Fundamentos de Probabilidad y Estadística

para Inteligencia Artificial

Estructura de la sesión

Estadística Descriptiva

- Medidas de Tendencia Central
- Medidas de Dispersión
- Correlación
- Ejemplos y aplicaciones para IA

Teoría de Probabilidad

- Conceptos Clave
- Probabilidad Conjunta, Marginal y Condicional
- Distribuciones de Probabilidad
- Inferencia Probabilística
- Ejemplos

¿Por qué...

necesitamos probabilidad y estadística en IA?

Incertidumbre

Modelar la Incertidumbre y la Toma de Decisiones

Estadística

Analizar Datos y Extraer Patrones

¿Por qué...

necesitamos probabilidad y estadística en IA?

Probabilidad

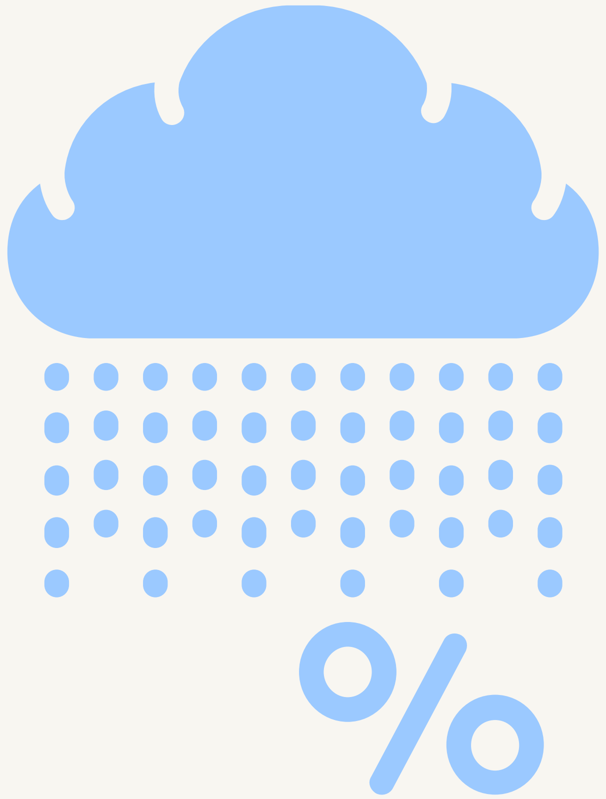
Algoritmos como clasificadores, Redes Neuronales Artificiales y Árboles de Decisión

Evaluar Modelos

¿Qué Tan Buenas Son Nuestras Predicciones?

¿Cómo la probabilidad y la estadística nos ayudan a tomar mejores decisiones con datos?

Imagina que sales de casa... No tienes certeza absoluta de si lloverá, pero puedes usar datos y probabilidades para tomar la mejor decisión.



Diferencia entre probabilidad (incertidumbre en eventos) y estadística (análisis de datos).

¿Debo llevar paraguas hoy?

Datos históricos y patrones climáticos

- ¿Cuántos días al año llueve en mi ciudad?
- ¿Llueve más en ciertas estaciones o meses?
- Si ayer llovió, ¿qué tan probable es que hoy también llueva?

Probabilidad condicional

Supongamos que:

- El pronóstico del tiempo dice que hay un 80% de probabilidad de lluvia.
- El cielo está nublado y hay alta humedad.

Decisión	¿Llovió?	Consecuencia
Llevaste paraguas	Si	No te mojas
Llevaste paraguas	No	Cargas un paraguas innecesariamente
No llevaste paraguas	Sí	Te mojas y arruinas tu ropa
No llevaste paraguas	No	No Cargas Peso Extra

Análisis de Datos

Estadística Descriptiva

es una rama de la estadística que se encarga de organizar, resumir y representar datos numéricos

Medidas de Tendencia Central

- Media, mediana y moda.

Medidas de Dispersión

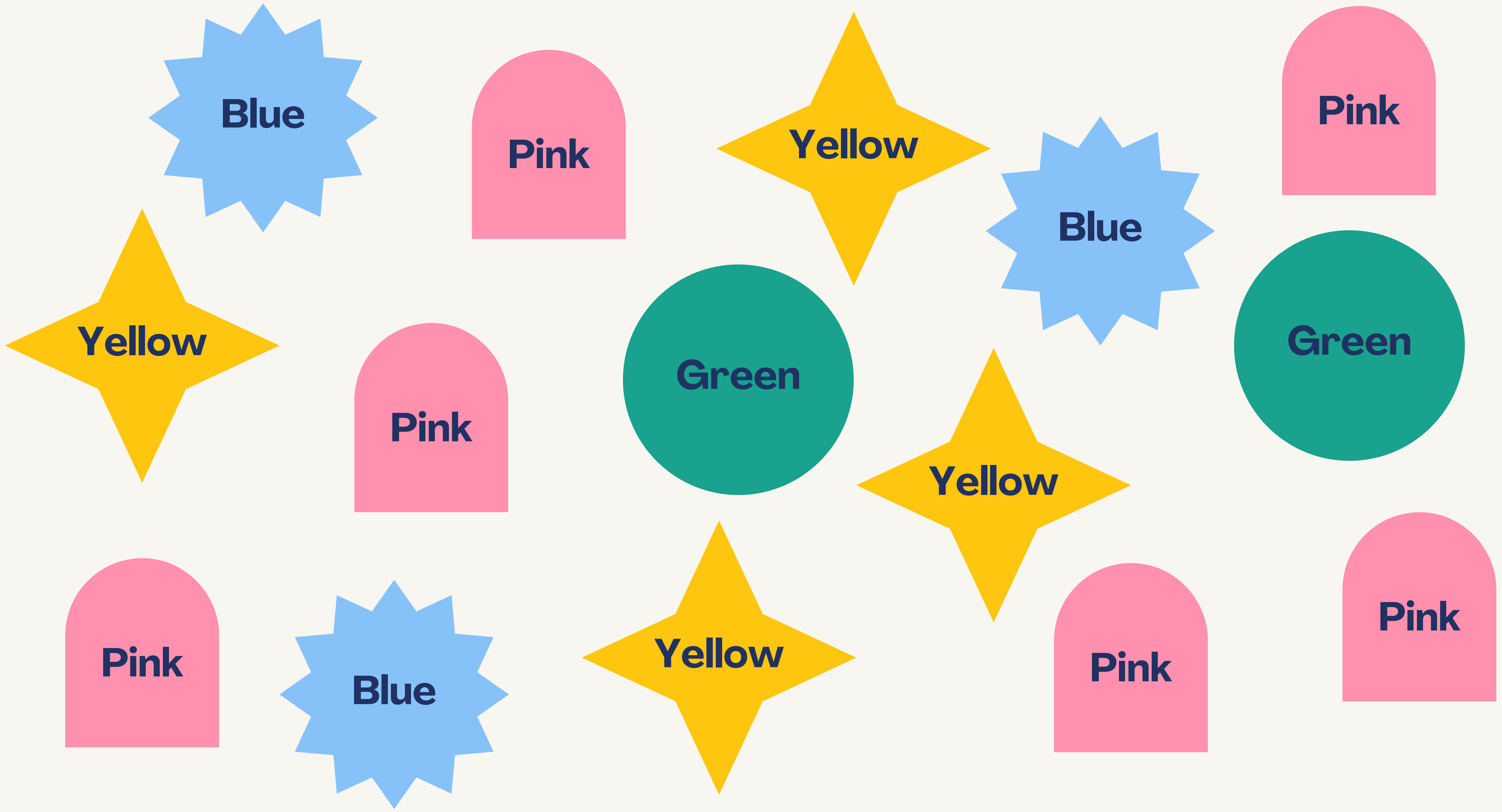
- Varianza y desviación estándar

Correlación y Covarianza

- Concepto de correlación: Relación entre variables.

Votación de Colores Favoritos

Supongamos que realizamos una encuesta a 100 personas sobre su color favorito y obtenemos los siguientes resultados:



¿Cómo analizar y organizar los resultados?

Preguntas a Resolver

¿Cuál es el color más popular?

¿Cuántos votos tiene un color en promedio?

¿Cuánta variabilidad hay entre los votos?

Estos mismos principios se aplican en Machine Learning para analizar datos antes de entrenar modelos

Análisis Exploratorio de Datos

Ahora aplicamos la estadística descriptiva para analizar estos datos:

Color	Número de personas
Azul	35
Rojo	20
Verde	15
Amarillo	10
Negro	10
Blanco	10

Which color is the mode?

Medidas de Tendencia Central:

- Media
- Mediana
- Moda

Medidas de Dispersión:

- Rango
- Varianza
- Desviación

Color	Número de personas
Azul	35
Rojo	20
Verde	15
Amarillo	10
Negro	10
Blanco	10

Ejemplo Varianza
Tiempos de entrega en dos sucursales de una empresa

Imagina que dos sucursales de una empresa entregan paquetes, y queremos analizar la rapidez en la entrega.

Sucursal 1:

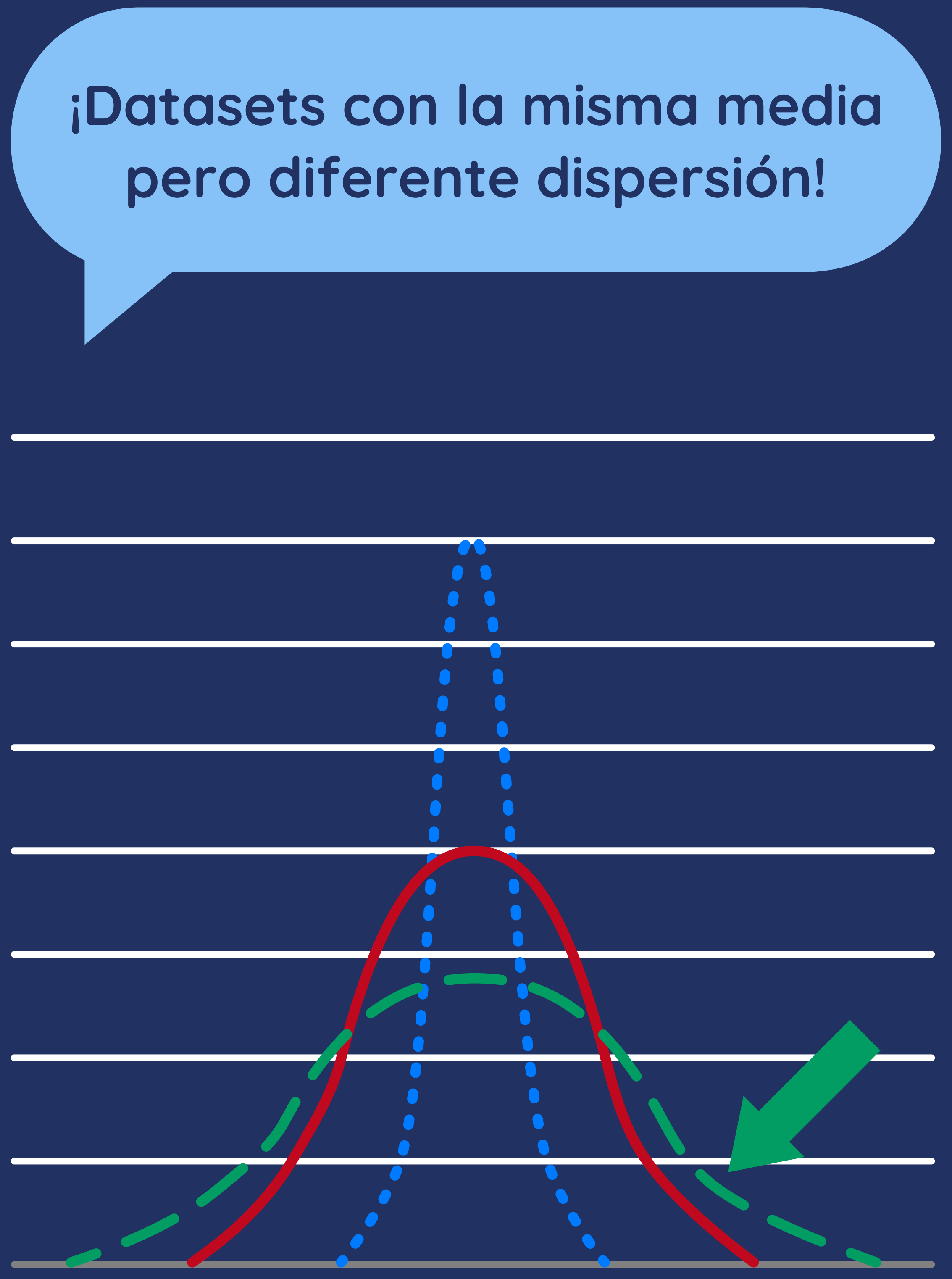
Tiempos en días: 5, 5, 5, 5, 5

- Media = 5 días
- Varianza = 0 (todas las entregas son en el mismo tiempo)

Sucursal 2:

Tiempos en días: 1, 3, 5, 7, 9

- Media = 5 días
- Varianza > 0 (algunas entregas son muy rápidas y otras tardan más)



Supongamos que, además de preguntar el color favorito, también preguntamos la edad de los encuestados y obtenemos la siguiente tabla:

Color Favorito	Número de Votos (X)	Edad Promedio (Y)
Azul	35	25
Rojo	20	30
Verde	15	22
Amarillo	10	35
Negro	10	28
Blanco	10	40

Relación entre variables

¿Existe una relación entre los votos y la edad?

La covarianza nos indica si existe una relación entre los votos y la edad.

Correlación

¿Qué tan fuerte es la relación?

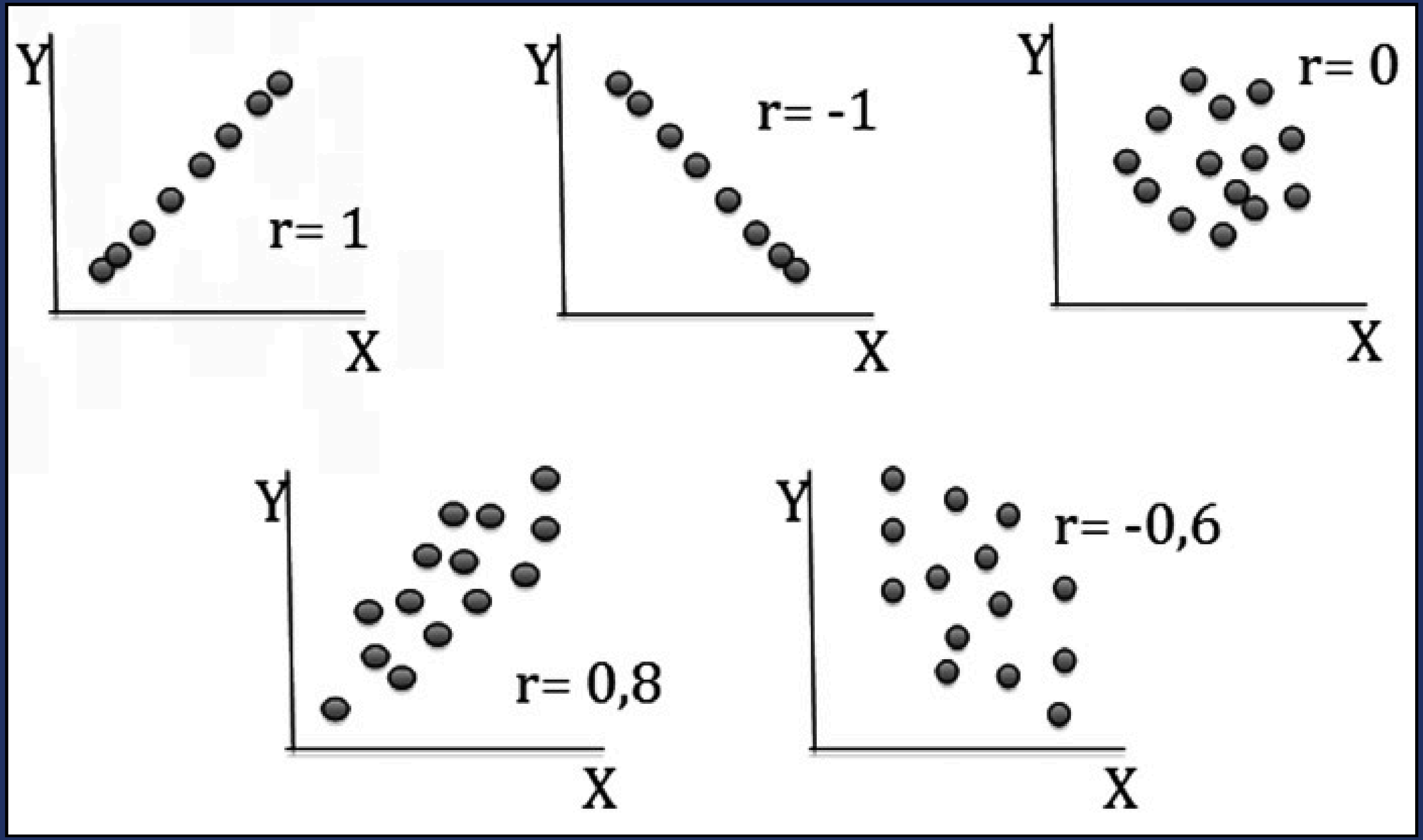
Estos cálculos ayudan a entender qué variables están relacionadas en un dataset, lo cual es útil para seleccionar características

Covarianza

Si Cov(X, Y) > 0, significa que a mayor número de votos, mayor edad promedio.
Si Cov(X, Y) < 0, significa que a mayor número de votos, menor edad promedio.
Si Cov(X, Y) ≈ 0, no hay relación entre los votos y la edad.

Correlación de Pearson

El resultado estará en el rango [-1, 1]:
✅ r > 0 → Correlación positiva (cuando sube X, sube Y)
❌ r < 0 → Correlación negativa (cuando sube X, baja Y)
⚠ r ≈ 0 → No hay relación



Con el ejemplo, ¿Cómo interpretar los valores?

- Si encontramos una correlación fuerte entre edad y popularidad de un color, podríamos predecir qué colores son más atractivos para diferentes grupos de edad.
- Si la correlación es débil, significa que los gustos por los colores son independientes de la edad.

Por lo tanto...

- La covarianza indica la dirección de la relación, pero no la intensidad.
- La correlación de Pearson nos da tanto la dirección como la fuerza de la relación de forma estandarizada.

¿Cómo analizar y organizar los resultados?

Presenting Results

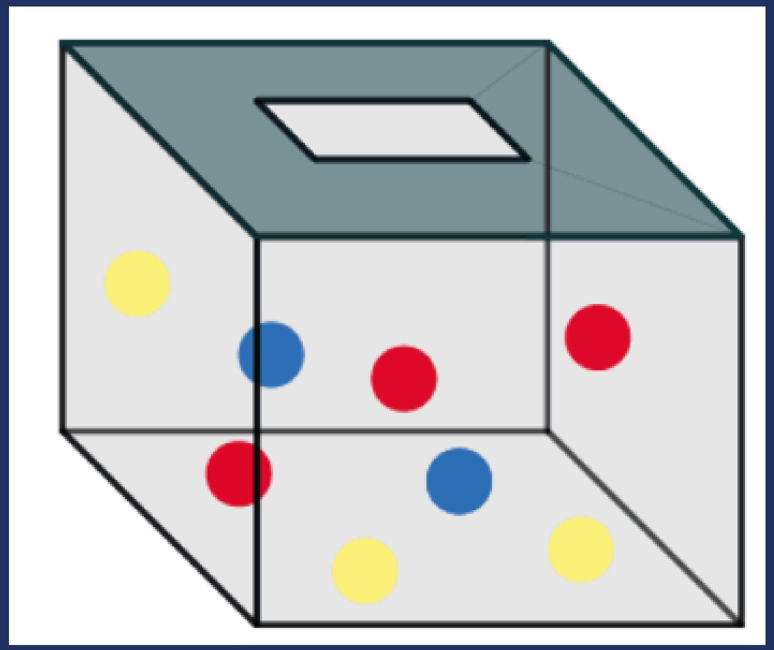
Fundamentos de Probabilidad

La probabilidad compuesta o conjunta es el cálculo de la probabilidad cuando un experimento de probabilidad simple se repite varias veces o se relaciona un experimento con otro.

Sucesos independientes:
Dos sucesos son independientes si la ocurrencia de uno no afecta la probabilidad de que ocurra el otro. En otras palabras, el hecho de que uno suceda no influye en la probabilidad del otro.

Sucesos dependientes:
Dos sucesos son dependientes si la ocurrencia de uno afecta la probabilidad de que ocurra el otro. En este caso, la probabilidad de un suceso cambia dependiendo de si el otro suceso ha ocurrido o no.

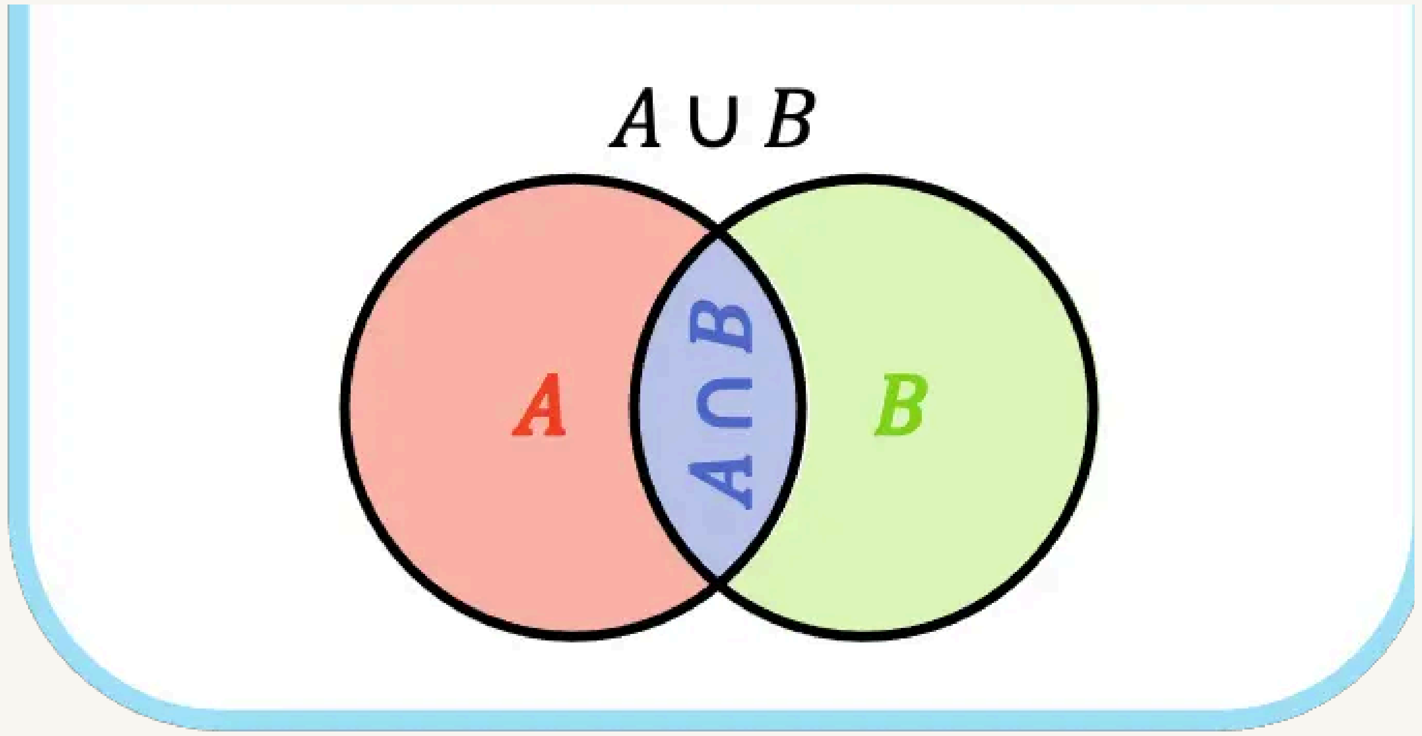
Sucesos Independientes vs Sucesos Dependientes



Reglas fundamentales de probabilidad

- **Regla de la suma:**

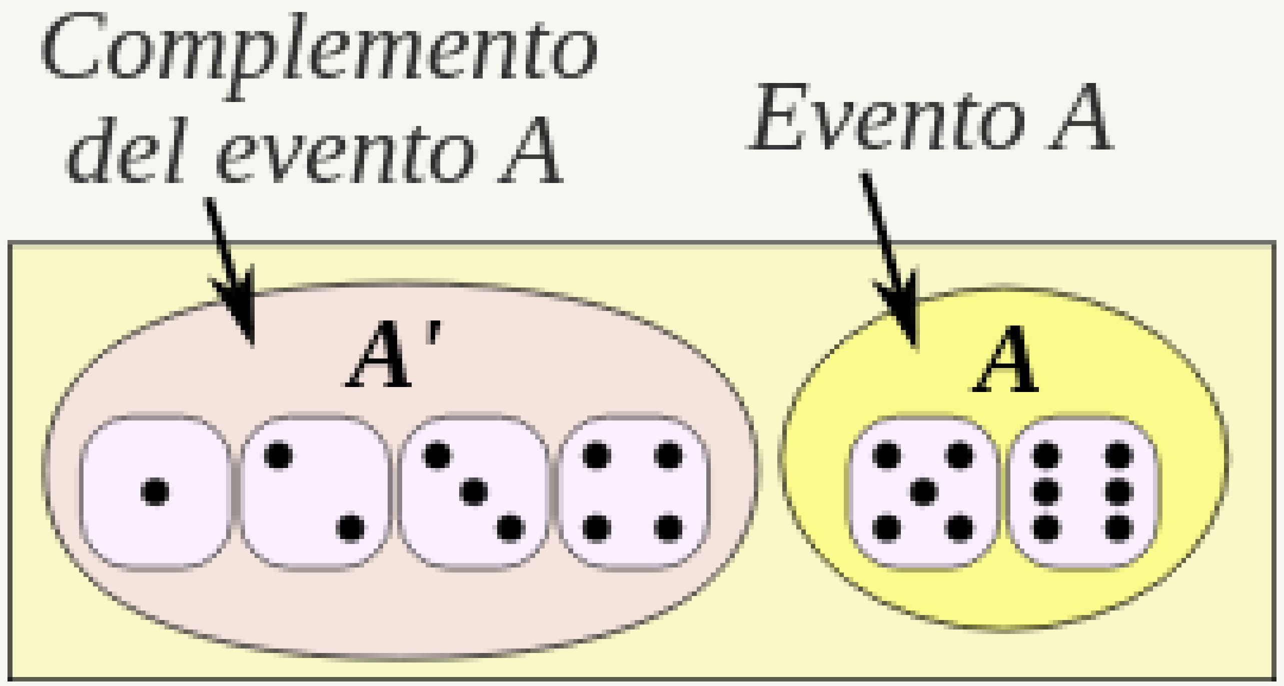
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$



Reglas fundamentales de probabilidad

- Regla del complemento:

$P(A^{c}) = 1 - P(A)$

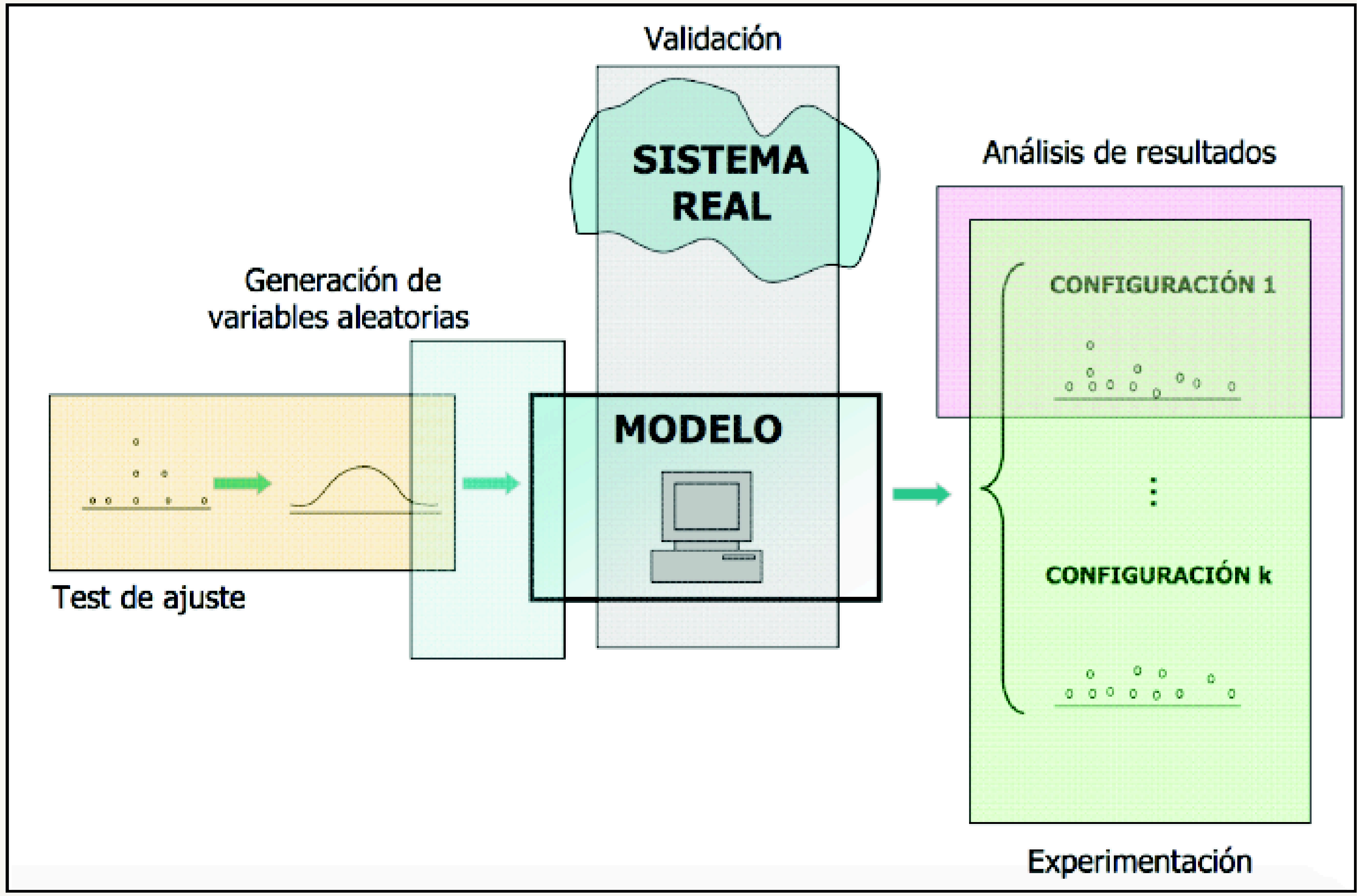


Distribución de Probabilidad

Describe cómo se reparten las probabilidades de los posibles valores de una variable aleatoria

En un contexto de simulación, cada elemento incierto del sistema, se modela mediante una variable aleatoria con una cierta distribución.

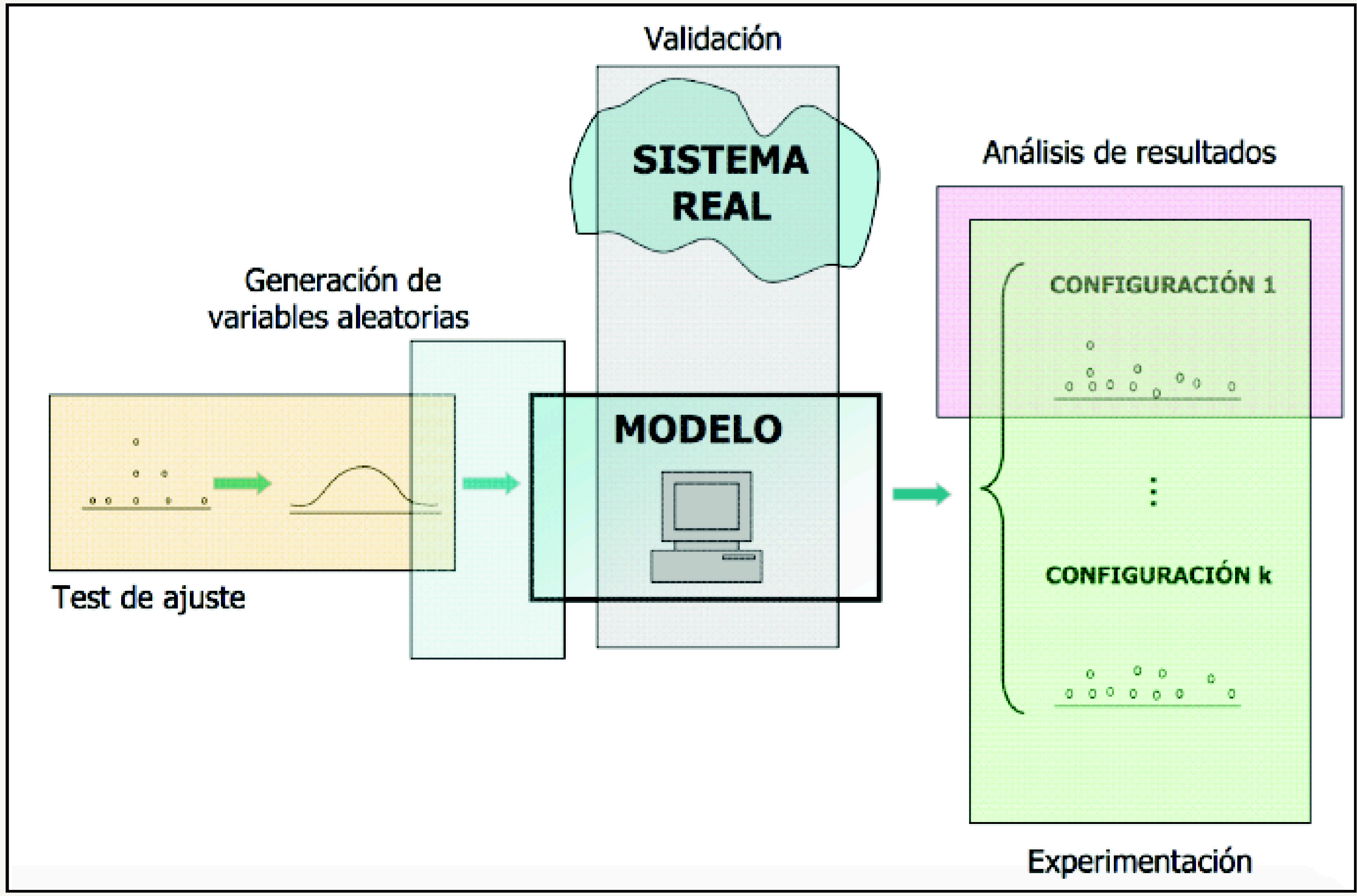
El uso de distribuciones en el modelo permite incorporar el comportamiento estocástico del sistema, imitando su variabilidad natural.



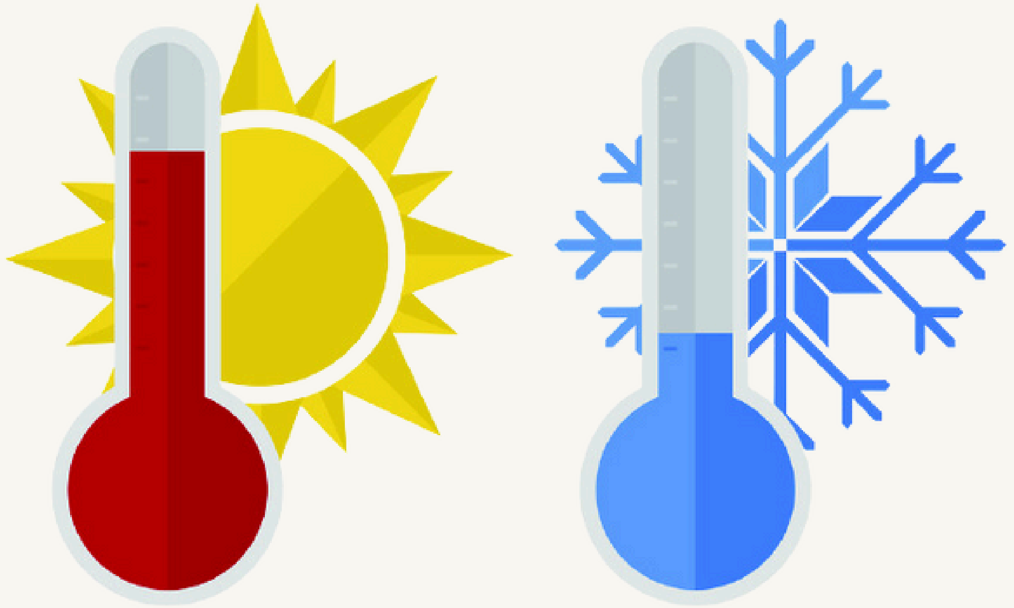
...es un modelo matemático que describe la incertidumbre sobre una variable aleatoria. Se usa para representar y predecir el comportamiento de datos en algoritmos de aprendizaje automático, modelos probabilísticos y redes neuronales.

Distribución de Probabilidad

1. Se observa el sistema real y se recolectan datos.
2. Se ajustan distribuciones para modelar las variables aleatorias.
3. Se crea un modelo computacional que integra esas distribuciones y la lógica del sistema.
4. Se realizan experimentos cambiando configuraciones.
5. Se analizan los resultados para tomar decisiones o proponer mejoras en el sistema real.



Distribuciones

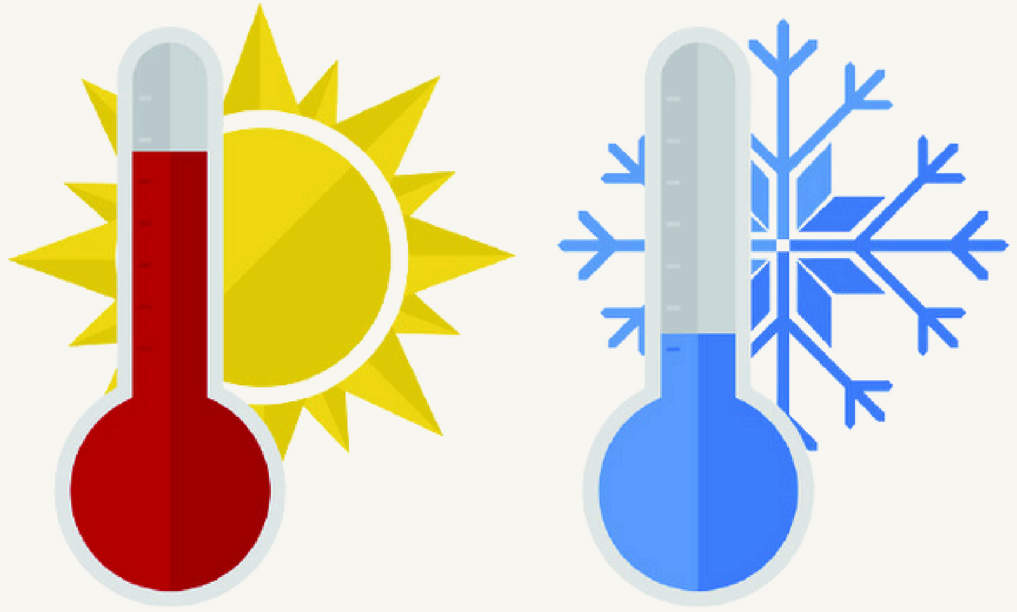


Estado	**Eventos**
Llueve	**9**
No Llueve	**12**

Eventos de lluvia

Probabilidad conjunta y marginal

Distribuciones

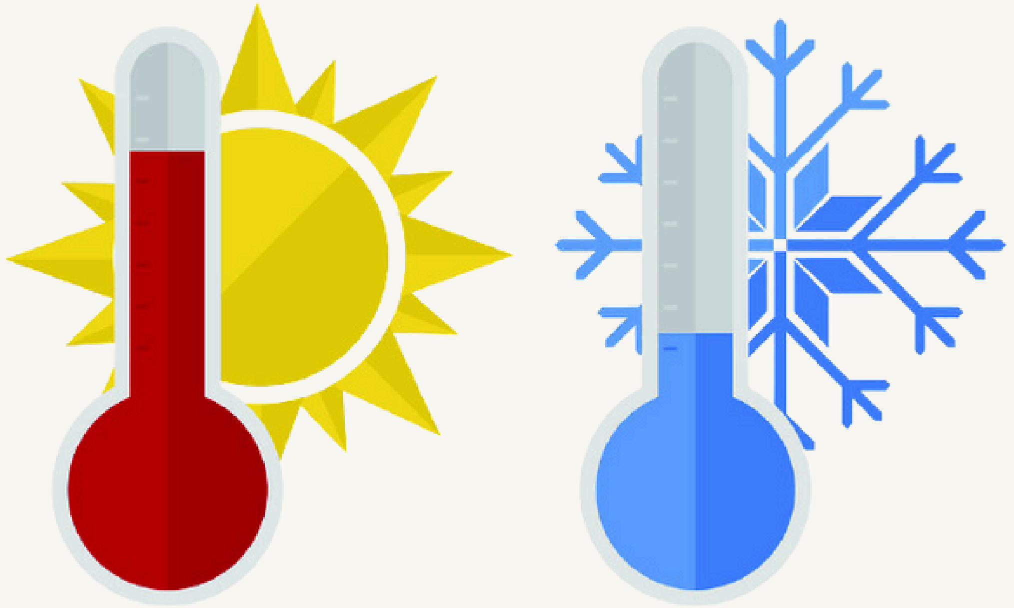


Estado	Eventos
Soleado	10
Nublado	11

Eventos del tiempo

Probabilidad conjunta y marginal

Tabla de Eventos



	Soleado	Nublado	
Llueve	2	7	9
No llueve	8	4	12
	10	11	

Probabilidad conjunta y marginal

Probabilidad Conjunta: Especifica la probabilidad de que dos eventos sucedan (NO necesariamente dependientes).

Probabilidad Marginal: Es la probabilidad de que una sola variable tome un valor específico, ignorando las demás variables.

Probabilidad Marginal

Var 1	Var 2	Probabilidad
Soleado	Llueve	2/21
Soleado	No Llueve	8/21
Nublado	Llueve	7/21
Nublado	No Llueve	4/21

Marginalizamos variable 2 ignorando la variable 1

Estado 2	Probabilidad
Llueve	9/21
No Llueve	12/21

Distribucion Marginal: Es la distribución de una de las variables (o un subconjunto) cuando no nos interesa el resto.

Se obtiene "integrando" o "sumando" la distribución conjunta sobre las variables que se ignoran.

Ejemplo: calcular la **probabilidad que llueva o no llueva**

NOTA: Una distribución es el conjunto completo de probabilidades para todos los valores/eventos, mientras que la probabilidad es la medida para un evento en particular.

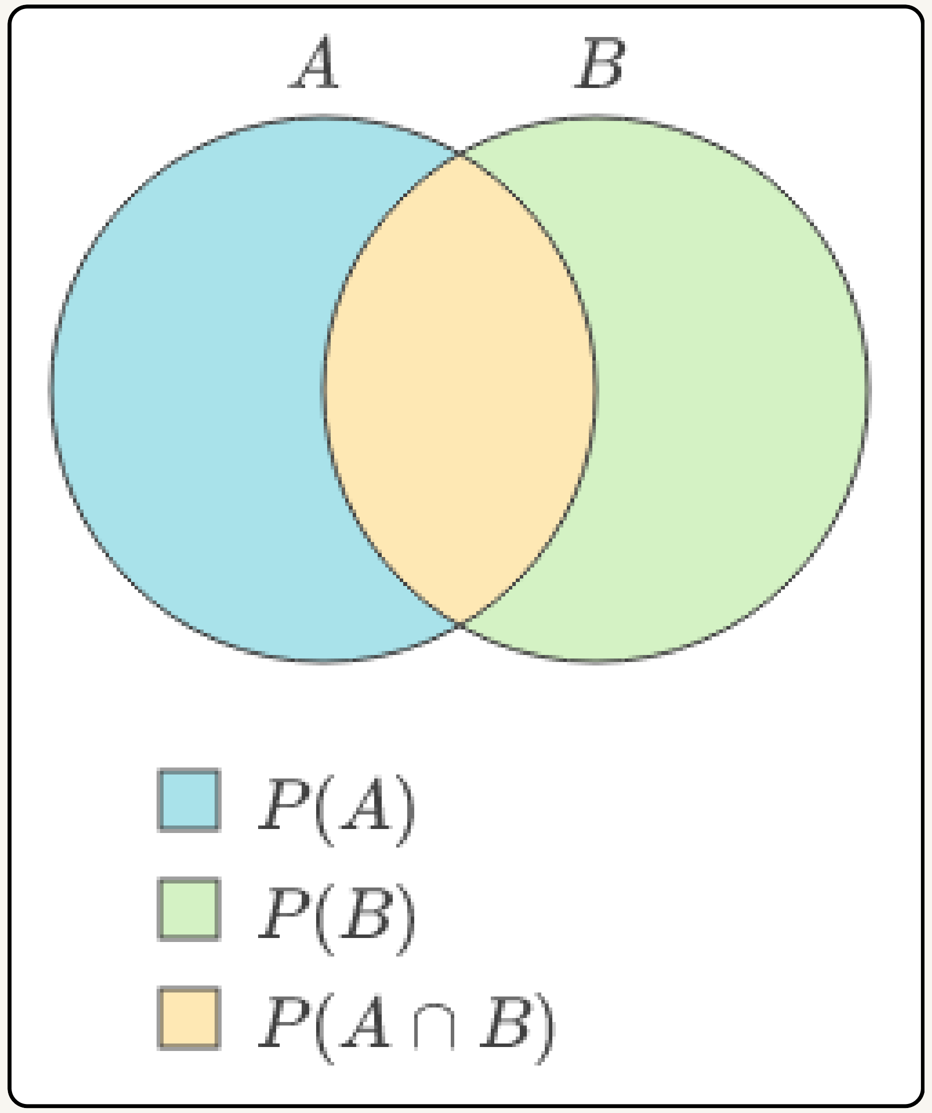
Probabilidad Condicional

Es la probabilidad de que ocurra un evento dado que otro ya ha ocurrido.

Se denota como:

P(A | B) <- Cual es la probabilidad que ocurra A dado que ocurrio B

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$



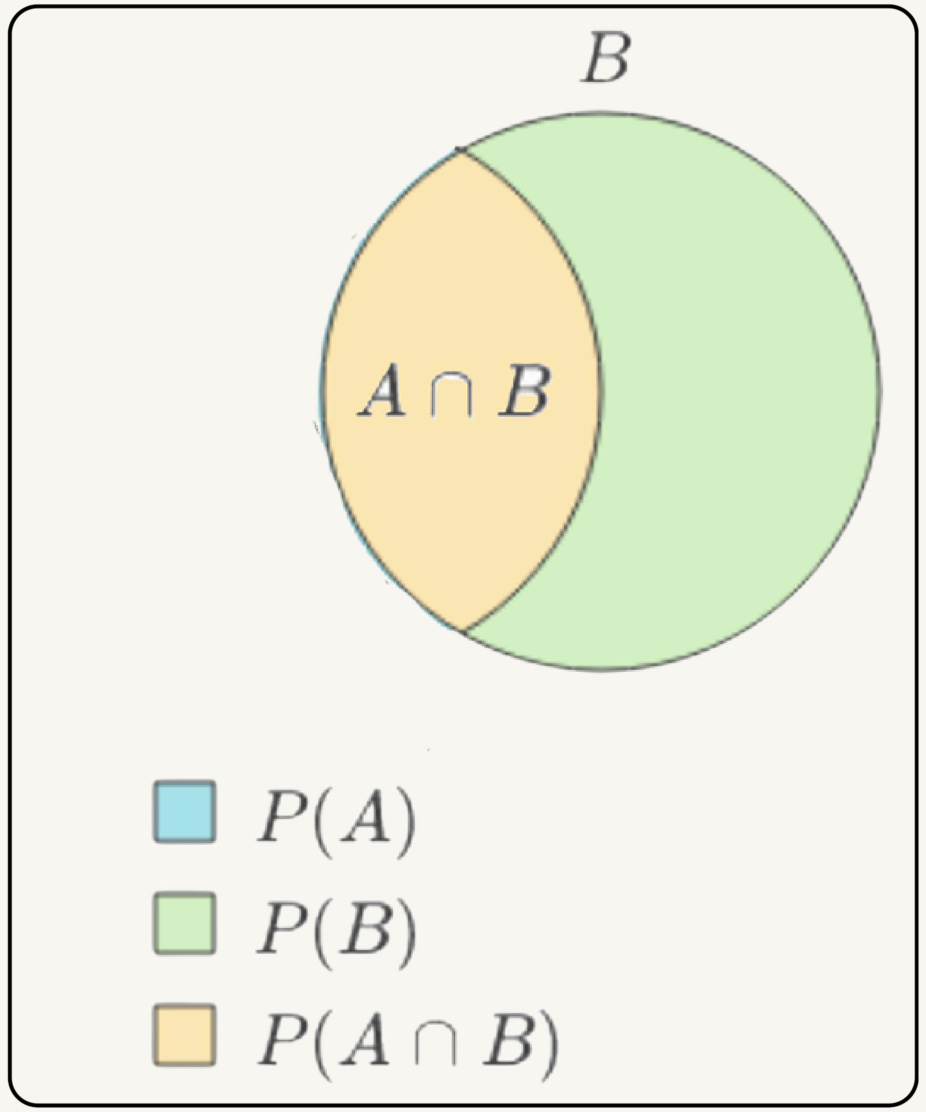
Probabilidad Condicional

Es la probabilidad de que ocurra un evento dado que otro ya ha ocurrido.

Se denota como:

P(A | B) <- Cual es la probabilidad que ocurra A dado que ocurrio B

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$



Probabilidad Condicional

Dada la tabla de probabilidad conjunta

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

Calcule:

1. $P(+x|+y)$
2. $P(-x|+y)$
3. $P(-y|+x)$

Para: $P(+x|+y)$ $P(A|B) = \frac{P(A \cap B)}{P(B)}$

$$P(+x|+y) = \frac{P(+x,+y)}{P(+y)}$$

NO TENEMOS P(+y) !!!!!

Probabilidad Condicional

Dada la tabla de probabilidad conjunta

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

Calcule:

1. $P(+x|+y)$
2. $P(-x|+y)$
3. $P(-y|+x)$

Para: $P(+x|+y)$ $P(A|B) = \frac{P(A \cap B)}{P(B)}$

$$P(+x|+y) = \frac{P(+x,+y)}{P(+y)}$$

NO TENEMOS P(+y) !!!!!
entonces:

Probabilidad marginal P(+y)

$$P(+y) = P(+x,+y)+P(-x,+y)=0.2+0.4=0.6$$

$$P(+x|+y) = \frac{P(+x,+y) = 0.2}{P(+y) = 0.6}$$

$$P(+x|+y) \approx 0.333$$

Probabilidad Condicional

Dada la tabla de probabilidad conjunta

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

Calcule:

1. P(+x|+y)
2. P(−x|+y)
3. P(−y|+x)

Para: P(−x|+y)

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

$$P(-x|+y) = \frac{P(-x,+y)}{P(+y)}$$

$$P(-x|+y) = \frac{P(-x,+y) = 0.4}{P(+y) = 0.6}$$

$$P(-x|+y) \approx 0.667$$

Probabilidad Condicional

Dada la tabla de probabilidad conjunta

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

Calcule:

1. $P(+x|+y)$
2. $P(-x|+y)$
3. $P(-y|+x)$

Para: $P(-y|+x)$ $P(A|B) = \frac{P(A \cap B)}{P(B)}$

$$P(-y|+x) = \frac{P(-y,+x)}{P(+x)}$$

NO TENEMOS P(+x) !!!!!
entonces:

Probabilidad marginal P(+x)

$P(+x) = P(+x,-y)+P(+x,-y)=0.2+0.3=0.5$

$$P(-y|+x) = \frac{P(-y,+x) = 0.3}{P(+x) = 0.5}$$

$P(-y|+x) = 0.6$

y si nos piden calcular ¿P(X|Y = -y)?

Supongamos la siguiente tabla de probabilidades

Var 1	Var 2	Probabilidad
Soleado	Llueve	0.3
Soleado	No Llueve	0.5
Nublado	Llueve	0.4
Nublado	No Llueve	0.6

Suma de probabilidades = 1.8

Normalizar dividiendo entre la suma total

Normalizacion

Es el proceso de asegurar que la suma de todas las probabilidades en un espacio dado sea igual a 1.

Normalizacion

Esto se logra ajustando las probabilidades para que puedan interpretarse como proporciones relativas.

Se parte de una función que no está garantizado que sume 1, y se fuerza a que lo haga dividiéndola entre la suma total de sus valores, obteniendo así una distribución válida.

Supongamos la siguiente tabla de probabilidades

Var 1	Var 2	Probabilidad	Normalizacion	Nueva Probabilidad
Soleado	LLueve	0.3	0.3/1.8	0.1667
Soleado	No Llueve	0.5	0.5/1.8	0.2778
Nublado	Llueve	0.4	0.4/1.8	0.2222
Nublado	No Llueve	0.6	0.6/1.8	0.3333
				1

la suma de las nuevas probabilidades es 1, por lo que la distribución está normalizada

INFERENCIA

Inferencia Probabilistica

Es un enfoque utilizado en inteligencia artificial, estadística y ciencias de la computación para modelar y razonar bajo incertidumbre.

Consiste en utilizar probabilidades para representar creencias sobre eventos desconocidos y actualizar esas creencias a medida que se obtiene nueva información (evidencia).

Su objetivo es calcular la probabilidad de ciertas hipótesis o eventos, dados los datos observados y un modelo probabilístico.

INFERENCIA

Inferencia Probabilistica

Incertidumbre:

- Modela situaciones donde no hay certeza absoluta, como: diagnóstico médico, predicción del clima, detección de spam.
- Ejemplo: ¿Cuál es la probabilidad de que un paciente tenga una enfermedad, dado un resultado positivo en una prueba?

INFERENCIA

Inferencia Probabilistica

Modelos Probabilísticos:

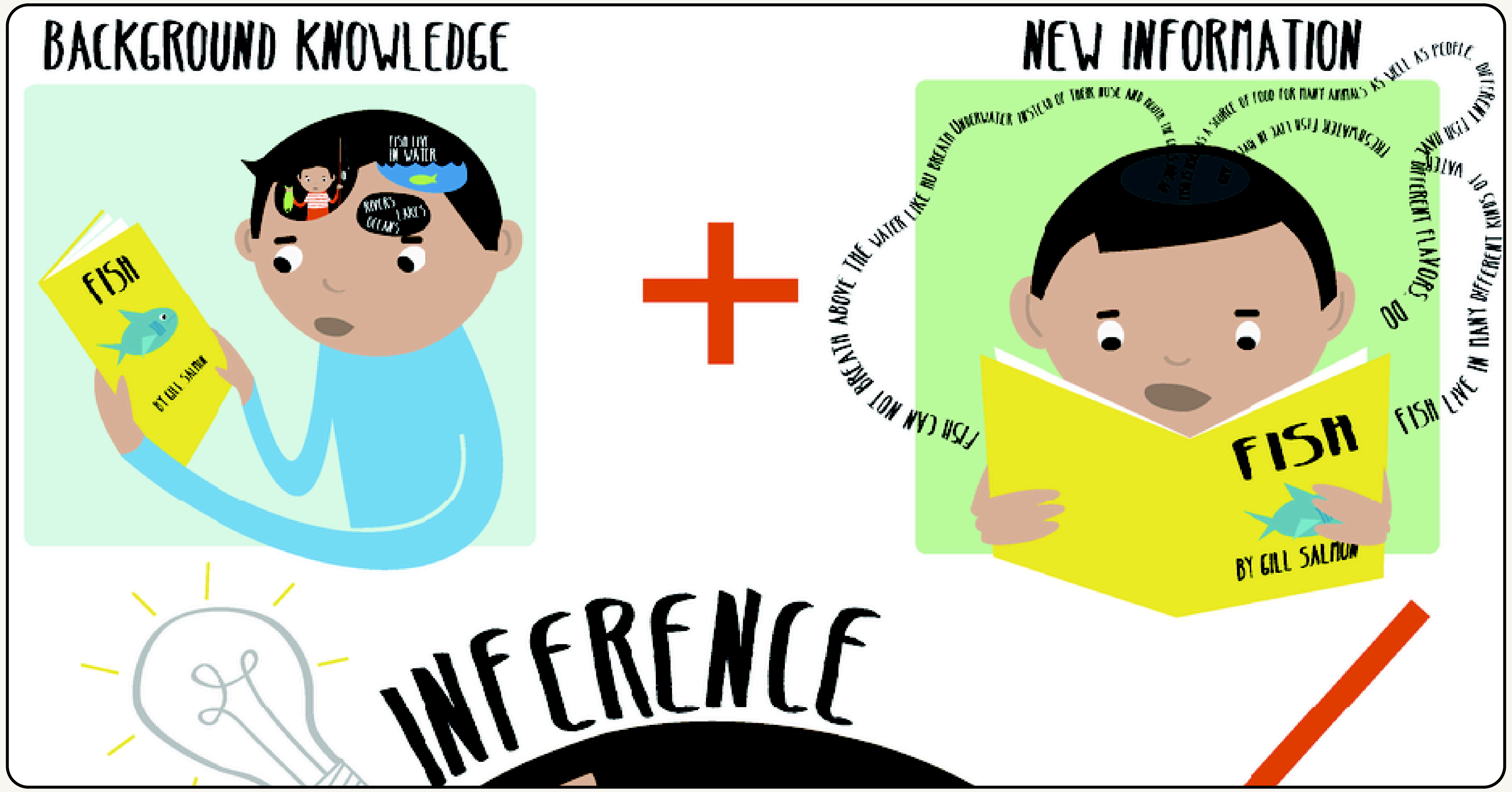
- Representan relaciones entre variables mediante distribuciones de probabilidad.
- Ejemplos: Redes Bayesianas, Cadenas de Markov, Modelos de Mezclas Gaussianas.

INFERENCIA

Inferencia Probabilistica

Actualización de Creencias:

- Usa reglas como el Teorema de Bayes para ajustar probabilidades ante nueva evidencia.



INFERENCIA

Inferencia por Enumeracion

Es un método exacto utilizado en probabilidad y redes bayesianas para calcular probabilidades marginales o condicionales enumerando y sumando todas las posibles combinaciones de variables ocultas (no observadas).

Consiste en:

- Identificar variables:
  - Separar variables en: consulta (X), evidencia (E), y ocultas (Y).

1. Calcular la distribución conjunta
2. Marginalizar variables ocultas
3. Normalizar

Ejemplo:

Supongamos que estamos modelando si una persona tiene gripe X, dado que tiene fiebre E, pero no sabemos si estuvo expuesta a un virus H

- X = ¿El paciente tiene gripe?
- E ?
- Y ?

INFERENCIA

Inferencia por Enumeracion

Supongamos los siguientes datos probabilisticos:

Estacion	Temp.	Clima	Prob.
Verano	calido	Soleado	0.30
Verano	calido	Lluvioso	0.05
Verano	frio	Soleado	0.10
Verano	frio	Lluvioso	0.05
Invierno	calido	Soleado	0.10
Invierno	calido	Lluvioso	0.05
Invierno	frio	Soleado	0.15
Invierno	frio	Lluvioso	0.20

Cual es la probabilidad de:

- P(Clima)
- P(Clima | Invierno)
- P(Clima | Invierno, Calido)

Estacion	Temp.	Clima	Prob.
Verano	calido	Soleado	0.30
Verano	calido	Lluvioso	0.05
Verano	frio	Soleado	0.10
Verano	frio	Lluvioso	0.05
Invierno	calido	Soleado	0.10
Invierno	calido	Lluvioso	0.05
Invierno	frio	Soleado	0.15
Invierno	frio	Lluvioso	0.20

Para P(Clima):

Identificacion de variables:

- Variables de consulta: Clima - C
- Variables de evidencia: Ninguna
- Variables de ocultas: Temperatura - T, Estacion E

P(C) = P(C=Soleado) , P(C=Lluvioso)

Marginalizar P(C)
P(C=Soleado) = 0.30 + 0.10 + 0.10 + 0.15 = 0.65

P(C=Lluvioso) = 0.05 + 0.05 + 0.05 + 0.20 = 0.35

Entonces P(C) es:

Clima	Prob.
Soleado	0.65
Lluvioso	0.35

Estacion	Temp.	Clima	Prob.
Verano	calido	Soleado	0.30
Verano	calido	Lluvioso	0.05
Verano	frio	Soleado	0.10
Verano	frio	Lluvioso	0.05
Invierno	calido	Soleado	0.10
Invierno	calido	Lluvioso	0.05
Invierno	frio	Soleado	0.15
Invierno	frio	Lluvioso	0.20

Para P(Clima | Invierno):

Identificacion de variables:

- Variables de consulta: Clima - C
- Variables de evidencia: Estacion - E = Invierno
- Variables de ocultas: Temperatura - T

P(C | E = Invierno) = P(Soleado | E = Invierno) , P(Lluvioso | E = Invierno)

$$P(C \mid E = \text{Invierno}) = \frac{P(C = \text{Soleado}, E = \text{Invierno})}{P(E = \text{Invierno})}, \frac{P(C = \text{Lluvioso}, E = \text{Invierno})}{P(E = \text{Invierno})}$$

Probabilidad total de Invierno (Marginalizacion):

P(Invierno) = 0.10 + 0.05 + 0.15 + 0.20 = 0.50

Probabilidades de cada clima en Invierno:

P(Soleado,Invierno) = 0.10 + 0.15 = 0.25

P(Lluvioso,Invierno) = 0.05 + 0.20 = 0.25

Estacion	Temp.	Clima	Prob.
Verano	calido	Soleado	0.30
Verano	calido	Lluvioso	0.05
Verano	frio	Soleado	0.10
Verano	frio	Lluvioso	0.05
Invierno	calido	Soleado	0.10
Invierno	calido	Lluvioso	0.05
Invierno	frio	Soleado	0.15
Invierno	frio	Lluvioso	0.20

Para P(Clima | Invierno):

$$P(C \mid E = \text{Invierno}) = \frac{P(C = \text{Soleado}, E = \text{Invierno})}{P(E = \text{Invierno})}, \frac{P(C = \text{Lluvioso}, E = \text{Invierno})}{P(E = \text{Invierno})}$$

Probabilidad total de Invierno (Marginalizacion):

P(Invierno) = 0.50

Probabilidades de cada clima en Invierno (Marginalizacion):

P(Soleado,Invierno) = 0.25
P(Lluvioso,Invierno) = 0.25

$$P(C \mid E = \text{Invierno}) = \frac{0.25}{0.5}, \frac{0.25}{0.5}$$

Estacion	Clima	Prob.
Invierno	Soleado	0.50
Invierno	Lluvioso	0.50

Estacion	Temp.	Clima	Prob.
Verano	calido	Soleado	0.30
Verano	calido	Lluvioso	0.05
Verano	frio	Soleado	0.10
Verano	frio	Lluvioso	0.05
Invierno	calido	Soleado	0.10
Invierno	calido	Lluvioso	0.05
Invierno	frio	Soleado	0.15
Invierno	frio	Lluvioso	0.20

Para P(Clima | Invierno, Calido):

Identificacion de variables:

- Variables de consulta: Clima - C
- Variables de evidencia: Estacion - E = Invierno, T = Calido
- Variables de ocultas: Ninguna

P(C | Calido, Invierno) = P(Soleado | Calido,Invierno) , P(Lluvioso | Calido, Invierno)

$$\mathbf{P(C \mid Calido, Invierno)} = \frac{P(C= Soleado,Cal\text{-}Invierno)}{P(Calido\text{-}Invierno)} , \frac{P(C= Lluvioso,Cal\text{-}Invierno)}{P(Calido\text{-}Invierno)}$$

Probabilidad total de Invierno (Marginalizacion):

P(Calido-Invierno) = 0.10 + 0.05 = 0.15

Probabilidades de cada clima en Invierno:

P(Soleado,Calido-Invierno) = 0.10 = 0.10

P(Lluvioso,Calido-Invierno) = 0.05 = 0.05

Estacion	Temp.	Clima	Prob.
Verano	calido	Soleado	0.30
Verano	calido	Lluvioso	0.05
Verano	frio	Soleado	0.10
Verano	frio	Lluvioso	0.05
Invierno	calido	Soleado	0.10
Invierno	calido	Lluvioso	0.05
Invierno	frio	Soleado	0.15
Invierno	frio	Lluvioso	0.20

Para P(Clima | E=Invierno, T=Calido):

$$P(C \mid E = \text{Invierno}) = \frac{P(C=\text{Soleado}, \text{Cal-Invierno})}{P(\text{Calido-Invierno})}, \frac{P(C=\text{Lluvioso}, \text{Cal-Invierno})}{P(\text{Calido-Invierno})}$$

Probabilidad total de Invierno (Marginalizacion):

P(Calido-Invierno) = 0.15

Probabilidades de cada clima en Invierno (Marginalizacion):

P(Soleado,Calido-Invierno) = 0.10
P(Lluvioso,Calido-Invierno) = 0.05

$$P(C \mid E = \text{Invierno}) = \frac{0.10}{0.15}, \frac{0.05}{0.15}$$

Estacion	Temp.	Clima	Prob.
Invierno	calido	Soleado	0.6667
Invierno	calido	Lluvioso	0.3333

Regla del Producto

Si sabemos que:

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

$$\downarrow$$

$$P(A,B) = P(A|B)\,P(B)$$

Regla del Producto

$$P(A,B) = P(A|B)\,P(B)$$

P(C)

Clima	Prob.
Soleado	0.8
Lluvioso	0.2

X

P(D | C)

Humedad	Clima	Prob.
Humedo	Soleado	0.10
Seco	Soleado	0.90
Humedo	Lluvioso	0.70
Seco	Lluvioso	0.30

P(D , C)

Humedad	Clima	Prob.
Humedo	Soleado	0.08
Seco	Soleado	0.72
Humedo	Lluvioso	0.14
Seco	Lluvioso	0.06

INFERENCIA

Regla de la Cadena

Es una herramienta en probabilidad y cálculo que permite descomponer probabilidades conjuntas complejas en multiplicaciones de probabilidades condicionales más simples.

$$P(A,B,C)=P(A)\cdot P(B \mid A)\cdot P(C \mid A,B)$$

Es decir, la probabilidad de que ocurran **A, B** y **C** juntos es igual a la probabilidad de A, multiplicada por la probabilidad de B dado A, multiplicada por la probabilidad de C dado A y B.

INFERENCIA

Regla de la Cadena

Ejemplo: Calcular la probabilidad de que usando la regla de la cadena:

1. Hoy esté nublado (N),
2. Luego llueva (L), y
3. Tu paraguas se moje (M).

$$P(N,L,M)=P(N)\cdot P(L \mid N)\cdot P(M \mid N,L)$$

- P(N): Probabilidad de que esté nublado.
- P(L|N): Probabilidad de lluvia si está nublado.
- P(M|N, L): Probabilidad de que el paraguas se moje si está nublado y llueve.

Universidad Rafael Landivar
Inteligencia Artificial
Primer Semestre 2025

Hoja de Trabajo No. 5

Fundamentos de Probabilidad

Eres un analista de datos en una empresa de marketing deportivo. Acaban de entregarte los resultados de una encuesta realizada a 1,000 personas en España, donde se les preguntó:

- Tipo de persona: ¿Eres adulto o niño?
- Preferencia alimenticia: ¿Prefieres comida saludable o no saludable?
- Equipo de fútbol: ¿Eres fanático del Barcelona o del Real Madrid?

Con estos datos, has construido una tabla de probabilidad conjunta que resume las preferencias combinadas de la muestra. Por ejemplo:

- El 15% de los encuestados son adultos que prefieren comida saludable y aman al FC Barcelona.
- El 27% de los niños prefieren comida no saludable y son fanáticos del Real Madrid

Estos números no son solo estadísticas frías, representan patrones ocultos que podrían ayudar a:

- Diseñar campañas de alimentación saludable dirigidas a grupos específicos.
- Crear promociones deportivas personalizadas según los gustos de los fanáticos.
- Entender cómo interactúan la edad, la alimentación y la pasión futbolística.

A B C

Persona	Equipo	Comida	Probabilidad conjunta
Adulto	Barcelona	Saludable	0.15
Adulto	Barcelona	No Saludable	0.10
Adulto	Real Madrid	Saludable	0.05
Adulto	Real Madrid	No Saludable	0.20
Niño	Barcelona	Saludable	0.12
Niño	Barcelona	No Saludable	0.08
Niño	Real Madrid	Saludable	0.03
Niño	Real Madrid	No Saludable	0.27

Resuelva las siguientes preguntas:

1) La probabilidad de que sea un Adulto
2) La probabilidad de un Equipo
3) Demuestre P(Real Madrid, No Saludable, Niño) = 0.27 utilizando regla de la cadena
4) Cual es la probabilidad que dado que una persona es adulta y coma saludable, resulte ser fanatico del Barcelona
5) Cual es la distribución de probabilidad para ser de un equipo si es un niño y come saludable

$$P(A,B,C) = P(A) \cdot P(B|A) \cdot P(C|A,B)$$

Universidad Rafael Landívar
Inteligencia Artificial
Primer Semestre 2025

Hoja de Trabajo 6

Descripción

En el marco del **Fin de Semana de Inteligencia Artificial**, organizado en colaboración con **Microsoft**, **Código Facilito** ha puesto a disposición una amplia oferta educativa que incluye más de **180 cursos especializados en Inteligencia Artificial y Ciencia de Datos**. Esta iniciativa busca facilitar el acceso a conocimientos clave en áreas como **Python, Machine Learning, Redes Neuronales, MLOps y fundamentos de IA**, entre otros:

- https://codigofacilito.com/fin/ia/cursos

El curso **"Probabilidad para Modelos de Aprendizaje"** es ideal para reforzar los conceptos de probabilidad para el Curso de Inteligencia Artificial.

Instrucciones

- Completa el curso "Probabilidad para Modelos de Aprendizaje" (1h 18m) disponible en la plataforma.
- Hay un ejercicio de Teorema de Bayes que se explica en dicho curso virtual, debes resolverlo y subir la respuesta, junto a la evidencia (captura de pantalla) al portal, en el apartado correspondiente.

Una vez que hayas completado el curso **"Probabilidad para Modelos de Aprendizaje"**, te invitamos a explorar otros cursos disponibles en la plataforma para seguir reforzando tus conceptos y ampliar tus habilidades en Inteligencia Artificial. Algunas opciones recomendadas incluyen:

- **Curso de Python Profesional** (8h 39m): Profundiza en el manejo de Python, uno de los lenguajes más utilizados en IA y ciencia de datos.
- **Curso de introducción a Machine Learning** (4h 9m): Aprende los fundamentos del aprendizaje automático y cómo aplicarlos en problemas reales.

- **Curso de Machine Learning con ScikitLearn** (6h 9m): Domina el uso de una de las bibliotecas más populares para el desarrollo de modelos de aprendizaje automático.
- **Curso de MLOps** (2h 53m): Conoce las mejores prácticas para implementar y mantener modelos de IA en producción.

Estos cursos te permitirán consolidar tus conocimientos y adquirir nuevas herramientas para enfrentar desafíos más avanzados en el campo de la IA.

Universidad Rafael Landívar
Inteligencia Artificial
Primer Semestre 2025

Hoja de Trabajo 7

Descripción

Una empresa de ciberseguridad ha desarrollado un sistema de inteligencia artificial para detectar intrusos en su red. Este sistema analiza patrones inusuales en el tráfico de datos y, basándose en algoritmos de aprendizaje automático, emite una alerta cuando sospecha de un ataque.

Según estadísticas históricas se tienen los siguientes datos:

- La probabilidad de que ocurra un intento real de intrusión en un día es del 1%:
  P(ataque)=0.01
- Si ocurre un ataque real, el sistema detecta y emite una alerta en el 95% de los casos:
  P(alerta|ataque)=0.95
- Si no hay ataque, el sistema puede generar una alerta errónea (falso positivo) en el 2% de los días:
  P(alerta|no ataque)=0.02

Utilizando estos datos, el teorema de Bayes y la probabilidad Conjunta, resuelva las siguientes preguntas:

1. Calcular la probabilidad de que, dado que se recibe una alerta, efectivamente se esté produciendo un intento de intrusión.
2. Calcular la probabilidad de que no se esté produciendo un ataque dado que **no** se ha recibido alerta, es decir
3. Determinar la probabilidad de que se emita una alerta en un día cualquiera
4. Si se recibe una alerta, ¿cuál es la probabilidad de que se trate de un falso positivo (es decir, no haya ataque)?
5. Se ha mejorado el sistema y se logra reducir la tasa de falsos positivos al 1%, manteniéndose el resto de los parámetros iguales. Calcular la nueva probabilidad de que se esté produciendo un ataque dado que se recibe una alerta

INFERENCIA Y REDES DE BAYES

PARA INTELIGENCIA ARTIFICIAL

AGENDA

- Teorema de Bayes
- Independencia de variables
- Redes Bayesianas
- D-Separation
- Inferencia por Enumeracion en una red bayesiana
- Eliminacion de variables

TEOREMA DE BAYES

Bayes

Thomas Bayes, matemático británico, estudió el problema de la determinación de la probabilidad de las causas a través de los efectos observados, su teorema se resuelve el problema conocido como **de la probabilidad inversa**

El teorema de Bayes tiene muchas aplicaciones, incluyendo **Aprendizaje Automático**, se usa en modelos de clasificación como el Naïve Bayes.

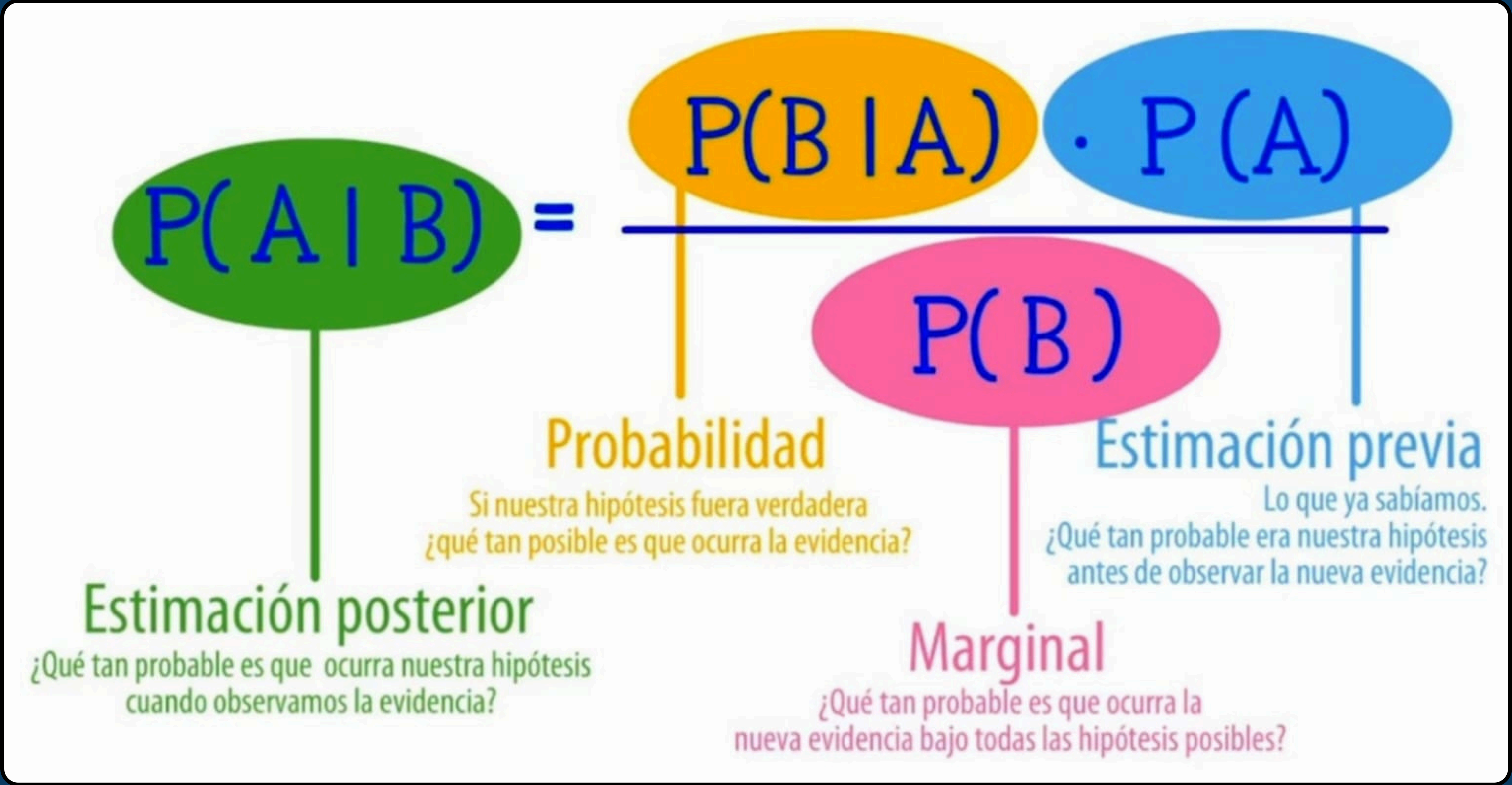
Teorema de Bayes

Es una regla matemática que nos dice cómo actualizar nuestras creencias cuando obtenemos nueva información.

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

$$P(Causa|Efecto) = \frac{P(Efecto|Causa)P(Causa)}{P(Efecto)}$$

Teorema de Bayes



Inferencia con Teorema de Bayes

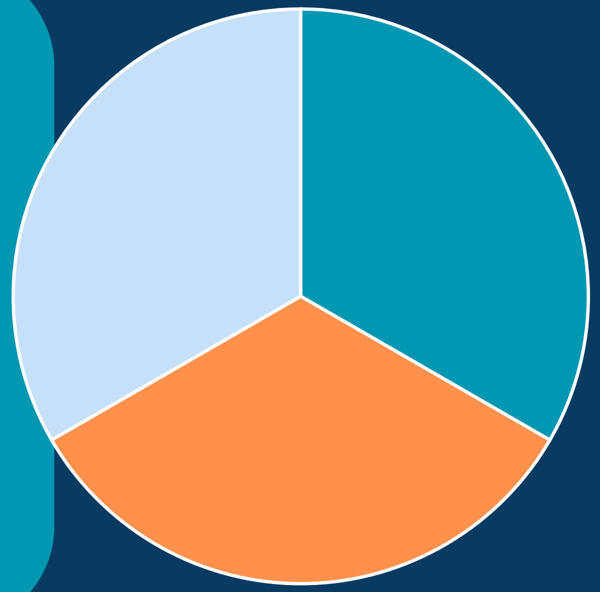
Segun datos del MSPAS durante 2016, 8 de cada 10 pacientes con Chinkunguya presentaban dolor de cuerpo entre sus sintomas.

En una muestra aleatoria de 10,000 Guatemaltecos se determinó que 1 de cada 10 presentaba síntomas de dolor de cuerpo, y existió unicamente un caso confirmado de Chinkunguya.

¿Cual es la probabilidad de tener Chinkunguya si tengo dolor de cuerpo?

Variables

- Dolor de Cuerpo = D (+d,−d)
- Chinkunguya = C (+c,−c)



Probabilidad de tener dolor de cuerpo (*P(+d)*)

que 1 de cada 10 presentaba síntomas de dolor de cuerpo

$$P(+d) = \frac{1}{10} = 0.10$$

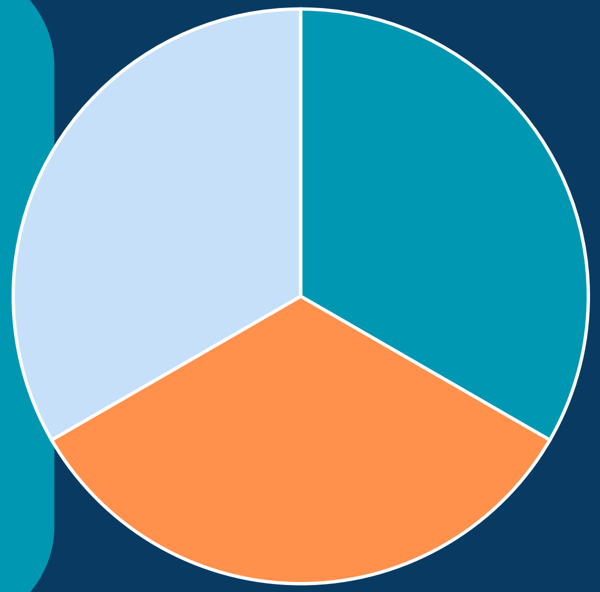
Probabilidad de tener Chinkunguya (*P(+c)*)

En una muestra aleatoria de 10,000 Guatemaltecos y existió unicamente un caso confirmado de Chinkunguya

$$P(+c) = \frac{1}{10,000} = 0.0001$$

Variables

- Dolor de Cuerpo = D (+d,−d)
- Chinkunguya = C (+c,−c)



Probabilidad de tener dolor de cuerpo dado que se tiene Chinkunguya (*P(+d | +c)*)

8 de cada 10 pacientes con Chinkunguya presentaban dolor de cuerpo entre sus síntomas

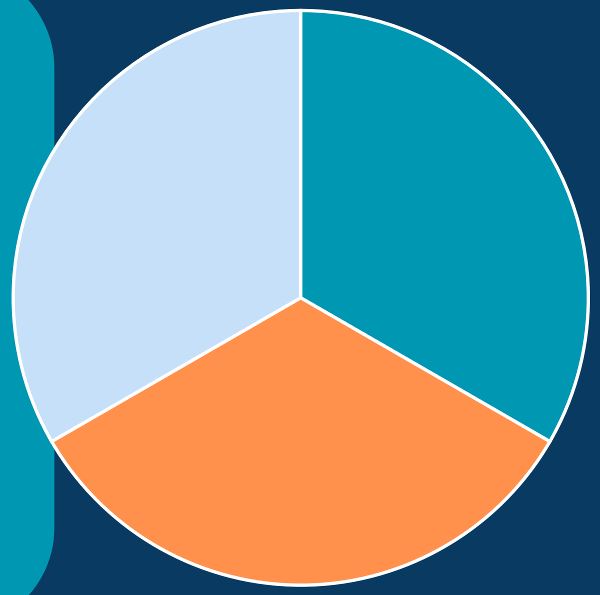
$$P(+d|+c) = \frac{8}{10} = 0.8$$

Entonces si tengo dolor de cuerpo, **¿Cuál es la probabilidad de tener Chinkunguya** (*P(+c | +d)*) ?

$$P(+c|+d) = \frac{P(+d|+c)P(+c)}{P(+d)} = \frac{0.8 * 0.0001}{0.1} = 0.0008$$

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

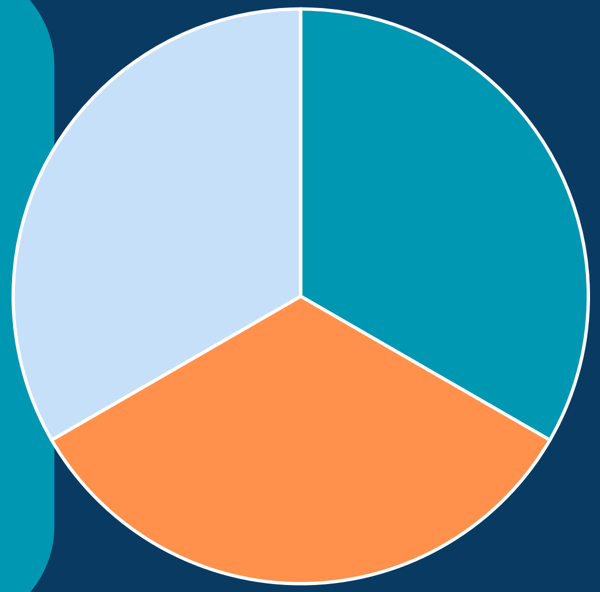
¿Que puede hacer el agente con los modelos?



Modelos probabilisticos describen como es el mundo, por tanto el agente puede:

- Razonar acerca de variables dada cierta evidencia **(inferir)**
- Explicación **(razonamiento diagnostico)**
- Predicción **(razonamiento causal)**

Ejercicio de inferencia con teorema de Bayes



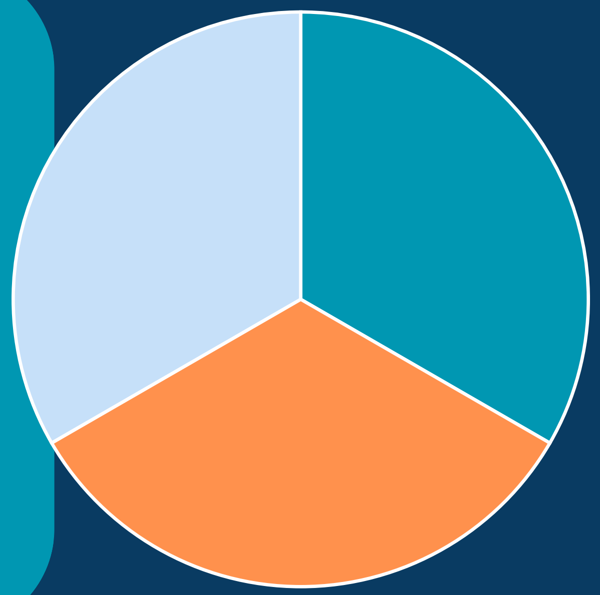
Clima	Prob.
Soleado	0.8
Lluvioso	0.2

D	C	Prob.
Mojado	Soleado	0.1
Seco	Soleado	0.9
Mojado	Lluvioso	0.7
Seco	Lluvioso	0.3

Distrubucion conjunta
o distribucion condicional?

¿P(C|D=seco)?

Si el suelo está seco, ¿qué tan probable es que el clima haya sido soleado o lluvioso?



Tenemos:

2 tipos de clima: Soleado y Lluvioso.

2 estados del suelo: Seco y Mojado.

Probabilidad total de suelo seco

Combinamos ambas opciones (soleado y lluvioso): $P(D = seco) = (0.9 \cdot 0.8) + (0.3 \cdot 0.2) = 0.72 + 0.06 = 0.78$

Aplicamos el Teorema de Bayes

Clima soleado dado que el suelo está seco $$P(C = soleado|D = seco) = \frac{0.9 \cdot 0.8}{0.78} = \frac{0.72}{0.78} \approx 0.9231$$

Clima lluvioso dado que el suelo está seco $$P(C = \underset{\text{lluvioso}}{\cancel{soleado}}|D = seco) = \frac{0.3 \cdot 0.2}{0.78} = \frac{0.06}{0.78} \approx 0.0769$$

Conclusión

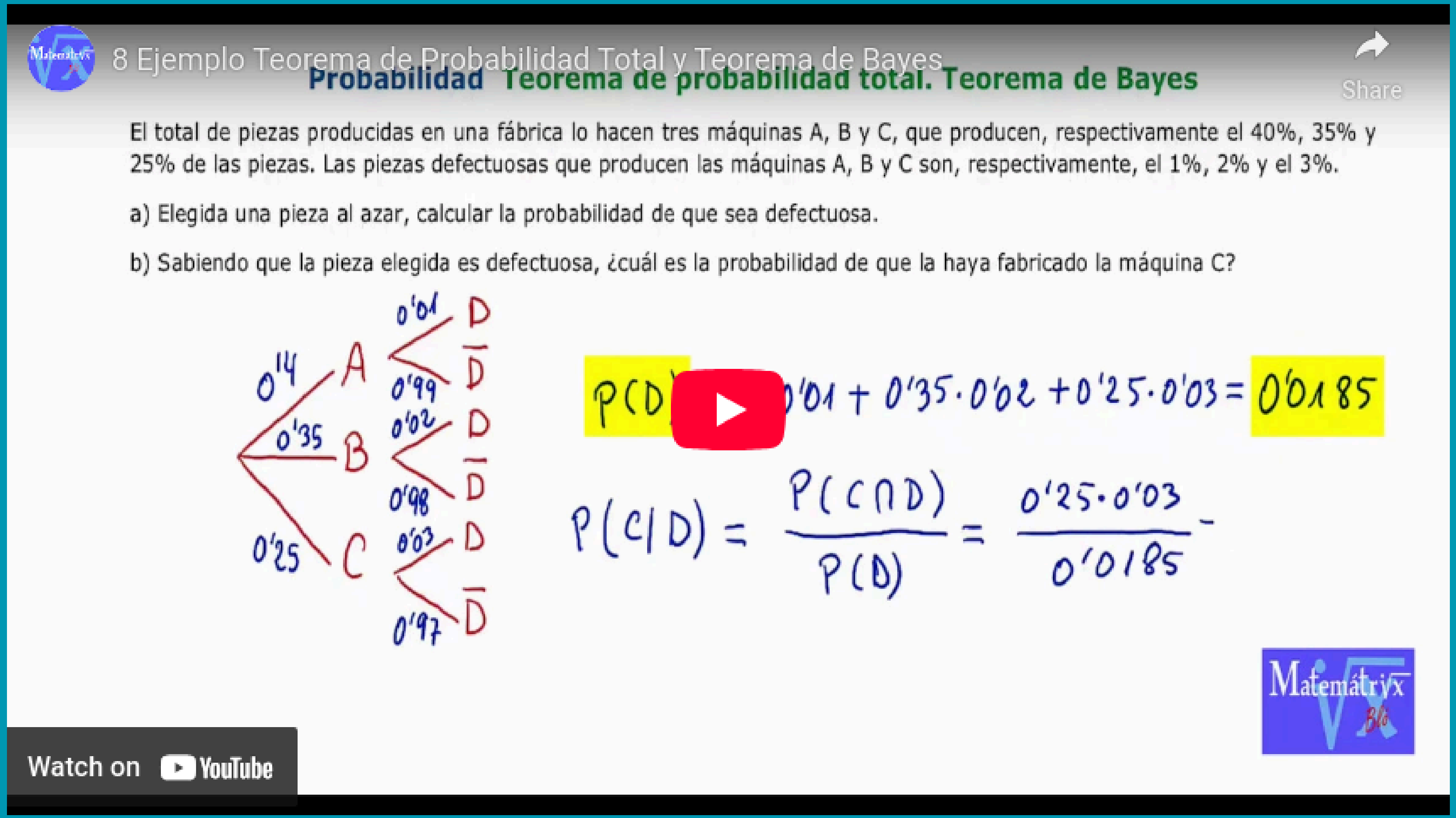
Si el suelo está seco, hay:

- 92.3% de probabilidad de que el clima sea soleado.
- 7.7% de probabilidad de que el clima sea lluvioso.

Teorema de probabilidad total

Es una herramienta que nos permite calcular la probabilidad de un evento considerando todas las maneras en que ese evento puede ocurrir, a través de una partición del espacio muestral.

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




Probabilidad condicional explicada de manera visual (Teorema de Bayes) | Khan Acade...
Share
Watch on YouTube

Resuelve...

En un sorteo recibe un premio si sacas una canica roja de un saco de 100 canicas, donde 20 canicas son rojas y el resto azules. De las canicas rojas 15 son chicas y 5 grandes, mientras que de las azules 70 son chicas y 10 grandes. Si puedes sentir el tamaño, ¿Qué tamaño te conviene sacar?

Datos del problema

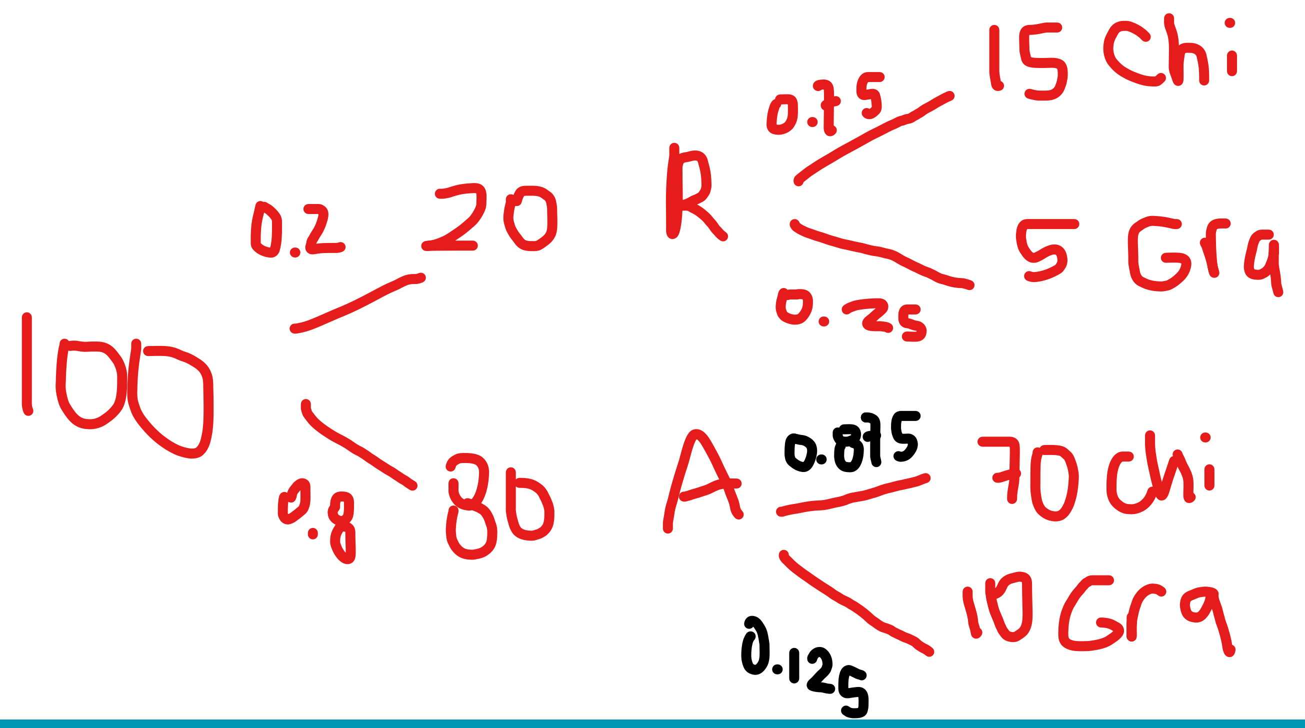
Total de canicas: 100

Canicas rojas: 20

- Chicas: 15
- Grandes: 5

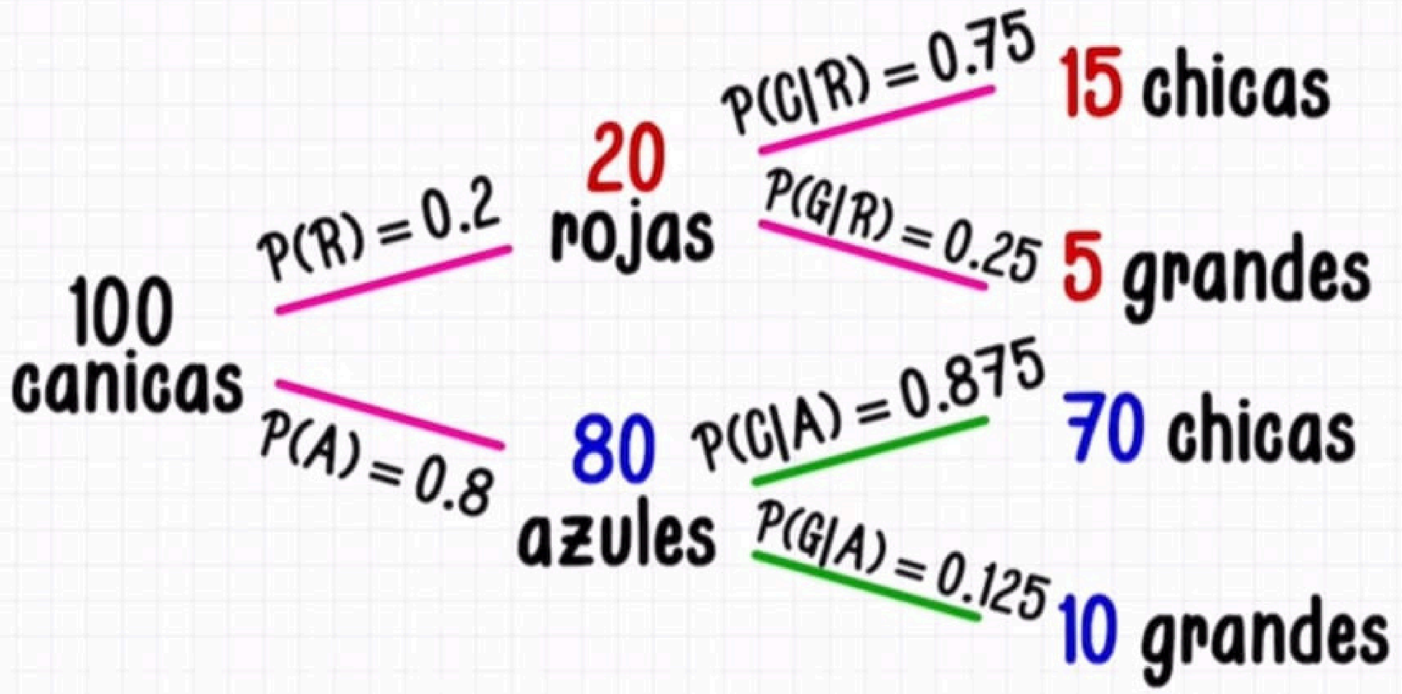
Canicas azules: 80

- Chicas: 70
- Grandes: 10



Resuelve...

Teorema de Bayes



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

$$P(C) = (0.2)(0.75) + (0.8)(0.875) = 0.15 + 0.7 = 0.85$$

$$P(G) = (0.2)(0.25) + (0.8)(0.125) = 0.05 + 0.1 = 0.15$$

$$P(R|C) = \frac{P(C|R)\cdot P(R)}{P(C)} = \frac{(0.75)(0.2)}{0.85} = \frac{0.15}{0.85} = 0.1765 \Big\} \ 17.65\%$$

$$P(R|G) = \frac{P(G|R)\cdot P(R)}{P(G)} = \frac{(0.25)(0.2)}{0.15} = \frac{0.05}{0.15} = 0.3333 \Big\} \ 33.33\%$$

INDEPENDENCIA DE VARIABLES

Independencia de Variables

Se refiere a que el comportamiento de una variable no afecta en absoluto al comportamiento de la otra.

Es decir, conocer el valor de una variable no proporciona ninguna información sobre el valor de la otra.

Decimos que son independientes si, para cualquier par de valores ***x*** e ***y***, se cumple:

$$\forall x, y \;\; P(x,y) = P(x)P(y) \dashrightarrow \;\; X \perp\!\!\!\perp Y$$

La independencia es una suposición fundamental en muchos modelos, ya que simplifica el análisis y el cálculo de probabilidades.

Independencia de Variables

Ejemplo: Lanzamiento de dos dados

Variables

- X = Resultado del primer dado (valores posibles: 1, 2, 3, 4, 5, 6)
- Y = Resultado del segundo dado (valores posibles: 1, 2, 3, 4, 5, 6)

Como es un dado justo:

P(X=x) = *1/6* para *x=1,2,3,4,5,6*

P(Y=y) = *1/6* para *y=1,2,3,4,5,6*

Es decir, cada número tiene la misma probabilidad de salir en cada dado.

Independencia de Variables

Como son dos dados independientes, la probabilidad de que el primer dado sea un 3 y el segundo dado sea un 5 es:

$$P(X=3 \text{ y } Y=5)=P(X=3)\cdot P(Y=5)$$

$$P(X=3 \text{ y } Y=5)=1/6\cdot 1/6=1/36$$

Esta es la fórmula de independencia: el producto de las probabilidades individuales (marginales) nos da la probabilidad conjunta.

El resultado del primer dado no afecta al resultado del segundo dado.
Por eso, saber el valor de X (primer dado) no cambia las probabilidades de Y (segundo dado).
Esto es independencia.

En este caso, tirar un dado no afecta al otro, por lo que:

$$P(X,Y)=P(X)P(Y) \Rightarrow X \perp Y$$

Independencia Condicional

Ocurre cuando dos variables son independientes entre sí al condicionar (fijar) una tercera variable

Si conocemos el valor de esa tercera variable, el conocimiento de una de las dos primeras no aporta información adicional sobre la otra.

$$\forall x, y, z \;\; P(x, y|z) = P(x|z)P(y|z) \dashrightarrow X \perp\!\!\!\perp Y|Z$$

Independencia Condicional

Ejemplo: Clima, paraguas la calle mojada

Variables

- X = La calle está mojada. (sí/no)
- Y = La gente lleva paraguas (sí/no)
- Z = El clima (lluvioso o soleado)

Si vemos a alguien con paraguas, es razonable pensar que probablemente la calle esté mojada (porque quizás llovió), así que:

$$P(X \mid Y) \neq P(X)$$

X y Y no son independientes, ya que ver un paraguas nos da información sobre la calle mojada.

Independencia Condicional

Si sabemos que el clima es lluvioso, entonces:

- Saber que alguien lleva paraguas ya no añade mucha información extra sobre si hay charcos, porque el clima (Z) ya explica eso.

Entonces:

- $P(X \mid Y, Z=\text{lluvioso}) = P(X \mid Z=\text{lluvioso})$
- X y Y son condicionalmente independientes dado Z.

Sin saber el clima, X y Y están relacionados.
Sabiendo el clima (Z), X y Y ya no dependen uno del otro. Toda la información sobre charcos y paraguas está explicada por el clima.

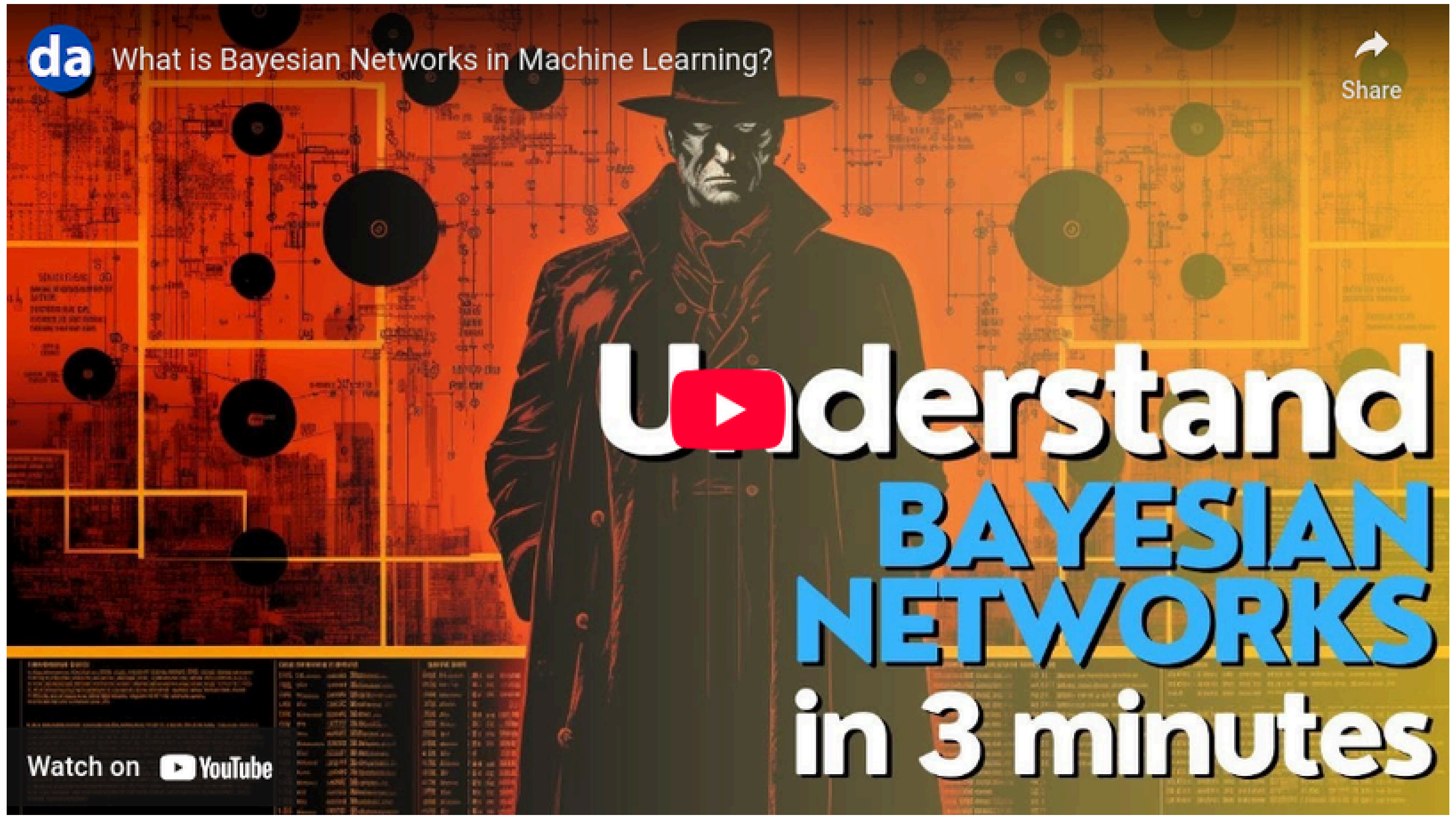
Diferencias Clave

Independencia	Independencia Condicional
$P(X, Y) = P(X)P(Y)$	$P(X, Y \mid Z)=P(X \mid Z) \cdot P(Y \mid Z)$
No depende de ninguna tercera variable	Involucra las variables X, Y y una tercera Z.
Ejemplo: Lanzar dos dados. El resultado de un dado no afecta al otro.	Ejemplo: El uso de paraguas (Y) y que el suelo esté mojado (X) son independientes si sabes que está lloviendo (Z).

- **Independencia: X y Y no se afectan.**
- **Independencia condicional: X y Y parecen relacionados, pero esa relación desaparece cuando conoces Z.**

REDES BAYESIANAS

Redes Bayesianas



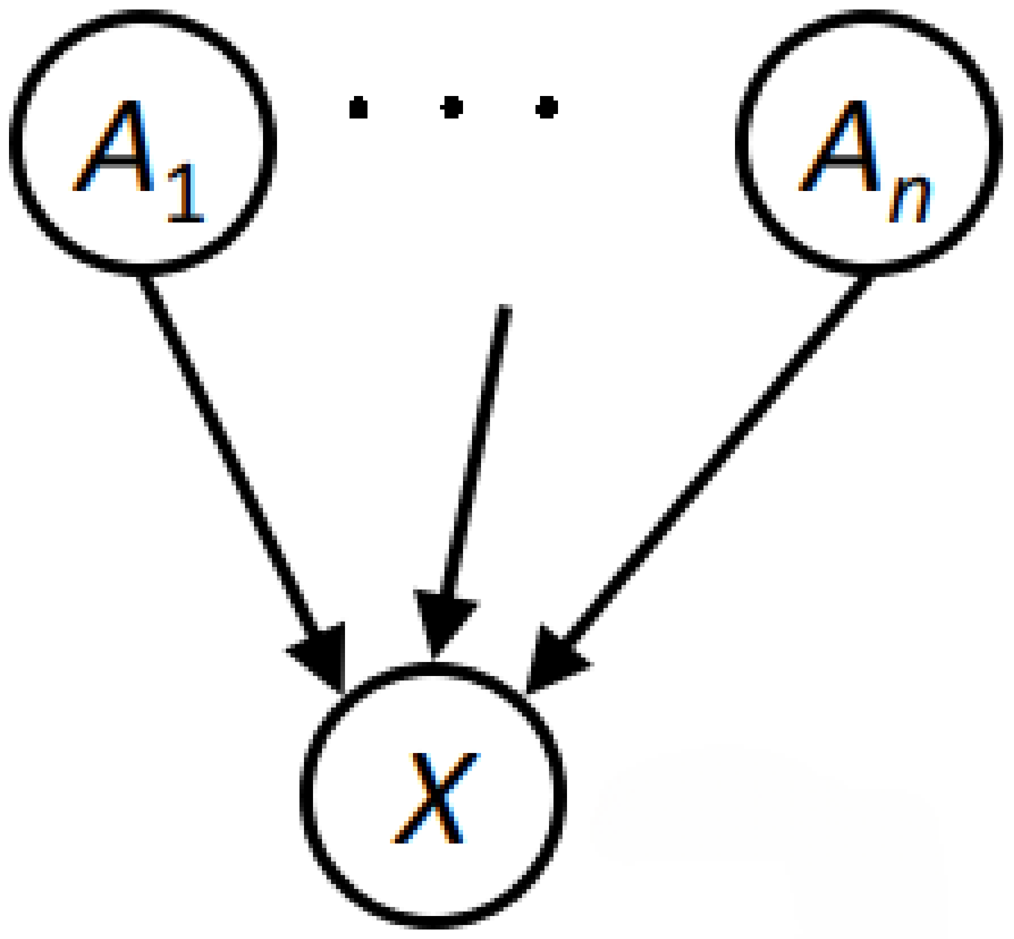
Se conoce también como:

- Red probabilística (Probabilistic Network)
- Red Causal (Causal Network)
- Red de Creencias (Belief Network)
- Mapa de Conocimiento (Knowledge Map)

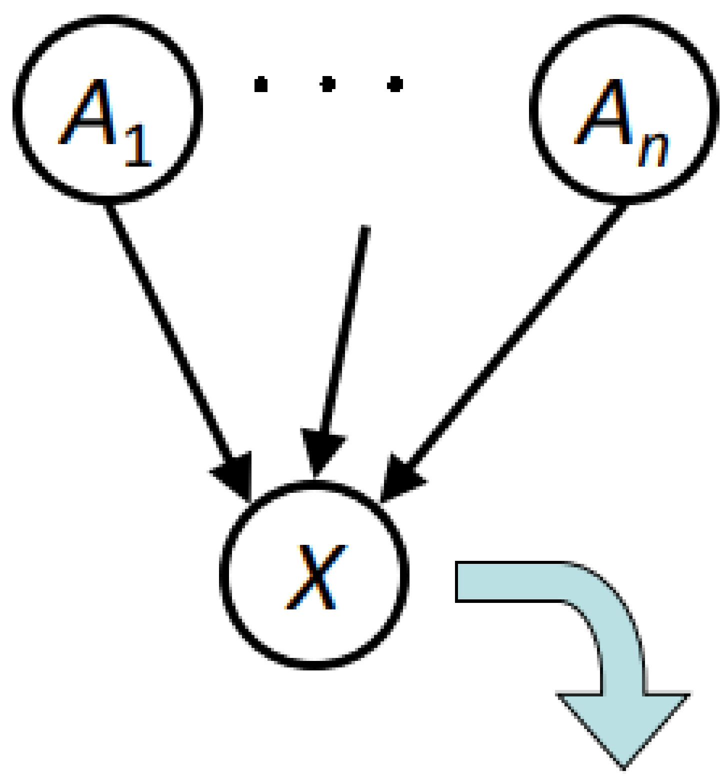
Redes Bayesianas

Es un modelo probabilístico gráfico que representa un conjunto de variables y las relaciones de dependencia condicional entre ellas, utilizando un grafo dirigido acíclico:

- Cada nodo representa una variable aleatoria
- Las flechas o conexiones entre nodos indican la dirección de la influencia o dependencia, muchas veces interpretada como una relación causal



Redes Bayesianas



$$P(X|A_1, \ldots A_n)$$

La red bayesiana utiliza el teorema de Bayes para actualizar las probabilidades de los eventos a medida que se incorpora nueva evidencia

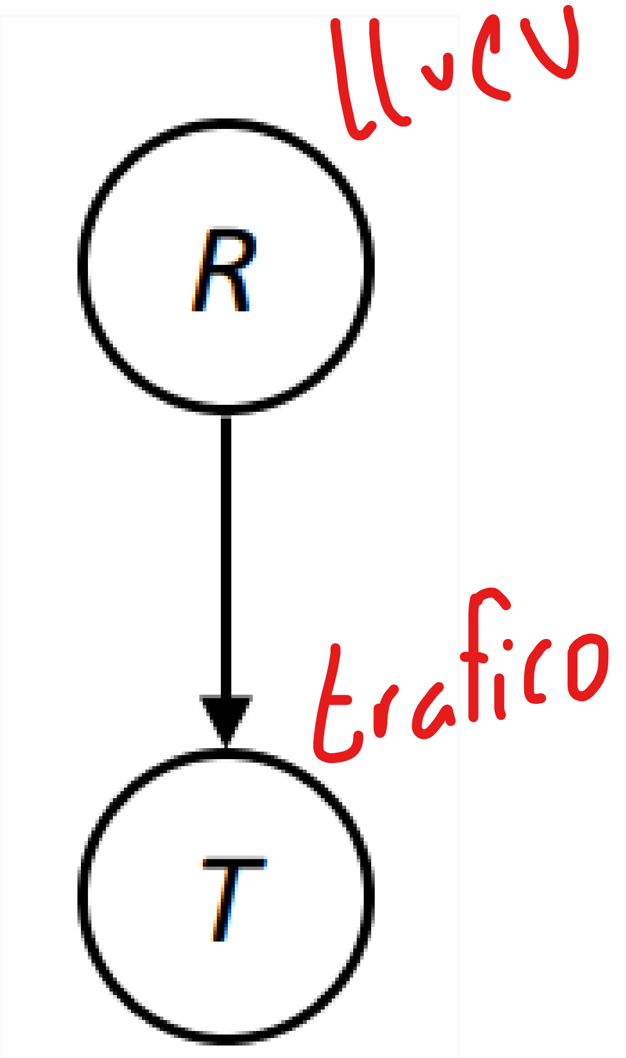
Describe distribuciones conjuntas complejas (modelos), usando distribuciones locales (probabilidades condicionales).

$$P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i|\text{parents}(X_i))$$

Ejemplo Redes Bayesianas – Tráfico y Lluvia

En una ciudad, se ha observado que la probabilidad de que llueva en un día cualquiera es de 1/4. Además, si llueve, la probabilidad de que haya tráfico es de 3/4, mientras que si no llueve, la probabilidad de tráfico es de 1/2.

Utilizando una red bayesiana, calcula la probabilidad de que llueva y no haya tráfico en un día determinado.



$P(R)$

+r	1/4
-r	3/4

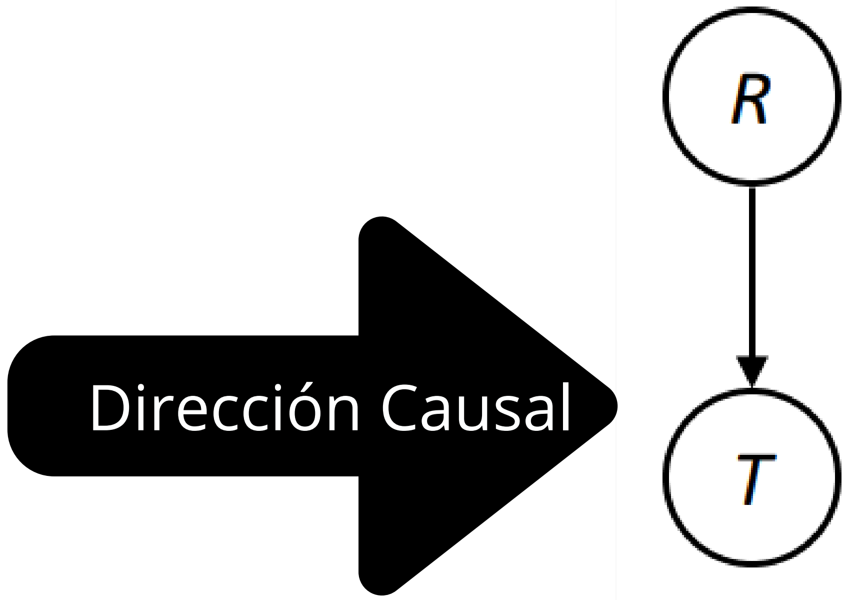
$P(T|R)$

+r	+t	3/4
	-t	1/4
-r	+t	1/2
	-t	1/2

Ejemplo Redes Bayesianas – Tráfico y Lluvia

Calcula la probabilidad de que llueva y no haya tráfico en un día determinado.

Utiliza la fórmula de la probabilidad conjunta para calcular P(+r,−t).



$P(R)$

+r	1/4
-r	3/4

$P(T|R)$

+r	+t	3/4
	-t	1/4
-r	+t	1/2
	-t	1/2

$$P(x_1, x_2, \dots x_n) = \prod_{i=1}^{n} P(x_i | \text{parents}(X_i))$$

$$P(+r, -t) = P(+r)P(-t| + r)$$

$$= \frac{1}{4} * \frac{1}{4} = \frac{1}{16}$$

Ejemplo Redes Bayesianas – Robo

Estás en el trabajo y recibes una llamada de tu vecino John, que te dice que la alarma de tu casa está sonando. Sin embargo, tu otra vecina Mary no te llama. A veces, la alarma se activa por pequeños terremotos.

La pregunta es: *¿Hay un ladrón?*

Variables

La red bayesiana modela estas variables:

- Burglar: ¿Hay un ladrón?
- Earthquake: ¿Hubo un terremoto?
- Alarm: ¿Está sonando la alarma?
- JohnCalls: ¿John te llama?
- MaryCalls: ¿Mary te llama?

Relación causal

La red refleja relaciones causales, es decir, cómo unas variables causan otras:

- Si hay un ladrón, puede activar la alarma.
- Si hay un terremoto, también puede activar la alarma.
- Si la alarma suena, eso puede hacer que:
  - John te llame.
  - Mary te llame.

Ejemplo Redes Bayesianas - Robo

B	P(B)
+b	0.001
-b	0.999

E	P(E)
+e	0.002
-e	0.998

A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

¿Qué te permite esta red?

- Calcular probabilidades conjuntas de todas las variables.
- Estimar, dado que John te llama pero Mary no, cuál es la probabilidad de que haya un ladrón.
- Actualizar tus creencias conforme obtienes evidencia (por ejemplo, si después te enteras que hubo un terremoto, ajustas la probabilidad de que haya un ladrón).

Ejemplo Redes Bayesianas - Robo

B	P(B)
+b	0.001
-b	0.999

E	P(E)
+e	0.002
-e	0.998

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

Para este problema en particular, las variables son:

- Burglary (B): Robo.
- Earthquake (E): Terremoto.
- Alarm (A): Alarma.
- John calls (J): Llamada de John.
- Mary calls (M): Llamada de Mary.

Ejemplo Redes Bayesianas - Robo

B	P(B)
+b	0.001
-b	0.999

E	P(E)
+e	0.002
-e	0.998

A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

El orden debe respetar la estructura de la red, es decir, si X es padre de Y, entonces X debe aparecer antes que Y en el orden:

- Burglary (B) y Earthquake (E) son variables independientes.
- Alarm (A) depende de B y E.
- John calls (J) y Mary calls (M) dependen de A.

Un orden válido es: B,E,A,J,M

Ejemplo Redes Bayesianas - Robo

B	P(B)
+b	0.001
-b	0.999

E	P(E)
+e	0.002
-e	0.998

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

¿Cuál es la probabilidad que se active la alarma, es debido a un robo y se llama únicamente a Mary?

- B: Robo (Burglary)
  - B=+b significa que hubo un robo
  - B=-b significa que no hubo un robo
- E: Terremoto (Earthquake)
  - E=+e terremoto
  - E=-e sin terremoto
- ...
- M: Mary llama (Mary calls)
  - M=+m Mary llama
  - M=-m Mary no llama

Ejemplo Redes Bayesianas - Robo

B	P(B)
+b	0.001
-b	0.999

E	P(E)
+e	0.002
-e	0.998

A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

Queremos encontrar la probabilidad de que la alarma se active debido a un robo y que solo Mary llame.

+b, +m, -e, +a, -j

Esto se puede expresar como:

P(+b, -e, +a, -j, +m)

Probabilidad de que la alarma suene, haya robo, no haya terremoto, Mary llame, y John no

Ejemplo Redes Bayesianas - Robo

$$P(+b, -e, +a, -j, +m) =$$

$$P(+b)P(-e)P(+a|+b, -e)P(-j|+a)P(+m|+a) =$$

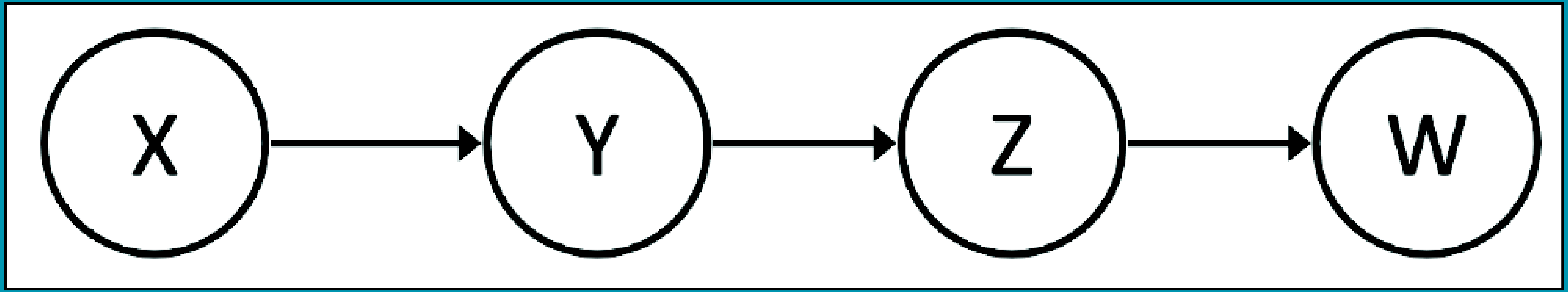
$$0.001 \times 0.998 \times 0.94 \times 0.1 \times 0.7$$

- La regla de la cadena codifica distribuciones conjuntas en secuencia de variables
- Las redes de Bayes asumen que la influencia directa son los padres
- Es comprobable a través del grafo

$$P(x_i|x_1 \cdots x_{i-1}) = P(x_i|parents(X_i))$$

D-SEPARATION

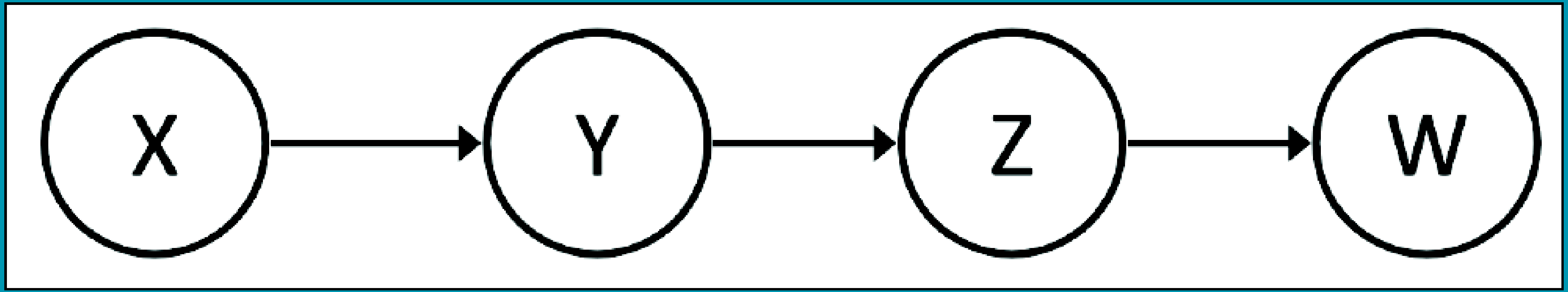
¿Son dos nodos independientes dada la evidencia?



Para determinar si dos nodos son independientes dado otro nodo, debemos ver si la influencia de uno sobre el otro se ve interrumpida por la evidencia.

¿Son dos nodos independientes dada la evidencia?

Digamos que queremos saber si "Estudiar" (X) y "Pasar el Examen" (W) son independientes, dado el conocimiento de "Confianza" (Z)



Estudio: Si el estudiante estudia o no.

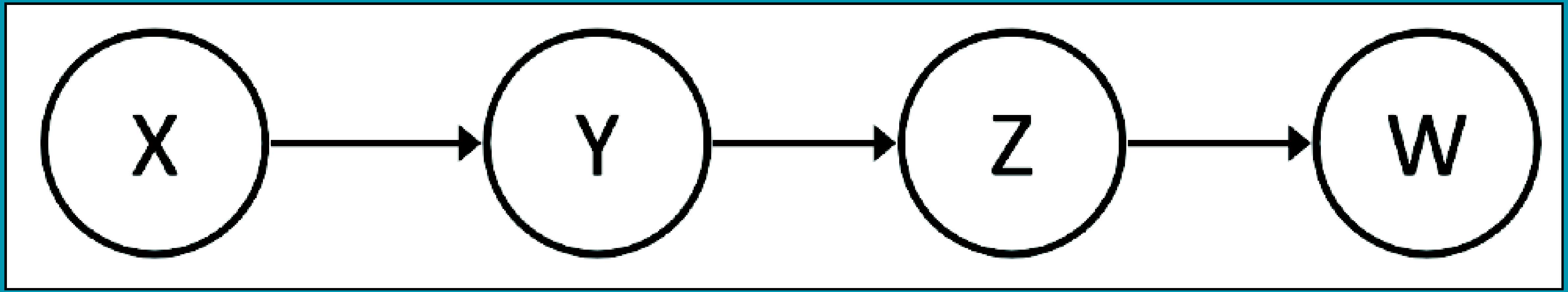
Dormir Bien: Si el estudiante durmió bien la noche anterior al examen.

Confianza: El nivel de confianza del estudiante al entrar al examen.

Pasar el Examen: Si el estudiante pasa o no el examen.

¿Son dos nodos independientes dada la evidencia?

Conociendo Z...



Si sabemos el nivel de confianza (Z), entonces la influencia de estudiar (X) en pasar el examen (W) se modula a través de la confianza (Z).

En este caso, estudiar (X) y pasar el examen (W) son independientes dado que conocemos la confianza (Z).

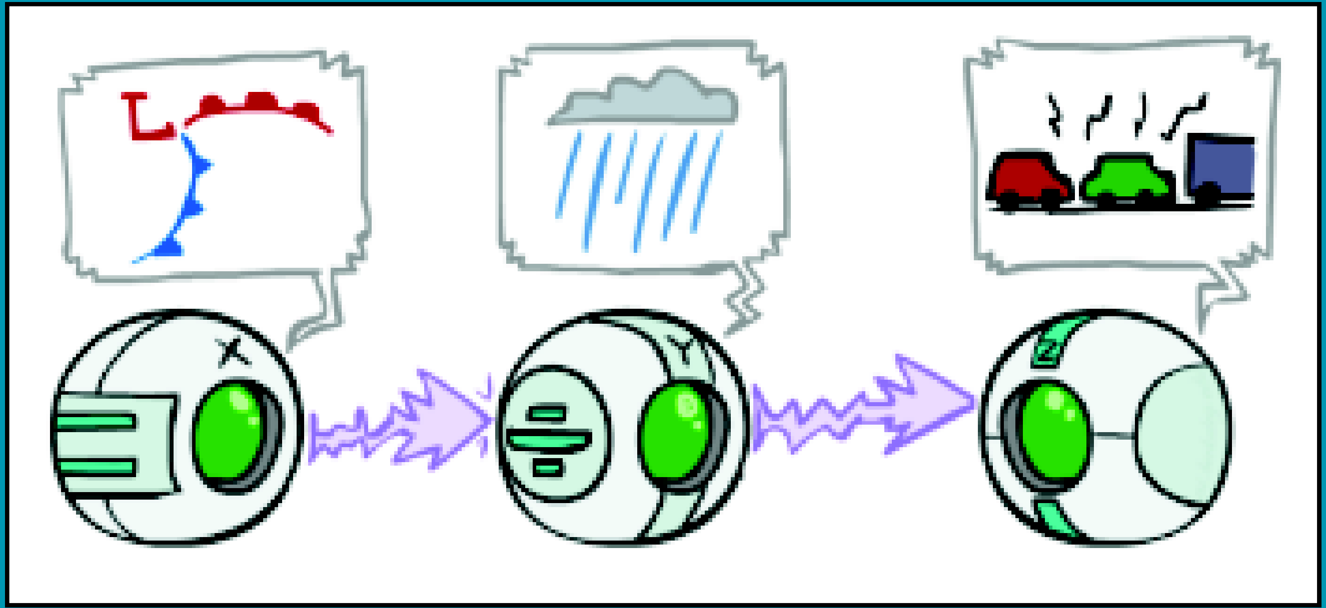
D-Separation

En una red bayesiana, dos nodos A y B están d-separados (es decir, son condicionalmente independientes) dado un conjunto de evidencia E, si todos los caminos entre A y B quedan “bloqueados” por E.

Existen tres patrones básicos a tener en cuenta:

- **Casual Chain**
- **Common Case**
- **Common Effect**

D-Separation - Casual Chain



$$P(x, y, z) = P(x)P(y|x)P(z|y)$$

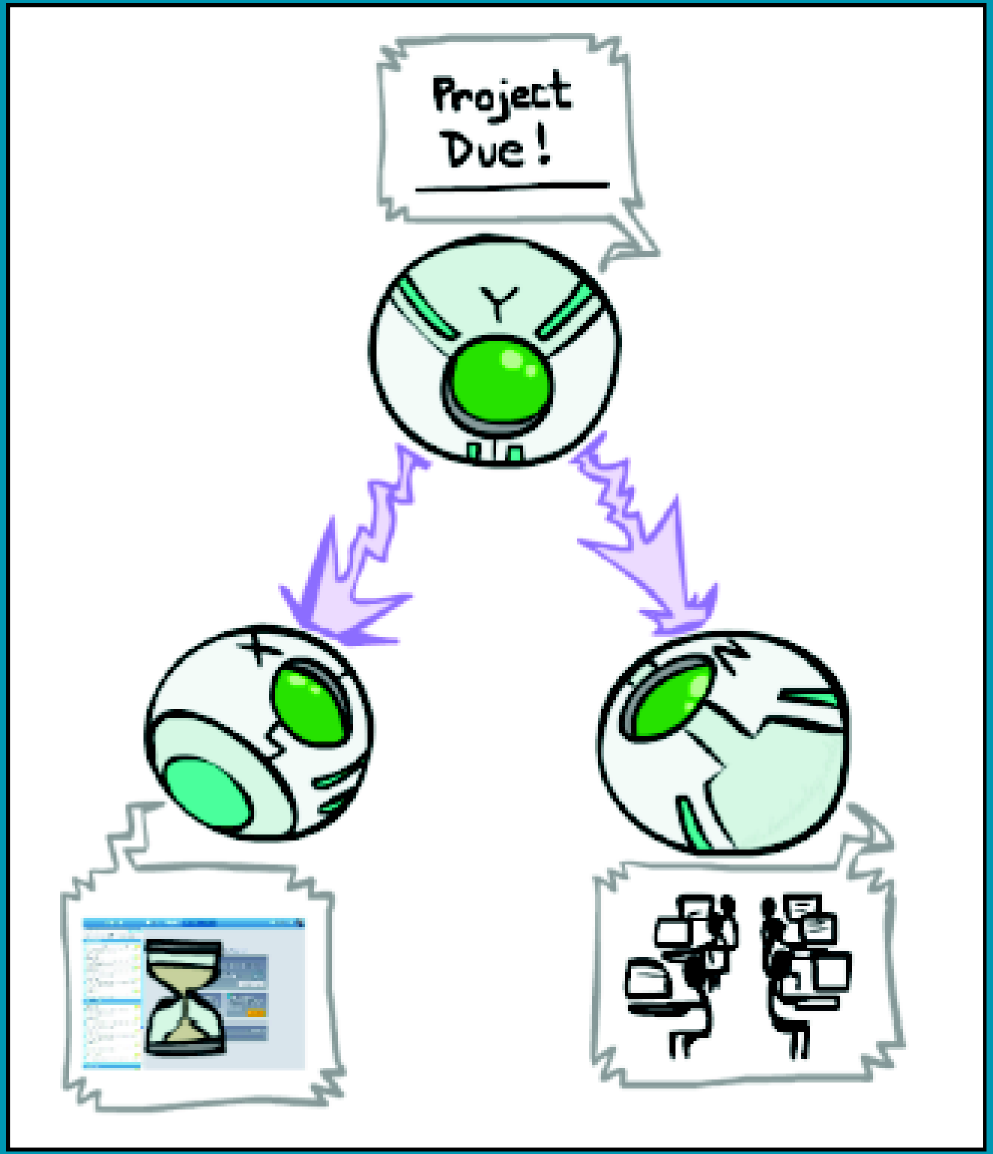
$$P(z|x,y) = \frac{P(x,y,z)}{P(x,y)}$$

$$= \frac{P(x)P(y|x)P(z|y)}{P(x)P(y|x)}$$

$$= P(z|y)$$

Sin condicionar en Y, el camino está abierto (dependencia).
Si condicionas en Y, bloqueas el camino (independencia).

D-Separation - Common Cause



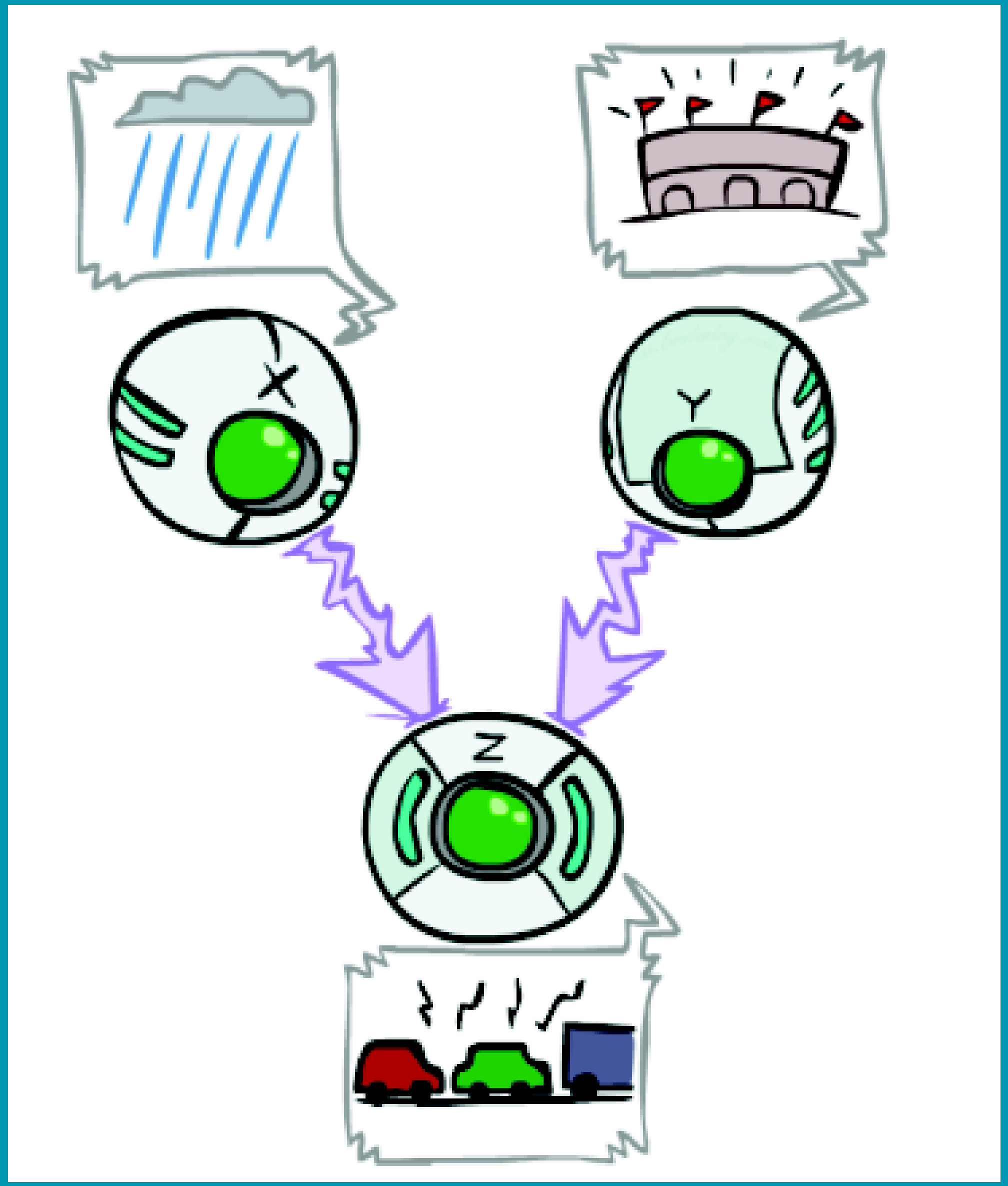
$$P(z|x,y) = \frac{P(x,y,z)}{P(x,y)}$$

$$= \frac{P(y)P(x|y)P(z|y)}{P(y)P(x|y)}$$

$$= P(z|y)$$

Sin condicionar en Y, el camino está abierto (dependencia).
Si condicionas en Y, bloqueas el camino (independencia).

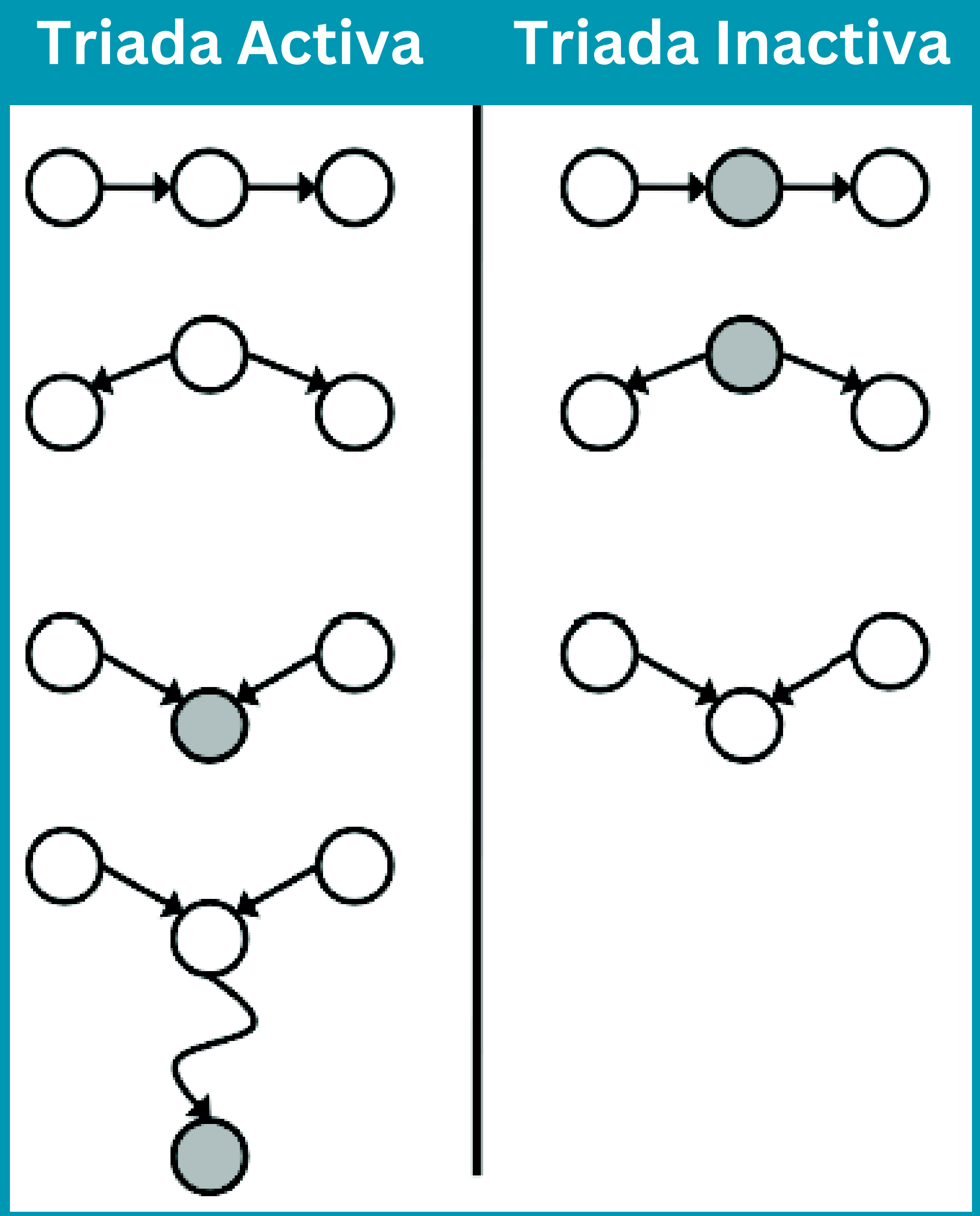
D-Separation – Common Effect



- Sin condicionar en Z, el camino está bloqueado (independencia).
- Si condicionas en Z (o en un descendiente de Z), se abre el camino (dependencia).

D-Separation

- Debe existir camino entre las variables
- Una vez se establece el(los) camino(s) se analizara por triadas
- Un camino es activo si todas las triadas son activas
- Si hay un camino activo, no hay independencia



INFERENCIA

Dada una Red Bayesiana, ¿Cuál es $P(X \mid e)$?

Inferencia

Probabilidad posterior

$$P(Q|E_1 = e_1, \ldots E_k = e_k)$$

$$\left.\begin{array}{ll} \text{Evidence variables:} & E_1 \ldots E_k = e_1 \ldots e_k \\ \text{Query* variable:} & Q \\ \text{Hidden variables:} & H_1 \ldots H_r \end{array}\right\} \begin{array}{c} X_1, X_2, \ldots X_n \\ \textit{All variables} \end{array}$$

Inferencia por Enumeracion

La inferencia por enumeración en redes bayesianas consiste en calcular la probabilidad conjunta de todas las variables relevantes mediante la suma (o marginalización) de todas las combinaciones posibles de valores para las variables no observadas.

Inferencia por Enumeracion

Supongamos que queremos resolver:

$$P(+b \mid +j, +m)$$

Entonces:

$$P(+b \mid +j, +m) = \frac{P(+b, +j, +m)}{P(+j, +m)}$$

$$P(+b, +j, +m) = \sum_a \sum_e P(+b)P(e)P(a \mid b, e)P(j \mid a)P(m \mid a)$$

Inferencia por Enumeracion

$$
\begin{aligned}
P(+b,+j,+m) = \quad & P(+b)P(+e)P(+a|+b,+e)P(+j|+a)P(+m|+a)+ \\
& P(+b)P(+e)P(-a|+b,+e)P(+j|-a)P(+m|-a)+ \\
& P(+b)P(-e)P(+a|+b,-e)P(+j|+a)P(+m|+a)+ \\
& P(+b)P(-e)P(-a|+b,-e)P(+j|-a)P(+m|-a)
\end{aligned}
$$

Calculamos toda (o casi toda la distribución conjunta) y luego enumeramos variables.

Inferencia por Enumeracion

Este enfoque es conceptualmente sencillo, presenta algunas desventajas importantes:

Complejidad Exponencial: La cantidad de combinaciones posibles crece exponencialmente con el número de variables.

Ineficiencia Computacional: Evaluar todas las combinaciones resulta impracticable para redes grandes.

Alto Consumo de Memoria: Guardar todas las combinaciones posibles consume una gran cantidad de memoria.

Redundancia de Cálculo: Muchas combinaciones son redundantes, ya que se vuelven a calcular probabilidades parciales repetidamente.

ELIMINACION DE VARIABLES

Eliminacion de Variables

La técnica de eliminación de variables aborda estas desventajas al evitar la enumeración completa de todas las combinaciones.

Agrupación de Sumas: Se agrupan y simplifican sumas parciales antes de realizar la marginalización.

Factorización Inteligente: Se explota la estructura de independencia de la red para calcular solo las combinaciones necesarias.

Uso de Factores: Se representan los factores condicionales en lugar de trabajar directamente con la distribución conjunta completa.

Eliminacion de Variables - Factor 1

Podemos identificar los diferentes factores o elementos que estan involucrados dentro de una red bayesiana

Distribución conjunta

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Distribución conjunta parcial (selected) P(x, Y)

$P(cold, W)$

T	W	P
cold	sun	0.2
cold	rain	0.3

Eliminacion de Variables - Factor 2

Condicional simple
(suma 1) P(Y |**x**)

$P(W|cold)$

T	W	P
cold	sun	0.4
cold	rain	0.6

Familia de condicionales P(Y |X)

$P(W|T)$

T	W	P
hot	sun	0.8
hot	rain	0.2
cold	sun	0.4
cold	rain	0.6

$P(W|hot)$

$P(W|cold)$

Eliminacion de Variables - Factor 3

Familia especificada (consecuencias) P(y|X)

$P(rain|T)$

T	W	P
hot	rain	0.2
cold	rain	0.6

$P(rain|hot)$
$P(rain|cold)$

Una vez identificados los elementos o factores dentro de la red bayesiana podemos empezar el proceso de eliminacion con el siguiente ejemplo

Ejemplo - Eliminacion de Variables

Hoy juegan los rojos, puede que haya trafico y llegue tarde a clase.

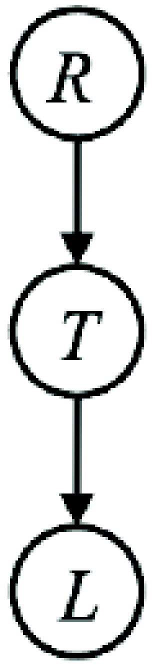
Variables

Juegan los rojos **(R)**
Trafico **(T)**
Llegue tarde a clase **(L)**

¿Llego tarde a clase?

Ejemplo - Eliminacion de Variables

Dada la red bayesiana



$P(R)$

+r	0.1
-r	0.9

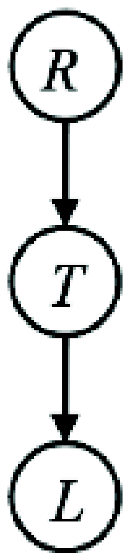
$P(T|R)$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

EV - Join



$P(R)$

+r	0.1
-r	0.9

$P(T|R)$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$P(L|T)$

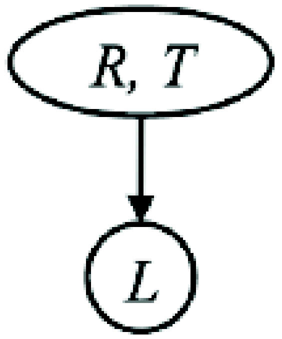
+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

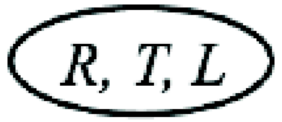
$P(R,T)$

+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9





$P(R,T,L)$

+r	+t	+l	0.024
+r	+t	-l	0.056
+r	-t	+l	0.002
+r	-t	-l	0.018
-r	+t	+l	0.027
-r	+t	-l	0.063
-r	-t	+l	0.081
-r	-t	-l	0.729

EV – Eliminacion – Suma

R, T, L

$P(R, T, L)$

+r	+t	+l	0.024
+r	+t	-l	0.056
+r	-t	+l	0.002
+r	-t	-l	0.018
-r	+t	+l	0.027
-r	+t	-l	0.063
-r	-t	+l	0.081
-r	-t	-l	0.729

T, L

$P(T, L)$

+t	+l	0.051
+t	-l	0.119
-t	+l	0.083
-t	-l	0.747

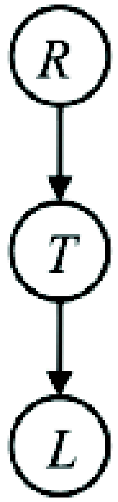
L

$P(L)$

+l	0.134
-l	0.886

EV - Alternativa

Join → **Eliminacion** → **Join** → **Eliminacion** →



$P(R)$

+r	0.1
-r	0.9

$P(T|R)$

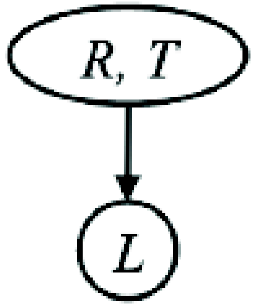
+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

$P(R,T)$

+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81

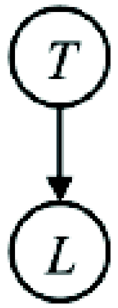


$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

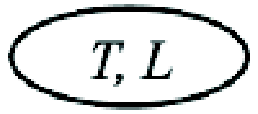
$P(T)$

+t	0.17
-t	0.83



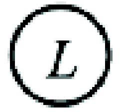
$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9



$P(T,L)$

+t	+l	0.051
+t	-l	0.119
-t	+l	0.083
-t	-l	0.747

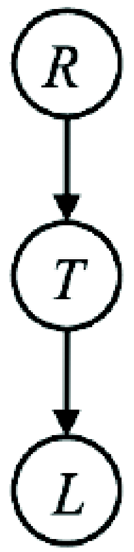


$P(L)$

+l	0.134
-l	0.886

Ejemplo – Eliminacion de Variables

Dada la red bayesiana



$P(R)$

+r	0.1
-r	0.9

$P(T|R)$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

Consultar

$$P(L|+r)$$

Ejemplo - Eliminacion de Variables

Aplicamos condicion simple P(L |+r), eliminamos -r

$P(+r)$

+r	0.1

$P(T|+r)$

+r	+t	0.8
+r	-t	0.2

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

Ejemplo – Eliminacion de Variables

Join

$P(+r)$

+r	0.1

$P(T|+r)$

+r	+t	0.8
+r	-t	0.2

P(T,+r)

+r	+t	0.08
+r	-t	0.02

Ejemplo – Eliminacion de Variables

Join

P(T,+r)

+r	+t	0.08
+r	-t	0.02

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

P(L,T,+r)

+r	+t	+l	0.024
+r	+t	-l	0.056
+r	-t	+l	0.02
+r	-t	-l	0.018

Ejemplo – Eliminacion de Variables

Eliminacion T

$$P(+r, L)$$

+r	+l	0.026
+r	-l	0.074

Ejemplo - Eliminacion de Variables

Normalizar

$$P(L \mid +r) = \frac{P(L,+r)}{P(+r)}$$

$P(+r, L)$

+r	+l	0.026
+r	-l	0.074

→

$P(L| + r)$

+l	0.26
-l	0.74

$$P(+r) = P(+r,+l) + P(+r,-l) = 0.026 + 0.074 = 0.1$$

P(+r) es la probabilidad marginal de +r, que se obtiene sumando P(L,+r) sobre todos los valores posibles de L. Así, los valores de la tabla se “reescalan” de modo que sumen 1, convirtiéndose en una dist. condicional

PROCEDIMIENTO GENERAL

1) Obtenemos una consulta

2) Iniciamos con los factores (probabilidades conjuntas c/ evidencia)

3) Mientras existan variables ocultas se hace join y eliminaciones

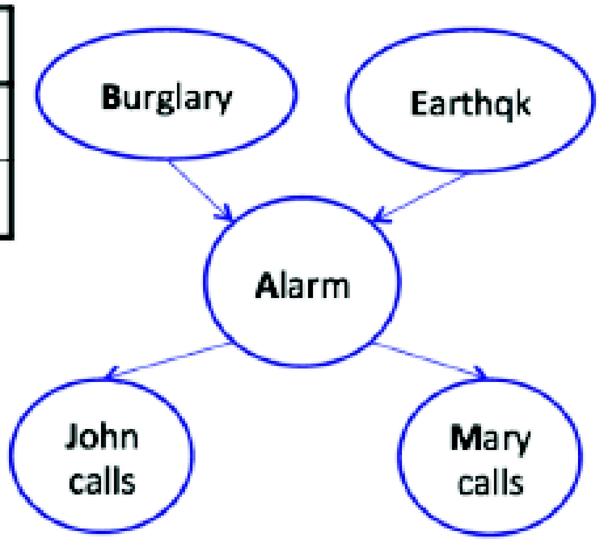
4) Normalizamos

Ejemplo – EV en Red Bayesiana – Robo

$$P(B|j,m)$$

Consulta

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

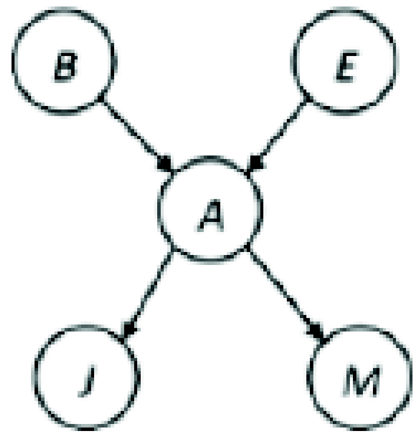
Ejemplo – EV en Red Bayesiana – Robo

Identificamos Variables

Consulta: B

Evidencia: j,m

Ocultas: E, A



Identificamos involucradas en la red segun variables

$$P(B|j,m) \propto P(B,j,m)$$

$P(B)$	$P(E)$	$P(A\|B,E)$	$P(j\|A)$	$P(m\|A)$

Ejemplo – EV en Red Bayesiana – Robo

Buscamos eliminar variables ocultas, empezando con A

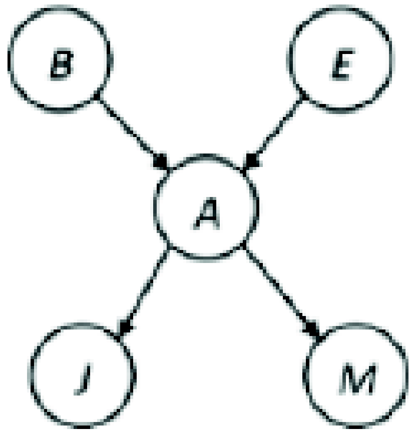
A

Join **Eliminacion**

$$\begin{matrix} P(A|B,E) \\ P(j|A) \\ P(m|A) \end{matrix} \xrightarrow{\times} P(j,m,A|B,E) \xrightarrow{\sum} P(j,m|B,E)$$

entonces

$$\boxed{P(B) \qquad P(E) \qquad P(j,m|B,E)}$$



Ejemplo – EV en Red Bayesiana – Robo

Continuamos eliminando E, que es una variable oculta

despues de remover A

$P(B) \qquad P(E) \qquad P(j, m|B, E)$

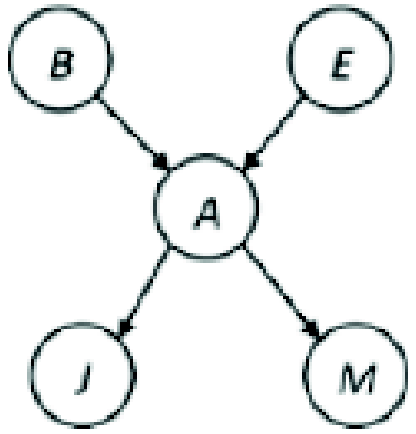
E

Join **Eliminacion**

$P(E)$

$P(j, m|B, E)$

$\xrightarrow{\times} P(j, m, E|B) \xrightarrow{\Sigma} P(j, m|B)$



despues de remover E

$P(B) \qquad P(j, m|B)$

Ejemplo – EV en Red Bayesiana – Robo

despues de remover E

$P(B)$ $P(j, m|B)$

B

Join

$P(B)$
$P(j, m|B)$

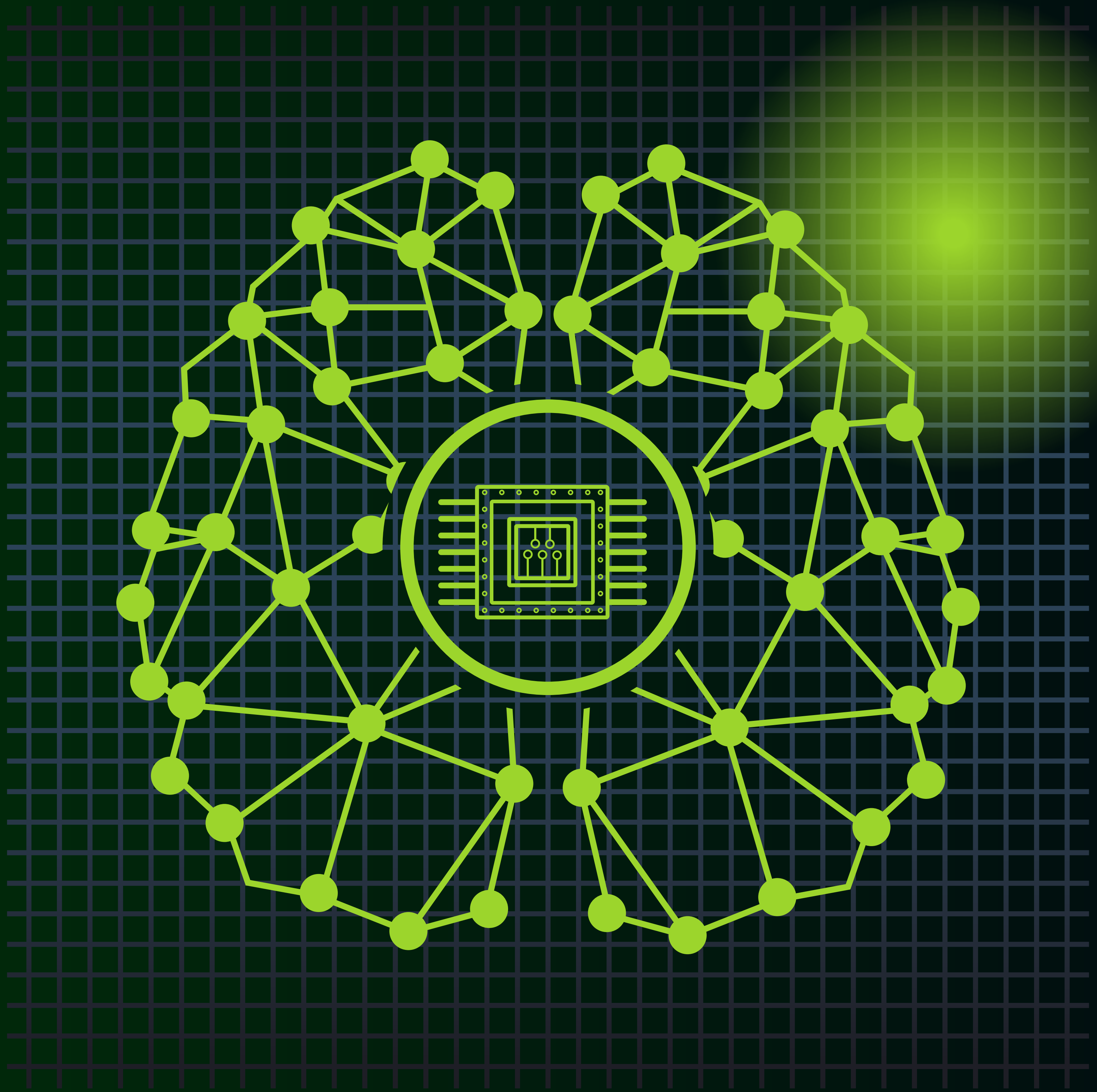
$\xrightarrow{\times} P(j, m, B) \xrightarrow{\text{Norm}} P(B|j, m)$

Ya tenemos las variables, solo queda hacer un ultimo Join y normalizar

Tipos de Inferencia en Redes Bayesianas

- **Inferencia Exacta**
  - Propósito y Función
    - Es el cálculo preciso de la probabilidad de una variable de interés dada un conjunto de evidencias.
    - Se basa en la teoría de la probabilidad y utiliza técnicas específicas para obtener resultados sin aproximaciones.
  - Métodos de Inferencia Exacta
    - Enumeración Completa: Se calcula la probabilidad marginalizando sobre todas las combinaciones de valores de las variables ocultas. Es efectivo en redes pequeñas, pero su complejidad crece exponencialmente con el número de variables.
    - Eliminación de Variables (Variable Elimination): Aplica la regla de Bayes y elimina variables innecesarias de manera eficiente, optimizando el cálculo en comparación con la enumeración completa.
    - Propagación en Árboles de Juntas (Junction Tree Algorithm): Convierte la red en una estructura manejable que permite inferencias eficientes, siendo el método más adecuado para redes con estructura de árbol.
  - Cuándo Usar Inferencia Exacta
    - ✔ Cuando la red tiene pocas variables o una estructura sencilla, como en forma de árbol.
    - ✔ Cuando es necesario obtener resultados precisos sin recurrir a aproximaciones.
- **Inferencia Aproximada**
  - Propósito y Función
    - Se utiliza cuando la inferencia exacta es computacionalmente inviable, como en redes con miles de nodos.
    - Los métodos de aproximación permiten estimar la probabilidad de interés sin necesidad de cálculos exactos.
  - Métodos de Inferencia Aproximada
    - Muestreo de Monte Carlo: Genera muestras aleatorias de la distribución de la red y utiliza estadísticas para estimar probabilidades. Incluye técnicas como Muestreo de Gibbs y Muestreo por Importancia.
    - Métodos de Propagación de Creencias (Belief Propagation - BP): Aproxima las probabilidades mediante el envío de mensajes entre nodos. Es eficaz en redes con estructura de árbol, pero no es exacto en redes con ciclos.
    - Aproximaciones Variacionales: Se basa en el uso de una distribución más simple para aproximar la verdadera distribución de la red, siendo común en modelos de Machine Learning.
  - Cuándo Usar Inferencia Aproximada
    - ✔ Cuando la red bayesiana es grande y la inferencia exacta resulta demasiado costosa.
    - ✔ Cuando se necesitan respuestas rápidas y una estimación es suficiente.

IA Y MACHINE LEARNING

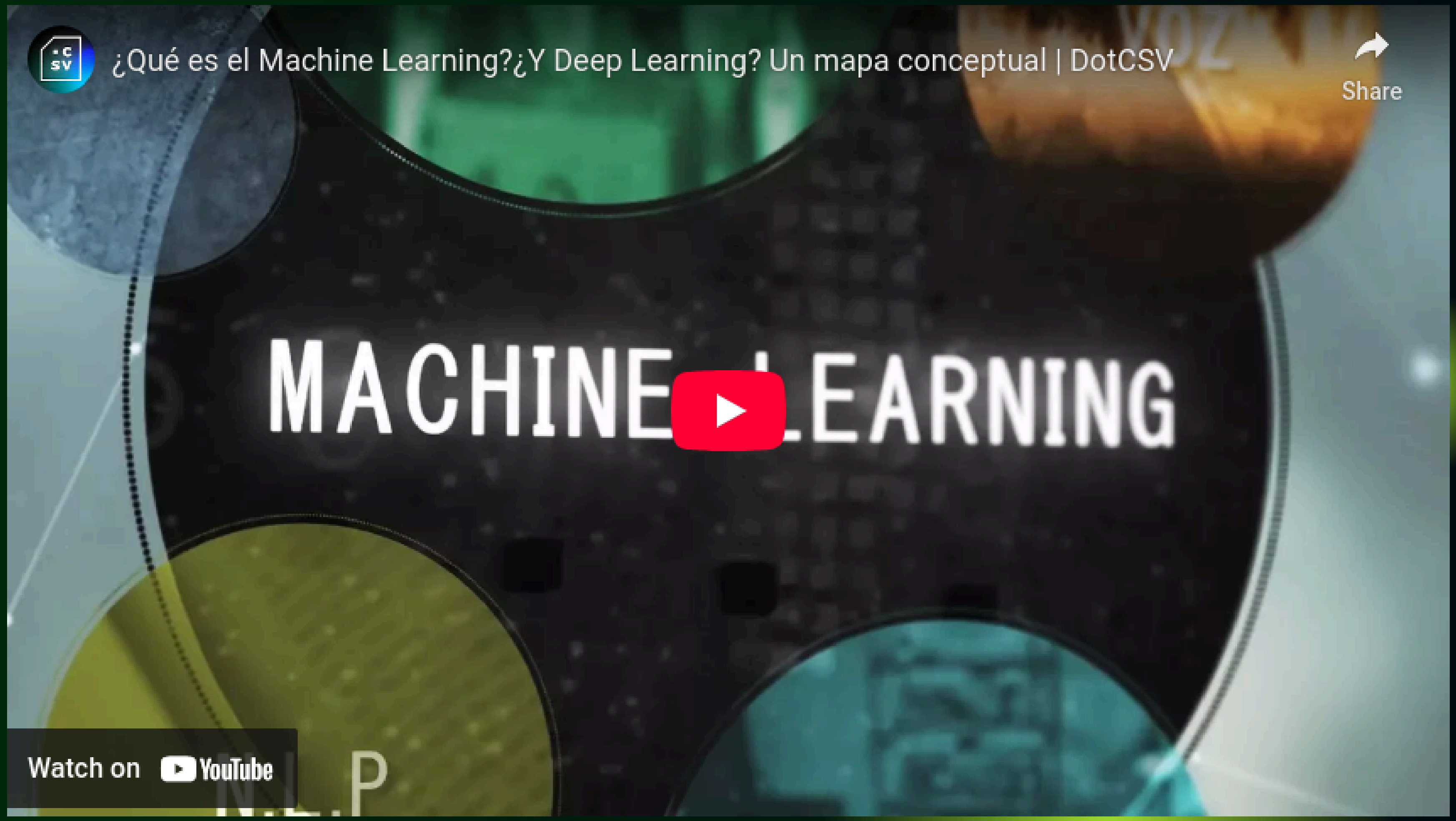


CONTENIDO

01 Aprendizaje Automático

02 Conjuntos de datos

03 Aplicaciones



INTRODUCCIÓN AL APRENDIZAJE AUTOMÁTICO

El aprendizaje automático (Machine Learning) es una rama de la inteligencia artificial que desarrolla algoritmos capaces de aprender automáticamente a partir de datos. Su objetivo es generalizar comportamientos o patrones para realizar predicciones o clasificaciones

01 Aprendizaje supervisado

Utiliza datos etiquetados (con entradas y salidas conocidas) para entrenar modelos.

02 Aprendizaje no supervisado

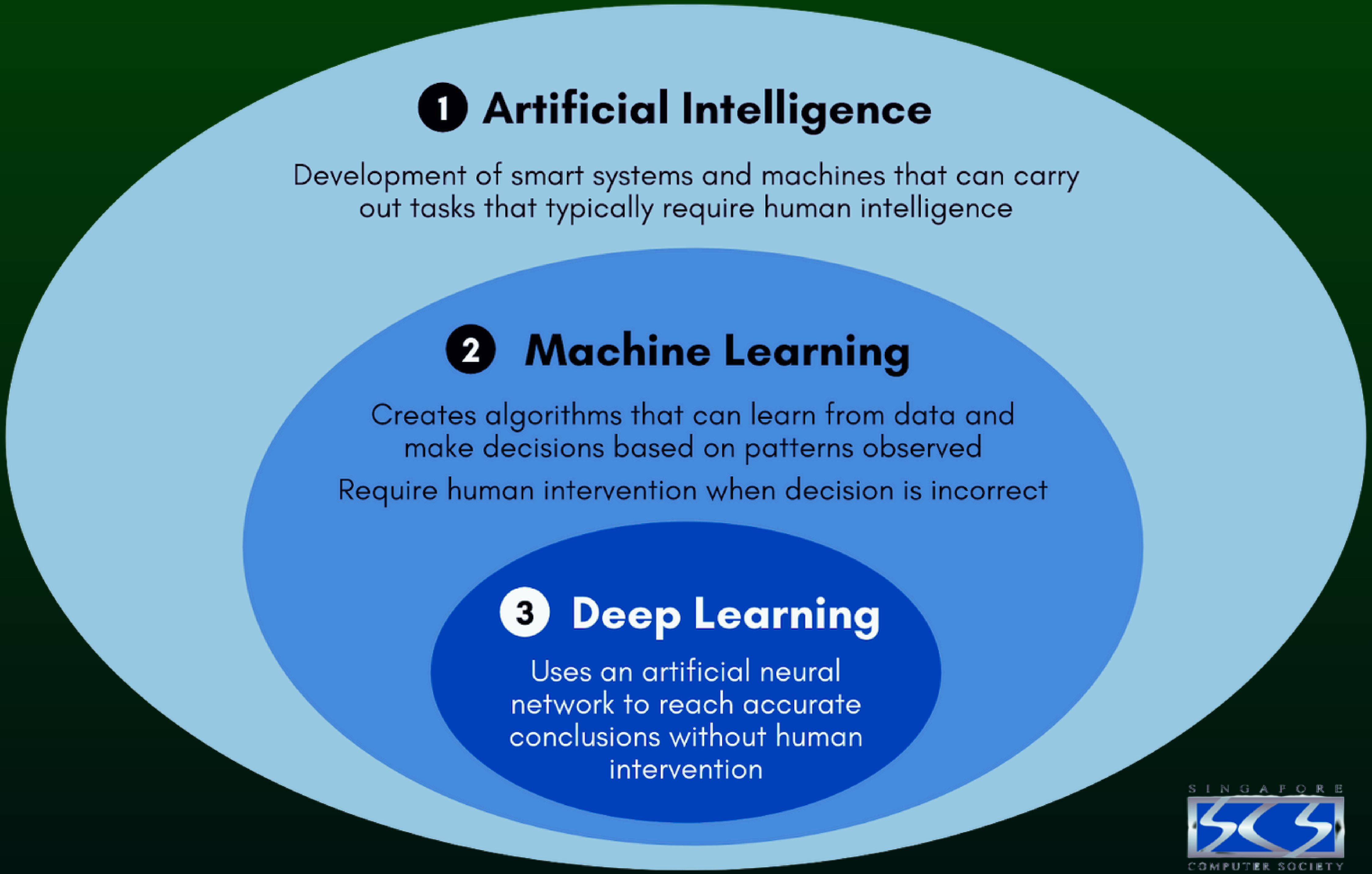
Trabaja con datos no etiquetados, buscando patrones ocultos.

03 Aprendizaje por refuerzo

Enseña a un agente a tomar decisiones mediante prueba y error, maximizando recompensas acumuladas

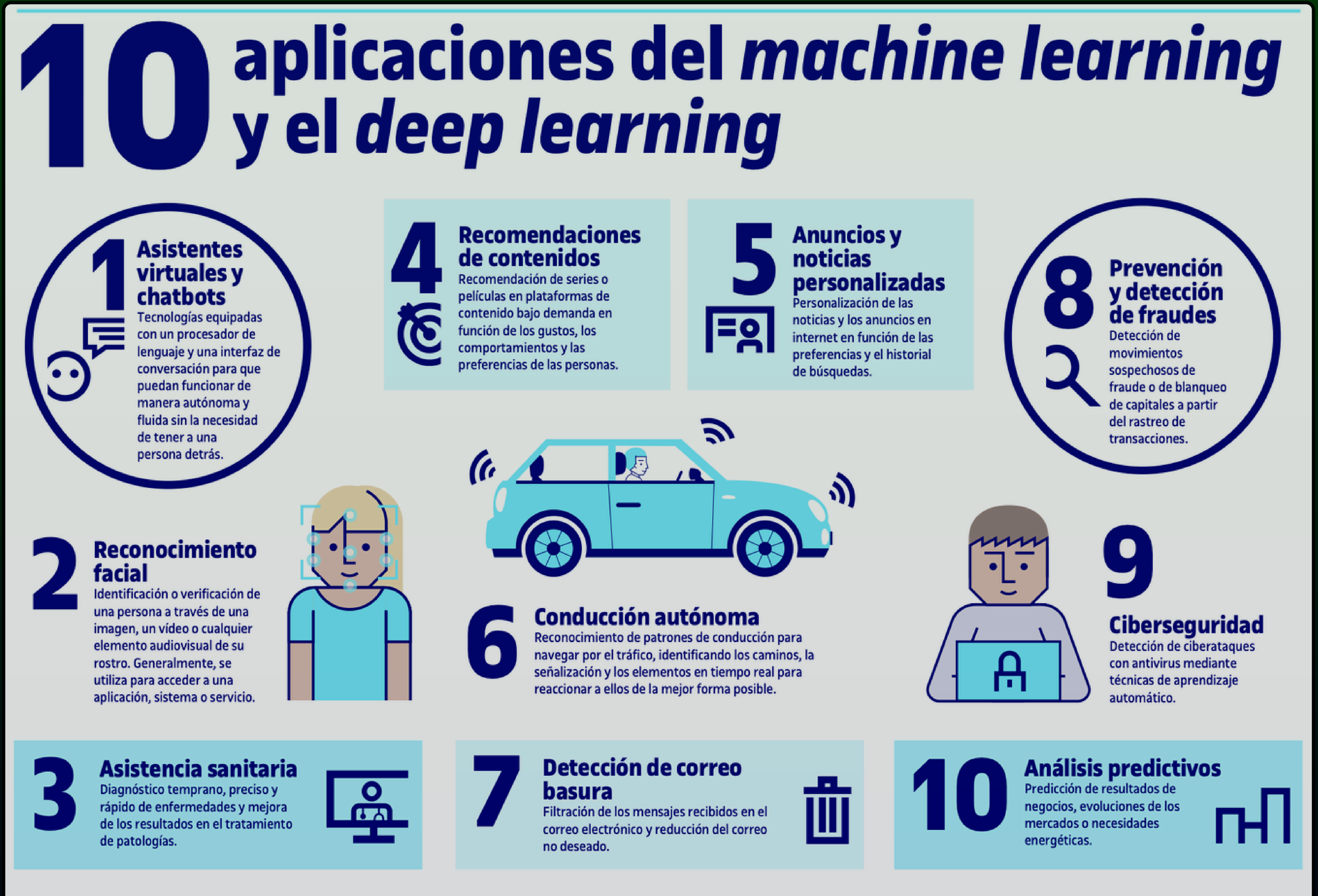
04 Deep Learning

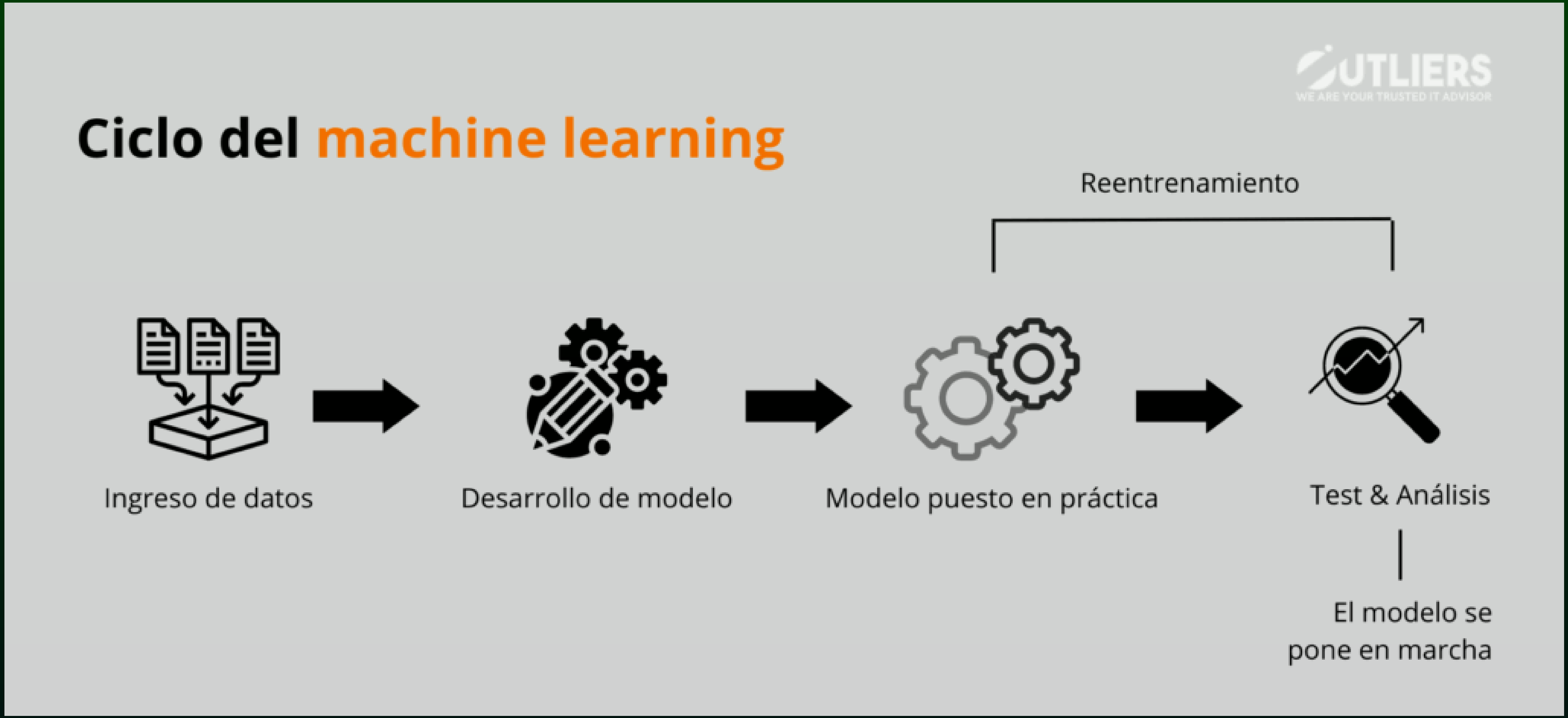
Los algoritmos emplean redes neuronales artificiales con múltiples capas, lo que permite el reconocimiento de patrones complejos y la toma de decisiones.

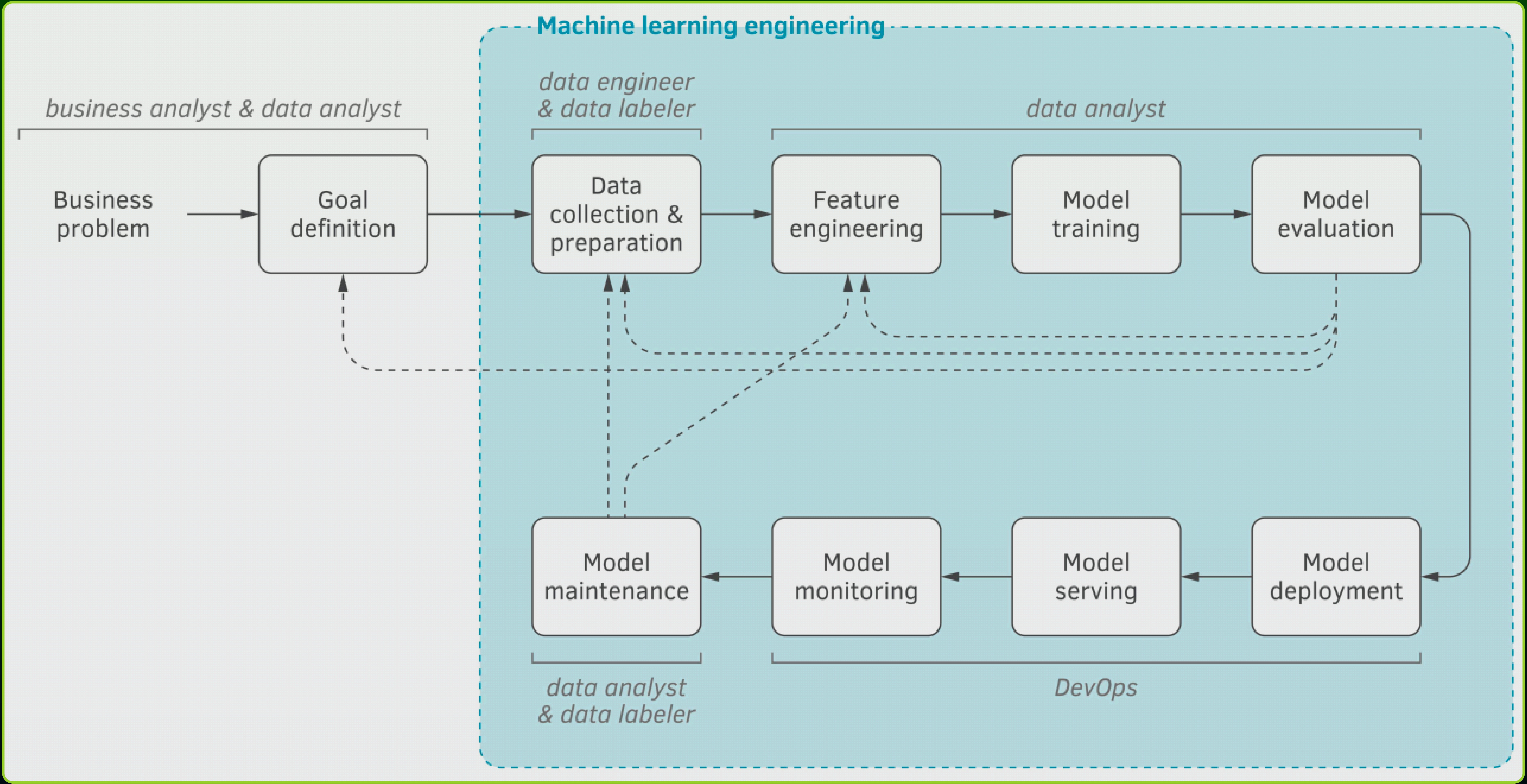


SINGAPORE COMPUTER SOCIETY

¿QUÉ SISTEMAS CONOCEN QUE UTILIZAN MACHINE LEARNING?




Ciclo del machine learning
Reentrenamiento
Ingreso de datos
Desarrollo de modelo
Modelo puesto en práctica
Test & Análisis
El modelo se pone en marcha


Machine learning engineering
business analyst & data analyst
data engineer
& data labeler
data analyst
Business
problem
Goal
definition
Data
collection &
preparation
Feature
engineering
Model
training
Model
evaluation
Model
maintenance
Model
monitoring
Model
serving
Model
deployment
data analyst
& data labeler
DevOps

APRENDIZAJE SUPERVISADO

El aprendizaje supervisado utiliza datos etiquetados para construir modelos que predicen resultados futuros. Se basa en una función *f(x)=y*, donde *x* son las variables independientes y *y* la respuesta

01 Regresión

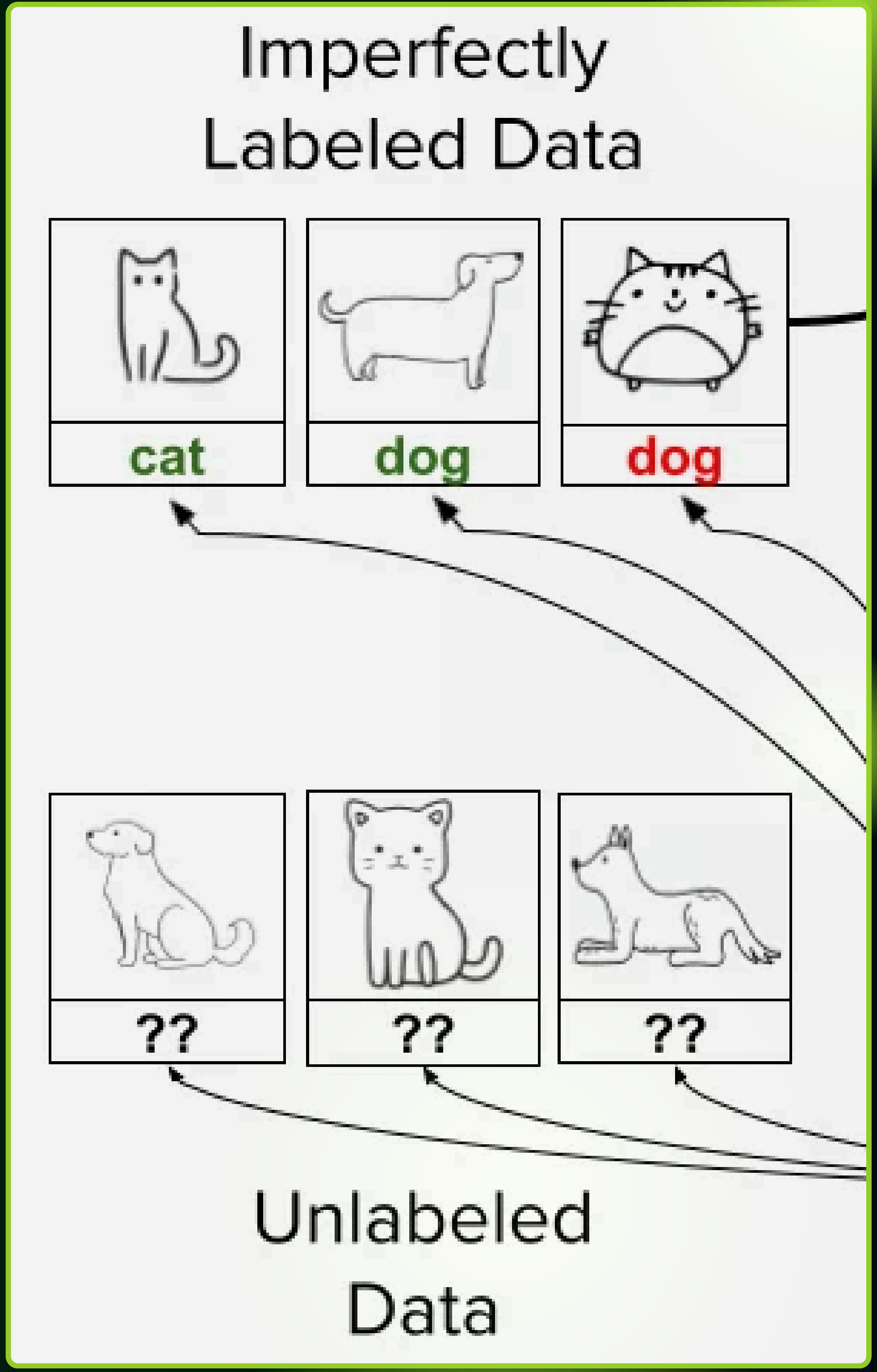
- Predicen valores numéricos continuos.
- Ejemplo: Estimar el precio de una casa según sus características (tamaño, ubicación).
- Representación gráfica: Relación entre variables independientes (x) y dependientes (y)

02 Clasificación

- Asignan categorías o clases a las observaciones.
- Ejemplo: Clasificar correos como spam/no spam o diagnosticar enfermedades.
- Tipos:
  - Binaria: Dos clases (spam/no spam).
  - Multiclase: Más de dos clases (diagnóstico médico).

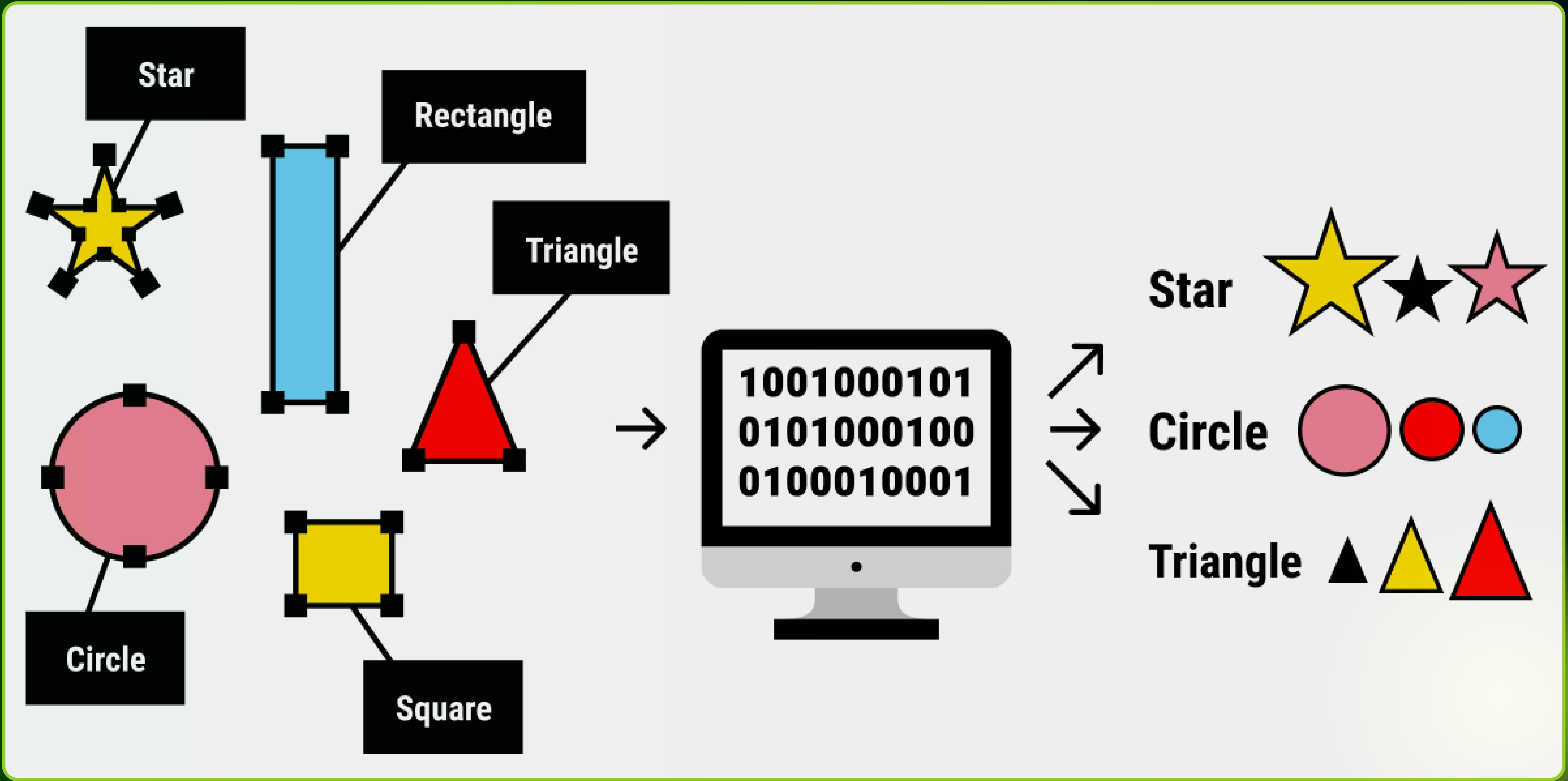
03 Métodos de ensamble

Son técnicas de aprendizaje automático que combinan múltiples modelos individuales para mejorar la precisión de las predicciones, en lugar de depender de un único modelo.



APRENDIZAJE SUPERVISADO

El aprendizaje supervisado utiliza datos etiquetados para construir modelos que predicen resultados futuros. Se basa en una función *f(X)=y*, donde *X* son las variables independientes y *y* la respuesta



CONJUNTO DE DATOS: DIAGNÓSTICO MÉDICO SIMPLIFICADO

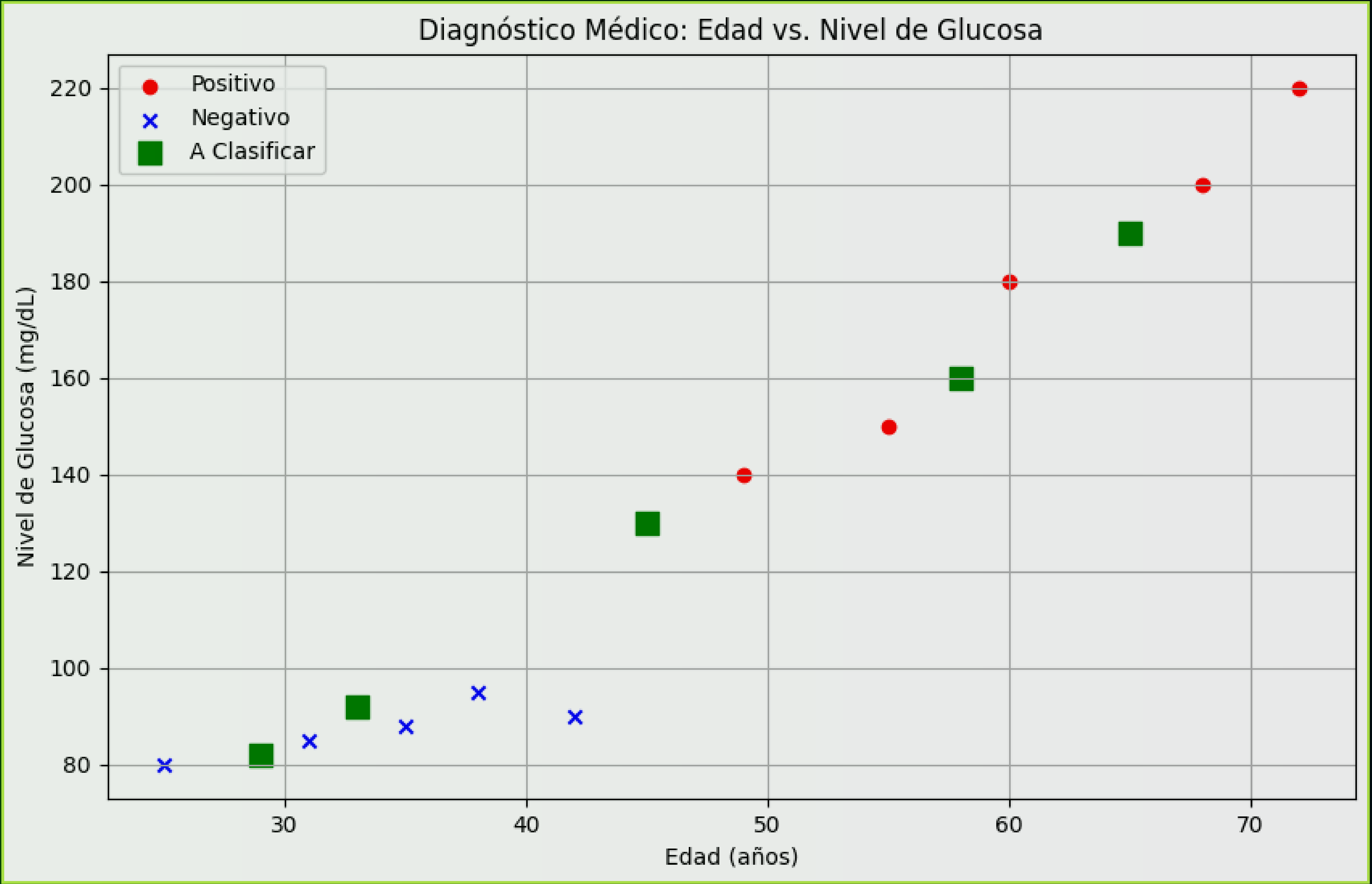
Este conjunto de datos simula características de pacientes con diagnóstico positivo o negativo para una enfermedad hipotética.

ID	Edad (años)	Presión Arterial (mmHg)	Nivel de Glucosa (mg/dL)	Diagnóstico
1	55	130	150	Positivo
2	42	120	90	Negativo
3	68	145	200	Positivo
4	31	115	85	Negativo
5	60	135	180	Positivo
6	38	118	95	Negativo
7	72	150	220	Positivo
8	25	110	80	Negativo
9	49	128	140	Positivo
10	35	116	88	Negativo

NUEVAS OBSERVACIONES PARA CLASIFICAR

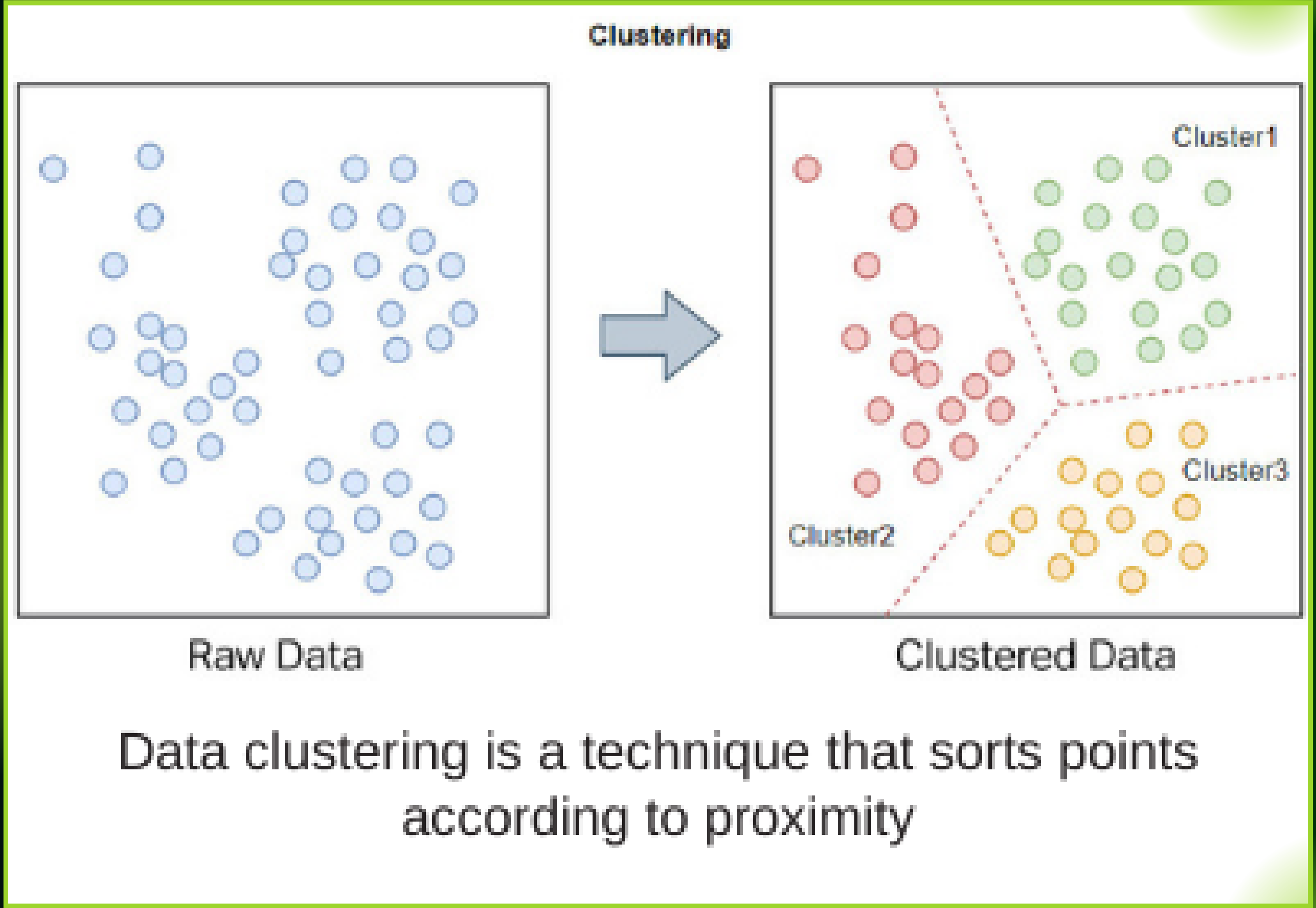
Basándose en los datos proporcionados anteriormente, intenten clasificarlas visualmente como "Positivo" o "Negativo":

ID	Edad (años)	Presión Arterial (mmHg)	Nivel de Glucosa (mg/dL)	Diagnóstico
11	58	132	160	?
12	29	112	82	?
13	65	140	190	?
14	45	125	130	?
15	33	117	92	?


Diagnóstico Médico: Edad vs. Nivel de Glucosa
Positivo
Negativo
A Clasificar
Nivel de Glucosa (mg/dL)
220
200
180
160
140
120
100
80
30
40
50
60
70
Edad (años)

APRENDIZAJE NO SUPERVISADO

Los algoritmos de aprendizaje no supervisado están diseñados para descubrir patrones y estructuras ocultos en datos no etiquetados.



01 Clustering

- Organiza datos similares en grupos.
- Identifica estructuras ocultas en los datos
- **Ejemplo:** Segmentación de clientes según comportamiento.

02 Detección de anomalías

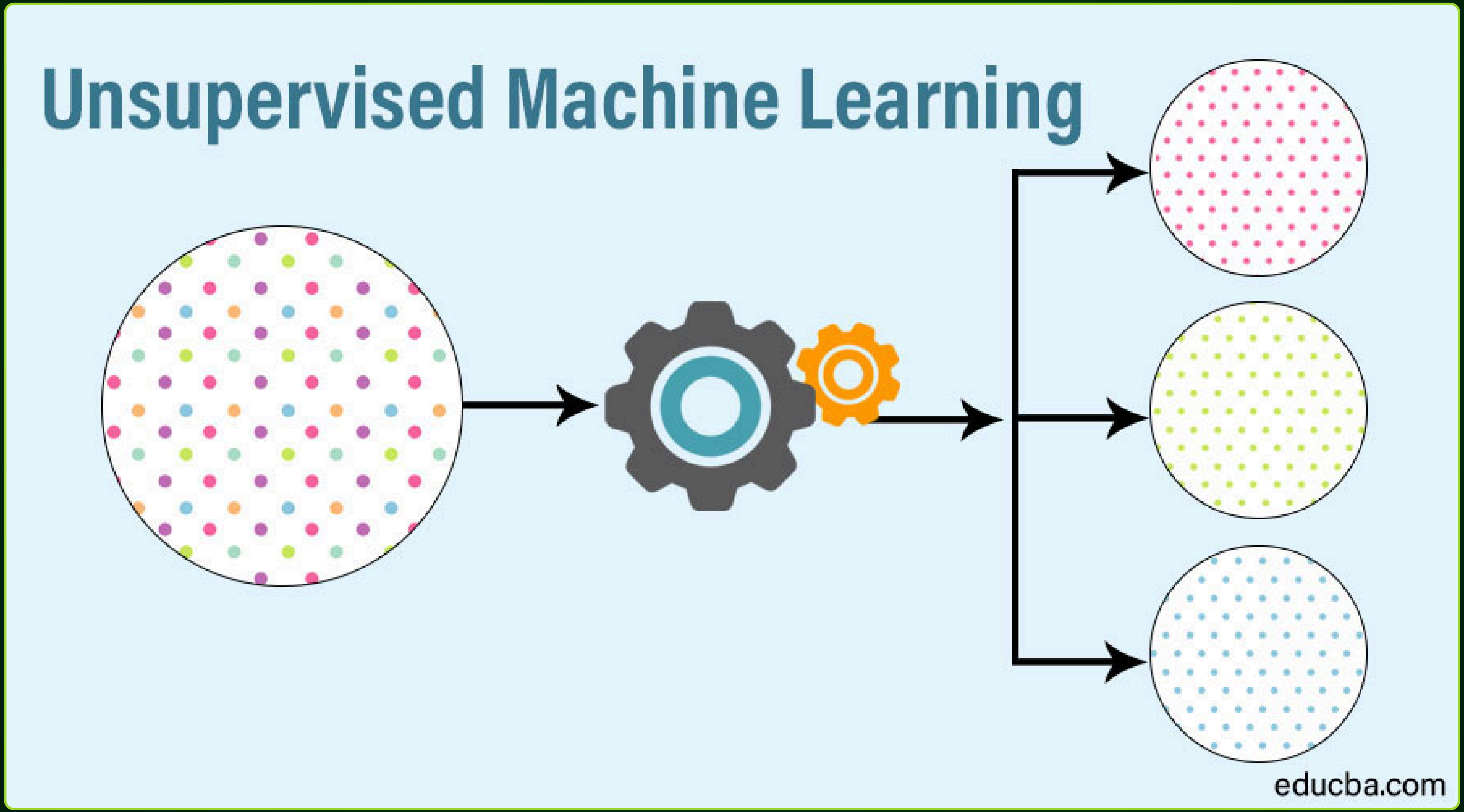
- Identifica observaciones significativamente diferentes al resto.
- **Ejemplo:** Detección de fraudes financieros o intrusos en sistemas informáticos.

03 Reducción de dimensionalidad

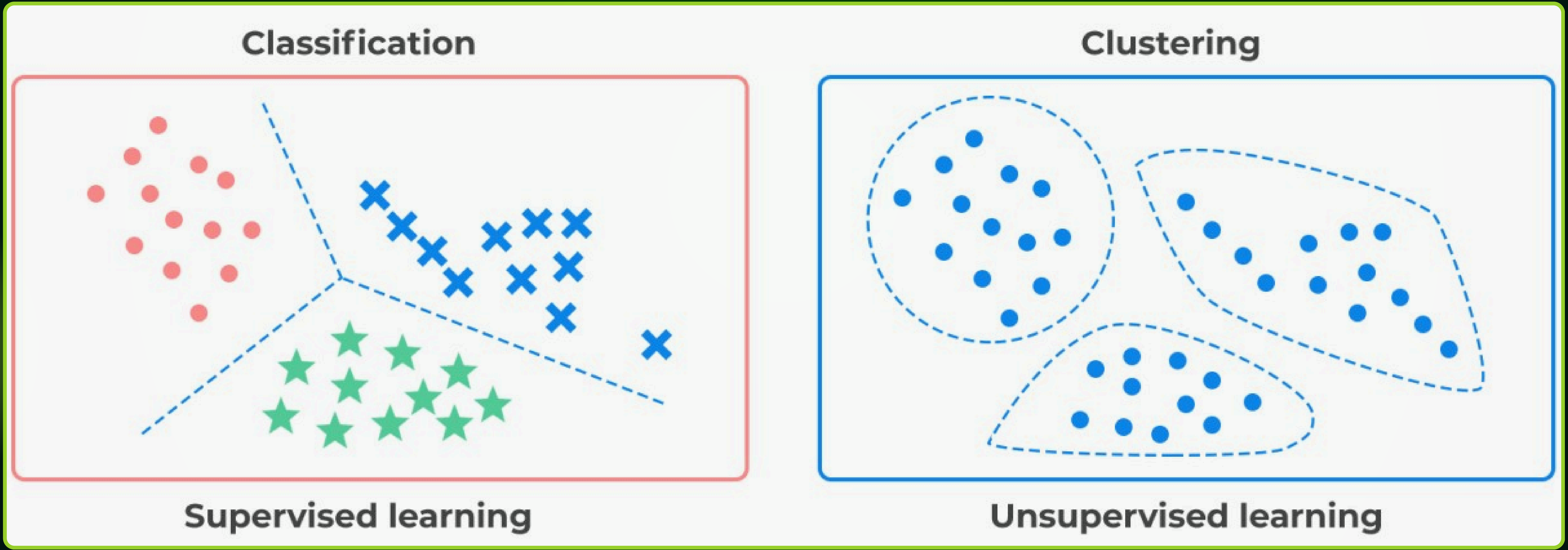
- Simplificación de datos de alta dimensión mediante la reducción del número de características, preservando al mismo tiempo la información esencial.
- **Ejemplo:** Encontrar las características del dataset del Titanic más relevantes para predecir la supervivencia.

APRENDIZAJE NO SUPERVISADO

Se utilizan a menudo para la exploración de datos, la reducción de dimensionalidad y la detección de anomalías.



APRENDIZAJE SUPERVISADO VS NO SUPERVISADO



APRENDIZAJE POR REFUERZO

El aprendizaje por refuerzo enseña a un agente a interactuar con un entorno tomando decisiones mediante prueba y error, interactuando con el entorno y recibiendo recompensas por las acciones deseadas. Se utiliza ampliamente en videojuegos, robótica y sistemas autónomos, permitiendo a los agentes adaptar y optimizar su comportamiento con el tiempo.

01 Función de Recompensa

La función de recompensa es un mecanismo que asigna un valor numérico a cada acción realizada por un agente en un entorno. Este valor refleja qué tan buena o mala fue la acción en relación con el objetivo final.

02 Política

La política define la estrategia del agente para seleccionar acciones en función del estado actual del entorno.

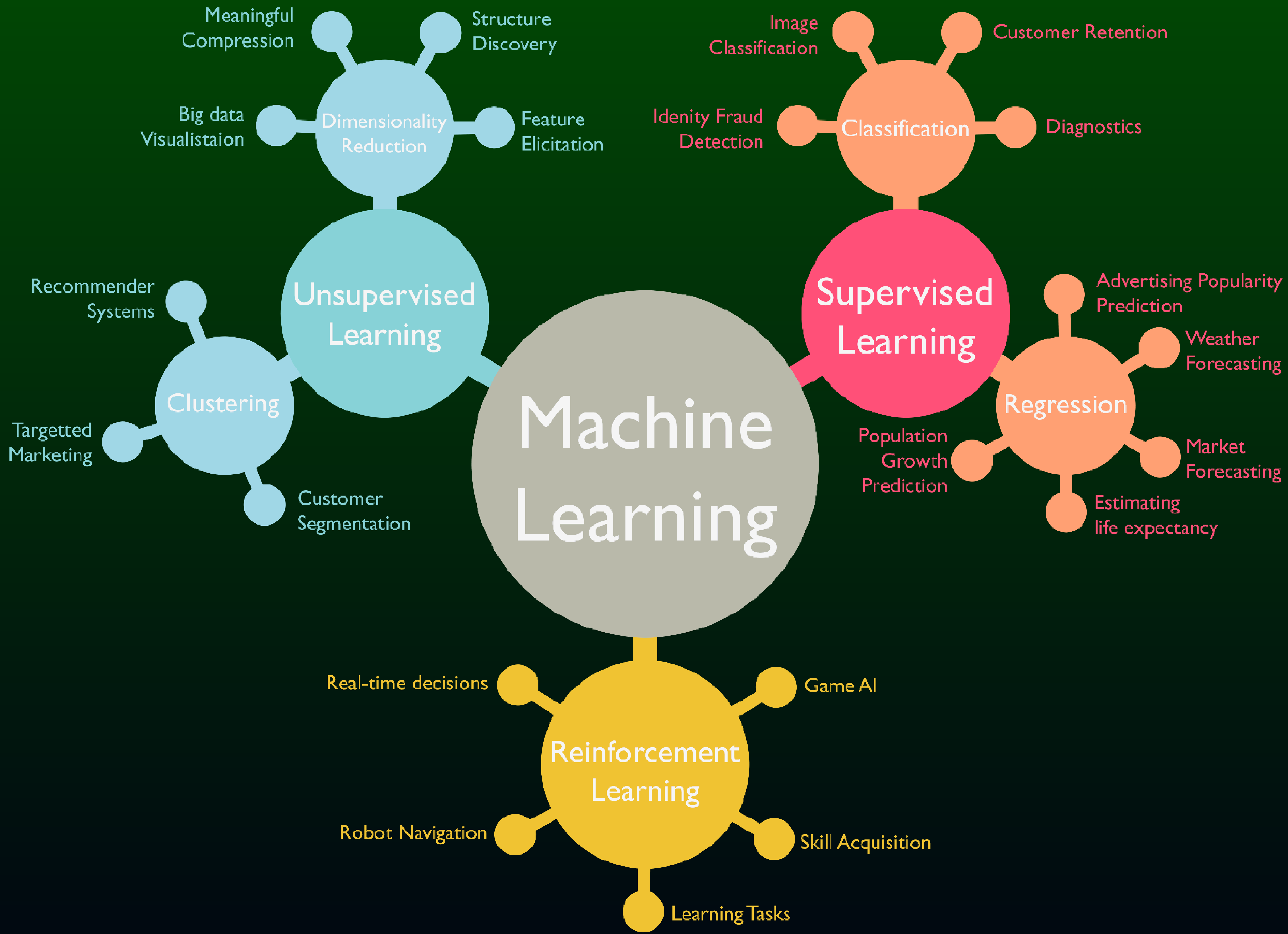
Ejemplo: En un robot que navega por un laberinto, la política podría ser "si hay una pared a la derecha, gira a la izquierda".

03 Exploración

La exploración implica que el agente pruebe nuevas acciones para descubrir más información sobre el entorno.

04 Función de Valor

La función de valor estima cuán beneficioso es estar en un estado particular o realizar una acción específica, considerando las recompensas futuras esperadas.


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

DEEP LEARNING

El aprendizaje profundo es una rama del aprendizaje automático que utiliza redes neuronales artificiales de múltiples capas para aprender patrones complejos a partir de datos. Estas redes se inspiran en la estructura y función del cerebro humano, lo que permite capacidades avanzadas en áreas como el reconocimiento de imágenes y el procesamiento del lenguaje natural.

01 ESTRUCTURA DE LA RED NEURONAL

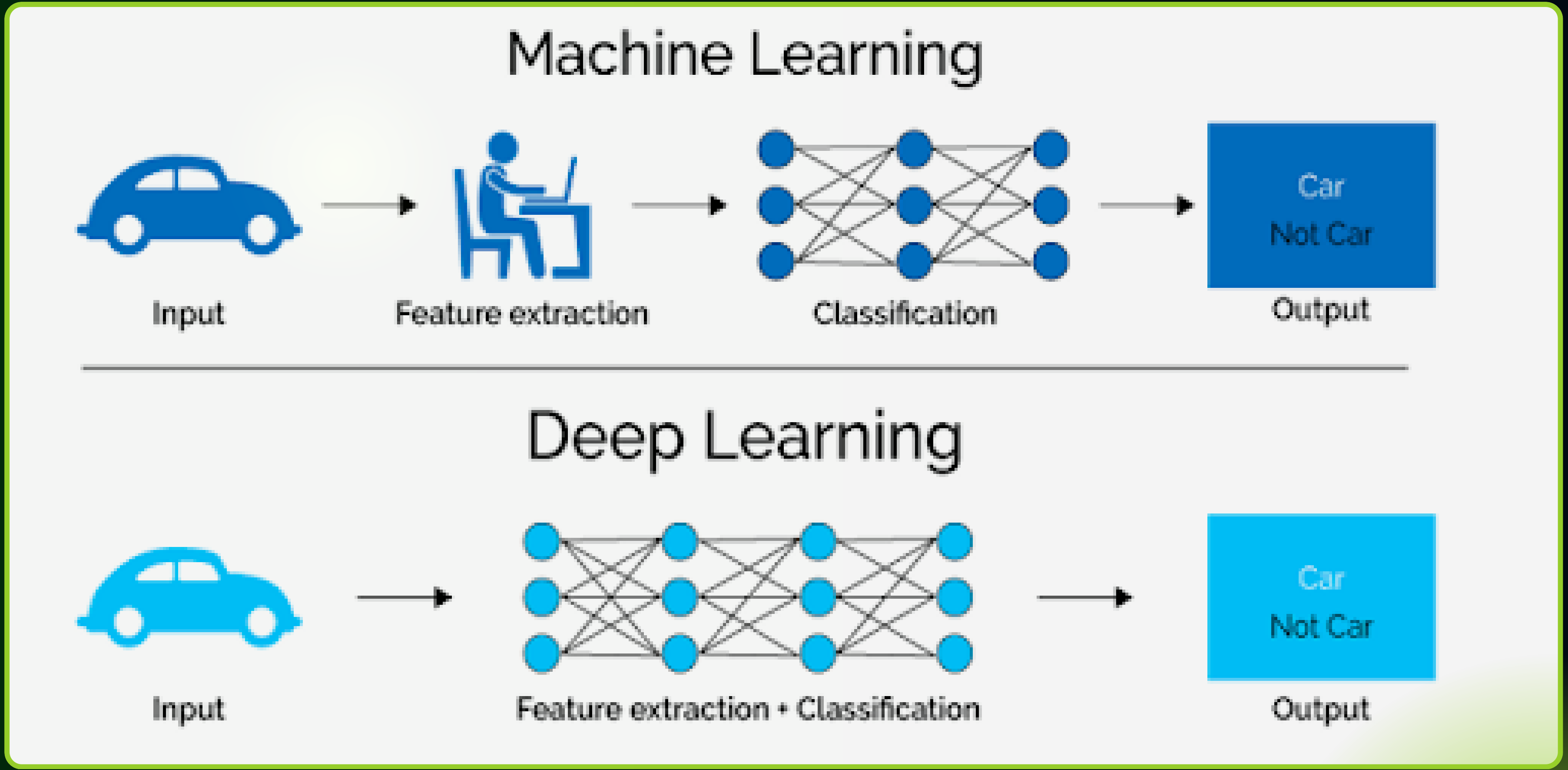
Consiste en nodos interconectados organizados en capas, y cada nodo realiza un cálculo simple.

02 FORWARD PROPAGATION

Los datos de entrada fluyen a través de la red, activando nodos y produciendo una salida.

03 BACKPROPAGATION

Ajustar los pesos de la red para minimizar la diferencia entre los resultados previstos y los reales.



CONJUNTOS DE DATOS Y VALIDACIÓN

Los conjuntos de datos son la base del aprendizaje automático, ya que proporcionan los ejemplos necesarios para entrenar, validar y probar los modelos. Se dividen en tres categorías principales:

01 Conjunto de Entrenamiento

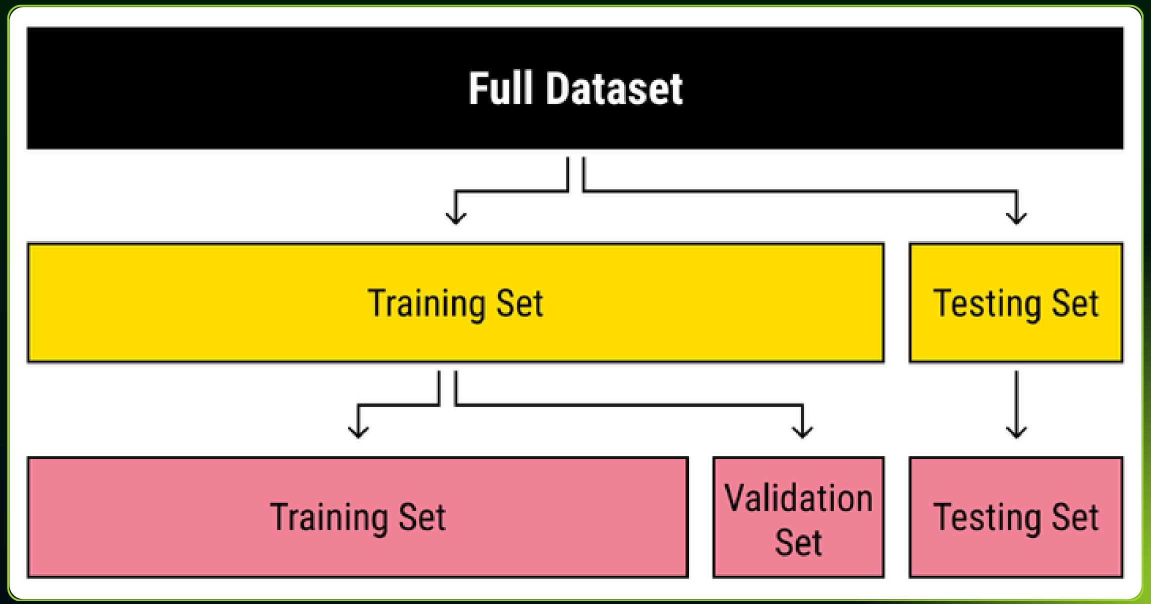
Es el subconjunto de datos utilizado para construir el modelo. Aquí el algoritmo aprende patrones y relaciones entre las variables independientes (x) y la variable objetivo (y).

02 Conjunto de Prueba (Test Set)

Es un subconjunto independiente que no se utiliza durante el entrenamiento del modelo.
Propósito: Evaluar la capacidad del modelo para generalizar a datos no vistos.

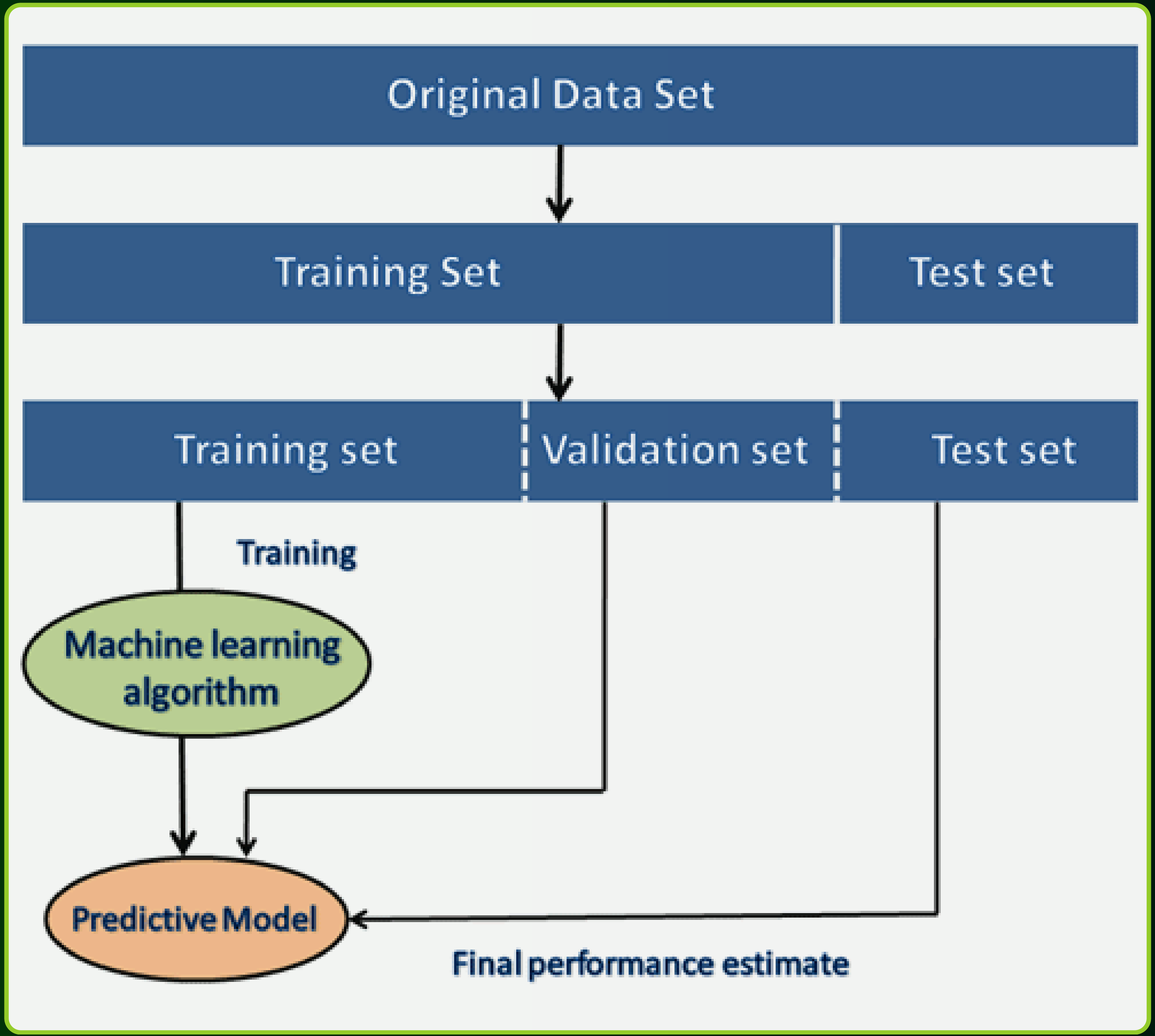
03 Conjunto de Validación

Es un subconjunto adicional utilizado durante el entrenamiento para ajustar hiperparámetros (por ejemplo, la profundidad de un árbol de decisión o la tasa de aprendizaje en redes neuronales). **Ejemplo:** Si se entrenan varios modelos con diferentes configuraciones, el conjunto de validación ayuda a elegir el más prometedor.



CONJUNTOS DE DATOS Y VALIDACIÓN

Conjuntos bien estructurados y técnicas como la validación cruzada son esenciales para construir modelos confiables que generalicen correctamente a nuevos datos.



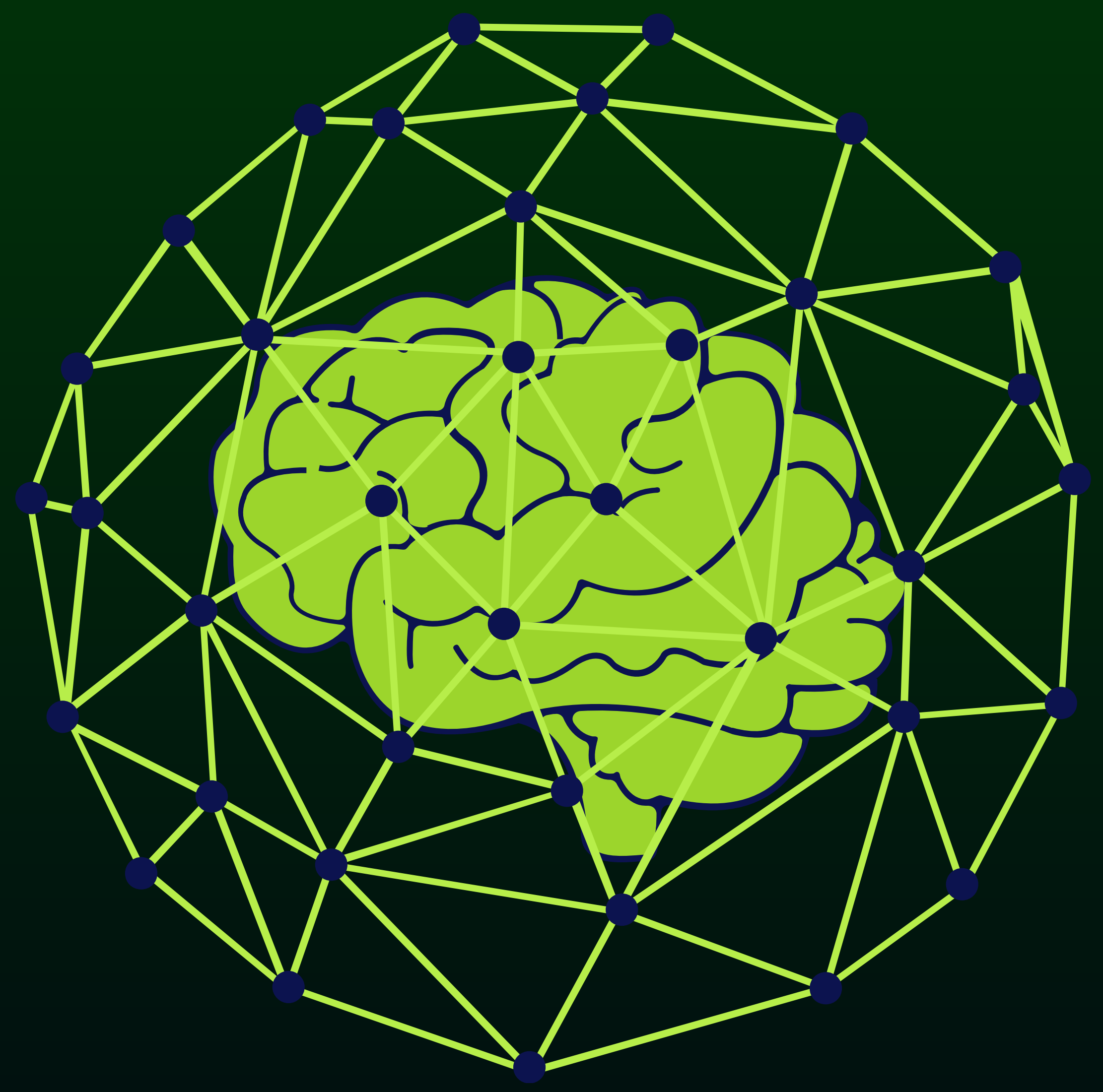
PROCESAMIENTO DEL LENGUAJE NATURAL

El procesamiento del lenguaje natural (PLN) permite a las computadoras comprender, interpretar y generar lenguaje humano. Utiliza técnicas de IA y ML para analizar texto y voz, lo que facilita tareas como la traducción automática, el análisis de sentimientos y el desarrollo de chatbots.

APLICACIONES DE TRADUCCIÓN AUTOMÁTICA	Ejemplos Traducir texto entre idiomas, como Google Translate
ANÁLISIS DE SENTIMIENTOS	Determinar el tono emocional del texto, como identificar reseñas positivas o negativas.
RESUMEN DE TEXTO	Resumir de forma concisa grandes cantidades de texto, como artículos de noticias o trabajos de investigación.
CHATBOTS	Desarrollar agentes conversacionales que interactúen con humanos, como chatbots de atención al cliente.

VISIÓN POR COMPUTADORA Y RECONOCIMIENTO DE IMÁGENES

La visión artificial permite a las computadoras "ver" e interpretar imágenes y vídeos. Implica el uso de algoritmos de IA y aprendizaje automático para analizar información visual, lo que facilita tareas como el reconocimiento de objetos, la clasificación de imágenes y el análisis de vídeos.



Reconocimiento de objetos

Identificar objetos dentro de imágenes, como automóviles, personas y animales.

Reconocimiento facial

Identificación de personas según sus rasgos faciales, utilizada con fines de seguridad y autenticación.

Clasificación de imágenes

Categorizar imágenes en diferentes clases, como identificar diferentes tipos de animales o plantas.

Análisis de vídeo

Analizar secuencias de vídeo para identificar patrones, rastrear objetos y detectar anomalías.

APLICACIONES DE IA Y ML EN EL MUNDO REAL

La IA y el aprendizaje automático están revolucionando diversas industrias, impactando nuestras vidas de innumerables maneras. Estas tecnologías ofrecen soluciones a problemas complejos, mejorando la eficiencia, la productividad y la toma de decisiones en diferentes sectores.

VEHÍCULOS AUTÓNOMOS

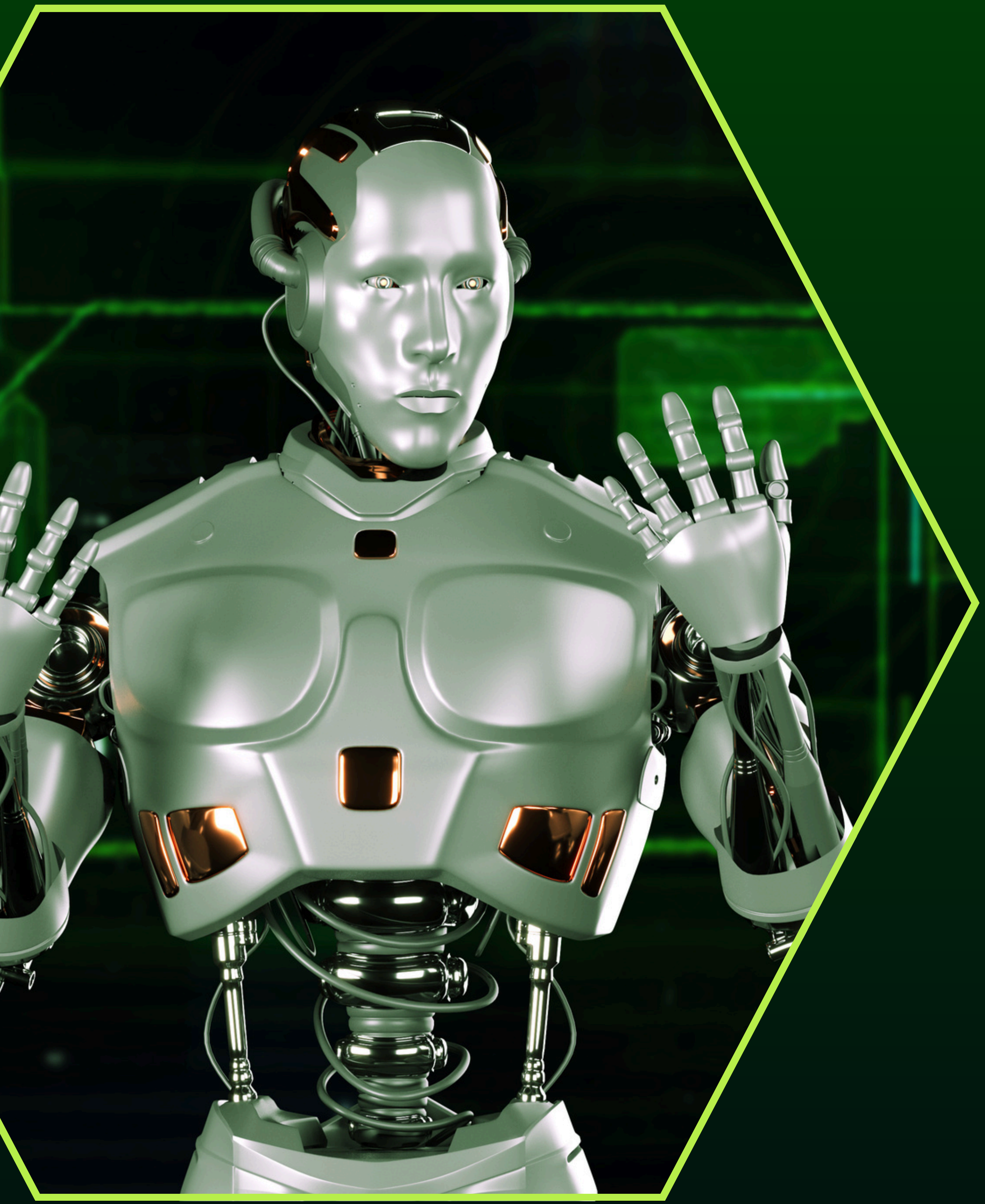
La IA impulsa los coches autónomos, permitiéndoles percibir su entorno, tomar decisiones y navegar con seguridad.

HEALTHCARE

La IA ayuda en el diagnóstico de enfermedades, el descubrimiento de fármacos y los planes de tratamiento personalizados, mejorando la atención y los resultados de los pacientes.

FINANZAS

La IA mejora el análisis financiero, la gestión de riesgos y la detección de fraudes, optimizando las decisiones de inversión y minimizando las pérdidas.

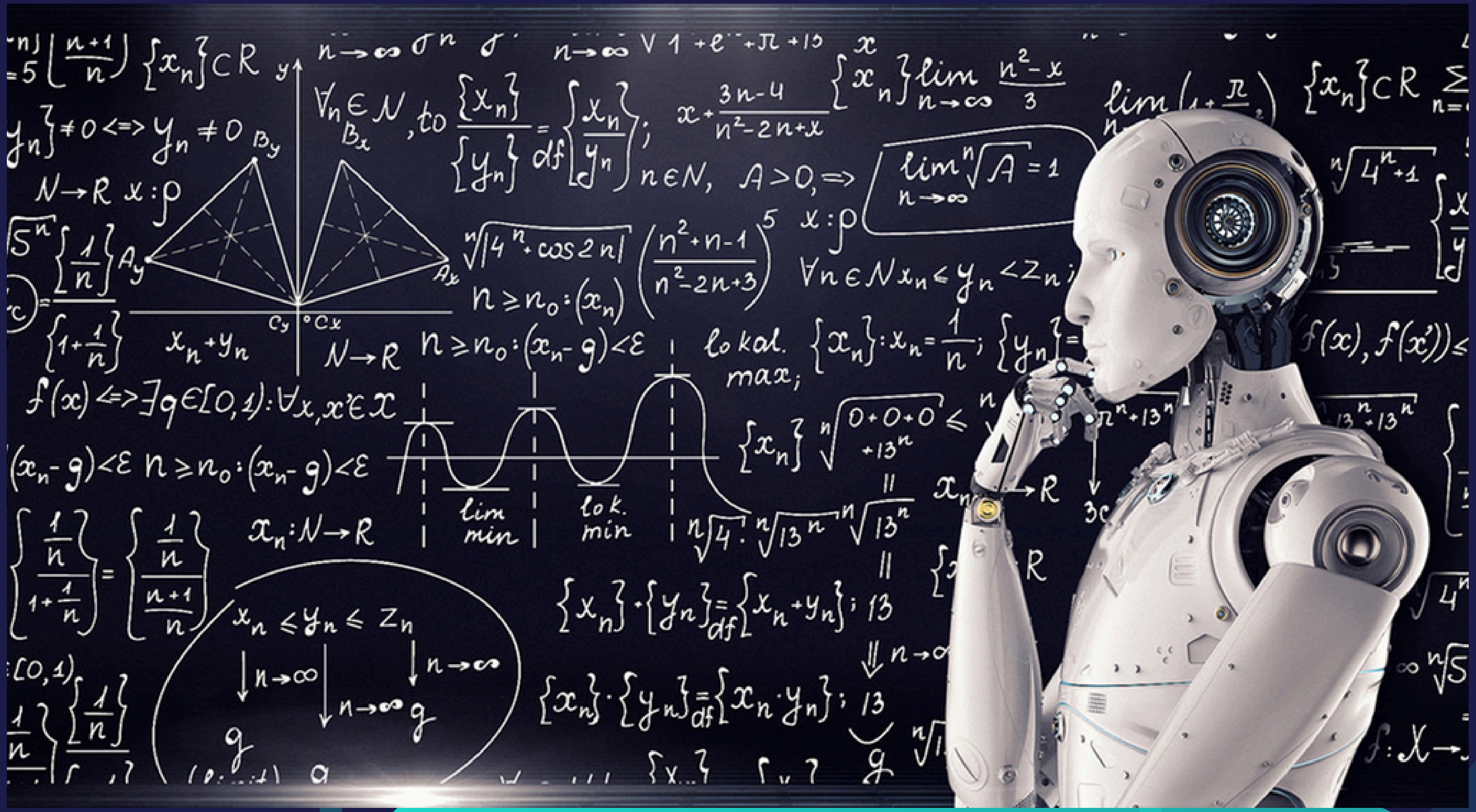


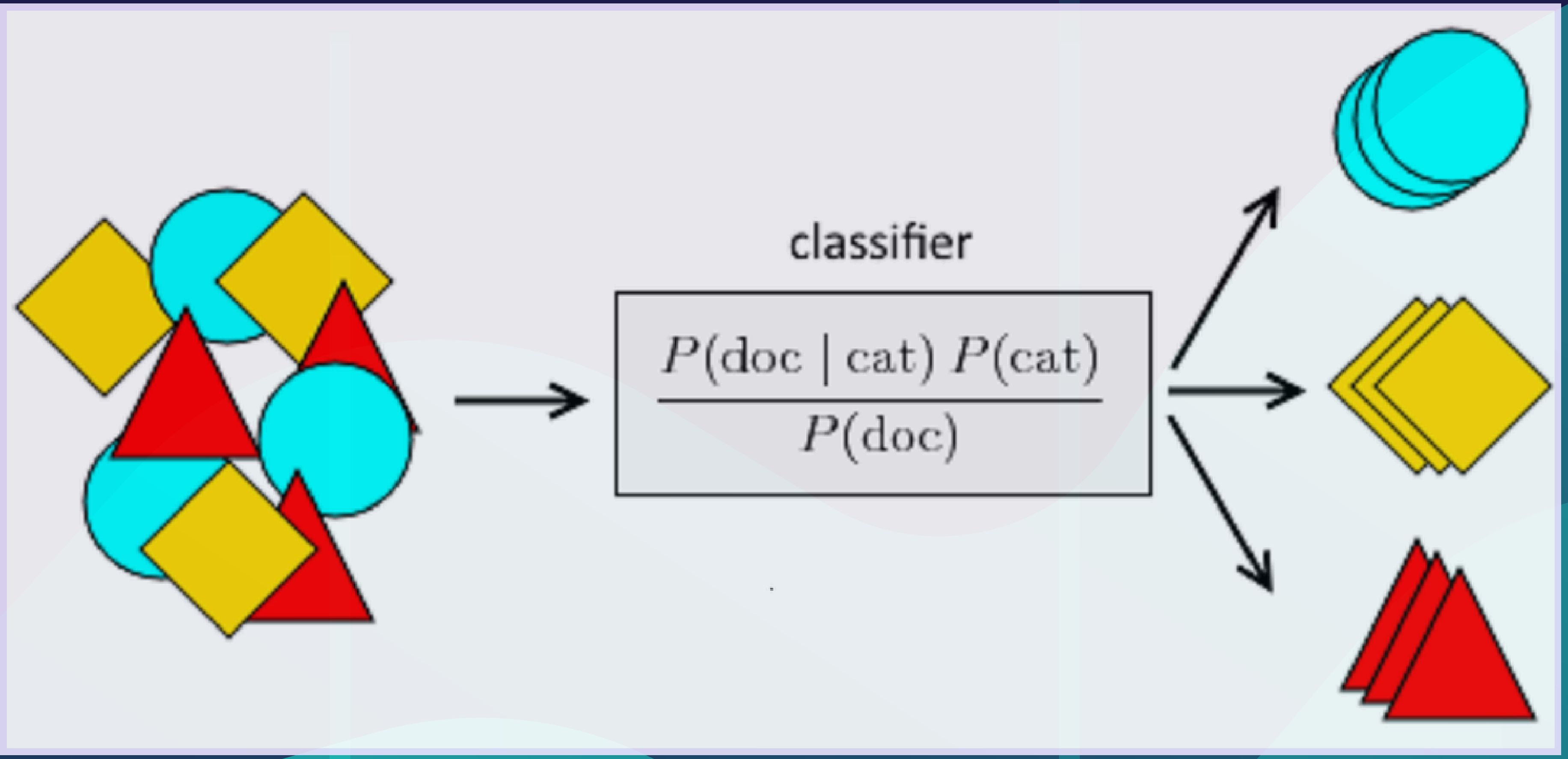
EL FUTURO DE LA IA Y EL APRENDIZAJE AUTOMÁTICO

El futuro de la IA y el aprendizaje automático es sumamente prometedor, con avances continuos previstos en áreas como el procesamiento del lenguaje natural, la visión artificial y la robótica. A medida que estas tecnologías evolucionen, transformarán aún más nuestras vidas, dando lugar a nuevas innovaciones, mayor eficiencia y un mundo más interconectado.

MACHINE LEARNING

Naive Bayes





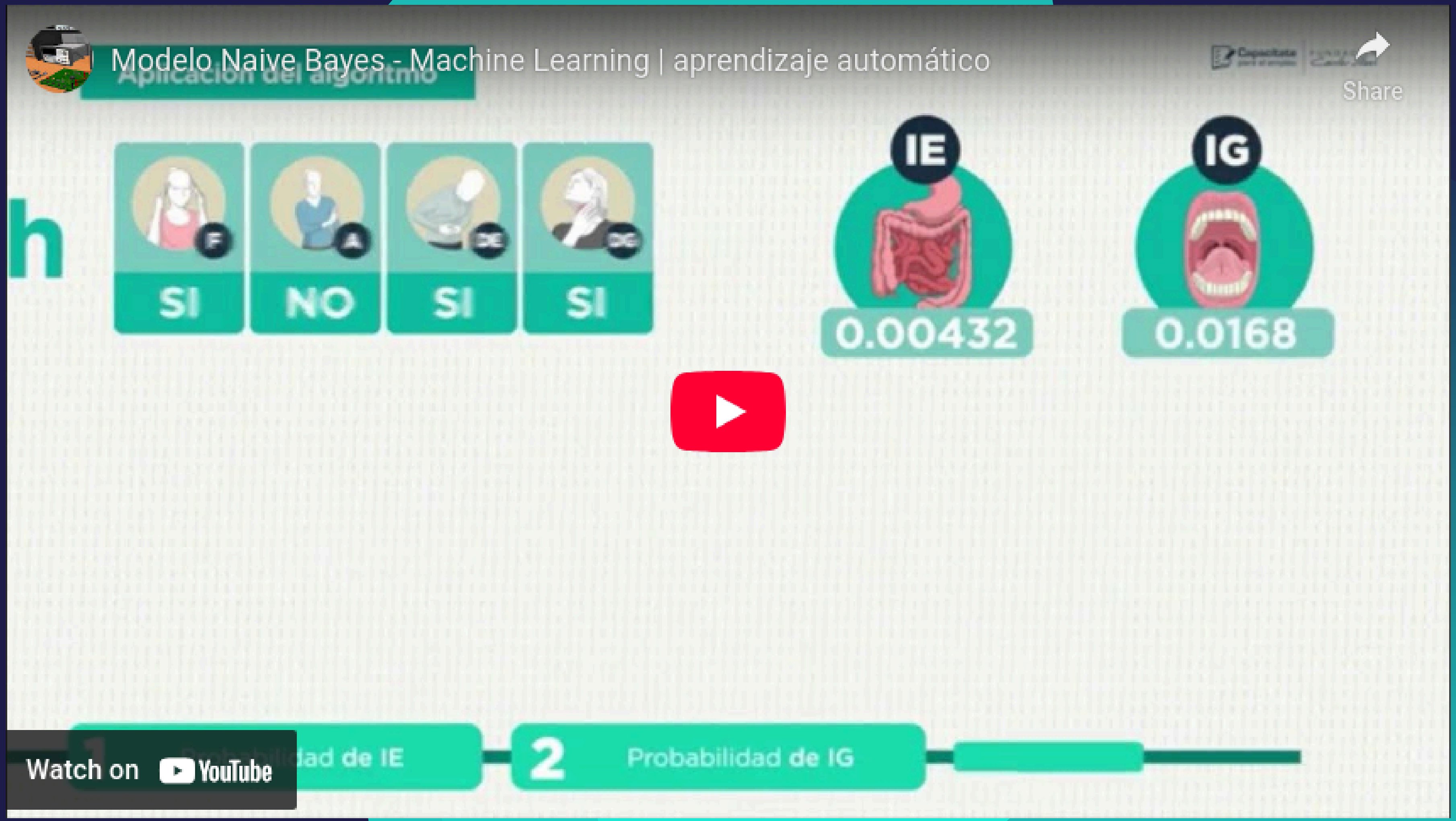
Agenda

- Naïve Bayes como Clasificador
- Clasificacion Basada en Modelos
- Ejemplos de clasificacion y aplicaciones
- Algortimo Naïve Bayes
- Maximum likelihood
- Over-fitting
- K-fold
- Laplace Smoothing
- Suma de Logaritmos

Naïve Bayes

Es una clase especial de algoritmo de clasificación de Aprendizaje Automatico, o Machine Learning

Proporcionan una manera fácil de construir modelos con un comportamiento muy bueno debido a su simplicidad



Naïve Bayes

El principio de parsimonia o navaja de Ockham en el contexto de Machine Learning y clasificadores como Naive Bayes, se refleja en la preferencia por modelos simples que generalicen bien en lugar de modelos complejos que puedan sobreajustarse a los datos.

En este caso se asume por simplificacion que:

Las características (features) son independientes entre sí, dado el resultado (clase).

Principio de simplicidad o parsimonia:
"En igualdad de condiciones, la explicación más sencilla, suele ser la correcta“.

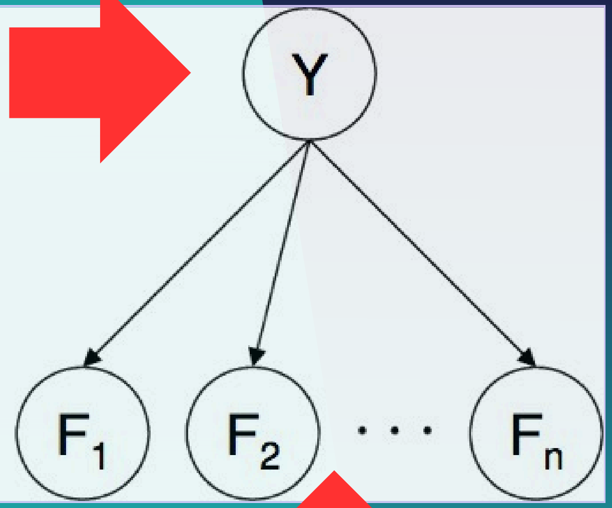
Naïve Bayes

Naïve Bayes, asume una red bayesiana donde los atributos (*features*) son independientes entre si, pero las features dependen de la clase

CLASE

CLASE

$$P(Y, F_1 \dots F_n) = P(Y) \prod_i P(F_i|Y)$$



PROB(CLASE)

FEATURES

FEATURES

- La **clase** representa la clasificacion a la que puede pertenecer una instancia
- La **feature** representa alguna de las caracteristicas o parametros que se tienen de esa instancia.

ES COMO UN CHEF QUE USA UNA RECETA (EL TEOREMA DE BAYES) PARA COMBINAR INGREDIENTES (CARACTERÍSTICAS) Y HACER UN PLATO (LA CLASIFICACIÓN)

Clasificacion basada en Modelos

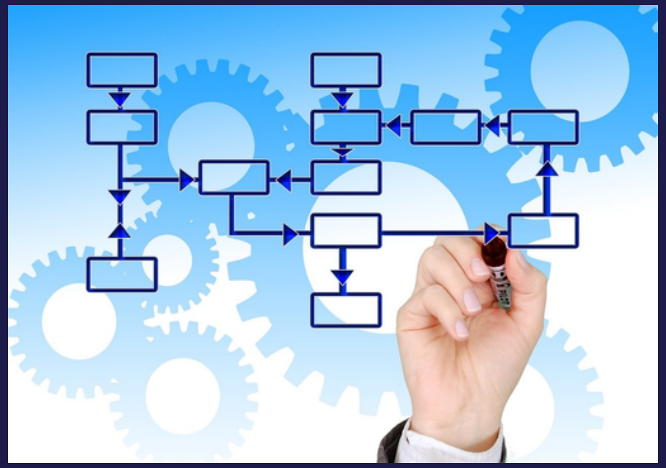
HASTA AHORA

¿Como hemos modelado sistemas con redes bayesianas?

Modelo

Variables

Definir parámetros o variables (evidencia, consulta, ocultas)



Estructura

Definir estructura mediante una Red bayesiana (BN) y Grafos

Probabilidades

Se usan las reglas de probabilidad condicional y el teorema de Bayes

Inferencia

Una vez modelada la red, se pueden responder las preguntas consultadas

Clasificación

Consiste en asignar una etiqueta o clase a una muestra de datos en función de sus características.

Los clasificadores se entrenan con datos etiquetados para aprender a predecir la clase de nuevas muestras.

Para clasificar necesitamos:

1. Recopilación de Datos
2. Preprocesamiento de Datos
    a. Limpiar los datos
    b. Escalar o normalizar las características
3. División del Conjunto de Datos
    a. Conjunto de entrenamiento (70%)
    b. Conjunto de prueba (15%)
    c. Conjunto de validacion (15%)
4. Selección del Modelo (Naive Bayes)
5. Entrenamiento del Modelo
6. Evaluación del Modelo

PERO...

¿Cómo evaluarían si un modelo es bueno o malo para clasificar?

¿Qué son las métricas de evaluación?

Definición

Herramientas para medir el rendimiento de un modelo.

Generalización

Un modelo con alta precisión en los datos de entrenamiento no necesariamente generaliza bien.

Decidir

Permiten comparar diferentes modelos y elegir el mejor.

Guían

Informan qué aspectos del modelo deben ajustarse para optimizar su rendimiento.

Matriz de Confusión

Una tabla que muestra las predicciones del modelo versus las clases reales.

		Valor Actual o Real	
		Positivo	Negativo
Valor Predicho	Positivo		
	Negativo		

Matriz de Confusión

Predicted Class	Actual Class
No Cancer	No Cancer
Has Cancer	Has Cancer
No Cancer	No Cancer
No Cancer	No Cancer
No Cancer	Has Cancer
Has Cancer	Has Cancer
No Cancer	No Cancer
Has Cancer	No Cancer
No Cancer	No Cancer
No Cancer	No Cancer
Has Cancer	No Cancer
No Cancer	No Cancer

		Actual Class	
		Has Cancer	No Cancer
Predicted Class	Has Cancer	2	2
	No Cancer	1	7

Evaluación del Modelo

Usar métricas como:

- ***Exactitud (Accuracy)***: mide qué tan bien el modelo clasifica correctamente

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

  - TP (True Positive): Clasificado correctamente como positivo.
  - TN (True Negative): Clasificado correctamente como negativo.
  - FP (False Positive): Clasificado incorrectamente como positivo (falso positivo).
  - FN (False Negative): Clasificado incorrectamente como negativo (falso negativo).

- **Precisión:** mide la proporción de ejemplos positivos predichos que realmente son positivos.

$$\text{Precision} = \frac{TP}{TP + FP}$$

¿Qué métrica usarías si te importa más evitar falsos positivos?

Evaluación del Modelo

Ejemplo comparativo

Tenemos un modelo de clasificación que predice si un tumor es maligno o benigno.

- Maligno: Tumor peligroso.
- Benigno: Tumor no peligroso.

Supongamos que tenemos 1000 tumores y el modelo realiza las siguientes predicciones:

1000 casos	**Predicción: Maligno**	**Predicción: Benigno**
Real: Maligno	30	70
Real:Benigno	20	880

Clasificacion

Ejemplo comparativo

El ***Accuracy*** mide qué tan bien el modelo clasifica correctamente en general, tanto malignos como benignos.

$$\text{Accuracy} = \frac{70 + 880}{70 + 880 + 20 + 30} = \frac{950}{1000} = 0.95 \quad (95\%)$$

El modelo acierta en el 95% de los casos totales.

Si el conjunto de datos está muy desbalanceado (por ejemplo, 950 benignos y 50 malignos), el modelo puede lograr un accuracy alto simplemente prediciendo todo como benigno, por tanto, ***<u>puede ser engañoso</u>***

Clasificacion

Ejemplo comparativo

La ***Precision*** mide qué porcentaje de los tumores clasificados como malignos realmente lo son.

$$\text{Precision} = \frac{70}{70+20} = \frac{70}{90} = 0.78 \quad (78\%)$$

De todos los tumores que el modelo clasificó como malignos, el 78% realmente lo eran.

Evaluación del Modelo

Usar métricas como:

- ***<u>Recall (Sensibilidad o Tasa de Verdaderos Positivos)</u>***: mide la capacidad del modelo para detectar correctamente los casos positivos (malignos).

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Recall} = \frac{70}{70+30} = \frac{70}{100} = 0.70 \quad (70\%)$$

El modelo identifica correctamente el 70% de los tumores malignos

- **F1-Score:** combina la ***Precision*** y el ***Recall*** en una única métrica para equilibrar ambos aspectos.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{F1-Score} = 2 \times \frac{0.78 \times 0.70}{0.78 + 0.70} = 2 \times \frac{0.546}{1.48} = 0.74 \quad (74\%)$$

Indica que el modelo logra un equilibrio razonable entre precisión y sensibilidad, útil en casos de clases desbalanceadas

¿QUÉ MÉTRICA PRIORIZARÍAS EN UN SISTEMA MÉDICO DONDE ES CRÍTICO DETECTAR TODAS LAS ENFERMEDADES?

Evaluación del Modelo

		Predicted Class		
		Positive	**Negative**	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	Specifitivy $\frac{TP}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TN + FP + FP + FN)}$

Evaluación del Modelo

¿Qué métrica usarías si te importa más evitar falsos positivos?

La **precisión** mide qué proporción de las predicciones positivas realizadas por el modelo son correctas. Penaliza los falsos positivos, ya que estos afectan directamente su valor.

- **Ejemplo:** En un sistema de detección de fraudes, es importante minimizar los falsos positivos para no bloquear transacciones legítimas.

¿Qué métrica priorizarías en un sistema médico donde es crítico detectar todas las enfermedades?

El **recall** mide qué proporción de los casos positivos reales son correctamente identificados por el modelo. En un sistema médico, detectar todas las enfermedades (minimizar los falsos negativos) es crucial, incluso si eso implica aceptar algunos falsos positivos.

- **Ejemplo:** En la detección de cáncer, es preferible que el modelo clasifique erróneamente algunos casos como positivos (aunque no lo sean) a que pase por alto un caso real de cáncer.

Resumen

Entonces:

Accuracy	Útil cuando las clases están equilibradas
Precision	Importante cuando el costo de un **falso positivo** es alto (como falsos diagnósticos positivos).
Recall	Importante cuando el costo de un **falso negativo** es alto (como pasar por alto un tumor maligno).
F1-Score	Útil cuando se necesita un equilibrio entre precisión y sensibilidad.

Ejemplo de clasificación

Ejemplo SPAM

Supongamos que tenemos un correo electronico en el cual queremos identificar correo valido y correo SPAM

Input

EMAIL

Ouput

SPAM

NO SPAM

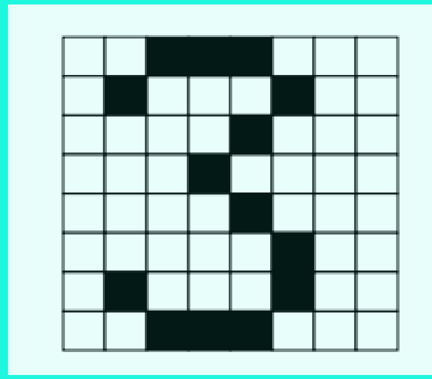
Features (Caracteristicas)

Palabras (free, nigeria, viagra, etc)

OBTENER UNA GRAN CANTIDAD DE EMAILS, CLASIFICAR CADA UNO DE ELLOS COMO SPAM O HAM, APRENDER DE LOS EMAILS PREVIAMENTE CLASIFICADOS

Ejemplo Reconocimiento de Numeros en Imagenes

Supongamos que queremos identificar numeros dentro de una imagen



Input

IMAGEN

Ouput

0

1

...

9

Features (Caracteristicas)

- Pixels (ON/OFF)
- Aspect Ratio (proporción entre el ancho y el alto)
- NumLoops (número de bucles cerrados)

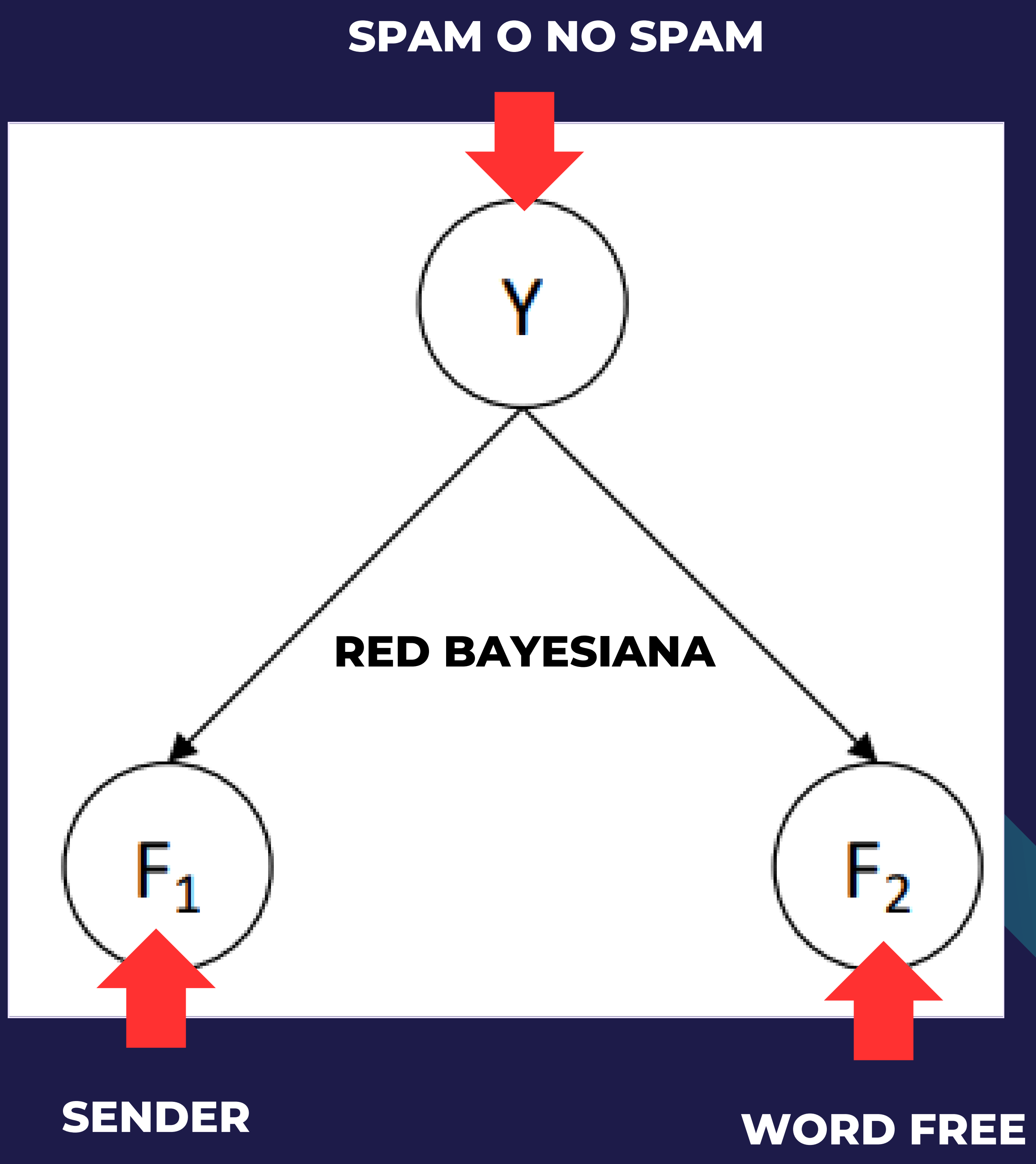
OBTENER UNA GRAN CANTIDAD DE MUESTRAS DE NÚMEROS PREVIAMENTE CLASIFICADOS

Aplicaciones de Clasificacion

- Detección spam
- OCR
- Diagnósticos médicos
- Detección de fraudes
- Ruteo de quejas
- Reconocimiento de Voz
- Segmentación de Clientes en Marketing
- Clasificación de Noticias
- etc...

Algortimo Naïve Bayes

Naïve Bayes



Para el correo electronico sabemos que hay dos clasificaciones esperadas: **SPAM y NO SPAM**

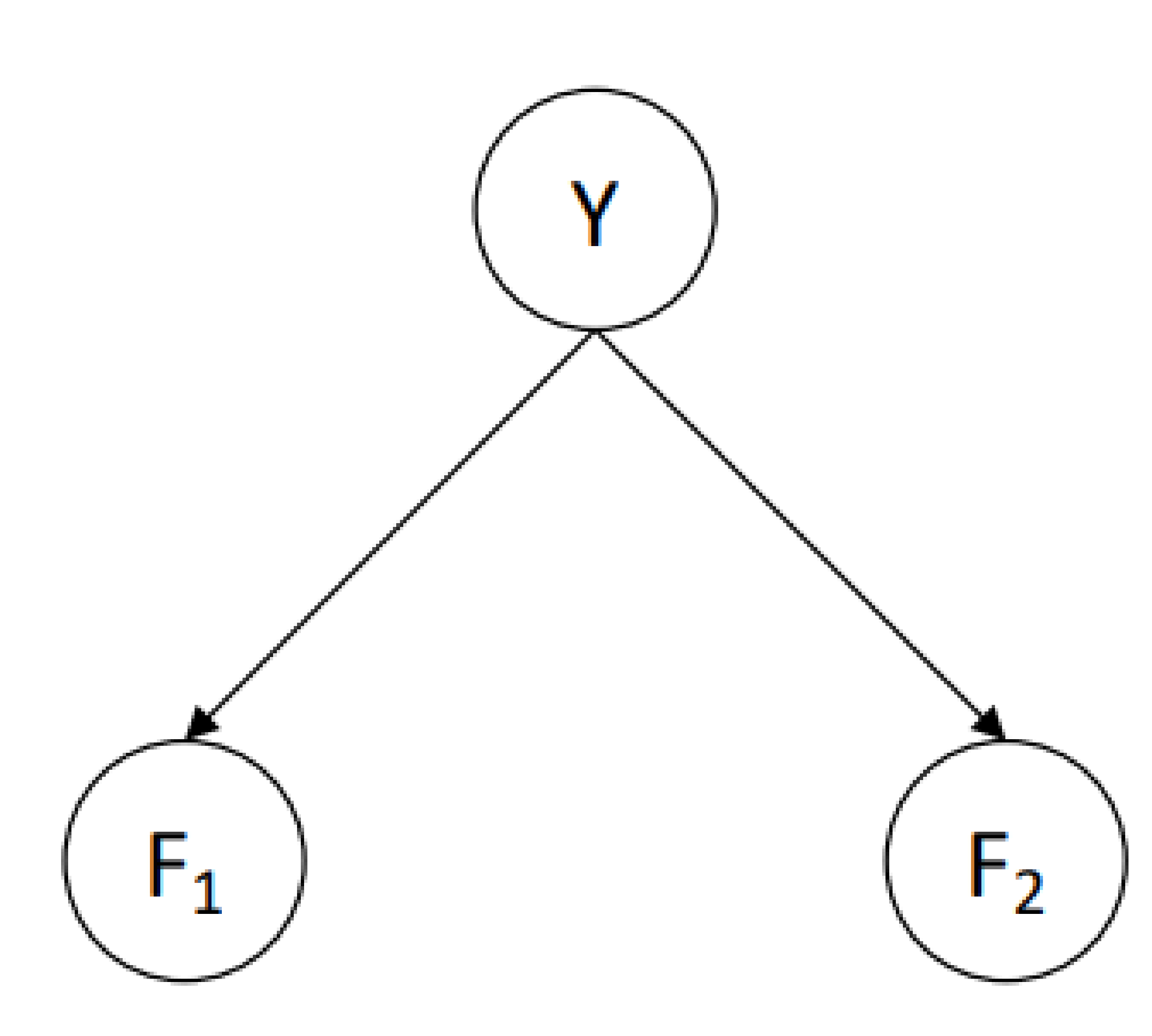
También para este ejemplo en los datos sabremos dos características:

- Quién envía el correo (F1)
- Numero de veces que aparece la palabra *FREE* (F2)

Consideramos ambas *features* independientes entre ellas

Naïve Bayes

Y: The label (spam or ham)	
Y	P(Y)
ham	?
spam	?



F_1: A feature (do I know the sender?)		
F_1	Y	$P(F_1\|Y)$
yes	ham	?
no	ham	?
yes	spam	?
no	spam	?

F_2: Another feature (# of occurrences of FREE)		
F_2	Y	$P(F_2\|Y)$
0	ham	?
1	ham	?
2	ham	?
0	spam	?
1	spam	?
2	spam	?

Naïve Bayes

DATOS DE ENTRENAMIENTO

Training Data		
#	Email Text	Label
1	Attached is my portfolio.	ham
2	Are you **free** for a meeting tomorrow?	ham
3	**Free** unlimited credit cards!!!!	spam
4	Mail $10,000 check to this address	spam
5	Sign up now for 1 **free** Bitcoin	spam
6	**Free** money **free** money	spam

Total de correos: **6**

Veces que aparece ham: **2**

P(ham) = 2/6 = 1/3 = 0.3333

Veces que aparece *FREE* en un correo etiquetado como ham:

- 0 veces: **1** **P(f2|ham) = 1/2 = 0.5**
- 1 veces: **1** **P(f2|ham) = 1/2 = 0.5**
- 2 veces: **0** **P(f2|ham) = 0**

Naïve Bayes

DATOS DE ENTRENAMIENTO

Training Data		
#	Email Text	Label
1	Attached is my portfolio.	ham
2	Are you **free** for a meeting tomorrow?	ham
3	**Free** unlimited credit cards!!!!	spam
4	Mail $10,000 check to this address	spam
5	Sign up now for 1 **free** Bitcoin	spam
6	**Free** money **free** money	spam

Total de correos: **6**

Veces que aparece spam: 4

P(spam) = 4/6 = 2/3 = 0.6666

Veces que aparece *FREE* en un correo etiquetado como spam:

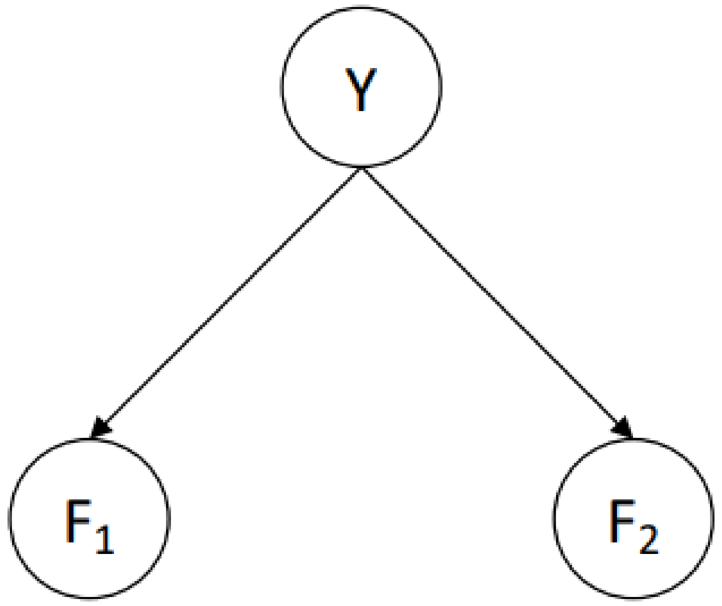
- 0 veces: **1** **P(f2|spam) = 1/4 = 0.25**
- 1 veces: **2** **P(f2|spam) = 2/4 = 0.5**
- 2 veces: **1** **P(f2|spam) = 1/4 = 0.25**

Naïve Bayes

F_2: # of occurrences of FREE		
F_2	Y	$P(F_2 \mid Y)$
0	ham	0.5
1	ham	0.5
2	ham	0.0
0	spam	0.25
1	spam	0.50
2	spam	0.25

Training Data		
#	Email Text	Label
1	Attached is my portfolio.	ham
2	Are you **free** for a meeting tomorrow?	ham
3	**Free** unlimited credit cards!!!!	spam
4	Mail $10,000 check to this address	spam
5	Sign up now for 1 **free** Bitcoin	spam
6	**Free** money **free** money	spam

Naïve Bayes



Y: The label (spam or ham)	
Y	P(Y)
ham	0.6
spam	0.4

F_1: A feature (do I know the sender?)		
F_1	Y	$P(F_1\|Y)$
yes	ham	0.7
no	ham	0.3
yes	spam	0.1
no	spam	0.9

F_2: Another feature (# of occurrences of FREE)		
F_2	Y	$P(F_2\|Y)$
0	ham	0.85
1	ham	0.07
2	ham	0.08
0	spam	0.75
1	spam	0.12
2	spam	0.13

Naïve Bayes

Suponga el mensaje conociendo la persona que envió (contacto):

“Free food in Soda 430 today”

Identifique si es spam o no spam

Sabemos que:

F1 = yes

F2 = 1

Probabilidad Conjunta

Que el correo sea Spam, sabiendo quien lo envio y que tiene la palabra Free una vez en el mensaje

P(Y = spam, F1 = yes, F2 = 1) = P(Y = spam) * P(F1 = yes | Y = spam) * P(F2 = 1 | Y = spam) = 0.4 * 0.1 * 0.12 = **0.0048**

Que el correo sea No Spam, sabiendo quien lo envio y que tiene la palabra Free una vez en el mensaje

P(Y = ham, F1 = yes, F2 = 1) = P(Y = ham) * P(F1 = yes | Y = ham) * P(F2 = 1 | Y = ham) = 0.6 * 0.7 * 0.07 **= 0.0294**

Naïve Bayes

Suponga el mensaje conociendo la persona que envió (contacto):

“Free food in Soda 430 today”

Identifique si es spam o no spam

Sabemos que:

F1 = yes

F2 = 1

Normalizar

P(Y = spam | F1 = yes, F2 = 1) = 0.0048 / (0.0048+0.0294) = **0.14**

P(Y = ham | F1 = yes, F2 = 1) = 0.0294 / (0.0048+0.0294) = **0.86**

14% de posibilidades de que el correo electrónico sea spam.

86% de posibilidades de que sea ham.

Maximum likelihood en Naïve Bayes

Maximum Likelihood

El algoritmo de Naive Bayes utiliza el principio de Máxima Verosimilitud (Maximum Likelihood) para estimar las probabilidades necesarias para realizar la clasificación.

El objetivo es encontrar la clase ***C*** que maximice la probabilidad posterior dado un conjunto de características ***X=(x1,x2,...,xn)***:

$$P(C \mid X) = \frac{P(X \mid C) \cdot P(C)}{P(X)}$$

Dado que P(X) es constante para todas las clases, basta con maximizar el numerador (o sea no necesitamos normalizar para identificar la clasificacion, basta con encontrar el mayor valor en el numerador):

$$P(C \mid X) \propto P(X \mid C) \cdot P(C)$$

Over-fitting

Over-fitting

Ocurre cuando un modelo de machine learning se ajusta demasiado bien a los datos de entrenamiento, capturando tanto los patrones reales como el ruido o las anomalías presentes en el conjunto de datos.

Como resultado, el modelo funciona muy bien con los datos de entrenamiento pero falla al generalizar con nuevos datos

¿Por qué ocurre el overfitting?

- Modelo demasiado complejo
- Demasiadas características
- Pocos datos de entrenamiento
- Entrenamiento excesivo

¿Cómo detectar el overfitting?

- **Diferencia de rendimiento entre entrenamiento y prueba:**
  - Alta exactitud (accuracy) en entrenamiento (por ejemplo, 98%).
  - Baja exactitud (accuracy) en prueba (por ejemplo, 60%).

- **Curvas de aprendizaje:**
  - La precisión de entrenamiento sigue aumentando, pero la precisión de validación se estanca o disminuye.

¿Cómo evitar el overfitting?

Aunque el algoritmo de Naive Bayes generalmente es menos propenso al overfitting en comparación con modelos más complejos como redes neuronales, aún puede ocurrir en ciertos escenarios, para esto se puede aplicar las siguientes tecnicas:

- **Suavizado de Laplace (Laplace Smoothing)**
- **Selección de características:** remover features irrelevantes o ruidosas que hagan aprender patrones falsos
- **Eliminación de características redundantes o correlacionadas** (features realmente dependientes)
- **Limpieza y preprocesamiento de datos:** Eliminar valores atípicos (outliers) y normalizar los datos puede reducir el riesgo.

¿Cómo evitar el overfitting?

- **Validación cruzada (Cross-Validation):** Realizar una validación cruzada, como **K-Fold***, ayuda a evaluar el modelo en diferentes subconjuntos de datos para detectar el overfitting de manera temprana.

- **Aumento de datos (Data Augmentation):** Si tienes pocos datos, el modelo puede memorizar ejemplos específicos. Generar ejemplos adicionales, especialmente en tareas de clasificación de imágenes o texto, puede ayudar a mejorar la generalización.

- **Priorización de clases balanceadas:** Si el conjunto de datos está muy desbalanceado, el modelo podría sobreajustarse a la clase mayoritaria.

K-Fold
Cross-Validation

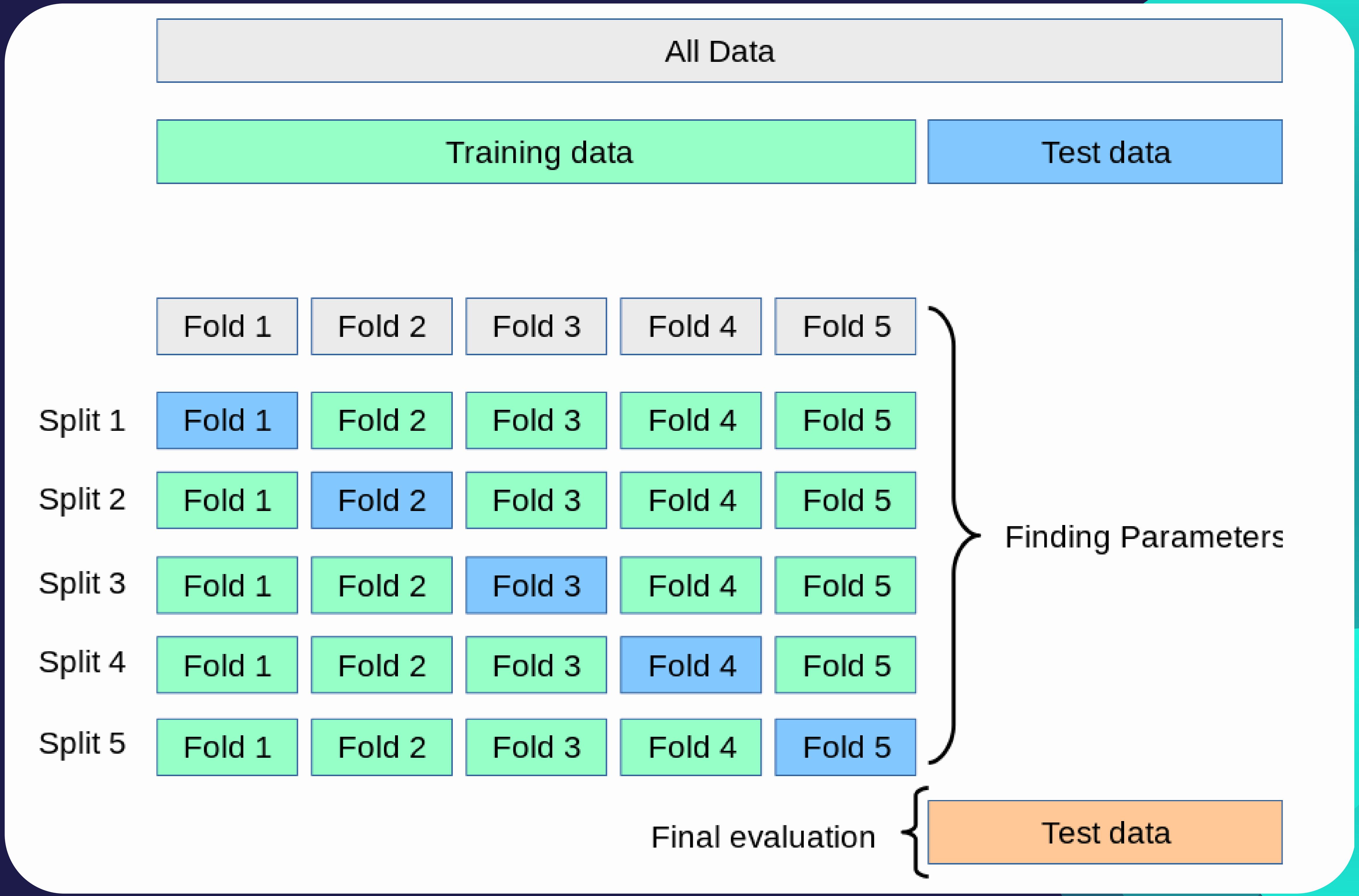
K-Fold Cross-Validation

Es una técnica utilizada para evaluar el rendimiento de un modelo de machine learning y evitar problemas como el overfitting. Su objetivo principal es asegurarse de que el modelo generalice bien a datos nuevos

Pasos

1. **Dividir los datos en K grupos (o "folds") de tamaño aproximadamente igual.**
   a. Por ejemplo, si se tienen 1000 datos y se elige K=5, cada grupo tendrá 200 datos
2. **Entrenamiento y prueba repetidos:**
   a. Se realizan K iteraciones en total.
   b. En cada iteración:
      i. (K-1) folds se utilizan para entrenar el modelo.
      ii. 1 fold se utiliza para probar el modelo.
3. **Promedio de resultados:**
   a. Después de entrenar y evaluar el modelo K veces, se promedian las métricas (como precisión, recall, F1-score, etc.) para obtener una estimación más robusta del rendimiento.

K-Fold Cross-Validation



Laplace Smoothing

Laplace Smoothing

Es una técnica que se utiliza en el algoritmo Naive Bayes para evitar el problema de las probabilidades nulas

¿Cuál es el problema?

Sabemos que: $$P(\text{clase} \mid \text{característica}) = \frac{P(\text{característica} \mid \text{clase}) \cdot P(\text{clase})}{P(\text{característica})}$$

El mayor problema surge al calcular $P(\text{característica} \mid \text{clase})$, ya que si alguna característica no aparece en una clase, el resultado será cero.

Si la **frecuencia** es cero, la probabilidad será cero, y eso hará que toda la probabilidad de la clase también sea cero.

Laplace Smoothing

El suavizado de Laplace añade un valor constante (generalmente 1) a cada recuento para asegurarse de que ninguna probabilidad sea cero.

$$P(\text{característica} \mid \text{clase}) = \frac{\text{Frecuencia}(\text{característica} \mid \text{clase}) + 1}{\text{Total de características en la clase} + V}$$

- V es el número total de características posibles.
- El valor "1" garantiza que incluso las características no observadas tengan una pequeña probabilidad.

Ejemplo Bag of Words usando Laplace Smoothing

Supongamos que tenemos un conjunto de datos de correos electrónicos clasificados como spam y no spam.

Email	Spam (1) / No Spam (0)	Palabras en el email
"oferta exclusiva"	1	oferta, exclusiva
"gana dinero fácil"	1	gana, dinero, facil
"saludos cordiales"	0	saludos, cordiales
"reunión de trabajo"	0	reunion, de, trabajo

Ejemplo Bag of Words usando Laplace Smoothing

Contemos las frecuencias de las palabras en cada clase (spam y no spam)

Spam:

- "oferta": 1
- "exclusiva": 1
- "gana": 1
- "dinero": 1
- "fácil": 1

No Spam:

- "saludos": 1
- "cordiales": 1
- "reunión": 1
- "trabajo": 1

Total de palabras en Spam: 5

Total de palabras en No Spam: 4

Ejemplo Bag of Words usando Laplace Smoothing

Probabilidades sin suavizado

Spam

Palabra	Frecuencia	P(Palabra)
"oferta"	1	1/5
"exclusiva"	1	1/5
"gana"	1	1/5
"dinero"	1	1/5
"fácil"	1	1/5

Ejemplo Bag of Words usando Laplace Smoothing

Probabilidades sin suavizado

Ham

Palabra	Frecuencia	P(Palabra)
"saludos"	1	1/4
"cordiales"	1	1/4
"reunión"	1	1/4
"trabajo"	1	1/4

Vocabulario Total:

$$V = \{\text{"oferta"}, \text{"exclusiva"}, \text{"gana"}, \text{"dinero"}, \text{"fácil"}, \text{"saludos"}, \text{"cordiales"}, \text{"reunión"}, \text{"trabajo"}\}$$

9 palabras unicas

Ejemplo Bag of Words usando Laplace Smoothing

Aplicando Suavizado

$$P(\text{palabra} \mid \text{clase}) = \frac{\text{Frecuencia de la palabra} + 1}{\text{Total de palabras en la clase} + V}$$

Spam:

$$P(\text{"oferta"} \mid \text{spam}) = \frac{1+1}{5+9} = \frac{2}{14} \approx 0.1429$$

$$P(\text{"exclusiva"} \mid \text{spam}) = \frac{1+1}{5+9} = \frac{2}{14} \approx 0.1429$$

$$P(\text{"gana"} \mid \text{spam}) = \frac{1+1}{5+9} = \frac{2}{14} \approx 0.1429$$

No Spam:

$$P(\text{"saludos"} \mid \text{no spam}) = \frac{1+1}{4+9} = \frac{2}{13} \approx 0.1538$$

$$P(\text{"cordiales"} \mid \text{no spam}) = \frac{1+1}{4+9} = \frac{2}{13} \approx 0.1538$$

Laplace Smoothing

Con nuestro proceso previamente calculado, tenemos la probabilidades ahora, supongamos que queremos clasificar un nuevo email: **“gana dinero”**

Queremos saber si es spam o no spam, usando Naive Bayes:

$$P(\text{Spam} \mid \text{gana, dinero}) = P(\text{Spam}) \cdot P(\text{gana} \mid \text{Spam}) \cdot P(\text{dinero} \mid \text{Spam})$$

$$P(\text{Spam}) = \frac{2}{4} = 0.5$$

$$P(\text{gana} \mid \text{Spam}) = 0.1429$$

$$P(\text{dinero} \mid \text{Spam}) = 0.1429$$

$$P(\text{Spam} \mid \text{gana, dinero}) = 0.5 \cdot 0.1429 \cdot 0.1429 \approx 0.0102$$

$$P(\text{No Spam}) = \frac{2}{4} = 0.5$$

$$P(\text{gana} \mid \text{No Spam}) = \frac{1}{13} \approx 0.0769$$

$$P(\text{dinero} \mid \text{No Spam}) = \frac{1}{13} \approx 0.0769$$

$$P(\text{No Spam} \mid \text{gana, dinero}) = 0.5 \cdot 0.0769 \cdot 0.0769 \approx 0.00295$$

Por Maximum Likelihood podemos conclur P(Spam|gana, dinero) > P(No Spam | gana, dinero) por tanto podemos concluir que el correo es SPAM

Suma de Logaritmos

Suma de Logaritmos

Los números muy pequeños pueden desaparecer (es decir, ser redondeados a cero) o pueden causar errores de precisión, lo que hace que los cálculos sean inexactos, el valor final puede ser tan pequeño que se pierde la información.

Una técnica comúnmente utilizada para evitar estos problemas es trabajar con logaritmos de las probabilidades en lugar de las probabilidades directamente.

Normal

$$P(\text{clase} \mid \text{características}) = \frac{P(\text{característica}_1 \mid \text{clase}) \cdot P(\text{característica}_2 \mid \text{clase}) \cdot \ldots}{P(\text{características})}$$

Logaritmo

$$P(\text{clase} \mid \text{características}) = \log P(\text{característica}_1 \mid \text{clase}) + \log P(\text{característica}_2 \mid \text{clase}) + \ldots$$