Análise de Sobrevivência: um estudo de caso em um Curso de Sistemas de Informação

Joubert Alexandrino de Souza[1], **Karin Satie Komati**[1], **Jefferson Oliveira Andrade**[1]

[1] Programa de Pós-graduação em Computação Aplicada (PPComp)
Instituto Federal do Espírito Santo (Ifes), Campus Serra,
Rodovia ES-010, Km 6,5, Manguinhos, Serra-ES

{joubert, kkomati, jefferson.andrade}@ifes.edu.br,

Abstract. *Dropping out of 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 technique combined with the construction of predictive models in the identification of the determinant elements of dropout in an undergraduate course 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 indicate the greatest risk of dropout is in the initial semesters of the course, while the identification of the determining characteristics of dropout makes it clear that the subjects of the first two semesters retain about 50% of the student population.*

Resumo. *A evasão do ensino superior é um problema grave que tem sido investigado há décadas e causa grandes danos aos indivíduos, instituições de ensino e à sociedade como um todo. Este artigo apresenta um estudo de caso sobre a aplicação da técnica de análise de sobrevivência combinada com a construção de modelos preditivos na identificação dos elementos determinantes da evasão em um curso de graduação em Sistemas de Informação de uma instituição pública de ensino superior no Brasil. Métodos de mineração de dados educacionais e modelagem probabilística foram aplicados aos dados dos alunos para modelar a conclusão esperada dos alunos do curso, semestre a semestre. Os resultados da análise de sobrevivência indicam que o maior risco de evasão está nos semestres iniciais do curso, enquanto a identificação das características determinantes da evasão deixa claro que as disciplinas dos dois primeiros semestres retêm cerca de 50% da população estudantil.*

1. Introdução

A evasão estudantil no ensino superior, doravante denominada evasão, é objeto de estudo de pesquisadores há várias décadas ao redor do mundo. O trabalho de Tinto [Tinto 1975] é uma das fontes conceituais mais importantes sobre o tema e define evasão a partir de duas perspectivas principais. A primeira delas é caracterizada pelo abandono, a qualquer tempo, da Instituição de Ensino Superior (IES) por um aluno regularmente matriculado em um de seus cursos. Já a segunda perspectiva diz respeito ao aluno que não abandona o curso, mas não obtêm o grau pretendido no tempo estipulado para a formação.

Estas perspectivas de evasão do ponto de vista de Silva et al. [Silva et al. 2007] podem ser mensuradas. Segundos os autores, a evasão total é "[...] o número de alunos que, tendo entrado num determinado curso, IES ou sistema de ensino, não obteve o diploma ao final de um certo número de anos". No contexto das IES, o problema da evasão diz respeito à interrupção dos estudos no processo formativo, que pode se manifestar em diferentes etapas, com diversas causas, e está associado a prejuízos para os indivíduos, para a IES e para um país. Ações de combate à evasão tem que ser parte integrante das agendas das IES e seus dirigentes, como forma de mitigar os impactos causados tanto para as instituições quanto para seus estudantes [Olaya et al. 2020].

Este trabalho utiliza uma abordagem holística sobre o estudo do fenômeno de evasão de um curso de graduação de Sistemas de Informações de uma instituição de ensino superior pública, o IFES (Instituto Federal do Espírito Santo). 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 com a análise de sobrevivência é de compreender melhor como a evasão, enquanto enxergada como um fenômeno, 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.
2. *Identificação de características mais relevantes na evasão dos alunos*. A conjectura 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.

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 [Wang et al. 2019]. O termo "sobrevivência" advém da área médica em que, por exemplo, a objetivo da análise é a taxa de sobrevivência de um tratamento de câncer após um determinado tempo. Neste trabalho, o evento é a evasão escolar. Das diferentes abordagens existentes, selecionou-se o método estatístico não paramétrico Kaplan-Meier [Ferreira and Patino 2016], que é usado para estimar a probabilidade de sobrevivência em vários intervalos de tempo e é um método indicado quando não há uma distribuição teórica conhecida.

Para a análise das características mais relevantes na evasão dos alunos, foram aplicadas técnicas de mineração de dados em bancos de dados educacionais para encontrar relacionamentos e padrões entre diferentes variáveis, área conhecida como Mineração de Dados Educacionais (MDE) [Kumar et al. 2017]. Uma parcela dos trabalhos em MDE tenta obter modelos preditivos que identifiquem a situação futura dos alunos a partir de registros passados [Franco et al. 2020]. O presente trabalho buscou entender as determinantes do abando escolar do curso em questão. Para compreender a relação entre os dados de entrada e o valores que se deseja prever, tirou-se o foco em determinar qual classificador trabalha melhor sobre o conjunto de dados para empregar os esforços em evidenciar as variáveis que representam os maiores desafios enfrentados pelos alunos durante a graduação. Foram usados três métodos: árvore de decisão, *gradient boosting* (GB) e XGBoost.

A base de dados usada neste trabalho contém informações reais de dados demográficos, dados de matrícula e a situação final de cada disciplina cursada pelo aluno (histórico escolar), seguindo o trabalho de Hellas et al. [Hellas et al. 2018], que é um trabalho sistemático de revisão de literatura que analisou 357 artigos publicados entre 2010 e 2018, os quais objetivavam predizer o desempenho dos alunos em cursos de computação. Uma de suas contribuições sugerem uma correlação entre dados de entrada e saída (que se deseja prever) e as características que frequentemente são usadas como dados de entrada para predizer os valores são os dados demográficos em geral, os dados de desempenho na escola secundária e os dados de desempenho durante o curso.

Este trabalho está organizado da seguinte maneira: na Seção 2 os trabalhos correlatos da literatura; a Seção 3 apresenta os materiais e métodos usados nos experimentos; a Seção 4 apresenta os resultados dos experimentos e por fim, o trabalho é finalizado na Seção 5, com as conclusões e trabalhos futuros.

2. Trabalhos Correlatos

Esta seção apresenta alguns trabalhos correlatos sobre evasão escolar em cursos de graduação. A seção é dividida em duas partes, uma que descreve artigos que usaram técnicas de análise de sobrevivência e outra que descreve artigos sobre identificação de características mais relevantes para o problema da evasão.

2.1. Análise de sobrevivência em cursos superiores

A proposta do artigo de Ameri et al. [Ameri et al. 2016] foi o de usar o modelo de riscos proporcionais de Cox e sua variação dependente do tempo para a previsão inicial de abandono escolar, que captura fatores que variam com o tempo e pode aproveitar essas informações para fornecer uma previsão mais precisa do abandono escolar. O método foi usado em uma base de dados reais coletados na Wayne State University de 2002 a 2009. A base de dados contém diferentes grupos de variáveis, como demográfica, histórico familiar, financeiro, informações sobre o ensino médio, matrícula na faculdade e créditos semestrais. Os resultados mostram que a proposta pode prever a evasão de alunos e semestre de evasão com alta exatidão e precisão em comparação com outros métodos de última geração.

O objetivo do trabalho de Costa et al. [Costa et al. 2018] foi analisar os condicionantes da evasão e da retenção de alunos do ensino superior em administração de uma universidade federal brasileira. A base de dados contém informações de 1.202 alunos ingressantes entre os anos de 2004 a 2009 que foram acompanhados até o ano de 2013. Há dados com relação ao tempo de permanência do aluno no curso, forma de saída (diplomação ou evasão), dados sócio-demográficos e sobre as características do curso. Para análise, foram utilizados as técnicas estatísticas de análise de sobrevivência, Cox e Kaplan-Meier. Como principais resultados, foi verificado que o número de semestres do curso, o desempenho do aluno, seu gênero, além da existência de reprovação e trancamento são fatores que explicam tanto o tempo de permanência quanto o risco de evasão. Na pesquisa foi constatado que variáveis relativas à idade no ingresso, estado civil, raça e natureza da escola de educação básica (pública ou privada) não demonstraram influência no tempo de conclusão ou evasão.

O trabalho de Saccaro et al. [Saccaro et al. 2019] analisa as bases de dados do Censo da Educação Superior dos anos de 2009 a 2014. Utiliza um método dividido em

duas partes, a primeira etapa é o Teste de Kaplan-Meier e a segunda etapa da análise consiste na utilização de uma técnica paramétrica, através da aplicação do modelo de Accelerated Failure Time (AFT), que estima as taxas de tempo – ou relações temporais. Como resultado, verificou-se que a evasão é maior nas instituições privadas. Além disso, ser homem e ter idade acima da média diminui o tempo de vida do indivíduo no ensino superior, enquanto que alunos contemplados com apoio financeiro apresentam uma maior retenção.

2.2. Análise de características mais relevantes na área de ensino

O trabalho de Oliveira Júnior et al. [de Oliveira Júnior et al. 2016] propõe um modelo de previsão da evasão escolar utilizando a criação e seleção de características oriundos de base de dados reais do sistema acadêmico da UTFPR, sendo selecionados os dados de alunos ingressantes pelo SISU dos cursos presenciais de graduação com oferta semestral. Os resultados experimentais apresentam que as características criadas (não existentes na base de dados inicial) foram relevantes para prever a evasão: (i) regressão do coeficiente de rendimento, que indica o coeficiente angular da equação de regressão linear do coeficiente de rendimento médio das disciplinas cursadas em cada semestre; (ii) dificuldade média das disciplinas cursadas pelo aluno; (iii) percentual de aprovação das disciplinas cursadas; (iv) total de semestres trancados; (v) empréstimos na biblioteca por semestre; além de um atributo da base inicial, o coeficiente de rendimento.

O trabalho de Bonaldo e Pereira [Bonaldo and Pereira 2016] identificou os fatores determinantes para a permanência ou afastamento de estudantes em IES privadas do Brasil. A população do estudo era composta por todos os estudantes dos cursos de graduação das universidades particulares do sudeste do Brasil, matriculados em qualquer área do conhecimento entre os semestres letivos de 2010 e 2015. A coleta de dados se deu por questionário eletrônico contendo 59 questões, denominado entrevista eletrônica, aplicado a 8.200 entrevistados, que gerou 1.294 respostas válidas. O método de regressão logística foi aplicado tendo como variável dependente a “evasão” e como variáveis independentes “sexo”, “evasão” “idade”, “nível de graduação da família”, “desempenho acadêmico do aluno”, “estado civil”, “filhos durante o curso” e “tipo de bolsa ou financiamento para suportar os custos dos estudos”. Os autores concluíram que idade, mudança de estado civil durante o curso, bolsa e solicitação de financiamento foram determinantes do abandono.

O trabalho de Carminati et al. [Carminati et al. 2020] se propõe a identificar o perfil dos alunos que evadem. A base de dados inicial é composta por 53 características, 37.212 instâncias de 10.960 alunos de todos os alunos dos cursos de direito e de engenharias no período de 2016-1 a 2018-2, totalizando seis semestres. As características envolvem aspectos demográficos, financeiros e acadêmicos. Foram usados as técnicas de árvore de decisão, Naïve Bayes, k-NN e rede neural Multilayer Perceptron. Os resultados preliminares identificaram alguns características que estão relacionados à evasão nos cursos de engenharia, tais como: semestre do ano (alunos têm maior probabilidade de evasão no primeiro semestre), assiduidade, notas (neste caso a mediana é mais importante que o valor médio) e número de créditos no semestre anterior ao qual está atualmente matriculado e alunos abaixo do 5º semestre têm maior tendência à evasão.

3. Materiais e Métodos

Nesta seção são apresentados os materiais e métodos usados nos experimentos. Inicia-se descrevendo a base de dados inicial proveniente de um sistema acadêmico em produção, isto é, são dados reais. Segue-se a descrição dos métodos referentes à análise de sobrevivência e os métodos para a identificação das características mais relevantes.

3.1. Base de dados coletada

O curso de Bacharelado em Sistemas de Informação (BSI) do IFES, 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] nas fichas 1.2 e 5.3, 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. A PNP teve início em 2017, e portanto não há dados anteriores de forma pública.

Tabela 1. Taxa de Evasão Anual em BSI

Ano Base	Ingressantes	Concluintes	Evasão Anual
2017	113 (2013)	28	27,8%
2018	102 (2014)	23	26,0%
2019	88 (2015)	1	9,0%
2020	90 (2016)	19	11,6%

Tabela 2. Conjunto de dados de alunos

(a) Conjunto de dados de alunos

Característica	Descrição
Matricula	Código da matrícula do aluno
Cor	Cor do aluno
Sexo	Sexo do aluno
Data_Nascimento	Data de Nascimento do aluno
Data_Conclusao_Medio	Conclusão do ensino médio
Situacao_Matricula	Situação da matrícula
Forma_Ingresso	Forma de ingresso
Periodo_Letivo_Ini	Período letivo inicial
Ano_Letivo_Ini	Ano letivo inicial
Ultimo_Periodo_Letivo	Período letivo final
Periodo_Let_Conclusao	Período letivo de conclusão
Ano_Let_Conclusao	Ano letivo de conclusão
Data_Conclusao_Curso	Data de conclusão do curso
Data_Colacao_Grau	Data da colação de grau
Periodo_Let	Período letivo inicial programado
Ano_Let	Ano letivo programado
Instituicao	Nome da instituição
Cod_Curso	Código do curso
Sigla_Curso	Sigla do curso
Descricao_Curso	Descrição do curso
Cod_Estrutura	Código da estrutura do curso
Descricao_Estrutura	Descrição da estrutura do curso
Cod_Matriz	Código da matriz do curso
Descricao_Matriz	Descrição da matriz do curso
Cod_Habilitacao	Códido da habilitação do curso
Desc_Habilitacao	Descrição da habilitação

(b) Conjunto de dados de disciplinas

Característica	Descrição
Cod_Instituicao	Código da instituição
Instituicao	Nome da instituição
Cod_Curso	Código do curso
Descricao_Curso	Descrição do curso
Cod_Turma	Código da turma
Sigla_Turma	Sigla da turma
Cod_Pauta	Código da pauta
Situacao_Pauta	Situação da pauta
Cod_Disciplina	Código da disciplina
Sigla_Disciplina	Sigla da disciplina
Descricao_Disciplina	Descrição da disciplina
Ano_Letivo	Ano de oferta da disciplina
Periodo_Letivo	Período de oferta da disciplina
Carga_Hor	Carga horária da disciplina
Hora_Aula	Horas aula da disciplina
Aulas_Previstas	Total de aulas previstas
Aulas_Previstas_Hoje	Aulas previstas até momento
Aulas_Dadas	Aulas ministradas
Total_Faltas_Hoje	Faltas do aluno
Frequencia_Momento	Frequência do aluno
Frequencia_Aulas_Prevista	Frequência relativa

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

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

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, apresentados por uma separação de linha horizontal na Tabela 2(a). Cada disciplina cursada pelo aluno representa um registro no conjunto de dados de disciplinas, Tabela 2(b), 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.

3.2. Análise de Sobrevivência

A análise de sobrevivência é essencialmente estudar a probabilidade de um determinado evento ocorrer em um dado período de tempo. Há cinco conceitos importantes nesta técnica:

- O **Evento** é o objeto da análise em termos de se ou quando vai ocorrer. Neste estudo o evento é a evasão do aluno.
- O **Tempo** é o período de tempo, contado à partir de uma determinada origem, até que o evento ocorra. A origem varia de acordo com o problema. Neste caso a origem no tempo é a matrícula do aluno no curso.
- A **Escala** é o tamanho do intervalo de tempo que será usado na análise, pode ser segundos, minutos, horas, dias, semanas, meses, anos, ou qualquer intervalo entre essas medidas. Para este estudo é o semestre.
- A **Censura**. Em alguns casos, o período do estudo se encerra e existem alguns indivíduos para os quais o evento nunca ocorreu como, por exemplo, o paciente não morreu durante o tempo de realização do estudo. Nestes casos costuma-se censurar os dados, ou seja, ao invés de"jogar o indivíduo fora", registra-se o indivíduo com o tempo máximo de duração do estudo [Chung et al. 1991]. Neste trabalho, a censura foi definida em 16 semestres, que é o prazo máximo de integralização do curso. Assim, todos os alunos que não evadiram até 16 semestres foram considerados como dados censurados, ou seja, foram registrados com tempo de evasão 16 semestres.
- A **Função de Sobrevivência** é a probabilidade do evento objeto do estudo não ter ocorrido no momento t. A função é definida como sendo $S(t) = Pr(T > t)$, onde T é o tempo de ocorrência do evento e t é o tempo máximo que se deseja observar [Wang et al. 2019]. Neste trabalho foi usado o estimador Kaplan-Meier.

Para realizar a análise de sobrevivência foi necessário produzir uma lista com os semestres em que aconteceram as evasões. Destarte, uma nova etapa de preparação de dados se sucederam originando o conjunto de dados **semestre evasão**. O campo *Id* corresponde apenas a um identificador de controle usado internamente no estudo. O campo *Matricula* é a referência alfanumérica que identifica o aluno e foi obtido a partir do conjunto de dados de alunos. O campo *Numsems_evade* é o número do semestre em que o aluno evadiu. Por exemplo, se o aluno cursou o primeiro semestre e não se matriculou em nenhuma disciplina no segundo semestre então `Numsems_evade = 1`; se ele esteve cursando disciplinas por quatro semestres desde que ingressou e no quinto semestre não se matriculou em nenhuma, então `Numsems_evade = 4`; se o aluno se graduar, então `Numsems_evade = 16`; se o aluno for jubilado por exceder o tempo máximo de integralização, então `Numsems_evade = 16`. Finalmente, 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. O estimador Kaplan-Meier é o método mais usado para estimar a curva de sobrevivência [Wang et al. 2019]. As probabilidades de ocorrência do evento são calculadas para um determinado ponto do tempo, e multiplica-se tais probabilidades por quaisquer probabilidades calculadas anteriormente para obter a estimativa final.

3.3. Identificação de Características mais Relevantes

O processo de identificação de características mais relevantes tem duas etapas, o pré-processamento da base de dados e os métodos de classificação. 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. A partir do conjunto de dados de alunos, manteve-se os atributos: *Matricula*, *Sexo*, *Cor*, *Data_Nacimento*, *Forma_Ingresso*, *Periodo_Letivo_Ini*, *Ano_Letivo_Ini* e *Situacao_Matricula*. As demais características do conjunto de dados de alunos foram descartadas. Quanto ao tratamento de dados ausentes, a característica *Cor* possuía muitos registros com dados ausentes. A eles foi imputado o valor mais frequente observado, a moda dos dados, por se tratar de dados categóricos.

De modo a favorecer o processo indutivo dos algoritmos aprendizado de máquina foi realizada a discretização dos atributos categóricos *Sexo*, *Cor* e *Forma_Ingresso*. 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.

Foi criada a característica Idade_Ingresso no conjunto de dados final **alunobsi**. Também foram criadas três novas características para cada valor único do atributo Sigla_Disciplina presente no conjunto de dados de disciplinas. Por exemplo, para a disciplina Introdução a Sistemas de Informação (código CSI_001), as três novas características resultantes foram: CSI_001_Aprovado, CSI_001_Rep_Falta, CSI_001_Rep_Nota. A característica com final _Aprovado indica se o aluno foi ou não aprovado na disciplina, a de final _Falta é o somatório das ocorrências de reprovação por falta pelo aluno na disciplina e a característica de final _Nota é a quantidade de vezes em que o aluno ficou reprovado por nota. Para disciplinas que o aluno não cursou foi atribuído o valor padrão 0. As demais características do conjunto de dados de disciplinas foram descartadas. Os 23 valores distintos do atributo alvo, Situacao_Matricula, foram mapeados numericamente para representar as situações alvo desejadas: **Concluídos (0)**, **Evasão (1)** e **Matriculados (2)**.

A análise exploratória dos dados revelou que o conjunto de dados estava desbalanceado do ponto de vista da variável alvo. A situação "Concluídos" possuía 102 registros, aproximadamente 8,72% dos registros. A situação "Evasão" possuía 677 registros, aproximadamente 57,92% dos registros. Por fim, a situação "Matriculados" possuía 390 registros totalizando cerca de 33,36% dos registros. Apesar deste desbalanceamento ter sido constatado, optou-se por não usar nenhuma estratégia de mitigação do desba-

lanceamento 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. Foram usados três métodos: árvore de decisão [Zhao et al. 2021], GB [Al Daoud 2019] e XGBoost [Chen et al. 2015]. Na árvore de decisão foi adotado o índice de ganho de informação, que é baseado no conceito de Entropia, onde quanto maior é a entropia maior é o ganho de informação que uma característica possui [Chung et al. 1991].

4. Experimentos e Resultados

Esta seção relata o estudo realizado e os resultados encontrados, divididos em duas partes, a análise de sobrevivência e a identificação das características mais relevantes na questão da evasão do aluno do BSI.

4.1. Análise de Sobrevivência

Para realizar as inferências sobre as ocorrências dos eventos de evasão foi usada a biblioteca Lifelines [Davidson-Pilon 2019] e o estimador *Kaplan–Meier* cuja tarefa é estimar a função de sobrevivência do conjunto de dados semestre evasão. A Figura 1 apresenta a função de sobrevivência e seus intervalos de confiança.

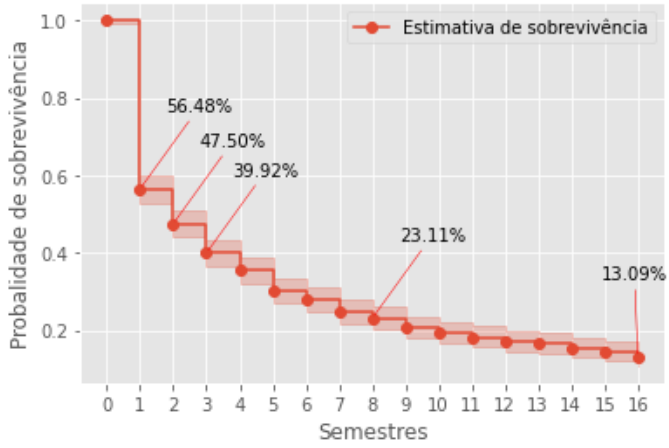

Figura 1. Função de sobrevivência com intervalo de confiança.

O gráfico apresenta as durações e as taxas de como ocorre o processo de evasão no curso de BSI. No eixo vertical tem-se a probabilidade dos alunos sobreviverem ao evento de evasão. No eixo horizontal tem-se o número de semestres de duração do curso. 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. Verifica-se que após 8 semestres de curso, tempo regular da formação, aproximadamente apenas 23% dos alunos conseguem sobreviver a eventos de evasão. Após o tempo regular do curso, os alunos são classificados como retidos. O tempo máximo de curso que é de 16 semestres, situação de jubilamento, 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.

4.2. Características mais relevantes

A identificação das características mais relevantes para a predição da evasão foi feita através do uso dos modelos de predição de árvore de decisão, GB e XGBoost. Este modelos, durante o processo de treinamento, intrinsecamente atribuem um *score* de relevância para cada característica do conjunto de dados. Este *score* indica quanto de informação aquela característica explica a predição da variável alvo.

Para implementação dos classificadores foi utilizada a biblioteca de aprendizado de máquina scikit-learn do Python [Pedregosa et al. 2011], com os hiper-parâmetros dos classificadores mantidos em seus valores padrão. Os classificadores foram treinados com 80% dos dados e testados com 20% dos dados restantes que foram divididos de modo estratificado, mantendo-se assim a proporcionalidade de distribuição das classes do atributo alvo, e obtiveram os seguintes valores de acurácia: 88,46% para a árvore de decisão, 90,59% para GB e 90,59% para XGBoost.

Tabela 3. 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 3 apresenta as sete características mais relevantes encontradas pelos classificadores ordenadas por ordem decrescente de entropia. É possível observar que as características CSI_040_Aprovado, CSI_004_Rep_Falta, e CSI_009_Aprovado (em negrito) aparecem na interseção dos conjuntos de característica mais relevantes de todos os três métodos. Além disso, para a árvore de decisão e GB também temos em comum as características (em itálico) Ano_Letivo_Ini e CSI_002_Rep_Nota; para os modelos de *árvore de decisão* e XGBoost em comum (em negrito e itálico) CSI_003_Rep_Falta; e para GB e XGBoost se repetem (em sublinhado) CSI_037_Aprovado e CSI_077_Aprovado.

Das três disciplinas que aparecem em comum como mais relevantes às três técnicas estudadas temos duas nos semestres iniciais e uma no último semestre do curso: CSI_004_Rep_Falta indica quantas vezes o aluno ficou reprovado por falta na disciplina de Fundamentos de Sistemas de Informação (1º semestres). Esta disciplina não é notória entre os alunos por sua dificuldade, deste modo, nossa conjectura é que este resultado reflete o fato de que os alunos que evadem estão deixando esta disciplina em segundo plano para priorizar disciplinas de maior dificuldade na matriz curricular, como Programação I, Cálculo I e Lógica. CSI_009_Aprovado representa a aprovação ou não na disciplina Programação II (2º semestre). Esta disciplina é obrigatória e é pré-requisito para todas as futuras disciplinas de programação e notória com uma disciplina difícil entre os alunos. Nossa conjectura é que caso o aluno consiga aprovação nesta disciplina, sua probabilidade de concluir o curso aumenta significativamente. O que está de acordo com o observado na análise de sobrevivência. CSI_040_Aprovado corresponde à obtenção de aprovação ou não na disciplina de Projeto de Diplomação II (8º semestres), que é a disciplina do 8º semestre curso relacionada ao Trabalho de Conclusão do curso (TCC).

As características que aparecem em comum nos modelos de árvore de decisão e de GB são, respectivamente, a de maior e menor escore de importância destes modelos: Ano_Letivo_Ini diz respeito ao ano letivo no qual o aluno iniciou os estudos; e CSI_002_Rep_Nota indica quantas vezes o aluno reprovou por nota na disciplina de Programação I (1º semestre). Esta disciplina é obrigatória e também é pré-requisito para todas as futuras disciplinas de programação. Nossa conjectura é que os alunos que experienciam repetidas retenções nesta disciplina e, consequentemente, não conseguem avançar no curso, eventualmente evadem.

Pela abordagem deste trabalho, foi possível constatar que os alunos mais vulneráveis são os calouros (primeiro ano de curso), que correm o maior risco de abandono escolar no início dos estudos, correspondendo a quase 50% de evasão. Portanto, a identificação precoce de alunos "em risco" é uma tarefa crucial que precisa ser tratada com eficácia. Como afirmado por Ameri et al. [Ameri et al. 2016], é importante não apenas classificar corretamente se um aluno vai abandonar o curso, mas também quando isso vai acontecer, pois a identificação do semestre é crucial para uma intervenção focada e evitar o evento. Também foi possível avaliar que as características mais relevantes encontradas materializaram-se como disciplinas, excetuando-se as características Ano_Letivo_Ini e Idade_Ingresso todas as demais são elementos da grade curricular. Estes resultados corroboram com o encontrado no trabalho de Carminati et al. [Carminati et al. 2020], em que alunos no 1º semestre de um curso de engenharia têm maior probabilidade de evasão. No artigo de revisão sistemática da literatura de Hellas et al. [Hellas et al. 2018], o trabalho cita que no 1º ano de curso, além das disciplinas, há questões como hábito de estudo, questões sociais e motivação. Já no final do curso, a transição para uma carreira profissional passa a ser um fator de evasão. Este trabalho não avaliou questões externas e se limitou aos dados existentes no sistema acadêmico da instituição.

5. Conclusões e Trabalhos Futuros

Este trabalho investigou o fenômeno da evasão no curso de Sistemas de Informação a partir de dados acadêmicos dos discentes. A associação de técnicas de mineração de dados educacionais para a identificação de características mais relevantes na evasão dos alunos com a técnica estatística de análise de sobrevivência se mostrou prática em identificar os elementos de retenção dos alunos e seu impacto no fluxo escolar. Para a análise de sobrevivência, o método estatístico não paramétrico Kaplan-Meier se mostrou adequado, dado que se partiu de uma distribuição teórica desconhecida. A decisão de se usar 3 métodos diferentes para a análise de quais disciplinas são mais relevantes na evasão e não apenas um método, foi interessante para avaliar as disciplinas em comum dos resultados.

Foi necessária a convalidação dos resultados obtidos junto a especialistas no domínio de conhecimento a fim de chancelar as descobertas, o que é extremamente importante no processo de descoberta do conhecimento. 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 (atual-

mente 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 com as mesmas características, de serem semestrais e que o aluno tem a escolha de disciplinas por semestre no momento da matrícula.

Ir além dos dados do aluno em seu curso, e poder identificar alunos em potencial risco de evasão na graduação a partir de seu desempenho no ensino médio é útil para o estabelecimento de medidas iniciais de enfrentamento por parte das IES. Para isto, fica evidenciado no trabalho de Nagy e Molontay [Nagy and Molontay 2018] a importância de se conhecer a história pregressa dos alunos, ou pelo menos poder coletar os dados do ENEM de cada aluno.

6. Agradecimentos

Os autores agradecem ao Ifes, apoio da FAPES e CAPES (proc 2021-2S6CD, nº FAPES 132/2021) do PDPG (Programa de Desenvolvimento da Pós-Graduação, Parcerias Estratégicas nos Estados). A profª Komati agradece ao CNPq pela Bolsa de Produtividade DT-2 (308432/2020-7) e à FAPES pelo Auxílio Taxa de Pesquisa (nº 293/2021).

Referências

Al Daoud, E. (2019). Comparison between xgboost, lightgbm and catboost using a home credit dataset. *International Journal of Computer and Information Engineering*, 13(1):6–10.

Ameri, S., Fard, M. J., Chinnam, R. B., and Reddy, C. K. (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*, pages 903–912, New YorkNYUnited States. ACM.

Bonaldo, L. and Pereira, L. (2016). Dropout: Demographic profile of brazilian university students. *Procedia Social and Behavioral Sciences*, 228:138–143.

Carminati, G., Augusto, R., Dallabrida, N., and Teive, R. (2020). Mineração de dados educacionais visando a identificação da evasão no ensino superior. In *Anais do Computer on the Beach (CoTB 2020)*, volume 11, pages 461–468.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., et al. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4.

Chung, C.-F., Schmidt, P., and Witte, A. D. (1991). Survival analysis: A survey. *Journal of Quantitative Criminology*, 7(1):59–98.

Costa, F. J. d., Bispo, M. d. S., and Pereira, R. d. C. d. F. (2018). Dropout and retention of undergraduate students in management: a study at a brazilian federal university. *RAUSP Management Journal*, 53:74–85.

Davidson-Pilon, C. (2019). lifelines: survival analysis in python. *Journal of Open Source Software*, 4(40):1317.

de Oliveira Júnior, J. G., Noronha, R. V., and Kaestner, C. A. A. (2016). Criação e seleção de atributos aplicados na previsão da evasão de curso em alunos de graduação. In *Anais do Computer on the Beach (CoTB 2016)*, pages 061–070.

Ferreira, J. C. and Patino, C. M. (2016). What is survival analysis, and when should i use it? *Jornal Brasileiro de Pneumologia*, 42(1):77–77.

Franco, J. J., de Almeida Miranda, F. L., Stiegler, D., Dantas, F. R., Brancher, J. D., and do Carmo Nogueira, T. (2020). Usando mineração de dados para identificar fatores mais importantes do enem dos últimos 22 anos. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 1112–1121. SBC.

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., and Liao, S. N. (2018). Predicting academic performance: a systematic literature review. In *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, pages 175–199.

Kumar, M., Singh, A., and Handa, D. (2017). Literature survey on educational dropout prediction. *International Journal of Education and Management Engineering*, 7(2):8.

Nagy, M. and Molontay, R. (2018). Predicting dropout in higher education based on secondary school performance. In *2018 IEEE 22nd international conference on intelligent engineering systems (INES)*, pages 000389–000394. IEEE.

Olaya, D., Vásquez, J., Maldonado, S., Miranda, J., and Verbeke, W. (2020). Uplift modeling for preventing student dropout in higher education. *Decision Support Systems*, page 113320.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Saccaro, A., França, M. T. A., and Jacinto, P. d. A. (2019). Fatores associados à evasão no ensino superior brasileiro: um estudo de análise de sobrevivência para os cursos das áreas de ciência, matemática e computação e de engenharia, produção e construção em instituições públicas e privadas. *Estudos Econômicos (São Paulo)*, 49:337–373.

Silva, Filho, R. L. L., Motejunas, P. R., Hipólito, O., and Lobo, M. B. d. C. M. (2007). A evasão no ensino superior brasileiro. *Cadernos de pesquisa*, 37(132):641–659.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45(1):89–125.

Wang, P., Li, Y., and Reddy, C. K. (2019). Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6):1–36.

Zhao, L., Lee, S., and Jeong, S.-P. (2021). Decision tree application to classification problems with boosting algorithm. *Electronics*, 10(16).