PREDIÇÃO DO DESEMPENHO ACADÊMICO DE ALUNOS DA GRADUAÇÃO UTILIZANDO MINERAÇÃO DE DADOS

Laci Mary Barbosa Manhães
Instituto Noroeste Fluminense/Universidade Federal Fluminense – INFES/UFF
Av. João Jasbick, S/N – Aeroporto. Santo Antônio de Pádua – RJ - Brasil
mary_manhaes@id.uff.br

Sérgio Manuel Serra da Cruz
Departamento de Computação/Universidade Federal Rural do Rio de Janeiro – ICE/UFRRJ
BR-465, Km 7 – Sala 80 – P1 CEP. 23.897-000 Seropédica – RJ - Brasil
serra@ufrrj.br

RESUMO

As altas taxas de evasão e retenção dos cursos de graduação das universidades públicas federais é um problema multifacetado. Este trabalho apresenta a arquitetura WAVE baseada em Mineração de Dados Educacionais (EDM). Seu objetivo é fornecer aos gestores educacionais das universidades, não especialista em mineração de dados, uma abordagem que ofereça acompanhamento semestral sobre o desempenho acadêmico dos graduandos e a predição dos que estão em risco de abandonar o sistema. A arquitetura e os modelos preditivos foram avaliados através de diversos estudos de casos que utilizaram dados reais de inúmeros cursos de graduação da UFRJ, durante um período de 16 anos. A abordagem é uma das primeiras que utiliza apenas dados acadêmicos invariantes no tempo sem considerar nenhum tipo de dado socioeconômico. Os estudos de casos comparam o desempenho de 12 algoritmos classificadores com o objetivo de identificar aqueles com maior precisão na identificação de alunos em risco de falhar na formação acadêmica.

Palavra-chave: Evasão; Estudantes; Algoritmos; Classificadores; Mineração de Dados Educacionais.

ABSTRACT

The high dropout and retention rates of undergraduate courses of federal public universities is a current and multifaceted problem. This work presents the WAVE architecture based on Educational Data Mining (EDM). Its goal is to provide educational managers of universities, non-specialist in data mining, an approach to follow the performance of students and predict those that are at risk of leaving the system. The architecture and the predictive models were tested through case studies using real data from students of various courses of UFRJ over a period of 16 years. The approach is one of the first to use only academic data that do not vary over time without regarding to any given social or economic data. The case studies compare the performance of 12 classification algorithms in order to identify those with greater accuracy to predict students at risk of academic failure.

Keywords: Dropout; Students; Algorithms; Classification; Educational Data Mining.

Como Citar:
MANHAES, Laci Mary B.; SOBRENOME, Nome Autor 2. Predição do Desempenho Acadêmico de Alunos da Graduação Utilizando Mineração de Dados. *In*: SIMPÓSIO DE PESQUISA OPERACIONAL E LOGÍSTICA DA MARINHA, 19., 2019, Rio de Janeiro, RJ. **Anais** […]. Rio de Janeiro: Centro de Análises de Sistemas Navais, 2019.

1. INTRODUÇÃO

A evasão universitária é um problema que permeia as Instituições de Ensino Superior públicas e privadas brasileiras. Ela gera desperdícios financeiros, sociais e acadêmicos. A evasão ocorre em diferentes instituições de ensino no mundo e estudos apontam que este fenômeno ocorre em diversos cursos de graduação (BAKER e YACEF, 2009, BAKER et al. 2011, ROMERO e VENTURA, 2013, PAIVA, 2014).

No contexto das Instituições Federais de Ensino Superior (IFES), verifica-se uma intensificação dos estudos sobre evasão e retenção no ensino superior nas últimas décadas. As IFES oferecem a cada ano um crescente número de vagas nos cursos de graduação, no entanto, o percentual de formados não acompanha esse quantitativo ou mesmo reduz a cada ano INEP (2017).

A retenção ou permanência prolongada é a situação em que o discente leva um tempo bem maior do que o previsto para completar a matriz curricular do curso, muitas vezes o aluno não consegue concluir e permanece vários anos matriculado no curso. A evasão e a retenção comprometem a taxa de sucesso na formação universitária, o planejamento para ocupação das vagas, gera custos e compromete a eficiência e produtividade do sistema universitário. A evasão e a retenção causam prejuízos sociais pois reduz a formação de profissionais de nível superior capacitados para o mercado de trabalho.

No Brasil, o INEP e as IFES analisam a questão evasão e retenção principalmente sob o prisma das análises estatísticas baseadas em conjuntos de dados históricos. No entanto, as análises preditivas e sob demanda que podem agilizar a tomada de decisão dentro do ambiente acadêmico ainda são um desafio em aberto (ROMERO e VENTURA, 2013). Atualmente, a identificação do estudante que está em situação de risco de evasão ou retenção é uma tarefa que depende fundamentalmente da dedicação ou da experiência do gestor. Em muitas IFES, o acompanhamento do desempenho dos alunos é feito de maneira subjetiva, reativa, empírica, morosa e não sistemática, depende primordialmente da vivência acadêmica de pequenos grupos de docentes. Os gestores acadêmicos contam com poucas ferramentas que possam auxiliar na identificação de quais alunos estão em risco de evasão ou retenção.

Esta pesquisa tem como objetivo propor uma solução computacional para apoiar nas questões relacionadas à evasão e retenção na IFES. Apresentamos a arquitetura WAVE baseada em Mineração de Dados Educacionais (EDM) que utiliza apenas dados acadêmicos. Sua função é fornecer aos gestores, não especialista em mineração de dados, uma abordagem sistemática de acompanhamento semestral sobre o desempenho acadêmico dos estudantes e oferecer predição dos que estão em risco de abandonar o sistema. A solução proposta é modular e foi concebida para ser genérica e acoplável aos sistemas de gestão acadêmica (SGA) que já existentes nas IFES sem a necessidade de alterá-los ou adaptá-los. Através dos experimentos apoiados pela arquitetura EDM WAVE, avaliamos os dados de milhares de alunos matriculados nas dezenas de cursos da UFRJ.

2. TRABALHOS RELACIONADOS

No âmbito da IFES, a evasão e a retenção estão intimamente associadas ao desperdício financeiro de recursos públicos, já que são alocados professores, funcionários, equipamentos e espaço-físico cuja capacidade total não é integralmente desfrutada. Por esse motivo, nos últimos anos, há um interesse crescente no uso de mineração de dados no âmbito da Educação.

Diversos autores enfatizam a necessidade de utilizar mineração de dados na análise aprofundada de dados gerados na educação, surge a mineração de dados educacionais (EDM) (BAKER e YACEF, 2009, BAKER et al., 2011, ROMERO e VENTURA, 2013, PAIVA, 2014). Para Romero e Ventura (2013) existe a necessidade de se projetar e criar ferramentas que facilitam o conhecimento do ambiente educacional. Particularmente, um dos temas de maior destaque é a necessidade de aplicar EDM na predição do desempenho dos estudantes. Os autores definem a predição como o ato de estimar o valor desconhecido de uma variável que descreve a situação acadêmica do estudante.

Baker et al. (2011) destacaram que a maior concentração de pesquisas sobre a EDM estava sendo realizada em instituições estrangeiras e faltam trabalhos relacionados no Brasil. Manhães (2011) utilizou EDM para analisar a evasão universitária, sendo um dos primeiros no Brasil nessa área (RODRIGUES et al., 2014). Basicamente, os atuais trabalhos correlatos em EDM buscam predizer o desempenho de estudantes seguem duas linhas principais: (i) identificar quais são os atributos mais relevantes para caracterizar os grupos de estudantes, ou (ii) identificar e comparar o desempenho de modelos preditivos, trata-se da seleção dos algoritmos computacionais mais apropriados, estimando o desempenho quando aplicados a dados educacionais.

Os artigos mais relevantes na análise dos atributos e identificação das características dos estudantes são Minaei-Bidgoli et al. (2006), a dissertação de Souza (2008), a tese de Kampff (2009). Atualmente existem poucos trabalhos abrangentes voltados para as IFES brasileiras. Os principais trabalhos da área direcionados a identificar e comparar o desempenho dos algoritmos foram apresentados por Hamalainen e Vinni (2006), Superby et al. (2006), Garcia et al. (2009), Dekker et al. (2009), Zafra et al. (2011), Cheewaprakobkit (2013).

Esta pesquisa, diferentemente das demais, apresenta uma investigação mais ampla, aprofundada, refinada e voltada para o cenário brasileiro. Ela é fundamentada na investigação de dados acadêmicos de um período de 16 anos dos cursos de graduação da UFRJ. Com base nos dados acadêmicos, identificamos diversos padrões de desempenho e os principais atributos que caracterizam esses padrões. Este artigo relata os principais experimentos realizados para se chegar aos melhores modelos preditivos. Os modelos preditivos são o resultado da aplicação dos algoritmos de aprendizagem de máquina sobre os dados coletados. Os melhores modelos permitem identificar com maior precisão os estudantes em risco de evasão ou retenção nos cursos de graduação. Este estudo levou em consideração somente dados acadêmicos comuns ao maior números de IFES, de modo que os experimentos aqui relatados possam ser reproduzidos em qualquer IFES no Brasil. Além dos experimentos, este artigo apresenta a arquitetura de software capaz de adicionar funcionalidades preditivas aos SGA legados das IFES (MANHÃES, 2015).

3. WAVE: UMA ARQUITETURA BASEADA EM MINERAÇÃO DE DADOS

A arquitetura EDM WAVE foi concebida para inferir periódica e sistematicamente o desempenho acadêmico dos estudantes de graduação (MANHÃES et al., 2011, 2012,

2014a, 2014b, 2014c, 2014d, 2015a, 2015b). Ela foi projetada com bases nos requisitos da mineração de dados educacionais (BAKER e YACEF 2009, BAKER et al. 2011, ROMERO e VENTURA 2013), sendo voltada para os gestores acadêmicos.

A arquitetura foi concebida para ser modular e acessar a base de dados dos SGA das IFES e agregar novos possibilidades analíticas e funcionalidades aos SGA, permitindo que gestores acadêmicos, não especialistas em EDM, possam interagir com o sistema e buscar informações que os auxiliem na tomada de decisão.

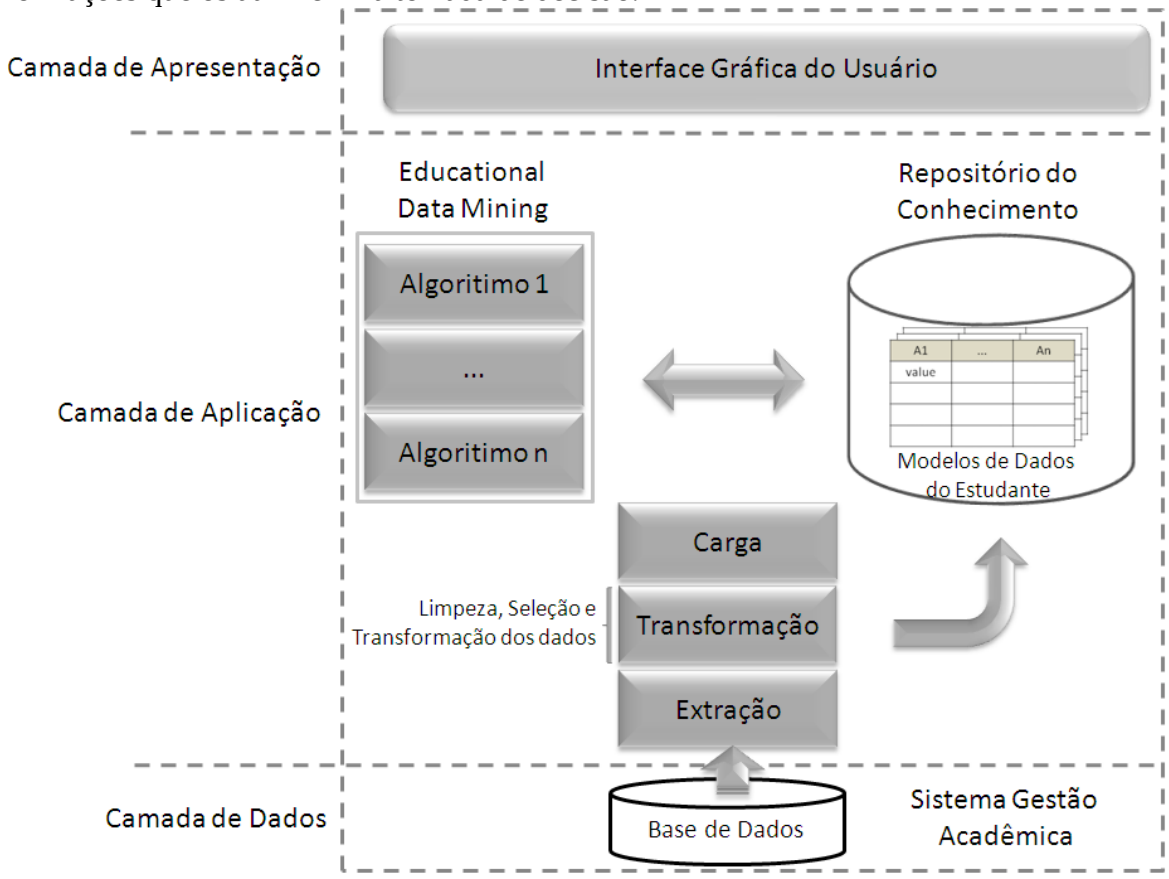

Figura 1. Representação conceitual da arquitetura EDM WAVE .

EDM WAVE foi implantada como uma arquitetura multicamadas. A Figura 1 ilustra e resume componentes da arquitetura e suas três camadas:

(1) A camada de dados consiste em bases de dados extraídos do SGA da IFES;

(2) A camada de aplicação gerencia as principais funcionalidades da arquitetura e as regras de processamento de dados. A camada de aplicação é constituída por três componentes: mecanismo de Extração-Transformação-Carga de dados (ETL), modelos preditivos (algoritmos) de EDM e repositório de conhecimento;

(3) A camada de apresentação é o nível mais alto. Ela é responsável por interagir com o gestor educacional que pode acessar o sistema diretamente usando suas interfaces.

4. METODOLOGIA E AVALIAÇÕES

A metodologia utilizada para conceber e avaliar a arquitetura EDM WAVE passou pela elaboração de vários estudos de casos (MANHÃES et al. 2011, 2012, 2014a, 2014b, 2014c, 2014d, 2015a, 2015b). Os experimentos realizados nos estudos de caso, ajudaram a compor o modelo de dado dos estudantes e o modelo de predição.

A Figura 2 mostra um esquema simples de predição, no qual as informações dos

estudantes constituem os dados de entrada, a partir dos dados de entrada o modelo pode produzir como saída uma predição.

Figura 2. Esquema simples de predição.

A precisão da predição, taxa de acertos, entre outras coisas, depende da qualidade e da quantidade de dados entrada. Esse conjunto e dados de entrada que permite aumentar a predição é chamado de modelo de dados. A geração do modelo de predição depende do modelo de dados e dos algoritmos de aprendizagem de máquina utilizados para aprendizagem do modelo. As avaliações utilizaram dados reais, porém sem identificação, de estudantes de vários cursos de graduação da UFRJ, os dados foram coletados durante de 16 anos e originalmente armazenados no SGA da UFRJ (SIGA).

4.1. MODELO DE DADOS

Os dados constituem parte indispensável no processo predição, no entanto a definição de quais dados utilizar para se obter a melhor predição constitui um verdadeiro desafio para a mineração de dados. Particularmente, quando os dados são educacionais e se pretende traçar um padrão de comportamento de estudante ao longo de um período durante o curso, esse desafio fica ainda maior. Alguns trabalhos relacionados utilizaram dados socioeconômicos para fazer a predição. No entanto, este tipo de dados é de coleta complexa (completude, corretude, opcionalidade) e são muito dependentes do contexto socioeconômico dos alunos. Essas informações não podem ser diretamente tratadas pelos algoritmos computacionais classificadores que constituem os modelos preditivos. Nosso trabalho, diferentemente dos relacionados, utilizou apenas dados invariantes e comuns a maioria das IFES.

A cada semestre letivo os SGA são alimentados com informações que variam com o tempo: notas das disciplinas, período da disciplina, CR (Coeficiente de Rendimento), CRA (Coeficiente de Rendimento Acumulado) entre outras informações dos discentes.

A partir dos dados originais do SGA, identificando os mais relevantes, realizamos diversas fases do tratamento de dados através de técnicas de ETL. A primeira fase cria uma série de novos atributos (atributos derivados). Desta forma, foi possível converter os atributos acadêmicos dos alunos originalmente armazenados nas bases de dados do SGA no formato para serem processados pelos algoritmos computacionais. Foi necessário redimensionar muitos atributos, entre eles o que identifica a situação do estudante no semestre letivo. Por exemplo, o SIGA apresenta vários valores para identificar a situação do aluno no semestre letivo (*AbandDef, Abandono, Aluno em Int, Ativa, Canc a Pedido, Canc Conc Int, Canc Dec Judic, Canc Faltou Matricula, Canc Opcao Curso, Canc out mot, Conclusao, Especial, Expulsao, Jubilamento, Mobilidade, Morte, Rem Aut, Rem Automatica, Rem Ex-offic, Rem Isenc Vest, Rem p/ Tranf, Rematricula, Trancada para o doutorado sanduiche, Trancamento, Trancamento Automatico, Trancamento Solicitado, Transferencia, Ult Prazo Tranc*), estes valores foram convertidos para apenas três valores: APROVADO, PAROU e ATIVO.

Ressaltamos que os dados originais dos SIGA não vieram com a identificação dos estudantes (matrícula) ou qualquer outro dado que possa identificar a pessoa. Para fins de

uso computacional foi criado um identificador genérico dos dados.

O resultado deste tratamento de dados permitiu projetar um modelo de dados conciso sobre os estudantes da IFES (Tabela 1), ele contém informações individuais relevantes sobre o desempenho acadêmico do estudante a cada semestre.

A predição do desempenho acadêmico dos estudantes depende fundamentalmente da escolha apropriada dos dados de entrada, dos atributos oriundos da base de dados dos SGA.

Tabela 1: Modelo de dados dos estudantes

Nº	Atributos	Descrição	Valor	Tipo
1	Id	Identificador genérico do dado	Código Id	String
2	IdIngresso	Identificador do ano e período em que o estudante ingressou na universidade	Código Id	String
3	IdCurso	Identificador do curso de graduação no qual o estudante está matriculado	Código Id	String
4	IdUnidadeCurso	Identificador da Escola, Instituto ou Faculdade onde o curso é oferecido na universidade	Código Id	String
5	CRA	Coeficiente Rendimento Acadêmico Acumulado (CRA) é a média de aproveitamento das disciplinas cursadas durante todo o curso de graduação	0 ... n	Numérico
6	(01S, 02S, ..., 13S) _Periodo	Período letivo identificado por (ano-semestre)	Código Id	String
7	(01S, 02S, ..., 13S) _SitPeriodo	Situação da matrícula do estudante no período	APROVADO, PAROU, ATIVO	String
8	(01S, 02S, ..., 13S) _CRPeriodo	Coeficiente de rendimento do período cursado	0 ... n	Numérico
9	(01S, 02S,..., 13S) _NoDisc	Número de disciplinas cursadas em cada período letivo	0 ... n	Numérico
10	(01S, 02S, ..., 13S) _NoAP	Número de disciplinas aprovadas em cada período letivo	0 ... n	Numérico
11	(01S, 02S, ..., 13S) _MediaAP	Média aritmética obtida nas disciplinas aprovadas em cada período letivo	0 ... 100	Numérico
12	(01S, 02S, ..., 13S) _NoRFM	Número de disciplinas reprovadas por falta e/ou média em cada período letivo	0 ... n	Numérico
13	(01S, 02S, ..., 13S) _NoRM	Número de disciplinas reprovadas por média em cada período letivo	0 ... n	Numérico
14	(1D, 2D, ...,7D) _Disciplina	Identifica as disciplinas do primeiro semestre da grade curricular do curso de graduação	Código disciplina	String
15	(1D, 2D, ...,7D) _Conceito	Armazena as notas (valor numérico) obtidas nas disciplinas da grade curricular do primeiro semestre letivo	0 ... 100	Numérico

16	(1D, 2D, ...,7D) _SitDisciplina	Situação da disciplina do primeiro semestre letivo: AP (Aprovado), RFM (Reprovado por Falta e Média), RM (Reprovado por Média) e RF (Reprovado por Falta)	AP, RM, RFM	String
17	atributo de classe	Atributo de classe é utilizado pelo algoritmo classificador para predizer a classe do exemplo.		

Através da arquitetura EDM WAVE, o gestor acadêmico pode definir (configurar) uma regra de progresso do curso de graduação segundo seus critérios. O estudante tem um progresso positivo quando seu desempenho em um determinado semestre do curso está acima de um padrão mínimo. Por exemplo, o estudante tem progresso quando foi aprovado em pelo menos quatro disciplinas. Esta regra de progresso pode ser o atributo de classe utilizado pelo algoritmo classificador para predizer o desempenho dos estudantes.

4.2. MODELO DE PREDIÇÃO

A identificação dos modelos mais adequados para predizer o desempenho acadêmico possui fundamental importância para solução do problema (BAKER et al. 2011). Neste estudo, foram selecionados os algoritmos classificadores de aprendizagem de máquina mais populares e amplamente utilizados em outros domínios (WITTEN e FRANK, 2005, KOTSIANTIS et al., 2007) para criar os modelos de predição utilizando dados educacionais.

A identificação dos algoritmos mais adequados para predizer o desempenho acadêmico possui fundamental importância para EDM (BAKER et al. 2011). Os algoritmos que apresentaram melhores resultados quando aplicados a dados educacionais são: (1) Statistical learning: BayesNet (BN) - implementação da rede bayesiana, e NaiveBayes (NB); (2) Decision trees: J48 - implementação do C4.5; (3) Learning rules: JRip (JR) - implementação do RIPPER; (4) Support Vector Machines: SVM with Poly Kernel (SVM1), e SVM with RBF Kernel (SVM2); (5) Perceptron-based: Multilayered Perceptrons (MP) - Rede neural Artificial; (6) Boosting: AdaBoost (AB); (7) Logistic regression model: SimpleLogistic (SL); (8) Decision table model: DecisionTable (DT); (9) One-level decision tree: OneR (OR); e (10) Randomized decision tree: RandomForest (RF).

O processo de construção de um modelo passa por duas fases distintas e interativas: a primeira, chamada de descritiva, requer um conjunto de dados de treinamento que serão utilizados pelo algoritmo classificador para construir o modelo descritivo dos dados. A segunda fase, chamada de preditiva, testa o modelo gerado na primeira fase utilizando novos dados, conjuntos de teste. Uma pessoa analisa os resultados, e verifica se o modelo atende ao propósito. O modelo é validado para ser usado com novos dados e obter a predição.

A Figura 3 mostra o esquema para construção do modelo. Vários algoritmos podem ser utilizados para criar os modelos que podem ser utilizados para solucionar o problema, no entanto, alguns modelos apresentam melhor desempenho.

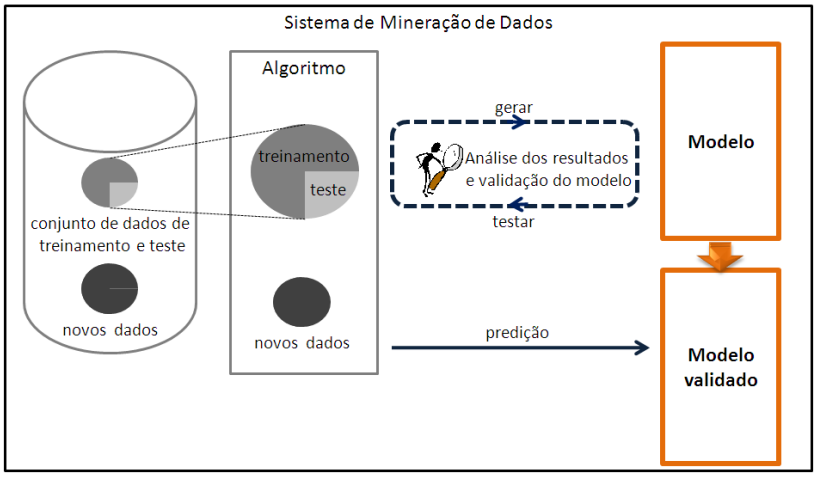

Figura 3: Esquema de construção, análise e validação do modelo preditivo de dados.

4.3. Estudos de Casos

Esta seção apresenta um resumo dos experimentos baseados em estudos de casos elaborados e desenvolvidos durante a tese de doutorado de Manhães (2015b). Os diversos experimentos permitiram avaliar os conjuntos de atributos para compor o modelo de dados (Tabela 1) e comparar o desempenho de 12 algoritmos classificadores afim de obter o melhor conjunto preditivo para a identificação dos alunos em risco de evasão e retenção nos cursos de graduação da UFRJ.

Os principais critérios de comparação utilizados foram: (i) as acurácias avaliadas individualmente; (ii) particionamento das bases de dados utilizando a validação cruzada com 10 conjuntos (10-fold cross validation) e dois conjuntos: treinamento e teste; (iii) as medidas estatísticas calculadas a partir da matriz de confusão: taxa de acerto da classe positiva (verdadeiro positivo), taxa de acerto da classe negativa (verdadeiro negativo); e (iv) os valores Kappa (Kappa de Cohen). A seguir serão relatados resumos dos estudos de casos:

Estudo de Caso 01 - Realizamos quatro experimentos com seguintes objetivos: (1) comparar o desempenho de 12 algoritmos considerando a acurácia dos classificadores; (2) verificar se a forma de particionar a base de dados interfere no desempenho dos algoritmos; (3) considerar a base de dados dividida em duas classes: não-concluintes e concluintes; (4) identificar os atributos mais relevantes para tratar o problema da evasão nos cursos de Engenharia Civil (MANHÃES et al., 2011, 2015b).

As formas de particionar a base de dados não interferem na acurácia dos algoritmos. Na análise dos atributos mais importantes para identificar o desempenho dos estudantes foram identificados: a nota na disciplina de Cálculo Diferencial e Integral I e o coeficiente de rendimento do primeiro semestre letivo.

A Figura 4 mostra os resultados das acurácias de 12 algoritmos classificadores para quatro experimentos distintos utilizando os dados do estudo de caso 1, estudantes do Curso de Engenharia Civil. O gráfico mostra que as acurácias de todos algoritmos observados foram superiores a 75%. O algoritmo OR possui uma acurácia superior a 84.0 para os três primeiros experimentos, no entanto, no experimento 4 a acurácia obtida é menor aproximadamente 80%. Os algoritmos DT, AB, OR e MP possuem acurácia semelhante para os três primeiros experimentos, mas a acurácia do experimento 4 é inferior com relação aos três primeiros experimentos. Os algoritmos MP, SVM1 e SVM2 possuem acurácias inferior a 80.0 para todos os experimentos. Os algoritmos SL, J48, BN e NB são os que apresentam maior homogeneidade entre as acurácias nos quatros experimentos e acima de 80%. O algoritmo Naive Bayes (NB) obteve acurácia superior a 80% em todos os quatro experimentos.

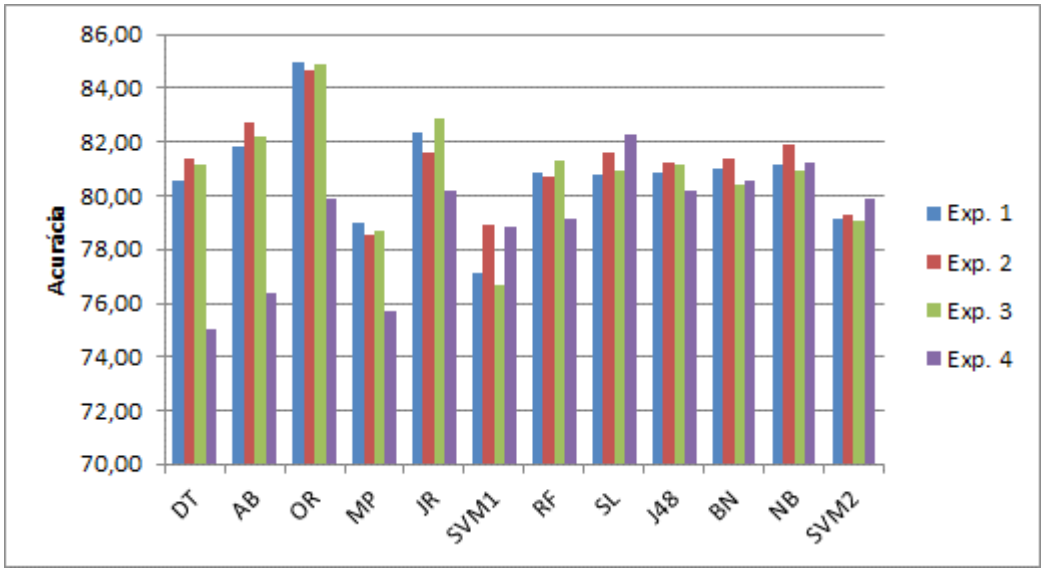

Figura 4: Gráfico das acurácias dos classificadores nos experimentos do estudo de caso 1.

Estudo de Caso 02 - Neste estudo, avaliamos quantitativamente os fatores (atributos dos dados) que influenciam o desempenho acadêmico dos estudantes de graduação da UFRJ. A correta predição depende da identificação desses atributos que caracterizam o desempenho dos estudantes ao longo da graduação. Após análise das tuplas da base de dados do SGA, identificamos três perfis de estudantes como: (i) cancelados - estudantes que interromperam o curso em algum período antes da formatura (evasão); (ii) ativos - estudantes que permaneceram matriculados além do prazo médio para conclusão do curso; e (iii) concluintes - estudantes que concluíram o curso de graduação.

Os objetivos do estudo de caso foram: (1) identificar os três perfis de estudantes preponderantes; (2) verificar quais algoritmos de mineração de dados são mais eficientes na predição dos três perfis; (3) identificar os principais valores dos atributos que diferenciam os três perfis; (4) identificar qual algoritmo possui melhor resultado e que possa ser convertido em visualizações gráficas para ser mais facilmente interpretado pelo gestor; (5) apresentar uma análise quantitativa dos principais atributos que diferenciam os três perfis.

Neste estudo de caso, foram comparamos o desempenho de 12 algoritmos em uma base de dados de 14 mil estudantes de 250 cursos de graduação oferecidos por 28 unidades acadêmicas da UFRJ. Os algoritmos demonstraram taxa de acerto geral (acurácia) superior a 78%. No entanto, os algoritmos mais sofisticados perdem na questão do tempo de processamento para construção dos modelos. Dentre os algoritmos avaliados, o Naive Bayes mostrou os melhores resultados e ofereceu maior facilidade na conversão dos modelos resultante para a análise quantitativa. O modelo gerado pelo classificador Naive Bayes apresentou informações importantes sobre o comportamento ao longo do curso dos três perfis distintos de estudantes (MANHÃES et al., 2012, 2014a, 2015b).

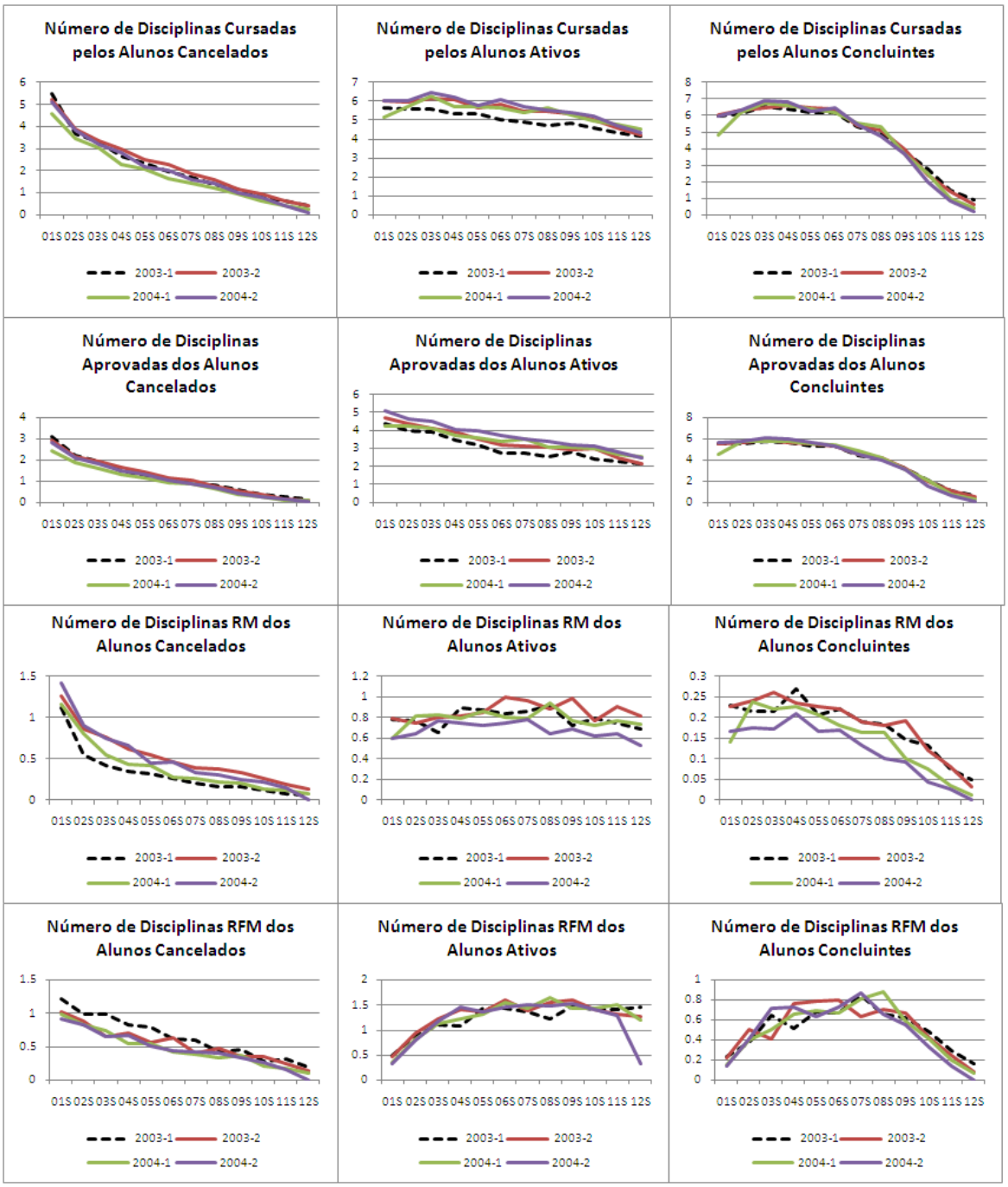

Figura 5: Da esquerda para a direita, temos os gráficos que mostram o número de disciplinas cursadas pelos estudantes: (a) cancelados, (b) ativos e (c) concluintes. Os gráficos que apresentam o número de disciplinas aprovadas para: (d) cancelados, (e) ativos e (f) concluintes. Os gráficos com o número de disciplinas RM para estudantes: (g) cancelados, (h) ativos e (i) concluintes. Os gráficos com o número de disciplinas RFM para estudantes: (j) cancelados, (k) ativos e (l) concluintes.

Estudos de Casos 03, 04, 05 - Nestes estudos, os objetivos foram avaliar a arquitetura EDM WAVE considerando os seguintes aspectos: o número de algoritmos para a camada de aplicação da arquitetura; o desempenho dos algoritmos para identificar os três perfis distintos dos estudantes; apresentar a predição do desempenho acadêmico para os

cinco primeiros semestres letivos para os cursos de graduação de base matemática (Engenharia Civil, Engenharia de Produção e Engenharia Mecânica e suas ênfases).

Os resultados obtidos mostraram que a arquitetura pode auxiliar na identificação do desempenho individual dos estudantes. Os algoritmos classificadores foram avaliados segundo a predição para três perfis distintos: (i) (APROVADO) indica que o estudante obteve aprovação em pelo menos uma disciplina no semestre letivo; (ii) (PAROU) indica que o estudante parou, ou seja, não se matriculou em nenhuma disciplina no semestre letivo; e (iii) (REPROVADO) indica estudantes que não registraram progresso no semestre, nenhuma aprovação nas disciplinas e reprovação (RFM) e/ou (RM) nas disciplinas cursadas no semestre letivo. As acurácias dos algoritmos foram superiores a 83%, considerando a avaliação dos algoritmos para os três perfis. A matriz de confusão para cada algoritmo mostrou alto índice de acerto para a classe (APROVADO) e (PAROU), mas baixo nível de acerto para a classe (REPROVADO). Observamos que a acurácia dos classificadores melhorou, ficando acima dos 89%, à medida que dados de vários semestres eram considerados como dado de entrada, ou seja, os algoritmos conseguem melhorar a taxa de predição ao longo do curso de graduação. O algoritmo *Naive Bayes* conseguiu prever a classe (REPROVADO) com os melhores índices (MANHÃES, 2015b).

Estudo de Caso 06 - Neste caso, o objetivo foi refazer os estudos de casos (03, 04 e 05) considerando apenas dois perfis para identificar a situação do estudante no semestre letivo. O valor do primeiro perfil é atribuído ao estudante que não obteve progresso do semestre (não-progresso-NP). O segundo perfil refere-se ao estudante que obteve progresso no semestre (progresso-P). O principal objetivo é identificar os algoritmos classificadores com uma taxa de acerto maior para a classe não-progresso, pois o estudante sem aprovação nas disciplinas do próximo semestre ou abaixo de uma regra de progresso, indica uma alta probabilidade de abandono do sistema de ensino.

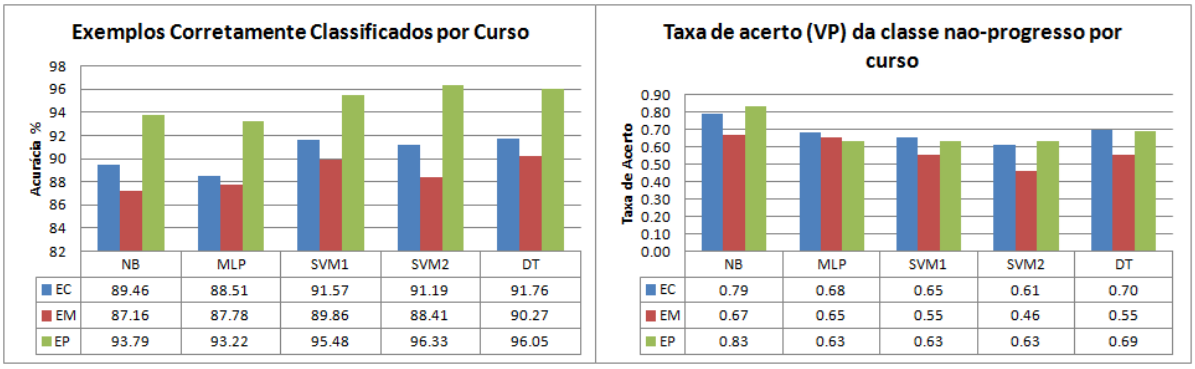

Figura 6: Da esquerda para direita temos os gráficos: (a) Exemplos corretamente classificados (acurácia) por curso de graduação. (b) Taxa de acerto da classe não-progresso por curso.

O gráfico da Figura 4.6a mostra a porcentagem de acerto dos classificadores (acurácia), observamos que a taxa de acerto foi elevada acima de 87% para todos os três cursos de graduação avaliados: engenharia civil (EC), engenharia mecânica (EM) e engenharia de produção (EP). O curso de Engenharia de Produção mostrou melhor resultado. O algoritmo classificador Naive Bayes obteve a melhor taxa para identificar corretamente alunos que não tiveram progresso segundo alguma regra (MANHÃES et al., 2014b, 2015b).

Estudo de Caso 07 - Avaliação das funcionalidades da Arquitetura EDM WAVE para os cursos de graduação historicamente associados com elevadas taxas de evasão e retenção, ou seja, cursos baseados em Matemática, Ciências e Tecnologias, a saber:

Engenharia Civil (EC), Engenharia Mecânica (EM), Engenharia de Produção (EP), Farmácia (FAR), Física (FIS) e Direito (DIR). Os cursos foram escolhidos porque pertencem a departamentos distintos da universidade e possuírem práticas pedagógicas diferenciadas. Além disso, os ingressantes desses cursos são considerados diferenciados, pois tais cursos possuem várias entradas de estudantes por ano/semestre.

Nos estudos de casos anteriores, verificou-se que o modelo de dados proposto e os algoritmos atendiam as demandas da predição do desempenho dos estudantes para muitos cursos de graduação da UFRJ. Verificamos experimentalmente que o algoritmo Naive Bayes mostrou um conjunto de características adequadas para utilizá-lo na arquitetura EDM WAVE.

Neste estudo, realizamos experimentos e análises mais aprofundadas sobre a utilização específica do algoritmo Naive Bayes para os cursos supracitados. Verificamos que o algoritmo apresentou a maior taxa de acerto para predizer quais estudantes não terão progresso no próximo semestre letivo, classe positiva (não-progresso). Além disso, o algoritmo apresentou um modelo de predição mais interpretável para os gestores. Desta forma, o resultado de sua predição pode ser facilmente convertido em gráficos possibilitando análise quantitativa ou outras formas de representar o conhecimento para os usuários da arquitetura EDM WAVE.

Os resultados obtidos nestes experimentos mostraram que a porcentagem dos exemplos corretamente rotulados considerando as duas classes do algoritmo para cada conjunto de teste dos seis cursos de graduação demostrou que a acurácia está em torno de 80%. As taxas de acerto da classe positiva (não-progresso) estão acima de 60%. Identificamos que a maior parte dos conjuntos de estudantes analisados neste estudo apresentou taxa de acerto do classificador para a classe negativa (progresso) superior a 80% (MANHÃES et al., 2014c, 2015b).

5. CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho propôs e avaliou quantitativamente uma solução de software modular baseada em técnicas de mineração de dados cujo objetivo é auxiliar gestores acadêmicos a prever o desempenho acadêmico dos estudantes de graduação e identificar aqueles que estão em risco de evadir do sistema de ensino.

As maiores dificuldades encontradas neste trabalho foram obter permissão de acesso aos dados e alteração do sistema SGA da UFRJ. Uma dificuldade adicional foi compreender o modelo de dados do SIGA já não havia documentação técnicas disponível. A finalidade da arquitetura EDM WAVE é operar acoplada aos sistemas SGA já existentes nas IFES, acredita-se que este tipo de abordagem requeira menores esforços e custos que o desenvolvimento de novos SGA.

Diferentemente dos trabalhos relacionados, os estudos apresentados foram baseados em dados acadêmicos de milhares de alunos e dezenas de cursos presenciais distintos, permitindo avaliar muitos algoritmos de aprendizado de máquina no contexto dos três principais perfis de alunos das IFES. Constatou-se que o modelo preditivo baseado no algoritmo *Naive Bayes* apresentou bons resultados no computo geral. Ele apresentou um modelo de classificação mais interpretável, além de ser facilmente convertido em gráficos facilitando o trabalho de gestores não especialistas em computação. Os resultados obtidos nos experimentos mostram que a arquitetura EDM WAVE é capaz de fornecer apoio para os gestores educacionais e tomadores de decisão gerenciais utilizando somente dados não variantes no tempo, além de apresentar um método sistemático e eficaz de monitoramento do progresso dos estudantes ao longo de cada semestre letivo.

Finalmente, no que tange aos possíveis benefícios diretos da utilização da arquitetura EDM WAVE por parte dos gestores são: (i) propor ações proativas ao identificar precocemente e ao longo do tempo, principalmente nos períodos iniciais, quais estudantes estão em risco de evasão e aqueles com possibilidade de permanecerem matriculados além do prazo médio para conclusão do curso; (ii) identificar os fatores de sucesso e insucesso específicos para cada curso; e (iii) promover discussões/ações focadas à detecção de gargalos e promoção de ajustes nos currículos dos cursos.

Como trabalhos futuros, consideramos ampliar os estudos experimentais e avaliar a arquitetura EDM WAVE para cursos de graduação oferecidos na modalidade on-line ou EAD. Além disso, poderão ser desenvolvidas novas interfaces gráficas não só para os gestores e como avaliar possíveis estratégias de emitir alertas e relatórios para os próprios estudantes.

6. REFERÊNCIAS BIBLIOGRÁFICAS

BAKER, R., ISOTANI, S., CARVALHO, A. **Mineração de Dados Educacionais: Oportunidades para o Brasil**. Revista Brasileira de Informática na Educação, 19(02), 3-13, 2011. Disponível em: http://dx.doi.org/10.5753/RBIE.2011.19.02.03.

BAKER, R.S.J.d., YACEF, K. **The State of Educational Data Mining in 2009: A Review and Future Visions**. Journal of Educational Data Mining (JEDM), v. 1, n. 1, October 2009, 3-17, 2009.

CHEEWAPRAKOBKIT, P. **Study of Factors Analysis Affecting Academic Achievement of Undergraduate Students in International Program**. In: Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS). Vol I, Hong Kong, March 13-15, 2013.

DEKKER G., PECHENIZKIY M., VLEESHOUWERS J. **Predicting Students Drop Out: A Case Study**. In Proceedings of the International Conference on Educational Data Mining, Cordoba, Spain, 2009.

GARCIA, E., ROMERO, C., VENTURA, S., GEA, M., DE CASTRO, C. **Collaborative Data Mining Tool for Education**. International Working Group on Educational Data Mining, 2009.

HAMALAINEN, W., VINNI, M. **Comparison of machine learning methods for intelligent tutoring systems**. in Proc. Int. Conf. Intell. Tutoring Syst., 525-534, 2006.

INEP. **Resumo Técnico do Censo da Educação Superior 2017**. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2017. Disponível em: http://portal.inep.gov.br/web/censo-da-educacao-superior/resumos-tecnicos.

KAMPFF, A. J. C. **Mineração de dados educacionais para geração de alertas em ambientes virtuais de aprendizagem como apoio à prática docente**. Tese de doutorado em Informática na Educação. Universidade Federal do Rio Grande do Sul, 2009.

KOTSIANTIS, S. B., ZAHARAKIS, I. D. AND PINTELAS, P. E. **Supervised machine learning: A review of classification techniques**, 3-24, 2007.

MANHÃES, L.M.B. et. al. **Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados**. Anais do Simpósio Brasileiro de Informática na Educação (XXII SBIE), V. 1. N. 1, 150-159, 2011.

MANHÃES, L.M.B. et. al. **Identificação dos Fatores que Influenciam a Evasão em Cursos de Graduação Através de Sistemas Baseados em Mineração de Dados: Uma**

Abordagem Quantitativa. Anais do VIII Simpósio Brasileiro de Sistemas de Informação (SBSI 2012) - Trilhas Técnicas, pp. 468-479, 2012.

MANHÃES, L.M.B., CRUZ, S.M.S., ZIMBRÃO, G. **The Impact of High Dropout Rates in a Large Public Federal Brazilian University: A Quantitative Approach Using Educational Data Mining**. In: 6th CSEDU 2014, Barcelona, Spain, 124-129, 2014a.

MANHÃES, L.M.B., CRUZ, S.M.S., ZIMBRÃO, G. **WAVE: an Architecture for Predicting Dropout in Undergraduate Courses using EDM**. In: Symposium of Applied Computing (SAC 2014), Gyeongju, Korea, 2014b.

MANHÃES, L.M.B., CRUZ, S.M.S., ZIMBRÃO, G. **Evaluating Performance and Dropouts of Undergraduates using Educational Data Mining**. The 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2014), Data Mining for Educational Assessment and Feedback workshop (ASSESS 2014), 2014c.

MANHÃES, L.M.B., CRUZ, S.M.S., ZIMBRÃO, G. **Investigating Withdraw of STEM Courses in a Brazilian University with EDM**. 2nd Symposium on knowledge Discovery, Mining and Learning (KDMiLe 2014), 2014d.

MANHÃES, L.M.B., CRUZ, S.M.S., ZIMBRÃO, G. **Towards Automatic Prediction of Student Performance in STEM Undergraduate Degree**. Programs. In: Symposium of Applied Computing (SAC'15), April 13–17, Salamanca, Spain, 2015a. http://dx.doi.org/10.1145/2695664.2695918.

MANHÃES, L.M.B., CRUZ, S.M.S., ZIMBRÃO, G. **Predição do Desempenho Acadêmico de Graduandos Utilizando Mineração de Dados Educacionais**. Tese de doutorado, 2015b.

MINAEI-BIDGOLI, B., TAN, P., KORTEMEYER G., PUNCH, W.F. Association analysis for a web-based educational system. Data Mining in E-Learning. WitPress. Southampton, Boston, 2006.

PAIVA, R., BITTENCOURT, I.I., SILVA, A.P., ISOTANI, S., JAQUES, P. **A Systematic Approach for Providing Personalized Pedagogical Recommendations Based on Educational Data Mining**. In: Int. Conf. on Intelligent Tutoring Syst., Honolulu. Lecture Notes in Computer Science, p. 362-367, 2014.

RODRIGUES, RODRIGO LINS, et al. **A literatura brasileira sobre mineração de dados educacionais**. Anais dos CBIE. Vol. 3. No. 1., 2014.

ROMERO, C., VENTURA, S. **Data Mining in Education**. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, In Press. V. 3, N. 1, 12-27, 2013.

SOUZA, S.L. **Evasão no ensino superior: um estudo utilizando a mineração de dados como ferramenta de gestão do conhecimento em um banco de dados referente à graduação de engenharia**. Dissertação de Mestrado, COPPE/UFRJ, Engenharia Civil, Rio de Janeiro, RJ, Brasil, 2008.

SUPERBY, J.F., VANDAMME, J-P., MESKENS, N. **Determination of factors influencing the achievement of the first-year university students using data mining methods**. In Proc. Int. Conf. Intelligent Tutoring System of the Workshop on Educational Data Mining, Taiwan, 2006, pp. 1-8, 2006.

WITTEN, I.H. AND FRANK, E. **Data Mining: Practical machine learning tools and techniques**. 2nd edition Morgan Kaufmann, San Francisco, 2005.

ZAFRA, A., ROMERO, C., VENTURA, S. **Multiple instance learning for classifying students in learning management systems**, 2011.