Algoritmo CART: Previsão do Desempenho na Matemática do Secundário

Nélia Pereira Cabete
Escola Básica 2º, 3º ciclos José Afonso- Alhos Vedros

Margarida G. M. S. Cardoso
Departamento de Métodos Quantitativos. Escola de Gestão-ISCTE
margarida.cardoso@iscte.pt

Resumo

O algoritmo CART-*Classification and Regression Trees* é aplicado na previsão das classificações de matemática associadas a uma amostra de alunos do ensino secundário. São modeladas, separadamente, as observações respeitantes a alunos do ensino secundário público e privado, considerando factores sócio-demográficos, factores específicos e factores pessoais. Obtém-se uma boa capacidade preditiva para os modelos propostos: 83,5% e 90,5%, estimativas da proporção de variância explicada, obtidas mediante validação cruzada, para os modelos do ensino público e privado, respectivamente. É ainda avaliada a importância relativa das variáveis preditivas nos modelos sublinhando-se, como principal, a média obtida pelos alunos às restantes disciplinas do secundário.

Palavras chave: árvores de regressão, algoritmo CART, previsão, ensino da matemática

Title: CART algorithm: Forecasting performance in high school mathematics

Abstract

In the present study we use the CART-Classification and Regression Trees algorithm to predict math grades based on a sample of high school students. Students from Public and Private schools are considered separately. Predictors include socio-demographics, personal attributes and some specific characteristics related to school. The models obtained have a good predictive capacity: proportion of grades' explained variance is 83,5% and 90,5% for regression trees referred to Public and Private schools, respectively. The relative importance of predictors is evaluated, the most important being the student's average grade referred to the remaining subjects (excluding mathematics).

Keywords: regression trees, CART algorithm, prediction, teaching of mathematics

1- Introdução

A importância do ensino da matemática no desenvolvimento dos indivíduos é amplamente aceite pela nossa comunidade. No entanto, o processo de ensino da Matemática está associado a algum insucesso, que ocorre ao nível de escola, sendo esta a disciplina que, por vezes, apresenta percentagens elevadas de classificações negativas, facto que, aquando do anúncio dos resultados dos exames nacionais, é bastante focalizado pelos órgãos de comunicação social.

O insucesso inerente à disciplina de Matemática e os resultados não satisfatórios dos alunos portugueses nesta disciplina, quando comparados com os de alunos da OCDE, são conhecidos por todos nós [Ramalho,G., 2001].

Diversos estudos têm procurado, ao longo do tempo, compreender, interpretar e justificar os desempenhos dos alunos portugueses assim como a relação dos alunos com a matemática e a sua aprendizagem ([Tavares, L. V., Graça, P.M. e Tavares, M. M.V., 2002], [Ponte, 2002], ou [Ramos,M., 2003], por exemplo).

No âmbito de estudos de comparação internacional de desempenhos na matemática há sobretudo a destacar o estudo internacional PISA 2000 que procurou estudar a literacia matemática de alunos de 15 anos de diferentes países. Entenda-se que a literacia matemática foi definida como a capacidade de identificar, de compreender e de se envolver em matemática e de realizar julgamentos bem fundamentados acerca do papel que a matemática desempenha na vida privada de cada indivíduo, na sua vida ocupacional e social, com colegas e familiares e na sua vida como cidadão construtivo, preocupado e reflexivo [OECD, 2002].

Comparando o desempenho em literacia matemática do conjunto dos países participantes, verifica-se que os resultados médios dos alunos portugueses estão abaixo dos resultados médios obtidos no espaço da OCDE. No entanto, verifica-se que não existem diferenças significativas entre o desempenho médio dos estudantes portugueses e o dos seus pares de Itália, da Letónia, da Polónia, da Grécia e do Luxemburgo. [Ramalho, G. (Coord.), 2001].

Quando se comparam os resultados obtidos pelos alunos portugueses nas várias regiões do país (NUT II), verifica-se que estes são bastante heterogéneos, sendo a zona de Lisboa e Vale do Tejo a que apresenta melhores resultados, embora o valor médio desta região continue a ser inferior ao da média da OCDE.

Comparando o desempenho em literacia matemática dos alunos portugueses por género sexual, constata-se que o seu desempenho não é muito diferente, sendo os rapazes que apresentam, em média, desempenhos significativamente melhores que os das raparigas.

Quando se entra em linha de conta com o ano lectivo frequentado pelos alunos portugueses de 15 anos, verifica-se que os que frequentam o 10º e 11º anos, ainda que estes últimos, em número muito reduzido, apresentam médias de desempenhos um pouco acima da média da OCDE, verificando-se uma grande diferença entre estes e os alunos que frequentam o 9º ano.

2- Objectivos

Neste trabalho pretende-se analisar o desempenho na Matemática de 10º e 11º anos procurando compreender os factores que a ele se associam e que poderão viabilizar uma previsão do mesmo desempenho. Considera-se, assim, como variável alvo do nosso estudo e portanto como variável dependente, a média da disciplina de matemática no 10º e 11º anos.

Da análise da diversa literatura definiram-se potenciais factores explicativos dos desempenhos dos alunos na disciplina de matemática: sócio-demográficos, específicos/escolares e pessoais (v. Figura 1). Assim, na análise que se propõe, consideram-se como variáveis independentes atributos diversos que se podem associar aos referidos factores.

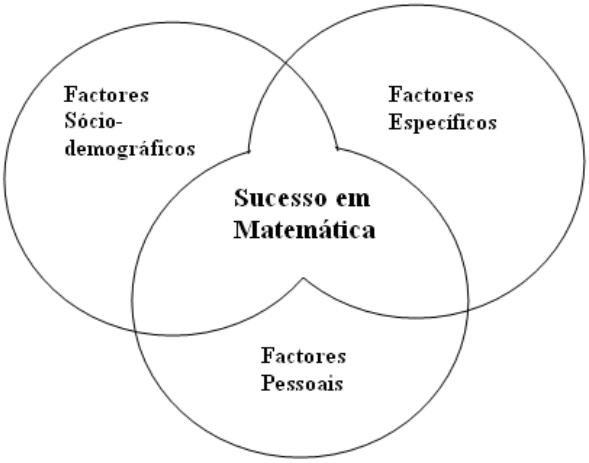

Figura 1: Variáveis explicativas do desempenho na disciplinas de Matemática

3- Metodologia

3.1- Recolha de dados

No intuito de explicar o desempenho na disciplina de matemática nos 10º e 11º anos procurou-se obter informação, junto de uma amostra de alunos do 12º ano acerca desta variável, assim como sobre potenciais factores explicativos: sócio-demográficos, específicos/escolares e pessoais.

A amostra utilizada é constituída por alunos que frequentavam o 12º ano no Curso Geral – Agrupamento 3 - Sócio Económico, durante o ano lectivo de 2003/2004, em escolas pertencentes à Direcção Regional de Lisboa e Vale do Tejo. A amostra recolhida foi de 271 alunos, estando 163 a frequentar o ensino público e 108 o ensino privado.

Procurou-se que as amostras recolhidas ilustrassem a diversidade do universo heterogéneo em estudo. Assim, embora as amostras obtidas não sejam aleatórias procurou-se cobrir a diversidade de escolas e da sua localização geográfica e, ainda, abranger escolas do ensino público e privado.

Como variável alvo deste estudo – variável dependente – considera-se a Média da disciplina de matemática no 10º e 11º anos (*Medmat*)

Como potenciais factores explicativos do desempenho na Matemática são consideradas variáveis relacionadas com o contexto social, económico e cultural, onde os alunos se inserem e são sujeitos activos, assim como variáveis relacionadas com as factores escolares e factores pessoais que possam influenciar o processo de ensino/ aprendizagem dos alunos.

Sendo assim, foram recolhidos dados sobre as variáveis que podem ser observadas nas Tabelas 1, 2 e 3.

Observe-se que a variável curso superior que o aluno pretende frequentar (*Curso*) se refere aos cursos superiores relacionados com o agrupamento em causa - cursos superiores de gestão, economia, marketing – tendo-se considerado uma alternativa que abrange os restantes cursos (*outros*).

Tabela 1: Variáveis explicativas: factores sócio-demográficos.

FACTORES SÓCIO- DEMOGRÁFICOS	
Variáveis	**Categorias**
Idade- Idade	
Sexo- Sexo	1- Masculino
	2- Feminino
Instpai- Grau de instrução do pai	1- Sem estudos
	2- Primária completa
	3- Ciclo preparatório
	4- 9º ano
	5- 12º ano
	6- Frequência curso médio/ superior
	7- Curso médio/ Bacharelato
	8- Curso superior
	9- Mestrado/ Doutoramento
Instmãe- Grau de instrução da mãe	1- Sem estudos
	2- Primária completa
	3- Ciclo preparatório
	4- 9º ano
	5- 12º ano
	6- Frequência curso médio/ superior
	7- Curso médio/ Bacharelato
	8- Curso superior
	9- Mestrado/ Doutoramento
Labpai- Situação laboral do pai	1- Trabalha por conta própria
	2- Trabalha por conta de outrém
	3- Temporariamente desempregado
	4- Reformado/pensionista
Labmãe- Situação laboral da mãe	1- Trabalha por conta própria
	2- Trabalha por conta de outrém
	3- Temporariamente desempregada
	4- Reformada/pensionista
	5- Dona de casa
Profpai- Profissão do pai	
Profmãe- Profissão da mãe	
Agregado- Número de pessoas do agregado familiar	
Rendimen- Rendimento mensal líquido do agregado familiar	1- até 500€
	2- 500-1000€
	3- 1000-1500€
	4- 1500-2000€
	5- 2000-3000€
	6- 3000-5000€
	7- mais de 5000€

Tabela 2:Variáveis explicativas: factores específicos

FACTORES ESPECÍFICOS	
Variáveis	**Categorias**
Escola- Escola que o aluno frequenta	
Ensino- Tipo de ensino- público ou privado	1- Privado
	2- Público
	3- Ambos
Recinf- Uso de recursos informáticos na sala de aula de matemática	1- Nunca
	2- Às vezes
	3- Frequentemente
Universi- Se pretende frequentar um curso superior	1- Sim
	2- Não
Curso- Curso superior que pretende frequentar	1- Gestão
	2- Economia
	3- Marketing
	4- Outro

Tabela 3: Variáveis explicativas: factores pessoais

FACTORES PESSOAIS	
Variáveis	**Categorias**
Reprov- Se o aluno já reprovou algum ano	1- Sim
	2- Não
Nreprov- Número de reprovações	
Meddis- Média das restantes disciplinas no 10º e 11º anos.	

3.2- Algoritmo CART

Neste trabalho propõe-se o uso do algoritmo CART - Classification and Regression Trees [Breiman, Friedman, Olshen e Stone, 1984] como metodologia de regressão não paramétrica para a previsão do desempenho em Matemática (medido mediante ***Medmat***).

As árvores de Regressão CART são essencialmente usadas para explicar e prever um determinado atributo – *Medmat*, neste caso - a partir de valores observados de atributos explicativos do mesmo (variáveis presentes em Tabela 1, Tabela 2 e Tabela 3). Este método permite ainda construir grupos homogéneos de indivíduos que são caracterizados pelos mesmos valores dos atributos (nós folha da árvore).

As árvores CART possuem a particularidade de serem árvores binárias, cuja leitura e interpretação é de fácil trato.

Este método é bastante utilizado em estudos multidimensionais, tendo a vantagem de ser bem sucedido em situações em que as variáveis explicativas são uma mistura de variáveis nominais, ordinais e contínuas. Para além desta, o modelo apresenta outras vantagens na sua aplicação, tais como: adaptar-se facilmente a dados omissos; ser invariante a transformações das variáveis, como a logaritmização das variáveis independentes, entre outras; não necessitar de satisfazer condições de aplicabilidade do modelo, como acontece nos modelos paramétricos. E, segundo os seus autores, os resultados do CART são bastante satisfatórios, nomeadamente em problemas não lineares.

A metodologia de regressão CART é desenvolvida em três etapas: i) o crescimento da árvore procedendo a diversas ramificações binárias no sentido de diminuir a diversidade da variável em estudo; ii) validação da árvore; iii) a interpretação da árvore de regressão proposta, na qual o papel da medida de importância relativa das variáveis preditivas proposta por Breiman et al [1984] deverá ser tido em conta.

3.2- Crescimento da árvore

A árvore de regressão CART, é obtida a partir de sucessivas divisões binárias do conjunto de dados - amostra de treino- através de uma medida de homogeneidade, que é usada para decidir qual a melhor variável de corte e valor de corte associados a cada nó.

Deste modo o processo de construção da árvore parte do geral para o particular.

Cada nó é dividido em dois nós descendentes, de modo que a heterogeneidade ou diversidade dos valores da variável dependente nestes nós seja mais reduzida do que no nó ascendente. Em cada divisão, para definir a melhor variável de corte é avaliada a redução da variância respeitante à variável alvo.

Todo este processo é recursivo, dado que cada novo nó obtido será considerado como um nó pai, ao qual será aplicado um novo critério de ramificação. Cada nova ramificação obtida origina uma árvore com menor variabilidade do que a árvore que a antecedia.

O crescimento da árvore pode, por vezes, ajustar-se demasiado bem aos valores da amostra de treino, o que pode causar algumas dificuldades na generalização do modelo obtido. Deste modo, é comum definirem-se regras de paragem de crescimento da árvore, que poderão, também, conduzir a uma mais fácil interpretação da mesma.

Algumas regras de paragem do crescimento de uma árvore são a consideração de um número máximo de níveis, a definição dos números mínimos de observações para nós a ramificar ou para nós descendentes e a imposição de um decréscimo mínimo da diversidade.

Uma vez terminada a construção da árvore, a previsão associada a um elemento que foi encaminhado para determinado nó folha será dada pela média no nó-folha onde esse elemento se enquadra (uma previsão que é igual para todos os elementos que pertençam ao mesmo nó-folha).

Finalmente, um indicador da precisão do modelo global será a soma ponderada das variâncias intra-nós-folha na amostra de treino. Contudo, a consideração desta estimativa de erro (erro de resubstituição) não considera a possibilidade de haver um sobreajustamento do modelo proposto.

3.4- Validação do modelo

A estimativa da qualidade do modelo proposto é obtida como resultado da aplicação do modelo a novos dados, ou, no caso de dados em número insuficiente, mediante validação cruzada.

O procedimento mais habitual na estimação do erro associado ao modelo envolve a criação de uma amostra de treino que é extraída aleatoriamente da amostra original e sobre a qual se vai desenvolver o modelo e a criação de uma amostra de teste que consiste num conjunto de restantes observações da amostra original, tipicamente de dimensão inferior, sobre a qual o modelo é testado.

Tendo em conta que o número de observações disponíveis nem sempre é suficiente para a constituição de amostras de treino e teste, o uso de uma validação cruzada surge como uma alternativa de validação do modelo quando a dimensão da amostra é reduzida.

A validação cruzada permite-nos calcular um erro mais realista para a árvore apresentada. O processo de cálculo do erro, usando a validação cruzada, é o seguinte:

- A amostra inicial é subdividida em V subamostras de dimensões aproximadamente iguais e nas quais, desejavelmente, as variáveis explicativas consideradas, têm uma distribuição semelhante.

- São construídas V árvores diferentes, utilizando para tal, (V-1)/V das observações, sendo as restantes 1/V observações utilizadas para avaliar o erro.

- A partir da construção das V árvores consegue-se calcular o erro associado a todas as observações

A consideração de V=10 no processo de validação cruzada é o método mais utilizado.

A proporção de variação explicada resultante de validação cruzada permite-nos obter uma estimativa da capacidade preditiva do modelo proposto:

$$1-\sum_{v=1}^{V}\frac{n_v}{n}\left(\frac{\sum_{i=1}^{n_v}(\hat{y}_i-y_i)^2}{\sum_{i=1}^{n_v}(y_i-\bar{y}_v)^2}\right)$$

No caso da presente aplicação, o facto de se dispor de uma amostra de dimensão reduzida impõe a adopção do processo de validação cruzada para obtenção de uma estimativa adequada do erro de previsão.

3.5- Medida de importância relativa das variáveis explicativas

Uma vez validado o modelo em árvore torna-se oportuna a sua interpretação. Breiman et al. [1984] propõem, como apoio à interpretação da árvore de decisão CART, uma medida *M* de importância das variáveis explicativas X_j usadas na construção da árvore.

Tendo em conta que as variáveis explicativas podem aparecer na previsão mascaradas por outras, isto é, não aparecer como responsáveis por ramificações mas proporcionar boas substitutas nessa tarefa, a sua importância deve ser medida atendendo a uma potencial contribuição para a previsão.

A medida *M* baseia-se na redução da diversidade desencadeada pelo uso da variável X_j em cada ramificação ou pelo seu potencial uso traduzido no conceito de ramificação substituta.

Assim, a medida *M* associada à variável X_j é dada pelo somatório das diversas reduções de diversidade associadas a essa variável em cada uma das ramificações, ou em possíveis ramificações substitutas.

Seja:

$$p(O)=\frac{n^{o}\text{ de observações no nó O}}{\text{nº total de observações}}$$

$$p(Oc)=\frac{n^{o}\text{ de observações no nó Oc descendente de O}}{\text{nº total de observações}}$$

$$c=1,2$$

π^j a partição por X_j que ramifica em O_1 e O_2 ($O_1 \cup O_2 = O$ e $O_1 \cap O_2 = \{\}$)

A medida *M* é dada por

$$M(X_j) = \sum_{O \in A} z^{jO} \Delta S^2(\pi^j; O)$$

onde

$z^{jO} = \begin{cases} 1 & \text{se } X_j \text{ ramifica o nó O ou } X_j \text{ é considerado ramificação substituta do nó O} \\ 0 & \text{caso contrário} \end{cases}$

e

$\Delta S^2(\pi^j; O)$ é o decréscimo da variância que resulta da ramificação do nó *O* pela variável X_j que proporciona a partição de O em O_1 e O_2 i.e.

$$\Delta S^2(\pi^j; O) = p(O) \times S^2(O) - \sum_{c=1}^{2} p(Oc) \times S^2(Oc)$$

4- Apresentação de Resultados

4.1- Árvores de regressão

No sentido de conhecer melhor a estrutura dos dados que se refere à amostra recolhida de alunos do secundário e de procurar determinar uma previsão para a variável alvo (medida de desempenho da matemática quantificada pela média de matemática nos 10º e 11º anos), procedeu-se à construção de vários modelos experimentais, baseados na metodologia CART.

Utilizando os diferentes grupos de factores explicativos foram ensaiadas diversas árvores, no sentido de perceber o contributo que cada grupo de factores, individualmente, poderia ter na explicação e previsão dos desempenhos dos alunos na disciplina de matemática de 10º e 11º anos. No entanto, utilizados separadamente, nenhum destes grupos de factores proporcionou a construção de um modelo com boa capacidades preditiva. Na árvore de previsão efectuada apenas com factores explicativos sócio-demográficos, por exemplo, a capacidade preditiva obtida, medida através da validação cruzada (procedimento *10-fold*), foi de 37%.

As árvores de regressão que se apresentam de seguida apresentam boa capacidade preditiva e referem-se ao conjunto de todos os factores explicativos. Na sua construção foram consideradas algumas regras de paragem específicas. Os parâmetros de aprendizagem que lhes correspondem apresentam-se na Tabela 4.

Tabela 4 – Parametrização das árvores de regressão

Regra de paragem	**Parametrização fixa**
Nº máximo de níveis da árvore	5 níveis
Nº mínimo de observações em nó *pai*	2 observações
Nº mínimo de observações em nó *filho*	1 observação
Decremento mínimo de variância numa ramificação	0,0001

4.2- Árvore associada a estudantes do ensino público

A árvore obtida sobre os dados referidos a estudantes do secundário do ensino público apresenta 46 nós, dos quais 26 são nós folha. A proporção de variância explicada pelo

modelo, medida através da validação cruzada é de, aproximadamente, 84%, como se pode observar na Tabela 5.

Tabela 5- Percentagem de variância explicada pelo modelo que inclui variáveis relacionadas com factores sócio-demográficos, específicos/escolares e pessoais - Ensino público.

	Amostra de treino	**Validação Cruzada (*10-fold*)**
Percentagem de variância explicada	87.4%	83.5%

Em resultado do cálculo da medida de importância relativa associada a cada variável explicativa (v.Tabela 6), verifica-se que a variável *média das restantes disciplinas nos 10º e 11º anos* é a que mais importância toma na previsão das classificações da disciplina de matemática para os mesmos anos. A importância desta variável revela-nos que alunos com melhores classificações nas restantes disciplinas também possuem melhores resultados na disciplina de matemática. Face a esta variável, as restantes variáveis tomam uma importância relativa muito pequena, havendo a destacar as variáveis *situação laboral do pai* e *grau de instrução do pai e da mãe*, como as variáveis que embora com uma contribuição muito mais pequena, também têm a sua importância na construção do modelo.

A situação laboral do pai, embora apareça como a segunda variável que mais contribui para as ramificações da árvore, apenas é determinante numa única ramificação, pelo que, por vezes, se encontra mascarada por outras variáveis explicativas.

Observe-se ainda, que as variáveis *sexo*, *curso superior* que os alunos pretendem tirar e *uso de recursos informáticos na sala de aula* apresentam, neste contexto e face a variáveis com grande capacidade explicativa, uma importância quase, ou mesmo nula.

No que diz respeito à última variável, *uso de recursos informáticos na sala de aula*, dado que este tipo de ferramenta didáctica ainda ser tão pouco utilizada na sala de aula, não se consegue perceber que impacto teria nas classificações da disciplina em estudo. Observe-se que apenas 19% dos alunos do ensino público dizem ter usado estes recursos na sala de aula.

Tabela 6-Importância relativa das variáveis relacionadas com factores sócio-demográficos, específicos/ escolares e pessoais, consideradas no modelo- Ensino público

Atributo Explicativo (X_i)	**M (X_i)**	**M normalizada**
Média das restantes disciplinas	5,507	100
Situação laboral do pai	0,8212	15
Grau de instrução do pai	0,6155	11
Grau de instrução da mãe	0,4804	9
Situação laboral da mãe	0,4305	8
Rendimento mensal líquido do agregado familiar	0,4266	8
Nº pessoas do agregado familiar	0,3974	7
Tipo de Ensino	0,1999	4
Idade	0,1992	4
Reprovações	0,1449	3
Número de reprovações	0,1368	2
Uso de recursos informáticos	0,048	1
Curso superior	0,0267	0
Pretende frequentar curso superior	0,0041	0
Sexo	0,0069	0

4.3- Resultados/avaliação da árvore efectuada para o ensino privado

A árvore obtida sobre os dados relativos ao grupo de estudantes do secundário em escolas do ensino privado apresenta 56 nós, dos quais 27 são nós folha. A proporção de variância explicada pelo modelo, medida através da validação cruzada, é de 90%, como se pode verificar na Tabela 7.

Tabela 7- Percentagem de variância explicada pelo modelo que inclui variáveis relacionadas com factores sócio-demográficos, específicos/escolares e pessoais - Ensino privado

	Amostra de treino	**Validação cruzada (*10-fold*)**
Percentagem de variância explicada	86.1%	90.5%

Em geral, verifica-se que os alunos que apresentam melhores *classificações nas restantes disciplinas* também apresentam melhores resultados na disciplina de matemática. Quanto mais elevado é *o grau de instrução da mãe*, melhores são os resultados obtidos na disciplina em estudo, sendo esta variável determinante de algumas das ramificações da árvore. O maior *uso de recursos informáticos na sala de aula* também está associado a melhores classificações na disciplina de matemática.

Verifica-se também que existe um grupo de alunos de dimensão razoável que apresenta uma previsão negativa para a média da disciplina de matemática nos 10º e 11º anos, grupo esse cuja *situação laboral do pai* é trabalhador por conta própria, e que, a um maior *rendimento líquido auferido pelo agregado familiar* não estão associados melhores desempenhos na disciplina.

Tabela 8: Importância relativa das variáveis relacionadas com factores sócio-demográficos, específicos/ escolares e pessoais, consideradas no modelo- Ensino privado

Atributo Explicativo (X_j)	**M (X_j)**	**M normalizada**
Média das restantes disciplinas	2,7455	100
Grau de instrução da mãe	1,0608	39
Idade	0,5721	21
Uso de recursos informáticos	0,5679	21
Situação laboral do pai	0,5263	19
Nº pessoas do agregado familiar	0,5293	19
Curso Superior	0,5126	19
Rendimento mensal líquido do agregado familiar	0,4889	18
Grau de instrução do pai	0,372	14
Situação laboral da mãe	0,3586	13
Tipo de ensino	0,1853	7
Sexo	0,1043	4
Reprovações	0,0875	3

No que respeita à importância relativa das variáveis explicativas, que pode ser observada na

Tabela 8, também neste modelo referente a alunos que frequentam o ensino privado, se verifica que a variável *média das restantes disciplinas no 10º e 11º anos*, é determinante para a previsão das classificações médias da disciplina de matemática para os mesmos anos. No modelo obtido, *o grau de instrução da mãe* continua a tomar uma importância razoável na determinação da previsão, cerca de 39 (na escala de 0 a 100). A *idade* e o *uso de recursos informáticos* são as variáveis explicativas que se seguem quanto à importância relativa que têm para o modelo. Observe-se que o uso de recursos informáticos é um pouco superior nas escolas privadas (24% dos alunos que frequentam o ensino privado dizem ter usado estes recursos nas aulas de matemática), pelo que esta variável já assume alguma importância na previsão das classificações de matemática.

5- Conclusões Gerais

O estudo desenvolvido procurou, de algum modo, contribuir para a investigação que tem ocorrido em torno da disciplina de matemática, dando especial ênfase ao desempenho dos alunos no ensino secundário, nomeadamente nos 10º e 11º anos.

Tabela 9 – Resultados comparativos das regressões CART para os estudantes de Ensino Público e Ensino Privado

Ensino Público	**Ensino Privado**
Percentagem de variância explicada(*10-fold*): **83,5%**	Percentagem de variância explicada(*10-fold*): **90,5%**
1- No modelo construído apenas com variáveis explicativas sócio demográficas, a variável que assume maior importância relativa é o - *Grau de instrução da mãe*. **2-** No modelo global, as variáveis que assumem maior importância relativa (M) são, numa escala de 0 a 100, respectivamente: - Média das restantes disciplinas (M=100) - Situação laboral do pai (M=15) - Grau de instrução do pai (M=11) - Grau de instrução da mãe (M=9) **2.1-** Principais conclusões Em geral, melhores classificações na disciplina de matemática estão associadas a melhores *médias das restantes disciplinas*, salvo alguns grupos de alunos que são excepção. A *situação laboral do pai*, embora seja a segunda variável com mais importância para a construção do modelo, aparece mascarada por outras variáveis, sendo apenas determinante na ramificação de um nó. O *rendimento mensal líquido* (M=8) auferido pelo agregado familiar não é determinante de melhores classificações na disciplina de matemática, assim como o *sexo* (M=0).	**1-** No modelo construído apenas com variáveis explicativas sócio demográficas, a variável que assume maior importância relativa é o - *Grau de instrução da mãe***.** **2-** No modelo global, as variáveis que assumem maior importância relativa (M) são, numa escala de 0 a 100, respectivamente: - Média das restantes disciplinas (M=100) - Grau de instrução da mãe (M=39) - Idade (M=21) - Uso de recursos informáticos (M=21) **2.1-** Principais conclusões: Em geral, melhores classificações na disciplina de matemática estão associadas a melhores *médias das restantes disciplinas*, a um maior *grau de instrução por parte da mãe* e ao *uso frequente de recursos informáticos na sala de aula*. O *rendimento mensal líquido*(M=18) auferido pelo agregado familiar não é muito determinante de melhores classificações na disciplina de matemática.

Neste trabalho pretendeu-se propor uma metodologia para prever as classificações da disciplina de matemática nos 10º e 11º anos, a partir de variáveis explicativas relacionadas com factores sócio-demográficos, específicos/escolares e pessoais, associados aos alunos. A metodologia proposta - Árvore de Regressão CART- tem em conta que aos múltiplos factores explicativos estão associados diferentes tipos de medição e que a relação entre estes factores e a variável alvo é complexa (não necessariamente linear).

As principais conclusões a retirar, estão expressas na

Tabela 9.

Verifica-se então que os alunos que apresentam melhores classificações médias na disciplina de matemática:

- apresentam, de um modo geral, melhores classificações médias nas restantes disciplinas

- são alunos cujas mães possuem um maior nível de habilitações literárias;

- são, em geral, mais novos;

- nas aulas, recorrem com maior frequência ao uso de recurso informáticos.

Note-se, a este propósito, que é destacada na previsão, a variável *médias nas restantes disciplinas*, o que de certo modo seria de esperar, não só atendendo aos aspectos substantivos do problema, mas também atendendo à natureza métrica desta medição. De facto, a maioria das restantes variáveis predictivas são de natureza qualitativa, o que de certo modo as pode colocar em desvantagem na corrida de "dividir para conquistar" uma boa previsão CART.

Estas conclusões podem ser observadas esquematicamente na Figura 2.

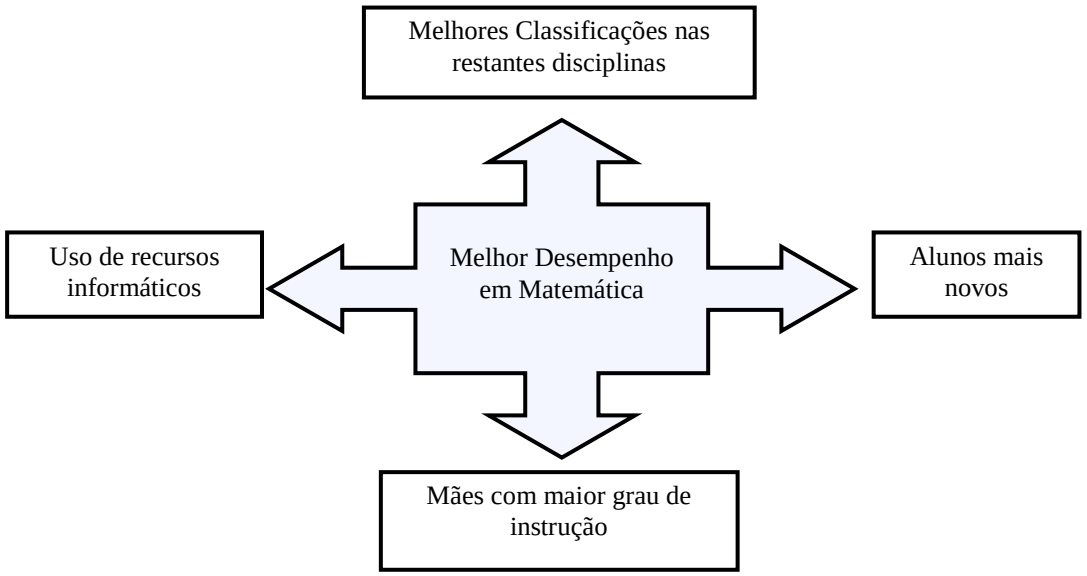

Figura 2: Variáveis relevantes na previsão das classificações de matemática do secundário

Em suma, a proposta de utilização das árvores de regressão CART- *Classification and Regression Trees* neste domínio do conhecimento, veio possibilitar a construção de modelos

explicativos do desempenho da matemática no secundário. Estes modelos revelam a importância (relativa) de alguns atributos específicos no desempenho da matemática.

Finalmente, embora os resultados apresentados se refiram apenas aos dados recolhidos, os modelos propostos têm associada uma boa capacidade preditiva estimada mediante validação cruzada: a proporção de variação das classificações de matemática explicada pelas árvores de regressão CART (aprendidas quer sobre observações de alunos do ensino público, quer do privado) encontra-se entre os 80% e os 90%.

Bibliografia

Breiman, L., Friedman, J. H., Olshen, R.A. e stone, C. J. (1984). Classification and Regression trees. Belmont, California: Wadsworth.

Cardoso, M. M., textos não publicados sobre CART.

Ponte, J.P. (2002). O ensino da Matemática em Portugal: Uma prioridade Educativa? Conferência realizada no Seminário sobre "O ensino da Matemática: Situação e Perspectivas", promovido pelo Conselho Nacional de Educação. Lisboa 2002.

Ramalho, G. (2001). Resultados do estudo internacional PISA 2000: Primeiro relatório nacional. Lisboa: Ministério da Educação, Gabinete de Avaliação Educacional (GAVE).

Ramalho, G. (2002). PISA 2000: Conceitos fundamentais em jogo na avaliação da literacia matemática e competências dos alunos portugueses. Lisboa: Ministério da Educação, Gabinete de Avaliação Educacional (GAVE).

Ramos, M.M.C. (2003) Matemática: A Bela ou o Monstro?, Tese de Doutoramento. Faculdade de Ciências da Universidade de Lisboa, Lisboa.

Tavares, L. V., Graça, P.M. e Tavares, M. M.V. (2002) Estudo SEDES: Assimetrias regionais do Desempenho Educativo, Observatório de prospectiva da Engenharia e da Tecnologia e Centro de Sistemas Urbanos e Regionais do I.S.T. Lisboa: I.S.T.

Anexos

Árvore para o Ensino Público

<table>
<tr><th>Nível 0</th><th>Nível 1</th><th>Nível 2</th><th>Nível 3</th><th>Nível 4</th><th>Nível 5</th></tr>
<tr><td rowspan="14">Nó 0 Média de matemática de 10º e 11º anos
Média 12.7 valores
[163]</td><td rowspan="10">Nó 1 Média das restantes disciplinas nos 10º e 11º anos ≤ 14.7 valores
Média 11.5 valores
[108]</td><td rowspan="5">Nó 3 Média das restantes disciplinas nos 10º e 11º anos ≤ 12.7 valores
Média 10.7 valores
[33]</td><td rowspan="3">Nó 7 Média das restantes disciplinas nos 10º e 11º anos ≤ 12.4 valores
Média 11 valores
[29]</td><td>Nó 15 Média das restantes disciplinas nos 10º e 11º anos ≤ 11.6 valores Média 10 valores
[9]</td><td>Nó 29 Média das restantes disciplinas nos 10º e 11º anos ≤ 11.20valores
Média 10.5 valores [6]</td></tr>
<tr><td>Nó 30 Média das restantes disciplinas nos 10º e 11º anos >11.20 valores
Média 9 valores [3]</td></tr>
<tr><td rowspan="2">Nó 16 Média das restantes disciplinas nos 10º e 11º anos >11.6 valores Média 11.5 valores
[20]</td><td>Nó 31 Média das restantes disciplinas nos 10º e 11º anos ≤ 12.05 valores
Média 12.4 valores [10]</td></tr>
<tr><td>Nó 32 Média das restantes disciplinas nos 10º e 11º anos >12.05 valores
Média 10.6 valores [10]</td></tr>
<tr><td colspan="2">Nó 8 Média das restantes disciplinas nos 10º e 11º anos >12.4 valores
Média 8 valores
[4]</td></tr>
<tr><td rowspan="7">Nó 4 Média das restantes disciplinas nos 10º e 11º anos >12.7 valores
Média 11.9 valores
[75]</td><td rowspan="4">Nó 9 Média das restantes disciplinas nos 10º e 11º anos ≤ 14.35 valores
Média 12.1valores
[64]</td><td rowspan="2">Nó 17 Média das restantes disciplinas nos 10º e 11º anos ≤ 13.55 valores
Média 11.7 valores
[43]</td><td>Nó 33 Média das restantes disciplinas nos 10º e 11º anos ≤ 13.20 valores
Média 12.5valores
[26]</td></tr>
<tr><td>Nó 34 Média das restantes disciplinas nos 10º e 11º anos >13.20 valores
Média 10.4 valores
[17]</td></tr>
<tr><td rowspan="2">Nó 18 Média das restantes disciplinas nos 10º e 11º anos >13.55 valores
Média 13 valores
[21]</td><td>Nó 35 Rendimento mensal líquido ≤ 1000€
Média 11.7 valores
[3]</td></tr>
<tr><td>Nó 36 Rendimento mensal líquido >1000€
Média 13.2 valores
[18]</td></tr>
<tr><td rowspan="2">Nó 10 Média das restantes disciplinas nos 10º e 11º anos >14.35 valores
Média 10.4 valores
[11]</td><td colspan="2">Nó 19 Média das restantes disciplinas nos 10º e 11º anos ≤ 14.45valores Média 11 valores
[4]</td></tr>
<tr><td colspan="2">Nó 20 Média das restantes disciplinas nos 10º e 11º anos >14.45 valores Média 10 valores
[7]</td></tr>
<tr><td rowspan="4">Nó 2 Média das restantes disciplinas nos 10º e 11º anos >14.7 valores
Média 15.14 valores
[55]</td><td rowspan="4">Nó 5
Média das restantes disciplinas nos 10º e 11º anos ≤ 16.7valores
Média 14valores
[34]</td><td rowspan="2">Nó 11 Média das restantes disciplinas nos 10º e 11º anos ≤ 15.05 valores Média 15.6 valores
[8]</td><td colspan="2">Nó 21 Média das restantes disciplinas nos 10º e 11º anos ≤ 14.95 valores Média 16 valores
[5]</td></tr>
<tr><td colspan="2">Nó 22 Média das restantes disciplinas nos 10º e 11º anos >14.95 valores Média 15valores
[3]</td></tr>
<tr><td rowspan="3">Nó 12 Média das restantes disciplinas nos 10º e 11º anos >15.05 valores
Média 13.5 valores
[26]</td><td rowspan="2">Nó 23 Média das restantes disciplinas nos 10º e 11º anos ≤ 15.7 valores Média 12.3 valores
[12]</td><td>Nó 37 Média das restantes disciplinas nos 10º e 11º anos ≤ 15.35 valores Média 13.6 valores
[5]</td></tr>
<tr><td>Nó 38 Média das restantes disciplinas nos 10º e 11º anos >15.35 valores
Média 11.4 valores
[7]</td></tr>
<tr><td>Nó 24 Média das restantes disciplinas nos 10º e 11º anos >15.7 valores Média 14.6 valores</td><td>Nó 39 Média das restantes disciplinas nos 10º e 11º anos ≤ 16.20 valores
Média 15.9 valores
[7]</td></tr>
</table>

<table>
<tr><td rowspan="8"></td><td rowspan="8"></td><td rowspan="2"></td><td rowspan="2"></td><td>[14]</td><td>Nó 40 Média das restantes disciplinas nos 10º e 11º anos >16.20 valores
Média 13.3 valores
[7]</td></tr>
<tr><td rowspan="6">Nó 6 Média das restantes disciplinas nos 10º e 11º anos >16.7 valores
Média 16.95 valores
[21]</td><td rowspan="4">Nó 13 Grau de instrução da mãe ≤ Frequência de curso médio / Superior
Média 16.1 valores
[13]</td><td>Nó 25 Grau de instrução da Mãe ≤ 12º ano
Média 16.8 valores
[11]</td><td>Nó 41 Grau de instrução do pai ≤ Ciclo preparatório
Média 17.7 valores
[6]</td></tr>
<tr><td></td><td>Nó 27 Grau de instrução do pai > Ciclo Preparatório
Média15.8 valores
[5]</td></tr>
<tr><td>Nó 26 Grau de instrução da mãe >12º ano
Média 12 valores
[2]</td><td>Nó 43 Tipo de ensino – Público
Média 11 valores [1]</td></tr>
<tr><td></td><td>Nó 44 Tipo de ensino – Ambos
Média 14 valores [1]</td></tr>
<tr><td rowspan="3">Nó 14 Grau de instrução da mãe > Frequência de curso médio / Superior Média 18 valores
[8]</td><td colspan="2">Nó 27 Grau de instrução do pai ≤ Frequência de curso médio / Superior
Média 20 valores
[1]</td></tr>
<tr><td>Nó 28 Grau de instrução do pai > Frequência de curso médio / Superior
Média 18 valores
[7]</td><td>Nó 45 Situação laboral do pai –Trabalhador por conta própria
Média 19 valores [1]</td></tr>
<tr><td></td><td>Nó 46 Situação laboral do pai – Trabalhador por conta de outrém
Média 17.8 valores [6]</td></tr>
</table>

Árvore para o Ensino Privado

Nível 0	Nível 1	Nível 2	Nível 3	Nível 4	Nível 5
Nó 0 Média de matemática de 10º e 11º anos Média 12.08 valores [108]	**Nó 1** Média das restantes disciplinas nos 10º e 11º anos $\leq$ 15.05 valores Média 11.12 valores [75]	**Nó 3** Média das restantes disciplinas nos 10º e 11º anos $\leq$ 13.95 valores Média 10.5 valores [47]	**Nó 7** Situação laboral do pai – Trabalhador por conta própria Média 9.7 valores [23]	**Nó 15** Rendimento mensal líquido $\leq$ 5000€ Média 10.4 valores [13]	**Nó 31** Idade $\leq$ 17 anos Média 11.8 valores [5]
					Nó 32 Idade>17 anos Média 9.5 valores [8]
				Nó 16 Rendimento mensal líquido >5000€ Média 8.7 valores [10]	**Nó 33** Média das restantes disciplinas nos 10º e 11º anos $\leq$ 12.85 valores Média 9.8 valores [5]
					Nó 34 Média das restantes disciplinas nos 10º e 11º anos >12.85 valores Média 7.6 valores [5]
			Nó 8 Situação laboral do pai – Trabalhador por conta de outrém ou reformado/pensionista Média 11.3 valores [24]	**Nó 17** Grau de instrução da mãe $\leq$ 9º ano Média 9.8 valores [6]	**Nó 35** Grau de instrução do pai $\leq$ Primária completa Média 10 valores [3]
					Nó 36 Grau de instrução do pai > primária Completa Média 9.7valores [3]
				Nó 18 Grau de instrução da mãe >9º ano Média 11.8 valores [18]	**Nó 37** Grau de instrução da mãe $\leq$ Curso Superior Média 11.6 valores [14]
					Nó 38 Grau de Instrução da mãe > Curso Superior Média12.5 valores [4]
		Nó 4 Média das restantes disciplinas nos 10º e 11º anos >13.95 valores Média 12.17 valores [28]	**Nó 9** Situação laboral da mãe –Trabalhador por conta de outrém Média 11.8 valores [[22]	**Nó 19** Tipo de ensino – Ambos Média 11 valores [5]	**Nó 39** Grau de instrução da mãe $\leq$ Curso Superior Média 10.8 valores [4]
					Nó 40 Grau de instrução da mãe > Curso Superior Média 12 valores [1]
				Nó 20 Tipo de ensino – Privado Média 12.1 valores [17]	**Nó 41**Agregado familiar $\leq$ 4 Média 11.6 valores [14]
					Nó 42 Agregado familiar > 4 Média 14 valores [3]
			Nó 10 Situação laboral da mãe –Trabalhador por conta própria, dona de casa ou temporariamente desempregada Média 13.5 valores [6]	**Nó 21** Situação laboral da mãe –Trabalhador por conta própria, dona de casa Média 13.2valores [5]	**Nó 43** Já reprovou ?- Não Média 13 valores [4]
					Nó 44 Já reprovou ?- Sim Média 14 valores[1]
				Nó 22 Situação laboral da mãe -temporariamente desempregada Média 15 valores [1]	
	Nó 2 Média das restantes disciplinas nos 10º e 11º anos >15.05 valores Média 14.27 valores [33]	**Nó 5** Média das restantes disciplinas nos 10º e 11º anos $\leq$ 16,20valores Média 13.1 valores [15]	**Nó 11** Grau de instrução da mãe $\leq$ 12º ano Média 11.7 valores [6]	**Nó 23** Uso de recursos informáticos- Nunca; às vezes Média 11 valores [5]	**Nó 45** Grau de instrução da mãe $\leq$ 9º ano Média 10 valores [2]
					Nó 46 Grau de instrução da mãe >9º ano Média 11.7 valores [3]
				Nó 24 Uso de recursos informáticos - frequentemente Média 15 valores [1]	
			Nó 12 Grau de instrução da Mãe >12º ano Média 14 valores [9]	**Nó 25** Curso universitário que pretendem seguir - Economia , Média 11.5 valores [2]	**Nó 47** Idade $\leq$ 17 anos Média 11 valores [1]
					Nó 48 Idade >17 anos Média 12 valores [1]
				Nó 26 Curso universitário que pretendem seguir - Gestão Média 14.7 valores [7]	**Nó 49** Tipo de ensino – Ambos Média 13 valores [1]
					Nó 50 Tipo de ensino – Privado Média 15 valores [6]

<table>
<tr><td rowspan="7"></td><td rowspan="7"></td><td rowspan="7">Nó 6 Média das restantes disciplinas nos 10º e 11º anos >16,20 valores
Média 15.3 valores
[18]</td><td rowspan="3">Nó 13 Situação laboral do pai –Trabalhador por conta própria
Média 14.5 valores
[6]</td><td>Nó 27 Uso de recursos informáticos- Nunca
Média 14 valores
[5]</td><td>Nó 51 Situação laboral da mãe –trabalhadora por conta de outrém, dona de casa
Média 14.3 valores
[4]</td></tr>
<tr><td></td><td>Nó 52 Situação laboral da mãe –trabalhadora por conta própria
Média 13 valores
[1]</td></tr>
<tr><td colspan="2">Nó 28 Uso de recursos informáticos- às vezes
Média 17 valores
[1]</td></tr>
<tr><td rowspan="4">Nó 14 Situação laboral do pai –Trabalhador por conta de outrém
Média 15.7 valores
[12]</td><td>Nó 29 Agregado familiar ≤ 4
Média 16.1 valores
[9]</td><td>Nó 53 Curso Universitário que pretendem seguir - Economia
Média 15.8 valores
[6]</td></tr>
<tr><td>Nó 54 Curso universitário que pretendem seguir - Marketing ou Gestão
Média 16.7 valores
[3]</td></tr>
<tr><td rowspan="2">Nó 30 Agregado familiar > 4
Média 14.3 valores
[3]</td><td>Nó 55 Média das restantes disciplinas nos 10º e 11º anos ≤ 17.35 valores
Média 15 valores
[1]</td></tr>
<tr><td>Nó 56 Média das restantes disciplinas nos 10º e 11º anos >17.35 valores
Média 10.5 valores
[2]</td></tr>
</table>