Programación 3
Introducción a la Ciencia de Datos y Aprendizaje Automático

Unidad 1

Fernando Lichtschein

¿Qué es la ciencia de datos?

CONVERTIR DATOS EN
INFORMACIÓN

¿Qué es la ciencia de datos?

ANALIZAR DATOS PARA TENER
PERCEPCIÓN
QUE SE TRADUZCA EN
ACCIÓN

¿Qué es la ciencia de datos?

IDENTIFICAR
TENDENCIAS,
PATRONES Y
CORRELACIONES

¿Qué es la ciencia de datos?

PONERLOS EN CONTEXTO, COMPRENDERLOS Y **APLICARLOS**

Operaciones sobre datos

- Obtener y procesar los datos crudos para convertirlos a un formato limpio
- Calcular e interpretar las variables estadísticas
- Crear visualizaciones y sacar conclusiones para su análisis y utilización
- Sugerir aplicaciones para la información y desarrollar implementaciones de aprendizaje automático

¿Por qué?

¿Por qué?

8 CPU 64 bit
1,8 GHz
4 GB RAM

¿Por qué?

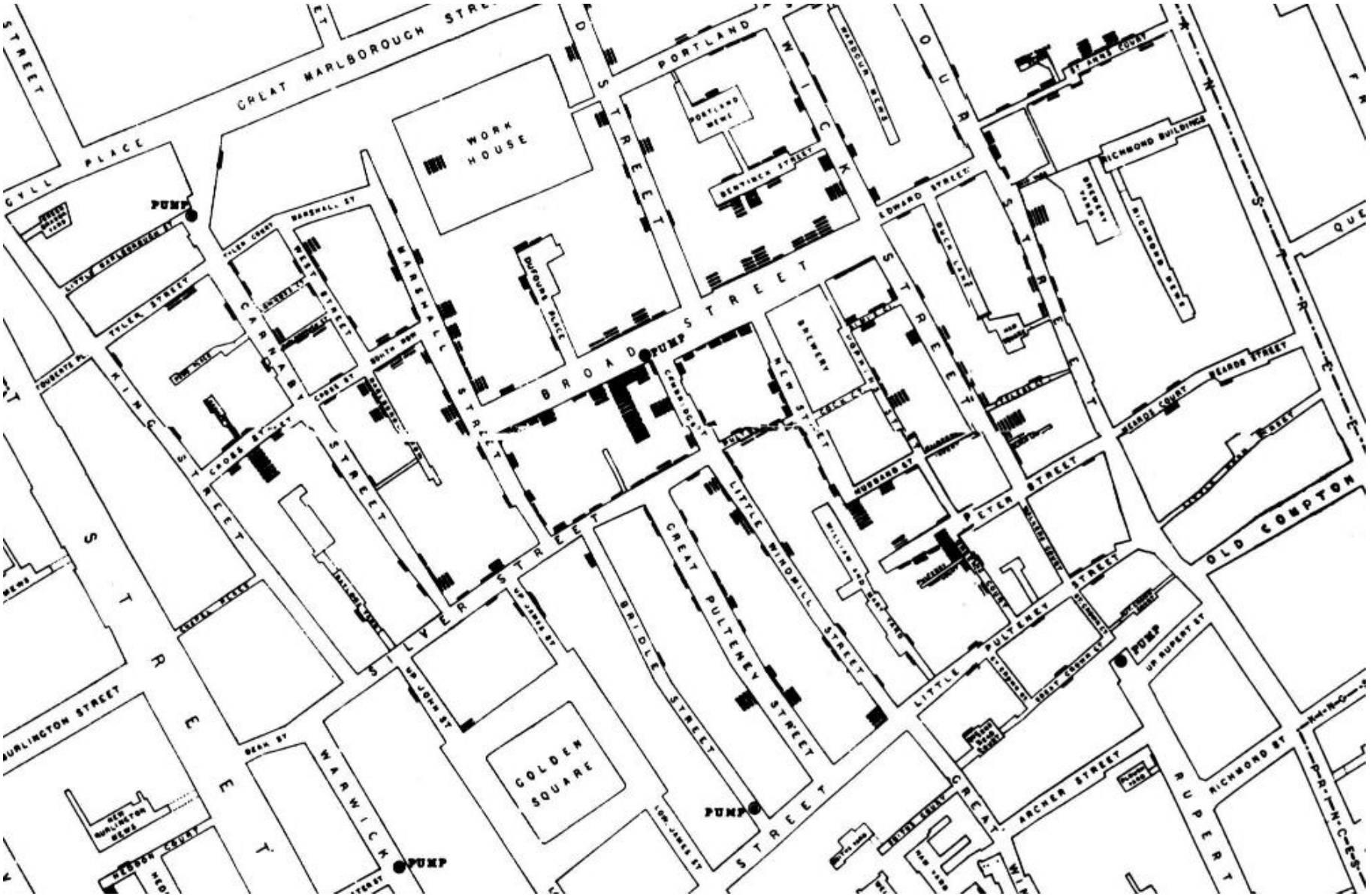
1 CPU 32 bit
2 MHz
128 KB RAM

8 CPU 64 bit
1,8 GHz
4 GB RAM

Estadística

- Comprender los distintos tipos de datos que se pueden encontrar
- Comprender los términos estadísticos clave:
  - Medidas de valor central
  - Medidas de variaciones en los datos
  - Distribución de los valores
- Dividir, agrupar y segmentar grupos de datos

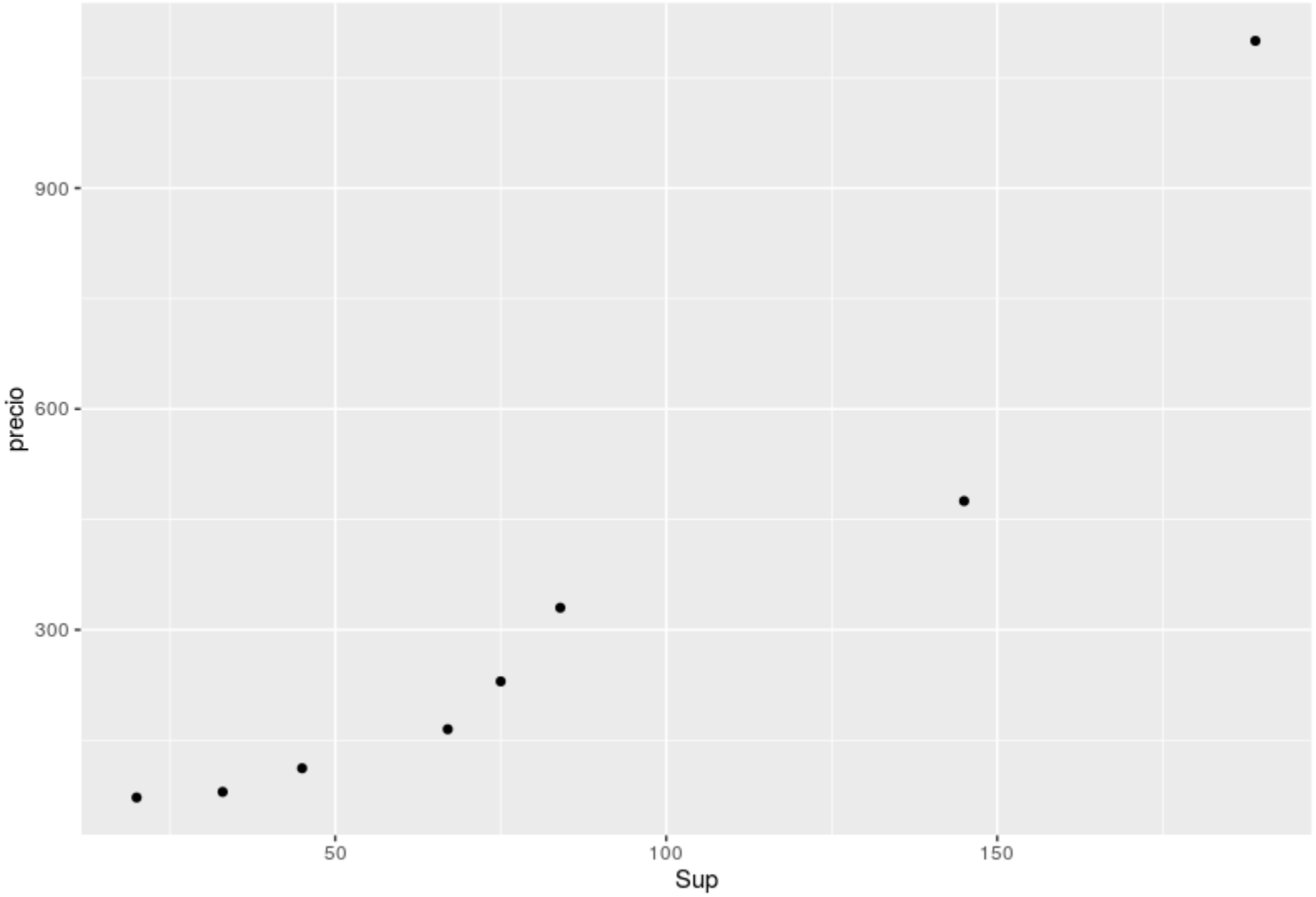
John Snow, 1854



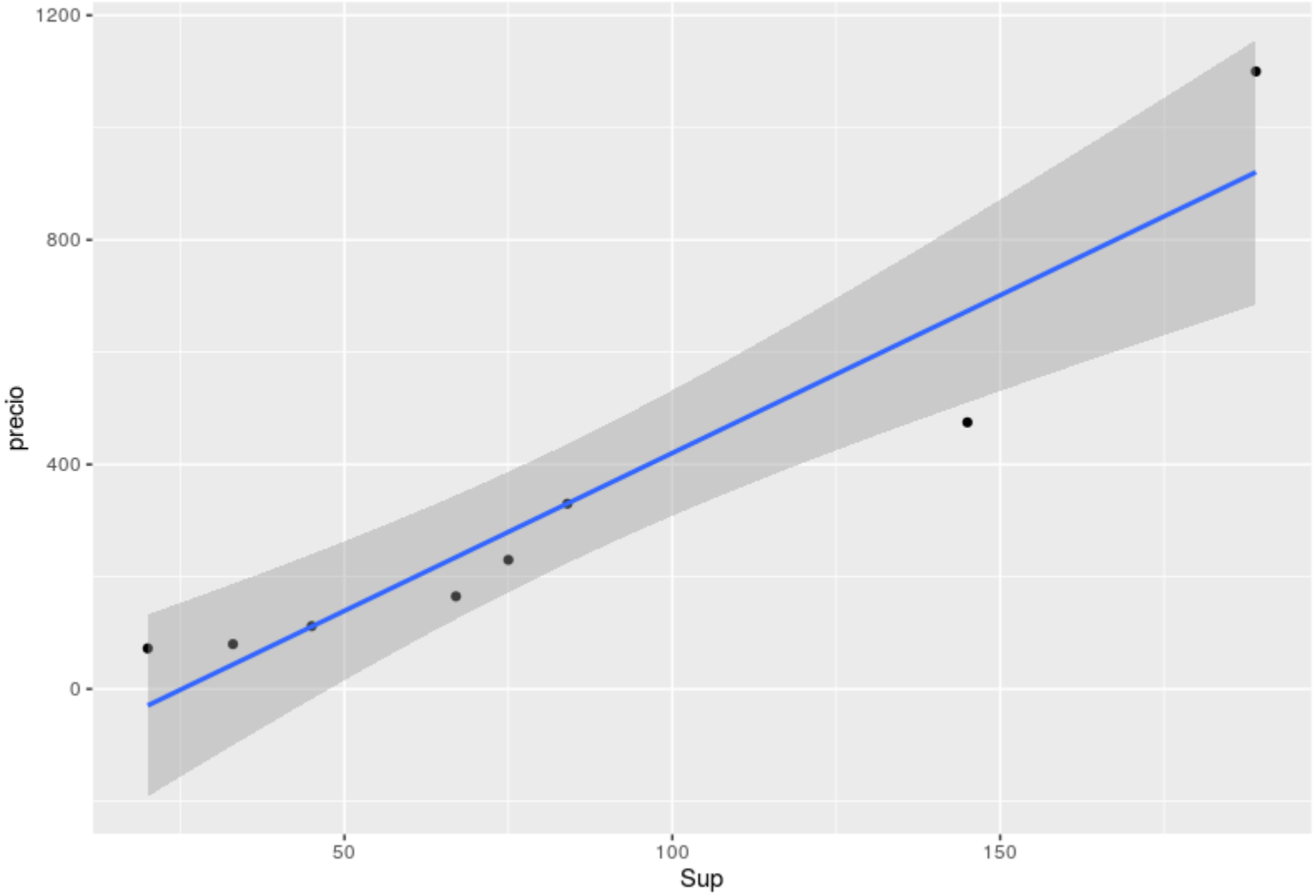
Datos de departamentos

Superficie	Precio
75	230.000
189	1.100.000
20	72.270
33	79.900
84	330.000
45	112.000
145	475.000
67	165.000

Datos de departamentos



Datos de departamentos



Datos de departamentos

Superficie	Ambientes	Precio
75	3	230.000
189	4	1.100.000
20	1	72.270
33	1	79.900
84	2	330.000
45	2	112.000
145	4	475.000
67	3	165.000

A **B** → **C**

Datos de departamentos

Superficie	Ambientes	Baños	Zona	Precio
75	3	2	Caballito	230.000
189	4	4	Palermo	1.100.000
20	1	1	Belgrano	72.270
33	1	1	Congreso	79.900
84	2	1	Belgrano	330.000
45	2	1	Monserrat	112.000
145	4	3	Belgrano	475.000
67	3	1	Congreso	165.000

Machine Learning vs. Data Science

Aprendizaje De Máquina
(Machine Learning)

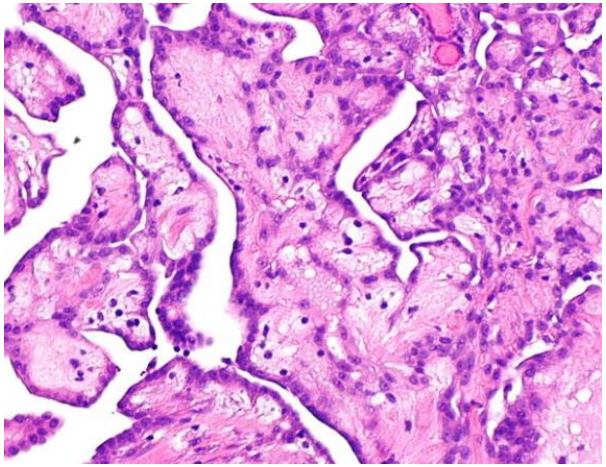
"Campo de estudio que da a las computadoras la habilidad de aprender sin ser programadas explícitamente."

Arthur Samuel - 1959

Ciencia de datos
(Data Science)

La ciencia de extraer conocimiento e información de los datos.

Tipos de problema



¿Estas células son cancerosas?

¿Estos hongos son comestibles?

Tipos de problema

- Clasificación

  El objetivo es ubicar a un objeto dentro de una categoría de acuerdo a sus características

- Regresión

  El resultado de esta operación es un valor numérico. Puede ser simple (una variable de entrada) o múltiple (varias variables de entrada)

Tipos de aprendizaje

- Supervisado
  - Los datos están clasificados mediante un identificador o “etiqueta”
- No supervisado
  - El algoritmo debe descubrir los patrones a partir de los mismos datos
- Por refuerzo
  - El algoritmo determina la estrategia de resolución a partir de los resultados obtenidos en intentos sucesivos

El juego de datos Iris

Iris setosa

Iris virginica

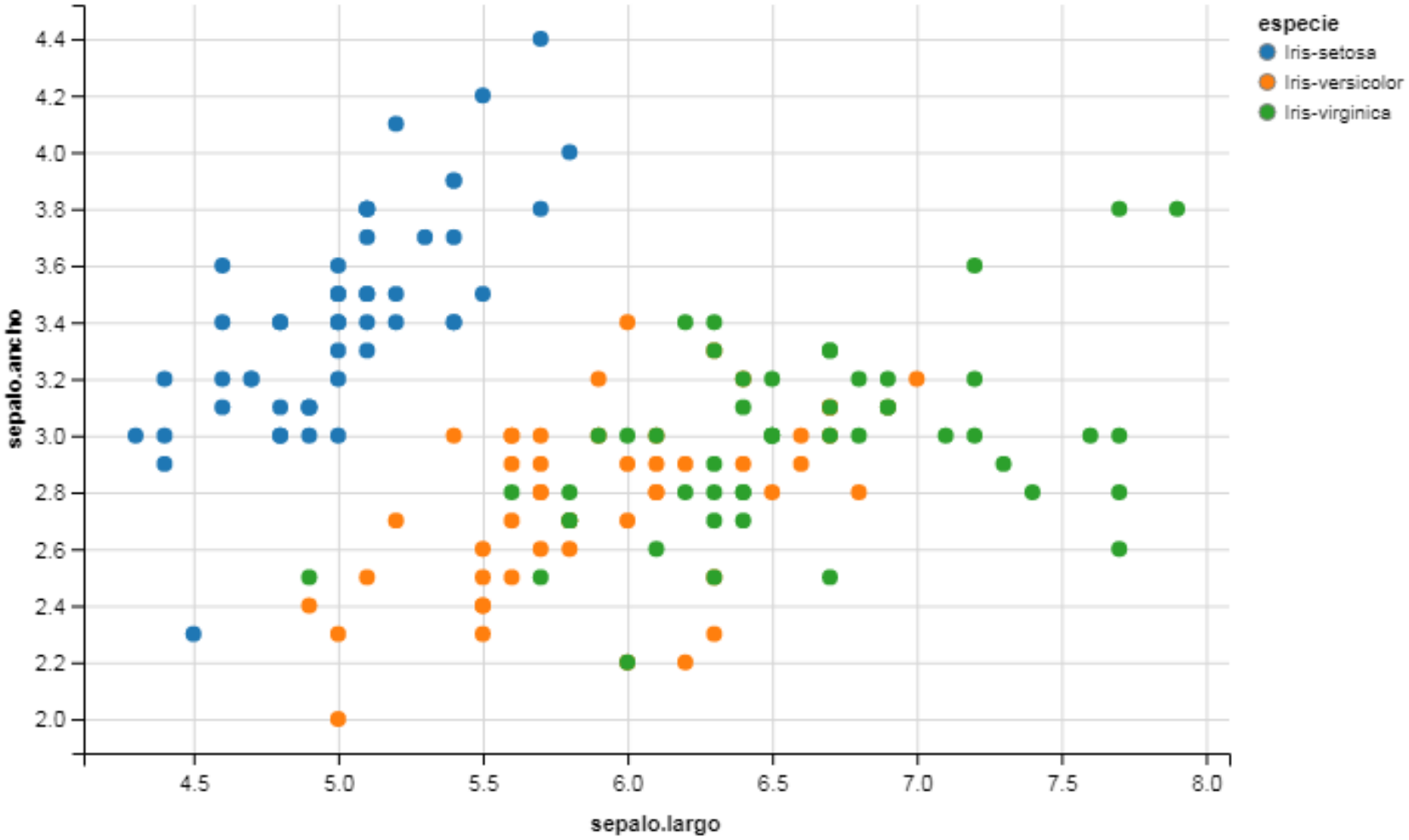
Iris versicolor

El juego de datos Iris

El juego de datos Iris

sepalo.largo	sepalo.ancho	petalo.largo	petalo.ancho	especie
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5.0	3.6	1.4	0.2	Iris-setosa
5.4	3.4	1.7	0.4	Iris-setosa
.	.	.	.	.

El juego de datos Iris

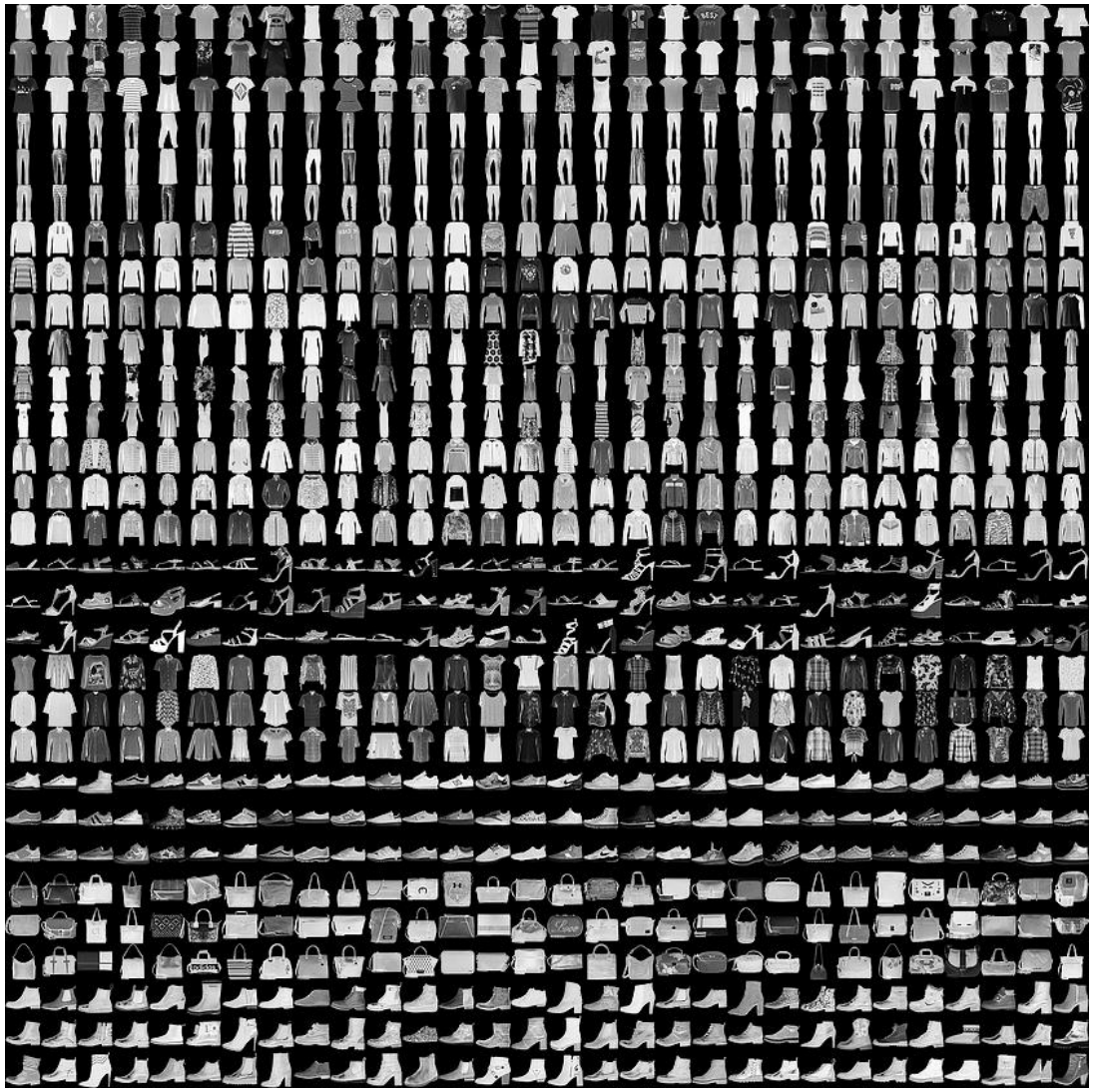


El juego de datos MNIST

(extracto)

http://yann.lecun.com/exdb/mnist/

El juego de datos Fashion-MNIST



https://github.com/zalandoresearch/fashion-mnist

Limpieza de Datos

Superficie	Ambientes	Baños	Zona	Precio
75		2	Caballito	230.000
189	4	4	Palermo	1.100.000
20	1	1	Belgrano	72.270
33	1	1	-	79.900
84	2	1	Belgrano	12
450	2	-	Monserrat	75.000
145	4	3	Belgrano	475.000
67	3	1	Congreso	165.000

Limpieza de Datos

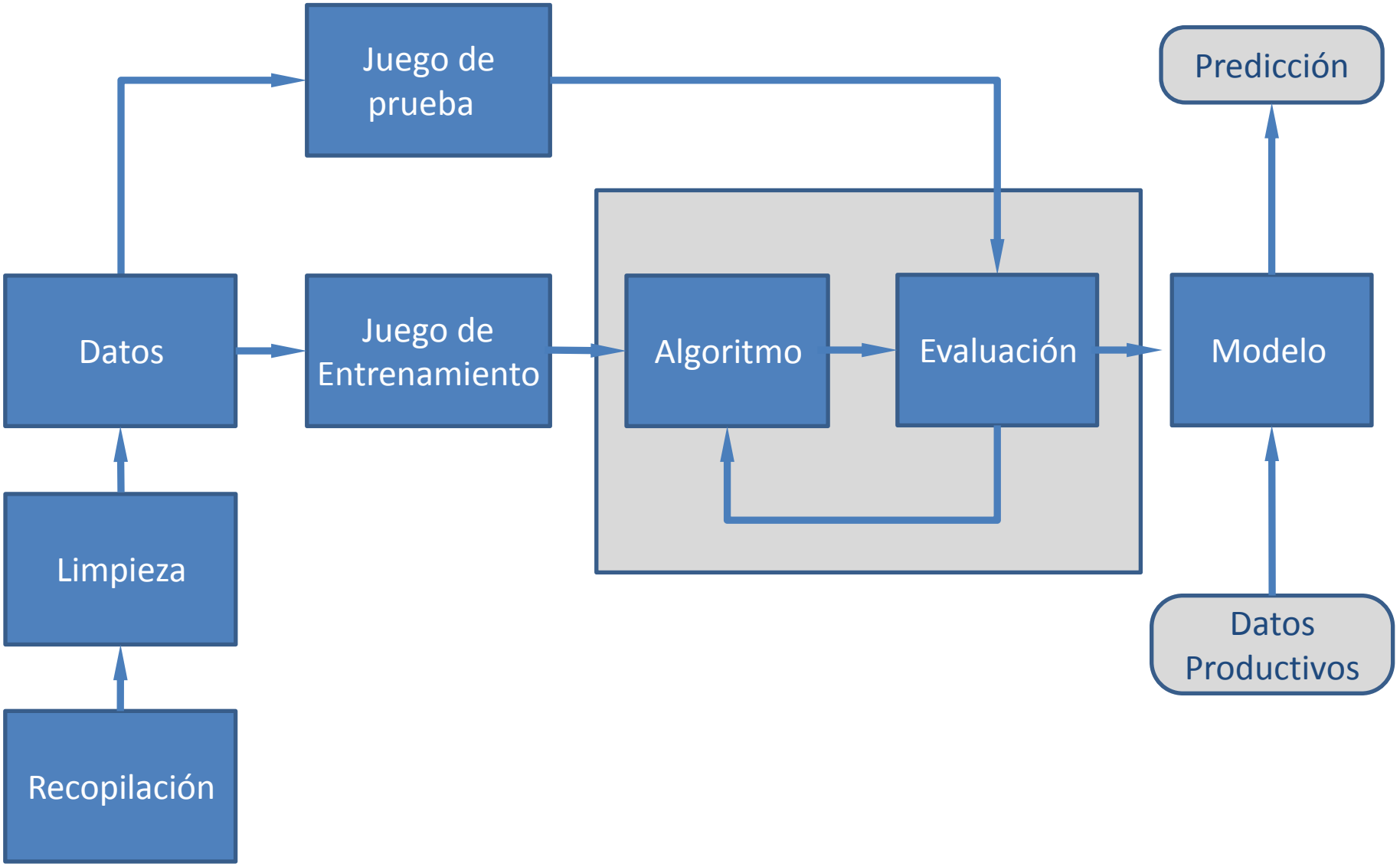
No siempre los datos pueden utilizarse sin tratarlos previamente. Puede haber:

- Datos faltantes
  - En algunas de las observaciones faltan datos
- Datos incorrectos
  - Algunos de los valores son erróneos. ¿Los puede identificar?

Flujo de datos de un proyecto

- El juego de datos de entrenamiento se utiliza para que el algoritmo seleccionado reconozca los patrones en los datos.
- El juego de datos de prueba sirve para ver cómo se comporta el algoritmo entrenado con datos de entrada desconocidos.
- Se obtienen dividiendo en dos partes al azar los datos originales en proporciones que generalmente son un 75% - 80% para entrenamiento y el resto para prueba
- Ocasionalmente se usa un tercer juego de validación para refinar el modelo

Flujo de datos de un proyecto



Matriz de confusión

Se utiliza para verificar el desempeño de un algoritmo de clasificación.

- Las *columnas* representan a *los valores reales* de los datos de prueba.
- Las *filas* de la matriz representan las *predicciones* del algoritmo

Matriz de confusión

Se utiliza para verificar el desempeño de un algoritmo de clasificación. La siguiente corresponde a una clasificación de tumores.

Datos predichos

Datos reales	Maligno	Benigno
Maligno	61	1
Benigno	2	36

El cuadro muestra que el modelo clasificó correctamente 61 casos como malignos (verdaderos positivos) y 36 como benignos (verdaderos negativos). Sin embargo, 2 benignos fueron clasificados como malignos (falso positivo) y un maligno fue clasificado como benigno (falso negativo).
En este contexto, “positivo” y “negativo” designan si una observación pertenece o no a la clase que estamos tratando de identificar, en este caso tumores malignos.

Programación 3
Introducción a la Ciencia de Datos

Preparación de Datos

Fernando Lichtschein

Preparación de juegos de datos

Supongamos que queremos entrenar un modelo para que reconozca el palo de un naipe.
Para eso armamos dos juegos de datos, uno para enseñarle al modelo y otro para verificar si el modelo aprendió.

(no se dibujan las 52 cartas por claridad)

♥	Corazones	Rey
♥	Corazones	Reina
♥	Corazones	As
♥	Corazones	Cinco
♦	Diamantes	Ocho
♦	Diamantes	Seis
♦	Diamantes	Rey
♦	Diamantes	Seis
♠	Pique	Nueve
♠	Pique	As
♠	Pique	Siete
♠	Pique	Dos
♣	Trébol	Reina
♣	Trébol	Jota
♣	Trébol	As
♣	Trébol	Siete

Preparación de juegos de datos

Datos de entrenamiento

♥	Corazones	Rey
♥	Corazones	Reina
♥	Corazones	As
♥	Corazones	Cinco
♦	Diamantes	Ocho
♦	Diamantes	Seis
♦	Diamantes	Rey
♦	Diamantes	Seis
♠	Pique	Nueve
♠	Pique	As
♠	Pique	Siete
♠	Pique	Dos

Datos de prueba

♣	Trébol	Reina
♣	Trébol	Jota
♣	Trébol	As
♣	Trébol	Siete

Preparación de juegos de datos

Esta separación no va a funcionar bien, mi modelo sólo va a haber aprendido a reconocer corazones, diamantes y piques. Cuando se le presenten los tréboles en la prueba (o en el uso real) probablemente los clasifique mal.

¿La solución? Mezclar los datos que forman ambos juegos.

♥	Corazones	Rey
♥	Corazones	Reina
♥	Corazones	As
♥	Corazones	Cinco
♦	Diamantes	Ocho
♦	Diamantes	Seis
♦	Diamantes	Rey
♦	Diamantes	Seis
♠	Pique	Nueve
♠	Pique	As
♠	Pique	Siete
♠	Pique	Dos

♣	Trébol	Reina
♣	Trébol	Jota
♣	Trébol	As
♣	Trébol	Siete

Preparación de juegos de datos

Ahora con los datos mezclados podemos separarlos en dos juegos. Otra alternativa es extraer los datos para ambos juegos de forma aleatoria.

♦	Diamantes	Rey
♥	Corazones	Reina
♠	Pique	As
♥	Corazones	Cinco
♣	Trébol	Reina
♦	Diamantes	Seis
♠	Pique	Siete
♣	Trébol	As
♠	Pique	Nueve
♦	Diamantes	Ocho
♣	Trébol	Siete
♥	Corazones	Rey
♦	Diamantes	Seis
♣	Trébol	Jota
♥	Corazones	As
♠	Pique	Dos

Preparación de juegos de datos

Ahora ambos juegos contienen la misma distribución de casos.

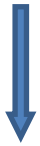
♦	Diamantes	Rey
♥	Corazones	Reina
♠	Pique	As
♥	Corazones	Cinco
♣	Trébol	Reina
♦	Diamantes	Seis
♠	Pique	Siete
♣	Trébol	As
♠	Pique	Nueve
♦	Diamantes	Ocho
♣	Trébol	Siete
♥	Corazones	Rey

♦	Diamantes	Seis
♣	Trébol	Jota
♥	Corazones	As
♠	Pique	Dos

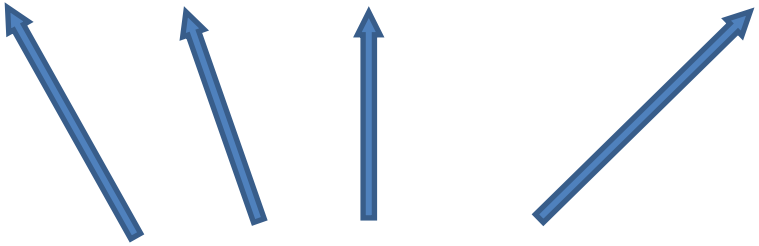
Programación 3
Modelos
y
Predicción

Modelo de predicción

Variable de salida



Var 1	Var 2	Var 3	Categoria	Var 4
5.1	3.5	1.4	Uno	0.2
4.9	3.0	1.4	Dos	0.2
4.7	3.2	1.3	Dos	0.2
4.6	3.1	1.5	Uno	0.2
5.0	3.6	1.4	Uno	0.2
5.4	3.4	1.7	Dos	0.4
.	.	.	.	.



Variables de entrada

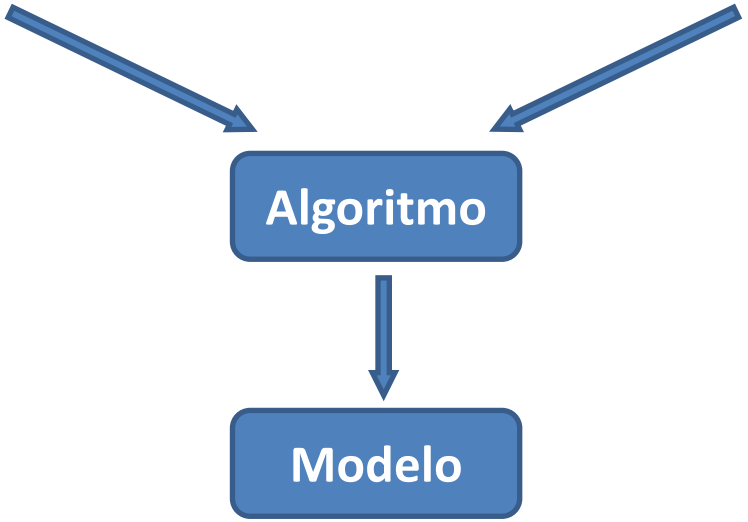
Creación del Modelo

Variables de entrada.

Var 1	Var 2	Var 3	Var 4
5.1	3.5	1.4	0.2
4.9	3.0	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5.0	3.6	1.4	0.2
5.4	3.4	1.7	0.4
.	.	.	.

Variable de salida

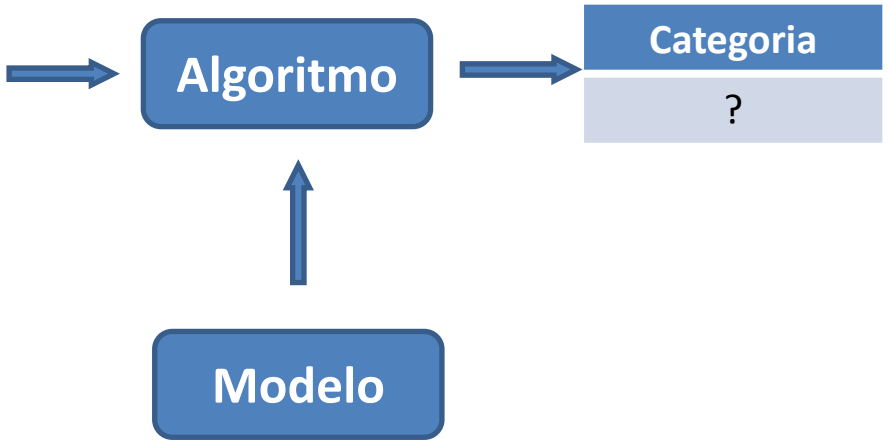
Categoria
Uno
Dos
Dos
Uno
Uno
Dos
.



Uso del Modelo

Variables de entrada.

Var 1	Var 2	Var 3	Var 4
5.1	3.5	1.4	0.2



Variable de salida

Categoria
?