UNIVERSIDADE DA BEIRA INTERIOR
Engenharia

Desenvolvimento de Modelos Analíticos de Apoio à Gestão em Instituições do Ensino Superior, com Recurso a Data Mining

Maria Prudência Gonçalves Martins

Tese para obtenção do Grau de Doutor em
Engenharia e Gestão Industrial
(3º ciclo de estudos)

Orientadores:
Prof. Doutora Vera Lúcia Miguéis Oliveira e Silva
Prof. Doutor Davide Sérgio Baptista da Fonseca

Covilhã, fevereiro de 2020

Dedicatória

À Isablinha e à Luisinha, as duas meninas dos meus olhos.

Ao Paulo, meu marido e meu porto seguro.

Agradecimentos

À Professora Doutora Vera Lúcia Miguéis Oliveira e Silva, endereço um agradecimento muito grande pela orientação científica, pela disponibilidade, pelo empenho e minucioso acompanhamento nos trabalhos de doutoramento. Agradeço-lhe também a simpatia, a sempre pronta cooperação e, especialmente, a forma tão agradável como sempre o fez.

Ao Professor Doutor Davide Sérgio Batista da Fonseca expresso uma profunda gratidão pelas orientações prestadas, pela disponibilidade, pela simpatia e por todas as diligências que se revelaram cruciais para o sucesso deste trabalho.

Ao Professor Doutor Albano Alves, Pró-Presidente do Instituto Politécnico de Bragança para os Sistemas de Informação e responsável pelo respetivo Centro de Desenvolvimento e Gestão de Dados, agradeço a cedência do conjunto de dados imprescindível à realização da investigação inerente a esta tese.

À minha família, sentido maior da minha existência, agradeço o melhor dos meus dias. Aos meus esmerados pais, Manuel Maria e Querubina, de quem muito me orgulho, um eterno obrigado pela educação, pelo amor e apoio incondicional que sempre me dispensaram. Ao Paulo, meu grande companheiro de vida, agradeço o amor, a dedicação e o suporte emocional transmitidos ao longo da nossa jornada de quase 3 décadas em comum. Agradeço-lhe também a preciosa ajuda na informática. Às minhas filhas, Isabel e Luísa, agradeço as melhores experiências de vida, os maiores motivos para sorrir e o ânimo que eu necessitei para alcançar esta meta.

Aos colegas de trabalho, muito obrigada pelas palavras de incentivo e em particular ao Pedro João Rodrigues e ao João Paulo Teixeira pelas impressões trocadas sobre as técnicas usadas na presente investigação. A todos os amigos um bem-haja pelos excelentes momentos de convívio e, especialmente, pela amizade e presença nos momentos mais difíceis.

Resumo

As instituições de ensino superior deparam-se atualmente com grandes desafios, derivados da concorrência na captação de novos alunos, da globalização da educação e das diretrizes das recentes políticas educativas, quer nacionais quer da união europeia, que exigem intervenções acrescidas em prol do sucesso educativo e da prevenção do abandono académico. Com o objetivo de contribuir para que essas instituições de ensino se transformem em organizações mais pró-ativas, capazes de enfrentar os atuais desafios, propõem-se, nesta tese, novos modelos analíticos de previsão, desenvolvidos com recurso a técnicas de *data mining*, que permitem estimar, com a devida antecedência, quer o nível de sucesso esperado no final do curso, quer a propensão do aluno para o abandono. Esses modelos vão permitir identificar quer os grupos de estudantes de maior risco, que venham a necessitar de uma maior atenção, quer os fatores que mais contribuam para o seu (in)sucesso, conhecimentos de importância primordial para que os agentes de gestão possam adotar as medidas e decisões estratégicas de promoção de sucesso académico mais adequadas.

Para prever o sucesso académico global do aluno é proposto um modelo de regressão baseado no algoritmo *random forest*. Para estimar a sua propensão para o abandono é desenvolvido um modelo de classificação que combina três das mais importantes técnicas de *data mining*, como são os casos dos algoritmos *random forest*, máquinas de vetores de suporte e redes neuronais artificiais. Com o objetivo de avaliar e garantir que as metodologias desenvolvidas e os modelos propostos possam ser utilizados em contextos reais, usam-se, como caso de estudo, os alunos de licenciatura duma instituição pública do ensino superior politécnico.

Para além da elevada capacidade de previsão evidenciada pelos modelos desenvolvidos e da própria dimensão e diversidade dos dados analisados, destacam-se, como contribuições diferenciadoras desta tese, os processos de seleção dos fatores explicativos do sucesso e do abandono académico. A tese também demonstra o potencial das técnicas de *data mining* quando aplicadas a bases de dados de grande dimensão provenientes de ambientes educacionais, podendo a abordagem metodológica seguida servir de guia a outras instituições de ensino, ajudando-as a perceber de que forma o *data mining* as poderá auxiliar na extração de conhecimento útil que suporte melhores decisões.

Palavras-chave

Aprendizagem automática, descoberta de conhecimento em bases de dados, *data mining*, *educational data mining*, previsão de sucesso/abandono académico, *random forest*, redes neuronais artificiais, máquinas de vetores de suporte.

Abstract

Higher education institutions are currently facing deep challenges stemming from the competition for attracting new students, the globalisation of education, and the directives from recent educational policies adopted both nationally and by the European Union, which require more interventions aiming the attainment of educational success and the prevention of students dropout. With the aim to contribute to the transformation of such institutions into more proactive organisations, capable of facing up to the current challenges, this thesis puts forward new analytical predictive models, developed by means of data mining techniques, which enable the early estimation of a student's expected level of success at the end of the degree course as well as their propensity to drop out. These models will enable the identification of major risk groups of students, who will need closer attention, and also the identification of the factors mostly contributing to their success or failure. The importance of such information is vital to enable decision-makers to take the most adequate measures and decisions in order to promote academic success.

In order to predict students' global academic success, we propose a regression model based on random forest algorithm. For the estimation of students' propensity to drop out, we developed a classification model which combines three of the most popular data mining techniques, namely random forest, support vector machines and artificial neural networks. Aiming to assess and ensure that the methodologies developed and the models proposed can be used in real contexts, the undergraduates of a public polytechnic higher education institution were used as a case study.

Besides the high predictive capacity demonstrated by the models developed and the dimension and diversity of the data analysed, other noteworthy differentiating contributions of this thesis are the innovative process of selection of the explanatory factors for academic success and students dropout. The thesis also shows the potential of data mining techniques when applied to large scale datasets deriving from educational environments, and the approach followed in this study may be used as a guideline to other educational institutions on how data mining can support the extraction of useful knowledge with a view to support better decisions.

Keywords

Machine learning, knowledge discovery in databases, data mining, educational data mining, prediction of academic success/dropout, random forest, artificial neural networks, support vector machines.

Índice

Lista de Figuras

Lista de Tabelas

Siglas e Acrónimos

A3ES	Agência de Avaliação e Acreditação do Ensino Superior
AEHS	*Adaptive Educational Hypermedia System*
AI	*Artificial Intelligence* (m.q. IA)
ANN	*Artificial Neural Networks* (m.q. RNA)
AODE	*Averaged One-Dependence Estimators*
APA	Ambiente Pessoal de Aprendizagem
AUC	Area Under ROC Curve
BD	Base de Dados
BKT	*Bayesian KT*
CART	*Classification And Regression Tree*
CE	Comissão Europeia
CET	Curso de Especialização Tecnológica
CFS	*Correlation based Feature Selection*
CGPA	*Cumulative Grade Point Average*
CNE	Conselho Nacional de Educação
CRISP-DM	*Cross Industry Standard Process for Data Mining*
CTeSP	Curso Técnico Superior Profissional
CURE	*Clustering Using REpresentatives*
DCBD	Descoberta de Conhecimento em Base de Dados
DGEEC	Direção Geral das Estatísticas da Educação e Ciência
DM	*Data Mining*
DT	*Decision Tree* (árvore de decisão)
ECTS	*European Credit Transfer System* (sistema europeu de acumulação e transferência de créditos)
EDM	*Educational Data Mining*
EM	*Expectation Maximization* (maximização de expectativas)
ESA	Escola Superior Agrária de Bragança
EsACT	Escola Superior de Comunicação, Administração e Turismo de Mirandela
ESE	Escola Superior de Educação de Bragança
ESSa	Escola Superior de Saúde de Bragança
ESTiG	Escola Superior de Tecnologia e Gestão de Bragança
FM	*Factorization Machines*
GMM	*Gaussian Mixture Model*
IA	Inteligência Artificial
IBk	*Instance-Bases learning with parameter k*
ICA	*Individual Component Analysis*
ID3	*Iterative Dichotomiser 3*

IDE	*Integrated Development Environment* (ambiente integrado de desenvolvimento)
IES	Instituição de Ensino Superior
IPB	Instituto Politécnico de Bragança
ITS	*Intelligent Tutoring System*
KDD	*Knowledge Discovery in Databases* (m.q. DCBD)
KEEL	*Knowledge Extraction based on Evolutionary Learning*
KNIME	*Konstanz Information Miner*
K-NN	*K-Nearest Neighbor* (k-vizinhos mais próximos)
LA	*Learning Analytics* (aprendizagem analítica)
LMS	*Learning Management System*
MCTES	Ministério da Ciência Tecnologia e Ensino Superior
MF	*Matrix Factorization* (fatorização de matrizes)
MLP	*Multi-Layer Perceptron*
MOOC	*Massive Open Online Course*
MSE	*Mean Square Error* (erro quadrático médio)
MVS	Máquinas de Vetores de Suporte
NB	*Naive Bayes*
NBT	NB *Tree*
OCDE	Organização para a Cooperação e Desenvolvimento Económico
PLMR	*Personalized Linear Multiple Regression* (regressão múltipla linear personalizada)
R	linguagem e IDE para cálculo estatístico e aprestações gráficas
RF	*Random Forest* (floresta aleatória)
RL	Regressão Logística
RMSE	*Root MSE* (raiz do erro quadrático médio)
RNA	Redes Neuronais Artificiais
ROA	Repositório de Objetos de Aprendizagem
ROC	*Receiver Operating Characteristic curve*
SGD	*Stochastic Gradient Descent*
SPSS	*Statistical Package for the Social Sciences*
SRP	Sistemas de Recomendações Pedagógicas
SVM	*Sport Vector Machines* (m.q. MVS)
TIC	*Tecnologias de Informação e Comunicação*
UBI	Universidade da Beira Interior
UC	Unidade Curricular
WEKA	*Waikato Environment for Knowledge Analysis*
WM	*Web Mining*

Capítulo 1

Introdução

1.1 Enquadramento

"Onde quer que se descuide a educação, o Estado sofre um golpe nocivo."

Aristóteles, filósofo grego, 384 – 322 a.C.

É indiscutível que uma formação académica de qualidade tem uma importância primordial no progresso científico, tecnológico, económico, cultural e social de qualquer país ou comunidade. Estudos diversos têm concluído que a produtividade, a competitividade e o progresso tecnológico estão estritamente correlacionados com o nível educacional dos cidadãos. De igual forma, o mesmo nível educacional tem vindo a ser identificado, ao longo dos tempos, como um fator determinante para o combate às desigualdades sociais, ao desemprego e aos baixos salários. Com efeito, Portugal é um dos países da Organização para a Cooperação e Desenvolvimento Económico (OCDE) onde o investimento na formação académica mais retorno garante. De acordo com o *Education At a Glance* de 2017 [39], o relatório anual mais abrangente sobre os sistemas educativos dos 34 países da OCDE, os diplomados portugueses auferem, em média, mais 69% do que os cidadãos que têm apenas o ensino secundário, quando a média comunitária e da OCDE não vão, respetivamente, além dos 53% e 54%.

Dada a relevância do sucesso educacional para o progresso das sociedades e para o bem-estar e prosperidade dos cidadãos, a promoção do desempenho académico dos estudantes e a diminuição dos índices de abandono no ensino superior, têm-se salientado como assuntos prioritários no âmbito das atuais políticas educativas, tanto no contexto nacional como a nível internacional. A Comissão Europeia (CE), com o intuito de estimular o interesse público pelo sucesso escolar, de responsabilizar e de escrutinar as Instituições de Ensino Superior (IES), elaborou um conjunto de recomendações [116] com o objetivo de sensibilizar a tutela e as IES dos países membros para a urgência de intervenções acrescidas em prol do sucesso educativo. Também em Portugal é comprovada a relevância da questão, através da resolução n.º 176/2017 da Assembleia da Republica [89], onde é recomendado ao XXI Governo Constitucional a adoção de um conjunto de medidas para combater o insucesso e o abandono escolar no ensino superior. Mais precisamente, com o objetivo de se "realizarem análises, inferirem variáveis explicativas e se estabelecerem comparações intra e inter-instituições", foi sugerido ao governo que, nos termos do n.º 5 do artigo 166.º da Constituição, "apresente anualmente um relatório sobre o abandono escolar no ensino superior, no cumprimento da Resolução da Assembleia da República n.º 60/2013, de 18 de abril".

Em junho de 2017, num estudo da Direção-Geral de Estatísticas da Educação e Ciência [12], é enfatizado que há ainda muito por fazer, para além do que já se fez no seio das IES nacionais, no que diz respeito à efetiva implementação das mencionadas diretrizes europeias e nacionais

em prol do sucesso escolar dos estudantes do ensino superior. Havendo ainda um longo caminho a percorrer para que o cumprimento da legislação a este respeito se concretize, e pelo facto da promoção do sucesso académico ser um problema cuja resolução é de elevada complexidade, os especialistas da DGEEC, no mesmo documento, também remetem para a conveniência de estabelecer padrões concetuais de análise e para a premência de desenvolver e aprofundar diagnósticos subordinados à prevenção do insucesso e do abandono académico.

O cumprimento das diretrizes das atuais políticas educativas, tanto de âmbito nacional como europeu, exige à gestão das IES uma resposta imediata, oportuna e eficaz. A persecução desse objetivo será facilitado com o desenvolvimento de novos modelos analíticos de apoio à Gestão, que possam contribuir para transformar as IES em organizações mais pró-ativas, capazes de antever oportunidades de melhoria com vista à otimização do serviço prestado aos seus estudantes.

É neste contexto que a atual investigação de doutoramento se enquadra, ao pretender desenvolver ferramentas de apoio à decisão nas IES, que permitam melhorar os seus processos e garantir que os seus recursos, humanos e materiais, possam ser geridos de uma forma mais eficiente em prol da melhoria do desempenho académico dos estudantes do ensino superior.

Com o objetivo de avaliar e garantir que as metodologias desenvolvidas e os modelos propostos possam ser utilizados em contextos reais, na investigação a desenvolver utilizar-se-à a informação reunida nas bases de dados dos serviços académicos do Instituto Politécnico de Bragança (IPB), uma instituição pública do ensino superior politécnico do nordeste de Portugal.

Uma vez que os modelos analíticos a desenvolver terão por base conjuntos de dados diversificados de grande dimensão e dimensionalidade, a investigação enquadrar-se-à num paradigma de investigação quantitativa com recurso a técnicas analíticas de grande alcance como as de *data mining*. A opção por esta abordagem justifica-se pelo facto de se tratar de instrumentos de análise e investigação de eficácia comprovada, aos quais está associado um vasto conjunto de potentes técnicas e algoritmos de análise automatizada, alicerçados nos tradicionais métodos quantitativos de análise de dados, oriundos da área disciplinar da estatística, da investigação operacional e otimização e da modelação matemática.

Pretende-se, por esta via, providenciar suporte informativo realista, consistente e de grande precisão, que possa contribuir para aumentar a eficiência e a eficácia organizacional das IES, designadamente, ao criar uma população estudantil mais instruída, com as qualificações devidas para corresponder às atuais necessidades do meio laboral, cada vez mais complexo, mais competitivo e em constante mutação.

O interesse em desenvolver estudos de previsão e explicação do desempenho académico aumentou consideravelmente no seio das IES, não só pela necessidade de cumprirem as diretrizes das atuais políticas educativas, mas também, pelo facto da melhoria dos resultados escolares dos estudantes se traduzir numa contribuição valiosa para a promoção do prestígio institucional, um fator determinante para a cativação e fixação de novos alunos, e, por conseguinte, para o aumento das receitas da instituição, quer as provenientes do financiamento do governo quer as que resultam diretamente das propinas dos estudantes.

As principais motivações e os objetivos gerais da investigação a desenvolver nesta tese de doutoramento são apresentados na secção que se segue.

1.2 Motivação e objetivos gerais

Promover a qualidade dos serviços prestados aos estudantes, por via do desenvolvimento de estudos de diagnóstico que contribuam para a definição de estratégias destinadas à melhoria do desempenho académico dos estudantes, são tarefas imprescindíveis no contexto atual das IES.
A previsão precoce e precisa do desempenho de aprendizagem dos estudantes, tal como a identificação em tempo útil dos estudantes propensos a um baixo desempenho académico e o reconhecimento dos fatores mais capazes de discriminar os alunos de sucesso dos de insucesso, poder-se-à revelar conhecimento de importância primordial para a delineação de intervenções de carácter preventivo eficazes.
Por esse motivo, o sucesso e o abandono no ensino superior português têm sido alvo de estudos e reflexões, tanto por parte dos órgãos governamentais e dos administradores das IES, como dos investigadores das Ciências Sociais e Humanas e especificamente das Ciências da Educação (Ferreira and Fernandes [37]. Embora este género de estudos tenham sido intensificados ao longo dos últimos anos, verifica-se que incidiram, sobretudo, em análises de cariz mais sociológico ou psicológico (Ferreira et al. [38]), como é exemplo o estudo de Évora [34], que perscrutou somente a influencia dos fatores do contexto familiar dos estudantes no desempenho académico, ou o estudo de Araújo [5], que investigou o abandono no ensino superior considerando apenas fatores da dimensão sociológica do estudante, ou o estudo de Miguel et al. [72], que investigou somente a relevância das variáveis psicológicas e comportamentais do aluno (e.g. motivacionais, ansiedade) como fatores de risco para o insucesso. Sendo inegável que o desempenho académico resulta da influência de múltiplos fatores de natureza diferente, que coexistem em simultâneo e interagem entre si, pelo facto de não terem sido investigados de forma conjunta não fica devidamente esclarecido qual é a importância diferenciada das múltiplas dimensões em interação e indissociáveis do estudante.
Para além disso, no contexto atual do ensino superior português, os estudos existentes neste âmbito (ver, por exemplo, Araújo [5], Almeida [2], Ferreira and Fernandes [37], Guimarães [44], Costa et al. [24], Miguel et al. [72], Évora [34], Ferreira et al. [38]) enquadram-se, maioritariamente, num paradigma de investigação de cariz qualitativo e construtivista, com recurso a entrevistas e/ou questionários como instrumentos de recolha de informação, mas são ainda muito escassos os estudos que usam métodos analíticos de maior alcance que suportem melhores decisões. Os estudos de cariz qualitativo e construtivista focam-se, essencialmente, na análise das perceções e dos motivos que levaram os estudantes inquiridos a abandonar a frequência dos curso onde ingressaram, ao invés de procurarem inferir quais são os fatores que prenunciam a decisão do estudante se desvincular da escola, ou de procurarem desenvolver modelos preditivos que providenciem conhecimento útil para a melhoria das decisões de gestão.
No que concerne à correspondente abordagem metodológica, os próprios autores desta índole de estudos também lhe sublinham algumas debilidades, designadamente:

- na maioria dos estudos sobre fatores de sucesso e de insucesso académico, é privilegiado um registo mais exploratório e descritivo, do que explicativo e propositivo (Costa et al. [24]);
- o tamanho reduzido das amostras em análise (Ferreira and Fernandes [37], Araújo [5]) não é suficientemente representativo para caracterizar devidamente a heterogeneidade de estudantes existentes no ensino superior;
- o tamanho reduzido das amostras em análise (Guimarães [44], Araújo [5]) e o fraco controlo das características pessoais dos indivíduos que constituem a amostra (Miguel et al. [72]),

como, por exemplo, a dificuldade de categorização social e a ausência de dados sobre o contexto escolar dos alunos (Araújo [5]), também leva a que a maioria das investigações não esclareça devidamente quais são os fatores mais capazes de discriminar os alunos de sucesso dos alunos em risco de fracasso, o que se configura como um mau prognóstico para a delineação de intervenções precoces (Belo [13]) que se possam revelar assertivas no combate ao insucesso;

- as pesquisas de cariz qualitativo têm naturalmente limitações na exploração e aprofundamento de variáveis mais específicas (Guimarães [44]), centrando-se mais sobre a compreensão e a interpretação dos factos do que em determinar as causas dos mesmos;

- o uso de métodos estatísticos exclusivamente correlacionais, que estas índoles de investigações apresentam, torna os seus resultados pouco robustos e pouco compreensivos do processo subjacente a este fenómeno (Miller et al. [75]).

As debilidades identificadas nos estudos existentes indiciam a necessidade de investigações de maior amplitude, que considerem conjuntos de dados de grande dimensão e dimensionalidade, dos quais decorra uma caracterização adequada quer da heterogeneidade de perfis estudantis existentes numa IES, quer das múltiplas dimensões indissociáveis dos estudantes que poderão ter influência no desempenho académico. Pretende-se, por essa via, desenvolver modelos analíticos realistas, consistentes, de elevada precisão e de fácil compreensão, que sejam capazes de capturar os padrões de sucesso e de insucesso da heterogeneidade de perfis estudantis existentes em toda a comunidade académica.

Uma grande parte da informação que permite caracterizar de uma forma bastante abrangente e precisa a heterogeneidade de perfis estudantis existentes numa comunidade académica do ensino superior, no que concerne às dimensões do contexto demográfico, socioeconómico e curricular de desempenho académico (pré e após ingresso no ensino superior), pode ser extraída do grande volume de dados digitais continuamente acumulados nas bases de dados dos serviços académicos das IES. A análise desses grandes conjuntos de dados digitais, designada análise de *big data*, é uma tendência de investigação crescente nas áreas de intervenção da engenharia e gestão industrial, como a gestão da qualidade, a gestão de operações e os sistemas de apoio à decisão. Efetivamente, se esses dados forem devidamente analisados, poderão providenciar a descoberta de conhecimento oculto e potencialmente útil, realista e de elevada precisão, que poderá melhorar os processos de decisão nas organizações. Para a identificação de padrões ocultos, de associações e relacionamentos desconhecidos entre as variáveis presentes no *big data*, emergiram há cerca de três décadas técnicas analíticas de grande alcance, denominadas técnicas de *data mining*. Trata-se de um paradigma de análise e investigação relativamente recente, ao qual está associado um vasto conjunto de poderosos métodos e algoritmos de análise automática com grande potencialidade. Por exemplo, têm a capacidade de processar, analisar e modelar grandes e complexos conjuntos de dados de natureza variada, estruturados ou não estruturados, procedentes de múltiplas fontes e em configurações heterogéneas.

Pelo facto de muitos dos algoritmos de *data mining* terem a sua génese na teoria da aprendizagem estatística, da investigação operacional e otimização, da modelação matemática e da inteligência artificial, 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 o processo de decisão nas organizações. Por esse motivo, nos últimos anos, sobretudo a partir de 2008, tem havido também um interesse crescente no uso do *data mining* em estudos relacionados com a melhoria dos resultados educacionais, a

qual é designada *educational data mining* (EDM) (Baker et al. [7]). De acordo com (Romero and Ventura [94]) o EDM apresenta-se como solução capaz de dar resposta à necessidade de sucesso académico dos estudantes e à melhoria contínua dos palcos onde a aprendizagem ocorre. Tal é a grande importância desta recente área de investigação que se estima que até ao ano de 2022 todos os estudos relacionadas com a melhoria dos resultados de aprendizagem envolverão o uso do *data mining* (Baker and Inventado [8]). Mas, apesar do promissor potencial da análise de *big data* apoiada pelo *data mining*, a maioria das IES não conseguiu analisar esses dados de forma a transformá-los em informações verdadeiramente valiosas (Miguéis et al. [71]). De facto, foi só na última década que o EDM começou a suscitar um indiscutível interesse junto da comunidade científica internacional, sendo ainda muito escassos os estudos existentes no contexto do ensino superior português. Portanto, a análise do *big data* reunido nas IES, por via destas novas técnicas analíticas de grande alcance, poder-se-à revelar uma vantagem de relevo para a melhoria da definição de estratégias de promoção educacional.

O IPB é particularmente atrativo para se desenvolver um estudo alicerçado nestas recentes metodologias de análise de dados, pelo facto de estar localizado numa região geográfica de baixa densidade populacional do Nordeste de Portugal, constrangimento ao qual se tem associado a dificuldade da fixação dos jovens estudantes na região.

Pelo facto de o EDM ser uma tendência de investigação emergente, considerado por alguns autores (e.g. Huebner [53], Kaur et al. [55]) estar ainda na fase de "infância", o estado da arte atual revela que ainda há muito a estudar e a considerar acerca das metodologias a aplicar nesta área de intervenção, da sua real capacidade de abrangência e também das possibilidades de aplicação (Romero and Ventura [94], Peña-Ayala [86], Sukhija et al. [110]]. Por conseguinte, existe uma necessidade crescente de estudos futuros relacionados, por forma a que o EDM se consolide como uma disciplina de investigação madura. Mais especificamente, a literatura revela a necessidade de aprofundamento dos estudos, a fim de se providenciar o desenvolvimento de modelos práticos e generalizáveis de aplicação em múltiplos contextos. De salientar, ainda, que a maioria dos estudos de *data mining* existentes no seio de instituições do ensino superior internacionais estão relacionados, predominantemente, com previsões de abandono numa ou noutra disciplina, frequentadas em contexto de aprendizagem do sistema de ensino à distância (*e-learning*). Esta constatação apela ao desenvolvimento de novos estudos no contexto das instituições de ensino presencial tradicional, cuja escassez poderá estar relacionada com a dificuldade de acesso aos dados, derivada, em parte, do cumprimento de políticas cada vez mais restritivas de proteção de dados.

As principais revisões de literatura no âmbito do EDM permitem, de facto, perceber a pertinência do trabalho de doutoramento que se propõe desenvolver. Na revisão dos 306 artigos científicos analisados por Romero and Ventura [94] no âmbito do EDM, apenas 36 se cingiam ao sistema de educação presencial tradicional. Este conjunto de 36 estudos estavam ainda subdivididos por 11 tópicos de investigação, pelo que apenas uma pequena minoria corresponde à previsão do desempenho académico em IES do sistema de ensino presencial tradicional. Também o estudo de revisão de literatura da autoria de Peña-Ayala [86] evidencia a predominância dos estudos em sistemas de ensino à distância baseados na *Web*, face aos estudos em sistemas de ensino presencial tradicional. Dos 222 estudos analisados neste contexto, apenas $20,7\%$ dizem respeito à modelação do desempenho dos estudantes, e são aplicados, predominantemente, em sistemas de aprendizagem baseados na *web* como o *Intelligent Tutoring System (ITS*) e *Learning Management System (LMS)*. O mesmo autor concluiu que 66% do conjunto de estudos orientados para a funcionalidade de modelação e avaliação de desempenho dos estudantes correspondem a abordagens ainda incipientes, facto que reforça ainda mais a necessidade de se virem a de-

senvolver estudos que sejam propícios a novos aprofundamentos analíticos. O trabalho que se pretende desenvolver tem então como objetivo central contribuir para o desenvolvimento desta área de investigação recente, denominada *Educational Data Mining*, através da proposta de novos modelos analíticos de apoio à Gestão de uma IES.

Na secção que se segue apresenta-se de forma mais minuciosa os objetivos específicos deste trabalho de doutoramento.

1.3 Objetivos específicos

Tal como descrito na secção anterior o objetivo central deste trabalho de doutoramento é o de explorar o uso de métodos analíticos na compreensão e (consequentemente) melhoramento dos ambientes de aprendizagem. Mais especificamente, visa usar o estado da arte das técnicas de *data mining* para prever o desempenho académico dos alunos e a formulação de recomendações para melhorar os níveis de desempenho académico dos mesmos. Dada a grande importância que a educação tem no desenvolvimento das sociedades, tem havido cada vez maior pressão para que os processos utilizados para garantir a qualidade do sistema educacional sejam revistos e reformulados. Consequentemente, tem havido um grande esforço e investimento das instituições de ensino na manutenção de sistemas de recolha e armazenamento de dados relacionados com o ensino. No entanto, as instituições deparam-se com a grande dificuldade que é tratar todos esses dados e transformá-los em informação verdadeiramente útil. A análise dos dados neste contexto é efetivamente promissora, pois permite às instituições descobrir conhecimento oculto de padrões dos alunos em ambientes educacionais. Neste enquadramento, o plano de trabalho de doutoramento abrange as seguintes atividades:

- Desenvolver uma revisão bibliográfica sobre os temas abordados no projeto de doutoramento. Esta revisão incidirá essencialmente no estado da arte das técnicas de *data mining*, sobretudo nas técnicas de previsão e nas teorias e métodos utilizados para temas abordados no projeto de doutoramento.

- Criação de modelos que consigam prever o sucesso académico dos alunos, com base em padrões de sucesso e de insucesso de todo o percurso académico. Os modelos devem explorar diferentes períodos temporais, tendo em vista avaliar o impacto que os eventos passados e os mais recentes têm no desempenho final dos alunos.

- Criação de modelos que consigam prever o abandono escolar e que identifiquem os estudantes propensos à evasão. No intuito de mitigar as perdas de receitas das instituições e de aumentar as qualificações dos estudantes, o conhecimento gerado deverá fundamentar intervenções atempadas com vista à diminuição dos índices de abandono.

- Com base nas conclusões que resultarem dos modelos preditivos, apontar sugestões de gestão, designadamente, ao nível da reafetação de recursos, do acompanhamento tutorial personalizado aos estudantes e demais práticas educativas orientadas para a maximização das taxas de graduação. Estas recomendações poderão vir a revelar-se o suporte necessário para que professores e gestores das instituições de ensino possam vir a reformular adequadamente os serviços educacionais prestados à comunidade estudantil.

1.4 Organização da tese

Para além deste capítulo introdutório, onde se enquadra o tema da investigação de doutoramento e se apresentam os objetivos globais e específicos da investigação, a tese encontra-se organizada da seguinte forma:

- No Capítulo 2 descreve-se, de modo sucinto, a génese de todo o processo de Descoberta de Conhecimento em Bases de Dados (DCBD). Com ênfase na fase do *data mining*, apresentam-se os principais métodos que lhe estão associados, realçando-se os mais recorrentes no âmbito do desenvolvimento de modelos de previsão para apoio à decisão em IES. A descrição efetuada providencia também alguns esclarecimentos que ajudarão a escolher os métodos mais adequados para os tópicos de investigação abordados.

- No Capítulo 3 concetualiza-se e caracteriza-se o *Educacional Data Mining*. Adicionalmente, é desenvolvida uma revisão de literatura através da qual se demonstra a importância crescente desta recente subárea de intervenção ao longo do tempo. São também referenciados e analisados os principais tópicos de investigação onde a mesma tem demonstrado um notável contributo, como instrumento de análise e apoio à gestão. Com ênfase na revisão dos estudos relacionados com a modelação de desempenho académico, são identificados os principais fatores explicativos do sucesso e do abandono, a fim de obter orientações para a investigação abordada nesta tese. A revisão a este género de estudos permitiu, igualmente, perceber quais os métodos e algoritmos de *data mining* mais usados neste tipo de modelação.

- No Capítulo 4 faz-se uma breve descrição do IPB, a IES usada como caso de estudo, caraterizando-se a sua oferta formativa, com especial destaque para os cursos conducentes ao grau de licenciatura, o ciclo de estudos abrangido pela presente investigação. Termina-se com a descrição das bases de dados disponibilizadas pelo IPB para objeto do atual estudo.

- No Capítulo 5 propõe-se um novo modelo analítico de regressão, desenvolvido com o intuito de prever de forma precoce o desempenho académico global dos estudantes de licenciatura do IPB, no *terminus* do seu percurso académico. Usando o algoritmo *random forest*, o modelo proposto identifica os principais fatores de sucesso e de insucesso dos estudantes, através dos quais a gestão institucional poderá desenhar ações de gestão que permitam mitigar o insucesso académico, promovendo uma melhor experiência educativa aos seus estudantes.

- No Capítulo 6 propõem-se dois novos modelos analíticos de classificação, que combinam 3 importantes técnicas de *data mining* (*random forest*, *support vector machines* e as redes neuronais artificiais) desenvolvidos com o objetivo de prever, de forma precoce, o abandono escolar dos discentes das licenciaturas do IPB. Esses modelos permitem, logo no final do 1º e 2º semestres, identificar os estudantes mais propensos à evasão académica, providenciando aos decisores institucionais informação oportuna e realista para poderem delinear, atempadamente, estratégias de combate à evasão.

- Por fim, apresentam-se no Capítulo 7 as principais conclusões e contribuições do trabalho desenvolvido e alguns apontamentos sobre as possíveis direções do trabalho a realizar futuramente, norteadas pela necessidade de melhoria contínua do desempenho das IES, tal como exigido pelas atuais políticas educativas, pela sociedade e pelo meio laboral.

Em apêndice pode ainda ser encontrada uma tabela com a síntese dos principais estudos subordinados ao estado da arte (Apêndice A), uma tabela com a síntese sobre a revisão de literatura relacionada com os fatores relevantes para a análise e previsão do desempenho académico com recurso ao EDM (apêndice B), a caracterização institucional do 1º Ciclo de Estudos lecionado no IPB (Apêndice C), a lista completa de tabelas e atributos presentes nas Bases de Dados disponibilizadas pelo IPB para objeto de estudo (Apêndice D) e os dados das simulações realizadas no âmbito da aplicação do método *forward search* (Apêndice E) aos dois modelos de previsão de abandono.

Capítulo 2

Data mining

2.1 Introdução

Torna-se cada vez mais urgente a necessidade de criação de uma nova geração de teorias computacionais e ferramentas, para auxiliar os agentes de decisão na extração de informação útil, a partir dos enormes volumes de dados digitais com que nos deparamos atualmente (Fayyad et al. [35]). Este capítulo, concetualiza todo o processo de Descoberta de Conhecimento em Bases de Dados (DCBD), com ênfase no *data mining*, dado ser considerada a principal e mais importante fase de todo o processo. O presente capítulo está organizado como se segue. Inicialmente, na Secção 2.2, descreve-se de forma sucinta o processo DCBD. Na Secção 2.3 concetualiza-se e apresenta-se uma revisão sobre o *data mining*. Na Secção 2.4 enunciam-se e descrevem-se os principais métodos que asseguram uma das tarefas de maior relevo de *data mining* e que é objeto de investigação nesta tese: a previsão. Conclui-se a secção com as principais técnicas de *data mining* que suportam os referidos métodos. Na Secção 2.5, apresentam-se as métricas de avaliação de desempenho preditivo mais populares em DCBD. Por fim, na Secção 2.6 apresenta-se um breve resumo e as principais conclusões deste capítulo.

2.2 Descoberta de conhecimento em base de dados

A origem e a permanente evolução dos sistemas informáticos propiciou a que as organizações procedessem à recolha, processamento, depósito e ordenação de toda a informação oriunda das suas atividades diárias, de um modo cómodo e rápido e a um custo acessível. Muitos autores de estudos de investigação (e.g. Lyman et al. [63], Witten et al. [117]) têm estimado a quantidade de dados gerados, armazenados e consumidos no mundo e, embora as estimativas sobre esses números possam variar, todas apontam para um aumento substancial a cada instante que passa. Essa grande e diversificada volumetria de dados gerados, que podem ser, estáticos ou dinâmicos, estruturados ou não estruturados, procedentes de múltiplas fontes e disponíveis em configurações heterogéneas (PhridviRaj and GuruRao [87]), traduziu-se num processo lento e de difícil agregação e análise (Oliveira [80]). Em simultâneo, conjuntamente com toda esta complexidade, derivada da dimensionalidade e crescimento massivo dos dados, persiste também a noção de que é ínfima a proporção de dados que são analisados, em comparação com os que são gerados, o que permite antever que muita informação, potencialmente útil, presente nesses grandes volumes de dados, não é devidamente aproveitada.

Numa economia global, caracterizada por uma elevada concorrência, as organizações que pretendam ser competitivas e diferenciadoras não podem prescindir da delineação de estratégias que lhe permitam extrair conhecimento crucial, atempado e eficiente para a sua gestão. Na

verdade, a criação dessas bases de dados substancialmente grandes só é pertinente se existir alguma forma eficaz para efetuar análises oportunas, inteligentes, exequíveis e fiáveis sobre os dados. Foi devido à necessidade premente de conceber e desenvolver potentes metodologias de extração de conhecimento útil, de grandes volumes de dados, que o processo de Descoberta de Conhecimento em Base de Dados (DCBD) emergiu. Este processo, que na linguagem anglo-saxónica é denominado de *Knowledge Discovery in Databases (KDD)*, designa uma área de conhecimento que começou a ser divulgada em 1989, no primeiro workshop relacionado com esta temática. Fayyad, Piatetsky-Shapiro, and Smyth [35], considerados os percursores e principais divulgadores desta área de investigação, apresentaram um estudo neste contexto onde enfatizaram que "a DCBD é o processo iterativo e interativo não trivial, de identificação válida, original, potencialmente útil, de padrões compreensíveis nos dados". Este processo iterativo e interativo, que considera uma ampla e muito complexa área de abordagem e análise, caracteriza-se, sobretudo, por uma grande flexibilidade e permeabilidade e ainda por uma grande capacidade de relacionamento interdisciplinar. É devido à sua flexibilidade que esta metodologia, de extração de conhecimento de grandes volumes de dados, tem sido aplicada a diversos sectores da economia, como a indústria, o comércio, as telecomunicações, a banca, a saúde, o desporto e mais recentemente também na área da educação. É também devido à sua grande capacidade de relacionamento interdisciplinar que o processo DCBD tem vindo a evoluir, a partir da conjunção de múltiplas áreas de pesquisa, como a estatística, a investigação operacional e optimização, a modelação matemática, a aprendizagem automática, a inteligência artificial, o reconhecimento de padrões, os sistemas de gestão de bases de dados, a visualização de dados e a computação (Cabena et al. [17], Turban et al. [113]). O objetivo unificador é extrair conhecimento de dados de alto nível a partir de dados de baixo nível no contexto de grandes conjuntos de dados (Fayyad et al. [35]).
A caracterização da DCBD como um processo complexo é evidenciada pelo ciclo de vida que empreende e que se perfaz numa ordem tendencialmente sequencial de fases e tarefas associadas entre si. As diferentes fases que integram este ciclo de vida podem ser observadas na Figura 2.1.

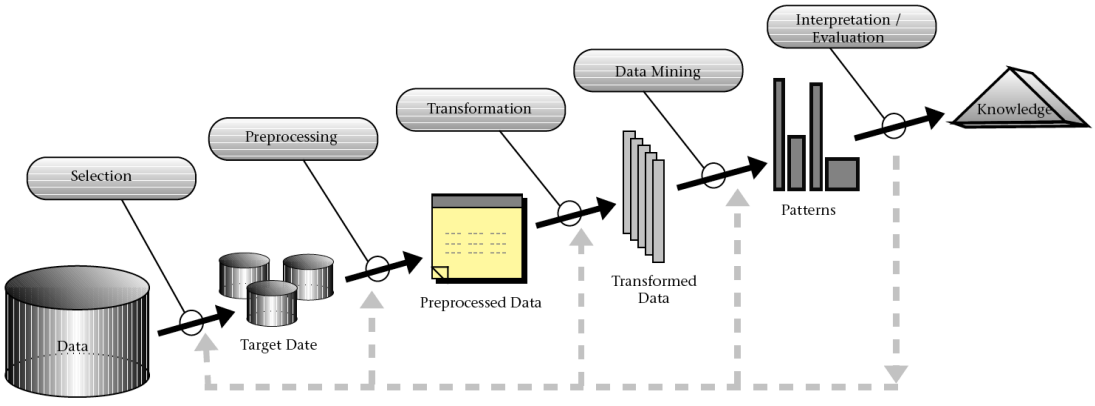


Figura 2.1: Sequência de fases que compõem todo o processo de DCBD (Fayyad et al. [35]).

As fases inerentes ao processo DCBD, com início na seleção dos dados de interesse e que culmina na extração de conhecimento, resumem-se de acordo com o que se segue (Fayyad et al. [35], Frawley et al. [40]):

- Seleção e caracterização dos dados: esta fase compreende a contextualização do projeto nas operações da organização, a compreensão da linguagem de negócio e a definição dos objetivos do projeto, de forma a selecionar os dados corretos a analisar, os atributos mais relevantes e o período de tempo mais apropriado. É nesta fase que se gera conhecimento de domínio que será utilizado durante todo o processo.

- Pré-processamento dos dados: usualmente, na fase que precede a transformação dos dados, é necessário conhecer, explorar e preparar adequadamente os dados. Este procedimento envolve a remoção do ruído, ou dos *outliers*, presentes nos dados, a recolha da informação necessária para modelar o ruído e a decisão sobre estratégias para lidar com valores em falta.

- Transformação dos dados: é no decorrer desta fase que se empreendem tarefas importantes e que virão a determinar a fiabilidade dos resultados obtidos. Após a filtragem dos dados, de acordo com o descrito na fase anterior, procede-se ao processamento dos dados no formato adequado para se poder proceder à aplicação de algoritmos de *data mining*. As transformações mais comuns são a normalização, a agregação e a discretização. Tendo em consideração os objetivos a que se destinam, as transformações compreendem a procura de características úteis nos dados, a utilização de métodos de transformação com vista à redução do número efetivo de variáveis em consideração e a procura de representações invariantes para os dados. É também no decorrer desta fase que se delineia o objetivo da prospeção de dados (previsão, descrição), tendo sempre em consideração o processo global de DCBD.

- *Data mining*: após uma preparação prévia dos dados, de acordo com as etapas anteriores, procede-se à seleção e aplicação de poderosos métodos e algoritmos de análise automatizada que sirvam o propósito da prospeção — a descoberta de padrões e relacionamentos que existam nos dados. Esta fase é realizada repetidamente e de forma iterativa e interativa.

- Interpretação e avaliação dos padrões: a derradeira fase do processo compreende a interpretação dos padrões descobertos e a avaliação da sua utilidade para a aplicação pretendida. É nesta fase que se consolida, ou não, o conhecimento descoberto e que se avalia a necessidade de novas iterações e de regressar a alguma(s) das fases anteriores para melhor interação ou documentação.

Tal como se observa na Figura 2.1 e de uma análise ao anteriormente mencionado, no âmbito da DCBD o conhecimento é o produto final de uma descoberta baseada em dados digitais. As fases que constituem todo este processo não obedecem, forçosamente, a um fluxo unidirecional, uma vez que não é estabelecida qualquer sequência rígida de ocorrência entre elas. Embora se possam perfazer de uma forma tendencialmente sequencial, pode-se voltar atrás, em sucessivas iterações, tudo depende dos resultados e do desempenho das outras fases e das tarefas que lhe estão associadas. Por conseguinte, não é possível saber *a priori* qual o número de iterações e a quantidade de ciclos necessários até à obtenção de conhecimento. O investigador, intimamente envolvido em todas as fases, avalia, de forma cuidada e com precisão, a necessidade de retornar a alguma das fases anteriores, para aprimorar os resultados obtidos, o que evidencia a iteratividade e interatividade do processo. O investigador tem também a função de delinear, com clareza, quais são os objetivos e as fronteiras inerentes à DCBD e quais são os métodos e algoritmos que se perspetivam como os mais adequados para os concretizar. É por via de

todo este processo, onde o investigador tem uma função crucial, que ocorre a transformação de dados brutos em conhecimento útil e compreensível, que poderá auxiliar e fundamentar a tomada de decisões, baseada em informação. Esta informação, escusa e presente em bases de dados de grande volumetria, é recuperada através da seleção e uso de métodos e algoritmos computacionais inerentes à fase do *data mining* (Fayyad et al. [36]). O conceito de *data mining* desenvolve-se com maior pormenor na secção que se segue.

2.3 O *data mining*

2.3.1 Concetualização do *data mining*

Como se expôs na secção anterior, o *data mining* é uma fase inerente ao processo DCBD, sendo considerada por diversos autores (e.g. Frawley et al. [40], Fayyad et al. [35], Han et al. [46], Witten et al. [117]) como a mais importante de todo o processo. Tal como o próprio processo DCBD, o *data mining* é também manifestamente uma área interdisciplinar, como se depreende do esquema da Figura 2.2.

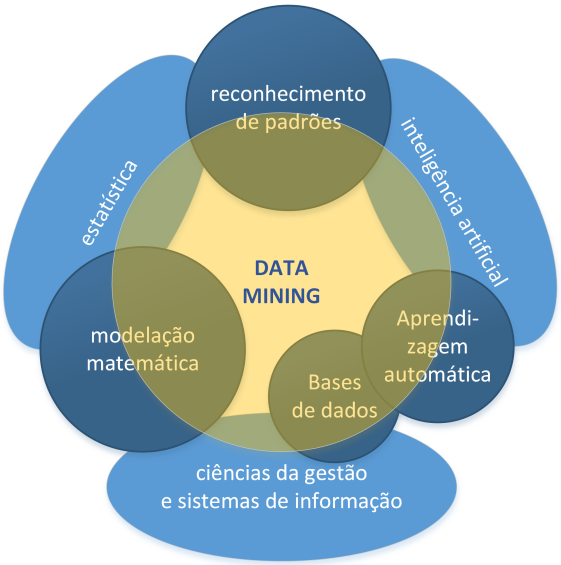


Figura 2.2: Interdisciplinariadade do *data mining* (adaptada de Turban et al. [113]).

Por ser considerada uma metodologia de conhecimento interdisciplinar, na literatura encontram-se algumas formas diferentes de a definir que, embora sejam, no essencial, convergentes entre si, variam consoante a área científica de proveniência dos autores. Por exemplo, os autores Fayyad et al. [35] e Štambuk and Konjevoda [108] dão ênfase à componente computacional, defendendo que o *data mining* é uma etapa no processo global de descoberta de conhecimento, em que se extrai, através da aplicação de algoritmos de aprendizagem automática e sob certas limitações computacionais, um conjunto de padrões e relacionamentos dos dados, que permitem revelar, de forma automática ou semi-automática, informação implícita que esteja presente em grandes base de dados. Em Hand et al. [47], a definição é apresentada sob uma perspetiva da área da estatística: o *data mining* é a análise de grandes conjuntos de dados a fim de encontrar relacionamentos inesperados e de resumir os dados de uma forma que eles sejam tão úteis quanto compreensíveis. Por sua vez, Revels and Nussbaumer [90] afirmam que o *data mining* utiliza sofisticados algoritmos matemáticos para analisar, automaticamente e sistema-

ticamente, uma grande quantidade de dados de forma a encontrar relacionamentos e avaliar a probabilidade de eventos futuros. Em Cabena et al. [17] a definição é dada sob a perspetiva da área das bases de dados: o *data mining* é um campo interdisciplinar que junta técnicas de aprendizagem automática, reconhecimento de padrões, estatística, bases de dados e visualização, no intuito de conseguir extrair informação útil de grandes bases de dados. Em Witten et al. [117] é mencionado que o *data mining* representa uma abordagem para extrair conhecimento escuso de informações contidas em base de dados e que faz parte de um processo mais amplo, a DCBD. Para Algarni [1] o *data mining* é uma poderosa ferramenta de inteligência artificial, com capacidade para descobrir informações úteis, através da análise de dados de muitas proveniências ou dimensões, categorizar essas informações e resumir as relações identificadas em bases de dados. De acordo com da Costa Côrtes et al. [27], é um processo altamente cooperativo entre homens e máquinas, que visa a exploração de grandes bases de dados, com o objetivo de extrair conhecimentos através do reconhecimento de padrões e relacionamentos entre variáveis, conhecimentos esses que possam ser obtidos por métodos comprovadamente fiáveis e validados pela sua expressividade estatística.
Em síntese, subjacente às definições apresentadas, é possível concluir que o *data mining* é um instrumento de gestão, suportado pela conjunção de diversos métodos científicos de conhecimento multidisciplinar, capaz de processar um enorme conjunto de dados, com o objetivo de extrair informação relevante e explícita, que possa ser útil para apoiar a tomada de decisões. É, portanto, todo um processo que permite aprender com os dados.

2.3.2 Objetivos, modelos e métodos de *data mining*

Um estudo de *data mining* requer a especificação de algumas componentes-chave que permitam adequar de forma correta a sua abordagem. Primeiro, é necessário definir convenientemente quais são os objetivos genéricos subjacentes à prospeção de dados. Depois, é necessário distinguir qual o tipo de modelo a ser gerado e quais as tarefas a executar. Por fim, selecionam-se os métodos e algoritmos mais adequados para concretizar os objetivos pretendidos. Nesta subsecção aclara-se o que se entende sobre cada uma das referidas componentes-chave para melhor perceção e interpretação dos conteúdos apresentados ao longo deste documento.

Objetivos e modelos de *data mining*

Os objetivos genéricos de *data mining* cingem-se à geração de modelos analíticos a partir dos registos em grandes conjuntos de dados. Um modelo analítico é um conceito que geralmente representa uma descrição formal de alguns aspetos da realidade, física ou social, com a finalidade de a compreender e de a comunicar (Mylopoulos et al. [77]). Dependendo do domínio do problema que representam, os modelos analíticos podem ser categorizados em dois grandes grupos:

modelos descritivos o objetivo principal da modelação descritiva é encontrar padrões frequentes, que possam explicar, ou generalizar, a estrutura intrínseca dos dados, incluindo os seus relacionamentos. Os padrões podem ser, por exemplo, presença de anomalias, tendências, agrupamentos entre objetos, associações e correlações entre variáveis. O seu intuito principal é analisar o que foi descoberto nos dados sob o ponto de vista da interpretação humana.

modelos preditivos o objetivo principal da modelação preditiva é estimar valores, desconhecidos ou futuros, de uma ou mais variáveis alvo de interesse, a partir de alguma combinação

de outras características presentes nos dados. Permite, portanto, antever circunstâncias futuras, tais como certas tendências e certos comportamentos. A variável objeto de previsão é designada variável alvo, variável resposta ou variável dependente (VD), enquanto que os atributos usados para a previsão são designadas variáveis independentes, preditivas ou explicativas.

De realçar, no entanto, que os modelos preditivos também poderão facultar uma descrição de um conjunto de dados e os modelos descritivos também poderão revelar previsões de eventos futuros. A distinção entre um e o outro é muito ténue e depende da forma de representação do próprio modelo e do objetivo pelo qual foi induzido. Mais especificamente, aqueles modelos cuja representação não é facilmente interpretável serão usados apenas para previsão. Por outro lado, os modelos preditivos cuja representação surge em forma de regras, ou nalguma estrutura facilmente interpretável também podem ser usados para descrever os dados.
Na Figura 2.3 apresenta-se um esquema ilustrativo de todo o processo de indução de um modelo analítico/preditivo a partir de um conjunto de dados. Para o modelo ser induzido, tal

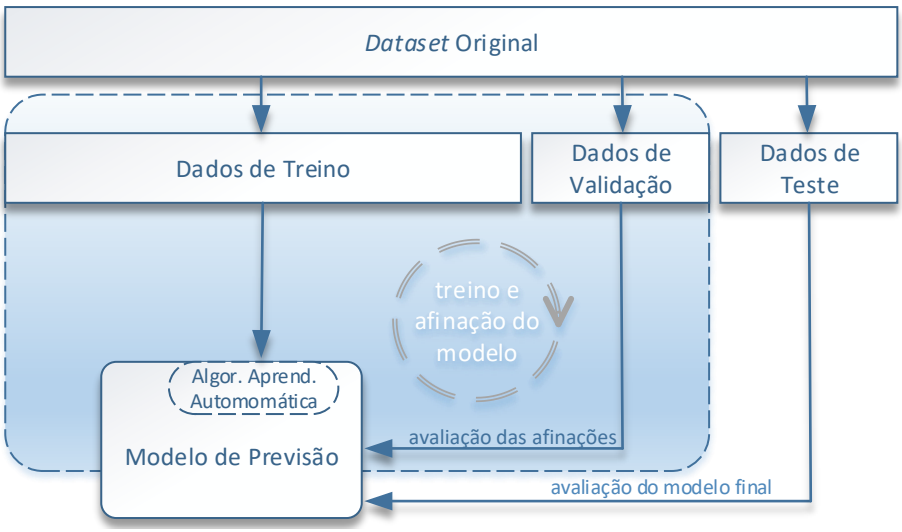


Figura 2.3: Indução dum modelo de previsão.

como ilustrado, o conjunto de dados para análise é normalmente dividido em três subconjuntos: de treino, de validação e de teste. O conjunto de treino é processado por um algoritmo de aprendizagem a fim de ser induzido o modelo. O segundo subconjunto serve para ajustar os hiperparâmetros responsáveis pela afinação (configuração) dos algoritmos que suportam o modelo. Por fim, concluído que estão o treino e a afinação do modelo, os dados do subconjunto de teste são mostrados pela pela primeira vez ao modelo encontrado, de forma a avaliar a sua verdadeira capacidade preditiva e em especial a sua capacidade de generalização. Se o modelo evidenciar elevada eficácia preditiva no conjunto de dados deixado para teste, assume-se que o modelo tem efetivamente boa capacidade de generalização, ou seja, será expectável que apresente bom desempenho com dados futuros e desconhecidos. Por outro lado, se o modelo apresentar um fraco desempenho com os dados do conjunto de teste, então o modelo criado não será adequado para efetuar a previsão. Perante esta segunda situação, é usual retornar-se à fase de pré-processamento para aprimorar os dados, ou, simplesmente, recorrer a outro algoritmo de aprendizagem.
Dependendo do género de aprendizagem no conjunto de treino, os modelos analíticos podem ser categorizados em supervisionados ou não supervisionados. No primeiro caso, o modelo en-

contrado permite prever um valor, ou um rótulo, que caraterize um novo exemplo, com base nos valores dos seus atributos de entrada. É, por conseguinte, necessário que na fase de aprendizagem a cada instância do conjunto de dados esteja associada a variável de saída (VD) e os atributos de entrada. Na aprendizagem não supervisionada as instâncias do conjunto de dados só estão caracterizados por atributos de entrada e não existe informação sobre o valor da VD associada a cada exemplo. A aprendizagem do modelo é efetuada descobrindo similaridades nos dados, isto é, formam-se agrupamentos de dados de acordo com as características semelhantes. Em qualquer tipo de aprendizagem, seja ou não supervisionada, um dos objetivos principais de *data mining* é a criação de modelos com capacidade de generalização. Trata-se de um conceito que traduz a capacidade de um algoritmo prever com precisão novos exemplos, ainda não observados, depois de ter construído um modelo com base num conjunto de dados de aprendizagem. Considera-se haver *overfitting* quando o modelo se ajusta demasiado aos dados de treino, comprometendo, dessa forma, a sua capacidade de generalização.

Métodos ou tarefas de *data mining*

As modelações descritiva e preditiva, a partir de grandes conjuntos de dados, requerem a aplicação de métodos e algoritmos de análise automatizada. Trata-se de instrumentos especialmente concebidos para encontrar e descrever formalmente os padrões estruturais presentes nos grandes conjuntos de dados. A Figura 2.4 apresenta a taxonomia proposta por Baker et al. [7], onde constam os métodos mais frequentemente usados pela comunidade de EDM.

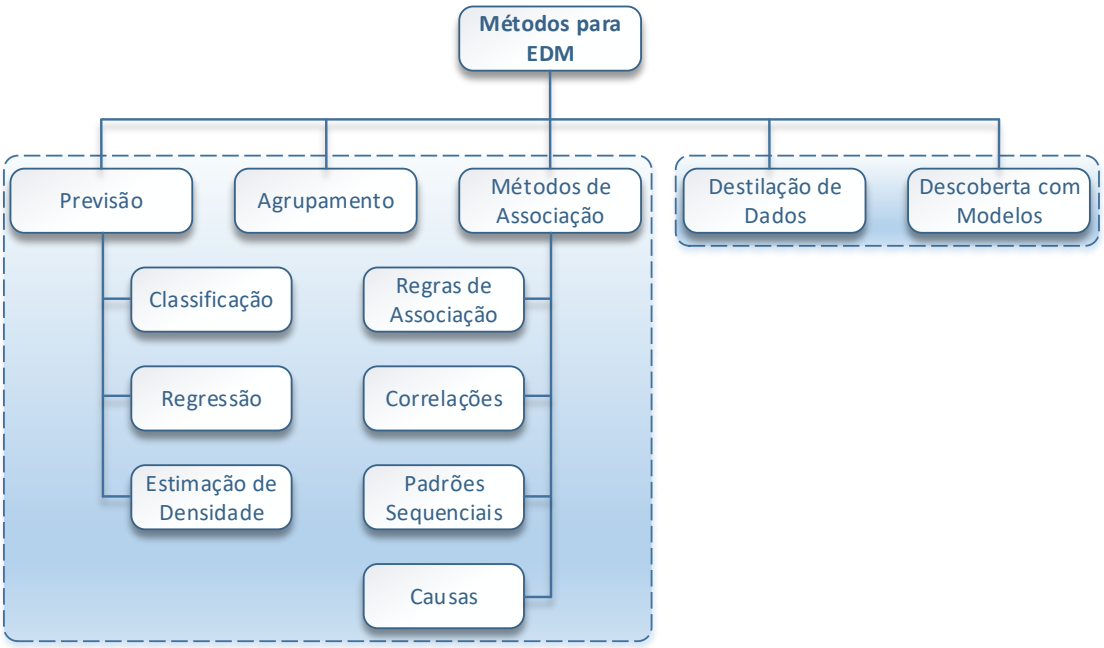


Figura 2.4: Taxonomia dos métodos para EDM (adaptada de Baker et al. [7]).

Os métodos de previsão, de regressão e de classificação, os métodos de associação e os métodos de *clustering* são transversais a todos os domínios de *data mining*. No entanto, outros métodos, como a destilação de dados para interpretação humana e descoberta com modelos têm especial destaque no contexto de EDM (Romero and Ventura [95]). Os métodos de agrupamento e de associação são dos mais populares quando o objetivo é obter modelos descritivos, enquanto os métodos de classificação e de regressão, são os métodos mais usuais, quando se pretende induzir modelos preditivos. A cada um destes métodos estão associados uma ampla variedade

de técnicas, também designados algoritmos. Wu et al. [118] apresentaram um estudo onde identificaram e descreveram os algoritmos mais influentes para a comunidade de pesquisa de *data mining*.

Uma vez que um modelo pode ser induzido por mais do que um algoritmo, a escolha do mais adequado é uma arte (Fayyad et al. [35]). Usualmente, não há *a priori* garantias de sucesso e fiabilidade na escolha do melhor algoritmo. Perante um grande conjunto de dados, os analistas vão selecionando as tarefas, métodos, e algoritmos de *data mining* a usar à medida que o processo decorre (Oliveira [80]). Essa seleção dependente sempre de dois aspetos muito importantes, que são o género de dados a analisar (PhridviRaj and GuruRao [87]) e os objetivos subjacentes à prospeção de dados. Por vezes, os investigadores, no intuito de selecionar o algoritmo mais adequado para a modelação, recorrem a uma metodologia de comparação, em que para o mesmo propósito e para o mesmo conjunto de dados, são usados vários algoritmos. Posteriormente, através de alguma métrica de avaliação de desempenho dos modelos encontrados, escolhe-se, aquele cujos resultados são melhores. Este procedimento de comparação permite, igualmente, averiguar se com o uso de diferentes métodos e algoritmos, resultados idênticos serão alcançados.

Na secção que se segue descrevem-se os métodos de previsão e os algoritmos a eles associados, por se tratarem dos métodos de *data mining* mais usados no contexto educacional e, particularmente, por serem aqueles aos quais se recorre no âmbito da investigação a desenvolver na presente tese.

2.4 Métodos de previsão

De acordo com Baker et al. [7] e com Romero and Ventura [95] os métodos de previsão, no contexto de *educational data mining*, usam-se, essencialmente, para dar resposta a duas questões de investigação. Nalguns casos, são usados para identificar as características do modelo importantes para a previsão, dando informações sobre a construção subjacente. Por exemplo, quando se pretende identificar as características/fatores dos estudantes que explicam o abandono/não abandono académico (e.g. Kotsiantis et al. [56], Delen [31], Nandeshwar et al. [78]), ou então quando se pretende identificar as características dos estudantes que explicam o seu sucesso no final do curso de graduação (e.g. Miguéis et al. [71], Natek and Zwilling [79], Aluko et al. [3]). Noutros casos, os métodos de previsão são usados para estimar o valor de saída duma ou mais variáveis alvo em contextos onde não seja possível, ou desejável, a obtenção direta do seu valor. Por exemplo, naturalmente, sempre que se pretenda prever eventos ou resultados futuros, ou até mesmo quando a obtenção direta do valor da variável alvo, ainda que possível, possa influenciar o próprio comportamento dessa variável.

Tal como se pode observar na taxonomia proposta por Baker et al. [7] e também corroborado por Romero and Ventura [95], no contexto do EDM são 3 os métodos de previsões mais comuns: os de classificação, de regressão e de estimativa de densidade. A principal diferença entre os modelos preditivos de classificação e os preditivos de regressão reside na caracterização da variável alvo de previsão. Na classificação as variáveis dependentes são categóricas ou discretas, enquanto que na regressão assumem valores contínuos. Na estimativa de densidade, a variável alvo de previsão é uma função de densidade de probabilidade.

Como em contexto de análise a dados educacionais a estimação de densidade é raramente utilizada, derivado da falta de independência estatística dos dados, apenas se descrevem, de forma sucinta, os métodos de classificação e de regressão.

2.4.1 Regressão

Os modelos de regressão têm como objetivo prever os valores futuros, ou desconhecidos, de uma ou mais variáveis numéricas contínuas, a partir de outros atributos presentes no conjunto de dados. Dito de outra forma, é um método que permite descrever a relação entre as variáveis alvo de previsão e as variáveis independentes ou explicativas. A análise de regressão tem como resultado uma função matemática que descreve o relacionamento entre essas variáveis e pode ser usada para estimar ou prever valores futuros de uma ou mais variáveis, quando se conhecem ou se supõem conhecidos os valores de outras variáveis. Na indução de modelos de regressão, os métodos mais populares, no contexto de EDM, são a regressão linear, as redes neuronais artificiais e as *support vector machines* (Baker et al. [7], Costa et al. [25]). No caso deste último método, são usadas versões adaptadas para regressão do algoritmo inicialmente desenvolvido para problemas de classificação. A regressão logística, em particular, é uma técnica estatística que pode ser usada para modelar resultados não contínuos. Permite avaliar a relação entre a variável categórica dependente e uma ou mais variáveis independentes. As probabilidades são estimadas usando a função *logit*.

2.4.2 Classificação

Classificar é a ação de separar um determinado objeto de acordo com as suas próprias características. Em *data mining* os métodos de classificação usam-se quando se pretende induzir um modelo classificador passível de descrever as classes de dados. Pretende-se que o modelo gerado atribua a cada um dos objetos/itens de dados, em função das suas variáveis explicativas, uma das classes categóricas pré-definidas (Romero and Ventura [95]). Como o atributo classe é conhecido na fase de treino diz-se que a fase de aprendizagem é supervisionada (Witten et al. [117]). O método de classificação é um dos mais frequentemente usados pela comunidade científica de *data mining*, (Wu et al. [118], Romero et al. [96]) quando se pretende prever valores do tipo categórico. Tem também muitas aplicações no contexto de EDM (Hämäläinen and Vinni [45]), entre as quais se destaca a sua popularidade no âmbito da previsão de desempenho dos estudantes (Shahiri et al. [103], Del Río and Insuasti [30]).
De acordo com Hämäläinen and Vinni [45], há quatro índoles de modelos classificadores que predominam no contexto de EDM:

- Quando se pretende classificar os estudantes relativamente ao (in)sucesso esperado no futuro. Por exemplo, quando se pretende estimar se o estudante é da classe reprovar/não reprovar de ano (e.g. Hoffait and Schyns [49]), ou reprovar/não reprovar a uma disciplina específica (e.g. Huang [52], Pascoal et al. [85], Costa et al. [26]).
- Quando se classifica o nível de sucesso esperado para o estudante no final do curso em categorias (e.g. excelente, bom, razoável, sofrível) [112, 79, 74, 3], ou quando se classifica o estado do aluno com sendo da classe progresso ou não progresso (Manhães [64]).
- Quando se tem como objetivo estimar se um estudante é da classe abandono/não abandono, ou quando se pretende classificar os estudantes relativamente ao risco de abandono escolar: baixo, médio e elevado (e.g. Kotsiantis et al. [56], Dekker et al. [29], Papamitsiou et al. [83]).
- Sempre que se pretende classificar as habilidades cognitivas do estudante, como por exemplo, sintomas de baixa motivação em cursos ministrados em *e-learning* (e.g. Romero and Ventura [94], Gray et al. [43]).

De acordo com Witten et al. [117], são 5 os principais algoritmos que induzem modelos classificadores: árvores de decisão (e o método que as combina, as *random forest*), classificadores baseados em regras, classificadores *bayesianos*, classificadores do vizinho mais próximo, redes neuronais artificiais e *support vector machines*. Descrevem-se de seguida as redes neuronais artificiais, as *random forest* e as *support vector machines*, por traduzirem três técnicas que têm exibido desempenhos muito competitivos no contexto do EDM — como se depreenderá da revisão de literatura apresentada no próximo capítulo desta tese — e em virtude de se tratar dos métodos selecionados para os modelos preditivos desenvolvimentos nesta tese (descritos nos Capítulos 5 e 6). Começa-se, no entanto, com uma breve descrição das árvores de decisão, por se tratarem dos classificadores mais populares no âmbito de EDM (Hämäläinen and Vinni [45], Shahiri et al. [103]), e uma vez que estão da génese das *random forest*.

2.4.3 Algoritmos de *data mining*

2.4.3.1 Árvores de decisão

Uma árvore de decisão (AD) apresenta um conjunto de regras de classificação organizadas segundo uma estrutura em forma de árvore. As regras subjacentes aos dados são representadas através de estruturas hierárquicas que dividem os dados de forma recursiva. Cada nó da árvore representa um critério de separação com base no valor que assuma um atributo específico e cada ramo que sai desse nó representa um dos possíveis resultados desse critério. Começa por ser escolhido para o nó raiz da árvore o atributo e o critério que melhor separe os dados de treino, após o qual, e segundo esse critério de separação, os dados são divididos em diferentes subconjuntos. Depois, para cada um desses subconjuntos cria-se um novo nó, descendente do primeiro, e aplica-se o mesmo procedimento, escolhendo o atributo e critério que melhor separe o subconjunto, e assim sucessivamente, até que os dados do subconjunto façam todos parte de uma mesma classe.
Para um melhor entendimento do funcionamento deste tipo de classificador, apresenta-se na Figura 2.5 um exemplo de árvore de decisão para classificação binária. Cada variável x_i repre-

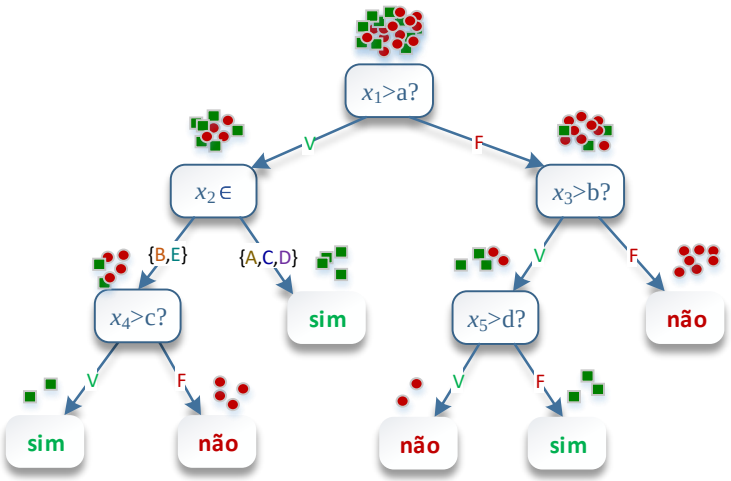


Figura 2.5: Árvore de decisão para classificação binária.

senta um dos atributos que caracterizam as instâncias do *dataset*. O atributo usado em cada nó e o respetivo critério de separação são escolhidos de forma a maximizar a separação, em dois subgrupos, das instâncias de classe positiva das de classe negativa. Um atributo pode ocorrer

mais do que uma vez na mesma árvore, podendo ser numérico ou categórico. No exemplo, a variável x_2 representa um atributo categórico que pode assumir 5 classes distintas (A, B, C, D e E), e todas as restantes são atributos, necessariamente, numéricos ou de tipo ordinal. Junta-se ainda a cada nó um conjunto de instâncias positivas e negativas, de forma ilustrar as sucessivas divisões que vão sofrendo os subconjuntos de dados à medida que lhes vão sendo aplicados os respetivos critérios de separação.
Uma árvore de decisão usa a estratégia de dividir para conquistar na resolução de problemas de decisão. Isto significa que um problema complexo é subdividido em problemas mais simples, aos quais, recursivamente, é aplicada a mesma estratégia. As soluções dos subproblemas encontram-se combinadas, na estrutura em árvore, de forma a darem resposta ao problema original. Trata-se de um algoritmo de treino supervisionado para proceder à classificação de cada uma das instâncias presentes no conjunto de dados inicial.
A literatura menciona diversos algoritmos de classificação que induzem árvores de decisão. Entre os mais recorrentes surgem o CART e o J48, resultando este segundo duma evolução dos algoritmos, seus antecessores, ID3 (*Iterative Dichotomiser 3*), C4.5 e C5.0. O ID3, assegurando o funcionamento típico de uma AD, desenha a árvore escolhendo, de forma sucessiva, o atributo que melhor divide os exemplos. O algoritmo C4.5 estende o algoritmo ID3 de forma a tratar dados com valores omissos e valores numéricos contínuos. Por sua vez, o algoritmo C5.0 é uma versão aprimorada do algoritmo C4.5, por ser mais eficiente computacionalmente, produzir regras mais assertivas e árvores de menor tamanho.
Depois de treinada, a árvore de decisão irá classificar cada exemplo, de acordo com o caminho que satisfizer as condições desde o nó raiz até ao nó terminal, sendo o exemplo classificado de acordo com a classe associada a esse último nó. Refira-se por fim que, após a criação da árvore, normalmente são-lhe aplicadas técnicas de "poda", de forma a expurgá-la de possíveis impurezas, contribuindo-se assim para que somente a informação considerada relevante seja usada na tomada de decisão.

Vantagens e desvantagens das árvores de decisão

Os algoritmos de árvores de decisão apresentam como principal vantagem o facto de produzirem regras de classificação fáceis de interpretar. Apresentam também outras vantagens como por exemplo, o facto de serem adaptáveis a objetivos de regressão, serem bastante eficientes na construção dos modelos, não serem dependentes da escala de variáveis e apresentarem robustez na presença de *outliers* e de atributos redundantes ou irrelevantes. A possibilidade de pequenas perturbações do conjunto de treino poderem provocar grandes alterações no modelo aprendido é a principal desvantagem que é apontada a esta técnica.

2.4.3.2 *Random forest*

O algoritmo *random forest* (RF), proposto por Breiman [14], é um método de aprendizagem indutiva baseada em comités, também designados métodos de conjunto ou ainda mistura de especialistas, que gera múltiplas árvores de decisão durante o treino. Com base na premissa de que um conjunto de classificadores fracos pode criar um classificador forte, as RF combinam os resultados de múltiplas árvores de decisão treinadas individualmente, com padrões de erro diferentes, para tentar otimizar o desempenho preditivo global.
Tal como se ilustra na Figura 2.6, para induzir individualmente cada uma das árvores da floresta, o algoritmo particiona recursivamente o conjunto de dados de treino inicial D, que contém n

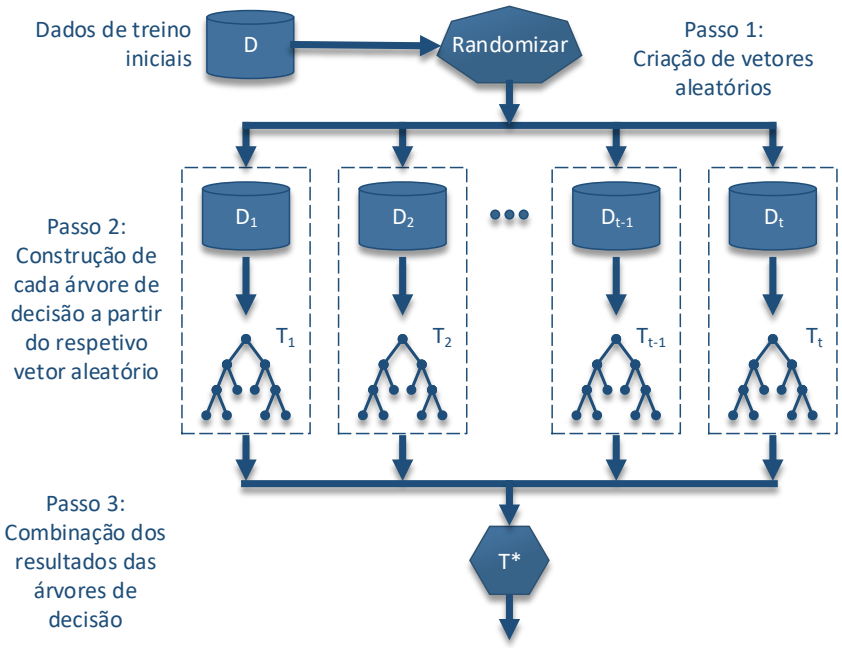


Figura 2.6: *Random forest*.

exemplos e d atributos, em múltiplos subconjuntos de treino de menor dimensão, $D_1 \dots D_t$, cada um obtido de forma independente e por re-amostragem aleatória com reposição do conjunto original. Cada uma destas múltiplas partições, que contém m exemplos e i atributos, em que $m < n$ e $i < d$, é utilizada para induzir uma diferente árvore da floresta. Os exemplos do conjunto de treino inicial D que não surjam no subconjunto de treino de uma árvore individual, os denominados dados *out-of-bag*, são utilizados como dados de teste para estimar o desempenho dessa árvore durante a fase de treino, conseguindo-se com isso uma estimativa mais fiável, uma vez que se tratam de dados novos. O processo de amostragem aleatória, quer dos exemplos, quer dos atributos, vai provocar o aumento da variabilidade das árvores da floresta, conseguindo-se, por essa via, reduzir a variância do modelo final e manter baixo o enviesamento do erro na generalização.

Para que as *random forest* possam induzir um modelo a partir de um conjunto de dados é necessário então definir o número de árvores de decisão (t) a serem geradas e o número de atributos a considerar (i), que o algoritmo selecionará de forma aleatória, para critério de decisão em cada nodo das árvores. O valor normalmente sugerido para o número de atributos aleatórios é a raiz quadrada do número total de atributos, em problemas de classificação, e um terço desse número em problemas de regressão. Para o número de árvores a gerar deve ser usado um valor elevado (Breiman [14]), para assegurar que qualquer uma das instâncias do conjunto de treino seja prevista pelo algoritmo algumas vezes. No final, o algoritmo tem ainda a capacidade de gerar uma lista com o *ranking* da importância dos atributos no desenvolvimento da floresta. Essa importância é determinada em função da importância acumulada de cada atributo nas sucessivas divisões dos nodos que vão compondo as árvores da floresta. São estas especificidades que caracterizam as *random forest* como um algoritmo mais poderoso quando comparado com uma árvore de decisão gerada isoladamente.

Uma vez as árvores construídas, o modelo estará pronto para ser usado na classificação de dados novos. Para classificar um exemplo de teste, alimenta cada uma das suas árvores especialistas com o vetor de entrada que represente esse exemplo, para que cada uma delas faça a sua clas-

sificação. É a classe mais escolhida pelas árvores aquela que será assumida para a classificação final do exemplo (tratando-se de um problema de regressão, o resultado final será uma média das previsões produzidas pelas árvores).

Vantagens e desvantagens das *random forest*

A capacidade de gerar estimativas com baixo enviesamento, a aptidão para a modelação de relações não lineares de elevada dimensão, a habilidade para tratar com maior eficácia características que tenham dados em falta, a resistência ao *overfitting*, a relativa robustez em relação a atributos com ruído, a aptidão para lidar com dados categóricos e contínuos, a possibilidade de estimar a importância das variáveis preditivas, o facto de conseguirem detetar interações entre as variáveis e ainda a rapidez de construção e eficiência preditiva, mesmo em conjuntos de elevada dimensão e dimensionalidade (Breiman [14], Liaw and Wiener [61], Spoon et al. [107]), são vantagens que as definem como uma abordagem capaz de produzir desempenhos preditivos muito competitivos. As desvantagens mais significativas que lhe são apontadas são a falta de reprodutibilidade para diferentes itens de dados, o facto de enviesarem os resultados quando existem características nominais com elevado número de atributos distintos e o facto dos modelos gerados serem de difícil interpretação (Spoon et al. [107]). Por fim, pode-se afirmar que as *random forest* são um dos algoritmos que melhores resultados oferece para um conjunto vasto de aplicações, ainda que envolvam conjuntos de dados de grande dimensão (muitas instâncias) e de elevada dimensionalidade (muitos atributos).

2.4.3.3 Redes neuronais artificiais

Uma Rede Neuronal Artificial (RNA) é uma técnica de aprendizagem indutiva que comporta uma inspiração de inteligência biológica, baseada na estrutura e no funcionamento do sistema nervoso, com o objetivo de simular a capacidade de aprendizagem do cérebro humano na aquisição de conhecimento. Trata-se de uma estrutura extremamente interconetada de unidades computacionais, designadas *nodos* ou *neurónios artificiais*, com capacidade de aprendizagem (Cortez and Neves [23]), representando, assim, aproximações simplificadas das redes de neurónios que se encontram no cérebro humano.
Os trabalhos iniciais com RNA foram propostos por McCulloch e Pitts em 1943. Mas foi principalmente na década de 80 do século XX que surgiram os avanços mais significativos, impulsionados, essencialmente, pelo avanço tecnológico dos computadores e pelo impacto das contribuições científicas da autoria de Hopfield [50]), sobre novas arquiteturas de RNA, e de Rumelhart et al. [101], sobre algoritmos de aprendizagem mais sofisticados, com destaque para o algoritmo por retro-propagação. A partir de então estavam reunidas as circunstâncias favoráveis ao desenvolvimento da área. Nos dias de hoje, a teoria das RNA vem-se consolidando, como uma nova e eficiente ferramenta para lidar com uma ampla classe de problemas complexos, em que extensas massas de dados devem ser modeladas e analisadas, envolvendo, simultaneamente, tanto os aspetos estatísticos e computacionais como os dinâmicos e de otimização (Kovács [57]). De forma sucinta, a formação e funcionamento de uma RNA são seguidamente apresentados.

Arquitetura das RNA

A componente de processamento fundamental de uma RNA é o neurónio. Na Figura 2.7 é apresentado um modelo simples de um único neurónio. Estas unidades, densamente interligadas através de um padrão de conexões, desempenham um papel muito simples que consiste em

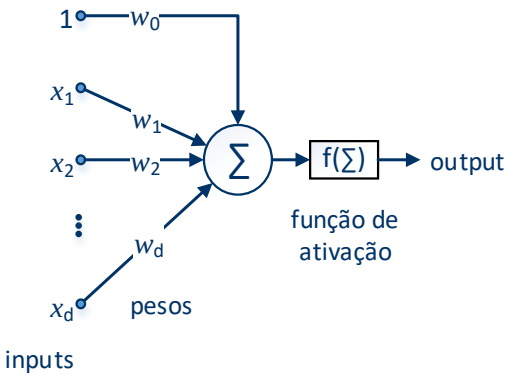


Figura 2.7: Aquitetura do neurónio.

receber sinais das ligações de entrada e com eles calcular um novo valor para ser enviado para a saída. Especificamente, cada terminal de entrada do neurónio, simulando um dendrite, recebe um valor. Os valores recebidos são então ponderados e somados, sendo depois o valor resultante, acrescido de um valor de *offset* (ou *bias*), usado para calcular o valor de saída do neurónio, em analogia com o processamento realizado pelo corpo celular ou soma. Esse cálculo é realizado por uma função específica, denominada função de ativação, que assume normalmente uma das formas típicas ilustradas na Figura 2.8. O *input* da função de ativação pode

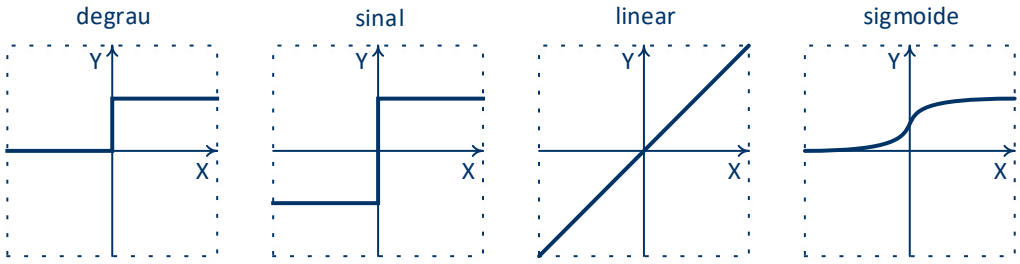


Figura 2.8: Funções de ativação típicas.

então ser expresso de forma simples pela seguinte expressão:

$$s = \mathbf{w}^t \cdot \mathbf{x} + w_0 = \sum_{i=1}^{d} w_i x_i + w_0, \quad (2.1)$$

em que $\mathbf{x} = [x_1, x_2, x_3, \ldots, x_d]^t$ é o vetor com o conjunto de d entradas do neurónio, $\mathbf{w} = [w_1, w_2, w_3, \ldots, w_d]^t$ o vetor de pesos associados e w_0 o valor de *offset* ou *bias*. São estes pesos, associados às entradas dos neurónios, que constituem o principal meio de armazenamento de informação numa rede.

O *output* do neurónio dependerá então da função de ativação que for escolhida. Com a função degrau, o sinal será 0 enquanto a entrada s for negativa (neurónio desativado), passando a 1 logo que se torne positiva (neurónio ativado). Já a função sigmoide representa uma aproximação contínua e diferençável da função degrau. O uso da função linear limitar-se-á a passar para a saída um sinal igual (ou, no mínimo, proporcional) ao de entrada e a função sinal converte a entrada nas saídas -1 e $+1$, em concordância com o sinal, negativo ou positivo, da entrada.

Estando os neurónios interligados em rede, as entradas $\mathbf{x}$ de um neurónico interior da rede não serão mais do que as saídas dos d neurónios que o precedam e que a ele estejam ligados. Já se for um neurónio da camada inicial, as entradas $\mathbf{x}$ corresponderão às entradas do próprio problema (vetor com os valores observados nos vários atributos) e tratando-se de um neurónio

da última camada, o seu valor de saída representará um dos valores estimados pela rede.
As várias unidades de processamento computacional, ou neurónios artificiais, estão dispostas numa ou mais camadas e interligadas por um grande número de conexões. O número de camadas, o número de neurónios em cada camada, o grau de conectividade e a presença ou não de conexões de retro-propagação definem a arquitetura, ou topologia, de uma RNA. Existem várias arquiteturas de RNA, ou topologias, com características diferentes. Uma das características mais diferenciadoras é se as redes contêm apenas ligações para a frente (redes progressivas ou *feedforword*), ou se existem ciclos de realimentação, formando nesse caso as redes com retro-propagação ou redes recorrentes. Geralmente são classificadas nas seguintes três categorias:

- Redes progressivas de uma só camada: apresentam sempre conexões unidirecionais, convergentes ou divergentes. Na sua forma mais simples uma rede é composta por uma "camada" de entrada, cujos valores de saída são fixados externamente, e por uma camada de saída, contendo todos os neurónios da rede (ver Fig. 2.9). Têm a limitação de apenas conseguirem classificar objetos linearmente separáveis, ou seja, se houver um hiperplano que separe os dados das duas classes.

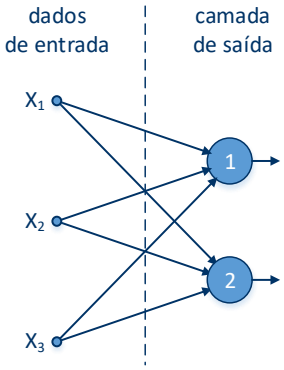


Figura 2.9: Redes neuronais de uma só camada.

- Redes progressivas multicamada: distinguem-se pelo facto de possuir uma ou mais camadas intermédias, designadas por camadas escondidas, cujos neurónios são também designados neurónios intermédios (ver Fig. 2.10). A função destes é de intervir de forma útil entre a entrada e a saída da rede. Ao se acrescentarem camadas intermédias está-se a aumentar a capacidade da rede para modelar funções de maior complexidade. As redes mais frequentemente usadas, desta categoria, são as *Multilayer Perceptron* (MLP).

- Redes recorrentes: são realimentadas das saídas para as entradas (ver Fig. 2.11). Devido a esta característica, respondem a estímulos de forma dinâmica, ou seja, após se aplicar uma nova entrada, a saída é recalculada e usada para modificar a entrada.

O primeiro passo para que as RNA possam induzir um modelo a partir de um conjunto de dados é definir a sua arquitetura. A escolha correta da função de ativação e da topologia da rede é decisiva para uma aprendizagem ou treino bem sucedido. Por exemplo, se a rede for muito pequena poderá ser incapaz de resolver o problema proposto. Por outro lado, se a rede for demasiado grande, poderá revelar-se incapaz de generalizar (*overfitting*). A arquitetura da rede é, por isso, determinante na capacidade de processamento e aprendizagem de uma RNA. Geralmente, a arquitetura mais promissora é encontrada através de um processo exaustivo de tentativa e erro, onde diferentes configurações são investigadas, comparadas e avaliadas para selecionar aquela que apresente melhor capacidade preditiva. Embora seja a abordagem mais

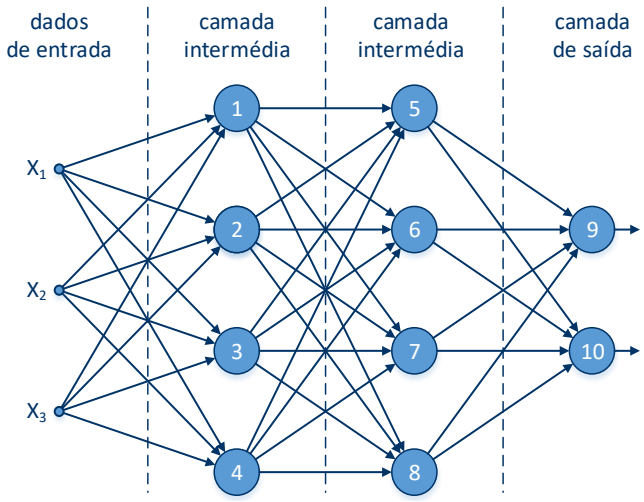


Figura 2.10: Redes neuronais multicamada.

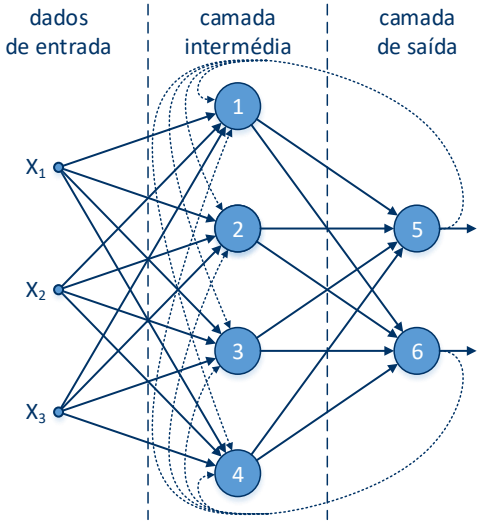


Figura 2.11: Redes neuronais recorrentes.

utilizada, esta procura cega tem a desvantagem de apresentar um elevado tempo de processamento. Alternativamente, pode recorrer-se a algoritmos que realizem uma procura de RNA eficientes.

Aprendizagem das RNA

As conexões entre os neurónios de uma RNA, que simulam as sinapses biológicas, possuem pesos associados, os quais desempenham, como já se referiu, um papel essencial na aquisição de conhecimento numa rede. São esses pesos que vão assumindo os valores adequados durante o processo de aprendizagem cujo objetivo principal é obter um conjunto de saídas desejadas e consistentes, a partir de um conjunto de entradas. Mais especificamente, após se atribuir um valor inicial aos pesos das ligações, normalmente por um processo aleatório, todos os vetores de observações presentes num conjunto de treino são apresentados à rede, por regra, mais do que uma vez (várias épocas). Sempre que um desses vetores é mostrado à rede, os neurónios da primeira camada escondida recebem como entradas as observações nele contidas. Esses neurónios produzem as respetivas saídas passado-as de seguida para os neurónios da camada

seguinte, e assim sucessivamente até se chegar à última camada de neurónios. O valor produzido à saída de cada um desses últimos neurónios é então comparado com o resultado que já é conhecido para o item em causa. A diferença entre os valores verificados à saída e os valores reais conhecidos indica o erro cometido pela rede para o item apresentado, e é esse erro que, de alguma forma, será levado em conta no ajuste dos pesos numa aprendizagem supervisionada. Durante todo este processo, os pesos da rede convergem de forma gradual para determinados valores, sendo, por regra, necessário apresentar à rede sucessivas épocas dos dados de treino, até que os vetores de entrada produzam as saídas desejadas.

A aprendizagem de uma RNA está intimamente relacionada com a forma pela qual se procede à modificação dos parâmetros, ou seja, como vão sendo definidos os valores dos pesos associados às conexões da rede que fazem com que o modelo obtenha cada vez melhor desempenho. Atualmente existe um conjunto diversificado de algoritmos de treino, que usam regras de aprendizagem diferentes, desenvolvidos de acordo com os principais paradigmas de aprendizagem: supervisionada, por reforço e não supervisionada. Por exemplo, o algoritmo de treino supervisionado mais popular para as redes MPL é o algoritmo de retro-propagação (*backpropagation*), que minimiza o erro quadrático entre os valores reais esperados e os valores produzidos pela rede. A ideia central subjacente a este algoritmo é a de que o erro produzido pelos neurónios das camadas escondidas é estimado retro-propagando, sucessivamente, os erros cometidos na camada de saída. Trata-se de um algoritmo iterativo que opera em duas fases, primeiro para a frente (*forward*) e de seguida para trás (*backward*), que se descrevem de forma muito sucinta como se segue. Na fase *forward*, um vetor do conjunto de treino é apresentado à rede, a qual calcula, com os pesos atuais, a saída correspondente e o erro associado; depois, na fase *backard*, o erro é sucessivamente retro-propagado, desde a camada de saída até à primeira camada escondida, para que a partir dos erros estimados para cada neurónio se encontre o ajuste adequado para os pesos das suas entradas.

Para um entendimento mais completo sobre os diferentes algoritmos propostos para o processo de aprendizagem das RNA, e sobre o seu funcionamento em geral, consultar, por exemplo, Cortez and Neves [23], Silva and Ribeiro [105], Gama et al. [42].

Em síntese, uma RNA é portanto definida pela sua arquitetura e algoritmo de aprendizagem. Enquanto a arquitetura está relacionada com o tipo e o número de unidades de processamento (neurónios) e com a forma como elas estão interligadas, a aprendizagem diz respeito às regras utilizadas para o ajuste dos pesos da rede. Por conseguinte, para se construir uma RNA é necessário começar por especificar o tipo e o número de neurónios e particularizar a forma como se interligam, ficando assim definida a sua topologia. Seguidamente, atribuem-se valores iniciais aos pesos das ligações (por norma aleatórios) e inicia-se com a aprendizagem da rede, fase em que cada um dos itens do conjunto de treino é iterativamente apresentado à rede. Para cada um desses itens de treino, os pesos são atualizados de forma a minimizar o erro entre a previsão da rede e o resultado conhecido. O processo é concluído quando os pesos convergem para valores adequados ao problema.

Refira-se, por fim, que existe normalmente um parâmetro importante associado aos algoritmos de aprendizagem, denominado taxa de aprendizagem, que permite acelerar ou desacelerar o treinamento da rede. Trata-se de uma constante entre 0 e 1 que o analista pode escolher. Quanto maior for essa constante, maior será a alteração dos pesos em cada iteração, conseguindo-se dessa forma aumentar a velocidade de aprendizagem, mas também a possibilidade de ocorrerem oscilações que tornem o algoritmo instável. Por outro lado, se a taxa de aprendizagem for demasiado baixa, a possibilidade do algoritmo convergir é de facto maior, mas, em contrapartida, a rede pode vir a demorar demasiado tempo a aprender. Uma outra

opção interessante, que tenta mitigar as dificuldades enunciadas, passa pela utilização de uma taxa variável que vá diminuindo o seu valor à medida que o algoritmo se vai aproximando do valor de convergência.

Vantagens e desvantagens das RNA

As RNA possuem várias características que justificam o seu bom desempenho preditivo e a sua popularidade. A capacidade de aprendizagem e generalização, o processamento maciçamente paralelo, a flexibilidade, a transparência, a linearidade, a adaptatividade, a resposta evidencial e a tolerância a falhas e ruídos, são apresentadas por vários autores (e.g. Cortez and Neves [23], Kovács [57], Haykin and Network [48], Han et al. [46]) como as suas principais vantagens. O elevado tempo de processamento, tal como as dificuldades associadas à parametrização da arquitetura da rede e à interpretação dos resultados obtidos, configuram as críticas mais frequentemente apontadas às RNA. Devido a esta ultima limitação, as RNA são usualmente designadas como técnicas do tipo caixa-preta.

2.4.3.4 Máquinas de vetores de suporte

Uma máquina de vetores de suporte (*Support Vector Machine – SVM*) é um algoritmo desenvolvido em 1995 por Vladimir Vapnik (Vapnik [114]), para objetivos de classificação binária. Tem a sua origem na aplicação de conceitos da teoria da aprendizagem estatística, tentando, por essa via, minimizar o erro real do problema e não apenas o erro que se pode inferir diretamente dos exemplos de treino.
Embora atualmente as SVM possam ser adaptadas para problemas de regressão (prever o valor de uma variável contínua em vez de se classificar) e para solucionar problemas multi-classe, apresentar-se-à somente a sua formulação original, a qual já abrange uma categoria muito importante dos problemas de previsão — os de classificação binária. Nestes problemas, perante duas classes, uma negativa e outra positiva, o objetivo de uma SVM é encontrar o hiperplano ótimo de separação entre elas. Para o efeito, o treino das SVM envolve a resolução de um problema de otimização quadrática, formulado com o objetivo de maximizar a margem de separação entre os itens das duas diferentes classes. Como se ilustra na Figura 2.12, através dum esquema representativo dum espaço de entrada de duas dimensões (duas variáveis preditivas), existem muitos hiperplanos que podem separar os conjuntos de pontos das duas classes, mas será aquele que possibilitar a maior margem de separação entre as duas classes que à partida oferecerá maior capacidade de generalização — hiperplano ótimo.

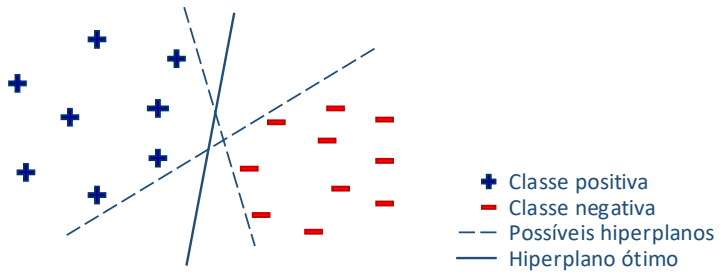


Figura 2.12: Possíveis planos de separação das classes positiva e negativa.

A Figura 2.13 ilustra uma construção geométrica do correspondente hiperplano ótimo num espaço de duas dimensões. Neste caso temos um plano ótimo que garante a máxima separação à

custa de 4 exemplos de treino, 2 negativos e 2 positivos, normalmente designados por vetores de suporte.

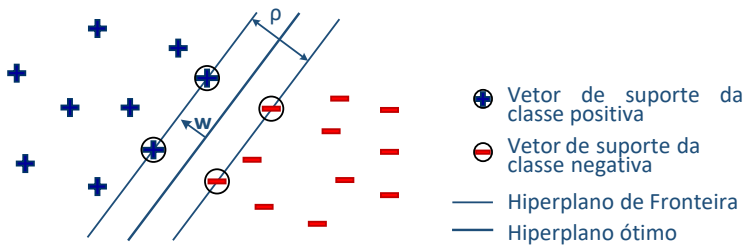


Figura 2.13: Hiperplano ótimo de separação e respetivos vetores de suporte.

A primeira formulação das SVM foi desenvolvida para lidar com dados linearmente separáveis. Uma classificação linear, capaz de lidar com esse tipo de problemas, pode ser representado por uma função linear $f(\mathbf{x})$ que, a partir do conjunto de variáveis explicativas $\mathbf{x}$ (vetor de dimensão d), produza como resultado um valor maior que zero sempre que aos valores observados de $\mathbf{x}$ esteja associada a classe positiva (designação convencionada para uma das duas categorias do classificador e simbolicamente representada $+1$) e um valor menor que zero sempre que, pelo contrário, esteja associada a classe negativa (-1). Essa função linear pode tomar a forma

$$f(\mathbf{x}) = \mathbf{w}^t \cdot \mathbf{x} + b = \sum_{i=1}^{d} w_i x_i + b, \quad (2.2)$$

em que $\mathbf{w}$ é o vetor de pesos (de dimensão d) e b o valor do enviesamento (*bias*), que em conjunto caracterizam o hiperplano ótimo. O vetor de pesos $\mathbf{w}$ define a direção perpendicular ao hiperplano, tal como se ilustra na Figura 2.13, e o parâmetro b tem como influência o deslocamento do hiperplano em direção a uma das classes, movendo-se paralelamente a si próprio.

SVM de margem rígida

Considerando que os dados são linearmente separáveis pode usar-se uma SVM de margem rígida para a classificação. O hiperplano ótimo é definido como

$$\mathbf{w}^t \cdot \mathbf{x} + b = 0. \quad (2.3)$$

Sendo as duas classes de dados completamente separáveis por um hiperplano, à semelhança dos exemplos ilustrados nas Figuras 2.12 e 2.13, é possível encontrar um par $(\mathbf{w}, b)$ que garanta a verificação das 2 inequações que se seguem, para qualquer que seja o conjunto de observações das variáveis explicativas $\mathbf{x}_i$,

$$\begin{aligned} \mathbf{w}^t \cdot \mathbf{x}_i + b \geq +1, \text{ para } y_i = +1 \\ \mathbf{w}^t \cdot \mathbf{x}_i + b \leq -1, \text{ para } y_i = -1 \end{aligned} \quad (2.4)$$

ou a verificação da restrição combinada equivalente

$$y_i(\mathbf{w}^t \cdot \mathbf{x}_i + b) \geq 1, \forall i = 1, \ldots, n \quad (2.5)$$

em que n é o número de observações do conjunto de treino usado para construir o classificador. Os classificadores lineares que separam em dois grupos o conjunto de treino possuem margem positiva, ou seja, garantem que não há nenhum exemplo de treino entre os hiperplanos de

fronteira $\mathbf{w}^t \cdot \mathbf{x} + b = +1$ e $\mathbf{w}^t \cdot \mathbf{x} + b = -1$ (também designados por hiperplanos canónicos e identificados na Figura 2.13 como limites da margem ρ). É devido à existência desta margem de exclusão (de exemplos de treino) que este tipo de classificador se designa por SVM de Margem Rígida.
O treinamento das SVM, tendo como objetivo a maximização da margem de separação ρ entre as duas classes dos exemplos de treino (entre os dois hiperplanos de fronteira), demonstra-se (Silva and Ribeiro [105]) que se atinge esse mesmo objetivo minimizando a norma do vetor de pesos $\mathbf{w}$. Dessa forma, o treinamento envolve o seguinte problema de otimização:

$$\begin{aligned} &\underset{\mathbf{w},b}{\text{Minimizar}} \ \|\mathbf{w}\|^2, \\ &\text{sob as restrições: } y_i(\mathbf{w}^t \cdot \mathbf{x}_i + b) \geq 1, \forall i = 1, \ldots, n. \end{aligned} \tag{2.6}$$

Depois de encontrado o par ótimo $(\mathbf{w}^*, b^*)$ durante a fase de treino, a função linear (2.2) será usada para classificar um qualquer exemplo $\mathbf{z}_j$ do conjunto de teste:

$$y_j = \begin{cases} +1, & \text{se } \mathbf{w}^{*t} \cdot \mathbf{z}_j + b^* \geq 0 \\ -1, & \text{se } \mathbf{w}^{*t} \cdot \mathbf{z}_j + b^* < 0 \end{cases} \tag{2.7}$$

SVM de margem suave

Na maioria dos problemas de classificação a solucionar os dados não são linearmente separáveis. As SVM lineares apresentadas anteriormente são incapazes de lidar com esses dados de treino mais gerais. Podem, no entanto, ser adaptadas de forma a conseguirem lidar também com esse tipo de problemas. É o que acontece com as SVM de Margem Suave, onde as restrições (2.5) que definem a margem de exclusão são relaxadas de modo a permitir que alguns exemplos dos dados de treino possam ficar dentro dessa margem ou mesmo ficarem do lado errado do hiperplano de separação. Isso é conseguido com a introdução de variáveis de tolerância ξ_i:

$$y_i(\mathbf{w}^t \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \ \ \xi_i \geq 0, \ \ \forall i = 1, \ldots, n \tag{2.8}$$

Ainda que as SVM de Margem Suave tolerem que alguns dos exemplos de treino violem a restrição da margem de exclusão (2.5), tentarão naturalmente minimizar a sua ocorrência.
Cada variável ξ_i representa a distância do exemplo "mal comportado" ao hiperplano de fronteira da sua classe. Tal como ilustrado na Figura 2.14, se o seu valor se situar entre 0 e 1 isso significa que o exemplo se encontra posicionado dentro da margem de "exclusão", mas caso supere a unidade, tratar-se-á de um erro de classificação, uma vez que o exemplo estará já do lado errado do hiperplano de separação.

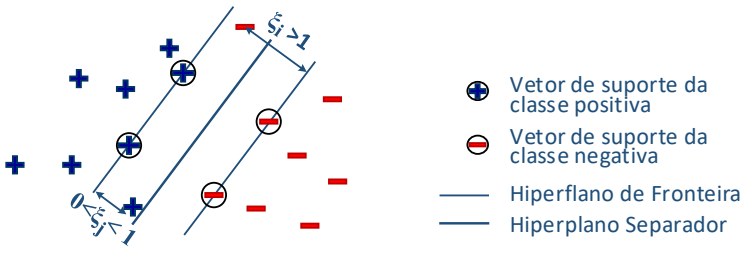


Figura 2.14: Posicionamento de exemplos de treino numa SVM de margem suave.

De forma a poder acomodar a possibilidade de existirem alguns erros de classificação, o pro-

blema de otimização quadrática leva em consideração, como novo termo a minimizar, a soma de todos os desvios ξ_i, passando, por isso, a ser formulado da seguinte forma:

$$\underset{\mathbf{w},b}{\text{Minimizar}}\ \frac{1}{2}\left\|\mathbf{w}\right\|^2 + C\left(\sum_{i=1}^{n}\xi_i\right), \qquad (2.9)$$
$$\text{sob as restrições: } y_i(\mathbf{w}^t\cdot\mathbf{x}_i+b)\geq 1-\xi_i,\ \xi_i\geq 0,\ \forall i=1,\ldots,n$$

onde o parâmetro C é um termo de regularização que atribui um peso à minimização dos erros de classificação em relação ao outro objetivo da otimização — o da maximização da margem de separação. Ou seja, este parâmetro C, que mais à frente nesta tese será também designado por 'parâmetro de custo', permite ao analista controlar a importância relativa que cada um desses dois objetivos — minimização dos erros vs maximização da margem — tem no processo de otimização levado a cabo durante a fase de treino, e controlar também, de alguma forma, a capacidade de generalização do classificador. Uma vez que, para valores pequenos de C as margens são tendencialmente maiores, será de esperar uma maior capacidade de generalização por parte do classificador, ainda que na fase de treino tenha permitido um maior número de erros de classificação.

Quanto à forma de resolver este problema de otimização, refira-se apenas que, à semelhança das SVM de Margem Rígida, passa pela utilização de multiplicadores de Lagrange.

SVM não lineares

Uma grande parte dos problemas do mundo real envolvem dados de treino para os quais não existe um hiperplano separador, por apresentarem estruturas inerentemente não lineares. Ainda que as SVM de margem suave consigam mitigar em parte essa dificuldade, não conseguem lidar bem com conjuntos de dados que tenham uma distribuição altamente não linear. Felizmente, uma característica atrativa das SVM é que podem facilmente ser transformadas em mecanismos de aprendizagem não linear. Para o efeito, as instâncias de entrada são normalmente mapeadas para um espaço de maior dimensão, designado espaço de características, onde já será possível definir hiperplanos que as separe linearmente.

Em geral, as transformações responsáveis pelo mapeamento dos conjuntos de treino, do seu espaço inicial para um novo espaço de dimensão em geral muito elevada, que torne as observações linearmente separáveis em duas classes, podem ser de grande complexidade ou até mesmo inviáveis. As SVM contornam esta dificuldade ao perceberem que a única operação que é necessário realizar no espaço de características é o cálculo de produtos internos e que, para determinados mapeamentos, esses produtos internos podem ser facilmente realizados através de funções conhecidas, designadas funções de Kernel. Representando por $\Phi(\cdot)$ a transformação responsável por um desses mapeamentos, o produto interno entre quaisquer dois vetores $\mathbf{x_i}$ e $\mathbf{x_i}$, depois de mapeados, pode ser obtido aplicando a função Kernel $K(\cdot)$ associada diretamente a esses dois vetores do espaço inicial, tal como se expressa na equação que se segue:

$$\Phi(\mathbf{x_i})^t\cdot\Phi(\mathbf{x}_j)=K(\mathbf{x_i},\mathbf{x}_j). \qquad (2.10)$$

Assim, na implementação das SVM não lineares, opta-se normalmente por um mapeamento a que esteja associada uma função de Kernel conhecida, e que respeite determinadas condições necessárias, não invocadas neste texto. Entre as funções kernel mais usadas, encontram-se a Polinomial, a RBF (*radial basis function*) e a Sigmoidal (redes neuronais sigmoidais), que apre-

sentam a seguinte forma:

$$\begin{aligned} K_{\text{Polinomial}}(\mathbf{x_i}, \mathbf{x}_j) &= \left(\gamma\, \mathbf{x_i}^t \cdot \mathbf{x}_j + k\right)^d \\ K_{\text{RBF}}(\mathbf{x_i}, \mathbf{x}_j) &= \exp\left(-\gamma \left\|\mathbf{x_i} - \mathbf{x}_j\right\|^2\right) \\ K_{\text{Sigmoidal}}(\mathbf{x_i}, \mathbf{x}_j) &= \tanh\left(\gamma\, \mathbf{x_i}^t \cdot \mathbf{x}_j + k\right) \end{aligned} \quad (2.11)$$

Refira-se por fim que para se usar a técnica de mapeamento numa SVM de margem rígida ou de margem suave, basta, no fundo, substituir todas as ocorrências de produtos internos pela função Kernel desejada.

Vantagens e desvantagens das SVM

As *support vector machines* têm recebido crescente atenção pela comunidade de *data mining* em virtude de possuírem características que favorecem a sua robustez e o seu bom desempenho preditivo comparativamente a outros algoritmos classificadores. O facto de serem consideravelmente tolerantes ao ruído, a convexidade do problema de otimização formulado no seu treino (que implica a existência de um único mínimo global), a precisão não depender da dimensão e da dimensionalidade dos dados e uma boa capacidade de generalização, torna-as muito apelativas para aplicações em diversos domínios. Entre as principais desvantagens das *support vector machines* encontram-se as dificuldades de parametrização e de interpretação do modelo criado, o facto de só lidarem com atributos numéricos e o tempo de processamento computacional ser elevado.

2.5 Avaliação de modelos preditivos

Tal como mencionado nas secções anteriores, são vários os métodos e algoritmos de *data mining* que podem ser usados na indução de modelos preditivos a partir de dados. Qualquer que seja o algoritmo selecionado com esse objetivo, surge sempre a necessidade de avaliar quão preciso é o seu desempenho. As métricas de avaliação de desempenho preditivo mais populares em DCBD apresentam-se nas subsecções que se seguem.

2.5.1 Métricas de avaliação de desempenho

Perante modelos preditivos de regressão, as métricas de avaliação de desempenho mais comuns são o coeficiente de determinação (R^2) e o erro da previsão. O R^2 mede a correlação entre os valores observados e os valores preditos e o erro mede o desvio das previsões em relação ao valor efetivo. Uma das medidas de erro mais usadas é o erro quadrático médio (MSE – *mean square error*) — ou a raiz quadrada desse valor (RMSE) —, que representa a média dos quadrados das distâncias entre os valores y_i conhecidos e o valores $\hat{f}(x_i)$ preditos pelo modelo:

$$\text{MSE}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2. \quad (2.12)$$

Perante modelos preditivos de classificação as medidas de avaliação de desempenho mais populares podem ser obtidas através da matriz de confusão. Trata-se de uma tabela que ilustra o número de previsões corretas e incorretas em cada classe. Para um determinado conjunto de dados, as linhas da matriz representam as classes verdadeiras, e as colunas representam as

classes preditas pelo modelo classificador. Logo, cada elemento m_{ij} de uma matriz de confusão apresenta o número de exemplos da classe i classificados como pertencentes à classe j. Para k classes a matriz de confusão tem então dimensão $k \times k$. A diagonal apresenta os casos corretamente classificados pelo modelo, enquanto os outros elementos correspondem aos erros cometidos nas suas previsões. Através da análise da matriz é possível calcular medidas que indicam quais as classes onde o algoritmo tem maior dificuldade de discriminar.
Embora as medidas a apresentar possam ser facilmente generalizadas a problemas multi-classe, por simplicidade, apresentar-se-à uma matriz relacionada com um problema de classificação binária. Considerando que uma classe é denotada positiva (+) e a outra é denotada como negativa (-), a correspondente matriz de confusão vem ilustrada na Tabela 2.1.

Tabela 2.1: Matriz de confusão para um problema de classificação binária.

		classe predita	
		+	-
classe verdadeira	+	VP	FN
	-	FP	VN

Em que:

- Verdadeiros Positivos (VP): representa o número de previsões positivas que estão corretas;
- Falsos Positivos (FP): representa o número de previsões positivas que estão incorretas;
- Falsos Negativos (FN): é o número de previsões negativas que estão incorretas;
- Verdadeiros Negativos (VN): é o número de previsões negativas que estão corretas.

As diferentes medidas de avaliação de desempenho que podem ser obtidas através da matriz de confusão são as seguidamente apresentadas:

- Taxa de Falsos Negativos (TFN): representa a taxa de erro na classe positiva. É uma medida da proporção de exemplos da classe positiva incorretamente classificados pelo preditor.

$$\text{TFN} = \frac{\text{FN}}{\text{VP} + \text{FN}} \quad (2.13)$$

- Taxa de Falsos Positivos (TFP): representa a taxa de erro na classe negativa. É uma medida da proporção de exemplos da classe negativa incorretamente classificados pelo preditor.

$$\text{TFP} = \frac{\text{FP}}{\text{FP} + \text{VN}} \quad (2.14)$$

- Taxa de erro: representa a percentagem de classificações incorretas do total de exemplos n, independentemente da direção do erro.

$$\text{erro} = \frac{\text{FP} + \text{FN}}{n} \quad (2.15)$$

- Taxa de acerto ou acurácia: representa a percentagem de classificações corretas do total de exemplos n, independentemente da direção do acerto. É calculada pela soma dos valores da diagonal principal da matriz, dividida pela soma dos valores de todos os elementos da matriz (n).

$$\text{acurácia} = \frac{\text{VP} + \text{VN}}{n} \quad (2.16)$$

- Precisão: representa a taxa de acerto entre os exemplos classificados pelo preditor como positivos.

$$\text{precisão} = \frac{\text{VP}}{\text{VP} + \text{FP}} \quad (2.17)$$

- Taxa de Verdadeiros Positivos (TVP) ou Sensibilidade (Recall): representa a proporção dos exemplos positivos que foram classificados corretamente pelo preditor.

$$\text{sensibilidade} = \text{recall} = \text{TVP} = \frac{\text{VP}}{\text{VP} + \text{FN}} \quad (2.18)$$

- Taxa de Verdadeiros Negativos (TVN) ou Especificidade: corresponde à taxa de acerto na classe negativa, ou seja, a proporção dos exemplos negativos que foram classificados corretamente pelo preditor. O complementar corresponde à taxa TFP.

$$\text{especificidade} = 1 - \text{TFP} = \frac{\text{VN}}{\text{VN} + \text{FP}} \quad (2.19)$$

- Fmedida: é a média harmónica ponderada da precisão e sensibilidade. Trata-se de uma medida única que valoriza os erros cometidos em qualquer dos sentidos (FP e FN).

$$\text{Fmedida} = \frac{(w+1) \times \text{precisao} \times \text{sensibilidade}}{w \times \text{precisao} + \text{sensibilidade}} \quad (2.20)$$

- F1: Caso os erros sejam valorizados de igual modo, a média armónica Fmedida deixa de ser ponderada ($w = 1$), passando a assumir a forma a que normalmente se designa F1.

$$\text{F1} = \frac{2 \times \text{precisao} \times \text{sensibilidade}}{\text{precisao} + \text{sensibilidade}} \quad (2.21)$$

De realçar que na avaliação de um modelo classificador dever-se-á ter em consideração qual a proporção de dados de cada classe, para se poder avaliar se o dataset é desbalanceado, a fim de escolher aquela que melhor se adequa.
Para além das medidas anteriormente apresentadas por via da da matriz de confusão, o coeficiente *kappa de Cohen* apresenta-se, igualmente, como uma métrica de grande utilidade quando se pretende validar uma classificação. Trata-se de uma medida que indica a concordância de dois examinadores entre as classificações de N elementos em duas categorias mutuamente exclusivas. O coeficiente *kappa de Cohen* pode variar entre 0 e 1, onde o 1 representa a concordância perfeita, valores próximos de zero indicam que a concordância é aquela esperada pelo acaso e valores inferiores a zero indicam que não existe concordância.
Ainda uma outra forma de avaliação dos algoritmos são as curvas ROC (*Receiver Operating Characteristic*), tal como as que se apresentam na Figura 2.15. Uma curva ROC é um gráfico que ilustra o desempenho de um modelo de classificação binário através da variação do limiar de discriminação entre elementos positivos e negativos. Na análise através de curvas ROC considera-se que um classificador é melhor que um outro se o seu ponto no espaço ROC se posiciona acima e à esquerda do ponto correspondente ao segundo classificador. Quando se comparam duas ou mais curvas, se estas não se intercetarem, aquela que mais se aproxima do ponto (0,1) corresponde ao melhor desempenho. No caso de ocorrerem interseções, cada algoritmo tem uma região com melhor desempenho. Também é usual compar o desempenho dos algoritmos em termos de uma medida única extraída da sua curva ROC: a área abaixo da curva ROC, designada, AUC (*Area*

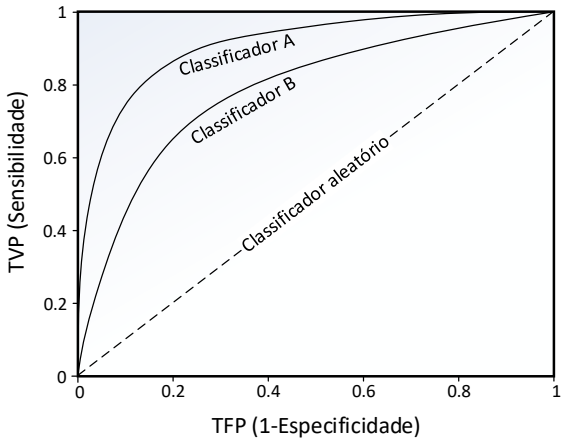


Figura 2.15: Exemplos de curvas ROC.

Under ROC Curve). A medida AUC produz valores entre 0 e 1. Valores mais próximos de 1 são considerados melhores, ou seja, quanto maior for a área sob a curva ROC maior será a precisão do algoritmo.

Cada uma das medidas apresentadas avalia de forma quantitativa um modelo de classificação, providenciando informação sobre a eficácia do método de aprendizagem. Nenhuma delas pode substituir a outra, pelo que, é recomendável que se recorra a várias em simultâneo. Por exemplo, o erro, bem como a acurácia, são medidas simples, mas no caso dos problemas desbalanceados não permitem evidenciar a diferença entre falsos positivos e falsos negativos. Neste caso, quando o número de exemplos de cada classe é muito diferente, é recomendável usar medidas que enfatizem uma ou outra medida de erro (FP ou FN) ou medidas que as combinem, como por exemplo a Fmedida. A literatura sugere que se calculem as medidas mencionadas usando o método de amostragem por validação cruzada, seguidamente apresentado.

2.5.2 Amostragem

Em DCBD, para se obterem estimativas de desempenho preditivo fiáveis, recorre-se a métodos de amostragem para induzir e avaliar a capacidade de generalização de um modelo. Alguns dos principais métodos de amostragem descritos na literatura são a aleatória, *houldout*, *bootstrap* e a validação cruzada *k-folds*. De entre todos elege-se apenas o de amostragem por validação cruzada *k-folds* (*k-folds cross validation*) pelo facto de se tratar do método que garante maior generalização (diminui o *overfitting*), sendo por isso, atualmente, o mais recorrente entre a comunidade científica de *data mining*.

No método de validação cruzada *K-folds* o conjunto de exemplos inicial é dividido em k subconjuntos, aproximadamente com os mesmos tamanhos. Os objetos de $k-1$ partições, designados por conjunto de treino, são utilizados na indução do modelo e na estimação dos seus parâmetros. A partição restante, denominada conjunto de teste, é usada na validação do modelo. O processo é repetido k vezes, utilizando em cada ciclo uma partição diferente para teste, tal como se ilustra na Figura 2.16. O desempenho final do modelo é obtido pela média dos desempenhos observados sobre cada subconjunto de teste, conseguindo-se desta forma uma estimativa de desempenho que se julga mais precisa.

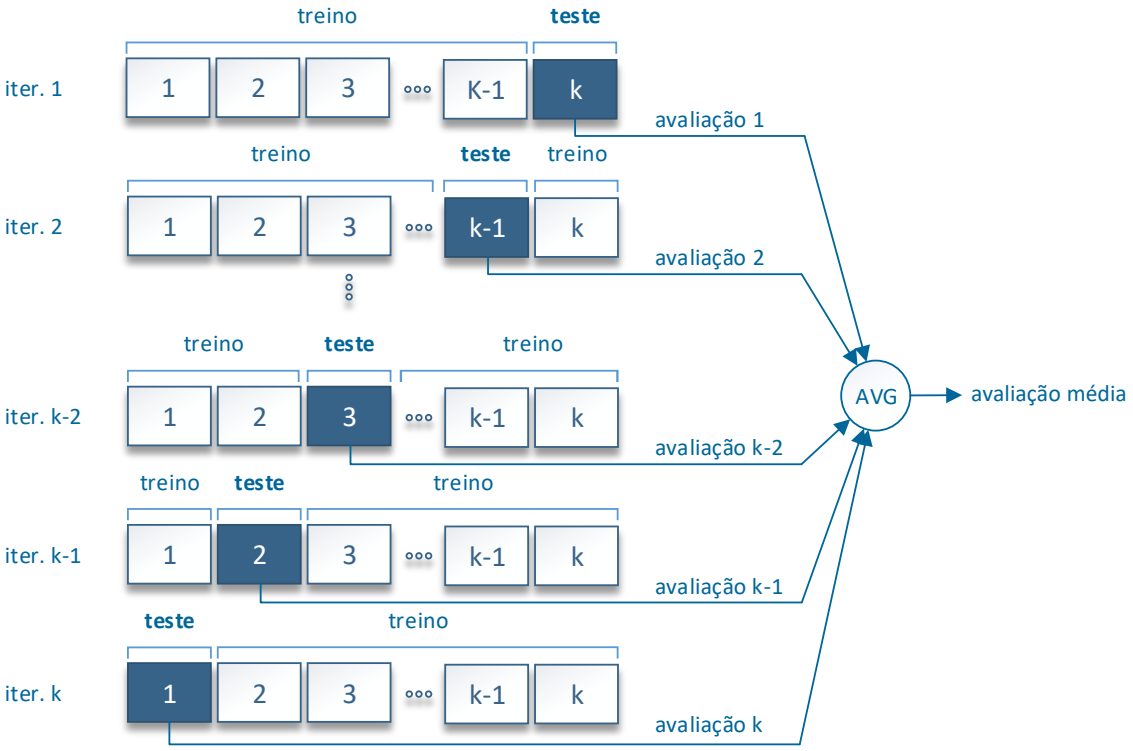


Figura 2.16: Técnica de validação cruzada *K-folds*.

2.6 Resumo e conclusão

As metodologias da DCBD conferem valor estratégico aos dados, transformando-os em conhecimento. Neste capítulo foi descrita, de forma sucinta, a génese da DCBD. Com ênfase na fase do *data mining*, descreveram-se os principais métodos que lhe estão associados, realçando-se os mais populares e recorrentes no contexto da análise e exploração de dados reunidos nas bases de dados dos serviços académicos das instituições educativas. Estes dados, se convenientemente estudados, apresentam grande potencial para fundamentar decisões ao nível dos órgãos de gestão das IES. A literatura demonstra que a previsão, conseguida pelos métodos de classificação e regressão, tem vindo a ser aplicada com sucesso no domínio educacional. Tem-se revelado de grande utilidade na previsão e compreensão do desempenho académico (e.g. sucesso e abandono), bem como na previsão de comportamentos futuros dos estudantes. Os métodos de regressão têm como objetivo prever os valores futuros, ou desconhecidos, de uma ou mais variáveis numéricas contínuas, a partir de outros atributos presentes no conjunto de dados. Portanto, essas técnicas podem ser usadas para prever o sucesso educacional, usualmente quantificado por métricas que envolvem a média do estudante no *terminus* do seu curso, como é feito no estudo descrito no Capítulo 5.

Os métodos de classificação são usados para prever alvos que representem classes de itens de dados predefinidas. Na classificação a variável a prever é uma variável discreta binária ou categórica. Portanto, essas técnicas podem ser usadas para prever se um aluno vai ou não abandonar os seus estudos, como é feito no estudo descrito no Capítulo 6.

Os objetivos de previsão e descrição podem ser alcançados usando uma variedade de métodos e de técnicas de *data mining*. Entre os algoritmos de *data mining* que suportam os métodos de classificação, mais frequentemente usados no domínio de EDM, encontram-se as árvores de decisão, *random forest*, redes neuronais artificiais e as máquinas de vetor de suporte.

Uma decisão crucial para o investigador de *data mining* está relacionada com a seleção dos métodos ou algoritmos mais adequados para induzir modelos analíticos a partir dos dados. Trata-se de uma questão de difícil resposta e nenhuma solução concreta pode ser dada *a priori*. A seleção entre um ou outro método pode trazer ganhos expressivos e ser determinante no sucesso das previsões. As próprias caraterísticas dos métodos existentes, as características dos dados e do problema que se pretende abordar, podem auxiliar na escolha do algoritmo a ser utilizado. A fim de selecionar os mais adequados, para desenvolver modelos analíticos de apoio à gestão de uma IES, apresentaram-se, para cada um deles, as suas vantagens e desvantagens. A relativa robustez em relação ao ruído, a capacidade de aprendizagem e generalização, a parametrização, a interpretabilidade do modelo gerado, a habilidade para tratar com maior exatidão características que tenham dados em falta, a resistência ao *overfitting*, a aptidão para lidar com dados categóricos e contínuos, a possibilidade de estimar a importância das variáveis preditivas, o facto de conseguirem detetar interações entre as variáveis, a rapidez de processamento e ainda a eficiência preditiva não depender da dimensão e dimensionalidade dos dados, são alguns dos critérios que os analistas consideram para eleger o método a usar durante o processo global de extração de conhecimento. Foram também estes os critérios que presidiram à escolha dos métodos e algoritmos a usar no âmbito da investigação inerente a este doutoramento.

Capítulo 3

Educational data mining

3.1 Introdução

Nos últimos anos tem havido um interesse crescente no uso do *data mining* para investigar questões relacionadas com a melhoria dos serviços prestados aos estudantes (e.g. promoção do desempenho académico) a qual é designada *educational data mining* (EDM) (Baker et al. [7]).
O objetivo principal deste capítulo é o de concetualizar e caracterizar o EDM, e com base nos estudos publicados aferir as abordagens, os métodos e as perspetivas atuais e futuras desta área disciplinar.
Para além desta secção introdutória o presente capítulo encontra-se estruturado como se segue. Na Secção 3.2 concetualiza-se e caracteriza-se o EDM. Na Secção 3.3 começa-se por efetuar uma análise às principais revisões sistemáticas de literatura existentes nesta área disciplinar, descrevendo o seu âmbito, os principais objetivos e as respetivas conclusões. Seguidamente, apresenta-se uma análise à literatura sobre a previsão de desempenho académico dos estudantes, com recurso a *data mining*. Na Secção 3.4 efetua-se finalmente uma revisão de literatura sobre os atributos que têm uma relação relevante com os modelos de previsão de desempenho académico com recurso ao *data mining*. Por fim, na Secção 3.5 apresenta-se um breve resumo e as principais conclusões deste capítulo.

3.2 Enquadramento

Nesta secção, além da concetualização e caracterização do EDM, aclara-se o enquadramento onde ele se concretiza e descrevem-se as suas componentes-chave: beneficiários e intervenientes, sistemas educacionais, abordagens ou aplicações e ferramentas ou software. Trata-se de uma descrição importante, ainda que sucinta, uma vez que providencia um esclarecimento acerca do valor agregado ao EDM e o papel que este desempenha, ou que poderá vir a desempenhar, em ambientes educacionais.

3.2.1 Concetualização do EDM

De acordo com o *website*[1] da Comunidade Internacional de *educational data mining* (EDM), o EDM é "uma disciplina emergente que visa o desenvolvimento de métodos que explorem os dados provenientes dos contextos educacionais para melhor entender os alunos e os ambientes em que eles aprendem". Mas a grande quantidade de dados, atualmente reunida nas bases de dados das instituições de ensino, excede a capacidade humana de analisar e extrair informações

[1] www.educationaldatamining.org

úteis, sem a ajuda de técnicas de análise automatizada (Algarni [1]). Por esse motivo, o EDM está relacionado com o desenvolvimento e aplicação de métodos computacionais poderosos, que permitam detetar padrões em grandes coleções de dados educacionais, que de outra forma seriam difíceis, ou mesmo impossíveis, de analisar (Romero et al. [92]). Face ao exposto, poder-se-á afirmar que este recente ramo de investigação assim definido, tem por suporte a aplicação dos métodos e algoritmos do *data mining* a dados provenientes de contextos educativos, para ajudar à melhor compreensão dos processos de ensino, de aprendizagem e de motivação dos estudantes (Baker and Yacef [10], Romero and Ventura [95], Huebner [53], Peña-Ayala [86], Algarni [1]). Por via dessa compreensão, esta área de intervenção que apresenta características muito específicas, visa também o desenvolvimento de modelos que se mostrem capazes de melhorar as experiências de aprendizagem e a eficiência e eficácia das instituições dedicadas ao ensino. De acordo com Romero et al. [92], o EDM apresenta-se como solução capaz de dar resposta à crescente pressão que as instituições de educação, sobretudo as académicas, têm vindo a sentir ao longo dos últimos anos, para se munirem de informação o mais atualizada e completa possível e que seja capaz, entre outras finalidades, de dar resposta à necessidade de sucesso académico dos estudantes e à melhoria contínua dos palcos onde a aprendizagem ocorre. Para isso, o EDM como parte do processo de DCBD, converte os dados brutos, gerados e armazenados em sistemas educacionais, em informações interessantes, úteis e interpretáveis, que podem ajudar a fundamentar a tomada de decisões em prol da promoção do desempenho em instituições dedicadas ao ensino.

3.2.2 Caracterização do EDM

De acordo com Huebner [53] e com Romero and Ventura [95], e tal como se pretende evidenciar através da Figura 3.1, a disciplina EDM está ancorada em várias disciplinas de referência, como é o caso dos sistemas de informação, da análise visual de dados, do *data mining*, da psicopedagogia, da psicologia cognitiva e da psicometria, e pode ser desenhada como o resultado de três áreas principais, que são as ciência da computação, a educação e a estatística.

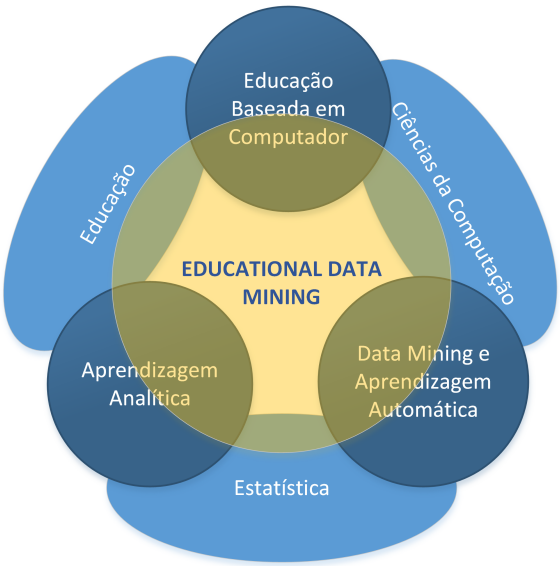


Figura 3.1: Principais áreas relacionadas com o *educational data mining* (adaptada de Romero and Ventura [95]).

A literatura refere que há cinco componentes-chave que caracterizam o EDM. São a diversidade

dos sistemas educacionais e respetivos dados analisados, os beneficiários e intervenientes no EDM, as tarefas associadas ao EDM, os métodos e algoritmos de *data mining* usados para implementar as tarefas de EDM e, por último, as ferramentas informáticas, ou software, usados para tratar os dados. Os métodos e algoritmos de *data mining* foram previamente descritos no Capítulo 2. Nesta subsecção, aclara-se o que se entende sobre cada um dos outros tópicos, para melhor perceção e interpretação dos conteúdos apresentados ao longo do presente capítulo.

Sistemas educacionais

O uso crescente da tecnologia nos processos de ensino e a evolução permanente dos sistemas informáticos, propiciou que se gerassem e armazenassem grandes volumes de dados digitais nos diversos sistemas educacionais. De acordo com Romero and Ventura [94], são três os principais sistemas educacionais de produção de dados e que, ao mesmo tempo, deles beneficiam:

- O primeiro é o ensino presencial tradicional, que designam como ensino offline. Este sistema caracteriza-se pelo contacto direto entre o professor e o aluno, sendo concretizado numa metodologia "face-to-face" (Romero and Ventura [94, p. 1]) e tem como mais valia o facto de se constituir como o campo, por excelência, do estudo das componentes psicológicas que induzem à apreensão de conhecimentos.
- O segundo sistema de ensino que os autores mencionam, é o *E-learning* e o *Learning Management System* (LMS), que além de providenciar instruções aos educandos, é também um veículo de comunicação, colaboração e administração de ferramentas tecnológicas, que garantem o funcionamento do sistema e que agregam a vantagem de poder reportar *feedbacks*, acerca do desempenho dos estudantes, que depois podem ser transportados para as bases de dados. Neste sistema, os professores não conseguem percecionar, de forma fácil, como os alunos se comportam e aprendem.
- Por último, o terceiro sistema educacional é denominado de *Intelligent Tutoring System* (ITS) e *Adaptive Educational Hypermedia System* (AEHS). Este é definido como um sistema alternativo ao método "tradicional" de colocar informações em rede sem grandes critérios associados, mas visa uma aproximação efetiva entre os conceitos a ensinar aos estudantes e as necessidades ou dificuldades de cada um deles.

É desta diversidade de sistemas educacionais de onde emanam os dados que se alojam nas plataformas que as ferramentas de *data mining* exploram.
Com base nos três sistemas educacionais acima referidos, Romero and Ventura [94] agruparam os estudos de EDM de acordo com o tipo de dados utilizados: ensino presencial tradicional, ensino à distância (e-learning), sistemas de gestão de aprendizagem, sistemas tutoriais inteligentes, sistemas educacionais adaptativos, questionários de testes e conteúdos dos textos.

Tarefas, abordagens ou tópicos de investigação de EDM

A finalidade com que o *data mining* é usado em ambientes educacionais é designada como tarefa, abordagem ou tópico de investigação de EDM. Recentemente, Bakhshinategh et al. [11] propõe uma nova taxonomia para categorizar as diferentes abordagens existentes na literatura de EDM. Tal como se observa na Figura 3.2, de acordo com o objetivo final do estudo, são 5 as categorias e 12 as subcategorias identificadas na taxonomia proposta. As referidas categorias caracterizam-se, de forma muito sucinta, de acordo com o que se segue (Bakhshinategh et al. [11]):

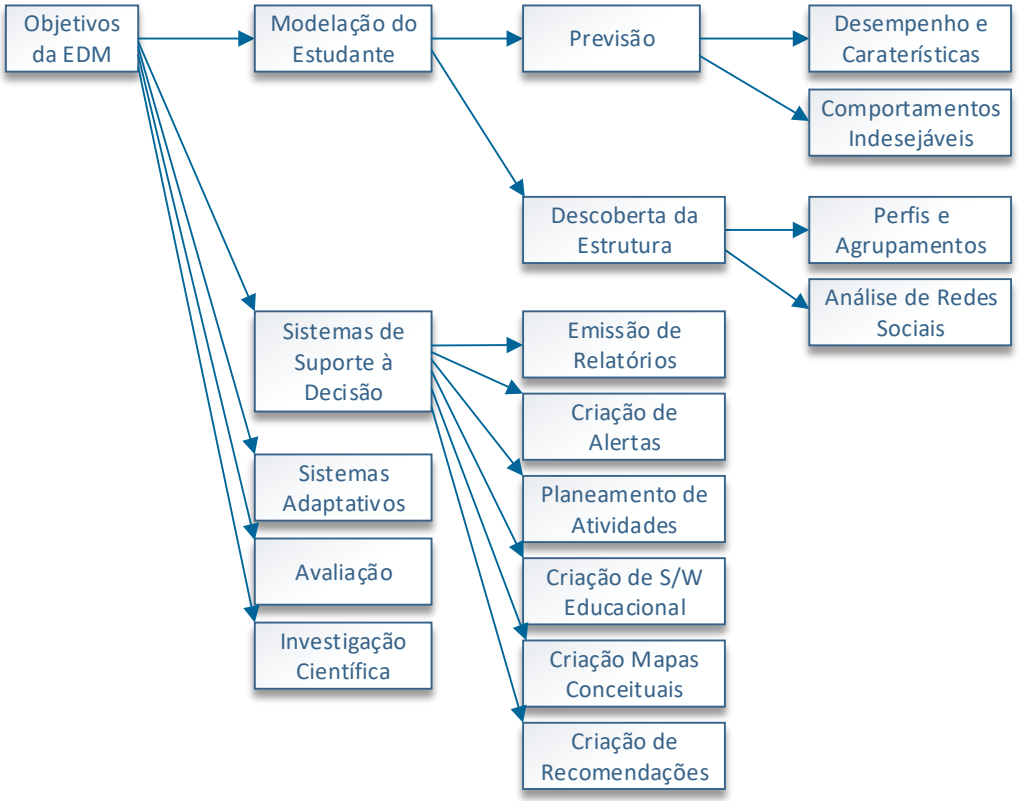


Figura 3.2: Taxonomia das aplicações do EDM (adaptada de Bakhshinategh et al. [11]).

- Modelação do Estudante: os estudos desta categoria têm como principal objetivo representar aspetos cognitivos que caracterizam o estudante. Por exemplo, representar emoções, o conhecimento de domínio, as estratégias de aprendizagem ou as habilidades a serem treinadas. Através dos estudos desta categoria é possível identificar, com a devida antecedência, os estudantes que necessitam de ajuda, tais como os de baixa motivação, ou com propensão ao insucesso ou abandono académico. São 2 as subcategorias na modelação de estudantes: previsão e descoberta de estrutura.

  - Na previsão, geralmente, é conhecido qual o atributo específico a prever. As variáveis alvo de previsão estão normalmente relacionadas com o desempenho académico ou com as características comportamentais dos estudantes, tais como: a média final de curso, a classificação em determinadas unidades curriculares, o abandono e o (in)sucesso académico, a participação em jogos de azar, hábitos de sono, a propensão para a entreajuda e para o trabalho corporativo.
  - Na descoberta da estrutura, o atributo a prever pode ser desconhecido, ou então, pode ser definido como uma estrutura, em vez de uma única propriedade. Por exemplo, quando se estima o perfil dos alunos com base em diferentes variáveis (e.g. comportamentais ou de conhecimentos), ou quando se usam essas informações para agrupar os alunos para diversos fins. Um outro exemplo é quando se fazem análises de redes sociais: neste caso o principal objetivo é obter um modelo sobre os alunos na forma de um grafo, que mostre diferentes relações possíveis entre eles. Por exemplo, a colaboração é uma propriedade atribuída à relação entre indivíduos e, para ser estudada, é necessário modelar os relacionamentos e os indivíduos.

- Sistemas de Suporte à Decisão: os estudos que se enquadram nesta categoria têm como

principal objetivo ajudar a aprimorar o processo de aprendizagem, para apoiar os intervenientes na tomada de decisões. São 6 as subcategorias onde se enquadram os estudos desta tipologia:

- Emissão de relatórios: o objetivo destes estudos é encontrar e destacar as informações relacionadas com as atividades do curso, que podem ser úteis para educadores e responsáveis pelos órgãos de gestão das escolas. Exemplos disso são: fornecer *feedback* sobre o desempenho ou sobre as características dos alunos, descrever as conexões e colaborações por meio de análise de redes sociais e criar relatórios a partir das informações de perfil extraídas com a ajuda de métodos de criação de perfil.
- Criação de alertas para os intervenientes no sistema educativo: os estudos desta subcategoria têm como objetivo principal prever as características dos alunos e detetar comportamentos indesejáveis, para criar alertas em tempo real. Exemplos de situações em que os alertas podem ser necessários são, por exemplo, os casos de baixa motivação, uso indevido de recursos e outras práticas pouco éticas.
- Planeamento de atividades: o objetivo dos estudos desta subcategoria é ajudar os intervenientes no planeamento das atividades. Por exemplo, ajudar os professores e administradores no planeamento de cursos futuros ou alocação de recursos.
- Criação de software educacional: o objetivo é ajudar o educador a criar ou desenvolver material didático do curso, de forma automática, usando as informações de uso do aluno. Por exemplo, fornecer conteúdo, vídeos, testes e outros materiais didáticos de aprendizagem.
- Desenvolvimento de mapas conceituais: os estudos desta categoria têm como objetivo desenvolver mapas conceituais, de vários aspetos, para ajudar os educadores a definir o processo de educação. Exemplos de mapas conceituais são: hierarquia de tópicos no material do curso, relações de habilidades e itens de teste, correlação de itens de teste e componentes de conhecimento.
- Criação de recomendações: o objetivo é poder fazer recomendações diretamente aos alunos em relação às suas atividades, tal como indicar *links* para visitas, ou solicitar trabalhos e ainda para poder adaptar conteúdos de aprendizagem. Os métodos mais comuns em sistemas de recomendação são filtragem colaborativa, métodos baseados em conteúdo, algoritmos baseados em regras de associação e abordagens híbridas. Outro método de gerar recomendações é usar a descoberta com modelos.

- Sistemas adaptativos: esta categoria de estudos está relacionada com o uso de sistemas inteligentes em contextos de aprendizagem baseada em computador, no qual seja necessário que o sistema se adapte às necessidades ou ao comportamento especifico do utilizador. Por exemplo, quando é necessário adaptar o material didático do curso, ou quando é necessário emitir conselhos em função do comportamento dos alunos.

- Avaliação: os estudos desta categoria têm como objetivo ajudar os educadores a diferenciar a proficiência dos alunos em níveis mais detalhados, por meio de métodos estáticos e dinâmicos, tais como testes e avaliações *online* e *offline*.

- Investigação científica: os estudos desta categoria têm como objetivo contribuir para o progresso do conhecimento científico relacionado com as aprendizagens dos estudantes, através do desenvolvimento de modelos sobre os alunos. Para o efeito, teorias e hipóteses

sobre o processo de aprendizagem e respetivas melhorias são estudadas na educação com recurso ao *data mining*.

Beneficiários e intervenientes do EDM

De acordo com Romero and Ventura [95], o EDM pode providenciar suporte aos principais intervenientes nos contextos educativos:

- Estudantes: poderão obter apoio em reflexões sobre a sua situação, *feedbacks* ou recomendações de boas práticas personalizadas, que respondam às suas necessidades e contribuam para melhorar o seu desempenho na aprendizagem.
- Docentes: poderão compreender melhor os processos de aprendizagem dos seus alunos, refletir sobre os seus próprios métodos de ensino e avaliar a sua eficácia na aprendizagem, ao usar configurações diferentes e metodologias de ensino diversificadas.
- Investigadores: desenvolvem e comparam métodos e algoritmos de prospeção de dados, de forma a que sejam encontrados, para cada abordagem ou problema educacional específico, aqueles que melhor se adequem.
- Administradores: com base em informação, poderão avaliar a melhor maneira de otimizar os recursos institucionais (humanos e materiais), as ofertas educativas, permitem elaborar planos de estudos e definir estratégias que visem a promoção educacional dos discentes.
- Organizações Educativas: contribuí para o aperfeiçoamento e agilização dos processos de tomada de decisão, como na identificação de melhores caminhos de custo-benefício para a promoção do desempenho.

Ferramentas ou software de EDM

Existe um vasto conjunto de ferramentas informáticas, ou software, que implementam os métodos e algoritmos de *data mining*. No *website www.Kdnuggts.com*[2] encontra-se uma listagem de todas essas ferramentas. Algumas são comerciais e outras são de acesso livre.
Slater et al. [106] divulgaram um estudo de investigação, onde descreveram 40 ferramentas informáticas que emergiram para pesquisa e prática de *data mining*. As ferramentas R, RapidMiner, WEKA, KEEL, KNIME, Orange e SPSS são apontadas, por estes autores, como as mais populares para a comunidade do EDM. Todas apresentam a particularidade de oferecem uma ampla gama de algoritmos e estruturas de modelação, que podem ser usados para prever processos e relacionamentos em dados educacionais. Slater et al. [106] advertem que é difícil antever qual das ferramentas será a ideal para conduzir todo o processo DCBD, desde o início ao fim, principalmente se o investigador não apresentar uma vasta experiência na área. Segundo estes autores, para selecionar uma delas de forma adequada, terá que ser considerado o perfil do analista de dados e também será necessário ter um bom entendimento sobre os objetivos a atingir. A principal dificuldade apontada por diversos autores (Romero and Ventura [93], Slater et al. [106], Sukhija et al. [110]) no uso destas ferramentas no contexto do EDM, prende-se com o facto dos dados, na sua origem, não estarem disponíveis num formato próprio para análise através destes softwares. Perante esta realidade, fica claro que são necessários

[2]Portal que disponibiliza informação sobre a DCBD/*data mining*, incluindo software, casos de estudo e sondagens.

mais desenvolvimentos, de modo a tornar as ferramentas de EDM mais acessíveis a utilizadores não especializados (Huebner [53]).
Informações adicionais acerca de algumas das especificidades das ferramentas mencionadas, como, por exemplo, formas de visualizar a informação, podem ser obtidas em Slater et al. [106] e em Camilo and Silva [18].

3.3 Trabalhos relacionados

Tal como mencionado na secção anterior, são diversas as categorias de abordagens de EDM descritas na literatura. Nesta secção, com o intuito de identificar quais as tendências de investigação atuais e futuras desta emergente área de investigação, começar-se-à por apresentar, em 3.3.1, uma análise exaustiva sobre as principais revisões sistemáticas da literatura existentes — para uma leitura mais rápida da análise efetuada, apresenta-se, no Apêndice A desta tese, a Tabela A.1, com a síntese dessa análise, descrevendo o âmbito, objetivos e conclusões das obras consultadas. Seguidamente, a fim de se obterem orientações para o tema central da investigação de doutoramento, apresentar-se-à em 3.3.2 uma revisão de literatura com ênfase na previsão do sucesso académico dos estudantes, com recurso ao *data mining*, e em 3.3.3 uma revisão com o foco na previsão do abandono escolar. Por fim, em 3.3.4, com o intuito de se demonstrar a aptidão e a contribuição do EDM na promoção da qualidade dos serviços prestados aos estudantes, apresentam-se dois estudos que se enquadram na categoria dos sistemas de apoio à decisão, um relacionado com os sistemas de recomendação pedagógica e o outro com ambientes pessoais de aprendizagem.

3.3.1 Revisões sistemáticas de literatura

As principais revisões sistemáticas da literatura de EDM (Romero and Ventura [93, 94, 95], Baker and Yacef [10], Huebner [53], Mohamad and Tasir [76], Papamitsiou and Economides [84], Peña-Ayala [86], Sukhija et al. [110], Algarni [1], Bakhshinategh et al. [11]), no seu conjunto, comprovam a importância crescente do EDM ao longo do tempo, referenciam e analisam os principais tópicos de investigação onde o EDM tem demonstrado um notável contributo, como instrumento de análise e apoio à gestão. Evidenciam, igualmente, a utilidade, o potencial e a eficácia dos métodos e algoritmos de *data mining* mais usados no âmbito dos estudos desenvolvidos. Uma análise aos principais estudos subordinados ao estado da arte é seguidamente apresentada.
A mais antiga revisão sistemática desta coleção foi publicada no ano de 2007, altura em que Romero and Ventura [93], considerados os padrinhos do EDM, pretenderam divulgar a aptidão do *data mining* na área da educação. Discorrendo sobre as tendências dos métodos de *data mining*, sobre a tipologia de dados utilizados, sobre as ferramentas mais comuns para os analisar e sobre os respetivos resultados educacionais, os autores começam por esclarecer que os métodos inerentes a esta área do conhecimento podem ser aplicados a dados provenientes dos sistemas educacionais de ensino presencial tradicional e de educação à distância, pela web. Para isso, os mesmos autores advertem que é necessário adequar separadamente a aplicação desses métodos, devido às especificidades desses sistemas educacionais, nomeadamente, pelo facto de que têm diferentes fontes de dados e objetivos (Romero and Ventura [93, p. 137]). Neste contexto, os autores explicam que ainda que a maioria dos métodos tradicionais do DM possam ser aplicadas a dados educacionais, outros têm que ser adaptados à situação concreta a avaliar e ao problema para o qual se pretende encontrar solução. Prosseguindo esta análise, os autores indicaram que

tipo de dados podem ser recolhidos, como podem ser pré-processados, como se podem aplicar os métodos e algoritmos de *data mining* e, por fim, ainda mostraram como se pode beneficiar do conhecimento descoberto por esta via.

Os 81 artigos, incluídos e analisados por Romero and Ventura [93] nesta revisão literária, foram agrupados de acordo com os métodos de *data mining* usados pelos investigadores e também em função dos sistemas educacionais intervencionados. Relativamente aos métodos de *data mining*, a que os investigadores mais recorreram, verifica-se que prevalecem a associação (43 artigos), a classificação e a visualização (28 artigos), embora apareçam alguns estudos que também comportam o *clustering* e o *text mining*. Ao nível da proveniência dos estudos em função dos sistemas educacionais intervencionados, é muito notória a predominância do sistema de ensino de educação à distância, em comparação com o número de estudos associados ao ensino presencial tradicional. Esta discrepância, evidencia que a adoção do EDM, na resolução de questões educativas nas instituições de ensino tradicional presencial, ainda está numa fase muito incipiente. Romero and Ventura [93] concluem que até à consolidação e sucesso desta jovem área de intervenção, muitos desenvolvimentos e trabalhos relacionados serão precisos, para que se alcance o nível de outras áreas, onde as abordagens de *data mining* estão já plenamente implantadas e consolidadas, como por exemplo, na medicina e no comércio eletrónico. Neste sentido, as linhas orientadoras para investigações futuras, cingem-se ao ajustamento dos métodos e algoritmos tradicionais de *data mining* e à necessidade de potenciar a evolução das ferramentas informáticas disponíveis para tratamento de dados, pois, segundo os autores, conviria que apresentassem uma interface mais intuitiva para o utilizador, de forma a que se promova uma utilização mais amigável (Romero and Ventura [93]).

Em 2009, Baker and Yacef [10], respetivamente o atual presidente da comunidade internacional do *educational data mining* e uma docente da Universidade de Sydney, instituição que edita o *International Journal of Educational Datamining*, efetuaram uma revisão da literatura relevante, tendo como linhas orientadoras, as áreas e as tendências de evolução do EDM, ao longo dos primeiros anos de aplicação das metodologias do *data mining* na educação. Ainda que já se tenham passado alguns anos após a publicação deste artigo de revisão, e o EDM se encontre atualmente numa situação de evolução muito diferente da reportada por via deste trabalho, decidiu-se pela sua integração na revisão bibliográfica desta tese por dois motivos. Primeiramente, favorece a compreensão da importância das investigações efetuadas e do corpo científico constituído para a evolução galopante do EDM. Numa segunda perspetiva, também deixa explícita a ligação intrínseca entre a evolução do EDM e os desenvolvimentos que a tecnologia tem vindo a permitir, na construção das bases de dados. Acresce ainda um outro argumento que também sustenta a integração deste artigo nesta revisão literária, que é o facto de se tratar do estudo de abertura do primeiro número do *Journal of Educational Data Mining*.

De acordo com o reportado por Baker and Yacef [10], antes do final da última década do século passado, o EDM socorria-se de sete métodos diferentes de *data mining*: a estatística, a visualização, a *web mining*, o *clustering*, a classificação, a associação e o *text mining*. Tal como apresentado para os métodos de *data mining* usados em contextos educativos, os autores quantificaram também as tarefas potenciais onde o EDM podia intervir, tendo concluído que eram quatro, até àquela altura. São elas, a previsão, o *clustering*, a associação e a destilação de dados para julgamento humano. No entanto, como sublinham Baker and Yacef [10], começava a emergir, nessa altura, uma nova estratégia de abordagem ao EDM que, introduzida como meio complementar a categorias mais populares, vinha servir como meio de confirmação dos resultados de análises mais sofisticadas, como, por exemplo, aquelas que envolviam subcategorias de materiais de aprendizagem ou que versavam sobre os diferentes tipos de comportamento

dos alunos. Era o início da descoberta através de novos modelos. A partir desta abordagem, regressiva à própria história científica do EDM, os autores também sublinharam que, naquele momento, se verificava que estas novas metodologias de análise a grandes bases de dados começava a ser usada para induzir modelos representativos da heterogeneidade de estudantes.

Sobressai da revisão de literatura elaborada por Baker and Yacef [10] a ideia da importância crescente do EDM na conquista de novas áreas de investigação, sendo exemplo fulcral deste movimento o *gaming the system* uma vez que a metodologia EDM, através de estudos também liderados por Baker et al. [9], demonstrou ter potencialidades para construir informação concreta, quantitativa e muito detalhada, dos estudantes que constroem o seu percurso académico em ambientes de aprendizagem interativos. Através deste estudo comprova-se a importância crescente do EDM ao longo do tempo e, da mesma forma, justifica-se a sua versatilidade e capacidade de adaptação à constante evolução tecnológica e ao contínuo aparecimento de bases de dados cada vez maiores e mais complexas.

Um outro estudo da autoria de Romero and Ventura [94], publicado em 2010, é uma versão melhorada, mais atualizada e mais abrangente, do estudo de Romero and Ventura [93], tanto no conteúdo abordado quanto na dimensão do conjunto de artigos considerados para a revisão. É apresentada uma ampla abordagem, que prima pela incidência e aprofundamento de conceitos úteis para a compreensão do EDM e também para a aferição da sua importância no contexto atual da educação. Os autores começam por aclarar o enquadramento onde o EDM se concretiza, aludindo às suas comunidades de pesquisa e respetivos sistemas educacionais.

Posteriormente, no intuito de evidenciar as inovações que o EDM tem vindo a promover, os autores deste artigo de revisão, reuniram 306 artigos científicos, dos quais 81 coincidem com os integrados no estudo de Romero and Ventura [93], publicados desde 1993 até ao ano de 2009, que subdividiram cronologicamente. Este vasto conjunto de artigos científicos, foi também agrupado e analisado de acordo com outros dois critérios: em função da proveniência do sistema educacional intervencionado e também em conformidade com as tarefas de EDM. Da contagem obtida e da abordagem ao estado da arte anterior a 2005, Romero and Ventura [94] sublinham que as comunidades de EDM mais ativas continuam a ser o *e-learning*, LMS, ITS/AEHS e que se verificou um crescimento exponencial de estudos e publicações, destacando o aparecimento do *Journal of Educational Data Mining* como meio difusor dos trabalhos relacionados com o EDM.

Aludindo à tipologia dos sistemas educacionais intervencionados, os autores revelam que as análises psicométricas e as análises relativas ao ensino presencial, comportam, sobretudo, estudos relacionados com o comportamento, o desempenho dos estudantes e os seus currículos. Ao contexto de ensino *e-learning*, os autores atribuem o surgimento das técnicas de *Web Mining* (WM), aplicadas aos dados armazenados acerca dos estudantes.

Relativamente ao ensino em contexto ITS e HST, é referido que também aos dados que daí provêm têm vindo a ser aplicadas metodologias de *data mining*, sendo que, neste contexto em particular, as técnicas mais usadas até ao momento têm sido os *log files* e os *user models*. Neste trabalho de revisão, os autores também deram particular atenção às inovações metodológicas que têm vindo a ser introduzidas no EDM. Por exemplo, a integração das estruturas de modelagem psicométricas nesta área de estudos, que Romero and Ventura [94], indo ao encontro da opinião Baker and Yacef [10], consideram ser categorias de DM pouco usuais e nem sempre consideradas como tal. No entanto, tendo em conta as já referidas características da educação, os métodos de *data mining* ganham nova perspetiva e, muitas vezes, torna-se necessário aprofundar estudos, com vista ao seu desenvolvimento, tendo como objetivo maior o cumprimento de tarefas que, também elas, têm vindo a ser alvo de evolução. Tal como referido em Baker and Yacef [10], as quatro principais tarefas do EDM eram a melhoria dos modelos de ensino, a

melhoria dos modelos dominantes, o estudo do suporte pedagógico que o software atual proporciona e a investigação científica com vista ao desenvolvimento do ensino e dos estudantes. Mas, tal como Romero and Ventura [94] fazem ver neste estudo de revisão, outros autores, como por exemplo Castro et al. [20] defendiam a ideia de que o EDM poderia ser usado em mais tarefas. Para Castro et al. [20] as aplicações de *data mining* na educação poderiam facilitar a avaliação do desempenho académico dos alunos e do seu grau de aprendizagem, permitiriam igualmente analisar evoluções ou retrocessos nas aprendizagens dos alunos inscritos em cursos de *e-learning*, serviriam também para a criação de relatórios de *feedback*, tanto para professores quanto para alunos, das aprendizagens feitas em plataformas eletrónicas e ainda teriam um papel crucial na identificação de comportamentos atípicos, ou indesejáveis, dos estudantes. Romero and Ventura [94] lembram as tarefas de EDM apontadas nestes dois estudos, para introduzir a sua própria ideia acerca desta temática e defender que o EDM serve para muito mais do que defende Baker and Yacef [10] ou mesmo Castro et al. [20]. Neste sentido, baseados na revisão de literatura suportada pelos 306 artigos que analisaram, apontaram as suas próprias tarefas que são: a criação de *feedbacks*; o favorecimento da criação de recomendações; a previsão da *performance* do aluno; a análise e a visualização de dados; a construção de modelos sobre os estudantes; a possibilidade de agrupar estudantes em função de determinadas características; a possibilidade de detetar tipos de comportamentos dos estudantes; a análise do comportamento na rede social onde o estudante se insere; o planeamento e calendarização de novas unidades curriculares e a construção de ferramentas eletrónicas direcionadas para a educação e o desenvolvimento de conceitos. Romero and Ventura [94] fundamentaram a afirmação sobre as tarefas de EDM por si apontadas, com base nas investigações que os autores dos 306 estudos visaram, referindo que, das mesmas, resultaram pelo menos 23 artigos para cada uma das primeiras 8 categorias mencionadas. As 4 últimas categorias foram as que tiveram menos referências, pautando-se todas elas por um número inferior a 15. De acordo com a explicação dos autores deste artigo de revisão de literatura, as preferências dos investigadores acerca do uso dado ao EDM, podem ter como justificação o facto das primeiras categorias serem mais antigas, portanto mais divulgadas, mas também salientou as preferências dos investigadores em relação às categorias mais recentes, sublinhando que as análises às redes sociais são as que reúnem o maior número de artigos entre as 4 menos referenciadas.

Para concretizar todas estas tarefas de EDM, Romero and Ventura [94] mencionam que os métodos de EDM mais usados continuam a ser a regressão, a classificação, a segmentação e a associação, sendo que os algoritmos mais usados são as árvores de decisão, as redes neuronais e as redes *bayesianas*. Por fim, antes de apresentarem os resultados que sobressaíram da sua revisão do "estado da arte", os autores citados corroboraram, uma vez mais, algo que já haviam assinalado no estudo de Romero and Ventura [93], como, por exemplo, a necessidade de ajustar aos dados de contextos educativos os métodos e algoritmos tradicionais de *data mining*, a necessidade de aperfeiçoar ferramentas informáticas específicas, especialmente desenvolvidas para analisar dados de contextos educativos e a necessidade de normalizar os dados de *input*, por forma a facilitar as tarefas de pré-processamento, de *data mining* e pós-processamento, da DCBD de sistemas educacionais. Ainda assim, devido ao elevado número de artigos científicos de EDM que têm progressivamente sido divulgados a cada ano que passa, Romero and Ventura [94] concluem que embora o EDM já não esteja a dar os primeiros passos, mas sim a aproximar-se da sua adolescência, carece ainda de muitos desenvolvimentos e investigações futuras, por forma a consolidar-se como uma área de investigação madura. De facto, tal como os autores lembram, o EDM na atualidade já não se restringe apenas ao uso ou benefício de professores e estudantes, uma vez que tem utilidade para as próprias instituições educativas, para os profis-

sionais responsáveis pela criação e desenvolvimento de planos de estudos e até mesmo para os Estados.

Outra revisão de literatura subscrita por Huebner [53] em 2013, abarca exclusivamente tópicos como a retenção e o abandono escolares, os sistemas pessoais de recomendação em contextos educativos e as formas como o *data mining* tem sido usado na análise de dados quando se pretende melhorar o sucesso dos aprendizes e os processos diretamente ligados à aprendizagem. Advogando que o tipo de pesquisa realizado no contexto de EDM é maioritariamente quantitativo, dada a utilização de técnicas de estatística, *machine learning* e inteligência artificial, os autores justificam que os estudos de previsão, de segmentação, de classificação e de associação, são o paradigma dominante nos modelos analíticos desenvolvidos no contexto das tarefas de EDM acima mencionadas. Indo de encontro à opinião de Romero and Ventura [94], Huebner [53] no seu estudo de revisão, também realça que a investigação no campo do EDM precisa de estudar formas de tornar os resultados do *data mining* mais generalizáveis, providenciando o desenvolvimento de modelos que possam ser usados em múltiplos contextos. Aponta ainda outras duas possíveis linhas para futuros trabalhos de investigação. Uma está relacionada com o desenvolvimento de sistemas de apoio à decisão e de sistemas de recomendação que minimizem a intervenção dos educadores. A outra está relacionada com o desenvolvimento de ferramentas que protejam a privacidade individual dos intervenientes ao mesmo tempo que possibilitam o EDM.

Um outro estudo também publicado em 2013, da autoria dos já citados Romero and Ventura [95], traz para o contexto do tema em estudo elementos de destaque face a outros estudos de revisão já aqui analisados. Visando estrategicamente a construção de um guia para quem pretenda aprofundar conhecimentos nesta área, o documento em análise fornece uma visão atualizada do estado atual dos conhecimentos em EDM. Os autores começaram por referenciar e analisar os principais tópicos de investigação na área de EDM e explicar o processo específico de descoberta de conhecimento em EDM. Neste contexto, os exemplos apontados pelos autores foram: (re)organização das aulas ou da avaliação, a colocação de materiais com base no uso e dados de desempenho; a identificação daqueles que poderão beneficiar de comentários, conselhos de estudo ou outros géneros de ajuda; a delineação de estratégias para ajudar os alunos a encontrar e pesquisar materiais e bibliografia útil.

Depois de referirem os principais tópicos de investigação do EDM, Romero and Ventura [95] enfatizaram que mais recentemente surgiram muitos outros objetivos específicos nesta área de intervenção, dependentes do ponto de vista do utilizador final e do problema a resolver. Alguns dos exemplos apontados pelos autores como problemas particulares foram: a criação de grupos de estudantes de acordo com as suas características pessoais e respetivos dados pessoais de aprendizagem; modelação de alunos para desenvolver e ajustar os seus modelos cognitivos; construção de material didático para ajudar os instrutores e administradores no planeamento de cursos futuros; a estimativa dos parâmetros dos modelos probabilísticos de dados para prever a probabilidade de eventos de interesse.

Para a concretização de todas estas tarefas de EDM, os autores indicam que os métodos mais populares continuam a ser a maioria das técnicas mais tradicionais de *data mining*, como a classificação e o *clustering*, apresentando todas elas um sucesso já comprovado no domínio educacional. No entanto, em virtude das características especiais dos dados únicos dos sistemas educacionais, Romero and Ventura [95] referem que outras técnicas menos comuns e mais apropriadas para o EDM foram surgindo. Por exemplo, análises de associação têm sido usadas para identificar relacionamentos nos padrões de comportamento e diagnóstico dos estudantes com dificuldades de aprendizagem; a análise de rede social pode ser usada para interpretar e

analisar a estrutura e relações em tarefas colaborativas e interações com ferramentas de comunicação; a deteção de *outliers* pode ser usada para identificar estudantes com dificuldades de aprendizagem, desvios nas ações ou comportamentos do aluno ou mesmo do educador e também para detetar processos de aprendizagem irregulares; a destilação de dados para julgamento humano pode ser usada para ajudar os educadores a visualizar e analisar os alunos em atividades de curso e informações de uso. Para linhas de investigação futura, Romero and Ventura [95] perspetivam que os educadores e as instituições desenvolvam uma cultura baseada em dados, a fim de serem melhoradas as decisões que promovam a eficiência das instituições dedicadas ao ensino. Como uma consequência prática desta necessidade, os investigadores demonstram cada vez mais interesse em repositórios de dados abertos e formatos de dados padrão para promover a troca de dados e modelos.

Num outro estudo, partindo da definição avançada em 2013 por Romero and Ventura [95], que relaciona o *data mining* com o desenvolvimento, a pesquisa e a aplicação informática de métodos que visam a procura de padrões coerentes e portadores de informação relevante em coleções de dados de tamanho considerável, Papamitsiou and Economides [84] elaboraram uma revisão sistemática da literatura que se distingue de trabalhos similares por considerar também a metodologia Learning Analytics (LA), defendendo a ideia de que ambas se complementam. Neste contexto, os autores apresentam a LA como uma estratégia de medição, agregação, análise e leitura de dados correspondentes aos estudantes e aos contextos em que estão integrados, com vista à produção de conhecimento que possa ser usado em função do desenvolvimento da área da educação. Apesar das semelhanças apontadas, ambos os métodos que constituem o protagonismo da Revisão de Literatura de Papamitsiou and Economides [84] diferem ao nível das técnicas, dos campos de observação e dos conhecimentos produzidos. O Learning Analytics é, segundo afirmam estes autores, mais abrangente e holístico, suscetível de oferecer uma visão globalizada dos contextos analisados. Em contraponto, o EDM tem por principal característica o facto de oferecer pontos de vista mais reduzidos, que resultam da possibilidade de análise dos componentes individuais encerrados nas bases de dados gigantes. É por via desta metodologia que se torna possível a descoberta de novos padrões e a alteração e aperfeiçoamento contínuo dos algoritmos existentes. Na equação que resulta da soma e subtração das semelhanças e diferenças apontadas para ambos os métodos, os autores em análise defendem que o EDM será tanto mais valioso quanto mais for considerado em complementaridade com o LA.

Partindo de uma pergunta de investigação específica, Papamitsiou and Economides [84] empreenderam um estudo que visava aferir quais os métodos mais usados na literatura existente, para determinar a eficácia da implementação do EDM. Os autores pretenderam também verificar em que medida é que os respetivos métodos têm vindo a contribuir para a implementação do EDM, como instrumento de análise. Na sua análise, foram considerados 40 estudos publicados entre 2008, ano que os autores consideram como o marco da implementação do EDM, e 2013. Neste conjunto de estudos constam alguns de natureza exploratória e experimental, alguns estudos de caso e revisões de literatura, que incidiam, sobretudo, na análise tecnológica e na componente matemática que caracteriza o método.

Numa primeira abordagem aos estudos selecionados, os autores desta revisão sistemática da literatura concluíram que a grande maioria dos estudos elaborados até à data determinada, adotaram o método *data mining* de classificação, seguindo-se o método de *clustering* e de regressão. Mais recentemente, mas ainda no âmbito da abordagem bibliográfica que se descreve, já foram identificados estudos que usavam novos métodos como análises de redes sociais, visualização, regras de associação e descobertas com outros modelos.

Quanto às métricas a que os autores destes 40 estudos mais recorreram, para aferir a *per-*

formance dos modelos apresentados, parecem prevalecer os de precisão, sensibilidade e de coerência. Todos estes instrumentos e métricas parecem ter sido aplicados com mais incidência em estudos que visavam analisar e classificar o modelo comportamental dos estudantes e também determinar formas capazes de prever o seu desempenho. Em alguns estudos pontuais, os autores foram ainda capazes de aplicar a metodologia com vista a determinar medidas de promoção da atenção dos professores em relação aos estudantes menos capazes, informando-os acerca dos seus progressos, em comparação com os outros colegas. Mas, tal como Papamitsiou and Economides [84] sublinharam, ao nível da discussão, o EDM não se limita à oferta desta tipologia de respostas, podendo vir a ser aproveitado, num futuro próximo, como instrumento de análise e investigação, nas mais variadas áreas da educação. Ele constitui uma ferramenta capaz de transformar qualquer clique, no contexto de uma plataforma tecnológica educacional, em informação valiosa, capaz de ser isolada, classificada e analisada por meio de métodos e algoritmos já existentes (Papamitsiou and Economides [84, p. 61]). Ao EDM os autores atribuem ainda o mérito da interseção entre a educação, a psicologia, a pedagogia e as ciências computacionais.

Cingindo-se à linha cronológica adotada por Papamitsiou and Economides [84], Peña-Ayala [86] empreendeu um estudo de revisão da literatura que considerou 240 estudos divididos em dois grupos heterogéneos, tanto na abordagem como em tamanho. O primeiro grupo analisado por este autor, que se estuda com maior destaque dado o facto de considerar uma temática que, se julga, precisar de aprofundamento no âmbito da investigação inerente a este doutoramento, era composto por 222 artigos em que o EDM foi abordado na perspetiva da sua concetualização, caracterização e aplicabilidade. O segundo recolheu 18 artigos em que o EDM era analisado sob o ponto de vista das suas ferramentas, aplicações e software.

Interessa sublinhar, desde logo, a discrepância dos estudos que até 2013 foram realizados acerca do EDM, parte deles considerando-o em termos teóricos e outros abordando-o numa perspetiva mais prática. Da mesma forma, é importante sublinhar o número total de artigos levantados durante o processo de *scopping*. Ambas as informações dão conta do indiscutível interesse que o EDM tem despertado junto da comunidade científica ao longo destes primeiros anos do século XXI e também da necessidade de aprofundamento dos estudos que visam a procura de modelos práticos de aplicação. No conjunto dos 222 artigos analisados na perspetiva das tarefas ou aplicações do EDM, Peña-Ayala [86] destaca que quase 82% dos estudos estão relacionados com as três versões de modelação de alunos e avaliação (comportamento, desempenho e geral). O complemento (18%) foi distribuído por abordagens no âmbito do suporte pedagógico e *feedback* dos alunos, domínio de conhecimento (habilidades a serem treinadas) e suporte a professores. No conjunto dos 18 artigos que deram ênfase à funcionalidade e instrumentos do EDM, os autores identificaram que 8, a maioria, apresentaram o EDM como meio de análise, 6 justificam-no como ferramenta de visualização de dados e os restantes consideraram esta metodologia como forma de extrair informação das bases de dados (Peña-Ayala [86, p. 1437]). Os últimos artigos considerados apontaram ainda o EDM como um mecanismo de suporte pedagógico e educacional e um meio para encetar esquemas de engenharia de variáveis e características a considerar em futuras análises. Uma vez que o autor empreendeu uma subdivisão dos artigos com base no ano da sua publicação, ficou evidenciada uma particularidade evolutiva do método que o conota a uma maior utilização enquanto modelo preditivo nos primeiros anos de abordagem da revisão de literatura realizada. Em pouco tempo, no entanto, o EDM viu os métodos de segmentação, de classificação e de regressão serem usados com alguma frequência.

Com o mesmo intuito, o de promover e valorizar o EDM enquanto ferramenta de análise e construção de estratégias nos processos de tomada de decisão ao nível do ensino académico, e

visando um objetivo mais estratégico que passa pela abordagem ao estado atual da aplicação da disciplina na Índia, Sukhija et al. [110] empreenderam um processo de investigação em que reuniram e analisaram 19 estudos que percorrem uma linha cronológica que vai desde o início do século XXI até 2015. De acordo com os autores desse estudo, o EDM é ainda uma disciplina em expansão, a que está associado um vasto conjunto de métodos e algoritmos de *data mining*, e que continua a exigir atenção por parte dos investigadores, sobretudo ao nível do alargamento do conjunto de algoritmos que permitam a conjunção (modelos híbridos/de conjunto) de técnicas de análise e agrupamento. Partindo da evidência de que no contexto educativo da Índia há ainda um longo caminho a percorrer para que a implementação e aceitação do EDM se concretize, Sukhija et al. [110] identificam, após analisarem estes 19 estudos, quatro lacunas associadas ao EDM. A principal diz respeito à indisponibilidade de bases de dados consistentes, suficientemente abrangentes, capazes de comportar todo o sistema educacional e o seu modo operandos. Fixados na importância da qualidade das bases de dados para o bom funcionamento e aproveitamento do EDM, os autores, sublinham ainda, como aspeto negativo a ter em conta, a falta de versatilidade das bases de dados que sustentam o funcionamento do *data mining* na educação e vão mais longe apontando as próprias ferramentas do EDM, em particular os seus algoritmos, como se tratando de instrumentos inflexíveis e ainda pouco aptos à utilização conjunta. Por último, os autores assinalam a falta de confiança que as entidades responsáveis, nomeadamente os governos, demonstram ainda ter nos resultados obtidos através do processo de *educational data mining*. Esta falta de confiança, constitui um entrave, quer na aplicação de estratégias e medidas a adotar num plano mais abrangente, quer na construção de bases de dados coletivas, integrativas e versáteis, que possam constituir um campo de análise à realidade educacional de países e até mesmo a nível mundial. De acordo com Sukhija et al. [110] o EDM pode vir a constituir-se num instrumento que permita aos professores, estudantes e administrações educativas beneficiar do melhor que eles próprios têm para oferecer [110, p. 35].

3.3.2 Previsão de desempenho académico

A previsão do desempenho académico pode providenciar conhecimento de importância primordial para a definição de estratégias de promoção educacional que se possam revelar assertivas. A previsão do desempenho académico consiste em estimar o valor futuro de uma variável de interesse a partir de outros fatores presentes no conjunto de dados. Como variáveis alvo de previsão de desempenho poder-se-à mencionar, por exemplo, uma qualquer métrica que quantifique o sucesso obtido no final do curso de graduação, como é o caso da própria média final de curso, classificações em determinadas unidades curriculares, e se aluno obtém ou não aprovação em determinadas unidades curriculares. Ainda que, num sentido mais lado, a previsão do desempenho académico possa abranger quer a previsão do sucesso académico quer a previsão do abandono escolar, os estudos analisados nesta secção relacionam-se quase invariavelmente apenas com o primeiro tópico de investigação. A revisão dos estudos de previsão de abandono, em particular, apresentar-se-á em separado na próxima secção.
Geralmente, os estudos sobre previsão de desempenho académico tentam de alguma forma dar resposta, pelo menos, a uma das duas seguintes questões de investigação:

- Quais são os fatores mais determinantes do desempenho académico?
- Quais são os algoritmos de *data mining* mais eficientes para a previsão do desempenho académico?

Nas duas subsecções que se seguem, apresenta-se uma revisão de literatura sobre alguns estudos desta índole com recurso ao EDM, focando-se a primeira delas na previsão do sucesso global do aluno e a segunda na previsão do sucesso em disciplinas específicas.

3.3.2.1 Previsão precoce do sucesso educacional dos estudantes no *terminus* da sua graduação

Romero et al. [96] desenvolveram uma ferramenta específica, para facilitarem o uso e a configuração das técnicas de *data mining* aos professores ou a outros utilizadores não especializados. A partir da análise dos dados gerados no Moodle, pelos alunos de 7 cursos ministrados em contexto *e-learning* da Universidade de Córdoba, foi desenvolvido um modelo de fácil compreensão e elevada precisão, para mostrar aos professores a previsão do desempenho académico dos estudantes no final dos seus cursos. Na análise do conjunto de dados foram usados métodos de classificação, de *clustering* e de regras de associação. Depois de terem sido aplicadas operações de pré-processamento, como a discretização e o balanceamento dos dados, foi verificado que alguns algoritmos melhoram o seu desempenho na classificação. Os autores destacaram que algumas das regras de associação mostraram que o número de questionários aprovados no Moodle foi o principal determinante do desempenho académico final dos estudantes. Foi igualmente sublinhado que algumas das regras geradas poderão ser usadas para ajudar os professores a decidir se as atividades escolares permitem melhorar os resultados escolares.
Num outro estudo empírico desenvolvido por Tiwari, Singh, and Vimal [112], realizado com dados dos estudantes do 2º ano de um curso tecnológico, de uma universidade nigeriana, foram identificadas relações entre o comportamento de aprendizagem dos estudantes e o seu desempenho académico. A informação sobre o desempenho académico considera as notas obtidas em trabalhos individuais e de grupo, as notas finais das disciplinas no semestre e a média do semestre. A assiduidade, a pontualidade e a frequência do horário de atendimento e de atividades extracurriculares foram as variáveis de cariz comportamental consideradas no estudo. Os métodos usados na abordagem seguida nesse estudo foram a associação, com o objetivo de encontrar relacionamentos entre os atributos, a classificação, para prever o desempenho, e o *clustering*, para agrupar os alunos em função das suas características semelhantes e do seu rendimento escolar. Os resultados mostraram que através de um conjunto de regras de associação do tipo *"if-then"* é possível identificar a relação entre os atributos comportamentais dos estudantes e as possíveis classificações de desempenho académico: excelente, bom, médio, fraco, mau. Por exemplo, se fossem baixas a frequência do estudante ao horário de atendimento no semestre, a pontuação nos exercícios propostos pelo professor *online* e a média final no semestre, então o aluno era classificado como fraco, com as medidas de suporte e confiança, respetivamente, iguais a 0.196 e 0.757. Estes valores indicam que no conjunto dos estudantes de engenharia em análise, 19.6% deles apresentaram baixo desempenho nas três componentes consideradas. Além disso, dos alunos com valores baixos nas 3 dimensões consideradas 75.7% tratar-se-ão de alunos efetivamente fracos (Tiwari et al. [112, p. 56]). Face aos resultados verificados, os autores concluíram que é possível construir suporte informativo realista a fim de fundamentar decisões no âmbito da definição de estratégias de promoção educacional dos estudantes. De acordo com os autores, o mesmo serviu ainda o propósito de afirmar o EDM como uma ferramenta que deve ser adotada pelas instituições de ensino, pelo facto de ser considerada uma mais-valia na construção de estratégias de promoção educacional, não obstante a necessidade firmada da continuação e melhoria dos estudos a serem realizadas acerca desta temática (Tiwari et al. [112]).
No estudo de Pandey and Taruna [82] foi defendida a necessidade de integrar classificadores e

preditores múltiplos, para aumentar a *performance* de um modelo desenvolvido para prever a média final de curso dos estudantes, logo no inicio do percurso académico. A metodologia proposta para o desenvolvimento do modelo preditivo de conjunto, designado KNNAD, um produto da regra de combinação de probabilidade foi usado para integrar 3 algoritmos complementares: Árvores de decisão (J48), k-vizinho mais próximo e agregação de estimadores de uma única dependência (AODE). O modelo assim desenvolvido foi aplicado e comparado através do teste *t de student* em três conjuntos de dados. O desempenho preditivo do modelo assim conseguido, ao ser comparado com cada um dos seus classificadores individuais, bem como com outros cinco de referência (*Naive Bayes, KSTAR, OneR, ZeroR e Naive Bayes Tree*), demonstrou ser mais efetivo e consistente do que qualquer um dos 8 algoritmos mencionados. Os autores concluíram que o estudo refletiu consistência ao nível da aferição de padrões comportamentais de desempenho académico dos estudantes de Engenharia, pelo que o mesmo pode ser aplicado a grupos académicos de outros contextos estudantis.

Natek and Zwilling [79] recorreram aos algoritmos *RepTree Model, J48 Model e M5P Model* com o objetivo de prever a taxa de graduação dos alunos recém ingressados na licenciatura de Informática da *International School for Social and Business Studies, Slovenia*. De forma a que o modelo preditivo pudesse refletir o mais possível a realidade foi considerado um amplo conjunto de variáveis preditivas tais como: o ano de estudo, o género, o ano de nascimento, situação de emprego (sim/não), tipo de ingresso (normal, atleta, militar), tipo de inscrição (primeiro/-repetição), tipo de frequência (tempo integral/parcial), classificações obtidas nas atividades propostas, pontuação em exames e classificação final. Os autores concluíram que os fatores que influenciaram de forma mais visível a média final dos estudantes estavam relacionados com a informação de acesso, demográfica e com as atividades extra-curriculares. Os dados relativos a 42 estudantes do primeiro ano e a 32 do segundo ano do curso foram usados para treinar os 3 algoritmos enquanto que os dados de 32 alunos do último ano académico foram utilizados para efeitos de teste e validação do modelo. Depois de terem escolhido diferentes combinações de variáveis preditivas, com base nas variáveis mencionadas, os autores constataram que o modelo que apresentou maior acurácia na previsão (90%) foi obtido com variáveis relacionadas com as classificações nos exames, constatando-se que todas as outras variáveis mencionadas se mostraram relevantes para a previsão, especialmente as relacionadas com a avaliação interna. Pelo facto dos resultados da investigação se terem revelado muito promissores, Natek and Zwilling [79] sugeriram à direção da universidade que fossem incorporadas ferramentas de *data mining*, como uma parte importante dos seus sistemas de gestão de conhecimento.

No estudo de Santos and Boticario [102] é reportado que através do algoritmo *apriori* de regras de associação foi encontrada uma relação significativa entre o sucesso alcançado pelos alunos no final do seu curso e o respetivo grau de comprometimento com as tarefas realizadas durante o processo educacional baseado em computador. Analisando a interação de 55 estudantes com a plataforma de aprendizagem, na participação dos fóruns, no preenchimento dos questionários de auto-avaliação, no carregamento da foto de perfil, na utilização do *chat*, na publicação de *posts* e na realização das tarefas obrigatórias, foi possível constatar que quanto maior for o grau de comprometimento com as tarefas solicitadas, maior será a probabilidade desse aluno obter maior média de curso. Estes resultados também foram consistentes com os obtidos com recurso ao algoritmo C4.5 (também chamado de J48) de árvores de decisão, ao ser confirmado que as árvores resultantes diferenciaram, claramente, os alunos que realizaram a auto-avaliação e tiveram sucesso no curso, daqueles que não o fizeram. Os autores concluíram que os resultados obtidos permitem abrir perspetivas para se poder providenciar um acompanhamento mais personalizado aos estudantes, principalmente, para os quais se prevê um baixo grau de compro-

metimento com as tarefas educacionais, quase desde o momento de ingresso no curso (Santos and Boticario [102]).

Num outro estudo, Campagni et al. [19] basearam-se em *clustering* e em técnicas de sequência de padrões para desenvolver um modelo que permitisse comparar os estudantes em geral com o padrão do estudante ideal, do curso de Ciências da Computação da Universidade de Florença, de Itália. No sistema de ensino universitário Italiano, os alunos podem realizar os exames finais em várias datas diferentes e, geralmente, têm alguma liberdade para escolher a ordem pela qual os realizam. Os autores pretenderam entender qual seria a sequência ideal para a realização dos exames e também qual seria o momento ideal para essa realização. Por exemplo, pretendia-se saber se os exames deveriam ocorrer logo após a conclusão do período letivo, ou, se deveriam ser adiados para uma data posterior.

Através da análise de *clustering* e da análise de sequência de padrões, foi possível revelar dois tipos de relacionamentos. Um entre a data e a ordem da realização dos exames com os desempenhos académicos. Outro entre a data e a ordem de realização dos exames com o número de anos necessários até à obtenção da graduação. Com base na análise de *clustering*, os alunos foram divididos em dois grupos: o grupo de estudantes que se formou relativamente rápido e com média de curso alta e o grupo de estudantes que obteve resultados piores. Pela análise de sequências de padrões foram identificadas várias sequências/ordens na realização dos exames, comparando-as, posteriormente, com a sequência ideal. Os resultados apurados permitiram demonstrar que 70% dos estudantes seguiam uma determinada ordem na realização dos exames. Foi essa mesma ordem que foi considerada como "quase ideal" e que serviu para a comparação. Neste sentido, foi criado um elemento que permitisse cruzar as informações das bases de dados (que continham atributos relacionados com o ano de inscrição na universidade, o currículo escolhido, a data e as notas de cada exame) com um plano estratégico desejado e antecipado. Partindo desta base de estudo, o plano metodológico construído foi aplicado a um estudo de caso, cuja amostra continha dados sobre 141 estudantes graduados, matriculados pela primeira vez no mencionado curso de Ciências da Computação, entre os anos letivos de 2001/2002 e 2007/2008. A mesma amostra serviu para validar os resultados obtidos, que se traduziram na ideia de que quantos mais estudantes se aproximavam do padrão do estudante ideal, melhores eram os resultados globais obtidos.

A partir dos resultados obtidos, foi possível proceder à apresentação de propostas efetivas para a promoção do desempenho académico dos estudantes. Entre outras propostas, foi apresentada a ideia de colocação de um tutor à disposição dos alunos que se atrasavam na realização dos exames, para o semestre seguinte, a fim de os ajudar a escolher o caminho ideal a seguir na realização dos exames. Outra consequência prática desta investigação foi o facto de terem incluído restrições adicionais entre os exames. Os autores adiantaram que ainda era prematuro avaliar os efeitos reais destas ações, mesmo que os resultados iniciais parecessem encorajadores, especialmente em termos de notas e atrasos.

Num outro estudo, realizado em contexto de aprendizagem mista (*e-learning* e presencial tradicional), Amrieh et al. [4] abordaram o EDM como um instrumento capaz de prever um modelo padronizado de comportamento estudantil. Através de uma metodologia que considerou os algoritmos de classificação redes neuronais artificiais, árvores de decisão e *Naive Bayes* (NB), e os métodos de conjunto *bagging*, *boosting* e o *random forest*, confirmou-se que das características comportamentais de aprendizagem dos estudantes depende o seu sucesso educacional no final do curso de graduação. No desenvolvimento do novo modelo, designado Características Comportamentais dos Estudantes (*student's behavioral features*), foi considerado um conjunto de dados relativo a 500 estudantes, caracterizados por um total de 16 fatores, distribuídos por

três categorias principais: demográficas, histórico académico e características comportamentais relacionadas com o processo de aprendizagem (em sala de aula e em contexto *e-learning*). No conjunto dos fatores relativos ao histórico académico dos estudantes constam as habilitações académicas prévias, as classificações nas diversas disciplinas frequentadas e o semestre frequentado pelo aluno (1º/2º). Relativamente às características comportamentais de aprendizagem dos estudantes, foi analisada a experiência dos alunos durante o processo educacional nos sistemas *e-learning* e o grau de satisfação dos pais com a escola. Na categoria das características demográficas dos alunos, foi estudada a influência do progenitor responsável, da nacionalidade, do género e do local de nascimento. Os autores sublinharam que a decisão de incluírem variáveis da categoria comportamental no novo modelo desenvolvido, como a participação em *chats* e em grupos de discussão, e a intervenção e participação em sala de aula, contribuiu para uma melhoria da capacidade preditiva do modelo avaliada em 21.1%, quando comparada com os resultados preditivos obtidos por aplicação do mesmo modelo sem inclusão desse mesmo tipo de atributos. Em relação ao desempenho preditivo dos modelos desenvolvidos pelos métodos de conjunto acima mencionados, a acurácia foi ainda 25.8% superior, quando comparada com os classificadores tradicionais. Posteriormente, com dados de teste, relativos a um conjunto de estudantes recém-ingressados, o modelo Características Comportamentais dos Estudantes evidenciou uma acurácia superior a 80%. Pelo facto do modelo desenvolvido ter demonstrado capacidade para espelhar um quadro fidedigno da realidade, Amrieh et al. [4] concluíram que se trata de um instrumento útil para a gestão das IES. Por exemplo, logo no momento do ingresso, o modelo pode ajudar a identificar os alunos propensos ao insucesso e, por conseguinte, será possível melhorar os processos de ensino da população identificada (idem).

Noutro estudo, Rubiano and Garcia [98] com recurso aos algoritmos *J48, PART e Ridor* de árvores de decisão, desenvolveram um modelo preditivo de classificação que permitiu, logo no inicio do 1º ano letivo, diferenciar os estudantes em função da estimativa da sua média final de curso. O modelo gerado estabeleceu 3 classes de estudantes: os de "risco iminente" de abandonar o curso de formação, os de "*performance* média" e os de "*performance* elevada". A metodologia apresentada foi desenvolvida e validada com registos de 932 estudantes do curso de engenharia de sistemas, de uma universidade colombiana. Para a classificação, foi considerada informação académica, demográfica e sócio-económica. A classe social dos estudantes demonstrou ser um fator altamente decisivo para o desempenho, aparecendo, de forma recorrente, como o primeiro atributo de classificação nos conjuntos de regras gerados pelos algoritmos de decisão em árvore. O estado civil parece ser o segundo atributo de classificação mais importante, verificando-se que a maioria das regras resultantes giram em torno do valor "solteiro". Outros atributos, como a idade, o género, a educação da mãe e os valores obtidos no ICFES[3], especialmente, em biologia, filosofia e línguas, também parecem influir, de forma relevante, no desempenho escolar. O classificador gerado pelo algoritmo *J48* foi o que demonstrou melhores resultados preditivos, uma vez que 78% das instâncias avaliadas foram corretamente classificadas. Os autores também verificaram que a eficácia preditiva dos modelos foi significativamente melhorada, após terem procedido a um adequado pré-processamento dos dados. Entre as tarefas efetuadas foi destacado o efeito da remoção dos atributos irrelevantes e a discretização dos dados.

Aluko et al. [3] recorreram à análise discriminante linear e ao algoritmo K-NN para prever a média final de curso de 101 alunos recém ingressados na licenciatura de arquitetura, numa universidade nigeriana. As 13 variáveis independentes usadas no desenvolvimento do modelo preditivo estavam todas relacionadas com o desempenho escolar pré-universitário dos estu-

[3]Exame usado nas universidades da Colômbia no processo de seleção de novos estudantes.

dantes. A análise discriminante linear mostrou que os fatores mais influentes para explicar o sucesso educacional, verificado no final do curso, foram as notas do exame de acesso ao ensino superior, o facto de serem ou não candidatos de entrada direta e as notas obtidas em matemática. A percentagem de casos corretamente classificados (50%) e o valor da estatística *kappa de Cohen* (-0.230) indicaram uma concordância fraca, entre a classe preditiva e a classe real, no conjunto de dados de teste. De acordo com os resultados verifica-se que o sucesso académico pré-universitário, principalmente, em matemática, física, química e língua local, é um bom preditor do sucesso académico na licenciatura em arquitetura. A análise discriminante linear mostrou que o sucesso educacional obtido no final do ensino secundário, como as notas do exame de acesso ao ensino superior, o facto de serem ou não candidatos de entrada direta e as notas de matemática, física, química e língua local são fortes preditores do sucesso obtido no final do curso de graduação.

Num outro estudo, para inferir a média final de curso dos estudantes, Cerezo et al. [21] usaram métodos de *clustering* para formar grupos de alunos, com comportamentos semelhantes e, se possível, com diferentes níveis de desempenho. Para a análise de *clustering* foi usado o algoritmo de maximização da expectativa (EM) e para a classificação do *cluster*, foi utilizado o algoritmo *k-means* como método confirmatório. Para testar a hipótese da diferença entre os *clusters* formados os autores efetuaram uma análise de variância (ANOVA). A metodologia proposta foi desenvolvida e validada com dados que caracterizam 140 alunos, do 3º ano da licenciatura em psicologia, da Universidade de *Oviedo*, do norte de Espanha. As 6 variáveis preditivas consideradas na análise, relacionadas com a procrastinação e com o tempo despendido pelo alunos no âmbito das suas atividades de aprendizagem, foram todas obtidas a partir das entradas no Moodle. A variável procrastinação representa o tempo, em dias, que o estudante demorou a realizar a tarefa solicitada, desde o momento em que a mesma foi disponibilizada no Moodle. As variáveis relacionadas com o tempo de aprendizagem representam o tempo total usado em tarefas de estudo $-$ o tempo total a estudar conteúdos teóricos, o tempo usado com atividades práticas, o tempo total no fórum, o número de palavras "postadas" no fórum e o número de ações relevantes no Moodle. O modelo gerado, pelo método de *clustering*, permitiu agrupar os 140 alunos em quatro *clusters* distintos, através dos quais foi possível analisar o desempenho escolar dos estudantes.

Os resultados obtidos sugerem que as 6 variáveis explicativas, além de serem relevantes para o agrupamento de alunos, também se revelaram essenciais para prever o desempenho escolar (Cerezo et al. [21]). Por exemplo, a variável procrastinação indica que os alunos que entregam a tarefa mais tarde têm menor probabilidade de sucesso escolar, enquanto que aqueles que dedicam mais tempo às tarefas solicitadas, apresentam, por norma, melhores resultados (idem).

Relativamente aos resultados obtidos através da metodologia ANOVA, em que a média final é a variável dependente e os diferentes *clusters* as variáveis independentes, foi possível verificar que existem diferenças estatisticamente significativas entre os quatro grupos de alunos, no que se refere à média final obtida no curso de graduação.

No estudo de Ruby and David [99] foi comparado o desempenho preditivo dos modelos suportados pelos algoritmos de classificação *ID3, J48, NBTree, RepTree, Multi Layer Perceptron (MLP), SimpleCart* e *Decision*, numa abordagem desenvolvida para a previsão precoce do desempenho académico dos estudantes no final do mestrado em Aplicação de Computadores. Os resultados comprovaram a superioridade preditiva do modelo desenvolvido pelo algoritmo MLP, ao ter demonstrado uma acurácia de $83,6\%$. O conjunto de dados usados na análise caracteriza 165 estudantes, através de 12 variáveis explicativas, relativamente às dimensões académica, pessoal e económica dos estudantes. Os autores concluíram que do conjunto das 12 variáveis

explicativas estudadas, são 7 as que justificam a capacidade preditiva do classificador MLP. No conjunto surge o rendimento familiar, as classificações obtidas nas disciplinas frequentadas anteriormente, a nota de licenciatura, o tipo de residência, o tempo médio de estudo, as notas obtidas em avaliações teóricas e as atividades extra-curriculares. Os autores concluíram que a metodologia proposta, ao permitir antever com a devida antecedência o *status* académico dos alunos, favorece o desenvolvimento de estratégias adequadas à promoção do desempenho, principalmente daqueles que apresentam maiores dificuldades (p.1091).

Os resultados do modelo desenvolvido por Asif et al. [6], com base nos algoritmos *Naive Bayes* e *Random Forest*, mostraram que é possível prever, com elevada precisão, o desempenho global dos estudantes no final de um curso de quatro anos, usando apenas indicadores de desempenho académico nas disciplinas pré-universitárias e nas disciplinas do 1º e 2º ano da universidade. Com recurso a árvores de decisão, os autores também identificaram quais as disciplinas dos primeiros anos de estudo que podem prenunciar um baixo rendimento académico global. Ao terem sido estudadas progressões típicas dos estudantes, para serem combinadas com os resultados do modelo de previsão, com técnicas de *clustering*, foram identificados dois importantes grupos de estudantes: os de baixo e os de elevado desempenho. Os autores concluíram que usando apenas a informação de desempenho académico (pré-ingresso e a obtida no final dos 2 primeiros anos académicos), prescindindo de informação sócio-económica e demográfica, é possível fornecer alertas oportunos para apoiar os alunos que mais precisam de promover o seu desempenho, bem como dar conselhos aos bons alunos.

Mais recentemente, no estudo de Miguéis et al. [71], com recurso aos algoritmos de classificação *random forest, decision trees, support vector machines, naive Bayes, bagged trees and boosted trees*, foi demonstrado que é possível prever, logo no final do 1º ano letivo, o sucesso final dos estudantes dos mestrados integrados na Faculdade de Engenharia da Universidade do Porto. Para inferir o sucesso no final dos cursos de mestrado com a duração de 5 anos, foi considerada informação de acesso ao ensino superior, demográfica, sócio-económica e de desempenho académico nos 2 primeiros semestres do percurso escolar dos estudantes. A informação analisada, que também serviu para testar a viabilidade da metodologia proposta, distribuída por 19 atributos, caracteriza um conjunto de 2469 estudantes em regime presencial tradicional. Dentre os algoritmos utilizados, o *random forest* foi aquele que apresentou os melhores resultados preditivos, com uma acurácia de $96,1\%$. A ordem de importância atribuída pelo mesmo algoritmo aos atributos preditivos que suportam o modelo, demonstrou que os fatores mais importantes para prever e explicar o nível de sucesso académico dos estudantes no final do curso de mestrado, são a média de acesso e das provas de ingresso ao ensino superior universitário, bem como a média obtida nas unidades curriculares do 1º e 2º semestres do primeiro ano letivo.

Os autores também propuseram uma estrutura de segmentação multi-classe, visando uma classificação precoce dos estudantes, baseada no desempenho observado no final do primeiro ano letivo e na propensão ao sucesso académico revelado pelo modelo preditivo. A estratégia de segmentação implementada, permite a delineação de ações educativas diferenciadas, por grupos de estudantes. Pelo conhecimento assim obtido, as instituições podem projetar estratégias oportunas, que atempadamente poderão fomentar alterações comportamentais, a fim de promover níveis mais elevados de desempenho académico (Miguéis et al. [71]).

3.3.2.2 Previsão precoce do sucesso educacional em unidades curriculares específicas

No seu estudo, Hoffait and Schyns [49] tiveram como objetivo prever, logo no momento da matricula, as notas finais nas disciplinas do 1º ano do ciclo de estudos, dos alunos recém ingressados

na Universidade de *Liège*, na Bélgica. O modelo usado para prever se os alunos completariam com sucesso o 1º ano do ciclo de estudos, foi desenvolvido com recurso aos algoritmos *random forest*, redes neuronais artificiais e regressão logística. Um novo modelo de conjunto desenvolvido para a previsão, que combina os 3 algoritmos, demonstrou melhor desempenho preditivo do que qualquer um dos 3 algoritmos de forma individual.

Com ênfase na identificação dos alunos mal sucedidos, o estudo de Bydžovská [16] teve como principal objetivo prever, no inicio de cada um dos semestres letivos, as notas das disciplinas no final do mesmo semestre. As duas abordagens metodológicas seguidas neste estudo, uma de classificação e outra de *clustering*, foram desenvolvidas e avaliadas com informação relativa a 138 unidades curriculares, ministradas na Universidade de *Masaryk*, na República Checa. Na abordagem de classificação foram identificados 2 grupos distintos de estudantes: os de sucesso (notas $1-3$) e os de insucesso (nota 4). Os algoritmos preditivos *random forest*, classificador baseado em regras (OneR), árvores de decisão (J48), IB1, *Naive Bayes* e *support vector machines* apresentaram todos resultados satisfatórios, embora as *support vector machines* tivessem revelado alguma superioridade ($Fmedida = 0.559$). Através da abordagem de classificação também foi possível concluir que os atributos relacionados com o comportamento social dos estudantes foram aqueles que se revelaram mais informativos para a previsão. Dentre todos, o número de vezes que um estudante frequentou uma qualquer disciplina com o mesmo professor, o curso frequentado ser o favorito e a interação com os seus pares, foram os mais informativos.

Na segunda abordagem, através de técnicas colaborativas de filtragem, baseando-se na semelhança entre itens, foi possível agrupar os estudantes com interesses e conhecimentos similares, para prever a propensão para o (in)sucesso nas disciplinas frequentadas no semestre em análise. Os resultados obtidos permitiram concluir que, em média, ambas as abordagens atingiram resultados muito semelhantes, pelo que, qualquer uma das duas pode ser aplicada, quando se pretende prever, logo no inicio do semestre, a classificação final para um período muito próximo que é o final desse mesmo semestre. Como o esforço e o trabalho dos estudantes, em cada semestre, depende sempre do conjunto de disciplinas às quais eles se inscrevem, os autores concluíram que o estudo pode ser benéfico, na medida em que será possível proceder à elaboração de recomendações destinadas a equilibrar o tempo e o esforço despendidos no estudo.

Num outro estudo, através de uma abordagem com os algoritmos de classificação *Naive Bayes, SMO, J48, REPTree e o Multilayer Perception*, Kaur et al. [55] propuseram-se a identificar quais serão os estudantes do último ano de estudos de uma escola secundária que apresentarão mais dificuldades de aprendizagem. A partir de registos de 152 alunos, foi possível verificar que o género, a instituição de ensino superior pretendida pelos estudantes, a propina privada (booleano), o facto do aluno possuir telemóvel, computador e acesso à rede em casa, notas das disciplina e a assiduidade, foram as características que permitiram diferenciar, por ordem de importância decrescente, os estudantes com maiores ou menores dificuldades de aprendizagem. Após uma análise comparativa à *performance* preditiva dos 5 classificadores utilizados, verificou-se que o *Multilayer Perception* demonstrou ser mais eficaz, ao atingir $Fmedida = 82\%$ quando todos os restantes ficaram abaixo de 70%. A investigação de Kaur et al. [55] revelou que por esta via se pode providenciar suporte à tomada de decisão baseada em conhecimento.

Huang [52] desenvolveu um conjunto de modelos preditivos de regressão, com o objetivo de estimar a nota final na disciplina *Engineering Dynamics*, ministrada na *Utah State University*, dos Estados Unidos. O conjunto de dados usado neste caso de estudo contém registos de 324 estudantes, caracterizados através de 8 variáveis, relacionadas com o seu desempenho académico, pré e após ingresso na universidade. A partir do conjunto de 8 variáveis preditivas, foram estudadas seis combinações dessas variáveis, a fim de encontrar os fatores que melhor expli-

cam os resultados obtidos na referida disciplina. O estudo realçou que a média acumulada, as classificações de 3 disciplinas pré-requisito e a classificação no primeiro teste parcelar são os principais fatores que explicam a nota final da disciplina em análise. Em relação ao desempenho dos 4 algoritmos de aprendizagem usados neste caso de estudo, verificou-se que os modelos desenvolvidos por rede de função de base radial (RBF) e por *support vector machines* têm melhor capacidade de generalização do que os modelos desenvolvidos por regressão linear múltipla e por redes neuronais artificiais multicamada do tipo *percepton*.

Num outro estudo, pelo facto do abandono académico nas licenciaturas de Computação estar fortemente associado às reprovações nas disciplinas iniciais de programação, Pascoal et al. [85] pretenderam estimar, de forma precoce, as classificações dos estudantes em 4 disciplinas de programação, ministradas nos 3 primeiros semestres dos cursos de computação da Universidade Federal da Paraíba, Brasil. Como atributos preditivos foram usados os resultados das provas do vestibular[4] e as classificações obtidas pelos estudantes nas disciplinas pré-requisito de cada uma das quatro disciplinas em análise. Os resultados obtidos evidenciaram um bom desempenho preditivo, para qualquer um dos modelos desenvolvidos pelos algoritmos de aprendizagem *IBk, random forest, BayesNET (BNET) e MultilayerPerceptron (MLP)*. Entre os quatros, verificou-se que as *random forest* foram as que apresentaram maior acurácia (84.7%) e que o IBk foi o que apresentou melhor resultado na taxa de verdadeiros positivos (VP), sempre superior a 87.6%, nas 4 previsões efetuadas. Os autores deram ênfase à taxa de VP ao mencionar que a mesma representa a percentagem dos alunos que provavelmente reprovarão nas disciplinas, cuja identificação é primordial para minimizar os índices de reprovação. Os autores enfatizaram que através da aplicação dos modelos de previsão proposto será possível definir ações pró-ativas, a fim de minimizar os índices de reprovação nas disciplinas de Introdução à Programação, Linguagens de Programação I e II e Estruturas de dados, que são aquelas que estão criticamente associadas ao abandono dos cursos de Computação (Pascoal et al. [85, p. 457]).

Num outro estudo, Papamitsiou et al. [83] desenvolveram uma investigação relativa à dinâmica comportamental dos estudantes, a partir da análise do tempo que os mesmos necessitam para resolver os problemas, num processo de avaliação realizado em computador. As variáveis temporais usadas para classificar os estudantes quanto ao seu desempenho na disciplina de Informática II foram o tempo total usado para responder corretamente, o tempo total usado para responder erroneamente e o tempo total de inatividade. O conjunto de dados, com registos de 301 estudantes a frequentar a disciplina de Informática II, ministrada no departamento de economia de uma universidade grega, foram processados pelos algoritmos *redes neuronais artificiais, support vector machines, Naive Bayes (NB), k-Nearest Neighbors (KNN) e treeBagger*. O rigor da previsão, quantificado pela $Fmedida$, variou entre o mínimo de 82%, no modelo NB, e o máximo de 88%, no modelo KNN. As baixas taxas de classificação errada também foram indicativas de um bom desempenho preditivo dos modelos desenvolvidos. Quanto a esta métrica, o algoritmo *treeBagger* foi o que apresentou os melhores resultados, embora os algoritmos KNN e NB também tenham alcançado resultados muito satisfatórios. Os resultados permitiram constatar que a capacidade preditiva é elevada, quando se relaciona o tempo usado na resolução dos problemas com o desempenho académico nesse mesmo teste. Os autores concluíram que através da metodologia proposta será possível saber, logo após o primeiro teste de avaliação, para que alunos de Informática II se antevê um desempenho escolar menos bom. Haverá, portanto, possibilidade de se delinearem, de forma antecipada, práticas educativas que possam promover o sucesso escolar.

[4] Exame usado pelas Universidades Brasileiras no processo de seleção de novos estudantes.

No estudo de Costa et al. [26] foi comparado o desempenho preditivo das árvores de decisão, redes neuronais artificiais, *support vector machines e Naive Bayes*, numa abordagem que pretendeu identificar, tão precoce quanto possível, os alunos mais suscetíveis ao insucesso na disciplina de Introdução à Programação, ministrada numa Universidade Federal Brasileira. O estudo comparativo foi efetuado com dois conjuntos de dados independentes. Um contém registos de 262 alunos inscritos na referida disciplina, ministrada no sistema de ensino à distância. O outro contém registos de 161 estudantes a frequentar a disciplina em regime presencial tradicional no *campus* académico. As variáveis explicativas presentes nos dois conjunto de dados caracterizam os estudantes nas suas dimensões sócio-económica e de desempenho académico (classificação no 1º exame). A comparação ao desempenho preditivo dos 4 algoritmos mencionados foi efetuada em duas fases. Os resultados da primeira fase, efetuada antes do pré-processamento dos dados e da afinação dos algoritmos, evidenciaram um desempenho preditivo a variar de 0.55 a 0.82 na disciplina ministrada no sistema de ensino à distância e a variar de 0.50 a 0.79 na disciplina do sistema de ensino presencial tradicional. Entre os 4 algoritmos, as árvores de decisão foram as que apresentaram melhor desempenho preditivo em ambas as fontes de dados ($Fmedida = 0.82$ e 0.79, respetivamente, para o ensino à distância e no *campus*).

Na segunda fase do estudo, após o adequado pré-processamento dos dados e/ou ajuste dos hiperparâmetros dos algoritmos em análise, as *support vector machines* ajustadas superaram a eficácia preditiva das árvores de decisão, redes neuronais artificiais e *Naive Bayes*, de maneira estatisticamente significativa. Com efeito, verificou-se que a meio do período letivo da disciplina ministrada no sistema de ensino à distancia, as *support vector machines* atingiram 92% de acerto na identificação precoce dos estudantes suscetíveis ao insucesso. Para a disciplina ministrada no *campus*, o acerto das *support vector machines* foi de 83%, quando decorrido $1/4$ do período letivo da disciplina. Os autores concluíram que a eficácia preditiva da maioria dos algoritmos atingiu melhores resultados, após um devido pré-processamento dos dados e/ou afinação dos algoritmos. Foi ainda possível verificar que, de entre os quatro, as *support vector machines* ajustadas superaram o desempenho dos restantes algoritmos, em ambas as fontes de dados, com significância estatística.

Sweeney et al. [111] desenvolveram o seu estudo, prevendo, no final de cada um dos semestres escolares, a classificação das disciplinas que seriam frequentadas pelo aluno no semestre seguinte. A informação histórica que suportou a investigação permitiu caracterizar professores, estudantes e as respetivas disciplinas, da universidade pública *George Mason University*, onde são ministradas 144 unidades curriculares, a uma população estudantil de 33.000 estudantes. Os estudantes foram caracterizados com dados demográficos (idade, raça, género, escola de proveniência, código postal) e de desempenho académico (notas dos exames, média em cada um dos semestres, média acumulada no semestre, o número de créditos a que se inscreveu em cada semestre, o número de créditos acumulados a que já se inscreveu no semestre e o número de semestres em que o aluno já esteve inscrito). Uma vez que a previsão pretendida tanto poderia ser encarada através de uma abordagem de classificação, como de regressão, devido ao mapeamento ordinal dos dados, os autores optaram por explorar os dois métodos em experiências preliminares. Concluíram que a exatidão da previsão obtida pela abordagem de classificação ficou aquém da obtida pela de regressão. Com ênfase na abordagem de regressão, foram desenvolvidos 4 modelos com base em métodos de linhas simples (MLS), métodos baseados em fatorização de matrizes (MF) e modelos de regressão comum. Depois de terem proposto uma nova métrica de seleção de recursos através do método FM, designada *Mean Absolute Deviation Importance (MADImp)*, os autores concluíram que o acerto do modelo preditivo aumentou em 26%, em comparação com a seleção tradicional apresentada pelo mesmo algoritmo.

Foi igualmente demonstrado, através de um novo modelo de conjunto, desenvolvido pela combinação dos algoritmos máquina de fatorização (FM) e *random forest* (RF), que o erro da previsão diminuiu em comparação com os 4 algoritmos clássicos. Os autores sublinharam, inclusive, que o sucesso preditivo do novo modelo de conjunto designado FM-RF, se deve ao facto de ter sido aplicada a nova técnica MADImp para a seleção de atributos. Pela aplicação do novo modelo FM-RF foi possível verificar que há uma forte relação entre as características dos professores e o desempenho dos estudantes nas avaliações efetuadas às disciplinas semestrais. Também foi possível observar que a área de residência dos estudantes foi o único fator da categoria dos demográficos que demonstrou ter alguma influência para prever e explicar as notas das disciplinas num determinado semestre. A média acumulada semestral também demonstrou ser menos explicativa para a previsão, do que as notas obtidas mais recentemente na universidade.

3.3.3 Previsão de abandono académico

A identificação precoce de estudantes propensos à evasão é crucial para a formulação de iniciativas de prevenção ao abandono. Nesta subsecção apresenta-se uma revisão de literatura, que permite demonstrar a contribuição e aptidão dos métodos de *data mining*, em previsões precoces do abandono escolar.

3.3.3.1 Previsão precoce do abandono académico em IES

Nandeshwar et al. [78] conseguiram identificar, logo no final do 1º semestre do plano de estudos, quais são os fatores que caracterizam os estudantes persistentes na Universidade *Kent State*, até adquirirem o seu grau académico. Para a investigação descrita neste estudo foi explorado um conjunto de dados com 100 variáveis explicativas, relacionadas com as dimensões demográfica, académica e sócio-económica dos estudantes da referida universidade. Após a análise e exploração do conjunto de dados, com recurso aos algoritmos de classificação *One-R, C4.5, ADTrees, Naive Bayes* e redes de polarização radial, os autores concluíram que o historial familiar (educacional e socioeconómico) e o historial de desempenho académico no ensino secundário, têm uma relação direta e positiva com aqueles que permaneceram na universidade até à conclusão do curso. Em contraponto, viver fora do *campus* demonstrou uma correlação negativa com os não desistentes da referida universidade. De salientar o facto de que nenhum dos atributos associados ao historial pós ingresso na universidade foram identificados pelos seletores de subconjuntos de recursos. O estudo demonstrou ainda que os bons níveis de desempenho alcançado pelos modelos preditivos, melhorou substancialmente após ter sido feito o adequado e cuidadoso pré-processamento de dados, como por exemplo, a remoção de atributos espúrios. Os autores sublinharam que pelo conhecimento obtido através da abordagem metodológica descrita, foi possível projetar ações estratégicas concretas e necessárias de combate à evasão. Entre outras propostas, foi recomendado que se prestasse apoio financeiro aos estudantes com rendimentos familiares escassos, que os alunos do primeiro ano fossem encorajados a viver no *campus* e que se lecionassem aulas de apoio suplementar, principalmente para os estudantes cujas classificações no ensino secundário indicassem falta de preparação prévia (p.14995).

A Tese de doutoramento de Manhães [64] apresenta uma abordagem de classificação para identificar, no inicio de cada um dos 5 primeiros semestres letivos do plano de estudos (em cursos de 12), quais são os estudantes em risco de abandonar os estudos na Universidade Federal do Rio de Janeiro (UFRJ). Prescindindo de informação de cariz sócio-económica e demográfica, cingindo-se apenas informação académica, que varia no tempo, o modelo proposto permite diferenciar,

do 1º ao 5º semestre letivo, os alunos em duas classes distintas: progresso ou não progresso. O critério usado para classificar um estudante como sendo de progresso foi ter tido pelo menos uma disciplina aprovada no semestre em análise. Caso contrário, a situação do estudante foi considerada como de não progresso, antevendo-se serem esses os que terão maior probabilidade de abandonar o seu curso. A informação usada para a classificação está relacionada com a assiduidade e com atributos académicos relativos ao progresso na aprendizagem (classificações obtidas nas disciplinas, número de disciplinas aprovadas/reprovadas em cada semestre, média acumulada no semestre em análise, ao longo dos primeiros 5 semestres curriculares). Para processamento do grande conjunto de dados, reunidos na instituição ao longo de 16 anos, entre 1994 e 2010, foram usados os 12 algoritmos de classificação disponíveis na biblioteca da ferramenta Weka (AdaBoost (AD), BayesNet (BN), DecisionTable (DT), J48, JRip (JR), MultilayerPerceptron (MP), NaiveBayes (NB), OneR (OR), RandomForest (RF), SimpleLogistic (SL), SVM com PolyKernel (SVM1) e SVM com RBF Kernel (SVM2)). Os resultados experimentais evidenciaram, para todos os classificadores mencionados, uma capacidade de acerto preditivo superior a 75%. Entre todos, o algoritmo *Naive Bayes* foi aquele que, em média, apresentou melhor resultado preditivo nos 5 semestres letivos em análise.

Os autores concluíram que através de um número reduzido de atributos, é possível prever, semestre a semestre, quais são os estudantes em risco iminente de evasão. Os resultados ainda permitiram demonstrar que o fator mais importante para a previsão de abandono é a nota da disciplina de Cálculo Diferencial e Integral I (Manhães [64, p. 85]). Nessa tese, com recurso a técnicas de visualização de *data mining*, foram apresentados os resultados dos algoritmos de forma mais interpretável. Por exemplo, através do algoritmo *Naive Bayes* a autora apresentou facilmente os resultados numéricos em forma de gráficos. Este procedimento permitiu auxiliar os gestores institucionais na interpretação dos resultados obtidos.

Também com o intuito de encontrar padrões e relações interessantes, que possam contribuir para a identificação dos fatores associados ao abandono escolar, Lehr et al. [60] exploraram informação de desempenho de aprendizagem (pré-universitário e do 1º ano académico), de cariz pessoal e sócio-económica, relativa a um conjunto de 972 alunos da *Riddle Aeronautical University* (ERAU).

Os seis algoritmos de aprendizagem utilizados no desenvolvimento dos modelos preditivos de abandono foram o Naive Bayes, o *k-nearest neighbors*, *random forest*, redes neuronais artificiais *multilayer perceptron*, árvores de decisão e regressão logística. Os autores concluíram que o modelo de regressão logística foi o que demonstrou a melhor eficácia de previsão, seguindo-se os modelos Naive Bayes e *multilayer perceptron*, ambos com eficácia superiores a 70%. Esta constatação evidencia que com pelo menos 70% de acerto, é possível identificar os alunos em risco de abandono, logo no final do 1º ano letivo. No conjunto dos 38 atributos preditivos analisados neste estudo, foi possível observar que a média do secundário, as classificações nas provas de ingresso ao ensino superior, a média do 1º ano na faculdade e a contribuição financeira facultada pela família são, por ordem decrescente, os principais fatores que melhor explicam a evasão discente. Por exemplo, pela análise de dados apresentada, verificou-se que os estudantes desistentes possuem uma média inferior em cerca de 0.9 valores, quando comparados com os alunos que persistem nos seus estudos. Mediante esta constatação, foi recomendado que se supervisionasse, de forma assídua, a média do 1º semestre, a fim de ser providenciado aconselhamento personalizado àqueles que venham a obter médias inferiores a 0.5 em relação aos restantes alunos.

3.3.3.2 Previsão precoce de desistências em unidades curriculares especificas

No estudo de Kotsiantis et al. [56] foi explorada informação demográfica e de desempenho académico, relativa a um conjunto de 354 estudantes, com o objetivo de prever de, forma precoce, quais os estudantes propensos ao abandono na disciplina de Introdução à Informática, ministrada no sistema de ensino à distância numa Universidade Grega. No conjunto dos atributos da categoria demográfica constam o género, idade, estado civil, número de filhos, ocupação, literacia informática e se o aluno tem ou não um trabalho associado a computadores. As notas dos estudantes nos dois primeiros trabalhos escritos e o facto de estarem presentes/ausentes nas duas primeiras reuniões presenciais, foram os atributos relacionados com a *performance* dos alunos. Para identificar quais são os fatores mais explicativos do abandono na referida disciplina, o conjunto de dados disponível foi processado em 5 iterações à medida que a informação ia ficando disponível.

Numa análise comparativa ao desempenho dos 6 algoritmos de aprendizagem usados neste caso de estudo (árvores de decisão, redes neuronais artificiais, algoritmos baseados em instâncias, regressão logística, *support vector machines* e Naive Bayes) o Naive Bayes foi o que revelou maior eficácia preditiva. O modelo resultante, aplicado ainda antes do meio do semestre letivo, demonstrou uma acurácia de 63% nas iterações iniciais, quando se usaram somente atributos demográficos e excedeu 83% na última iteração quando se usaram todos os atributos disponíveis (demográficos e os de *performance*).

A investigação de Márquez-Vera et al. [65] apresentou um modelo de classificação suficientemente confiável que permitiu prever, ainda antes do meio do 1º semestre letivo, o abandono escolar dos estudantes a frequentar o último ano do ensino secundário, matriculados na disciplina Trabalhos Preparatórios da Universidade Autónoma de Zacatecas, no México. Para o desenvolvimento do modelo foi considerado um conjunto de dados relativo a 419 estudantes, caracterizados de forma bastante abrangente por 60 variáveis explicativas das dimensões académicas e sócio-económica dos estudantes. Com o objetivo de se poder efetuar a previsão do abandono no momento mais precoce possível do percurso escolar, os autores foram avaliando, de forma iterativa, o desempenho dos algoritmos preditivos ao longo de 7 momentos temporais distintos. A avaliação do modelo ia sendo efetuada, em cada um desses momentos, mediante a inclusão de novas variáveis explicativas, conforme iam ficando disponíveis ao longo do período letivo. Os autores concluíram que entre a 4ª e a 6ª semanas letivas, altura em que já se dispôs de informação de cariz académico (assiduidade e desempenho), é possível prever com elevado acerto quais os estudantes do ensino secundário que não prosseguem para o ensino superior.

A fim de serem identificados os atributos mais informativos para a previsão da classe abandono/não abandono, foi efetuado, em cada uma das iterações, um estudo de seleção de atributos preditivos, que contribuiu para melhorar o acerto da previsão. Por essa via concluíram que os atributos mais influentes na previsão do abandono escolar, são, do primeiro ao último instante de recolha, respetivamente: média do ensino secundário, número de alunos por turma, idade, frequência do horário manhã/tarde, situação de empregabilidade e nível de escolaridade da mãe, o consumo regular de álcool e tabagismo, a assiduidade, local normalmente utilizado para estudar, nível de motivação, notas obtidas nas disciplinas de matemática, ciências sociais e humanidades. Numa análise que comparou o desempenho preditivo dos algoritmos usados neste estudo (Naives Bayes, *support vector machines*, *instance-based lazy learning* (IBK) e Jrip, e regras de classificação), foi verificado que o algoritmo ICRM2 de classificação interpretável foi aquele que demonstrou melhor desempenho na classificação.

No estudo de Delen [31] foram propostos modelos analíticos de classificação que demonstraram

elevada capacidade para prever, logo no final do 1º semestre académico, se os alunos recém ingressados numa universidade americana iam ou não desistir de estudar no final do seu primeiro ano letivo. O conjunto de dados usados na investigação contém registos relativos a 16.066 alunos, inscritos pela primeira vez na universidade, entre os anos de 2004 a 2008, caracterizados quanto às suas dimensões financeira, demográfica e de desempenho académico (no ensino secundário e no 1º semestre letivo).
Depois de terem sido comparados quanto à sua capacidade preditiva, 3 modelos desenvolvidos por métodos de conjunto (*bagging, busting e information fusion)* e 4 modelos desenvolvidos por classificadores individuais (redes neuronais artificiais, *support vector machines*, árvores de decisão e regressão logística), os autores concluíram que os modelos de conjunto superaram sempre o desempenho dos individuais. Nestes últimos, a acurácia obtida, por ordem decrescente, foi de 81.18%, 80.65%, 79.85% e de 79.85%, respetivamente, para os modelos de *support vector machines*, árvores de decisão, redes neuronais artificiais e de regressão logística. Após uma análise de sensibilidade, efetuada com o objetivo de identificar os fatores mais importantes para a previsão, foi revelado que o sucesso educacional obtido no ensino secundário e o suporte financeiro providenciados aos estudantes (familiar ou de bolsas de estudo) são os fatores mais importantes para explicar o abandono no final do 1º ano letivo dos estudantes. Estas previsões podem providenciar suporte à tomada de decisão para que, fundamentadamente, se possam elaborar programas de intervenção adequados e determinantes para a diminuição dos índices de abandono escolar, logo no 1º ano dos cursos de graduação.
Num outro estudo mais recente, Burgos et al. [15] apresentaram uma abordagem baseada em regressão logística, redes neuronais artificiais, *support vector machines* e num conjunto probabilístico simplificado com classificador *fuzzy adaptive resonance theory mapping*, a qual permitiu prever com um acerto de 97.13%, o abandono numa disciplina de informática, ministrada ao longo de 20 semanas, em contexto de ensino *e-learning*. O conjunto de dados usado neste estudo contém registos de 104 estudantes, caracterizados por 12 variáveis explicativas, relacionadas com as classificações obtidas nas 12 atividades de avaliação realizadas ao longo do curso no Moodle. Após uma análise comparativa à capacidade preditiva demonstrada pelos 4 modelos de classificação, os autores constataram que o modelo de regressão logística foi aquele que demonstrou maior exatidão, com uma acurácia de 97.13%, quando se usa a informação disponível entre as semanas de 9 a 13. Os modelos suportados pelas *support vector machines* e pelas redes neuronais artificiais apresentaram também uma boa capacidade preditiva, ao terem demonstrado uma acurácia de 94.23% e 85.58%, respetivamente. A proposta dos autores enfatizou o modelo incremental logístico, não só pelo facto de se ter relevado o mais eficaz, mas também por poder ser utilizado em diferentes instantes de tempo, conforme a informação vai ficando disponível. Por exemplo, não é possível incluir como atributo de entrada a nota obtida na última atividade de avaliação do curso, antes de ser realizada (Burgos et al. [15]). Baseados no conhecimento obtido pelo modelo de previsão proposto foi possível elaborar um plano de tutoria personalizado, aos estudantes do ano letivo 2014/15, com propensão para o abandono. Os autores sublinharam que a taxa de desistência reduziu 14%, relativamente aos anos académicos anteriores, onde nenhum plano de prevenção de abandono foi elaborado e aplicado (idem).

3.3.4 Sistemas de recomendação pedagógica e ambientes pessoais de aprendizagem

Os sistemas de recomendações pedagógicas (SRP) e os ambientes pessoais de aprendizagem (APA) estão diretamente ligados ao EDM (Huebner [53]). Os SRP servem para providenciar recomen-

dações e orientações personalizadas, a cada estudante individualmente, a fim de promover o seu sucesso educacional. De acordo com Romero and Ventura [95] os SRP são a melhor maneira de mostrar resultados, informações, bibliografia, explicações, recomendações e comentários para utilizadores que não são especialistas em *data mining*, como, por exemplo, os professores. Assim, em vez de se mostrar o modelo obtido, dever-se-á privilegiar uma lista de sugestões ou conclusões sobre os resultados e a melhor forma de os aplicar. Por vezes, os estudantes optam por aprender os conteúdos programáticos relativos às unidades curriculares através da internet, sem a interação professor-aluno. O APA é uma ferramenta que proporciona a aprendizagem online. Por via do APA, o aluno pode recolher os dados e toda a informação online, ajudando-o a interagir com os conteúdos da disciplina de forma fácil (Romero and Ventura [94]). Nesta secção descrevem-se alguns estudos científicos que evidenciam a aptidão e contribuição do *data mining* em tarefas de SRP e APA.

O estudo de Su et al. [109] aplicou o *data mining* para providenciar conteúdos de aprendizagem rápidos, dinâmicos e personalizados em dispositivos móveis, para se poder atender às diversas necessidades dos utilizadores. Os autores propuseram um *Personalized Learning Content Adaptation Mechanism* (PLCAM), para responder ao diverso e crescente número de pedidos de utilizadores. O *PLCAM* pode gerir eficientemente um histórico com um grande número de solicitações de utilizadores, entregar de forma inteligente e com com maior fiabilidade conteúdos de aprendizagem personalizados, através de um Repositório de Objetos de Aprendizagem (ROA). Uma versão de conteúdos adaptada, pode então ser preparada para um próximo pedido similar. Todas estas funcionalidades foram implementadas com recurso a árvores de decisão de adaptação de conteúdos (CADT) de *data mining*, que permitem determinar de forma eficiente o conteúdo adaptado de um ROA. Os autores destacaram duas principais contribuições. Primeiro, o esquema de adaptação de conteúdos propostos, extensível de forma a permitir representar dados relevantes, que podem ser processados por técnicas de *data mining*. Segundo, a elaboração de um esquema de gestão de adaptação de conteúdos de aprendizagem, proposta com o objetivo de pesquisar, recuperar e manter os dados históricos, bem como a elaboração de um processo de decisão que determine a adaptação, de forma eficiente, de uma versão de conteúdo personalizada e especificamente adaptada para o aluno.

No estudo de de Marcos et al. [28] foi investigada a relação entre o sucesso educacional e a estrutura de uma rede social, formada num sistema de ensino que utiliza "gamificação"[5]. O conjunto de dados considerado na investigação contém informação de um grupo de 161 estudantes, a frequentar a disciplina Qualificação para Utilizadores de Técnicas de Informação Computacional, das licenciaturas de Economia, Administração de Empresas e Ciências da Vida. Para análise, foram consideradas várias variáveis explicativas, como a proximidade central, a excentricidade, centralidade de interação, *ranking* de páginas, coeficiente de *clustering*. A nota final da disciplina foi a variável dependente do modelo. No decorrer do estudo, realizado durante um semestre, em 2013, foram testados e analisados vários métodos, como as medidas de centralidade básica, os algoritmos de análise de links, análise de *clusters*, coeficiente de correlação, análise fatorial e a regressão linear. Em contexto de aferição da capacidade preditiva dos modelos gerados, de salientar que as medidas dos algoritmos de análise de links (*hub*, autoridade e *pagerank*) só evidenciaram resultados significativos na análise de correlação. Para além disso, a análise de componentes principais revelou que essas medidas estão no mesmo componente principal que a maioria das medidas de centralidade (de Marcos et al. [28]). Esta constatação sugere que as medidas de centralidade têm mais poder preditivo do que as me-

[5]Termo derivado do inglês *gamification*, que designa o uso de jogos, para promover a aprendizagem e a motivação dos estudantes, através de ferramentas tecnológicas e interativas.

didas obtidas a partir de algoritmos de análise de links (idem). Os autores sublinharam que a posição do aluno na rede social parece influenciar o desempenho da aprendizagem na referida disciplina.

3.4 Fatores explicativos do desempenho académico

A literatura sugere que existe uma grande diversidade de fatores que, direta ou indiretamente, influem no desempenho académico dos estudantes do ensino superior. Em 3.4.1 enunciam-se as diversas categorias, bem como os respetivos fatores, que permitem, de um modo geral, caracterizar os estudantes do ensino superior. Em 3.4.2 apresenta-se uma revisão de literatura sobre os fatores que demonstraram relevância para a análise e previsão do desempenho académico, com recurso ao EDM — apresenta-se, no Apêndice B desta tese, a Tabela B.1 com a síntese da revisão efetuada, descrevendo o âmbito, objetivos e conclusões das obras consultadas.

3.4.1 Categorias de fatores determinantes do desempenho académico

Nos sistemas de informação de uma IES são armazenados dados que permitem caracterizar, com precisão, cada um dos seus estudantes, toda a sua oferta formativa e outros elementos intervenientes no processo educacional. De acordo com a literatura (e.g. Shahiri et al. [103], Del Río and Insuasti [30], Kumar et al. [58]), os atributos associados aos modelos preditivos de desempenho académico, baseados em técnicas de EDM, podem ser categorizados, tal como esquematizado na Figura 3.3, em:

Figura 3.3: Fatores determinantes do desempenho académico.

- Dados demográficos: inclui idade, género, estado civil, etnia, local de residência, naturalidade, nacionalidade e incapacidades/deficiências dos estudantes;
- Historial familiar e socioeconómico: formação académica e profissão dos pais, situação de empregabilidade do próprio e da família;

- Historial no ensino secundário: média no secundário, notas obtidas em disciplinas específicas, ano de conclusão do ensino secundário e escola de proveniência;
- Dados de acesso ao ensino superior: média de ingresso, classificação em provas específicas, média das provas de ingresso, tipo de ingresso e fase de ingresso;
- Historial no ensino superior: data de ingresso, curso de ingresso, tipo de ingresso, tipo de frequência, média de curso, média obtida em semestres específicos, notas a disciplinas específicas, notas em trabalhos específicos, ano de conclusão do curso, professor responsável por cada disciplina, método de avaliação a cada disciplina, número de unidades de crédito em que se inscreve em cada semestre/ano e concessão de bolsa de estudo;
- Historial comportamental: participação em grupos de discussão, número de *links* visitados, participação em sala de aula, acesso aos anúncios, frequência do horário de atendimento, número total de mensagens postadas em fóruns de discussão, comportamento de estudo e tempo total despendido no mesmo (*online e offline*);
- Atividades extra-curriculares: estudo de música, prática desportiva, interesses culturais, entre outros;
- Fatores psicométricos: perfil psicológico, motivação do aluno, apoio familiar, interação social com os colegas.

3.4.2 Revisão de literatura aos fatores determinantes do desempenho académico

Tal como anteriormente mencionado, os estudos de EDM, com ênfase na análise e previsão do desempenho académico, identificam uma grande diversidade de fatores que o podem influenciar. Na literatura encontram-se alguns estudos de revisão (e.g. Shahiri et al. [103], Del Río and Insuasti [30] e de Kumar et al. [58]), elaborados no intuito de apurar quais os fatores mais relevantes para a previsão e explicação do desempenho académico dos estudantes do ensino superior. Segue-se uma análise aos três estudos referenciados.

Shahiri et al. [103], a partir de uma análise a 30 artigos científicos, publicados entre 2002 e janeiro de 2015, concluiram que são 6 os atributos usados com mais frequência, em abordagens que visam a previsão de desempenho a partir do EDM. Em primeiro lugar, surge a média da pontuação acumulada (*cumulative grade point average*, CGPA), praticamente a par dos vários elementos de avaliação interna vistos de forma discriminada. Shahiri et al. [103] deram ênfase à preferência dos investigadores pelo uso da CGPA, por ter sido identificada como o atributo de maior capacidade preditiva, quando comparada com os restantes, em 10 dos 30 estudos em análise. Segundo os autores, tal deve-se ao facto dela derivar um valor tangível para determinar o futuro educacional e a perspetiva de carreira profissional, para além de ainda fornecer informações que podem servir de indicador de potencial académico. A CGPA foi ainda defendida como um atributo capaz de determinar o grau de sobrevivência dos estudantes no ensino, ou seja, de aferir a sua continuidade, ou desistência, num certo curso ou área de estudos. Os atributos de avaliação interna, indicativos do desempenho académico discriminado no ensino superior é, segundo os autores, o segundo mais frequente. A ele estão agregados os valores obtidos em trabalhos escolares, exames, trabalhos de laboratório, testes e assiduidade. Depois destes dois tipos de atributos, de acordo com Shahiri et al. [103], os que foram mais frequentemente escolhidos pelos investigadores dos 30 estudos que suportam esta revisão sistemática, são as

características demográficas e a avaliação externa. As características demográficas incluem, o género, a idade, o historial familiar e de incapacidades dos alunos. As características de avaliação externa correspondem a pontuações obtidas em exames finais de determinadas disciplinas do ensino secundário, à média final do secundário, e à nota em exames de acesso ao ensino superior. Entre as características demográficas, é dada ênfase ao género, designadamente porque, de acordo com alguns autores, os estudantes do sexo feminino são mais disciplinados e têm estratégias de estudo mais assertivas do que os estudantes do sexo masculino (Shahiri et al. [103, p. 417]). Por fim, o terceiro grupo de atributos que os autores consideraram mais importantes, no conjunto dos 6 mais usados, são os relacionados com atividades extra-curriculares e com a interação social dos estudantes. Acrescem a estes três principais grupos de atributos, outros, também usados por autores analisados neste contexto que, embora não tenham sido considerados com tanta frequência, como os atrás apontados, também contribuíram para explicar a previsão de desempenho dos estudantes. Entre eles, sublinham-se os fatores psicométricos, reportados em 4 dos 30 artigos, aos quais estão associados o perfil psicológico, a motivação e a influência familiar do estudante.

Por via do reagrupamento dos artigos referenciados, em função dos algoritmos de classificação utilizados, os autores evidenciaram também, para cada algoritmo, os atributos associados assim como a correspondente capacidade preditiva. De entre todos, as árvores de decisão (AD) foi o algoritmo usado com mais frequência, justificada pela sua simplicidade de utilização, capacidade de lidar com dados numéricos e categóricos, facilidade na interpretação, para além de ainda terem a capacidade de revelar pequenos pormenores, em bases de dados consideravelmente grandes. Entre os 10 artigos que reportaram o uso de AD, o que demonstrou apresentar melhor acurácia (91%) foi o estudo de Jishan et al. [54] que considerou como único atributo preditor a CGPA. Os estudos de Natek and Zwilling [79] e o de Elakia and Aarthi [32] apresentaram uma acurácia de 90%, tendo sido usados, respetivamente, um conjunto de 3 e 4 variáveis preditivas. Ao primeiro estudo, foram associadas a avaliação interna, características demográficas e atividades extra-curriculares e para o segundo estudo foram consideradas a CGPA, avaliação externa, características demográficas e atividades extra-curriculares. Para além dos atributos usados nestes três estudos, de uma análise global, constata-se que as AD foram também aplicadas a bases de dados com atributos de médias ponderadas finais, fatores psicométricos, bolsa de estudo e notas obtidas em disciplinas especificas. Entre todos, o grupo de atributos com menor influência nos modelos de previsão de desempenho obtidos com recurso às árvores de decisão foram os fatores psicométricos.

Shahiri et al. [103] consideram que as redes neuronais artificiais também se revelaram um algoritmo popular para o EDM, apresentando como principal vantagem a capacidade de detetar todas as interações possíveis entre as variáveis preditivas. Para além disso, é ainda salientado que este algoritmo revelou, em alguns dos estudos analisados, ter capacidade para detetar padrões em bases de dados complexas e não lineares. O estudo de Kumar and Vijayalakshmi [59] que considerou em junção os atributos de avaliação interna e de avaliação externa, foi o que apresentou a maior acurácia (98%) entre os 30 analisados, pelo que Shahiri, Husain, et al. [103] sustentam nestas características, a afirmação de que as "redes neuronais artificiais é um dos melhores algoritmos de previsão de que se tem conhecimento". Os autores mencionaram ainda que, em alguns dos estudos, as redes neuronais artificiais foram aplicadas em paralelo com as AD, para permitir análises comparativas que visavam identificar qual dos dois se revelaria mais preciso. Outros atributos também analisados noutros estudos, onde foi aplicado este algoritmo, foram as datas de admissão, as atitudes dos estudantes para com as estratégias de auto aprendizagem, características demográficas e fatores psicométricos. À semelhança das AD a

pior acurácia (69%) com recurso a redes neuronais artificiais foi revelada num estudo cujo único preditor se relacionava com questões psicométricas.

De entre os vários algoritmos usados no conjunto da literatura analisada na revisão de Shahiri et al. [103], o *K-Nearest Neighbor* foi o que comprovou demorar menos tempo de processamento computacional a identificar características, como, alunos de aprendizagem lenta, alunos médios, bons alunos e excelentes alunos. Para além disso, indica uma boa precisão na estimativa do padrão detalhado para a progressão do aluno no ensino superior. Por outro lado, a mais alta acurácia (83%), verificada em modelos de previsão de desempenho, no conjunto de artigos que usaram este algoritmo, obteve-se num estudo com a combinação de três atributos, que são a avaliação interna, CGPA e atividades extra-curriculares.

Uma outra particularidade que sobressaiu da revisão de literatura de Shahiri et al. [103] foi a verificação de que o algoritmo *support vector machines* é o mais indicado para a realização de previsões do desempenho escolar em conjuntos de dados de tamanho médio ou pequeno. De facto, os autores destacam que um dos artigos incluídos, assinado por Shovon and Haque [104] sustenta mesmo a afirmação de que este algoritmo apresenta uma boa capacidade de generalização e é mais rápido no precessamento do conjunto de dados do que todos os outros métodos. Também Gray et al. [43] defenderam as vantagens advindas das *support vector machines* e sublinharam-lhes a capacidade de obter elevados níveis de desempenho na previsão de estudantes em risco de falhar os objetivos académicos, ou seja, aqueles cujo sucesso não será positivo.

Num outro estudo de revisão, Del Río and Insuasti [30] analisaram 51 artigos, divulgados desde 2011 até agosto de 2016, para identificar quais são os preditores do desempenho académico dos estudantes do sistema presencial tradicional. Neste conjunto de estudos, embora a grande maioria visasse a previsão da média final de curso (CGPA), constam também alguns sobre a previsão da transição/não transição de ano letivo. Para inferir sobre este género de variáveis alvo, Del Río and Insuasti [30] concluem que os autores dos 51 estudos analisados neste contexto, usaram como variáveis preditivas indicadores do desempenho académico no ensino superior, em combinação com mais um outro tipo de atributo, em 51.8% dos estudos. Em 37.5% dos artigos analisados, foi considerada para análise somente informação sobre o desempenho académico obtido já no ensino superior.

Entre os métodos de *data mining*, usados na tarefa de EDM que deu ênfase a esta revisão literária, Del Río and Insuasti [30] destacaram o de classificação, dado que foi reportado em 71.4% das investigações referenciadas. Os métodos de segmentação e de regras de associação, foram os que se seguiram, tendo incidido, respetivamente, em 8.9% e 7.1% dos estudos. Ainda no contexto dos métodos usados, em jeito de conclusão, os autores, sublinham que se trata de instrumentos muito fiáveis, pois, se forem corretamente aplicados, está comprovado que providenciam elevados níveis de desempenho (Del Río and Insuasti [30]).

Na mesma linha de pensamento de Shahiri et al. [103] e de Del Río and Insuasti [30], mas com ênfase nas elevadas taxas de abandono no ensino superior em geral, e na Índia em particular, também Kumar et al. [58] pretenderam identificar os principais preditores do desempenho académico e a determinação dos principais métodos de *data mining* usados nesse âmbito.

Pesquisando via *Springer Link, Researchgate, Springer LinkIEEE, Xplore, biblioteca digital ACM, Elsevier, Science Direct* e de acordo com os critérios definidos no processo de *scoping*, os autores reuniram e analisaram 14 estudos subordinados ao tema, publicados entre 2009 a 2016.

Focados principalmente nos 10 artigos relacionados com previsão do abandono escolar, a abordagem de EDM protagonista neste artigo de revisão, Kumar et al. [58] concluíram que os investigadores analisados recorreram, essencialmente, aos métodos de classificação e associação,

com cerca de 50% dos trabalhos de investigação estudados a reportarem o uso destes dois métodos. Os algoritmos identificados neste conjunto de 10 estudos com ênfase no abandono, foram o CART, o C4.5, o JRip, o RF, o NB, o redes neuronais artificiais, o C5.0, o SVM, o algoritmo ID3 melhorado, o ICRM2, e análises de regressão linear e logística. Na maioria dos estudos de classificação, dentre os citados, prevalecem os algoritmos CART, C4.5, J48 e JRip, e em estudos de regressão predomina o RF (Kumar et al. [58, p. 457]).
No que diz respeito à identificação dos diferentes atributos estudados em abordagens de previsão do desempenho dos alunos com recurso ao EDM, Kumar et al. [58] contabilizaram os sociais, os demográficos, os pessoais e familiares e os de desempenho académico, quer no ensino secundário quer no ensino superior. De salientar que Kumar et al. [58] sublinharam ainda que, na maioria dos casos, a média de acesso, o nível de escolaridade e a ocupação dos pais, e uma metodologia de ensino pobre são os principais fatores que afetam o resultado escolar dos alunos. Uma vez que a ênfase da revisão literária foi o abandono escolar, Kumar et al. [58] também destacaram os atributos que os autores de 10 estudos revelaram como os mais informativos para a previsão do abandono. Por ordem decrescente, enumeraram: o género, a estrutura familiar e a profissão dos pais, más metodologias de ensino adotadas, nível educacional dos pais e se o estudante tem vícios (álcool, tabaco, pílulas, solventes, drogas), a necessidade da execução de tarefas domésticas e o estado civil.
Kumar et al. [58] concluíram que os resultados dos estudos sobre a previsão do desempenho académico são muito úteis, tanto para a adoção de novas metodologias pedagógicas, como para o desenvolvimento e a implementação de novas regras e regulamentos no ensino. Para além disso, a previsão do abandono escolar é tida como uma tarefa importante e desafiadora, para os investigadores, professores, instituições de ensino e até para os decisores políticos, confirmando-se a elevada contribuição do EDM em tarefes desta tipologia, nomeadamente, quando se pretende identificar as características associadas aos estudantes que abandonam a sua formação (idem).

3.5 Resumo e conclusão

Ao longo deste capítulo foi apresentada uma revisão de literatura, regressiva à própria história científica do EDM, que permite compreender a sua evolução e contribuição para o entendimento dos processos de ensino e de aprendizagem dos estudantes.
Apesar da relevância que o EDM tem adquirido, como é percetível a partir das conclusões dos trabalhos apresentados, alguns autores destacam algumas lacunas que dificultam o processo de consolidação do EDM. Por exemplo, Huebner [53] adverte que a investigação nesta área desenvolve-se de forma isolada, não se conhecendo, com exatidão, a forma como as instituições têm implementado as metodologias que visam contribuir para a promoção da aprendizagem dos seus alunos, ou os respetivos processos educacionais. Peña-Ayala [86] aponta, igualmente, como aspeto penalizador, a falta de reconhecimento daquelas que são as verdadeiras capacidades do EDM, para ampliar e melhorar as conquistas tradicionais dos sistemas educacionais. Através do seu estudo de revisão de literatura, concluiu que 66% do conjunto de estudos com ênfase na modelação e avaliação de desempenho dos estudantes correspondem a abordagens ainda incipientes. O mesmo autor realça ainda que a maioria das abordagens existentes relacionam-se, essencialmente, com a implementação do *data mining* para explorar assuntos educacionais, não dando propriamente contributos para o progresso do EDM. Sukhija et al. [110] referem a indisponibilidade de dados consistentes e suficientemente abrangentes. Sublinham, também, que as próprias ferramentas de EDM continuam a exigir a atenção dos investigadores, sobretudo

ao nível do aumento e utilização dos métodos de conjunto. Também Romero and Ventura [95] incitam a que no seio das IES seja desenvolvida uma cultura baseada em dados, a fim de ser melhorada a qualidade das decisões de gestão.

Da revisão de literatura apresentada ao longo do presente capítulo, também é de assinalar que a generalidade dos estudos de revisão sistemática de literatura analisados evidenciam a predominância de investigações relacionadas com o sistema de ensino à distância, em comparação com as efetuados no ensino presencial tradicional. Esta constatação apela ao desenvolvimento de novos estudos no contexto das IES de ensino presencial tradicional.

De realçar, também, que os estudos subordinados à previsão do sucesso educacional são efetuados ao nível de um determinado curso especifico, ou então, ao nível de certas unidades curriculares. No entanto, não foi possível encontrar nenhum estudo, que contemplasse a previsão do sucesso educacional global dos estudantes, ao nível do universo de uma IES onde fossem ministradas diversas áreas educativas.

Sobressai, igualmente, que em relação aos métodos e algoritmos mais usados no âmbito de EDM, prevalecem os de previsão, conseguida pela regressão e classificação, seguindo-se, com muito menor incidência, o *clustering* e a associação. Dentre os métodos de classificação, sobressaí que há relativamente menos artigos que reportam o uso de métodos de conjunto, o que abre perspetivas para que se investigue a pertinência da sua utilização no âmbito do EDM. De igual modo, há relativamente menos artigos que têm recorrido a métodos de visualização de *data mining*. Tratando-se de um método que muito contribuí para melhorar a perceção dos padrões ou relacionamentos descobertos, dever-se-á considerar, de igual forma, o seu uso, em complemento aos outros métodos.

A revisão de literatura apresentada também permite identificar os fatores que se revelaram determinantes para prever e explicar o desempenho académico. Neste contexto, sobressai que os relacionados com indicadores de desempenho académico, pré e pós-ingresso no ensino superior, têm sido, globalmente, os que apresentam maior poder explicativo nas previsões efetuadas. De salientar, no entanto, que há um conjunto de outros fatores que não têm sido reportados, mas que poderão, eventualmente, desempenhar um papel significativo na previsão do desempenho. Por exemplo, informação sobre o círculo de amigos — que poderão influenciar os hábitos de sono, os métodos de estudo, a assiduidade às aulas, a frequência de horário de atendimento —, o tamanho da turma, o ambiente de sala de aula, a tecnologia usada, o tipo de avaliação e as características dos professores, são dimensões que também poderão ter um vínculo direto com o desempenho académico. Mas, tal como foi referido pela generalidade dos autores, outras e novas tendências, como por exemplo, o suporte aos alunos e aos professores, a padronização de métodos de *data mining*, a padronização do conjunto de dados a estudar e até as próprias ferramentas de *data mining*, reivindicam o foco e o interesse da comunidade de EDM.

Por todos estes motivos, a generalidade dos autores realça a necessidade de desenvolvimentos metodológicos futuros, com vista à obtenção de modelos práticos de aplicação. Por conseguinte, existe uma necessidade de intensificar os estudos de investigação, analisando outros grupos de estudantes, de outros cursos e instituições, a fim de promover e valorizar o EDM como instrumento de análise e construção de estratégias nos processos de ensino.

Face às debilidades identificadas, nos Capítulos 5 e 6 são propostos novos modelos analíticos de apoio à gestão de IES, com vista à produção de conhecimento científico que possa contribuir para o progresso desta recente área de investigação, denominada *educational data mining*.

Capítulo 4

Caso de estudo: o Instituto Politécnico de Bragança

4.1 Introdução

Tendo-se como objetivo explorar a aplicação de técnicas de *data mining* na compreensão e previsão do desempenho académico dos alunos que frequentam as licenciaturas do IPB, neste capítulo, apresenta-se uma breve descrição da instituição usada como caso de estudo e da sua oferta formativa, dando-se especial destaque à caracterização dos cursos conducentes ao grau de licenciatura. Termina-se com a descrição das bases de dados disponibilizadas pelo IPB para objeto do atual estudo e com a apresentação da ferramenta, ou *software*, selecionada para aplicar os algoritmos de *data mining*.

4.2 Caraterização do IPB[1]

4.2.1 Descrição geral

O Instituto Politécnico de Bragança (IPB) é uma instituição pública de ensino superior que tem por missão a criação, transmissão e difusão do conhecimento técnico-científico e do saber de natureza profissional, através da articulação do estudo, do ensino, da investigação orientada e do desenvolvimento experimental. Fundado em 1983, o IPB é constituído por cinco escolas; quatro no Campus de Bragança e uma em Mirandela:

- Escola Superior Agrária de Bragança (ESA);
- Escola Superior de Comunicação, Administração e Turismo de Mirandela (EsACT);
- Escola Superior de Educação de Bragança (ESE);
- Escola Superior de Saúde de Bragança (ESSa);
- Escola Superior de Tecnologia e Gestão de Bragança (ESTiG).

Tipicamente, cada escola possui uma organização matricial, onde cada departamento leciona unidades curriculares (UCs) de vários cursos.
A sua atividade abrange uma vasta área do saber e da tecnologia, nomeadamente, as artes, comunicação e multimédia, o turismo, desporto e lazer, a educação e a formação de professores,

[1] A caraterização apresentada foi essencialmente baseada na informação disponível no portal do IPB (http://portal3.ipb.pt).

a saúde e protecção social, as ciências empresariais e o direito, as ciências agrárias e recursos naturais e as tecnologias. O IPB consolidou a sua dimensão em cerca de 7000 estudantes e concretizou a adequação ao Processo de Bolonha através da oferta de cerca de uma centena de formações de cursos de especialização tecnológica, licenciaturas e mestrados. O IPB é hoje em dia uma instituição multicultural, onde 10% dos seus estudantes possuem nacionalidade não portuguesa. O IPB oferece a todos os seus alunos, nacionais e internacionais, uma oportunidade única de estudar numa instituição criativa e inovadora e de desfrutar de um ambiente académico, cultural e de enquadramento paisagístico verdadeiramente único.

O IPB é a única instituição da região que tem conseguido atrair e fixar jovens qualificados, provenientes de outras regiões, nomeadamente do litoral, contrariando a tendência verificada nas décadas anteriores à sua consolidação como instituição de ensino superior. A sua população estudantil representa cerca de 20% da população do concelho de Bragança e mais de 30% da do perímetro urbano. Entre os pontos fortes do IPB poder-se-á apontar a forte componente de ensino experimental e prático dos seus cursos, a existência de um corpo docente altamente qualificado nas diversas áreas de estudo e de linhas de investigação com elevado índice de produtividade científica. Para além disto, no contexto europeu, o IPB encontra-se bem posicionado no *ranking* das instituições de ensino superior com maior receção de professores em mobilidade Erasmus e com maior mobilidade de alunos. Um dos seus principais pontos fracos será a sua localização numa região interior e de baixa densidade populacional e todos os constrangimentos associados, nomeadamente, a ainda insuficiente ligação às empresas e ao mercado de trabalho e a dificuldade na fixação de jovens diplomados na região, atendendo ao ainda débil tecido empresarial. O elevado insucesso escolar e abandono que se vêm verificando na instituição também não serão alheios a este tipo de constrangimentos.[2]

4.2.2 Cursos do IPB conducentes ao grau de licenciatura

A oferta formativa do IPB é atualmente composta por mais de uma centena de cursos e ciclos de estudos, incluindo cursos de especialização tecnológica (CETs), cursos técnicos superiores profissionais (CTeSPs), licenciaturas, pós-graduações e pós-licenciaturas e ciclos de estudos de mestrado.

A componente de maior importância da oferta formativa do IPB é, claramente, o 1º Ciclo de Estudos, que confere aos seus diplomados o grau de licenciatura, abrangendo todas as áreas de formação do Ensino Superior, adequadas ao Processo de Bolonha e de acordo com o regime europeu de créditos (ECTS).

O ciclo de estudos conducente ao grau de licenciado tem 180 créditos e uma duração normal de seis semestres curriculares de trabalho dos alunos. Excetuam-se do disposto no número anterior os casos em que seja indispensável, para o acesso ao exercício de determinada atividade profissional, uma formação de até 240 créditos, com uma duração normal de até sete ou oito semestres curriculares de trabalho, em consequência de normas jurídicas expressas, nacionais ou da União Europeia, ou de uma prática consolidada em instituições de referência de ensino superior do espaço europeu. Para uma caracterização mais completa do 1º Ciclo de Estudos típico ministrado no IPB, consultar o Apêndice C.

[2] Deduções baseadas em relatórios da autoavaliação institucional de licenciaturas do IPB para a Agência de Avaliação e Acreditação do Ensino Superior (A3ES).

4.3 Descrição da base de dados do IPB

Para a recolha dos dados que vieram a ser objeto de estudo neste trabalho de doutoramento, contou-se com a preciosa colaboração do Pró-Presidente do IPB para os Sistemas de Informação, responsável institucional pelos Serviços de Informática do IPB e, em particular, pelo seu Centro de Desenvolvimento e Gestão de Dados. Com a sua intermediação foi possível reunir um conjunto vasto de informação relacionada com o IPB, distribuída por três bases de dados reacionais MySQL:

acesso – base de dados com os dados de acesso dos Concursos Nacionais de Acesso ao Ensino Superior, incluindo também alguns dados do historial do ensino secundário;

inqueritos – base de dados com as respostas dos alunos aos inquéritos preenchidos no ato da matrícula nos cursos do IPB, destinados a recolher dados familiares e socioeconómicos do aluno.

sa – base de dados dos Serviços Académicos do IPB, com os dados académicos e demográficos de toda a sua população estudantil e os relacionados com toda a sua oferta formativa.

Na Tabela 4.1 mostra-se o conjunto de tabelas que integram as três bases de dados, objeto do presente estudo, com indicação, para cada tabela, do número de observações, número de atributos e duma breve descrição do respetivo conteúdo. A lista completa dos atributos de cada

Tabela 4.1: Conjunto de tabelas integradas nas bases de dados disponibilizadas

bd	nome da tabela	nºobs×nºatrib	conteúdo
acesso	candidato	501.489×3	identificação do candidato
	candidatura	2.842.597×12	dados de candidatura
	curso	9.826×6	cursos
	estabelecimento	1.401×4	estabelecimentos de ensino superior
	exame	203×4	identificação dos exames
	exame_pi	218×4	identificação das prov. ingresso e exame
	nota_exame	419.006×6	notas de exame
	nota_exame_pi	27.202×6	notas das provas de ingresso
	pi	40×3	identificação das provas de ingresso
inqueritos	alunos_estat	16.151×13	dados familiares e socioeconómicos
	nivel_escolar	13×2	tipos de níveis de escolaridade
	sit_profissinal	10×2	tipos de situações profissionais
	tipo_profissao	12×3	tipos de profissões
sa	alunos	36.656×12	identificação e caraterização dos alunos
	concelhos	309×3	concelhos do país
	cursos	330×5	cursos do IPB
	disciplinas	15.150×8	unidades curriculares
	distritos	30×2	distritos do país
	epocas	47×2	épocas de avaliação
	escolas	9×3	escolas do IPB
	freguesias	4.257×4	freguesias do país
	lect_ini	140.794×13	dados de inscrição nos anos letivos
	matriculas	57.576×6	dados de matrícula
	notas	1.638.361×11	notas dos alunos
	opcoes	6.262×7	unidades curriculares opcionais
	planos	677×5	planos estudos dos cursos
	stat_medias_conclusivos	25.034×6	média final de curso dos diplomados
	tipo_estatutos	12×2	tipos de estatutos do aluno
	tipo_frequencias	13×2	tipos de frequência do aluno
	tipo_ingresso	23×2	tipos de ingresso
	tipo_nota	14×3	tipos de nota

tabela pode, por sua vez, ser consultada na Tabela D.1 do Apêndice D desta tese.
Para uma rápida perceção do período a que os dados reportam, na Tabela 4.2, mostra-se, de uma forma gráfica, os intervalos de tempo a que reportam os dados contidos nas diferentes BDs disponibilizadas. Para cada BD, apenas são mostrados os pares tabela/atributo de natureza cronológica (datas e anos, civis ou letivos).

Tabela 4.2: Intervalos de tempo a que reportam os dados contidos nas tabelas das bases de dados.

BD	Tabela	Atributo	Período de	Período a	Anos abrangidos
acesso	exame	ano	2007	2008	
		ano_ex	2006	2008	
	exame_pi	ano	2007	2008	
		ano_ex	2006	2008	
	nota_exame	ano	2007	2008	
		ano_ex	2006	2008	
	pi	ano	2007	2008	
	candidato	ano	2007	2015	
	candidatura	ano	2007	2015	
	curso	ano	2007	2015	
	estabelecimento	ano	2007	2015	
inqueritos	alunos_estat	[a]	1986	2016	
sa	lect_ini	ano_lect	1986	2016	
		data	8/10/86	18/1/17	
	matriculas	ano_mat	1986	2016	
	notas	ano_lect	1986	2016	
		data	19/1/87	1/12/16	
	opcoes	ano_lect	1994	2016	
	stat_medias_concl	ano_lect	1988	2015	

86 87 88 ----- 94 ---------- 06 07 08 --- 12 13 14 15 16 17

[a] Ano de matrícula do aluno inquirido.

4.4 Ferramentas e ambiente de desenvolvimento adotados no presente trabalho

Tal como mencionado no Capítulo 2 existe um vasto conjunto de ferramentas informáticas, ou software, que implementam os algoritmos de data mining. De entre as disponíveis a principal ferramenta escolhida para a presente investigação foi o software estatístico R, uma linguagem de programação de alto nível, interpretada e de tipo matricial, vocacionada para o cálculo estatístico e a análise de dados, cuja popularidade não para de aumentar. Tal opção justifica-se também pelo facto de correr nas várias plataformas e ambientes computacionais e estar disponível em *open source* [88], para além de ser frequentemente apresentada como uma das ferramentas mais utilizadas pela comunidade científica da área de *data mining*[3]. Tem também a particularidade de permitir aplicar uma ampla gama de algoritmos e estruturas de modelação de dados que podem ser usados para efeitos de previsão e para estabelecer relações em dados educacionais (Slater et al. [106]).
Na investigação que se descreve nesta tese, todos o cálculos computacionais relacionados com o processo de *data mining* foram realizados no RStudio, um ambiente integrado de desenvolvimento (IDE) para o R, que inclui, entre o seu vasto conjunto de ferramentas e funcionalidades, um editor que possibilita a execução direta de código R. O RStudio é um software que corre nas várias plataformas e está disponível quer em *open source* quer em versões comerciais [97]. Para a aplicação dos métodos de *data mining* em particular, usaram-se, como extensões ao próprio R, *packages* específicos, disponibilizados para o efeito pela comunidade científica. Foi ainda

[3] Ver, por exemplo, website www.Kdnuggts.com.

usado o *package* `RMySQL` [81] — uma interface para a base de dados MySQL — na importação para o R, a partir do MySQL Server, das tabelas com os dados entretanto tratados.

Capítulo 5

Previsão de sucesso académico global

5.1 Introdução

A previsão do desempenho académico é uma das mais antigas e populares tarefas de EDM (Romero and Ventura [95]), que continua a ser amplamente realizada (Del Río and Insuasti [30]), em virtude de poder ser muito útil na antecipação das dificuldades de aprendizagem dos estudantes e na antevisão de oportunidades de melhoria no seio das IES.
Neste enquadramento, o presente capítulo apresenta um novo modelo analítico de regressão, desenvolvido com o objetivo de prever, de forma precoce, o desempenho académico global dos estudantes das licenciaturas do IPB. Em simultâneo, são também identificados e apresentados os principais fatores que explicam o (in)sucesso educacional global dos mesmos estudantes.
Face ao conhecimento obtido por via do modelo apresentado, este capítulo também sugere algumas recomendações promotoras do sucesso educacional e das taxas de graduação dos estudantes, e, por consequência, da eficiência institucional.
A estrutura do restante capítulo é a que se segue. Na Secção 5.2 apresenta-se a motivação e os objetivos do estudo. Na Secção 5.3 apresenta-se a metodologia desenvolvida face aos objetivos estabelecidos. Na Secção 5.4 descreve-se todo o pré-processamento necessário à completa definição do modelo de dados a usar, para se passar, na Secção 5.5, a explorar modelos de previsão baseados no algoritmo *random forest* que melhores resultados apresentam. Na Secção 5.6 são discutidos os resultados obtidos face aos procedimentos adotados. Por fim, na Secção 5.7 apresenta-se um breve resumo do capítulo e são destacadas as contribuições da investigação para a literatura de EDM. Em consonância são apresentadas algumas sugestões que possam contribuir para a promoção do desempenho académico.

5.2 Motivação e objetivos

No relatório recentemente divulgado pelo Conselho Nacional da Educação [73], sobre o estado da educação no ano de 2016 em Portugal, é sublinhada a necessidade e a importância de se continuar a apostar em estratégias promotoras do sucesso escolar. É igualmente realçado, no mesmo relatório, que o sistema educacional apresenta grandes dificuldades em recuperar alunos que apresentam desempenhos escolares negativos. Estes apontamentos, de tão elevada relevância no seio de uma IES, têm sido alvo de sérias preocupações e reflexões ao nível da gestão do IPB, sobretudo, porque a eficiência e a eficácia institucional dependem muito do sucesso escolar dos seus discentes. A fim de delinear estratégias promotoras de sucesso académico eficazes, a gestão institucional necessita de informação concreta, oportuna e atempada sobre o potencial de desempenho dos estudantes e sobre os principais fatores que o expliquem. Foi devido à

inexistência de qualquer estudo desta natureza, no contexto da instituição usada como caso de estudo, que surgiu a necessidade de desenvolver um modelo preditivo de desempenho académico capaz de gerar conhecimento útil que possa auxiliar e fundamentar a tomada de decisões. Mais precisamente, pretende-se prever, logo numa fase precoce, o sucesso académico global do aluno no *terminus* do seu percurso escolar. Para o efeito, desenvolve-se um novo modelo preditivo de regressão, que use como variáveis explicativas, dados académicos, demográficos, socioeconómicos e de acesso ao ensino superior, dos estudantes de licenciatura do IPB.
O sucesso académico, é inferido a partir de uma variável contínua expressa na equação 5.1. Esta variável acomoda dimensões relativas à média final de curso obtida pelo estudante e à repetição, ou não, de inscrições nas disciplinas.
A metodologia adotada no âmbito deste estudo visa a aplicação de métodos e algoritmos de *data mining*. A opção por esta abordagem é fundamentada pelo facto da revisão de literatura, apresentada nos Capítulos 2 e 3, ter comprovado o potencial e a elevada contribuição dos referidos métodos para o entendimento do tema. Esta metodologia é apresentada na secção que se segue.

5.3 Metodologia

Com o objetivo de prever, de forma precoce, o desempenho dos estudantes de licenciatura do IPB no *terminus* do seu percurso académico, começa-se por selecionar, através dum estudo comparativo, a melhor combinação de categorias de variáveis que irão ser consideradas no modelo de previsão adotado. Encontrada a melhor combinação de categorias, procura-se, numa fase seguinte, ajustar melhor o conjunto de variáveis, retirando do modelo todas aquelas que não se revelem verdadeiramente importantes para a previsão pretendida.
O *dataset* que suporta a investigação é obtido a partir das 3 bases de dados disponibilizadas, previamente descritas na Tabela 4.1 do Capítulo anterior.
Para a aplicação dos métodos de *data mining*, recorreu-se ao *package* `randomForest` [61, 62], uma extensão do R que contém algoritmos de regressão e classificação baseados no método *random forest* (RF), implementados originalmente por Breiman [14], e ainda ao *package* `caret` [41], como ferramenta de apoio na criação das partições do *dataset* usadas para treino e teste na validação cruzada, a técnica de *resampling* usada no processo de aferição da capacidade do modelo proposto.
Com efeito, todas as simulações em que se testou o modelo preditivo foram realizadas através do método de validação cruzada *k-fold*, previamente descrito na Secção 2.5.2. Mais concretamente, no caso do estudo que se desenvolveu optou-se por uma validação cruzada de 10 partições ($k = 10$). Para os parâmetros de configuração do método randomForest usaram-se os seus valores definidos por omissão, destacando-se, de entre eles, o número de árvores a construir (`ntree=500`) e o número de variáveis que serão aleatoriamente escolhidas, como candidatas a servirem de critério na divisão realizada em cada um dos nodos (`mtry=max(floor(ncol(x)/3), 1)` $-$ aproximadamente um terço do nº de preditores). De salientar que vários outros valores foram testados, quer para o número de árvores quer para o número de variáveis escolhidas como candidatas em cada divisão, não se conseguindo, no entanto, identificar uma combinação que resultasse em melhores níveis de desempenho dos modelos. Ainda assim, convém realçar, que em investigações que se pretendam mais profundas, dever-se-á realizar a afinação adequada desses parâmetros, em particular do `mtry`, através de uma metodologia mais sistematizada, seguindo o procedimento normalmente adotado na literatura de referência.

Refira-se, por fim, que foram usados, como métricas de avaliação de desempenho do modelo proposto, o coeficiente de determinação R^2 e o erro quadrático médio RMSE, previamente descritos na Secção 2.5.1.

5.4 Definição do modelo de dados dos estudantes de licenciatura

Com o objetivo de desenvolver o modelo de dados que permita identificar os principais fatores que possam influenciar de alguma forma o sucesso académico, nesta fase, procedeu-se à recolha de toda a informação pertinente sobre os estudantes. Esta informação inclui a que caracteriza os ambientes escolares, familiares e socioeconómicos em que os mesmos estão, ou estiveram, inseridos. A informação reunida provém das diferentes BDs disponibilizadas, conforme referenciadas nas Tabelas 4.1 e 4.2.

5.4.1 Definição do indicador de sucesso

O ingresso dos estudantes nas licenciaturas do IPB pode ocorrer em diferentes condições e nem todos os que ingressam na instituição são bem sucedidos na obtenção do grau académico, pelas mais variadas razões. Mesmo entre os que concluem com sucesso a sua formação, é possível identificar perfis bastante diferenciados.
Um indicador de sucesso académico deve ser, tanto quanto possível, imune a todas as especificidades do estudante que não estejam propriamente relacionadas com o seu desempenho académico. Por outro lado, um bom desempenho académico não se deve resumir à obtenção de uma classificação média elevada nas unidades curriculares concluídas pelo aluno. Deve, igualmente, refletir outros fatores reveladores de sucesso, como, a fração de inscrições em unidades curriculares que tenham resultado em aprovações (percentagem de "tentativas" bem sucedidas) e o tempo usado, em semestres frequentados, por cada unidade de ECTS aprovada. Assim, propõe-se que o desempenho académico dos estudantes seja aferido tendo em conta o indicador apresentado na expressão 5.1. Uma vez que se entende que o tempo gasto pelo aluno já está, de certa forma, a ser contabilizado no rácio das inscrições bem sucedidas, optou-se por não considerar este último fator, por forma a evitar uma dupla penalização — basta lembrar que são as reprovações às unidades curriculares que, por norma, prolongam no tempo o percurso escolar do aluno.
Assim, um possível indicador de desempenho académico do aluno, que evidencie as contribuições enunciadas, pode ser expresso simplesmente por:

$$\texttt{vd} = \texttt{media} \times \frac{\texttt{ects\_aprov}}{\texttt{ects\_aprov+ects\_reprov}} \quad (5.1)$$

onde:

- `media` é a média ponderada das notas obtidas nas unidades curriculares já concluídas;
- `ects_aprov` é o nº de ECTS concluídos com sucesso;
- `ects_reprov` o nº de ECTS em que o aluno se inscreveu sem que tenha conseguido aprovação.

Por exemplo, de acordo com a métrica proposta, um aluno que, em todas as unidades curriculares, tenha conseguido aprovação apenas à segunda inscrição, verá o seu desempenho ser reduzido a metade.

Observe-se que a fórmula considerada pode ser usada para avaliar o desempenho do aluno em qualquer período do seu percurso escolar. Quer no momento de conclusão do seu curso, por forma a obter-se o desempenho final, quer numa fase ainda intermédia do mesmo, ou, até mesmo, em semestres específicos. Estas avaliações intermédias e parcelares, podendo ser calculadas ainda numa fase prematura do percurso escolar do aluno, revelar-se-ão úteis quando se definir o conjunto de variáveis preditivas indicativas da evolução de desempenho do aluno nos seus primeiros semestres.

5.4.2 Seleção e limpeza dos dados

Na Tabela 4.2 destaca-se, na área retangular mais escurecida, o subconjunto de dados (tabelas e intervalos de observações) que foi selecionado para o estudo que se pretende elaborar. De salientar que fazem igualmente parte do subconjunto de dados selecionado, muitos outros atributos e tabelas que, ao não incluírem datas, não são mostrados, como é o caso, por exemplo, de tabelas e atributos relacionados com a identificação dos alunos, cursos, etc.
Optou-se por considerar o período temporal compreendido desde 2007/2008 a 2015/2016, perfazendo 9 anos letivos consecutivos. Nos 7 primeiros (de 2007/2008 a 2013/2014) ingressam na instituição 7 grupos distintos de estudantes que têm a possibilidade de concluir o ciclo de estudos de licenciatura — que, como se sabe, e salvo raras exceções, têm a duração de 3 anos letivos.
O horizonte temporal definido corresponde ao maior período de tempo em que há observações registadas para a quase totalidade dos atributos presentes nas base de dados. Só para um subconjunto muito restrito de atributos não existem observações nesse período, tendo sido por isso desconsiderados. Trata-se de atributos com informações muito específicas, relacionadas com o desempenho do aluno discriminado nas várias provas de ingresso. Toda essa informação acaba por estar presente, ainda que de uma forma mais agregada, noutros atributos considerados no estudo, como será o caso da nota de ingresso, média do secundário e média das provas de ingresso.
De salientar que, por facilidade de representação e exposição, os anos letivos são muitas vezes identificados, tanto no presente trabalho como nas BDs originais, não pelo par de anos que lhes correspondem mas sim unicamente pelo ano em que se iniciam. É por essa razão que a área selecionada na tabela, ao terminar no ano 2015, também incluí todo o ano letivo 2015/2016.
Optou-se por restringir o estudo apenas aos cursos de licenciatura, por se tratar do *core* da oferta formativa da instituição (cf. §4.2.2) e por abranger um conjunto de dados mais completo, quer na quantidade de observações registadas nas BDs por ano letivo, quer no horizonte temporal a que os dados se reportam. Como se percebe da Tabela 5.1, as matrículas em cursos de licenciatura correspondem a 71% do total da amostra, quando, por exemplo, quer as matrículas em mestrados quer as matriculas em CETs não vão além dos 8% da amostra. Com a Tabela 5.1 pretende-se explicitar o tamanho e, em especial, a representatividade, no conjunto de dados original, da amostra que é selecionada para objeto de estudo.
O que melhor identifica o elemento central do estudo é a matrícula, e não o aluno, uma vez que será sobre o percurso curricular subsequente a cada matrícula que incidirá a análise de desenho que se pretende elaborar, com recurso a métodos de *data mining* — basta recordar que um aluno pode ter várias matrículas na instituição. Assim, as quantidades mostradas na Tabela 5.1 referem-se ao número de matrículas por cada subconjunto de dados considerado do *dataset* e discriminadas por tipo de formação. Para uma correta interpretação dos dados tabelados, deve-se considerar que cada subconjunto de matrículas, identificado na primeira

coluna, representa um subconjunto das matrículas da linha anterior, ou seja, é um subconjunto mais restritivo que o anterior. Por exemplo, pelo valor da célula da 4ª coluna e última linha percebe-se que estão registadas nas BDs 6544 matrículas em cursos de licenciatura, concluídas no período 2007/08—2015/16 e iniciadas dentro do período 2007/08—2013/14. É precisamente este conjunto de matrículas — que virá ainda a ser reduzido para 4530 na fase de limpeza e uniformização de dados — que será alvo de estudo na previsão de sucesso académico que se pretende elaborar.

Tabela 5.1: Número de matrículas nos subconjuntos de dados selecionados (cada subconjunto de matrículas considerado está contido no subconjunto anterior).

Subconjunto de matrículas	Total		Licenciat.		Mestrados		CETs		TeSPs		Pós-Grad.		Bacharel.	
todas	55000	100%	39048	71%	4156	8%	4642	8%	1025	2%	640	1%	5489	10%
iniciadas no período 2007/08—2015/16	26487	48%	16838	31%	3804	7%	4576	8%	632	1%	637	1%	0	0%
iniciadas no período 2007/08—2013/14	20520	37%	13381	24%	2936	5%	3729	7%	0	0%	474	1%	0	0%
concluídas no período 2007/08—2015/16	10825	20%	6544	12%	1277	2%	2633	5%	0	0%	371	1%	0	0%

Os valores mostrados na Tabela 5.1, de certa forma, refletem a primeira seleção de dados que foi necessária realizar sobre a amostra inicial (área selecionada da Tabela 4.2), a qual se traduziu na remoção dos seguintes subconjuntos de dados:

1. todos os dados associados a matrículas de alunos que tenham ocorrido em anos letivos anteriores a 2007/2008 ou posteriores a 2015/2016, passando a amostra a representar 48% de todas as matrículas da BD — 3ª linha da Tabela 5.1;
2. todos os dados associados a matrículas de alunos que tenham ocorrido nos anos letivos 2014/2005 ou 2015/2016, uma vez que os dados disponíveis, para esses alunos, nem sequer chegam a abarcar 3 anos letivos, o número mínimo de anos que necessitam para poderem concluir os seus cursos, passando assim a amostra para 37% de todas as matrículas da BD — 4ª linha da Tabela 5.1;
3. todos os dados associados a matrículas de alunos que não tenham concluído os seus cursos no período de análise 2007/2008—2015/2016, pelo facto de se tratar duma previsão de sucesso, passando a amostra para 20% da sua dimensão inicial — 5ª linha da Tabela 5.1;
4. todos os dados associados a matrículas de alunos em cursos não conducentes ao grau de licenciatura (cursos de especialização tecnológica (CETs), cursos técnicos superiores profissionais (TeSPs), pós-graduações, pós-licenciaturas e ciclos de estudos de mestrado), ficando-se, finalmente, com um conjunto de 6.544 matrículas, o que equivale a 12% do total de matrículas registadas nas BDs — 4ª e 5ª células da 5ª linha da Tabela 5.1.

Da subamostra de 6.544 matrículas a que se chegou nesta primeira seleção foram ainda excluídas mais de 2 mil matrículas, em resultado quer da necessária limpeza de dados, quer da necessidade de uma maior uniformização dos mesmos, destacando-se, entre essas operações de remoção, as seguintes:

1. Procedeu-se à limpeza de dados omissos, excluindo-se as matrículas e alunos, e todos os dados a si associados, que não tinham a sua informação académica completa. Foi o caso, por exemplo, das matrículas em cursos com um nº de ECTS inferior a 180 ou então com disciplinas com o nº de ECTS indefinido, por se considerar resultarem de informação errónea ou, no mínimo, incompleta.

2. Optou-se por excluir as matrículas, e dados associados, relativos a alunos que não tivessem sequer frequentado 6 semestres letivos, visando dessa forma uma maior uniformização dos resultados.

3. Removeram-se duas matrículas, e todos os dados a si associados, que estavam ainda alocadas à antiga designação da Escola de Mirandela com o código 3047 — cf. Tabela 5.2. Como nenhuma das outras designações desatualizadas das escolas, que aparecem na BD `sa`, tinham matrículas ou outros dados associados, optou-se por manter apenas as designações atuais das cinco escolas do IPB, tal como descritas (a negrito) na Tabela 5.2.

Tabela 5.2: Conjunto de nomes usados na BD para designação das escolas do IPB (a **negrito** as designações atuais)

cod_escola	abreviatura	escola
3040	IPB	Instituto Politécnico de Bragança
3041	ESAB	**Escola Superior Agrária de Bragança**
3042	ESEB	**Escola Superior de Educação de Bragança**
3043	ESTiG	**Escola Superior de Tecnologia e Gestão de Bragança**
3044	ESTGM	Escola Superior de Tecnologia e de Gestão — Polo de Mirandela
3045	EsACT	**Escola Superior de Comunicação, Administração e Turismo**
3047	ESTGM	Escola Superior de Tecnologia e de Gestão de Mirandela
7014	ESEnB	Escola Superior de Enfermagem de Bragança
7015	ESSB	**Escola Superior de Saúde de Bragança**

4. Removeram-se da tabela `disciplinas` da BD `sa` todas as disciplinas que não tinham notas lançadas no período considerado.

5. Visando uma maior uniformização dos dados e com o objetivo de garantir aos dados analisados um significativo nível de contribuição na formação do indicador de sucesso, optou-se por desconsiderar matrículas com uma elevada taxa de ECTS creditados de forma "automática" (i.e., sem envolver frequência e avaliação do aluno). Para o efeito, foram selecionadas apenas as matrículas (linhas) que não tivessem mais de $1/3$ dos seus ECTS assim creditados.

Ainda a montante de todo o processo de seleção e limpeza de dados descrito, foram também removidas do *dataset* original todas as matrículas em unidades curriculares específicas, mais precisamente, as assinaladas na BD original como matrículas "Avulso" ou como "Extra-curriculares e Complementares". Por esse motivo, os valores apresentados na Tabela 5.1 já só incluem, como pretendido, as matrículas em cursos — daí ter-se partido de um total de 55.000 matrículas, quando na BD original perfaziam as 57.576 (cf. Tabela 4.1). Para além dos dados de matrícula, também os alunos sem outras matrículas para além das "Extracurriculares e Complementares" ou "Avulso" foram igualmente excluídos da BD, assim como, todos os dados com eles relacionados.

Depois de realizadas todas as operações de remoção, o número de observações constantes na tabela `matriculas` baixou de 6544 para 4530, significando então que o *dataset* passou a ser composto por 4530 matrículas e toda a informação a elas associada.

5.4.3 Pré-processamento

Após a filtragem dos dados descrita na fase anterior, procedeu-se ao pré-processamento ainda em ambiente MySQL, com recurso ao editor MySQL Workbench[1], com o intuito de os preparar para posterior aplicação dos algoritmos de *data mining*. Efetuaram-se as seguintes transformações:

1. Uma vez que associadas a cada aluno podem existir várias matrículas (mudanças de curso, reingressos ou múltiplas licenciaturas na instituição), para facilitar a separação e o seguimento dos diferentes percursos curriculares de um mesmo aluno, considerou-se oportuno a criação dum atributo chave que viesse a permitir identificar de forma inequívoca cada uma das matrículas. Como nos reingressos, o par `aluno/curso` volta a ser o mesmo, foi necessário juntar à chave o atributo `tipo_de_ingresso`, de forma a diferenciar esse tipo de matrículas. Com esta opção, passou-se a considerar que os reingressos dão origem a novos percursos curriculares, com as UCs aprovadas na matrícula anterior a serem creditadas na matrícula de reingresso.

   Cada terno `aluno/curso/tipo_de_ingresso`, representando uma matrícula distinta (percurso curricular distinto), passou então a ser identificado por um código único. Sendo 38.013 o maior número mecanográfico encontrado no corpo de alunos e confirmando-se que nenhum dos alunos possui mais de 9 matrículas na instituição, de forma a permitir a bijetividade e a fácil reversibilidade da transformação em causa, optou-se por criar o código identificador de matrícula, designado `id_mat`, da seguinte forma: se o aluno tiver apenas uma matrícula na instituição, o identificador da matrícula será o próprio número do aluno; caso contrário, será dado pela expressão

   $$id\_mat = (num\_aluno + 100.000) * 10 + b, \quad (5.2)$$

   com o dígito b a ser usado para distinguir as diferentes matrículas que o aluno possa ter. Este novo atributo, que dará origem a uma nova coluna na tabela com o registo das matrículas, terá então o formato $1aaaaab$, em que os cinco dígitos $aaaaa$ representam o número mecanográfico do aluno e o dígito b um dos cursos desse aluno (mais concretamente, um dos pares `curso/tipo_de_ingresso`).

2. Acrescentou-se às tabelas `notas`, `lect_ini` (tabela com o registo dos dados de inscrição dos alunos nos diferentes anos letivos) e `stat_medias_conclusivas` (tabela com o registo das médias finais das matrículas concluídas com sucesso), da BD `sa`, uma nova coluna com o atributo id_mat, o qual terá a função de servir de chave estrangeira na identificação das matrículas associadas e, dessa forma, dispensar a utilização da chave composta `aluno/curso/tipo_de_ingresso`.

3. Modificaram-se alguns nomes, quer de atributos quer de tabelas, para normalização, diferenciação (não colisão de nomes) ou no sentido de providenciar uma descrição mais elucidativa do significado dos seus conteúdos.

4. Acrescentou-se, como novo atributo, a idade que o estudante tinha no ato da sua matrícula, calculada com base no atributo data de nascimento, presente na base de dados.

[1] Ferramenta de desenvolvimento integrado em ambiente gráfico, para bases de dados MySQL, disponibilizada em `https://www.mysql.com/products/workbench/`.

5. Considerou-se conveniente passar a identificar os concelhos e freguesias, quer de proveniência, quer de naturalidade do aluno, com códigos únicos dentro do contexto global da base de dados. Esta decisão justifica-se pelo facto de em cada um dos distritos, os concelhos e freguesias serem identificados pelos códigos 1, 2, $..n$, levando a que haja múltiplos concelhos e freguesias com o mesmo código, em virtude de os mesmos se repetirem quando se muda de distrito ou de concelho, respetivamente. Relativamente aos distritos esta questão não se colocou dado que têm sempre códigos diferentes entre eles. Assim, optou-se por incluir no código de concelho o prefixo 'código de distrito' e no código de freguesia o prefixo 'código de distrito + código de concelho'. Como nem o número de concelhos por distrito, nem o número de freguesias por concelho, chegam a atingir os 3 dígitos (menos 100 ocorrências), aos referidos atributos foram aplicadas as seguintes transformações inteiras:

$$cod\_concelho \rightarrow cod\_distrito \times 100 + cod\_concelho, \tag{5.3}$$

$$cod\_freguesia \rightarrow (cod\_distrito \times 100 + cod\_concelho) \times 100 + cod\_freguesia. \tag{5.4}$$

6. Na BD `sa` original uma parte significativa das notas aparece várias vezes, uma por cada plano de estudos que vigorou durante o período de tempo em que o aluno frequentou o curso. Tratando-se, já por si, de um conjunto de dados muito extenso, ao incluir múltiplas instâncias das mesmas notas (5 em alguns dos casos) a tabela com o registo das notas assume proporções consideráveis — refira-se a título de exemplo as $1.638.361$ instâncias que compõem a tabela original. Mas ainda mais crítico que a quantidade de dados que possa resultar da repetição das notas é, sem dúvida, o risco de se poder vir a considerar nos algoritmos as várias ocorrências de uma mesma nota como se notas distintas se tratassem.

   Na verdade, sempre que um novo plano de estudos de um curso entra em vigor, todas as notas já registadas dos alunos que se encontrem a frequentar esse curso são importadas para esse novo plano de estudos, passando a coexistir na mesma tabela as notas do plano mais recente com as que foram lançadas em planos anteriores. Assim, como o último plano de estudos (de um curso) que o aluno frequentou integra sempre todas as suas notas, o problema das repetições foi eficazmente resolvido deixando ficar na tabela de notas apenas as associadas ao último plano de estudos de cada matrícula. Todas as outras foram removidas.

7. Para cada matrícula (`id_mat`), determinou-se a quantidade de ECTS que resultaram de equivalências ou doutras formas de creditação, acrescentado-se a respetiva coluna na tabela com os dados de cada matrícula (`stat_medias_conclusivos`).

Idealmente, um estudo de previsão de desempenho académico dever-se-ia realizar no preciso momento em que o aluno se matricula no seu curso. Porém, se se puder protelar um pouco esse estudo, para um momento mais avançado do seu percurso escolar, talvez haja possibilidade de retirar significativos dividendos quanto ao grau de acerto da previsão que se pretenda efetuar. Este pressuposto advém do facto de se considerar que o nível de sucesso que o aluno possa apresentar em fases intermédias do seu percurso escolar, poderá vir a revelar-se um bom preditor do seu desempenho futuro. Foi por se acreditar na importância que as avaliações parcelares de desempenho possam ter quando usadas como variáveis preditivas, que se criou uma nova tabela MySQL (`estat_semestrais`), com as suas colunas definidas pelas variáveis descritas na Tabela 5.3, e que tem o propósito de agregar, em cada uma das suas linhas, os resultados semestrais dos alunos e, em particular, um indicador de sucesso semestral dos mesmos.

Para calcular os valores dos diferentes atributos da tabela foi desenvolvido um conjunto diverso de *queries* MySQL, tendo-se algumas delas revelado algo complexas. Segue-se uma descrição mais detalhada dos atributos da tabela cujo significado possa suscitar maiores dúvidas:

`id_mat` Código inteiro que identifica a matrícula (e o próprio aluno) associada ao percurso escolar objeto de análise.

`nucr` Número de unidades curriculares em que o aluno se inscreveu no semestre considerado, sem ter conseguido aprovação, quer por ter reprovado, faltado, desistido ou simplesmente por não ter sido admitido à avaliação.

`navalr` Número de avaliações a que o aluno efetivamente se submeteu no semestre, em que tenha reprovado ou desistido, contabilizando, nomeadamente, cada uma das múltiplas avaliações a que se tenha submetido numa mesma unidade curricular.

`ects_reprov` Soma dos ECTS das unidades curriculares reprovadas no semestre, ou seja, as que foram contabilizadas no atributo `nucr`.

`ects_aprov` Soma dos ECTS das unidades curriculares aprovadas no semestre, ou seja, as que foram contabilizadas no atributo `nuca`.

`média` Média aritmética da nota das unidades curriculares aprovadas no semestre.

`vd` Variável que, usando uma métrica equivalente à da variável dependente global (expressa na equação 5.1), quantifica o desempenho do aluno no semestre considerado.

Na verdade, ainda antes da construção da Tabela 5.3, com os resultados semestrais, houve a necessidade de dar um tratamento adequado às unidades curriculares anuais. Ainda que a tipologia normal da estrutura curricular das licenciaturas do IPB seja baseada em unidades curriculares semestrais, existem de facto cursos que incluem, ainda que em reduzido número (cerca de 10% das unidades curriculares), algumas unidades curriculares anuais, como é o caso das cadeiras de final de curso de projeto ou estágio. Por isso, e uma vez que o estudo é baseado em resultados semestrais, houve a necessidade de converter as unidades curriculares anuais em semestrais. Optou-se por dividir cada unidade curricular anual em duas semestrais — uma de cada semestre letivo —, ficando ambas com a mesma nota classificativa, mas com metade dos ECTS da unidade curricular anual.
Uma outra opção, alternativa à transformação de dados que se realizou, teria sido manter a integridade da unidade curricular anual, passando-a, por inteiro, para o segundo semestre letivo. Repare-se que isso resolveria por completo um ponto fraco importante da solução que desdobra a cadeira em duas: a impossibilidade de se considerar, para o cálculo das variáveis preditivas, a nota da unidade curricular do 1º semestre que resulta do desdobramento, a não ser *a posteriori* — será sempre necessário esperar pelo fim do ano letivo, para se saber a nota dessa cadeira do "1º semestre". Mas a deslocação de toda a cadeira para o 2º semestre também teria o grave inconveniente de provocar, artificialmente, o desbalanceamento do esforço do aluno entre os dois semestres letivos, podendo ter um efeito indesejado, por exemplo, num indicador de sucesso que viesse a contabilizar o nº de ECTS realizados por semestre.
Não sendo as cadeiras anuais em número significativo, e surgindo quase sempre apenas a partir do 5º semestre escolar do aluno (cadeiras de final de curso), entendeu-se não ter, a transformação de dados, uma tal influência nos resultados, que viesse a justificar o desenvolvimento de todo o processamento para este segundo tipo de transformação.
Na fase que se seguiu às tarefas de pré-processamento descritas, os dados foram preparados com o objetivo de os exportar para o R, ambiente onde foram analisados.

Tabela 5.3: Lista de variáveis com os resultados semestrais dos alunos.

atributo	significado
id_mat	código identificador da matrícula
ano	ano letivo
semestre	semestre letivo
nuca	nº de UCs aprovadas no semestre
ects_aprov	nº de ECTS aprovados no semestre
min	nota mínima das UCs aprovadas no semestre
max	nota máxima das UCs aprovadas no semestre
media	nota média das UCs aprovadas no semestre
nucr	nº de UCs reprovadas no semestre
navalr	nº de avaliações sem aprovação no semestre
ects_reprov	nº de ECTS reprovados no semestre
vd	medida de desempenho no semestre

Preparação dos dados para o R

Embora nesta fase da investigação o foco do estudo seja analisar até que ponto os resultados intercalares de desempenho do aluno poderão, eles próprios, ser usados como variáveis preditivas do seu sucesso escolar final, nada impede que, no modelo preditivo, se venham também a incluir, conjuntamente com os dados semestrais, também outras variáveis preditivas de âmbito "intemporal", relacionadas com características do aluno, como o seu ambiente familiar, social, escolar, etc. Pensa-se, mesmo, que a capacidade preditiva do modelo virá melhorada com a combinação destes dois grupos de atributos.

Visando a transferência dos dados para o R, procedeu-se à integração, numa mesma tabela, das variáveis semestrais da Tabela 5.3 com um conjunto de varáveis intemporais — cujos valores não se alteram ao longo do percurso escolar do aluno — de natureza académica, demográfica, sócio-económica e de acesso ao ensino superior. Obteve-se, dessa forma, uma nova tabela MySQL (designada `semestres`) com as variáveis descritas na Tabela 5.4, ou seja, com todas as potenciais variáveis preditivas a considerar no estudo e com um novo atributo, `sem_escolar_s`, identificativo do semestre escolar do aluno, para além da variável dependente (`vd`) que será usada nos modelos de previsão como indicador de sucesso.

Para um completo esclarecimento sobre o verdadeiro significado das variáveis que constituem a Tabela 5.4, convém, em primeiro lugar, enfatizar o significado de "semestre escolar". Como se sabe, no seu percurso escolar, o aluno frequenta um determinado número de semestres, não necessariamente igual aos previstos pela estrutura curricular do curso. Esses semestres frequentados pelo aluno são então os designados semestres escolares (do aluno), sendo cada um deles identificado pelo seu número de ordem no conjunto de todos os semestres frequentados pelo aluno, ordenados cronologicamente — por exemplo, um aluno pode ter estado no 2º semestre letivo de 2010/2011 a frequentar o 1º ano do plano curricular do curso, mas já no seu 4º semestre escolar. Semestres em que o aluno não se tenha inscrito em qualquer unidade curricular não foram considerados semestres escolares. Foram simplesmente ignorados.

Uma vez que se optou por agregar numa mesma tabela atributos de dados intemporais com atributos de dados semestrais, que se reportam apenas àquilo que se passou num dado semestre escolar do aluno, achou-se conveniente, para uma melhor identificação das duas famílias de atributos, acrescentar o sufixo "_s" ao nome de todas as variáveis que apenas reportam dados ou resultados do semestre escolar considerado. Todas os atributos não sufixados reportam, por isso, informação considerada intemporal no contexto do atual estudo. Na família de atributos intemporais, poder-se-á ainda considerar dois subgrupos importantes: os dados que são conhe-

Tabela 5.4: Lista de variáveis exportadas para o R.

id	atributo	cat	tipo	min..max	significado
0	**sem_escolar_s**[a]		discreto	1..6	*semestre escolar do aluno em análise*
1	**ano_curricular_s**	C	discreto	1..4	*ano curricular do aluno no sem. escolar considerado*
2	**ano_s**	C	discreto	07..15	*ano letivo do semestre escolar considerado*
3	**bolseiro_s**	C	contínuo[e]	0..1	*o aluno foi bolseiro no semestre escolar?*
4	**cod_estatuto_s**[b]	C	nominal	1..5	*tipo de estatuto do aluno no semestre escolar*
5	**cod_freq_tipo_s**[b]	C	nominal	1..7	*tipo de frequência do aluno no semestre escolar*
6	**dir_associativo_s**	C	contínuo[e]	0..1	*o aluno foi dirigente associativo no sem. escolar?*
7	**ects_aprov_s**	C	discreto	0..60	*nº de ECTS aprovados no semestre escolar*
8	**ects_reprov_s**	C	discreto	0..60	*nº de ECTS reprovados no semestre escolar*
9	**max_s**	C	discreto	0..20	*nota máxima das UCs aprovadas no semestre escolar*
10	**media_s**	C	contínuo	0..20	*nota média das UCs aprovadas no semestre escolar*
11	**min_s**	C	discreto	0..20	*nota mínima das UCs aprovadas no semestre escolar*
12	**navalr_s**	C	discreto	0..18	*nº de avaliações sem aprovação no semestre escolar*
13	**nuca_s**	C	discreto	0..10	*nº de UCs aprovadas no semestre escolar*
14	**nucr_s**	C	discreto	0..10	*nº de UCs reprovadas no semestre escolar*
15	**semestre_s**[b]	C	discreto	1..2	*semestre letivo do semestre escolar considerado*
16	**vd12_s**[c]	C	contínuo	-20..20	*diferença de desempenho do 1º para o 2º semestre*
17	**vd23_s**[c]	C	contínuo	-20..20	*diferença de desempenho do 2º para o 3º semestre*
18	**vd34_s**[c]	C	contínuo	-20..20	*diferença de desempenho do 3º para o 4º semestre*
19	**vd45_s**[c]	C	contínuo	-20..20	*diferença de desempenho do 4º para o 5º semestre*
20	**vd56_s**[c]	C	contínuo	-20..20	*diferença de desempenho do 5º para o 6º semestre*
21	**ano_mat**	M	discreto	07..13	*ano da matrícula*
22	**cod_curso**	M	nominal	1..51	*código do curso*
23	**cod_escola**	M	nominal	1..5	*código da escola*
24	**ects_cred_tx**	M	discreto	0..100	*fração de ECTS que foram creditados ao aluno*
25	**ects_curso**	M	discreto	180..240	*número de ECTS do curso*
26	**tipo_ing**	M	nominal	1..9	*tipo de ingresso*
27	**conc**[d]	D	nominal	1..215	*concelho de proveniência do aluno*
28	**conc_n**[d]	D	nominal	1..209	*concelho de naturalidade*
29	**deslocado**	D	lógico	0..1	*o aluno está deslocado da sua residência habitual?*
30	**dist**	D	nominal	1..28	*distrito de proveniência do aluno*
31	**dist_n**	D	nominal	1..27	*distrito de naturalidade*
32	**freg**[d]	D	nominal	1..1542	*freguesia de proveniência*
33	**freg_n**[d]	D	nominal	1..1468	*freguesia de naturalidade*
34	**idade**	D	discreto	17..61	*idade no ato da matrícula*
35	**nacionalidade**	D	nominal	1..15	*nacionalidade do aluno*
36	**sexo**	D	nominal	1..2	*género*
37	**cod_prof_aluno**	S	nominal	1..12	*profissão do aluno*
38	**cod_prof_mae**	S	nominal	1..12	*profissão da mãe*
39	**cod_prof_pai**	S	nominal	1..12	*profissão do pai*
40	**nivel_esc_mae**	S	ordinal	1..13	*nível de escolaridade da mãe*
41	**nivel_esc_pai**	S	ordinal	1..13	*nível de escolaridade do pai*
42	**sit_prof_aluno**	S	nominal	1..10	*situação profissional do aluno*
43	**sit_prof_mae**	S	nominal	1..10	*situação profissional da mãe*
44	**sit_prof_pai**	S	nominal	1..9	*situação profissional do pai*
45	**fase**	A	ordinal	1..3	*fase de acesso*
46	**media_acesso**	A	contínuo	0..200	*nota de acesso ao ensino superior*
47	**n10_11_acesso**	A	contínuo	0..200	*média dos 10º e 11º anos*
48	**n12_acesso**	A	contínuo	0..200	*média do 12º ano*
49	**opcao_acesso**	A	ordinal	1..6	*ordem da opção na candidatura ao curso*
50	**ordem_acesso**	A	discreto	1..322	*ordem de acesso entre os colocados no curso*
51	**pi_acesso**	A	contínuo	0..200	*nota média das provas de ingresso*
52	**vd**[a]		contínuo	0..20	*var. dependente com o desempenho final do aluno*

[a] Variável não preditiva.
[b] Variável não existente nos modelos de dados com resultados semestrais acumulados.
[c] Cada variável $\texttt{vd}ij\texttt{\_s}$ apenas está presente nos modelos de dados com resultados acumulados de j ou mais semestres.
[d] Variável categórica que se veio a desconsiderar no estudo por assumir mais de 52 valores distintos, o número de níveis máximo permitido pelo algoritmo *random forest* usado.
[e] Ainda que tipicamente de natureza binária, valor considerado contínuo de forma a poder traduzir o rácio de ocorrências no conjunto dos semestres agregados.

cidos logo no ato da matrícula e que, por isso, poderão ser usados como variáveis preditivas, e aqueles que, sendo apenas conhecidos na conclusão da licenciatura, poderão vir a ser usados na construção de indicadores de sucesso.
Na Tabela 5.4 é então possível identificar variáveis preditivas semestrais (sufixadas com "_s"), variáveis preditivas intemporais e variáveis conclusivas (apenas a última da tabela) — ainda que não incluídas na tabela, há a considerar mais 3 importantes variáveis conclusivas, precisamente as variáveis `media`, `ects_aprov` e `ects_reprov`, usadas na construção da variável dependente (cf. equação 5.1). Por outro lado, no âmbito do estudo que se pretende realizar, houve o cuidado de categorizar (3ª coluna da tabela) as variáveis preditivas, de acordo com a sua tipologia, em 5 classes distintas: C — Curriculares; M — de Matrícula; D — Demográficas; S — Socioeconómicas; A — de Acesso.
Note-se que, no contexto deste estudo, as variáveis com dados curriculares (C) confundem-se com as variáveis semestrais (sufixadas com "_s"). Foi pela importância que este subconjunto de varáveis representa no modelo preditivo que se optou por subdividir os dados académicos em dois subgrupos: os curriculares (C), de valores variáveis e acumuláveis ao logo percurso escolar, e os dados de matrícula (M), que são logo estabelecidos no momento de ingresso. Passa-se assim a dispor de 5 subgrupos de variáveis preditivas, facilmente referenciáveis através das letras [C, M, D, S, A].
Segue-se uma descrição mais detalhada dos atributos da tabela cujo significado possa suscitar maiores dúvidas:

`sem_escolar_s` Número de ordem identificativo do semestre escolar analisado e a que se referem os dados das variáveis sufixadas por "_s"; É este atributo que, conjuntamente com o `id_mat` (chave composta), permite identificar/distinguir as diferentes instâncias (linhas) da tabela `semestres` — após a completa definição da tabela, retirou-se-lhe este segundo atributo, por ser dispensável em todo o processo subsequente de previsão de sucesso.

`nucr_s` Número de unidades curriculares em que o aluno se inscreveu no semestre escolar considerado, sem ter conseguido aprovação, quer por ter reprovado, faltado, desistido ou simplesmente por não ter sido admitido à avaliação.

`navalr_s` Número de avaliações a que o aluno efetivamente se submeteu no semestre escolar, em que tenha reprovado ou desistido, contabilizando, nomeadamente, cada uma das múltiplas avaliações a que se tenha submetido numa mesma unidade curricular.

`ects_reprov_s` Soma dos ECTS das unidades curriculares reprovadas no semestre escolar, ou seja, as que foram contabilizadas no atributo `nucr_s`.

`ects_cred_tx` Fração de ECTS do curso (em pontos percentuais) que foram creditados sem que o aluno se tenha submetido a qualquer avaliação (ECTS de unidades curriculares aprovadas por equivalência ou com recurso a outros tipos de creditações). Optou-se pelo rácio em vez da contagem simples dos ECTS, uma vez que nem todos os planos de estudos das licenciaturas são de 180 ECTS — 5 das 51 licenciaturas analisadas, todas da Escola Superior de Saúde de Bragança, prevêm planos estudos de 240 ECTS, correspondentes a 4 anos de formação.

`vd` Variável dependente usada como indicador de sucesso, que utiliza a métrica expressa na equação 5.1 para quantificar o desempenho global do aluno. A variável `ects_reprov` usada na equação corresponde à soma dos ECTS das unidades curriculares em que o aluno se inscreveu ao longo de todo o seu percurso escolar no curso, sem ter conseguido aprovação,

quer por ter reprovado, faltado, desistido ou simplesmente por não ter sido admitido à avaliação (i. e., somatório de todos os `ects_reprov_s`) — repare-se que podem muito bem estar a ser contabilizados ECTS de unidades curriculares às quais o aluno tenha posteriormente obtido aprovação.

A Tabela 5.4, que contém as variáveis a exportar para o R, naturalmente, integra dados de tabelas das três BDs analisadas. De todas, a BD `inqueritos` é claramente, e como se compreende, aquela que tem os seus dados mais incompletos.
Sabe-se também que uma parte ainda significativa dos alunos (cerca de $1/3$ dos que concluíram no período de análise) ingressa nas licenciaturas do IPB através de concursos especiais de acesso. Logo, os atributos provenientes da BD `acesso` relacionados com esses alunos vão estar omissos na tabela que se pretende exportar para o R, uma vez que essa BD coleciona apenas dados referentes ao Concurso Nacional de Acesso.
Após o processamento que se acabou de descrever, dispõe-se já, em formato adequado e na forma integrada, de todos os dados necessários para ser poderem vir a usar, como variáveis preditivas, resultados curriculares que o aluno tenha obtido num qualquer dos seus primeiros 6 semestres. Todavia, será expectável que a capacidade preditiva do modelo aumente ainda significativamente, se, em vez de resultados semestrais, se venham a incluir, no modelo preditivo, variáveis que agreguem os resultados acumulados até cada um desses semestres. Tendo em vista esse objetivo, preparou-se mais um conjunto de tabelas para serem exportadas para o R, contendo, cada uma delas, essencialmente as mesmas variáveis da Tabela 5.4, mas onde os valores do subconjunto de varáveis preditivas semestrais (variáveis da categoria C) passaram a reportar resultados, não de um semestre, mas sim resultados acumulados de vários semestres. Assim, para além da tabela `semestres`, com os resultados semestrais e composta por 27.180 linhas, será também exportado para o R, ambiente onde serão aplicadas técnicas de *data mining*, um conjunto de 5 tabelas, cada uma com 4530 linhas, que, para além das variáveis preditivas intemporais (categorias M, D, S, e A) e da variável conclusiva (`vd`), presentes na Tabela 5.4, inclui um conjunto próprio de variáveis curriculares (C), em concordância com a descrição que se segue:

`semestres12` – tabela que usa para as varáveis curriculares os resultados acumulados do 1º e 2º semestres escolares;

`semestres123` – tabela que usa para as varáveis curriculares os resultados acumulados do 1º, 2º e 3º semestres escolares;

`semestres1234` – tabela que usa para as varáveis curriculares os resultados acumulados do 1º, 2º, 3º e 4º semestres escolares;

`semestres12345` – tabela que usa para as varáveis curriculares os resultados acumulados do 1º, 2º, 3º, 4º e 5º semestres escolares;

`semestres123456` – tabela que usa para as varáveis curriculares os resultados acumulados do 1º, 2º, 3º, 4º, 5º e 6º semestres escolares;

Nestas cinco tabelas excluiu-se a coluna da variável `semestre_s`, por não fazer sentido falar-se em semestre letivo, dado tratar-se de um conjunto agregado de vários semestres escolares, bem como a coluna da variável `sem_ecolar_s`, uma vez que os dados acumulados de cada uma das tabelas reportar ao mesmo conjunto de semestres escolares. Excluíram-se, de igual modo, as variáveis `cod_freq_tipo_s` e `cod_estatuto_s`, por se tratarem de códigos inteiros não agregáveis. Em contrapartida, essas novas tabelas passaram a incluir variáveis da família `vdij_s`, que

traduzem a evolução do desempenho do aluno entre semestres. Segue-se, por fim, uma descrição do significado, ou da forma, como foram obtidos os valores das variáveis com resultados agregados menos triviais:

`ano_s` — ano letivo do último semestre agregado;

`ano_curricular_s` — ano curricular frequentado pelo aluno no último semestre agregado;

`bolseiro_s` — percentagem dos semestres agregados em que o aluno teve bolsa;

`dir_associativo_s` — percentagem dos semestres agregados em que o aluno teve estatuto de dirigente associativo;

`vdij_s` — diferença de desempenho do aluno do *i*-ésimo para o *j*-ésimo semestre escolar. Em rigor, $vd_{ij\_s} = vd_j - vd_i$, com $j = i + 1$ e $i = 1..5$, em que vd_n é o desempenho do aluno no seu enésimo semestre, calculado com recurso a uma métrica equivalente à da variável dependente global `vd` (c.f. equação 5.1). A inclusão desta família de variáveis foi motivada pelo propósito de se fornecer ao modelo preditivo informação acerca da evolução do aluno, ao longo dos semestres agregados em cada um dos estudos. Assim, cada variável `vdij_s` apenas estará presente nos modelos de dados com resultados acumulados de *j* ou mais semestres.

Importação para o R

Em contexto de ambiente RStudio, começou-se por importar do MySQL Server as 6 tabelas com os resultados do pré-processamento já realizado — embora essas tabelas, dentro do R, tomem a forma de *data frames*[2], por facilidade de linguagem e compreensão, continuar-se-ão a designar, nesta tese, simplesmente por tabelas. Após a sua importação, tendo em vista a aplicação dos algoritmos de *data mining*, foi necessário converter para tipo categórico, nominal ou ordinal, os atributos que se considera terem essa tipologia, em concordância com a classificação assumida na 4 coluna da Tabela 5.4, dado que todos esses atributos provêm do MySQL Server ou como *strings*[3] ou em formato numérico. Seguidamente, extraíram-se para tabelas separadas todas as observações (linhas) da tabela `semestres` respeitantes a cada um dos 6 primeiros semestres escolares, resultando, dessa forma, 6 novas tabelas de 4530 linhas cada: `semestre1`, `semestre2`, …, `semestre6`.

O modelo de dados é agora composto por 11 tabelas (`semestre1`, `semestre2`, …, `semestre6`, `semestres12`, `semestres123`, …, `semestres123456`), que contêm, cada uma delas, todas as variáveis preditivas que serão objeto de estudo, as quais, como anteriormente explicado, pode-se assumir que se enquadram nas seguintes 5 categorias: curriculares (C), de matrícula (M), demográficas (D), socioeconómicos (S) e de acesso ao ensino superior (A). Enquanto que os três primeiros subgrupos de variáveis provêm essencialmente da BD `sa` (a exceção será a variável `deslocado`, que provém da BD `inqueritos`), as socioeconómicas e as de acesso tiveram origem nas BDs `inqueritos` e `acesso`, respetivamente.

Em contexto de seleção e caracterização dos dados, constatou-se que nem todos os alunos registados na BD `sa` possuem dados de acesso e nem todos responderam ao inquérito de recolha

[2] Um *data frame* é uma importante estruturas de dados da linguagem R (classe `data.frame`), usada para representar tabelas de dados. À semelhança de uma qualquer tabela MySQL, é formada por uma lista de variáveis, todas com o mesmo número de observações, e onde cada coluna representa uma variável e cada linha uma instância das variáveis.

[3] `Character` em R e `varchar` em MySQL.

de dados socioeconómicos. Por isso, se se pretender que o modelo de previsão envolva o uso de todas as variáveis preditivas disponíveis, então o conjunto de matrículas terá que ser ainda mais selecionado, excluindo-lhe todas aquelas que contenham dados incompletos. Dessa forma, o tamanho da amostra reduzir-se-á a menos de metade, passando de 4.530 para 2.159 matrículas. No caso de só se pretender usar, para além dos dados académicos (C+M) e demográficos (D), os dados de acesso ao ensino superior (A), o tamanho da amostra passa para 3.016. Mas se em vez dos dados de acesso, forem usados, para variáveis preditivas, os dados socioeconómicos (S), então a amostra já será formada por 3.109 matrículas, dado que as matrículas a excluir são agora as que não possuam dados dos inquéritos.

5.5 Aplicação do algoritmo de *data mining random forest*

5.5.1 O algoritmo *random forest*

Conforme discutido no Capítulo 2, existe uma grande variedade de algoritmos *data mining* desenvolvidos para objetivos de previsão. Neste estudo optou-se por basear o modelo preditivo no algoritmo *random forest*, proposto por Breiman [14]. De acordo com a revisão de literatura desenvolvida por Kumar et al. [58], trata-se do método mais popular entre os investigadores que visavam a previsão de desempenho académico através de métodos de regressão. Também diversos estudos de investigação demonstraram que, em tarefas de previsão de sucesso final dos estudantes, as *random forest* superam claramente outras técnicas, em termos de eficácia preditiva (ver, por exemplo, na Secção 3.3, a revisão aos estudos Delen [31], Pascoal et al. [85], Miguéis et al. [71], Amrieh et al. [4]). Mas mais importante ainda, permite uma boa interpretação dos resultados que produz, em contraste com outras técnicas, como é o caso das Redes Neuronais e Máquinas de Vetores de Suporte, que são consideradas autênticas caixas pretas. De facto, apresenta uma interessante funcionalidade, que é a capacidade de quantificar o grau de importância dos atributos preditivos em análise, na explicação da previsão pretendida. Concretizando para o caso do presente trabalho, permite ordenar a importância de cada dimensão do aluno na determinação do seu sucesso académico (ver, por exemplo, figuras 5.5 e 5.6). A partir desta informação será possível identificar, por ordem de importância, fatores associados ao sucesso educacional dos estudantes de licenciatura do IPB, com vista a identificar os estudantes em risco que possam beneficiar de alguma intervenção ou apoio pedagógico. Essa mesma funcionalidade possibilita que se excluam do modelo atributos considerados pouco relevantes.
A ordem de importância estabelecida sobre a relevância de cada característica na determinação do sucesso académico de um aluno, pode ser usada por decisores académicos para identificar fatores importantes para prever o sucesso dos alunos e por exemplo para averiguar a necessidade de ações de tutoria em função do perfil dos alunos.

5.5.2 Estudo comparativo de diferentes grupos de preditores usando resultados semestrais acumulados

Na análise exploratória de dados que se pretendeu desenvolver, optou-se por manter fixa a configuração do algoritmo `randomForest`, parametrizado como se descreveu na secção anterior, para o foco do estudo incidir no conjunto de variáveis preditivas que lhe darão suporte.
Na Figura 5.1 apresenta-se um esquema ilustrativo que pretende caraterizar o modelo de previsão arquitetado no âmbito do atual estudo. Nele 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 um dos 6 primeiros semestres escolares do aluno — ainda que não totalmente percetível no esquema, apenas uma das 6 entradas de dados curriculares (C) deve ser considerada em cada execução do algoritmo (entradas mutuamente exclusivas).

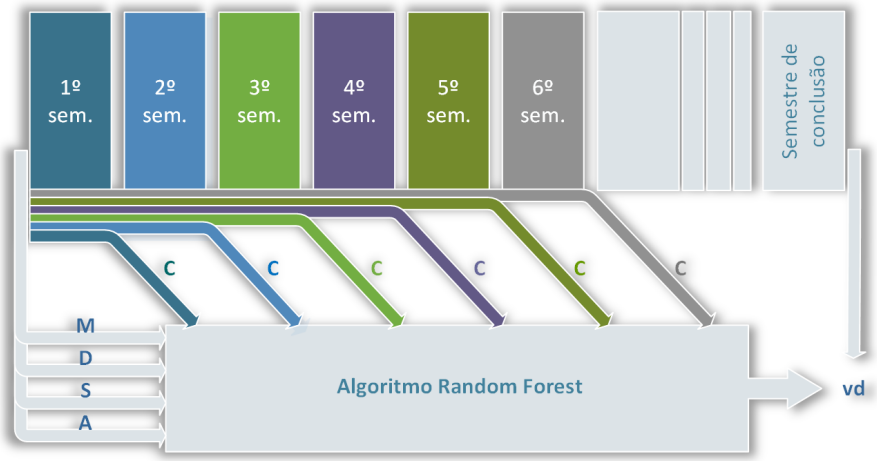


Figura 5.1: Esquema ilustrativo do estudo comparativo realizado.

Como se sabe, o desempenho global 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 perceber o impacto dessas dimensões na capacidade do modelo. Na Tabela 5.5 mostram-se as diferentes categorias de atributos escolhidas em cada um desses estudos. Ainda que se pretenda desenvolver um estudo suficientemente exaustivo, entende-se não ser necessário abranger todas as combinações possíveis dos 5 grupos de variáveis, que perfaz um total de 31 possibilidades,

$$N_{max} = \sum_{n=1}^{5} {}^{5}C_n = 31. \quad (5.5)$$

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, ou seja, o estudo onde se usam todos os grupos, à exceção do grupo C. Com esta simplificação, o número de estudos reduz-se a quase metade, uma vez que, estando esse grupo de variáveis (quase) sempre presente, o número de estudos passa a ser dado essencialmente pelas combinações entre os restantes 4 grupos. Em rigor, contabilizando também a possibilidade de não escolha de nenhum desses 4 grupos (estudo C), serão 17, os estudos a concretizar,

$$N = \sum_{n=0}^{4} {}^{4}C_n + 1 = 17. \quad (5.6)$$

Na Tabela 5.6 são mostrados, para cada um dos estudos, os coeficientes de determinação (R^2)

Tabela 5.5: Combinações de grupos de variáveis preditivas usadas nos estudos realizados.

estudo	mnem.	**C**urriculares	de **M**atrícula	**D**emográficos	**S**ocioeconómicos	de **A**cesso
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**	✓				

que se obtiveram com a aplicação do algoritmo de previsão *random forest* aos dados selecionados, ao fim de cada um dos 6 primeiros semestres escolares do aluno. Para elucidar melhor em que consistiram os diferentes estudos e o verdadeiro significado dos resultados tabelados, tome-se, como exemplo, o Estudo 8: o *input* do algoritmo *random forest* (ver Figura 5.1) 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. É a média ponderada dos coeficientes de determinação desses 6 estudos que se apresenta na última coluna da tabela. 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 uma das métricas 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 que para muitos dos alunos, esse é o semestre em que concluem a sua formação, tornando completamente irrelevante a capacidade preditiva que o modelo possa evidenciar nesse momento.
Observando os valores tabelados, listados por ordem decrescente do valor médio do R^2 (última coluna), algumas considerações podem, desde já, ser avançadas:

- Não foi o modelo que se "alimenta" da totalidade das variáveis (Estudo 1 – CMDSA) que apresentou melhores capacidades preditivas. Na verdade, 6 outros modelos conseguiram iguais ou melhores desempenhos, com um menor número de variáveis preditivas.
- No gráfico da Figura 5.2 é notória a grande diferença de desempenho entre os 8 modelos mais precisos e os restantes — repare-se na proximidade dos valores médios dos coeficientes de determinação (R^2) nos primeiros 8 estudos e na sua repentina diminuição entre o Estudo 3 e o 12 —. É percetível o que diferencia os 8 estudos mais precisos dos restantes: precisamente, a inclusão de todos os dados académicos, os quais, como previamente apresentado, integram os dados curriculares e os dados de matrícula, ou seja, os dois subgrupos C e M. Se a quebra repentina do R^2 entre os estudos 3 e 12 se deve claramente à perda

Tabela 5.6: Coeficiente de determinação R^2 do modelo de previsão, para diferentes grupos de variáveis preditivas, e em função do semestre escolar do aluno.

	mnem.	1º sem	2º sem	3º sem	4º sem	5º sem	6º sem	Média[a]
Estudo 13	**CM**	80.4	86.5	92.0	94.3	96.6	97.9	88.4
Estudo 8	**CMS**	80.7	86.8	91.7	93.9	96.4	97.7	88.4
Estudo 4	**CMSA**	80.7	86.5	91.6	93.9	96.3	97.7	88.3
Estudo 1	**CMDSA**	80.3	86.3	91.3	93.7	96.2	97.6	88.1
Estudo 2	**CMDS**	80.3	86.4	91.4	93.7	96.3	97.7	88.1
Estudo 7	**CMD**	79.8	86.3	91.6	94.0	96.5	97.8	88.1
Estudo 9	**CMA**	79.9	86.2	91.6	94.0	96.4	97.7	88.1
Estudo 3	**CMDA**	79.7	86.1	91.4	93.8	96.3	97.7	87.9
Estudo 12	**CSA**	70.5	78.8	86.6	89.9	94.2	96.6	81.8
Estudo 15	**CS**	70.6	78.7	86.5	89.7	94.1	96.5	81.8
Estudo 5	**CDSA**	70.5	78.5	86.4	89.8	94.2	96.5	81.7
Estudo 10	**CDS**	70.7	78.3	86.2	89.6	94.1	96.4	81.6
Estudo 17	**C**	70.3	78.0	86.6	90.0	94.4	96.7	81.6
Estudo 16	**CA**	69.6	78.2	86.6	90.1	94.4	96.7	81.5
Estudo 11	**CDA**	69.8	78.2	86.3	89.9	94.3	96.6	81.4
Estudo 14	**CD**	70.1	77.8	86.3	89.6	94.2	96.6	81.4
Estudo 6	**MDSA**	64.4	64.4	64.4	64.4	64.4	64.4	64.4
Média[b]		75.2	82.4	89.0	91.9	95.3	97.2	84.9

[a] Média ponderada dos valores semestrais, com pesos 6, 5, ..., 2, 1, para os semestres 1º, 2º, ..., 5º, 6º, respetivamente.
[b] Valor médio sem contabilizar o Estudo 6.

do subgrupo M dos dados académicos, a que sobressai entre os estudos 14 e 6, também perfeitamente percetível no gráfico da Figura 5.2, fica a dever-se à perda do outro subgrupo dos dados académicos, os dados de natureza curricular C. Este segundo decaimento também evidencia de forma clara a grande relevância do subgrupo C em relação ao M, uma vez que o primeiro dos subgrupos dos dados académicos é substituído pelo segundo quando se passa do estudo 14 para o 6.

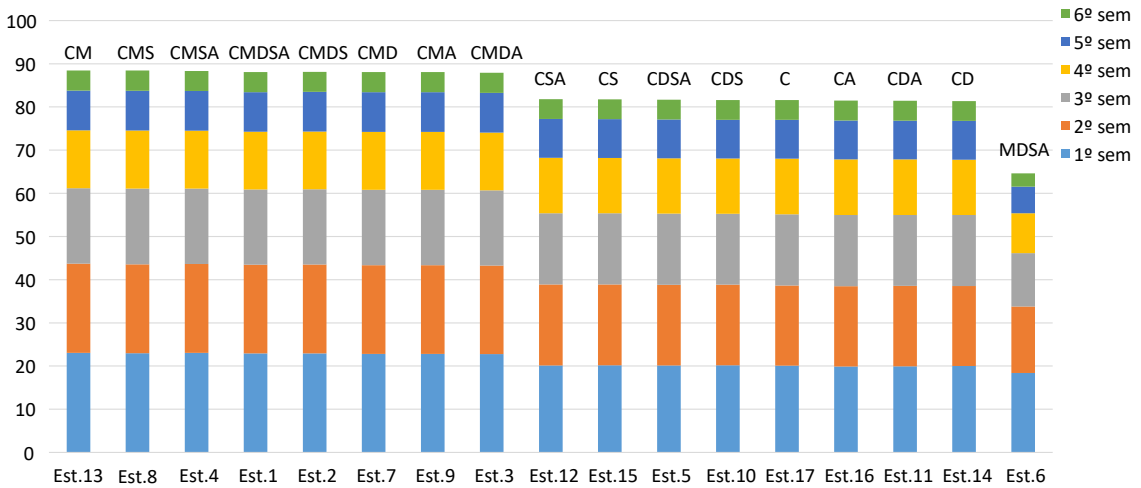


Figura 5.2: Coeficientes de determinação médios para os diferentes grupos de variáveis preditivas.

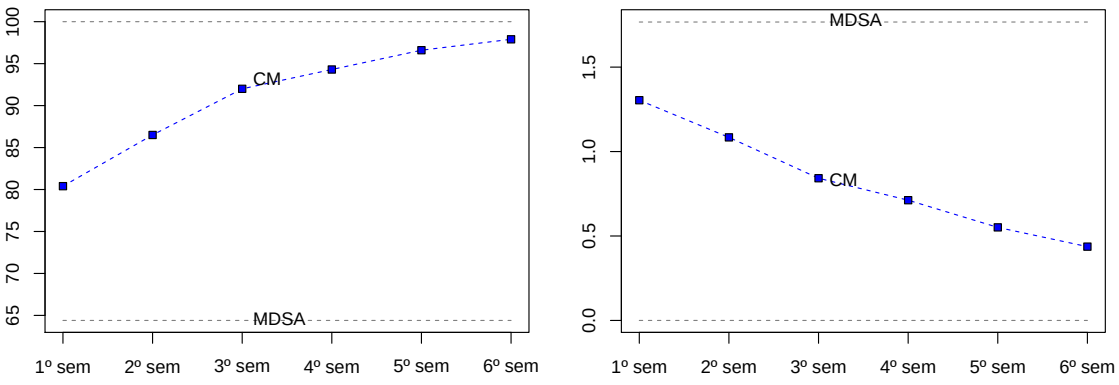
- Os dados académicos (subgrupos CM, Estudo 13), só por si, justificam o melhor resultado dos estudos (R^2=88.4%), obtido em *ex-aequo* com o Estudo 8 (CMS).
- Se se excluir o Estudo 6, o único que não inclui dados curriculares, e agruparem-se os

restantes estudos em 8 pares, onde a única diferença que distingue os 2 estudos de cada par é a inclusão, ou não, dos atributos demográficos D — (8-CMS,2-CMDS), (13-CM,7-CMD), etc. — é possível percecionar a pertinência desse subgrupo de dados. Como, dentro de cada par, os modelos apresentam praticamente o mesmo R^2, sendo, inclusive, tendencialmente maior nos que excluem o grupo D, fica por demais evidente a incapacidade preditiva deste conjunto de dados. Portanto, fica-se com a clara convicção de que o grupo de variáveis demográficas considerado não representa uma contribuição significativa para a capacidade preditiva do modelo.

- Conclusão idêntica pode ser retirada para os dados de acesso ao Ensino Superior (grupo A). Se se reparar nos pares dos estudos que se distinguem apenas pela inclusão ou não do grupo A, rapidamente se percebe que a inclusão deste grupo de atributos, em vez de contribuir positivamente, penaliza mesmo o desempenho do modelo. Apenas o par (5-CDSA,10-CDS) aparenta ser a exceção, mas note-se que a diferença de desempenho médio dos dois modelos acaba por ser negligenciável, sendo mesmo o modelo sem dados de acesso o mais preciso no 1º dos semestres.

- Ocorre exatamente o contrário com os dados socioeconómicos (grupo S). Se se tiverem em conta os pares dos estudos que se distinguem apenas pela inclusão ou não do grupo S, percebe-se que este grupo de atributos contribui efetivamente para aumentar o desempenho do modelo. Apenas no melhor par do teste (13-CM,8-CMS) a sua influência aparenta ser negligenciável.

- Como esperado, a capacidade preditiva do modelo aumenta consistentemente com o avanço do percurso escolar do aluno, como é bem notório nos gráficos da Figura 5.3, que mostram o desempenho do modelo suportado unicamente por dados académicos (CM).



(a) Coeficiente Determinação (%)

(b) Erro quadrático médio residual

Figura 5.3: Desempenho do modelo preditivo CM nos 6 primeiros semestres do aluno.

- Os resultados do Estudo 6, 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 dos modelos mais precisos dispara de 64.4% para valores superiores a 80% e 86%, respetivamente.

Na Tabela 5.7, por sua vez, são mostrados, para os primeiros 6 semestres escolares do aluno, os erros quadráticos médios residuais (RMSE) cometidos pelos 17 modelos do estudo. Para uma mais fácil confrontação de resultados, os estudos são apresentados por ordem crescente do valor médio do RMSE (última coluna). Pelos valores tabelados, percebe-se que os dois estudos que

Tabela 5.7: Erro quadrático médio residual do modelo de previsão, para diferentes grupos de variáveis preditivas, e em função do semestre escolar do aluno.

	mnem.	1º sem	2º sem	3º sem	4º sem	5º sem	6º sem	Média[a]
Estudo 13	**CM**	1.304	1.084	0.842	0.712	0.551	0.437	0.966
Estudo 9	**CMA**	1.324	1.099	0.861	0.728	0.565	0.450	0.983
Estudo 8	**CMS**	1.313	1.094	0.873	0.750	0.581	0.462	0.986
Estudo 4	**CMSA**	1.313	1.100	0.875	0.748	0.581	0.460	0.988
Estudo 7	**CMD**	1.335	1.106	0.873	0.743	0.572	0.456	0.993
Estudo 3	**CMDA**	1.340	1.113	0.881	0.749	0.580	0.462	1.000
Estudo 1	**CMDSA**	1.335	1.116	0.893	0.765	0.594	0.471	1.006
Estudo 2	**CMDS**	1.339	1.121	0.895	0.766	0.592	0.475	1.008
Estudo 17	**C**	1.596	1.372	1.070	0.926	0.694	0.528	1.210
Estudo 12	**CSA**	1.598	1.356	1.077	0.934	0.709	0.544	1.211
Estudo 15	**CS**	1.595	1.356	1.081	0.944	0.720	0.553	1.214
Estudo 16	**CA**	1.616	1.367	1.071	0.923	0.696	0.533	1.215
Estudo 11	**CDA**	1.615	1.371	1.087	0.936	0.703	0.544	1.221
Estudo 14	**CD**	1.604	1.382	1.086	0.945	0.708	0.542	1.222
Estudo 5	**CDSA**	1.606	1.370	1.089	0.943	0.714	0.554	1.222
Estudo 10	**CDS**	1.603	1.376	1.099	0.954	0.726	0.563	1.227
Estudo 6	**MDSA**	1.769	1.769	1.769	1.769	1.769	1.769	1.769
Média[b]		1.465	1.236	0.978	0.842	0.643	0.502	1.105

[a] Média ponderada dos valores semestrais, com pesos 6, 5, ..., 2, 1, para os semestres 1º, 2º, ..., 5º, 6º, respetivamente.
[b] Valor médio sem contabilizar o Estudo 6.

apresentaram, em *ex-aequo*, o melhor coeficiente de determinação R^2, surgem agora como 1º e 3º estudos com erros residuais mais baixos, respetivamente RMSE=0.966 para o modelo CM e RMSE=0.986 para o CMS. Com estes novos resultados, e no seguimento das considerações anteriores, considera-se então adequado propor para o modelo preditivo que se pretende arquitetar os grupos de variáveis do Estudo 13 (CM). Esta opção justifica-se pelo facto de serem aquelas que produzem os mais elevados coeficientes de determinação e, simultaneamente, os menores erros residuais, o qual usa, como se sabe, apenas 2 categorias de variáveis, das 5 possíveis.

Ainda que se pretendesse, na escolha do melhor modelo, dar especial importância às qualidades preditivas que os mesmos possam apresentar nas fases mais precoces do percurso escolar do aluno, o confronto das duas métricas conduziria a resultados inconclusivos: sendo o Estudo 8 (CMS) a destacar-se dos restantes nos dois primeiros semestres, no que concerne aos coeficientes de determinação associados, é o Estudo 13 (CM) que apresenta menores erros residuais nesses mesmos semestres — RMSE=1.304 no 1º semestre e RMSE=1.084 no 2º, quando o grupo CMS fica pelos valores 1.313 e 1.094, respetivamente.

Com estes 17 estudos conseguiram-se já excluir três "importantes"[4] subgrupos de variáveis: os socioeconómicos, os demográficos e os dados de acesso. No passo seguinte tentar-se-á, dentro dos dois grupos que se mantiveram, excluir as variáveis que apresentem uma influência negligenciável na capacidade do modelo preditivo.

[4] Em rigor, foi exatamente por não serem importantes que esses subgrupos de variáveis foram excluídos.

5.5.3 Ajuste do modelo preditivo suportado unicamente por dados académicos (modelo CM Ajustado)

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: o rigor das suas previsões e o grau de antecipação com que as consegue obter. Por conseguinte, nesta fase do trabalho, procura-se ajustar e estudar, com maior profundidade, o modelo do estudo que mostrou ser globalmente mais eficaz, aplicado logo ao fim do 1º semestre escolar do aluno, modelo que se ilustra no esquema da Figura 5.4. Em concreto, tentar-se-á retirar do conjunto de variáveis preditivas CM (com o subgrupo C a incluir apenas resultados curriculares do 1º semestre do aluno), todas aquelas que não contribuam positiva e significativamente para o desempenho do modelo.

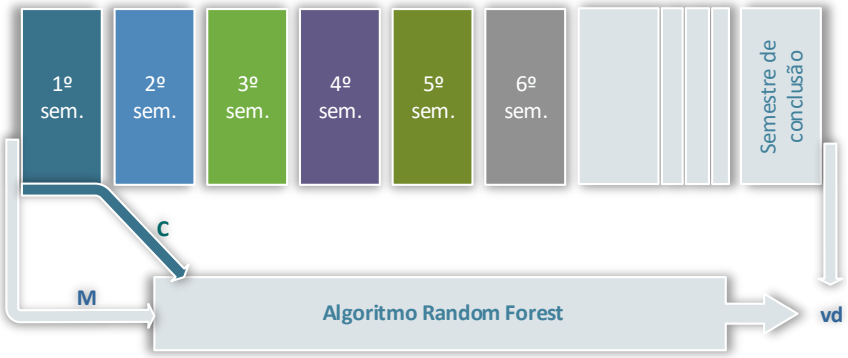


Figura 5.4: Esquema ilustrativo do modelo preditivo CM ajustado com base nos dados curriculares do 1º semestre escolar.

Como já foi referido anteriormente, tanto os dados académicos como os demográficos, provindo ambos da mesma BD dos Serviços Académicos do IPB, encontram-se completos. Atendendo a que o modelo CM prescindiu dos dados socioeconómicos e de acesso, as categorias com dados omissos num número muito significativo de estudantes — todos aqueles que não preencheram o inquérito no ato de matrícula e todos os que não ingressaram pelo Concurso Nacional de Acesso, respetivamente —, passa-se a dispor de uma amostra mais abrangente de matrículas com dados completos (para o conjunto de preditores previstos). Ao não incluir dados dessas categorias, o número de observações das tabelas com as varáveis preditivas mais que duplica, passando de 2159 para 4530, a dimensão total da amostra. Será com este novo *dataset* que se estudará o modelo CM. Acredita-se mesmo que ao se usar uma amostra de matrículas maior neste segundo ajuste do modelo, o mesmo apresentará maior capacidade de generalização.

Para uma melhor perceção do impacto que terá o alargamento da amostra no comportamento do modelo, apresenta-se, na Tabela 5.8, o desempenho do modelo CM nos primeiros 6 semestres escolares do aluno, usando para as suas variáveis preditivas, quer o *dataset* anterior, quer o novo conjunto de 4530 observações.

Observando os valores tabelados, constata-se que o desempenho do modelo decresce ligeiramente com o aumento do conjunto de observações, nos primeiros semestres do percurso escolar do aluno, invertendo-se depois essa tendência nos semestres mais avançados. Este efeito estará, muito provavelmente, relacionado com as especificidades do novo conjunto de observações que foi acrescentado. Enquanto que o primeiro conjunto de observações relacionava-se apenas com

Tabela 5.8: Desempenho do modelo CM para os *datasets* de 2159 e 4530 observações.

	2159 obs.		4530 obs.	
	R^2	RMSE	R^2	RMSE
1º sem	80.4	1.304	79.0	1.339
2º sem	86.5	1.084	85.6	1.113
3º sem	92.0	0.842	91.0	0.885
4º sem	94.3	0.712	94.0	0.724
5º sem	96.6	0.551	96.9	0.527
6º sem	97.9	0.437	98.3	0.394
Média[(a)]	88.4	0.966	87.6	0.989

[(a)] Média ponderada dos valores semestrais, com pesos 6, 5, ..., 2, 1, para os semestres 1º, 2º, ..., 5º, 6º, respetivamente.

alunos que ingressaram pelo contingente geral (facto pelo qual, tinham dados de acesso — dados do conjunto A), os novos dados estão associados a outras formas de ingresso, consideradas especiais. Como os alunos que não entram pelo regime geral provêm, na sua maioria, de outros cursos, conseguem, em média, quase o triplo dos ECTS creditados nos restantes — em concreto, verificou-se que, na amostra considerada, os alunos que entram pelo regime geral têm em média 3.86% dos ECTS creditados, quando a média para todos os outros é de 11.13%. Como os ECTS creditados são tendencialmente de UCs iniciais dos planos de estudos dos cursos, o aumento das creditações acaba por ter um maior impacto nos primeiros semestres do aluno, tornando-os mais atípicos. Portanto, pensa-se ser essa maior atipicidade a razão do menor acerto do modelo nos primeiros semestres escolares do aluno.

Antes de se avançar para um processo mais sistematizado de afinação do modelo CM, há dois atributos do conjunto de preditores que requerem uma atenção especial, uma vez que poderão apresentar uma forte correlação com outros também incluídos no *dataset*. Como as UCs dos planos estudos das licenciaturas do IPB são, na sua esmagadora maioria, de 6 ECTS — e as exceções têm a ver, quase sempre, com as UCs dos últimos semestres curriculares, como será o caso das cadeiras de projeto e de estágio —, contabilizar-se o número de UCs que são aprovadas ou reprovadas não será muito diferente de efetuar-se essa contabilidade em número de ECTS. Assim, pensa-se que talvez os atributos `nuca_s` e `nucr_s` pouco ou nada acrescem ao modelo preditivo, em relação aos atributos `ects_aprov_s` e `ects_reprov_s`, respetivamente. Depois de se analisar devidamente a pertinência destes dois atributos, que se passará a referenciar por 'iteração 0' do processo de ajuste do modelo, tentar-se-á, em várias iterações, avaliar o impacto das diferentes dimensões relativas ao estudante, para averiguar se há possibilidade de exclusão de algumas dessas dimensões, entre aquelas que se revelem menos informativas. Estrategicamente opta-se por um processo iterativo, em vez de se excluírem de uma só vez todas as dimensões que possam ser prescindíveis, atendendo à correlação que possa existir entre elas.

Tal como referido na secção anterior, as *random forest* já fazem, de alguma forma, uma seleção interna de variáveis. No entanto, após algumas simulações preliminares, considerou-se pertinente manter o processo iterativo de seleção de variáveis, uma vez que os resultados mostraram não serem exatamente iguais aos obtidos sem seleção de variáveis. Por exemplo, pelo processo iterativo, o `ano_mat` e `ano_curricular_s` constam no conjunto das 11 variáveis selecionadas finais, em detrimento das variáveis `navalr_s` e `max_s`, tal como se constata da Tabela 5.10.

Também pela análise da mesma tabela, é possível observar que se altera a ordem de importância entre as variáveis `cod_curso` e `ects_reprov_s`, e entre `ects_cred_tx` e `ects_aprov_s`, ao fim

das 5 iterações. Isso significa que o processo iterativo pode efetivamente conduzir a resultados diferentes. Para uma melhor clarificação do que se acabou de afirmar, repare-se que, se o objetivo fosse, por exemplo, chegar às 4 variáveis mais significativas, obter-se-iam diferentes resultados consoante se se partisse do conjunto inicial de 18 variáveis ou do conjunto de 11 variáveis a que se chegou na 5ª iteração (`cod_curso`, `ects_reprov_s`, `media_s` e `ects_cred_tx`, no primeiro caso e `cod_curso`, `ects_reprov_s`, `media_s` e `ects_aprov_s` no segundo). Visando uma fundamentação mais esclarecedora da opção pelo ajuste iterativo do modelo e da pré-seleção de categorias que o precedeu, apresenta-se, na discussão dos resultados, a Subsecção 5.6.1 com os resultados que se obteriam se se excluíssem de uma só vez todas as dimensões consideradas prescindíveis.

Segue-se uma breve descrição das iterações desenvolvidas no processo de seleção das variáveis.

Iteração 0 - Começou-se então por correr o algoritmo `randomForest` para o *dataset* com dados curriculares do 1º semestre, sem as variáveis `nuca_s` e `nucr_s`, e os resultados foram o esperado: $R^2 = 79.2$ e RMSE $= 1.334$, valores que superam, ainda que ligeiramente, os que se obtiverem com a inclusão dessas duas variáveis ($R^2 = 79.0$ e RMSE $= 1.339$, cf. Tabela 5.8). Optou-se, por essa razão, por retirar definitivamente do *dataset* as variáveis `nuca_s` e `nucr_s`, passando as restantes a revelar os valores de importância representados no gráfico da Figura 5.5.

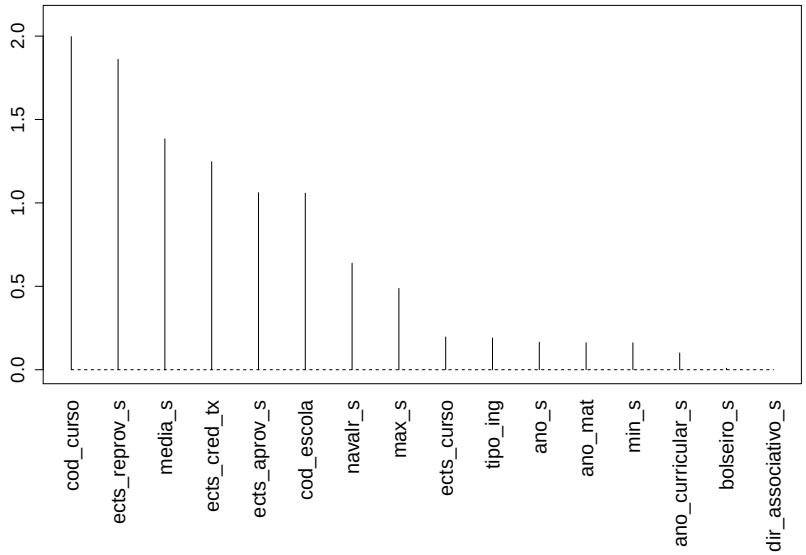


Figura 5.5: Importância das variáveis do modelo CM, com dados curriculares do 1º semestre e para um *dataset* de 4530 observações.

Iteração 1 - Observando o gráfico da Figura 5.5, percebe-se que há pelo menos 2 variáveis preditivas cuja influência no desempenho do modelo parece negligenciável: as variáveis `dir_associativo_s` e `bolseiro_s`. Optou-se, por isso, por excluí-las e confirmou-se que o desempenho do modelo se mantém praticamente inalterado: $R^2 = 79.2$ e RMSE $= 1.335$.

Na Figura 5.6 é mostrado o gráfico com a importância das 14 variáveis que se mantiveram no *dataset*. Os pequenos segmentos de linha horizontais, que sinalizam os valores anteriores à exclusão das duas variáveis, mostram que as importâncias das que ficaram se alteraram de forma desigual.

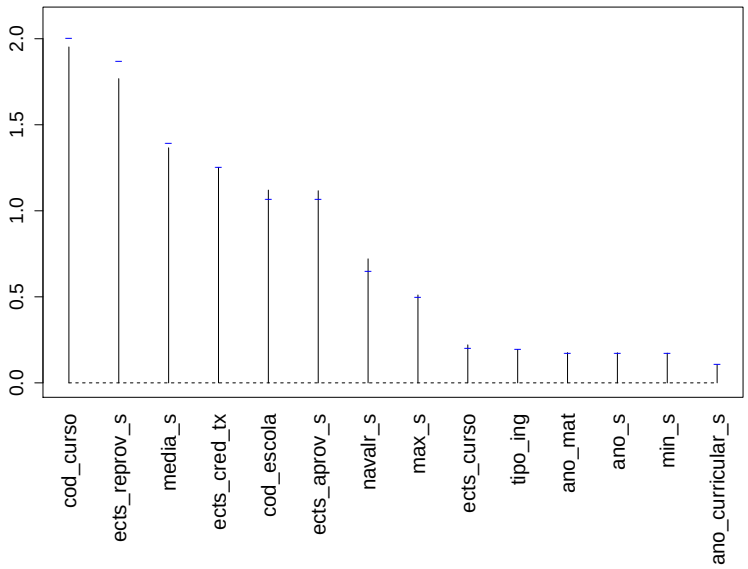


Figura 5.6: Importância das variáveis depois de excluídas as duas menos importantes.

Iterações 2-4 - Repetiu-se sucessivas vezes o procedimento adoptado na iteração 1, enquanto foi possível ir retirando do *dataset* as variáveis de menor importância sem comprometer o desempenho do modelo. Atendendo a que em cada uma dessas iterações que se seguiram já não foi possível identificar, de forma tão clara, as variáveis de influência mais negligenciável, optou-se por retirar apenas uma variável de cada vez, de entre as menos importantes. Dada a natureza do algoritmo RF e o seu processo construtivo, nem sempre a exclusão da menos importante de todas gerou os melhores resultados. Os dados relevantes que caraterizam essas várias iterações encontram-se resumidos na Tabela 5.9. Por sua vez, as importâncias das variáveis usadas nas várias iterações de ajustamento do modelo são mostradas na Tabela 5.10 e representadas graficamente na Figura 5.7.

Tabela 5.9: Remoção de variáveis do *dataset* CM.

	variáveis excluídas		nº var.	R^2	RMSE
iter. 0	`nuca_s`	`nucr_s`	16	79.2	1.334
iter. 1	`bolseiro_s`	`dir_associativo_s`	14	79.2	1.335
iter. 2	`min_s`		13	79.2	1.333
iter. 3	`max_s`		12	79.3	1.329
iter. 4	`navalr_s`		11	79.5	1.326

Observando a Tabela 5.9, constata-se que a perda de variáveis manteve, ou melhorou ligeiramente, o desempenho do modelo. Após a iteração 4, na perspetiva de se avançar para uma próxima iteração de ajuste do modelo, começou-se por seguir um procedimento análogo ao das iterações precedentes, percebendo-se que, nesse caso, a remoção de qualquer uma das 5 variáveis de menor importância (correspondentes ao 5 menores valores da última coluna da Tabela 5.10) conduziria sempre a um decréscimo da capacidade do modelo. Ainda que esse decréscimo tenha sido pouco significativo em algumas das remoções decidiu-se, seguindo uma abordagem conservativa, não continuar com o processo de eliminação de variáveis.
Em síntese, foi possível retirar do *dataset* 7 dos seus 18 atributos, sem que isso afetasse negativamente o desempenho do modelo, conseguindo-se, pelo contrário, um ligeiro aumento, ainda que pouco significativo. As 11 variáveis, que no seu conjunto justificam a capacidade preditiva

Tabela 5.10: Importância das variáveis usadas nas várias iterações de ajustamento do modelo CM.

atributo	iter. 0	iter. 1	iter. 2	iter. 3	iter. 4
cod_curso	1.997	1.952	2.065	2.159	2.011
ects_reprov_s	1.861	1.768	1.858	1.994	2.387
media_s	1.384	1.366	1.520	1.816	1.785
ects_cred_tx	1.247	1.252	1.323	1.415	1.454
ects_aprov_s	1.061	1.116	1.120	1.207	1.461
cod_escola	1.058	1.120	1.104	1.170	1.230
navalr_s	0.638	0.721	0.704	0.784	—
max_s	0.488	0.511	0.545	—	—
ects_curso	0.196	0.221	0.226	0.262	0.359
tipo_ing	0.190	0.194	0.213	0.242	0.274
ano_s	0.164	0.175	0.187	0.209	0.241
ano_mat	0.162	0.176	0.191	0.203	0.239
min_s	0.162	0.173	—	—	—
ano_curricular_s	0.101	0.106	0.124	0.134	0.176
bolseiro_s	0.009	—	—	—	—
dir_associativo_s	0.002	—	—	—	—

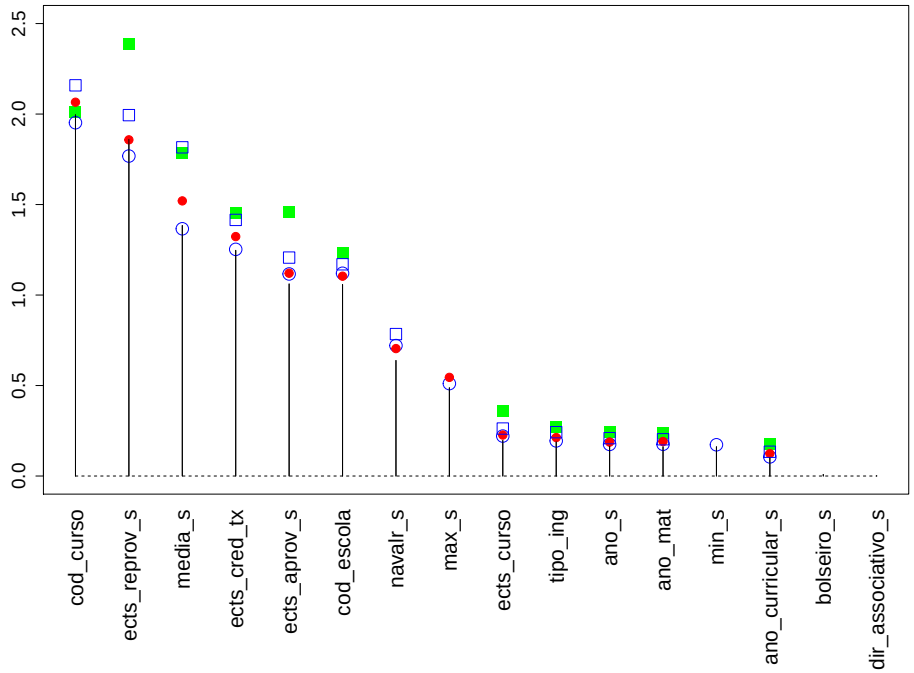


Figura 5.7: Importância das variáveis usadas nas várias iterações de ajustamento do modelo CM: (|) iteração 0; (○) iteração 1; (●) iteração 2; (□) iteração 3; (■) iteração 4.

do modelo, e que se revelaram, por isso, mais influentes para explicar o sucesso educacional dos estudantes de licenciatura do IPB, foram, por ordem decrescente de importância, as seguintes: `ects_reprov_s`, `cod_curso`, `media_s`, `ects_aprov_s`, `ects_cred_tx`, `cod_escola`, `ects_curso`, `tipo_ing`, `ano_s`, `ano_mat` e `ano_curricular_s`. Para uma mais fácil referenciação, este modelo preditivo de 11 variáveis, o principal resultado do estudo, passará a ser identificado nesta tese por modelo "CM Ajustado".

Para melhor percepção do verdadeiro impacto que tem a redução para 11 preditores no comportamento do modelo CM ao fim de cada um dos primeiros 6 semestres escolares do aluno, apresenta-se, na Tabela 5.11, o desempenho do modelo "CM Ajustado", usando para as suas

variáveis preditivas, quer os *datasets* de 2159 observações em que se baseou a seleção de categorias de preditores, quer o conjuntos completos de 4530 observações. É esclarecedora a

Tabela 5.11: Desempenho do modelo CM Ajustado para os *datasets* de 2159 e 4530 observações (cf. Tabela 5.8).

	2159 obs.		4530 obs.	
	R^2	RMSE	R^2	RMSE
1º sem	80.7	1.298	79.5	1.326
2º sem	86.7	1.081	86.2	1.094
3º sem	92.3	0.832	91.6	0.860
4º sem	94.7	0.694	94.5	0.698
5º sem	96.9	0.537	97.1	0.514
6º sem	97.9	0.442	98.3	0.400
Média[(a)]	88.7	0.958	88.2	0.971

[(a)] Média ponderada dos valores semestrais, com pesos 6, 5, ..., 2, 1, para os semestres 1º, 2º, ..., 5º, 6º, respetivamente.

confrontação dos resultados tabelados com os do modelo CM "não ajustado", entretanto apresentados na Tabela 5.8. À exceção do 6º semestre — aquele em que a capacidade preditiva se revela de menor importância —, em que os desempenhos são equiparados, apraz constatar que em todos os restantes, e independentemente da dimensão do *dataset*, o ajustamento para 11 variáveis conduz a um aumento efetivo na capacidade do modelo de previsão. De realçar que este incremento de desempenho apenas estaria garantido para o 1º dos semestres, uma vez que foi nos resultados desse 1º semestre escolar que se baseou toda a afinação que conduziu ao modelo "CM Ajustado". Ou seja, procurou-se o melhor subconjunto de preditores do modelo CM com o objetivo de aumentar a sua capacidade de previsão logo ao fim do 1º semestre escolar do aluno; esse mesmo subconjunto de preditores acaba por se revelar também adequado para os restantes semestres.

Na Figura 5.8 apresenta-se a função densidade da variável `vd` para o conjunto de 4530 matrículas consideradas. Como se pode constatar, a moda da variável dependente situa-se em torno do

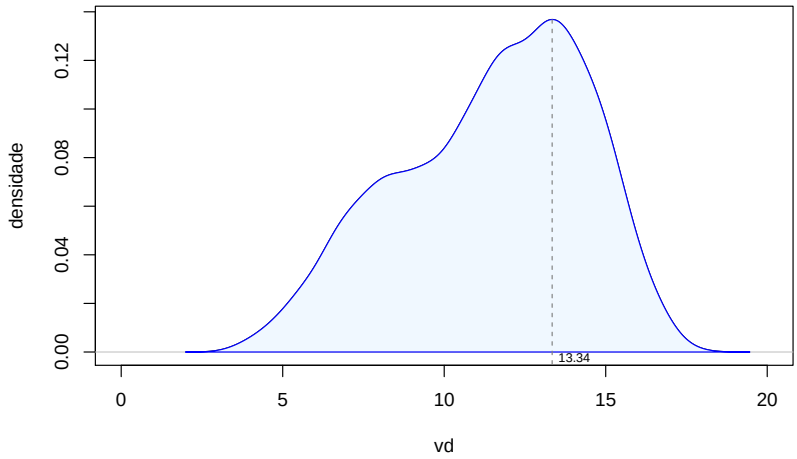


Figura 5.8: Função densidade da variável dependente `vd` para o conjunto de 4530 matrículas considerado no estudo do modelo CM Ajustado.

valor 13.34, e apenas assume valores abaixo de 10 porque a mesma, como se sabe, é afetada pelo rácio de aprovações — ver eq. (5.1). Três matrículas que apresentam desempenhos finais

alinhados com a moda são as que incluem, por exemplo, os seguintes perfis curriculares, no 1º semestre escolar: `media_s` $= 13.5/14.4/15.0$, `ects_aprov_s` $= 30/25/17$ e `ects_reprov_s` $= 0/5/13$.

Do grupo de variáveis a que se chegou, é o nº de ECTS reprovados (`ects_reprov_s`) que maior influência terá na capacidade preditiva do modelo. De forma a explicitar melhor esse tipo de influência, apresenta-se, na Figura 5.9, o gráfico da função densidade da variável dependente `vd` para os diferentes valores da variável `ects_reprov_s` (número de ECTS reprovados no 1º semestre escolar do aluno), depois de discretizada para múltiplos de 6, o nº de ECTS correspondente a uma UC típica.

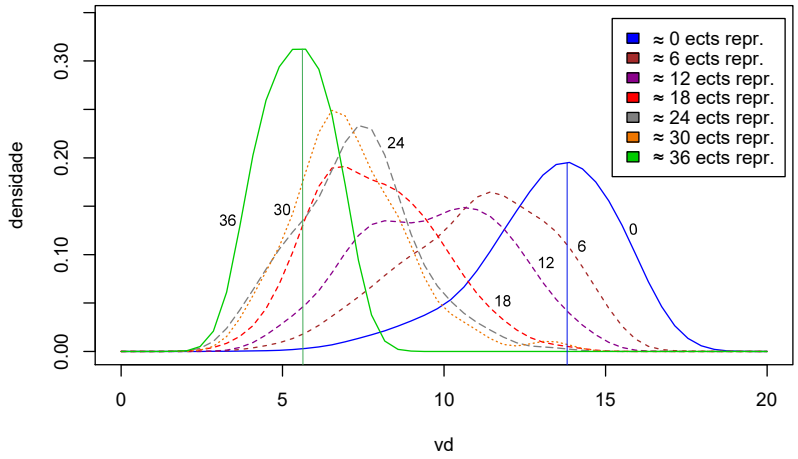


Figura 5.9: Função densidade da variável dependente `vd`, por número de ECTS reprovados no 1º semestre escolar do aluno.

Como rapidamente se perceberá pela análise do gráfico, existe de facto uma forte correlação (negativa) entre a variável preditiva `ects_reprov_s` e a variável dependente `vd`. Repare-se que à medida que aumenta `ects_reprov_s`, os valores da `vd` distribuem-se cada vez mais à esquerda na escala de 0 a 20 do eixo horizontal. Por exemplo, se para o subconjunto de alunos que não reprovam em nenhuma UC no 1º semestre a moda da `vd` aproxima-se dos 14 valores (≈ 13.8), já para aqueles que reprovam o equivalente a 6 UCs logo no 1º semestre, a moda da `vd` não chega sequer aos 6 valores (≈ 5.6).

De salientar que uma das principais evidências dos resultados obtidos, que à partida seria algo difícil de antever, foi o facto de a variável `ects_reprov_s` ter-se revelado bem mais informativa, na previsão de sucesso, que propriamente a variável `ects_aprov_s`, que ocupa apenas a 4º posição no grupo de variáveis mais influentes. Esse facto é evidenciado pela maior proximidade entre as funções densidade que aparecem na Figura 5.10, referentes aos valores (depois de discretizados) da variável `ects_aprov_s` (número de ECTS aprovados no 1º semestre escolar do aluno).

A maior importância da variável `ects_reprov_s` na capacidade preditiva do modelo em relação à variável `ects_aprov_s` poderá ter a explicação que se segue. Um aluno, que no 1º semestre reprove a muitas UCs (`ects_reprov_s` elevado) dificilmente se revelará um bom aluno, uma vez que inicia o seu percurso escolar com um desempenho claramente deficitário. Já a um aluno que no 1º semestre obtenha aprovação a poucas UCs (`ects_aprov_s` baixo) não se deve associar necessariamente um mau desempenho, pois poderá até ter conseguido aprovação, ainda que a poucas, a todas a que se inscreveu — por exemplo, quando lhe são creditadas muitas UCs do 1º semestre. Será este o facto pelo qual a função densidade da `vd` para alunos com 0 ECTS

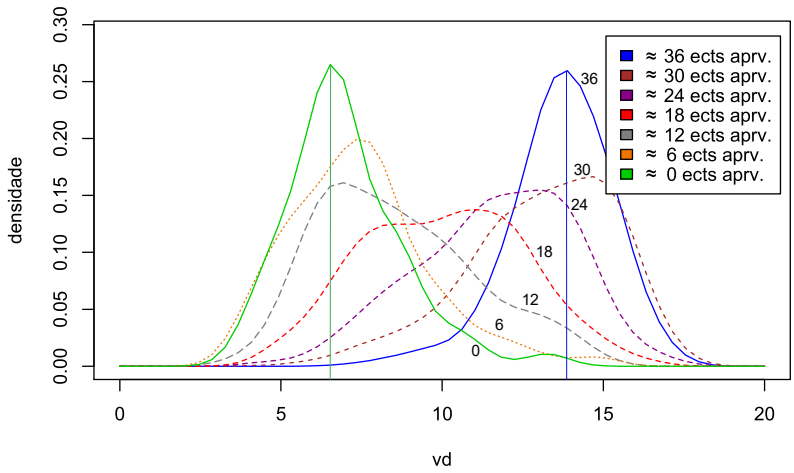


Figura 5.10: Função densidade da variável dependente `vd`, por número de ECTS aprovados no 1º semestre escolar do aluno.

aprovados (Figura 5.10) surge claramente mais à direita e mais estendida que a função densidade de todos aqueles que apresentam maior número de ECTS reprovados (Figura 5.9).

Já na situação em que as variáveis darão indicação de bom desempenho (`ects_reprov_s` baixo e `ects_aprov_s` elevado), o mesmo tipo de argumentação, da usada anteriormente, aponta para que sejam os ECTS aprovados a dar uma contribuição preditiva maior: um aluno que no 1º semestre obtenha aprovação a muitas UCs é necessariamente um aluno com bom desempenho, enquanto que um aluno que reprove a poucas UCs poderá, ainda assim, ter tido um desempenho deficiente, uma vez que poderá ter resultado, por exemplo, de ter muitas cadeiras já creditadas no 1º semestre — será essa a explicação para o maior achatamento da função densidade associada aos alunos com 0 ECTS reprovados (Figura 5.9) quando comparada com a função densidade de alunos com 36 ECTS aprovados (Figura 5.10). Mas ainda que isso aconteça, o facto de lhe terem sido já creditadas muitas UCs, torna-o, por essa via, num aluno com alguma propensão para o sucesso. Poderá, eventualmente, ser esta particularidade que explique o facto de não haver uma diferenciação assim tão notória entre a moda da função densidade para alunos com 36 ECTS aprovados (Figura 5.10) e a moda da função densidade para aqueles que apresentem menor número de ECTS reprovados (Figura 5.9). Ainda que os valores da `vd` se encontrem mais concentrados na 2ª das funções, a moda não difere muito entre os dois grupos de valores.

Em síntese, as variáveis `ects_reprov_s` e `ects_aprov_s` dão uma indicação do sucesso do aluno no seu 1º semestre com um grau de significância pouco diferenciado, mas é a 1ª das variáveis a dar uma indicação mais significativa do insucesso do aluno.

Tentou-se perceber as particularidades das variáveis `ects_reprov_s` e `ects_aprov_s`, que ajudassem a explicar, respetivamente, as suas 1ª e 4ª posições, nas variáveis preditivas mais importantes. Observando agora as restantes variáveis do grupo das 6 mais influentes, uma vez que as mesmas se destacam claramente, quanto à sua importância, das outras 5 das 11 selecionadas (do 1º grupo de variáveis para o 2º, a importância decresce de 1.230 para 0.359, como pode ser confirmado consultando a última coluna da Tabela 5.10).

Logo na 2ª posição surge o código que identifica o curso (`cod_curso`), revelando que o curso que frequenta o aluno acaba por ter também uma forte influência no seu sucesso escolar. Para melhor se perceber essa influência, mostra-se na Figura 5.11 as funções densidade da variável `vd` para os vários cursos da ESTiG, a escola que apresenta pior desempenho, como se verá já a seguir.

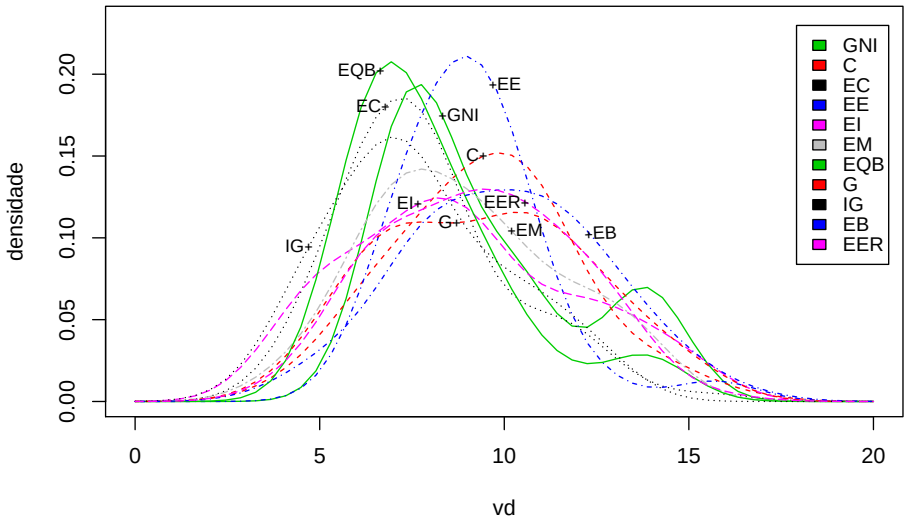


Figura 5.11: Função densidade da variável dependente `vd`, por cursos da ESTiG.

Já na 3ª posição não será de estranhar que surja a média do aluno no seu 1º semestre (`media_s`). Trata-se de um atributo que contribui diretamente para a medida de desempenho global do aluno, considerada, neste estudo, como indicador de sucesso. Também não será estranho não se assumir como o atributo mais informativo do desempenho do aluno, dado que apenas contabiliza as notas das UCs concluídas com sucesso, ignorando por completo todas as eventuais reprovações que o aluno possa ter.

Quanto à 5ª posição das variáveis preditivas mais importantes, a mesma é ocupada pela taxa de ECTS creditados ao aluno (`ects_cred_tx`) logo no início da sua formação, revelando-se também ela, e como esperado, um preditor com uma importante influência no sucesso escolar do aluno. Um aluno que já tenha à partida parte das suas UCs creditadas, terá, naturalmente, alguma propensão para o sucesso.

A terminar o grupo das variáveis mais influentes, surge, na 6ª posição, o código que identifica a escola (`cod_escola`), revelando que a escola que frequenta o aluno acaba por ter também ela uma forte influência no seu sucesso escolar. Para melhor se perceber esta influência, apresenta-se na Figura 5.12 a função densidade da variável dependente `vd` para cada escola em separado. Percebe-se claramente que entre as 5 escolas há duas que se destacam pelo sucesso escolar dos seus alunos, que, em média, é bastante superior ao das restantes. Tal evidência pode indiciar a possibilidade de diferenciação de estratégias promotoras do sucesso escolar a adotar por cada uma das escolas para mitigar o insucesso. Aparentemente, as três escolas de menor sucesso poderão ter a necessidade de medidas mais ativas e permanentes do que as outras duas.

5.5.4 Estudo do modelo preditivo CM Ajustado suportado por dados semestrais não acumulados

O modelo de previsão de sucesso que foi objeto de estudo foi suportado por um conjunto de dados que integrava, como parte importante, os resultados curriculares semestrais acumulados ao fim de cada um dos 6 primeiros semestres escolares do aluno. Desse estudo resultou, como anteriormente descrito, o modelo suportado unicamente por 11 preditores, que passámos a designar por "CM Ajustado". Para completar a investigação que se vem descrevendo, estudou-se o desempenho do modelo CM Ajustado usando, para o subconjunto de dados curriculares do

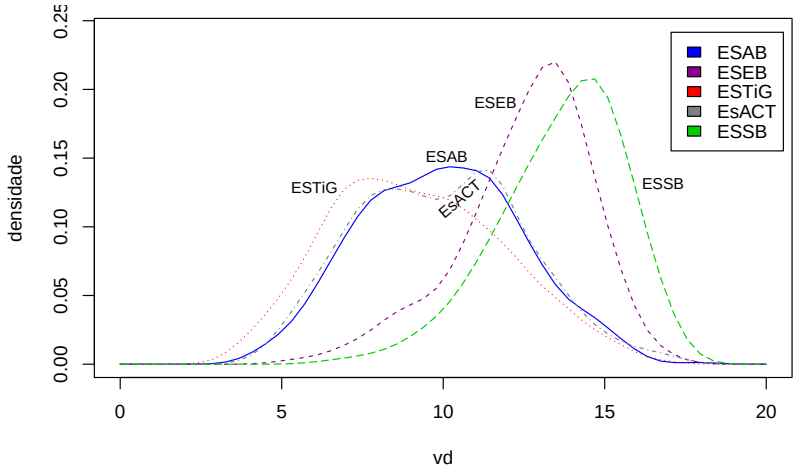


Figura 5.12: Função densidade da variável dependente `vd`, por escolas.

dataset, os resultados semestrais "não acumulados", de cada um dos 6 semestres escolares do aluno, tal como ilustrado no esquema da Figura 5.13 — com as 6 entradas de dados curriculares (C) desse esquema mutuamente exclusivas.

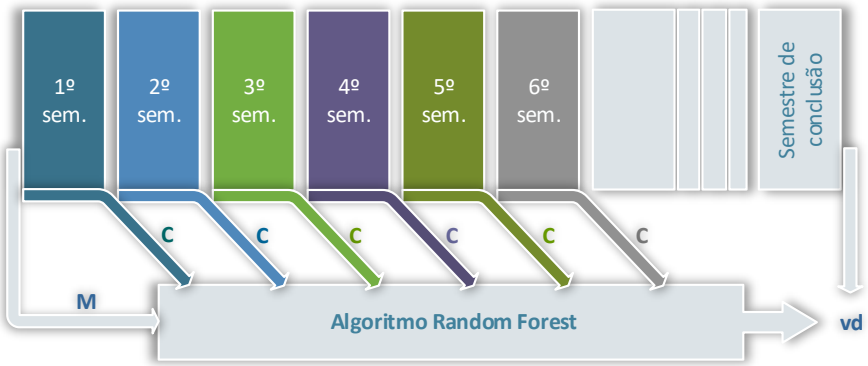


Figura 5.13: Esquema ilustrativo do modelo preditivo CM suportado por dados não acumulados.

Do *dataset* inicial fazem parte 3 atributos curriculares (`cod_estatuto_s`, `cod_freq_tipo_s` e `semestre_s` — atributos 4, 5 e 15 da Tabela 5.4) que não foram usados pelo modelo suportado por dados semestrais acumulados, precisamente por se tratarem de valores não acumuláveis. Embora a nova tipologia de dados curriculares permita que se inclua agora essas variáveis, constatou-se que, com elas, o desempenho do modelo CM Ajustado para o 1º semestre não melhorou, chegando mesmo a piorar, ainda que ligeiramente ($R^2 = 79.4$ e RMSE $= 1.328$). Mas mais sintomático são os valores quase nulos da importância dessas 3 variáveis reportados pelo modelo, e representados no gráfico da Figura 5.14 — confrontar com os valores do modelo CM ajustado anteriormente (no gráfico, pequenos segmentos de linha horizontais).

Não se perspetivando, assim, qualquer contribuição positiva na capacidade preditiva do modelo, optou-se por não incluir as 3 variáveis de valores não acumuláveis. A decisão de continuar com as mesmas 11 variáveis preditivas, tem também a vantagem de se poderem confrontar os resultados do modelo de dados não acumulados com os previamente obtidos no modelo de dados acumulados.

Correu-se então o algoritmo *random forest* para 6 *datasets* distintos, cada um contendo as mes-

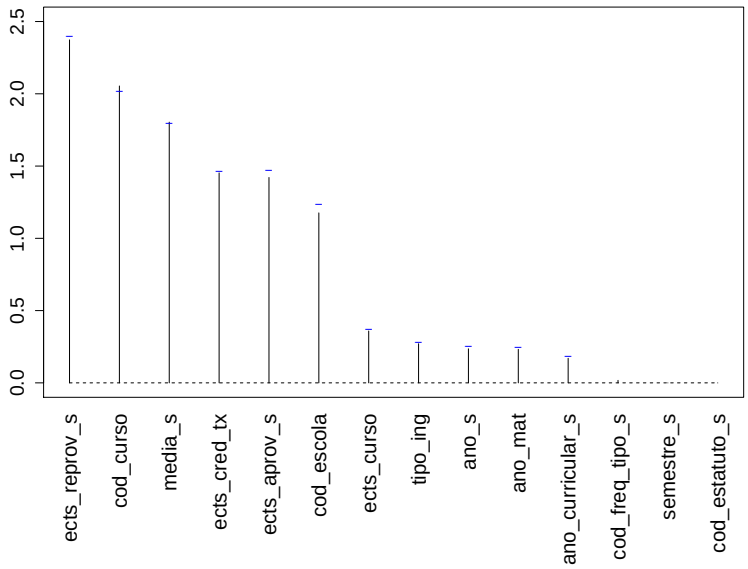


Figura 5.14: Importância das variáveis do modelo CM Ajustado depois de incluídas as três variáveis de valores não acumuláveis (3 últimas do gráfico).

mas 11 variáveis do CM Ajustado, mas com dados curriculares de um semestre diferente (ver Figura 5.13). Os resultados produzidos pelo modelo encontram-se na 2ª e 3ª colunas da Tabela 5.12. Observando esses resultados, percebe-se, desde logo, que, muito naturalmente, a

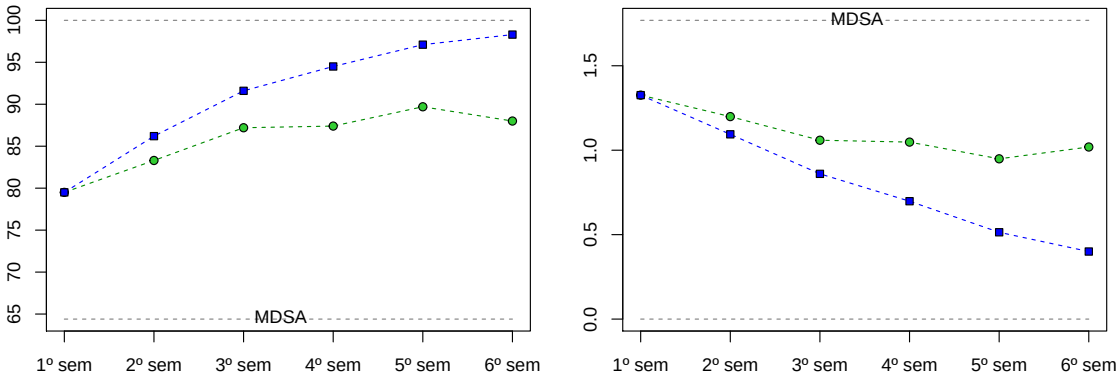
Tabela 5.12: Desempenho do modelo CM Ajustado com dados curriculares não acumulados.

	CM ajust (4530 obs)		CM ajust (2159 obs)		CMDSA (2159 obs)	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
1º sem	79.5	1.326	80.7	1.298	80.1	1.346
2º sem	83.3	1.199	84.2	1.179	83.6	1.230
3º sem	87.2	1.059	88.8	1.008	87.5	1.098
4º sem	87.4	1.048	87.9	1.038	87.0	1.106
5º sem	89.7	0.949	90.7	0.909	89.7	0.982
6º sem	88.0	1.019	89.3	0.972	88.2	1.044
Média[(a)]	84.4	1.155	85.5	1.125	84.6	1.188

[(a)] Média ponderada dos valores semestrais, com pesos 6, 5, ..., 2, 1, para os semestres 1º, 2º, ..., 5º, 6º, respetivamente.

capacidade preditiva do modelo é agora menor, quando já não se dispõe dos resultados curriculares acumulados — apenas dizem respeito ao semestre em causa. Ainda mais elucidativa será uma ilustração gráfica sobre os resultados dos dois modelos, com e sem dados curriculares acumulados. Efetivamente, através da Figura 5.15 é claramente visível o efeito que a acumulação dos resultados curriculares tem no incremento da capacidade preditiva do modelo.

Ainda assim, também o desempenho do modelo sem dados acumulados parece apresentar uma ligeira tendência de melhoria com o avanço dos semestres, em especial até ao 3º deles. Esta caraterística será o reflexo do tempo que o aluno leva a integrar-se e a adaptar-se ao sistema de ensino. É sobretudo durante os primeiros 2 semestres (1º ano) que o aluno vai adquirindo os métodos de estudo e todo um comportamento que o caracterizará ao longo do restante percurso escolar. Trata-se, de facto, de um ano de adaptação e de descoberta para o estudante, sendo o primeiro dos semestres, como esperado e em alinhamento com a literatura, aquele onde a capacidade preditiva se revela menor. Será apenas no seu 3º semestre que o aluno atinge a



(a) Coeficiente de Determinação (%)

(b) Erro quadrático médio

Figura 5.15: Desempenho do modelo CM Ajustado, quando suportado por: (■) dados curriculares semestrais acumulados; (●) dados curriculares semestrais não acumulados.

estabilidade comportamental que leva a que os seus resultados curriculares passem a ser mais indiciadores daquele que será o seu desempenho académico global.

Verifica-se também que o semestre onde os modelos atingem maior capacidade preditiva é o quinto e não o sexto, porventura, contrariamente ao que se estaria à espera. O sexto é um semestre algo particular, que estará, por isso, menos alinhado com o desempenho académico global do estudante. Para uma parte significativa dos estudantes, tratar-se-á do último semestre curricular, cujo plano de estudos inclui, por norma, unidades curriculares mais atípicas e específicas de fim de curso, como por exemplo: projetos, estágios e seminários.

Tentou-se também perceber até que ponto o modelo CM Ajustado se revelaria adequado para a nova família de *datasets* com dados não acumulados. Com o objetivo de o comparar com o modelo CMDSA, que integra a totalidade das variáveis (42 atributos — os da Tabela 5.4, incluindo os atributos 4, 5 e 15, de valores não acumuláveis, e excluindo quer os atributos 16 a 20, relacionados com a evolução entre semestres, quer os atributos 27, 28, 32 e 33, devido o seu elevado números de categorias), correu-se também o algoritmo para os seis *datasets* depois da filtragem de dados para excluir todas as observações com dados incompletos — sem dados de acesso ou de cariz socioeconómico. Dessa forma, foi possível comparar os dois modelos suportados pelas mesmas 2159 observações.

Como se constata da Tabela 5.12 (colunas 4 a 7), o modelo CM Ajustado continua a revelar um bom desempenho, mesmo depois de se ter mudado o conjunto de dados curriculares — supera o CMDSA em todos os semestres escolares. Quanto à diminuição da eficácia preditiva verificada no modelo CM Ajustado, quando se passa de 2159 para 4530 observações, a justificação será a já anteriormente avançada, aquando da análise que se fez aos resultados da Tabela 5.8.

5.6 Discussão de Resultados

Por via do modelo gerado foi possível identificar quais as dimensões mais explicativos do sucesso académico final. Através desse conhecimento é viável identificar o grupo de estudantes de maior risco de insucesso, o que permitirá a delineação de políticas promotoras de sucesso escolar.

Houve a pretensão de averiguar qual o poder explicativo do modelo logo no dia em que o aluno ingressa na instituição. Foi possível concluir que o desempenho preditivo conseguido nesse momento, ao não incluir dados curriculares, fica pelos 64.4% (Estudo 6, Tabela 5.6), passando para

os 80.3% (1º sem. do Estudo 1, Tabela 5.6) logo no final do primeiro semestre escolar, altura em que o modelo passou já a contar com os primeiros resultados curriculares do aluno. Esta constatação confirma a importância do fator desempenho académico semestral já anteriormente demonstrada por Manhães [64]. Estudos exploratórios preliminares já tinham, logo, demonstrado que as dimensões do contexto académico do estudante, e em particular do seu contexto curricular, eram determinantes para a previsão pretendida. De salientar que são ainda mais reveladores os principais resultados que depois se obtiveram, pois não poderiam estar mais alinhados com essa evidência: todos os 11 atributos que se revelaram mais significativos para a previsão, pertencem, sem exceção, à categoria da dimensão académica.
Sendo a abordagem seguida na seleção de caraterísticas (*feature selection*), que permitiu excluir por completo categorias de variáveis, um aspeto que diferencia o presente trabalho face à literatura existente será interessante perceber-se, de forma objetiva, se os resultados assim obtidos conduzem a um modelo com maior ou menor capacidade preditiva, quando comparados com os obtidos pela abordagem tradicional de seleção direta das variáveis preditivas, tendo em conta a sua ordem de importância estabelecida pelo algoritmo *random forest*. Na subsecção que se segue descreve-se esse estudo comparativo.

5.6.1 Comparação com o método de seleção direta de preditores

Na Figura 5.16 encontram-se representados graficamente os valores de importância das variáveis preditivas atribuídas pelo algoritmo *random forest*, quando aplicado ao fim do primeiro semestre e suportado por todas as categorias de variáveis (1º sem. do Estudo 1, Tabela 5.6). Para uma

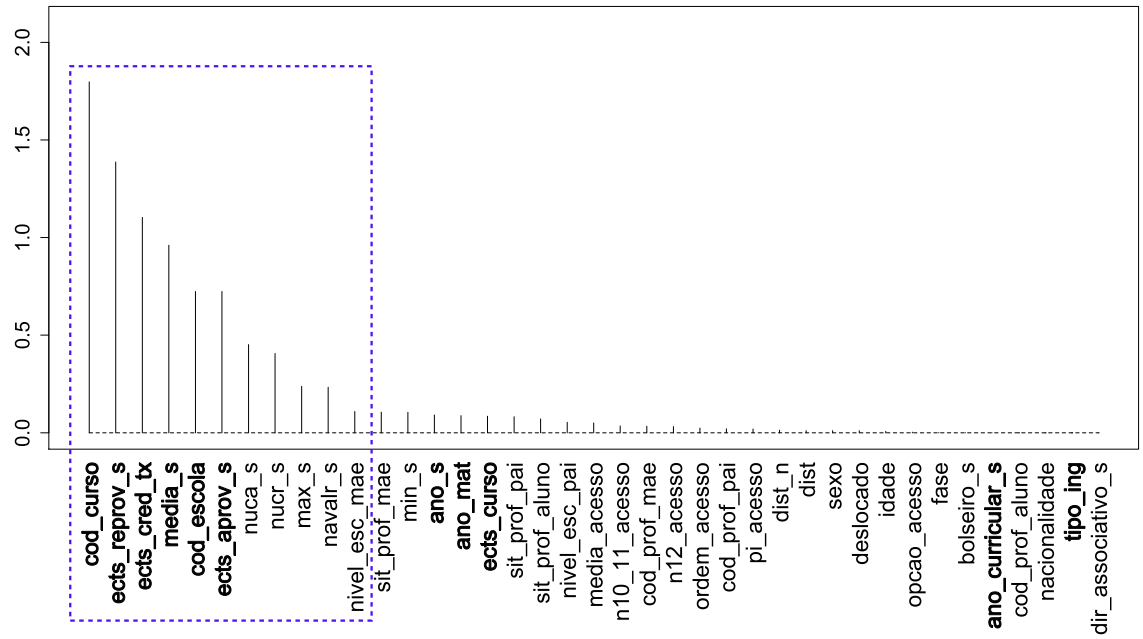


Figura 5.16: Importância das variáveis do modelo com todos os possíveis preditores aplicados aos dados curriculares do 1º semestre e para um *dataset* de 2159 observações.

mais fácil comparação de resultados, nesse mesmo gráfico delimitam-se numa área retangular, de contorno a tracejado, as 11 variáveis com maior ordem de importância para o algoritmo *random forest* e assinalam-se a negrito as 11 variáveis eleitas pela abordagem seguida no âmbito

do presente trabalho, que começa, como previamente descrito, pela pré-seleção de categorias e termina com o ajuste iterativo do modelo, culminando no modelo CM Ajustado. Com base nesse gráfico, é possível, desde já, avançar que os resultados que se obteriam com a seleção direta de preditores seriam claramente distintos dos que se obtiveram neste trabalho, no CM Ajustado. Repare-se que, nas 11 primeiras variáveis, que seriam selecionadas de uma só vez, constariam 5, a `nuca_s`, `nucr_s`, `max_s`, `navalr_s` e `nivel_esc_mae`, em detrimento das variáveis `ano_s`, `ano_mat`, `ects_curso`, `ano_curricular_s` e `tipo_ing`, que surgiram no conjunto das 11 variáveis do modelo CM Ajustado.
Conclui-se, portanto, que as duas abordagens conduzem a resultados efetivamente distintos. A fim de se perceber qual delas garantirá maior eficácia no momento de se efetuar a previsão de sucesso escolar correu-se o algoritmo *random forest* usando para variáveis explicativas precisamente as 11 variáveis com maior ordem de importância (variáveis delimitadas pelo contorno retangular a tracejado da Figura 5.16) entre todas as categorias de variáveis (CMDSA). É precisamente a esta solução que se designou "método de seleção direta de preditores", atendendo a que os mesmos são selecionados entre todas as categorias de variáveis e numa única iteração, a qual passará a ser conhecida mais resumidamente por modelo "CMDSA melhor 11".
Na Figura 5.17 é mostrado o gráfico com a importância das 11 variáveis que se mantiveram no modelo. Como esperado, a importância de todas essas variáveis aumentou consideravelmente

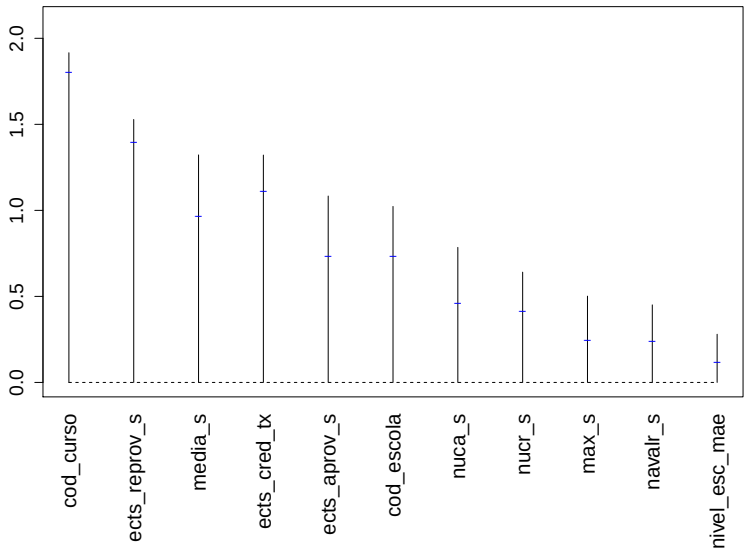


Figura 5.17: Importância das primeiras 11 variáveis depois de excluídas de forma "cega" todas as outras.

em relação aos valores que as mesmas apresentavam quando integradas no conjunto completo de variáveis CMDSA, representados no gráfico por pequenos segmentos de linha horizontais. A própria ordem de importância das variáveis alterou-se. Porém, o que será mesmo de realçar é que essa ordem também se alterou em relação às 6 variáveis em comum do modelo CM Ajustado, como facilmente se depreende, comparando a ordem das 6 primeiras variáveis da Figura 5.17 com a ordem dessas mesmas variáveis no modelo CM Ajustado, a seguir listadas por ordem decrescente da sua importância: `ects_reprov_s`, `cod_curso`, `media_s`, `ects_aprov_s`, `ects_cred_tx`, `cod_escola`. Esta constatação é mais uma evidência de que as duas abordagens conduzem efetivamente a resultados distintos.
Mas mais do que olhar para a importância relativa das variáveis explicativas e para a divergência dos resultados, há especial interesse em comparar o desempenho da capacidade de previsão

das duas abordagens: a que conduziu ao CM Ajustado, seguida neste trabalho, e a seleção direta de preditores, que designamos "CMDSA melhor 11". Nesse sentido, apresentam-se na Tabela 5.13 os coeficientes de determinação e os erros quadráticos médios residuais obtidos com esses dois modelos. Nessa tabela começa-se por mostrar os valores de desempenho do

Tabela 5.13: Confrontação de desempenho do modelo CM Ajustado com a seleção direta de preditores.

modelo	#preditores	semestre	#dataset	R^2	RMSE
CM Ajustado	11	1	4530	79.5	1.326
CM Ajustado	11	1	2159	80.7	1.298
CMDSA melhor 11	11	1	2159	79.5	1.333
CM Ajustado	11	1..6	6×2159	88.7*	0.958*
CMDSA melhor 11	11	1..6	6×2159	88.0*	0.984*

* Média ponderada que valoriza os resultados dos primeiros semestres, equivalente à que se usou para o cálculo dos valores médios apresentados nas Tabelas 5.6 e 5.7.

modelo CM Ajustado, obtidos com o *dataset* de 4530 matrículas que foi usado na sua definição e ajustamento. Mas atendendo a que o modelo "CMDSA melhor 11" usou um subconjunto desse *dataset*, constituído por apenas 2159 matrículas, houve a necessidade de correr o CM Ajustado também com esse conjunto mais restrito de observações, de forma a tornar os resultados mais comparáveis. Como se poderá então verificar pelos valores tabelados, o CM Ajustado apresenta efetivamente melhor desempenho, com $R^2 = 80.7$ e RMSE$= 1.298$, quando o melhor que conseguiu o modelo "CMDSA melhor 11" foi 79.5 e 1.333, respetivamente. De salientar, também, que a sua maior capacidade de acerto não se verifica apenas ao fim do 1º semestre. Ainda que tenha sido ajustado para esse semestre, o modelo CM Ajustado supera também o modelo obtido de forma tradicional (seu rival) no conjunto dos 6 primeiros semestres escolares do aluno, tal como confirmam os valores de desempenho médio mostrados nas duas últimas linhas da tabela.
Com o intuito de inferir se há evidências de diferenças estatisticamente significativas entre o desempenho dos modelos CM Ajustado e "CMDSA melhor 11", recorreu-se à metodologia dos testes de hipóteses, efetuando o teste paramétrico *t student* à diferença dos valores esperados para duas amostras emparelhadas. Depois de calculados os módulos dos erros de previsão dos dois modelos — que passaremos a designar simplesmente por resíduos — e de confrontadas as duas hipóteses subjacentes ao teste, expressas na forma:

H0: média dos resíduos do CM Ajustado $\geq$ média dos resíduos do "CMDSA melhor 11";

H1: média dos resíduos do CM Ajustado $<$ média dos resíduos do "CMDSA melhor 11";

obteve-se o valor de prova do teste de 0.04% pelo que, a um nível de confiança de 95%, há evidências estatísticas no sentido de afirmar que o valor esperado dos resíduos do modelo CM Ajustado é efetivamente inferior ao valor esperado dos resíduos do modelo "CMDSA melhor 11". Esta evidência vem corroborar que a precisão do modelo obtido pela nova abordagem proposta para a seleção dos preditores mais explicativos da previsão de desempenho global dos estudantes é efetivamente superior à precisão do modelo obtido pela abordagem tradicional de seleção direta de preditores.
Fica assim demonstrado o interesse e a pertinência da abordagem adotada neste trabalho para a seleção das dimensões mais explicativas do sucesso do aluno. Em síntese, para além de possibilitar a exclusão completa de categorias de variáveis e o possível consequente alargamento da amostra de alunos, a abordagem proposta introduz um incremento na capacidade preditiva do modelo em relação ao método que habitualmente é usado, o da seleção direta de preditores.

Como nota final, refira-se, ainda, que ao usar uma amostra de maior dimensão, o modelo obtido pela abordagem proposta, para além de mais preciso, acaba também por ser mais generalizável − enquanto que as 11 variáveis do modelo "CMDSA melhor 11" foram selecionadas com base em 2159 observações, as 11 do CM Ajustado foram obtidas a partir de todas as 4530 observações da amostra, acreditando-se, por isso, que o $R^2 = 80.7$ e RMSE$= 1.298$, para além de revelarem melhores desempenhos, traduzem resultados mais conservadores do que os valores 79.5 e 1.333, de R^2 e RMSE respetivamente.

5.7 Resumo e principais contribuições

Neste capítulo, usando o algoritmo *random forest*, propôs-se um modelo para prever o sucesso académico global dos estudantes de licenciatura do IPB, aquando do *terminus* do seu percurso académico.

Partindo-se de um *dataset* real contendo, entre os seus dados, resultados curriculares semestrais acumulados ao longo dos 6 primeiros semestres do aluno, começou-se por categorizar o conjunto de todas as potenciais variáveis preditivas em cinco grupos distintos, designados por dados *curriculares* (C), *de matrícula* (M), *demográficas* (D), *socioeconómicas* (S) e *de acesso* (A).

Aplicando o algoritmo de *data mining random forest* a diferentes combinações das categorias de variáveis estabelecidas, percebeu-se que quer os dados demográficos, quer os socioeconómicos, quer mesmo os de acesso ao ensino superior em nada contribuem para o desempenho do modelo preditivo, conseguindo-se, assim, passar de um *dataset* inicial de 44 variáveis para um *dataset* de apenas 23[5] variáveis preditivas, que foi designado por CM, dado ser composto unicamente por dados académico, retirados quer do contexto curricular (C) do aluno quer do seu ato de matrícula (M).

Percebeu-se também a grande influência dos dados académicos do tipo curricular no desempenho do modelo, mesmo que se limitem a refletir unicamente resultados do 1º semestre escolar do aluno. É esta particularidade que abre boas perspetivas para que a previsão de sucesso do aluno se possa efetuar numa fase ainda precoce do seu percurso escolar.

Seguidamente, usando para os dados curriculares (C) apenas os resultados do 1º semestre escolar do aluno, considerou-se a ordem de importância das variáveis mostrada pelo algoritmo *random forest*, para se proceder a um ajustamento mais minucioso do modelo CM, conseguindo-se chegar, dessa forma, a um modelo final suportado por apenas 11 variáveis explicativas e ainda com a sua capacidade preditiva reforçada. Conseguiu-se, depois, demonstrar a mais-valia dessa abordagem, que divide o processo de seleção de variáveis em duas fases, em relação ao método que habitualmente é usado, o da seleção direta dos preditores.

Adicionalmente, mostrou-se que o modelo CM ajustado a 11 variáveis mantém a sua adequabilidade, mesmo quando usado, para o subconjunto de dados curriculares, os resultados semestrais não acumulados. A utilização do modelo nesse tipo de dados também contribuiu para a perceção do tempo que tipicamente um aluno demora a adaptar-se ao sistema de ensino.

Entre as principais contribuições do presente trabalho para a literatura de EDM começar-se-á por apontar, por exemplo, a própria dimensão do problema estudado: bases de dados com tabelas de milhões de observações, grupos de alunos bastante heterogéneo, de mais de meia centena de licenciaturas, cobrindo as mais diversas áreas educacionais, e a integração de perto de meia

[5] Isso no caso de se incluírem todas as variáveis da família `vdij_s`, o que só acontecerá no *dataset* com os dados curriculares acumulados ao 6º semestre. Tratando-se, por exemplo, do *dataset* do 1º semestre, o nº de variáveis do modelo CM é de apenas 18.

centena de variáveis explicativas. Dito de outra forma, o estudo abarcou o universo de uma IES com 5 escolas, ao invés de se prever o sucesso esperado para os estudantes num só curso específico tal como é feito nos estudos já publicados. Tal metodologia possibilita que o modelo seja generalizável a outras instituições, sobretudo às do subsistema politécnico localizadas no interior do país, as quais apresentarão, provavelmente, o mesmo tipo de constrangimentos da instituição estudada.

Interessa igualmente sublinhar que, tanto quanto se conhece da literatura especializada, um novo fator curricular é considerado pela primeira vez: o tipo de escola. De salientar, neste contexto, que foi um dos fatores que mais se destacou no âmbito da previsão efetuada.

De relevar, também, que o tipo de abordagem adotada parece diferenciar-se daquela que usualmente tem sido seguida em trabalhos relacionados com a mesma temática — a de identificação das características do estudante que melhor expliquem o seu sucesso. Mais precisamente, no caso do presente trabalho, a seleção das dimensões relativas ao aluno explicativas do seu sucesso processou-se em duas fases distintas. Primeiro, selecionaram-se as categorias de atributos que melhor explicam o sucesso do aluno, conseguindo-se, com isso, um primeiro ajuste que permitiu eliminar grupos completos de variáveis. Depois, numa segunda fase, procedeu-se a um ajuste mais fino na seleção dos atributos que não foram excluídos na primeira fase. Destaca-se que com a abordagem mencionada foi possível, para além de se reduzir a "praga" da dimensionalidade dos dados, excluir por completo categorias de variáveis, sem se ter perdido a capacidade preditiva do modelo. Esta caraterística revela-se de especial importância, uma vez que concorre para a diminuição da multidisciplinaridade dos preditores, baixando, por essa via, alguma da complexidade do processo de previsão. E, mais importante ainda, possibilita o alargamento do estudo a uma amostra mais abrangente, como por exemplo, nos casos em que se excluam categorias com dados omissos — repare-se que no caso estudado, a não inclusão dos dados das categorias A e S, que sem se encontram em falta na maior parte dos registos, mais que duplicou o número de observações das varáveis preditivas, passando de 2159 para 4530, a dimensão total da amostra. Ou seja, é a abordagem seguida que viabiliza que o estudo de previsão se possa efetuar sobre a totalidade dos estudantes de licenciatura, e não apenas naqueles que, cumulativamente, tenham entrado pelo Concurso Nacional de Acesso (categoria A) e tenham respondido ao inquérito de recolha de dados socioeconómicos (categoria S).

Por fim, refira-se que também o facto de se fazer uma análise com dados agregados e outra com dados não agregados, consubstancia um aspeto diferenciador do presente trabalho, face aos estudos já publicados.

As deduções efetuadas por via da investigação desenvolvida, evidenciam, claramente, que através de um único algoritmo é possível extrair, de grandes conjuntos de dados, conhecimento útil que pode melhorar a tomada de decisões numa IES. 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.

Face aos resultados observados, poder-se-á indicar, como uma medida importante de promoção do sucesso académico, a supervisão atenta do número de ECTS reprovados pelo aluno, principalmente nos primeiros semestres do seu percurso escolar e com especial atenção para as matrículas em cursos com maior propensão para o insucesso, tal como sugestionam as duas variáveis que se revelaram mais explicativas, `ects_reprov_s` e `cod_curso`. Assim, será possível providenciar o devido acompanhamento aos estudantes com apetência para o insucesso. Por exemplo, motivá-los a adotar estratégias e rotinas de estudo mais assertivas, a frequentarem o horário de atendimento e a não faltarem às aulas, poderão fazer com que o insucesso académico seja mitigado.

Capítulo 6

Previsão de abandono académico

6.1 Introdução

"As Instituições de Ensino Superior não têm ferramentas para prevenir o abandono escolar",

João Pedro Videira, presidente da Federação Académica do Porto,
in Jornal público, 4 de abril de 2018.

O abandono escolar no ensino superior é um problema de índole académica, económica, política e social, de grande visibilidade, que tem motivado inúmeros especialistas, de múltiplas áreas do saber, a prever a sua ocorrência e a perscrutar as razões do fenómeno. A aplicação de métodos de *data mining*, em estudos desta índole, é uma tendência de investigação emergente que se tem revelado muito eficaz e promissora (Romero and Ventura [95], Papamitsiou and Economides [84], Peña-Ayala [86], Burgos et al. [15]). Neste enquadramento, no presente capítulo, propõem-se dois modelos analíticos de classificação, desenvolvidos com o objetivo de prever, de forma precoce e precisa, quais são os estudantes das licenciaturas do IPB mais propensos à evasão académica. Pretende-se, igualmente, identificar quais são os principais fatores determinantes do abandono, a fim de se poder providenciar informação realista, oportuna e atempada que permita aos decisores institucionais, de forma fundamentada, a delineação de estratégias destinadas à diminuição dos índices de evasão discente, bem como ao aumento das qualificações dos estudantes.
A estrutura do restante capítulo é a que se segue. Na Secção 6.2 apresenta-se a motivação e os objetivos do estudo; na Secção 6.3 explicita-se a metodologia desenvolvida, descrevendo os procedimentos seguidos; na Secção 6.4, descreve-se o pré-processamento que foi necessário realizar para a definição do modelo de dados a usar na previsão do abandono, para se passar, na Secção 6.5, à afinação dos 3 algoritmos de *data mining* em que se basearão os modelos de previsão a usar na extração de conhecimento a partir do modelo de dados gerado. Em 6.6 procura-se identificar o conjunto de fatores que melhor explicam o abandono escolar e na Secção 6.7 é proposta uma forma de cálculo da importância relativa dos mesmos. Por fim, na Secção 6.8, analisam-se os resultados obtidos e apresentam-se as principais conclusões da investigação. Face aos resultados obtidos, por via dos modelos desenvolvidos, propõem-se também algumas sugestões que possam vir a contribuir para a diminuição dos índices de abandono académico.

6.2 Motivação e objetivos

De acordo com um estudo da Direção-Geral de Estatística da Educação e Ciência (DGEEC) [33], de março de 2018, subordinado ao tema "Percursos no Ensino Superior", nas IES portuguesas, a

taxa de abandono dos estudantes de licenciatura com duração de três anos, é de 29%. Os dados oficiais publicados pela DGEEC enfatizam, também, que mais de metade (54%) dos alunos que entram no ensino superior com média de acesso de 10 valores não termina a licenciatura.
É igualmente sublinhado, no mesmo documento, que é precisamente nos Institutos Superiores Politécnicos públicos, no conjunto das IES públicas e privadas dos dois subsistemas de ensino superior, onde menos alunos (44%) acabam a licenciatura nos três anos previstos.
No caso concreto do IPB os índices de abandono são ainda mais preocupantes. Com efeito, uma pesquisa nas bases de dados em estudo permitiu concluir que nas licenciaturas a taxa de abandono média é de 38% e atinge mesmo os 41% nas que têm duração de três anos. De realçar, ainda, que cerca de 55% dos alunos que entram no IPB com média de acesso de 10 valores não acaba a sua licenciatura. Em virtude destas índoles de insucesso afetarem um elevado número de estudantes, repercutindo efeitos nefastos para eles, para as suas famílias, para a reputação e sustentação das escolas e para a sociedade em geral, é cada vez mais urgente a necessidade de delinear estratégias preventivas, precoces e precisas, propícias à diminuição dos índices de evasão discente.
É neste contexto que se pretende desenvolver dois modelos analíticos de classificação, a fim de se identificar, logo no final do 1º e do 2º semestre do primeiro ano do curso de graduação, os alunos do IPB com propensão ao abandono académico, cumprindo-se assim um dos principais objetivos específicos avançados para esta tese: a criação de modelos de previsão que fundamentem decisões destinadas à melhoria da qualidade dos serviços prestados aos estudantes. Em simultâneo, pretende-se identificar e apresentar, por ordem de importância, os principais fatores que determinam o abandono. Para esse efeito, considerar-se-à um elevado e diversificado conjunto de variáveis explicativas que a literatura tem identificado como determinantes do abandono, designadamente, as de desempenho académico pré e após ingresso no ensino superior, as demográficas e as socioeconómicas. O conhecimento obtido por via dos modelos desenvolvidos poder-se-à revelar fundamental para a definição de estratégias de gestão centradas na diminuição dos índices de evasão dos estudantes. A relevância em desenvolver um estudo desta índole resulta, desde logo, da inexistência de qualquer estudo científico sobre a previsão do abandono académico, na instituição usada como caso de estudo.
Face aos objetivos estabelecidos, para o presente estudo elege-se uma abordagem de *data mining*, pelo facto de a revisão de literatura apresentada nos Capítulos 2 e 3 ter demonstrado o elevado potencial dos métodos que lhe são associados, como instrumento de investigação e análise, em estudos desta índole.
Na secção que se segue descreve-se a metodologia usada no desenvolvimento dos modelos de previsão de abandono escolar.

6.3 Metodologia

A metodologia a desenvolver no presente capítulo tem como objetivo o desenvolvimento de dois novos modelos analíticos de classificação, que permitam prever, em dois momentos distintos do primeiro ano do percurso académico, quais serão os estudantes das licenciatura do IPB que apresentam maior propensão para o abandono. A metodologia proposta também visa avaliar se através do desenvolvimento de modelos de conjunto, que combinem a integração de três algoritmos preditivos, se consegue minimizar o erro da previsão de abandono.
Para o efeito, começa-se por explorar a combinação de três das técnicas *data mining* mais usadas em problemas de classificação, já previamente descritas no Capítulo 2 desta tese: *random*

forest (§2.4.3.2), redes neuronais artificiais (§2.4.3.3) e *support vector machines* (§2.4.3.4). Em *data mining* não existe um método único que garanta o melhor dos desempenhos em todos os problemas, pois o desempenho de um modelo dependerá sempre das especificidades quer do problema estudado quer do *dataset* adotado ([100]). A combinação dos métodos surge então como solução pragmática para se tentar tirar partido das especificidades das diferentes técnicas de previsão.

Se na previsão de Sucesso Académico Global (§5) dos estudantes de licenciatura do IPB, a metodologia adotada visou, essencialmente, a identificação do conjunto de variáveis que melhor explicavam a capacidade de previsão do modelo desenvolvido pelo algoritmo Random Forest, no presente Capítulo, o foco do estudo incide, também, na identificação e configuração de diferentes técnicas de classificação. 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*), indicador previamente descrito na Secção 2.5.1 desta tese.

No presente estudo desenvolvem-se, usando as técnicas e a abordagem mencionadas, dois modelos de previsão de abandono distintos, um para ser aplicado ao fim do 1º semestre escolar e o outro ao fim do 1º ano letivo do aluno (ao fim do seu 2º semestre escolar). No desenvolvimento de cada um desses modelos, começa-se por treinar e afinar, de forma separada, os três algoritmos de *data mining* suportados pelo conjunto completo de variáveis preditivas disponíveis. Seguidamente, com o intuito de se poder comparar a capacidade preditiva dos modelos completos com a dos modelos de variáveis selecionadas (feature selection), tenta-se ajustar o conjunto de variáveis que suportam o modelo, selecionando apenas as que se revelem importantes para a previsão pretendida. Para esse efeito, adota-se no presente estudo, 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 ainda candidatas a entrar leve a um maior incremento no valor médio de desempenho dos três algoritmos de *data mining*. 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.

Depois de encontradas as variáveis mais explicativas, averigua-se, de seguida, qual é 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 *data mining* usados no estudo.

Tanto na fase de *forward search*, 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 *data mining* para cada um dos conjuntos diferentes de variáveis independentes que se considere.

Com o intuito de se modelar de uma forma mais fiel o processo de previsão real que se pretende implementar no futuro, opta-se, neste estudo, por definir os subconjuntos de treino, de validação e de teste por um particionamento temporal do *dataset* original. 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. Repare-se que em contexto real, a propensão para o abandono de um determinado aluno será sempre inferida através de um modelo de previsão treinado e validado com conjuntos de observações de anos letivos anteriores. Pelo facto de se optar por um particionamento do tipo temporal, está-se a optar por um particionamento simples, excluindo-se, desta forma, a aplicação da técnica de validação cruzada. Ainda que a validação cruzada seja atualmente uma técnica muito usada pela comunidade científica de *data mining*, é nos *datasets* de pequena dimensão que mais se faz sentir a sua verdadeira

capacidade. O *dataset* do atual estudo, sendo composto por um número bastante elevado de observações, permite que se prescinda da técnica de validação cruzada, conseguindo-se, também com isso, uma redução significativa do esforço computacional envolvido. Trata-se de uma vantagem particularmente importante atendendo à elevada complexidade computacional que envolve a aplicação da técnica *forward search* adotada.
Embora os dados disponíveis para o presente estudo abarquem 9 anos letivos, de 2007 (2007/2008) a 2015 (2015/2016), apenas são consideradas matriculas iniciadas entre 2007 e 2013, por forma a minimizar o desbalanceamento no tempo do *dataset*. Tendo em conta a distribuição das matrículas ao longo do período de análise, opta-se neste estudo pelo seguinte particionamento:

dados de treino – matrículas iniciadas entre 2007 e 2010, (59% do total de observações);

dados de validação – matrículas iniciadas em 2011 (19% do total de observações);

dados de teste – matrículas iniciadas em 2012 e 2013 (23% do total de observações).

Os dados de validação, conjuntamente 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 futuramente em contexto real, as classificações atribuídas pelos três algoritmos de *data mining* são combinadas de forma a conseguir-se um único veredicto. Propõe-se, como forma de combinação, que seja sempre escolhido o valor mais predominante das três classificações.
Refira-se, por fim, que o presente estudo de previsão de abandono académico é feito quer com dados recolhidos ao fim do 1º semestre escolar, quer com dados recolhidos ao fim do 2º semestre escolar, incorporando este segundo *dataset* os seus dados curriculares na forma agregada (acumulada). Na Figura 6.1 apresenta-se um esquema que ajuda a ilustrar o modelo de previsão desenvolvido. Nele encontram-se representadas as diferentes categorias de variáveis preditivas usadas como *input* dos algoritmos de *data mining*, tal como identificadas na Tabela 6.1. 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. 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.
Todos o cálculos computacionais relacionados com o processo de *data mining* inerente à investigação descrita neste capítulo, são realizados em R, no ambiente RStudio e usa-se, como extensão ao próprio R, *packages* específicos para aplicação das três técnicas de *data mining* de classificação. Já todo o pré-processamento que foi necessário realizar para este novo estudo de previsão foi feito em MySQL.

6.4 Preparação do modelo de dados para a previsão de abandono

6.4.1 Caraterização da variável alvo a prever

A variável alvo de previsão neste estudo será do tipo binário, assumindo o seu valor um dos seguintes significados: ‘abandona’ ou ‘não abandona’. Neste contexto dever-se-á sempre enten-

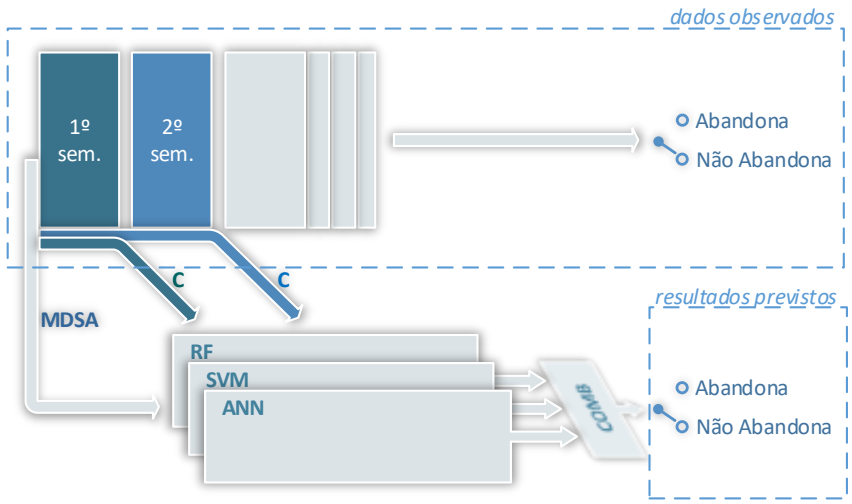


Figura 6.1: Esquema ilustrativo do modelo de previsão desenvolvido.

der a afirmação 'não abandona' como sinónimo de que o aluno 'conclui' o seu curso — repare-se que num sentido mais literal da frase, poderíamos afirmar o mesmo relativamente a um aluno que ainda esteja a frequentar um curso. Refira-se, por outro lado, que na base de dados inicial não se encontra registada qualquer informação que sinalize explicitamente os abandonos. Essa informação terá que ser de alguma forma inferida a partir de outros dados do percurso académico do aluno dentro do período em análise. Nesse sentido, e após consultado o principal responsável pelo Sistema de Informação do IPB, entendeu-se adequado passar a classificar como 'abandono' todas as matrículas referentes a alunos que não tenham qualquer inscrição válida no ano letivo a seguir ao período em análise (ano letivo de 2016/2017, tal como se depreende da Tabela 4.2) e, cumulativamente, não tenha ainda concluído o seu curso, nem, tão pouco, mudado para outro curso nesse mesmo ano letivo. Cada matrícula cujos dados (curriculares e extra-curriculares associados ao aluno) constem no *dataset* inicial poderá então ser classificada como 'abandona' (o aluno abandona), 'não abandona' (o aluno conclui) ou de desfecho indefinido. Nesta última classe estarão, entre outras, todas as matrículas com inscrição válida no ano letivo 2016/2017, pois trata-se dos alunos que continuavam a frequentar a instituição após o período de análise e sobre os quais não se sabe, como é óbvio, o seu trajeto subsequente. Todas estas matrículas foram naturalmente excluídas do *dataset*. De seguida, descrevem-se com maior objetividade os critérios adotados para a correta classificação do desfecho de cada uma das matrículas.

As matrículas que são concluídas com sucesso são facilmente identificáveis uma vez que o próprio *dataset* inclui colunas com a média e data de conclusão. Porém, como se sabe, uma franja importante dos alunos que ingressam na instituição acabam, pelas mais diversas razões, por interromper os seus cursos, nem sempre se traduzindo essas interrupções em verdadeiros abandonos. Entre aqueles que acabam por não concluir o seu curso é possível identificar cinco perfis distintos:

1. o aluno transfere-se para outro curso do IPB;
2. o aluno abandona o curso, ingressando num outro curso do IPB anos mais tarde;
3. o aluno abandona o curso, reingressando no mesmo curso anos mais tarde;
4. o aluno transfere-se para outra instituição de ensino;

5. o aluno abandona o ensino superior, nunca reingressando à instituição.

Destes perfis, percebe-se que o aluno pode abandonar o curso mantendo-se ou reingressando mais tarde na instituição ou, pelo contrário, decidir por abandonar definitivamente o IPB, desfechos com consequências completamente distintas para a instituição, que passarão a ser designados nesta tese simplesmente por "abandono do curso" e "abandono do IPB", respetivamente. Uma vez que é o "abandono do IPB" que verdadeiramente preocupa a instituição, será esse o significado que se dará à variável alvo "abandona". Sempre que não se refira explicitamente tratar-se de abandono de curso, o termo "abandona" deverá passar a ser entendido invariavelmente como "abandona o IPB".
À exceção dos dois últimos perfis, que serão classificados naturalmente como abandonos, os restantes requerem uma análise mais cuidada, uma vez que, nesses casos, o aluno acaba por vir a ter uma nova matrícula na instituição, que poderá inclusive não ser a última. Consequentemente, é relevante definir como classificar o desfecho de cada uma dessas matrículas do aluno. Uma vez que nos três perfis de abandono em causa, o aluno acaba por se manter ou reingressar mais tarde na instituição, considerou-se adequado classificar como 'indefinido' o desfecho deste tipo de matrículas. Se é verdade que o aluno não conclui o curso, não é menos verdade que, neste tipo de matrículas, o aluno também não abandona, pelo menos definitivamente, o IPB, que é o que interessa verdadeiramente avaliar. Ao classificar-se como de desfecho indefinido, essas matrículas são excluídas do *dataset*, tal como já se referiu anteriormente.

6.4.2 Pré-processamento e seleção de dados

O modelo de dados que reúna os principais fatores que poderão influenciar de alguma forma a propensão para o abandono, não será muito diferente do que foi usado para a previsão de sucesso global, o estudo que foi descrito no capítulo anterior. Nesta secção tentar-se-á relevar todo o processamento que ainda assim foi necessário realizar, derivado das especificidades próprias deste novo estudo de previsão, relacionado com o abandono. O *dataset* inclui essencialmente as mesmas variáveis explicativas (cf. Tabela 6.1), e as mesmas observações, relacionadas com o ambiente escolar, familiar e socioeconómico do aluno, com as diferenças que derivam do pré-processamento que é descrito a seguir.
Para o estudo de previsão de abandono académico que se pretende agora elaborar, usa-se como ponto de partida o subconjunto de dados composto pelas 13.381 matrículas a que se refere a 4ª célula da 4ª linha da Tabela 5.1, ou seja, todas as matrículas em licenciaturas que tiveram início entre os anos letivos 2007/2008 e 2013/2014. À semelhança do que se fez na previsão de sucesso, não se inclui nesta amostra inicial matrículas que se tenham iniciado nos últimos dois anos letivos do período temporal analisado, respetivamente 2014/2015 e 2015/2016, uma vez que os dados disponíveis, para esses alunos, nem sequer chegam a abarcar os 3 anos letivos previstos nos planos de curso da esmagadora maioria das licenciaturas do IPB.
Para além das tarefas de limpeza, filtragem e uniformização já asseguradas no pré-processamento do estudo de previsão de sucesso, que reduziu o *datadset* de 13.381 para 12.025 matrículas, procedeu-se a ajustamentos adicionais do *dataset* com o intuito de o preparar para posterior aplicação dos algoritmos de *data mining* no âmbito da previsão de abandono. Seguem-se as transformações mais relevantes que foi ainda necessário efetuar:

1. Adicionou-se a nova coluna binária `abandona` para a variável alvo, cujos valores foram calculados tendo em conta os critérios já descritos: atribui-se o valor `false` (o aluno conclui) nas matrículas (linhas) cujo atributo `ano_conclusao` assuma um valor não nulo;

nas restantes, atribui-se o valor `null` (desfecho indefinido) sempre que se sucedam novas matrículas no IPB do mesmo aluno ou quando o aluno se encontre inscrito no ano letivo 2016/2017; é nas matrículas não classificadas nem com o valor `false` nem com o valor `null` que se considera haver abandono, colocando-se assim no respetivo atributo o valor `true` (o aluno não esteve inscrito em 2016/2017, não concluiu o curso nem tem matrículas posteriores).

2. Removeram-se todas as matrículas de desfecho indefinido (atributo `abandona` com valor nulo), passando o *dataset* a ser composto por 10.135 matrículas e toda a informação a elas associada.

3. De modo a baixar o número de níveis da variável categórica `cod_curso`, optou-se por não considerar no estudo os 4 cursos com menor número de matrículas, respetivamente 5, 8, 9 e 11, excluindo-se então do *dataset* todas as observações associadas a essas matrículas. Com esta transformação, o número de níveis da variável categórica `cod_curso` passou de 55 (número de cursos do IPB) para 51 (número de cursos considerados no estudo). Esta decisão justifica-se pelo facto dos *packages* usados para aplicação das três técnicas de *data mining* de classificação, particularmente as *random forest* — que limita a 53 o número de níveis permitidos —, terem dificuldade em lidar com variáveis categóricas com elevado número de níveis.

Refira-se ainda que no presente estudo não se aplicaram os filtros que foram usados na uniformização dos dados da previsão de sucesso e que levaram à exclusão de matrículas com menos de 6 semestres letivos frequentados e matrículas com elevadas taxas de ECTS automaticamente creditados.
Após o processamento descrito dispõe-se, na forma integrada, dos dados extra-curriculares associados ao aluno bem como dos resultados curriculares acumulados ao fim de cada um dos dois primeiros semestres escolares por si frequentados, para poderem ser usados, como variáveis explicativas, nos modelos de previsão de abandono. Dados esses, distribuídos por 2 tabelas que, para além das variáveis preditivas intemporais (categorias M, D, S, e A da Tabela 6.1) e da variável alvo `abandona`, incluem então um conjunto próprio de variáveis curriculares (categoria C da Tabela 6.1), em concordância com a descrição que se segue:

`semestre1` – tabela com 39 colunas (38 variáveis preditivas mais a variável alvo) e 9.989 linhas, que usa para as variáveis curriculares os resultados recolhidos ao fim do 1º semestre escolar do aluno;

`semestres12` – tabela com 42 colunas (41 variáveis preditivas mais a variável alvo) e 9.596 linhas, que usa para as variáveis curriculares os resultados acumulados do 1º e 2º semestres escolares do aluno;

Ainda que os *datasets* `semestre1` e `semestres12` sejam compostos por 9.989 e 9.596 observações cada, apenas, respetivamente, 3.373 e 3.344 dessas observações são consideradas nos estudos que envolvam todas as categorias de preditores, uma vez que apenas essas contêm dados completos para todas as variáveis envolvidas.
Como se referiu na secção anterior, esses *datasets* são ainda subdivididos em 3 subconjuntos, para serem usados no treino, validação e teste dos diferentes modelos de previsão estudados e cujas dimensões são mostradas na Tabela 6.2.

Tabela 6.1: Variáveis explicativas usadas na previsão de abandono.

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[b]	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**[a]	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**[a]	C	nominal	1..7	*tipo de frequência do aluno no 2º sem. escolar*
7	**dir_associativo_s**	C	contínuo[b]	0..1	*o aluno foi dirigente associativo no sem. 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**[a]	C	contínuo	-20..20	*diferença de desempenho do 1º para o 2º semestre*
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*

[a] Variável não existente nos modelos de dados com resultados recolhidos ao fim do 1º semestre escolar.
[b] Ainda que tipicamente de natureza binária, valor considerado contínuo de forma a poder traduzir o rácio de ocorrências no conjunto dos semestres agregados.

Tabela 6.2: Dimensão dos conjuntos usados para treino, validação e teste.

dataset	total	treino (2007 a 2010)	validação (2011)	teste (2012 e 2013)
semestre1	3373	58.7%	18.5%	22.8%
semestres12	3344	58.8%	18.6%	22.5%

6.5 Afinação dos modelos com todos os preditores

Na investigação que se descreve neste capítulo são usadas três técnicas de classificação: os algoritmos *random forest*, redes neuronais artificiais e máquinas de vetores de suporte. Nesta primeira fase do estudo, de forma a se conseguir uma primeira perceção da capacidade preditiva do modelo baseado nessas 3 técnicas, começa-se por suportar os algoritmos em todas as variáveis preditivas disponíveis, e que foram já enumeradas na Tabela 6.1. Com o intuito de se maximizar a capacidade do modelo de previsão, procedeu-se à afinação dos três algoritmos de classificação para ambos os *datasets* `semestre1` e `semestres12`, avaliando o desempenho dos algoritmos para diferentes valores de alguns dos seus hyperparâmetros mais importantes. Na Tabela 6.3 podem ser encontrados os resultados das simulações que foram realizadas nesse processo de afinação e, para uma mais fácil identificação, os melhores valores obtidos encontram-se sublinhados (e posteriormente resumidos na Tabela 6.4). Para cada valor dos hyperparâmetros, é então mostrado o desempenho do algoritmo de classificação, através do valor AUC, quer para os dados (de validação) recolhidos ao fim do 1º semestre (*dataset* `semestre1`), quer para os dados (de validação) recolhidos ao fim do 2º semestre (*dataset* `semestre12`).
Tendo em conta os valores ótimos encontrados, resumidos na Tabela 6.4, 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 desempenhos nas outras duas técnicas de classificação.

Tabela 6.3: Resultados das simulações para afinação dos três modelos de classificação.

RF			SVM			ANN			
mtry	AUC		cost	AUC		size	decay	AUC	
	1ºsem	2ºsem		1ºsem	2ºsem			1ºsem	2ºsem
1	0.8484	0.8676	2^{-5}	0.8531	0.8655	1	$10^{-8/3}$	0.8303	0.8342
2	0.8445	0.8807	2^{-3}	0.8470	0.8612	1	$10^{-7/3}$	0.8418	0.8851
3	0.8510	0.8889	2^{-1}	0.8530	0.8638	1	$10^{-6/3}$	0.8515	0.8950
4	0.8530	0.8766	2^{1}	0.8517	0.8611	1	$10^{-5/3}$	0.8520	0.8909
5	0.8548	0.8805	2^{3}	0.8528	0.8792	1	$10^{-4/3}$	0.8504	0.8938
6	0.8494	0.8750	2^{5}	0.8554	0.8636	1	$10^{-3/3}$	0.8555	0.8954
7	0.8411	0.8786	2^{7}	0.8534	0.8585	1	$10^{-2/3}$	0.8639	0.8986
8	0.8374	0.8696	2^{9}	0.8471	0.8542	1	$10^{-1/3}$	0.8719	0.9007
9	0.8493	0.8730	2^{11}	0.5871	0.8608	1	$10^{0/3}$	0.8768	0.8993
10	0.8447	0.8658	2^{13}	0.8500	0.8776	2	$10^{-8/3}$	0.8457	0.8867
11	0.8413	0.8640	2^{15}	0.8487	0.8789	2	$10^{-7/3}$	0.8491	0.8695
12	0.8359	0.8665				2	$10^{-6/3}$	0.7969	0.8135
13	0.8391	0.8616				2	$10^{-5/3}$	0.8356	0.8656
14	0.8347	0.8663				2	$10^{-4/3}$	0.8405	0.8626

Tabela 6.3: (continuação da página anterior)

RF			SVM			ANN			
mtry	AUC		cost	AUC		size	decay	AUC	
	1ºsem	2ºsem		1ºsem	2ºsem			1ºsem	2ºsem
15	0.8342	0.8634				2	$10^{-3/3}$	0.8258	0.8856
16	0.8356	0.8574				2	$10^{-2/3}$	0.8521	0.8682
17	0.8353	0.8556				2	$10^{-1/3}$	0.8562	0.8923
18	0.8326	0.8499				2	$10^{0/3}$	0.8747	0.8973
19	0.8324	0.8565				5	$10^{-8/3}$	0.8590	0.8152
20	0.8286	0.8560				5	$10^{-7/3}$	0.8008	0.8484
21	0.8298	0.8511				5	$10^{-6/3}$	0.7631	0.8354
22	0.8314	0.8532				5	$10^{-5/3}$	0.8245	0.8412
23	0.8306	0.8535				5	$10^{-4/3}$	0.8101	0.8379
24	0.8280	0.8502				5	$10^{-3/3}$	0.8203	0.8510
25	0.8289	0.8483				5	$10^{-2/3}$	0.7973	0.8677
26	0.8213	0.8523				5	$10^{-1/3}$	0.8571	0.8793
27	0.8247	0.8509				5	$10^{0/3}$	0.8715	0.8942
28	0.8290	0.8413				10	$10^{-8/3}$	0.8100	0.8332
29	0.8259	0.8439				10	$10^{-7/3}$	0.7714	0.8359
30	0.8186	0.8428				10	$10^{-6/3}$	0.7956	0.8545
31	0.8207	0.8465				10	$10^{-5/3}$	0.8140	0.8399
32	0.8217	0.8434				10	$10^{-4/3}$	0.8164	0.8558
33	0.8232	0.8383				10	$10^{-3/3}$	0.8201	0.8807
34	0.8249	0.8468				10	$10^{-2/3}$	0.8067	0.8838
35	0.8258	0.8433				10	$10^{-1/3}$	0.8513	0.8901
36	0.8208	0.8438				10	$10^{0/3}$	0.8743	0.8960
37	0.8206	0.8418				20	$10^{-8/3}$	0.8262	0.8419
38	0.8204	0.8412				20	$10^{-7/3}$	0.8217	0.8573
39		0.8361				20	$10^{-6/3}$	0.8216	0.8594
40		0.8507				20	$10^{-5/3}$	0.8362	0.8588
41		0.8366				20	$10^{-4/3}$	0.8041	0.8692
						20	$10^{-3/3}$	0.8519	0.8880
						20	$10^{-2/3}$	0.8365	0.8812
						20	$10^{-1/3}$	0.8539	0.8772
						20	$10^{0/3}$	0.8728	0.8991
						50	$10^{-8/3}$	0.8339	0.8663
						50	$10^{-7/3}$	0.8362	0.8823
						50	$10^{-6/3}$	0.8237	0.8670
						50	$10^{-5/3}$	0.8396	0.8219
						50	$10^{-4/3}$	0.8435	0.8797
						50	$10^{-3/3}$	0.8455	0.8802
						50	$10^{-2/3}$	0.8615	0.8841
						50	$10^{-1/3}$	0.8599	0.8861
						50	$10^{0/3}$	0.8787	0.8973
						60	$10^{-8/3}$	0.8227	0.8503
						60	$10^{-7/3}$	0.8206	0.8596
						60	$10^{-6/3}$	0.8429	0.8754

Tabela 6.3: (continuação da página anterior)

RF			SVM			ANN			
mtry	AUC		cost	AUC		size	decay	AUC	
	1ºsem	2ºsem		1ºsem	2ºsem			1ºsem	2ºsem
						60	$10^{-5/3}$	0.8469	0.8818
						60	$10^{-4/3}$	0.8404	0.8771
						60	$10^{-3/3}$	0.8258	0.8971
						60	$10^{-2/3}$	0.8466	0.8906
						60	$10^{-1/3}$	0.8633	0.8925
						60	$10^{0/3}$	0.8785	0.8993
						80	$10^{-8/3}$	0.8387	0.8786
						80	$10^{-7/3}$	0.8408	0.8857
						80	$10^{-6/3}$	0.7943	0.8814
						80	$10^{-5/3}$	0.8634	0.8876
						80	$10^{-4/3}$	0.8570	0.8874
						80	$10^{-3/3}$	0.8683	0.8864
						80	$10^{-2/3}$	0.8487	0.8917
						80	$10^{-1/3}$	0.8492	0.8916
						80	$10^{0/3}$	0.8767	0.8994
						100	$10^{-8/3}$	0.8376	0.8692
						100	$10^{-7/3}$	0.8577	0.8777
						100	$10^{-6/3}$	0.8428	0.8887
						100	$10^{-5/3}$	0.8516	0.8822
						100	$10^{-4/3}$	0.8507	0.8891
						100	$10^{-3/3}$	0.8464	0.8932
						100	$10^{-2/3}$	0.8561	0.9006
						100	$10^{-1/3}$	0.8680	0.8942
						100	$10^{0/3}$	0.8719	0.9005

Tabela 6.4: Valores ótimos encontrados para os três modelos de classificação.

		RF		SVM		ANN		
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

6.5.1 Algoritmo *random forest*

Para aplicação do algoritmo de classificação *random forest* (RF) usou-se o método `randomForest`, do *package* R com o mesmo nome [61, 62], baseado na implementação original em Fortran de Breiman [14], que tinha sido já utilizado no problema de regressão tratado no capítulo anterior. Neste algoritmo de classificação ajustou-se o hyperparâmetro `mtry` — número de variáveis escolhidas aleatoriamente para o critério de divisão em cada nodo das árvores do RF — testando-se a capacidade do modelo para todos os seus possíveis valores, entre 1 e o número total de variáveis independentes. Para o número de árvores a construir, `ntree`, fixou-se o valor 200, dado não se prever que valores mais elevados produzissem incrementos de desempenho que justificassem o consequente agravamento do esforço computacional associado. Para os restantes parâmetros

de configuração do método usaram-se os seus valores definidos por omissão. Como se constata através dos valores tabelados, o melhor desempenho que se conseguiu com o modelo de previsão baseado no algoritmo *random forest* foi um AUC de 0.8548 para `mtry=5`, com os dados do 1º semestre, e um AUC de 0.8889 para `mtry=3`, com os dados recolhidos ao fim do 2º semestre.

6.5.2 Algoritmo máquinas de vetores de suporte

Para implementação do modelo de previsão baseado nas SVM (Support Vector Machines) recorreu-se ao *package* R `e1071`, da autoria de David Meyer et al. [70], que disponibiliza uma interface para o *package* `libsvm` anteriormente desenvolvido em C++ por Chih-Chung Chang e Chih-Jen Lin [22] e que consiste numa biblioteca de funções que facilitam a aplicação do algoritmo SVM nos diferentes tipos de problemas. Antes da aplicação do algoritmo SVM aos dados em análise, houve sempre o cuidado de normalizar para o intervalo $[0, 1]$ todos os atributos de tipo numérico, de forma a evitar dificuldades de cálculo desnecessárias e, principalmente, para evitar que atributos com grande amplitude de valores se sobreponham aos de reduzida amplitude. O desempenho das SVM depende do tipo de *kernel* selecionado, dos parâmetros de configuração do próprio *kernel* e do parâmetro de penalização do erro, designado neste trabalho por parâmetro de custo. Como se está perante *datasets* de elevada dimensão, de acordo com Hsu et al. [51], pode-se optar por um *kernel* linear e, dessa forma, confinar a afinação do modelo apenas à procura do melhor valor para o parâmetro de custo. Ainda segundo os mesmos autores, para procurar o melhor valor para esse parâmetro pode ser usada uma sequência de valores com crescimento exponencial do tipo $2^{-5}, 2^{-3}, \ldots, 2^{15}$. A partir dos resultados apresentados na Tabela 6.3 é possível concluir que o modelo baseado nas SVM apresenta um desempenho muito semelhante ao do modelo baseado nas *random forest*. Concretamente, o melhor desempenho foi agora conseguido com o parâmetro custo a assumir os valores 2^5 e 2^3, para os dados recolhidos, respetivamente, ao fim do 1º e 2º semestres, a que corresponderam os valores de AUC 0.8554 e 0.8792.

6.5.3 Algoritmo redes neuronais artificiais

Por fim, usou-se o *package* R `nnet`, para implementação do modelo de classificação baseado em redes neuronais (ANN – *artificial neural networks*). Trata-se de um *package* com algoritmos que implementam redes neuronais *feed-forward* de uma única camada escondida, desenvolvidos por Brian Ripley e William Venables [91, 115]. Tal com se fez no algoritmo SVM, e como recomendado na generalidade da literatura, também no modelo baseado nas redes neuronais artificiais começa-se por normalizar para o intervalo $[0, 1]$ todos os atributos de tipo numérico, de modo a evitar que aqueles que assumam valores de maior amplitude venham a ter uma maior preponderância no processo de treinamento. Como a configuração de uma rede neuronal requer o ajuste de um conjunto elevado de parâmetros, decidiu-se limitar a níveis considerados adequados a complexidade do processo de afinação, assumindo-se logo à partida a função logística como função de ativação, fixando em 100 o número de épocas usadas para treino e permitindo que os pesos das conexões da rede fossem sempre iniciados com valores aleatórios — opções assumidas por omissão pelo `nnet`. Ainda assim, a capacidade de previsão da rede é influenciada por dois importantes parâmetros: o número de neurónios que compõem a camada escondida (parâmetro `size`) e o decaimento da taxa de aprendizagem (parâmetro `decay`). No processo de afinação deste novo modelo foram então testadas, num esquema de *grid search*, todas as combinações de valores para esse par de parâmetros, com o número de neurónios a assumir os valores

$\{1, 2, 5, 10, 20, 50, 60, 80, 100\}$ e o fator de decaimento a assumir cada um dos valores da sequência de crescimento exponencial $10^{-8/3}, 10^{-7/3}, 10^{-6/3}, \ldots, 1$. Como se percebe da Tabela 6.3, a rede neuronal com melhor desempenho ao fim do 1º semestre é formada por um único neurónico na camada escondida e treinada com um decaimento na sua taxa de aprendizagem de $10^{-1/3}$, e a rede neuronal com melhor desempenho ao fim do 2º semestre contém 50 neurónios na sua camada escondida e é treinada com um decaimento de 1 na sua taxa de aprendizagem.
Pelos resultados tabelados é ainda possível concluir que, entre as três técnicas de classificação analisadas, foram as redes neuronais que demonstraram melhor *performance*, em ambos os *datasets* de validação considerados: apresenta ao fim do 1º semestre um AUC de 0.8787, algo acima dos 0.8554 das SVM, e ao fim do 2º semestre um AUC de 0.9007, também ligeiramente acima dos 0.8889 das *random forest*.

6.6 Seleção dos principais fatores explicativos do abandono

A seleção das variáveis de entrada é uma parte importante do processo de desenvolvimento de um qualquer modelo de previsão, uma vez que, para além de permitir, desde logo, identificar os principais fatores que expliquem as variáveis alvo, permite mitigar o impacto negativo que as variáveis menos informativas possam ter no desempenho global do modelo, ajudar a eliminar variáveis redundantes, diminuir a complexidade computacional do modelo e facilitar, quer a sua interpretabilidade, quer a sua aplicação em contexto real.
Com o objetivo de identificar os principais fatores que expliquem o abandono escolar, é usado neste trabalho o método de Seleção Progressiva (*forward search*), uma das várias estratégias conhecidas, destinadas a selecionar as variáveis de entrada de um modelo de previsão — para uma interessante revisão sobre as diferentes metodologias de seleção de variáveis, analisadas para o caso das redes neuronais artificiais, consultar [69].
A Seleção Progressiva é um método de procura incremental que vai selecionando as variáveis mais explicativas, uma a uma, até que a inclusão de uma nova variável não se traduza num aumento de desempenho do modelo. Nesse processo de seleção, as variáveis vão sendo escolhidas, naturalmente, em função do comportamento de um modelo de previsão específico que se esteja a desenvolver. Por conseguinte, sendo três os modelos em desenvolvimento no presente estudo, seria de esperar, como resultado final do processo de seleção, três subconjuntos de variáveis selecionadas, não necessariamente coincidentes. Estando em causa, no fundo, três processos de seleção, seria interessante conseguir-se, de alguma forma, combiná-los no sentido de ser alcançado um só subconjunto das variáveis mais explicativas que, de certa forma, reflita a contribuição dos três modelos. Com esse objetivo propõe-se uma forma combinada de seleção progressiva de variáveis, que se passa a descrever.
Para cada variável explicativa disponível, treina-se, separadamente, cada um dos 3 modelos, suportando-os unicamente nessa variável, e registam-se os valores do AUC que se vão obtendo. Seleciona-se, como primeira variável mais explicativa, aquela que maximizar o valor médio dos AUCs dos 3 modelos — é este critério de otimização, que tendo por base o desempenho agregado dos 3 modelos, permite combinar num único resultado o processo de seleção. Seguidamente, volta-se a treinar cada um dos modelos, mas desta vez suportados por cada um dos pares de variáveis formados pela variável já fixada e cada uma das restantes, escolhendo-se como 2ª variável a do par que traduzir um maior AUC médio. Este processo é depois repetido para todos os ternos de variáveis que se obtêm com a adição de uma nova variável às duas primeiras, e assim sucessivamente, adicionando-se em cada iteração uma nova variável às previamente selecio-

nadas. A procura finaliza quando a inclusão de uma nova variável não melhore o desempenho médio dos modelos. Com esta abordagem consegue-se, como pretendido, chegar a um único subconjunto de variáveis, consideradas mais explicativas.

Estando-se a lidar com *datasets* de elevada dimensionalidade e com um grande número de observações (dezenas de variáveis com milhares de observações), o nível do esforço computacional exigido pelas diferentes soluções, acaba por ser um fator determinante nas decisões a tomar. O método de Seleção Progressiva foi o escolhido para este estudo, precisamente, por se tratar de uma das abordagens computacionalmente mais eficientes. No seu processo de seleção, os modelos de previsão acabam por ser treinados com subconjuntos de variáveis de entrada relativamente pequenos: começa por considerar subconjuntos de uma só variável e, testando subconjuntos cada vez maiores, termina logo que o subconjunto "ótimo" seja atingido, que em geral ficará aquém do número de variáveis disponíveis à partida, especialmente quando este número assuma valores elevados. Relembra-se, no entanto, que este método, tal como a generalidade dos restantes, não garante soluções ótimas, pois não considera todas as combinações possíveis das variáveis candidatas. A exceção será mesmo o método de seleção exaustiva, mas como se percebe, com o aumento do número de variáveis de entrada a sua aplicação rapidamente deixa de ser exequível. Além disso, devido à natureza incremental da sua pesquisa, a Seleção Progressiva pode eventualmente ignorar combinações de variáveis bastante informativas que, ao não serem relevantes individualmente, não são selecionadas. Em todo o caso, acredita-se tratar-se de um dos métodos mais eficazes entre os que apresentam o mesmo nível de exigência computacional.

Com o objetivo de se conseguir resultados mais consolidados, ao longo de todo o processo de seleção das variáveis opta-se por voltar a afinar os modelos, sempre que o subconjunto de variáveis de entradas sofra alterações, garantindo-se assim que o desempenho é sempre determinado com o melhor conjunto de hyperparâmetros. Nessas inúmeras afinações, excetuando o hyperparâmetro `size` das ANN, todos os outros são testados com a mesma gama de valores que se usou na secção anterior, aquando da afinação dos modelos baseados em todas as variáveis disponíveis. Não sendo expectável que o processo iterativo de acumulação de novas variáveis explicativas se prolongue muito para além da dezena de variáveis, e atendendo ao enorme esforço computacional envolvido, opta-se neste caso por não testar as redes neuronais artificiais com mais de 20 neurónios na sua camada escondida ($\texttt{size} \leq 20$).

Procedeu-se então à seleção de variáveis, pelo método de Seleção Progressiva combinada, para os modelos de previsão suportados quer pelos dados recolhidos ao fim do primeiro semestre escolar do aluno, quer pelos dados recolhidos ao fim do segundo, tendo resultado das simulações produzidas nesses processos iterativos os valores reportados, respetivamente, nas Tabelas de E.1a a E.1h e de E.2a a E.2m, incluídas no Apêndice E desta tese. Cada uma dessas tabelas contém os resultados obtidos, com os dados de validação, numa iteração específica, ou seja, os hyperparâmetros ótimos e AUCs obtidos com os 3 algoritmos usados para cada uma das variáveis que se junta ao subconjunto de entrada já selecionado, dispostos por ordem decrescente do valor médio do AUC. É então a primeira linha de cada uma dessas tabelas que, reportando o melhor resultado obtido na respetiva iteração, contém a variável eleita, a ser integrada no subconjunto de variáveis explicativas já selecionadas. Essas primeiras linhas encontram-se reunidas, em forma de resumo, nas Tabelas 6.5 e 6.6.

Como se observa pela Tabela 6.5, respeitante ao *dataset* do 1º semestre, o processo de seleção de variáveis terminou à 8ª iteração, significando que a inclusão de uma 8ª variável, qualquer que ela fosse das ainda disponíveis, não incrementaria o AUC médio, a medida usada no critério de otimização. Conclui-se assim, com a abordagem seguida, que ao fim do 1º semestre escolar

Tabela 6.5: Variáveis selecionadas pela aplicação do método *forward search* ao *dataset* do 1º semestre.

			RF		SVM		ANN		
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
8º	dir_associativo_s	0.8671	3	0.8695	2^{-3}	0.8451	20	$10^{-3/3}$	0.8867

Tabela 6.6: Variáveis selecionadas pela aplicação do método *forward search* ao *dataset* dos 2 primeiros semestres.

			RF		SVM		ANN		
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

do aluno, são 7 as variáveis que mais explicam a sua propensão para o abandono, 4 das quais de natureza curricular, 2 representando dados de matrícula e uma de cariz demográfico. Já para os dados recolhidos ao fim do 2º semestre o processo de seleção terminou num subconjunto de variáveis mais alargado. Tal como mostrado na Tabela 6.6, são 12 as variáveis que se revelam mais explicativas ao fim do 2º semestre, não estando nelas representada apenas a categoria M, respeitante aos dados de matrícula.

6.6.1 Caraterização dos modelos de previsão encontrados

Dispõe-se já do subconjunto de variáveis consideradas mais explicativas ao fim do 1º semestre escolar e do subconjunto de variáveis consideradas mais explicativas ao fim do 2º semestre escolar, que darão suporte aos modelos de previsão de abandono escolar a propor. Cada um desses modelos é no fundo um conjunto de 3 modelos de previsão distintos. Em vez de se optar por um único algoritmo de classificação, usa-se uma combinação de três desses algoritmos, de modo a tirar proveito de três técnicas de classificação distintas. Uma vez que cada um dos algoritmos apresenta um viés diferenciado dos restantes, será de esperar algum efeito de compensação desse viés num modelo que combine as três técnicas. Gerando cada um dos 3 algoritmos uma classificação própria, houve a necessidade de combinar numa só as classificações produzidas. Tentaram-se várias formas de combinação, avaliando a sua eficácia com os dados de validação. Percebeu-se, com agrado, que a forma que conduz ao melhor desempenho é aquela que é mais comum entre os *ensemble methods*, e que consiste em assumir para a variável alvo

a classe mais "votada", ou seja, a que mais ocorre entre as 3 classificações produzidas pelas diferentes técnicas.

De acordo com os resultados obtidos, resumidos na Tabela 6.5, o modelo de previsão a usar ao fim do 1º semestre, que se passa a identificar pela mnemónica `var7s1` (modelo suportado por 7 variáveis do *dataset* obtido ao fim do 1º semestre), é suportado pelas 7 variáveis de entrada `ects_aprov_s`, `ects_cred_tx`, `ects_reprov_s`, `cod_escola`, `media_s`, `sexo` e `bolseiro_s`, e os 3 algoritmos de classificação que integra são configurados com os hyperparâmetros ótimos `mtry` $= 2$ (algoritmo RF), `cost` $= 2^{-3}$ (algoritmo SVM), `size` $= 20$ e `decay` $= 10^{-5/3}$ (algoritmo ANN).

Por sua vez, os resultados resumidos Tabela 6.6, sugerem que o modelo de previsão a usar no fim do 2º semestre, que se passa a identificar simplesmente por `var12s12` (modelo suportado por 12 variáveis do *dataset* obtido ao fim do 2º semestre), seja suportado pelas 12 variáveis de entrada `ects_reprov_s`, `cod_prof_mae`, `n10_11_acesso`, `idade`, `media_s`, `vd12_s`, `sit_prof_mae`, `sit_prof_aluno`, `nacionalidade`, `dir_associativo_s`, `nivel_esc_pai` e `cod_prof_aluno`, com os 3 algoritmos de classificação que integra configurados com os hyperparâmetros ótimos `mtry` $= 2$ (algoritmo RF), `cost` $= 2^{11}$ (algoritmo SVM), `size` $= 1$ e `decay` $= 1$ (algoritmo ANN).

Para uma mais fácil referenciação e compreensão da descrição que se segue nesta tese, 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 ao fim do 1º semestre) e `var41s12` (modelo suportado por todas as 41 variáveis do *dataset* obtido ao fim do 2º semestre) — ver Figura 6.2 para um melhor entendimento dos modelos de previsão que foram desenvolvidos e dos diferentes subconjuntos de preditores que lhes dão suporte.

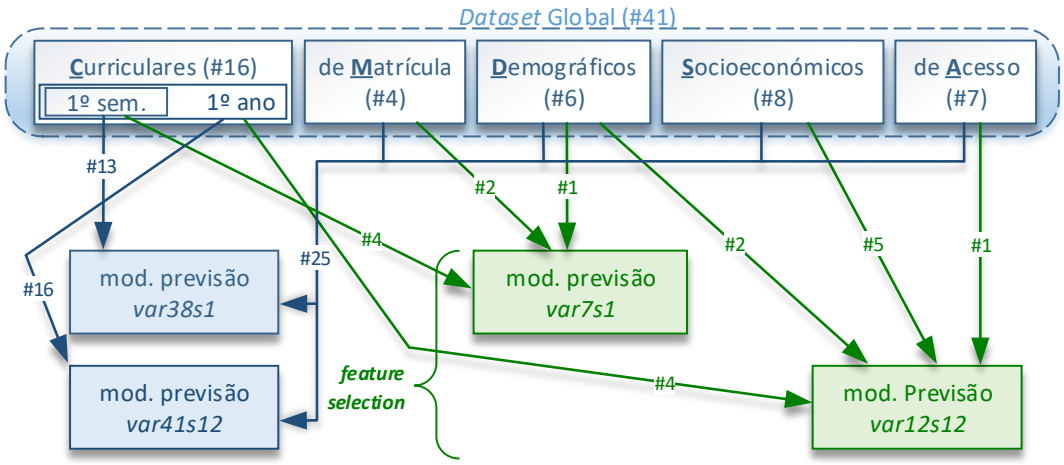
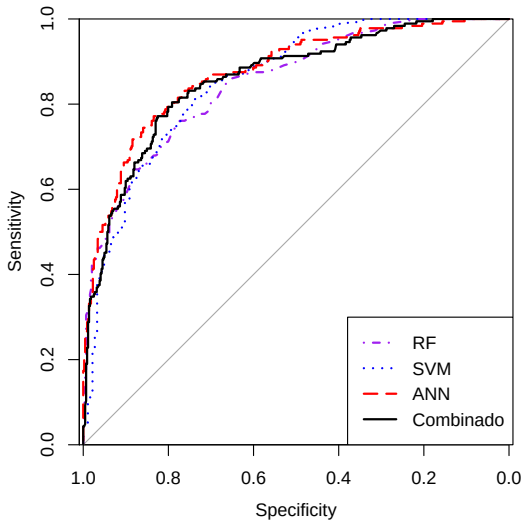


Figura 6.2: Diagrama ilustrativo dos subconjuntos de preditores usados nos diferentes modelos de previsão.

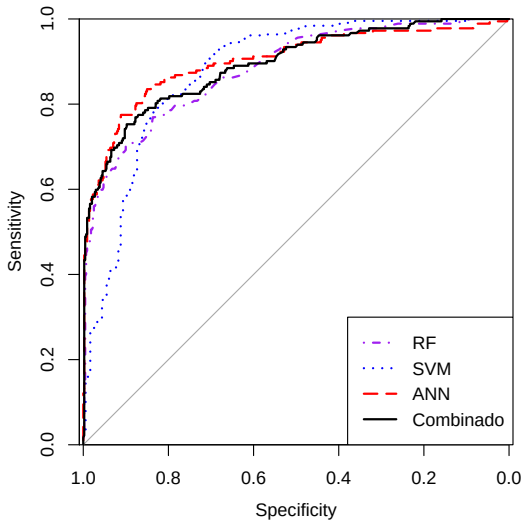
Com os dois modelos encontrados, `var7s1` e `var12s12`, para além de outras vantagens que resultam do simples facto de se passar a lidar com menos variáveis, conseguiu-se incrementar ligeiramente os valores do AUC. O desempenho, com os dados de validação, passou de 0.8630 (Tabela 6.4) para 0.8672 (Tabela 6.5) no 1º modelo, e de 0.8896 (Tabela 6.4) para 0.8977 (Tabela 6.6) no segundo. Refira-se, no entanto, que pouco significado terá este incremento de desempenho na capacidade preditiva real dos modelos, se o mesmo não se vier a confirmar na avaliação a efetuar, com dados de teste, na próxima secção do presente capítulo.

De forma a se ter uma outra perspetiva do comportamento de cada um dos modelos evidenci-

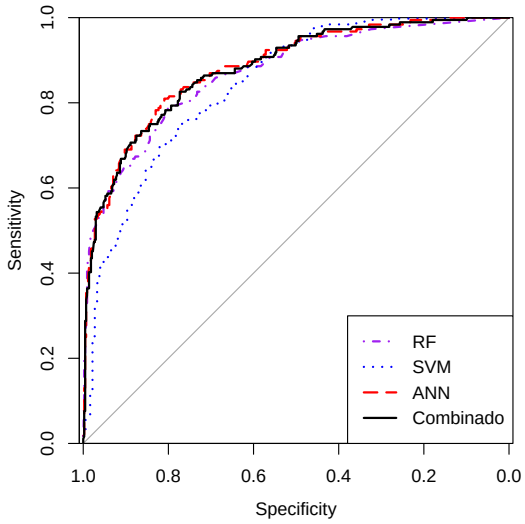
ado com os dados de validação, construíram-se as respetivas curvas ROC. Na Figura 6.3 estão representadas as curvas ROC quer para o modelo global, que combina os três algoritmos de classificação (legendado na figura com a designação 'Combinado'), quer para cada um desses três algoritmos, visto de forma isolada. Se se tiver em conta apenas os modelos com as suas



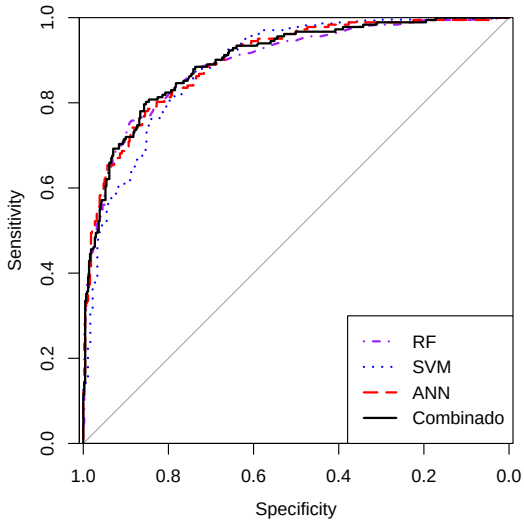
(a) `var38s1` – todas as variáveis/dados do 1º sem.



(b) `var41s12` – todas as variáveis/dados dos 2 sem.



(c) `var7s1` – 7 variáveis/dados do 1º sem.



(d) `var12s12` – 12 variáveis/dados dos 2 sem.

Figura 6.3: Curvas ROC dos modelos de classificação construídas com dados de validação.

variáveis selecionadas — gráficos c) e d) —, o modelo que combina o resultado dos três classificadores parece apresentar um desempenho superior àquele que será o desempenho médio dos classificadores em separado, aproximando-se mesmo do desempenho do melhor deles. Este resultado ajuda a reforçar a confiança no modelo combinado e em particular na sua forma de agregar num só os resultados dos três classificadores.

Comparando as curvas do gráfico a) com as do gráfico c) e as do gráfico b) com as do gráfico d), percebe-se que os modelos de previsão que perderam parte das suas variáveis apresentam um desempenho ligeiramente superior. Quando se comparam os gráficos a) com b) e c) com d), é também percetível uma ligeira melhoria no desempenho dos modelos quando são aplicados ao

dataset com dados recolhidos ao fim do 2º semestre.
Não sendo suficientemente claras nem evidentes, nos gráficos da Figura 6.3, as interpretações anteriores, para uma melhor perceção do que se acabou de afirmar, sobrepõem-se, na Figura 6.4, as curvas ROC dos 4 modelos combinados. Neste novo gráfico percebe-se que a curva que mais

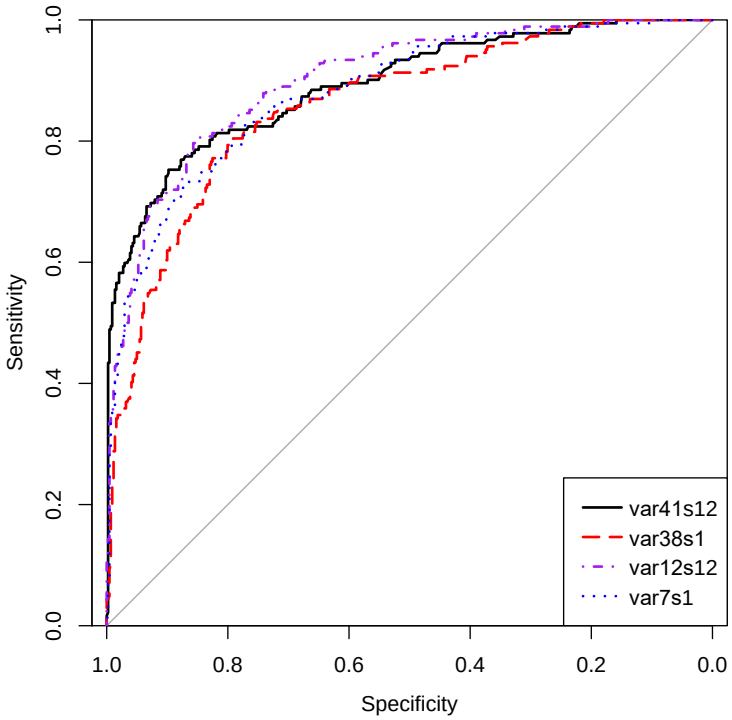


Figura 6.4: Curvas ROC dos modelos de classificação combinada construídas com dados de validação.

se aproxima dos eixos $\texttt{especificidade} = 1$ e $\texttt{sensibilidade} = 1$ é a do modelo `var12s12`, e a que menos se aproxima é a do modelo `var38s1`. Estes dois factos, combinados, corroboram as afirmações anteriores de que o desempenho do modelo de previsão aumenta ligeiramente com a seleção das variáveis de entrada e com a escolha dos dados do 2º semestre para o *dataset*.
As análises e interpretações acabadas de expressar, visaram essencialmente obter, com os dados de validação, uma primeira perceção do comportamento dos modelos, que permitiu de alguma forma validar as soluções encontradas e servirá também de termo de comparação na avaliação da capacidade de generalização dos modelos propostos. Será na secção que se segue que os modelos de previsão serão verdadeiramente postos à prova, com a sua aplicação a dados completamente novos.

6.6.2 Avaliação da capacidade de generalização dos modelos propostos

Com o objetivo de avaliar a sua capacidade de generalização, os modelos de previsão desenvolvidos são agora aplicados aos conjuntos de observações deixados para teste. Tratando-se de dados de anos letivos posteriores aos usados para treino e validação, que nunca foram mostrados a qualquer dos modelos, permitem, de alguma forma, antever em que medida conseguirão esses modelos cumprir a sua missão quando se virem a confrontar com dados completamente novos. Cada um dos quatro modelos desenvolvidos, descritos na secção anterior, foi então usado para prever os abandonos escolares registados nos dados de teste, tendo-se obtido os resulta-

dos expressos na Tabela 6.7, a qual também inclui, para efeitos de comparação, os resultados anteriormente obtidos com os dados de validação. Comparando o AUC médio verificado com os

Tabela 6.7: Aplicação dos modelos encontrados aos conjuntos de observações deixados para teste.

			RF			SVM			ANN			
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

dados de teste com o que se tinha obtido com os dados de validação, constata-se, como já esperado, alguma perda da capacidade preditiva dos modelos, baixando o valor do AUC 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 mostram que os modelos em que se procedeu à seleção de variáveis (modelos `var7s1` e `var12s12`), embora se tenham revelado ligeiramente mais precisos que os outros nos dados de validação, infelizmente são aqueles que apresentam pior eficácia com os dados de teste. Isto 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, já anteriormente enunciadas.

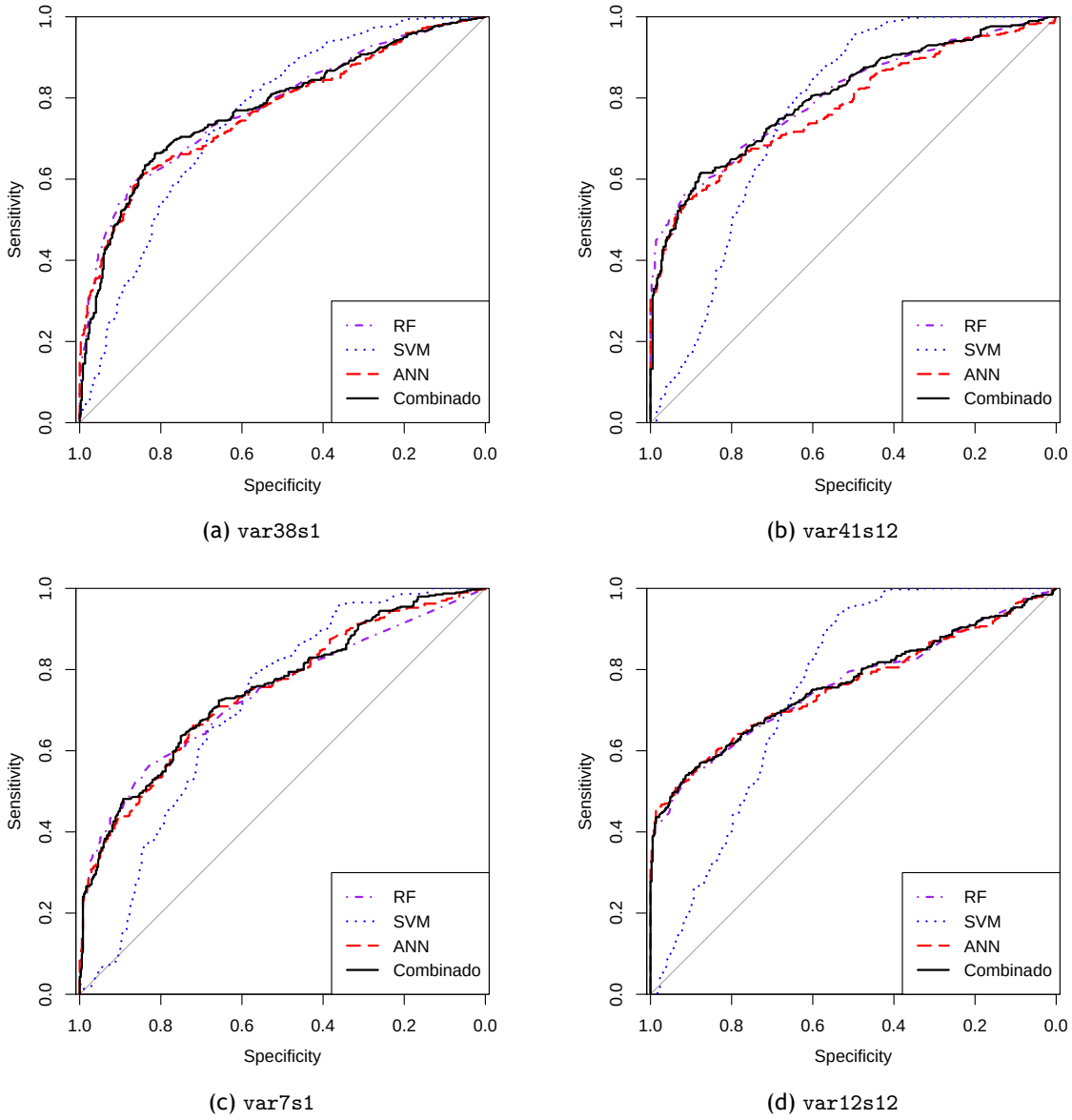
Mais interessante ainda será avaliar o comportamento do modelo combinado, quando exposto aos dados de teste. Na Tabela 6.8 são apresentados os valores AUC obtidos com o modelo combinado, bem como a média e melhores valores AUC obtidos com os 3 algoritmos de classificação que o integram, para efeitos de comparação. Os resultados tabelados permitem, também eles,

Tabela 6.8: Comparação do desempenho do modelo combinado com o desempenho dos 3 algoritmos de classificação que o integram, usando dados de teste.

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 (ANN)	0.7303	0.7495
var12s12	0.7605 (RF)	0.7582	0.7633

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. Este não deixa de ser também ele um aspeto revelador, até porque, como se sabe, não existe um algoritmo de classificação único capaz garantir o melhor desempenho em todos os problemas — a melhor *performance* de um algoritmo dependerá sempre das especificidades quer do problema estudado quer do *dataset* adotado.

A perda de capacidade preditiva dos modelos quando aplicados a dados novos é também percetível graficamente. É, pelo menos, isso que revelam as curvas ROC das Figuras 6.5 e 6.6 (produzidas com dados de teste) quando comparadas, respetivamente, com as das Figuras 6.3 e 6.4 (produzidas com dados de validação). Invariavelmente, as curvas produzidas com dados de teste ficam algo mais distantes dos eixos que caraterizam a condição de otimalidade absoluta ($\texttt{especificidade} = 1$ e $\texttt{sensibilidade} = 1$).



(a) `var38s1` (b) `var41s12`

(c) `var7s1` (d) `var12s12`

Figura 6.5: Curvas ROC dos modelos de classificação construídas com dados de teste.

Tal como os 3 algoritmos de classificação em separado mostraram, também o modelo combinado apresenta um melhor desempenho para dados de teste quando suportado por todas as variáveis explicativas disponíveis (modelos var38s1 e var41s12) — um AUC acima dos 0.77 na previsão dos abandonos escolares ao fim do 1º semestre escolar (modelo var38s1), e um AUC acima dos 0.80 na previsão realizada ao fim do 2º semestre (modelo var41s12).

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 ao fim 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).

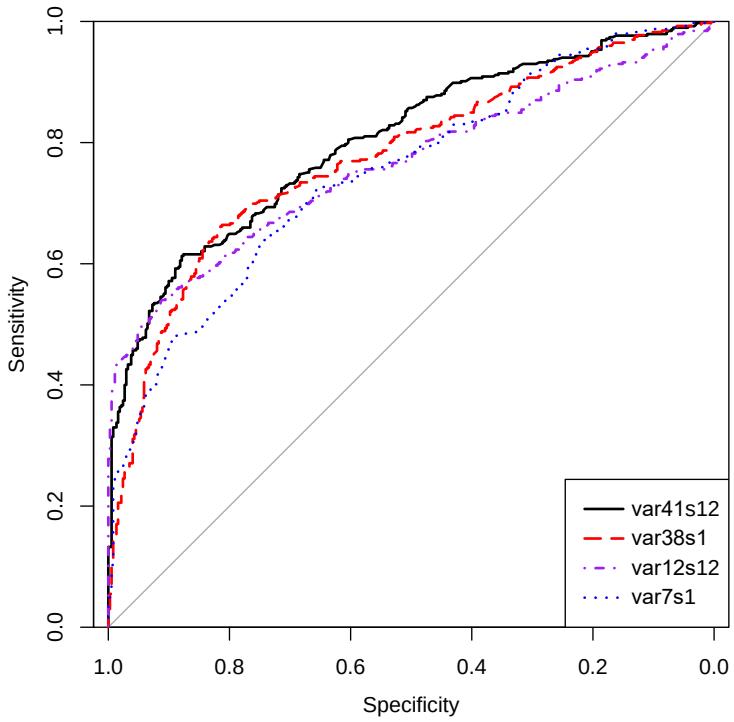


Figura 6.6: Curvas ROC dos modelos de classificação combinada construídas com dados de teste.

6.7 Importância relativa das variáveis explicativas selecionadas

Um interessante estudo adicional que se poderá ainda fazer com os modelos de previsão suportados pelas variáveis mais explicativas, $var7s1$ e $var12s12$, será tentar quantificar, de alguma forma, a importância relativa que cada uma das suas variáveis tem na explicação da variável resposta. Com essa informação adicional, para além de se identificar o subconjunto de fatores que mais influenciam a capacidade preditiva do modelo, conseguir-se-á também estabelecer entre eles uma ordem de relevância em termos de sua importância na previsão. Para esse efeito, propõe-se uma abordagem simples que, recorrendo a uma técnica de análise de sensibilidade, tenta definir a importância da variável medindo a sensibilidade do modelo a essa variável.

A importância relativa de uma variável de entrada na explicação da variável resposta pode então ser vista como sendo a razão, em valores percentuais, entre a perda de acurácia que resulte da não inclusão no modelo dessa variável 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 será o nível de importância dessa mesma variável. Representando por $V(f)$ a medida de acurácia dum dado modelo f e por f_{-i} esse modelo sem a sua i-ésima variável, a importância relativa dessa i-ésima variável poderá ser facilmente expressa pela fórmula:

$$S_i = \frac{V(f) - V(f_{-i})}{V(f)} \times 100. \quad \textbf{(6.1)}$$

No presente estudo, a medida de acurácia $V(\cdot)$ é expressa pelo valor do AUC do modelo combinado, o qual, como já se referiu, assume para a variável alvo a classe mais "votada" pelos 3 algoritmos de classificação que o integram.

A importância relativa das variáveis independentes foi então determinada para os modelos

$var7s1$ e $var12s12$, de acordo com a equação (6.1). Para o cálculo do desempenho do modelo sem a sua i-ésima variável, $V(f_{-i})$, começou-se por reajustar, com os dados de validação, os hyperparâmetros dos 3 algoritmos de classificação sem essa variável, determinando-se depois, com os dados de teste, o valor do AUC do modelo combinado resultante. Para a medida de desempenho $V(f)$, dos modelos completos $var7s1$ e $var12s12$, usaram-se, respetivamente, os valores 0.7495 e 0.7633, anteriormente determinados e já apresentados na Tabela 6.8.
Nas Tabelas 6.9 e 6.10 apresentam-se os resultados obtidos para os dois modelos de previsão, dispostos por ordem decrescente do valor calculado para a importância da variável.

Tabela 6.9: Importância relativa das variáveis explicativas do modelo `var7s1`.

	Comb.	RF		SVM		ANN			
var. retirada	AUC	mtry	AUC	cost	AUC	size	decay	AUC	S_i
ects_reprov_s	0.7249	3	0.7110	2^{-5}	0.7128	10	$10^{-8/3}$	0.7429	3.28
ects_aprov_s	0.7351	2	0.7391	2^{-5}	0.6826	5	$10^{-5/3}$	0.7365	1.92
ects_cred_tx	0.7353	2	0.7167	2^{-1}	0.7057	20	$10^{-5/3}$	0.7370	1.89
media_s	0.7403	3	0.7417	2^{9}	0.6891	10	$10^{-6/3}$	0.7425	1.23
sexo	0.7442	2	0.7306	2^{-5}	0.7047	20	$10^{-6/3}$	0.7453	0.71
cod_escola	0.7481	2	0.7313	2^{-5}	0.7002	5	$10^{-7/3}$	0.7406	0.19
bolseiro_s	0.7673	1	0.7436	2^{-3}	0.7064	20	$10^{-3/3}$	0.7730	-2.37

Tabela 6.10: Importância relativa das variáveis explicativas do modelo `var12s12`.

	Comb.	RF		SVM		ANN			
var. retirada	AUC	mtry	AUC	cost	AUC	size	decay	AUC	S_i
ects_reprov_s	0.7078	3	0.7122	2^{15}	0.6850	5	$10^{-1/3}$	0.6875	7.27
sit_prof_mae	0.7547	3	0.7652	2^{11}	0.7697	1	$10^{0/3}$	0.7520	1.13
idade	0.7573	2	0.7571	2^{13}	0.7489	1	$10^{0/3}$	0.7562	0.79
vd12_s	0.7581	4	0.7568	2^{9}	0.7563	1	$10^{0/3}$	0.7599	0.68
nacionalidade	0.7608	3	0.7684	2^{11}	0.6925	1	$10^{0/3}$	0.7570	0.33
nivel_esc_pai	0.7609	2	0.7634	2^{11}	0.7618	1	$10^{0/3}$	0.7551	0.31
n10_11_acesso	0.7617	2	0.7628	2^{15}	0.7578	1	$10^{0/3}$	0.7591	0.21
sit_prof_aluno	0.7618	2	0.7626	2^{13}	0.7657	1	$10^{0/3}$	0.7627	0.20
media_s	0.7634	4	0.7717	2^{9}	0.7614	1	$10^{0/3}$	0.7626	-0.01
cod_prof_aluno	0.7647	2	0.7627	2^{11}	0.7513	1	$10^{0/3}$	0.7624	-0.18
cod_prof_mae	0.7661	3	0.7639	2^{9}	0.7578	1	$10^{0/3}$	0.7729	-0.37
dir_associativo_s	0.7680	2	0.7692	2^{11}	0.7588	1	$10^{0/3}$	0.7603	-0.62

Os resultados expostos em cada uma das linhas dessas tabelas dizem respeito ao modelo sem a variável que surge na primeira coluna, sendo o da última coluna a importância relativa dessa mesma variável, calculada pela fórmula (6.1), e tal como se acabou de descrever. Os valores AUC de cada um dos algoritmos de classificação que integram o modelo são incluídos nas tabelas a título meramente informativo, pois apenas os valores AUC do modelo combinado (2ª coluna) são usados no cálculo da importância S_i. De recordar ainda que todos os valores AUC tabelados foram calculados com os dados de teste, tendo os valores dos hyperparâmetros sido previamente ajustados com os dados de validação.
O aspeto que, desde logo, sobressai em ambos os modelos, é a existência de variáveis com impacto negativo na previsão do abandono (ver Figuras 6.7 e 6.8). Ainda que não seja o resultado que há partida mais se desejaria, não se deverá cair na tentação de retirar do modelo essas variáveis. Estando em causa a importância que as variáveis revelaram na previsão num determinado conjunto de dados de teste, nada garante que essas mesmas variáveis não venham a ter um comportamento distinto num outro conjunto de teste, aliás, à semelhança do que aconteceu

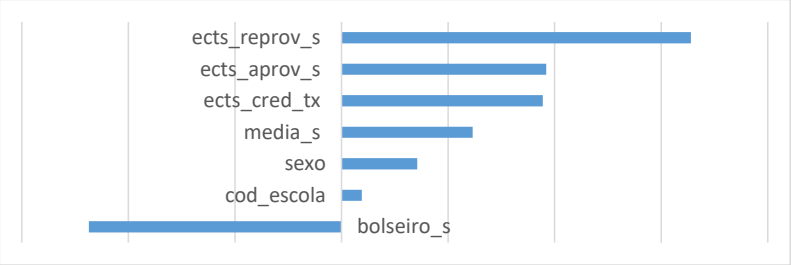


Figura 6.7: Importância relativa das variáveis explicativas do modelo `var7s1`.

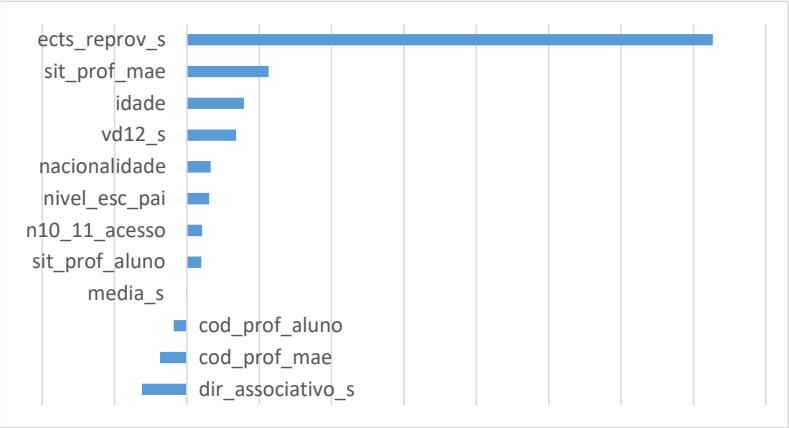


Figura 6.8: Importância relativa das variáveis explicativas do modelo `var12s12`.

com os dados de validação, aquando da sua seleção — foram selecionadas precisamente por revelarem uma contribuição positiva na previsão dos abandonos nos dados de validação. Retirar-se variáveis nesta fase significaria estar-se, 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 6.8, 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, tal como se depreende da Figura 6.7. Repare-se que a variável `media_s`, ao não contabilizar a avaliação de unidades curriculares em que o aluno tenha reprovado, não dará uma indicação tão consistente do desempenho curricular do mesmo, como aquela que se retira das variáveis `ects_aprov_s` e `ects_reprov_s`. Daí surgir apenas na 4ª posição do primeiro modelo e apresentar mesmo uma contribuição negligenciável no segundo.

Será interessante perceber-se por que razão a variável predominante no segundo modelo não surge com o mesmo destaque no primeiro. Na verdade, os resultados curriculares ao fim do 1º semestre assumem, de certa forma, um caráter intercalar, não definitivo. São os resultados curriculares ao fim do 2º semestre que, ao corresponderem a resultados finais do ano letivo, traduzem um balanço real daquele que foi o verdadeiro desempenho do aluno, com consequências

para o seu percurso escolar imediato. À luz desse entendimento, não será difícil aceitar que, no final do 1º semestre, o 'número de ECTS reprovados' não seja muito mais informativo do que, por exemplo, o 'número de ECTS aprovados', quanto ao real sucesso do aluno. Repare-se que, tendo o aluno já conseguido aprovação num número específico de ECTS, esse será certamente um indicador de sucesso incontornável. Já o facto de ter reprovado num número específico de ECTS pode não vir a traduzir-se necessariamente em insucesso, uma vez que, não raras as vezes, o aluno, alarmado com os resultados do 1º semestre, consegue vir ainda a recuperar a sua *performance* até ao final do ano letivo, empenhando-se nas unidades curriculares do 2º semestre ou, até mesmo, na recuperação das unidades curriculares do 1º semestre, recorrendo a épocas de avaliação de recurso ou especiais. No final do 2º semestre, o 'número de ECTS reprovados' será muito mais informativo, uma vez que, estando-se já perante resultados finais, traduz um indicador de insucesso incontornável. Já para o 'número de ECTS aprovados' é possível identificar, pelo menos, uma situação em que o mesmo não dará uma indicação válida de desempenho: a de um aluno que, tendo conseguido creditação para todas as unidades curriculares do 1º ano, acaba por não ter qualquer ECTS aprovado ao fim do 2º semestre. Mesmo não tendo ECTS aprovados, não se pode, portanto, concluir que tenha tido um mau desempenho. A razão para não constar nas 12 variáveis do modelo $var12s12$ uma variável tão "importante" como a `ects_aprov_s`, dever-se-á, muito provavelmente, ao facto de a mesma apresentar um elevado grau de correlação com a variável `ects_reprov_s`, já incluída no modelo.
Não se desenvolvem considerações adicionais, que reflitam o comportamento das restantes variáveis dos modelos, por se considerar pouco representativo os resultados que com elas se obtiveram. A componente estocástica que caracteriza cada um dos algoritmos de classificação do modelo combinado, só por si, é responsável por uma variabilidade considerável em todas essas variáveis. Refira-se apenas que correram-se várias vezes os modelos, sempre com a mesma configuração e com os mesmos *datasets*, mas não se fixando o valor de partida (*seed*) da função geradora de números (pseudo) aleatórios, tendo os resultados evidenciado consistência apenas nos aspetos que foram já objeto de reflexão. Invariavelmente, surgem em ambos os modelos variáveis com contribuição negativa; confirma-se a grande predominância da variável `ects_reprov_s` no segundo modelo e o posicionamento das variáveis `ects_reprov_s` e `bolseiro_s` como a mais e a menos explicativas no primeiro modelo de previsão, mantendo-se sempre negativa a contribuição desta última variável.

6.8 Conclusões e discussão de resultados

Com o objetivo de se poder identificar, de forma precoce, quais os alunos das licenciaturas do IPB que apresentam maior propensão para o abandono académico, desenvolveu-se um modelo de previsão suportado pelos dados recolhidos ao fim do 1º semestre escolar do aluno e um outro suportado pelos dados acumulados até ao fim do seu 2º semestre escolar. Trata-se, portanto, de dois modelos preditivos distintos, cuja capacidade de previsão é aferida em diferentes instantes do percurso académico do estudante: o primeiro logo no final do 1º semestre e o segundo no final do primeiro ano do curso de graduação.
Apresentou-se como solução, um modelo combinado, que tira partido de três importantes técnicas de *data mining*, como são os algoritmos de classificação *random forest*, *support vector machines* e as redes neuronais artificiais. Ainda que composto por três algoritmos independentes, conseguiu-se encontrar um esquema, baseado no método *forward search*, com o qual foi possível chegar a um único subconjunto de variáveis, consideradas mais explicativas, para su-

porte do modelo. De acordo com essa abordagem, são 7 as variáveis que, ao fim do 1º semestre escolar, se revelam mais explicativas do abandono e 12 as variáveis mais explicativas ao fim do 2º semestre.

Depois de encontradas as variáveis mais explicativas, conseguiu-se estabelecer entre elas uma ordem de relevância na previsão, medindo-se o seu impacto na explicação da variável alvo, através de uma técnica simples de análise de sensibilidade. Por via dessa análise, 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.

Como se teve como objetivo, nesta parte do trabalho, explorar as múltiplas dimensionalidades indissociáveis do aluno, que pudessem explicar o seu abandono, optou-se por expurgar o *dataset* dos seus dados omissos, excluindo dois importantes conjuntos de alunos: os que não entraram pelo concurso nacional de acesso e os que, por qualquer razão, não preencheram o inquérito no ato de matrícula. Com esta opção, manteve-se, como se pretendia, o conjunto completo de variáveis explicativas disponíveis, mas acabou por se reduzir a cerca de $1/3$ a dimensão dos *datasets* iniciais. Assim, uma outra opção interessante, a ser considerada em trabalho futuro, seria manter todos os alunos e retirar do *dataset* as 2 categorias de variáveis com dados omissos — a socioeconómica e a de acesso. Dito de outra forma, vendo os *datasets* como tabelas, a ideia seria retirar as suas células vazias eliminando as colunas em vez de se eliminarem as linhas que contêm essas células. Só por essa via será possível estender o modelo de previsão desenvolvido a todos os alunos do IPB.

Como trabalho futuro, será também pertinente investigar a aplicabilidade, no problema de abandono agora tratado, do método de seleção variáveis que foi usado no estudo de previsão do sucesso académico global desenvolvido no capítulo anterior, e que passava por selecionar, numa primeira fase, categorias de variáveis e só depois, numa segunda fase, variáveis individuais. Será interessante comparar os resultados obtidos por essa via com os que foram agora alcançados com a aplicação do método convencional *forward search* (adaptado para o modelo combinado).

Uma vez que a investigação científica tem revelado que o absentismo às aulas é um forte prenúncio do abandono, será igualmente pertinente averiguar qual a importância da assiduidade dos estudantes do IPB para a explicação de abandono nesta academia. Para o efeito, sugere-se ao IPB a recolha sistematizada dos dados que venham a permitir essa análise.

De acordo com os dois modelos desenvolvidos, verificou-se que a contribuição de cada uma das cinco dimensões em análise, para a explicação do abandono no IPB, é significativamente diferenciada entre si. Tal como anteriormente referido, são principalmente dois fatores da dimensão curricular do estudante que explicam o abandono académico. 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 baixo rendimento académico de aprendizagem. Uma vez que os fatores que melhor prenunciam a decisão dos estudantes se desvincularem do IPB estão relacionados com o seu desempenho curricular, sugere-se, por exemplo, a atribuição de um professor conselheiro aos alunos que o modelo preveja abandonarem ou, no caso de não ser usado o modelo de previsão desenvolvido, aos que apresentem piores resultados nos ECTS aprovados/reprovados, a fim de lhes serem incutidas aspirações académicas e de lhes ser providenciado um acompanhamento tutorial personalizado.

É relevante notar que do conjunto dos 7 fatores do modelo $var7s1$, que se revelaram significati-

vos para a previsão do abandono no final do 1º semestre, apenas o 'género' não está relacionado com o contexto académico do estudante. Já para o modelo $var12s12$ verifica-se que apenas a dimensão matricula não está representada no conjunto das 12 variáveis que se revelaram mais informativas, assumindo a dimensão curricular do aluno, por via da sua variável ECTS reprovados, uma relevância incomparável com as restantes.

É interessante assinalar que do conjunto de fatores que caracterizam os estudantes relativamente ao historial de desempenho académico pré-ingresso no ensino superior (dimensão Acesso), apenas a média das classificações obtidas no 10º e 11º anos de escolaridade revelou algum poder explicativo para a previsão de abandono no final do 2º semestre. Embora a literatura mencione que a média de acesso ao ensino superior é um forte preditor de sucesso académico no ensino superior (Miguéis et al. [71], Aluko et al. [3]), o motivo de não ter sido capturada pelos modelos apresentados, poder-se-à justificar pelo facto da média de acesso da grande maioria dos estudantes do IPB apresentar uma baixa dispersão entre si.

De acordo com o modelo $var12s12$, a dimensão socioeconómica dos estudantes é a segunda mais bem representada na explicação do abandono académico. Cinco fatores, dos oito analisados nessa dimensão, também parecem contribuir de alguma forma para a explicação do fenómeno do abandono no IPB, designadamente, a profissão e respetiva situação de empregabilidade quer da mãe quer do próprio aluno, e o nível de escolaridade do pai. Esta conclusão sugere que o rendimento económico familiar poderá influenciar na decisão de abandono.

Em relação à dimensão demográfica do estudante, o género é o único fator a ter em conta aquando da definição de ações preventivas de combate ao abandono durante o 1º semestre, enquanto que no fim do 2º semestre os fatores que se revelaram significativos, nesta dimensão, foram a nacionalidade e a idade dos estudantes.

A principal contribuição do estudo apresentado neste capítulo para a literatura de EDM está relacionada com a originalidade do método proposto e com o elevado desempenho dos modelos obtidos por essa via.

Refira-se, por fim, que os resultados e consequentes reflexões que se acabaram de apresentar não deverão ser vistos com excessiva rigidez, atendendo à componente estocástica que caracteriza qualquer um dos algoritmos envolvidos e aos exíguos diferenciais que por vezes estiveram na base das decisões que ditaram o rumo da otimização. De facto, durante o processo de seleção das variáveis mais explicativas, algumas das escolhidas apresentaram um incremento preditivo muito pouco diferenciado das restantes. Repare-se, a título de exemplo, que a variável `cod_escola`, do modelo do 1º semestre, foi escolhida em detrimento da variável `cod_prof_mae`, com base nos seus valores quase indiferenciáveis de AUC médio, respetivamente 0.8453 e 0.8452 (ver Tabela E.1d incluída no Apêndice E desta tese). Também no modelo do 2º semestre, por exemplo, `cod_escola` foi escolhida em detrimento de `cod_prof_mae`, com base nos valores 0.8899 e 0.8898, respetivamente (ver Tabela E.2f do Apêndice E). Assim, não será de estranhar que a simples alteração das observações do *dataset*, ou até mesmo a própria natureza estocástica dos algoritmos, leve a que seja outra a variável escolhida numa dada iteração da *forward search* e, podendo com isso, influenciar, inclusivamente, toda a seleção das variáveis subsequentes.

Capítulo 7

Conclusões finais

Nesta tese exploraram-se algumas das atuais técnicas de *data mining* no contexto educacional, a fim de desenvolver um conjunto de modelos de previsão de desempenho escolar, que permitem estimar, com a devida antecedência, quer o nível de sucesso final, quer a propensão para o abandono, dos alunos das licenciaturas do Instituto Politécnico de Bragança (IPB), uma instituição de ensino superior politécnico do interior do país. Na construção desses modelos, usaram-se três dos algoritmos mais populares entre a comunidade científica de *data mining*, como são as *random forest*, as *support vector machines* e as redes neuronais artificiais. Mostrou-se como estes algoritmos podem revelar-se particularmente adequados na produção de conhecimento que suporte melhores decisões de gestão numa IES.
Tendo sido já apresentadas, ao longo da tese, conclusões parcelares para cada um dos estudos realizados, recuperam-se agora apenas aqueles que se pensa serem os principais resultados alcançados e apontam-se, numa subsecção seguinte, que outras investigações se perspetivam para um futuro próximo.

A fase de pré-processamento apresentou-se, desde logo, como uma etapa bastante desafiadora deste estudo, dada a dimensão e especificidades do conjunto de registos de dados com que se teve que lidar no desenvolvimento dos modelos de previsão. Tratou-se de um *dataset* 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 do IPB e onde cada estudante é caracterizado por cerca de meia centena de variáveis explicativas.

No estudo descrito no Capítulo 5 desta tese propôs-se um modelo analítico de regressão, baseado no algoritmo *random forest*, desenvolvido com o objetivo de prever, de forma precoce, o desempenho académico global dos estudantes de licenciatura do IPB. Ao invés de se seguir o procedimento normalmente adotado de se delimitar a previsão a um só curso específico, o modelo foi desenvolvido envolvendo a mais de meia centena de licenciaturas do IPB, que cobrem as mais diversas áreas educacionais. Desta forma é possível dotar a instituição de uma ferramenta única, capaz de acomodar a diversidade das dinâmicas educativas e a heterogeneidade do universo dos estudantes presentes na academia.
A seleção de caraterísticas que melhor explicam o sucesso do aluno processou-se em duas etapas. Começando-se por aplicar o algoritmo de *data mining random forest* a diferentes combinações de categorias de variáveis, constatou-se que quer os dados demográficos, quer os socioeconómicos, quer mesmo os de acesso ao ensino superior em pouco ou nada contribuem para a capacidade do modelo preditivo, conseguindo-se, assim, reduzir a quase metade o número de variáveis preditivas. Percebeu-se, dessa forma, a grande influência que têm os dados curri-

culares na eficácia do modelo, mesmo que se limitem a refletir unicamente resultados do 1º semestre escolar do aluno. É esta particularidade que abre boas perspetivas para que a previsão de sucesso do aluno se possa vir a efetuar numa fase ainda precoce do seu percurso escolar. Numa segunda etapa do processo de seleção de variáveis, e usando apenas os resultados do 1º semestre escolar do aluno, foi possível reduzir ainda mais o número de variáveis que suportam o modelo de previsão, num ajuste mais minucioso do modelo de previsão que se traduziu na exclusão individual de variáveis de importância considerada negligenciável.
A abordagem seguida permitiu isolar 11 variáveis explicativas, a partir de um conjunto inicial de cerca de meia centena, que só por si justificam a capacidade preditiva do modelo, e que se poderão vir a revelar fundamentais na definição de estratégias de gestão mais assertivas centradas na promoção do sucesso académico. O modelo desenvolvido permitiu assim identificar os fatores de (in)sucesso dos estudantes, permitindo, nomeadamente, perceber que os fatores do contexto curricular de desempenho académico do estudante são determinantes para a previsão pretendida, o que confirma resultados já anteriormente demonstrados na literatura. Um desses fatores com influência no sucesso do aluno é o tipo de escola, considerado pela primeira vez na literatura de EDM. 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.
O tipo de abordagem adotada na identificação das características do estudante que melhor explicam o seu sucesso parece distanciar-se daquela que usualmente tem sido seguida em trabalhos relacionados com a mesma temática. No caso do presente trabalho, num primeiro ajuste do modelo, começou-se por selecionar as dimensões do aluno que melhor explicam o seu sucesso, eliminando-se, com isso, grupos completos de variáveis. Com esta abordagem foi então possível, logo numa primeira fase, reduzir a elevada dimensionalidade dos dados, sem se ter perdido a capacidade preditiva do modelo. A eliminação de dimensões completas do aluno reveste-se de particular importância, uma vez que contribui para a redução do número de categorias às quais pertencem os fatores explicativos, baixando-se, com isso, alguma da complexidade do processo de previsão e, mais importante, tornando possível estender o estudo a outros contextos, onde nem todas as categorias de variáveis inicialmente consideradas neste estudo estejam presentes.

Também 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, no estudo descrito no Capítulo 6, dois modelos de previsão distintos, um suportado pelos dados recolhidos ao fim 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 *data mining*: os algoritmos de classificação *random forest*, as *support vector machines* e as redes neuronais artificiais. 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, ao fim 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 sobre as variáveis selecionadas, 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 escolar baixo rendimento académico. Sugere-se, por exemplo, a atribuição de um professor conselheiro aos alunos que o modelo preveja abandonarem ou, no caso de não ser usado o modelo de previsão desenvolvido, aos que apresentem piores resultados nos ECTS aprovados/reprovados, a fim de lhes serem incutidas aspirações académicas e de lhes ser providenciado um acompanhamento tutorial personalizado. 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.
Do conjunto de fatores que caracterizam os estudantes relativamente ao historial de desempenho académico pré-ingresso no ensino superior, apenas a média das classificações obtidas no 10º e 11º anos de escolaridade revelou algum poder explicativo para a previsão de abandono no final do 2º semestre. Embora a literatura mencione que, por exemplo, a média de acesso ao ensino superior é um forte preditor de sucesso académico no ensino superior, o motivo de não ter sido capturada pelos modelos desenvolvidos, poder-se-à justificar pelo facto da média de acesso dos estudantes do IPB apresentar uma baixa dispersão entre si.

Conclui-se com algumas considerações finais e resumindo aqueles que foram, de uma forma geral, os principais resultados obtidos e as principais contribuições para a literatura de EDM.
Através do conhecimento obtido, quer no estudo de previsão de abandono quer no de previsão de sucesso académico, é possível identificar grupos de estudantes de maior risco, o qual permitirá aos gestores institucionais proceder à definição orientada de estratégias educacionais e tutoriais em prol da eficácia e da eficiência educativa. Os resultados do estudo de previsão de abandono, em particular, revelam-se de capital importância uma vez que, no caso concreto do IPB, os índices de abandono são muito preocupantes, chegando a cerca de 40% os alunos que acabam por não concluir a sua licenciatura. Esse conhecimento será fundamental para a delineação de medidas preventivas urgentes, precoces e precisas, que levem à diminuição dos índices de evasão escolar.
Para além do elevado desempenho evidenciado pelos modelos desenvolvidos e da própria dimensão e diversidade dos dados analisados, destaca-se o carácter inovador do processo de seleção de variáveis realizado no estudo de previsão de sucesso escolar e a originalidade do método combinado proposto para o modelo de previsão de abandono.
Esta tese também demonstra o potencial das técnicas *data mining* quando aplicadas a grandes bases de dados do contexto educacional, podendo mesmo, a metodologia apresentada, servir de guia às IES que pretendam extrair conhecimento dos grandes conjuntos de dados provenientes dos seus processos de ensino e de aprendizagem.
Embora a metodologia apresentada de aplicação de técnicas de *data mining* possa vir a ser replicada no âmbito de outras instituições de ensino superior — sendo esse claramente também um dos objetivos da presente investigação —, uma parte importante dos resultados que se obtiveram neste estudo não é generalizável ao contexto geral do ensino superior, uma vez que se refere a uma realidade muito específica, não representativa 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 con-

segue captar a mesma heterogeneidade de alunos das instituições de grandes centros urbanos do litoral. Por exemplo, no estudo de previsão de sucesso excluíram-se duas categorias de variáveis explicativas, as demográficas e as de acesso. Não será de estranhar que, em instituições com outra capacidade de captação de alunos, e onde heterogeneidade dos mesmos possa ser bem mais evidente, essas variáveis se venham a revelar fatores importantes para o sucesso. Quanto muito, os resultados apresentados poderão espelhar realidades de instituições de ensino superior que reúnam condições similares às do IPB, como será o caso de outros institutos politécnicos do interior do país, localizados longe dos grandes centros urbanos.
Como nota final, refira-se que os resultados e consequentes reflexões que se foram apresentando nesta tese não devem ser vistos com excessiva rigidez, atendendo à variabilidade das caraterísticas dos *datasets* e até mesmo à componente estocástica que caracteriza os algoritmos de *data mining* usados, associada aos exíguos diferenciais que por vezes estiveram na base das decisões que ditaram o rumo da otimização dos modelos.

7.1 Perspetivas de trabalho futuro

Nesta investigação foram analisados, usando técnicas de *data mining*, os dados que caracterizam os estudantes das licenciaturas do Instituto Politécnico de Bragança, nas suas diferentes dimensões, com o objetivo de se desenvolverem modelos analíticos de apoio à gestão. Um trabalho que se tentará desenvolver no futuro será, naturalmente, replicar estes mesmos estudos nos cursos de mestrado e de técnicos superiores profissionais (CTeSP), os outros dois ciclos de estudos que integram a oferta formativa do IPB.
De modo a complementar a investigação descrita no presente estudo, poder-se-á também 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, que venham a permitir a adoção de estratégias de promoção educacional individualizadas para cada escola e para cada curso.
Uma vez que se adotaram diferentes abordagens no estudo de previsão de sucesso académico e no estudo de previsão de abandono escolar, sugere-se, para trabalho futuro, o desenvolvimento de investigações adicionais que permitam perceber o nível de desempenho que se consegue alcançar caso se troquem as abordagens entre esses dois estudos. Mais especificamente, no caso da previsão de sucesso, em vez do modelo a desenvolver se basear apenas no algoritmo *random forest*, será interessante o mesmo vir a combinar a integração de vários métodos de *data mining*, à semelhança do que já se fez na previsão de abandono. Mas não menos interessante será perceber-se até que ponto a abordagem original que se adotou na seleção das variáveis (*feature selection*) da previsão de sucesso — que começa por encontrar a melhor combinação de categorias de variáveis e termina excluindo individualmente as menos explicativas — não dará também ela bons resultados na seleção das variáveis da previsão de abandono escolar, quando comparada com o *forward search* ou com outros métodos de seleção convencionais.
Neste trabalho, os dados omissos foram excluídos retirando do *dataset* original todas as observações (matrículas) que não contivessem informação completa para todas as dimensionalidades do aluno (atributos). Com essa opção acabaram por ser excluídos do estudo de previsão de abandono dois importantes subgrupos de alunos: os que não entraram pelo concurso nacional de acesso — pois não dispõem de dados de acesso — e os que, por qualquer razão, não preencheram o inquérito no ato de matrícula — e que por isso não dispunham de dados socioeconómicos. Dessa forma, manteve-se, como se pretendia, o conjunto completo de variáveis explicativas dis-

poníveis, mas acabou por se reduzir a cerca de $1/3$ a dimensão do *dataset* inicial, como referido na Secção 6.8. Assim, uma outra opção interessante, a ser considerada em trabalho futuro, será manter todos os alunos e, em contrapartida, retirar do *dataset* as duas categorias de variáveis com dados omissos: a socioeconómica e a de acesso. Ou seja, vendo os *datasets* como tabelas, a ideia será retirar as suas células vazias eliminando as colunas em vez de se eliminarem as linhas que contêm essas células. Só por essa via será possível estender o modelo de previsão de abandono desenvolvido a todos os alunos de licenciatura do IPB. Refira-se, por fim, que essa questão não se coloca, pelo menos com a mesma pertinência, em relação à previsão de sucesso, uma vez que o próprio processo de seleção de variáveis adotado nesse estudo acabou por excluir automaticamente essas duas categorias problemáticas de variáveis, viabilizando a aplicação do respetivo modelo de previsão a todos os alunos do *dataset*, tal como se veio a fazer na Secção 5.5.3.

Uma vez que a investigação científica tem revelado que o absentismo às aulas é um forte prenúncio do abandono, será igualmente pertinente averiguar qual a importância da assiduidade dos estudantes do IPB para a explicação de abandono nesta academia. Para o efeito, sugere-se ao IPB a recolha sistematizada dos dados que venham a permitir essa análise. Considerar-se-á também, em pesquisas futuras, o uso de outros fatores que ajudem a explicar o desempenho do aluno e que não foram considerados neste estudo, como, por exemplo, condições de saúde, hábitos de sono, estado civil parental, círculo de amigos e redes sociais.

Naturalmente, este estudo poderá ainda ser ampliado, quer considerando outras fontes de dados, que caracterizem os estudantes de IES com outras especificidades, quer usando outras técnicas de *data mining*, como será o caso dos métodos de associação, das técnicas de *clustering* e das técnicas de visualização.

O desenvolvimento de uma aplicação informática *user-friendly*, que venha a permitir, no futuro, automatizar a aplicação integrada dos modelos de previsão propostos nesta tese, será o desfecho natural e esperado que complementará o trabalho agora apresentado.

7.2 Consignação

Os principais resultados emanados desta tese foram já objeto de divulgação e validação junto da comunidade científica. Mais concretamente, os resultados do estudo de previsão do sucesso escolar global do aluno, de previsão de abandono escolar, bem como um resumo da revisão de literatura efetuada no âmbito desta tese, foram apresentados em conferências internacionais de referência[1] e publicados nas respetivas atas [66, 67] e num capítulo de livro [68]:

- em junho de 2018, *A data mining approach to predict undergraduate students' performance*, "2018 13th Iberian Conference on Information Systems and Technologies (CISTI)", Cáceres, Espanha [66];
- em junho de 2018, *Educational data mining: A literature review*, "2018 13th Iberian Conference on Information Systems and Technologies (CISTI)", Cáceres, Espanha [67];
- em fevereiro de 2019, *A data mining approach for predicting academic success – a case study*, "International Conference on Information Technology & Systems (ICITS19)", Quito, Equador [68].

[1] Indexadas à ISI Web of Knowledge e à SCOPUS.

- em fevereiro de 2020, *Previsão do abandono académico numa instituição de ensino superior com recurso a data mining*, "International Conference on Information Technology & Systems (ICITS20)", Bogotá, Colombia. (Aceite para publicação na RISTI[2])

[2]Revista Ibérica de Sistemas e Tecnologias de Informação.

Bibliografia

[1] Abdulmohsen Algarni. Data mining in education. *International Journal of Advanced Computer Science and Applications*, 7:456–461, 2016. 13, 38, 43

[2] Bruno Miguel Rocha Almeida. *Gapar ou não gapar, importa explorar e apoiar! A influência de atividades exploratórias e de suporte na transição e adaptação ao ensino superior*. PhD thesis, Universidade de Lisboa, 2017. 3

[3] Ralph Olusola Aluko, Olumide Afolarin Adenuga, Patricia Omega Kukoyi, Aliu Adebayo Soyingbe, and Joseph Oyewale Oyedeji. Predicting the academic success of architecture students by pre-enrolment requirement: using machine-learning techniques. *Construction Economics and Building*, 16(4):86–98, 2016. 16, 17, 54, 140

[4] Elaf Abu Amrieh, Thair Hamtini, and Ibrahim Aljarah. Mining educational data to predict student's academic performance using ensemble methods. *International Journal of Database Theory and Application*, 9(8):119–136, 2016. 53, 54, 91

[5] Tânia Daniela da Silva Araújo. *O abandono escolar no ensino superior: trajetos e projetos: uma análise sociológica*. PhD thesis, Universidade do Minho, 2018. 3, 4

[6] Raheela Asif, Agathe Merceron, Syed Abbas Ali, and Najmi Ghani Haider. Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, 113:177–194, 2017. 56

[7] RSJD Baker et al. Data mining for education. *International encyclopedia of education*, 7 (3):112–118, 2010. 5, 15, 16, 17, 37

[8] Ryan Shaun Baker and Paul Salvador Inventado. Educational data mining and learning analytics. In *Learning analytics*, pages 61–75. Springer, 2014. 5

[9] Ryan Shaun Baker, Albert T Corbett, and Kenneth R Koedinger. Detecting student misuse of intelligent tutoring systems. In *International Conference on Intelligent Tutoring Systems*, pages 531–540. Springer, 2004. 45

[10] Ryan SJD Baker and Kalina Yacef. The state of educational data mining in 2009: A review and future visions. *JEDM-Journal of Educational Data Mining*, 1(1):3–17, 2009. 38, 43, 44, 45, 46, 157

[11] Behdad Bakhshinategh, Osmar R Zaiane, Samira ElAtia, and Donald Ipperciel. Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies*, 23(1):537–553, 2018. 39, 40, 43

[12] Rui Banha. Promoção do sucesso dos alunos nas instituições de ensino superior em portugal: medidas observadas nos respetivos sítios da internet. Direção-Geral de Estatísticas da Educação e Ciência (DGEEC), Junho 2017. 1

[13] Pedro Belo. Avaliação das expetativas e das vivências académicas na transição para o ensino superior. *Revista Portuguesa de Pedagogia*, pages 95–113, 2016. 4

[14] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 19, 20, 21, 78, 91, 125

[15] Concepción Burgos, María L Campanario, David de la Peña, Juan A Lara, David Lizcano, and María A Martínez. Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering*, 2017. 63, 115

[16] Hana Bydžovská. A comparative analysis of techniques for predicting student performance. In *Proceedings of the 9th International Conference on Educational Data Mining 2016*, 2016. 57

[17] Peter Cabena, Pablo Hadjinian, Rolf Stadler, Jaap Verhees, and Alessandro Zanasi. *Discovering data mining: from concept to implementation*. Prentice-Hall, Inc., 1998. 10, 13

[18] Cássio Oliveira Camilo and João Carlos da Silva. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. *Universidade Federal de Goiás (UFC)*, pages 1–29, 2009. 43

[19] Renza Campagni, Donatella Merlini, Renzo Sprugnoli, and Maria Cecilia Verri. Data mining models for student careers. *Expert Systems with Applications*, 42(13):5508–5521, 2015. 53

[20] Félix Castro, Alfredo Vellido, Àngela Nebot, and Francisco Mugica. Applying data mining techniques to e-learning problems. In *Evolution of teaching and learning paradigms in intelligent environment*, pages 183–221. Springer, 2007. 46

[21] Rebeca Cerezo, Miguel Sánchez-Santillán, M Puerto Paule-Ruiz, and J Carlos Núñez. Students' lms interaction patterns and their relationship with achievement: A case study in higher education. *Computers & Education*, 96:42–54, 2016. 55

[22] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011. 126

[23] Paulo Cortez and José Neves. Redes neuronais artificiais. Escola de Engenharia, Universidade do Minho, 2000. 21, 25, 26

[24] António Firmino da Costa, João Teixeira Lopes, and Ana Caetano. Percursos de estudantes no ensino superior. fatores e processos de sucesso e insucesso. *Lisboa: Mundos Sociais*, 2014. 3

[25] Evandro Costa, Ryan SJd Baker, Lucas Amorim, Jonathas Magalhães, and Tarsis Marinho. Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. *Jornada de Atualização em Informática na Educação*, 1(1):1–29, 2013. 17

[26] Evandro B Costa, Baldoino Fonseca, Marcelo Almeida Santana, Fabrísia Ferreira de Araújo, and Joilson Rego. 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, 2017. 17, 59

[27] Sérgio da Costa Côrtes, Rosa Maria Porcaro, and Sérgio Lifschitz. *Mineração de dados-funcionalidades, técnicas e abordagens*. PUC, 2002. 13

[28] Luis de Marcos, Eva García-López, Antonio García-Cabot, José-Amelio Medina-Merodio, Adrián Domínguez, José-Javier Martínez-Herráiz, and Teresa Diez-Folledo. Social network analysis of a gamified e-learning course: Small-world phenomenon and network metrics as predictors of academic performance. *Computers in Human Behavior*, 60:312–321, 2016. 64

[29] Gerben Dekker, Mykola Pechenizkiy, and Jan Vleeshouwers. Predicting students drop out: A case study. In *Educational Data Mining 2009*, 2009. 17

[30] César A Del Río and Julio A Pineda Insuasti. Predicting academic performance in traditional environments at higher-education institutions using data mining: A review. *Ecos de la Academia*, 2016(7), 2016. 17, 65, 66, 68, 77, 161

[31] Dursun Delen. A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, 49(4):498–506, 2010. 16, 62, 91

[32] Gayathri Elakia and Naren J Aarthi. Application of data mining in educational database for predicting behavioural patterns of the students. *Elakia et al,/(IJCSIT) International Journal of Computer Science and Information Technologies*, 5(3):4649–4652, 2014. 67

[33] P Engrácia and JO Baptista. Percursos no ensino superior: Situação após quatro anos dos alunos inscritos em licenciaturas de três anos. *Direção-Geral de Estatísticas da Educação e Ciência. Lisboa*, 2018. 115

[34] Gustavo Silva Évora. Sucesso escolar nos alunos de origem cabo-verdiana: o caso dos alunos que ingressam no ensino superior. Faculdade de Ciências Sociais e Humanas, Universidade Nova de Lisboa, 2013. 3

[35] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996. 9, 10, 12, 16

[36] Usama M Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy. *Advances in knowledge discovery and data mining*, volume 21. AAAI press Menlo Park, 1996. 12

[37] Filomena Ferreira and Preciosa Fernandes. Fatores que influenciam o abandono no ensino superior e iniciativas para a sua prevenção: O olhar de estudantes. *Educação, Sociedade & Culturas*, 45:177–197, 2015. 3

[38] J. Brites Ferreira, Graça Maria Seco, Fernando Canastra, Isabel Simões Dias, and Maria Odilia Abreu. (in)sucesso académico no ensino superior: Conceitos, factores e estratégias de intervenção. *Revista Revista Iberoamericana de Educación Superior*, II, 4:28–40, 2011. ISSN ISSN 2007-2872. URL `www.redalyc.org/articulo.oa?id=299124247002`. 3

[39] Organization for Economic Cooperation and Development. *Education at a glance 2017: OECD indicators*. OECD, 2017. 1

[40] William J Frawley, Gregory Piatetsky-Shapiro, and Christopher J Matheus. Knowledge discovery in databases: An overview. *AI magazine*, 13(3):57, 1992. 10, 12

[41] Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Can-

dan, and Tyler Hunt. *caret: Classification and Regression Training*, 2017. URL `https://CRAN.R-project.org/package=caret`. R package version 6.0-76. 78

[42] João Gama, André Carlos Ponce de Leon Carvalho, Katti Faceli, Ana Carolina Lorena, Márcia Oliveira, et al. *Extração de conhecimento de dados: data mining*. Edições Sílabo, Lisbon, Portugal, 2nd edition, 2015. 25

[43] Geraldine Gray, Colm McGuinness, and Philip Owende. An application of classification models to predict learner progression in tertiary education. In *Advance Computing Conference (IACC), 2014 IEEE International*, pages 549–554. IEEE, 2014. 17, 68

[44] Sandra Carina Machado Guimarães. *Estudar e Aprender no Ensino Superior: A Experiência do Aluno Novel de Engenharia Informática*. PhD thesis, Universidade da Beira Interior, Covilhã, Portugal, 2016. 3, 4

[45] Wilhelmiina Hämäläinen and Mikko Vinni. Classifiers for educational data mining. *Handbook of Educational Data Mining*, pages 57–74, 2010. 17, 18

[46] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011. 12, 26

[47] David J Hand, Heikki Mannila, and Padhraic Smyth. *Principles of data mining*. MIT press, 2001. 12

[48] Simon Haykin and Neural Network. A comprehensive foundation. *Neural networks*, 2 (2004):41, 2004. 26

[49] Anne-Sophie Hoffait and Michael Schyns. Early detection of university students with potential difficulties. *Decision Support Systems*, 101:1–11, 2017. 17, 56

[50] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982. 21

[51] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University, Taipei 106, Taiwan, July 2003. URL `http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf`. Last updated: May 19, 2016. 126

[52] Shaobo Huang. *Predictive modeling and analysis of student academic performance in an engineering dynamics course*. PhD thesis, Utah State University, 2011. 17, 57

[53] Richard A Huebner. A survey of educational data-mining research. *Research in higher education journal*, 19, 2013. 5, 38, 43, 47, 63, 69, 158

[54] Syed Tanveer Jishan, Raisul Islam Rashu, Naheena Haque, and Rashedur M Rahman. Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique. *Decision Analytics*, 2(1):1, 2015. 67

[55] Parneet Kaur, Manpreet Singh, and Gurpreet Singh Josan. Classification and prediction based data mining algorithms to predict slow learners in education sector. *Procedia Computer Science*, 57:500–508, 2015. 5, 57

[56] Sotiris B Kotsiantis, CJ Pierrakeas, and Panayiotis E Pintelas. Preventing student dropout in distance learning using machine learning techniques. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 267–274. Springer, 2003. 16, 17, 62

[57] Zsolt László Kovács. *Redes neurais artificiais*. Editora Livraria da Fisica, 2002. 21, 26

[58] Mukesh Kumar, A.J. Singh, and Disha Handa. Literature survey on educational dropout prediction. *International Journal of Education and Management Engineering (IJEME)*, 7 (2):8–19, 2017. 65, 66, 68, 69, 91, 161

[59] S Anupama Kumar and MN Vijayalakshmi. Prediction of the students recital using classification technique. *IFRSA's International journal of computing (IIJC)*, 1(3), 2011. 67

[60] Steven Lehr, Hong Liu, Sean Kinglesmith, Alex Konyha, Natalia Robaszewska, and Jacob Medinilla. Use educational data mining to predict undergraduate retention. In *Advanced Learning Technologies (ICALT), 2016 IEEE 16th International Conference on*, pages 428–430. IEEE, 2016. 61

[61] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL `http://CRAN.R-project.org/doc/Rnews/`. 21, 78, 125

[62] Andy Liaw and Matthew Wiener. randomForest for classification and regression in R. `https://cran.r-project.org/web/packages/randomForest`, 2015. 78, 125

[63] Peter Lyman, Hal R Varian, K Swearingen, P Charles, N Good, LL Jordan, and J Pal. How much information? 2003. *URL http://www2. sims. berkeley. edu/research/projects/how-much-info-2003*, 2005. 9

[64] Laci Mary Barbosa Manhães. *Predição Do Desempenho Acadêmico De Graduandos Utilizando Mineração De Dados Educacionais*. PhD thesis, Tese Doutorado). Universidade Federal do Rio de Janeiro, 2015. 17, 60, 61, 109

[65] Carlos Márquez-Vera, Alberto Cano, Cristobal Romero, Amin Yousef Mohammad Noaman, Habib Mousa Fardoun, and Sebastian Ventura. Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, 33(1):107–124, 2016. 62

[66] Maria PG Martins, Vera L Migueis, and DSB Fonseca. A data mining approach to predict undergraduate students' performance. In *2018 13th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–7. IEEE, 2018. doi: 10.23919/CISTI.2018.8399175. 145

[67] Maria PG Martins, Vera L Migueis, and DSB Fonseca. Educational data mining: A literature review. In *2018 13th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–6. IEEE, 2018. doi: 10.23919/CISTI.2018.8399281. 145

[68] Maria PG Martins, Vera L Miguéis, DSB Fonseca, and Albano Alves. A data mining approach for predicting academic success – a case study. In *Information Technology and Systems - Proceedings of ICITS 2019*, volume 918 of *Advances in Intelligent Systems and Computing*, pages 45–56. Springer, 2019. 145

[69] Robert May, Graeme Dandy, and Holger Maier. Review of input variable selection methods for artificial neural networks. In *Artificial neural networks-methodological advances and biomedical applications*. InTech, 2011. 127

[70] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2017. URL https://CRAN.R-project.org/package=e1071. R package version 1.6-8. 126

[71] Vera L Miguéis, Ana Freitas, Paulo JV Garcia, and André Silva. Early segmentation of students according to their academic performance: A predictive modelling approach. *Decision Support Systems*, 115:36–51, 2018. 5, 16, 56, 91, 140

[72] Rita Ramos Miguel, Daniel Rijo, and Luiza Nobre Lima. Fatores de risco para o insucesso escolar: a relevância das variáveis psicológicas e comportamentais do aluno. *Revista portuguesa de pedagogia*, pages 127–143, 2014. 3

[73] Manuel Miguéns. Estado da educação 2016. Conselho Nacional de Educação. Lisboa, 2017. 77

[74] Vera Miguéis, André Silva, Ana Freitas, and Paulo Garcia. A data mining approach for identifying actionable students for mitigation. Faculdade de Engenharia da Universidade do Porto, 2016. Submitted for publication. 17

[75] L Dee Miller, Leen-Kiat Soh, Ashok Samal, Kevin Kupzyk, and Gwen Nugent. A comparison of educational statistics and data mining approaches to identify characteristics that impact online learning. *Journal of Educational Data Mining*, 7(3):117–150, 2015. 4

[76] Siti Khadijah Mohamad and Zaidatun Tasir. Educational data mining: A review. *Procedia-Social and Behavioral Sciences*, 97:320–324, 2013. 43

[77] John Mylopoulos, Lawrence Chung, and Brian Nixon. Representing and using nonfunctional requirements: A process-oriented approach. *IEEE Transactions on software engineering*, 18(6):483–497, 1992. 13

[78] Ashutosh Nandeshwar, Tim Menzies, and Adam Nelson. Learning patterns of university student retention. *Expert Systems with Applications*, 38(12):14984–14996, 2011. 16, 60

[79] Srečko Natek and Moti Zwilling. Student data mining solution–knowledge management system related to higher education institutions. *Expert systems with applications*, 41 (14):6400–6407, 2014. 16, 17, 52, 67

[80] Vera Lúcia Miguéis Oliveira. *Analytical customer relationship management in retailing supported by data mining techniques*. PhD thesis, Universidade do Porto, 2012. 9, 16

[81] Jeroen Ooms, David James, Saikat DebRoy, Hadley Wickham, and Jeffrey Horner. *RMySQL: Database Interface and 'MySQL' Driver for R*, 2016. URL https://cran.r-project.org/web/packages/RMySQL. R package version 0.10.9. 75

[82] Mrinal Pandey and S. Taruna. Towards the integration of multiple classifier pertaining to the student's performance prediction. *Perspectives in Science*, 8:364 – 366, 2016. ISSN 2213-0209. doi: http://dx.doi.org/10.1016/j.pisc.2016.04.076. URL http://www.sciencedirect.com/science/article/pii/S2213020916300982. Recent Trends in Engineering and Material Sciences. 51

[83] Zacharoula Papamitsiou, Eirini Karapistoli, and Anastasios 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, pages 299–303. ACM, 2016. 17, 58

[84] Zacharoula K Papamitsiou and Anastasios A Economides. Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Educational Technology & Society*, 17(4):49–64, 2014. 43, 48, 49, 115, 159

[85] Túlio Albuquerque Pascoal, Daniel Miranda de Brito, and Thaís Gaudencio do Rêgo. Uma abordagem para a previsão de desempenho de alunos de computação em disciplinas de programação. *Nuevas Ideas en Informática Educativa TISE*, 2015:454–458, 2015. 17, 58, 91

[86] Alejandro Peña-Ayala. Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications*, 41(4):1432–1462, 2014. 5, 38, 43, 49, 69, 115, 159

[87] MSB PhridviRaj and CV GuruRao. Data mining–past, present and future–a typical survey on data streams. *Procedia Technology*, 12:255–263, 2014. 9, 16

[88] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org/`. 74

[89] Resolução da Assembleia da República n.º 176/2017 de 2 de Agosto. Diário da República nº 148 – 1ª Série. Ministério da Ciência Tecnologia e Ensino Superior. Lisboa, 2017. 1

[90] Mark A Revels and Hélène Nussbaumer. Data mining and data warehousing in the airline industry. *Academy of Business Research Journal*, 3, 2013. 12

[91] Brian D Ripley. *Pattern recognition and neural networks*. Cambridge university press, 2007. 126

[92] Crist√≥ bal Romero, Sebasti√° n Ventura, Mykola Pechenizkiy, and Ryan SJD Baker. *Handbook of educational data mining*. CRC Press, 2010. 38, 157

[93] Cristóbal Romero and Sebastián Ventura. Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, 33(1):135–146, 2007. 42, 43, 44, 45, 46, 157

[94] Cristóbal Romero and Sebastián Ventura. Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618, 2010. 5, 17, 39, 43, 45, 46, 47, 64

[95] Cristóbal Romero and Sebastián Ventura. Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):12–27, 2013. 15, 16, 17, 38, 42, 43, 47, 48, 64, 70, 77, 115, 158

[96] Cristóbal Romero, Sebastián Ventura, Pedro G Espejo, and César Hervás. Data mining algorithms to classify students. In *Educational Data Mining 2008*, 2008. 17, 51

[97] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2017. URL `http://www.rstudio.com/`. 74

[98] Sandra Milena Merchan Rubiano and Jorge Alberto Duarte Garcia. Analysis of data mining techniques for constructing a predictive model for academic performance. *IEEE Latin America Transactions*, 14(6):2783–2788, 2016. 54

[99] Jai Ruby and Dr K David. Analysis of influencing factors in predicting students performance using mlp-a comparative study. *International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization)*, 3 (2):10851092, 2015. 55

[100] Edilberto Ruiz and Fabio H. Nieto. A note on linear combination of predictors. *Statistics & Probability Letters*, 47(4):351 – 356, 2000. ISSN 0167-7152. doi: https://doi.org/10.1016/S0167-7152(99)00177-7. URL http://www.sciencedirect.com/science/article/pii/S0167715299001777. 117

[101] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533, 1986. 21

[102] Olga C Santos and Jesus G Boticario. User-centred design and educational data mining support during the recommendations elicitation process in social online learning environments. *Expert Systems*, 32(2):293–311, 2015. 52, 53

[103] Amirah Mohamed Shahiri, Wahidah Husain, et al. A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, 72:414–422, 2015. 17, 18, 65, 66, 67, 68, 161

[104] Md Hedayetul Islam Shovon and Mahfuza Haque. Prediction of student academic performance by an application of k-means clustering algorithm. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2(7), 2012. 68

[105] Catarina Silva and Bernardete Ribeiro. *Aprendizagem Computacional em Engenharia*. Imprensa da Universidade de Coimbra/Coimbra University Press, 2018. 25, 28

[106] Stefan Slater, Srećko Joksimović, Vitomir Kovanovic, Ryan S Baker, and Dragan Gasevic. Tools for educational data mining a review. *Journal of Educational and Behavioral Statistics*, page 1076998616666808, 2016. 42, 43, 74

[107] Kelly Spoon, Joshua Beemer, John C Whitmer, Juanjuan Fan, James P Frazee, Jeanne Stronach, Andrew J Bohonak, and Richard A Levine. Random forests for evaluating pedagogy and informing personalized learning. *Journal of Educational Data Mining*, 8(2): 20–50, 2016. 21

[108] Nikola Štambuk and Paško Konjevoda. The role of independent test set in modeling of protein folding kinetics. In *Software Tools and Algorithms for Biological Systems*, pages 279–284. Springer, 2011. 12

[109] Jun-Ming Su, Shian-Shyong Tseng, Huan-Yu Lin, and Chun-Han Chen. A personalized learning content adaptation mechanism to meet diverse user needs in mobile learning environments. *User modeling and user-adapted interaction*, 21(1):5–49, 2011. 64

[110] Karan Sukhija, Manish Jindal, and Naveen Aggarwal. The recent state of educational data mining: A survey and future visions. In *MOOCs, Innovation and Technology in Education (MITE), 2015 IEEE 3rd International Conference on*, pages 354–359. IEEE, 2015. 5, 42, 43, 50, 69, 159

[111] Mack Sweeney, Huzefa Rangwala, Jaime Lester, and Aditya Johri. Next-term student performance prediction: A recommender systems approach. *Journal of Educational Data Mining*, 8(1):22–51, 2016. 59

[112] Mahendra Tiwari, Randhir Singh, and Neeraj Vimal. An empirical study of applications of data mining techniques for predicting student performance in higher education. *International Journal of Computer Sciences and mobile Computing*, 2(2):53–57, 2013. 17, 51

[113] Efraim Turban, Ramesh Sharda, and Dursun Delen. *Decision support and business intelligence systems*. Pearson Education India, ninth edition, 2011. 10, 12

[114] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013. 26

[115] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL `http://www.stats.ox.ac.uk/pub/MASS4`. ISBN 0-387-95457-0. 126

[116] Johan J Vossensteyn, Andrea Kottmann, Benjamin WA Jongbloed, Franciscus Kaiser, Leon Cremonini, Bjorn Stensaker, Elisabeth Hovdhaugen, and Sabine Wollscheid. Dropout and completion in higher education in europe: Main report. Technical report, European Union, 2015. 1

[117] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016. 9, 12, 13, 17, 18

[118] Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, S Yu Philip, et al. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37, 2008. 16, 17

Apêndice A

Síntese das revisões sistemáticas da literatura

Tabela A.1: Síntese das revisões sistemáticas da literatura.

Obra	Objetivo	Conclusões
Romero and Ventura [93] (Analisam 81 estudos, publicados entre 1995 e 2005)	Identificar: tendências de investigação da área; sistemas educacionais; métodos e algoritmos de *data mining* mais usados.	Notória predominância dos estudos em contexto *e-learning* face à adoção do EDM no ensino tradicional presencial. Enfatizam a necessidade de desenvolvimento de trabalhos relacionados, até à consolidação do EDM como área de investigação.
Baker and Yacef [10] (Analisam 48 estudos, publicados entre 2000 e 2009)	Identificar as tendências de evolução do EDM. Determinar quais os métodos e algoritmos de DM mais usados em tarefas de EDM.	São quatro as principais tarefas de EDM: a melhoria dos modelos de ensino; a melhoria dos modelos dominantes, o estudo do suporte pedagógico que os softwares proporcionam e a investigação científica com vista ao desenvolvimento do ensino e dos estudantes. O EDM socorre-se de sete métodos diferentes: a estatística; a visualização; a *web mining*; o *clustering*; a classificação; a associação e o *text mining*.
Romero et al. [92] (Analisam 306 estudos, publicados entre 1993 a 2009)	Aferir a importância do DM no contexto atual da educação. Evidenciar as inovações que o EDM tem vindo a promover. Identificar as funcionalidades onde o DM pode ser útil na área de educação.	Reforçam a necessidade de aprofundar estudos, mais unificados e colaborativos, por forma ao EDM consolidar-se como área de investigação madura. Funcionalidades: a criação de *feedbacks*; o favorecimento da criação de recomendações; a previsão do desempenho; a análise e a visualização de dados; a construção de modelos sobre os estudantes; a possibilidade de agrupar estudantes em função de determinadas características; a possibilidade de detetar tipos de comportamentos dos estudantes; a análise do comportamento na rede social onde o estudante se insere; o planeamento e construção de ferramentas eletrónicas direcionadas para a educação e o desenvolvimento de conceitos. Para concretizar todas estas tarefas de EDM, os métodos mais usados continuam a ser a regressão, a classificação, o *clustering* e a associação. Os algoritmos mais usados são as árvores de decisão, as redes neuronais e as redes *bayesianas*.

Tabela A.1: (continuação da página anterior)

Obra	Objetivo	Conclusões
		O EDM não se restringe apenas ao uso ou benefício de professores e estudantes, uma vez que tem utilidade para as próprias instituições educativas, para os profissionais responsáveis pela criação e desenvolvimento de planos de estudos e até mesmo para os Estados.
Romero and Ventura [95] (Analisam 67 estudos, publicados entre 2001 e 2012)	Construir um guia orientador para quem pretende desenvolver estudos na área de EDM. Fornecer uma visão atualizada do estado dos conhecimentos em EDM. Identificação dos tópicos de investigação.	Os principais tópicos de investigação na área de EDM: – (Re)Organização das aulas ou da avaliação, a colocação de materiais com base no uso e dados de desempenho; a identificação daqueles que poderão beneficiar de comentários, conselhos de estudo ou outros géneros de ajuda; a delineação de estratégias para ajudar os alunos a encontrar e pesquisar materiais e bibliografia útil. – A criação de grupos de estudantes de acordo com as suas características de aprendizagem; modelação de alunos para desenvolver e ajustar os seus modelos cognitivos; construção de material didático para ajudar os instrutores e administradores no planeamento de cursos futuros. Indicam como linhas de orientação futura: – A necessidade de desenvolver uma cultura baseada em dados que permita tomar decisões destinadas a promover a eficiência das instituições. Entre possíveis soluções para este problema, apontam o uso de sistemas de suporte à decisão, mecanismos de recomendação e algoritmos de DM que autonomizem e facilitem aos instrutores todo o processo de EDM. – Reforçam uma vez mais a necessidade crescente de estudos de replicação para testar generalizações mais amplas.
Huebner [53] (Analisa 36 estudos publicados entre 2002 e 2012)	Identificar as formas como o *data mining* tem sido usado quando se pretende melhorar o sucesso dos aprendizes e os processos diretamente ligados à aprendizagem. Identificar oportunidades de Investigação.	Abarca exclusivamente tópicos como a retenção e o abandono escolar, os sistemas pessoais de recomendação em contextos educativos e as formas como o *data mining* tem sido usado quando se pretende melhorar o sucesso dos aprendizes ou otimizar os processos diretamente relacionados com a aprendizagem. Os métodos de previsão, de *clustering*, de classificação e de associação, são o paradigma dominante nos modelos analíticos desenvolvidos. Realçam três necessidades: – Estudar formas de tornar os resultados de *data mining* mais generalizáveis, providenciando o desenvolvimento de modelos que possam ser usados em múltiplos contextos; – O desenvolvimento de sistemas de apoio à decisão e de sistemas de recomendação que minimizem a intervenção dos educadores; – O desenvolvimento de ferramentas que protejam a privacidade individual dos intervenientes, ao mesmo tempo que possibilitam a EDM.

Tabela A.1: (continuação da página anterior)

Obra	Objetivo	Conclusões
Papamitsiou and Economides [84] (Analisam 40 estudos, publicados entre 2008 e 2013)	Identificar quais os métodos mais usados na literatura existente, para determinar a eficácia da implementação do EDM. Verificar até que ponto os métodos têm contribuído para a implementação do EDM como instrumento de análise e investigação.	Entre os métodos mais adotados surge em primeiro lugar o método de classificação, seguindo-se o método de *clustering* e de associação. Mais recentemente, já foram identificados estudos que comportavam novos métodos como, o *text mining*, o *association rule mining*, o *social network analysis*, o *discovery with models* e a visualização. Todos estes instrumentos terão sido aplicados com mais incidência em estudos de modelação comportamental dos estudantes e também na determinação de formas mais eficazes de prever o seu desempenho.
Peña-Ayala [86] (Analisam 240 estudos, publicados entre 2010 e o 1º período de 2013)	Identificar sistemas educacionais; tópicos de investigação; e métodos e algoritmos usados.	No conjunto dos 222 artigos analisados na perspetiva da caracterização e das funcionalidades do EDM, quase 82% dos estudos estão relacionados com as três versões de modelação de alunos: comportamental, desempenho e geral. O complemento (18%) foi distribuído por abordagens no âmbito do suporte pedagógico e *feedback* dos alunos, domínio de conhecimento (habilidades a serem treinadas) e suporte a professores. Nos 18 artigos que deram ênfase quer à funcionalidade quer aos instrumentos do EDM, os autores identificaram que 8 (a maioria) apresentaram o EDM como meio de análise, 6 justificam-no como ferramenta de visualização de dados e os restantes consideraram esta metodologia como forma de extrair informação de bases de dados. É também realçada a necessidade de aprofundamento dos estudos que visem a procura de modelos práticos de aplicação.
Sukhija et al. [110] (Analisam 19 estudos, publicados entre 2001 e 2015)	Promover e valorizar o EDM enquanto ferramenta de análise e construção de estratégias nos processos de tomada de decisão nas instituições académicas.	O EDM é ainda uma disciplina em expansão, a que está associado um vasto conjunto de métodos e algoritmos, e que continua a exigir atenção por parte dos investigadores, sobretudo ao nível do alargamento do conjunto de algoritmos que permitam a hibridação das técnicas de análise e agrupamento. Concluem que o EDM pode vir a constituir-se num instrumento que permita aos professores, estudantes e administrações educativas beneficiar do melhor que eles próprios tenham para oferecer.

Apêndice B

Síntese das revisões aos fatores determinantes do desempenho académico

Tabela B.1: Síntese das revisões aos fatores determinantes do desempenho académico.

Obra	Objetivo	Conclusões
Shahiri et al. [103]	Revisão sobre os preditores de desempenho académico.	Concluem que são 6 os atributos usados com mais frequência: CGPA (Cumulative Grade Point Average); indicadores de avaliação interna pós ingresso no ensino superior; características demográficas; indicadores de avaliação externa pré-ingresso; características relacionadas com a interação social dos estudantes.
(Analisam 30 estudos, publicados entre 2002 e janeiro de 2015)	Determinação dos algoritmos mais usados na mesma tarefa.	Os algoritmos Decision Tree (DT), Artificial Neural Networks (ANN), Naive Bayes (NB), K-Nearest Neighbor (K-NN) e as Support Vector Machines (SVM) foram, por ordem decrescente de importância, os mais usados nos trabalhos analisados.
Del Río and Insuasti [30]	Revisão sobre os preditores de desempenho académico, mas delimitado ao sistema presencial tradicional.	Para inferir a média final de curso, os investigadores usaram apenas indicadores de desempenho académico após o ingresso no ensino superior, em 37.5% dos estudos, e em combinação com mais um outro tipo de atributo, em 51.8% dos casos.
(Analisam 51 estudos, publicados entre 2011 e agosto de 2016)	Determinação dos métodos e algoritmos de *data mining* usados na tarefa que deu ênfase à revisão literária.	Entre os métodos de *data mining* mais usados, os autores destacaram o de classificação, dado que foi reportado em 71.4% das investigações referenciadas. Os métodos de agrupamento e de regras de associação foram os que se seguiram, tendo incidido, respetivamente, em 8.9% e 7.1% dos estudos.
Kumar et al. [58] (Analisam 14 estudos, publicados entre 2009 a 2016)	Revisão sobre os preditores de desempenho académico. Determinação dos métodos e algoritmos usados na mesma tarefa.	Na maioria dos estudos revistos, a média de acesso, o nível educacional e ocupação dos pais, e uma metodologia de ensino pobre são os principais fatores que afetam o resultado dos alunos. Os investigadores recorreram, essencialmente, aos métodos de classificação e associação, com cerca de 50% dos trabalhos referenciados a reportarem estes dois métodos.

Tabela B.1: (continuação da página anterior)

Obra	Objetivo	Conclusões
	Identificar as funcionalidades onde o EDM pode ser útil.	Identificam as seguintes funcionalidades do EDM: análise de padrões comportamentais de estudo em cursos online, previsão dos resultados académicos dos alunos, previsão do *ranking* do aluno, análise dos hábitos de aprendizagem online, análise de cursos MOOC, previsões de progressos ou retrocessos dos estudantes e, com maior predominância, a previsão de abandono escolar. Sublinharam ainda que a previsão do abandono escolar é tida como uma tarefa importante e desafiadora, para os investigadores, professores, instituições de ensino e até para os decisores políticos, confirmando-se a elevada contribuição do EDM em tarefas desta tipologia, nomeadamente quando se pretende identificar as características dos estudantes propensos ao abandono.

Apêndice C

Caracterização do 1º Ciclo de Estudos lecionado no IPB[1]

C.1 Estrutura curricular

A estrutura curricular respeita integralmente os princípios do Processo de Bolonha relativos à duração de 3 anos para o 1.º ciclo e permite o acesso ao mercado de trabalho e ingresso no 2.º ciclo para prosseguimento de estudos, estando organizada do seguinte modo: 6 semestres curriculares (3 anos); 20 semanas de estudo por semestre, a tempo inteiro (40 por ano); 40 horas totais por semana; 810 horas totais por semestre (1620 por ano); 180 créditos do ECTS (30 por semestre), correspondendo 1 crédito a 27 horas.

C.2 Condições de acesso e ingresso

Os estudantes podem candidatar-se ao 1º ciclo de estudos do IPB através do regime geral de acesso, dos regimes especiais de acesso e de concursos especiais, conforme consta da descrição do Sistema de Ensino Superior Português, disponibilizada pelo NARIC (http://www.dges.mctes.pt/DGES/pt/Reconhecimento/NARICENIC/) e apresentada na seção 8 do Suplemento do Diploma.

Concurso Nacional de Acesso

As candidaturas para acesso às licenciaturas do IPB dos estudantes titulares do ensino secundário são efetuadas através de Concurso Nacional de Acesso ao Ensino Superior, coordenado pela Direção Geral do Ensino Superior (DGES). Os candidatos deverão satisfazer as condições de acesso: ter aprovação num curso de ensino secundário ou habilitação legalmente equivalente, ter aprovação nas provas de ingresso exigidas e satisfazer os pré-requisitos, quando exigidos (no IPB, são exigidos pré-requisitos em todos os ciclos de estudos de licenciatura da Escola Superior de Saúde, qualquer que seja o concurso ou regime de acesso. Os pré-requisitos exigidos são do tipo de "seleção" e do Grupo A; ver legislação).

Concursos Especiais de Acesso

São organizados pelo IPB e destinam-se:

[1]Parte significativa da caraterização apresentada foi extraída ou baseada em informação disponível no portal do IPB (http://portal3.ipb.pt) e em relatórios da autoavaliação institucional de licenciaturas do IPB para a Agência de Avaliação e Acreditação do Ensino Superior (A3ES).

1. aos estudantes provenientes de cursos de especialização tecnológica (titulares de um diploma de especialização tecnológica, DETs);
2. aos estudantes provenientes de cursos técnicos superiores profissionais (titulares de um diploma de técnico superior profissional, DTeSPs);
3. aos candidatos aprovados nas provas destinadas a avaliarem a capacidade para a frequência do ensino superior dos maiores de 23 anos;
4. aos titulares de cursos médios e superiores.

Regimes de Reingresso e Mudança de Instituição/Curso

São organizados pelo IPB e destinam-se aos estudantes já inscritos no ensino superior, nacional ou estrangeiro:

1. Mudança de Instituição/Curso no Ensino Superior: quando um estudante do IPB ou de outra instituição de ensino superior, nacional ou estrangeira, pretende mudar para outro curso do IPB, tendo havido ou não interrupção de inscrição no ensino superior;
2. Reingresso: quando um estudante do IPB, após uma interrupção dos estudos num determinado curso, pretende reingressar no mesmo curso ou em curso que lhe tenha sucedido;

Estudantes Internacionais

O concurso especial para Estudantes Internacionais é organizado pelo IPB e destina-se aos estudantes de nacionalidade não portuguesa com Estatuto de Estudante Internacional. Possuem Estatuto de Estudante Internacional todos os estudantes que não têm nacionalidade portuguesa, com exceção dos nacionais de um Estado membro da União Europeia e os estudantes que, não sendo nacionais de um Estado membro da União Europeia, residam legalmente em Portugal há mais de dois anos. Os estudantes de nacionalidade não portuguesa não abrangidos pelo Estatuto de Estudante Internacional poderão ingressar num ciclo de estudos de Licenciatura do IPB através do Concurso Nacional de Acesso para estudantes com habilitações estrangeiras.

C.3 Estruturas e mecanismos de garantia da qualidade para o ciclo de estudos

Mecanismos de garantia da qualidade

Os mecanismos para a garantia da qualidade do ciclo de estudos baseiam-se em 4 instrumentos principais:

- modelos próprios de fichas de unidade curricular (UC) e de sumários, e para a publicação de documentação de apoio aos alunos, suportados por plataformas Web;
- relatório anual da comissão de curso, elaborado nos moldes definidos pelo conselho permanente do IPB, que reflete as atividades desenvolvidas em torno do ciclo de estudos e as preocupações dos alunos e dos docentes responsáveis pela lecionação das UCs;

- relatório de atividades da Escola, que é incluído no relatório de atividades do IPB, para aprovação pelo conselho geral do IPB, e onde são comparados e analisados indicadores variados — procura, taxas de sucesso, abandono, eficiência educativa, empregabilidade, etc. — para todos os cursos da Escola;
- relatório institucional sobre a concretização do Processo de Bolonha, no qual é analisada, de forma integrada, a evolução de todos os ciclos de estudos do IPB.

Responsáveis pela implementação dos mecanismos de garantia da qualidade

A implementação dos mecanismos de garantia da qualidade do ciclo de estudos compreende 3 níveis distintos de responsabilidade:

- diretor de curso, que é o responsável pela elaboração do relatório anual da comissão de curso;
- diretor da Escola, que é o responsável pela elaboração do relatório de atividades da Escola;
- vice-presidente do IPB para os assuntos académicos, que é o responsável pela elaboração do relatório institucional sobre a concretização do Processo de Bolonha e pelas plataformas Web de suporte à elaboração de fichas de unidade curricular (UC) e de sumários e à publicação de documentação de apoio aos alunos.

Procedimentos para a recolha de informação, acompanhamento e avaliação periódica do ciclo de estudos

A recolha de informação é efetuada fundamentalmente através de:

- inquéritos aos alunos para caracterização das entradas, avaliação do funcionamento das unidades curriculares (UCs), monitorização da carga de trabalho exigida, avaliação do nível de articulação entre matérias;
- inquéritos aos docentes para avaliação: da preparação dos alunos, do nível de articulação entre matérias e do número de créditos de cada UC;
- inquéritos aos empregadores para validação da adequação das competências dos diplomados às reais necessidades das empresas;
- inquéritos aos ex-alunos para aferir o grau de satisfação relativamente às competências e a adequação do emprego ao diploma;
- recolha automática, ao nível do sistema de informação da Instituição, de dados relativos ao sucesso escolar e ao abandono e de elementos para caracterização da utilização de ferramentas online e da frequência e acompanhamento de aulas;
- recolha de taxas de empregabilidade, tendo por base informação dos centros de emprego.

Utilização dos resultados das avaliações na definição de ações de melhoria

Os resultados das avaliações são tornados públicos, para discussão generalizada ao nível da comunidade académica e para conhecimento de futuros alunos, através do sítio web da Instituição. As comissões de curso e as comissões científicas refletem sobre as questões mais específicas do

ciclo de estudos, solicitando, aos departamentos, alterações ao nível das UCs e, caso tal se justifique, propondo alterações ao plano de estudos. Os departamentos analisam questões específicas das UCs pelas quais são responsáveis, implementando as melhorias que sejam necessárias. O Conselho Permanente da Escola debate questões transversais aos departamentos, acordando medidas de uniformização. O Conselho Pedagógico aprova alterações ao regulamento pedagógico e propõe medidas para melhoria do sucesso escolar. O Conselho Técnico-Científico aprova alterações aos planos de estudos e à forma como os docentes são alocados às UCs e pronuncia-se sobre a fixação de vagas e continuidade do ciclo de estudos.

C.4 Ambiente de ensino/aprendizagem

Estruturas e medidas de apoio pedagógico e de aconselhamento sobre o percurso académico dos estudantes

O acompanhamento dos alunos é efetuado, em primeira linha, pelos docentes da cada unidade curricular, que disponibilizam no seu horário 4 horas semanais (extra horário letivo) para atendimento pedagógico dos alunos. As comissões de curso e as comissões científicas organizam regularmente sessões de esclarecimento, nomeadamente em relação às saídas profissionais e à motivação dos alunos para o desenvolvimento de um percurso académico coerente. O Gabinete de Relações Internacionais e o Gabinete de Imagem e apoio ao aluno da Instituição são responsáveis pelo desenvolvimento de campanhas de divulgação de oportunidades de mobilidade internacional e de estágios em contexto de trabalho.

Medidas para promover a integração dos estudantes na comunidade académica

A integração dos alunos começa logo no ato de matrícula, com a entrega de informação diversa e realização de sessões individualizadas de esclarecimento e orientação, por parte de elementos do gabinete de imagem e apoio ao aluno, que durante esse período se encontram em permanência nos serviços académicos da Instituição. No fim do período de matrículas é organizada a receção oficial dos novos alunos, com a presença de todos os órgãos de gestão da Instituição e das Escolas, do provedor do estudante e de todos os responsáveis das associações de estudantes e da associação académica. A Associação de Estudantes de cada Escola e o núcleo de estudantes do ciclo de estudos, em coordenação com a Direção, desempenham também um papel importante no esclarecimento e integração dos novos alunos, no que respeita à especificidade da Escola. A comissão de curso, que integra docentes e alunos, é responsável pelo acompanhamento dos novos alunos ao longo de todo o ano.

Estruturas e medidas de aconselhamento sobre as possibilidades de financiamento e emprego

Os alunos da Instituição têm ao seu dispor um Gabinete de Empreendedorismo que ministra um programa de formação extra curricular, direcionado para as temáticas da criação e financiamento de negócios. O programa de formação inclui matérias como: Inovação, Estratégia, Desenho de Processos, Microeconomia, Análise de Investimentos, Formalidades e Financiamentos, Marketing, e Estudos de Mercado e Oportunidades. Dispõem, ainda, de um espaço para incubar os seus projetos empresariais e onde são assessorados em matéria de aconselhamento e consultoria empresarial. Está ainda ao dispor dos alunos, uma plataforma eletrónica,

http://comunidade.ipb.pt, que possibilita a gestão dos currículos e a consulta de todas as ofertas de emprego que chegam à Instituição.

Utilização dos resultados de inquéritos de satisfação dos estudantes na melhoria do processo ensino/aprendizagem

O Conselho Pedagógico de cada Escola promove, semestralmente, a realização de inquéritos pedagógicos. Os alunos, anonimamente, respondem a questões relacionadas com o funcionamento de cada unidade curricular e a questões sobre o desempenho dos docentes. As questões são de resposta fechada, cabendo ao aluno selecionar um nível de satisfação. Aos alunos que não frequentam as aulas é solicitado que indiquem as razões que os levam a tal. Os resultados do tratamento estatístico das respostas aos inquéritos são distribuídos aos docentes, aos coordenadores de departamento e aos diretores de curso, para efeitos de reflexão crítica. Ao nível dos departamentos e das comissões de curso, são analisados especialmente os casos com avaliações mais negativas, para definição de estratégias de convergência relativamente às práticas avaliadas de forma mais positiva pelos alunos.

Apêndice D

Conjunto de tabelas e atributos presentes nas bases de dados disponibilizadas

Tabela D.1: Lista de tabelas e atributos presentes nas Bases de Dados originais (Todos os dados pessoais presentes na última coluna foram ficcionados, de forma a manter-se a privacidade dos mesmos).

BD	Tabela (instâncias)	Atributo	1ª instância
acesso	candidato (501.489)	ano	*2007*
		bi	*123456789*
		nome	*Maria Martins*
	candidatura (2.842.597)	ano	*2007*
		fase	*1*
		cod_estab	*123*
		cod_curso	*1234*
		ordem	*1*
		bi	*123456789*
		media	*184.3*
		opcao	*2*
		pi	*191.5*
		n12	*177.0*
		n10_11	*177.0*
		colocacao	*0*
	curso (9.826)	ano	*2007*
		codigo	*1234*
		cod_estab	*123*
		nome	*Engenharia Civil*
		grau	*L*
		vagas	*40*
	estabelecimento (1.401)	ano	*2007*
		codigo	*123*
		nome	*Esc. Sup. de Tecnologia e Gestão*
		instituicao	*Instituto Politécnico de Bragança*
	exame (203)	ano	*2007*
		ano_ex	*2006*
		codigo	*102*
		nome	*Biologia*
	exame_pi (218)	ano	*2007*

Table D.1: (continuação da página anterior)

BD	Tabela (instâncias)	Atributo	1ª instância
acesso		ano_ex	*2006*
		cod_exame	*102*
		cod_pi	*22*
	nota_exame (419.006)	ano	*2008*
		bi	*123456789*
		cod_exame	*123*
		ano_ex	*2008*
		fase_ex	*1*
		nota	*175.0*
	nota_exame_pi (27.202)	ano	*2008*
		bi	*123456789*
		cod_exame	*321*
		ano_ex	*2008*
		fase_ex	*1*
		nota	*175.0*
	pi (40)	ano	*2007*
		codigo	*1*
		nome	*Alemão*
inqueritos	alunos_estat (16.151)	cod_aluno	*12345*
		nota_ingresso	*123.0*
		deslocado	*1*
		nivel_esc_pai	*19*
		nivel_esc_mae	*19*
		sit_prof_pai	*10*
		sit_prof_mae	*10*
		sit_prof_aluno	*17*
		cod_prof_pai	*11*
		cod_prof_mae	*11*
		cod_prof_aluno	*17*
		esc_sec	*Esc. Sec. Miguel Torga*
		local_esc_sec	*Bragança*
	nivel_escolar (13)	nivel_esc	*11*
		descricao	*Sabe ler s/ 4º ano de escolaridade*
	sit_profissinal (10)	cod_sit	*19*
		descricao	*Não Disponível*
	tipo_profissao (12)	cod_prof	*21*
		descricao	*Não disponível*
		observacao	
sa	alunos (36.656)	cod_aluno	*1234*
		nome	*Maria Martins*
		sexo	*F*
		data_nasc	*01/01/2000*
		bi_n	*123456789*
		nacionalidade	*Portugal*
		cod_freguesia	*1*

Table D.1: (continuação da página anterior)

BD	Tabela (instâncias)	Atributo	1ª instância
sa		cod_concelho	*2*
		cod_distrito	*3*
		cod_freguesia_natural	*10*
		cod_concelho_natural	*11*
		cod_distrito_natural	*12*
	concelhos (309)	cod_distrito	*1*
		cod_concelho	*11*
		concelho	*Mealhada*
	cursos (330)	cod_escola	*3040*
		cod_curso	*9118*
		curso	*Mestrado em Contabilidade*
		grau	*M*
		cod_curso_dges	*9118*
	disciplinas (15.150)	n_plano	*148*
		cod_curso	*9118*
		cod_escola	*3040*
		n_disciplina	*1101*
		disciplina	*Finanças Empresariais*
		ano	*1*
		semestre	*1*
		ects	*0.0*
	distritos (30)	cod_distrito	*2*
		distrito	*Beja*
	epocas (47)	cod_epoca	*1*
		epoca	*Normal Fev/Março*
	escolas (9)	cod_escola	*3043*
		abreviatura	*ESTiG*
		escola	*Esc. Sup. Tecn. Gest. de Bragança*
	freguesias (4.257)	cod_concelho	*1*
		cod_distrito	*1*
		cod_freguesia	*1*
		freguesia	*Agadão*
	lect_ini (140.794)	cod_aluno	*12345*
		n_plano	*123*
		cod_curso	*1234*
		cod_escola	*3043*
		ano_lect	*2000*
		data	*10/10/2000*
		ano_curricular	*1*
		bolseiro	*0*
		dir_associativo	*0*
		anulou	*0*
		anulou_data	
		cod_freq_tipo	*1*
		cod_estatuto	*0*

Table D.1: (continuação da página anterior)

BD	Tabela (instâncias)	Atributo	1ª instância
sa	matriculas (57.576)	cod_curso	*1234*
		cod_escola	*3043*
		cod_aluno	*12345*
		ano_mat	*2000*
		fase	*1*
		cod_tipo_ing	*1*
	notas (1.638.361)	n_disciplina	*1234*
		n_plano	*10*
		cod_curso	*123*
		cod_escola	*3043*
		cod_aluno	*12345*
		cod_tipo_nota	*3*
		cod_epoca	*1*
		ano_lect	*2000*
		data	*02/04/2001*
		nota	*10*
		n_opcao	*0*
	opcoes (6.262)	n_disciplina	*1104*
		n_plano	*148*
		cod_curso	*9118*
		cod_escola	*3040*
		n_opcao	*1*
		ano_lect	*1999*
		opcao	*Economia da Empresa*
	planos (677)	cod_curso	*490*
		cod_escola	*3043*
		n_plano	*38*
		nome	*Normal*
		ramo	
	stat_medias_conclusivos (25.034)	ano_lect	*2000*
		cod_aluno	*12345*
		cod_curso	*123*
		n_plano	*10*
		media	*15*
		cod_escola	*3043*
	tipo_estatutos (12)	cod_estatuto	*1*
		estatuto	*Agente de Ensino*
	tipo_frequencias (13)	cod_freq_tipo	*1*
		freq_tipo	*Ordinário*
	tipo_ingresso (23)	cod_tipo_ing	*1*
		tipo_ingresso	*Regime Geral*
	tipo_nota (14)	cod_tipo_nota	*1*
		tipo_nota	*Aprovado*
		positivo	*1*

Apêndice E

Resultados da aplicação do método *forward search*

Tabela E.1: Aplicação do método *forward search* ao *dataset* do 1º semestre.

(a) Procura da 1ª variável.

		RF		SVM		ANN		
variável	AUC médio	mtry	AUC	cost	AUC	size	decay	AUC
ects_aprov_s	0.7955	1	0.7359	2^{-1}	0.8198	5	$10^{-8/3}$	0.8308
nuca_s	0.7780	1	0.6863	2^{-5}	0.8238	1	$10^{-8/3}$	0.8238
max_s	0.7628	1	0.6863	2^{-5}	0.8011	1	$10^{-8/3}$	0.8011
media_s	0.7529	1	0.6873	2^{-5}	0.7858	1	$10^{-8/3}$	0.7858
ects_reprov_s	0.7317	1	0.6241	2^{-5}	0.7855	1	$10^{-8/3}$	0.7855
nucr_s	0.7283	1	0.6162	2^{-5}	0.7844	1	$10^{-8/3}$	0.7844
navalr_s	0.7256	1	0.6192	2^{-5}	0.7783	20	$10^{-8/3}$	0.7793
min_s	0.7133	1	0.6802	2^{-5}	0.7299	1	$10^{-8/3}$	0.7299
cod_curso	0.6063	1	0.5052	2^{-5}	0.5913	1	$10^{-1/3}$	0.7224
n10_11_acesso	0.5924	1	0.5441	2^{-5}	0.6164	5	$10^{-6/3}$	0.6167
n12_acesso	0.5879	1	0.5301	2^{-5}	0.6164	20	$10^{-8/3}$	0.6171
cod_escola	0.5779	1	0.5000	2^{-5}	0.5667	1	$10^{-8/3}$	0.6670
media_acesso	0.5722	1	0.5128	2^{-5}	0.6005	10	$10^{-8/3}$	0.6033
nivel_esc_pai	0.5542	1	0.5177	2^{1}	0.5516	1	$10^{-8/3}$	0.5932
nivel_esc_mae	0.5526	1	0.5164	2^{13}	0.5651	1	$10^{-1/3}$	0.5762
sit_prof_mae	0.5494	1	0.5320	2^{7}	0.5498	1	$10^{-8/3}$	0.5664
ects_cred_tx	0.5491	1	0.5177	2^{-5}	0.5498	20	$10^{-8/3}$	0.5797
idade	0.5488	1	0.5182	2^{-5}	0.5641	1	$10^{-8/3}$	0.5641
dist	0.5478	1	0.5190	2^{-5}	0.5584	2	$10^{-6/3}$	0.5660
cod_prof_mae	0.5475	1	0.5066	2^{9}	0.5418	1	$10^{0/3}$	0.5943
sexo	0.5461	1	0.5000	2^{-5}	0.5000	1	$10^{-8/3}$	0.6382
ordem_acesso	0.5421	1	0.5005	2^{-3}	0.5630	1	$10^{-8/3}$	0.5630
dist_n	0.5404	1	0.5033	2^{1}	0.5561	20	$10^{0/3}$	0.5619
bolseiro_s	0.5395	1	0.5000	2^{-5}	0.5000	1	$10^{-8/3}$	0.6185
pi_acesso	0.5386	1	0.5001	2^{5}	0.5512	10	$10^{-8/3}$	0.5644
ects_curso	0.5359	1	0.5000	2^{-5}	0.5000	1	$10^{-8/3}$	0.6078
cod_prof_pai	0.5345	1	0.5025	2^{5}	0.5450	1	$10^{-8/3}$	0.5561
fase	0.5335	1	0.5156	2^{-5}	0.5000	1	$10^{-8/3}$	0.5848
sit_prof_aluno	0.5313	1	0.5211	2^{3}	0.5432	1	$10^{-8/3}$	0.5295
cod_freq_tipo1_s	0.5188	1	0.5188	2^{-5}	0.5188	1	$10^{-8/3}$	0.5188
sit_prof_pai	0.5183	1	0.5011	2^{11}	0.5135	1	$10^{-8/3}$	0.5402
cod_prof_aluno	0.5176	1	0.5188	2^{15}	0.5161	10	$10^{0/3}$	0.5181
opcao_acesso	0.5126	1	0.5000	2^{15}	0.5103	1	$10^{-8/3}$	0.5276
dir_associativo_s	0.5057	1	0.5000	2^{-5}	0.5000	1	$10^{-8/3}$	0.5170
deslocado	0.5050	1	0.5000	2^{-5}	0.5000	1	$10^{-8/3}$	0.5149
nacionalidade	0.5039	1	0.5009	2^{11}	0.5077	1	$10^{-5/3}$	0.5032
cod_estatuto1_s	0.5027	1	0.5027	2^{9}	0.5027	1	$10^{-8/3}$	0.5027
ano_curricular_s	0.5000	1	0.5000	2^{-5}	0.5000	1	$10^{-8/3}$	0.5000

Tabela E.1: Aplicação do método *forward search* ao *dataset* do 1º semestre (continuação).

(b) Procura da 2ª variável (ects_aprov_s/?).

		RF		SVM		ANN		
variável	AUC médio	mtry	AUC	cost	AUC	size	decay	AUC
ects_cred_tx	0.8236	1	0.7988	2^{9}	0.8243	20	$10^{-7/3}$	0.8476
cod_prof_pai	0.8224	1	0.8101	2^{13}	0.8215	5	$10^{-8/3}$	0.8356
pi_acesso	0.8202	1	0.8031	2^{5}	0.8239	5	$10^{-7/3}$	0.8337
n12_acesso	0.8187	1	0.8054	2^{-3}	0.8237	20	$10^{-6/3}$	0.8270
navalr_s	0.8169	2	0.7756	2^{9}	0.8279	2	$10^{-5/3}$	0.8473
nucr_s	0.8169	2	0.7746	2^{9}	0.8310	5	$10^{-4/3}$	0.8451
ects_reprov_s	0.8165	1	0.7714	2^{11}	0.8328	5	$10^{-3/3}$	0.8452
n10_11_acesso	0.8156	1	0.8047	2^{13}	0.8142	5	$10^{-7/3}$	0.8278
nivel_esc_pai	0.8155	2	0.8009	2^{15}	0.8165	1	$10^{-5/3}$	0.8292
nivel_esc_mae	0.8151	1	0.7977	2^{-5}	0.8254	10	$10^{0/3}$	0.8223
media_s	0.8144	2	0.7871	2^{-5}	0.8149	20	$10^{-8/3}$	0.8411
cod_curso	0.8141	1	0.7598	2^{-1}	0.8275	5	$10^{-1/3}$	0.8549
media_acesso	0.8140	1	0.7850	2^{5}	0.8254	5	$10^{-7/3}$	0.8316
ordem_acesso	0.8139	1	0.7848	2^{11}	0.8264	5	$10^{-7/3}$	0.8306
sit_prof_mae	0.8124	2	0.8047	2^{1}	0.8071	1	$10^{-1/3}$	0.8255
cod_prof_mae	0.8122	1	0.7898	2^{11}	0.8223	20	$10^{0/3}$	0.8246
sexo	0.8063	2	0.7588	2^{3}	0.8304	20	$10^{-6/3}$	0.8297
bolseiro_s	0.8062	1	0.7679	2^{11}	0.8241	10	$10^{-8/3}$	0.8266
sit_prof_pai	0.8050	2	0.7695	2^{-5}	0.8197	2	$10^{-7/3}$	0.8259
deslocado	0.8046	2	0.7636	2^{11}	0.8209	5	$10^{-7/3}$	0.8293
max_s	0.8038	1	0.7669	2^{-5}	0.8112	20	$10^{-8/3}$	0.8333
idade	0.8036	1	0.7644	2^{1}	0.8226	5	$10^{-8/3}$	0.8239
fase	0.8028	1	0.7524	2^{7}	0.8263	20	$10^{-7/3}$	0.8295
dist	0.8021	1	0.7571	2^{5}	0.8245	5	$10^{0/3}$	0.8247
nuca_s	0.8010	2	0.7453	2^{9}	0.8258	20	$10^{-6/3}$	0.8320
dist_n	0.8006	2	0.7535	2^{7}	0.8250	20	$10^{-1/3}$	0.8232
min_s	0.8003	2	0.7378	2^{15}	0.8259	20	$10^{-8/3}$	0.8372
ects_curso	0.7997	1	0.7444	2^{5}	0.8233	20	$10^{-8/3}$	0.8314
dir_associativo_s	0.7996	2	0.7526	2^{15}	0.8220	5	$10^{-8/3}$	0.8244
cod_estatuto1_s	0.7984	1	0.7468	2^{5}	0.8200	5	$10^{-8/3}$	0.8283
opcao_acesso	0.7975	2	0.7570	2^{5}	0.8185	5	$10^{-5/3}$	0.8169
nacionalidade	0.7973	2	0.7363	2^{9}	0.8249	20	$10^{-8/3}$	0.8308
cod_escola	0.7972	2	0.7593	2^{-3}	0.7961	2	$10^{-1/3}$	0.8363
cod_prof_aluno	0.7948	1	0.7368	2^{11}	0.8212	10	$10^{-7/3}$	0.8265
ano_curricular_s	0.7932	1	0.7400	2^{-5}	0.8198	5	$10^{-5/3}$	0.8198
sit_prof_aluno	0.7914	2	0.7281	2^{15}	0.8214	10	$10^{-5/3}$	0.8247
cod_freq_tipo1_s	0.7892	1	0.7358	2^{-3}	0.8127	2	$10^{-5/3}$	0.8193

Tabela E.1: Aplicação do método *forward search* ao *dataset* do 1º semestre (continuação).

(c) Procura da 3ª variável (ects_aprov_s/ects_cred_tx/?).

		RF		SVM		ANN		
variável	AUC médio	mtry	AUC	cost	AUC	size	decay	AUC
ects_reprov_s	0.8399	1	0.8308	2^{3}	0.8236	10	$10^{-8/3}$	0.8653
media_s	0.8361	1	0.8279	2^{-5}	0.8256	10	$10^{-8/3}$	0.8549
nucr_s	0.8346	2	0.8163	2^{13}	0.8284	10	$10^{-8/3}$	0.8591
navalr_s	0.8323	2	0.8154	2^{-3}	0.8253	20	$10^{-8/3}$	0.8562
sexo	0.8321	1	0.8130	2^{7}	0.8350	5	$10^{-8/3}$	0.8483
pi_acesso	0.8316	2	0.8279	2^{1}	0.8275	20	$10^{-6/3}$	0.8395
n10_11_acesso	0.8304	1	0.8183	2^{-3}	0.8267	20	$10^{-7/3}$	0.8462
cod_curso	0.8296	1	0.7939	2^{-1}	0.8310	20	$10^{-2/3}$	0.8640
media_acesso	0.8294	1	0.8146	2^{1}	0.8283	20	$10^{-6/3}$	0.8453
max_s	0.8293	1	0.8119	2^{-3}	0.8262	5	$10^{-8/3}$	0.8499
n12_acesso	0.8289	2	0.8139	2^{11}	0.8293	10	$10^{-5/3}$	0.8434
bolseiro_s	0.8282	1	0.8124	2^{1}	0.8263	10	$10^{-8/3}$	0.8458
cod_escola	0.8280	1	0.8189	2^{-3}	0.8045	10	$10^{-8/3}$	0.8606
cod_prof_mae	0.8276	1	0.8169	2^{13}	0.8281	2	$10^{-4/3}$	0.8376
ects_curso	0.8275	1	0.8102	2^{15}	0.8236	10	$10^{-8/3}$	0.8487
dir_associativo_s	0.8272	2	0.8019	2^{13}	0.8280	10	$10^{-7/3}$	0.8516
cod_prof_pai	0.8268	1	0.8175	2^{13}	0.8261	2	$10^{-5/3}$	0.8369
sit_prof_mae	0.8267	1	0.8219	2^{13}	0.8226	5	$10^{-4/3}$	0.8356
nivel_esc_pai	0.8263	2	0.8182	2^{11}	0.8265	5	$10^{-6/3}$	0.8343
fase	0.8254	1	0.8068	2^{-3}	0.8255	20	$10^{-6/3}$	0.8440
nuca_s	0.8253	1	0.7928	2^{13}	0.8315	10	$10^{-6/3}$	0.8518
ordem_acesso	0.8250	1	0.8138	2^{11}	0.8210	20	$10^{-5/3}$	0.8403
idade	0.8250	2	0.8101	2^{3}	0.8249	5	$10^{-8/3}$	0.8401
nacionalidade	0.8242	1	0.7931	2^{-5}	0.8321	20	$10^{-8/3}$	0.8476
nivel_esc_mae	0.8240	1	0.8166	2^{13}	0.8234	5	$10^{-4/3}$	0.8320
cod_estatuto1_s	0.8229	1	0.7958	2^{9}	0.8266	10	$10^{-7/3}$	0.8464
deslocado	0.8227	1	0.7945	2^{15}	0.8260	10	$10^{-8/3}$	0.8475
cod_prof_aluno	0.8219	1	0.7980	2^{13}	0.8266	10	$10^{-8/3}$	0.8411
cod_freq_tipo1_s	0.8216	1	0.7979	2^{1}	0.8159	20	$10^{-7/3}$	0.8508
dist_n	0.8207	2	0.8037	2^{13}	0.8220	5	$10^{-5/3}$	0.8363
sit_prof_aluno	0.8206	2	0.7924	2^{13}	0.8278	5	$10^{-7/3}$	0.8415
ano_curricular_s	0.8204	2	0.7942	2^{13}	0.8243	20	$10^{-6/3}$	0.8427
dist	0.8201	1	0.7980	2^{15}	0.8217	5	$10^{-6/3}$	0.8406
opcao_acesso	0.8188	1	0.8022	2^{5}	0.8238	10	$10^{-7/3}$	0.8304
sit_prof_pai	0.8186	1	0.8000	2^{15}	0.8257	5	$10^{-4/3}$	0.8302
min_s	0.8168	2	0.7919	2^{-3}	0.8077	10	$10^{-7/3}$	0.8509

Tabela E.1: Aplicação do método *forward search* ao *dataset* do 1º semestre (continuação).

(d) Procura da 4ª variável (ects_aprov_s/ects_cred_tx/ects_reprov_s/?).

		RF		SVM		ANN		
variável	AUC médio	mtry	AUC	cost	AUC	size	decay	AUC
cod_escola	0.8453	2	0.8226	2^{-5}	0.8404	10	$10^{-4/3}$	0.8728
cod_prof_mae	0.8452	1	0.8433	2^{13}	0.8409	10	$10^{-2/3}$	0.8515
sexo	0.8447	1	0.8367	2^{-3}	0.8344	10	$10^{-8/3}$	0.8632
cod_freq_tipo1_s	0.8447	1	0.8333	2^{13}	0.8268	20	$10^{-8/3}$	0.8739
pi_acesso	0.8443	2	0.8439	2^{15}	0.8316	20	$10^{-6/3}$	0.8573
media_s	0.8442	1	0.8352	2^{-5}	0.8371	10	$10^{-8/3}$	0.8601
dir_associativo_s	0.8433	1	0.8345	2^{-1}	0.8307	20	$10^{-8/3}$	0.8647
nacionalidade	0.8430	1	0.8319	2^{9}	0.8307	20	$10^{-8/3}$	0.8663
ordem_acesso	0.8430	1	0.8340	2^{13}	0.8351	20	$10^{-7/3}$	0.8598
n12_acesso	0.8428	1	0.8372	2^{13}	0.8285	20	$10^{-8/3}$	0.8628
n10_11_acesso	0.8426	2	0.8429	2^{1}	0.8253	5	$10^{-7/3}$	0.8595
ects_curso	0.8416	1	0.8337	2^{-5}	0.8248	10	$10^{-8/3}$	0.8663
cod_prof_pai	0.8406	2	0.8330	2^{11}	0.8350	5	$10^{-2/3}$	0.8537
max_s	0.8405	3	0.8256	2^{-5}	0.8370	5	$10^{-3/3}$	0.8590
cod_curso	0.8402	1	0.8068	2^{-3}	0.8427	10	$10^{-2/3}$	0.8711
sit_prof_mae	0.8400	1	0.8402	2^{11}	0.8324	2	$10^{-8/3}$	0.8474
dist	0.8393	4	0.8241	2^{13}	0.8386	10	$10^{-2/3}$	0.8552
cod_estatuto1_s	0.8389	1	0.8375	2^{3}	0.8219	20	$10^{-8/3}$	0.8573
media_acesso	0.8387	2	0.8395	2^{-1}	0.8225	5	$10^{-7/3}$	0.8543
fase	0.8386	1	0.8341	2^{15}	0.8270	10	$10^{-7/3}$	0.8548
navalr_s	0.8384	1	0.8354	2^{-1}	0.8261	20	$10^{-8/3}$	0.8537
bolseiro_s	0.8378	2	0.8248	2^{-5}	0.8308	20	$10^{-8/3}$	0.8580
dist_n	0.8378	4	0.8207	2^{15}	0.8373	20	$10^{-3/3}$	0.8555
deslocado	0.8375	3	0.8350	2^{1}	0.8185	20	$10^{-8/3}$	0.8591
sit_prof_aluno	0.8364	1	0.8256	2^{11}	0.8288	20	$10^{-7/3}$	0.8548
cod_prof_aluno	0.8363	2	0.8139	2^{13}	0.8328	20	$10^{-8/3}$	0.8621
nivel_esc_mae	0.8355	3	0.8373	2^{11}	0.8299	20	$10^{-2/3}$	0.8393
idade	0.8352	2	0.8217	2^{13}	0.8290	20	$10^{-6/3}$	0.8550
nuca_s	0.8347	2	0.8124	2^{1}	0.8289	20	$10^{-6/3}$	0.8627
min_s	0.8344	2	0.8039	2^{-5}	0.8283	20	$10^{-8/3}$	0.8711
nucr_s	0.8340	3	0.8146	2^{1}	0.8248	20	$10^{-7/3}$	0.8626
nivel_esc_pai	0.8335	1	0.8304	2^{13}	0.8287	10	$10^{-3/3}$	0.8415
opcao_acesso	0.8313	1	0.8236	2^{-3}	0.8233	5	$10^{-3/3}$	0.8471
ano_curricular_s	0.8306	2	0.8120	2^{13}	0.8273	20	$10^{-7/3}$	0.8525
sit_prof_pai	0.8276	3	0.8256	2^{-1}	0.8148	20	$10^{-4/3}$	0.8425

Tabela E.1: Aplicação do método *forward search* ao *dataset* do 1º semestre (continuação).

(e) Procura da 5ª variável (ects_aprov_s/ects_cred_tx/ects_reprov_s/cod_escola/?).

		RF		SVM		ANN		
variável	AUC médio	mtry	AUC	cost	AUC	size	decay	AUC
media_s	0.8559	1	0.8443	2^{-3}	0.8455	10	$10^{-5/3}$	0.8778
sit_prof_mae	0.8551	1	0.8548	2^{1}	0.8435	2	$10^{-7/3}$	0.8670
cod_estatuto1_s	0.8520	1	0.8502	2^{-5}	0.8356	5	$10^{-5/3}$	0.8700
sit_prof_aluno	0.8511	1	0.8408	2^{-5}	0.8429	20	$10^{-4/3}$	0.8697
media_acesso	0.8510	2	0.8583	2^{3}	0.8239	20	$10^{-4/3}$	0.8708
n12_acesso	0.8508	2	0.8435	2^{15}	0.8315	10	$10^{-4/3}$	0.8774
nacionalidade	0.8505	1	0.8391	2^{-5}	0.8408	5	$10^{-4/3}$	0.8714
max_s	0.8503	1	0.8348	2^{-3}	0.8407	5	$10^{-5/3}$	0.8753
pi_acesso	0.8500	2	0.8503	2^{1}	0.8302	5	$10^{-6/3}$	0.8694
ects_curso	0.8499	1	0.8372	2^{-5}	0.8415	10	$10^{-5/3}$	0.8711
n10_11_acesso	0.8497	1	0.8466	2^{5}	0.8329	20	$10^{-4/3}$	0.8697
cod_freq_tipo1_s	0.8486	1	0.8428	2^{-5}	0.8325	20	$10^{-4/3}$	0.8704
sit_prof_pai	0.8477	2	0.8412	2^{-5}	0.8372	2	$10^{-3/3}$	0.8647
dist	0.8476	2	0.8436	2^{-5}	0.8279	20	$10^{-2/3}$	0.8713
ordem_acesso	0.8474	2	0.8351	2^{1}	0.8393	20	$10^{-2/3}$	0.8679
ano_curricular_s	0.8465	1	0.8294	2^{-5}	0.8403	10	$10^{-4/3}$	0.8699
nucr_s	0.8464	1	0.8316	2^{-5}	0.8366	20	$10^{-4/3}$	0.8710
cod_prof_mae	0.8464	1	0.8514	2^{-5}	0.8199	5	$10^{-3/3}$	0.8678
idade	0.8460	2	0.8425	2^{-1}	0.8244	20	$10^{-4/3}$	0.8712
nuca_s	0.8460	1	0.8263	2^{-3}	0.8408	20	$10^{-4/3}$	0.8708
nivel_esc_pai	0.8448	1	0.8396	2^{-3}	0.8369	2	$10^{-8/3}$	0.8578
sexo	0.8447	1	0.8418	2^{15}	0.8209	20	$10^{-4/3}$	0.8715
bolseiro_s	0.8440	1	0.8378	2^{11}	0.8198	10	$10^{-4/3}$	0.8745
cod_prof_aluno	0.8440	2	0.8228	2^{-5}	0.8405	10	$10^{-4/3}$	0.8688
cod_prof_pai	0.8438	1	0.8481	2^{15}	0.8171	20	$10^{-2/3}$	0.8662
fase	0.8429	1	0.8415	2^{-1}	0.8170	20	$10^{-3/3}$	0.8701
dir_associativo_s	0.8427	1	0.8384	2^{-5}	0.8172	20	$10^{-4/3}$	0.8724
deslocado	0.8425	1	0.8358	2^{-5}	0.8278	10	$10^{-5/3}$	0.8639
navalr_s	0.8423	4	0.8352	2^{-5}	0.8185	20	$10^{-4/3}$	0.8731
nivel_esc_mae	0.8413	2	0.8488	2^{-5}	0.8193	20	$10^{-2/3}$	0.8557
dist_n	0.8407	2	0.8341	2^{-5}	0.8164	10	$10^{-2/3}$	0.8716
min_s	0.8406	2	0.8164	2^{-3}	0.8319	10	$10^{-4/3}$	0.8735
opcao_acesso	0.8381	1	0.8301	2^{-5}	0.8225	20	$10^{-2/3}$	0.8617
cod_curso	0.7524	1	0.8262	2^{-5}	0.5606	10	$10^{-2/3}$	0.8703

Tabela E.1: Aplicação do método *forward search* ao *dataset* do 1º semestre (continuação).

(f) Procura da 6ª variável (ects_aprov_s/ects_cred_tx/ects_reprov_s/cod_escola/media_s/?).

		RF		SVM		ANN		
variável	AUC médio	mtry	AUC	cost	AUC	size	decay	AUC
sexo	0.8621	1	0.8641	2^{-3}	0.8456	20	$10^{-3/3}$	0.8766
media_acesso	0.8588	1	0.8578	2^{-1}	0.8393	20	$10^{-3/3}$	0.8792
cod_prof_pai	0.8586	2	0.8574	2^{-5}	0.8436	2	$10^{-7/3}$	0.8748
dir_associativo_s	0.8580	2	0.8483	2^{-3}	0.8449	10	$10^{-5/3}$	0.8809
pi_acesso	0.8580	1	0.8552	2^{-5}	0.8414	5	$10^{-3/3}$	0.8774
n12_acesso	0.8576	1	0.8575	2^{-1}	0.8388	20	$10^{-3/3}$	0.8765
sit_prof_aluno	0.8566	1	0.8511	2^{-5}	0.8422	20	$10^{-4/3}$	0.8766
nacionalidade	0.8566	4	0.8466	2^{-5}	0.8462	10	$10^{-5/3}$	0.8770
ano_curricular_s	0.8565	2	0.8455	2^{-3}	0.8455	10	$10^{-4/3}$	0.8785
sit_prof_mae	0.8564	1	0.8584	2^{3}	0.8402	20	$10^{-2/3}$	0.8707
n10_11_acesso	0.8564	1	0.8511	2^{-1}	0.8398	10	$10^{-6/3}$	0.8782
bolseiro_s	0.8563	2	0.8456	2^{-5}	0.8435	20	$10^{-6/3}$	0.8798
idade	0.8560	1	0.8529	2^{-5}	0.8414	5	$10^{-6/3}$	0.8738
cod_estatuto1_s	0.8560	3	0.8435	2^{-3}	0.8455	10	$10^{-8/3}$	0.8790
cod_prof_mae	0.8554	1	0.8493	2^{-5}	0.8420	10	$10^{-2/3}$	0.8749
dist	0.8551	2	0.8434	2^{-5}	0.8431	20	$10^{-2/3}$	0.8789
cod_prof_aluno	0.8551	1	0.8461	2^{-3}	0.8459	10	$10^{-3/3}$	0.8734
ects_curso	0.8550	1	0.8502	2^{-5}	0.8379	5	$10^{-5/3}$	0.8768
max_s	0.8548	1	0.8448	2^{-5}	0.8437	10	$10^{-5/3}$	0.8760
nuca_s	0.8544	3	0.8466	2^{-1}	0.8421	5	$10^{-4/3}$	0.8744
dist_n	0.8540	1	0.8437	2^{-5}	0.8450	20	$10^{-2/3}$	0.8733
cod_freq_tipo1_s	0.8539	1	0.8410	2^{-5}	0.8454	10	$10^{-6/3}$	0.8754
nivel_esc_mae	0.8532	1	0.8567	2^{-3}	0.8403	20	$10^{-2/3}$	0.8625
nucr_s	0.8531	1	0.8412	2^{7}	0.8410	10	$10^{-4/3}$	0.8769
deslocado	0.8528	1	0.8443	2^{-5}	0.8424	2	$10^{-3/3}$	0.8717
sit_prof_pai	0.8526	1	0.8483	2^{3}	0.8429	10	$10^{-2/3}$	0.8666
ordem_acesso	0.8526	2	0.8398	2^{3}	0.8394	5	$10^{-3/3}$	0.8785
navalr_s	0.8522	1	0.8402	2^{-3}	0.8405	20	$10^{-4/3}$	0.8759
nivel_esc_pai	0.8510	2	0.8429	2^{-3}	0.8439	5	$10^{-3/3}$	0.8663
opcao_acesso	0.8493	1	0.8440	2^{-3}	0.8349	5	$10^{-4/3}$	0.8690
min_s	0.8439	2	0.8379	2^{-5}	0.8187	10	$10^{-4/3}$	0.8750
fase	0.7740	1	0.8525	2^{-5}	0.5923	2	$10^{-2/3}$	0.8772
cod_curso	0.7608	1	0.8390	2^{-5}	0.5697	10	$10^{-2/3}$	0.8737

Tabela E.1: Aplicação do método *forward search* ao *dataset* do 1º semestre (continuação).

(g) Procura da 7ª variável (ects_aprov_s/ects_cred_tx/ects_reprov_s/cod_escola/media_s/sexo/?).

		RF		SVM		ANN		
variável	AUC médio	mtry	AUC	cost	AUC	size	decay	AUC
bolseiro_s	0.8672	2	0.8702	2^{-3}	0.8440	20	$10^{-5/3}$	0.8872
cod_prof_pai	0.8633	1	0.8701	2^{-5}	0.8437	20	$10^{-2/3}$	0.8762
cod_prof_mae	0.8628	2	0.8677	2^{-5}	0.8440	20	$10^{-2/3}$	0.8768
dir_associativo_s	0.8623	2	0.8641	2^{-1}	0.8443	20	$10^{-3/3}$	0.8784
fase	0.8620	2	0.8675	2^{-1}	0.8427	5	$10^{-1/3}$	0.8759
cod_freq_tipo1_s	0.8614	2	0.8668	2^{-1}	0.8421	5	$10^{-5/3}$	0.8754
sit_prof_aluno	0.8614	2	0.8692	2^{13}	0.8420	20	$10^{-3/3}$	0.8731
pi_acesso	0.8613	3	0.8630	2^{-5}	0.8437	20	$10^{-3/3}$	0.8773
cod_prof_aluno	0.8612	2	0.8654	2^{-1}	0.8442	5	$10^{-2/3}$	0.8739
cod_estatuto1_s	0.8612	2	0.8651	2^{-3}	0.8454	5	$10^{-2/3}$	0.8729
nacionalidade	0.8611	2	0.8613	2^{-5}	0.8458	2	$10^{-2/3}$	0.8764
dist	0.8609	2	0.8595	2^{-5}	0.8460	5	$10^{-2/3}$	0.8772
ects_curso	0.8604	2	0.8621	2^{-5}	0.8428	5	$10^{-2/3}$	0.8763
ano_curricular_s	0.8604	3	0.8623	2^{-3}	0.8456	2	$10^{-2/3}$	0.8732
idade	0.8598	2	0.8619	2^{-3}	0.8433	10	$10^{-8/3}$	0.8742
deslocado	0.8597	2	0.8626	2^{-5}	0.8428	5	$10^{-7/3}$	0.8738
nivel_esc_mae	0.8595	2	0.8662	2^{-5}	0.8455	2	$10^{-1/3}$	0.8669
navalr_s	0.8594	2	0.8608	2^{-3}	0.8430	2	$10^{-2/3}$	0.8746
sit_prof_mae	0.8594	2	0.8709	2^{3}	0.8360	5	$10^{-2/3}$	0.8714
nucr_s	0.8594	3	0.8580	2^{-1}	0.8439	20	$10^{-3/3}$	0.8763
nuca_s	0.8594	1	0.8584	2^{-5}	0.8440	10	$10^{-4/3}$	0.8758
max_s	0.8593	2	0.8598	2^{-5}	0.8417	5	$10^{-3/3}$	0.8763
ordem_acesso	0.8585	1	0.8567	2^{13}	0.8400	10	$10^{-3/3}$	0.8789
dist_n	0.8575	2	0.8567	2^{11}	0.8447	10	$10^{-1/3}$	0.8710
n10_11_acesso	0.8574	1	0.8647	2^{-1}	0.8311	5	$10^{-4/3}$	0.8762
n12_acesso	0.8567	1	0.8633	2^{-1}	0.8298	10	$10^{-3/3}$	0.8772
media_acesso	0.8567	2	0.8610	2^{-5}	0.8317	5	$10^{-4/3}$	0.8775
nivel_esc_pai	0.8567	2	0.8556	2^{-1}	0.8440	10	$10^{-2/3}$	0.8706
opcao_acesso	0.8567	1	0.8579	2^{-3}	0.8415	2	$10^{-5/3}$	0.8706
sit_prof_pai	0.8561	2	0.8636	2^{1}	0.8349	2	$10^{-3/3}$	0.8696
min_s	0.8556	2	0.8560	2^{-3}	0.8357	10	$10^{-3/3}$	0.8751
cod_curso	0.8132	1	0.8503	2^{5}	0.7080	20	$10^{-2/3}$	0.8814

Tabela E.1: Aplicação do método *forward search* ao *dataset* do 1º semestre (continuação).

(h) Procura da 8ª variável (ects_aprov_s/ects_cred_tx/ects_reprov_s/cod_escola/media_s/sexo/bolseiro_s/$\times$).

		RF		SVM		ANN		
variável	AUC médio	mtry	AUC	cost	AUC	size	decay	AUC
dir_associativo_s	0.8671	3	0.8695	2^{-3}	0.8451	20	$10^{-3/3}$	0.8867
pi_acesso	0.8666	2	0.8736	2^{-3}	0.8434	10	$10^{-3/3}$	0.8827
cod_prof_aluno	0.8664	3	0.8703	2^{-3}	0.8441	10	$10^{-3/3}$	0.8849
sit_prof_mae	0.8655	2	0.8763	2^{11}	0.8423	5	$10^{-2/3}$	0.8778
cod_prof_mae	0.8654	2	0.8744	2^{-5}	0.8433	5	$10^{-2/3}$	0.8786
nacionalidade	0.8653	2	0.8692	2^{-3}	0.8445	10	$10^{-6/3}$	0.8821
nucr_s	0.8648	2	0.8700	2^{-3}	0.8431	10	$10^{-4/3}$	0.8814
ects_curso	0.8645	2	0.8676	2^{-5}	0.8428	10	$10^{-4/3}$	0.8829
cod_estatuto1_s	0.8644	3	0.8656	2^{-3}	0.8444	10	$10^{-4/3}$	0.8832
deslocado	0.8643	2	0.8682	2^{-5}	0.8433	20	$10^{-3/3}$	0.8815
idade	0.8640	2	0.8671	2^{-1}	0.8437	10	$10^{-4/3}$	0.8812
ano_curricular_s	0.8634	4	0.8666	2^{-3}	0.8440	10	$10^{-5/3}$	0.8797
cod_prof_pai	0.8631	2	0.8716	2^{-3}	0.8415	20	$10^{-1/3}$	0.8763
nuca_s	0.8629	2	0.8643	2^{-5}	0.8418	5	$10^{-8/3}$	0.8826
cod_freq_tipo1_s	0.8623	2	0.8636	2^{-3}	0.8435	20	$10^{-4/3}$	0.8799
min_s	0.8617	3	0.8640	2^{-5}	0.8387	20	$10^{-5/3}$	0.8825
navalr_s	0.8617	3	0.8657	2^{-5}	0.8385	20	$10^{-4/3}$	0.8810
n12_acesso	0.8615	1	0.8699	2^{-5}	0.8318	5	$10^{-6/3}$	0.8829
sit_prof_pai	0.8615	2	0.8657	2^{3}	0.8426	20	$10^{-2/3}$	0.8762
max_s	0.8612	2	0.8641	2^{-5}	0.8357	20	$10^{-4/3}$	0.8838
nivel_esc_mae	0.8611	3	0.8675	2^{-3}	0.8458	10	$10^{-2/3}$	0.8701
media_acesso	0.8608	4	0.8660	2^{-3}	0.8334	20	$10^{-4/3}$	0.8831
dist	0.8608	2	0.8624	2^{-3}	0.8441	10	$10^{-1/3}$	0.8759
dist_n	0.8608	2	0.8593	2^{-5}	0.8457	20	$10^{-2/3}$	0.8774
opcao_acesso	0.8603	2	0.8638	2^{5}	0.8419	20	$10^{-2/3}$	0.8752
n10_11_acesso	0.8600	2	0.8714	2^{-5}	0.8291	10	$10^{-4/3}$	0.8797
ordem_acesso	0.8589	2	0.8595	2^{-1}	0.8352	10	$10^{-3/3}$	0.8821
nivel_esc_pai	0.8589	2	0.8584	2^{-3}	0.8442	5	$10^{-2/3}$	0.8741
sit_prof_aluno	0.8462	3	0.8713	2^{3}	0.7846	20	$10^{-4/3}$	0.8827
cod_curso	0.8300	3	0.8503	2^{9}	0.7567	5	$10^{-3/3}$	0.8831
fase	0.7814	1	0.8711	2^{-5}	0.5910	10	$10^{-4/3}$	0.8821

Tabela E.2: Aplicação do método *forward search* ao *dataset* dos 2 primeiros semestres.

(a) Procura da 1ª variável.

		RF		SVM		ANN		
variável	AUC médio	mtry	AUC	cost	AUC	size	decay	AUC
ects_reprov_s	0.8568	1	0.8119	2^{-5}	0.8793	1	$10^{-8/3}$	0.8793
ects_aprov_s	0.8434	1	0.7748	2^{-5}	0.8777	1	$10^{-8/3}$	0.8777
nucr_s	0.8328	1	0.7579	2^{-5}	0.8702	1	$10^{-8/3}$	0.8702
nuca_s	0.8284	1	0.7326	2^{-5}	0.8763	1	$10^{-8/3}$	0.8763
navalr_s	0.8219	1	0.7262	2^{-5}	0.8697	1	$10^{-8/3}$	0.8697
media_s	0.7663	1	0.7098	2^{-5}	0.7946	1	$10^{-8/3}$	0.7945
max_s	0.7646	1	0.6906	2^{-5}	0.8015	1	$10^{-8/3}$	0.8015
min_s	0.6617	1	0.6563	2^{-5}	0.6643	1	$10^{-8/3}$	0.6643
vd12_s	0.6450	1	0.7107	2^{-5}	0.6122	1	$10^{-8/3}$	0.6122
cod_curso	0.6082	1	0.5053	2^{7}	0.5879	2	$10^{0/3}$	0.7314
n12_acesso	0.5897	1	0.5390	2^{5}	0.6148	5	$10^{-6/3}$	0.6152
n10_11_acesso	0.5857	1	0.5272	2^{-3}	0.6148	10	$10^{-6/3}$	0.6152
cod_escola	0.5775	1	0.5000	2^{-5}	0.5655	1	$10^{-8/3}$	0.6670
media_acesso	0.5697	1	0.5004	2^{-1}	0.6033	10	$10^{-8/3}$	0.6054
nivel_esc_pai	0.5569	1	0.5150	2^{11}	0.5655	1	$10^{-8/3}$	0.5901
idade	0.5501	1	0.5188	2^{-5}	0.5658	1	$10^{-8/3}$	0.5658
pi_acesso	0.5494	1	0.5248	2^{-5}	0.5575	20	$10^{-7/3}$	0.5659
nivel_esc_mae	0.5493	1	0.5166	2^{13}	0.5558	1	$10^{-1/3}$	0.5754
ects_cred_tx	0.5475	1	0.5144	2^{-5}	0.5510	10	$10^{-8/3}$	0.5772
sexo	0.5462	1	0.5000	2^{-5}	0.5000	1	$10^{-8/3}$	0.6385
dist	0.5435	1	0.5156	2^{5}	0.5456	10	$10^{-7/3}$	0.5693
ordem_acesso	0.5414	1	0.5034	2^{-5}	0.5604	1	$10^{-8/3}$	0.5604
cod_prof_mae	0.5410	1	0.5065	2^{13}	0.5320	1	$10^{-1/3}$	0.5845
dist_n	0.5409	1	0.5062	2^{7}	0.5541	5	$10^{-1/3}$	0.5623
cod_prof_pai	0.5391	1	0.5000	2^{15}	0.5396	1	$10^{-8/3}$	0.5777
bolseiro_s	0.5391	1	0.5000	2^{-5}	0.5000	1	$10^{-8/3}$	0.6172
ects_curso	0.5358	1	0.5000	2^{-5}	0.5000	1	$10^{-8/3}$	0.6073
sit_prof_mae	0.5331	1	0.5140	2^{5}	0.5234	1	$10^{-3/3}$	0.5619
sit_prof_pai	0.5315	1	0.5011	2^{3}	0.5465	1	$10^{-8/3}$	0.5468
fase	0.5293	1	0.5000	2^{-5}	0.5000	1	$10^{-8/3}$	0.5878
sit_prof_aluno	0.5270	1	0.5214	2^{5}	0.5294	1	$10^{-2/3}$	0.5303
cod_prof_aluno	0.5208	1	0.5190	2^{15}	0.5218	1	$10^{-8/3}$	0.5217
cod_freq_tipo1_s	0.5191	1	0.5191	2^{-5}	0.5191	1	$10^{-8/3}$	0.5191
cod_freq_tipo2_s	0.5191	1	0.5191	2^{-5}	0.5191	1	$10^{-8/3}$	0.5191
opcao_acesso	0.5109	1	0.5000	2^{9}	0.5096	5	$10^{0/3}$	0.5232
dir_associativo_s	0.5058	1	0.5000	2^{-5}	0.5000	1	$10^{-8/3}$	0.5174
deslocado	0.5043	1	0.5000	2^{-5}	0.5000	1	$10^{-8/3}$	0.5128
nacionalidade	0.5035	1	0.5010	2^{9}	0.5061	1	$10^{-8/3}$	0.5033
cod_estatuto1_s	0.5027	1	0.5027	2^{3}	0.5027	1	$10^{-8/3}$	0.5027
cod_estatuto2_s	0.5027	1	0.5027	2^{3}	0.5027	1	$10^{-8/3}$	0.5027
ano_curricular_s	0.5000	1	0.5000	2^{-5}	0.5000	1	$10^{-8/3}$	0.5000

Tabela E.2: Aplicação do método *forward search* ao *dataset* dos 2 primeiros semestres (continuação).

(b) Procura da 2ª variável (ects_reprov_s/?).

		RF		SVM		ANN		
variável	AUC médio	mtry	AUC	cost	AUC	size	decay	AUC
cod_prof_mae	0.8740	1	0.8564	2^{9}	0.8818	5	$10^{-1/3}$	0.8839
n10_11_acesso	0.8724	1	0.8525	2^{1}	0.8817	10	$10^{-8/3}$	0.8830
n12_acesso	0.8714	1	0.8506	2^{1}	0.8817	20	$10^{-8/3}$	0.8818
media_s	0.8712	1	0.8504	2^{-5}	0.8759	20	$10^{-5/3}$	0.8873
pi_acesso	0.8710	1	0.8537	2^{3}	0.8793	10	$10^{-6/3}$	0.8798
sit_prof_pai	0.8704	1	0.8399	2^{15}	0.8831	2	$10^{-4/3}$	0.8882
max_s	0.8696	1	0.8458	2^{-5}	0.8744	2	$10^{-6/3}$	0.8886
nivel_esc_pai	0.8691	2	0.8495	2^{11}	0.8758	2	$10^{0/3}$	0.8819
sit_prof_mae	0.8682	1	0.8383	2^{11}	0.8831	5	$10^{-3/3}$	0.8834
vd12_s	0.8681	1	0.8423	2^{1}	0.8808	2	$10^{-4/3}$	0.8812
fase	0.8677	1	0.8399	2^{-5}	0.8804	10	$10^{-7/3}$	0.8828
ordem_acesso	0.8675	1	0.8390	2^{-3}	0.8819	1	$10^{-8/3}$	0.8814
cod_prof_pai	0.8669	2	0.8445	2^{7}	0.8781	2	$10^{-2/3}$	0.8781
media_acesso	0.8666	1	0.8437	2^{-1}	0.8774	20	$10^{-8/3}$	0.8789
opcao_acesso	0.8651	1	0.8317	2^{1}	0.8766	5	$10^{-8/3}$	0.8870
nivel_esc_mae	0.8650	2	0.8385	2^{15}	0.8844	2	$10^{0/3}$	0.8721
idade	0.8639	1	0.8300	2^{-3}	0.8785	10	$10^{-8/3}$	0.8831
navalr_s	0.8616	2	0.8218	2^{11}	0.8814	1	$10^{-8/3}$	0.8816
nucr_s	0.8614	1	0.8229	2^{1}	0.8794	5	$10^{-8/3}$	0.8819
sit_prof_aluno	0.8614	1	0.8229	2^{5}	0.8810	5	$10^{-4/3}$	0.8803
ects_cred_tx	0.8607	2	0.8323	2^{-5}	0.8637	10	$10^{-7/3}$	0.8861
sexo	0.8605	1	0.8244	2^{3}	0.8732	5	$10^{-8/3}$	0.8840
dist_n	0.8593	1	0.8178	2^{-5}	0.8817	2	$10^{-3/3}$	0.8785
dist	0.8589	1	0.8135	2^{-5}	0.8828	2	$10^{-2/3}$	0.8805
nacionalidade	0.8587	2	0.8116	2^{9}	0.8837	1	$10^{0/3}$	0.8809
ano_curricular_s	0.8586	2	0.8172	2^{-5}	0.8793	1	$10^{-8/3}$	0.8793
nuca_s	0.8584	2	0.8054	2^{-5}	0.8828	10	$10^{-7/3}$	0.8868
bolseiro_s	0.8581	1	0.8141	2^{-5}	0.8787	20	$10^{-8/3}$	0.8816
dir_associativo_s	0.8572	1	0.8117	2^{9}	0.8800	10	$10^{-5/3}$	0.8798
cod_estatuto1_s	0.8571	2	0.8095	2^{15}	0.8836	10	$10^{-7/3}$	0.8782
cod_estatuto2_s	0.8570	2	0.8095	2^{15}	0.8836	1	$10^{-8/3}$	0.8781
ects_aprov_s	0.8565	1	0.8081	2^{-5}	0.8795	2	$10^{-8/3}$	0.8818
deslocado	0.8559	2	0.8198	2^{-5}	0.8715	2	$10^{-4/3}$	0.8764
cod_prof_aluno	0.8553	2	0.8073	2^{7}	0.8794	2	$10^{0/3}$	0.8793
ects_curso	0.8543	2	0.8170	2^{9}	0.8645	2	$10^{-3/3}$	0.8815
cod_freq_tipo1_s	0.8528	1	0.7997	2^{15}	0.8791	20	$10^{-6/3}$	0.8796
cod_freq_tipo2_s	0.8528	1	0.7997	2^{15}	0.8791	20	$10^{-6/3}$	0.8796
cod_curso	0.8515	1	0.7841	2^{-1}	0.8798	2	$10^{-2/3}$	0.8907
cod_escola	0.8442	1	0.8033	2^{-5}	0.8458	20	$10^{-6/3}$	0.8835
min_s	0.8437	2	0.7998	2^{-5}	0.8470	20	$10^{-5/3}$	0.8844

Tabela E.2: Aplicação do método *forward search* ao *dataset* dos 2 primeiros semestres (continuação).

(c) Procura da 3ª variável (ects_reprov_s/cod_prof_mae/?).

		RF		SVM		ANN		
variável	AUC médio	mtry	AUC	cost	AUC	size	decay	AUC
n10_11_acesso	0.8797	1	0.8749	2^{-1}	0.8811	5	$10^{-1/3}$	0.8832
n12_acesso	0.8787	1	0.8721	2^{7}	0.8802	10	$10^{-7/3}$	0.8838
vd12_s	0.8784	1	0.8625	2^{9}	0.8849	1	$10^{0/3}$	0.8879
sit_prof_aluno	0.8775	2	0.8624	2^{15}	0.8823	2	$10^{0/3}$	0.8876
deslocado	0.8772	3	0.8684	2^{13}	0.8796	20	$10^{-1/3}$	0.8836
idade	0.8770	1	0.8632	2^{7}	0.8827	5	$10^{-1/3}$	0.8850
nuca_s	0.8767	1	0.8597	2^{-3}	0.8841	2	$10^{-1/3}$	0.8862
pi_acesso	0.8764	1	0.8685	2^{-1}	0.8788	5	$10^{-1/3}$	0.8818
sit_prof_pai	0.8763	1	0.8567	2^{11}	0.8842	2	$10^{0/3}$	0.8878
nucr_s	0.8762	1	0.8616	2^{11}	0.8841	2	$10^{-3/3}$	0.8828
navalr_s	0.8760	2	0.8640	2^{1}	0.8811	5	$10^{-1/3}$	0.8829
max_s	0.8758	1	0.8561	2^{11}	0.8832	5	$10^{-1/3}$	0.8881
sit_prof_mae	0.8756	2	0.8602	2^{11}	0.8791	2	$10^{0/3}$	0.8876
opcao_acesso	0.8749	1	0.8670	2^{7}	0.8772	2	$10^{-1/3}$	0.8806
sexo	0.8742	1	0.8536	2^{11}	0.8821	20	$10^{-1/3}$	0.8870
nivel_esc_pai	0.8736	1	0.8617	2^{9}	0.8737	10	$10^{-1/3}$	0.8855
nacionalidade	0.8727	2	0.8487	2^{3}	0.8848	2	$10^{-1/3}$	0.8845
ano_curricular_s	0.8726	2	0.8517	2^{9}	0.8818	10	$10^{-1/3}$	0.8842
ects_cred_tx	0.8725	2	0.8688	2^{13}	0.8653	2	$10^{-1/3}$	0.8834
ects_aprov_s	0.8725	1	0.8533	2^{-3}	0.8815	10	$10^{-1/3}$	0.8827
media_acesso	0.8722	1	0.8538	2^{5}	0.8804	20	$10^{-4/3}$	0.8825
bolseiro_s	0.8721	1	0.8507	2^{7}	0.8810	5	$10^{-1/3}$	0.8845
cod_prof_aluno	0.8715	2	0.8485	2^{15}	0.8826	10	$10^{-1/3}$	0.8834
fase	0.8714	1	0.8489	2^{7}	0.8817	2	$10^{0/3}$	0.8837
ordem_acesso	0.8714	1	0.8469	2^{7}	0.8823	10	$10^{-1/3}$	0.8848
media_s	0.8713	1	0.8466	2^{-5}	0.8790	10	$10^{-1/3}$	0.8885
cod_estatuto2_s	0.8709	2	0.8509	2^{11}	0.8801	2	$10^{-1/3}$	0.8817
cod_prof_pai	0.8707	1	0.8511	2^{15}	0.8797	2	$10^{0/3}$	0.8815
cod_estatuto1_s	0.8707	2	0.8509	2^{11}	0.8801	5	$10^{-1/3}$	0.8812
dir_associativo_s	0.8706	1	0.8438	2^{9}	0.8834	5	$10^{-1/3}$	0.8846
nivel_esc_mae	0.8699	1	0.8639	2^{15}	0.8710	2	$10^{0/3}$	0.8749
min_s	0.8695	2	0.8407	2^{9}	0.8817	2	$10^{-1/3}$	0.8860
ects_curso	0.8684	2	0.8432	2^{7}	0.8780	20	$10^{-1/3}$	0.8839
dist_n	0.8653	2	0.8312	2^{3}	0.8815	20	$10^{-1/3}$	0.8833
cod_freq_tipo1_s	0.8647	1	0.8287	2^{15}	0.8814	10	$10^{-1/3}$	0.8840
cod_freq_tipo2_s	0.8647	1	0.8287	2^{15}	0.8814	10	$10^{-1/3}$	0.8840
dist	0.8634	2	0.8262	2^{7}	0.8818	2	$10^{0/3}$	0.8823
cod_escola	0.8613	2	0.8514	2^{5}	0.8471	5	$10^{-3/3}$	0.8853
cod_curso	0.7980	1	0.8456	2^{-5}	0.6600	2	$10^{0/3}$	0.8882

Tabela E.2: Aplicação do método *forward search* ao *dataset* dos 2 primeiros semestres (continuação).

(d) Procura da 4ª variável (ects_reprov_s/cod_prof_mae/n10_11_acesso/?).

		RF		SVM		ANN		
variável	AUC médio	mtry	AUC	cost	AUC	size	decay	AUC
idade	0.8855	1	0.8904	2^{5}	0.8818	5	$10^{-1/3}$	0.8842
nucr_s	0.8853	1	0.8877	2^{7}	0.8817	2	$10^{-5/3}$	0.8865
nuca_s	0.8846	1	0.8809	2^{-3}	0.8837	20	$10^{-4/3}$	0.8892
media_s	0.8843	1	0.8858	2^{-5}	0.8793	10	$10^{-4/3}$	0.8877
max_s	0.8832	1	0.8817	2^{11}	0.8810	20	$10^{-1/3}$	0.8870
navalr_s	0.8830	1	0.8867	2^{1}	0.8803	5	$10^{-1/3}$	0.8820
ects_aprov_s	0.8827	1	0.8834	2^{-1}	0.8810	10	$10^{-4/3}$	0.8836
sit_prof_aluno	0.8815	1	0.8770	2^{1}	0.8816	2	$10^{0/3}$	0.8860
pi_acesso	0.8815	1	0.8809	2^{-1}	0.8788	20	$10^{-4/3}$	0.8849
nacionalidade	0.8813	1	0.8754	2^{5}	0.8845	2	$10^{-1/3}$	0.8842
dir_associativo_s	0.8813	2	0.8763	2^{1}	0.8818	20	$10^{-5/3}$	0.8858
ano_curricular_s	0.8808	1	0.8790	2^{5}	0.8801	10	$10^{-1/3}$	0.8832
bolseiro_s	0.8801	1	0.8765	2^{3}	0.8793	2	$10^{-5/3}$	0.8845
opcao_acesso	0.8797	2	0.8824	2^{13}	0.8758	20	$10^{-1/3}$	0.8810
n12_acesso	0.8796	1	0.8706	2^{1}	0.8816	2	$10^{-5/3}$	0.8867
vd12_s	0.8788	1	0.8666	2^{9}	0.8832	10	$10^{-1/3}$	0.8866
cod_prof_aluno	0.8788	2	0.8733	2^{1}	0.8804	10	$10^{-1/3}$	0.8827
cod_freq_tipo1_s	0.8788	2	0.8734	2^{3}	0.8795	5	$10^{-1/3}$	0.8834
cod_freq_tipo2_s	0.8788	2	0.8734	2^{3}	0.8795	5	$10^{-1/3}$	0.8834
deslocado	0.8787	1	0.8785	2^{5}	0.8755	20	$10^{-1/3}$	0.8821
cod_estatuto1_s	0.8786	2	0.8743	2^{15}	0.8809	5	$10^{-1/3}$	0.8805
cod_estatuto2_s	0.8785	2	0.8743	2^{15}	0.8809	5	$10^{-1/3}$	0.8803
sit_prof_mae	0.8781	2	0.8727	2^{9}	0.8751	1	$10^{-1/3}$	0.8866
ects_cred_tx	0.8778	2	0.8831	2^{15}	0.8682	2	$10^{-1/3}$	0.8822
media_acesso	0.8777	1	0.8706	2^{-1}	0.8791	20	$10^{-4/3}$	0.8834
fase	0.8775	2	0.8680	2^{7}	0.8807	2	$10^{-1/3}$	0.8839
sexo	0.8775	2	0.8708	2^{13}	0.8770	10	$10^{-1/3}$	0.8848
nivel_esc_pai	0.8770	1	0.8699	2^{11}	0.8770	5	$10^{0/3}$	0.8843
sit_prof_pai	0.8765	2	0.8689	2^{7}	0.8737	1	$10^{0/3}$	0.8871
min_s	0.8765	2	0.8693	2^{15}	0.8740	20	$10^{-4/3}$	0.8860
ordem_acesso	0.8738	1	0.8547	2^{11}	0.8830	2	$10^{-1/3}$	0.8837
ects_curso	0.8727	2	0.8606	2^{1}	0.8746	20	$10^{-1/3}$	0.8829
nivel_esc_mae	0.8723	2	0.8685	2^{11}	0.8737	2	$10^{0/3}$	0.8746
dist_n	0.8717	1	0.8506	2^{1}	0.8810	20	$10^{-1/3}$	0.8835
cod_prof_pai	0.8715	1	0.8601	2^{5}	0.8729	2	$10^{0/3}$	0.8817
dist	0.8713	1	0.8510	2^{5}	0.8811	2	$10^{0/3}$	0.8818
cod_escola	0.8639	2	0.8798	2^{9}	0.8286	10	$10^{-2/3}$	0.8834
cod_curso	0.8048	2	0.8589	2^{-5}	0.6678	2	$10^{0/3}$	0.8875

Tabela E.2: Aplicação do método *forward search* ao *dataset* dos 2 primeiros semestres (continuação).

(e) Procura da 5ª variável (ects_reprov_s/cod_prof_mae/n10_11_acesso/idade/?).

		RF		SVM		ANN		
variável	AUC médio	mtry	AUC	cost	AUC	size	decay	AUC
media_s	0.8883	1	0.8897	2^{13}	0.8821	2	$10^{-6/3}$	0.8931
nuca_s	0.8874	1	0.8880	2^{-3}	0.8844	20	$10^{-3/3}$	0.8897
navalr_s	0.8866	1	0.8936	2^{5}	0.8820	10	$10^{-1/3}$	0.8842
nacionalidade	0.8865	2	0.8857	2^{5}	0.8863	20	$10^{-4/3}$	0.8875
nucr_s	0.8854	1	0.8892	2^{11}	0.8832	10	$10^{-4/3}$	0.8838
ects_aprov_s	0.8852	1	0.8892	2^{-3}	0.8816	20	$10^{-3/3}$	0.8849
sit_prof_aluno	0.8852	2	0.8864	2^{5}	0.8824	2	$10^{0/3}$	0.8868
cod_freq_tipo1_s	0.8845	2	0.8876	2^{5}	0.8813	5	$10^{-1/3}$	0.8846
cod_freq_tipo2_s	0.8845	2	0.8876	2^{5}	0.8813	5	$10^{-1/3}$	0.8846
ects_cred_tx	0.8844	1	0.8953	2^{13}	0.8680	2	$10^{-6/3}$	0.8900
max_s	0.8841	1	0.8851	2^{13}	0.8790	20	$10^{-1/3}$	0.8882
dir_associativo_s	0.8841	2	0.8851	2^{3}	0.8825	5	$10^{-1/3}$	0.8845
cod_prof_aluno	0.8839	1	0.8856	2^{5}	0.8822	5	$10^{-1/3}$	0.8840
min_s	0.8831	2	0.8825	2^{11}	0.8805	2	$10^{-1/3}$	0.8861
opcao_acesso	0.8827	2	0.8884	2^{7}	0.8772	20	$10^{-1/3}$	0.8826
ano_curricular_s	0.8824	2	0.8795	2^{5}	0.8823	5	$10^{-1/3}$	0.8853
cod_estatuto2_s	0.8823	2	0.8855	2^{5}	0.8790	10	$10^{-3/3}$	0.8825
cod_estatuto1_s	0.8821	2	0.8855	2^{5}	0.8790	5	$10^{-1/3}$	0.8817
vd12_s	0.8819	1	0.8721	2^{9}	0.8856	10	$10^{-1/3}$	0.8880
bolseiro_s	0.8819	2	0.8799	2^{3}	0.8807	2	$10^{-4/3}$	0.8851
fase	0.8818	2	0.8783	2^{11}	0.8828	5	$10^{-1/3}$	0.8843
deslocado	0.8817	2	0.8836	2^{15}	0.8776	20	$10^{-1/3}$	0.8838
sit_prof_mae	0.8812	3	0.8812	2^{3}	0.8754	1	$10^{-1/3}$	0.8871
pi_acesso	0.8812	3	0.8795	2^{5}	0.8780	10	$10^{-4/3}$	0.8860
nivel_esc_pai	0.8804	3	0.8806	2^{9}	0.8756	5	$10^{0/3}$	0.8849
sit_prof_pai	0.8801	3	0.8768	2^{1}	0.8755	1	$10^{0/3}$	0.8880
n12_acesso	0.8801	1	0.8737	2^{5}	0.8820	5	$10^{-1/3}$	0.8844
media_acesso	0.8799	1	0.8756	2^{-1}	0.8799	5	$10^{-4/3}$	0.8842
ects_curso	0.8795	2	0.8772	2^{13}	0.8778	20	$10^{-1/3}$	0.8836
sexo	0.8794	1	0.8727	2^{15}	0.8795	20	$10^{-1/3}$	0.8861
cod_prof_pai	0.8783	2	0.8739	2^{15}	0.8786	2	$10^{0/3}$	0.8825
ordem_acesso	0.8777	2	0.8630	2^{9}	0.8845	10	$10^{-3/3}$	0.8857
dist_n	0.8756	2	0.8587	2^{3}	0.8827	20	$10^{-1/3}$	0.8852
dist	0.8747	1	0.8587	2^{5}	0.8825	10	$10^{0/3}$	0.8829
nivel_esc_mae	0.8737	1	0.8715	2^{9}	0.8742	10	$10^{0/3}$	0.8755
cod_escola	0.8284	1	0.8914	2^{-5}	0.7061	5	$10^{-4/3}$	0.8878
cod_curso	0.8086	2	0.8754	2^{-3}	0.6627	5	$10^{-1/3}$	0.8878

Tabela E.2: Aplicação do método *forward search* ao *dataset* dos 2 primeiros semestres (continuação).

(f) Procura da 6ª variável (ects_reprov_s/cod_prof_mae/n10_11_acesso/idade/media_s/?).

		RF		SVM		ANN		
variável	AUC médio	mtry	AUC	cost	AUC	size	decay	AUC
vd12_s	0.8899	2	0.8867	2^{15}	0.8894	1	$10^{-5/3}$	0.8935
nivel_esc_pai	0.8898	2	0.8973	2^{13}	0.8819	1	$10^{-6/3}$	0.8902
sit_prof_aluno	0.8898	1	0.8930	2^{11}	0.8857	5	$10^{0/3}$	0.8907
nucr_s	0.8893	1	0.8960	2^{11}	0.8827	1	$10^{-7/3}$	0.8892
sit_prof_mae	0.8889	1	0.8917	2^{11}	0.8827	1	$10^{-2/3}$	0.8922
nuca_s	0.8887	1	0.8891	2^{-5}	0.8878	5	$10^{-4/3}$	0.8893
ects_cred_tx	0.8887	2	0.8979	2^{-5}	0.8787	2	$10^{-3/3}$	0.8893
navalr_s	0.8883	1	0.8946	2^{-5}	0.8816	1	$10^{-6/3}$	0.8887
fase	0.8883	1	0.8913	2^{13}	0.8841	2	$10^{0/3}$	0.8895
nacionalidade	0.8881	2	0.8872	2^{11}	0.8872	2	$10^{-1/3}$	0.8900
sit_prof_pai	0.8879	1	0.8866	2^{13}	0.8845	1	$10^{-1/3}$	0.8925
dir_associativo_s	0.8874	1	0.8872	2^{13}	0.8846	10	$10^{-4/3}$	0.8903
n12_acesso	0.8873	2	0.8884	2^{13}	0.8846	5	$10^{-3/3}$	0.8890
cod_prof_aluno	0.8872	2	0.8880	2^{11}	0.8851	10	$10^{-1/3}$	0.8885
ano_curricular_s	0.8868	2	0.8912	2^{-5}	0.8800	10	$10^{-1/3}$	0.8893
deslocado	0.8868	2	0.8876	2^{11}	0.8845	20	$10^{-1/3}$	0.8883
bolseiro_s	0.8868	1	0.8904	2^{13}	0.8811	10	$10^{-1/3}$	0.8887
cod_estatuto1_s	0.8867	2	0.8908	2^{13}	0.8819	10	$10^{-4/3}$	0.8873
pi_acesso	0.8867	1	0.8893	2^{11}	0.8832	20	$10^{-1/3}$	0.8875
ects_aprov_s	0.8865	1	0.8901	2^{-5}	0.8838	10	$10^{-4/3}$	0.8855
cod_estatuto2_s	0.8865	2	0.8908	2^{13}	0.8819	20	$10^{-1/3}$	0.8867
cod_freq_tipo1_s	0.8863	1	0.8865	2^{15}	0.8840	10	$10^{-1/3}$	0.8883
cod_freq_tipo2_s	0.8863	1	0.8865	2^{15}	0.8840	10	$10^{-1/3}$	0.8883
min_s	0.8862	1	0.8839	2^{13}	0.8856	2	$10^{-6/3}$	0.8890
opcao_acesso	0.8860	2	0.8955	2^{-5}	0.8759	20	$10^{-1/3}$	0.8867
ects_curso	0.8859	2	0.8851	2^{13}	0.8845	20	$10^{-1/3}$	0.8881
ordem_acesso	0.8855	2	0.8780	2^{13}	0.8871	1	$10^{-5/3}$	0.8912
sexo	0.8855	3	0.8897	2^{-5}	0.8759	10	$10^{-1/3}$	0.8908
cod_prof_pai	0.8850	1	0.8881	2^{9}	0.8791	2	$10^{0/3}$	0.8879
nivel_esc_mae	0.8835	2	0.8878	2^{15}	0.8817	2	$10^{0/3}$	0.8809
max_s	0.8828	1	0.8825	2^{11}	0.8772	20	$10^{-1/3}$	0.8888
media_acesso	0.8822	2	0.8792	2^{-5}	0.8796	20	$10^{-1/3}$	0.8878
dist	0.8816	2	0.8729	2^{11}	0.8839	1	$10^{0/3}$	0.8882
dist_n	0.8798	1	0.8723	2^{-5}	0.8803	20	$10^{-1/3}$	0.8868
cod_escola	0.8732	1	0.8977	2^{-5}	0.8319	5	$10^{-2/3}$	0.8900
cod_curso	0.8180	1	0.8837	2^{-5}	0.6772	2	$10^{0/3}$	0.8929

Tabela E.2: Aplicação do método *forward search* ao *dataset* dos 2 primeiros semestres (continuação).

(g) Procura da 7ª variável (ects_reprov_s/cod_prof_mae/n10_11_acesso/idade/media_s/vd12_s/?).

		RF		SVM		ANN		
variável	AUC médio	mtry	AUC	cost	AUC	size	decay	AUC
sit_prof_mae	0.8909	1	0.8934	2^{11}	0.8844	1	$10^{-1/3}$	0.8950
sit_prof_aluno	0.8908	1	0.8921	2^{11}	0.8863	1	$10^{-1/3}$	0.8939
nacionalidade	0.8906	1	0.8880	2^{13}	0.8900	1	$10^{-4/3}$	0.8939
ano_curricular_s	0.8906	1	0.8915	2^{11}	0.8869	1	$10^{-5/3}$	0.8934
fase	0.8905	2	0.8864	2^{15}	0.8909	1	$10^{-6/3}$	0.8943
ects_cred_tx	0.8905	1	0.9003	2^{-5}	0.8798	5	$10^{-2/3}$	0.8915
sexo	0.8903	3	0.8875	2^{15}	0.8889	20	$10^{-1/3}$	0.8944
nivel_esc_pai	0.8901	2	0.8936	2^{13}	0.8818	1	$10^{-6/3}$	0.8947
n12_acesso	0.8895	1	0.8863	2^{11}	0.8871	2	$10^{-5/3}$	0.8952
cod_estatuto1_s	0.8894	1	0.8864	2^{13}	0.8875	2	$10^{-5/3}$	0.8944
navalr_s	0.8893	1	0.8875	2^{13}	0.8872	1	$10^{-5/3}$	0.8934
nucr_s	0.8893	1	0.8860	2^{11}	0.8868	2	$10^{-8/3}$	0.8950
opcao_acesso	0.8892	1	0.8911	2^{11}	0.8840	2	$10^{-3/3}$	0.8925
sit_prof_pai	0.8889	1	0.8874	2^{13}	0.8831	1	$10^{-1/3}$	0.8961
cod_prof_aluno	0.8888	1	0.8847	2^{13}	0.8892	20	$10^{-1/3}$	0.8923
deslocado	0.8886	2	0.8870	2^{11}	0.8864	1	$10^{-1/3}$	0.8924
cod_estatuto2_s	0.8885	1	0.8864	2^{13}	0.8875	2	$10^{-8/3}$	0.8915
dir_associativo_s	0.8882	1	0.8888	2^{-5}	0.8809	2	$10^{-5/3}$	0.8948
pi_acesso	0.8882	2	0.8846	2^{15}	0.8889	10	$10^{-1/3}$	0.8909
bolseiro_s	0.8880	2	0.8843	2^{11}	0.8866	1	$10^{-5/3}$	0.8930
ordem_acesso	0.8874	1	0.8797	2^{15}	0.8871	1	$10^{-5/3}$	0.8954
media_acesso	0.8870	2	0.8884	2^{15}	0.8808	20	$10^{-1/3}$	0.8918
ects_aprov_s	0.8869	1	0.8884	2^{-5}	0.8841	5	$10^{-1/3}$	0.8883
cod_freq_tipo1_s	0.8869	2	0.8842	2^{13}	0.8827	2	$10^{-4/3}$	0.8937
cod_freq_tipo2_s	0.8869	2	0.8842	2^{13}	0.8827	2	$10^{-4/3}$	0.8937
min_s	0.8868	1	0.8835	2^{15}	0.8848	1	$10^{-5/3}$	0.8921
nuca_s	0.8865	1	0.8805	2^{-5}	0.8881	10	$10^{-1/3}$	0.8909
nivel_esc_mae	0.8858	1	0.8928	2^{11}	0.8800	2	$10^{0/3}$	0.8845
ects_curso	0.8855	2	0.8820	2^{15}	0.8821	1	$10^{-5/3}$	0.8924
cod_prof_pai	0.8854	3	0.8859	2^{-5}	0.8784	2	$10^{0/3}$	0.8921
dist	0.8818	2	0.8725	2^{15}	0.8819	1	$10^{0/3}$	0.8908
max_s	0.8814	1	0.8804	2^{-5}	0.8715	20	$10^{-1/3}$	0.8922
dist_n	0.8802	2	0.8709	2^{11}	0.8808	5	$10^{0/3}$	0.8890
cod_escola	0.8763	2	0.8904	2^{-5}	0.8463	10	$10^{-1/3}$	0.8923
cod_curso	0.8417	1	0.8803	2^{3}	0.7486	1	$10^{0/3}$	0.8962

Tabela E.2: Aplicação do método *forward search* ao *dataset* dos 2 primeiros semestres (continuação).

(h) Procura da 8ª variável (ects_reprov_s/cod_prof_mae/n10_11_acesso/idade/media_s/vd12_s/sit_prof_mae/?).

		RF		SVM		ANN		
variável	AUC médio	mtry	AUC	cost	AUC	size	decay	AUC
sit_prof_aluno	0.8921	2	0.8952	2^{9}	0.8846	2	$10^{-1/3}$	0.8964
deslocado	0.8919	2	0.8940	2^{11}	0.8873	1	$10^{-1/3}$	0.8945
navalr_s	0.8917	1	0.8971	2^{11}	0.8832	1	$10^{-8/3}$	0.8949
bolseiro_s	0.8913	2	0.8936	2^{13}	0.8855	1	$10^{-6/3}$	0.8949
nivel_esc_pai	0.8910	1	0.8925	2^{9}	0.8823	2	$10^{-8/3}$	0.8981
fase	0.8906	2	0.8921	2^{11}	0.8839	1	$10^{-7/3}$	0.8958
ects_curso	0.8902	2	0.8883	2^{15}	0.8881	20	$10^{0/3}$	0.8941
dir_associativo_s	0.8900	2	0.8915	2^{11}	0.8842	1	$10^{-6/3}$	0.8944
cod_freq_tipo1_s	0.8899	2	0.8910	2^{11}	0.8836	1	$10^{-5/3}$	0.8951
cod_freq_tipo2_s	0.8899	2	0.8910	2^{11}	0.8836	1	$10^{-5/3}$	0.8951
nuca_s	0.8899	1	0.8906	2^{-3}	0.8855	20	$10^{0/3}$	0.8935
opcao_acesso	0.8896	3	0.8919	2^{11}	0.8854	2	$10^{-1/3}$	0.8916
nacionalidade	0.8896	4	0.8873	2^{11}	0.8852	1	$10^{-3/3}$	0.8964
n12_acesso	0.8896	1	0.8955	2^{-5}	0.8784	1	$10^{-1/3}$	0.8950
pi_acesso	0.8896	2	0.8980	2^{15}	0.8768	1	$10^{-8/3}$	0.8941
ects_cred_tx	0.8891	2	0.8948	2^{-5}	0.8807	2	$10^{-1/3}$	0.8920
ano_curricular_s	0.8891	2	0.8910	2^{15}	0.8813	1	$10^{-1/3}$	0.8950
media_acesso	0.8890	1	0.8892	2^{11}	0.8834	1	$10^{-1/3}$	0.8944
sit_prof_pai	0.8889	1	0.8832	2^{13}	0.8860	1	$10^{0/3}$	0.8975
cod_estatuto1_s	0.8887	2	0.8900	2^{15}	0.8829	20	$10^{0/3}$	0.8932
cod_estatuto2_s	0.8887	2	0.8900	2^{15}	0.8829	20	$10^{0/3}$	0.8932
nivel_esc_mae	0.8887	2	0.8960	2^{9}	0.8826	1	$10^{0/3}$	0.8875
ordem_acesso	0.8885	2	0.8829	2^{9}	0.8851	1	$10^{-7/3}$	0.8976
cod_prof_aluno	0.8883	1	0.8876	2^{9}	0.8830	1	$10^{-1/3}$	0.8944
ects_aprov_s	0.8880	1	0.8924	2^{-5}	0.8805	20	$10^{0/3}$	0.8911
min_s	0.8878	2	0.8901	2^{13}	0.8780	1	$10^{-8/3}$	0.8953
sexo	0.8873	4	0.8847	2^{11}	0.8805	20	$10^{0/3}$	0.8968
nucr_s	0.8873	1	0.8861	2^{15}	0.8803	1	$10^{-6/3}$	0.8955
cod_prof_pai	0.8867	3	0.8862	2^{9}	0.8787	1	$10^{0/3}$	0.8951
max_s	0.8865	2	0.8840	2^{13}	0.8809	1	$10^{0/3}$	0.8944
dist	0.8849	3	0.8751	2^{11}	0.8875	1	$10^{0/3}$	0.8922
dist_n	0.8845	2	0.8766	2^{11}	0.8848	1	$10^{0/3}$	0.8920
cod_escola	0.8800	2	0.8958	2^{-5}	0.8518	20	$10^{-1/3}$	0.8922
cod_curso	0.8387	1	0.8781	2^{9}	0.7416	2	$10^{0/3}$	0.8964

Tabela E.2: Aplicação do método *forward search* ao *dataset* dos 2 primeiros semestres (continuação).

(i) Procura da 9ª variável (ects_reprov_s/cod_prof_mae/n10_11_acesso/idade/media_s/vd12_s/ sit_prof_mae/sit_prof_aluno/?).

		RF		SVM		ANN		
variável	AUC médio	mtry	AUC	cost	AUC	size	decay	AUC
nacionalidade	0.8941	2	0.8928	2^{13}	0.8912	2	$10^{-1/3}$	0.8983
bolseiro_s	0.8939	2	0.8971	2^{15}	0.8882	1	$10^{-2/3}$	0.8964
nivel_esc_pai	0.8924	2	0.8944	2^{9}	0.8832	1	$10^{-1/3}$	0.8997
cod_estatuto1_s	0.8924	2	0.8978	2^{15}	0.8837	2	$10^{-2/3}$	0.8955
cod_estatuto2_s	0.8921	2	0.8978	2^{15}	0.8837	1	$10^{0/3}$	0.8947
nuca_s	0.8916	1	0.8969	2^{-5}	0.8836	1	$10^{0/3}$	0.8944
fase	0.8914	3	0.8925	2^{11}	0.8848	1	$10^{-1/3}$	0.8969
deslocado	0.8914	5	0.8894	2^{15}	0.8886	1	$10^{-1/3}$	0.8962
sexo	0.8911	1	0.8947	2^{15}	0.8809	20	$10^{0/3}$	0.8977
cod_prof_aluno	0.8910	2	0.8919	2^{9}	0.8844	1	$10^{-1/3}$	0.8968
cod_freq_tipo1_s	0.8910	2	0.8970	2^{9}	0.8806	1	$10^{-4/3}$	0.8954
cod_freq_tipo2_s	0.8910	2	0.8970	2^{9}	0.8806	1	$10^{-4/3}$	0.8954
nucr_s	0.8910	1	0.8936	2^{13}	0.8829	1	$10^{-5/3}$	0.8964
min_s	0.8906	1	0.8951	2^{15}	0.8810	1	$10^{-7/3}$	0.8959
cod_prof_pai	0.8905	2	0.8916	2^{9}	0.8839	1	$10^{-1/3}$	0.8962
n12_acesso	0.8903	2	0.8925	2^{11}	0.8817	2	$10^{-7/3}$	0.8968
pi_acesso	0.8903	2	0.8996	2^{11}	0.8762	1	$10^{-1/3}$	0.8951
ano_curricular_s	0.8901	2	0.8905	2^{9}	0.8835	1	$10^{-1/3}$	0.8963
dir_associativo_s	0.8899	2	0.8911	2^{15}	0.8828	2	$10^{0/3}$	0.8957
nivel_esc_mae	0.8893	2	0.9002	2^{9}	0.8798	1	$10^{0/3}$	0.8878
opcao_acesso	0.8892	2	0.8966	2^{9}	0.8785	2	$10^{0/3}$	0.8924
ects_aprov_s	0.8892	1	0.8970	2^{-5}	0.8778	2	$10^{-2/3}$	0.8928
ordem_acesso	0.8891	1	0.8833	2^{11}	0.8860	1	$10^{-4/3}$	0.8981
ects_cred_tx	0.8890	1	0.8959	2^{-3}	0.8774	10	$10^{-1/3}$	0.8938
navalr_s	0.8888	1	0.8908	2^{9}	0.8799	1	$10^{-2/3}$	0.8957
sit_prof_pai	0.8885	2	0.8848	2^{9}	0.8827	1	$10^{0/3}$	0.8979
cod_escola	0.8880	2	0.9006	2^{-5}	0.8682	2	$10^{-2/3}$	0.8953
ects_curso	0.8878	2	0.8871	2^{13}	0.8813	20	$10^{0/3}$	0.8951
media_acesso	0.8876	2	0.8882	2^{13}	0.8782	2	$10^{-1/3}$	0.8963
max_s	0.8868	5	0.8884	2^{9}	0.8757	1	$10^{-1/3}$	0.8962
dist	0.8843	3	0.8786	2^{11}	0.8805	1	$10^{0/3}$	0.8938
dist_n	0.8833	3	0.8801	2^{9}	0.8765	1	$10^{0/3}$	0.8933
cod_curso	0.8406	1	0.8832	2^{3}	0.7399	2	$10^{0/3}$	0.8987

Tabela E.2: Aplicação do método *forward search* ao *dataset* dos 2 primeiros semestres (continuação).

(j) Procura da 10ª variável (ects_reprov_s/cod_prof_mae/n10_11_acesso/idade/media_s/vd12_s/ sit_prof_mae/sit_prof_aluno/nacionalidade/?).

		RF		SVM		ANN		
variável	AUC médio	mtry	AUC	cost	AUC	size	decay	AUC
dir_associativo_s	0.8959	2	0.9007	2^{15}	0.8903	2	$10^{0/3}$	0.8967
deslocado	0.8958	3	0.9020	2^{9}	0.8885	2	$10^{-2/3}$	0.8970
nivel_esc_pai	0.8958	2	0.8974	2^{9}	0.8898	1	$10^{-2/3}$	0.9002
bolseiro_s	0.8955	2	0.8982	2^{11}	0.8912	1	$10^{-2/3}$	0.8971
cod_prof_pai	0.8942	2	0.8976	2^{9}	0.8889	1	$10^{0/3}$	0.8959
cod_prof_aluno	0.8940	2	0.8933	2^{11}	0.8909	1	$10^{0/3}$	0.8978
sexo	0.8936	2	0.8939	2^{15}	0.8881	10	$10^{0/3}$	0.8987
sit_prof_pai	0.8935	2	0.8901	2^{9}	0.8918	1	$10^{0/3}$	0.8988
ano_curricular_s	0.8929	3	0.8958	2^{9}	0.8846	2	$10^{-1/3}$	0.8983
fase	0.8928	3	0.8922	2^{15}	0.8881	1	$10^{-3/3}$	0.8982
navalr_s	0.8927	2	0.8933	2^{11}	0.8880	1	$10^{-1/3}$	0.8969
nivel_esc_mae	0.8927	2	0.9025	2^{13}	0.8885	1	$10^{0/3}$	0.8870
n12_acesso	0.8926	4	0.8945	2^{9}	0.8860	1	$10^{-1/3}$	0.8973
cod_freq_tipo1_s	0.8925	2	0.8964	2^{9}	0.8845	1	$10^{-2/3}$	0.8967
cod_freq_tipo2_s	0.8925	2	0.8964	2^{9}	0.8845	1	$10^{-2/3}$	0.8967
ects_cred_tx	0.8925	3	0.8997	2^{11}	0.8822	10	$10^{-1/3}$	0.8957
cod_estatuto1_s	0.8922	3	0.8949	2^{9}	0.8864	1	$10^{0/3}$	0.8955
cod_estatuto2_s	0.8922	3	0.8949	2^{9}	0.8864	1	$10^{0/3}$	0.8955
nucr_s	0.8921	2	0.8933	2^{15}	0.8858	1	$10^{-3/3}$	0.8971
ordem_acesso	0.8918	3	0.8831	2^{15}	0.8923	1	$10^{-3/3}$	0.8998
ects_curso	0.8914	2	0.8913	2^{15}	0.8870	1	$10^{0/3}$	0.8959
nuca_s	0.8910	1	0.8910	2^{-5}	0.8872	20	$10^{0/3}$	0.8947
opcao_acesso	0.8904	3	0.8975	2^{11}	0.8803	5	$10^{0/3}$	0.8933
media_acesso	0.8902	2	0.8906	2^{15}	0.8833	1	$10^{-1/3}$	0.8967
pi_acesso	0.8901	1	0.8956	2^{9}	0.8788	1	$10^{-1/3}$	0.8958
max_s	0.8900	3	0.8886	2^{11}	0.8843	1	$10^{0/3}$	0.8971
ects_aprov_s	0.8898	1	0.8908	2^{13}	0.8866	20	$10^{0/3}$	0.8920
dist	0.8898	2	0.8822	2^{11}	0.8928	1	$10^{0/3}$	0.8945
min_s	0.8891	2	0.8877	2^{9}	0.8828	1	$10^{-5/3}$	0.8969
cod_escola	0.8887	2	0.8986	2^{-5}	0.8722	10	$10^{-1/3}$	0.8953
dist_n	0.8854	2	0.8832	2^{-5}	0.8785	1	$10^{0/3}$	0.8945
cod_curso	0.8387	3	0.8881	2^{7}	0.7294	2	$10^{0/3}$	0.8986

Tabela E.2: Aplicação do método *forward search* ao *dataset* dos 2 primeiros semestres (continuação).

(k) Procura da 11ª variável (ects_reprov_s/cod_prof_mae/n10_11_acesso/idade/media_s/vd12_s/sit_prof_mae/sit_prof_aluno/nacionalidade/dir_associativo_s/?).

		RF		SVM		ANN		
variável	AUC médio	mtry	AUC	cost	AUC	size	decay	AUC
nivel_esc_pai	0.8975	2	0.9025	2^{11}	0.8906	1	$10^{0/3}$	0.8993
fase	0.8957	3	0.8982	2^{13}	0.8919	1	$10^{-1/3}$	0.8970
bolseiro_s	0.8953	2	0.9058	2^{9}	0.8838	1	$10^{-2/3}$	0.8963
nucr_s	0.8953	3	0.8990	2^{15}	0.8907	1	$10^{-3/3}$	0.8962
cod_prof_aluno	0.8950	2	0.9004	2^{11}	0.8876	1	$10^{0/3}$	0.8972
n12_acesso	0.8947	3	0.8989	2^{13}	0.8888	1	$10^{-1/3}$	0.8964
sexo	0.8944	2	0.8990	2^{9}	0.8858	20	$10^{0/3}$	0.8984
cod_prof_pai	0.8944	3	0.8950	2^{13}	0.8921	1	$10^{0/3}$	0.8961
ano_curricular_s	0.8940	2	0.8961	2^{11}	0.8894	1	$10^{-1/3}$	0.8964
navalr_s	0.8936	2	0.8918	2^{15}	0.8927	1	$10^{-1/3}$	0.8965
cod_estatuto2_s	0.8935	2	0.8985	2^{9}	0.8867	2	$10^{0/3}$	0.8952
cod_estatuto1_s	0.8935	2	0.8985	2^{9}	0.8867	2	$10^{0/3}$	0.8952
deslocado	0.8935	4	0.8973	2^{9}	0.8871	1	$10^{0/3}$	0.8960
sit_prof_pai	0.8932	2	0.8911	2^{9}	0.8910	1	$10^{0/3}$	0.8975
ects_cred_tx	0.8927	2	0.9007	2^{-3}	0.8818	10	$10^{-1/3}$	0.8957
nuca_s	0.8925	2	0.8947	2^{-3}	0.8880	20	$10^{0/3}$	0.8947
opcao_acesso	0.8921	3	0.8988	2^{13}	0.8847	5	$10^{0/3}$	0.8927
pi_acesso	0.8919	2	0.8940	2^{11}	0.8865	1	$10^{-1/3}$	0.8950
min_s	0.8918	3	0.8946	2^{9}	0.8847	1	$10^{0/3}$	0.8960
cod_freq_tipo1_s	0.8916	3	0.8943	2^{15}	0.8848	1	$10^{0/3}$	0.8958
cod_freq_tipo2_s	0.8916	3	0.8943	2^{15}	0.8848	1	$10^{0/3}$	0.8958
ects_curso	0.8912	3	0.8929	2^{11}	0.8853	20	$10^{0/3}$	0.8955
ordem_acesso	0.8912	2	0.8830	2^{9}	0.8917	1	$10^{-5/3}$	0.8988
media_acesso	0.8911	3	0.8917	2^{13}	0.8856	1	$10^{0/3}$	0.8961
nivel_esc_mae	0.8906	4	0.8965	2^{9}	0.8888	1	$10^{0/3}$	0.8865
max_s	0.8901	2	0.8919	2^{9}	0.8824	1	$10^{0/3}$	0.8961
cod_escola	0.8900	3	0.9020	2^{-5}	0.8721	5	$10^{-1/3}$	0.8959
ects_aprov_s	0.8884	2	0.8910	2^{11}	0.8816	2	$10^{-2/3}$	0.8925
dist	0.8877	2	0.8826	2^{13}	0.8862	1	$10^{0/3}$	0.8944
dist_n	0.8870	2	0.8804	2^{13}	0.8863	1	$10^{0/3}$	0.8943
cod_curso	0.8421	3	0.8903	2^{7}	0.7368	2	$10^{0/3}$	0.8993

Tabela E.2: Aplicação do método *forward search* ao *dataset* dos 2 primeiros semestres (continuação).

(l) Procura da 12ª variável (ects_reprov_s/cod_prof_mae/n10_11_acesso/idade/media_s/vd12_s/sit_prof_mae/sit_prof_aluno/nacionalidade/dir_associativo_s/nivel_esc_pai/?).

		RF		SVM		ANN		
variável	AUC médio	mtry	AUC	cost	AUC	size	decay	AUC
cod_prof_aluno	0.8977	2	0.9051	2^{11}	0.8883	1	$10^{0/3}$	0.8997
min_s	0.8975	2	0.9054	2^{9}	0.8884	1	$10^{0/3}$	0.8988
nucr_s	0.8972	2	0.9016	2^{11}	0.8914	1	$10^{-5/3}$	0.8986
sexo	0.8967	2	0.9028	2^{11}	0.8875	20	$10^{0/3}$	0.8999
n12_acesso	0.8961	3	0.9008	2^{13}	0.8882	1	$10^{0/3}$	0.8993
fase	0.8957	2	0.8953	2^{15}	0.8925	1	$10^{0/3}$	0.8992
ano_curricular_s	0.8956	2	0.8979	2^{9}	0.8896	1	$10^{0/3}$	0.8993
deslocado	0.8955	4	0.9000	2^{11}	0.8886	1	$10^{0/3}$	0.8979
cod_freq_tipo1_s	0.8955	2	0.8962	2^{15}	0.8918	1	$10^{0/3}$	0.8985
cod_freq_tipo2_s	0.8955	2	0.8962	2^{15}	0.8918	1	$10^{0/3}$	0.8985
cod_estatuto1_s	0.8950	3	0.8992	2^{9}	0.8875	1	$10^{0/3}$	0.8984
cod_estatuto2_s	0.8950	3	0.8992	2^{9}	0.8875	1	$10^{0/3}$	0.8982
nuca_s	0.8949	2	0.8978	2^{15}	0.8904	20	$10^{0/3}$	0.8966
ects_curso	0.8946	3	0.8972	2^{13}	0.8893	1	$10^{0/3}$	0.8974
opcao_acesso	0.8944	2	0.9022	2^{11}	0.8857	1	$10^{-4/3}$	0.8954
sit_prof_pai	0.8943	4	0.8975	2^{11}	0.8856	1	$10^{0/3}$	0.8998
navalr_s	0.8943	2	0.8931	2^{9}	0.8911	1	$10^{-1/3}$	0.8987
cod_prof_pai	0.8942	2	0.8941	2^{15}	0.8899	1	$10^{0/3}$	0.8986
ects_cred_tx	0.8941	3	0.9056	2^{-3}	0.8822	5	$10^{-1/3}$	0.8944
media_acesso	0.8940	2	0.8975	2^{9}	0.8860	1	$10^{0/3}$	0.8985
bolseiro_s	0.8938	5	0.9002	2^{9}	0.8828	1	$10^{-8/3}$	0.8986
ordem_acesso	0.8938	2	0.8881	2^{9}	0.8914	1	$10^{-6/3}$	0.9018
pi_acesso	0.8937	3	0.8997	2^{9}	0.8838	1	$10^{0/3}$	0.8975
nivel_esc_mae	0.8914	5	0.8990	2^{9}	0.8864	2	$10^{0/3}$	0.8888
ects_aprov_s	0.8914	2	0.8965	2^{15}	0.8836	20	$10^{0/3}$	0.8941
max_s	0.8911	3	0.8921	2^{11}	0.8823	1	$10^{0/3}$	0.8987
dist_n	0.8900	2	0.8853	2^{13}	0.8880	1	$10^{0/3}$	0.8966
dist	0.8899	2	0.8874	2^{9}	0.8859	1	$10^{0/3}$	0.8964
cod_escola	0.8866	4	0.9027	2^{-5}	0.8605	1	$10^{0/3}$	0.8966
cod_curso	0.8211	2	0.8927	2^{-1}	0.6710	1	$10^{0/3}$	0.8995

Tabela E.2: Aplicação do método *forward search* ao *dataset* dos 2 primeiros semestres (continuação).

(m) Procura da 13ª variável (ects_reprov_s/cod_prof_mae/n10_11_acesso/idade/media_s/vd12_s/sit_prof_mae/sit_prof_aluno/nacionalidade/dir_associativo_s/nivel_esc_pai/cod_prof_aluno/×).

		RF		SVM		ANN		
variável	AUC médio	mtry	AUC	cost	AUC	size	decay	AUC
min_s	0.8962	3	0.9045	2^{9}	0.8849	1	$10^{0/3}$	0.8993
n12_acesso	0.8961	5	0.8981	2^{11}	0.8907	1	$10^{0/3}$	0.8996
ano_curricular_s	0.8958	3	0.8987	2^{15}	0.8891	1	$10^{0/3}$	0.8997
fase	0.8958	4	0.8998	2^{9}	0.8884	1	$10^{0/3}$	0.8993
deslocado	0.8958	3	0.9024	2^{9}	0.8865	1	$10^{0/3}$	0.8984
navalr_s	0.8952	3	0.8962	2^{9}	0.8907	1	$10^{0/3}$	0.8986
cod_prof_pai	0.8946	3	0.8968	2^{9}	0.8884	1	$10^{0/3}$	0.8988
bolseiro_s	0.8946	3	0.9018	2^{11}	0.8836	1	$10^{-1/3}$	0.8983
nuca_s	0.8945	2	0.8972	2^{11}	0.8894	1	$10^{0/3}$	0.8969
nucr_s	0.8943	2	0.8981	2^{9}	0.8869	1	$10^{0/3}$	0.8978
cod_freq_tipo1_s	0.8941	2	0.8975	2^{9}	0.8854	1	$10^{0/3}$	0.8995
cod_freq_tipo2_s	0.8941	2	0.8975	2^{9}	0.8854	1	$10^{0/3}$	0.8995
sit_prof_pai	0.8938	2	0.8968	2^{11}	0.8844	1	$10^{0/3}$	0.9003
ects_aprov_s	0.8937	2	0.8990	2^{13}	0.8877	20	$10^{0/3}$	0.8944
opcao_acesso	0.8932	3	0.9045	2^{11}	0.8795	10	$10^{0/3}$	0.8957
cod_estatuto1_s	0.8931	2	0.8965	2^{9}	0.8841	1	$10^{0/3}$	0.8986
cod_estatuto2_s	0.8931	2	0.8965	2^{9}	0.8841	1	$10^{0/3}$	0.8986
ects_cred_tx	0.8927	4	0.9002	2^{-3}	0.8813	2	$10^{0/3}$	0.8966
max_s	0.8926	3	0.8981	2^{9}	0.8805	1	$10^{0/3}$	0.8992
sexo	0.8924	4	0.8986	2^{9}	0.8784	20	$10^{0/3}$	0.9003
pi_acesso	0.8923	4	0.8947	2^{11}	0.8843	1	$10^{0/3}$	0.8978
ordem_acesso	0.8917	3	0.8864	2^{11}	0.8880	1	$10^{0/3}$	0.9007
nivel_esc_mae	0.8917	4	0.9018	2^{9}	0.8841	2	$10^{0/3}$	0.8892
media_acesso	0.8916	2	0.8941	2^{13}	0.8818	1	$10^{0/3}$	0.8988
ects_curso	0.8915	2	0.8992	2^{11}	0.8768	1	$10^{0/3}$	0.8984
dist	0.8887	3	0.8852	2^{9}	0.8840	1	$10^{0/3}$	0.8969
dist_n	0.8871	2	0.8840	2^{11}	0.8801	1	$10^{0/3}$	0.8973
cod_escola	0.8863	2	0.9084	2^{5}	0.8526	5	$10^{0/3}$	0.8978
cod_curso	0.8229	3	0.8930	2^{-3}	0.6764	1	$10^{0/3}$	0.8994