Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
http://br-ie.org/pub/index.php/rbie

Submission: 20/06/2020; 1st round notif.: 30/07/2020; New version: 27/08/2020; 2nd round notif.: 20/10/2020
Camera ready: 16/11/2020; Edition review: 17/11/2020; Available online: 20/11/2020; Published: 20/11/2020

Aplicação de Técnicas de Aprendizado de Máquina Para Predição de Risco de Evasão Escolar em Instituições Públicas de Ensino Superior no Brasil

Title: Machine Learning Applied to Academic Drop Out Prediction in Brazilian Public Universities

Leonardo de Almeida Teodoro
Aluno PIBIC – Programa Institucional de Bolsas de Iniciação Científica – Centro Federal de Educação Tecnológica Celso Suckow da Fonseca – CEFET/RJ – Campus Nova Friburgo, Nova Friburgo, RJ, Brasil
leonardo.teodoro@aluno.cefet-rj.br

Marco André Abud Kappel
Centro Federal de Educação Tecnológica Celso Suckow da Fonseca – CEFET/RJ – Campus Nova Friburgo, Nova Friburgo, RJ, Brasil
marco.kappel@cefet-rj.br

Resumo

As instituições públicas de ensino superior do Brasil enfrentam taxas de evasão anual preocupantes. Torna-se de extrema importância, então, o reconhecimento do perfil de alunos com maior probabilidade de evadir, levando em consideração características dos estudantes e das universidades em que eles se encontram matriculados, para que planos de medidas públicas sejam construídos de maneira a reduzir estas taxas. Nesse contexto, o presente trabalho tem como objetivo a identificação dos padrões característicos de alunos com maior tendência a abandonar o ensino público superior, assim como a identificação dos atributos mais determinantes nestes padrões. Para isso, foram aplicadas cinco técnicas de aprendizado de máquina nos dados de educação superior do INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira): Naive Bayes, K-Nearest Neighbors, Árvores de Decisão, Random Forest e Redes Neurais. Dentre elas, o melhor resultado foi obtido pela técnica Random Forest, que alcançou uma taxa de acerto de aproximadamente 80% das previsões de evasão. O modelo construído indicou que algumas das características mais determinantes na evasão de um aluno são a idade, a participação em atividades extracurriculares e a carga horária total do curso. A principal contribuição do presente trabalho vem na forma da identificação das variáveis mais importantes para a previsão de evasão. Espera-se que os resultados aqui apresentados possibilitem o desenvolvimento de estratégias de redução de evasão focadas no suporte a estudantes que se encontram nos padrões característicos identificados.

Palavras-Chave: *Predição de evasão; Aprendizado de Máquina; Evasão Escolar; Análise de evasão escolar; Extração de características.*

Abstract

Brazilian public educational institutions face worrisome annual dropout rates. Therefore, it is essential to recognize students' profiles who are most likely to drop out, considering their characteristics and universities. In this context, the present work aims to identify students with a higher tendency to drop out of public universities and the most determining features for this prediction. To achieve this goal, five machine learning techniques were applied to INEP's education data: Naive Bayes, K-Nearest Neighbors, Decision Trees, Random Forest and Neural Networks. The best result was obtained by the Random Forest technique, which achieved a success rate of approximately 80% of the evasion predictions. The developed model indicates that some of the most determinant characteristics in students' dropout prevision are age, participation in extra-curricular activities, and the course's total hours. The main contribution of the present work is the identification of the most important characteristics for dropout prediction. The presented results can be used to motivate the development of dropout reduction strategies, focused on the support of students that fit the identified characteristic patterns.

Keywords: *Dropout prediction; Machine Learning; Scholar dropout; Scholar dropout analysis; Feature extraction.*

Cite as: Teodoro, L. A. & Kappel, M. A. A. (2020). Machine Learning Applied to Academic Drop Out Prediction in Brazilian Public Universities (Aplicação de Técnicas de Aprendizado de Máquina para Predição de Risco de Evasão Escolar em Instituições Públicas de Ensino Superior no Brasil). Brazilian Journal of Computers in Education (Revista Brasileira de Informática na Educação - RBIE), 28, 838-863. DOI: 10.5753/RBIE.2020.28.0.838

1 Introdução

A redução da taxa de evasão de alunos é um dos grandes desafios enfrentados pelas instituições públicas de ensino do Brasil, na atualidade (Prestes & Fialho, 2018). Nesse contexto, a identificação de padrões e perfis de alunos com maior risco de evasão revela-se fundamental para a construção de um plano de mitigação voltado à redução da probabilidade de abandono do curso pelo estudante, envolvendo programas de assistência permanente para o fortalecimento do vínculo acadêmico. De acordo com a Sinopse Estatística da Educação Superior, publicada pelo INEP (Instituto nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2018), as universidades públicas brasileiras, em sua totalidade, receberam 2.152.752 alunos em 2017. Por outro lado, segundo o levantamento, apenas 947.606 alunos concluíram os estudos universitários no mesmo ano. Com o número de alunos matriculados aumentando a cada ano (Costa & Dias, 2016), a desproporção na razão entrada/saída nas universidades brasileiras pode configurar o desperdício de recursos públicos, gerando problemas econômicos, acadêmicos e sociais (Filho, Motejunas, Hipólito, & Lobo, 2007), (Sales, Balby, & Cajueiro, 2016).

A aplicação de técnicas de Aprendizado de Máquina tem se tornado uma das formas mais eficazes para a determinação e classificação de padrões em massas de dados nos dias de hoje. Com a utilização de técnicas e estratégias desta natureza, é possível realizar inferências direcionadas à previsão do risco de evasão de alunos em instituições públicas de ensino superior no Brasil. Este processo, quando aplicado à área da Educação, é conhecido como Mineração de dados Educacionais (Baker, Isotani, & Carvalho, 2011).

Técnicas como Redes Neurais são capazes de processar grandes quantidades de dados e identificar padrões que não são facilmente determinados por humanos. Estes padrões podem ser utilizados para gerar um modelo computacional que permite a classificação de um aluno como provável concluinte ou provável evasão, dadas as suas características e as do curso em que está inserido, ou pretende se inserir. Além disso, técnicas como Árvores de Decisão e Random Forest permitem uma avaliação interna do modelo, possibilitando a identificação dos atributos mais determinantes na classificação do estudante (Gislason, Benediktsson, & Sveinsson, 2006), (Vlahou, Schorge, Gregory, & Coleman, 2003).

Na literatura, é possível encontrar alguns trabalhos que realizaram a aplicação de técnicas de Mineração de Dados e Aprendizado de Máquina para a análise da evasão escolar (Nascimento, Junior, & Roberta, 2018), (Ferreira, 2015), (Santos, Siebra, & Oliveira, 2014). Porém, normalmente, essas análises são realizadas dentro de um escopo específico de uma base de dados local, como de uma universidade (Manhães, Cruz, Costa, Zavaleta, & Zimbrão, 2011), (Pinheiro, Silva, & Souza, 2018) ou curso (Reis, Cunha, & Spritzer, 2012). No presente trabalho, os dados do INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira) foram filtrados de forma a incluir todas as instituições públicas de ensino superior no Brasil.

O presente trabalho possui como objetivo principal investigar e identificar padrões nas características de estudantes de graduação, a partir da aplicação de técnicas de Aprendizado de Máquina a uma base de dados públicos do INEP, de forma a possibilitar a previsão do risco de um aluno desistir de seus estudos em uma instituição de ensino superior pública. As características que serão estudadas envolvem tanto dados demográficos dos alunos, como sexo, idade e forma de ingresso na instituição, como também atributos da universidade e curso em que ingressou, como área do curso, infraestrutura presente e percentual de docentes mestres e doutores. Como contribuição principal, o presente trabalho pretende enumerar as características mais determinantes para a classificação de um aluno como possível concluinte ou possível desistente.

Os resultados deste estudo poderão servir como subsídio à tomada de decisão na construção de políticas de combate à evasão escolar, direcionadas e focadas em padrões possivelmente não-

triviais, dificilmente identificados em uma quantidade massiva de dados como a disponibilizada pelo INEP.

A seção 2 apresenta trabalhos relacionados com o tema abordado, de forma a contextualizar o trabalho e revisar o estado da arte. Em seguida, a seção 3 explica o processo adotado para a obtenção e preparação dos dados, descreve a etapa de pré-processamento e as técnicas de aprendizado de máquina adotadas, além de listar as principais métricas usadas para a avaliação dos algoritmos. Depois, a seção 4 apresenta os resultados referentes ao desempenho das técnicas e os atributos apontados como mais importantes segundo o classificador mais preciso. Também apresenta uma análise exploratória das variáveis identificadas, detalhando suas relações com a evasão escolar. Por fim, a seção 5 contém as conclusões do presente trabalho.

2 Trabalhos Relacionados

A base de dados do INEP possui uma extensa quantidade de informações sobre as instituições de ensino públicas e privadas de todos os níveis, de forma ampla e de acesso livre. Devido ao tamanho e à complexidade dos dados contidos na base completa do INEP, muitos trabalhos estudam casos específicos de um curso ou universidade.

O trabalho de (Rodrigues, Brackmann, & Barone, 2015) utilizou dados do curso de Ciência da Computação da Universidade Federal do Rio Grande do Sul e modelos conceituais para identificar os fatores mais importantes para a evasão. O estudo conclui que o tempo previsto para o término do curso não é respeitado pela maioria dos alunos. Além disso, o número de alunos desligados com ingresso no primeiro semestre letivo de cada ano é maior do que o daqueles com ingresso no segundo semestre letivo. O trabalho também aponta que há uma incompatibilidade entre as exigências dos cursos na questão de conhecimentos que deveriam ter sido obtidos, por parte dos alunos, nos ensinos anteriores ao superior. O estudo revela uma tendência de evasão prematura nos cursos estudados, indicada pela diminuição de inscritos no vestibular dos cursos. Por fim, os autores apontam como fatores importantes para a redução da evasão: o investimento em atividades de monitoria, a capacitação e a educação dos docentes, a maior flexibilização de horários de disciplinas, o aumento da oferta de cursos noturnos, o aumento da oferta de bolsas de pesquisa, de extensão, de assistência e de monitoria, e de estágios remunerados. A maior parte destas conclusões são confirmadas nos resultados do presente trabalho.

O problema de previsão computacional da evasão no ensino superior foi abordado pelo trabalho de (Santos, Menezes, Carvalho, & Montesco, 2019), em que o desempenho de diferentes técnicas de aprendizado de máquina foi comparado. Os dados utilizados foram extraídos da base da Universidade Federal de Sergipe, englobando apenas três cursos, todos da área de Computação. O classificador que obteve os melhores resultados, no caso, foi o Random Forest, que alcançou uma acurácia média de 0,69. Com o mesmo objetivo, o trabalho de (Martins, Carvalho, & Carvalho, 2017) utilizou dados referentes especificamente aos alunos da Universidade Federal Fluminense, filtrados da base do INEP. Os autores relatam que alcançaram um *recall* de 0,71, usando esta como a principal métrica de avaliação do algoritmo. Além disso, eles descrevem que fatores geográficos são importantes para prever a evasão, e que a escolha de um segundo curso no SiSu (Sistema de Seleção Unificada), o principal método de ingresso nas instituições públicas de ensino superior para os candidatos participantes do Exame Nacional do Ensino Médio (ENEM), pode ter relevância no cálculo do risco de evasão. O estudo desenvolvido por (Delen, 2011) alcança 0,81 de acurácia na previsão da evasão em uma base de dados contendo estudantes de uma universidade pública dos Estados Unidos. A base utilizada continha informações relativas a características acadêmicas, financeiras e demográficas dos estudantes. O estudo concluiu que as variáveis acadêmicas e financeiras foram as mais importantes para realizar a previsão de evasão. Outros estudos procuram construir classificadores eficazes para a evasão em bases e cursos

específicos, mas não apresentam análises sobre as variáveis mais determinantes para sua previsão (Sales, Balby, & Cajueiro, 2016), (Rigo, Cambruzzi, Barbosa, & Cazella, 2014), (Manhães, Cruz, Costa, Zavaleta, & Zimbrão, 2011), (Sarker, Tiropanis, & Davis, 2014).

O problema da evasão também já foi estudado por outras áreas, como a Psicologia. O estudo realizado por (Ambiel, 2015) utilizou um modelo psicométrico para analisar dados extraídos de uma pesquisa realizada com alunos e ex-alunos de instituições públicas e particulares. Os resultados apontam que os principais aspectos psicológicos que motivam a decisão podem ser descritos por 7 categorias: motivos institucionais, pessoais, relacionados à falta de suporte, relacionados à carreira, relacionados ao desempenho acadêmico, interpessoais e relacionados à autonomia. Na discussão final, o autor aponta a necessidade de estudos que ampliem o conhecimento científico sobre as variáveis que podem causar evasão do Ensino Superior.

Muitos trabalhos buscam alcançar este objetivo, mas o tema não possui uma resposta única consolidada. O trabalho de (Meedech, Iam-On, & Boongoen, 2016) utiliza uma base com dados de desempenho acadêmico de estudantes tanto na universidade, quanto em seus históricos na educação básica, e indica que o número de notas baixas no primeiro semestre universitário é de grande importância. O trabalho desenvolvido por (Manrique, Casanova, Nunes, Nurmikko-Fuller, & Marino, 2019) também faz uso de dados de performance acadêmica, obtidos a cada período em uma base com 2.175 estudantes dos cursos de Administração e Arquitetura em uma universidade brasileira. Dados de desempenho acadêmico também são valorizados pelo trabalho de (Zhang, Oussena, Clark, & Kim, 2010), que utiliza uma base de dados da Thames Valley University, em Londres. Já o estudo de (Araque, Roldán, & Salguero, 2009) diz que outras características, como forma de ingresso, variáveis culturais, sociais e o próprio curso escolhido, possuem importância em um modelo de predição de evasão. O trabalho de (daCosta, SouzaBispo, & Pereira, 2018) busca gerar mais evidências para verificar hipóteses com respeito à evasão. Para isso, utilizam uma base de dados da Universidade Federal da Paraíba sobre os alunos do curso de Administração, com o intuito de investigar as principais variáveis que tem relação com a evasão e a permanência dos estudantes. É apontado que, na universidade em questão, o risco de evasão é consistentemente maior para o sexo masculino do que para o sexo feminino. Por outro lado, o estudo desenvolvido por (Bonaldo & Pereira, 2016), realizado com o uso dados de universidades privadas do sudeste brasileiro, indicou que o gênero do estudante não foi uma variável significativa na previsão da evasão. Também foi relatado que a idade é um fator determinante, indicando que quanto mais jovem o estudante, menor a chance de evadir.

Como a maior parte dos estudos é restrita a uma base de dados específica de um curso ou universidade, espera-se que o presente trabalho contribua para um enriquecimento dos estudos do tema e revele uma perspectiva geral sobre os principais fatores relacionados à evasão em universidades públicas no Brasil.

3 Metodologia

O INEP disponibiliza, anualmente, uma massiva quantidade de dados relativos a todos os cursos de graduação do país (INEP, 2019). Esses dados, de domínio público, reúnem características de alunos, docentes, cursos e instituições de ensino que vão, desde atributos acadêmicos, como área dos cursos, até particularidades de infraestrutura, como a presença de laboratórios. O procedimento adotado no presente trabalho envolve o processamento destes dados e a aplicação de técnicas computacionais que permitam a previsão da probabilidade de evasão de um estudante. Um resumo da metodologia adotada no presente trabalho pode ser visto na Figura 1. Todas as implementações deste trabalho foram desenvolvidas com o uso da linguagem Python e das ferramentas disponíveis no pacote *scikit-learn* (Pedregosa, et al., 2011).

Inicialmente, foi realizado um pré-processamento dos dados do INEP, de forma a organizar a estrutura das informações para a aplicação das técnicas escolhidas. Foram removidos dados faltantes e desconsideradas as características irrelevantes para o estudo em questão, como dados de baixa variância ou relativos a instituições privadas ou de nível médio. Após uma etapa de balanceamento, a base foi separada em duas partes: dados para treinamento e dados para testes. A base de treinamento, que representa os dados efetivamente usados no processo de aprendizagem, foi composta por 75% dos dados, selecionados aleatoriamente. A base de testes foi composta pelos 25% restantes. Ela foi usada como dados novos, nunca vistos pelo sistema.

As técnicas de Aprendizado de Máquina foram aplicadas na base de treinamento, para que os melhores valores para os parâmetros de configuração de cada método fossem identificados. Os modelos treinados foram testados na base de testes e algumas métricas como acurácia e *recall* foram contabilizadas para medir a eficácia de cada técnica. Após a identificação dos melhores parâmetros, o melhor resultado de cada técnica foi legitimado por uso da validação cruzada. Por fim, uma análise qualitativa foi feita, a partir da comparação dos resultados obtidos com estudos científicos conceituais sobre a evasão escolar existentes na literatura (Filho, Motejunas, Hipólito, & Lobo, 2007), (Bastos & Gomes, 2016), (Silva & Imran, 2015).

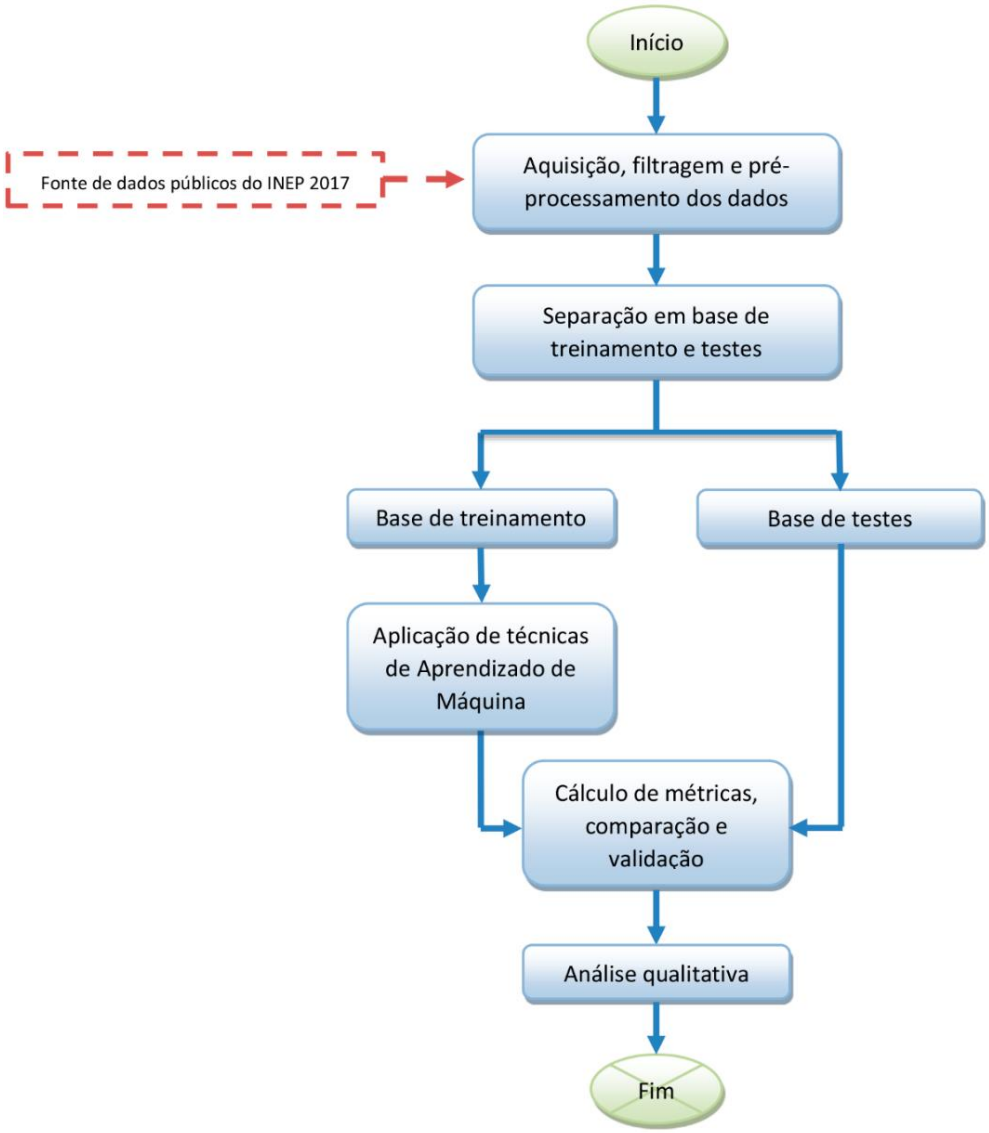

Figura 1: Fluxograma da metodologia.

3.1 Obtenção e Preparação dos Dados

Para a montagem da base de dados utilizada no presente trabalho, foi necessário explorar as relações entre as tabelas DM_ALUNO (dados de todos os alunos que já se matricularam em instituições de ensino superior), DM_CURSO (dados de todos os cursos superiores cadastrados no MEC), DM_DOCENTE (dados de todos os docentes das instituições de ensino superior), DM_IES (dados de todas as instituições de ensino superior), DM_LOCAL_OFERTA (dados de todos os locais de oferta dos cursos superiores) e TB_AUX_AREA_OCDE (dados de todas as áreas relativas aos cursos superiores, de acordo com a classificação internacional Eurostat/ Unesco/ OCDE), que são disponibilizadas pelo INEP. A relação entre estas tabelas pode ser vista na Figura 2, na forma de um diagrama entidade-relacionamento resumido, envolvendo apenas os atributos que identificam as ligações entre as tabelas.

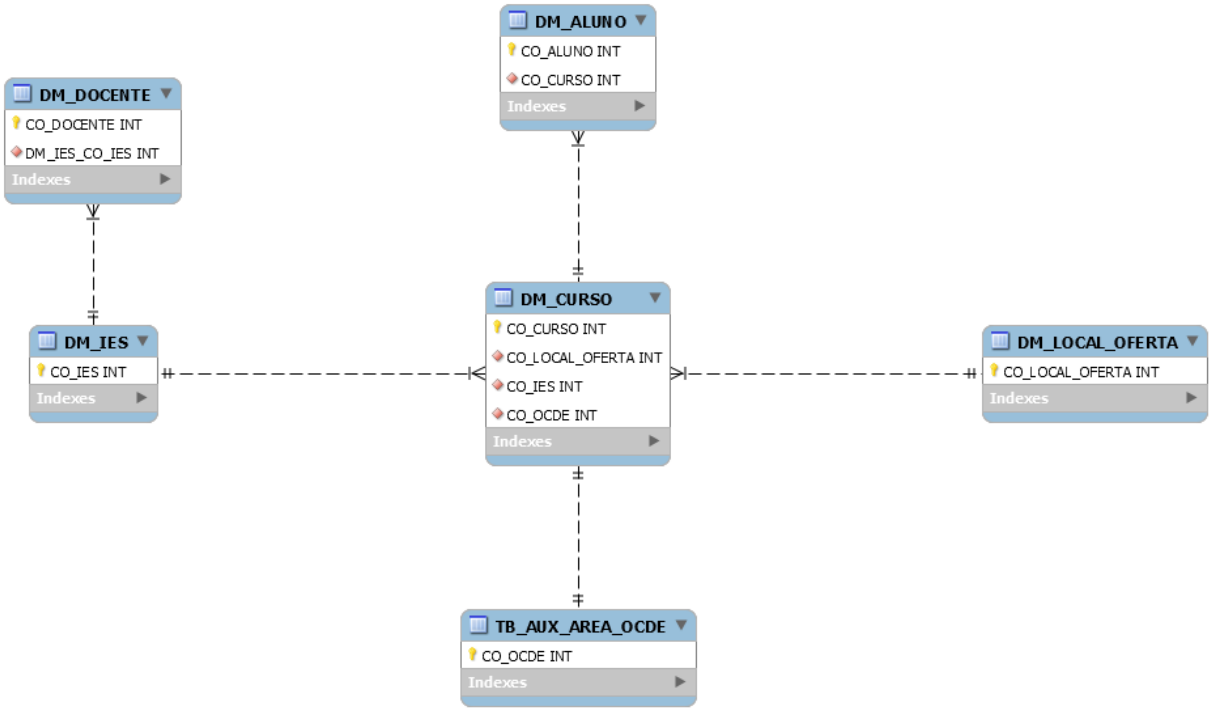

Figura 2: Diagrama entidade-relacionamento dos microdados disponibilizados pelo INEP.

Com o objetivo de investigar e identificar padrões nas características globais de estudantes de graduação, de forma a possibilitar a previsão do risco de um aluno desistir de seus estudos em uma instituição de ensino superior pública, o escopo deste trabalho envolve apenas os dados que representam alunos de instituições públicas que falharam ou tiveram sucesso em sua formação. Foram descartados, por exemplo, dados referentes a alunos de instituições privadas e de alunos que foram desligados de seus cursos por falecimento.

Para respeitar tais restrições, os dados foram filtrados e associados de forma que uma tabela única fosse construída. Esta nova tabela tem como finalidade unificar o máximo de dados possível quanto aos alunos de interesse da pesquisa. Primeiro, os dados completos dos alunos em questão, armazenados na tabela DM_ALUNO, foram filtrados de acordo com os atributos correspondentes ao código da categoria administrativa da instituição de ensino do aluno (TP_CATEGORIA_ADMINISTRATIVA) e ao código referente à situação na qual o aluno se encontra (TP_SITUACAO). A categoria administrativa da instituição é uma variável categórica nominal, que define a instituição como pública federal, pública estadual, pública municipal, privada com fins lucrativos, privada sem fins lucrativos, privada confessional e especial. Foram filtrados os registros correspondentes apenas a instituições públicas. Já a situação na qual o aluno se encontra, também uma variável categórica nominal, pode assumir os seguintes valores:

"cursando", "matrícula trancada", "desvinculado do curso", "transferido para outro curso da mesma IES", "formado" ou "falecido". Os registros referentes aos alunos que estão com o curso em andamento, foram transferidos para outro curso da mesma IES, ou faleceram, foram removidos da base utilizada, por serem irrelevantes ao estudo. Os casos em que o aluno teve a matrícula trancada ou foi desvinculado do curso foram considerados casos de falha na formação, caracterizando evasão, enquanto os alunos que se encontram formados foram considerados casos de sucesso. Com isso, o atributo TP_SITUACAO foi utilizado como a variável classe no problema.

Em seguida, os dados completos das instituições citadas, envolvendo dados financeiros e estruturais armazenados na tabela DM_IES, também foram filtrados pelo atributo TP_CATEGORIA_ADMINISTRATIVA, para que fossem consideradas apenas as instituições públicas. A partir da tabela DM_DOCENTE, foram calculados vários novos atributos para compor a nova tabela unificada das instituições de ensino. Com isso, foi possível identificar características referentes aos docentes de cada instituição como, por exemplo, a idade média dos docentes, o percentual de professores doutores e o total de professores substitutos presentes no curso.

Além destas informações, os dados relativos à área dos cursos foram agregados aos dados dos alunos e os dados dos locais de oferta incorporados aos dados dos cursos. Essas últimas estruturas foram associadas aos elementos relativos aos discentes dos cursos. Por fim, todas as características foram agregadas para formar a nova estrutura, utilizada no trabalho, que fornece, em cada registro, todas as informações disponibilizadas pelo INEP referentes a cada aluno e seu local de estudo. Considerando estes registros, a base de dados possuía 64% de alunos caracterizados como evasão, ou seja, um desbalanceamento. Um dos obstáculos mais comuns em problemas de classificação é o desbalanceamento da base de dados (Fernández, Galar, & Krawczyk, 2018). Para suprimir este problema, foi adotada a estratégia de balanceamento por *undersampling* aleatório. Esta estratégia foi escolhida por ter se mostrado eficaz em bases com um número suficientemente grande de registros (Fernández, Galar, & Krawczyk, 2018). O total de registros presentes na base final balanceada é de 376.746, número que corresponde ao total de alunos considerados neste estudo. Nesta base, foi realizada ainda uma etapa simples de seleção de variáveis, em que foram removidas colunas que apresentavam muitos valores nulos, inválidos ou com baixa variância. No total, a base construída e utilizada no presente trabalho ficou com 160 variáveis previsoras, correspondentes às características, ou atributos, dos alunos, e uma única variável de classe, que pode possuir dois valores: sucesso ou evasão.

3.2 Pré-processamento dos Dados

Nesta etapa, os dados foram tratados de forma que os algoritmos obtivessem o melhor resultado possível. Para isso, duas técnicas principais foram utilizadas: a transformação de valores não numéricos para numéricos e a padronização dos dados.

3.2.1 Codificação dos Valores

A codificação de valores categóricos para numéricos faz-se necessária porque, frequentemente, as técnicas de aprendizado de máquina envolvem cálculos numéricos como distâncias e probabilidades, o que se torna impraticável se os dados representam uma característica não quantitativa.

Com o propósito de realizar esta transformação, os métodos da classe *LabelEncoder*, disponível na biblioteca *scikit-learn*, foram aplicados. Esta ferramenta faz a codificação dos dados em valores pertencentes ao intervalo [0, v], em que v é o número de categorias existentes na base para o atributo em questão, decrescido de 1.

3.2.2 Padronização

Quando o problema abordado possui muitas variáveis é comum que suas escalas sejam diferentes. Isso acontece por conta das diferentes grandezas que representam características normalmente não comparáveis, como renda e idade. Porém, para os algoritmos de aprendizado de máquina, é desejável que as variáveis previsoras estejam em escalas similares.

Com o intuito de melhor extrair a importância de cada característica na decisão da classe, um algoritmo básico de padronização foi aplicado aos dados, para colocá-los em uma mesma escala. A Equação 1 mostra o cálculo usado para a padronização de uma variável x, que leva em conta o quão diferente um dado é de sua média. Nesse cálculo, considera-se z a variável padronizada, conforme a Equação 1.

$$z = \frac{x - \bar{x}}{S} \qquad (1)$$

onde x é uma variável cuja média é $\bar{x}$ e o desvio padrão é S.

3.3 Técnicas de Aprendizado de Máquina

Neste trabalho, foram aplicadas, no total, cinco diferentes técnicas de Aprendizado de Máquina: Naive Bayes, K-Nearest Neighbors, Árvores de Decisão, Random Forest e Redes Neurais. Como o objetivo principal deste trabalho é a identificação das variáveis mais determinantes na classificação de um estudante como evasão ou sucesso, a utilização de diferentes técnicas foi necessária para que o melhor resultado possível para a classificação fosse alcançado. Ressalta-se, ainda, que o presente trabalho não procura comparar as técnicas em si, mas, apenas, encontrar o melhor resultado para o problema abordado. Estas técnicas foram escolhidas por sua ampla utilização em problemas de classificação. Para cada uma das técnicas, diferentes configurações de parâmetros foram testadas a fim de melhorar os resultados e reduzir a ocorrência de *overfitting*.

O algoritmo Naive Bayes (NB) é um classificador probabilístico que usa o teorema de Bayes para calcular suas estimativas. Esse método pressupõe total independência entre as características dos dados (Downey, 2012). Tal suposição raramente coincide com a realidade, porém, ainda assim, o algoritmo tem uma boa performance em diversos casos de aprendizado de máquina supervisionado (Granik & Mesyura, 2017).

A técnica k-*Nearest Neighbors* (kNN) admite que registros com a mesma classe possuem características similares. Para classificar os registros, a distância entre o registro novo e os demais é calculada. Geralmente, a técnica alcança bons resultados, mas pode apresentar um custo computacional considerável (Bruce & Bruce, 2017).

As Árvores de Decisão podem ser interpretadas como disjunções de conjunções (Mitchell, 1997). O algoritmo utilizado para a construção das árvores leva em consideração a importância dos atributos associando o mais importante ao nodo raiz, o início da árvore (Müller & Guido, 2017). O cálculo da importância das variáveis, neste contexto, pode ser entendido como o ganho de informação, que mede o quão bom um atributo é em separar as instâncias do conjunto de treinamento de acordo com suas classes (Mitchell, 1997).

O método *Random Forest*, em sua abordagem quanto um classificador, gera um modelo que é basicamente um agrupamento de Árvores de Decisão, onde cada árvore possui características próprias (Breiman, 2001). A abordagem usada no presente trabalho é a descrita em (Müller & Guido, 2017).

A denominação Redes Neurais é atribuída a uma classe de algoritmos que configuram um ramo da Inteligência Artificial (Gardner & Dorling, 1998), (Lerner, et al., 1994). Um dos componentes dessa família de algoritmos são as redes *Multilayer Perceptrons* (MLP), que se baseiam em uma série de camadas compostas por neurônios: a camada de entrada, as camadas

escondidas e a camada de saída. As redes MLP são consideradas um tipo relativamente simples de Rede Neural e, de acordo com (Müller & Guido, 2017), podem ser interpretadas como "generalizações de modelos lineares que executam vários estágios de processamento para chegar a uma decisão".

3.4 Métricas de Avaliação e Validação

Para a obtenção de resultados com valor estatístico, as métricas foram avaliadas através da aplicação de uma validação cruzada. Essa técnica realiza o treinamento dos modelos em vários subconjuntos dos dados, assim, revela se há ou não um padrão para ser generalizado pelos algoritmos de Aprendizado de Máquina.

A versão da técnica de validação utilizada neste trabalho é uma variação da técnica k-Fold, identificada como *Stratified k-Fold Cross-Validation* (Müller & Guido, 2017). Ambas as técnicas geram k grupos (*splits*) a partir da base de dados original. Em seguida, em cada *split* é realizada uma seleção de um grupo que atuará como dados de teste, enquanto o resto será utilizado para o treinamento do modelo. Por fim, calcula-se o resultado a partir da média simples entre os resultados obtidos em cada *split*. A principal vantagem da versão *Stratified* é que, ao invés de selecionar aleatoriamente quais são os registros que farão parte de cada grupo, a seleção é feita de maneira que a proporção de classes do conjunto original seja mantida em cada *split*. Dessa forma, cada *split* fica com o mesmo balanceamento que o conjunto de dados original. No presente trabalho, foi feita a divisão em 5 *splits*.

Para avaliar os resultados obtidos pelos classificadores treinados e identificar aquele que apresentou os melhores resultados na previsão de evasão, a Matriz de Confusão foi calculada para cada caso. A Matriz de Confusão separa as previsões em quatro valores: verdadeiros positivos (VP), verdadeiros negativos (VN), falsos positivos (FP) e falsos negativos (FN), como mostra a Equação 2.

$$MC = \begin{pmatrix} VP & FN \\ FP & VN \end{pmatrix} \quad (2)$$

A terminologia "positivo", no caso do presente artigo, representa um caso de evasão. Os verdadeiros positivos são os valores que realmente correspondem a uma evasão e foram previstos como tal. Os falsos positivos são os valores que correspondem a uma não-evasão e foram previstos como evasão. Os verdadeiros negativos são os valores que correspondem a uma não-evasão e foram previstos como tal. Os falsos negativos são os valores que correspondem a uma evasão, mas foram previstos como não-evasão. A partir da Matriz de Confusão, foram calculadas as métricas detalhadas a seguir.

3.4.1 Acurácia

A acurácia representa a taxa de acertos em relação ao total de amostras. A Equação 3 mostra como calcular a acurácia.

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \quad (3)$$

3.4.2 Precisão

Essa métrica é particularmente útil em cenários cujo foco é limitar o número de falsos positivos. Seu cálculo revela quantos dos dados classificados como positivos realmente são positivos. A Equação 4 mostra como calcular a precisão.

$$Precisão = \frac{VP}{VP + FP} \quad (4)$$

3.4.3 Recall

Também conhecida como "Taxa de Verdadeiros Positivos", é geralmente utilizada em contextos nos quais o objetivo é identificar todos os positivos. O resultado mostra quantos dos dados positivos foram capturados pelas previsões positivas. No contexto deste trabalho, esta é uma métrica importante, tendo em vista que é mais problemático deixar de oferecer planos de auxílio aos alunos que provavelmente evadirão do que oferecê-los aos que vão ter sucesso. A Equação 5 mostra como calcular o *recall*.

$$Recall = \frac{VP}{VP + FN} \quad (5)$$

3.4.4 F1-Score

Quando um balanço entre a precisão e o recall é necessário, a métrica F1-Score é usada. Ela representa a média harmônica entre essas duas medidas. Além disso, proporciona uma boa capacidade de avaliação em casos de classificação binária. É adequada em situações que precisam da análise de todo o quadro com a mesma importância e não apenas de um de seus aspectos. A Equação 6 mostra como calcular o *F1-Score*.

$$F = 2 * \frac{Precisão * Recall}{Precisão + Recall} \quad (6)$$

3.4.5 Curva ROC e AUC

Este indicador possibilita uma avaliação mais completa quanto à qualidade das predições do classificador, levando em consideração a taxa de verdadeiros positivos e a taxa de falsos positivos (Müller & Guido, 2017). A curva ROC (*Receiver Operating Characteristics*) ilustra o limiar da capacidade de discriminação de um classificador binário, variando o critério de aceitação de um classificador. O gráfico da curva representa a variação da Taxa de Verdadeiros Positivos (TVP), dada pela Equação 7, em relação à Taxa de Falsos Positivos (TFP), dada pela Equação 8.

$$TVP = \frac{VP}{VP + FN} \quad (7)$$

$$TFP = \frac{FP}{FP + VN} \quad (8)$$

Para resumir a qualidade mensurada pela curva, é comum a utilização da métrica AUC (*Area Under the Curve*). Com isso, é possível comparar os classificadores utilizando um único escalar, a área abaixo da curva ROC.

4 Resultados e Discussões

A utilização de diferentes técnicas de classificação teve como objetivo alcançar o melhor resultado possível para a previsão de evasão. Analisando o melhor classificador, torna-se possível identificar as variáveis com maior importância em seu funcionamento interno, ou seja, as mais determinantes na identificação da evasão.

4.1 Desempenho dos Classificadores

Cada método de Aprendizado de Máquina utilizado no presente trabalho possui uma série de parâmetros de configuração. Estes fatores podem influenciar diretamente os resultados das técnicas, podendo causar problemas como *overfitting* ou *underfitting*. Os valores escolhidos para os parâmetros foram os melhores encontrados a partir de uma série de testes individuais, levando em conta os resultados tanto na base de testes quanto na de treinamento, para evitar os problemas mencionados.

A Tabela 1 mostra quais foram as melhores configurações de métodos e tratamentos de dados que obtiveram os melhores resultados no presente trabalho.

Tabela 1: Parâmetros de configuração das técnicas utilizadas.

Técnica	**Configuração**
Naive Bayes	-
KNN	K = 10
Árvores de Decisão	Profundidade máxima = 18
Random Forest	Profundidade máxima = 22 Número de características usadas = 70 Número de árvores = 70
Redes Neurais	Neurônios na camada de entrada: 100 Neurônios na camada escondida: 40 Neurônios na camada de saída: 20

A Tabela 2 mostra as Matrizes de Confusão médias resultantes do processo de validação cruzada. A taxa de incerteza correspondente ao desvio padrão obtido após a aplicação da validação cruzada também é mostrada, juntamente com os valores médios referentes aos resultados obtidos. Cada elemento da matriz está no formato: [valor médio] ± [desvio padrão]. Quanto maior o desvio padrão, menos confiança podemos ter na generalização alcançada pela técnica para casos diferentes. Os resultados médios obtidos por cada uma das técnicas, calculados a partir das Matrizes de Confusão, podem ser encontrados na Figura 3. Nela, o desvio padrão obtido para cada métrica é representado por barras de erro. Além disso, na Figura 4 estão dispostas as curvas ROC médias obtidas para cada método. É possível observar que todas as técnicas obtiveram níveis de incerteza relativamente baixos, o que fortalece as previsões realizadas pelos modelos treinados.

Tabela 2: Matrizes de Confusão obtidas pela validação cruzada de cada técnica.

Técnica	**Matriz de Confusão**
Naive Bayes	$\begin{pmatrix} 34806{,}8 \pm 224{,}1 & 2867{,}8 \pm 178{,}0 \\ 26017{,}2 \pm 200{,}5 & 11657{,}4 \pm 177{,}1 \end{pmatrix}$
KNN	$\begin{pmatrix} 29005{,}6 \pm 112{,}0 & 8669{,}0 \pm 83{,}2 \\ 10746{,}2 \pm 77{,}0 & 26928{,}4 \pm 102{,}1 \end{pmatrix}$
Árvores de Decisão	$\begin{pmatrix} 27703{,}0 \pm 258{,}5 & 9971{,}6 \pm 321{,}4 \\ 8165{,}0 \pm 203{,}1 & 29509{,}6 \pm 130{,}8 \end{pmatrix}$
Random Forest	$\begin{pmatrix} 28910{,}0 \pm 83{,}4 & 8764{,}6 \pm 89{,}2 \\ 6898{,}4 \pm 103{,}2 & 30776{,}2 \pm 95{,}6 \end{pmatrix}$
Redes Neurais	$\begin{pmatrix} 28713{,}4 \pm 121{,}8 & 8961{,}2 \pm 47{,}1 \\ 7606{,}0 \pm 88{,}4 & 30068{,}6 \pm 128{,}2 \end{pmatrix}$

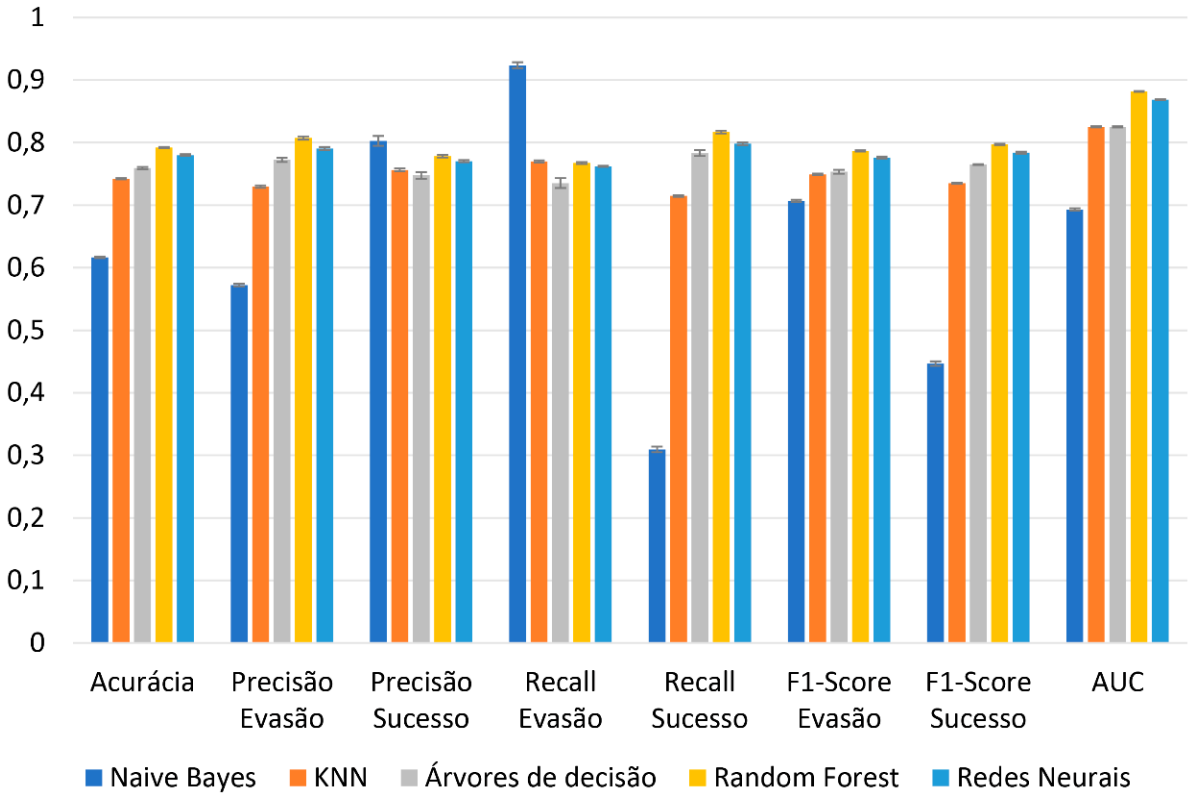

Figura 3: Resultados obtidos pela validação cruzada de cada técnica.

Ao analisar as métricas mostradas na Figura 3, infere-se que o método *Naive Bayes* foi capaz de fazer as melhores previsões de verdadeiros positivos, ou seja, dos casos nos quais o aluno provavelmente irá evadir. Por outro lado, dentre todas as técnicas aplicadas, este método apresentou os piores resultados com relação a previsão de sucesso do aluno. Como consequência, percebe-se que o AUC médio obtido para o Naive Bayes foi o mais baixo, assim como o F1-Score. Ou seja, o método não é eficiente quando levamos em conta a proporção entre previsões corretas e incorretas de evasão. Além disso, este algoritmo não revela diretamente quais características foram importantes para gerar as previsões e, portanto, não obtém informações úteis para a criação de planos de mitigação.

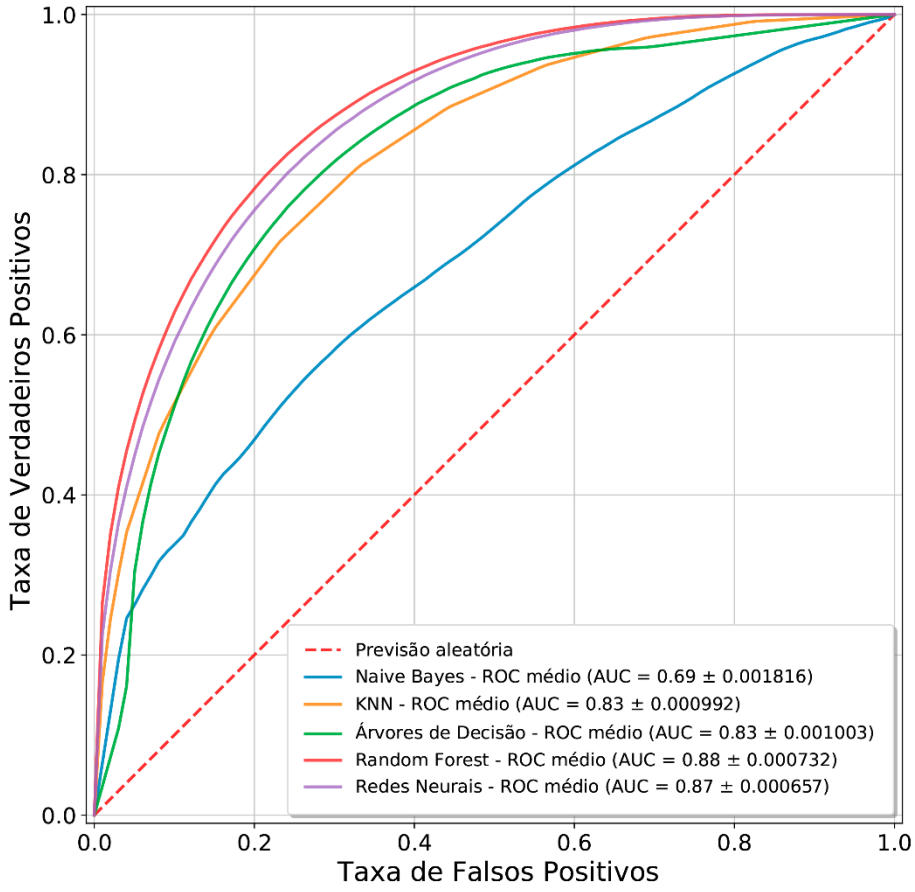

Figura 4: Curvas ROC médias para cada classificador.

Mesmo com bons resultados para o valor de *recall* em relação as evasões, o método kNN obteve resultados medianos para a precisão e acurácia, alcançando um nível de resultados similar ao das Árvores de Decisão em AUC. Porém, o processo para a realização de uma previsão através do kNN é de grande custo computacional, pois a inserção de um novo registro na base gera a necessidade do cálculo da distância para todos os outros registros. Como a base utilizada possui muitos registros, a falta de escalabilidade da técnica se fez presente. Além disso, a qualidade do resultado possui alto grau de dependência com o parâmetro k.

Como esperado, os resultados das Árvores de Decisão foram bons, mas inferiores aos do *Random Forest*. A grande vantagem destas técnicas é que o modelo criado permite a avaliação direta das características mais importantes para a classificação de um estudante. Por ser um algoritmo mais simples, a construção de uma única Árvore de Decisão apresenta um custo computacional menor que o do *Random Forest*, mas um resultado inferior. O valor médio de AUC obtido pelas Árvores de Decisão foi tão alto quando o kNN.

Já o modelo treinado pelas Redes Neurais obteve resultados próximos dos melhores obtidos no presente trabalho. Além disso, se mostrou uma das técnicas com maior capacidade de generalização, tendo em vista que seus valores de desvio padrão foram os mais baixos. Apesar de alcançar resultados excelentes para a previsão assertiva da evasão, como mostra o AUC obtido pela técnica, as Redes Neurais não fornecem facilmente informações utilizáveis para a compreensão das causas relacionadas à evasão.

Por fim, o modelo construído pelo *Random Forest* foi o que obteve o melhor desempenho em geral. Como mostra sua curva ROC, este classificador é o que permite a escolha de um critério mais alto de aceitação. Mesmo com um resultado para *recall* um pouco mais baixo que o do *Naive Bayes*, por exemplo, o método se mostrou mais estável e confiável levando em consideração todas as outras métricas, que mostram eficácia tanto em suas previsões de evasão quanto de sucesso.

Além disso, esse classificador ainda se adequa bem ao problema, pois fornece dados que tornam possível a geração de um relatório sobre quais atributos foram os mais importantes para as suas previsões. Por estes motivos, é a técnica cujos resultados foram os mais importantes para esse trabalho. A importância de um atributo, neste caso, pode ser avaliada de acordo com o ganho de informação total normalizado produzido pela característica em questão. Dessa forma, a soma total das importâncias de todos os atributos, em valor percentual, é igual a 100%.

A Figura 5 mostra os 30 atributos mais importantes para a previsão de evasão para um estudante em uma instituição pública de ensino superior no Brasil. Juntos, os 30 atributos chegam a 78,87% de importância, de acordo com o modelo construído pela técnica *Random Forest*. As previsões da técnica alcançaram um AUC de 0,88, além de uma acurácia média de 0,79. Isso significa que, quando o modelo treinado classifica um aluno como provável evasão ou sucesso, ele tem aproximadamente 80% de chance de estar correto. Pode-se dizer, então, que os atributos listados na Figura 5 são os principais responsáveis pela previsão de evasão de um aluno. A Tabela 3 mostra a descrição de cada um dos 30 atributos identificados como mais importantes.

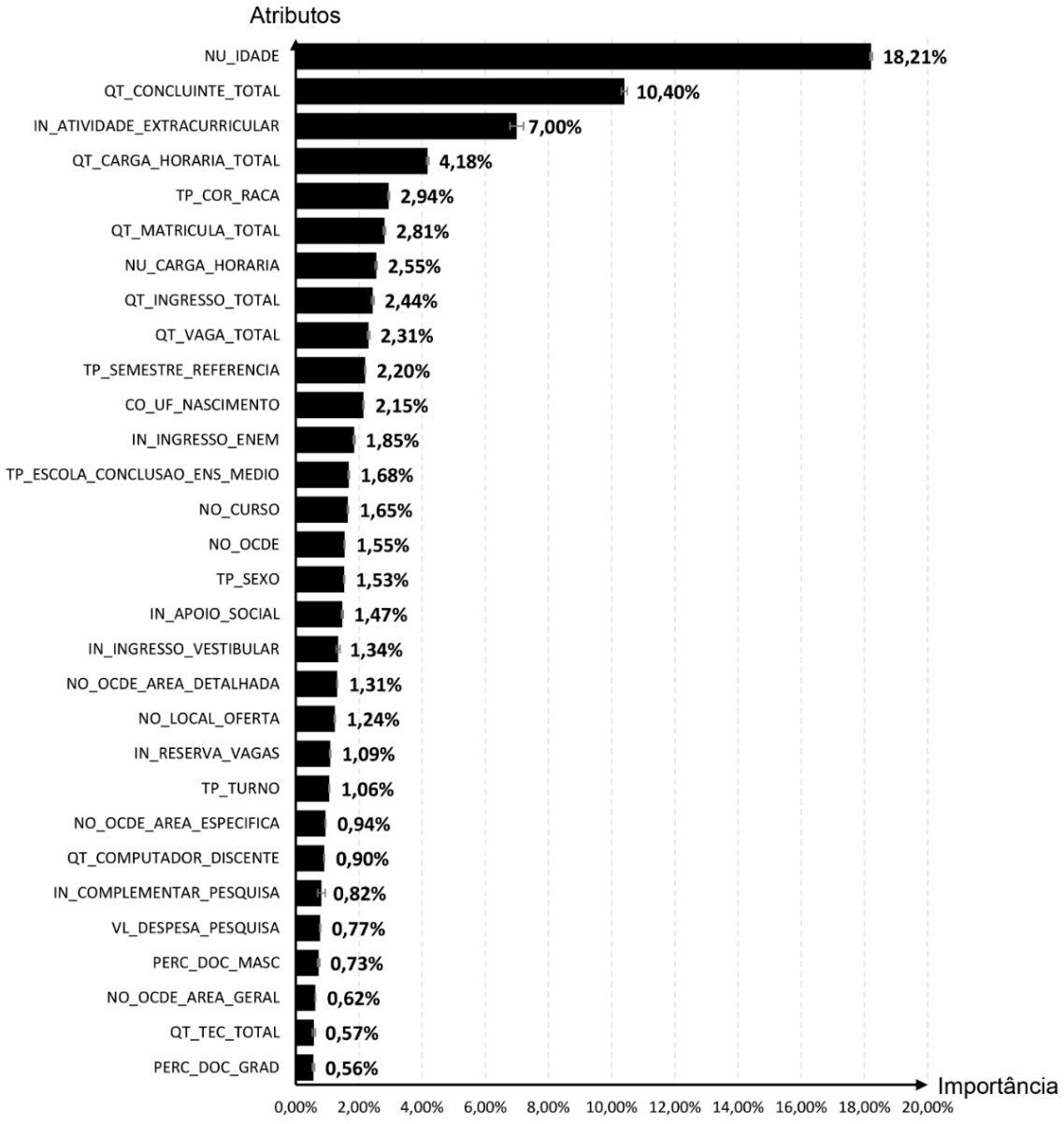

Figura 5: Lista ordenada dos valores médios das importâncias dos trinta atributos considerados mais determinantes para a previsão de evasão pelo classificador Random Forest.

Tabela 3: Descrição das trinta variáveis correspondentes aos atributos mais determinantes para a previsão de evasão.

Identificador	Descrição
NU_IDADE	Idade que o aluno completou no ano de referência do Censo.
QT_CONCLUINTE_TOTAL	Número total de concluintes no curso.
IN_ATIVIDADE_EXTRACURRICULAR	Participação em algum tipo de atividade extracurricular. (0 - não, 1 - sim)
QT_CARGA_HORARIA_TOTAL	Total da carga horária dos componentes curriculares do curso.
TP_COR_RACA	Código que identifica a cor/raça do aluno.
QT_MATRICULA_TOTAL	Número total de matrículas no curso.
NU_CARGA_HORARIA	Carga horária mínima do curso.
QT_INGRESSO_TOTAL	Número de ingressantes no curso no ano de referência.
QT_VAGA_TOTAL	Quantidade de vagas totais oferecidas no curso.
TP_SEMESTRE_REFERENCIA	Semestre de referência do preenchimento do vínculo do curso.
CO_UF_NASCIMENTO	Código da Unidade da Federação de nascimento do aluno.
IN_INGRESSO_ENEM	Informa se o aluno ingressou no curso pelo ENEM. (0 - não, 1 - sim)
TP_ESCOLA_CONCLUSAO_ENS_MEDIO	Tipo de escola que o aluno concluiu ensino médio.
NO_CURSO	Nome do curso.
NO_OCDE	Nome de identificação do curso de acordo com a OCDE.
TP_SEXO	Código que informa o sexo do aluno.
IN_APOIO_SOCIAL	Informa se o aluno recebe algum tipo de apoio social na forma de moradia, transporte, alimentação, material didático e bolsas. (0 - não, 1 - sim)
IN_INGRESSO_VESTIBULAR	Informa se o aluno ingressou no curso por vestibular. (0 - não, 1 - sim)
NO_OCDE_AREA_DETALHADA	Nome da área detalhada do curso de acordo com a OCDE.
NO_LOCAL_OFERTA	Nome do local de oferta definido pela instituição
IN_RESERVA_VAGAS	Informa se o aluno participa de programa de reserva de vagas.
TP_TURNO	Código do turno do curso ao qual o aluno está vinculado.
NO_OCDE_AREA_ESPECIFICA	Nome da área específica conforme padrão Eurostat/Unesco/OCDE.
QT_COMPUTADOR_DISCENTE	Informa a quantidade de computadores para uso dos discentes.
IN_COMPLEMENTAR_PESQUISA	Informa se o aluno participa de atividade extracurricular de pesquisa. (0 - não, 1 - sim)
VL_DESPESA_PESQUISA	Informa as despesas com Pesquisa e Desenvolvimento da IES.
PERC_DOC_MASC	Percentual de docentes do sexo masculino.

NO_OCDE_AREA_GERAL	Nome da área geral conforme adaptação da classificação internacional Eurostat/ Unesco/ OCDE
QT_TEC_TOTAL	Número de funcionários técnico-administrativos.
PERC_DOC_GRAD	Percentual de docentes com graduação (maior grau).

4.2 Análise Exploratória das Variáveis Identificadas

Analisando os dados mostrados na Figura 5, é possível perceber que, apesar de alguns atributos pessoais dos alunos estarem presentes dentre os mais determinantes para a previsão da probabilidade de um aluno desistir de um curso de graduação, os atributos mais importantes, em sua maioria, são da própria instituição de ensino em que o aluno se matriculou.

Para analisar como as variáveis numéricas se relacionam com a taxa de evasão, a correlação de Pearson entre elas e a evasão dos estudantes foi calculada, gerando os resultados mostrados na Figura 6.

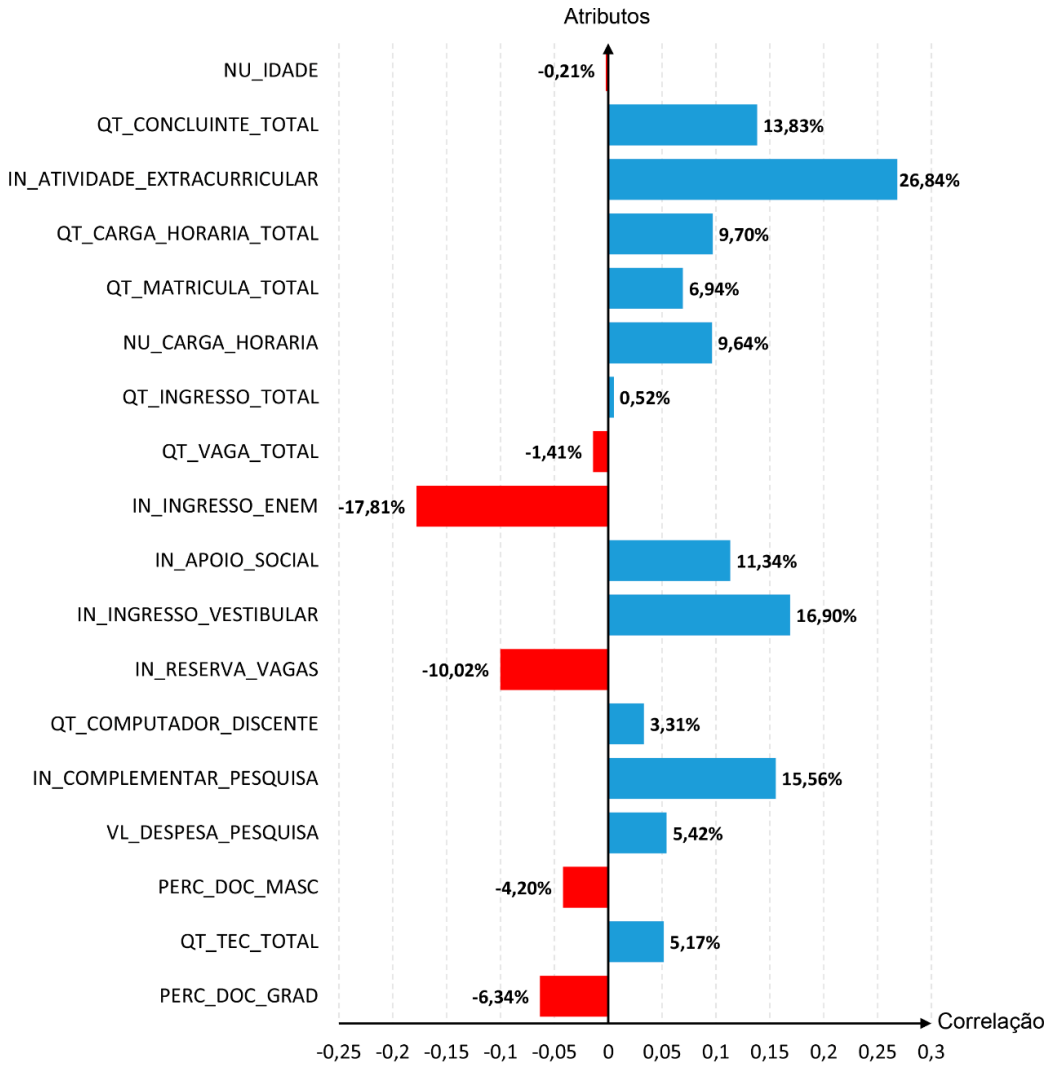

Figura 6: Correlação de Pearson entre as variáveis numéricas identificadas como mais importantes pelo classificador Random Forest e a variável de evasão dos alunos.

Os valores mostrados na Figura 6 mostram a correlação linear entre as variáveis numéricas e a evasão dos alunos. Como a variável de evasão pode assumir os valores ordinais 0 e 1,

representando “evasão” e “sucesso”, respectivamente, uma correlação positiva indica que a característica está relacionada com o sucesso do aluno, enquanto uma correlação negativa indica que a característica está relacionada com a evasão. Isso não significa que existe dependência ou causalidade direta entre as características e a evasão, mas ajuda a identificar os padrões característicos de alunos que possuem maior chance de evasão, e que devem ser alvos de políticas públicas de alguma forma de assistência. Por exemplo, um dos atributos determinantes, de acordo com a Figura 5, é a participação do aluno em atividades extracurriculares. A Figura 6 mostra uma forte correlação positiva desta característica com a variável de evasão, ou seja, pode-se dizer que se o aluno realiza atividades extracurriculares, sua probabilidade de obter sucesso na formação é maior. De acordo com o indicado na Figura 6, e como foi apontado por (Lamers, Santos, & Toassi, 2017), o fato do aluno se dedicar a atividades que vão além da sala de aula possui forte relação com sua probabilidade de conclusão do curso. Outra característica com forte correlação positiva é a quantidade de alunos que já concluíram com sucesso um curso. Este fato indica que um curso já consolidado, que possui um histórico favorável de formação de profissionais, irá oferecer maiores chances para um aluno se formar.

Os trabalhos de (Bastos & Gomes, 2016) e (Lamers, Santos, & Toassi, 2017) revelam que a desmotivação para o estudo devido a práticas tradicionais de ensino também possuem alto grau de correlação com as taxas de evasão. Estes resultados também foram corroborados pelo grau de importância dado, no presente trabalho, ao atributo relativo ao exercício de atividades extracurriculares e de pesquisa. Os resultados na Figura 6 mostram alguns atributos também ligados, possivelmente, à motivação dos estudantes, como a carga horária dos cursos, a quantidade total de alunos matriculados e quantos a universidade já foi capaz de formar. Além disso, foi exposto por (Filho, Motejunas, Hipólito, & Lobo, 2007) que a relação candidato vaga do curso é inversamente proporcional a taxa de evasão. Os resultados do presente trabalho demonstram, também, que a quantidade de vagas total oferecida por um curso tem correlação negativa com a variável de evasão de alunos.

Quanto às características dos docentes apontadas na Figura 5 como mais importantes, é possível observar que, levando em consideração a margem definida pelo desvio padrão, tanto o nível de escolaridade, quanto o gênero dos docentes, possuem importâncias próximas. De acordo com a correlação mostrada na Figura 6, é possível presumir que o investimento na formação docente e o incentivo à diversidade de gênero no quadro funcional de professores podem ser caminhos interessantes a serem seguidos. A quantidade total de técnicos administrativos na instituição em que o aluno irá ingressar, e o investimento total em pesquisa, também são exemplos de dados das instituições que entraram na lista das trinta características mais determinantes. Análises similares podem ser feitas com respeito à correlação das outras variáveis ordinais apresentadas na Figura 6.

A forma de ingresso também se mostrou uma variável importante para a previsão de evasão. De acordo com a Figura 6, alunos que ingressaram a partir do ENEM possuem mais chance de evasão que aqueles que ingressaram pelo vestibular. A forma de seleção utilizada pelo ENEM pode ser uma explicação para este fato, pois muitos alunos acabam sendo chamados inicialmente para uma segunda opção e, quando reclassificados para o curso de primeira opção, desistem da vaga anterior, caracterizando evasão (Martins, Carvalho, & Carvalho, 2017). A variável TP_SEMESTRE_REFERENCIA também foi apontada como uma das variáveis mais importantes para classificação da evasão, como indicava o trabalho de (Martins, Carvalho, & Carvalho, 2017).

É importante ressaltar que a correlação de Pearson revela apenas a correlação linear entre duas variáveis. Algumas vezes, as variáveis podem possuir uma correlação não-linear, que não é facilmente observada. Por exemplo, a variável NU_IDADE possui uma correlação muito baixa com a variável de evasão, mas foi apontada pelo classificador como variável de grande importância. Para compreender como a idade está realmente relacionada com a evasão, uma

análise um pouco mais detalhada se faz necessária. A Figura 7 mostra um *box-plot* da variável NU_IDADE, separada entre casos de sucesso ou evasão.

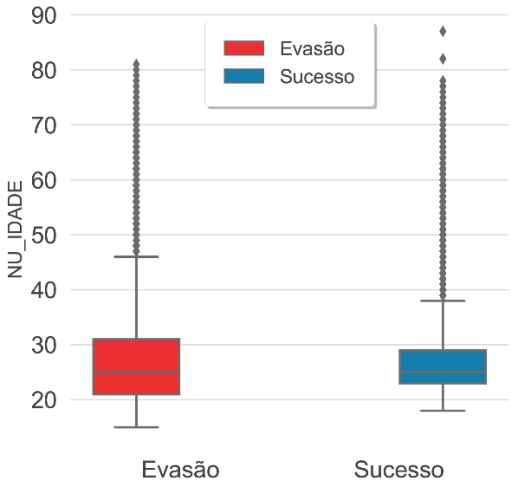

Figura 7: *Box-plot* da varável NU_IDADE, separada em casos de evasão e sucesso.

Percebe-se, ao analisar a Figura 7, que, realmente, não existe uma relação clara de correlação linear entre a idade dos alunos e a evasão, tendo em vista que a média da idade dos alunos que caracterizam evasão ou sucesso é bastante próxima. Porém, a amplitude reduzida dos quartis mostrados no *box-plot* dos casos de sucesso permite a observação de que a faixa de idade dos alunos que obtém sucesso está mais "concentrada" em torno da média que a dos alunos que caracterizam evasão. Investigando um pouco mais, podemos visualizar no histograma mostrado na Figura 8 que a idade dos alunos com evasão ou sucesso seguem distribuições de probabilidade diferentes. Este fato mostra a capacidade do classificador de identificar padrões normalmente difíceis de serem observados. Analisando a Figura 8, vemos que a probabilidade de evasão é maior em jovens com idade menor ou igual a 21 anos. Além disso, é interessante notar que, a partir dos 29 anos, a probabilidade de evasão se torna maior que a de sucesso. A maior parte dos alunos que obtém sucesso em sua formação estão dentro de uma faixa entre 22 e 28 anos.

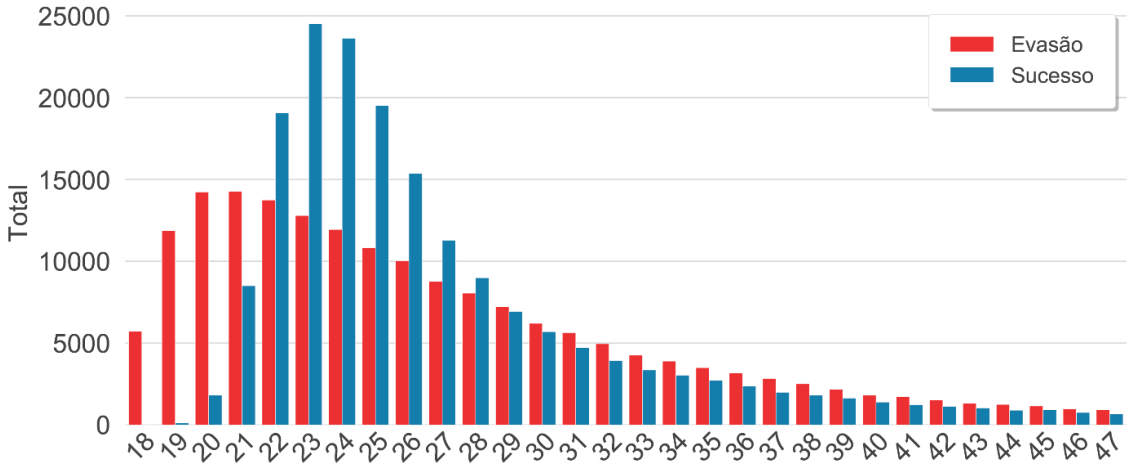

Figura 8: Histograma da variável NU_IDADE, separada em casos de evasão e sucesso.

Como as variáveis categóricas nominais não podem ser relacionadas com a evasão pela correlação de Pearson, torna-se necessária uma análise individual para entender a relação delas com a evasão. A presença, na Figura 5, de atributos que descrevem os cursos e suas áreas indica,

por exemplo, que diferentes cursos apresentam níveis de dificuldade diferentes e, com isso, maiores ou menores taxas de evasão, como é de se esperar. A Figura 9 ilustra esse fato, mostrando a proporção de casos de evasão e sucesso nos 20 cursos com maior número de alunos. A análise de (Filho, Motejunas, Hipólito, & Lobo, 2007) indica que a área de conhecimento associada ao curso escolhido pelo aluno realmente tem relação direta com a taxa de evasão, fato aqui corroborado pela presença das variáveis relativas ao tipo de curso e à sua área dentre as mais determinantes na previsão de evasão.

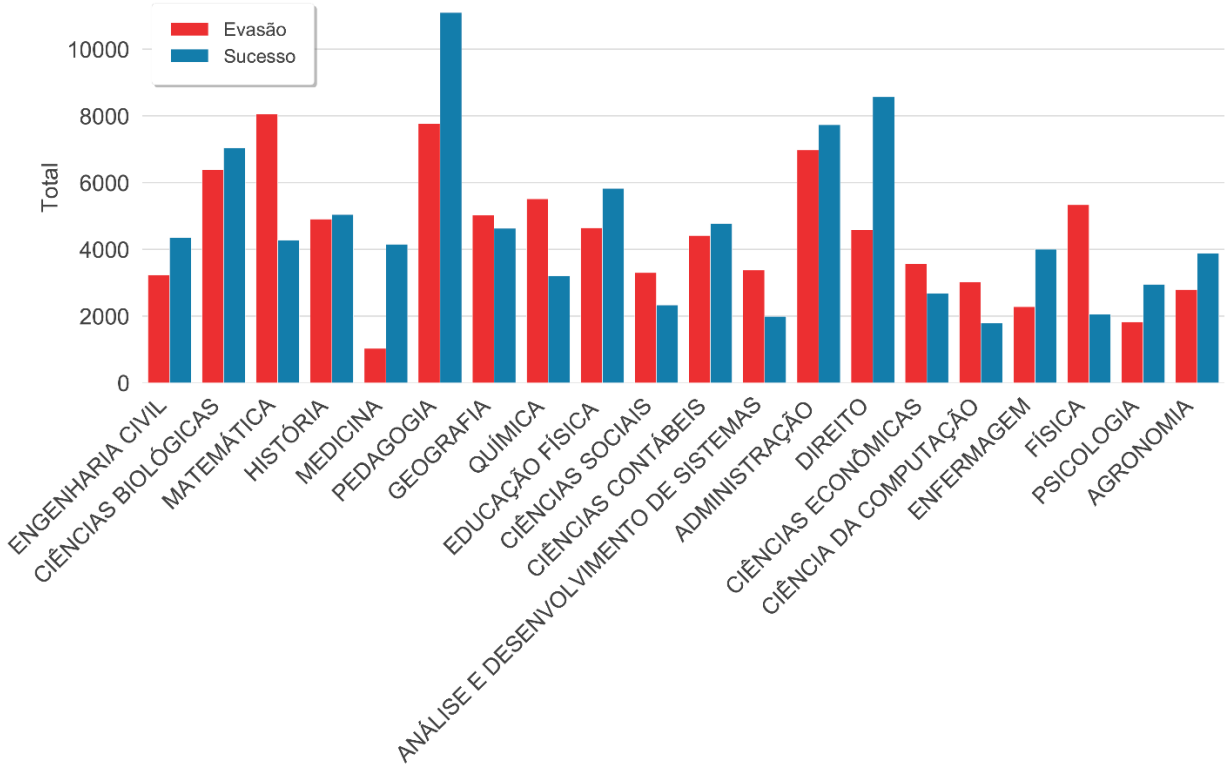

Figura 9: Total de casos de evasão e sucesso para os 20 cursos com maior número de alunos na base de dados.

Estudos pedagógicos evidenciam que o contexto socioeconômico e a diversidade cultural dos alunos têm grande influência na evasão do ensino superior (Reis, Cunha, & Spritzer, 2012), (Filho, Motejunas, Hipólito, & Lobo, 2007), (Bastos & Gomes, 2016). Os resultados mostrados na Figura 5 corroboram estes estudos, tendo em vista a importância observada para as variáveis que dizem respeito ao sexo, idade, tipo de escola em que foi cursado o ensino médio, classificação racial e presença de programas de apoio social. Os gráficos mostrados nas Figuras 10, 11 e 12 mostram a proporção de casos de evasão e sucesso para as variáveis sexo, raça e tipo de escola em que os alunos presentes na base concluíram o ensino médio. O local de nascimento do aluno também é uma variável demográfica apontada como importante pelo classificador. Para analisar esta informação, a Figura 13 mostra um mapa que indica a proporção de evasão de acordo com a variável CO_UF_NASCIMENTO, que corresponde ao estado em que o aluno nasceu. Estas figuras permitem a identificação dos grupos demográficos e regiões do país que precisam de mais atenção no desenvolvimento de políticas públicas de mitigação da evasão.

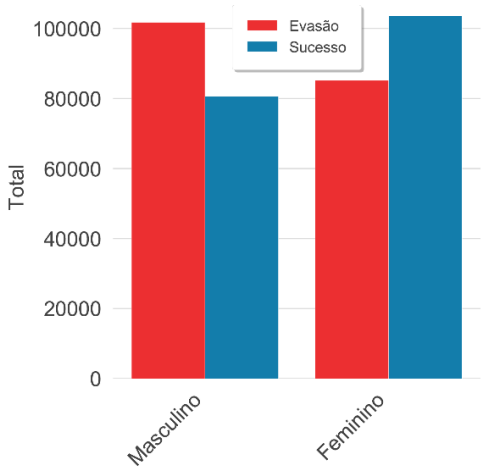

Figura 10: Total de casos de evasão e sucesso de acordo com o sexo dos alunos.

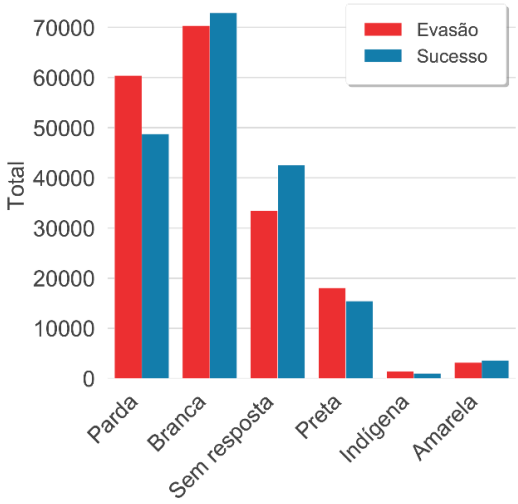

Figura 11: Total de casos de evasão e sucesso de acordo com a raça declarada dos alunos.

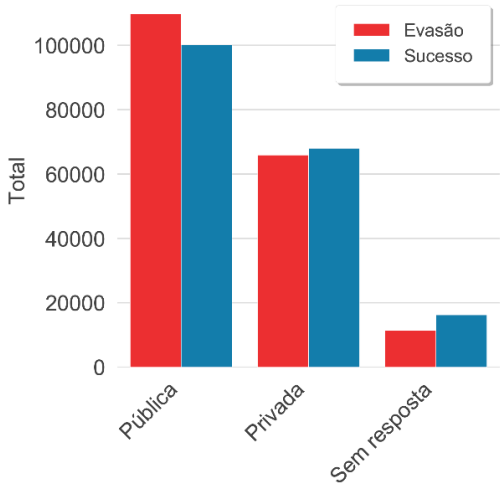

Figura 12: Total de casos de evasão e sucesso de acordo com o tipo de escola que o aluno concluiu ensino médio.

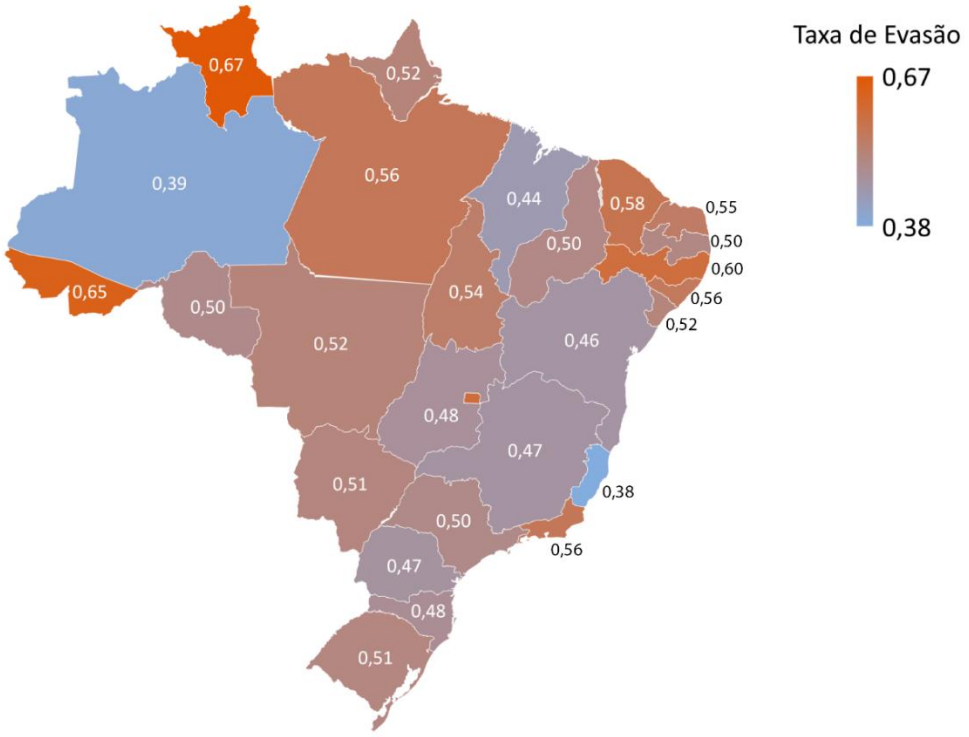

Figura 13: Taxa de evasão de acordo com o estado em que o aluno nasceu.

Utilizando os dados disponibilizados pelo INEP, não foi possível avaliar a influência que o desempenho dos alunos na progressão de seus cursos possui nas taxas de evasão. Para isso, seriam necessárias informações como notas, coeficientes de rendimento e dados de frequência. Em contrapartida, muitos trabalhos presentes na literatura consideram apenas aspectos locais, como notas e disciplinas do aluno em cursos específicos, para a previsão da evasão (Santos, Siebra, & Oliveira, 2014), (Manhães, Cruz, Costa, Zavaleta, & Zimbrão, 2011), (Silva & Imran, 2015). O estudo elaborado por (Pinheiro, Silva, & Souza, 2018), por exemplo, leva em consideração dados do histórico escolar de alunos, como notas, coeficientes de rendimento, frequência nas aulas, tempo de curso, a forma de ingresso e o turno do curso. Algumas destas características também foram identificadas como importantes no presente trabalho, como o turno do curso e forma de ingresso. A influência do turno do curso com a evasão pode ser visualizada na Figura 14.

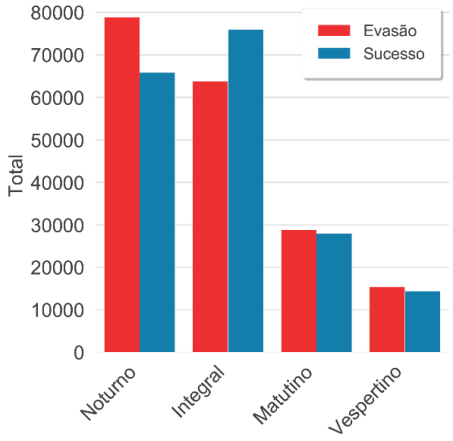

Figura 14: Total de casos de evasão e sucesso de acordo com o turno do curso em que o aluno se matriculou.

Assim, ao ser possível confirmar a maior parte dos resultados aqui obtidos com a análise de trabalhos conceituais relacionados, pode-se dizer que o presente trabalho revela as principais características, em um aspecto geral, dos alunos com maior risco de evasão. Os resultados também mostram que, utilizando a técnica *Random Forest,* é possível treinar um modelo capaz de ser usado, com aceitável grau de certeza, para a identificação de estudantes que precisam ser alvos de programas de incentivo aos estudos. Além disso, a análise exploratória das variáveis possibilitou o entendimento dos padrões e perfis dos grupos mais críticos de alunos que precisam de mais atenção no desenvolvimento de políticas públicas de combate à evasão. Os atributos apresentados na Figura 5 podem, então, ser considerados importantes para a orientação do combate ao problema em foco, com medidas como ampliação de planos de apoio social e incentivos à pesquisa e à adesão dos alunos a atividades extracurriculares, além de investimentos na formação dos professores.

5 Conclusões

As realizações de medidas para aumentar a taxa de sucesso de um aluno em um curso superior demandam a identificação dos pontos principais nos quais os estudantes e as IES podem estar negligenciando. Assim, a identificação dos padrões de um aluno que provavelmente irá falhar na conclusão do ensino superior é um problema de grande importância. Os resultados do presente trabalho revelaram as principais características do estudante que potencialmente influenciam de alguma forma na evasão escolar. Assim, torna-se possível para organizações governamentais o embasamento para a elaboração de planos de mitigação de evasão baseados nos padrões encontrados.

Primeiramente, os dados foram obtidos através da filtragem da base pública disponibilizada pelo INEP, da qual foram selecionados os estudantes da rede pública do ensino superior. Após uma fase inicial de filtragem de dados, seleção de variáveis, balanceamento e pré-processamento da base, os classificadores foram treinados e testados por um procedimento de validação cruzada para aquisição e validação dos resultados. A utilização de diferentes técnicas foi feita apenas com o objetivo de encontrar o melhor resultado possível para a previsão e, a partir dele, determinar as variáveis mais importantes para a classificação correta da evasão. Finalmente, foram levantados os trinta atributos mais relevantes e determinantes para a previsão da evasão, de acordo com o melhor classificador obtido. Uma análise exploratória das variáveis foi desenvolvida para identificar a relação que as variáveis possuem com a evasão, juntamente com o estabelecimento de um diálogo entre este trabalho e estudos acerca da evasão no ensino superior do Brasil.

A principal contribuição do presente trabalho vem na forma da identificação das variáveis mais importantes para a previsão de evasão ou sucesso de um aluno. Além disso, as relações entre as variáveis identificadas e a evasão foram estudadas, de forma que o presente estudo possa servir como base para futuros planos de mitigação da evasão que venham a ser desenvolvidos. Com isso, espera-se que os resultados aqui apresentados contribuam para o direcionamento apropriado de esforços na área da Educação, possibilitando o desenvolvimento de estratégias de redução de evasão focadas no suporte a estudantes que se encontram nos padrões característicos identificados. Porém, é importante lembrar que as análises aqui desenvolvidas se referem apenas aos dados de alunos de instituições públicas de ensino superior do Brasil. É possível que os resultados aqui obtidos possam ser extrapolados para outras categorias de instituições, níveis de ensino ou até por universidades do exterior, mas futuros trabalhos precisam ser desenvolvidos para confirmar esta hipótese.

Para trabalhos posteriores, resultados ainda mais interessantes podem ser obtidos ao levar em consideração não só os atributos utilizados aqui, presentes na base de dados do INEP, como

também as características específicas do desempenho dos alunos, como notas e coeficientes de rendimentos obtidos durante o período letivo. Dessa forma, unificar ambas as abordagens pode ser de grande importância na logística do combate à evasão.

Agradecimentos

Os autores agradecem ao CEFET-RJ pelo apoio no desenvolvimento desta pesquisa.

Referências

Ambiel, R. A. (2015). Construção da Escala de Motivos para Evasão do Ensino Superior. *Avaliação Psicológica, 14*(1), 41-52. doi:10.15689/ap.2015.1401.05 [GS Search]

Araque, F., Roldán, C., & Salguero, A. (2009). Factors influencing university drop out rates. *Computers & Education*, 563-574. doi:10.1016/j.compedu.2009.03.013 [GS Search]

Baker, R. S., Isotani, S., & Carvalho, A. M. (24 de Agosto de 2011). Mineração de Dados Educacionais: Oportunidades para o Brasil. *Revista Brasileira de Informática na Educação, 19*(2), 3-13. doi:10.5753/RBIE.2011.19.02.03 [GS Search]

Bastos, A., & Gomes, C. (2016). A evasão escolar no Ensino Técnico - Um estudo de caso do CEFET-RJ. *Educação e Cultura Contemporânea, 13*(32), 217-234. doi:10.5935/2238-1279.20160049 [GS Search]

Bonaldo, L., & Pereira, L. N. (2016). Dropout: Demographic profile of Brazilian university students. *Procedia - Social and Behavioral Sciences, 228*, 138-143. doi:10.1016/j.sbspro.2016.07.020 [GS Search]

Breiman, L. (2001, October). Random Forests. (R. E. Schapire, Ed.) *Machine Learning, 45*, 5-32. doi:10.1023/A:1010933404324 [GS Search]

Bruce, P., & Bruce, A. (2017). *Practical Statistics for Data Scientists: 50 Essential Concepts Using R and Python.* Sebastopol: O'Reilly. [GS Search]

Costa, S. L., & Dias, S. M. (2016). A permanência no ensino superior e as estratégias institucionais de enfrentamento da evasão. *Jornal de Políticas Educacionais, 9*(17/18), 51-60. doi:10.5380/jpe.v9i17/18.38650 [GS Search]

daCosta, F. J., SouzaBispo, M. d., & Pereira, R. d. (2018, March). Dropout and retention of undergraduate students in management: a study at a Brazilian Federal University. *RAUSP Management Journal, 53*(1), 74-85. doi:10.1016/j.rauspm.2017.12.007 [GS Search]

Delen, D. (12 de agosto de 2011). Predicting Student Attrition with Data Mining Methods. *Journal of College Student Retention: Research, Theory & Practice, 13*(1), 17-35. doi:10.2190/CS.13.1.b [GS Search]

Downey, A. (2012). *Think Bayes: Bayesian statistics in python.* Needham, Massachusetts, Estados Unidos da América: Green Tea Press. Fonte: https://greenteapress.com/wp/think-bayes/, Acesso em 10 de agosto de 2020. [GS Search]

Fernández, A., Galar, M., & Krawczyk, B. (2018). *Learning from Imbalanced Data Sets.* Gewerbestrasse, Switzerland: Springer. doi:10.1007/978-3-319-98074-4 [GS Search]

Ferreira, G. (2015). Investigação acerca dos fatores determinantes para a conclusão do Ensino Fundamental utilizando Mineração de Dados Educacionais no Censo Escolar da Educação Básica do INEP 2014. *Workshops do IV Congresso Brasileiro de Informática na Educação*

(pp. 1034-1043). Maceió: Sociedade Brasileira de Computação – SBC. doi:10.5753/cbie.wcbie.2015.1034 [GS Search]

Filho, R. L., Motejunas, P. R., Hipólito, O., & Lobo, M. B. (Setembro de 2007). A Evasão no Ensino Superior Brasileiro. *Cadernos de Pesquisa, 37*(132), 641-659. doi:10.1590/S0100-15742007000300007 [GS Search]

Gardner, M., & Dorling, S. R. (1998, August 1). Artificial neural networks (The multilayer perceptron)—A review of applications in the atmospheric sciences. *Atmospheric Environment, 32*(14-15), 2627-2636. doi:10.1016/S1352-2310(97)00447-0 [GS Search]

Gislason, P., Benediktsson, J., & Sveinsson, J. (2006, March). Random Forests for land cover classification. *Pattern Recognition Letters, 27*(4), 294-300. doi:10.1016/j.patrec.2005.08.011 [GS Search]

Granik, M., & Mesyura, V. (2017). Fake news detection using naive Bayes classifier. *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)* (pp. 900-903). Kyiv: IEEE. doi:10.1109/UKRCON.2017.8100379 [GS Search]

INEP. (20 de Janeiro de 2019). Acesso em 20 de janeiro de 2019, disponível em Portal INEP: http://portal.inep.gov.br/web/guest/dados

Instituto nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (20 de setembro de 2018). *Sinopse Estatística da Educação Superior 2017.* Acesso em 10 de agosto de 2020, disponível em Inep: http://inep.gov.br/sinopses-estatisticas-da-educacao-superior

Lamers, J., Santos, B., & Toassi, R. (2017). Retenção e evasão no ensino superior público: Estudo de caso em um curso norturno de odontologia. *Educação em Revista*, 33, 1-26. doi:10.1590/0102-4698154730 [GS Search]

Lerner, B., Levinstein, M., Rosenberg, B., Guterman, H., Dinstein, I., & Romem, Y. (1994). Feature Selection and Chromosome Classification Using a Multilayer Perceptron Neural Network. *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)* (pp. 3540-3545). Orlando: Institute of Electrical and Electronics Engineers. doi:10.1109/ICNN.1994.374905 [GS Search]

Manhães, L., Cruz, S., Costa, R., Zavaleta, J., & Zimbrão, G. (21 a 25 de Novembro de 2011). Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados. *Simpósio Brasileiro de Informática na Educação*, 150-159. Fonte: https://www.br-ie.org/pub/index.php/sbie/article/view/1585 [GS Search]

Manrique, R., Casanova, M. A., Nunes, B. P., Nurmikko-Fuller, T., & Marino, O. (2019). An Analysis of Student Representation, Representative Features and Classification Algorithms to Predict Degree Dropout. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (pp. 401–410). New York, NY, USA: Association for Computing Machinery. doi:10.1145/3303772.3303800 [GS Search]

Martins, L. C., Carvalho, R. N., & Carvalho, R. S. (2017). Early prediction of college attrition using data mining. *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 1075-1078). Cancun: IEEE. doi:10.1109/ICMLA.2017.000-6 [GS Search]

Meedech, P., Iam-On, N., & Boongoen, T. (2016). Prediction of Student Dropout Using Personal Profile and Data Mining Approach. In P.-A. S. Lavangnananda K. (Ed.), I*ntelligent and Evolutionary Systems. Proceedings in Adaptation, Learning and Optimization* (Vol. 5, pp. 143-155). Springer. doi:10.1007/978-3-319-27000-5_12 [GS Search]

Mitchell, T. M. (1997). *Machine Learning.* McGraw-Hill Science/Engineering/Math. [GS Search]

Müller, A., & Guido, S. (2017). Introduction to Machine Learning with Python: A GUIDE FOR DATA SCIENTISTS. Em A. Müller, & S. Guido, *Introduction to Machine Learning with Python: A GUIDE FOR DATA SCIENTISTS* (pp. 68-74,282-284). Sebastopol: O'Reilly. [GS Search]

Nascimento, R., Junior, G., & Roberta, F. (Julho de 2018). Mineração de Dados Educacionais: Um Estudo Sobre Indicadores da Educação em Bases de Dados do INEP. *RENOTE - Revista Novas Tecnologias na Educação , 16*(1), 1-11. doi:10.22456/1679-1916.85989 [GS Search]

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*(85), 2825−2830. Fonte: https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html [GS Search]

Pinheiro, M., Silva, J., & Souza, B. (2018). Aprendizado de Máquina Aplicado à Análise de Evasão no Ensino Superior. *Computer on the beach*, 512-521. Fonte: https://siaiap32.univali.br/seer/index.php/acotb/article/view/12810 [GS Search]

Prestes, E. M., & Fialho, M. G. (2018). Evasão na educação superior e gestão institucional: o caso da Universidade Federal da Paraíba. *Ensaio: Avaliação e Políticas Públicas em Educação, 26*(100), 869-889. doi:10.1590/s0104-40362018002601104 [GS Search]

Reis, V., Cunha, P., & Spritzer, I. (2012). Evasão no Ensino Superior de Engenharia no Brasil: Um estudo de caso no Cefet/RJ. *XL Congresso Brasileiro de Educação em Engenharia*. [GS Search]

Rigo, S. J., Cambruzzi, W., Barbosa, J. L., & Cazella, S. i. (2014). Aplicações de Mineração de Dados Educacionais e Learning Analytics com foco na evasão escolar: oportunidades e desafios. *Revista Brasileira de Informática na Educação, 22*(1). doi:10.5753/RBIE.2014.22.01.132 [GS Search]

Rodrigues, F. S., Brackmann, C. P., & Barone, D. A. (2015). Estudo da Evasão no Curso de Ciência da Computação da UFRGS. *Revista Brasileira de Informática na Educação, 23*(1), 97-109. doi:10.5753/RBIE.2015.23.01.97 [GS Search]

Sales, A., Balby, L., & Cajueiro, A. (2016, August). Exploiting Academic Records for Predicting Student Drop Out: a case study in Brazilian higher education. *Journal of Information and Data Management, 7*(2), 166-180. Fonte: https://periodicos.ufmg.br/index.php/jidm/article/view/343 [GS Search]

Santos, K. J., Menezes, A. G., Carvalho, A. B., & Montesco, C. A. (2019). Supervised Learning in the Context of Educational Data Mining to Avoid University Students Dropout. *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)* (pp. 207-208). Maceió: IEEE. doi:10.1109/ICALT.2019.00068 [GS Search]

Santos, R., Siebra, C., & Oliveira, E. (2014). Uma Abordagem Temporal para Identificação Precoce de Estudantes de Graduação a Distância com Risco de Evasão em um AVA utilizando Árvores de Decisão. *Congresso Brasileiro de Informática na Educação* (pp. 262-271). Dourados: Sociedade Brasileira de Computação – SBC. doi:10.5753/cbie.wcbie.2014.262 [GS Search]

Sarker, F., Tiropanis, T., & Davis, H. C. (2014). Linked data, data mining and external open data for better prediction of at-risk students. *2014 International Conference on Control, Decision and Information Technologies (CoDIT)* (pp. 652-657). Metz: IEEE. doi:10.1109/CoDIT.2014.6996973 [GS Search]

Silva, J., & Imran, H. (Dezembro de 2015). Um estudo sobre as variáveis para predição de alunos não concluintes em cursos suportados por Ambientes Virtuais de Ensino e Aprendizagem.

RENOTE - Revista Novas Tecnologias na Educação, 13(2). doi:10.22456/1679-1916.61427 [GS Search]

Vlahou, A., Schorge, J., Gregory, B., & Coleman, R. (2003). Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data. *Journal of Biomedicine and Biotechnology*, 308-314. Fonte: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC521504/ [GS Search]

Zhang, Y., Oussena, S., Clark, T., & Kim, H. (2010). Use Data Mining to Improve Student Retention in Higher Education - A Case Study. *ICEIS 2010 - Proceedings of the 12th International Conference on Enterprise Information Systems*, (pp. 190-197). Madeira. Fonte: https://dblp.uni-trier.de/db/conf/iceis/iceis2010-1.html [GS Search]