



Analítica de datos

Aprendizaje automático

Jorge Bedoya

Aprendizaje

Temas a tratar:



1. Introducción



2. Tareas de aprendizaje



3. Modelos



4. Evaluación de modelos



5. Modelos Supervisados

Predicción
Clasificación



6. Modelos No supervisados

Clustering
Asociaciones
Correlaciones
Técnicas de reducción de datos



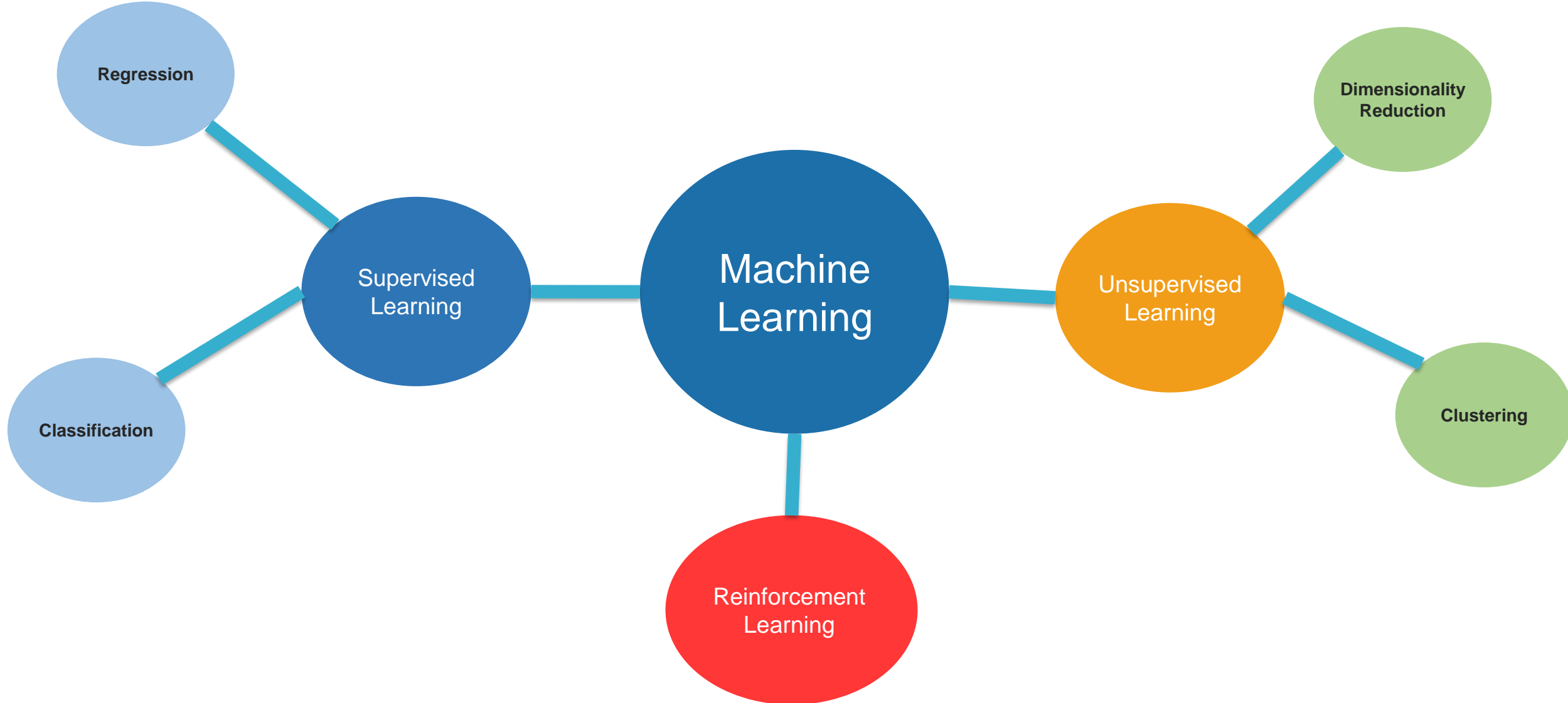
7. Otros temas de aprendizaje

Series temporales
Minería de texto
Detección de datos atípicos



<https://animalmascota.com/fotos-de-mamas-oso-ensenando-sus-crias/mamas-oso-ensenando-a-sus-crias1/>

1. Introducción



2. Tareas de aprendizaje

Supervisado

- Regresión
- Clasificación

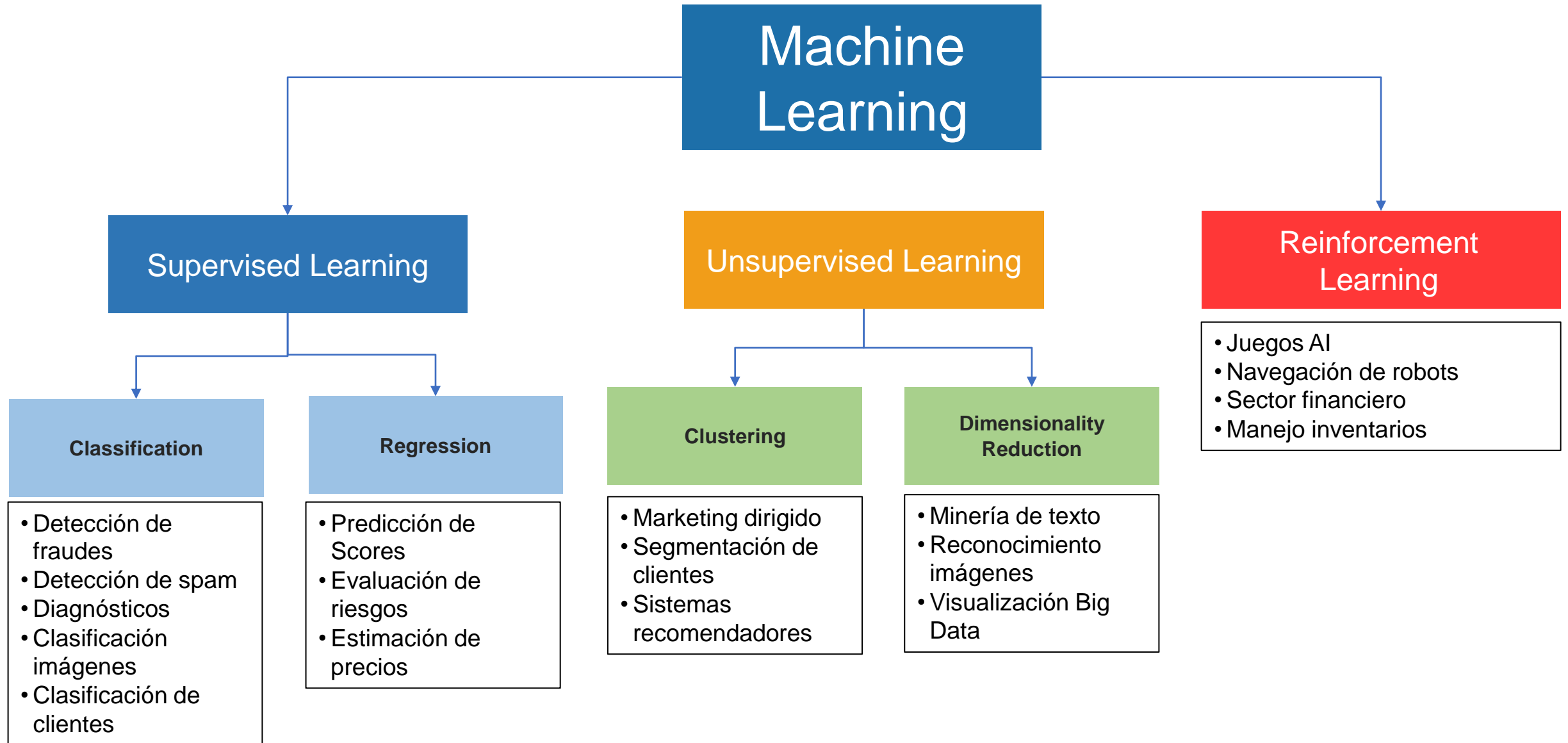
No supervisado

Descriptivo

- Análisis exploratorio
- Agrupamiento (o Clustering)
- Reducción de la dimensionalidad

- Correlaciones (y dependencias)
- Asociaciones

2. Tareas de aprendizaje



Aprendizaje supervisado (predictivo)

Objetivo:

Crear una función capaz de predecir el valor correspondiente a cualquier objeto de entrada válida después de haber visto una serie de ejemplos (datos de entrenamiento).

- Existe conocimiento previo (tiene una variable de salida)
- El resultado de la función puede ser un valor numérico ([regresión](#)) o una etiqueta de clase ([clasificación](#)).

Aprendizaje no supervisado (descriptivo)

Objetivo:

Comprender los datos: la relación entre las variables y entre las instancias (ejemplos)

- No hay un conocimiento a priori (No tiene un atributo de salida)
- Comúnmente requiere un proceso posterior
- El resultado es:
 - **Asociaciones** y **dependencias** (variables categóricas)
 - **Correlaciones** (variables numéricas)
 - **Agrupaciones** (relaciones entre instancias)
- ¿Qué variables le aportan al modelo?
- ¿Qué transformaciones me permiten reducir la dimensionalidad?

3. Modelos y evaluación de modelos

Video Modelos



3. Modelos

Modelo*

Permiten comprender los datos, sus atributos y relaciones:

- Paramétricos
- No paramétricos

* <https://www.youtube.com/watch?v=Sb8XVheowVQ&list=PL-Ogd76BhmcDxef4liOGXGXLL-4h65bs4&index=5>

<https://www.youtube.com/watch?v=CELZmc56v4I>

3. Paramétricos

Construyen la función que aproxima los datos de entrenamiento a la variable objetivo con un número fijo de parámetros

Por ejemplo, un algoritmo de regresión lineal tiene la forma:

$$a_0 + a_1x_1 + a_2x_2 + \dots + a_nx_n$$

Modelos de regresión lineal
(simple y múltiple)

Ventajas:

- Fáciles de entender
- Entrenamiento suele ser rápido

Desventajas:

- Limitar la complejidad del modelo generado.

3. No Paramétricos

No presuponen una forma concreta en el modelo a generar

Ventajas:

- Más flexibles y dando generalmente mejor resultado

- k-nearest neighbors
- Árboles de decisión
- Support Vector Machine

Desventajas:

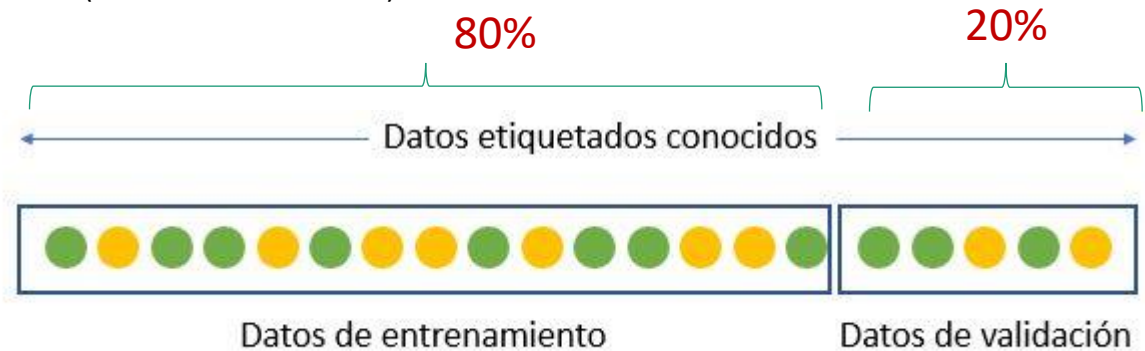
- Requieren más datos para su entrenamiento y resultando más lentos
- Son más proclives al sobreentrenamiento y más difíciles de interpretar.

4. Evaluación de modelos

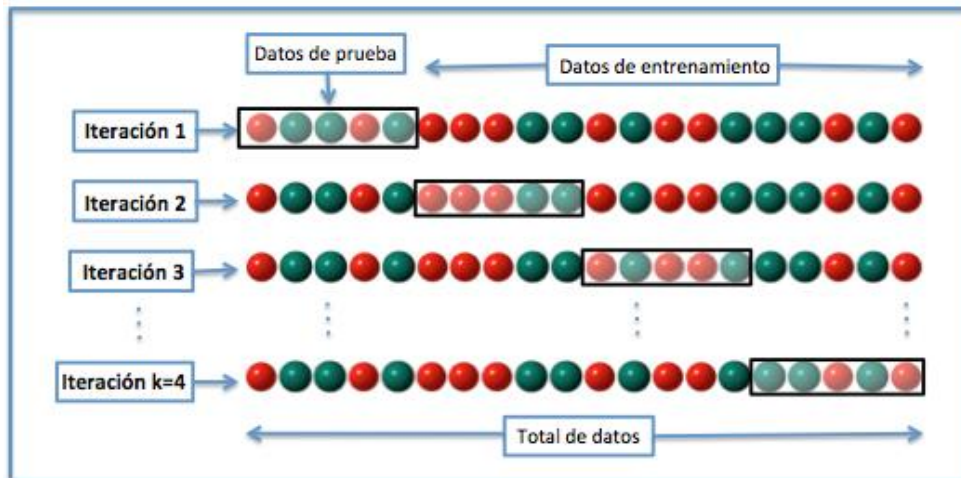
✓ Validación modelos:

- ✓ Miden la eficacia de un modelo

Método de retención
(*holdout method*)



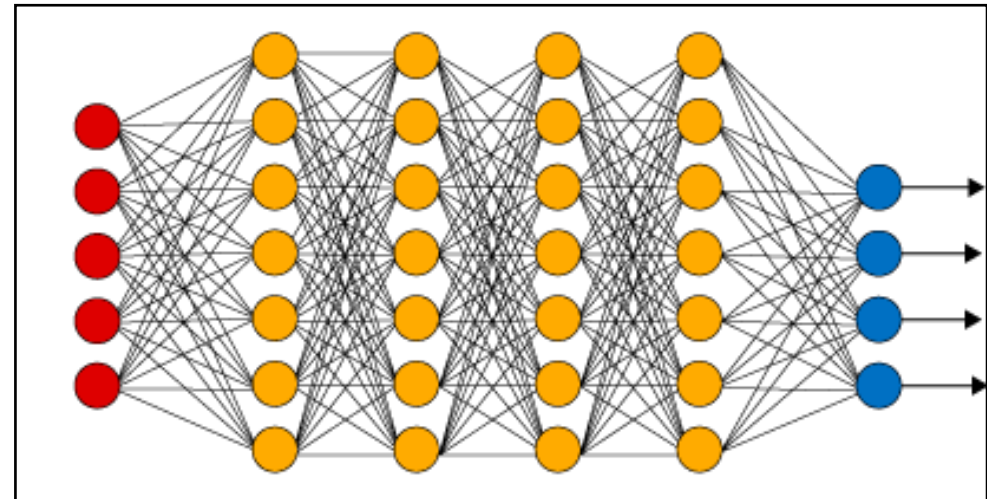
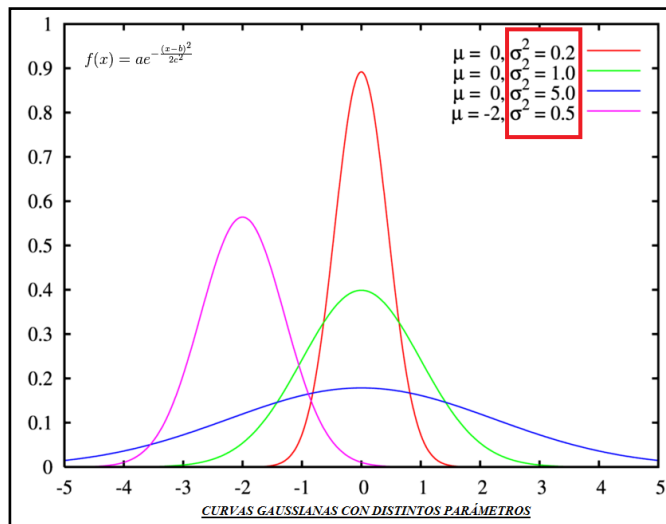
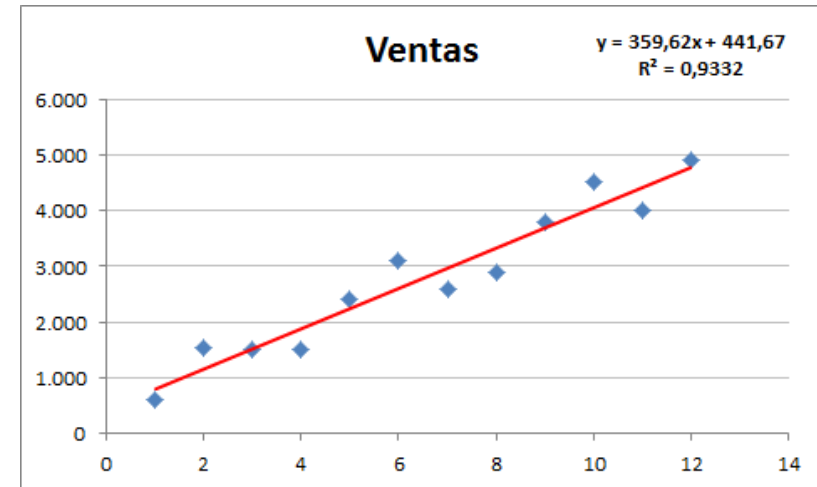
Validación cruzada
(*cross validation*)



4. Evaluación de modelos

✓ Ajuste de un modelo

- Calcular los parámetros de un modelo
- $y=a+bx$, donde a y b son los parámetros



4. Evaluación de modelos

✓ Matriz de confusión:

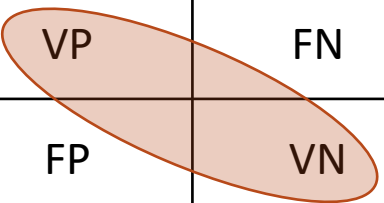
		PREDICCIÓN	
		Positivo	Negativo
REAL	Positivo	VP	FN
	Negativo	FP	VN

Clasificación

4. Evaluación de modelos

✓ Matriz de confusión:

		PREDICCIÓN	
		Positivo	Negativo
REAL	Positivo	VP	FN
	Negativo	FP	VN



Clasificación

✓ Exactitud: proporción de instancias identificadas correctamente entre todas las instancias

4. Evaluación de modelos

✓ Matriz de confusión:

		PREDICCIÓN	
		Positivo	Negativo
REAL	Positivo	VP	FN
	Negativo	FP	VN

Clasificación

✓ Tasa de errores: proporción de instancias identificadas incorrectamente entre todas las instancias

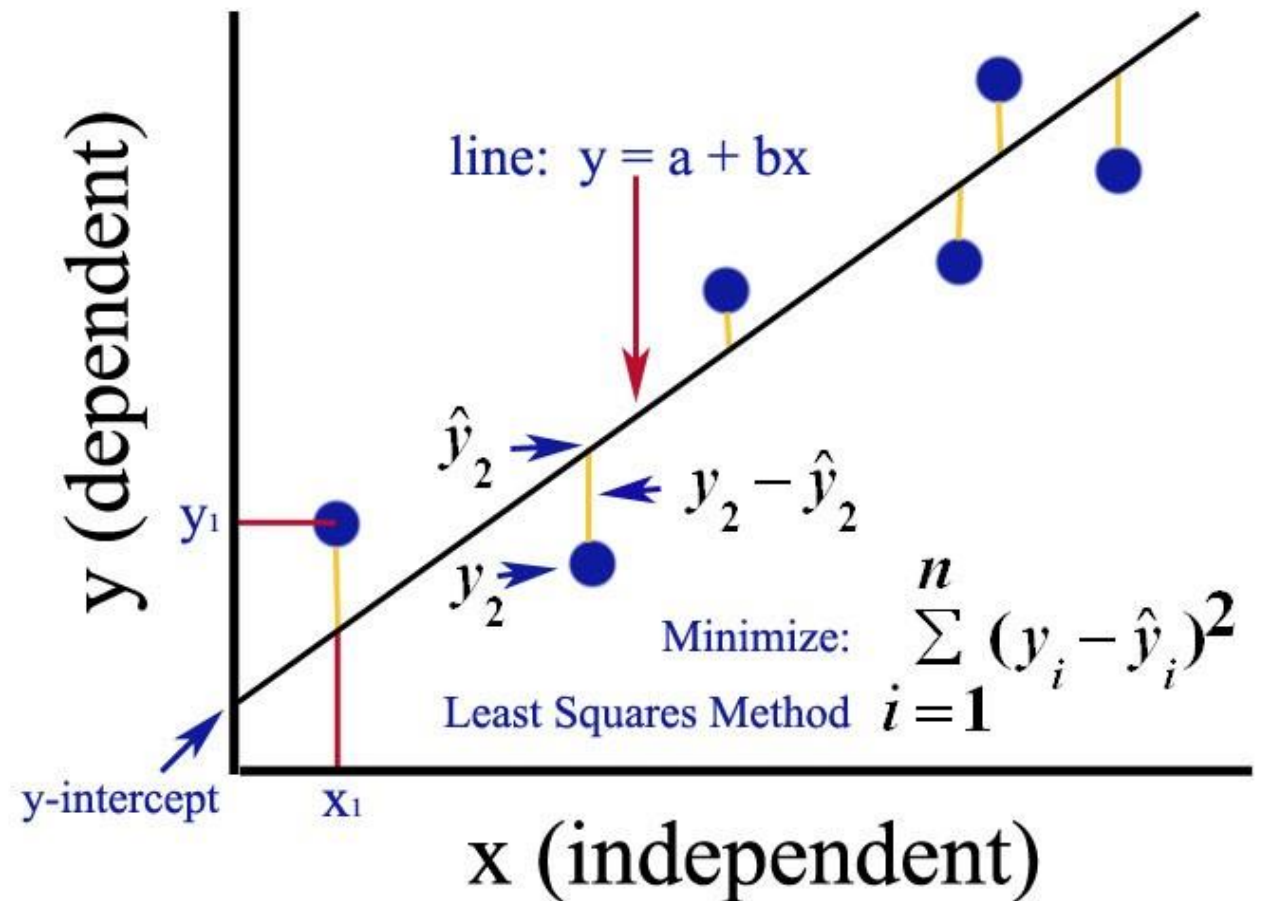
4. Evaluación de modelos

Regresión

Funciones de valor residual (Regresión):
Diferencia entre el valor predicho
(o *score*) y el valor real.

Error medio cuadrado (Mean squared error) o MSE

$$MSE = \frac{1}{N} \sum_{i=1}^N (f(x_i) - y_i)^2$$



5. Modelos Supervisados



Predicción

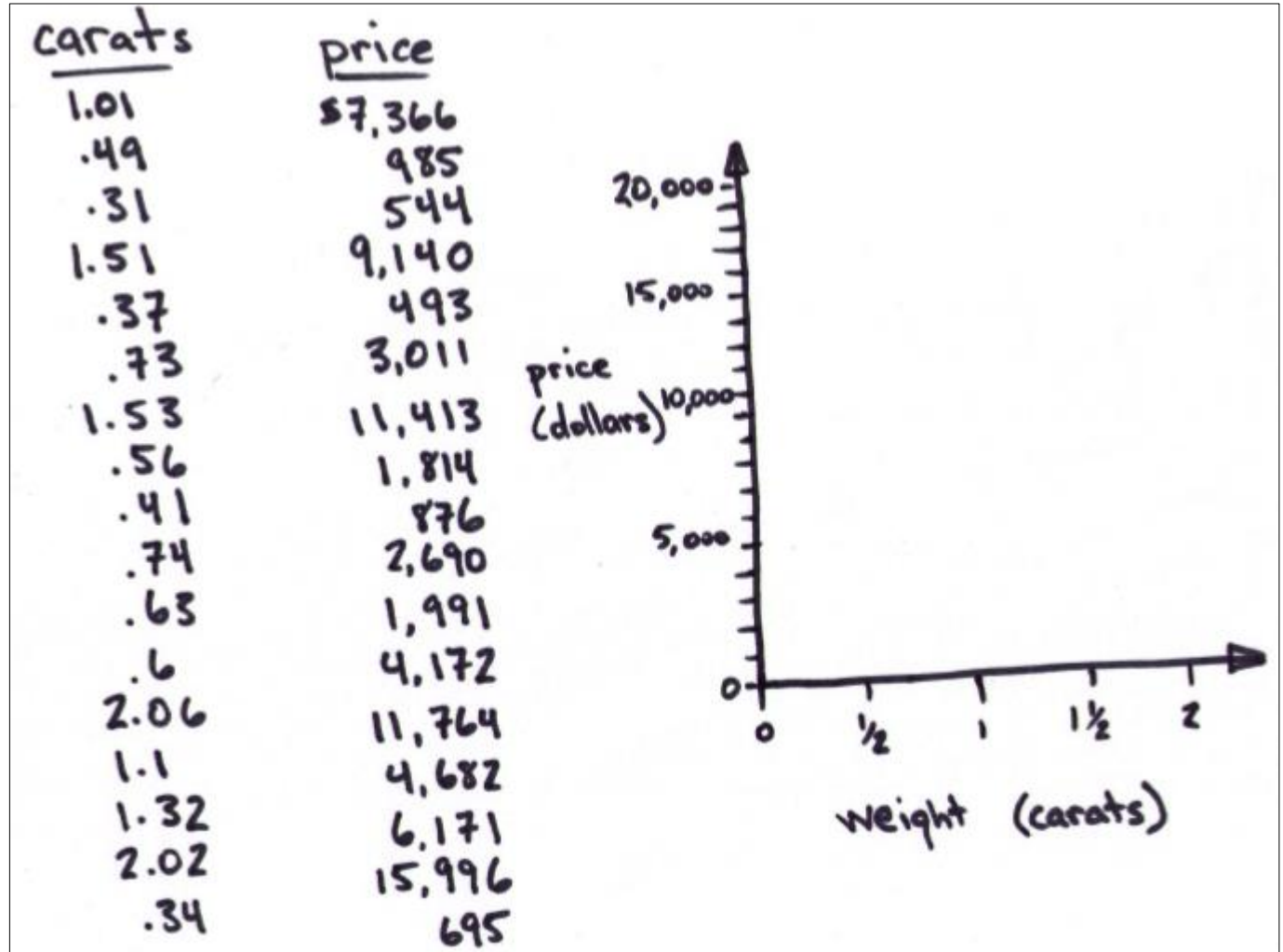
- ✓ **Regresión lineal:**
- ✓ Técnica estadística para predecir valores de una variable continua dependiente con base en valores de una variable independiente
- ✓ **¿Cuántos?**
 - ¿Cuántos twits vamos a recibir si existen por lo menos 10 influenciadores relevantes?
 - ¿Cuántos email van a abrir?



Predicción

✓ Regresión lineal:

✓ Diamantes:



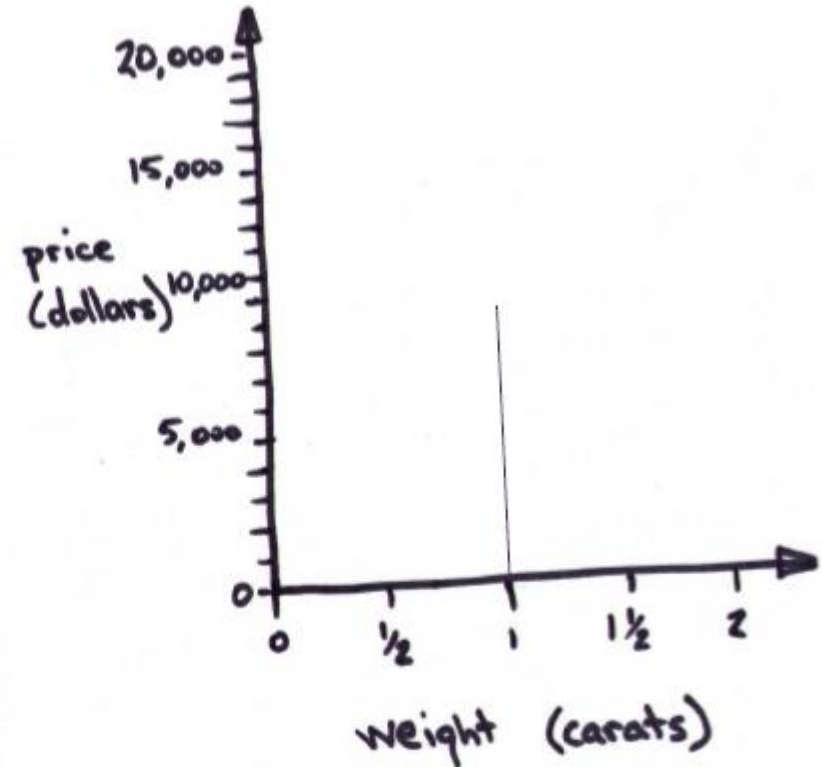
Predicción

✓ Regresión lineal:

✓ Diamantes:



<u>carats</u>	<u>price</u>
1.01	\$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695



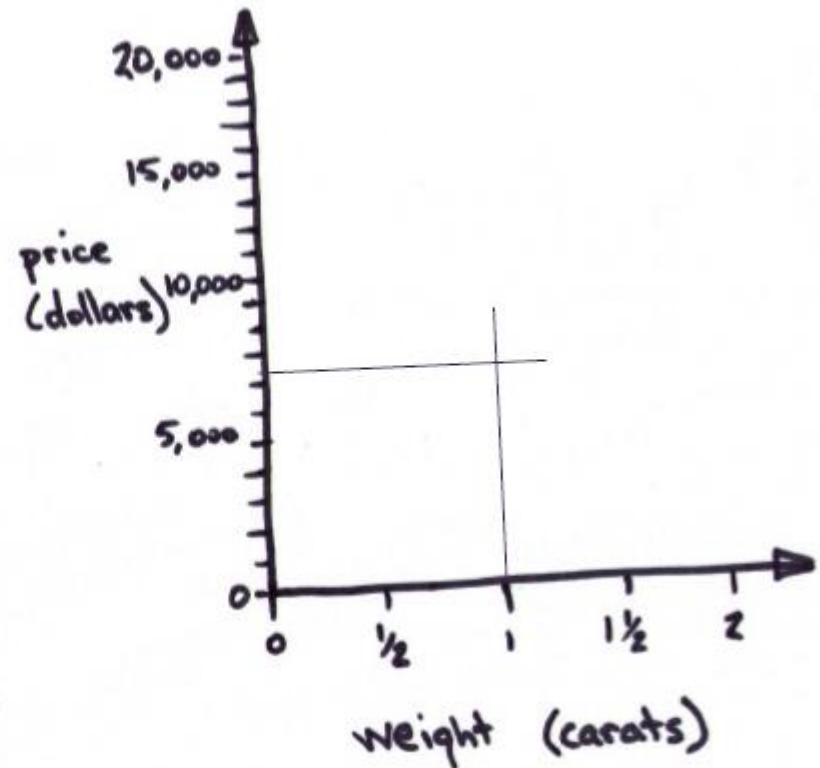
Predicción

✓ Regresión lineal:

✓ Diamantes:



<u>carats</u>	<u>price</u>
1.01	57,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695



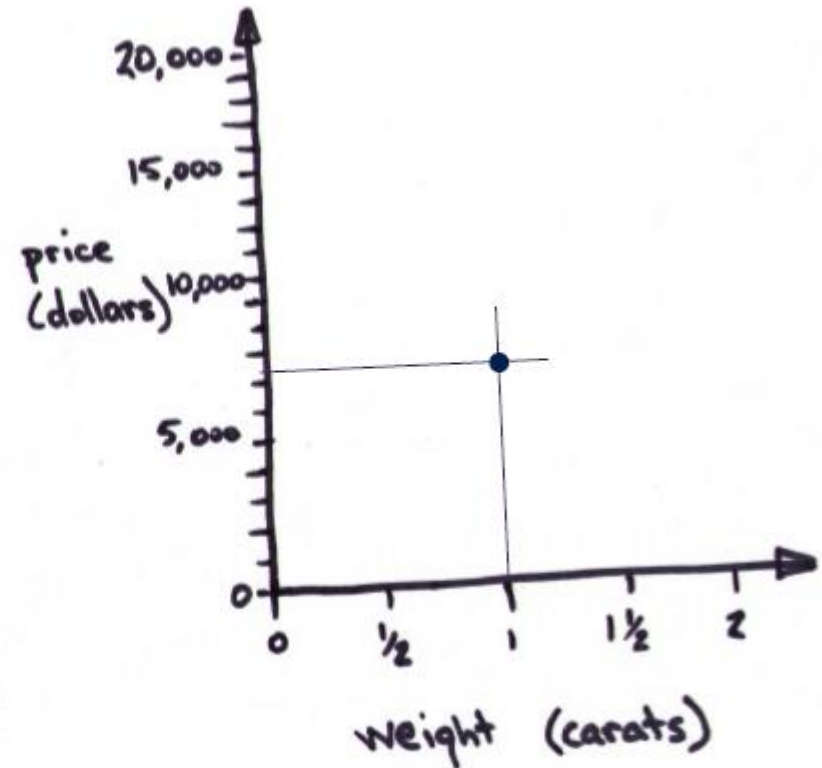
Predicción

✓ Regresión lineal:

✓ Diamantes:



<u>carats</u>	<u>price</u>
1.01	\$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695



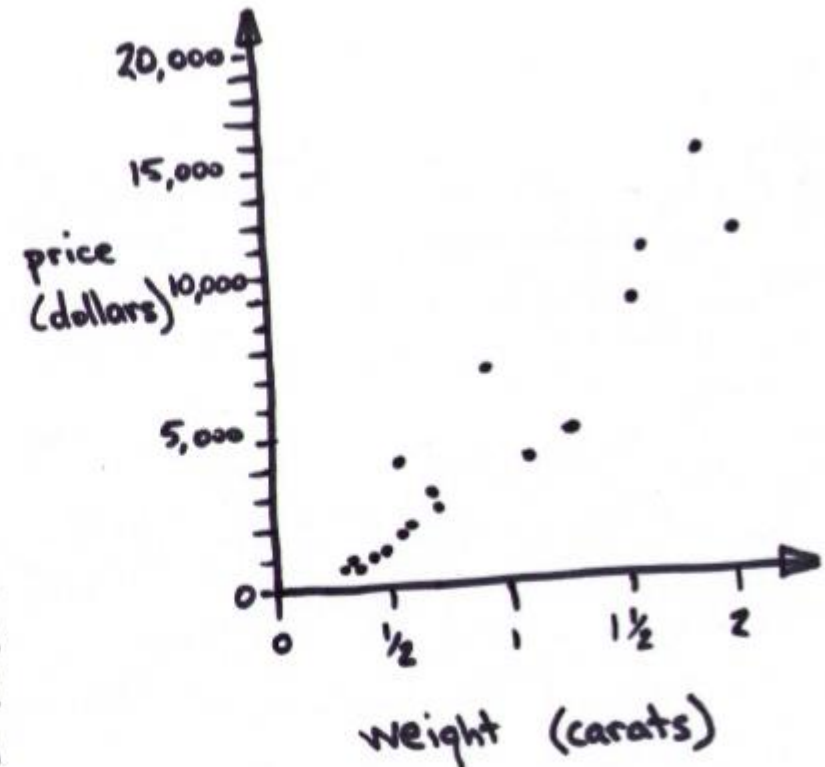
Predicción

✓ Regresión lineal:

✓ Diamantes:



<u>carats</u>	<u>price</u>
1.01	\$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695



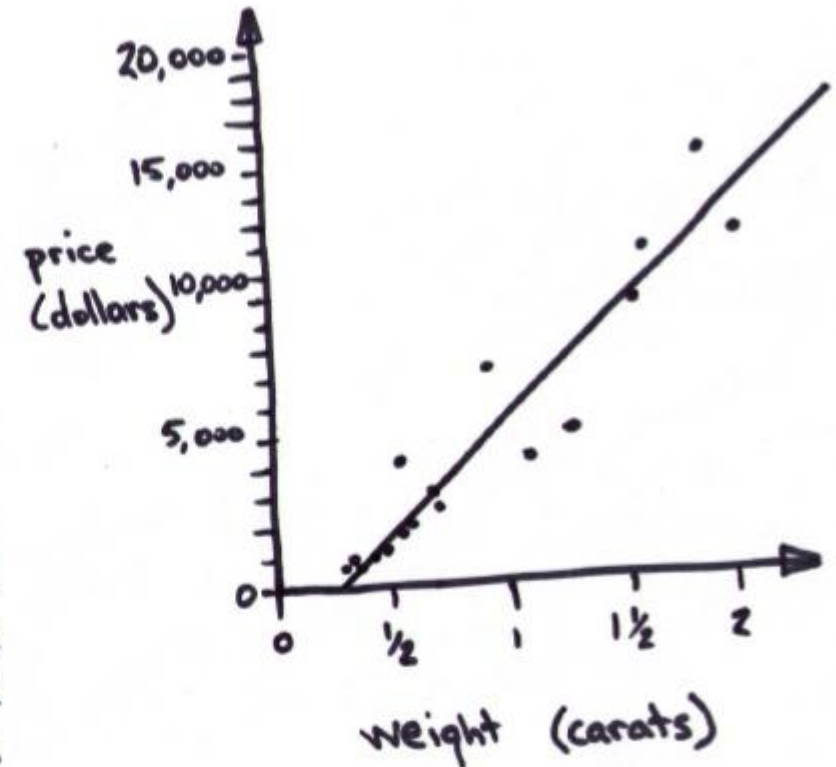
Predicción

✓ Regresión lineal:

✓ Diamantes:



<u>carats</u>	<u>price</u>
1.01	\$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695



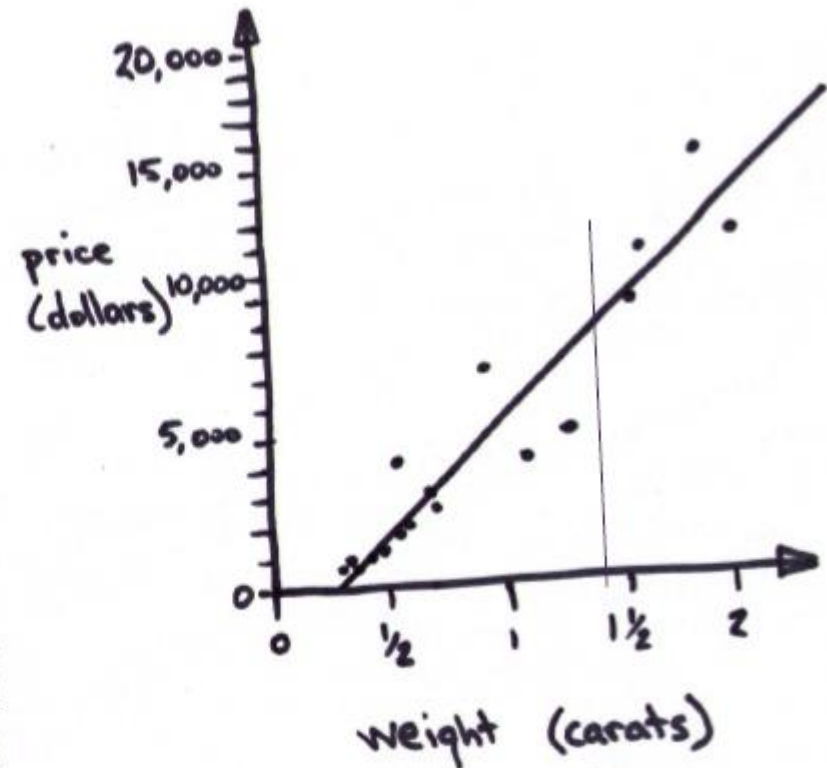
Predicción

✓ Regresión lineal:

✓ Diamantes:



<u>carats</u>	<u>price</u>
1.01	\$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695



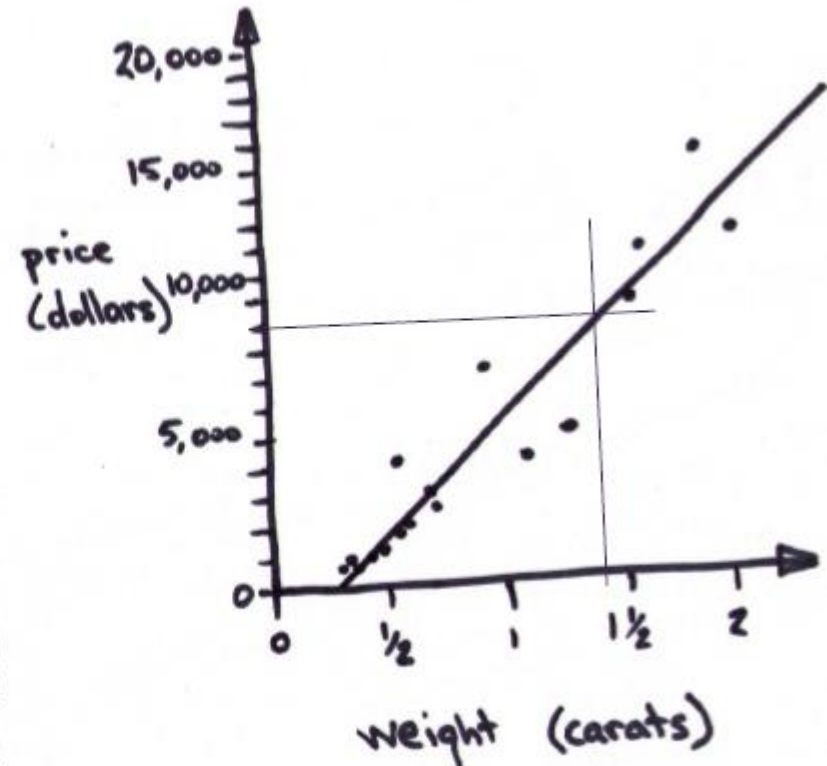
Predicción

✓ Regresión lineal:

✓ Diamantes:



<u>carats</u>	<u>price</u>
1.01	\$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695



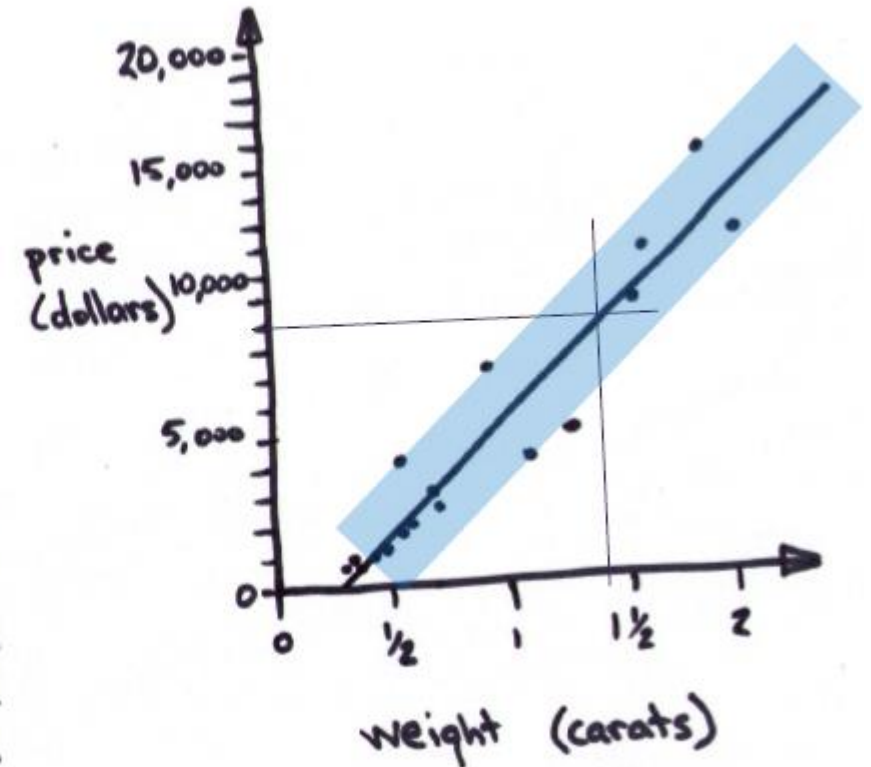
Predicción

✓ Regresión lineal:

✓ Diamantes:



<u>carats</u>	<u>price</u>
1.01	\$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695



Predicción

✓ Técnicas:

✓ Regresion Lineal

- Proceso estadístico para estimar las relaciones entre variables
- Ayuda a entender cómo el valor de la variable dependiente varía al cambiar el valor de una de las variables independientes
- Se ve afectado por los valores atípicos

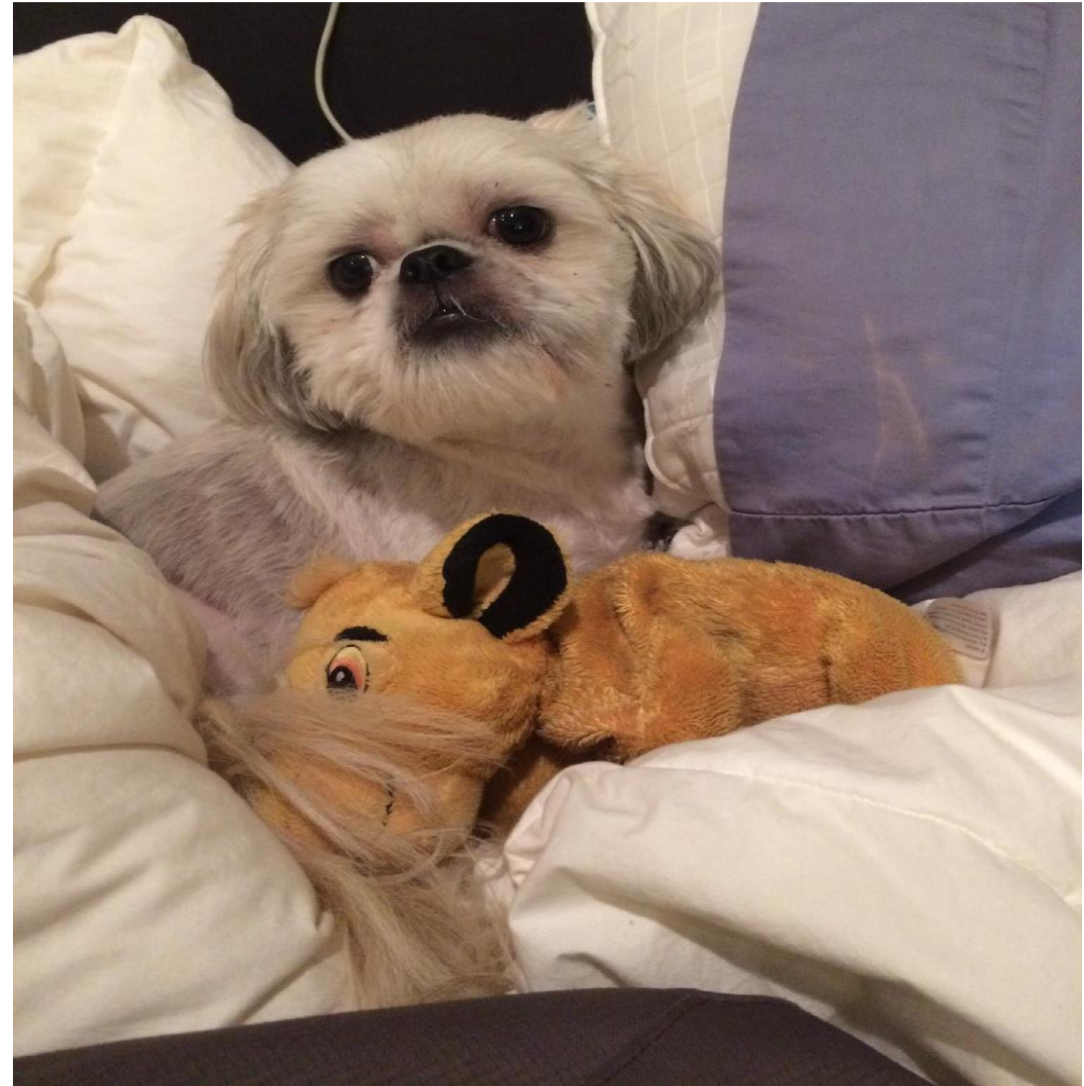
✓ Maquinas de vectores de soporte

✓ Redes Neuronales

✓ Árboles de decision

Clasificación

- ✓ Técnica que permite identificar una instancia a qué clase pertenece
- ✓ ¿Cuál categoría?
 - ¿La imagen es un gato o un perro?
 - ¿El cliente es perfil de riesgo alto, medio o bajo?
 - ¿El tweet es positivo, negativo o neutro?



Clasificación

✓ Técnicas:

✓ K-NN: K-vecinos más cercanos

- Se basa en similitud (distancia)
- Buen desempeño en instancias difíciles de explicar
- Requiere gran cantidad de memoria
- Se ve afectado por datos atípicos

✓ Regresión logística

- Probabilidad de que una instancia pertenezca o no a una clase

✓ Naïve Bayes

- Se basa en probabilidades
- Es capaz de tener en cuenta las características que parecen insignificantes (características independientes)
- Permite seleccionar las mejores instancias
- Poca información de falsos positivos y negativos
- Usa solo valores categóricos
- Modelos eficientes y rápidos

Clasificación

✓ Técnicas:

✓ Árboles de decisión

- Divide el problema en partes
- Los modelos son fáciles de comprender
- Requieren definir criterio de parada (Pre-poda, post-poda)
- Si se requiere postpoda requiere muchos recursos

✓ Reglas de clasificación

- Modelos basados en reglas
- Fácil de comprender
- Características nominales
- Pueden ser utilizadas para identificar datos atípicos

✓ Maquinas de vectores de soporte

- Buscar un hiperplano que separe lo mejor posible las clases.
- Pueden usar muchos tipos de **funciones del núcleo** que permiten encontrar una separación no lineal de las clases

✓ Redes Neuronales

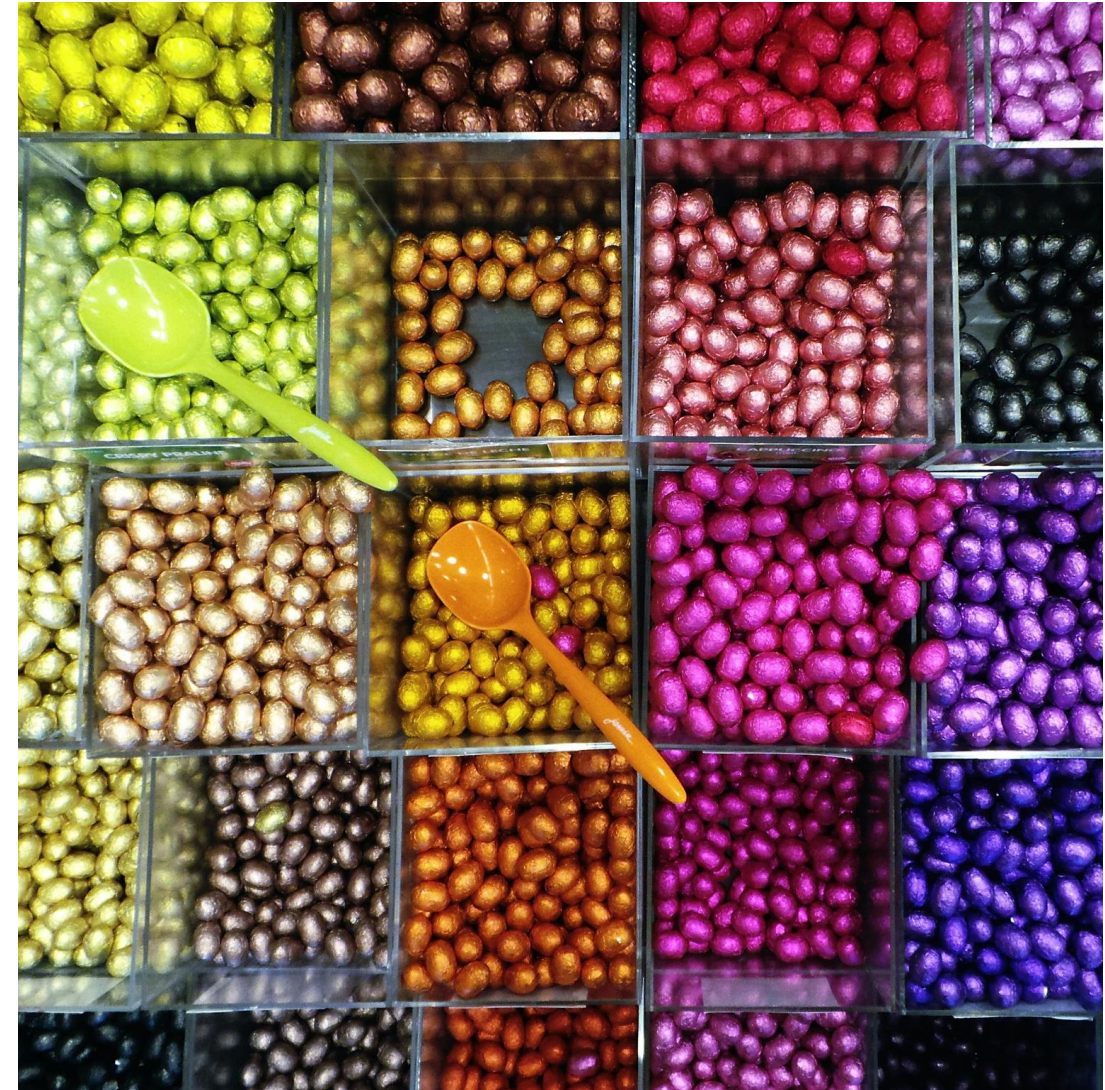
- Gran capacidad de ser utilizada como un mecanismo de función de aproximación arbitraria que "aprende" a partir de datos observados
- No es fácilmente comprensible el por qué de su respuesta

A group of lemurs, likely ring-tailed lemurs, are gathered together in a natural setting. They have white bodies with black and white striped tails and faces. Many of them are looking directly at the camera with bright orange eyes. The background is dark and out of focus, showing some tree branches.

6. Modelos NO Supervisados

Clustering (Agrupamiento)

- ✓ Técnica que permite clasificar de acuerdo con propiedades de grupos homogéneos (agrupación natural)
- ✓ ¿Cuáles grupos?
 - ¿Cuáles compradores tienen gustos similares?
 - ¿Cuáles temas se están hablando en redes sociales?
 - ¿Cuáles tiendas son similares?



Clustering

✓ Técnicas:

✓ K-medias

- Se debe definir el número de grupos (k)
- Se emplea el algoritmo de K-medias
- Cada grupo debe ser analizado y etiquetado manualmente

✓ BDSCAN

✓ Hierarchical clustering

- Agglomerative Clustering
- Divisivo

Los resultados del *hierarchical clustering* pueden representarse como un árbol en el que las ramas representan la jerarquía con la que se van sucediendo las uniones de *clusters*

✓ SPECTRAL CLUSTERING

Correlaciones

- ✓ La **correlación** es una medida estadística que nos ayuda a entender cómo dos variables se relacionan entre sí.
- ✓ ¿Existe una **dependencia** entre las 2 o mas variables?
 - Correlación entre el consumo de tabaco y el riesgo de cáncer de pulmón
 - Correlación entre el nivel de educación y el ingreso salarial



Correlaciones

✓ Técnicas:

✓ Coeficiente de Correlación de Pearson

- Mide la relación lineal entre dos variables continuas.
- Varía entre -1 y +1.

✓ Coeficiente de Correlación de Spearman

- Evalúa la relación entre variables ordinales o no lineales.
- Utiliza rangos en lugar de valores exactos.
- Útil cuando los datos no siguen una distribución normal

✓ Coeficiente de Correlación de Kendal

Asociaciones,
dependencias o
correlaciones



Asociaciones

Reglas de asociación (Apriori)

- ✓ Reglas accionables que son fáciles de entender y ofrecen conocimientos accionables.
- ✓ Ej: {colchón} → {almohada}
- ✓ Reglas triviales que son claras, pero dan algo de valor adicional.
- ✓ Ej: {zapatos} → {correa}
- ✓ Reglas inexplicables que no son claras y no ofrecen ningún conocimiento práctico.
- ✓ Ej: {pañales} → {cerveza}.



Técnicas de reducción de datos

Reducción de la dimensionalidad:

- ✓ **Selección de características**
 - Selección hacia adelante
 - Eliminación hacia atrás
 - PCA
 - RFE
 - Inducción del árbol de decisión
- ✓ **Extracción de características**

Discretización de los datos:

- ✓ **Binning**
- ✓ **Agrupamiento**



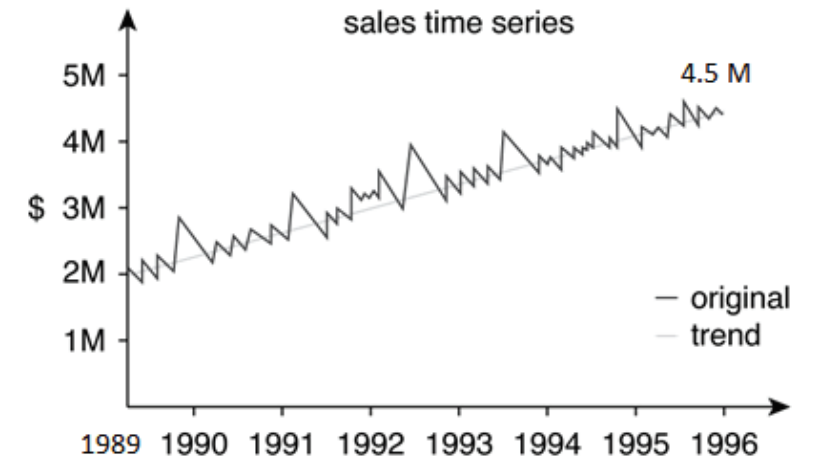
7. Otros temas de aprendizaje

Series temporales

Series temporales

Agrupar una serie de datos recopilados cronológicamente en intervalos de tiempos constantes

- Tendencia
- Estacionalidad



Minería de texto

Técnicas de minería de texto para:

- Recuperar información
- Clasificar documentos, según categorías conocidas
- Encontrar grupos de documentos similares
- Análisis de sentimientos



Minería de texto

Técnicas

Bolsas de palabras

Ej: Doc 1: Better three hours too soon than a minute too late
Doc 2: Better a witty fool than a foolish wit

	Better	three	hours	minute	late	fool	wit
Doc 1	1	1	1	1	1	0	0
Doc 2	1	0	0	0	0	2	2

Frecuencias de términos

Frecuencia inversa de documento

Se deben aplicar técnicas de normalización, stemming y eliminación de stopwords



Minería de texto

Técnicas

N-gramas

Bi-gramas	Tri-gramas
smoking_patient	smoking_patient_ with
patient_with	
with_lung	patient_with_lung
lung_cancer	with_lung_cancer



Extracción de entidades con nombre:

- Nombre: Personas, lugares, empresas
- Patrón: coordenadas, códigos
- Conceptos: un automóvil, un humano
- Hechos: vínculos entre entidades
- Sentimientos: Actitudes, gestos o emociones

Detección de datos atípicos (outliers)

Se diferencia del ruido este en que este no es suficientemente importante para marcarlo como atípico

¿Cuál es extraño?

- ¿Un cliente puede registrar tantas facturas en un día?
- ¿Un cliente puede comprar tanto en un día?



Detección de datos atípicos (outliers)

Tipos:

Global o puntual: Un dato suficientemente inconsistente

Contextual: es consistente dentro de un contexto

Colectivo: es atípico si está combinado con otro dato similar sin condiciones sin contextos



Detección de datos atípicos (outliers)

Técnicas:

Estadísticas

- Operan con base en un ajuste de distribución

Ej: Valores que se encuentren a 3 desviaciones estándar son atípicos
- Valores que no pertenecen a un bin o pertenece al bin de puntuación alta
- Valores que no pertenecen al rango intercuartil

Basadas en distancias

- Agrupamiento K-medias
- K-NN

Detección de datos atípicos (outliers)

Técnicas:

Supervisadas

- Con algunos datos de ejemplo desarrollar un modelo de detección de datos atípicos

Semi-supervisadas

- Primero se realiza un agrupamiento
- Todos los puntos o clusters individuales que no pertenezcan a un cluster son considerados atípicos

THANK YOU

GRACIAS
ARIGATO
SHUKURIA

DANKSCHEEN
JUSPAXAR

TASHAKKUR ATU
YAQHANYELAY
SUKSAMA
EKHMET
MEHRBANI
MAAKE
GRAZIE
MEHRBANI
PALDIES
BOLZİN
MERCI

BIYAN
SHUKRIA

TINGKI
MINMONCHAR

SPASSIBO
SNACHALHUYA
NUHUN
CHALTU
WABEEJA
MAITEKA
HUI
YUSPAGARATAM
UNALCHEESH
SPASIBO
DENKAUJA
HEHACHALHYA
MEERS
ATTO
ANHA
DHANYABAD
SAHCO
MERASTAWHY
GAEJTHO
AGUYJE
FAKAARUE
KOMAPSUMNIDA
LAH
BAIKA
TAVYAPUCH
MEDAWAGSE