Análise de perfis de desempenho académico pelo método CHAID

José L. C. Verdasca
Universidade de Évora, 2018

Introdução

Existe abundante literatura sobre a influência de fatores contextuais no desempenho académico dos alunos e das escolas[1]. O modelo de análise esquematizado na figura 1, incorpora, no seu delineamento, algumas das conclusões desses estudos relacionadas com a importância de fatores contextuais no desempenho escolar dos alunos e das unidades orgânicas escolares nos seus diversos níveis de análise. Deste modo, no esquema apresentado configura-se um conjunto de interações múltiplas entre variáveis resultado (desempenho académico em Matemática) e um conjunto de preditores de natureza sociográfica (género, idade), de origem familiar (capital escolar e situação socioeconómica dos pais), de itinerário escolar (repetência), de organização escolar (turma) e de comportamento/absentismo escolar (faltas). Algumas das variáveis relacionadas com atitudes e comportamentos escolares podem ser do tipo híbrido e circular, no sentido de que podem ser já elas próprias consequência (resultado) de efeitos de fatores contextuais e, simultaneamente, preditores diretos do desempenho académico. Está, de certo modo, neste registo a variável 'Faltas3P_1213', entre outras, cuja maior ou menor intensidade decorre da conjugação de certas caraterísticas contextuais dos alunos e, enquanto indicador de absentismo escolar, será, por certo, condicionante dos resultados académicos e da qualidade desses resultados.

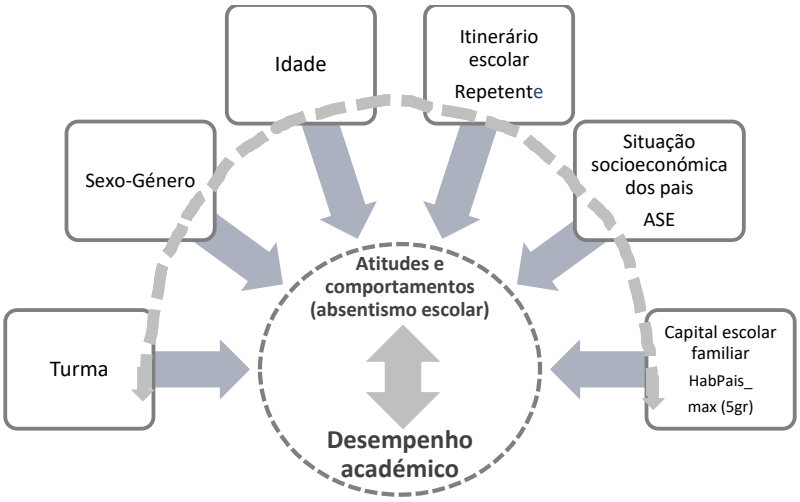

Figura 1- Esquematização das interações

Método

A análise dos perfis de desempenho académico foi realizada pelo método das árvores de classificação, através do algoritmo *CHAID* (*Chi-square Automatic Interaction Detector*)[2], tendo por base uma amostra de alunos do 6º ano de escolaridade, dos quais se obteve informação registada

[1] Referindo apenas alguns dos estudos e relatórios institucionais mais recentes e de abrangência nacional, veja-se, por exemplo: CNE, 2013, 2014; Canto e Castro *et al*, 2014; Justino *et al*, 2014. No que concerne à turma, enquanto unidade-sujeito de análise, a literatura é, desta perspetiva, bem mais escassa e os estudos existentes são distanciados no tempo e, de algum modo, circunscritos a um determinado território (Verdasca, 2002).

[2] Ver, por exemplo, a este propósito, Pestana & Gageiro (2009) e IBM-SPSS (2012).

em base de dados (Base K6_1213) relativamente a potenciais variáveis explicativas do desempenho académico em Matemática.

Resultados

Apresentam-se de seguida o quadro resumo com as respetivas especificações e resultados decorrentes da aplicação do algoritmo EXHAUSTIVE CHAID.

Quadro I – Resumo e especificações do modelo arbóreo

Model Summary

Specifications	Growing Method	EXHAUSTIVE CHAID
	Dependent Variable	r_M3P_1213
	Independent Variables	Turma, Sexo, Idade, Repetente, ASE, HabPais_max, Faltas_1P_1213
	Validation	Cross Validation
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	2
	Minimum Cases in Child Node	1
Results	Independent Variables Included	Repetente, Idade, Faltas_1P_1213, Sexo, HabPais_max
	Number of Nodes	13
	Number of Terminal Nodes	8
	Depth	3

No diagrama da árvore de classificação gerada pelo algoritmo CHAID (ver Figura 2), os nós estão representados através de caixas que contêm a informação do número de alunos e respetiva percentagem com nível positivo ou com nível negativo a Matemática. Em cada caixa, a categoria modal prevista da variável de resposta aparece sombreada a cinzento. Na primeira caixa (nó 0), está assinalada a sombreado a categoria 'Nível Positivo', mostrando que na amostra geral há uma maior probabilidade de ocorrerem níveis positivos em Matemática (41/77*100=53,2%) do que níveis negativos (36/77*100=46,8%).

O primeiro nível de profundidade da árvore obtém-se através da condição 'ser ou não repetente', revelando que, do total dos 77 alunos da amostra, 49 (63,6%) não são repetentes e concluindo-se ser esta variável a que melhor prevê os alunos com nível positivo em Matemática ao gerar a primeira partição da amostra geral nas duas categorizações possíveis (sim ou não) da variável 'Repetente'. Apesar de no primeiro nível de profundidade não existirem nós terminais, mas apenas nós intermédios, retira-se do nó 1, relativo ao segmento dos alunos não repetentes, que o algoritmo CHAID lhe atribui 81,6% de probabilidade em pertencer à categoria de alunos com nível positivo em Matemática. Retira-se ainda deste nó que 98% (40/41*100) da totalidade dos alunos da amostra geral com nível positivo na disciplina de Matemática se deve a alunos não repetentes.

No segundo nível de profundidade emergem como variáveis estatisticamente mais importantes na subsegmentação dos ramos do primeiro nível, a idade e o capital escolar da família (HabPais) e surgem os nós 5, 6 e 7, respetivamente com 5, 26 e 2 casos, como primeiros nós terminais e com uma probabilidade de 80%, 0% e 50% de incluir alunos com positiva a Matemática. Os respetivos perfis dos alunos que integram estes nós podem ser descritos do seguinte modo: *Perfil (Nó 5)- alunos do 6º ano não repetentes com idade superior a 12 anos; Perfil (Nó 6)- alunos repetentes cujos pais têm habilitações académicas iguais ou inferiores ao 9º ano ou habilitações desconhecidas; Perfil (Nó 7)- alunos repetentes cujos pais têm o 12º ano de escolaridade*.

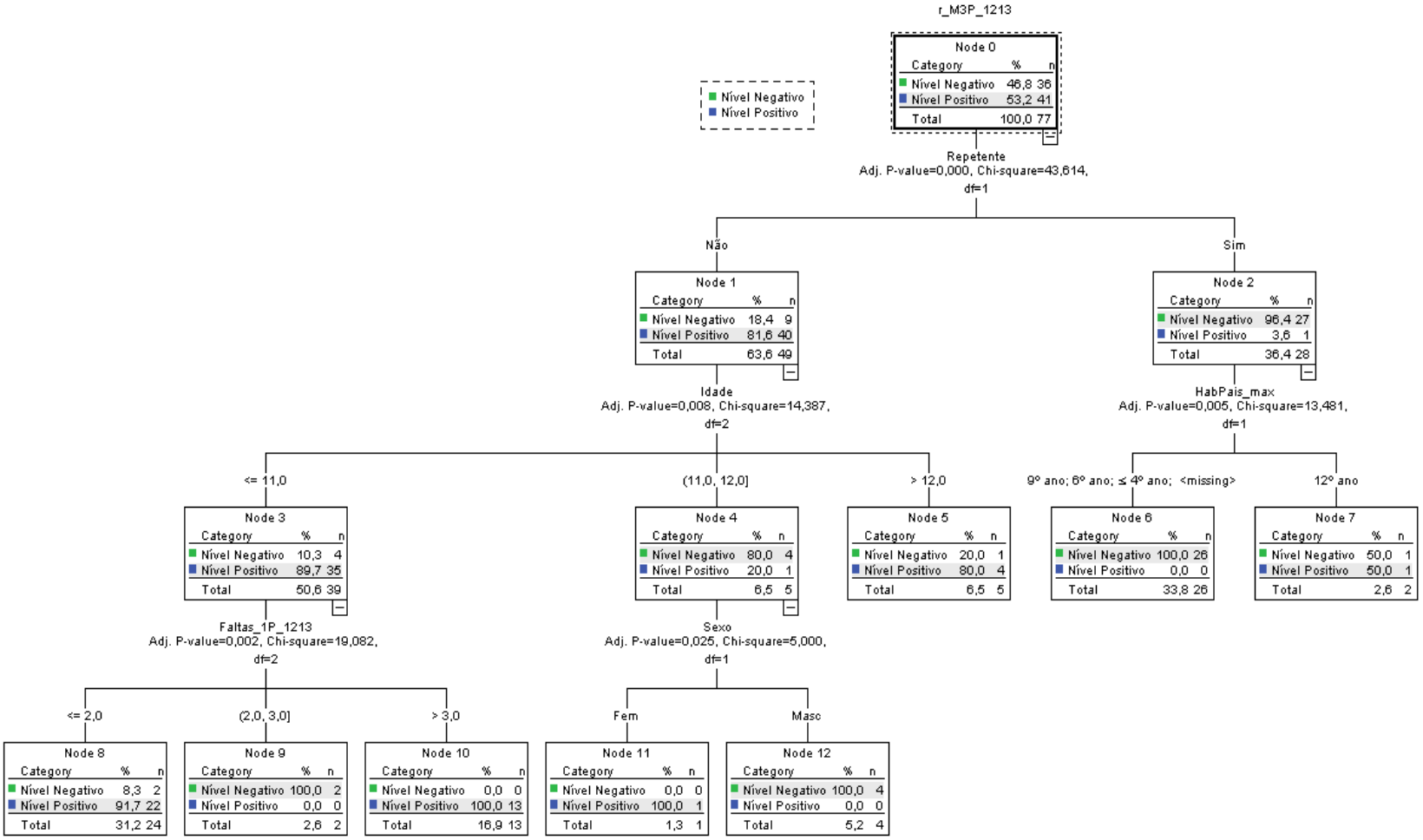

Figura 2- Estrutura hierárquica do desempenho académico em Matemática de alunos K6

Os 40 alunos não repetentes com positiva a Matemática e com idade igual ou inferior a doze anos distribuem-se por cinco nós terminais (8, 9, 10, 11 e 12). Face à categoria alvo sinalizada (alunos com nível positivo), como decorre da análise do quadro II, são de destacar pelo seu grau de concentração relativo (*Index*) os nós 10, 11, 8 e 5, com registos de aproximadamente 188%, 188%, 172% e 150%, respetivamente. Para além da coluna *Index*, estão disponíveis no quadro II outras medidas informativas sobre os ganhos absolutos e relativos da categoria alvo por nó terminal, designadamente as indicadas nas colunas *Node*, *Gain* e *Response*.

Quadro II – Ganhos na categoria alvo por nó terminal

Gains for Nodes

Node	Node		Gain		Response	Index
	N	Percent	N	Percent		
10	13	16,9%	13	31,7%	100,0%	187,8%
11	1	1,3%	1	2,4%	100,0%	187,8%
8	24	31,2%	22	53,7%	91,7%	172,2%
5	5	6,5%	4	9,8%	80,0%	150,2%
7	2	2,6%	1	2,4%	50,0%	93,9%
6	26	33,8%	0	0,0%	0,0%	0,0%
12	4	5,2%	0	0,0%	0,0%	0,0%
9	2	2,6%	0	0,0%	0,0%	0,0%

Growing Method: EXHAUSTIVE CHAID
Dependent Variable: r_M3P_1213

As colunas *Node* contêm informação sobre o número de elementos de cada nó e o seu peso relativo na amostra dos 77 alunos K_6 (6º ano de escolaridade). O nó 6, com 26 alunos, é o nó com maior peso relativo representando cerca de um terço da amostra total de alunos em análise; ao contrário, o nó 11 com apenas um caso, é o nó com menor peso relativo amostral, representando apenas 1,3% do total.

As colunas *Gain* incluem o número de elementos da categoria alvo em cada nó e o seu peso relativo na subamostra dos 41 alunos com nível positivo a Matemática. Assim, o nó 8, com 22 alunos com positiva a Matemática, destaca-se claramente dos restantes, pois, contém 53,7% (22/41*100) da subamostra geral de alunos da categoria 'nível positivo'. A contrastar, os nós 6, 12 e 9, com zero alunos com positiva na disciplina de Matemática e a que correspondem ganhos de 0%.

A coluna *Response* representa o peso relativo da categoria alvo intranó. Os valores de 100% registados nos nós 10 e 11, significam que nestes nós a frequência absoluta da categoria alvo é igual ao número de total de casos que compõem esses nós. Por outro lado, e em contraste com os nós anteriores, estão os nós 6, 12 e 9, com zero registos na categoria alvo, e a que correspondem pesos relativos intranó de 0%.

Outros indicadores relevantes para a análise são os que decorrem do quadro das classificações e do risco (Quadros III-A e III-B), os quais disponibilizam informação sobre o número de previsões corretas e incorretas e sobre o índice global de risco estimado. Assim, por exemplo, há 100% [(4+1+22+13+1)/41] dos alunos corretamente classificados com positiva a Matemática, correspondentes, respetivamente, aos nós terminais 5, 7, 8, 10 e 11 e o seu peso no total da amostra representa 58,4% [(5+2+24+13+1)/77].

Quadros III-A e III-B

Classification

Observed	Predicted		
	Nível Negativo	Nível Positivo	Percent Correct
Nível Negativo	32	4	88,9%
Nível Positivo	0	41	100,0%
Overall Percentage	41,6%	58,4%	94,8%

Growing Method: EXHAUSTIVE CHAID
Dependent Variable: r_M3P_1213

Risk

Method	Estimate	Std. Error
Resubstitution	,052	,025
Cross-Validation	,156	,041

Growing Method: EXHAUSTIVE CHAID
Dependent Variable: r_M3P_1213

Recorrendo ao Quadro III-A (*Classification*), as classificações corretas correspondem à soma dos valores da diagonal principal [(32+41)/77*100=94,8%] e as classificações incorretas (risco estimado) correspondem à soma dos valores da diagonal secundária [(0+4)/77*100=5,2%]. Por outro lado, o erro padrão (*Std. Error*=0,025) permite a construção de intervalos de confiança, que a 95% projetam um risco de classificações incorretas compreendido entre 0,003% e 10,1% (0,052±1,96*0,025) e com validação cruzada entre 7,6% e 23,6%.

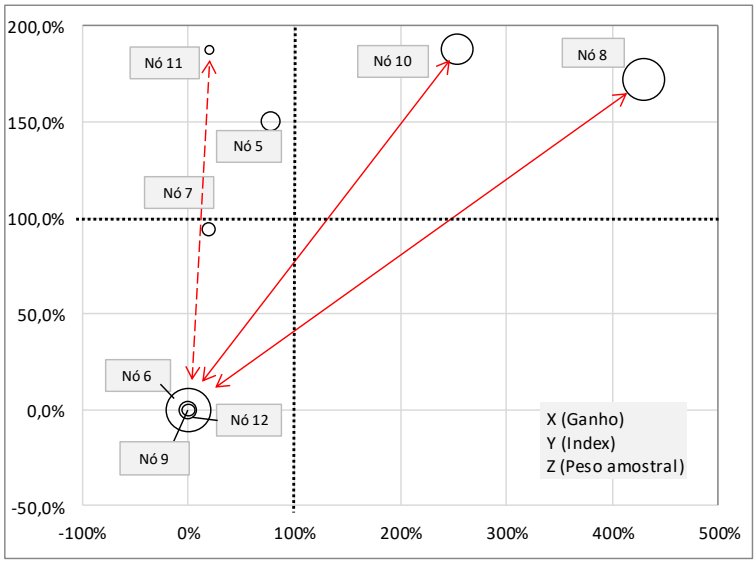

Figura 3- Projeção gráfica dos pares de maior contraste pelo critério dos quadrantes ímpares e máxima amplitude

Retomando os dados constantes do quadro II (Ganhos na categoria alvo por nó terminal) e esmiuçando um pouco mais a análise, pelo seu contraste, em termos de concentração (*index*) e ganho (*gain*) relativamente à categoria alvo (% de níveis positivos em Matemática), por um lado e, por outro, pelo seu peso relativo na amostra, pelo critério dos quadrantes ímpares e máxima amplitude os pares de nós 10-6 e 8-12 assumem particular relevância interpretativa ao constituírem-se como os pares de maior contraste e aos quais se poderá ainda acrescentar o par 11-9, por apresentar um diferencial *índex* superior a 100%, apesar do não cumprimento do critério dos quadrantes ímpares, todavia, cumprindo o critério da pertença a quadrantes diferentes (Verdasca, 2013). Com efeito, a inclusão do

terceiro par aumenta a representatividade da análise em termos amostrais de 83% para 91%. Por outro lado, no que concerne à formação dos pares, no caso em análise os valores de 0% em *index* e ganho, deslocou para o peso amostral o critério prioritário, daí a sequência de entrada dos nós 6, 12 e 9.

Em jeito de conclusão, da comparação dos perfis dos elementos dos pares constituídos pelo critério *índex* "grau de concentração relativo da categoria alvo do respetivo nó face à categoria alvo do nó inicial" ressalta que: i) relativamente ao primeiro par (nó 10 vs nó 9), os seus elementos distinguem-se apenas pelo número de faltas dadas no 1º período, ainda que um dos elementos tenha uma reduzida representatividade em termos da amostra; ii) no caso do segundo par (nó 12 vs nó 11), a variável distintiva é o género e também neste caso ambos os elementos apresentam reduzida representatividade amostral; iii) no caso do 3º par (nó 8 vs nó 6), este representa 64% dos casos e as caraterísticas dos seus elementos dão-lhe configurações em termos de perfil bastante diferentes, ou seja, no caso do nó 8, 22 dos 24 alunos que o integram obtiveram positiva a Matemática, não são repetentes, têm idades iguais ou inferiores a 11 anos e registaram no máximo duas faltas à disciplina, enquanto que os 26 alunos que compõem o nó 6 todos eles tiveram negativa, são repetentes e os seus pais têm como habilitações académicas o 9º ano de escolaridade ou menos. Estes resultados vêm, por um lado, ao encontro do esquema inicial de interações apresentado na figura 1 e, por outro lado, acrescentam-lhe, no contexto da presente amostra e neste nível de ensino, uma certa hierarquização explicativa dos fatores preditores do desempenho académico em Matemática.

Referências

Justino, D. e Miguéns, M. (dir. e coord.) (2014). *Estado da Educação 2013*. Lisboa: CNE.

Justino, D. e Miguéns, M. (dir. e coord.) (2013). *Estado da Educação 2012*. Lisboa: CNE.

Justino, D., Pascueiro, L., Franco, L., Santos, R., Almeida, S. e Batista, S. (2014). *Atlas da Educação – Contextos sociais e locais do sucesso e insucesso: Portugal 1991-2012*. Lisboa: CESNOVA.

Castro, L., Santos, J., Pereira, T. e Vitorino, A. (2014). *Modelos para comparação estatística dos resultados académicos em escolas de contexto análogo: Painel de dados para apoio à avaliação externa das escolas.* Lisboa: MEC-DGEEC.

IBM SPSS (2012). *Decision Trees 21*. (ftp://public.dhe.ibm.com/software/analytics/spss/documentation /statistics/21.0/en/client/Mauals/IBMSPSSDecisionTrees.pdf)

Pestana, M. e Gageiro, J. (2009). *Análise Categórica, Árvores de Decisão e Análise de Conteúdo em Ciências Sociais e da Saúde com o SPSS*. Lisboa: Lidel, Edições Técnicas.

Verdasca, J. (2002). *Desempenho escolar, dinâmicas de evolução e elementos configuracionais estruturantes. Os casos do 2.º e 3.º ciclos do ensino básico nos municípios de Évora e Portel*. Évora: Universidade de Évora.

Verdasca, J. (2013). *Instrumentos metodológicos e de análise exploratória de dados*. Coletânea de textos e documentos de apoio à disciplina de Métodos e Técnicas de Administração Educacional. Évora: UEvora (polic.)