Aprendizagem e Extração do Conhecimento

PL06 · CRISP – DM | Construção de Modelos

Mestrado em Engenharia Biomédica

Hugo Peixoto – hpeixoto@di.uminho.pt

Conteúdo autoral; design e estruturação suportados por IA.

O caminho de hoje

PARTE 1 – CONSTRUÇÃO DE MODELOS

- ✓ Correlações
- ✓ Regras de Associação
- ✓ Regressão Linear
- ✓ Classificação

PARTE I · CONSTRUÇÃO DE MODELOS

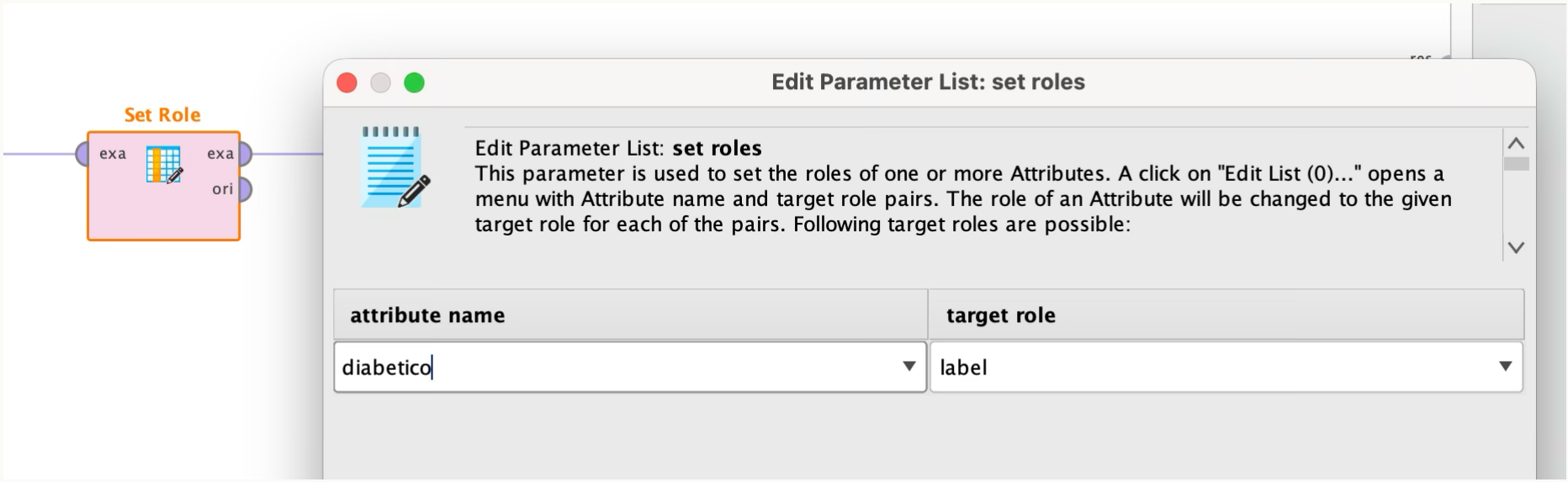
CRISP-DM


01
Compreensão do Negócio
✓ concluído
02
Compreensão dos Dados
✓ concluído
03
Preparação dos Dados
✓ concluído
04
Construção dos Modelos
← estamos aqui
05
Avaliação de Performance

CLASSIFICAÇÃO

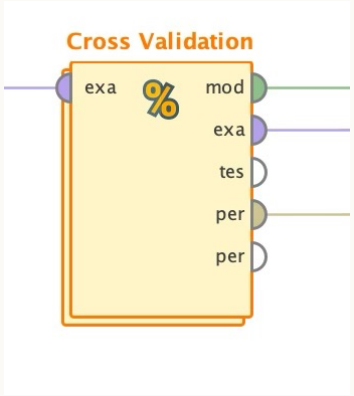
Classificação

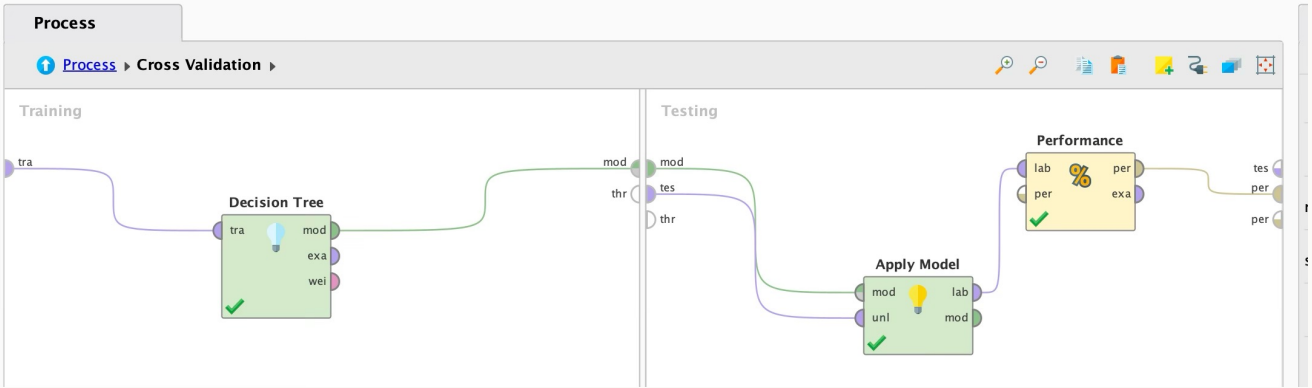
- Vamos usar o pre-processamento anterior!
- Garantir que o nosso label é o diabético: o que queremos definir!



Classificação

- Vamos usar o Modelo de Árvores de Decisão!
- Utilizar cross-validation!





Classificação

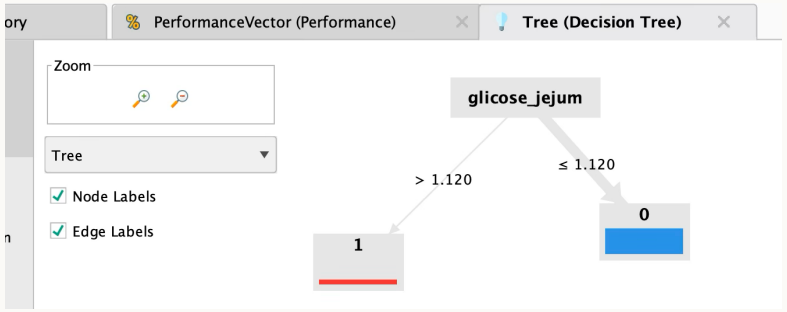
Resultados – com normalização:

accuracy: 99.67% +/- 1.05% (micro average: 99.66%)

	true 0	true 1	class precision
pred. 0	257	1	99.61%
pred. 1	0	38	100.00%
class recall	100.00%	97.44%	



Alguma coisa errada?



Classificação

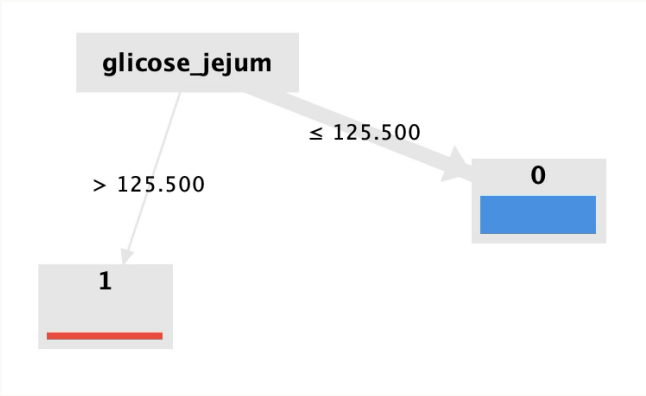
Resultados – sem normalização:

accuracy: 99.67% +/- 1.05% (micro average: 99.66%)

	true 0	true 1	class precision
pred. 0	257	1	99.61%
pred. 1	0	38	100.00%
class recall	100.00%	97.44%	



Alguma coisa errada?

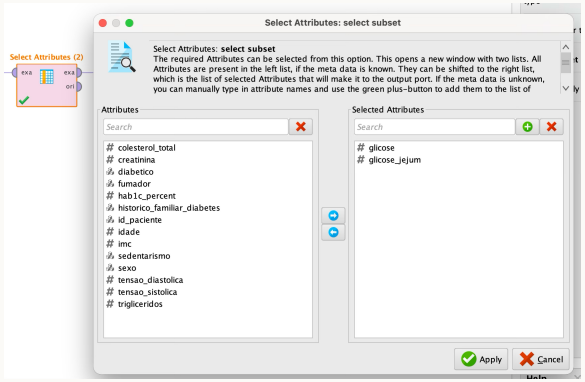


Classificação - Quizz

Observar o fluxo todo e descobrir um possível data leakage.

- Como corrigir?

Classificação



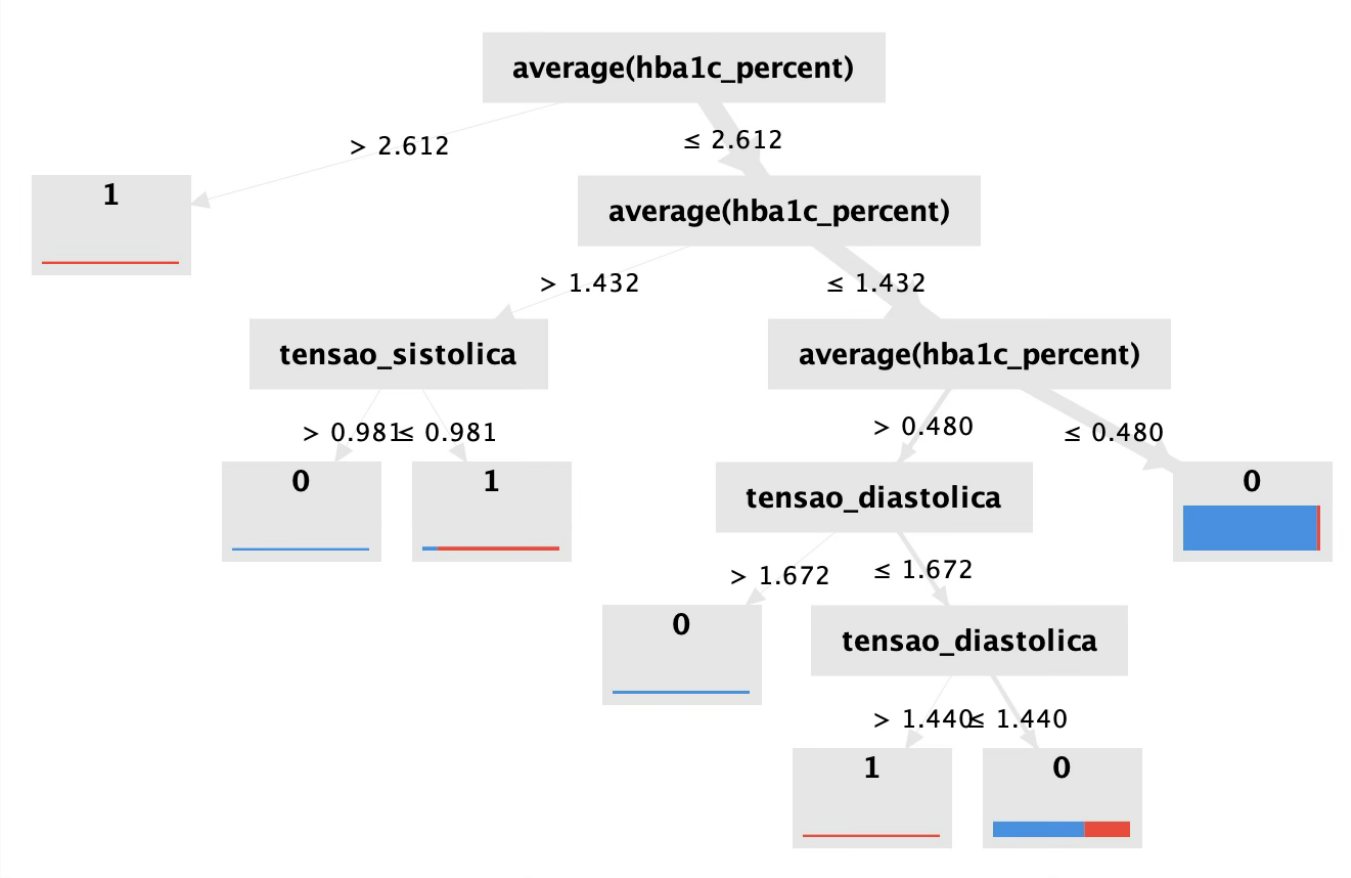
accuracy: 81.10% +/- 8.17% (micro average: 81.08%)

	true 0	true 1	class precision
pred. 0	232	31	88.21%
pred. 1	25	8	24.24%
class recall	90.27%	20.51%	

Accuracy: 81,10% ± 8,17%

A percentagem de pacientes classificados corretamente (diabéticos e não diabéticos juntos), em média nos folds do Cross Validation. O ±8,17% diz que há alguma variação entre folds, mas não é alarmante.

Classificação



Classificação

accuracy: 81.10% +/- 8.17% (micro average: 81.08%)

	true 0	true 1	class precision
pred. 0	232	31	88.21%
pred. 1	25	8	24.24%
class recall	90.27%	20.51%	

Comparação com a baseline:

Dos 296 pacientes nesta matriz, 257 são true 0 (não diabéticos). Um modelo "ingénuo" que prevê sempre 0 acertaria 259/298 = **86,9%**. O modelo deu 81,10%

Pior do que não fazer nada!!

Classificação

accuracy: 81.10% +/- 8.17% (micro average: 81.08%)

	true 0	true 1	class precision
pred. 0	232	31	88.21%
pred. 1	25	8	24.24%
class recall	90.27%	20.51%	

Matriz de confusão:

- **232 verdadeiros negativos** — não diabéticos corretamente identificados.
- **8 verdadeiros positivos** — diabéticos corretamente identificados.
- **25 falsos positivos** — não diabéticos marcados como diabéticos.
- **31 falsos negativos** — diabéticos que o modelo não detetou.

Classificação

accuracy: 81.10% +/- 8.17% (micro average: 81.08%)

	true 0	true 1	class precision
pred. 0	232	31	88.21%
pred. 1	25	8	24.24%
class recall	90.27%	20.51%	

Class recall:

- **Classe 0 (não diabéticos): 90,27%**
  - O modelo é muito bom a identificar quem não tem diabetes.
- **Classe 1 (diabéticos): 20,51%**
  - O modelo só deteta **1 em cada 5** diabéticos reais.
  - Os outros 4 em 5 passam-lhe ao lado.

Classificação

accuracy: 81.10% +/- 8.17% (micro average: 81.08%)

	true 0	true 1	class precision
pred. 0	232	31	88.21%
pred. 1	25	8	24.24%
class recall	90.27%	20.51%	

Class precision:

Classe 0: 88,21%

- Quando o modelo diz "não diabético", está certo 9 em cada 10 vezes.

Classe 1: 24,24%

- Quando o modelo diz "diabético", está errado **3 em cada 4 vezes**.

Classificação

accuracy: 81.10% +/- 8.17% (micro average: 81.08%)

	true 0	true 1	class precision
pred. 0	232	31	88.21%
pred. 1	25	8	24.24%
class recall	90.27%	20.51%	

O desvio (±8,17%) também merece nota: é grande, sugerindo que o desempenho varia bastante entre os folds do Cross Validation: possivelmente porque a classe minoritária (39 diabéticos em 296) fica muito pouco representada em cada fold individual, tornando a estimativa instável.

Classificação - Avaliação

- Accuracy de 81,1%, mas **inferior à baseline ingénua** (86,8% só prevendo "não diabético")
- Recall da classe diabética: apenas 20,5% — o modelo deteta só 1 em cada 5 diabéticos reais
- **31 falsos negativos:** pacientes diabéticos classificados como saudáveis — o erro mais grave clinicamente
- **Causa provável:** desequilíbrio de classes (13% diabéticos), já identificado na compreensão dos dados e não compensado no modelo

- **Conclusão:** este modelo, tal como está, não cumpre o objetivo de negócio de triagem precoce — necessário testar balanceamento de classes (ex.: SMOTE, undersampling, ou ajustar pesos de classe) antes de avançar para deployment

Classificação – Quizz!

Porque é que um modelo pode ter accuracy razoável e ainda assim ser clinicamente inútil?

Aprendizagem e Extração do Conhecimento

PL06 · CRISP – DM | Construção de Modelos

Mestrado em Engenharia Biomédica

Hugo Peixoto – hpeixoto@di.uminho.pt

Conteúdo autoral; design e estruturação suportados por IA.