Detecção precoce de estudantes em risco de evasão usando dados administrativos e aprendizagem de máquina

José Ahirton Batista Lopes Filho [1], Ismar Frango Silveira[1]

ahirtonlopes@gmail.com, ismar.silveira@mackenzie.br

[1] Universidade Presbiteriana Mackenzie, Programa de Pós Graduação em Engenharia Elétrica e Computação, Rua da Consolação, 930, Consolação, São Paulo, Brasil.

DOI: 10.17013/risti.n.pi-pf

Resumo: Altas taxas de evasão são um problema sério e bastante comum em vários países do globo, seja em ambientes de ensino-aprendizagem tradicionais quanto em ambientes de *e-learning*, tanto na educação privada quanto pública. A alta destas taxas costuma ter um impacto negativo em todos os perfis envolvidos: estudantes, instituições e o público em geral; haja visto que, não obstante a perda de ganho educacional do próprio aluno que se evade, há também a perda monetária para o sistema em questão, o estigma social e os sentimentos de não adequação que podem estar associados a tal evasão, além da perda nos âmbitos culturais, sociais e interpessoais advindos dos processos educacionais. Portanto, sistemas de detecção precoce quanto a evasão escolar têm ganhado mais destaque, principalmente quanto a possibilidade de serem um arcabouço para a elaboração de novas políticas públicas, ao também auxiliar no melhor entendimento quanto as prováveis causas para essa evasão.

Palavras-chave: aprendizagem de máquina; detecção de risco de evasão; aprendizagem supervisionada; classificação; mineração de dados educacionais.

Early detection of students at dropout risk using administrative data and machine learning

Abstract: High dropout rates are a serious and widespread problem in many countries around the globe, both in traditional and e-learning environments, in both private and public education. Higher rates tend to have a negative impact on all profiles involved: students, institutions and the general public; noting that, despite the student's own loss of educational gain, there is also

monetary loss to the system in question, social stigma and feelings of inadequacy that may be associated with such dropout, and loss of cultural, social and interpersonal spheres arising from educational processes. Therefore, early detection systems regarding school dropout have gained more prominence, especially regarding the possibility of being a framework for the elaboration of new public policies, while also helping to better understand the probable causes for this dropout.

***Keywords**:* machine learning; dropout prediction; supervised learning; classification; educational data mining.

1. Introdução

A evasão, principalmente nos ensinos médio (Sansone, 2017) e superior (Berens, Oster, Schneider & Burghoff, 2018), é uma questão fundamental nos mais variados sistemas educacionais; segundo a Organização para Cooperação e Desenvolvimento Econômico, OECD (OECD, 2017), a taxa de conclusão do ensino médio em países como Índia (30,4%), Bélgica (31,4%), México (56,8%), Argentina (65,2%), Brasil (67,3%), Portugal (84,6%), Espanha (80,8%) e até mesmo Estados Unidos (85,0%) estão abaixo da média entre os países membro da OECD (86.2%) e bem longe de países como Chile (92.1%), Itália (95.6%) e Japão (97.6%).

Ainda, em muitos desses países, existem diferenças substanciais quanto a gênero, raça, localização geográfica bem como diferentes características socioeconômicas que podem ser chave para o processo de evasão de alunos (U.S. Department of Education, 2016). Fora isso, é importante salientar que a verificação de taxas de graduação por si só pode mascarar outras questões e serem uma medida ruim de como os jovens estão preparados para o trabalho e o ensino superior, visto que fatores tais como escalas de notas mais brandas e programas de recuperação de crédito ou ensino especial podem influir em tais taxas.

As instituições e o próprio governo costumam ter um imenso interesse em melhorias quanto a evasão escolar; além disso, não conseguir se formar no ensino médio pode ter altos custos, já que apenas 12% de todos os empregos na economia exigirão menos do que um diploma do ensino médio até 2020 (Carnevale, Smith & Strohl, 2013). Mesmo a definição de evasão diferindo entre pesquisadores, governos e instituições, em qualquer caso, se uma instituição perde um estudante por qualquer motivo, a instituição tem uma taxa de retenção menor (Marquez-Vera, Romero & Ventura, 2011).

Assim, a identificação precoce de alunos vulneráveis, propensos a desistir de seus cursos, é tida como crucial para o sucesso de qualquer estratégia de retenção escolar fora que, nos últimos tempos, tem havido cada vez mais estudos quanto à correlação de educação para com benefícios não pecuniários que vão desde a saúde

à felicidade, confiança e prazer no trabalho (Oreopoulos, 2007; Oreopoulos e Salvanes, 2011).

Prever a evasão de estudantes é então uma questão importante na educação porque diz respeito a muitos alunos em escolas e instituições individuais em todo o mundo, e geralmente resulta em perda financeira geral, taxas de graduação mais baixas e reputação escolar inferior aos olhos de todos os envolvidos (Neild, Balfanz e Herzog, 2007). Ainda, as ações a serem tomadas para a redução das taxas de evasão escolar têm que satisfazer os seguintes critérios: elas precisam ser eficientes em termos de custo e direcionadas aos alunos necessitados (Berens, Oster, Schneider & Burghoff, 2018).

Ou seja, primeiro, os estudantes em risco precisam ser identificados, idealmente usando informações disponíveis (dados administrativos). Em segundo lugar, os alunos e o apoio dos alunos devem ser correspondidos e as intervenções precisam ser avaliadas. Em terceiro lugar, o sistema deve ser dinâmico e auto-ajustável (Berens, Oster, Schneider & Burghoff, 2018).

Assim sendo, para se tentar reduzir o problema supracitado, é necessário detectar os alunos que estão em risco o mais cedo possível e, assim, fornecer cuidados para evitar com que esses alunos abandonem seus estudos e também deve haver intervenção de maneira rápida, de modo a facilitar a retenção do aluno (Heppen & Bowles, 2008; Marquez-Vera, Romero & Ventura, 2011). Para tanto, surge então a figura dos sistemas de detecção precoce de evasão (do inglês, *early detection systems* – EDS ou *early warning systems* - EWS), para a correta predição do risco de evasão de alunos, tendo em vista intervenção cada vez mais individualizada.

Este artigo pretende então contribuir para os três pontos apontados por Berens et al. (2018), ao passo que é apresentado um EDS que, baseado em um arcabouço generalista o qual pode se usar de diferentes atributos construídos a partir de dados administrativos, tem como objetivo identificar possíveis alunos em situação de risco de evasão, tendo como base informações disponíveis referentes ao ensino público estadual no estado de São Paulo, Brasil (dados.educacao.sp.gov.br).

Assim sendo, por gerar um percentual individualizado para cada estudante com o que pode ser considerado seu grau de risco quanto a evasão escolar, além de poder ser utilizado em qualquer momento da vida acadêmica do aluno, como por exemplo no caso de novos alunos e transferências, tal sistema também pode ser considerado como uma ferramenta de apoio para intervenções futuras tais como via SMS, mensagem em aplicativos *mobile*, *e-mail* e contatos telefônicos (*nudging*) (Cunha, Lichand, Madeira & Bettinger, 2016).

Também pode ser considerado dinâmico e auto-ajustável tendo em vista usar dados educacionais comuns em sistemas públicos estaduais brasileiros, o que faz com que o sistema implementado tenha grandes possibilidades de uso em todas as escolas públicas dos diferentes estados brasileiros, além de ter a capabilidade de, ao final

de cada bimestre escolar brasileiro, ser atualizado com os dados mais recentes de alunos tais como notas, frequência, idade, defasagem idade-série, bem como diversos dados demográficos e censitários referentes às diferentes escolas.

Uma vez implementado, o sistema pode não se restringir a uma amostra de estudantes mas também ser usado de modo a monitorar grupos de estudantes individuais, diferentes programas de estudo, grupos de alunos inteiros, alunos que se encontram em diferentes turnos e escolas e, se desejado, até mesmo estudantes individuais.

Assim, nota-se que a existência de um EDS pode servir como ponto de partida para uma pesquisa mais minuciosa a respeito de evasão escolar bem como fornecer valiosos *insights* tendo em vista apoiar cada vez mais os processos estratégicos e de tomada de decisões nos diferentes sistemas educacionais (Gaebel, Hauschildt, Mühleck & Smidt, 2012) e, no caso em específico, servir como guia para a possível elaboração de novas políticas públicas governamentais.

O EDS em questão, por exemplo, nos permite estudar os efeitos de alterações nos currículos escolares e programas educacionais à influência de diferentes fatores como repasse destinado às escolas, tamanho da escola bem como pode auxiliar no monitoramento da eficiência de medidas de intervenção, programas de ajuda e medidas de apoio com relação a alunos prestes a evadir (Cunha, Lichand, Madeira & Bettinger, 2016).

Para a correta validação do *framework* e do EDS implementado, ele foi treinado e testado tendo por base informações dos quatro últimos bimestres do ano de 2019, advindas da Secretaria de Educação do Estado de São Paulo, com base nos mais de 2.836.000 alunos matriculados da sexta a nona série do ensino fundamental bem como do primeiro ao terceiro anos do ensino médio (logo tanto anos finais do ensino fundamental quanto ensino médio), de modo que os resultados aqui contidos englobam todas as 82.752 escolas do sistema público estadual que atendem tais alunos.

Tendo em vista reprodutibilidade e a exposição de uma melhor comparação com outros estudos da literatura, o EDS implementado foi desenvolvido e testado tendo por base quatro diferentes algoritmos da área de aprendizagem de máquina dentre os mais comuns na literatura atual para classificação binária, sendo eles *Decision Jungle*, Regressão Logística, *Bayes Point Machine* e *Decision Forest*, para cada bimestre.

Os resultados aferidos, a partir de treinamento e teste em dados 2018 via *decision jungle*, por exemplo, indicam que 91,45% das evasões são corretamente identificadas ao final do primeiro bimestre (7.531 verdadeiro positivos e 704 falsos negativos para a classe de evadidos), usando ambos, dados quanto a desempenho acadêmico do aluno bem como demográficos e censitários quanto a escola; além disso a precisão para a classe de evadidos tende a aumentar no segundo e terceiro

bimestres, à medida que novos dados de desempenho dos alunos ficam disponíveis no final de cada bimestre.

Após o quarto bimestre a predição tem seu pior desempenho quanto a acurácia total, ainda que mantendo alto *recall* e, portanto, ainda sendo de grande valia para a correta predição da classe de evadidos; neste estudo também são apresentados os atributos mais significativos para cada diferente classificação, de modo a demonstrar que o acesso a dados de desempenho, frequência e de censo são fundamentais para a correta predição quanto a evasão escolar.

Para efeito de roteiro o artigo está organizado da seguinte forma. A Seção 2 a seguir apresenta a problematização conhecida como aprendizagem de máquina para desenvolvimento de sistemas de detecção precoce de evasão bem como apresenta uma visão geral sobre o estado atual de alguns desses sistemas desenvolvidos ao redor do mundo, por meio da apresentação e discussão de trabalhos referência e breve explanação dos diferentes algoritmos de aprendizagem de máquina implementados e bases utilizadas em diferentes EDS.

Na Seção 3 é apresentada a metodologia de implementação do EDS desenvolvido, baseado numa proposta de *framework* se utilizando de aprendizagem de máquina; assim englobando desde a coleta de dados, o pré-processamento, a extração e utilização de diferentes atributos (baseados em dados administrativos) até a problematização da utilização de tais atributos para a construção de um modelo de classificação supervisionado via quatro diferentes algoritmos. Na Seção 4 há discussão dos resultados obtidos junto a uma comparação com outros trabalhos que se utilizaram de modelos de classificação supervisionada e, finalmente, na Seção 5 são expostas as conclusões e trabalhos futuros.

2. Sistemas de detecção precoce de evasão

Os sistemas voltados para a detecção precoce de evasão escolar costumam se utilizar de dados administrativos relevantes, tanto demográficos quanto de desempenho acadêmico, tais como dados pessoais (idade, gênero dentre outros), dados de educação prévia, notas e frequência escolar. Basicamente um sistema para detecção precoce pressupõe então a existência de dados que possam ser implementados e executados, de preferência sem o envolvimento do pessoal engajado diretamente nas instituições de ensino, facilitando consideravelmente os requisitos legais em relação às leis de proteção de dados.

Como citado anteriormente, tais sistemas podem ainda serem utilizados tendo em vista o monitoramento de grupos individuais de alunos, a recepção a novos programas de educacionais, ementas e (ou) políticas públicas, o acompanhamento de grupos de alunos inteiros e, se desejado, até mesmo a situação de alunos individuais. Assim, a existência de um EDS pode oferecer um importante ponto de partida para a pesquisa em evasão bem como oferecer *insights* importantes para a

administração de instituições de ensino e pode servir de base para futuras intervenções e construção de políticas e boas práticas para a redução no número de estudantes evadidos e, logo, aumento na retenção (Gaebel, Hauschildt, Mühleck & Smidt, 2012).

Podem, portanto, servir de apoio a processos decisórios estratégicos, táticos e operacionais das diferentes instituições. Por exemplo, a existência de um EDS permite se estudar os efeitos das mudanças nos programas de estudo e cursos, a influência de barreiras de entrada, tais como, por exemplo, taxas de estudo, características demográficas e sociais, bem como pode servir para se monitorar a eficiência das medidas de intervenção e programas de ajuda a estudantes em risco (Berens, Oster, Schneider & Burghoff, 2018).

Na literatura, a proposta de Tinto (1975), na qual um modelo pensado quanto a evasão no Ensino Superior deveria levar em conta fatores tais como *background* familiar, fatores individuais e educação prévia como fatores pré requisito para tal evasão, é tida como um dos primeiros *frameworks* teóricos na área. No artigo são discutidas possíveis causas raízes do processo de evasão e, principalmente, é atestado que um dos fatores primordiais para que um aluno esteja ou não em processo de evasão é justamente a interação entre estudantes e o ambiente de ensino, durante o processo de apredizagem.

Tal interação pode então ser entendida como a integração acadêmica, associada à performance e ao desenvolvimento intelectual, além de integração social, haja visto as diferentes interações estudante-estudante e estudante-professor. A partir dos trabalhos de Spady (1970) e Tinto (1975), e baseado em seus resultados, tem-se então o trabalho de Kember (1989) o qual, já voltado para o contexto de ensino a distância, introduziu uma nuance de custo-benefício ao tentar explicar o processo de tomada de decisão por parte dos estudantes, ao decidir a não terminar ou a continuar seus estudos.

Mais especificamente, a partir deste trabalho, notou-se que as variáveis que culminam na decisão dos estudantes, em evadirem ou não, podem mudar dinamicamente ao longo do tempo bem como existem várias ocasiões no processo de ensino aprendizagem onde os estudantes podem se sentir forçados a tomar tal decisão.

Já no trabalho de Bean e Metzner (1985) foi proposto um modelo conceitual o qual relacionou o processo de evasão a atributos tais como performance acadêmica (notas), variáveis acadêmicas (tais como hábitos de estudo), traços psicológicos (tais como eficácia, satisfação, pressão dentre outros) e variáveis ambientais (status econômicos, horários de trabalho e outros) além de características pré requisito como idade, gênero e local de residência do estudante, os quais podem estar, direta ou indiretamente, ligados ao processo de evasão.

Mais ainda, desde a década de 90, estudos como o de Seidman (1996), a respeito da retenção de alunos, mostram que a identificação precoce de estudantes em risco, além da manutenção de uma intervenção intensiva e contínua, é uma das chaves para se reduzir os níveis de evasão. Assim, a construção de metodologias voltadas para sistemas de detecção precoce (EDS) são vistos como uma boa solução para se detectar estudantes com alto risco o mais cedo possível e, de fato, tomar as medidas cabíveis dentro do contexto em que tal EDS está inserido.

O conceito de um EDS em si também não é novidade, ainda mais na era da tomada de decisão orientada a dados; visto que são basicamente qualquer sistema projetado para alertar tomadores de decisão sobre possíveis perigos. Sua finalidade é permitir a prevenção do problema antes que ele se torne um perigo real (Grasso, 2009). No domínio educacional, um EDS consiste num conjunto de procedimentos e instrumentos para a detecção precoce de indicadores de estudantes em risco de evasão escolar e envolve também a implementação de intervenções apropriadas para os fazer retornar as atividades educacionais (Heppen & Bowles, 2008), sendo então uma linha de pesquisa bastante interessante no que diz respeito a mineração de dados educacionais.

Na maioria dos trabalhos analisados esses indicadores são, em sua grande maioria, os aspectos de desempenho acadêmico dos alunos, os quais podem refletir com precisão o risco de desistência correspondente a cada um deles em um determinado momento. Entretanto, detectar esses indicadores ou fatores é realmente difícil, porque não há uma única razão para os alunos se evadirem dos sistemas educacionais e, de fato, evasão pode ser encarada como um problema dito multi-fatorial, também conhecido na literatura como o problema dos “mil fatores” (Hernandez, 2002).

Os EDS então, em sua grande maioria, são desenvolvidos de modo a ter observação regular de tais indicadores específicos tendo em vista acompanhamento do desempenho escolar dos alunos antes de se evadirem. Nos últimos anos, o esforço para se criar sistemas EDS mais assertivos para contextos educacionais aumentou e, atualmente, existem vários exemplos de EDS implementados em diferentes países. Alguns dos EDS que analisados neste trabalho foram os seguintes:

O Centro Nacional de Ensino Médio dos EUA, por exemplo, propôs, ainda em 2008, um guia e um EDS (Heppen & Bowles, 2008). Tal EDS é baseado em um modelo no Microsoft Excel e em apenas dois indicadores (desempenho do estudante e sua frequência no curso). A partir desta ferramenta, o Departamento de Educação de Delaware, nos EUA, implementou um EDS nos estados de Chicago, Colorado e Texas usando um modelo multivariável para determinar quais indicadores tinham a correlação mais forte com a evasão por parte dos alunos (Uekawa, Merola, Fernandez & Porowski, 2010).

Outro artigo de interesse, ainda a partir da realidade norte-americana, é o de Sansone (2017), o qual visou a criação de um modelo que identifica os alunos que

correm o risco de desistir usando informações referentes ao primeiro ano do ensino médio; Já para a realidade mexicana, pode-se citar o trabalho realizado pela Subsecretaria de Ensino Médio do México, a qual definiu várias diretrizes para acompanhar a educação de jovens estudantes e desenvolveu um EDS baseado em um arquivo do Microsoft Excel (Maldonado-Ulloa, Sancén-Rodríguez, Torres-Valades & Murillo-Pazaran, 2011).

Tal EDS tem o objetivo de gerar alertas, a partir de três indicadores principais (absenteísmo, baixo desempenho e comportamento ou conduta problemáticos), tendo limiares críticos específicos os quais são níveis nos quais pode se considerar que a probabilidade de evasão é geralmente maior. Ainda a partir da realidade mexicana, também foi analisado o trabalho de Carlos Márquez-Vera et al. o qual se utilizou de dados referentes a 419 estudantes do ensino médio mexicano (Marquez-Vera, Romero & Ventura, 2011);

No continente europeu, três países (Áustria, Croácia e Inglaterra) desenvolveram um EDS focado no monitoramento sistemático de absenteísmo e resultados (por meio das notas dos estudantes) (Vassiliou, 2013). Já no estudo dinamarquês de Şara et al. (2015), a partir de dados administrativos do sistema MaCom Lectio, foi utilizada uma amostra significantemente maior que na maioria dos estudos aqui expostos, sendo usadas as informações referentes a 36.299 alunos para treinamento e 36.299 para testes;

Já o estudo alemão de Berens et al. (2018) se utilizou de análise de regressão e métodos de aprendizado de máquina tais como redes neurais, árvores de decisão e o algoritmo AdaBoost para a identificação de características dos alunos que podem distinguir potenciais desistentes dos graduados. O EDS apresentado nesse estudo foi testado e aplicado em uma universidade estadual de porte médio com 23.000 alunos e em uma universidade privada de médio porte, de ciências aplicadas, com 6.700 estudantes;

Tendo em vista que as técnicas baseadas em estatística e, principalmente, na utilização de aprendizagem de máquina, têm sido cada vez mais usadas para se prever evasão escolar, vê-se então a importância de se propor um *framework* generalista para a criação de um EDS, ao se analisar diferentes atributos quanto a desempenho, frequência e várias outras características de censo e demográficas.

Nas próximas seções especial atenção é dada à comparação dos resultados obtidos a partir do desenvolvimento de quatro diferentes algoritmos de aprendizagem de máquina, a partir da utilização de tal *framework*, além de mais detalhes em como essa problematização tem sido abordada nos diferentes estudos bem como contextualização com relação as diferentes referências metodológicas observadas na literatura tendo em vista prover um arcabouço conciso de observações, principalmente quanto a técnicas e atributos, que possam vir a ser utilizadas para a implementação de EDS cada vez mais generalistas, dinâmicos e assertivos.

3. Metodologia de Implementação

3.1. Framework

O presente estudo visa a utilização de diferentes técnicas e algoritmos em aprendizagem de máquina tendo em vista a construção de um modelo de predição que, com base no treinamento em dados de anos anteriores, pode ser utilizado para predição quanto ao ano anterior, ao final de cada bimestre ou qualquer outro período de tempo estipulado. Este acompanhamento deve se dar de forma espaçada mas que, ainda assim, possa atingir o objetivo de identificar possíveis casos de evasão antes mesmo que o fato aconteça.

A essência deste estudo é então a construção de um modelo de classificação binário no qual, a partir de dados administrativos previamente aferidos ao final de cada ano letivo, as amostras podem ser categorizadas em duas classes, a saber: a classe de evadidos e a classe de retidos. As variáveis , quando da avaliação de uma única amostra são compostas de duas partes: a atribuição de entrada X = (x_1, x_2, ..., x_n) e a atribuição de categoria Y; o processo de construção de um modelo de classificação é o estabelecimento de uma função de mapeamento na forma y = *f(*X*)* Na qual a função pode ser usada para determinar a atribuição de categoria Y de uma amostra de acordo com a atribuição de entrada de amostra X, estabelecendo assim um padrão aferido para predição quanto a evasão escolar. A estrutura geral do estudo é mostrada na Figura 1 a seguir:

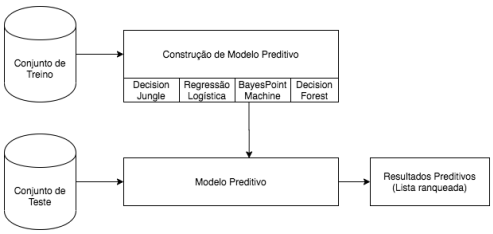

Figura 1 – *Framework* para detecção precoce de estudantes em risco de evasão

A metodologia proposta pode ser dividida em alguns passos fundamentais: 1 – extração de atributos relacionados a evasão e existência de dados prévios rotulados, a partir de dados administrativos, para a construção dos classificadores como conjunto de dados de treinamento; 2 – utilização dos dados para treinamento dos modelos de predição por meio dos algoritmos de aprendizagem de máquina *Decision Jungle*, Regressão Logística, *Bayes Point Machine* e *Decision Forest*; 3 – extração de outra seção de dados dos sistemas de dados administrativos para construir um conjunto de dados de teste e alimentação de tais amostras reais do modelo de previsão gerado anteriormente; 4 - utilização das amostras do modelo

de previsão para executar previsões no conjunto de dados de teste e avaliar os resultados de previsão gerados.

3.2. Descrição dos dados

Os dados usados no presente artigo são referentes a Secretaria de Educação do Estado de São Paulo (educacao.sp.gov.br), mais especificamente, disponibilizados em sua plataforma de dados abertos (dados.educacao.sp.gov.br), além de dados de censo (ibge.gov.br), da sexta a nona série do ensino fundamental bem como do primeiro ao terceiro anos do ensino médio (logo tanto anos finais do ensino fundamental quanto ensino médio) do ensino público estadual presencial, no estado de São Paulo, Brasil.

A partir desses dados foram extraídos 50 diferentes atributos (descritos na subseção a seguir) para os mais de 2.836.000 alunos regularmente matriculados, englobando todas as 82.752 escolas do sistema público estadual que atendem tais alunos. A partir dos dados rotulados, a taxa de evasão aferida se aproxima de 3% de nossa base de alunos para treinamento de nossos algoritmos, entretanto, a literatura mostra que altas taxas de evasão são um problema costumeiro nos mais diversos sistemas educacionais brasileiros.

No sistema educacional paulista, a nível de sistema, o aluno estar considerado evadido ou não (como visto anteriormente, $Y=1$) se dá quando o aluno deixa de frequentar as aulas em determinado bimestre, entretanto, no sistema, há a possibilidade desse aluno voltar a frequentar as aulas normalmente, enquanto o abandono é considerado como uma evasão por maior tempo.

Também é de interesse notar que, muitas vezes, uma evasão pode ter ocorrido e não ter sido corretamente registrada em sistema, portanto, além de ser um classificador binário, o produto do *framework* aqui proposto é um lista ranqueada do risco de evasão de nossos estudantes bem como há, na seção posterior de resultados, exemplo de comparação a respeito dos alunos em risco preditos por um dos modelos desenvolvidos bem como o número de evadidos e abandonos registrados em sistema.

Um análise descritiva dos dados de treinamento aferidos acabou mostrando uma grande correlação entre o desempenho acadêmico e frequência com a evasão além de que foi verificado que a o número de alunos evadidos por bimestre aumenta após as férias escolares, de acordo com o observado na literatura. Assim, na Tabela 1 abaixo estão descritos os números de alunos com dados utilizados como ambos, conjunto de treinamento, testes e para predição em dados 2019 a partir do modelo treinado e testado com dados 2018, para cada um dos quatro bimestres analisados:

	Conjunto de treinamento (Dados 2018)	**Conjunto de testes (Dados 2018)**	**Predição (Dados 2019)**
1B	2.292.851	573.213	2.866.064

2B	2.269.342	567.336	2.774.251
3B	2.269.336	567.334	2.777.538
4B	2.269.328	567.332	2.778.631

Tabela 1 – Descrição do número de amostras utilizadas para treinamento, teste e predição para cada bimestre analisado (anos 2018 para treinamento e teste e 2019 para validação de predição)

O *framework* proposto, baseado na utilização de informações dos quatro diferentes bimestres no ano, faz com que tenhamos então quatro intervalos de inspeção, onde fazemos a verificação das taxas de evasão e abandono bem como da taxa de risco de evasão, além de que, em tais intervalos, ações podem ser tomadas pelos órgãos competentes de modo a haver uma busca ativa dos alunos em risco de evasão (tendo em vista atestar o porquê do aluno estar em processo de evasão e fornecer acompanhamento personalizado) ou ao menos um programa de alerta junto a pais e responsáveis (*nudging*) (Cunha, Lichand, Madeira & Bettinger, 2016).

A partir de um método de partição e seleção aleatória, e de acordo com as prática tradicionais na área de aprendizagem de máquina, os dados foram divididos em conjuntos de treinamento e testes na proporção de 80:20. As taxas de evasão dos dois conjuntos de dados foram consistentes após a partição. Os atributos utilizados, os quais podem ser verificados na seção a seguir, em um total de 50, são todos referentes a desempenho acadêmico, frequência, dados censitários e sociais (de cadastro, quando da matrícula) do aluno além de informações referentes a escola frequentada por ele. Os atributos de características individuais para cada aluno foram todos obtidos a partir de sistema administrativos enquanto “repasse por aluno”, “defasagem ano/série“ e outros atributos numéricos são calculados via sistema. As notas de português, matemática, geografia e história bem como a frequência são as mesmas registradas em sistema pelos professores responsáveis.

Tendo em vista melhoria na predição há pré-processamento quanto a dados ausentes; os dados binários e categóricos foram preenchidos com a moda, ou seja, valores mais comuns para cada atributo, e dados numéricos foram preenchidos com a média, sendo que variáveis numéricas contínuas foram discretizadas com base no requisito de cada um dos algoritmos utilizados. Dadas as características de desequilíbrio (classe de evadidos consistindo em apenas 3% do total de nossos dados) dos conjuntos de dados, para garantir que a informação quanto a classe majoritária de dados não fosse perdida, os dados da classe minoritária (evadidos) foram duplicados de modo a equilibrar os dados do conjunto de treinamento (*oversampling*).

3.3. Atributos

Neste estudo, os atributos usados para se prever o risco de evasão escolar, além de devidamente correlacionadas ao abandono escolar, como explicitado

anteriormente, também precisam ser obtidos por meio de dados administrativos. Assim, as primeiras informações a serem utilizadas para extração de atributos foram características individuais dos alunos, tais como idade, defasagem (distorção idade/série), nota no IDESP (http://idesp.edunet.sp.gov.br/), notas em matemática, geografia, português e história, além de frequência nessas mesmas disciplinas dentre outros mais.

Por sua vez, a partir das informações quanto a escola a qual determinado aluno atende, consegue-se então se utilizar também de atributos da escola, tendo em vista discretizar ainda mais o que é de fato um aluno em evasão, tais como número de domicílios na região da escola, rendimento médio domiciliar, idade média das pessoas responsáveis além de outras características referentes ao entorno da escola, e da própria escola, tais como número de estudantes matriculados, repasse por escola dentre outros. O foco em tais características deveu-se às características individuais dos alunos serem definidas como fator pré-requisito para o processo de evasão escolar em todos os modelos teóricos vistos na literatura. Os atributos de desempenho acadêmico e de frequência dos alunos, os quais costumam ser tradicionais medidas do sistema de ensino em questão, também têm grande importância no *framework* apresentado.

Apesar da associação com o abandono escolar, ambos os dados, seja de evasão ou abandono, também presentes no sistema administrativo, tendo em vista a construção de nosso modelo de predição, não foram utilizados como atributos nesse estudo, apenas de modo a corroborar os resultados obtidos. Assim, foram utilizados dois conjuntos de atributos (referentes a alunos e escolas) como entrada de nosso modelo preditivo; conjuntos esses selecionados com base em sua relação com a evasão, bem como na facilidade de aquisição desses dados nos sistemas de informação disponibilizados. Os códigos e descrição dos atributos utilizados podem ser encontrados no seguinte *link*: https://github.com/AhirtonLopes/Phd.--Project/blob/master/Detalhamento%20-%20Atributos%20Evasao%20Escolar.xlsx

4. Discussão e resultados obtidos

Após a correta divisão dos dados disponíveis em conjuntos de treinamento e teste na proporção de 80:20, como visto na Tabela 1 anterior com, respectivamente, 2.292.851 e 573.213 estudantes no primeiro bimestre, 2.269.342 e 567.336 para o segundo bimestre, 2.269.336 e 567.334 estudantes para o terceiro e 2.269.328 e 567.332 para o quarto bimestre. As amostras de treinamento são utilizadas de modo a alimentar quatro algoritmos, sendo eles Decision Jungle, Regressão Logística, Bayes Point Machine e Decision Forest; os testes e resultados das predições para o conjunto de teste são mostrados nas Tabelas 2 e 3 a seguir.

	Acurácia	Precisão	Recall	F1 Score
	Decision Jungle (DT)			
1B	0,843	0,078	0,915	0,143
2B	0,871	0,095	0,934	0,172
3B	0,877	0,101	0,959	0,183
4B	0,829	0,076	0,978	0,142
	Regressão Logística (RL)			
1B	0,766	0,053	0,915	0,101
2B	0,775	0,057	0,944	0,108
3B	0,857	0,082	0,872	0,150
4B	0,742	0,049	0,924	0,094
	Bayes Point Machine (BPM)			
1B	0,715	0,045	0,943	0,087
2B	0,755	0,053	0,949	0,100
3B	0,753	0,053	0,961	0,101
4B	0,542	0,030	0,993	0,059
	Decision Forest (DF)			
1B	0,982	0,322	0,253	0,284
2B	0,983	0,408	0,373	0,390
3B	0,983	0,429	0,515	0,468
4B	0,980	0,354	0,501	0,415

Tabela 2 – Resultados de acurácia, precisão, *recall* e F1 *score* quando da utilização dos quatro algoritmos para treinamento e teste no conjunto de treinamento (dados 2018)

	Verdadeiros Positivos (VP)	Falso Negativos (FN)	Falso Positivos (FP)	Verdadeiro Negativos (VN)
	Decision Jungle (DJ)			
1B	7.531	704	89.236	475.742
2B	7.611	534	72.711	486.480
3B	7.832	337	69.635	489.530

4B	8.004	181	96.908	462.239
	Regressão Logística (RL)			
1B	7.534	701	133.449	431.529
2B	7.685	460	126.982	432.209
3B	7.126	1.043	79.940	479.225
4B	7.561	624	145.814	413.333
	Bayes Point Machine (BPM)			
1B	7.762	473	162.931	402.047
2B	7.731	414	138.325	420.866
3B	7.848	321	139.740	419.425
4B	8.128	57	259.732	299.415
	Decision Forest (DF)			
1B	2.086	6.149	4.394	560.584
2B	3.039	5.106	4.406	554.785
3B	4.211	3.958	5.597	553.568
4B	4.100	4.085	7.487	551.660

Tabela 3 – Matrizes de confusão quando da utilização dos quatro algoritmos para treinamento e teste no conjunto de treinamento (dados 2018)

Ao observarmos tanto os resultados para acurácia, precisão, *recall* e *F1 score*, da Tabela 2, quando a matriz de confusão da Tabela 3, percebemos que, em geral, buscou-se ter algoritmos genéricos com alta assertividade para a classe de evadidos (Y=1), tendo em vista que, nesse tipo de sistema educacional, maior importância é dada em agir rapidamente, principalmente em casos cujo risco de evasão é mais elevado, importando mais a correta identificação de evadidos (evidenciados aqui por altos valores de *recall*) que a acurácia total. A partir do *framework* sugerido vemos que, para esse conjunto de dados em particular, no qual temos poucos dados quanto a evasão para treinamento, a precisão é geralmente baixa.

A taxa de acurácia geral acabou sendo mais alta quando da utilização de Decision Forest (DF), bem como foram verificados os maiores valores quanto a acurácia total do modelo (98,3%) enquanto os valores de *recall* ficaram equilibrados numa faixa entre 87,2% (Regressão logística - RL, para o terceiro bimestre) e 99,3% (*Bayes point machine* - BPM, para o quarto bimestre). Como dito anteriormente, o objetivo do *framework* e do EDS desenvolvido, foi identificar possíveis processo de evasão. De modo geral, altos valores de *recall*, como os alcançados nesse estudo,

quando em referência à classe de evadidos, reflete a eficácia geral dos modelos de predição na classificação da classe de evadidos, ou seja, o total de instâncias relevantes classificadas corretamente.

Em geral, excetuando-se o uso de Decision Forest, as demais técnicas são efetivas ao classificar corretamente alunos em risco de evasão escolar onde, comparativamente, os algoritmos baseados em técnicas de Decision Jungle e Regressão Logística, tendo em vista os altos valores de *recall*, evidenciados pelo número de verdadeiros positivos (para a classe de evadidos).

A precisão em geral, para esse tipo de problema, acaba sendo prejudicada tendo em vista a distribuição natural de dados, ou seja, dados quanto a evasão escolar não suficientemente compreensíveis, principalmente quanto mais atributos sejam utilizados. Visto que os dados utilizados aqui foram dados administrativos, a única barreira é de fato o que é ou não passível de coleta em plataformas oficiais, principalmente durante o período de matrícula. O conjunto de predição, também mencionado anteriormente, foi utilizado para avaliar, tanto os resultados da previsão (onde temos quantidade total de alunos em determinada faixa de risco de evasão, quais destes são alunos de ensino fundamental anos finais, quais destes são alunos de ensino médio e quantos já estavam registrados em sistema como abandonos) quanto uma seleção dos 15 atributos considerados mais significantes dentre os utilizados, resultados estes se utilizando de regressão logística, para o primeiro bimestre (Tabelas 4 e 5, respectivamente).

Risco de Evasão	***QTD***	***FUND2***	***EM***	***ABANDONO***
60% - 65%	65.406	2.4071	41.183	1.370
65% - 70%	60.244	2.1715	38.412	1.522
70% - 75%	57.459	20.001	37.356	1.842
75% - 80%	55.153	18.398	36.664	2.087
80% - 85%	54.005	17.240	36.700	2.670
85% - 90%	55.006	16.723	38.197	3.626
90% - 95%	58.003	16.415	41.518	5.460
95% - 100%	85.554	19.290	66.202	7.708
Total	**2.773.980**	**1.461.430**	**1.307.198**	**34.777**

Tabela 4 – Resultados de classificação para o conjunto de testes (dados 2019)

ATRIBUTO	**RELEVÂNCIA PARA A ACURÁCIA**
VL_NOTA_LP	0.036207

VL_NOTA_GEO	0.034959
VL_NOTA_MAT	0.033355
VL_NOTA_HIST	0.032414
TX_EVASAO	0.031892
VL_DISTORCAO	0.024777
QT_ALUNO_EVADIDO	0.010054
NR_SERIE	0.007287
CD_TURNO	0.006262
NM_TURNO	0.006262
TX_FREQ_LP	0.005016
TX_FREQ_MAT	0.00402

Tabela 5 – Resultados quanto a ranqueamento da importância dos atributos utilizados (via ganho de informação) para o melhor resultado de classificação em nosso conjunto de treinamento (dados 2018)

5. Conclusões e trabalhos futuros

O presente artigo buscou, por meio de uma pesquisa compreensiva da literatura no uso de aprendizagem de máquina, como uma abordagem supervisionada, para a classificação quanto ao risco de evasão escolar, por meio de dados administrativos utilizando tanto características individuais dos alunos quanto das escolas as quais eles frequentam como atributos de entrada de nosso modelo preditivo. Quatro métodos foram testados por meio de algoritmos de aprendizagem de máquina implementados como um EDS (DJ, RL, BPM e DF), de modo a serem utilizados para a elaboração de modelos preditivos de classificação, de modo supervisionado, com intervalos e avaliações bimestre a bimestre. Os resultados mostraram que ao menos três dos modelos construídos são efetivos na predição quanto a evasão escolar, tendo em vista maior retenção, com foco na detecção da classe de evadidos; dentre estes os baseados em DJ e RL tiveram resultados mais equilibrados para uso em produção (tendo em vista melhor classificação da classe de evadidos).

A partir de um sistema previamente implementado, uma lista atualizada com informações de alunos em maior risco de evasão se encontra disponível para todos os perfis relacionados aos serviços de apoio à aprendizagem dos alunos, de modo que melhor acompanhamento individualizado possa acontecer e que possíveis novas políticas públicas, voltadas a retenção de alunos, possam ser testadas e implementadas. Os resultados parciais comprovam que o EDS desenvolvido tem um grande potencial como ferramental baseada em mineração de dados

educacionais. O estudo tem um grande valor prático, sendo chave para um posterior processo de alerta e acompanhamento personalizado de alunos em situação de evasão, fazendo com que haja critérios tangíveis para que professores e demais perfis possam, por meio de uma ferramenta simples, ter informação de maneira disponível antes que o processo ocorra. Assim, o método proposto já se encontra em utilização e disponível para todos os perfis de professores e coordenação de todas as escolas estaduais do Estado de São Paulo – SP, Brasil.

Melhorias no f*ramework* e no EDS desenvolvido passam pelo teste de mais atributos, assim como otimização de algoritmos e correta seleção quanto a atributos mais significativos; tais novos atributos são, principalmente, advindos de uma triangularização de dados quanto a localização do estudante, a qual, mesmo que de maneira anonimizada, pode nos trazer dados importantes não mais somente do entorno da escola, mas sim da vizinhança onde mora cada aluno, fornecendo então atributos cada vez mais individualizados.

Referências

Sansone, D. (2017). Beyond Early Warning Indicators: High School Dropout and Machine Learning. http://dx.doi.org/10.2139/ssrn.3062317

Berens J., Schneider K., Görtz S., Oster S., Burghoff J. (2018). Early Detection of Students at Risk – Predicting Student Dropouts Using Administrative Student Data and Machine Learning Methods, CESifo Working Paper Series 7259, CESifo Group Munich. https://ideas.repec.org/p/ces/ceswps/_7259.html

OECD (2020), Secondary graduation rate (indicador). doi: 10.1787/b858e05b-en

Public High School Graduation Rate Reaches New High, But Gaps Persist. (2016). U.S. Department of Education, Institute of Education Sciences, Washington, DC.

Carnevale A. P., Smith, N. and Strohl, J. (2013). Recovery - Job Growth and Education Requirements through 2020, Center on Education and the Workforce, Georgetown University, Washington, DC.

Marquez-Vera C., Romero C. and Ventura S. (2011). Predicting School Failure Using Data Mining, Educational Data Mining Conference, Eindhoven, Netherlands, 271–275.

Oreopoulos, P. and Salvanes, K. G. (2011). Priceless: the nonpecuniary benefits of schooling, Journal of Economic Perspective, Vol. 25, pp. 159–184.

Neild, R.C., R. Balfanz and l. Herzog. (2007). An early warning system, Educational leadership. Association or supervision and curriculum development.,1–7.

Heppen, J.B. e Bowles, S. (2008). Developing early warning systems to identify potential high school dropouts, National High School Center, American Institutes for Research.,1–13.

Gaebel, M., Hauschildt, K., Mühleck, K. & Smidt, H. (2012). Tracking Learners' and Graduates' Progression Paths. TRACKIT. EUA Publications.

Tinto, V. (1975). Dropouts from higher education: a theoretical synthesis of the recent literature. Review of Educational Research, 45: 89-125. http://dx.doi.org/10.3102/00346543045001089

Spady, W.G. (1970). Interchange 1: 64. https://doi.org/10.1007/BF02214313

Kember, D. (1989). A longitudinal-process model of dropout from distance education. The Journal of Higher Education, 1989, 60 (3): 278-301. http://dx.doi.org/10.2307/1982251

Bean, J. P. e Metzner, B. S. (1985). A conceptual model of non traditional undergraduate student attrition. Review of educational Research, 55(4): 485-540.

Seidman, A.(1996). Spring retention revisited: RET=EId+(E+I+C) Iv, College and University, 71,18–20.

Grasso, V.F. (2009). Early warning systems: state-of-art analysis and future directions, Draft report United Nations Environment Programme (UNEP), 1,1–66.

Hernandez, M.M. (2002). Causas del Fracaso Escolar; XIII Congreso de la Sociedad Española de Medicina del Adolescente, España,1–5.

Uekawa K., Merola S., Fernandez F., Porowski A. (2010). Creating an early warning system: predictors of dropout in Delaware. Regional Educational LaboratoryMid Atlantic, 1,1–50.7.

Maldonado-Ulloa P.Y., Sancén-Rpdríguez A.J., Torres- Valades M., Murillo-Pazaran B. (2011). Secretaria de Educación Pública de Mexico. Programa Síguele. Sistema de Alerta Temprana, Lineamientos de Operación. 1–18.

Vassiliou, A. (2013). Early warning systems in Europe: practice, methods and lessons, Thematic Working Group on Early School Leaving.,1–17 .

Şara, N-B., Halland, R., Igel, C., & Alstrup, S. (2015). High-school dropout prediction using machine learning: a Danish large-scale study. In M. Verleysen (Ed.), Proceedings. ESANN 2015: 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (pp. 319-324). i6doc.com.