GOVERNO DO ESTADO DO ESPÍRITO SANTO
SECRETARIA DE ECONOMIA E PLANEJAMENTO – SEP
INSTITUTO JONES DOS SANTOS NEVES – IJSN

RELATÓRIO



Preditor do Abandono Escolar

Estudos Educacionais

Instituto Jones dos Santos Neves

GOVERNO DO ESTADO DO ESPÍRITO SANTO

Julho de 2023

Instituto Jones dos Santos Neves

Preditor do Abandono Escolar: estudos educacionais.

Vitória, ES, 2023. 34p. il. Tab. (Relatório).

1.Educação. 2. Abandono escolar. 3. Estudos educacionais. 5. Espírito Santo (Estado).

I. Pereira, Guilherme Armando de Almeida (Bolsista). II. Demura, Kiara de Deus. III. Título.

As opiniões emitidas são exclusivas e de inteira responsabilidade do (s) autor (es), não exprimindo necessariamente, o ponto de vista do Instituto Jones dos Santos Neves ou da Secretária de Estado de Economia e Planejamento do governo do Estado do Espirito Santo.

GOVERNO DO ESTADO DO ESPÍRITO SANTO
José Renato Casagrande

VICE-GOVERNADORIA
Ricardo Ferraço

SECRETARIA DE ECONOMIA E PLANEJAMENTO – SEP
Álvaro Rogério Duboc Fajardo

INSTITUTO JONES DOS SANTOS NEVES – IJSN

Diretor Presidente
Pablo Silva Lira

Diretoria de Estudos e Pesquisas
Pablo Medeiros Jabor

Diretoria de Integração e Projetos Especiais
Antonio Ricardo Freislebem da Rocha

Diretoria de Gestão Administrativa
Katia Cesconeto de Paula

Coordenação Geral
Kiara de Deus Demura

Elaboração
Guilherme Armando de Almeida Pereira (Bolsista Fapes)
Kiara de Deus Demura

Bibliotecária
Rosana Mariano Chagas

Instituto Jones dos Santos Neves

Sumário

1. Introdução 5

2. Revisão de literatura 7

3. Metodologia 11

3.1. Regressão logística 11

3.2. Seleção de variáveis 13

3.2.1. Forward/backward stepwise selection 13

3.2.2. Shrinkage methods 13

3.3. Métricas 15

3.3.1. Métricas para avaliação das previsões 15

3.3.2. Métricas para avaliação do ajuste do modelo 16

3.4. Banco de dados 18

4. Estudo de casos 19

4.1. Banco de dados 19

4.2. Modelo de referência 20

4.3. Resultados 21

4.3.1. Caso 2 – 1º ano do ensino médio em 2019 21

4.3.2. Caso 2 – 1º ano do ensino médio em 2020 29

5. Conclusões 31

Referências 33

1. Introdução

É de conhecimento comum que a educação é um dos pilares fundamentais para o desenvolvimento social e econômico de uma nação. Os impactos da educação sobre os mais diversos parâmetros de uma sociedade vêm sendo tema de diversos trabalhos empíricos ao longo dos anos. Psacharopoulos e Patrinos (2018) avaliaram 1.120 estimativas provenientes de diferentes trabalhos científicos realizados em 139 países distintos. Em linhas gerais, os autores concluem que o retorno privado (individual) de um ano extra de educação é de aproximadamente 9% ao ano. Além destes, os autores identificam que os retornos sociais da educação também são positivos e subestimados, uma vez que há dificuldade do cômputo das externalidades positivas e dos benefícios não monetários tais como a redução da mortalidade. Outras características encontradas na revisão sistemática de literatura que merecem destaque são: (i) mulheres possuem taxas de retornos maiores; (ii) países de baixa e média renda possuem taxas de retorno maiores; (iii) os retornos são maiores nos setores privados, o que suporta a hipótese que a educação impacta a produtividade da economia.

Adicionalmente, Burgess (2016) advoca ainda que os benefícios da educação possuem um alcance maior do que se pode imaginar uma vez que estão associados também a melhoria da saúde, redução da gravidez precoce, redução do desemprego, redução das taxas de criminalidade etc.

O elevado abandono escolar[1] é um dos principais desafios que gestores e educadores encontram normalmente no dia a dia. Altas taxas de abandono estão associadas à baixos salários, maior grau de vulnerabilidade durante crises econômicas, uma pior saúde, aumento do envolvimento com o crime e encarceramento e aumento das taxas de mortalidade. Do ponto de vista social, os impactos das altas taxas de abandono passam por uma redução da arrecadação de impostos, aumento dos gastos com programas sociais e dificuldade em atrair negócios que demandam trabalhadores altamente qualificados (PARR; BONITZ, 2015; WOOD et al., 2017).

[1] O abandono escolar ocorre quando determinado aluno passa a não frequentar a escola durante o andamento do ano letivo.

Segundo o módulo Educação[2] da Pesquisa Nacional por Amostra de Domicílios Contínua (PNADC) divulgada pelo Instituto Brasileiro de Geografia e Estatística (IBGE), aproximadamente 20,2% dos indivíduos de 14 a 29 anos no Brasil não completaram o ensino médio. Isso representa 10,1 milhões de pessoas que abandonaram a escola antes do térmico ou nunca a frequentou. De acordo com a pesquisa, a principal razão por terem abandonado ou nunca frequentado a escola é atribuída majoritariamente a necessidade de trabalhar. O abandono não atinge os estudantes de forma linear. Este é mais frequente entre homens (58,3%) do que entre mulheres (41,7%). Tendo em vista a cor ou raça, a discrepância entre grupos é ainda maior uma vez que pretos ou pardos constituem 71,7% total de abandonos enquanto em brancos essa taxa é de 27,3%.

As razões pelas quais ocorrem o abandono é bem documentada na literatura e estão associadas principalmente a características socioeconômicas dos alunos, questões familiares, desempenho escolar, características das escolas e professores (GIUBERTI et al., 2018).

Neste contexto, identificar alunos em situação de abandono é fundamental para que gestores e educadores sejam capazes de formular ações de modo a mitigar esse problema. Diversos modelos para previsão de abandono vêm sendo propostos na literatura, dentre estes pode-se citar, professores (GIUBERTI et al., 2018; BAYER et al., 2012; ADELMAN et al., 2018; WOOD et al., 2017; MARTINHO; NUNES; MINUSSI, 2013). Tais modelos são construídos com base nas mais variadas ferramentas a exemplo de modelos de inteligência artificial, modelos estatísticos, *growth mixture modeling* entre outros. É importante destacar, contudo, que não há a prevalência de uma técnica sobre a outra. O modelo mais adequado para cada caso depende de um conjunto de fatores tais como janela temporal analisada, estrutura dos dados, tipos de variáveis, faixa etária dos estudantes etc.

Nesse sentido, este estudo busca apresentar os modelos estudados e testados para predizer o abandono escolar no âmbito da parceria “Estudos Educacionais” entre o Instituto Jones dos Santos Neves (IJSN), a Secretaria de Estado da Educação (SEDU) e a

[2] https://biblioteca.ibge.gov.br/visualizacao/livros/liv101736_informativo.pdf

Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (FAPES). Assim, os dados compreenderam os estudantes da rede estadual de ensino do Espírito Santo, sendo este produto desenvolvido com o intuito de auxiliar a gestão educacional realizada pela SEDU.

Isto posto, este relatório está organizado da seguinte forma. Na Seção 2 apresentamos a revisão de literatura sobre modelos de previsão relacionadas ao abandono evasão e reprovação escolar. A Seção 3 traz as ferramentas matemáticas utilizadas nesse relatório, isto é, as descrições dos modelos e métricas para avaliação. Em seguida, na Seção 4 são apresentados os resultados, as análises dos modelos estimados e as correspondentes performance em termos preditivos. Por fim, na Seção 5 são feitas as considerações finais e propostas de novos estudos.

2. Revisão de literatura

A tarefa de predizer a situação que antecede o abandono de determinado estudante não é simples pois este resulta de um conjunto de características socioeconômicas, estruturais, escolares, históricas e psicológicas de cada indivíduo. Diferentes abordagens para entender esse problema têm sido desenvolvidas ao longo dos anos, sendo as mais recorrentes as baseadas em modelos estatísticos (*statistical learning*) e de aprendizado de máquinas (*machine learning*).

Embora possa parecer uma tarefa única, a construção de um modelo de previsão de abandono deve atentar para as especificidades de cada momento do ciclo escolar. Não é razoável supor que as razões do abandono de uma criança no começo de sua vida escolar sejam as mesmas de uma adolescente que está a iniciar o ensino médio.

De acordo com Sara et al. (2015) comparar modelos de previsão de abandono/evasão escolar não é uma tarefa simples, uma vez que os trabalhos existentes divergem em relação aos níveis educacionais analisados, informações dos estudantes, características socioeconômicas, infraestrutura das escolas, critérios de avaliação, horizonte de previsão etc. Por esses motivos, diversos modelos são propostos na literatura onde os

mais variados níveis educacionais e as mais variadas informações sobre os estudantes são empregadas.

Sara et al. (2015) desenvolveram um modelo de classificação para predizer se estudantes iram abandonar o ensino médio (*high-school*) na Dinamarca. Foram considerados 72.598 estudantes que haviam completado ao menos os primeiros seis meses das correspondentes séries e o objetivo era predizer o abandono para os próximos 3 meses subsequentes. Os autores utilizaram as principais técnicas de aprendizado estatístico e de máquinas tais como *support vector machines* (SVM), *random forest*, *classification and regression tree (CART)* e *naive Bayes*. O modelo r*andom forest obteve* a melhor acurácia (93,5%) e a melhor *area under the receiver operating characteristc curve* (AUC).

Costa et al. (2017) empregaram diferentes técnicas de *educational data mining* (EDM) tais como redes neurais, *decision tree*, SVM e *naives bayes*. O intuído era identificar os estudantes prováveis à reprovação em cursos introdutórios de programação ofertados nas universidades públicas onde os autores lecionam. Adicionalmente, os autores empregaram técnicas de pré-processamento de dados pois identificaram que quando o número de elementos em cada classe é muito destoante, os algoritmos de previsão tendem a aprender mais com a base na classe predominante. As estimativas foram realizadas em dois cursos distintos: um curso à distância e o outro presencial. Ao compararem o efeito das técnicas de pré-processamento sobre o desempenho dos modelos, os resultados encontrados foram ambíguos, uma vez que melhorou o desempenho apenas para um caso. Por fim, os autores concluem que as técnicas empregadas são importantes ferramentas para identificar reprovações e que o SVM foi superior com base nas métricas analisadas.

Bayer et al. (2012) propuseram modelos de previsão para abandono e reprovação para os alunos dos cursos de computação aplicada da Universidade Masaryk baseados em modelos de *machine learning*. O banco de dados empregado continha, além das informações básicas sobre estudantes e professores, informações relativas ao acesso de estudantes às plataformas computacionais ofertadas pelos professores durante os

cursos, e-mails trocados, arquivos compartilhados, fóruns, etc. Baseados nessas informações comportamentais, os autores propuseram a elaboração de novas variáveis utilizando *social network analysis* que, posteriormente, foram utilizadas nos modelos de previsão.

Dekker, Pechenizkiy e Vleeshouwers (2009) utilizaram métodos de *data mining* para predizer o abandono do curso de engenharia elétrica na Holanda após o primeiro semestre ou mesmo antes deles começarem o programa. A acurácia[3] variou entre 75% e 80%.

Sandoval-Palis et al. (2020) avaliaram o desempenho de modelos de regressão logística e redes neurais artificiais para previsão do abandono nos cursos de nivelamento da *Escuela Politécnica Nacional*. O banco de dados disponível consistia em informações acadêmicas e socioeconômicas sobre 2097 alunos. O modelo de redes neurais apresentou melhor desempenho preditivo.

Adelman et al. (2018) empregaram modelos de probabilidade linear para desenvolver um sistema de previsão de abandono para a Guatemala e Honduras com base em informações sociodemográficas, escola, município etc. Em seu trabalho, os autores tiveram como foco as séries que apresentavam maior número de abandono. Os resultados indicam que modelos simples muita das vezes podem se comportar de maneira satisfatória. Além disso, os autores ilustram como a inclusão de efeitos fixos melhora o desempenho preditivo nos casos analisados no artigo.

Jiménez-Gómez et al. (2015) testaram diferentes técnicas de *data mining* para previsão de alunos em situação de reprovação de uma escola em Córdoba, Espanha, com base em informações institucionais tais como registros acadêmicos, dados pessoais, presença, etc. Diferentemente dos trabalhos anteriores, o modelo proposto era estimado a cada semestre de modo que com o passar dos semestres, novas informações eram incorporadas ao banco de dados. Um dos objetivos era determinar o melhor momento para se realizar a previsão. Além disso, os resultados indicam que os modelos

[3] Ver item 3.3.1

empregados apresentam desempenho satisfatório, embora não há um modelo que se destaque em termos de performance.

Márquez-Vera et al. (2016) propuseram uma metodologia baseada em programação genética, a qual considera o fato de existir classes com diferentes números de observação para previsão de abandono. Além disso, os autores exploraram a dinâmica do problema de modo que o modelo é estimado e reestimado diversas vezes ao longo do ano. O modelo foi testado em 419 estudantes matriculados na *Academic Unit Preparatoria* na Autonomous University of Zacatecas. Os resultados indicam que a metodologia proposta é válida para a previsão do abandono no ensino médio.

Cisterna, Prats e Muñoz (2017) desenvolveram uma ferramenta baseada nas principais ferramentas de aprendizado de máquinas (regressão logística, *naive* bayes, SVM, *random forest* e *adaptive boosting*) para previsão de evasão, notas e recomendações de curso personalizadas. O sistema foi testado em 4.434 estudantes de direito, matemática e ciência da computação da Universidade de Barcelona entre os anos de 2009 e 2014. Assim como em outros casos, não é possível estabelecer um modelo geral, já que para diferentes cursos cada técnica determinada técnica é mais adequada. Para o curso de direito, *random forest* e *adaptive boosting* foram os melhores. Por outro lado, para matemática e ciência da computação, a regressão logística e *naive bayes* foram os que apresentaram os melhores desempenhos.

Alguns trabalhos também foram desenvolvidos em âmbito nacional. Do Nascimento et al. (2018) e Do Nascimento, Fagundes e De Souza (2021) empregaram métodos de aprendizado estatístico para predizer índices de abandono nas escolas com as informações públicas fornecidas pelo INEP[4]. Os resultados apontam para a utilização do SVM como sendo a mais adequada para os dados analisados.

Da Cunha, Moura e Analide (2016) utilizaram análise de agrupamento (*clustering analysis*) para detectar quais as principais razões que estariam associadas ao abandono e reprovação dos estudantes no Instituto Federal de Educação, Ciência e Tecnologia do

[4] INEP (2019). Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. http:// portalinep. gov. br/. Acesso em: 11 dez. 2022

Rio Grande do Norte (IFRN). Dessa forma, o objetivo era traçar o perfil do aluno propenso a abandoar o curso. O banco de dados consistia em registros acadêmicos, além de questionários preenchidos por estudantes e professores. Os resultados indicam que estudantes oriundos de escolas públicas, com renda familiar menor que um salário-mínimo, com baixo desempenho escolar, que moram com seus pais, que estão desempregados ou são menores de idade constituem o perfil de alunos que acabam por não concluírem o ciclo acadêmico.

Martinho, Nunes e Minussi (2013) desenvolveram um modelo Fuzzy-ARTMAP Neural Network com base em informações acadêmicas e socioeconômicas dos estudantes dos cursos de tecnologia do Instituto Federal do Mato Grosso (IFMT). Os resultados indicam uma acurácia de aproximadamente 85% e uma taxa de acerto para abandono em torno de 92%.

3. Metodologia

Nesta seção apresentamos as ferramentas matemáticas e estatísticas utilizadas ao longo desse trabalho. Inicialmente introduziremos o modelo de regressão logística e os métodos de seleção de variáveis. Por fim, algumas métricas para avaliação dos modelos e procedimentos de reamostragem usados para calibragem dos modelos são descritos.

3.1. Regressão logística

Antes de introduzirmos a formulação do modelo de regressão logística, iremos definir a variável de resposta y. Neste caso, adotaremos a seguinte definição:

$$y_i = \begin{cases} 0, & caso\ o\ aluno\ i\ não\ abandone\ a\ escola. \\ 1, & caso\ o\ aluno\ i\ abandone\ a\ escola. \end{cases}$$

Isto posto, neste trabalho, iremos empregar a regressão logística para o caso binário onde o interesse primordial é estabelecer $Prob(y_i = 1/\boldsymbol{X_i})$. Em outras palavras, o objetivo é estimar a probabilidade de abandono de determinado estudante

condicionada à um conjunto de variáveis explicativas, doravante representadas por $\boldsymbol{X}$. Assim, o modelo identifica padrões existentes em diversas variáveis tais como condições estruturais da escola, socioeconômicas, familiares e de desempenho dos estudantes, e infere a probabilidade de abandono do mesmo.

Formalmente, a regressão logística (*logit*) é dada por:

$$Prob(y = 1/X) = F(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) = F(\beta_0 + \boldsymbol{\beta^T} X)$$

onde $F(.)$ é uma função de distribuição acumulada logística dada por $F(u) = \frac{exp(u)}{1+ exp(u)}$ de modo que $0 < F(u) < 1,\ u \in \boldsymbol{\mathcal{R}}$.

A estimação do modelo é realizada por meio da maximização da função de verossimilhança com o suporte de algoritmos de otimização não linear.

Observe que a estimação do modelo é realizada com base em um conjunto $\boldsymbol{X} = (x_1, \dots, x_k)$ de variáveis que fazem sentido ao problema e que são pré-selecionadas pelo analista. Todavia, com o aumento da capacidade de processamento dos computares e também a reconhecida importância da coleta, armazenamento e tratamento de dados, o número de variáveis disponíveis para pesquisas empíricas é cada vez maior. Atualmente, não é raro nos defrontarmos com problemas onde o número de variáveis disponíveis (w) é muito maior do que o número de variáveis necessárias ao modelo (k). Assim, precisamos selecionar apenas as variáveis mais relevantes para a construção do modelo.

Uma alternativa é empregar métodos para seleção de variáveis tais como os métodos de regularização (*shrinkage methods*) ou os métodos de busca exaustiva, como o (*stepwise selection*). Outra possibilidade é a utilização de técnicas para redução/agrupamento de variáveis, a exemplo da análise de componentes principais (*principal component analysis*).[5]

[5] Mais detalhes podem ser encontrados em James et al. (2013) e Hastie, Tibshirani e Friedman (2017).

3.2. Seleção de variáveis

Neste relatório, optou-se por empregar duas abordagens para seleção de variáveis construídas sob diferentes conceitos. O primeiro método empregado, *stepwise selection* (*forward* ou *backward*), consiste num algoritmo de busca intensiva e sequencial onde variáveis são adicionadas/removidas do modelo com base em algum critério. Por outro lado, a segunda abordagem utilizada consiste no emprego de métodos de regularização tais como *Least Absolute Shrinkage Selection* (LASSO). Estes permitem a estimação do modelo e a seleção das variáveis simultaneamente ao inserirmos termos de penalização na função de verossimilhança.

3.2.1. Forward/backward stepwise selection

Os algoritmos de *stepwise selection* são métodos sequenciais de seleção de variáveis onde a cada iteração uma variável é adicionada/removida do modelo de acordo com alguma regra. No *forward stepwise selection* o algoritmo inicia-se com um modelo simples onde apenas o intercepto está presente e, a cada iteração, uma nova variável é adicionada ao modelo com base em algum critério como por exemplo o *Bayesian Information Criterion* (BIC) ou *Akaike Information Criterion* (AIC), (HASTIE; TIBSHIRANI; FRIEDMAN, 2017).

Por outro lado, o *backward stepwise selection* funciona de forma oposta pois parte de um modelo mais complexo, onde todas as variáveis e o intercepto estão presentes. A cada iteração, uma variável é removida com base em alguma métrica de ajuste.[6]

3.2.2. Shrinkage methods

Os *shrinkage* ou *regularization methods* são métodos de estimação que impõem penalidades aos parâmetros do modelo de modo a tornar os coeficientes das variáveis explicativas não relevantes próximos ou iguais à zero. Existem diversos métodos de regularização sendo os principais o Lasso, a *rigde regression* e o *elastic net*.

[6] Ver nota 5.

A função de verossimilhança de uma regressão logística com *Lasso* em sua forma Lagrangiana é dada por:

$$Máx \left\{ \sum_{i=1}^{N} [y_i(\beta_0 + \beta' x_i) - \log\left(1 + e^{\beta_0 + \beta' x_i}\right)] - \lambda \sum_{j=1}^{k} |\beta_j| \right\}$$

Nesta equação, $\lambda \geq 0$ é o parâmetro de penalização (*tunning parameter*) que controla o tamanho da regularização (*shrinkage*), isto é, a força na qual os parâmetros $\beta_j's$ são forçados a zero. Quanto maior o λ, mais elementos de β (aqueles associados as variáveis com menor contribuição a função objetivo) se tornarão iguais a zero. Dessa forma, o modelo final será constituído por subconjunto das variáveis explicativas mais relevantes em termos da função de verossimilhança. Se $\lambda = 0$, a equação acima é um simples modelo *logit.*

Outros métodos de regularização podem ser construídos ao empregarmos diferentes termos de penalização. Por exemplo, a regularização *ridge* é obtida ao utilizarmos o termo $\lambda \sum_{j=1}^{k} \beta_j^2$. Neste caso, o modelo resultante será constituído por todas as variáveis disponíveis. Contudo, as variáveis menos importantes terão seus coeficientes próximos de zero. Por outro lado, o *elastic-net* pode ser entendido como uma combinação linear convexa entre a regularização *ridge* e o *lasso*. Assim, alguns coeficientes das variáveis menos relevantes serão próximos de zero enquanto outros serão iguais a zero. Matematicamente, a regularização *elastic-net* é dada por $\lambda \sum_{j=1}^{k} \left(\alpha \beta_j^2 + (1-\alpha)|\beta_j|\right)$.[7]

Vale destacar que nos métodos de regularização expostos, a definição de λ é fundamental e realizada de antemão à estimação dos coeficientes $\beta_j's$. Em termos práticos, testa-se diferentes valores e seleciona-se aquele que apresentar o melhor ajuste.

[7] Ver nota 5.

3.3. Métricas

3.3.1. Métricas para avaliação das previsões

Nesta seção apresentamos as métricas que serão empregadas para a avaliação do desempenho preditivo dos modelos. Para isto, introduziremos a matriz de confusão, Tabela 1, onde é possível observar os erros e acertos das classificações realizadas pelo modelo. Essa matriz pode ser analisada da seguinte forma: As linhas estão relacionadas aos valores verdadeiros e as colunas estão relacionadas aos valores previstos. Dessa forma, podemos estabelecer o número de elementos classificados corretamente com o *status* 0 (verdadeiro negativo[8]) e o número de elementos classificados corretamente com o *status* 1. Dois erros são passíveis de serem cometidos. O erro falso negativo ocorre quando o verdadeiro *status* é 1 e o modelo acaba atribuindo (predizendo) o *status* como sendo igual à 0. O erro *falso positivo* ocorre quando o modelo classifica o indivíduo com *status* igual a 1, porém o verdadeiro status é zero.

Tabela 1 – Exemplo de matriz de confusão

		***Status* previsto**	
		0	1
***Status* Verdadeiro (Real)**	0	*Verdadeiro Negativo (VN)*	*Falso Positivo (FP)*
	1	*Falso Negativo (FN)*	*Verdadeiro Positivo (VP)*

Elaboração: Estudos Educacionais (IJSN).

Com base nestas definições as seguintes métricas podem ser definidas:

i. *Acurácia*. Definida como o percentual de previsões corretamente realizadas. Matematicamente é estimada da seguinte forma: (VN + VP)/(VN + VP + FN + VP);

ii. *Sensibilidade*. Calculada como VP/(FN+VP). Intuitivamente, a sensibilidade mensura o número de previsões de *status* igual 1 realizadas corretamente diante do total de indivíduos que de fato possuem *status* igual a 1.

iii. *Especificidade*. Mede o número de previsões de *status* igual a 0 realizadas corretamente diante do total de indivíduos que de fato possuem *status* igual a 0. Formalmente é obtida dada por VN/(VN + FP).

[8] É comum atribuirmos também os temos *negativo* e *positivo* para a variável dicotômica da seguinte forma $y_i = \{0\ (negativo), 1\ (positivo)\}$.

iv. *Precisão*. Obtida pela razão de previsões corretas iguais a 1 e o total de previsões iguais a 1. Formalmente, é estimada por VP/(VP+FP).

v. F1. Pode ser entendida como uma espécie de média entre precisão e sensibilidade. Matematicamente é dada por 2VP/(2VP + FP + FN).

É importante observar que tais métricas permitem uma avaliação mais robusta do modelo, já que em muitas situações classificações erradas geram impactos diferentes sobre o problema em questão. Tendo como exemplo o abandono escolar, as consequências de se cometer um falso negativo, isto é, classificar um estudante como não provável ao abandono quando de fato ele é um estudante que abandonará os estudos, terá consequências negativas enormes uma vez que nenhuma intervenção seria realizada para reverter esse cenário. Por outro lado, na situação de um falso positivo, alguma intervenção seria realizada para evitar o abandono de um aluno que de fato não abandonaria o curso. Isso pode representar algum custo, porém o objetivo final de manter o aluno na escola seria alcançado.

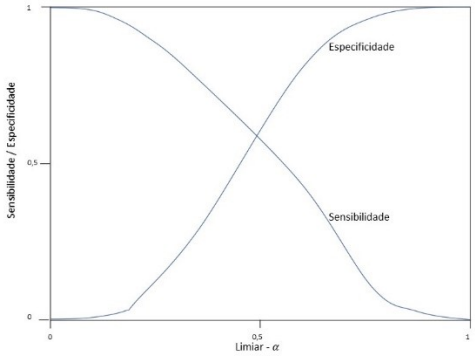
3.3.2. Métricas para avaliação do ajuste do modelo

3.3.2.1. Curva característica de operação do receptor (ROC curve – receiver operation characteristic curve)

Como já apresentado, diversas métricas tais como especificidade e sensibilidade podem ser calculadas a partir de um modelo de classificação binária. Contudo, os resultados de um modelo *logit* são fornecidos em termos de probabilidade e não em termos de variáveis dicotômicas, $y = \{0, 1\}$. Por esse motivo é necessário especificarmos alguma regra para essa transformação. Usualmente, esta transformação pode ser definida como $\hat{y}_i = 1$, se $Prob(y_i = 1|\boldsymbol{X}) \geq \gamma$ **e** $\hat{y}_i = 0$, se $Prob(y_i = 1|\boldsymbol{X}) < \gamma$. Neste caso, o limiar de discriminação α funciona como ponto de corte da classificação. Não é difícil perceber então que a definição do limiar impactará o desempenho preditivo do modelo. Assim, pode-se selecionar, por exemplo, o limiar γ que maximiza a especificidade e a sensibilidade simultaneamente. Para mais detalhes, veja (HOSMER JR.; LEMESHOW; STURDIVANT, 2013).

A Figura 1 abaixo ilustra o comportamento da especificidade e sensibilidade de determinado modelo para diferentes limiares, $\gamma = [0, 1]$. Como pode ser observado, o aumento da especificidade está associado à queda da sensibilidade. Isso ocorre pois o incremento do limiar aumenta a quantidade de previsões corretas com *status* igual a 0, o que altera o numerador da fórmula da especificidade tornando-o maior. Consequentemente isso diminui o numerador da equação da sensibilidade e, naturalmente, o reduz. Por esse motivo, sensibilidade e especificidade estão relacionados.

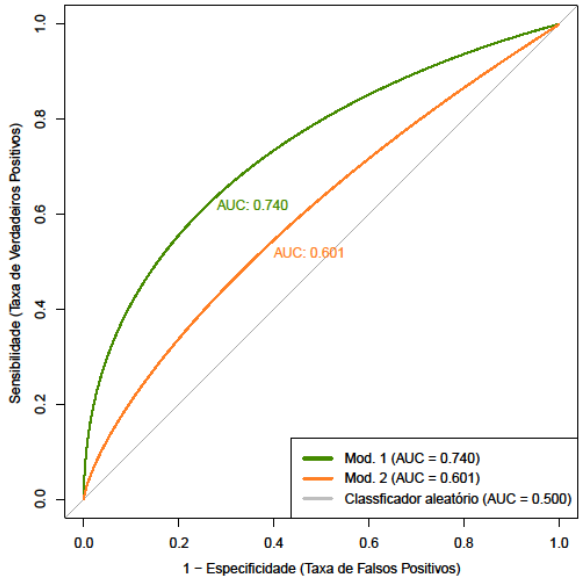
Figura 1 – Curva de sensibilidade versus especificidade para $\gamma \in [0,1]$



Elaboração: Estudos Educacionais (IJSN).
Nota: Em linhas gerais, o α que maximiza a sensibilidade e a especificidade é obtido de forma aproximada pelo ponto de interseção entre as curvas de sensibilidade e especificidade.

Uma medida mais geral acerca do desempenho preditivo do modelo pode ser obtida por meio da curva *receiver operating characteristic curve* (ROC), Figura 2. Esta é uma curva construída ao calcularmos a sensibilidade (taxa de verdadeiros positivos) e um menos a especificidade (taxa de falsos positivos). Com base na ROC é possível obter uma medida global relativa à performance de um modelo de classificação binário. Para tal, basta calcularmos a área sob a curva que assumirá valores entre 0 e 1. Um modelo que erra todas as classificações terá uma área igual a zero. Por outro lado, um modelo que acerta todas as classificações apresentará uma área igual a um. Uma área igual a 0,5 corresponde a performance de um classificador aleatório. Dessa forma, modelos com áreas inferiores a 0,5 devem ser descartados.

Figura 2 – Ilustração de uma curva de operação do receptor (Curva ROC)



Elaboração: Estudos Educacionais (IJSN).
Nota: AUC significa area under the curve. Quanto mais próximo de um, melhor o modelo. Um AUC de 0,5 é obtido por um classificador aleatório. Modelos com AUC inferiores a 0,5 não são recomendados.

3.4. Banco de dados

A validação cruzada, baseada em métodos de reamostragem, tem como objetivo fornecer informações adicionais sobre métricas e/ou parâmetros do modelo considerando a incerteza amostral do problema. Em linhas gerais consiste em sortear diversas subamostras a partir de uma amostra geral (dados iniciais) e reestimar para cada iteração o modelo e as correspondentes métricas.

Existem diferentes algoritmos para validação cruzada sendo um dos mais empregados o *K-fold cross-validation*. Este, divide a amostra em *k* grupos (*folds*) de tamanho iguais. Em seguida, seleciona-se *k-1* grupos para estimação e avalia-se o seu desempenho no grupo remanescente. Este procedimento é repetido *k* vezes até que cada um dos *k* grupos iniciais tenham servidos como "conjunto fora da amostra"[9].[10]

[9] Uma amostra pode ser dividia em dois grupos denominados "fora da amostra" e "dentro da amostra". O conjunto "dentro da amostra" é aquela no qual o modelo de interesse é estimado. O conjunto fora da amostra é aquele no qual o modelo estimado previamente é testado.
[10] Para mais detalhes ver Hastie, Tibshirani e Friedman (2017).

Exemplificando, suponha que queiramos estimar a precisão de um modelo qualquer. Dessa forma, esta seria obtida ao tomarmos a média das precisões estimadas em cada uma das *k* repetições da validação cruzada, ou seja,

$$Precisão_{k-fold} = \frac{1}{k}\sum_{i=1}^{k} Precisão_k$$

Com esse procedimento, temos uma melhor ideia sobre como o modelo se comportaria num conjunto independente de dados. De forma similar, podemos empregar essa técnica também para selecionar configurações e/ou hiperparâmetros tais como o limiar para a transformação das probabilidades resultados de um modelo de classificação binária ou o próprio termo de regularização λ, por exemplo.

4. Estudo de casos

4.1. Banco de dados

O banco de dados utilizado neste relatório origina-se do Sistema Estadual de Gestão Escolar do Espírito Santo (SEGES), é composto por informações individuais dos estudantes da Rede Estadual de Ensino, e a base Situação com classificação de rendimento dos estudantes a partir de cruzamentos realizados pelo INEP.

Com base nesse banco de dados, algumas variáveis foram criadas. O conjunto total de variáveis empregadas nesse relatório é formado por: idade; sexo; turno; nota média da escola em matemática e português; nota média da turma em matemática e português; notas do aluno em português, matemática, história, geografia, física, química e biologia; proporção de faltas em português, matemática, história, geografia, física, química e biologia; cor/raça; *dummies* que indicam se o aluno trocou de turno, turma, escola ou município; quantas mudanças de turno, turma, escola e município o aluno realizou e; o número de escolas, municípios, turno e turma o aluno frequentou no ano.

4.2. Modelo de referência

Em Giuberti et al. (2018) é possível observar o modelo para previsão do abandono escolar no Espírito Santo desenvolvido pelo IJSN ao longo dos últimos anos. Tal ferramenta utiliza o modelo *Logit* com variáveis referentes às características pessoais e de desempenho dos alunos disponíveis no primeiro trimestre do ano letivo para a previsão do abandono ao longo do ano. Como o objetivo é, ao início de cada ano, gerar uma lista com os alunos mais prováveis ao abandono, o modelo é calibrado com base nas informações do primeiro trimestre do ano anterior (t) e, ao início do ano do ano seguinte (t+1), as previsões são elaboradas com base nas novas informações coletadas no primeiro trimestre.

Duas especificações do modelo *Logit* foram analisadas em Giuberti et al. (2018). A primeira, denominada modelo parcimonioso, foi construída considerando apenas variáveis administrativas sendo estas a idade, sexo, turno, notas médias da escola, da turma e nota individual nas disciplinas de português e matemática, além da proporção de faltas. Para a segunda especificação foram adicionadas variáveis relacionadas à infraestrutura e características socioeconômicas das escolas. Além destas, o modelo considera ainda o nível socioeconômico da escola, regional (SRE), localidade (rural ou urbana), índice de complexidade da gestão (ICG), indicador de esforço docente (IED) e indicador de adequação da formação do docente (AFD).

Com base nas análises realizadas, os autores concluem que o modelo parcimonioso é o mais adequado para a análise. Além disso, por considerar apenas dados administrativos sob responsabilidade da Secretaria de Educação do Espírito Santo, a gestão dos dados é mais simples, uma vez que não depende de dados oriundos de outros órgãos. Por esses motivos, neste relatório, optou-se por utilizar também o modelo parcimonioso descrito em Giuberti et al. (2018) como modelo de referência.

4.3. Resultados

Dois estudos de caso foram realizados nesse relatório. No primeiro consideramos apenas o ano de 2019, tanto para a estimação do modelo quanto para a validação. No segundo estudo de caso, consideramos o modelo estimado em 2019 e observamos o seu desempenho em 2020. Este exercício é como de fato a ferramenta do preditor de abandono funcionará, uma vez que ao iniciarmos o ano, precisamos utilizar o modelo calibrado no ano anterior.

Em todos os casos 8 diferentes especificações foram testadas. Primeiramente, denominado modelo 1, temos a formulação parcimoniosa como estabelecido em Giuberti et al. (2018). Os modelos 2 e 3 foram construídos com base nos algoritmos *forward* e *backward stepwise selection*, respectivamente. Em seguida, adicionamos efeitos fixos de escola aos modelos 2 e 3. O objetivo é verificar se a heterogeneidade não observada das escolas (condições específicas de cada escola) é informação relevante para o abandono. Por fim, os modelos 6, 7 e 8 foram construídos via métodos de regularização (*lasso*, *ridge* e *elastic-net*) onde *dummies* para as escolas estavam disponíveis para seleção.

É importante observar que os modelos aqui estimados são relativos ao 1º ano do ensino médio na rede Estadual do Espírito Santo. Uma vez que efeitos fixos de escola foram testados, apenas escolas existentes nos dois anos, isto é, 2019 e 2020 foram analisadas. No total, 273 escolas foram consideradas, o que representa uma amostra com 34.954 alunos. Desse quantitativo, 937 estudantes abandonaram ao longo do ano 2019. Para o ano de 2020, 1207.

A seguir, apresentamos os resultados para os dois estudos de caso realizados.

4.3.1. Caso 2 – 1º ano do ensino médio em 2019

Neste estudo de caso consideramos apenas as informações relativas ao primeiro semestre de 2019 para verificar o abandono ao longo de 2019. Para isto, separamos o banco de dados em dois conjuntos: o de treinamento com aproximadamente 85% da amostra onde o modelo é estimado e o de validação (teste) com 15% da amostra. É

importante ressaltar que o percentual de abandono presente na totalidade do bando de dados foi preservado nesses dois conjuntos. Além disso, o conjunto de validação não é utilizado para a estimação dos parâmetros, sendo o seu único objetivo verificar a performance do modelo em um conjunto novo de dados.

Em todos os modelos, as métricas de desempenho e o limiar γ foram obtidas por meio de *k-fold cross-validation* com *k=10.* Além destes, tal procedimento também é crucial no caso dos modelos com regularização para a definição das constantes de regularização α e λ. Na Tabela 2 é possível observar as especificações dos modelos estimados e os correspondentes métodos de seleção/estimação empregado.

Com relação ao modelo de referência, podemos notar que os modelos 2 e 3 são muito similares, possuindo pequenas diferenças no que tange às variáveis empregadas. Podemos perceber, por exemplo, que os algoritmos de *stepwise* tendem a considerar apenas os turnos de tarde e/ou integral quando comparados ao modelo de referência que considera todos os turnos. Além disso, as *dummies* para cor/raça e quantidade de mudança de turma, escola e/ou município também são incorporadas aos modelos.

Tendo em vista os métodos de regularização, os correspondentes parâmetros estimados foram $\lambda = 0{,}001$ para o Lasso, $\lambda = 0{,}0081$ para o *ridge* e $\alpha = 0{,}1$ e $\lambda = 0{,}00005$ para *elastic-net*. Tais parâmetros, como já mencionado, foram obtidos por meio de validação cruzada ao minimizarmos a acurácia. É importante observar que em todos os casos, os λs selecionados são muito próximos de zero, o que significa que o termo de penalização possui pouco valor na função de verossimilhança. Vale lembrar que quando $\lambda = 0$, os modelos se equiparam a um simples modelo *logit.*

Tabela 2 – Modelos e especificações

Especificação	Método de Seleção	Variáveis selecionadas
Modelo 1	Modelo de referência	▪ Idade, sexo, turno da manhã, tarde, noite e integral; ▪ Nota média da escola[11] em português e matemática; ▪ Nota média da turma[12] em português e matemática; ▪ Nota média do aluno[13] em português, matemática, história, geografia, física, química e biologia; ▪ Proporção de faltas[14] em português, matemática, história, geografia, física, química e biologia.
Modelo 2	*Forward stepwise*	▪ Idade; turnos da tarde e integral; ▪ Nota média da escola em português e matemática; ▪ Nota média da turma em português e matemática; ▪ Nota média do aluno em português, matemática, história, geografia, química, física e biologia; ▪ Proporção de faltas em matemática, português, história, geografia e química. ▪ *Dummy* para raça/cor branca e não declarada; ▪ Quantidade de mudanças de município que o estudante realizou.
Modelo 3	*Backward stepwise*	▪ Idade; turno integral; ▪ Nota média da escola em português e matemática; ▪ Nota média da turma em português e matemática; ▪ Nota média do aluno em português, matemática, história, geografia, química, física e biologia; ▪ Proporção de faltas em matemática, português, química e história; ▪ *Dummy* para raça/cor indígena, parda e preta; ▪ Quantidade de mudança de turma, escola e município.
Modelo 4	-	▪ Modelo 2 com adição de efetivos fixos para as escolas.
Modelo 5	-	▪ Modelo 3 com adição de efeitos fixos para as escolas.
Modelo 6	LASSO	▪ Idade, turnos da tarde e integral; ▪ Notas do primeiro trimestre em química, língua portuguesa, matemática, biologia, física, geografia e química. ▪ Proporção de faltas em matemática, português, química, história, geografia e biologia; ▪ Quantidade de mudança de município; ▪ Dummy para raça/cor branca e indígena; ▪ 98 *dummies* para as escolas.
Modelo 7	Rigde	▪ Todas as variáveis disponíveis, incluindo *dummies* para todas as escolas.
Modelo 8	Elastic-Net	▪ Todas as variáveis disponíveis com exceção da variável quantidade de mudança de município e variáveis *dummy* relativas a 4 escolas.

Elaboração: Estudos Educacionais (IJSN).

A Tabela 3 resume as variáveis explicativas, os correspondentes sinais e quais são significativas a um nível de significância de 5%. Nesta tabela, o sinal "+" indica que um aumento na correspondente variável aumenta a probabilidade de abandono. Por outro lado, o símbolo "-" indica que um aumento na variável reduz a probabilidade de abandono. Sinais destacados em vermelho indicam que a variável não é

[11] Nota média da escola calculada com base nas informações do 1º trimestre.
[12] Nota média da turma calculada com base nas informações do 1º trimestre.
[13] Nota média do aluno calculada com base nas informações do 1º trimestre.
[14] No 1º trimestre.

estatisticamente significativa. Ademais, espaços em branco indicam que determinado modelo não utiliza a correspondente variável.

De modo geral, podemos perceber que a variável idade aparece em todos os modelos com sinal positivo. Quanto mais velho for o estudante, considerando determinada série, maior a probabilidade de abandono. As notas médias dos estudantes também foram empregadas em todos os modelos. Como pode ser observado, quanto maior a nota dos alunos, menores serão as probabilidades de abandono. Exceção diz respeito a variável nota do aluno em história. Esta não é estatisticamente significante ou não foi selecionada em 5 modelos. Além disso, quando selecionada, aparece com sinal contrário ao esperado.

Com relação a proporção de faltas dos estudantes, estas foram variáveis bastante utilizadas nos modelos. Como pode ser visualizado na Tabela 3, um aumento no número de faltas aumenta a chance de abandono.

As variáveis relativas à raça também se mostraram relevantes ao problema. No geral, as raças indígenas, pardas e pretas aumentam a probabilidade de abandono. Por outro lado, a raça branca diminui esta probabilidade.

Com base nessas informações podemos estimar um perfil para o abandono. Assim, pessoas com idade acima ao esperado para a sua série, baixas notas, número elevado de faltas e não brancas são mais propensas ao abandono.

A variável turno integral aparenta diminuir a probabilidade de abandono enquanto que o turno da noite aumenta. As notas médias da turma e da escola em português e em matemática aparecem com sinais opostos. Fato que carece de maiores investigações para entender as razões pelas quais isto ocorre. Por fim, as variáveis relativas a mudanças de escola, turma, turno e cidade foram selecionadas poucas vezes, o que é um indício de baixa relevância para o problema.

Vale a pena destacar que os sinais dos modelos 7 e 8 devem ser analisados com certo cuidado. O modelo 7, por exemplo, não elimina nenhuma variável, mesmo que não relevante ao problema. O método de penalização utilizado aproxima o valor do

coeficiente de variáveis não relevantes à zero. O método de penalização que elimina variáveis não relevantes foi empregado no Modelo 6 apenas. O Modelo 8, como já mencionado, pode ser entendido como um balanceamento entre os métodos do Modelo 6 e Modelo 8. Dessa forma, é possível que variáveis não relevantes sejam integradas ao modelo, porém com coeficientes próximos de zero.

Tabela 3 – Associação entre as variáveis explicativas e a probabilidade de abandono

	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5	Modelo 6	Modelo 7	Modelo 8
Idade	+	+	+	+	+	+	+	+
Sexo Masculino	+						+	+
Nota média da escola (português)	-	-	-	-	-		-	-
Nota média da escola (matemática)	+	+	+	-	+		-	-
Turno tarde	-						+	+
Turno da noite	+	+		+		+	+	+
Integral	-	-	-	-	-	-	-	-
Nota média da turma (português)	+	-	+	+	+		+	+
Nota média da turma (matemática)	-	-	-	-	-		-	-
Nota do aluno (português)	-	-	-	-	-	-	-	-
Nota do aluno (matemática)	-	-	-	-	-	-	-	-
Nota do aluno (história)	+	+	+	+	+		+	+
Nota do aluno (geografia)	-	-	-	-	-	-	-	-
Nota do aluno (biologia)	-	-	-	-	-	-	-	-
Nota do aluno (química)	-	-	-	-	-	-	-	-
Nota do aluno (física)	-	-	-	-	-	-	-	-
Proporção de faltas (português)	+	+	+	+	+	+	+	+
Proporção de faltas (matemática)	+	+	+	+	+	+	+	+
Proporção de faltas (história)	+	+	+	+	+	+	+	+
Proporção de faltas (geografia)	+	+		+		+	+	+
Proporção de faltas (biologia)	+					+	+	+
Proporção de faltas (química)	+	+	+	+	+	+	+	+
Proporção de faltas (física)	-						-	-
Raça branca	+	-		-		-	-	-
Raça Indígena	+		+		+	+	+	+
Raça não declarada	+	-		-			+	+
Raça parda	+		+		+		+	+
Raça preta	+		+		+		+	+
Dummy de mudança de escola							-	-
Quantidade de mudança de escola			-		-		-	-
Quantidade de escola							-	-
Dummy de mudança município							+	+
Quantidade de mudança de município		-		-		-	-	-
Quantidade de município			-		-			
Dummy de mudança turma							+	+
Quantidade de mudança de turma							+	+
Quantidade de turma			+		+		+	+
Dummy de mudança turno							-	+
Quantidade de mudança de turno							+	-
Quantidade de turno							+	-
Efeitos Fixos de Escola	Não	Não	Não	Sim	Sim	Parcial	Sim	Sim

Elaboração: Estudos Educacionais (IJSN).
Fonte: dados administrativos (SEDU, 2022).

Análise do desempenho preditivo. Tendo em vista o desempenho preditivo dos modelos utilizados no conjunto de treinamento, as Tabelas 5 e 6, apresentam as métricas obtidas para os conjuntos de treinamento e teste. Tais métricas foram construídas com base nos limiares estimados por meio de validação cruzada (*k-fold cross validation*), as quais podem ser observadas na Tabela 4.

Como pode ser observado, os modelos 1, 2 e 3 não apresentam grandes diferenças em termos das métricas empregadas. Por outro lado, nos modelos 4 e 5, onde há a inclusão de variáveis *dummy* para captar a heterogeneidade não observada das escolas, podemos verificar que tal recurso melhorou o desempenho preditivo em todas as métricas analisadas.

Uma das possíveis explicações para tal fenômeno seria o fato que algumas características particulares existentes na escola ou na redondeza, tais como infraestrutura, violência entre outras podem estar afetando a decisão dos estudantes de abandonarem os estudos.

Com relação aos modelos 6, 7 e 8, estes apresentam resultados bem similares aos modelos 4 e 5. Dessa forma, podemos perceber que, em linhas gerais, as informações disponíveis no banco de dados do SEGES são relevantes para a previsão de abandono. Além disso, embora com pequenas diferenças entre as especificações, o comportamento em termos preditivos, dentro da amostra, é similar. Diferenças interessantes ocorrem quando adicionamos efeitos fixos de escola, fazendo com que haja diferenças entre os modelos. Assim, é possível observar dois grupos de performance. O primeiro contém os modelos 1, 2 e 3, enquanto o segundo é composto pelos modelos 4, 5, 6, 7 e 8.

Tabela 4 – Limiares para as classificações da previsão

Especificação	Limiar
Modelo 1	0,0251
Modelo 2	0,0256
Modelo 3	0,0256
Modelo 4	0,0354
Modelo 5	0,0347
Modelo 6	0,0300
Modelo 7	0,0340
Modelo 8	0,0357

Elaboração: Estudos Educacionais (IJSN).
Fonte: dados administrativos (SEDU, 2022).

Tabela 5 – Performance dos modelos – conjunto de treinamento

Especificação	Acurácia	Sensibilidade	Especificidade	Precisão	F1	AUC
Modelo 1	0,828	0,829	0,828	0,119	0,208	0,907
Modelo 2	0,830	0,829	0,830	0,122	0,212	0,907
Modelo 3	0,835	0,835	0,835	0,124	0,216	0,906
Modelo 4	0,887	0,887	0,887	0,181	0,300	0,956
Modelo 5	0,886	0,885	0,886	0,178	0,296	0,957
Modelo 6	0,866	0,865	0,866	0,154	0,261	0,943
Modelo 7	0,882	0,882	0,882	0,172	0,288	0,952
Modelo 8	0,890	0,890	0,890	0,186	0,307	0,959

Elaboração: Estudos Educacionais (IJSN).
Fonte: dados administrativos (SEDU, 2022).

Tabela 6 – Performance dos modelos – conjunto de teste

Especificação	Acurácia	Sensibilidade	Especificidade	Precisão	F1	AUC
Modelo 1	0,826	0,814	0,826	0,114	0,200	0,897
Modelo 2	0,823	0,807	0,823	0,111	0,196	0,903
Modelo 3	0,836	0,785	0,837	0,117	0,204	0,905
Modelo 4	0,888	0,807	0,891	0,169	0,279	0,945
Modelo 5	0,878	0,878	0,878	0,165	0,277	0,944
Modelo 6	0,866	0,800	0,867	0,142	0,241	0,922
Modelo 7	0,876	0,807	0,878	0,154	0,259	0,955
Modelo 8	0,886	0,885	0,886	0,176	0,293	0,945

Elaboração: Estudos Educacionais (IJSN).
Fonte: dados administrativos (SEDU, 2022).

Adicionalmente, com o intuito de investigarmos a contribuição individual de cada variável para o modelo, estimamos a área sob a curva ROC para cada variável. Esta área representa a nossa medida de contribuição. A Tabela 7 abaixo apresenta o *ranking* das variáveis para cada um dos oito modelos. Vale a pena ressaltar que a importância das variáveis *dummy* para os correspondentes modelos foram estimadas, porém omitidas para uma melhor apresentação dos resultados. As 5 variáveis explicativas (excluindo as *dummies* para escolas, quando presentes) mais importantes para cada modelo foram destacadas em amarelo.

As variáveis relativas à proporção de faltas estão entre as mais relevantes em diversos modelos. Além destas, idade e turno integral também são listas como principais em mais de um modelo. Vale a pena ressaltar que nos modelos onde há a inserção de variáveis *dummy*, as variáveis explicativas do SEGES não são assumem as primeiras posições. Isso indica que, neste ano, os abandonos podem estar associados também a características dos colégios e não somente a dos alunos.

Tabela 7 – Ranking das contribuições das variáveis para a previsão

	Mod. 1	Mod. 2	Mod. 3	Mod. 4	Mod. 5	Mod. 6	Mod. 7	Mod. 8
Idade	22	1	1	28	115	115	312	306
Sexo Masculino	23						289	297
Nota média da escola (port.)	10	21	23	203	242		305	289
Nota média da escola (mat.)	8	2	3	51	117		295	293
Turno tarde	14						308	287
Turno da noite	9	19		30		69	155	212
Integral	3	3	4	290	217	83	254	204
Nota média da turma (port.)	12	13	15	35	184		302	291
Nota média da turma (mat.)	13	11	13	260	258		310	302
Nota do aluno (port.)	20	9	12	37	89	112	303	299
Nota do aluno (mat.)	16	10	20	39	114	111	300	298
Nota do aluno (hist.)	21	18	22	43	141		311	303
Nota do aluno (geo.)	19	16	16	41	262	113	304	301
Nota do aluno (bio.)	17	7	17	29	122	108	292	295
Nota do aluno (quí.)	15	6	7	32	157	110	294	296
Nota do aluno (fís.)	18	12	11	38	133	114	301	300
Proporção de faltas (port.)	6	15	14	36	20	48	95	222
Proporção de faltas (mat.)	1	4	2	31	53	24	26	160
Proporção de faltas (hist.)	4	8	8	34	77	54	93	203
Proporção de faltas (geo.)	5	20		33		55	109	232
Proporção de faltas (bio.)	11					47	108	207
Proporção de faltas (quí.)	2	5	5	40	46	61	77	227
Proporção de faltas (fís.)	7						290	249
Raça branca		14		44		102	296	307
Raça Indígena			19		223	98	60	256
Raça não declarada		22		277			275	279
Raça parda			10		144		291	286
Raça preta			18		215		306	280
Dummy de mud. de escola							287	304
Quantidade de mud. de escola			9		132		307	278
Quantidade de escola							271	163
Dummy de mud. município							274	277
Quantidade de mud. de município		17		42		94	293	312
Quantidade de município			21		284		264	236
Dummy de mud. turma							272	268
Quantidade de mud. de turma							286	259
Quantidade de turma			6		117		273	208
Dummy de mudança turno							281	270
Quantidade de mud. de turno							298	281
Quantidade de turno							282	269
Número de parâmetros	24	23	23	296	296	116	308	313

Elaboração: Estudos Educacionais (IJSN).
Fonte: dados administrativos (SEDU, 2022).

4.3.2. Caso 2 – 1º ano do ensino médio em 2020

O segundo estudo de caso consiste em utilizar o modelo calibrado com os dados do primeiro trimestre em 2019 para predizer a situação de abandono ao longo de 2020. Isso ocorre pois, em termos operacionais, o preditor deve ser utilizado assim que as primeiras informações relativas aos alunos estiveram disponíveis.

Novamente, avaliamos os 8 modelos estimados até então. Vale destacar que a estimação e os parâmetros obtidos no Estudo de caso 1 são os considerados para a previsão do ano de 2020.

A Tabela 8 apresenta os resultados das métricas consideradas para a previsão. Como pode se notas, o padrão observado no estudo de caso 1 não ocorre para o ano de 2020. Não há uma distinção clara entre dois grupos, um formado pelos modelos 1, 2 e 3 e outro formado pelos modelos restantes. Além disso, a depender da métrica adotada, os resultados são ambíguos. Por exemplo, o modelo 1 tem a pior acurácia, porém a maior sensibilidade. Uma possibilidade para seleção de modelos nesses casos seria a utilização da métrica F1, que pode ser entendida como uma espécie de média entre sensibilidade e precisão, ou a área sob a curva (AUC) que é uma medida independente do limiar γ.

Com relação a inclusão das variáveis *dummy*, pode-se perceber que no geral, estas aumentam a acurácia da previsão. Contudo, tais variáveis não possuem o mesmo impacto sobre a sensibilidade, talvez a métrica mais importante para o problema em questão.

Tendo em vista os métodos de regularização estes foram capazes de aumentar a acurácia, precisão, especificidade e F1. Por outro lado, a sensibilidade dos modelos 7 e 8 foram menores do que nos modelos mais simples.

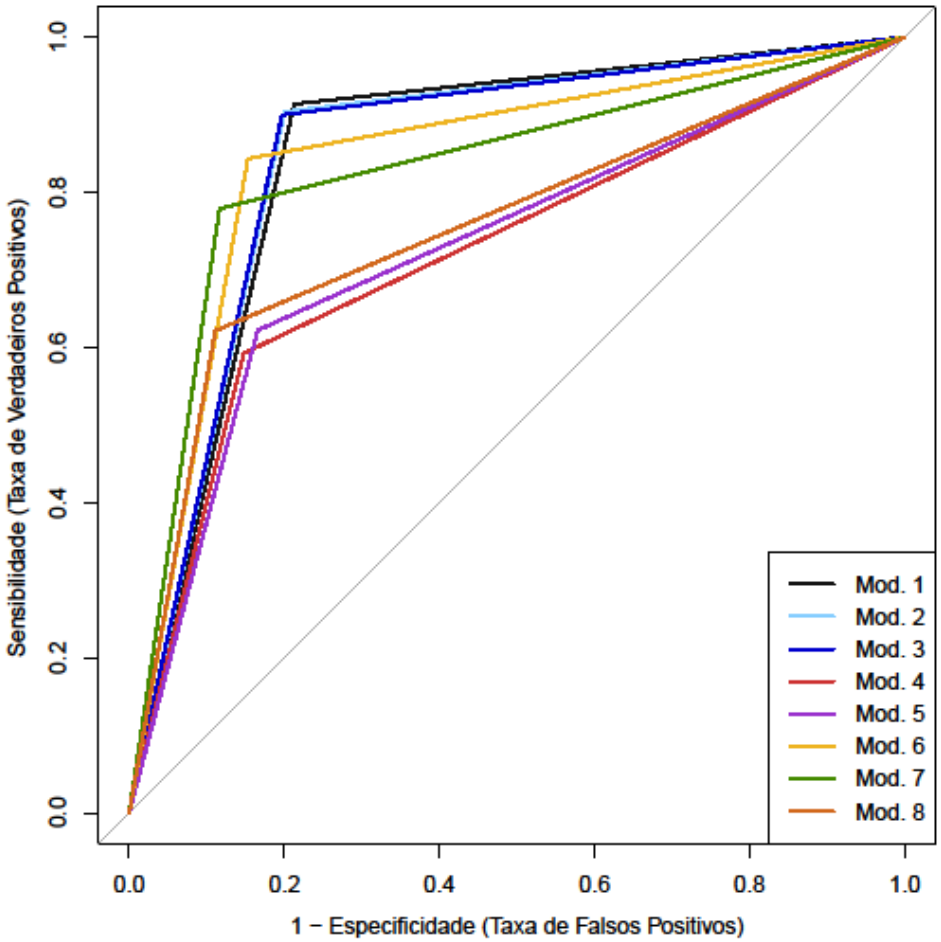
A Figura 3 apresenta a Curva ROC para os 8 modelos considerados. Por meio desta métrica também é possível observar que os modelos 4, 5 e 8 apresentam os menores AUC enquanto que os modelos restantes apresentam AUC, a princípio, próximos.

Tabela 8 – Performance dos modelos para o ano de 2020

Especificação	Acurácia	Sensibilidade	Especificidade	Precisão	F1	AUC
Modelo 1	0,791	0,913	0,787	0,132	0,231	0,850
Modelo 2	0,803	0,903	0,799	0,138	0,239	0,851
Modelo 3	0,806	0,899	0,802	0,139	0,241	0,851
Modelo 4	0,843	0,592	0,852	0,124	0,206	0,722
Modelo 5	0,826	0,622	0,834	0,117	0,197	0,728
Modelo 6	0,846	0,843	0,846	0,163	0,274	0,845
Modelo 7	0,879	0,778	0,883	0,191	0,307	0,831
Modelo 8	0,879	0,621	0,889	0,166	0,262	0,755

Elaboração: Estudos Educacionais (IJSN).
Fonte: dados administrativos (SEDU, 2022).

Figura 3 – Curva ROC para os 8 modelos estimados



Elaboração: Estudos Educacionais (IJSN).
Fonte: dados administrativos (SEDU, 2022).

Para verificarmos se as diferenças observadas entre as áreas são relevantes ou não, realizamos o teste proposto por DeLong, DeLong e Clarke-Pearson (1988) para verificar se as AUCs são estatisticamente significantes. Os p-valores dos testes podem sem

observados na Tabela 9 abaixo. A hipótese nula indica que as diferenças entre as áreas são estatisticamente iguais a zero, enquanto a hipótese alternativa indica que as diferenças são estatisticamente significantes.

Com base nos p-valores e considerando um nível de significância de 5%, podemos concluir que a performance em termos de AUC entre os modelos 1, 2, 3 e 6 são estatisticamente iguais e superiores aos outros modelos. Dessa forma, podemos perceber que o aumento da complexidade dos modelos não acarreta diretamente ganhos preditivos. Além disso, os efeitos positivos da inclusão de efeitos fixos das escolas observados no estudo de Caso 1 não é verificado neste momento.

Tabela 9 – P-valores dos testes de DeLong

	Mod. 2	Mod. 3	Mod. 4	Mod. 5	Mod. 6	Mod. 7	Mod. 8
Mod. 1	0,5612	0,7235	0,0001	0,001	0,237	0,0003	0,0001
Mod. 2		0,8235	0,0001	0,001	0,166	0,0002	0,0001
Mod. 3			0,0001	0,001	0,188	0,0003	0,0001
Mod. 4				0,097	0,001	0,0001	0,0001
Mod. 5					0,001	0,0001	0,0001
Mod. 6						0,0005	0,0001
Mod. 7							0,0001

Elaboração: Estudos Educacionais (IJSN).
Fonte: dados administrativos (SEDU, 2022).

5. Conclusões

O presente relatório teve por objetivo a avaliação de métodos de seleção de variáveis que podem ser utilizados simultaneamente aos modelos *Logit*, já empregados pelo IJSN como ferramenta para previsão do abandono escolar no Espírito Santo.

Ao total, sete diferentes especificações foram estimadas. Estas foram construídas com base nos métodos *stepwise selection* e nos métodos de regularização (*lasso*, *ridge* e *elastic-net*). Além desses métodos, também optamos por avaliar se a modelagem de efeitos não observados das escolas seria capaz de melhorar o desempenho preditivo dos modelos propostos.

Todos os modelos foram comparados com o modelo de referência, atualmente utilizado pelo IJSN, com base em métricas calculadas a partir da matriz de confusão e por meio da área sob a curva ROC (AUC).

Dois estudos de casos foram realizados. O primeiro foi elaborado com os dados de 2019 e os modelos foram avaliados nos conjuntos de treinamento e validação. No segundo estudo de caso, utilizamos o modelo calibrado em 2019 para realizar a previsão para 2020. Os resultados são antagônicos. Enquanto no primeiro estudo de caso os modelos mais complexos, tanto do ponto de vista de métodos de seleção quanto da inclusão de efeitos fixos de escolas se comportaram melhor, no segundo estudo de caso, modelos mais simples e com menos parâmetros foram os mais adequados.

Além disso, os resultados também indicam que os dados administrativos fornecidos pelo SEGES são capazes de predizer o abandono de forma satisfatória, em consonância com outros modelos existentes ao redor do mundo.

Os modelos estimados também permitiram a especificação de um perfil para o abandono. Pessoas com idade acima ao esperado para a sua série, baixas notas, número elevado de faltas e não brancas são mais propensas ao abandono.

Direcionamentos para trabalhos futuros passam por utilizar bancos de dados adicionais e testar novas variáveis. A aplicação de questionários a estudantes e professores pode ser uma fonte relevante de informação. Tal abordagem já foi empregada com sucesso em pesquisas similares. Outra possibilidade é o uso de modelos de inteligência artificial. Uma característica bem marcante do problema do abandono diz respeito ao desbalanceamento de classes uma vez que a situação de não abandono é muito maior do que o número de abandonos. Há relatos na literatura que tal desbalanceamento pode vir a gerar vieses durante a calibração do modelo. Por esse motivo, testar diferentes soluções para tal questão pode vir a melhorar o desempenho preditivo do modelo.

Referências

ADELMAN, Melissa et al. Predicting school dropout with administrative data: new evidence from Guatemala and Honduras. **Education Economics**, v. 26, n. 4, p. 356-372, 2018.

BAYER, Jaroslav et al. Predicting Drop-Out from Social Behaviour of Students. **International Educational Data Mining Society**, 2012.

BURGESS, Simon M. Human capital and education: The state of the art in the economics of education. **IZA Discussion Paper**, n. 9885, 2016.

COSTA, Evandro B. et al. Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. **Computers in Human Behavior**, p. 247-256, 2017.

DA CUNHA, José Antônio; MOURA, Elionai; ANALIDE, Cesar. Data mining in academic databases to detect behaviors of students related to school dropout and disapproval. In: **New Advances in Information Systems and Technologies: Volume 2**. Springer International Publishing, 2016. p. 189-198.

DEKKER, Gerben W.; PECHENIZKIY, Mykola; VLEESHOUWERS, Jan M. Predicting Students Drop Out: A Case Study. **International Conference on Educational Data Mining (EDM)**. Cordoba, Espanha: EDM, 2009. 10p.

DELONG, Elisabeth R.; DELONG, DAVID M.; CLARKE-PEARSON, Daniel L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. **Biometrics**, v. 44, n. 3, p. 837-845, 1988.

GIUBERTI, Ana Carolina et al. Modelo de predição do abandono escolar: Pesquisa Aplicada à Gestão da Educação. Texto para discussão nº 58, Vitória – ES: Instituto Santos dos Jones Neves (IJSN), 2018.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN Jerome. **The elements of statistical learning: data mining, inference, and prediction.** 2. ed. New York: Springer series in statistics, 2017.

HOSMER JR., David W.; LEMESHOW, Stanley; STURDIVANT, Rodney X. **Applied logistic regression**. John Wiley & Sons, 2013.

JAMES, Gareth et al. **An introduction to statistical learning with applications in R**. New York: Springer, 2013.

JIMÉNEZ-GÓMEZ, Manuel Ángel et al. Discovering clues to avoid middle school failure at early stages. In: **Proceedings of the Fifth International Conference on Learning Analytics and Knowledge**. 2015. p. 300-304.

MARTINHO, Valquiria Ribeiro De Carvalho; NUNES, Clodoaldo; MINUSSI, Carlos Roberto. An intelligent system for prediction of school dropout risk group in higher education classroom based on artificial neural networks. In: **2013 IEEE 25th International Conference on Tools with Artificial Intelligence**. IEEE, 2013. p. 159-166.

DO NASCIMENTO, Rafaella LS; FAGUNDES, Roberta A. de A.; DE SOUZA, Renata MCR. Statistical learning for predicting school dropout in elementary education: A comparative study. **Annals of Data Science**, p. 1-28, 2021.

DO NASCIMENTO, Rafaella Leandra Souza et al. Educational data mining: An application of regressors in predicting school dropout. In: **Machine Learning and Data Mining in Pattern Recognition: 14th International Conference, MLDM 2018, New York, NY, USA, July 15-19, 2018, Proceedings, Part II 14**. Springer International Publishing, 2018. p. 246-257.

PARR, Alyssa K.; BONITZ, Verena S. Role of family background, student behaviors, and school-related beliefs in predicting high school dropout. **The Journal of Educational Research**, v. 108, n. 6, p. 504-514, 2015.

PSACHAROPOULOS, George; PATRINOS, Harry Anthony. Returns to investment in education: a decennial review of the global literature. **Education Economics**, v. 26, n. 5, p. 445-458, 2018.

CISTERNA, Sergi Rovira; PRATS, Eloi Puertas I.; MUÑOZ, Laura Igual. Data-driven System to Predict Academic Grades and Dropout. **PLoS One**, vol. 12, n. 2, 2017.

SANDOVAL-PALIS, Iván et al. Early dropout prediction model: A case study of university leveling course students. **Sustainability**, v. 12, n. 22, p. 9314, 2020.

SARA, Nicolae-Bogdan et al. High-School Dropout Prediction Using Machine Learning: A Danish Large-scale Study. **ESANN**, p. 319-324, 2015.

WOOD, Laura et al. Predicting dropout using student-and school-level factors: An ecological perspective. **School Psychology Quarterly**, v. 32, n. 1, p. 35, 2017.