Ingeniería en Ciencias de la Computación

Estadística para Ciencias de la Computación

Remigio Hurtado

2019

Introducción

Esta asignatura es parte de las integraciones curriculares del proyecto. Se relaciona curricularmente con Probabilidad y Estadística integrando los Fundamentos Teóricos mediante los elementos relacionados con procesos estocásticos, cadenas de Markov, el análisis multivariado, y en la praxis profesional la posibilidad de resolver y generar varios tipos de programas informáticos que brinden soluciones innovadoras para problemas complejos y requerimientos especiales de los usuarios.

Objetivos:

1. Describir distribuciones de funciones lineales de variables aleatorias mediante momentos
2. Identificar y calcular soluciones a problemas que involucran procesos estocásticos
3. Diseñar cadenas de Markov para problemas básicos
4. Identificar las técnicas de análisis multivariado

UNIDAD 1. MOMENTOS

Temas:

- Repaso de probabilidad
- Funciones generatrices

1.1 REPASO DE PROBABILIDAD

La probabilidad es la ciencia de la incertidumbre. Da reglas matemáticas precisas que permiten comprender y analizar nuestra propia ignorancia. Ese conocimiento nos ayuda a hacer predicciones, tomar decisiones, valorar los riesgos e incluso ganar dinero.

Modelo de probabilidad

1. **S**: es el espacio muestral (conjunto no vacío) S= {lluvia, nieve, despejado}

2. Conjunto de **sucesos**: subconjuntos de S. Se les puede asignar una probabilidad.

Subconjuntos: {lluvia, nieve} {lluvia} Ø={} es vacío

3. **P**: medida de probabilidad. Probabilidad a cada suceso A. P(A).

P(A)= número real no negativo entre 0 y 1

P(Ø)=0

P(S)=1

P es aditiva -> $P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots$

P([a-b])=b-a siempre que $0 \leq a \leq b \leq 1$. La probabilidad del intervalo es simplemente su longitud -> Distribución Uniforme en [0,1]

1.1 REPASO DE PROBABILIDAD

Diagramas de Venn y subconjuntos: proporcionan un método gráfico muy útil para representar el espacio muestral y subconjuntos del mismo.

Probabilidad Uniforme

S si es finito, una posible medida de probabilidad es la probabilidad uniforme.

La probabilidad uniforme **asigna una probabilidad 1/|S| a cada resultado**. |S| es el número de elementos de S o cardinalidad de S.

Por aditividad: P(A)= |A|/|S| Se requiere determinar los tamaños de los conjuntos A y S (mediante principios de combinatoria: principio del producto, enumeración de permutaciones, enumeración de subconjuntos)

Ejemplos:

1 Moneda: S = {cara, cruz} |S|=2 P({cara})=P({cruz})=1/2

3 Monedas: S = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT} |S|=8

P({HHH})=1/8 P({HHH,TTT})=2/8=1/4

Exactamente 2 caras= P({HHT,HTH,THH})=1/8 + 1/8 + 1/8 = 3/8

1.1 REPASO DE PROBABILIDAD

Probabilidad condicional e independencia

S = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT} P(s)=1/8 para todo s ϵ S

¿Probabilidad de que la primera moneda sea cara?

P ({HHH,HHT,HTH,HTT})=4/8 =1/2 =**0.5**

Información extra = Condición = sabemos que 2 de 3 monedas han salido cara. Esto cambia nuestra información disponible, es decir nuestro grado de ignorancia. Cambian las probabilidades.

P({HHT,HTH,THH})=1 P(s)=1/3

P(cara en la primera moneda | cara en dos monedas) = P({HHT,HTH})=1/3+1/3=2/3

| significa "bajo la condición" o "dado que"

P(suceso | condición)

P(A|B) = representa la fracción de veces que ocurre A sabiendo que ha ocurrido B.

P(A|B) = $\frac{P(A \cap B)}{P(B)}$ El cociente entre la probabilidad que ocurra A y B (ambos) y la prob. que ocurra B.

P(A|B) = P({HHT,HTH}) / P({HHT,HTH,THH}) = $\frac{2/8}{3/8}$ = 2/3 = **0.666 (AUMENTA LA PROBABILIDAD)**

¿Cruz en la primera moneda (A) dado que 2 monedas han salido cara (B)?

P(A|B) = P({THH}) / P({HHT,HTH,THH}) = $\frac{1/8}{3/8}$ = 1/3 = **0.333 (DISMINUYE LA PROBABILIDAD)**

1.1 REPASO DE PROBABILIDAD

Probabilidad condicional e independencia

Fórmula del producto $P(A|B) = \frac{P(A \cap B)}{P(B)}$ -> $P(A \cap B) = P(B)\,P(A|B) = P(A)\,P(B|A)$

Ley de la probabilidad total

$P(B) = P(A_1)\,P(B|A_1) + P(A_2)\,P(B|A_2) + \ldots$

$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + \cdots$

Ejemplo: A1=60% chicas, A2=40% chicos, B=PELO LARGO: 30% chicas, 20% chicos

$P(B) = P(A_1)\,P(B|A_1) + P(A_2)\,P(B|A_2) = (0.6)(0.3) + (0.4)(0.2) = 0.26 = 26\%$

Teorema de Bayes: cuando se conoce P(A), P(B) y P(B|A), se desea conocer P(A|B)

$$P(A|B) = \frac{P(A)}{P(B)}\,P(B|A)$$

Sistemas de 2 etapas (ver ejemplo 1.5.2 Pag 42 [1]):

Fórmula del producto: para calcular probabilidades conjuntas de ambas etapas

Ley de la probabilidad total: para calcular probabilidades de la segunda etapa.

Teorema de Bayes: para calcular probabilidades de la primera etapa.

Sucesos independientes: si $P(A \cap B) = P(A)P(B)$.

Sucesos cuya ocurrencia no afecta la probabilidad de los otros. Es decir, P(A|B)=P(A) y P(B|A)=P(B)

1.1 REPASO DE PROBABILIDAD

Variables aleatorias

Existen formas más sencillas de representar una asignación particular de probabilidad que P. Formas más convenientes para trabajar con P.

Una variable aleatoria asigna un valor numérico a cada posible resultado s de S.

S={lluvia, nieve, despejado} Variable aleatoria X

X=3 si llueve, X=6 si nieva, X=-2,7 si esta despejado

Una variable aleatoria es una función definida sobre el espacio muestral S que asigna valores del conjunto R^1 de todos los números reales.

Dado que las variables aleatorias se definen como funciones de un resultado s y dado también que el resultado s se supone aleatorio (es decir, toma **distintos valores con diferentes probabilidades**), se sigue que el valor de una variable aleatorio será a su vez aleatorio (como su nombre implica).

1.1 REPASO DE PROBABILIDAD

Distribuciones de variables aleatorias

Si X es una variable aleatoria, ¿cuál es la probabilidad de que X tome un valor concreto x?

X=x cuando el resultado se escoge de modo que X(s)=x

Ejemplo: X=3, X=6, X=-2.7

- P(X=3)=P(lluvia)=**0.4**, P(X=6)=P(nieve)=**0.15**, P(X=-2.7)=P(despejado)=**0.45**
- P(X ∈ {3,6})= P(X=3) + P(X=6) = 0.4+0.15 = **0.55**
- P(X<5)= P(X=3) + P(-2.7) = 0.4+0.45 = **0.85**

La distribución de una variable aleatoria X es el conjunto de probabilidades P(X ∈ B) de X pertenecientes a diversos conjuntos.

Ejemplo: ¿Cuál es la distribución de X?

P(x ∈ B) 0.4 si 3 ∈ B, 0.15 si 6 ∈ B, 0.45 si -2.7 ∈ B

P(x ∈ B)= 0.4 I_B (3) += 0.15 I_B(6) += 0.45 I_B (-2.7) ← **La Distribución** $I_B(x)=1$ si $x \in B$, $I_B(x)=0$ si $x \notin B$

→Resulta engorrosa la forma de obtener la distribución para todos los subconjuntos B. Existen funciones más sencillas para obtener una distribución de probabilidad: **funciones de distribución acumulada, funciones de probabilidad y funciones de densidad de probabilidad.**

1.1 REPASO DE PROBABILIDAD

Distribuciones discretas

Una variable aleatoria es discreta si $\sum_x \boldsymbol{P}(\boldsymbol{X} = \boldsymbol{x}) = \mathbf{1}$, es decir si sus probabilidades son iguales a determinados valores.

Principales distribuciones: Uniforme discreta, Degenerada, Bernoulli, Binomial, Geométrica, Binomial negativa, Poisson, Hipergeométrica, Multinomial, Polya, De dos puntos.

Toda la información sobre la distribución de X esta contenida en su función de probabilidad, siempre y cuando sepamos que X es una variable aleatoria discreta.

Distribuciones continuas

Una variable aleatoria es continua si **P(X=x)=0**. Ninguna de las probabilidades puede tener un valor predeterminado para un valor de la variable.

Distribución Uniforme [0,1] P(aXB)=b-a siempre que 0 a b 1 P(X<0)=P(X>1)=0

$f(x) = \begin{cases} 1, & 0 \leq x \leq 1 \\ 0, & otro\ caso \end{cases}$ $\mathrm{P(a \leq X \leq b)} = \int_a^b f(x)dx$ → **Ventaja:** Modificando la función se obtienen otras distribuciones continuas.

$f(x)$ es una función densidad si $f(x) \geq 0$ para todo x ∈ R1 y $\int_{-\infty}^{\infty} f(x)dx = 1$

Principales distribuciones: Uniforme continua, Exponencial, Gamma, Normal, T de Student, Chi-cuadrada, F de Fisher, Erlang, Cauchy, Beta, Laplace, Log-normal, Rayleigh, Weibull, Maxwell, Valor extremo.

1.1 REPASO DE PROBABILIDAD

Valores esperados

El valor esperado de una variable aleatoria es el **valor medio** que esta variable puede tomar.

Ejemplo:

- si la mitad de veces X=0, y la otra mitad X=10, E(X)=5
- si 1/3 Y=6, 2/3 Y=15, entonces (1/3)(6) + (2/3)(15)=2+10=12 E(X)=12

Caso Discreto: E(X)= $\mu_x = \sum_{x\epsilon R^1} x\ P(X = x) = \sum_{x\epsilon R^1} x\ Px(x) = \sum_i x_i\ Pi$

Ejemplo:

P(W=-3)=0.2, P(W=-11)=0.7, P(W=31)=0.1

E(W)=(-3)(0.2)+(-11)(0.7)+(31)(0.1)= -0.6-7.7+3.1= -5.2

Caso totalmente continuo: E(X)=$\int_{-\infty}^{\infty} x f_x(x) dx$

El **valor medio μ_x de X es una medida de la posición** de la distribución de probabilidad de X. La media se desplaza con la distribución de probabilidad.

Los valores de una o dos variables aleatorias también pueden calcularse como la suma de los productos de los valores de la función por sus probabilidades. E(XY)= $\sum_z z\ P(XY = z)$ → Pag 157 [1]

1.1 REPASO DE PROBABILIDAD

Varianza, covarianza y correlación

Estos valores nos permiten obtener información sobre la distribución de variables aleatorias.

El valor medio de X será E(X), esto no nos dice nada sobre el modo en que X tiende a ser E(X).

La **varianza de una variable aleatoria:** σ_x^2 = var(X) = E((x- μ_x)^2). **Desviación estándar:** $\sigma_x = \sqrt{var(X)}$ → evita efecto escala varianza

La varianza es una medida de lo dispersa que es la distribución de X, o cuan aleatoria es X, o cuanto varía X.

Ejemplo:

$$P_X(x) = \begin{cases} 1, & x = 10 \\ 0, & otro\ caso \end{cases}$$

$$P_Y(y) = \begin{cases} \frac{1}{2}, & Y = 5 \\ \frac{1}{2}, & Y = 15 \\ 0, & otro\ caso \end{cases}$$

E(X)=E(Y)=10

Var(X) = (10-10)^2(1)=0

Var(Y) = (5-10)^2 (1/2) + (15-10)^2 (1/2)=25

X e Y tienen el mismo valor esperado. La varianza de Y es mucho mayor que la de X. Y es más aleatoria (cambiante) que X, varía más de lo que hace X.

1.1 REPASO DE PROBABILIDAD

Varianza, covarianza y correlación

Propiedades de la varianza:

a) Var (X)≥ 0

b) a y b dos números reales var(aX+b) = a^2var(X)

c) Var(X) = E(X^2) - $(\mu_x)^2$ = E(X^2) - E(X)^2

d) Var (X) $\leq$ E(X^2)

El **efecto de la varianza** es la dispersión de cada distribución respecto a su valor medio.

Si la varianza aumenta, aumenta la dispersión.

Si la varianza disminuye, la distribución resulta más “concentrada” sobre el valor de la media.

1.1 REPASO DE PROBABILIDAD

Varianza, covarianza y correlación

La **Covarianza** determina la relación entre dos variables aleatorias.

$$\mathbf{Cov(X,Y)=E((X-\mu_x)(Y-\mu_y))}$$

Ejemplo:

$$\mathbf{P_{X,Y}(x,y)=}\begin{cases}\frac{1}{2} & x=3, y=4\\ \frac{1}{3} & x=3, y=6\\ \frac{1}{6} & x=5, y=6\\ 0 & otro\ caso\end{cases}$$

E(X)= (3)(1/2) + (3)(1/3) + (5)(1/6) = 10/3

E(Y)= (4)(1/2) + (6)(1/3) + (6)(1/6) = 5

Cov (X,Y)=E((X-10/3)(Y-5))= (3-10/3)(4-5)/2 + (3-10/3)(6-5)/3 + (5-10/3)(6-5)/6 = 1/3

Linealidad: Cov(X,Y)= E(XY) - E(X)E(Y)

Si X e Y son independientes: Cov(X,Y)=0. Cov(X,Y)=0 no implica independencia en todos los casos.

1.1 REPASO DE PROBABILIDAD

Varianza, covarianza y correlación

La **Correlación (coeficiente de correlación)** es una medida del grado de existencia de una relación lineal entre X e Y.

$$corr(X,Y) = \frac{cov(X,Y)}{Sd(X)Sd(Y)} = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}}$$

$$-1 \leq corr(X,Y) \leq 1$$

$corr(X,Y) = 1$ → Y aumenta X aumenta

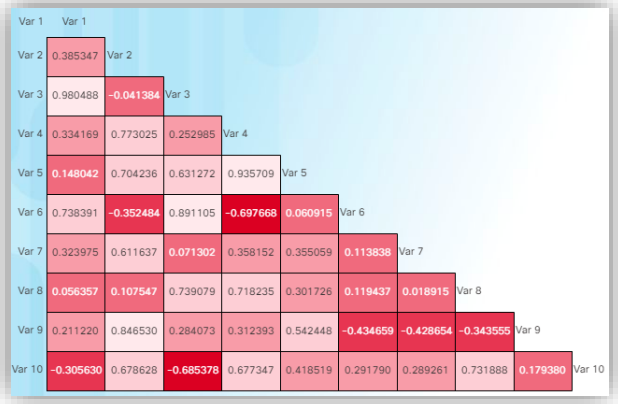
$corr(X,Y) = -1$ → Z disminuye X aumenta

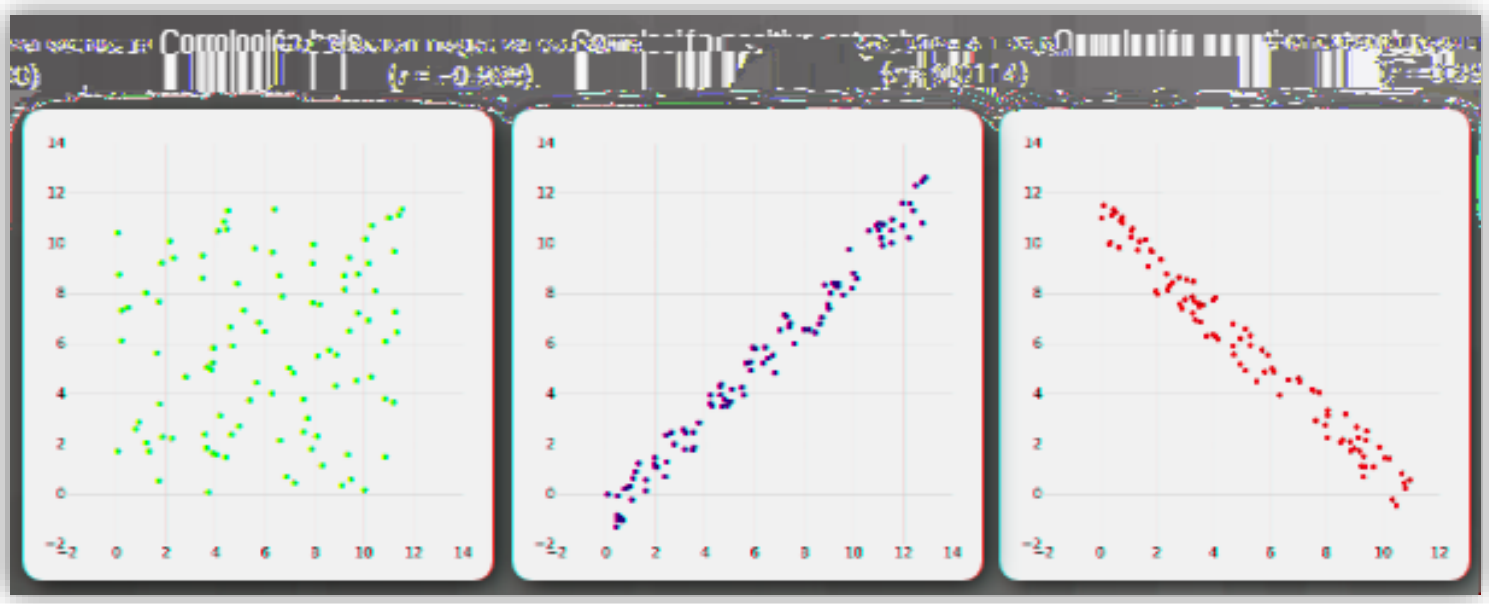
Recursos para análisis de correlación y varianza:
https://rpsychologist.com/d3/correlation/

Mapa de Calor

Python
import seaborn as sns

Práctica 3.1.5.5





1.2 FUNCIONES GENERATRICES

Función de distribución acumulada de X $F_X(x)=P(X\leq x)$, contiene toda la información de la distribución de probabilidad de X. En muchas situaciones, no es tan fácil calcular la media, varianza y otros momentos de una variable directamente. A menudo se pueden usar funciones generadoras para derivar los momentos indirectamente.

Otras funciones también contienen información:

- Función generatriz de probabilidad
- Función generatriz de momentos

Función generatriz de probabilidad

$$r_x(t)= E(t^x) \qquad t \in R^1$$

$$r_X(0)= P(X=0)$$

$$r'_X(0)= P(X=1)$$

$$r''_X(0)= 2P(X=2)$$

...

$$r_X^{(k)}(0)=k!\ P(X=k)$$

Una vez que conocemos la función generatriz de probabilidad $r_x(t)$, podemos calcular todos los valores de probabilidad P(X=0), P(X=1), P(X=2), etc.

1.2 FUNCIONES GENERATRICES

Función generatriz de probabilidad

Ejemplo:

Distribución Binomial (n,θ): i éxitos de n ensayos. θ es la probabilidad de éxito.

$$r_x(t) = E(t^x) = \sum_{i=0}^{n} P(X=i)\, t^i = \sum_{i=0}^{n} \binom{n}{i} \theta^i (1-\theta)^{n-1} t^i = (t\theta + 1 - \theta)^n$$

$$r_X(0) = P(X=0) = (1-\theta)^n$$

$$r'_X(0) = P(X=1) = n(1-\theta)^{n-1}\theta$$

$$r''_X(0) = 2P(X=2) = n(n-1)(1-\theta)^{n-2}\theta^2$$

Podemos calcular todas las probabilidades de X a partir de r_x (al menos en el caso discreto).

Si X e Y toman valores {0,1,2,…} y $r_X = r_Y$ -> X=Y tienen la misma distribución.

Sólo se utiliza la función generadora de probabilidades para variables discretas y no negativas. Una función más general, que se puede utilizar para cualquiera variable es la función generadora de momentos.

1.2 FUNCIONES GENERATRICES

Función generatriz de momentos

$$m_x(s) = r_x(e^s) = E(e^{sX}) \quad s \in R^1$$

Una vez conocida la función generatriz de momentos $m_x(s)$, podemos calcular todos los momentos $E(X)$, $E(X^2)$, $E(X^3)$, etc.

$$m_x(0) = 1$$

$$m'_x(0) = E(X)$$

$$m''_x(0) = E(X^2)$$

$$\dots$$

$$m^{(k)}_x = E(X^k)$$

Ejemplo:

Distribución Binomial (n,θ): $r_Y(t) = (t\theta + 1 - \theta)^n$ → $m_Y(s) = r_y(e^s) = (e^s\,\theta + 1 - \theta)^n$

$m'_Y(s) = n\theta e^s(\theta e^s - \theta + 1)^{(n-1)}$ → $m'_y(0) = E(Y) = n\,\theta$

$m''_Y(s) = n\theta e^s(\theta e^s - \theta + 1)^{(n-2)}(n\theta e^s - \theta + 1)$ → $m''_Y(0) = E(Y^2) = n^2\theta^2 - n\theta^2 + n\theta$

$Var(Y) = E(Y^2) - (E(Y))^2 = n^2\theta^2 - n\theta^2 + n\theta - n^2\theta^2 = -n\theta^2 + n\theta = n\theta\,(1 - \theta)$

1.2 FUNCIONES GENERATRICES

Función generatriz de momentos

$$m_x(s) = r_x(e^s) = E(e^{sX}) \quad s \in R^1$$

Una vez conocida la función generatriz de momentos $m_x(s)$, podemos calcular todos los momentos $E(X)$, $E(X^2)$, $E(X^3)$, etc.

$$m_x(0) = 1$$

$$m'_x(0) = E(X)$$

$$m''_x(0) = E(X^2)$$

...

$$m^{(k)}_x = E(X^k)$$

Ejemplo:

Distribución Exponencial (λ): $m_x(s) = E(e^{sX}) = \int_{-\infty}^{\infty} e^{sX} f_x(x)dx = \int_0^{\infty} e^{sX} \lambda e^{-\lambda x} dx = \lambda(\lambda - s)^{-1}$

$m'_x(0) = E(X) = 1/\lambda$

$m''_x(0) = E(X^2) = 2/(\lambda^2)$

$Var(X) = E(X^2) - (E(X))^2 = 2/(\lambda^2) - 1/\lambda = 1/\lambda^2$

1.2 FUNCIONES GENERATRICES

Momentos:

- **Primer momento E(X):** la media de la variable aleatoria. Medida de la posición sobre la recta real en que se localiza el centro de probabilidad de X, al menos cuando la distribución es unimodal (un solo máximo) y no sea demasiado sesgada.
- **Segundo momento $E(X^2)$:** junto con el primer momento, nos da la varianza a través de la relación $Var(X)= E(X^2) – (E(X))^2$
- Los dos primeros momentos nos informan de la posición y dispersión (o grado de concentración) de la distribución respecto a su media.
- Lo momentos de orden superior también proporcionan información sobre la distribución. Con el tercer momento se obtiene el **Sesgo (Oblicuidad/Asimetría/Skewness)**. Con el cuarto momento se obtiene la **Curtosis (Apuntamiento/Kurtosis)**. La mayoría de distribuciones poseen función generatriz de momentos.
- Las funciones generatrices pueden ayudarnos con las distribuciones compuestas.

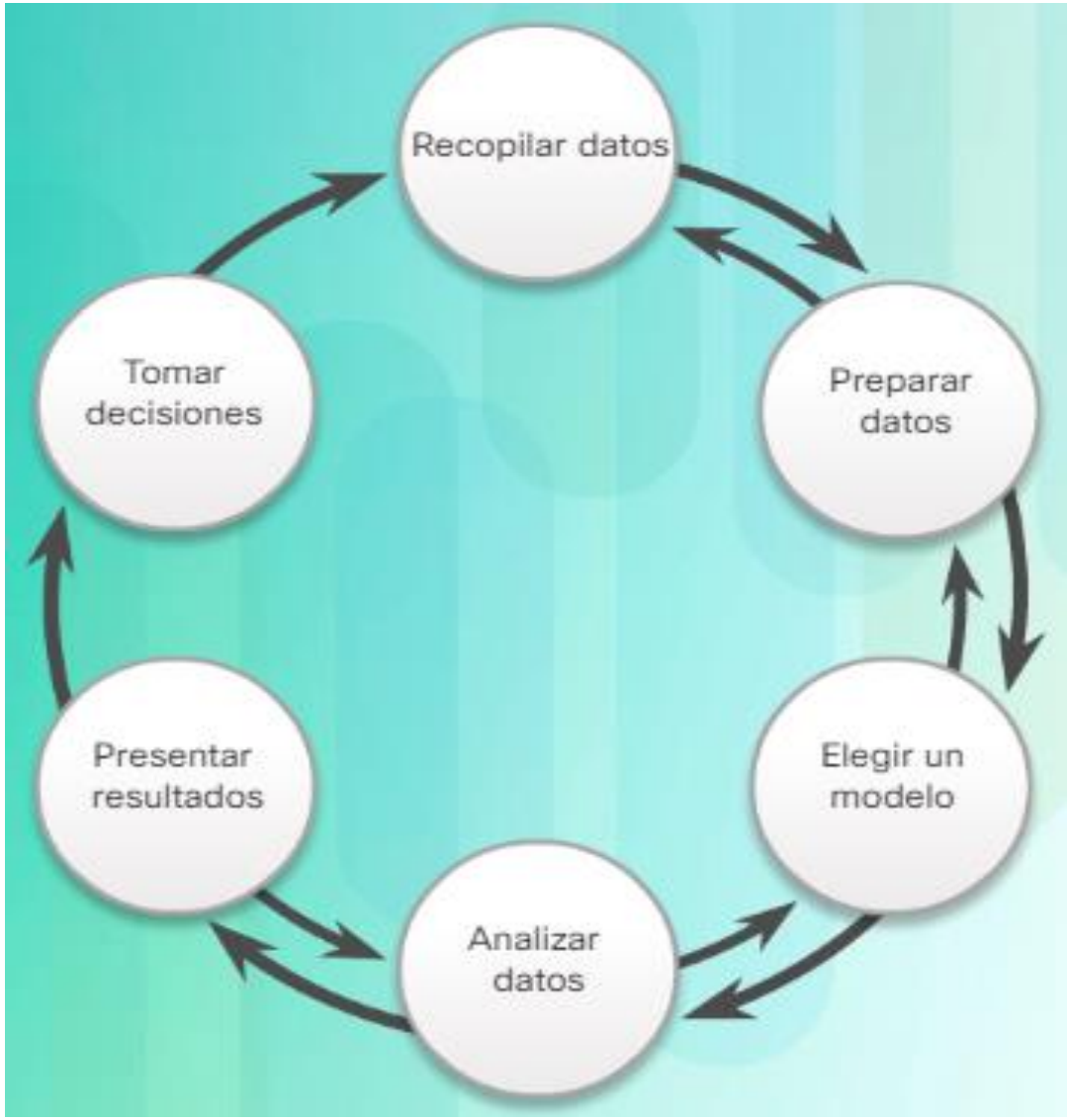
UNIDAD 4. ANALISIS MULTIVARIANTE

2 INTRODUCCIÓN

Las máquinas, las personas y las cosas **generan datos a un ritmo sin precedentes**.

El surgimiento de estos grandes conjuntos de datos **requiere métodos, tecnologías e infraestructura más avanzados para manejar los datos y convertirlos en información procesable.** Los datos ya no se pueden guardar en unas pocas máquinas o procesar con una sola herramienta.

Existen **muchas metodologías para realizar análisis de datos**, incluido el proceso estándar a través de la industria para minería de datos (CRISP-DM) utilizado por más del 40 % de los analistas de datos. Cerca del 27 % de los analistas de datos utilizan su propia metodología. El resto usa diversas metodologías.



2 PYTHON

Existen diversos programas que se utilizan para formatear datos, limpiarlos, analizarlos y visualizarlos. Python se creó en 1991 como un lenguaje fácil de aprender con muchas bibliotecas utilizadas para manipulación de datos, aprendizaje automático y visualización de datos. Python es un lenguaje flexible que está creciendo y volviéndose cada vez más importante para la ciencia de datos gracias a su flexibilidad y facilidad de aprendizaje.

Jupyter Notebook permite que la instrucción y la programación formen parte del mismo archivo (libreta de anotaciones). Es fácil alterar código en las computadoras portátiles y experimentar cómo diferentes códigos se pueden utilizar para manipular, analizar y visualizar datos.

Bibliotecas:

NumPy: agrega soporte para matrices, tiene muchas funciones matemáticas.

Pandas: agrega soporte para las tablas y las series de tiempo, se utiliza para manipular y limpiar datos, entre otras acciones. Agrega estructuras de datos de alto rendimiento y herramientas para el análisis de grandes conjuntos de datos.

Matplotlib: agrega soporte para la visualización de datos. Matplotlib es una biblioteca de trazado de gráficos capaz de crear desde simples diagramas de línea hasta complejos diagramas 3D y de contorno.

Práctica 1.3.2.9: Desafío de Python

2 Toma de decisiones

Los cuatro aspectos de los datos masivos son: volumen, velocidad, variedad y veracidad.

Las metodologías estadísticas incorporadas en aplicaciones les facilitan la tarea a los analistas de datos para interpretar y utilizar datos masivos a fin de tomar mejores decisiones. **Las herramientas de análisis de datos modernas permiten extraer y transformar los datos sin procesar para presentar un conjunto mucho más pequeño de datos de calidad.** Sin embargo, los datos por sí mismos no son información significativa; se deben analizar y luego presentar en un formato que se pueda interpretar.

Por **ejemplo**, en **política**, es habitual que los analistas de datos extraigan la información relevante para su candidato. En los **negocios**, un analista de datos puede descubrir tendencias de mercado que permitan a una empresa moverse un paso adelante de la competencia.

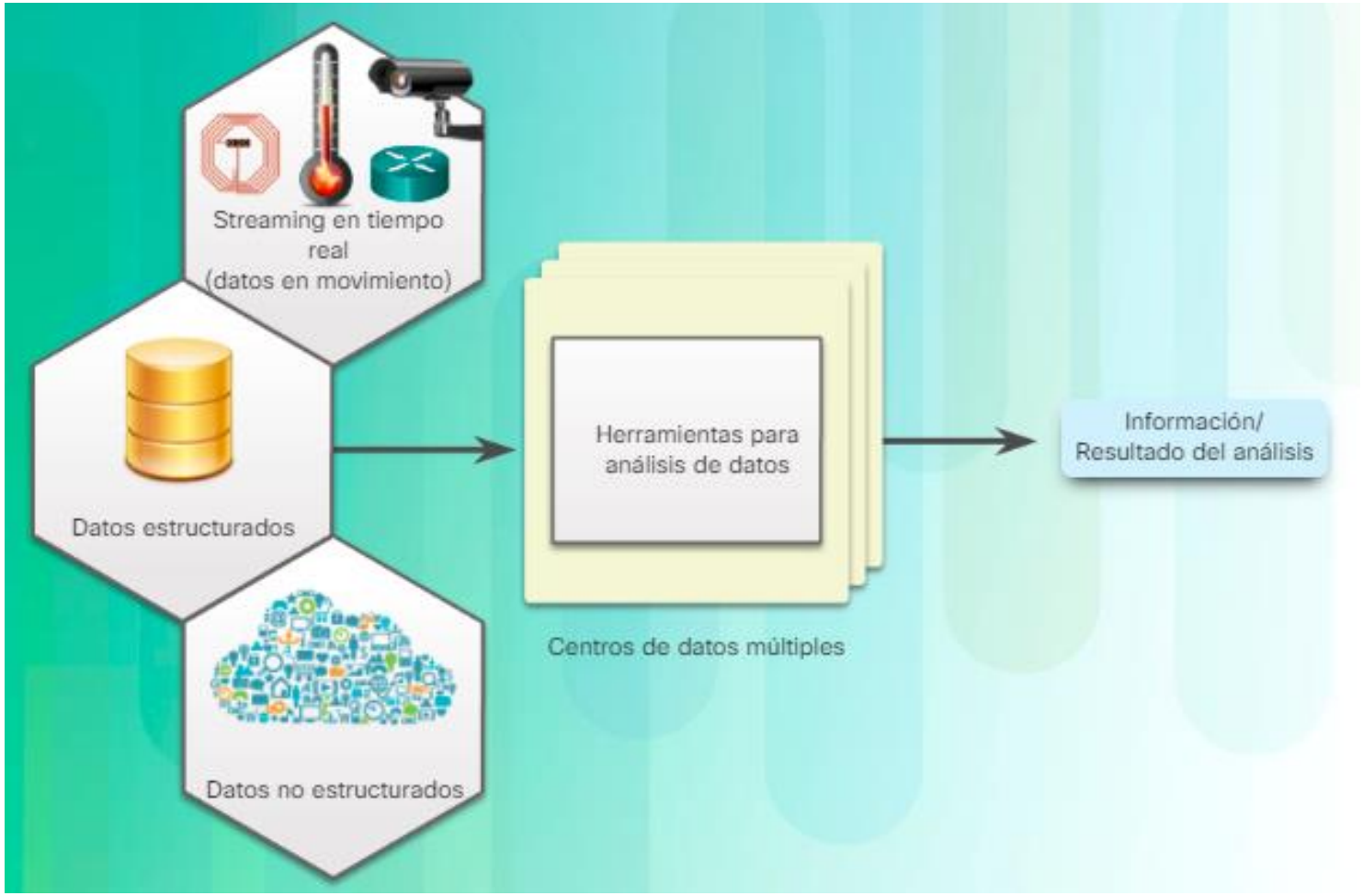
2 Tipos de análisis

El **análisis descriptivo** se utiliza para identificar las características principales de un conjunto de datos. El análisis descriptivo depende únicamente de datos históricos para brindar informes regulares sobre eventos que ya se han producido.
El **análisis predictivo** intenta predecir qué podría ocurrir con cierto grado de seguridad, en función de datos y estadísticas. El análisis predictivo se puede utilizar para deducir datos faltantes y establecer una línea de tendencia futura según los últimos datos. Usa modelos y predicciones de simulación para sugerir qué podría ocurrir.
El **análisis prescriptivo** predice resultados y sugiere medidas de acción que lograrán el mayor beneficio para la empresa u organización.

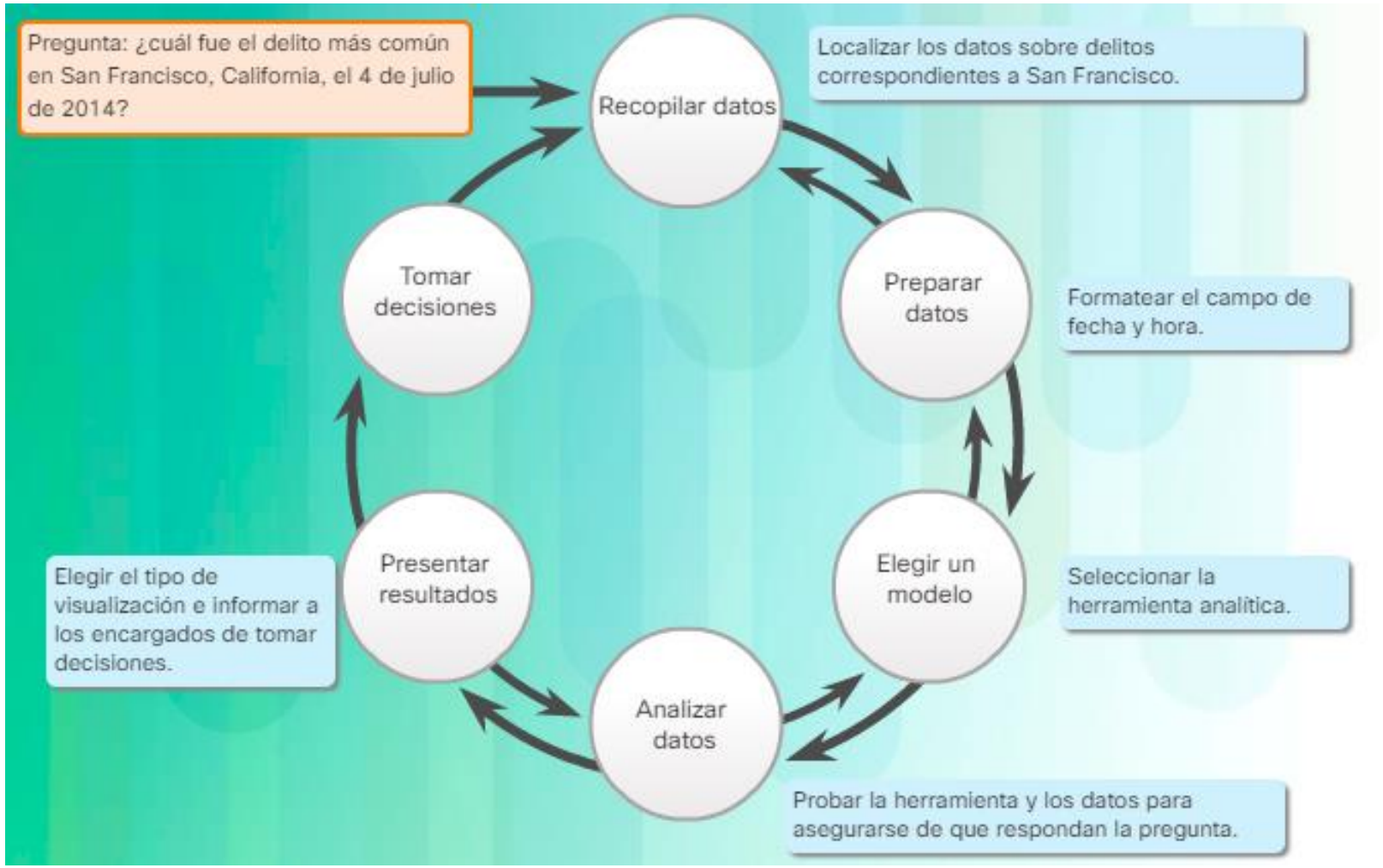
Tipo	Tareas	Preguntas
Descriptiva	Generación de informes estándar	¿Qué ocurrió?
	Informes ad hoc	¿Qué cantidad, con qué frecuencia, dónde?
	Consultas de datos	Exactamente, ¿cuál es el problema?
Predictivo	Simulación	¿Qué podria suceder?
	Pronósticos	¿Qué sucederia si continúan estas tendencias?
	Modelado predictivo	¿Qué sucederá a continuación?
Prescriptiva	Optimización	¿Cómo podemos obtener el mejor resultado?
	Optimizaciones bajo incertidumbre	¿Cómo podemos lograr el mejor resultado, dada la variabilidad?

2 Analítica de nueva generación

A medida que los conjuntos de datos crecen en volumen, velocidad y variedad, la complejidad del almacenamiento, procesamiento y agregado de datos se convierte en un desafío para las herramientas analíticas tradicionales (Excel, SPSS, etc). Los grandes conjuntos de datos se pueden distribuir y procesar a través de dispositivos físicos múltiples y geográficamente dispersos o en la nube. Las herramientas de datos masivos, como **Hadoop y Apache Spark**, son necesarias para que estos grandes conjuntos de datos permitan el análisis en tiempo real y el modelado predictivo.



2 Ejemplo del ciclo del análisis de datos



Recopilación de datos en Internet con web scraping o API RESTful (HTTP y JSON).

Proceso de **web scraping** es un proceso automatizado que utiliza un bot o un rastreador web. Los datos específicos se recopilan y se copian de la web a una base de datos o una hoja de cálculo. Los datos pueden luego analizarse fácilmente.

2 Preparación de datos

Extracción de datos: extraer datos con diferentes formatos, ejemplo de BDs NOSQL y BDs Relacionales para posteriormente transformar a un formato único.

Transformación de datos: combinar datos de varios orígenes, agregar, clasificar, determinar nuevos valores que se calculan de datos agregados y luego aplicar reglas de validación.

Carga de datos: los datos transformados se cargan en la base de datos de destino (pueden sobrescribir los anteriores o cargar nuevos con fechas actuales).

Práctica 2.2.4.5: Analizar un conjunto de datos (Preparar datos, remover columnas y NAN)

2 Python comunicación externa

Escritura y lectura de archivos

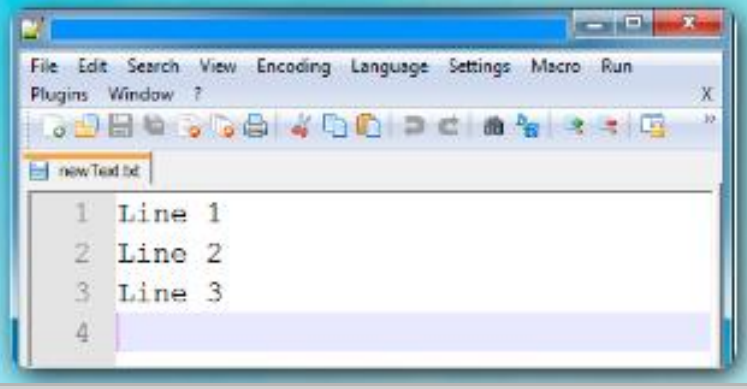
```
myFile = open('newText.txt', "w")
myFile.close()
```

```
myFile = open('newText.txt', "a")
myFile.write('Line 1\n')
myFile.write('Line 2\n')
myFile.write('Line 3\n')
```

```
myFile.close()
myFile = open('newText.txt', "r")
myFile.read()
```

```
'Line1\nLine2\nLine3\n'
```

Contenido del archivo en un editor de texto.



2 Python comunicación externa

Práctica 2.5.1.4: Análisis de datos del medidor de Internet

Ejecución de los comandos del sistema operativo Linux

```
'''Importar la biblioteca de subproceso que es necesaria para la comunicación con
aplicaciones externas'''

import subprocess

#Ahora ejecutamos el proceso de ping a partir de la shell:

pingCmd = 'ping -c 127.0.0.1'
process = subprocess.Popen(pingCmd.split(), stdout=subprocess.PIPE)

'''crea un objeto para mantener el resultado del proceso y dividir los elementos de
salida en una lista"

process_output = process.communicate()[0]
process_output = process_output.split()

#ver el contenido del objeto de resultado
print process_output
```

```
['PING'. '127.0.0.1', '(127.0.0.1)', '56(84)', 'bytes', 'of', 'data.', '64', 'bytes',
 'from', '127..0.0.1', 'icmp_seq=1', 'ttl=64', 'time=0.094', 'ms', '64', 'bytes',
 'from', '127.0.0.1', icmp_seq=2', 'ttl=64', 'time=0.052', 'ms', '---', '127.0.0.1',
 'ping', 'statistics', '---', '2', 'packets', 'transmitted', '2', 'received,', '0%',
 'packet', 'loss,', 'time', '999ms', 'rtt', 'min/avg/max/mdev', '=',
 '0.052/0.073/0.094/0.021', 'ms']
```

```
#Ver los cinco primeros elementos de la lista
 process_output[0:5]
```

```
['PING', '127.0.0.1', '(127.0.0.1)', '56(84)', 'bytes']
```

2 Python y SQL

Prácticas 2.5.2.4 y 2.5.2.5: Trabajar con Python y SQLite

2 Análisis de datos exploratorio

El análisis de datos exploratorio es un **conjunto de procedimientos diseñados para producir resúmenes descriptivos y gráficos de datos con el concepto de es posible que los resultados revelen patrones interesantes**. Es un proceso de detección que nos permite a veces crear una hipótesis sobre los datos. Permite la detección de nuevas preguntas que deben contestarse.

El análisis de datos exploratorio proporciona una manera útil de examinar los datos **para determinar si hay relaciones existentes entre los datos obtenidos o recopilados o si hay problemas con los datos**.

2 Observaciones, variables y valores

Conjunto de datos

Variables

Observación

Raza	Color	Tamaño	Peso (kg)
Caniche	blanco	grande	30
Schnauzer	gris	medio	15
Yorkie	marrón-negro	pequeño	3
Labrador	negro	grande	30
Pitbull	habano	medio	20
Cacatúa	habano	grande	30

Conjunto de observaciones
Valores o puntos de datos

2 Tipos de datos

Categóricos:

Nominales: estas son variables compuestas por dos o más categorías cuyo valor se asigna basado en la identidad del objeto. Algunos ejemplos son sexo, color de ojos o tipo de animal.

Ordinales: estas son variables compuestas por dos o más categorías en las que el orden es importante en el valor. Algunos ejemplos son el rango de clases de los estudiantes o las escalas de las encuestas de satisfacción (insatisfecho, neutro, satisfecho).

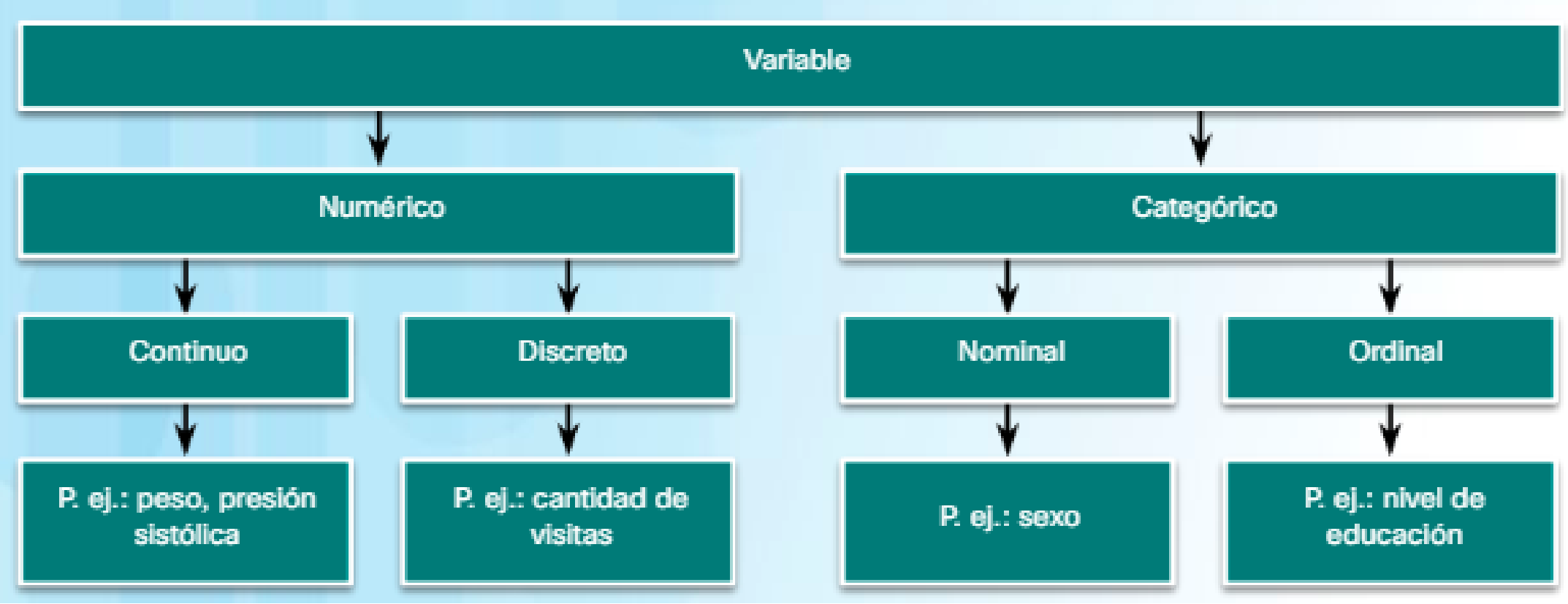
Numéricos:

Continuas: estas son variables que son cuantitativas y pueden medirse a lo largo de una secuencia o un rango de valores. Existen dos tipos de variables continuas; las variables de intervalo puede tener cualquier valor dentro del rango de valores. Algunos ejemplos son temperatura o tiempo. Las variables de relaciones son las variables de intervalo especiales donde un valor de cero (0) significa que no hay ninguna variable. Entre los ejemplos se incluyen ingresos o el volumen de ventas.

Discretas: estos tipos de variables continuas son cuantitativos pero tienen un valor específico de un conjunto de valores finito. Los ejemplos incluyen el número de sensores habilitados en una red, o el número de automóviles en un estacionamiento.

2 Tipos de datos

Algunos métodos estadísticos y visualizaciones de datos están diseñados para trabajar mejor con ciertos tipos de datos que con otros. Cómo se muestran mejor los resultados del análisis dependerá del tipo de variables utilizadas en los datos.



2 Términos del análisis de datos

Término	Definición
Valor	El monto asumido por una variable en una observación específica
Punto de datos	El conjunto de valores para una observación específica
Datos no estructurados	Requiere un procesamiento considerable que tenga significado
Datos estructurados	La naturaleza y el significado de los datos se comprenden
Análisis exploratorio	Un proceso de descubrimiento que puede crear hipótesis para guiar el rumbo del análisis formal
Variable	Características de algo que se ha observado
Conjunto de datos	Conjunto de dos o más variables, con sus valores, que se han combinado entre sí

2 Estadística y Aprendizaje Automático

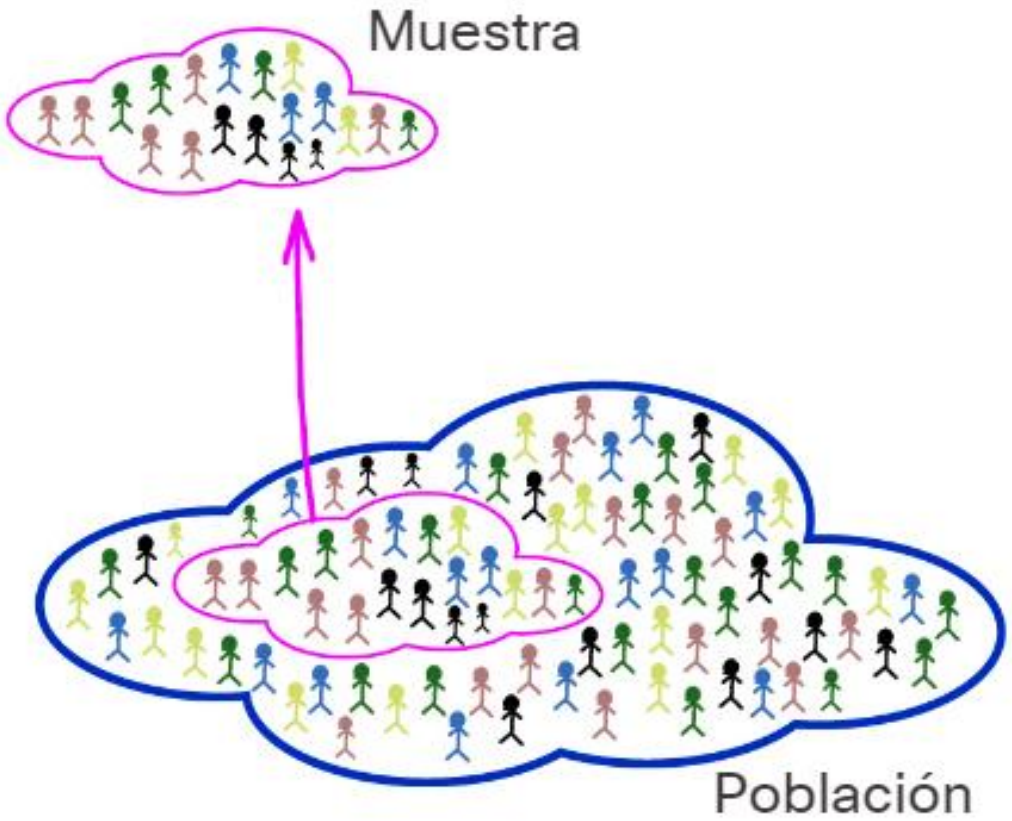
Las **estadísticas** son la recopilación y el análisis de datos mediante técnicas matemáticas. Esto también incluye la interpretación de datos y la presentación de las conclusiones. Otro uso de las estadísticas es detectar los patrones o las relaciones entre las variables y evaluar esos patrones para ver con qué frecuencia surgen.

El análisis adopta generalmente un dominio de herramientas más amplio que las estadísticas. El análisis utiliza las herramientas de modelado matemático en las estadísticas además de otras formas de análisis, como el **aprendizaje automático**. También puede implicar la necesidad de trabajar con conjuntos de datos muy grandes, que incluyen datos no estructurados.

2 Estadística: población y muestra

Una **población** es un grupo de entidades similares como personas, objetos o eventos, que comparten un conjunto de características común que se puede utilizar para fines estadísticos o de investigación.

Se puede utilizar para el análisis un grupo representativo de la población. Este grupo se denomina **muestra**. Las muestras se eligen a menudo para representar la mayor población posible de alguna manera. Si este es el caso, debe tenerse particular cuidado al seleccionar la muestra para asegurarse de que todas las características necesarias de la población estén representadas.



2 Estadística Descriptiva y Estadística Inferencial

Las **estadísticas descriptivas** se usan para describir o resumir los valores y las observaciones de un conjunto de datos. Pueden responder preguntas como por ejemplo:

¿Cuán dispersos están los datos?

¿Hay valores que ocurren con más frecuencia que otros?

¿Cuál es el valor más pequeño o más grande?

¿Hay tendencias en particular?

Ejemplo: Si una persona alcanzó sus objetivos de estado físico en seis de los diez días, entonces tuvo un 60 % de éxito.

Si bien las estadísticas descriptivas muestran el estado actual o histórico de la población observada, no tienen en cuenta la comparación de los grupos, las conclusiones que se deben extraer, o las predicciones que se harán sobre otros conjuntos de datos que no están en la población. No podemos deducir que la persona tiene una salud deficiente solo porque pudo cumplir su objetivo el 60 % del tiempo. Tampoco podemos utilizar el conjunto de datos para que esta persona prediga el rendimiento físico de otras personas con características similares. Aquí es donde las **estadísticas inferenciales** se vuelven importantes.

2 Estadística Descriptiva

Al tratar con una gran cantidad de datos que provienen de diversas fuentes, pueden producirse muchos problemas. A veces los puntos de datos pueden estar dañados, estar incompletos o faltar por completo. Las **estadísticas descriptivas pueden ayudar a determinar qué parte de los datos de la muestra sirve para el análisis** y a identificar los criterios para quitar los datos que resulten inadecuados o problemáticos.

2 Estadística Inferencial

Las **estadísticas inferenciales** consisten en el proceso de recopilar, analizar e interpretar los datos recolectados de un ejemplo para hacer generalizaciones o predicciones sobre una población. Como se utiliza una muestra representativa en lugar de los datos de la población total en sí, debe tenerse en cuenta la posibilidad de que los grupos determinados elegidos para el estudio o el entorno en el que se realiza el estudio no sean un fiel reflejo de las características del grupo mayor. Al utilizar estadísticas inferenciales, deben responderse las preguntas de cuán cerca están los datos inferidos de los datos reales y cuán seguros podemos estar en la conclusión. Por lo general, estos tipos de análisis incluirán diferentes técnicas de muestreo para reducir el error y aumentar la confianza en las generalizaciones sobre las conclusiones. El tipo de técnica de muestreo utilizado dependerá del tipo de datos.

2 Análisis Multivariante

Definición: El estudio estadístico de varias variables medidas en elementos (observaciones) de una población. Proporciona métodos objetivos para conocer cuantas variables indicadoras (a veces se denominan factores) son necesarias para describir una realidad compleja y determinar su estructura. Es importante entender la estructura de dependencia entre variables, ya que las relaciones entre variables son las que permiten resumirlas en variables indicadoras, encontrar grupos no aparentes por las variables individuales o clasificar casos complejos.

Objetivos específicos:

1. Resumir el conjunto de variables en pocas nuevas variables construidas como transformaciones de las originales, con la mínima pérdida de información. Al tener menos variables (factores/indicadores) estas se pueden representar gráficamente y comparar. Los indicadores pueden llegar a interpretarse y mejorar el conocimiento de la realidad.
2. Encontrar grupos en los datos (si existen). Se descubren grupos sin conocer previamente la cantidad de grupos.
3. Clasificar nuevas observaciones en grupos definidos.
4. Relacionar dos conjuntos de variables.

2 Análisis Multivariante

Aplicaciones:

- En Medicina: identificar tumores mediante imágenes.

- En Ciencias de la Computación: para resumir información, diseñar sistemas de clasificación automática y de reconocimiento de patrones.

- En Agricultura: clasificar cultivos por Teledetección (fotos aéreas, imágenes satelitáles)

- En Psicología: determinar los factores que componen la inteligencia humana.

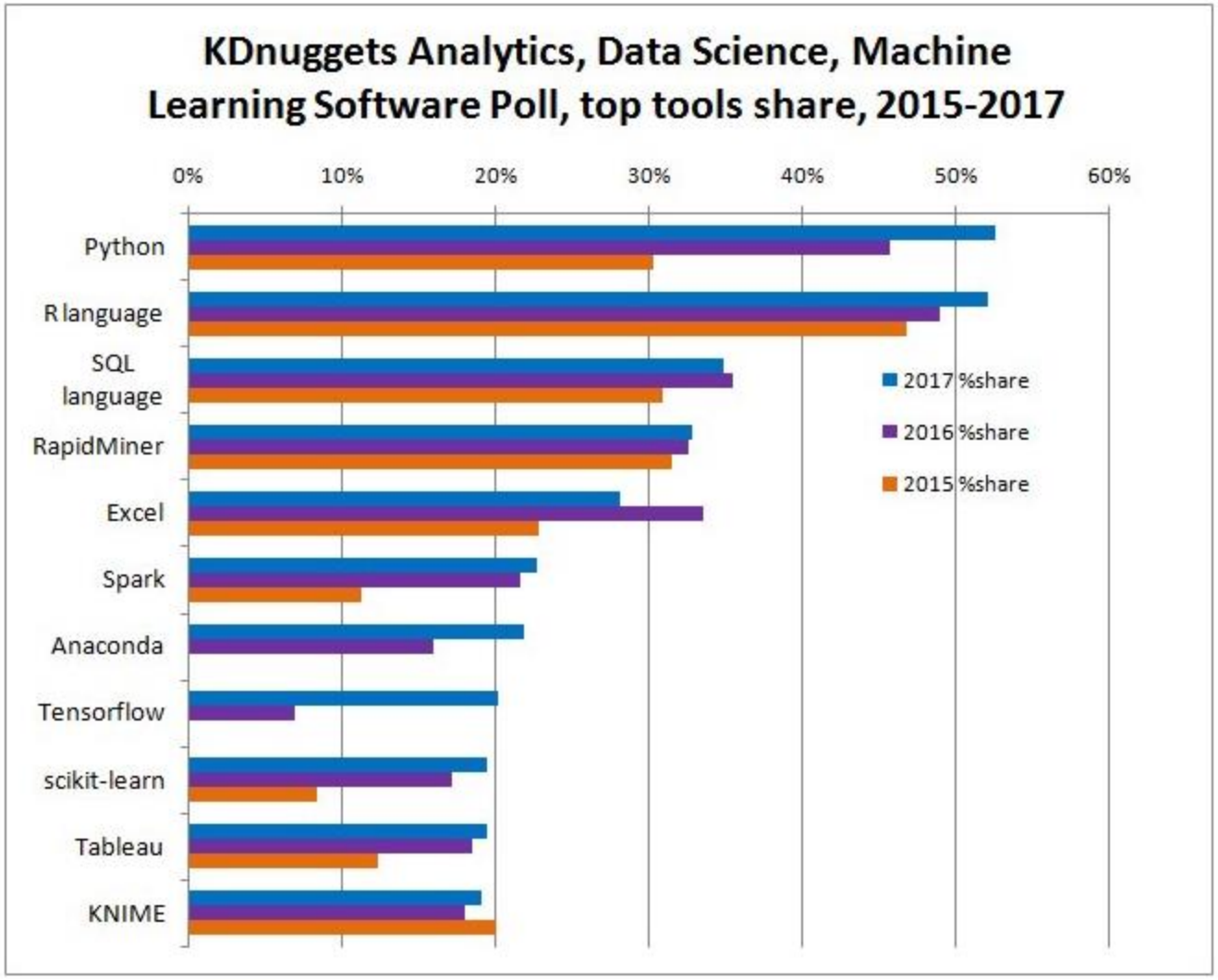
- Otras.

Herramientas:

Statgraphics, Minitab, SPSS, S-PLUS, R, MATLAB, GAUSS, OCTAVE, PYTHON.

2 Análisis Multivariante: Herramientas

https://www.kdnuggets.com/2017/05/poll-analytics-data-science-machine-learning-software-leaders.html



2 Preparación de datos: Limpieza y análisis de Datos

CO	x_1	x_2	x_3
A	1	0	0
V	0	1	0
C	0	0	1
N	0	0	0

Coding de datos categóricos (código binario): Clases A (azul), V (verde), C (cafe) y N (negro). P=4 variables, que son P-1=3 variables binarias (x1,x2,x3). Este procedimiento siempre puede aplicarse pero puede lógicamente dar lugar a muchas variables. Conviene entonces ver si podemos agrupar las clases o categorías para evitar tener variables que casi siempre toman el mismo valor (cero si la categoría es poco frecuente o uno si lo es mucho).

Naturalmente la variable CO podría también haberse codificado dando valores numéricos arbitrarios a las categorías, por ejemplo, A=1, V=2, C=3, N=4, pero esta codificación tiene el inconveniente de sugerir una graduación de valores que puede no existir. Sin embargo, cuando los atributos pueden interpretarse en función de los valores de una variable continúa tiene más sentido codificarla con números que indiquen el orden de las categorías. Por ejemplo, si tenemos empresas pequeñas, medianas y grandes, en función del número de trabajadores, tienen sentido codificarlas con los números 1, 2, y 3, aunque conviene siempre recordar que estos números sólo tienen un sentido de orden.

Normalización – transformación al rango entre 0 y 1
Scaling – transformación a umbrales mínimo y máximo (pueden ser diferentes de 0 y 1)
Estandarización - transformación de datos (mean=0, varianza=1)

2 Preparación de datos: Analizar distribución de datos

Existen varias formas de resumir los datos mediante estadísticas descriptivas. Se puede buscar la **distribución real de los datos, las medidas de la tendencia central o las medidas de rangos**. En un nivel básico, la distribución es una asociación simple entre un valor y la cantidad o el porcentaje de veces que aparece en una muestra de datos. Las distribuciones son útiles para comprender las características de una muestra de datos.

Las **distribuciones de frecuencia** consisten en todos los valores únicos de una variable y la cantidad de veces que aparecen los valores en el conjunto de datos. En **distribuciones de probabilidad**, en lugar de frecuencias, se utiliza la proporción de veces que se presenta el valor en los datos.

Las **funciones de distribución de probabilidad** permiten representar la forma de distribución completa del conjunto de datos utilizando solo un pequeño grupo de parámetros, como la media y la variación. Una función de distribución de probabilidad que se adapta en particular para representar muchos eventos que se producen en la naturaleza es la **gausiana, o distribución normal**, que es **simétrica y acampanada**.

Otras distribuciones no son simétricas. El pico del gráfico podría estar a la izquierda o la derecha del centro. Esta propiedad de una distribución se denomina **asimetría**. Algunas distribuciones tienen dos picos y se conocen como **bimodales**. Los extremos derecho e izquierdo del gráfico de distribución se conocen como **colas**. La media, o promedio, se utiliza actualmente para describir muchas distribuciones.

Calificación original del estudiante

Estudiante	Cuestionario 1 (10 puntos)
Estudiante 1	6
Estudiante 2	7
Estudiante 3	7
Estudiante 4	8
Estudiante 5	7
Estudiante 6	9
Estudiante 7	10
Estudiante 8	8
Estudiante 9	7
Estudiante 10	5

Distribución de calificación

Puntaje	Frequency of Score	Probability of Score
1	0	0
2	0	0
3	0	0
4	0	0
5	1	0.1
6	1	0.1
7	4	0,4
8	2	0.2
9	1	0.1
10	1	0.1

2 Preparación de datos: Limpieza y análisis de Datos

Práctica Estadística Descriptiva con Python: Tabla de Correlación

Práctica 3.2.1.6

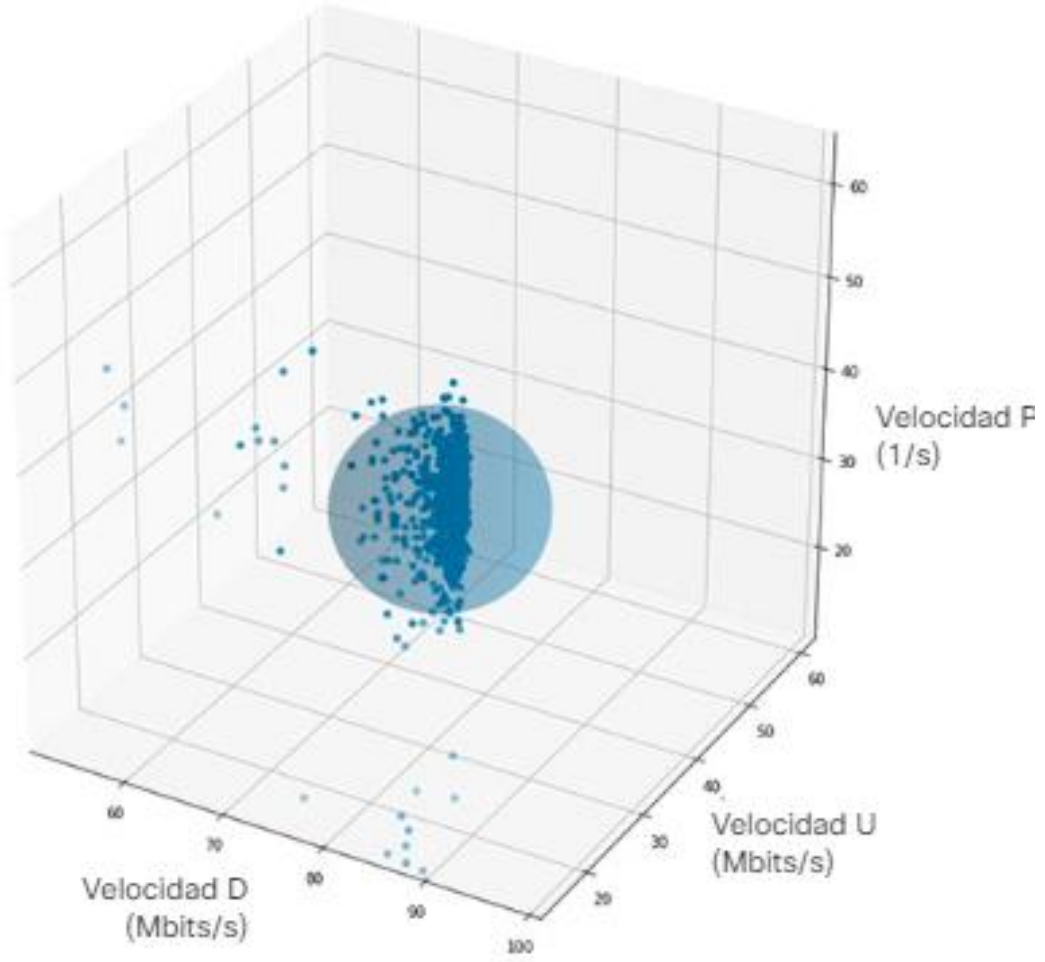
En esta práctica de laboratorio, aprenderá a utilizar la biblioteca de pandas para realizar los pasos preliminares que son necesarios antes de realizar cualquier análisis de datos. Esto incluye la eliminación de valores faltantes, el cambio de formato de datos y la ejecución de análisis estadístico preliminar. Después de limpiar los datos, utilizará matplotlib para la exploración y visualización de datos.

2 Preparación de datos: Detección de anomalías/Datos atípicos

Las anomalías pueden representar datos que son anómalos, o valores que son anómalos. Los datos pueden dañarse o distorsionarse mediante muchos factores durante la medición, la transmisión o el almacenamiento. Estos valores se consideran valores atípicos. Se desvían tanto de los valores esperados que podrían distorsionar los resultados del análisis. Estas consideraciones se suelen eliminar del conjunto de datos después de un estudio detallado.

Para detectar anomalías, deberá identificar el límite de la decisión que define si un punto de datos es normal o es una anomalía. Para ello, primero se normalizan los datos de la distancia al establecer el trayecto más lejano a 1. Luego se determina un umbral entre 0 y 1 que defina el umbral para el límite de la decisión.

Práctica 4.3.2.4 Normalización y Detección de anomalías (Visualización en 3D)



2 Principales técnicas de análisis multivariante

Con la disponibilidad de CPU y GPU de alto rendimiento, es bastante posible resolver problemas de regresión, clasificación, agrupación y otros problemas relacionados mediante el aprendizaje automático y los modelos de aprendizaje profundo. Sin embargo, todavía hay varios factores que causan cuellos de botella en el rendimiento al desarrollar dichos modelos. Un gran número de características en el conjunto de datos es uno de los factores que afectan tanto el tiempo de entrenamiento como la precisión de los modelos de aprendizaje automático. Entrenar a los modelos con el número original de variables, puede llegar a tardar días o semanas si el número de variables es demasiado alto. Hay varias alternativas para lidiar con una gran cantidad de características en un conjunto de datos.

- Reducir el número de variables mediante la fusión de variables correlacionadas.
- Extraer las características más importantes del conjunto de datos que son responsables de la varianza máxima en la salida.

Se utilizan diferentes técnicas estadísticas para este propósito, p. Ej. Análisis discriminante lineal, análisis factorial y análisis de componentes principales.

2 Principales técnicas de análisis multivariante

- **Análisis de Componentes Principales (PCA)**
- **Análisis Discriminante Lineal (LDA)**
- **Análisis con Regresión Multivariable**
- **Análisis Factorial (SVD/MF)**
- **Análisis de Grupos (CLUSTERING)**

2 Análisis de Componentes Principales (PCA)

El análisis de componentes principales, o PCA, es una técnica estadística para convertir datos de alta dimensión en datos de baja dimensión, mediante la selección de las características más importantes que capturan la máxima información sobre el conjunto de datos. Las características se seleccionan sobre la base de la varianza que causan en la salida. La característica que causa la mayor variación es el primer componente principal. La característica que es responsable de la segunda varianza más alta se considera el segundo componente principal, y así sucesivamente. Es importante mencionar que los componentes principales no tienen ninguna correlación entre sí (Al obtener la tabla de correlación entre los componentes principales, los resultados son aproximadamente cero).

Ventajas de PCA

- El tiempo de entrenamiento de los algoritmos se reduce significativamente con menos variables.

- No siempre es posible analizar datos en altas dimensiones. Por ejemplo, si hay 100 características en un conjunto de datos. El número total de gráficos de dispersión necesarios para visualizar los datos sería 100 (100-1) / 2 = 4950. En la práctica, no es posible analizar los datos de esta manera.

2 Análisis de Componentes Principales (PCA)

Es imperativo mencionar que un conjunto de características debe normalizarse antes de aplicar PCA. Por ejemplo, si un conjunto de características tiene datos expresados en unidades de Kilogramos, Años luz o Millones, la escala de variación es enorme en el conjunto de entrenamiento. Si se aplica PCA en un conjunto de características de este tipo, las cargas resultantes para las características con alta variación también serán grandes. Por lo tanto, los componentes principales se desviarán hacia características con alta varianza, lo que conducirá a resultados falsos.

PCA es una técnica estadística y solo se puede aplicar a datos numéricos. Por lo tanto, se requiere que las características categóricas se conviertan en características numéricas antes de poder aplicar PCA.

Revisar material adicional:

- Notebooks de JUPITER sobre PCA.
- Enlace: https://stackabuse.com/implementing-pca-in-python-with-scikit-learn/
- Enlace: https://stackabuse.com/implementing-lda-in-python-with-scikit-learn/

2 Análisis Discriminante Lineal (LDA)

PCA y LDA son técnicas de transformación lineal. Sin embargo, **PCA es no supervisado** mientras **LDA es una técnica de reducción supervisada**.

PCA no tiene ninguna preocupación con las etiquetas de clase. En palabras simples, PCA resume el conjunto de características sin depender de la salida. PCA intenta encontrar las direcciones de la varianza máxima en el conjunto de datos. En un gran conjunto de características, hay muchas características que son simplemente duplicadas de las otras características o tienen una alta correlación con las otras características. Tales características son básicamente redundantes y pueden ser ignoradas. **El rol de PCA es encontrar características altamente correlacionadas o duplicadas y crear un nuevo conjunto de características donde exista una correlación mínima entre las características o, en otras palabras, un conjunto de características con la máxima varianza entre las características. Dado que la variación entre las características no depende de la salida, PCA no toma en cuenta las etiquetas de salida.**

2 Análisis Discriminante Lineal (LDA)

A diferencia de PCA, LDA intenta reducir las dimensiones del conjunto de características mientras retiene la información que discrimina en las clases de salida.

LDA intenta encontrar un límite de decisión alrededor de cada grupo de una clase. A continuación, proyecta los puntos de datos a nuevas dimensiones de manera que los **grupos estén lo más separados posible entre sí** y los **elementos individuales dentro de un grupo estén lo más cerca posible del centroide del grupo**. Las nuevas dimensiones se clasifican en función de su capacidad para maximizar la distancia entre los grupos y minimizar la distancia entre los puntos de datos dentro de un grupo y sus centroides. **Estas nuevas dimensiones forman los discriminantes lineales del conjunto de características.**

2 Análisis Discriminante Lineal (LDA)

En el caso de datos distribuidos uniformemente, LDA casi siempre se desempeña mejor que PCA. Sin embargo, **si los datos están muy sesgados (distribuidos irregularmente) se recomienda utilizar PCA** ya que la LDA puede estar sesgada hacia la clase mayoritaria.

Finalmente, es beneficioso que **PCA se pueda aplicar a los datos etiquetados y no etiquetados**, ya que no se basa en las etiquetas de salida. Por otro lado, LDA requiere clases de salida para encontrar discriminantes lineales y, por lo tanto, requiere datos etiquetados.

2 Introducción al Aprendizaje Automático

El aprendizaje automático aborda los desafíos y las oportunidades presentados por el análisis de datos masivos para modelar datos existentes y predecir los resultados futuros.

Video de Google:

https://youtu.be/_rdINNHLYaQ

Definición: un conjunto de métodos que pueden detectar automáticamente patrones en los datos, y luego utilizar los patrones detectados para predecir los datos futuros, o realizar otros tipos de toma de decisiones bajo incertidumbre".

Por **ejemplo**, un programa informático es diseñado por un servicio de video para recomendar películas que podrían gustarles a los usuarios individuales. El algoritmo analiza las películas que los espectadores han visto ya y las películas que las personas con preferencias similares de visualización calificaron con buena puntuación. El objetivo es mejorar la satisfacción del cliente con el servicio de video

Aplicaciones: Los métodos de aprendizaje automático se han aplicado a una amplia variedad de aplicaciones que incluyen reconocimiento del habla, diagnósticos médicos, automóviles que conducen solos, motores de recomendación de ventas y muchos otros.

2 Tipos de análisis de aprendizaje automático

Estos algoritmos pueden dividirse en dos categorías principales: supervisados y no supervisados.

Los algoritmos de aprendizaje automáticos supervisados son los algoritmos de aprendizaje automático más utilizados para el análisis predictivo. Estos algoritmos dependen de conjuntos de datos que fueron procesados por los expertos humanos (por lo tanto, se usa la palabra "supervisión"). Los algoritmos luego aprenden cómo realizar las mismas tareas de procesamiento de forma independiente en los nuevos conjuntos de datos.

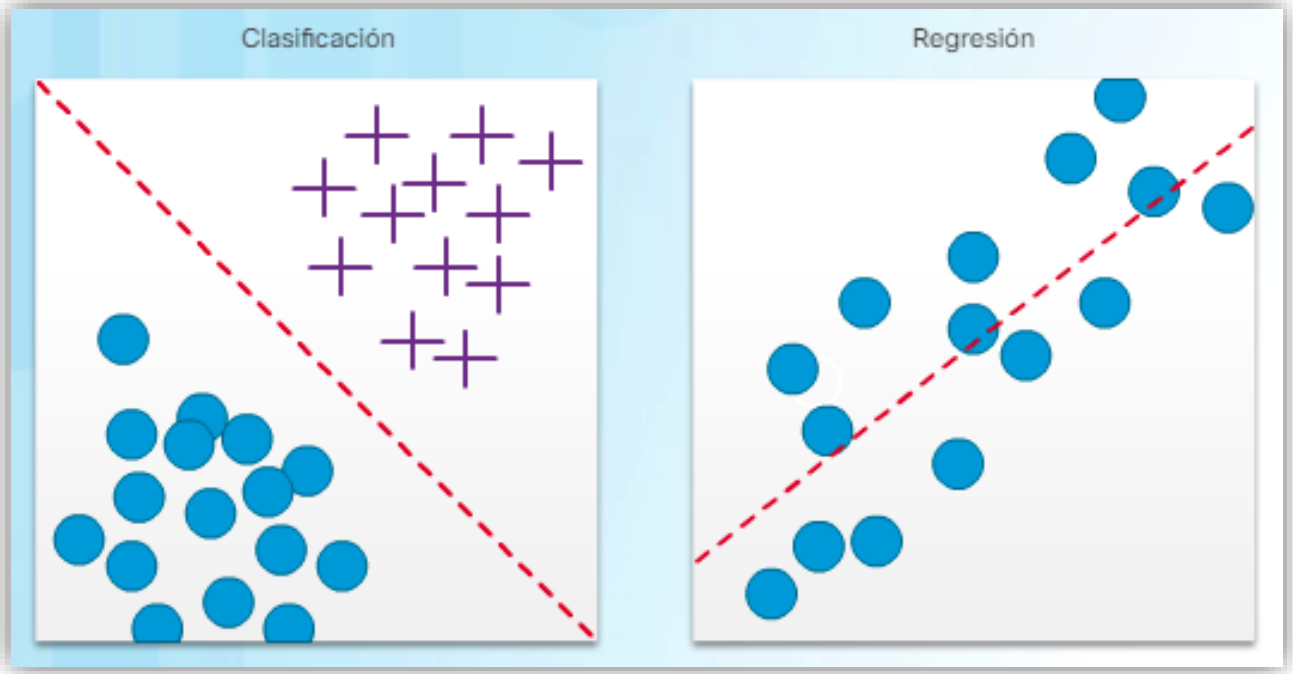
En particular, los **métodos supervisados** se utilizan para resolver **problemas de regresión y clasificación**.

Los algoritmos de **aprendizaje automático no supervisados** no requieren expertos humanos de los que aprender, sino que descubren patrones en los datos de forma autónoma. Algunos ejemplos de problemas resueltos con métodos no supervisados son el **agrupamiento y la asociación.**

2 Tipos de análisis de aprendizaje automático: Aprendizaje Supervisado

Problemas de regresión: son el cálculo de las relaciones matemáticas entre una variable continua y una o más variables. Esta relación matemática luego puede utilizarse para calcular los valores de una variable desconocida dados los valores conocidos de las demás. Los ejemplos de regresión son el cálculo de la posición del automóvil y su velocidad con el GPS, la predicción de la trayectoria de un tornado con datos meteorológicos, o la predicción del valor futuro de una acción mediante datos históricos y otras fuentes de información. Para mostrar mentalmente el ejemplo más simple de regresión, imagine dos variables, cuyos valores se muestran como los puntos en un diagrama bidimensional similar a la imagen de la derecha de la Figura 1. La ejecución de la regresión significa encontrar la línea que interpola mejor los valores. La línea puede tomar varias formas y se expresa como función de regresión. Una función de regresión le permite estimar el valor de una variable dado el valor de la otra, para los valores que no se han obtenido antes.

Problemas de clasificación: se utilizan cuando la variable desconocida es discreta. Por lo general, el problema comprende el cálculo al cual, de un conjunto de clases predefinidas, pertenece un ejemplo específico. Los ejemplos típicos de clasificación son reconocimiento de la imagen, o diagnóstico de las patologías de exámenes médicos, o identificación de rostros en una imagen. Una interpretación visual de un problema de clasificación se puede considerar en dos dimensiones, donde los puntos que pertenecen a diferentes clases se marcan con otro símbolo, similar a la imagen a la izquierda de la Figura. 1. El algoritmo “aprende” ejemplos de la ubicación y la forma de la línea fronteriza entre las clases. Esta línea fronteriza luego puede utilizarse para clasificar nuevos ejemplos.



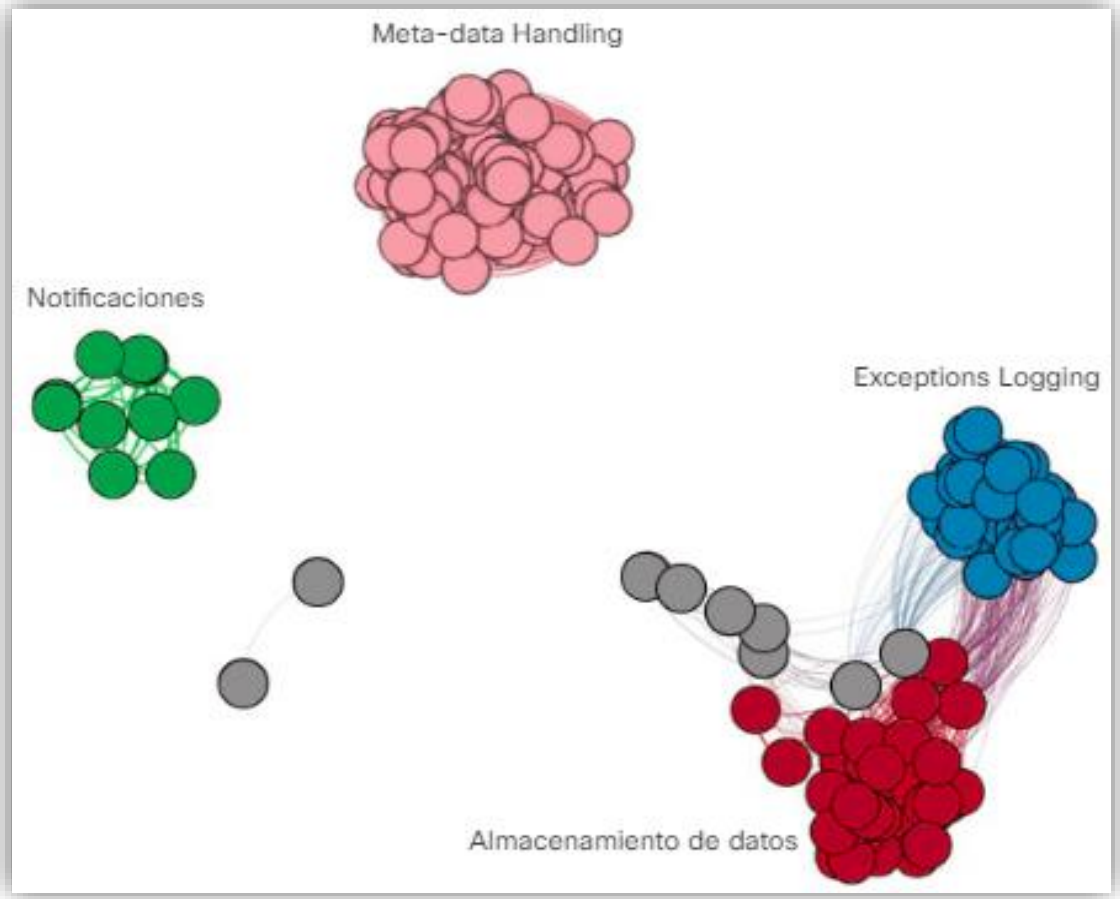
2 Tipos de análisis de aprendizaje automático: Aprendizaje No Supervisado

Métodos de agrupamiento: estos se pueden ver como la detección automática de grupos de ejemplos que tienen características similares, que pueden indicar posiblemente el hecho de que un miembro del grupo pertenece a una clase bien definida. Por ejemplo, los algoritmos de agrupamiento se utilizan para identificar grupos de usuarios basados en su historial de compras en línea, y luego envían avisos dirigidos a cada miembro.

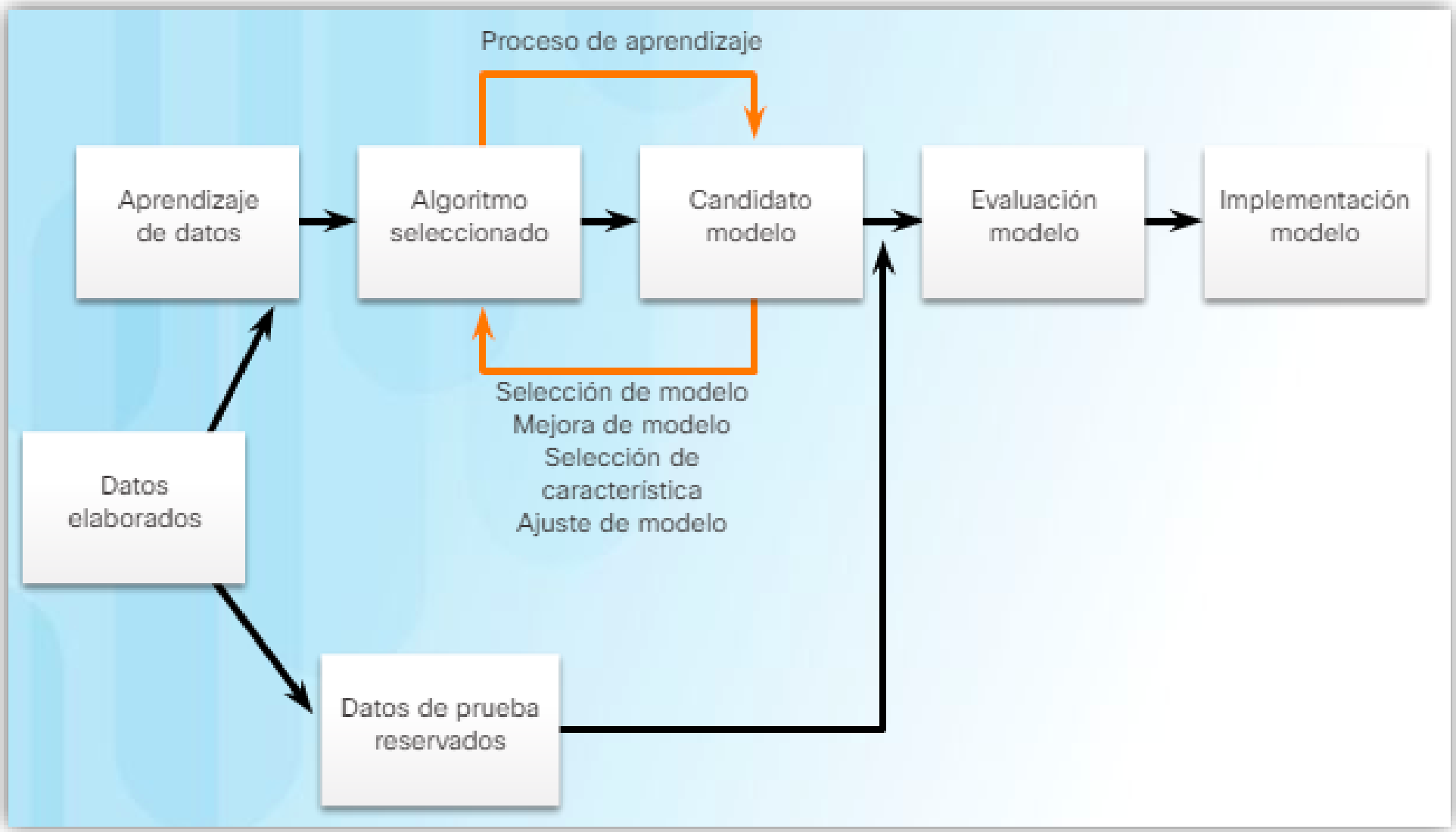
Métodos de asociación: estos son un problema muy relevante para los comerciantes en línea, y consisten en detectar grupos de elementos que se observan con frecuencia en conjunto. Se usan para sugerir compras adicionales a un usuario, según el contenido de su carrito de compras.

Resumen detallado de diez de los algoritmos de aprendizaje automático más utilizados:
https://www.kdnuggets.com/2016/08/10-algorithms-machine-learning-engineers.html/2

KDnuggets es un excelente recurso para los científicos de datos y los aprendices de científicos de datos.



2 Aprendizaje Automático: Proceso



2 Aprendizaje Automático: Aplicaciones

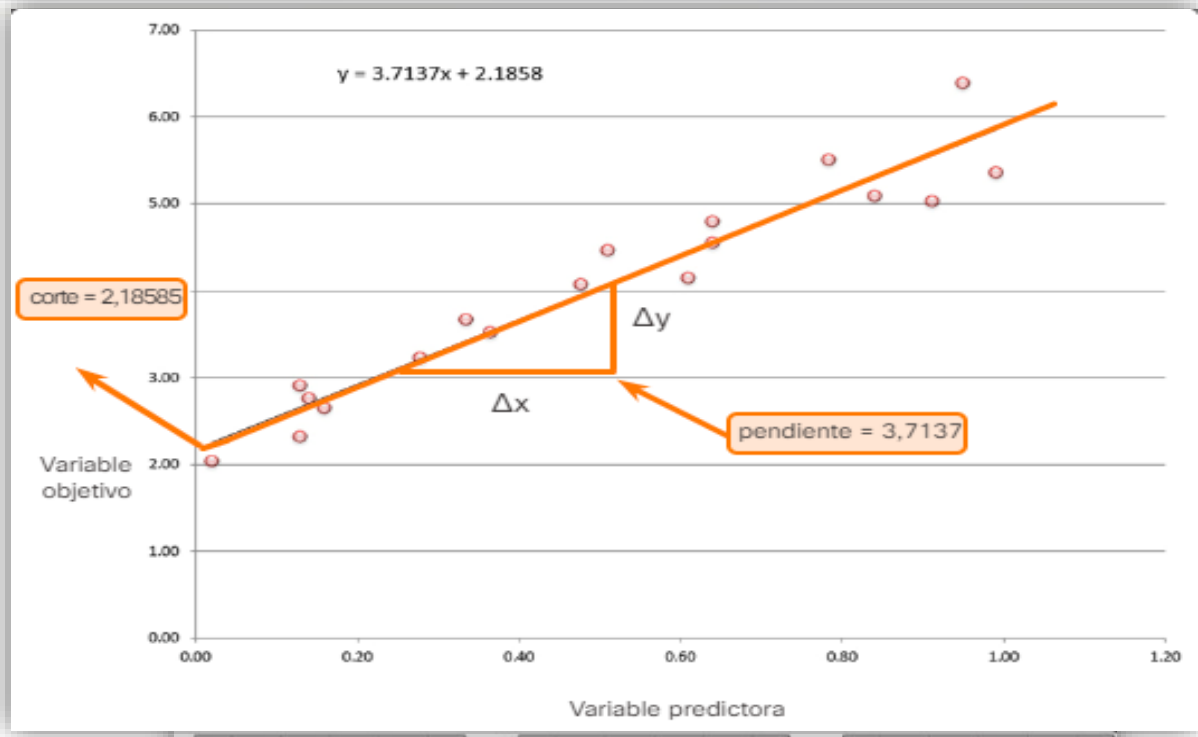
2 Aprendizaje Automático: Regresión Lineal

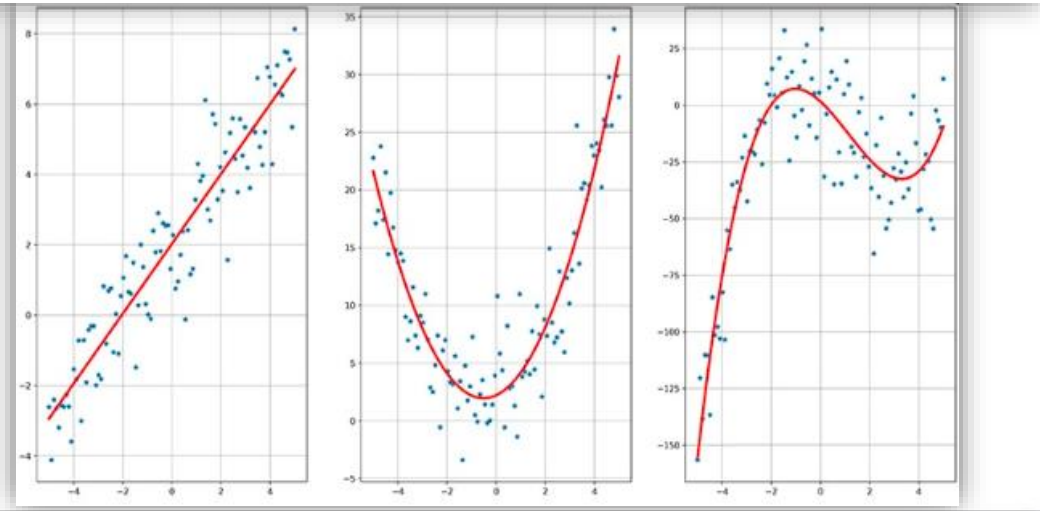
Los métodos más comunes de regresión se denominan regresiones lineales. Estos son los más simples desde el punto de vista del cálculo y las matemáticas; y por lo tanto, representan la primera opción para un analista de datos que presenta un problema de regresión. A pesar del nombre, la regresión lineal no implica pasar una línea a través de puntos de datos**. El término lineal significa que la función de regresión intentará siempre adaptarse a los datos mediante un promedio ponderado de otras funciones, ya sea que aquellas funciones sean lineales o no**. La propiedad de linealidad simplifica el cálculo de los parámetros del modelo de regresión y, al mismo tiempo, permite que prácticamente se use cualquier forma para responder a las observaciones. El **caso más simple de regresión lineal consta de ajustar una línea recta. Esto también se conoce como modelo lineal simple.**

Un alto porcentaje de correlación de Pearson indica que un modelo lineal simple es un buen candidato para adaptarse a los datos. **El proceso de regresión, en este caso, consiste en encontrar de la pendiente y la intercepción de la línea que minimiza la suma de las distancias entre la línea y todos los puntos de datos**. Al utilizar modelos lineales, el algoritmo más común para calcular estos parámetros modelo óptimos se denomina **requisitos mínimos cuadráticos**.

En la Figura se pueden ver tres conjuntos de datos, cada uno con una variable objetivo y una variable predictor. En los tres casos, se puede observar cómo, a pesar del ruido que afecta las observaciones, hay una línea clara que captura la relación subyacente entre las variables. La línea roja representa el modelo de regresión lineal que minimiza la distancia de todas las observaciones. Los modelos se obtuvieron mediante regresión lineal.

***Práctica 4.1.2.4 (regresión lineal) y 4.2.2.5 (Evaluación de errores: MSE, MAE)**





2 Aprendizaje Automático: Clasificación

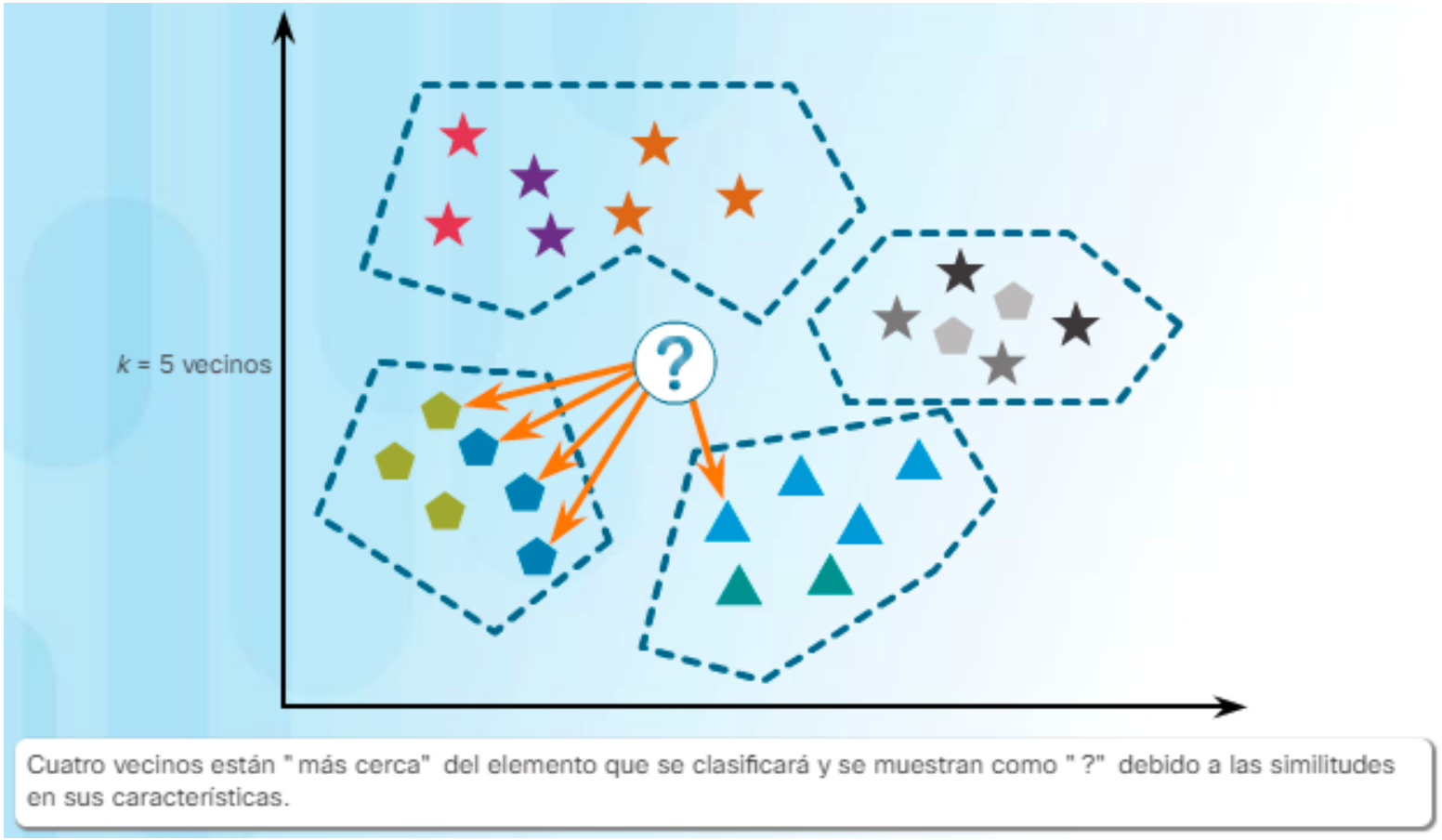
La clasificación es otro problema de aprendizaje automático común que se adecua a la categoría de aprendizaje supervisado. Mejoras constantes se han obtenido en la última década, especialmente en el dominio del reconocimiento de imagen. **La clasificación se puede considerar un problema de regresión cuando la variable objetivo es discreta y representa una clase en la cual un experto humano ha clasificado la muestra de datos.**

Es común, en los problemas de clasificación, proporcionar no solo un conjunto de puntos de datos de ejemplo de cada clase, sino también **establecer cuáles son las características de cada punto de datos más útiles para estimar la clase correspondiente**. La definición de funciones relevantes es un paso importante que, a excepción de los algoritmos muy avanzados como la profundidad de aprendizaje, se basa en el conocimiento del experto humano

Por ejemplo, una empresa de viajes en Internet está interesada en ofrecer una calificación de fiabilidad para los vuelos que encuentra para los clientes. Mediante el error de prueba de los diferentes modelos, se ha determinado **qué variables entre todas las del conjunto de datos son más relevantes para las clasificaciones**. Esto también se conoce como las **variables con el poder discriminante más alto**. Solo estas funciones relevantes se extraen de los datos y se utilizan para entrenar el clasificador.

2 Aprendizaje Automático: Algoritmos de Clasificación

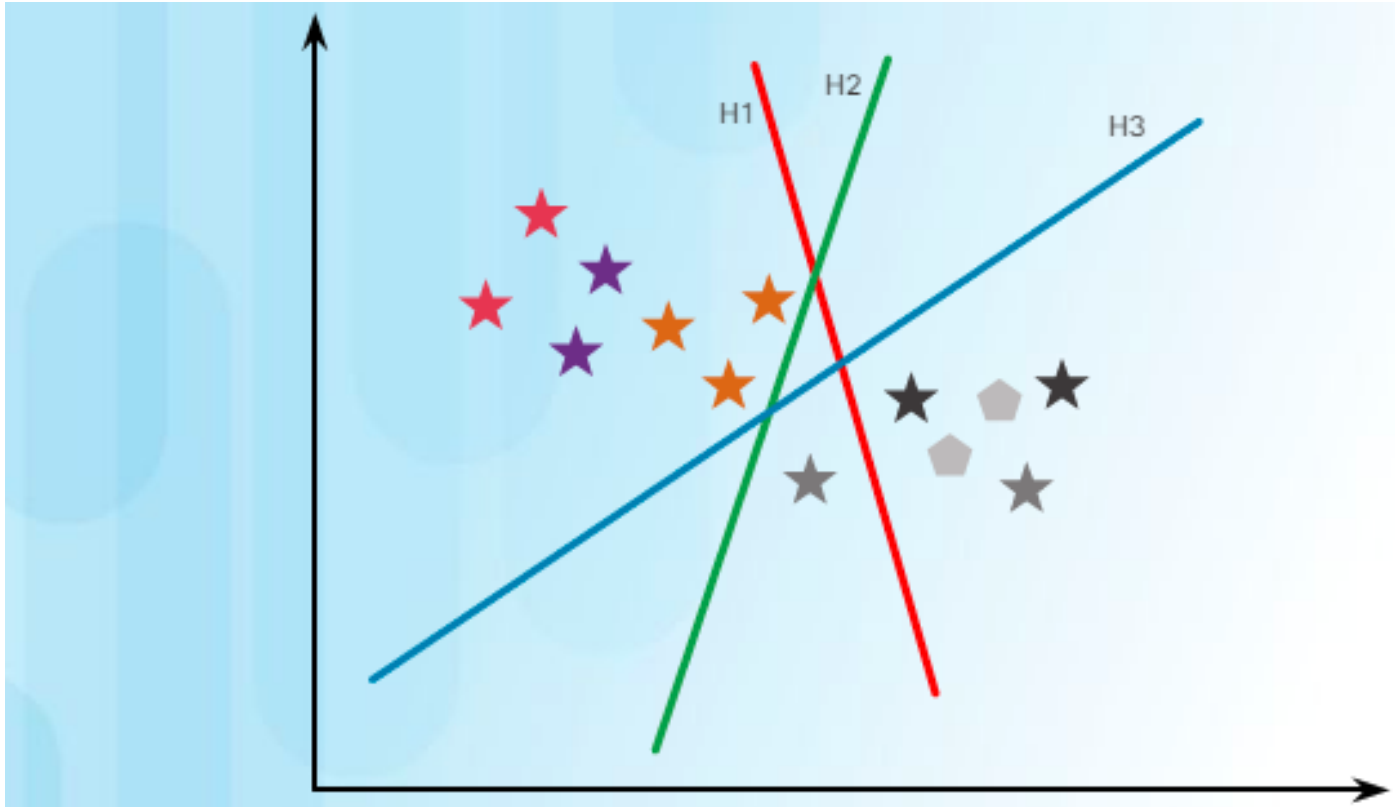
k-nearest neighbor (k-NN): k-NN es posiblemente el clasificador más simple, que utiliza la distancia entre los ejemplos de entrenamiento como medida de similitud. La distancia entre los puntos representa la diferencia entre los valores de sus funciones. Dado un nuevo punto de datos, un clasificador k-NN debe ver los puntos de entrenamiento más cercanos. La clase predicha para el nuevo punto será la clase más común entre los k neighbors.



Cuatro vecinos están "más cerca" del elemento que se clasificará y se muestran como "?" debido a las similitudes en sus características.

2 Aprendizaje Automático: Algoritmos de Clasificación

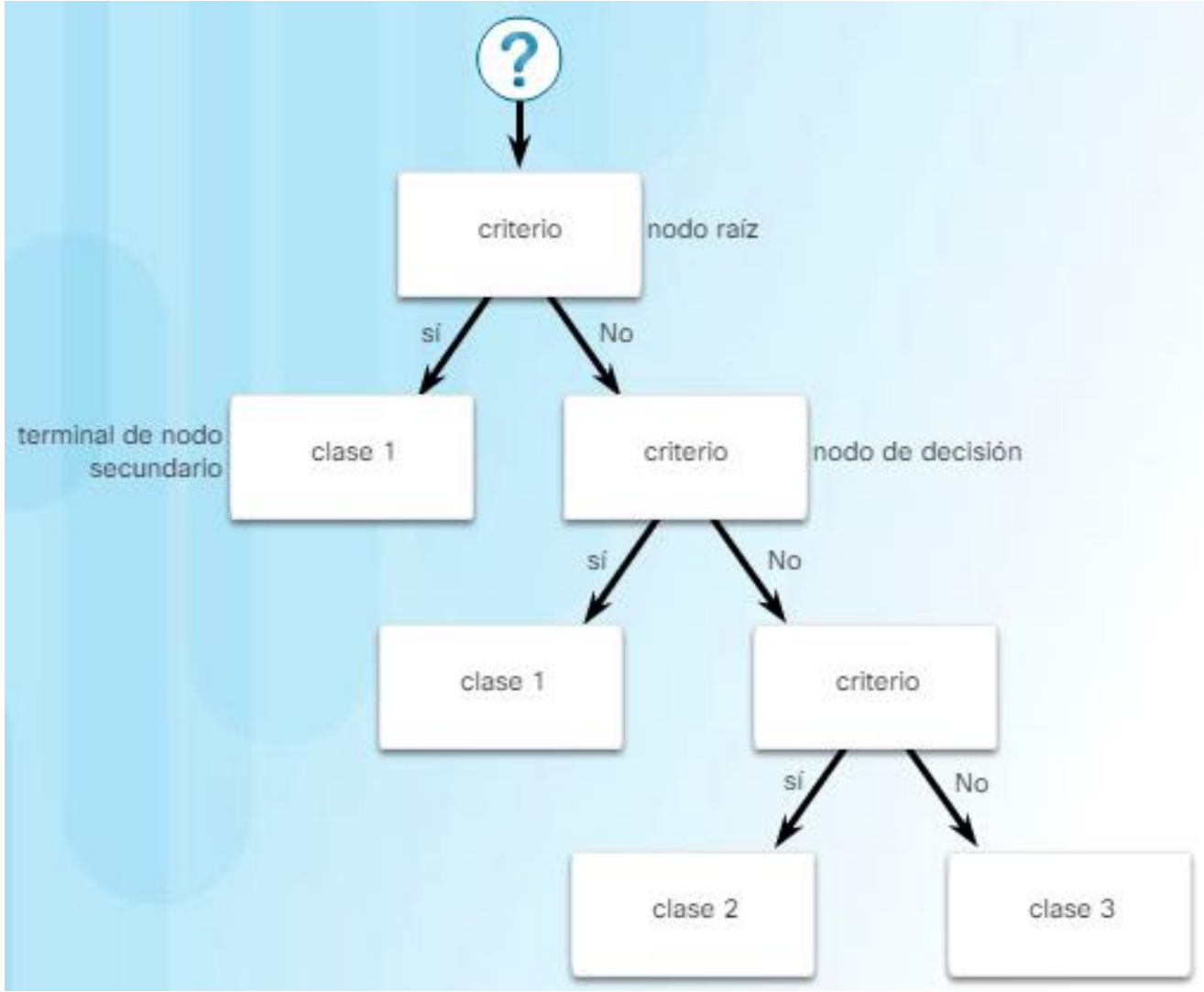
Máquinas de vector de soporte (SVM): las máquinas de vector de soporte (SVM), que se muestran en la Figura 3, son ejemplos de clasificadores de aprendizaje automático supervisados. En lugar de basar la asignación de membresía de la categoría en distancias de otros puntos, las máquinas de vector de soporte computan la frontera, o el hiperplano, que mejor separa los grupos. En la figura, el H3 es el hiperplano que maximiza la distancia entre puntos de entrenamiento de las dos clases, visibles en color o en blanco y negro. Cuando se presenta un nuevo punto de datos, se clasifica según si se encuentra en un lado o en el otro de H3.



Tres líneas divisorias posibles en el espacio de datos para definir clases.
- H1 no separa los grupos.
- H2 separa los grupos pero está más cerca de uno que del otro.
- H3 separa los grupos con el margen más ancho posible entre ellos.

2 Aprendizaje Automático: Algoritmos de Clasificación

Árboles de decisión: los árboles de decisión representan un problema de clasificación como un conjunto de decisiones basadas en los valores de las funciones. Cada nodo del árbol representa un umbral sobre el valor de una función, y parte los ejemplos de entrenamiento en dos grupos más pequeños. El proceso de decisión se repite sobre todas las características, con lo que el árbol crece hasta que una manera óptima de dividir los ejemplos se computa. La clasificación de un nuevo ejemplo luego puede obtenerse siguiendo las ramas del árbol según los valores de sus funciones. Una vista simplificada de un árbol de decisión binario y de los tipos de nodos se muestra en la figura.



2 Aprendizaje Automático: Aplicaciones de Clasificación

Los algoritmos de clasificación tienen muchas aplicaciones. He aquí varios ejemplos:

Evaluación de riesgos: los sistemas de clasificación se pueden utilizar para determinar cuáles de muchos factores contribuyen a la probabilidad de diversos riesgos. Por ejemplo, varios factores pueden utilizarse para clasificar a los usuarios de seguros de vehículo en categorías de bajo, medio y alto riesgo y para ajustar las primas que los conductores pagan según el nivel de riesgo.

Diagnósticos médicos: los sistemas de clasificación pueden utilizar preguntas orientadas para construir un árbol de decisión que pueda ayudar a diagnosticar varias enfermedades y riesgos de enfermedades. Los sistemas de clasificación de aprendizaje automático también pueden realizar el análisis preliminar de una gran cantidad de imágenes de diagnóstico, y señalar las condiciones sospechosas para la revisión de los médicos.

Reconocimiento de imagen: por ejemplo, en el reconocimiento de escritura a mano, un sistema puede trabajar desde la tarea de identificar números manuscritos. Los números 0 a 9 se pueden considerar como clases. El clasificador recibe un ejemplo grande de números manuscritos, que se ha etiquetado en cada instancia con el número real representado. El clasificador busca las funciones que probablemente estén presentes y sean únicas para cada uno de los números.

Práctica 4.1.3.5: Clasificador del árbol de decisión

2 Aprendizaje Automático: Método Científico

El método científico es un proceso de seis pasos:

Paso 1: Hacer preguntas sobre una observación como qué, cuándo, cómo o por qué.

Paso 2: Hacer una investigación.

Paso 3: Formar las hipótesis de esta investigación.

Paso 4: Probar las hipótesis con la experimentación.

Paso 5: Analizar los datos de los experimentos para sacar una conclusión.

Paso 6: Comunicar los resultados del proceso.

2 Aprendizaje Automático: Validez y Fiabilidad

Tipos de validez:

Validez de construcción: ¿el estudio mide realmente lo que afirma medir?
Validez interna: ¿el experimento se diseñó correctamente? ¿Incluye todos los pasos del método científico?
Validez externa: ¿las conclusiones se pueden aplicar a otras situaciones u otras personas en otros lugares en otro momento? ¿Hay otras relaciones causales en el estudio que puedan explicar los resultados?
Validez de conclusión: según las relaciones en los datos, ¿las conclusiones del estudio son razonables?

Un **experimento o un estudio fiable** significa que otra persona puede repetirlo y acceder a los mismos resultados.

Tipos de fiabilidad:

Fiabilidad de calificación interna: ¿con cuánta similitud diferentes personas obtienen resultados en la misma prueba?
Fiabilidad de prueba y nueva prueba: ¿cuánta variación hay entre los resultados de una persona que realiza una prueba múltiples veces?
Fiabilidad de formas paralelas: ¿cuánta similitud hay en los resultados de dos pruebas diferentes construidas a partir del mismo contenido?
Fiabilidad de consistencia interna: ¿cuál es la variación de los resultados para elementos diferentes en la misma prueba?

¿Entonces cómo pueden estar razonablemente seguros de que el sistema de clasificación funcionará con datos que no se hayan procesado antes? Se recurre a un método denominado validación cruzada. La **validación cruzada** es donde se entrena el algoritmo utilizando solo un ejemplo de datos seleccionado aleatoriamente, denominado conjunto de entrenamiento. Luego, el modelo se analiza en el resto de los datos, denominado conjunto de validación. El rendimiento de la clasificación que un sistema de clasificación muestra en el conjunto de entrenamiento generalmente es mayor que el del conjunto de validación. Sin embargo, esto representa de mejor manera cómo el algoritmo se comporta con ejemplos que no haya procesado antes.

2 Aprendizaje Automático: scikit-learn

Scikit-learn es una biblioteca popular de aprendizaje automático. Esta biblioteca contiene muchas herramientas útiles para el análisis de datos y se basa en NumPy, SciPy y matplotlib.

Práctica 4.3.1.4 Regresión lineal del contador de Internet

**Parte 1: Importar las bibliotecas **
**Parte 2: Visualizar los datos históricos **
**Parte 3: Crear un modelo de regresión lineal simple - Polinomio de primer orden **
**Parte 4: Crear un modelo de regresión lineal simple - Polinomios de orden superior **
Parte 5: Calcular los errores
**Parte 6: Crear un modelo de regresión no lineal - Crecimiento exponencial **
Parte 7: Comparar los modelos

2 Visualización de datos

Ejemplos de cómo presentar datos completos en formas simples:
https://www.import.io/post/8-fantastic-examples-of-data-storytelling/

https://www.thinkwithgoogle.com/marketing-resources/data-measurement/tell-meaningful-stories-with-data/

https://www.thinkwithgoogle.com/marketing-resources/data-measurement/data-to-insights-blueprint-for-your-business/

Hojas de Estilo (el mismo código en línea para varios diagramas):
https://matplotlib.org/users/customizing.html

Plotly es una herramienta en línea poderosa que se puede utilizar para generar rápidamente hermosas visualizaciones de datos.
https://plot.ly/python/

Gráfico de búrbuja (para más de dos variables)
https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen

Práctica 5.2.3.8 Visualización de datos en Excel

Folium permite tomar los marcos de datos de Python y mostrarlos en un mapa interactivo de Leaflet.

Folium Tilesets: Un tileset es un conjunto de tramas o datos de vector que puede mostrar un mapa en los dispositivos móviles o en un navegador.

Práctica 5.3.1.4 Visualización Avanzada de los datos

2 Ingeniería de Datos

La ingeniería de datos generalmente implica un sistema de información empresarial basado en computadoras, que captura o genera, procesa, almacena, distribuye y analiza información (datos).

Bases de datos no relacionales

A partir de 1990 un nuevo tipo de bases de datos no relacionales fue desarrollado a partir de la observación de que el paradigma emergente de programación orientado a objetos también podría ser una base de datos modelo. Se creó el sistema de administración de la base de datos orientado a objetos (OODBMS). OODBMS no se adoptó ampliamente y no pudo reemplazar las bases de datos relacionales.

A principios del 2000, el surgimiento de la red 2.0, el comercio electrónico y las empresas como Google clarificaron que las bases de datos relacionales no pueden resolver el volumen y la velocidad de las peticiones de búsqueda en la red. Para responder a esta demanda, **Google desarrolló el sistema de archivos distribuidos de Google (GFS), el algoritmo distribuido MapReduce de procesamiento paralelo y las bases de datos distribuidas BigTable NoSQL**. En 2004, Jeffrey Dean y Sanjay Ghemawat de Google publicaron un paper de seminario llamado “MapReduce: procesamiento de datos simplificado en los clústeres grandes” que cambió para siempre la manera de guardar y de procesar un conjunto de datos grande. Ese informe fue la inspiración para dos programadores, **Doug Cutting y Mike Cafarella, quienes crearon Apache Hadoop**. Después de esto, el enfoque de MapReduce formó la base para el desarrollo del ecosistema de Hadoop de Yahoo y la base de datos de HBase, así como la base de datos NoSQL de los pares de valor clave de dínamo de Amazon. Hadoop, a su vez, ayudó a promover el desarrollo de otras tecnologías y aplicaciones de datos masivos.

NoSQL son las familias grandes de bases de datos que no confían en el enfoque de la base de datos relacional de las tablas enlazadas. Las bases de datos NoSQL pueden utilizar un enfoque de almacenamiento de valor clave en lugar de un enfoque basado en tablas relacional. Otras bases de datos NoSQL almacenan datos como documentos estructurados en formatos XML o JSON. Las bases de datos NoSQL son mucho más rápidas que las bases de datos relacionales, y pueden importar datos no estructurados. Las bases de datos NoSQL están diseñadas para escalarse de forma horizontal, lo que significa que la capacidad de almacenamiento y de administración puede aumentarse con solo agregar otras máquinas al grupo. Los sistemas más populares NoSQL incluyen MongoDB, Couchbase, Riak, Memcached, Redis, CouchDB, Hazelcast, Apache Cassandra, HBase y Dynamo, que son todos productos de software de código abierto.

2 Ingeniería de Datos

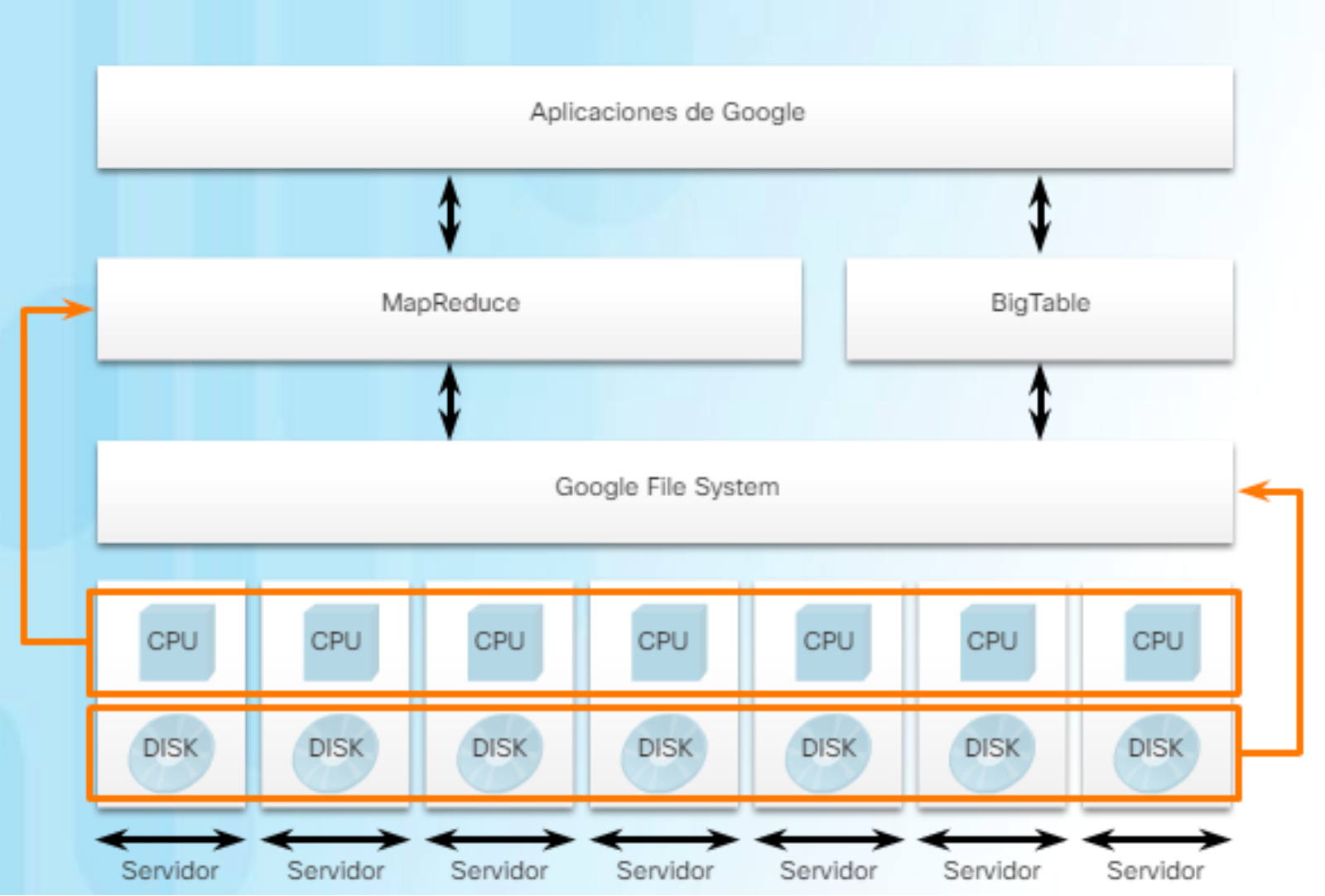
Analista de negocios: Un analista de negocio es la persona que pueda estudiar una empresa o industria y luego formular una pregunta específica. Los analistas de negocio representan los expertos de datos que trabajan con las partes interesadas de la empresa para determinar los motivos de inquietud.

Analista de datos: Los analistas de datos consultan y procesan datos, proporcionan informes, resumen y muestran datos.

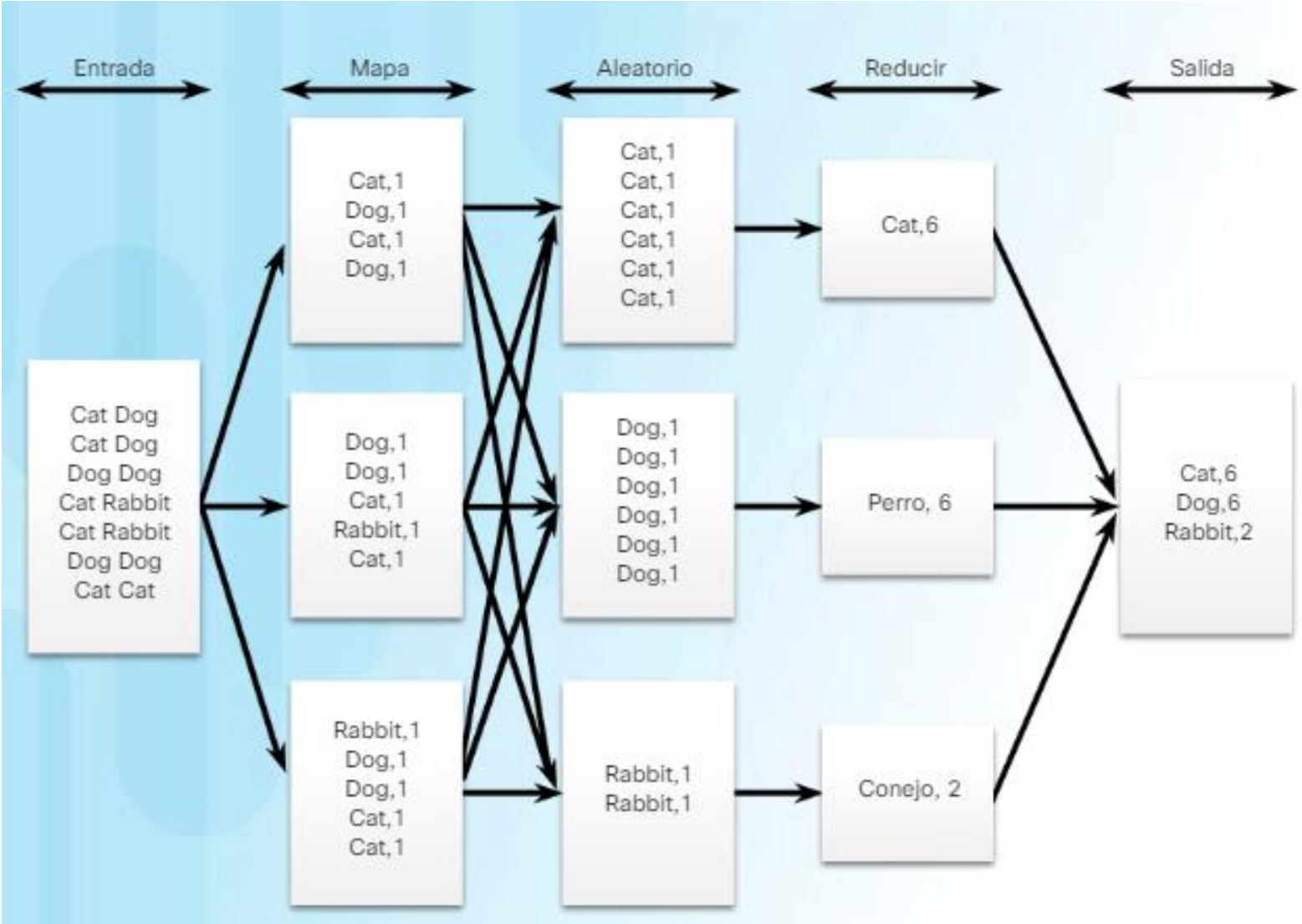
Científico de datos: Un científico de datos toma datos sin procesar y los transforma en información importante. Los científicos de datos aplican las estadísticas, el aprendizaje automático y los enfoques analíticos para responder las preguntas esenciales de la empresa.

Ingeniero de datos: El ingeniero de datos crea una infraestructura que admite datos masivos. Diseña y construye la plataforma en la que todos estos datos se almacenan y se procesan. Los ingenieros de datos también administran todos estos datos. Aseguran la accesibilidad y la disponibilidad por parte de los científicos de datos y los analistas de datos.

2 Ingeniería de Datos

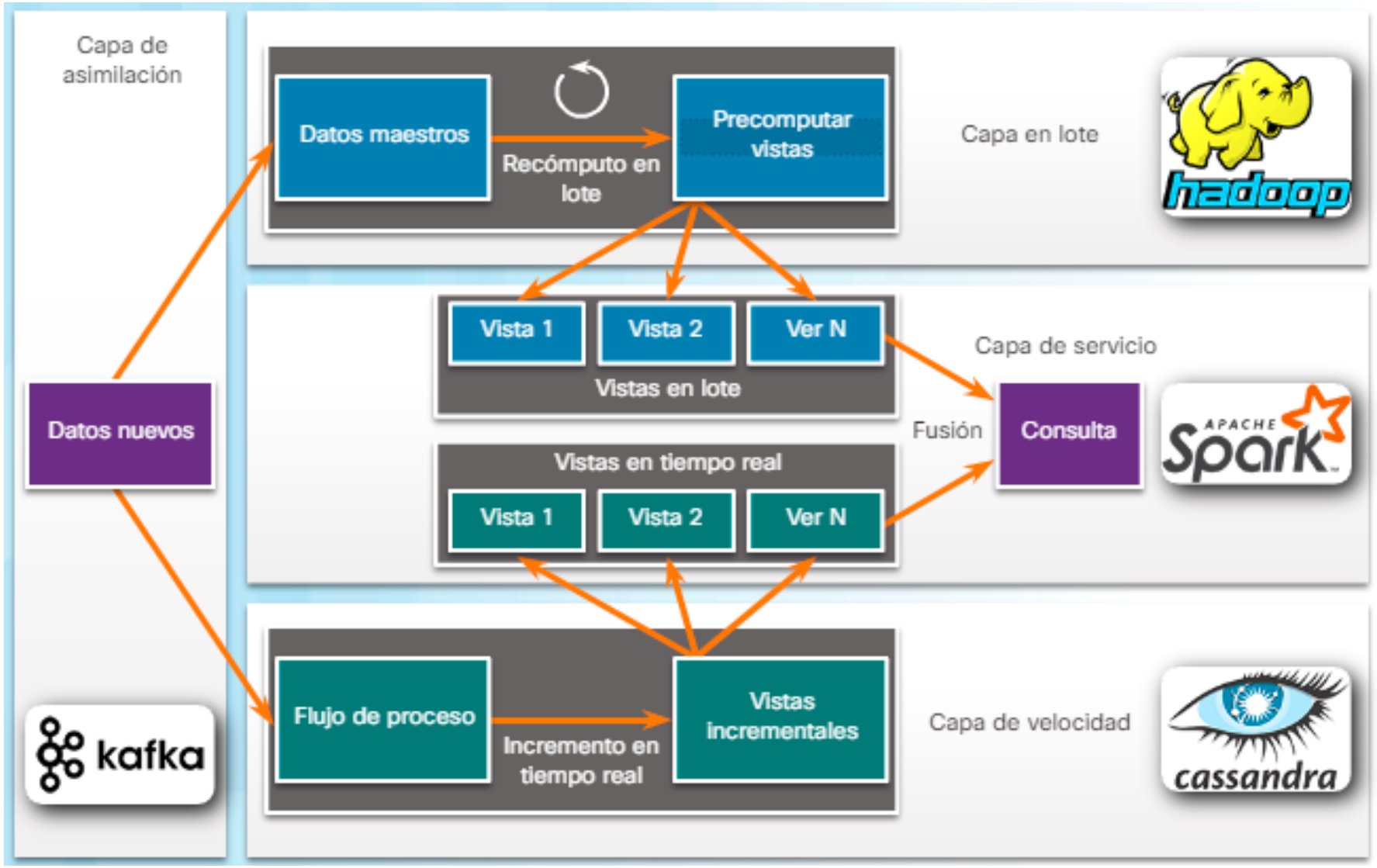


2 Ingeniería de Datos: Cómo funciona Hadoop: MapReduce



2 Ingeniería de Datos: Arquitecturas

2 Ingeniería de Datos: Arquitecturas



2 Ingeniería de Datos: Imágenes Digitales como datos

Práctica de laboratorio 6.4.1.4: Detección de sonrisa

En esta práctica de laboratorio, se aplicará un modelo de aprendizaje automático para identificar caras sonrientes en las imágenes digitales que han sido capturadas con Raspberry Pi. El modelo de aprendizaje automático se ha capacitado en las imágenes digitales y está compuesto por un conjunto de datos que se han rotulado como caras sonrientes o no sonrientes. Este modelo de aprendizaje automático luego se aplica en tiempo real a imágenes que se han capturado mediante Raspberry Pi para identificar caras sonrientes.

UNIDAD 3. PROCESOS ESTOCÁSTICOS

2.1 INTRODUCCIÓN Y PROCESOS DE RENOVACIÓN

Antes de la era de los datos masivos, el rol del tiempo en el análisis de datos estaba limitado a cuánto tiempo se tardaba en compilar un conjunto de datos de fuentes dispares, o a cuánto tiempo tomaba ejecutar un conjunto de datos mediante cálculos. Con los datos masivos, el tiempo se vuelve importante de otras maneras, porque gran parte del valor de los datos deriva de crear oportunidades de acción inmediata.

Los sensores, los consumidores, los usuarios de redes sociales, los motores, el mercado de valores y lo que sea que esté conectado a una red genera datos a un ritmo sin precedentes. Estos datos no solo están creciendo en cantidad; también están cambiando en tiempo real. El análisis de datos también se debe realizar en tiempo real mientras se recopilan los datos.

UNIDAD 4. CADENAS DE MARKOV

REFERENCIAS

[1] Evans, M. Rosenthal, J. (2013) Probabilidad y estadística la ciencia de la incertidumbre. Reverté.

[2] Forsyth, D. (2018). Probability and Statistics for Computer Science. Springer.

[3] Enlace calcular derivadas: https://www.calculadora-de-derivadas.com/