XXXVI CONGRESO NACIONAL
XXII CONGRESO INTERNACIONAL
de Informática y Computación de la ANIEI

Universidad Veracruzana

Facultad de ESTADÍSTICA E INFORMÁTICA

Haciendo que una máquina aprenda: árboles de decisión

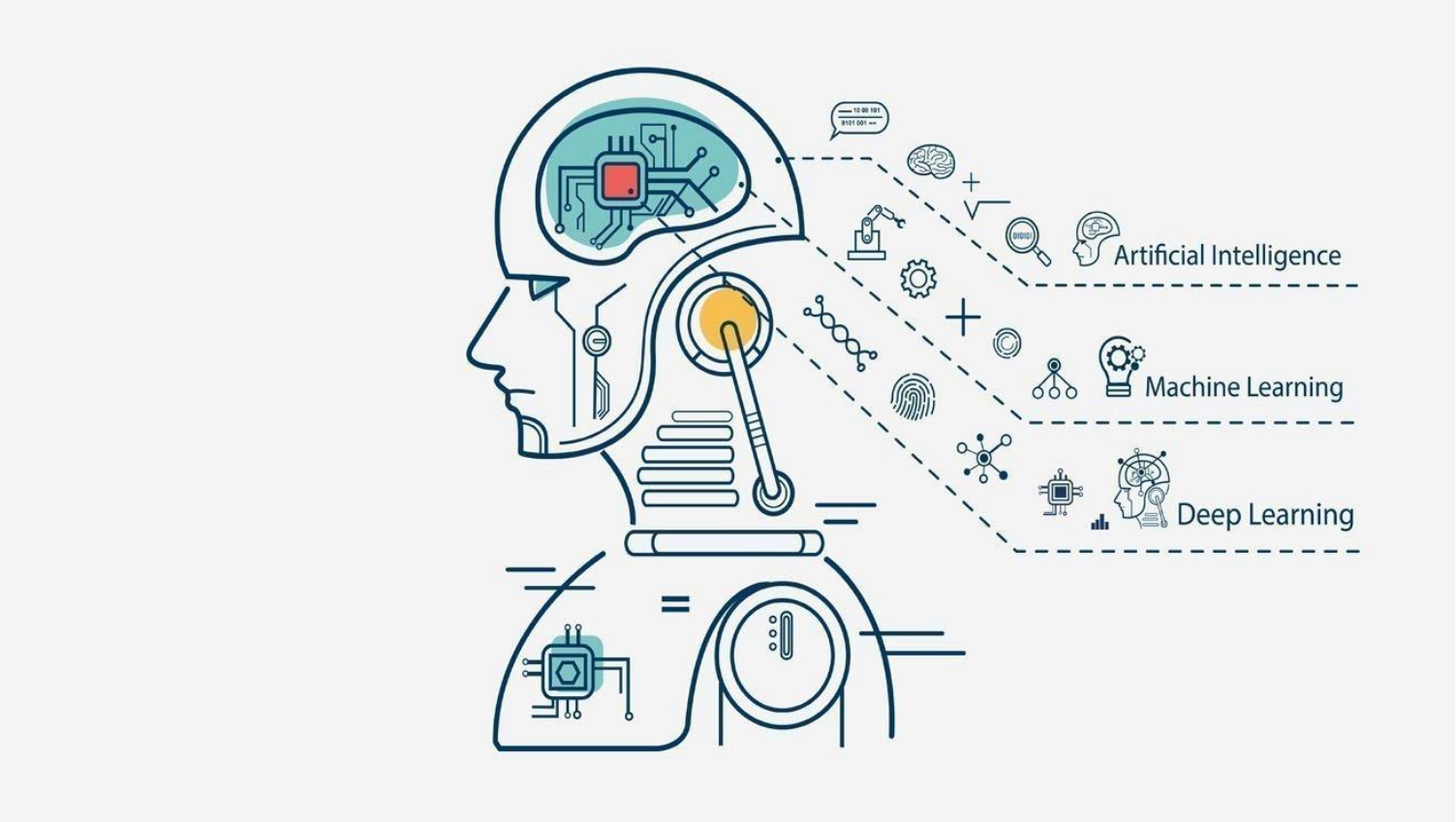
Prof. Saúl Domínguez-Isidro
Universidad Veracruzana
sauldominguez@uv.mx

40 años ANIEI
Impulsando el Talento en TI

Temario

I. Introducción
II. Aprendizaje Máquina
III. Conceptos básicos
IV. Construcción de un Árbol de Decisión
V. Algoritmos y métricas
VI. Ejercicios Prácticos
VII. Sesión de Preguntas y Respuestas

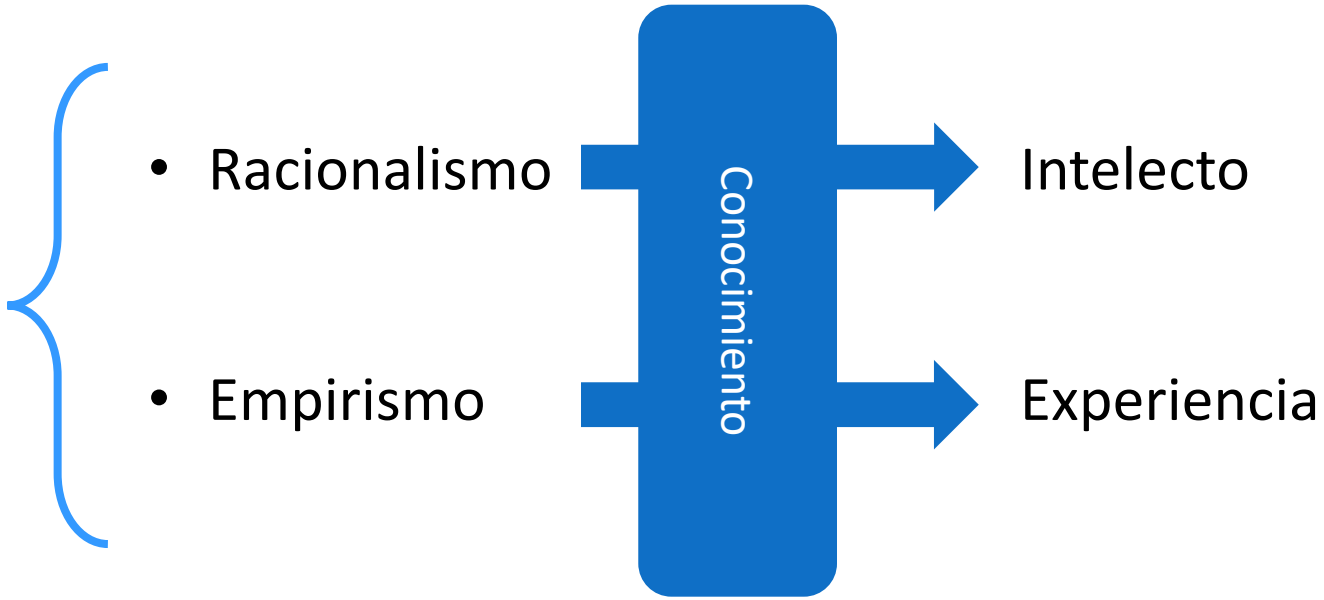
¿Una máquina puede aprender?



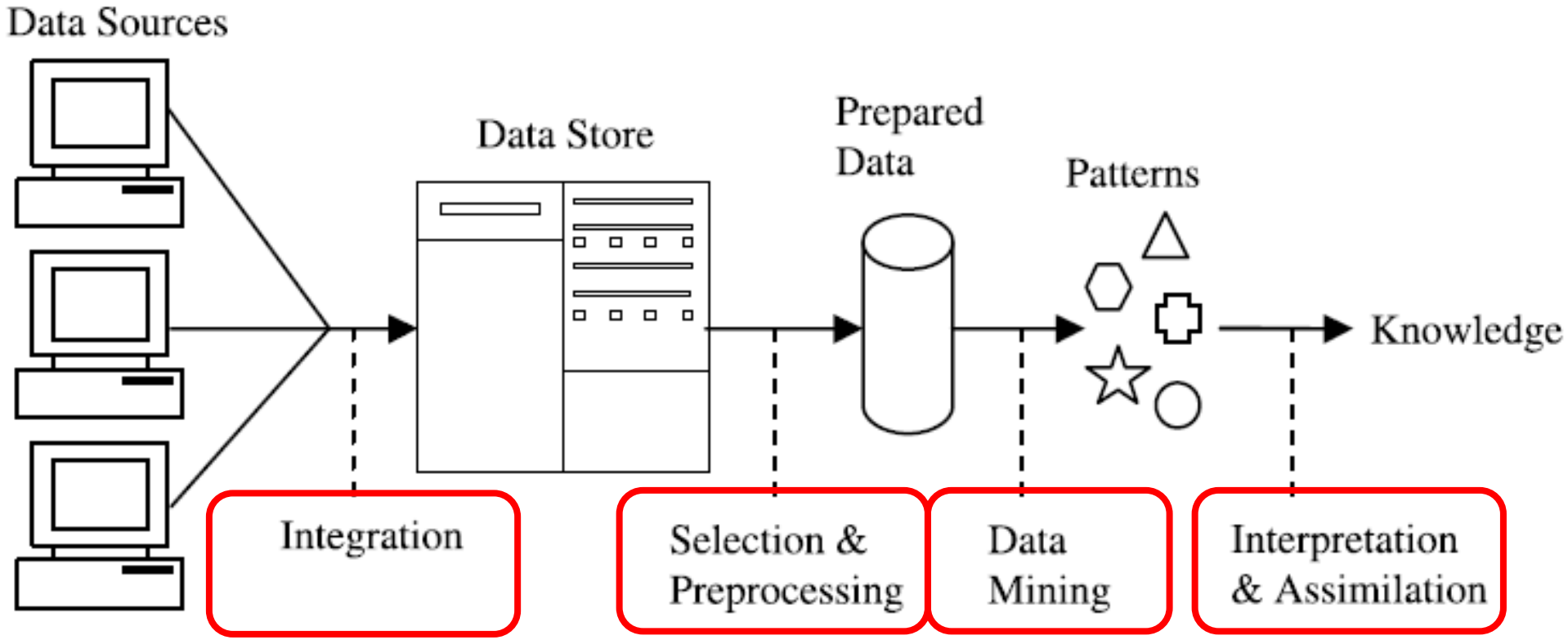
¿cómo aprende una máquina?

Aprender

Acción de **adquirir y retener conocimiento**,
habilidad o información sobre alguna materia.

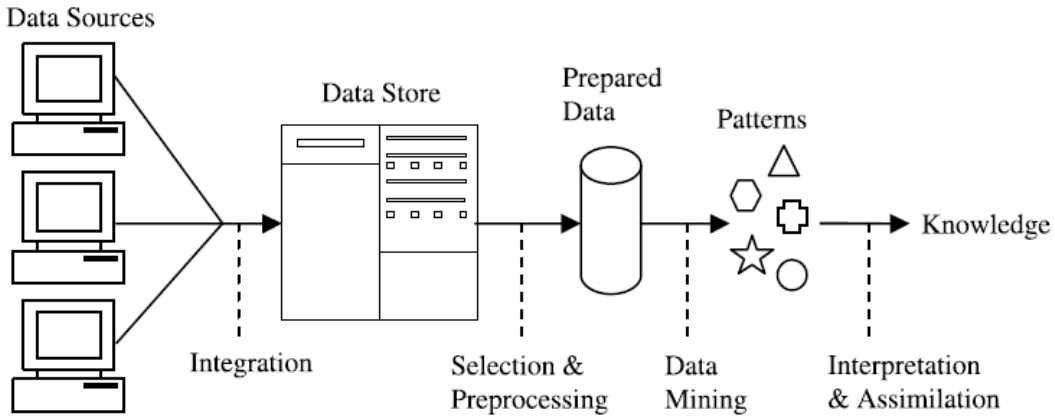


Descubrimiento del Conocimiento en bases de datos (KDD)



Patrones en datos

¿Y cómo se expresan los patrones?

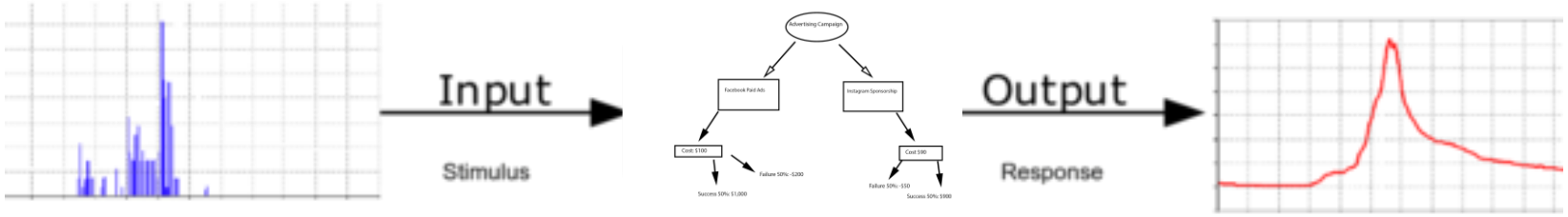


Patrones de caja negra

1. como una caja negra cuyas entrañas son efectivamente incomprensibles

Patrones estructurales

2. como una caja transparente cuya construcción revela la estructura del patrón.



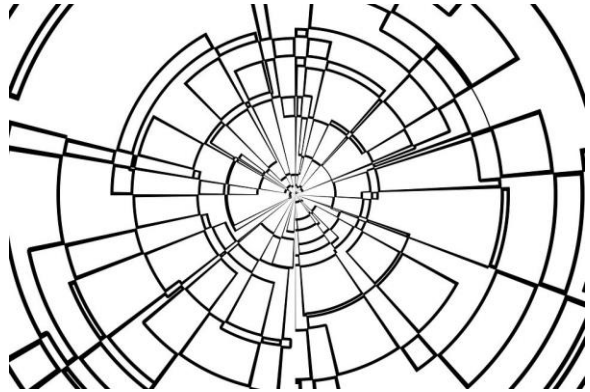
Diferencias

La diferencia entre ambas expresiones es: **si los patrones que se extraen están representados o no** en términos de una estructura que puede ser examinada, razonada y utilizada para informar futuras decisiones.

Los llamamos estructurales capturan la estructura de decisión de manera explícita. En otras palabras, ayudan a explicar algo sobre los datos.

Describiendo patrones estructurales

- ¿Qué se entiende por patrones estructurales?
- ¿Cómo los describes?
- ¿qué forma toma la entrada?



Responderemos estas preguntas a modo de ilustración en lugar de intentar definiciones formales

Lentes de contacto

	Edad	Prescripción de anteojos	Astigmatismo	Tasa de producción de lágrimas	Lentes recomendadas
1	joven	miope	no	reducida	no
2	joven	miope	no	normal	suave
3	joven	miope	si	reducida	no
4	joven	miope	si	normal	duro
5	joven	hipermetrope	no	reducida	no
6	joven	hipermetrope	no	normal	suave
7	joven	hipermetrope	si	reducida	no
8	joven	hipermetrope	si	normal	duro
9	pre-presbicia	miope	no	reducida	no
10	pre-presbicia	miope	no	normal	suave
11	pre-presbicia	miope	si	reducida	no
12	pre-presbicia	miope	si	normal	duro
13	pre-presbicia	hipermetrope	no	reducida	no
14	pre-presbicia	hipermetrope	no	normal	suave
15	pre-presbicia	hipermetrope	si	reducida	no
16	pre-presbicia	hipermetrope	si	normal	no
17	presbicia	miope	no	reducida	no
18	presbicia	miope	no	normal	no
19	presbicia	miope	si	reducida	no
20	presbicia	miope	si	normal	duro
21	presbicia	hipermetrope	no	reducida	no
22	presbicia	hipermetrope	no	normal	suave
23	presbicia	hipermetrope	si	reducida	no
24	presbicia	hipermetrope	si	normal	no

1. edad del paciente: (1) joven, (2) pre-presbicia, (3) presbicia
2. prescripción de gafas: (1) miope, (2) hipermétrope
3. astigmático: (1) no, (2) sí
4. tasa de producción de lágrimas: (1) reducida, (2) normal

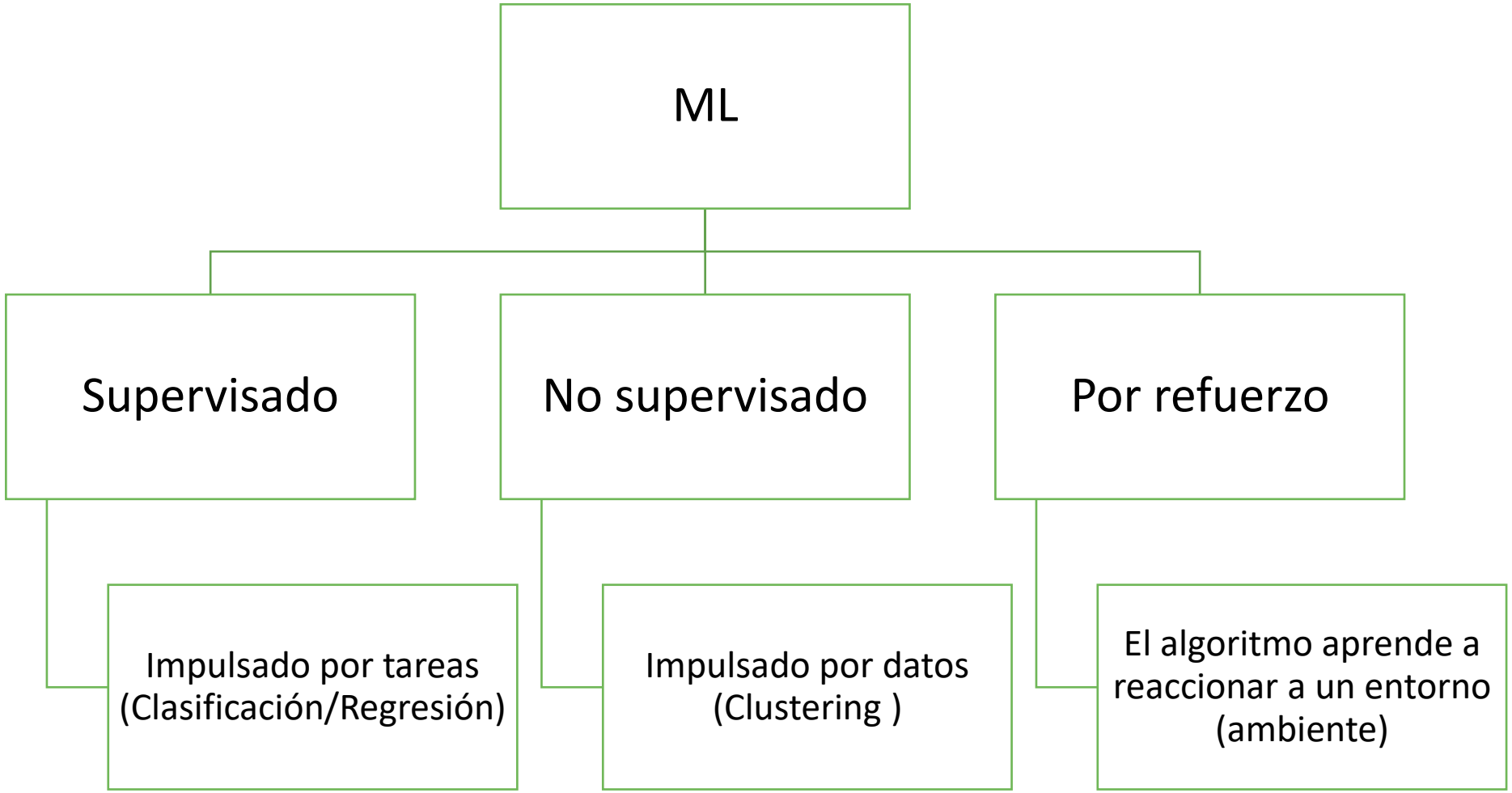
Estructura

- Regla

SI tasa_producción_lágrimas = reducida **ENTONCES** recomendación=no

EN OTRO CASO, SI edad=joven **Y** astigmatico=no **ENTONCES** recomendacion=suave

Técnicas de ML



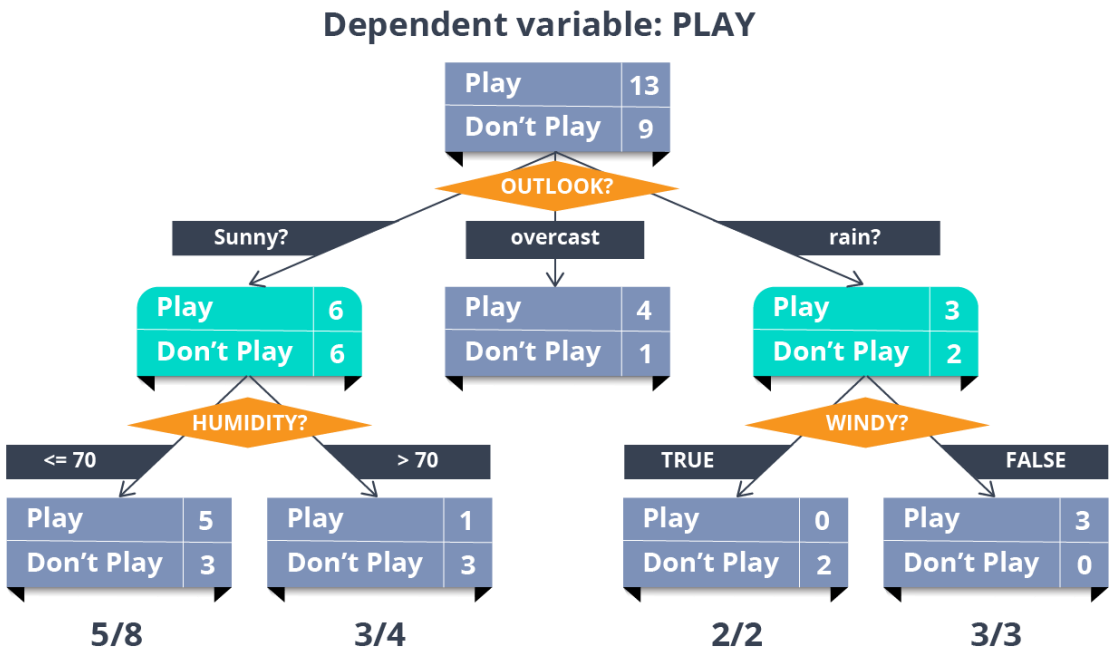
Farrukh Akhtar (2017) Practical Reinforcement Learning: Develop self-evolving, intelligent agents with OpenAI Gym, Python, and Java

Árbol de Decisión

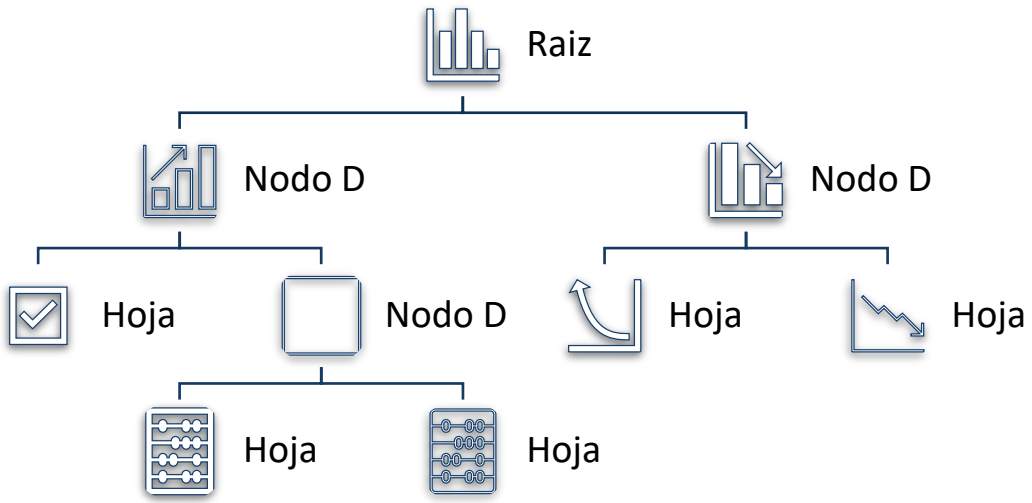
Son diagramas con construcciones lógicas, similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema.

Están compuestos por nodos interiores, nodos terminales y ramas que emanan de los nodos interiores.

Cada nodo interior en el árbol contiene una prueba de un atributo, y cada rama representa un valor distinto del atributo.



Conceptos



Nodo raíz: Nodo principal, representa una de las características más importantes del problema.

Nodo de decisión: Así se le denomina al nodo que tiene asociados dos o más sub-nodos

Nodo terminal (leaf): Es un nodo que no tiene divisiones

División (split): Proceso de dividir un nodo en dos o más sub-nodos

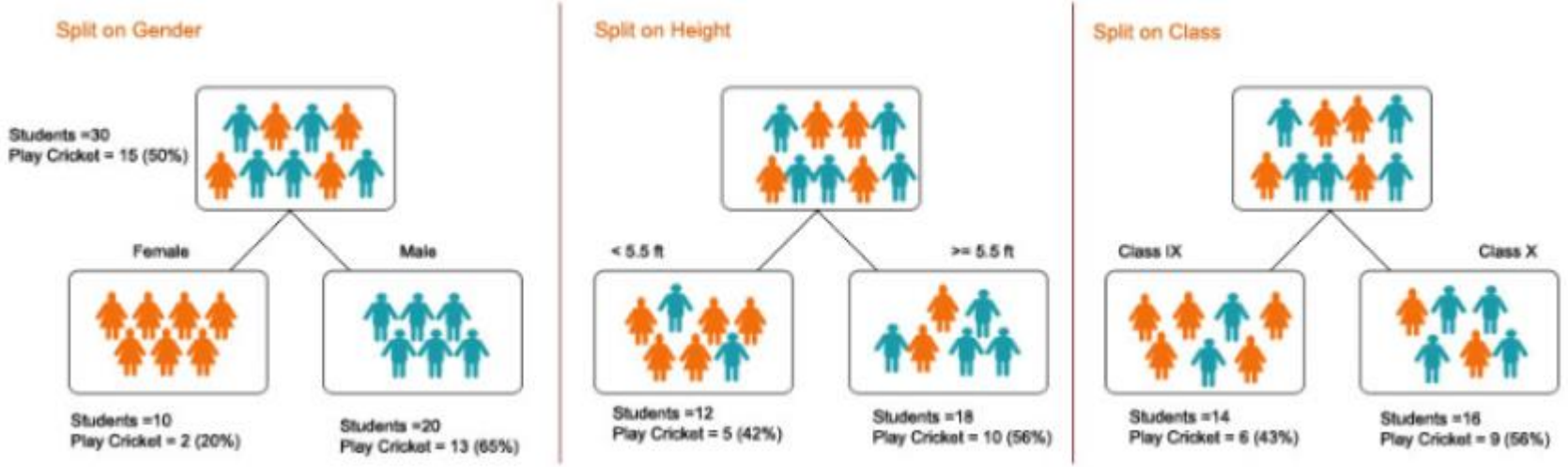
Poda (pruning): Proceso de remover sub-nodos de un nodo de decisión

Rama (sub-árbol): Es un subsección de un DE

Nodo padre e hijo: cuando un nodo es dividido en sub-nodos se les conoce como padre de los sub-nodos, mientras que los sub-nodos son los hijos del nodo padre.

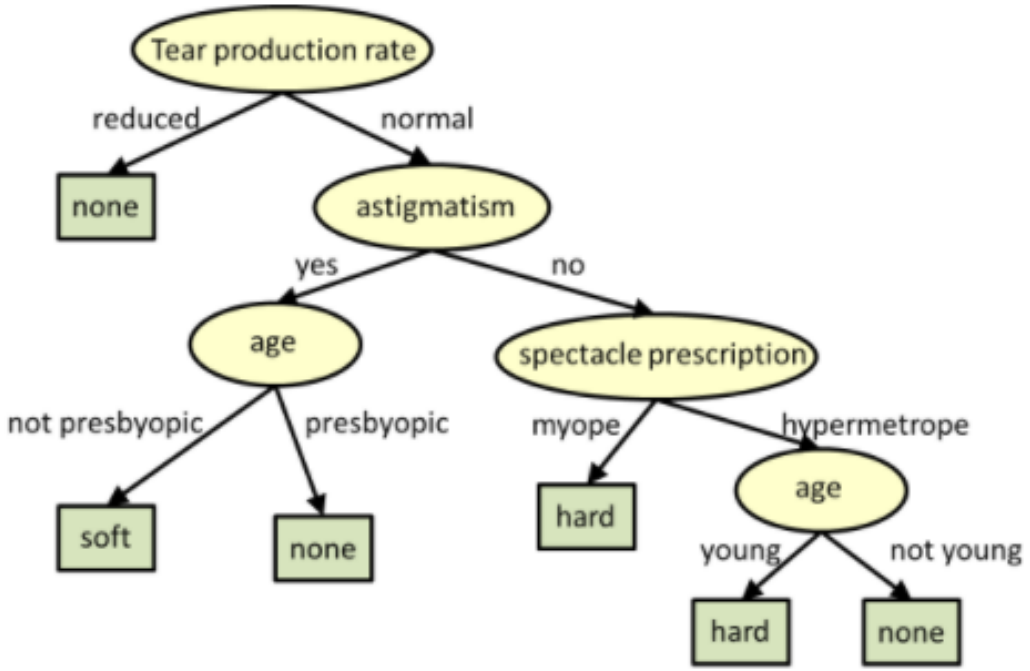
Cómo funciona un DT

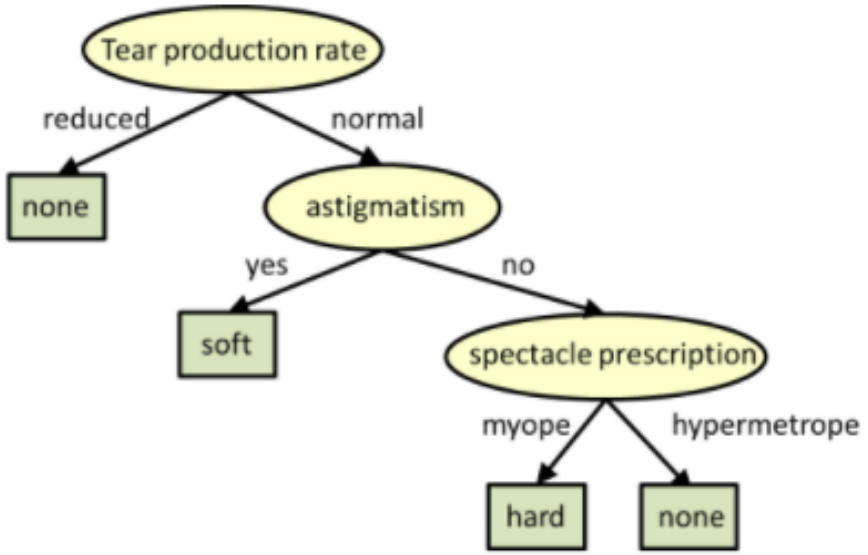
- Los criterios para realizar divisiones estratégicas afectan en gran medida la precisión de un árbol.
- Los criterios de decisión son diferentes para los árboles de clasificación y regresión.



Generalización de un DT

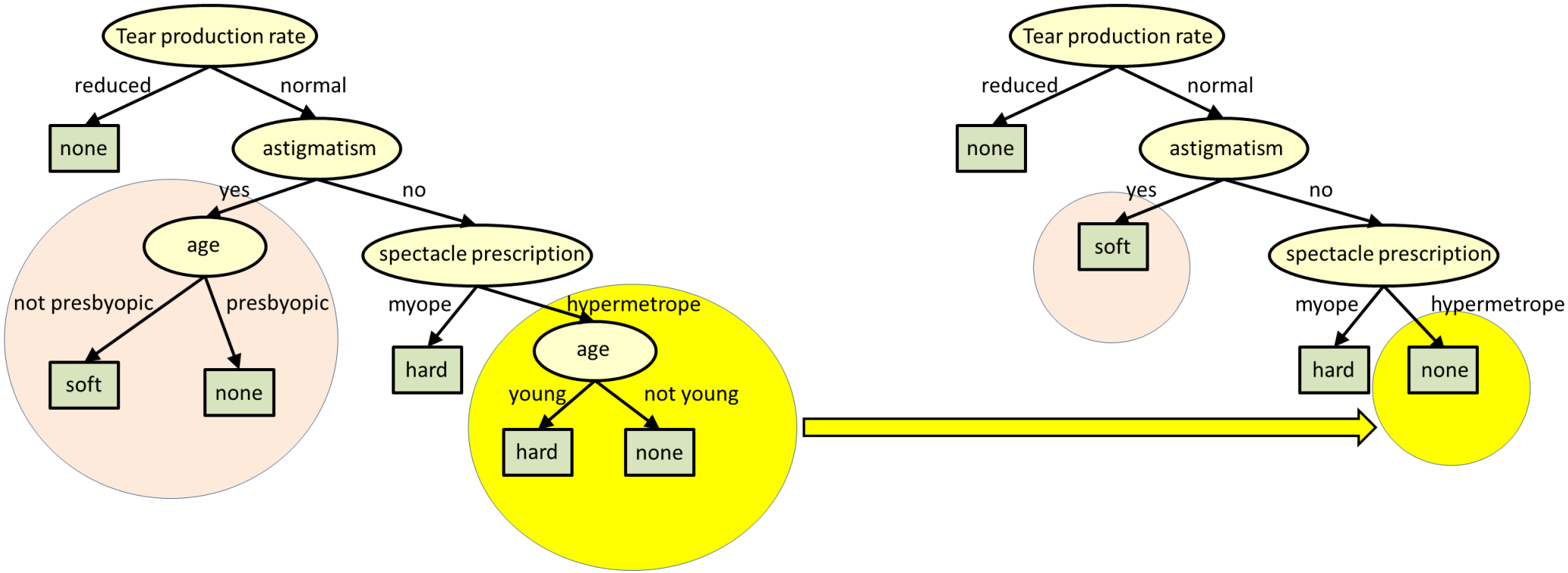
- Los árboles de decisión que se entrenan con cualquier dato de entrenamiento corren el riesgo de sobreajustar los datos de entrenamiento.





Poda

La técnica más simple es podar las partes del árbol que resulten en la menor ganancia de información. Este procedimiento no requiere ningún dato adicional y solo basa la poda en la información que ya está calculada cuando el árbol se construye a partir de datos de entrenamiento.



Conjunto de datos Jugadores de Golf

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
3	Overcast	Hot	High	Weak	46
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
7	Overcast	Cool	Normal	Strong	43
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
10	Rain	Mild	Normal	Weak	46
11	Sunny	Mild	Normal	Strong	48
12	Overcast	Mild	High	Strong	52
13	Overcast	Hot	Normal	Weak	44
14	Rain	Mild	High	Strong	30

- Avg = 39.78
- Std Dev = 9.32

Analizando característica “Outlook”

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
11	Sunny	Mild	Normal	Strong	48
3	Overcast	Hot	High	Weak	46
7	Overcast	Cool	Normal	Strong	43
12	Overcast	Mild	High	Strong	52
13	Overcast	Hot	Normal	Weak	44
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
10	Rain	Mild	Normal	Weak	46
14	Rain	Mild	High	Strong	30

- Avg = 35.2
- Std Dev = 7.78

- Avg = 46.25
- Std Dev = 3.49

- Avg = 39.2
- Std Dev = 10.87

Resumen de Std.Dev. de la característica Outlook

Outlook	Std.Dev	Instancias
Overcast	3.49	4
Rain	10.87	5
Sunny	7.78	5

- desviación estándar ponderada para Outlook = $\frac{4}{14}3.49 + \frac{5}{14}10.87 + \frac{5}{14}7.78 = 7.66$

- reducción de la desviación estándar para Outlook = 9.32 – 7.66 = 1.66

Analizando característica “Temp.”

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
3	Overcast	Hot	High	Weak	46
13	Overcast	Hot	Normal	Weak	44
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
7	Overcast	Cool	Normal	Strong	43
9	Sunny	Cool	Normal	Weak	38
4	Rain	Mild	High	Weak	45
8	Sunny	Mild	High	Weak	35
10	Rain	Mild	Normal	Weak	46
11	Sunny	Mild	Normal	Strong	48
12	Overcast	Mild	High	Strong	52
14	Rain	Mild	High	Strong	30

- Std Dev = 8.95
- Std Dev = 10.51
- Std Dev = 7.65

Resumen de Std.Dev. de la característica Temp.

Temperature	Std.Dev	Instancias
Hot	8.95	4
Cool	10.51	4
Mild	7.65	6

- desviación estándar ponderada para Temperature = $\frac{4}{14}8.95 + \frac{4}{14}10.51 + \frac{6}{14}7.65 = 8.84$

- reducción de la desviación estándar para Temperature = 9.32 – 8.84 = 0.47

Analizando característica “Humidity”

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
3	Overcast	Hot	High	Weak	46
4	Rain	Mild	High	Weak	45
8	Sunny	Mild	High	Weak	35
12	Overcast	Mild	High	Strong	52
14	Rain	Mild	High	Strong	30

- Std Dev = 9.36

5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
7	Overcast	Cool	Normal	Strong	43
9	Sunny	Cool	Normal	Weak	38
10	Rain	Mild	Normal	Weak	46
11	Sunny	Mild	Normal	Strong	48
13	Overcast	Hot	Normal	Weak	44

- Std Dev = 8.73

Resumen de Std.Dev. de la característica Humidity

Humidity	Std.Dev	Instancias
High	9.36	7
Normal	8.73	7

- desviación estándar ponderada para Humidity = $\frac{7}{14}9.36 + \frac{7}{14}8.73 = 9.04$

- reducción de la desviación estándar para Humidity = 9.32 – 9.04 = 0.27

Analizando característica “Wind”

Day	Outlook	Temp.	Humidity	Wind	Golf Players
2	Sunny	Hot	High	Strong	30
6	Rain	Cool	Normal	Strong	23
7	Overcast	Cool	Normal	Strong	43
11	Sunny	Mild	Normal	Strong	48
12	Overcast	Mild	High	Strong	52
14	Rain	Mild	High	Strong	30

1	Sunny	Hot	High	Weak	25
3	Overcast	Hot	High	Weak	46
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
10	Rain	Mild	Normal	Weak	46
13	Overcast	Hot	Normal	Weak	44

- Std Dev = 10.59

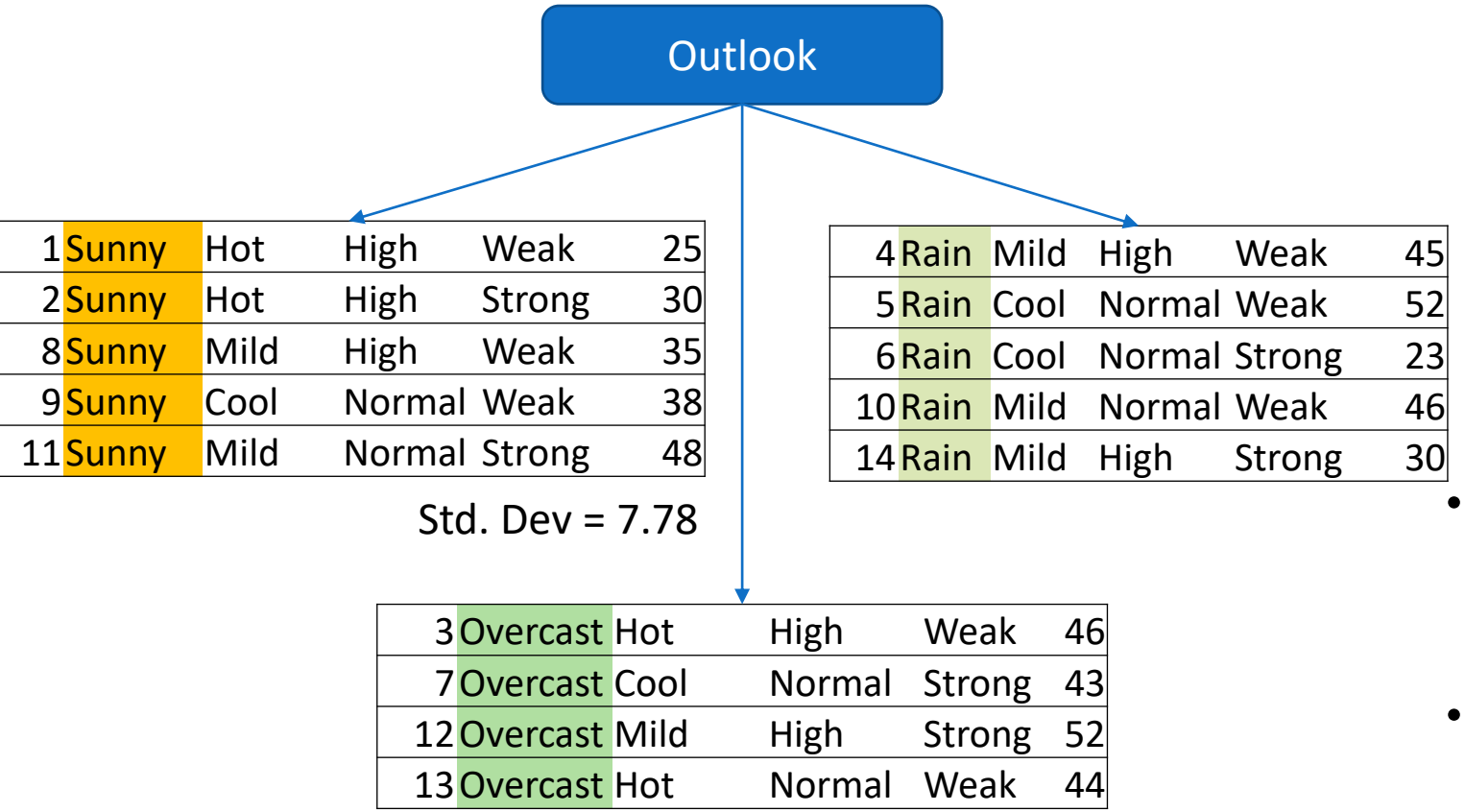
- Std Dev = 7.87

Resumen de Std.Dev. de la característica Wind

Wind	Std.Dev	Instancias
Strong	10.59	6
Weak	7.87	8

- desviación estándar ponderada para Wind = $\frac{6}{14}10.59 + \frac{8}{14}7.87 = 9.03$
- reducción de la desviación estándar para Wind = 9.32 – 9.03 = 0.29

Determinando el nodo Raiz



Característica	Reducción Std.Dev
Outlook	1.66
Temperature	0.47
Humidity	0.27
Wind	0.29

- De acuerdo a la reducción de la desviación estándar, la característica que la maximiza es Outlook, por lo tanto,
- Outlook será nuestro primer nodo de decisión (nodo raíz)

Definiendo nodos para Outlook:Sunny

1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
11	Sunny	Mild	Normal	Strong	48

Std. Dev = 7.78

Temp.	Std.Dev	Instancias
Hot	2.5	2
Cool	0	1
Mild	6.5	2

Std. Dev ponderada 3.6

Humidity	Std.Dev	Instancias
High	4.08	3
Normal	5	2

Std. Dev ponderada 4.45

Wind	Std.Dev	Instancias
Strong	9	2
Weak	5.56	3

Std. Dev ponderada 6.93

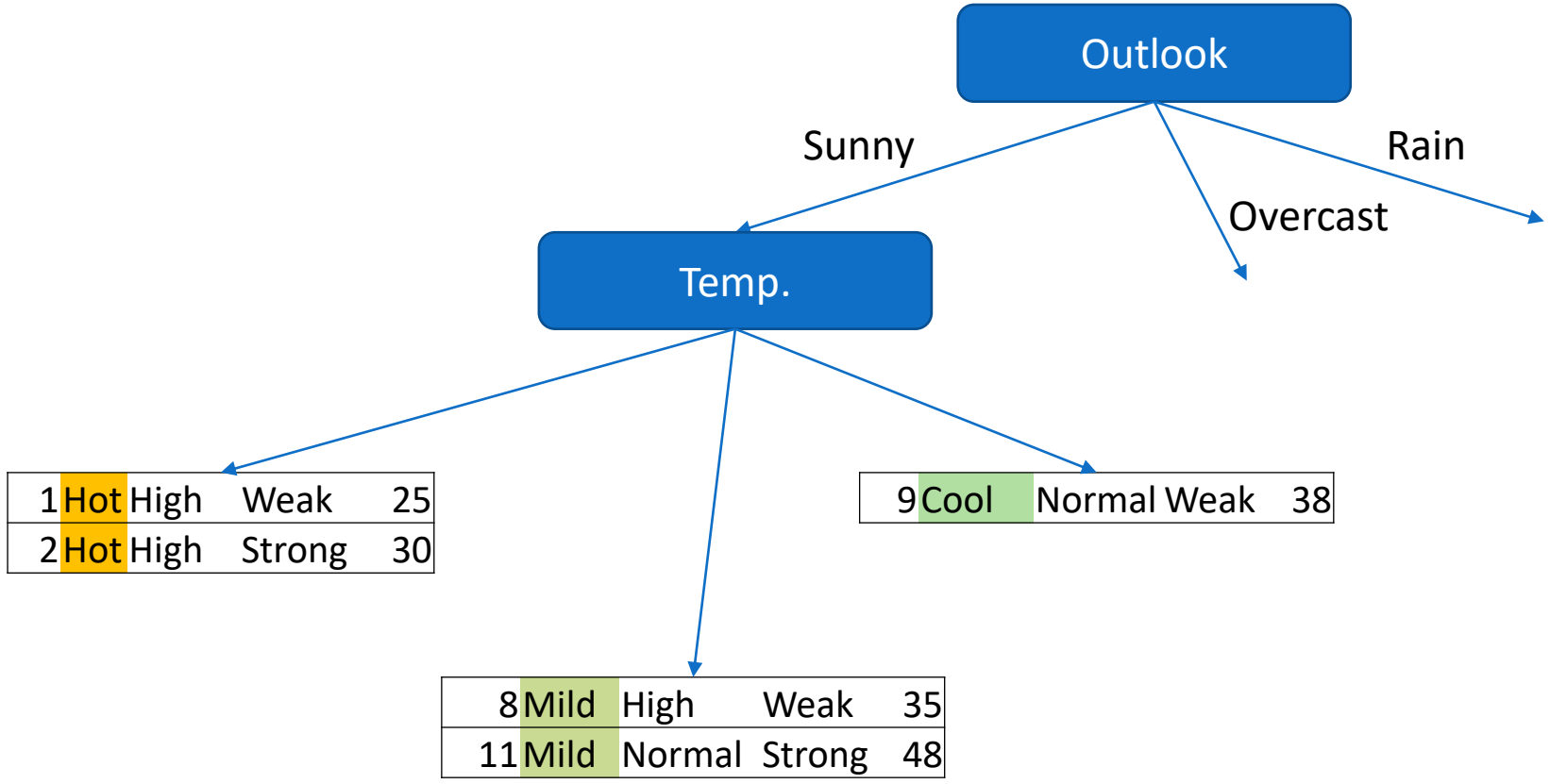
Característica	Reducción Std.Dev
Temp.	4.18
Humidity	3.33
Wind	0.85

Análisis de rama Temp.

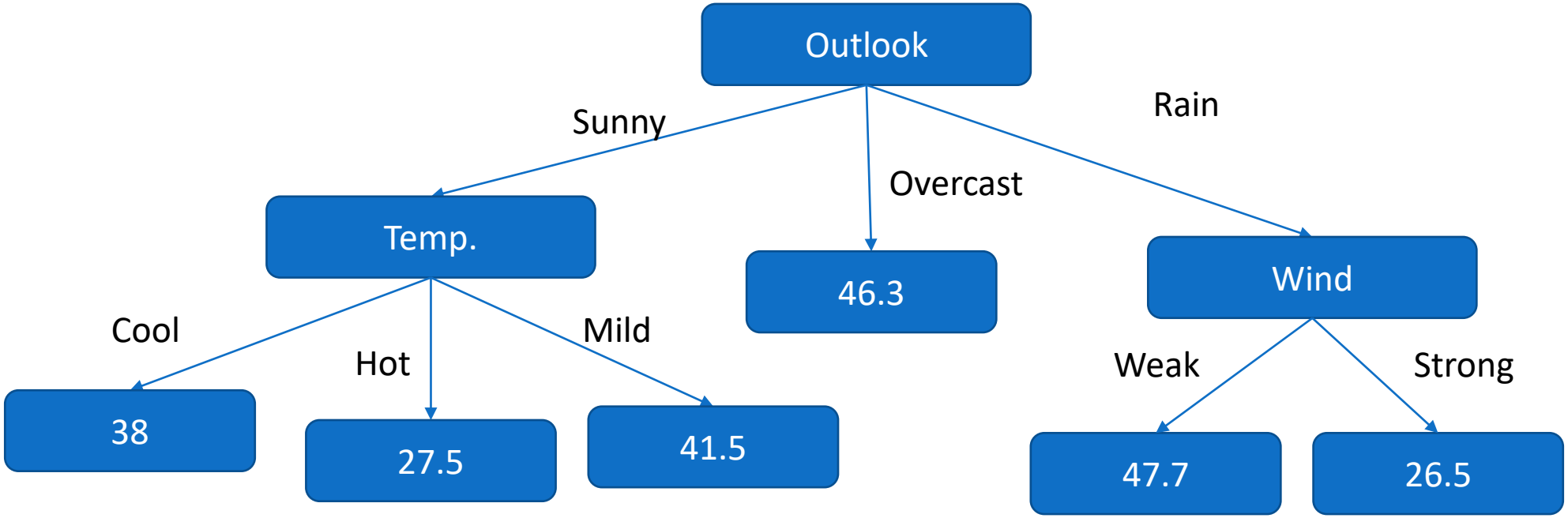
La rama Cool tiene una instancia en su conjunto de datos secundarios. Podemos decir que si Outlook es Sunny y la temperatura es fresca, entonces habría 38 jugadores de golf.

¿Pero qué hay de la rama Hot? Todavía hay 2 instancias. ¿Deberíamos agregar otra rama para viento débil y viento fuerte?
No, no deberíamos. Porque esto provoca un ajuste excesivo.

Deberíamos terminar la construcción de ramas, por ejemplo, si hay menos de cinco instancias en el conjunto de datos secundario. O la desviación estándar del conjunto de datos secundario puede ser inferior al 5% de todo el conjunto de datos.



Forma final del DT



Algoritmos para la generación de DT

- ID3 → Extensión del algoritmo D3
- C4.5 → Es un ID3 mejorado
- CART → Classification and Regression Tree
- CHAID → Detección automática de interacción chi-cuadrado. Realiza divisiones de varios niveles al calcular árboles de clasificación
- MARS → splines de regresión adaptativa multivariante

Algoritmo ID3

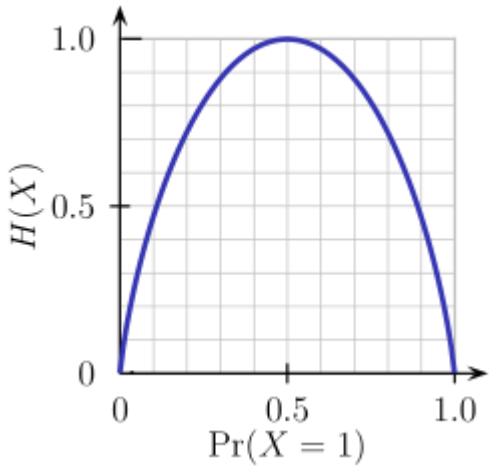
1. Comienza con el conjunto original S como nodo raíz.
2. En cada iteración del algoritmo, recorre el atributo no utilizado del conjunto S y calcula la entropía (H) y la ganancia de información (IG) de este atributo.
3. Luego selecciona el atributo que tiene la menor entropía o la mayor ganancia de información.
4. Luego, el conjunto S se divide por el atributo seleccionado para producir un subconjunto de los datos.
5. El algoritmo continúa recurriendo a cada subconjunto, considerando solo los atributos nunca seleccionados antes.

Medidas para la selección de características

- Entropía
- Information Gain (IG)
- Índice Gini
- Gain Ratio
- Reducción en varianza
- Chi-Square

Entropía (H)

- La entropía es una medida de la aleatoriedad de la información que se procesa. Cuanto mayor sea la entropía, más difícil será sacar conclusiones de esa información.



$$H(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

Play	
Si	No
9	5

H(Play) = H(9, 5)

$= \mathrm{H}\left(\frac{9}{14}, \frac{5}{14}\right) = -\left(\frac{9}{14}\log_2\frac{9}{14}\right) - \left(\frac{5}{14}\log_2\frac{5}{14}\right)$

= 0.94

Entropía para múltiples atributos

$$H(S,X) = \sum_{c \in X} P(c)H(c)$$

		Play		Instancias
		Si	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5
				14

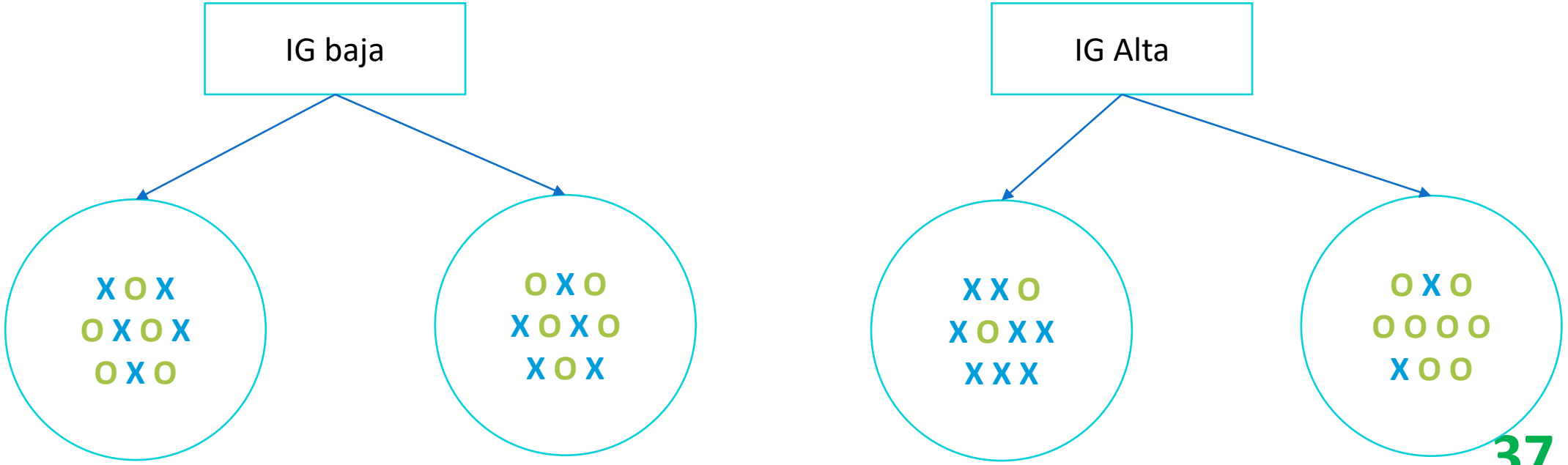
E(Play, Outlook) = P(Sunny)*H(2,3) + P(Overcast)*H(4,0) + P(Rainy)*H(2,3)

$$= \left(\frac{5}{14}\right) * 0.971 + \left(\frac{4}{14}\right) * 0.0 + \left(\frac{5}{14}\right) * 0.971$$

= 0.693

Information Gain

- Propiedad estadística que mide qué tan bien una característica determinada separa las muestras de entrenamiento de acuerdo con su clasificación de destino.



Information Gain

$$IG(S, X) = H(S) - H(S, X)$$

IG(Play, Outlook) = E(Play) – E(Play, Outlook)

= 0.940 – 0.693

= 0.247

$$IG = Entropy(antes) - \sum_{j=1}^{k} Entropy(j, después)$$

Índice de Gini

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2$$

- Gini trabaja con variables target categóricos binarias

- Un valor más alto del índice de Gini implica una mayor desigualdad, una mayor heterogeneidad.

Tasa de Ganacia

$$Gain\ Ratio = \frac{IG}{SplitInfo} \qquad SplitInfo = \sum_{j=1}^{k} w_j \log_2 w_j$$

- Gain Ratio resuelve el problema IG al tener en cuenta el número de ramas que resultarían antes de realizar la división.

Reducción de Varianza

$$Varianza = \frac{\sum(X - \bar{X})^2}{n}$$

- La reducción de la varianza es un algoritmo que se utiliza para las variables *target* continuas (problemas de regresión).
- Este algoritmo utiliza la fórmula estándar de varianza para elegir la mejor división.
- La división con menor varianza se selecciona como criterio para dividir la población.

Chi-Square

$$X^2 = \sum \frac{(O - E)^2}{E}$$

- O ← valor observado
- E ← valor esperado

- Se calcula mediante la suma de cuadrados de las diferencias estandarizadas entre las frecuencias observadas y esperadas de la variable *target*.

Ventajas de los DT

- Si la relación entre las variables dependientes e independientes está bien aproximada por un modelo lineal, la regresión lineal superará al modelo basado en árboles.
- Si hay una alta no linealidad y una relación compleja entre las variables dependientes e independientes, un modelo de árbol superará a un método de regresión clásico.
- Si necesita construir un modelo que sea fácil de explicar a las personas, un modelo de árbol de decisiones siempre funcionará mejor que un modelo lineal.