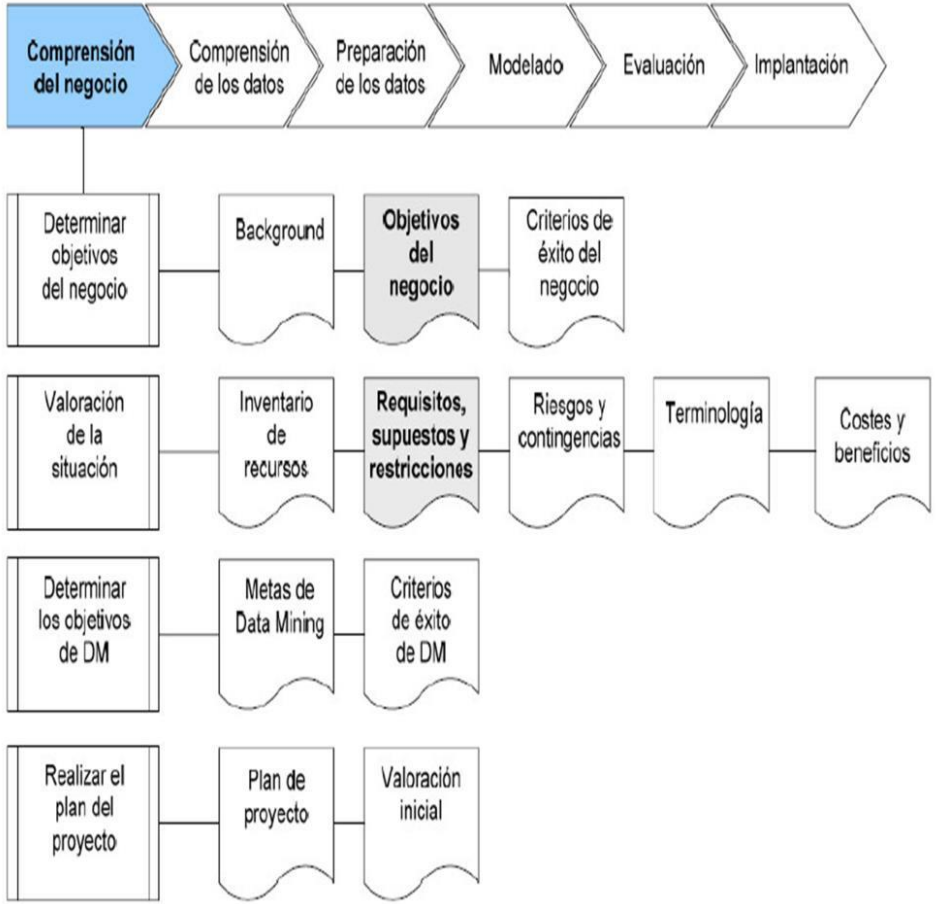
ID	Segment_Target	Var_Target	Var_X1	Var_X2	Var_X3	Var_X4	Var_X5	Var_X6
1	Segment 1	1	-0.243257655	216	952.4800	1	4	3
2	Segment 2	1	1.696358794	191	633.4949	0	7	4
3	Segment 3	1	0.561226988	192	637.5107	0	6	3
4	Segment 1	1	-1.673888687	205	927.2513	0	8	3
5	Segment 2	0	-0.315746538	200	988.0877	0	2	3
6	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

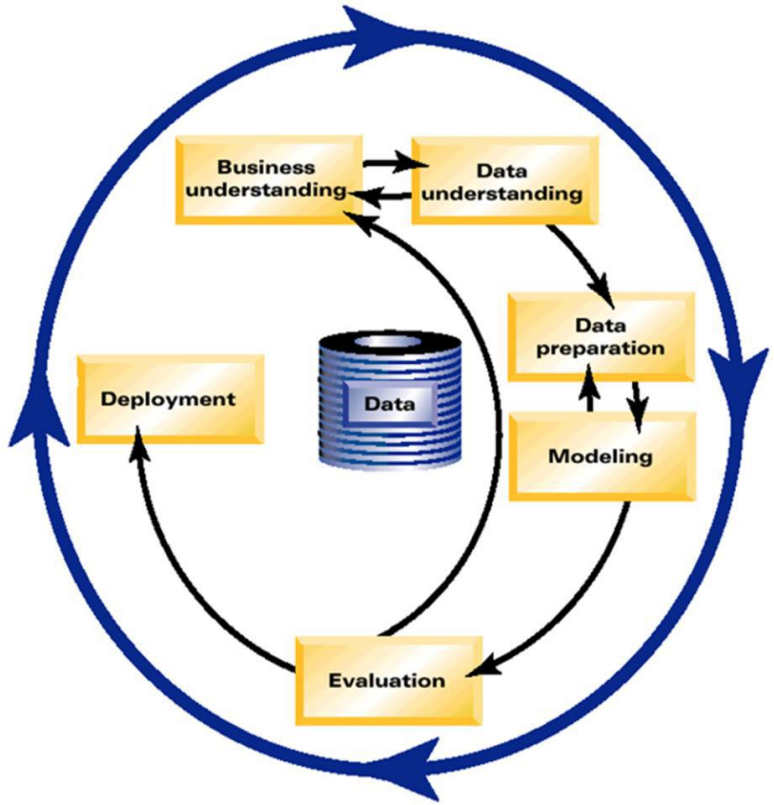
Proceso ML

2. LA ANALÍTICA EN LOS NEGOCIOS

Proceso ML

METODOLOGÍA CRISP - DM





Otras aplicaciones

Deserción Académica

Detección de Fraudes

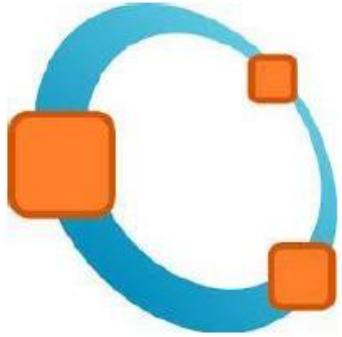
Fuga de Clientes

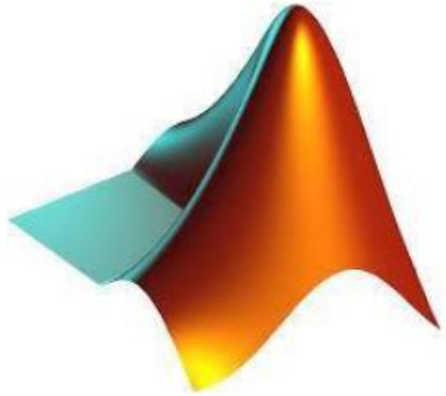
Generar Perfiles o Grupos

Lavado de Activos

Venta Cruzada

PANORAMA TECNOLÓGICO : SOFTWARES MACHINE LEARNING

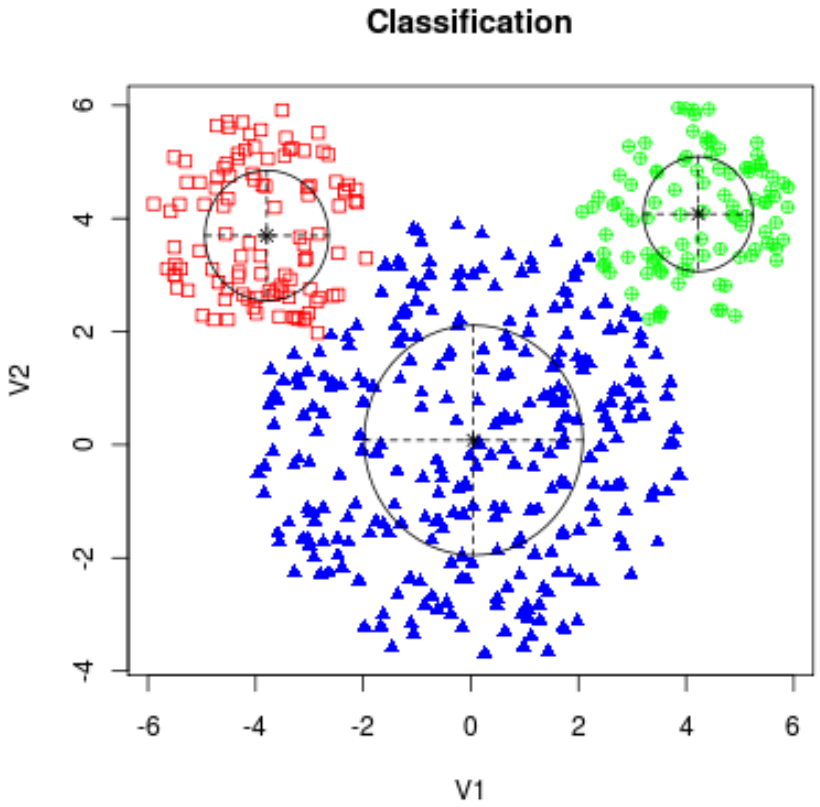




Casos

Clasificación

La clasificación, que es la tarea de asignar objetos a una de varias categorías predefinidas, es un problema generalizado que abarca muchas aplicaciones diversas.



Ejemplo de problemas de clasificación

Tarea	Conjunto de atributos, x	Etiqueta de clase, y
Categorizar mensajes de correo electrónico	Características extraídas del encabezado y contenido del mensaje de correo electrónico	spam o no spam
Identificación de células tumorales	Características extraídas de las imágenes por resonancia magnética	células malignas o benignas
Catalogación de galaxias	Características extraídas de imágenes de telescopio	Galaxias elípticas, espirales o de forma irregular

Clasificación : Definición

La clasificación es la tarea de aprender una función de destino *f* que asigna cada conjunto de atributos *X* a una de las etiquetas de clase predefinidas *Y*.

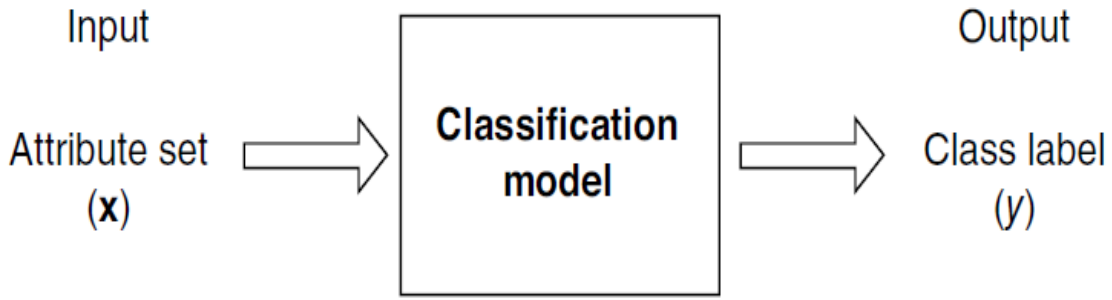
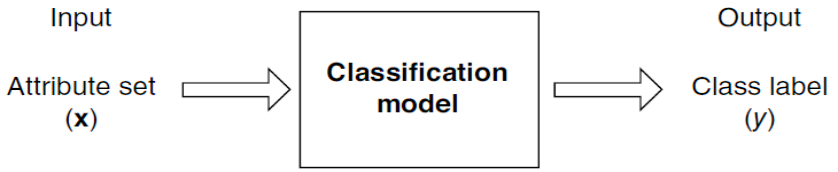


Figura 1: La clasificación es la tarea de mapear un conjunto de atributos de entrada *x* en su etiqueta de clase *y*.

Clasificación



ID	Casa propia	Estado civil	Ingreso anual	Default
1	Si	Soltero	125	No
2	No	Casado	100	No
3	No	Soltero	70	No
4	Si	Casado	120	No
5	No	Divorciado	95	Si
6	No	Casado	60	No
7	Si	Divorciado	220	No
8	No	Soltero	85	Si
9	No	Casado	75	No
10	No	Soltero	90	Si
11	No	Soltero	55	No
12	Si	Casado	80	No
13	Si	Divorciado	110	Si
14	No	Soltero	95	No
15	No	Casado	67	Si

Training Set

ID	Casa propia	Estado civil	Ingreso anual	Default
1	Si	Soltero	125	No
2	No	Casado	100	No
3	No	Soltero	70	No
4	Si	Casado	120	No
5	No	Divorciado	95	Si
6	No	Casado	60	No
7	Si	Divorciado	220	No
8	No	Soltero	85	Si
9	No	Casado	75	No
10	No	Soltero	90	Si

Test Set

ID	Casa propia	Estado civil	Ingreso anual	Default
11	No	Soltero	55	
12	Si	Casado	80	
13	Si	Divorciado	110	
14	No	Soltero	95	
15	No	Casado	67	

Enfoque del modelo de clasificación

Clasificación de datos es un proceso de dos pasos:

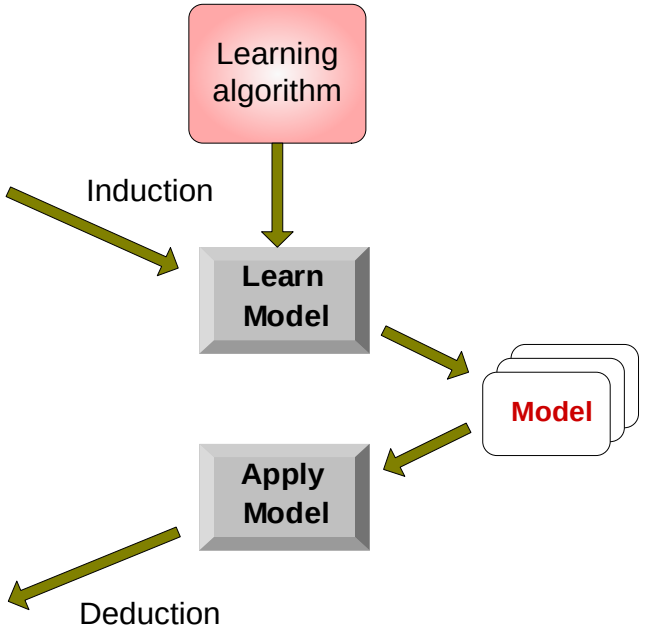
- **Paso de aprendizaje** donde se construye un modelo de clasificación
- **Paso de clasificación (prueba)** donde el modelo se utiliza para predecir las etiquetas de clase para los datos dados.

Training Set

ID	Casa propia	Estado civil	Ingreso anual	Default
1	Si	Soltero	125	No
2	No	Casado	100	No
3	No	Soltero	70	No
4	Si	Casado	120	No
5	No	Divorciado	95	Si
6	No	Casado	60	No
7	Si	Divorciado	220	No
8	No	Soltero	85	Si
9	No	Casado	75	No
10	No	Soltero	90	Si

Test Set

ID	Casa propia	Estado civil	Ingreso anual	Default
11	No	Soltero	55	
12	Si	Casado	80	
13	Si	Divorciado	110	
14	No	Soltero	95	
15	No	Casado	67	



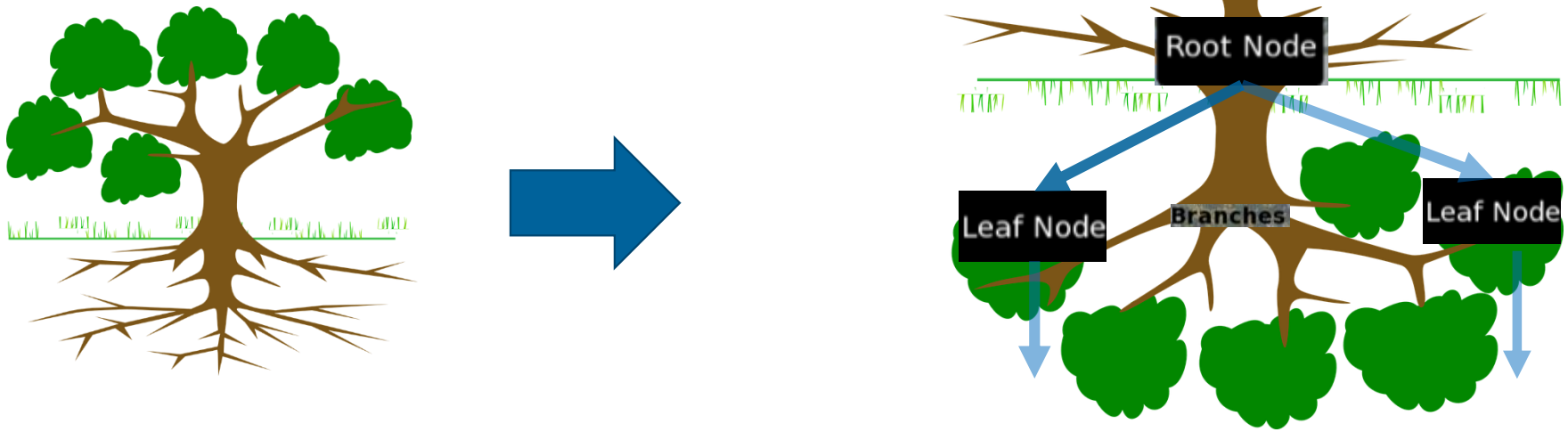
Técnicas de clasificación

1. Métodos basados en árboles Decision
2. Métodos basados en reglas
3. Neural Networks
4. Nearest-neighbor (Vecinos mas cercanos)
5. Naïve Bayes and Bayesian Belief Networks
6. Support Vector Machines (MSV)

Inducción del árbol de decisión

UN árbol de decisión es una **estructura de árbol** similar a un diagrama de flujo.

Donde el nodo superior en un árbol es, **raíz nodo**. cada **rama** representa un resultado de la prueba, cada **nodo interno** denota una prueba en un atributo, y cada **nodo terminal** tiene una etiqueta de clase.



CONJUNTOS DE RESULTADOS POSIBLES

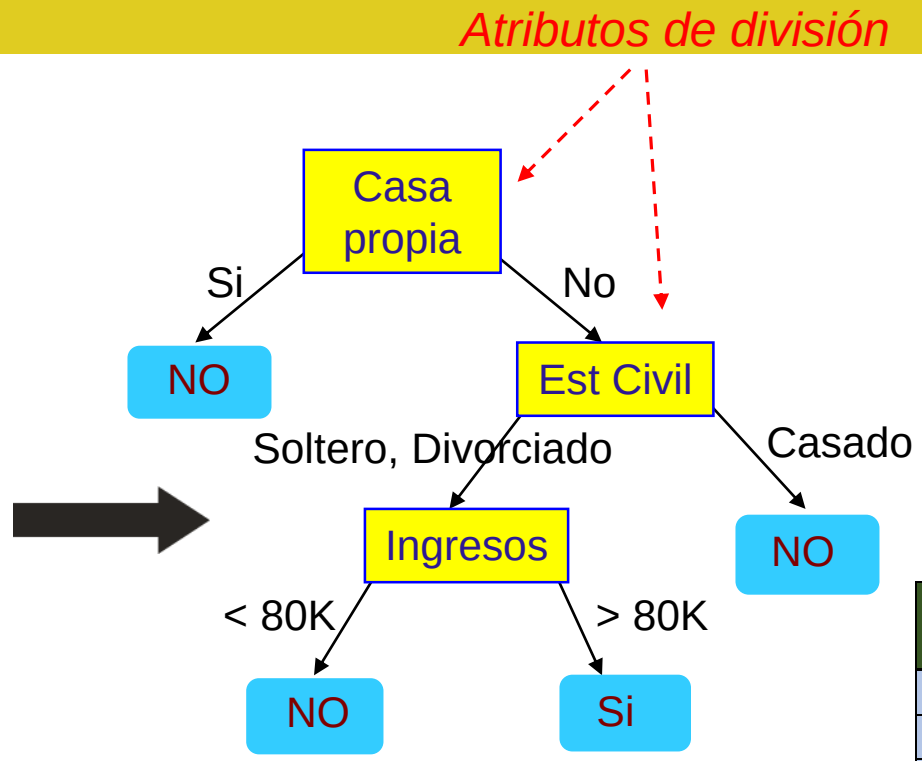
¿Cómo se utilizan los árboles de decisión para la clasificación?

1. Dado un conjunto de datos de entrada (*x*) Para la cual la etiqueta de clase asociada es desconocida (*y*), los valores de atributo de los datos de entrada se comparan con el árbol de decisión.
2. Se rastrea una ruta desde la raíz a un nodo hoja, que contiene la predicción de clase para tales datos.
3. Los árboles de decisión se pueden convertir fácilmente en reglas de clasificación.

¿Cómo se utilizan los árboles de decisión para la clasificación?

	Categorica	Categorica	Continua	clase
ID	Casa propia	Estado civil	Ingreso anual	Default
1	Si	Soltero	125	No
2	No	Casado	100	No
3	No	Soltero	70	No
4	Si	Casado	120	No
5	No	Divorciado	95	Si
6	No	Casado	60	No
7	Si	Divorciado	220	No
8	No	Soltero	85	Si
9	No	Casado	75	No
10	No	Soltero	90	Si

Training Data



Modelo: Árbol de desición

Test Set

ID	Casa propia	Estado civil	Ingreso anual	Default
11	No	Soltero	55	
12	Si	Casado	80	
13	Si	Divorciado	110	
14	No	Soltero	95	
15	No	Casado	67	

Árbol de decisión

1. Existen muchos algoritmos, la gran mayoría varia de acuerdo a su criterio de división
    1. Hunt’s Algorithm (one of the earliest)
    2. CART
    3. CHAID
    4. ID3 (Machine learning)
    5. C4.5 (Sucesor de ID3)
    6. SLIQ,SPRINT

Algoritmo de un árbol de decisión

Algorithm: Generate_decision_tree. Generate a decision tree from the training tuples of data partition, *D*.

Input:

- Data partition, *D*, which is a set of training tuples and their associated class labels;
- *attribute_list*, the set of candidate attributes;
- *Attribute_selection_method*, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a *splitting_attribute* and, possibly, either a *split-point* or *splitting subset*.

Output: A decision tree.

Method:

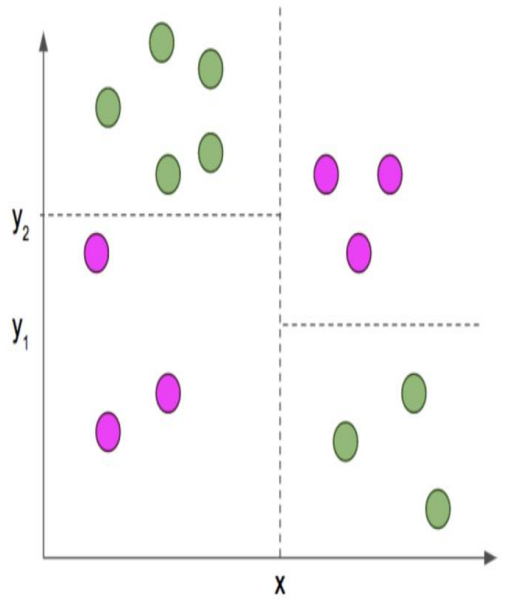
```
(1)  create a node N;
(2)  if tuples in D are all of the same class, C, then
(3)      return N as a leaf node labeled with the class C;
(4)  if attribute_list is empty then
(5)      return N as a leaf node labeled with the majority class in D; // majority voting
(6)  apply Attribute_selection_method(D, attribute_list) to find the "best" splitting_criterion;
(7)  label node N with splitting_criterion;
(8)  if splitting_attribute is discrete-valued and
         multiway splits allowed then // not restricted to binary trees
(9)      attribute_list ← attribute_list − splitting_attribute; // remove splitting_attribute
(10) for each outcome j of splitting_criterion
     // partition the tuples and grow subtrees for each partition
(11)     let Dj be the set of data tuples in D satisfying outcome j; // a partition
(12)     if Dj is empty then
(13)         attach a leaf labeled with the majority class in D to node N;
(14)     else attach the node returned by Generate_decision_tree(Dj, attribute_list) to node N;
     endfor
(15) return N;
```

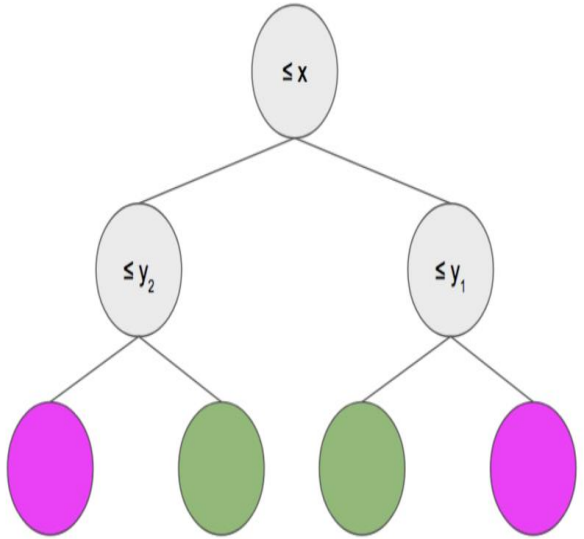
Construcción de un árbol de decisión

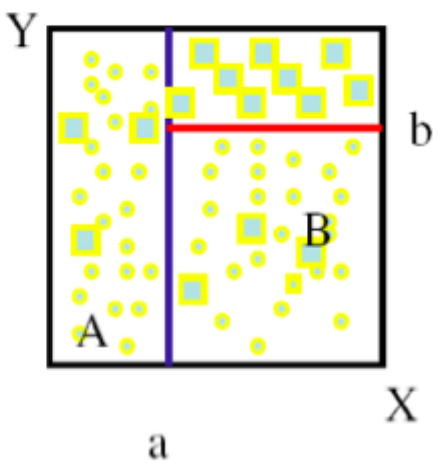
1. Un árbol de decisión particiona el espacio de variables predictoras en un conjunto de hiperrectangulos y en cada uno de ellos ajusta un modelo sencillo, generalmente una constante. Es decir, y = c, donde y es la variable de respuesta.
2. La construcción de un árbol de decisión se basa en cuatro elementos
    1. Un conjunto de preguntas binarias Q de la forma $\{x \in A\}$
    2. El método usado para particionar los nodos.
    3. La estrategia requerida para el crecimiento del árbol.
    4. La asignación de cada nodo terminal a una clase de la variable respuesta.

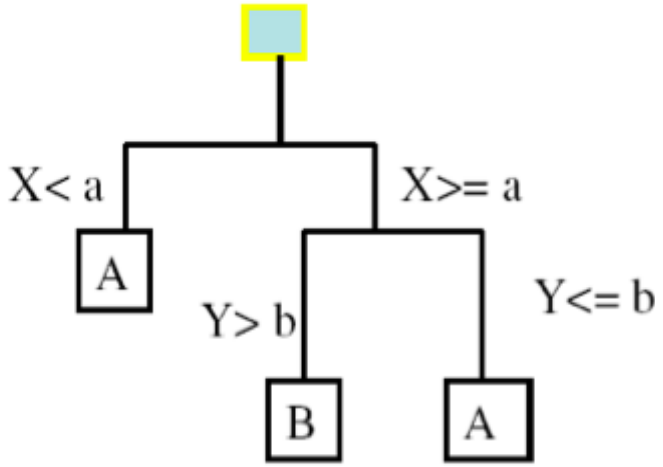
Las diferencias entre los algoritmos para construir arboles se hallan en la regla para particionar los nodos, la estrategia para podar los arboles, y el tratamiento de valores perdidos (“missing values")

Construcción de un árbol de decisión









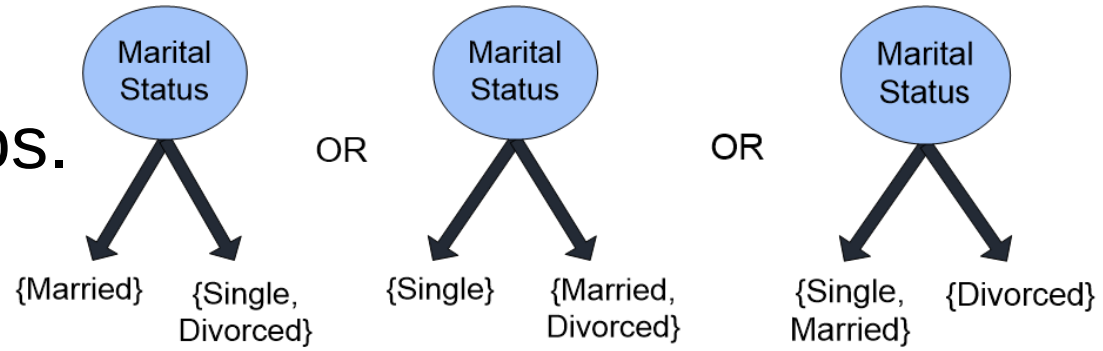
Construcción de un árbol de decisión

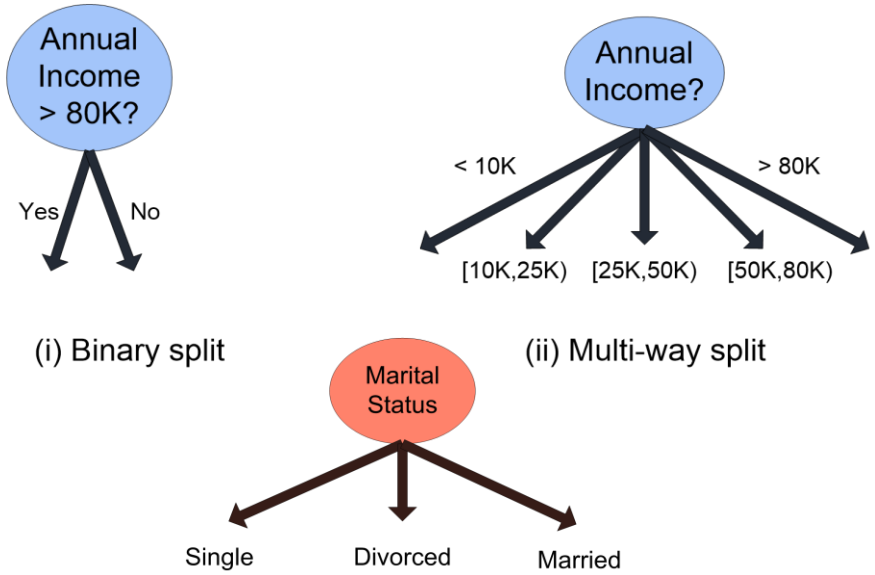
1. Estrategia:
    1. Dividir los registros basados en un atributo de prueba que optimice cierto criterio.
2. Discusión:
    1. Determinar como dividir los registros
        1. ¿Cómo especificar la condición a evaluar?
        2. ¿Cómo determinar la mejor partición?
3. Determinar el criterio de parada.

Métodos para expresar condiciones de prueba

1. Depende de los tipos de atributos.
   1. Binario
   2. Nominal
   3. Ordinal
   4. Continuo

5. Depende de la cantidad de maneras de dividir
   1. División de 2 vías
   2. División de múltiples vías





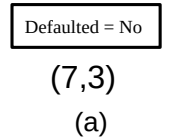
Algoritmo de Hunt’s

1. En el algoritmo de Hunt, un árbol de decisión crece de manera recursiva dividiendo (particionando) los registros de entrenamiento en subconjuntos sucesivamente más puros.
2. D_t es el conjunto de registros de entrenamiento que están asociados con los nodos *t* y *y* = {y_1, y_2,. . . , y_c} sean las etiquetas de clase. La siguiente es una definición recursiva del algoritmo de Hunt.
    1. Paso 1: Si todos los registros en D_t pertenecen a la misma clase y_t, entonces *t* es un nodo terminal etiquetado como y_t.
    2. Paso 2: si D_t contiene registros que pertenecen a más de una clase, se selecciona una condición de prueba de atributo para dividir los registros en subconjuntos más pequeños. Se crea un nodo hijo para cada resultado de la condición de prueba y los registros en D_t se distribuyen a los hijos según los resultados. El algoritmo se aplica recursivamente a cada nodo secundario.

Algoritmo de Hunt's

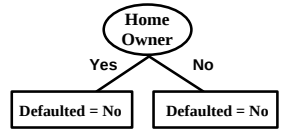
	Categorica	Categorica	Continua	clase
ID	Casa propia	Estado civil	Ingreso anual	Default
1	Si	Soltero	125	No
2	No	Casado	100	No
3	No	Soltero	70	No
4	Si	Casado	120	No
5	No	Divorciado	95	Si
6	No	Casado	60	No
7	Si	Divorciado	220	No
8	No	Soltero	85	Si
9	No	Casado	75	No
10	No	Soltero	90	Si

Training Data

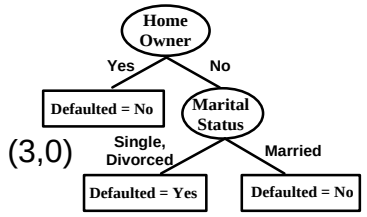


(7,3)

(a)

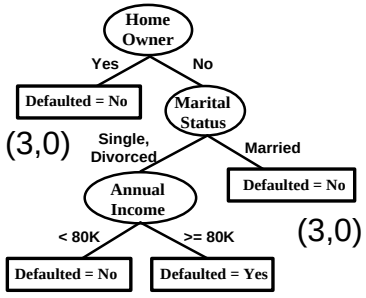


(3,0) (b) (4,3)



(3,0) (1,3) (3,0)

(c)



(3,0) (3,0) (1,0) (0,3)

(d)

Ejemplo 1:

Evaluación de un modelo

La evaluación del rendimiento de un modelo de clasificación se basa en los recuentos de registros de pruebas pronosticados correcta e incorrectamente por el modelo. Estos recuentos se tabulan en una tabla conocida como matriz de confusión.

- **Matriz de confusión**

  Matriz de confusión para un problema

  de clasificación.

  Es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea.

		Predicted Class	
		$Class = 1$	$Class = 0$
Actual Class	$Class = 1$	f_{11}	f_{10}
	$Class = 0$	f_{01}	f_{00}

Evaluación de un modelo

Aunque una matriz de confusión proporciona la información necesaria para determinar qué tan bien funciona un modelo de clasificación, resumir esta información con un solo número haría que sea más conveniente comparar el rendimiento de los diferentes modelos.

Accuracy:

- La evaluación del Accuracy de un modelo de clasificación se basa en los recuentos de registros de pruebas pronosticados correcta e incorrectamente por el modelo.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}.$$

Evaluación de un modelo

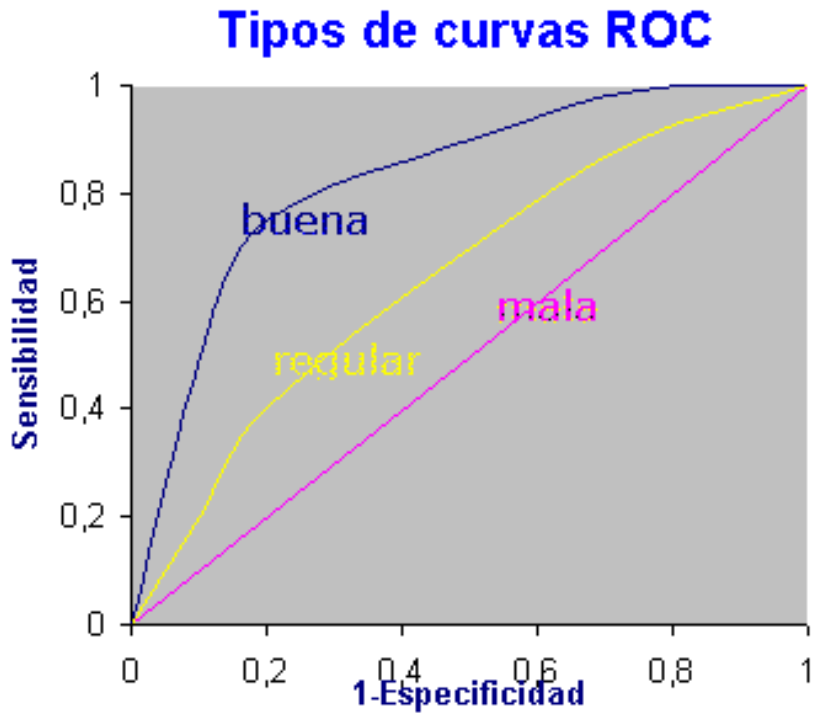
Error rate

- Se evalúa la tasa de error de un modelo o algoritmo
- La mayoría de los algoritmos de clasificación buscan modelos que logren la mayor precisión, o de manera equivalente, la tasa de error más baja cuando se aplica al conjunto de prueba.

$$\text{Error rate} = \frac{\text{Number of wrong predictions}}{\text{Total number of predictions}} = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}.$$

Indicadores : Curva de ROC

1. Una curva ROC es una representación gráfica de la sensibilidad en función de los falsos positivos (complementario de la especificidad) para distintos puntos de corte. Un parámetro para evaluar la bondad de la prueba es el área bajo la curva que tomará valores entre 1 (prueba perfecta) y 0,5 (prueba inútil).



La Librería rpart

En el R se puede utilizar la librería rpart con la función rpart cuya sintaxis es:

1. rpart(formula, data,method)

Argumentos:

formula: formula indicando las variable respuesta y las predictoras.

data: conjunto de datos a ser utilizado.

se usa **class** para árboles de decisión.

Opciones de control de rpart

1. minsplit: fija el número mínimo de observaciones en un nodo para que este sea dividido. Esta opción por defecto es 20.
2. minbucket: indica el número mínimo de observaciones en cualquier nodo terminal. Por defecto esta opción es el valor redondeado de minsplit/3.
3. cp: parámetro de complejidad. Indica que si el criterio de impureza no es reducido en mas de cp*100% entonces se para.
4. Por defecto cp=.01. Es decir, la reducción en la impureza del nodo terminal debe ser de al menos 1% de la impureza inicial.
5. maxdepth: condiciona la profundidad máxima del arbol. Por defecto está establecida como 30.

La Librería rpart

Para gracar el árbol se utiliza:

1. plot(objeto,margin=0.25)
2. text(objeto,use.n=T)

Argumentos:

1. objeto: salida de rpart.
2. margin: margen del gráfico del árbol.
3. use.n : si es T adiciona al gráfico cuantos elementos hay de cada clase.

La Librería rpart

Para realizar predicciones se utiliza:

predict(object,newdata,type)

Argumentos:

1. predict(object,newdata,type)
2. newdata: conjunto de datos con los cuales se va predecir
3. type: tipo de salida (“class” retorna la clase, “prob” retorna matriz de probabilidades de las clases)

Gracias…