Mineração de Dados para Identificação de Alunos com Alto Risco de Evasão: Um Estudo de Caso

Luciano Antonio Digiampietri*, Fabio Nakano, Marcelo de Souza Lauretto

Sistemas de Informação
Escola de Artes, Ciências e Humanidades da Universidade de São Paulo

* Autor para correspondência: luciano.digiampietri@gmail.com

RESUMO

A evasão estudantil no ensino superior é uma questão crítica nas universidades brasileiras. Existem diversos fatores potenciais relacionados a esse fenômeno, sendo alguns deles mitigáveis pelas instituições por meio de aperfeiçoamentos nos cursos e políticas de apoio aos alunos. Neste trabalho, apresentamos uma metodologia baseada em mineração de dados para o acompanhamento e identificação precoce dos estudantes com grande potencial de desistência ou desligamento compulsório. Nossa abordagem utiliza exclusivamente o histórico de desempenho nas disciplinas do primeiro ano do curso, dispensando fontes externas de dados mais escassas ou de difícil obtenção. Tais instrumentos podem balizar a tomada de ações individuais direcionadas a alunos em risco, bem como o planejamento de ações futuras. Apresentamos um estudo de caso conduzido sobre o histórico escolar de alunos do Bacharelado em Sistemas de Informação da USP, com índices de acerto superiores a 90%.

Palavras-chave: Predição de Evasão Universitária; Mineração de Dados em Educação.

ABSTRACT

The student dropout in higher education is a critical issue for the Brazilian universities. Among the several known factors associated with this phenomenon, some of them are manageable by the universities, through improvements in courses and student support policies. This work introduces a method based on data mining for students monitoring and early identification of those with high dropout risk. Our approach uses only the academic performance of students in the first year in the course, requiring no external data sources – which are not always available. The proposed approach supports decision making for specific actions with the students, as well as planning of future actions. We present a case study conducted at the Bachelor in Information Science of the University of São Paulo, which achieves accuracy higher than 90%.

Keywords: University Dropout Forecasting; Education Data Mining.

Introdução

A evasão de estudantes no ensino superior é uma questão crítica nas universidades brasileiras, tendo como consequências a formação de profissionais abaixo da capacidade instalada (o que limita, portanto, o impacto social das instituições), frustração dos alunos que não conseguem concluir sua graduação e significativo desperdício de recursos (HIPÓLITO, 2011).

Particularmente na Universidade de São Paulo, a razão entre alunos concluintes e o total de vagas oferecidas em 2013 foi de apenas 12,8%, com uma média de alunos concluintes por docente ativo de 1,24 (USP, 2014). Na Escola de Artes, Ciências e Humanidades (EACH-USP), a taxa estimada de evasão[1] em 2013 foi de 12%, enquanto no Bacharelado em Sistemas de Informação esse número foi de 13,3%. Isso significa dizer que, em média, de cada 180 alunos ingressantes anualmente no curso, aproximadamente 24 não conseguiram se graduar.

AMBIEL (2015) aponta, a partir de revisão de literatura, alguns dos fatores mais comumente relacionados com a evasão: baixa qualidade do ensino recebido antes de entrar na graduação; insatisfação com as relações sociais estabelecidas com colegas, professores e funcionários da instituição; não oferecimento de atividades extracurriculares; necessidade de trabalhar ou dependência financeira

para custear os estudos; características sociodemográficas familiares; falta de orientação para escolha do curso; e defasagem entre o término do ensino médio e o ingresso na graduação. O autor construiu, a partir de 81 itens que versavam sobre motivos que poderiam influenciar a decisão de uma pessoa de desistir de seu curso superior, uma escala de motivos para evasão do ensino superior, obtendo quatro agrupamentos de motivos: 1) Motivos institucionais, que englobam a qualidade do corpo docente, seu relacionamento com alunos, falta de oferecimento de certos serviços e aspectos de infraestrutura da instituição; 2) Motivos pessoais, que englobam incertezas a respeito de estar no curso certo e aspectos familiares; 3) Motivos relacionados a dificuldades financeiras e/ou dificuldade de conciliar os estudos e o trabalho; 4) Motivos relacionados a incertezas quanto à carreira futura, tanto em termos de realização pessoal como aspectos do mercado de trabalho.

BARKER *et al.* (2014) apresentam um estudo envolvendo cursos de computação de 14 universidades norte-americanas, com o objetivo de determinar os principais fatores que levam um estudante a evadir, de forma que esta potencial evasão possa ser predita e, talvez, evitada. Dentre as conclusões sobre os dados analisados, os autores revelam que um dos principais fatores para se evitar a evasão é o uso de atribuições significativas e relevantes aos estudantes; ou seja, se o discente não perceber a relevância de uma disciplina ou qualquer outra atividade exigida em sua formação e/ou atuação profissional, ele facilmente poderá perder o interesse pela disciplina e, eventualmente, pelo curso. Outros fatores incluem uma boa interação/relacionamento entre cada aluno e o corpo docente, bem como o desenvolvimento de atividades em grupo.

O conhecimento dos fatores institucionais que induzem a evasão tem sido de fundamental importância para tomadas de ações visando a aumentar o envolvimento dos alunos e diminuir o percentual de reprovações e evasão dentro do Bacharelado em Sistemas de Informação da USP. DIGIAMPIETRI *et al.* (2013) apresentam duas atividades desenvolvidas no curso, centradas em algoritmos e programação, para aumentar o envolvimento dos discentes no curso: a criação do campeonato anual de programação para calouros (BXCOMP) e a criação das disciplinas optativas de Desafios de Programação I e II que, em cada aula, apresentam um ambiente semelhante a competições como a Maratona de Programação[2]. Além da elevada avaliação positiva entre os graduandos participantes e do aumento de diferentes habilidades relatadas por estes em questionários de autoavaliação, os autores relatam que tais iniciativas incentivaram os alunos a participarem de competições nacionais. DIGIAMPIETRI *et al.* (2015) apresentam uma análise quantitativa dos impactos das mudanças em pré-requisitos na retenção dos alunos, algumas das quais já implementadas e outras em processo de implantação. LAURETTO *et al.* (2015) trazem os resultados de uma pesquisa com formandos e egressos do curso, conduzida com o objetivo de levantar informações sobre o grau de inserção profissional desses egressos e seu nível de satisfação em relação à formação recebida.

Adicionalmente às ações globais, revela-se importante a construção de instrumentos para acompanhar individualmente a trajetória dos estudantes no curso, a fim de identificar precocemente aqueles com grande potencial de desistência ou desligamento compulsório. Tais instrumentos têm importância tanto na tomada de ações individuais (aconselhamento e orientação do aluno) como também na previsão de retenções futuras, planejamento de oferecimento de turmas extras, entre outras.

Ao longo deste trabalho, o termo evasão será adotado para classificar o estudante que não conseguiu concluir a graduação, devido a abandono ou a desligamento compulsório (MANHÃES *et al.*, 2014).

Neste trabalho, apresentamos uma metodologia para a identificação de alunos com elevado risco de evasão, com base em seu histórico escolar nas disciplinas do primeiro ano do curso. Essa metodologia é baseada na mineração de dados, definida sucintamente como um conjunto de técnicas e procedimentos para identificação e representação de

padrões relevantes a partir de conjuntos de dados (WITTEN *et al*., 2011).

Trabalhos correlatos

Além das técnicas usuais para identificação de fatores ligados à evasão, a utilização de métodos estatísticos e computacionais para prever o desempenho de alunos também é bastante explorada na literatura. Duas vertentes principais são a previsão da nota final do aluno em uma disciplina (abordagem adotada especialmente no ensino a distância, em que se pode mensurar o nível de participação e desempenho do aluno durante a disciplina) e a previsão da evasão do curso. Um exemplo da primeira abordagem pode ser encontrado no trabalho de MEIER *et al.* (2016). São relatados a seguir alguns trabalhos com abordagem similar à proposta neste artigo, e os respectivos resultados obtidos para efeitos de comparação.

SILVA & ADEODATO (2012) propõem um modelo baseado em regressão logística que busca identificar, ao final do segundo semestre do curso, alunos com potencial risco de evasão. O estudo de caso analisou dados de seis cursos da Universidade Federal de Pernambuco. Vinte e uma novas variáveis foram derivadas dos dados originais, com o propósito de capturar indicadores de alto nível que fossem comuns a todos os cursos. Alguns exemplos das novas variáveis são: notas médias no 1º e 2º semestres; variação na taxa de reprovação do 1º para o 2º semestre; taxa de aprovação nas provas finais nos dois semestres etc. O procedimento atingiu um desempenho de 0,84 para a medida da área sob a curva ROC (em uma escala de 0 a 1).

MANHÃES *et al.* (2014) apresentam uma arquitetura para monitorar o progresso acadêmico de estudantes e avaliar o risco de evasão deles. Os autores conduzem um estudo de caso em três cursos tradicionais de engenharia, usando como base as notas e frequências em cada disciplina, bem como a situação dos alunos (ativo/desligado) em cada semestre. A partir dos dados originais, alguns indicadores adicionais foram derivados, tais como: quantidade de disciplinas em que o aluno foi aprovado no semestre; nota média nas disciplinas com aprovação; número de disciplinas em que o aluno foi reprovado por nota; média geral do aluno no semestre. Os autores observaram desempenho ligeiramente superior dos algoritmos de árvores de decisão, com acurácias variando de 90,3% a 96,1%.

LAN-OM & BOOGOEN (2014) observam que os bancos de dados tendem a incluir muitas variáveis redundantes (principalmente pela criação de novas variáveis derivadas a partir dos dados originais), e, em tais condições, alguns métodos de classificação têm seu desempenho degradado. Para mitigar esse problema, os autores propõem uma etapa de transformação dos dados antes da aplicação dos métodos de classificação, a fim de eliminar a redundância e aumentar a acurácia dos classificadores. Em uma análise comparativa envolvendo seis técnicas de transformação de dados e quatro algoritmos de classificação, a melhor acurácia obtida foi de aproximadamente 92%, combinando uma técnica de comitês de agrupamentos para transformação dos dados e redes neurais artificiais para classificação.

Metodologia

A metodologia deste trabalho foi dividida em seis atividades, que serão descrita a seguir: (i) obtenção de dados brutos; (ii) organização dos dados; (iii) identificação das informações de interesse; (iv) seleção da amostra; (v) execução de experimentos para a seleção de atributos e classificação; e (vi) análise dos resultados.

Na *obtenção de dados brutos*, foram obtidos 1.896 históricos escolares no formato PDF dos alunos e ex-alunos do bacharelado em Sistemas de Informação da EACH-USP no período de 2005 (início do curso) até 2015.

Para a *organização dos dados*, inicialmente os arquivos PDF foram convertidos para o formato texto mediante a utilização do conversor automático da ferramenta Adobe Acrobat Reader. Em seguida, foi utilizado um programa desenvolvido pelos autores para corrigir alguns problemas da conversão automática, bem como para extrair informações de interesse.

As *informações de interesse*, para cada aluno ou ex-aluno do curso, consideradas pelo presente trabalho são: ano de ingresso, situação atual (matrícula ativa, egresso, matrícula cancelada etc.), cada uma das disciplinas cursadas pelo aluno, sua frequência e nota. Ao todo, os 1.896 alunos cursaram 62.295 disciplinas.

Para a realização da identificação de padrões e estatísticas sobre os históricos escolares, faz-se necessário utilizar uma *amostra* composta por dados cujo resultado seja conhecido. Para este estudo, os resultados esperados são: o aluno conseguiu se formar (egresso) ou o aluno teve a matrícula cancelada. Dos 1.896 históricos obtidos, 627 correspondem a egressos, 400 a matrículas canceladas, 733 a alunos com matrícula ativa e 136 a alunos em outras situações (por exemplo, transferência de curso, reingresso e falecimento). Assim, foram utilizados dados de 1.027 históricos como amostra (egressos mais alunos com matrícula cancelada).

Foram realizados experimentos de *seleção de atributos*, isto é, identificação de quais as características mais relevantes para se prever a probabilidade de o aluno conseguir se formar, e experimentos de *classificação*, considerando três informações referentes às disciplinas obrigatórias do curso: a nota obtida pela primeira vez que o aluno cursou a disciplina; o semestre em que o aluno obteve nota 3,0 ou superior em cada disciplina; e o semestre em que ele foi aprovado em cada uma das disciplinas. Para o presente trabalho, tratou-se a predição como um problema de classificação binária, no qual, dadas as informações do histórico de um aluno, o classificador tenta identificar se esse aluno irá ou não se formar. Para este trabalho, foram utilizados os seletores de atributos e o classificador Rotation Forest, disponíveis no arcabouço Weka (HALL *et al.*, 2009).

Os *resultados* foram analisados considerando os atributos (disciplinas) mais relevantes para o problema de se prever se o aluno se formará ou não. A classificação foi avaliada de acordo com sua acurácia geral (isto é, a taxa de acertos do classificador ao estimar, com base nas informações de algumas disciplinas, se o aluno irá ou não se formar). Para o cálculo da acurácia (percentual de classificações corretas) foi utilizada a validação cruzada com dez subconjuntos, técnica que consiste em particionar o conjunto de dados em dez subconjuntos; destes, nove são utilizados para treinamento do algoritmo e um para teste. O processo é repetido dez vezes, alternando-se o conjunto utilizado para os destes.

Resultados

Nesta subseção são apresentados os resultados dos seletores de atributos e do classificador, considerando três informações referentes às disciplinas cursadas pelos alunos: a nota obtida na primeira vez que o aluno cursou a disciplina; o semestre em que o aluno obteve nota 3,0 ou superior em cada disciplina; e o semestre em que ele foi aprovado em cada uma das disciplinas. Os resultados apresentados a seguir consideraram apenas as cinco disciplinas específicas ministradas por professores do curso no primeiro ano do Bacharelado em Sistemas de Informação, a saber: ACH0021-Tratamento e Análise de Dados/Informações; ACH2001-Introdução à Programação; ACH2011-Cálculo I; ACH2002-Introdução à Análise de Algoritmos; e ACH2012-Cálculo II.

Seleção de Atributos

Considerando-se a nota obtida na primeira vez em que o aluno cursou cada uma das disciplinas analisadas, as disciplinas consideradas mais informativas (em relação às chances de o aluno se formar) identificadas pelos seletores de atributos foram, em ordem: ACH2001, ACH2011 e ACH2002.

Já em termos do semestre em que o aluno obteve nota 3,0 ou superior em cada disciplina, a situação se altera um pouco. Introdução à Análise de Algoritmos (ACH2002) passa da terceira posição para a segunda. Assim, as três disciplinas mais informativas, segundo os seletores, são: ACH2001, ACH2002 e ACH0021.

Por fim, considerando o semestre em que o aluno foi aprovado em cada uma das cinco disciplinas, a ordem de importância delas, segundo os seletores de atributo é a seguinte: ACH2001,

ACH2002 e ACH2012. Isto é, têm-se inicialmente as duas disciplinas da área de computação/programação do curso seguidas por Cálculo II (que não aparecia entre os três primeiros resultados dos seletores nas análises anteriores).

Predição de Desempenho/Classificação:

Nesta subseção, apresentam-se os resultados da acurácia do classificador Rotation Forest utilizando validação cruzada em dez subconjuntos, considerando tanto disciplinas individuais como subconjuntos das disciplinas analisadas.

A Tabela 1 traz os resultados de acurácia considerando as disciplinas individualmente e alguns subconjuntos de disciplinas de acordo com os três aspectos analisados neste artigo: nota obtida na primeira vez em que o aluno cursou cada uma das disciplinas, semestre em que o aluno obteve ao menos 3,0 em cada disciplina e semestre em que ele foi aprovado na disciplina. As células são coloridas em uma escala do vermelho para o verde, de acordo com os valores de acurácia. Observa-se que, individualmente, a disciplina Introdução à Análise de Algoritmos (ACH2002) é a que permite uma classificação mais precisa, seguida por Cálculo II (ACH2012). Ainda se observando as disciplinas individualmente, é possível detectar que a informação responsável por tornar a classificação mais precisa foi o semestre em que o aluno foi aprovado em ACH2002 com acurácia acima de 89% (isto é, ao se saber em qual semestre o aluno foi aprovado nesta disciplina é possível prever, com 89% de precisão, se o aluno irá ou não se formar).

Ao se analisarem os conjuntos formados por mais de uma disciplina, observa-se que os melhores conjuntos são aqueles formados por todas as disciplinas (obtendo acurácia de 91,72% para o semestre em que o aluno obteve aprovação nessas disciplinas) e o conjunto formado por ACH2002 e ACH2012 (obtendo acurácia de 91,63% considerando o mesmo tipo de informação).

Conclusão

Dentre as diversas atividades que visam a reduzir a evasão está a identificação precoce dos alunos com maior probabilidade de desistir do curso superior em que estão matriculados.

No presente trabalho foi apresentada uma metodologia baseada em mineração de dados para classificar alunos quanto ao risco de evasão, a partir do histórico escolar destes graduandos. Foram analisados os históricos de mais de mil alunos do Bacharelado de Sistemas de Informação da EACH-USP, sendo possível predizer o resultado do aluno no curso com uma acurácia superior a 90% ao se considerar em que semestre ele conseguiu ser aprovado, por exemplo, em duas disciplinas obrigatórias de seu primeiro ano de curso: Cálculo II e Introdução à Análise de Algoritmos. O desempenho obtido neste artigo é similar aos encontrados na literatura.

A utilização de classificadores ou regressores pode ser empregada pelas Comissões Coordenadoras de Curso não apenas para identificar se o aluno está ou não em risco de evasão, mas também para tentar quantificar esse risco, o que possibilitaria às

	ACH0021	ACH2001	ACH2011	ACH2002	ACH2012	ACH0021, ACH2001, ACH2011	ACH0021, ACH2001, ACH2011, ACH2002, ACH2012	ACH2001, ACH2011	ACH2002, ACH2012	ACH2001, ACH2002
Nota Obtida	65,92%	74,10%	69,13%	82,28%	75,95%	70,30%	79,07%	71,18%	79,84%	78,09%
Semestre da obtenção da nota 3,0	67,09%	76,83%	76,05%	84,62%	82,47%	81,01%	88,32%	81,30%	87,83%	84,32%
Semestre de aprovação	68,06%	76,34%	76,05%	89,09%	86,76%	79,65%	91,72%	79,55%	91,63%	88,22%

Tabela 1 – Resultados do classificador.

CoCs o desenvolvimento de medidas específicas para conjuntos de alunos com riscos diferentes.

Agradecimentos

Agradecemos à Pró-Reitoria de Graduação da USP e à CoC do Bacharelado de Sistemas de Informação da EACH-USP.

Notas

1 Taxa de evasão estimada pela percentagem de alunos matriculados em 2012 que, não tendo se formado, também não se matricularam em 2013 (SILVA FILHO *et al.*, 2007); cálculos realizados com base nas informações do Anuário Estatístico da USP (USP, 2014) e do relatório *EACH em Números* (LATIF, 2013).

2 Trata-se da Maratona de Programação promovida pela Sociedade Brasileira de Computação. Cf. <http://maratona.ime.usp.br/>

Referências Bibliográficas

AMBIEL, Rodolfo A. M. "Construção da Escala de Motivos para Evasão do Ensino Superior". *Aval. Psicol.*, Itatiba, vol. 14, n. 1, abr. 2015, pp. 41-52.

BARKER, Lecia; HOVEY, C. L.; THOMPSON, L. D. "Results of a Large-Scale, Multi-institutional Study of Undergraduate Retention in Computing". Frontiers in Education Conference (FIE), IEEE, 2014, pp. 1-8.

DIGIAMPIETRI, Luciano Antonio; PERES, S.M.; NAKANO, F.; ROMAN, N.T.; WAGNER, P.K.; SILVA, B.B.; TEODORO, B.; SILVA-JUNIOR, D.F.P.; PEREIRA, G.V.A.; BORGES, G.O.; PEREIRA, G.R.; SANTOS, M.V.; BAKLISKY, M.; BARROS, V.A. "Complementando o Aprendizado em Programação: Revisitando Experiências no Curso de Sistemas de Informação da USP". *iSys: Revista Brasileira de Sistemas de Informação*, vol. 6, 2013, pp. 5-29.

DIGIAMPIETRI, Luciano Antonio; LAURETTO, M.S.; NAKANO, F. "Análise do Histórico Escolar dos Estudantes Visando à Adaptação Curricular e do Processo de Ensino e Avaliação". In: 1º CONGRESSO DE GRADUAÇÃO DA UNIVERSIDADE DE SÃO PAULO, 2015. *Resumos.* São Paulo: Universidade de São Paulo, 2015, pp. 265-266.

DIGIAMPIETRI, Luciano Antonio; LAURETTO, M.S.; NAKANO, F. "Aperfeiçoamento da Estrutura Curricular de um Bacharelado em Sistemas de Informação: Metodologia e Resultados sobre a Análise de Pré-requisitos". Submetido. 2016

HALL, Mark; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I H. "The WEKA Data Mining Software: An Update". *SIGKDD Explorations*, vol. 11, n. 1, 2009.

HIPÓLITO, O. "O Gargalo do Ensino Superior Brasileiro: Depoimento". [27 abr. 2011]. *Carta Capital.* Entrevista concedida a Fernando Vives.

LAN-ON, Natthakan & BOONGOEN, Tossapon. "Using Cluster Ensemble to Improve Classification of Student Dropout in Thai University". In: *Proceedings of the 7th International Conference on and Advanced Intelligent Systems* (ISIS), 2014, pp. 452-457.

LATIF, Sumaia Abdel. *EACH em Números*. São Paulo: EACH-USP, 2013. Disponível em: <http://each.uspnet.usp.br/site/download/each-numeros-graduacao-2012.pdf>. Acesso em 29 mar. 2016.

LAURETTO, Marcelo de Souza; DIGIAMPIETRI, L. A.; NAKANO, F. "Pesquisa com Formandos e Egressos do Bacharelado em Sistemas de Informação da EACH-USP: Resultados Preliminares". In: 1º CONGRESSO DE GRADUAÇÃO DA UNIVERSIDADE DE SÃO PAULO, 2015. *Resumos.* São Paulo: Universidade de São Paulo, 2015, pp. 243-244.

MANHÃES, Laci Mary Barbosa; CRUZ, S.M.S.; ZIMBRÃO, G. "WAVE: An Architecture for Predicting Dropout in Undergraduate Courses Using EDM". 29th ANNUAL ACM SYMPOSIUM ON APPLIED COMPUTING (SAC '14). Nova York: ACM, 2014. pp. 243-247.

MEIER, Yannick; XU, J.; ATAN, O.; SHCAAR, M. van der. "Predicting Grades". *IEEE Trans on Signal Processing*. Hoes Lane, Piscataway, NJ, vol. 64, n. 4, 2016, pp. 959-972.

SALAZAR URIBE, Juan Carlos; LOPERA GOMEZ, C. M.; JARAMILLO ELORZA, M. C. "Identification of Factors that Affect the Loss of Student Status Using a Logit Survival Model for Discrete Data". *Dyna rev.fac.nac.minas,* Medellín, vol. 79, n. 171, fev. 2012.

SILVA, Hadautho Roberto Barros da & ADEODATO, P. J. L. "A Data Mining Approach for Preventing Undergraduate Students Retention". In: THE 2012 INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS (IJCNN), 2012, pp. 1-8.

SILVA FILHO, Roberto Leal Lobo; MOTEJUNAS, R.; HIPOLITO, O.; LOBO, M.B.C.M. "A Evasão no Ensino Superior Brasileiro". *Cadernos de Pesquisa*, São Paulo, vol. 37, n. 132, set./dez. 2007, pp. 641-659.

UNIVERSIDADE DE SÃO PAULO. *Anuário Estatístico USP*. São Paulo: VREA/USP, 2014. Disponível

em: <https://uspdigital.usp.br/anuario/br/acervo/AnuarioUSP_2014.pdf>. Acessado em 28 mar. 2016.

WITTEN, Ian H.; FRANK, E. HALL, M.A. *Data Mining*. *Practical Machine Learning Tools and Techniques*. 3. ed. Burlington, Massachusetts: Morgan Kaufmann, 2011.

Publicado em 05/07/2016.