Evasão em Nível de Disciplinas na Ciência da Computação na Universidade Federal do ABC

Bruno Aristimunha, Jairo da Silva Freitas Júnior, Harlen Costa Batagelo, Denise Goya, Carlos da Silva dos Santos

Universidade Federal do ABC - Av. dos Estados, 5001 - Bangú, Santo André - SP, 09210-580

b.aristimunha@gmail.com, jairo.freitas@aluno.ufabc.edu.br, harlen.batagelo@ufabc.edu.br, denise.goya@ufabc.edu.br, carlos.ssantos@ufabc.edu.br

Resumo: Este trabalho apresenta um estudo de caso que identificou padrões na evasão (reprovações por faltas) ao nível de aluno através de algoritmos de aprendizado de máquina. Os conjuntos de dados analisados foram coletados diretamente do sistema acadêmico de uma Universidade Federal, sendo posteriormente pré-processados e agrupados. Os dados foram minerados no contexto de classificação, utilizando os algoritmos interpretáveis de Florestas Aleatórias e Regressão Logística.

1. Introdução

A evasão no Ensino Superior possui grande destaque em pesquisas de gestão educacional [1]. Dentre as metas do Plano Nacional da Educação (PNE) do Brasil há a redução da evasão em diferentes níveis escolares [2]. Esse fenômeno multifacetado possui categorização e interpretações diversas [1,3]. No Ensino Superior, esse processo está diretamente ligado aos atributos internos e externos das Instituições de Ensino Superior (IES), sendo o primeiro dependente da infraestrutura, corpo docente e assistência estudantil; e o segundo, relacionados ao indivíduo, salientando-se vocação, aspectos socioeconômicos e problemas de ordem pessoal [4].

A evasão é um problema histórico dos cursos superiores brasileiros públicos; em 1996, a Comissão Especial de Estudo sobre Evasão nas Universidades Públicas identificou três categorias: (i) evasão de curso, situação na qual o aluno desliga-se sem concluir a grade, característico em transferências internas ou reingresso em outro curso; (ii) evasão da Instituição, quando o indivíduo realiza a transferência para outra instituição ou quando é aprovado em outra IES; e (iii) desistência, condição na qual o estudante não continua os seus estudos [5].

Sobre a amplitude desse evento, adotamos a definição de [6, 7] ao entender que a evasão não se restringe apenas ao evento de desligamento da universidade, mas ao abandono ou reprovação do estudante em qualquer etapa do curso (ou em uma disciplina), independente do motivo. Independente da causa, as perdas pela evasão acarretam custos para a Universidade e, se pública, o recurso investido perde o seu potencial de retorno.

Ao implementar o Bacharelado Interdisciplinar (BI) em 2006, a Universidade Federal do ABC (UFABC) possibilita ao estudante a escolha de formação em um curso superior dentro de uma grande área do conhecimento. Essa ação pode acarretar estranhamentos no relacionamento com os cursos, uma vez que o BI exige que o estudante dialogue e se identifique nas diferentes áreas do conhecimento. Ao término do BI o estudante ainda pode escolher entre os cursos pós-BI da instituição. Esses cursos pós-bacharelados são sobrepostos aos Bacharelados Interdisciplinares e, desse modo, independente do curso, há uma oferta de disciplinas que não compõem o seu currículo usualmente. Se, por um lado, esse modelo pode minimizar a evasão da Universidade ou de seus cursos, por outro, esse processo de redescoberta do curso acarreta em um tempo maior de permanência na instituição, maximizando a possibilidade de retenção.

Nesse sistema, na UFABC, o estudante opta em que momento irá cursar as disciplinas obrigatórias, limitadas e livres do BI, e/ou se irá cursar as do pós-BI. Essa liberdade gera padrões de formações singulares nas trajetórias acadêmicas (históricos), tendo em vista que, por período acadêmico, o aluno precisa escolher quais disciplinas cursar entre as ofertadas. Cada disciplina é estruturada de forma diferente considerando os Projetos Pedagógicos dos cursos da instituição. Uma mesma disciplina pode ser obrigatória a um (ou mais) curso(s) pós-BI, mas ser livre ao BI.

Nesse contexto, há insuficiência de dados que possibilite a análise e compreensão dos elementos que podem causar a evasão, seja em disciplinas ou no curso como um todo. Como uma disciplina pode pertencer a todos os cursos da

universidade ou a apenas um, se faz necessário o estudo de abandono e retenção a nível individualizado, tendo em vista que cada trajetória acadêmica é singular.

Dada a complexidade do tema no contexto da UFABC, emerge a necessidade da explicação do evento com instrumentos robustos para um acompanhamento da retenção do estudante (reprovações) e seus sintomas. Esse processo de descoberta de conhecimento útil pode ser englobado na área de Mineração de Dados Educacionais. Aqui, pode-se, por meio de um processo de descoberta de conhecimento em uma base de dados, descobrir os seus fatores latentes. Assim, o objetivo desse trabalho é avaliar os padrões obtidos por diferentes classificadores interpretáveis buscando o evento da evasão em disciplina (Reprovação por falta) dos estudantes que já estabeleceram algum vínculo com o Bacharelado em Ciência da Computação.

2. Materiais e Métodos

Os dados analisados foram extraídos do sistema acadêmico da Universidade Federal do ABC (UFABC). A estrutura curricular da UFABC permite que alunos cursem disciplinas de um curso sem estarem formalmente vinculados a este. Para analisar o desempenho dos alunos do Bacharelado em Ciência da Computação (BCC) e também o desempenho médio em disciplinas de computação (independentemente do curso do aluno), nós optamos por construir três conjuntos de dados distintos para análise. O primeiro conjunto contém os alunos que possuem ou já possuíram a reserva de vaga no BCC e abrange todas as disciplinas já cursadas por esses. O segundo conjunto é definido como o histórico de toda pessoa que já cursou alguma disciplina ofertada pelo Bacharelado em Ciência da Computação. O terceiro conjunto conjunto consiste de um recorte do segundo, em que consideramos apenas as disciplinas que possuem algum vínculo com o curso (Obrigatórias e Limitadas).

Em cada conjunto de dados, um exemplo (instância) corresponde a uma matricula realizada por um estudante em determinada disciplina e o respectivo resultado (Aprovado, Reprovado por Conceito, Reprovado por Frequência). Os atributos preditores vinculados a cada exemplo são: SITUAÇÃO_NA_UNI, COD_CURSO, COD_DISCI, NOME_DISCI, CONCEITO, SITUACAO_DISCI, PERIODO, ANO, COD_TURMA, T_SEMANAL, P_SEMANAL, I_SEMANAL. O recorte temporal corresponde ao segundo quadrimestre de 2019. O início do histórico abrange de 2006 a 2019.

Com as bases de dados definidas, realizamos um pré-processamento dos dados para a remoção de eventuais valores faltantes. Foram criadas nessa etapa as seguintes variáveis derivativas relacionadas ao estudante e a disciplina. As variáveis criadas relacionadas aos estudantes são: tempo na universidade até o curso da disciplina em quadrimestres (QUAD_CURSADO); Idade quando cursou a disciplina (IDADE); Quantidade normalizada de créditos cursados no quadrimestre (Z_SCORE_CRED_QUAD); e Coeficiente de Rendimento acumulado até o quadrimestre anterior (CR_ATE_QUAD_ANTERIOR). Das variáveis criadas no contexto das disciplinas reunimos a porcentagem de reprovação e abandono por disciplina para todos os oferecimentos - (PORC_REPROV, PORC_ABANDO); e se o oferecimento ocorreu no Turno Noturno - (NOTURNO). Além disso, redimensionamos as variáveis através de uma normalização Min-Max, de modo a evitar um mau condicionamento dos dados.

Para estabelecer uma base de comparação com outros métodos, foi criado um classificador Base que decide aleatoriamente o rótulo de cada exemplo, de acordo com as probabilidades de classe (positiva/negativa) observadas no conjunto de treinamento. Cada conjunto de dados foi recortado em partições de treino e teste na proporção 80%/20%. Dado a natureza desbalanceada do problema, após o recorte das partições, nós realizamos um balanceamento do conjunto de treino empregando um método de *Oversampling*. O método escolhido realiza a criação sintética de instâncias da classe minoritária, *Synthetic Minority Oversampling* (SMOTE).

Para avaliação dos classificadores, foram calculados: acurácia, acurácia balanceada, precisão, sensibilidade (*recall*) e área sob a curva ROC para cada um dos modelos. O ajuste dos modelos de classificação empregou validação cruzada com dez (10) folhas (10-*fold cross validation*).

Foram aplicados três (3) modelos para analisar a melhor performance, sendo eles: Florestas Aleatórias e Regressão Logística (com penalização L1 e L2). No processo de validação cruzada, o método de regressão logística e a floresta aleatória otimizaram a métrica área sobre a curva ponderada. Performamos cem estimativas em cada folha da validação cruzada nas Florestas Aleatórias, observado o critério de *Gini*.

3. Resultados e Discussão

Observamos que o fenômeno de evasão em disciplina em todos os recortes experimentais são raros. No primeiro conjunto ele ocorre em 5.01% da disciplinas, no segundo há à ocorrência de 4.61%, e no último conjunto há 11.93%. No conjunto de dados com reserva de vaga temos um total de 31528 instâncias. Conforme exibido na tabela 1, temos

que todos os métodos não apresentam um resultado satisfatório na precisão. Com a relação das métricas podemos analisar que os classificadores possuem uma *recall* mediano, acertando boa porcentagem dos casos que apontam, mas dado uma precisão baixa temos que os métodos apontam muito falso positivos, mas dado o contexto de prevenção do fenômeno, o custo torna-se adequado. O segundo conjunto possui 701558 objetos, com os mesmos atributos do anterior. Já o terceiro conjunto, um recorte do segundo, possui 56475 instâncias (7.74% do segundo conjunto). Os resultados do segundo conjunto de dados preservam os padrões observados no primeiro. Há um aumento de precisão no último conjunto, correspondendo ao dobro do valor obtido anteriormente em todos os métodos. Todos os métodos ultrapassam o *baseline* estabelecido. Já com somente as disciplinas da computação temos que o cenário análogo ao recorte dos alunos que só cursaram disciplinas da Computação. Os melhores métodos foram a Regressão Logística, e a Floresta Aleatória.

	Alunos com Reserva de Vaga				**Alunos que já cursaram disciplina da Ciência da Computação**				**Alunos que já cursaram disciplinas da Ciência da Computação, *somente essas disciplinas***			
	Baseline	**RL L1**	**RL L2**	**FA**	**Baseline**	**RL L1**	**RL L2**	**FA**	**Baseline**	**RL L1**	**RL L2**	**FA**
Acurácia	0.4987	0.7530	0.7628	**0.9387**	0.5006	0.7333	0.7512	**0.8958**	0.4952	0.7075	0.7244	**0.8715**
Acurácia Balanceada	0.5064	0.7481	**0.7579**	0.6640	0.5002	0.7442	**0.7528**	0.692	0.497	0.7111	**0.7199**	0.6756
Precisão	0.0498	0.1334	0.1397	**0.3670**	0.0707	0.1765	0.1873	**0.3287**	0.1094	0.2327	0.2445	**0.4197**
Recall	0.5149	0.7426	**0.7525**	0.3597	0.4997	**0.7570**	0.7547	0.4547	0.4992	**0.7157**	0.7141	0.4241
ROC-AUC	0.5064	0.7481	**0.7579**	0.6640	0.5002	0.7442	**0.7528**	0.6920	0.4970	0.7111	**0.7199**	0.6756

Tabela 1: Métodos por Métricas nos conjuntos de dados. Na tabela temos que Regressão Logística - RL, com penalidade L1 ou L2; Floresta Aleatória - FA, e modelo Base Aleatório *Dummy*.

Para além de obter um algoritmo que possua bons valores em métricas, o método deve possuir uma interpretabilidade para possibilitar um diálogo com a literatura. Observamos na Figura 1 o ranque por atributo dado um teste de permutação (*Mean Decrease Accuracy - MDA*) para o melhor método, a regressão logística - L2. A melhor métrica, segundo o teste, consiste da ponderada até o quadrimestre anterior, indicando que o rendimento acumulado até então discrimina a ocorrência da evasão em disciplina. Chama atenção que a quantidade normalizada de créditos - (*Z_SCORE*) intercala com o padrão de abandono característico de cada disciplina - (*PORC_ABANDO*), isso indica que os alunos que possuem reserva de vaga, ou seja, já concluíram ou estão de via de conclusão do Bacharelado Interdisciplinar possuem um padrão de evasão dado número de créditos que estão cursando. A quantidade de quadrimestre cursados pelo aluno - (*QUAD_CURSADO*) aparece como uma das melhores variáveis para distinguir o conjunto de dados. Na outra extremidade, nós temos que o número de créditos Teóricos da disciplina - (*T_SEMANAL*), o quadrimestre - (*PERIODO*), e se a disciplina ocorre no noturno possui baixa importância, a depender do conjunto.

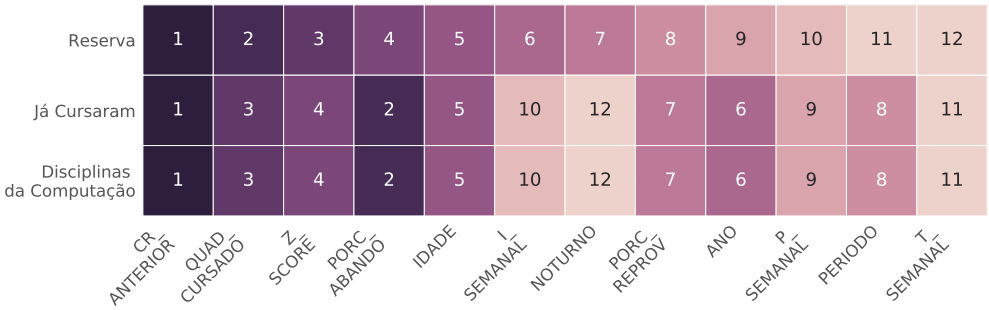

Figura 1: Posto da importância dos atributos dado pelo teste de permutação. Resultados nos diferentes conjunto de dados.

O atributo *PORC_ABANDO* apresentar uma alta importância para o melhor algoritmo, uma possivelmente explicação proposta em [8] diz que se discente não observa relevância do conteúdo em sua formação, este está suscetível à perda do interesse na disciplina, e eventualmente do curso. Para tanto, para sanar esse problema sugere-se que as disciplinas que apresentam maiores taxas de evasão em disciplina sejam encaminhadas para um processo de restruturação curricular.

Outro achado que também dialoga com a literatura está contido na importância do rendimento até o quadrimestre anterior, em [9] temos as variáveis mais discriminaram se os alunos evadiriam o curso são: notas médias no 1º e 2º Semestre. Nesse sentido, temos que o efeito observado nos primeiros anos se mantém para toda a graduação. A importância observada na variável *IDADE* destoa dos achados da literatura encontrados [10]. O tempo na universidade (*QUAD_CURSADO*) nos diz que todas as questões discutidas no amadurecimento no tempo de curso, também se fazem presentes [10].

4. Conclusão

A evasão ao nível de disciplina é um fenômeno complexo, tanto para o estudante, quanto para a instituição. Através desse trabalho apontamos direções para um estudo quantitativo com métodos de aprendizado de máquina para o processo de descoberta de padrões. No contexto de uma universidade interdisciplinar, com currículo flexível e singular, emergem questões que não se fazem presentes em outras instituições de ensino. Neste sentido, deve-se explorar técnicas que busquem associação nos conjuntos dados para melhor entendimento do macro cenário universitário.

Os experimentos obtiveram uma taxa de acurácia máxima de 93.87%, acurácia balanceada de 75.79%, precisão máxima de 41.97%, *recall* de 75.70% e área sobre a curva máxima de 75.79%. Muito além desses números, esse trabalho apresenta relações presentes no conjunto de dados que podem auxiliar no processo de tomada de decisão na gestão da universidade. Como trabalhos futuros, pretendemos investigar na literatura pedagógica os indícios que explicam esse acontecimento da evasão ao nível de disciplina, e analisar tais fatores nos dados universitários. Os resultados encontrados aqui refletem o recorte experimental e também as limitações dos dados, em outras palavras, não sendo generalizáveis para fora do curso de Ciência da Computação.

5. Referências

[1] R. S. Freitas, "A ocorrência da evasão do ensino superior: uma análise das diferentes formas de mensurar," Master's thesis, Programa de Pós-Graduação em Educação da Faculdade de Educação da Universidade Estadual de Campinas, 2016.

[2] M. A. A. S. Aguiar, "Avaliação do plano nacional de educação 2001-2009: questões para reflexão," *Educação & Sociedade*, vol. 31, no. 112, 2010.

[3] M. P. Bardagi, *Evasão e comportamento vocacional de universitários: estudo sobre desenvolvimento de carreira na graduação*. PhD thesis, Universidade Federal do Rio Grande do Sul. Instituto de Psicologia. Curso de Pós-Graduação em Psicologia do Desenvolvimento., 2007.

[4] A. S. Paredes and E. R. Durham, "Evasão do terceiro grau em curitiba," *Núcleo de Pesquisas sobre Ensino Superior Universidade de São Paulo*, 1994.

[5] BRASIL/MEC/SESU/ABRUEM/ANDIFES, "Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas," tech. rep., MEC, 1996. Comissão Especial de Estudos sobre a Evasão nas Universidades Públicas Brasileiras.

[6] M. d. C. Maia, F. d. S. Meirelles, and S. K. Pela, "Análise dos índices de evasão nos cursos superiores a distância do brasil," in *Anais do XI Congresso Internacional de Educação à Distância. Salvador; Bahia*, 2004.

[7] E. M. dos Santos and J. D. de Oliveira Neto, "Evasão na educação a distância: identificando causas e propondo estratégias de prevenção," *Revista Paidéi@-Revista Científica de Educação a Distância*, vol. 2, no. 2, 2009.

[8] L. Barker, C. L. Hovey, and L. D. Thompson, "Results of a large-scale, multi-institutional study of undergraduate retention in computing," in *2014 IEEE Frontiers in Education Conference (FIE) Proceedings*, pp. 1–8, IEEE, 2014.

[9] Hadautho Roberto Barros da Silva and Paulo Jorge Leitão Adeodato, "A data mining approach for preventing undergraduate students retention," in *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, June 2012.

[10] I. O. Pappas, M. N. Giannakos, and L. Jaccheri, "Investigating factors influencing students' intention to dropout computer science studies," in *Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education*, pp. 198–203, ACM, 2016.