Mineração de Dados Educacionais Aplicada à Identificação de Variáveis Associadas à Evasão e Retenção

Diego da Costa do Couto, Ádamo Lima de Santana

Laboratório de Inteligência Computacional e Pesquisa Operacional (LINC) – Universidade Federal do Pará (UFPA)
Caixa Postal 479 – 66.075-110 – Belém – PA – Brasil

{diegocouto,adamo}@ufpa.br

Abstract. *This paper applies classification algorithms in a large database with the purpose of diagnosing the causes of two problems faced in Brazilian universities, college dropout and retention. The accuracies of many algorithms were measured with a focus on verifying the ability to correctly classify available instances. Results showed that the Bayesian Network method reached an overall precision approximately 86% and it is considered a very satisfactory solution for the discovery and representation of knowledge about academic performance of undergraduate students, especially those who are willing to give up or extrapolate the deadline for completing to the course.*

Resumo. *Este artigo aplica algoritmos de classificação em uma grande base de dados com finalidade de diagnosticar as causas de dois problemas enfrentados em universidades brasileiras, a evasão e a retenção. Foram mensuradas acurácias de diversos algoritmos com foco em verificar a capacidade de classificar corretamente as instâncias disponíveis. Os resultados apontaram que o método Rede Bayesiana atingiu precisão geral de aproximadamente 86% sendo considerada uma solução bastante satisfatória para descoberta e representação do conhecimento acerca do desempenho acadêmico dos alunos da graduação, especialmente aqueles propensos a desistir ou extrapolar o prazo para conclusão do curso.*

1. Introdução

A educação superior está em ascensão no Brasil. O Censo da Educação Superior revelou que até o ano de 2013 existiam cerca de 32.049 cursos de graduação em todo país, distribuídos entre os graus bacharelado, licenciatura e tecnológico nas modalidades de ensino presencial e a distância. Em 2014, houve um acréscimo de aproximadamente 2,5% (32.878) no número dos cursos ofertados em 2.368 Instituições de Ensino Superior (IES). Vale destacar que entre 2003 e 2014, a matrícula na educação superior registrou aumento de 96,5% [INEP 2014a]. Estas constatações corroboram os avanços em termos quantitativos da educação superior no país nas iniciativas privada e estatal. Contudo, ressalta-se que gestores devem continuamente avaliar se este aumento em quantidade se converteu igualmente em qualidade, ao estudante, à instituição de ensino superior e à sociedade.

Os levantamentos realizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), formatados no Censo da Educação Superior, também apontam descompasso entre os números de matrícula, ingressantes, cursos e

concluintes. Constata-se que houve queda no número de concluintes em todas as organizações acadêmicas, representando redução de 6,4% para as faculdades, 6,8% em relação aos centros universitários, 4,4% tendo em conta as universidades e 26,0% considerando Institutos Federais (IFs) e Cefets [INEP 2014b]. Essas informações denotam um importante diagnóstico: o aumento na quantidade de vagas não está impactando diretamente na permanência do aluno até a sua formatura. Esta problemática, conhecida como evasão resulta em vagas ociosas ou remanescentes, as quais se destinam a outros processos de seleção.

O Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais (REUNI), instituído pelo Decreto no 6.096, de 24 de abril de 2007 [Governo Federal 2007], torna explícita em suas diretrizes gerais [MEC 2007] a preocupação do Governo Federal acerca do problema da evasão, sob a asserção "os índices de evasão de estudantes nos cursos de graduação atingem, em alguns casos, níveis alarmantes". Outra medida adotada pelo Governo Federal brasileiro foi a Lei Nº 12.089 de novembro de 2009 [Governo Federal 2009], a qual proíbe a ocupação de duas vagas, simultaneamente, pela mesma pessoa em cursos de graduação de instituições públicas do ensino superior. Infere-se que a lei visa minimizar os casos nos quais discentes, por desinteresse ou vários outros motivos, abandonem um dos cursos ou demorem mais que o tempo normal para concluírem os estudos, sendo este último fenômeno chamado de retenção.

Segundo [Silva Filho et al. 2007] a evasão estudantil no ensino superior, de modo geral, causa desperdícios de ordem social, acadêmica e econômica. Os reflexos deste entrave no setor público de ensino se manifestam quando os recursos são aplicados sem o devido retorno à sociedade. Enquanto que no ramo privado, os empresários perdem receitas e aumentam os gastos com manutenção de infraestrutura de ensino. Em ambos os casos, a evasão implica em ociosidade de professores, funcionários, equipamentos e espaços físicos. A desistência tem consequências diretas no cotidiano do estudante, visto que este não consegue a qualificação necessária para atuar na área pretendida e, em outros casos, não retorna à IES em busca de novas oportunidades. Segundo os pontos de vista dos autores [Tinto 1975], [Tinto 1987] e [Andriola 2009], as causas da evasão emanam da falta de integração com ambiente acadêmico e social da instituição.

Considerando que os problemas da evasão e retenção possuem inúmeras causas e consequências negativas para estudantes, instituições de ensino e comunidades nas quais esses indivíduos estão inseridos, este trabalho tem como objetivo a criação de subsídios que auxiliem gestores da instituição de ensino superior a identificar alunos, dos cursos de graduação, em situação de vulnerabilidade à evasão ou à retenção dentro dos seus ambientes de aprendizagem. Dentre os subsídios importantes à gestão, destacam-se: previsão de quais alunos são propensos a desistir ou permanecer além do tempo estipulado pelo currículo; representação desta informação; e identificar quais atributos, dentre os disponíveis, são mais relevantes durante a classificação desse aluno.

Pretende-se alcançar estes objetivos pela utilização da Descoberta de Conhecimento em Base de Dados (*Knowledge Discovery in Database* – KDD) que representa um "processo não-trivial de extração de informações implícitas, previamente desconhecidas e potencialmente úteis a partir de dados" [Frawley et al. 1992]. Uma das suas etapas, denominada de Mineração de Dados (*Data Mining*) [Fayyad et al. 1996], onde ocorre a extração de padrões dos dados através do uso de algoritmos específicos,

foi empregada para verificar a relação entre as variáveis e a problemática explicitada. A etapa de *Data Mining* pode ser aplicada em diversas áreas [Han et al. 2012] [Goldschmidt e Passos 2005] [Luo 2008] [Fayyad et al. 1996] desde que estas possuam razoáveis volumes de dados históricos.

Foram testados algoritmos, durante a etapa de *Data Mining*, a partir da tarefa de classificação, que define-se como "o processo de atribuir, a uma determinada informação recebida, o nome de uma classe à qual ela pertence" [Rich e Knight 1993] ou ainda constrói um modelo ou classificador [Han et al. 2012]. Dentro do contexto aplicado, a classificação permite presumir a situação (classe) do estudante na universidade, dado um conjunto de atributos a respeito desse aluno. Avaliaram-se métricas relativas ao desempenho dos classificadores, cujas características possam atender aos requisitos associados ao objetivo deste trabalho, com o intuito de testá-los e, posteriormente, selecioná-los à resolução do problema pesquisado.

Este trabalho está organizado da seguinte forma: A Seção 2 apresenta os trabalhos correlatos. Por sua vez, na Seção 3 é apresentada a base de dados utilizada neste trabalho. Na Seção 4 será apresentada proposta de aplicação. Em seguida, Seção 5 serão discutidos os resultados. Na Seção 6 apresentação das considerações finais.

2. Trabalhos Correlatos

O sítio [JEDM 2016] define Mineração de Dados Educacionais (*Educational Data Mining* – EDM) como uma disciplina emergente cujo objetivo está no desenvolvimento de métodos para explorar os dados provenientes de cenários educacionais e como essas metodologias são empregadas para compreender os alunos nos seus ambientes de aprendizagem. Argumenta-se a existência de um aumento considerável no interesse por pesquisas valendo-se de EDM [Sachin e Vijay 2012], nesta perspectiva, [Romero e Ventura 2010] elaboraram um trabalho relativo ao estado da arte da Mineração de Dados Educacionais, no qual são discutidas 235 publicações mais relevantes até o ano de 2009. Os autores verificaram crescimento exponencial no número de publicações ao longo dos últimos anos, destacando o aparecimento de jornais e edições de livros especializados no assunto.

[Baker et al. 2011] ratificam os argumentos supramencionados, afirmando, em adição, que a área de EDM vem crescendo rapidamente em países da Europa e nos EUA, salientando ainda o fortalecimento das pesquisas no Brasil durante a última década. Destas constatações, deduz-se que a comunidade científica está cada vez mais interessada nesse campo emergente de investigação, o que o torna tendência promissora de investimentos e pesquisas em âmbito científico e acadêmico [Baker 2009].

[Cortez e Silva 2008] obtiveram dados no período letivo de 2005 e 2006 de escolas públicas de Portugal. Os atributos constituíram-se de registros coletados de relatórios emitidos pelo sistema escolar e questionários com perguntas sobre aspectos sociais, demográficos e emocionais dos estudantes. A finalidade dos autores era predizer o desempenho escolar nas disciplinas básicas de Matemática e Português. Os autores trabalharam com três metodologias: classificação binária, multi-classe e regressão. Os algoritmos testados foram Árvore de Decisão (*Decision Tree* – DT), *Random Forest*, Redes Neurais Artificiais (RNA) e *Support Vector Machine* (SVM). Os resultados atingidos foram satisfatórios, a exemplo, para o teste com classes binárias a

árvore de decisão conseguiu a melhor taxa de acerto (93, 0%). [Cortez e Silva 2008] priorizaram a geração de conhecimento especialista, os autores descobriram importantes regras das árvores de decisão.

No Brasil, as investigações em Mineração de Dados Educacionais se consolidaram em 2012, na ocasião, [Manhães et al. 2012] elaboraram um estudo de caso para avaliar a evasão em 155 cursos de graduação ofertados por 28 unidades da UFRJ. Para a pesquisa em discussão, foram selecionados dados acadêmicos dos discentes que ingressaram nos dois semestres letivos dos anos de 2003 e 2004. Além da acurácia, a interpretabilidade dos resultados foi um dos requisitos considerados à escolha do método apropriado para solucionar a problemática. Neste contexto, o classificador *Naive Bayes* foi escolhido, pois conseguiu atingir precisão global superior a 80%. As contribuições da pesquisa citada também foram publicadas com outros resultados e métodos de avaliações em [Manhães et al. 2014b] [Manhães et al. 2014a, Manhães et al. 2015].

O nosso trabalho, proposto neste artigo, possui similaridades com aqueles discutidos anteriormente, visto que, por exemplo, vale-se de algoritmos classificadores para detecção de um padrão que classifique às instâncias corretamente quanto à evasão e retenção em âmbito acadêmico. Contudo diferencia-se dos demais nos seguintes aspectos: i) aplicação em uma grande base de dados, composta por quase 100 mil amostras, pois a maioria dos trabalhos usam data sets com algumas centenas de registros; ii) análise sobre todos os cursos de graduação, enquanto muitos trabalhos avaliam cursos ou disciplinas de maneira isolada. Além disso, este trabalho visa fortalecer o campo de EDM, uma vez que esta área é nova, há poucos estudos nacionais e exerce grande influência na resolução de problemas atrelados ao desempenho escolar.

3. Base de Dados

O Sistema Integrado de Gestão de Atividades Acadêmicas (SIGAA) é parte dos Sistemas Institucionais Integrados de Gestão (SIG) e informatiza os procedimentos da área acadêmica através de módulos como: graduação, pós-graduação (*stricto e lato sensu*), ensino técnico, entre outros. O SIGAA foi adquirido pela Universidade Federal do Pará (UFPA), por meio de um contrato firmado com a Universidade Federal do Rio Grande do Norte (UFRN), além disso, outras Instituições Federais de Ensino Superior (IFES) e Institutos Federais (IF) também adquiriram o produto a fim de promover integração entre sistemas, processar dados e oferecer serviços da área fim através de interface *web*.

Os dados selecionados à pesquisa são registros acadêmicos, oriundos do SIGAA, referentes a 157.298 discentes dos cursos de graduação, ingressantes até o ano de 2016, da Universidade Federal do Pará, totalizando 175.779 amostras. Desta quantidade, as tuplas inconsistentes ou que continham valores nulos foram removidas, permanecendo 98.698 linhas. A Tabela 1 mostra os 31 atributos selecionados e os seus respectivos significados.

3.1. Pré-processamento e transformação de dados

Os significados dos atributos de 1 a 12, considerados intuitivos, podem ser consultados nas descrições dispostas na Tabela 1. As variáveis de 13 a 19 representam os indicadores de rendimento acadêmico acumulado, a saber: Média de Conclusão (MC),

Média de Conclusão Normalizada (MCN), Índice de Rendimento Acadêmico (IRA) ou Coeficiente de Rendimento Geral (CRG), Índice de Eficiência em Carga Horária (IECH), Índice de Eficiência em Períodos Letivos (IEPL), Índice de Eficiência Acadêmica (IEA) e Índice de Eficiência Acadêmica Normalizado (IEAN). Essas métricas quantificam o desempenho dos alunos da graduação e nos cálculos consideram-se dados do histórico acadêmico, tais como: quantidades de reprovações, aprovações, trancamentos, cargas horárias acumuladas e esperadas para integralização do curso, entre outros. As fórmulas para cálculo desses indicadores e os seus respectivos significados estão dispostos no Regimento da Graduação da Universidade Federal do Rio Grande do Norte [UFRN 2013].

Tabela 1. Atributos selecionados à pesquisa

Nº	Variável	Descrição
1	sexo	Sexo que o discente pertence
2	idade	Idade que o aluno ingressou no curso
3	interior	Informa se o discente estuda no campus capital ou em um dos campi do interior do estado
4	turno	Turno no qual o discente estuda
5	forma_ingresso	Forma de seleção pela qual o discente ingressou na universidade
6	numero_trancamento	Número de vezes que o discente trancou a matrícula
7	numero_vinculos	Número de vezes que o discente fez outras graduações (vínculos) até o ingresso no curso atual
8-10	perc_ch_{tipo}	Percentual das cargas horárias prática, teórica e de estágio
11	sem_ordem	O percentual das disciplinas cursadas fora da ordem proposta pelo currículo do discente
12	primeiro_semestre_ocorr	Informa qual o semestre que o discente cursou pela primeira vez uma disciplina fora de ordem
13-19	indices_academicos	Representam os indicadores de desempenho acadêmico
20-22	prob_indices	Refere-se a probabilidade de um discente formado nos últimos 5 anos possuir índice acadêmico maior ou igual ao aluno avaliado
23-30	perc_{conceito}_{avaliacao}	Refere-se ao percentual de um conceito conseguido pelo discente dentro do período avaliado
31	status	Denota a situação (classe) a qual o estudante pertence

Os atributos de 20 a 22 denotam a probabilidade de um discente formado nos últimos cinco anos possuir um dos índices acadêmicos igual ou superior aos demais alunos pertencentes ao mesmo curso e matriz curricular. Foram usados os indicadores MC, IRA e IEA, uma vez que estes, em suas definições matemáticas e conceituais, aferem a eficiência do aluno durante o seu percurso acadêmico.

A média das notas obtidas pelo estudante em cada disciplina, em um período letivo, é convertida em conceito, definido segundo a escala apresentada na Tabela 2. As variáveis indexadas de 23 a 30 referem-se ao percentual de um determinado conceito de acordo com o período de avaliação, seja este geral (acumulado por todo o curso) ou para

o primeiro ano cursado. Por exemplo, a variável *perc_ins_primeiro_ano* denota o percentual de conceitos do tipo INS referente ao primeiro ano de graduação.

Finalmente, o atributo 31 representa a classe à qual o discente pertence, cujos possíveis valores são: "Formado", "Evadido" e "Retido". Os alunos considerados na classe "Formado" são aqueles que conseguiram integralizar a carga horária prevista pelo curso. Por sua vez, o rótulo "Evadido" remete-se aos alunos que, por decisão própria ou processo de prescrição previsto em regimento da instituição, abandonaram a graduação. Os estudantes com matrículas ativas no SIGAA, porém que ultrapassaram um ano do prazo de conclusão estabelecido no currículo do curso foram classificados como "Retido". Existem na base de dados 65.758 (66,63%) amostras referentes a classe dos alunos formados; 25.581 (25,92%) dos registros, pertencem aqueles que desistiram dos estudos; e por fim, os alunos em retenção são menos representativos, 7.359 (7,46%).

Tabela 2. Correspondência entre a média das notas e o conceito

Conceito	Intervalo da média
Insuficiente (INS)	[0-4,99]
Regular (REG)	[5-6,99]
Bom (BOM)	[7-8,99]
Excelente (EXC)	[9-10]

4. Aplicação Proposta

Durante a etapa de *Data Mining*, foram testados algoritmos classificadores, a partir disso analisou-se a precisão global (acurácia) de cada um deles, para finalmente selecionar aquele que obteve uma taxa de acerto aceitável. Considerou-se ainda à seleção do algoritmo dois critérios: a representação dos resultados e o quanto esta informação pode ser interpretada por especialistas e usuários inseridos no domínio. Para estas finalidades, a Rede Bayesiana se mostra uma importante ferramenta, pelos seguintes aspectos: representação gráfica da relação entre estados; a rede expressa o conhecimento especialista acerca do domínio; e os resultados numéricos (probabilidades) podem ser visualizados através de gráficos.

A estratégia utilizada para segmentar a base de dados em conjuntos de treinamento e testes, destinados a estimar precisão e confiabilidade do modelo construído pelo classificador, foi a validação cruzada com k conjuntos estratificada (*stratified k-fold cross-validation*), por ser uma das mais empregadas em mineração de dados [Han et al. 2012].

Os algoritmos de aprendizado supervisionado [Rezende 2005] empregados nesta pesquisa estão disponíveis na ferramenta de código aberto (*open source*) Weka [Weka 2017]. Os classificadores estão divididos de acordo com as seguintes abordagens: árvores de decisão, probabilísticos, baseados em instâncias, baseados em funções e redes neurais artificiais. A Tabela 3 apresenta todos os métodos experimentados, as respectivas abordagens de construção do modelo e a configuração dos parâmetros de execução.

5. Resultados

5.1. Análise de desempenho dos algoritmos

A Tabela 4 apresenta os 9 algoritmos e as métricas usadas: tempos para treinar e testar modelo, acurácia e coeficiente Kappa. Os resultados mostram que a melhor solução foi conseguida através do indutor *Random Forest* cuja acurácia superou 87%, não obstante o algoritmo *Bayesian Network* revelou precisão global próxima de 86% e tempos aceitáveis para construção e testes do modelo, além disso este algoritmo obteve valor de estatística Kappa igual a 0,6961, considerado um nível substancial de concordância interobservador [Viera e Garrett 2005]. Destaca-se que aplicações nas quais o tempo de processamento é considerado requisito crucial ao domínio, soluções como *Multilayer Perceptron* e SVM são consideradas inviáveis, embora apresentem boas taxas de acerto.

Tabela 3. Abordagens para construção dos modelos, algoritmos classificadores e parâmetros de execução

Abordagens	**Algoritmos**	**Parâmetros**
Probabilístico	*Naive Bayes*	Não se aplica
Probabilístico	Redes Bayesianas (*Bayesian Network*)	Algoritmo de construção da rede: K2; Máximo número de pais em cada nós 5
Baseado em funções	*Support Vector Machine* (SVM)	Função *kernel* gaussiana: $\exp(-\gamma * \|u\text{-}v\|^2)$; C = 1; γ = 1/k, seja k o número de instâncias.
Baseado em instâncias (*Instanced-based learning*)	*K-Nearest Neighbor* (KNN)	K=1
Redes Neurais Artificiais	*Multilayer Perceptron*	Tipo *Backpropagation*; Função de ativação sigmóide; Número de épocas = 500; Taxa de aprendizado = 0.3; *Momentum Rate* = 0.2
Árvores de Decisão	C4.5	Mínimo de instâncias por folha = 2; Limite de confiança para *prunning* = 25%
Árvores de Decisão	*Random Tree*	Número de iterações = 100; Profundidade máxima da árvore ilimitada; Mínimo de instâncias por folha = 1
Árvores de Decisão	*Random Forest*	Número de iterações = 100; Profundidade máxima da árvore ilimitada; Mínimo de instâncias por folha = 1
Árvores de Decisão	*Classification And Regression Trees* (CART)	Número máximo de instâncias em nós terminais = 2

Diferentemente do método *Naive Bayes*, por exemplo, que serve como um classificador natural, a rede Bayesiana necessita ter uma boa precisão para ser aplicado ao domínio, e os testes comprovaram a sua eficiência quando comparada às técnicas

clássicas. A escolha pela rede Bayesiana é satisfatória aos objetivos desta pesquisa, porquanto, neste experimento: não penalizou tempo de construção e testes do modelo; demonstrou taxa de acerto adequada se confrontada as demais; e agrega conhecimento especialista sobre o domínio em representação gráfica. Diante do exposto, apresentam-se, na Subseção 5.2, a geração da RB e o conhecimento extraído da própria topologia da rede, por intermédio da inferência probabilística.

Tabela 4. Métricas de desempenho geral dos classificadores

Algoritmos	Tempo para treino (segundos)	Tempo para teste (segundos)	Acurácia (%)	Kappa
Naive Bayes	0.31	2.14	78.7736	0.5688
Bayesian Network	6.77	1.26	85.865	0.6961
KNN	0.29	1153.27	83.8041	0.6483
SVM	1418.69	1288.21	86.6938	0.6999
Multilayer Perceptron	4739.78	3.56	86.2054	0.7048
C4.5	1.62	1.44	86.2449	0.6984
Random Tree	0.55	1.48	80.5021	0.5924
Random Forest	12.31	8.24	87.1102	0.7118
CART	236.39	0.63	86.5104	0.7023

5.2. Análise da evasão e retenção via Redes Bayesianas

Foram selecionados os 14 atributos mais relevantes, dispostos na Figura 1, além da classe (*status*), de acordo com ganho de informação [Han et al. 2012]. Após a redução no número de variáveis, aferiu-se novamente a acurácia do algoritmo *Bayesian Network*, apresentando precisão de 83,5%, ratificando a sua robustez. O algoritmo de busca gulosa (*greedy search*) K2 [Cooper e Herskovits 1992] foi empregado para construção da topologia da rede, atingindo-se maior precisão global com o parâmetro de número esperado de pais por nó definido a 5.

A Figura 1 mostra a rede Bayesiana resultante, as cores de fundo dos nodos estão diferentes com propósito de agrupar os tipos das variáveis conforme seus significados no domínio de aplicação. Os elementos em laranja, verdes e azuis são respectivamente, índices acadêmicos, probabilidades de o discente possuir o valor índice acadêmico menor ou igual aos formados, e os percentuais avaliados para determinado conceito (notas).

A partir da rede apresentada na Figura 1, percebe-se que o desempenho estudante depende diretamente do número de trancamentos, do Índice de Eficiência Acadêmico (IEA), da probabilidade em relação a este índice, e do percentual de reprovações durante todo o curso. Com efeito, essas relações fazem sentido, uma vez que o aumento do número de trancamentos e baixo IEA implica, consequentemente, em uma elevação na probabilidade do índice (*probabilidade_iea*); isto é, o discente fica abaixo da expectativa de conclusão dos estudos em tempo hábil, ou ainda de concluir do curso. Vale destacar que, a partir da topologia apresentada, outras interpretações são possíveis e válidas no contexto pesquisado.

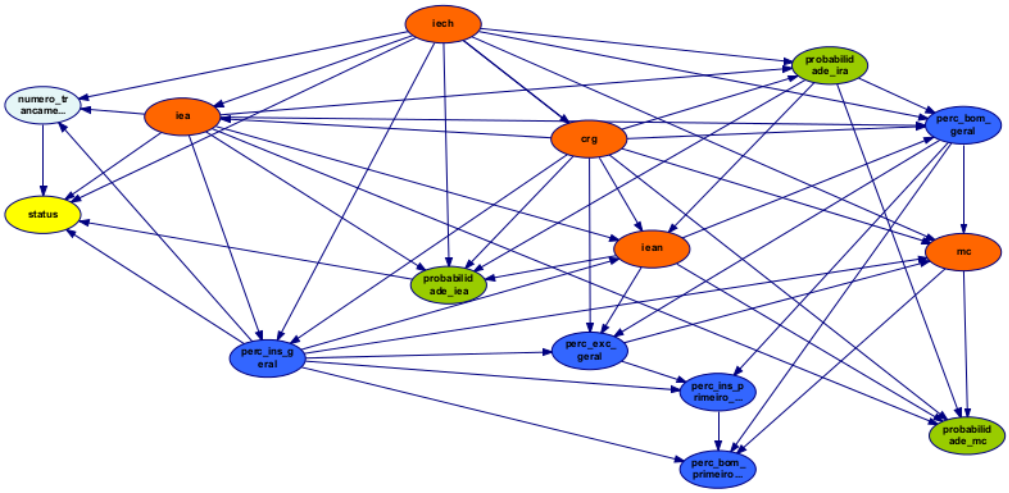

Figura 1. Rede Bayesiana construída para analisar a evasão e retenção no âmbito educacional

Para a inferência Bayesiana, escolheram-se os atributos IEA (*iea*) e número de trancamentos (*numero_trancamento*). A Tabela 5 evidencia as probabilidades, em cada categoria de estudantes, dadas evidências por classe (estado) dos atributos apresentados. Os estados, de todas as variáveis, foram conseguidos por intermédio da discretização com distribuição uniforme de frequência. Os atributos *numero_trancamento* e *iea* tiveram, respectivamente, três e quatro intervalos para conversão de dados contínuos em discretos, estas quantidades foram determinadas após análise dos dados.

Tabela 5. Probabilidades das três classes de estudantes segundo a inferência probabilística para os atributos IEA (iea) e número de trancamentos (numero_trancamentos)

Classes de alunos	Nº de Trancamentos	IEA			
		[0-3,95]	(3,905-6,435]	(6,435-7,975]	(7,975-10]
Formado	0	0.239	0.74	0.91	0.946
	1	0.092	0.518	0.675	0.706
	>1	0.024	0.252	0.421	0.436
Retido	0	0.153	0.116	0.045	0.016
	1	0.128	0.107	0.077	0.042
	>1	0.071	0.075	0.089	0.077
Evadido	0	0.608	0.144	0.045	0.038
	1	0.78	0.376	0.248	0.252
	>1	0.905	0.672	0.49	0.487

Os resultados na Tabela 5 mostram que o aumento do IEA e a redução do número de trancamentos são determinantes para o aluno concluir os estudos com êxito. Constata-se que o estudante que possui IEA superior a 7,975 e nenhum trancamento de matrícula possui aproximadamente 95% (0,946) de possibilidades para formar-se. Nota-se que a ocorrência de 1 trancamento e índice de eficiência acadêmica de regular a bom (6,435 a 7,975), constitui uma faixa limítrofe, pois estudantes nesta situação têm 67,5% de serem diplomados. Após 1 trancamento os discentes, com desempenhos de bom a

excelente, possuem 44% (0,436) de chances para concluírem os estudos sem entrar em situação de retenção.

Os graduandos que não trancaram a matrícula e possuem índice de eficiência acadêmica inferior a 3,95 apresentam 15,3% de chances à retenção. Dos resultados, observa-se que um estudante tende a extrapolar o prazo de conclusão proposto em currículo quando realiza 1 trancamento e possui um IEA entre 3,905 e 6,335. Estas situações são, empiricamente, observáveis em ambientes universitários, geralmente os alunos com baixo rendimento acadêmico (insuficiente ou regular) e que trancaram a matrícula, extrapolam o tempo de permanência na graduação em decorrência do acumulo de disciplinas o qual dificulta ainda mais a conclusão dos estudos.

Inversamente ao fato notado aos estudantes formados, a redução do IEA e o aumento do número de trancamentos são fortes indícios da evasão. Os resultados obtidos evidenciam que há uma alta probabilidade (90,5%) de o aluno abandonar o curso, caso realize mais de um trancamento da matrícula e tenha desenvolvido um IEA insuficiente (3,95). Evidencia-se com um IEA maior que 3,95 e no máximo 1 trancamento, o discente possuirá probabilidade menor a 38% de abandonar a graduação, uma situação tida segura para o discente concluir os estudos, ademais do risco da retenção.

6. Considerações Finais

Esta pesquisa utilizou mineração de dados sobre uma base de dados com quase cem mil registros acadêmicos dos discentes de graduação para entender as causas associadas ao abandono dos estudos e a permanência além do prazo estipulado para conclusão do curso. Neste trabalho foram testados nove algoritmos classificadores, sendo que o método *Random Forest*, apresentou a melhor acurácia, superior a 87%. Contudo, priorizou-se a escolha de um classificador capaz de possuir fácil representação de resultados e que esta possa expressar o conhecimento do especialista sobre o domínio estudado. Nessa perspectiva, o classificador *Bayesian Network* foi escolhido e, ratificou sua escolha por também obter desempenho satisfatório, visto que sua precisão global ultrapassou 85%.

A Rede Bayesiana construída mediante o uso do método de buscas K2 viabilizou a extração de importantes conhecimentos a respeito dos problemas analisados, permitindo a sua vinculação ao índice de eficiência acadêmica e a interrupção da matrícula em período letivo. Os resultados alcançados não são exaustivos, dessa forma outras pesquisas são necessárias para consolidar as respostas acerca dos principais fatores ligados à evasão e retenção em âmbito universitário. Como trabalhos futuros, serão realizadas novas investigações com atributos adicionais de caráter socioeconômicos, fato que permitirá relacionar o desempenho acadêmico a situações de vulnerabilidades sociais e, principalmente, quantificar o impacto dessa dependência, o que propiciará aos gestores a criação de mecanismos eficazes de combate à evasão e retenção.

Referências

Andriola, W. (2009). “Fatores associados à evasão discente na Universidade Federal do Ceará (UFC) de acordo com as opiniões de docentes e de coordenadores de cursos”.

Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 7(4):343-356.

Baker, R., Isotani, S., e Carvalho, A. (2011). "Mineração de Dados Educacionais: Oportunidades para o Brasil". Revista Brasileira de Informática na Educação, 19(2):3-13.

Baker, R. S. (2009). "Data Mining for Education". International Encyclopedia of Education, 3.

Cooper, G. F. e Herskovits, E. (1992). "A Bayesian Method for the Induction of Probabilistic Networks from Data". Mach. Learn., 9(4):309-347.

Cortez, P. e Silva, A. (2008). Using data mining to predict secondary school student performance. In *Proceedings of 5th Annual Future Business Technology Conference*, pages 5-12, Porto, Portugal.

Fayyad, U. M., PiatetskyShapiro, G., Smyth, P., e Uthurusamy, R., editors (1996). Advances In *Knowledge Discovery and Data Mining. American Association for Artificial Intelligence*, Menlo Park, CA, USA.

Frawley, W., PiatetskyShapiro, G., e Matheus, C. (1992). "Knowledge Discovery in Databases: An Overview". AI Magazine, pages 57-70.

Goldschmidt, R. e Passos, E. (2005). Data Mining: Um Guia Prático. Editora Campus.

Governo Federal (2007). "DECRETO Nº 6.096, DE 24 DE ABRIL DE 2007". https://www.planalto.gov.br/ccivil_03/_ato2007-2010/2007/decreto/d6096.htm .[Online; Acessado em 06/02/2017].

Governo Federal (2009). "Lei nº 12.089 de 11 de novembro de 2009". http://www.planalto.gov.br/ccivil_03/_Ato2007-2010/2009/Lei/L12089.htm. [Online; Acessado em 15/04/2016].

Han, J., Kamber, M., e Pei, J. (2012). Data Mining: Concepts and Techniques. Morgan Kaufmann, 3 edition.

INEP (2014a). "Censo da Educação Superior 2014 Notas Estatísticas". http://download.inep.gov.br/educacao_superior/censo_superior/documentos/2015/notas_sobre_o_censo_da_educacao_superior_2014.pdf. [Online; Acessado em 16/04/2016].

INEP (2014b). "Resumo Técnico Censo da Educação Superior 2013". http://download.inep.gov.br/download/superior/censo/2013/resumo_tecnico_censo_educacao_superior_2013.pdf. [Online; Acessado em 16/04/2016].

JEDM (2016). "Journal of Educational Data Mining". http://www.educationaldatamining.org/JEDM. [Online; Acessado em 26/01/2016].

Luo, Q. (2008). Advancing Knowledge Discovery and Data Mining. In *Knowledge Discovery and Data Mining*, 2008. WKDD 2008. First International Workshop on, pages 3-5.

Manhães, L. M. B., da Cruz, S. M. S., Costa, R. J. M., Zavaleta, J., e Zimbrão, G (2012). Identificação dos Fatores que Influenciam a Evasão em Cursos de Graduação

Através de Sistemas Baseados em Mineração de Dados: Uma Abordagem Quantitativa. In *Simpósio Brasileiro de Sistemas de Informação*, pages 468-479.

Manhães, L. M. B., da Cruz, S. M. S., Zavaleta, J., e Zimbrão, G. (2014a). "Investigating Withdraw of STEM Courses in a Brazilian University with EDM". Symposium on Knowledge Discovery, Mining and Learning (KDMILE), pages 1-8.

Manhães, L. M. B., da Cruz, S. M. S., Zavaleta, J., e Zimbrão, G. (2014b). "The Impact of High Dropout Rates in a Large Public Brazilian University". CSEDU – 6th International Conference on Computer Supported Education, pages 126-129.

Manhães, L. M. B., da Cruz, S. M. S., Zavaleta, J., e Zimbrão, G. (2015). "Towards Automatic Prediction of Student Performance in STEM Undergraduate Degree Programs". pages 247–253. Symposium of Applied Computing (SAC 2015), SAC.

MEC (2007). "Reestruturação e Expansão das Universidades Federais: Diretrizes Gerais". http://portal.mec.gov.br/sesu/arquivos/pdf/diretrizesreuni.pdf. [Online; Acessado em 15/04/2016].

Rezende, S. O. (2005). Sistemas Inteligentes: Fundamentos e Aplicações. Manole. Rich, E. e Knight, K. (1993). Inteligência Artificial. Makron Books.

Romero, C. e Ventura, S. (2010). "Educational Data Mining: A Review of the State of the Art". Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 40(6):601-618.

Sachin, R. e Vijay, M. (2012). A Survey and Future Vision of Data Mining in Educational Field. In *Advanced Computing Communication Technologies (ACCT), 2012 Second International Conference on*, pages 96–100.

Silva Filho, R. L. L., Motejunas, P. R., Hipólito, O., e de Carvalho Melo Lobo, M. B. (2007). A Evasão no Ensino Superior Brasileiro. In *Cadernos de Pesquisa*, volume 37, pages 641-659. Fundação Carlos Chagas.

Tinto, V. (1975). "Dropout from higher education: a theoretical synthesis of recent research". Review of Educational Research, pages 89–125.

Tinto, V. (1987). Leaving college: rethinking the causes of student attrition. University of Chigago Press.

UFRN (2013). "Resolução Nº 171/2013 CONSEPE – Regulamento dos Cursos Regulares de Graduação da Universidade Federal do Rio Grande do Norte".

Viera, A. e Garrett, J. (2005). "Understanding interobserver agreement: The kappa statistic". Family Medicine, 37(5):360-363.

Weka (2017). "Weka 3: Data Mining Software in Java". http://www.cs.waikato.ac.nz/ml/weka/ [Online; Acessado em 06/02/2017].