É Possível Prever Evasão com Base Apenas no Desempenho Acadêmico?

Carlos Henrique D. C. Santos[1], **Simone de Lima Martins**[1], **Alexandre Plastino**[1]

Abstract. *One of the major problems of high education in Brazil is the elevated dropout rate of students. In this work, we apply Data Mining techniques, more specifically, classification techniques, to predict and try to avoid dropouts. The predictive models are generated based only on the performance of the students in the subjects taken. Also, n different models are created, from which the i-th model, $1 \leq i \leq n$, is capable of predicting, at the end of a student's i-th semester, whether he or she will drop out or graduate in the future. The experiments conducted with a real database with data from students of a Brazilian university showed that the models are capable of achieving predictive accuracy between 79.31% and 98.25%.*

Resumo. *Um dos grande problemas do ensino superior no Brasil é o alto índice de evasão dos estudantes. Neste trabalho, aplicamos técnicas de Mineração de Dados, mais especificamente, técnicas de classificação, para prever e tentar evitar a evasão. Os modelos preditivos são gerados apenas com base no desempenho dos estudantes nas disciplinas cursadas. São criados n diferentes modelos, dos quais o i-ésimo modelo, $1 \leq i \leq n$, é capaz de prever, ao fim do i-ésimo semestre de um estudante, se ele ou ela irá evadir os se formar no futuro. Os experimentos realizados com uma base de dados real, sobre estudantes de uma universidade brasileira, mostraram que os modelos são capazes de atingir acurácia preditiva entre 79,31% e 98,25%.*

1. Introdução

A evasão no ensino superior representa um problema que gera altos gastos para as universidades anualmente [Hess 2018], além de tornar a contribuição das universidades menos efetiva no âmbito social [da Silva and Marques 2017]. Por esses fatos, torna-se imprescindível conseguir evitar os altos índices de evasão atuais. Uma forma é tentar prever a evasão dos alunos para evitar que ela aconteça. Por isso, sistemas computacionais podem ser muito importantes quando utilizados para identificar automaticamente alunos com tendências à evasão, permitindo ajudá-los [Alban and Mauricio 2019, Santos et al. 2018].

Nosso principal objetivo é utilizar técnicas de Mineração de Dados para tentar identificar alunos com comportamento semelhante a alunos que evadiram no passado. Utilizaremos técnicas de classificação, mais especificamente, indução de árvores de decisão para gerar modelos capazes de prever se um aluno irá evadir ou irá se formar. A partir desses modelos, idealizamos um Sistema de Previsão de Evasões para tornarmos nossa proposta mais efetiva.

Uma importante contribuição do trabalho – sua principal questão de pesquisa – é mostrar que podemos obter modelos com boas capacidades preditivas com base apenas no desempenho dos alunos nas disciplinas cursadas até o momento da previsão.

DOI: 10.5753/sbie.2021.218105

Outra contribuição deste trabalho é a proposta de geração de, não apenas um, mas de diferentes modelos preditivos, que deverão ser aplicados dependendo do período em que o aluno se encontra no curso. Dessa forma, aplica-se um modelo mais adequado de acordo com a etapa em que o aluno está no curso.

Especificamos e discutimos também todas as fases necessárias para a geração dos modelos preditivos. Desde o pré-processamento dos dados, passando pela criação das bases de treinamento e teste, até a efetiva avaliação dos modelos gerados, os quais irão compor o Sistema de Previsão de Evasões.

O artigo está dividido em outras cinco seções. A Seção 2 apresenta os trabalhos relacionados que utilizam Mineração de Dados para previsão de evasão. Na Seção 3, apresentamos o Sistema de Previsão de Evasões. Os métodos para criação e avaliação dos modelos preditivos são descritos na Seção 4. Na Seção 5, são apresentados os resultados obtidos e, como exemplo, uma das árvores de decisão geradas. Finalmente, na Seção 6, apresentamos as conclusões gerais e os próximos passos.

2. Trabalhos Relacionados

Neste trabalho, considera-se evasão quando o estudante desiste definitivamente do curso, em qualquer etapa, conforme utilizado em [da Silva et al. 2017]. Tanto no contexto internacional como nacional, o problema de evasão afeta as instituições públicas e privadas. Um estudo realizado nos Estados Unidos mostrou que, em 2016, somente 28% dos estudantes de faculdades iriam finalizar seus cursos [Hess 2018]. Isso significa que, por ano, quase 2 milhões de estudantes que iniciam a faculdade evadem nos Estados Unidos. No Brasil, temos uma evasão média de 40%, sendo que em alguns cursos pode chegar a 67% [da Silva and Marques 2017].

Nas últimas décadas, sistemas da área de Tecnologia da Informação vêm sendo aplicados cada vez mais na área de Educação, seja em cursos presenciais ou remotos. Diante da impossibilidade de se analisar manualmente a grande quantidade de dados gerados por esses sistemas, ferramentas têm sido propostas e utilizadas para analisar automaticamente esses dados, permitindo que estudantes, professores e administradores obtenham novas visões sobre o comportamento dos atores do sistema educacional.

A área de Mineração de Dados Educacionais (MDE) visa converter os dados brutos de sistemas educacionais em informações úteis para a prática e a pesquisa educacional [Moissa et al. 2015, Romero and Ventura 2020]. Em linhas gerais, o processo de aplicação de MDE é um ciclo composto das seguintes fases: (i) definição do tipo de dado que pode ser coletado de acordo com o ambiente educacional (presencial, remoto ou híbrido), (ii) coleta dos dados brutos, (iii) pré-processamento (seleção de atributos e anonimização de dados pessoais), (iv) especificação da técnica de Mineração de Dados a ser utilizada e (v) interpretação e aplicação do novo conhecimento [Moissa et al. 2015, Romero and Ventura 2020].

Alguns trabalhos apresentam estudos específicos sobre a utilização de MDE para predição de evasão do ensino superior. Santos et al. [Santos et al. 2018] analisam alguns destes trabalhos publicados entre os anos de 2013 e 2018. Os autores mostram que diferentes técnicas foram utilizadas, como Árvores de Decisão, Algoritmos de Regressão, Redes Neurais e outros, mas que os resultados obtidos ainda não atingiram um poder pre-

ditivo satisfatório. Nas abordagens estudadas, a previsão da evasão se concentrava principalmente em dados acadêmicos, sendo que algumas pesquisas indicam que é importante combinar abordagens baseadas em dados acadêmicos e não acadêmicos.

Em [Alban and Mauricio 2019], realizou-se um estudo sobre trabalhos publicados entre os anos de 2006 e 2018. Analisando esses trabalhos, os autores concluíram que os fatores que influenciam a evasão podem ser classificados em cinco dimensões: pessoal, acadêmico, econômico, social e institucional. Em relação às técnicas de Mineração de Dados utilizadas, aproximadamente 79% dos estudos avaliados usaram classificadores baseados em árvores de decisão, devido à sua flexibilidade no processamento de dados de natureza numérica e categórica, e à facilidade de interpretação dos resultados pelos tomadores de decisão.

Os trabalhos desenvolvidos em [Aulck et al. 2019, Sales et al. 2016] abordam predição de evasão para alunos de diversos cursos. Aulck et al. [Aulck et al. 2019] utilizaram dados acadêmicos apenas do primeiro ano de 66.060 alunos de todos os cursos de uma universidade. A formatura e a continuação dos estudos dos alunos no segundo ano pôde ser prevista usando dados dos registros da universidade. Os dados demográficos e de pré-entrada apresentaram menor poder preditivo que os dados acadêmicos dos estudantes.

O estudo realizado em [Sales et al. 2016] gerou dois modelos para predição de evasão de alunos de uma universidade: global, que utilizava todos os dados acadêmicos semestrais dos alunos, e específico, que utilizava os dados de um determinado curso. O modelo global obteve melhores valores de *precision* e *recall* do que o modelo específico, devido principalmente ao desbalanceamento das classes encontrado nas bases de dados usadas no modelo específico.

Rai e Jain [Rai and Jain 2013] utilizaram os classificadores ID3 e J48 para gerar modelos para prever a evasão de um curso de graduação em Ciência da Computação após o primeiro semestre. Os dados preliminares de 220 estudantes, escolhidos aleatoriamente, foram coletados através de questionários. As principais razões detectadas para a evasão de estudantes foram: fatores pessoais, fatores educacionais relacionados a problemas de aprendizagem e a cursos difíceis, e o ambiente do campus.

Em [de Brito et al. 2014], utilizaram-se técnicas de Mineração de Dados para predizer o desempenho dos alunos no primeiro período do curso de graduação em Ciência da Computação, através das suas notas de ingresso no vestibular. Segundo os autores, o desempenho nas disciplinas do primeiro período influencia diretamente na evasão do aluno. Os resultados mostraram que é possível inferir o desempenho dos estudantes com uma acurácia superior a 70%.

Carrano et al. [Carrano et al. 2019] utilizaram técnicas de Mineração de Dados para criar modelos preditivos para identificar alunos com risco de evasão, e para identificar atributos relevantes relacionados à evasão. Foi realizada uma avaliação global incluindo várias áreas do conhecimento e uma avaliação para cada área de conhecimento específica. O estudo mostrou que os atributos acadêmicos relacionados ao desempenho, assiduidade e satisfação foram os mais influentes na evasão.

Observamos nestes estudos que vários fatores sócio-econômicos e de desempenho são utilizados para prever evasão. Nosso objetivo neste trabalho é verificar se é possível prever evasão somente com dados de desempenho dos alunos durante curso.

3. Sistema de Previsão de Evasões

A ideia principal do nosso sistema é prever, ao fim de cada semestre cursado pelo aluno, se este evadirá ou não, a partir apenas do seu desempenho nas disciplinas cursadas até então. Utilizamos algoritmos de classificação para criar modelos capazes de prever se um aluno se formará ou se irá evadir em algum momento no futuro. Buscando, dessa forma, encontrar alunos em risco de evasão que possam se beneficiar de algum tipo de apoio.

O objetivo geral é utilizar os modelos para informar aos dirigentes, como por exemplo a coordenação do curso, sobre os alunos em risco de evasão, de forma que alguma ação possa ser tomada para tentar ajudá-los. É importante ressaltar que esse tipo de informação é extremamente sensível e deve ser tratada com cuidado, sendo imprescindível a realização de discussões sobre maneiras seguras de sua análise e divulgação.

O sistema será composto por n modelos preditivos, onde n é o número recomendado de períodos que o aluno deve cursar para se formar. O i-ésimo modelo, $1 \leq i \leq n$, é utilizado para prever se um aluno que entrou no curso há i semestres irá, no futuro, se formar ou evadir. O número de modelos preditivos pode ser definido pela coordenação do curso. No nosso caso de estudo, geramos 10 modelos considerando os alunos de graduação do curso de Ciência da Computação. Neste caso, por exemplo, o terceiro modelo preditivo será usado para prever a evasão de alunos que já cursaram três períodos, enquanto que o sexto modelo será usado para os alunos que cursaram seis períodos.

Para treinar e testar cada um dos n modelos preditivos, é necessário construir n bases de treinamento e teste (BTTs), uma para cada período. Todas essas n bases devem ser compostas por dados de alunos que já tenham se formado ou evadido no passado, pois é necessário saber a classe a que cada elemento (aluno) da base pertence: evadido ou formado. A i-ésima BTT será composta por alunos que tenham cursado pelo menos i semestres. Os atributos da i-ésima base serão definidos pelo resultado obtido por esses alunos nas disciplinas cursadas.

A Figura 1 representa os processos de geração, teste e utilização do Sistema de Previsão de Evasões. As BTTs são obtidas diretamente do banco de dados da universidade, contendo informações históricas sobre o desempenho dos alunos nas disciplinas cursadas. Como será visto mais adiante, cada BTT é dividida em duas partes: a base de treinamento, que será utilizada para gerar o modelo, e a de teste, que será utilizada para estimar o poder preditivo do modelo gerado. Nessa divisão, os alunos que fazem parte da base de teste devem ser de períodos posteriores aos dos alunos que fazem parte da base de treinamento. Dessa forma, teremos um processo de avaliação dos modelos preditivos mais próximo ao funcionamento real do sistema.

Além do poder preditivo estimado com a base de teste, será executada uma validação cruzada com 10 partições sobre a base de treinamento, que oferecerá outra estimativa do poder preditivo dos modelos. Após criados, e tendo a eficácia validada, os modelos serão utilizados, ao fim de cada semestre, para prever a evasão ou formatura de todos os alunos em curso. O i-ésimo modelo será utilizado para prever o comportamento futuro de um aluno que acabou de concluir o seu i-ésimo semestre na universidade. Para garantir a qualidade do poder preditivo dos modelos ao longo do tempo, uma vez que o comportamento dos alunos pode variar, é importante atualizar as bases de treinamento e teste periodicamente. Portanto, novas evasões e formaturas, assim como o desempenho em dis-

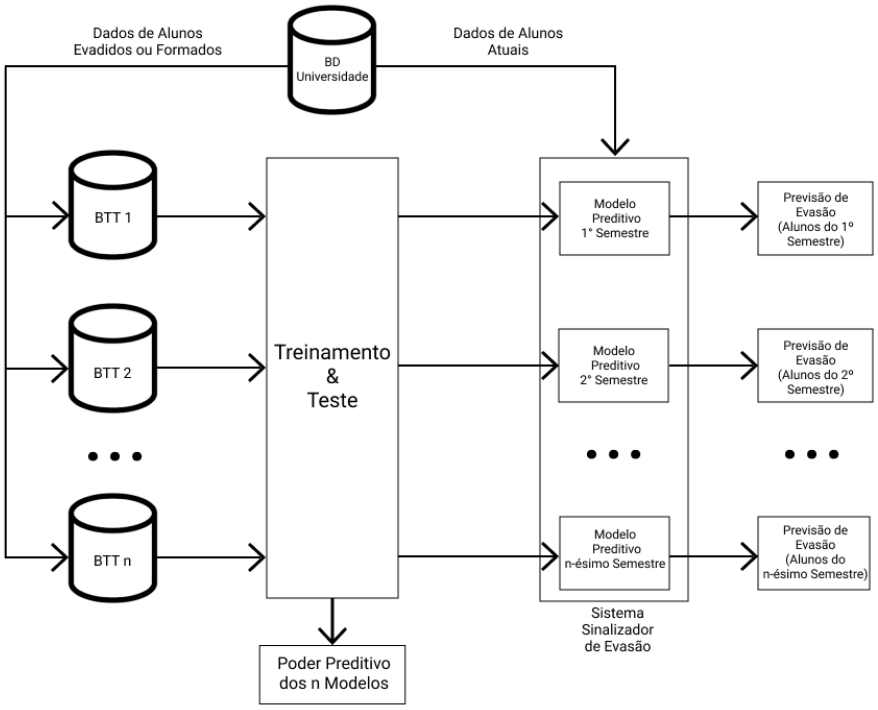

Figura 1. Treinamento, teste e utilização do Sistema de Previsão de Evasões

ciplinas recentemente cursadas, devem ser considerados de forma a manter os modelos sempre atualizados.

Nas próximas seções, serão explicados os processos de criação das bases de treinamento e teste, de geração dos modelos e de avaliação dos mesmos.

4. Criação e Teste dos Modelos

4.1. Criação das Bases de Treinamento e Teste

Inicialmente, são obtidos do banco de dados da universidade os seguintes dados sobre os alunos: curso do aluno, semestre de ingresso, semestre de desvinculação (tanto em caso de formatura quanto evasão), situação atual e, para cada disciplina cursada: código, semestre em que foi cursada, e situação final.

A partir da situação atual de cada aluno, é definida sua classe: "Evasão"ou "Formatura". Os alunos serão classificados na classe "Evasão" nas seguintes situações: "Cancelamentos por Abandono", "Cancelamentos por Solicitação Oficial", "Cancelamentos por Mudança de Curso", "Cancelamentos por Insuficiência de Aproveitamento" e "Cancelamentos por Limite de Permanência".

Os alunos serão classificados na classe "Formatura" com a situação "Formado". Não foram considerados nas BTTs os alunos nas seguintes situações: "Cursando", "Matriculas Trancadas", "Transferência", "Cancelamentos por Rematrícula" e "Motivos Especiais de Cancelamento".

Para a definição das BTTs, é importante também definir o intervalo de tempo do estudo, ou seja, o intervalo de tempo dentro do qual os alunos que farão parte do

estudo ingressaram na universidade. Esse intervalo é definido por um par de períodos [A,B]. Portanto, um aluno fará parte do estudo se tiver ingressado na universidade entre os períodos A e B, inclusive. No nosso estudo de caso, farão parte do estudo os alunos que tiverem ingressado na universidade entre 2008.1 e 2019.1, inclusive.

Como visto na seção anterior, são criados n modelos preditivos no sistema. O i-ésimo modelo é gerado para prever se um aluno que acabou de concluir o seu i-ésimo semestre na universidade irá evadir ou se formar no futuro. A i-ésima BTT será composta por alunos que tenham cursado pelo menos i semestres, ou seja, se matriculado na universidade há i semestres. Portanto, as diferentes BTTs terão quantidades distintas de alunos.

Em relação aos atributos, a BTT do primeiro semestre deverá conter, para cada disciplina cursada pelos alunos desta BTT no primeiro semestre, um atributo referente à situação final dos alunos na respectiva disciplina. A BTT do segundo semestre deverá conter os atributos referentes às disciplinas da grade do curso que foram cursadas pelos alunos desta BTT até o segundo semestre, e assim por diante.

É importante observar que, ao gerar, por exemplo, a BTT do quarto semestre, não serão registrados os resultados das disciplinas cursadas pelos alunos dessa BTT após o quarto semestre. Além disso, só farão parte da BTT do quarto semestre os alunos que cursaram pelo menos quatro semestres, ou seja, aqueles que não evadiram antes.

Portanto, a i-ésima BTT será uma tabela onde os elementos (linhas) representam seu alunos e os atributos (colunas) representam as disciplinas cursadas por pelo menos um dos seus alunos, dentro do seu i-ésimo semestre.

Para cada aluno, o valor de cada atributo, ou seja, o valor associado a cada disciplina da BTT, poderá ser: (i) "Não Cursada": indica que o aluno não cursou a disciplina até o semestre da BTT em questão, (ii) "Reprovado": indica que o aluno já cursou a disciplina e foi reprovado (por nota ou por frequência), e (iii) "Aprovado": Indica que o aluno cursou a disciplina e foi aprovado (podendo ter sido reprovado antes ou não).

4.2. Filtro de Disciplinas Irrelevantes

Um dos problemas encontrados na criação das BTTs foi o fato de existirem alunos que cursaram disciplinas antes do período que era esperado que elas fossem cursadas. Por exemplo, nas BTTs do 1º e 2º semestres, eram criados atributos referentes a diversas disciplinas avançadas no curso, que foram antecipadas por alguns poucos alunos. Muitas vezes, eram alunos que vieram de outras instituições e que aproveitavam várias disciplinas assim que entravam no curso. Esses dados poderiam enviesar os resultados uma vez que esses alunos teriam sido considerados aprovados em um grande número de disciplinas logo no primeiro período e avançariam mais rapidamente no curso, representando um comportamento atípico.

Para resolver esse problema, resolvemos contabilizar o número de estudantes que cursaram cada uma das disciplinas, analisando cada BTT individualmente. Após essa contabilização, removemos das BTT as disciplinas que não tinham sido cursadas por pelo menos 3% dos alunos da base. Dessa forma, evitamos que disciplinas cursadas por poucos alunos sejam utilizadas para treinamento. Removemos também os alunos que cursaram essas disciplinas, tentando evitar que o comportamento fora do comum desses alunos in-

fluencie nos resultados. O ponto de corte utilizado (3%) foi obtido de forma experimental. O valor foi iniciado em 1% e incrementado até que as BTTs não possuíssem disciplinas de semestres avançados.

Uma disciplina só não é removida ao ser cursada por menos de 3% dos alunos se a sua posição na grade curricular for menor ou igual ao semestre correspondente da BTT em questão. Por exemplo, caso uma disciplina do 4º período na grade tenha sido cursada por menos de 3% dos alunos da BTT do 5º semestre, ela não será removida.

4.3. Divisão das BTTs em Treinamento e Teste

Após a criação das BTTs, elas são separadas em duas partes: base de treinamento e base de teste. A base de treinamento será utilizada para gerar o modelo, que será avaliado com a base de teste.

No nosso estudo de caso, criamos 10 BTTs, cada uma delas contendo alunos que ingressaram dentro da janela temporal considerada entre 2008.1 e 2019.1, inclusive. Para cada BTT, a base de treinamento foi formada por alunos que ingressaram entre 2008.1 e 2012.2, inclusive. Já a base de teste foi formada pelos alunos que ingressaram entre 2013.1 e 2019.1.

As bases de teste foram geradas com alunos entre 2013.1 e 2019.1 e, portanto, houve uma predominância por alunos que evadiram. Pois, se eles estão na base, é porque provavelmente evadiram, pois alunos que entraram mais recentemente ainda não tiveram tempo para se formar. Então foi preciso realizar um ajuste da distribuição de classes.

O ajuste na distribuição de classes nas bases de teste foi realizado da seguinte forma. Como a quantidade de alunos que evadiram era consideravelmente maior que a de formados, foi necessário reduzir a quantidade de evasões. Essa redução foi feita de maneira que a proporção de evasões e formaturas da respectiva base de treinamento fosse mantida. Por exemplo: na base de treinamento do 1º semestre temos 61.41% de evasões e 38.59% de formaturas. Dessa forma, na base de teste do 1º semestre o número de evasões foi reduzido de modo a manter aproximadamente essa mesma distribuição de valores.

A Tabela 1 apresenta como ficou a transformação de cada BTT em base de treinamento e base de teste. A primeira coluna representa o semestre. As seguintes três colunas representam a quantidade de alunos de cada base de treinamento, a quantidade de alunos que evadiram e a quantidade de alunos que se formaram, respectivamente. As próximas três colunas representam as mesmas informações para cada base de teste. Observa-se que, ao longo dos semestres, o número de alunos se reduz. Isso é natural pois, devido às evasões, é maior o número de alunos que cursaram um semestre do que o número dos que cursaram 10 semestres. Nota-se também que, em cada semestre, a distribuição de alunos entre evasões e formaturas da base de teste é aproximadamente a mesma da base de treinamento. Isso se deve ao ajuste realizado.

5. Resultados

Após a geração das bases de treinamento e teste, foram gerados os 10 modelos. Foi utilizado o algoritmo CART de indução de árvores de decisão [J. Han and Pei 2012], implementado na linguagem R. Escolhemos trabalhar com árvores de decisão por se tratar

Tabela 1. Evasões e Formaturas nas Bases de Treinamento e Teste

	Bases de Treinamento			Bases de Teste		
Semestre	Qtd.	Nº Evasões	Nº Formados	Qtd.	Nº Evasões	Nº Formados
1º	482	296 (61.41%)	186 (38.59%)	85	52 (61.18%)	33 (38.82%)
2º	466	278 (59.66%)	188 (40.34%)	87	52 (59.77%)	35 (40.23%)
3º	462	276 (59.74%)	186 (40.26%)	87	52 (59.77%)	35 (40.23%)
4º	454	270 (59.47%)	184 (40.53%)	84	50 (59.52%)	34 (40.48%)
5º	440	258 (58.64%)	182 (41.36%)	77	45 (58.44%)	32 (41.56%)
6º	426	234 (54.93%)	192 (45.07%)	80	44 (55.00%)	36 (45.00%)
7º	379	182 (48.02%)	197 (51.98%)	68	33 (48.53%)	35 (51.47%)
8º	342	149 (43.57%)	193 (56.43%)	64	28 (43.75%)	36 (56.25%)
9º	295	118 (40.00%)	177 (60.00%)	57	23 (40.35%)	34 (59.65%)
10º	221	77 (34.84%)	144 (65.16%)	39	14 (35.90%)	25 (64.10%)

de um modelo que, em geral, apresenta um bom poder preditivo e cujas predições podem ser explicadas, característica importante no contexto educacional.

A primeira avaliação do algoritmo CART nas 10 bases de treinamento foi realizada por validação cruzada com 10 partições. A Tabela 2 apresenta os resultados obtidos. A primeira coluna indica a base utilizada. A segunda indica a acurácia obtida. As três seguintes indicam *precision*, *recall* e *F-measure* referentes à classe Evasão. As outras três representam as medidas referentes à classe Formatura. Observa-se um desempenho satisfatório de acurácia, que varia entre 75,93% (1º semestre) e 90,06% (8º semestre), estando sempre bem acima do percentual da classe majoritária de cada semestre.

Tabela 2. Resultados da Validação Cruzada nas Bases de Treinamento

		Evasão			Formatura		
Base	Acurácia	Precision	Recall	F-Measure	Precision	Recall	F-Measure
1º sem.	75.93%	81.69%	78.38%	80.00%	67.68%	72.04%	69.79%
2º sem.	78.54%	81.56%	82.73%	82.14%	73.91%	72.34%	73.12%
3º sem.	80.30%	83.15%	84.06%	83.60%	75.96%	74.73%	75.35%
4º sem.	86.56%	88.00%	89.63%	88.81%	84.36%	82.07%	83.20%
5º sem.	85.46%	86.74%	88.76%	87.74%	83.52%	80.77%	82.12%
6º sem.	85.45%	87.39%	85.90%	86.64%	83.16%	84.90%	84.02%
7º sem.	87.07%	85.95%	87.36%	86.65%	88.14%	86.80%	87.47%
8º sem.	90.06%	90.78%	85.91%	88.28%	89.55%	93.26%	91.37%
9º sem.	87.46%	84.03%	84.75%	84.39%	89.77%	89.27%	89.52%
10º sem.	86.88%	83.33%	77.92%	80.54%	88.59%	91.67%	90.10%

A segunda avaliação visa medir a capacidade preditiva de cada modelo, gerado a partir de toda a base de treinamento, realizando o teste com a respectiva base de teste. Essa avaliação simula melhor a aplicação dos modelos em dados reais pois respeita a questão temporal: alunos da base de teste ingressaram na universidade após os alunos da base de treinamento. A Tabela 3 apresenta os resultados obtidos e possui estrutura semelhante à da Tabela 2. Novamente, observa-se um desempenho satisfatório de acurácia, entre 79,31% (3º semestre) e 98,25% (9º semestre). Alguns modelos obtiveram *precision* de 100% para a classe Evasão, o que indica que todas as previsões de evasão estavam corretas. Considerando o modelo do 3º semestre, 80,35% das previsões de evasão estavam corretas.

Para a classe Formatura, a medida *precision* variou de 75,68% a 97,22%. Em relação à medida *recall*, alguns modelos obtiveram o valor de 100% para a classe Formatura, o que indica que todas as formaturas foram corretamente preditas. Porém, considerando o modelo do 3º semestre, apenas 68,57% das formaturas foram corretamente preditas. Para a classe Evasão, a medida *recall* variou de 82,22% a 96,43%.

Tabela 3. Resultados Obtidos com a Base de Teste

		Evasão			Formatura		
Modelo	Acurácia	Precision	Recall	F-Measure.	Precision	Recall	F-Measure.
1º sem.	83.53%	89.58%	82.69%	86.00%	75.68%	84.85%	80.00%
2º sem.	83.91%	83.93%	90.39%	87.04%	83.87%	74.29%	78.79%
3º sem.	79.31%	80.36%	86.54%	83.33%	77.42%	68.57%	72.73%
4º sem.	90.48%	100.00%	84.00%	91.30%	80.95%	100.00%	89.47%
5º sem.	89.61%	100.00%	82.22%	90.24%	80.00%	100.00%	88.89%
6º sem.	93.75%	97.56%	90.91%	94.12%	89.74%	97.22%	93.33%
7º sem.	97.06%	100.00%	93.94%	96.88%	94.59%	100.00%	97.22%
8º sem.	96.88%	96.43%	96.43%	96.43%	97.22%	97.22%	97.22%
9º sem.	98.25%	100.00%	95.65%	97.78%	97.14%	100.00%	98.55%
10º sem.	97.44%	100.00%	92.86%	96.30%	96.15%	100.00%	98.04%

Como podemos perceber, o poder preditivo dos modelos aumenta com o passar dos semestres. Esse efeito ocorre pelo fato de os modelos dos primeiros semestres terem menos atributos (disciplinas) do que os modelos dos períodos mais avançados.

A Figura 2 mostra a árvore de decisão gerada a partir da base de treinamento do 4º semestre. Observam-se facilmente as regras obtidas e que serão utilizadas para prever as evasões e formaturas. Por exemplo, o modelo estima que um aluno que acabou de cursar seu 4º semestre irá evadir se ainda não tiver cursado ou tiver sido reprovado nas disciplinas Programação Estruturada, Cálculo 2A e Probabilidade & Estatística. As árvores geradas ajudarão a explicar o funcionamento dos modelos para os coordenadores de cursos.

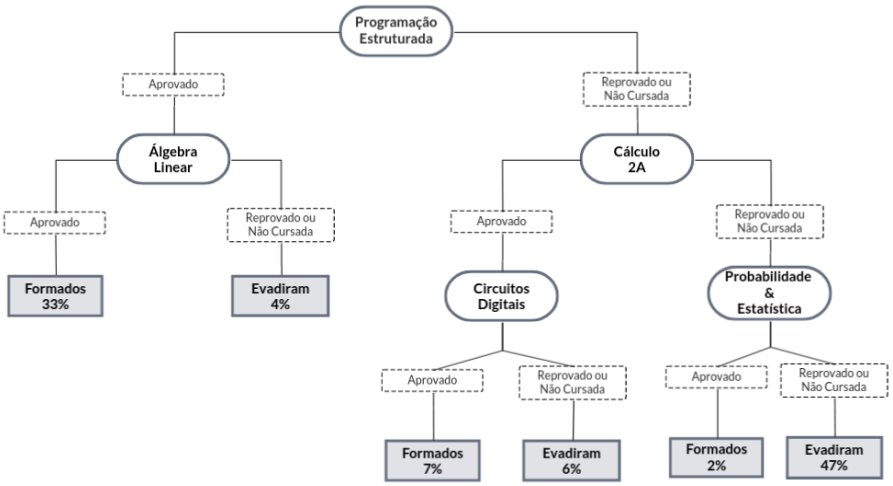

Figura 2. Árvore de decisão gerada para a base do 4º semestre

Após a avaliação dos modelos gerados pelo CART, utilizamos o algoritmo Random Forest [Breiman 2001], também implementado na linguagem R, para gerar outros 10 modelos e viabilizar uma comparação de desempenho dos dois algoritmos.

A Tabela 4 mostra que, por meio da validação cruzada realizada nas bases de treinamento, em termos de acurácia, o algoritmo Random Forest teve um melhor desempenho em 9 dos 10 semestres, com pior desempenho no quarto semestre. Para as bases de teste, a Tabela 5 mostra que Árvore de Decisão teve melhor desempenho que Random Forest em 6 dos 10 semestres, a estratégia Random Forest teve um desempenho melhor em 2 dos 10 semestres, e ambas empataram em 2 semestres.

Tabela 4. Comparação entre Árvore de Decisão (AD) e Random Forest (RF) por validação cruzada

Sem.	AD	RF
1º	75.93%	**79.67%**
2º	78.54%	**81.55%**
3º	80.30%	**82.47%**
4º	**86.56%**	86.12%
5º	85.58%	**87.44%**
6º	85.45%	**88.73%**
7º	87.07%	**89.97%**
8º	90.06%	**90.94%**
9º	87.46%	**90.85%**
10º	86.88%	**91.86%**

Tabela 5. Comparação entre Árvore de Decisão (AD) e Random Forest (RF) para as bases de teste

Sem.	AD	RF
1º	**83.53%**	81.18%
2º	**83.91%**	82.76%
3º	79.31%	**89.66%**
4º	**90.48%**	**90.48%**
5º	**89.61%**	88.31%
6º	93.75%	**96.25%**
7º	**97.06%**	**97.06%**
8º	**96.88%**	95.31%
9º	**98.25%**	92.98%
10º	**97.44%**	94.87%

Podemos observar que os modelos tiveram desempenhos opostos considerando a validação cruzada e o teste com as bases de teste, e para ambos seus valores de acurácia foram altos. Como Árvore de Decisão é um modelo mais simples de ser compreendido e interpretado, e teve um melhor desempenho que Random Forest no cenário em que o teste é feito com alunos que ingressaram depois dos alunos que foram utilizados no treinamento, ele foi o escolhido para compor o Sistema de Controle de Evasão.

6. Conclusões

Mostramos, neste trabalho, que é possível obter modelos preditivos capazes de estimar, com eficácia, a evasão no ensino superior a partir apenas de dados relativos ao desempenho dos alunos nas disciplinas cursadas. Foram obtidas, na avaliação considerada mais fiel à utilização real dos modelos, acurácias entre 79,31% e 98,25% . Observou-se que a indução de árvores de decisão levou a modelos eficientes e de fácil compreensão.

A metodologia para geração destes modelos preditivos pode ser realizada para qualquer outro curso de qualquer outra universidade, ou mesmo em escolas de ensino básico ou médio, pois só depende do desempenho dos alunos nas disciplinas em cada período do curso.

Os próximos passos deste trabalho serão: (i) disponibilização dos modelos gerados às instâncias universitárias competentes, (ii) levar em consideração a questão da privacidade dos dados, e (iii) avaliar métodos de classificação alternativos visando o aumento do poder preditivo dos modelos gerados.

Referências

[Alban and Mauricio 2019] Alban, M. and Mauricio, D. (2019). Predicting university dropout through data mining: A systematic literature. *Indian Journal of Science and Technology*, 12(4):1–12.

[Aulck et al. 2019] Aulck, L. S., Nambi, D., Velagapudi, N., Blumenstock, J. E., and West, J. D. (2019). Mining university registrar records to predict first-year undergraduate attrition. In *Proceedings of the 12th International Conference on Educational Data Mining*.

[Breiman 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[Carrano et al. 2019] Carrano, D., Tuler, E., Infante, C., and Rocha, L. (2019). Combinando técnicas de mineração de dados para melhorar a detecção de indicadores de evasão universitária. In *Anais do XXX Simpósio Brasileiro de Informática na Educação*.

[da Silva et al. 2017] da Silva, D. R., de Lima Martins, S., and Maciel, C. (2017). Identification and systematization of indicatives and data mining techniques for detecting evasion in distance education. In *Proceedings of Twelfth Latin-American Conference on Learning Technologies (LACLO)*.

[da Silva and Marques 2017] da Silva, H. F. D. and Marques, W. (2017). Evasão na educação superior no brasil: Desafio à gestão acadêmica. *Quaestio*, 19(1):197–208.

[de Brito et al. 2014] de Brito, D. M., de Almeida Júnior, I. A., Queiroga, E. V., and do Rêgo, T. G. (2014). Predição de desempenho de alunos do primeiro período baseado nas notas de ingresso utilizando métodos de aprendizagem de máquina. In *Anais do XXV Simpósio Brasileiro de Informática na Educação*.

[Hess 2018] Hess, F. (2018). The college dropout problem. *Forbes*.

[J. Han and Pei 2012] J. Han, M. K. and Pei, J. (2012). *Data Mining: Concepts and Techniques*. Morgan Kaufmann, third edition.

[Moissa et al. 2015] Moissa, B., Gasparini, I., and Kemczinski, A. (2015). Educational data mining versus learning analytics: estamos reinventando a roda? um mapeamento sistemático. In *Anais do XXVI Simpósio Brasileiro de Informática na Educação*.

[Rai and Jain 2013] Rai, S. and Jain, A. K. (2013). Students' dropout risk assessment in undergraduate courses of ict at residential university - a case study. *International Journal of Computer Applications*, 84(14):31–36.

[Romero and Ventura 2020] Romero, C. and Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, 10(3):e1355.

[Sales et al. 2016] Sales, A., Balby, L., and Cajueiro, A. (2016). Exploiting academic records for predicting student drop out: A case study in brazilian higher education. *Journal of Information and Data Management*, 7(2):166–166.

[Santos et al. 2018] Santos, G. A. S., Bordignon, A. L., Oliveira, S. L. G., Haddad, D. B., Brandão, D. N., and Belloze, K. T. (2018). A brief review about educational data mining applied to predict student's dropout. In *Anais da V Escola Regional de Sistemas de Informação do Rio de Janeiro*, pages 86–91. SBC.