



Licap

Formação Cientista de Dados



PUC Minas

Formação do Cientista de Dados

Aplicações e Estratégias – Módulo Básico

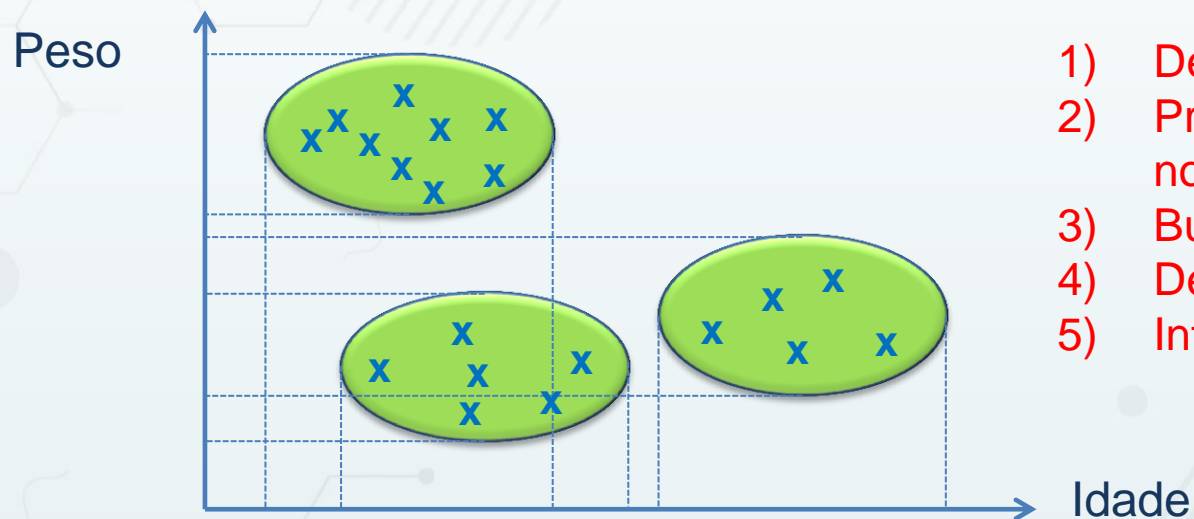
Luis Enrique Zárate

Aplicações e Estratégias

1. Principais técnicas de data mining (Agrupamento, Classificação e Associação, exemplos ilustrativos)
2. Marketing – CRM
3. Recursos Humanos
4. Setor Financeiro
5. Medicina e Saúde
6. Educação
7. Agro-negócio
8. Indústria 4.0

Principais técnicas para Data Mining

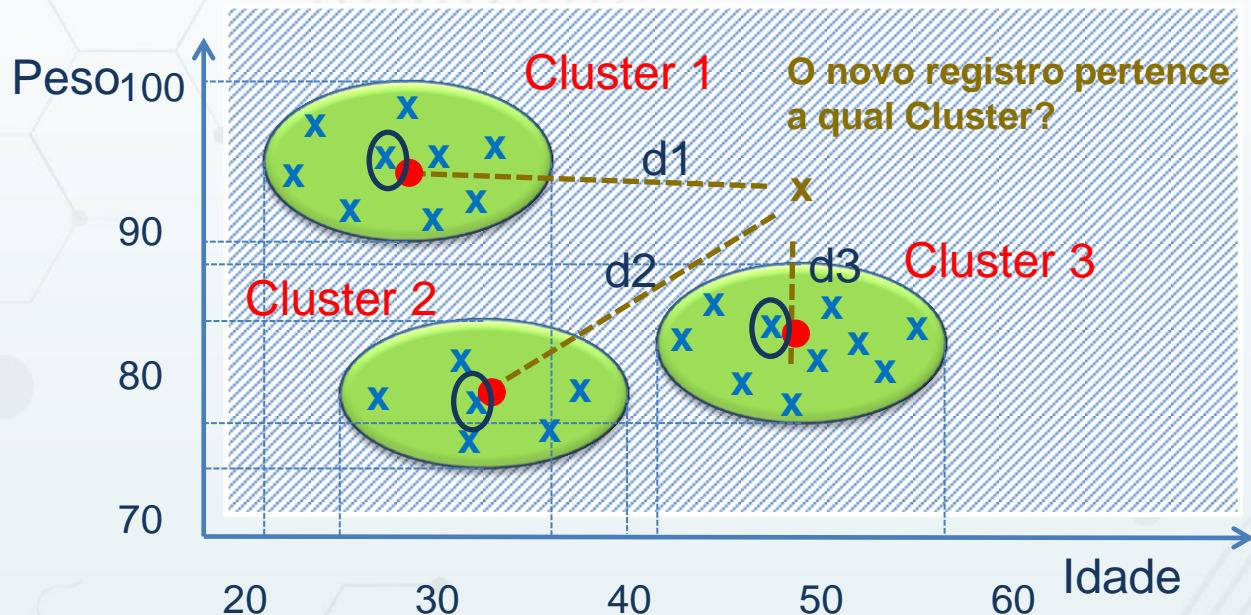
Técnicas de Agrupamento



Etapas da Aplicação:

- 1) Definição do conjunto de dados.
- 2) Pré-processamento e normalização dos dados.
- 3) Busca dos Agrupamentos;
- 4) Descrição dos Grupos.
- 5) Interpretação do modelo.

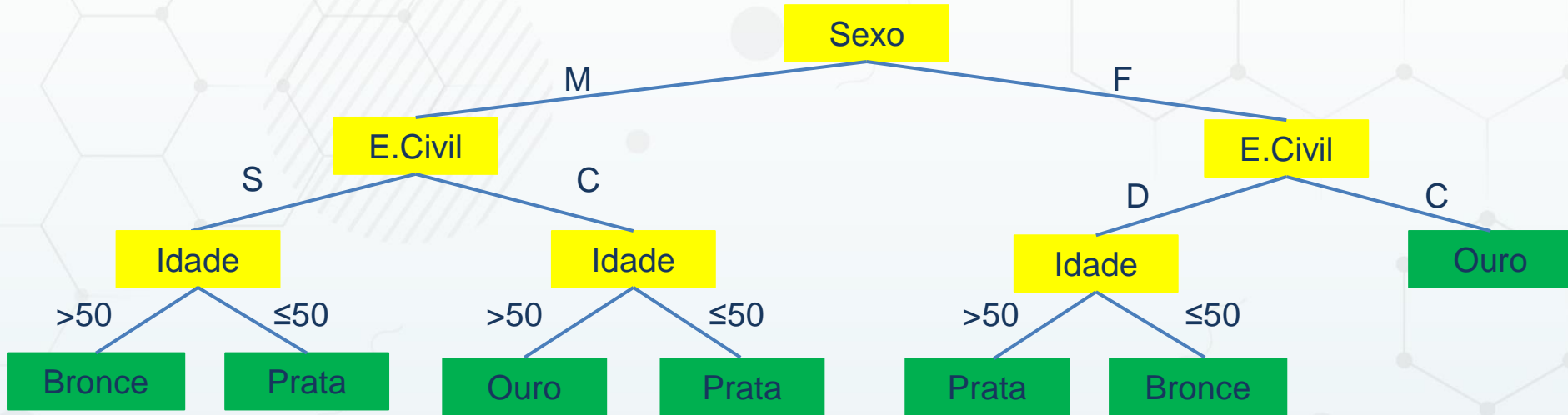
Técnicas de Agrupamento



O novo registro, de acordo ao nosso modelo, pertencerá ao cluster com:

- 1) Menor distância ao centro do mesmo (Centroide ●). No exemplo, ao cluster 3, (K-means)
- 2) Menor distância ao Medóide ○ (objeto mais próximo, em média, aos demais objetos do cluster) (k-medóides)
- 3) Mais próximo do objeto caracterizado pela Mediana. Ex. Para o atributo $a=1,2,3,4,5$ Mediana(a) = 3 (k-medianas)
- 4) Para dados categóricos k-modes.

Técnicas de Classificação – Árvore de decisão



Etapas da Aplicação:

- 1) Definição do conjunto de dados;
- 2) Pré-processamento e categorização;
- 3) Construção da árvore para o maior Ganho de Informação;
- 4) Identificação das regras de decisão;
- 5) Interpretação do modelo

Técnicas de Associação

Transc	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 100
#1	X			X	X		X	
#2	X		X		X	X	X	X
#3	X	X	X	X	X			X
...	X	X	X		X		X	
#n		X				X		X

Se compra ITEM 1 e ITEM 3 então compra ITEM 4, suporte = 80%, Confiança = 60%

Etapas da Associação para atividades de Data mining:

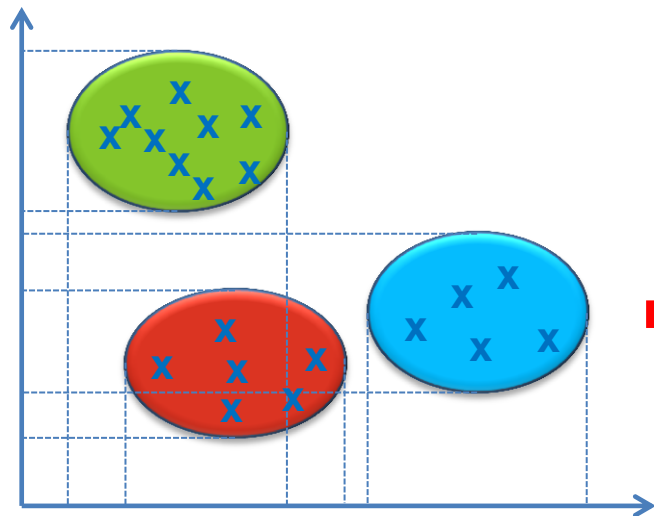
- 1) Definição do conjunto de dados;
- 2) Pré-processamento e binarização;
- 3) Identificação de itens frequentes;
- 4) Construção de regras frequentes;
- 5) Interpretação das regras de associação

Aplicações e Estratégias



Marketing-CRM:

Descobrir padrões de consumo para diferentes perfis desconhecidos de clientes.



Trans	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 100
#1	X			X			X	

#	Trans	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 100
#	#1	X			X	X		X	

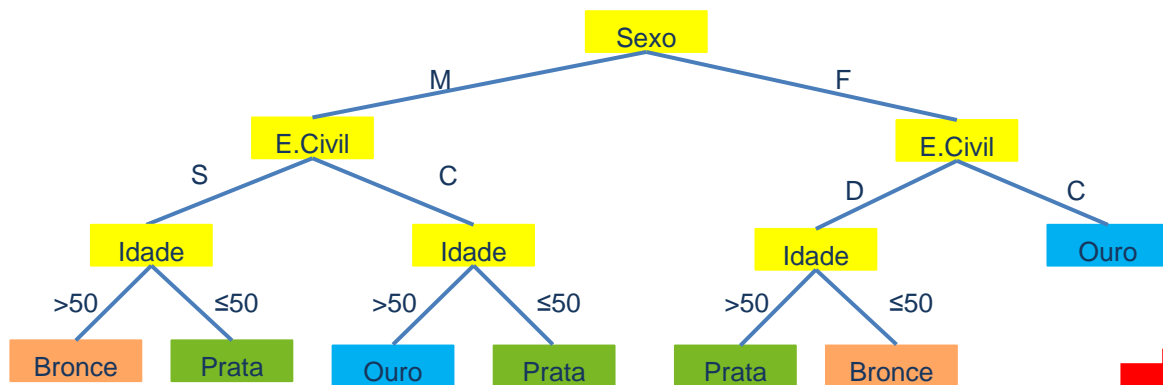
#	Trans	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 100
#	#1	X			X	X		X	
#	#2	X		X		X	X	X	X
#	#3	X	X	X	X	X			X
#	...	X	X	X		X		X	
#	#n		X				X		X

Aplicações e Estratégias



Marketing-CRM:

Descobrir padrões de consumo, de grupos de clientes com perfis conhecidos.



Trans	Ítem 1	Ítem 2	...	Ítem 100
#1	X		X	
#2		X		X

Trans	Ítem 1	Ítem 2	...	Ítem 100
#1	X		X	
#2	X			X

Trans	Ítem 1	Ítem 2	...	Ítem 100
#1	X		X	
#2	X		X	X
#3	X	X		X
...	X	X	X	
#n		X		X

Aplicações e Estratégias



Marketing-CRM:

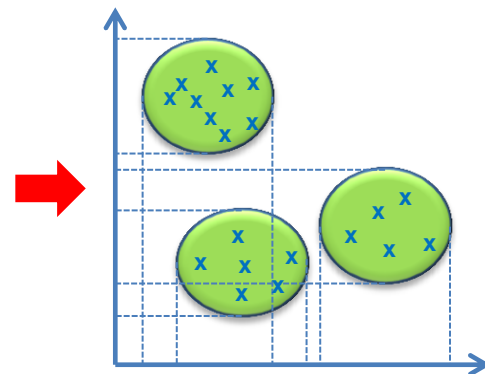
Identificar perfis de clientes a partir de dados pessoais e de padrões de consumo.



Registro	Idade	Sexo	EC	...	Ítem 1	Ítem 2	Ítem 3	Ítem 100
#1	37	M	S	...	X		X	
#2	45	F	C	...	X	X	X	X
...
#n	53	M	C	...		X		X

Categorização do Conjunto de Dados

Registro	Idade	Sexo	EC	...	Categoria ítems A	Categoria ítems B	...	Categoria ítems N
#1	37	M	S	...	X		...	
#2	45	F	C	...	X	X	...	X
...
#n	53	M	C	...		X	...	X

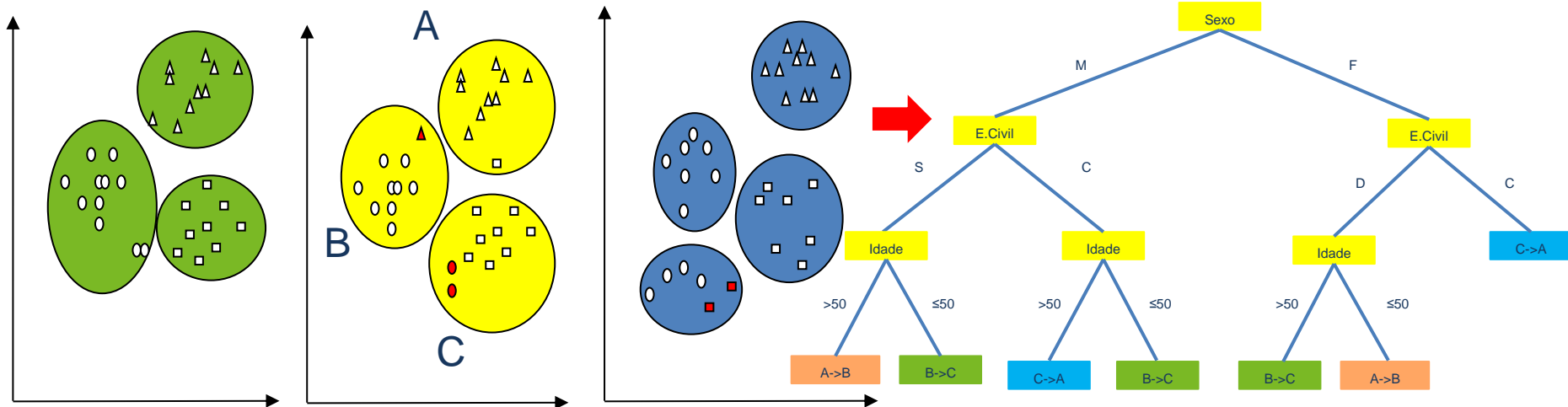


Aplicações e Estratégias



Marketing-CRM:

Identificar o perfil de clientes que se movimentam entre grupos, ou que tendem a abandonar a fidelidade. Identificar e acompanhar temporalmente grupos de clientes emergentes e em extinção.



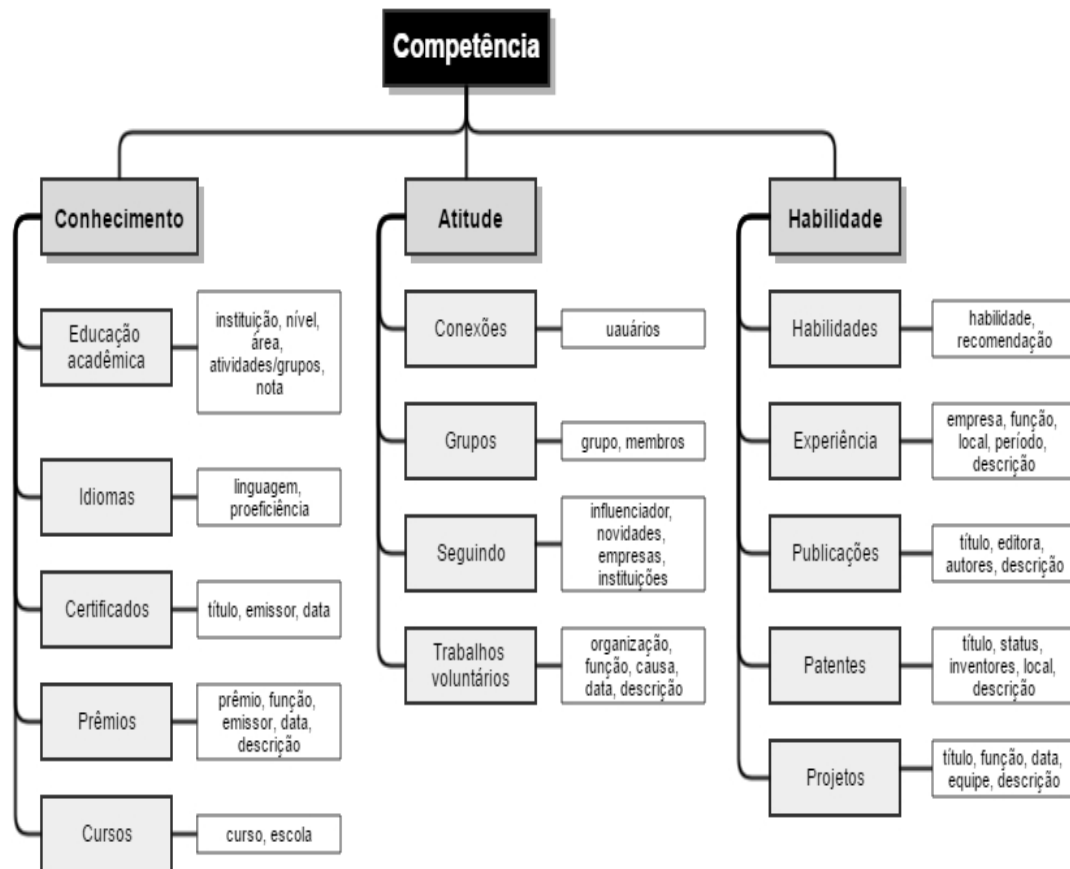
Aplicações e Estratégias

Recursos humanos:

Descoberta de padrões de habilidades e competências de perfis profissionais.

Profissão Analista Sênior	Eng. Soft	Data Analytic	...	BI
CV #1	X		X	
CV #2	X			X
...	X	X	X	
CV #n		X		X

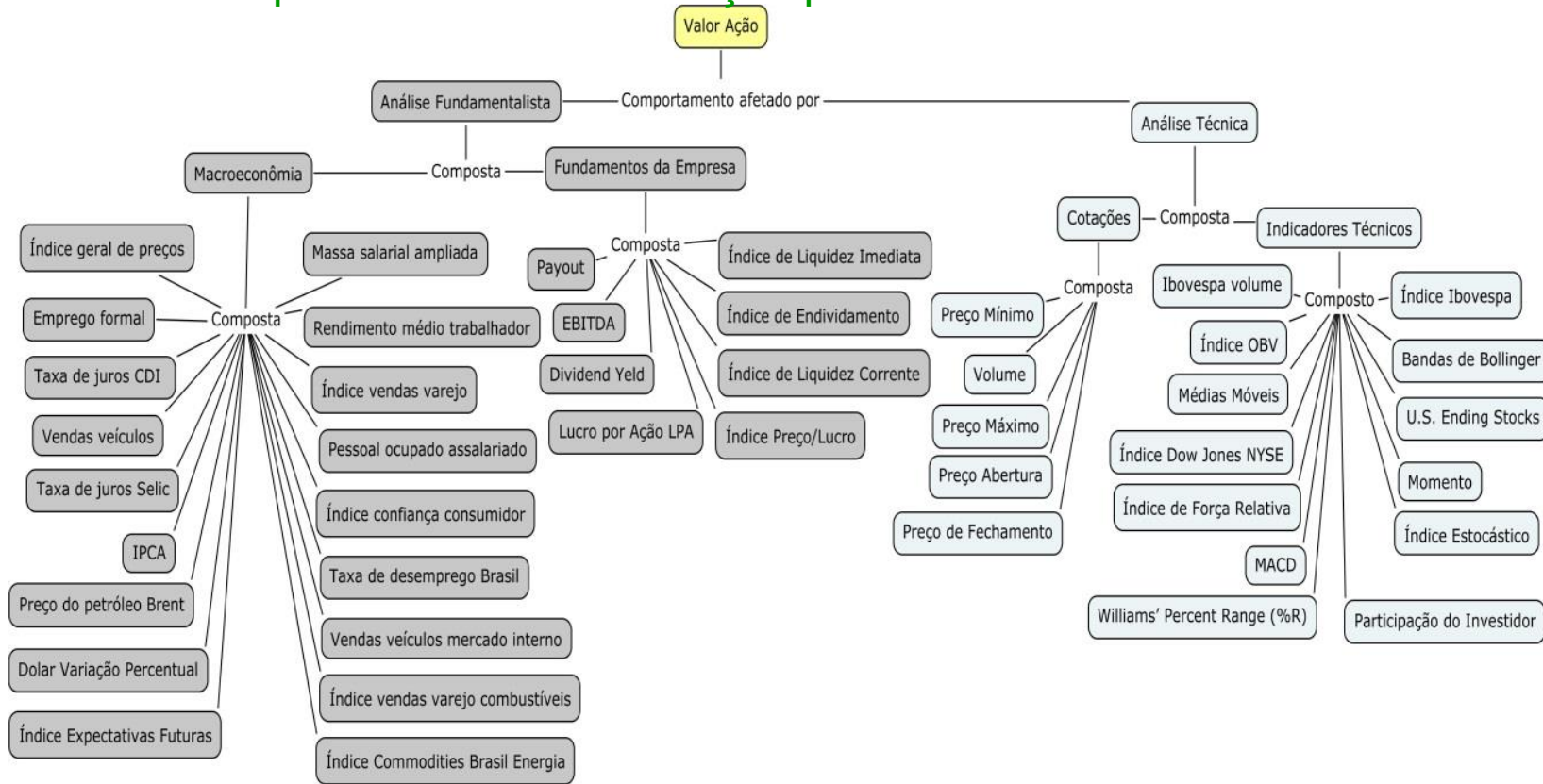
**Se Eng. Soft e BI
então Data Analytic,
suporte = 80%, Confiança = 60%**



Aplicações e Estratégias

Financeiro:

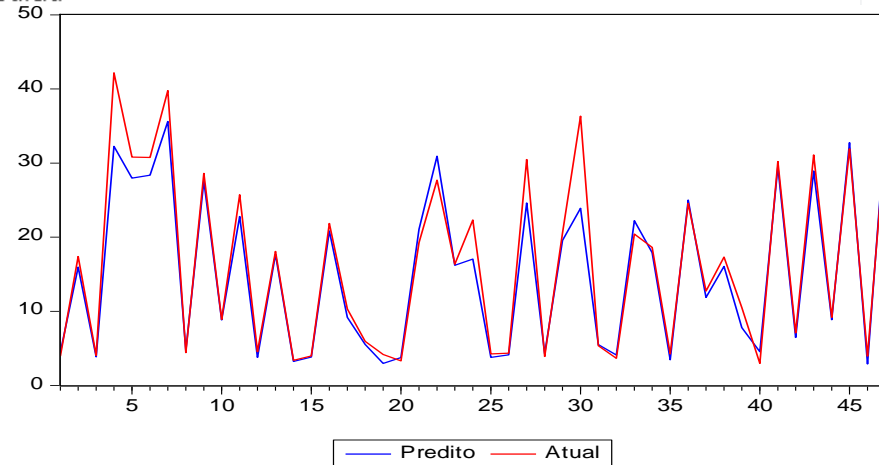
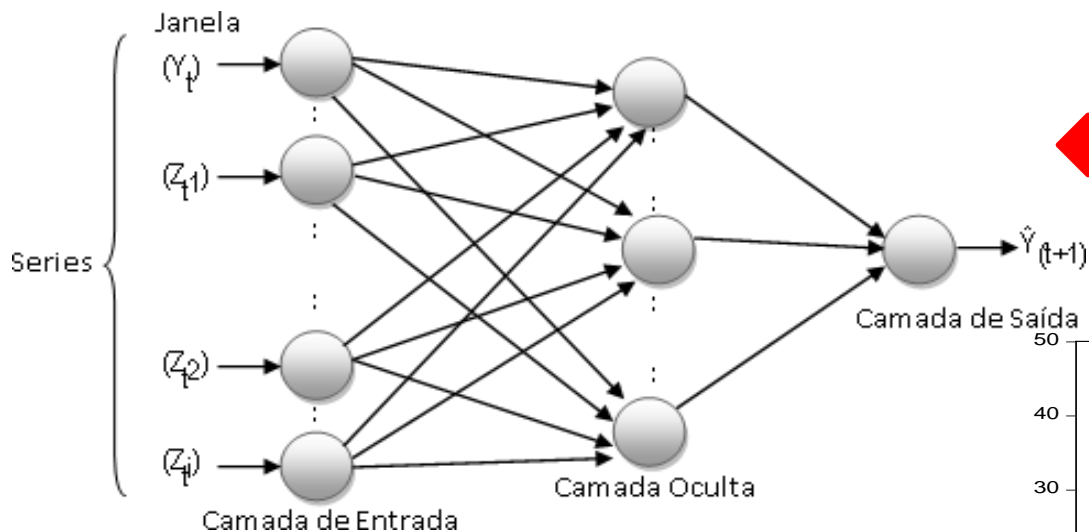
Modelos de previsão do valor da ação para o mercado financeiro.



Aplicações e Estratégias

Financeiro:

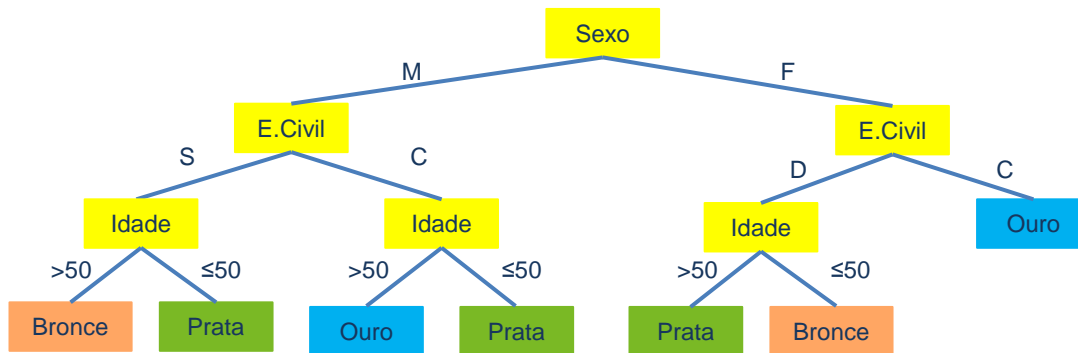
Modelos de previsão do valor da ação para o mercado financeiro.



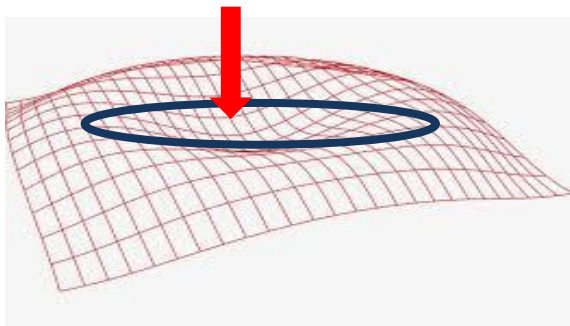
Aplicações e Estratégias

Financeiro:

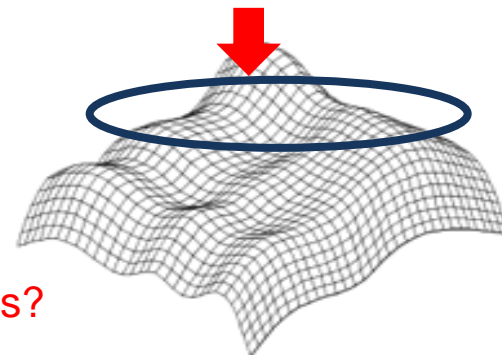
Identificar perfis de clientes para liberação de crédito.



O quanto nosso modelo não está polarizado?



O quanto nosso modelo gera respostas consistente?



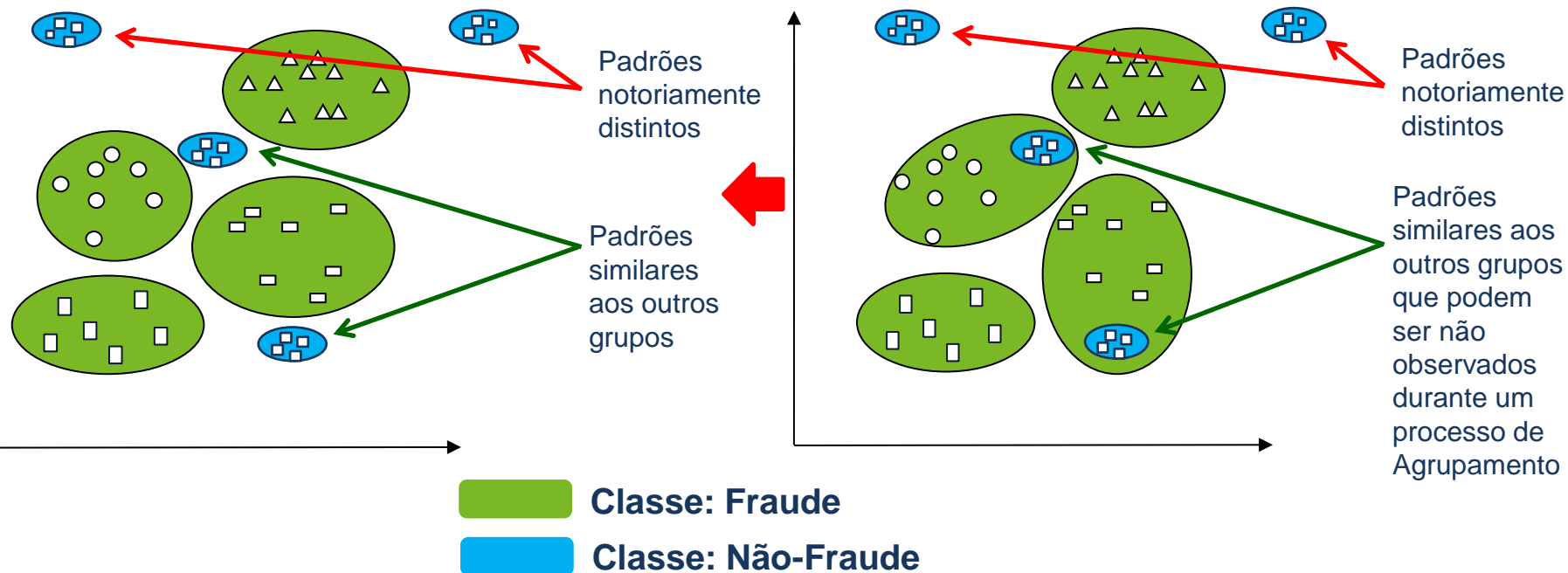
- Modelos Representativos?
- Modelos Consistentes?
- Modelos Parcialmente Consistentes?

Aplicações e Estratégias



Financeiro:

Detectar fraudes no uso de cartões de crédito e empréstimos bancários.



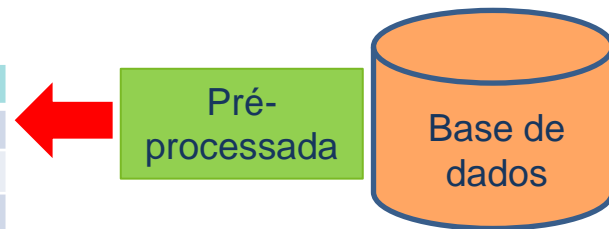
Aplicações e Estratégias



Medicina e saúde:

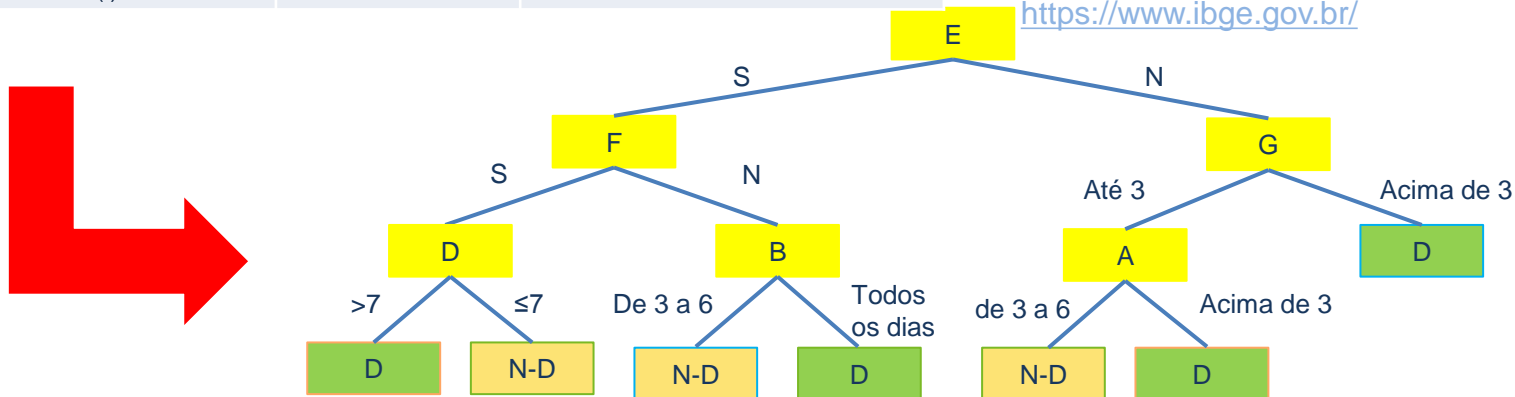
Descoberta de padrões de doenças - Depressão.

Atributos ou Variáveis	Tipo de variável	Valores
Duas Semanas com problemas de Sono (A)	Categórica	0/até 3/de 3 a 6/todos os dias
Duas semanas com cansaço (B)	Categórica	0/até 3/de 3 a 6/todos os dias
Lentidão ou agitação nas últimas semanas (C)	Categórica	0/até 3/de 3 a 6/todos os dias
Frequência do uso de remédios para dormir (D)	Númérico	0...14 dias
Sofreu violência de pessoa conhecida (E)	Dicotômica	S/N
Sofreu violência de pessoa desconhecida (F)	Dicotômica	S/N
Falta de apetite (G)	Categórica	0/até 3/de 3 a 6/todos os dias
Atividade física semanal (I)	Dicotômica	S/N



A base de dados possui **205546** instâncias com **942** variáveis que caracterizam aspectos sobre a saúde e estilo de vida de cidadãos no Brasil. Fonte IBGE.

<https://www.ibge.gov.br/>

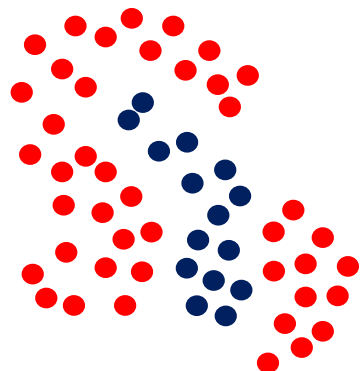


Aplicações e Estratégias

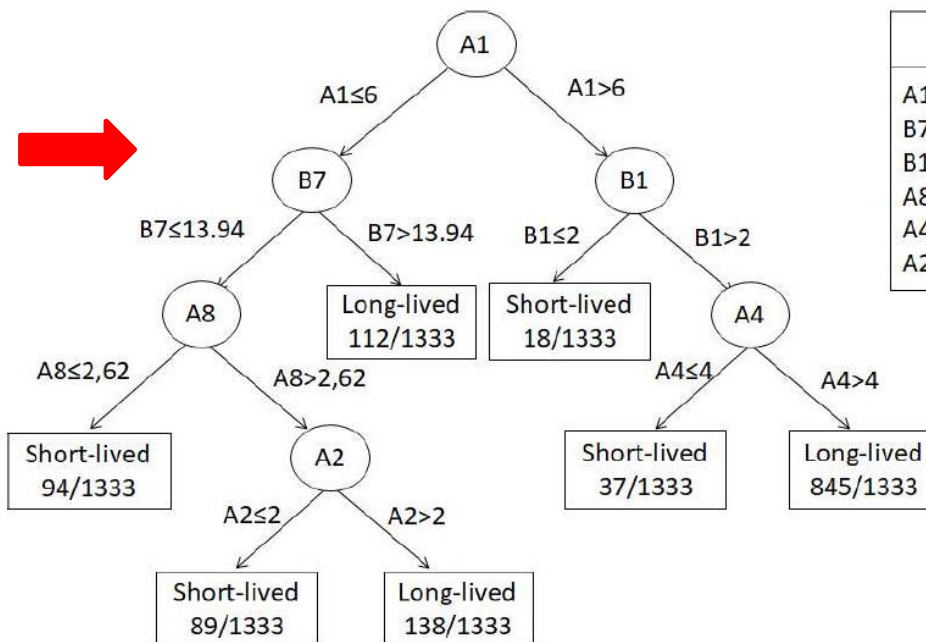


Medicina e saúde:

Identificar padrões de longevidade em populações.



Agrupamentos com
formatos arbitrários



Legend

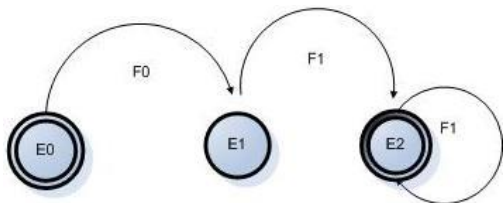
A1: Frequency of physical activity
B7: Negative feelings questionnaire
B1: Depression symptoms questionnaire
A8: Alcohol and Tobacco consumption
A4: Medical history questionnaire
A2: Serious health issues



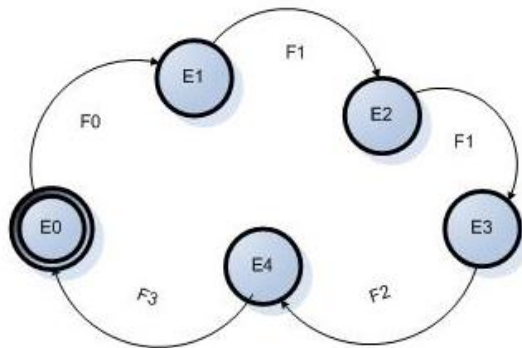
Aplicações e Estratégias

Medicina e saúde:

Doenças Epidemiológicas ou Crônicas



Contaminado mas não doente



Doente em tratamento

Cenários a serem analisados:

Esses estados representam a situação em que um indivíduo se encontra antes, durante e depois da aquisição do bacilo.

E0 – indivíduo saudável

E1 – indivíduo contaminado

E2 – indivíduo contaminado mas não doente

E3 – indivíduo doente

E4 – doente em tratamento

E5 – Falecimento do paciente

E6 – indivíduo doente com rejeição do tratamento

Transições dos cenários apresentados

Essas transições representam as ações que levaram às mudanças de estados

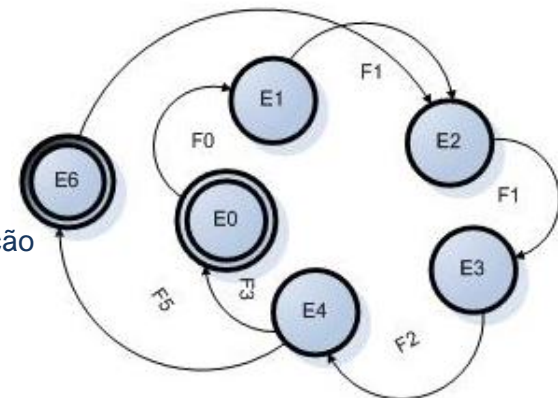
F0 – exposição do bacilo pelo ar

F1 – fatores endógenos e/ou exógenos

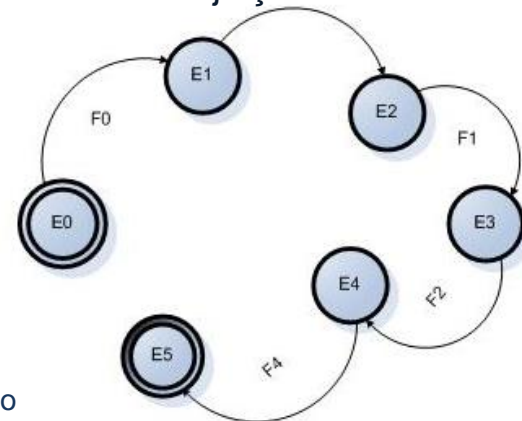
F2 – detecção da doença

F3 – conclusão do tratamento

F4 – negligência do tratamento relativa aos sintomas ou ao tratamento



Doente com rejeição ao tratamento



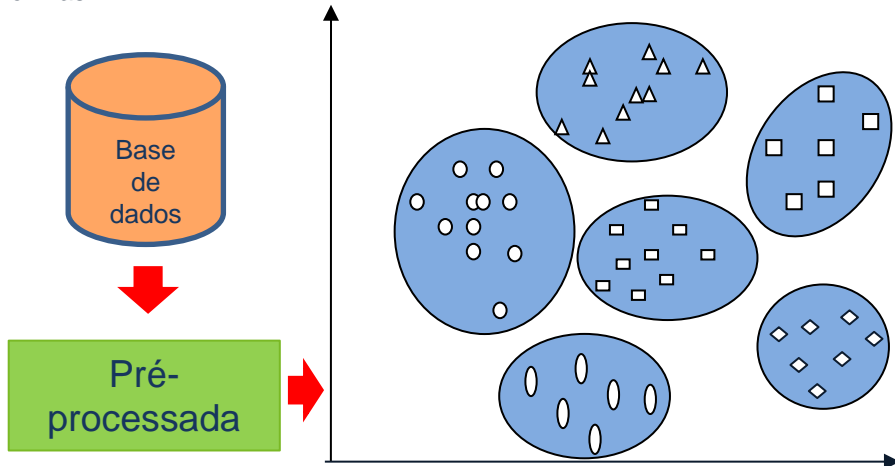
Doente com óbito

Aplicações e Estratégias

Medicina e saúde:

Identificar grupos para estudos de vulnerabilidade social.

A base de dados possui **195.582** instâncias com **177** variáveis que caracterizam aspectos sociais de indivíduos no Brasil.

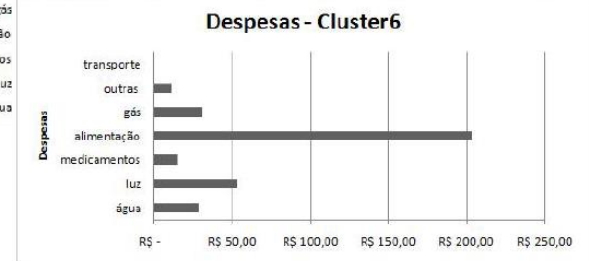


CATEGORIA	TIPO	%
Localidade	Urbana	92,7
Programa de assistência	PSF	94,2

CATEGORIA	TIPO	%
Localidade	Urbana	98,9
Programa de assistência	Outros Programas	88,8

CATEGORIA	TIPO	%
Localidade	Urbana	96,8
Programa de assistência	PSF	84,8

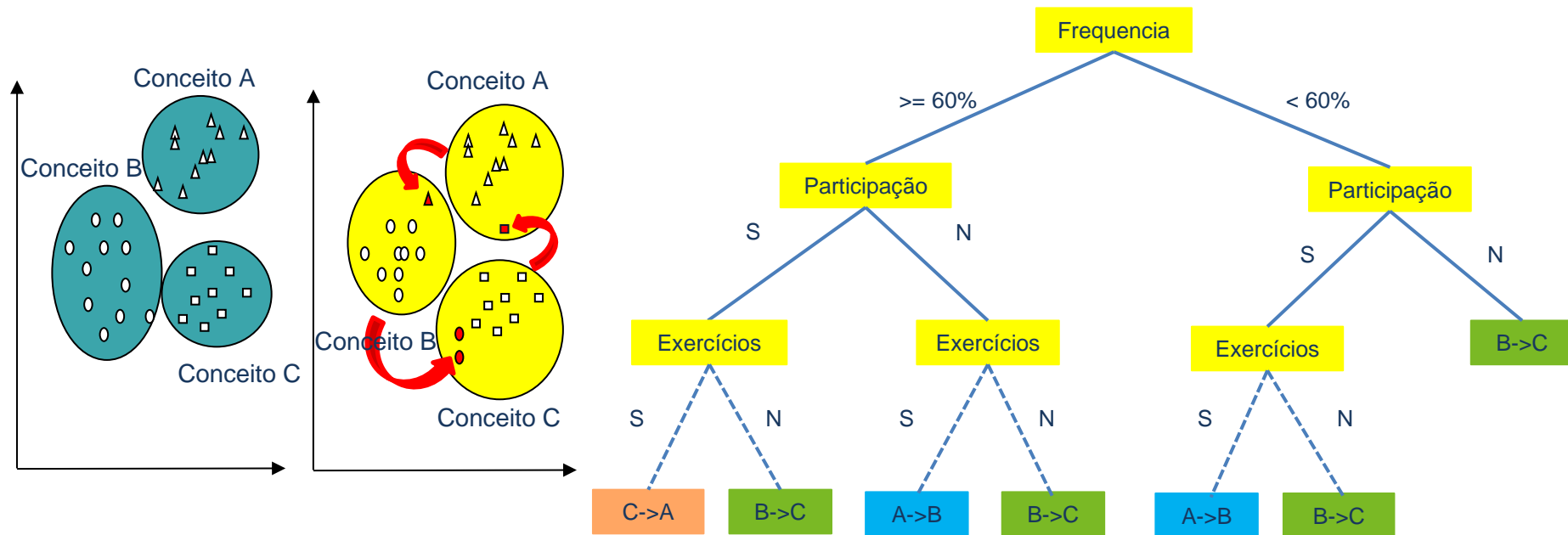
CATEGORIA	TIPO	%
Localidade	Urbana	93,8
Programa de assistência	PSF	93,8
Tipo do domicílio	Casa	87,2
Situação do domicílio	Próprio	93,8
Material de construção	Alvenaria	96,4
Abastecimento de água	Público	99,4
Tratamento de água	Filtração	70,4
Tratamento de água	Cloração	28,1
Iluminação	Relógio Próprio	98,4
Escoamento	Público	94,8
Destino do Lixo	Coletado	92,3



Aplicações e Estratégias

Educação:

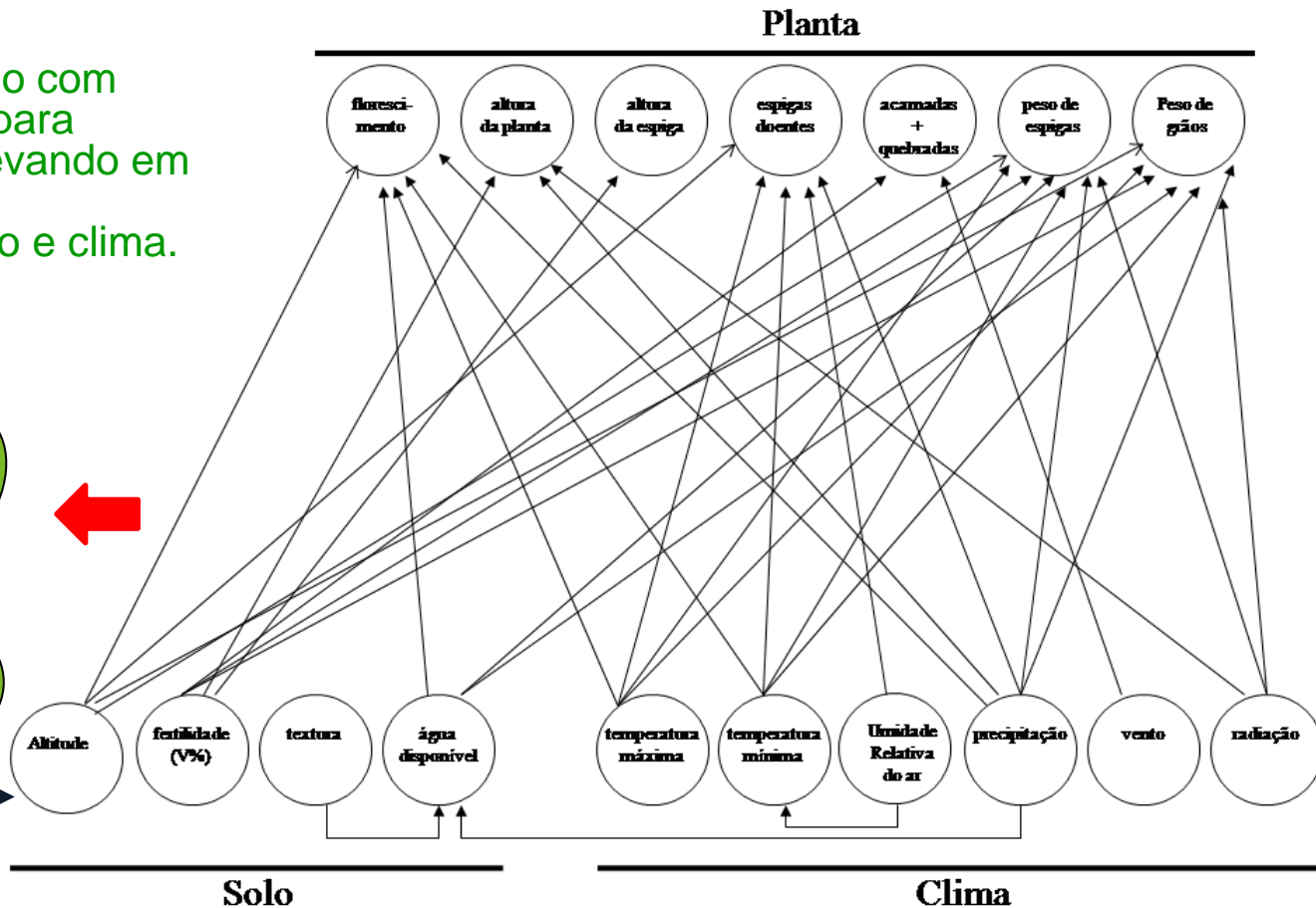
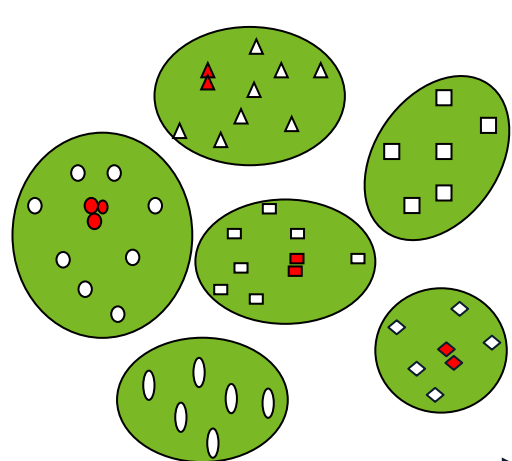
Identificar ações que torna-se efetivas temporalmente no processo de ensino-aprendizado.



Aplicações e Estratégias

Agro-negócio:

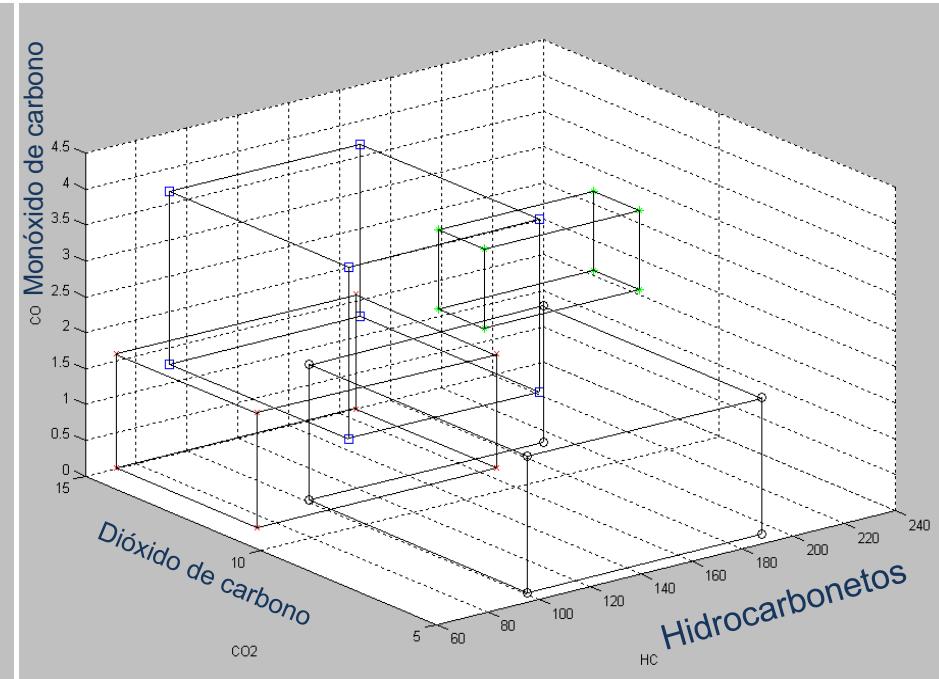
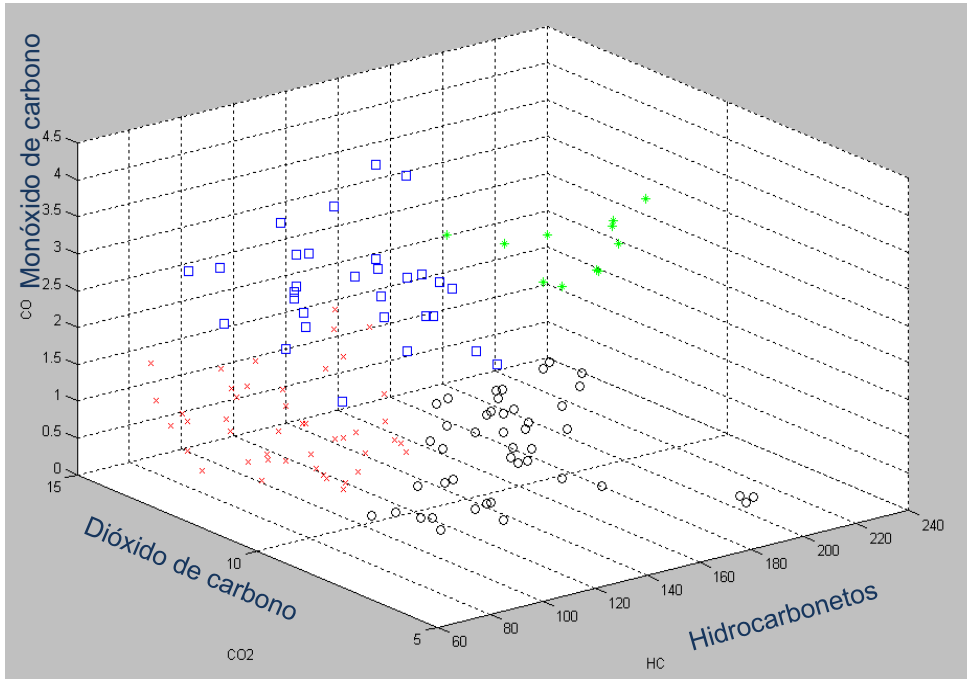
Identificar locais de plantio com características similares para expansão de cultivares levando em consideração dados da produtividade, planta, solo e clima.



Aplicações e Estratégias

Industria 4.0:

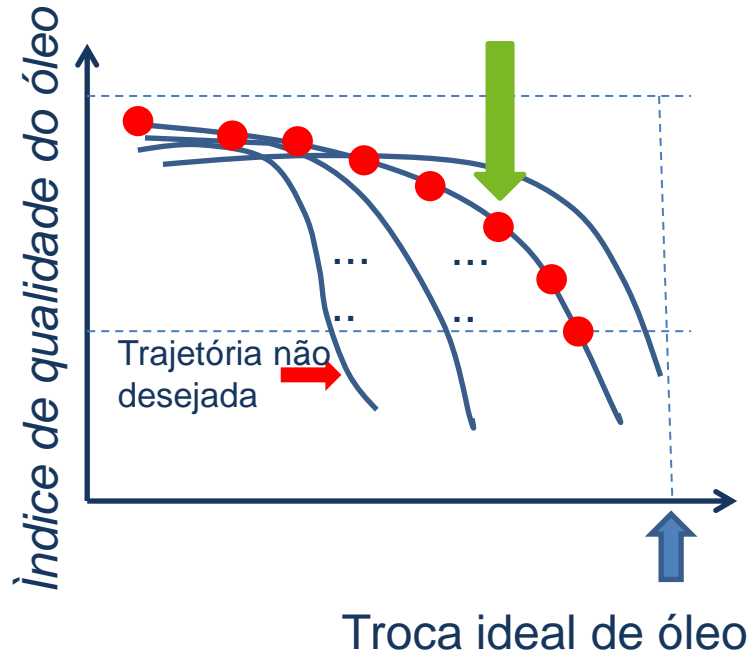
Descoberta de padrões de falhas em motores para industria automovilística.



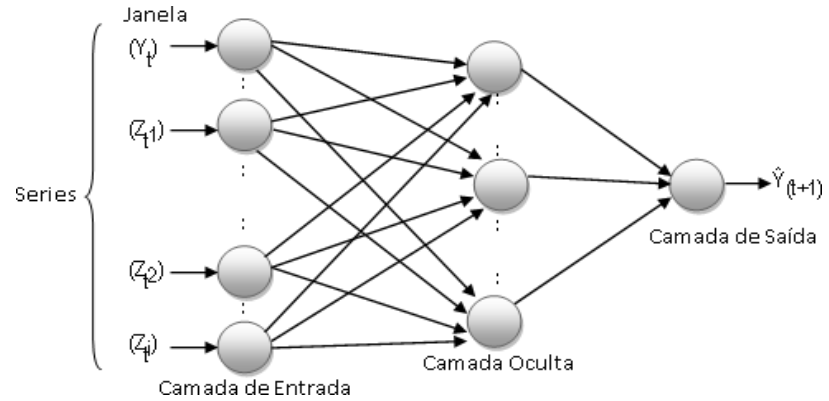
Aplicações e Estratégias

Industria 4.0:

- Previsão do consumo de óleo em maquinárias industriais.



Como evoluíram as condições mecânicas e de operação antes de alcançar uma determinada condição de operação que levou a troca prematura de óleo?



Aplicações e Estratégias

Indústria 4.0:

Modelagem não convencional de processos siderúrgicos.

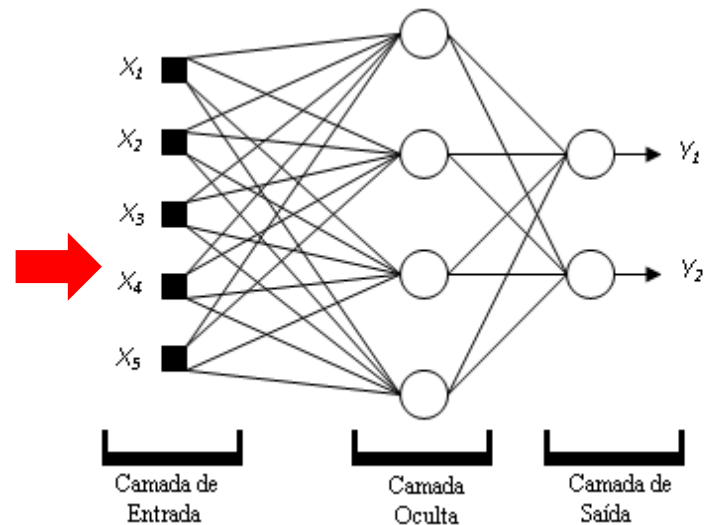
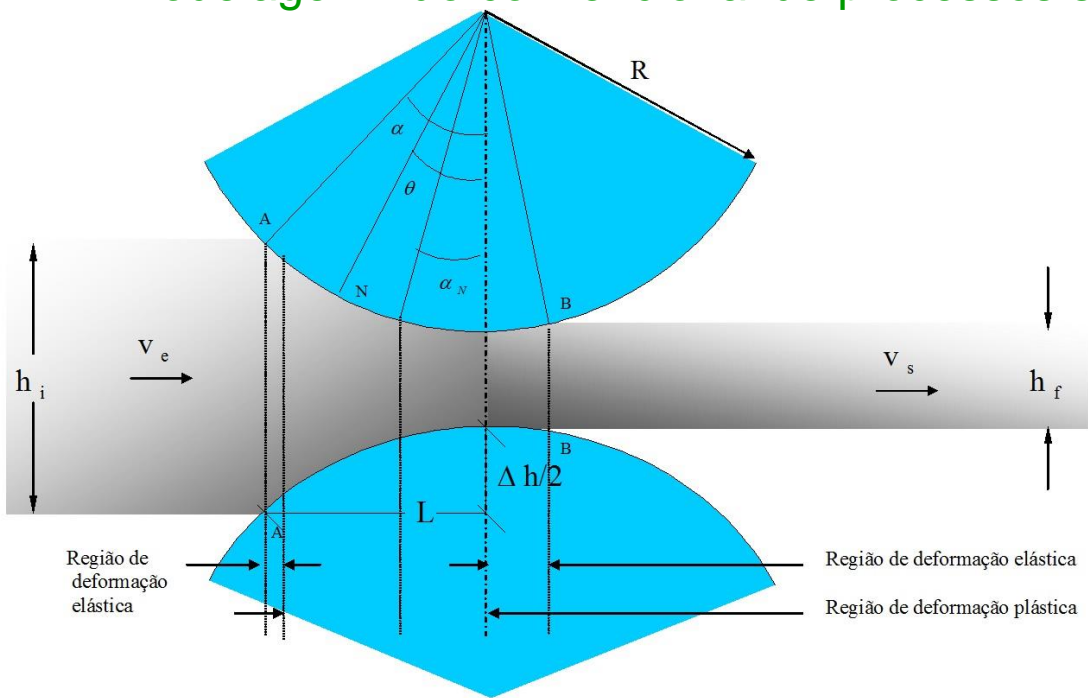


Figura I.1 Variáveis Geométricas do Processo de Laminação



Licap

Formação Cientista de Dados

Obrigado!



PUC Minas