Revista da Estatística UFOP, Vol VI, 2017, ISSN 2237-8111

ÁRVORES DE DECISÃO E SUA IMPORTÂNCIA DIDÁTICA NO CONTEXTO ESCOLAR

Domingos Silva[1]

Resumo: As árvores de decisão têm vindo a ganhar popularidade junto da comunidade de estatísticos. Na base desta aceitação está o modo simples de acesso ao conhecimento, capaz de ser facilmente entendível pela generalidade das pessoas, ao mesmo tempo que são um excelente meio de representar classes, bem como outras informações/atributos presentes no conjunto de dados. De forma a determinar as diferenças ao nível da prática desportiva (variável resposta), na presença de algumas variáveis independentes (ano de escolaridade, curso, sexo, motivação para as disciplinas de Educação Física, Português e Matemática), aplicou-se a metodologia das árvores de decisão, tendo-se selecionado uma amostra aleatória de estudantes do ensino secundário. Foi usado o método CHAID. Os resultados mostram uma definição dos requisitos associados à prática e não prática de desporto, explicados, sobretudo, pelas variáveis motivação para a disciplina de EF e sexo. O modelo global é válido, explicativo e com capacidade preditiva.
Palavra-chave: árvore de decisão, CHAID, motivação, ensino secundário.

Abstract: The decision trees have been gaining popularity among statistical community. In the basis of this acceptance is the simple way to access knowledge, able to be easily understood by the majority of people, at the same time they are an excellent way to represent classes, as well other information/attributes present in dataset. In order to determine the differences in sports practice (dependent variable) in the presence of some independent variables (grade, course, sex, motivation to Physical Education, Portuguese and Mathematics subjects), it was applied the decision trees methodology, and secondary students were randomly included in the study. It was used CHAID method. The results showed a definition of the requirements associated with sports and non-sports practice, mainly explained by the motivation for PE subject and sex variables. The global model is valid, explanatory, and with predictive power.
Key-words: decision tree, CHAID, motivation, secondary school.

1. Introdução

As árvores de decisão têm vindo a ganhar popularidade junto da comunidade de estatísticos e de muitos investigadores de diferentes ramos do conhecimento. Na base desta aceitação está o facto de permitirem explorar, representar, identificar e classificar numa estrutura em árvore, as sequências de decisões e acontecimentos incertos suscetíveis de ocorrerem num problema de decisão, possibilitando, a partir daí, a adoção da estratégia mais adequada para a resolução desse problema (Quinlan, 1986, 1999; McLachlan, 2004). Genericamente, uma árvore de decisão é uma representação simples e esquemática do conhecimento e da forma mais eficiente de construir classificadores que definem classes com base nos atributos de um certo conjunto de dados (Quinlan, 1986). Ou seja, são uma técnica estatística que apresenta os resultados numa estrutura recursiva, e que são usadas na construção de modelos de análise de dados e na sua classificação (segmentação), a partir de uma variável resposta (variável de interesse) e de variáveis independentes (variáveis explicativas).

[1]conceitoestatistico@gmail.com

Uma das mais importantes características da árvore de decisão situa-se no seu tipo de representação, cuja estrutura hierárquica assenta numa árvore invertida, i.e., da raiz para as folhas, ou do geral para o particular. Por sua vez, a representação hierárquica ao proporcionar uma progressão da análise, atua como meio de previsão/classificação (Quinlan, 1986, 1999). Em todos os níveis da árvore são tomadas decisões acerca da estrutura do nível seguinte, até que sejam atingidos os nós terminais. Diversos autores (Quinlan, 1986; Banfield et al., 2007) referem-se à utilização destes modelos como sendo baseados no princípio de "dividir para conquistar". Isto significa que em cada nível da árvore, um problema mais complexo de previsão/classificação é decomposto em problemas mais simples, traduzindo-se na geração de nós descendentes, onde a heterogeneidade da variável resposta é atenuada, originando previsões com menos riscos para cada um dos nós gerados. Este método é supervisionado, onde a variável resposta é explicada à custa das variáveis independentes medidas em qualquer escala. Quando a variável resposta é de natureza contínua, as árvores de decisão designam-se de árvores de regressão (modelo de predição com regressão); quando é qualitativa, designam-se de árvores de classificação (modelo de predição com classificação). Neste último caso, o objetivo da previsão visa determinar a classe a que uma certa observação pertence.

Algumas das principais vantagens (Utgoff, 1989; Quinlan, 1999; Lewis, 2000; Deng et al., 2011), são: (i) simples de compreender e interpretar, (ii) ausência dos pressupostos clássicos dos modelos paramétricos, (iii) podem ser usadas diversas variáveis independentes e em diferentes escalas de medida, (iv) permite a adição de novos cenários possíveis, ajudando a determinar os piores, os melhores e os valores esperados para diferentes cenários, (v) possui técnicas próprias para lidar com os *missings*, (vi) não necessita de transformar as variáveis, (vii) proporciona a integração de relações complexas entre a variável resposta e as variáveis independentes, que não apenas uma relação linear, (viii) os resultados são facilmente interpretáveis, (ix) podem ser combinadas com outras técnicas de decisão.

Como principais desvantagens (Utgoff, 1989; Quinlan, 1999; Lewis, 2000; Deng et al., 2011), apontam-se: (i) são relativamente instáveis, uma vez que pequenas perturbações podem ocasionar grandes alterações no modelo que se pretende, (ii) fragmentação de conceito, i.e., podem ocorrer replicas de subárvores.

Dos vários algoritmos existentes (e.g., AID, *Automatic Interaction Detection*; ID3, *Iterative Dichotomizer 3*; C4.5; CHAID, *Chi-square Automatic Interaction Detection*; CART, *Classification and Regression Tree*; QUEST, *Quick, Unbiased, Efficient Statistical Tree*; THAID, *Theta Automatic Interaction Detection*), o mais usado é o CHAID, o qual se baseia na estatística do quiquadrado de Pearson onde os dados estão dispostos numa tabela de contingência entre as categorias da variável resposta e as categorias das variáveis independentes (se forem usadas variáveis contínuas, estas são previamente discretizadas em classes). O CHAID é bastante eficiente para a segmentação ou crescimento da árvore, onde as variáveis têm poder de predição (Magdison, 1993). Uma das principais vantagens deste algoritmo reside no facto do crescimento da árvore ser interrompido antes da ocorrência de *overfitting*, i.e., não tem tratamentos como o da poda. Por sua vez, a principal desvantagem do CHAID situa-se na necessidade de ter que lidar com grandes quantidades de dados de modo a ser possível garantir que a quantidade de observações em cada nó seja significativa (Hoare, 2004).

2. Objetivo

Com este trabalho pretendemos aplicar a metodologia das árvores de decisão, mais propriamente árvores de classificação, a um conjunto de 7 variáveis, das quais consideramos a «prática desportiva» como a variável resposta (tendo-se assinalado a opção de resposta "Não" como categoria alvo para identificar as suas características) e as restantes variáveis são independentes, conforme Tabela 1.

Tabela 1 – Variáveis em estudo.

Variáveis	Categorias	Tipo
Prática desportiva	sim/não	Dependente
Ano de escolaridade	10º/11º/12º	Independente
Curso	CCT/CCSE/CCSH/CTD	Independente
Sexo	masculino/feminino	Independente
Motivação para a disciplina de Português	reduzida/média/elevada	Independente
Motivação para a disciplina de Matemática	reduzida/média/elevada	Independente
Motivação para a disciplina de Educação Física	reduzida/média/elevada	Independente

3. Amostra

A amostra aleatória é formada por n=274 sujeitos, dos quais n=118 são do sexo masculino (43.1%) e n=156 do sexo feminino (56.9%), de idades compreendidas entre os 15.4 e os 20.6 anos (17.2 ±1.05), estudantes do ensino secundário (10º, 11º e 12º anos de escolaridade) de escolas públicas portuguesas.

4. Análise exploratória de dados

As características da amostra são mostradas na Tabela 2, onde se apresenta a frequência absoluta (n) e relativa (%) das variáveis categóricas do estudo: prática desportiva, ano de escolaridade, curso, sexo, motivação para as disciplinas de Educação Física (EF), Português (PTG) e Matemática (MAT), quer em termos globais quer por categoria da variável resposta.

Tabela 2 – Frequência absoluta (n) e relativa (%) de sujeitos da amostra relativamente à prática desportiva, ano de escolaridade, curso, sexo, motivação para as disciplinas de Educação Física, Português e Matemática.

	Análise Global		Prática Desportiva			
			Sim		Não	
	n	**%**	***n***	**%**	***n***	**%**
+ Prática desportiva						
Sim	108	39.4	---	---	---	---
Não	166	60.6	---	---	---	---
+ Ano de escolaridade						
10º ano	74	27.0	27	25.0	47	28.4
11º ano	81	29.6	33	30.6	48	28.8
12º ano	119	43.4	48	44.4	71	42.8
+ Curso						
Curso de Ciências e Tecnologias (CCT)	204	74.5	74	68.5	130	78.4
Curso de Ciências Socioeconómicas (CCSE)	11	4.0	3	2.8	8	4.8
Curso de Ciências Sociais e Humanas (CCSH)	27	9.9	7	6.5	20	12.0
Curso Tecnológico de Desporto (CTD)	32	11.7	24	22.2	8	4.8
+ Sexo						
Masculino	118	43.1	62	57.4	56	33.7
Feminino	156	56.9	46	42.6	110	66.3
+ Motivação para a disciplina de Educação Física (EF)						
Baixa	4	1.5	---	---	4	2.4
Média	80	29.2	18	16.7	62	37.3
Elevada	190	69.3	90	83.3	100	60.3

	Análise Global		Prática Desportiva			
			Sim		Não	
	n	%	*n*	%	*n*	%
+ Motivação para a disciplina de Português (PTG)						
Baixa	30	10.9	14	13.0	16	9.6
Média	196	71.5	73	67.6	123	74.1
Elevada	48	17.5	21	19.4	27	16.3
+ Motivação para a disciplina de Matemática (MAT)						
Baixa	83	30.3	32	29.6	51	30.7
Média	133	48.5	57	52.8	76	45.8
Elevada	58	21.2	19	17.6	39	23.5

5. Características da categoria alvo

A Tabela 3, apresenta o sumário da árvore de decisão, onde estão identificadas quer as especificações quer os resultados.

As especificações indicam o método de crescimento da árvore CHAID, a variável dependente e as variáveis independentes, o nível máximo de profundidade "Maximum Tree Depth" e o número mínimo de casos estabelecidos para os nós "Parent Node" e "Child Node".

Os resultados mostram a existência de 2 níveis de profundidade, cujas variáveis estatisticamente significativas na explicação do perfil de não ser desportista são a motivação a EF e o sexo, distribuídos por 5 nós, dos quais 3 são terminais.

Tabela 3 – Resumo da árvore de decisão: especificações e resultados obtidos.

Specifications	Growing Method	CHAID
	Dependent Variable	Prática.Desportiva
	Independent Variables	Ano, Curso, Sexo, Motivação EF, Motivação PORT, Motivação MAT
	Validation	None
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	Motivação EF, Sexo
	Number of Nodes	5
	Number of Terminal Nodes	3
	Depth	2

Pela Figura 1, observa-se que os cinco nós se repartem em caixas que contêm informação do número e percentagem de alunos do ensino secundário relativamente à prática desportiva, tal como é possível observar no diagrama da árvore. Três nós não têm ramificação (1, 3, 4), pelo que se designam de nós terminais.

No diagrama da árvore de decisão pode observar-se, a sombreado em cada caixa, a categoria prevista da variável resposta pelo método CHAID, que neste caso corresponde aos valores modais, dado não se terem definidos os custos de classificações incorretas.

A primeira caixa corresponde ao nó zero ou de raiz, que assinala a sombreado a categoria alvo "Não" da variável resposta (prática desportiva), mostrando que na amostra há uma maior probabilidade $60.6\% = \left(\frac{166}{274} \times 100\right)$ de ocorrerem não praticantes de desporto, pelo que os praticantes de desporto "Sim" são de 39.4%.

O primeiro nível de profundidade da árvore obtém-se através da motivação para a disciplina de Educação Física (Motivação EF), denotando ser esta a variável que melhor prevê os não praticantes de desporto ($\chi^2_{(1)} = 16.413; p < 0.001$) , segmentando-se a amostra em apenas duas das suas modalidades.

No primeiro nível a interpretação recai unicamente sobre o nó 1, não apenas por ser um nó terminal, mas também porque o CHAID lhe atribui 78.6% de probabilidade de pertencer à categoria alvo de não praticantes de desporto.
Retira-se deste nó que, por ser o nó terminal do primeiro nível, cerca de $39.8\% = \left(\frac{66}{166} \times 100\right)$ da totalidade dos não-desportistas se deve a alunos com baixa (n=4) e média (n=80) motivação EF, independentemente das restantes variáveis.
O segundo nível de profundidade mostra que o "sexo" é a variável que melhor prevê os não praticantes de desporto $\chi^2_{(1)} = 10.218; p = 0.001)$. O CHAID atribui cerca de $52.6\% = \left(\frac{100}{190} \times 100\right)$ de probabilidade de pertencer à categoria de não-desportista. Do nó 3, retira-se que cerca de $55.5\% = \left(\frac{61}{110} \times 100\right)$ da totalidade de moças não-desportistas, e do nó 4 que cerca de $69.6\% = \left(\frac{39}{56} \times 100\right)$ dos rapazes não-desportistas, se deve a alunos com alta motivação a EF. Os nós 3 e 4, embora terminais mostram probabilidades de 35.8% e 58.9% de incluírem praticantes de desporto, em moças e rapazes, respetivamente, ou 64.25% e 41.1% de probabilidade de incluir não-desportistas, em moças e rapazes, respetivamente. Embora o método CHAID preveja para o nó 4 a existência modal de praticantes de desporto, este nó tem associado um erro de classificação incorreta em não praticantes de desporto de 41.1%.

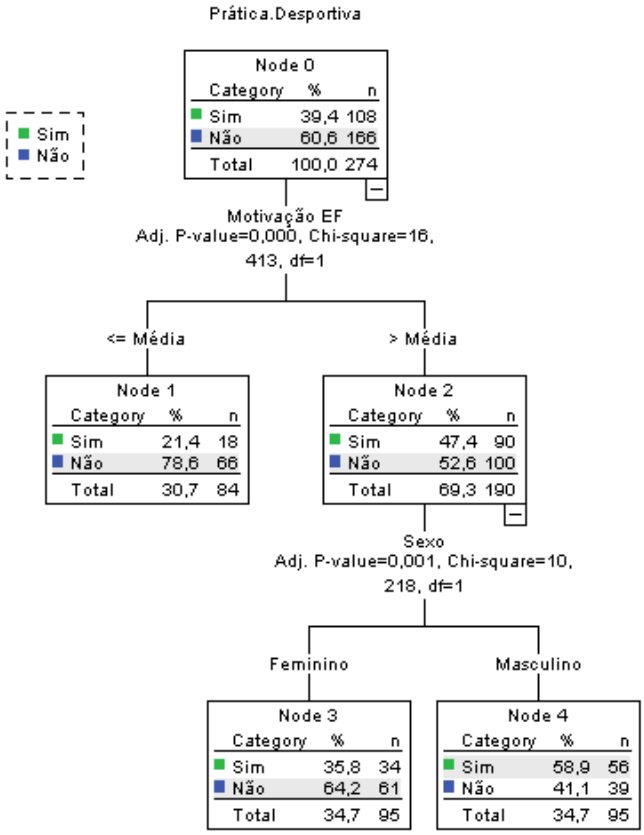


Figura 1 – Diagrama da árvore de decisão CHAID.

6. Classificação e risco

A Tabela de classificação (Tabela 4) informa sobre o número de previsões corretas e incorretas feitas pelo modelo. Assim, existem cerca de $76.5\% \left(= \frac{127}{166} \times 100\right)$ dos alunos

corretamente classificados como não-desportistas, que correspondem ao nó 1 (n=66) e nó 3 (n=61), sendo que o seu peso no total da amostra representa cerca de $65.3\% \left(= \frac{84+95}{274} \times 100\right)$.

As classificações corretas correspondem à soma dos valores da diagonal principal, ou seja, $\left(\frac{56+127}{274} \times 100 = 66.8\%\right)$, ao passo que as classificações incorretas, que no modelo sem custos coincidem com o *Risk Estimate*, correspondem à soma dos valores da diagonal secundária, ou seja, $\left(\frac{52+39}{274} \times 100 = 33.2\%\right)$.

Em resumo:

- Existem n=127 alunos classificados corretamente pelo CHAID como não-desportistas, provenientes do nó 1 (n=66) e do nó 3 (n=61).
- Existem n=56 alunos classificados corretamente pelo CHAID como desportistas provenientes do nó terminal 4 (n=56).
- Existem n=52 e n=39 alunos mal classificados, Sim/Não e Não/Sim, respetivamente, que corresponde a 33.2% do total. Os valores obtêm-se pela diferença entre a variável de resposta observada "Prática Desportiva" com a prevista "Predicted Value" pelo CHAID. Na presença de um resultado nulo, tal significa que não existe erro (Tabela 6).

Tabela 4 – Tabela de classificação (*classification*).

	Predicted			
Observed	**Sim**	**Não**	**Percent Correct**	**TOTAL**
Sim	56	52	51.9%	108
Não	39	127	76.5%	166
Overall Percentage	34.7%	65.3%	66.8%	274

A Tabela de Risco (Tabela 5) faz a comparação dos valores observados com os previstos pelo método CHAID. Esta tabela informa que, neste modelo sem custos, a taxa geral de classificações incorretas, é de 33.2%. Por sua vez, co9m apoio no erro-padrão (se=0.028) para o modelo sem custos, foram construídos intervalos de confiança a 90%, 95% e 99%.

$$IC90\% : 0.332 \pm 1.645 \times 0.028 =]0.286; 0.378[$$
$$IC95\% : 0.332 \pm 1.960 \times 0.028 =]0.277; 0.387[$$
$$IC99\% : 0.332 \pm 2.576 \times 0.028 =]0.260; 0.404[$$

Assim, para o modelo sem custos, o risco de classificações incorretas situa-se entre 28.6% e 37.8% para o IC90%, entre 27.7% e 38.7% para o IC95% e entre 26% e 40.4% para o IC99%.

Tabela 5 – Tabela do Risco (*Risk Estimate*).

Method	**Estimate**	**Std. Error**
Resubstitution	0.332	0.028

Pela Tabela 6, referente à frequência absoluta e relativa dos erros, verifica-se que n=183 observações (66.8%) não apresentam erro, ou seja, a diferença Prática Desportiva – Predicted Value=0. A categoria sem erro corresponde à soma das classificações corretas, i.e., Sim/Sim (n=56) + Não/Não (n=127) (Tabela 4).

Tabela 6 – Frequência absoluta (n), relativa (%) dos erros.

	n	**%**
-1 Desportista	52	19.0
0 Sem erro	183	66.8
1 Não-desportista	39	14.2
Total	274	100.0

7. Ganhos na categoria alvo

A categoria *Gains for Nodes* apenas surge porque se elegeu uma categoria alvo, para o efeito "não-desportista". Eventualmente, poder-se-ia eleger mais do que uma categoria alvo, o que não foi o caso. A Tabela 7 refere-se aos ganhos por nó, resumindo a informação prevista em termos absolutos e relativos nos nós terminais referentes à categoria alvo.

Nas colunas *Node* observa-se o número de elementos de cada nó e o seu peso relativo na amostra dos *n*=274 alunos do ensino secundário. Por exemplo, o nó 3 com *n*=95 alunos do sexo feminino e o nó 4 com *n*=95 alunos do sexo masculino, representa cada um cerca de $34.7\% = \left(\frac{95}{274} \times 100\right)$. Ou seja, *n*=95 moças e *n*=95 rapazes, independentemente da prática desportiva, têm motivação alta a EF. Contudo, recorrendo à Tabela 2, percebemos que destes, *n*=90 são desportistas e *n*=100 são não-desportistas.

As colunas *Gain* incluem o número de sujeitos da categoria alvo em cada nó e o seu peso na amostra dos *n*=166 alunos não-desportistas. Por exemplo, o nó 3 tem *n*=61 moças não-desportistas que representam cerca de $36.7\% = \left(\frac{61}{166} \times 100\right)$, ao passo que o nó 4 tem 39 moças não-desportistas que representam cerca de $23.5\% = \left(\frac{39}{166} \times 100\right)$.

A coluna *Response* refere-se à percentagem da categoria alvo por nó. Uma vez mais, o nó 3 tem cerca de $64.2\% = \left(\frac{61}{95} \times 100\right)$ de moças não-desportistas, e o nó 4 cerca de $41.1\% = \left(\frac{39}{95} \times 100\right)$.

A coluna *Index* é um índice que compara a proporção de não-desportistas por nó com a respetiva proporção na amostra. Novamente, o nó 3 tem o *Index* $106\% = \left(\frac{64.2\%}{60.6\%} \times 100\right)$ superior a 100%, o que significa que este nó tem maior concentração de moças não-desportistas. Opostamente, o nó 4 com o *Index* $67.8\% = \left(\frac{41.1\%}{60.6\%} \times 100\right)$ é o que regista a menor concentração de rapazes não-desportistas.

Tabela 7 – Frequência absoluta (*n*), relativa (%) dos ganhos por nó (*Gains for Nodes*).

	Node		Gain			
Node	***n***	**%**	***n***	**%**	**Response**	**Index**
1	84	30.7%	66	39.8%	78.6%	129.7%
3	95	34.7%	61	36.7%	64.2%	106.0%
4	95	34.7%	39	23.5%	41.1%	67.8%

A Tabela 8, ganhos por percentis (*Gains for Percentiles*), é similar à Tabela 7, expressando os nós terminais em função dos percentis. Dado que os percentis são estatísticas de ordem que acumulam até si uma certa percentagem, os nós apresentam-se por ordem decrescente de importância relativamente à presença de não-desportistas, conforme se pode observar nas colunas *Response* ou *Index* quer da Tabela 8 quer da Tabela 7.

Por exemplo, o P40 corresponde a $40\% \times 274 = 109.6\,(n \sim 110)$ sujeitos, os quais compreendem todos os sujeitos do nó 1 e uma parte de sujeitos do nó 3, como se mostra 110 = 84 (nó 1) + 26 (nó 3). A coluna de ganhos (*gains*) neste percentil tem a presença de $n = 82\,(= \underbrace{66}_{nó\,1} + \underbrace{26}_{parte\ do\ nó\ 3} \times 64.2\%)$ não-desportistas, os quais correspondem a ganhos de cerca de $49.4\% = \left(\frac{82}{166} \times 100\right)$.

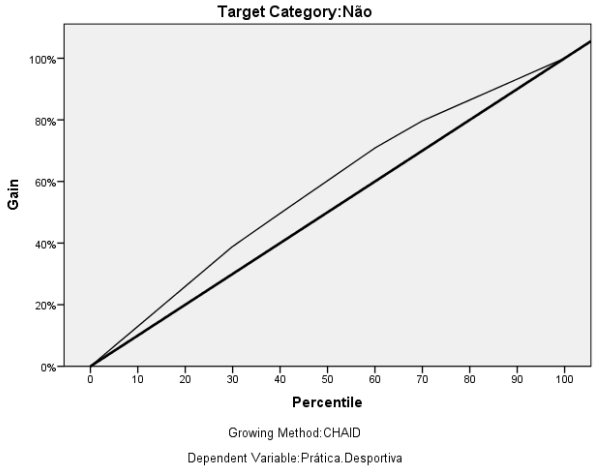
A coluna *Response* mostra que, para P40 (tomado como exemplo), 40% da amostra contém $74.55\% = \left(\frac{82}{110} \times 100\right)$ de não-desportistas. Por sua vez, a coluna *Index* mostra que 40% da amostra (recordamos que foi assumido como exemplo, P40) contém maior concentração de não-desportistas, uma vez que $123.06\% = \left(\frac{74.55\%}{60.58\%} \times 100\right)$.

Tabela 8 – Ganhos por percentis (*Gains for Percentiles*).

			Gain			
Percentile	**Nodes**	***n***	***n***	**%**	**Response**	**Index**
10	1	27	22	13.25%	81.48%	134.50%
20	1	55	43	25.90%	78.18%	129.05%
30	1	82	65	39.16%	79.27%	130.85%
40	1; 3	110	82	49.40%	74.55%	123.06%
50	3	137	100	60.24%	72.99%	120.49%
60	3	164	118	71.08%	71.95%	118.77%
70	3; 4	192	132	79.52%	68.75%	113.49%
80	4	219	144	86.75%	65.75%	108.53%
90	4	247	155	93.37%	62.75%	103.58%
100	4	274	166	100.00%	60.58%	100.00%

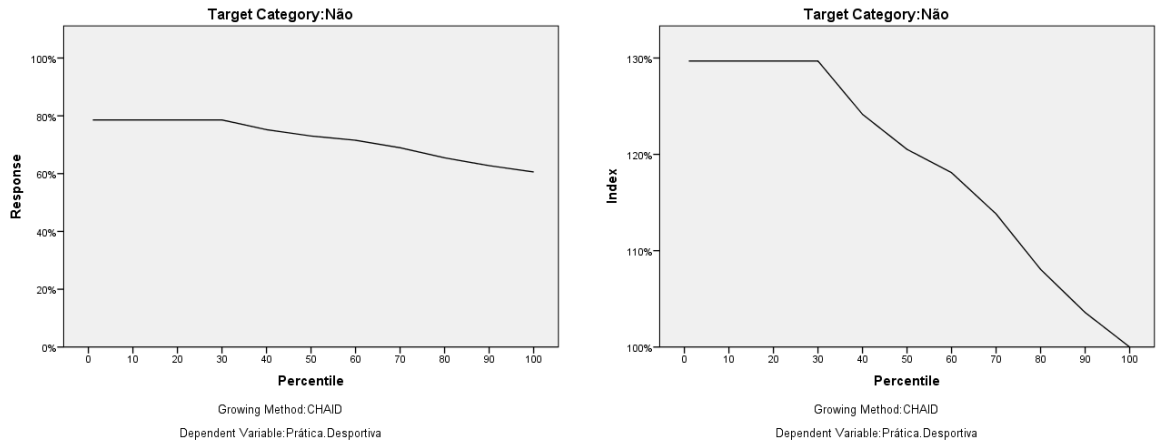
A Figura 2 mostra os percentis em ganhos da categoria alvo: "não praticante de desporto". A linha reta oblíqua traçada no gráfico indica não haver explicação sobre a categoria alvo. Assim, quanto maior for o afastamento da curva dessa linha mais explicativo é o modelo sobre a categoria. Neste estudo, verifica-se que o modelo proporciona alguma explicação sobre a categoria de não-desportistas.

Figura 2 – Percentis em ganhos da categoria alvo.



Pelas Figuras 3, observa-se que o andamento dos gráficos das respostas (esquerda) e dos índices (direita) é descendente, uma vez que os primeiros percentis acumulam os maiores valores, dado estarem ordenados por ordem decrescente de importância da categoria alvo (não-desportistas).

Figura 3 – Evolução dos percentis em ganhos da categoria alvo: respostas (esquerda) e índice (direita).



8. Validação

Construir árvores de decisão implica sempre a possibilidade de ocorrerem erros de classificações incorretas, suscetíveis de aumentar quando aplicados em novas amostras, pelo que é recomendável a validação do modelo CHAID. Caso o modelo seja generalizável, então a previsão com base nestas variáveis explicativas produz os mesmos resultados em qualquer outra subamostra da mesma população.

Das várias metodologias existentes (e.g., *crossvalidation, split-sample validation*: (i) *use random assignment – training sample*, (ii) *use variable*), optamos pela validação cruzada. Este procedimento consiste em dividir os dados iniciais em *n* subamostras (neste estudo, optamos por 10 subamostras) que, dentro do possível, contenham o mesmo número de sujeitos. Por sua vez, as *n* subamostras originam *m* árvores (m_1,m_2,...,m_n). Cada árvore é construída com base nos sujeitos das restantes *n*-1 subamostras, as quais desempenham o papel de treino (*training*) com a restante que lhe serve de teste (*test*) para validar os dados estimando o erro associado às classificações incorretas.

De acordo com o IBM SPSS (2012), a divisão em 10 subamostras é considerada um bom compromisso entre a precisão e a complexidade do modelo. Paralelamente, Banfield et al. (2007) recomendam a validação cruzada quando $n<1000$, o que é o caso do presente estudo (n=274).

A tabela *Risk* (Tabela 9), compara a taxa de risco para a amostra global (*Resubstitution*) com a taxa média dos erros estimados em cada subamostra (*Cross-Validation*). Se os valores forem semelhantes, então a informação da validação possibilita a aplicação do modelo a outras amostras provenientes da mesma população. Neste estudo, o método de validação cruzada tem uma taxa de classificações incorretas ligeiramente superior ao modelo sem custos (~2.6%), ou seja, 35.8% contra 33.2%. Por sua vez, o erro-padrão da validação cruzada (*se*=0.029) é semelhante ao erro-padrão do modelo sem custos (*se*=0.028). Assim, podemos afirmar que a validação pelo procedimento de *cross-validation* é suficientemente realista, apesar do valor de estimativa ser ligeiramente superior ao valor de estimativa do modelo sem custos.

Tabela 9 – Tabela do Risco (*Risk*)

Method	Estimate	Std. Error
Resubstitution	0.332	0.028
Cross-Validation	0.358	0.029

9. Previsão

Depois da validação do modelo, o mesmo pode ser usado noutras amostras para prever os resultados. Foram adicionadas 3 novas variáveis, as quais proporcionam as seguintes informações:

nod_001	o número previsto do nó
pre_001	a categoria prevista de Y
prb_001	a probabilidade prevista

Tomemos como exemplo a Figura 5. A primeira linha para as variáveis nod_001=1, pre_001=2 e prb_001=0.79, indica que este sujeito se localiza no nó 1 e, com 79% de probabilidade de ser não-desportista (codificação: 1-desportista; 2-não desportista).

	NodeID	PredictedValue	PredictedProbability_1	PredictedP...	Erros	nod_001	pre_001	prb_001
1	1	2	,21	,79	0	1	2	,79
2	3	2	,36	,64	0	3	2	,64
3	1	2	,21	,79	0	1	2	,79
4	3	2	,36	,64	0	3	2	,64
5	1	2	,21	,79	0	1	2	,79
6	1	2	,21	,79	0	1	2	,79
7	1	2	,21	,79	0	1	2	,79
8	1	2	,21	,79	-1	1	2	,79
9	3	2	,36	,64	-1	3	2	,64
10	3	2	,36	,64	0	3	2	,64

Figura 5 – Observação dos 10 primeiros casos.

```
/* Node 1 */.
DO IF (VALUE(MOTIV_EF) LE 2).
COMPUTE nod_001 = 1.
COMPUTE pre_001 = 2.
COMPUTE prb_001 = 0.785714.
END IF.
EXECUTE.

/* Node 3 */.
DO IF (SYSMIS(MOTIV_EF) OR (VALUE(MOTIV_EF) GT 2)) AND (SYSMIS(Sexo) OR
VALUE(Sexo) NE 1).
COMPUTE nod_001 = 3.
COMPUTE pre_001 = 2.
COMPUTE prb_001 = 0.642105.
END IF.
EXECUTE.

/* Node 4 */.
DO IF (SYSMIS(MOTIV_EF) OR (VALUE(MOTIV_EF) GT 2)) AND (VALUE(Sexo) EQ 1).
COMPUTE nod_001 = 4.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.589474.
END IF.
EXECUTE.
```

10. Conclusão

O objetivo deste trabalho foi aplicar a metodologia das árvores de decisão, usando o método CHAID, a um conjunto de variáveis relacionadas com a prática desportiva (variável dependente) (sim/não, sendo "não" a categoria de referência), o sexo e variáveis do envolvimento escolar (variáveis independentes). De uma forma geral (i) foram identificadas duas variáveis estatisticamente significativas na explicação do perfil de "não-desportista": motivação para EF e sexo, (ii) no 1º nível de profundidade, a variável "motivação para a disciplina de EF" é a que melhor prevê os não-desportistas $\chi^2_{(1)} = 16.413; p < 0.001)$, no 2º nível de profundidade, a variável "sexo" é a que melhor prevê os não-desportistas $\chi^2_{(1)} = 10.218; p = 0.001)$, (iii) em termos globais, existem n=91 sujeitos mal classificados (33.2%), dos quais [observado/predito] n=52 (48.1%) são Sim/Não e n=39 (23.5%) são Não/Sim, (iv) para o modelo sem custos, considerando um IC95%, o risco de classificações incorretas situa-se entre (27.7%; 38.7%), (v) o modelo é explicativo sobre a categoria alvo (não-desportista), (vi) não obstante a validação cruzada produzir um valor de estimativa ligeiramente superior ao valor da estimativa do modelo sem custos, pode-se afirmar que o modelo é válido e com capacidade preditiva.

Referências

[1]BANFIELD, R.E.; HALL, L.O.; BOWYER, K.W.; KEGELMEYER, W.P. A Comparison of Decision Tree Ensemble Creation Techniques. IEEE Transactions on Pattern Analysis and Machine Intelligence, v.29(1), p.1-8, Jan. 2007.

[2]DENG,H.; RUNGER, G.; TUV, E. Bias of importance measures for multi-valued attributes and solutions. Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN, 2011).

[3]HOARE, R. Using CHAID for Classification Problems. New Zealand Statistical Association Conference. Wellington, New Zealand, p.115-120, Jul. 2004.

[4]IBM SPSS. IBM SPSS Decision Trees 21, 2012.

[5]LEWIS, R.J. An introduction to classification and regression tree (CART) analysis. Annual Meeting of the Society for Academy Emergency Medicine, San Francisco, California. USA, p.1-14, 2000.

[6]MAGDISON, J. The Use of the New Ordinal Algorithm in CHAID to Target Profitable Segments. The Journal of Database Marketing, v.1, p.29-48, 1993.

[7]MCLACHLAN, G. Discriminant Analysis and Statistical Pattern Recognition. John Wiley & Sons, New York, 2004.

[8]QUINLAN, J.R. Introduction of decision trees. Machine Learning, v.1, p.81-106, 1986.

[9]QUINLAN, J.R. Simplifying decision trees. International Journal of Human-Computer Studies, v.51(2), p.497-510, Aug. 1999.

[10]UTGOFF, P.E. Incremental induction of decision trees. Machine learning, v.4(2), p.161-186, 1989.