risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Iberian Journal of Information Systems and Technologies

Recebido/Submission: 03/10/2019
Aceitação/Acceptance: 29/01/2020

Previsão do abandono académico numa instituição de ensino superior com recurso a *data mining*

Maria P. G. Martins[1,3], Vera L. Migueis[2], D. S. B. Fonseca[3], Paulo D. F. Gouveia[1]

prud@ipb.pt, **vera.migueis@fe.up.pt, davide@ubi.pt, pgouveia@ipb.pt**

[1] Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal

[2] Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal

[3] CISE – Electromechatronic Systems Research Centre, Universidade da Beira Interior, Calçada Fonte do Lameiro, 6201-001 Covilhã, Portugal

Pages: 188–203

Resumo: Este estudo propõe dois modelos preditivos de classificação que permitem identificar, logo no final do 1º e do 2º semestres escolares, os estudantes de licenciatura de uma instituição de ensino superior mais propensos ao abandono académico. A metodologia proposta, que combina 3 algoritmos populares de *data mining*, como são as *random forest*, as máquinas de vetores de suporte e as redes neuronais artificiais, para além de contribuir para a assertividade da previsão, permite identificar por ordem de relevância os principais fatores que prenunciam o abandono académico. Os resultados empíricos demonstram ser possível reduzir para cerca de 1/4 as 4 dezenas de potenciais preditores do abandono, e mostram serem essencialmente dois, do contexto curricular do estudante, a explicarem essa propensão. Esse conhecimento revela-se de importância primordial para que os agentes de gestão possam adotar as medidas e decisões estratégicas mais propícias à diminuição dos índices de evasão discente.

Palavras-chave: *Educational data mining*; previsão do abandono académico; *random forest*; máquinas de vetor de suporte; redes neuronais artificiais.

Prediction of academic dropout in a higher education institution using data mining

Abstract: This study proposes two predictive models of classification that allow to identify, at the end of the 1st and 2nd semesters, the undergraduate students of a higher education institution more prone to academic dropout. The proposed methodology, which combines 3 popular data mining algorithms, such as random forest, support vector machines and artificial neural networks, in addition to contributing to predictive performance, allows to identify the main factors behind academic dropout. The empirical results show that it is possible to reduce to about 1/4 the 4 tens potential predictors of dropout, and show that there are essentially two predictors, concerning student's curriculum context, that explain this propensity. This knowledge is useful for decision-makers to adopt the most appropriate strategic measures and decisions in order to reduce student dropout rates.

Keywords: Educational data mining; prediction academic dropout; random forest; support vector machines; artificial neural networks.

1. Introdução

Dada a relevância do sucesso educacional para o progresso científico, tecnológico, económico e cultural de qualquer país ou sociedade, as recentes políticas educativas, tanto de âmbito nacional como da União Europeia, têm exigido às Instituições do Ensino Superior (IES) intervenções acrescidas em prol da promoção do desempenho académico e da diminuição dos índices de evasão discente.

A criação de modelos de previsão de desempenho académico que permitam identificar, com elevado desempenho e de forma precoce, os estudantes propensos ao abandono, bem como os principais fatores que prenunciam o fenómeno, são ações que poderão ter uma importância crucial para delinear intervenções que se possam revelar eficazes no combate à evasão discente. Com esse intuito, com recurso a ferramentas de *data mining* (DM), no presente artigo são propostos dois modelos de classificação para previsão do abandono, um deles para ser aplicado logo no final do 1º semestre e o outro no final do 1º ano letivo.

A fim de garantir uma caraterização adequada, quer da heterogeneidade de perfis estudantis existentes no ensino superior, quer das múltiplas dimensões indissociáveis do aluno que poderão influir no seu desempenho de aprendizagem, no desenvolvimento dos modelos, são considerados conjuntos de dados reais, de grande dimensão e dimensionalidade, provenientes das bases de dados do Instituto Politécnico de Bragança (IPB), uma IES do interior do país, com 5 escolas, onde são ministradas mais de meia centena de licenciaturas, que cobrem as mais diversas áreas educacionais.

2. *Data Mining* Educacional

A análise e exploração de grandes conjuntos de dados, designada análise de *big data*, requer técnicas analíticas de grande alcance, como as de DM. Trata-se de técnicas que possibilitam a indução de modelos analíticos, de natureza preditiva ou descritiva, de fácil compreensão, consistentes, de elevada precisão e bastante realistas, que se têm revelado muito promissores para melhorar os processos de decisão nas organizações. Por esse motivo, nos últimos anos, tem havido um interesse crescente no uso de DM em estudos relacionados com a melhoria dos resultados educacionais, a qual é designada *educacional data mining* (EDM) (Baker, 2010).

Entre as 5 categorias de estudos existentes na literatura de EDM, identificadas na taxonomia recentemente proposta por (Bakhshinategh, Zaiane, ElAtia, & Ipperciel, 2018), destaca-se a modelação do estudante, onde se enquadra a previsão do desempenho académico. Por via deste tipo de estudos é possível estimar, de forma precoce, o valor futuro, ou desconhecido, de algumas variáveis de interesse, contínuas ou categóricas, que permitem inferir o desempenho académico a partir de outros fatores presentes em conjuntos de dados.

Como variáveis contínuas de interesse, poder-se-á mencionar, por exemplo, qualquer indicador que infira o sucesso educacional esperado para os estudantes no final do seu curso, como é o caso de uma métrica que inclua a média final de curso – e.g. (Miguéis, Freitas, Garcia, & Silva, 2018), (Natek & Zwilling, 2014), (Aluko, Adenuga, Kukoyi, Soyingbe, & Oyedeji, 2016), (Martins, Miguéis, Fonseca, & Alves, 2019) – ou

as classificações em determinadas unidades curriculares – e.g. (Huang, 2011), (Pascoal, Brito, & Rêgo, 2015), (Costa, Fonseca, Santana, Araújo, & Rego, 2017). Como variáveis de interesse do tipo categórico poder-se-ão mencionar, por exemplo, aquelas que informem se o estudante abandona ou não a sua formação académica (e.g. (Delen, 2010), (Nandeshwar, Menzies, & Nelson, 2011)), se obtém ou não aprovação no final do ano letivo (e.g. (Hoffait & Schyns, 2017)), ou então, se desiste, ou não, de unidades curriculares específicas (e.g. (Kotsiantis, Pierrakeas, & Pintelas, 2003), (Burgos, et al., 2017)).

Por via desta tipologia de estudos é possível identificar, com a devida antecedência, os estudantes que necessitem de apoio, tais como os de baixa motivação, ou com propensão ao insucesso ou abandono académico. Têm também a utilidade de possibilitar a identificação dos fatores que mais contribuem para explicar o (in)sucesso e o abandono académico. Por exemplo, no caso concreto dos estudos subordinados à previsão do abandono académico – o âmbito do presente artigo –, os autores do estudo (Nandeshwar, Menzies, & Nelson, 2011), após uma análise e exploração de um conjunto de dados com informação de cariz demográfico, académico e socioeconómico, processada com recurso aos algoritmos de classificação *One-R, C4.5, ADTrees, Naive Bayes* e redes de polarização radial, concluíram que o historial familiar (educacional e socioeconómico) e o historial de desempenho académico no ensino secundário têm uma relação direta com os estudantes que permaneceram na universidade *Kent State* até à conclusão da sua graduação. Num outro estudo, (Manhães, 2015), prescindindo de informação de cariz socioecónomico e demográfico, e cingindo-se apenas a informação académica que varia no tempo, é proposto um modelo que permite diferenciar, do 1º ao 5º semestres letivos (em cursos de 12 semestres), os alunos do curso de engenharia civil da Universidade do Rio de Janeiro em duas classes distintas: progresso ou não progresso. Os resultados experimentais evidenciaram, para todos os 12 algoritmos classificadores existentes na ferramenta *Weka*, uma capacidade de acerto preditivo superior a 75%.

3. Dados e Metodologia

3.1. Modelo de dados

Para a criação dos modelos que consigam prever o abandono académico dos alunos de licenciatura do IPB usa-se um conjunto vasto de informação curricular e extracurricular do aluno, registada no Sistema de Informação da instituição entre 2007/2008 e 2015/2016. Opta-se por restringir o estudo apenas aos cursos de licenciatura, na sua esmagadora maioria ciclos de 3 anos, por se tratar da principal oferta formativa da instituição e por abranger um conjunto de dados mais completo.

Tendo-se como objetivo o desenvolvimento de dois modelos preditivos distintos, um para ser aplicado no final do 1º semestre e o outro no final do 1º ano letivo, prepararam-se dois *datasets* distintos, designados *semestre1 e semestre12*, que integram, para além

da variável alvo a prever, 38 e 41 variáveis preditivas, respetivamente. Depois da limpeza de dados e de outras tarefas de pré-processamento, os 2 *datasets* em questão passaram a incorporar, respetivamente, 3373 e 3344 matrículas de licenciaturas iniciadas entre os anos letivos 2007/08 e 2013/14.

3.2. Metodologia

A variável alvo a prever é do tipo binário, assumindo o seu valor um dos seguintes significados: 'abandona' ou 'não abandona'. No atual contexto dever-se-á sempre entender a afirmação 'não abandona' como sinónimo de que o aluno 'conclui' o seu curso. Note-se que se classificou como 'abandono' todas as matrículas referentes a alunos que não tenham qualquer inscrição válida no ano 2016/2017 (último ano letivo com matrículas registadas na base de dados do IPB no período de desenvolvimento deste trabalho) e, cumulativamente, não tenham ainda concluído o seu curso, nem, tão pouco, mudado para outro curso nesse mesmo ano letivo. Todas as matrículas de desfecho indefinido foram naturalmente excluídas do *dataset*.

No que concerne aos potenciais fatores preditores do abandono académico, considerou-se essencialmente a mesma tipologia de fatores mencionados nos trabalhos relacionados de relevo, como os de cariz demográfico, socioeconómico e do contexto curricular dos estudantes. Estes fatores são passíveis de serem classificados em dois importantes subgrupos: os resultados curriculares acumulados e os dados 'intemporais' (i.e., todos aqueles cujos valores não se alteram ao longo do percurso escolar do aluno). Na Tabela 1 apresentam-se todos os fatores preditivos considerados neste estudo[1]. Cada um dos potenciais fatores preditivos do abandono académico é classificado (3a coluna da tabela) de acordo com a sua natureza, em cinco categorias distintas: curriculares (C), de matricula (M), demográficos (D), socioeconómicos (S) e de acesso (A). De salientar que é a categoria C que representa os resultados curriculares acumulados até ao final do 1º ou do 2º semestres, sendo todas as restantes categorias compostas unicamente por dados intemporais.

Na Figura 1 apresenta-se um esquema que ilustra os modelos de previsão a desenvolver. Nele encontram-se representadas as diferentes categorias de fatores preditivos do abandono, usadas como *input* dos algoritmos de DM, tal como identificadas na Tabela 1. Como se procura ilustrar, para o grupo de variáveis curriculares (C) são usados os resultados acumulados ao fim do 1º e 2º semestres escolares do aluno. Mais concretamente, desenvolvem-se dois modelos distintos, um suportado pelos dados curriculares recolhidos ao fim do 1º semestre e o outro suportado pelos dados curriculares recolhidos ao fim do 2º semestre, incorporando este segundo *dataset* os seus dados curriculares na forma agregada (acumulada).

[1] Para uma mais fácil identificação, o nome dos atributos curriculares (os que formam o 1º subgrupo importante de fatores) terminam com o sufixo "_s".

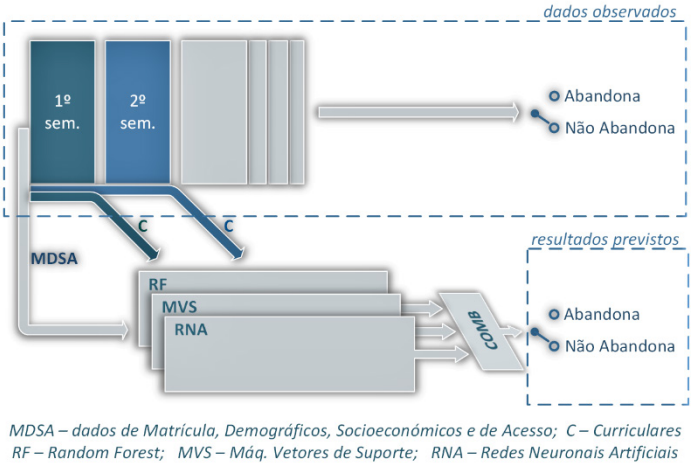

Figura 1 – Esquema ilustrativo do modelo de previsão

id	atributo	cat	tipo	min.. max	significado
1	*ano_curricular_s*	C	discreto	1..4	ano curricular do aluno no sem. escolar considerado
2	*bolseiro_s*	C	contínuo[1]	0..1	o aluno foi bolseiro no semestre escolar?
3	*cod_estatuto1_s*	C	nominal	1..5	tipo de estatuto do aluno no 1º sem. escolar
4	*cod_estatuto2_s*[2]	C	nominal	1..5	tipo de estatuto do aluno no 2º sem. escolar
5	*cod_freq_tipo1_s*	C	nominal	1..7	tipo de frequência do aluno no 1º sem. escolar
6	*cod_freq_tipo2_s*[3]	C	nominal	1..7	tipo de frequência do aluno no 2º sem. escolar
7	*dir_associativo_s*	C	contínuo[2]	0..1	o aluno foi dirigente associativo no semestre escolar?
8	*ects_aprov_s*	C	discreto	0..60	nº de ECTS aprovados no semestre escolar
9	*ects_reprov_s*	C	discreto	0..60	nº de ECTS reprovados no semestre escolar
10	*max_s*	C	discreto	0..20	nota máxima das UCs aprovadas no semestre escolar
11	*media_s*	C	contínuo	0..20	nota média das UCs aprovadas no semestre escolar
12	*min_s*	C	discreto	0..20	nota mínima das UCs aprovadas no semestre escolar
13	*navalr_s*	C	discreto	0..18	nº de avaliações sem aprovação no semestre escolar
14	*nuca_s*	C	discreto	0..10	nº de UCs aprovadas no semestre escolar
15	*nucr_s*	C	discreto	0..10	nº de UCs reprovadas no semestre escolar
16	*vd12_s*[3]	C	contínuo	-20..20	diferença de desempenho do 1º para o 2º semestre

id	atributo	cat	tipo	min.. max	significado
17	*cod_curso*	M	nominal	1..51	código do curso
18	*cod_escola*	M	nominal	1..5	código da escola
19	*ects_cred_tx*	M	discreto	0..100	fração de ECTS que foram creditados ao aluno
20	*ects_curso*	M	discreto	180..240	número de ECTS do curso
21	*deslocado*	D	binário	0..1	o aluno está deslocado da sua residência habitual?
22	*dist*	D	nominal	1..28	distrito de proveniência do aluno
23	*dist_n*	D	nominal	1..27	distrito de naturalidade
24	*idade*	D	discreto	17..61	idade no ato da matrícula
25	*nacionalidade*	D	nominal	1..15	nacionalidade do aluno
26	*sexo*	D	nominal	1..2	género
27	*cod_prof_aluno*	S	nominal	1..12	profissão do aluno
28	*cod_prof_mae*	S	nominal	1..12	profissão da mãe
29	*cod_prof_pai*	S	nominal	1..12	profissão do pai
30	*nivel_esc_mae*	S	ordinal	1..13	nível de escolaridade da mãe
31	*nivel_esc_pai*	S	ordinal	1..13	nível de escolaridade do pai
32	*sit_prof_aluno*	S	nominal	1..10	situação profissional do aluno
33	*sit_prof_mae*	S	nominal	1..10	situação profissional da mãe
34	*sit_prof_pai*	S	nominal	1..9	situação profissional do pai
35	*fase*	A	ordinal	1..3	fase de acesso
36	*media_acesso*	A	contínuo	0..200	nota de acesso ao ensino superior
37	*n10_11_acesso*	A	contínuo	0..200	média dos 10º e 11º anos
38	*n12_acesso*	A	contínuo	0..200	média do 12º ano
39	*opcao_acesso*	A	ordinal	1..6	ordem da opção na candidatura ao curso
40	*ordem_acesso*	A	ordinal	1..322	ordem de acesso entre os colocados no curso
41	*pi_acesso*	A	contínuo	0..200	nota média das provas de ingresso

Tabela 1 – Variáveis explicativas usadas na previsão do abandono

Com a sigla MDSA pretende-se representar o conjunto de todas as restantes categorias de variáveis, comuns aos dois modelos a desenvolver: variáveis de Matrícula, Demográficas, Socioeconómicas e de Acesso (ao ensino superior).

Cada um dos *datasets*, *semestre1 e semestre12,* é ainda subdividido em 3 subconjuntos, para serem usados nas fases de aprendizagem, validação e teste dos modelos de previsão desenvolvidos. Com o intuito de se modelar de uma forma mais fiel o processo de previsão real que se pretende implementar no futuro, opta-se por um particionamento temporal. Com esta opção procura-se garantir que a previsão de abandono relativamente a uma dada matrícula se faça sempre a partir de dados já observados no passado, que será isso que acontecerá em contexto real. Tendo em conta a distribuição das matrículas ao longo do período em análise (iniciadas entre 2007/08 e 2013/14), opta-se pelo particionamento caracterizado na Tabela 2.

dataset	total	Treino (2007/08 a 2010/11)	Validação (2011/12)	Teste (2012/13 e 2013/14)
semestre1	3373	58.7%	18.5%	22.8%
semestres12	3344	58.8%	18.6%	22.5%

Tabela 2 – Dimensão dos conjuntos usados para treino, validação e teste

É também objetivo do atual estudo avaliar se através da geração de um modelo de conjunto, que integre os 3 algoritmos preditivos, se consegue minimizar o erro de previsão do abandono. Para o efeito, explora-se a combinação de 3 dos mais populares algoritmos de DM em abordagens de classificação: as *random forests* (RF), as máquinas de vetores de suporte (MVS) e as redes neuronais artificiais (RNA). A opção por estes 3 algoritmos justifica-se em virtude da literatura de EDM demonstrar que apresentam desempenhos competitivos (capacidade de aprendizagem e generalização) face a outro tipo de classificadores.

No desenvolvimento de cada um dos dois modelos propostos, começa-se por treinar e afinar, de forma separada, os três algoritmos de DM suportados pelo conjunto completo de variáveis preditivas disponíveis (38 e 41 variáveis).

Seguidamente, com o intuito de se identificarem os fatores mais explicativos do abandono, tenta-se ajustar o conjunto de variáveis que suportam o modelo, selecionando apenas as que se revelem importantes para a previsão pretendida (*feature selection*). Para esse efeito, adota-se o método de seleção progressiva (*forward search*), em que as variáveis vão sendo selecionadas uma a uma, num processo iterativo, juntando-se sempre às já selecionadas aquela que das restantes conduza a um maior incremento no valor médio de desempenho dos três algoritmos de DM — é este critério de otimização, que tendo por base o desempenho agregado dos 3 modelos, permite combinar num único resultado o processo de seleção. O processo termina no momento que esse incremento seja nulo ou negativo. Desta forma, ainda que se usem três algoritmos, apenas se obtém como resultado um único subconjunto de variáveis explicativas para cada um dos 2 modelos.

Depois de encontradas as variáveis mais explicativas, averigua-se qual a importância relativa das mesmas, medindo-se o impacto de cada uma delas, na explicação da variável alvo, através de uma técnica de análise de sensibilidade que combina os três algoritmos de DM usados no estudo. Mais concretamente, a importância relativa de uma qualquer variável de entrada na explicação da variável resposta é encarada como sendo a razão entre a perda de acurácia que resulte da sua não inclusão e a acurácia do modelo com todas as suas variáveis de entrada. Dessa forma, quanto maior for a deterioração do modelo com a exclusão da variável específica, maior é o nível de importância dessa variável.

Tanto na fase de *feature selection*, como nas simulações relacionadas com o cálculo da importância relativa das variáveis explicativas, tem-se sempre o cuidado de reafinar cada um dos algoritmos de DM para cada um dos conjuntos diferentes de variáveis independentes que se considere. Os dados de validação, juntamente com os de treino, são usados para todo o processo de afinação dos modelos e de seleção das variáveis mais

explicativas. Já os dados de teste são deixados de parte, para que, através deles, se possa proceder, no fim, à avaliação da capacidade de generalização das soluções que vierem a ser encontradas. Nessa avaliação final, tal como se espera que venha a acontecer em contexto real, as classificações atribuídas pelos três algoritmos de DM são combinadas de forma a conseguir-se um único veredicto. Propõe-se, como forma de combinação, que seja sempre escolhido o valor mais votado entre as três classificações.

Para comparação de desempenho das diferentes configurações dos modelos de previsão considerados neste estudo usa-se, como métrica de avaliação, o valor AUC (*Area Under ROC Curve*). Uma medida que enfatiza a especificidade e a sensibilidade do modelo classificador, que no caso em estudo, se traduzirá, respetivamente, nas proporções de 'abandonos' e de 'não abandonos' classificados corretamente pelo modelo.

4. Implementação e Resultados

Para um mais fácil entendimento dos resultados apresentados nas próximas secções, mostra-se na Figura 2 um esquema que tenta ilustrar os modelos de previsão desenvolvidos e, em particular, os diferentes subconjuntos de preditores que lhes dão suporte.

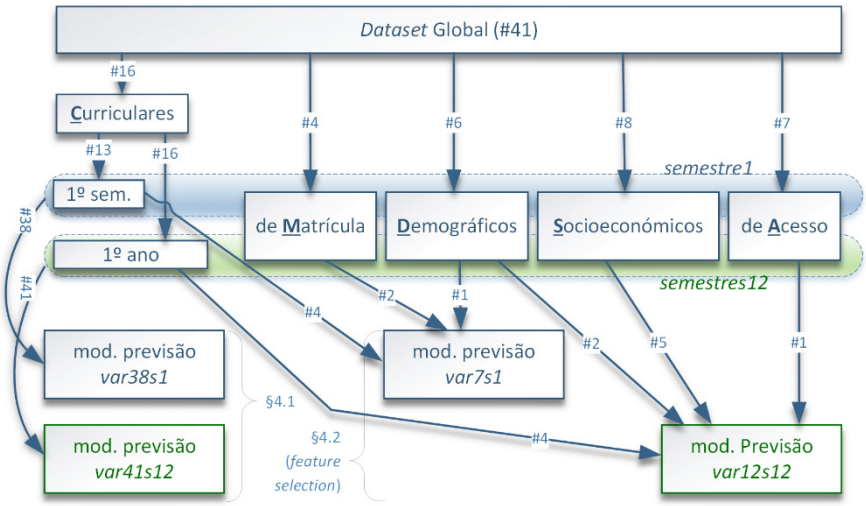

Figura 2 – Diagrama ilustrativo dos subconjuntos de preditores usados nos diferentes modelos de previsão

4.1. Treinamento em separado dos 3 algoritmos de DM, suportados por todas as variáveis independentes

De forma a se conseguir uma primeira perceção da capacidade preditiva dos modelos, começou-se por treinar separadamente cada um dos três algoritmos de DM, sempre suportados pelo conjunto completo das potenciais variáveis explicativas. Com o intuito de se maximizar a capacidade de previsão, houve o cuidado de afinar os três algoritmos de

classificação para ambos os *datasets semestre1* e *semestres12*, avaliando o desempenho desses algoritmos para diferentes valores de alguns dos seus hyperparâmetros mais importantes. Na Tabela 3 podem ser encontrados os melhores valores obtidos nesse processo de afinação[2].

Tendo em conta os valores ótimos encontrados, percebe-se, como esperado, que a capacidade preditiva do modelo aumenta quando são utilizados para o *dataset* os dados curriculares acumulados ao fim do 2º semestre escolar do aluno (0.8896, quando com dados do 1º semestre não vai além de 0.8630). A partir da mesma tabela, também é possível, desde logo, concluir que é o algoritmo redes neuronais artificiais que demonstra ter maior capacidade preditiva no contexto do problema que está a ser tratado, não se notando diferença significativa de desempenho nas outras duas técnicas de classificação.

		RF		MVS		RNA		
dataset	**AUC médio**	**mtry**	**AUC**	**cost**	**AUC**	**size**	**decay**	**AUC**
semestre1	0.8630	5	0.8548	2^{5}	0.8554	50	1	0.8787
semestres12	0.8896	3	0.8889	2^{3}	0.8792	1	$10^{-1/3}$	0.9007

Tabela 3 – Valores ótimos encontrados para os três modelos de classificação

4.2. Seleção dos principais fatores explicativos do abandono

Depois de concluída a fase de treinamento e afinação dos 3 algoritmos usados, procedeu-se, com os mesmos dados de treino e validação, a um ajustamento mais minucioso do conjunto de variáveis que suportam os modelos, selecionando, apenas, as que se revelam mais explicativas do abandono. Os resultados das simulações produzidas nos processos iterativos de seleção são reportados nas tabelas 4 e 5.

			RF		MVS		RNA		
ord.	**variável**	**AUC médio**	**mtry**	**AUC**	**cost**	**AUC**	**size**	**decay**	**AUC**
1ª	*ects_aprov_s*	0.7955	1	0.7359	2^{-1}	0.8198	5	$10^{-8/3}$	0.8308
2ª	*ects_cred_tx*	0.8236	1	0.7988	2^{9}	0.8243	20	$10^{-7/3}$	0.8476
3ª	*ects_reprov_s*	0.8399	1	0.8308	2^{3}	0.8236	10	$10^{-8/3}$	0.8653
4ª	*cod_escola*	0.8453	2	0.8226	2^{-5}	0.8404	10	$10^{-4/3}$	0.8728
5ª	*media_s*	0.8559	1	0.8443	2^{-3}	0.8455	10	$10^{-5/3}$	0.8778
6ª	*sexo*	0.8621	1	0.8641	2^{-3}	0.8456	20	$10^{-3/3}$	0.8766
7ª	*bolseiro_s*	0.8672	2	0.8702	2^{-3}	0.8440	20	$10^{-5/3}$	0.8872
8ª	*dir_associativo_s*	0.8671	3	0.8695	2^{-3}	0.8451	20	$10^{-3/3}$	0.8867

Tabela 4 – Variáveis selecionadas pela aplicação do método *forward search* combinado ao *dataset* do 1º semestre

[2] Hyperparâmetros considerados na afinação: mtry – número de variáveis escolhidas aleatoriamente para o critério de divisão em cada nodo das árvores das RF; cost – fator de penalização do erro; size – número de neurónios que compõem a camada escondida das RNA; decay – decaimento da taxa de aprendizagem.

Desses resultados conclui-se que no final do 1º semestre escolar do aluno, são 7 as variáveis que mais explicam a sua propensão para o abandono (modelo de previsão que se passará a identificar pela mnemónica *var7s1*), 4 das quais da dimensão curricular, 2 representando dados de matrícula e uma de cariz demográfico – cf. Figura 2. Já para os dados recolhidos ao fim do 2º semestre, o processo de seleção terminou num subconjunto de variáveis mais alargado. Tal como mostra a Tabela 5, são 12 as variáveis que se revelam mais explicativas ao fim do 2º semestre (modelo que se passará designar *var12s12*), não estando nelas representada apenas a categoria M, respeitante aos dados de matrícula – cf. Figura 2.

			RF		MVS		RNA		
ord.	**variável**	**AUC médio**	**mtry**	**AUC**	**cost**	**AUC**	**size**	**decay**	**AUC**
1ª	*ects_reprov_s*	0.8568	1	0.8119	2^{-5}	0.8793	1	$10^{-8/3}$	0.8793
2ª	*cod_prof_mae*	0.8740	1	0.8564	2^{9}	0.8818	5	$10^{-1/3}$	0.8839
3ª	*n10_11_acesso*	0.8797	1	0.8749	2^{-1}	0.8811	5	$10^{-1/3}$	0.8832
4ª	*idade*	0.8855	1	0.8904	2^{5}	0.8818	5	$10^{-1/3}$	0.8842
5ª	*media_s*	0.8883	1	0.8897	2^{13}	0.8821	2	$10^{-6/3}$	0.8931
6ª	*vd12_s*	0.8899	2	0.8867	2^{15}	0.8894	1	$10^{-5/3}$	0.8935
7ª	*sit_prof_mae*	0.8909	1	0.8934	2^{11}	0.8844	1	$10^{-1/3}$	0.8950
8ª	*sit_prof_aluno*	0.8921	2	0.8952	2^{9}	0.8846	2	$10^{-1/3}$	0.8964
9ª	*nacionalidade*	0.8941	2	0.8928	2^{13}	0.8912	2	$10^{-1/3}$	0.8983
10ª	*dir_associativo_s*	0.8959	2	0.9007	2^{15}	0.8903	2	$10^{0/3}$	0.8967
11ª	*nivel_esc_pai*	0.8975	2	0.9025	2^{11}	0.8906	1	$10^{0/3}$	0.8993
12ª	*cod_prof_aluno*	0.8977	2	0.9051	2^{11}	0.8883	1	$10^{0/3}$	0.8997
13ª	*min_s*	0.8962	3	0.9045	2^{9}	0.8849	1	$10^{0/3}$	0.8993

Tabela 5 – Variáveis selecionadas pela aplicação do método *forward search* combinado ao dataset dos 2 primeiros semestres

Para uma mais fácil referenciação, os modelos de previsão que integram todas as variáveis de entrada disponíveis serão doravante identificados simplesmente por *var38s1* (modelo suportado por todas as 38 variáveis do *dataset* obtido no final do 1º semestre) e *var41s12* (modelo suportado pelas 41 variáveis do 2º semestre), nomenclatura já usada no diagrama da Figura 2.

4.3. Avaliação da capacidade de generalização dos modelos encontrados

Concluídos os processos de treino, afinação e *feature selection*, os modelos obtidos são avaliados com base no subconjunto de teste, de forma a avaliar a sua verdadeira capacidade preditiva e em especial a sua capacidade de generalização. Cada um dos quatro modelos desenvolvidos, descritos nas 2 secções anteriores, foram então usados para prever os abandonos escolares registados no subconjunto de dados de teste, tendo-se obtido os resultados expressos na Tabela 6, a qual também inclui, para efeitos de comparação, os resultados anteriormente obtidos com os dados de validação.

				RF			MVS				RNA	
modelo	AUC médio		mtry	AUC		cost	AUC		size	decay	AUC	
	valid	teste		valid	teste		valid	teste			valid	teste
var38s1	0.8630	0.7661	5	0.8548	0.7778	2^{5}	0.8554	0.7533	50	1	0.8787	0.7671
var41s12	0.8896	0.7825	3	0.8889	0.8040	2^{3}	0.8792	0.7639	1	$10^{-1/3}$	0.9007	0.7795
var7s1	0.8672	0.7303	2	0.8702	0.7384	2^{-3}	0.8440	0.7074	20	$10^{-5/3}$	0.8872	0.7451
var12s12	0.8977	0.7582	2	0.9051	0.7605	2^{11}	0.8883	0.7548	1	1	0.8997	0.7592

Tabela 6 – Aplicação dos modelos encontrados aos dados deixados para teste

Comparando o AUC médio obtido com os dados de teste, com o que se tinha obtido com os dados de validação, constata-se, como já esperado, que os valores do AUC baixam - cerca de uma décima nos modelos que usam todas as variáveis disponíveis e perto de 1.4 décimas nos modelos que sofreram redução de variáveis. Estes mesmos resultados evidenciam que os modelos em que se procedeu à seleção dos fatores mais explicativos (modelos var7s1 e var12s12), embora se tenham revelado melhores que os modelos var38s1 e var41s12 nos dados de validação, infelizmente, são aqueles que apresentam pior eficácia com os dados de teste. Esta constatação revela que a redução de variáveis nos modelos teve como consequência alguma diminuição da sua capacidade de generalização. Ainda que se preveja algum agravamento no desempenho desses modelos quando aplicados em contexto real, a utilidade e pertinência do processo de seleção de variáveis não estará em causa, dado proporcionar um conjunto de outras vantagens, designadamente, permitir, desde logo, identificar os principais fatores que expliquem as variáveis alvo, ajudar a eliminar variáveis redundantes, diminuir a complexidade computacional do modelo e facilitar, quer a sua interpretabilidade, quer a sua aplicação em contexto real.

No que respeita ao comportamento do modelo combinado, quando exposto aos dados de teste, a Tabela 7 apresenta os respetivos valores de AUC obtidos, bem como a média e melhores valores de AUC conseguidos com os 3 algoritmos de classificação que o integram, para efeitos de comparação.

modelo	melhor AUC	AUC médio	AUC mod. Combinado
var38s1	0.7778 (RF)	0.7661	0.7757
var41s12	0.8040 (RF)	0.7825	0.8032
var7s1	0.7451 (RNA)	0.7303	0.7495
var12s12	0.7605 (RF)	0.7582	0.7633

Tabela 7 – Comparação do desempenho do modelo combinado com o desempenho dos 3 algoritmos de classificação que o integram, usando dados de teste

Os resultados tabelados permitem, também eles, reafirmar que o modelo combinado apresenta um desempenho superior ao desempenho médio dos classificadores em

separado. Repare-se, inclusivamente, que praticamente iguala o desempenho do melhor dos classificadores, qualquer que seja o conjunto de variáveis de entrada e *dataset* considerado. De forma a se ter uma outra perspetiva do comportamento de cada um dos 4 modelos combinados de previsão do abandono, sobrepõem-se na Figura 3 as respetivas curvas ROC.

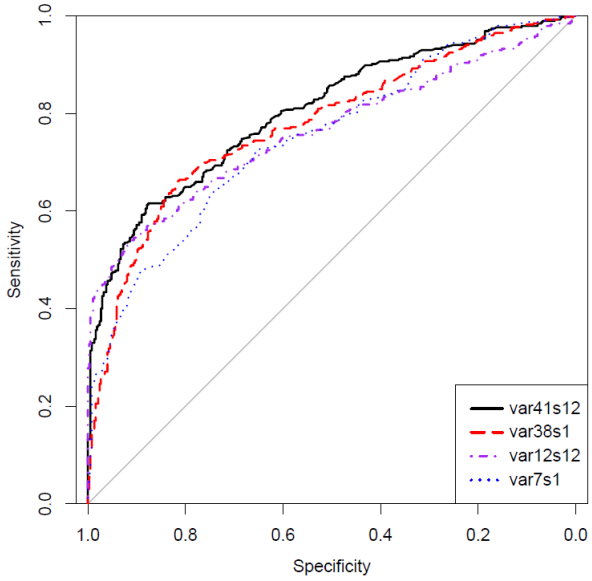

Figura 3 – Curvas ROC dos modelos de classificação combinada

Não sendo facilmente percetíveis, na figura, as diferenças de desempenho entre os vários classificadores, é possível, ainda assim, perceber-se que a curva que mais se aproxima do vértice que caracteriza a condição de otimalidade (especificidade=1 e sensibilidade=1) é a do modelo *var41s12*, e a que menos se aproxima é a do modelo var7s1. Estes dois factos, combinados, corroboram constatações já feitas anteriormente de que o desempenho do modelo de previsão diminui ligeiramente com a seleção das variáveis de entrada (*var7s1* e *var12s12*) e aumenta com a escolha para dados de *input* os resultados do 2º semestre (*var41s12* e *var12s12*).

Ainda que os modelos com variáveis de entrada selecionadas percam alguma da sua capacidade de generalização face aos modelos *var38s1* e *var41s12*, não deixa de ser interessante o nível de desempenho que, mesmo assim, revelam ter quando aplicados a dados completamente novos. Pelos resultado obtidos, será de esperar um AUC em torno dos 0.75 na previsão dos abandonos escolares, quando realizada no final do 1º semestre escolar e a partir de 7 variáveis explicativas (modelo *var7s1*), e um AUC um pouco maior (acima dos 0.76) na previsão realizada ao fim do 2º semestre, com base em 12 variáveis (modelo *var12s12*) – cf. Tabela 7.

4.4. Importância relativa dos principais fatores explicativos

Depois de encontrado o conjunto dos fatores mais explicativas do abandono, a fim de diferenciar qual a influência das múltiplas dimensionalidades do estudante, estabeleceu-se entre elas uma ordem de relevância, em termos da sua importância para a previsão.

A importância relativa dos fatores explicativos foi então determinada para os modelos *var7s1* e *var12s12*, encontrando-se os seus valores representados, respetivamente, nos gráficos das figuras 4 e 5.

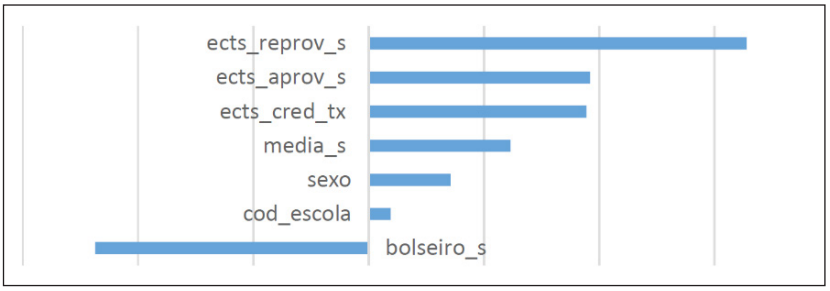

Figura 4 – Importância relativa das variáveis explicativas do modelo *var7s1*

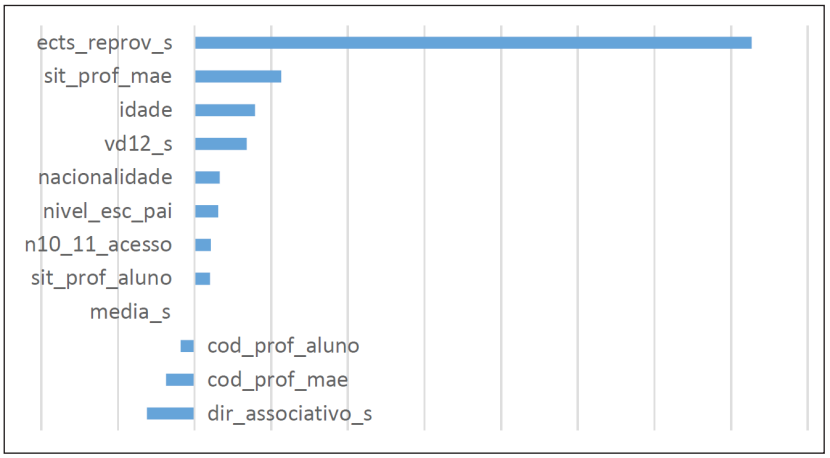

Figura 5 – Importância relativa das variáveis explicativas do modelo var12s12

O aspeto que, desde logo, sobressai em ambos os modelos, é a existência de variáveis com impacto negativo na previsão do abandono. Ainda que não seja o resultado que há partida mais se esperaria, não se deverá cair na tentação de remover essas variáveis, pois estar-se-ia, erradamente, a reconfigurar os modelos com base em dados de teste. Outra particularidade que também sobressai dos resultados obtidos, é a elevada preponderância

que uma única variável tem sobre as restantes no modelo de previsão aplicado ao fim do 2º semestre. Trata-se da variável *ects_reprov_s* (número de ECTS reprovados), um dos principais indicadores conhecidos de desempenho curricular do aluno. Esse facto, exposto de forma evidente na representação gráfica da Figura 5, mostra a grande influência que o desempenho curricular do aluno, no seu ainda curto percurso escolar, já tem na sua propensão para o abandono. Também os resultados do modelo aplicado ao fim do 1º semestre revelam, de alguma forma, esse tipo de influência. Ainda que não se destaquem das restantes, as variáveis *ects_reprov_s* e *ects_aprov_s* (número de ECTS aprovados) acabam por ser as duas variáveis mais explicativas.

A razão para não constar nas 12 variáveis do modelo var12s12 uma variável tão "importante" como a *ects_aprov_s*, dever-se-á ao facto de a mesma apresentar um elevado grau de correlação com a variável *ects_reprov_s*, já incluída no modelo. Acredita-se que esta é também a razão para a variável *ects_reprov_s* não assumir maior preponderância no modelo var7s1.

5. Conclusões

Com o objetivo de se tentar identificar, logo numa fase precoce, os alunos das licenciaturas do IPB que apresentem maior propensão para o abandono académico, desenvolveram-se dois modelos de previsão distintos, um suportado pelos dados recolhidos no final do 1º semestre escolar do aluno e o outro suportado pelos dados acumulados até ao final do seu 2º semestre escolar. Para cada um dos modelos foi desenvolvida uma solução que combina três importantes técnicas de DM: os algoritmos de classificação RF, MVS e RNA. Ainda que composto por três algoritmos independentes, conseguiu-se encontrar um esquema, baseado no método *forward search*, através do qual foi possível chegar a um único subconjunto de variáveis, consideradas mais explicativas, para suporte do modelo combinado. Com essa metodologia, foram 7 as variáveis que, no final do 1º semestre escolar, se revelaram mais explicativas do abandono e 12 as que se revelaram mais explicativas no final do 2º semestre.

Através de um estudo posterior de análise de sensibilidade percebeu-se a grande influência que tem, na propensão para um futuro abandono, o desempenho curricular que o aluno apresenta logo no seu 1º ano de formação. Essa influência é evidenciada no final do 1º semestre pela importância do número de ECTS quer aprovados quer reprovados, e no final do 2º semestre unicamente pela importância do número de ECTS reprovados, mas neste caso assumindo esse fator uma elevada preponderância sobre todos os restantes. Por conseguinte, ambos os modelos indiciam que as prioridades de intervenção da gestão académica, no combate ao abandono, deverão estar centradas numa supervisão atenta, sobretudo, sobre os estudantes que apresentem logo no seu primeiro ano baixo rendimento académico.

Ainda que já se esperasse, através do conhecimento empírico, que o número de ECTS (ou o número de unidades curriculares) que o aluno consegue ou não concluir logo no seu primeiro ano de estudos fosse um importante preditor do abandono, esta investigação vem confirmar, com recurso a métodos científicos, a veracidade dessa relação, em consonância com outros estudos já publicados.

Os resultados deste estudo revelam-se de crucial importância uma vez que, no caso concreto do IPB, os índices de abandono são muito preocupantes, chegando a cerca de 40% os alunos que, no conjunto de dados analisado, acabam por não concluir a sua licenciatura. O conhecimento emanado deste estudo será fundamental para a delineação de medidas preventivas urgentes, precoces e precisas, que levem à diminuição dos índices de evasão escolar.

Agradecimentos

Este trabalho foi suportado pela Fundação para a Ciência e Tecnologia (FCT) através do Projeto UID/EEA/04131/2019. Agradece-se igualmente ao IPB, e em particular ao seu pró-presidente para os Sistemas de Informação, Prof. Doutor Albano Alves, pela disponibilização dos dados analisados no presente estudo.

Referências

Aluko, R. O., Adenuga, O. A., Kukoyi, P. O., Soyingbe, A. A., & Oyedeji, J. O. (2016). Predicting the academic success of architecture students by pre-enrolment requirement: using machine-learning techniques. *Construction Economics and Building, 16(4)*, 86-98.

Baker, R. S. (2010). Data mining for education. *International encyclopedia of education, 7(3)*, 112-118.

Bakhshinategh, B., Zaiane, O. R., ElAtia, S., & Ipperciel, D. (2018). Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies, 23(1)*, 537-553.

Burgos, C., Campanario, M. L., Peña, D. d., Lara, J. A., Lizcano, D., & Martínez, M. A. (2017). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering*.

Costa, E. B., Fonseca, B., Santana, M. A., Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior, 73*, 247-256.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems, 49(4)*, 498-506.

Hoffait, A.-S., & Schyns, M. (2017). Early detection of university students with potential difficulties. *Decision Support Systems, 101*, 1-11.

Huang, S. (2011). *Predictive modeling and analysis of student academic performance in an engineering dynamics course*. (Doctoral Thesis in Engineering Education), Utah State University, Logan, Utah, United States.

Kotsiantis, S. B., Pierrakeas, C., & Pintelas, P. E. (2003). Preventing student dropout in distance learning using machine learning techniques. *In International Conference on Knowledge-Based and Intelligent Information and Engineering Systems* (pp. 267-274). Springer.

Manhães, L. M. (2015). *Predição Do Desempenho Acadêmico De Graduandos Utilizando Mineração De Dados Educacionais*. (Doctoral Thesis) Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil.

Martins, M., Miguéis, V., Fonseca, D., & Alves, A. (2019). A Data Mining Approach for Predicting Academic Success - A Case Study. Em *Information Technology and Systems. ICITS 2019. Advances in Intelligent Systems and Computing* (Vol. 918, pp. 45-56). Springer.

Miguéis, V. L., Freitas, A., Garcia, P. J., & Silva, A. (2018). Early segmentation of students according to their academic performance: A predictive modelling approach. *Decision Support Systems, 115*, 36-51.

Nandeshwar, A., Menzies, T., & Nelson, A. (2011). Learning patterns of university student retention. *Expert Systems with Applications, 38(12)*, 14984-14996.

Natek, S., & Zwilling, M. (2014). Student data mining solution-knowledge management system related to higher education institutions. *Expert systems with applications, 41(14)*, 6400-6407.

Pascoal, T. A., Brito, D. M., & Rêgo, T. G. (2015). Uma abordagem para a previsão de desempenho de alunos de computação em disciplinas de programação. *Nuevas Ideas en Informática Educativa TISE, 2015*, 454-458.