FIAP

MBA

MBA EM DATA SCIENCE & AI

STATISTICS WITH R

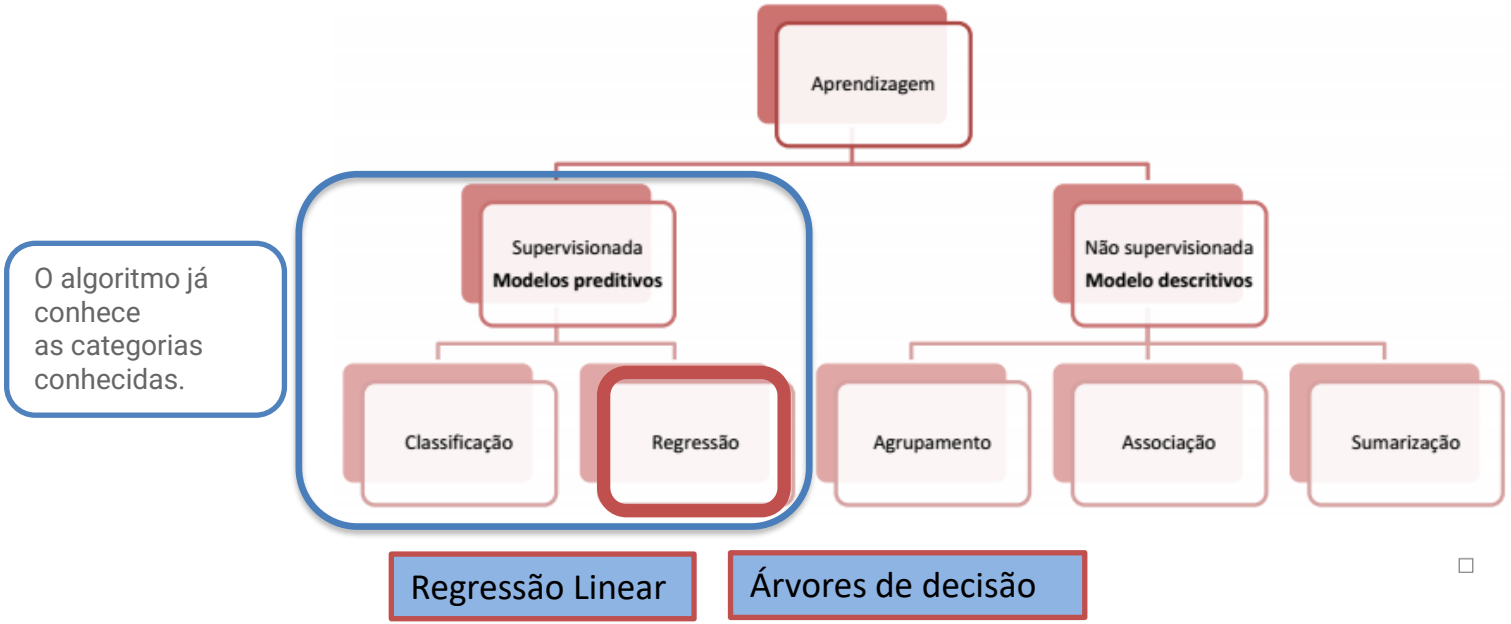
AULA 6

Árvores de decisão

ÁRVORES DE DECISÃO

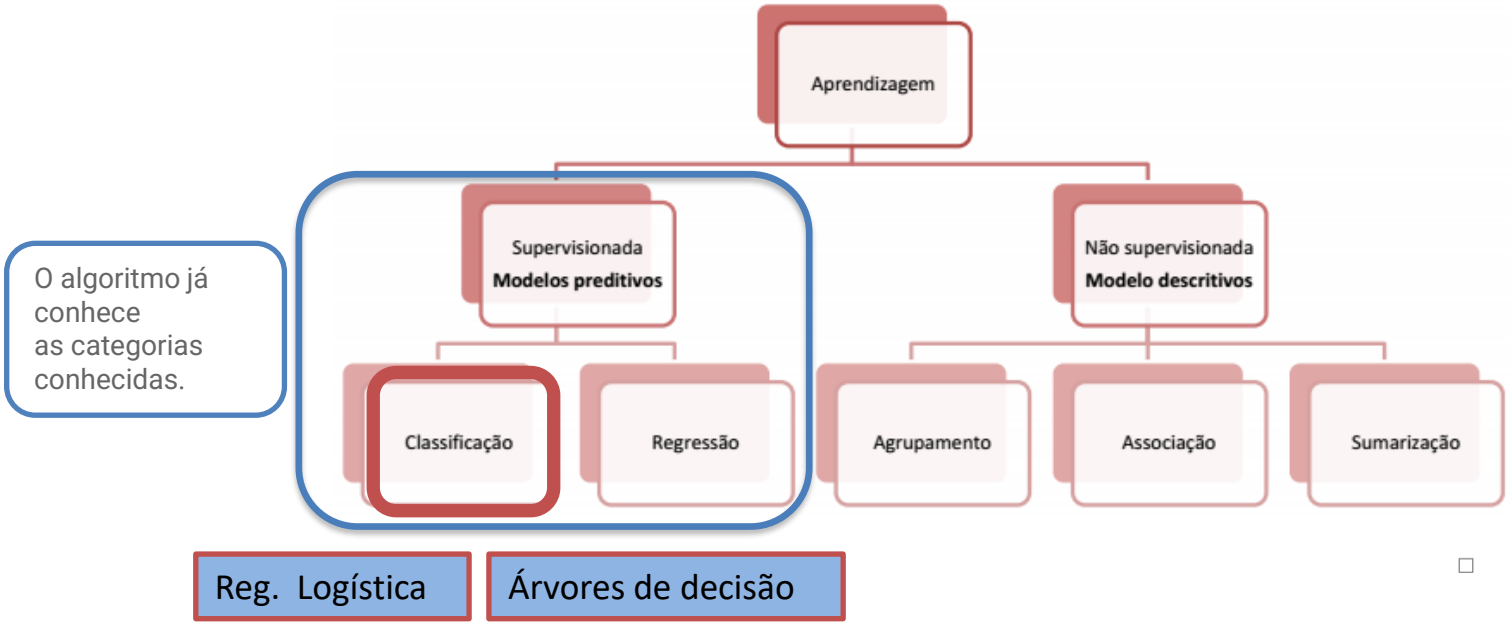
Aprendizagem supervisionada

As técnicas aprendizagem de máquina envolvem diversas finalidades, podendo ser supervisionadas ou não supervisionadas.

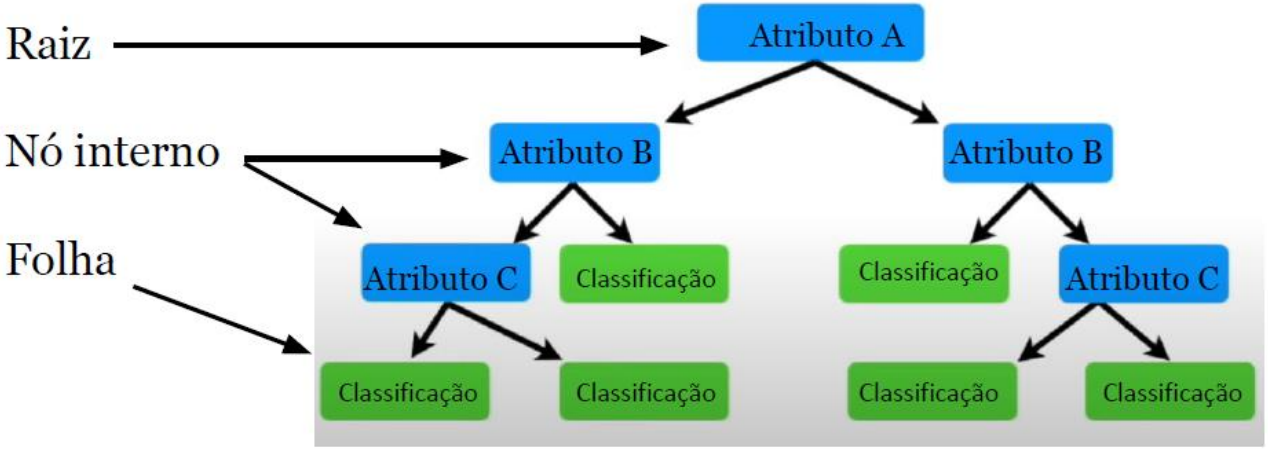


Aprendizagem supervisionada

As técnicas aprendizagem de máquina envolvem diversas finalidades, podendo ser supervisionadas ou não supervisionadas.



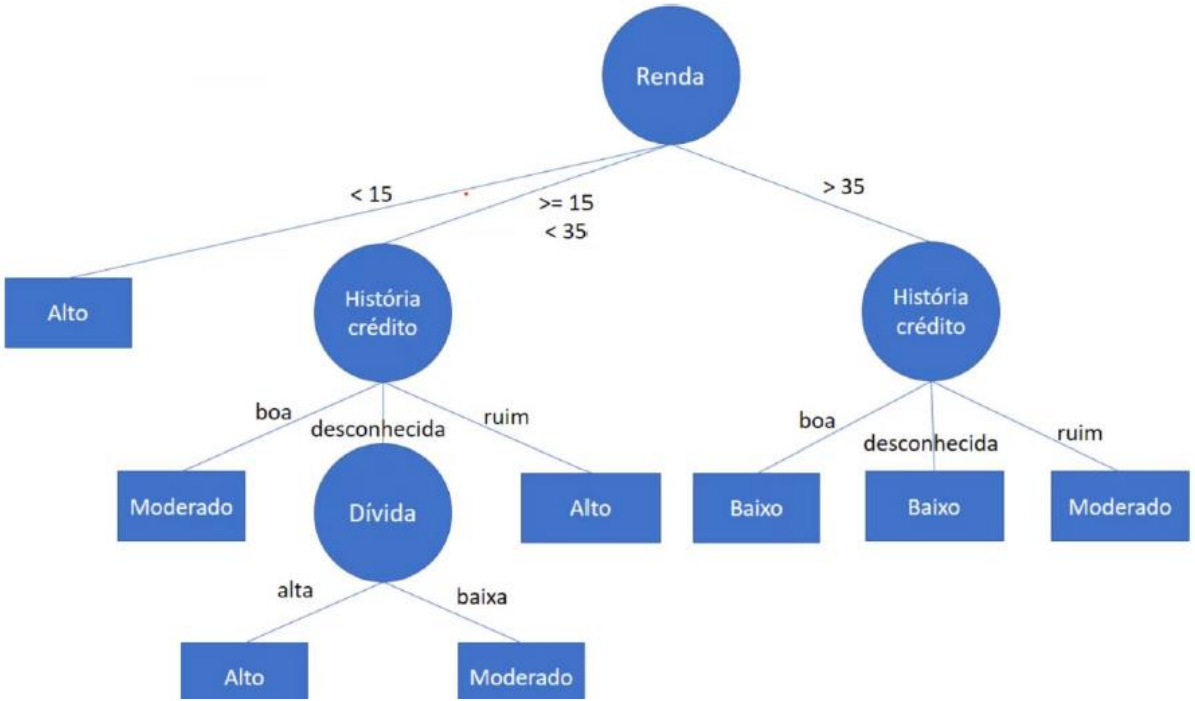
Árvores de Decisão



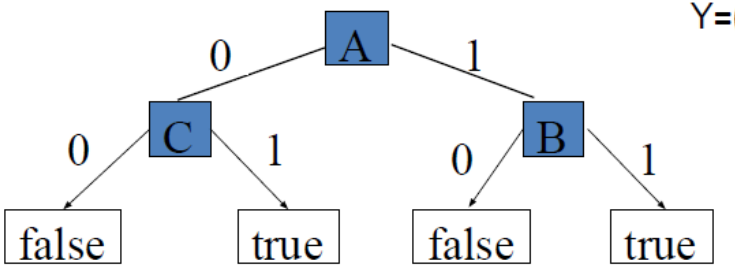
Base de Crédito

História do crédito	Dívida	Garantias	Renda anual	Risco
Ruim	Alta	Nenhuma	< 15.000	Alto
Desconhecida	Alta	Nenhuma	>= 15.000 a <= 35.000	Alto
Desconhecida	Baixa	Nenhuma	>= 15.000 a <= 35.000	Moderado
Desconhecida	Baixa	Nenhuma	> 35.000	Alto
Desconhecida	Baixa	Nenhuma	> 35.000	Baixo
Desconhecida	Baixa	Adequada	> 35.000	Baixo
Ruim	Baixa	Nenhuma	< 15.000	Alto
Ruim	Baixa	Adequada	> 35.000	Moderado
Boa	Baixa	Nenhuma	> 35.000	Baixo
Boa	Alta	Adequada	> 35.000	Baixo
Boa	Alta	Nenhuma	< 15.000	Alto
Boa	Alta	Nenhuma	>= 15.000 a <= 35.000	Moderado

Anatomia da Árvore

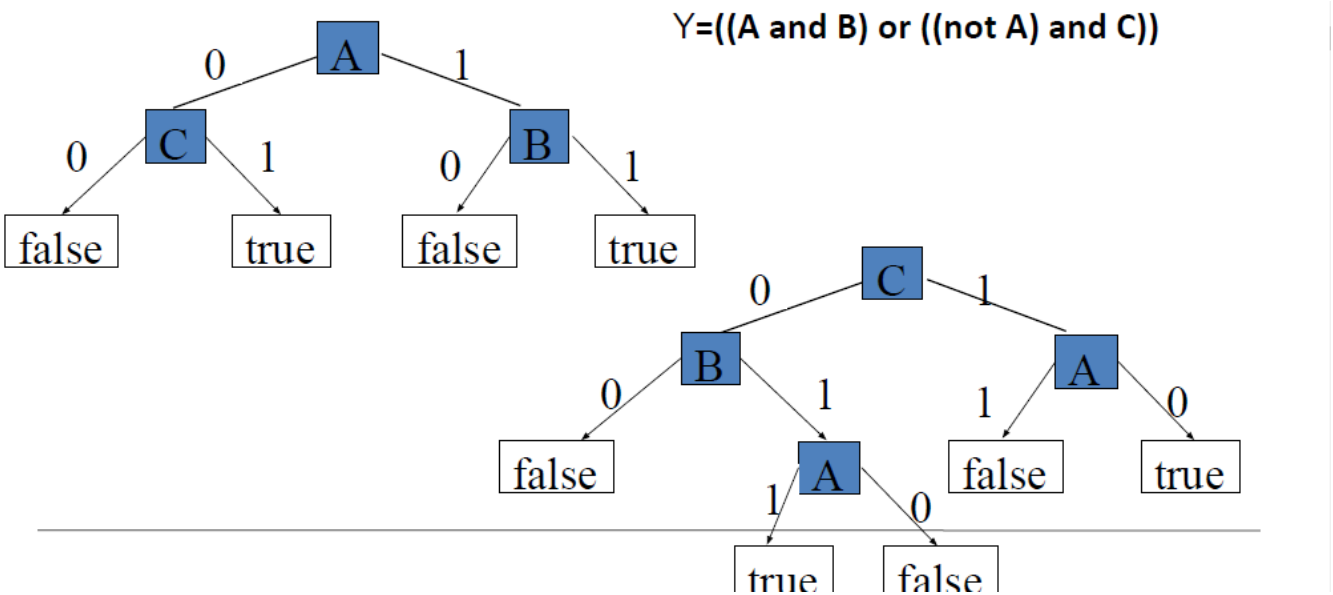


Mesmo Conceito



Y=((A and B) or ((not A) and C))

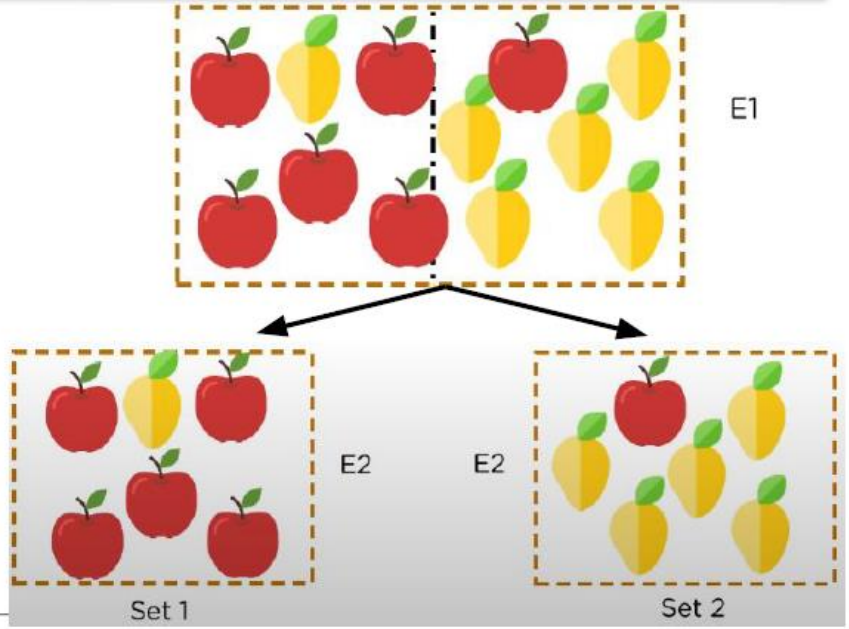
Mesmo Conceito, diferentes representações



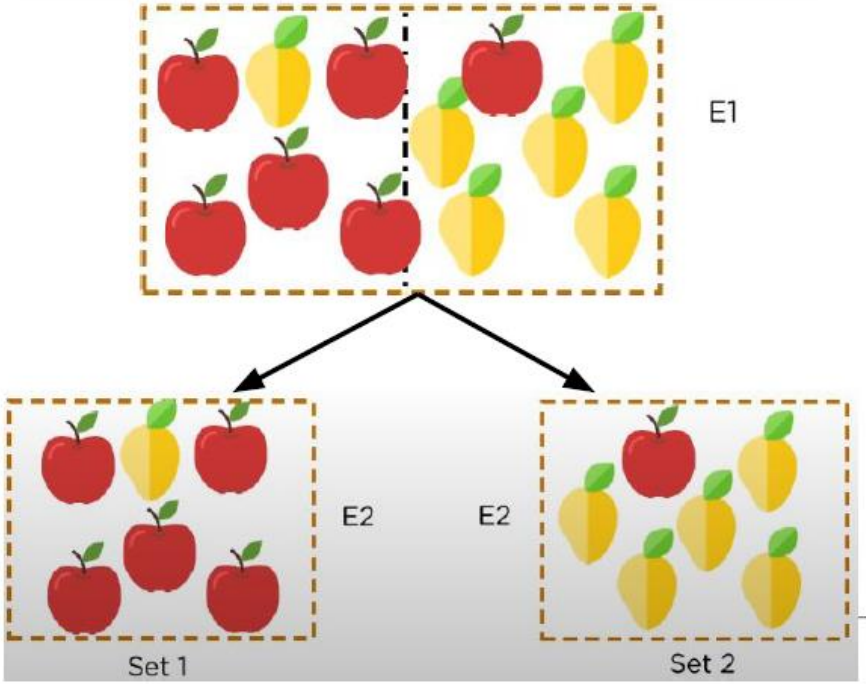
Quais atributos escolher?

Quais variáveis vão ser usadas em X ordem? E qual vai ser a raiz?

Precisamos da variável que vai *separar* a base de dados o máximo possível.



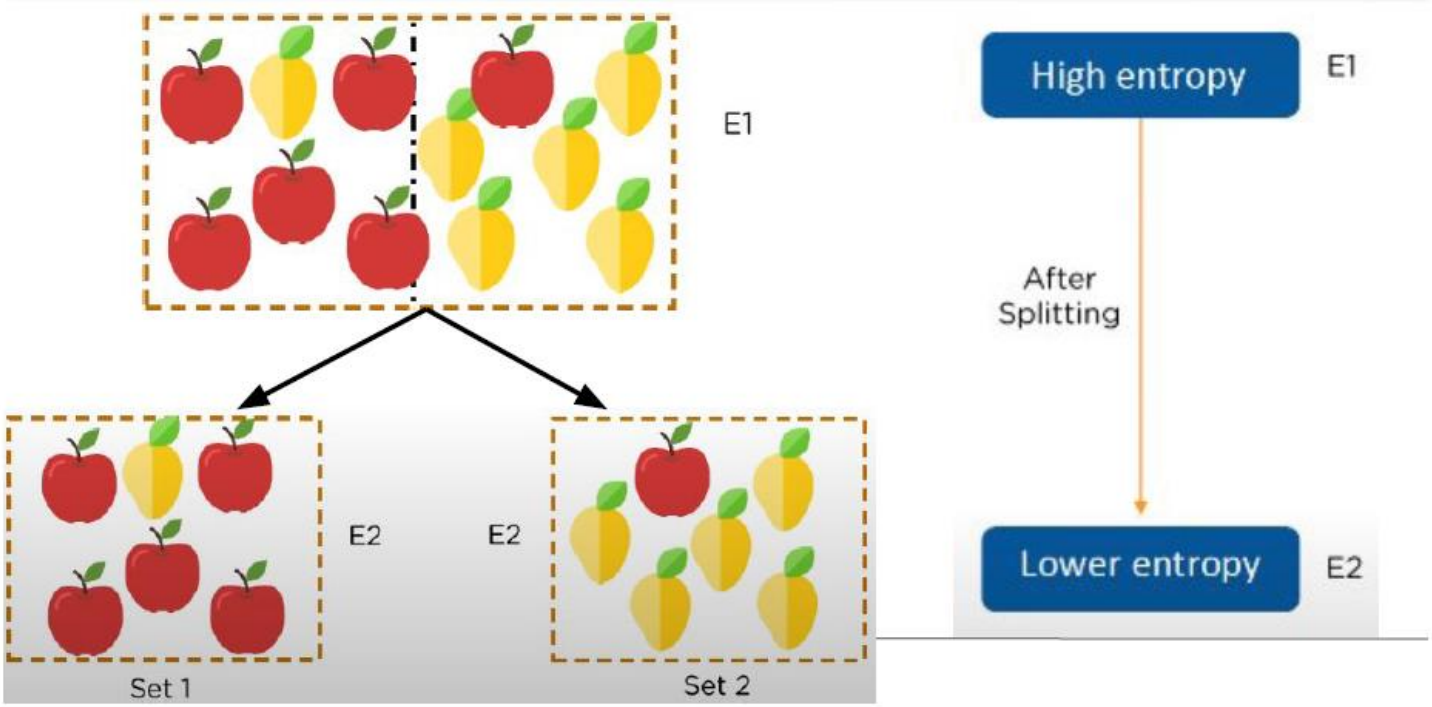
Quais atributos escolher ?



Entropia

Medição da aleatoriedade e impredictibilidade da base de dados

Quais atributos escolher?



Entropia

A *quantidade esperada de informação* quando observando uma

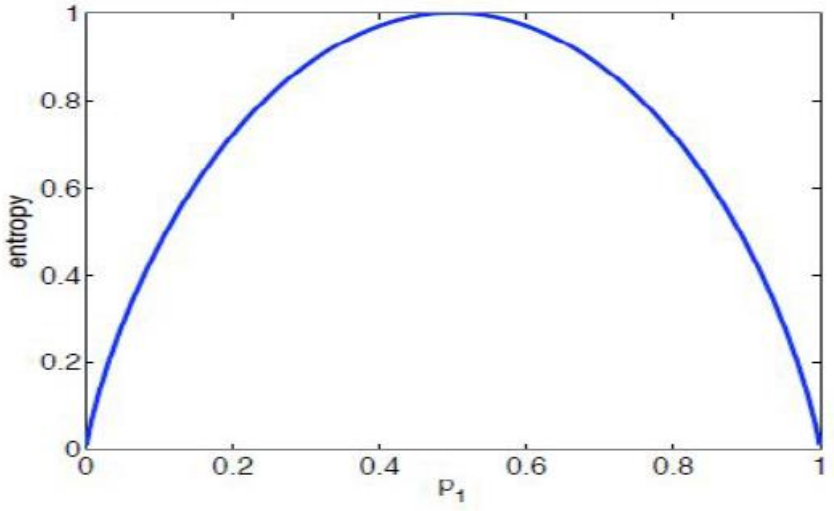
$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

Se X tem 8 possíveis resultados igualmente prováveis então: $H(X) == -\sum_i 1/8 \log_2 1/8 = 3$ bits

Entropia

Quanto menos uniforme e mais as probabilidades tendem a 0 ou 1, **menor** é a entropia

Distribuição:	-uniforme
Entropia:	-menor
Nó:	+puro



Ganho de Informação

Derivada da *teoria da informação* de Claude E. Shannon em 1932

"A Mathematical Theory of Communication"

Esse vai ser nosso critério de avaliação

Ganho da Informação

Ganho de informação - Gain(S,A):

- redução esperada da entropia (da incerteza) após a divisão

S = Base antes da divisão / A = Base após a divisão

Ganho(S,A) = Informação antes da divisão - após a divisão

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Ganho da Informação

Ganho de informação - Gain(S,A):

- redução esperada da entropia (da incerteza) após a divisão

S = Base antes da divisão / A = Base após a divisão

Ganho(S,A) = Informação antes da divisão - após a divisão

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Ganho da Informação

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

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Cálculo da entropia

Risco
Alto
Alto
Moderado
Alto
Baixo
Baixo
Alto
Moderado
Baixo
Baixo
Alto
Moderado
Baixo
Alto

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

Alto = 6/14

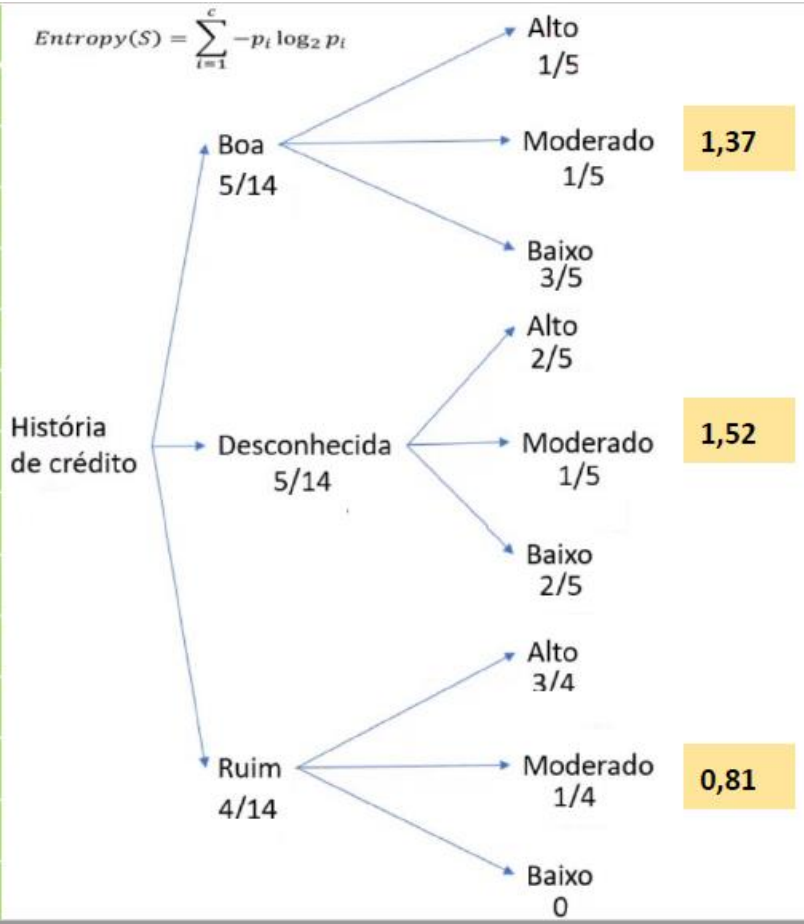
Moderado = 3/14

Baixo = 5/14

E(s) = -6/14 * log(6/14; 2) – 3/14 * log(3/14; 2) – 5/14 * log(5/14; 2) = **1,53**

Cálculo da entropia

História do crédito	Risco
Ruim	Alto
Desconhecida	Alto
Desconhecida	Moderado
Desconhecida	Alto
Desconhecida	Baixo
Desconhecida	Baixo
Ruim	Alto
Ruim	Moderado
Boa	Baixo
Boa	Baixo
Boa	Alto
Boa	Moderado
Boa	Baixo
Ruim	Alto



1. Calcular a Entropia da variável em cada combinação
2. Calcular o Ganho de Informação para essa variável

Ganho da Informação

História do crédito	Risco
Ruim	Alto
Desconhecida	Alto
Desconhecida	Moderado
Desconhecida	Alto
Desconhecida	Baixo
Desconhecida	Baixo
Ruim	Alto
Ruim	Moderado
Boa	Baixo
Boa	Baixo
Boa	Alto
Boa	Moderado
Boa	Baixo
Ruim	Alto

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Ganho(Hist.Crédito) = 1,53 - (5/14 * 1,37) - (5/14 * 1,52) - (4/14 * 0,81)
= 0,26

Quais atributos escolher

História de crédito: 0,26 Dívida: 0,06 Garantias: 0,20 **Renda: 0,66**

História do crédito	Dívida	Garantias	Renda anual	Risco
Ruim	Alta	Nenhuma	< 15.000	Alto
Desconhecida	Alta	Nenhuma	>= 15.000 a <= 35.000	Alto
Desconhecida	Baixa	Nenhuma	>= 15.000 a <= 35.000	Moderado
Desconhecida	Baixa	Nenhuma	> 35.000	Alto
Desconhecida	Baixa	Nenhuma	> 35.000	Baixo
Desconhecida	Baixa	Adequada	> 35.000	Baixo
Ruim	Baixa	Nenhuma	< 15.000	Alto
Ruim	Baixa	Adequada	> 35.000	Moderado
Boa	Baixa	Nenhuma	> 35.000	Baixo
Boa	Alta	Adequada	> 35.000	Baixo
Boa	Alta	Nenhuma	< 15.000	Alto
Boa	Alta	Nenhuma	>= 15.000 a <= 35.000	Moderado

Outro métodos

- Impureza de Gini

$$I_G(n) = 1 - \sum_{i=1}^{J} (p_i)^2$$

- Teste qui-quadrado

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Comandos no R

- Você vai precisar dos pacotes:
  - Rpart
  - Rpart.plot
  - Caret

Modelando no R

```
library(rpart)
library(rpart.plot)



mod1 <- rpart(formula, data = data)

rpart.plot(mod1)
```

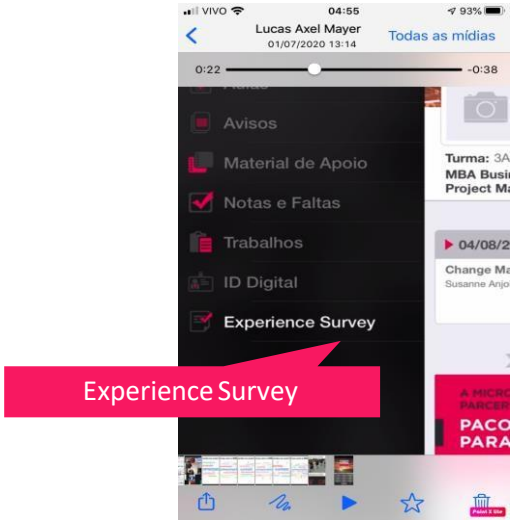
Avaliando o modelo no R

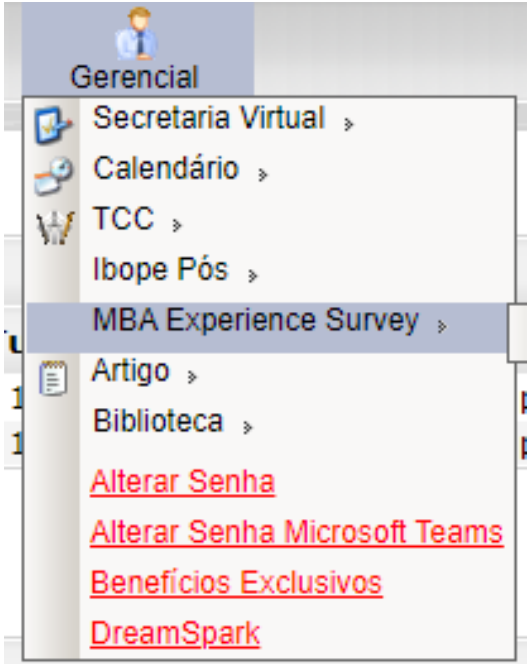
```
library(caret)

#probabilidades
predict(mod1, treino, type = "prob")

#classes
predict(mod1, treino, type = "class")

#medindo
treino$y <- predict(mod1, treino, type = "class")
teste$y <- predict(mod1, teste, type = "class")

confusionMatrix(treino$y, treino$target)
confusionMatrix(teste$y, teste$target)
```

O que você achou da aula de hoje?

Pelo aplicativo da FIAP

(Entrar no FIAPP, e no menu clicar em Experience Survey)





OBRIGADO

in/lafphd

FIAP MBA+

Copyright © 2019 | Professor (a) Nome do Professor
Todos os direitos reservados. Reprodução ou divulgação total ou parcial deste documento, é expressamente proibido sem consentimento formal, por escrito, do professor/autor.


FIAP