

Enhancing the Interpretability of Cardiovascular Disease Classifiers using Born-Again Tree

Luís Guilherme S. N. A. Magalhães

Eduardo Corrêa Gonçalves

Escola Nacional de Ciências Estatísticas (ENCE/IBGE)

KDMile 2024

Outline

Introdução

Trabalhos Relacionados

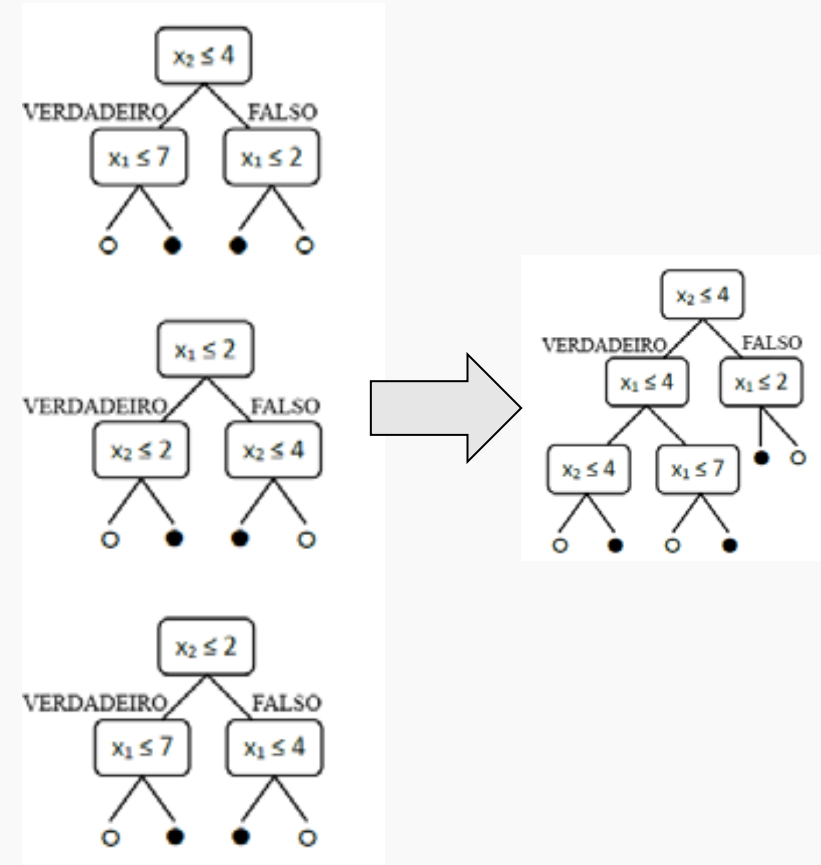
Born-Again Tree Ensembles

Experimentos

Conclusões e Trabalhos Futuros

Introdução (1/2)

- **Objetivo do trabalho:**
 - Avaliar o algoritmo **Born-Again Tree Ensembles (BA)** na classificação do risco de **doenças cardiovasculares**.
- **O que é BA ?** (Vidal e Schiffer, 2020)
 - Algoritmo que **transforma** ...
 - ... uma **Random Forest**
 - ... em uma **única Árvore de Decisão** com o mesmo poder preditivo.



Introdução (2/2)

- O que são doenças cardiovasculares (DCV) ?
 - Termo geral para **condições que afetam o coração ou vasos sanguíneos**.
 - **Principal causa de morte** no mundo, de acordo com a Organização Pan-Americana de Saúde (OPAS).
 - **Taxa de Prevalência** de DCV no **Brasil em 2021** (Oliveira et al., 2024) :
 - 7,6% para homens
 - 6,3% para mulheres
 - Taxa de mortalidade por 100 mil habitantes = 348,5.


Trabalhos Relacionados (1/2)

- Previsão de DCV - Abordagem 1: Métodos Baseados em Escores e Equações
 - Fáceis de serem usados e interpretados.
 - Muito difundidos na área médica.
 - **Ex.:** Calculadora de Risco Cardiovascular da OPAS.

Estimate cardiovascular risk

[Go to the Clinical Pathway](#)

10-year CV risk: 3%



Input data

Country	Brasil
Gender	Male
Age	51
Smoker	No
Systolic pressure	110 mmHg
Cholesterol	No
Weight	58 kg
Height	160 cm
BMI	22 kg/m2

What would happen if...

Smoker	Systolic pressure (mmHg)	Weight (kg)
<input type="button" value="Yes"/> <input checked="" type="button" value="No"/>	<input type="text" value="110"/>	<input type="text" value="58"/>

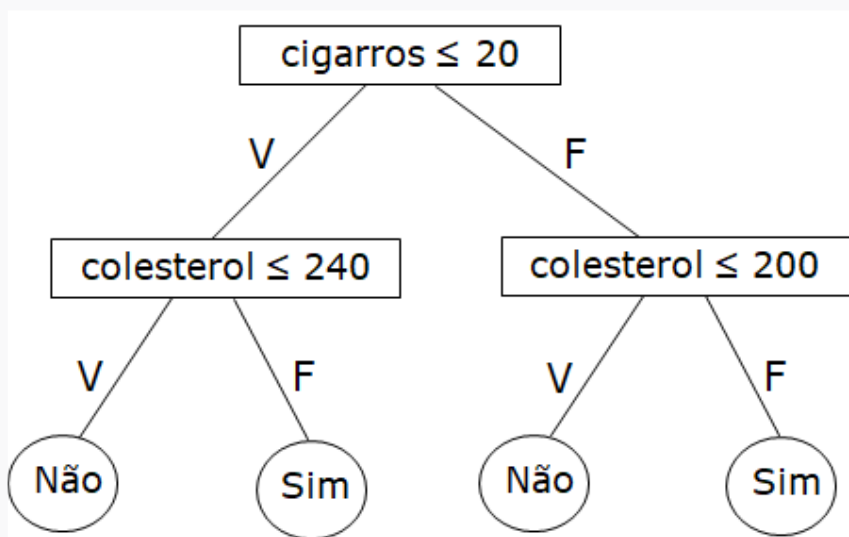
[Recalculate](#) [New CVD risk estimation](#)

Trabalhos Relacionados (2/2)

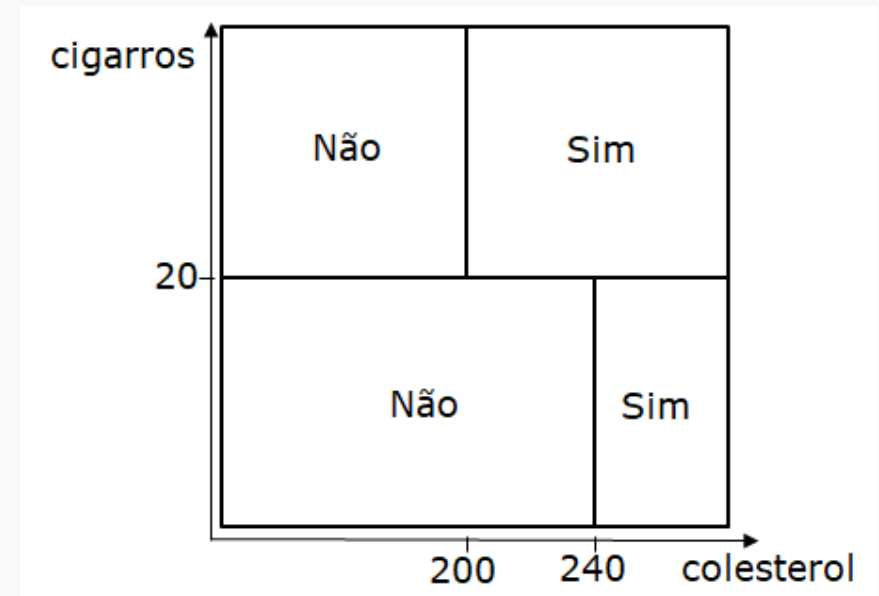
- **Previsão de DCV - Abordagem 2:**
Mineração de Dados / Aprendizado de Máquina
 - Um grande número combinações de fatores podem estar envolvidos no desenvolvimento de DCV.
 - Por isso, redes neurais e *ensembles* têm sido empregados prever o risco de DCV.
 - **Vantagem:** acurácia alta.
 - **Desvantagem:** modelos caixa-preta.
 - Neste trabalho, empregaremos o **BA** para:
 - Produzir um modelo **interpretável**, com a **mesma acurácia** de uma Random Forest.

Born-Again Tree Ensembles (BA) (1/3)

- Como funciona o BA?
 - As regras geradas por uma árvore de decisão (AD) dividem o espaço de atributos em diferentes **regiões**.
 - **Exemplo:** AD p/ classificar risco alto para DCV em função da taxa de colesterol e consumo de cigarros.



AD



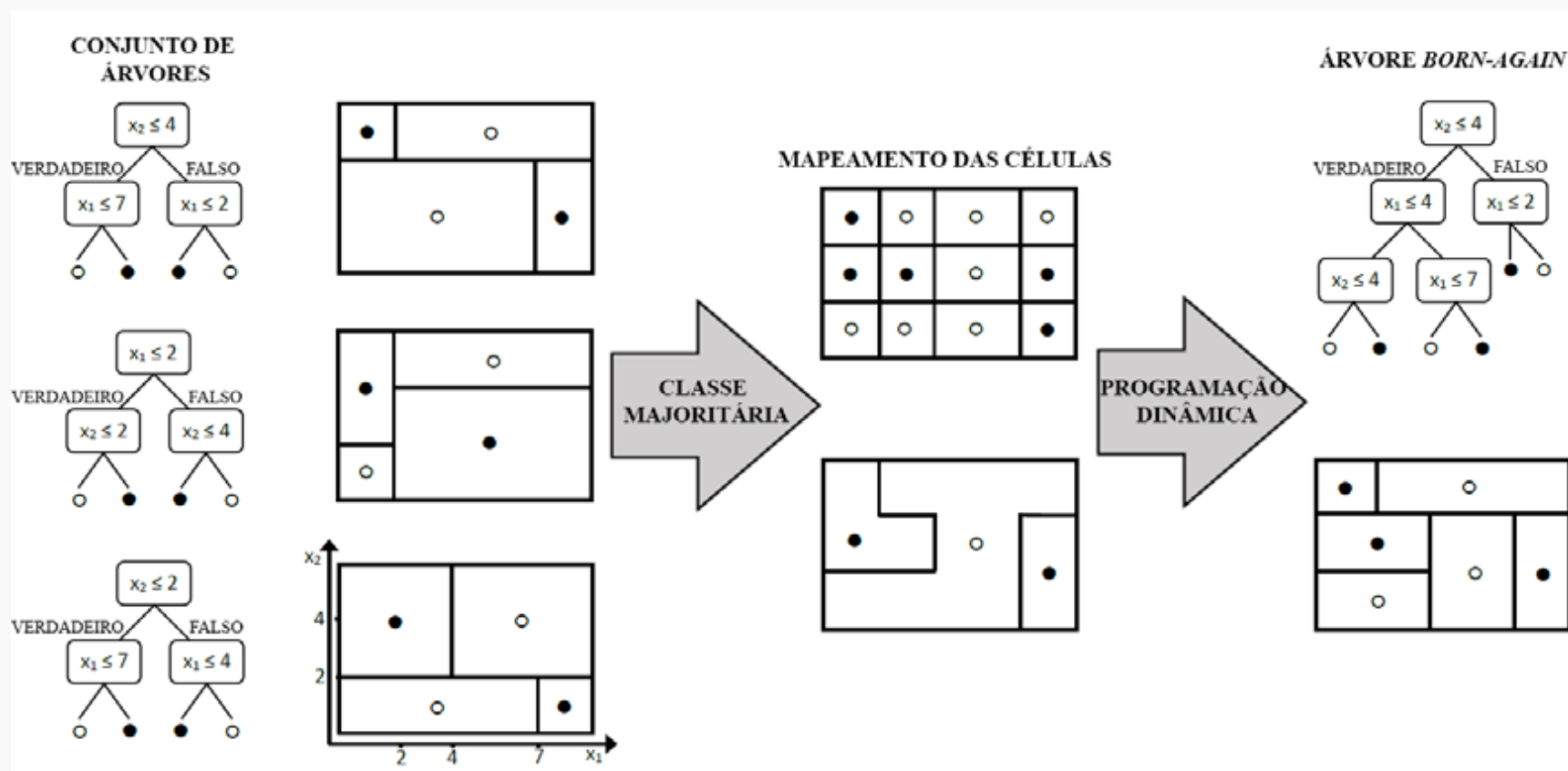
Regiões do espaço de atributos definidas pela AD

Born-Again Tree Ensembles (BA) (2/3)

- Como funciona o BA?
 - O BA processa as regiões definidas por cada árvore de uma RF para transformar a RF em uma única AD.
 - Ela é chamada de **árvore born-again**.
 - Possui o **mesmo desempenho preditivo** da RF.
 - O BA é o **primeiro algoritmo exato** para transformar uma RF em uma única AD:
 - Com o mesmo poder preditivo.
 - E fiel ao modelo RF em todo o seu espaço de atributos.

Born-Again Tree Ensembles (BA) (3/3)

- Como funciona o BA?
 - **Exemplo:** RF com 3 árvores e 2 atributos (x_1 e x_2)



Base de Dados (1/2)

- **STULONG/ENTRY**

- 1.417 pacientes europeus, sexo masculino, acima de 35 anos
 - Dados pessoais
 - Hábitos gerais
 - Exames físicos e laboratoriais.
- Classificados por especialistas em 3 grupos de risco para DCV:
 - baixo (grupo 0) – sem fator de risco, sem DCV
 - médio (grupo 1) – com fator(es) de risco, mas não têm DCV
 - alto (grupo 2) – alguma DCV já identificada

Base de Dados (2/2)

- **STULONG/ENTRY**

- Foram selecionados 6 atributos frequentemente apontados como fatores de risco.

Atributo	Categorias
Fumante	não fumante; 1-4 cig/dia; 5-14; 15-20; ≥ 21
Pressão Sanguínea	normal; normal/alta; alta
Colesterol	desejável; limítrofe; alto
Educação	fundamental; médio; especialização; superior
Faixa Etária	35-39; 40-44; 45-49; ≥ 50
IMC	baixo peso; normal; excesso de peso; obesidade mórbida

- Classes desbalanceadas:
 - baixo risco (grupo 0) – 22% da base
 - médio risco (grupo 1) – 69%
 - alto risco (grupo 2) – 9%

Resultados (1/6)

- Experimentos

- Comparação **CART x RF x BA** na base de dados **STULONG**

- Exp. 1: **desempenho preditivo** (CART x RF e BA)
 - Exp. 2: **comparação das ADs** (CART x BA)

- Setup Experimental

- *holdout* com 90% treino e 10% teste.
 - precisão, revocação e F1 (por classe).

- **CART** e **RF**: implementações da scikit-learn

- Árvore c/ máx. de 8 nós-folha e 3 níveis
 - RF com 10 árvores.

- **BA**: implementação da página do projeto

- recebe como entrada RF gerada pela scikit-learn e produz BA como saída

Resultados (2/6)

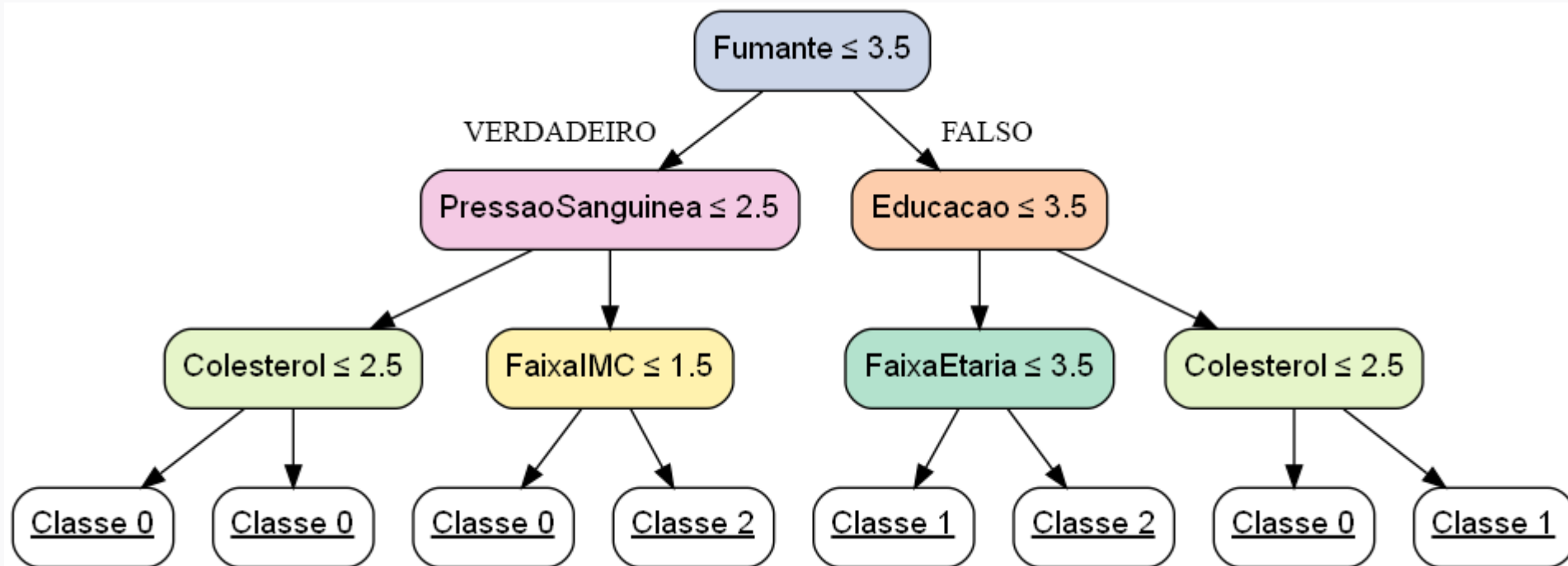
- (1) Desempenho preditivo por classe: CART x RF x BA

Grupo de Risco	Precisão		Revocação		F1	
	CART	RF e BA	CART	RF e BA	CART	RF e BA
baixo (classe 0)	0,685	0,711	0,850	0,960	0,759	0,817
médio (classe 1)	0,700	0,793	0,630	0,690	0,663	0,738
alto (classe 2)	0,480	0,705	0,367	0,550	0,414	0,618

- BA e RF têm mesmo desempenho preditivo.
 - ... pois BA é representação diferente da mesma função de decisão da RF.
- BA e RF superam CART em todas as medidas nas 3 classes.
- F1 obtido p/ BA e RF é 20% ao do CART na classe minoritária (classe 2).

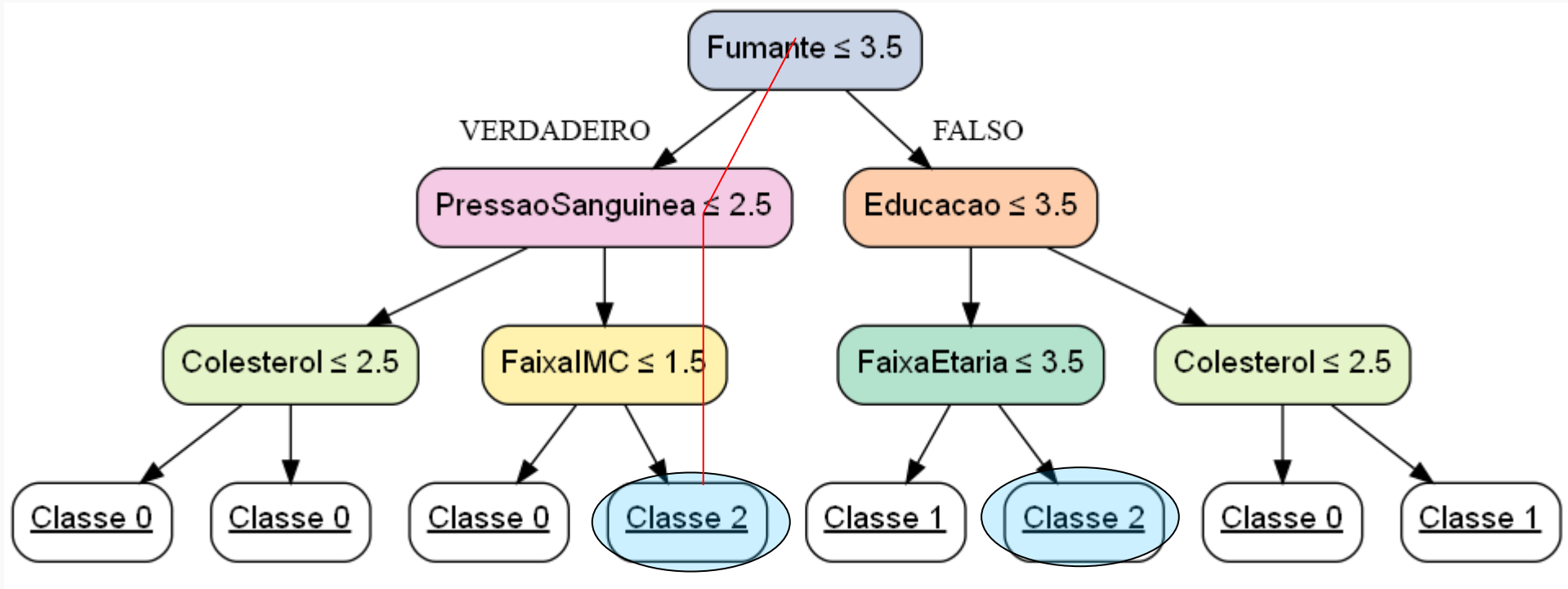
Resultados (3/6)

- (2) Comparação das ADs Geradas (CART x BA)
 - AD do CART



Resultados (4/6)

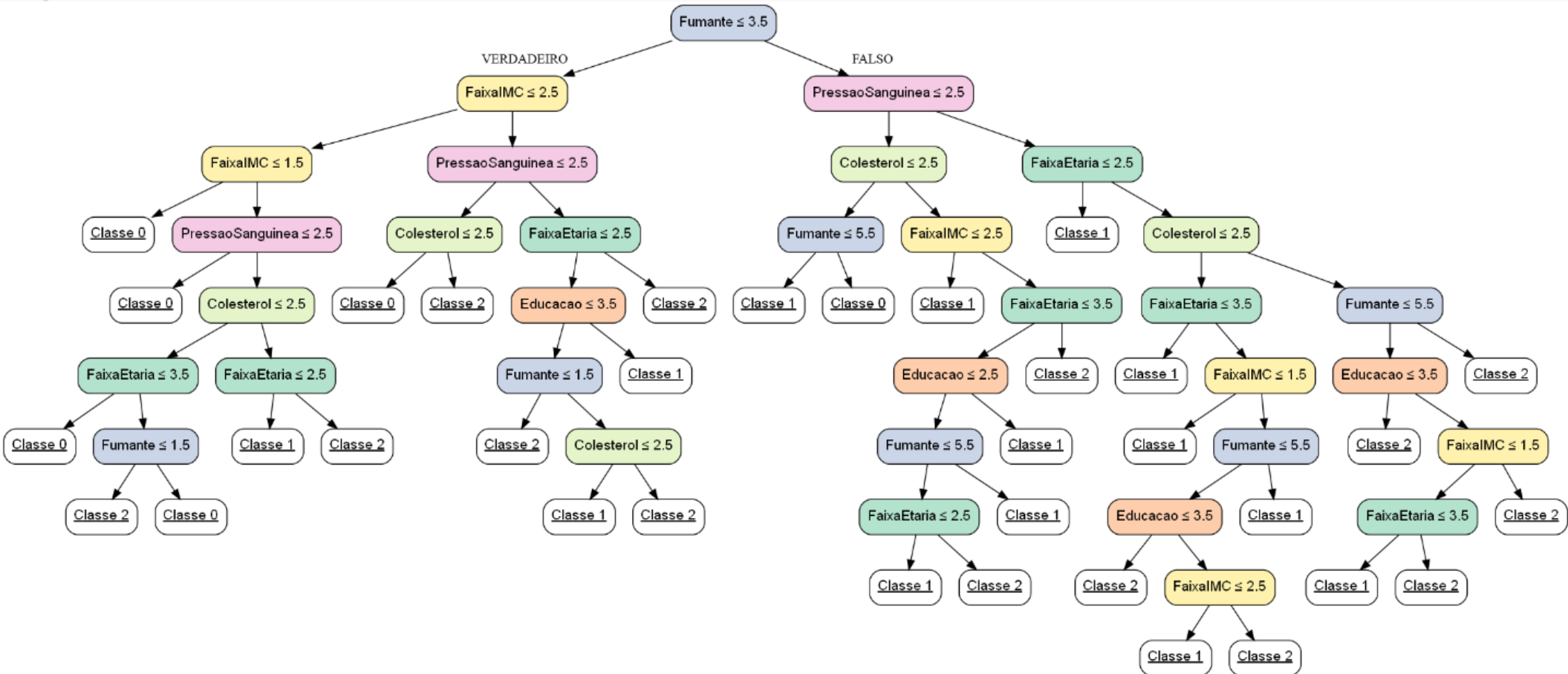
- (2) Comparação das ADs Geradas (CART x BA)
 - AD do CART



- 8 regras, apenas 2 envolvendo Classe 2 (risco alto)
- Uma delas é:
 - (Fuma ≤ 15cig/dia) & (Pressão = “alta”) & (IMC ≥ “normal”) → Risco “Alto”
- Nenhuma das regras cobre especificamente pacientes não fumantes.

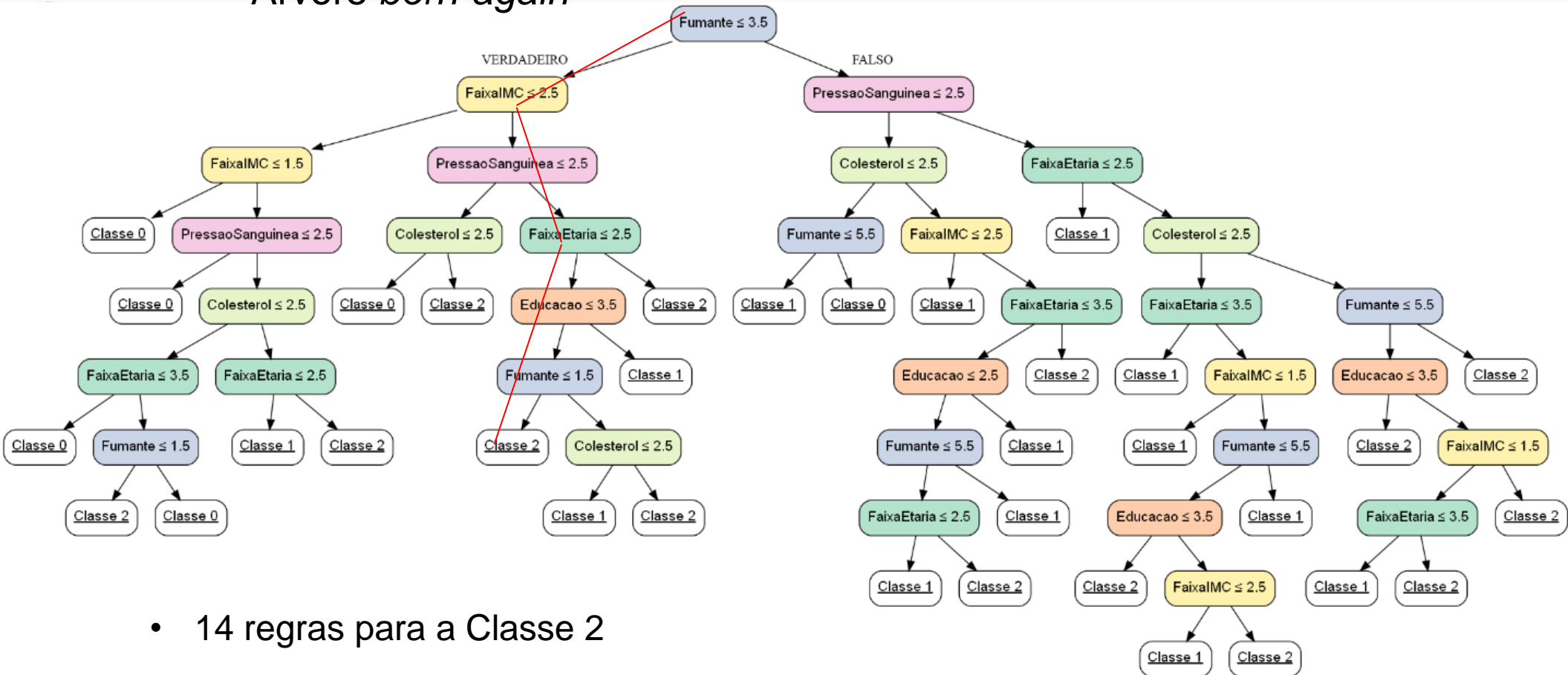
Resultados (5/6)

- **(2) Comparação das ADs Geradas (CART x BA)**
 - Árvore *born-again*



Resultados (6/6)

- (2) Comparação das ADs Geradas (CART x BA)
 - Árvore *born-again*



- 14 regras para a Classe 2
- “Fumante” continua na raiz.
- Uma regra para não fumante foi gerada:
 - (Fuma = “não”) & (IMC > “normal”) & (Pressão > “normal”) & (idade < 45) & (Educação ≠ “superior”) → Risco “Alto”

Comentários Finais

- **Resumo**

- Na base de dados STULONG, a **árvore born-again** gerada pelo **BA**:
 - Obteve o mesmo poder preditivo da RF.
 - Produz classificações tão fáceis de interpretar quanto a calculadora de risco da OPAS.
 - Possui poder descritivo superior a AD produzida pelo CART.

- **Trabalhos Futuros**

- Avaliar a **eficiência** do BA em bases de dados mais volumosas.
- Aplicar as regras geradas na entrada da calculadora de risco e comparar os resultados.



Obrigado !!!!