Avaliação da evasão universitária com técnicas de Aprendizagem de Máquina

Thiago Gomes Veríssimo[1]*; Douglas Augusto de Paula[2]

[1] Universidade de São Paulo. Analista de Sistema. Rua do Lago, 717 – Cidade Universitária; 05508-080 São Paulo, SP, Brasil
[2] Mestre em Controladoria e Contabilidade pela FEA/USP
*autor correspondente: thiago.verissimo@usp.br

Avaliação da evasão universitária com técnicas de "Machine Learning"

Resumo

Este trabalho investiga a aplicação de técnicas de Aprendizagem de Máquina ou no termo em inglês "Machine Learning" [ML] no contexto educacional e em particular na predição da evasão universitária. O desligamento de alunos no curso universitário é um fenômeno que se busca evitar, pois suas consequências são negativas para a sociedade e para o indivíduo evadido, entretanto, dada a natureza variada dos possíveis motivos que podem levar à evasão, é árdua a tarefa na concepção de estratégias para sua diminuição, neste contexto, técnicas de ML podem dar suporte na criação dessas estratégias. Esse projeto enquadra-se na linha de pesquisa interdisciplinar denominada "Learning Analytics" [LA]. Em levantamento bibliográfico prévio identificou-se que grande parte das pesquisas envolvendo predição da evasão universitária utilizam principalmente dados de cursos de graduação das áreas de exatas, neste projeto estudou-se um curso da área de humanidades. A partir dos resultados obtidos, mostrou-se que aplicação de ML em dados educacionais pode ser uma ferramenta auxiliar de suporte de tomada de decisão e de apoio as políticas educacionais para manutenção do quadro discente. Os modelos de ML utilizados produziram taxas de acertos acima de 70%. Usaram-se cinco algoritmos de ML supervisionados no contexto de classificação e após o procedimento de validação cruzada avaliou-se a capacidade preditiva de cada modelo segundo a área da curva "Receiver Operating Caracteristic" (ROC): "Gradiente Boosting" (82,3%), Rede Neural (81,8), "Random Forest" (80,3), Árvore de Decisão (79,9%) e Regressão Logística (79,7%). Modelos mais flexíveis do ponto de vista de ajuste aos dados como "Gradiente Boosting" e Rede Neural forneceram as maiores taxas de acertos. Regressão Logística apresentou a menor área ROC, entretanto, além da identificação de indivíduos com alta chance de evasão, é importante também identificar suas causas. Neste sentido, considerou-se que o fornecimento de uma métrica para avaliação da contribuição de cada variável preditiva na evasão, como a significância estatística fornecida pela regressão logística, é um parâmetro importante na consideração do melhor modelo. Escolheu-se variáveis preditivas referentes ao relacionamento acadêmico entre discentes e instituição, como quantidades de aprovações e reprovações em disciplinas, auxílios financeiros, entre outros. Considerou-se dados dos 3 primeiros anos da graduação. Encontrou-se que a quantidade de auxílios estudantis recebido pelos discentes não tiveram significância estatística para explicar a evasão universitária. A quantidade de reprovações em disciplinas obrigatórias no primeiro ano da graduação teve o maior peso na evasão. Variáveis referentes aos primeiros anos tiveram maior peso no geral, mostrando que possíveis intervenções com o objetivo de diminuir evasão poderiam ter como foco o primeiro ano. Por fim, não apenas um algoritmo em particular, mas a combinação deles possibilitou um entendimento mais abrangente dos motivos da evasão universitária.

Palavras-chave: evasão universitária; learning analytics; aprendizado de máquina; educação; ciência de dados.

Introdução

A evasão no ensino superior é um problema global e de difícil solução. O impacto de altas taxas de desistências em cursos de graduação não se limita à vida privada dos indivíduos desistentes, tampouco se restringe às instituições de ensino. O efeito da evasão atinge a sociedade, a economia e o Estado, pois ao mesmo tempo em que o investimento monetário se dissipa, seja público ou privado, a sociedade não recebe os possíveis benefícios

dos esforços empregados na formação educacional dos indivíduos evadidos (Rodrigues et al., 2021).

Embora o foco dessa análise seja o ensino superior, a evasão representa um problema em todas fases do processo educacional. Filho e Silveira (2021) fizeram um levantamento sistemático sobre a evasão no âmbito de escolas secundárias do Brasil e com o auxílio de Ciência de Dados construíram modelos que fornecem probabilidades de alunos evadirem, baseando-se nos dados históricos de concluintes e evadidos. No contexto secundário, tais modelos são conhecidos como “Early Warning System” [EWS]. Ainda segundo Filho e Silveira (2021) simples intervenções, como contatos telefônicos, podem diminuir a taxas de evasão em muitos casos, no caso de crianças e adolescentes. Já no contexto universitário, a identificação precoce de indivíduos na graduação com alta probabilidade de evasão poderia ajudar gestores na criação de estratégias para mitigar ou ao menos minimizar o problema.

O uso de Aprendizagem de Máquina ou em inglês “Machine Learning” [ML] enquanto ferramenta de suporte tem ganhado espaço na última década na predição de evasão (Mduma et al., 2019), apesar da qualidade dos dados educacionais ainda ser um fator limitante para criação de modelos com alto poder preditivo. Mesmo com a ausência ou baixa qualidade de dados, pesquisas de ML e educação têm ganhado destaques nas últimas décadas (Rafiq et al., 2021), como, por exemplo, no caso brasileiro o trabalho desenvolvido por Rodrigues et al. (2021) que utilizando dados de instituições de ensino superior federais brasileiras criaram modelos de ML para predição de evasão e encontraram acurácia de 82,2%.

Várias áreas do conhecimento se debruçam a respeito das circunstâncias que envolvem os indivíduos evadidos, como pedagogia, psicologia, sociologia, saúde pública, dentre outras. Aqui fez-se uma abordagem baseada em Ciência de Dados. De forma simplificada agrupam-se os motivos que causam evasão universitária em três dimensões: fatores internos à instituição de ensino; fatores externos à instituição de ensino; e fatores individuais. As variáveis preditivas que contribuem para evasão podem ser de difícil acesso dependendo da dimensão que se está interessado, pois os dados podem não existirem ou podem não estarem sistematizadas e organizadas em algum banco de dados de fácil acesso. No geral, as variáveis explicativas da dimensão fatores internos à instituição de ensino são as de mais fácil acesso, pois normalmente estão organizadas para o devido funcionamento de sistemas de controle da dinâmica acadêmica.

No Brasil, trabalhos como o de Digiampietri et al. (2016) e de Jesus et al. (2021) aplicaram técnicas de ML para predição de evasão em cursos de graduação na área de exatas. Em levantamento bibliográfico para este projeto identificou-se que na literatura científica recente ainda são poucas análises similares com cursos da área de humanidades. Este projeto se propôs a investigar numa perspectiva de Ciência de Dados, e de forma mais

restrita na área de pesquisa denominada “Learning Analytics” (Ferguson, 2012), o fenômeno da evasão de discentes do curso superior de graduação em Letras da Universidade de São Paulo [USP], com enfoque em dados acadêmicos dos três primeiros anos do curso.

O objetivo desse Trabalho de Conclusão de Curso [TCC] consistiu em prover possíveis caminhos e sugestões para que atores envolvidos no processo educacional possam desenvolver estratégias para lidar com a problemática da evasão universitária, usando para tal, o ferramental provido por técnicas de Ciência de Dados, e em especial, ML. Para alcançar tal objetivo aplicaram-se diversos modelos que fornecessem probabilidades de evasão, além disso, identificaram-se variáveis preditivas com maiores ou menores pesos nessas probabilidades, informações essas extremamente valiosas para gestores sob uma ótica institucional.

A gestão educacional é a responsável pela criação de estratégias que buscam aprimorar a qualidade do ensino e diminuir a evasão, esteja essa gestão em um âmbito local da universidade ou no âmbito mais amplo, como política de Estado. Deste modo, explorar dados que possam dar suporte ao desenvolvimento dessas estratégias é essencial para que se tenha um resultado mais efetivo.

Este projeto foi submetido e aprovado pelo Conselho de Ética e Pesquisa (CEP) e está registrado sob o número CAAE de 56238121.8.0000.0138.

Material e Métodos

Foram analisados dados no nível observacional, ou seja, de cada discente, extraídos do sistema administrativo que opera a gestão de discentes do curso de graduação em Letras da Faculdade de Filosofia, Letras e Ciências Humanas (FFLCH) da Universidade de São Paulo (USP), com recorte de 15 anos, englobando discentes que ingressaram a partir de 2000 até 2014. O banco de dados base do sistema administrativo é do tipo relacional, normalizado e contém informações acadêmicas detalhadas dos discentes desde o início do vínculo até a fim, tanto para casos de concluintes ou de evadidos. Contém ainda alguns dados pregressos ao vínculo, como tipo de instituição de ensino da escola secundária, notas em exames de ingresso, entre outros. A seleção de variáveis preditivas é uma etapa importante e crucial para a construção de modelos (Mduma et al., 2019), neste caso, tratando-se de um banco de dados relacional e normalizado, os dados estão armazenados de forma esparsa nas tabelas, e consequentemente para cada variável escolhida como explicativa, exigiu-se a identificação das tabelas correspondentes e respectiva junção.

O projeto foi executado nas seguintes etapas:

Fase 1 - “Extract, Transform and Load” (ETL): Estudou-se a estrutura dos dados, criaram-se consultas e “scripts” para automação da extração dos dados e preparação em um formato tabular. Levantaram-se possíveis informações individuas dos discentes que foram usadas como preditivas, como idade no ingresso, quantidade de aprovações ou reprovações nas disciplinas, discernimento do peso das reprovações de disciplinas obrigatórias ou não na grade curricular, estado civil, gênero declarado, estado civil e quantidades de auxílios de permanência estudantil.

Fase 2 - Pré-processamento dos dados: Análise descritiva e sumarização dos dados.

Fase 3: Extração do conhecimento: Aplicação dos modelos de ML. Avaliação dos parâmetros e respectivos testes estatísticos.

Fase 4: Validação: análise do conhecimento obtido, avaliações dos algoritmos treinados. Julgamento conceitual e subjacente dos resultados e comparação com a literatura científica correlata.

Os dados de 33.195 discentes ingressantes no curso de graduação em Letras entre os anos de 2000 e 2014 foram extraídos do sistema acadêmico e organizados no formato “wide”, isto é, observações nas linhas e variáveis nas colunas. As 20 variáveis selecionadas para esse estudo estão apresentadas na Tabela 1.

Tabela 1. Variáveis selecionadas neste estudo

Sigla	Descrição
evadido	sim ou não
idade	Idade em anos no momento do ingresso
sexo	Feminino (F) e Masculino (M)
cor	Branca, Amarela, Não informada, Indígena, Parda, Preta/negra
ec	Divorciado, Casado, Solteiro, Outro, Separado judicialmente, União Estável, Viúvo
ay1	Quantidade de auxílios estudantis no primeiro ano da graduação
ay2	Quantidade de auxílios estudantis no segundo ano da graduação
ay3	Quantidade de auxílios estudantis no terceiro ano da graduação
o1a	Quantidade de aprovações em disciplinas obrigatórias no primeiro ano da graduação
o2a	Quantidade de aprovações em disciplinas obrigatórias no segundo ano da graduação
o3a	Quantidade de aprovações em disciplinas obrigatórias no terceiro ano da graduação
no1a	Quantidade de aprovações em disciplinas não obrigatórias no primeiro ano da graduação
no2a	Quantidade de aprovações em disciplinas não obrigatórias no segundo ano da graduação
no3a	Quantidade de aprovações em disciplinas não obrigatórias no terceiro ano da graduação
o1r	Quantidade de reprovações em disciplinas obrigatórias no primeiro ano da graduação
o2r	Quantidade de reprovações em disciplinas obrigatórias no segundo ano da graduação
o3r	Quantidade de reprovações em disciplinas obrigatórias no terceiro ano da graduação
no1r	Quantidade de reprovações em disciplinas não obrigatórias no primeiro ano da graduação
no2r	Quantidade de reprovações em disciplinas não obrigatórias no segundo ano da graduação
no3r	Quantidade de reprovações em disciplinas não obrigatórias no terceiro ano da graduação

Fonte: Dados originais da pesquisa

A escolha para seleção das variáveis teve como premissa a investigação de possíveis fatores que pudessem ter impacto na evasão considerando a vida universitária nos três primeiros anos do curso, assim as aprovações e reprovações foram quantificadas em disciplinas categorizadas em dois tipos: obrigatórias e não obrigatórias na grade curricular correspondente do curso.

A USP oferece auxílios para permanência estudantil em diversos formatos e abrange áreas como moradia, alimentação, ajuda financeira, dentre outros. No intuito de capturar a influência de auxílios optou-se por contabilizar a quantidade de auxilio indistintamente em cada um dos três primeiros anos. Os auxílios têm duração máxima de 1 ano, com a possibilidade de renovação, a depender das regras de cada auxílio. As variáveis explicativas foram escolhidas de forma a evitar a multicolinearidade.

Aplicaram-se cinco algoritmos de classificação: Regressão Logística, Árvore de Decisão, “Random Forest” (RF), “Gradiente Boosting” (GB) e Redes Neurais.

Para o primeiro deles, a regressão logística, foi realizada uma abordagem de cunho estatístico, isto é, analisaram-se a significância estatística dos testes referentes aos parâmetros e a eficiência global do modelo, removeram-se os parâmetros não estatisticamente significativos com apoio da metodologia *stepwise*, que testa combinações de parâmetros para manter somente aqueles que quando em conjunto apresentam testes estatisticamente significantes.

Os algoritmos de Árvore de Decisão, RF, GB e Redes Neurais são flexíveis o suficiente para se ajustarem perfeitamente aos dados, situação denominada como “overfitting”. O problema do “overfitting” se manifesta quando a capacidade preditiva nos dados usado para treinamento é alta, porém quando aplicado em novas observações, a capacidade preditiva diminui drasticamente. O procedimento usado para identificar situação de “overfitting” foi o de validação cruzada, que consiste em dividir os dados em dois grupos: treinamento e teste, usados na construção do modelo, e na validação, respectivamente. Foram separadas aleatoriamente 80% das observações para treinamento e 20% para teste. Toda análise foi realizada no software R, versão 4.1.3.

Fundamentação da Regressão Logística

O fenômeno estudado nesse trabalho, evasão universitária, tem natureza qualitativa e respectiva variável dependente do tipo categórica e dicotômica, sendo que a ocorrência do evento se dá quando o discente é evadido e a não ocorrência quando é concluinte. Dada a natureza do problema, uma possível solução é o modelo de regressão logística binária

(Fávero e Belfiore, 2017), com o que se estima a probabilidade de ocorrência do evento (p_i) a partir do comportamento e combinação linear das variáveis explicativas (x_i). A Equação 1 apresenta a combinação linear ainda sem a probabilidade explicitada.

$$Z_i = \alpha + \beta_1 x_{1_i} + \beta_2 x_{2_i} + \cdots + \beta_k x_{k_i} \quad (1)$$

em que i indica cada observação das k variáveis explicativas. Z_i é a variável dependente com domínio contínuo variando de menos infinito até infinito, comumente chamada de *logito*. Para encontrarmos de fato a probabilidade de ocorrência do evento assume-se que Z_i é dado pelo logaritmo natural da chance, como definido pela Equação 2.

$$Z_i = ln\left(\frac{p_i}{1-p_i}\right) \quad (2)$$

sendo p_i a probabilidade de ocorrência do evento. Isolando a probabilidade p_i em função logito Z_i na Equação 2 chega-se à equação final da probabilidade, Equação 3.

$$p_i = \frac{1}{1+e^{-Z_i}} \quad (3)$$

Etapas para avaliação da eficiência do modelo de Regressão Logística

Considerando Y_i como a ocorrência (1) ou não (0) do evento na observação i, a função de verossimilhança L é uma função custo, ou seja, aquela que usamos para avaliar a performance do modelo e é definida para n observações na Equação 4.

$$L = \prod_{i=1}^{n} [p_i^{Y_i}(1-p_i)^{1-Y_i}] \quad (4)$$

Interessa-nos estimar os parâmetros β_k da equação do logito Z_i que maximizam o logaritmo da função de verossimilhança L, ou seja, LL. Cada parâmetro β_k pode ou não ser estatisticamente significante para explicar o comportamento da variável resposta, evasão, e conforme incluímos ou excluímos variáveis preditivas no modelo, parâmetros que não eram estatisticamente significantes podem se tornar significantes ou contrário. Seria muito custoso

o processo para fazer todas combinações possíveis e procurar todas variáveis preditivas que em conjunto sejam estatisticamente significantes, assim optou-se pela utilização de um procedimento automático, que testasse todas combinações, conhecido como “stepwise”, e implementado no método “step” do pacote stats do R, ”, versão 4.2.0. Nesta versão, a função “step” apresenta um pequeno “bug”, que mesmo quando definimos as colunas categóricas como do tipo “factor” há um problema para remoção dos β_k não significativos. Para contornar esse problema utilizou-se o pacote “fastDummies”, versão 1.6.3, para transformação das colunas categóricas em colunas binárias, em que cada tipo de categoria se transforma em uma coluna binária, com zero representando a ausência da categoria e um a presença. Na regressão logística foram usadas as funções “glm” e “step” do pacote “stats”, versão 4.2.0.

Árvore de Decisão

A árvore de decisão é um algoritmo de tomada de decisões sequências e condicionais que traça possíveis caminhos que levam a variável resposta. É normalmente representada em uma estrutura de árvore invertida, com a raiz na parte de cima e as folhas na base. Cada nó da árvore é um ponto de tomada de decisão baseada no valor de alguma variável explicativa. A priori, seria possível traçar o caminho exato para a variável resposta, situação de “overfitting”. Para evitar o “overfitting”, define-se um parâmetro que controla a flexibilidade da árvore de decisão, chamado de custo. Custo com valor zero indica uma árvore que pode ajustar-se perfeitamente aos dados de treinamento. Quanto maior o custo, menos flexível a árvore de decisão. Foi utilizado o pacote “rpart” versão 4.1.16 (Breiman et al., 1984).

Random Forest

“Random Forest” é um algoritmo baseado na construção de múltiplos modelos aleatórios. “Bootstraping” é uma técnica de criação de modelos baseados em amostragem com reposição nas observações. “Aggregation” é uma técnica de agregação de modelos sob uma métrica em particular. A combinação do “bootstraping” e “aggregation” gera uma classe de modelos conhecida como “Bagging”. Random Forest é um tipo de “bagging” que usa modelos do tipo árvores de decisão e acrescenta aleatoriedade nas escolhas de variáveis explicativas. Utilizou-se o pacote “randomForest” versão 4.7.1 (Breiman, 2001).

Gradient Boosting

"Boosting" é um algoritmo que gera modelos sequências que tentam melhorar os erros dos modelos anteriores. Na prática ajusta-se o modelo nos ruídos gerados no modelo anterior, de forma a minimizar os erros. O "Gradient Boosting" (GD) é uma variação de "boosting" que usa árvores de decisão como modelo. Utilizou-se o pacote "xgboost" versão 1.6.0 (Friedman et al., 2000).

Rede Neural

Rede neural é um algoritmo baseado em combinações lineares sucessivas que ocorrem dentro de nós e em camadas. Na prática, cada nó funciona como um atenuador de sinal, pois pode aumentar ou diminuir os pesos das variáveis explicativas nas sucessivas camadas. Devido à natureza do nó em atenuar os pesos, dá-se a ele o nome de neurônio, dada a similaridade com os neurônios do cérebro humano, que possuem a função de transmissores de sinais elétricos para controle das atividades mentais ou físicas. O pacote "nnet" versão 7.3.17 foi utilizado (Ripley, 1996).

Resultados e Discussão

Análise Descritiva e Exploratória

O anuário estatístico da USP disponibiliza dados oficiais agregados da rotina acadêmica (Anuário USP, 2022). Nos anuários mais recentes, é possível consultar a taxa de evasão dos cursos de graduação de toda universidade. Em 2017, 2018, 2019 e 2020 as taxas para o curso de Letras foram 25,9%, 36,5%, 40,2%, 33,6% respectivamente. Nota-se que esses anos são posteriores ao período considerado nesse estudo.

Para o recorte realizado, 2000 até 2014, a proporção total de evasão foi de 38,3% contra 61,7% de concluintes. Na Figura mostra a taxa anual da evasão no curso de Letras considerando o ano de ingresso:

Figura 1. Taxa anual da evasão no curso de Letras da USP considerando o ano de ingresso entre 2000 e 2014
Fonte: Resultados originais da pesquisa

A idade média dos discentes evadidos foi μ_e 23,3 anos e dos concluintes μ_c 21,0 anos. Para verificar se em um nível de significância de 5% essas médias são iguais, aplicou-se um teste estatístico com as seguintes hipóteses:

- Hipótese Nula: $\mu_e = \mu_c$
- Hipótese Alternativa: $\mu_e \neq \mu_c$

Supondo uma distribuição normal para as idades, o "p-value" encontrado foi 0. Considerando um nível de significância de 5%, pode-se rejeitar a hipótese nula, e as médias das idades dos discentes evadidos μ_e e concluintes μ_c são diferentes. Na Figura 2 é apresentado os histogramas da densidade da distribuição de idades para os respectivos grupos (evadidos e concluintes):

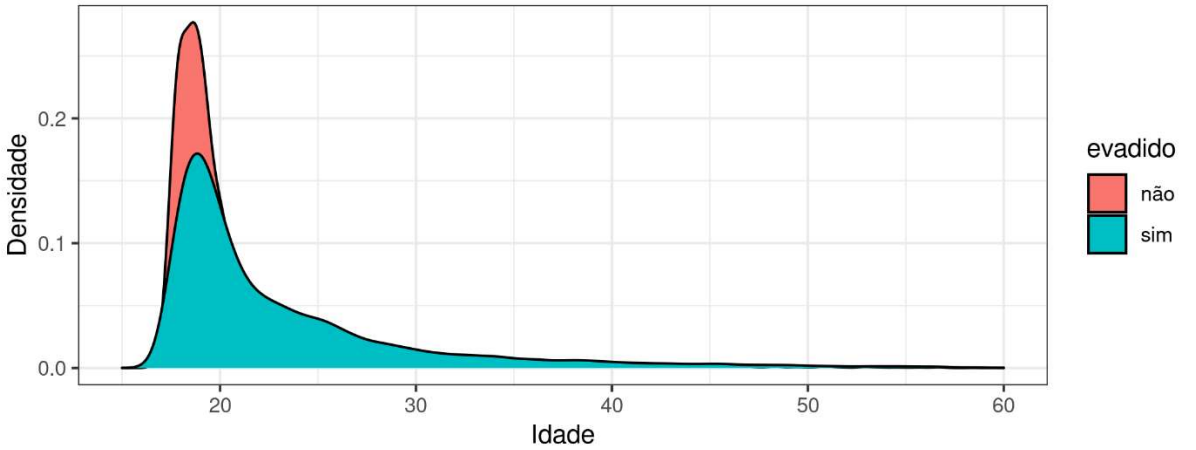

Figura 2. Histogramas da densidade da distribuição de idades para evadidos e concluintes do curso de graduação em Letras da USP entre 2000 e 2014.
Fonte: Resultados originais da pesquisa

Regressão Logística

No universo de 19 variáveis explicativas, três são categóricas: sexo, cor e estado civil. Criou-se colunas artificiais e binárias para cada tipo de categoria, sendo o critério de escolha da referência a variável mais frequente, que para sexo foi feminino, para estado civil foi Solteiro e para cor foi Branca. Após este processo a base de dados final ficou com 28 variáveis explicativas.

Criaram-se três modelos logísticos:

- M1 - Modelo Nulo: Só contém o termo constante alfa e nenhum beta. Será usado para avaliação da significância estatística global do modelo
- M2 - Modelo Completo: Com todas as 19 variáveis explicativas (28 contando as colunas artificiais categóricas) da base de dados, independente das respectivas significâncias estatísticas
- M3 - Modelo Stepwise: Somente com variáveis explicativas que foram estatisticamente significantes após aplicação do método stepwise.

O Modelo Nulo apresentou logaritmo da verossimilhança $\mathrm{LL}_{\mathrm{nulo}}$ de -22073,67. O Modelo Completo, com as 28 variáveis explicativas, apresentou logaritmo da verossimilhança $\mathrm{LL}_{\mathrm{completo}}$ de -17253,63 e por fim o Modelo Stepwise selecionou 18 variáveis explicativas estatisticamente significantes à 95% de confiança e apresentou logaritmo da verossimilhança $\mathrm{LL}_{\mathrm{stepwise}}$ de -17261,41. O Modelo Stepwise e o Modelo Completo apresentaram logaritmo da verossimilhança similares e usaremos o teste qui-quadrado para verificar se há diferença entre ambos.

O teste qui-quadrado (χ^2) é adequado para modelos com parâmetros estimados por método de máxima verossimilhança e o usaremos para comparação entre os modelos, definindo-o assim $\chi^2 = -2(\mathrm{LL}_{\mathrm{modelo}_1} - \mathrm{LL}_{\mathrm{mode}\ _2})$, sendo que o $\chi^2{}_{\mathrm{crítico}}$ será relativo distribuição do tipo χ^2 com números de graus de liberdade equivalente a diferença entre os números de graus de liberdade do modelo 1 e do modelo 2.

Inicialmente vamos comparar o Modelo Completo e o Modelo Stepwise para verificação se há diferença estatística entre eles. Baseado nos valores de $\mathrm{LL}_{\mathrm{stepwise}}$ e $\mathrm{LL}_{\mathrm{completo}}$ elaborou-se as seguintes hipóteses:

- Hipótese Nula (H_0): $\mathrm{LL}_{\mathrm{stepwise}} - \mathrm{LL}_{\mathrm{completo}} = 0$. A diferença entre $\mathrm{LL}_{\mathrm{stepwise}}$ e $\mathrm{LL}_{\mathrm{completo}}$ é nula, ou seja, os dois modelos são estatisticamente iguais.
- Hipótese Alternativa (H_a): $\mathrm{LL}_{\mathrm{stepwise}} - \mathrm{LL}_{\mathrm{completo}} \neq 0$. A diferença entre $\mathrm{LL}_{\mathrm{stepwise}}$ e $\mathrm{LL}_{\mathrm{completo}}$ **não** é nula, ou seja, os dois modelos são estatisticamente diferentes.

O valor encontrado de χ^2 entre o Modelo Completo e o Modelo Stepwise foi 15,5. Como há 28 variáveis explicativas no Modelo Completo e 18 variáveis explicativas no Modelo Stepwise, a diferença dos graus de liberdade nos dois modelos é de 10. Numa distribuição χ^2 com 10 graus de liberdade a área até 15,5, ou seja, o "p-value" é 0,113. Considerando um nível de significância de 5%, e dado que 11,3% > 5%, não rejeita-se a hipótese nula e assume-se que não há diferença estatística entre o Modelo Completo e o Modelo Stepwise.

O modelo nulo é aquele que só contém a constante alfa (α) e nenhum beta (β_k). Usou-se tal modelo para avaliar se pelo menos um β_k é estatisticamente significante para explicar nossa variável resposta, a evasão universitária. Aplicou-se o mesmo procedimento anteriormente usado entre o Modelo Completo e o Modelo Stepwise para comparar o Modelo Nulo ao Modelo Stepwise. A hipótese nula será aquela em que nenhum dos parâmetros são estatisticamente significantes:

- Hipótese Nula (H_0): $H_0: \beta_1 = \beta_2 \dots \beta_k = 0$

A hipótese alternativa é aquela em que ao menos um dos β_k é estatisticamente significante:

- Hipótese Alternativa (H_a): Existe pelo menos um β_k

O χ^2 entre o Modelo Nulo ao Modelo Stepwise, $\chi^2 = -2(LL_{nulo} - LL_{stepwise})$, foi de 9624,521 e o respectiva área na distribuição χ^2 ("p-value") com 18 graus de liberdade foi zero. Considerando um nível de significância de 5%, rejeita-se a hipótese nula e considera-se a hipótese alternativa, de que existe pelo menos um β_k.

O modelo matemático de Regressão Logística da como resposta a probabilidade de ocorrência do evento de interesse. Cabe ao condutor(a) da análise a definição do ponte de corte (p_{cutoff}) a partir do qual a probabilidade será escolhida como ocorrência do evento ou não, dependendo do contexto da pesquisa. A mudança do ponto de corte impacta diretamente em três métricas comumente empregadas para avaliação do modelo:

- Acurácia: taxa de acertos
- Sensibilidade: taxa de acerto do modelo para as observações classificadas como "evento"
- Especificidade: taxa de acerto do modelo para as observações classificadas como "não evento".

A Tabela 2 apresenta a matriz de classificação para Regressão Logística considerando-se um p_{cutoff} de 50%.

Tabela 2. Tabela de confusão para Regressão Logística e p_{cutoff} de 50%

		Observado	
		Evadido	Concluinte
Previsto	Evadido	6824	2192
	Concluinte	5852	18327

Fonte: Resultados originais da pesquisa

Respectivos cálculos manuais da Sensibilidade, Especificidade e Acurácia:

- Sensibilidade: 6824/(6824+5852) = 53,8%
- Especificidade: 18327/(18327+2192) = 89,3%
- Acurácia: (6824+18327)/33195 = 75,7%

Se p_{cutoff} fosse alterado, obter-se-ia outra tabela de classificação e portanto outros valores de sensibilidade, especificidade e acurácia. Pode-se escolher um p_{cutoff} que maximiza a acurácia ou a sensibilidade ou a especificidade, mas não é possível maximizar as três métricas simultaneamente.

No gráfico a seguir podemos verificar como a sensibilidade e a especificidade variam entre si de forma inversa com a mudança do p_{cutoff}.

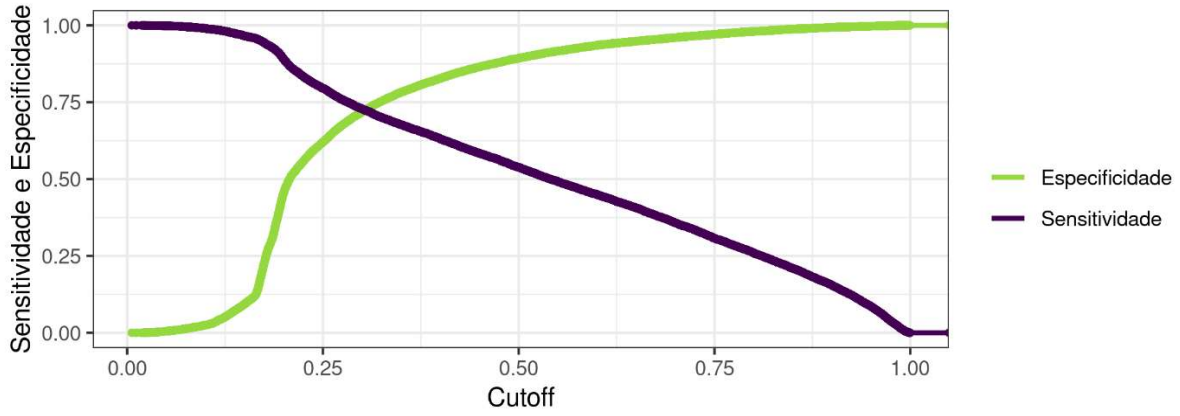

Figura 3. Sensibilidade versus Especificidade na Regressão Logística
Fonte: Resultados originais da pesquisa

A curva “Receiver Operating Caracteristic” (ROC) é um gráfico da sensibilidade em função de 1-especificidade que apresenta o formato convexo, sendo largamente utilizada para comparação entre diferentes algoritmos de ML, pois a área abaixo da curva é uma métrica de eficiência global do modelo para fins de previsão, já considerando todas possibilidades de p_{cutoff}. Por considerar todas as possibilidades de p_{cutoff} a área abaixo da curva ROC tem a vantagem de ser uma métrica independente de um p_{cutoff} em particular, o que não ocorre com a sensitividade, especificidade e acurácia.

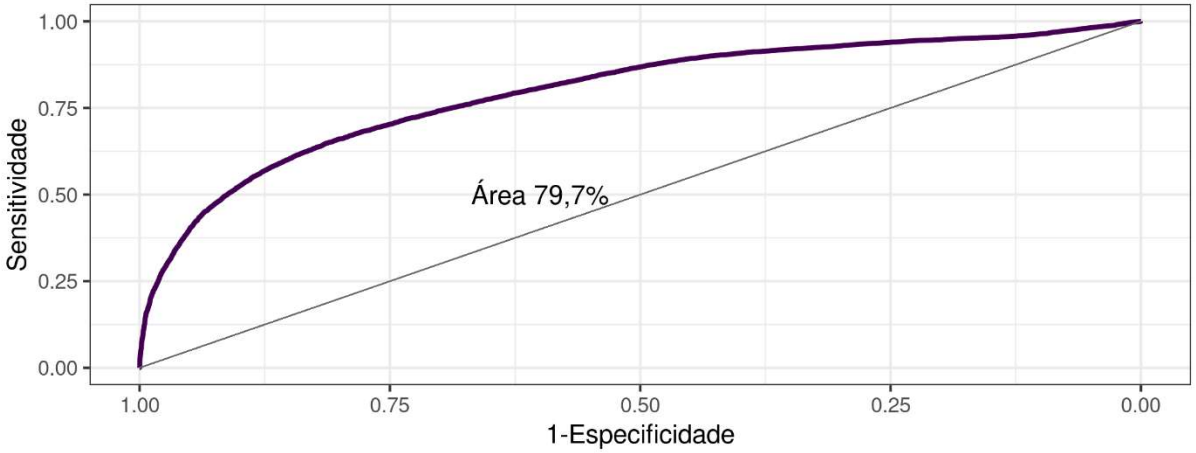

Figura 4. Curva ROC para Regressão Logística
Fonte: Resultados originais da pesquisa

A Tabela 3 apresenta as intensidades dos parâmetros estatisticamente significantes após o procedimento de stepwise.

Tabela 3. Parâmetros da Regressão Logística após o procedimento de stepwise

Variável	Est.	2.5%	97.5%	z val.	p
(Intercept)	-2.5186	-2.6254	-2.4118	-46.2280	0.0000
idade	0.0487	0.0442	0.0532	21.0977	0.0000
o1a	0.1925	0.1829	0.2020	39.3792	0.0000
o1r	0.4960	0.4725	0.5195	41.4162	0.0000
no1r	0.3720	0.2566	0.4873	6.3193	0.0000
o2a	-0.0233	-0.0347	-0.0120	-4.0443	0.0001
o2r	0.1600	0.1356	0.1843	12.8701	0.0000
no2a	-0.1113	-0.1405	-0.0821	-7.4593	0.0000
no2r	0.0870	0.0062	0.1678	2.1094	0.0349
o3a	-0.1472	-0.1590	-0.1353	-24.3408	0.0000
o3r	0.0588	0.0334	0.0841	4.5431	0.0000
no3a	-0.1520	-0.1715	-0.1325	-15.3157	0.0000
no3r	0.1189	0.0640	0.1737	4.2480	0.0000
ay2	-0.1088	-0.1645	-0.0532	-3.8319	0.0001
sexo_M	-0.1410	-0.1944	-0.0877	-5.1796	0.0000
cor_Não informada	0.1641	0.0486	0.2796	2.7845	0.0054
ec_Casado	-0.4832	-0.5994	-0.3670	-8.1513	0.0000
ec_Divorciado	-0.3827	-0.7486	-0.0168	-2.0501	0.0404
ec_Outro	0.3946	0.2501	0.5392	5.3521	0.0000

Fonte: Resultados originais da pesquisa

As colunas artificiais binárias oriundas das categorias de estado civil: Solteiro, Separado judicialmente, União Estável e Viúvo não foram estatisticamente significantes, assim como todas categorias de cor, com exceção da opção de cor “não informada”. Também não foram estatisticamente significantes as colunas no1a (quantidade de aprovações em disciplinas não obrigatórias no primeiro ano da graduação), ay1 e ay3 (quantidade de auxílios estudantis no primeiro e terceiro ano da graduação). Em termos dos coeficientes das variáveis

preditivas para evasão, o1r (quantidade de reprovações em disciplinas obrigatória no primeiro ano da graduação) obteve o maior peso, de 0,496. No segundo ano (o2r), o peso diminuiu para 0,16 e no terceiro (o3r) para 0,05. Esse padrão de diminuição das intensidades dos pesos ao longo dos anos ocorre para as demais variáveis, evidenciando assim que uma possível estratégia com o objetivo de diminuir evasão poderia ser melhorando o suporte didático aos alunos em disciplinas do início da graduação, em especial as obrigatórias, que apresentaram os maiores pesos. A maior intensidade negativa em módulo foi -0,48 para variável ec_Casado. Como a referência da categoria estado civil foi solteiro, pode-se dizer que, em termos de interpretação da regressão logística, uma mudança na variável estado civil de Solteiro para Casado diminui o logaritmo natural da chance de evasão em 48%, para divorciado essa diminuição é de 38%.

Árvore de Decisão

Iniciou-se o algoritmo de Árvore de Decisão com custo zero, que permite uma árvore com ajustes totalmente flexíveis, uma árvore livre. Essa árvore permite nós com apenas uma observação. A Tabela 4 apresenta a predição do modelo de árvore de decisão livre aplicada ao conjunto de treinamento.

Tabela 4. Tabela de confusão para Árvore de Decisão Livre nos dados de treinamento e p_{cutoff} de 50%

		Observado	
		Evadido	Concluinte
Previsto	Evadido	9648	85
	Concluinte	499	16324

Fonte: Resultados originais da pesquisa

Com acurácia de 97,8%, sensitividade de 95,1%, especificidade de 99,5% e curva ROC de 99,7%. Apesar de aparentar um ótimo ajuste, quando aplicamos o modelo de árvore de decisão livre para predição na base de teste obtemos uma nova tabela de classificação, mostrada na Tabela 5.

Tabela 5. Tabela de confusão para Árvore de Decisão Livre nos dados de teste e p_{cutoff} de 50%

		Observado	
		Evadido	Concluinte
Previsto	Evadido	1475	960
	Concluinte	1054	3150

Fonte: Resultados originais da pesquisa

Os novos valores das métricas avaliativas do modelo diminuíram intensamente. A nova acurácia ficou em 69,6%, a sensitividade em 58,3%, a especificidade 76,6% e a curva ROC 65,3%. A árvore de decisão livre se ajustou muito bem aos dados do treinamento, mas teve uma performance nos dados de testes bem pior, ou seja, o modelo de árvore de decisão livre apresentou característica de “overfitting”. O custo zero usado na criação da árvore de decisão anterior atrapalha a generalização do modelo fora dos dados de treinamento. O próximo passo é encontrar um valor ideal de custo para controlar a flexibilidade da árvore de decisão de modo que ela possa ser generalizada para fora dos dados do conjunto de treinamento.

Pode-se remover da árvore de decisão nós que atrapalham a generalização, pois dizem respeito a características exclusivas dos dados de treinamento. Em busca do custo ideal, ou seja, que evite ao máximo “overfitting”, testaremos várias possibilidades de custos. Uma técnica de validação cruzada largamente usada para procurar o custo ideal chama-se “k-fold”, que consiste, para múltiplos custos fixos, dividir a base de treino em k partes e recursivamente criar árvores de decisão com k-1 partes restantes, validando-os na parte isolado para validação, podendo esse processo ser repetido inúmeras vezes. Para cada iteração a árvore gerada é armazenada em uma grade que nos possibilitará escolher o melhor resultado a partir de uma métrica pré-estabelecida, como por exemplo, o erro relativo. Intitula-se esse processo de encontrar o custo ideal para construção da árvore de poda e no final então obtém-se uma árvore podada. A Tabela 6 apresenta 76 árvores de decisões para 76 custos fixos (CP) com os respectivos erros relativos e número de segmentações (quebra condicionais de nós) dentro das árvores:

Tabela 6. Resultado da validação cruzada para 76 árvores de decisões criadas para 76 custos fixos (CP) com respectivos erros relativos e números de segmentações

id	CP	nsplit	rel error	xerror	xstd
1	2.6914e-01	0	1.000000	1.00000	0.0078035
2	1.9710e-02	1	0.730856	0.73086	0.0072050
3	1.2565e-02	3	0.691436	0.69843	0.0071037
4	4.9276e-03	7	0.639696	0.65202	0.0069461
5	4.7305e-03	8	0.634769	0.64384	0.0069167
6	3.4493e-03	9	0.630038	0.64167	0.0069089
7	3.1536e-03	10	0.626589	0.63684	0.0068912
8	1.9710e-03	11	0.623435	0.63260	0.0068756
9	1.8725e-03	14	0.617522	0.62964	0.0068646
10	1.7739e-03	15	0.615650	0.62738	0.0068562

11	1.3797e-03	17	0.612102	0.62659	0.0068532
12	1.1826e-03	22	0.605204	0.62353	0.0068417
13	1.0841e-03	27	0.599290	0.62353	0.0068417
14	1.0512e-03	30	0.596038	0.62294	0.0068395
15	9.8551e-04	33	0.592885	0.62334	0.0068410
16	8.5411e-04	35	0.590914	0.62137	0.0068336
17	6.8986e-04	38	0.588351	0.62137	0.0068336
18	6.6522e-04	39	0.587661	0.62166	0.0068347
19	5.9131e-04	43	0.585000	0.62117	0.0068328
20	5.5846e-04	51	0.580270	0.62068	0.0068309
21	4.9276e-04	54	0.578595	0.61959	0.0068268
22	4.4348e-04	60	0.575638	0.61949	0.0068265
23	4.2706e-04	73	0.569035	0.62393	0.0068432
24	3.9421e-04	80	0.565882	0.62452	0.0068455
25	3.6957e-04	95	0.559968	0.62334	0.0068410
26	3.5478e-04	107	0.555139	0.62383	0.0068429
27	3.4493e-04	112	0.553366	0.62373	0.0068425
28	3.1536e-04	121	0.550113	0.62393	0.0068432
29	2.9565e-04	128	0.547748	0.62107	0.0068324
30	2.7594e-04	160	0.537696	0.62373	0.0068425
31	2.6280e-04	178	0.531783	0.62748	0.0068566
32	2.4638e-04	186	0.529615	0.62925	0.0068632
33	2.3652e-04	230	0.517099	0.62886	0.0068617
34	2.2995e-04	236	0.515522	0.62787	0.0068580
35	2.1681e-04	264	0.508229	0.62807	0.0068588
36	1.9710e-04	276	0.505568	0.63260	0.0068756
37	1.8068e-04	479	0.462008	0.63477	0.0068836
38	1.7739e-04	498	0.458165	0.64255	0.0069121
39	1.7246e-04	509	0.456194	0.64206	0.0069103
40	1.6895e-04	558	0.445945	0.64226	0.0069110

41	1.6425e-04	601	0.437075	0.64226	0.0069110
42	1.5768e-04	620	0.433823	0.64768	0.0069306
43	1.4783e-04	627	0.432640	0.65152	0.0069444
44	1.4079e-04	825	0.401104	0.66059	0.0069764
45	1.3797e-04	899	0.387602	0.66157	0.0069798
46	1.3140e-04	927	0.383266	0.66345	0.0069863
47	1.2671e-04	1084	0.358924	0.66601	0.0069952
48	1.2319e-04	1092	0.357741	0.66739	0.0070000
49	1.1826e-04	1125	0.353602	0.66857	0.0070040
50	1.1498e-04	1173	0.347196	0.67025	0.0070098
51	1.1263e-04	1198	0.344240	0.67064	0.0070112
52	9.8551e-05	1205	0.343451	0.69676	0.0070983
53	8.7601e-05	2477	0.210111	0.70050	0.0071104
54	8.6232e-05	2486	0.209323	0.70435	0.0071227
55	8.4473e-05	2498	0.208140	0.70435	0.0071227
56	7.8841e-05	2514	0.206761	0.71095	0.0071436
57	7.3913e-05	2581	0.201340	0.71154	0.0071455
58	7.0394e-05	2720	0.190401	0.72159	0.0071768
59	6.5701e-05	2745	0.188529	0.72297	0.0071810
60	6.1595e-05	3102	0.161723	0.72682	0.0071928
61	5.9131e-05	3129	0.160047	0.72908	0.0071997
62	5.6315e-05	3190	0.156204	0.72928	0.0072003
63	5.4751e-05	3201	0.155514	0.72928	0.0072003
64	4.9276e-05	3210	0.155021	0.75618	0.0072795
65	4.4796e-05	4545	0.085838	0.75727	0.0072826
66	4.3801e-05	4560	0.085148	0.76062	0.0072921
67	4.2236e-05	4594	0.083473	0.76348	0.0073001
68	3.9421e-05	4638	0.081502	0.76446	0.0073029
69	3.6957e-05	4764	0.075687	0.76436	0.0073026
70	3.2850e-05	4790	0.074702	0.77481	0.0073316

71	2.6878e-05	5154	0.060609	0.77609	0.0073351
72	2.4638e-05	5165	0.060313	0.77875	0.0073424
73	1.9710e-05	5243	0.058244	0.77954	0.0073445
74	1.6425e-05	5258	0.057948	0.78191	0.0073509
75	1.2319e-05	5276	0.057653	0.78200	0.0073512
76	0.0000e+00	5284	0.057554	0.78200	0.0073512

Fonte: Resultados originais da pesquisa

Na representação gráfica da Tabela 6 disponível na Figura 5 pode-se verificar o comportamento do erro relativo em função da variação do custo e identificar a região que os menores erros relativos.

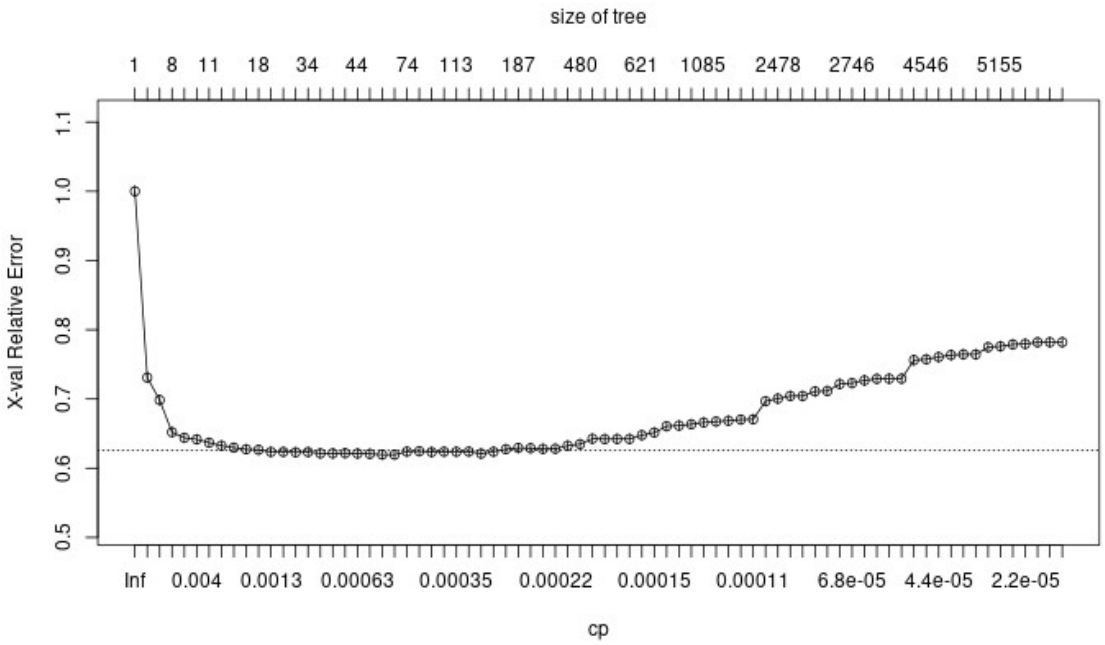

Figura 5. Erros relativo da validação cruzada para 76 árvores de decisões criadas para 76 custos fixos
Fonte: Resultados originais da pesquisa

O custo que apresentou menor erro relativo foi 0,0004434808. Esse foi o custo usado da árvore podada. A Tabela 7 apresenta a tabela confusão da Árvore de Decisão Podada aplicada na base de treinamento.

Tabela 7. Tabela de confusão para Árvore de Decisão Podada nos dados de treinamento e p_{cutoff} de 50%

		Observado	
		Evadido	Concluinte
Previsto	Evadido	5701	1395
	Concluinte	4446	15014

Fonte: Resultados originais da pesquisa

A Árvore de Decisão Podada nos dados de treinamento apresentou acurácia de 78,0%, sensitividade de 56,2%, especificidade de 91,5% e curva ROC de 80,6%. Aplicando-se a Árvore de Decisão Podada aos dados separados inicialmente para teste gerou-se a tabela de confusão apresentada na Tabela 8.

Tabela 8. Tabela de confusão para Árvore de Decisão Podada nos dados de teste e p_{cutoff} de 50%

		Observado	
		Evadido	Concluinte
Previsto	Evadido	1388	420
	Concluinte	1141	3690

Fonte: Resultados originais da pesquisa

A Árvore de Decisão Podada nos dados de teste apresentou acurácia de 76,5%, sensitividade de 54,9%, especificidade de 89,8% e curva ROC de 79,9%. Considerando a acurácia e a área da curva ROC, há muito menos discrepância entre a base de treinamento e teste no modelo de Árvore de Decisão Podada do que no modelo de Árvore de Decisão Livre, forte indicativo que árvore podada sofre menos do efeito de "overfitting" que a árvore livre. A Figura 6 apresenta o gráfico comparando as quatros curvas ROC geradas para árvore de decisão livre e podada.

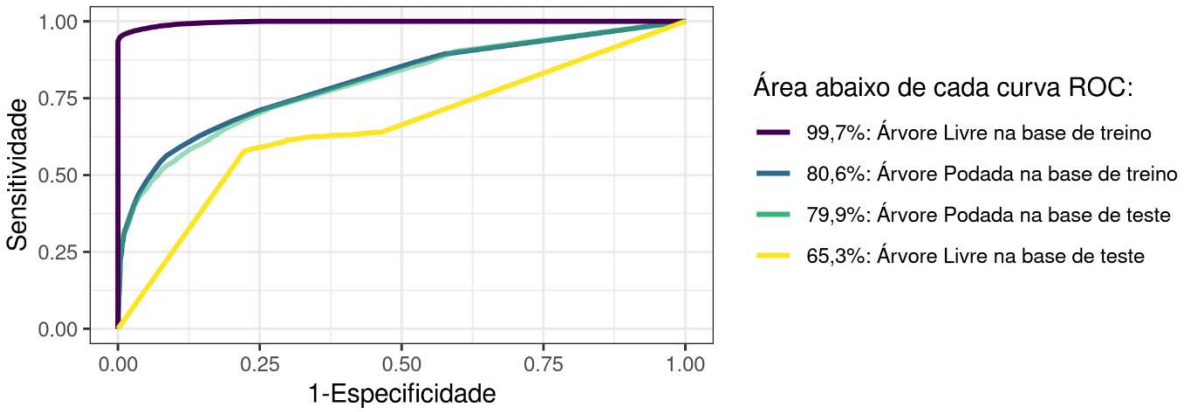

Figura 6. Curvas ROC para modelos de Árvore de Decisão Livre e Podada
Fonte: Resultados originais da pesquisa

Diferente da Regressão Logística, a Árvore de Decisão não nos oferece um procedimento para avaliação da significância estatística de cada variável explicativa, mas oferece uma métrica chamada importância, calculada para cada variável, que é conceitualmente um tipo de medida que mede a diminuição da impureza em cada nó da árvore e que pode ser interpretada também com base na redução correspondente da capacidade preditiva do modelo quando a variável em questão é removida. Segue-se a importância para cada variável obtidas na árvore podada:

Tabela 9. Importância das variáveis explicativas na Árvore de Decisão Podada

Variável	Importância
o1r	2036,6
o3a	1214,9
o2a	935,5
o1a	903,8
o2r	778,5
o3r	629,6
no3a	600,4
no2a	228,1
no2r	151,3
no3r	147,7
no1r	102,8
idade	96,3
cor	36,6
ec	8,0
no1a	3,2
ay2	3,0
ay1	2,4
sexo	1,5
ay3	1,5

Fonte: Resultados originais da pesquisa

As variáveis ay1 e ay3, quantidade de auxílios estudantis no primeiro e terceiro ano da graduação, tiveram importância baixa perante as demais, resultado que concorda conceitualmente com o da Regressão Logística, que considerou essas variáveis não estatisticamente significantes. A variável o1r, quantidade de reprovações em disciplinas obrigatória na graduação, teve a maior importância, concordando novamente com a Regressão Logística, que deu maior peso no coeficiente de o1r.

Random Forest

Inicialmente construiu-se o modelo RF, com o pacote “randomForest”, definindo o hiperparâmetro “ntree” de 50, pois desejou-se 50 árvores em cada iteração. O hiperparâmetro “mtry”, que define o número máximo de variáveis que serão amostradas, como 19, pois optou-se testar todas combinações de seleções aleatórias de variáveis, mesmo incluíndo-se todas que temos disponíveis, ou seja, 19. A Tabela 10 apresenta a tabela de confusão da aplicação da RF nos dados de treinamento.

Tabela 10. Tabela de confusão para RF nos dados de treinamento e p_{cutoff} de 50%

		Observado	
		Evadido	Concluinte
Previsto	Evadido	9862	45
	Concluinte	285	16364

Fonte: Resultados originais da pesquisa

RF nos dados de treinamento apresentou acurácia de 98,8%, sensitividade de 97,2%, especificidade de 99,7% e curva ROC de 99,6%. Aplicando o modelo RF na base de teste obtém-se a Tabela 11.

Tabela 11. Tabela de confusão para RF nos dados de teste e p_{cutoff} de 50%

		Observado	
		Evadido	Concluinte
Previsto	Evadido	1497	655
	Concluinte	1032	3455

Fonte: Resultados originais da pesquisa

Assim como ocorreu na Árvore de Decisão Livre, a capacidade preditiva caiu drasticamente nos dados de teste na RF, com 74,6% de acurácia, 59,2% de sensitividade, 84,1% de especificidade e 80,3% de área abaixo da curva ROC. A discrepância entre o valor 99,6% para 80,3% de área abaixo da curva ROC entre a base de treinamento e a base de teste, indicam a ocorrência de "overfitting". Aplicou-se novamente o procedimento de validação cruzada "k-fold", com 4 folds, duas repetições e com o auxílio do método "train", do pacote "caret", versão 6.0.86 e indicação de "grid" no parâmetro "search". Como buscou-se o maior valor da curva ROC, especificou-se a utilização da função "twoClassSummary" no parâmetro "summaryFunction". A Figura 7 apresenta o gráfico dos valores das áreas ROC em função das quantidades de variáveis preditivas e permite identificar a quantidade de variáveis que geraram a maior área ROC, no caso, 8 variáveis, com 80,1% de área ROC.

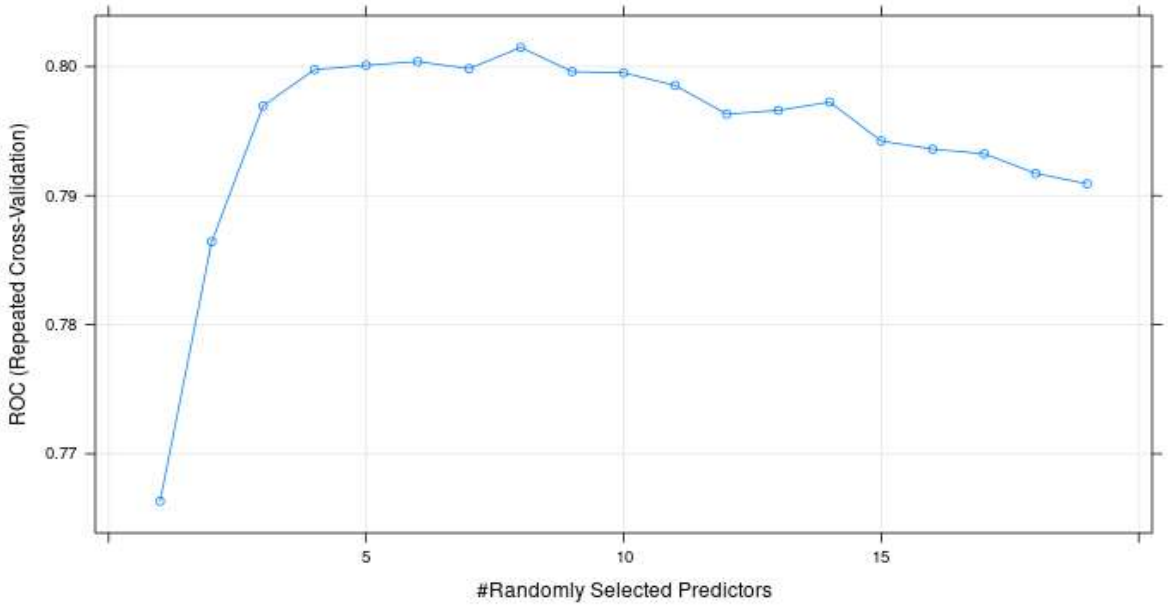

Figura 7. Valores das áreas ROC para RF em função de diferentes quantidades de variáveis preditivas
Fonte: Resultados originais da pesquisa

Os valores das áreas ROC das tentativas apresentada no gráfico da Figura 7 podem ser acessados na Tabela 12 com o acréscimo da sensitividade e especificidade para cada tentativa, pois dependendo dos requisitos do projeto, poder-se-iam escolher essas métricas para definir o melhor modelo, e não necessariamente a área da curva ROC.

Tabela 12. Valores da área ROC, sensitividade e especificidade para RF com amostragem nas variáveis explicativas variando de 1 a 19.

mtry	ROC	Sens	Spec
1	0.7663323	0.9489916	0.3174859
2	0.7864233	0.9272045	0.4831005
3	0.7969556	0.9081603	0.5269538
4	0.7997648	0.8939604	0.5463681
5	0.8001000	0.8854291	0.5530694
6	0.8003834	0.8784813	0.5578981
7	0.7998417	0.8755866	0.5647971
8	0.8014820	0.8701628	0.5688874
9	0.7996065	0.8673898	0.5668165
10	0.7995124	0.8654397	0.5696760
11	0.7985322	0.8633065	0.5687882
12	0.7963174	0.8617525	0.5700200
13	0.7966045	0.8589493	0.5714988
14	0.7972477	0.8552925	0.5774124
15	0.7942307	0.8497776	0.5775103
16	0.7936016	0.8507217	0.5780030
17	0.7932450	0.8463647	0.5805173
18	0.7917192	0.8439575	0.5783978
19	0.7909139	0.8408799	0.5786939

Fonte: Resultados originais da pesquisa

A Tabela 12 apresenta o resultado, em forma de tabela de confusão, do modelo Random Forest após a validação cruzada, apelidado de “Random Forest Gridsearch” (RFG), aplicado na base de treinamento:

Tabela 12. Tabela de confusão para Random Forest Gridsearch nos dados de treinamento e p_{cutoff} de 50%

		Observado	
		Evadido	Concluinte
Previsto	Evadido	8695	109
	Concluinte	1452	16300

Fonte: Resultados originais da pesquisa

RFG aplicado aos dados de treinamento apresentou acurácia de 94,1%, sensitividade de 85,7%, especificidade de 99,4% e curva ROC de 95,4%. A Tabela 13 apresenta a tabela de confusão para o RFG no conjunto de dados de testes:

Tabela 13. Tabela de confusão para Random Forest Gridsearch nos dados de teste e p_{cutoff} de 50%

		Observado	
		Evadido	Concluinte
Previsto	Evadido	1471	543
	Concluinte	1058	3567

Fonte: Resultados originais da pesquisa

RFG aplicado aos dados de teste apresentou acurácia de 75,9%, sensitividade de 58,2%, especificidade de 86,8% e curva ROC de 80,4%. Diferente da árvore de decisão, o procedimento de validação cruzada no caso da Random Forest não minimizou os efeitos o “overfitting”. A Figura 8 apresenta o gráfico com as quatros curvas ROC para as RF obtidas:

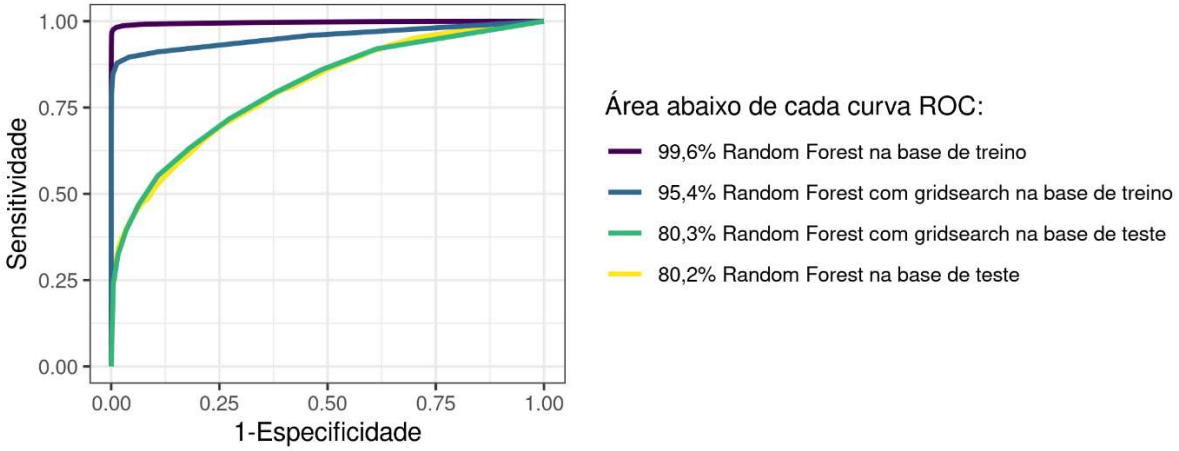

Figure 8. Curvas ROC para modelos de Random Forest e Random Forest Gridsearch
Fonte: Resultados originais da pesquisa

Gradient Boosting

No “Gradient Boostring” (GB) construiu-se o modelo diretamente com validação cruzada configurando como métrica de escolha do melhor modelo a área da curva ROC. A Tabela 14 apresenta a tabela de confusão para GB no conjunto de dados de treinamento.

Tabela 14. Tabela de confusão para Gradient Boosting nos dados de treinamento e p_{cutoff} de 50%

		Observado	
		Evadido	Concluinte
Previsto	Evadido	5661	1365
	Concluinte	4486	15044

Fonte: Resultados originais da pesquisa

GB aplicado aos dados de treinamento apresentou acurácia de 77,9%, sensitividade de 55,8%, especificidade de 91,7% e curva ROC de 83,8%. A Tabela 15 apresenta a tabela de confusão para o GB no conjunto de dados de testes.

Tabela 15. Tabela de confusão para Gradient Boosting nos dados de teste e p_{cutoff} de 50%

		Observado	
		Evadido	Concluinte
Previsto	Evadido	1377	358
	Concluinte	1152	3752

Fonte: Resultados originais da pesquisa

GB aplicado aos dados de teste apresentou acurácia de 77,3%, sensitividade de 54,4%, especificidade de 91,3% e curva ROC de 82,3%. O gráfico apresentado na Figura 9 mostra que a performance do GB foi similar tanto no conjunto de testes como no de treinamento, explicitando assim pouco efeito de “overfitting”.

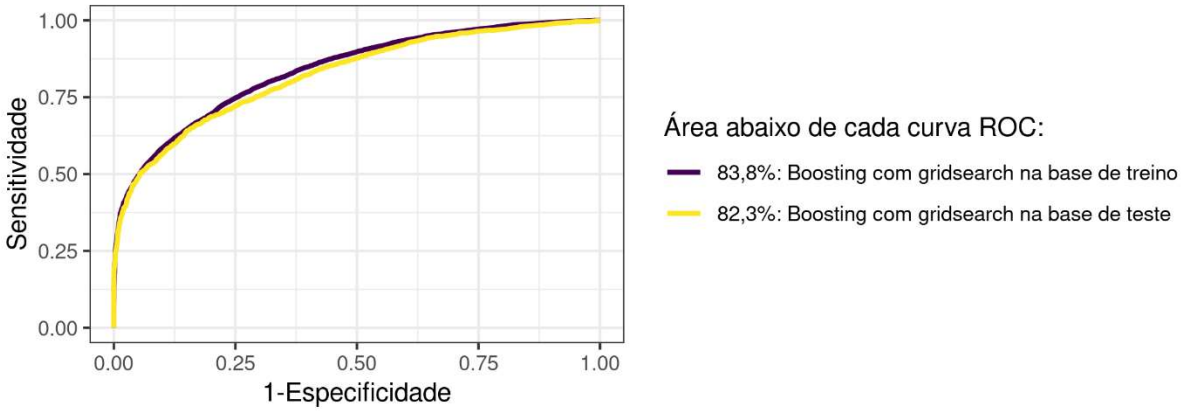

Figure 9. Curvas ROC para o modelo Gradient Boosting
Fonte: Resultados originais da pesquisa

Rede Neural

Assim como para Regressão Logística, as colunas categóricas foram transformadas em binárias para cada categoria na preparação dos dados para o algoritmo de Redes Neurais (RN). Além disso, com intuito de diminuir a intensidade dos ruídos, os valores de todas colunas foram padronizados segundo mínimo/(máximo-mínimo). Como no caso do GB, aplicou-se diretamente o procedimento de validação cruzada definindo como métrica de escolha do melhor modelo a área abaixo da curva ROC. O pacote usado, nnet, só permite a criação de uma RN com apenas uma camada escondida. A Tabela 16 apresenta o resultado em forma de tabela de confusão do modelo de RN aplicado no conjunto de dados de treinamento.

Tabela 16. Tabela de confusão para RN nos dados de treinamento e p_{cutoff} de 50%

		Observado	
		Evadido	Concluinte
Previsto	Evadido	5461	1419
	Concluinte	4686	14990

Fonte: Resultados originais da pesquisa

A RN aplicada aos dados de treinamento apresentou acurácia de 77,3%, sensitividade de 54,5%, especificidade de 91,3% e curva ROC de 82,5%. A Tabela 17 apresenta a classificação da RN aplicada na base de teste:

Tabela 17. Tabela de confusão para RN nos dados de teste e p_{cutoff} de 50%

		Observado	
		Evadido	Concluinte
Previsto	Evadido	1386	356
	Concluinte	1143	3754

Fonte: Resultados originais da pesquisa

A RN aplicada aos dados de teste apresentou acurácia de 77,4%, sensitividade de 54,8%, especificidade de 91,3% e curva ROC de 81,8%. A Figura 10 apresenta o gráfico das curvas ROC para base de treinamento e base de teste, e assim como no caso do GB, observa-se que a performance foi similar tanto no conjunto de testes como no de treinamento, apresentado pouco efeito de “overfitting”.

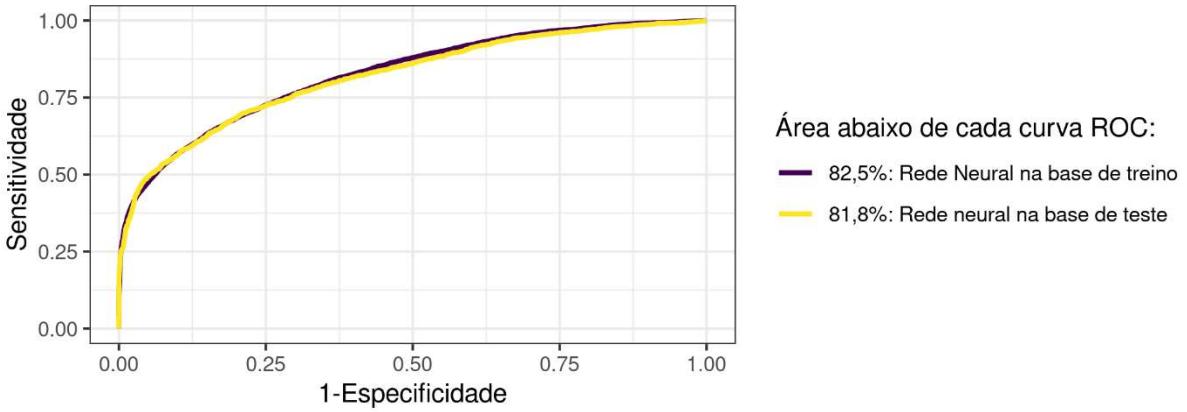

Figura 10. Curvas ROC para o modelo de Rede Neural
Fonte: Resultados originais da pesquisa

As métricas e performances preditivas encontradas nos 5 modelos selecionados, em particular os valores das áreas abaixo das curvas ROC estão de acordo com a literatura científica correlata. Digiampietri et al. (2016) utilizando o algoritmo “Rotation Forest” obtiveram 84% na área da curva ROC ao estudarem predição de evasão universitária em cursos de exatas da graduação. Bezerra et al. (2016) obteve área da curva ROC máxima de 69% aplicando diferentes modelos no contexto de predição de evasão de alunos em escolas secundárias brasileiras.

Considerações Finais

A taxa de evasão universitária é um dos sinais preocupante que indicam necessidade de intervenções para a melhoria da qualidade de educação. A capacidade de predizer indivíduos com altas chances de evasão ou identificar fatores relevantes que levam à evasão, possibilita gestores educacionais a atuarem de forma mais fundamentadas em suas políticas de permanência. Este trabalho avaliou cinco modelos de classificação e identificou, nesta ordem, os modelos com maior capacidade preditiva, segundo a área ROC: “Gradient Boosting” (82,3%), Rede Neural (81,8), Random Forest (80,3), Árvore de Decisão (79,9%) e Regressão Logística (79,7%), sendo que não foi possível minimizar o efeito de “overfitting” na Random Forest. A Regressão Logística e a Árvore de Decisão, apesar da baixa capacidade

preditiva perante os demais modelos, permitiram identificar mais facilmente variáveis que não contribuíram significativamente para evasão universitária, como foi o caso da quantidade de auxílios estudantis recebida pelos discentes, em que ambos modelos concordaram que essa variável não contribuiu significativamente para explicar a evasão. Além disso, identificou-se que o progresso acadêmico no primeiro ano de graduação tem maior peso na evasão.

Por fim, é importante pontuar que o ferramental provido por técnicas de ML, como as usadas neste trabalho, contribui com informações valiosíssimas para entendimento da evasão, porém se constitui apenas de um suporte inicial para o enfrentamento dessa problemática, que deve envolver outras áreas do conhecimento e considerações de aspectos que não são passíveis de inclusão em modelos de ML.

Referências

Breiman, L. 2001. Random Forests. Machine Learning 45: 5-32.

Breiman L.; Friedman J. H.; Olshen R. A.; Stone, C. J. 1984. Classification and Regression Trees. Wadsworth. 1ed. Routledge. Oxfordshire, United Kingdom.

Digiampietri, L.A.; Nakano, F; Lauretto, M.S. 2016. Mineraçao de dados para identificaçao de alunos com alto risco de evasao: Um estudo de caso. Revista de Graduação USP 1: 17-23.

Fávero, L. P.; & Belfiore, P. 2017. Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®. 1ed. Elsevier Brasil. São Paulo, SP, Brasil.

Ferguson, R. 2012. Learning analytics: drivers, developments and challenges. International Journal of Technology Enhanced Learning 4: 304-317.

Filho, J.A.B.L; Silveira, I.F. 2021. Detecção precoce de estudantes em risco de evasão usando dados administrativos e aprendizagem de máquina. Revista Ibérica de Sistemas e Tecnologias de Informação 40: 480-495.

Friedman J.; Hastie T.; Tibshirani R. 2000. Additive logistic regression: a statistical view of boosting. The Annals of Statistics 28: 337-407.

Jesus, H.O.; Rodriguez, L.C; Junior, A.D.O.C. 2021. Predição de Evasão Escolar na Licenciatura em Computação. Revista Brasileira de Informática na Educação 29: 255-272.

Mduma, N.; Kalegele, K.; Machuve, D. 2019. Machine learning approach for reducing students dropout rates. International Journal of Advanced Computer Research 9: 156-169.

Rafiq, M.A.; Rabbi, A.M; Ahammad, R., 2021. A data science approach to Predict the University Students at risk of semester dropout: Bangladeshi University Perspective. 5th International Conference on Trends in Electronics and Informatics: 1350-1354.

Ripley, B. D. 2007. Pattern Recognition and Neural Networks. 1ed. Cambridge University Press. Cambridge, United Kingdom

Rodrigues, L.M.; Moraes, E.A.P.; Pinto, R.C. 2021. Análise preditiva para identificação de alunos suscetíveis à evasão escolar. Brazilian Journal of Development 7: 71631-71643.

Universidade de São Paulo [USP]. 2022. Anuário Estatístico. Disponível em: <https://uspdigital.usp.br/anuario/>. Acesso em 19 mai. 2022.