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

RELATÓRIO

Fotografia: Freepik

Preditor do Abandono Escolar

Avaliação dos métodos de balanceamento e modelos de machine learning

Instituto Jones dos Santos Neves

GOVERNO DO ESTADO DO ESPÍRITO SANTO
Secretaria de Economia e Planejamento

Dezembro de 2024

Instituto Jones dos Santos Neves

Preditor do Abandono Escolar: avaliação dos métodos de balanceamento e modelos de machine learning.

Vitória, ES, 2024. 37p.; il. tab. (Relatório).

1.Educação. 2. Abandono Escolar. 3. Educação – Machine Learning. 4. Educação – Preditor. 5. Espírito Santo (Estado) I. Pereira, Guilherme Armando de Almeida (Fapes). II. Demura, Kiara de Deus. III. Pinto, Rosângela Vargas Davel. IV. Título.

As opiniões emitidas são exclusivas e de inteira responsabilidade do (os) 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
Antônio Ricardo F. da Rocha

Diretoria de Gestão Administrativa
Katia Cesconeto de Paula

Coordenação Geral
Kiara de Deus Demura

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

Revisão
Rosângela Vargas Davel Pinto – SEDU

Bibliotecário
Rosana Mariano Chagas

Fotografia Capa
Fotografia gerada por IA (Freepik)

Instituto Jones dos Santos Neves

Sumário

1. Introdução

O abandono escolar refere-se à interrupção prematura da educação formal, estando suas causas relacionadas a fatores socioeconômicos, familiares, individuais e institucionais.

Os impactos do abandono são diversos, tanto do ponto de vista individual quanto para a sociedade como um todo. A evasão, por conseguinte, perpetua e agrava as desigualdades sociais, contribuindo para o ciclo de pobreza e aumento da vulnerabilidade social, pois limita as oportunidades de emprego e desenvolvimento pessoal. Além disso, a evasão está associada a baixos salários, maiores riscos de envolvimento em atividades ilícitas, baixa produtividade, aumento do desemprego e subemprego. Do ponto de vista social, os custos associados a alta evasão estão relacionados com aumento dos gastos com assistência social, cuidados com a saúde, justiça criminal e dificuldade de atrair negócios que demandem trabalhadores altamente qualificados (PARR; BONITZ, 2015; BURGESS, 2016; UNICEF, 2017; WOOD et al., 2017).

Dessa forma, o desenvolvimento de uma ferramenta de previsão do abandono escolar é fundamental, pois auxilia gestores e educadores a identificarem previamente os estudantes mais propensos ao abandono, permitindo assim a elaboração de ações para mitigar esse seríssimo problema e evitar a evasão escolar. Algumas evidências sobre a efetividade de tais ferramentas são comentadas em UNICEF (2017).

Diversos modelos vêm sendo desenvolvidos na literatura, sendo sua grande maioria baseados em modelos de aprendizado estatístico ou *machine learning* tais como a regressão logística, o *Support Vector Machine* (SVM), o *Random Forest*, o *Classification And Regression Tree* (CART), o *Naive Bayes* e as Redes Neurais Artificiais (RNAs) (DEKKER et al., 2009; BAYER et al., 2012; SARA et al., 2015; KNOWLES, 2015; JIMÉNEZ-GÓMEZ et al., 2015; COSTA et al., 2017; ROVIRA et al., 2017; SANDOVAL-PALIS et al., 2020; ULDALL; ROJAS, 2022; RODRÍGUEZ, et al. 2023; PEREIRA, et al. 2024).

Esses modelos, muita das vezes, são estimados a partir de dados desbalanceados, isto é, quando a proporção entre as classes é desequilibrada. De fato, essa é uma

característica recorrente em estudos de modelos de predição do abandono escolar, pois o número de estudantes que abandonam é muito menor em relação ao número de estudantes que não abandonam. De acordo com Fernández et al. (2018), em situações como essa, é possível observar diversos trabalhos onde os modelos apresentaram desempenho inferior ao esperado. A literatura refere-se a problemas como este como "*imbalanced classification problems*".

Para lidar com este tipo de problema, uma alternativa consiste no pré-processamento de dados para que as classes fiquem equilibradas. Existem atualmente diversos algoritmos para tal tarefa, dentre eles pode-se citar o *Random Under-Sampling* (RUS), o *Synthetic Minority Over-Sampling Technique* (SMOTE) (CHAWLA et al., 2002) e o *Random Over-Sampling Examples* (ROSE) (MENARDI; TORELLI, 2014) entre outros.

Apesar da uma possível inclinação ao uso de tais algoritmos, para previsão do abandono, há diversos modelos com desempenhos satisfatórios que não consideram o problema do balanceamento tais como em Martinho, Nunes e Minussi (2013), Sara et al. (2015), Adelman et al. (2018), Freitas et al. (2020), Sandoval-Palis et al. (2020) e Pereira et al. (2024).

Por outro lado, é possível encontrar na literatura trabalhos que incorporam automaticamente os métodos de balanceamento em seus preditores do abandono (MARQUEZ-VERA et al., 2016; ROVIRA; PUERTAS; IGUAL, 2017; DEL BONIFRO et al., 2020). Outros autores preferem avaliar diversos métodos de balanceamento e selecionar o mais adequado para o seu caso, mesmo sem considerar a não utilização de tais métodos (SELIM; REZK, 2023; VILLAR; ANDRADE, 2024).

Finalmente, há trabalhos investigando a eficiência das metodologias de balanceamento, comparando modelos treinados com e sem esses algoritmos (MARQUEZ-VERA et al., 2016; COSTA et al., 2017; ROVIRA; PUERTAS; IGUAL, 2017; BARROS et al., 2019; LEE; CHUNG, 2019; OROOJI; CHEN, 2019; DEL BONIFRO et al., 2020; CHO; YU; KIM, 2023; KIM et al., 2023; PSATHAS; CHATZIDAKI; DEMETRIADIS, 2023; SELIM; REZK, 2023; WONGVORACHAN; HE; BULUT, 2023; VILLAR; ANDRADE, 2024).

A principal conclusão que podemos trazer desses trabalhos é que não há uma solução global para o problema do desbalanceamento da amostra. Alguns modelos foram construídos sem abordar tal tema e obtiveram bons resultados, outros autores encontraram evidências a favor dos métodos de balanceamento, enquanto outros não. Nesse sentido, testar métodos distintos e escolher o mais adequado seria o caminho mais adequado para a construção de um bom modelo de predição do abandono escolar.

Isto posto, este relatório, produto da parceria Estudos Educacionais[1], investiga a eficiência de métodos de balanceamento da amostra para a previsão do abandono escolar do ensino médio da rede estadual do estado do Espírito Santo. Para isso, avaliamos três métodos de balanceamento construídos a partir de conceitos metodológicos distintos. Além disso, como será mais bem detalhado ao longo deste texto, os algoritmos de pré-processamento são utilizados antes da etapa de estimação do modelo de predição. Assim, podemos testar os métodos de balanceamento em conjunto não só com a regressão logística[2], mas também com outros métodos de *machine learning*. Dessa forma, o segundo objetivo deste trabalho consiste na avaliação de métodos adicionais de *machine learning*, tais como SVM, *Random Forest, Naive Bayes, C5.0* e RNAs.

Os resultados indicam que o balanceamento da amostra é capaz de melhorar o desempenho preditivo dos modelos. O SMOTE e o RUS se apresentaram como os métodos com melhor desempenho. Contudo, ressalta-se que algumas configurações de balanceamento pioraram os resultados. Assim, sempre que os modelos forem estimados, é necessária uma avaliação prévia sobre qual a melhor forma de balancear a amostra para aquele conjunto de dados específico.

Tendo em vista os modelos adicionais de *machine learning*, pode-se perceber que nenhum método considerado apresentou resultados significativamente superiores à regressão logística. Além disso, os modelos com as melhores performances foram a regressão logística, o *Naive Bayes* e as RNAs. Dessa forma, considerando o custo

[1] Parceria entre a Secretaria de Estado da Educação (SEDU) do Espírito Santo, o Instituto Jones dos Santos Neves (IJSN) e a Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (Fapes).
[2] No primeiro relatório o modelo de previsão foi construído a partir de uma regressão logística com métodos de regularização para seleção de variáveis.

computacional, a fácil interpretação do modelo, bem como o seu bom desempenho, recomenda-se a utilização da regressão logística como modelo de classificação.

Este relatório está organizado da seguinte forma. A Seção 2 apresenta a revisão de literatura, enquanto os métodos de balanceamento, bem como os modelos de classificação, são apresentados na Seção 3. Na Seção 4 são apresentados os dois estudos de caso realizados. No primeiro, avaliamos a regressão logística estimada em amostras balanceadas e desbalanceadas. No segundo estudo de caso, testamos diversos modelos de *machine learning* calibrados em amostras balanceadas e desbalanceadas. Por fim, são realizadas as conclusões e apontamentos para estudos futuros.

2. Revisão de literatura

Do ponto de vista teórico, amostras desbalanceadas devem receber o tratamento matemático adequado, pois podem impactar a performance dos modelos de classificação. No entanto, empiricamente, nem sempre isso é necessário. Apesar dos dados de abandono serem um exemplo clássico de desbalanceamento, é importante destacar que existem vários modelos de previsão que não abordam essa questão, mas ainda assim, possuem resultados preditivos satisfatórios (MARTINHO; NUNES; MINUSSI, 2013; SARA et al., 2015; ADELMAN et al., 2018; SANDOVAL-PALIS et al., 2020; FREITAS et al., 2020; PEREIRA, et al. 2024).

Por outro lado, diversas pesquisas incorporaram automaticamente algoritmos de pré-processamento em suas metodologias. Marquez-Vera et al. (2016) desenvolveram um modelo baseado em programação genética. O modelo proposto foi comparado com os classificadores disponíveis no *Weka*[3] em conjunto com o *Synthetic Minority Over-Sampling Technique* (SMOTE) para previsão da evasão escolar na Unidade Acadêmica Preparatória da Universidade Autônoma de Zacatecas no México. Rovira, Puertas e Igual (2017) propuseram um modelo para a Universidade de Barcelona, Espanha, aplicando primeiramente o SMOTE para balancear os dados e, posteriormente, testando diversos

[3] Software de Machine Learning em Java: https://www.cs.waikato.ac.nz/ml/weka/index.html

modelos de *machine learning*[4]. De modo similar, Del Bonifro et al. (2020) balancearam a amostra por amostragem aleatória e analisaram distintos métodos de *machine learning*. Os modelos foram testados em pseudo-dados anonimizados de 15.000 alunos matriculados em diversos cursos de graduação.

Outros trabalhos tiveram o foco na comparação de diferentes métodos de balanceamento, buscando o mais adequado para o correspondente caso. Por exemplo, para dados do Egito, Selim e Rezk (2023) testaram mais de 12 técnicas de balanceamento. Entre as técnicas empregadas estão SMOTE, *Adaptive Synthetic* (ADASYN), e *Random Over-Sampling* (ROS). Contudo, os autores não compararam o desempenho do modelo sem os métodos de balanceamento. Os resultados indicaram que o ROS combinado com o NearMiss-3 melhorou o desempenho em termos da F1-Score e do erro tipo II[5].

Villar e Andrade (2024), para uma instituição de ensino superior, avaliam diferentes algoritmos de aprendizado supervisionados e não supervisionados em conjunto com técnicas de balanceamento (SMOTE e ADASYN). Os resultados indicaram que o SMOTE melhorou significativamente a acurácia do modelo.

Por fim, outro grupo de investigação consistiu na comparação entre modelos treinados com e sem algoritmos de balanceamento de dados. Lee e Chung (2019) empregaram o SMOTE associado aos modelos tradicionais de *machine learning* para análise de 165.715 estudantes do ensino médio da Coreia do Sul. Os resultados são ambíguos, uma vez que, dependendo das métricas consideradas, os métodos do balanceamento melhoraram ou não o desempenho.

Resultados semelhantes foram encontrados por Costa et al. (2017) que empregaram o SMOTE e avaliaram o desempenho de quatro modelos de *machine learning*[6] para predição da reprovação acadêmica em cursos introdutórios de programação de uma universidade pública brasileira. Os autores analisaram dois conjuntos de dados independentes: um relacionado aos cursos presenciais e outro relacionado aos cursos

[4] Regressão logística, Gaussian Naive Bayes, SVM, *Random Forest* e adaptive boosting.
[5] Ver Apêndice A – Métricas para as avaliações das previsões.
[6] SVM, árvore de decisão, RNA e, ingênuo Bayes.

on-line. Em termos de desempenho, em relação aos cursos *on-line*, o balanceamento melhorou a qualidade da identificação, enquanto o efeito inverso ocorreu nos cursos presenciais.

Ainda no Brasil, Barros et al. (2019) testaram três modelos de previsão[7] e compararam os seus desempenhos quando treinandos com amostras balanceadas e desbalanceadas. Os métodos de balanceamento utilizados foram *Random Under-Sampling* (RUS), SMOTE e ADASYN. Para o estudo, foram considerados 7.718 alunos do ensino integrado (ensino médio com formação em educação profissional) do Instituto Federal do Rio Grande do Norte. Os resultados indicaram que técnicas de balanceamento puderam aumentar significativamente o desempenho de modelos preditivos.

Por outro lado, Orooji e Chen (2019), utilizando dados administrativos da Louisiana, avaliaram distintas abordagens para lidar com o desbalanceamento, além das técnicas de pré-processamento de dados tais como *Case Weighting* e *Cost-Sensitive Analysis*. Os resultados são ambíguos uma vez que o balanceamento produziu efeitos positivos sobre a métrica sensibilidade, mas reduziu a precisão do modelo, enquanto os modelos treinados sem balanceamento obtiveram uma melhor precisão, porém uma baixa sensibilidade.

Wongvorachan, He e Bulut (2023) compararam várias técnicas de balanceamento em dados com diferentes proporções de classes, isto é, dados com desbalanceamento moderado e desbalanceamento extremo. A análise ocorre com os dados do *High School Longitudinal Study* de 2009, dos Estados Unidos da América. Os autores compararam métodos tais como ROS, RUS, SMOTE e um método híbrido entre RUS e SMOTE. O modelo de classificação utilizado foi o *Random Forest.* Além disso, os autores avaliaram o desempenho das *Random Forests* sem o balanceamento dos dados, concluindo que o balanceamento pode auxiliar o desempenho dos modelos.

Por outro lado, Cho, Yu e Kim (2023) propuseram um modelo para a Universidade *Sahmyook*, na República da Coreia. Eles testaram diferentes modelos de *machine*

[7] Árvores de decisão, RNA e *Balanced Bagging*.

learning[8] em conjunto com vários métodos de balanceamento tais como SMOTE, ADASYN e *Borderline*-SMOTE. Os resultados indicaram que o desempenho de todos os modelos diminuiu após a aplicação das técnicas de balanceamento, exceto o modelo de *Random Forest* e, além disso, nesse caso a melhora não foi significativa.

Psathas, Chatzidaki e Demetriadis (2023) consideram a combinação de diferentes técnicas de balanceamento, tais como SMOTE, *Borderline*-SMOTE e ADSYN, juntamente com algoritmos tradicionais de *machine learning*[9] para a previsão da evasão em cursos *on-line* (*Massive Open Online Courses* – MOOCs). Além disso, os autores também avaliaram o modelo sem balancear a amostra. Os resultados indicaram que o balanceamento pode melhorar significativamente o desempenho dos modelos. No entanto, os autores alertam que combinações distintas entre métodos de classificação e algoritmos de balanceamento devem ser testadas, uma vez que não há garantia de que todas as configurações melhorarão o desempenho. Nessa mesma linha, Kim et al. (2023) propuseram um sistema de previsão da evasão escolar para a Universidade Nacional de *Gyeongsang*, na Coreia do Sul. Para esta tarefa, os autores testaram várias combinações de algoritmos de balanceamento e métodos de classificação de aprendizado de máquina.

Como se vê, não há solução única e geral para o problema dos dados desbalanceados. Alguns modelos foram construídos sem lidar com essa característica e alcançaram bons resultados; outros autores encontraram evidências a favor dos métodos de balanceamento, enquanto outros não. Nesse sentido, os trabalhos demonstram a importância de se testar métodos distintos e escolher o mais adequado para cada caso.

[8] Regressão Logística, Árvore de Decisão, *Random Forest*, SVM, RNA Profunda e LightGBM (*Light Gradient Boosting Machine*).
[9] Regressão logística, SVM e Naive Bayes.

3. Ferramental matemático

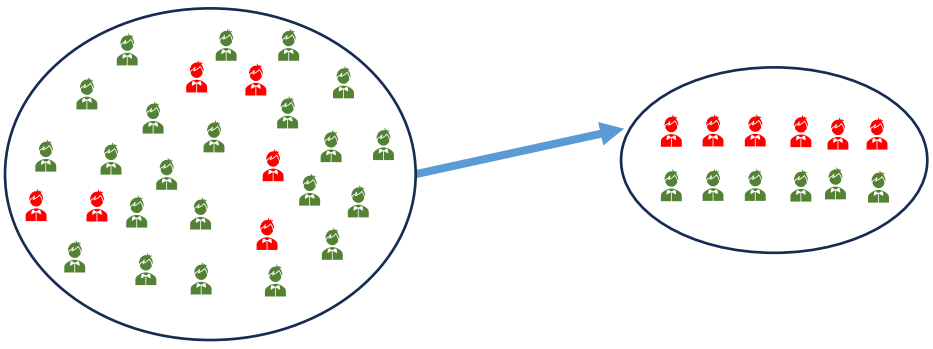
3.1. Métodos de balanceamento

Nesta seção, apresentamos brevemente os métodos empregados para o balanceamento da amostra nesta pesquisa: *Random Under-Sampling* (RUS), *Synthetic Minority Over-Sampling Technique* (SMOTE), e *Random Over-Sampling Examples* (ROSE). Para uma compreensão mais abrangente desses métodos, recomendamos Chawla et al. (2002), He e Garcia (2009), Ma e He (2013), Menardi e Torelli (2014) e Fernández et al. (2018).

Random Under-Sampling (RUS)

Envolve a construção de uma amostra balanceada, removendo aleatoriamente os indivíduos da classe majoritária, reduzindo assim seu tamanho e preservando todos os elementos da classe minoritária. Como resultado, essa abordagem diminui o tamanho da amostra. A Figura 1 ilustra essa abordagem simples.

Figura 1 – Exemplo de amostra balanceada

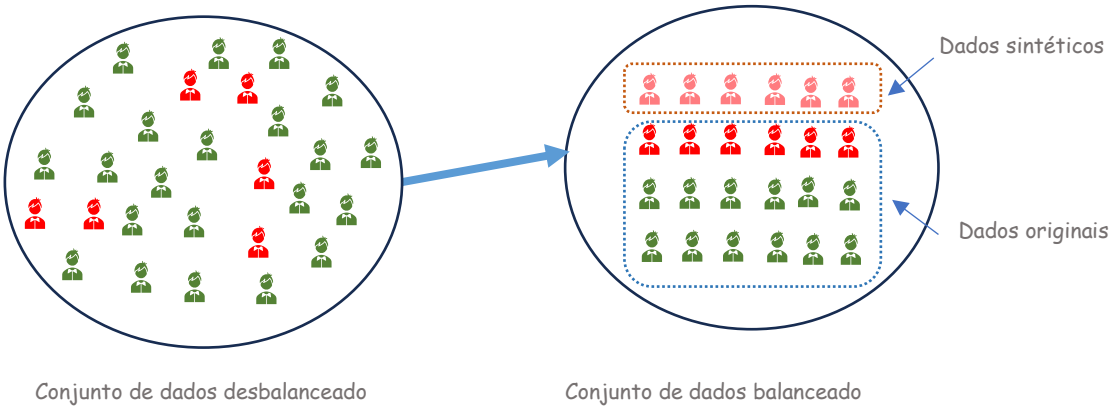


Elaboração: Estudos Educacionais/IJSN.
Nota: Os elementos verdes são eliminados por amostragem aleatória. A amostra resultante contém, em geral, o mesmo número de elementos de ambas as classes. Após essa etapa, estima-se o modelo de classificação utilizando o conjunto de dados balanceado.

Synthetic Minority Over-Sampling Technique (SMOTE)

Chawla et al. (2002) propuseram interpolar elementos vizinhos da classe minoritária para criar elementos sintéticos. Consequentemente, o algoritmo aumenta a classe minoritária incorporando dados simulados, enquanto remove aleatoriamente elementos da classe majoritária. Segundo os autores, essa abordagem resulta em uma região de decisão maior e menos específica, aumentando a qualidade da estimativa. A Figura 2 ilustra a construção da amostra final.

Figura 2 – Exemplo de uma amostra balanceada usando o SMOTE

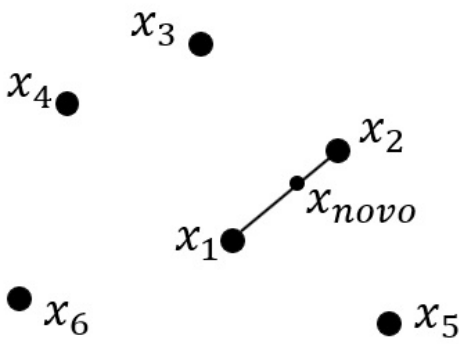


Elaboração: Estudos Educacionais/IJSN.
Nota: Os elementos representados em vermelho claro são gerados via SMOTE e adicionados ao conjunto de dados. Simultaneamente, alguns elementos da classe majoritária são descartados por sorteio aleatório.

A Figura 3 ilustra um processo de geração de uma nova instância sintética. Suponha que o vetor x_1 contenha todas as variáveis relativas a um indivíduo (instância) e seja proveniente da classe minoritária. Em seguida, procura-se o vizinho ou os vizinhos mais próximos[10]. Na Figura 3, o vizinho mais próximo é representado por x_2. Assim, computa-se a distância ente x_1 e x_2, sendo a nova observação sintética dada por $x_{novo} = x_1 + \alpha(x_2 - x_1)$ onde α é um número aleatório entre 0 e 1.

[10] Geralmente utiliza-se 5 vizinhos mais próximos.

Figura 3 – Criação de um dado sintético via SMOTE



Elaboração: Estudos Educacionais/IJSN.

Em outras palavras, considere que $\mathrm{x}_1 = [1, 8, 4]$ e que $\mathrm{x}_2 = [5, 8, 7]$. Dessa forma, a nova observação (vetor) sintética é dada por:

$$x_{novo} = \begin{bmatrix} 1 \\ 8 \\ 4 \end{bmatrix} + \alpha \left(\begin{bmatrix} 1 \\ 8 \\ 4 \end{bmatrix} - \begin{bmatrix} 5 \\ 8 \\ 7 \end{bmatrix} \right), \quad \alpha \in (0, 1) \tag{1}$$

Com esse princípio, novas observações que respeitam as características iniciais das classes minoritárias são geradas. Vale lembrar que no SMOTE a classe majoritária é reduzida por meio de sorteios aleatórios de forma similar ao RUS. Para uma descrição pormenorizada, veja Chawla et al. (2002).

Random Over-Sampling Examples (ROSE)

Este método foi proposto por Menardi e Torelli (2014) e se baseia nos princípios do SMOTE. A principal diferença está na geração dos dados sintéticos. No ROSE, os autores substituem a técnica de interpolação por núcleos (*kernel*) estatísticos. Assim, segundo Menardi e Torelli (2014), este processo de simulação de dados sintéticos corresponde ao processo de geração de dados artificiais a partir de uma densidade *kernel* estimada para $f(\boldsymbol{x}|y_j),\ j = 0, 1$, em que $f(\cdot)$ indica a função de densidade condicional.

Isso permite uma expansão da região de decisão para além da interpolação, resultando em dados simulados que são diferentes do original, mas que são equiprováveis do ponto de vista estatístico. Como resultado, há uma expansão da região de decisão (MENARDI; TORELLI, 2014).

3.2. Modelos de classificação

Os métodos de balanceamento da amostra constituem uma etapa que pode ser entendida como pré-processamento dos dados. Assim, após a aplicação dos métodos apresentados na subseção anterior, é necessário que utilizemos algum modelo de classificação.

Esta seção apresenta, de forma resumida, os modelos utilizados como preditores do abandono neste trabalho. Os modelos utilizados são: i) Regressão logística; ii) *Random Forest*; iii) *C5.0*; iv) *Naive Bayes*; v) *Support Vector Machine* (SVM), e; vi) Redes Neurais Artificiais (RNAs). Vale lembrar que esses modelos serão estimados, ao longo deste relatório, tanto em amostras balanceadas quanto não balanceadas.

A variável de interesse, isto é, aquela na qual iremos fazer as classificações é definida da seguinte forma:

$$y_i = \begin{cases} 0 \text{ (negativo)}, & caso\ o\ aluno\ i\ não\ abandone\ a\ escola. \\ 1 \text{ (positivo)}, & caso\ o\ aluno\ i\ abandone\ a\ escola. \end{cases} \quad (2)$$

A seguir, é realizada uma breve introdução a cada um dos modelos.

Regressão logística

A regressão logística para o caso binário pode ser descrita como

$$Prob(Y = 1/\boldsymbol{X}) = F(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k) \quad (3)$$

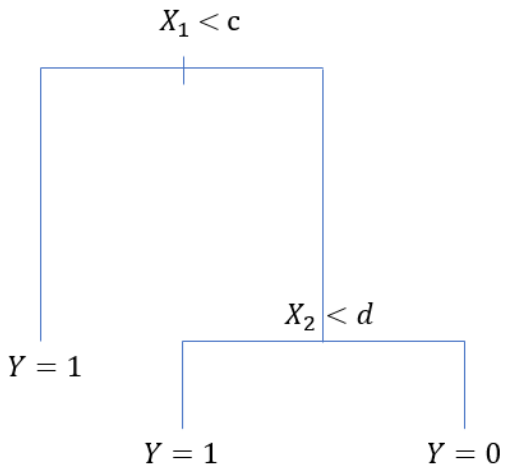
onde $F(.)$ é uma função de distribuição acumulada logística dada por $F(u) = \frac{exp(u)}{1+ exp(u)}$ de modo que $0 < F(u) < 1,\ u \in \mathcal{R}$. Para a seleção das variáveis utilizou-se o LASSO. A grande vantagem de tal abordagem reside no fato que tal ferramenta é capaz de selecionar as variáveis mais relevantes simultaneamente à estimação dos parâmetros. Assim, o modelo final é constituído por um subconjunto das variáveis explicativas mais relevantes. Para mais detalhes veja James et al. (2013) e Hastie, Tibshirani e Friedman (2017).

Random Forest & C5.0

O *Random Forest*, ou em tradução livre, árvores aleatórias são métodos de classificação baseados em árvores de decisão que surgem a partir da ideia de estratificação e segmentação do espaço dos preditores (JAMES et al., 2013).

Para ilustrarmos a ideia das árvores de classificação (*classification tree*), vamos utilizar a Figura 4, onde pretende-se classificar a variável Y a partir de X_1 e X_2. Neste exemplo, X_1 é a variável mais importante para a previsão. Quando X_1 é menor que determinado valor c, classificamos Y como 1. Quanto X_1 é maior que c, a classificação passa a depender também da variável X_2. Assim, quando $X_1 > c\ e\ X_2 \leq d$, classificamos Y também como 1. Por outro lado, quando $X_1 > c\ e\ X_2 > d$, classificamos Y como 0.

Figura 4 – Exemplo de uma árvore de decisão com duas variáveis explicativas



Elaboração: Estudos Educacionais/IJSN.

Tanto o *Random Forest* quanto o *C5.0* são algoritmos que exploram a essência das árvores de decisão para a classificação, sendo a diferença entre os modelos dada pela forma na qual as árvores são construídas e implementadas. Para mais detalhes, veja James et al. (2013) e Kuhn e Johnson (2013).

Naive Bayes

O *Naive Bayes* é um modelo de classificação de base probabilística com origem na inferência bayesiana. Um dos seus principais pressupostos é a independência entre as variáveis explicativas. Dessa forma, a probabilidade de abandono pode ser obtida da seguinte forma:

$$Prob(Y = 1/X = x) = \frac{\pi_1 \text{Prob}(X = x/Y = 1)}{\sum_{i=0}^{1} \pi_i \text{Prob}\,(X = x/Y = i)} \qquad (4)$$

onde $\pi_i, i = \{0,1\}$ é a probabilidade *a priori* da classe i. Essas probabilidades, por exemplo, podem ser obtidas ao se calcular a proporção de alunos que abandonam e não abandonam a partir dos dados históricos. Neste trabalho, as probabilidades apresentadas na equação (3) são construídas a partir de uma distribuição Gaussiana, contudo, pode-se utilizar outras distribuições. Para mais detalhes, veja James et al. (2013).

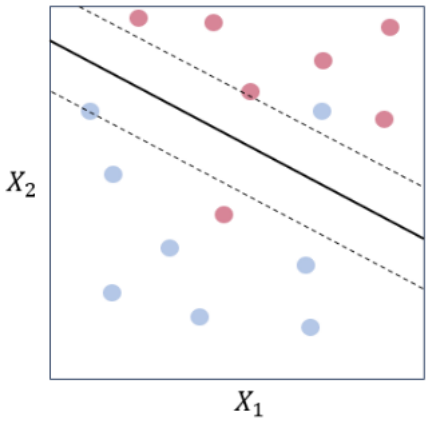
Support Vector Machine (SVM)

Os SVMs surgiram na ciência da computação nos anos 90 e ganharam popularidade desde então (JAMES et al., 2013). Inicialmente, esses modelos constroem hiperplanos para classificar um vector X em classes. A Figura 5 apresenta um exemplo com duas variáveis. Como pode ser observado, existem duas classes separadas por duas linhas tracejadas e uma linha contínua. A ideia do SVM é construir esses limites de modo que as linhas tracejadas possuam a menor distância para a classe mais próxima enquanto a linha central contínua possui a maior distância possível em relação a cada um dos pontos

mais próximos de cada classe. Obviamente, essa classificação não é perfeita, aceitando o modelo alguns erros de classificação. Contudo, o objeto é a construção de fronteiras que classifiquem corretamente o maior número possível de observações.

Na Figura 5, as linhas apresentadas expressam fronteiras lineares, contudo, nem sempre tais limites serão dessa forma. É importante destacar que os SVMs também são capazes de criar fronteiras não lineares. Para mais detalhes, veja James et al. (2013).

Figura 5 – SVM ajustado para uma pequeno conjunto de dados com duas classes



Elaboração: Estudos Educacionais/IJSN.
Nota: Classes representadas pelas cores azul e vermelha. O hiperplano separando os grupos é representado pela linha contínua enquanto as linhas tracejadas são denominadas margens. Como pode ser observado, o SVM apresentado é capaz de classificar corretamente a maior parte dos pontos. Além disso, é importante destacar que os SVMs também são capazes de produzir limites não lineares para classificação.

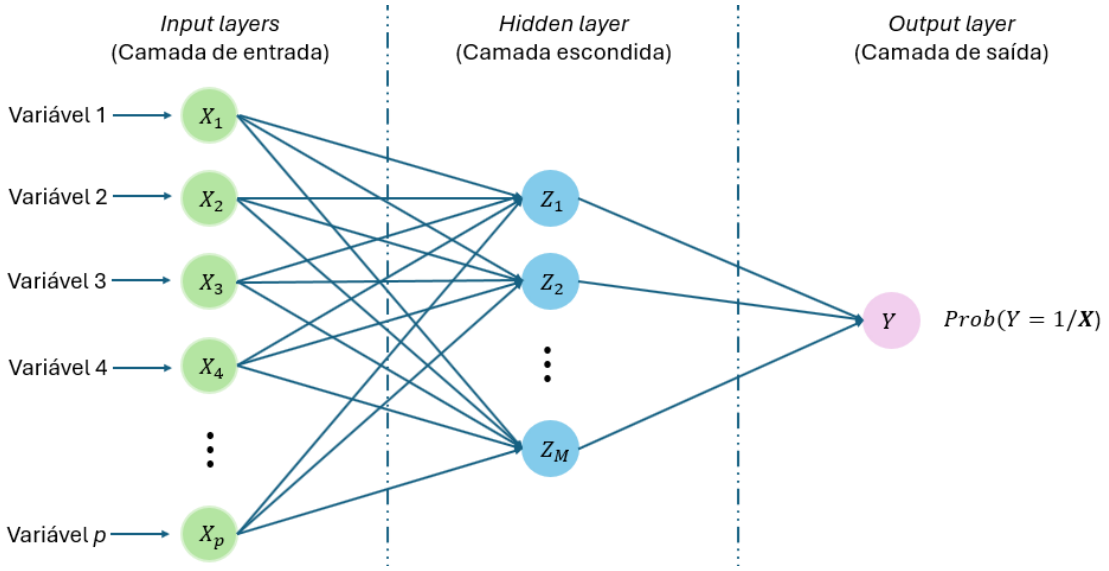
Redes Neurais Artificias (RNAs)

As RNAs são aproximadores universais de funções contínuas e, por esse motivo, são uma ferramenta extremamente flexível para a construção de modelos de classificação/previsão. Contudo, uma de suas desvantagens é a redução significativa da interpretação do modelo (HASTIE; TIBSHIRANI; FRIEDMAN, 2017). Por esse motivo, muitos classificam as RNAs como *black-box.*

Em linha gerais, as RNAs são compostas por neurônios artificiais distribuídos por camadas e conectados entre si. Atualmente, existem diversos modelos de RNAs, diferenciando-se entre si por meio do tipo de conexão entre os neurônios, número de camadas, algoritmo de aprendizado para a estimação, funcionalidades, entre outros.

Neste trabalho, empregamos uma RNA denominada *Single-Hidden Neural Network.* A Figura 6 ilustra o diagrama dessa RNA. Como pode ser observado, essa estrutura pode ser entendida como um modelo de classificação em dois estágios. Na camada de entrada da RNA estão os neurônios (*inputs*) relacionados às variáveis explicativas $X_1, \dots, X_p$. Posteriormente, na camada escondida (*hidden layer*), estão as variáveis (neurônios) $Z_1, \dots, Z_M$ que são construídas a partir de uma combinação linear das variáveis explicativas. Por fim, a variável Y (variável prevista ou *target*) é modelada novamente como uma combinação linear de $Z_1, \dots, Z_M$. Para mais detalhes, veja Hastie, Tibshirani e Friedman (2017).

Figura 6 – Diagrama de uma RNA com apenas uma camada escondida



Elaboração: Estudos Educacionais/IJSN.
Nota: Os nós representam os neurônios enquanto as arrestas representam as conexões (sinapses) entre os neurônios.

4. Resultados

Esta seção apresenta o banco de dados utilizado, bem como os resultados dos dois estudos de caso realizados.

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 Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP).

Dessa forma, essas fontes de dados permitem a elaboração de um banco de dados que contenha informações sobre características físicas, socioeconômicas e de desempenho dos estudantes. Além dessas, também é possível obter informações sobre a infraestrutura da escola e sobre as turmas. Todas as variáveis dizem respeito ao final do primeiro trimestre do correspondente ano. A Tabela 1 lista as variáveis empregadas nesse estudo.

Tabela 1 – Variáveis consideradas

(continua)

Categoria	Variáveis	
Pessoal	▪ Idade. ▪ Sexo. ▪ Cor/raça. ▪ Zona residencial. ▪ Necessidade especial?	▪ Turno. ▪ Quantidade de mudança de escola. ▪ Quantidade de mudança de turma. ▪ Quantidade de mudança de turno. ▪ Quantidade de mudança de município.
Desempenho acadêmico	▪ Notas em: ▪ geografia; ▪ português; ▪ matemática; ▪ sociologia; ▪ filosofia; ▪ educação física; ▪ história; ▪ biologia; ▪ física; ▪ química; ▪ artes.	▪ Proporção de faltas em: ▪ geografia; ▪ português; ▪ matemática; ▪ sociologia; ▪ filosofia; ▪ educação física; ▪ história; ▪ biologia; ▪ física; ▪ química; ▪ artes.

Tabela 2 – Variáveis consideradas

(continuação)

Categoria	Variáveis	
Infraestrutura da Escola	▪ Variáveis binárias para identificar na escola a existência de: ▪ almoxarifado; ▪ área verde; ▪ auditório; ▪ biblioteca; ▪ laboratórios de ciência e informática; ▪ pátio coberto; ▪ parque infantil; ▪ quadra de esportes; ▪ sala de leitura; ▪ secretaria; ▪ sala para atendimento especial; ▪ refeitório; ▪ equipamento multimídia; ▪ computador; ▪ equipamento de copiadora;	▪ equipamento de impressora; ▪ equipamento de DVD; ▪ equipamento de som; ▪ equipamento de TV; ▪ banda larga; ▪ internet; ▪ alimentação. ▪ Variável indicando se o prédio da escola é compartilhado. ▪ Quantidade de: ▪ salas utilizadas; ▪ equipamentos multimídia disponíveis; ▪ equipamento de DVD; ▪ equipamentos de som e equipamentos de tv.
Características de desempenho da turma e escola	▪ Taxa de distorção idade-série da turma. ▪ Taxa de distorção idade-série do turno da escola. ▪ Taxa de distorção idade-série da escola. ▪ Nota média da turma em português. ▪ Nota média da turma em matemática.	▪ Nota média da escola em português. ▪ Nota média da escola em matemática. ▪ Estudantes totais da turma. ▪ Média de estudantes por turma do turno da escola. ▪ Média de estudantes por turma da escola.

Elaboração: Estudos Educacionais/IJSN.

Apenas para evidenciar a questão do desbalanceamento dos dados, a Tabela 2 apresenta o número de abandonos nas séries do ensino médio da rede estadual do Espírito Santo com base nos dados supracitados. Como pode ser observado, a proporção de estudantes que não abandonam a escola é consideravelmente superior, criando assim um exemplo de dados desbalanceados.

Tabela 3 – Número de abandonos para os anos de 2019 e 2020

Classificação	1ª série		2ª série		3ª série	
	2019	2020	2019	2020	2019	2020
Abandona	751	1.025	423	463	214	161
Não Abandona	35.606	34.164	23.313	25.509	19.091	19.178

Elaboração: Estudos Educacionais/IJSN.

4.2. Estudo de caso I – Investigando o problema do balanceamento

Esta seção investiga se o desempenho preditivo pode ser melhorado com técnicas de balanceamento. Assim, utilizaremos três métodos de balanceamento para equilibrar a amostra[11]: *Random Under-Sampling* (RUS)[12], *Synthetic Minority Over-Sampling Technique* (SMOTE) e *Random Over-Sampling Examples* (ROSE). Em seguida, utilizaremos as amostras balanceadas para estimação do modelo de classificação (regressão logística). A título de comparação, estimaremos também a regressão logística na amostra desbalanceada (original). Com isso, podemos avaliar se o balanceamento auxilia na identificação dos estudantes em risco de abandono.

Os dados para o treinamento são relativos ao 1º trimestre de 2019 da 1ª, 2ª e 3ª séries do ensino médio da rede estadual do Espírito Santo. Após a estimação do modelo, realizamos a previsão para o ano de 2020. O procedimento para uso do preditor, neste caso, pode ser organizado da seguinte forma:

1. Com os dados do 1º trimestre de 2019, realizamos o balanceamento;
2. Em posse de uma amostra balanceada, estima-se a regressão logística;
3. Em seguida, fazemos as previsões para o ano de 2020 com base nos dados relativos ao 1º trimestre de 2020. É importante destacar que esses dados não são utilizados para estimação do preditor.

Diferentes possibilidades em termos de proporções de classes foram avaliadas. Inicialmente optou-se por um balanceamento perfeito onde cada classe possuía 50% das observações. Em seguida, testou-se uma proporção de 25%-75%, isto é, a amostra balanceada sendo constituída por 25% dos estudantes que abandonam e 75% de estudantes que não abandonam.

Além disso, também investigamos se aumentar o tamanho da amostra com dados de anos anteriores a 2019 traria alguma melhoria em termos de previsão. Para isto, em

[11] Os métodos de balanceamento são aplicados ao banco de dados antes da estimação do modelo. Dessa forma, primeiro aplica-se o balanceamento e, posteriormente, estima-se o modelo.
[12] São sinônimos: *Random Under-Sampling*, *Downsampling* e *DownSample*.

alguns casos, estimamos o modelo com base numa amostra composta pelos anos de 2018 e 2019.

Todas as configurações testadas para cada série de ensino médio podem ser elencadas da seguinte maneira:

A. Amostra sem balanceamento, considerando apenas o ano de 2019;
B. Amostra sem balanceamento, considerando informações de 2018 e 2019;
C. Balanceamento RUS com proporção de 50% das classes, apenas com o ano de 2019;
D. Balanceamento RUS com proporção de 75% (não abandono) e 25% (abandono) em 2019;
E. Balanceamento RUS com proporção de 50% das classes, considerando os anos de 2018 e 2019;
F. Balanceamento RUS com proporção de 75% (não abandono) e 25% (abandono), considerando os anos de 2018 e 2019;
G. Balanceamento SMOTE com proporção de 50% das classes, apenas com o ano de 2019;
H. Balanceamento SMOTE com proporção de 75% (não abandono) e 25% (abandono) em 2019;
I. Balanceamento SMOTE com proporção de 50% das classes, considerando os anos de 2018 e 2019;
J. Balanceamento SMOTE com proporção de 75% (não abandono) e 25% (abandono), considerando os anos de 2018 e 2019;
K. Balanceamento ROSE com proporção de 50% das classes, apenas com o ano de 2019;
L. Balanceamento ROSE com proporção de 75% (não abandono) e 25% (abandono) em 2019;
M. Balanceamento ROSE com proporção de 50% das classes, considerando os anos de 2018 e 2019;
N. Balanceamento ROSE com proporção de 75% (não abandono) e 25% (abandono), considerando os anos de 2018 e 2019.

Os hiperparâmetros da regressão logística foram definidos via *10-fold cross-validation* e *grid search* (JAMES et al., 2013). A métrica considerada para seleção do melhor modelo foi a área sob a curva característica de operação do receptor (*Receiver Operating Characteristic Curve* – curva ROC). Os modelos foram implementados via pacote Caret (*Classification And Regression Training*), desenvolvido por (KUHN, 2008). Além disso, utilizamos um Intel(R) Core (TM) i7-10700 CPU @2.90GHz com 16GB de RAM com processamento paralelo via o pacote *doParallel.*

As Figuras 7, 8, 9, 10 e 11 apresentam os resultados para as métricas consideradas. A definição das métricas, bem como suas intepretações, podem ser encontradas no Apêndice A.

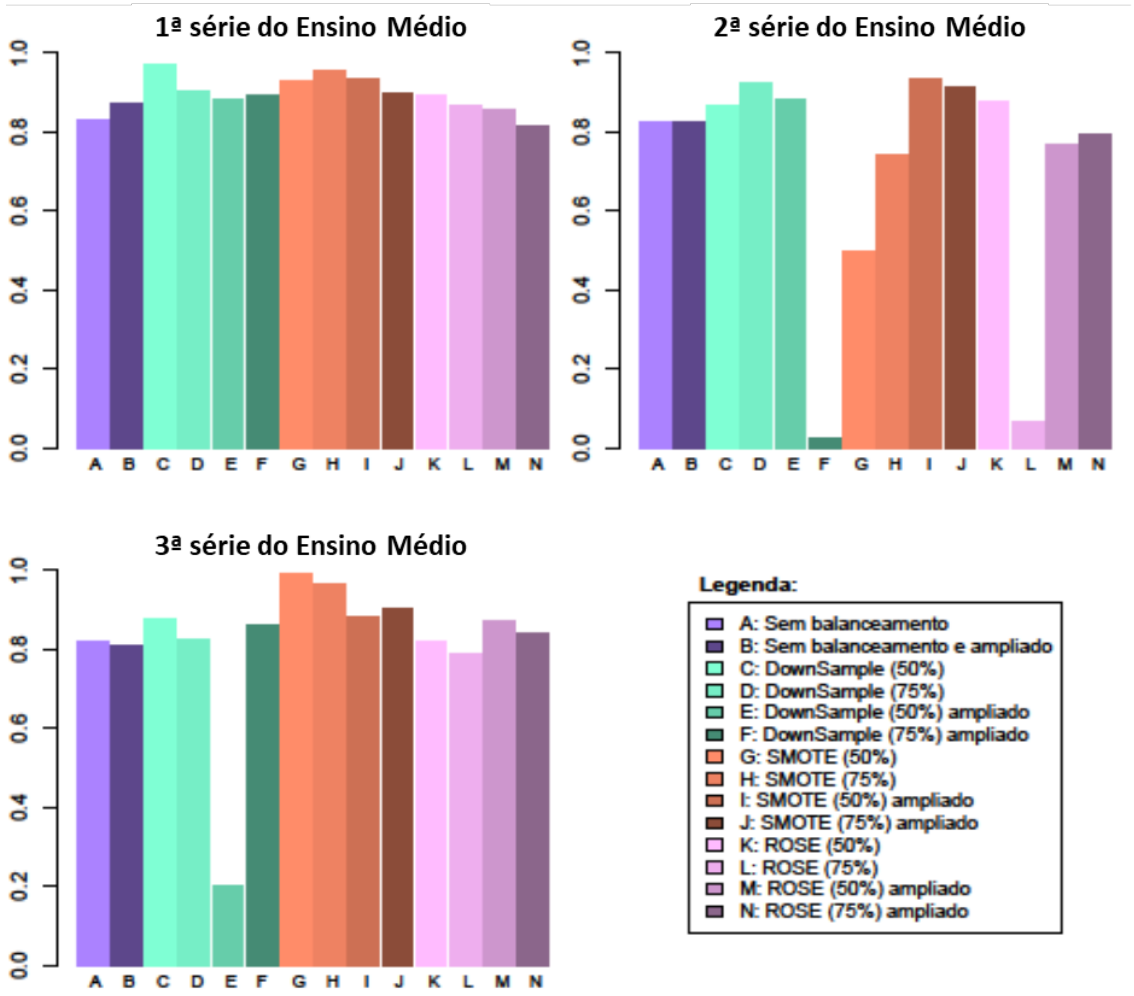
Tendo em vista a acurácia, a Figura 7 apresenta os resultados para as três séries de ensino. Como pode ser observado, o balanceamento melhorou o desempenho na maioria das configurações testadas.

Para a 1ª série do ensino médio, todas as configurações de SMOTE e *Downsampling* (RUS) foram superiores aos modelos estimados com dados desbalanceados. Com relação às 2ª e 3ª séries, é possível encontrar também configurações de SMOTE e *Downsampling* capazes de melhorar a performance preditiva do modelo. Além disso, se observarmos apenas as melhores configurações é possível perceber que SMOTE e *Downsampling* estão entre as melhores performances em todas as séries do ensino médio.

Com relação ao ROSE, esta técnica apresentou desempenho inferior ao SMOTE e *Downsampling*, não sendo capaz de superá-los. Quando comparado aos modelos estimados sem balanceamento, para as três séries há uma configuração de ROSE capaz de melhorar a previsão do abandono.

Por fim, em termos de acurácia, ampliar a base de treinamento (aumentar o número de anos do histórico) não refletiu em melhorias em termos de previsão.

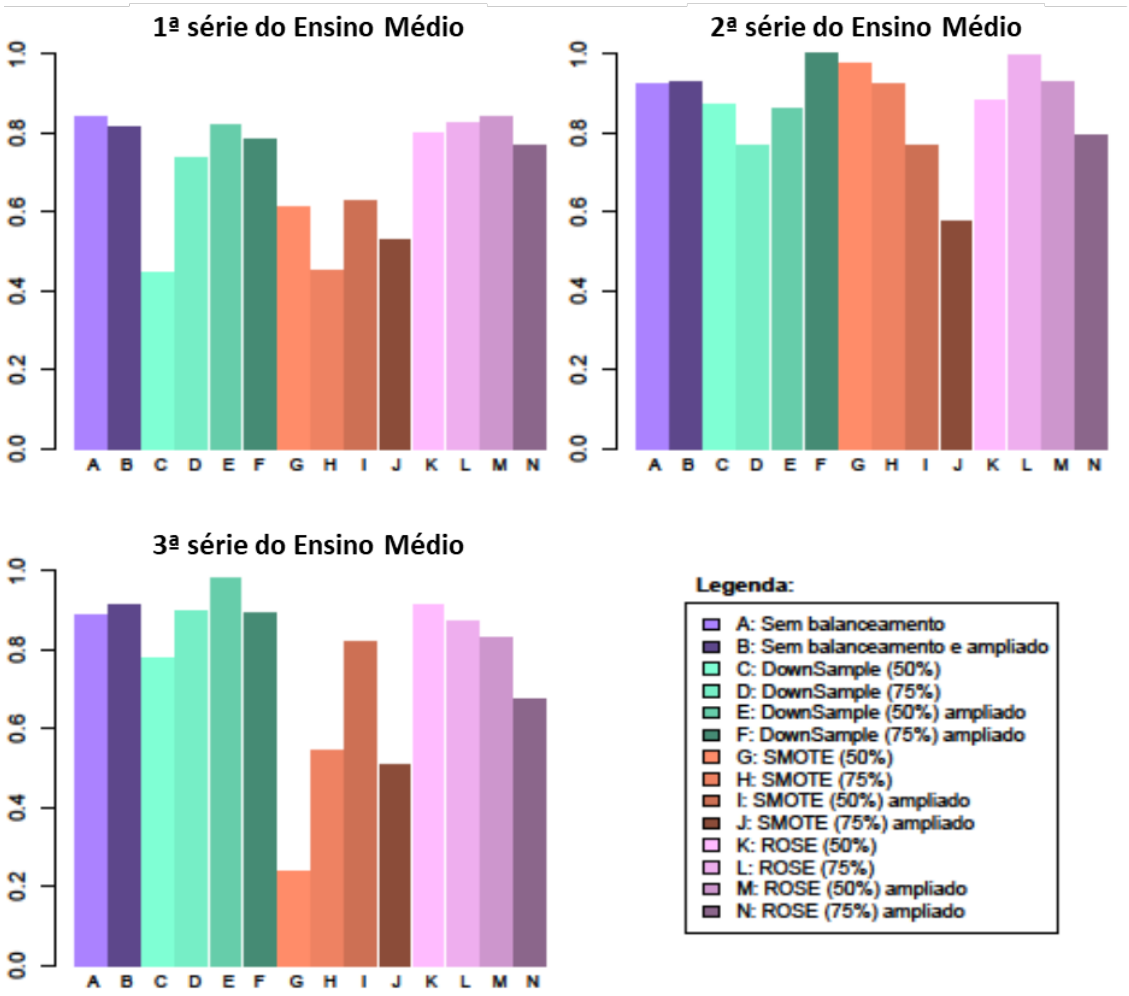
Figura 7 – Acurácia



Elaboração: Estudos Educacionais/IJSN.
Nota: As barras indicam a acurácia encontrada. Quanto mais próximas de um, melhor o modelo. As barras foram categorizadas em tons de cores específicas. Os modelos estimados em dados sem balanceamento são representados por tons de roxo, as barras verdes indicam o Downsampling enquanto as barras laranjas representam as diferentes configurações de SMOTE. Por fim, a cor rosa foi reservada para o ROSE.

A Figura 8 apresenta os resultados para a métrica sensibilidade. Para a 1ª série do ensino médio, observa-se que nenhum método de balanceamento foi capaz de melhorar as previsões. Tal resultado, contudo, não é observado para a 2ª e a 3ª séries onde existem configurações capazes de melhorar o desempenho preditivo. É importante destacar o número considerável de configurações em que o balanceamento piorou o desempenho do modelo. Isto pode ocorrer pois os métodos empregados descartam observações da classe majoritária aleatoriamente. Dessa forma, em alguns casos, pode haver o descarte de observações relevantes.

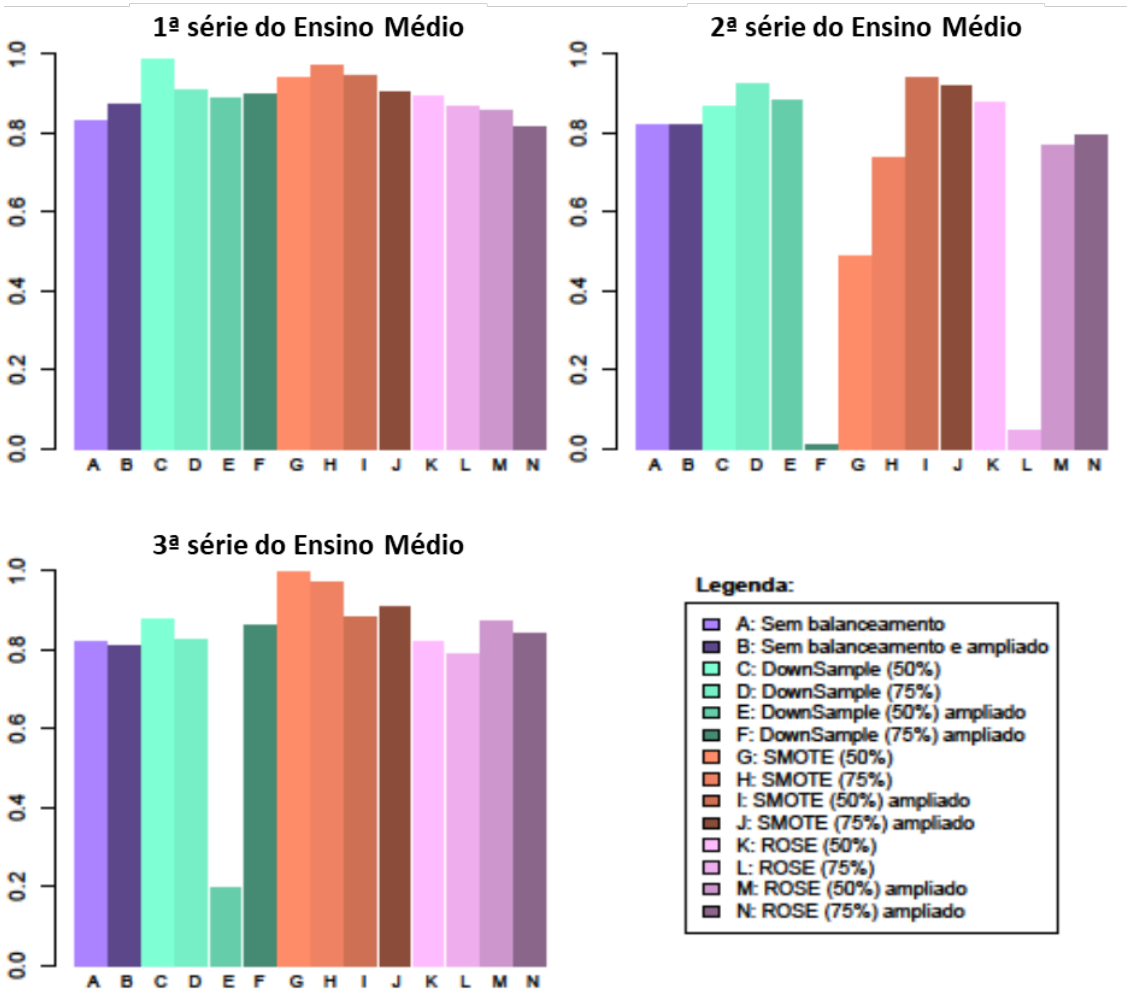
Figura 8 – Sensibilidade



Elaboração: Estudos Educacionais/IJSN.
Nota: As barras indicam a sensibilidade encontrada. Quanto mais próximas de um, melhor o modelo. As barras foram categorizadas em tons de cores específicas. Os modelos estimados em dados sem balanceamento são representados por tons de roxo, as barras verdes indicam o Downsampling enquanto as barras laranjas representam as diferentes configurações de SMOTE. Por fim, a cor rosa foi reservada para o ROSE.

Tendo em vista a especificidade, a Figura 9 apresenta os correspondentes resultados. Para a 1ª série, praticamente todas as configurações melhoram o desempenho. Por outro lado, para a 2ª série, algumas configurações (C, D, E, I, J e K) melhoraram a performance. Para a 3ª série, o SMOTE foi o melhor método em todas as configurações testadas. Por fim, é importante notar que há casos em que o desempenho preditivo foi impactado negativamente.

Figura 9 – Especificidade

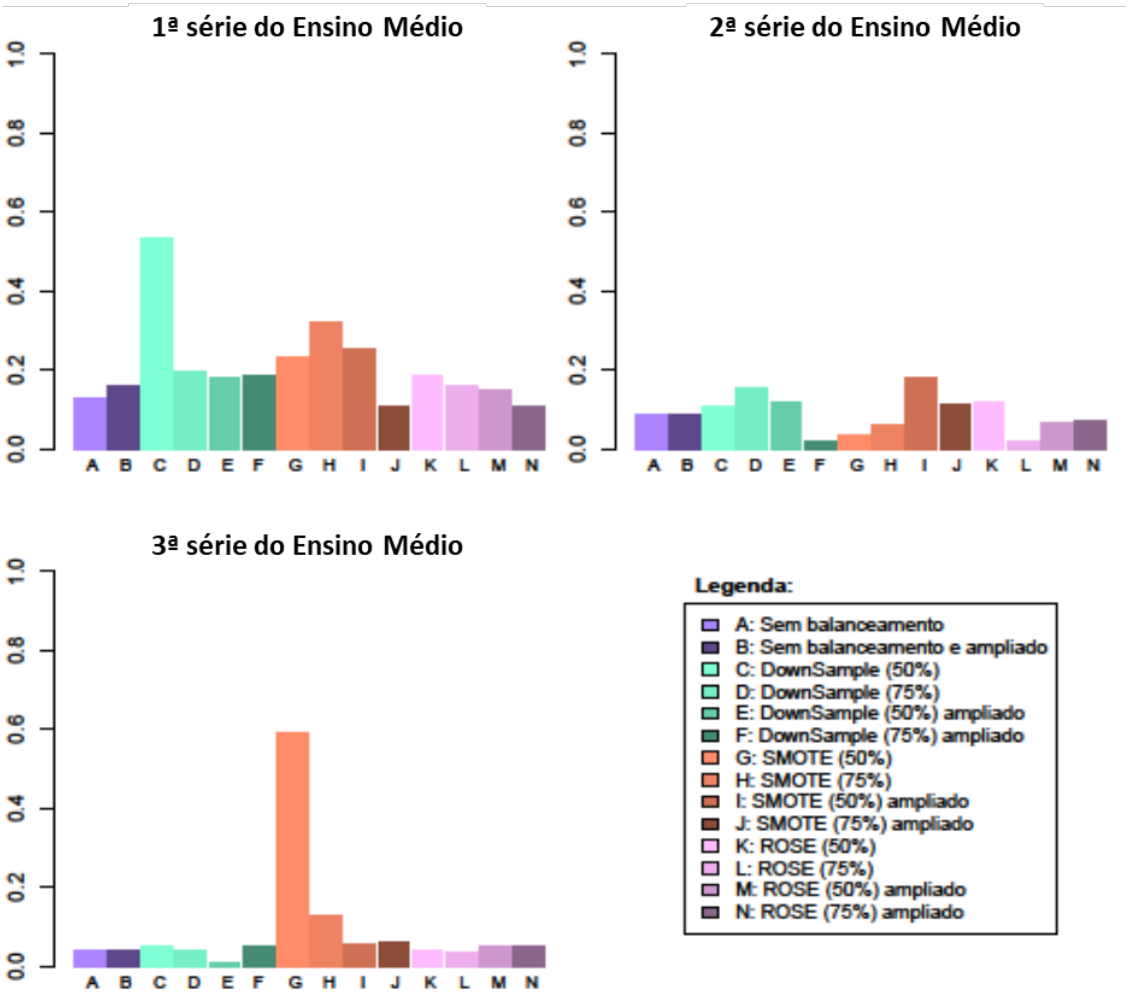


Elaboração: Estudos Educacionais/IJSN.
Nota: As barras indicam a especificidade encontrada. Quanto mais próximas de um, melhor o modelo. As barras foram categorizadas em tons de cores específicas. Os modelos estimados em dados sem balanceamento são representados por tons de roxo, as barras verdes indicam o Downsampling enquanto as barras laranjas representam as diferentes configurações de SMOTE. Por fim, a cor rosa foi reservada para o ROSE.

Os resultados para a precisão são apresentados por meio da Figura 10. Para a 1ª série, o *Downsampling* melhorou o desempenho para todas as configurações avaliadas. A configuração *C* apresentou significativas melhoras. Além desta, três das quatro configurações do SMOTE também apresentaram resultados superiores aos métodos sem balanceamento. Tendo em vista a 2ª série, algumas configurações melhoraram marginalmente a performance preditiva. As únicas configurações com melhorias

significativas foram o SMOTE 50% ampliado e *Downsampling* 75%. Para a 3ª série, a configuração SMOTE 50% apresentou significativas melhorias.

Figura 10 – Precisão

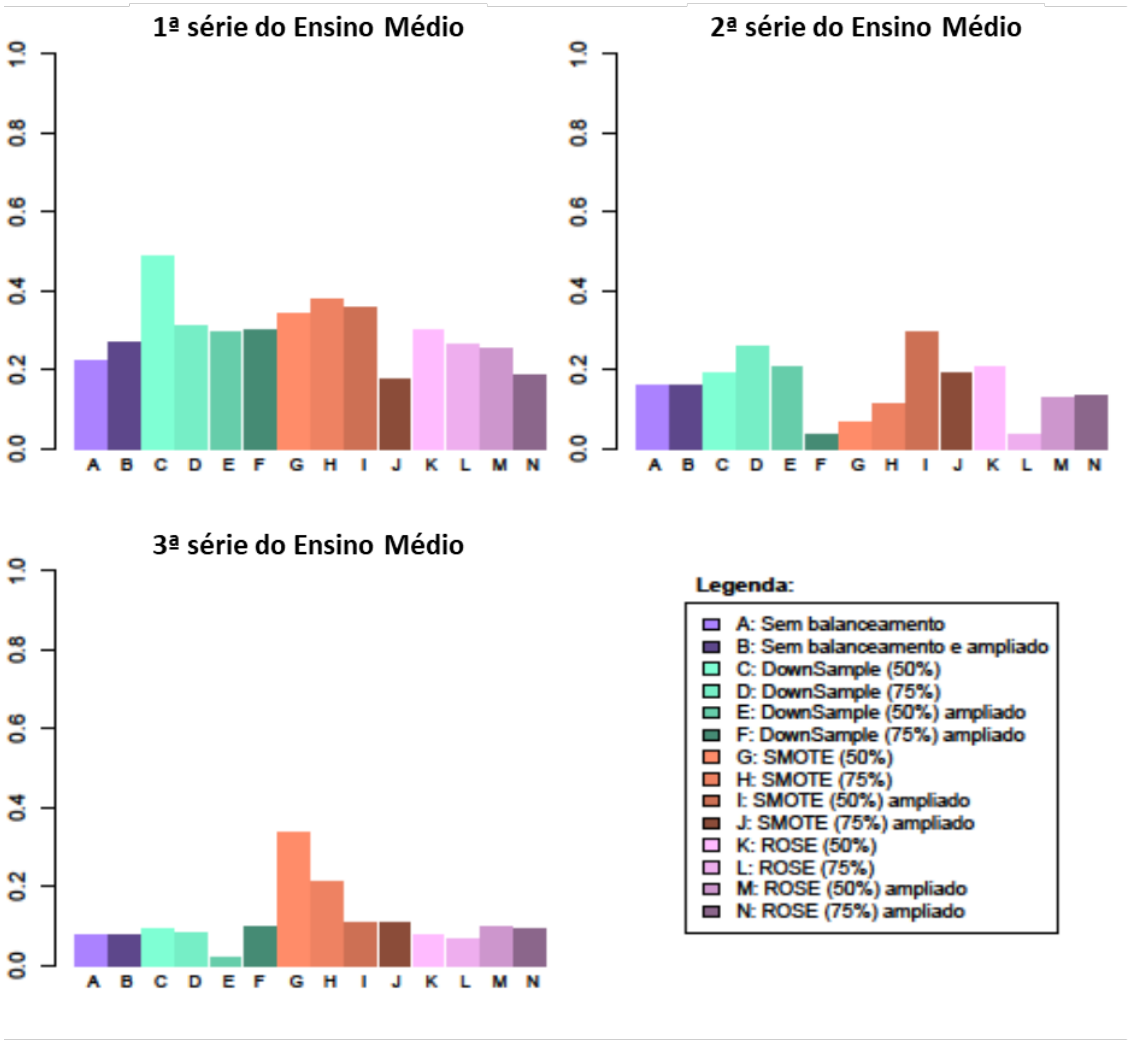


Elaboração: Estudos Educacionais/IJSN.
Nota: As barras indicam a precisão encontrada. Quanto mais próximas de um, melhor o modelo. As barras foram categorizadas em tons de cores específicas. Os modelos estimados em dados sem balanceamento são representados por tons de roxo, as barras verdes indicam o Downsampling enquanto as barras laranjas representam as diferentes configurações de SMOTE. Por fim, a cor rosa foi reservada para o ROSE.

Por fim, a Figura 11 apresenta a métrica F1–Score. É possível perceber que para a 1ª série, a maioria dos métodos melhorou o desempenho preditivo, com destaque para o *Downsampling* 50%. Para a 2ª série, algumas configurações melhoraram marginalmente a performance preditiva, sendo as únicas configurações com melhorias significativas o SMOTE 50% ampliado e *Downsampling* 75%. Ao analisarmos a 3ª série, é possível notar

que a configuração SMOTE 50% apresentou significativas melhoras. Em todos os outros casos, os desempenhos foram similares. Por outro lado, a configuração *Downsampling* 50% ampliado piorou o desempenho do modelo.

Figura 11 – F1-Score



Elaboração: Estudos Educacionais/IJSN.
Nota: As barras indicam o F1-Score obtido. Quanto mais próximas de um, melhor o modelo. As barras foram categorizadas em tons de cores específicas. Os modelos estimados em dados sem balanceamento são representados por tons de roxo, as barras verdes indicam o Downsampling enquanto as barras laranjas representam as diferentes configurações de SMOTE. Por fim, a cor rosa foi reservada para o ROSE.

Os melhores resultados estão compilados na Tabela 3. Os métodos de balanceamento apresentaram as melhores performances em todas as séries e métricas analisadas, sendo a única exceção a métrica sensibilidade para a 1ª série do ensino médio.

Além disso, embora o método ROSE seja capaz de melhorar o desempenho preditivo em certos contextos, os resultados indicam a superioridade dos métodos *Downsampling* e/ou SMOTE. No geral, as configurações com balanceamento de 50% foram as melhores, sendo as segundas melhores configurações obtidas em amostras balanceadas na proporção 25%-75%.

Os resultados indicam que o balanceamento é uma ferramenta útil para melhorar o desempenho preditivo. Contudo, não é possível estabelecer um método global para todas as séries e métricas. Além disso, recomenda-se cautela em seu emprego. Análises são necessárias sempre que o método for utilizado, uma vez que em certos (e poucos) casos o balanceamento piorou o desempenho preditivo dos modelos.

Tabela 3 – Melhores configurações por série do ensino médio

Métrica	1ª série		2ª série		3ª série	
	1ª posição	2ª posição	1ª posição	2ª posição	1ª posição	2ª posição
Acurácia	Downsampling – 50%	SMOTE – 75%	SMOTE – 50% ampl.	Downsampling – 75%	SMOTE – 50%	SMOTE – 75%
Sensibilidade	Sem bal.	ROSE – 50% ampl.	Downsampling – 75% ampl.	ROSE – 75%	Downsampling – 50% ampl.	Sem bal. – ampl.
Especificidade	Downsampling – 50%	SMOTE – 75%	SMOTE – 50% ampl.	Downsampling – 75%	SMOTE – 50%	SMOTE – 75%
Precisão	Downsampling – 50%	SMOTE – 75%	SMOTE – 50% ampl.	Downsampling – 75%	SMOTE – 50%	SMOTE – 75%
F1-Score	Downsampling – 50%	SMOTE – 75%	SMOTE – 50% ampl.	Downsampling – 75%	SMOTE – 50%	SMOTE – 75%

Elaboração: Estudos Educacionais/IJSN.

4.3. Estudo de caso II – Modelos adicionais de *machine learning*

Além dos métodos de balanceamento utilizados simultaneamente com a regressão logística, nesta seção testamos diferentes classificadores e os comparamos com a regressão logística. Assim, empregamos os seguintes modelos adicionais de *machine learning*: i) *Random Forest*; ii) *C5.0*; iii) *Support Vector Machine* (SVM); iv) Redes Neurais Artificias (RNA); e v) *Naive Bayes*.

De forma similar ao primeiro estudo de caso, comparamos os resultados desses modelos nos cenários de dados balanceados e dados desbalanceados. Neste segundo estudo de caso, apenas o balanceamento via SMOTE foi considerado, uma vez que os resultados anteriores indicam sua adequação e, além disso, pelo fato de ser considerado o padrão-ouro para balanceamento da amostra.

As etapas de construção do modelo, bem como as variáveis presentes no banco de dados, seguem as apresentadas no estudo de caso I. Contudo, para o treinamento dos modelos consideramos o ano de 2018 e as previsões foram realizadas para o ano de 2019. Além disso, realizamos apenas previsões para a 1ª série do ensino médio, por ser a série com maior abandono. Dessa forma, para cada modelo, utilizamos o SMOTE[13] para balancear a amostra e comparamos os resultados com os mesmos modelos estimados em amostras desbalanceadas.

Os hiperparâmetros dos modelos foram definidos via *10-fold cross-validation* e *grid search* (JAMES et al., 2013). A métrica considerada para seleção do melhor modelo foi a área sob a curva característica de operação do receptor (*Receiver Operating Characteristic Curve - ROC curve*). Todos os modelos foram implementados via pacote *Caret* (Classification and regression Training), desenvolvido por (KUHN, 2008). Além disso, utilizamos um Intel(R) Core (TM) i7-10700 CPU @2.90GHz com 16GB de RAM com processamento paralelo via o pacote *doParallel.*

A Tabela 4 apresenta as métricas consideradas, bem como os respectivos tempos computacionais. Como pode ser observado, o custo computacional varia consideravelmente, sendo os modelos *Random Forest* e SVM os mais demorados, cerca de 2 horas e 54 minutos e 2 horas e 41 minutos, respectivamente. Além desses, o *C5.0* e a RNA possuem tempos significativos de processamento. É importante destacar que o balanceamento reduz consideravelmente o tempo de estimação dos modelos, uma vez que a amostra balanceada possui menos observações.

[13] As amostras, neste estudo de caso, foram equilibradas na proporção 50%-50%.

Por outro lado, a regressão logística e o *Naive Bayes* são os modelos com o menor tempo computacional, tanto em amostras balanceadas quanto em amostras não balanceadas.

Se avaliarmos pela métrica F1-Score, métrica calculada a partir da precisão e sensibilidade, perceber-se que o balanceamento melhorou o desempenho preditivo de todos os modelos, quando comparados aos respectivos pares. A única exceção ocorre como o *Naive Bayes* onde o balanceamento possui resultado próximo ao não balanceamento.

Em termos de performance, os métodos baseados em *árvores de decisão* (*Random Forest* e *C5.0*) apresentaram os piores resultados, sendo os métodos cujo os modelos estimam probabilidades, isto é, regressão logística, *Naive Bayes* e RNA, os que apresentaram os melhores resultados.

Considerando a complexidade de estimação dos modelos, o que elava os custos computacionais, e os desempenhos preditivos, recomenda-se o uso da regressão logística. Além do bom desempenho preditivo, a regressão logística possui baixo custo computacional e seus parâmetros são interpretáveis, fazendo com que o usuário possa entender também as causas do abandono.

Tabela 4 – Comparação entre os diferentes modelos de classificação: resultados para a 1ª série do ensino médio

(continua)

Modelo	Acurácia	Sensibilidade	Especificidade	Precisão	F1-Score	Tempo (segundos)
Random Forest	0,978	0,047	0,998	0,363	0,084	10.479 seg.
***Random Forest* & SMOTE**	0,941	0,440	0,950	0,158	0,233	492 seg.
C5.0	0,962	0,219	0,978	0,175	0,195	3.751 seg.
C5.0 & SMOTE	0,947	0,338	0,961	0,154	0,211	271 seg.
SVM	0,528	0,644	0,526	0,027	0,053	9.697 seg.
SVM & SMOTE	0,914	0,525	0,922	0,123	0,200	553 seg.
RNA	0,884	0,549	0,891	0,096	0,161	2.468 seg.
RNA & SMOTE	0,898	0,523	0,906	0,105	0,175	506 seg.
Naive Bayes*	0,833	0,695	0,836	0,082	0,146	49 seg.

Tabela 4 – Comparação entre os diferentes modelos de classificação: resultados para a 1ª série do ensino médio

(continuação)

Modelo	Acurácia	Sensibilidade	Especificidade	Precisão	F1-Score	Tempo (segundos)
Naive Bayes & SMOTE*	0,835	0,675	0,838	0,081	0,144	12 seg.
Regressão logística	0,901	0,719	0,851	0,092	0,164	21 seg.
Regressão logística & SMOTE	0,901	0,596	0,907	0,119	0,199	6 seg.

Elaboração: Estudos Educacionais/IJSN.
Nota: (*) Estimados apenas com variáveis contínuas.

5. Conclusões

Este relatório foi composto por dois estudos de caso. No primeiro, testou-se diferentes métodos de balanceamento da amostra em conjunto com a regressão logística para verificarmos se tal abordagem é capaz de melhorar o desempenho preditivo da regressão logística estimada na amostra original (desbalanceada).

Os resultados indicam que o balanceamento é uma ferramenta útil para melhorar o desempenho preditivo. Contudo, não é possível estabelecer um método global que irá performar melhor em todos as séries e métricas. Além disso, recomenda-se cautela no momento de seu emprego. Análises são necessárias sempre que o método for empregado, uma vez que em certos (e poucos) casos o balanceamento piorou o desempenho preditivo dos modelos. Os métodos empregados reduzem a amostra majoritária por meio de descarte de observações sorteadas aleatoriamente. Dessa forma, tais métodos podem descartar observações importantes ao problema e, por isso, piorar o desempenho preditivo. Dessa forma, para estudos futuros, recomenda-se a avaliação de novas metodologias de balanceamento que não estejam sujeitas ao descarte de observações relevantes.

Tendo em vista o segundo estudo de caso, onde diversos modelos de *machine learning* foram testados em amostras balanceadas e desbalanceadas, os resultados indicam que os métodos baseados em árvores de decisão (*Random Forest* e *C5.0*) apresentaram os piores resultados. Ademais, os métodos cujos modelos estimam probabilidades, isto é,

regressão logística, *Naive Bayes* e Redes Neurais Artificiais (RNA), são os que apresentaram os melhores resultados. Além disso, o tempo computacional varia bastante entre os diferentes modelos, sendo a regressão logística e o *Naive Bayes* os mais rápidos.

Dessa forma, considerando a complexidade da estimação dos modelos, o que elava os custos computacionais, e os desempenhos preditivos, recomenda-se a uso da regressão logística. Além do bom desempenho preditivo, seus parâmetros são interpretáveis, fazendo com que o usuário possa entender também as razões do abandono.

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.

BARROS, Thiago M. et al. Predictive Models for Imbalanced Data: A School Dropout Perspective. **Education Sciences**, v. 9, n. 4, 2019.

BAYER, Jaroslav et al. Predicting dropout from social behaviour of students. **International Conference on Educational Data Mining (EDM)**. Chania, Greece, 2012.

BURGESS, Simon M. Human capital and education: The state of art in the economics of education. **IZA Discussion paper**, n. 9885, Bonn: Institute for the Study of Labour, 2016.

CHAWLA, et al. SMOTE: synthetic minority over-sampling technique. **Journal of artificial intelligence research**, v. 16, n. 1, 2002.

CHO, Choong H.; YU, Yang W.; KIM, HYEON G. A Study on Dropout Prediction for University Students Using Machine Learning. **Applied Sciences**, v. 13, n. 21, 2023.

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.

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.

DEL BONIFRO, Francesca et al. Student Dropout Prediction. **Artificial Intelligence in Education. AIED 2020. Lecture Notes in Computer Science**, por Bittencourt, Ig I. et al., Springer, 2020.

FERNÁNDEZ, Alberto et al. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. **Journal of artificial intelligence research**, v. 61, n. 1, p. 863-905, 2018.

FREITAS, Francisco A. da S. et al. IoT system for school dropout prediction using machine learning techniques based on socioeconomic data. **Electronics**, v. 9, n. 10, 2020.

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

HE, Haibo; GARCIA, Edwardo A. Learning from imbalanced data. **IEEE Transactions on knowledge and data engineering**, vol. 21, n. 9, 2009.

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

KIM, Sangyn et al. Student Dropout Prediction for University with High Precision and Recall. **Applied Sciences**, v. 13, n. 10, 2023.

KNOWLES, Jared E. Of needles and haystacks: Building and accurate statewide dropout early warnings system in Wisconsin. **Journal of Educational Data Mining**, vol. 3, n.3, p.18-67, 2015.

KUHN, Max. Building Predictive Models in R Using the caret Package. **Journal of Statistical Software**, v. 28, n. 5, p. 1-26, 2008.

KUHN, Max; JOHNSON, Kjell. **Applied Predictive Modeling**. Springer, 2013.

LEE, Sunbok; CHUNG, Jae Y. The Machine Learning-Based Dropout Early Warning System for Improving the Performance of Dropout Prediction. **Applied Sciences**, v. 9, n. 15, 2019.

MA, Yunqian; HE, Haibo. **Imbalanced learning: foundatinos, algorithms and applications**. IEEE Press - Wiley, 2013.

MARQUEZ-VERA, Carlos, et al. Early dropout prediction using data mining: a case study with high school students. **Expert Systems**, v. 33, n. 1, 2016.

MARTINHO, Valquíria R.C.; NUNES, Clodoaldo; MINUSSI, Carlos R. An intelligent system for prediction of school dropout risk group in higher education classroom based on artificial neural networks. **2013 IEEE 25th International Conference on Tools with Artificial Intelligence**, p. 159—166, 2013.

MENARDI, Giovanna; TORELLI, Nicola. Training and assessing classification rules with imbalanced data. **Data mining and knowledge discovery**, v. 28, p. 92-122, 2014.

OROOJI, Marmar; CHEN, Jianhua. Predicting Louisiana Public High School Dropout through Imbalanced Learning Techniques. ***2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA).*** **IEEE**, 2019.

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.

PSATHAS, Georgios; CHATZIDAKI, Theano K.; DEMETRIADIS, Stavros N. Learning, Predictive Modeling of Student Dropout in MOOCs and Self-Regulated. **Computers**, v. 12. N. 10, 2023.

PEREIRA, Guilherme A.A.; DEMURA, K.D; NUNES, I.C.; DE PAULA, K.C.; LIRA, P.S. An early warning system for school dropout in the State od Espírito Santo: a machine learning approach with Variable selection methods. **Pesquisa Operacional**, v. 44, 2024

RODRÍGUEZ, Patrício et al. Methodology to design, develop, and evaluate machine learning models for predicting dropout in school systems: the case of Chile. **Education and Information Technologies**, v. 28, pp- 10103-10149, 2023.

ROVIRA, Sergi; PUERTAS, Eloi; IGUAL, Laura. Data-driven system to predict academic grades and dropout. **PLoS one**, 2017.

SANDOVAL-PALIS, Iván. Early Dropout Prediction Model: A Case Study of University Leveling Course Students. **Sustainability**, v. 12, n. 22, 2020.

SARA, Nicole-Bogdan et al. High-School Dropout Prediction Using Machine Learning: A Danish Large-scale Study. **ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence.** Bruges (Belgium), 2015

SELIM, Kamal S.; REZK, Sahar Saeed. On predicting school dropouts in Egypt: A machine learning approach. **Education and Information Technologies**, v. 28, p. 9235-9266, 2023.

ULDALL, Jerome S.; ROJAS, Cristian G. An application of machine learning in public policy early warning prediction of school dropout in the chilean public education system. **Multidisciplinary Business Review**, v. 15, n. 1, 2022.

UNICEF. *Early Warning Systems for Students at Risk of Dropping out (UNICEF Series on Education Participation and Dropout Preention).* 2017.

VILLAR, Alice; ANDRADE, Carolina, R.V. Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study. **Discover Artificial Intelligence**, 2024.

WONGVORACHAN, Tarid; HE, Surina; BULUT, Okan. A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. **Information**, v. 14, n. 1, 2023.

WOOD, Laura et al. Predicting dropout using student-and school-level factors: An ecological perspective." **School Psychology Quarterly**, v. 32, n. 1, p. 35-49, 2017.

APÊNDICE

Apêndice A – Métricas para as avaliações das previsões

A Tabela 5 apresenta a matriz de confusão na qual as métricas utilizadas nesse trabalho são utilizadas.

Tabela 5 – 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 nesta matriz, as seguintes métricas podem ser definidas:

i. Acurácia: definida como o percentual de previsões corretamente realizadas e matematicamente é estimada como (VN + VP)/(VN + VP + FN + VP).
ii. Sensibilidade: dada por VP/(FN+VP) e 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).
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-Score: definida como 2VP/(2VP + FP + FN). Esta pode ser entendida como uma espécie de média entre precisão e sensibilidade. A precisão mensura quantas previsões da classe “abandono (1)” foram feitas corretamente. Por outro lado, a sensibilidade mensura quantas instâncias da classe “abandono (1)” presente no banco de dados foram identificadas corretamente pelo modelo. Tanto a precisão quanto a sensibilidade possuem um trade-off inerente em suas formulações. Por exemplo, a sensibilidade de um modelo pode ser artificialmente aumentada caso haja um aumento do número de previsões da classe “abandono (1)”. Consequente, esse aumento artificial acarreta num aumento do número de falsos positivos (FP), que compõe parte do denominador da métrica precisão. Dessa forma, um aumento da sensibilidade geralmente implica numa redução da precisão. A métrica F1 combina ambos, de modo que maximizar F1 significa maximizar a precisão e a sensibilidade simultaneamente.