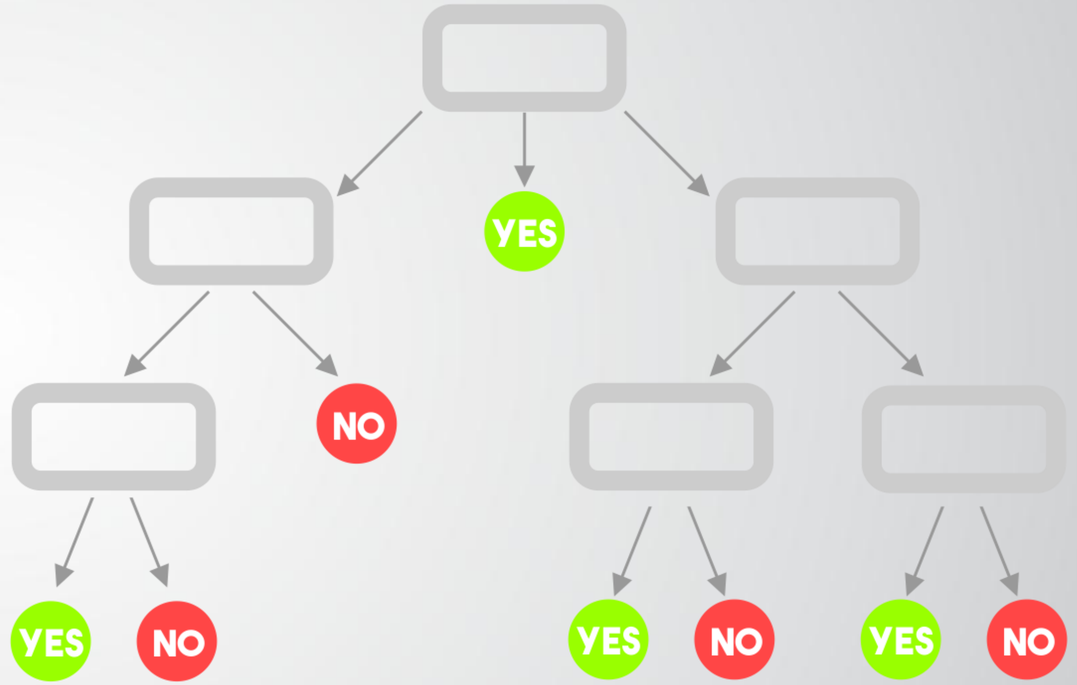
Sdco

/ CANALSANDECO

DECISION TREE

ÁRVORE DE DECISÃO

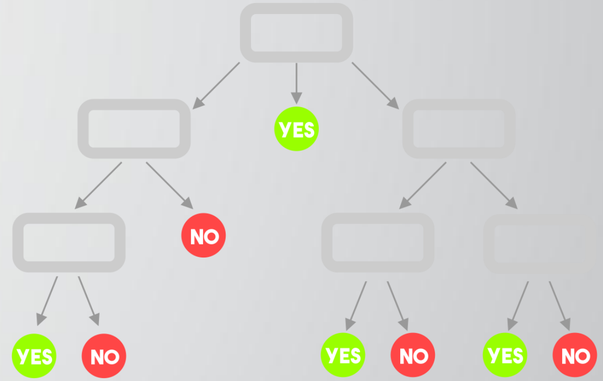


/CANALSANDECO

DECISIONTREE

MÉTODO BASEADO EM ÁRVORES

APRENDIZAGEM DE ÁRVORES DE DECISÃO É UM DOS MÉTODOS DE APRENDIZAGEM MAIS PRÁTICOS E MAIS UTILIZADOS PARA A APRENDIZAGEM INDUTIVA.

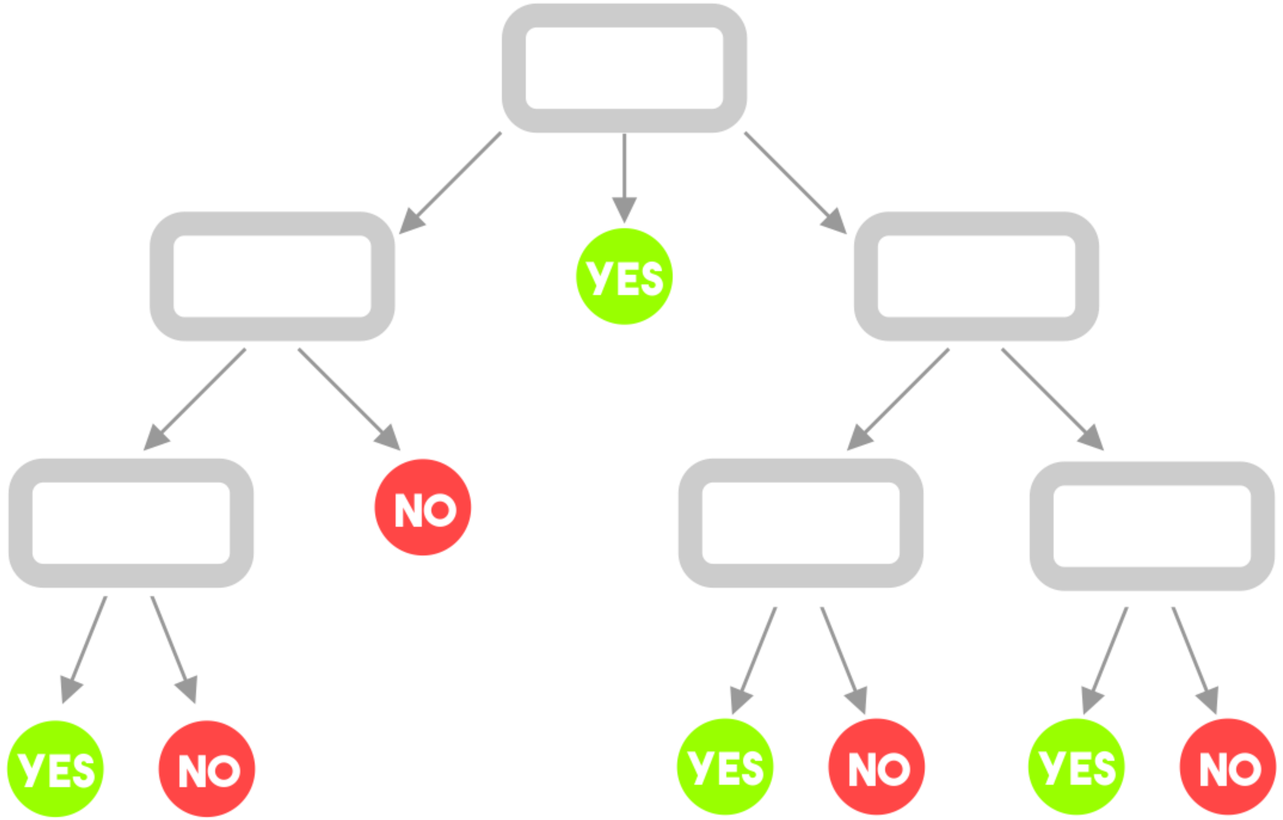


CLASSIFICAÇÃO

DECISIONTREE

CARACTERÍSTICAS

O CONHECIMENTO
É BASEADO
EM ÁRVORES

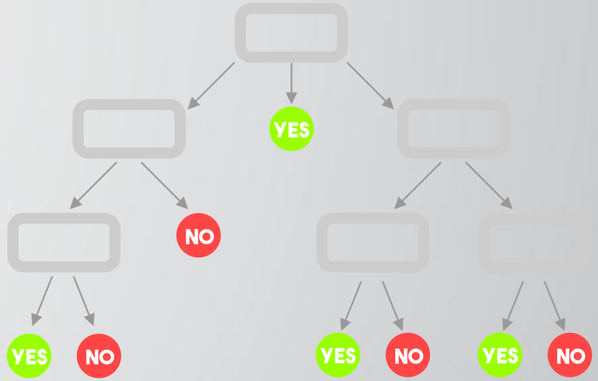


CLASSIFICAÇÃO

DECISIONTREE

CARACTERÍSTICAS

ALGORITMOS DE ÁRVORES DE DECISÃO E REGRAS ADQUIREM CONHECIMENTO SIMBÓLICO A PARTIR DE DADOS DE TREINAMENTO



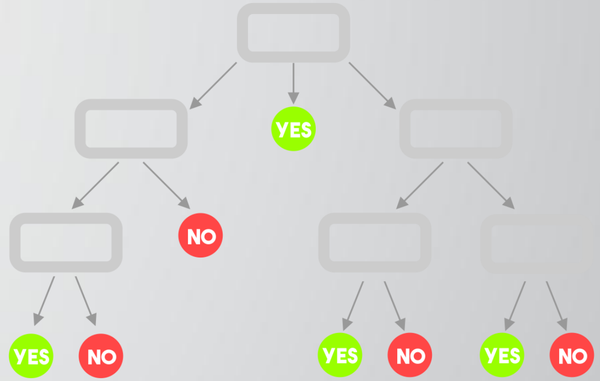
CLASSIFICAÇÃO

DECISIONTREE

CARACTERÍSTICAS

AMPLA CLASSE DE ALGORITMOS DE APRENDIZADO

EXEMPLO: ID3, C4.5, CART,...



CLASSIFICAÇÃO

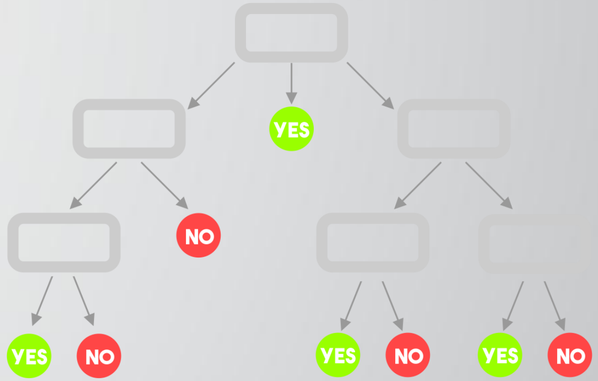
DECISIONTREE

CARACTERÍSTICAS

UTILIZA UM *VÍES INDUTIVO:*

PREFERÊNCIA POR ÁRVORES

MENORES.

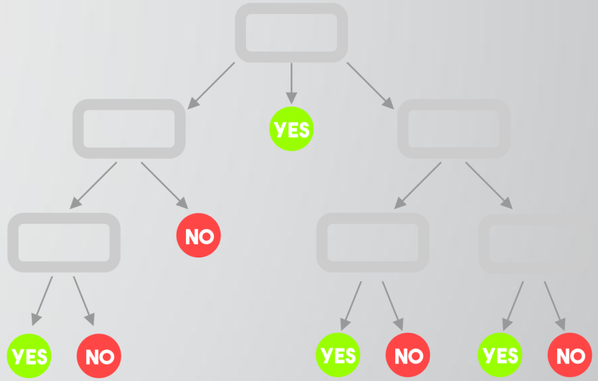


CLASSIFICAÇÃO

DECISIONTREE

CARACTERÍSTICAS

ÁRVORES DE DECISÃO TAMBÉM PODEM SER REPRESENTADAS COMO CONJUNTOS DE REGRAS ***SE–ENTÃO (IF–THEN).***



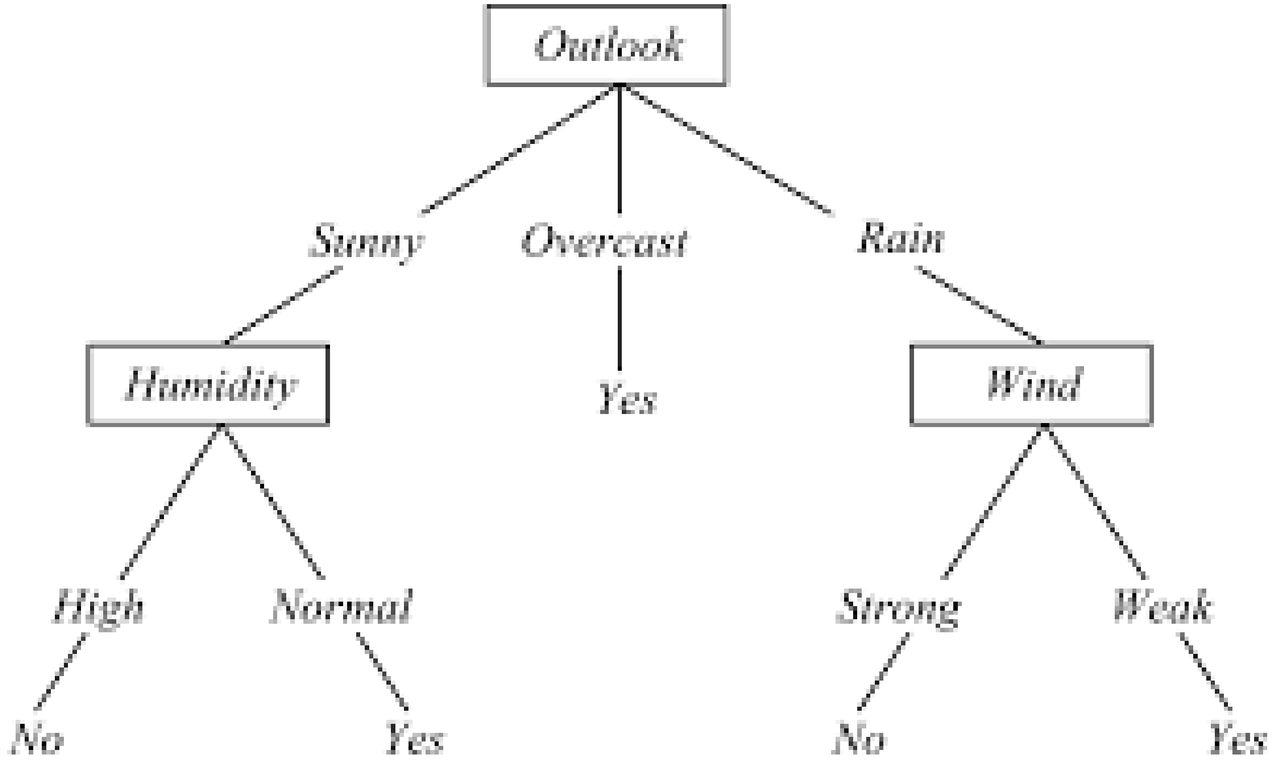
CLASSIFICAÇÃO

DECISIONTREE

EXEMPLO

VAMOS JOGAR

TÊNIS?



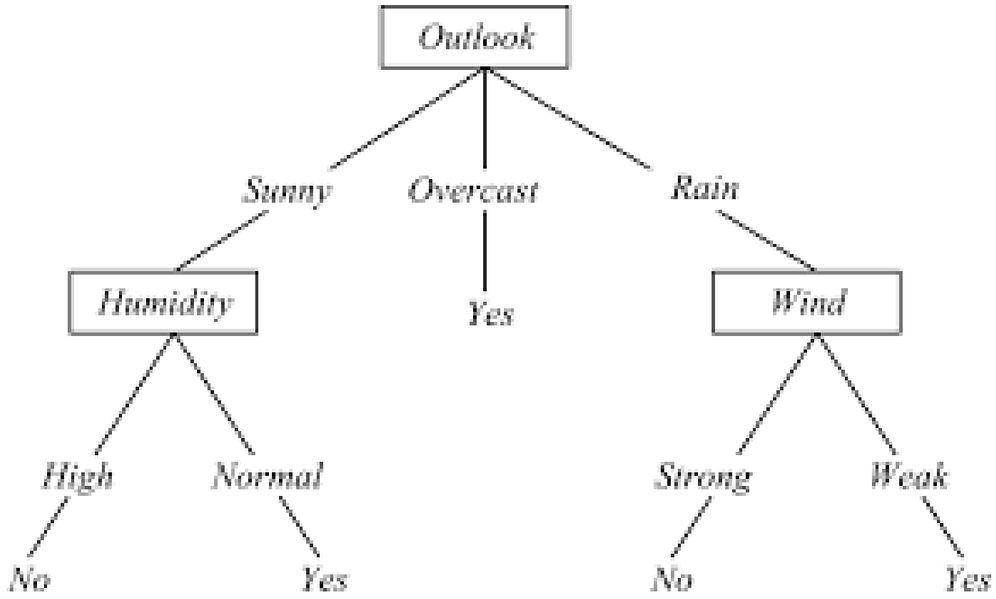
SÓ JOGAREMOS QUANDO

(Outlook=Sunny && Humidity=Normal)
|| (Outlook=Overcast)
|| (Outlook=Rain && Wind=Weak)

DECISIONTREE

REPRESENTAÇÃO

- CADA NÓ INTERNO TESTA UM ATRIBUTO
- CADA RAMO CORRESPONDE AO VALOR DO ATRIBUTO
- CADA FOLHA ATRIBUI UMA CLASSIFICAÇÃO

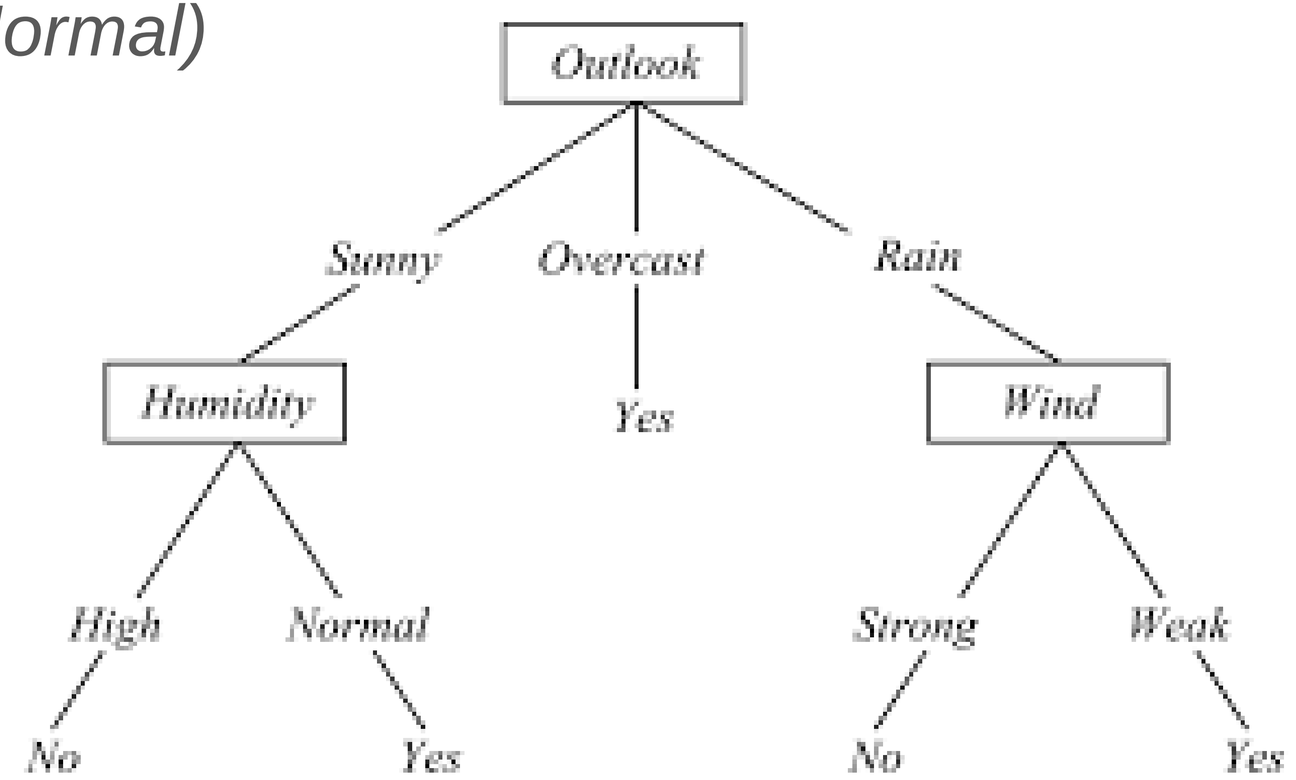


CLASSIFICAÇÃO

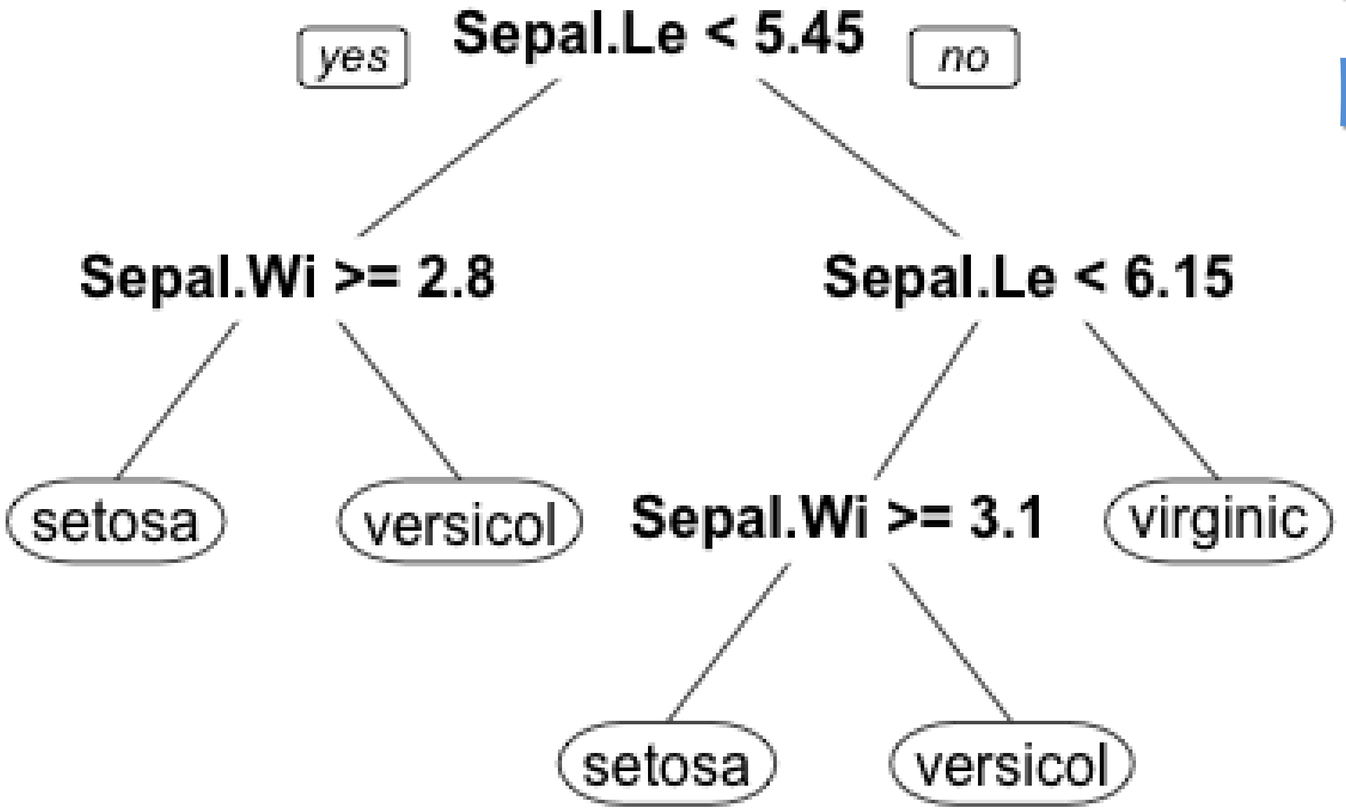
DECISIONTREE

SÓ JOGAREMOS QUANDO

(*Outlook=Sunny* ***AND*** *Humidity=Normal)*
OR
(*Outlook=Overcast)*
OR
(*Outlook=Rain* ***AND*** *Wind=Weak)*



DECISIONTREE



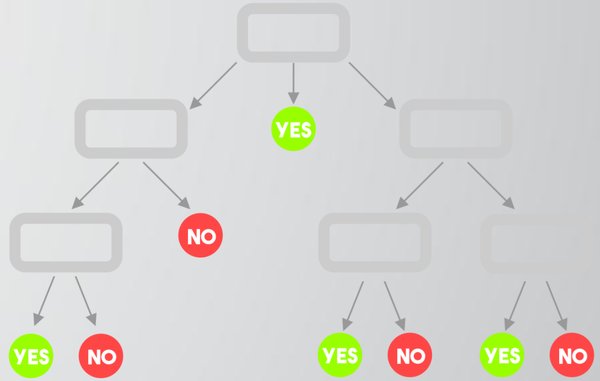
IRIS FLOWER PROBLEMS

CLASSIFICAÇÃO

DECISIONTREE – ID3

ID3

ALGORITMO BÁSICO PARA
APRENDIZAGEM
DE ÁRVORES DE DECISÃO



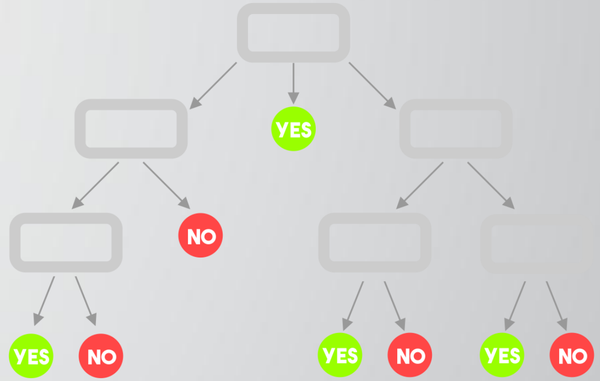
CANALSANDECO

CLASSIFICAÇÃO

DECISIONTREE – ID3

O ALGORITMO *ID3* "*APRENDE*" *ÁRVORES DE DECISÃO* CONSTRUINDO–AS DE CIMA PARA BAIXO (*TOP–DOWN*).

COMEÇANDO COM SEGUINTE QUESTÃO:



CLASSIFICAÇÃO

QUEM É O ATRIBUTO RAIZ?

/CANALSANDECO

O QUE TIVER MELHOR GANHO DE INFORMAÇÃO

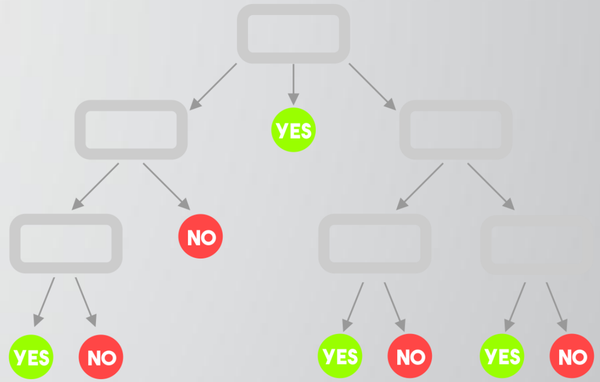
/CANALSANDECO

DECISIONTREE – ID3

ENTROPIA

NÍVEL DE BAGUNÇA

CARACTERIZA A (IM)PUREZA DE UMA COLEÇÃO ARBITRÁRIA DE EXEMPLOS.



/CANALSANDECO

CLASSIFICAÇÃO

ENTROPIA

Dado uma coleção *S contendo exemplos + e – de algum* conceito alvo, a *entropia de S relativa a esta* classificação booleana é:

$$Entropia\ (S) \equiv -p_+ \log_2 p_+ - p_- \log_2 p_-$$

p+ é a proporção de exemplos positivos em S
p– é a proporção de exemplos negativos em S

EXEMPLO

Exemplo: Sendo *S* *uma coleção de 14 exemplos* de algum conceito booleano, incluindo 9 exemplos positivos e 5 negativos [9+, 5–].

Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

ENTROPIA - EXEMPLO

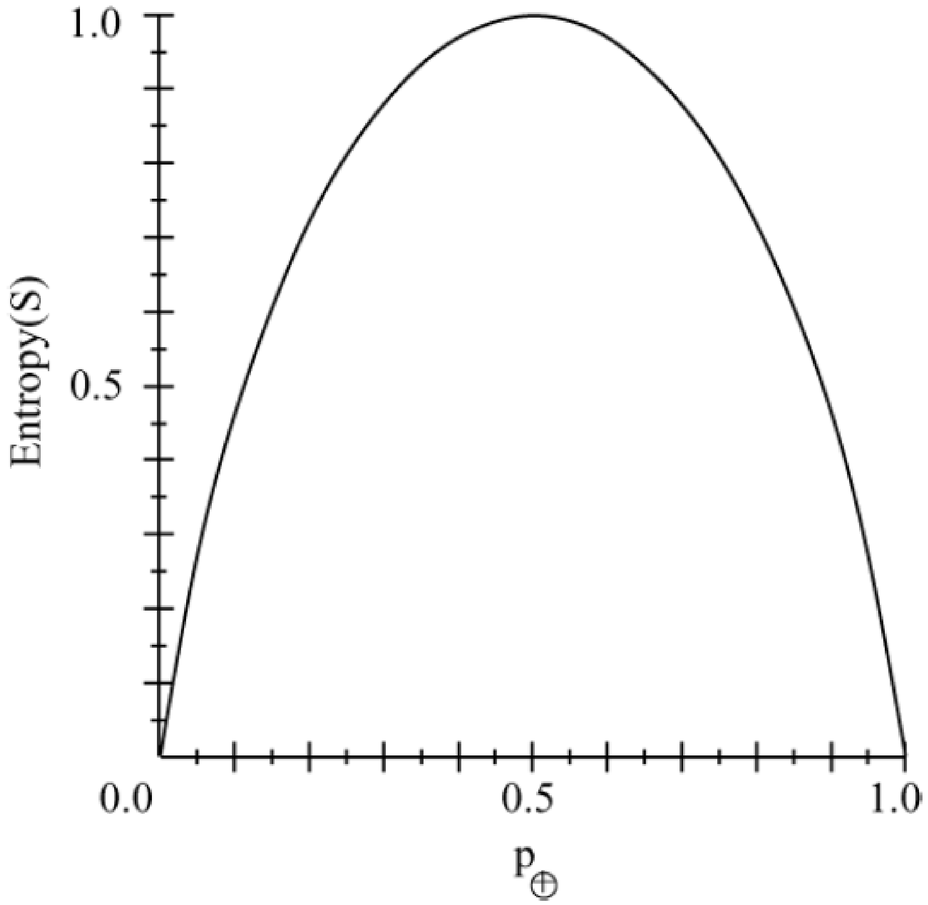
Exemplo: Sendo *S uma coleção de 14 exemplos* de algum conceito booleano, incluindo 9 exemplos positivos e 5 negativos [9+, 5–].

$$Entropia\ ([9+,5-]) = -(\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14})$$

$$= 0.940$$

ENTROPIA - VARIAÇÃO

A função *entropia relativa a uma classificação* booleana, como a proporção, *p+ de exemplos* positivos varia entre 0 e 1.



ENTROPIA - MULTICLASSES

Generalizando para o caso de um atributo alvo aceitar *c diferentes valores, a entropia de S* relativa a esta classificação *c–classes é definida* como:

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

onde *pi é a proporção de S pertencendo a classe i.*

AGORA

O GANHO DE INFORMAÇÃO

GANHO DE INFORMAÇÃO

Redução esperada na entropia devido a ordenação sobre *A*, *ou seja, a redução* esperada na entropia causada pela partição dos exemplos de acordo com este atributo *A.*

$$Gain\,(S,A) \equiv Entropia\ (S) - \sum_{v \in Valores\ (A)} \frac{|S_v|}{|S|} Entropia\ (S_v)$$

DECISIONTREE

EXEMPLO

Vamos calcular o ganho de informação do atributo *Wind*

Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

DECISIONTREE

Todo conjunto

S [9+,5-]

WIND

Weak [6+,2-]

Strong [3+,3-]

Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

DECISIONTREE

GANHO DE NFORMAÇÃO

Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

$$Values(Wind) = Weak, Strong$$

$$S = [9+, 5-]$$

$$S_{Weak} \leftarrow [6+, 2-]$$

$$S_{Strong} \leftarrow [3+, 3-]$$

$$Gain(S, Wind) = Entropy(S) - \sum_{v \in \{Weak, Strong\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

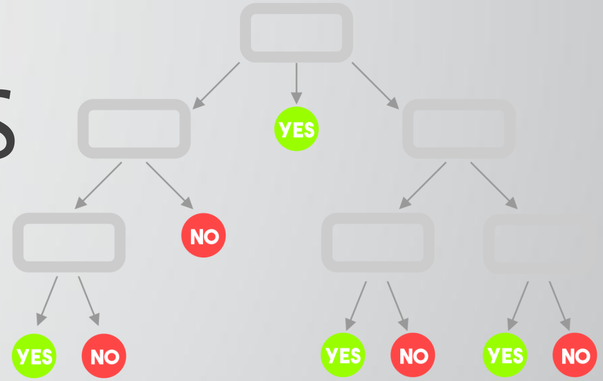
$$= Entropy(S) - (8/14) Entropy(S_{Weak}) - (6/14) Entropy(S_{Strong})$$

$$= 0.940 - (8/14)0.811 - (6/14)1.00$$

$$= 0.048$$

DECISIONTREE – ID3

COMO O OBJETIVO DE ENCONTRAR A **RAIZ DA ÁRVORE**, CALCULAMOS O GANHO DE INFORMAÇÃO PARA TODOS OS ATRIBUTOS DO DATASET



CLASSIFICAÇÃO

GANHO DE INFORMAÇÃO

Gain (S, Outlook) = 0.246

Gain(S, Humidity) = 0.151

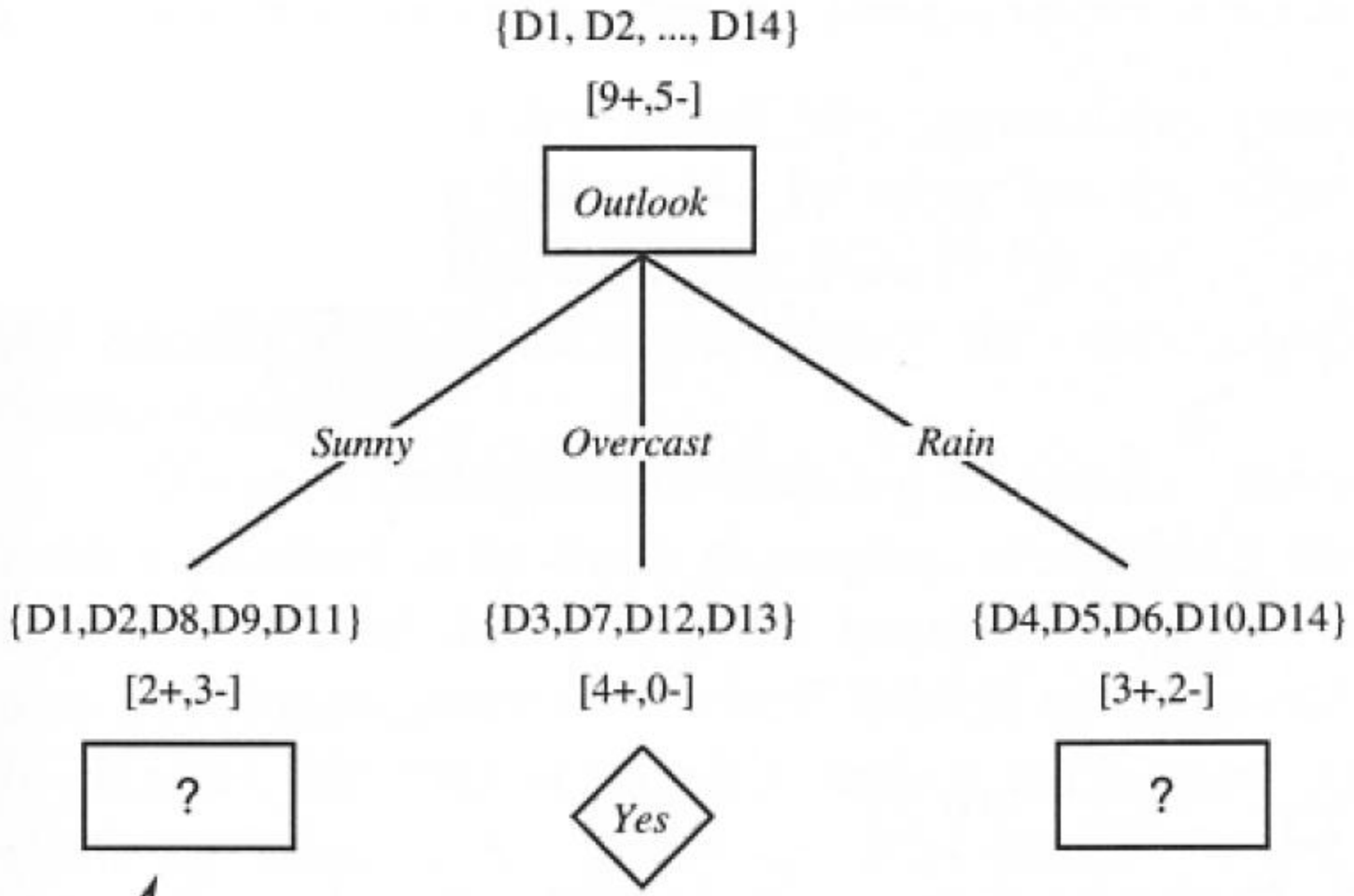
Gain (S, Wind) = 0.048

Gain (S, Temperature) = 0.029

Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

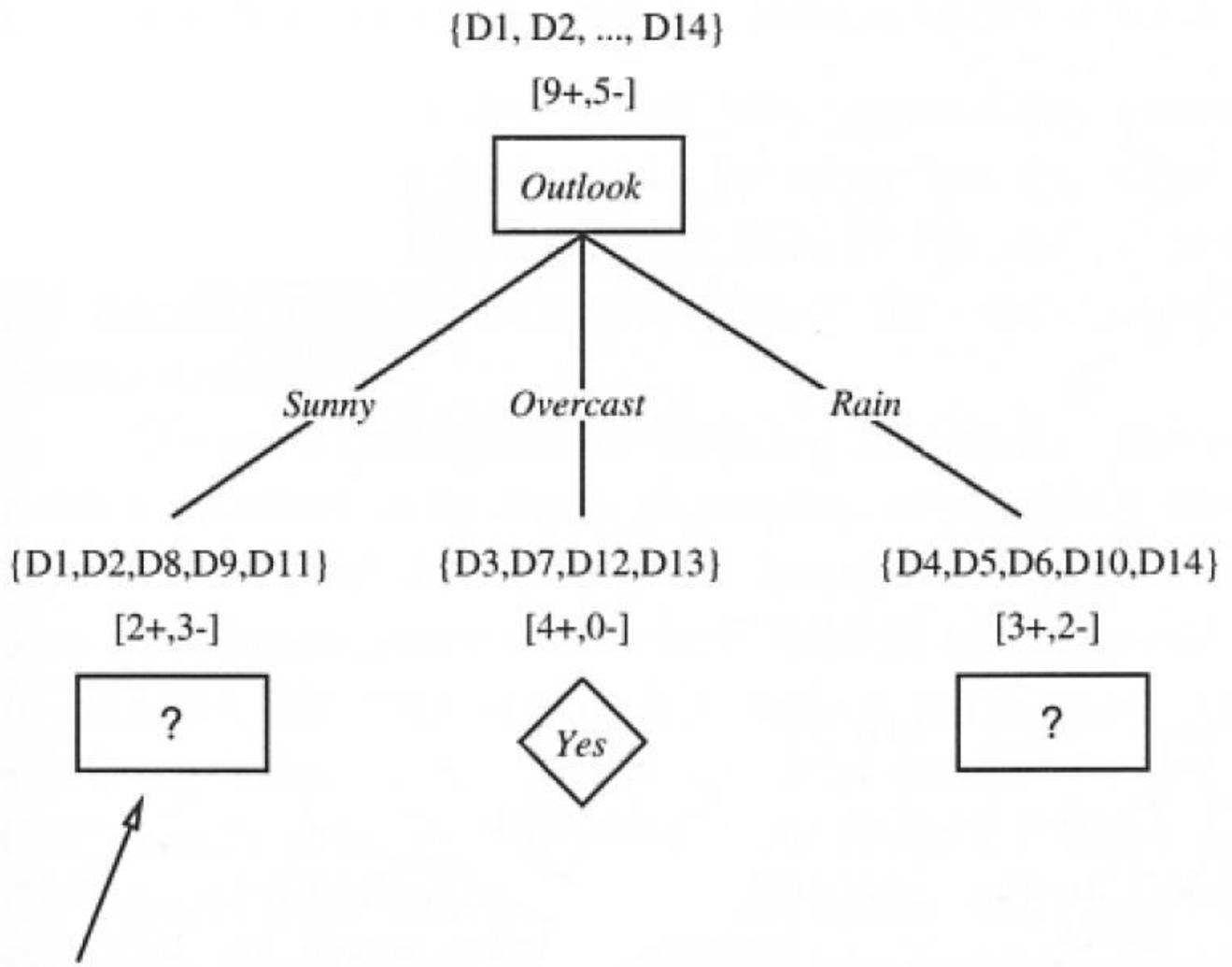
RAIZ DA ÁRVORE

Montar a raiz da árvore e distribuir seus ramos.



QUE ATRIBUTO VEM AQUI?

DECISIONTREE



GANHO DE INFORMAÇÃO

LEVANDO EM CONSIDERAÇÃO SOMENTE

OUTLOOK(Sunny)

$S_{sunny} = \{D1,D2,D8,D9,D11\}$

$Gain\ (S_{sunny}, Humidity) = .970 - (3/5)\ 0.0 - (2/5)\ 0.0 + .970$

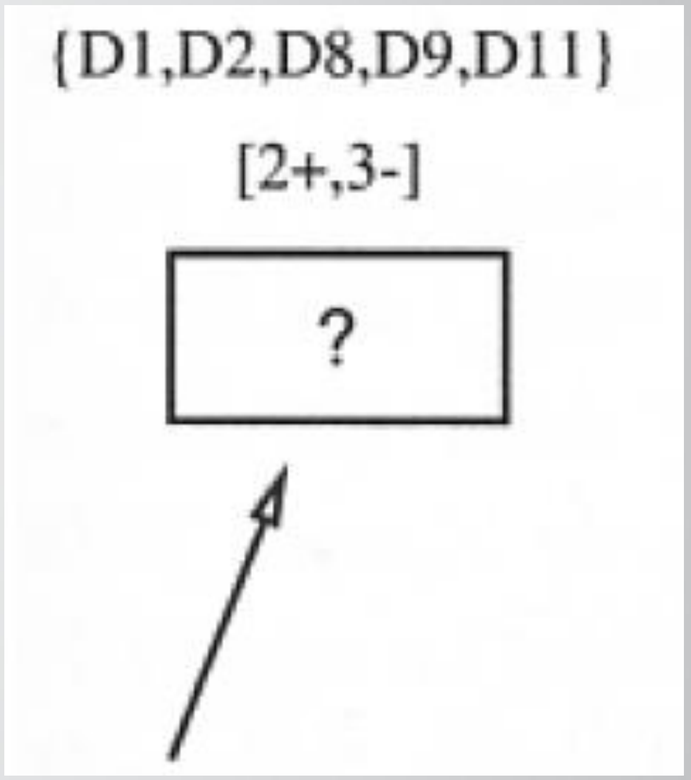
$Gain\ (S_{sunny}, Temperature) = .970 - (2/5)\ 0.0 - (2/5)\ 1.0 - (1/5)\ 0.0 = .570$

$Gain\ (S_{sunny}, Wind) = .970 - (2/5)\ 1.0 - (3/5)\ .918 = .019$

DECISIONTREE – ID3

RECURSIVAMENTE

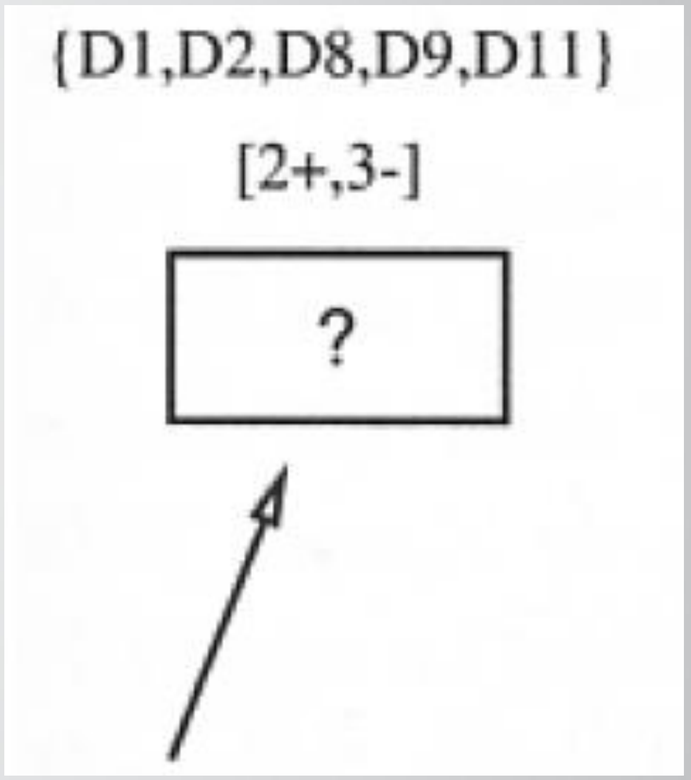
REPETE-SE O PROCESSO DE SELECIONAR UM NOVO ATRIBUTO E PARTICIONAR OS EXEMPLOS DE TREINAMENTO É REPETIDO PARA CADA NÓ DESCENDENTE NÃO TERMINAL.



DECISIONTREE – ID3

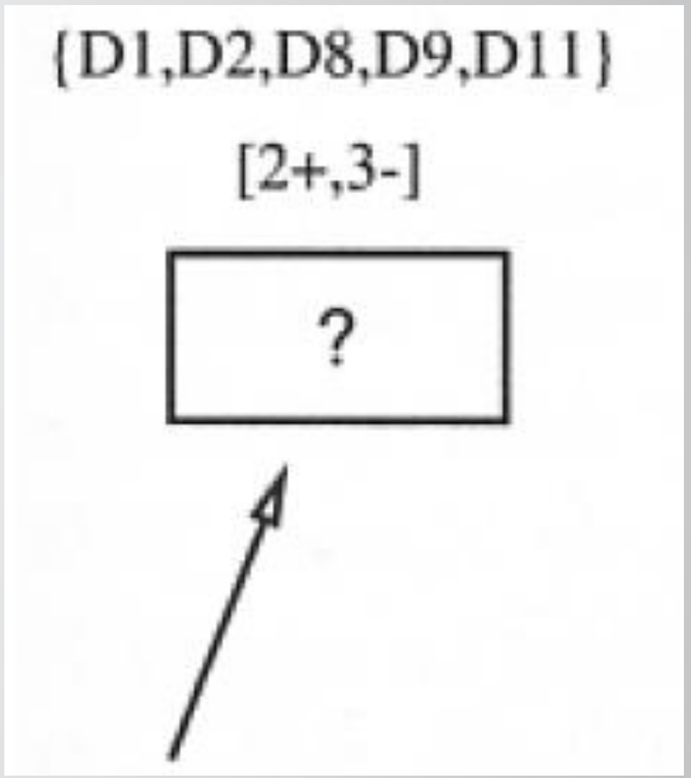
RECURSIVAMENTE

SÃO UTILIZADOS SOMENTE OS EXEMPLOS DE TREINAMENTO ASSOCIADOS COM ESTE NÓ.



DECISIONTREE – ID3

ATRIBUTOS QUE FORAM INCORPORADOS ANTERIORMENTE A ÁRVORE SÃO EXCLUÍDOS. QUALQUER ATRIBUTO DEVE APARECER SOMENTE UMA VEZ AO LONGO DE QUALQUER CAMINHO NA ÁRVORE.



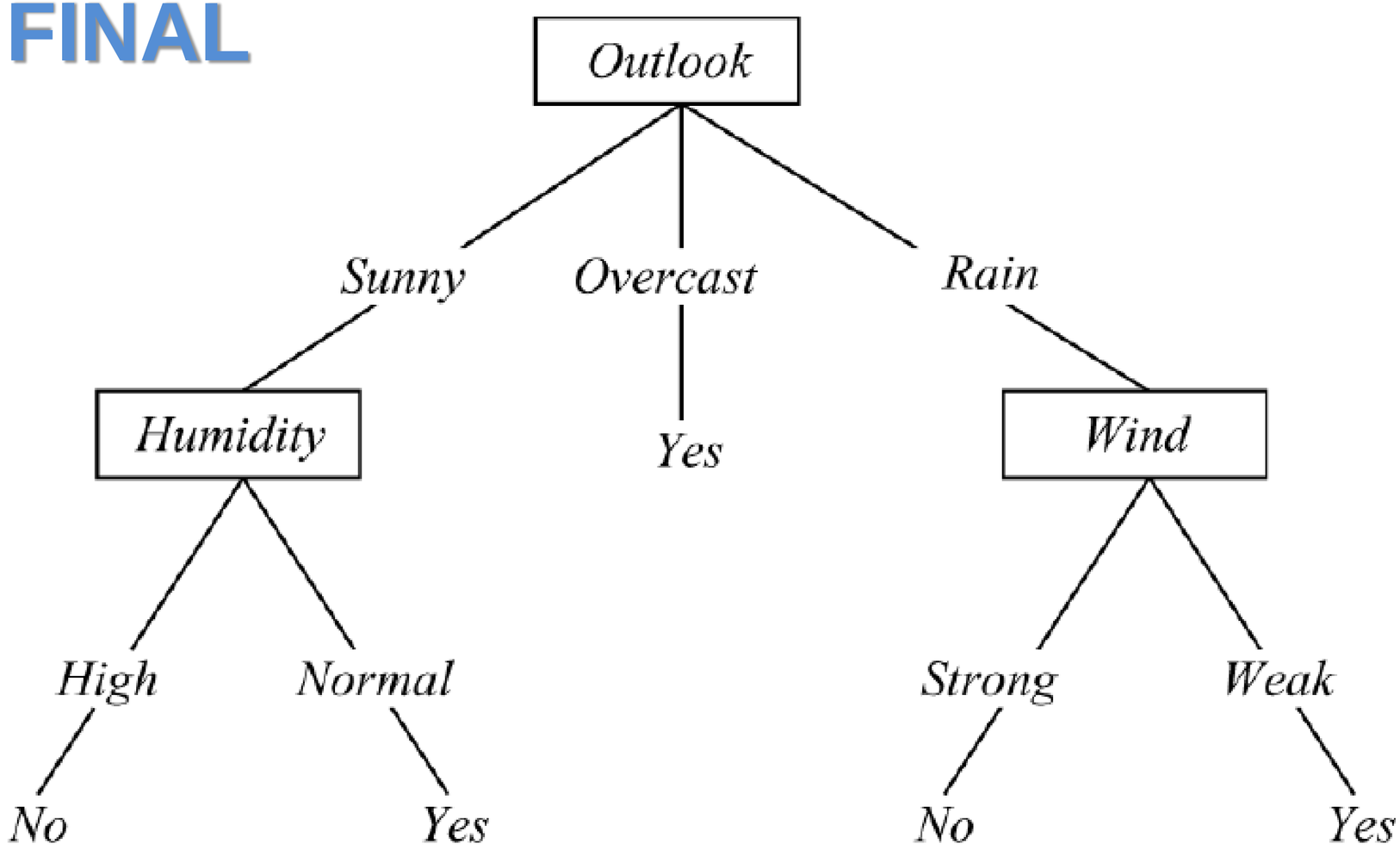
ÁRVORE - Recursivamente

Para cada ramo da árvore o processo continua até que uma das seguintes condições seja atendida:

1. Todos os atributos já estejam incluídos ao longo deste caminho da árvore;

2. Os exemplos de treinamento associados com este nó folha tenham todos o mesmo valor de atributo alvo.

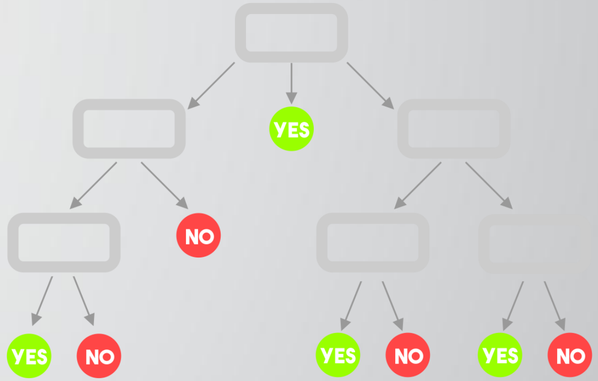
ÁRVORE FINAL



DECISIONTREE – ID3

ESPAÇO DE HIPÓTESES

O MÉTODO DE APRENDIZAGEM ***ID3*** *PODE SER CARACTERIZADO* COMO UM MÉTODO DE BUSCA EM UM ESPAÇO DE HIPÓTESES, POR UMA HIPÓTESE QUE SE AJUSTA AOS EXEMPLOS DE TREINAMENTO.

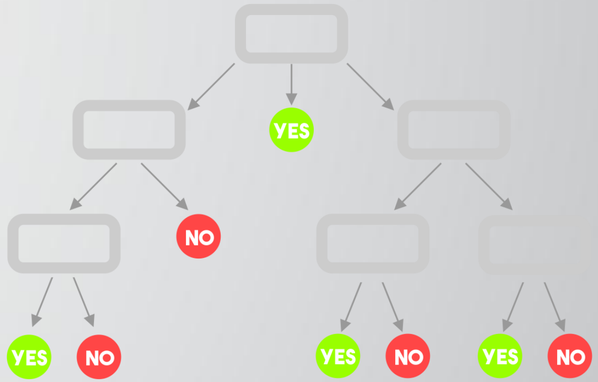


CLASSIFICAÇÃO

DECISIONTREE – ID3

ESPAÇO DE HIPÓTESES

O ESPAÇO DE HIPÓTESES BUSCADO PELO *ID3 É O CONJUNTO DE* ÁRVORES DE DECISÃO POSSÍVEIS.



CLASSIFICAÇÃO

SCIKIT LEARN

```
from sklearn import tree

clf = tree.DecisionTreeClassifier()

clf.fit(Features, classes)

clf.predict(new_object)
```

http://scikit-learn.org/stable/modules/tree.html