Um Estudo sobre Motivadores e Deterrentes de Evasão no Curso de Sistemas de Informação

Joubert Alexandrino de Souza
IFES - Instituto Federal do Espírito Santo, Serra, ES, Brasil
joubert@ifes.edu.br

Karin Satie Komati
IFES - Instituto Federal do Espírito Santo, Serra, ES, Brasil
kkomati@ifes.edu.br

Jefferson Oliveira Andrade
IFES - Instituto Federal do Espírito Santo, Serra, ES, Brasil
jefferson.andrade@ifes.edu.br

ABSTRACT

Dropping out in higher education is a serious problem that has been investigated for decades and causes great harm to individuals, educational institutions, and society as a whole. This article presents a case study on the application of the survival analysis combined with the construction of predictive models in the identification motivating and deterrent elements of dropout in an undergraduate program in Information Systems at a public institution of higher education in Brazil. Methods of educational data mining and probabilistic modeling were applied to student data to model students' expected completion of the course, semester by semester. The results of the survival analysis are consistent with the literature and indicate the greatest risk of dropout is in the initial semesters of the course, while the identification of characteristics of dropout makes it clear that the subjects of the first two semesters are the biggest barriers of the course retaining about 50% of the student population.

KEYWORDS

Dropout in Higher Education, Survival Analysis, Educational Data Mining

1 INTRODUÇÃO

A evasão estudantil no ensino superior tem objeto de estudo há várias décadas Tinto [7]. Embora o fenômeno da evasão possa se manifestar a qualquer tempo, ele ocorre de forma mais severa no primeiro ano dos cursos [5], e somente nos Estados Unidos da América (EUA), gerou um prejuízo de cerca de US$ 9 bilhões de dólares para os governos estaduais e federal entre 2003 e 2008, impactando 1 em cada 3 alunos [2]. Índices de evasão semelhantes foram verificados no Brasil por Silva et al. [6] utilizando dados do Censo da Educação Superior produzido pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP).

A evasão se caracteriza como um problema de múltiplos fatores. Gilioli [4] analisou se o Sistema de Seleção Unificada (SiSU) teria contribuído de forma determinante para o aumento da evasão nas Instituições Federais de Ensino Superior (IES mantidas pelo setor público federal), concluindo, porém, não poder determinar tal relação dissociada de outros fatores como, por exemplo, as greves ocorridas nas instituições e as condições socio-econômicas dos alunos. No tocante às IES privadas do Brasil, concluiu-se que as causas mais significativas para evasão foram o estado civil, a idade e a falta de financiamento [3]. No trabalho de Hoffmann et al. [5], dentre os motivos apurados como causas da evasão, observa-se que a maioria têm origem em questões da própria IES, seguido de questões econômico-financeiras e de questões pessoais do estudante, sendo que tais causas podem ocorrer de forma concomitante.

2 SOLUÇÃO PROPOSTA

A fim de compreender melhor o impacto do fenômeno de evasão e para identificar os principais elementos motivadores de evasão em um curso superior, decidiu-se por realizar dois tipos de análise distintos:

(1) *Análise de sobrevivência de evasão dos alunos.* O objetivo da análise de sobrevivência é de compreender melhor como a evasão molda a expectativa de conclusão de curso por parte dos alunos, i.e., qual é a probabilidade, em um determinado momento, de que um aluno venha a concluir o curso. A análise de sobrevivência é uma ferramenta estatística que analisa e modela dados em que o resultado é o tempo até a ocorrência de um evento de interesse [8]. Neste trabalho, o evento é a evasão escolar.
(2) *Identificação de características mais relevantes na evasão dos alunos.* A hipótese de pesquisa é que a capacidade de seleção de características dos modelos preditivos pode ser utilizada como um bom indicador de quais são os elementos que se apresentam como os maiores motivadores ou deterrentes para a evasão. Para buscar determinar quais variáveis do processo de formação acadêmica tem maior impacto como motivadores ou como deterrente na evasão foram construídos modelos preditivos utilizando-se métodos do tipo *wrappers* e/ou *embedded* que possuem a propriedade de autoseleção de características. Foram usados três métodos: árvore de decisão, Gradient Boosting e XGBoost.

2.1 Base de dados coletada

O curso de Sistemas de Informação (SI) de uma IES pública brasileira, desde sua concepção em 2008 até os presentes dias, apresenta taxas de evasão anual muito altas. A Tabela 1 compila os dados da Taxa de Evasão Anual obtidos na Plataforma Nilo Peçanha (PNP)[1], para o anos base de 2017, 2018, 2019 e 2020. É importante ressaltar que os dados referentes ao ano de 2017, são de ingressantes do ano de 2013 e o tempo regular do curso é de 8 semestres.

[1] http://plataformanilopecanha.mec.gov.br/2020.html

Tabela 1: Taxa de Evasão Anual em BSI

Ano Base	Ingressantes	Concluintes	Evasão Anual
2017	113 (2013)	28	27, 8%
2018	102 (2014)	23	26%
2019	88 (2015)	1	9%
2020	90 (2016)	19	11,6%

O conjunto de dados foi obtido anonimizado junto à IES e estava dividido em dois subconjuntos: o conjunto de dados de alunos e o conjunto de dados de disciplinas. Ao todo o banco de dados continha 1.169 registros únicos de alunos matriculados entre os anos de 2008 e 2021. As características do conjunto de dados de alunos eram compostas de dados demográficos, dados do ensino médio (data da conclusão do ensino médio), dados da matrícula e dados do curso. Cada disciplina cursada pelo aluno representa um registro no conjunto de dados de disciplinas cujas características eram compostas de dados da matrícula (matrícula e situação da matrícula) que não são apresentados na tabela, dados da instituição e do curso e dados da disciplina.

2.2 Análise de Sobrevivência

Para realizar a análise de sobrevivência foi necessário produzir uma lista com os semestres em que aconteceram as evasões, originando o conjunto de dados **semestre evasão**, que contém Id, Matricula, Numsems_evade e Situacao. O campo *Numsems_evade* é o número do semestre em que o aluno evadiu. O campo *Situacao* é a codificação da ocorrência do evento a ser observado, sendo que o código inteiro 0 denota que não ocorreu evasão e 1 que ocorreu evasão.

A análise de sobrevivência usa probabilidade condicional, ou seja, a probabilidade de sobreviver até o tempo t, dado que um sujeito estava vivo no início de um intervalo de tempo especificado. Utilizou-se o estimador Kaplan-Meier para estimar a curva de sobrevivência [8].

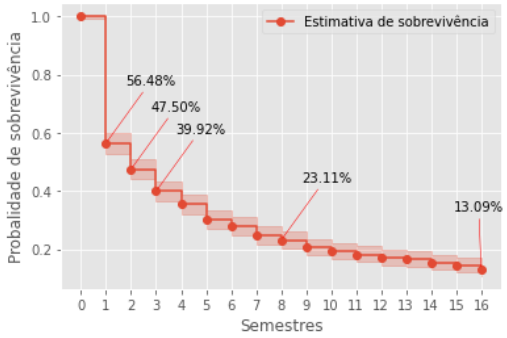

Figura 1: Função de sobrevivência para evasão.

A Figura 1 apresenta a função de sobrevivência e seus intervalos de confiança, já que a curva de sobrevivência é estimada. O gráfico apresenta as durações e as taxas de como ocorre o processo de evasão no curso de BSI. Observa-se que no primeiro semestre, a taxa de sobreviver a eventos e evasão é de 56,48%, no terceiro semestre é próximo à 40% (39,92%), ou seja, aproximadamente 60% dos alunos evadem até o início do terceiro semestre. Após 8 semestres de curso, tempo regular da formação, aproximadamente apenas 23% dos alunos conseguem sobreviver à eventos de evasão. O tempo máximo de curso que é de 16 semestres, e a taxa nesse momento é de cerca de 13%. Para esta população de alunos a média de sobrevivência calculada pelo estimador foi de apenas 2 semestres, ou seja, o fenômeno da evasão age de forma mais severa no período de 2 semestres haja visto a taxa de mortalidade de mais de 50% desta população neste período.

2.3 Seleção de Características

O processo de identificação de características mais relevantes tem duas etapas: (i) pré-processamento da base de dados e (ii) experimentação com os métodos de classificação.

2.3.1 Pré-processamento dos Dados. Os conjuntos de dados iniciais passaram por vários processos: integração de dados, tratamento de dados ausentes, criação de características, mapeamento de valores, discretização dos dados e análise de desbalanceamento. A base de dados resultante do processo foi denominada de alunobsi. As características do conjunto de dados de disciplinas foram usadas na criação de novas características. Todos os registros das disciplinas que foram ministradas após semestre letivo 2019/2 foram excluídas. Este procedimento foi necessário para remover os efeitos da pandemia do Coronavírus (COVID-19) dos dados acadêmicos.

Os 23 valores distintos do atributo alvo, *Situacao_Matricula*, foram consolidados em apenas 3 valores distintos: (0) concluído; (1) evasão; (2) matriculado. A análise exploratória dos dados revelou que o conjunto de dados estava desbalanceado do ponto de vista da variável alvo conforme pode ser visto na Figura 2.

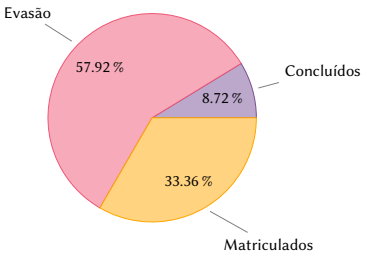

Figura 2: Distribuição da situação de matrícula.

Apesar deste desbalanceamento ter sido constatado, optou-se por não usar nenhuma estratégia de mitigação do desbalanceamento para os modelos preditivos utilizados nos experimentos. Porém, tomou-se o cuidado de realizar amostragens de dados estratificadas com o intuito de garantir a distribuição proporcional de classes.

Concluída a etapa de preparação dos dados, o conjunto de dados resultante, alunobsi, contém 1.169 registros e 202 características.

2.3.2 Métodos de Classificação. A identificação das características mais relevantes para a predição da evasão foi feita através do uso dos modelos de predição *Árvore de Decisão*, *Gradiente Boosting* e *XGBoost*, todos do tipo *ensemble*. Este modelos intrinsecamente atribuem um *score* de relevância para cada característica do conjunto de dados. Este *score* indica, a grosso modo, quanta informação aquela característica carrega para a predição da variável alvo.

Os classificadores foram treinados com 80% dos dados e testados com 20% dos dados restantes. Os dados de treinamento e teste foram gerados de modo estratificado. Foram obtidos os seguintes valores de acurácia: 88,46% para a árvore de decisão, 90,59% para Gradient Boosting e 90,59% para XGBoost.

Tabela 2: Características mais relevantes de acordo com o escore de entropia.

	Árvore de Decisão	Gradient Boosting	XGBoost
1	Ano_Letivo_Ini	Ano_Letivo_Ini	CSI_037_Aprovado
2	CSI_040_Aprovado	CSI_040_Aprovado	CSI_004_Rep_Falta
3	CSI_009_Aprovado	CSI_009_Aprovado	CSI_040_Aprovado
4	CSI_004_Rep_Falta	CSI_004_Rep_Falta	CSI_043_Rep_Falta
5	Idade_Ingresso	CSI_037_Aprovado	CSI_009_Aprovado
6	CSI_003_Rep_Falta	CSI_077_Aprovado	CSI_077_Aprovado
7	CSI_002_Rep_Nota	CSI_002_Rep_Nota	CSI_003_Rep_Falta

A Tabela 2 apresenta as sete características mais relevantes encontradas pelos classificadores ordenadas por ordem decrescente de *score* de entropia. É possível observar que as características CSI_040_Aprovado, CSI_004_Rep_Falta, e CSI_009_Aprovado (em vermelho) aparecem como mais relevantes nos três métodos. Para *Árvore de Decisão* e *Gradient Boosting* temos em comum (em azul) Ano_Letivo_Ini e CSI_002_Rep_Nota. Para *Árvore de Decisão* e *XGBoost* temos em comum (em verde) CSI_003_Rep_Falta; e para *Gradiente Boosting* e *XGBoost* se repetem (em roxo) CSI_037_Aprovado e CSI_077_Aprovado.

2.4 Discussão dos Resultados

Das três disciplinas que aparecem em comum como mais relevantes às três técnicas estudadas temos duas nos semestres iniciais (CSI_004 – Fundamentos de Sistemas de Informação, 1º semestre; CSI_009 – Programação II, 2º semestre) e uma no último semestre do curso (CSI_040 – Projeto de Diplomação II, 8º semestre). Acreditamos que as duas disciplinas do início do curso representam elementos motivadores de evasão, enquanto a disciplina do final do curso atua como sinalizador da conclusão do curso.

As características em comum nos modelos de *Árvore de Decisão* e de *Gradiente Boosting* são Ano_Letivo_Ini, que diz respeito ao ano letivo no qual o aluno iniciou os estudos, e CSI_002_Rep_Nota que indica quantas vezes o aluno reprovou por nota na disciplina de Programação I. A análise do modelo de Árvore de Decisão revelou que 100% dos alunos do grupo Concluídos ingressou antes de 2016, daí a importância da característica Ano_Letivo_Ini. Já a disciplina de Programação I é prerrequisito de todas as outras do eixo de programação. conjecturamos que os alunos que ficam retidos nesta disciplina tem maior propensão a evadirem.

Os modelos de *Árvore de Decisão* e *XGBoost* apresentaram em comum apenas a característica CSI_003_Rep_Falta que indica quantas vezes o aluno reprovou por falta na disciplina de Lógica. Os modelos *Gradient Boosting* e *XGBoost* apresentaram duas características em comum: CSI_037_Aprovado que indica aprovação na disciplina de Comércio Eletrônico, e CSI_077_Aprovado indica a aprovação na disciplina optativa de Língua Brasileira de Sinais. Em ambos os casos, a análise dos modelos demonstrou que estas características estão sendo utilizadas para identificar as estudantes que concluem o curso.

3 CONSIDERAÇÕES FINAIS

Este trabalho indica, em consonância com a literatura da área, que os maiores riscos de evasão afetam os alunos nos dois semestres iniciais do curso, correspondendo à quase 50% de evasão do curso. Portanto, a identificação precoce de alunos "em risco" é crucial para a efetividade das ações de permanência dos alunos Ameri et al. [1]. A associação de técnicas de mineração de dados educacionais com a análise de sobrevivência se mostrou útil em identificar os elementos de retenção dos alunos e seu impacto no itinerário acadêmico dos estudantes.

Considerando a severidade da evasão nos dois primeiros semestre do curso, atinge cerca de 50% da população estudantil, e considerando as disciplinas que foram consideradas como barreiras à formação, espera-se que de posse de tais informações a IES possa envidar esforços adicionais para revisar essas componentes curriculares de modo a favorecer o processo formativo.

Um dos trabalhos futuros é melhorar a questão do ensino aprendizagem nas disciplinas de Programação I e II, diminuindo o tamanho das turmas para 20 alunos (atualmente são 40), e com isso, possibilitar que os professores possam aumentar a atenção às dificuldades dos alunos de forma mais individualizada. Além de incluir um sistema de monitoria para reforçar a assistência para além do horário de aulas. Outros passos são na direção de entender o comportamento individual das variáveis exploratórias no modelo de sobrevivência para tentar identificar quais características dos alunos têm maior influência no processo de evasão. Uma vez que se identificou quais são as maiores barreiras do curso e quando surgem, acreditamos que o complemento lógico ao estudo será entender o porquê da suscetibilidade dos alunos a tais barreiras. Além de aplicar esta proposta para outros cursos da IES.

RECONHECIMENTO

Os autores agradecem ao apoio do IFES via edital Propós da PRPPG. A profª Komati agradece ao CNPq pela Bolsa de Produtividade DT-2 (308432/2020-7) e à FAPES pela concessão de Taxa de Pesquisa (nº FAPES 293/2021).

REFERÊNCIAS

[1] Sattar Ameri, Mahtab J Fard, Ratna B Chinnam, and Chandan K Reddy. 2016. Survival analysis based framework for early prediction of student dropouts. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, New YorkNYUnited States, 903–912.

[2] Lovenoor Aulck, Nishant Velagapudi, Joshua Blumenstock, and Jevin West. 2016. Predicting student dropout in higher education. *arXiv preprint arXiv:1606.06364* (2016).

[3] Luciane Bonaldo and Luis Pereira. 2016. Dropout: Demographic profile of Brazilian university students. *Procedia Social and Behavioral Sciences* 228 (2016), 138–143.

[4] Renato de Sousa Porto Gilioli. 2016. Evasão em instituições federais de ensino superior no Brasil: expansão da rede, Sisu e desafios. *Brasília: Câmara dos Deputados* (2016), 49.

[5] Ivan Londero Hoffmann, Raul Ceretta Nunes, and Felipe Martins Muller. 2019. As informações do Censo da Educação Superior na implementação da gestão do conhecimento organizacional sobre evasão. *Gestão & Produção* 26, 2 (2019).

[6] Roberto Leal Lobo Silva, Filho, Paulo Roberto Motejunas, Oscar Hipólito, and Maria Beatriz de Carvalho Melo Lobo. 2007. A evasão no ensino superior brasileiro. *Cadernos de pesquisa* 37, 132 (2007), 641–659.

[7] Vincent Tinto. 1975. Dropout from Higher Education: A Theoretical Synthesis of Recent Research. *Review of Educational Research* 45, 1 (1975), 89–125. https://doi.org/10.3102/00346543045001089 arXiv:https://doi.org/10.3102/00346543045001089

[8] Ping Wang, Yan Li, and Chandan K. Reddy. 2019. Machine Learning for Survival Analysis: A Survey. *ACM Computing Surveys (CSUR)* 51, 6, Article 110 (feb 2019), 36 pages. https://doi.org/10.1145/3214306