Uma Metodogia de Data Mining para Prever o Desempenho de Estudantes de Licenciatura

A Data Mining Approach to Predict Undergraduate Students' Performance

Maria P. G. Martins
Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Bragança
Bragança, Portugal
CISE - Centro de Investigação em Sistemas Electromecatrónicos, Universidade da Beira Interior
Covilhã, Portugal

Vera L. Miguéis, INESC TEC
Faculdade de Engenharia da Universidade do Porto
Porto, Portugal

D. S. B. Fonseca, *Member*, *IEEE*
CISE - Centro de Investigação em Sistemas Electromecatrónicos, Universidade da Beira Interior
Covilhã, Portugal

Resumo — **No presente artigo apresenta-se uma metodologia desenvolvida com base no algoritmo *random forest*, para prever precocemente e de forma rigorosa o desempenho académico de graduação dos estudantes de uma instituição de ensino superior politécnico. A abordagem seguida permitiu isolar 11 variáveis explicativas, a partir de um conjunto inicial de cerca de meia centena, que garantem uma boa capacidade preditiva do modelo (R^2=0.79). Estas variáveis revelam aspetos fundamentais para a definição de estratégias de gestão centradas na promoção do sucesso académico.**

Palavras Chave - data mining educacional, random forest, regressão, sucesso académico.

Abstract — **This paper presents a methodology based on random forest algorithm to predict the undergraduate academic performance of students from a polytechnic institution. The approach followed enabled to select 11 explanatory variables, starting from an initial set of around fifty, which allow to obtain a good predictive performance (R^2=0.79). These variables reveal crucial aspects for the definition of management strategies focused on promoting academic success.**

Keywords - educational data mining, random forest, regression, academic success.

I. INTRODUÇÃO

A antecipação de dificuldades de aprendizagem e a recuperação de alunos com desempenhos negativos são temas de elevada relevância e reflexão ao nível da gestão de instituições educativas. Neste contexto, e com o intuito de providenciar uma ferramenta que venha a auxiliar os agentes de decisão do Instituto Politécnico de Bragança (IPB) desenvolveu-se um modelo preditivo de regressão, baseado no algoritmo *random forest* [1]. Mais precisamente, pretendeu-se antever o (in)sucesso educacional global dos estudantes no términus do seu percurso académico e em simultâneo identificar os principais fatores que o possam explicar. A metodologia adotada na seleção desses fatores foi desenvolvida em duas fases distintas. Logo num primeiro momento foi possível reduzir a "praga" da dimensionalidade dos dados sem se perder a capacidade assertiva do modelo, deixando-se para uma segunda fase um ajuste mais seletivo dos fatores de sucesso.

II. TRABALHOS RELACIONADOS

A utilização de técnicas de data mining em análise de dados provenientes dos processos de ensino e de aprendizagem dos estudantes é uma área de investigação emergente designada Educational Data Mining (EDM) [2][3][4]. Trata-se de uma área que tem despertado um indiscutível interesse junto da comunidade científica, principalmente ao longo desta última década do século XXI. O interesse advém do facto de se terem revelado metodologias de grande alcance e comprovado sucesso quando há a necessidade de promover o sucesso académico dos estudantes e a melhoria contínua dos palcos onde a aprendizagem ocorre [5].

De entre toda a tipologia de tarefas onde o EDM tem sido útil destaca-se a previsão do desempenho académico dos estudantes. Os estudos sobre a previsão da média final de curso têm contribuído para identificar quais os fatores determinantes do (in)sucesso. Por exemplo, Natek and Zwilling [6] concluíram que os fatores que influenciaram de forma mais visível a média final dos estudantes de licenciatura em Informática estavam

relacionados com a informação de acesso, demográfica e com as atividades extracurriculares. Os dados foram processados pelos algoritmos de classificação *RepTree Model*, *J48 Model* e *M5P Model*, e estavam relacionados com 42 alunos do 1º ano académico, 32 do 2º e 32 do 3º. Em [7] foi encontrada uma relação significativa entre o (in)sucesso na conclusão do curso e o grau de comprometimento dos alunos com as tarefas do processo educacional, por via do algoritmo *apriori* de regras de associação. A realização da autoavaliação, o uso dos fóruns e chats, e o preenchimento do questionário de aprendizagem foram as ações que mais afetaram o (in)sucesso dos estudantes.

Num outro estudo, Amrieh, Hamtini and Aljarah [8] usando métodos de conjunto, como o *bagging, boosting e o random forest,* demonstraram que há uma forte relação entre os atributos comportamentais relacionados com a capacidade de aprendizagem e os resultados académicos. Nesta investigação foi considerada uma amostra de 500 estudantes, caracterizados por um total de 16 atributos, classificados em três categorias principais: o histórico académico dos estudantes, as suas características demográficas e as suas características comportamentais, ou atitudes, de aprendizagem durante o processo educacional. Os autores sublinharam que a decisão de incluírem variáveis da categoria comportamental como, por exemplo, participação em chats e em grupos de discussão, intervenção e participação em sala de aula, no modelo criado, contribuiu para uma melhoria avaliada em 21.1% da capacidade preditiva do modelo, quando comparada com os resultados obtidos por aplicação do mesmo modelo sem inclusão desse mesmo tipo de atributos. Em [9] os autores recorreram à análise discriminante linear e ao algoritmo *K-NN* para prever a média final de curso de 101 alunos recém-ingressados no curso de arquitetura, de uma universidade nigeriana. As 13 variáveis independentes usadas para a previsão estavam todas relacionadas com o desempenho pré-universitário dos estudantes. A análise discriminante linear mostrou que os fatores mais influentes para a previsão foram as notas do exame de acesso universitário, o facto de serem ou não candidatos de entrada direta e as notas obtidas em Matemática I.

Num outro estudo, Papamitsiou, Karapistoli and Economides [10] recorrendo aos algorimos *ANN; SVM, NB, K-NN e treeBagger*, desenvolveram uma investigação relativa à dinâmica comportamental dos alunos, com base no tempo que os alunos levam a resolver os problemas num processo de avaliação realizado em computador. Os resultados do estudo demonstraram, com elevada precisão, que o tempo usado na resolução do teste tem uma relação direta com o desempenho efetivo verificado nesse mesmo teste. Os dados analisados caracterizam 301 estudantes inscritos na disciplina Informática II, do departamento de economia de uma universidade grega.

Os resultados do recente estudo de Asif, Merceron, Ali and Haider [11] mostraram que através dos algoritmos *Naive Bayes e Random Forest Trees*, é possível prever com elevada precisão o desempenho global da graduação de um curso de quatro anos, usando apenas classificações pré-universitárias e as notas das disciplinas do 1º e 2º ano da universidade.

A literatura de EDM demonstra que a identificação dos fatores de sucesso e de insucesso escolar e a antecipação do desempenho dos estudantes, pode providenciar conhecimento útil que fundamente e apoie a tomada de decisões destinadas a melhorar a aprendizagem da comunidade estudantil e a eficiência das instituições dedicadas ao ensino.

III. METODOLOGIA E MODELO DE DADOS

Para a criação do modelo preditivo que consiga prever o sucesso académico dos alunos no términus do seu percurso académico, optou-se por explorar dados de um universo estudantil de diversas áreas educacionais, a frequentar cerca de meia centena de cursos de uma instituição com 5 escolas, ao invés de se seguir o procedimento mais usual de delimitar a previsão a um só curso específico. Desta forma pretende-se dotar a instituição de uma ferramenta única capaz de acomodar a heterogeneidade do universo de estudantes e as dinâmicas educativas. Pretende-se que esta ferramenta venha a ser usada a um nível central pelos agentes de decisão que têm capacidade para desenhar ações para mitigar o insucesso académico, promovendo uma melhor experiência educativa aos seus estudantes.

Embora existam vários algoritmos de previsão, neste estudo optou-se por basear o modelo preditivo no algoritmo, proposto por Breiman [1]. O mesmo autor sublinha as qualidades que o caracterizam, destacando a sua capacidade de modelar relações não-lineares de alta dimensionalidade, de lidar com variáveis independentes contínuas, de resistência ao *overfitting* e de apresentar uma boa tolerância ao ruído. Adicionalmente, de acordo com estudos similares, o *random forest* tem superado outras técnicas quanto à capacidade preditiva demonstrada, além de apresentar a interessante funcionalidade de permitir ordenar a importância dos preditores que suportam o modelo. Concretizando para o caso do presente estudo, a ordem de importância estabelecida poderá ser usada pelos decisores académicos para identificar os fatores explicativos do sucesso dos alunos e também para averiguar a necessidade de ações diferenciadas tendo em conta o perfil dos mesmos.

Seguindo o procedimento que normalmente é adotado, todas as simulações em que se testou o modelo preditivo foram realizadas através do método de validação cruzada *k-fold* (com k = 10). Como métrica de avaliação do modelo utilizou-se o coeficiente de determinação, R^2, que mede a correlação entre os valores observados e os valores preditos.

Para aferir o desempenho académico do estudante usou-se, como indicador de sucesso, a variável dependente introduzida na expressão (1),

$$vd = media \times ects\_aprov /(ects\_aprov+ects\_reprov) \quad (1)$$

onde *media* é a média ponderada das notas obtidas nas unidades curriculares (UCs) concluídas, *ects_aprov* o nº de ECTS concluídos com sucesso e ects_*reprov* o nº de ECTS em que o aluno se inscreveu sem que tenha conseguido aprovação. Desta forma, a métrica leva em conta, não apenas a classificação média do aluno, mas também a fração de inscrições em unidades curriculares em que tenha conseguido aprovação (rácio de "tentativas" bem-sucedidas).

No que respeita à inclusão das variáveis preditivas de sucesso académico, considerou-se, essencialmente, a mesma tipologia de variáveis que foi usada nos trabalhos relacionados

de referência, designadamente, dados académicos, de cariz sociodemográfico e de acesso ao ensino superior. Não foi possível incluir variáveis de cariz comportamental e de atividades extracurriculares por não ser recolhida esse tipo de informação na instituição usada como caso de estudo. É possível classificar as variáveis consideradas em dois importantes subgrupos: variáveis com resultados curriculares semestrais acumulados e varáveis 'intemporais' – variáveis cujos valores não se alteram ao longo do percurso escolar do aluno. A Tabela I apresenta todas as variáveis preditivas consideradas no estudo.

TABELA I. CONJUNTO INICIAL DE VARIÁVEIS PREDITIVAS.

Atributo	cat	significado
ano_curricular_s	C	ano do aluno no semestre escolar (s.e.)
ano_s	C	ano letivo do s.e. considerado
bolseiro_s	C	foi bolseiro no s.e.?
dir_associativo_s	C	foi dirigente associativo no s.e.?
ects_aprov_s	C	nº de ECTS aprovados no s.e.
ects_reprov_s	C	nº de ECTS reprovados no s.e.
max_s	C	nota máxima das UCs aprovadas no s.e.
media_s	C	nota média das UCs aprovadas no s.e.
min_s	C	nota mínima das UCs aprovadas no s.e.
navalr_s	C	nº de avaliações sem aprovação no s.e.
nuca_s	C	nº de UCs aprovadas no s.e.
nucr_s	C	nº de UCs reprovadas no s.e.
ano_mat	M	ano da matrícula
cod_curso	M	código do curso
cod_escola	M	código da escola
ects_cred_tx	M	fração de ECTS creditados ao aluno
ects_curso	M	número de ECTS do curso
tipo_ing	M	tipo de ingresso
conc	D	concelho de proveniência do aluno
conc_n	D	concelho de naturalidade
deslocado	D	está deslocado da sua residência habitual?
dist	D	distrito de proveniência do aluno
dist_n	D	distrito de naturalidade
freg	D	freguesia de proveniência
freg_n	D	freguesia de naturalidade
idade	D	idade no ato da matrícula
nacionalidade	D	nacionalidade do aluno
sexo	D	género
cod_prof_aluno	S	profissão do aluno
cod_prof_mae	S	profissão da mãe
cod_prof_pai	S	profissão do pai
nivel_esc_mae	S	nível de escolaridade da mãe
nivel_esc_pai	S	nível de escolaridade do pai
sit_prof_aluno	S	situação profissional do aluno
sit_prof_mae	S	situação profissional da mãe
sit_prof_pai	S	situação profissional do pai
fase	A	fase de acesso
media_acesso	A	nota de acesso ao ensino superior
n10_11_acesso	A	média dos 10º e 11º anos
n12_acesso	A	média do 12º ano
opcao_acesso	A	ordem da opção na candidatura ao curso
ordem_acesso	A	ordem de acesso entre os colocados no curso
pi_acesso	A	nota média das provas de ingresso

Foi possível considerar o período temporal compreendido entre 2007/2008 e 2015/2016, perfazendo 9 anos letivos consecutivos. Optou-se por restringir o estudo apenas aos cursos de licenciatura, por se tratar do core da oferta formativa da instituição e por abranger um conjunto de dados mais completo. Depois da limpeza de dados e de outras tarefas de pré-processamento, o *data set* em que este estudo incide passou a respeitar a 2159 matrículas em cursos de licenciatura, concluídas no período 2007/08–2015/16 e iniciadas dentro do período 2007/08– 2013/14.

Partindo-se de um *data set* real, houve o cuidado de classificar (2ª coluna da tabela) cada uma das variáveis preditivas, de acordo com a sua natureza, numa das seguintes categorias: *curriculares* (C), *de matrícula* (M), *demográficas* (D), *socioeconómicas* (S) e *de acesso* (A). Passou-se assim a dispor de 5 subgrupos de variáveis preditivas, facilmente referenciáveis através das letras C, M, D, S e A. Também os atributos com dados semestrais (todos os da categoria C) são facilmente distinguíveis dos restantes (dados intemporais) pelo sufixo "_s". Note-se que os dados curriculares (C) referem-se a resultados de desempenho académico semestrais acumulados ao fim de cada um dos 6 primeiros semestres do aluno.

A seleção das dimensões relativas ao aluno, explicativas do seu sucesso escolar, processou-se em duas fases distintas. Primeiro, procedeu-se, em §IV.A, à seleção das dimensões do aluno que melhor explicam o seu sucesso, conseguindo-se, com isso, um primeiro ajuste que permitiu eliminar grupos completos de variáveis. Seguidamente, em §IV.B, procedeu-se a um ajuste mais fino na seleção dos atributos que não foram excluídos na primeira fase. Com a abordagem mencionada foi possível reduzir a "praga" da dimensionalidade dos dados, sem se ter perdido a capacidade preditiva do modelo.

Na Fig.1 apresenta-se um esquema ilustrativo que pretende caraterizar o modelo de previsão arquitetado no âmbito do atual estudo. Nesse esquema são evidenciadas as diferentes categorias de variáveis preditivas usadas como *input* do algoritmo *random forest*. Como se tenta ilustrar, para o grupo de variáveis curriculares (C) são usados os resultados acumulados ao fim de cada um dos 6 primeiros semestres escolares do aluno. Note-se que apenas uma das 6 entradas de dados curriculares (C) é considerada em cada execução do algoritmo (entradas mutuamente exclusivas).

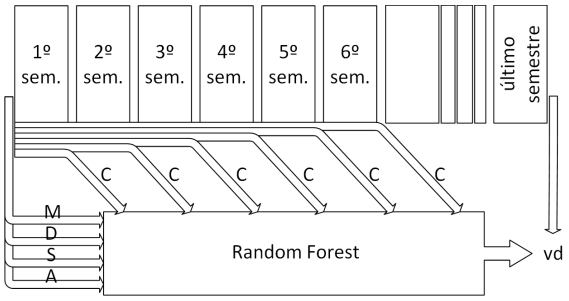

Figura 1. Esquema ilustrativo do estudo comparativo realizado.

IV. IDENTIFICAÇÃO DOS FATORES DE SUCESSO

Na análise exploratória de dados, opta-se por manter fixa a configuração do algoritmo *random forest*, para o foco do estudo incidir no conjunto de variáveis preditivas que lhe darão suporte.

A. Seleção de Categorias de Preditores

Como se sabe, a assertividade dum modelo de previsão depende muito do conjunto de variáveis preditivas que forem consideradas na análise. E o melhor modelo nem sempre é o que inclui todas as variáveis disponíveis. Embora seja do conhecimento geral que as *random forest* fazem uma seleção interna de variáveis, resolveu-se, ainda assim, fazer um teste de inclusão, ou não, de categorias de variáveis, a fim de se perceber o impacto das várias dimensões na capacidade do modelo.

Na Tabela II apresentam-se as diferentes categorias de atributos escolhidas em cada um desses estudos.

TABELA II. CATEGORIAS DE VARIÁVEIS PREDITIVAS USADAS NOS ESTUDOS REALIZADOS.

		Categoria				
Estudo	***mnem.***	Curriculares	de Matrícula	Demográficos	Socioeconómicos	de Acesso
1	CMDSA	✓	✓	✓	✓	✓
2	CMDS	✓	✓	✓	✓	
3	CMDA	✓	✓	✓		✓
4	CMSA	✓	✓		✓	✓
5	CDSA	✓		✓	✓	✓
6	MDSA		✓	✓	✓	✓
7	CMD	✓	✓	✓		
8	CMS	✓	✓		✓	
9	CMA	✓	✓			✓
10	CDS	✓		✓	✓	
11	CDA	✓		✓		✓
12	CSA	✓			✓	✓
13	CM	✓	✓			
14	CD	✓		✓		
15	CS	✓			✓	
16	CA	✓				✓
17	C	✓				

Ainda que o objetivo seja desenvolver um estudo exaustivo, entendeu-se não ser necessário abranger todas as combinações possíveis dos 5 grupos de variáveis, que perfaz um total de 2^5-1 = 31 possibilidades. Na verdade, sendo os dados curriculares (C) o grupo de preditores claramente mais determinante, antevê-se grande dificuldade de precisão preditiva a um qualquer modelo que o não venha a incluir. Por isso, apenas se considera um caso particular onde não se usa esse grupo de variáveis: o caso MDSA (Estudo 6). Com esta simplificação, o número de estudos reduziu-se a quase metade, ou mais concretamente a $2^4 + 1 = 17$.

Na Tabela III são apresentados, para cada um dos estudos, os coeficientes de determinação (R^2) obtidos com a aplicação do algoritmo de previsão *random forest* aos dados dos grupos selecionados, ao fim de cada um dos 6 primeiros semestres escolares do aluno.

TABELA III. COEFICIENTE DE DETERMINAÇÃO R^2 PARA DIFERENTES GRUPOS DE VARIÁVEIS PREDITIVAS, E EM FUNÇÃO DO SEMESTRE ESCOLAR CONSIDERADO.

		Semestre Escolar						
	mnem.	***1º***	***2º***	***3º***	***4º***	***5º***	***6º***	***média***
Est. 8	CMS	80.8	86.7	91.9	94.3	96.6	97.9	88.6
Est. 2	CMDS	80.9	86.5	91.7	94.1	96.6	97.8	88.5
Est. 13	CM	80.2	86.5	91.9	94.4	96.7	98.0	88.4
Est. 7	CMD	80.3	86.4	91.7	94.2	96.6	97.9	88.3
Est. 4	CMSA	80.5	86.3	91.6	94.0	96.5	97.8	88.3
Est. 1	CMDSA	80.6	86.4	91.5	93.9	96.4	97.8	88.3
Est. 3	CMDA	80.0	86.1	91.5	94.0	96.4	97.7	88.0
Est. 9	CMA	79.6	86.0	91.5	94.1	96.5	97.8	87.9
Est. 10	CDS	71.5	78.8	86.7	90.0	94.3	96.6	82.1
Est. 5	CDSA	71.1	78.7	86.7	90.1	94.4	96.6	82.0
Est. 15	CS	71.2	78.8	86.7	89.9	94.3	96.7	82.0
Est. 12	CSA	70.8	78.8	86.7	90.0	94.3	96.7	81.9
Est. 14	CD	70.7	78.5	86.7	90.0	94.4	96.7	81.9
Est. 11	CDA	70.3	78.6	86.7	90.1	94.4	96.6	81.8
Est. 17	C	70.3	78.0	86.6	90.0	94.4	96.7	81.6
Est. 16	CA	69.5	78.3	86.5	90.1	94.4	96.7	81.4
Est. 6	MDSA	63.8	63.8	63.8	63.8	63.8	63.8	63.8
Média		75.5	82.5	89.2	92.1	95.5	97.3	85.1

Para um melhor entendimento dos resultados tabelados, tome-se, como exemplo, o Estudo 8: o input do algoritmo *random forest* (ver Fig.1) resumiu-se aos grupos de variáveis curriculares (C), de matrícula (M) e socioeconómicas (S); neste, como nos restantes estudos (à exceção do Estudo 6, que não inclui dados curriculares), correu-se 6 vezes o algoritmo *random forest*, de forma a usarem-se para dados curriculares os resultados acumulados ao fim de cada um dos 6 semestres. Na última coluna da tabela apresenta-se a média ponderada dos coeficientes de determinação desses 6 estudos. Optou-se por uma média ponderada dos R^2 semestrais, com pesos 6, 5, ..., 2, 1, para os semestres 1º, 2º, ..., 5º, 6º, respetivamente, de forma a

valorizar os resultados dos primeiros semestres em detrimento dos obtidos em momentos mais avançados do percurso escolar do aluno. Sendo a métrica de desempenho que será considerada na escolha do melhor modelo, entendeu-se que seria adequado valorizar a capacidade preditiva do modelo evidenciada logo no 1º semestre 6 vezes mais do que a demonstrada ao fim do 6º – repare-se, por exemplo, que para os alunos que concluem a sua formação em 3 anos, é completamente irrelevante a capacidade preditiva que o modelo possa apresentar ao fim do 6º semestre. Por sua vez, a média do R^2 para cada semestre, apresentada na última linha da mesma tabela, não contabiliza o Estudo 6, uma vez que não inclui dados curriculares, os que mais contribuem para a capacidade de acerto do modelo.

Observando os valores tabelados, listados por ordem decrescente do valor médio do R^2 (última coluna), pode, desde logo, concluir-se que:

- Não foi o modelo que se "alimenta" da totalidade das variáveis (Estudo 1 – CMDSA) que apresentou melhores capacidades preditivas. Na verdade, 5 outros modelos conseguiram iguais ou melhores desempenhos, com um menor número de variáveis preditivas.
- É notória a grande diferença de desempenho entre os 8 modelos melhor classificados e os restantes – repare-se na sua repentina diminuição entre o Estudo 9 e o 10. Se essa quebra repentina se deve claramente à perda dos dados de matrícula (M), a que sobressai entre os estudos 16 e 6 fica a dever-se à perda do outro subgrupo de dados académicos, os de natureza curricular (C).
- Como não poderia deixar de ser, a assertividade do modelo aumenta consistentemente com o avanço do percurso escolar do aluno.
- Os resultados do Estudo 6 (MDSA), que é o único que não inclui atributos da categoria C, estando muito aquém dos restantes, confirmam claramente que são os dados curriculares do aluno que mais contribuem para a capacidade de acerto do modelo, ao qual não será, certamente, alheio o facto de serem também aqueles de que não se dispõe no início do percurso académico. Em todo o caso, apraz-se constatar que mesmo numa fase ainda muito precoce do percurso escolar do aluno, que são os seus 1º e 2º semestres, o coeficiente de determinação médio do modelo com melhor desempenho dispara de 63.8% para 80.8% e 86.7%, respetivamente.

Atendendo aos resultados tabelados, e no seguimento das correspondentes considerações anteriores, parece pertinente propor para o modelo preditivo que se pretende arquitetar os grupos de variáveis do Estudo 8. Conseguem os mais elevados coeficientes de determinação, usando apenas 3 categorias de variáveis, das 5 possíveis.

A corroborar a escolha do grupo CMS (Estudo 8), como o melhor dos estudos, está ainda o facto desse conjunto de variáveis ter-se também destacado dos restantes nas suas qualidades preditivas logo em fases precoces do percurso escolar do aluno. Se no 1º semestre fica a uns residuais 0.1% do melhor coeficiente de determinação dos estudos (80.9%), no 2º é mesmo esse o grupo que apresenta melhor desempenho (86.7%).

Com o estudo comparativo que se realizou conseguiram-se já excluir dois "importantes" grupos de variáveis: os demográficos e os dados de acesso. Seguidamente, tentar-se-á, dentro dos grupos que se mantiveram, excluir variáveis que apresentem uma influência negligenciável no desempenho do modelo preditivo.

B. Ajuste Adicional do Modelo – Seleção de Preditores

O desempenho de um qualquer modelo de previsão que se venha a propor será tanto mais valorizado quanto mais precoce for o momento em que ele possa vir a ser aplicado. Na verdade, a relevância preditiva dum modelo, assenta em duas vertentes cruciais: a veracidade das suas previsões e o grau de antecipação com que as consegue obter. Por isso, nesta fase do trabalho, procura-se ajustar com maior precisão o modelo CMS (o que mostrou ser globalmente mais assertivo), quando aplicado logo ao fim do 1º semestre escolar do aluno, tal como se tenta ilustrar no esquema da Fig. 2. Em concreto, tentar-se-á retirar do conjunto de variáveis preditivas CMS (com o subgrupo C a incluir apenas resultados curriculares do 1º semestre do aluno), todas aquelas que não contribuam positiva e significativamente para a qualidade do modelo.

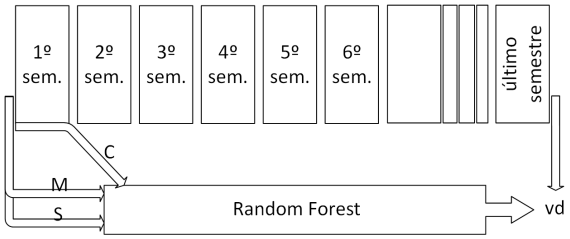

Figura 2. Modelo preditivo CMS suportado por dados do 1º semestre escolar.

Uma vez que o modelo CMS prescindiu dos dados de acesso, passou-se a dispor de uma amostra mais abrangente de matrículas com dados completos. Perante a possibilidade de incluir no estudo matrículas sem dados de acesso, o tamanho do *data set* com que se ajustará o modelo CMS passou, por essa razão, de 2159 para 3109, tendo a assertividade do modelo decrescido ligeiramente no 1º dos semestres, de R^2= 80.8 para 78.7.

Ainda antes de se avançar para um processo mais sistematizado de afinação do modelo CMS, correu-se o algoritmo *random forest* para o *data set* sem as variáveis *nuca_s* e *nucr_s*, por se considerar terem, as mesmas, uma forte correlação, respetivamente, com os atributos *ects_aprov_s* e *ects_reprov_s*. Depois de se confirmar a pertinência da exclusão desses dois atributos, avaliou-se, em sucessivas iterações, a possibilidade de exclusão de novas variáveis, entre aquelas que se revelassem menos informativas. Os dados relevantes que caraterizam essas várias iterações encontram-se resumidos na Tabela IV.

TABELA IV. REMOÇÃO DE VARIÁVEIS DO MODELO CMS.

	variáveis excluídas			#var	R^2
iter. 0	nuca_s	nucr_s		24	78.9
iter. 1	cod_prof_aluno	bolseiro_s	dir_associativo_s	21	78.9
iter. 2	cod_prof_mae	cod_prof_pai	ano_curricular_s	18	79.0
iter. 3	tipo_ing	sit_prof_pai	sit_prof_aluno	15	78.7
iter. 4	nivel_esc_pai	sit_prof_mae	min_s	12	79.0
iter. 5	ects_curso			11	79.0

As 11 variáveis, que no seu conjunto justificam a capacidade preditiva do modelo, e que se revelaram, por isso, mais influentes no sucesso educacional dos estudantes de licenciatura do IPB, foram, por ordem decrescente de importância (indicada entre parêntesis), as seguintes: *ects_reprov_s* (2.006), *media_s* (1.525), *cod_escola* (1.451), *ects_aprov_s* (1.350), *ects_cred_tx* (1.126), *navalr_s* (0.863), *cod_curso* (0.738), *max_s* (0.636), *nível_esc_mae* (0.370), *ano_s* (0.351) e *ano_mat* (0.345).

V. CONCLUSÕES

Neste estudo, usando o algoritmo *random forest*, propôs-se um modelo de previsão de sucesso académico global dos estudantes das licenciaturas do Instituto Politécnico de Bragança, aquando do términus do seu percurso académico. Ao invés de se seguir o procedimento normalmente adotado de se delimitar a previsão a um só curso específico, o modelo foi desenvolvido a partir de um *data set* real de grande dimensão, envolvendo registos de grupos de alunos bastante heterogéneos, provenientes de mais de meia centena de licenciaturas que cobrem as mais diversas áreas educacionais ministradas nas cinco escolas da instituição e onde cada estudante é caracterizado por cerca de meia centena de variáveis explicativas. Desta especificidade surgiu a possibilidade de se estudar a influencia de um novo fator curricular considerado na literatura pela primeira vez: o tipo de escola. Os resultados obtidos permitiram concluir que da escola frequentada pelos alunos também depende o seu sucesso. Esta conclusão indicia que para mitigar o insucesso poderá ser necessário adotar estratégias de promoção educacional diferenciadas por escolas. Um resultado semelhante poder-se-ia eventualmente esperar com a variável "tipo de departamento", contudo a mesma não foi considerada no estudo, uma vez que cada uma das escolas da instituição possui uma organização matricial, em que cada departamento leciona unidades curriculares em quase todos os cursos.

A ordem de importância atribuída pelo *random forest* permitiu identificar os fatores de (in)sucesso dos estudantes, permitindo, nomeadamente, perceber que os fatores do contexto curricular (C) de desempenho académico do estudante são determinantes para a previsão pretendida, o que confirma resultados já anteriormente demonstrados por [12]. De salientar que dos 11 atributos que se revelaram significativos para a previsão, apenas um deles, o nível educacional da mãe, não pertence à categoria dos curriculares. Esta relação foi particularmente defendida por [13], que afirmou que são as mães quem mais influenciam o sucesso académico dos discentes.

Através do conhecimento obtido é viável identificar grupos de estudantes de maior risco, o que permitirá a delineação de políticas promotoras de sucesso escolar. Através do conhecimento assim obtido, os gestores institucionais poderão proceder à definição de estratégias educacionais e tutoriais em prol da eficácia e da eficiência educativa. Também o tipo de abordagem adotada na identificação das características do estudante que melhor explicam o seu sucesso parece distanciar-se um pouco daquela que usualmente tem sido seguida em trabalhos relacionados com a mesma temática. No caso do presente trabalho, a seleção dessas caraterísticas processou-se em duas fases distintas. Primeiro, a seleção das dimensões do aluno que melhor explicam o seu sucesso permitiu um primeiro ajuste do modelo ao se eliminarem grupos completos de variáveis. Depois, num ajuste mais fino, procedeu-se à seleção dos atributos entre os que não foram excluídos nessa primeira fase. Com esta abordagem foi então possível, logo numa primeira fase, reduzir a "praga" da dimensionalidade dos dados sem se ter perdido a capacidade preditiva do modelo.

Para complementar a investigação descrita no presente estudo, poder-se-á apontar, como linhas de orientação futura, o desenvolvimento de outras análises ainda mais minuciosas, ao nível de cada uma das escolas e de cada um dos cursos , de forma a delinear estratégias de promoção educacional individualizadas em função das necessidades dos estudantes.

Embora a metodologia apresentada possa ser replicada no âmbito de outras instituições de ensino superior, uma parte importante dos resultados que se obtiverem neste estudo não é generalizável ao contexto geral do ensino superior, uma vez que teve por base uma amostra de dados não representativos desse contexto mais alargado. Tendo-se utilizado como caso de estudo o IPB, uma instituição do subsistema de ensino superior politécnico, localizada numa região interior de baixa densidade populacional, a mesma não consegue captar a mesma heterogeneidade de alunos das instituições de grandes centros urbanos do litoral. Quanto muito, os resultados apresentados poderão espelhar realidades de instituições de ensino superior que reúnam condições similares às do IPB, como serão o caso de outros Institutos Politécnicos do interior do país, localizados longe dos grandes centros urbanos.

AGRADECIMENTOS

Este trabalho foi suportado pela Fundação para a Ciência e Tecnologia (FCT) através do Projeto UID/EEA/04131/2013. 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 BIBLIOGRÁFICA

[1] L. Breiman, Random forests, Machine learning, vol. 45(1), 2001, pp. 5–32.

[2] R. S. J. D. Baker and K. Yacef, The state of educational data mining in 2009: A review and future visions, JEDM-Journal of Educational Data Mining, vol. 1(1), 2009, pp. 3–17.

[3] C. Romero and S. Ventura, Educational data mining: A survey from 1995 to 2005, Expert systems with applications, vol. 33(1), 2007, pp. 135–146.

[4] R. A. Huebner, A survey of educational data-mining research, Research in higher education journal, vol. 19, 2013.

[5] C. Romero, S. Ventura, M. Pechenizkiy, and R. S. J. D. Baker, Handbook of educational data mining, CRC Press, 2010.

[6] S. Natek and M. Zwilling, Student data mining solution-knowledge management system related to higher education institutions, Expert systems with applications, vol. 41(14), 2014, pp. 6400–6407.

[7] O. C. Santos and J. G. Boticario, User-centred design and educational data mining support during the recommendations elicitation process in social online learning environments, Expert Systems, vol. 32(2), 2015, pp. 293–311.

[8] E. A. Amrieh, T. Hamtini, and I. Aljarah, Mining educational data to predict student's academic performance using ensemble methods, International Journal of Database Theory and Application, vol. 9(8), 2016, pp. 119–136.

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

[10] Z. Papamitsiou, E. Karapistoli, and A. A. Economides, Applying classification techniques on temporal trace data for shaping student behavior models, in Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, ACM, 2016, pp. 299–303.

[11] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, Analyzing undergraduate students' performance using educational data mining, Computers & Education, vol. 113, 2017, pp. 177-194.

[12] L. Manhães, Predicting Academic Performance of Undergraduate Students Using Educational Data Mining (*Predição do Desempenho Académico de Graduandos Utilizando Mineração de Dados Educacionais*), PhD thesis (*Tese Doutorado*), Universidade Federal do Rio de Janeiro, 2015.

[13] A. Agus and Z. K. M. Makhbul, An empirical study on academic achievement of business students in pursuing higher education: An emphasis on the influence of family backgrounds, in International Conference on the Challenges of Learning and Teaching in a Brave New World, Hatyai, Thailand, 2002, p. 168.