Machine Learning

Professor Helder Vieira

Tópicos da primeira aula

1. Apresentação e visão geral
2. Conversa sobre Machine Learning
3. Árvores de decisão: intuição
4. Entropia e Gini
5. Mão na massa!

O professor

Cientista de Dados no Itaú Unibanco

Formado em Engenharia Elétrica pela UNICAMP

Especialização em Inteligência Artificial na Universidade de Bologna, Itália

Louco por matemática

São Paulino ☹

Conteúdo do módulo

24/11	Árvores
26/11	KNN
29/11	RandomForest
01/12	Boosting
03/12	Pipelines
06/12	Feature Selection + CV
08/12	Exercícios 1
10/12	Mini Projeto 1
13/12	MLP
15/12	RNN
17/12	LSTM
10/01	Bag Of Words
12/01	TF IDF
14/01	Word2vec
17/01	Exercícios 2
19/01	Mini Projeto 2
21/01	Kmeans e DBScan
26/01	Discussões

O que é Machine Learning?

Dados

Fonte primária e material que subsidia todo o trabalho do cientista.

Análise

Etapa de construção das hipóteses e de entendimento do negócio.

Modelo

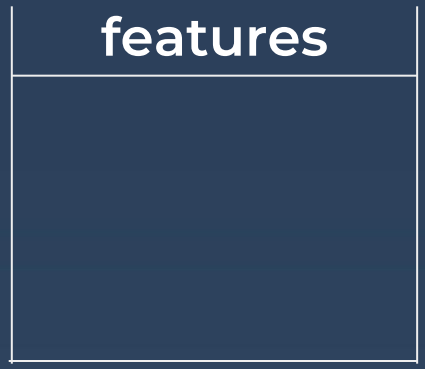
Abstração dos fenômenos reais num ambiente simplificado e matemático.

Produto

O que gera valor para o negócio.


interpretabilidade
Árvore de decisão
Regressão Logística
LSTM
capacidade de predição

Não supervisionado



- Metodologias em desenvolvimento
+ Dados abundantes
- Resultados instáveis

Supervisionado

+ Modelos mais avançados
+ Desenvolvimento mais simples
+ Resultados consistentes
- Dados escassos

Supervisionado

features	target
	0
	1
	1
	0

Target categórico:
CLASSIFICAÇÃO

features	target
	15
	32
	18
	47

Target numérico:
REGRESSÃO

Supervisionado

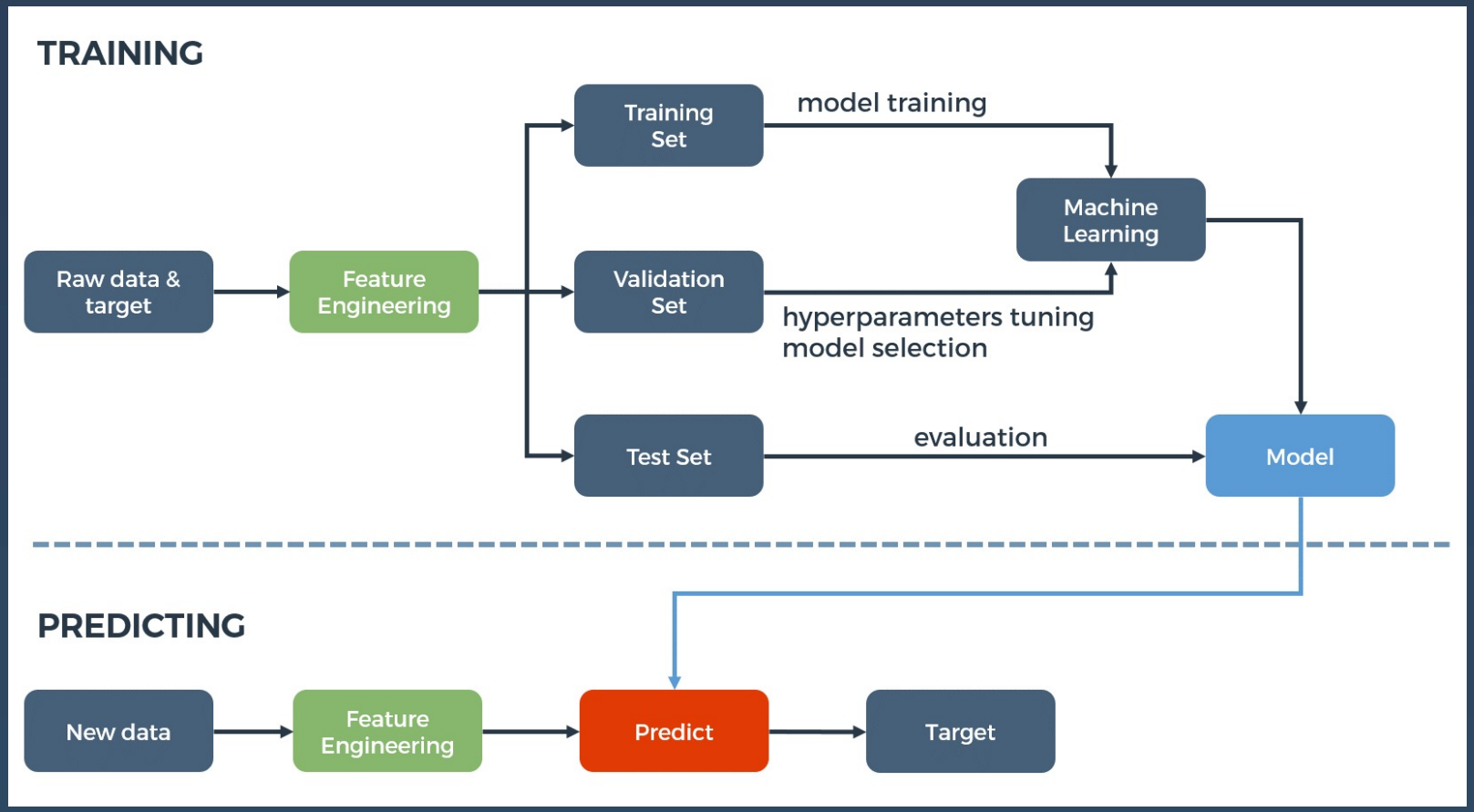
features	target
	0.2
	0.8
	0.7
	0.1

Target categórico:
CLASSIFICAÇÃO

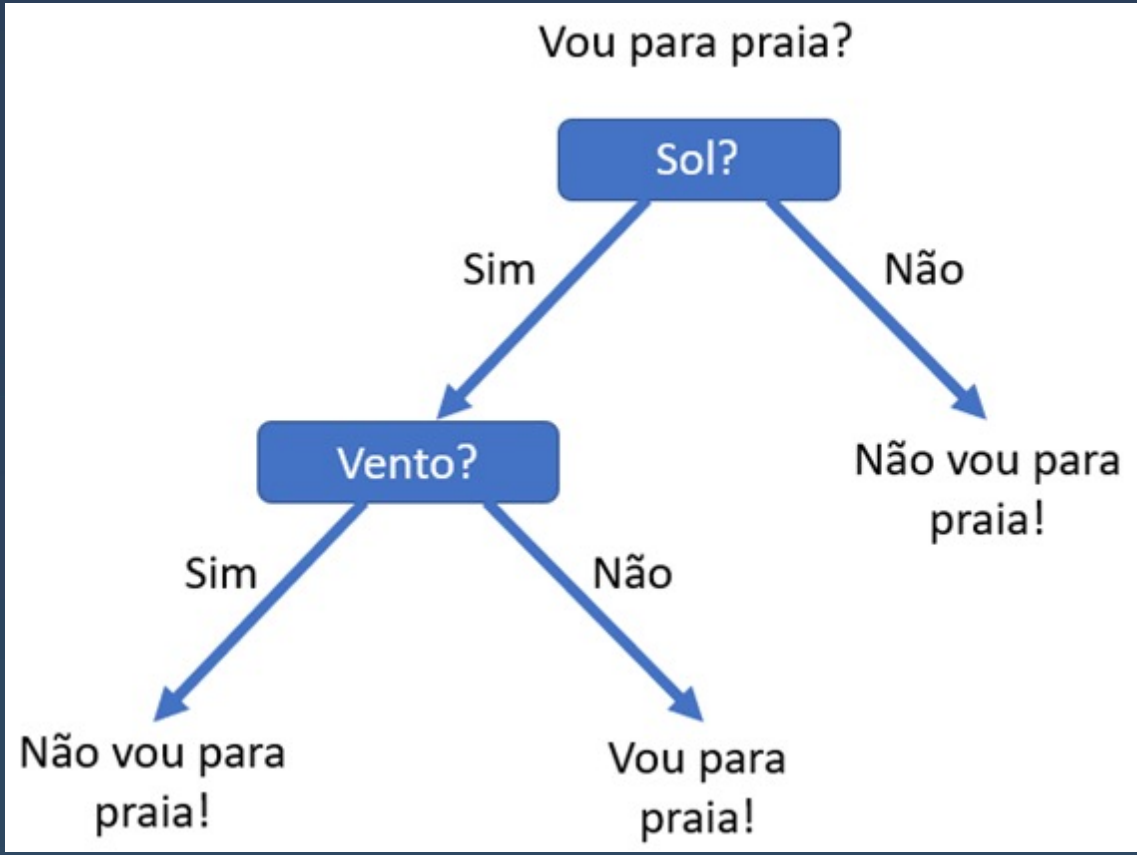
features	target
	15
	32
	18
	47

Target numérico:
REGRESSÃO

Avaliação


TRAINING
Raw data & target
Feature Engineering
Training Set
Validation Set
Test Set
model training
Machine Learning
hyperparameters tuning
model selection
evaluation
Model
PREDICTING
New data
Feature Engineering
Predict
Target

Árvore de decisão: intuição


Vou para praia?
Sol?
Sim
Não
Vento?
Não vou para praia!
Sim
Não
Não vou para praia!
Vou para praia!

temp	chuva	dia	jogo
14	sim	seg	não
28	não	ter	sim
.	.	.	.
.	.	.	.
.	.	.	.
24	não	qui	sim
11	sim	sab	não

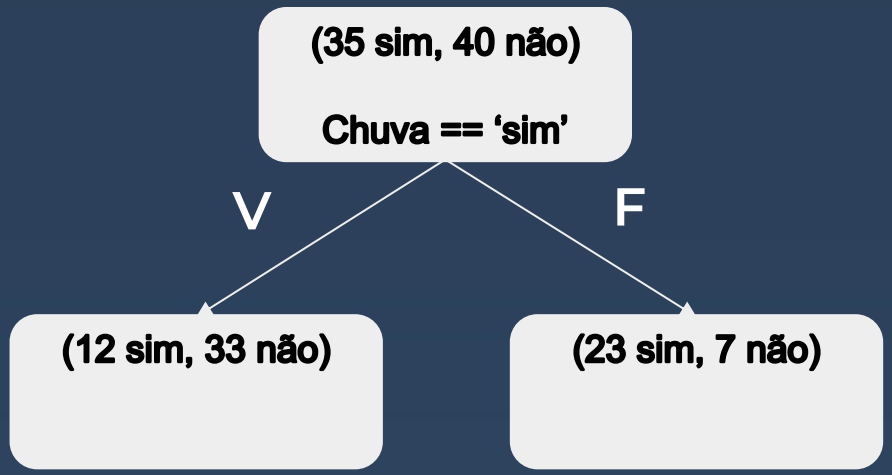
(35 sim, 40 não)

temp	chuva	dia	jogo
14	sim	seg	não
28	não	ter	sim
.	.	.	.
.	.	.	.
.	.	.	.
24	não	qui	sim
11	sim	sab	não

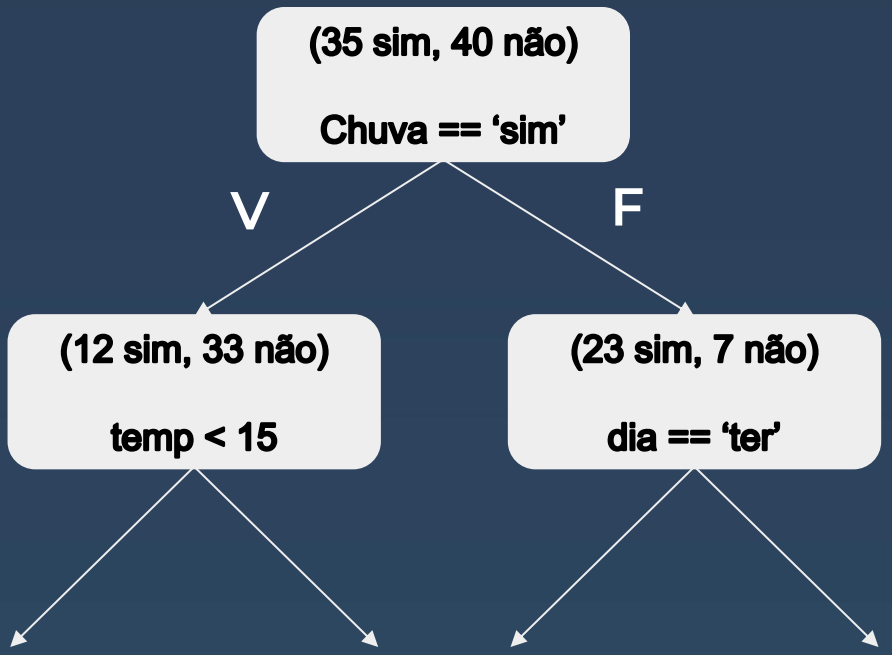
(35 sim, 40 não)

Chuva == ‘sim’

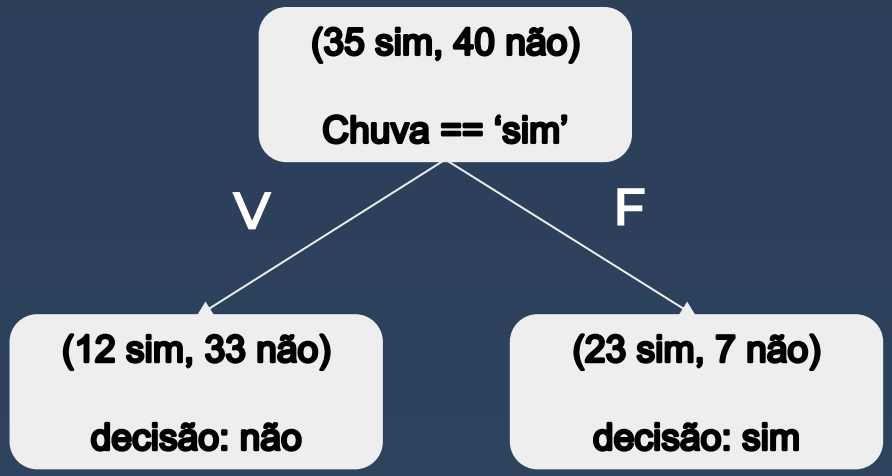
temp	chuva	dia	jogo
14	sim	seg	não
28	não	ter	sim
.	.	.	.
.	.	.	.
.	.	.	.
24	não	qui	sim
11	sim	sab	não



temp	chuva	dia	jogo
14	sim	seg	não
28	não	ter	sim
.	.	.	.
.	.	.	.
.	.	.	.
24	não	qui	sim
11	sim	sab	não



temp	chuva	dia	jogo
14	sim	seg	não
28	não	ter	sim
.	.	.	.
.	.	.	.
.	.	.	.
24	não	qui	sim
11	sim	sab	não



Árvore de decisão: entropia e gini

Medindo a homogeneidade

temp	chuva	dia	jogo
14	sim	seg	não
28	não	ter	sim
.	.	.	.
.	.	.	.
.	.	.	.
24	não	qui	sim
11	sim	sab	não

(35 sim, 40 não)

Gini

$$gini = 1 - \sum_j p_j^2$$

$$gini = 1 - (\frac{35}{35+40})^2 - (\frac{45}{35+40})^2 = 0.42$$

Medindo a homogeneidade

temp	chuva	dia	jogo
14	sim	seg	não
28	não	ter	sim
.	.	.	.
.	.	.	.
.	.	.	.
24	não	qui	sim
11	sim	sab	não

(35 sim, 40 não)

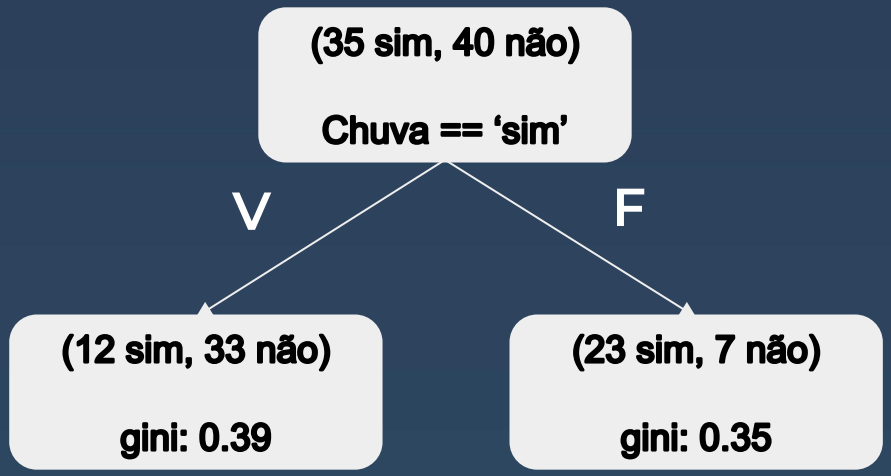
Entropia

$$entropia = \sum_j -p_j log_2 p_j$$

$$entropia = -\frac{35}{75} log_2 \frac{35}{75} - \frac{40}{75} log_2 \frac{40}{75} = 0.99$$

Medindo a homogeneidade

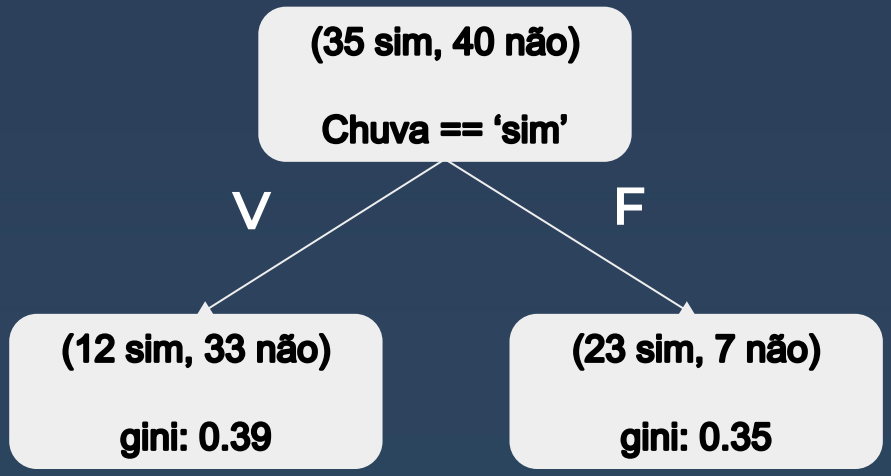
temp	chuva	dia	jogo
14	sim	seg	não
28	não	ter	sim
.	.	.	.
.	.	.	.
.	.	.	.
24	não	qui	sim
11	sim	sab	não



gini ponderado:
((12+33)*0.39 + (23+7)*0.35)/75 = 0.37

Medindo a homogeneidade: escolha da variável

temp	chuva	dia	jogo
14	sim	seg	não
28	não	ter	sim
.	.	.	.
.	.	.	.
.	.	.	.
24	não	qui	sim
11	sim	sab	não

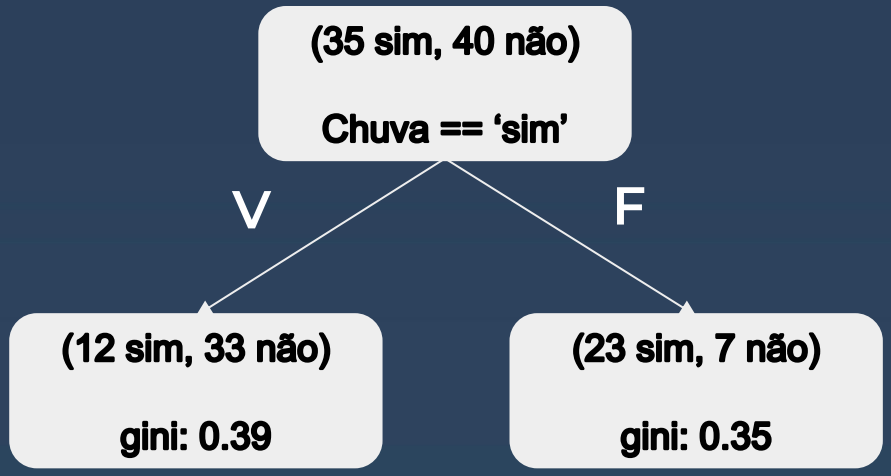


gini ponderado:
((12+33)*0.39 + (23+7)*0.35)/75 = 0.37

variável	gini pond.
temp	0.41
chuva	0.37
dia	0.40

Medindo a homogeneidade: escolha da variável

temp	chuva	dia	jogo
14	sim	seg	não
28	não	ter	sim
.	.	.	.
.	.	.	.
.	.	.	.
24	não	qui	sim
11	sim	sab	não

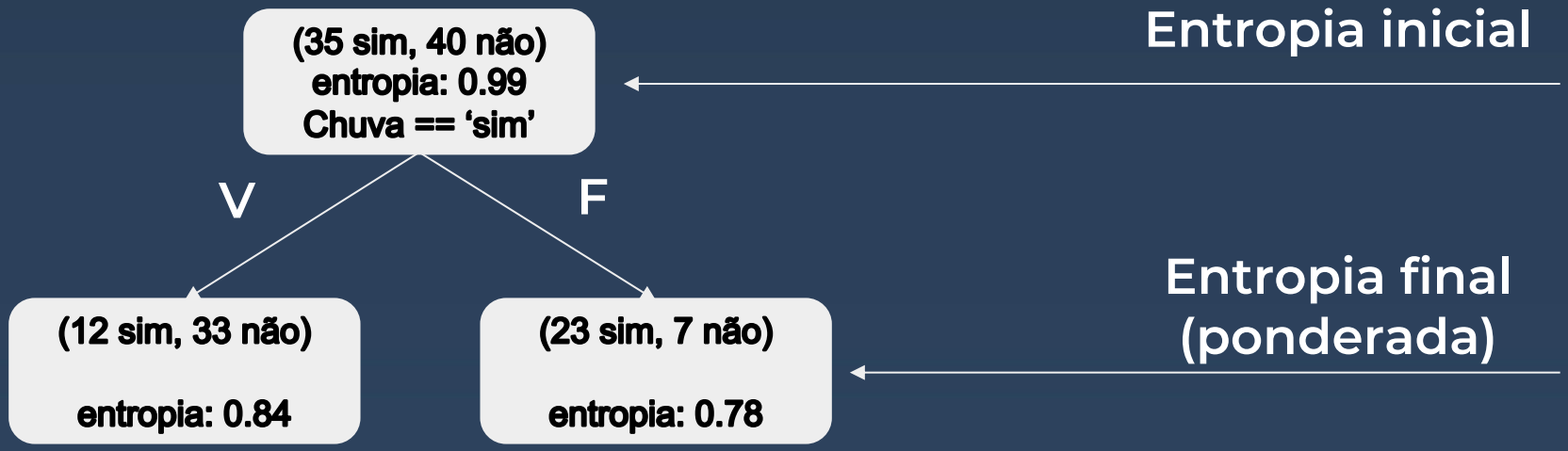


gini ponderado:
((12+33)*0.39 + (23+7)*0.35)/75 = 0.37

variável	gini pond.
temp	0.41
chuva	0.37
dia	0.40

Escolhe o menor!

Medindo a homogeneidade: escolha da variável



Ganho de informação = entropia inicial – entropia final
Ganho de informação = 0.17

variável	ganho inf.
temp	0.10
chuva	0.17
dia	0.14

Escolhe o maior!